

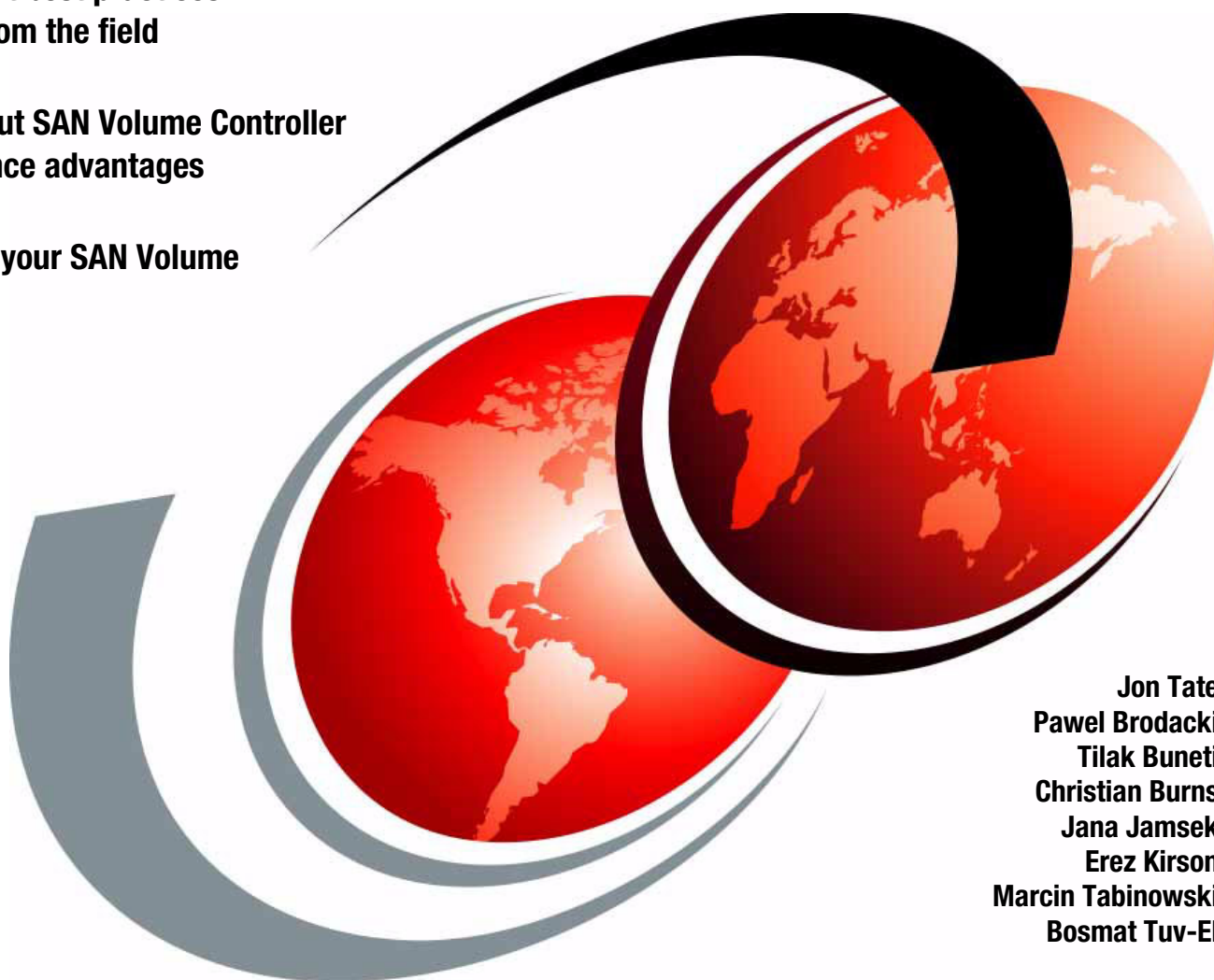
IBM System Storage SAN Volume Controller and Storwize V7000

Best Practices and Performance Guidelines

Read about best practices
learned from the field

Learn about SAN Volume Controller
performance advantages

Fine-tune your SAN Volume
Controller



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

Redbooks



International Technical Support Organization

Best Practices and Performance Guidelines

September 2014

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

Fourth Edition (September 2014)

This edition applies to Version 7, Release 2, of the IBM System Storage SAN Volume Controller and Storwize V7000.

© Copyright International Business Machines Corporation 2008, 2014. 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	xv
Trademarks	xvi
IBM Redbooks promotions	xvii
Preface	xix
The team who wrote this book	xix
Now you can become a published author, too!	xxii
Comments welcome.	xxii
Stay connected to IBM Redbooks	xxiii
Summary of changes	xxv
September 2014, Fourth Edition	xxv
Part 1. Configuration guidelines and preferred practices	1
Chapter 1. Updates in IBM System Storage SAN Volume Controller	3
1.1 Enhancements and changes in SAN Volume Controller V5.1	4
1.2 Enhancements and changes in SAN Volume Controller V6.1	5
1.3 Enhancements and changes in SAN Volume Controller V6.2	7
1.4 Enhancements and changes in SAN Volume Controller V6.3	9
1.5 Enhancements and changes in SAN Volume Controller V6.4	11
1.6 Enhancements and changes in SAN Volume Controller V7.1	12
1.7 Enhancements and changes in SAN Volume Controller V7.2	14
Chapter 2. SAN topology	17
2.1 SAN topology of the SAN Volume Controller/Storwize	18
2.1.1 Redundancy	18
2.1.2 Topology basics	19
2.1.3 ISL oversubscription	20
2.1.4 Single switch SAN Volume Controller/Storwize SANs	21
2.1.5 Basic core-edge topology	21
2.1.6 Four-SAN, core-edge topology	22
2.1.7 Common topology issues	24
2.1.8 Stretched Cluster	27
2.1.9 Enhanced Stretched Cluster	30
2.2 SAN switches	31
2.2.1 Selecting SAN switch models	31
2.2.2 Switch port layout for large SAN edge switches	31
2.2.3 Switch port layout for director-class SAN switches	32
2.2.4 Virtual channels	32
2.2.5 IBM System Storage and IBM b-type SANs	35
2.2.6 IBM System Storage and Cisco SANs	37
2.2.7 SAN routing and duplicate worldwide node names	38
2.3 Zoning	38
2.3.1 Types of zoning	39
2.3.2 Prezoning tips and shortcuts	40
2.3.3 SAN Volume Controller internode communications zone	41
2.3.4 SAN Volume Controller/Storwize storage zones	41

2.3.5	SAN Volume Controller/Storwize host zones	44
2.3.6	Standard SAN Volume Controller/Storwize zoning configuration	46
2.3.7	Zoning with multiple SAN Volume Controller/Storwize clustered systems	50
2.3.8	Split storage subsystem configurations.	50
2.4	Switch domain IDs	50
2.5	Distance extension for remote copy services	50
2.5.1	Optical multiplexors.	51
2.5.2	Long-distance SFPs or XFPs	51
2.5.3	Fibre Channel over IP	51
2.5.4	Native IP replication with 7.2 SAN Volume Controller/Storwize code version . . .	53
2.6	Tape and disk traffic that share the SAN	53
2.7	Switch interoperability	53
2.8	IBM Tivoli Storage Productivity Center	54
2.9	iSCSI support	54
2.9.1	iSCSI initiators and targets	54
2.9.2	iSCSI Ethernet configuration.	55
2.9.3	Security and performance.	55
2.9.4	Failover of port IP addresses and iSCSI names	55
2.9.5	iSCSI protocol limitations	56
2.10	SAS support	56
Chapter 3.	SAN Volume Controller and Storwize V7000 Cluster	59
3.1	Advantages of virtualization	60
3.1.1	SAN Volume Controller features.	61
3.2	Scalability of SAN Volume Controller clustered systems	61
3.3	Scalability of Storwize V7000	62
3.3.1	Advantage of multiclustered systems versus single-clustered systems	62
3.3.2	Growing or splitting SAN Volume Controller clustered systems	64
3.3.3	Adding or upgrading SAN Volume Controller node hardware.	67
3.4	Clustered system upgrade	68
Chapter 4.	Back-end storage	71
4.1	Controller affinity and preferred path.	72
4.2	Round Robin Path Selection.	73
4.3	Considerations for DS4000 and DS5000 series	74
4.3.1	Setting the DS4000 and DS5000 series so that both controllers have the same worldwide node name.	74
4.3.2	Balancing workload across DS4000 and DS5000 series controllers	74
4.3.3	Ensuring path balance before MDisk discovery	75
4.3.4	Auto-Logical Drive Transfer for the DS4000 and DS5000 series (firmware version before 7.83.x)	75
4.3.5	Asymmetric Logical Unit Access for the DS4000 and DS5000 series (firmware 7.83.x and later).	76
4.3.6	Selecting array and cache parameters	76
4.3.7	Logical drive mapping	78
4.4	Considerations for DS8000 series	78
4.4.1	Balancing workload across DS8000 series controllers	78
4.4.2	DS8000 series ranks to extent pools mapping	79
4.4.3	Mixing array sizes within a storage pool	80
4.4.4	Determining the number of controller ports for the DS8000 series.	81
4.4.5	LUN masking	81
4.4.6	WWPN to physical port translation	83
4.5	Considerations for IBM XIV Storage System	83

4.5.1 Cabling considerations	83
4.5.2 Host options and settings for XIV systems	85
4.5.3 Number and size of the Volumes	86
4.5.4 Restrictions	87
4.6 Considerations for IBM Storwize V7000/V5000/V3700	87
4.6.1 Cabling and zoning	87
4.6.2 Defining internal storage	88
4.6.3 Configuring Storwize storage systems	88
4.7 Considerations for IBM FlashSystem	89
4.7.1 Physical FC port connection and zoning	89
4.7.2 Logical configuration	91
4.7.3 Extent size and storage pools	91
4.7.4 Volumes	91
4.8 Considerations for third-party storage with EMC Symmetrix DMX and Hitachi Data Systems	92
4.9 Medium error logging	93
4.10 Mapping physical LBAs to volume extents	93
4.11 Identifying storage controller boundaries by using the IBM Tivoli Storage Productivity Center	93
Chapter 5. Storage pools and managed disks	95
5.1 Availability considerations for storage pools	96
5.2 Selecting storage subsystems	96
5.3 Selecting the storage pool	97
5.3.1 Selecting the number of arrays per storage pool	97
5.3.2 Selecting LUN attributes	99
5.3.3 Considerations for Storwize family systems	100
5.3.4 Considerations for the IBM XIV Storage System	100
5.4 Quorum disk considerations	102
5.5 Tiered storage	104
5.6 Adding MDisks to existing storage pools	106
5.6.1 Checking access to new MDisks	106
5.6.2 Persistent reserve	106
5.6.3 Renaming MDisks	107
5.7 Rebalancing extents across a storage pool	107
5.7.1 Installing prerequisites and the SAN Volume ControllerTools package	108
5.7.2 Running the extent balancing script	108
5.8 Removing MDisks from existing storage pools	111
5.8.1 Migrating extents from the MDisk to be deleted	111
5.8.2 Verifying the identity of an MDisk before removal	112
5.8.3 Correlating the back-end volume with the MDisk	112
5.9 Remapping managed MDisks	120
5.10 Controlling extent allocation order for volume creation	121
5.11 Moving an MDisk between SAN Volume Controller clusters	122
5.12 MDisk group considerations when Real-time Compression is used	124
Chapter 6. Volumes	125
6.1 Overview of volumes	126
6.1.1 Striping compared to sequential type	126
6.1.2 Thin-provisioned volumes	126
6.1.3 Space allocation	127
6.1.4 Compressed volumes	127
6.1.5 Thin-provisioned volume	127

6.1.6	Limits on virtual capacity of thin-provisioned volumes	128
6.1.7	Testing an application with a thin-provisioned volume	128
6.2	Volume mirroring	129
6.2.1	Creating or adding a mirrored volume	129
6.2.2	Availability of mirrored volumes	129
6.2.3	Mirroring between controllers	130
6.3	Creating volumes	130
6.3.1	Selecting the storage pool	131
6.3.2	Changing the preferred node within an I/O group	132
6.3.3	Non-Disruptive volume move	133
6.4	Volume migration	135
6.4.1	Image-type to striped-type migration	136
6.4.2	Migrating to image-type volume	137
6.4.3	Migrating with volume mirroring	138
6.5	Preferred paths to a volume	138
6.5.1	Governing of volumes	139
6.6	Cache mode and cache-disabled volumes	141
6.6.1	Underlying controller remote copy with SAN Volume Controller cache-disabled volumes	142
6.6.2	Using underlying controller FlashCopy with SAN Volume Controller cache disabled volumes	143
6.6.3	Changing the cache mode of a volume	143
6.7	Effect of a load on storage controllers	145
6.8	Setting up FlashCopy services	146
6.8.1	Making a FlashCopy volume with application data integrity	147
6.8.2	Making multiple related FlashCopy volumes with data integrity	149
6.8.3	Creating multiple identical copies of a volume	151
6.8.4	Creating a FlashCopy mapping with the incremental flag	151
6.8.5	Using thin-provisioned FlashCopy	152
6.8.6	Using FlashCopy with your backup application	152
6.8.7	Migrating data by using FlashCopy	153
6.8.8	Summary of FlashCopy rules	154
6.8.9	IBM Tivoli Storage FlashCopy Manager	155
6.8.10	IBM System Storage Support for Microsoft Volume Shadow Copy Service	155
Chapter 7	Remote copy services	157
7.1	Introduction to remote copy services	158
7.1.1	Common terminology and definitions	159
7.1.2	Intercluster link	161
7.2	SAN Volume Controller remote copy functions by release	161
7.2.1	Remote copy in SAN Volume Controller V7.2	161
7.2.2	Dual physical links with active/standby for use in two or more I/O groups environments	164
7.2.3	Remote copy features by release	165
7.3	Terminology and functional concepts	171
7.3.1	Remote copy partnerships and relationships	172
7.3.2	Global Mirror control parameters	172
7.3.3	Global Mirror partnerships and relationships	173
7.3.4	Asynchronous remote copy	175
7.3.5	Understanding remote copy write operations	175
7.3.6	Asynchronous remote copy	176
7.3.7	Global Mirror write sequence	177
7.3.8	Write ordering	177

7.3.9 Colliding writes	177
7.3.10 Link speed, latency, and bandwidth	178
7.3.11 Choosing a link cable of supporting Global Mirror applications	180
7.3.12 Remote copy volumes: Copy directions and default roles	180
7.4 Intercluster link	181
7.4.1 SAN configuration overview	181
7.4.2 Switches and ISL oversubscription	182
7.4.3 Zoning	183
7.4.4 Distance extensions for the intercluster link	183
7.4.5 Optical multiplexors	184
7.4.6 Long-distance SFPs and XFPs	184
7.4.7 Fibre Channel IP conversion	184
7.4.8 Configuration of intercluster links	184
7.4.9 Link quality	185
7.4.10 Hops	186
7.4.11 Buffer credits	187
7.5 Global Mirror design points	188
7.5.1 Global Mirror parameters	189
7.5.2 The chcluster and chpartnership commands	190
7.5.3 Distribution of Global Mirror bandwidth	190
7.5.4 1920 errors	194
7.6 Global Mirror planning	194
7.6.1 Rules for using Metro Mirror and Global Mirror	194
7.6.2 Planning overview	195
7.6.3 Planning specifics	196
7.7 Global Mirror use cases	198
7.7.1 Synchronizing a remote copy relationship	198
7.7.2 Global Mirror relationships, saving bandwidth, and resizing volumes	199
7.7.3 Master and auxiliary volumes and switching their roles	200
7.7.4 Migrating a Metro Mirror relationship to Global Mirror	201
7.7.5 Multicluster mirroring	201
7.7.6 Performing three-way copy service functions	205
7.7.7 When to use storage controller Advanced Copy Services functions	207
7.7.8 Using Metro Mirror or Global Mirror with FlashCopy	207
7.7.9 Global Mirror upgrade scenarios	208
7.8 Intercluster Metro Mirror and Global Mirror source as an FC target	209
7.9 States and steps in the Global Mirror relationship	211
7.9.1 Global Mirror states	212
7.9.2 Disaster recovery and Metro Mirror and Global Mirror states	214
7.9.3 State definitions	214
7.10 1920 errors	216
7.10.1 Diagnosing and fixing 1920 errors	216
7.10.2 Focus areas for 1920 errors	217
7.10.3 Recovery	221
7.10.4 Disabling the glintolerance feature	222
7.10.5 Cluster error code 1920 checklist for diagnosis	223
7.11 Monitoring remote copy relationships	223
Chapter 8. Hosts	225
8.1 Configuration guidelines	226
8.1.1 Host levels and host object name	226
8.1.2 The number of paths	226
8.1.3 Host ports	227

8.1.4	Port masking	228
8.1.5	Host to I/O group mapping	228
8.1.6	Volume size as opposed to quantity	228
8.1.7	Host volume mapping	229
8.1.8	Server adapter layout	232
8.1.9	Availability versus error isolation.	233
8.2	Host pathing	233
8.2.1	Preferred path algorithm	233
8.2.2	Path selection	234
8.2.3	Path management.	234
8.2.4	Dynamic reconfiguration	235
8.2.5	Nondisruptive Volume migration between I/O groups.	237
8.3	I/O queues.	240
8.3.1	Queue depths	240
8.4	Multipathing software	242
8.5	Host clustering and reserves.	242
8.5.1	Clearing reserves	243
8.5.2	SAN Volume Controller MDisk reserves	244
8.6	AIX hosts.	244
8.6.1	HBA parameters for performance tuning	244
8.6.2	Configuring for fast fail and dynamic tracking	246
8.6.3	Multipathing.	246
8.6.4	SDD	247
8.6.5	SDDPCM	247
8.6.6	SDD compared to SDDPCM.	248
8.7	Virtual I/O Server.	249
8.7.1	Methods to identify a disk for use as a virtual SCSI disk	250
8.7.2	UDID method for MPIO.	250
8.7.3	Backing up the virtual I/O configuration	251
8.8	Windows hosts	251
8.8.1	Clustering and reserves	251
8.8.2	Tunable parameters	252
8.8.3	Changing back-end storage LUN mappings dynamically	252
8.8.4	Guidelines for disk alignment by using Windows with SAN Volume Controller volumes	252
8.9	Linux hosts	252
8.9.1	SDD compared to DM-MPIO.	253
8.9.2	Tunable parameters	253
8.9.3	I/O Scheduler	253
8.10	Solaris hosts	254
8.10.1	Solaris MPxIO	254
8.10.2	Symantec Veritas Volume Manager	254
8.10.3	ASL specifics for SAN Volume Controller.	255
8.10.4	SDD pass-through multipathing	255
8.10.5	DMP multipathing	256
8.10.6	Troubleshooting configuration issues	256
8.11	VMware server	256
8.11.1	Multipathing solutions supported	257
8.11.2	Multipathing configuration maximums.	257
8.12	Mirroring considerations	258
8.12.1	Host-based mirroring.	258
8.13	Monitoring	258
8.13.1	Automated path monitoring.	259

8.13.2 Load measurement and stress tools	260
Part 2. Performance preferred practices.	261
Chapter 9. Performance highlights for SAN Volume Controller V7.2	263
9.1 SAN Volume Controller continuing performance enhancements	264
9.2 FlashSystem 820 Performance	265
9.3 Solid-State Drives and Easy Tier	266
9.3.1 Internal SSD redundancy	267
9.3.2 Performance scalability and I/O groups	267
9.4 Real-Time Performance Monitor	268
Chapter 10. Back-end storage performance considerations	269
10.1 Workload considerations	270
10.2 Tiering	271
10.3 Storage controller considerations	271
10.3.1 Back-end I/O capacity	272
10.4 Array considerations	282
10.4.1 Selecting the number of LUNs per array	282
10.4.2 Selecting the number of arrays per storage pool	283
10.5 I/O ports, cache, and throughput considerations	284
10.5.1 Back-end queue depth	284
10.5.2 MDisk transfer size	285
10.6 SAN Volume Controller extent size	288
10.7 SAN Volume Controller cache partitioning	290
10.8 IBM DS8000 series considerations	291
10.8.1 Volume layout	291
10.8.2 Cache	296
10.8.3 Determining the number of controller ports for DS8000 series	296
10.8.4 Storage pool layout	298
10.8.5 Extent size	303
10.9 IBM XIV considerations	303
10.9.1 LUN size	303
10.9.2 I/O ports	305
10.9.3 Storage pool layout	306
10.9.4 Extent size	307
10.10 Storwize V7000 considerations	307
10.10.1 Volume setup	307
10.10.2 I/O ports	310
10.10.3 Storage pool layout	312
10.10.4 Extent size	314
10.11 DS5000 series considerations	314
10.11.1 Selecting array and cache parameters	314
10.11.2 Considerations for controller configuration	316
10.11.3 Mixing array sizes within the storage pool	317
10.11.4 Determining the number of controller ports for DS4000	317
10.11.5 Performance considerations with FlashSystem	317
Chapter 11. IBM System Storage Easy Tier function.	319
11.1 Overview of Easy Tier	320
11.2 Easy Tier concepts	320
11.2.1 SSD arrays and MDisk	320
11.2.2 Disk tiers	321
11.2.3 Single tier storage pools	321

11.2.4	Multitier storage pools	322
11.2.5	Easy Tier process	322
11.2.6	Easy Tier operating modes	323
11.2.7	Easy Tier activation	324
11.3	Easy Tier implementation considerations	324
11.3.1	Prerequisites	324
11.3.2	Implementation rules	325
11.3.3	Easy Tier limitations	325
11.4	Measuring and activating Easy Tier	326
11.4.1	Measuring by using the Storage Advisor Tool	326
11.5	Activating Easy Tier with the SAN Volume Controller CLI	329
11.5.1	Initial cluster status	329
11.5.2	Turning on Easy Tier evaluation mode	330
11.5.3	Creating a multitier storage pool	331
11.5.4	Setting the disk tier	333
11.5.5	Checking the Easy Tier mode of a volume	334
11.5.6	Final cluster status	334
11.6	Activating Easy Tier with the SAN Volume Controller GUI	335
11.6.1	Setting the disk tier on MDisks	335
11.6.2	Checking Easy Tier status	338
Chapter 12.	Applications	339
12.1	Application workloads	340
12.1.1	Transaction-based workloads	340
12.1.2	Throughput-based workloads	341
12.1.3	Storage subsystem considerations	341
12.1.4	Host considerations	341
12.2	Application considerations	341
12.2.1	Transaction environments	342
12.2.2	Throughput environments	342
12.2.3	Performance tuning	343
12.3	Data layout overview	343
12.3.1	Storage virtualization layers	343
12.3.2	Virtualized storage characteristics	344
12.3.3	Storage, OS, and application administrator roles	345
12.3.4	General data layout guidelines	346
12.3.5	Throughput workloads	348
12.3.6	LVM volume groups and logical volumes	349
12.4	Database storage	350
12.5	Data layout with the AIX Virtual I/O Server	350
12.5.1	Overview	350
12.5.2	Data layout strategies	351
12.6	Volume size	351
12.7	Failure domains	352
12.8	More resources	352
12.8.1	IBM System Storage Interoperation Center	352
12.8.2	Techdocs - the Technical Sales Library	352
12.8.3	DB2 white papers	353
12.8.4	Oracle white papers	354
12.8.5	Diskcore and Tapecore mailing lists	354

Part 3. Management, monitoring, and troubleshooting 355

Chapter 13. Monitoring 357

13.1 Analyzing the SAN Volume Controller and Storwize Family Storage Systems by using Tivoli Storage Productivity Center	359
13.1.1 Analyzing with the Tivoli Storage Productivity Center 5.2 web-based GUI. . . .	359
13.2 Considerations for performance analysis	380
13.2.1 SAN Volume Controller considerations.	380
13.2.2 Storwize V7000 considerations.	381
13.3 Top 10 reports for SAN Volume Controller and Storwize V7000	382
13.3.1 I/O Group Performance for SAN Volume Controller and Storwize V7000	384
13.3.2 Node Cache Performance for SAN Volume Controller and Storwize V7000 . .	400
13.3.3 Viewing the Managed Disk Group Performance report for SAN Volume Controller by using the stand-alone GUI.	413
13.3.4 Top Volume Performance reports.	422
13.3.5 Port Performance reports for SAN Volume Controller and Storwize V7000 . .	433
13.4 Reports for fabric and switches.	439
13.4.1 Switches reports	439
13.4.2 Switch Port Data Rate Performance.	440
13.5 Case studies	441
13.5.1 Server performance problem	442
13.5.2 Disk performance problem in a Storwize V7000 subsystem.	446
13.5.3 Top volumes response time and I/O rate performance reports.	455
13.5.4 Performance constraint alerts	457
13.5.5 Monitoring and diagnosing performance problems for a fabric Viewing	465
13.5.6 Verifying the SAN Volume Controller and Fabric configuration by using Topology Viewer.	470
13.5.7 Verifying the SAN Volume Controller and Fabric configuration by using the Tivoli Storage Productivity Center 5.2 web-based GUI Data Path tools	475
13.6 Monitoring in real time by using the SAN Volume Controller or Storwize V7000 GUI	477
13.7 Manually gathering SAN Volume Controller statistics.	479
Chapter 14. Maintenance	485
14.1 Automating the documentation for SAN Volume Controller/Storwize and SAN environment	486
14.1.1 Naming conventions	486
14.1.2 SAN fabrics documentation	489
14.1.3 SAN Volume Controller and Storwize family products	491
14.1.4 Storage	492
14.1.5 Technical Support information	492
14.1.6 Tracking incident and change tickets	493
14.1.7 Automated support data collection	494
14.1.8 Subscribing to SAN Volume Controller/Storwize support.	494
14.2 Storage management IDs	495
14.3 Standard operating procedures.	496
14.3.1 Allocating and deallocating volumes to hosts	496
14.3.2 Adding and removing hosts in SAN Volume Controller/Storwize	497
14.4 SAN Volume Controller/Storwize code upgrade	497
14.4.1 Preparing for the upgrade.	498
14.4.2 SAN Volume Controller upgrade from V5.1 to V6.2	504
14.4.3 Upgrading SAN Volume Controller clusters/Storwize systems that are participating in Metro Mirror or Global Mirror	506
14.4.4 SAN Volume Controller/Storwize upgrade	507
14.4.5 Storwize family systems disk drive upgrade	507
14.5 SAN modifications.	509
14.5.1 Cross-referencing HBA WWPNS.	509

14.5.2	Cross-referencing LUN IDs	511
14.5.3	HBA replacement	511
14.6	Hardware upgrades for SAN Volume Controller	512
14.6.1	Adding SAN Volume Controller nodes to an existing cluster	513
14.6.2	Upgrading SAN Volume Controller nodes in an existing cluster	514
14.6.3	Moving to a new SAN Volume Controller cluster	514
14.7	Adding expansion enclosures to Storwize family systems	515
14.8	More information	517
Chapter 15.	Troubleshooting and diagnostics	519
15.1	Common problems	520
15.1.1	Host problems	520
15.1.2	SAN Volume Controller problems	520
15.1.3	SAN problems	522
15.1.4	Storage subsystem problems	522
15.2	Collecting data and isolating the problem	524
15.2.1	Host data collection	524
15.2.2	SAN Volume Controller data collection	527
15.2.3	SAN data collection	532
15.2.4	Storage subsystem data collection	536
15.3	Recovering from problems	539
15.3.1	Solving host problems	540
15.3.2	Solving SAN Volume Controller problems	541
15.3.3	Solving SAN problems	544
15.3.4	Solving back-end storage problems	545
15.4	Mapping physical LBAs to volume extents	549
15.4.1	Investigating a medium error by using lsvdisklba	549
15.4.2	Investigating thin-provisioned volume allocation by using lsmdisklba	549
15.5	Medium error logging	550
15.5.1	Host-encountered media errors	550
15.5.2	SAN Volume Controller-encountered medium errors	551
15.5.3	Replacing a bad disk	552
15.5.4	Health status during upgrade	552

Part 4. Practical examples 553

Chapter 16.	SAN Volume Controller scenarios	555
16.1	SAN Volume Controller upgrade with CF8 nodes and internal solid-state drives	556
16.2	Handling Stuck SAN Volume Controller Code Upgrades	568
16.3	Moving an AIX server	568
16.4	Migrating to a new SAN Volume Controller by using Copy Services	570
16.5	SAN Volume Controller scripting	575
16.5.1	Connecting to SAN Volume Controller by using predefined SSH connection	575
16.5.2	Scripting toolkit	579
16.6	Migrating AIX cluster volumes off DS4700	579
16.6.1	Preparation	580
16.6.2	Importing image mode volumes	582
16.6.3	Data migration	582
16.6.4	Final configuration	583
16.7	Easy Tier and FlashSystem planned outages	584
16.8	Changing LUN ID presented to a VMware ESXi host	585
Chapter 17.	IBM Real-time Compression	593
17.1	Overview	594

17.2 What is new in version 7.2	594
17.3 Evaluate data types for estimated compression savings by using the Comprestimator utility	595
17.4 Evaluate workload by using Disk Magic sizing tool	596
17.5 Configure a balanced system	596
17.6 Verify available CPU resources	597
17.7 Compressed and non-compressed volumes in the same MDisk group	598
17.8 Application benchmark results	599
17.8.1 Synthetic workloads	599
17.9 Standard benchmark tools	600
17.10 Compression with FlashCopy	600
17.11 Compression with Easy Tier	600
17.12 Compression on SAN Volume Controller with Storwize V700	601
17.13 Related Publications	601
Appendix A. IBM i considerations	603
IBM i Storage management	604
Single level storage	604
Planning for IBM i capacity	604
Connecting SAN Volume Controller or Storwize to IBM i	605
Native connection	605
Connection with VIOS_NPIV	605
Connection with VIOS virtual SCSI	606
Setting of attributes in VIOS	606
Preparing SAN Volume Controller or Storwize storage for IBM i	607
Disk drives for IBM i	607
Defining LUNs for IBM i	608
Data layout	608
Solid-state drives	609
Sizing Fibre Channel adapters in IBM i and VIOS	609
Zoning SAN switches	610
Boot from SAN	610
IBM i mirroring	611
IBM i Multipath	611
Copy services considerations	612
Related publications	613
IBM Redbooks publications	613
Other resources	614
Referenced websites	614
Help from IBM	615
Index	617

Notices

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

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

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

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

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

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

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

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

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

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

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

COPYRIGHT LICENSE:

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

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at <http://www.ibm.com/legal/copytrade.shtml>

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

AIX®	FlashCopy®	Redbooks®
alphaWorks®	FlashSystem™	Redpaper™
BladeCenter®	GPFS™	Redbooks (logo)  ®
Cognos®	HACMP™	Service Request Manager®
DB2®	IBM®	Storwize®
developerWorks®	IBM FlashSystem™	System p®
DS4000®	IBM Flex System™	System Storage®
DS6000™	Nextra™	System x®
DS8000®	POWER®	System z®
Easy Tier®	PowerHA®	SystemMirror®
Enterprise Storage Server®	PowerVM®	Tivoli®
eServer™	Real-time Compression™	XIV®

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

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

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

Find and read thousands of IBM Redbooks publications

- ▶ Search, bookmark, save and organize favorites
- ▶ Get up-to-the-minute Redbooks news and announcements
- ▶ Link to the latest Redbooks blogs and videos

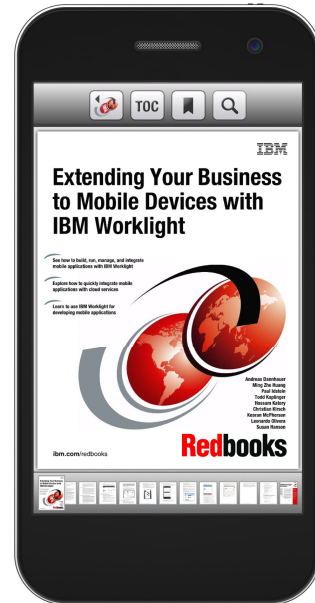
Get the latest version of the Redbooks Mobile App



iOS

Download
Now

Android



Promote your business in an IBM Redbooks publication

Place a Sponsorship Promotion in an IBM® Redbooks® publication, featuring your business or solution with a link to your web site.

Qualified IBM Business Partners may place a full page promotion in the most popular Redbooks publications. Imagine the power of being seen by users who download millions of Redbooks publications each year!



ibm.com/Redbooks

About Redbooks → Business Partner Programs

THIS PAGE INTENTIONALLY LEFT BLANK

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 V7.2.

This book begins with a look at the latest developments with SAN Volume Controller and Storwize V7000 and reviews the changes in the previous versions of the product. 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. Finally, this book highlights several scenarios that demonstrate the preferred practices and performance guidelines.

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. For more information, see the following publications:

- ▶ *Implementing the IBM System Storage SAN Volume Controller V7.2*, SG24-7933
- ▶ *Implementing the IBM Storwize V7000 V7.2*, SG24-7938
- ▶ *Real-time Compression in SAN Volume Controller and Storwize V7000*, REDP-4859
- ▶ *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
- ▶ *Introduction to Storage Area Networks and System Networking*, SG24-5470

The team who wrote this book

This book was produced by a team of specialists from around the world working for the International Technical Support Organization (ITSO), at the IBM Tel Aviv, Israel office.



Jon Tate is a Project Manager for IBM System Storage SAN Solutions at the International Technical Support Organization (ITSO), San Jose Center. Before joining the ITSO in 1999, he worked in the IBM Technical Support Center, providing Level 2 support for IBM storage products. Jon has 28 years of experience in storage software and management, services, and support. He is an IBM Certified IT Specialist and an IBM SAN Certified Specialist. He is also the UK Chairman of the Storage Networking Industry Association.



Pawel Brodacki is an IT Specialist working for IBM Integrated Technology Services at IBM Polska, where he is involved in designing, delivering, and supporting IT infrastructure solutions. Pawel is an IBM Certified IT Specialist with over 15 years of experience. He specializes in infrastructure and virtualization. His experience includes SAN, storage, highly available systems, disaster recovery solutions, IBM xSeries, Power and Blade servers, and several types of operating systems (Linux, IBM AIX®, and Microsoft Windows). Pawel holds certifications from IBM, Red Hat, and VMware.



Tilak Buneti is an IBM Real-time Compression™ Development Support Engineer based in North Carolina, US, and has over 15 years of experience working in Storage and IT fields. He joined IBM directly as a professional and holds a Bachelor degree in Electronics and Communication Engineering from Jawaharlal Nehru Technological University, Hyderabad, India. He has expertise in various technologies that are used in NAS, SAN, backup, and storage optimization technologies. He has certifications for CCNA, MCSE, NACP, and NACA. In his current role, he is responsible for worldwide product support for IBM Real-time Compression and documentation updates.



Christian Burns is an IBM Storage Solution Architect based in New Jersey. As a member of the Storage Solutions Engineering team in Littleton, MA, he works with clients, IBM Business Partners, and IBM employees worldwide, designing, and implementing storage solutions that include various IBM products and technologies. Christian's areas of expertise include IBM Real-time Compression, SAN Volume Controller, XIV®, and IBM FlashSystem™. Before joining IBM, Christian was the Director of Sales Engineering at IBM Storwize before it became IBM Storwize. He brings over a decade of industry experience in the areas of sales engineering, solution design, and software development. Christian holds a BA degree in Physics and Computer Science from Rutgers College.



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.



Erez Kirson is a Technical Sales Specialist for EMET BARMOR, a premier IBM Business Partner in Israel. As a member of the pre-sales team in Israel, he works with clients and IBM employees worldwide, designing, and implementing complex storage solutions that include various storage products and technologies, especially IBM products, such as Storwize V7000, XIV, SAN Volume Controller, and GPFS™. Erez has 12 years experience of IT support, technical sales, knowledge of operating systems, SAN, NAS, and IBM products. His current responsibility is to assist the marketing and sales team with proof of concept, complex design, and technical sales strategy.



Marcin Tabinowski works as an IT Specialist in STG Lab Services in Poland. He has over eight years of experience in designing and implementing IT solutions that are based on storage and POWER® systems. His main responsibilities are architecting, consulting, implementing, and documenting projects including storage systems, SAN networks, Power Systems, disaster recovery, virtualization, and data migration. Pre-sales, post sales, and training are also part of his everyday duties. Martin holds many certifications that span different IBM storage products and Power Systems. He also holds an MSC of Computer Science from Wroclaw University of Technology, Poland.



Bosmat Tuv-El is a Manager of Development Support for IBM Real-time Compression in Israel. Bosmat has 10 years of IT, QA, and support experience in Storage systems and Networking. She joined IBM through the acquisition of Storwize in 2010. She manages the worldwide product support team for IBM Real-time Compression that provides analysis of complex customer problems and works to improve the overall customer experience through product improvements and documentation. Bosmat graduated from the Open University of Israel with a BA in Computer Science and Management.

We thank the following people for their contributions to this project:

- ▶ The development and product field engineer teams in Hursley, England
- ▶ The RtC development and L3 teams in Tel Aviv, Israel, and United States
- ▶ The following authors of the previous edition of this book:

Katja Gebuhr
Ivo Gomilsek
Ronda Hruby
Mary Lovelace
Paulo Neto
Jon Parkes
Otavio Rocha Filho
Leandro Torolho

- ▶ The following people for their contributions:

Andrew Martin
Katja Gebuhr
Markus Standau
Barry Whyte
Bill Wiegand

- ▶ Ann Lund from the ITSO

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:

<http://www.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:

<http://www.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 that were 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 SG24-7521-03
for *Best Practices and Performance Guidelines*
as created or updated on August 4, 2016.

September 2014, Fourth Edition

This revision reflects the addition of new information since version 6.2. Chapter 1, “Updates in IBM System Storage SAN Volume Controller” on page 3 contains a list of the updates to the previous releases of SAN Volume Controller and Storwize V7000.

- 19 November 2015: changed note box wording in 4.7.4, “Volumes” on page 91 to:

On SAN Volume Controller/Storwize, disable the cache for basic volumes that are created in storage pools that are composed of only FlashSystem MDisks. You can disable the cache when the volume is created or later. This does not apply to the volumes in multitier storage pools or volumes utilising advanced SVC function. Always enable cache for compressed vdisks. For most other advanced SVC function (FlashCopy, Mirroring, Space Efficient) best latency may be achieved with cache enabled.



Part 1

Configuration guidelines and preferred practices

This part describes the latest developments for IBM System Storage SAN Volume Controller V6.2 and reviews the changes in the previous versions of the product. 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.

This part includes the following chapters:

- ▶ Chapter 1, “Updates in IBM System Storage SAN Volume Controller” on page 3
- ▶ Chapter 2, “SAN topology” on page 17
- ▶ Chapter 3, “SAN Volume Controller and Storwize V7000 Cluster” on page 59
- ▶ Chapter 4, “Back-end storage” on page 71
- ▶ Chapter 5, “Storage pools and managed disks” on page 95
- ▶ Chapter 6, “Volumes” on page 125
- ▶ Chapter 7, “Remote copy services” on page 157
- ▶ Chapter 8, “Hosts” on page 225



Updates in IBM System Storage SAN Volume Controller

This chapter summarizes the enhancements in the IBM System Storage SAN Volume Controller since V4.3. It also describes the terminology that changed over previous releases of SAN Volume Controller.

This chapter includes the following sections:

- ▶ Enhancements and changes in SAN Volume Controller V5.1
- ▶ Enhancements and changes in SAN Volume Controller V6.1
- ▶ Enhancements and changes in SAN Volume Controller V6.2
- ▶ Enhancements and changes in SAN Volume Controller V6.3
- ▶ Enhancements and changes in SAN Volume Controller V6.4
- ▶ Enhancements and changes in SAN Volume Controller V7.1
- ▶ Enhancements and changes in SAN Volume Controller V7.2

1.1 Enhancements and changes in SAN Volume Controller V5.1

The following major enhancements and changes were introduced in SAN Volume Controller V5.1:

- ▶ New capabilities with the 2145-CF8 hardware engine

SAN Volume Controller offers improved performance capabilities by upgrading to a 64-bit software kernel. With this enhancement, you can use cache increases, such as 24 GB, that are provided in the new 2145-CF8 hardware engine. SAN Volume Controller V5.1 runs on all SAN Volume Controller 2145 models that use 64-bit hardware, including Models 8F2, 8F4, 8A4, 8G4, and CF8. The 2145-4F2 node (32-bit hardware) is not supported in this version.

SAN Volume Controller V5.1 also supports optional solid-state drives (SSDs) on the 2145-CF8 node, which provides a new ultra-high-performance storage option. Each 2145-CF8 node supports up to four SSDs with the required serial-attached SCSI (SAS) adapter.

- ▶ Multitarget reverse IBM FlashCopy® and Storage FlashCopy Manager

With SAN Volume Controller V5.1, reverse FlashCopy support is available. With reverse FlashCopy, FlashCopy targets can become restore points for the source without breaking the FlashCopy relationship and without waiting for the original copy operation to complete. Reverse FlashCopy supports multiple targets and, therefore, multiple rollback points.

- ▶ 1 Gb iSCSI host attachment

SAN Volume Controller V5.1 delivers native support of the iSCSI protocol for host attachment. However, all internode and back-end storage communications still flow through the Fibre Channel (FC) adapters.

- ▶ I/O group split in SAN Volume Controller across long distances

With the option to use 8 Gbps Longwave (LW) Small Form Factor Pluggables (SFPs) in the SAN Volume Controller 2145-CF8, SAN Volume Controller V5.1 introduces the ability to split an I/O group in SAN Volume Controller across long distances.

- ▶ Remote authentication for users of SAN Volume Controller clusters

SAN Volume Controller V5.1 provides the Enterprise Single Sign-on client to interact with an LDAP directory server, such as IBM Tivoli® Directory Server or Microsoft Active Directory.

- ▶ Remote copy functions

The number of cluster partnerships increased from one to a maximum of three partnerships. That is, a single SAN Volume Controller cluster can have partnerships of up to three clusters at the same time. This change allows the establishment of multiple partnership topologies that include star, triangle, mesh, and daisy chain.

The maximum number of remote copy relationships increased to 8,192.

- ▶ Increased maximum virtual disk (VDisk) size to 256 TB

SAN Volume Controller V5.1 provides greater flexibility in expanding provisioned storage by increasing the allowable size of VDIs from the former 2 TB limit to 256 TB.

- ▶ Reclaiming unused disk space by using space-efficient VDIs and VDisk mirroring

SAN Volume Controller V5.1 enables the reclamation of unused allocated disk space when you convert a fully allocated VDisk to a space-efficient virtual disk by using the VDisk mirroring function.

- New reliability, availability, and serviceability (RAS) functions

The RAS capabilities in SAN Volume Controller are further enhanced in V5.1. Administrators benefit from better availability and serviceability of SAN Volume Controller through automatic recovery of node metadata, with improved error notification capabilities (across email, syslog, and SMNP). Error notification supports up to six email destination addresses. Also, quorum disk management improved with a set of new commands.

- Optional second management IP address configured on eth1 port

The existing SAN Volume Controller node hardware has two Ethernet ports. Until SAN Volume Controller V4.3, only one Ethernet port (eth0) was used for cluster configuration. In SAN Volume Controller V5.1, a second, new cluster IP address can be optionally configured on the eth1 port.

- Added interoperability

Interoperability is now available with new storage controllers, host operating systems, fabric devices, and other hardware. For an updated list, see *V5.1.x - Supported Hardware List, Device Driver and Firmware Levels for SAN Volume Controller*, S1003553, which is available at this website:

<https://www.ibm.com/support/docview.wss?uid=ssg1S1003553>

- Withdrawal of support for 2145-4F2 nodes (32-bit)

SAN Volume Controller V5.1 supports only SAN Volume Controller 2145 engines that use 64-bit hardware. Therefore, support is withdrawn for 32-bit 2145-4F2 nodes.

- Up to 250 drives, running only on 2145-8A4 nodes, allowed by SAN Volume Controller Entry Edition

The SAN Volume Controller Entry Edition uses a per-disk-drive charge unit and now can be used for storage configurations of up to 250 disk drives.

1.2 Enhancements and changes in SAN Volume Controller V6.1

SAN Volume Controller V6.1 has the following major enhancements and changes:

- A newly designed user interface (similar to IBM XIV Storage System)

The SAN Volume Controller Console has a newly designed GUI that now runs on the SAN Volume Controller and can be accessed from anywhere in the network by using a web browser. The interface includes several enhancements, such as greater flexibility of views, display of running command lines, and improved user customization within the GUI. Customers who use Tivoli Storage Productivity Center and IBM Systems Director can use integration points with the new SAN Volume Controller console.

- New licensing for SAN Volume Controller for XIV (5639-SX1)

Product ID 5639-SX1, IBM SAN Volume Controller for XIV Software V6, is priced by the number of storage devices (also called *modules* or *enclosures*). It eliminates the appearance of double charging for features that are bundled in the XIV software license. Also, you can combine this license with a per TB license to extend the usage of SAN Volume Controller with a mix of back-end storage subsystems.

- ▶ Service Assistant

SAN Volume Controller V6.1 introduces a new method for performing service tasks on the system. In addition to performing service tasks from the front panel, you can service a node through an Ethernet connection by using a web browser or command-line interface (CLI). The web browser runs a new service application that is called the *Service Assistant*. All functions that were previously available through the front panel are now available from the Ethernet connection, with the advantages of an easier to use interface and remote access from the cluster. You also can run Service Assistant commands through a USB flash drive for easier serviceability.

- ▶ IBM System Storage Easy Tier function added at no charge

SAN Volume Controller V6.1 delivers IBM System Storage Easy Tier, which is a dynamic data relocation feature that allows host transparent movement of data among two tiers of storage. This feature includes the ability to automatically relocate volume extents with high activity to storage media with higher performance characteristics. Extents with low activity are migrated to storage media with lower performance characteristics. This capability aligns the SAN Volume Controller system with current workload requirements, which increases overall storage performance.

- ▶ Temporary withdrawal of support for SSDs on the 2145-CF8 nodes

At the time of this writing, 2145-CF8 nodes that use internal SSDs are not supported by V6.1.0.x code (fixed in version 6.2).

- ▶ Interoperability with new storage controllers, host operating systems, fabric devices, and other hardware

For an updated list, see V6.1 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller, S1003697, which is available at this website:

<https://www.ibm.com/support/docview.wss?uid=s5g1S1003697>

- ▶ Removal of 15-character maximum name length restrictions

SAN Volume Controller V6.1 supports object names up to 63 characters. Previous levels supported only up to 15 characters.

- ▶ SAN Volume Controller code upgrades

The SAN Volume Controller console code is now removed. You need to update only the SAN Volume Controller code. The upgrade from SAN Volume Controller V5.1 requires usage of the former console interface or a command line. After the upgrade is complete, you can remove the existing ICA console application from your SSPC or master console. The new GUI is started through a web browser that points the SAN Volume Controller IP address.

- ▶ SAN Volume Controller to back-end controller I/O change

SAN Volume Controller V6.1 allows variable block sizes, up to 256 KB against 32 KB supported in the previous versions. This change is handled automatically by the SAN Volume Controller system without requiring any user control.

- ▶ Scalability

The maximum extent size increased four times to 8 GB. With an extent size of 8 GB, the total storage capacity that is manageable for each cluster is 32 PB. The maximum volume size increased to 1 PB. The maximum number of worldwide node names (WWNN) increased to 1,024, which allows up to 1,024 back-end storage subsystems to be virtualized.

► **SAN Volume Controller and Storwize V7000 interoperability**

The virtualization layer of IBM Storwize V7000 is built upon the IBM SAN Volume Controller technology. SAN Volume Controller V6.1 is the first version that is supported in this environment.

To coincide with new and existing IBM products and functions, several common terms changed and are incorporated in the SAN Volume Controller information. Table 1-1 shows the current and previous usage of the changed common terms.

Table 1-1 SAN Volume Controller Version 6.1 terminology mapping table

Term in SAN Volume Controller V6.1	Term in previous versions of SAN Volume Controller	Description
Event	Error	A significant occurrence to a task or system. Events can include completion or failure of an operation, a user action, or the change in state of a process.
Host mapping	VDisk-to-host mapping	The process of controlling which hosts have access to specific volumes within a cluster.
Storage pool	Managed disk group	A collection of storage capacity that provides the capacity requirements for a volume.
Thin provisioning (thin-provisioned)	Space efficient	The ability to define a storage unit (full system, storage pool, and volume) with a logical capacity size that is larger than the physical capacity that is assigned to that storage unit.
Volume	Virtual disk (VDisk)	A discrete unit of storage on disk, tape, or other data recording medium that supports some form of identifier and parameter list, such as a volume label or I/O control.

1.3 Enhancements and changes in SAN Volume Controller V6.2

SAN Volume Controller V6.2 has the following enhancements and changes:

► **Support for SAN Volume Controller 2145-CG8**

The new 2145-CG8 engine contains 24 GB of cache and four 8 Gbps FC host bus adapter (HBA) ports for attachment to the SAN. The 2145-CG8 autonegotiates the fabric speed on a per-port basis and is not restricted to run at the same speed as other node pairs in the clustered system. The 2145-CG8 engine can be added in pairs to an existing system that consists of 64-bit hardware nodes (8F2, 8F4, 8G4, 8A4, CF8, or CG8) up to the maximum of four pairs.

► **10 Gb iSCSI host attachment**

The new 2145-CG8 node comes with the option to add a dual port 10 Gb Ethernet adapter, which can be used for iSCSI host attachment. The 2145-CG8 node also supports the optional use of SSD devices (up to four). However, both options cannot coexist on the same SAN Volume Controller node.

- Real-time performance statistics through the management GUI

Real-time performance statistics provide short-term status information for the system. The statistics are shown as graphs in the management GUI. Historical data is kept for about five minutes. Therefore, you can use Tivoli Storage Productivity Center to capture more detailed performance information to analyze mid-term and long-term historical data, and to have a complete picture when you develop best-performance solutions.

- SSD RAID at levels 0, 1, and 10

Optional SSDs are not accessible over the SAN. Their usage is done through the creation of RAID arrays. The supported RAID levels are 0, 1, and 10. In a RAID 1 or RAID 10 array, the data is mirrored between SSDs on two nodes in the same I/O group.

- Easy Tier for use with SSDs on 2145-CF8 and 2145-CG8 nodes

SAN Volume Controller V6.2 restarts support of internal SSDs by allowing Easy Tier to work with internal Subsystem Device Driver (SDD) storage pools.

- Support for a FlashCopy target as a remote copy source

In SAN Volume Controller V6.2, a FlashCopy target volume can be a source volume in a remote copy relationship.

- Support for the VMware vStorage API for Array Integration (VAAI)

SAN Volume Controller V6.2 fully supports the VMware VAAI protocols. An improvement that comes with VAAI support is the ability to dramatically offload the I/O processing that is generated by performing a VMware Storage vMotion.

- CLI prefix removal

The **svctask** and **svcinfo** command prefixes are no longer necessary when you issue a command. If you have existing scripts that use those prefixes, they continue to function.

- Licensing change for the removal of a physical site boundary

The licensing for SAN Volume Controller systems (formerly *clusters*) within the same country and that belong to the same customer can be aggregated in a single license.

- FlashCopy license on the main source volumes

SAN Volume Controller V6.2 changes the way the FlashCopy is licensed so that SAN Volume Controller now counts as the main source in FlashCopy relationships. Previously, if cascaded FlashCopy was set up, multiple source volumes had to be licensed.

- Interoperability with new storage controllers, host operating systems, fabric devices, and other hardware

For an updated list, see *V6.2 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller*, S1003797, which is available at this website:

<https://www.ibm.com/support/docview.wss?uid=ssg1S1003797>

- Exceeding entitled virtualization license 45 days from the installation date for migrating data from one system to another

With the benefit of virtualization, customers can bring new storage systems into their storage environment and quickly and easily migrate data from their existing storage systems to the new storage systems by using SAN Volume Controller. To facilitate this migration, IBM customers can temporarily (45 days from the date of installation of the SAN Volume Controller) exceed their entitled virtualization license for migrating data from one system to another.

Table 1-2 shows the current and previous usage of one changed common term.

Table 1-2 SAN Volume Controller Version 6.2 terminology mapping table

Term in SAN Volume Controller V6.2	Term in previous versions of SAN Volume Controller	Description
Clustered system or system	Cluster	A collection of nodes that is placed in pairs (I/O groups) for redundancy, which provides a single management interface.

1.4 Enhancements and changes in SAN Volume Controller V6.3

SAN Volume Controller V6.3 has the following enhancements and changes:

- ▶ Enhanced Replication via Global Mirror with Change Volumes (GMCV)

Enhancements to Global Mirror with the SAN Volume Controller V6.3.0 are designed to provide new options to help administrators balance network bandwidth requirements and recovery point objectives (RPOs) for applications. SAN Volume Controller now supports higher RPO times, which provide the option to use a lower-bandwidth link between mirrored sites. This lower-bandwidth remote mirroring uses space-efficient, FlashCopy targets as sources in remote copy relationships to increase the time that is allowed to complete a remote copy data cycle.

- ▶ Metro Mirror and Global Mirror Replication between SAN Volume Controller and Storwize V7000 systems

With SAN Volume Controller and Storwize V7000 running V6.3.x, the Storwize V7000 can act as a SAN Volume Controller Metro Mirror or Global Mirror partner system. SAN Volume Controller V6.3 introduces a new cluster property that is called *layer*. Storwize V7000 is in “replication layer” mode or “storage layer” mode, while SAN Volume Controller is always in “replication layer” mode. Storwize V7000 is in “storage layer” mode by default, and can be switched to “replication layer” by using the **svctask chcluster -layer replication** command. After it is changed to “replication layer” mode, the Storwize V7000 can then be used to create a remote copy relationship with a SAN Volume Controller cluster.

- ▶ Automatically shrink Thin Provisioned Volumes

For thin provisioned volumes, the real capacity starts small and grows as data is written to the volume. If data was deleted in previous versions, real capacity did not automatically shrink, but it can be manually shrunk. This might be problematic with FlashCopy (FC) mappings. In SAN Volume Controller V6.3, starting a FlashCopy mapping automatically shrinks used and real capacity to zero for all thin provisioned volumes that are used in the FC mapping.

- ▶ Mirrored Volume Time-out Enhancements

SAN Volume Controller V6.3 introduces the option of configurable timeout settings for each mirrored volume. The default setting for the `mirror_write_priority` property of the volume is “latency”, and uses a short timeout, which prioritizes low host latency. This property can be changed to “redundancy, where a longer timeout is used to prioritize redundancy.

► Support for Round-Robin multipathing to external storage systems

Before V6.3, all I/O to an external MDisk on a storage system was via a single port on that controller. The selected port changed if that port became unavailable. In V6.3, an I/O is submitted by using one path per target port per MDisk per node. Paths are chosen according to port groups that are presented by the storage system and the I/O is sent in parallel to all target ports. This enables I/O to an MDisk to progress in a “round robin” fashion, with the following potential benefits:

- Performance improvement
- Spreading the I/O across multiple storage system ports
- Balancing the number of preferred paths per node port both within each port group and across the system as a whole
- Improved resilience to certain storage system failures
- Faster detection of path failures

For more information about the specific external storage systems that are supported by Round-Robin multipathing in V6.3, see *V6.3 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller, S1003907*, which is available at this website:

<https://www.ibm.com/support/docview.wss?uid=ssg1S1003907>

► Stretched Cluster enhancements

SAN Volume Controller V6.3.0 introduces the ability to extend the distance between SAN Volume Controller nodes in a Stretched Cluster (Split I/O Group) configuration. Although the extended distances depend on application latency restrictions, this function now enables enterprises to access and share a consistent view of data simultaneously across data centers. This function also enables enterprises to relocate data across disk array vendors and tiers, both inside and between data centers at full metro distances.

► Enhanced LDAP authentication support

V6.3 introduces support for direct authentication to and LDAP server. Authentication via Tivoli Integrated Portal is still supported, but no longer required.

► Support for CLI password authentication

V6.3 introduces support for CLI authentication that uses a password. Authentication via SSH key is still supported, but no longer required.

► Storwize V7000 support for other drive types

V6.3 introduces support for the following new drive types:

- Storwize V7000 3 TB 3.5-inch 7.2 K RPM Near-Line SAS drive
- Storwize V7000 200 GB and 400 GB 2.5-inch SSD

► GUI Enhancements

V6.3 introduces various GUI-related enhancements in the following categories:

- Usability:
 - Per-column grid filtering support
 - New tree table views (MDisks by pool, FC consistgrp, and RC consistgrp)
 - New XIV style status pods
 - Recommended actions and events panels combined
 - Easy Tier icon badge for easy tier pools

- New cluster features:
 - Quorum disk management support (SAN Volume Controller)
 - Native LDAP support
 - Support for Global Mirror with Change volumes
- Performance monitoring: Read and write latency statistics for volumes and MDisk
- Interoperability with new storage controllers, host operating systems, fabric devices, and other hardware

For an updated list, see *V6.3 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller*, S1003907, which is available at this website:

<https://www.ibm.com/support/docview.wss?uid=ssg1S1003907>

1.5 Enhancements and changes in SAN Volume Controller V6.4

SAN Volume Controller V6.4 has the following enhancements and changes:

► FCoE Support

With V6.4, SAN Volume Controller systems with 10 Gbps Ethernet ports now support attachment to next-generation CEE networks by using FCoE. This support enables SAN Volume Controller connections to servers for host attachment and to other SAN Volume Controller systems for clustering or for mirroring by using Fibre Channel or FCoE interfaces that use these networks. The same ports can also be used for iSCSI server connections.

► Non-Disruptive Volume Movement across clustered systems

V6.4 enhances data mobility, with greater flexibility for nondisruptive volume moves. Previous versions of SAN Volume Controller provided the ability to move volumes nondisruptively between the nodes in an I/O group. Version 6.4 supports moving volumes anywhere within a clustered system without disruption of host access to storage.

► Real-time Compression

V6.4 is designed to improve storage efficiency by supporting real-time compression for block storage, which is designed to improve efficiency by compressing data by as much as 80%, which enables storage for up to five times as much data in the same physical disk space. Unlike other approaches to compression, IBM Real-time Compression is used with active primary data, such as production databases and email systems, which expands the range of candidate data that can benefit from compression. IBM Real-time Compression operates as data is written to disk, which avoids the need to store uncompressed data while awaiting compression.

► Storwize V7000 clustering

V6.4 allows for the clustering of multiple V7000 control enclosures. As with SAN Volume Controller, Storwize V7000 clustering works on the notion of I/O groups. With V7000, an I/O group is a control enclosure and its associated expansion enclosures. A V7000 clustered system can consist of 2-4 I/O groups.

► Support for direct host attachment

► Updated thin provisioned volume grain size

For improved performance and interaction with Easy Tier, the default grain size of a thin provisioned volume was changed to 256 KB from the previous default of 32 KB.

- ▶ Extended support for SCSI-3 persistent reservations
Other persistent reserve functions allow GPFS to use persistent reserves on a Storwize V7000 or SAN Volume Controller system.
- ▶ Interoperability
For an updated list, see V6.4 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller, S100411 1, which is available at this website:
<https://www.ibm.com/support/docview.wss?uid=ssg1S1004111>

1.6 Enhancements and changes in SAN Volume Controller V7.1

SAN Volume Controller V7.1 has the following enhancements and changes:

- ▶ Increased Number of Host Objects
SAN Volume Controller V7.1 increases number of host objects per I/O group from 256 to 512 and the per cluster limit from 1024 to 2048. The increased host objects can be used for any host type subject to limit restrictions for that host type; for example, iSCSI names/IQNs (iSCSI Qualified Names).
- ▶ Increased Number of Host WWPNNs
SAN Volume Controller V7.1 increases the officially supported number of host WWPNNs per I/O group to 2048 and per cluster to 8192. This specifically benefits AIX LPM configurations and environments that use NPIV to map volumes to virtual WWPNNs. This increase applies to native FC or FCoE WWPNNs.
- ▶ Increased Number of Volumes per Host
SAN Volume Controller V7.1 increase volumes per host from 512 to 2048, and the Increase in volumes per host is available to any host operating system, subject to that host's OS limits. Increase in volumes per host is applicable to FC and FCoE host attachment types and does not apply to iSCSI attached servers/hosts.
- ▶ Support for the direct attachment of AIX hosts
- ▶ Support for more drive types
SAN Volume Controller V7.1 introduces support for the following drives:
 - 4 TB NL_SAS 7.2 K RPM 3.5-inch LFF drives (supported on Storwize V7000 2076-x12 LFF enclosures only; supported by Flex System V7000 when external expansion enclosure model 2076-212 is used and connected to the Flex System V7000 control enclosure. It is not supported on Storwize V3700/3500 models).
 - 1.2 TB SAS 10K RPM 2.5-inch SFF drive that is supported on Storwize V7000, Storwize V3700/3500, and Flex System V7000 SFF enclosures. For Storwize V3700/3500 LFF control or expansion enclosures, this 2.5-inch drive is available on an LFF carrier as is the 2.5-inch SAS 15 K RPM 300 GB drive and the 2.5-inch SAS 10K RPM 900 GB drive.
- ▶ Second Fibre Channel HBA support
SAN Volume Controller V7.1 adds support for the other 4-port 8 Gbps Fibre Channel HBA that is available in feature code AHA7 on 2145-CG8 hardware.

► Port masking

The addition of more Fibre Channel HBA ports that are introduced with feature code AHA7 allow clients to optimize their SAN Volume Controller configuration by using dedicated ports for certain system functions. However, the addition of these ports necessitates the ability to ensure traffic isolation. As such, SAN Volume Controller V7.1 introduces port masking.

Traffic types that you might want to isolate by using port masking are shown in the following examples:

- Local node-to-node communication
- Replication traffic

► Support for Easy Tier with compressed volumes

Easy Tier is a performance optimization function that automatically migrates “hot” extents that belong to a volume to MDisks that better meet the performance requirements of that extent. The Easy Tier function can be turned on or off at the storage pool level and at the volume.

Real-time Compression is a feature of SAN Volume Controller that addresses all of the requirements of primary storage data reduction, including performance and the use of purpose-built compression technology, which allow for data reduction of up to 80%.

In practice, clients find that their target workloads for these two features have a significant overlap. Before SAN Volume Controller Storage Software version 7.1, the use of these two features was mutually exclusive at the volume level. SAN Volume Controller V7.1 introduces support for the concurrent use of Easy Tier and Real-time Compression on the same volume.

► Enhanced flexibility in modifying Remote Copy relationships

SAN Volume Controller V7.1 introduces the ability to change between Metro Mirror and Global Mirror (with or without change volumes) without requiring a full resync of all data from the primary volume to the secondary volume.

► Storwize V3700 support for Remote Copy

SAN Volume Controller V7.1 introduces support for Remote Copy on Storwize V3700 systems, which allows for remote replication between any combination of the following systems:

- SAN Volume Controller
- Storwize V7000
- Flex System V7000
- Storwize V3700
-

► Interoperability

For an updated list, see *V7.1 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller*, S1004392, which is available at this website:

<https://www.ibm.com/support/docview.wss?uid=ssg1S1004392>

1.7 Enhancements and changes in SAN Volume Controller V7.2

SAN Volume Controller V7.2 has the following enhancements and changes:

- ▶ Remote Mirroring over IP communication links

The Remote Mirroring function (also referred to as *Metro/Global Mirror*) is now supported by using Ethernet communication links. Storwize Family Software IP replication uses innovative Bridgeworks SANSlide technology to optimize network bandwidth and utilization. This new function enables the use of lower speed and lower-cost networking infrastructure for data replication. When integrated into IBM Storwize Family Software, Bridgeworks' SANSlide technology uses artificial intelligence to help optimize network bandwidth utilization and adapt to changing workload and network conditions. This technology can improve remote mirroring network bandwidth utilization up to three times, which can enable clients to deploy a less costly network infrastructure or speed remote replication cycles to enhance disaster recovery effectiveness.

- ▶ Enhanced Stretched Cluster for SAN Volume Controller

Before this release, stretched cluster configurations did not provide manual failover capability, and data that is being sent across a long-distance link had the potential to be sent twice. The addition of site awareness in Storwize Family Software V7.2 routes I/O traffic between SAN Volume Controller nodes and storage controllers to optimize the data flow, and it polices I/O traffic during a failure condition to allow for a manual cluster invocation to ensure consistency. The use of stretched cluster continues to follow all the same hardware installation guidelines as previously announced and found in the product documentation. Use of enhanced stretched cluster is optional, and existing stretched cluster configurations continues to be supported.

- ▶ Performance improvements for asynchronous remote mirroring

Enhancements to the asynchronous remote mirroring function enable improved throughput of remotely replicated data.

- ▶ Improved efficiency in drive firmware update process

In previous software releases, updating internal drive firmware was a serial process that required the user to update each drive individually. V7.2 introduces a drive firmware update command, **svctask applydrivesoftware**, that allows for the updating of multiple drives with a single CLI command.

- ▶ vSphere API for Storage Awareness (VASA)

Storwize Family Software V7.2 enables users to get more capability out of their VMware environments by being a provider for the vSphere API for Storage Awareness.

- ▶ Data Migration by using SAS connectivity on Storwize V3500, V3700, and V5000

Data migration support is standard on all Storwize V3500, V3700, and V5000 systems and this function can now be performed by using SAS connectivity to help you easily and nondisruptively migrate data from IBM System Storage DS3200 and DS3500 systems onto Storwize V3500, V3700, and V5000.

- ▶ Enhanced monitoring capabilities with the new IBM Storage Mobile Dashboard

With V7.2, a new mobile application was released that allows for monitoring and health check functionality for SAN Volume Controller and Storwize Family storage systems. The application is available for free from the Apple App Store.

- Improved performance and efficiency for Real-time Compression with the introduction of the Random Access Compression Engine (RACE) 2.2

V7.2 introduces the following improvements to the Real-time Compression functionality:

- Up to 3x higher sequential write throughput, which allows for faster VMware vMotion operations and sequential copy operations and more VMware vMotion sessions in parallel
 - 35% higher throughput (IOPS) in intensive DB OLTP workloads
 - 35% lower compression CPU usage for the same workload compared to V7.1
- Interoperability

For an updated list, see *V7.2 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller*, S1004453, which is available at this website:

<https://www.ibm.com/support/docview.wss?uid=ssg1S1004453>



SAN topology

The IBM System Storage SAN Volume Controller and Storwize family systems have unique SAN fabric configuration requirements that differ from what you might be used to in your storage infrastructure. A quality SAN configuration can help you achieve a stable, reliable, and scalable SAN Volume Controller/Storwize installation. Conversely, a poor SAN environment can make your SAN Volume Controller/Storwize experience considerably less pleasant.

This chapter helps to tackle this topic that is based on experiences from the field. Although many other SAN configurations are possible (and supported), this chapter highlights the preferred configurations.

This chapter includes the following sections:

- ▶ SAN topology of the SAN Volume Controller/Storwize V7000
- ▶ SAN switches
- ▶ Zoning
- ▶ Switch domain IDs
- ▶ Distance extension for remote copy services
- ▶ Fabric Virtual channels
- ▶ Tape and disk traffic that share the SAN
- ▶ Switch interoperability
- ▶ IBM Tivoli Storage Productivity Center
- ▶ iSCSI support
- ▶ SAS support

SAN design: If you are planning for a SAN Volume Controller installation, you must be knowledgeable about general SAN design principles. For more information about SAN design, limitations, caveats, and updates that are specific to your SAN Volume Controller environment, see the following publications:

- ▶ *IBM System Storage SAN Volume Controller V6.4.1 - 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.641.doc/mlt_relatedinfo_224agr.html
- ▶ *V7.2 Configuration Limits and Restrictions for IBM System Storage SAN Volume Controller*, S1004510, which is available at this website:
<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004510>

For updated documentation before you implement your solution, see the IBM System Storage SAN Volume Controller Support Portal at this website:

[http://www.ibm.com/support/entry/portal/Overview/Hardware/System_Storage/Storage_software/Storage_virtualization/SAN_Volume_Controller_\(2145\)](http://www.ibm.com/support/entry/portal/Overview/Hardware/System_Storage/Storage_software/Storage_virtualization/SAN_Volume_Controller_(2145))

For updated documentation and information about Storwize family systems, see the following IBM Storwize Support Portals:

- ▶ Storwize V3700
http://www-947.ibm.com/support/entry/portal/product/system_storage/disk_systems/entry-level_disk_systems/ibm_storwize_v3700?productContext=-124971743
- ▶ Storwize V5000
http://www-947.ibm.com/support/entry/portal/product/system_storage/disk_systems/mid-range_disk_systems/ibm_storwize_v5000?productContext=-2033461677
- ▶ Storwize V7000
[http://www-947.ibm.com/support/entry/portal/product/system_storage/disk_systems/mid-range_disk_systems/ibm_storwize_v7000_\(2076\)?productContext=-1546771614](http://www-947.ibm.com/support/entry/portal/product/system_storage/disk_systems/mid-range_disk_systems/ibm_storwize_v7000_(2076)?productContext=-1546771614)

2.1 SAN topology of the SAN Volume Controller/Storwize

The topology requirements for the SAN Volume Controller/Storwize do not differ too much from any other storage device. What makes the SAN Volume Controller/Storwize unique is that it can be configured with many hosts, which means that you must carefully consider SAN scalability. Also, because the SAN Volume Controller/Storwize often serves so many hosts, it is essential to avoid poor SAN design by planning thoroughly in advance of the implementation phase.

2.1.1 Redundancy

One of the fundamental SAN requirements for SAN Volume Controller/Storwize is to create two (or more) separate SANs that are not connected to each other over Fibre Channel (FC) in any way. The easiest way is to construct two SANs that are mirror images of each other.

Note: SAN Volume Controller/Storwize can be connected to up to four separate fabrics.

Technically, the SAN Volume Controller/Storwize supports the use of a single SAN (appropriately zoned) to connect the entire SAN Volume Controller/Storwize. However, we recommend that you do not use this design in any production environment. Based on experience from the field, do not use this design in development environments either because a stable development platform is important to programmers. Also, an extended outage in the development environment can have an expensive business effect. However, for a dedicated storage test platform, it might be acceptable.

Redundancy through Cisco virtual SANs or Brocade Virtual Fabrics

Although virtual SANs (VSANs) and Virtual Fabrics can provide a logical separation within a single appliance, they do not replace the hardware redundancy. All SAN switches are known to suffer from hardware or unrecoverable software failures. Furthermore, separate redundant fabrics into different noncontiguous racks and feed them from redundant power sources.

Note: You can use virtual fabrics or virtual SANs for public and private fabrics in a SAN Volume Controller Enhanced Stretched Cluster or Stretched Cluster scenario.

2.1.2 Topology basics

Regardless of the size of your SAN Volume Controller/Storwize installation, apply the following practices to your topology design:

- ▶ Connect all SAN Volume Controller/Storwize node ports in a clustered system to the same SAN switches 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.
- ▶ Make sure that high-bandwidth utilization servers (such as tape backup servers) are on the same SAN switches as the SAN Volume Controller/Storwize node ports. Placing these servers on a separate switch can cause unexpected SAN congestion problems. Also, placing a high-bandwidth server on an edge switch wastes ISL capacity.
- ▶ If possible, plan for the maximum size configuration that you expect your SAN Volume Controller/Storwize installation to reach. The design of the SAN can change radically for a larger numbers of hosts. Modifying the SAN later to accommodate a larger-than-expected number of hosts might produce a poorly designed SAN. Moreover, it can be difficult, expensive, and disruptive to your business. Planning for the maximum size does not mean that you must purchase all of the SAN hardware initially. It requires you to design only the SAN in consideration of the expected maximum size.
- ▶ Always deploy at least one extra ISL per switch. If you do not, you are exposed to consequences from complete path loss (bad) to fabric congestion (even worse).

The SAN Volume Controller/Storwize does not permit the number of hops between the SAN Volume Controller clustered system and the hosts to exceed three hops. Exceeding three hops is typically not a problem, especially with SAN director class switches.

- ▶ Because of the nature of FC, avoid inter-switch link (ISL) congestion. Although FC (and the SAN Volume Controller/Storwize) can handle a host or storage array that becomes overloaded, the mechanisms in FC for dealing with congestion in the fabric are ineffective under 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 SAN Volume Controller or Storwize.

When an Ethernet network becomes congested, the Ethernet switches discard frames for which no room is available. When an FC network becomes congested, the FC switches stop accepting more frames until the congestion clears and occasionally drop frames. This congestion quickly moves “upstream” in the fabric and clogs the end devices (such as the SAN Volume Controller/Storwize) 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 the inability of SAN Volume Controller/Storwize nodes to communicate with storage subsystems or to mirror their write caches because you have a single congested link that leads to an edge switch.

- ▶ 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 directors fabrics.

2.1.3 ISL oversubscription

IBM System Storage SAN Volume Controller V6.4.1 - Software Installation and Configuration Guide, GC27-2286 specifies a suggested maximum host port to ISL ratio of 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 114 MBps (8 Gbps) or 228 MBps (16 Gbps). If you do not expect most of your hosts to reach anywhere near that value, you can request an exception to the ISL oversubscription rule, which is known as a *request for price quotation* (RPQ) or SCORE request from your IBM marketing representative. Before you request an exception, consider the following factors:

- ▶ Consider your peak loads, not your average loads. For example, although a database server might use only 20 MBps during regular production workloads, it might perform a backup at far higher data rates.
- ▶ Congestion to one switch in a large fabric can cause performance issues throughout the entire fabric, including traffic between SAN Volume Controller 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, as described in 2.1.2, “Topology basics” on page 19. You must avoid congestion if an ISL fails because of such issues as a SAN switch line card or port blade failure.
- ▶ Exceeding the standard 7:1 oversubscription ratio requires you to implement fabric bandwidth threshold alerts. If your ISLs exceed 70%, schedule fabric changes to distribute the load further.
- ▶ Consider the bandwidth consequences of a complete fabric outage. Although a complete fabric outage is a rare event, insufficient bandwidth can turn a single SAN outage into a total access loss event.
- ▶ Consider the bandwidth of the links. It is common to have ISLs run faster than host ports, which reduce the number of required ISLs.

The RPQ process involves a review of your proposed SAN design to ensure that it is reasonable for your proposed environment.

2.1.4 Single switch SAN Volume Controller/Storwize SANs

The most basic SAN Volume Controller/Storwize topology consists of a single switch per SAN. This switch can range from a 16-port 1U switch for a small installation of a few hosts and storage devices, to a director with hundreds of ports. This design has the advantage of simplicity and is a sufficient architecture for small-to-medium SAN Volume Controller/Storwize installations.

The preferred practice is to use a multislotted director-class single switch over setting up a core-edge fabric that is made up solely of lower-end switches. As described in 2.1.2, “Topology basics” on page 19, keep the maximum planned size of the installation in mind if you decide to use this architecture. If you run too low on ports, expansion can be difficult.

2.1.5 Basic core-edge topology

The core-edge topology (as shown in Figure 2-1 on page 22) 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. If they are multislotted directors, they are fitted with at least a few oversubscribed line cards or port blades because most hosts do not require line-speed bandwidth or anything close to it. ISLs must not be on oversubscribed ports.

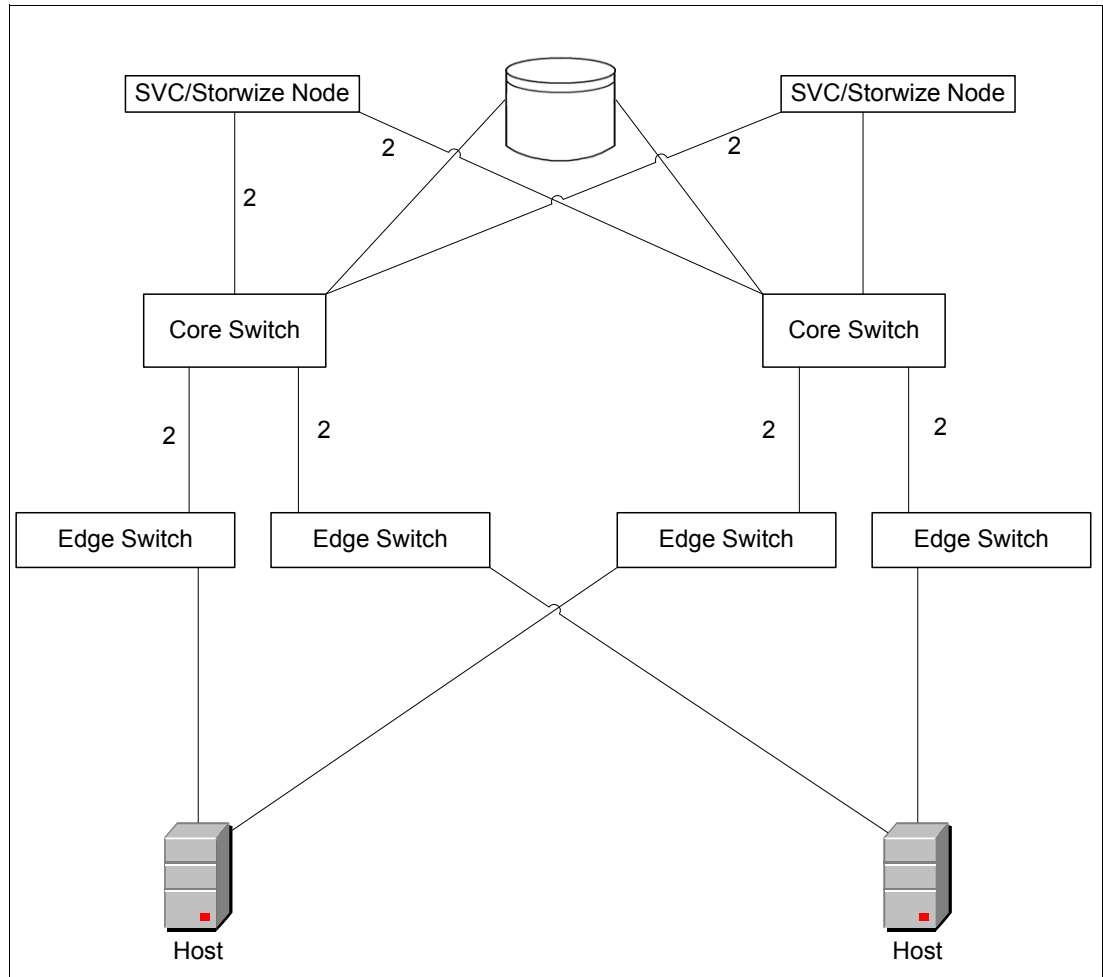


Figure 2-1 Core-edge topology

2.1.6 Four-SAN, core-edge topology

For installations where a core-edge fabric made up of multislot director-class SAN switches is insufficient, the SAN Volume Controller/Storwize can be attached to four SAN fabrics instead of the normal two SAN fabrics. 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.

As shown in Figure 2-2 on page 23, the SAN Volume Controller/Storwize is attached to each of four independent fabrics. The storage subsystem that is used also connects to all four SAN fabrics, even though this design is not required.

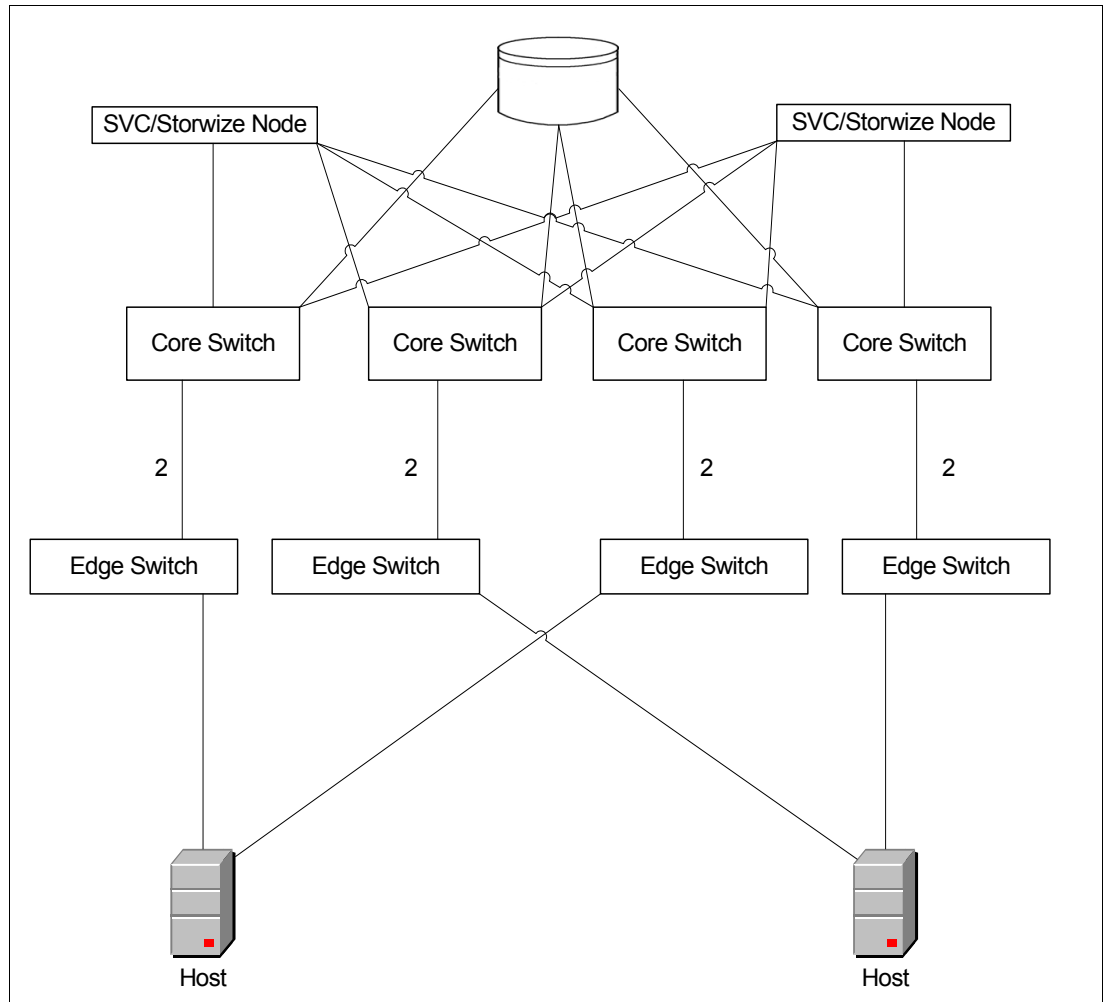


Figure 2-2 Four-SAN core-edge topology

Although some clients simplify management by connecting the SANs into pairs with a single ISL, do not use this design. With only a single ISL connecting fabrics, a small zoning mistake can quickly lead to severe SAN congestion.

SAN Volume Controller/Storwize as a SAN bridge: With the ability to connect a SAN Volume Controller/Storwize to four SAN fabrics, you can use the SAN Volume Controller/Storwize as a bridge between two SAN environments (with two fabrics in each environment). This configuration is useful for sharing resources between SAN environments without merging them. Another use is if you have devices with different SAN requirements in your installation.

When you use the SAN Volume Controller/Storwize as a SAN bridge, pay attention to any restrictions and requirements that might apply to your installation.

2.1.7 Common topology issues

You can encounter several common topology problems.

Accidentally accessing storage over ISLs

A common topology mistake in the field is to have SAN Volume Controller/Storage paths from the same node to the same storage subsystem on multiple core switches that are linked together, as shown in Figure 2-3. This problem is encountered in environments where the SAN Volume Controller/Storage is not the only device that accesses the storage subsystems.

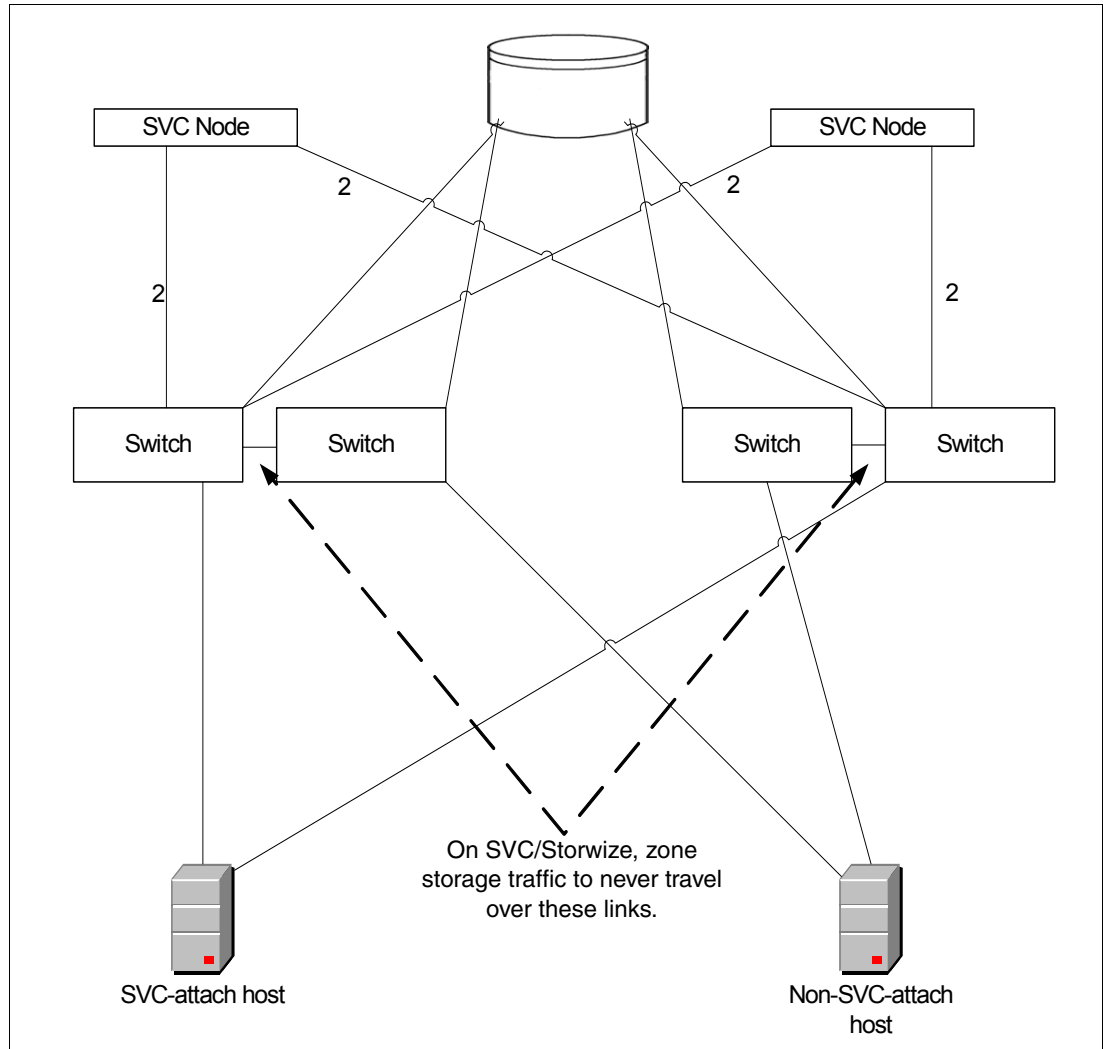


Figure 2-3 Spread out disk paths

If you have this type of topology, you must zone the SAN Volume Controller/Storage so that it detects only paths to the storage subsystems on the same SAN switch as the SAN Volume Controller/Storage nodes. You might consider implementing a storage subsystem host port mask here.

Restrictive zoning: With this type of topology, you must have more restrictive zoning than what is described in 2.3.6, “Standard SAN Volume Controller/Storage zoning configuration” on page 46.

Because of the way that the SAN Volume Controller/Storwize load balances traffic between the SAN Volume Controller nodes and MDisks, the amount of traffic that transits your ISLs is unpredictable and varies significantly. You can use Cisco VSANs or Brocade Traffic Isolation Zones to dedicate an ISL to high-priority traffic. However, internode and SAN Volume Controller/Storwize to back-end storage communication must never cross ISLs.

Important: The SAN Volume Controller/Storwize traffic to storage devices can fill up your ISLs if you have many storage devices ports that are accessed via ISL, especially with the Round Robin Path Selection that was introduced in v6.3 of SAN Volume Controller/Storwize code. Therefore, remember this when you are planning.

Intentionally accessing storage subsystems over an ISL

The practice of intentionally accessing storage subsystems over an ISL goes against the SAN Volume Controller/Storwize configuration guidelines. The reason is that the consequences of SAN congestion to your storage subsystem connections can be severe. Use only this configuration in SAN migration scenarios. If you do use this configuration, closely monitor the performance of the SAN. For most configurations, trunking is required and ISLs must be regularly monitored to detect failures.

I/O group switch splitting with SAN Volume Controller/ Storwize

Clients often want to attach another I/O group to an existing SAN Volume Controller/Storwize to increase the capacity of the SAN Volume Controller/Storwize, but they lack the switch ports to do so. In this situation, you have the following options:

- ▶ Completely overhaul the SAN during a complicated and painful redesign.
- ▶ Add a new switch and ISL to the new I/O group. The new switch is connected to the original switch, as shown in Figure 2-4 on page 26.

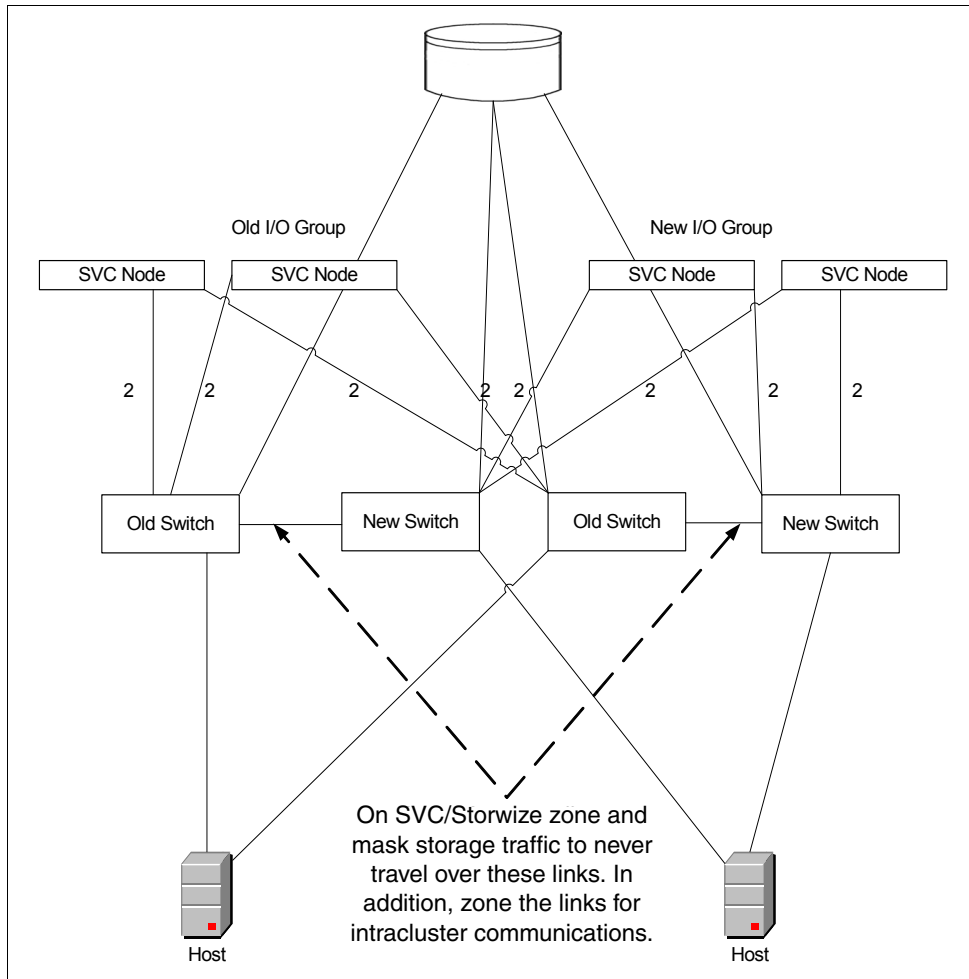


Figure 2-4 I/O group splitting

This design is a valid configuration, but you must take the following precautions:

- ▶ Do not access the storage subsystems over the ISLs. As described in “Accidentally accessing storage over ISLs” on page 24, zone and LUN mask the SAN and storage subsystems. With this design, your storage subsystems need connections to the old and new SAN switches.
- ▶ Have two dedicated ISLs between the two switches on each SAN with no data traffic traveling over them. Use this design because, if this link becomes congested or lost, you might experience problems with your SAN Volume Controller/Storwize clustered system if issues occur at the same time on the other SAN. If possible, set a 5% traffic threshold alert on the ISLs so that you know whether a zoning mistake allowed any data traffic over the links.

Important: Do not use this configuration to perform mirroring between I/O groups within the same clustered system. Also, for SAN Volume Controller, never split the two nodes in an I/O group between various SAN switches within the same SAN fabric if you do not use the SAN Volume Controller Stretched Cluster scenario.

By using the optional 8 Gbps longwave (LW) small form factor pluggables (SFPs) in the 2145-CF8 and 2145-CG8, you can split a SAN Volume Controller I/O group across long distances, as described in 2.1.8, “Stretched Cluster” on page 27.

2.1.8 Stretched Cluster

For high availability, you can stretch a SAN Volume Controller clustered system across three locations and mirror the data. A stretched clustered system configuration locates first node at the first site, second node of the same IO group at the second site, and the active quorum disk at a third site. If communication is lost between the primary and secondary sites, the site with access to the active quorum disk continues to process transactions. If communication is lost to the active quorum disk, an alternative quorum disk at another site can become the active quorum disk.

Note: This chapter does not apply to Storwize family systems because you cannot “stretch” the Storwize control enclosure.

To configure a stretched clustered system, follow these rules:

- ▶ Directly connect each SAN Volume Controller node to one or more SAN fabrics at the primary and secondary sites. Sites are defined as independent power domains that might fail independently. Power domains can be in the same room or across separate physical locations.

- ▶ Use a third site to house a quorum disk.

The storage system that provides the quorum disk at the third site must support extended quorum disks. Storage systems that provide extended quorum support are listed on the IBM System Storage SAN Volume Controller Support page at this website:

[http://www.ibm.com/support/entry/portal/Troubleshooting/Hardware/System_Storage/Storage_software/Storage_virtualization/SAN_Volume_Controller_\(2145\)](http://www.ibm.com/support/entry/portal/Troubleshooting/Hardware/System_Storage/Storage_software/Storage_virtualization/SAN_Volume_Controller_(2145))

- ▶ Place independent storage systems at the primary and secondary sites. In addition, use volume mirroring to mirror the host data between storage systems at the two sites.
- ▶ Use longwave FC connections on SAN Volume Controller nodes that are in the same I/O group and that are separated by more than 100 meters (109 yards). You can purchase an LW SFP transceiver as an optional SAN Volume Controller component. The SFP transceiver must be one of the LW SFP transceivers that are listed at the IBM System Storage SAN Volume Controller Support page at this website:
[http://www.ibm.com/support/entry/portal/Troubleshooting/Hardware/System_Storage/Storage_software/Storage_virtualization/SAN_Volume_Controller_\(2145\)](http://www.ibm.com/support/entry/portal/Troubleshooting/Hardware/System_Storage/Storage_software/Storage_virtualization/SAN_Volume_Controller_(2145))
- ▶ Alternatively, you can use ISL between SAN switches in both sites. It is mandatory to use separate physical switches or virtual fabrics/vsans for node-to-node communication.
- ▶ Connect half of SAN Volume Controller ports of each node to the public fabric and the other half to the private fabric.
- ▶ Dedicate ISLs in private fabrics to node-to-node communication.
- ▶ Avoid the use of ISLs in paths between SAN Volume Controller nodes and external storage systems. If this situation is unavoidable, follow the workarounds that are described in 2.1.7, “Common topology issues” on page 24.
- ▶ Do not use a single switch at the third site because it can lead to the creation of a single fabric rather than two independent and redundant fabrics. A single fabric is an unsupported configuration.
- ▶ Connect SAN Volume Controller nodes in the same system to the same Ethernet subnet.
- ▶ Ensure that a SAN Volume Controller node is in the same rack as the 2145 UPS or 2145 UPS-1U that supplies its power.

- Consider the physical distance of SAN Volume Controller nodes as related to the service actions. Some service actions require physical access to all SAN Volume Controller nodes in a system. If nodes in a split clustered system are separated by more than 100 meters, service actions might require multiple service personnel.

Figure 2-5 shows a stretched clustered system configuration. When used with volume mirroring, this configuration provides a high availability solution that is tolerant of failure at a single site.

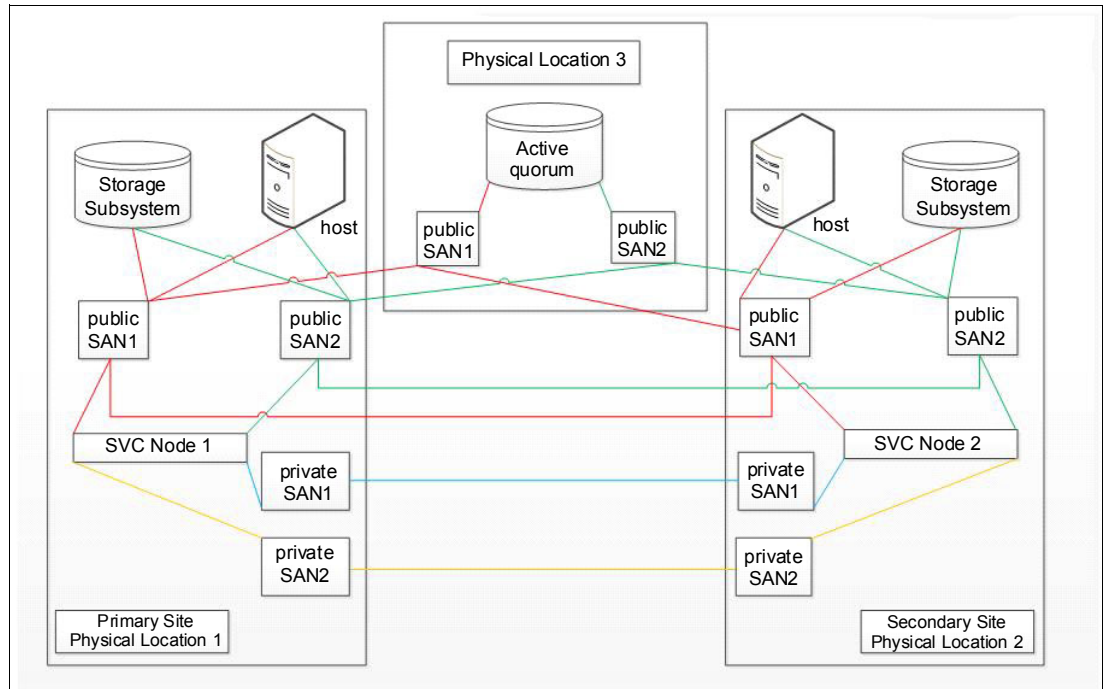


Figure 2-5 Stretched clustered system with physical switches and a quorum disk at a third site

If you do not have enough SAN switches to create two public and two private fabrics, you can use Brocade Virtual Fabrics or Cisco Virtual SANs, as shown in Figure 2-6 on page 29.

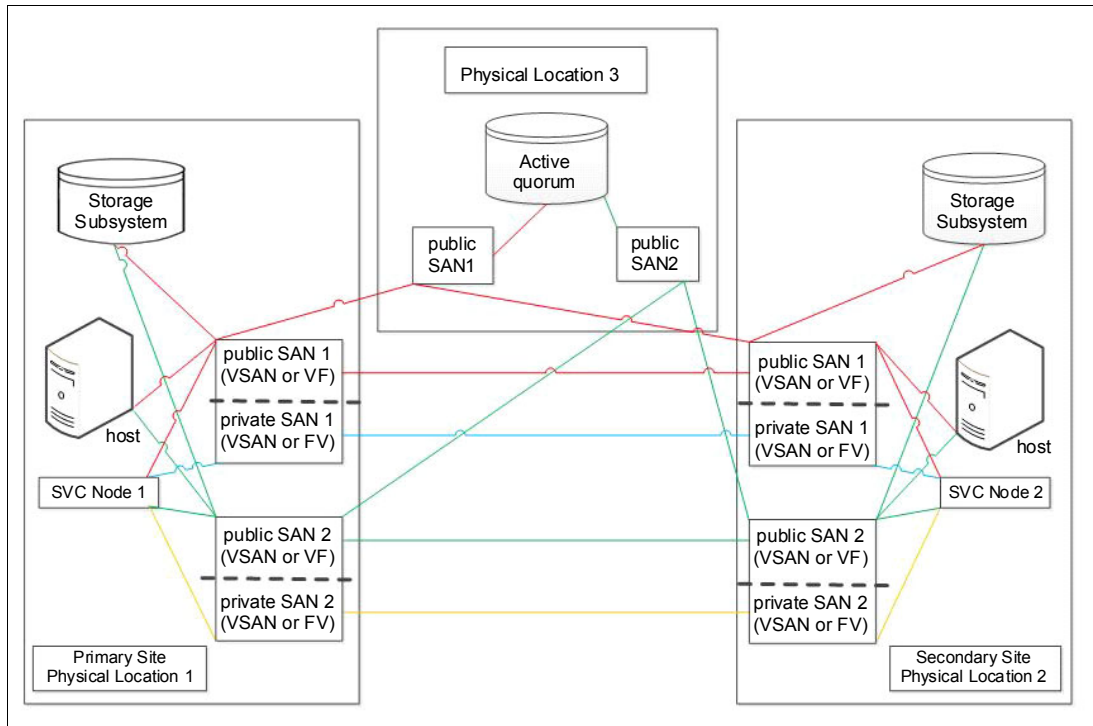


Figure 2-6 SAN Volume Controller stretched cluster with VSANs or Virtual Fabrics

Quorum placement

A stretched clustered system configuration locates the active quorum disk at a third site. If communication is lost between the primary and secondary sites, the site with access to the active quorum disk continues to process transactions. If communication is lost to the active quorum disk, an alternative quorum disk at another site can become the active quorum disk.

Although you can configure a system of SAN Volume Controller nodes to use up to three quorum disks, only one quorum disk can be elected to solve a situation where the system is partitioned into two sets of nodes of equal size. The purpose of the other quorum disks is to provide redundancy if a quorum disk fails before the system is partitioned.

Important: Do not use solid-state drive (SSD) physical disks or managed disks for quorum disk purposes if the SSD lifespan depends on write workload.

Configuration summary

Generally, when the nodes in a system are split among sites, configure the SAN Volume Controller system in the following way:

- ▶ Site 1 has half of the SAN Volume Controller system nodes and one quorum disk candidate.
- ▶ Site 2 has half of the SAN Volume Controller system nodes and one quorum disk candidate.
- ▶ Site 3 has the active quorum disk.

Disable the dynamic quorum configuration by using the **chquorum** command with the **override yes** option.

For more information about Stretched Cluster, see *IBM SAN and SVC Stretched Cluster and VMware Solution Implementation*, SG24-8072.

2.1.9 Enhanced Stretched Cluster

Version 7.2 of SAN Volume Controller code introduced several improvements to the stretched cluster scenario. Now, SAN Volume Controller can be configured in such way that it is aware of its topology. To do so, you must change the system's topology from standard to stretched and then change the nodes' and controllers' sites in a way that nodes and controllers in the same physical location have the same site configured. Then, you can create mirrored volumes in standard way with each copy being placed in separate site.

Note: If you have a running stretched cluster configuration, you can convert it to enhanced stretched cluster nondisruptively after a SAN Volume Controller code upgrade to version 7.2. The conversion can be done only in the command-line interface (CLI).

Another enhancement is the **satask overridequorum** command. By using this command, you can recover from rolling disaster scenarios where one site suffered a disaster and second site, which won the tie break, suffered another disaster. With the manual override quorum feature, you can restore access to volumes on the site that suffered first disaster and now is back online.

Enhanced stretched cluster has the following advantages over normal stretched cluster topology:

- ▶ Connections between local controller MDisks to remote SAN Volume Controller nodes in another site are ignored.
- ▶ Data is transferred between sites a minimum number of times.
- ▶ Write cache destage to the local controller is performed by the local node, even if this node is not an owner of particular volume.
- ▶ Reads are issued to the local copy of a volume.
- ▶ Manual failover is possible because of the manual quorum override feature.

Best practice: If you implement an Enhanced Stretched Cluster scenario, consider the following preferred practices:

- ▶ Use enhanced stretched cluster functionality that was introduced in the v7.2 of SAN Volume Controller code.
- ▶ Use at least two IO groups with each node of an IO group that is placed in separate site.
- ▶ For each mirrored volume, set the `mirrorwritepriority` parameter to `latency`.
- ▶ Manually configure quorum MDisks.

For more information about IBM SAN Volume Controller Enhanced Stretched Cluster, see *IBM SAN and SVC Enhanced Stretched Cluster and VMware Solution Implementation*, SG248211, which is available at this website:

<http://www.redbooks.ibm.com/redpieces/abstracts/sg248211.html?Open>

2.2 SAN switches

You must make several considerations when you select the FC SAN switches for use with your SAN Volume Controller/Storwize installation. To meet design and performance goals, you must understand the features that are offered by the various vendors and associated models.

2.2.1 Selecting SAN switch models

In general, SAN switches come in two classes: fabric switches and directors. Although the classes are normally based on the same software code and Application Specific Integrated Circuit (ASIC) hardware platforms, they have differences in performance and availability.

Directors feature a slotted design and have component redundancy on all active components in the switch chassis (for example, dual-redundant switch controllers). A SAN fabric switch (or a SAN switch) normally has a fixed-port layout in a non-slotted chassis. (An exception is the IBM and Cisco MDS 9200 series, for example, which features a slotted design). Regarding component redundancy, both fabric switches and directors are normally equipped with redundant, hot-swappable environmental components (power supply units and fans).

In the past, when you selected a SAN switch model, you had to consider oversubscription on the SAN switch ports. Here, *oversubscription* refers to a situation in which the combined maximum port bandwidth of all switch ports is higher than what the physical switch internally can switch. For directors, this number can vary for different line card or port blade options. For example, a high port-count module might have a higher oversubscription rate than a low port-count module because the capacity toward the switch backplane is fixed.

With the latest generation of SAN switches (fabric switches and directors), this issue is less important because of increased capacity in the internal switching. This situation is true for both switches with an internal crossbar architecture and switches that are realized by an internal core or edge ASIC lineup.

For modern SAN switches (fabric switches and directors), processing latency from an ingress to egress port is low and is normally negligible.

When you select the switch model, try to consider the future SAN size. It is better to initially get a director with only a few port modules instead of implementing multiple smaller switches. Having a high port-density director instead of several smaller switches also saves ISL capacity and, therefore, ports that are used for interswitch connectivity.

IBM sells and supports SAN switches from the major SAN vendors that are listed in the following product portfolios:

- ▶ IBM System Storage and Brocade b-type SAN portfolio
- ▶ IBM System Storage and Cisco SAN portfolio

2.2.2 Switch port layout for large SAN edge switches

Users of smaller, non-bladed SAN fabric switches generally do not need to be concerned with which ports go where. However, users of multislot directors must pay attention to where the ISLs are in the switch.

Generally, ensure that the ISLs (or ISL trunks) are on separate port modules within the switch to ensure redundancy. Also, spread out the hosts evenly among the remaining line cards in the switch. Remember to locate high-bandwidth hosts on the core switches directly.

2.2.3 Switch port layout for director-class SAN switches

Each SAN switch vendor has a selection of line cards or port blades that are available for their multislotted director-class SAN switch models. Some of these options are oversubscribed, and some of them have full bandwidth that is available for the attached devices. For your core switches, use only line cards or port blades where the full line speed that you expect to use is available. For more information about the full line card or port blade option, contact your switch vendor.

To help prevent the failure of any line card from affecting performance or availability, spread out your SAN Volume Controller/Storwize ports, storage ports, ISLs, and high-bandwidth hosts evenly among your line cards.

2.2.4 Virtual channels

When it comes to connecting SAN Volume Controller/Storwize to FC ports in SAN switches, you must consider the concept of virtual channels, especially if you use ISLs. When you connect two switches with an ISL, this ISL is partitioned into several virtual channels. Four of those virtual channels are designated to transfer data and are assigned five buffer credits for each virtual channel. The IDs of those virtual channels are 2, 3, 4 and 5. The selection of ports that are sending data on a particular virtual channel through ISL is based on the destination port address. The switch assigns physical ports to virtual channel by the last two bits of binary port address. Because of this configuration, we have the following mapping:

- ▶ Ports that end with 00 are assigned virtual channel 2
- ▶ Ports that end with 01 are assigned virtual channel 3
- ▶ Ports that end with 10 are assigned virtual channel 4
- ▶ Ports that end with 11 are assigned virtual channel 5

This configuration applies to all ports in a particular switch, as shown in the following examples:

- ▶ Port 0 (0000) is assigned virtual channel 2
- ▶ Port 1 (0001) is assigned virtual channel 3
- ▶ Port 2 (0010) is assigned virtual channel 4
- ▶ Port 3 (0011) is assigned virtual channel 5
- ▶ Port 4 (0100) is assigned virtual channel 2
- ▶ Port 5 (0101) is assigned virtual channel 3, and so on

When you connect SAN Volume Controller or Storwize to the switch, avoid connecting it to the ports on the same virtual channel; for example, do not connect SAN Volume Controller/Storwize to ports 0, 4, 8, 12, 16, and so on. This might lead to the buffer credit starvation, which can cause congestion and the drop of frames on this particular virtual channel, even if other virtual channels on the same ISL work without any problem.

Figure 2-7 on page 33 shows the correct and incorrect connection schema for non-director class switches.

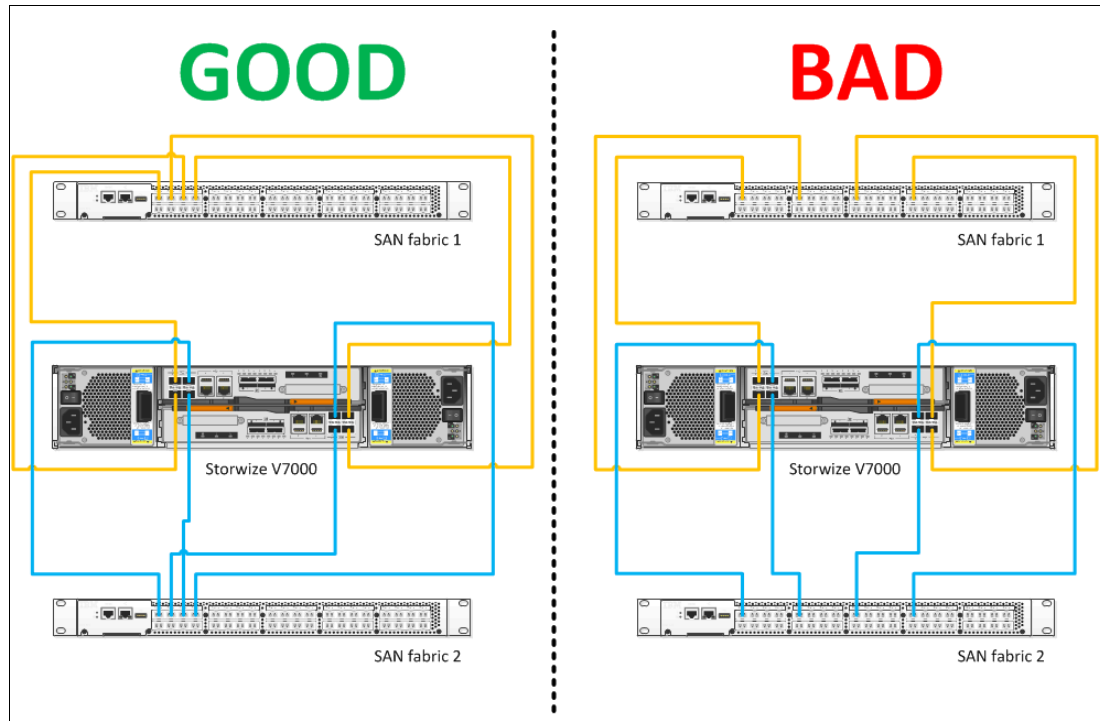


Figure 2-7 Difference between correct and incorrect connection schema

As shown on the left side of Figure 2-7, all ports of Storwize V7000 are connected to the following separate virtual channels in each fabric:

- Fabric 1:
 - Node canister 1, port 1 Ø switch port 0, port ID 0000 0000, virtual channel 2
 - Node canister 2, port 1 Ø switch port 1, port ID 0000 0001, virtual channel 3
 - Node canister 1, port 3 Ø switch port 2, port ID 0000 0010, virtual channel 4
 - Node canister 2, port 3 Ø switch port 3, port ID 0000 0011, virtual channel 5
- Fabric 2:
 - Node canister 1, port 2 Ø switch port 0, port ID 0000 0000, virtual channel 2
 - Node canister 2, port 2 Ø switch port 1, port ID 0000 0001, virtual channel 3
 - Node canister 1, port 4 Ø switch port 2, port ID 0000 0010, virtual channel 4
 - Node canister 2, port 4 Ø switch port 3, port ID 0000 0011, virtual channel 5

In the right side of Figure 2-7, the wrong schema is shown because all of the ports are connected to the following same virtual channel in each fabric:

- Fabric 1:
 - Node canister 1, port 1 Ø switch port 0, port ID 0000 0000, virtual channel 2
 - Node canister 2, port 1 Ø switch port 8, port ID 0000 1000, virtual channel 2
 - Node canister 1, port 3 Ø switch port 16, port ID 0001 0000, virtual channel 2
 - Node canister 2, port 3 Ø switch port 24, port ID 0001 1000, virtual channel 2
- Fabric 2:
 - Node canister 1, port 2 Ø switch port 0, port ID 0000 0000, virtual channel 2
 - Node canister 2, port 2 Ø switch port 8, port ID 0000 1000, virtual channel 2
 - Node canister 1, port 4 Ø switch port 16, port ID 0001 0000, virtual channel 2
 - Node canister 2, port 4 Ø switch port 24, port ID 0001 1000, virtual channel 2

A similar situation occurs with director class SAN switches. The best way is to connect each SAN Volume Controller/Storwize port to separate virtual channels on separate port blade, which is called *diagonal connection*.

Figure 2-8 shows the correct and incorrect cabling for director class switches.

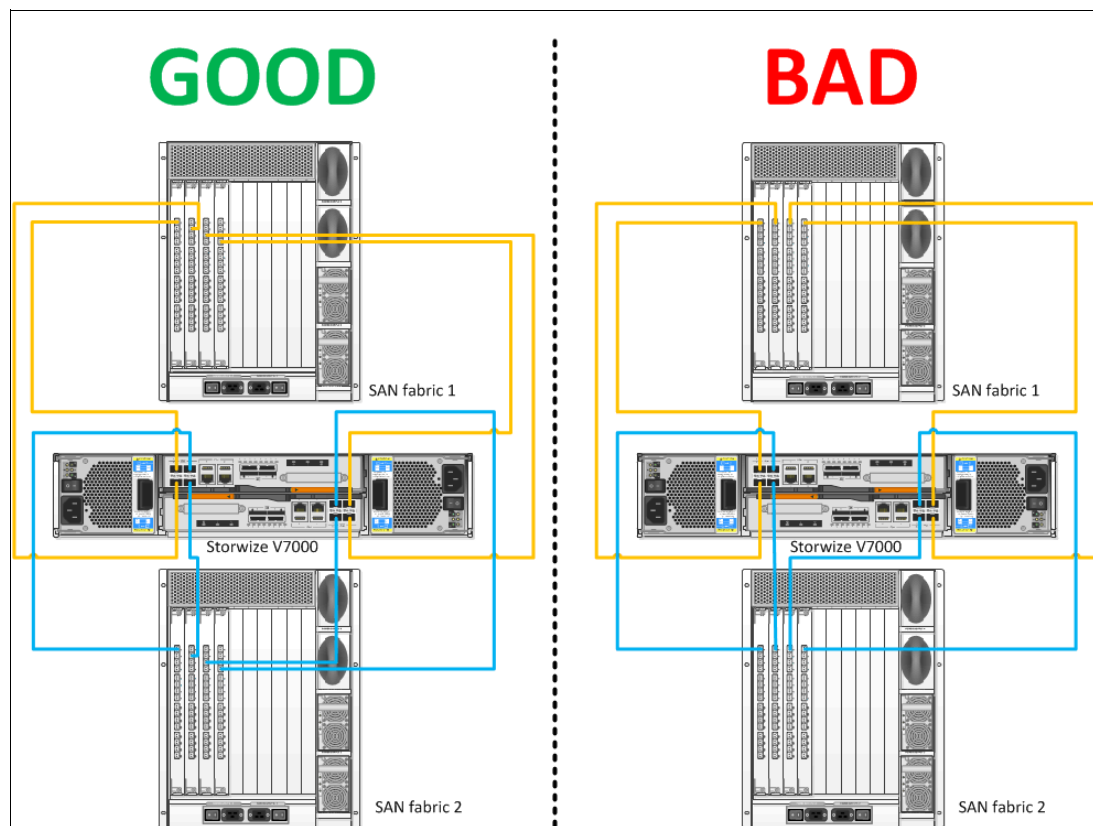


Figure 2-8 Difference between diagonal and incorrect connection schema

As shown in the left side of Figure 2-8, all ports of Storwize V7000 are connected to the following separate virtual channels and to separate port blades in each fabric:

- Fabric 1:
 - Node canister 1, port 1 Ø switch blade1/port 0, port ID 0000 0000, virtual channel 2
 - Node canister 2, port 1 Ø switch blade2/port 1, port ID 0000 0001, virtual channel 3
 - Node canister 1, port 3 Ø switch blade3/port 2, port ID 0000 0010, virtual channel 4
 - Node canister 2, port 3 Ø switch blade4/port 3, port ID 0000 0011, virtual channel 5
- Fabric 2:
 - Node canister 1, port 2 Ø switch blade1/port 0, port ID 0000 0000, virtual channel 2
 - Node canister 2, port 2 Ø switch blade2/port 1, port ID 0000 0001, virtual channel 3
 - Node canister 1, port 4 Ø switch blade3/port 2, port ID 0000 0010, virtual channel 4
 - Node canister 2, port 4 Ø switch blade4/port 3, port ID 0000 0011, virtual channel 5

In the right side of Figure 2-8, the schema is wrong because all ports are connected to the same virtual channel in each of the following fabric even if they are in separate port blades:

- Fabric 1:
 - Node canister 1, port 1 Ø switch blade1/port 0, port ID 0000 0000, virtual channel 2
 - Node canister 2, port 1 Ø switch blade2/port 0, port ID 0000 0000, virtual channel 2

- Node canister 1, port 3 Ø switch blade3/port 0, port ID 0000 0000, virtual channel 2
- Node canister 2, port 3 Ø switch blade4/port 0, port ID 0000 0000, virtual channel 2
- Fabric 2:
 - Node canister 1, port 2 Ø switch blade1/port 0, port ID 0000 0000, virtual channel 2
 - Node canister 2, port 2 Ø switch blade2/port 0, port ID 0000 0000, virtual channel 2
 - Node canister 1, port 4 Ø switch blade3/port 0, port ID 0000 0000, virtual channel 2
 - Node canister 2, port 4 Ø switch blade14/port 0, port ID 0000 0000, virtual channel 2

Best practice: Always connect all SAN Volume Controller/Storwize Fibre Channel ports to separate virtual channels, if possible.

For more information about virtual channels, see *Brocade Fabric OS Administrator's Guide version 7.2*, which is available at this website:

http://www.brocade.com/downloads/documents/product_manuals/B_SAN/FOS_AdminGd_v720.pdf

More information also is available at Seb's sanblog, which is available at this website:

https://www.ibm.com/developerworks/community/blogs/sanblog/entry/how_to_not_connect_an_svc_in_a_core_edge_brocade_fabric16?lang=en

2.2.5 IBM System Storage and IBM b-type SANs

Several practical features of IBM System Storage and IBM b-type SAN switches are available.

Fabric Watch

If the SAN Volume Controller/Storwize relies on a healthy, properly functioning SAN, consider the use of the Fabric Watch feature in newer Brocade-based SAN switches. Fabric Watch is a SAN health monitor that enables real-time proactive awareness of the health, performance, and security of each switch. It automatically alerts SAN managers to predictable problems to help avoid costly failures. It tracks a wide range of fabric elements, events, and counters.

By using Fabric Watch, you can configure the monitoring and measuring frequency for each switch and fabric element and specify notification thresholds. Whenever these thresholds are exceeded, Fabric Watch automatically provides notification by using several methods, including email messages, SNMP traps, log entries, or posts alerts to IBM Network Advisor.

The components that Fabric Watch monitors are grouped into the following classes:

- Environment, such as temperature
- Fabric, such as zone changes, fabric segmentation, and E_Port down
- Field Replaceable Unit, which provides an alert when a part replacement is needed
- Performance Monitor; for example, RX and TX performance between two devices
- Port, which monitors port statistics and takes actions (such as port fencing) that are based on the configured thresholds and actions
- Resource, such as RAM, flash, memory, and processor
- Security, which monitors different security violations on the switch and takes action that is based on the configured thresholds and their actions
- SFP, which monitor the physical aspects of an SFP, such as voltage, current, RXP, TXP, and state changes in physical ports

By implementing Fabric Watch, you benefit by improved high availability from proactive notification. You also can reduce troubleshooting and root cause analysis (RCA) times. Fabric Watch is an optionally licensed feature of Fabric OS. However, it is already included in the base licensing of the new IBM System Storage b-series switches.

Bottleneck detection

A bottleneck is a situation where the frames of a fabric port cannot get through as fast as they should. In this condition, the offered load is greater than the achieved egress throughput on the affected port.

The bottleneck detection feature does not require any other license. It identifies and alerts you to ISL or device congestion and device latency conditions in the fabric. By using bottleneck detection, you can prevent degradation of throughput in the fabric and to reduce the time it takes to troubleshoot SAN performance problems. Bottlenecks are reported through RAS log alerts and SNMP traps, and you can set alert thresholds for the severity and duration of the bottleneck. Starting in Fabric OS 6.4.0, you configure bottleneck detection on a per-switch basis, with per-port exclusions.

The following types of bottleneck detection are available in Brocade b-type switches:

- ▶ Congestion bottleneck detection, which measures utilization of fabric links.
- ▶ Latency bottleneck detection, which indicates of buffer credits starvation.

You can enable bottleneck detection by using the **bottleneckmon** command in the CLI.

Best practice: To spot SAN problems as soon as possible, it is advised to upgrade b-type switches to at least version 7.0 of FOS and enable bottleneck detection.

Virtual Fabrics

Virtual Fabrics adds the capability for physical switches to be partitioned into independently managed logical switches. Implementing Virtual Fabrics has several advantages, such as hardware consolidation, improved security, and resource sharing by several customers.

The following IBM System Storage platforms are Virtual Fabrics capable:

- ▶ SAN768B, SAN768B-2
- ▶ SAN384B, SAN384B-2
- ▶ SAN96B-5
- ▶ SAN80B-4
- ▶ SAN48B-5

To configure Virtual Fabrics, you do not need to install any more licenses.

Fibre Channel routing and Integrated Routing

Fibre Channel routing (FC-FC) is used to forward data packets between two or more (physical or virtual) fabrics while maintaining their independence from each other. Routers use headers and forwarding tables to determine the best path for forwarding the packets. This technology allows the development and management of large heterogeneous SANs, which increases the overall device connectivity.

FC routing has the following advantages:

- ▶ Increases the SAN connectivity interconnecting (not merging) several physical or virtual fabrics.
- ▶ Shares devices across multiple fabrics.
- ▶ Centralizes management.
- ▶ Smooths fabric migrations during technology refresh projects.
- ▶ When used with tunneling protocols (such as FCIP), allows connectivity between fabrics over long distances.

By using the Integrated Routing licensed feature, you can configure 8 Gbps FC ports of SAN768B and SAN384B platforms (among other platforms) as EX_Ports (or VEX_Ports) that support FC routing. By using switches or directors that support the Integrated Routing feature with the respective license, you do not need to deploy external FC routers or FC router blades for FC-FC routing.

For more information about IBM System Storage and Brocade b-type products, see the following IBM Redbooks publications:

- ▶ *Implementing an IBM b-type SAN with 8 Gbps Directors and Switches*, SG24-6116
- ▶ *IBM System Storage b-type Multiprotocol Routing: An Introduction and Implementation*, SG24-7544

2.2.6 IBM System Storage and Cisco SANs

Several practical features of IBM System Storage and Cisco SANs are available.

Port channels

To ease the required planning efforts for future SAN expansions, ISLs or port channels can be made up of any combination of ports in the switch. With this approach, you do not need to reserve special ports for future expansions when you provision ISLs. Instead, you can use any free port in the switch to expand the capacity of an ISL or port channel.

Cisco VSANs

By using VSANs, you can achieve an improved SAN scalability, availability, and security by allowing multiple FC SANs to share a common physical infrastructure of switches and ISLs. These benefits are achieved based on independent FC services and traffic isolation between VSANs. By using Inter-VSAN Routing (IVR), you can establish a data communication path between initiators and targets on different VSANs without merging VSANs into a single logical fabric.

If VSANs can group ports across multiple physical switches, you can use enhanced ISLs to carry traffic that belongs to multiple VSANs (VSAN trunking).

The main VSAN implementation advantages are hardware consolidation, improved security, and resource sharing by several independent organizations. You can use Cisco VSANs with inter-VSAN routes to isolate the hosts from the storage arrays. This arrangement provides little benefit for a great deal of added configuration complexity. However, VSANs with inter-VSAN routes can be useful for fabric migrations that are not from Cisco vendors onto Cisco fabrics, or for other short-term situations.

VSANs can also be useful if you have a storage array that is direct attached by hosts with some space virtualized through the SAN Volume Controller/Storwize. In this case, use separate storage ports for the SAN Volume Controller/Storwize and the hosts. Do not use inter-VSAN routes to enable port sharing.

2.2.7 SAN routing and duplicate worldwide node names

The SAN Volume Controller has a built-in service feature that attempts to detect if two SAN Volume Controller nodes are on the same FC fabric with the same worldwide node name (WWNN). When this situation is detected, the SAN Volume Controller restarts and turns off its FC ports to prevent data corruption. This feature can be triggered erroneously if a SAN Volume Controller port from fabric A is zoned through a SAN router so that a SAN Volume Controller port from the same node in fabric B can log in to the fabric A port.

To prevent this situation from happening, ensure that the routing configuration is correct whenever you are implementing advanced SAN FCR functions.

2.3 Zoning

Because the SAN Volume Controller/Storwize differs from traditional storage devices, properly zoning the SAN Volume Controller/Storwize into your SAN fabric is a source of misunderstanding and errors. Despite the misunderstandings and errors, zoning the SAN Volume Controller/Storwize into your SAN fabric is not complicated.

Important: Errors that are caused by improper SAN Volume Controller/Storwize zoning are often difficult to isolate. Therefore, create your zoning configuration carefully.

Basic SAN Volume Controller/Storwize zoning entails the following tasks:

1. Create the internode communications zone for the SAN Volume Controller. Although this zone is not necessary for Storwize family systems, it is highly recommended to have one.
2. Create a clustered system for the SAN Volume Controller/Storwize.
3. Create a SAN Volume Controller/Storwize Ø Back-end storage subsystem zones.
4. Assign back-end storage to the SAN Volume Controller/Storwize.
5. Create a host Ø SAN Volume Controller/Storwize zones.
6. Create host definitions on the SAN Volume Controller/Storwize.

The zoning scheme that is described in the following section is slightly more restrictive than the zoning that is described in *IBM System Storage SAN Volume Controller V6.4.0 - Software Installation and Configuration Guide*, GC27-2286. The Configuration Guide is a statement of what is supported. However, this Redbooks publication describes the preferred way to set up zoning, even if other ways are possible and supported.

2.3.1 Types of zoning

Modern SAN switches have three types of zoning available: port zoning, WWNN 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.

The use of WWNN zoning is not recommended for many reasons. For hosts, the WWNN is often based on the WWPN of only one of the host bus adapters (HBAs). If you must replace the HBA, the WWNN of the host changes on both fabrics, which results in access loss. In addition, it makes troubleshooting more difficult because you have no consolidated list of which ports are supposed to be in which zone. Therefore, it is difficult to determine whether a port is missing.

IBM and Brocade SAN Webtools users

If you use the IBM and Brocade Webtools GUI to configure zoning, do not use the WWNNs. When you look at the tree of available WWNs, the WWNN is always presented one level higher than the WWPNs (see Figure 2-9). Therefore, make sure that you use a WWPN, not the WWNN.

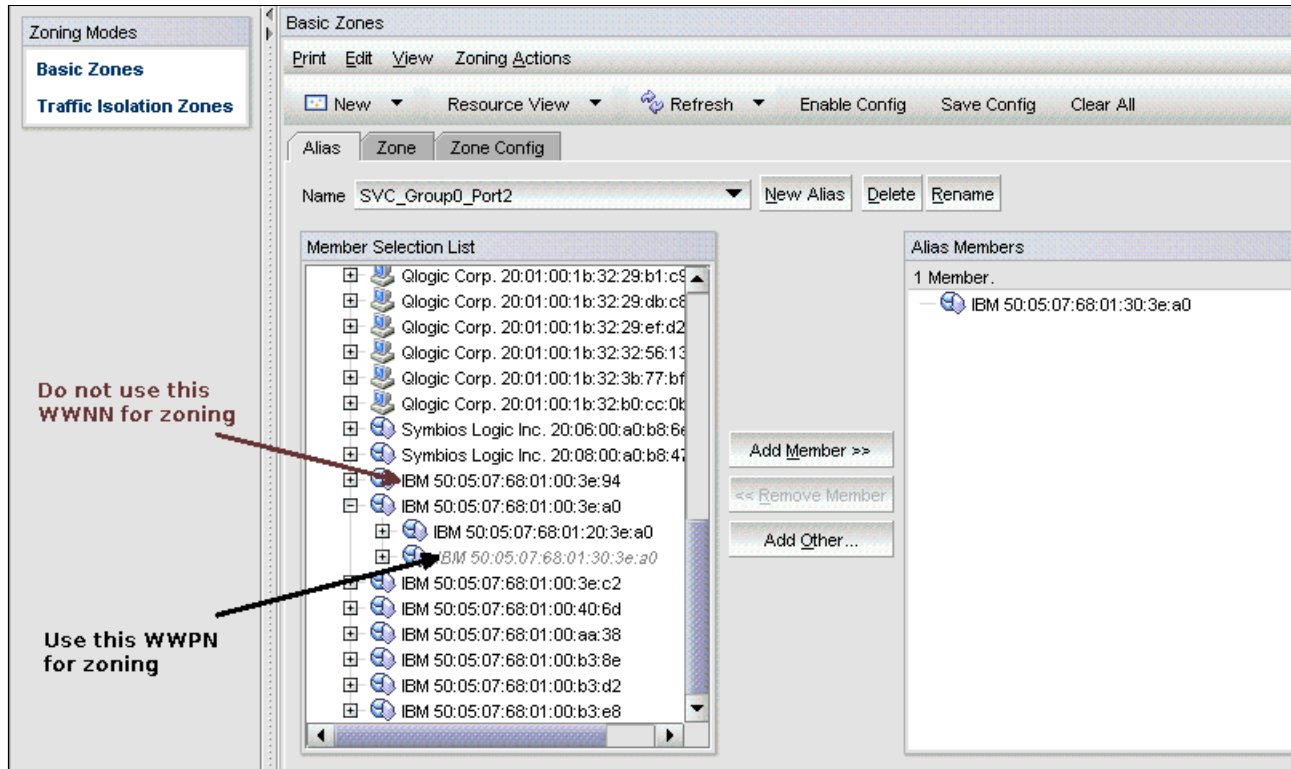


Figure 2-9 IBM and Brocade Webtools zoning

2.3.2 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 detailing. Then, use it consistently whenever you change the environment.

For more information about a SAN Volume Controller/Storwize naming convention, see 14.1.1, “Naming conventions” on page 486.

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 make 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). Create the following zone aliases:

- ▶ One zone alias that holds all the SAN Volume Controller/Storwize node ports on each fabric.
- ▶ One zone alias for each storage subsystem (or controller blade for DS4x00 units).
- ▶ One zone alias for each I/O group port pair (it must contain the first node in the I/O group, port X, and the second node in the I/O group, port X).

You can omit host aliases in smaller environments, as we did in the lab environment that was used for this IBM Redbooks publication. Figure 2-10 on page 41 shows some aliases examples.

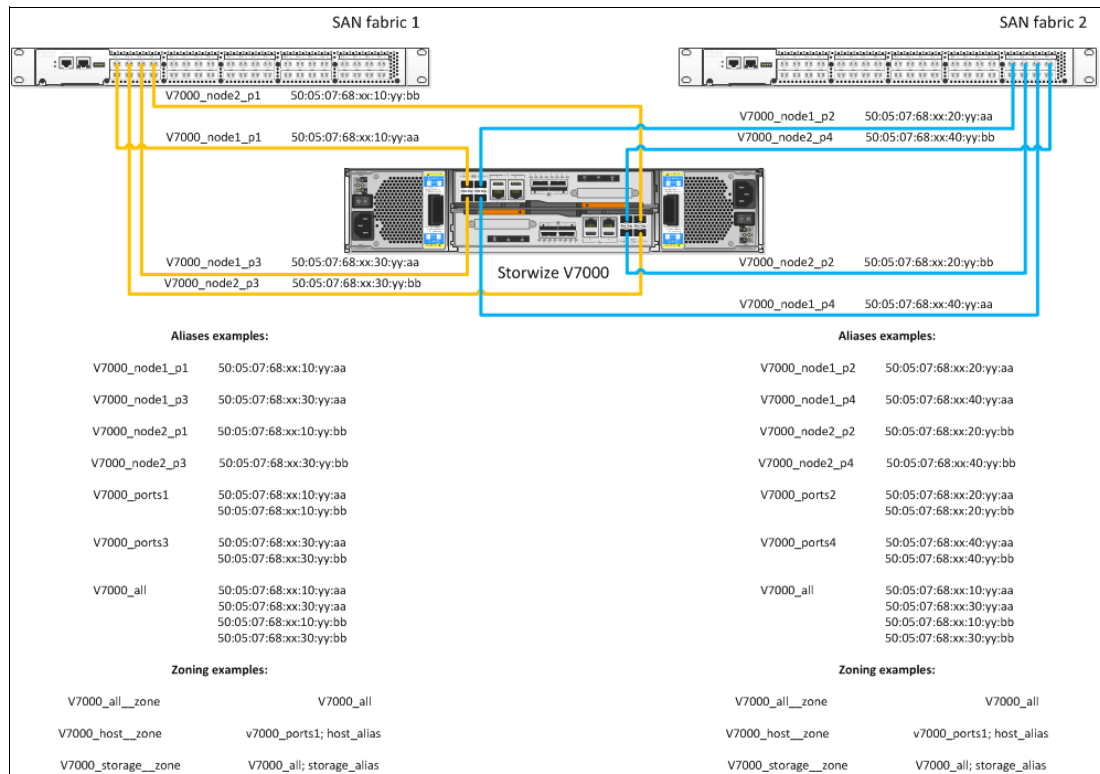


Figure 2-10 Different SAN Volume Controller/Storwize aliasing examples

2.3.3 SAN Volume Controller internode communications zone

The internode communications zone must contain every SAN Volume Controller node port on the SAN fabric. Although it overlaps with the storage zones that you create, it is convenient to have this zone as “fail-safe” in case you make a mistake with your storage zones.

When you configure zones for communication between nodes in the same system, the minimum configuration requires that all FC ports on a node detect at least one FC port on each other node in the same system. You cannot reduce the configuration in this environment.

2.3.4 SAN Volume Controller/Storwize storage zones

Avoid zoning different vendor storage subsystems together. The ports from the storage subsystem must be split evenly across the dual fabrics. Each controller might have its own preferred practice.

All nodes in a system must detect the same ports on each back-end storage system. Operation in a mode where two nodes detect a different set of ports on the same storage system is degraded, and the system logs errors that request a repair action. This situation can occur if inappropriate zoning is applied to the fabric or if inappropriate LUN masking is used.

IBM System Storage DS4000 and DS5000 series

Each IBM System Storage DS4000® and DS5000 series storage controller consists of two separate blades. Do not place these two blades in the same zone if you attached them to the same SAN (see Figure 2-11). Storage vendors other than IBM might have a similar preferred practice. For more information, contact your vendor.

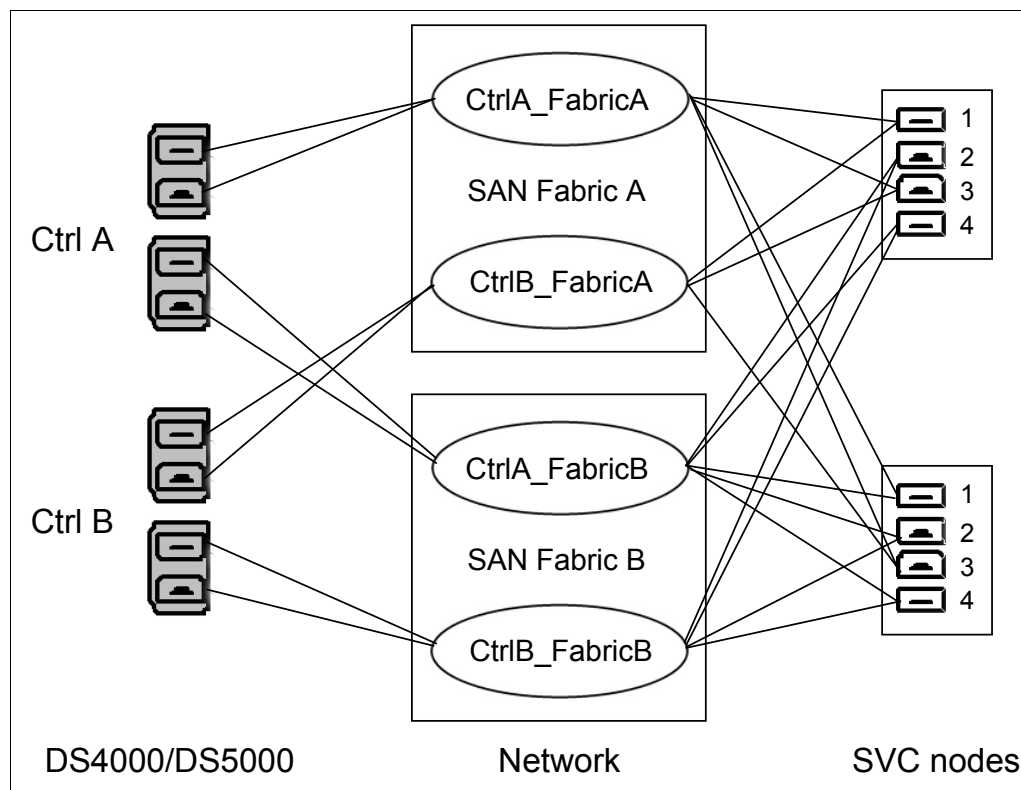


Figure 2-11 Zoning a DS4000 or DS5000 series as a back-end controller

For more information about zoning the IBM System Storage IBM DS4000 or IBM DS5000 series within the SAN Volume Controller/Storwize, see *IBM Midrange System Storage Implementation and Best Practices Guide*, SG24-6363.

XIV storage subsystem

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

Decide which XIV ports you are going to use for connectivity with the SAN Volume Controller/Storwize. If you do not use and do not plan to use XIV remote mirroring, you *must* change the role of port 4 from *initiator* to *target* on all XIV interface modules. You must also use ports 1 and 3 from every interface module in the fabric for the SAN Volume Controller/Storwize attachment. Otherwise, use ports 1 and 2 from every interface module instead of ports 1 and 3. Each HBA port on the XIV Interface Module is designed and set to sustain up to 1,400 concurrent I/Os. However, port 3 sustains only up to 1,000 concurrent I/Os if port 4 is defined as initiator.

Figure 2-12 on page 43 shows how to zone an XIV frame as a SAN Volume Controller storage controller.

Tip: Only single rack XIV configurations are supported by SAN Volume Controller/Storage. Multiple single racks can be supported where each single rack is seen by SAN Volume Controller/Storage as a single controller.

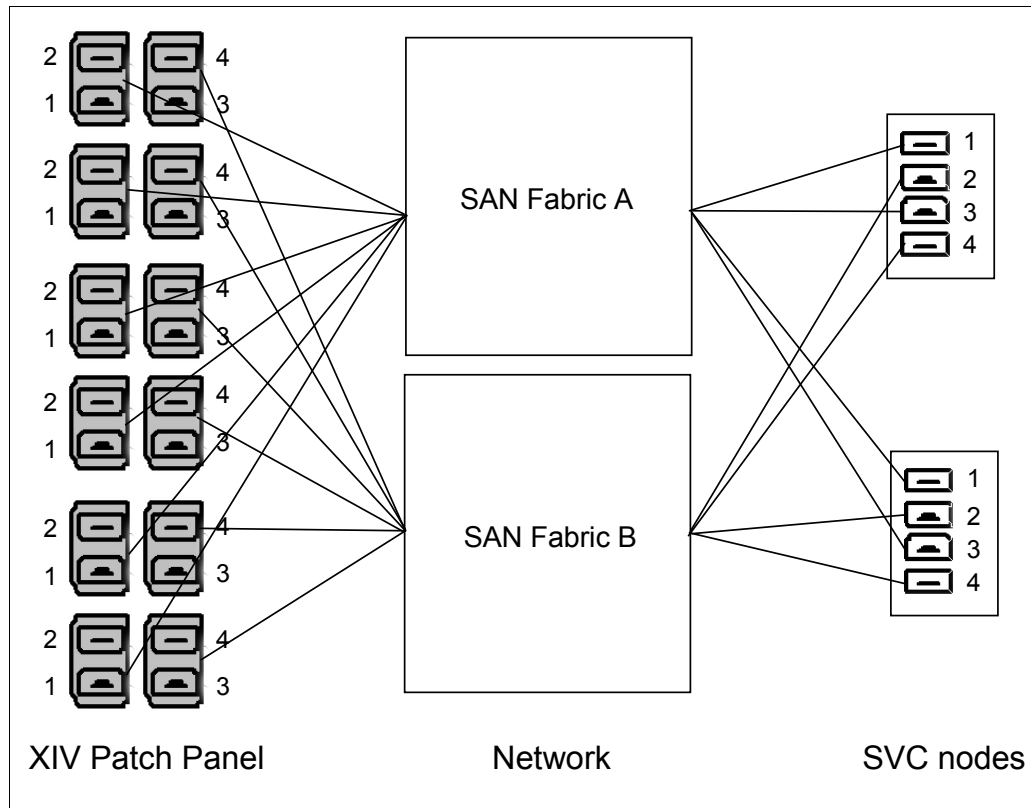


Figure 2-12 Zoning an XIV as a back-end controller

Storwize V7000 storage subsystem

Storwize external storage systems can present volumes to a SAN Volume Controller or to another Storwize system. To zone the Storwize as a back-end storage controller of SAN Volume Controller, every SAN Volume Controller node must have the same Storwize view as a minimum requirement, which must be at least one port per Storwize canister. However, zone all SAN Volume Controller and Storwize ports together in each fabric for best performance and availability.

If you want to virtualize one Storwize by other Storwize, you must change the *layer* of the upper Storwize. By default, SAN Volume Controller includes the layer of *replication* and Storwize with layer of *storage*. Volumes from the storage layer can be presented to replication layer and are seen on the replication layer as MDisk, but not vice versa. Storage layer cannot see a replication layer's MDisk.

SAN Volume Controller layer of replication cannot be changed, so you cannot virtualize SAN Volume Controller behind Storwize. However, layer of Storwize can be changed from storage to replication and from replication to storage. If you want to virtualize one Storwize behind another, the upper Storwize must have layer of replication and the lower Storwize must have the layer of 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 relations, and zoning between Storwize and other Storwize or SAN Volume Controller. Then, use the command `chsystem -layer` to set the layer of the system.

Figure 2-13 shows how you can zone the SAN Volume Controller with the minimum Storwize ports.

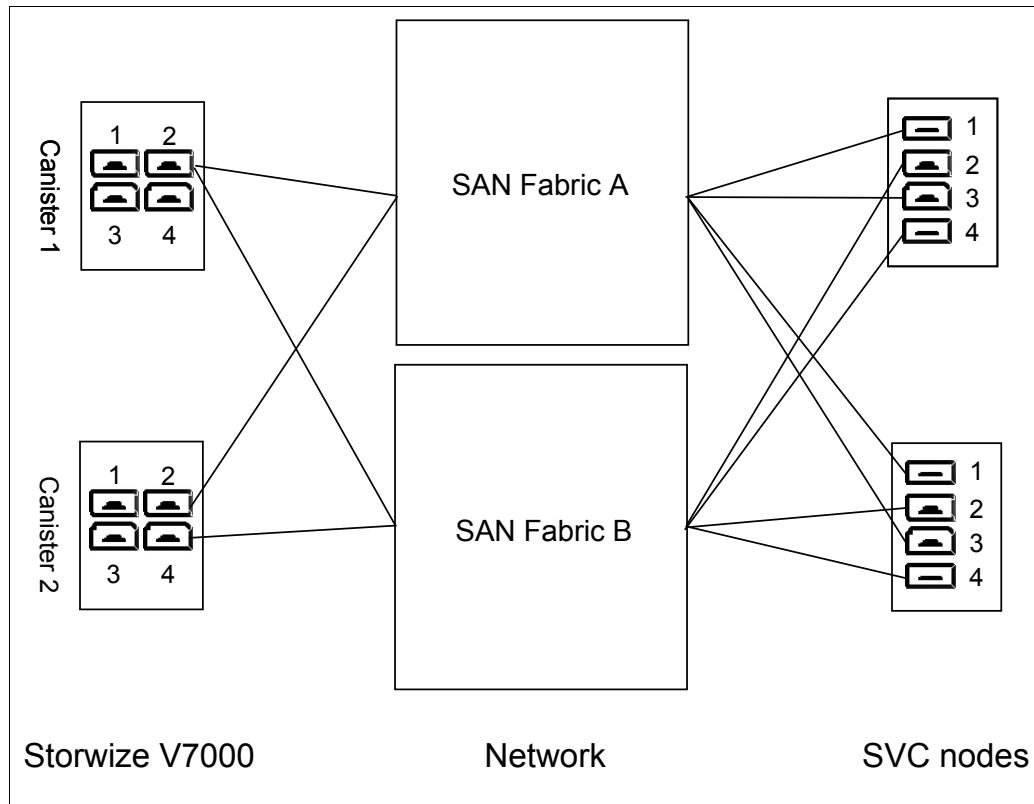


Figure 2-13 Zoning a Storwize V7000 as a back-end controller

2.3.5 SAN Volume Controller/Storwize host zones

Each host port must have a single zone. This zone must contain the host port and *one* port from each SAN Volume Controller/Storwize node that the host must access. Although two ports from each node per SAN fabric are in a usual dual-fabric configuration, ensure that the host accesses only one of them, as shown in Figure 2-14 on page 45.

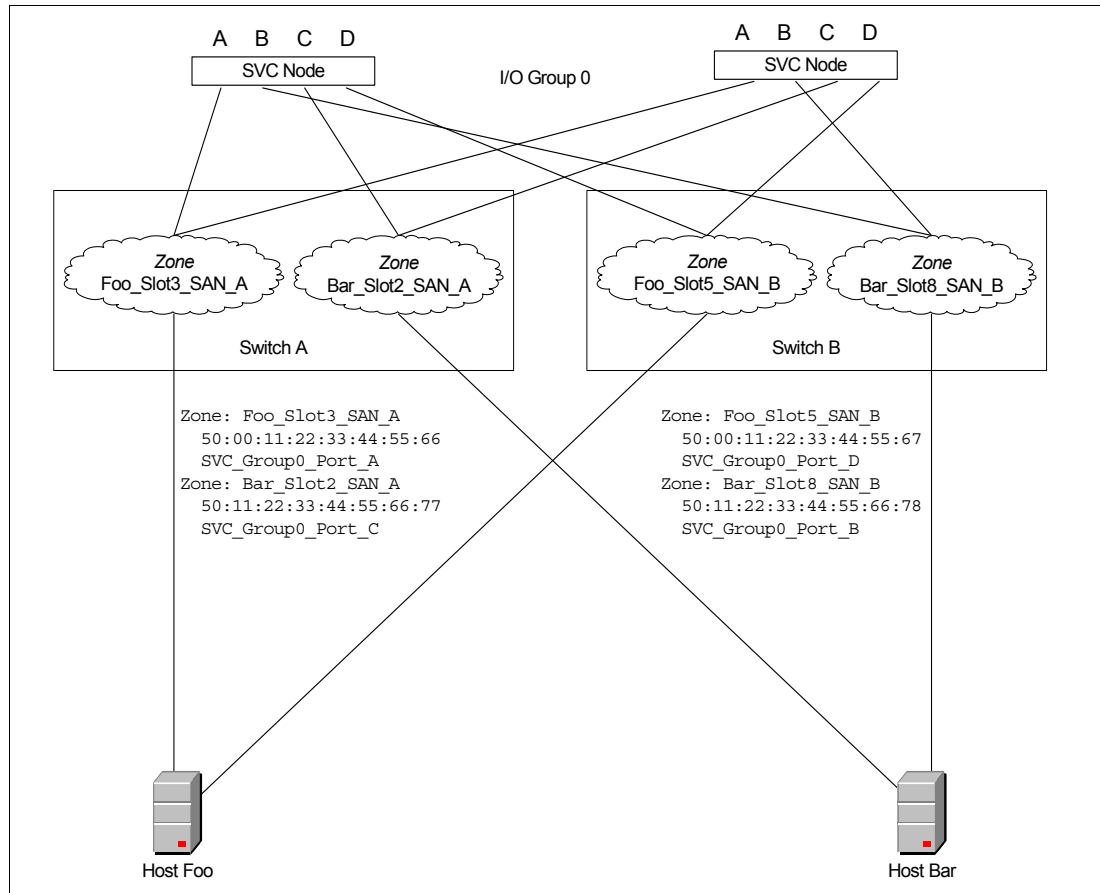


Figure 2-14 Typical host to SAN Volume Controller zoning

This configuration provides four paths to each volume, which is the number of paths per volume for which Subsystem Device Driver (SDDPCM and SDDDSM) multipathing software and the SAN Volume Controller/Storwize are tuned.

For more information about the placement of many hosts in a single zone as a supported configuration in some circumstances, see *IBM System Storage SAN Volume Controller V6.4.0 - Software Installation and Configuration Guide*, GC27-2286. Although this design usually works, instability in one of your hosts can trigger various impossible-to-diagnose problems in the other hosts in the zone. For this reason, you need only a single host in each zone (single initiator zones).

A supported configuration is to have eight paths to each volume. However, this design provides no performance benefit and, in some circumstances, reduces performance. Also, it does not significantly improve reliability nor availability.

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

Hosts with four or more host bus adapters

If you have four HBAs in your host instead of two HBAs, more planning is required. Because eight paths are 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, one node solely serves as a backup node for any specific volume. That is, a preferred node scheme is used. The load is never be balanced 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.

2.3.6 Standard SAN Volume Controller/Storwize zoning configuration

This section provides an example of a “standard” zoning configuration for a SAN Volume Controller clustered system. The setup has two I/O groups, two storage subsystems, and eight hosts, as shown in Figure 2-15. Although this applies to Storwize family systems and the zoning configuration must be duplicated on both SAN fabrics, only the zoning for the SAN Volume Controller and SAN named “SAN A” is shown and explained.

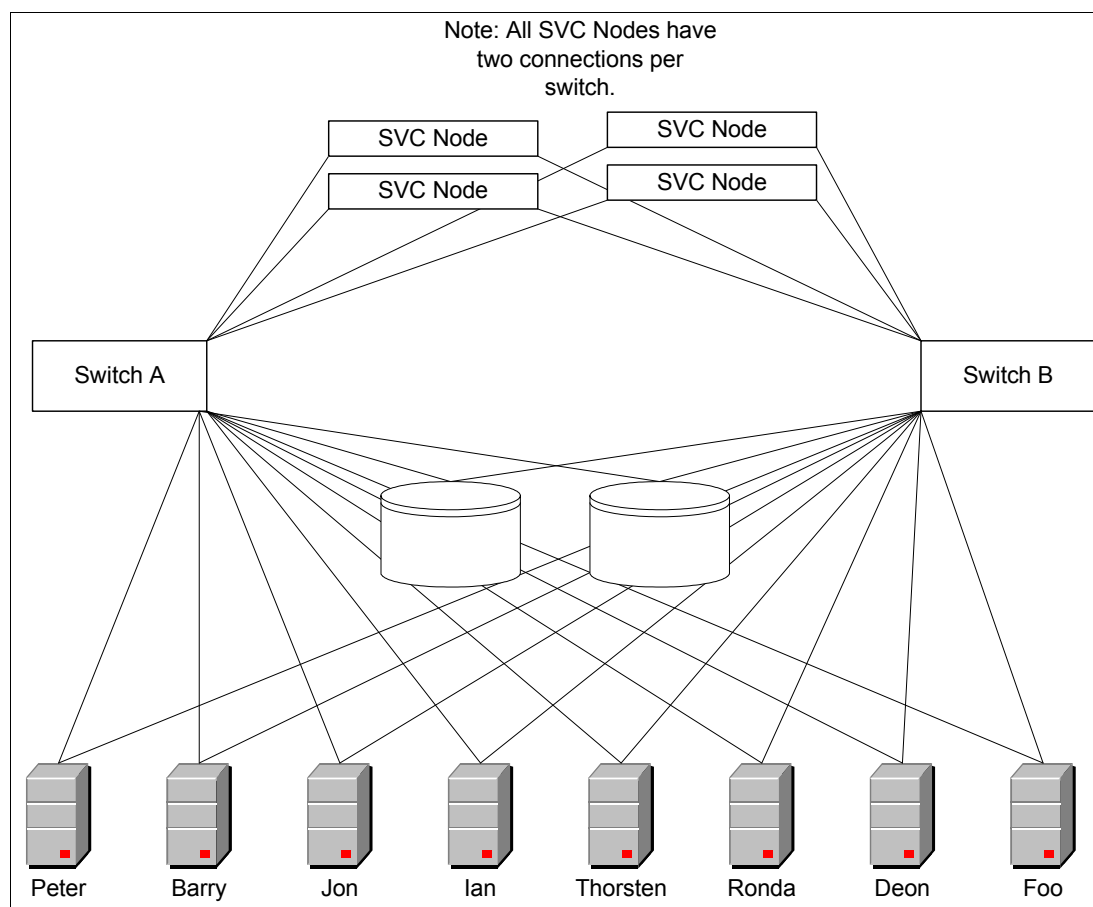


Figure 2-15 SAN Volume Controller SAN

Aliases

Unfortunately, you cannot nest aliases. Therefore, several of the WWPNs appear in multiple aliases. Also, your WWPNs might not look like the ones in the example; some were created when this book was written.

Some switch vendors do not allow multiple-member aliases, but you can still create single-member aliases. Although creating single-member aliases does not reduce the size of your zoning configuration, it still makes it easier to read than a mass of raw WWPNs.

For the alias names, “SAN_A” is appended on the end where necessary to distinguish that these alias names are the ports on SAN A. This system helps if you must troubleshoot both SAN fabrics at one time.

Clustered system alias for SAN Volume Controller

The SAN Volume Controller has a predictable WWPN structure, which helps make the zoning easier to “read.” It always starts with 50:05:07:68 (see Example 2-1) and ends with two octets that distinguish which node is which. The first digit of the third octet from the end identifies the port number in the following way:

- ▶ 50:05:07:68:01:4x:xx:xx refers to port 1.
- ▶ 50:05:07:68:01:3x:xx:xx refers to port 2.
- ▶ 50:05:07:68:01:1x:xx:xx refers to port 3.
- ▶ 50:05:07:68:01:2x:xx:xx refers to port 4.

Example 2-1 SAN Volume Controller clustered system alias

```
SVC_Cluster_SAN_A:  
50:05:07:68:01:40:37:e5  
50:05:07:68:01:10:37:e5  
50:05:07:68:01:40:37:dc  
50:05:07:68:01:10:37:dc  
50:05:07:68:01:40:1d:1c  
50:05:07:68:01:10:1d:1c  
50:05:07:68:01:40:27:e2  
50:05:07:68:01:10:27:e2
```

The clustered system alias that is created is used for the internode communications zone and for all back-end storage zones. It is also used in any zones that you need for remote mirroring with another SAN Volume Controller clustered system (not be addressed in this example).

SAN Volume Controller I/O group ‘port pair’ aliases

I/O group port pair aliases (see Example 2-2) are the basic building blocks of the host zones. Because each HBA is supposed to detect only a single port on each node, these aliases are included. To have an equal load on each SAN Volume Controller node port, you must roughly alternate between the ports when you create your host zones.

Example 2-2 I/O group port pair aliases

```
SVC_Group0_Port1:  
50:05:07:68:01:40:37:e5  
50:05:07:68:01:40:37:dc
```

```
SVC_Group0_Port3:  
50:05:07:68:01:10:37:e5  
50:05:07:68:01:10:37:dc
```

```
SVC_Group1_Port1:  
50:05:07:68:01:40:1d:1c  
50:05:07:68:01:40:27:e2
```

```
SVC_Group1_Port3:  
50:05:07:68:01:10:1d:1c
```

Storage subsystem aliases

The first two aliases that are shown in Example 2-3 are similar to what you might see with an IBM System Storage DS4800 storage subsystem with four back-end ports per controller blade. As shown in Example 2-3, we created different aliases for each blade to isolate the two controllers from each other, as suggested by the DS4000 and DS5000 development teams.

Because the IBM System Storage DS8000® has no concept of separate controllers (at least, not from the SAN viewpoint), we placed all the ports on the storage subsystem into a single alias, as shown in Example 2-3.

Example 2-3 Storage aliases

```
DS4k_23K45_Blade_A_SAN_A
20:04:00:a0:b8:17:44:32
20:04:00:a0:b8:17:44:33
```

```
DS4k_23K45_Blade_B_SAN_A
20:05:00:a0:b8:17:44:32
20:05:00:a0:b8:17:44:33
```

```
DS8k_34912_SAN_A
50:05:00:63:02:ac:01:47
50:05:00:63:02:bd:01:37
50:05:00:63:02:7f:01:8d
50:05:00:63:02:2a:01:fc
```

Zones

When you name your zones, do not give them identical names as aliases. For the environment that is described in this book, we use the following sample zone set, which uses the defined aliases as described in “Aliases” on page 40.

SAN Volume Controller internode communications zone

This zone is simple, as shown in Example 2-4. It contains only a single alias (which happens to contain all of the SAN Volume Controller node ports) and this zone overlaps with every storage zone. Even so, it is good to have it as a fail-safe because of the dire consequences that occur if your clustered system nodes ever completely lose contact with one another over the SAN.

Example 2-4 SAN Volume Controller clustered system zone

```
SVC_Cluster_Zone_SAN_A:
SVC_Cluster_SAN_A
```

SAN Volume Controller storage zones

Recall that we put each storage controller (and, for the DS4000 and DS5000 controllers, each blade) in a separate zone, as shown in Example 2-5 on page 49.

Example 2-5 SAN Volume Controller storage zones

```
SVC_DS4k_23K45_Zone_Blade_A_SAN_A:  
SVC_Cluster_SAN_A  
DS4k_23K45_Blade_A_SAN_A
```

```
SVC_DS4k_23K45_Zone_Blade_B_SAN_A:  
SVC_Cluster_SAN_A  
DS4K_23K45_BLADE_B_SAN_A
```

```
SVC_DS8k_34912_Zone_SAN_A:  
SVC_Cluster_SAN_A  
DS8k_34912_SAN_A
```

SAN Volume Controller host zones

We did not create aliases for each host because each host appears only in a single zone. Although a “raw” WWPN is in the zones, an alias is unnecessary because it is obvious where the WWPN belongs.

All of the zones refer to the slot number of the host rather than “SAN_A.” If you are trying to diagnose a problem (or replace an HBA), you must know on which HBA you must work.

For IBM System p® hosts, we also appended the HBA number into the zone name to make device management easier. Although you can get this information from the SDD, it is convenient to have it in the zoning configuration.

We alternate the hosts between the SAN Volume Controller node port pairs and between the SAN Volume Controller I/O groups for load balancing. However, you might want to balance the load that is based on the observed load on ports and I/O groups, as shown in Example 2-6.

Example 2-6 SAN Volume Controller host zones

```
WinPeter_Slot3:  
21:00:00:e0:8b:05:41:bc  
SVC_Group0_Port1
```

```
WinBarry_Slot7:  
21:00:00:e0:8b:05:37:ab  
SVC_Group0_Port3
```

```
WinJon_Slot1:  
21:00:00:e0:8b:05:28:f9  
SVC_Group1_Port1
```

```
WinIan_Slot2:  
21:00:00:e0:8b:05:1a:6f  
SVC_Group1_Port3
```

```
AIXRonda_Slot6_fcs1:  
10:00:00:00:c9:32:a8:00  
SVC_Group0_Port1
```

```
AIXThorsten_Slot2_fcs0:  
10:00:00:00:c9:32:bf:c7  
SVC_Group0_Port3
```

```
AIXDeon_Slot9_fcs3:  
10:00:00:00:c9:32:c9:6f  
SVC_Group1_Port1
```

```
AIXFoo_Slot1_fcs2:  
10:00:00:00:c9:32:a8:67  
SVC_Group1_Port3
```

Best practice: Although we used raw WWPNs for this example, the preferred practice is to *always* use aliases for your WWPNs and name them in the meaningful manner.

2.3.7 Zoning with multiple SAN Volume Controller/Storwize clustered systems

Unless two separate SAN Volume Controller/Storwize systems participate in a mirroring relationship, configure all zoning so that the two systems do not share a zone. If a single host requires access to two different clustered systems, create two zones with each zone to a separate system. The back-end storage zones must also be separate, even if the two clustered systems share a storage subsystem. You also must zone separate 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.

2.3.8 Split storage subsystem configurations

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

2.4 Switch domain IDs

Ensure that all switch domain IDs are unique between *both* fabrics and that the switch name incorporates the domain ID. Having a unique domain ID makes troubleshooting problems much easier in situations where an error message contains the Fibre Channel ID of the port with a problem. For example, have all domain IDs in first fabric starting with 10 and all domain IDs in second fabric starting with 20.

2.5 Distance extension for remote copy services

To implement remote copy services over a 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 7.2 version of 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.

Distance extension: If possible, use distance extension only for links between SAN Volume Controller clustered systems. Do not use it for intracluster system communication. Technically, distance extension is supported for relatively short distances, such as a few kilometers (or miles). For information about why this arrangement should not be used, see *IBM System Storage SAN Volume Controller Restrictions*, S1003799.

2.5.1 Optical multiplexors

Optical multiplexors can extend your SAN up to hundreds of kilometers (or miles) 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.

2.5.2 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.

2.5.3 Fibre Channel over IP

Fibre Channel over IP (FCIP) conversion is by far the most common and least expensive form of distance extension. It is also a form of distance extension that is more complicated to configure and relatively subtle errors can have severe performance implications.

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 SAN384B-2/SAN768B-2 directors or SAN06B-R FCIP router. 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 always is the case. The design of FC is more sensitive to congestion than most IP applications. You do not want a spyware problem or a spam 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, which 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 (GMVC), make sure that you 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 GM environment within one SAN Volume Controller or one Storwize system, without partnership with another. Use the **chsystem** command with 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 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 7.2*

2.5.4 Native IP replication with 7.2 SAN Volume Controller/Storwize code version

Starting with version 7.2 of SAN Volume Controller/Storwize code, it is possible to implement native IP-based replication. *Native* means that SAN Volume Controller nor Storwize needs 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 7, “Remote copy services” on page 157.

To enable the native IP replication, SAN Volume Controller/Storwize V7.2 implements the Bridgeworks SANslide network optimization technology. For more information about this solution, see *IBM Storwize V7000 and SANslide Implementation*, REDP-5023, which is available at this website:

<http://www.redbooks.ibm.com/abstracts/redp5023.html?Open>

2.6 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.

2.7 Switch interoperability

The 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.

The 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. In general, Brocade interoperates with McDATA under special circumstances. For more information, contact your IBM marketing representative. (*McDATA* refers to the switch products that are sold by the McDATA Corporation before their acquisition by Brocade Communications Systems). QLogic and IBM BladeCenter® FCSM also can work with Cisco.

Do not interoperate Cisco switches with Brocade switches now, except during fabric migrations and only if you have a back-out plan in place. Also, do not connect the QLogic or BladeCenter FCSM to Brocade or McDATA. When you connect 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.

2.8 IBM Tivoli Storage Productivity Center

You can use IBM Tivoli Storage Productivity Center to create, administer, and monitor your SAN fabrics. You do not need to take any extra steps to use it to administer a SAN Volume Controller/Storwize SAN fabric as opposed to any other SAN fabric. For information about Tivoli Storage Productivity Center, see Chapter 13, “Monitoring” on page 357.

For more information, see the following IBM Redbooks publications:

- ▶ *IBM Tivoli Storage Productivity Center V4.2 Release Guide*, SG24-7894
- ▶ *SAN Storage Performance Management Using Tivoli Storage Productivity Center*, SG24-7364

In addition, contact your IBM marketing representative or see the IBM Tivoli Storage Productivity Center Information Center, which is available at this website:

<http://pic.dhe.ibm.com/infocenter/tivihelp/v59r1/index.jsp>

2.9 iSCSI support

iSCSI is a block-level protocol that encapsulates SCSI commands into TCP/IP packets and uses an existing IP network instead of requiring expensive FC HBAs and SAN fabric infrastructure. Since SAN Volume Controller V5.1.0, iSCSI is an alternative to FC host attachment. Nevertheless, all inter-node communications and SAN Volume Controller/Storwize to back-end storage communications (or even with remote clustered systems) are established through the FC or FCoE links.

2.9.1 iSCSI initiators and targets

In an iSCSI configuration, the iSCSI host or server sends requests to a node. The host contains one or more initiators that attach to an IP network to start requests to and receive responses from an iSCSI target. Each initiator and target are given a unique iSCSI name, such as an iSCSI qualified name (IQN) or an extended unique identifier (EUI). An IQN is a 223-byte ASCII name; an EUI is a 64-bit identifier. An iSCSI name represents a worldwide unique naming scheme that is used to identify each initiator or target in the same way that WWNNs are used to identify devices in an FC fabric.

An *iSCSI target* is any device that receives iSCSI commands. The device can be an end node, such as a storage device, or it can be an intermediate device, such as a bridge between IP and FC devices. Each iSCSI target is identified by a unique iSCSI name. The SAN Volume Controller/Storwize can be configured as one or more iSCSI targets. Each node that has one or both of its node Ethernet ports configured becomes an iSCSI target.

To transport SCSI commands over the IP network, an iSCSI driver must be installed on the iSCSI host and target. The driver is used to send iSCSI commands and responses through a network interface controller (NIC) or an iSCSI HBA on the host or target hardware.

2.9.2 iSCSI Ethernet configuration

A clustered system management IP address is used for access to the SAN Volume Controller command-line interface/Storwize (CLI), Console (Tomcat) GUI, and the CIM object manager (CIMOM). Each clustered system has one or two clustered system IP addresses. These IP addresses are bound to Ethernet port one and port two of the current configuration nodes.

You can configure a service IP address per clustered system or per node, and the service IP address is bound to Ethernet port one. Each Ethernet port on each node can be configured with one iSCSI port address. Onboard Ethernet ports can be used for management service or for iSCSI I/O. If you are using IBM Tivoli Storage Productivity Center or an equivalent application to monitor the performance of your SAN Volume Controller/Storwize clustered system, separate this management traffic from iSCSI host I/O traffic. For example, use node port 1 for management traffic and use node port 2 for iSCSI I/O.

2.9.3 Security and performance

All Storwize family systems and all engines that are SAN Volume Controller V7.2 capable support iSCSI host attachments. However, with the new 2145-CG8 node, you can add 10-Gigabit Ethernet connectivity with two ports per Storwize controller enclosure or SAN Volume Controller hardware engine to improve iSCSI connection throughput.

Use a private network between iSCSI initiators and targets to ensure the required performance and security. By using the **cfgportip** command that configures a new port IP address for a node or port, you can set the maximum transmission unit (MTU). The default value is 1500, with a maximum of 9000. With an MTU of 9000 (jumbo frames), you can save CPU utilization and increase efficiency. It reduces the overhead and increases the payload. Jumbo frames provide improved iSCSI performance. You must configure jumbo frames on all network devices on the route from SAN Volume Controller/Storwize to hosts.

Hosts can use standard NICs or converged network adapters (CNAs). For standard NICs, use the operating system iSCSI host-attachment software driver. CNAs can offload TCP/IP processing, and some CNAs can offload the iSCSI protocol. These intelligent adapters release CPU cycles for the main host applications.

For a list of supported software and hardware iSCSI host-attachment drivers, see the following resources:

- For SAN Volume Controller, see *V7.2.x Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller*, S1004453, which is available at this website:

http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004453#_V3K

- For Storwize family systems, see *Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for Flex System V7000 and IBM Storwize V3500, V3700 and V5000*, S1004515, which is available at this website:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004515>

2.9.4 Failover of port IP addresses and iSCSI names

FC host attachment relies on host multipathing software to provide high availability if a node that is in an I/O group is lost. iSCSI allows failover without host multipathing. To achieve this type of failover, the partner node in the I/O group takes over the port IP addresses and iSCSI names of the failed node.

When the partner node returns to the online state, its IP addresses and iSCSI names failback after a delay of 5 minutes. This method ensures that the recently online node is stable before the host uses it for I/O again.

The **svcinfo lsportip** command lists a node's own IP addresses and iSCSI names and the addresses and names of its partner node. The addresses and names of the partner node are identified by the **failover** field that is set to **yes**. The **failover_active** value of **yes** in the **svcinfo lsnode** command output indicates that the IP addresses and iSCSI names of the partner node failed over to a particular node.

2.9.5 iSCSI protocol limitations

When you use an iSCSI connection, consider the following iSCSI protocol limitations:

- ▶ No Service Location Protocol support is available for discovery.
- ▶ Header and data digest support is provided only if the initiator is configured to negotiate.
- ▶ Only one connection per session is supported.
- ▶ A maximum of 256 iSCSI sessions per SAN Volume Controller iSCSI target is supported.
- ▶ Only Error Recovery Level 0 (session restart) is supported.
- ▶ The behavior of a host that supports FC and iSCSI connections and accesses a single volume can be unpredictable and depends on the multipathing software.
- ▶ A maximum of four sessions can come from one iSCSI initiator to a SAN Volume Controller/Storwize iSCSI target.

2.10 SAS support

With the introduction of the newest Storwize family member (at the time of this writing, Storwize V5000 and 7.1 code version), it is possible to directly connect Storwize V3700 or Storwize V5000 to hosts that are equipped with SAS host interface cards (HIC). Both Storwize family members support only direct SAS host attachment. Storwize V3700 can be connected to up to three hosts and Storwize V5000 up to two hosts. Both Storwize storage systems use new mini SAS HD ports; Storwize V7000 uses mini SAS port.

The port difference is shown in Figure 2-16 on page 57.

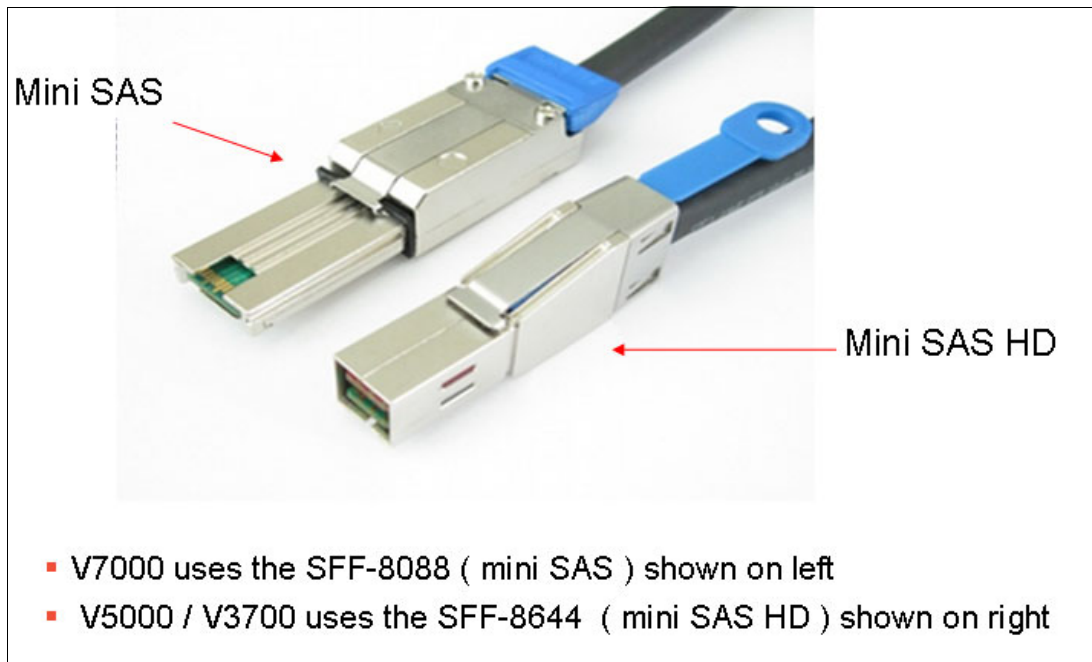


Figure 2-16 Difference between Storwize V7000 and Storwize V3700/V5000 SAS connectors

Storwize V3700 and V5000 have one, four-port SAS per canister node. In Storwize V3700, one of those ports is used for connecting expansion drawers in one chain and three ports are used for host connection. In Storwize V5000, two ports are used for connecting expansion drawers in two chains and two ports are used to connect hosts. Each host must have at least one HIC with two SAS ports because it must be connected to both canister nodes in Storwize V3700 or V5000.

The proper cabling is shown in Figure 2-17.

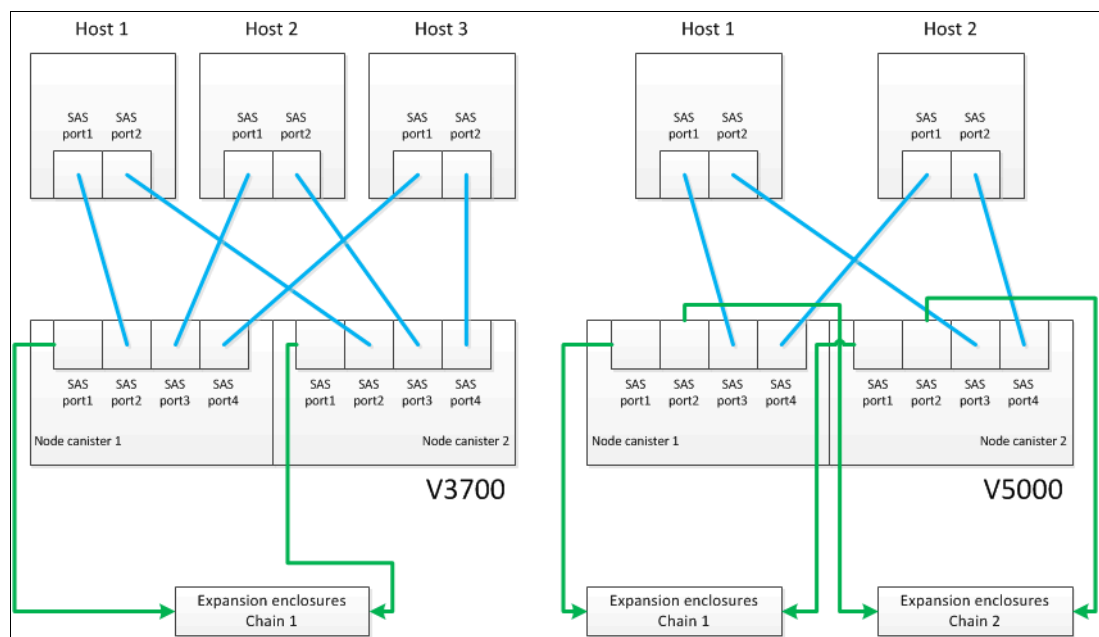


Figure 2-17 Storwize V3700/V5000 possible SAS connections

New GUI options and CLI commands for defining SAS-connected hosts were added to address this new feature.

Also, with 7.2 code version, it is possible to directly connect DS3200 or DS3500 to Storwize V3700 or Storwize V5000 for data migration. As with host connection, there is no support for SAS switches, only direct connection. The external storage SAS connection is not for general virtualization purpose; instead, it is only for data migration. Migrated DS3200/DS3500 storage systems can be directly connected to the SAS host ports in each of Storwize V3700/V5000 node canisters.

For more information about supported servers and storage systems, see *Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for Flex System V7000 and IBM Storwize V3500, V3700 and V5000*, S1004515, which is available at this website:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004515>



SAN Volume Controller and Storwize V7000 Cluster

This chapter highlights the advantages of virtualization and the optimal time to use virtualization in your environment. This chapter also describes the scalability options for the IBM Storwize V7000 and SAN Volume Controller and when to grow or split a SAN Volume Controller clustered system.

This chapter includes the following sections:

- ▶ Advantages of virtualization
- ▶ Scalability of SAN Volume Controller clustered systems
- ▶ Scalability of IBM Storwize V7000
- ▶ Clustered system upgrade

3.1 Advantages of virtualization

The IBM System Storage SAN Volume Controller (as shown in Figure 3-1) enables a single point of control for disparate, heterogeneous storage resources.



Figure 3-1 SAN Volume Controller CG8 model

By using the SAN Volume Controller, you can join capacity from various heterogeneous storage subsystem arrays into one pool of capacity for better utilization and more flexible access. This design helps the administrator to control and manage this capacity from a single common interface instead of managing several independent disk systems and interfaces. The SAN Volume Controller also can improve the performance and efficiency of your storage subsystem array. This improvement is possible by introducing 24 GB of cache memory in each node and the option of the use of internal solid-state drives (SSDs) with the IBM System Storage Easy Tier function.

By using SAN Volume Controller virtualization, users can move data nondisruptively between different storage subsystems. This feature can be useful, for example, when you replace an existing storage array with a new one or when you move data in a tiered storage infrastructure.

By using the Volume mirroring feature, you can store two copies of a volume on different storage subsystems. This function helps to improve application availability if a failure occurs or disruptive maintenance occurs to an array or disk system. Moreover, the two mirror copies can be placed at a distance of 10 km (6.2 miles) when you use longwave (LW) small form factor pluggables (SFPs) with a split-clustered system configuration.

As a virtualization function, thin-provisioned volumes allow provisioning of storage volumes that is based on future growth that requires only physical storage for the current utilization. This feature is best for host operating systems that do not support logical volume managers.

In addition to remote replication services, local copy services offer a set of copy functions. Multiple target FlashCopy volumes for a single source, incremental FlashCopy, and Reverse FlashCopy functions enrich the virtualization layer that is provided by SAN Volume Controller. FlashCopy is commonly used for backup activities and is a source of point-in-time remote copy relationships. Reverse FlashCopy allows a quick restore of a previous snapshot without breaking the FlashCopy relationship and without waiting for the original copy. This feature is convenient, for example, after a failing host application upgrade or data corruption. In such a situation, you can restore the previous snapshot almost instantaneously.

If you are presenting storage to multiple clients with different performance requirements, you can create a tiered storage environment and provision storage with SAN Volume Controller.

3.1.1 SAN Volume Controller features

SAN Volume Controller offers the following features:

- ▶ Combines capacity into a single pool
- ▶ Manages all types of storage in a common way from a common point
- ▶ Improves storage utilization and efficiency by providing more flexible access to storage assets
- ▶ Reduces the physical storage usage when you allocate volumes or convert allocating volumes (formerly volume disks [VDisks]) for future growth by enabling thin provisioning
- ▶ Provisions capacity to applications easier through a new GUI that is based on the IBM XIV interface
- ▶ Provides real-time compression engine to reduce physical storage usage
- ▶ Improves performance through caching, optional SSD utilization, and striping data across multiple arrays
- ▶ Creates tiered storage pools
- ▶ Optimizes SSD storage efficiency in tiering deployments with the Easy Tier feature
- ▶ Provides advanced copy services over heterogeneous storage arrays
- ▶ Removes or reduces the physical boundaries or storage controller limits that are associated with any vendor storage controllers
- ▶ Insulates host applications from changes to the physical storage infrastructure
- ▶ Allows data migration among storage systems without interruption to applications
- ▶ Brings common storage controller functions into the storage area network (SAN) so that all storage controllers can be used and can benefit from these functions
- ▶ Delivers low-cost SAN performance through 1 Gbps and 10 Gbps iSCSI host attachments in addition to Fibre Channel (FC)
- ▶ Enables a single set of advanced network-based replication services that operate in a consistent manner, regardless of the type of storage that is used
- ▶ Improves server efficiency through VMware vStorage APIs, which offloads some storage-related tasks that were previously performed by VMware
- ▶ Improves VMware disaster recovery by integrated SRM plug-in for VMware
- ▶ Enables a more efficient consolidated management with plug-ins to support Microsoft System Center Operations Manager (SCOM) and VMware vCenter

3.2 Scalability of SAN Volume Controller clustered systems

The SAN Volume Controller is highly scalable and can be expanded up to eight nodes in one clustered system. An I/O group is formed by combining a redundant pair of SAN Volume Controller nodes (IBM System x® server-based). Highly available I/O groups are the basic configuration element of a SAN Volume Controller clustered system.

The most recent SAN Volume Controller node (2145-CG8) includes a four-port 8 Gbps-capable host bus adapter (HBA), which allows the SAN Volume Controller to connect and operate at a SAN fabric speed of up to 8 Gbps. It also contains 24 GB of cache memory that is mirrored with the cache of the counterpart node.

In scenarios of compressed environments, it is recommended to add another CPU and 24 GB of RAM. The other CPU and more RAM are used by the system for compression only and not for total performance or read/write cache.

Another four-port, 8 Gbps-capable host bus adapter (HBA) can be added to each node to improve total performance and inter-connect connectivity

Adding I/O groups to the clustered system linearly increases system performance and bandwidth. An entry level SAN Volume Controller configuration contains a single I/O group. The SAN Volume Controller can scale out to support four I/O groups, 1024 host servers, and 8192 volumes (formerly VDisks). This flexibility means that SAN Volume Controller configurations can start small with an attractive price to suit smaller clients or pilot projects, and can grow to manage large storage environments up to 32 PB of virtualized storage.

3.3 Scalability of Storwize V7000

Because the Storwize V7000 is based on the same code level as SAN Volume Controller, many of the characteristics apply. The Storwize V7000 can be expanded to a maximum of four controllers for a maximum of four I/O groups. RAM and CPU cannot be expanded within the Storwize V7000 cluster; therefore, the compression environment should be carefully sized.

3.3.1 Advantage of multiclustered systems versus single-clustered systems

When a configuration limit is reached or when the I/O load reaches a point where a new I/O group is needed, you must decide whether to grow your SAN Volume Controller configuration or add new I/O groups to a SAN Volume Controller clustered system.

Monitor CPU performance

If CPU performance is related to I/O performance and the system concern is related to excessive I/O load, consider monitoring the clustered system nodes. You can monitor the clustered system nodes by using the real-time performance statistics GUI or by using the Tivoli Storage Productivity Center to capture more detailed performance information. You can also use the *unofficially supported* **svcmon** tool, which is available at this website:

<http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS3177>

When the processors consistently become 70% busy, decide whether to add nodes to the clustered system and move part of the workload onto the new nodes, or to move several volumes to a different, less busy I/O group.

The following activities affect CPU utilization:

- ▶ Volume activity

The preferred node is responsible for I/Os for the volume and coordinates sending the I/Os to the alternative node. Although both systems exhibit similar CPU utilization, the preferred node is a little busier. To be precise, a preferred node is always responsible for destaging writes for the volumes that it owns. Therefore, skewing preferred ownership of volumes toward one node in the I/O group leads to more destaging, which also leads to more work on that node.

- ▶ Cache management

The purpose of the cache component is to improve performance of read and write commands by holding part of the read or write data in the memory of SAN Volume Controller. The cache component must keep the caches on both nodes consistent because the nodes in a caching pair have physically separate memories.

- ▶ Mirror Copy activity

The preferred node is responsible for coordinating copy information to the target and for ensuring that the I/O group is current with the copy progress information or change block information. When Global Mirror is enabled, another 10% of overhead occurs on I/O work because of the buffering and general I/O overhead of performing asynchronous Peer-to-Peer Remote Copy (PPRC).

- ▶ Processing I/O requests for thin-provisioned volumes increases SAN Volume Controller CPU overheads.

After you reach the performance or configuration maximum for an I/O group, you can add performance or capacity by attaching another I/O group to the SAN Volume Controller clustered system.

Limits for a SAN Volume Controller I/O group

Table 3-1 on page 64 shows the current maximum limits for one SAN Volume Controller I/O group. Reaching one of the limits on a SAN Volume Controller system that is not fully configured might require the addition of a new pair of nodes (I/O group).

Table 3-1 Storwize/SAN Volume Controller Maximum configurations for an I/O group

Objects	Maximum number	Comments
SAN Volume Controller nodes	8	The nodes are arranged as four I/O groups.
Storwize V7000 Controllers	4	Each Controller includes one I/O group
I/O groups	4	Each group contains two nodes on SAN Volume Controller and one controller on Storwize V7000.
Volumes per I/O group	2048	The I/O group includes managed-mode and image-mode volumes.
Host IDs per I/O group	256	A host object can contain FC ports and iSCSI names.
Host ports (FC and iSCSI) per I/O group	512	N/A
Metro Mirror or Global Mirror volume capacity per I/O group	1024 TB	A per I/O group limit of 1024 TB is placed on the amount of primary and secondary volume address space that can participate in Metro Mirror or Global Mirror relationships. This maximum configuration uses all 512 MB of bitmap space for the I/O group and no FlashCopy bitmap space is available. (The default is 40 TB, which uses 20 MB of bitmap memory.)
FlashCopy volume capacity per I/O group	1024 TB	This capacity is a per I/O group limit on the amount of FlashCopy mappings that use bitmap space from an I/O group. This maximum configuration uses all 512 MB of bitmap space for the I/O group and no Metro Mirror or Global Mirror bitmap space is available. (The default is 40 TB, which uses 20 MB of bitmap memory.)

3.3.2 Growing or splitting SAN Volume Controller clustered systems

Growing a SAN Volume Controller clustered system can be done concurrently, up to a maximum of eight SAN Volume Controller nodes per I/O groups. Table 3-2 shows an extract of the total configuration limits for a SAN Volume Controller clustered system.

Table 3-2 Maximum limits of a SAN Volume Controller clustered system

Objects	Maximum number	Comments
SAN Volume Controller nodes	8	The nodes are arranged as four I/O groups.
MDisks	4096	The maximum number refers to the logical units that can be managed by SAN Volume Controller. This number includes disks that are not configured into storage pools.
Volumes (formerly VDisks) per system	8192	The system includes managed-mode and image-mode volumes. The maximum requires an 8-node clustered system.

Objects	Maximum number	Comments
Total storage capacity manageable by SAN Volume Controller	32 PB	The maximum requires an extent size of 8192 MB.
Host objects (IDs) per clustered system	1024	A host object can contain FC ports and iSCSI names.
Total FC ports and iSCSI names per system	2048	N/A
Total 8 Gbps Fibre Channel ports per system	128	Each IO/Group maximum is 16 Fibre Channel ports; 128 can be reached in an 8-node cluster
Total CPU cores per system	96 cores	Maximum configuration for compression
Total RAM per system	384 GB RAM	Maximum configuration for compression

If you exceed one of the current maximum configuration limits for the fully deployed SAN Volume Controller clustered system, you scale out by adding a SAN Volume Controller clustered system and distributing the workload to it.

Because the current maximum configuration limits can change, see for the current SAN Volume Controller restrictions that is in *IBM System Storage SAN Volume Controller 7.2.x Configuration Limits and Restrictions*, S1004510, which is available at this website:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004510>

By splitting a SAN Volume Controller system or having a secondary SAN Volume Controller system, you can implement a disaster recovery option in the environment. With two SAN Volume Controller clustered systems in two locations, work continues even if one site is down. By using the SAN Volume Controller Advanced Copy functions, you can copy data from the local primary environment to a remote secondary site. The maximum configuration limits also apply.

Another advantage of having two clustered systems is the option of using the SAN Volume Controller Advanced Copy functions. Licensing is based on the following factors:

- ▶ The total amount of storage (in GB) that is virtualized.
- ▶ The Metro Mirror and Global Mirror capacity that is in use (primary and secondary).
- ▶ The FlashCopy source capacity that is in use.

In each case, the number of terabytes (TBs) to order for Metro Mirror and Global Mirror is the total number of source TBs and target TBs that are participating in the copy operations. Because FlashCopy is licensed, SAN Volume Controller now counts as the main source in FlashCopy relationships.

Requirements for growing the SAN Volume Controller clustered system

Before you add an I/O group to the existing SAN Volume Controller clustered system, you must make the following high-level changes:

- ▶ Verify that the SAN Volume Controller clustered system is healthy, all errors are fixed, and the installed code supports the new nodes.
- ▶ Verify that all managed disks are online

- ▶ If you are adding a node that was used previously, consider changing its worldwide node name (WWNN) before you add it to the SAN Volume Controller clustered system. For more information, see Chapter 3, “SAN Volume Controller user interfaces for servicing your system” in *IBM System Storage SAN Volume Controller Troubleshooting Guide*, GC27-2284-01.
- ▶ Install the new nodes and connect them to the local area network (LAN) and SAN.
- ▶ Power on the new nodes.
- ▶ Include the new nodes in the internode communication zones and in the back-end zones.
- ▶ Use LUN masking on back-end storage LUNs (managed disks) to include the worldwide port names (WWPNs) of the SAN Volume Controller nodes that you want to add.
- ▶ Add the SAN Volume Controller nodes to the clustered system
- ▶ Check the SAN Volume Controller status, including the nodes, managed disks, and (storage) controllers.

For more information about adding an I/O group, see “Replacing or adding nodes to an existing clustered system” in the *IBM System Storage SAN Volume Controller Software Installation and Configuration Guide*, GC27-2286-01.

Splitting the SAN Volume Controller clustered system

Splitting the SAN Volume Controller clustered system might become a necessity if the maximum number of eight SAN Volume Controller nodes is reached and you have one or more of the following requirements:

- ▶ To grow the environment beyond the maximum number of I/Os 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 that is described in the “*BM System Storage SAN Volume Controller 7.2.0 Configuration Limits and Restrictions*, S1004510, which is available at this website:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004510>

By splitting the clustered system, you no longer have one SAN Volume Controller clustered system that handles all I/O operations, hosts, and subsystem storage attachments. The goal is to create a second SAN Volume Controller clustered system so that you can equally distribute all of the workload over the two SAN Volume Controller clustered systems.

Approaches for splitting

You can choose from the following approaches to split a SAN Volume Controller clustered system:

- ▶ Create a SAN Volume Controller clustered system, attach storage subsystems and hosts to it, and start putting workload on this new SAN Volume Controller clustered system. This option is probably the easiest approach from a user perspective.
- ▶ Create a SAN Volume Controller clustered system and start moving workload onto it. To move the workload from an existing SAN Volume Controller clustered system to a new SAN Volume Controller clustered system, you can use the Advanced Copy features, such as Metro Mirror and Global Mirror.

Outage: This option move involves an outage from the host system point of view because the WWPN from the subsystem (SAN Volume Controller I/O group) changes.

This option is more difficult, involves more steps (replication services), and requires more preparation in advance. For more information about this option, see Chapter 7, “Remote copy services” on page 157.

- Use the volume “managed-mode-to-image-mode” migration to move workload from one SAN Volume Controller clustered system to the new SAN Volume Controller clustered system. You migrate a volume from managed mode to image mode and reassign the disk (LUN masking) from your storage subsystem point of view. Then, you introduce the disk to your new SAN Volume Controller clustered system and use the image mode to manage mode migration.

Outage: This scenario also invokes an outage to your host systems and the I/O to the involved SAN Volume Controller volumes.

This option involves the longest outage to the host systems; therefore, it is not a preferred option. For more information about this option, see Chapter 6, “Volumes” on page 125.

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 side and spare bandwidth is on the other side.

3.3.3 Adding or upgrading SAN Volume Controller node hardware

Consider a situation where you have a clustered system of six or fewer nodes of older hardware, and you purchased new hardware. In this case, you can choose to start a new clustered system for the new hardware or add the new hardware to the old clustered system. Both configurations are supported.

Although both options are practical, add the new hardware to your existing clustered system if, in the short term, you are not scaling the environment beyond the capabilities of this clustered system.

By using the existing clustered system, you maintain the benefit of managing only one clustered system. Also, if you are using mirror copy services to the remote site, you might be able to continue to do so without adding SAN Volume Controller nodes at the remote site.

Upgrading hardware

You have a few choices to upgrade existing SAN Volume Controller system hardware. Your choice depends on the size of the existing clustered system.

Up to six nodes

If your clustered system has up to six nodes, the following options are available:

- Add the new hardware to the clustered system, migrate volumes to the new nodes, and then retire the older hardware when it is no longer managing any volumes. This method requires a brief outage to the hosts to change the I/O group for each volume.
- Swap out one node in each I/O group at a time and replace it with the new hardware. Contact an IBM service support representative (IBM SSR) to help you with this process. You can perform this swap without an outage to the hosts.

Up to eight nodes

If your clustered system has eight nodes, the following options are available:

- ▶ Swap out a node in each I/O group (one at a time) and replace it with the new hardware. Contact an IBM SSR to help you with this process. You can perform this swap without an outage to the hosts, and you need to swap a node in one I/O group at a time. Do not change all I/O groups in a multi-I/O group clustered system at one time.
- ▶ Move the volumes to another I/O group so that all volumes are on three of the four I/O groups. You can then remove the remaining I/O group with no volumes and add the new hardware to the clustered system.

As each pair of new nodes is added, volumes can then be moved to the new nodes, leaving another old I/O group pair that can be removed. After all the old pairs are removed, the last two new nodes can be added, and, if required, volumes can be moved onto them.

Unfortunately, this method requires several outages to the host because volumes are moved between I/O groups. This method might not be practical unless you must implement the new hardware over an extended period and the first option is not practical for your environment.

Combination of the six node and eight node upgrade methods

You can mix the two options that were described for upgrading SAN Volume Controller nodes.

New SAN Volume Controller hardware provides considerable performance benefits with each release and substantial performance improvements were made since the first hardware release. Depending on the age of your SAN Volume Controller hardware, the performance requirements might be met by only six or fewer nodes of the new hardware.

If this situation fits, you can use a mix of the steps that are described in the six-node and eight-node upgrade methods. For example, use an IBM SSR to help you upgrade one or two I/O groups and then move the volumes from the remaining I/O groups onto the new hardware.

For more information about replacing nodes nondisruptively or expanding an existing SAN Volume Controller clustered system, see *IBM System Storage SAN Volume Controller Software Installation and Configuration Guide Version 6.2.0*, GC27-2286-01.

3.4 Clustered system upgrade

The SAN Volume Controller clustered system and Storwize V7000 performs a concurrent code update. During the automatic upgrade process, each system node is upgraded and restarted sequentially, while its I/O operations are directed to the partner node. This way, the overall concurrent upgrade process relies on I/O group high availability and host multipathing driver. Although the SAN Volume Controller/Storwize V7000 code upgrade is concurrent with multiple host components, such as operating system level, multipath driver, or HBA driver might require updating, which causes the host operating system to be restarted. Plan up front the host requirements for the target SAN Volume Controller code.

Certain concurrent upgrade paths are available only through an intermediate level. For more information, see *SAN Volume Controller Concurrent Compatibility and Code Cross Reference*, S1001707, which is available at this website:

<https://www.ibm.com/support/docview.wss?uid=ssg1S1001707>

Updating the SAN Volume Controller/Storwize V7000 code

Although the SAN Volume Controller code update is concurrent, complete the following steps in advance:

1. Before you apply a code update, ensure that no problems are open in your SAN Volume Controller/Storwize V7000 SAN, or storage subsystems. Use the Run maintenance procedure on the SAN Volume Controller/Storwize V7000 and fix the open problems first. For more information, see 15.3.2, “Solving SAN Volume Controller problems” on page 541.
2. Check your host dual pathing. From the host point of view, make sure that all paths are available. Missing paths can lead to I/O problems during the SAN Volume Controller code update. For more information about hosts, see Chapter 8, “Hosts” on page 225. Also confirm that no hosts have a status of degraded.
3. Run the `svc_snap -c` command and copy the `tgz` file from the clustered system. The `-c` flag enables running a fresh `config_backup` (configuration backup) file.
4. Schedule a time for the SAN Volume Controller/Storwize V7000 code update during low I/O activity.
5. Upgrade the Master Console GUI before the SAN Volume Controller I/O group.
6. Allow the SAN Volume Controller/Storwize V7000 code update to finish before you make any other changes in your environment.
7. Allow at least one hour to perform the code update for a single SAN Volume Controller/Storwize V7000 I/O group and 30 minutes for each additional I/O group. In a worst-case scenario, an update can take up to two and a half hours, which implies that the SAN Volume Controller code update also updates the BIOS, SP, and the SAN Volume Controller service card.

Important: The concurrent code upgrade might appear to stop for some time (up to an hour) if it is upgrading a low-level BIOS. *Never power off* during a concurrent code upgrade unless you are instructed to do so by IBM service personnel. If the upgrade encounters a problem and fails, the upgrade is backed out.

New features are not available until all of the nodes in the clustered system are at the same level. Features that depend on a remote clustered system, such as Metro Mirror or Global Mirror, might not be available until the remote cluster is at the same level.



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:

- ▶ Controller affinity and preferred path
- ▶ Considerations for DS4000 and DS5000 series
- ▶ Considerations for DS8000 series
- ▶ Considerations for IBM XIV Storage System
- ▶ Considerations for IBM Storwize V7000/V5000/V3700
- ▶ Considerations for IBM FlashSystem
- ▶ Considerations for third-party storage with EMC Symmetrix DMX and Hitachi Data Systems
- ▶ Medium error logging
- ▶ Mapping physical LBAs to volume extents
- ▶ Identifying storage controller boundaries by using the IBM Tivoli Storage Productivity Center

4.1 Controller affinity and preferred path

This section describes the architectural differences between common storage subsystems in terms of *controller affinity* (also referred to as *preferred controller*) and *preferred path*. In this context, *affinity* refers to the controller in a dual-controller subsystem that is assigned access to the back-end storage for a specific LUN under nominal conditions (both active controllers). *Preferred path* refers to the host-side connections that are physically connected to the controller that has the assigned affinity for the corresponding LUN that is being accessed.

All storage subsystems that incorporate a dual-controller architecture for hardware redundancy employ the concept of affinity. For example, if a subsystem has 100 LUNs, 50 of those LUNs have an affinity to controller 0, and 50 of them have an affinity to controller 1. Only one controller is serving any specific LUN at any specific time. However, the aggregate workload for all LUNs is evenly spread across both controllers. Although this relationship exists during normal operation, each controller can control all 100 LUNs if a controller failure occurs.

For the IBM System Storage DS4000 or DS5000 series, preferred path is important because Fibre Channel (FC) cards are integrated into the controller. This architecture allows dynamic multipathing and active/standby pathing through FC cards that are attached to the same controller and an alternative set of paths. The alternative set of paths is configured to the other controller that is used if the corresponding controller fails.

For example, if each controller is attached to hosts through two FC ports, 50 LUNs use the two FC ports in controller 0, and 50 LUNs use the two FC ports in controller 1. If either controller fails, the multipathing driver fails the 50 LUNs that are associated with the failed controller over to the other controller, and all 100 LUNs use the two ports in the remaining controller. The DS4000/DS5000 series differs from the IBM System Storage DS8000 series because it can transfer ownership of LUNs at the LUN level as opposed to the controller level.

For the DS8000 series, the concept of preferred path is not used because FC cards are outboard of the controllers. Therefore, all FC ports are available to access all LUNs, regardless of cluster affinity. Although cluster affinity still exists, the network between the outboard FC ports and the controllers performs the appropriate controller routing. This approach is different from the DS4000/DS5000 series, in which controller routing is performed by the multipathing driver on the host, such as with Subsystem Device Driver Path Control Module (SDDPCM) and Redundant Disk Array Controller (RDAC).

4.2 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 (DS8000) 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 case of storage systems without preferred controller like XIV or DS8000, each MDisk uses all available FC ports of that storage controller. In the case of active-passive systems, such as DS4000 or DS5000 series, each MDisk uses all available ports from preferred 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 I/Os to many backend FC ports in parallel. Particularly, the sequential I/O within a single extent can benefit from this feature.

Additionally, round robin path selection improves resilience to certain storage system failures. For example, if one of the backend storage system FC ports has some performance problems, the I/O to MDisk is sent via other ports. Moreover, because I/Os 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 this writing, the Round Robin Path Selection is supported on the following storage systems:

- ▶ IBM Storwize V3700, V5000, and V7000
- ▶ IBM FlashSystem 710, 720, 810, and 820
- ▶ IBM DS8100, DS8300, DS8700, DS8800, and DS8870
- ▶ IBM XIV
- ▶ IBM DS5020, DS5100, and DS5300
- ▶ IBM DS4800
- ▶ EMC Symmetrix (including DMX and VMAX)
- ▶ Fujitsu Eternus
- ▶ Violin Memory 3100, 3200, and 6000

For more information about the latest updates of this list, see *V7.2.x Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller*, which is available at this website:

http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004453#_V3K

4.3 Considerations for DS4000 and DS5000 series

When you configure the controller for IBM System Storage DS4000 or DS5000 series, you must remember the considerations that are described in this section.

4.3.1 Setting the DS4000 and DS5000 series so that both controllers have the same worldwide node name

The SAN Volume Controller/Storwize recognizes that the DS4000 and DS5000 series controllers belong to the same storage system unit if they both have the same worldwide node name (WWNN). You can choose from several methods to determine whether the WWNN is set correctly for SAN Volume Controller/Storwize. From the SAN switch GUI, you can check whether the worldwide port name (WWPN) and WWNN of all devices are logged in to the fabric. You confirm that the WWPN of all DS4000 or DS5000 series host ports are unique but that the WWNNs are identical for all ports that belong to a single storage unit.

You can obtain the same information from the Controller section when you view the Storage Subsystem Profile from the Storage Manager GUI. This section lists the WWPN and WWNN information for each host port, as shown in the following example:

```
World-wide port identifier: 20:27:00:80:e5:17:b5:bc  
World-wide node identifier: 20:06:00:80:e5:17:b5:bc
```

If the controllers are set up with different WWNNs, run the **SameWWN.script** script that is bundled with the Storage Manager client download file to change it.

Attention: This procedure is intended for initial configuration of the DS4000 or DS5000 series. Do *not* run the script in a live environment because all hosts that access the storage subsystem are affected by the changes.

4.3.2 Balancing workload across DS4000 and DS5000 series controllers

When you create arrays, spread the disks across multiple controllers and alternating slots within the enclosures. This practice improves the availability of the array by protecting against enclosure failures that affect multiple members within the array. It also improves performance by distributing the disks within an array across drive loops. You spread the disks across multiple enclosures and alternating slots within the enclosures by using the manual method for creating an array.

Figure 4-1 on page 75 shows a Storage Manager view of a 2+p array that is configured across enclosures. Here, you can see that each of the three disks is represented in a separate physical enclosure and that slot positions alternate from enclosure to enclosure.

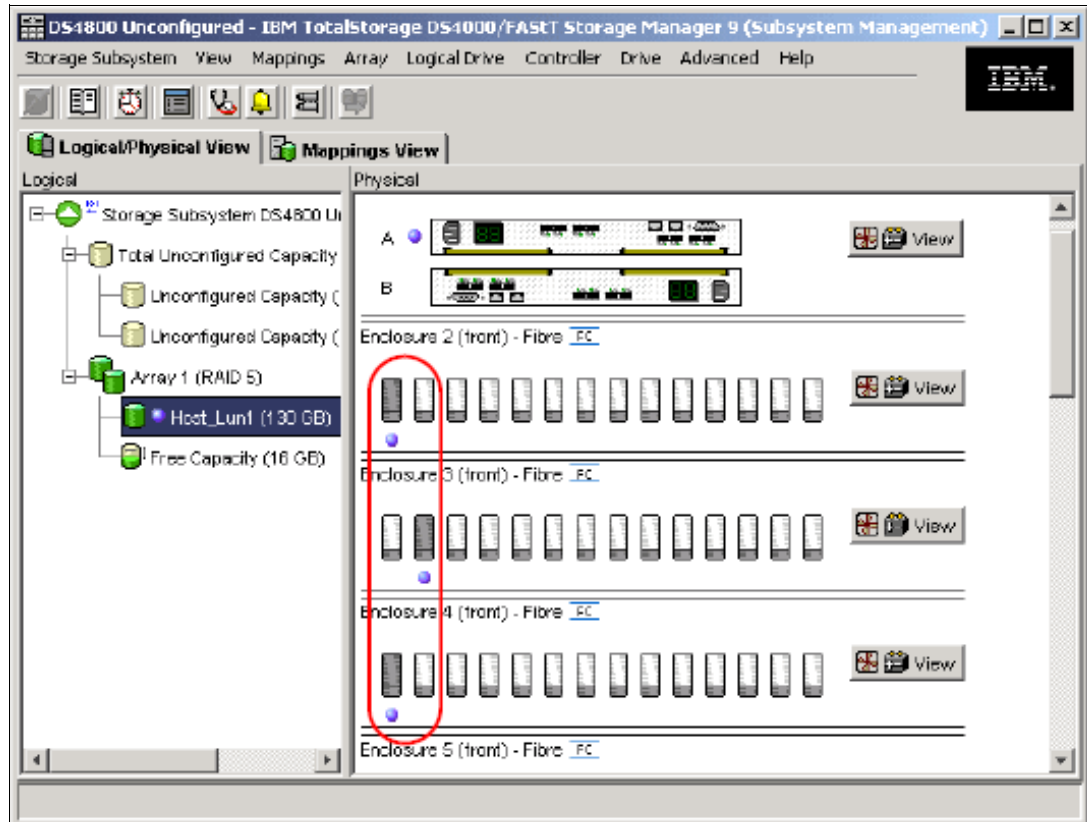


Figure 4-1 Storage Manager view

4.3.3 Ensuring path balance before MDisk discovery

Before you perform MDisk discovery, properly balance LUNs across storage controllers. Failing to properly balance LUNs across storage controllers in advance can result in a suboptimal pathing configuration to the back-end disks, which can cause a performance degradation. You must also ensure that storage subsystems have all controllers online and that all LUNs were distributed to their preferred controller (local affinity). Pathing can always be rebalanced later, but often not until after lengthy problem isolation occurs.

If you discover that the LUNs are not evenly distributed across the dual controllers in a DS4000 or DS5000, you can dynamically change the LUN affinity. However, the SAN Volume Controller/Storwize moves them back to the original controller, and the storage subsystem generates an error message that indicates that the LUN is no longer on its preferred controller. To fix this situation, run the `svctask detectmdisk` SAN Volume Controller/Storwize command, or use the Detect MDisk GUI option. SAN Volume Controller/Storwize queries the DS4000 or DS5000 again and accesses the LUNs through the new preferred controller configuration.

4.3.4 Auto-Logical Drive Transfer for the DS4000 and DS5000 series (firmware version before 7.83.x)

The DS4000 and DS5000 series have a feature that is called Auto-Logical Drive Transfer (ADT), which allows logical drive-level failover as opposed to controller level failover. When you enable this option, the DS4000 or DS5000 series moves LUN ownership between controllers according to the path that is used by the host.

For the SAN Volume Controller/Storwize, the ADT feature is enabled by default when you select **IBM TS SAN VCE** as the host type.

IBM TS SAN VCE: When you configure the DS4000 or DS5000 series for SAN Volume Controller or Storwize attachment, select the **IBM TS SAN VCE** host type so that the SAN Volume Controller/Storwize can properly manage the back-end paths. If the host type is incorrect, SAN Volume Controller/Storwize reports error 1625 (incorrect controller configuration).

For more information about checking the back-end paths to storage controllers, see Chapter 15, “Troubleshooting and diagnostics” on page 519.

4.3.5 Asymmetric Logical Unit Access for the DS4000 and DS5000 series (firmware 7.83.x and later)

With the firmware version of 7.83.x, the new setting was introduced called Asymmetric Logical Unit Access (ALUA). ALUA replaces the ADT setting for SAN Volume Controller/Storwize starting with DS4000/DS5000 series firmware v7.83.x.

With ALUA compatible storage systems, the controllers are no longer active-passive but act as active-active. LUNs still have a controller affinity; however, if all preferred paths fail, multipath driver redirects all of the I/O to the non-preferred paths, which is the non-preferred controller. In cases where preferred controller works, all I/O is redirected from non-preferred controller to the preferred controller. The controllers do not change the ownership of the LUNs if this condition lasts less than 5 minutes. After 5 minutes, non-preferred controller stops redirecting I/O to preferred controller and takes ownership of the LUNs.

ALUA features the following advantages:

- ▶ Boot from SAN does not fail if the boot LUN is not on preferred path.
- ▶ Eliminates LUN failover/fallback if there are transitory path interruptions.
- ▶ Prevents LUN trashing in clustered environments.
- ▶ I/O can be sent to both controllers.

Note: IBM SAN Volume Controller and Storwize family storage systems are ALUA capable.

For more information about ALUA, see *Installation and Host Support Guide v10.8 - IBM System Storage DS Storage Manager*, which is available at this website:

<https://www-947.ibm.com/support/entry/myportal/docdisplay?ln docid=MIGR-5090826&brandind=5000028>

4.3.6 Selecting array and cache parameters

When you define the SAN Volume Controller array and cache parameters, you must consider the settings of the array width, segment size, and cache block size.

DS4000 and DS5000 series array width

With Redundant Array of Independent Disks 5 (RAID 5) arrays, determining the number of physical drives to place into an array always presents a compromise. Striping across a larger number of drives can improve performance for transaction-based workloads. However, striping can also have a negative effect on sequential workloads.

A common mistake that people make when they select an array width is the tendency to focus only on the capability of a single array to perform various workloads. However, you must also consider in this decision the aggregate throughput requirements of the entire storage server. Many physical disks in an array can create a workload imbalance between the controllers because only one controller of the DS4000 or DS5000 series actively accesses a specific array.

When you select array width, you must also consider its effect on rebuild time and availability.

A larger number of disks in an array increases the rebuild time for disk failures, which can have a negative effect on performance. Also, more disks in an array increase the probability of having a second drive fail within the same array before the rebuild completion of an initial drive failure, which is an inherent exposure to the RAID 5 architecture.

Best practice: For the DS4000 or DS5000 series, use array widths of 4+p and 8+p for RAID5, and 4+4 or 8+8 for RAID10.

Segment size

With direct-attached hosts, considerations are often made to align device data partitions to physical drive boundaries within the storage controller. For the SAN Volume Controller/Storwize, aligning device data partitions to physical drive boundaries within the storage controller is less critical. The reason is based on the caching that the SAN Volume Controller/Storwize provides and on the fact that less variation is in its I/O profile, which is used to access back-end disks.

For the SAN Volume Controller/Storwize, the only opportunity for a full stride write occurs with large sequential workloads. In that case, the larger the segment size is, the better. However, larger segment sizes can adversely affect random I/O. The SAN Volume Controller/Storwize and controller cache hide the RAID 5 write penalty for random I/O well. Therefore, larger segment sizes can be accommodated. The primary consideration for selecting segment size is to ensure that a single host I/O fits within a single segment to prevent access to multiple physical drives.

Testing demonstrated that the best compromise for handling all workloads is to use a segment size of 256 KB.

Best practice: Use a segment size of 256 KB as the best compromise for all workloads.

Cache block size

The size of the cache memory allocation unit can be 4 K, 8 K, 16 K, or 32 K. Earlier models of the DS4000 system that use the 2 Gb FC adapters have their block size configured as 4 KB by default. For the newest models (on firmware 7.xx and later), the default cache memory is 8 KB.

In SAN Volume Controller/Storwize code version 6.x, the maximum I/O size to external MDisks was changed 32 - 256 KB. Additionally, SAN Volume Controller/Storwize cache algorithms tries to put as much data as it can into one I/O to backend storage. For best performance, we advise that you set the cache block size to larger than the default value. This is especially true when compressed volumes are used. All I/O from compressed volumes to back-end storage is always 32 KB.

Best practice: Set the cache block size to 32 KB and use the IBM TS SAN VCE host type to establish the correct cache block size for the SAN Volume Controller cluster or Storwize storage system.

Table 4-1 shows the values for SAN Volume Controller/Storwize and DS4000 or DS5000 series.

Table 4-1 SAN Volume Controller/Storwize values

Models	Attribute	Value
SAN Volume Controller/Storwize	Extent size (MB)	256
SAN Volume Controller/Storwize	Managed mode	Striped
DS4000 or DS5000 series	Segment size (KB)	256
DS4000 ^a series	Cache block size (KB)	4 KB (default)
DS5000 series	Cache block size (KB)	32 KB
DS4000 or DS5000 series	Cache flush control	80/80 (default)
DS4000 or DS5000 series	Readahead	1 (enabled)
DS4000 or DS5000 series	RAID 5	4+p, 8+p, or both
DS4000 or DS5000 series	RAID 6	8+P+Q
DS4000 or DS5000 series	RAID10	4+4 or 8+8

a. For the newest models (on firmware 7.xx and later), use 32 KB.

4.3.7 Logical drive mapping

You must map all logical drives to the single host group that represents the entire SAN Volume Controller cluster or Storwize storage system. You cannot map LUNs to certain nodes or ports in the SAN Volume Controller cluster and exclude other nodes or ports.

Access LUN provides in-band management of a DS4000 or DS5000 series and must be mapped only to hosts that can run the Storage Manager Client and Agent. The SAN Volume Controller or Storwize ignores the Access LUN if the Access LUN is mapped to it. Even so, remove the Access LUN from the SAN Volume Controller/Storwize host group mappings.

Important: Never map the Access LUN as LUN 0.

4.4 Considerations for DS8000 series

Although all recommendations in this chapter are true for SAN Volume Controller and Storwize family storage systems, there is little chance that DS8000 series is virtualized behind one of Storwize storage systems for a long time. Sometimes, it might be virtualized by Storwize only for data migration purposes. Because of this, only the SAN Volume Controller name is used in this chapter. When the DS8000 is configured for SAN Volume Controller, you remember several considerations that are described in this section.

4.4.1 Balancing workload across DS8000 series controllers

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

The DS8000 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.

Example 4-1 shows the correct configuration that 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.

Example 4-1 Output of the lsarray command

```
dscli> lsarray -l
Date/Time: Nov 20, 2013 10:15:43 AM CEST IBM DSCLI Version:5.2.410.299 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: Nov 20, 2013 10:20:23 AM CEST IBM DSCLI Version: 5.2.410.299 DS:
IBM.2107-75L2321
ID Group State datastate Array RAIDtype extpoolID extpoolnam stgtype exts usedexts
=====
```

R0	0	Normal	Normal	A0	5	P0	extpool10	fb	779	779
R1	1	Normal	Normal	A1	5	P1	extpool11	fb	779	779
R2	0	Normal	Normal	A2	5	P2	extpool12	fb	779	779
R3	1	Normal	Normal	A3	5	P3	extpool13	fb	779	779
R4	1	Normal	Normal	A4	5	P5	extpool15	fb	779	779
R5	0	Normal	Normal	A5	5	P4	extpool14	fb	779	779
R6	1	Normal	Normal	A6	5	P7	extpool17	fb	779	779
R7	0	Normal	Normal	A7	5	P6	extpool16	fb	779	779

4.4.2 DS8000 series ranks to extent pools mapping

When you configure the DS8000 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 DS8000 series extent pool striping

The old and most common approach is to map one rank to one extent pool, which provides good control for volume creation. It ensures that all volume allocation from the selected extent pool come from the same rank.

The extent pool striping feature became available with the R3 microcode release for the DS8000 series. With this feature, a single DS8000 series volume can be striped across all the ranks in an extent pool. The function is often referred as *extent pool striping*. 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; no reallocation is possible.

To use the storage pool striping feature, your DS8000 series layout must be well-planned from the initial DS8000 series configuration to using all resources in the DS8000 series. 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.

Because the SAN Volume Controller stripes across MDisk, without proper planning you might end up with a double striping issue: one striping on extent pool level in DS8000 series and another striping on an MDisk in SAN Volume Controller or Storwize. For more information, see *IBM DS8870 Architecture and Implementation*, SG24-8085, which is available at this website:

<http://www.redbooks.ibm.com/abstracts/sg248085.html?Open>

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.

Cache

For the DS8000, you cannot tune the array and cache parameters. The arrays are 6+p or 7+p. This configuration depends on whether the array site contains a spare and whether the *segment size* (contiguous amount of data that is written to a single disk) is 256 KB for fixed block volumes. Caching for the DS8000 series is done on a 64 KB track boundary.

Note: Because of the SAN Volume Controller aggressive cache prefetch algorithms, sometimes it might be beneficiary to turn off SAN Volume Controller cache prefetch and allow the DS8000 series controllers sophisticated cache algorithms do the prefetching. Extraordinary caution must be taken here because cache prefetch is a SAN Volume Controller system-wide parameter. Turning it off means turning off prefetch to all volumes from all backend storage systems. Therefore, this should be done *only* with coordination with IBM support.

4.4.3 Mixing array sizes within a storage pool

Mixing array sizes within a storage pool in general is not of concern. Testing shows no measurable performance differences between selecting all 6+p arrays and all 7+p arrays as opposed to mixing 6+p arrays and 7+p arrays. In fact, mixing array sizes can help balance workload because it places more data on the ranks that have the extra performance capability that is provided by the eighth disk. A small exposure is if an insufficient number of the larger arrays are available to handle access to the higher capacity. To avoid this situation, ensure that the smaller capacity arrays do not represent more than 50% of the total number of arrays within the storage pool.

Best practice: When you mix 6+p arrays and 7+p arrays in the same storage pool, avoid having smaller capacity arrays that comprise more than 50% of the arrays.

If multi-rank extent pools with 6+p and 7+p arrays are used, we recommend turning on EasyTier Auto Rebalance feature. With DS8000 series, EasyTier works even for single tier extent pools; therefore, it manages rebalancing extents between the arrays in extend pool, if needed.

4.4.4 Determining the number of controller ports for the DS8000 series

With the introduction of a Round Robin path selection mechanism in version 6.3 of SAN Volume Controller code, the preferred practice is to configure 16 controller ports from one DS8000. (SAN Volume Controller supports a maximum of 16 ports.) Additionally, use no more than two ports of each of the four-port adapters of the DS8000 series, unless you have DS8700 or later.

The DS8000 series populates FC adapters across 2 - 8 I/O enclosures, depending on the configuration. Each I/O enclosure represents a separate hardware domain.

Ensure that adapters that are configured to different SAN networks do not share I/O enclosure as part of the goal of keeping redundant SAN networks isolated from each other.

Best practices: Consider the following preferred practices:

- ▶ Configure 16 ports per DS8000 series.
- ▶ Configure a maximum of two ports per one DS8000 series adapter, unless you have DS8700 or later.
- ▶ Configure adapters across redundant SANs from different I/O enclosures.

4.4.5 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 DS8000 series controllers perform LUN masking that is based on the volume group. Example 4-2 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 4-2 Output of the showvolgrp command

```
dsccli> showvolgrp V0
Date/Time: November 21, 2013 4:27:20 PM PDT IBM DSCLI Version: 7.6.10.511 DS: IBM.2107-75L3001
Name SVCCF8
ID V0
Type SCSI Mask
Vols 1001 1002 1003 1004 1005 1006 1007 1008 1101 1102 1103 1104 1105 1106 1107 1108
```

Example 4-3 on page 82 shows output for the **lshostconnect** command from the DS8000 series. In this example, you can see that all eight ports of the two-node cluster are assigned to the same volume group (V0) and, therefore, are assigned to the same four LUNs.

Example 4-3 Output for the lshostconnect command

```
dsccli> lshostconnect
Date/Time: November 21, 2013 4:30:15 PM PDT IBM DSCLI Version: 7.6.10.511 DS: IBM.2107-75L3001
Name      ID      WWPN      HostType Profile      portgrp volgrpID ESSIOport
=====
SVCCF8_N1P1 0000 500507680140BC24 - San Volume Controller 0 V0 I0003,I0103
SVCCF8_N1P2 0001 500507680130BC24 - San Volume Controller 0 V0 I0003,I0103
SVCCF8_N1P3 0002 500507680110BC24 - San Volume Controller 0 V0 I0003,I0103
SVCCF8_N1P4 0003 500507680120BC24 - San Volume Controller 0 V0 I0003,I0103
SVCCF8_N2P1 0004 500507680140BB91 - San Volume Controller 0 V0 I0003,I0103
SVCCF8_N2P3 0005 500507680110BB91 - San Volume Controller 0 V0 I0003,I0103
SVCCF8_N2P2 0006 500507680130BB91 - San Volume Controller 0 V0 I0003,I0103
SVCCF8_N2P4 0007 500507680120BB91 - San Volume Controller 0 V0 I0003,I0103
dsccli>
```

From Example 4-3, 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 4-4 shows the general details of the output storage controller by using the SAN Volume Controller command-line interface (CLI).

Example 4-4 Output of the lscontroller command

```
IBM_2145:svccf8:admin>svcinfo lscontroller DS8K75L3001
id 1
controller_name DS8K75L3001
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 75L3001FFFF
allow_quorum yes
WWPN 500507630500C74C
path_count 16
max_path_count 16
WWPN 500507630508C74C
path_count 16
max_path_count 16
IBM_2145:svccf8:admin>
```

Example 4-4 on page 82 also shows that the Managed Disk Link Count is 16 and the storage controller port details. The path_count represents a connection from a single node to a single LUN. Because this configuration has 2 nodes and 16 LUNs, you can expect to see a total of 32 paths, with all paths evenly distributed across the available storage ports. This configuration was validated and is correct because 16 paths are on one WWPN and 16 paths on the other WWPN, for a total of 32 paths.

4.4.6 WWPN to physical port translation

Storage controller WWPNs can be translated to physical ports on the controllers for isolation and debugging purposes. You can also use this information to validate redundancy across hardware boundaries. Example 4-5 shows the WWPN to physical port translations for the DS8000.

Example 4-5 DS8000 WWPN format

WWPN format for DS8000 = 50050763030XXYNNN

XX = adapter location within storage controller

Y = port number within 4-port adapter

NNN = unique identifier for storage controller

I/O Bay	B1				B2				B3				B4			
Slot	S1	S2	S4	S5	S1	S2	S4	S5	S1	S2	S4	S5	S1	S2	S4	S5
XX	00	01	03	04	08	09	0B	0C	10	11	13	14	18	19	1B	1C

I/O Bay	B5				B6				B7				B8			
Slot	S1	S2	S4	S5	S1	S2	S4	S5	S1	S2	S4	S5	S1	S2	S4	S5
XX	20	21	23	24	28	29	2B	2C	30	31	33	34	38	39	3B	3C

Port	P1	P2	P3	P4
Y	0	4	8	C

4.5 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.

4.5.1 Cabling considerations

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

To use the combined capabilities of SAN Volume Controller/Storwize and XIV, connect two ports from every interface module into the fabric for SAN Volume Controller/Storwize use. You must decide which ports you want to use for the connectivity.

Note: With XIV Gen3, you do not have to change the role of the fourth FC port from initiator to target as it was with XIV Gen2. You must make this change only when port 4 is used for native attach.

You must use ports 1 and 3 from every interface module and connect them into the fabric for SAN Volume Controller/Storage use.

Best practice: Use ports 1 and 3 because they belong to two separate HBA cards. Connect ports 1 and 3 to the separate SAN fabrics. Zone all XIV ports with all SAN Volume Controller/Storage ports in one large zone in each SAN fabric.

Figure 4-2 shows a two-node cluster that uses redundant fabrics.

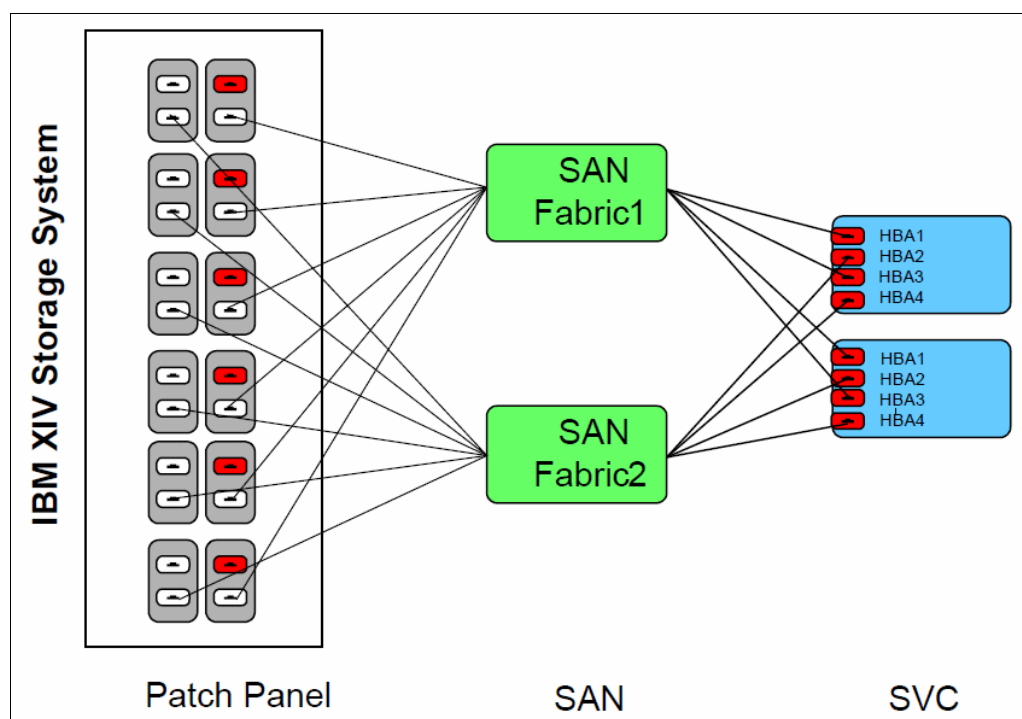


Figure 4-2 Two-node redundant SAN Volume Controller cluster configuration

SAN Volume Controller/Storage supports a maximum of 16 ports from any disk system. The XIV system supports 8 - 24 FC ports, depending on the configuration (6 - 15 modules). Table 4-2 indicates port usage for each XIV system configuration.

Table 4-2 Number of SAN Volume Controller ports and XIV modules

Number of XIV modules	XIV modules with FC ports	Number of FC ports available on XIV	Ports used per card on XIV	Number of SAN Volume Controller ports used
6	Module 4 and 5	8	1	4
9	Module 4, 5, 7 and 8	16	1	8
10	Module 4, 5, 7 and 8	16	1	8
11	Module 4, 5, 7, 8 and 9	20	1	10
12	Module 4, 5, 7, 8 and 9	20	1	10
13	Module 4, 5, 6, 7, 8 and 9	24	1	12
14	Module 4, 5, 6, 7, 8 and 9	24	1	12

Number of XIV modules	XIV modules with FC ports	Number of FC ports available on XIV	Ports used per card on XIV	Number of SAN Volume Controller ports used
15	Module 4, 5, 6, 7, 8 and 9	24	1	12

Port naming convention

The port naming convention for XIV system ports is WWPNN: 5001738NNNNRRMP, where:

- ▶ 001738 is the registered identifier for XIV
- ▶ NNNNN is the serial number in hex
- ▶ RR is the rack ID (01)
- ▶ M is the module ID (4 - 9)
- ▶ P is the port ID (0 - 3)

4.5.2 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 WWPNN. An XIV node is considered to be a single SCSI target. Up to 256 XIV nodes can be presented to each SAN Volume Controller/Storwize port. Each SAN Volume Controller/Storwize host object that is created within the XIV System must be associated with the same LUN map because each LUN can be assigned to a single map only.

From a SAN Volume Controller/Storwize perspective, an XIV Type Number 2810 controller can consist of more than one WWPNN; 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 2810

A single host instance can be created for use in defining and then implementing the SAN Volume Controller/Storwize. 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.

Complete the following steps to create the SAN Volume Controller/Storwize host definition:

1. Select **Add Cluster**.
2. Enter a name for the SAN Volume Controller/Storwize host definition.
3. Select **Add Host**.
4. Enter a name for the first node instance. Then, click the **Cluster** drop-down menu and select the SAN Volume Controller/Storwize cluster that you created.
5. Repeat steps 1 - 4 for each instance of a node in the cluster.
6. Right-click a node instance and select **Add Port**. Figure 4-3 on page 86 shows that four ports per node were added to ensure that the host definition is accurate.

Name	Type	Cluster	Access
SVC	default		
SVC_node01	default	SVC_iogrp01	
500507680110A730	FC		
500507680120A730	FC		
500507680130A730	FC		
500507680140A730	FC		
SVC_node02	default	SVC_iogrp01	
500507680110A78A	FC		
500507680120A78A	FC		
500507680130A78A	FC		
500507680140A78A	FC		

Figure 4-3 SAN Volume Controller/Storwize host definition on IBM XIV Storage System

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.

4.5.3 Number and size of the Volumes

Because XIV system has no raid arrays or ranks, the usual approach of one RAID = one LUN does not work here. To calculate the number and size of the volumes that are mapped to SAN Volume Controller/Storwize, use following formula:

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

where:

- ▶ M is the number of volumes that were created on XIV and presented to SAN Volume Controller/Storwize.
- ▶ P is the number of XIV host ports that are zoned to SAN Volume Controller/Storwize.
- ▶ C is the maximum queue depth SAN Volume Controller/Storwize uses for each XIV port. This depth is set internally in SAN Volume Controller/Storwize code to 1000.
- ▶ N is the number of nodes in SAN Volume Controller/Storwize cluster.
- ▶ Q is the maximum queue depth for each MDisk. This depth is set internally in SAN Volume Controller/Storwize code to 60.

For example, for a two-node SAN Volume Controller cluster or one Storwize control enclosure and fully populated XIV system with 12 host ports that are zoned to SAN Volume Controller/Storwize, we see the following formula:

$$M = ((12 \times 1000) / 2 / 60 = 100$$

Now, depending on XIV disk drives size, we can estimate the volume size. For a fully populated XIV with 2 TB volumes, the total usable XIV capacity is approximately 160 TB. Dividing this number by the number of volumes (100), we get the volume size of 1600 GB.

For more information about XIV and SAN Volume Controller/Storwize sizing and configuration, see *IBM XIV Gen3 with IBM System Storage SAN Volume Controller and Storwize V7000*, REDP-5063, which is available at this website:

<http://www.redbooks.ibm.com/redpieces/abstracts/redp5063.html?open>

4.5.4 Restrictions

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

Clearing SCSI reservations and registrations

Do *not* use the `vol_clear_keys` command to clear SCSI reservations and registrations on volumes that are managed by SAN Volume Controller/Storwize.

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. Thin provisioning is not supported for use with SAN Volume Controller/Storwize.

4.6 Considerations for IBM Storwize V7000/V5000/V3700

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

4.6.1 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. There is no need 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 backend Storwize system likely does not use this function. All remote copy services functions should be used from the front-end SAN Volume Controller/Storwize system.

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.

On the SAN switches, create one large zone per fabric, with all backend Storwize and SAN Volume Controller or front-end Storwize FC ports in it.

4.6.2 Defining internal storage

When you plan to attach a Storwize V7000/V5000/V3700 to the SAN Volume Controller or another Storwize V7000/V5000 system, create the arrays (MDisks) manually (by using the CLI), instead of using the Storwize default settings. For RAID10, select half disk drives from enclosure in first chain and second half of disk drives from second chain (for better availability). Although drive selection is not a concern for RAID5 or RAID6, ensure that the selected drives are in the 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.

When you define Storwize internal storage, create a one-to-one relationship. That is, create one storage pool to one MDisk (array) to one volume. Then, map the volume to the SAN Volume Controller/Storwize host.

The Storwize systems can have a mixed disk drive type, such as solid-state drives (SSDs), serial-attached SCSI (SAS), and nearline SAS. 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 same characteristic.

Note: One exception is when you want to use Easy Tier on SAN Volume Controller. Then, create SSD storage pool and volume on Storwize, map it to SAN Volume Controller, and add it to pool with SAS MDisks.

For example, assume that you have two Storwize arrays. One array (model A) is configured as a RAID 5 that uses 300 GB SAS drives. The other array (model B) is configured as a RAID 5 that uses 2 TB Nearline SAS drives. When you map to the SAN Volume Controller, assign model A to one specific storage pool (model A), and assign model B to another specific storage pool (model B).

Important: The extent size value for SAN Volume Controller should be 1 GB. The extent size value for the V7000 should be 256 MB. These settings stop potential negation of stripe on stripe. For more information, see the blog post “Configuring IBM Storwize V7000 and SAN Volume Controller for Optimal Performance” at this website:

https://www.ibm.com/developerworks/mydeveloperworks/blogs/storagevirtualization/entry/configuring_ibm_storwize_v7000_and_svc_for_optimal_performance_part_121?lang=en

4.6.3 Configuring Storwize storage systems

Storwize external storage systems can present volumes to a SAN Volume Controller. Additionally, Storwize can present volumes to another Storwize. However, to do this, you must change the **layer** of the Storwize system.

Note: By default, SAN Volume Controller is in the replication layer and Storwize is in the storage layer. This means SAN Volume Controller can virtualize (but cannot create) a partnership with Storwize. The layer of SAN Volume Controller cannot be changed.

If you change the Storwize layer to replication, it can virtualize another Storwize and can create a partnership with SAN Volume Controller or other Storwize in replication layer. To change a Storwize layer, make sure that it cannot see any other Storwize or SAN Volume Controller in the SAN fabric, which means you must remove all remote copy relationships and all zoning first.

Complete the following steps to configure the Storwize system:

1. On the backend Storwize system, define a host object, and then add all WWPNs from the SAN Volume Controller or front-end Storwize to it.
2. On the backend Storwize system, create host mappings between each volume and the SAN Volume Controller or front-end Storwize host object that you created.

The volumes that are presented by the backend Storwize system are displayed in the SAN Volume Controller or front-end Storwize MDisk view. The back-end Storwize system is displayed in the SAN Volume Controller or front-end Storwize view with a vendor ID of IBM and a product ID of 2145.

4.7 Considerations for IBM FlashSystem

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

4.7.1 Physical FC port connection and zoning

Each model of IBM FlashSystem family has two, 2-port, 8 Gb FC HBA cards and these ports should be connected to the SAN network in the following manner:

- ▶ Connect FlashSystem first ports from both HBAs to first SAN fabric and second ports from both HBAs to second SAN fabric.
- ▶ Zone each FlashSystem port to a pair of ports from both SAN Volume Controller/Storwize nodes. If you have more than one IO group, repeat this set up for all nodes in all IO groups.
- ▶ If you have SAN Volume Controller with more HBA cards, dedicate half of the ports from each SAN Volume Controller card only to the FlashSystem.

The recommended connections for SAN Volume Controller nodes with one HBA are shown in Figure 4-4 on page 90, where blue connections are in the first fabric and the red connections are in the second fabric.

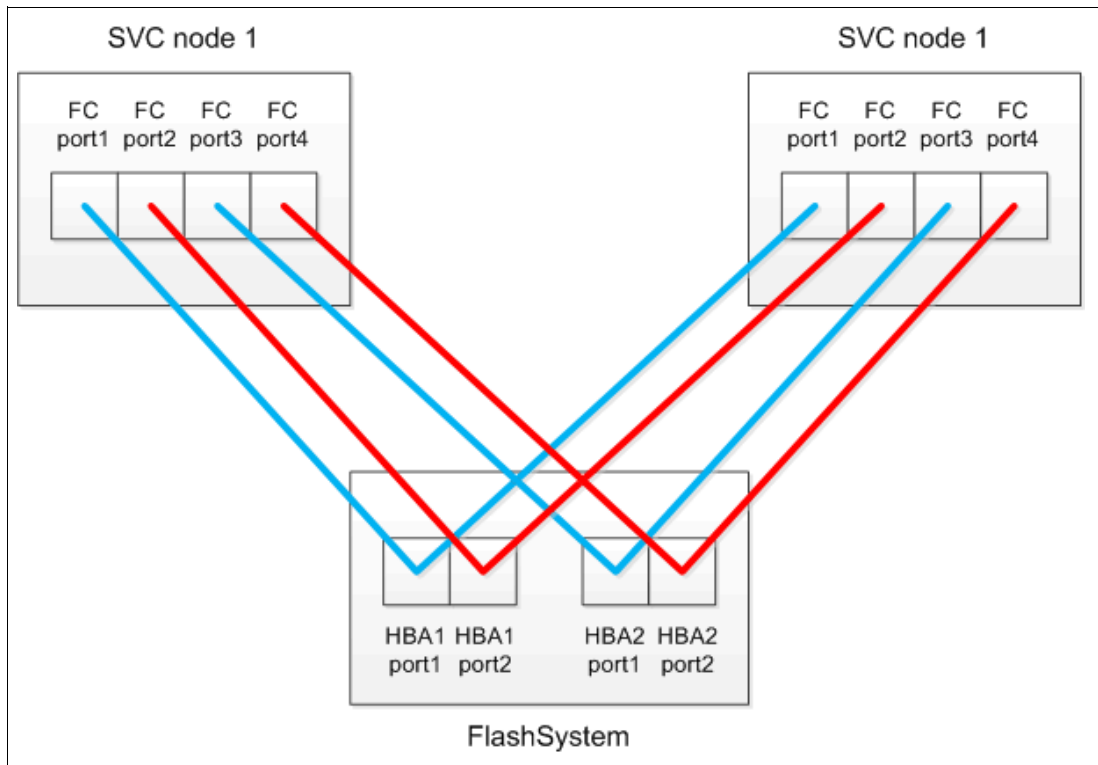


Figure 4-4 FlashSystem to SAN Volume Controller zoning

In the case of two HBAs in SAN Volume Controller nodes, the recommended connections are shown in Figure 4-5, where blue connections are in the first fabric and the red connections are in the second fabric.

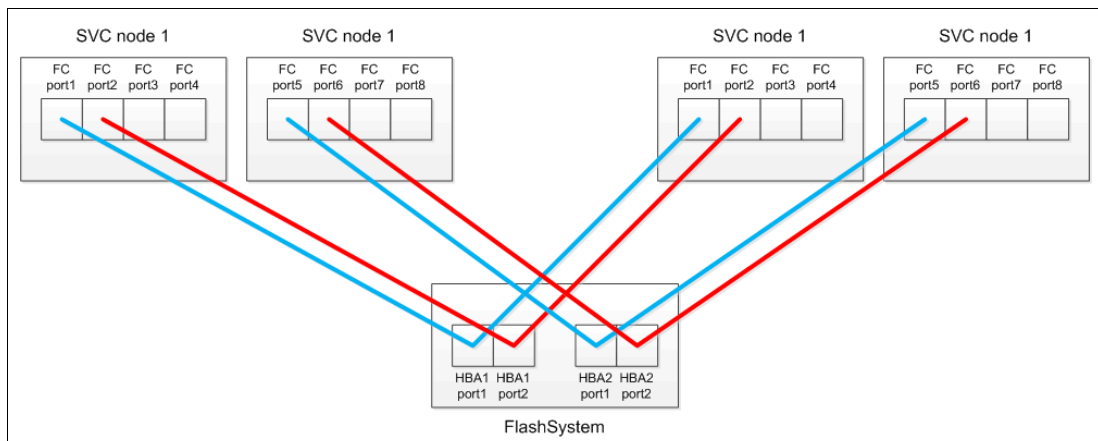


Figure 4-5 FlashSystem to SAN Volume Controller zoning with two HBA cards

Best practice: For FlashSystems, use SAN Volume Controller with two HBA cards and dedicate two ports of each card to FlashSystem. You cannot create storage zones with ports 7 and 8; therefore, use ports 3 and 4 to connect other storage systems. Create host zones with ports 3, 4, 7, or 8.

4.7.2 Logical configuration

As with XIV, there is no RAID configuration on the IBM FlashSystems; therefore, no one array = one LUN preferred practice. Instead, you must create volumes from the available capacity, which depends on the FlashSystem model. As a preferred practice, create 4 - 16 volumes of the same size with the sector value of 512B.

After the volume is created, you must create an access group policy to map volumes to SAN Volume Controller/Storwize ports. Map all the volumes to all four FlashSystem ports and all zoned SAN Volume Controller/Storwize ports.

4.7.3 Extent size and storage pools

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. If you do not have any other storage pools, you can leave the default extent size, which in version 7.x of SAN Volume Controller/ Storwize code equals 1 GB.

Always create only one storage pool per backend storage system and place all volumes from this system into its pool. The only exception is Easy Tier. If IBM FlashSystem is to be used with Easy Tier, you must add all FlashSystem MDisk to the existing storage pool with other MDisk characteristics (preferably SAS).

Note: To have SAN Volume Controller/Storwize to properly recognize FlashSystem MDisk as flash disks, remember to change the MDisk type from `generic_hdd` to `generic_ssd`.

4.7.4 Volumes

To fully use all SAN Volume Controller/Storwize resources, at least eight volumes should be created 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 the FlashSystem 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 a cache) to the FlashSystem can take 100 - 200 microseconds. If you use volumes from the pure FlashSystem storage pool, it is better to create cache-disabled volumes.

Best Practice: On SAN Volume Controller/Storwize, disable the cache for basic volumes that are created in storage pools that are composed of only FlashSystem MDisk. You can disable the cache when the volume is created or later. This does not apply to the volumes in multitier storage pools or volumes utilizing advanced SVC function. Always enable cache for compressed vdisks. For most other advanced SVC function (FlashCopy, Mirroring, Space Efficient) best latency may be achieved with cache enabled.

If you have a mirrored volume and one copy comes from FlashSystem MDisk but the other copy comes from spinning-type MDisk, it is better to have *primary* copy set to the FlashSystem MDisk. 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. 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 of FlashSystem MDisk and change the **mirrorwritepriority** setting to **latency**.

For more information, see the following resources:

- ▶ *IBM SAN Volume Controller and IBM FlashSystem 820: Best Practices and Performance Capabilities*, REDP-5027, which is available at this website:
<http://www.redbooks.ibm.com/abstracts/redp5027.html?Open>
- ▶ *Implementing the IBM SAN Volume Controller and FlashSystem 820*, SG24-8172, which is available at this website:
<http://www.redbooks.ibm.com/abstracts/sg248172.html?Open>

4.8 Considerations for third-party storage with EMC Symmetrix DMX and Hitachi Data Systems

Although many third-party storage options are available (supported), this section highlights the pathing considerations for EMC Symmetrix/DMX and Hitachi Data Systems (HDS). For EMC Symmetrix/DMX and HDS, some storage controller types present a unique WWNN and WWPN for each port. This action can cause problems when it is attached to the SAN Volume Controller/Storwize because the SAN Volume Controller/Storwize enforces a WWNN maximum of four per storage controller. Because of this behavior, you must group the ports if you want to connect more than four target ports to a SAN Volume Controller/Storwize.

For information about specific models, see *IBM System Storage SAN Volume Controller 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.641.doc/mlt_relatedinfo_224agr.html

4.9 Medium error logging

Medium errors on back-end MDiskS can be encountered by host I/O and by SAN Volume Controller background functions, such as volume migration and FlashCopy. If a SAN Volume Controller receives a medium error from a storage controller, it attempts to identify which logical block addresses (LBAs) are affected by this MDisk problem. It also records those LBAs as having virtual medium errors.

If a medium error is encountered on a read from the source during a migration operation, the medium error is logically moved to the equivalent position on the destination. This action is achieved by maintaining a set of bad blocks for each MDisk. Any read operation that touches a bad block fails with a medium error SCSI. If a destage from the cache touches a location in the medium error table and the resulting write to the managed disk is successful, the bad block is deleted.

On the Storwize MDiskS, medium error is handled automatically because Storwize knows its disk drives. It scans all disks all the time and if finds some errors it can recover data from parity (for RAID5, RAID6) or from copy on another drive in pair (for RAID10).

For more information about how to troubleshoot a medium error, see Chapter 15, “Troubleshooting and diagnostics” on page 519.

4.10 Mapping physical LBAs to volume extents

It is possible to find the volume extent that a physical MDisk LBA maps to and to find the physical MDisk LBA to which the volume extent maps. This function can be useful in several situations, as shown in the following examples:

- ▶ If a storage controller reports a medium error on a logical drive, but SAN Volume Controller did not yet take the MDiskS offline, you might want to establish which volumes are affected by the medium error.
- ▶ When you investigate application interaction with thin-provisioned volumes, determine whether a volume LBA is allocated. If an LBA is allocated when it was not intentionally written to, the application might not be designed to work well with thin volumes.

The output of the `svcinfo lsmdisklba` and `svcinfo lsvdisklba` commands varies depending on the type of volume (such as thin-provisioned versus fully allocated) and type of MDisk (such as quorum versus non-quorum). For more information, see *IBM System Storage SAN Volume Controller 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.641.doc/mlt_relatedinfo_224agr.html

4.11 Identifying storage controller boundaries by using the IBM Tivoli Storage Productivity Center

You might often want to map the virtualization layer to determine which volumes and hosts are using resources for a specific hardware boundary on the storage controller. An example is when a specific hardware component, such as a disk drive, is failing and the administrator is interested in performing an application-level risk assessment.

Information that is learned from this type of analysis can lead to actions that are taken to mitigate risks, such as scheduling application downtime, performing volume migrations, and initiating FlashCopy. By using IBM Tivoli Storage Productivity Center, mapping the virtualization layer can be done quickly. Also, Tivoli Storage Productivity Center can help to eliminate mistakes that can be made by using a manual approach.

Figure 4-6 shows how a failing disk on a storage controller can be mapped to the MDisk that is being used by an SAN Volume Controller cluster. To display this panel, click **Physical Disk** **RAID5 Array** **Logical Volume** **MDisk**.

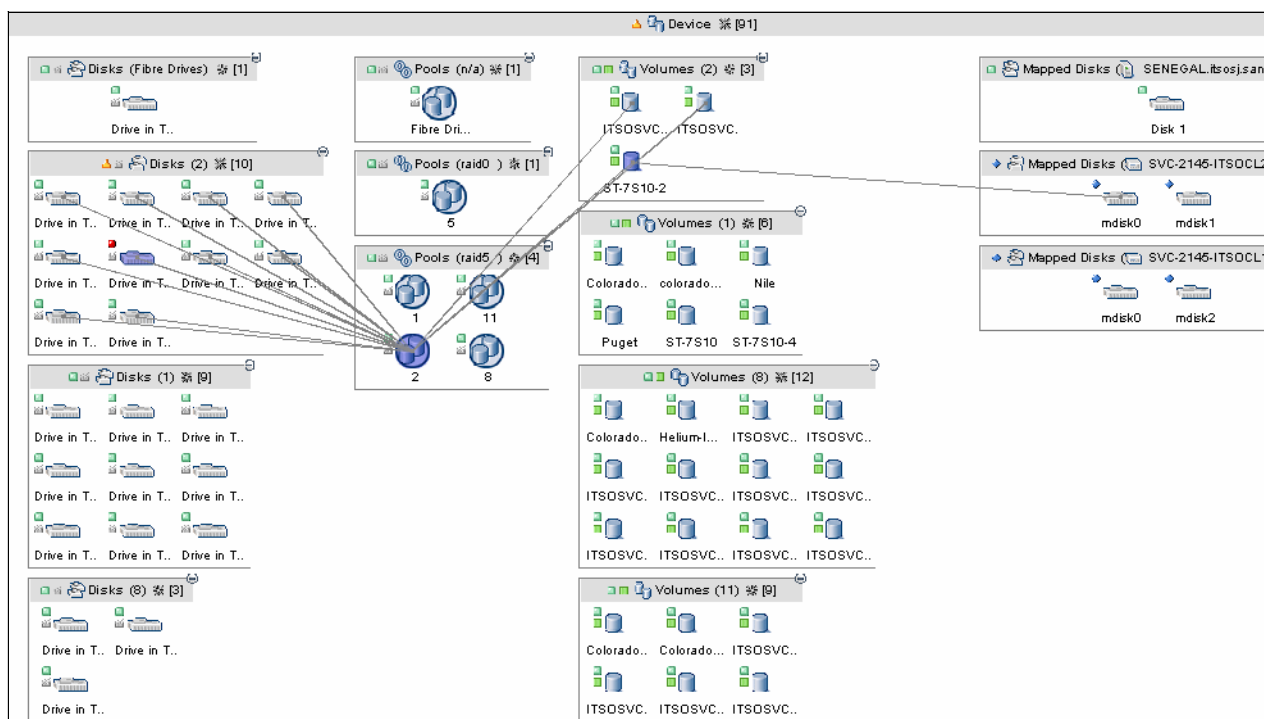


Figure 4-6 Mapping MDisk

Figure 4-7 completes the end-to-end view by mapping the MDisk through the SAN Volume Controller to the attached host. Click **MDisk** **MDGroup** **VDisk** **Host disk**.

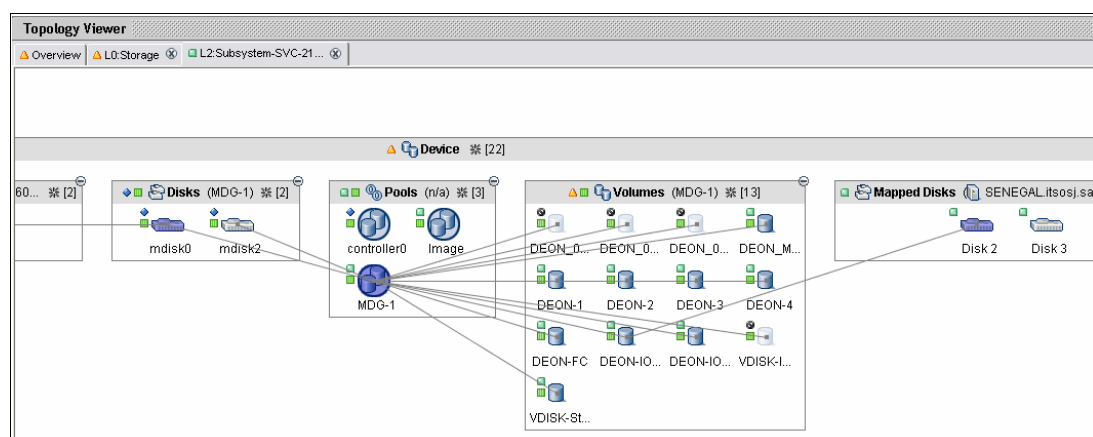


Figure 4-7 Host mapping



Storage pools and managed disks

This chapter highlights considerations when you are planning storage pools for an IBM 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
- ▶ 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
- ▶ Moving an MDisk between SAN Volume Controller clusters or Storwize systems
- ▶ MDisk group (Storage pool) considerations when using Real-time Compression

5.1 Availability considerations for storage pools

Although the SAN Volume Controller provides many advantages through the consolidation of storage, you must understand the availability implications that storage subsystem failures can have on availability domains within the SAN Volume Controller cluster. The SAN Volume Controller 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 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 the SAN Volume Controller 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 5.3.4, “Considerations for the IBM XIV Storage System” on page 100.

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

- ▶ Each storage subsystem must be used with only a single SAN Volume Controller cluster.
- ▶ Each storage pool must contain only MDisk from a single storage subsystem. An exception exists when you are working with IBM System Storage Easy Tier. For more information, see Chapter 11, “IBM System Storage Easy Tier function” on page 319.
- ▶ Each storage pool must contain MDisk from no more than approximately 10 storage subsystem arrays.

5.2 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 SAN Volume Controller 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 SAN Volume Controller. 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 RAID type that is selected.

Performance is also a determining factor, where adding a SAN Volume Controller as a front-end results in considerable gains. Another factor is the ability of your storage subsystems to be scaled up or scaled out. For example, IBM System Storage DS8000 series is a scale-up architecture that delivers “best of breed” performance per unit, and the IBM System Storage DS5000 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 SAN Volume Controller 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.

5.3 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.

Start by using one storage pool per application volume. Then, split the volumes across other storage pools if you observe that this specific storage pool is saturated.

Cluster capacity: For most clusters, a 1 - 2 PB capacity is sufficient. In general, use 256 MB. For larger clusters, use 512 MB as the standard extent size. Alternatively, when you are working with the XIV system, use an extent size of 1 GB.

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.

5.3.1 Selecting the number of arrays per storage pool

The capability to stripe across disk arrays is the most important performance advantage of the SAN Volume Controller. 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 it is often difficult to determine what is required in terms of performance, the tendency is to add too many arrays to a single storage pool, which increases the failure domain as described in 5.1, “Availability considerations for storage pools” on page 96.

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.

More arrays in the storage pool have more of an effect with lower performing storage controllers. For example, fewer arrays are required from a DS8000 than from a DS4000 to achieve the same performance objectives. Table 5-1 shows the number of arrays per storage pool that is appropriate for general cases. When it comes to performance, exceptions can exist. For more information, see Chapter 10, “Back-end storage performance considerations” on page 269.

Table 5-1 Number of arrays per storage pool

Controller type	Arrays per storage pool
IBM DS4000 series or DS5000 series	4 - 60
IBM DS6000™ series or DS8000 series	4 - 60
IBM DS3500 series	4 - 60
IBM Storwize V7000	4 - 60
IBM XIV Storage Systems	4 - 60
IBM FlashSystems	8 - 60

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 high cost in usable capacity, and in most cases, RAID 5 is the best overall choice.

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 makes the most sense. If RAID 10 shows a service time advantage over RAID 5, the importance of that advantage must be weighed against its additional cost.

5.3.2 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.

Table 5-2 provides guidelines for array provisioning on IBM storage subsystems.

Table 5-2 Array provisioning

Controller type	LUNs per array
IBM DS4000 series or DS5000 series	1
IBM DS6000 series or DS8000 series	1 - 2
IBM DS3500 series	1
IBM Storwize V7000	1
IBM XIV Storage Systems	1
IBM FlashSystems	1

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 (known to the SAN Volume Controller as *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.

Remember that the SAN Volume Controller 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 the introduction of volume mirroring in SAN Volume Controller V4.3, you can protect against the loss of a storage pool by mirroring a volume across multiple storage pools. For more information, see Chapter 6, "Volumes" on page 125.

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

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

5.3.3 Considerations for Storwize family systems

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

- ▶ Storwize as back-end storage system to SAN Volume Controller or another Storwize
- ▶ Storwize as front-end storage systems to hosts

In a case where Storwize is a back-end controller for SAN Volume Controller or another Storwize, see 4.6, “Considerations for IBM Storwize V7000/V5000/V3700” on page 87.

If you have Storwize as 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 a 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.

5.3.4 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: Although SAN Volume Controller V6.2 and higher supports MDisks up to 1 PB, at the time of the writing of this book in 2013, there was no support available for MDisks that are larger than 2 TB on the XIV system. However, further enhancements were made in the subsequent releases and starting with SAN Volume Controller V7.1 and higher, support for MDisks larger than 2 TB on the XIV system was added.

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

Because the XIV configuration grows 6 - 15 modules, use the SAN Volume Controller rebalancing script to restripe volume extents to include new MDisks. For more information, see 5.7, “Rebalancing extents across a storage pool” on page 107.

For a fully populated rack with 12 ports and 2 TB drives, create 48 volumes of 1632 GB each.

Tip: Always use the largest volumes possible.

Table 5-3 shows the number of 1632 GB LUNs that are created, depending on the XIV capacity that is populated with 2 TB drives.

Table 5-3 Values that use the 1632 GB LUNs

Number of LUNs (MDisks) at 1632 GB each	XIV system TB used	XIV system TB available capacity
16	26.1	27
26	42.4	43
30	48.9	50
33	53.9	54
37	60.4	61
40	65.3	66
44	71.8	73
48	78.3	79

The best use of the SAN Volume Controller virtualization solution with the XIV Storage System 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 SAN Volume Controller, each physical XIV system should have a separate storage pool. This design provides a good queue depth on the SAN Volume Controller 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 SAN Volume Controller/Storwize, see the following resources:

- ▶ *XIV Gen3 with SAN Volume Controller and Storwize V7000*, REDP-5063, which is available at this website:
<http://www.redbooks.ibm.com/redpieces/abstracts/redp5063.html?Open>
- ▶ “Can you use SAN Volume Controller with XIV as storage?”, which is available at this website:
https://www.ibm.com/developerworks/community/blogs/storage_redbooks/entry/can_you_use_svc_with_xiv_as_storage?lang=en_us

5.4 Quorum disk considerations

When back-end storage is initially added to an SAN Volume Controller cluster as a storage pool, three quorum disks are automatically created by allocating space from the assigned MDisk. 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 SAN Volume Controller 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 MDisk. You can change placement of all three quorums to external MDisk or you can have some quorums on internal drives and some on the external MDisk. 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 MDisk 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.

Important: Do not assign internal SAN Volume Controller solid-state drives (SSD) as a quorum disk.

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 SAN Volume Controller GUI or from the SAN Volume Controller command-line interface (CLI).

To list SAN Volume Controller cluster quorum MDisk and to view their number and status, run the **svcinfo lsquorum** command, as shown in Example 5-1.

Example 5-1 The lsquorum command

```
IBM_2145:ITS0-CLS4:admin>svcinfo lsquorum
```

quorum_index	status	id	name	controller_id	controller_name	active	object_type
0	online	0	mdisk0	0	ITS0-4700	yes	mdisk
1	online	1	mdisk1	0	ITS0-4700	no	mdisk
2	online	2	mdisk2	0	ITS0-4700	no	mdisk

To move one SAN Volume Controller quorum MDisk from one MDisk to another or from one storage subsystem to another, use the **svctask chquorum** command, as shown in Example 5-2 on page 103.

Example 5-2 The chquorum command

```
IBM_2145:ITSO-CLS4:admin>svctask chquorum -mdisk 9 2
```

```
IBM_2145:ITSO-CLS4:admin>svcinfolsqquorum
```

quorum_index	status	id	name	controller_id	controller_name	active	object_type
0	online	0	mdisk0	0	ITSO-4700	yes	mdisk
1	online	1	mdisk1	0	ITSO-4700	no	mdisk
2	online	2	mdisk9	1	ITSO-XIV	no	mdisk

As you can see in Example 5-2, quorum index 2 moved from mdisk2 on ITSO-4700 controller to mdisk9 on ITSO-XIV controller.

Tip: Although the **setquorum** command (deprecated) still works, use the **chquorum** command to change the quorum association.

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 SAN Volume Controller 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 an SAN Volume Controller 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 5-4 provides the number of extents that are reserved for quorum use by extent size.

Table 5-4 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

5.5 Tiered storage

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

In a *single-tiered storage pool*, the MDisks must have the following characteristics 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).
- ▶ The disk subsystems that provide the MDisks must have similar characteristics; for example, maximum I/O operations per second (IOPS), response time, cache, and throughput.

- The MDisk that are used are of the same size and are, therefore, MDisk that provide the same number of extents. If this requirement is not feasible, you must check the distribution of the extents of the volumes in that storage pool.

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 `generic_hdd` and `generic_ssd` MDisk.

A multitiered storage pool, therefore, contains MDisk with various characteristics, as opposed to a single-tier storage pool. However, each tier must have MDisk of the same size and MDisk that provide the same number of extents. Multi-tiered storage pools are used to enable the automatic migration of extents between disk tiers by using the SAN Volume Controller Easy Tier function. For more information about IBM System Storage Easy Tier, see Chapter 11, “IBM System Storage Easy Tier function” on page 319.

It is likely that the MDisk (LUNs) that are presented to the SAN Volume Controller cluster have various performance attributes because of the type of disk or RAID array on which they are installed. The MDisk can be on a 15 K RPM Fibre Channel (FC) or serial-attached SCSI (SAS) disk, a nearline SAS, Serial Advanced Technology Attachment (SATA), or SSDs.

Therefore, a storage tier attribute is assigned to each MDisk, with the default of `generic_hdd`. With SAN Volume Controller V6.2, a new tier 0 level disk attribute is available for SSDs, and it is known as *generic_ssd*.

There are two types of storage tier: `generic_ssd` or `generic_hdd`. For Storwize, when you create an array with SSD drives, the MDisk becomes `generic_ssd` by default. If you create an array with normal HDD drives (SAS or NL-SAS), the MDisk becomes `generic_hdd` by default.

If you present an external MDisk to SAN Volume Controller or Storwize, it becomes `generic_hdd` by default, even if that external MDisk was built by using SSD drives or a flash memory system; for example, where it is presented from IBM FlashSystem storage system. You can change the MDisk tier only for MDisk that are presented from external storage systems.

You can also define storage tiers by using storage controllers of varying performance and availability levels. Then, you can easily provision them based on host, application, and user requirements.

Remember that a single storage tier can be represented by multiple storage pools. For example, if you have a large pool of tier 3 storage that is provided by many low-cost storage controllers, it is sensible to use several storage pools. The use of several storage pools prevents a single offline volume from taking all of the tier 3 storage offline.

When multiple storage tiers 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.

5.6 Adding MDisks to existing storage pools

Before you add MDisks to existing storage pools, first ask why you are adding these MDisks. If MDisks are being added to the SAN Volume Controller cluster to provide more capacity, consider adding them to a new storage pool. Adding MDisks to storage pools reduces the reliability characteristics of the storage pool and risk, and destabilizes the storage pool if hardware problems exist with the new LUNs. If the storage pool is meeting its performance objectives, in most cases, add the MDisks to new storage pools rather than adding the MDisks to existing storage pools.

Important: Do not add an MDisk to a storage pool if you want to create an image mode volume from the MDisk that you are adding. If you add an MDisk to a storage pool, the MDisk becomes managed, and extent mapping is not necessarily one-to-one anymore. This means you cannot create an image mode volume any more and all data on this MDisk is lost.

5.6.1 Checking access to new MDisks

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

In SAN Volume Controller V4.2.1, 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 SAN Volume Controller cluster suffers a software error during the MDisk test.

Image-mode MDisks are not tested before they are added to a storage pool because an offline image-mode MDisk does not take the storage pool offline.

5.6.2 Persistent reserve

A common condition where MDisks can be configured by SAN Volume Controller (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. You do not see this condition on the DS4000 system when they were attached by using Redundant Disk Array Controller (RDAC) because RDAC does not implement persistent reserve.

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.

5.6.3 Renaming MDiskS

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

When multiple tiers of storage are on the same SAN Volume Controller 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 MDiskS that associates the MDisk with its corresponding controller and array within the controller; for example, DS8K_R5_12345.

5.7 Rebalancing extents across a storage pool

Adding MDiskS to existing storage pools can result in reduced performance across the storage pool because of the extent imbalance that occurs and the potential to create hot spots within the storage pool. After you add MDiskS to storage pools, rebalance extents across all available MDiskS by using the CLI by manual command entry. Alternatively, you can automate rebalancing the extents across all available MDiskS by using a Perl script, which is available as part of the SAN Volume ControllerTools package. The package is available from the IBM alphaWorks® website.

There was a new release of SAN Volume ControllerTools package that was released in April 2013 with support of Storwize products and Easy Tier and is available for download at this website:

<https://www.ibm.com/developerworks/community/groups/service/html/communityview?communityUuid=18d10b14-e2c8-4780-bace-9af1fc463cc0>

If you want to manually balance extents, you can use the following CLI commands to identify and correct extent imbalance across storage pools. Remember that the **svcinfo** and **svctask** prefixes are no longer required:

- ▶ **lsmdiskextent**
- ▶ **migrateexts**
- ▶ **lsmigrate**

The following section describes how to use the script from the SAN Volume ControllerTools package to rebalance extents automatically. You can use this script on any host with Perl and an SSH client installed. The next section also describes how to install it on a Windows Server 2003 server.

5.7.1 Installing prerequisites and the SAN Volume ControllerTools package

For this test, SAN Volume ControllerTools is installed on a Windows Server 2003 server. The installation has the following major prerequisites:

- ▶ PuTTY

This tool provides SSH access to the SAN Volume Controller cluster. If you are using a SAN Volume Controller Master Console or an SSPC server, PuTTY is already installed. If not, you can download PuTTY from this website:

<http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html>

The easiest package to install is the “Windows installer,” which installs all the PuTTY tools in one location.

- Perl

Perl packages for Windows are available from several sources. For this Redbooks publication, ActivePerl was used, which you can download at no charge from this website:

<http://www.activestate.com/Products/activeperl/index.mhtml>

The latest SAN Volume ControllerTools package is available from this alphaWorks website:

<http://www.alphaworks.ibm.com/tech/svctools>

The SAN Volume ControllerTools package is a compressed file that you can extract to a convenient location. For example, for this book, the file was extracted to C:\SVCTools on the Master Console. The extent balancing script requires the following key files:

- ▶ The `SVCToolsSetup.doc` file, which explains the installation and use of the script in detail.
- ▶ The `lib\IBM\SVC.pm` file, which must be copied to the Perl lib directory.

With ActivePerl installed in the C:\Perl directory, copy it to C:\Perl\lib\IBM\SVC.pm.

- The `examples\balance\balance.pl` file, which is the rebalancing script.

5.7.2 Running the extent balancing script

The storage pool on which the script was tested was unbalanced because it was recently expanded from four MDisks to eight MDisks. Example 5-3 shows that all of the volume extents are on the original four MDisks.

Example 5-3 The lsmdiskextent script output that shows an unbalanced storage pool

```

IBM_2145:itso$vscc11:admin>lsmdisk -filtervalue "mdisk_grp_name=itso_ds45_18gb"
id name      status mode      mdisk_grp_id mdisk_grp_name capacity ctrl_LUN_#      controller_name UID
0 mdisk0 online managed      1          itso_ds45_18gb 18.0GB 0000000000000000 itso_ds4500
600a0b80001744310000011a4888478c0000000000000000000000000000000000000000
1 mdisk1 online managed      1          itso_ds45_18gb 18.0GB 00000000000000001 itso_ds4500
600a0b800017443100000119488847780000000000000000000000000000000000000000
2 mdisk2 online managed      1          itso_ds45_18gb 18.0GB 00000000000000002 itso_ds4500
600a0b800017443100000118488847580000000000000000000000000000000000000000
3 mdisk3 online managed      1          itso_ds45_18gb 18.0GB 00000000000000003 itso_ds4500
600a0b8000174431000001174888473e0000000000000000000000000000000000000000
4 mdisk4 online managed      1          itso_ds45_18gb 18.0GB 00000000000000004 itso_ds4500
600a0b800017443100000116488847260000000000000000000000000000000000000000
5 mdisk5 online managed      1          itso_ds45_18gb 18.0GB 00000000000000005 itso_ds4500
600a0b8000174431000001154888470c0000000000000000000000000000000000000000
6 mdisk6 online managed      1          itso_ds45_18gb 18.0GB 00000000000000006 itso_ds4500
600a0b800017443100000114488846ec00000000000000000000000000000000000000000
7 mdisk7 online managed      1          itso_ds45_18gb 18.0GB 00000000000000007 itso_ds4500
600a0b800017443100000113488846c00000000000000000000000000000000000000000

```

```

IBM_2145:itsosvcc11:admin>svcinfolsmdiskextent mdisk0
id          number_of_extents copy_id
0           64                0
2           64                0
1           64                0
4           64                0
IBM_2145:itsosvcc11:admin>svcinfolsmdiskextent mdisk1
id          number_of_extents copy_id
0           64                0
2           64                0
1           64                0
4           64                0
IBM_2145:itsosvcc11:admin>svcinfolsmdiskextent mdisk2
id          number_of_extents copy_id
0           64                0
2           64                0
1           64                0
4           64                0
IBM_2145:itsosvcc11:admin>svcinfolsmdiskextent mdisk3
id          number_of_extents copy_id
0           64                0
2           64                0
1           64                0
4           64                0
IBM_2145:itsosvcc11:admin>svcinfolsmdiskextent mdisk4
IBM_2145:itsosvcc11:admin>svcinfolsmdiskextent mdisk5
IBM_2145:itsosvcc11:admin>svcinfolsmdiskextent mdisk6
IBM_2145:itsosvcc11:admin>svcinfolsmdiskextent mdisk7

```

The **balance.pl** script is then run on the Master Console by using the following command:

```
C:\SVCTools\examples\balance>perl balance.pl itso_ds45_18gb -k "c:\icat.ppk" -i 9.43.86.117 -r -e
```

where:

► **itso_ds45_18gb**

Indicates the storage pool to be rebalanced.

► **-k "c:\icat.ppk"**

Gives the location of the PuTTY private key file, which is authorized for administrator access to the SAN Volume Controller cluster.

► **-i 9.43.86.117**

Gives the IP address of the cluster.

► **-r**

Requires that the optimal solution is found. If this option is not specified, the extents can still be unevenly spread at completion. However, not specifying **-r** often requires fewer migration commands and less time. If time is important, you might not want to use **-r** at first, but then rerun the command with **-r** if the solution is not good enough.

► **-e**

Specifies that the script runs the extent migration commands. Without this option, it merely prints the commands that it might run. You can use this option to check that the series of steps is logical before you commit to migration.

In this example, with 4 x 8 GB volumes, the migration completed within around 15 minutes. You can use the **svcinfolsmigrate** command to monitor progress. This command shows a percentage for each extent migration command that is issued by the script.

After the script completed, check that the extents are correctly rebalanced. Example 5-4 shows that the extents were correctly rebalanced in the example for this book. In a test run of 40 minutes of I/O (25% random, 70/30 read/write) to the four volumes, performance for the balanced storage pool was around 20% better than for the unbalanced storage pool.

Example 5-4 Output of the lsmdiskextent command that shows a balanced storage pool

```

IBM_2145:itsosvcc11:admin>svcinfolsmdiskextentmdisk0
id          number_of_extents  copy_id
0           32                0
2           32                0
1           32                0
4           32                0
IBM_2145:itsosvcc11:admin>svcinfolsmdiskextentmdisk1
id          number_of_extents  copy_id
0           32                0
2           32                0
1           32                0
4           31                0
IBM_2145:itsosvcc11:admin>svcinfolsmdiskextentmdisk2
id          number_of_extents  copy_id
0           32                0
2           32                0
1           32                0
4           32                0
IBM_2145:itsosvcc11:admin>svcinfolsmdiskextentmdisk3
id          number_of_extents  copy_id
0           32                0
2           32                0
1           32                0
4           32                0
IBM_2145:itsosvcc11:admin>svcinfolsmdiskextentmdisk4
id          number_of_extents  copy_id
0           32                0
2           32                0
1           32                0
4           33                0
IBM_2145:itsosvcc11:admin>svcinfolsmdiskextentmdisk5
id          number_of_extents  copy_id
0           32                0
2           32                0
1           32                0
4           32                0
IBM_2145:itsosvcc11:admin>svcinfolsmdiskextentmdisk6
id          number_of_extents  copy_id
0           32                0
2           32                0
1           32                0
4           32                0
IBM_2145:itsosvcc11:admin>svcinfolsmdiskextentmdisk7
id          number_of_extents  copy_id
0           32                0
2           32                0
1           32                0
4           32                0

```

The use of the extent balancing script

To use the extent balancing script, consider the following points:

- ▶ Migrating extents might have a performance impact if the SAN Volume Controller or (more likely) the MDisk are already at the limit of their I/O capability. The script minimizes the impact by using the minimum priority level for migrations. Nevertheless, many administrators prefer to run these migrations during periods of low I/O workload, such as overnight.
- ▶ You can use command-line options other than **balance.pl** to tune how extent balancing works. For example, you can exclude certain MDisk or volumes from rebalancing. For more information, see the `SVCToolsSetup.doc` file in the `svctools.zip` file.
- ▶ Because the script is written in Perl, the source code is available for you to modify and extend its capabilities. If you want to modify the source code, make sure that you pay attention to the documentation in Plain Old Documentation (POD) format within the script.

5.8 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.

5.8.1 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 5-5 shows how to list the volumes that have extents on a MDisk by using the CLI.

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

```
IBM_2145:itsosvcc11:admin>svcinfo lsmdiskextent 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 the SAN Volume Controller to automatically move all used extents on the MDisk to the remaining MDisks in the storage pool.

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

5.8.2 Verifying the identity of an MDisk before removal

MDisks must appear to the SAN Volume Controller cluster as unmanaged before their controller LUN mapping is removed. Unmapping LUNs from the SAN Volume Controller 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 that are described in “MDisks and storage pools” on page 487, 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 SAN Volume Controller 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 SAN Volume Controller 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 5.8.3, “Correlating the back-end volume with the MDisk” on page 112.

5.8.3 Correlating the back-end volume with the MDisk

The correct correlation between the back-end volume (LUN) with the SAN Volume Controller MDisk is crucial to avoid mistakes and possible outages. You can correlate the back-end volume with MDisk for DS4000 series, DS8000 series, XIV, and V7000 storage controllers.

DS4000 volumes

Identify the DS4000 volumes by using the Logical Drive ID and the LUN that is associated with the host mapping. The example in this section uses the following values:

- ▶ Logical drive ID: 600a0b80001744310000c60b4e2eb524
- ▶ LUN: 3

To identify the logical drive ID by using the Storage Manager Software, on the **Logical/Physical View** tab, right-click a volume, and select **Properties**. The Logical Drive Properties window (see Figure 5-1 on page 113) opens.

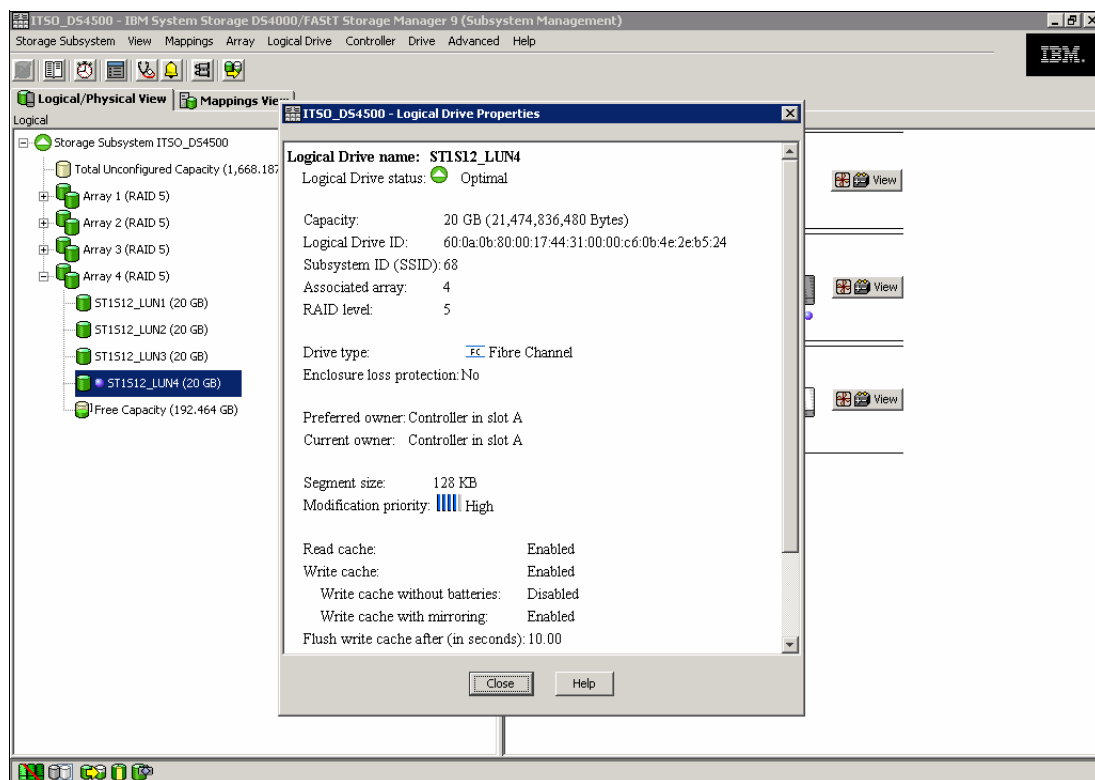


Figure 5-1 Logical Drive Properties window for DS4000

To identify your LUN, on the **Mappings View** tab, select your SAN Volume Controller host group and then look in the LUN column in the right pane, as shown in Figure 5-2.

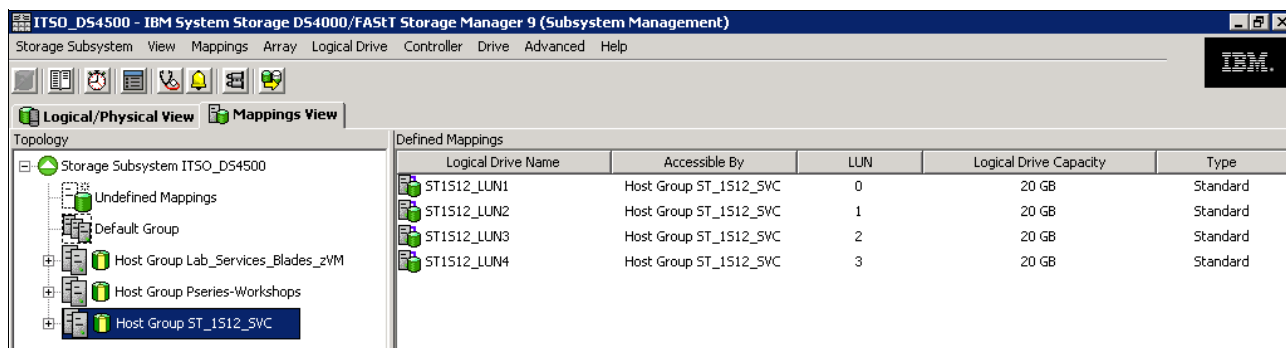


Figure 5-2 Mappings View tab for DS4000

Complete the following steps to correlate the LUN with your corresponding MDisk:

1. Review the MDisk details and the UID field. The first 32 bits of the MDisk UID field (600a0b80001744310000c60b4e2eb524) must be the same as your DS4000 logical drive ID.
2. Make sure that the associated DS4000 LUN correlates with the SAN Volume Controller ctrl_LUN_#. For this task, convert your DS4000 LUN in hexadecimal and check the last two bits in the SAN Volume Controller ctrl_LUN_# field. In the example that is shown in Figure 5-3 on page 114, it is 0000000000000003.

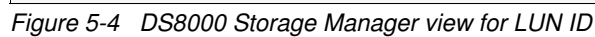
[illegible]

Figure 5-5 MDisk details for DS8000 volume

You can also identify the storage controller from the Storage Subsystem field as DS8K75L3001, which was manually assigned.

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 5-6 shows the Volume Properties dialog box that opens.

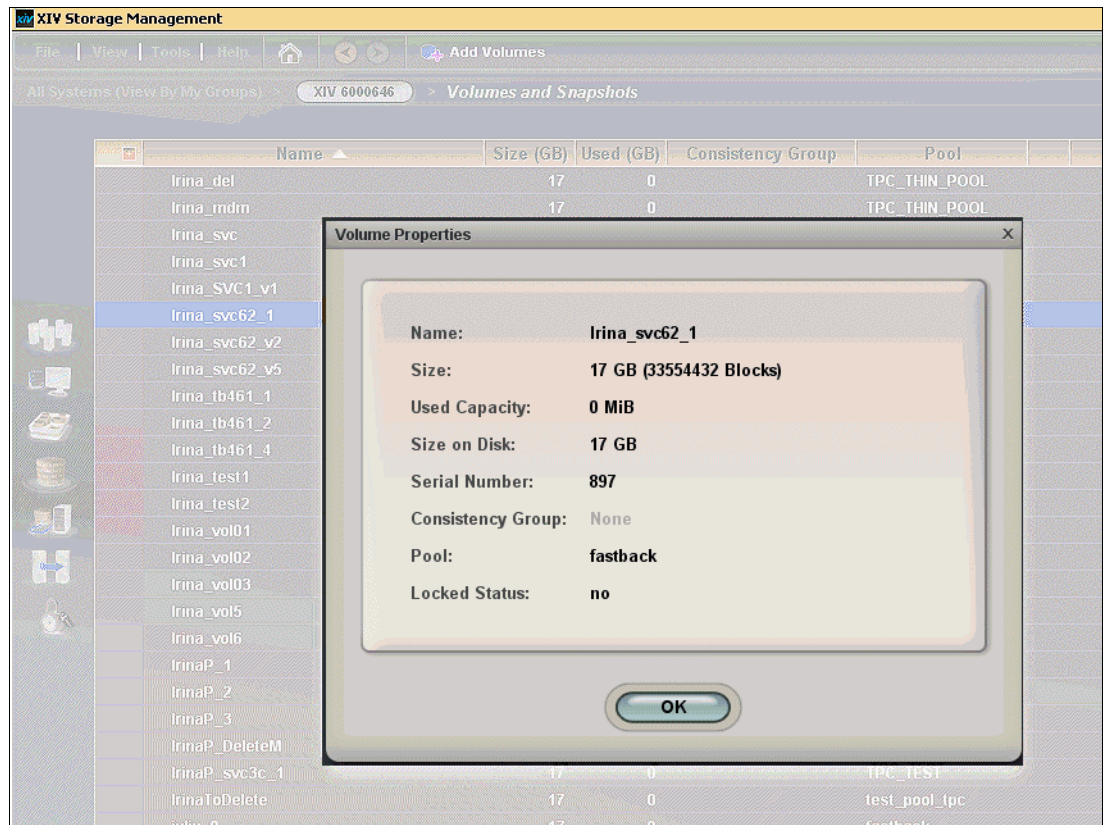


Figure 5-6 XIV Volume Properties dialog box

To identify your LUN, in the Volumes by Hosts view, expand your SAN Volume Controller host group and then review the LUN column, as shown in Figure 5-7 on page 117.

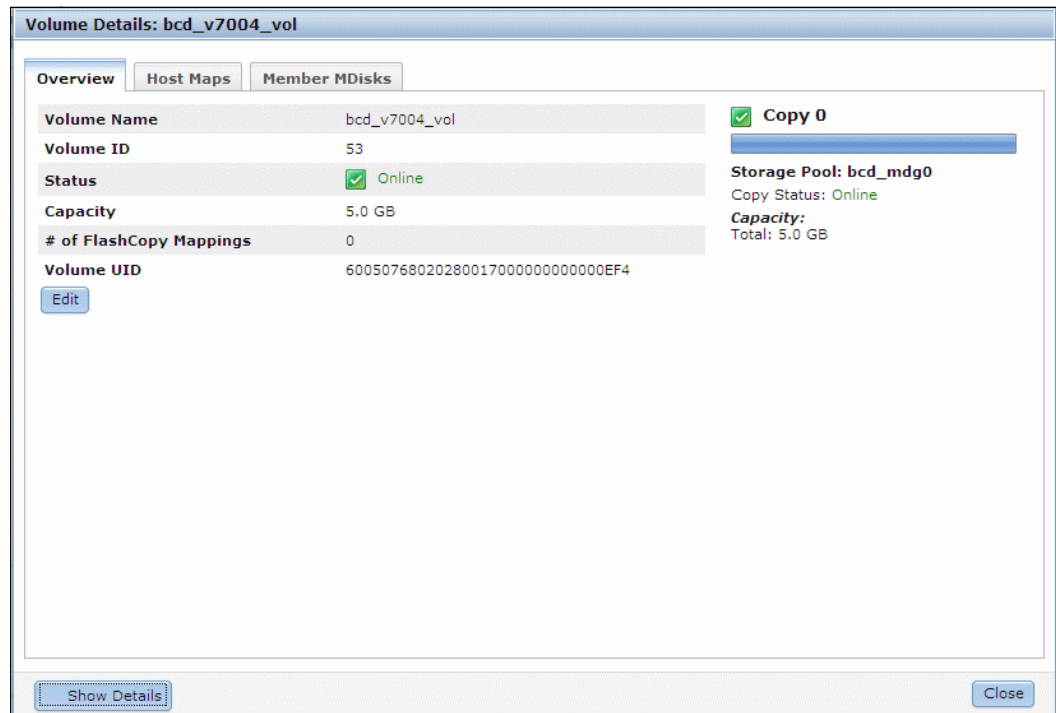


Figure 5-9 Storwize Volume details

2. On the Host Maps tab (see Figure 5-10), check the SCSI ID number for the specific volume. This value is used to match the SAN Volume Controller ctrl_LUN_# (in hexadecimal format).

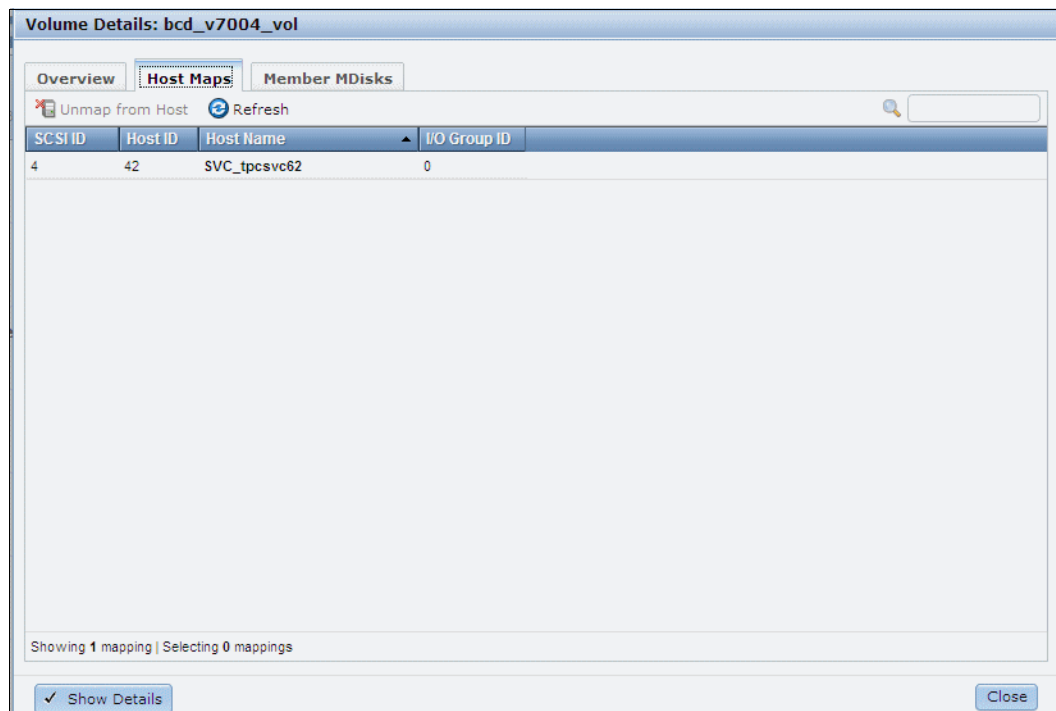


Figure 5-10 Storwize Volume Details for Host Maps

3. On the SAN Volume Controller side, review the MDisk details (see Figure 5-11) and compare the MDisk UID field with the Storwize Volume UID. The first 32 bits should be the same.

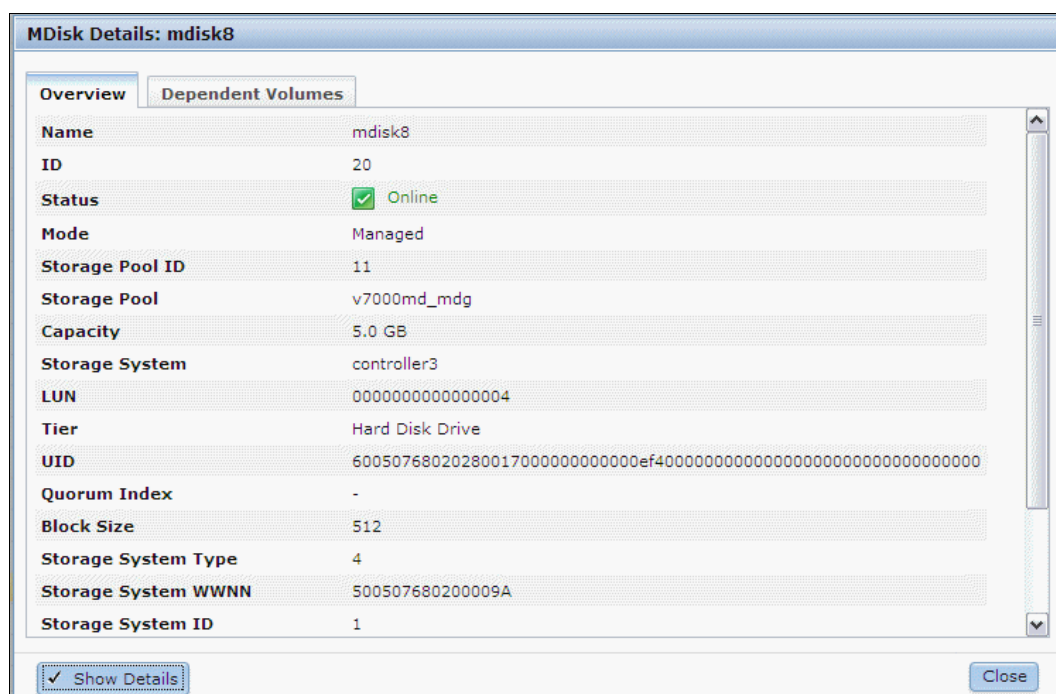


Figure 5-11 SAN Volume Controller MDisk Details for Storwize volumes

4. Double-check that the SAN Volume Controller ctrl_LUN_# is the Storwize SCSI ID number in hexadecimal format. In this example, the number is 0000000000000004.

5.9 Remapping managed MDisks

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

If the LUN is mapped back with different attributes, the SAN Volume Controller 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. Figure 5-1 on page 113 shows the Logical Drive Properties panel for a DS4000 logical drive and includes the logical drive ID (UID) and SSID.

5.10 Controlling extent allocation order for volume creation

When you create a virtual disk, you might want to control the order in which extents are allocated across the MDisks in the storage pool to balance workload across controller resources. For example, you can alternate extent allocation across DA pairs and even and odd extent pools in the DS8000.

For this reason, plan the order that the MDisk are included on the storage pools because the extents allocation follows the sequence that the MDisks were added.

Tip: When volumes are created, the MDisk that contains the first extent is selected by a pseudo-random algorithm. Then, the remaining extents are allocated across MDisks in the storage pool in a round-robin fashion in the order in which the MDisks were added to the storage pool and according with MDisks free extents available.

Table 5-5 shows the initial discovery order of six MDisks. Adding these MDisks to a storage pool in this order results in three contiguous extent allocations that alternate between the even and odd extent pools, as opposed to alternating between extent pools for each extent.

Table 5-5 Initial discovery order

LUN ID	MDisk ID	MDisk name	Controller resource DA pair/extent pool
1000	1	mdisk01	DA2/P0
1001	2	mdisk02	DA6/P16
1002	3	mdisk03	DA7/P30
1100	4	mdisk04	DA0/P9
1101	5	mdisk05	DA4/P23
1102	6	mdisk06	DA5/P39

To change extent allocation so that each extent alternates between even and odd extent pools, the MDisks can be removed from the storage pool and then added again to the storage pool in the new order.

Table 5-6 shows how the MDisks were added back to the storage pool in their new order so that the extent allocation alternates between even and odd extent pools.

Table 5-6 MDisks that were added again

LUN ID	MDisk ID	MDisk name	Controller resource DA pair or extent pool
1000	1	mdisk01	DA2/P0
1100	4	mdisk04	DA0/P9
1001	2	mdisk02	DA6/P16
1101	5	mdisk05	DA4/P23
1002	3	mdisk03	DA7/P30
1102	6	mdisk06	DA5/P39

The following options are available for volume creation:

► Option A

Explicitly select the candidate MDisks within the storage pool that is used (through the CLI only). When you are explicitly selecting the MDisk list, the extent allocation goes round-robin across the MDisks in the order that they are represented in the list that starts with the first MDisk in the list, as shown in the following examples:

- Example A1: Creating a volume with MDisks from the explicit candidate list order md001, md002, md003, md004, md005, and md006

The volume extent allocations then begin at md001 and alternate in a round-robin manner around the explicit MDisk candidate list. In this case, the volume is distributed in the order md001, md002, md003, md004, md005, and md006.

- Example A2: Creating a volume with MDisks from the explicit candidate list order md003, md001, md002, md005, md006, and md004

The volume extent allocations then begin at md003 and alternate in a round-robin manner around the explicit MDisk candidate list. In this case, the volume is distributed in the order md003, md001, md002, md005, md006, and md004.

► Option B

Do not explicitly select the candidate MDisks within a storage pool that is used (through the CLI or GUI). When the MDisk list is not explicitly defined, the extents are allocated across MDisks in the order that they were added to the storage pool and the MDisks that receive the first extent are randomly selected.

For example, you create a volume with MDisks from the candidate list order md001, md002, md003, md004, md005, and md006. This order is based on the definitive list from the order in which the MDisks were added to the storage pool. The volume extent allocations then begin at a random MDisk starting point. (Assume md003 is randomly selected.) The extent allocations alternate in a round-robin manner around the explicit MDisk candidate list that is based on the order in which they were originally added to the storage pool. In this case, the volume is allocated in the order md003, md004, md005, md006, md001, and md002.

When you create striped volumes that specify the MDisk order (if not well-planned), you might have the first extent for several volumes in only one MDisk. This situation can lead to poor performance for workloads that place a large I/O load on the first extent of each volume or that create multiple sequential streams.

Important: When you perform administrative tasks daily, create the striped volumes without specifying the MDisk order.

5.11 Moving an MDisk between SAN Volume Controller clusters

You might want to move an MDisk to a separate SAN Volume Controller cluster. Before you begin this task, consider the following alternatives:

- Use Metro Mirror or Global Mirror to copy the data to a remote cluster.
- Attach a host server to two SAN Volume Controller clusters and use host-based mirroring to copy the data.
- Use storage controller-based copy services. If you use storage controller-based copy services, make sure that the volumes that contain the data are image-mode *and* cache-disabled.

If none of these options are appropriate, complete the following steps to move an MDisk to another cluster:

1. Ensure that the MDisk is in image mode rather than striped or sequential mode.
If the MDisk is in image mode, the MDisk contains only the raw client data and not any SAN Volume Controller metadata. If you want to move data from a non-image mode volume, use the **svctask migratetoimage** command to migrate to a single image-mode MDisk. For a thin-provisioned volume, image mode means that all metadata for the volume is present on the same MDisk as the client data, which not readable by a host, but it can be imported by another SAN Volume Controller cluster.
2. Remove the image-mode volumes from the first cluster by using the **svctask rmvdisk** command.

The -force option: You must *not* use the **-force** option of the **svctask rmvdisk** command. If you use the **-force** option, data in the cache is not written to the disk, which might result in metadata corruption for a thin-provisioned volume.

3. Verify that the volume is no longer displayed by entering the **svcinfo lsvdisk** command. You must wait until the volume is removed to allow cached data to destage to disk.
4. Change the back-end storage LUN mappings to prevent the source SAN Volume Controller cluster from detecting the disk, and then make it available to the target cluster.
5. Enter the **svctask detectmdisk** command on the target cluster.
6. Import the MDisk to the target cluster:
 - If the MDisk is *not* a thin-provisioned volume, use the **svctask mkvdisk** command with the **-image** option.
 - If the MDisk *is* a thin-provisioned volume, use the following options:
 - **-import** instructs the SAN Volume Controller to look for thin volume metadata on the specified MDisk.
 - **-rsize** indicates that the disk is thin-provisioned. The value that is given to **-rsize** must be at least the amount of space that the source cluster that is used on the thin-provisioned volume. If it is smaller, an 1862 error is logged. In this case, delete the volume and enter the **svctask mkvdisk** command again.

The volume is now online. If it is not online and the volume is thin-provisioned, check the SAN Volume Controller error log for an 1862 error. If present, an 1862 error indicates why the volume import failed (for example, metadata corruption). Then, you might be able to use the **repairsevdiskcopy** command to correct the problem.

5.12 MDisk group considerations when Real-time Compression is used

IBM recommends that compressed and non-compressed volumes not be part of the same MDisk group (storage pool) because by mixing volume types, compressed and non-compressed volumes share the same cache partition, which is not recommended.

For more information, see *Real-time Compression in SAN Volume Controller and Storwize V7000*, REDP-4859.



Volumes

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

This chapter includes the following sections:

- ▶ Overview of volumes
- ▶ Volume mirroring
- ▶ Creating volumes
- ▶ Volume migration
- ▶ Preferred paths to a volume
- ▶ Non-Disruptive volume move (NDVM)
- ▶ Cache mode and cache-disabled volumes
- ▶ Effect of a load on storage controllers
- ▶ Setting up FlashCopy services

6.1 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.

6.1.1 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.

6.1.2 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 System Storage SAN Volume Controller 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, the SAN Volume Controller 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 email 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.

6.1.3 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 32 KB, 64 KB, 128 KB, or 256 KB (default).

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. For more information about thin-provisioned FlashCopy, see 6.8.5, “Using thin-provisioned FlashCopy” on page 152.

6.1.4 Compressed volumes

Compressed volume is a thin provision volume. The compression technology is implemented into the SAN Volume Controller and Storwize V7000 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 this website:

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

6.1.5 Thin-provisioned volume

Thin-provisioned volumes save 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.

Some file systems (for example, New Technology File System [NTFS]) write to the whole volume before the overwrite the deleted files. Other file systems reuse space in preference to allocating new space.

File system problems can be moderated by tools, such as “defrag,” or by managing storage by using host Logical Volume Managers (LVMs).

The thin-provisioned volume also depends on how applications use the file system. For example, some applications delete log files only when the file system is nearly full.

For more information about performance, see Part 2, “Performance preferred practices” on page 261.

6.1.6 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 6-1 shows the maximum thin-provisioned volume virtual capacities for an extent size.

Table 6-1 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 6-2 show the maximum thin-provisioned volume virtual capacities for a grain size.

Table 6-2 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

6.1.7 Testing an application with a thin-provisioned volume

To help you understand what works with thin-provisioned volumes, complete the following test:

1. Create a thin-provisioned volume with autoexpand turned off.

2. Test the application.
 - If the application and thin-provisioned volume do *not* work well, the volume fills up. In the worst case, it goes offline.
 - If the application and thin-provisioned volume work well, the volume does not fill up and remains online.
3. Configure warnings and monitor how much capacity is being used.
4. If necessary, the user can expand or shrink the real capacity of the volume.
5. If you determine that the combination of the application and the thin-provisioned volume works well, enable autoexpand.

6.2 Volume mirroring

With the volume mirroring feature, you can create a volume with one or two copies, which provides a simple RAID 1 function. Therefore, a volume has two physical copies of its data. These copies can be in the same storage pools or in different storage pools (with different extent sizes of the storage pool). The first storage pool that is specified contains the primary copy.

If a volume is created with two copies, both copies use the same virtualization policy, as with any other volume. You can have two copies of a volume with different virtualization policies. Combined with thin-provisioning, each mirror of a volume can be thin-provisioned, compressed, or fully allocated and in striped, sequential, or image mode.

A mirrored volume has all of the capabilities of a volume and the same restrictions. For example, a mirrored volume is owned by an I/O group, which is similar to any other volume.

The volume mirroring feature also provides a *point-in-time copy* function that is achieved by splitting a copy from the volume.

6.2.1 Creating or adding a mirrored volume

When a mirrored volume is created and the format is specified, all copies are formatted before the volume comes online. The copies are then considered *synchronized*. Alternatively, if you select the **no synchronization** option, the mirrored volumes are *not synchronized*.

Not synchronizing the mirrored volumes might be helpful in the following cases:

- ▶ If you know that the formatted MDisk space is used for mirrored volumes
- ▶ If synchronization of the copies is not required

6.2.2 Availability of mirrored volumes

Volume mirroring provides a low RAID level, RAID 1 to protect against controller and storage pool failure. By having a low RAID level for volume mirroring, you can create a volume with two copies that are in different storage pools. If one storage controller or storage pool fails, a volume copy is unaffected if it is placed on a different storage controller or in a different storage pool.

For FlashCopy usage, a mirrored volume is only online to other nodes if it is online in its own I/O group and if the other nodes are visible to the same copies as the nodes in the I/O group. If a mirrored volume is a source volume in a FlashCopy relationship, asymmetric path failures or a failure of the I/O group for the mirrored volume can cause the target volume to be taken offline.

6.2.3 Mirroring between controllers

An advantage of mirrored volumes is having the volume copies on different storage controllers or storage pools. Normally, the read I/O is directed to the primary copy, but the primary copy must be available and synchronized.

Important: For the preferred practice and best performance, place all the primary mirrored volumes on the same storage controller, or you might see a performance impact. Selecting the copy that is allocated on the higher performance storage controller maximizes the read performance of the volume.

The write performance is constrained by the lower performance controller because writes must complete to both copies before the volume is considered to be written successfully.

6.3 Creating volumes

To create volumes, follow the procedure that is described in *Implementing the IBM System Storage SAN Volume Controller V6.3*, 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 6.3.3, “Non-Disruptive volume move” on page 133.

Tip: Migrating volumes across I/O groups is 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 the SAN Volume Controller nodes within an I/O group. The *preferred node* cannot easily be changed. If you must change the preferred node, see 6.3.2, “Changing the preferred node within an I/O group” on page 132.

Performance tip: The minimal volume size should be determined by the following formula:
number of mdisks * extent size

For example:

8 mdisks * 256 MB Extent size = 2 GB

Volumes that are smaller than 2 GB are not evenly distributed across all MDisks. For optimal performance, volume size should be the product of 2 GB. This enables full stripe across all of the MDisks and spindles of the storage pool.

- ▶ 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 is 2048.
- ▶ The maximum number of volumes per cluster is 8192 (eight-node cluster).
- ▶ 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 SAN Volume Controller can virtualize. 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. A default value set is no longer available. Extent size is set during the managed disk group creation.

Important: You can migrate volumes only between storage pools that have the same extent size, except for mirrored volumes. The two copies can be in different storage pools with different extent sizes.

As described in 6.1, “Overview of volumes” on page 126, 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.

Important: To avoid negatively affecting system performance, you must thoroughly understand the data layout and workload characteristics if you use sequential mode over striping.

6.3.1 Selecting the storage pool

You can use the SAN Volume Controller to create tiers of storage, where each tier has different performance characteristics.

When volumes are created for a new server, all of the volumes for this specific server should be on a unique storage pool. Later, if you observe that the storage pool is saturated or that your server demands more performance, move some volumes to another storage pool or move all the volumes to a higher tier storage pool. By having volumes from the same server in more than one storage pool, you are increasing the availability risk if any of the storage pools that are related to that server goes offline.

6.3.2 Changing the preferred node within an I/O group

Currently, *no* nondisruptive method is available to change the preferred node within an I/O group. The correct way is to migrate the volume to a recovery group and migrate back with the preferred node. Complete the following steps to use this method:

1. Cease I/O operations to the volume.
2. Disconnect the volume from the host operating system. For example, in Windows, remove the drive letter.
3. On the SAN Volume Controller, unmap the volume from the host.
4. On the SAN Volume Controller, validate that the volume cache is empty. It can take up to 2 minutes for the cache to destage after I/O operations are stopped. Example 6-1 shows cache status for volume TEST_1.

Example 6-1 Cache status on volume TEST1

```
IBM_2145:svccg8:admin>svcinfolsvdisk TEST_1
id 2
name TEST_1
IO_group_id 0
IO_group_name io_grp0
status online
...
fast_write_state empty
...
```

5. On the SAN Volume Controller, move the volume to a temporary I/O group. Example 6-2 and Example 6-3 show the recovery group status and the volume migration to a temporary recover group.

Example 6-2 View recovery group status

```
IBM_2145:svccg8:admin>svcinfolgsiogrp
id name          node_count vdisk_count host_count
0 io_grp0         2          29          9
1 io_grp1         0          0          9
2 io_grp2         0          0          9
3 io_grp3         0          0          9
4 recovery_io_grp 0      0          0          0
```

Example 6-3 Migrate volume to a temporary recovery group

```
IBM_2145:svccg8:admin>svcinfolmovevdisk -force -iogrp recovery_io_grp TEST_1
IBM_2145:svccg8:admin>svcinfolsiogrp
id name          node_count vdisk_count host_count
0 io_grp0         2          28          9
1 io_grp1         0          0          9
2 io_grp2         0          0          9
3 io_grp3         0          0          9
4 recovery_io_grp 0      1          0          0
```

6. On the SAN Volume Controller migrate the volume back to the original I/O group while adding the preferred node options. Example 6-4 shows how to migrate by using the node option.

Example 6-4 Migrating by using the node option

```
IBM_2145:svccg8:admin>svctask movevdisk -iogrp io_grp0 -node node2 TEST_1
```

7. Resume I/O operations on the host.

6.3.3 Non-Disruptive volume move

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. For more information about supported operating systems, see this website:

http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004453#_NDVM

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.

You can also use the management GUI to move volumes between I/O groups nondisruptively. In the management GUI, select **Volumes**  **Volumes**. On the Volumes panel, select the volume that you want to move and the select **Actions**  **Move to Another I/O Group**. The wizard guides you through the steps to move a volume to another I/O group, including any changes to hosts that are required. For more information, click **Need Help** on the associated management GUI panels.

Review the following supported operating system website before you perform any actions:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004510>

To move a volume between I/O groups by using the CLI, complete the following steps:

1. Run the **addvdiskaccess -iogrp iogrp id/name volume id/name** command.
2. Run the **movevdisk -iogrp destination iogrp -node new preferred node volume id/name** command.
3. Run the appropriate commands on the hosts that are mapped to the volume to detect the new paths to the volume in the destination I/O group.

4. After you confirm that the new paths are online, remove access from the old I/O group by running the **rmvdiskaccess -iogrp iogrp id/name volume id/name** command.
5. Run the appropriate commands on the hosts that are mapped to the volume to remove the paths to the old I/O group.

For more information about nondisruptive volume movement in Linux, see the “Host” section of this website:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004510>

Migrating a volume to a new I/O group

Complete the following steps to migrate a volume to a new I/O group:

1. Quiesce all I/O operations for the volume.
2. Determine the hosts that use this volume and ensure that it is properly zoned to the target SAN Volume Controller I/O group.
3. Stop or delete any FlashCopy mappings or Metro Mirror or Global Mirror relationships that use this volume.
4. To check whether the volume is part of a relationship or mapping, run the **svcinfo lsvdisk vdiskname/id** command, where **vdiskname/id** is the name or ID of the volume. Example 6-5 shows that the **vdiskname/id** filter of the **lsvdisk** command is **TEST_1**.

Example 6-5 Output of the lsvdisk command

```
IBM_2145:svccg8:admin>svcinfo lsvdisk TEST_1
id 2
name TEST_1
IO_group_id 0
IO_group_name io_grp0
status online
mdisk_grp_id many
mdisk_grp_name many
capacity 1.00GB
type many
formatted no
mdisk_id many
mdisk_name many
FC_id
FC_name
RC_id
RC_name
vdisk_UID 60050768018205E12000000000000002
...
```

5. Look for the **FC_id** and **RC_id** fields. If these fields are *not* blank, the volume is part of a mapping or a relationship.

Migrating a volume between I/O groups

Complete the following steps to migrate a volume between I/O groups:

1. Cease I/O operations to the volume.
2. Disconnect the volume from the host operating system. For example, in Windows, remove the drive letter.
3. Stop any copy operations.

4. To move the volume across I/O groups, run the `svctask movevdisk -iogrp io_grp1 TEST_1` command.

This command does not work when data is in the SAN Volume Controller cache that must be written to the volume. After 2 minutes, the data automatically destages if no other condition forces an earlier destaging.

5. On the host, rediscover the volume. For example, in Windows, run the **rescan** command, and then mount the volume or add a drive letter. For more information, see Chapter 8, “Hosts” on page 225.
6. Resume copy operations as required.
7. Resume I/O operations on the host.

After any copy relationships are stopped, you can move the volume across I/O Groups with the following command in a SAN Volume Controller:

```
svctask movevdisk-iogrp newiogrpname/id vdiskname/id
```

Where `newiogrpname/id` is the name or ID of the I/O group to which you move the volume, and `vdiskname/id` is the name or ID of the volume.

For example, the following command moves the volume that is named `TEST_1` from its existing I/O group, `io_grp0`, to `io_grp1`:

```
IBM_2145:svccg8:admin>svctask movevdisk-iogrp io_grp1 TEST_1
```

Migrating volumes between I/O groups can be a potential issue if the old definitions of the volumes are not removed from the configuration before the volumes are imported to the host. Migrating volumes between I/O groups is not a dynamic configuration change. However, you must shut down the host before you migrate the volumes. Then, follow the procedure in Chapter 8, “Hosts” on page 225 to reconfigure the SAN Volume Controller volumes to hosts. Remove the stale configuration and restart the host to reconfigure the volumes that are mapped to a host.

For information about how to dynamically reconfigure the SDD for the specific host operating system, see *Multipath Subsystem Device Driver: User's Guide*, GC52-1309.

Important: Do *not* move a volume to an offline I/O group for any reason. Before you move the volumes, you must ensure that the I/O group is online to avoid any data loss.

The command that is shown in step 4 on page 135 does not work if any data is in the SAN Volume Controller cache that must first be flushed out. A **-force** flag is available that discards the data in the cache rather than flushing it to the volume. If the command fails because of outstanding I/Os, wait a few minutes after which the SAN Volume Controller automatically flushes the data to the volume.

Attention: The use of the **-force** flag can result in data integrity issues.

6.4 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 6-3 on page 136.

Table 6-3 Migration types and associated commands

Storage pool-to-storage pool type	Command
Managed-to-managed or Image-to-managed	migratevdisk
Managed-to-image or Image-to-image	migratetoimage

Migrating a volume from one storage pool to another is nondisruptive to the host application by using the volume. Depending on the workload of the SAN Volume Controller, 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.

6.4.1 Image-type to striped-type migration

When you are migrating existing storage into the SAN Volume Controller, the existing storage is brought in as *image-type volumes*, which means that the volume is based on a single MDisk. In general, migrate the volume to a striped-type volume, which is striped across multiple MDisks and, therefore, across multiple RAID arrays when it is practical. You generally expect to see a performance improvement by migrating from an image-type volume to a striped-type volume. Example 6-6 shows the image mode migration command.

Example 6-6 Image mode migration 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 V6.3*, SG24-7933.

You can monitor the migration process by using the `svcinfolsmigrate` command, as shown in Example 6-7.

Example 6-7 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>
```


6.4.2 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. 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 SAN Volume Controller and to migrate data out of virtualization. Image-type volumes cannot be expanded.

The reason for migrating a volume to an image type volume often is to move the data on the disk to a nonvirtualized environment. This operation is also carried out so that you can change the preferred node that is used by a volume. For more information, see 6.3.2, “Changing the preferred node within an I/O group” on page 132.

To migrate a striped-type volume to an image-type volume, you must migrate to an available unmanaged MDisk. The destination MDisk must be greater than or equal to the size of the volume that you want to migrate. Regardless of the mode in which the volume starts, the volume is reported as being in managed mode during the migration. Both of the MDisk that are involved are reported as being in image mode during the migration. If the migration is interrupted by a cluster recovery, the migration resumes after the recovery completes.

Complete the following steps to migrate a striped-type volume to an image-type volume:

1. To determine the name of the volume to be moved, run the **svcinio lsvdisk** command.

Example 6-8 shows the results of running the command.

Example 6-8 The lsvdisk output

```
IBM_2145:svccg8:admin>svcinio lsvdisk -delim :
id:name:IO_group_id:IO_group_name:status:mdisk_grp_id:mdisk_grp_name:capacity:type:FC_id:FC_name:RC_id:RC_name:vdisk_UID
:fc_map_count:copy_count:fast_write_state:se_copy_count
0:NYBIXTDB02_T03:0:io_grp0:online:3:MDG4DS8KL3331:20.00GB:striped::::60050768018205E12000000000000000:0:1:empty:0
1:NYBIXTDB02_2:0:io_grp0:online:0:MDG1DS8KL3001:5.00GB:striped::::60050768018205E12000000000000007:0:1:empty:0
2:TEST_1:0:io_grp0:online:many:many:1.00GB:many::::60050768018205E12000000000000002:0:2:empty:0
3:Migrate_sample:0:io_grp0:online:2:MDG1DS4K:2.00GB:striped::::60050768018205E12000000000000012:0:1:empty:0
```

2. To migrate the volume, get the name of the MDisk to which you migrate it by using the command that is shown in Example 6-9.

Example 6-9 The lsmdisk command output

```
IBM_2145:svccg8:admin>lsmdisk -delim :
id:name:status:mode:mdisk_grp_id:mdisk_grp_name:capacity:ctrl_LUN_#:controller_name:UID:tier
0:D4K_ST1S12_LUN1:online:managed:2:MDG1DS4K:20.0GB:0000000000000000:DS4K:600a0b8000174233000071894e2eccaf00000000000000
0000000000000000:generic_hdd
1:mdisk0:online:array:3:MDG4DS8KL3331:136.2GB::::generic_ssd
2:D8K_L3001_1001:online:managed:0:MDG1DS8KL3001:20.0GB:4010400100000000:DS8K75L3001:6005076305ffc74c00000000000010010000
00000000000000000000000000000000:generic_hdd
...
33:D8K_L3331_1108:online:unmanaged:::20.0GB:4011400800000000:DS8K75L3331:6005076305ffc747000000000000110800000000000000
0000000000000000:generic_hdd
34:D4K_ST1S12_LUN2:online:managed:2:MDG1DS4K:20.0GB:00000000000000001:DS4K:600a0b80001744310000c6094e2eb4e40000000000000
0000000000000000000000:generic_hdd
```

From this command, you can see that D8K_L3331_1108 is the candidate for the image type migration because it is unmanaged.

3. Run the **migratetoimage** command (as shown in Example 6-10) to migrate the volume to the image type.

Example 6-10 The migratetoimage command

```
IBM_2145:svccg8:admin>svctask migratetoimage -vdisk Migrate_sample -threads 4  
-mdisk D8K_L3331_1108 -mdiskgrp IMAGE_Test
```

4. If no unmanaged MDisk is available to which to migrate, remove an MDisk from a storage pool. Removing this MDisk is possible only if enough free extents are on the remaining MDisk that are in the group to migrate any used extents on the MDisk that you are removing.

6.4.3 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.

6.5 Preferred paths to a volume

For I/O purposes, SAN Volume Controller nodes within the cluster are grouped into pairs, which are called *I/O groups*. A single pair is responsible for serving I/O on a specific volume. One node within the I/O group represents the preferred path for I/O to a specific volume. The other node represents the nonpreferred path. This preference alternates between nodes as each volume is created within an I/O group to balance the workload evenly between the two nodes.

The SAN Volume Controller 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 the SAN Volume Controller. 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, the SAN Volume Controller 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 6.3.3, “Non-Disruptive volume move” on page 133.

SDD is aware of the preferred paths that SAN Volume Controller 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.

6.5.1 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 MDisks. 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 MDisks. 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.

Before you run the **chvdisk** command, run the **lsvdisk** command (as shown in Example 6-11) against the volume that you want to throttle to check its parameters.

Example 6-11 The lsvdisk command output

```
IBM_2145:svccg8:admin>svcinfolsvdisk TEST_1
id 2
name TEST_1
IO_group_id 0
IO_group_name io_grp0
status online
mdisk_grp_id many
mdisk_grp_name many
```

```
capacity 1.00GB
type many
formatted no
mdisk_id many
mdisk_name many
FC_id
FC_name
RC_id
RC_name
vdisk_UID 60050768018205E12000000000000002
throttling 0
preferred_node_id 2
fast_write_state empty
cache readwrite
...
```

The throttle setting of zero indicates that no throttling is set. After you check the volume, you can then run the **chvdisk** command.

To modify the throttle setting, run the following command:

```
svctask chvdisk -rate 40 -unitmb TEST_1
```

Running the **lsvdisk** command generates the output that is shown in Example 6-12.

Example 6-12 Output of the lsvdisk command

```
IBM_2145:svccg8:admin>svcinfolsvdisk TEST_1
id 2
name TEST_1
IO_group_id 0
IO_group_name io_grp0
status online
mdisk_grp_id many
mdisk_grp_name many
capacity 1.00GB
type many
formatted no
mdisk_id many
mdisk_name many
FC_id
FC_name
RC_id
RC_name
vdisk_UID 60050768018205E12000000000000002
virtual_disk_throttling (MB) 40
preferred_node_id 2
fast_write_state empty
cache readwrite
...
```

This example shows that the throttle setting (**virtual_disk_throttling**) is 40 MBps on this volume. If you set the throttle setting to an I/O rate by using the I/O parameter (which is the default setting) you do not use the **-unitmb** flag, as shown in the following example:

```
svctask chvdisk -rate 2048 TEST_1
```

As shown in Example 6-13, the throttle setting has no unit parameter, which means that it is an I/O rate setting.

Example 6-13 The chvdisk command and lsvdisk output

```
IBM_2145:svccg8:admin>svctask chvdisk -rate 2048 TEST_1
IBM_2145:svccg8:admin>svcinfolsvdisk TEST_1
id 2
name TEST_1
IO_group_id 0
IO_group_name io_grp0
status online
mdisk_grp_id many
mdisk_grp_name many
capacity 1.00GB
type many
formatted no
mdisk_id many
mdisk_name many
FC_id
FC_name
RC_id
RC_name
vdisk_UID 60050768018205E12000000000000002
throttling 2048
preferred_node_id 2
fast_write_state empty
cache readwrite
...
```

I/O governing rate of zero: An I/O governing rate of 0 (displayed as `virtual_disk_throttling` in the command-line interface [CLI] output of the `lsvdisk` command) does not mean that zero IOPS (or MBps) can be achieved. It means that no throttle is set.

6.6 Cache mode and cache-disabled volumes

You use cache-disabled 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.

6.6.1 Underlying controller remote copy with SAN Volume Controller 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 SAN Volume Controller as image mode disks. The SAN Volume Controller cache must be disabled. You can access the source or the target of the remote copy from a host directly, rather than through the SAN Volume Controller. You can use the SAN Volume Controller copy services with the image mode volume that represents the primary site of the controller remote copy relationship. Do not use SAN Volume Controller copy services with the volume at the secondary site because the SAN Volume Controller does not detect the data that is flowing to this LUN through the controller.

Figure 6-1 shows the relationships between the SAN Volume Controller, the volume, and the underlying storage controller for a cache-disabled volume.

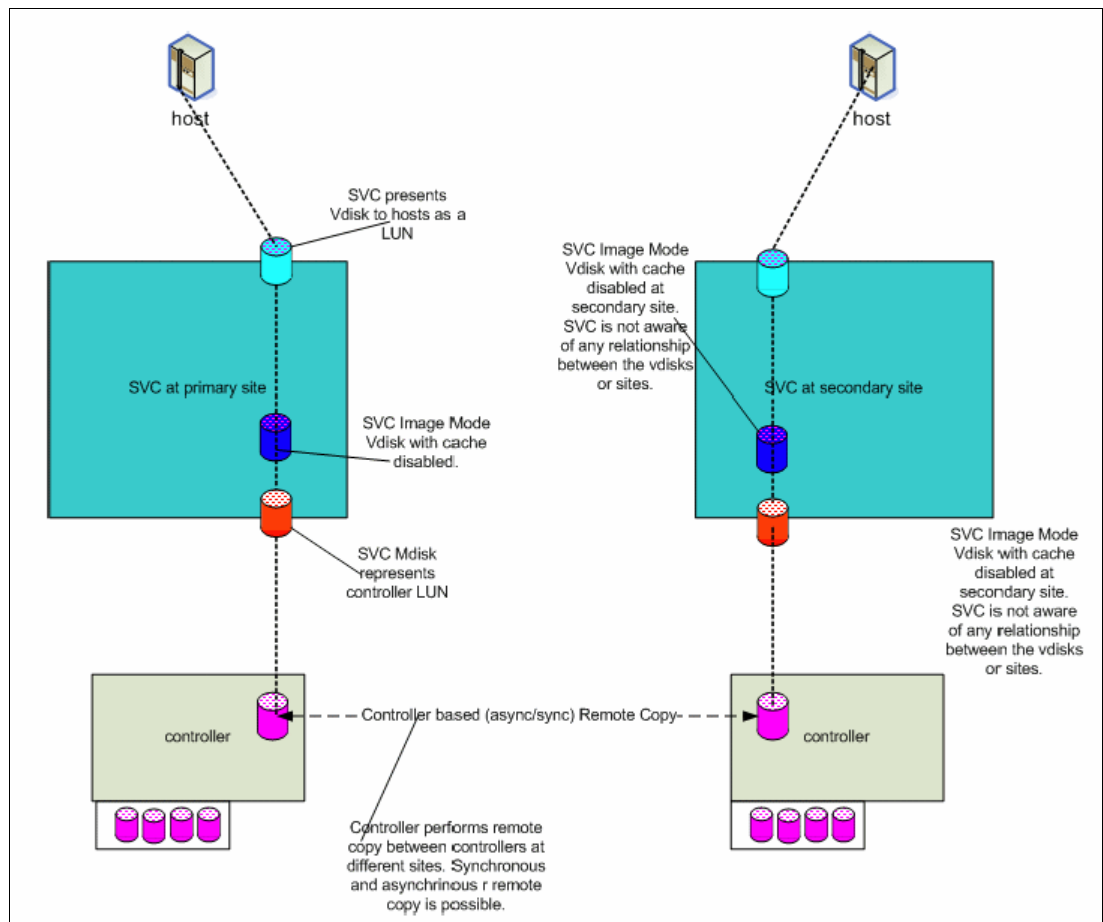


Figure 6-1 Cache-disabled volume in a remote copy relationship

6.6.2 Using underlying controller FlashCopy with SAN Volume Controller 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 the SAN Volume Controller as image mode disks, as shown in Figure 6-2. The SAN Volume Controller cache must be disabled. You can access the source or the target of the FlashCopy from a host directly rather than through the SAN Volume Controller.

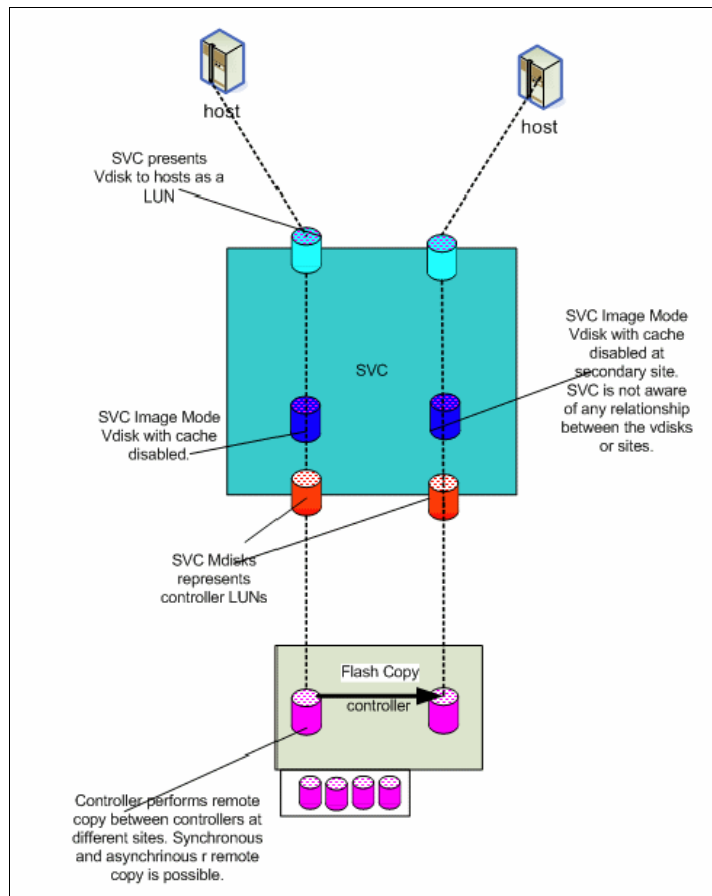


Figure 6-2 FlashCopy with cache-disabled volumes

6.6.3 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 6-14 on page 144 shows an image volume `VDISK_IMAGE_1` that changed the cache parameter after it was created.

Example 6-14 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
name VDISK_IMAGE_1
IO_group_id 0
IO_group_name io_grp0
status online
mdisk_grp_id 5
mdisk_grp_name IMAGE_Test
capacity 20.00GB
type image
formatted no
mdisk_id 33
mdisk_name D8K_L3331_1108
FC_id
FC_name
RC_id
RC_name
vdisk_UID 60050768018205E120000000000000014
throttling 0
preferred_node_id 1
fast_write_state empty
cache readwrite
udid
fc_map_count 0
sync_rate 50
copy_count 1
se_copy_count 0
...

IBM_2145:svccg8:admin>svctask chvdisk -cache none VDISK_IMAGE_1
IBM_2145:svccg8:admin>svcinfolsvdisk VDISK_IMAGE_1
id 9
name VDISK_IMAGE_1
IO_group_id 0
IO_group_name io_grp0
status online
mdisk_grp_id 5
mdisk_grp_name IMAGE_Test
capacity 20.00GB
type image
formatted no
mdisk_id 33
mdisk_name D8K_L3331_1108
FC_id
FC_name
RC_id
RC_name
vdisk_UID 60050768018205E120000000000000014
throttling 0
preferred_node_id 1
fast_write_state empty
```



```
cache none
udid
fc_map_count 0
sync_rate 50
copy_count 1
se_copy_count 0
...
```

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.

6.7 Effect of a load on storage controllers

The SAN Volume Controller can share the capacity of a few MDisk to many more volumes (and, thus, are assigned to hosts that are generating I/O). As a result, a SAN Volume Controller can generate more I/O than the storage controller normally received if a SAN Volume Controller was not in the middle. Adding FlashCopy to this situation can add more I/O to a storage controller and to the I/O that hosts are generating.

When you define volumes for hosts, consider the load that you can put onto a storage controller to ensure that you do not overload a storage controller. Assuming that a typical physical drive can handle 150 IOPS (a Serial Advanced Technology Attachment [SATA] might handle slightly fewer than 150 IOPS), you can calculate the maximum I/O capability that a storage pool can handle. Then, as you define the volumes and the FlashCopy mappings, calculate the maximum average I/O that the SAN Volume Controller receives per volume before you start to overload your storage controller.

From the example of the effect of FlashCopy on I/O, we can make the following assumptions:

- ▶ An MDisk is defined from an entire array. That is, the array provides only one LUN, and that LUN is given to the SAN Volume Controller as an MDisk.
- ▶ Each MDisk that is assigned to a storage pool is the same size and same RAID type and comes from a storage controller of the same type.
- ▶ MDisk from a storage controller are entirely in the same storage pool.

The raw I/O capability of the storage pool is the sum of the capabilities of its MDisk. For example, five RAID 5 MDisk with eight component disks on a typical back-end device have the following I/O capability:

$$5 \times (150 \times 7) = 5250$$

This raw number might be constrained by the I/O processing capability of the back-end storage controller.

FlashCopy copying contributes to the I/O load of a storage controller, which you must consider. The effect of a FlashCopy adds several loaded volumes to the group; therefore, a weighting factor can be calculated to make allowance for this load.

The effect of FlashCopy copies depends on the type of I/O that is taking place. For example, in a group with two FlashCopy copies and random writes to those volumes, the weighting factor is $14 \times 2 = 28$.

Table 6-4 on page 146 shows the total weighting factor for FlashCopy copies.

Table 6-4 FlashCopy weighting

Type of I/O to the volume	Effect on I/O	Weight factor for FlashCopy
None or very little	Insignificant	0
Reads only	Insignificant	0
Sequential reads and writes	Up to 2 x the number of I/Os	2 x F
Random reads and writes	Up to 15 x the number of I/Os	14 x F
Random writes	Up to 50 x the number of I/Os	49 x F

Therefore, to calculate the average I/O per volume before overloading the storage pool, use the following formula:

$$\text{I/O rate} = (\text{I/O Capability}) / (\text{No volumes} + \text{Weighting Factor})$$

By using the example storage pool as defined earlier in this section, consider a situation in which you add 20 volumes to the storage pool and that storage pool can sustain 5250 IOPS, and two FlashCopy mappings also have random reads and writes. In this case, the average I/O rate is calculated by the following formula:

$$5250 / (20 + 28) = 110$$

Therefore, if half of the volumes sustain 200 I/Os and the other half of the volumes sustain 10 I/Os, the average is still 110 IOPS.

Summary

As you can see from the examples in this section, Tivoli Storage Productivity Center is a powerful tool for analyzing and solving performance problems. To monitor the performance of your system, you can use the read and response times parameter for volumes and MDisk. This parameter shows everything that you need in one view and it is the key day-to-day performance validation metric. You can easily notice if a system that usually had 2 ms writes and 6 ms reads suddenly has 10 ms writes and 12 ms reads and is becoming overloaded. A general monthly check of CPU usage shows how the system is growing over time and highlights when you must add an I/O group (or cluster).

In addition, rules apply to OLTP-type workloads, such as the maximum I/O rates for back-end storage arrays. However, for batch workloads, the maximum I/O rates depend on many factors such as workload, backend storage, code levels, and security.

6.8 Setting up FlashCopy services

Regardless of whether you use FlashCopy to make one target disk or multiple target disks, consider the application and the operating system. Even though the SAN Volume Controller can make an exact image of a disk with FlashCopy at the point that you require, it is pointless if the operating system or the application cannot use the copied disk.

Data that is stored to a disk from an application normally goes through the following steps:

1. The application records the data by using its defined application programming interface. Certain applications might first store their data in application memory before they send it to disk later. Normally, subsequent reads of the block that are being written get the block in memory if it is still there.

2. The application sends the data to a file. The file system that accepts the data might buffer it in memory for a period.
3. The file system sends the I/O to a disk controller after a defined period (or even based on an event).
4. The disk controller might cache its write in memory before it sends the data to the physical drive.
If the SAN Volume Controller is the disk controller, it stores the write in its internal cache before it sends the I/O to the real disk controller.
5. The data is stored on the drive.

At any point, any number of unwritten blocks of data might be in any of these steps and are waiting to go to the next step.

Also, the order of the data blocks that were created in step 1 might not be in the same order that was used when the blocks are sent to steps 2, 3, or 4. Therefore, at any point, data that arrives in step 4 might be missing a vital component that was not yet sent from step 1, 2, or 3.

FlashCopy copies often are created with data that is visible from step 4. Therefore, to maintain application integrity, any I/O that is generated in step 1 must make it to step 4 when the FlashCopy is started when a FlashCopy is created. There must not be any outstanding write I/Os in steps 1, 2, or 3. If write I/Os are outstanding, the copy of the disk that is created at step 4 is likely to be missing those transactions. If the FlashCopy is to be used, these missing I/Os can make it unusable.

6.8.1 Making a FlashCopy volume with application data integrity

Complete the following steps to create FlashCopy copies:

1. Verify which volume your host is writing to as part of its day-to-day usage. This volume becomes the source volume in our FlashCopy mapping.
2. Identify the size and type (image, sequential, or striped) of the volume. If the volume is an image mode volume, you must know its size in bytes. If it is a sequential or striped mode volume, its size, as reported by the SAN Volume Controller GUI or SAN Volume Controller CLI, is sufficient.

To identify the volumes in an SAN Volume Controller cluster, run the **svcinfo lsvdisk** command, as shown in Example 6-15.

Example 6-15 Running the command line to see the type of the volumes

```
IBM_2145:svccg8:admin>svcinfo lsvdisk -delim :
id:name:IO_group_id:IO_group_name:status:mdisk_grp_id:mdisk_grp_name:capacity:type:FC_id:FC_name:RC_id:RC_n
ame:vdisk_UID:fc_map_count:copy_count:fast_write_state:se_copy_count
0:NYBIXTDB02_T03:0:io_grp0:online:3:MDG4DS8KL3331:20.00GB:striped:::::60050768018205E12000000000000000:0:1:
empty:0
1:NYBIXTDB02_2:0:io_grp0:online:0:MDG1DS8KL3001:5.00GB:striped:::::60050768018205E12000000000000007:0:1:emp
ty:0
3:Vdisk_1:0:io_grp0:online:2:MDG1DS4K:2.00GB:striped:::::60050768018205E12000000000000012:0:1:empty:0
9:VDISK_IMAGE_1:0:io_grp0:online:5:IMAGE_Test:20.00GB:image:::::60050768018205E12000000000000014:0:1:empty:
0
...
```

If you want to put `Vdisk_1` into a FlashCopy mapping, you do not need to know the byte size of that volume because it is a striped volume. Creating a target volume of 2 GB is sufficient. The `VDISK_IMAGE`, which is used in our example, is an image-mode volume. In this case, you must know its exact size in bytes.

Example 6-16 uses the **-bytes** parameter of the **svcinfo lsvdisk** command to find its exact size. Therefore, you must create the target volume with a size of 21474836480 bytes, not 20 GB.

Example 6-16 Finding the size of an image mode volume by using the CLI

```
IBM_2145:svccg8:admin>svcinfo lsvdisk -bytes VDISK_IMAGE_1
id 9
name VDISK_IMAGE_1
IO_group_id 0
IO_group_name io_grp0
status online
mdisk_grp_id 5
mdisk_grp_name IMAGE_Test
capacity 21474836480
type image
formatted no
mdisk_id 33
mdisk_name D8K_L3331_1108
FC_id
FC_name
RC_id
RC_name
vdisk_UID 60050768018205E12000000000000014
...
```

3. Create a target volume of the required size as identified by the source volume. The target volume can be an image, sequential, or striped mode volume. The only requirement is that it must be the same size as the source volume. The target volume can be cache-enabled or cache-disabled.
4. Define a FlashCopy mapping, making sure that you have the source and target disks that are defined in the correct order. If you use your newly created volume as a source and the existing host volume as the target, you corrupt the data on the volume if you start the FlashCopy.
5. As part of the define step, specify a copy rate of 0 - 100. The copy rate determines how quickly the SAN Volume Controller copies the data from the source volume to the target volume.

When you set the copy rate to 0 (NOCOPY), SAN Volume Controller copies to the target volume (if it is mounted, read/write to a host) only the blocks that changed since the mapping was started on the source volume.

6. Run the prepare process for FlashCopy mapping. This process can take several minutes to complete because it forces the SAN Volume Controller to flush any outstanding write I/Os that belong to the source volumes to the disks of the storage controller. After the preparation completes, the mapping has a *Prepared* status and the target volume behaves as though it was a cache-disabled volume until the FlashCopy mapping is started or deleted.

You can perform step 1 on page 147 to step 5 when the host that owns the source volume performs its typical daily activities (that is, no downtime). During the prepare process (step 6, which can last several minutes), there might be a delay in I/O throughput because the cache on the volume is temporarily disabled.

FlashCopy mapping effect on Metro Mirror relationship: If you create a FlashCopy mapping where the source volume is a target volume of an active mirror relationship, you add more latency to that existing Metro Mirror relationship. You might also affect the host that is using the source volume of that Metro Mirror relationship as a result.

The reason for the additional latency is that FlashCopy prepares and disables the cache on the source volume, which is the target volume of the Metro Mirror relationship. Therefore, all write I/Os from the Metro Mirror relationship must commit to the storage controller before the completion is returned to the host.

7. After the FlashCopy mapping is prepared, quiesce the host by forcing the host and the application to stop I/Os and flush any outstanding write I/Os to disk. This process is different for each application and for each operating system.

One way to quiesce the host is to stop the application and unmount the volume from the host.

You must perform this step when the application I/O is stopped (or suspended). Steps 8 and 9 complete quickly, and application unavailability is minimal.

8. When the host completes its flushing, start the FlashCopy mapping. The FlashCopy starts quickly (at most, a few seconds).
9. After the FlashCopy mapping starts, unquiesce your application (or mount the volume and start the application). The cache is now re-enabled for the source volumes. The FlashCopy continues to run in the background and ensures that the target volume is an exact copy of the source volume when the FlashCopy mapping was started.

The target FlashCopy volume can now be assigned to another host, and it can be used for read or write even though the FlashCopy process is not completed.

Hint: You might intend to use the target volume on the same host (as the source volume is) at the same time that the source volume is visible to that host. You might need to perform more preparation steps to enable the host to access volumes that are identical.

6.8.2 Making multiple related FlashCopy volumes with data integrity

Where a host has more than one volume and those volumes are used by one application, you might need to perform FlashCopy consistency across all disks at the same time to preserve data integrity. The following examples are situations in which you might need to perform this consistency:

- A Windows Exchange server has more than one drive, and each drive is used for an Exchange Information Store. For example, the exchange server has a D drive, an E drive, and an F drive. Each drive is a SAN Volume Controller volume that is used to store different information stores for the Exchange server.

Thus, when a “snap copy” of the exchange environment is performed, all three disks must be flashed at the same time. This way, if they are used during recovery, no information store has more recent data on it than another information store.

- A UNIX relational database has several volumes to hold different parts of the relational database. For example, two volumes are used to hold two distinct tables, and a third volume holds the relational database transaction logs.

Again, when a snap copy of the relational database environment is taken, all three disks must be in sync. That way, when they are used in a recovery, the relational database is not missing any transactions that might occur if each volume was copied by using FlashCopy independently.

To ensure that data integrity is preserved when volumes are related to each other, complete the following steps:

1. Ensure that your host is writing to the volumes as part of its daily activities. These volumes become the source volumes in the FlashCopy mappings.
2. Identify the size and type (image, sequential, or striped) of each source volume. If any of the source volumes is an image mode volume, you must know its size in bytes. If any of the source volumes are sequential or striped mode volumes, their size, as reported by the SAN Volume Controller GUI or SAN Volume Controller command line, is sufficient.
3. Create a target volume of the required size for each source that is identified in step 2. The target volume can be an image, sequential, or striped mode volume. The only requirement is that they must be the same size as their source volume. The target volume can be cache-enabled or cache-disabled.
4. Define a FlashCopy consistency group. This consistency group is linked to each FlashCopy mapping that you defined; therefore, that data integrity is preserved between each volume.
5. Define a FlashCopy mapping for each source volume, making sure that you defined the source disk and the target disk in the correct order. If you use any of your newly created volumes as a source and the volume of the existing host as the target, you destroy the data on the volume if you start the FlashCopy.

When the mapping is defined, link this mapping to the FlashCopy consistency group that you defined in the previous step.

As part of defining the mapping, you can specify the copy rate of 0 - 100. The copy rate determines how quickly the SAN Volume Controller copies the source volumes to the target volumes. When you set the copy rate to 0 (NOCOPY), SAN Volume Controller copies only the blocks that changed on any volume since the consistency group was started on the source volume or the target volume (if the target volume is mounted read/write to a host).

6. Prepare the FlashCopy consistency group. This preparation process can take several minutes to complete because it forces the SAN Volume Controller to flush any outstanding write I/Os that belong to the volumes in the consistency group to the disk of the storage controller. After the preparation process completes, the consistency group has a *Prepared* status, and all source volumes behave as though they were cache-disabled volumes until the consistency group is started or deleted.

You can perform step 1 on page 150 - step 6 on page 150 when the host that owns the source volumes is performing its typical daily duties (that is, no downtime). During the prepare step (which can take several minutes) you might experience a delay in I/O throughput because the cache on the volumes is temporarily disabled.

More latency: If you create a FlashCopy mapping where the source volume is a target volume of an active Metro Mirror relationship, this mapping adds latency to that existing Metro Mirror relationship. It also can affect the host that is using the source volume of that Metro Mirror relationship as a result. The reason for the added latency is that the preparation process of the FlashCopy consistency group disables the cache on *all* source volumes, which might be target volumes of a Metro Mirror relationship. Therefore, all write I/Os from the Metro Mirror relationship must commit to the storage controller before the complete status is returned to the host.

7. After the consistency group is prepared, quiesce the host by forcing the host and the application to stop I/Os and to flush any outstanding write I/Os to disk. This process differs for each application and for each operating system. One way to quiesce the host is to stop the application and unmount the volumes from the host.

You must perform this step when the application I/O is stopped (or suspended). However, steps 8 and 9 complete quickly and application unavailability is minimal.

8. When the host completes its flushing, start the consistency group. The FlashCopy start completes quickly (at most, in a few seconds).
9. After the consistency group starts, unquiesce your application (or mount the volumes and start the application), at which point the cache is re-enabled. FlashCopy continues to run in the background and preserves the data that existed on the volumes when the consistency group was started.

The target FlashCopy volumes can now be assigned to another host and used for read or write, even though the FlashCopy processes were not completed.

Hint: Consider a situation where you intend to use any target volumes on the same host as their source volume at the same time that the source volume is visible to that host. In this case, you might need to perform more preparation steps to enable the host to access volumes that are identical.

6.8.3 Creating multiple identical copies of a volume

Since the release of SAN Volume Controller 4.2, you can create multiple point-in-time copies of a source volume. These point-in-time copies can be made at different times (for example, hourly) so that an image of a volume can be captured before a previous image completes.

If you are required to have more than one volume copy that is created at the same time, use FlashCopy consistency groups. By placing the FlashCopy mappings into a consistency group (where each mapping uses the same source volumes), each target is an identical image of all the other volume FlashCopy targets when the FlashCopy consistency group is started.

With the volume mirroring feature, you can have one or two copies of a volume. For more information, see 6.2, “Volume mirroring” on page 129.

6.8.4 Creating a FlashCopy mapping with the incremental flag

By creating a FlashCopy mapping with the incremental flag, only the data that changed since the last FlashCopy was started is written to the target volume. This function is necessary in cases where you want, for example, a full copy of a volume for disaster tolerance, application testing, or data mining. It greatly reduces the time that is required to establish a full copy of the source data as a new snapshot when the first background copy is completed. In cases where clients maintain fully independent copies of data as part of their disaster tolerance strategy, the use of incremental FlashCopy can be useful as the first layer in their disaster tolerance and backup strategy.

6.8.5 Using thin-provisioned FlashCopy

By using the thin-provisioned volume feature, which was introduced in SAN Volume Controller 4.3, FlashCopy can be used in a more efficient way. A thin-provisioned volume allows for the late allocation of MDisk space. Thin-provisioned volumes present a virtual size to hosts. The real storage pool space (that is, the number of extents x the size of the extents) that is allocated for the volume might be considerably smaller.

Thin volumes that are used as target volumes offer the opportunity to implement a thin-provisioned FlashCopy. Thin volumes that are used as a source volume and a target volume can also be used to make point-in-time copies.

You use thin-provisioned volumes in a FlashCopy relationship in the following scenarios:

- ▶ Copy of a thin source volume to a thin target volume
The background copy copies only allocated regions, and the incremental feature can be used for refresh mapping (after a full copy is complete).
- ▶ Copy of a fully allocated source volume to a thin target volume
For this combination, you must have a zero copy rate to avoid fully allocating the thin target volume.

Default grain size: The default values for grain size are different. The default value is 256 KB for a thin-provisioned volume and 256 KB for FlashCopy mapping.

You can use thin volumes for cascaded FlashCopy and multiple target FlashCopy. You can also mix thin volumes with normal volumes, which can also be used for incremental FlashCopy. However, the use of thin volumes for incremental FlashCopy makes sense only if the source and target are thin-provisioned.

For more information, see *IBM System Storage SAN Volume Controller and Storwize V7000 Replication Family Services*, SG24-7574-02, which is available at this website:

<http://www.redbooks.ibm.com/abstracts/sg247574.html?Open>

The exception is where the thin target volume becomes a production volume (which is subjected to ongoing heavy I/O). In this case, use the 256 KB thin-provisioned grain size to provide better long-term I/O performance at the expense of a slower initial copy.

FlashCopy grain size: Even if the 256 KB thin-provisioned volume grain size is chosen, it is still beneficial to keep the FlashCopy grain size to 64 KB. Then, you can still minimize the performance impact to the source volume, even though this size increases the I/O workload on the target volume. Clients with large numbers of FlashCopy and remote copy relationships might still be forced to choose a 256 KB grain size for FlashCopy because of constraints on the amount of bitmap memory.

6.8.6 Using FlashCopy with your backup application

If you are using FlashCopy with your backup application and you do not intend to keep the target disk after the backup completes, create the FlashCopy mappings by using the **NOCOPY** option (background copy rate = 0).

If you intend to keep the target so that you can use it as part of a quick recovery process, you might choose one of the following options:

- ▶ Create the FlashCopy mapping with the **NOCOPY** option initially. If the target is used and migrated into production, you can change the copy rate at the appropriate time to the appropriate rate to copy all the data to the target disk. When the copy completes, you can delete the FlashCopy mapping and delete the source volume, freeing the space.
- ▶ Create the FlashCopy mapping with a *low copy rate*. The use of a low rate might enable the copy to complete without affecting your storage controller, which leaves bandwidth available for production work. If the target is used and migrated into production, you can change the copy rate to a higher value at the appropriate time to ensure that all data is copied to the target disk. After the copy completes, you can delete the source, which frees the space.
- ▶ Create the FlashCopy with a *high copy rate*. Although this copy rate might add more I/O burden to your storage controller, it ensures that you get a complete copy of the source disk as quickly as possible.

By using the target on a different storage pool, which, in turn, uses a different array or controller, you reduce your window of risk if the storage that provides the source disk becomes unavailable.

With multiple target FlashCopy, you can now use a combination of these methods. For example, you can use the **NOCOPY** rate for an hourly snapshot of a volume with a daily FlashCopy that uses a high copy rate.

6.8.7 Migrating data by using FlashCopy

SAN Volume Controller FlashCopy can help with data migration, especially if you want to migrate from a controller (and your own testing reveals that the SAN Volume Controller can communicate with the device). Another reason to use SAN Volume Controller FlashCopy is to keep a copy of your data behind on the old controller to help with a back-out plan. You might use this method if you want to stop the migration and revert to the original configuration.

Complete the following steps to use FlashCopy to help migrate to a new storage environment with minimum downtime so that you can leave a copy of the data in the old environment if you must back up to the old configuration:

1. Verify that your hosts are using storage from an unsupported controller or a supported controller that you plan on retiring.
2. Install the new storage into your SAN fabric and define your arrays and LUNs. Do not mask the LUNs to any host. You mask them to the SAN Volume Controller later.
3. Install the SAN Volume Controller into your SAN fabric and create the required SAN zones for the SAN Volume Controller nodes and SAN Volume Controller to see the new storage.
4. Mask the LUNs from your new storage controller to the SAN Volume Controller. Run the **svctask detectmdisk** command on the SAN Volume Controller to discover the new LUNs as MDisks.
5. Place the MDisks into the appropriate storage pool.
6. Zone the hosts to the SAN Volume Controller (and maintain their current zone to their storage) so that you can discover and define the hosts to the SAN Volume Controller.
7. At an appropriate time, install the IBM SDD onto the hosts that soon use the SAN Volume Controller for storage. If you tested to ensure that the host can use SDD and the original driver, you can perform this step anytime before the next step.
8. Quiesce or shut down the hosts so that they no longer use the old storage.

9. Change the masking on the LUNs on the old storage controller so that the SAN Volume Controller is now the only user of the LUNs. You can change this masking one LUN at a time. This way, you can discover them (in the next step) one at a time and not mix up any LUNs.
10. Run the **svctask detectmdisk** command to discover the LUNs as MDisks. Then, run the **svctask chmdisk** command to give the LUNs a more meaningful name.
11. Define a volume from each LUN and note its exact size (to the number of bytes) by running the **svcinfo lsvdisk** command.
12. Define a FlashCopy mapping and start the FlashCopy mapping for each volume by following the steps that are described in 6.8.1, “Making a FlashCopy volume with application data integrity” on page 147.
13. Assign the target volumes to the hosts and then restart your hosts. Your host sees the original data with the exception that the storage is now an IBM SAN Volume Controller LUN.

You now have a copy of the existing storage and the SAN Volume Controller is not configured to write to the original storage. Therefore, if you encounter any problems with these steps, you can reverse everything that you did, assign the old storage back to the host, and continue without the SAN Volume Controller.

By using FlashCopy, any incoming writes go to the new storage subsystem and any read requests that were not copied to the new subsystem automatically come from the old subsystem (the FlashCopy source).

You can alter the FlashCopy copy rate (as appropriate) to ensure that all the data is copied to the new controller.

After FlashCopy completes, you can delete the FlashCopy mappings and the source volumes. After all the LUNs are migrated across to the new storage controller, you can remove the old storage controller from the SAN Volume Controller node zones and then, optionally, remove the old storage controller from the SAN fabric.

You can also use this process if you want to migrate to a new storage controller and not keep the SAN Volume Controller after the migration. In step 2 on page 153, make sure that you create LUNs that are the same size as the original LUNs. Then, in step 11, use image mode volumes. When the FlashCopy mappings are completed, you can shut down the hosts and map the storage directly to them, remove the SAN Volume Controller, and continue on the new storage controller.

6.8.8 Summary of FlashCopy rules

To summarize, 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. If the target volumes are mapped to hosts, they have I/O errors.
- ▶ A volume cannot be a source in one FlashCopy mapping and a target in another FlashCopy mapping.
- ▶ A volume can be the source for up to 256 targets.

Starting with SAN Volume Controller V6.2.0.0, 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.

6.8.9 IBM Tivoli Storage FlashCopy Manager

The management of many large FlashCopy relationships and consistency groups is a complex task without a form of automation for assistance. IBM Tivoli FlashCopy Manager V2.2 provides integration between the SAN Volume Controller and Tivoli Storage Manager for Advanced Copy Services. It provides application-aware backup and restore by using the SAN Volume Controller FlashCopy features and function.

For more information about how IBM Tivoli Storage FlashCopy Manager interacts with the IBM System Storage SAN Volume Controller, see *IBM SAN Volume Controller and IBM Tivoli Storage FlashCopy Manager*, REDP-4653.

For more information about IBM Tivoli Storage FlashCopy Manager, see this website:

<http://www.ibm.com/software/tivoli/products/storage-flashcopy-mgr/>

6.8.10 IBM System Storage Support for Microsoft Volume Shadow Copy Service

The SAN Volume Controller 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 SAN Volume Controller 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 SAN Volume Controller 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 *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933.



Remote copy services

This chapter describes the preferred practices for using the remote copy services Metro Mirror and Global Mirror. The main focus is on intercluster Global Mirror relationships. For more information about the implementation and setup of IBM System Storage SAN Volume Controller, including remote copy and intercluster link (ICL), see *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933.

This chapter includes the following sections:

- ▶ Introduction to remote copy services
- ▶ SAN Volume Controller remote copy functions by release
- ▶ Terminology and functional concepts
- ▶ Intercluster link
- ▶ Global Mirror design points
- ▶ Global Mirror planning
- ▶ Global Mirror use cases
- ▶ Intercluster Metro Mirror and Global Mirror source as an FC target
- ▶ States and steps in the Global Mirror relationship
- ▶ 1920 errors
- ▶ Monitoring remote copy relationships

7.1 Introduction to remote copy services

The general application of a remote copy service is to maintain two copies of a data set. Often, the two copies are separated by some distance (which is why the term *remote* is used to describe the copies) but having remote copies is not a prerequisite.

As implemented by SAN Volume Controller, remote copy services can be configured in the form of Metro Mirror or Global Mirror. Both are based on two or more independent SAN Volume Controller clusters that are connected on a Fibre Channel (FC) fabric (intracluster Metro Mirror, which is a single cluster in which remote copy relationships exist). The clusters are configured in a remote copy partnership over the FC fabric. They connect (FC login) to each other and establish communications in the same way as though they were nearby on the same fabric. The only difference is in the expected latency of the communication, the bandwidth capability of the ICL, and the availability of the link as compared with the local fabric.

Local and remote clusters in the remote copy partnership contain volumes in a one-to-one mapping that are configured as a *remote copy relationship*. This relationship maintains the two identical copies. Each volume performs a designated role. The *local volume* functions as the source (and services runtime host application I/O) and the *remote volume* functions as the target, which shadows the source and is accessible as read-only.

SAN Volume Controller offers the following remote copy solutions that are based on distance (they differ by implication mode of operation):

- Metro Mirror (synchronous mode)

This mode is used over metropolitan distances (less than 5 km). Before foreground writes (writes to the target volume) and mirrored foreground writes (shadowed writes to the target) are acknowledged as complete to the host application, they are committed at the local and remote cluster.

Tip: This solution ensures that the target volume is fully up-to-date, but the application is fully exposed to the latency and bandwidth limitations of ICL. Where this remote copy solution is truly remote, it might have an adverse effect on application performance.

- Global Mirror (asynchronous mode)

This mode of operation allows for greater intercluster distance and deploys an asynchronous remote write operation. Foreground writes at the local clusters are started in normal run time, where their associated mirrored foreground writes at the remote cluster are started asynchronously. Write operations are completed on the target volume (local cluster) and are acknowledged to the host application before they are completed at the source volume (remote cluster).

Regardless of which mode of remote copy service is deployed, operations between clusters are driven by the background and foreground write I/O processes.

Background write synchronization and *resynchronization* writes I/O across the ICL (which is performed in the background) to synchronize source volumes to target mirrored volumes on a remote cluster. This concept is also referred to as a *background copy*.

Foreground I/O reads and writes I/O on a local SAN, which generates a *mirrored foreground write I/O* that is across the ICL and remote SAN.

When you consider a remote copy solution, you must consider each of these processes and the traffic that they generate on the SAN and ICL. You must understand how much traffic the SAN can take (without disruption) and how much traffic your application and copy services processes generate.

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 cluster and storage configuration, and the ICL.

7.1.1 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 cluster or master cluster
The cluster 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.
- ▶ Intercluster link
The remote inter-switch link (ISL) between the local and remote clusters. It must be redundant and provide dedicated bandwidth for remote copy processes.
- ▶ Remote cluster or auxiliary cluster
The cluster 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.

Performance loss: A performance loss in the foreground write I/O is a result of ICL latency.

► 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 cluster. Mirrored foreground writes are processed asynchronously at the remote cluster, but in a committed sequential order as determined and managed by the Global Mirror remote copy process.

Performance loss: Performance loss in the foreground write I/O is minimized by adopting an asynchronous policy to run a mirrored foreground write I/O. The effect of ICL latency is reduced. However, a small increase occurs in processing foreground write I/O because it passes through the remote copy component of the SAN Volume Controllers software stack.

► Global Mirror Change Volume

Holds earlier consistent revisions of data when changes are made. A change volume can be created for the master volume and the auxiliary volume of the relationship.

Figure 7-1 shows some of the concepts of remote copy.

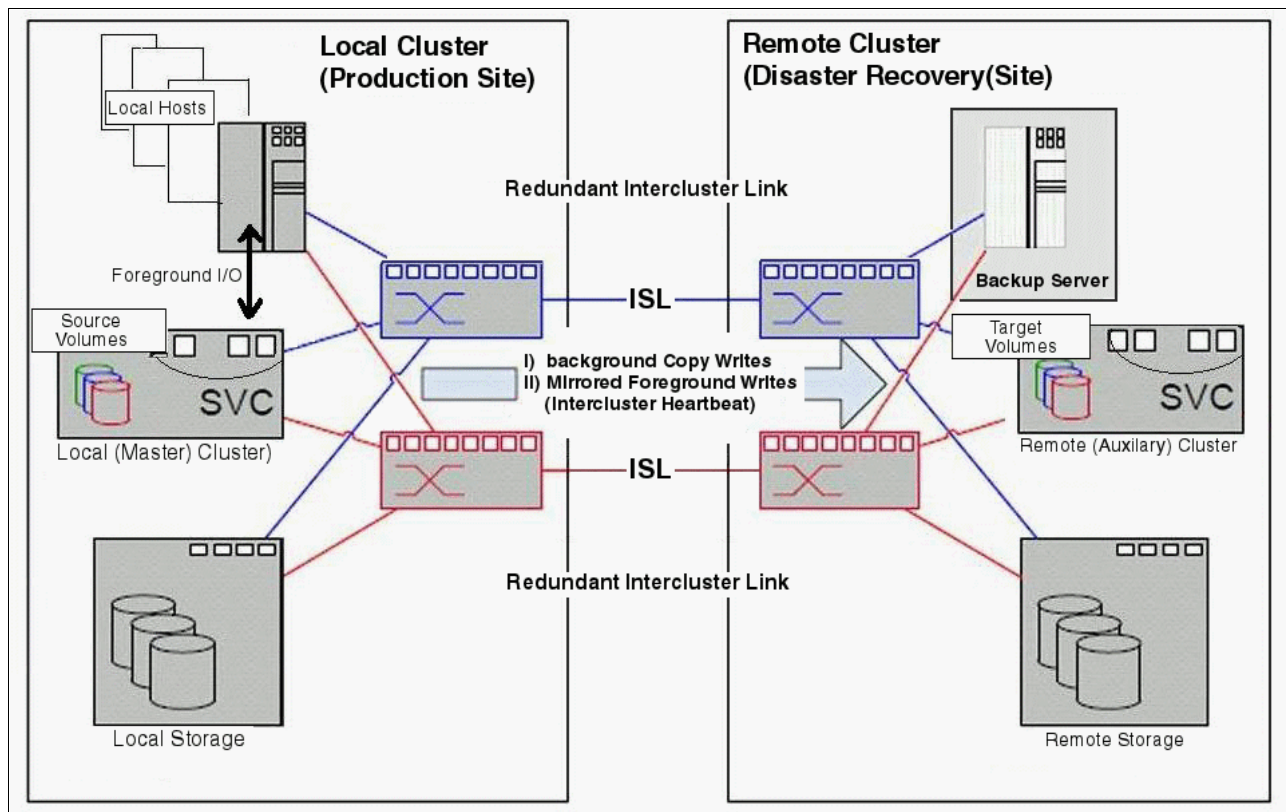


Figure 7-1 Remote copy components and applications

A successful implementation of an intercluster remote copy service depends on quality and configuration of the ICL (ISL). The ICL must provide a dedicated bandwidth for remote copy traffic.

7.1.2 Intercluster link

The ICL 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 intercluster 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.

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.
- ▶ Global Mirror Partnership Bandwidth (parameter): The rate at which background write synchronization is attempted (unit of MBps).

Intercluster communication 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 clusters. A *dedicated amount* of the link bandwidth is required for the exchange of heartbeat messages and the initial configuration of intercluster 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
- ▶ Intercluster communication (*heartbeat messaging*)

7.2 SAN Volume Controller remote copy functions by release

This section describes the new remote copy functions in SAN Volume Controller V7.2 and then in SAN Volume Controller by release.

7.2.1 Remote copy in SAN Volume Controller V7.2

SAN Volume Controller V7.2 has several enhancements for remote copy.

IP Replication

Before V7.2, remote copy services between remote SAN Volume Controller/Storwize storage systems had to use FC network connections. By using V7.2 provides, users can configure remote replication by using a 1 Gbit Ethernet connection without FCIP routers. That is, the SAN Volume Controller/Storwize V7000 now offers native IP Replication.

This feature supports all remote copy modes with the normal remote copy license. IP replication in V7.2 virtualization software includes Bridgeworks SANSlide network optimization technology, which enhances the parallelism of data transfer by using multiple virtual connections (VC) and by that improves WAN connection use. These virtual connections share the same IP link and addresses and send more packets across other virtual connections. For more information about SANSlide technology, see *IBM Storwize V7000 and SANSlide Implementation*, RED-5023, which is available at this website:

<http://www.redbooks.ibm.com/redpapers/pdfs/redp5023.pdf>

Remote Copy Port Group

Port Group is a set of local and remote Ethernet ports with associated TCP/IP addresses to establish a session over the IP link. A maximum of two links can be connected between the remote and local sites. A path is established between the local and remote ports when the IP partnership is first configured. To create a port group, a minimum of one local and one remote port is required.

The remote copy port group can be configured in the following options:

- ▶ Single physical link with active/standby ports
- ▶ Dual physical links with all ports active and no standby
- ▶ Dual physical links with active/standby, for two or more I/O groups environments

Single physical link with active/standby ports

With this configuration, two IP addresses must be configured for each I/O group. If H1 fails, a new session with H2 is established with the IP address of H2 and M1 or M2.

Figure 7-2 shows a single physical link with active/standby ports when there is only one I/O group in each system.

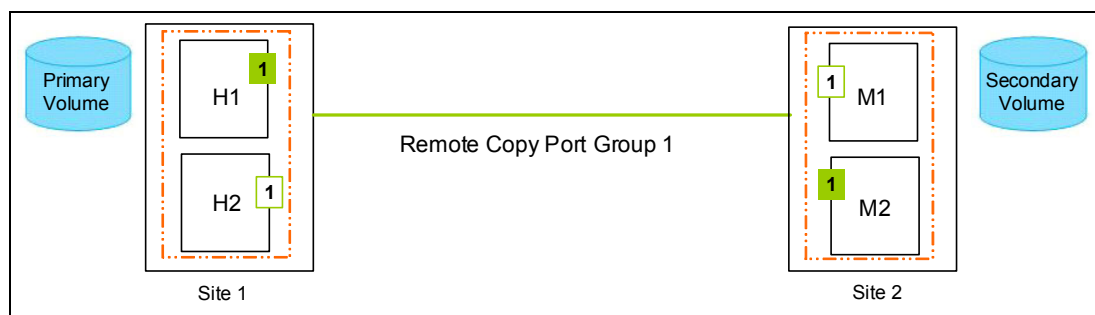


Figure 7-2 Single physical link

Figure 7-3 on page 163 shows single physical link in two I/O groups system. With this configuration, remote copy port group on each system includes four IP addresses. If H1 node fails, the session between H1 and M2 fails and the system automatically establishes another session between H2, H3, or H4 and M1, M2, M3, or M4.

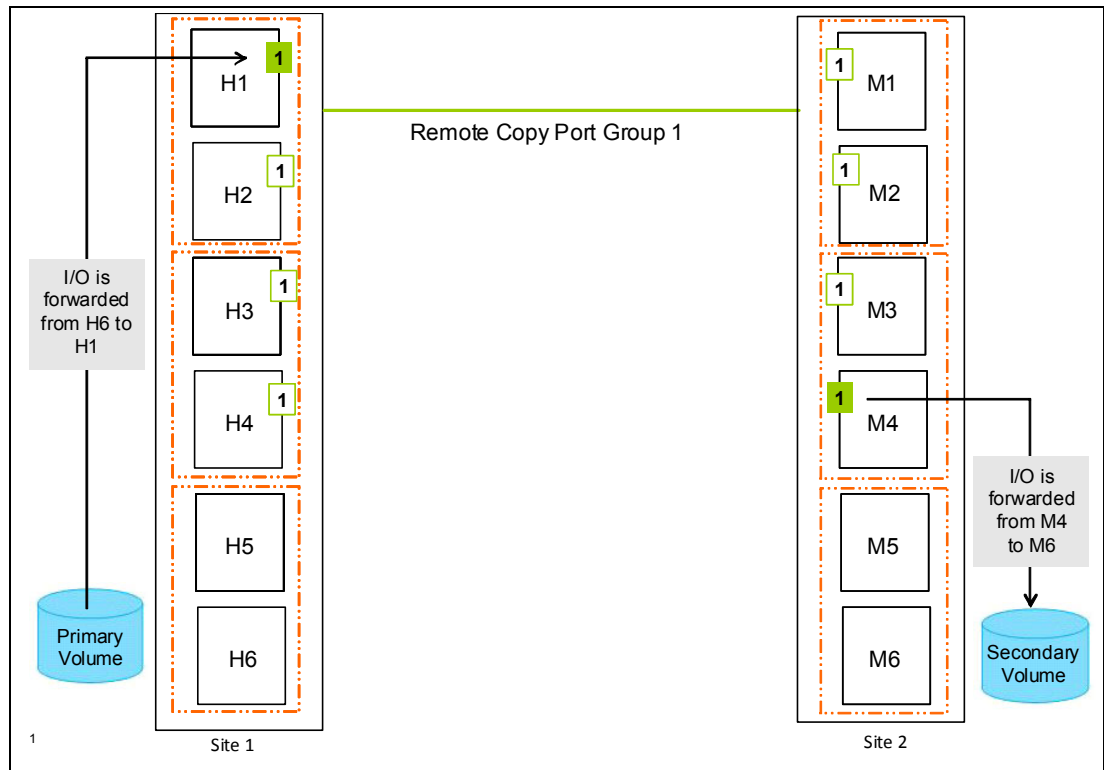


Figure 7-3 Single physical link with two I/O groups

Note: For systems with more than two I/O groups, there is maximum of four IP addresses than can be configured; therefore, only four nodes can participate in the remote copy port group.

Dual physical links with all ports active and no standby ports

With this configuration, there is no redundancy in case of node failure and only half of the bandwidth is available for replication. If there is a node failure, Global Mirror or Metro Mirror replication does not operate properly.

Figure 7-4 shows dual physical links with all ports active.

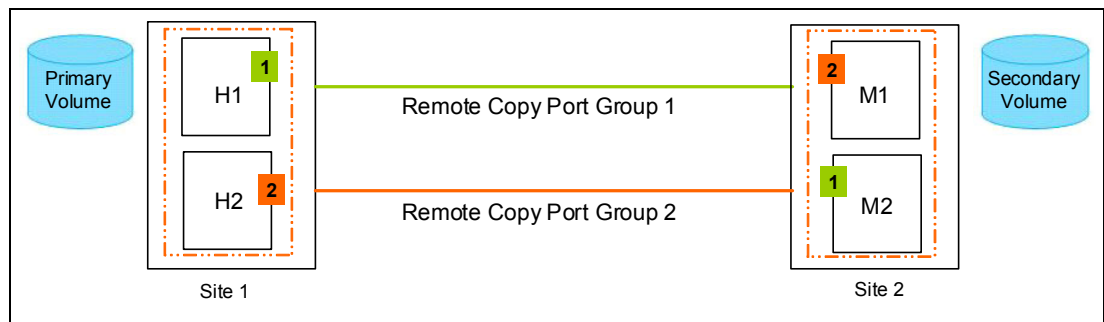


Figure 7-4 Dual physical links with all ports active

Note: It is recommended to cover node failure situations. In single I/O groups, use two ports of the local node and two ports of the remote node in the same remote copy port group, as shown in Figure 7-4 on page 163.

7.2.2 Dual physical links with active/standby for use in two or more I/O groups environments

Each remote copy port group on each system includes two IP addresses. When the port group is initially configured, the system establishes the pairings that are used.

If H1 node fails, the session between H1 and M2 fails and the system automatically establishes another session between H3 and M2 or M4 because they are all in the same remote copy port group with H1.

Figure 7-5 shows dual physical links with active/standby.

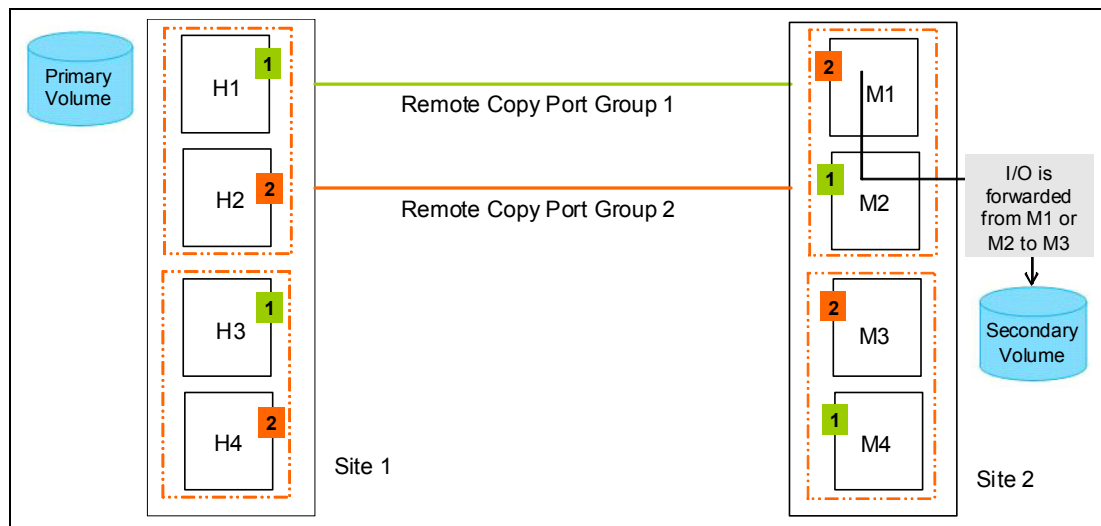


Figure 7-5 Dual physical links with active/standby

Differences between IP and FC/FCoE remote copy partnership

IP and FC/FCoE remote copy partnership includes the following differences:

- ▶ FC/FCoE remote clusters are discovered automatically via Fabric Name Service while IP remote clusters are discovered by using the **mkippartnership** command.
- ▶ FC/FCoE partnerships establish all-to-all active paths (limited only by zoning), while IP partnerships establish active/standby paths in one or two remote copy port groups.

Note: A system can be in an IP partnership with one other system and in an FC partnership with others. For example, create partnership A ∅ B that uses IP replication and B ∅ C ∅ D that uses FC replication.

Configuration recommendations for IP replication

Consider the following configuration recommendations for IP replication:

- ▶ Ports 3260 and 3265 are used by IP Replication. Port 3260 is the port that is used by the systems to initially discover each other. Port 3265 is for the actual IP Replication sessions

that are used to transmit data. Make sure that these ports are opened in the Firewall to configure IP replication.

- ▶ Do not mix iSCSI host I/O and IP partnership traffic. The recommendation is to use different ports for each.
- ▶ IP replication might use CPU resources. The first compressed volume is enabled in an I/O group, which results in the reallocation of CPU cores. If you choose to create an IP partnership on a Storwize V7000 system that has compressed volumes and the expected throughput is more than 100 MBps in the inter-site link, it is recommended to configure ports for the IP partnership in I/O groups that do not contain compressed volumes. For more information about the CPU allocation when compression is used, see Chapter 17, “IBM Real-time Compression” on page 593.

IP replication limitations

IP Replication features the following limitations:

- ▶ 1 Gb and 10 Gb cannot be mixed in the same port group.
- ▶ If IPv6 is used for IP replication, the management IPs on both systems should have IPv6 addresses that have connectivity with each other
- ▶ If IPv4 is used for IP replication, the management IPs on both systems should have IPv4 addresses that have connectivity with each other
- ▶ You can have only two direct attach links (not less or more) and both need to be on the same I/O group
- ▶ NAT (Network Address Translation) between systems that are being configured in an IP Partnership group is not supported.

Optimized Global Mirror processing in second site

In version 7.2, the Global Mirror processing is optimized to allow more parallelism. This improvement delivers almost four times the number of writes that are sustained on secondary system volumes that are used in regular Global Mirror relationship.

Note: Before you upgrade to V7.2, all Global Mirror relationships must be stopped. Metro Mirror and Global Mirror with Change Volume relationships can be left running.

7.2.3 Remote copy features by release

SAN Volume Controller added various remote copy features for Global Mirror and Metro Mirror by code release.

Global Mirror has the following features by release:

- ▶ Release V4.1.1: Initial release of Global Mirror (asynchronous remote copy)
- ▶ Release V4.2 changes and the following features are added:
 - Increased size of nonvolatile bitmap space and can be copied to the virtual disk (VDisk) space to 16 TB
 - Allowance for 40 TB of remote copy per I/O group
- ▶ Release V5.1: Introduced Multicluster Mirroring
- ▶ Release V6.2: Allowance for a Metro Mirror or Global Mirror disk to be a FlashCopy target

Metro Mirror has the following features by release:

- ▶ Release V1.1: Initial release of remote copy
- ▶ Release V2.1: Initial release as Metro Mirror
- ▶ Release V4.1.1: Changed algorithms to maintain synchronization through error recovery to use the same nonvolatile journal as Global Mirror
- ▶ Release V4.2:
 - Increased the size of nonvolatile bitmap space and can be copied to the VDisk space to 16 TB
 - Allowance for 40 TB of remote copy per I/O group
- ▶ Release 5.1: Introduced multiple cluster mirroring, as described in “Multiple cluster mirroring” on page 166
- ▶ Release 6.2: Allowance for a Metro Mirror or Global Mirror disk to be a FlashCopy target
- ▶ Release 6.3:
 - Global Mirror using Change Volume with configurable Recovery Point Objective (RPO).
 - New cluster property that is called Layer. For more information about the Layers concept, see Chapter 3, Section 3 of *IBM System Storage SAN Volume Controller and Storwize V7000 Replication Family Services*, SG24-7574, which is available at this website:
<http://www.redbooks.ibm.com/redbooks/pdfs/sg247574.pdf>
- ▶ Release 6.4: Allowance of Storwize V7000 in Replication layer to virtualize a Storwize V7000 in Storage layer. For more information, see “Replication between different layers” on page 169.
- ▶ Release 7.1: Support remote copy with V3700. Therefore, remote copy is supported by every combination of SAN Volume Controller, Storwize V7000, FlexSystem, and Storwize V3700.
- ▶ Release 7.2:
 - IP replication
 - Optimized Global Mirror processing in second site.

Multiple cluster mirroring

Multiple cluster mirroring enables Metro Mirror and Global Mirror partnerships up to a maximum of four SAN Volume Controller clusters. The rules that govern a Metro Mirror and Global Mirror relationships remain unchanged. That is, a volume can exist only as part of a single Metro Mirror and Global Mirror relationship, and Metro Mirror and Global Mirror are supported within the same overall configuration.

An advantage to multiple cluster mirroring is that customers can use a single disaster recovery site from multiple production data sites to help in the following situations:

- ▶ Implementing a consolidated disaster recovery strategy
- ▶ Moving to a consolidated disaster recovery strategy

Figure 7-6 shows the supported and unsupported configurations for multiple cluster mirroring.

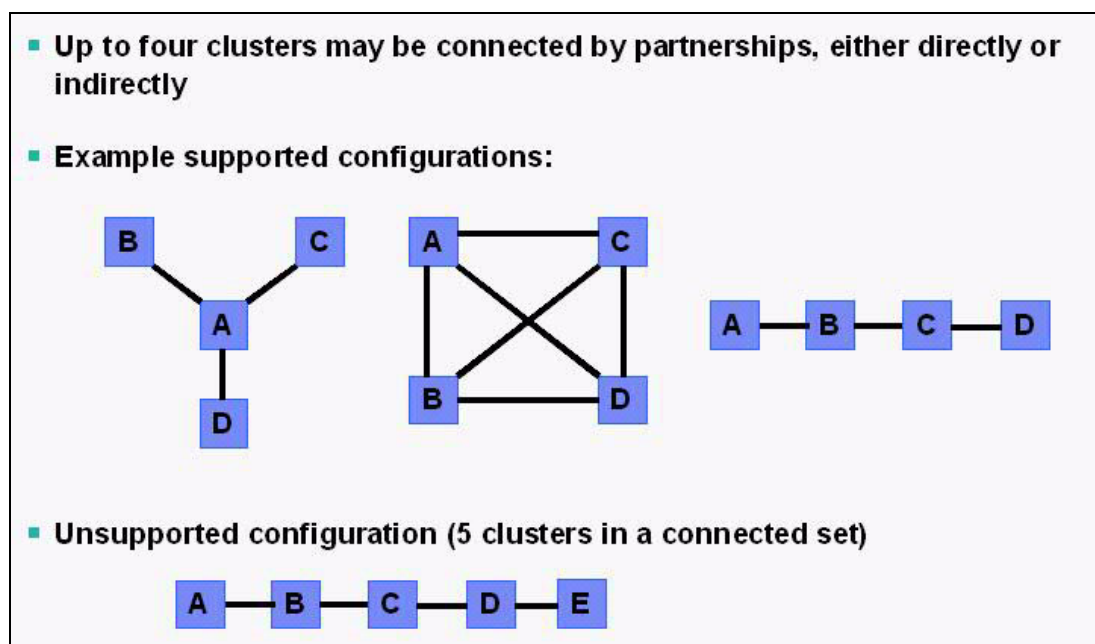


Figure 7-6 Supported multiple cluster mirroring topologies

Improved support for Metro Mirror and Global Mirror relationships and consistency groups

With SAN Volume Controller V5.1, the number of Metro Mirror and Global Mirror remote copy relationships that can be supported increases from 1024 to 8192. This increase provides improved scalability regarding increased data protection, and greater flexibility so that you can use fully the new multiple Cluster Mirroring possibilities.

Consistency groups: You can create up to 256 consistency groups, and all 8192 relationships can be in a single consistency group, if required.

Zoning considerations

The zoning requirements were revised, as described in 7.4, “Intercluster link” on page 181. For more information, see *Nodes in Metro or Global Mirror Inter-cluster Partnerships May Reboot if the Inter-cluster Link Becomes Overloaded*, S1003634, which is available at this website:

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

FlashCopy target volumes as remote copy source volumes

Before the release of SAN Volume Controller V6.1, a Metro Mirror and Global Mirror source volume cannot be part of a FlashCopy relationship. Conceptually, a configuration of this type is advantageous because, it can reduce the time in some disaster recovery scenarios in which the Metro Mirror and Global Mirror relationship is in an inconsistent state.

FlashCopy target volume as remote copy source scenario

A Global Mirror relationship exists between a source volume A and a target volume B. When this relationship is in a consistent-synchronized state, an incremental FlashCopy is taken that provides a point-in-time record of consistency. A FlashCopy of this nature can be made regularly, as shown in Figure 7-7.

Incremental FlashCopy: An incremental FlashCopy is used in this scenario because after the initial instances of FlashCopy are successfully started, all subsequent executions do not require a full background copy. The incremental parameter means that only the areas of disk space where data changed since the FlashCopy mapping was completed are copied to the target volume, which speeds up FlashCopy completion.

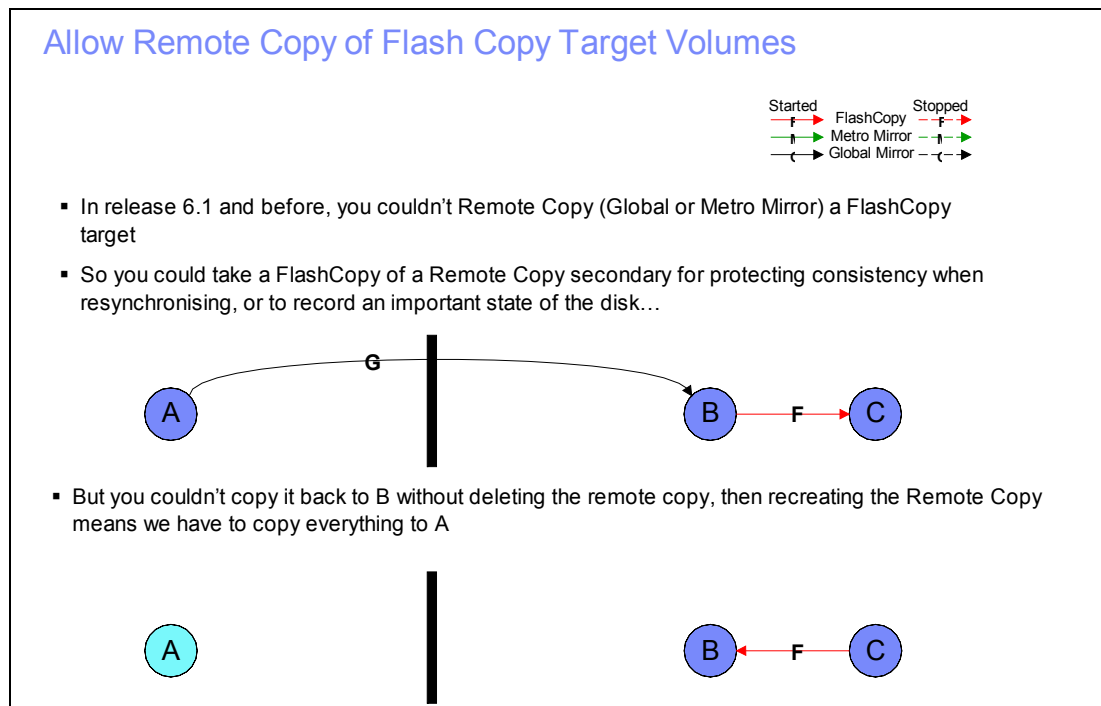


Figure 7-7 Remote copy of FlashCopy target volumes

If corruption occurs on source volume A or the relationship stops and becomes inconsistent, you might want to recover from the last incremental FlashCopy that was taken. Unfortunately, recovering SAN Volume Controller versions before 6.2 means destroying the Metro Mirror and Global Mirror relationship. In this case, the remote copy does not need to be running when a FlashCopy process changes the state of the volume. If both processes were running concurrently, a volume might be subject to simultaneous data changes.

Destruction of the Metro Mirror and Global Mirror relationship means that a complete background copy is required before the relationship is again in a consistent-synchronized state. In this case, the host applications are unprotected for an *extended* period.

With the release of 6.2, the relationship does not need to be destroyed, and a consistent-synchronized state can be achieved more quickly. That is, host applications are unprotected for a *reduced* period.

Remote copy: SAN Volume Controller supports the ability to make a FlashCopy copy away from a Metro Mirror or Global Mirror source or target volume. That is, volumes in remote copy relationships can act as source volumes of FlashCopy relationship.

However, when you prepare a FlashCopy mapping, the SAN Volume Controller puts the source volumes in a temporary cache-disabled state. This temporary state adds latency to the remote copy relationship. I/Os that are normally committed to the SAN Volume Controller cache must now be directly committed as destaged to the back-end storage controller.

FlashCopy Dependency Chain

When a point-in-time copy of a point-in-time copy is taken, a dependency chain of FlashCopy targets is created, which is called a *dependency chain*. A dependency chain approach is used because the number of concurrent point-in-time copies for a single source can (in theory) be unbounded. Unlike the traditional approach, the dependency chain approach a single host I/O to the source volume results in a magnification of one read and two writes to the back-end storage, regardless of the number point-in-time copies. The drawback of the use of dependency chains is that target volumes are now dependent on each other and it causes the following problems:

- ▶ A single user write I/O to the storage ends up in two write operations to the back-end storage.
- ▶ To remove a target volume from a dependency chain, you must copy all of the data off the target that is removed to the first target on which it is dependent.

For example, consider the dependency chain that is shown in Figure 7-8. To remove target volume C, we must copy any data on C that is not on B to B. In this case, there are two grains of data on C (at grain 4 and 5). B has its own copy of data at grain 5. Therefore, we must clean the data at grain 4 on volume C to volume B, which is referred to as *cleaning*.

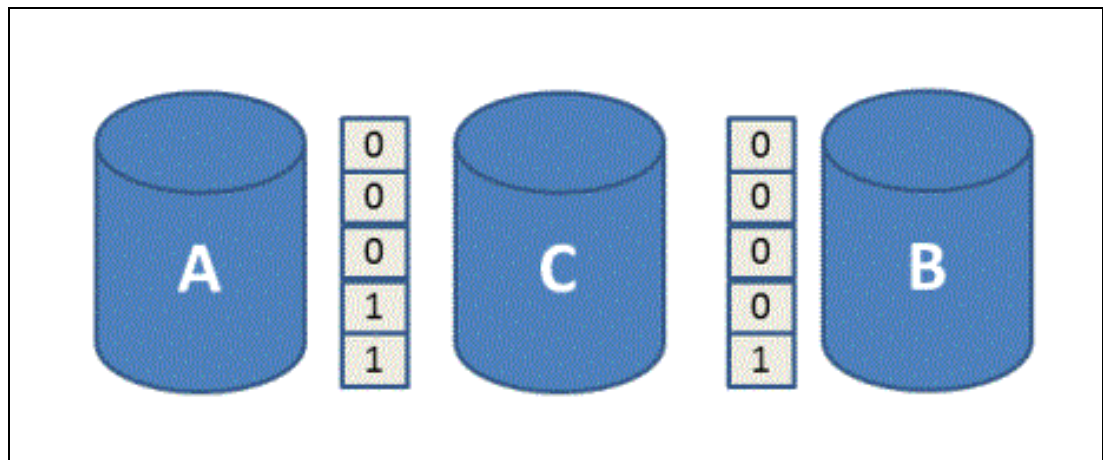


Figure 7-8 Dependency chain

Replication between different layers

Layers concept was first introduced in version 6.3. It distinguishes between storage layers, by default, Storwize V7000, and replication layer, SAN Volume Controller. Since version 6.4, it is possible to do remote copy between SAN Volume Controller and Storwize V7000, which are in different 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 V7000, the system must meet the following prerequisites:

- ▶ The Storwize V7000 must not have any defined host objects and must not be presenting any volumes to an SAN Volume Controller as MDisks.
- ▶ The Storwize V7000 must not be visible to any other SAN Volume Controller or Storwize V7000 in the SAN fabric (this might require SAN zoning changes).

Changing a Storwize V7000 from Storage layer to Replication layer can be performed only by using the CLI. After you confirm that these prerequisites are met, run the following command:

```
chsystem -layer replication
```

Figure 7-9 shows an example for possible replication. SAN Volume Controller uses a V7000 as a backend storage controller and replicates to a different V7000 (SAN Volume Controller = replication, backend V7000 = storage, remote V7000 = replication).

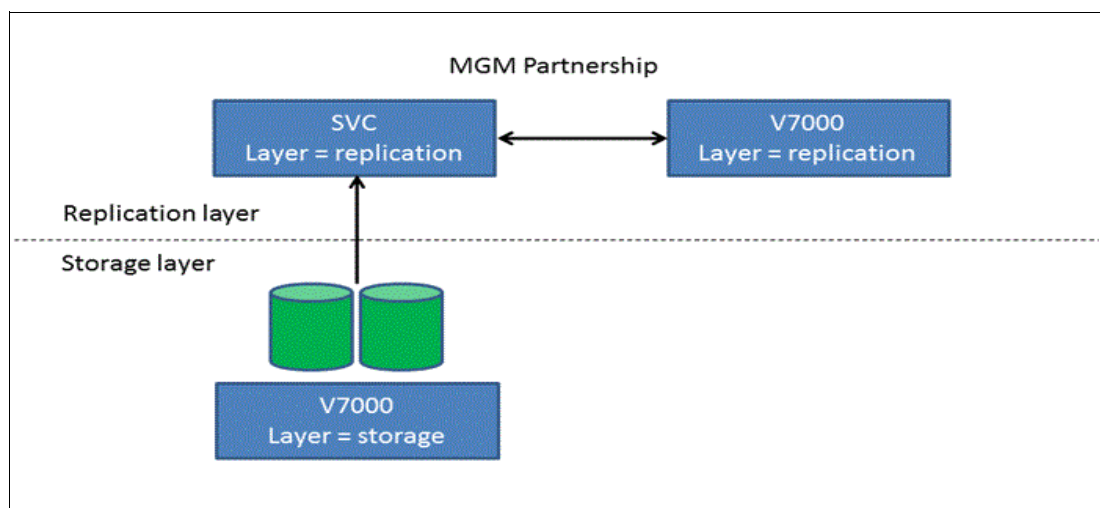


Figure 7-9 Replication example

Figure 7-10 on page 171 shows an example of replication between two Storwize V7000 system in replication layer and Storwize V3700 is the backend storage.

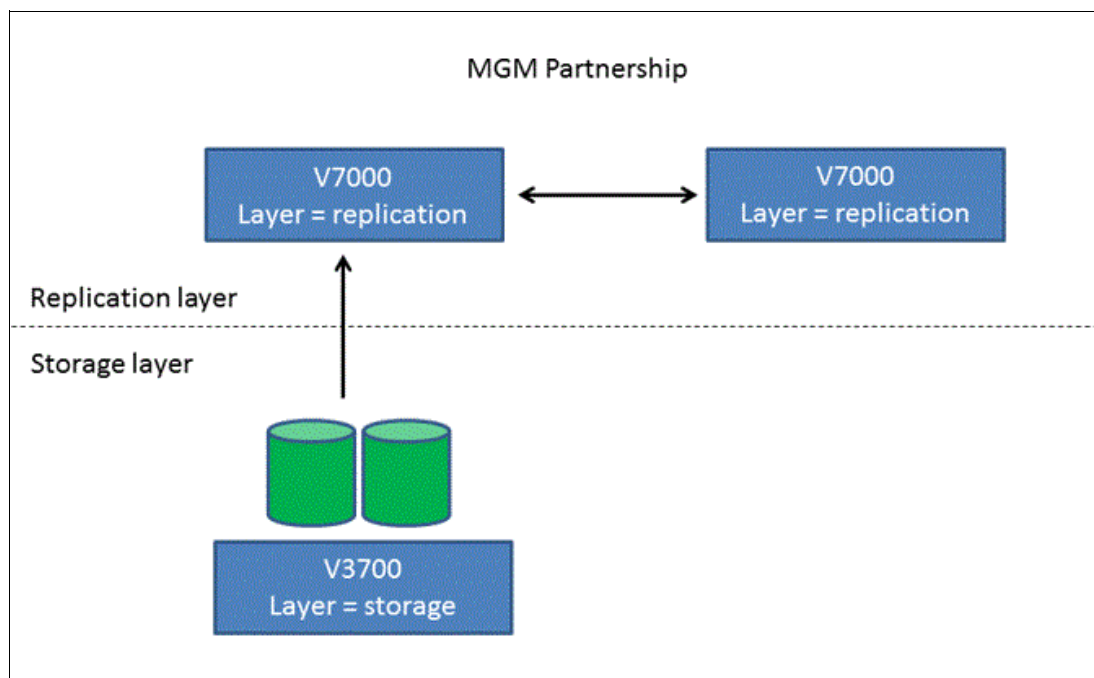


Figure 7-10 Replication example that uses Storwize V3700

Figure 7-11 shows an example for replication between two Storwize V7000 when both of them are in the Storage layer.

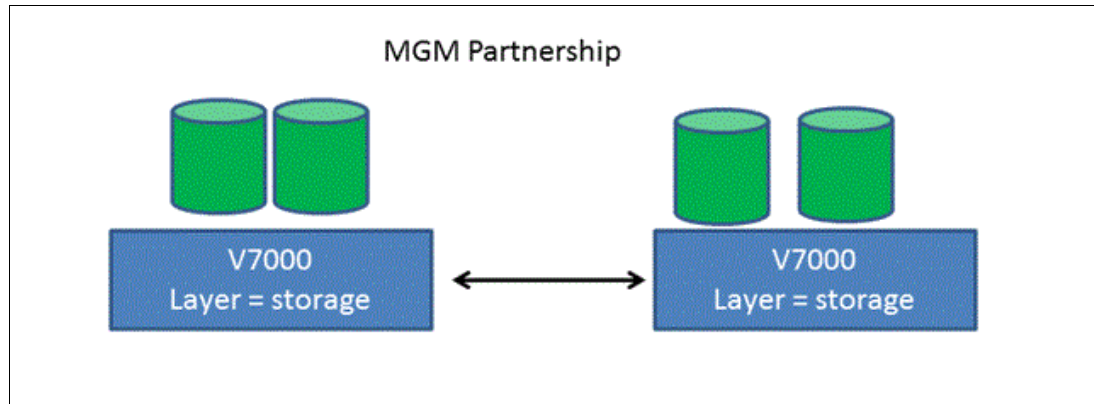


Figure 7-11 Replication example that uses Storwize V7000 at the storage layer

7.3 Terminology and functional concepts

As presented in this section, the functional concepts define how SAN Volume Controller implements remote copy. In addition, the terminology that is presented describes and controls the functionality of SAN Volume Controller. These terms and concepts build on the definitions that were outlined in 7.1.1, “Common terminology and definitions” on page 159 and introduce information about specified limits and default values.

For more information about setting up remote copy partnerships and relationships or about administering remote copy relationships, see *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933.

7.3.1 Remote copy partnerships and relationships

A remote copy partnership is made between a local and remote cluster by using the **mkpartnership** command. This command defines the operational characteristics of the partnership. You must consider the following two following important parameters of this command:

- ▶ **Bandwidth:** The rate at which background write synchronization or resynchronization is attempted.
- ▶ **gminktolerance:** The amount of time, in seconds, that a Global Mirror partnership tolerates poor performance of the ICL before adversely affecting the foreground write I/O.

Mirrored foreground writes: Although mirrored foreground writes are performed asynchronously, they are inter-related at a Global Mirror process level with foreground write I/O. Slow responses along the ICL can lead to a backlog of Global Mirror process events, or an inability to secure process resource on remote nodes. In turn, the ability of Global Mirror to process foreground writes is delayed; therefore, it causes slower writes at the application level.

The following features further define the **bandwidth** and **gminktolerance** parameters that are used with Global Mirror:

- ▶ **relationship_bandwidth_limit:** The maximum resynchronization limit, at relationship level.
- ▶ **gm_max_hostdelay:** The maximum acceptable delay of host I/O that is attributable to Global Mirror.

7.3.2 Global Mirror control parameters

The following parameters control the Global Mirror processes:

- ▶ **bandwidth**
- ▶ **relationship_bandwidth_limit**
- ▶ **gminktolerance**
- ▶ **gm_max_hostdelay**

The Global Mirror partnership **bandwidth** parameter specifies the rate (in MBps) at which the background write resynchronization processes are attempted. That is, it specifies the total bandwidth that the processes use.

With SAN Volume Controller V5.1.0, the granularity of control at a volume relationship level for Background Write Resynchronization also can be modified by using the **relationship_bandwidth_limit** parameter. Unlike its co-parameter, this parameter has a default value of 25 MBps. The parameter defines at a cluster-wide level the maximum rate at which background write resynchronization of an individual source-to-target volume is attempted. Background write resynchronization is attempted at the lowest level of the combination of these two parameters.

Background write resynchronization: The term *background write resynchronization*, when used with SAN Volume Controller, is also referred to as *Global Mirror Background copy* in this book and in other IBM publications.

Although asynchronous Global Mirror adds overhead to foreground write I/O, it requires a dedicated portion of the interlink bandwidth to function. Controlling this overhead is critical to foreground write I/O performance and is achieved by using the **gmlinktolerance** parameter. This parameter defines the amount of time that Global Mirror processes can run on a poorly performing link without adversely affecting foreground write I/O. By setting the **gmlinktolerance** time limit parameter, you define a safety valve that suspends Global Mirror processes so that foreground application write activity continues at acceptable performance levels.

When you create a Global Mirror Partnership, the default limit of 300 seconds (5 minutes) is used, but you can adjust this limit. The parameter can also be set to 0, which effectively turns off the safety valve, meaning that a poorly performing link might adversely affect foreground write I/O.

The **gmlinktolerance** parameter does not define what constitutes a poorly performing link. It also does not explicitly define the latency that is acceptable for host applications.

With the release of V5.1.0, you define what constitutes a poorly performing link by using the **gmmxhostdelay** parameter. With this parameter, you can specify the maximum allowable overhead increase in processing foreground write I/O (in milliseconds) that is attributed to the effect of running Global Mirror processes. This threshold value defines the maximum allowable impact that Global Mirror operations can add to the response times of foreground writes on Global Mirror source volumes. You can use the parameter to increase the threshold limit from its default value of 5 milliseconds. If this threshold limit is exceeded, the link is considered to be performing poorly, and the **gmlinktolerance** parameter becomes a factor. The Global Mirror link tolerance timer starts counting down.

7.3.3 Global Mirror partnerships and relationships

A *Global Mirror partnership* is a partnership that is established between a master (local) cluster and an auxiliary (remote) cluster, as shown in Figure 7-12.

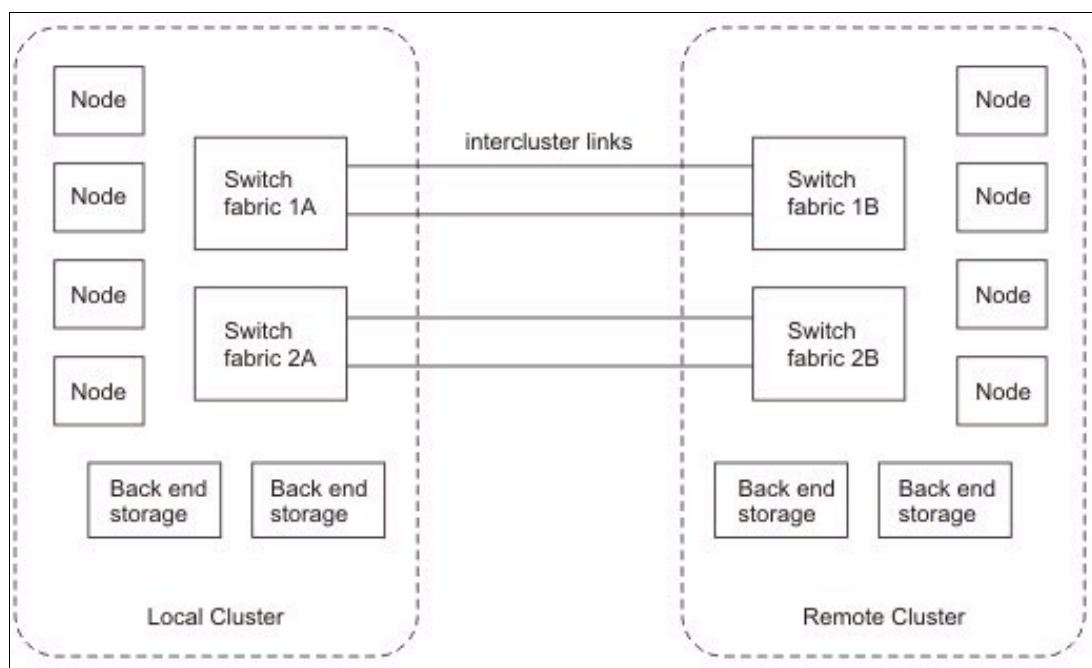


Figure 7-12 Global Mirror partnership

The mkpartnership command

The **mkpartnership** command establishes a one-way Metro Mirror or Global Mirror relationship between the local cluster and a remote cluster. When you create a partnership, the client must set a remote copy bandwidth rate (in MBps). This rate specifies the proportion of the total ICL bandwidth that is used for Metro Mirror and Global Mirror background copy operations.

Tip: To establish a fully functional Metro Mirror or Global Mirror partnership, you must issue the **mkpartnership** command from both clusters.

The mkrcrelationship command

When the partnership is established, a Global Mirror relationship can be created between volumes of equal size on the master (local) and auxiliary (remote) clusters.

The volumes on the local cluster are *master volumes* and have an initial role as the “source” volumes.

The volumes on the remote cluster are defined as *auxiliary volumes* and have the initial role as the “target” volumes.

Tips: After the initial synchronization is complete, you can change the copy direction. Also, the role of the master and auxiliary volumes can swap. That is, the source becomes the target.

Like FlashCopy volumes can be maintained as consistency groups.

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. The relationship provides this support without requiring the hosts that are connected to the local cluster to wait for the full round-trip delay of the long-distance ICL. That is, it provides the same function as Metro Mirror remote copy, but over longer distance by using links with a higher latency.

Tip: Global Mirror is an asynchronous remote copy service.

Asynchronous writes: Writes to the target volume are made asynchronously. The host that writes to the source volume provides the host with confirmation that the write is complete before the I/O completes on the target volume.

Intracenter versus intercenter

Although Global Mirror is available for intracenter, it has no functional value for production use. Intracenter 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 cluster.

Intercenter Global Mirror operations require a minimum of a pair of SAN Volume Controller clusters that are connected by several ICLs.

Hop limit: When a local fabric and a remote fabric are connected for Global Mirror purposes, the ISL hop count between a local node and a remote node must not exceed seven hops.

7.3.4 Asynchronous remote copy

Global Mirror is an asynchronous remote copy technique. In asynchronous remote copy, write operations are completed on the primary site and the write acknowledgment is sent to the host before it is received at the secondary site. An update of this write operation is sent to the secondary site at a later stage. The update can perform a remote copy over distances that exceed the limitations of a synchronous remote copy.

7.3.5 Understanding remote copy write operations

This section describes the remote copy write operations concept.

Normal I/O writes

Schematically, you can consider SAN Volume Controller as several software components that are arranged in a software stack. I/Os pass through each component of the stack. The first three components define how SAN Volume Controller processes I/O regarding the following areas:

- SCSI target and how the SAN Volume Controller volume is presented to the host
- Remote copy and how remote copy processes affect I/O (including Global Mirror and Metro Mirror functions)
- Cache and how I/O is cached

Host I/O to and from volumes that are not in Metro Mirror and Global Mirror relationships pass transparently through the remote copy component layer of the software stack, as shown in Figure 7-13.

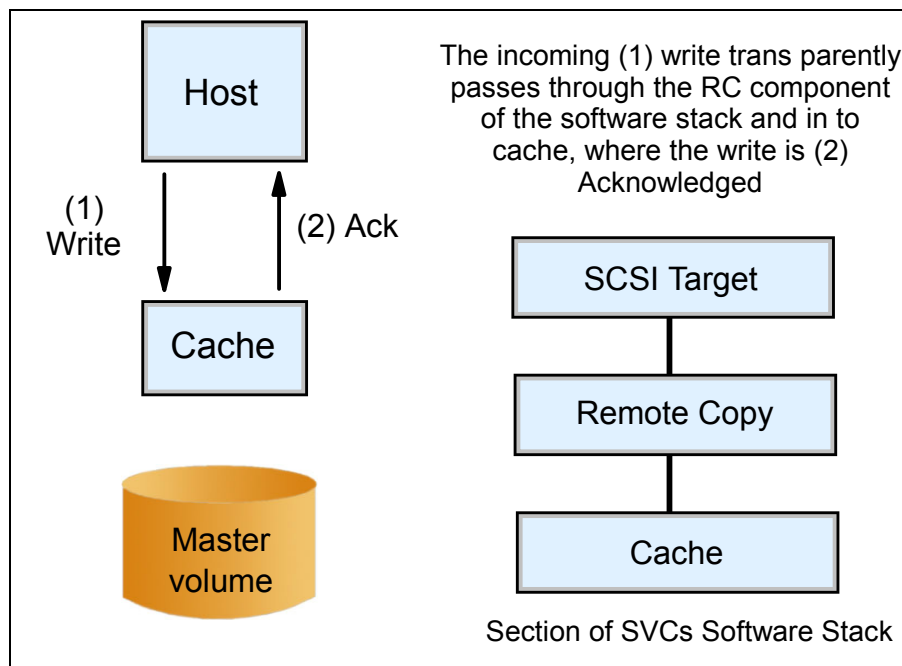


Figure 7-13 Write I/O to volumes that are not in remote copy relationships

7.3.6 Asynchronous remote copy

Although Global Mirror is an asynchronous remote copy technique, foreground writes at the local cluster and mirrored foreground writes at the remote cluster are not wholly independent of one another. SAN Volume Controller 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 cluster, and therefore, is not subject to the latency of the long-distance link.

Point-in-time consistency: A consistent image is defined as *point-in-time consistency*.

Figure 7-14 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.

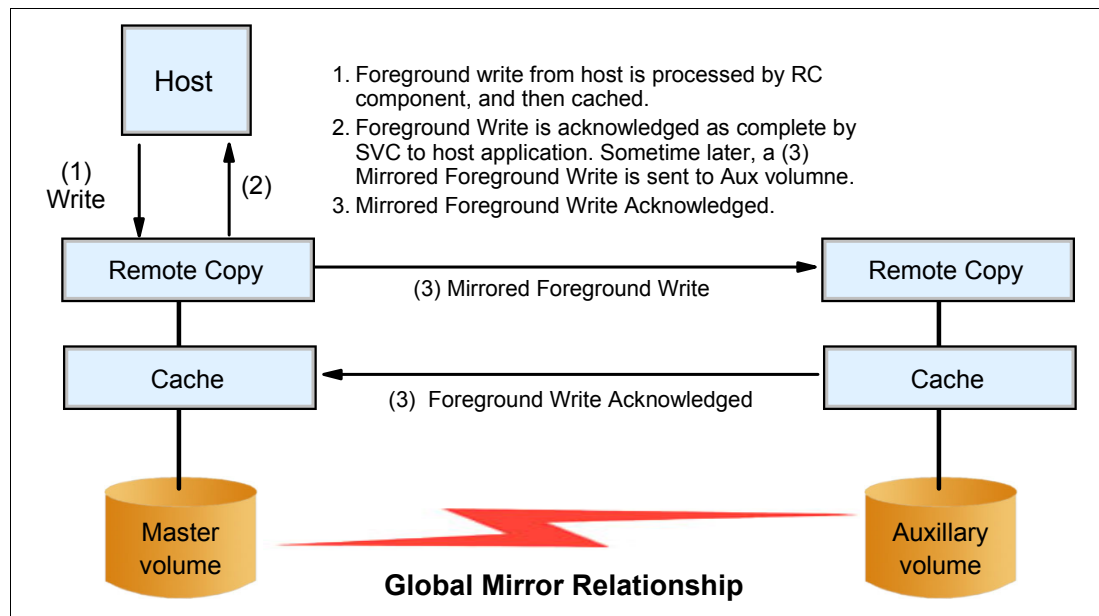


Figure 7-14 Global Mirror relationship write operation

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 VDisk, 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.

7.3.7 Global Mirror write sequence

The Global Mirror algorithms maintain a consistent image on the auxiliary always. To achieve this consistent image, the following tasks must be completed:

- ▶ They identify the sets of I/Os that are active concurrently at the master.
- ▶ They assign an order to those sets
- ▶ They apply those sets of I/Os in the assigned order at the secondary.

As a result, Global Mirror maintains the features of write ordering and read stability.

The multiple I/Os within a single set are applied concurrently. The process that marshals the sequential sets of I/Os operates at the secondary cluster, and therefore, is not subject to the latency of the long-distance link. These two elements of the protocol ensure that the throughput of the total cluster can be grown by increasing the cluster size and maintaining consistency across a growing data set.

In a failover scenario where the secondary site must become the master source of data, certain updates might be missing at the secondary site. Therefore, any applications that use this data must have an external mechanism, such as a transaction log replay, to recover the missing updates and to reapply them.

7.3.8 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.

7.3.9 Colliding writes

Colliding writes are defined as new write I/Os that overlap existing “active” write I/Os.

Before SAN Volume Controller 4.3.1, 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 SAN Volume Controller 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 7-15 shows a colliding write sequence.

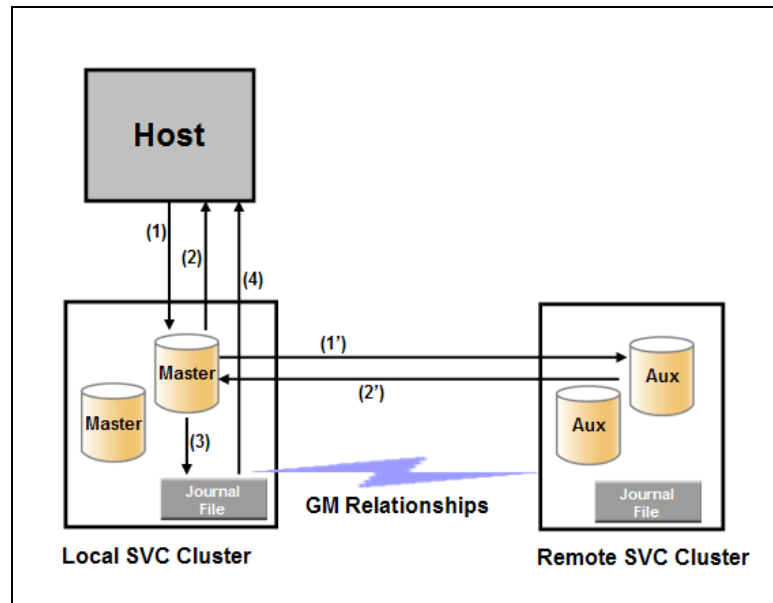


Figure 7-15 Colliding writes

The following numbers correspond to the numbers that are shown in Figure 7-15:

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.

7.3.10 Link speed, latency, and bandwidth

The concepts of link speed, latency, and bandwidth are described in this section.

Link speed

The speed of a communication link (*link speed*) determines how much data can be transported and how long the transmission takes. The faster the link is, the more data can be transferred within an amount of time.

Latency

Latency is the time that is taken by data to move across a network from one location to another location and is measured in milliseconds. The longer the time is, the greater the performance impact. Latency depends on the speed of light ($c = 3 \times 10^8 \text{ m/s}$, vacuum = 3.3 microsec/km (microsec represents microseconds, which is one millionth of a second)). The bits of data travel at about two-thirds of the speed of light in an optical fiber cable.

However, some latency is added when packets are processed by switches and routers and are then forwarded to their destination. Although the speed of light might seem infinitely fast, latency becomes a noticeable factor over continental and global distances. Distance has a direct relationship with latency. Speed of light propagation dictates about one milliseconds of latency for every 100 km. For some synchronous remote copy solutions, even a few milliseconds of more delay can be unacceptable. Latency is a difficult challenge because bandwidth and spending more money for higher speeds reduces latency.

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.

Bandwidth

Regarding FC networks, bandwidth is the network capacity to move data as measured in millions of bits per second (Mbps) or billions of bits per second (Gbps). In storage terms, bandwidth measures the amount of data that can be sent in a specified amount of time.

Storage applications issue read and write requests to storage devices. These requests are satisfied at a certain speed that is commonly called the *data rate*. Usually disk and tape device data rates are measured in bytes per unit of time and not in bits.

Most modern technology storage device LUNs or volumes can manage sequential sustained data rates in the order of 10 MBps to 80 - 90 MBps. Some manage higher rates.

For example, an application writes to disk at 80 MBps. If you consider a conversion ratio of 1 MB to 10 Mb (which is reasonable because it accounts for protocol overhead), the data rate is 800 Mb.

Always check and make sure that you correctly correlate MBps to Mbps.

Attention: When you set up a Global Mirror partnership, the use of the `mkpartnership` command with the `-bandwidth` parameter *does not* refer to the general bandwidth characteristic of the links between a local and remote cluster. Instead, this parameter refers to the background copy (or write resynchronization) rate, as determined by the client that the ICL can sustain.

7.3.11 Choosing a link cable of supporting Global Mirror applications

The ICL bandwidth is the networking link bandwidth and often is measured and defined in Mbps. For Global Mirror relationships, the link bandwidth must be sufficient to support all intercluster traffic, including the following types of traffic:

- ▶ Background write resynchronization (or background copy)
- ▶ Intercluster node-to-node communication (heartbeat control messages)
- ▶ Mirrored foreground I/O (associated with local host I/O)

Requirements: Adhere to the following requirements:

- ▶ Set the Global Mirror Partnership bandwidth to a value that is less than the sustainable bandwidth of the link between the clusters.

If the Global Mirror Partnership bandwidth parameter is set to a higher value than the link can sustain, the initial background copy process uses all available link bandwidth.

- ▶ Both ICLs, as used in a redundant scenario, must provide the required bandwidth.
- ▶ Starting with SAN Volume Controller V5.1.0, you must set a bandwidth parameter when you create a remote copy partnership.

For more considerations about these rules, see 7.5.1, “Global Mirror parameters” on page 189.

7.3.12 Remote copy volumes: Copy directions and default roles

When you create a Global Mirror 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 7-16 on page 181). The ability to change roles is used to facilitate disaster recovery.

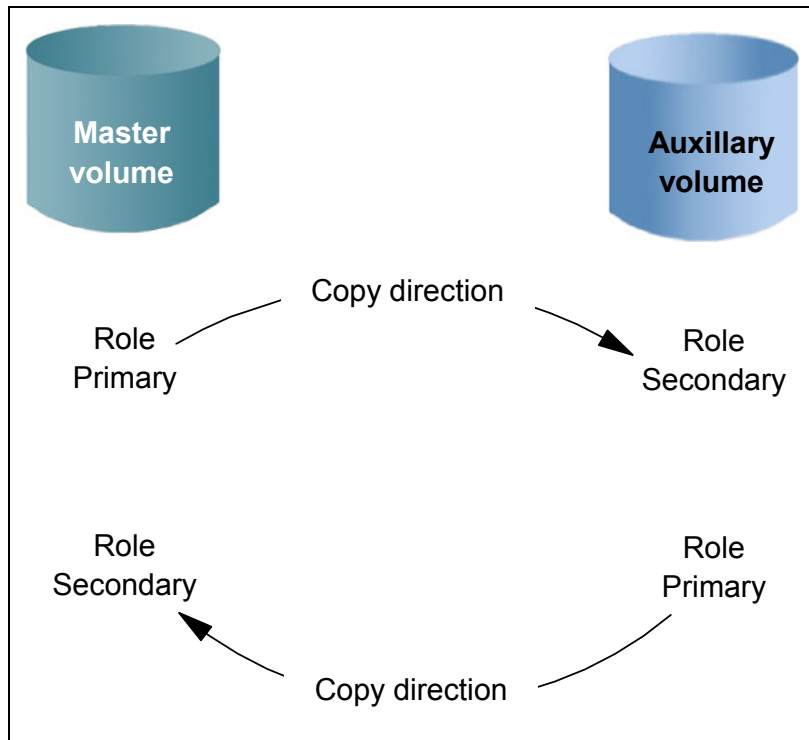


Figure 7-16 Role and direction changes

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.

7.4 Intercluster link

Global Mirror partnerships and relationships do not work reliably if the SAN fabric on which they are running is configured incorrectly. This section focuses on the ICL, which is a part of a SAN that encompasses local and remote clusters, and the critical part ICL plays in the overall quality of the SAN configuration.

7.4.1 SAN configuration overview

You must remember several considerations when you use the ICL in a SAN configuration.

Redundancy

The ICL must adopt the same policy toward redundancy as for the local and remote clusters 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 ICL. Although FC (and the SAN Volume Controller) 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 SAN Volume Controller.

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 (such as the SAN Volume Controller) 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 SAN Volume Controller 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.

7.4.2 Switches and ISL oversubscription

The *IBM System Storage SAN Volume Controller - Software Installation and Configuration Guide*, SC23-6628, specifies a suggested maximum host port to ISL ratio of 7:1. With modern 4 Gbps or 8 Gbps SAN switches, this ratio implies an average bandwidth (in one direction) per host port of approximately 57 MBps (4 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 SAN Volume Controller 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.

7.4.3 Zoning

Zoning requirements were revised, as described in *Nodes in Metro or Global Mirror Inter-cluster Partnerships May Reboot if the Inter-cluster Link Becomes Overloaded*, S1003634, which is available at this website:

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

Although Multicluster Mirroring is supported since SAN Volume Controller V5.1, it increases the potential to zone multiple clusters (nodes) in usable configurations. Therefore, do not use this configuration.

Abstract

SAN Volume Controller nodes in Metro Mirror or Global Mirror intercluster partnerships can experience lease expiry reboot events if an ICL to a partner system becomes overloaded. These reboot events can occur on all nodes simultaneously, which leads to a temporary loss of host access to volumes.

Content

If an ICL becomes severely and abruptly overloaded, the local Fibre Channel fabric can become congested if no FC ports on the local SAN Volume Controller nodes can perform local intracluster heartbeat communication. This situation can result in the nodes that experience lease expiry events, in which a node reboots to attempt to re-establish communication with the other nodes in the system. If all nodes lease expire simultaneously, this situation can lead to a loss of host access to volumes during the reboot events.

Workaround

Default zoning for intercluster Metro Mirror and Global Mirror partnerships now ensures that, if link-induced congestion occurs, only two of the four Fibre Channel ports on each node can be subjected to this congestion. The remaining two ports on each node remain unaffected, and therefore, can continue to perform intracluster heartbeat communication without interruption.

Adhere to the following revised guidelines for zoning:

- ▶ For each node in a clustered system, zone only two Fibre Channel ports to two FC ports from each node in the partner system. That is, for each system, you have two ports on each SAN Volume Controller node that has only local zones (not remote zones).
- ▶ If dual-redundant ISLs are available, split the two ports from each node evenly between the two ISLs. For example, zone one port from each node across each ISL. Local system zoning must continue to follow the standard requirement for all ports, on all nodes, in a clustered system to be zoned to one another.

7.4.4 Distance extensions for the intercluster link

To implement remote mirroring over a distance, 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.

SAN Volume Controller cluster links: Use distance extension only for links between SAN Volume Controller clusters. Do not use it for intracluster links. Technically, distance extension is supported for relatively short distances, such as a few kilometers (or miles). For more information about why this arrangement should not be used, see *IBM System Storage SAN Volume Controller Restrictions*, S1003903.

7.4.5 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.

7.4.6 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*.

7.4.7 Fibre Channel IP conversion

Fibre Channel IP conversion 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 FC IP 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. You do not want a spyware problem or a spam attack on an IP network to disrupt your SAN Volume Controller.

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). If you do not specify *megabytes*, you can end up with an impressive 155 Mbps OC-3 link that supplies only 15 MBps or so to your SAN Volume Controller. With the suggested safety margins included, this link is not fast at all.

7.4.8 Configuration of intercluster links

IBM tested several Fibre Channel extender and SAN router technologies for use with the SAN Volume Controller. For the list of supported SAN routers and FC extenders, see the support page at this website:

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

Link latency considerations

If you use one of the Fibre Channel extenders or routers, you must test the link to ensure that the following requirements are met before you place SAN Volume Controller traffic onto the link:

- ▶ For SAN Volume Controller 4.1.0.x, round-trip latency between sites must not exceed 68 ms (34 ms one way) for FC extenders or 20 ms (10 ms one way) for SAN routers.
- ▶ For SAN Volume Controller 4.1.1.x and later, the round-trip latency between sites must not exceed 80 ms (40 ms one way).

The latency of long-distance links depends on the technology that is used. Typically, for each 100 km (62.1 miles) of distance, 1 ms is added to the latency. For Global Mirror, the remote cluster can be up to 4,000 km (2,485 miles) away.

When you test your link for latency, consider current and future expected workloads, including any times when the workload might be unusually high. You must evaluate the peak workload by considering the average write workload over a period of 1 minute or less, plus the required synchronization copy bandwidth.

Link bandwidth that is used by internode communication

SAN Volume Controller uses part of the bandwidth for its internal SAN Volume Controller intercluster heartbeat. The amount of traffic depends on how many nodes are in each of the local and remote clusters. Table 7-1 shows the amount of traffic (in megabits per second) that is generated by different sizes of clusters.

Table 7-1 SAN Volume Controller intercluster heartbeat traffic (megabits per second)

Local or remote cluster	Two nodes	Four nodes	Six nodes	Eight nodes
Two nodes	2.6	4.0	5.4	6.7
Four nodes	4.0	5.5	7.1	8.6
Six nodes	5.4	7.1	8.8	10.5
Eight nodes	6.7	8.6	10.5	12.4

These numbers represent the total traffic between the two clusters when *no* I/O is occurring to a mirrored volume on the remote cluster. Half of the data is sent by one cluster, and half of the data is sent by the other cluster. The traffic is divided evenly over all available ICLs. 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.

7.4.9 Link quality

The optical properties of the fiber optic cable influence the distance that can be supported. A decrease in signal strength occurs along a fiber optic cable. As the signal travels over the fiber, it is attenuated, which is caused by absorption and scattering and often is expressed in decibels per kilometer (dB/km). Some early deployed fiber supports the telephone network, which is sometimes insufficient for today's new multiplexed environments. If you are supplied dark fiber by a third-party vendor, you normally specify that they must not allow more than a specified loss in total (xdB).

Tip: SCSI write-over Fibre Channel 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 more 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 msec of more service time. At 100 km, it becomes 2 msec of more service time.

The decibel (dB) is a convenient way to express an amount of signal loss or gain within a system or the amount of loss or gain that is caused by a component of a system. When signal power is lost, you never lose a fixed amount of power. The rate at which you lose power is not linear. Instead, you lose a portion of power; that is, one half, one quarter, and so on, which makes it difficult to add up the lost power along a signal's path through the network if you measure signal loss in watts.

For example, a signal loses half of its power through a bad connection. Then, it loses another quarter of its power on a bent cable. You cannot add $\frac{1}{2}$ plus $\frac{1}{4}$ ($\frac{1}{2} + \frac{1}{4}$) to find the total loss. You must *multiply* $\frac{1}{2}$ by $\frac{1}{4}$ ($\frac{1}{2} \times \frac{1}{4}$), which makes calculating large network dB loss time-consuming and difficult. However, decibels are logarithmic so that you can easily calculate the total loss or gain characteristics of a system by adding them up (they scale logarithmically). If your signal gains 3 dB, the signal doubles in power. If your signal loses 3 dB, the signal divides the power into equal parts.

The decibel is a ratio of signal powers. You must have a reference point. For example, you can state, "There is a 5 dB drop over that connection." But you cannot state, "The signal is 5 dB at the connection." A decibel is not a measure of signal strength. Instead, it is a measure of signal power loss or gain.

A *decibel milliwatt* (dBm) is a measure of signal strength. People often confuse dBm with dB. A dBm is the signal power in relation to 1 milliwatt. A signal power of zero dBm is 1 milliwatt, a signal power of 3 dBm is 2 milliwatts, 6 dBm is 4 milliwatts, and so on. The more negative the dBm goes, the closer the power level gets to zero. Do not be misled by the minus signs because they have nothing to do with signal direction.

A good link has a small rate of frame loss. A retransmission occurs when a frame is lost, which directly impacts performance. SAN Volume Controller aims to support retransmissions at 0.2 or 0.1.

7.4.10 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 timeout value and the resource allocation timeout value.

For more information about fabric values, see *IBM TotalStorage: SAN Product, Design, and Optimization Guide*, SG24-6384. If these times become excessive, the fabric experiences undesirable timeouts. It is considered that every extra hop adds about 1.2 microseconds of latency to the transmission.

Currently, SAN Volume Controller 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 SAN Volume Controller.

7.4.11 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.

Tip: To maintain acceptable performance, one buffer credit is required for every 2 km of distance that is covered. 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.

At 1 Gbps, a frame occupies 4 km of fiber. In a 100 km link, you can send 25 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 25 frames before you receive the first ACK.

You need at least 50 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.

A *host bus adapter* (HBA), with a buffer credit of 64 that communicates with a switch port that has only eight buffer credits, can read at full performance over a greater distance than it can write. The reason is that the HBA can send a maximum of only eight buffers to the switch port on the writes; however, the switch can send up to 64 buffers to the HBA on the reads.

7.5 Global Mirror design points

SAN Volume Controller supports the following features of Global Mirror:

- ▶ Asynchronous remote copy of volumes that are dispersed over metropolitan scale distances.
- ▶ Implementation of a Global Mirror relationship between volume pairs.
- ▶ Intracluster Global Mirror, where both volumes belong to the same cluster (and I/O group). However, this function is better suited to Metro Mirror.
- ▶ Intercluster Global Mirror, where each volume belongs to its separate SAN Volume Controller cluster. A SAN Volume Controller cluster can be configured for partnership with 1 - 3 other clusters, which is referred to as *Multicluster Mirroring* (introduced in V5.1).

Attention: Clusters that run on SAN Volume Controller V6.1.0 or later cannot form partnerships with clusters that run on V4.3.1 or earlier.

- ▶ Concurrent usage of intercluster and intracluster Global Mirror within a cluster for separate relationships.
- ▶ No required control network or fabric to be installed to manage Global Mirror. For intercluster Global Mirror, the SAN Volume Controller maintains a control link between the two clusters. This control link controls the state and coordinates the updates at either end. The control link is implemented on top of the same FC fabric connection that the SAN Volume Controller uses for Global Mirror I/O.

ICL bandwidth: Although not separate, this control does require a dedicated portion of ICL bandwidth.

- ▶ A configuration state model that maintains the Global Mirror configuration and state through major events, such as failover, recovery, and resynchronization.
- ▶ Flexible resynchronization support to resynthesized volume pairs that experienced write I/Os to both disks and to resynchronize only those regions that are known to change.
- ▶ Colliding writes.
- ▶ Application of a delay simulation on writes that are sent to auxiliary volumes (optional feature for Global Mirror).
- ▶ Write consistency for remote copy. This way, when the primary VDisk and the secondary VDisk are synchronized, the VDIsks stay synchronized even if a failure occurs in the primary cluster or other failures that cause the results of writes to be uncertain.

7.5.1 Global Mirror parameters

Several commands and parameters help to control remote copy and its default settings. You can display the properties and features of the clusters by using the **svcinfo lsccluster** and **svctask chcluster** commands. Also, you can change the features of clusters by using the **svctask chcluster** command.

The following features are of particular importance regarding Metro Mirror and Global Mirror:

- The Partnership bandwidth parameter (Global Mirror)

This parameter specifies the rate, in MBps, at which the (background copy) write resynchronization process is attempted. From V5.1 onwards, this parameter has *no* default value (previously 50 MBps).

- Optional: The **relationship_bandwidth_limit** parameter

This optional parameter specifies the new background copy bandwidth in the range 1 - 1000 MBps. The default is 25 MBps. This parameter operates cluster-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 VDisk rates attempted.

Important: Do not set this value higher than the default without establishing that the higher bandwidth can be sustained.

- Optional: The **gm_link_tolerance** parameter

This optional parameter specifies the length of time, in seconds, for which an inadequate ICL is tolerated for a Global Mirror operation. The parameter accepts values of 60 - 400 seconds in increments of 10 seconds. The default is 300 seconds. You can disable the link tolerance by entering a value of zero for this parameter.

Important: For later releases, there is no default setting. You must explicitly define this parameter.

- Optional: The **gmmaxhostdelay max_host_delay** parameter

These optional parameters specify the maximum time delay, in milliseconds, above which the Global Mirror link tolerance timer starts counting down. The threshold value determines the impact that Global Mirror operations can add to the response times of the Global Mirror source volumes. You can use these parameters to increase the threshold from the default value of 5 milliseconds.

- Optional: The **gm_inter_cluster_delay_simulation** parameter

This optional parameter specifies the intercluster delay simulation, which simulates the Global Mirror round-trip delay between two clusters in milliseconds. The default is 0. The valid range is 0 - 100 milliseconds.

- Optional: The **gm_intra_cluster_delay_simulation** parameter

This optional parameter specifies the intracluster delay simulation, which simulates the Global Mirror round-trip delay in milliseconds. The default is 0. The valid range is 0 - 100 milliseconds.

7.5.2 The **chcluster** and **chpartnership** commands

The **chcluster** and **chpartnership** commands (as shown in Example 7-1) alter the Global Mirror settings and the cluster and partnership level.

Example 7-1 Alter Global Mirror settings

```
svctask copartnership -bandwidth 20 cluster1
svctask copartnership -stop cluster1
```

For more information about the use of Metro Mirror and Global Mirror commands, see *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933, or use the command-line help option (-h).

7.5.3 Distribution of Global Mirror bandwidth

The Global Mirror bandwidth resource is distributed within the cluster. You can optimize the distribution of volumes within I/O groups at the local and remote clusters to maximize performance.

Although defined at a cluster 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 cluster to each node on the remote cluster. 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 cluster. The more nodes that exist at the remote cluster, 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 clusters.

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 Global Mirror to route I/O between clusters.

The SAN Volume Controller node that receives a write for a volume is normally the preferred node of the volume. For volumes in a Global Mirror 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.

Tip: The preferred node for a volume cannot be changed nondisruptively or easily after the volume is created.

Each node of the remote cluster has a fixed pool of Global Mirror system resources for *each* node of the primary cluster. 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 cluster are set so that every combination of primary node and secondary node is used, Global Mirror performance is maximized.

Figure 7-17 shows an example of Global Mirror resources that are not optimized. Volumes from the local cluster are replicated to the remote cluster, where all volumes with a preferred node of node 1 are replicated to the remote cluster, where the target volumes also have a preferred node of node 1.

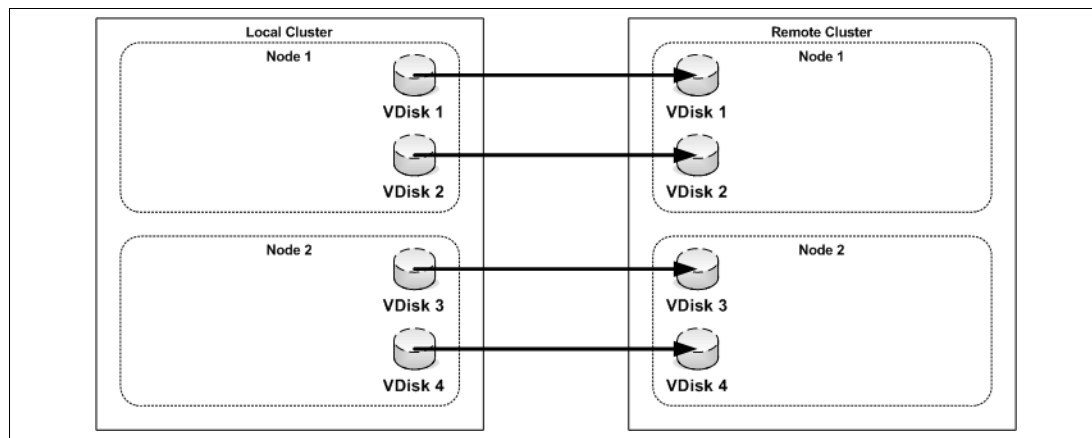


Figure 7-17 Global Mirror resources that are not optimized

With this configuration, the resources for remote cluster node 1 that are reserved for local cluster node 2 are not used. The resources for local cluster node 1 that are used for remote cluster node 2 also are not used.

If the configuration changes to the configuration that is shown in Figure 7-18, all Global Mirror resources for each node are used and SAN Volume Controller Global Mirror operates with better performance than this configuration.

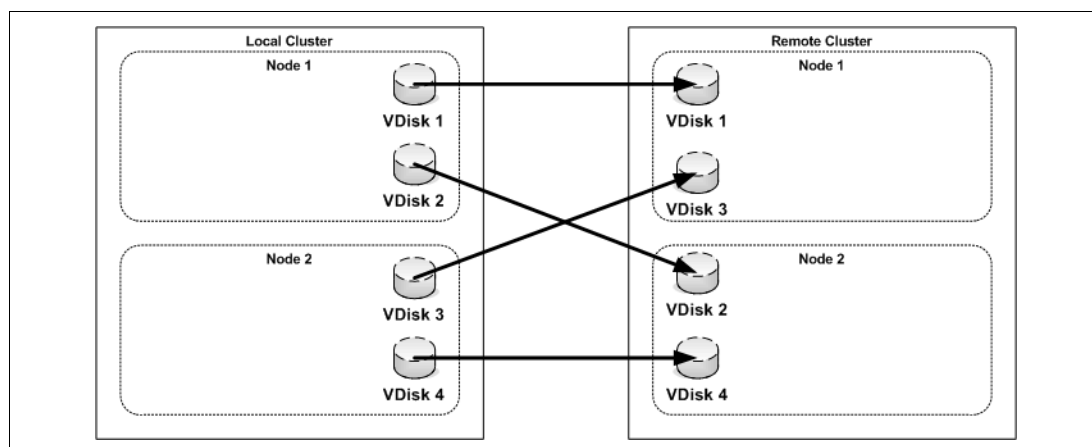


Figure 7-18 Optimized Global Mirror resources

Effect of the Global Mirror Bandwidth parameter on foreground I/O latency

The Global Mirror 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 background copy bandwidth is set too high (compared to the ICL capacity), the synchronous secondary writes of foreground I/Os delay and increase the foreground I/O latency as perceived by the applications.

- Increasing latency of foreground I/O

If the Global Mirror bandwidth parameter is set too high for the actual ICL capability, the background copy resynchronization writes use too much of the ICL. 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 Global Mirror bandwidth parameter (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 Global Mirror bandwidth parameter (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 ICL, 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
- ICL bandwidth
- Auxiliary storage

Changes in the environment, or loading of it, can affect the foreground I/O. SAN Volume Controller provides the client with a means to monitor, and a parameter to control, how foreground I/O is affected by running remote copy processes. SAN Volume Controller code 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. This “cut-off valve” parameter is called **gmlinktolerance**.

Internal monitoring and the gmlinktolerance parameter

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 Metro Mirror and 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 average over a period of 10 seconds, this period is treated as being overloaded.

Global Mirror uses the **maxhostdelay** and **gmlinktolerance** parameters to monitor Global Mirror protocol backlogs in the following ways:

- ▶ Users set the **maxhostdelay** and **gmlinktolerance** parameters to control how software responds to these delays. The **maxhostdelay** 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 over half of these writes is greater than the **maxhostdelay** 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.

If the link is overloaded for several consecutive seconds greater than the **gmlinktolerance** value, a 1920 error (or other Global Mirror error code) is recorded against the volume that used the most Global Mirror resource over recent time.

A period without overload decrements the count of consecutive periods of overload. Therefore, an error log is also raised if, over any period, the amount of time in overload exceeds the amount of nonoverloaded time by the **gmlinktolerance** parameter.

Bad periods and the gmlinktolerance parameter

The **gmlinktolerance** parameter is defined in seconds. Bad periods are assessed at intervals of 10 seconds. The maximum bad period count is the **gmlinktolerance** parameter value that is divided by 10.

With a **gmlinktolerance** value of 300, the maximum bad period count is 30. When 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.

I/O assessment within bad periods

Within each sample period, I/Os are assessed. The proportion of bad I/O to good I/O is calculated. If the proportion exceeds a defined value, the sample period is defined as a bad period. A consequence is that, under a light I/O load, a single bad I/O 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.

Edge case

The worst possible situation is achieved by setting the `gm_max_host_delay` and `gmlinktolerance` 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 `gmlinktolerance` value, `gm_max_host_delay` setting, or I/O load might reduce the risk of encountering this edge case.

7.5.4 1920 errors

The SAN Volume Controller Global Mirror process aims to maintain a low response time of foreground writes even when the long-distance link has a high response time. It monitors how well it is doing compared to the goal by measuring how long it is taking to process I/O.

Specifically, SAN Volume Controller measures the locking and serialization part of the protocol that takes place when a write is received. It compares this information with how much time the I/O is likely to take if Global Mirror processes were not active. If this extra time is consistently greater than 5 ms, Global Mirror determines that it is not meeting its goal and shuts down the most bandwidth-consuming relationship. This situation generates a 1920 error and protects the local SAN Volume Controller from performance degradation.

I/O information: Debugging 1920 errors requires detailed information about I/O at the primary and secondary clusters, in addition to node-to-node communication. As a minimum requirement, I/O stats must be running, covering the period of a 1920 error on *both* clusters, and if possible, Tivoli Storage Productivity Center statistics must be collected.

7.6 Global Mirror planning

When you plan for Global Mirror, you must keep in mind the considerations that are outlined in the following sections.

7.6.1 Rules for using Metro Mirror and Global Mirror

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

- ▶ For V6.2 and earlier, you cannot have FlashCopy targets in a Metro Mirror or Global Mirror relationship. Only FlashCopy sources can be in a Metro Mirror or Global Mirror relationship. For more information, see 7.2.1, “Remote copy in SAN Volume Controller V7.2” on page 161.
- ▶ 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 intracluster Metro Mirror or Global Mirror only between volumes in the same I/O group.

- ▶ 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).
- ▶ When you use SAN Volume Controller Global Mirror, ensure that all components in the SAN switches, remote links, and storage controllers can sustain the workload that is generated by application hosts or foreground I/O on the primary cluster. They must also sustain workload that is generated by the following remote copy processes:
 - Mirrored foreground writes
 - Background copy (background write resynchronization)
 - Intercluster heartbeat messaging
- ▶ You must set the Ignore Bandwidth parameter (which controls the background copy rate) to a value that is appropriate to the link and secondary back-end storage.
- ▶ Global Mirror is not supported for cache-disabled volumes that are participating in a Global Mirror relationship.
- ▶ Use a SAN performance monitoring tool, such as IBM Tivoli Storage Productivity Center, to continuously monitor the SAN components for error conditions and performance problems.
- ▶ Have IBM Tivoli Storage Productivity Center alert you when a performance problem occurs or if a Global Mirror (or Metro Mirror) link is automatically suspended by SAN Volume Controller. A remote copy relationship that remains stopped without intervention can severely affect your recovery point objective. Also, restarting a link that was suspended for a long time can add burden to your links while the synchronization catches up.
- ▶ Set the **gmlinktolerance** parameter of the remote copy partnership to an appropriate value. The default value of 300 seconds (5 minutes) is appropriate for most clients.
- ▶ If you plan to perform SAN maintenance that might affect SAN Volume Controller Global Mirror relationships, complete the following tasks:
 - Select a maintenance window where application I/O workload is reduced during the maintenance.
 - Disable the **gmlinktolerance** feature or increase the **gmlinktolerance** value, meaning that application hosts might see extended response times from Global Mirror volumes.
 - Stop the Global Mirror relationships.

7.6.2 Planning overview

Ideally, consider the following areas on a holistic basis and test them by running data collection tools before you go live:

- ▶ The ICL
- ▶ Peak workloads at the primary cluster
- ▶ Back-end storage at both clusters

Before you start with SAN Volume Controller remote copy services, consider any overhead that is associated with their introduction. You must fully know and understand your current infrastructure. Specifically, you must consider the following items:

- ▶ ICL or link distance and bandwidth
- ▶ Load of the current SAN Volume Controller clusters and of the current storage array controllers

Bandwidth analysis and capacity planning for your links helps to define how many links you need and when you need to add more links to ensure the best possible performance and high availability. As part of your implementation project, you can identify and then distribute hot spots across your configuration, or take other actions to manage and balance the load.

You must consider the following areas:

- ▶ If your bandwidth is so little, you might see an increase in the response time of your applications at times of high workload.
- ▶ The speed of light is less than 300,000 km/s, which is less than 300 km/ms on fiber optic cable. Therefore, the data must go to the other site and then an acknowledgment must come back. Add any possible latency times of some active components on the way and you get approximately 1 ms of overhead per 100 km for write I/Os.

Metro Mirror adds extra latency time because of the link distance to the time of write operation.

- ▶ Determine whether your current SAN Volume Controller cluster or clusters can handle the extra load.

Problems are not always related to remote copy services or ICL, but rather to hot spots on the disks subsystems. Be sure to resolve these problems. Can your auxiliary storage handle the added workload that it receives? It is basically the same back-end workload that is generated by the primary applications.

7.6.3 Planning specifics

You can use Metro Mirror and Global Mirror between two clusters, as explained in this section.

Remote copy mirror relationship

A *remote copy mirror relationship* is a relationship between two volumes of the same size. Management of the remote copy mirror relationships is always performed in the cluster where the source volume exists. However, you must consider the performance implications of this configuration because write data from all mirroring relationships is transported over the same ICLs.

Metro Mirror and Global Mirror respond differently to a heavily loaded, poorly performing link. Metro Mirror usually maintains the relationships in a consistent synchronized state, meaning that primary host applications start to detect poor performance as a result of the synchronous mirroring that is being used.

However, Global Mirror offers a higher level of write performance to primary host applications. With a well-performing link, writes are completed asynchronously. If link performance becomes unacceptable, the link tolerance feature automatically stops Global Mirror relationships to ensure that the performance for application hosts remains within reasonable limits.

Therefore, with active Metro Mirror and Global Mirror relationships between the same two clusters, Global Mirror writes might suffer degraded performance if Metro Mirror relationships use most of the ICL capability. If this degradation reaches a level where hosts that write to Global Mirror experience extended response times, the Global Mirror relationships can be stopped when the link tolerance threshold is exceeded. If this situation happens, see 7.5.4, “1920 errors” on page 194.

Supported partner clusters

This section provides the following considerations for intercluster compatibility regarding SAN Volume Controller release code and hardware types:

- ▶ Clusters that run V6.1 or later cannot form partnerships with clusters that run on V4.3.1 or earlier.

Back-end storage controller requirements

The storage controllers in a remote SAN Volume Controller cluster must be provisioned to allow for 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

The performance of applications at the primary cluster can be limited by the performance of the back-end storage controllers at the remote cluster.

To maximize the number of I/Os that applications can perform to Global Mirror and Metro Mirror volumes, complete the following tasks:

- ▶ Ensure that Global Mirror and Metro Mirror volumes at the remote cluster are in dedicated managed disk groups. The managed disk groups must not contain nonmirror volumes.
- ▶ Configure storage controllers to support the mirror workload that is required of them, which might be achieved in the following ways:
 - Dedicating storage controllers to only Global Mirror and Metro Mirror volumes
 - Configuring the controller to ensure sufficient quality of service for the disks that are used by Global Mirror and Metro Mirror
 - Ensuring that physical disks are not shared between Global Mirror or Metro Mirror volumes and other I/O
 - Verifying that MDisks within a mirror managed disk group must be similar in their characteristics (for example, Redundant Array of Independent Disks [RAID] level, physical disk count, and disk speed)

Technical references and limits

The Metro Mirror and Global Mirror operations support the following functions:

- ▶ Intracluster copying of a volume, in which both VDisks belong to the same cluster and I/O group within the cluster
- ▶ Intercluster copying of a Disk, in which one Disk belongs to a cluster and the other Disk belongs to a different cluster

Tip: A cluster can participate in active Metro Mirror and Global Mirror relationships with itself and up to three other clusters.

- ▶ Concurrent usage of intercluster and intracluster Metro Mirror and Global Mirror relationships within a cluster.
- ▶ Bidirectional ICL, meaning that it can copy data from cluster A to cluster B for one pair of VDisks and copy data from cluster B to cluster A for a different pair of VDisks.
- ▶ Reverse copy for a consistent relationship.
- ▶ Consistency groups support to manage a group of relationships that must be kept synchronized for the same application.

This support also simplifies administration because a single command that is issued to the consistency group is applied to all the relationships in that group.

- Support for a maximum of 8192 Metro Mirror and Global Mirror relationships per cluster.

7.7 Global Mirror use cases

This section describes the common uses cases of Global Mirror.

7.7.1 Synchronizing a remote copy relationship

You can choose from three methods to establish (or synchronize) a remote copy relationship.

Full synchronization after the Create method

The *full synchronization after Create method* is the default method. It is the simplest in that it requires no other administrative activity apart from issuing the necessary SAN Volume Controller commands.

A CreateRelationship with CreateConsistent state set to FALSE. Start the remote copy relationship with CLEAN parameter set to FALSE

However, in some environments, the available bandwidth make this method unsuitable.

Synchronization before the Create method

In the *synchronization before Create method*, the administrator must ensure that the master and auxiliary virtual disks contain identical data before a relationship is created.

The administrator can perform this check in 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 master or auxiliary volume before the relationship is established. The administrator must then issue the following settings:

- A CreateRelationship with CreateConsistent state set to TRUE
- A Start the relationship with Clean set to FALSE

This method has an advantage over the full synchronization method, in that it does not require all the data to be copied over a constrained link. However, if the data must be copied, the master and auxiliary disks cannot be used until the copy is complete, which might be unacceptable.

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.

Quick synchronization after Create method

In the *quick synchronization after Create method*, the administrator must still copy data from the master to auxiliary volume. However, the data can be used without stopping the application at the master volume.

This method has the following flow:

- A CreateRelationship issued with CreateConsistent set to TRUE.

- ▶ A Stop (Relationship) is issued with EnableAccess set to TRUE.
- ▶ A tape image (or other method of transferring data) is used to copy the entire master volume to the auxiliary volume after the copy is complete,
- ▶ The relationship is restarted with Clean set to TRUE.

With this technique, only the data that changed since the relationship was created, including all regions that were incorrect in the tape image, are copied by remote copy from the master and auxiliary volumes.

Attention: As described in “Synchronization before the Create method” on page 198, you must perform the copy step correctly. Otherwise, the auxiliary volume is useless, although remote copy reports it as *synchronized*.

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.

7.7.2 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 SAN Volume Controller 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 VDisks, 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 SAN Volume Controller to maintain the Metro Mirror or Global Mirror. This example does not require downtime for the hosts that use the source VDisks.

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 VDisks normally. No Metro Mirror relationship nor Global Mirror relationship is defined yet.
Identify all the VDisks that become the source VDisks in a Metro Mirror relationship or in a Global Mirror relationship.
2. Establish the SAN Volume Controller cluster relationship with the target SAN Volume Controller.

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 SAN Volume Controller from performing an initial sync.

Attention: If you do not use the **-sync** option, all of these steps are redundant because the SAN Volume Controller performs a full initial synchronization anyway.

2. Stop each mirror relationship by using the **-access** option, which enables write access to the target VDisks. 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 Tivoli Storage) to make an image backup of the volume.

Change tracking: Although the source is being modified while you are copying the image, the SAN Volume Controller 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 SAN Volume Controller applies all of the changes that occurred since the relationship stopped in step 2 on page 200. 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 SAN Volume Controller 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 SAN Volume Controller stops write access to the volume while the mirror relationship is running.
6. Start your Metro Mirror and Global Mirror relationships. While the mirror relationship catches up, the target volume is not usable at all. When it reaches *Consistent Copying* status, your remote volume is ready for use in a disaster.

7.7.3 Master and auxiliary volumes and switching their roles

When you create a Global Mirror relationship, the master volume is initially assigned as the master, and the auxiliary volume is initially assigned as the auxiliary. This design implies that the initial copy direction is mirroring the master volume to the auxiliary volume. After the initial synchronization is complete, the copy direction can be changed if appropriate.

In the most common applications of Global Mirror, the master volume contains the production copy of the data and is used by the host application. The auxiliary volume contains the mirrored copy of the data and is used for failover in disaster recovery scenarios.

Tips: Consider the following points:

- ▶ A volume can be part of only one Global Mirror relationship at a time.
- ▶ A volume that is a FlashCopy target cannot be part of a Global Mirror relationship.

7.7.4 Migrating a Metro Mirror relationship to Global Mirror

A Metro Mirror relationship can be changed to a Global Mirror relationship or a Global Mirror relationship to a Metro Mirror relationship. However, this procedure requires an outage to the host and is successful only if you can ensure that no I/Os are generated to the source or target volumes by completing the following steps:

1. Ensure that your host is running with volumes that are in a Metro Mirror or Global Mirror relationship. This relationship is in the *Consistent Synchronized* state.
2. Stop the application and the host.
3. (Optional) Unmap the volumes from the host to ensure that no I/O can be performed on these volumes. If currently outstanding write I/Os are in the cache, you might need to wait at least 2 minutes before you unmap the volumes.
4. Stop the Metro Mirror or Global Mirror relationship and ensure that the relationship stops with a *Consistent Stopped* status.
5. Delete the current Metro Mirror or Global Mirror relationship.
6. Create the Metro Mirror or Global Mirror relationship. Ensure that you create it as *synchronized* to stop the SAN Volume Controller from resynchronizing the volumes. Use the `-sync` flag with the `svctask mkrcrelationship` command.
7. Start the new Metro Mirror or Global Mirror relationship.
8. Remap the source volumes to the host if you unmapped them in step 3.
9. Start the host and the application.

Attention: If the relationship is not stopped in the consistent state, those changes are never mirrored to the target volumes. The same is true if any host I/O occurs between stopping the old Metro Mirror or Global Mirror relationship and starting the new Metro Mirror or Global Mirror relationship. As a result, the data on the source and target volumes is not the same, and the SAN Volume Controller is unaware of the inconsistency.

7.7.5 Multicluster mirroring

The concept of multicluster mirroring was introduced with SAN Volume Controller V5.1.0. Previously, mirroring was limited to a one-to-one only mapping of clusters.

Each SAN Volume Controller cluster can maintain up to three partner cluster relationships, which allows as many as four clusters to be directly associated with each other. This SAN Volume Controller partnership capability enables the implementation of disaster recovery solutions.

Figure 7-19 shows a multiple cluster mirroring configuration.

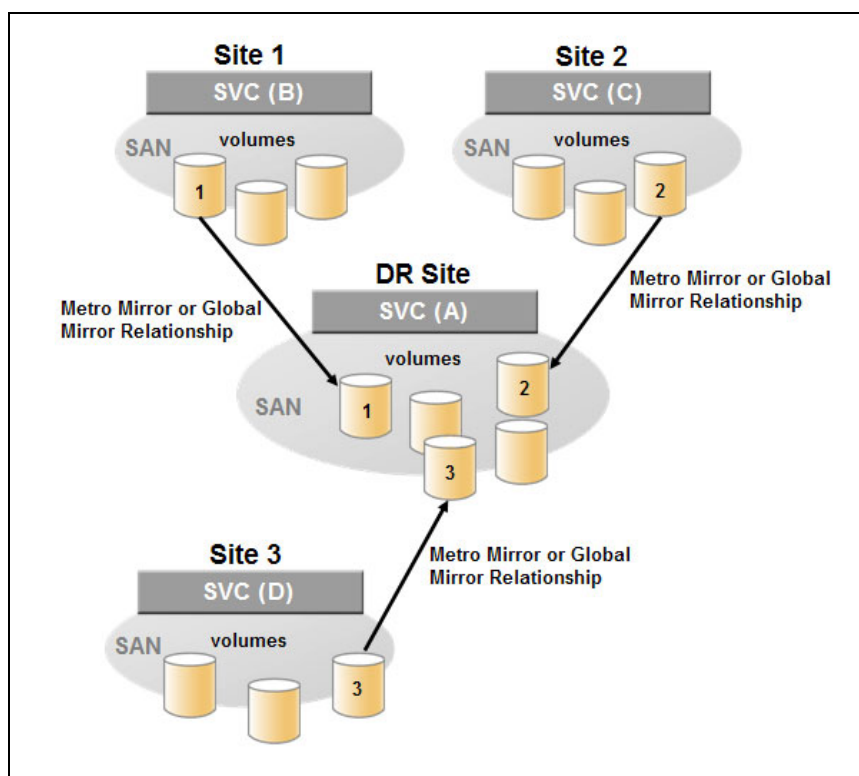


Figure 7-19 Multiple cluster mirroring configuration

Note: The following software-level restrictions apply to multiple cluster mirroring:

- ▶ Partnership between a cluster that runs V6.1 and a cluster that runs V4.3.1 or earlier is not supported.
- ▶ Clusters in a partnership where one cluster is V6.1 and the other cluster is running V4.3.1 cannot participate in more partnerships with other clusters.
- ▶ Clusters that are all running V6.1 or V5.1 can participate in up to three cluster partnerships.

Object names: SAN Volume Controller V6.1 supports object names up to 63 characters. Previous levels supported only up to 15 characters. When SAN Volume Controller V6.1 clusters are partnered with V4.3.1 and V5.1.0 clusters, various object names are truncated at 15 characters when displayed from V4.3.1 and V5.1.0 clusters.

Supported multiple cluster mirroring topologies

Multiple cluster mirroring allows for various partnership topologies as shown in the examples in this section.

Star topology: A-B, A-C, and A-D

Figure 7-20 shows four clusters in a star topology, with cluster A at the center. Cluster A can be a central disaster recovery site for the three other locations.

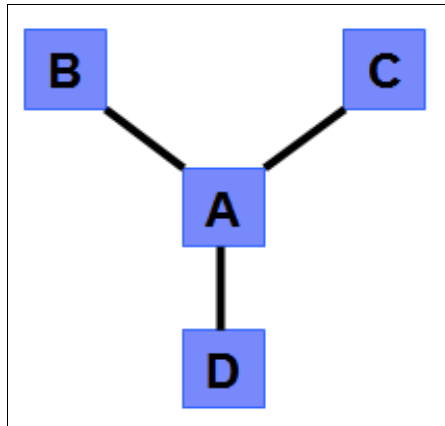


Figure 7-20 SAN Volume Controller star topology

Using a star topology, you can migrate applications by using the following process:

1. Suspend application at A.
2. Remove the A-B relationship.
3. Create the A-C relationship (or alternatively, the B-C relationship).
4. Synchronize to cluster C, and ensure that the following A-C relationship is established:
 - A-B, A-C, A-D, B-C, B-D, and C-D
 - A-B, A-C, and B-C

Triangle topology: A-B, A-C, and B-C

Figure 7-21 shows three clusters in a triangle topology. A potential use case might be that data center B is migrating to data center C, considering that data center A is the host production site and that data centers B and C are the disaster recovery sites.

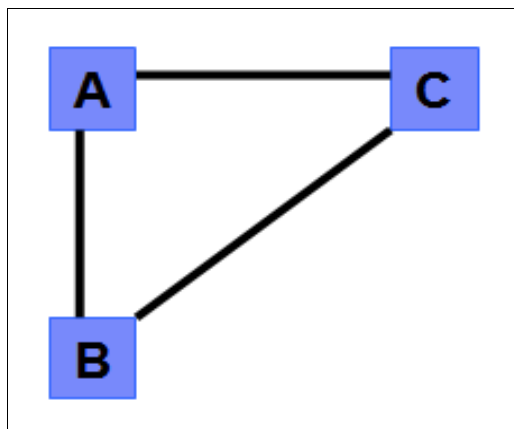


Figure 7-21 SAN Volume Controller triangle topology

By using the cluster-star topology, you can migrate different applications at different times by using the following process:

1. Suspend the application at data center A.
2. Take down the A-B data center relationship.

3. Create an A-C data center relationship (or a B-C data center relationship).
4. Synchronize to data center C and ensure that the A-C data center relationship is established.

Migrating different applications over a series of weekends provides a phased migration capability.

Fully connected topology: A-B, A-C, A-D, B-D, and C-D

Figure 7-22 is a fully connected mesh where every cluster has a partnership to each of the three other clusters, which allows volumes to be replicated between any pair of clusters.

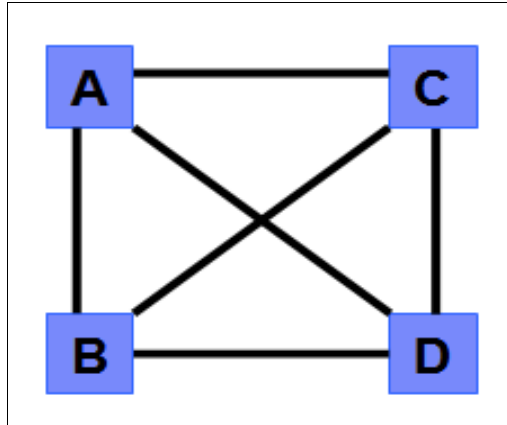


Figure 7-22 SAN Volume Controller fully connected topology

Attention: Create this configuration only if relationships are needed between every pair of clusters. Restrict intercluster zoning only to where it is necessary.

Daisy chain topology: A-B, A-C, and B-C

Figure 7-23 shows a daisy-chain topology.

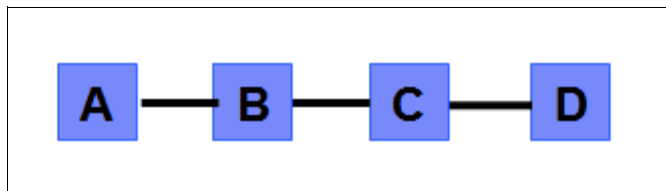


Figure 7-23 SAN Volume Controller daisy-chain topology

Although clusters can have up to three partnerships, volumes can be part of only one remote copy relationship; for example, A-B.

Unsupported topology: A-B, B-C, C-D, and D-E

Figure 7-24 shows an unsupported topology where five clusters are indirectly connected. If the cluster can detect this unsupported topology at the time of the fourth `mkpartnership` command, this command is rejected with an error message, which sometimes is not possible. In this case, an error is displayed in the error log of each cluster in the connected set.

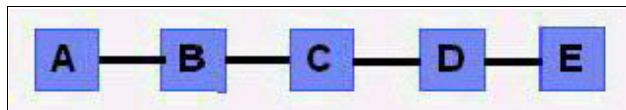


Figure 7-24 Unsupported SAN Volume Controller topology

7.7.6 Performing three-way copy service functions

Three-way copy service functions that use SAN Volume Controller 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 by combining SAN Volume Controller copy services (with image mode cache-disabled volumes) and native storage controller copy services. Both relationships are active, as shown in Figure 7-25.

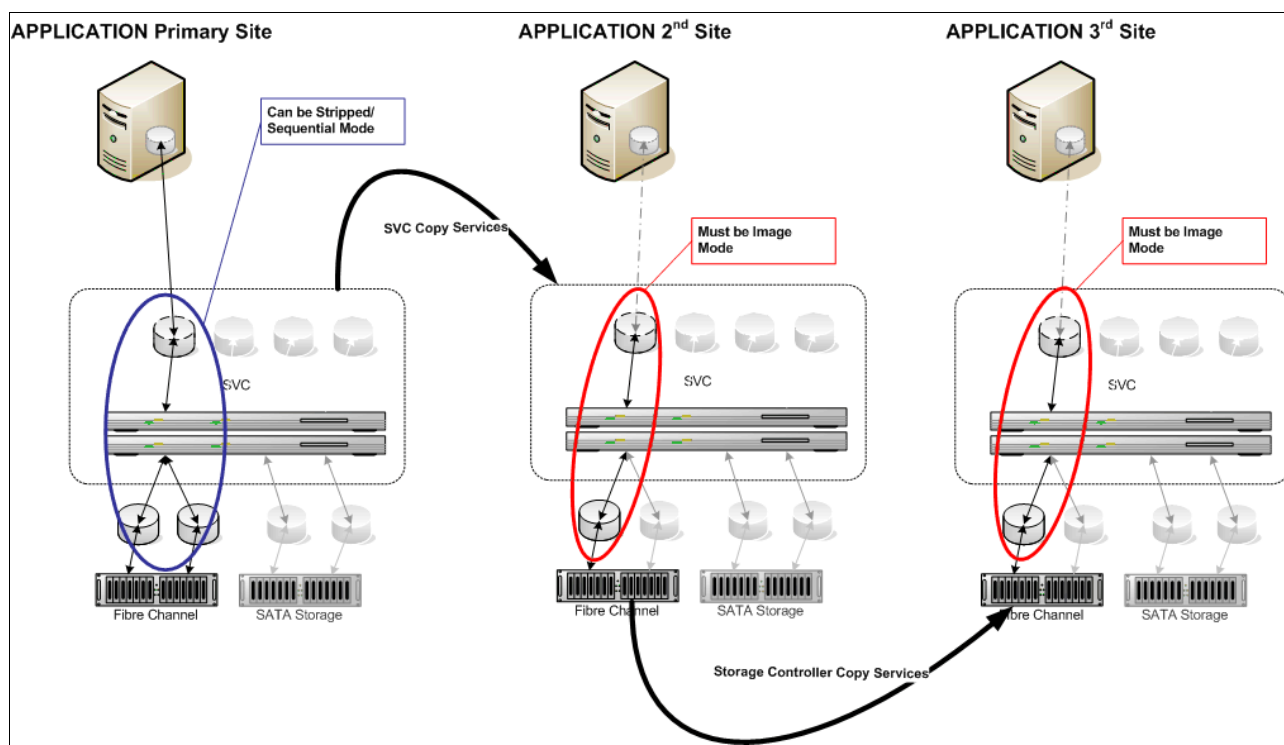


Figure 7-25 Using three-way copy services

Important: The SAN Volume Controller supports copy services between only *two* clusters.

In Figure 7-25, the primary site uses SAN Volume Controller copy services (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.

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>

Storage controller is unaware of the SAN Volume Controller

When you use the copy services function in a storage controller, the storage controller has no knowledge that the SAN Volume Controller exists and that the storage controller uses those disks on behalf of the real hosts. Therefore, when you allocate source volumes and target volumes in a point-in-time copy relationship or a remote mirror relationship, ensure that you choose them *in the correct order*. If you accidentally use a source logical unit number (LUN) with SAN Volume Controller data on it as a target LUN, you can corrupt that data.

If that LUN was a managed disk (MDisk) in a managed disk group with striped or sequential volumes on it, the managed disk group might be brought offline. This situation, in turn, makes all of the volumes that belong to that group go offline.

When you define LUNs in a point-in-time copy or a remote mirror relationship, verify that the SAN Volume Controller is not visible to the LUN (by masking it so that no SAN Volume Controller node can detect it). Alternatively, if the SAN Volume Controller must detect the LUN, ensure that LUN is an *unmanaged MDisk*.

As part of its Advanced Copy Services function, the storage controller might take a LUN offline or suspend reads or writes. The SAN Volume Controller does not understand why this happens. Therefore, the SAN Volume Controller might log errors when these events occur.

Consider a case in which you mask target LUNs to the SAN Volume Controller and rename your MDisks as you discover them, and the Advanced Copy Services function prohibits access to the LUN as part of its processing. In this case, the MDisk might be discarded and rediscovered with an MDisk name that is assigned by SAN Volume Controller.

Cache-disabled image mode volumes

When the SAN Volume Controller 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 use the LUN for any other type of SAN Volume Controller Volume, you risk data loss of the data on that LUN. You might also bring down all volumes in the managed disk group to which you assigned that LUN (MDisk).

If you leave caching enabled on a volume, the underlying controller does not receive any write I/Os as the host writes them. The SAN Volume Controller 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.

7.7.7 When to use storage controller Advanced Copy Services functions

The SAN Volume Controller provides greater flexibility than using only native copy service functions.

Regardless of the storage controller behind the SAN Volume Controller, you can use the Subsystem Device Driver (SDD) to access the storage. As your environment changes and your storage controllers change, the use of SDD negates the need to update device driver software as those changes occur.

The SAN Volume Controller can provide copy service functions between any supported controller to any other supported controller, even if the controllers are from different vendors. By using this capability, you can use a lower class or cost of storage as a target for point-in-time copies or remote mirror copies.

By using SAN Volume Controller, you can move data around without host application interruption, which is helpful especially when the storage infrastructure is retired and new technology becomes available.

However, some storage controllers can provide more copy service features and functions compared to the capability of the current version of SAN Volume Controller. If you require usage of those other features, you can use them and the features that the SAN Volume Controller provides by using cache-disabled image mode VDisks.

7.7.8 Using Metro Mirror or Global Mirror with FlashCopy

With SAN Volume Controller, you can use a volume in a Metro Mirror or Global Mirror relationship as a source volume for FlashCopy mapping. You cannot use a volume as a FlashCopy mapping target that is already in a Metro Mirror or Global Mirror relationship.

When you prepare FlashCopy mapping, the SAN Volume Controller places the source volume in a temporary cache-disabled state. This temporary state adds latency to the Metro Mirror relationship because I/Os that are normally committed to SAN Volume Controller memory now must be committed to the storage controller.

One way to avoid this latency is to temporarily stop the Metro Mirror or Global Mirror relationship before you prepare FlashCopy mapping. When the Metro Mirror or Global Mirror relationship is stopped, the SAN Volume Controller records all changes that occur to the source volumes. Then, it applies those changes to the target when the remote copy mirror is restarted.

To temporarily stop the Metro Mirror or Global Mirror relationship before you prepare the FlashCopy mapping, complete the following steps:

1. Stop each mirror relationship by using the **-access** option, which enables write access to the target volumes. You need this access later.
2. 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 Tivoli Storage Manager) to make an image backup of the volume.

Tracking and applying the changes: Although the source is modified when you copy the image, the SAN Volume Controller is tracking those changes. The image that you create might include part of the changes and is likely to miss part of the changes.

When the relationship is restarted, the SAN Volume Controller applies all changes that occurred since the relationship stopped in step 1. After all the changes are applied, you have a consistent target image.

3. 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 this 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 SAN Volume Controller starts running and maintaining the Metro Mirror and Global Mirror.

4. Unmount the target volumes from your host. When you start the Metro Mirror and Global Mirror relationship later, the SAN Volume Controller stops write access to the volume when the mirror relationship is running.
5. Start your Metro Mirror and Global Mirror relationships. While the mirror relationship catches up, the target volume is unusable. When it reaches the *Consistent Copying* status, your remote volume is ready for use in a disaster.

7.7.9 Global Mirror upgrade scenarios

When you upgrade cluster software where the cluster participates in one or more intercluster relationships, upgrade only one cluster at a time. That is, do not upgrade the clusters concurrently.

Attention: Upgrading both clusters concurrently is not monitored by the software upgrade process.

Allow the software upgrade to complete one cluster before it is started on the other cluster. Upgrading both clusters 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.

Intercluster Metro Mirror and Global Mirror compatibility cross-reference

For more information about a compatibility table for intercluster Metro Mirror and Global Mirror relationships between SAN Volume Controller code levels, see *SAN Volume Controller Inter-cluster 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>

If clusters are at the same code level, the partnership is supported. If clusters are at different code levels, check the compatibility according to the table in Figure 7-26 by completing the following steps:

1. Select the higher code level from the column on the left side of the table.
2. Select the partner cluster code level from the row on the top of the table.

Figure 7-26 also shows intercluster Metro Mirror and Global Mirror compatibility.

Inter-cluster Metro Mirror and Global Mirror support:										
Cluster code levels	4.3.0	4.3.1	5.1.0	6.1.0	6.2.0	6.3.0	6.4.0	6.4.1	7.1.0	7.2.0
4.3.0	✓	-	-	-	-	-	-	-	-	-
4.3.1	✓	✓	-	-	-	-	-	-	-	-
5.1.0	✓	✓	✓	-	-	-	-	-	-	-
6.1.0	✗	✓ *	✓ *	✓ *	-	-	-	-	-	-
6.2.0	✗	✓ *	✓ *	✓ *	✓ *	-	-	-	-	-
6.3.0	✗	✗	✓ *	✓ *	✓ *	✓	-	-	-	-
6.4.0	✗	✗	✓ *	✓ *	✓ *	✓	✓	-	-	-
6.4.1	✗	✗	✓ *	✓ *	✓ *	✓	✓	✓	-	-
7.1.0	✗	✗	✓ *	✓ *	✓ *	✓	✓	✓	✓	-
7.2.0	✗	✗	✓ *	✓ *	✓ *	✓	✓	✓	✓	✓

✓ * If performing Global/Metro Mirror between SVC and a Storwize Family System then both systems must be running 6.3.0 or later.

Figure 7-26 Intercluster Metro Mirror and Global Mirror compatibility

If all clusters are running software V5.1 or later, each cluster can be partnered with up to three other clusters, which support Multicuster Mirroring. If a cluster is running a software level of V5.1 or earlier, each cluster can be partnered with only one other cluster.

7.8 Intercluster Metro Mirror and Global Mirror source as an FC target

The inclusion of Metro Mirror and Global Mirror source as an FC 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.
- ▶ The I/O group for the FlashCopy mappings must be the same as the I/O group for the FlashCopy target volume.

Figure 7-27 shows a Metro Mirror or Global Mirror and FlashCopy relationship before SAN Volume Controller V6.2.

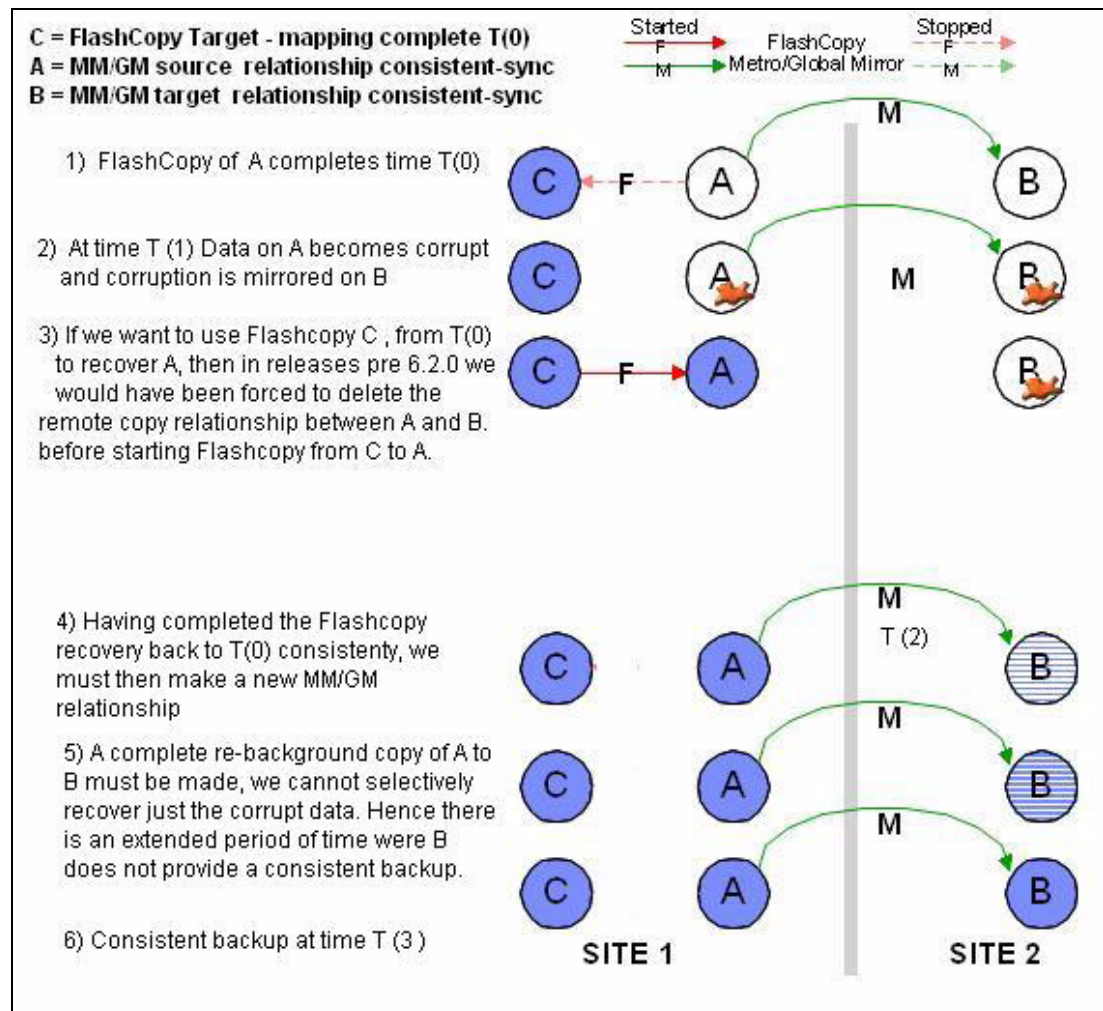


Figure 7-27 Metro Mirror or Global Mirror and FlashCopy relationship before SAN Volume Controller V6.2

Figure 7-28 shows a Metro Mirror or Global Mirror and FlashCopy relationship with SAN Volume Controller V6.2.

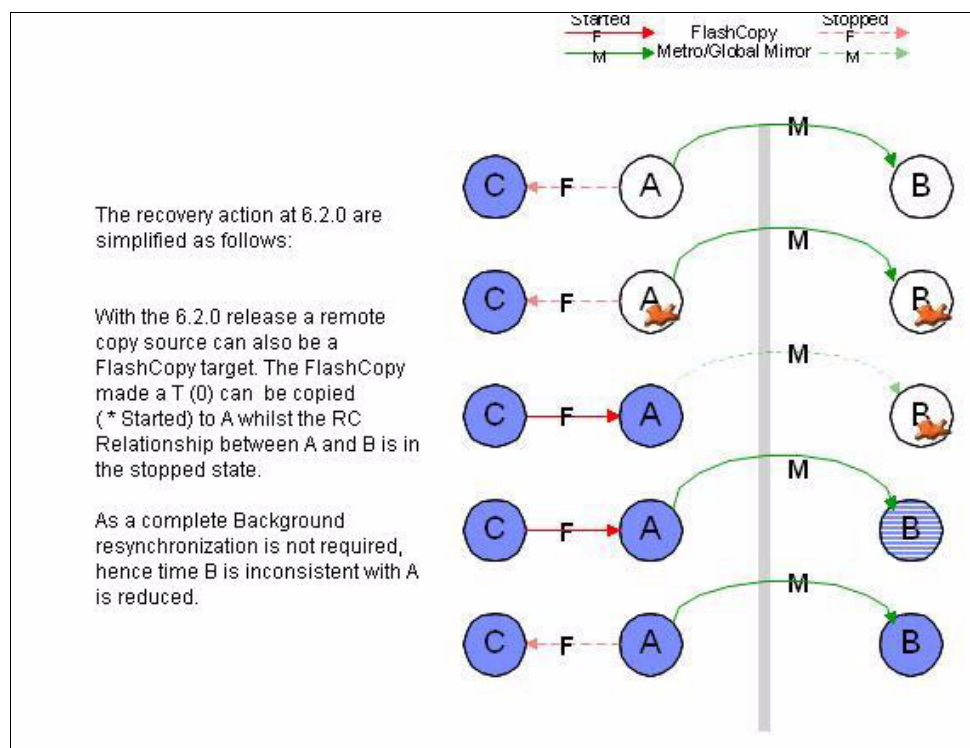


Figure 7-28 Metro Mirror or Global Mirror and FlashCopy relationships with SAN Volume Controller V6.2

7.9 States and steps in the Global Mirror relationship

A Global Mirror relationship has various states and actions that allow for or lead to changes of state. You can create Global Mirror relationships as “Requiring Synchronization” (default) or as “Being Synchronized.” For simplicity, this section considers single relationships, not consistency groups.

Requiring full synchronization (after creation)

Full synchronization after creation is the *default method*, and therefore, the simplest method. However, in some environments, the bandwidth that is available makes this method unsuitable.

The following commands are used to create and start a Global Mirror relationship of this type:

- ▶ A Global Mirror relationship is created by using the **mkrcrelationship** command (without the **-sync** flag).
- ▶ A new relationship is started by using the **startrcrelationship** command (without the **-clean** flag).

Synchronized before creation

When you make a synchronized Global Mirror relationship, you specify that the source volume and target volume are in sync. That is, they contain identical data at the point at which you start the relationship. There is no requirement for background copying between the volumes.

In this method, the administrator must ensure that the source and target volumes contain identical data before the relationship is created. There are two ways to ensure that the source and master volumes contain identical data:

- Both volumes are created with the security delete (**-fmtdisk**) feature to make all data zero.
- A complete tape image (or other method of moving data) is copied from the source volume to the target volume before you start the Global Mirror relationship. With this technique, do not allow I/O on the source or target before the relationship is established.

Then, the administrator must run the following commands:

- To ensure that a Global Mirror relationship is created, run the **mkrcrelationship** command with the **-sync** flag.
- To ensure that a new relationship is started, run the **startcrelationship** command with the **-clean** flag.

Attention: If you do not correctly perform these steps, Global Mirror can report the relationship as consistent when it is not, which creates a data loss or data integrity exposure for hosts that access the data on the auxiliary volume.

7.9.1 Global Mirror states

Figure 7-29 shows the steps and states regarding the Global Mirror relationships that are synchronized, and those relationships that require synchronization after creation.

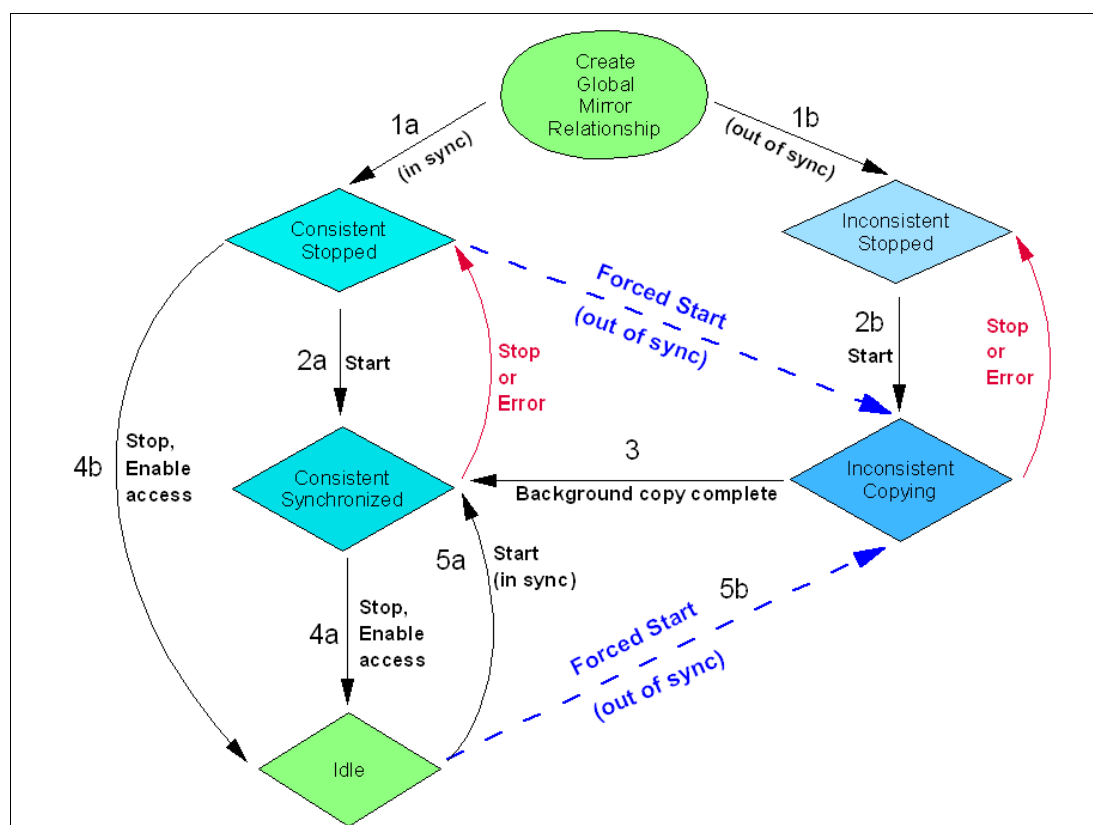


Figure 7-29 Global Mirror states diagram

Global Mirror relationships: Synchronized states

The Global Mirror relationship is created with the **-sync** option and the Global Mirror relationship enters the ConsistentStopped state (1a).

When a Global Mirror relationship starts in the ConsistentStopped state, it enters the ConsistentSynchronized state (2a). This state implies that no updates (write I/O) were performed on the master volume when in the ConsistentStopped state.

Otherwise, you must specify the **-force** option, and the Global Mirror relationship then enters the InconsistentCopying state when the background copy is started.

Global Mirror relationships: Out of Synchronized states

The Global Mirror relationship is created without specifying that the source and target volumes are in sync, and the Global Mirror relationship enters the InconsistentStopped state (1b). When a Global Mirror relationship starts in the InconsistentStopped state, it enters the InconsistentCopying state when the background copy is started (2b). When the background copy completes, the Global Mirror relationship transitions from the InconsistentCopying state to the ConsistentSynchronized state (3).

With the relationship in a consistent synchronized state, the target volume now contains a copy of source data that can be used in a disaster recovery scenario. The consistent synchronized state persists until the relationship is stopped for system administrative purposes or an error condition is detected (typically, a 1920 condition).

A Stop condition with enable access

When a Global Mirror relationship is stopped in the ConsistentSynchronized state (where specifying the **-access** option enables write I/O on the auxiliary volume), the Global Mirror relationship enters the Idling state, which is used in disaster recovery scenarios (4a).

To enable write I/O on the auxiliary volume, when the Global Mirror relationship is in the ConsistentStopped state, enter the **svctask stopprcrelationship** command with the **-access** option. Then, the Global Mirror relationship enters the Idling state (4b).

Tip: A forced start from ConsistentStopped or Idle changes the state to InconsistentCopying.

Stop or Error

When a remote copy relationship is stopped (intentionally or because of an error), a state transition is applied. For example, the Metro Mirror relationships in the ConsistentSynchronized state enter the ConsistentStopped state. The Metro Mirror relationships in the InconsistentCopying state enter the InconsistentStopped state. If the connection is broken between the SAN Volume Controller clusters in a partnership, all intercluster Metro Mirror relationships enter a Disconnected state.

You must be careful when you restart relationships that are in an idle state because auxiliary volumes in this state can process read and write I/O. If an auxiliary volume is written to when in an idle state, the state of relationship is implicitly altered to inconsistent. When you restart the relationship, you must change the direction of the relationship if you want to preserve any write I/Os that occurred on the auxiliary volume.

Starting from Idle

When you start a Metro Mirror relationship that is in the Idling state, you must specify the **-primary** argument to set the copy direction (5a). Because no write I/O was performed (to the master volume or auxiliary volume) when in the Idling state, the Metro Mirror relationship enters the ConsistentSynchronized state.

If write I/O was performed to the master volume or auxiliary volume, you must specify the **-force** option (5b). The Metro Mirror relationship then enters the InconsistentCopying state when the background copy is started.

7.9.2 Disaster recovery and Metro Mirror and Global Mirror states

A secondary (target volume) does not contain the requested data to be useful for disaster recovery purposes until the background copy is complete. Until this point, all new write I/O since the relationship started is processed through the background copy processes. As such, it is subject to sequence and ordering of the Metro Mirror and Global Mirror internal processes, which differ from the real-world ordering of the application.

At background copy completion, the relationship enters a ConsistentSynchronized state. All new write I/O is replicated as it is received from the host in a consistent-synchronized relationship. The primary and secondary volumes are different only in regions where writes from the host are outstanding.

In this state, the target volume is also available in read-only mode. As shown in Figure 7-29 on page 212, a relationship can enter from ConsistentSynchronized in either of the following states:

- ▶ ConsistentStopped (state entered when posting a 1920 error)
- ▶ Idling

Both the source and target volumes have a common point-in-time consistent state, and both are made available in read/write mode. Write available means that both volumes can service host applications, but any additional writing to volumes in this state causes the relationship to become inconsistent.

Tip: Moving from this point usually involves a period of inconsistent copying and, therefore, loss of redundancy. Errors that occur in this state become more critical because an inconsistent stopped volume does not provide a known consistent level of redundancy. The inconsistent stopped volume is unavailable in respect to read-only or write/read.

7.9.3 State definitions

States are portrayed to the user for consistency groups or relationships. This section describes these states and the major states to provide guidance about the available configuration commands.

The InconsistentStopped state

The InconsistentStopped state is a connected state. In this state, the master is accessible for read and write I/O, but the auxiliary is inaccessible for read or write I/O. A copy process must be started to make the auxiliary consistent. This state is entered when the relationship or consistency group is in the InconsistentCopying state and suffers a persistent error or receives a **stop** command that causes the copy process to stop.

A **start** command causes the relationship or consistency group to move to the InconsistentCopying state. A **stop** command is accepted, but has no effect.

If the relationship or consistency group becomes disconnected, the auxiliary side transitions to the InconsistentDisconnected state. The master side transitions to the IdlingDisconnected state.

The InconsistentCopying state

The InconsistentCopying state is a connected state. In this state, the master is accessible for read and write I/O, but the auxiliary is inaccessible for read or write I/O. This state is entered after a **start** command is issued to an InconsistentStopped relationship or consistency group. This state is also entered when a forced start is issued to an Idling or ConsistentStopped relationship or consistency group. In this state, a background copy process runs, which copies data from the master to the auxiliary volume.

In the absence of errors, an InconsistentCopying relationship is *active*, and the copy progress increases until the copy process completes. In certain error situations, the copy progress might freeze or regress. A persistent error or **stop** command places the relationship or consistency group into the InconsistentStopped state. A **start** command is accepted, but has no effect.

If the background copy process completes on a stand-alone relationship or on all relationships for a consistency group, the relationship or consistency group transitions to the ConsistentSynchronized state.

If the relationship or consistency group becomes disconnected, the auxiliary side transitions to the InconsistentDisconnected state. The master side transitions to the IdlingDisconnected state.

The ConsistentStopped state

The ConsistentStopped state is a connected state. In this state, the auxiliary contains a consistent image, but it might be out of date regarding the master. This state can arise when a relationship is in the ConsistentSynchronized state and experiences an error that forces a consistency freeze. It can also arise when a relationship is created with CreateConsistentFlag set to true.

Normally, after an I/O error, subsequent write activity causes updates to the master, and the auxiliary is no longer synchronized (set to false). In this case, consistency must be given up for a period to re-establish synchronization. You must use a **start** command with the **-force** option to acknowledge this situation and the relationship or consistency group transitions to the InconsistentCopying state. Run this command only after all of the outstanding events are repaired.

In the unusual case where the master and auxiliary are still synchronized (perhaps after a user stop and no further write I/O is received), a **start** command takes the relationship to the ConsistentSynchronized state. No **-force** option is required. Also, in this unusual case, a **switch** command is permitted that moves the relationship or consistency group to the ConsistentSynchronized state and reverses the roles of the master and the auxiliary.

If the relationship or consistency group becomes disconnected, the auxiliary side transitions to the ConsistentDisconnected state. The master side transitions to the IdlingDisconnected state.

An informational status log is generated whenever a relationship or consistency group enters the ConsistentStopped state with a status of *Online*. The ConsistentStopped state can be configured to enable an SNMP trap and provide a trigger to automation software to consider running a **start** command after a loss of synchronization.

The ConsistentSynchronized state

The ConsistentSynchronized state is a connected state. In this state, the master volume is accessible for read and write I/O. The auxiliary volume is accessible for read-only I/O. Writes that are sent to the master volume are sent to the master and auxiliary volumes. Either successful completion must be received for both writes. The write must be failed to the host, or a state must transition out of the ConsistentSynchronized state before a write is completed to the host.

A **stop** command takes the relationship to the ConsistentStopped state. A **stop** command with the **-access** parameter takes the relationship to the Idling state. A **switch** command leaves the relationship in the ConsistentSynchronized state, but reverses the master and auxiliary roles. A **start** command is accepted, but has no effect.

If the relationship or consistency group becomes disconnected, the same transitions are made as for the ConsistentStopped state.

The Idling state

The Idling state is a connected state. The master and auxiliary disks operate in the master role. The master and auxiliary disks are accessible for write I/O. In this state, the relationship or consistency group accepts a **start** command. Global Mirror maintains a record of regions on each disk that received write I/O when in the Idling state. This record is used to determine the areas that must be copied after a **start** command.

The **start** command must specify the new copy direction. This command can cause a loss of consistency if either volume in any relationship received write I/O, which is indicated by the synchronized status. If the **start** command leads to loss of consistency, you must specify a **-force** parameter.

After a **start** command, the relationship or consistency group transitions to the ConsistentSynchronized state if no loss of consistency occurs or to the InconsistentCopying state if a loss of consistency occurs.

Also, while in this state, the relationship or consistency group accepts a **-clean** option on the **start** command. If the relationship or consistency group becomes disconnected, both sides change their state to IdlingDisconnected.

7.10 1920 errors

Several mechanisms can lead to remote copy relationships stopping. Recovery actions are required to start them again.

7.10.1 Diagnosing and fixing 1920 errors

The SAN Volume Controller 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 ICL, unexpectedly higher foreground host I/O workload, or a permanent error because of a hardware failure. It is also possible that not all relationships are affected and that multiple 1920 errors can be posted.

Internal 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 VDisk, 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.

Users set the **maxhostdelay** and **gmlinktolerance** parameters to control how software responds to these delays. The **maxhostdelay** 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 over half of these writes are greater than the **maxhostdelay** parameter, that sample period is marked as bad. Software keeps a running count of bad periods. When a bad period occurs, this count goes up by one. When a good period occurs, this count goes down by one to a minimum value of 0. The **gmlinktolerance** parameter dictates the maximum allowable count of bad periods. The **gmlinktolerance** parameter is given in seconds in intervals of 10 seconds. The value of the **gmlinktolerance** parameter is divided by 10 and is used as the maximum bad period count. Therefore, if the value is 300, the maximum bad period count is 30. After this count is reached, the 1920 error is issued.

Bad periods do not need to be consecutive. For example, 10 bad periods, followed by five good periods, followed by 10 bad periods might result in a bad period count of 15.

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 cluster 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 Tivoli Storage Productivity Center with a statistics monitoring interval of 5 minutes. Also, turn on the internal statistics gathering function, **I0stats**, in SAN Volume Controller. Although not as powerful as Tivoli Storage Productivity Center, **I0stats** can provide valuable debug information if the **snap** command gathers system configuration data close to the time of failure.

7.10.2 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 ICL
- ▶ Primary storage and remote storage
- ▶ SAN Volume Controller nodes (internode communications, CPU usage, and the properties and state of remote copy volumes that are associated with remote copy relationships)

To debug, you must obtain information from the following components to ascertain their health at the point of failure:

- ▶ Switch logs (confirmation of the state of the link at the point of failure)
- ▶ Storage logs
- ▶ System configuration information from the master and auxiliary clusters for SAN Volume Controller (by using the **snap** command), including the following types:
 - I/O stats logs, if available
 - Live dumps, if they were triggered at the point of failure
- ▶ Tivoli Storage Productivity Center statistics (if available)

Important: Contact IBM Level 2 Support for assistance in collecting log information for 1920 errors. IBM Support personnel can provide collection scripts that you can use during problem recreation or that you can deploy during proof-of-concept activities.

Data collection for diagnostic purposes

A successful diagnosis depends on the collection of the following data at both clusters:

- ▶ The **snap** command with livedump (triggered at the point of failure)
- ▶ I/O Stats running
- ▶ Tivoli Storage Productivity Center (if possible)
- ▶ The following information and logs from other components:
 - ICL and switch details:
 - Technology
 - Bandwidth
 - Typical measured latency on the ICL
 - 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 VDisks that are involved
 - RAID rebuilds
 - Whether 1920 errors are associated with Workload Peaks or Scheduled Backup

Intercluster link

For diagnostic purposes, ask the following questions about the ICL:

- ▶ Was link maintenance being performed?

Consider the hardware or software maintenance that is associated with ICL; for example, updating firmware or adding more capacity.

- ▶ Is the ICL overloaded?

You can find indications of this situation by using statistical analysis with the help of I/O stats, Tivoli Storage Productivity Center, or both, to examine the internode communications, storage controller performance, or both. By using Tivoli Storage Productivity Center, you can check the storage metrics before for the Global Mirror relationships were stopped, which can be tens of minutes depending on the **gmLinktolerance** parameter.

Diagnose the overloaded link by using the following methods:

- High response time for internode communication

An overloaded long-distance link causes high response times in the internode messages that are sent by SAN Volume Controller. 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.

- Storage metrics (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.

- Storage metrics (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 cluster 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 cluster.

In this situation, ask the following questions:

- Are the storage controllers at the remote cluster overloaded (pilfering slowly)?

Use Tivoli Storage Productivity Center to obtain the back-end write response time for each MDisk at the remote cluster. 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.

Tip: Any of the MDisk on the remote back-end storage controller that are providing poor response times can be the underlying cause of a 1920 error. For example, the response prevents application I/O from proceeding at the rate that is required by the application host and the **gmlinktolerance** parameter is issued, which causes the 1920 error.

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.

SAN Volume Controller node hardware

For the SAN Volume Controller node hardware, the possible cause of the 1920 errors might be from a heavily loaded primary cluster. If the nodes at the primary cluster are heavily loaded, the internal Global Mirror lock sequence messaging between nodes (which is used to assess the added effect of running Global Mirror) exceeds the `gm_max_host_delay` parameter (default 5 ms). If this condition persists, a 1920 error is posted.

Important: For analysis of a 1920 error, contact your IBM service support representative (SSR) regarding the effect of the SAN Volume Controller node hardware and loading.

Level 3 Engagement is the highest level of support. It provides analysis of SAN Volume Controller clusters for overloading.

Use Tivoli Storage Productivity Center and I/O stats to check the following areas:

- ▶ Port-to-local node send response time
- ▶ Port-to-local node send queue time

A high response (>1 ms) indicates a high load, which is a possible contribution to a 1920 error.

- ▶ SAN Volume Controller node CPU utilization

An excess of 50% is higher than average loading and a possible contribution to a 1920 error.

SAN Volume Controller volume states

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.

7.10.3 Recovery

After a 1920 error occurs, the Global Mirror auxiliary VDisks 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 VDisks on the secondary cluster is inconsistent, and your applications cannot use the VDisks as backup disks.

Tip: If the relationship stopped in a consistent state, you can use the data on the auxiliary volume at the remote cluster 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.

Restarting after a 1920 error

Example 7-2 shows a script to help restarts Global Mirror consistency groups and relationships that stopped after a 1920 error was issued.

Example 7-2 Script for restarting Global Mirror

```
svcinfolsrcconsistgrp -filtervalue state=consistent_stopped -nohdr -delim : | while IFS=: read id name mci
mci mcni aci acn p state junk; do
    echo "Restarting group: $name ($id)"
    svctask startsrcconsistgrp -force $name
    echo "Clearing errors..."
    svcinfo lsrcerrlogbyrcconsistgrp -unfixed $name | while read id type fixed snmp err_type node seq_num
junk; do
    if [ "$id" != "id" ]; then
        echo "Marking $seq_num as fixed"
        svctask cherrstate -sequencenumber $seq_num
    fi
done
done

svcinfolsrcrelationship -filtervalue state=consistent_stopped -nohdr -delim : | while IFS=: read id name
mci mcni mvi mvn aci acn avi avn p cg_id cg_name state junk; do
    if [ "$cg_id" == "" ]; then
        echo "Restarting relationship: $name ($id)"
        svctask startsrcrelationship -force $name
        echo "Clearing errors..."
        svcinfo lsrcerrlogbyrcrelationship -unfixed $name | while read id type fixed snmp err_type node
seq_num junk; do
            if [ "$id" != "id" ]; then
                echo "Marking $seq_num as fixed"
                svctask cherrstate -sequencenumber $seq_num
            fi
        done
    fi
done
done
```

7.10.4 Disabling the gmlinktolerance feature

You can 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 VDisks.
- ▶ 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.

7.10.5 Cluster error code 1920 checklist for diagnosis

Metro Mirror (remote copy) stops because of a persistent I/O error. This error might be caused by problems on the following components:

- ▶ Primary cluster (including primary storage)
- ▶ Secondary cluster (including auxiliary storage)
- ▶ ICL

The problem might occur for the following reasons:

- ▶ A component failure.
- ▶ A component that becomes unavailable or that features reduced performance because of a service action.
- ▶ The decreased performance of a component to a level where the Metro Mirror or Global Mirror relationship cannot be maintained.
- ▶ A change in the performance requirements of the applications that use Metro Mirror or Global Mirror.

This error is reported on the primary cluster when the copy relationship is not progressing sufficiently over a period. Therefore, if the relationship is restarted before all of the problems are fixed, the error might be reported again when the period expires. The default period is 5 minutes.

Use the following checklist as a guide to diagnose and correct the error or errors:

- ▶ On the primary cluster that reports the error, correct any higher priority errors.
- ▶ On the secondary cluster, review the maintenance logs to determine whether the cluster was operating with reduced capability at the time the error was reported. The reduced capability might be because of a software upgrade, hardware maintenance to a 2145 node, maintenance to a back-end disk subsystem, or maintenance to the SAN.
- ▶ On the secondary 2145 cluster, correct any errors that are not fixed.
- ▶ On the ICL, review the logs of each link component for any incidents that might cause reduced capability at the time of the error. Ensure that the problems are fixed.
- ▶ On the primary and secondary cluster that report the error, examine the internal I/O stats.
- ▶ On the ICL, examine the performance of each component by using an appropriate SAN productivity monitoring tool to ensure that they are operating as expected. Resolve any issues.

7.11 Monitoring remote copy relationships

You monitor your remote copy relationships by using Tivoli Storage Productivity Center. For information about a process that uses Tivoli Storage Productivity Center, see Chapter 13, “Monitoring” on page 357.

To ensure that all SAN components perform correctly, use a SAN performance monitoring tool. Although a SAN performance monitoring tool is useful in any SAN environment, it is important when you use an asynchronous mirroring solution, such as Global Mirror for SAN Volume Controller. You must gather performance statistics at the highest possible frequency.

If your VDisk or MDisk configuration changed, restart your Tivoli Storage Productivity Center performance report to ensure that performance is correctly monitored for the new configuration.

If you are using Tivoli Storage Productivity Center, monitor the following information:

- ▶ Global Mirror secondary write lag
You monitor the Global Mirror secondary write lag to identify mirror delays.
- ▶ Port-to-remote node send response
Time must be less than 80 ms (the maximum latency that is supported by SAN Volume Controller Global Mirror). A number in excess of 80 ms suggests that the long-distance link has excessive latency, which must be rectified. One possibility to investigate is that the link is operating at maximum bandwidth.
- ▶ 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 cluster. 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.
- ▶ CPU utilization percentage
CPU utilization must be below 50%.
- ▶ Sum of Back-end write response time and Write queue time for Global Mirror MDisk at the *remote cluster*
The time must be less than 100 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.
- ▶ Sum of Back-end write response time and Write queue time for Global Mirror MDisk at the *primary cluster*
Time must also be less than 100 ms. If response time is greater than 100 ms, the application hosts might see extended response times if the cache of the SAN Volume Controller becomes full.
- ▶ Write data rate for Global Mirror managed disk groups at the remote cluster
This data rate indicates the amount of data that is being written by Global Mirror. If this number approaches the ICL bandwidth or the storage controller throughput limit, further increases can cause overloading of the system. Therefore, monitor this number appropriately.

Hints and tips for Tivoli Storage Productivity Center statistics collection

Analysis requires Tivoli Storage Productivity Center Statistics (CSV) or SAN Volume Controller Raw Statistics (XML). You can export statistics from your Tivoli Storage Productivity Center instance. Because these files become large quickly, you can limit this situation. For example, you can filter the statistics files so that individual records that are below a certain threshold are not exported.

Default naming convention: IBM Support has several automated systems that support analysis of Tivoli Storage Productivity Center data. These systems rely on the default naming conventions (file names) that are used. The default name for Tivoli Storage Productivity Center files is `StorageSubsystemPerformance ByXXXXXX.csv`, where `XXXXXX` is the I/O group, managed disk group, MDisk, node, or volume.



Hosts

You can monitor host systems that are attached to the SAN Volume Controller 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.

The most important part of tuning, troubleshooting, and performance is the host that is attached to a SAN Volume Controller. 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 SAN Volume Controller V7.2 Information Center and Guides, which are available at this website:

<http://pic.dhe.ibm.com/infocenter/svc/ic/index.jsp>

This chapter includes the following sections:

- ▶ Configuration guidelines
- ▶ Host pathing
- ▶ I/O queues
- ▶ Multipathing software
- ▶ Host clustering and reserves
- ▶ AIX hosts
- ▶ Virtual I/O Server
- ▶ Windows hosts
- ▶ Linux hosts
- ▶ Solaris hosts
- ▶ VMware server
- ▶ Mirroring considerations
- ▶ Monitoring

8.1 Configuration guidelines

When the SAN Volume Controller 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.

8.1.1 Host levels and host object name

When a new host is configured to the SAN Volume Controller, 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 SAN Volume Controller at the preferred levels, see *V7.2 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller*, S1004453, which is available at this website:

<http://www-01.ibm.com/support/docview.wss?uid=ssglS1004453>

When you are creating the host, use the host name from the host as the host object name in the SAN Volume Controller to aid in configuration updates or problem determination in the future. If multiple hosts share an identical set of disks, you can create them with a single host object with multiple ports (worldwide port names [WWPNs]) or as multiple host objects.

8.1.2 The number of paths

Based on our general experience, it is best to limit the total number of paths from any host to the SAN Volume Controller. 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 *V7.2.0 Configuration Limits and Restrictions for IBM Storwize V7000*, S1004510, which is available at this website:

<http://www-01.ibm.com/support/docview.wss?uid=ssglS1004510>

The most important reason to limit the number of paths that are available to a host from the SAN Volume Controller 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 SAN Volume Controller releases, host configuration information is available by using the *IBM System Storage SAN Volume Controller 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 7.2 and earlier, this information is now consolidated into the IBM System Storage SAN Volume Controller Information Center, which is available at this website:

<http://publib.boulder.ibm.com/infocenter/svc/ic/index.jsp>

We measured the effect of multipathing on performance, as shown in Table 8-1. As the table shows, the differences in performance are minimal, but the differences can reduce performance by almost 10% for specific workloads. These numbers were produced with an AIX host that is running IBM Subsystem Device Driver (SDD) against the SAN Volume Controller. The host was tuned specifically for performance by adjusting queue depths and buffers.

We tested a range of reads and writes, random and sequential, cache hits, and misses at transfer sizes of 512 bytes, 4 KB, and 64 KB.

Table 8-1 shows the effects of multipathing in IBM System Storage SAN Volume Controller 2145-8G4.

Table 8-1 Effect of multipathing on write performance

Read/write test	Four paths	Eight paths	Difference
Write Hit 512 b Sequential IOPS	81 877	74 909	-8.6%
Write Miss 512 b Random IOPS	60 510.4	57 567.1	-5.0%
70/30 R/W Miss 4 K Random IOPS	130 445.3	124 547.9	-5.6%
70/30 R/W Miss 64 K Random MBps	1 810.8138	1 834.2696	1.3%
50/50 R/W Miss 4 K Random IOPS	97 822.6	98 427.8	0.6%
50/50 R/W Miss 64 K Random MBps	1 674.5727	1 678.1815	0.2%

Although these measurements were taken with SAN Volume Controller 2145-8G4, hardware and software performance does change release to release, and the figures that are shown in Table 8-1 provide an example of the difference that multipathing can make.

8.1.3 Host ports

When you are using host ports that are connected to the SAN Volume Controller, limit the number of physical ports to two ports on two different physical adapters. Each port is zoned to one target port in each SAN Volume Controller 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 SAN Volume Controller 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 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.

8.1.4 Port masking

You can use a port mask to control the node target ports that a host can access. The port mask applies to logins from the host port that are associated with the host object. You can use this capability to simplify the switch zoning by limiting the SAN Volume Controller ports within the SAN Volume Controller configuration, rather than using direct one-to-one zoning within the switch. This capability can simplify zone management.

The port mask is a 4-bit field that applies to all nodes in the cluster for the particular host. For example, a port mask of 0001 allows a host to log in to a single port on every SAN Volume Controller node in the cluster, if the switch zone also includes host ports and SAN Volume Controller node ports.

8.1.5 Host to I/O group mapping

An *I/O grouping* consists of two SAN Volume Controller 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 SAN Volume Controller 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 V7.2 Configuration Limits and Restrictions for IBM Storwize V7000, S1004510, which is available at this website:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004510>

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 Tivoli Storage Productivity Center, to determine whether the host is I/O group-limited. If more I/O groups are needed for the 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 SAN Volume Controller 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.

8.1.6 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.

8.1.7 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 SAN Volume Controller 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 SAN Volume Controller, 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 8-1 shows the syntax of the **lshostvdiskmap** command that is used to determine the SCSI ID and the WWPN of volumes.

Example 8-1 The lshostvdiskmap command

```
svcinflshostvdiskmap -delim
```

Example 8-2 shows the results of using the **lshostvdiskmap** command.

Example 8-2 Output of using the lshostvdiskmap command

```
svcinflsvdiskhostmap -delim : EEXCLS_HBin01 id:name:SCSI_id:host_id:host_name:wwpn:vdisk_UID
950:EEXCLS_HBin01:14:109:HDMCENTEX1N1:10000000C938CFDF:600507680191011D480000000000466
950:EEXCLS_HBin01:14:109:HDMCENTEX1N1:10000000C938D01F:600507680191011D480000000000466
950:EEXCLS_HBin01:13:110:HDMCENTEX1N2:10000000C938D65B:600507680191011D480000000000466
950:EEXCLS_HBin01:13:110:HDMCENTEX1N2:10000000C938D3D3:600507680191011D480000000000466
950:EEXCLS_HBin01:14:111:HDMCENTEX1N3:10000000C938D615:600507680191011D480000000000466
950:EEXCLS_HBin01:14:111:HDMCENTEX1N3:10000000C938D612:600507680191011D480000000000466
950:EEXCLS_HBin01:14:112:HDMCENTEX1N4:10000000C938CFBD:600507680191011D480000000000466
950:EEXCLS_HBin01:14:112:HDMCENTEX1N4:10000000C938CE29:600507680191011D480000000000466
950:EEXCLS_HBin01:14:113:HDMCENTEX1N5:10000000C92EE1D8:600507680191011D480000000000466
950:EEXCLS_HBin01:14:113:HDMCENTEX1N5:10000000C92EDFFE:600507680191011D480000000000466
```

In this example, VDisk 10 has a unique device identifier (UID, which is represented by the UID field) of 6005076801958001500000000000000A (see Example 8-3), but the SCSI_id that host2 used for access is 0.

Example 8-3 VDisk 10 with a UID

```
id:name:SCSI_id:vdisk_id:vdisk_name:wwpn:vdisk_UID
2:host2:0:10:vdisk10:00000000000000ACA:6005076801958001500000000000000A
2:host2:1:11:vdisk11:00000000000000ACA:6005076801958001500000000000000B
2:host2:2:12:vdisk12:00000000000000ACA:6005076801958001500000000000000C
2:host2:3:13:vdisk13:00000000000000ACA:6005076801958001500000000000000D
2:host2:4:14:vdisk14:00000000000000ACA:6005076801958001500000000000000E
```

If you are using IBM multipathing software (SDD or 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 8-4 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 8-4 Host mapping for one host from two I/O groups

```
IBM_2145:ITSOCL1:admin>svcinfolshostvdiskmap 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 8-5 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 8-5 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. For more information, see 8.2.4, “Dynamic reconfiguration” on page 235.

8.1.8 Server adapter layout

If your host system has multiple internal I/O busses, place the two adapters that are used for SAN Volume Controller cluster access on two different I/O busses to maximize the availability and performance.

8.1.9 Availability versus error isolation

Balance availability through the multiple paths by using a SAN to the two SAN Volume Controller nodes, not by using error isolation. Normally, users add more paths to a SAN to increase availability, which leads to the conclusion that you want all four ports in each node zoned to each port on the host. However, based on our experience, it is better to limit the number of paths so that the software error recovery software within a switch or a host can manage the loss of paths quickly and efficiently.

Therefore, it is beneficial to keep the span out from the host port through the SAN to a SAN Volume Controller port to one-to-one as much as possible. Limit each host port to a different set of SAN Volume Controller ports on each node. This approach keeps the errors within a host isolated to a single adapter if the errors come from a single SAN Volume Controller port or from one fabric, which makes isolation to a failing port or switch easier.

8.2 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 SAN Volume Controller nodes that present the volume. Most operating systems present each path to a volume as a separate storage device. Therefore, the SAN Volume Controller, 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.

8.2.1 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 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 the SAN Volume Controller 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-owner node, it must send extra messages to the owner 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. Therefore, performance is enhanced by accessing the volume through the preferred node.

IBM multipathing software (SDD, SDDPCM, or SDDDSM) 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)

8.2.2 Path selection

Multipathing software uses many algorithms to select the paths that are used for an individual I/O for each volume. For enhanced performance with most host types, load balance the I/O between only preferred node paths under normal conditions. The load across the host adapters and the SAN paths is balanced by alternating the preferred node choice for each volume. Use care when you are allocating volumes with the SAN Volume Controller console GUI to ensure adequate dispersion of the preferred node among the volumes. If the preferred node is offline, all I/O goes through the nonpreferred node in write-through mode.

Table 8-2 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 8-2 The 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 8-3 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 8-2).

Table 8-3 The 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 8-4 shows the effect of the use of the nonpreferred paths versus the preferred paths on read performance.

Table 8-4 Random (1 TB) 4 Kb read response time (4.1 nodes, microseconds)

Preferred node (owner)	Nonpreferred node	Delta
5,074	5,147	73

Table 8-5 shows the effect of the use of nonpreferred nodes on write performance.

Table 8-5 Random (1 TB) 4 Kb write response time (4.2 nodes, microseconds)

Preferred node (owner)	Nonpreferred node	Delta
5,346	5,433	87

IBM SDD, SDDDSM, and SDDPCM software recognize the preferred nodes and use the preferred paths.

8.2.3 Path management

The SAN Volume Controller design is based on multiple path access from the host to both SAN Volume Controller 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 (SDD, SDDPCM, and SDDDSM), this monitoring is done by using the **datapath query device** or **pcmpath query device** commands.

Fast node reset

SAN Volume Controller 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 SAN Volume Controller V4.2 and later

When a SAN Volume Controller node is reset, the node ports do *not* disappear from the fabric. Instead, the node keeps the ports alive. From a host perspective, SAN Volume Controller stops responding to any SCSI traffic. Any query to the switch name server finds that the SAN Volume Controller ports for the node are still present, but any FC login attempts (for example, PLOGI) are ignored. This state persists for 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 SAN Volume Controller node.

8.2.4 Dynamic reconfiguration

Many users want to dynamically reconfigure the storage that is connected to their hosts. With the SAN Volume Controller, you can virtualize the storage behind the SAN Volume Controller so that a host sees only the SAN Volume Controller volumes that are presented to it. The host can then add or remove storage dynamically and reallocate it by using volume and MDisk changes.

After you decide to virtualize your storage behind a SAN Volume Controller, complete the following steps:

1. Use image mode migration to move the existing back-end storage behind the SAN Volume Controller. This process is simple, seamless, and requires the host to be gracefully shut down.
2. Rezone the SAN for the SAN Volume Controller to be the host.
3. Move the back-end storage LUNs to the SAN Volume Controller as a host.
4. Rezone the SAN for the SAN Volume Controller as a back-end device for the host.

The LUNs are now managed as SAN Volume Controller image mode volumes. You can then migrate these volumes to new storage or move them to striped storage anytime in the future with no host affect.

However, sometimes users want to change the volume presentation of SAN Volume Controller to the host. Do not change the SAN Volume Controller volume presentation to the host dynamically because this process is error-prone. However, you can change the volume presentation of SAN Volume Controller to the host by remembering several key issues.

Hosts do not dynamically reprobe storage unless they are prompted by an external change or by the users manually, which causes rediscovery. Most operating systems do not notice a change in a disk allocation automatically. Information is saved about the device database, such as the Windows registry or the AIX Object Data Manager (ODM) database that is used.

Adding new volumes or paths

Normally, adding new storage to a host and running the discovery methods (such as the **cfgmgr** command) are safe because no old, leftover information is required to be removed. Scan for new disks, or run the **cfgmgr** command several times if necessary to see the new disks.

Removing volumes and later allocating new volumes to the host

The problem surfaces when a user removes a hostmap on the SAN Volume Controller during the process of removing a volume. After a volume is unmapped from the host, the device becomes unavailable and the SAN Volume Controller reports that no such disk is on this port. The use of the **datapath query device** command after the removal shows a closed, offline, invalid, or dead state, as shown in Example 8-6 and Example 8-7.

Example 8-6 Datapath query device on a Windows host

DEV#:	0	DEVICE NAME:	Disk1 Part0	TYPE:	2145	POLICY:	OPTIMIZED
SERIAL:	60050768018201BEE000000000000041						
=====							
Path#	Adapter/Hard Disk			State	Mode	Select	Errors
0	Scsi	Port2	Bus0/Disk1 Part0	CLOSE	OFFLINE	0	0
1	Scsi	Port3	Bus0/Disk1 Part0	CLOSE	OFFLINE	263	0

Example 8-7 Datapath query device on an AIX host

DEV#: 189 DEVICE NAME: vpath189 TYPE: 2145 POLICY: Optimized
SERIAL: 600507680000009E68000000000007E6
=====

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	fscsi0/hdisk1654	DEAD	OFFLINE	0	0
1	fscsi0/hdisk1655	DEAD	OFFLINE	2	0
2	fscsi1/hdisk1658	INVALID	NORMAL	0	0
3	fscsi1/hdisk1659	INVALID	NORMAL	1	0

The next time that a new volume is allocated and mapped to that host, the SCSI ID is reused if it is allowed to set to the default value. Also, the host can confuse the new device with the old device definition that is still left over in the device database or system memory.

You can get two devices that use identical device definitions in the device database, such as in Example 8-8. Both vpath189 and vpath190 have the same hdisk definitions, but they contain different device serial numbers. The **fscsi0/hdisk1654** path exists in both vpaths.

Example 8-8 vpath sample output

DEV#: 189 DEVICE NAME: vpath189 TYPE: 2145				POLICY:	Optimized
SERIAL: 600507680000009E68000000000007E6					
=====					
Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	fscsi0/hdisk1654	CLOSE	NORMAL	0	0
1	fscsi0/hdisk1655	CLOSE	NORMAL	2	0
2	fscsi1/hdisk1658	CLOSE	NORMAL	0	0

3	fscsi1/hdisk1659	CLOSE	NORMAL	1	0
DEV#:	190	DEVICE NAME:	vpath190	TYPE:	2145
				POLICY:	Optimized
SERIAL:	600507680000009E6800000000007F4				
=====					
Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	fscsi0/hdisk1654	OPEN	NORMAL	0	0
1	fscsi0/hdisk1655	OPEN	NORMAL	6336260	0
2	fscsi1/hdisk1658	OPEN	NORMAL	0	0
3	fscsi1/hdisk1659	OPEN	NORMAL	6326954	0

The multipathing software (SDD) recognizes that a new device is available because it issues an **inquiry** command at configuration time and reads the mode pages. However, if the user did not remove the stale configuration data, the ODM for the old hdisks and vpaths remains and confuses the host because the SCSI ID, not the device serial number mapping, changed. To avoid this situation, remove the hdisk and vpath information from the device configuration database before you map new devices to the host and run discovery, as shown by the commands in the following example:

```
rmdev -dl vpath189
rmdev -dl hdisk1654
```

To reconfigure the volumes that are mapped to a host, remove the stale configuration and restart the host.

Another process that might cause host confusion is expanding a volume. The SAN Volume Controller communicates to a host through the SCSI check condition “mode parameters changed.” However, not all hosts can automatically discover the change and might confuse LUNs or continue to use the old size.

For more information about supported hosts, see *IBM System Storage SAN Volume Controller V6.2.0 - Software Installation and Configuration Guide*, GC27-2286.

8.2.5 Nondisruptive 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 nondisruptively. 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 8-9 shows the Redhat Enterprise Linux 6.5 before I/O group migration. For this example, the Storwize V7000/SAN Volume Controller caching I/O group is io_grp0.

Example 8-9 Native Linux multipath display before I/O group migration

```
[root@RHEL_65 ~]# multipath -ll
mpathb (360050768028100550000000000000fd) dm-2 IBM,2145
size=100G features='1 queue_if_no_path' hwhandler='0' wp=rw
|-+- policy='round-robin 0' prio=50 status=active
|  - 0:0:0:0 sdb 8:16 active ready running
|  - 1:0:0:0 sde 8:64 active ready running
|  - 1:0:1:0 sdf 8:80 active ready running
|  ~ 0:0:7:0 sdi 8:128 active ready running
~+- policy='round-robin 0' prio=10 status=enabled
|  - 0:0:1:0 sdc 8:32 active ready running
|  - 0:0:2:0 sdd 8:48 active ready running
|  - 1:0:2:0 sdg 8:96 active ready running
|  ~ 1:0:3:0 sdh 8:112 active ready running
```

Complete the following steps:

1. Run the following commands to enable VDisk ndvm access for both I/O groups, io_grp0 and io_grp1:

svctask 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 8-10.

Example 8-10 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.
```

For more information about Online rescanning of LUNs on Linux hosts, see this website:

http://pic.dhe.ibm.com/infocenter/storwize/ic/index.jsp?topic=%2Fcom.ibm.storwize.v7000.doc%2Fsvc_linux_onlinerescan.html

3. Validate that the new paths are detected by Redhat Enterprise Linux 6.5, as shown in Example 8-11.

Example 8-11 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 8-12.

Example 8-12 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 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
   |- 0:0:3:0 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
```

8.3 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 SAN Volume Controller) have host adapter parameters and queue depth parameters. Algorithms are also available within multipathing software, such as the `qdepth_enable` attribute.

8.3.1 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 the “Queue depth in Fibre Channel hosts” topic in the IBM SAN Volume Controller Version 6.4 Information Center, which is available at this website:

http://publib.boulder.ibm.com/infocenter/svc/ic/index.jsp?topic=/com.ibm.storage.svc.console.doc/svc_FCqueuedepth.html

You must consider queue depth control for the overall SAN Volume Controller I/O group to maintain performance within the SAN Volume Controller. You must also control it on an individual host adapter basis, LUN basis to avoid taxing the host memory, or physical adapter resources basis. The AIX host attachment scripts define the initial queue depth setting for AIX. Other operating system queue depth settings are specified for each host type in the information center if they are different from the defaults.

For more information, see the “Host attachment” topic in the IBM SAN Volume Controller Version 6.4 Information Center at:

http://publib.boulder.ibm.com/infocenter/svc/ic/index.jsp?topic=/com.ibm.storage.svc.console.doc/svc_hostattachmentmain.html

For AIX host attachment scripts, see the download results for System Storage Multipath Subsystem Device Driver, which is available at this website:

<http://www.ibm.com/support/dlsearch.wss?rs=540&q=host+attachment&tc=ST52G7&dc=D410>

Queue depth control within the host is accomplished by limits that are placed by the adapter resources for handling I/Os and by setting a queue depth maximum per LUN. Multipathing software also controls queue depth by using different algorithms. SDD recently made an algorithm change in this area to limit queue depth individually by LUN, not an overall system queue depth limitation.

The host I/O is converted to MDisk I/O as needed. The SAN Volume Controller submits I/O to the back-end (MDisk) storage as does any host. The host allows user control of the queue depth that is maintained on a disk. SAN Volume Controller controls the queue depth for MDisk I/O without any user intervention. After SAN Volume Controller submits I/Os and has “Q” IOPS outstanding for a single MDisk (waiting for Q I/Os to complete), it does not submit any more I/O until some I/O completes. That is, any new I/O requests for that MDisk are queued inside SAN Volume Controller.

Figure 8-1 shows the effect on host volume queue depth for a simple configuration of 32 volumes and one host.

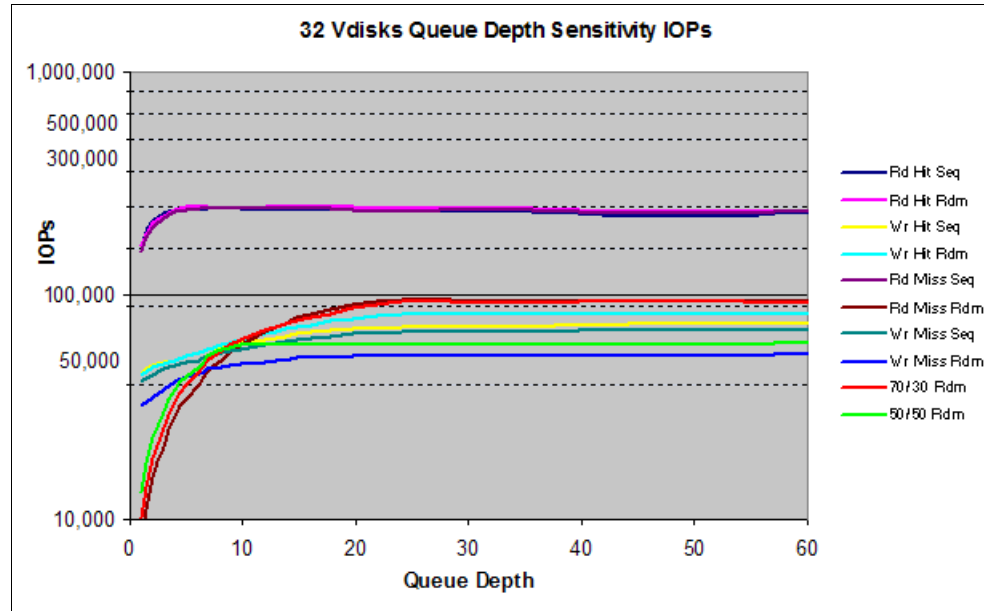


Figure 8-1 IOPS compared to queue depth for 32 volumes tests on a single host in V4.3

Figure 8-2 shows queue depth sensitivity for 32 volumes on a single host.

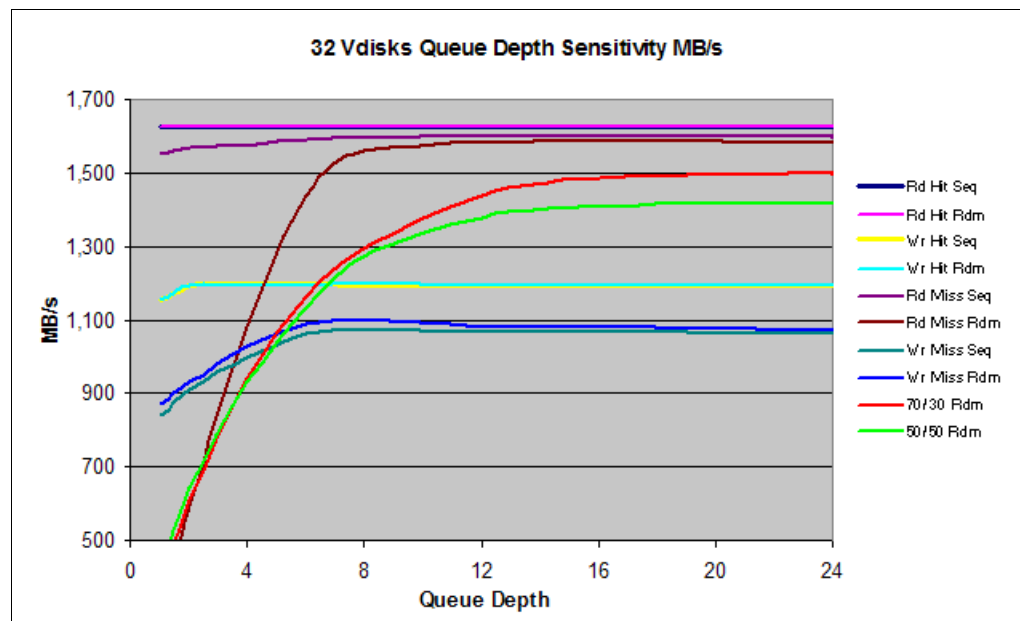


Figure 8-2 MBps compared to queue depth for 32 volume tests on a single host in V4.3

Although these measurements were taken with V4.3 code, the effect that queue depth has on performance is the same regardless of the SAN Volume Controller code version.

8.4 Multipathing software

The SAN Volume Controller 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 *V7.2 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller*, S1004453, which is available at this website:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004453>

Previous preferred levels of host software packages are also tested for SAN Volume Controller V4.3 and allow for flexibility in maintaining the host software levels regarding the SAN Volume Controller software version. Depending on your maintenance schedule, you can upgrade the SAN Volume Controller before you upgrade the host software levels or after you upgrade the software levels.

8.5 Host clustering and reserves

To prevent hosts from sharing storage inadvertently, establish a storage reservation mechanism. The mechanisms for restricting access to SAN Volume Controller 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 the SAN Volume Controller 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.

8.5.1 Clearing reserves

It is possible to accidentally leave a reserve on the SAN Volume Controller volume or on the SAN Volume Controller MDisk during migration into the SAN Volume Controller 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 **lquerypr** (AIX SDD host) and **pcmquerypr** (AIX SDDPCM host) commands. Another tool is a menu-driven Windows SDD or SDDDSM tool.

The Windows Persistent Reserve Tool is called **PRTool.exe** and is installed automatically when SDD or SDDDSM is installed in the `C:\Program Files\IBM\Subsystem Device Driver\PRTool.exe` directory.

You can clear the SAN Volume Controller volume reserves by removing all the host mappings when SAN Volume Controller code is at V4.1 or later.

Example 8-13 shows how to determine whether a reserve is on a device by using the AIX SDD **lquerypr** command on a reserved hdisk.

Example 8-13 The lquerypr command

```
[root@ktazp5033]/reserve-checker-> lquerypr -vVh /dev/hdisk5
connection type: fscsi0
open dev: /dev/hdisk5
Attempt to read reservation key...
Attempt to read registration keys...
Read Keys parameter
      Generation : 935
      Additional Length: 32
      Key0 : 7702785F
      Key1 : 7702785F
      Key2 : 770378DF
      Key3 : 770378DF
Reserve Key provided by current host = 7702785F
Reserve Key on the device: 770378DF
```

Example 8-13 shows that the device is reserved by a different host. The advantage of using the **vV** parameter is that the full persistent reserve keys on the device are shown, in addition to the errors if the command fails.

Example 8-14 shows a failing **pcmquerypr** command to clear the reserve and the error.

Example 8-14 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 2.6.2.1 README FOR AIX*, which is available at this website:

<ftp://ftp.software.ibm.com/storage/subsystem/aix/2.6.2.1/sddpcm.readme.2.6.2.1.txt>

8.5.2 SAN Volume Controller 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 the SAN Volume Controller. You cannot clear a leftover reserve on a SAN Volume Controller MDisk from the SAN Volume Controller. 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.

8.6 AIX hosts

This section describes various topics that are specific to AIX.

8.6.1 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 the SAN Volume Controller 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 hdisk 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_xfr_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.

8.6.2 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.

8.6.3 Multipathing

When the AIX operating system was first developed, multipathing was not embedded within the device drivers. Therefore, each path to a SAN Volume Controller volume was represented by an AIX `hdisk`.

The SAN Volume Controller host attachment script **`devices.fcp.disk.ibm.rte`** sets up the predefined attributes within the AIX database for SAN Volume Controller disks. These attributes changed with each iteration of the host attachment and AIX technology levels. Both SDD and Veritas DMP use the `hdisk`s for multipathing control. The host attachment is also used for other IBM storage devices. The host attachment allows AIX device driver configuration methods to properly identify and configure SAN Volume Controller (2145), IBM DS6000 (1750), and IBM System Storage DS8000 (2107) LUNs.

For more information about supported host attachments for SDD on AIX, see *Host Attachments for SDD on AIX*, S4000106, which is available at this website:

http://www.ibm.com/support/docview.wss?rs=540&context=ST52G7&dc=D410&q1=host+attachment&uid=s5g1S4000106&loc=en_US&cs=utf-8&lang=en

8.6.4 SDD

IBM Subsystem Device Driver multipathing software was designed and updated consistently over the last decade and is a mature multipathing technology. The SDD software also supports many other IBM storage types, such as the 2107, that are directly connected to AIX. SDD algorithms for handling multipathing also evolved. Throttling mechanisms within SDD controlled overall I/O bandwidth in SDD Releases 1.6.1.0 and earlier. This throttling mechanism evolved to be single vpath-specific and is called `qdepth_enable` in later releases.

SDD uses persistent reserve functions and places a persistent reserve on the device in place of the legacy reserve when the volume group is **varyon**. However, if IBM High Availability Cluster Multi-Processing (IBM HACMP™) is installed, HACMP controls the persistent reserve usage, depending on the type of **varyon** used. Also, the enhanced concurrent volume groups have no reserves. The **varyonvg -c** command is for enhanced concurrent volume groups, and **varyonvg** for regular volume groups that use the persistent reserve.

Datapath commands are a powerful method for managing the SAN Volume Controller storage and pathing. The output shows the LUN serial number of the SAN Volume Controller volume and which vpath and hdisk represent that SAN Volume Controller LUN. Datapath commands can also change the multipath selection algorithm. The default is load balance, but the multipath selection algorithm is programmable. When SDD is used, load balance by using four paths. The datapath query device output shows a balanced number of selects on each preferred path to the SAN Volume Controller, as shown in Example 8-15.

Example 8-15 Datapath query device output

DEV#:	12	DEVICE NAME:	vpath12	TYPE:	2145	POLICY:	Optimized
SERIAL:	60050768018B810A88000000000000E0						
=====							
Path#	Adapter/Hard Disk		State	Mode	Select	Errors	
0	fscsi0/hdisk55		OPEN	NORMAL	1390209	0	
1	fscsi0/hdisk65		OPEN	NORMAL	0	0	
2	fscsi0/hdisk75		OPEN	NORMAL	1391852	0	
3	fscsi0/hdisk85		OPEN	NORMAL	0	0	

Verify that the selects during normal operation are occurring on the preferred paths by using the following command:

```
datapath query device -l
```

Also, verify that you have the correct connectivity.

8.6.5 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 SAN Volume Controller 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*, S4000203, which is available at this website:

http://www.ibm.com/support/docview.wss?rs=540&context=ST52G7&dc=D410&q1=host+attachmment&uid=s5g1S4000203&loc=en_US&cs=utf-8&lang=en

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. For SDD, you see the preferred path immediately after you configure it.

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.

8.6.6 SDD compared to SDDPCM

You might choose SDDPCM over SDD for several reasons. SAN boot is much improved with native MPIO-SDDPCM software. Multiple Virtual I/O Servers (VIOs) are supported. Certain applications, such as Oracle ASM, do not work with SDD.

Also, with SDD, all paths can go to the dead state, which improves HACMP and Logical Volume Manager (LVM) mirroring failovers. With SDDPCM, one path always remains open even if the LUN is not available. This design causes longer failovers.

With SDDPCM that is using HACMP, enhanced concurrent volume groups require the *no reserve policy* for concurrent and non-concurrent resource groups. Therefore, HACMP uses a software locking mechanism instead of implementing persistent reserves. HACMP that is used with SDD uses persistent reserves that are based on the type of `varyonvg` that was run.

SDDPCM pathing

SDDPCM `pcmpath` commands are the best way to understand configuration information about the SAN Volume Controller storage allocation. Example 8-16 shows the amount of information that can be determined from the `pcmpath query device` command about the connections to the SAN Volume Controller from this host.

Example 8-16 The `pcmpath query device` command

DEV#:	0	DEVICE NAME:	hdisk0	TYPE:	2145	ALGORITHM:	Load Balance
SERIAL:	6005076801808101400000000000037B						
=====							
Path#	Adapter/PathName	State	Mode	SelectErrors			
0	fscsi0/path0	OPEN	NORMAL	155009	0		
1	fscsi1/path1	OPEN	NORMAL	155156	0		

In this example, both paths are used for the SAN Volume Controller connections. These counts are not the normal select counts for a properly mapped SAN Volume Controller, and two paths is an insufficient number of paths. Use the `-1` option on the `pcmpath query device` command to check whether these paths are both preferred paths. If they are preferred paths, one SAN Volume Controller node must be missing from the host view.

The use of the `-1` option shows an asterisk on both paths, which indicates that a single node is visible to the host (and is the nonpreferred node for this volume), as shown in the following examples:

0*	fscsi0/path0	OPEN	NORMAL	9795 0
1*	fscsi1/path1	OPEN	NORMAL	9558 0

This information indicates a problem that must be corrected. If zoning in the switch is correct, perhaps this host was rebooted when one SAN Volume Controller node was missing from the fabric.

Veritas

Veritas DMP multipathing is also supported for the SAN Volume Controller. Veritas DMP multipathing requires certain AIX APARS and the Veritas Array Support Library (ASL). It also requires a certain version of the host attachment script `devices.fcp.disk.ibm.rte` to recognize the 2,145 devices as hdisks rather than MPIIO hdisks. In addition to the normal ODM databases that contain hdisk attributes, the following Veritas file sets contain configuration data:

- ▶ `/dev/vx/dmp`
- ▶ `/dev/vx/rdmp`
- ▶ `/etc/vxX.info`

Storage reconfiguration of volumes that are presented to an AIX host require cleanup of the AIX hdisks and these Veritas file sets.

8.7 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 the SAN Volume Controller) 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 hdisk cannot be served up from multiple VIOSs. LV VSCSI hdisks are in LVM volume groups on the VIOS and cannot span PVs in that volume group or be striped LVs. Because of these restrictions, use PV VSCSI hdisks.

Multipath support for SAN Volume Controller attachment to Virtual I/O Server is provided by SDD or MPIIO with SDDPCM. Where Virtual I/O Server SAN Boot or dual Virtual I/O Server configurations are required, only MPIIO with SDDPCM is supported. Because of this restriction with the latest SAN Volume Controller-supported levels, use MPIIO with SDDPCM.

For more information, see *V6.2 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller*, which is available at this website:

https://www.ibm.com/support/docview.wss?uid=ssg1S1003797#_VIOS

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 is, on a physical SCSI device (LUN) with user data on it that is in a SAN environment, can this device be allocated to a VIOS and then provisioned to a client partition and used by the client “as is”?

The answer is *no*. This function is not supported as of this writing. 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.

8.7.1 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.

8.7.2 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 SDD 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.

8.7.3 Backing up the virtual I/O configuration

Complete the following steps to back up the virtual I/O configuration:

1. Save the volume group information from the virtual I/O client (PVIDs and volume group names).
2. Save off the disk mapping, PVID, and LUN ID information from *all* VIOSs. In this step, you map the VIOS hdisk (typically, a hdisk) to the virtual I/O client hdisk and you save at least the PVID information.
3. Save off the physical LUN to host LUN ID information about the storage subsystem for when you reconfigure the hdisk (typically).

After all the pertinent mapping data is collected and saved, you can back up and reconfigure your storage and then restore it by using AIX commands.

Back up the volume group data on the virtual I/O client.

For rootvg, the supported method is a **mksysb** and an installation, or **savevg** and **restvg** for nonrootvg.

8.8 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>

8.8.1 Clustering and reserves

Windows SDDDSM uses the persistent reserve functions to implement Windows clustering. A stand-alone Windows host does not use reserves.

For more information about how a cluster works, see the Microsoft article “How the Cluster service reserves a disk and brings a disk online”, which is available at this website:

<http://support.microsoft.com/kb/309186/>

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.

8.8.2 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 according to your manufacturer's instructions, see the "Hosts running the Microsoft Windows Server operating system" topic in the IBM SAN Volume Controller Version 7.2 Information Center, which is available at this website:

http://publib.boulder.ibm.com/infocenter/svc/ic/index.jsp?topic=/com.ibm.storage.svc.console.doc/svc_FChostswindows_cover.html

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 8.3.1, "Queue depths" on page 240.

For IBM FASTT FC2-133 (and HBAs that are QLogic based), the queue depth is known as the *execution throttle*, which can be set by using the QLogic SANSurfer tool or in the BIOS of the HBA that is QLogic based by pressing Ctrl+Q during the startup process.

8.8.3 Changing back-end storage LUN mappings dynamically

Unmapping a LUN from a Windows SDD or SDDDSM server and then mapping a different LUN that uses the same SCSI ID can cause data corruption and loss of access. For more information about the reconfiguration procedure, see this website:

http://www-1.ibm.com/support/docview.wss?rs=540&context=ST52G7&uid=ssg1S1003316&loc=en_US&cs=utf-8&lang=en

8.8.4 Guidelines for disk alignment by using Windows with SAN Volume Controller volumes

You can find the preferred settings for best performance with SAN Volume Controller when you use Microsoft Windows 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

8.9 Linux hosts

IBM is transitioning SAN Volume Controller multipathing support from IBM SDD to Linux native DM-MPIO multipathing. Veritas DMP is also available for certain kernels. For more information about which versions of each Linux kernel require SDD, DM-MPIO, and Veritas DMP support, see "V7.2 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller", which is available at this website:

http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004453#_RH60

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-01.ibm.com/support/knowledgecenter/STPVGU_7.3.0/com.ibm.storage.svc.console.730.doc/svc_linux_settings.html?lang=en

8.9.1 SDD compared to DM-MPIO

For more information about the multipathing choices for Linux operating systems, see the white paper, *Considerations and Comparisons between IBM SDD for Linux and DM-MPIO*, from SDD development, 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

8.9.2 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 8.3.1, “Queue depths” on page 240. In addition, the SAN Volume Controller Information Center provides maximums per HBA adapter or type. For more information, see this website:

<http://pic.dhe.ibm.com/infocenter/svc/ic/index.jsp>

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 SAN Volume Controller Information Center, which is available at this website:

http://publib.boulder.ibm.com/infocenter/svc/ic/index.jsp?topic=/com.ibm.storage.svc431.console.doc/svc_linover_1dcv35.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>

8.9.3 I/O Scheduler

There are several I/O schedulers that are included in modern Linux distributions. Changing I/O scheduler can improve overall performance that is based on the workload. There are three types of scheduler: complete fair queuing (CFQ), NOOP (first in, first out), and deadline. NOOP is good for random I/O workloads. Deadline is good for high-throughput applications. It is recommended to choose between NOOP or deadline and not work with CFQ.

I/O scheduler can be applied globally through `/etc/grub.conf` options by using `elevator=<scheduler>`, which applies to all devices. Individual devices can be assigned by using `/sys/block/<device>/queue/scheduler`. To set the deadline scheduler to `/dev/dm-1`, run the following command:

```
# echo deadline > /sys/block/dm-1/queue/scheduler
```

It is recommended to validate with the application manufacture for best recommendations.

8.10 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 “V6.2 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller” at this website:

https://www.ibm.com/support/docview.wss?uid=ssg1S1003797#_Sun58

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.

8.10.1 Solaris MPxIO

SAN boot and clustering support is available for V5.9 and V5.10, depending on the multipathing driver and HBA choices. Support for load balancing of the MPxIO software was included with SAN Volume Controller code level V4.3.

If you want to run MPxIO on your Sun SPARC host, configure your SAN Volume Controller host object with the type attribute set to `tpgs`, as shown in the following example:

```
svctask mkhost -name new_name_arg -hbawwpn wwpn_list -type tpgs
```

In this command, `-type` specifies the type of host. Valid entries are `hpx`, `tpgs`, or `generic`. The `tpgs` option enables an extra target port unit. The default is `generic`.

For more information about configuring MPxIO software for operating system V5.10 and using SAN Volume Controller volumes, see “Administering Multipathing Devices through `mpathadm` Commands” at this website:

http://download.oracle.com/docs/cd/E19957-01/819-0139/ch_3_admin_multi_devices.htm
1

8.10.2 Symantec Veritas Volume Manager

When you are managing IBM SAN Volume Controller 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 SAN Volume Controller requires for error recovery.

Use the commands that are shown in Example 8-17 on page 255 to determine the basic configuration of a Symantec Veritas server.

Example 8-17 Determining the Symantec Veritas server configuration

```
pkginfo -l (lists all installed packages)
showrev -p |grep vxvm (to obtain version of volume manager)
vxddladm listsupport (to see which ASLs are configured)
vxdisk list
vxdmpadm listctrl all (shows all attached subsystems, and provides a type where
possible)
vxdmpadm getsubpaths ctrl=cX (lists paths by controller)
vxdmpadm getsubpaths dmpnodename=cctxdxs2' (lists paths by LUN)
```

The commands that are shown in Example 8-18 and Example 8-19 determine whether the SAN Volume Controller is properly connected and show at a glance, which ASL is used (native DMP ASL or SDD ASL).

Example 8-18 show what you see when Symantec Volume Manager correctly accesses the SAN Volume Controller by using the SDD pass-through mode ASL.

Example 8-18 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 8-19 shows what you see when SAN Volume Controller is configured by using native DMP ASL.

Example 8-19 SAN Volume Controller 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
```

8.10.3 ASL specifics for SAN Volume Controller

For SAN Volume Controller, ASLs are developed by using DMP multipathing or SDD pass-through multipathing. SDD pass-through is described here for legacy purposes only.

8.10.4 SDD pass-through multipathing

For more information about SDD pass-through, see “Veritas Enabled Arrays - ASL for IBM SAN Volume Controller on Veritas Volume Manager 3.5 and 4.0 using SDD (VPATH) for Solaris” at this website:

<http://www.symantec.com/business/support/index?page=content&id=TECH45863>

The use of SDD is no longer supported. Replace SDD configurations with native DMP.

8.10.5 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 **vxctl enable** command. To list the ASLs that are active, enter the following command:

```
vxddladm listsupport
```

8.10.6 Troubleshooting configuration issues

Example 8-20 shows that the appropriate ASL is not installed or the system is enabling the ASL. The key is the enclosure type **OTHER_DISKS**.

Example 8-20 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

8.11 VMware server

To determine the various VMware ESX levels that are supported, see “V7.2 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller” at this website:

https://www.ibm.com/support/docview.wss?uid=s5g1S1003797#_VMVAAI

On this website, you can also find information about the available support in V7.2 of VMware vStorage API for Array Integration (VAAI).

SAN Volume Controller V7.2 adds support for VMware vStorage APIs. SAN Volume Controller 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 SAN Volume Controller Version 7.2 Information Center at this website:

http://publib.boulder.ibm.com/infocenter/svc/ic/index.jsp?topic=/com.ibm.storage.svc.console.doc/svc_vmwrequirements_21layq.html

8.11.1 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:

- ▶ Fixed-path
- ▶ Round-robin

VMware multipathing was improved to use the SAN Volume Controller preferred node algorithms starting with V4.0. Preferred paths are ignored in VMware versions before V4.0. The 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. Both fixed-path and round-robin algorithms are modified with V4.0 and later to honor the SAN Volume Controller preferred node that is discovered by using the **TPGS** command. Path failover is automatic in both cases. If the round-robin algorithm is used, path fallback might not return to a preferred node path. Therefore, manually check pathing after any maintenance or problems occur.

Update: From vSphere version 5.5 and on, VMware multipath driver fully supports SAN Volume Controller/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.

8.11.2 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

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>

For more information about VMware and SAN Volume Controller, VMware storage, and zoning recommendations, HBA settings, and attaching volumes to VMware, see *Implementing the IBM System Storage SAN Volume Controller V7.2*, SG24-7933, which is available at this website:

<http://www.redbooks.ibm.com/abstracts/sg247933.html?Open>

8.12 Mirroring considerations

As you plan how to fully use the various options to back up your data through mirroring functions, consider how to keep a consistent set of data for your application. A consistent set of data implies a level of control by the application or host scripts to start and stop mirroring with host-based mirroring and back-end storage mirroring features. It also implies a group of disks that must be kept consistent.

Host applications have a certain granularity to their storage writes. The data has a consistent view to the host application only at certain times. This level of granularity is at the file system level, not at the SCSI read/write level. The SAN Volume Controller ensures consistency at the SCSI read/write level when its mirroring features are in use. However, a host file system write might require multiple SCSI writes. Therefore, without a method of controlling when the mirroring stops, the resulting mirror can miss a portion of a write and appear to be corrupted. Normally, a database application has methods to recover the mirrored data and to back up to a consistent view, which is applicable if a disaster that breaks the mirror. However, for nondisaster scenarios, you must have a normal procedure to stop at a consistent view for each mirror to easily start the backup copy.

8.12.1 Host-based mirroring

Host-based mirroring is a fully redundant method of mirroring that uses two mirrored copies of the data. Mirroring is done by the host software. If you use this method of mirroring, place each copy on a separate SAN Volume Controller cluster.

Mirroring that is based on SAN Volume Controller is also available. If you use SAN Volume Controller mirrors, ensure that each copy is on a different back-end controller-based managed disk group.

8.13 Monitoring

A consistent set of monitoring tools is available when IBM SDD, 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 8-21 on page 259 shows the output of the **datapath query devstats** command for two devices.

Example 8-21 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 8-22 shows the use of two adapters.

Example 8-22 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
```

8.13.1 Automated path monitoring

In many situations, a host can lose one or more paths to storage. If the problem is isolated to that one host, it might go unnoticed until a SAN issue occurs that causes the remaining paths to go offline. An example is a switch failure or a routine code upgrade, which can cause a loss-of-access event and seriously affect your business.

To prevent this loss-of-access event from occurring, many clients implement automated path monitoring by using SDD commands and common system utilities. For example, a simple command string, such as the following example, in a UNIX system can count the number of paths, as shown in the following example:

```
datapath query device | grep dead | wc
```

You can combine this command with a scheduler, such as **cron**, and a notification system, such as an email, to notify SAN administrators and system administrators if the number of paths to the system changes.

8.13.2 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 **perfmon.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>

AIX on IBM System p has available the following wikis about performance tools for users:

- Performance Monitoring Tools:

<http://www.ibm.com/collaboration/wiki/display/WikiPtype/Performance+Monitoring+Tools>

- nstress:

<http://www.ibm.com/developerworks/wikis/display/WikiPtype/nstress>

Xdd is a tool that is used to measure and analyze disk performance characteristics on single systems or clusters of systems. Thomas M. Ruwart from I/O Performance, Inc. designed this tool to provide consistent and reproducible performance of a sustained transfer rate of an I/O subsystem. Xdd is a command line-based tool that grew out of the UNIX community and was ported to run in Windows environments. Xdd is a no charge software program that is distributed under a GNU General Public License.

The bXdd distribution comes with all the source code that is necessary to install Xdd and the companion programs for the **timeserver** and the **gettime** utility programs.

For more information about how to use these measurement and test tools, see *IBM Midrange System Storage Implementation and Best Practices Guide*, SG24-6363.



Part 2

Performance preferred practices

This part highlights preferred practices for IBM System Storage SAN Volume Controller. It includes the following chapters:

- ▶ Chapter 9, “Performance highlights for SAN Volume Controller V7.2” on page 263
- ▶ Chapter 10, “Back-end storage performance considerations” on page 269
- ▶ Chapter 11, “IBM System Storage Easy Tier function” on page 319
- ▶ Chapter 12, “Applications” on page 339



Performance highlights for SAN Volume Controller V7.2

This chapter highlights the latest performance improvements that are achieved by IBM System Storage SAN Volume Controller code release V7.2, the new SAN Volume Controller node hardware models CG8, and the new SAN Volume Controller Performance Monitoring Tool.

This chapter includes the following sections:

- ▶ SAN Volume Controller continuing performance enhancements
- ▶ FlashSystem 820 Performance
- ▶ FlashSystem 820 Performance
- ▶ Real-Time Performance Monitor

9.1 SAN Volume Controller continuing performance enhancements

Since IBM introduced SAN Volume Controller in May 2003, it continually improved its performance to meet increasing client demands. The SAN Volume Controller hardware architecture, which is based in the IBM eServer™ xSeries servers, allows for fast deployment of the latest technological improvements available, such as multi-core processors, increased memory, faster Fibre Channel interfaces, and optional features.

Table 9-1 lists and compares the main specifications of each SAN Volume Controller node model.

Table 9-1 SAN Volume Controller node models specifications

SAN Volume Controller node model	xSeries model	Processors	Memory	FC Ports and speed	Solid-state drives (SSDs)	iSCSI
4F2	x335	2 Xeon	4 GB	4@2 Gbps	-	-
8F2	x336	2 Xeon	8 GB	4@2 Gbps	-	-
8F4	x336	2 Xeon	8 GB	4@4 Gbps	-	-
8G4	x3550	2 Xeon 5160	8 GB	4@4 Gbps	-	-
8A4	x3250M2	1 dual-core Xeon 3100	8 GB	4@4 Gbps	-	-
CF8	x3550M2	1 quad-core Xeon E5500	24 GB	4@8 Gbps	Up to 4x 146 GB ^a	2x 1 Gbps
CG8	x3550M3	1 quad-core Xeon E5600	24 GB	4@8 Gbps	Up to 4x 146 GB ^a	2x 1 Gbps 2x 10 Gbps ^a
CG8	x3550M3	1 Hexa-core Xeon 5600 2 Hexa-core Xeon 5600 ^b	24 GB 48 GB ^b	4@8 Gbps 8@ 8Gbps ^a	Up to 4x800 GB ^a	2x 1 Gbps 2x 10 Gbps ^a

a. Item is optional. In the CG8 model, a node can have one of the following components: SSDs or 10 Gbps iSCSI or HBA interfaces.

b. Recommended for compression environments.

In January 2012, a SAN Volume Controller with the eight nodes model CG8 that is running code V6.2 delivered 520,043.99 SPC-1 IOPS. For more information about these benchmarks, see the following resources:

- *SPC Benchmark 1 Full Disclosure report: IBM System Storage SAN Volume Controller v6.2 with IBM Storwize V7000 DISK storage:*
http://www.storageperformance.org/benchmark_results_files/SPC-1/IBM/A00113_IBM_SVC-6.2_Storwize-V7000/a00113_IBM_SVC-v6.2_Storwize-V7000_SPC-1_full-disclosure.pdf
- *SPC Benchmark 1 Full Disclosure Report: IBM System Storage SAN Volume Controller V5.1 (6-node cluster with 2 IBM DS8700S):*
http://www.storageperformance.org/benchmark_results_files/SPC-1/IBM/A00087_IBM_DS8700_SVC-5.1-6node/a00087_IBM_DS8700_SVC5.1-6node_full-disclosure-r1.pdf

Also, see the following Storage Performance Council website for the latest published SAN Volume Controller benchmarks:

<http://www.storageperformance.org/home/>

When you consider Enterprise Storage solutions, raw I/O performance is important, but it is not the only consideration. To date, IBM shipped more than 22,500 SAN Volume Controller engines, which run in more than 7,200 SAN Volume Controller systems.

In 2008 and 2009, across the entire installed base, the SAN Volume Controller delivered better than “five nines” (99.999%) availability. For more information about the SAN Volume Controller, see the following IBM SAN Volume Controller website:

<http://www.ibm.com/systems/storage/software/virtualization/svc>

9.2 FlashSystem 820 Performance

IBM FlashSystem storage systems deliver high performance, efficiency, and reliability to various storage environments, which helps to address performance issues with the most important applications and infrastructure. These storage systems can complement or replace traditional hard disk arrays for many business-critical applications that require high performance or low latency. Such applications include online transaction processing (OLTP), business intelligence (BI), online analytical processing (OLAP), virtual desktop infrastructures (VDIs), high-performance computing (HPC), and content delivery solutions (such as cloud storage and video-on-demand).

Known existing flash-based technologies, such as PCIe flash cards, serial-attached SCSI (SAS), or Serial Advanced Technology Attachment (SATA) solid-state drives (SSDs) traditionally are inside individual servers. Such drives are limited in that they deliver more performance capability only to the dedicated applications that are running on the server, and often are limited in capacity.

Hybrid shared storage systems that use flash and spinning disk technology at the same time offer the potential to improve performance for a wide range of tasks. However, in products of this type, the internal resources of the system (bus, PCI adapters, and so on) are shared between SSD drives and spinning disks, which limits the performance that can be achieved by using flash technology.

Because shared data storage devices that are designed around flash technology, IBM FlashSystem storage systems deliver performance beyond that of most traditional arrays, even those that incorporate SSDs or other flash technology. FlashSystem storage systems can also be used as the top tier of storage, alongside traditional arrays in tiered storage architectures, such as SAN Volume Controller or Storwize V7000 storage virtualization platforms that use IBM Easy Tier functionality. Additionally, IBM FlashSystem storage systems have sophisticated reliability features, such as Variable Stripe Redundant Array of Independent Disks (RAID) that often are not present on locally attached flash devices.

The IBM FlashSystem portfolio includes shared flash storage systems, SSD devices that are provided in disk storage systems, and server-based flash devices.

For more information, see the IBM FlashSystem storage systems home page at this website:

<http://www.ibm.com/systems/storage/flash/>

Peak performance measurements

Table 9-2 shows SAN Volume Controller capabilities in the following workloads:

- ▶ 100% read random 4 K
- ▶ 100% write random 4 K
- ▶ 70% read 30% write (70/30) random 4 KB
- ▶ 8 node SAN Volume Controller cluster (cache disabled)
- ▶ 3 FlashSystem 820

Table 9-2 Peak IOPS while maintaining under 1 ms response time

	Peak IOPS under 1 ms	Per SAN Volume Controller IO groups/single 820 peak IOPS under 1 ms
Read Miss IOPS	1,400,000	350,000
Write Miss IOPS	1,000,000	250,000
70/30 Miss IOPS	1,350,000	337,500

The goal of the test is to demonstrate SAN Volume Controller capabilities of a total configuration and per I/O group.

For more information, see *IBM SAN Volume Controller and IBM FlashSystem 820: Best Practices and Performance Capabilities*, REDP-5027, which is available at this website:

<http://www.redbooks.ibm.com/abstracts/redp5027.html>

9.3 Solid-State Drives and Easy Tier

SAN Volume Controller V7.2 radically increased the number of possible approaches you can take with your managed storage. These approaches included introducing the use of SSDs internally to the SAN Volume Controller nodes and in the managed array controllers. They also included introducing Easy Tier to automatically analyze and make the best use of your fastest storage layer.

SSDs are much faster than conventional disks, but also are more costly. SAN Volume Controller node model CF8 already supported internal SSDs in code version 5.1.

For information about the preferred configuration and use of SSDs in SAN Volume Controller V6.2 (installed internally in the SAN Volume Controller nodes or in the managed storage controllers), see the following chapters:

- ▶ Chapter 10, “Back-end storage performance considerations” on page 269
- ▶ Chapter 11, “IBM System Storage Easy Tier function” on page 319
- ▶ Chapter 12, “Applications” on page 339

Tip: This book includes guidance about fine-tuning your existing SAN Volume Controller and extracting optimum performance, in I/Os per second and in ease of management. Many other scenarios are possible that are not described here. If you have a highly demanding storage environment, contact your IBM marketing representative and Storage Techline for more guidance. They have the knowledge and tools to provide you with the best-fitting, tailor-made SAN Volume Controller solution for your needs.

9.3.1 Internal SSD redundancy

To achieve internal SSD redundancy with SAN Volume Controller V7.2 if a node failure occurs, a scheme was needed in which the SSDs in one node are mirrored by a corresponding set of SSDs in its partner node. The preferred way to accomplish this task was to define a striped managed disk group to contain the SSDs of a node to support an equal number of primary and secondary VDisk copies. The physical node location of each primary VDisk copy should match with the node assignment of that copy and the node assignment of the VDisk. This arrangement ensures minimal traffic requirements between nodes, and a balanced load across the mirrored SSDs.

SAN Volume Controller V6.2 introduced the use of arrays for the internal SSDs that can be configured according to the use you intend to give them. Table 9-3 shows the possible RAID levels to which you can configure your internal SSD arrays.

Usage information: SAN Volume Controller version 5.1 supports the use of internal SSDs as managed disks, whereas SAN Volume Controller V6.2 uses them as array members. Internal SSDs are *not* supported in SAN Volume Controller V6.1.

To learn about an upgrade approach when SSDs are used in SAN Volume Controller version 5.1, see Chapter 16, “SAN Volume Controller scenarios” on page 555.

Table 9-3 RAID levels for internal SSDs

RAID level (GUI Preset)	What you need	When to use it	For best performance
RAID-0 (Striped)	1 - 4 drives, all in a single node.	When Volume Mirror is on external MDisks.	A pool should contain only arrays from a single I/O group.
RAID-1 (Easy Tier)	2 drives, one in each node of the I/O group.	When Easy Tier is used or both mirrors are on SSDs.	An Easy Tier pool should contain only arrays from a single I/O group. The external MDisks in this pool should be used only by the same I/O group.
RAID-10 (Mirrored)	4 - 8 drives, equally distributed among each node of the I/O group.	When multiple drives are used for a volume.	A pool should contain only arrays from a single I/O group. Preferred over Volume Mirroring.

9.3.2 Performance scalability and I/O groups

Because an SAN Volume Controller cluster handles the I/O of a particular volume by the pair of nodes (I/O group) it belongs to, its performance scalability when adding nodes is linear. That is, under normal circumstances, you can expect a four-node cluster to drive about twice as much I/O or throughput as a two-node cluster. This concept is valid if you do not reach a contention or bottleneck in other components, such as back-end storage controllers or SAN links.

However, try to keep your I/O workload balanced across your SAN Volume Controller nodes and I/O groups as even as possible to avoid situations where one I/O group experiences contention and another has idle capacity. If you have a cluster with different node models, you can expect the I/O group with newer node models to handle more I/O than the other ones, but exactly how much more is unknown. For this reason, try to keep your SAN Volume Controller cluster with similar node models. For more information about various approaches to upgrading, see Chapter 14, “Maintenance” on page 485.

Plan carefully the distribution of your servers across your SAN Volume Controller I/O groups and the volumes of one I/O group across its nodes. Reevaluate this distribution whenever you attach another server to your SAN Volume Controller. Use the Performance Monitoring Tool that is described in 9.4, “Real-Time Performance Monitor” on page 268 to help with this task.

9.4 Real-Time Performance Monitor

SAN Volume Controller code V7.2 includes a Real-Time Performance Monitor window. It displays the main performance indicators, which include CPU usage and throughput at the interfaces, volumes, IP replication, and MDisks. Figure 9-1 shows an example of write to V7000 cluster.

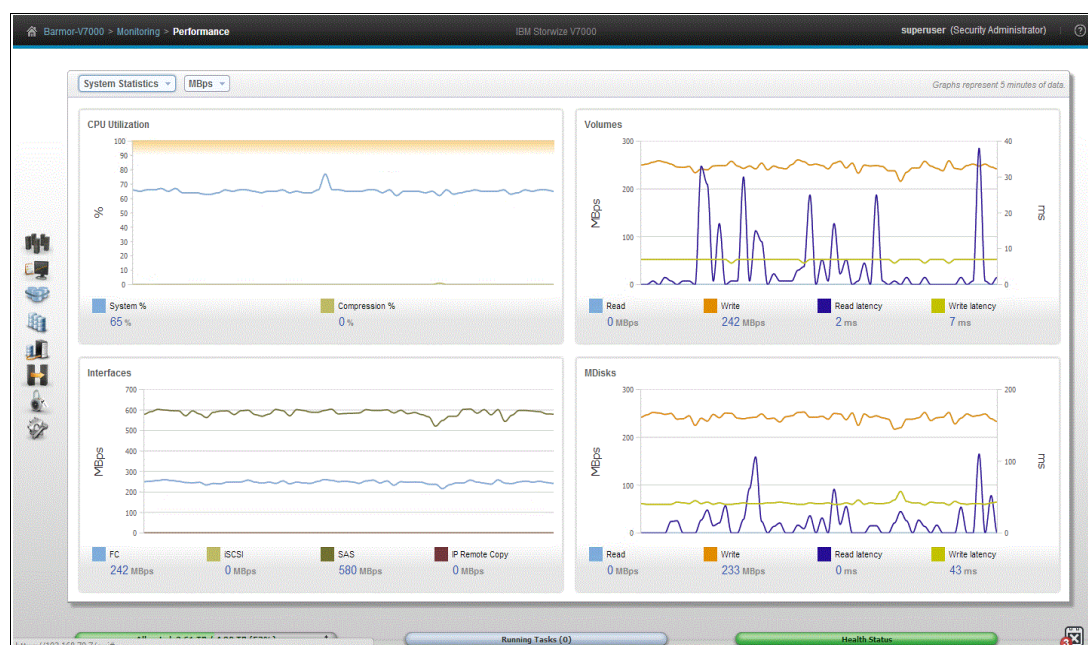


Figure 9-1 Real-Time Performance Monitor example of volume migration

Check this display periodically for possible hot spots that might be developing in your SAN Volume Controller environment. To view this window in the GUI, go to the home page, and select **Performance** in the upper-left menu. The SAN Volume Controller GUI begins plotting the charts. After a few moments, you can view the graphs.

Position your cursor over a particular point in a curve to see details, such as the actual value and time for that point. SAN Volume Controller plots a new point every 5 seconds and it shows you the last 5 minutes of data. You can also change the System Statistics setting in the upper-left corner to see details for a particular node.

The SAN Volume Controller Performance Monitor does not store performance data for later analysis. Instead, its display shows what happened in the last 5 minutes only. Although this information can provide valuable input to help you diagnose a performance problem in real time, it does not trigger performance alerts or provide the long-term trends that are required for capacity planning. For those tasks, you need a tool, such as IBM Tivoli Storage Productivity Center, to collect and store performance data for long periods and present you with the corresponding reports. For more information about this tool, see Chapter 13, “Monitoring” on page 357.



Back-end storage performance considerations

Proper back-end sizing and configuration are essential to achieving optimal performance from the SAN Volume Controller environment.

This chapter describes performance considerations for back-end storage in the IBM System Storage SAN Volume Controller implementation. It highlights the configuration aspects of back-end storage to optimize it for use with the SAN Volume Controller, and examines generic aspects and storage subsystem details.

This chapter includes the following sections:

- ▶ Workload considerations
- ▶ Tiering
- ▶ Storage controller considerations
- ▶ Array considerations
- ▶ I/O ports, cache, and throughput considerations
- ▶ SAN Volume Controller extent size
- ▶ SAN Volume Controller cache partitioning
- ▶ IBM DS8000 series considerations
- ▶ IBM XIV considerations
- ▶ Storwize V7000 considerations
- ▶ DS5000 series considerations
- ▶ Performance considerations with FlashSystems

10.1 Workload considerations

Most applications meet performance objectives when average response times for random I/O are in the range of 2 - 15 milliseconds. However, response-time sensitive applications (typically transaction-oriented) cannot tolerate maximum response times of more than a few milliseconds. You must consider availability in the design of these applications. Be careful to ensure that sufficient back-end storage subsystem capacity is available to prevent elevated maximum response times.

Sizing performance demand: You can use the Disk Magic application to size the performance demand for specific workloads. Disk Magic is available at this website:

<http://www.intellimagic.net>

Batch and OLTP workloads

Clients often want to know whether to mix their batch and online transaction processing (OLTP) workloads in the same managed disk group. Batch and OLTP workloads might require the same tier of storage. However, in many SAN Volume Controller installations, multiple managed disk groups are in the same storage tier so that the workloads can be separated.

You often can mix workloads so that the maximum resources are available to any workload when needed. However, batch workloads are a good example of the opposite viewpoint. A fundamental problem exists with letting batch and online work share resources. That is, the amount of I/O resources that a batch job can use is often limited only by the amount of I/O resources that are available.

To address this problem, it can help to segregate the batch workload to its own managed disk group. However, segregating the batch workload to its own managed disk group does not necessarily prevent node or path resources from being overrun. Those resources might also need to be considered if you implement a policy of batch isolation.

For SAN Volume Controller, an alternative is to cap the data rate at which batch volumes are allowed to run by limiting the maximum throughput of a VDisk. For more information about this approach, see 6.5.1, “Governing of volumes” on page 139. Capping the data rate at which batch volumes are allowed to run can let online work benefit from periods when the batch load is light, and limit the affect when the batch load is heavy.

Much depends on the timing of when the workloads run. If you have mainly OLTP during the day shift and the batch workloads run at night, problems normally do not occur with mixing the workloads in the same managed disk group. If you run the two workloads concurrently and if the batch workload runs with no cap or throttling and requires high levels of I/O throughput, segregate the workloads onto different managed disk groups. The managed disk groups should be supported by different back-end storage resources.

Importance of proper sizing: The SAN Volume Controller can greatly improve the overall capacity and performance usage of the back-end storage subsystem by balancing the workload across parts of it, or across the whole subsystem.

You must size the SAN Volume Controller environment properly on the back-end storage level because virtualizing the environment cannot provide more storage than is available on the back-end storage, especially with cache-unfriendly workloads.

10.2 Tiering

You can use the SAN Volume Controller to create tiers of storage in which each tier has different performance characteristics by including only managed disks (MDisks) that have the same performance characteristics within a managed disk group. Therefore, if you have a storage infrastructure with, for example, three classes of storage, you create each volume from the managed disk group that has the class of storage that most closely matches the expected performance characteristics of the volume.

Because migrating between storage pools (or managed disk groups) is nondisruptive to users, it is easy to migrate a volume to another storage pool if the performance is different than expected.

Tip: If you are uncertain about in which storage pool to create a volume, use the pool with the lowest performance first and then move the volume up to a higher performing pool later, if required.

10.3 Storage controller considerations

Storage virtualization provides greater flexibility in managing the storage environment. In general, you can use storage subsystems more efficiently than when they are used alone. SAN Volume Controller achieves this improved and balanced usage with the use of striping across back-end storage subsystem resources. Striping can be done on the entire storage subsystem, on part of the storage subsystem, or across more storage subsystems.

Tip: Perform striping across back-end disks of the same characteristics. For example, if the storage subsystem has 100 15 K Fibre Channel (FC) drives and 200 7.2 K SATA drives, do not stripe across all 300 drives. Instead, have two striping groups, one with 15 K FC drives and the other with 7.2 SATA drives.

SAN Volume Controller sits in the middle of the I/O path between the hosts and the storage subsystem and acts as a storage subsystem for the hosts. Therefore, it can also improve the performance of the entire environment because of the additional cache usage, which is especially true for cache-friendly workloads.

SAN Volume Controller acts as the host toward storage subsystems. For this reason, apply all standard host considerations. The main difference between the SAN Volume Controller usage of the storage subsystem and the host's usage of it is that, with SAN Volume Controller, only one device is accessing it. With the use of striping, this access provides evenly used storage subsystems. The even utilization of a storage subsystem is achievable only through a proper setup. To achieve even utilization, storage pools must be distributed across all available storage subsystem resources, including drives, I/O buses, and RAID controllers.

The SAN Volume Controller environment can serve to the hosts only the I/O capacity that is provided by the back-end storage subsystems and its internal solid-state drives (SSDs).

10.3.1 Back-end I/O capacity

To calculate what the SAN Volume Controller environment can deliver in terms of I/O performance, you must consider several factors. The following factors are important when the I/O capacity of the SAN Volume Controller back-end is calculated:

► RAID array I/O performance

RAID arrays are created on the storage subsystem as a placement for LUNs that are assigned to the SAN Volume Controller as MDisks. The performance of a particular RAID array depends on the following parameters:

- The type of drives that are used in the array (for example, 15 K FC, 10 K SAS, 7.2 K SATA, SSD)
- The number of drives that are used in the array
- The type of RAID used (that is, RAID 10, RAID 5, RAID 6)

Table 10-1 shows conservative “rule of thumb” numbers for random I/O performance that can be used in the calculations.

Table 10-1 Disk I/O rates

Disk type	Number of input/output operations per second (IOPS)
SSD	20,000
SAS 15 K	180
FC 15 K	160
SAS 10 K	150
FC 10 K	120
NL_SAS 7.2 K	100
SATA 7.2 K	80

The next parameter to consider when you calculate the I/O capacity of a RAID array is the write penalty. Table 10-2 shows the write penalty for various RAID array types.

Table 10-2 RAID write penalty

RAID type	Number of sustained failures	Number of disks	Write penalty
RAID 5	1	N+1	4
RAID 10	Minimum 1	2 x N	2
RAID 6	2	N+2	6

RAID 5 and RAID 6 do not suffer from the write penalty if full stripe writes (also called *stride writes*) are performed. In this case, the write penalty is 1.

With this information and the information about how many disks are in each array, you can calculate the read and write I/O capacity of a particular array.

Table 10-3 shows the calculation for I/O capacity. In this example, the RAID array has eight 15 K FC drives.

Table 10-3 RAID array (eight drives) I/O capacity

RAID type	Read only I/O capacity (IOPS)	Write-only I/O capacity (IOPS)
RAID 5	$7 \times 160 = 1120$	$(8 \times 160)/4 = 320$
RAID 10	$8 \times 160 = 1280$	$(8 \times 160)/2 = 640$
RAID 6	$6 \times 160 = 960$	$(8 \times 160)/6 = 213$

In most of the current generation of storage subsystems, write operations are cached and handled asynchronously, meaning that the write penalty is hidden from the user. However, heavy and steady random writes can create a situation in which write cache destage is not fast enough. In this situation, the speed of the array is limited to the speed that is defined by the number of drives and the RAID array type. The numbers in Table 10-3 on page 273 cover the worst-case scenario and do not consider read or write cache efficiency.

► Storage pool I/O capacity

If you are using a 1:1 LUN (SAN Volume Controller managed disk) to array mapping, the array I/O capacity is already the I/O capacity of the managed disk. The I/O capacity of the SAN Volume Controller storage pool is the sum of the I/O capacity of all managed disks in that pool. For example, if you have 10 managed disks from the RAID arrays with eight disks as used in the example, the storage pool has the I/O capacity as shown in Table 10-4.

Table 10-4 Storage pool I/O capacity

RAID type	Read only I/O capacity (IOPS)	Write-only I/O capacity (IOPS)
RAID 5	$10 \times 1120 = 11200$	$10 \times 320 = 3200$
RAID 10	$10 \times 1280 = 12800$	$10 \times 640 = 6400$
RAID 6	$10 \times 960 = 9600$	$10 \times 213 = 2130$

The I/O capacity of a RAID 5 storage pool ranges from 3200 IOPS when the workload pattern on the RAID array level is 100% write, and 11200 when the workload pattern is 100% read. This workload pattern is caused by a SAN Volume Controller toward the storage subsystem. Therefore, it is not necessarily the same as it is from the host to the SAN Volume Controller because of the SAN Volume Controller cache usage.

If more than one managed disk (LUN) is used per array, each managed disk receives a portion of the array I/O capacity. For example, you have two LUNs per 8-disk array and only one of the managed disks from each array is used in the storage pool. Then, the 10 managed disks have the I/O capacity that is listed in Table 10-5.

Table 10-5 Storage pool I/O capacity with two LUNs per array

RAID type	Read only I/O capacity (IOPS)	Write-only I/O capacity (IOPS)
RAID 5	$10 \times 1120/2 = 5600$	$10 \times 320/2 = 1600$
RAID 10	$10 \times 1280/2 = 6400$	$10 \times 640/2 = 3200$
RAID 6	$10 \times 960/2 = 4800$	$10 \times 213/2 = 1065$

The numbers in Table 10-5 on page 273 are valid if both LUNs on the array are evenly used. However, if the second LUNs on the arrays that are participating in the storage pool are idle storage pool capacity, you can achieve the numbers that are shown in Table 10-4. In an environment with two LUNs per array, the second LUN can also use the entire I/O capacity of the array and cause the LUN used for the SAN Volume Controller storage pool to get less available IOPS.

If the second LUN on those arrays is also used for the SAN Volume Controller storage pool, the cumulative I/O capacity of two storage pools in this case equals one storage pool with one LUN per array.

► **Storage subsystem cache influence**

The numbers for the SAN Volume Controller storage pool I/O capacity that is calculated in Table 10-5 on page 273 did not consider caching on the storage subsystem level, but only the raw RAID array performance.

Similar to the hosts that are using SAN Volume Controller and that have the read/write pattern and cache efficiency in its workload, the SAN Volume Controller also has the read/write pattern and cache efficiency toward the storage subsystem. The following example shows a host-to-SAN Volume Controller I/O pattern:

70:30:50 - 70% reads, 30% writes, 50% read cache hits

Read related IOPS generated from the host IO = Host IOPS x 0.7 x 0.5

Write related IOPS generated from the host IO = Host IOPS x 0.3

Table 10-6 shows the relationship of the host IOPS to the SAN Volume Controller back-end IOPS.

Table 10-6 Host to SAN Volume Controller back-end I/O map

Host IOPS	Pattern	Read IOPS	Write IOPS	Total IOPS
2000	70:30:50	700	600	1300

The total IOPS from Table 10-6 is the number of IOPS that is sent from the SAN Volume Controller to the storage pool on the storage subsystem. Because the SAN Volume Controller is acting as the host toward the storage subsystem, we can also assume that we have some read/write pattern and read cache hit on this traffic.

As shown in Table 10-6, the 70:30 read/write pattern with the 50% cache hit from the host to the SAN Volume Controller is causing an approximate 54:46 read/write pattern from the SAN Volume Controller traffic to the storage subsystem. If you apply the same read cache hit of 50%, you get the 950 IOPS that is sent to the RAID arrays, which are part of the storage pool, inside the storage subsystem, as shown in Table 10-7.

Table 10-7 SAN Volume Controller to storage subsystem I/O map

SAN Volume Controller IOPS	Pattern	Read IOPS	Write IOPS	Total IOPS
1300	54:46:50	350	600	950

I/O considerations: These calculations are valid only when the I/O that is generated from the host to the SAN Volume Controller generates exactly one I/O from the SAN Volume Controller to the storage subsystem. If the SAN Volume Controller is combining several host I/Os to one storage subsystem I/O, higher I/O capacity can be achieved.

Also, I/O with a higher block size decreases RAID array I/O capacity. Therefore, it is possible that combining the I/Os does not increase the total array I/O capacity as viewed from the host perspective. The drive I/O capacity numbers that are used in the preceding I/O capacity calculations are for small block sizes, that is, 4 K - 32 K.

To simplify this example, assume that number of IOPS that is generated on the path from the host to the SAN Volume Controller and from the SAN Volume Controller to the storage subsystem remains the same.

If you assume the write penalty, Table 10-8 shows the total IOPS toward the RAID array for the previous host example.

Table 10-8 RAID array total utilization

RAID type	Host IOPS	SAN Volume Controller IOPS	RAID array IOPS	RAID array IOPS with write penalty
RAID 5	2000	1300	950	$350+4*600 = 2750$
RAID 10	2000	1300	950	$350+2*600 = 1550$
RAID 6	2000	1300	950	$350+6*600 = 3950$

Based on these calculations, we can create a generic formula to calculate available host I/O capacity from the RAID/storage pool I/O capacity. Assume that you have the following parameters:

- ▶ R: Host read ratio (%)
- ▶ W: Host write ratio (%)
- ▶ C1: SAN Volume Controller read cache hits (%)
- ▶ C2: Storage subsystem read cache hits (%)
- ▶ WP: Write penalty for the RAID array
- ▶ XIO: RAID array/storage pool I/O capacity

You can then calculate the host I/O capacity (HIO) by using the following formula:

$$HIO = XIO / (R*C1*C2/1000000 + W*WP/100)$$

The host I/O capacity can be lower than storage pool I/O capacity when the denominator in the preceding formula is greater than 1.

To calculate at which write percentage in I/O pattern (W) the host I/O capacity is lower than the storage pool capacity, use the following formula:

$$W \leq 99.9 / (WP - C1 \times C2/10000)$$

Write percentage (W) mainly depends on the write penalty of the RAID array. Table 10-9 on page 276 shows the break-even value for W with a read cache hit of 50% on the SAN Volume Controller and storage subsystem level.

Table 10-9 W percentage break-even

RAID type	Write penalty (WP)	W percentage break-even
RAID 5	4	26.64%
RAID 10	2	57.08%
RAID 6	6	17.37%

The W percentage break-even value from Table 10-9 is a useful reference about which RAID level to use if you want to maximally use the storage subsystem back-end RAID arrays from the write workload perspective.

With the preceding formulas, you can also calculate the host I/O capacity, for the example storage pool from Table 10-4 on page 273 with the 70:30:50 I/O pattern (read:write:cache hit) from the host side and 50% read cache hit on the storage subsystem.

Table 10-10 shows the results.

Table 10-10 Host I/O example capacity

RAID type	Storage pool I/O capacity (IOPS)	Host I/O capacity (IOPS)
RAID 5	112000	8145
RAID 10	128000	16516
RAID 6	9600	4860

This formula assumes that no I/O grouping is on the SAN Volume Controller level. With SAN Volume Controller code 6.x, the default back-end read and write I/O size is 256 K. Therefore, a possible scenario is that a host might read or write multiple (for example, 8) aligned 32 K blocks from or to the SAN Volume Controller. The SAN Volume Controller might combine this to one I/O on the back-end side. In this situation, the formulas might need to be adjusted. Also, the available host I/O for this particular storage pool might increase.

FlashCopy

The use of FlashCopy on a volume can generate more load on the back-end. When a FlashCopy target is not fully copied or when copy rate 0 is used, the I/O to the FlashCopy target causes an I/O load on the FlashCopy source. After the FlashCopy target is fully copied, read/write I/Os are served independently from the source read/write I/O requests.

The combinations that are shown in Table 10-11 are possible when copy rate 0 is used or the target FlashCopy volume is not fully copied and I/Os are run in an uncopied area.

Table 10-11 FlashCopy I/O operations

I/O operation	Source volume write I/Os	Source volume read I/Os	Target volume write I/Os	Target volume read I/Os
1x read I/O from source	0	1	0	0
1x write I/O to source	1	1	1	0
1x write I/O to source to the already copied area (copy rate > 0)	1	0	0	0
1x read I/O from target	0	1	0	Redirect to the source

I/O operation	Source volume write I/Os	Source volume read I/Os	Target volume write I/Os	Target volume read I/Os
1x read I/O from target from the already copied area copy rate > 0)	0	0	0	1
1x write I/O to target	0	1	1	0
1x write I/O to target to the already copied area copy rate > 0)	0	0	1	0

In some I/O operations, you might experience multiple I/O overheads, which can cause performance degradation of the source and target volume. If the source and the target FlashCopy volume share the back-end storage pool (as shown in Figure 10-1), this situation further influences performance.

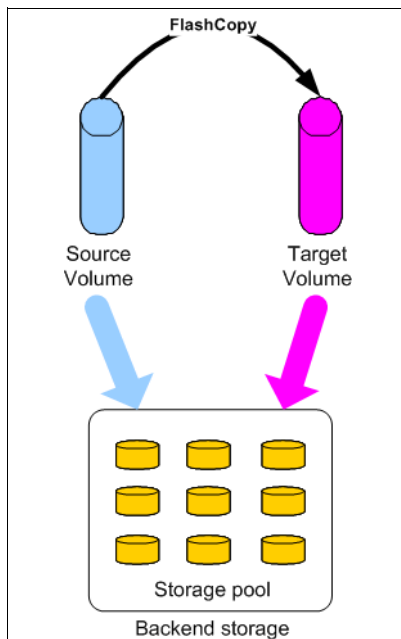


Figure 10-1 FlashCopy source and target volume in the same storage pool

When frequent FlashCopy operations are run and you do not want too much effect on the performance of the source FlashCopy volumes, place the target FlashCopy volumes in a storage pool that does not share the back-end disks. If possible, place them on a separate back-end controller, as shown in Figure 10-2 on page 278.

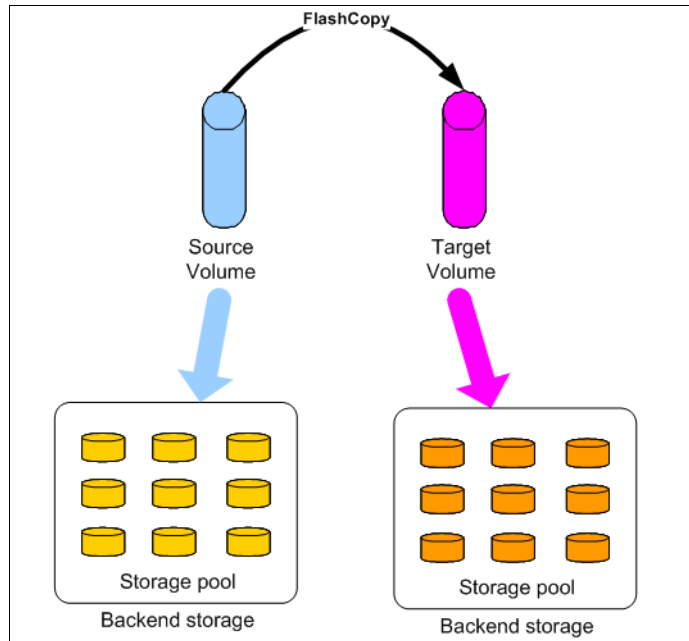


Figure 10-2 Source and target FlashCopy volumes in different storage pools

When you need heavy I/O on the target FlashCopy volume (for example, the FlashCopy target of the database can be used for data mining), wait until FlashCopy copy is completed before the target volume is used.

If volumes that participate in FlashCopy operations are large, the copy time that is required for a full copy is not acceptable. In this situation, use the incremental FlashCopy approach. In this setup, the initial copy lasts longer, and all subsequent copies copy only changes because of the FlashCopy change tracking on source and target volumes. This incremental copying is performed much faster, and it is usually in an acceptable time frame so that you have no need to use target volumes during the copy operation, as shown in Figure 10-3 on page 279.

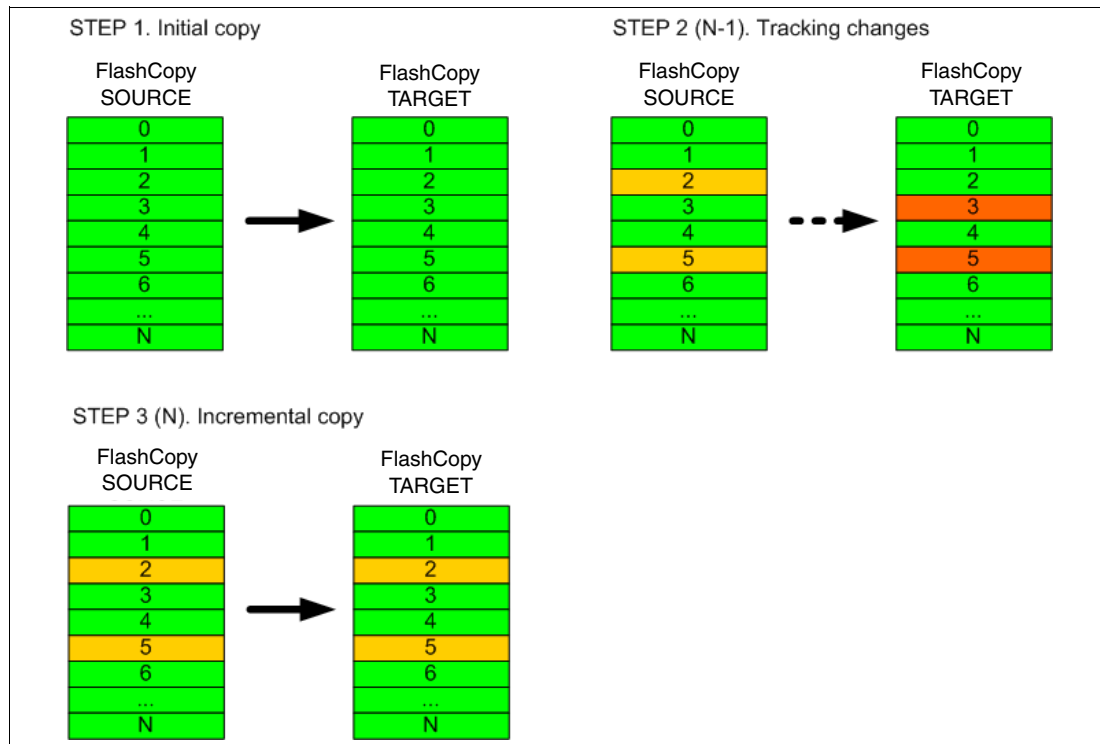


Figure 10-3 Incremental FlashCopy for performance optimization

This approach achieves minimal impact on the source FlashCopy volume.

Thin provisioning

The thin provisioning function also affects the performance of the volume because it generates more I/Os. Thin provisioning is implemented by using a B-Tree directory that is stored in the storage pool, as is the actual data. The real capacity of the volume consists of the virtual capacity and the space that is used for the directory, as shown in Figure 10-4.

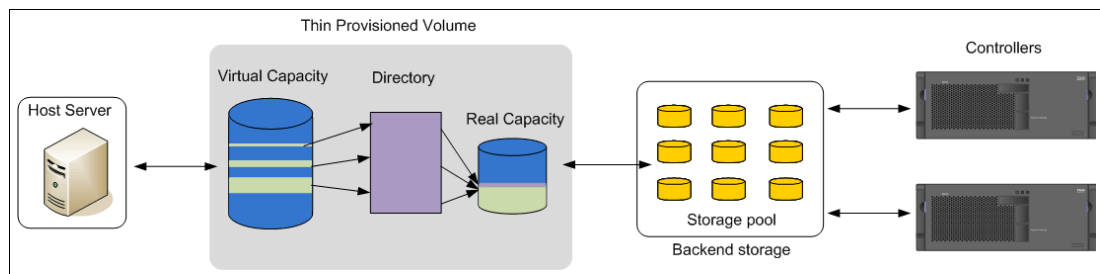


Figure 10-4 Thin provisioned volume

Thin-provisioned volumes can have the following I/O scenarios:

- Write to a region that is not allocated:
 - a. Directory lookup indicates that the region is not allocated.
 - b. The SAN Volume Controller allocates space and updates the directory.
 - c. The data and the directory are written to disk.

- ▶ Write to an allocated region:
 - a. Directory lookup indicates that the region is already allocated.
 - b. The data is written to disk.
- ▶ Read to region that is not allocated (unusual):
 - a. Directory lookup indicates that the region is not allocated.
 - b. The SAN Volume Controller returns a buffer of 0x00s.
- ▶ Read to an allocated region:
 - a. Directory lookup indicates that the region is allocated.
 - b. The data is read from disk.

Single-host I/O requests to the specified thin-provisioned volume can result in multiple I/Os on the back end because of the related directory lookup. Consider the following key elements when you use thin-provisioned volumes:

- ▶ If possible, use striping for all thin provisioned volumes across many back-end disks. If thin-provisioned volumes are used to reduce the number of required disks, striping can also result in a performance penalty on those thin-provisioned volumes.
- ▶ Thin-provisioned volumes require more I/O capacity because of the directory lookups. For truly random workloads, this can generate twice as much workload on the back-end disks.
- ▶ The directory I/O requests are two-way write-back cached, the same as fast-write cache. This means that some applications perform better because the directory lookup is served from the cache.
- ▶ Thin-provisioned volumes require more CPU processing on the SAN Volume Controller nodes, so the performance per I/O group is lower. The rule of thumb is that I/O capacity of the I/O group can be only 50% when only thin-provisioned volumes are used.
- ▶ A smaller grain size can have more influence on performance because it requires more directory I/O.

Use a larger grain size (256 K) for the host I/O where larger amounts of write data are expected.

Thin provisioning and FlashCopy

Thin-provisioned volumes can be used in FlashCopy relations as a space-efficient function that provides the capability for thin provisioned volumes, as shown in Figure 10-5.

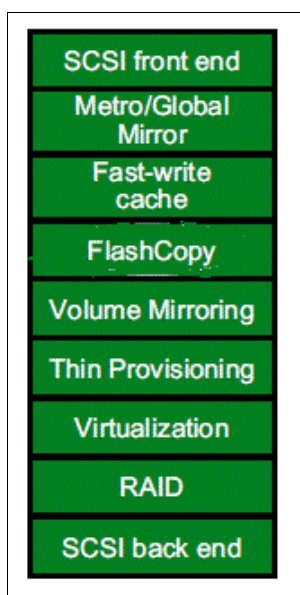


Figure 10-5 SAN Volume Controller I/O facilities

For some workloads, the combination of thin provisioning and the FlashCopy function can significantly affect the performance of target FlashCopy volumes, which is related to the fact that FlashCopy starts to copy the volume from its end. When the target FlashCopy volume is thin provisioned, the last block is physically at the beginning of the volume allocation on the back-end storage, as shown in Figure 10-6.

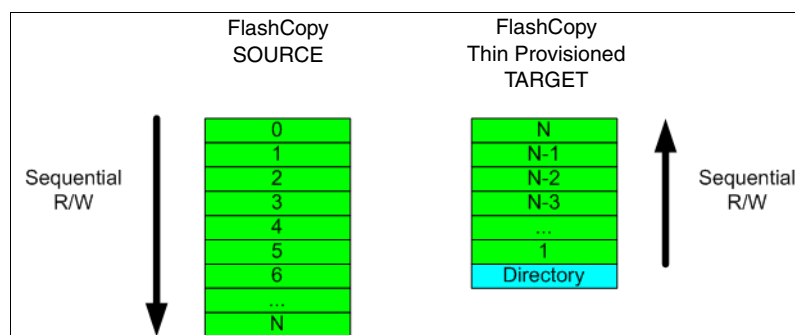


Figure 10-6 FlashCopy thin provisioned target volume

With a sequential workload (as shown in Figure 10-6), the data is on the physical level (back-end storage) read/write from the end to the beginning. In this case, the underlying storage subsystem cannot recognize a sequential operation, which causes performance degradation on that I/O operation.

Considerations for Compressed volumes

Starting with SAN Volume Controller release v6.4.1.5 and v7.1.0.1 and above, there is a new cache destage mode that was introduced that can be enabled for improved performance if the backend storage response time is slow. For more information about this new mode and other preferred practices to follow when compression is used, see Chapter 17, “IBM Real-time Compression” on page 593.

To enable this feature, use the **compressiondestagemode on** command.

Note: The **compressiondestagemode** tunable parameter is also available with v7.2 and we recommend that you work with IBM Support to enable this feature.

Compression on SAN Volume Controller with Storwize V7000 as backend storage

If you have a SAN Volume Controller system setup with Storwize V7000 as backend storage, and if you plan to implement compression, it is recommended to configure compression volumes on the SAN Volume Controller system and not on the backend storage Storwize V7000 system.

10.4 Array considerations

To achieve optimal performance of the SAN Volume Controller environment, you must understand how the array layout is selected.

10.4.1 Selecting the number of LUNs per array

Configure LUNs to use the entire array, which is especially true for midrange storage subsystems where multiple LUNs that are configured to an array result in significant performance degradation. The performance degradation is attributed mainly to smaller cache sizes and the inefficient use of available cache, which defeats the ability of the subsystem to perform “full stride writes” for RAID 5 arrays. Also, I/O queues for multiple LUNs that are directed at the same array tend to overdrive the array.

Higher-end storage controllers, such as the IBM System Storage DS8000 series, make this situation much less of an issue by using large cache sizes. Large array sizes might require the creation of multiple LUNs because of LUN size limitations. However, on higher-end storage controllers, most workloads show the difference between a single LUN per array to be negligible compared to multiple LUNs per array.

For midrange storage controllers, have one LUN per array because it provides the optimal performance configuration. In midrange storage controllers, LUNs are often owned by one controller. One LUN per array minimizes the effect of I/O collision at the drive level. I/O collision can happen with more LUNs per array, especially if those LUNs are not owned by the same controller and when the drive pattern on the LUNs is different.

Consider the manageability aspects of creating multiple LUNs per array configurations. Use care with the placement of these LUNs so that you do not create conditions where over-driving an array can occur. Also, placing these LUNs in multiple storage pools expands failure domains considerably, as described in 5.1, “Availability considerations for storage pools” on page 96.

Table 10-12 on page 283 provides guidelines for array provisioning on IBM storage subsystems.

Table 10-12 Array provisioning

Controller type	LUNs (Managed disks) per array
IBM System Storage DS3000/4000/5000	1
IBM Storwize V7000	1
IBM System Storage DS6000	1
IBM System Storage DS8000	1 - 2
IBM XIV Storage System series	N/A
IBM FlashSystem	N/A

10.4.2 Selecting the number of arrays per storage pool

The capability to stripe across disk arrays is one of the most important performance benefits of the SAN Volume Controller; 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 it is usually difficult to determine what is required in terms of performance, the tendency is to add far too many arrays to a single storage pool, which increases the failure domain as described in 5.1, “Availability considerations for storage pools” on page 96.

It is also worthwhile to consider the effect of an aggregate workload across multiple storage pools. It is clear that striping workload across multiple arrays has a positive effect on performance when you are talking about dedicated resources; however, the performance gains diminish as the aggregate load increases across all available arrays. For example, if you have eight arrays and are striping across all eight arrays, your performance is much better than if you were striping across only four arrays. However, if the eight arrays are divided into two LUNs each and are also included in another storage pool, the performance advantage drops as the load of SP2 approaches that of SP1. When the workload is spread evenly across all storage pools, there is no difference in performance.

More arrays in the storage pool have more of an effect with lower-performing storage controllers because of the cache and RAID calculation constraints. This is true because RAID often is calculated in the main processor, not on the dedicated processors. For example, we require fewer arrays from a DS8000 than we do from a DS5000 to achieve the same performance objectives. This difference is primarily related to the internal capabilities of each storage subsystem and varies based on the workload. Table 10-13 shows the number of arrays per storage pool that is appropriate for general cases. However, there can always be exceptions when it comes to performance.

Table 10-13 Number of arrays per storage pool

Controller type	Arrays per storage pool
IBM System Storage DS3000, DS4000, or DS5000	4 - 24
IBM Storwize V7000	4 - 24
IBM System Storage DS6000	4 - 24
IBM System Storage DS8000	4 - 12
IBM XIV Storage System series	4 - 12
IBM FlashSystem	4 - 16

As shown in Table 10-13 on page 283, the number of arrays per storage pool is smaller in high-end storage subsystems. This number is related to the fact that those subsystems can deliver higher performances per array, even if the number of disks in the array is the same. The performance difference is because of multilayer caching and specialized processors for RAID calculations.

Consider the following points:

- ▶ You must consider the number of MDisk per array and the number of arrays per managed disk group to understand aggregate managed disk group loading effects.
- ▶ You can achieve availability improvements without compromising performance objectives.

Before V6.2 of the SAN Volume Controller code, the SAN Volume Controller cluster used only one path to the managed disk. All other paths were standby paths. When managed disks are recognized by the cluster, active paths are assigned in round-robin fashion. To use all eight ports in one I/O group, at least eight managed disks are needed from a particular back-end storage subsystem. In the setup of one managed disk per array, you need at least eight arrays from each back-end storage subsystem. A new path management was introduced in v6.2. For more information about the new round robin path selection, see 4.2, “Round Robin Path Selection” on page 73.

10.5 I/O ports, cache, and throughput considerations

When you configure a back-end storage subsystem for the SAN Volume Controller environment, you must provide enough I/O ports on the back-end storage subsystems to access the LUNs (managed disks).

The storage subsystem (SAN Volume Controller in this case) must have adequate IOPS and throughput capacities for achieve the appropriate performance level on the host side. Although the SAN Volume Controller greatly improves the utilization of the storage subsystem and increases performance, the back-end storage subsystems must have sufficient capability to handle the load.

The back-end storage must have enough cache for the installed capacity, especially because the write performance greatly depends on a correctly sized write cache.

10.5.1 Back-end queue depth

The SAN Volume Controller submits I/O to the back-end (MDisk) storage in the same fashion as any direct-attached host. For direct-attached storage, the queue depth is tunable at the host and is often optimized based on specific storage type and various other parameters, such as the number of initiators. For the SAN Volume Controller, the queue depth is also tuned; however, the optimal value that is used is calculated internally.

The exact algorithm that is used to calculate queue depth is subject to change. The details that are presented in this section might not stay the same. However, the summary is true of SAN Volume Controller V4.3.

The algorithm features the following parts:

- ▶ A per MDisk limit
- ▶ A per controller port limit, as shown in the following example:

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

where:

- ▶ If $Q > 60$, then $Q=60$ (maximum queue depth is 60)
- ▶ If $Q < 3$, then $Q=3$ (minimum queue depth is 3)

In this algorithm, the following values are used:

- ▶ Q: The queue for any MDisk in a specific controller
- ▶ P: Number of WWPNs that is visible to SAN Volume Controller in a specific controller
- ▶ N: Number of nodes in the cluster
- ▶ M: Number of MDisk that is provided by the specific controller
- ▶ C: A constant. C varies by the following controller types:
 - FASTT = 500
 - EMC CLARiiON = 250
 - DS4700, DS4800, DS6K, and DS8K = 1000
 - XIV Gen1 = 450
 - XIV Gen2 and above = 900
 - Any other controller = 500

When the SAN Volume Controller is submitted and has Q I/Os outstanding for a single MDisk (that is, it is waiting for Q I/Os to complete), it does not submit any more I/O until part of the I/O completes. New I/O requests for that MDisk are queued inside the SAN Volume Controller, which is unwanted and indicates that back-end storage is overloaded.

The following example shows how a 4-node SAN Volume Controller cluster calculates the queue depth for 150 LUNs on a DS8000 storage controller that uses six target ports:

$$Q = ((6 \text{ ports} \times 1000/\text{port})/4 \text{ nodes})/150 \text{ MDisk} = 10$$

With the sample configuration, each MDisk has a queue depth of 10.

SAN Volume Controller V4.3.1 introduced dynamic sharing of queue resources that is based on workload. MDisk with high workload can now borrow unused queue allocation from less-busy MDisk on the same storage system. Although the values are calculated internally and this enhancement provides for better sharing, consider queue depth in deciding how many MDisk to create.

10.5.2 MDisk transfer size

The size of I/O that the SAN Volume Controller performs to the MDisk depends on where the I/O originated.

Host I/O

In SAN Volume Controller versions before V6.x, the maximum back-end transfer size that results from host I/O under normal I/O is 32 KB. If host I/O is larger than 32 KB, it is broken into several I/Os sent to the back-end storage, as shown in Figure 10-7 on page 286. For this example, the transfer size of the I/O is 256 KB from the host side.

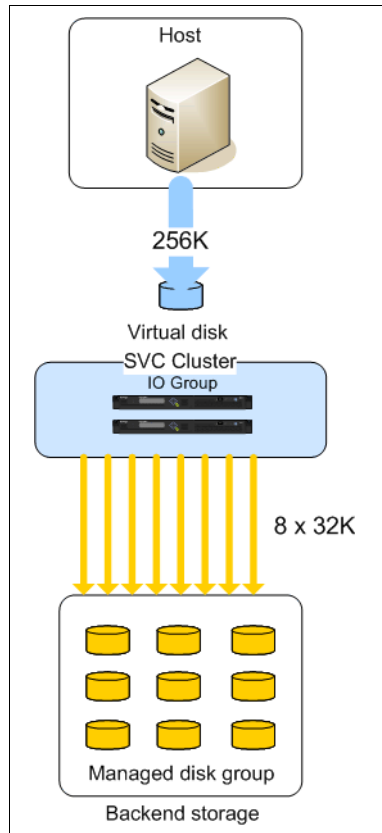


Figure 10-7 SAN Volume Controller back-end I/O before V6.x

In such cases, I/O utilization of the back-end storage ports can be multiplied compared to the number of I/Os coming from the host side. This situation is especially true for sequential workloads where I/O block size tends to be bigger than in traditional random I/O.

To address this situation, the back-end block I/O size for reads and writes was increased to 256 KB in SAN Volume Controller versions 6.x, as shown in Figure 10-8 on page 287.

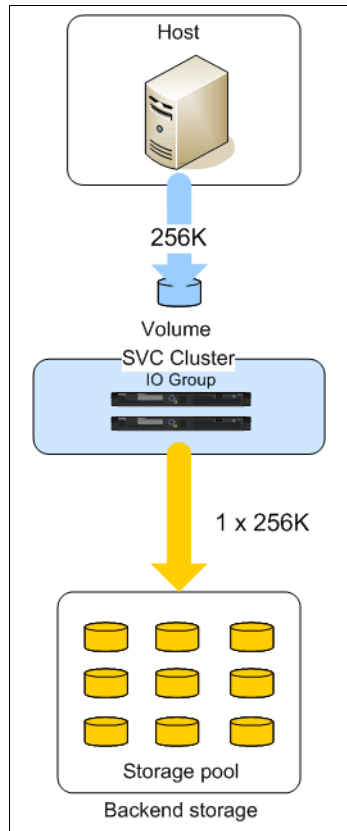


Figure 10-8 SAN Volume Controller back-end I/O with V6.x

Internal cache track size is 32 KB. Therefore, when the I/O comes to the SAN Volume Controller, it is split to the adequate number of the cache tracks. For the preceding example, this number is eight 32 KB cache tracks.

Although the back-end I/O block size can be up to 256 KB, the particular host I/O can be smaller. As such, read or write operations to the back-end managed disks can range 512 bytes - 256 KB. The same is true for the cache because the tracks are populated to the size of the I/O. For example, the 60 KB I/O might fit in two tracks, where first track is fully populated with 32 KB and second track holds only 28 KB.

If the host I/O request is larger than 256 KB, it is split into 256 KB chunks where the last chunk can be partial depending on the size of I/O from the host.

FlashCopy I/O

The transfer size for FlashCopy can be 64 KB or 256 KB for the following reasons:

- ▶ The grain size of FlashCopy is 64 KB or 256 KB.
- ▶ Any size write that changes data within a 64 KB or 256 KB grain results in a single 64-KB or 256-KB read from the source and write to the target.

Thin provisioning I/O

The use of thin provisioning also affects the backed transfer size, which depends on the granularity at which space is allocated. The granularity can be 32 KB, 64 KB, 128 KB, or 256 KB. When grain is initially allocated, it is always formatted by writing 0x00s.

Coalescing writes

The SAN Volume Controller coalesces writes up to the 32-KB track size if writes are in the same tracks before destage. For example, if 4 KB is written into a track, another 4 KB is written to another location in the same track. This track moves to the bottom of the least recently used (LRU) list in the cache upon the second write, and the track now contains 8 KB of actual data. This system can continue until the track reaches the top of the LRU list and is then destaged. The data is written to the back-end disk and removed from the cache. Any contiguous data within the track is coalesced for the destage.

For sequential writes, the SAN Volume Controller does not use a caching algorithm for “explicit sequential detect,” which means coalescing of writes in SAN Volume Controller cache has a random component to it. For example, 4 KB writes to VDisks translates to a mix of 4-KB, 8-KB, 16-KB, 24-KB, and 32-KB transfers to the MDisk, which reduces probability as the transfer size grows.

Although larger transfer sizes tend to be more efficient, this varying transfer size has no effect on the ability of the controller to detect and coalesce sequential content to achieve “full stride writes.”

For sequential reads, the SAN Volume Controller uses “prefetch” logic for staging reads that is based on statistics that are maintained on 128 MB regions. If the sequential content is sufficiently high enough within a region, prefetch occurs with 32 KB reads.

10.6 SAN Volume Controller extent size

The SAN Volume Controller extent size defines the following important parameters of the virtualized environment:

- ▶ Maximum size of the volume
- ▶ Maximum capacity of the single managed disk from the back-end systems
- ▶ Maximum capacity that can be virtualized by the SAN Volume Controller cluster

Table 10-14 lists the possible values with the extent size.

Table 10-14 SAN Volume Controller extent sizes

Extent size (MB)	Maximum non-thin provisioned volume capacity in GB	Maximum thin provisioned volume capacity in GB	Maximum MDisk capacity in GB	Total storage capacity manageable per system
16	2048 (2 TB)	2000	2048 (2 TB)	64 TB
32	4096 (4 TB)	4000	4096 (4 TB)	128 TB
64	8192 (8 TB)	8000	8192 (8 TB)	256 TB
128	16,384 (16 TB)	16,000	16,384 (16 TB)	512 TB
256	32,768 (32 TB)	32,000	32,768 (32 TB)	1 PB
512	65,536 (64 TB)	65,000	65,536 (64 TB)	2 PB
1024	131,072 (128 TB)	130,000	131,072 (128 TB)	4 PB
2048	262,144 (256 TB)	260,000	262,144 (256 TB)	8 PB

Extent size (MB)	Maximum non-thin provisioned volume capacity in GB	Maximum thin provisioned volume capacity in GB	Maximum MDisk capacity in GB	Total storage capacity manageable per system
4096	262,144 (256 TB)	262,144	524,288 (512 TB)	16 PB
8192	262,144 (256 TB)	262,144	1,048,576 (1024 TB)	32 PB

The size of the SAN Volume Controller extent also defines how many extents are used for a particular volume. The example of two different extent sizes that is shown in Figure 10-9 shows that fewer extents are required with a larger extent size.

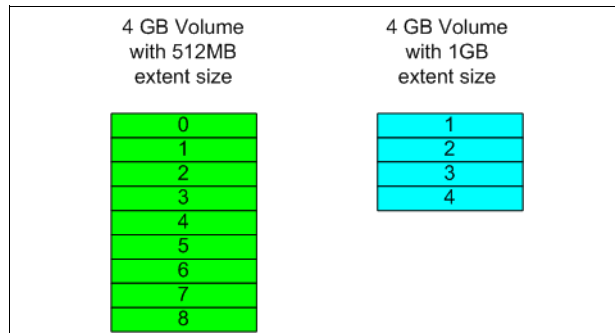


Figure 10-9 Different extent sizes for the same volume

The extent size and the number of managed disks in the storage pool define the extent distribution in case of striped volumes. The example in Figure 10-10 shows two different cases. In one case, the ratio of volume size and extent size is the same as the number of managed disks in the storage pool. In the other case, this ratio is not equal to the number of managed disks.

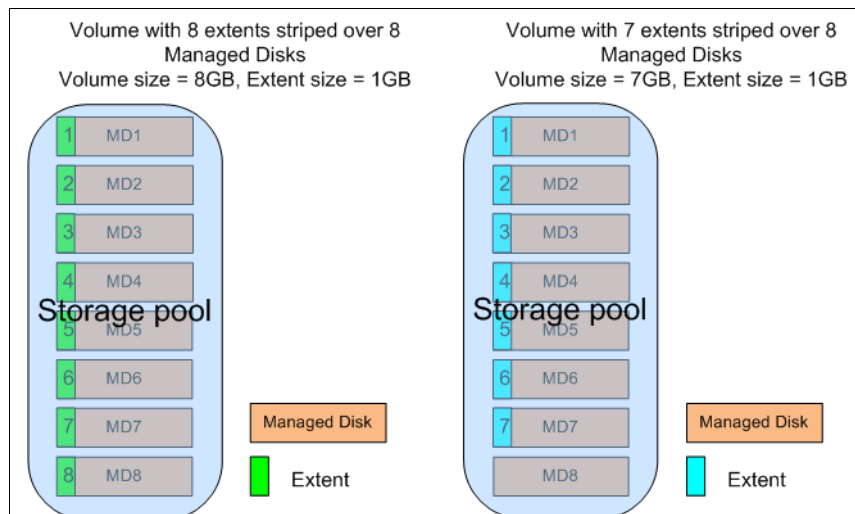


Figure 10-10 SAN Volume Controller extents distribution

For even storage pool utilization, align the size of volumes and extents so that even extent distribution can be achieved. Because the volumes often are used from the beginning of the volume, performance improvements are not gained, which is also valid only for non-thin provisioned volumes.

Tip: Align the extent size to the underlying back-end storage; for example, an internal array stride size (if possible) in relation to the whole cluster size.

10.7 SAN Volume Controller cache partitioning

In a situation where more I/O is driven to a SAN Volume Controller node than can be sustained by the back-end storage, the SAN Volume Controller cache can become exhausted. This situation can occur even if only one storage controller is struggling to cope with the I/O load, but it also affects traffic to others. To avoid this situation, SAN Volume Controller cache partitioning provides a mechanism to protect the SAN Volume Controller cache from overloaded and misbehaving controllers.

The SAN Volume Controller cache partitioning function is implemented on a per storage pool basis. That is, the cache automatically partitions the available resources on a per storage pool basis.

The overall strategy is to protect the individual controller from overloading or faults. If many controllers (or in this case, storage pools) are overloaded, the overall cache can still suffer. Table 10-15 shows the upper limit of write cache data that any one partition (or storage pool) can occupy.

Table 10-15 Upper limit of write cache data

Number of storage pools	Upper limit
1	100%
2	66%
3	40%
4	30%
5 or more	25%

The effect of SAN Volume Controller cache partitioning is that no single storage pool occupies more than its upper limit of cache capacity with write data. Upper limits are the point at which the SAN Volume Controller cache starts to limit incoming I/O rates for volumes that are created from the storage pool.

If a particular storage pool reaches the upper limit, it experiences the same result as a global cache resource that is full. That is, the host writes are serviced on a one-out, one-in basis as the cache destages writes to the back-end storage. However, only writes that are targeted at the full storage pool are limited; all I/O that is destined for other (non-limited) storage pools continues normally.

Read I/O requests for the limited storage pool also continue normally. However, because the SAN Volume Controller is destaging write data at a rate that is greater than the controller can sustain (otherwise, the partition does not reach the upper limit), reads are serviced as slowly.

The key point to remember is that the partitioning is limited only on write I/Os. In general, a 70:30 or 50:50 ratio of read-to-write operations is observed. However, some applications or workloads can perform 100% writes. However, write cache hits are much less of a benefit than read cache hits. A write always hits the cache. If modified data is in the cache, it is overwritten, which might save a single destage operation. However, read cache hits provide a much more noticeable benefit, which saves seek and latency time at the disk layer.

In all benchmarking tests that are performed, even with single active storage pools, good path SAN Volume Controller I/O group throughput is the same as before SAN Volume Controller cache partitioning was introduced.

For information about SAN Volume Controller cache partitioning, see *IBM SAN Volume Controller 4.2.1 Cache Partitioning*, REDP-4426.

10.8 IBM DS8000 series considerations

This section describes SAN Volume Controller performance considerations when the DS8000 series is used as back-end storage.

10.8.1 Volume layout

Volume layout considerations as related to the SAN Volume Controller performance are described in this section.

Ranks-to-extent pools mapping

When you configure the DS8000 series, the following approaches are available for the rank-to-extent pools mapping exist:

- ▶ One rank per extent pool
- ▶ Multiple ranks per extent pool by using DS8000 series storage pool striping

The most common approach is to map one rank to one extent pool. This approach provides good control for volume creation because it ensures that all volume allocation from the selected extent pool come from the same rank.

The storage pool striping feature became available with the R3 microcode release for the DS8000 series. It effectively means that a single DS8000 series volume can be striped across all ranks in an extent pool (the function is often referred as “extent pool striping”). Therefore, if an extent pool includes more than one rank, a volume can be allocated by using free space from several ranks. That is, storage pool striping can be enabled only at volume creation. No reallocation is possible.

The storage pool striping feature requires your DS8000 series layout to be well-planned from the beginning to use all resources in the DS8000 series. If the layout is not well-planned, storage pool striping might cause severe performance problems. An example might be configuring a heavily loaded extent pool with multiple ranks from the same DA pair. Because the SAN Volume Controller stripes across MDisk, the storage pool striping feature is not as relevant here as when it accesses the DS8000 series directly.

Regardless of which approach is used, a minimum of two extent pools must be used to fully and evenly use DS8000 series. A minimum of two extent pools is required to use both servers (server0 and server1) inside the DS8000 series because of the extent pool affinity to those servers.

The decision about which type of ranks-to-extent pool mapping to use depends mainly on the following factors:

- The DS8000 model that is used for back-end storage (DS8100, DS8300, DS8700, or DS8800)
- The stability of the DS8000 series configuration
- The microcode that is installed or can be installed on the DS8000 series

One rank to one extent pool

When the DS8000 series physical configuration is static from the beginning, or when microcode 6.1 or later is not available, use one rank-to-one extent pool mapping. In such a configuration, also define one LUN per extent pool, if possible. The DS8100 and DS8300 do not support LUNs that are larger than 2 TB. If the rank is larger than 2 TB, define more than one LUN on that particular rank. That is, two LUNs might share the back-end disks (spindles), which you must consider for performance planning, as shown in Figure 10-11.

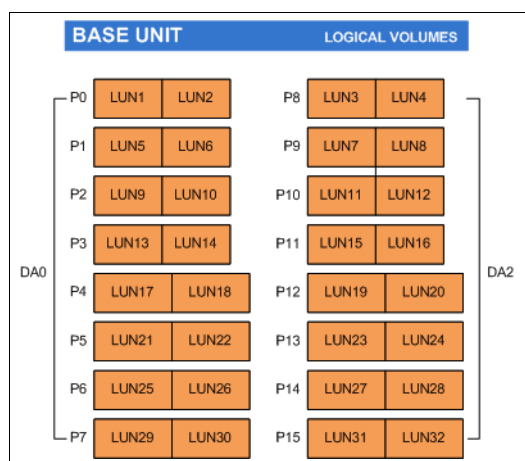


Figure 10-11 Two LUNs per DS8300 rank

The DS8700 and DS8800 models do not have the 2-TB limit. Therefore, use a single LUN-to-rank mapping, as shown in Figure 10-12.

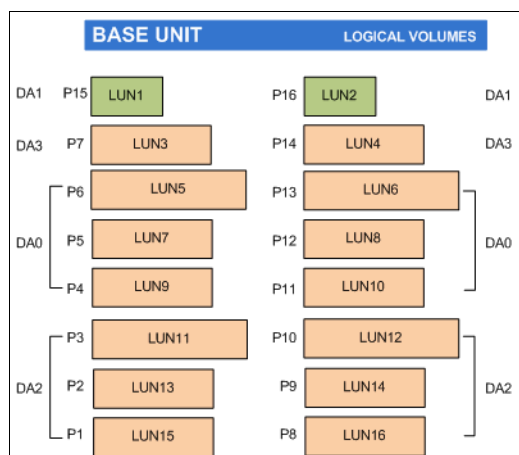


Figure 10-12 One LUN per DS8800 rank

In this setup, we have as many extent pools as ranks, and extent pools might be evenly divided between both internal servers (server0 and server1).

With both approaches, the SAN Volume Controller is used to distribute the workload across ranks evenly by striping the volumes across LUNs.

A benefit of one rank to one extent pool is that physical LUN placement can be easily determined when it is required, such as in performance analysis.

The drawback of such a setup is that, when ranks are added and then integrated into existing SAN Volume Controller storage pools, existing volumes must be restriped manually or by using scripts.

Multiple ranks in one extent pool

When DS8000 series microcode level 6.1 or later is installed or available, and the physical configuration of the DS8000 series changes during the lifecycle (more capacity is installed), use storage pool striping with two extent pools for each disk type. Two extent pools are required to balance the use of processor resources, as shown in Figure 10-13.

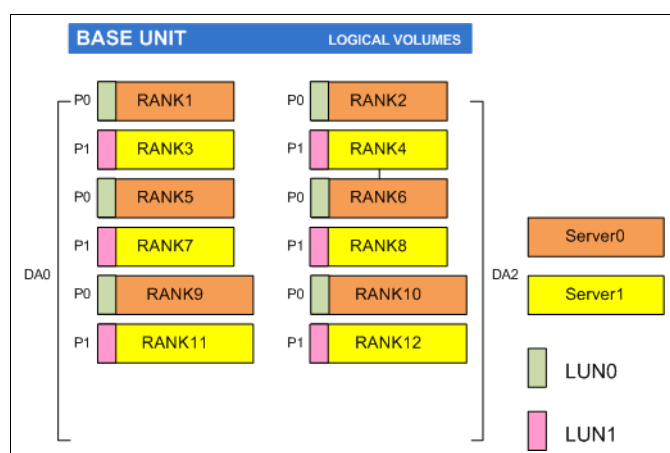


Figure 10-13 Multiple ranks in extent pool

With this design, you must define the LUN size so that each has the same number of extents on each rank (extent size of 1 GB). In the previous example, the LUN might have a size of $N \times 10$ GB. With this approach, the utilization of the DS8000 series on the rank level might be balanced.

If another rank is added to the configuration, the existing DS8000 series LUNs (SAN Volume Controller managed disks) can be rebalanced by using the DS8000 series Easy Tier manual operation so that the optimal resource utilization of DS8000 series is achieved. With this approach, you do not need to restripe volumes on the SAN Volume Controller level.

Extent pools

The number of extent pools on the DS8000 series depends on the rank setup. A minimum of two extent pools is required to evenly use both servers inside DS8000. In all cases, an even number of extent pools provides the most even distribution of resources.

Device adapter pair considerations for selecting DS8000 series arrays

The DS8000 series storage architectures access disks through pairs of device adapters (DA pairs), with one adapter in each storage subsystem controller. The DS8000 series scales from two to eight DA pairs.

When possible, consider adding arrays to storage pools based on multiples of the installed DA pairs. For example, if the storage controller contains six DA pairs, use 6 or 12 arrays in a storage pool with arrays from all DA pairs in a managed disk group.

Balancing workload across DS8000 series controllers

When you configure storage on the IBM System Storage DS8000 series disk storage subsystem, ensure that ranks on a device adapter (DA) pair are evenly balanced between odd and even extent pools. Failing to balance the ranks can result in a considerable performance degradation because of uneven device adapter loading.

The DS8000 series 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. Ranks that belong to an odd-numbered extent pool have an affinity to server1.

Figure 10-14 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.

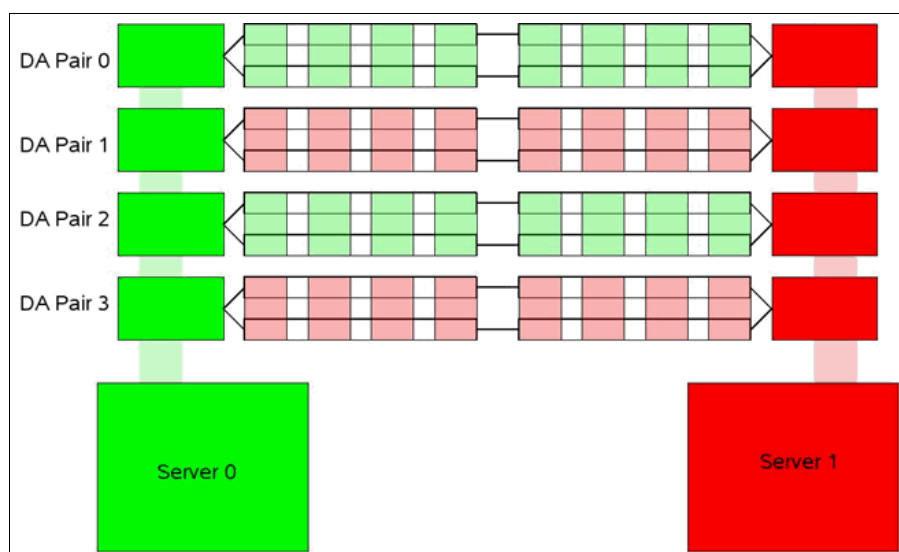


Figure 10-14 DA pair reduced bandwidth configuration

Example 10-1 on page 295 shows what this invalid configuration looks like from the CLI output of the `lsarray` and `lsrank` commands. The arrays that are on the same DA pair contain the same group number (0 or 1), meaning that they have affinity to the same DS8000 series server. Here, server0 is represented by group0, and server1 is represented by group1.

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

Example 10-1 Command output for the lsarray and lsrank commands

```

dscli> lsarray -l
Date/Time: Aug 8, 2008 8:54:58 AM CEST IBM DSCLI Version:5.2.410.299 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: Aug 8, 2008 8:52:33 AM CEST IBM DSCLI Version: 5.2.410.299 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 10-15 shows a correct configuration that balances the workload across all four DA pairs.

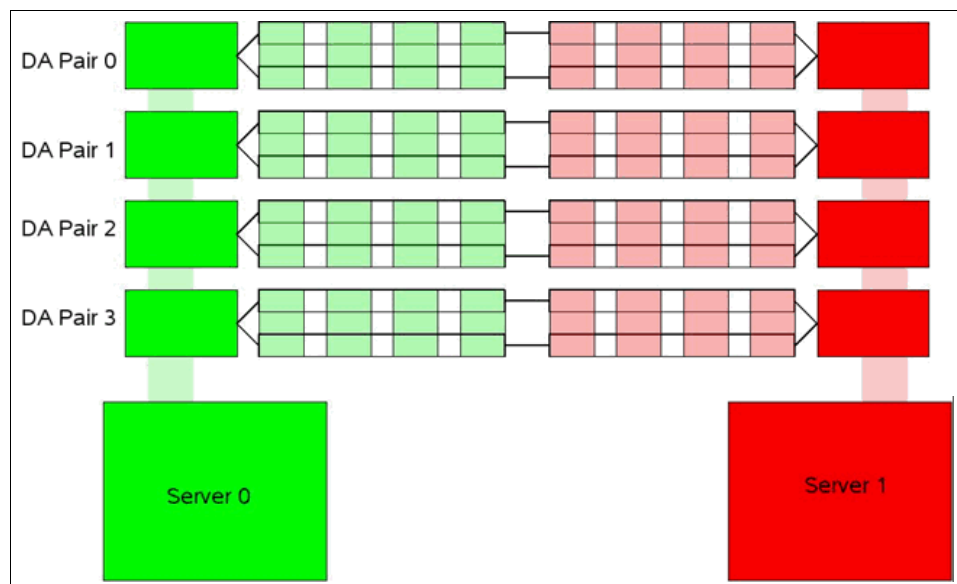


Figure 10-15 DA pair correct configuration

Example 10-2 on page 296 shows how this correct configuration looks from the CLI output of the **lsrank** command. The configuration from the **lsarray** output remains unchanged. Arrays that are on the same DA pair are now split between groups 0 and 1.

Reviewing arrays A0 and A4 again now shows that they have different affinities (A0 - group0, A4 - group1). To achieve this correct configuration (compared to Example 10-1), array A4 now belongs to an odd-numbered extent pool (P5).

Example 10-2 Command output

```
dsccli> lsrank -l
Date/Time: Aug 9, 2008 2:23:18 AM CEST IBM DSCCLI Version: 5.2.410.299 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      1 Normal Normal    A4        5    P5        extpool5   fb     779     779
R5      0 Normal Normal    A5        5    P4        extpool4   fb     779     779
R6      1 Normal Normal    A6        5    P7        extpool7   fb     779     779
R7      0 Normal Normal    A7        5    P6        extpool6   fb     779     779
```

10.8.2 Cache

For the DS8000 series, you cannot tune the array and cache parameters. The arrays are 6+p or 7+p, depending on whether the array site contains a spare and whether the *segment size* (contiguous amount of data that is written to a single disk) is 256 KB for fixed block volumes. Caching for the DS8000 series is done on a 64-KB track boundary.

10.8.3 Determining the number of controller ports for DS8000 series

Configure a minimum of four 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 48 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.

For smaller DS8000 configurations, four controller ports are sufficient.

Also, use no more than two ports of each of the DS8000 series 4-port adapters. When the DS8000 series 8-port adapters are used, use no more than four ports.

Table 10-16 shows the number of DS8000 series ports and adapters that are based on rank count and adapter type.

Table 10-16 Number of ports and adapters

Ranks	Ports	Adapters
2 - 16	8	2 - 4 (2/4-port adapter)
16 - 48	16	4 - 8 (2/4-port adapter), 2-4 (8-port adapter)
> 48	16	8 - 16 (2/4-port adapter), 4-8 (8-port adapter)

The DS8000 series populates Fibre Channel adapters across two to eight I/O enclosures, depending on configuration. Each I/O enclosure represents a separate hardware domain.

Ensure that adapters that are configured to different SAN networks do not share the I/O enclosure as part of our goal of keeping redundant SAN networks isolated from each other.

Figure 10-16 shows an example of DS8800 series connections with 16 I/O ports on eight 8-port adapters. In this case, two ports per adapter are used.



Figure 10-16 DS8800 series with 16 I/O ports

Figure 10-17 shows an example of DS8800 series connections with four I/O ports on two 4-port adapters. In this case, two ports per adapter are used.

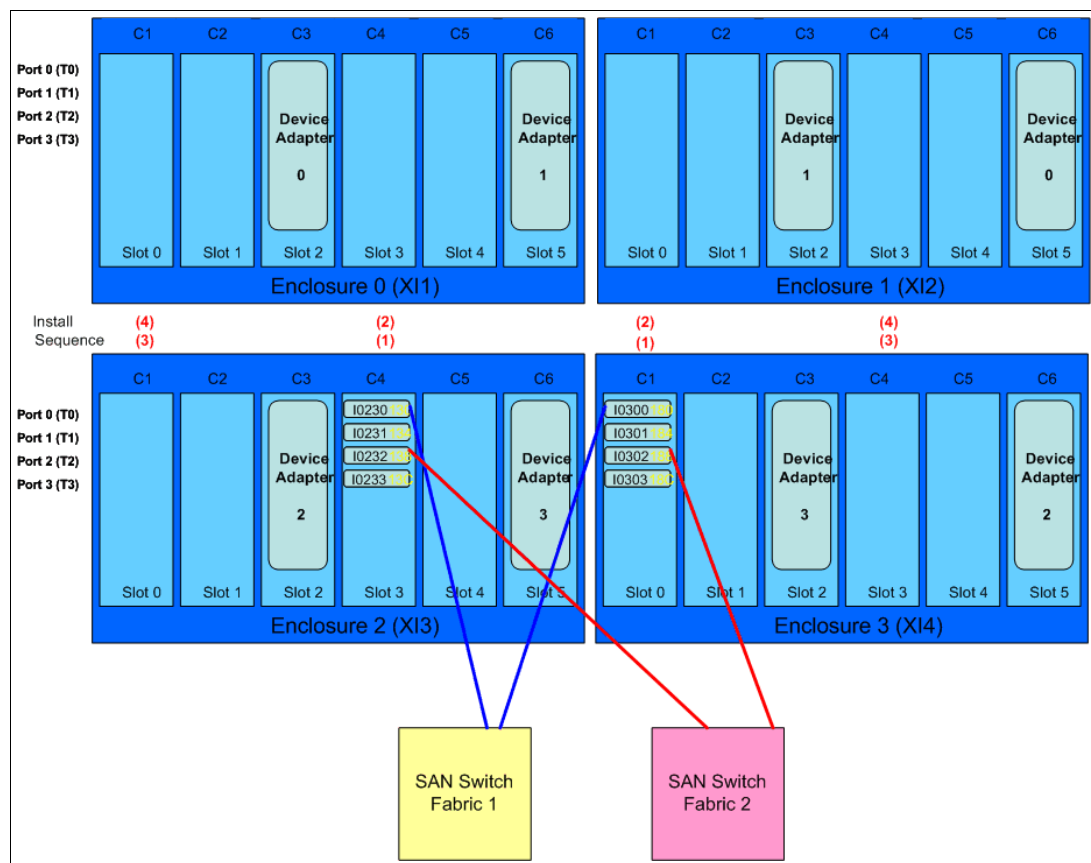


Figure 10-17 DS8000 series with four I/O ports

Preferred practices: Consider the following preferred practices:

- ▶ Configure a minimum of four ports per DS8000 series.
- ▶ Configure 16 ports per DS8000 series when more than 48 ranks are presented to the SAN Volume Controller cluster.
- ▶ Configure a maximum of two ports per 4-port DS8000 series adapter and four ports per 8-port DS8000 series adapter.
- ▶ Configure adapters across redundant SAN networks from different I/O enclosures.

10.8.4 Storage pool layout

The number of SAN Volume Controller storage pools from DS8000 series primarily depends on the following factors:

- ▶ The type of different disks that are installed in the DS8000 series
- ▶ The number of disks in the array:
 - RAID 5: 6+P+S
 - RAID 5: 7+P
 - RAID 10: 2+2+2P+2S
 - RAID 10: 3+3+2P

These factors define the performance and size attributes of the DS8000 series LUNs that act as managed disks for SAN Volume Controller storage pools. The SAN Volume Controller storage pool should have MDisks with the same characteristic for performance and capacity, which is required even for DS8000 series utilization.

Tip: Describe the main characteristics of the storage pool in its name. For example, the pool on DS8800 series with 146 GB 15K FC disks in RAID 5 might have the name DS8800_146G15KFCR5.

Figure 10-18 shows an example of a DS8700 series storage pool layout that is based on disk type and RAID level. In this case, ranks with RAID5 6+P+S and 7+P are combined in the same storage pool, and RAID10 2+2+2P+2S and 3+3+2P are combined in the same storage pool.

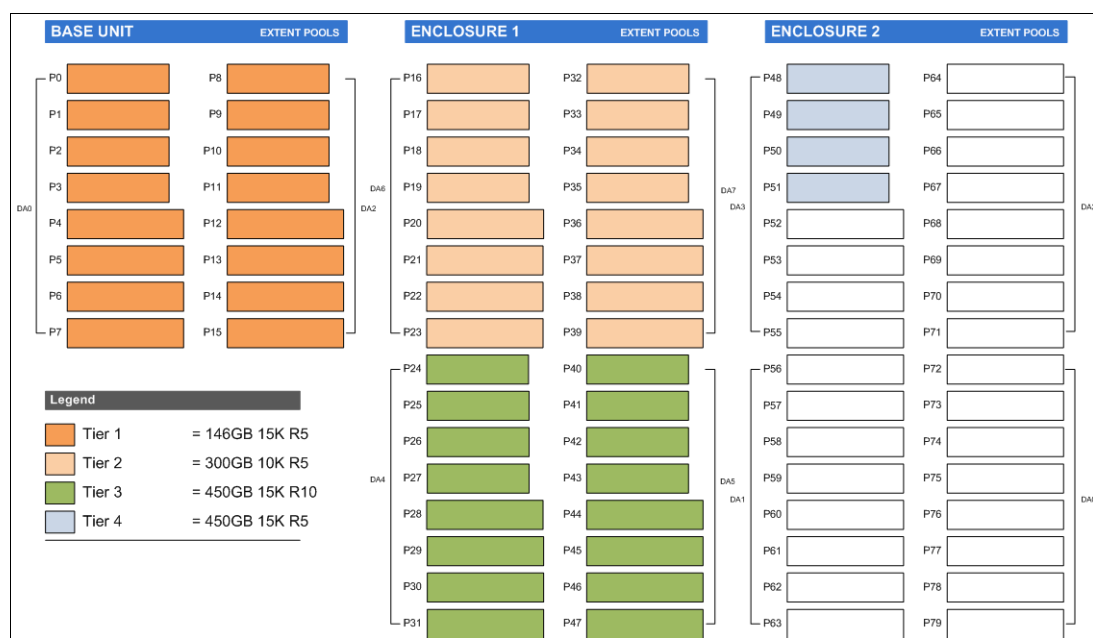


Figure 10-18 DS8700 series storage pools that are based on disk type and RAID level

With this approach, some parts of volumes or some volumes might be striped only over MDs (LUNs) that are on the arrays or ranks where no spare disk is available. Because those MDs have one spindle more, this approach can also compensate for the performance requirements because more extents are placed on them.

Such an approach simplifies the management of the storage pools because it allows for a smaller number of storage pools to be used.

Four storage pools are defined in the following scenario:

- ▶ 145 GB 15K R5 - DS8700_146G15KFCR5
- ▶ 300 GB 10K R5 - DS8700_300G10KFCR5
- ▶ 450 GB 15K R10 - DS8700_450G15KFCR10
- ▶ 450 GB 15K R5 - DS8700_450G15KFCR5

To achieve an optimized configuration from the RAID perspective, the configuration includes storage pools that are based on the number of disks in the array or rank, as shown in Figure 10-19 on page 300.

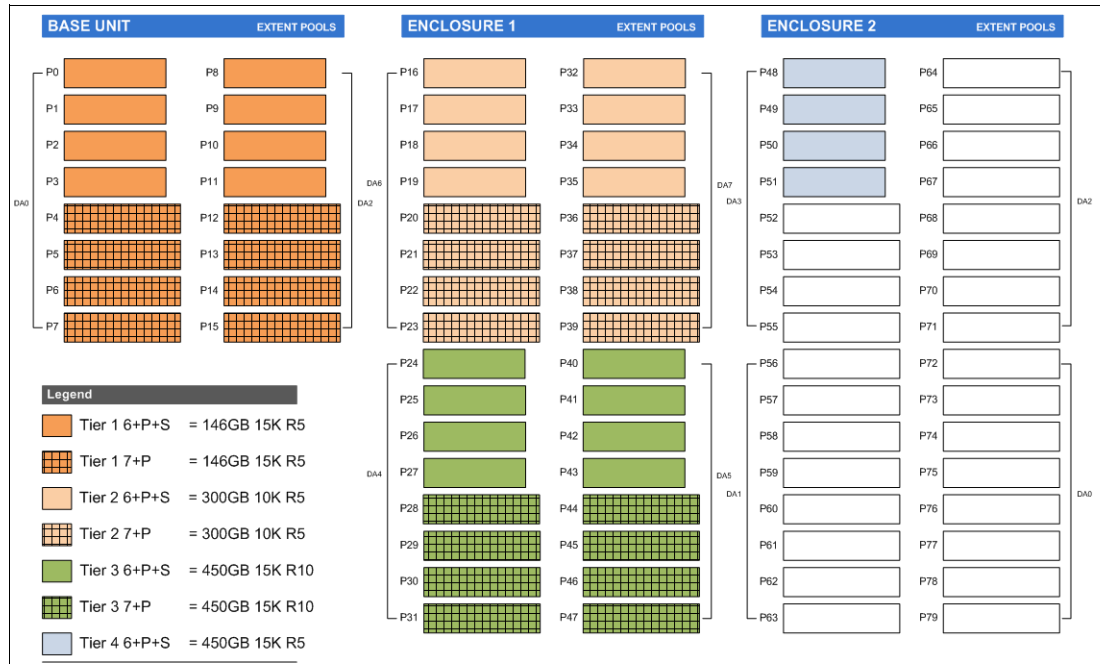


Figure 10-19 DS8700 storage pools with exact number of disks in the array/rank

With this setup, seven storage pools are defined instead of four. The complexity of management increases because more pools must be managed. From the performance perspective, the back end is completely balanced on the RAID level.

Configurations with so many different disk types in one storage subsystem are uncommon. One DS8000 series system often has a maximum of two types of disks, and different types of disks are installed in different systems. Figure 10-20 shows an example of such a setup on DS8800 series.

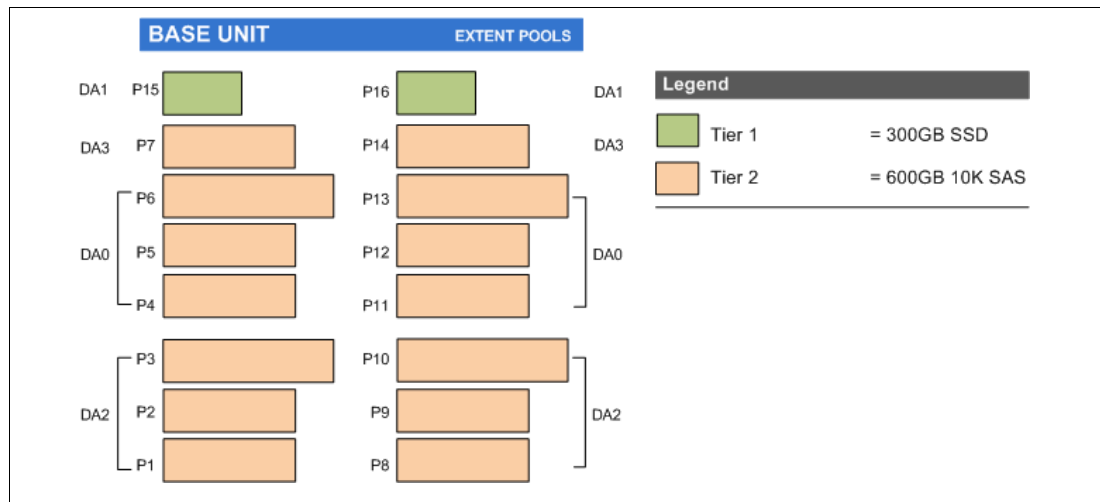


Figure 10-20 DS8800 series storage pool setup with two types of disks

Although it is possible to span the storage pool across multiple back-end systems, as shown in Figure 10-21, keep storage pools bound inside single DS8000 series for availability.

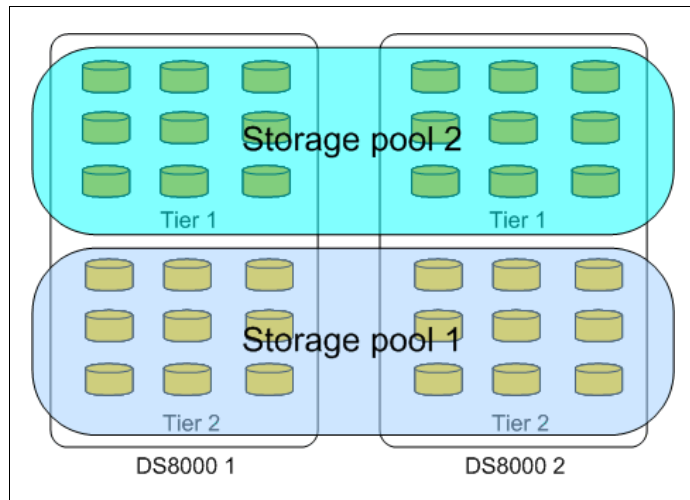


Figure 10-21 DS8000 series spanned storage pool

Preferred practices: Consider the following preferred practices:

- ▶ Use the same type of arrays (disk and RAID type) in the storage pool.
- ▶ Minimize the number of storage pools. If a single type or two types of disks are used, two storage pools can be used per DS8000 series:
 - One for RAID 6+P+S
 - One for RAID 7+P if RAID5 is used

Also, the same for RAID 10 is used with 2+2+2P+2S and 3+3+2P.

- ▶ Spread the storage pool across both internal servers (server0 and server1). Use LUNs from extent pools that have affinity to server0 and those LUNs with affinity to server1 in the same storage pool.
- ▶ Where performance is not the main goal, a single storage pool can be used with mixing LUNs from array with different number of disks (spindles).

Figure 10-22 shows a DS8800 series with two storage pools for 6+P+S RAID5 and 7+P arrays.

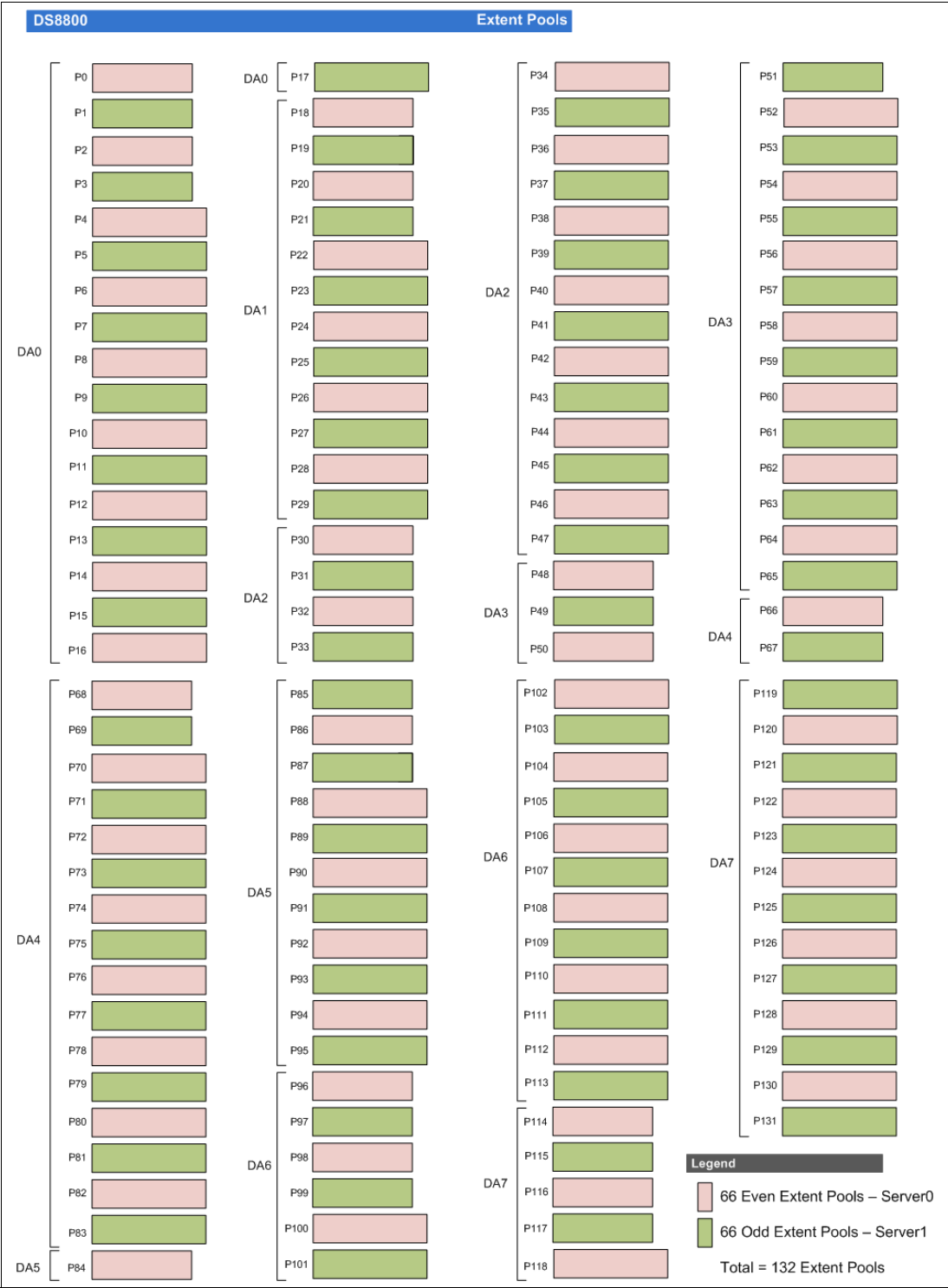


Figure 10-22 Three-frame DS8800 series with RAID 5 arrays

10.8.5 Extent size

Align the extent size with the internal DS8000 series extent size, which is 1 GB. If the SAN Volume Controller cluster size requires a different extent size, this size prevails.

10.9 IBM XIV considerations

This section describes SAN Volume Controller performance considerations when you use the IBM XIV as back-end storage.

10.9.1 LUN size

The main benefit of the XIV Storage System is that all LUNs are distributed across all physical disks. The volume size is the only attribute that is used to maximize the space usage and to minimize the number of LUNs.

The XIV system can grow 6 - 15 installed modules, and it can have 1 TB, 2 TB, or 3 TB disk modules. The maximum LUN size that can be used on the SAN Volume Controller is 2 TB. A maximum of 511 LUNs can be presented from a single XIV system to the SAN Volume Controller cluster. The SAN Volume Controller does not support dynamic expansion of LUNs on the XIV.

Use the following LUN sizes:

- ▶ 1-TB disks: 1632 GB (see Table 10-17)
- ▶ 2-TB disks (Gen3): 1669 GB (see Table 10-18 on page 304)
- ▶ 3-TB disks (Gen3): 2185 GB (see Table 10-19 on page 304)
- ▶ 4-TB disks (Gen3) (see Figure 10-23 on page 305)

Table 10-17, Table 10-18 on page 304, and Table 10-19 on page 304 show the number of managed disks and the available capacity that is based on the number of installed modules.

Table 10-17 XIV with 1-TB disks and 1632-GB LUNs

Number of XIV modules installed	Number of LUNs (MDisks) at 1632 GB each	IBM XIV System TB used	IBM XIV System TB capacity available
6	16	26.1	27
9	26	42.4	43
10	30	48.9	50
11	33	53.9	54
12	37	60.4	61
13	40	65.3	66
14	44	71.8	73
15	48	78.3	79

Table 10-18 lists the data for XIV with 2-TB disks and 1669-GB LUNs (Gen3).

Table 10-18 XIV with 2-TB disks and 1669-GB LUNs (Gen3)

Number of XIV modules installed	Number of LUNs (MDisks) at 1669 GB each	IBM XIV System TB used	IBM XIV System TB capacity available
6	33	55.1	55.7
9	52	86.8	88
10	61	101.8	102.6
11	66	110.1	111.5
12	75	125.2	125.9
13	80	133.5	134.9
14	89	148.5	149.3
15	96	160.2	161.3

Table 10-19 lists the data for XIV with 3-TB disks and 2185-GB LUNs (Gen3).

Table 10-19 XIV with 3-TB disks and 2185-GB LUNs (Gen3)

Number of XIV modules installed	Number of LUNs (MDisks) at 2185 GB each	IBM XIV System TB used	IBM XIV System TB capacity available
6	38	83	84.1
9	60	131.1	132.8
10	70	152.9	154.9
11	77	168.2	168.3
12	86	187.9	190.0
13	93	203.2	203.6
14	103	225.0	225.3
15	111	242.5	243.3

Figure 10-23 shows XIV with 4TB disks (Gen3).

Rack Configuration								
Total number of modules (Configuration type)	6 partial	9 partial	10 partial	11 partial	12 partial	13 partial	14 partial	15 full
Number of data modules	4	5	6	6	7	7	8	9
Number of active interface modules	2	4	4	5	5	6	6	6
Module 9 state		Disabled	Disabled	Enabled	Enabled	Enabled	Enabled	Enabled
Module 8 state		Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Module 7 state		Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Module 6 state	Disabled	Disabled	Disabled	Disabled	Disabled	Enabled	Enabled	Enabled
Module 5 state	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Module 4 state	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
FC ports	8	16	16	20	20	24	24	24
iSCSI ports	6	14	14	18	18	22	22	22
Number of disks	72	108	120	132	144	156	168	180
Usable capacity (1 / 2 / 3 / 4 TB)	28 TB	44 TB	51 TB	56 TB	63 TB	67 TB	75 TB	81 TB
	55 TB	88 TB	102 TB	111 TB	125 TB	134 TB	149 TB	161 TB
	84 TB	132 TB	154 TB	168 TB	190 TB	203 TB	225 TB	243 TB
	112 TB	177 TB	207 TB	225 TB	254 TB	272 TB	301 TB	325 TB
# of 6 core CPUs (one per Module)	6	9	10	11	12	13	14	15
Memory (24 GB per module w 12/3TB)	144	216	240	264	288	312	336	360
Memory (48 GB per module w/4TB)	288	432	480	528	576	624	672	720
{Optional} SSD Cache - TB	2.4	3.6	4.0	4.4	4.8	5.2	5.6	6.0
Power (kVA) -Gen3.2 / SSD	2.4 / 2.6	3.5 / 3.8	3.9 / 4.2	4.2 / 4.5	4.6 / 4.9	4.9 / 5.3	5.3 / 5.7	5.6 / 6.1

Figure 10-23 XIV with 4TB disks

If XIV is not configured with the full capacity initially, you can use the SAN Volume Controller rebalancing script to optimize volume placement when capacity is added to the XIV.

10.9.2 I/O ports

XIV supports 8 - 24 FC ports, depending on the number of modules that are installed. Each module has two dual-port FC cards. Use one port per card for SAN Volume Controller use. With this setup, the number of available ports for SAN Volume Controller use is in the range of 4 - 12 ports, as shown in Table 10-20.

Table 10-20 XIV FC ports for SAN Volume Controller

Number of XIV modules installed	XIV modules with FC ports	Total available FC ports	Ports used per FC card	Port available for the SAN Volume Controller
6	4, 5	8	1	4
9	4, 5, 7, 8	16	1	8
10	4, 5, 7, 8	16	1	8
11	4, 5, 7, 8, 9	20	1	10
12	4, 5, 7, 8, 9	20	1	10
13	4, 5, 6, 7, 8, 9	24	1	12
14	4, 5, 6, 7, 8, 9	24	1	12
15	4, 5, 6, 7, 8, 9	24	1	12

As shown in Table 10-20 on page 305, the SAN Volume Controller 16-port limit for storage subsystem is not reached.

To provide redundancy, connect the ports available for SAN Volume Controller use to dual fabrics. Connect each module to separate fabrics. Figure 10-24 shows an example of preferred practice SAN connectivity.

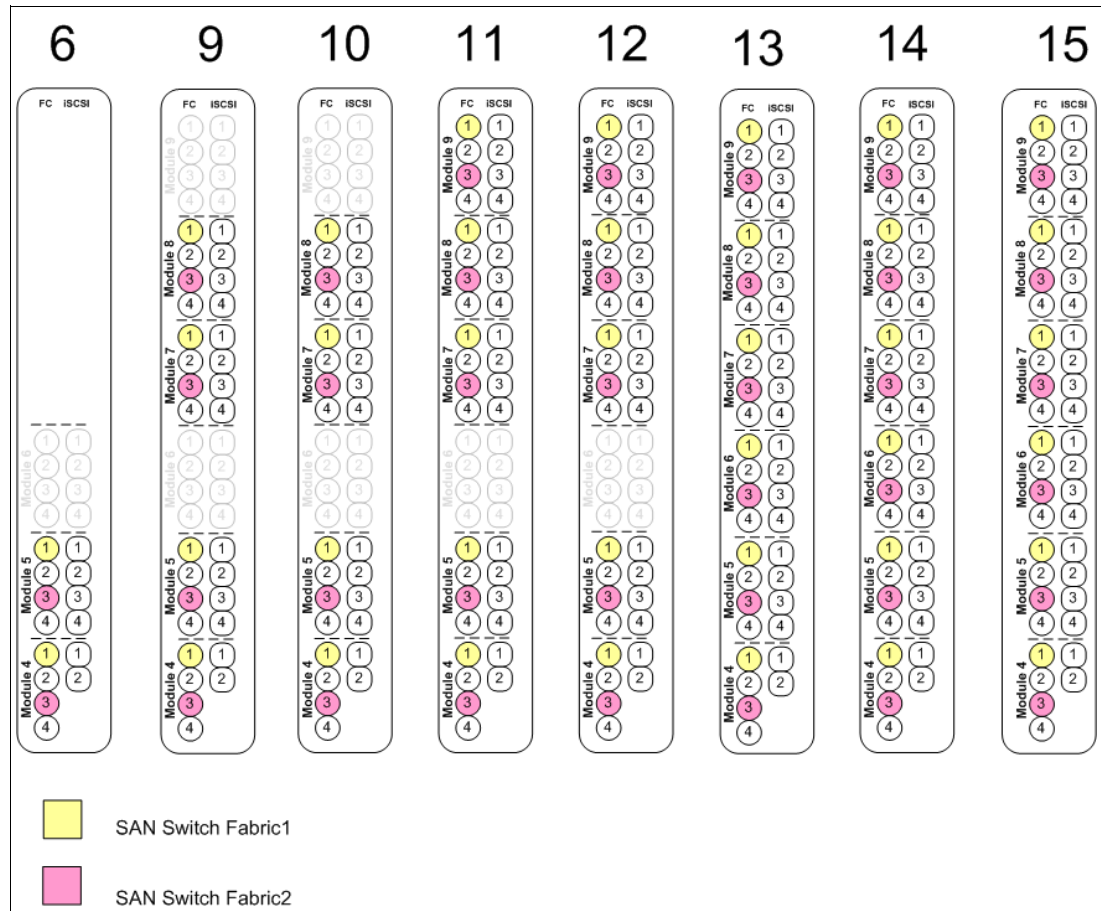


Figure 10-24 XIV SAN connectivity

Host definition for the SAN Volume Controller on an XIV system

Use one host definition for the entire SAN Volume Controller cluster, define all SAN Volume Controller WWPNs to this host, and map all LUNs to it.

You can use the cluster definition with each SAN Volume Controller node as a host. However, the LUNs that are mapped must have their LUN ID preserved when they are mapped to the SAN Volume Controller.

10.9.3 Storage pool layout

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

10.9.4 Extent size

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.

10.10 Storwize V7000 considerations

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 this writing) provides more cache (24 GB per node versus 8 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.

10.10.1 Volume setup

When Storwize V7000 is used as the back-end storage system for the SAN Volume Controller, its main function is to provide RAID capability.

For the Storwize V7000 setup in a SAN Volume Controller environment, define one storage pool with one volume per Storwize V7000 array. With this setup, you avoid striping over striping. Striping might be performed only on the SAN Volume Controller level. Each volume is then presented to the SAN Volume Controller as managed disk, and all MDs from the same type of disks in Storwize V7000 must be used in one storage pool on the SAN Volume Controller level.

The optimal array sizes for SAS disks are 6+1, 7+1, and 8+1. The smaller array size is mainly for RAID rebuild times. The performance has no other implications with bigger array sizes; for example, 10+1 and 11+1.

Figure 10-25 shows an example of the Storwize V7000 configuration with optimal smaller arrays and non-optimal larger arrays.

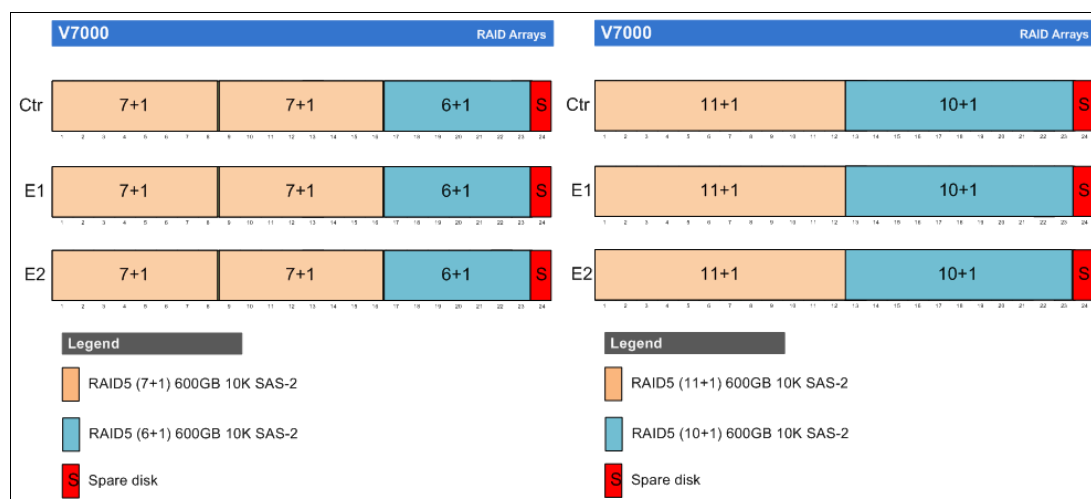


Figure 10-25 Storwize V7000 array for SAS disks

As shown in Figure 10-25, one hot spare disk was used for enclosure, which is not a requirement. However, it is helpful because it provides symmetrical usage of the enclosures. At a minimum, use one hot spare disk per SAS chain for each type of disk in the Storwize V7000. If more than two enclosures are present, you must have at least two HS disks per SAS chain per disk type, if those disks occupy more than two enclosures. Figure 10-26 shows a Storwize V7000 configuration with multiple disk types.

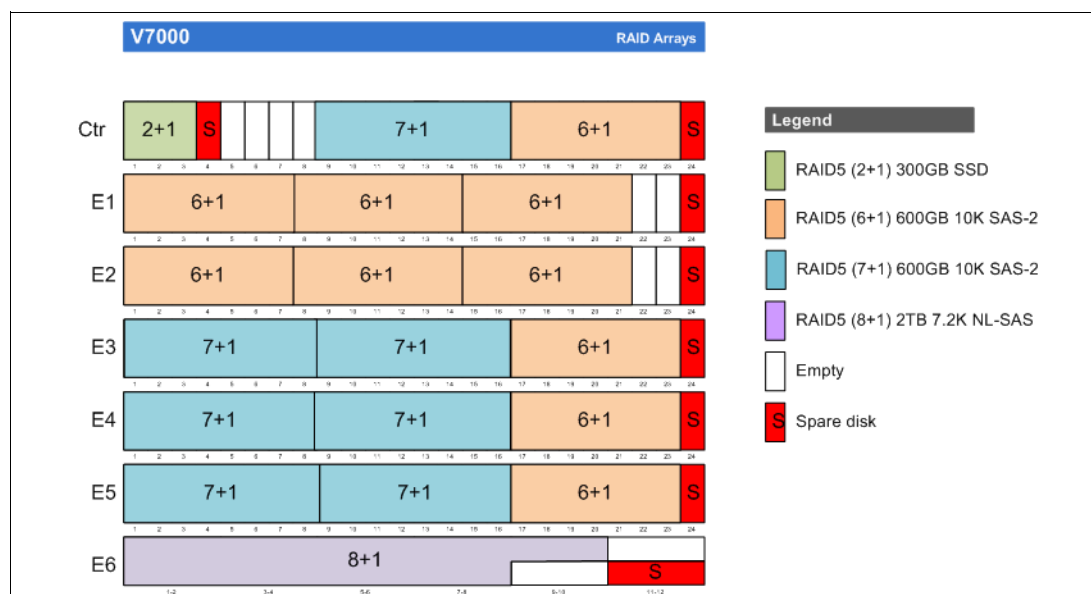


Figure 10-26 Storwize V7000 with multiple disk types

When you define a volume on the Storwize V7000 level, use the default values. The default values define a 256-KB strip size (the size of the RAID chunk on that disk), which is in line with the SAN Volume Controller back-end I/O size, which in V6.1 is above 256 KB. For example, the use of a 256 KB strip size gives a 2-MB stride size (the whole RAID chunk size) in an 8+1 array.

Storwize V7000 also supports large NL-SAS drives (2 TB and 3 TB). The use of those drives in RAID 5 arrays can produce significant RAID rebuild times, even several hours. Therefore, use RAID 6 to avoid double failure during the rebuild period. Figure 10-27 shows this type of setup.



Figure 10-27 Storwize V7000 RAID6 arrays

Tip: Make sure that volumes that are defined on Storwize V7000 are distributed evenly across all nodes.

10.10.2 I/O ports

Each Storwize V7000 has four FC ports for a host access. These ports are used by the SAN Volume Controller to access the volumes on Storwize V7000. A minimum configuration is to connect each Storwize V7000 canister node to two independent fabrics, as shown in Figure 10-28.

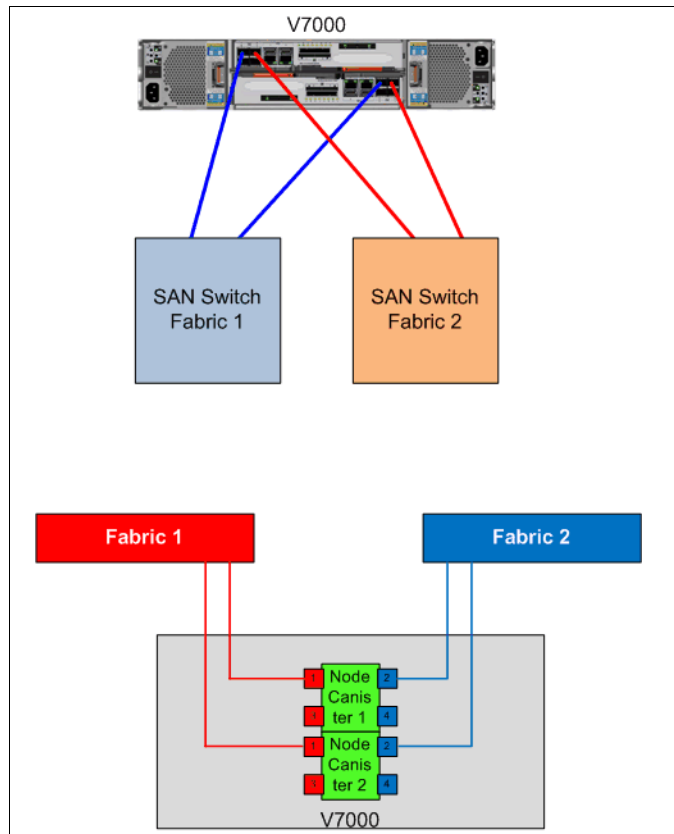


Figure 10-28 Storwize V7000 two connections per node

In this setup, the SAN Volume Controller can access a Storwize V7000 with a two-node configuration over four ports. Such connectivity is sufficient for Storwize V7000 environments that are not fully loaded.

However, if the Storwize V7000 is hosting capacity that requires more than two connections per node, use four connections per node, as shown in Figure 10-29 on page 311.

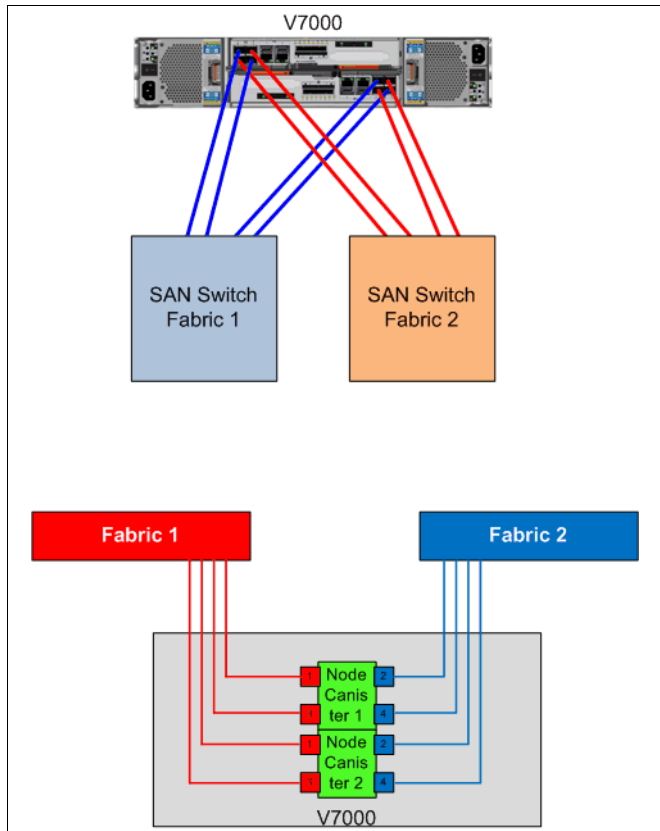


Figure 10-29 Storwize V7000 four connections per node

With a two-node Storwize V7000 setup, eight target connections are provided from the SAN Volume Controller perspective. This number is well below the 16 target ports that is the current SAN Volume Controller limit for back-end storage subsystems.

Previously, the limit in the Storwize V7000 configuration was four-node cluster. With this configuration of four connections to the SAN, the limit of 16 target ports is reached. As such, this configuration might still be supported. Figure 10-30 shows an example of the configuration. However, the specified limit is no longer true and we can have configurations supporting eight-node cluster with four control enclosures with four IO groups.

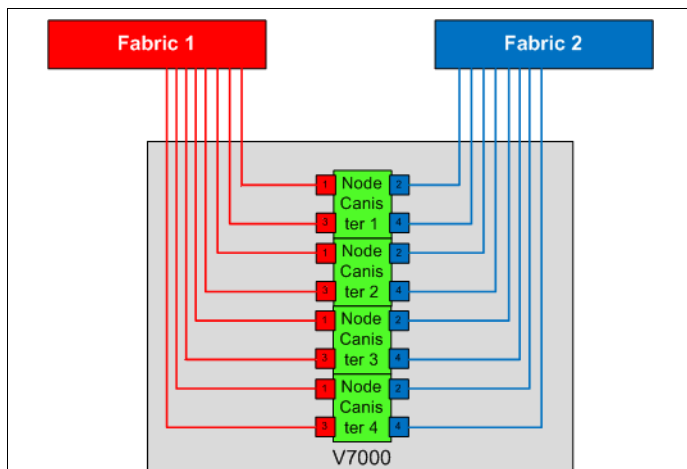


Figure 10-30 Four-node Storwize V7000 setup

Redundancy consideration: At a minimum, connect two ports per node to the SAN with connections to two redundant fabrics.

10.10.3 Storage pool layout

As with any other storage subsystem where different disk types can be installed, use the volumes with the same characteristics (size, RAID level, and rotational speed) in a single storage pool on the SAN Volume Controller level. Also, use the single storage pool for all volumes of the same characteristics.

For an optimal configuration, use the exact number of disks in the storage pool. For example, if you have 7+1 and 6+1 arrays, you can use two pools, as shown in Figure 10-31.

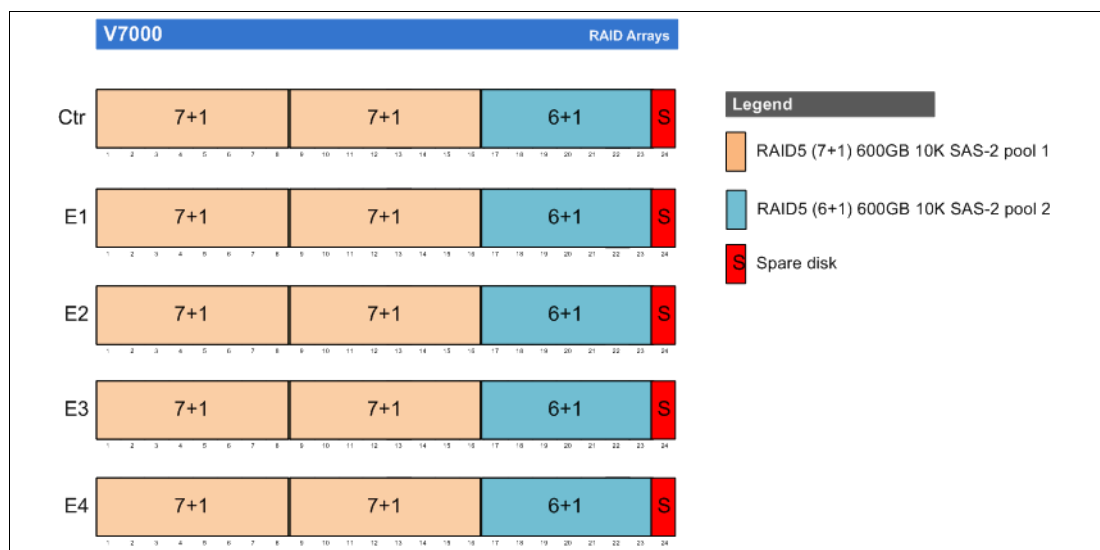


Figure 10-31 Storwize V7000 storage pool example with two pools

This example has a hot spare disk in every enclosure, which is not a requirement. To avoid having two pools for the same disk type, create an array configuration that is based on the following rules:

- ▶ Number of disks in the array:
 - 6+1
 - 7+1
 - 8+1
- ▶ Number of hot spare disks: Minimum of two

Based on the array size, the following symmetrical array configuration is possible as a setup for a five-enclosure Storwize V7000:

- ▶ 6+1 - 17 arrays (119 disks) + 1 x hot spare disk
- ▶ 7+1 - 15 arrays (120 disks) + 0 x hot spare disk
- ▶ 8+1 - 13 arrays (117 disks) + 3 x hot spare disks

The 7+1 array does not provide any hot spare disks in the symmetrical array configuration, as shown in Figure 10-32.

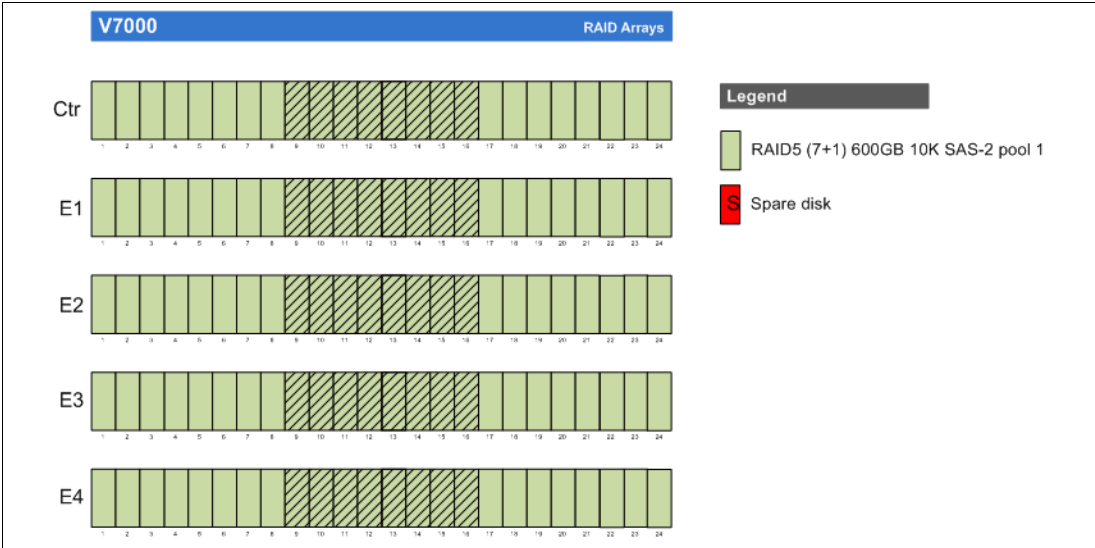


Figure 10-32 Storwize V7000 7+1 symmetrical array configuration

The 6+1 arrays provide a single hot spare disk in the symmetrical array configuration, as shown in Figure 10-33. It is not a preferred value for the number of hot spare disks.

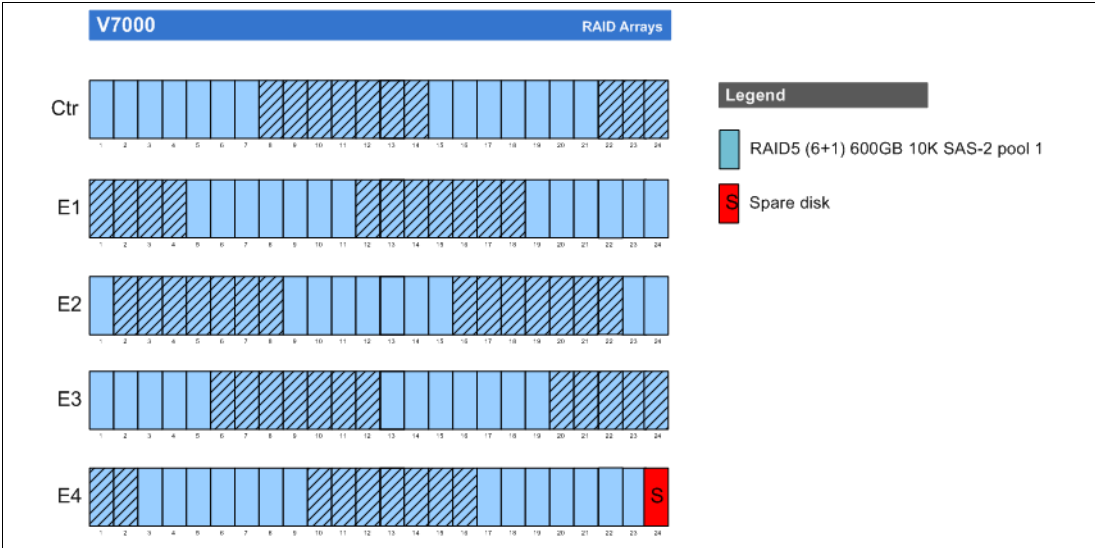


Figure 10-33 Storwize V7000 6+1 symmetrical array configuration

The 8+1 arrays provide three hot spare disks in the symmetrical array configuration, as shown in Figure 10-34 on page 314. These arrays are within the recommended value range for the number of hot spare disks (two).

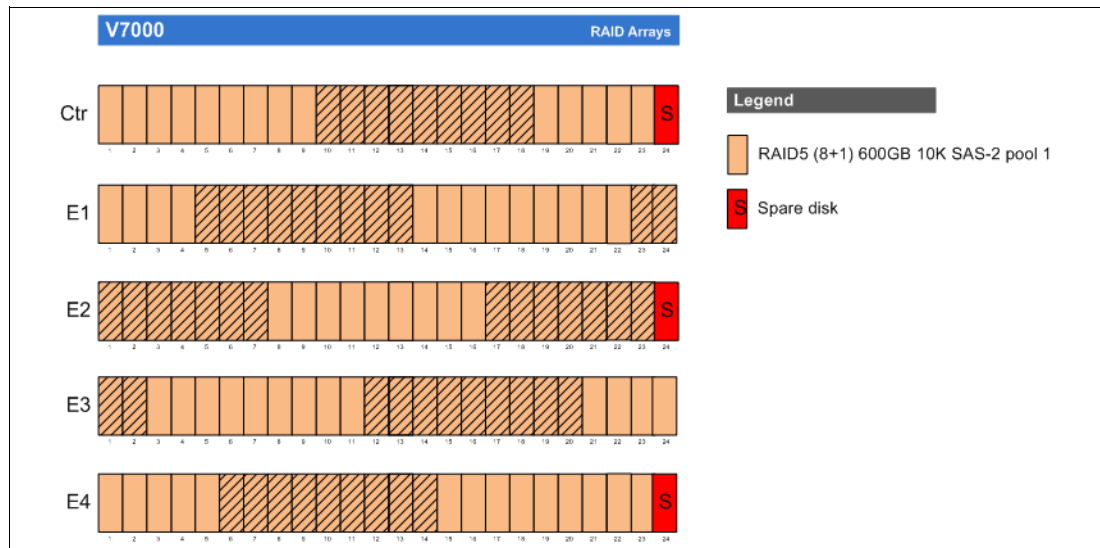


Figure 10-34 Storwize V7000 8+1 symmetrical array configuration

The best configuration for a single storage pool for the same type of disk in a five-enclosure Storwize V7000 is an 8+1 array configuration.

Tip: A symmetrical array configuration for the same disk type provides the least possible complexity in a storage pool configuration.

10.10.4 Extent size

To optimize capacity, use an extent size of 1 GB. Although you can use smaller extent sizes, this size limits the amount of capacity that can be managed by the SAN Volume Controller cluster. No performance benefit is gained by using smaller or larger extent sizes.

10.11 DS5000 series considerations

The considerations for DS5000 series also apply to the DS3000 and DS4000 series.

10.11.1 Selecting array and cache parameters

This section describes optimum array and cache parameters.

DS5000 array width

With Redundant Array of Independent Disks 5 (RAID 5) arrays, determining the number of physical drives to put into an array always presents a compromise. Striping across a larger number of drives can improve performance for transaction-based workloads. However, striping can also have a negative effect on sequential workloads.

A common mistake that is often made when selecting array width is the tendency to focus only on the capability of a single array to perform various workloads. But, you must also consider the aggregate throughput requirements of the entire storage server. Many physical disks in an array can create a workload imbalance between the controllers because only one controller of the DS5000 actively accesses a specific array.

When you select the array width, consider its effect on rebuild time and availability. A larger number of disks in an array increases the rebuild time for disk failures, which can have a negative effect on performance. Also, having more disks in an array increases the probability of having a second drive failure within the same array before the rebuild of an initial drive failure completes. This exposure is inherent to the RAID 5 architecture.

Preferred practice: For the DS5000, use array widths of 4+p and 8+p.

Segment size

With direct-attached hosts, considerations are often made to align device data partitions to physical drive boundaries within the storage controller. For the SAN Volume Controller, aligning device data partitions to physical drive boundaries within the storage controller is less critical. The reason is based on the caching that the SAN Volume Controller provides, and on the fact that less variation is in its I/O profile, which is used to access back-end disks.

Because the maximum destage size for the SAN Volume Controller is 256 KB, it is impossible to achieve full stride writes for random workloads. For the SAN Volume Controller, the only opportunity for full stride writes occurs with large sequential workloads, and in that case, the larger the segment size, the better. Larger segment sizes can adversely affect random I/O, however. The SAN Volume Controller and controller cache hide the RAID 5 write penalty for random I/O well, and therefore, larger segment sizes can be accommodated. The primary consideration for selecting segment size is to ensure that a single host I/O fits within a single segment to prevent accessing multiple physical drives.

Preferred practice: Use a segment size of 256 KB as the best compromise for all workloads.

Cache block size

The DS4000 uses a 4-KB cache block size by default. However, it can be changed to max 32KB with the latest firmware.

For earlier models of DS4000 that use 2-Gb FC adapters, the 4-KB block size performed better for random I/O, and 16 KB block size performs better for sequential I/O. However, because most workloads contain a mix of random and sequential I/O, the default values prove to be the best choice. For the higher-performing DS4700 and DS4800, the 4-KB block size advantage for random I/O becomes harder to see.

Because most client workloads involve at least some sequential workload, the best overall choice for these models is the 16KB block size.

Preferred practice: For the DS5/4/3000, set the cache block size to 16KB. For use with SAN Volume Controller/Storwize, the maximum cache block size can be set to 32 KB running the latest firmware.

Table 10-21 summarizes the SAN Volume Controller and DS5000 values.

Table 10-21 SAN Volume Controller values

Models	Attribute	Value
SAN Volume Controller	Extent size (MB)	256
SAN Volume Controller	Managed mode	Striped
DS5000	Segment size (KB)	256

Models	Attribute	Value
DS5000	Cache block size (KB)	32 KB
DS5000	Cache flush control	80/80 (default)
DS5000	Readahead	1
DS5000	RAID 5	4+p, 8+p

10.11.2 Considerations for controller configuration

This section describes some considerations for a controller configuration.

Balancing workload across DS5000 controllers

When you are creating arrays, spread the disks across multiple controllers and alternating slots within the enclosures. This practice improves the availability of the array by protecting against enclosure failures that affect multiple members within the array, and improves performance by distributing the disks within an array across drive loops. You spread the disks across multiple controllers and alternating slots within the enclosures by using the manual method for array creation.

Figure 10-35 shows a Storage Manager view of a 2+p array that is configured across enclosures. Here, each disk of the three disks is represented in a separate physical enclosure, and slot positions alternate from enclosure to enclosure.

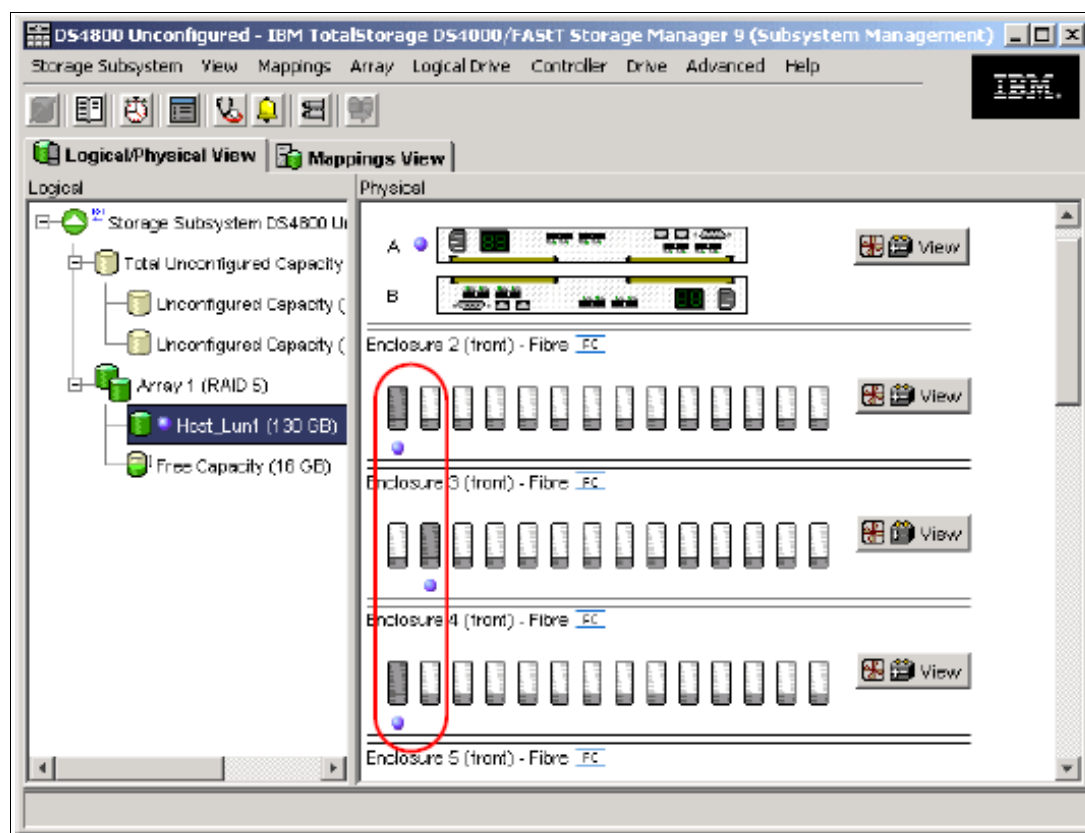


Figure 10-35 Storage Manager

10.11.3 Mixing array sizes within the storage pool

Mixing array sizes within the storage pool in general is not of concern. Testing shows no measurable performance differences between selecting all 6+p arrays and all 7+p arrays as opposed to mixing 6+p arrays and 7+p arrays. In fact, mixing array sizes can help balance workload because it places more data on the ranks that have the extra performance capability that is provided by the eighth disk. A small exposure is if an insufficient number of the larger arrays is available to handle access to the higher capacity. To avoid this situation, ensure that the smaller capacity arrays do not represent more than 50% of the total number of arrays within the storage pool.

Preferred practice: When 6+p arrays and 7+p arrays are mixed in the same storage pool, avoid having smaller capacity arrays comprise more than 50% of the arrays.

10.11.4 Determining the number of controller ports for DS4000

The DS4000 must be configured with two ports per controller, for a total of four ports per DS4000.

10.11.5 Performance considerations with FlashSystem

For more information, see the following publications:

- ▶ *IBM SAN Volume Controller and IBM FlashSystem 820: Best Practices and Performance Capabilities*, REDP-5027
- ▶ *Implementing IBM FlashSystem 840*, SG24-8189
- ▶ *Implementing FlashSystem 840 with SAN Volume Controller*, TIPS1137



IBM System Storage Easy Tier function

This chapter describes the function that is provided by the IBM System Storage EasyTier feature of the SAN Volume Controller for disk performance optimization. It also describes how to activate the EasyTier process for evaluation purposes and for automatic extent migration.

This chapter includes the following sections:

- ▶ Overview of Easy Tier
- ▶ Easy Tier concepts
- ▶ Easy Tier implementation considerations
- ▶ Measuring and activating Easy Tier
- ▶ Activating Easy Tier with the SAN Volume Controller CLI
- ▶ Activating Easy Tier with the SAN Volume Controller GUI

11.1 Overview of Easy Tier

Determining the amount of I/O activity that occurs on a SAN Volume Controller extent and when to move the extent to an appropriate storage performance tier is usually too complex a task to manage manually. Easy Tier is a performance optimization function that overcomes this issue. It automatically migrates or moves extents that belong to a volume between MDisk storage tiers.

Easy Tier monitors the I/O activity and latency of the extents on all volumes with the Easy Tier function that is turned on in a multitier storage pool over a 24-hour period. It then creates an extent migration plan that is based on this activity and dynamically moves high activity (or *hot extents*) to a higher disk tier within the storage pool. It also moves extents whose activity dropped off (or “cooled”) from the high-tier MDisk back to a lower-tiered MDisk. Because this migration works at the extent level, it is often referred to as *sub-LUN tiering*.

Turning Easy Tier on and off: The Easy Tier function can be turned on or off at the storage pool level and at the volume level.

To experience the potential benefits of the use of Easy Tier in your environment before you install solid-state drives (SSDs) or virtualize external Flash-based storage, such as IBM FlashSystem storage systems, turn on the Easy Tier function for a *single level storage pool*. Next, turn on the Easy Tier function for the volumes within that pool, which starts monitoring activity on the volume extents in the pool.

Easy Tier creates a migration report every 24 hours on the number of extents that might be moved if the pool were a multitiered storage pool. Although Easy Tier extent migration is not possible within a single tier pool, the Easy Tier statistical measurement function is available.

Attention: Image mode and sequential volumes are *not* candidates for Easy Tier automatic data placement. Compressed volumes *cannot* be used with Easy Tier in software version 6.4 or earlier. Compressed volumes *are* candidates for Easy Tier in software version 7.1 and later.

Depending on the specific storage platform model, Easy Tier is a no-charge feature or it can be optionally licensed.

11.2 Easy Tier concepts

This section describes the concepts that underpin Easy Tier functionality.

11.2.1 SSD arrays and MDisks

The SSDs are treated no differently by the SAN Volume Controller than hard disk drives (HDDs) regarding RAID arrays or MDisks.

The individual SSDs in the storage that is managed by the SAN Volume Controller 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 overhead, with two logical SSDs used for parity only. A LUN is created on the array and is then presented to the SAN Volume Controller as a normal managed disk (MDisk).

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 protection above the RAID level by using volume mirroring.

In the example disk tier pool that is shown in Figure 11-2 on page 322, you can see the SSD MDisks presented from the SSD disk arrays.

11.2.2 Disk tiers

The MDisks (LUNs) presented to the SAN Volume Controller cluster are likely to have different performance attributes because of the type of disk or RAID array on which they are installed. The MDisks can be on 15 K RPM Fibre Channel or SAS disk, Nearline SAS or SATA, or even SSDs or IBM FlashSystem storage.

Therefore, a storage tier attribute is assigned to each MDisk. The default is generic_hdd. With SAN Volume Controller V6.1, a new disk tier attribute is available for SSDs and is known as generic_ssd.

The SAN Volume Controller does not automatically detect SSD MDisks. Instead, all external MDisks initially are put into the generic_hdd tier by default. Then, the administrator must manually change the SSD tier to generic_ssd by using the command-line interface (CLI) or GUI.

11.2.3 Single tier storage pools

Figure 11-1 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 or mirrored volume can be measured by Easy Tier, but no action to optimize the performance occurs.

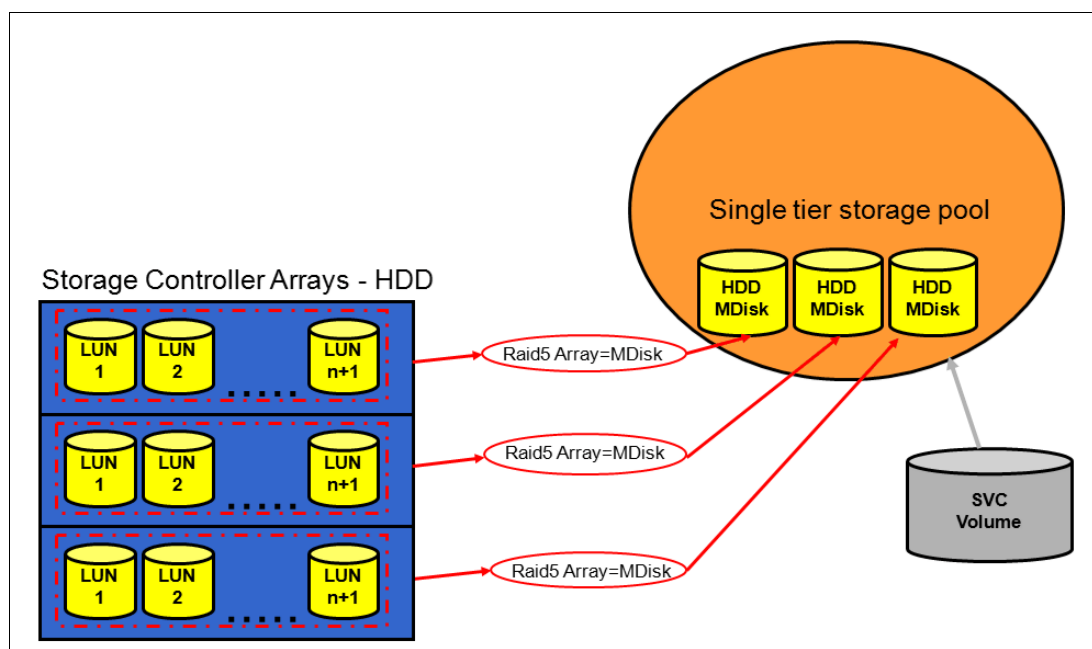


Figure 11-1 Single tier storage pool with striped volume

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, and disk revolutions per minute (RPMs), and controller performance characteristics.

11.2.4 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 generic_hdd and generic_ssd MDisks.

Figure 11-2 shows a scenario in which a storage pool is populated with two different MDisk types: one belonging to an SSD array and one belonging to an HDD array. Although this example shows RAID 5 arrays, other RAID types can be used.

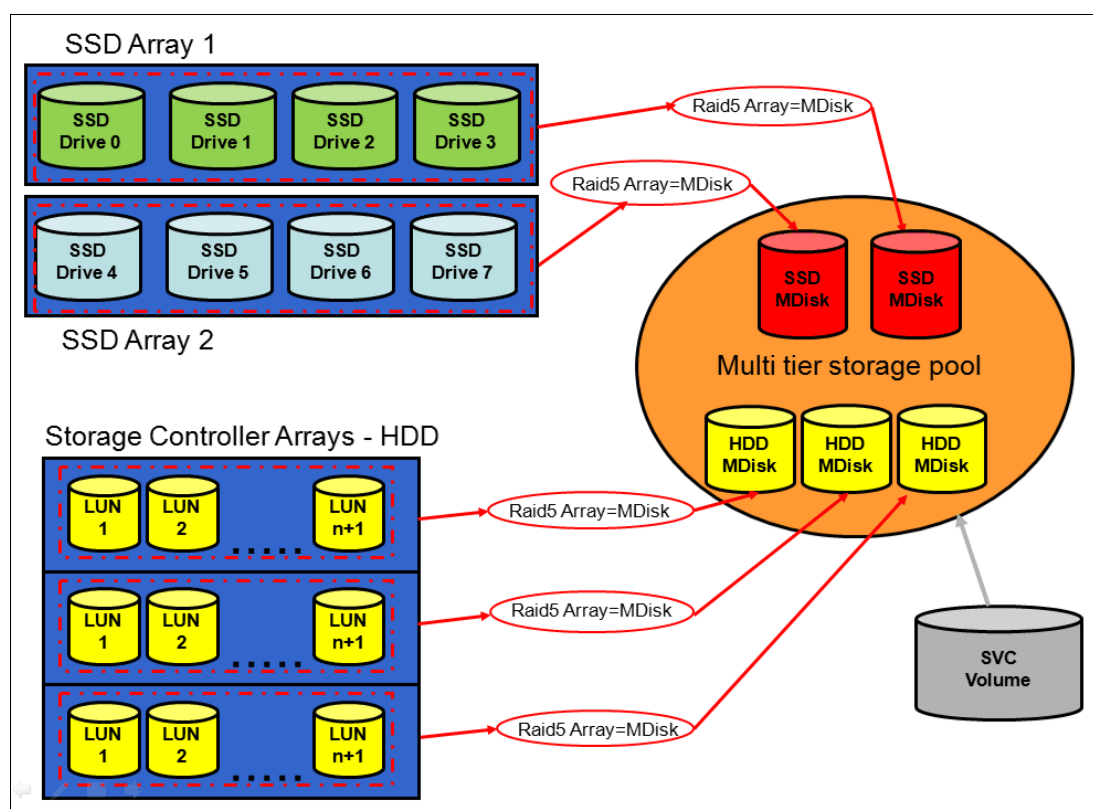


Figure 11-2 Multitier storage pool with striped volume

Adding SSD to the pool means that more space is also now available for new volumes or volume expansion.

11.2.5 Easy Tier process

The Easy Tier function has the following 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 and thus considers only I/Os of up to 64 KB as migration candidates.

This process is efficient and adds negligible processing overhead to the SAN Volume Controller 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 (SSD) tier.

This process also identifies extents that must be migrated back to a lower (HDD) tier.

- ▶ **Data Migration Planner**

By using the extents that were previously identified, the Data Migration Planner step builds the extent migration plan for the storage pool.

- ▶ **Data Migrator**

This process involves the actual movement or migration of the volume's extents up to, or down from, the high disk tier. The extent migration rate is capped so that a maximum of up to 30 MBps is migrated, which equates to around 3 TB a day that is migrated between disk tiers.

When it relocates volume extents, Easy Tier performs the following tasks:

- ▶ It attempts to migrate the most active volume extents up to SSD first.

To ensure that a free extent is available, you might need to first migrate a less frequently accessed extent back to the HDD.

- ▶ A previous migration plan and any queued extents that are not yet relocated are abandoned.

11.2.6 Easy Tier operating modes

Easy Tier features the following main operating modes:

- ▶ Off mode
- ▶ Evaluation or measurement only mode
- ▶ Automatic Data Placement or extent migration mode

Easy Tier off mode

With Easy Tier turned off, no statistics are recorded and no 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 often is 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 SAN Volume Controller nodes. This file can be offloaded from the SAN Volume Controller nodes with PSCP `-load` or by using the GUI, as described in 11.4.1, "Measuring by using the Storage Advisor Tool" on page 326. 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 MDisk with different disk tiers, thus being a multitiered storage pool.

Dynamic data movement is not apparent to the host server and application users of the data, other than providing improved performance. Extents are automatically migrated, as described in 11.3.2, “Implementation rules” on page 325.

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 SSD and a prediction of performance improvement that can be gained if more SSD arrays are available.

11.2.7 Easy Tier activation

To activate Easy Tier, set the Easy Tier value on the pool and volumes as shown in Figure 11-3. The defaults are set in favor of Easy Tier. For example, if you create a storage pool, the `-easytier` value is `auto`. If you create a volume, the value is `on`.

Figure 11-3 Easy Tier parameter 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	Off	inactive (see note 2)
Off	Two	On	inactive (see note 2)
Auto (see note 5)	One	Off	inactive (see note 2)
Auto (see note 5)	One	On	inactive (see note 2)
Auto (see note 5)	Two	Off	measured (see note 3)
Auto (see note 5)	Two	On	active (see note 1)
On (see note 5)	One	Off	measured (see note 3)
On (see note 5)	One	On	measured (see note 3)
On (see note 5)	Two	Off	measured (see note 3)
On (see note 5)	Two	On	active (see note 1)
Note: 1. If the volume copy is in image or sequential mode or is being migrated then the volume copy Easy Tier status will be <i>measured</i> instead of <i>active</i> . 2. When the volume copy status is <i>inactive</i> , no Easy Tier functions are enabled for that volume copy. 3. When the volume copy status is <i>measured</i> , the Easy Tier function operates in evaluation mode for that volume. 4. When the volume copy status is <i>active</i> , the Easy Tier function operates in automatic data placement mode for that volume. 5. The default Easy Tier setting for a storage pool is <i>auto</i> , and the default Easy Tier setting for a volume copy is <i>on</i> . This means that Easy Tier functions will be disabled for storage pools with a single tier, and that automatic data placement mode will be enabled for all striped volume copies in a storage pool with two tiers.			

For more information about the use of these parameters, see 11.5, “Activating Easy Tier with the SAN Volume Controller CLI” on page 329, and 11.6, “Activating Easy Tier with the SAN Volume Controller GUI” on page 335.

11.3 Easy Tier implementation considerations

This section describes considerations to remember before you implement Easy Tier.

11.3.1 Prerequisites

No Easy Tier license is required for the SAN Volume Controller. Easy Tier comes as part of the V6.1 code. For Easy Tier to migrate extents, you must have disk storage available that has different tiers; for example, a mix of SSD and HDD.

11.3.2 Implementation rules

Remember the following implementation and operation rules when you use the IBM System Storage Easy Tier function on the SAN Volume Controller:

- ▶ 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 such 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. Therefore, the number of extents that Easy Tier migrates to the SSD tier might be different for each copy.

- ▶ If possible, the SAN Volume Controller creates new volumes or volume expansions by using extents from MDisk from the HDD tier. However, it uses extents from MDisk from the SSD tier, if necessary.
- ▶ 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 if 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.

11.3.3 Easy Tier limitations

When you use IBM System Storage Easy Tier on the SAN Volume Controller, Easy Tier has 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 the SAN Volume Controller 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 SAN Volume Controller 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 image mode.
Easy Tier automatic data placement does not support image mode. When a volume with Easy Tier automatic data placement mode active is migrated to 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. Easy Tier supports evaluation mode for image mode volumes.
- ▶ Compressed volumes *cannot* be used with Easy Tier in software version 6.4 or earlier. Compressed volumes *are* candidates for Easy Tier in software version 7.1 and later.

Preferred practices: Consider the following preferred practices:

- ▶ Always set the storage pool **-easytier** value to on rather than to the default value auto. This setting makes it easier to turn on evaluation mode for existing single tier pools, and no further changes are needed when you move to multitier pools. For more information about the mix of pool and volume settings, see “Easy Tier activation” on page 324.
- ▶ The use of Easy Tier can make it more appropriate to use smaller storage pool extent sizes.

11.4 Measuring and activating Easy Tier

You can measure and activate Easy Tier, as described in the following sections.

11.4.1 Measuring by using the Storage Advisor Tool

The IBM Storage Advisor Tool (STAT) is a command-line tool that runs on Windows systems. It takes input from the dpa_heat files that are created on the SAN Volume Controller nodes and produces a set of HTML files that contain activity reports. The advisor tool is an application that creates an HTML file that contains a report. For more information, see “IBM Storage Tier Advisor Tool” at this website:

<http://www.ibm.com/support/docview.wss?uid=ssg1S4000935>

For more information about the Storage Advisor Tool, contact your IBM representative or IBM Business Partner.

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 following CLI command:

```
svcinfo lsdumps node_id | node_nameWhere node_id | node_name is the node ID or name to list the available dpa_heat data files.
```

Next, perform the normal PSCP **-load** download process, as shown in the following example:

```
pscp -unsafe -load saved_putty_configuration
admin@cluster_ip_address:/dumps/dpa_heat.node_name.date.time.data
your_local_directory
```

Using the GUI

If you prefer to use the GUI, click **Settings** $\varnothing$ **Support** page, as shown in Figure 11-4. If the page is not displaying a list of individual log files, click **Show full log listing**.

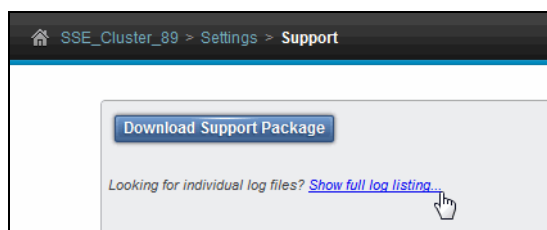


Figure 11-4 Accessing the “dpa_heat” file in the SAN Volume Controller 7.2 GUI

Next, right-click the row for the dpa_heat file and choose **Download**, as shown in Figure 11-5.

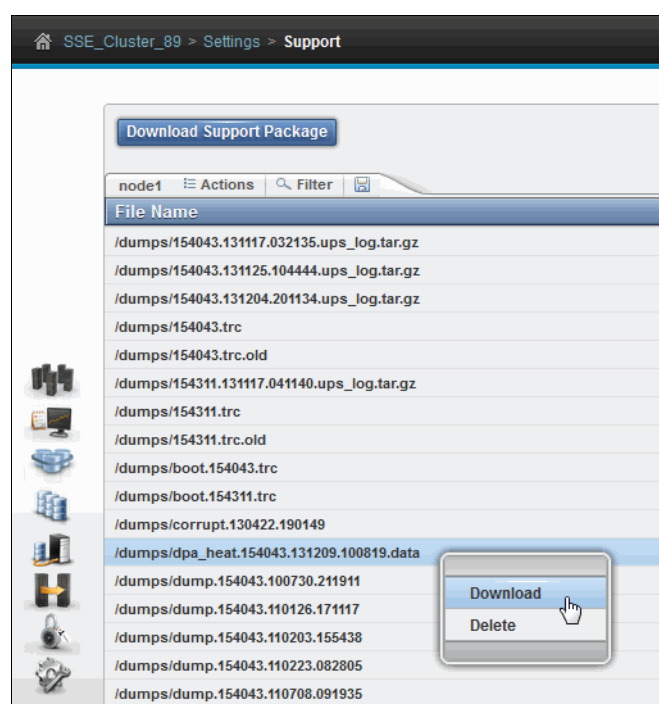


Figure 11-5 Downloading the dpa_heat file in the SAN Volume Controller 7.2 GUI

The file is downloaded to your local workstation.

Running the tool

You run the tool from a command line or terminal session by specifying up to two input dpa_heat file names and directory paths, as shown in the following example:

```
C:\Program Files\IBM\STAT>STAT dpa_heat.nodenummer.yymmdd.hhmmss.data
```

The index.html file is then created in the STAT base directory. When it is opened with your browser, a summary page is displayed, as shown in Figure 11-6.

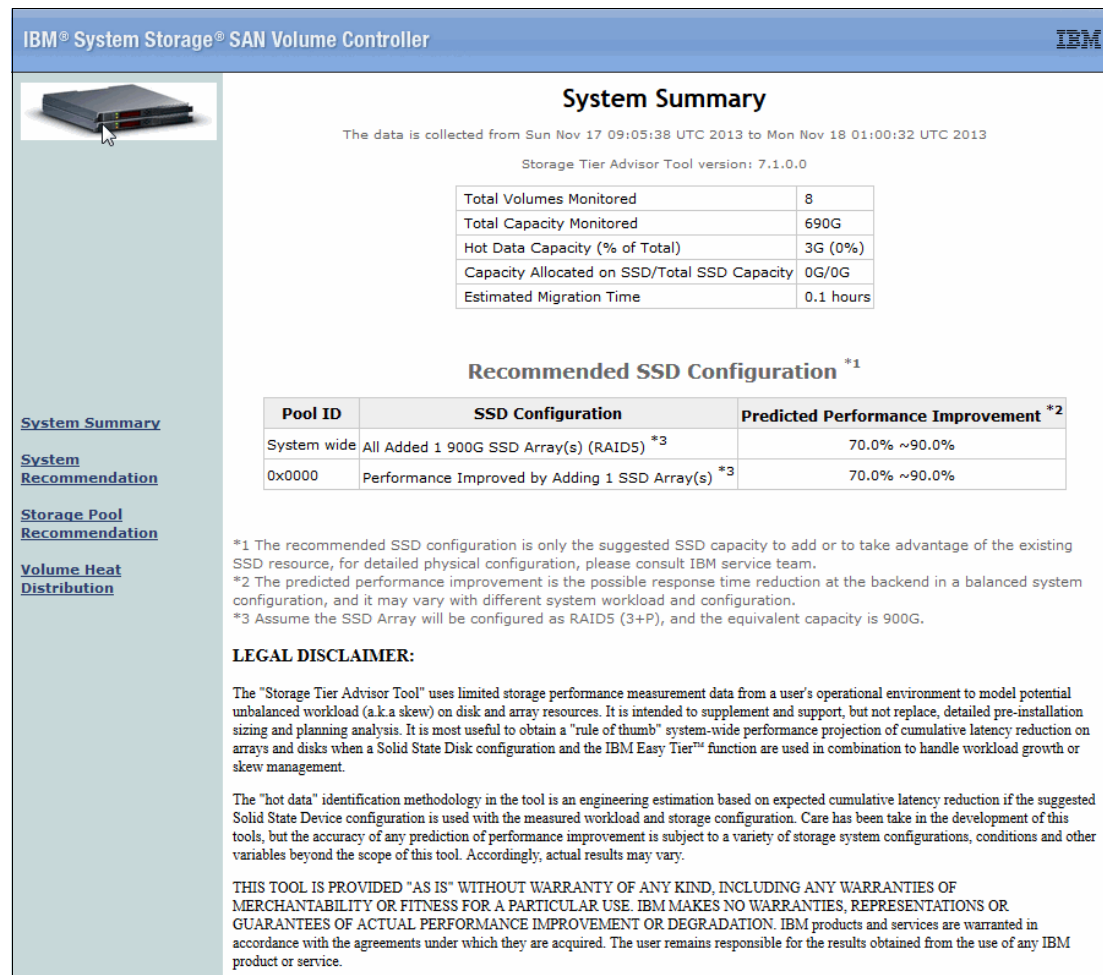


Figure 11-6 STAT System Summary page

The distribution of hot data and cold data for each volume is shown in the volume heat distribution report. The report displays the portion of the capacity of each volume on SSD (red), and HDD (blue), as shown in Figure 11-7.

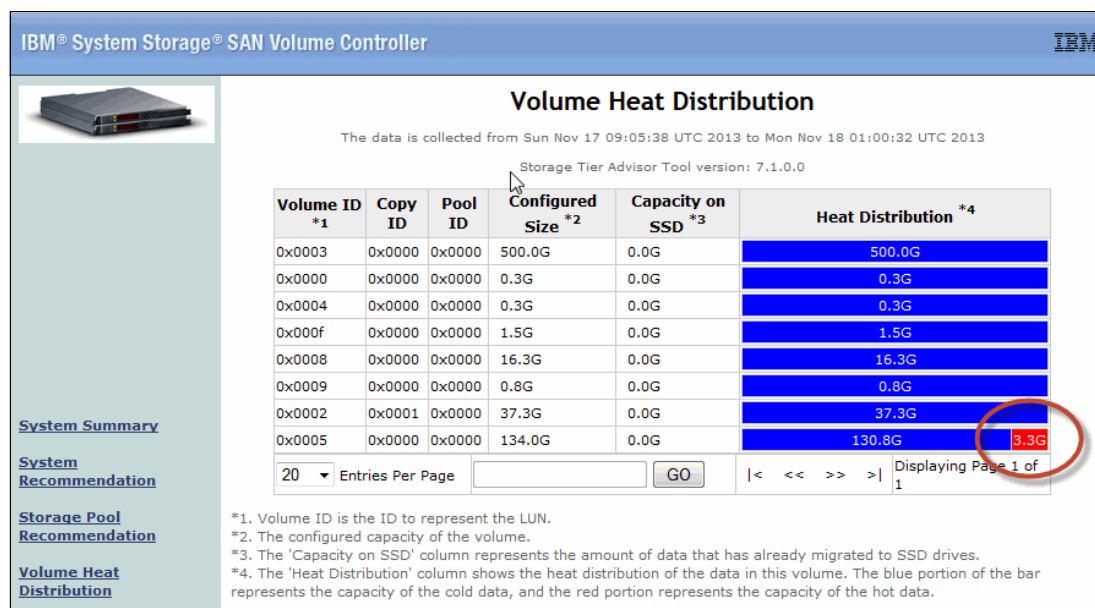


Figure 11-7 STAT Volume Heat Distribution page

11.5 Activating Easy Tier with the SAN Volume Controller CLI

This section describes how to activate Easy Tier by using the SAN Volume Controller CLI. The example is based on the storage pool configurations as shown in Figure 11-1 on page 321 and Figure 11-2 on page 322.

The environment is an SAN Volume Controller cluster with the following resources available:

- ▶ 1 x I/O group with two 2145-CF8 nodes
- ▶ 8 x external 73 GB SSDs (4 x SSD per RAID5 array)
- ▶ 1 x external Storage Subsystem with HDDs

Deleted lines: Many lines that were not related to Easy Tier were deleted in the command output or responses in the examples that are shown in the following sections so that you can focus only on information that is related to Easy Tier.

11.5.1 Initial cluster status

Example 11-1 on page 330 shows the SAN Volume Controller cluster characteristics before you add multitiered storage (SSD with HDD) and begin the Easy Tier process. The example shows two different tiers available in our SAN Volume Controller cluster: generic_ssd and generic_hdd. Now, zero disk is allocated to the generic_ssd tier; therefore, it is showing a capacity of 0.00 MB.

Example 11-1 SAN Volume Controller cluster characteristics

```
IBM_2145:ITS0-CLS5:admin>svcinfolcluster
id          name          location partnership bandwidth id_alias
0000020060800004 ITS0-CLS5 local          0000020060800004

IBM_2145:ITS0-CLS5:admin>svcinfolcluster 0000020060800004
id 0000020060800004
name ITS0-CLS5
.
tier generic_ssd
tier_capacity 0.00MB
tier_free_capacity 0.00MB
tier generic_hdd
tier_capacity 18.85TB
tier_free_capacity 18.43TB
```

11.5.2 Turning on Easy Tier evaluation mode

Figure 11-1 on page 321 shows an existing single tier storage pool. To turn on Easy Tier evaluation mode, set **-easytier on** for both the storage pool and the volumes in the pool. Table 11-3 on page 324 shows how to check the required mix of parameters that are needed to set the volume Easy Tier status to measured.

Example 11-2 shows turning on Easy Tier evaluation mode for both the pool and volume so that the extent workload measurement is enabled. First, you check the pool. Then, you change it. You then repeat the steps for the volume.

Example 11-2 Turning on Easy Tier evaluation mode

```
IBM_2145:ITS0-CLS5:admin>svcinfolmdiskgrp -filtervalue "name=Single*"
id name status mdisk_count vdisk_count easy_tier easy_tier_status
27 Single_Tier_Storage_Pool online 3 1 off inactive

IBM_2145:ITS0-CLS5:admin>svcinfolmdiskgrp Single_Tier_Storage_Pool
id 27
name Single_Tier_Storage_Pool
status online
mdisk_count 3
vdisk_count 1
.
easy_tier off
easy_tier_status inactive
.
tier generic_ssd
tier_mdisk_count 0
.
tier generic_hdd
tier_mdisk_count 3
tier_capacity 200.25GB

IBM_2145:ITS0-CLS5:admin>svctask chmdiskgrp -easytier on Single_Tier_Storage_Pool
IBM_2145:ITS0-CLS5:admin>svcinfolmdiskgrp Single_Tier_Storage_Pool
id 27
name Single_Tier_Storage_Pool
status online
mdisk_count 3
```

```

vdisk_count 1
.
easy_tier on
easy_tier_status active
.
tier generic_ssd
tier_mdisk_count 0
.
tier generic_hdd
tier_mdisk_count 3
tier_capacity 200.25GB

```

----- Now Repeat for the Volume -----

```

IBM_2145:ITS0-CLS5:admin>svcinfolsvdisk -filtervalue "mdisk_grp_name=Single*"
id name status mdisk_grp_id mdisk_grp_name capacity type
27 ITS0_Volume_1 online 27 Single_Tier_Storage_Pool 10.00GB striped

```

```

IBM_2145:ITS0-CLS5:admin>svcinfolsvdisk ITS0_Volume_1
id 27
name ITS0_Volume_1
.

```

```

easy_tier off
easy_tier_status inactive
.
tier generic_ssd
tier_capacity 0.00MB
.
tier generic_hdd
tier_capacity 10.00GB

```

```

IBM_2145:ITS0-CLS5:admin>svctask chvdisk -easytier on ITS0_Volume_1
IBM_2145:ITS0-CLS5:admin>svcinfolsvdisk ITS0_Volume_1
id 27
name ITS0_Volume_1
.

```

```

easy_tier on
easy_tier_status measured
.
tier generic_ssd
tier_capacity 0.00MB
.
tier generic_hdd
tier_capacity 10.00GB

```

11.5.3 Creating a multitier storage pool

With the SSD candidates placed into an array, you now need a pool in which to place the two tiers of disk storage. If you already have an HDD single tier pool, you must know the existing MDiskgrp ID or name.

In this example, a storage pool is available within which we want to place the SSD arrays, Multi_Tier_Storage_Pool. After you create the SSD arrays (which appear as MDisks), they are placed into the storage pool, as shown in Example 11-3.

The storage pool easy_tier value is set to auto because it is the default value that is assigned when you create a storage pool. Also, the SSD MDisks default tier value is set to generic_hdd, and not to generic_ssd.

Tip: Internal SSD MDisks in SAN Volume Controller and Storwize family storage systems are automatically recognized upon detection and their tier value is set to generic_ssd. However, this is not the case for external SSD MDisks. The tier value of any external MDisk is set to generic_hdd when the MDisk is initially detected. You must manually change the tier value to activate Easy Tier.

Example 11-3 Multitier pool creation

```
IBM_2145:ITS0-CLS5:admin>svcinfolsmdiskgrp -filtervalue "name=Multi*"
id name status mdisk_count vdisk_count capacity easy_tier easy_tier_status
28 Multi_Tier_Storage_Pool online 3 1 200.25GB auto inactive

IBM_2145:ITS0-CLS5:admin>svcinfolsmdiskgrp Multi_Tier_Storage_Pool
id 28
name Multi_Tier_Storage_Pool
status online
mdisk_count 3
vdisk_count 1
.
easy_tier auto
easy_tier_status inactive
.
tier generic_ssd
tier_mdisk_count 0
.
tier generic_hdd
tier_mdisk_count 3

IBM_2145:ITS0-CLS5:admin>svcinfolsmdisk
mdisk_id mdisk_name status mdisk_grp_name capacity raid_level tier
299 SSD_Array_RAID5_1 online Multi_Tier_Storage_Pool 203.6GB raid5 generic_hdd
300 SSD_Array_RAID5_2 online Multi_Tier_Storage_Pool 203.6GB raid5 generic_hdd

IBM_2145:ITS0-CLS5:admin>svcinfolsmdisk SSD_Array_RAID5_2
mdisk_id 300
mdisk_name SSD_Array_RAID5_2
status online
mdisk_grp_id 28
mdisk_grp_name Multi_Tier_Storage_Pool
capacity 203.6GB
.
raid_level raid5
tier generic_hdd

IBM_2145:ITS0-CLS5:admin>svcinfolsmdiskgrp -filtervalue "name=Multi" ""
id name mdisk_count vdisk_count capacity easy_tier easy_tier_status
28 Multi_Tier_Storage_Pool 5 1 606.00GB auto inactive
```

```

IBM_2145:ITS0-CLS5:admin>svcinfolsmdiskgrp Multi_Tier_Storage_Pool
id 28
name Multi_Tier_Storage_Pool
status online
mdisk_count 5
vdisk_count 1
.
easy_tier auto
easy_tier_status inactive
.
tier generic_ssd
tier_mdisk_count 0
.
tier generic_hdd
tier_mdisk_count 5

```

11.5.4 Setting the disk tier

As shown in Example 11-3 on page 332, external MDisks that are detected have a default disk tier of `generic_hdd`. Easy Tier is also still inactive for the storage pool because we do not yet have a true multidisk tier pool. To activate the pool, set the SSD MDisks to their correct `generic_ssd` tier. Example 11-4 shows how to modify the SSD disk tier.

Example 11-4 Changing an SSD disk tier to generic_ssd

```

IBM_2145:ITS0-CLS5:admin>svcinfolsmdisk SSD_Array_RAID5_1
id 299
name SSD_Array_RAID5_1
status online
.
tier generic_hdd

IBM_2145:ITS0-CLS5:admin>svctask chmdisk -tier generic_ssd SSD_Array_RAID5_1
IBM_2145:ITS0-CLS5:admin>svctask chmdisk -tier generic_ssd SSD_Array_RAID5_2

IBM_2145:ITS0-CLS5:admin>svcinfolsmdisk SSD_Array_RAID5_1
id 299
name SSD_Array_RAID5_1
status online
.
tier generic_ssd

IBM_2145:ITS0-CLS5:admin>svcinfolsmdiskgrp Multi_Tier_Storage_Pool
id 28
name Multi_Tier_Storage_Pool
status online
mdisk_count 5
vdisk_count 1
.
easy_tier auto
easy_tier_status active
.
tier generic_ssd
tier_mdisk_count 2
tier_capacity 407.00GB

```

```
.
tier generic_hdd
tier_mdisk_count 3
```

11.5.5 Checking the Easy Tier mode of a volume

To check the Easy Tier operating mode on a volume, display its properties by using the **lsvdisk** command. An automatic data placement mode volume has its pool value set to on or auto, and the volume set to on. The CLI volume `easy_tier_status` is displayed as `active`, as shown in Example 11-5.

An evaluation mode volume has the pool and volume value set to on. However, the CLI volume `easy_tier_status` is displayed as `measured`, as shown in Example 11-2 on page 330.

Example 11-5 Checking a volume `easy_tier_status`

```
IBM_2145:ITS0-CLS5:admin>svcinfolsvdisk ITS0_Volume_10
id 28
name ITS0_Volume_10
mdisk_grp_name Multi_Tier_Storage_Pool
capacity 10.00GB
type striped
.
easy_tier on
easy_tier_status active
.
tier generic_ssd
tier_capacity 0.00MB
tier generic_hdd
tier_capacity 10.00GB
```

The volume in the example is measured by Easy Tier, and a hot extent migration is performed from the HDD tier MDisk to the SSD tier MDisk. Also, the volume HDD tier `generic_hdd` still holds the entire capacity of the volume because the `generic_ssd` capacity value is 0.00 MB. The allocated capacity on the `generic_hdd` tier gradually changes as Easy Tier optimizes the performance by moving extents into the `generic_ssd` tier.

11.5.6 Final cluster status

Example 11-6 shows the SAN Volume Controller cluster characteristics after you add multitiered storage (SSD with HDD).

Example 11-6 SAN Volume Controller multitier cluster

```
IBM_2145:ITS0-CLS5:admin>svcinfolcluster ITS0-CLS5
id 000002006A800002
name ITS0-CLS5
.
tier generic_ssd
tier_capacity 407.00GB
tier_free_capacity 100.00GB

tier generic_hdd
tier_capacity 18.85TB
tier_free_capacity 10.40TB
```

You now have two different tiers available in our SAN Volume Controller cluster: generic_ssd and generic_hdd. Now, extents are used on the generic_ssd tier and the generic_hdd tier. See the free_capacity values.

However, you cannot determine from this command if the SSD storage is being used by the Easy Tier process. To determine whether Easy Tier is actively measuring or migrating extents within the cluster, you must view the volume status, as shown in Example 11-5 on page 334.

11.6 Activating Easy Tier with the SAN Volume Controller GUI

This section describes how to activate Easy Tier by using the web GUI. This simple example uses a single storage pool that contains an MDisk that is presented from an XIV Gen 2 Storage System and a MDisk that is presented from an IBM FlashSystem 820 Storage System.

The environment is an SAN Volume Controller cluster with the following resources available:

- ▶ 1 x I/O group with two 2145-CF8 nodes
- ▶ 1 x external 400 GB FlashSystem LUN
- ▶ 1 x external 2.45TB XIV Gen 2 Storage System LUN

11.6.1 Setting the disk tier on MDisks

If we browse to the MDisks by Pools page (as shown in Figure 11-8), we can see that our storage pool set the Tier value of both MDisks to Hard Disk Drive.

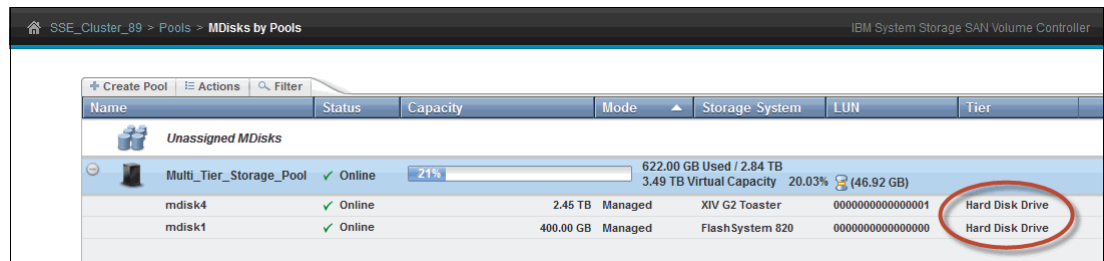


Figure 11-8 MDisks by Pools view, showing both MDisks with Tier set to Hard Disk Drive

When you view the properties of the storage pool, you can see that Easy Tier is inactive, as shown in Figure 11-9.

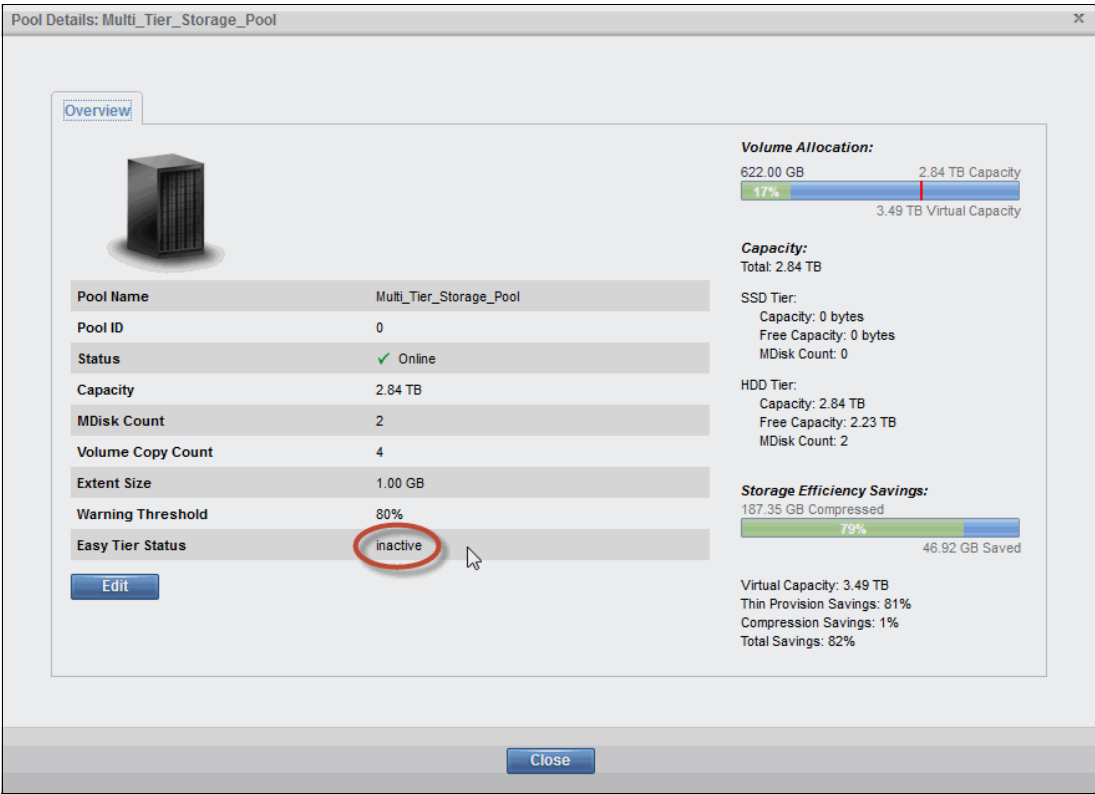


Figure 11-9 Pool properties page showing Easy Tier inactive status

Therefore, for Easy Tier to take effect, you must change the disk tier. Right-click the selected MDisk and choose **Select Tier**, as shown in Figure 11-10.

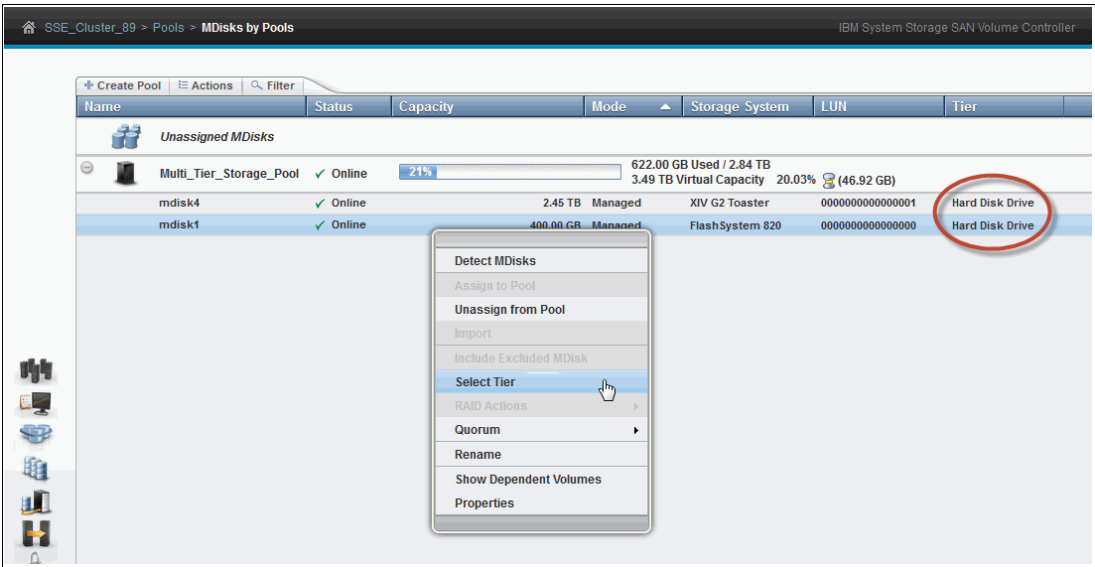


Figure 11-10 Selecting the Tier of an MDisk

Now set the MDisk Tier to **Solid-State Drive**, as shown in Figure 11-11.

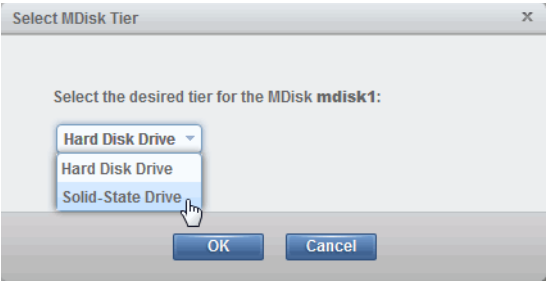


Figure 11-11 Setting Solid-State Drive tier value in the GUI

The MDisks now have the correct tier values, as shown in Figure 11-12.

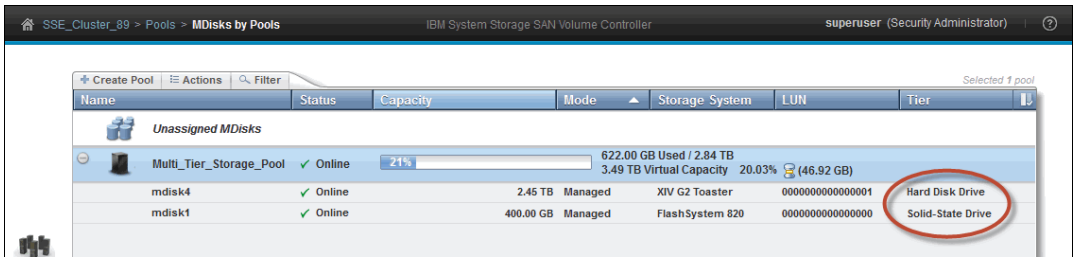


Figure 11-12 MDisks by Pools page, showing the correct tier values for the MDisks in the pool

11.6.2 Checking Easy Tier status

Now that the FlashSystem MDisk is correctly identified in the pool as an SSD, the Easy Tier function becomes active, as shown in Figure 11-13. For easy identification, the GUI displays a unique icon for storage pools in which Easy Tier is active.

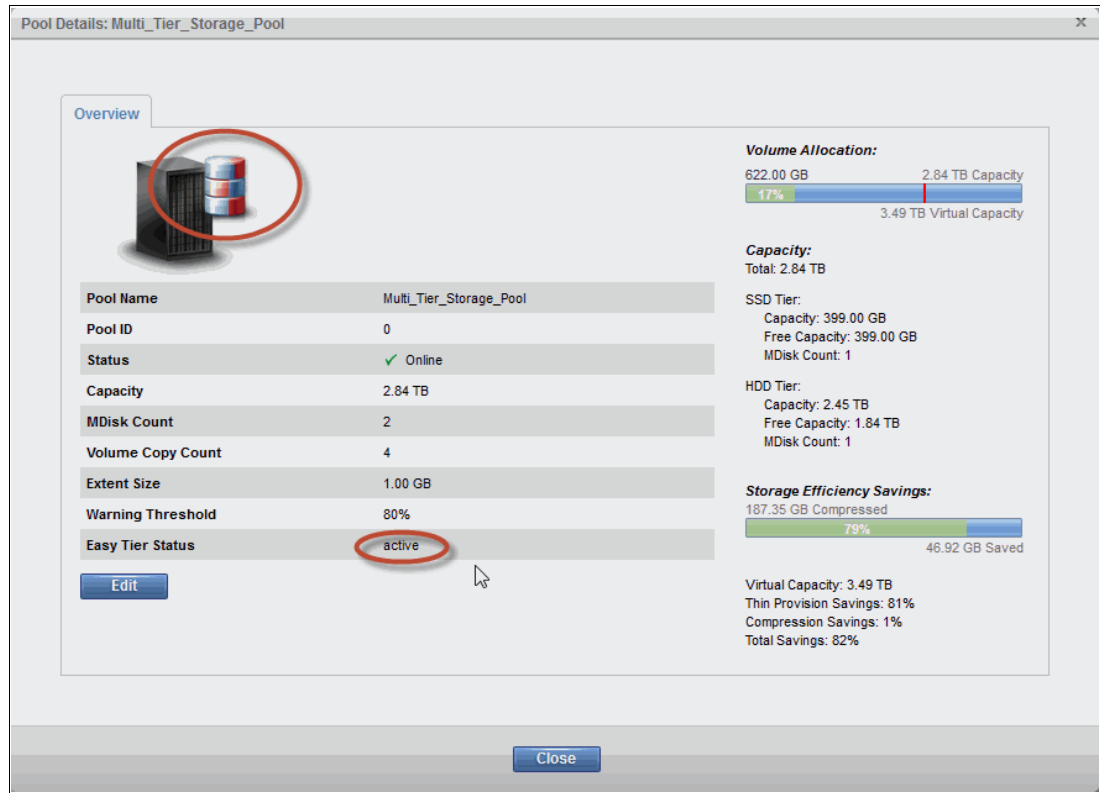


Figure 11-13 Storage pool with Easy Tier active

After the pool has an Easy Tier active status, the automatic data relocation process begins for the volumes in the pool, which occurs because the default Easy Tier setting for volumes is on. If we browse to the Volumes by Pool page, we can verify that Easy Tier is now active on the volumes in the pool, as shown in Figure 11-14.

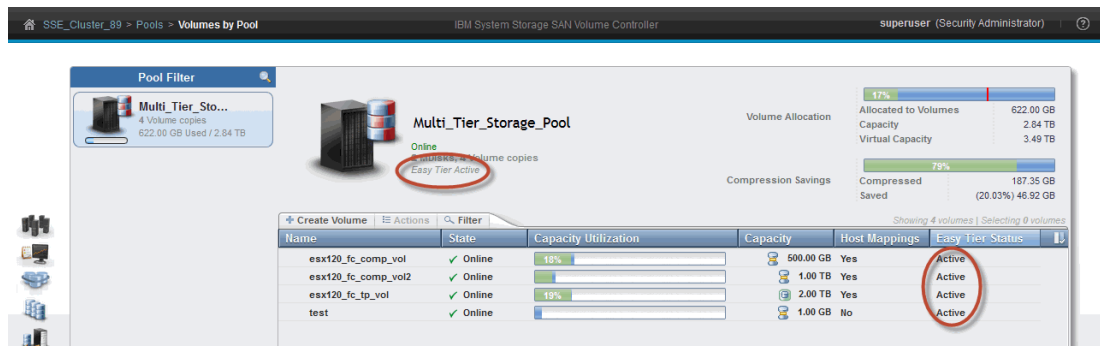


Figure 11-14 Viewing the Easy Tier status of volumes in a pool



Applications

This chapter provides information about laying out storage for the best performance for general applications; specifically, IBM AIX Virtual I/O Servers (VIOS), and IBM DB2® databases. Although most of the specific information is directed to hosts that are running the IBM AIX operating system, the information is also relevant to other host types.

This chapter includes the following sections:

- ▶ Application workloads
- ▶ Application considerations
- ▶ Data layout overview
- ▶ Database storage
- ▶ Data layout with the AIX Virtual I/O Server
- ▶ Volume size
- ▶ Failure domains

12.1 Application workloads

The following types of data workload (data processing) are possible:

- ▶ Transaction-based workloads
- ▶ Throughput-based workloads

These workloads are different by nature and must be planned for in different ways. Knowing and understanding how your host servers and applications handle their workload is an important part of being successful with your storage configuration efforts and the resulting performance.

A workload that is characterized by a high number of transactions per second and a high number of I/Os per second (IOPS) is called a *transaction-based* workload.

A workload that is characterized by a large amount of data that is transferred (normally with large I/O sizes) is called a *throughput-based* workload.

These two workload types are conflicting in nature and, therefore, require different configuration settings across all components that comprise the storage infrastructure. Generally, I/O (and therefore application) performance is optimal when I/O activity is evenly spread across the I/O subsystems that are dedicated to servicing the workload.

The following sections describe each type of workload in greater detail and explain what you can expect to encounter in each case.

12.1.1 Transaction-based workloads

High performance transaction-based environments cannot be created with a low-cost model of a storage system. Transaction process rates heavily depend on the number of operations per seconds that the back-end storage can deliver. This, in turn, depends on the technology (SSD versus HDD) and number of physical drives that are available for the storage subsystem controllers to use for parallel processing of host I/Os. The performance requirements of the workload are an important factor that you must consider when you are deciding on technology, number of physical drives that you need, and the potential use of Easy Tier technology.

Generally, transaction-intense applications also use a small *random data block pattern* to transfer data. With this type of data pattern, having more back-end drives enables more host I/Os to be processed simultaneously. The reason is that read cache is less effective than write cache, and the misses must be retrieved from the physical disks. Also, applications usually block on data reads, while delayed writes can be effectively hidden by the use of write cache.

In many cases, slow transaction performance problems can be traced directly to “hot” files that cause a bottleneck on a critical component (such as a single physical disk). This situation can occur even when the overall storage subsystem sees a fairly light workload. For this class of problems, it might be helpful to create a multitier storage pool and use the Easy Tier functionality of Storwize family systems (for more information, see Chapter 11, “IBM System Storage Easy Tier function” on page 319). When bottlenecks occur, they can be difficult and frustrating to resolve manually. Because workload content can continually change throughout the course of the day, these bottlenecks can be mysterious in nature. They can appear and disappear or move over time from one location to another location.

12.1.2 Throughput-based workloads

Throughput-based workloads are seen with applications or processes that require massive amounts of data sent. Such workloads often use large *sequential* blocks to reduce disk latency.

Few physical drives are needed to reach adequate I/O performance than with transaction-based workloads. For example, 20 - 28 physical drives are normally enough to reach maximum I/O throughput rates with the IBM System Storage DS4000 series of storage subsystems. In a throughput-based environment, read operations use the storage subsystem cache to stage greater chunks of data at a time to improve overall performance. Throughput rates heavily depend on the internal bandwidth of the storage subsystem. Newer storage subsystems with broader bandwidths can reach higher numbers and bring higher rates to bear.

12.1.3 Storage subsystem considerations

The selected storage subsystem model *must* support the required I/O workload. In addition to availability concerns, adequate performance must be ensured to meet the requirements of the applications, which include evaluation of the disk drive modules (DDMs) that are used and whether the internal architecture of the storage subsystem is sufficient.

The DDM characteristics must match the needs. In general, transaction-based workloads benefit from the use of SSDs or high rotation speed HDDs because for *random* I/Os, the important characteristic of the drive is the seek time. For throughput-based workloads, HDDs with lower rotation speed might be sufficient because of the sequential I/O nature.

As for the subsystem architecture, newer generations of storage subsystems have larger internal caches, higher bandwidth busses, and more powerful storage controllers.

12.1.4 Host considerations

When you discuss performance, you must consider more than the performance of the I/O subsystem only. Many settings within the host can affect the overall performance of the system and the applications it is running. Several of the performance-affecting settings and parameters that are available within the host operating system and the host bus adapters are described in Chapter 8, “Hosts” on page 225.

Many aspects of the configuration must be considered to ensure that we are not focusing on a result rather than the cause.

12.2 Application considerations

When you gather data for planning from the application side, first consider the workload type for the application.

If multiple applications or workload types share the system, you must know the type of workloads of each application. If the applications have both types or are mixed (transaction-based and throughput-based), you must know which workload is the most critical. Many environments have a mix of transaction-based and throughput-based workloads, and the transaction performance often is considered the most critical.

However, in some environments (for example, a Tivoli Storage Manager backup environment), the streaming high throughput workload of the backup is the critical part of the operation. Although a transaction-centered workload, the backup database is a less critical workload.

12.2.1 Transaction environments

Applications that use high transaction workloads are known as *online transaction processing* (OLTP) systems. Examples of these systems are database servers and mail servers.

If you have a database, you tune the server type parameters and the logical drives of the database to meet the needs of the database application. If the host server has a secondary role of performing nightly backups for the business, you need another set of logical drives. You must tune these drives for high throughput for the best backup performance you can get within the limitations of the parameters of the mixed storage subsystem.

The following sections describe the traits of a transaction-based application.

You can expect to see a high number of transactions and a fairly small I/O size. A typical transaction-based application is a database, which displays different I/O patterns for logs and table space access. Different databases use different I/O sizes for their logs; however, in all cases, the logs are generally high write-oriented workloads. For table spaces, most databases use 4 KB - 16 KB I/O size. In some applications, larger chunks (for example, 64 KB) are moved to host application cache memory for processing. Understanding how your application handles its I/O is critical to laying out the data properly on the storage server.

In many cases, the database *table space* is a large file that is made up of small blocks of data records. The records are normally accessed by using small I/Os of a random nature, which can result in about a 50% cache miss ratio. For this reason and to not waste cache space with unused data, plan for the Storwize family system to read and write data into cache in small chunks (use striped volumes with smaller extent sizes).

Another point to consider is whether the typical I/O is read or write. Most OLTP environments often have a mix of approximately 70% reads and 30% writes. However, the transaction logs of a database application have a much higher write ratio and, therefore, perform better in a different storage pool. Therefore, place the logs on a separate virtual disk (volume), which for best performance must be on a different storage pool that is defined to better support the heavy write need. Mail servers also frequently have a higher write ratio than read ratio.

Preferred practice: To avoid placing database table spaces, journals, and logs on the same back-end storage logical unit number (LUN) or RAID array, do not collocate them on the same MDisk or storage pool.

12.2.2 Throughput environments

With throughput workloads, you have fewer transactions, but much larger I/Os. I/O sizes of 128 K or greater are normal, and these I/Os often are sequential in nature. Applications that typify this type of workload are imaging, video servers, seismic processing, high performance computing (HPC), and backup servers.

With large-size I/O, it is better to use large cache blocks to write larger chunks into cache with each operation. You want the sequential I/Os to take as few back-end I/Os as possible and to get maximum throughput from them. Therefore, carefully decide how to define the logical drive and how to disperse the volumes on the back-end storage MDisks.

Many environments have a mix of transaction-oriented workloads and throughput-oriented workloads. Unless you measured your workloads, assume that the host workload is mixed and use Storwize family system striped volumes over several MDisks in a storage pool.

12.2.3 Performance tuning

For many, if not most, purposes, the following preferred practices provide the system with more than adequate performance. If you plan on tuning performance for a specific system, it is important to determine what is the bottleneck before you start tuning. To this end, you must measure system performance and use monitoring tools to analyze the OS, SAN fabric, and the Storwize family system performance during the test. Trying to guess the nature of the bottleneck can cause you to spend resources ineffectively or not lead to actual increase of performance.

12.3 Data layout overview

This section describes the document data layout from an operating system perspective. The objective is to help ensure that OS and storage administrators who are responsible for allocating storage understand how to lay out storage data, consider the virtualization layers, and avoid the performance problems and hot spots that can occur with poor data layout. The goal is to balance I/Os evenly across the physical disks in the back-end storage subsystems.

Specifically, you see how to lay out storage for DB2 applications as a useful example of how an application might balance its I/Os. The host data layout can have various implications, which are based on whether you use image mode or striped mode volumes for Storwize family system. You must also consider whether the storage pool you use employs Easy Tier.

12.3.1 Storage virtualization layers

Considered from the bottom up, the storage in a typical environment that is using Storwize family system includes the following layers:

- ▶ Disk drives (mechanical or solid-state based).
- ▶ RAID arrays of various levels (RAID 10, RAID 5, and so on). Some vendors use the term *array* to refer to a storage subsystem. Arrays often contain 2 - 32 disks; most often around 10 disks.
- ▶ LUNs (parts or complete RAID arrays) that are presented to and managed by Storwize family systems as MDisks.
- ▶ MDisk Groups combining one or more MDisks into a storage pool. MDisk Group is logically divided into extents.
- ▶ VDisks that are made of extents and presented to hosts.
- ▶ Operating system logical volume manager (LVM) combines VDisks that are presented to the system into a logical volume group. LVM volume group is logically divided into extents (the terminology differs for specific operating systems).

- Extents from the LVM volume group are used to define logical volumes. LVM logical volumes are used to store data on a file system that is defined on the volume or directly on the volume. Realistically sized logical volumes are backed up by many MDisk Group extents. This means that an LVM logical volume is striped across many MDisks, even if the OS LVM uses linear allocation policy for LVM extents.

The schematic representation of the storage virtualization layers is shown in Figure 12-1. Understanding of the layers of storage virtualization is important when you are making decisions about data layout in a system.

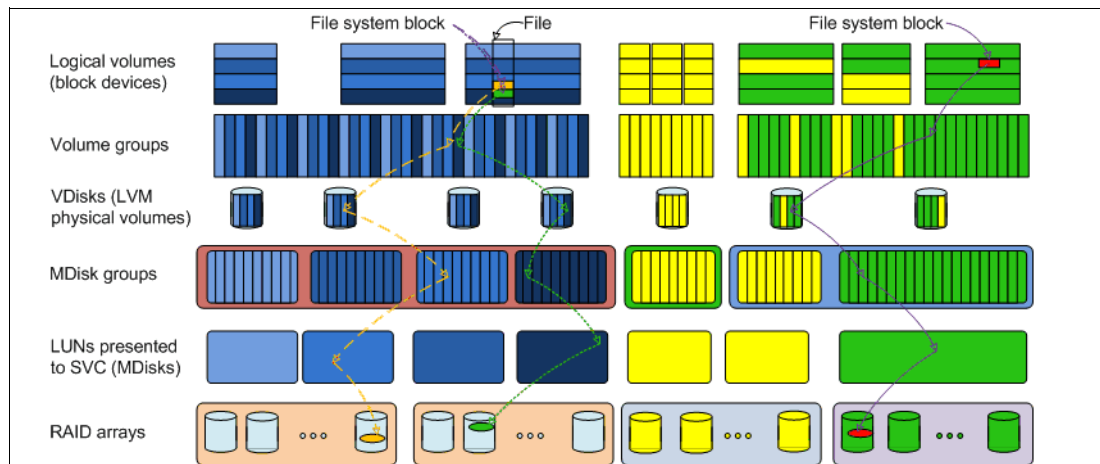


Figure 12-1 Storage virtualization layers

12.3.2 Virtualized storage characteristics

For decades, systems storage was based on mechanical drives that use rotational media. The introduction of multi-layer storage virtualization, solid-state based storage and technologies, such as Easy Tier and Real-Time Compression, introduced a new frame of reference when performance characteristics of storage subsystems are considered.

The following typical assumptions are made when data layout is considered:

- Sequential access to the storage is faster than random access.
- Full-stripe writes are preferred for RAID levels with defined parity (RAID 5 and RAID 6); therefore, the RAID chunk size and number of drives should be tuned to obtain the optimal stripe width.
- It is beneficial for performance to use many smaller volumes rather than fewer larger volumes.
- Seek operations (movement of the hard disk drive head assembly to the appropriate region of the disk) are time-consuming and therefore, the operating system should queue and reorder I/O operations (for example, the use of the elevator algorithm) to minimize disk head movements when data is stored or retrieved.

In a modern storage system, some or all of these assumptions can be false.

One of the main characteristics of solid storage technology is that random access to the data is not slower than sequential. For some storage solutions, the opposite might be true: back-end storage that uses SSDs performs better under workload that is characterized by random data access patterns than when it is accessed sequentially.

For Storwize family systems, the maximum I/O size for destage to back-end storage is 256 KB. Therefore, full stripe writes to a back-end storage are not feasible. However, for storage that is virtualized by a Storwize family system, this is not a problem. Even for internal storage, this is not a major concern. If there are many disk drive modules in the array, whether the I/Os that are issued by the hosts are full-stripe write or a partial-stripe write is not a major performance concern.

Multiple disks are presented to the operating system to enable sending I/Os in parallel to multiple independent disks. With virtualized storage, multiple disks that are presented to a host can be backed up by the same MDisk group and, ultimately, the same physical disks. In such circumstances, the only performance gains can be realized at the OS level, if the OS can process I/Os to the disks in parallel at all levels of the storage stack. Often, I/Os still are serialized at the FC card driver level.

In tests that were performed on AIX hosts, some increase in performance was observed when the number of physical volumes was increased from 1 to 2 and then to 4. Further, the increase of the number of physical volumes did not result in substantial performance increase. The performance gain can be to some extent explained by increased effective queue depth for the host (number of I/Os in flight at any moment). To realize performance gains that result from the usage of multiple storage systems in parallel, it is important to make sure that multiple VDisks that are presented to the host originate from separate back-end storage systems.

At the same time, OS-level storage layout can be simplified by presenting fewer larger volumes to the host, while making sure that these volumes are backed by multiple back-end storage systems. Such a setup realizes the benefits of parallelization of I/Os at the Storwize family system layer, which helps to take full benefit from investment in the storage systems and reduce planning complexity of the storage level at the OS level.

In virtualized storage environments, the OS has little to no knowledge of the mapping between logical addressing (LBA of a block) and the physical layout of data on physical disk drive modules.

For example, in case of thinly provisioned volumes (including volumes that are using Real-time Compression technology), sequential storage access as seen by the OS can result in random access pattern at the disk module level. Conversely, because of the temporal data locality, multiple random accesses to storage can result in few sequential accesses at the storage level. The problem can be exacerbated when the host's disk is virtualized by a hypervisor, such as Virtual I/O Server (VIOS) or a VMware host. In these circumstances, multiple hosts are issuing I/Os to the same volume simultaneously; therefore, I/O access pattern optimization that is performed at the level of distinct OS instances has little bearing on the access pattern at the physical disks level.

12.3.3 Storage, OS, and application administrator roles

Storage administrators control the configuration of the back-end storage subsystems and their RAID arrays (RAID type and number of disks in the array). Storage administrators normally also decide the layout of the back-end storage LUNs (MDisks), Storwize family system storage pools, volumes, and assignment of volumes to hosts.

OS administrators control the OS Logical Volume Manager (LVM) configuration, namely placement of Storwize family system volumes (LUNs) into LVM volume groups (VG). They also create logical volumes (LVs) and file systems within the VGs. These administrators have no control over where particular files or directories are placed within an LV, unless only one file or directory is in the LV.

Some applications (database engines in particular) also have built-in mechanisms for I/O management with which their administrators can control I/O access patterns. This ability is often used to send multiple I/Os in parallel by using striping across LVs that are available to the application.

Together, the storage administrator, OS administrator, and application administrator control how data is distributed among physical disks.

12.3.4 General data layout guidelines

When you lay out data on a Storwize family system back-end storage for general applications, use striped volumes across storage pools that consist of similar-type MDisk within a storage tier with as few MDisk as possible per RAID array. This general-purpose guideline applies to most Storwize family system back-end storage configurations and it removes a significant data layout burden for storage administrators.

Another important consideration is location of *failure domains* in the back-end storage. A failure domain is defined as what is affected if you lose a RAID array (a Storwize family system MDisk). If there is an MDisk failure, all of the volumes that are defined in the storage pool that contains this MDisk are affected. To reduce the size of a failure domain, it is beneficial to minimize the number of MDisk and storage systems in a specific storage pool.

Consider also that spreading the I/Os across multiple back-end storage systems has a performance benefit and a management benefit. It improves performance and reduces the workload on storage administrators because fewer entities exist in the managed systems.

Note: Increased performance and simplified storage management comes at a price of increasing the size of the failure domain. It is important to balance these aspects of storage system when you are designing the data layout.

If a company has several lines of business (LOBs), it can be an acceptable solution to align failure boundaries with LOBs so that each LOB has a unique set of back-end storage. Therefore, for each set of back-end storage (a group of one or more storage pools), we create volumes only within a well-defined set of back-end storage systems. This approach is beneficial because the failure domain is limited to a LOB, and performance and storage management is handled as a unit for the LOB independently.

Do *not* create striped volumes that are striped across different sets of back-end storage within the same storage tier. The use of different sets of back-end storage unbalances the I/O, which results in the nondeterministic performance of volumes that are defined in the pool and might limit the performance of volumes that are using the slowest back-end storage system.

For Storwize family storage system configurations in which you must use image mode volumes, the back-end storage configuration should consist of one LUN (and, therefore, one image mode volume) per array. Alternatively, equal number of LUNs should be defined per each array in the storage system. This way, it is possible to distribute the I/O workload evenly across the underlying physical disks of the arrays.

Note: Consider the following points about the preferred general data layout:

- ▶ Evenly balance I/Os across all physical disks (one method is by striping the volumes).
- ▶ To maximize sequential throughput, use a maximum range of physical disks for each LV when defining a logical volume on AIX host (by using the `mk1v -e x` AIX command).
- ▶ MDisk and volume sizes:
 - Create one MDisk per RAID array.
 - Create volumes that are based on the space that is needed and ensure that they are backed by a back-end storage system with sufficient performance.
- ▶ When you need more space on the server, dynamically extend the volume on the Storwize family storage system and then use the appropriate OS command to see the increased size in the system.

Use striped mode volumes if there is no recommendation to the contrary from the vendor of the application that is using the volumes. Striped volumes are the all-purpose volumes that provide good performance for most applications.

Manual, explicit I/O balancing for a specific application is also an option. However, this approach requires in-depth knowledge of the application and extensive testing. Therefore, this approach is much more time-consuming and the performance gains might not justify the time spent on fine-tuning the storage layout. Modifying the storage system (for example, increasing of the size of the volumes), might require another iteration of tests. Changing workload characteristics of a manually tuned system can result in an unbalanced configuration with suboptimal performance.

Some applications stripe their data across the underlying disks; for example, IBM DB2, IBM GPFS, and Oracle ASM. These types of applications might require more data layout considerations, as described in 12.3.6, “LVM volume groups and logical volumes” on page 349.

Storwize family systems striped mode volumes

Use striped mode volumes for applications that do not already stripe their data across disks. Striping at the VDisk level allows the OS administrator not to use striping at the LVM level. At the same time, you can use the available performance of the back-end storage system.

To ensure balanced access to all back-end storage systems and maximize parallelization of I/O processing, create VDIs that are striped evenly across all underlying MDIs in the storage pool. The size of the VDI should be a product of the number of MDIs that are defined in the storage pool and the extend size of the pool. You achieve best results with this approach if all MDIs in the pool have equal sizes.

Creating volumes that are striped across all RAID arrays in a storage pool simplifies OS level LVM configuration (there is no need for LV striping across physical volumes). Creating volumes that are striped across all RAID arrays in a storage pool is an excellent approach for most general applications.

Use striped volumes with the following considerations:

- ▶ Use extent sizes of 64 MB to maximize sequential throughput when necessary. A comparison of extent size and capacity is shown in Table 12-1 on page 348.

Table 12-1 Extent size versus maximum storage capacity

Extent size	Maximum storage capacity of Storwize family system cluster
16 MB	64 TB
32 MB	128 TB
64 MB	256 TB
128 MB	512 TB
256 MB	1 PB
512 MB	2 PB
1 GB	4 PB
2 GB	8 PB

- Use striped volumes when the number of volumes does not matter.
- Use striped volumes when the number of VGs does not affect performance.
- Use striped volumes when sequential I/O rates are greater than the sequential rate for a single RAID array on the back-end storage. Extremely high sequential I/O rates might require a different layout strategy.
- Use striped volumes when you prefer the use of large LUNs on the host.

For information about how to use large volumes, see 12.6, “Volume size” on page 351.

12.3.5 Throughput workloads

The goal for throughput workloads is to maximize the amount of data that is transferred per unit of time with an assumption of a sequential access pattern. The ideal situation is when each I/O stream is directed to a dedicated set of physical disk modules. Such workload is typical for some backup systems, extract, transform, and load (ETL) applications, and several scientific or engineering applications. However, often it is not appropriate for DB2 or Tivoli Storage Manager.

If you want to tune the system manually, you might consider creating image mode volumes on the Storwize family storage system and present as many of them to the host because many threads are performing I/O operations. Then, each thread can have a dedicated storage and sequential OS level access patterns translate to sequential access to the physical storage and performance is optimal. However, this scheme requires that all threads generate a similar amount of I/O activity or that you must tune the performance of each image mode volume to the requirements of a specific thread. This scheme also breaks down if the workload characteristic changes; for example, if the number of threads increases (you must provide a new image mode volume for the new threads) or decreases (you have unused storage resources).

In practice, you can achieve a similar level of performance by using Storwize family system striped volumes and by presenting the host with fewer, larger volumes. This approach simplifies the data layout at the OS level and provides good performance because of I/O access parallelization that is done at the Storwize family system level.

Tivoli Storage Manager uses a similar scheme as DB2 to spread out its I/O, but it also depends on ensuring that the number of client backup sessions is equal to the number of Tivoli Storage Manager storage volumes or containers. The perfect situation for Tivoli Storage Manager is some number of client backup sessions that go to some number of containers (with each container on a separate RAID array).

To summarize, if you have a good understanding of the application's I/O characteristics and can control the destination of data streams that are generated by the application, you might consider explicitly balancing the I/Os. In this case, use image mode volumes from Storwize family system to direct applications sequential data streams to independent physical disk sets. However, the use of Storwize family system striped volumes provides similar performance while simplifying OS level storage administration complexity. It is also less vulnerable to workload characteristics changes.

12.3.6 LVM volume groups and logical volumes

Without a Storwize family system managing the back-end storage, the administrator must ensure that the host operating system aligns its partitions or *slices* with physical storage systems. Misalignment can result in numerous boundary crossings that are responsible for unnecessary multiple drive I/Os. Certain operating systems perform this alignment automatically, and you must know the alignment boundary that they use. However, other operating systems might require manual intervention to set their start point to a value that aligns them.

With a Storwize family system, managing the storage and presenting it to the hosts storage alignment is not a problem. Although misalignment between OS level storage configuration and the back-end storage results in performance penalty, such problems are not reported by users of Storwize family systems.

Understanding how your host-based volume manager (if used) defines and uses the VDisks that are presented to the host is also an important part of the data layout. Volume managers group VDisks (physical volumes) into volume groups. The volume manager then creates volumes by carving up the logical drives into *partitions* (sometimes referred to as *slices* or *extents*, depending on the operating system) and then builds a logical volume from them by striping or concatenating them to achieve the wanted volume size.

How partitions are selected for use and laid out can vary from system to system. In all cases, you must ensure that spreading the partitions is done in a manner with which the maximum I/Os can be made available to the logical drives in the group. This can be achieved by striping the logical volume across all physical volumes that are available in the volume group. However, this approach can make it more difficult to add volumes to the volume group. Therefore, an easier solution is to present striped VDisks to the host and define non-striped volumes at the OS logical volume manager level. This approach gives optimum performance and simplifies the management of storage at the OS level.

Logical volumes in a specific volume group should be expected to compete for resources at the physical storage level. Therefore, when the application requires that the I/Os from distinct streams be issued to different physical disks, separate hierarchies of volume groups, VDisks, or MDisk should be created for each I/O stream.

12.4 Database storage

In a world with networked and highly virtualized storage, correct database storage design can seem like a dauntingly complex task for a DBA or system architect to accomplish.

Poor database storage design can have a significant negative effect on a database server. Processors are so much faster than physical disks that it is common to find poorly performing database servers that are I/O bound and underperforming by many times their potential.

Fortunately, it is not necessary to get database storage design perfectly correct. Understanding the makeup of the storage stack and manually tuning the location of database tables and indexes on parts of different physical disks is generally not achievable or maintainable by the average DBA in today's virtualized storage world. However, it is also no longer necessary so that good performance is realized. It is also unnecessary to manage multiple small volumes and define complex storage layouts in a database. In practice, the use of several (approximately four) sufficiently large volumes that are backed by sufficiently performing back-end storage results in good performance and significantly simplifies storage administration.

Simplicity is the key to good database storage design. The basics involve ensuring enough physical disks to keep the system from becoming I/O bound.

For more information, basic guidance, advice for a healthy database server, and easy-to-follow preferred practices in database storage, see "Best Practices: Database Storage" at this website:

<http://www.ibm.com/developerworks/data/bestpractices/databasesstorage/>

12.5 Data layout with the AIX Virtual I/O Server

This section describes strategies that you can use to achieve the best I/O performance by evenly balancing I/Os across physical disks when the VIOS is used.

12.5.1 Overview

In setting up storage at a VIOS, several possibilities exist for creating volumes and serving them to VIO clients (VIOCs). The first consideration is to create sufficient storage for each VIOC. Less obvious, but equally important, is obtaining the best use of the storage. Performance and availability are also significant. Internal Small Computer System Interface (SCSI) disks (which are used for the VIOS operating system) and SAN disks often are available. Availability for disk is usually handled by RAID on the SAN or by SCSI RAID adapters on the VIOS.

Here, it is assumed that any internal SCSI disks are used for the VIOS operating system and possibly for the operating systems of the VIOC. Furthermore, the applications are configured so that the limited I/O occurs to the internal SCSI disks on the VIOS and to the rootvg of the VIOC. If you expect your rootvg might have a significant IOPS rate, you can configure it in the same manner as for other application VGs later.

Remember to keep Storwize family system drivers (SDDPCM) on the VIOS system up to date. The use of outdated drivers deprives you of benefits of updates that are in the new version and potentially exposes your VIOS client to bugs that were fixed in new versions.

VIOS restrictions

You can create the following types of volumes on a VIOS:

- **NPIV volumes**

By using NPIV volumes, you can natively present storage from a Storwize family storage system to a logical partition defined in a Power system. This approach allows the OS administrator to access the storage system by using its native drivers and, for example, the use of its FlashCopy capabilities. When an NPIV mapping is defined for a logical partition, VIOS defines two sets of WWPN addressees for the client. If you plan to use live partition mobility, always remember to map the storage to both sets of WWPN addressees. The drawback of the NPIV approach is that you need to maintain OS level drivers for the storage on each LPAR, which is using NPIV to access the storage system.

- **Physical volume (PV) VSCSI hdisks**

PV VSCSI hdisks are entire LUNs from the VIOS perspective that are presented to the client, which sees them as physical volumes.

If you are concerned about failure of a VIOS and configured redundant VIOS for that reason, you must use PV VSCSI or NPIV hdisks.

- **Logical volume (LV) VSCSI hdisks**

An LV VSCSI hdisk is a logical volume that is presented by a VIOS to its client, who sees it as a physical volume. LV VSCSI hdisks cannot be served from multiple VIOSs. LV that is presented to a client cannot span physical volumes or be a striped logical volume.

VIOS queue depth

The default value of `queue_depth` parameter for VSCSI hdisks in a VIOS client is set to three. This limits the system to approximately 300 (assuming an average I/O service time of 10 ms). Therefore, it is advisable to tune the `queue_depth` parameter to a higher value to get the IOPS bandwidth that is needed. When possible, set the queue depth of the VIOC hdisks to match that of the VIOS hdisk to which it maps. Changing the queue depth setting is a disruptive operation; therefore, you must schedule a maintenance window if you want to modify this setting in a production system.

For or IBM i client logical partitions, the queue depth value is 32 and cannot be changed.

12.5.2 Data layout strategies

You can use the Storwize family system or AIX LVM (with appropriate configuration of VSCSI disks at the VIOS) to balance the I/Os across the back-end physical disks. The simpler solution is to use the capabilities of the Storwize family system because then you can present striped VDisks to the VIOS and remove the burden of load balancing from the OS administrators.

12.6 Volume size

The use of few larger volumes instead of many smaller volumes might require more disk buffers and larger `queue_depths` per volume to obtain optimum performance. However, a significant benefit is using less AIX memory and fewer path management resources. Therefore, tune the `queue_depth`, `num_cmd_elements`, and `max_xfer_size` parameters to appropriately large values. These changes require a system restart, so plan for a maintenance window if you must change them in a production system.

12.7 Failure domains

As described in 12.3.4, “General data layout guidelines” on page 346, consider failure domains in the back-end storage configuration. If all LUNs are spread across all physical disks (by LVM or Storwize family system volume striping), and you experience a single RAID array failure, you lose access to all your data. Therefore, in some situations, you might want to limit the spread for certain applications or groups of applications or use volume mirroring at the Storwize family system or LVM level.

When you are determining your failure domains, pay attention to the system configuration and application dependencies. You might have a group of applications where, if one application fails, none of the applications can perform any productive work.

When you are implementing the Storwize family system, limiting the size of a failure domain can be achieved through an appropriate storage pool layout. For more information about failure boundaries in the back-end storage configuration, see Chapter 5, “Storage pools and managed disks” on page 95.

12.8 More resources

This section describes some other resources that might be useful.

12.8.1 IBM System Storage Interoperation Center

The IBM System Storage Interoperation Center (SSIC) is available at this website:

<http://www-03.ibm.com/systems/support/storage/ssic/interoperability.wss>

The SSIC is an invaluable resource when you design a solution that includes Storwize family storage product. You should refer to this site to verify that all elements of the solution are compatible or to identify the requirements of the environment that must be met to ensure that the environment functions correctly.

12.8.2 Techdocs - the Technical Sales Library

IBM provides public access to a resource that is called Techdocs - the Technical Sales Library, which is available at this website:

<http://www-03.ibm.com/support/techdocs/atmastr.nsf/Web/TechDocs>

This website is a gateway to a collection of technical white papers that can be valuable when you are preparing for the design and implementation of a solution that uses the Storwize family storage system.

The documents that are listed next are from this library. It is not exhaustive list and you are encouraged to search through the Technical Sales Library on your own. Information that is found in the white papers that are dedicated to SAN Volume Controller often pertains to Storwize systems and vice versa.

12.8.3 DB2 white papers

The following DB2 white papers are available:

- ▶ *Reference Architecture: SAP on IBM eX5 enterprise systems, Storwize V7000 / SAN Volume Controller & VMware:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101998>
- ▶ *IBM Storage Reference Architecture for SAP landscapes featuring IBM FlashSystem:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102361>
- ▶ *Best Practices for Tivoli Storage FlashCopy Manager in a SVC Metro Mirror environment:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102196>
- ▶ *Quick-start guide to FlashCopy Manager for SAP with IBM Storwize V7000, IBM SAN Volume Controller or IBM DS8000 Storage System:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101627>
- ▶ *Inter-data center workload mobility and failover with VMware vSphere and IBM System Storage SAN Volume Controller:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101923>
- ▶ *IBM SAN Volume Controller Thin Provisioning and Oracle ASM: SVC Space Efficient VDisks:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101862>
- ▶ *IBM SAN Volume Controller 6.3 Advanced Copy Services for Backup and Recovery of Oracle 11.2 RAC/ASM Databases:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102080>
- ▶ *Planning for Easy Tier with IBM Storwize V7000 and SAN Volume Controller:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102295>
- ▶ *IBM SVC Stretched Cluster for Oracle RAC:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102280>
- ▶ *IBM System Storage Architecture and Configuration Guide for SAP HANA TDI (tailored datacenter integration) V1.4:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102347>
- ▶ *Accelerate with ATS: New Hardware Extensions for SVC CG8 Nodes:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS5183>
- ▶ *SVC Node Replacement and Node Addition Procedures:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD104437>
- ▶ *Protecting Exchange 2013 with TSM FlashCopy Manager 3.2.1 on IBM Storwize family products and SVC:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102328>
- ▶ *Accelerate with ATS: Best Practices for VMware using IBM storage:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS4982>
- ▶ *IBM Storwize V7000 Exchange Server 2010 NetBackup 7 VSS Solution:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101762>
- ▶ *IBM Storwize V7000 SQL Server NetBackup 7 VSS Solution:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101763>

- ▶ *How to configure IBM VSS hardware provider on VMware ESXi 5 for FlashCopy (IBM Tivoli FCM):*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD106013>
- ▶ *PowerHA Hardware Support Matrix:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD105638>

12.8.4 Oracle white papers

The following Oracle white papers are available:

- ▶ *Performance benefits of IBM Storwize V7000 with Easy Tier for Oracle 11g workload:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101838>
- ▶ *Performance benefits of IBM Storwize V7000 with IBM Easy Tier for Oracle ASM:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101990>
- ▶ *Deploying Oracle RAC 11g R2 on IBM Storwize V7000 with IBM GPFS 3.3:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101991>
- ▶ *Oracle DB 11gR2 and RAC on IBM Flex System™ p460 Compute Nodes with PowerVM® and IBM Storwize V7000 Storage System:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102124>
- ▶ *Oracle Database 11g Release 2 with Oracle RAC on IBM Flex System Compute Nodes with PowerVM and IBM Storwize V7000 storage system:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102342>
- ▶ *Deploying Oracle 11g RAC Release 2 with IBM Storwize V7000 on Red Hat Enterprise Linux:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101772>
- ▶ *Disaster Recovery solution for Oracle 11gR2 database using Metro Mirror and Global Mirror features of IBM Storwize V7000:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101881>
- ▶ *Safely upgrading an Oracle database using Remote Copy on the IBM Storwize V7000 storage system:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102121>
- ▶ *IBM Storwize V7000 Real-time Compression volumes with Oracle 11g R2 databases:*
<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102149>

12.8.5 Diskcore and Tapecore mailing lists

Open to IBM employees and IBM Business Partners only, the Diskcore and Tapecore mailing lists are for discussions or requests about disk or tape storage and do not replace (or allow you to bypass) normal methods for technical support.

Many experts in their fields from within and without IBM post to these mailing lists and are happy to answer questions at this website:

<https://www.ibm.com/developerworks/community/blogs/diskcore/?lang=en>



Part 3

Management, monitoring, and troubleshooting

This part provides information about preferred practices for monitoring, managing, and troubleshooting your installation of SAN Volume Controller.

This part includes the following chapters:

- ▶ Chapter 13, “Monitoring” on page 357
- ▶ Chapter 14, “Maintenance” on page 485
- ▶ Chapter 15, “Troubleshooting and diagnostics” on page 519



Monitoring

Tivoli Storage Productivity Center offers several reports that you can use to monitor SAN Volume Controller and Storwize family products and identify performance problems. Tivoli Storage Productivity Center version 5.2 is a major release that provides improvements to the web-based user interface that is designed to offer ease of use access to your storage environment. This interface also provides a common appearance that is based on the current user interfaces for IBM XIV Storage System, IBM Storwize V7000, and IBM System Storage SAN Volume Controller. For more information about Tivoli Storage Productivity Center version 5.2, see this website:

http://pic.dhe.ibm.com/infocenter/tivihelp/v59r1/topic/com.ibm.tpc_V521.doc/tpc_kc_homepage.html

This chapter describes how to use the product for monitoring. It includes examples of misconfiguration and failures. Then, it explains how you can identify them in Tivoli Storage Productivity Center by using the Topology Viewer and performance reports. This chapter also shows how to collect and view performance data directly from the SAN Volume Controller.

You should always use the latest version of Tivoli Storage Productivity Center that is supported by your SAN Volume Controller code. Tivoli Storage Productivity Center is often updated to support new SAN Volume Controller features. If you have an earlier version of Tivoli Storage Productivity Center installed, you might still be able to reproduce the reports that are described in this chapter, but some data might not be available.

This chapter includes the following sections:

- ▶ Analyzing the SAN Volume Controller and Storwize Family Storage Systems by using Tivoli Storage Productivity Center
- ▶ Considerations for performance analysis
- ▶ Top 10 reports for SAN Volume Controller and Storwize V7000
- ▶ Reports for fabric and switches
- ▶ Case studies
- ▶ Monitoring in real time by using the SAN Volume Controller or Storwize V7000 GUI
- ▶ Manually gathering SAN Volume Controller statistics

Note: In Tivoli Storage Productivity Center version 5.2, certain reporting and monitoring capabilities are only available via the Tivoli Storage Productivity Center stand-alone GUI, while others might be available in the stand-alone GUI or web-based GUI. In some scenarios, the version 5.2 stand-alone GUI might provide more robust reporting and monitoring capabilities than the web-based GUI. Where possible in this chapter, we include examples that use the Tivoli Storage Productivity Center version 5.2 web-based GUI. Otherwise, examples are provided that use the Tivoli Storage Productivity Center version 5.2 stand-alone GUI.

13.1 Analyzing the SAN Volume Controller and Storwize Family Storage Systems by using Tivoli Storage Productivity Center

In this section, we describe analyzing by using the Tivoli Storage Productivity Center 5.2 web-based GUI.

13.1.1 Analyzing with the Tivoli Storage Productivity Center 5.2 web-based GUI

Tivoli Storage Productivity Center provides a large amount of detailed information about SAN Volume Controller and Storwize family products. The examples in this section show how to access this information for a SAN Volume Controller with the Tivoli Storage Productivity Center web-based GUI. The examples assume that the SAN Volume Controller cluster was added to Tivoli Storage Productivity Center. For more information about the installation, configuration, and administration of Tivoli Storage Productivity Center (including how to add a storage system), see this website:

http://pic.dhe.ibm.com/infocenter/tivihelp/v59r1/topic/com.ibm.tpc_V521.doc/tpc_kc_homepage.html

Viewing Storage Systems

In the Tivoli Storage Productivity Center 5.2 web-based GUI, browse to the Storage Systems view, as shown in Figure 13-1.



Figure 13-1 Browsing to the Storage Systems view

Tip: All tabular data views within the Tivoli Storage Productivity Center version 5.2 web-based GUI can be exported to CSV, PDF, or HTML file versions by using the Actions menu, as shown in Figure 13-2 on page 360.

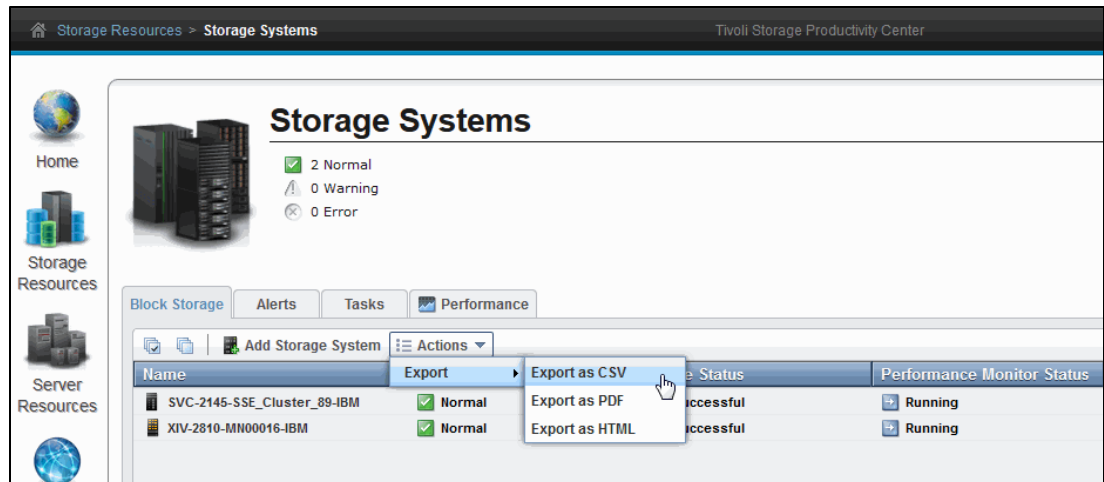


Figure 13-2 Exporting tabular reports

In the Storage Systems view, double-click the wanted storage system, as shown in Figure 13-3.



Figure 13-3 Storage Systems view

The Overview window of the SAN Volume Controller storage system is displayed, as shown in Figure 13-4. From this window, you can browse to various reports about many aspects of the storage system.

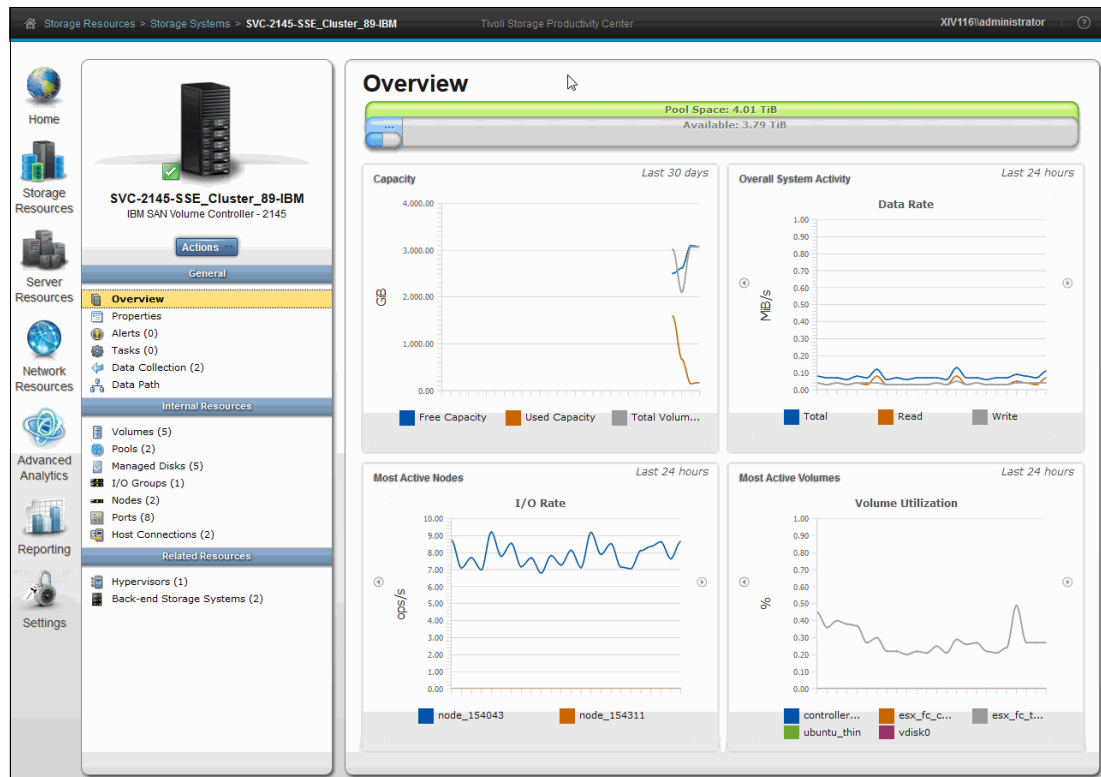


Figure 13-4 Overview page of a SAN Volume Controller

Viewing SAN Volume Controller or Storwize storage pools)

Click **Pools** under the Internal Resources section, as shown in Figure 13-5.



Figure 13-5 Accessing the Pools view for a storage system

The Pools page opens for this storage system, which displays the storage pools for this system in a tabular format with various details, as shown in Figure 13-6.

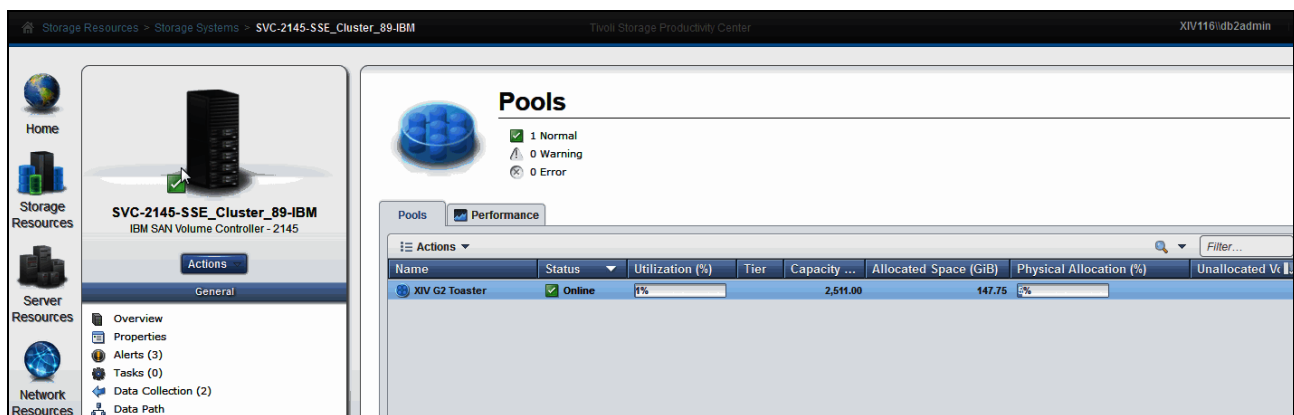


Figure 13-6 The Pools page for a storage system

Right-click any of the column headings to modify the columns that are displayed, as shown in Figure 13-7 on page 363.

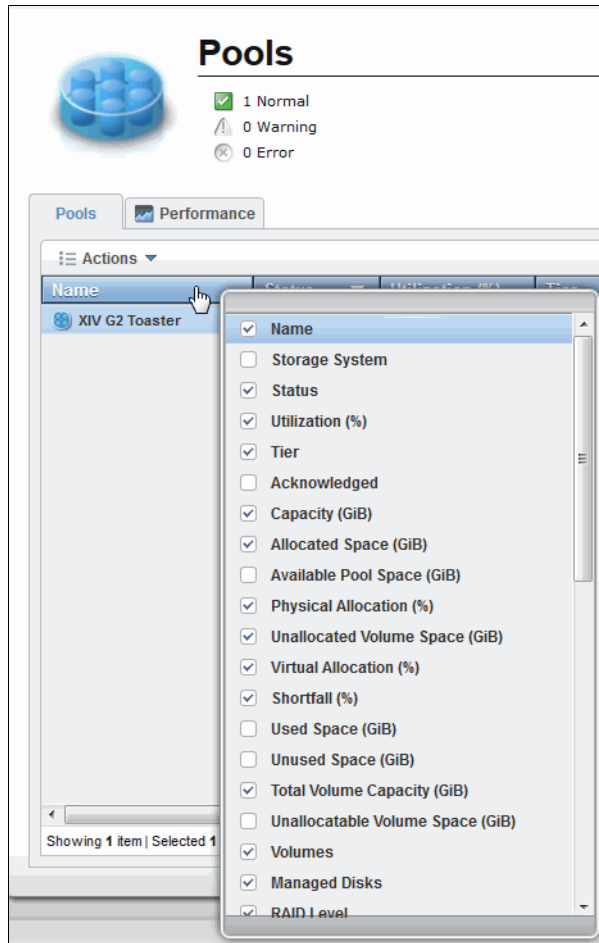


Figure 13-7 Modifying Pools column headings

Tip: You can modify the columns that are displayed in any tabular view in the Tivoli Storage Productivity Center 5.2 web GUI by right-clicking the column headings, as shown in Figure 13-7.

By using the Pools window, you can view the following details about the pools:

- Name

The name of a pool that uniquely identifies it within a storage system.

- Storage System

The name of the storage system that contains a pool. This name was defined when the storage system was added to Tivoli Storage Productivity Center. If a name was not defined, the ID of the storage system is displayed.

- Status

The status of a pool. Statuses include Online, Offline, Warning, Error, and Unknown. Use the status to determine the condition of a pool, and if any actions must be taken. For example, if a pool has an Error status, take immediate action to correct the problem.

► Utilization (%)

The average daily utilization of the pool. The utilization of a pool is based on an estimate of the average daily utilization of storage resources, such as controllers, device adapters, and hard disks. The value for the utilization of the pool is estimated based on the performance data that was collected on the previous day.

► Tier

The tier level of pools on storage virtualizers. If the pool is not assigned a tier level, the cell is left blank. To set or change the tier level, select one or more pools, right-click, and then select **Set Tier**.

► Acknowledged

Shows whether the status of a pool is acknowledged. An acknowledged status indicates that the status was reviewed and is resolved or can be ignored. An acknowledged status is not used the status of related, higher-level resources is determined.

For example, if the status of a pool is Error, the status of the storage system that contains it is also Error. If the Error status for the pool is acknowledged, its status is not used to determine the overall status of the associated storage system. In this case, if the other internal resources of the storage system are Normal, the status of the storage system is also Normal.

► Capacity (GB)

The total amount of storage space in a pool. For XIV systems, this value represents the physical (hard) capacity of the pool, not the virtual (soft) capacity. For other storage systems, this value might also include overhead space if the pool is unformatted.

► Allocated Space (GB)

The amount of space that is reserved for all the volumes in a pool, which includes thin-provisioned and standard volumes.

The space that is allocated for thin-provisioned volumes is less than their virtual capacity, which is shown in the Total Volume Capacity column. If a pool does not contain thin-provisioned volumes, this value is the same as Total Volume Capacity.

This value is equal to Used Space for the following storage systems:

- Storage systems that are not SAN Volume Controller and Storwize V7000.
- SAN Volume Controller and Storwize V7000 storage systems that are not thin-provisioned.

► Available Pool Space (GB)

The amount of space in a pool that is not reserved for volumes.

► Physical Allocation (%)

The percentage of physical space in a pool that was reserved for volumes. This value is always less than or equal to 100% because you cannot reserve more physical space than is available in a pool. The following formula is used to calculate this value:

$$\text{Allocated Space} / \text{Pool Capacity} * 100$$

For example, if the physical allocation percentage is 25% for a total storage pool size of 200 GB, the space that is reserved for volumes is 50 GB.

The first section of the bar in the Pools window uses the color blue and a percent (%) sign to represent the physical allocation percentage. The second section of the bar uses the color gray to represent the pool capacity. Hover the mouse pointer over the percentage bar to view the following values:

- Allocated Space: The amount of space that is reserved for all the volumes in a pool, which includes both thin-provisioned and standard volumes.
- Capacity: The total amount of space in a pool.

► Unallocated Volume Space (GB)

The amount of the Total Volume Capacity in the pool that is not allocated.

- Virtual Allocation (%): The percentage of physical space in a pool that was committed to the total virtual capacity of the volumes in the pool. In thin-provisioned environments, this percentage exceeds 100% if a pool is over committed (over provisioned). The following formula is used to calculate this value:

$$\text{Total Volume Capacity} / \text{Pool Capacity} * 100$$

This value is available only for pools with thin-provisioned volumes.

For example, if the allocation percentage is 200% for a total storage pool size of 15 GB, the virtual capacity that is committed to the volumes in the pool is 30 GB. This configuration means that twice as much space is committed than is physically contained in the pool. If the allocation percentage is 100% for the same pool, the virtual capacity that is committed to the pool is 15 GB. This configuration means that all the physical capacity of the pool is allocated to volumes.

An allocation percentage that is higher than 100% is considered aggressive because there is insufficient physical capacity in the pool to satisfy the maximum allocation for all the thin-provisioned volumes in the pool. In such cases, you can use the value for Shortfall (%) to estimate how critical the shortage of space is for a pool.

Hover the mouse pointer over the percentage bar to view the following values:

- Total Volume Capacity: The total storage space on all the volumes in a pool. For thin-provisioned volumes, this value includes virtual space.
- Capacity: The total amount of space in a pool.

- Shortfall (%): The percentage of the remaining unallocated volume space in a pool that is not available to be allocated. The higher the percentage, the more critical the shortfall of pool space. This percentage is available only for a pool with thin-provisioned volumes.

The following formula is used to calculate this value:

$$\text{Unallocatable Space} / \text{Committed but Unallocated Space} * 100$$

You can use this percentage to determine when the amount of over-committed space in a pool is at a critically high level. Specifically, if the physical space in a pool is less than the committed virtual space, the pool does not have enough space to fulfill the commitment to virtual space. This value represents the percentage of the committed virtual space that is not available in a pool. As more space is used over time by volumes while the pool capacity remains the same, this percentage increases.

For example, the remaining physical capacity of a pool is 70 GB, but 150 GB of virtual space was committed to thin-provisioned volumes. If the volumes are using 50 GB, there is still 100 GB committed to the volumes (150 GB - 50 GB) with a shortfall of 30 GB (70 GB remaining pool space - 100 GB remaining commitment of volume space to the volumes).

Because the volumes are overcommitted by 30 GB based on the available space in the pool, the shortfall is 30% when the following calculation is used:

$$\frac{100 \text{ GB unallocated volume space} - 70 \text{ GB remaining pool space}}{100 \text{ GB unallocated volume space}} * 100$$

The first section of the bar in the Pools window uses the color blue and a percent (%) sign to represent the shortfall percentage. The second section of the bar uses the color gray to represent the unallocated volume space. Hover the mouse pointer over the percentage bar to view the following values:

- Unallocated Volume Space: The amount of the Total Volume Capacity in the pool that is not allocated.
- Available Pool Space: The amount of space in a pool that is not reserved for volumes.

► **Used Space (GB)**

The amount of allocated space that is used by the volumes in a pool, which includes thin-provisioned and standard volumes.

For SAN Volume Controller and Storwize V7000, you can pre-allocate thin-provisioned volume space when the volumes are created. In these cases, the Used Space might be different from the Allocated Space for pools that contain thin-provisioned volumes. For pools with compressed volumes on SAN Volume Controller and Storwize V7000, the Used Space reflects the size of compressed data that is written to disk. As the data changes, the Used Space can, at times, be less than the Allocated Space.

For pools with volumes that are not thin provisioned or compressed in SAN Volume Controller, Storwize V7000, and other storage systems, the values for Used Space and Allocated Space are equal.

This value is accurate as of the most recent time that Tivoli Storage Productivity Center collected data about a pool. Because data collection is run on a set schedule and the used space on volumes can change rapidly, the value in this column might not be 100% accurate for the current state of volumes.

► **Unused Space (GB)**

The amount of space that is allocated to the volumes in a pool and is not yet used. The following formula is used to calculate this value:

$$\text{Allocated Space} - \text{Used Space}$$

This value is available only for SAN Volume Controller and Storwize V7000 pools.

► **Total Volume Capacity (GB)**

The total storage space on all the volumes in a pool, which includes thin-provisioned and standard volumes. For thin-provisioned volumes, this value includes virtual space.

► **Unallocatable Volume Space (GB)**

The amount of space by which the Total Volume Capacity exceeds the physical capacity of a pool. The following formula is used to calculate this value:

$$\text{Total Volume Capacity} - \text{Pool Capacity}$$

In thin-provisioned environments, it is possible to over commit (over provision) storage in a pool by creating volumes with more virtual capacity than can be physically allocated in the pool. This value represents the amount of volume space that cannot be allocated based on the current capacity of the pool.

► **Volumes**

The number of volumes that are allocated from a pool.

► **Managed Disks**

The number of managed disks that are assigned to a pool. This value is available only for SAN Volume Controller and Storwize V7000 pools.

► **RAID Level**

The RAID level of the pool, such as RAID 5 and RAID 10. The RAID level affects the performance and fault tolerance of the volumes that are allocated from the pool. In some cases, there might be a mix of RAID levels in a pool. The RAID levels in a mixed pool are shown in a comma-separated list.

► **Extent Size (MB)**

The extent granularity that was specified when a pool was created. Smaller extent sizes limit the maximum size of the volumes that can be created in a pool, but minimize the amount of potentially wasted space per volume. This value is available only for SAN Volume Controller and Storwize V7000 pools.

► **Compression savings (%)**

The percentage of capacity saving in a pool by using compression feature. This information is available only for SAN Volume Controller and Storwize V7000 volumes.

► **Solid-State**

Shows whether a pool contains solid-state disk drives. If a pool contains solid-state disks and other disks, the value Mixed is shown.

► **Assigned Volume Space (GB)**

The space on all the volumes in a pool that are mapped or assigned to host systems. For a thin-provisioned pool, this value includes the virtual capacity of thin-provisioned volumes, which might exceed the total space in the pool.

► **Unassigned Volume Space (GB)**

The space on all the volumes in a pool that are not mapped or assigned to host systems. For a thin-provisioned pool, this value includes the virtual capacity of thin-provisioned volumes, which might exceed the total space in the pool.

► **Easy Tier**

Shows how the Easy Tier function is enabled on a pool. The following values are possible:

- Enabled/Inactive
- Enabled/Active
- Auto/Active
- Auto/Inactive
- Disabled
- Unknown

► **Tier Capacity SSD (GB)**

The total storage space on the solid-state drives that are participating in, or are eligible to participate in, the Easy Tier optimization for a pool.

► **Tier Capacity HDD (GB)**

The total storage space on the hard disk drives that are participating in, or are eligible to participate in, the Easy Tier optimization for a pool.

► **Tier Available Space SSD (GB)**

The unused storage space on the solid-state drives in a pool.

► **Tier Available Space HDD (GB)**

The unused storage space on the hard disk drives in a pool.

► **Back-End Storage System Type**

The type of storage system that is providing storage space to a pool. This value is available only for SAN Volume Controller, Storwize V7000, and Storwize V7000 Unified pools.

If the back-end storage system was not probed, the value in this column is blank. To manually select a type of storage system to help calculate the approximate read I/O capability of the pool, right-click a pool in the list and select **View Properties**. On the Back-end Storage tab in properties notebook, click **Edit**.

Tip: The value in the Read I/O Capability column is not calculated until you select values for the other columns that are related to back-end storage and save your changes.

► **Back-End Storage RAID Level**

The RAID level of the volumes on the back-end storage system that are providing storage space to a pool. This value is available only for SAN Volume Controller, Storwize V7000, and Storwize V7000 Unified pools.

If the back-end storage system was not probed, the value in this column is blank. To manually define a RAID level to help calculate the approximate read I/O capability of the pool, right-click a pool in the list and select **View Properties**. On the Back-end Storage tab in properties notebook, click **Edit**.

Tip: The value in the Read I/O Capability column is not calculated until you select values for the other columns that are related to back-end storage and save your changes.

► **Back-End Storage Disk Type**

The class and speed of the physical disks that contribute to the volumes on the back-end storage system. This value is available only for SAN Volume Controller, Storwize V7000, and Storwize V7000 Unified pools.

If the back-end storage system was not probed, the value in this column is blank. To manually define a disk type to help calculate the approximate read I/O capability of the pool, right-click a pool in the list and select **View Properties**. On the Back-end Storage tab in properties notebook, click **Edit**.

Tip: The value in the Read I/O Capability column is not calculated until you select values for the other columns that are related to back-end storage and save your changes.

► **Back-End Storage Disks**

The number of physical disks that contribute to the volumes on the back-end storage system. This value is available only for SAN Volume Controller, Storwize V7000, and Storwize V7000 Unified pools.

If the back-end storage system was not probed, the value in this column is blank. To manually define the number of disks to help calculate the approximate read I/O capability of the pool, right-click a pool in the list and select **View Properties**. On the Back-end Storage tab in properties notebook, click **Edit**.

Tip: The value in the Read I/O Capability column is not calculated until you select values for the other columns that are related to back-end storage and save your changes.

► Read I/O Capability

The projected maximum number of I/O operations per second for a pool. This value is calculated based on the values in the following fields:

- Back-End Storage System Type
- Back-End Storage RAID Level
- Back-End Storage Disk Type
- Back-End Storage Disks fields

This value is available only for SAN Volume Controller, Storwize V7000, and Storwize V7000 Unified pools.

If the back-end storage system was not probed, the value in this column is blank. To help calculate an approximate read I/O capability for the pool, you must manually define values for the columns that are related to back-end storage.

► Capacity Pool

The name of the capacity pool to which the storage pool is assigned.

► Last Data Collection

The most recent date and time when data was collected about the storage system that contains a storage pool.

► Custom tag 1, 2, and 3

Any user-defined text that is associated with a pool. This text can be included as a report column when you generate reports for the pool.

You can tag a pool to satisfy custom requirements of a service class. A service class can specify up to three tags. To provide a service class, storage resources must have all the same tags that are specified by the service class. If a pool is not tagged, any tags on the containing storage system also apply to the pool for determining whether it satisfies the custom requirements of a service class.

To edit the custom tags, right-click it and select **View Properties**. On the properties notebook, click **Edit**.

To view detailed information about a specific storage pool, double-click the wanted storage pool in the Pools tab in the right pane. A new window opens that includes various tabs for displaying information about the pool, as shown in Figure 13-8 on page 370.

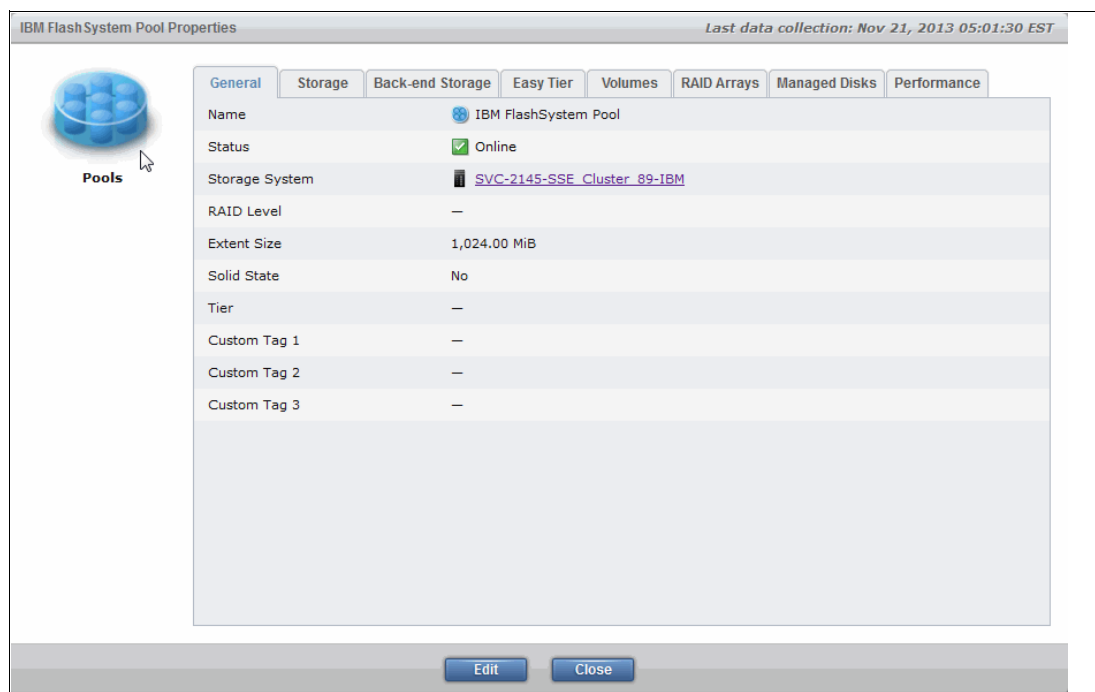


Figure 13-8 Viewing detailed pool information

Accessing the Managed Disks view

To access the Managed Disks view for a storage system, complete the following steps:

1. Click **Managed Disks** under the Internal Resources section, as shown in Figure 13-9.



Figure 13-9 Accessing the Managed Disks view for a storage system

The **Managed Disks** page for this storage system opens, which displays the managed disks for this system in a tabular format with various details, as shown in Figure 13-10.



Figure 13-10 The Managed Disks page for a storage system

2. Right-click any of the column headings to modify the columns that are displayed, as shown in Figure 13-11.

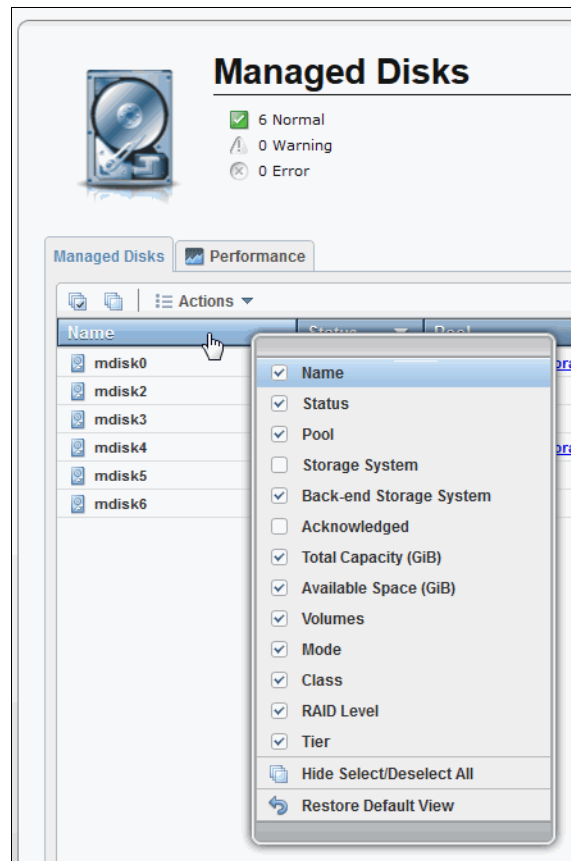


Figure 13-11 Modifying Managed Disks column headings

By using the Managed Disks window, you can view the following details about the managed disks:

- Name

The name of a managed disk that uniquely identifies it within a storage system.

- Status

The status of a managed disk. The following statuses are available:

- Online
- Offline
- Error
- Unknown

Use the status to determine the condition of a managed disk, and if any actions must be taken. For example, if a managed disk has an Offline status, its associated pool also is offline and you should take immediate action to correct the problem.

- Pool

The name of the storage pool (if any) to which the managed disk belongs.

- Storage System

The SAN Volume Controller or Storwize system that is managing this managed disk.

- Back-end Storage System
The name of the back-end storage system that is providing the storage for this managed disk.
- Acknowledged
Shows whether the status of a managed disk is acknowledged. An acknowledged status indicates that the status was reviewed and is resolved or can be ignored.
- Total Capacity (GB)
The total storage space of the managed disk.
- Available Capacity (GB)
The available capacity on the managed disk.
- Volumes
The number of volumes that might have extents on this managed disk.
- Mode
The mode of the managed disk. Possible values are Unmanaged, Managed, and Image.
- RAID Level
The RAID level of the managed disk.
- Tier
The managed disk tier, as defined in the SAN Volume Controller or Storwize system. Possible values are Hard Disk Drive and Solid-State Drive.

To view detailed information about a specific managed disk, double-click the wanted MDisk in the Managed Disks tab in the right pane. A new window opens that includes various tabs for displaying information about the MDisk, as shown in Figure 13-12.

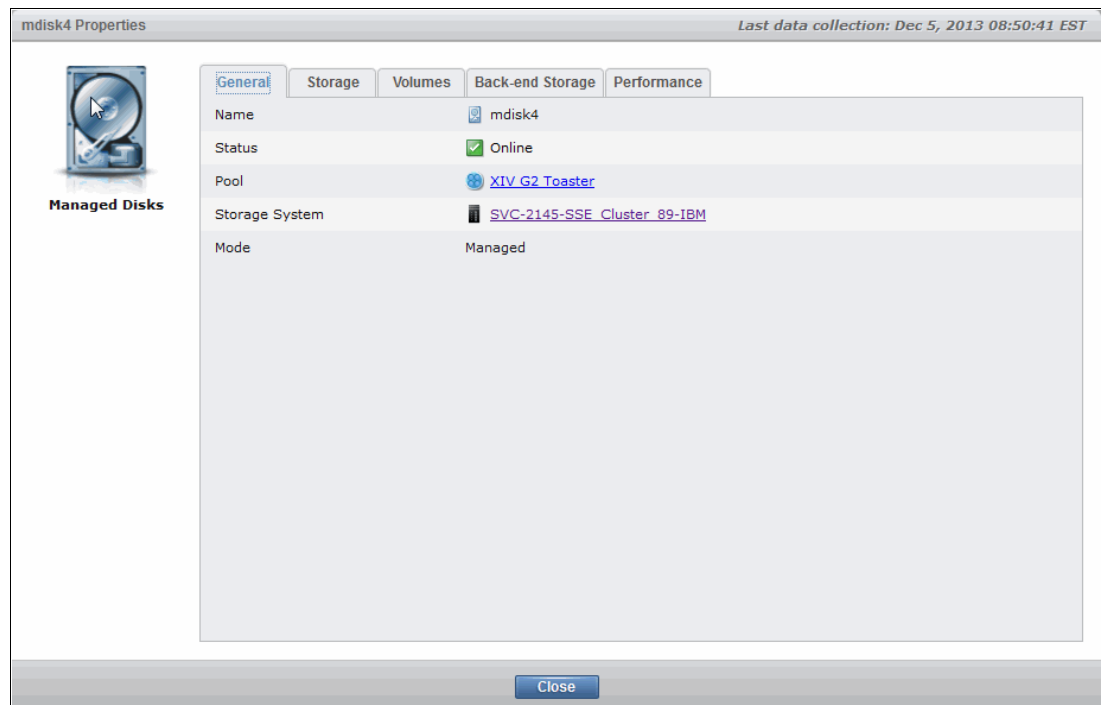


Figure 13-12 Viewing detailed managed disk information

Accessing the Volumes view

To access the volumes view, complete the following steps:

1. Click **Volumes** under the Internal Resources section, as shown in Figure 13-13.

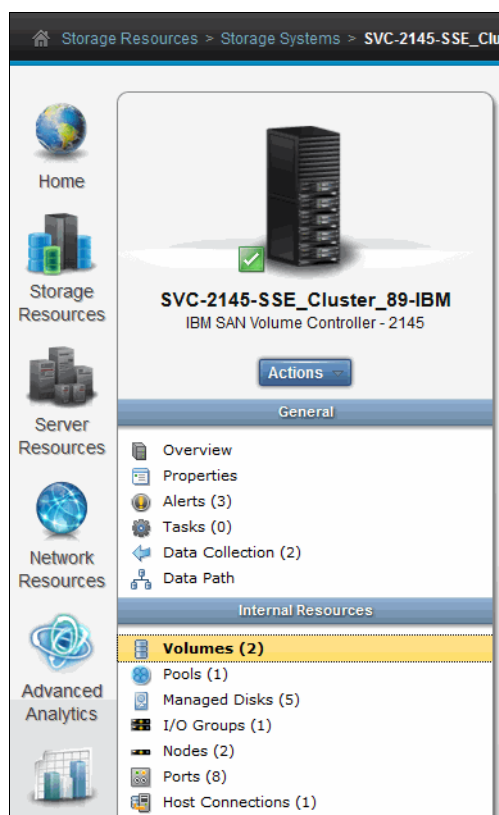


Figure 13-13 Accessing the Volumes view for a storage system

The **Volumes** window for this storage system opens, which displays the volumes for this system in a tabular format with various details, as shown in Figure 13-14.



Figure 13-14 The Volumes page for a storage system

2. Right-click any of the column headings to modify the columns that are displayed, as shown in Figure 13-15.

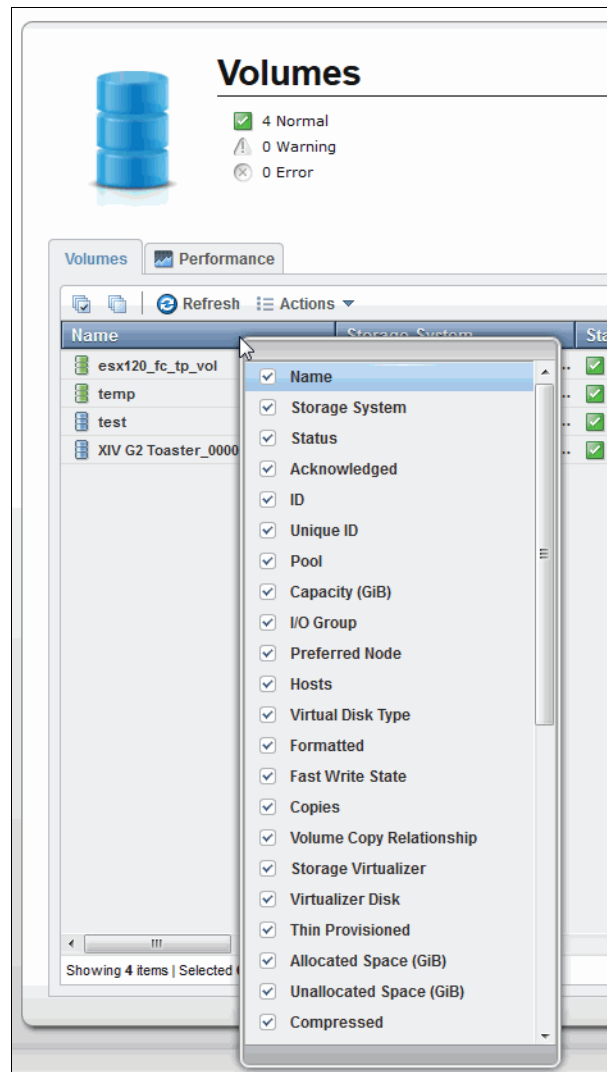


Figure 13-15 Modifying Volumes column headings

By using the Volumes window, you can view the following details about the volumes:

- Name

The name that was assigned to a volume when it was created.

- Storage System

The name of the storage system that contains a volume. This name was defined when the storage system was added to Tivoli Storage Productivity Center. If a name was not defined, the ID of the storage system is displayed.

- Status

The status of a volume. The following statuses are available:

- Normal
- Warning
- Error
- Unknown

Use the status to determine the condition of the volume, and if any actions must be taken. For example, if a volume has an Error status, take immediate action to correct the problem.

- Acknowledged

Shows whether the status of a volume is acknowledged. An acknowledged status indicates that the status was reviewed and is resolved or can be ignored. An acknowledged status is not used when the status of related, higher-level resources is determined.

For example, if the status of a volume is Error, the status of the related storage system also is Error. If the Error status of the volume is acknowledged, its status is not used to determine the overall status of the storage system. In this case, if the other internal resources of the storage system are Normal, the status of the storage system is also Normal.

- ID

The identifier for a volume, such as a serial number or internal ID.

- Unique ID

The ID that is used to uniquely identify a volume across multiple storage systems.

- Pool

The name of the storage pool in which a volume is a member.

- Capacity (GB)

The total amount of storage space that is committed to a volume. For thin-provisioned volumes, this value represents the virtual capacity of the volume. This value might also include overhead space if the pool is unformatted.

- I/O Group

The name of the I/O Group to which a volume is assigned.

- Preferred Node

The name of the preferred node within the I/O Group to which a volume is assigned.

- Hosts

The name of the host to which a volume is assigned. This name is the host name as defined on the storage system, and is not the name of the server that was discovered by a Storage Resource agent. If more than one host is assigned, the number of hosts is displayed. For storage systems that are managed by a CIM agent, the host name in this column might not match the configured host name on the storage system. Instead, the host name might be replaced by the WWPN of the host port or text that is generated by the CIM agent.

- Virtual Disk Type

The type of virtual disk with which a volume was created, such as sequential, striped, or image.

- Formatted

Shows whether a volume is formatted.

- Fast Write State

Shows the cache state for a volume, such as empty, not empty, corrupt, and repairing. The corrupt state indicates that you must recover the volume by using one of the **recovervdisk** commands for the storage system. The repairing state indicates that repairs that were started by a **recovervdisk** command are in progress.

– Copies

The number of secondary copies (virtual disk copies) for a volume. The primary copy of a virtual disk is not counted as a mirror.

– Volume Copy Relationship

Shows whether a volume is in a replication relationship that creates a snapshot or point-in-time copy of the volume on a specified target volume. A volume can be a source, target, or both a target for one copy pair and a source for a different copy pair. In storage systems, this relationship might be referred to as a FlashCopy, snapshot, or point-in-time copy relationship. A volume can have one of the following properties:

- Source: The volume is the source of the relationship.
- Target: The volume is the target of the relationship.
- Source and Target: The volume is a target for one copy pair and a source for a different copy pair.
- None: The volume is not part of any volume copy relationship.

– Storage Virtualizer

The name of the storage virtualizer that is managing a volume. A storage virtualizer is a storage system that virtualizes storage space from internal storage or from another storage system. Examples of storage virtualizers include SAN Volume Controller and Storwize V7000. A value is displayed in this column only if the volume is managed by a storage virtualizer and Tivoli Storage Productivity Center collected data about that virtualizer.

– Virtualizer Disk

The managed disk for the virtualizer that corresponds to a volume.

– Thin Provisioned

Shows whether a volume is a thin-provisioned volume and the type of thin-provisioning that is used for the volume. A thin-provisioned volume is a volume with a virtual capacity that is different from its real capacity. Not all the storage capacity of the volume is allocated when the volume is created, but is allocated over time, as needed.

– Allocated Space (GB)

The amount of space that is reserved for a volume. The space that is allocated for a thin-provisioned volume is less than its virtual capacity. The value for Allocated Space is equal to Used Space for SAN Volume Controller and Storwize V7000 storage systems that are not thin-provisioned

– Unallocated Space (GB)

The amount of space that is not reserved for a thin-provisioned volume. This value is determined by using the following formula:

Capacity - Allocated Space

– Compressed

Shows whether a storage volume is compressed.

– Compression Savings (%)

The percentage of capacity saving in a volume by using the compression feature.

– Physical Allocation (%)

The percentage of physical space that is reserved for a volume. This value cannot be greater than 100% because it is not possible to reserve more physical space than is available.

This value is determined by using the following formula:

$$\text{Allocated Space} / \text{Capacity} * 100$$

For example, if the space that is reserved for volumes is 50 GB for a volume size of 200 GB, physical allocation is 25%.

The first section of the bar in the Pools window uses the color blue and a percent (%) sign to represent the physical allocation percentage. The second section of the bar uses the color gray to represent the volume capacity.

- Shortfall (%)

The percentage of the remaining unallocated volume space in a pool that is not available to be allocated to a volume. The higher the percentage, the more critical the shortfall of space. A warning icon is shown if this percentage is greater than 100%. This information is available only for thin-provisioned volumes. This value is determined by using the following formula:

$$\text{Unallocatable Volume Space} / \text{Unallocated Volume Space} * 100$$

- The first section of the bar uses the color blue and a percent (%) sign to represent the shortfall percentage. The second section of the bar uses the color gray to represent the unallocated volume space.

- Used Allocated Space (%)

The percentage of reserved space for a volume that is being used. This value cannot be greater than 100% because a volume cannot use more space than is allocated. This value is determined by using the following formula:

$$\text{Used Space} / \text{Allocated Space} * 100$$

- Used Space (GB)

The amount of allocated space that is used by a volume. For SAN Volume Controller and Storwize V7000, you can pre-allocate thin-provisioned volume space when the volumes are created. In these cases, the Used Space might be different from the Allocated Space. For compressed volumes on SAN Volume Controller and Storwize V7000, the Used Space reflects the size of compressed data that is written to disk. As the data changes, the Used Space might at times be less than the Allocated Space.

For volumes that are not thin provisioned in SAN Volume Controller, Storwize V7000, and other storage systems, the values for Used Space and Allocated Space are equal. This value is accurate as of the most recent time that data was collected about a volume. Data collection is run on a set schedule and the used space on a volume can change rapidly. Therefore, the value in this column might not reflect the current state of a volume.

- Unused Space (GB)

The amount of space that is allocated to a volume and is not yet used. This value is determined by using the following formula:

$$\text{Allocated Space} - \text{Used Space}$$

- Grain Size (KB)

The grain size with which a thin-provisioned volume was created. This value is typically 32, 64, 128, or 256 KB. Larger grain sizes maximize performance, whereas smaller grain sizes maximize space efficiency. Grain sizes also limit the maximum virtual space of the volume.

– Warning Level (%)

The warning level that was defined when a thin-provisioned volume was created. This value is measured in MB (10^{20} bytes) or a percentage of the total, virtual capacity of the volume. The storage system generates a warning if the used capacity of a volume grows enough to exceed the specified threshold.

– Auto Expand

Shows whether a thin-provisioned volume automatically expands its allocated capacity as more of its space is used.

– Easy Tier

Shows how the Easy Tier function is enabled on a volume. The following values are possible:

- Enabled/Inactive
- Enabled/Active
- Auto/Active
- Auto/Inactive
- Disabled
- Unknown

– Tier Capacity SSD (GB)

The total storage space on the solid-state drives that are participating in, or are eligible to participate in, the Easy Tier optimization for a volume.

– Tier Capacity HDD (GB)

The total storage space on the hard disk drives that are participating in, or are eligible to participate in, the Easy Tier optimization for a volume.

– Last Data Collection

The most recent date and time when data was collected about the storage system that contains a volume.

– Service Class

The name of the block storage service class that is associated with the volume. A block storage service class typically represents a particular quality of service.

– Original Capacity Pool

The name of the capacity pool from which a volume was provisioned. The storage resource on which the volume was created might no longer be a member of the capacity pool.

– Ticket

A ticket identifier that was associated with the volume for tracking purposes. The ticket identifier was specified when the volume was created.

To view detailed information about a volume, double-click the wanted volume in the Volumes tab in the right pane. A new window opens that includes various tabs for displaying information about the volume, as shown in Figure 13-16 on page 380.

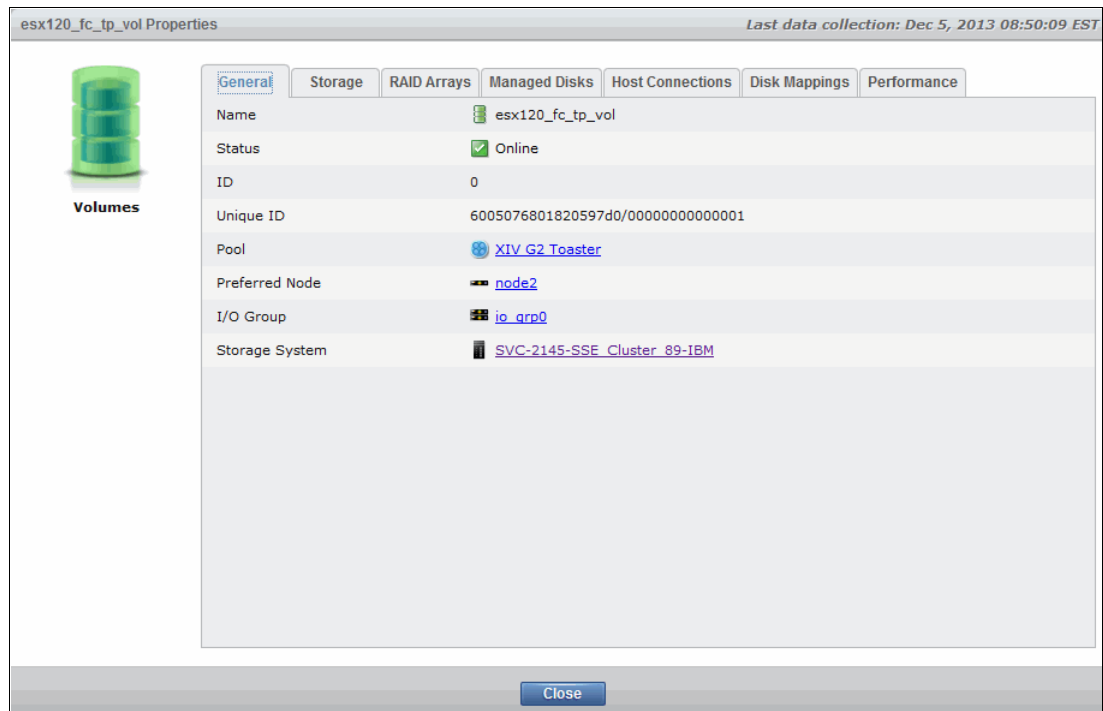


Figure 13-16 Viewing detailed volume information

13.2 Considerations for performance analysis

When you start to analyze the performance of your environment to identify a performance problem, you identify all of the components and then verify the performance of these components. This section describes the considerations for a SAN Volume Controller environment and for a Storwize V7000 environment.

13.2.1 SAN Volume Controller considerations

For the SAN Volume Controller environment, you identify all of the components between the two systems, and then you verify the performance of the smaller components.

SAN Volume Controller traffic

Traffic between a host, the SAN Volume Controller nodes, and a storage controller adheres to the following path:

1. The host generates the I/O and transmits it on the fabric.
2. The I/O is received on the SAN Volume Controller node ports.
3. If the I/O is a write I/O:
 - a. The SAN Volume Controller node writes the I/O to the SAN Volume Controller node cache.
 - b. The SAN Volume Controller node sends a copy to its partner node to write to the cache of the partner node.
 - c. If the I/O is part of a Metro Mirror or Global Mirror, a copy must go to the secondary virtual disk (VDisk) of the relationship.

- d. If the I/O is part of a FlashCopy and the FlashCopy block was not copied to the target VDisk, the action must be scheduled.
4. If the I/O is a read I/O:
 - a. The SAN Volume Controller must check the cache to see whether the Read I/O is already there.
 - b. If the I/O is not in the cache, the SAN Volume Controller must read the data from the physical LUNs (managed disks).
5. At some point, write I/Os are sent to the storage controller.
6. To reduce latency on subsequent read commands, the SAN Volume Controller might also perform read-ahead I/Os to load the cache.

SAN Volume Controller performance guidance

You must have at least two managed disk groups: one for key applications and another for everything else. You might want more managed disk groups if you want to separate different device types, such as RAID 5 versus RAID 10 or SAS versus nearline SAS (NL-SAS).

For SAN Volume Controller, follow these development guidelines for IBM System Storage DS8000:

- ▶ One MDisk per extent pool
- ▶ One MDisk per storage cluster
- ▶ One managed disk group per storage subsystem
- ▶ One managed disk group per RAID array type (RAID 5 versus RAID 10)
- ▶ One MDisk and managed disk group per disk type (10K versus 15 K RPM or 146 GB versus 300 GB)

In some situations, such as the following examples, you might want multiple managed disk groups:

- ▶ Workload isolation
- ▶ Short-stroking a production managed disk group
- ▶ Managing different workloads in different groups

13.2.2 Storwize V7000 considerations

In a Storwize V7000 environment, identify all of the components between the Storwize V7000, the server, and the back-end storage subsystem if they are configured in that manner. Alternatively, identify the components between Storwize V7000 and the server. Then, verify the performance of all of components.

Storwize V7000 traffic

Traffic between a host, the Storwize V7000 nodes, direct-attached storage, or a back-end storage controller traverses the same storage path:

1. The host generates the I/O and transmits it on the fabric.
2. The I/O is received on the Storwize V7000 canister ports.
3. If the I/O is a write I/O:
 - a. The Storwize V7000 node canister writes the I/O to its cache.
 - b. The preferred canister sends a copy to its partner canister to update the partner canister's cache.
 - c. If the I/O is part of a Metro or Global Mirror, a copy must go to the secondary volume of the relationship.

- d. If the I/O is part of a FlashCopy and the FlashCopy block was not copied to the target volume, this action must be scheduled.
- 4. If the I/O is a read I/O:
 - a. The Storwize V7000 must check the cache to see whether the Read I/O is already in the cache.
 - b. If the I/O is not in the cache, the Storwize V7000 must read the data from the physical MDisks.
- 5. At some point, write I/Os are destaged to Storwize V7000 MDisks or sent to the back-end SAN-attached storage controllers.
- 6. The Storwize V7000 might also do some data optimized, sequential detect pre-fetch cache I/Os to preinstall the cache if the next read I/O was determined by the Storwize V7000 cache algorithms. This approach benefits the sequential I/O when compared with the more common least recently used (LRU) method that is used for nonsequential I/O.

Storwize V7000 performance guidance

You must have at least two storage pools for internal MDisks and two for external MDisks from external storage subsystems. Whether built from internal or external MDisks, each storage pool provides the basis for a general-purpose class of storage or for a higher performance or high availability class of storage.

You might want more storage pools if you want to separate different device types, such as RAID 5 versus RAID 10, or SAS versus NL-SAS.

For Storwize V7000, adhere to the following development guidelines:

- ▶ One managed disk group per storage subsystem
- ▶ One managed disk group per RAID array type (RAID 5 versus RAID 10)
- ▶ One MDisk and managed disk group per disk type (10K versus 15 K RPM, or 146 GB versus 300 GB)

In some situations, such as the following examples, you might want to use multiple managed disk groups:

- ▶ Workload isolation
- ▶ Short-stroking a production managed disk group
- ▶ Managing different workloads in different groups

13.3 Top 10 reports for SAN Volume Controller and Storwize V7000

The top 10 reports from Tivoli Storage Productivity Center are a common request. This section summarizes which reports to create, and in which sequence to begin your performance analysis for a SAN Volume Controller or Storwize V7000 virtualized storage environment.

Use the following top 10 reports and in the order that is shown in Figure 13-17 on page 383:

- ▶ Report 1: I/O Group Performance
- ▶ Report 2: Module/Node Cache Performance report
- ▶ Reports 3 and 4: Managed Disk Group Performance
- ▶ Report 5: Top Active Volumes Cache Hit Performance
- ▶ Report 6: Top Volumes Data Rate Performance

- ▶ Report 7: Top Volumes Disk Performance
- ▶ Report 8: Top Volumes I/O Rate Performance
- ▶ Report 9: Top Volumes Response Performance
- ▶ Report 10: Port Performance

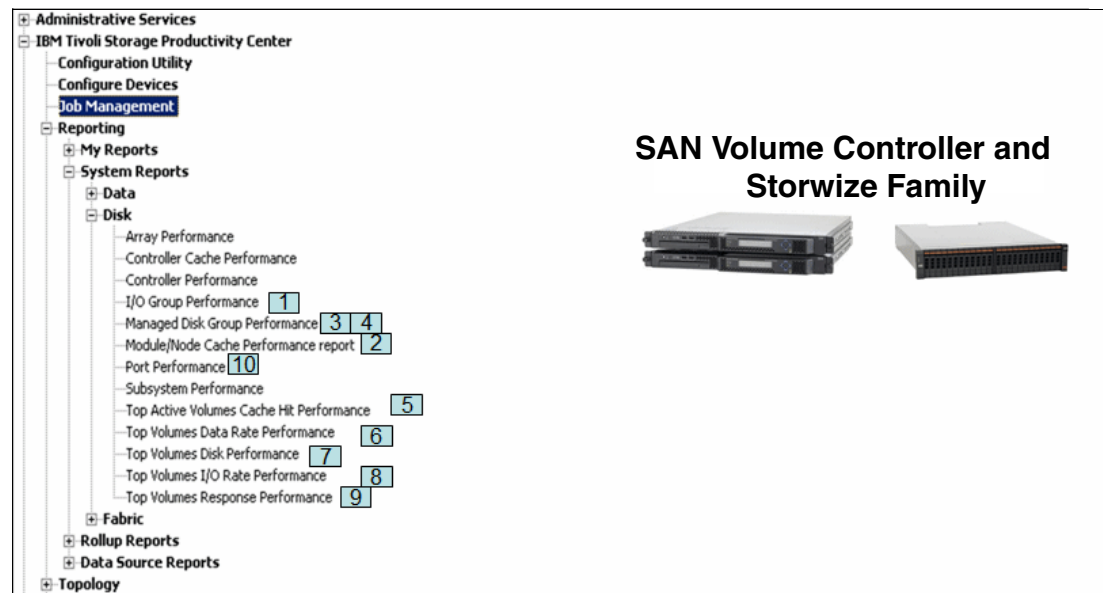


Figure 13-17 Sequence for running the top 10 reports

In other cases, such as performance analysis for a particular server, you follow another sequence, starting with Managed Disk Group Performance. By using this approach, you can quickly identify MDisk and VDisks that belong to the server that you are analyzing.

To view system reports that are relevant to SAN Volume Controller and Storwize V7000, click **IBM Tivoli Storage Productivity Center** $\varnothing$ **Reporting** $\varnothing$ **System Reports** $\varnothing$ **Disk**.

I/O Group Performance and Managed Disk Group Performance are specific reports for SAN Volume Controller and Storwize V7000. Module/Node Cache Performance is also available for IBM XIV. These reports are shown in Figure 13-18.

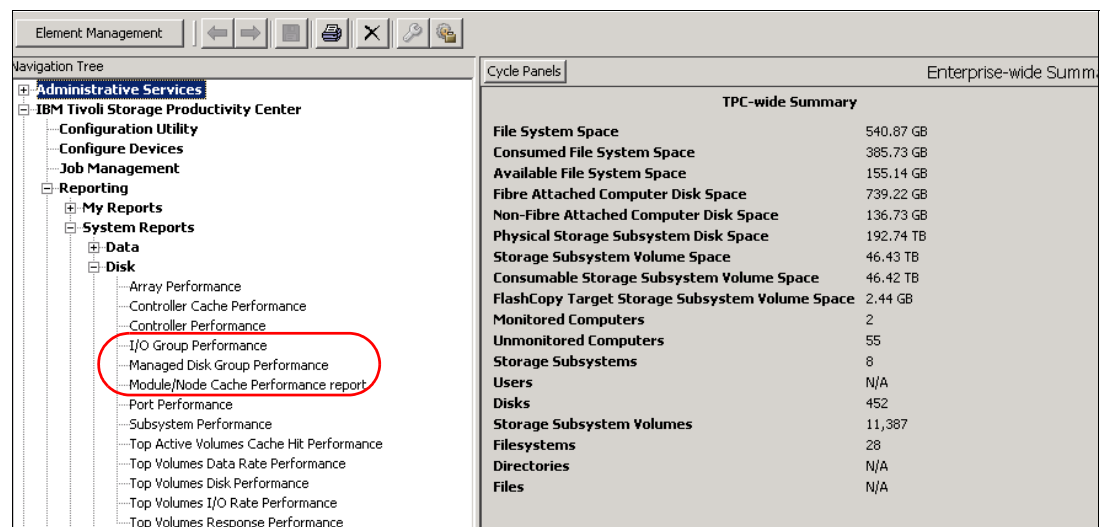


Figure 13-18 System reports for SAN Volume Controller and Storwize V7000

Figure 13-19 shows a sample structure to review basic SAN Volume Controller concepts about SAN Volume Controller structure and then to proceed with performance analysis at the component levels.

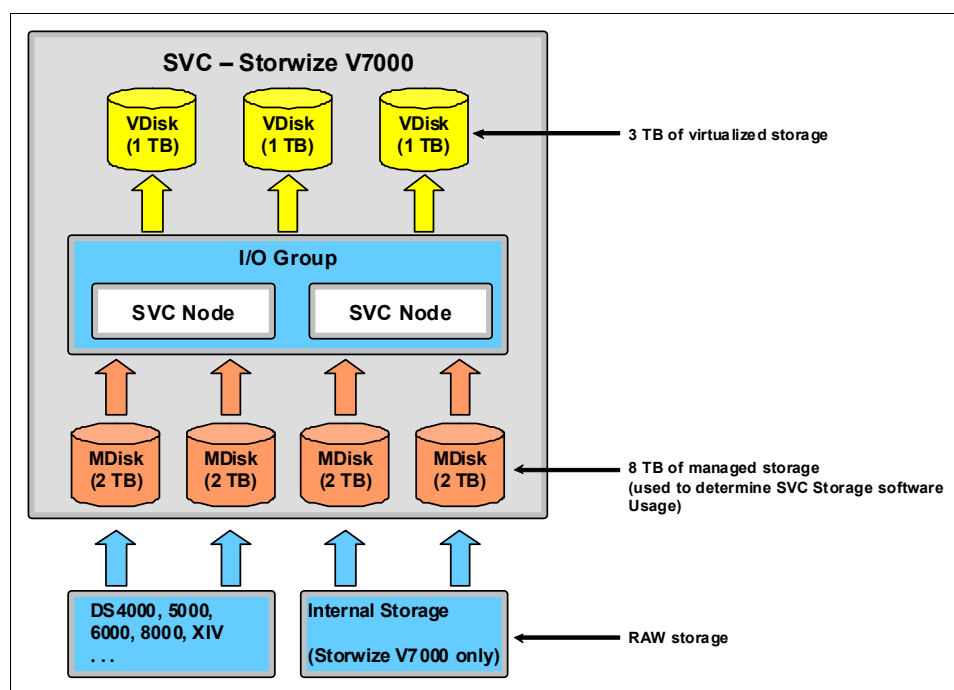


Figure 13-19 SAN Volume Controller and Storwize V7000 sample structure

Note: By using Tivoli Storage Productivity Center version 5.2, you can generate some of the reports that were identified in this section in the stand-alone GUI or the web-based GUI. Where applicable, examples are provided by using both methods.

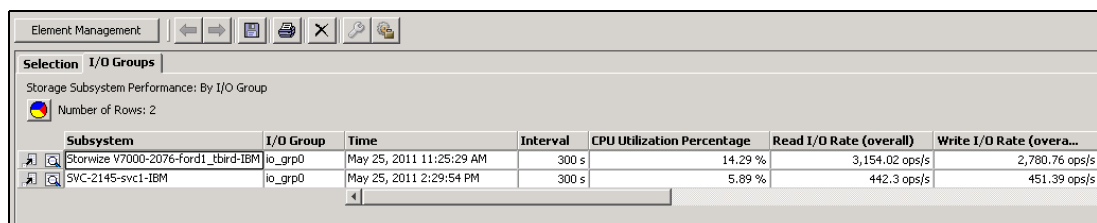
13.3.1 I/O Group Performance for SAN Volume Controller and Storwize V7000

In this section, we describe the I/O group performance monitoring for SAN Volume Controller and Storwize V7000.

I/O Group Performance by using the stand-alone GUI

Tip: For SAN Volume Controllers with multiple I/O groups, a separate row is generated for every I/O group within each SAN Volume Controller.

In our lab environment, data was collected for a SAN Volume Controller with a single I/O group. In Figure 13-20 on page 385, the scroll bar at the bottom of the table indicates that you can view more metrics.

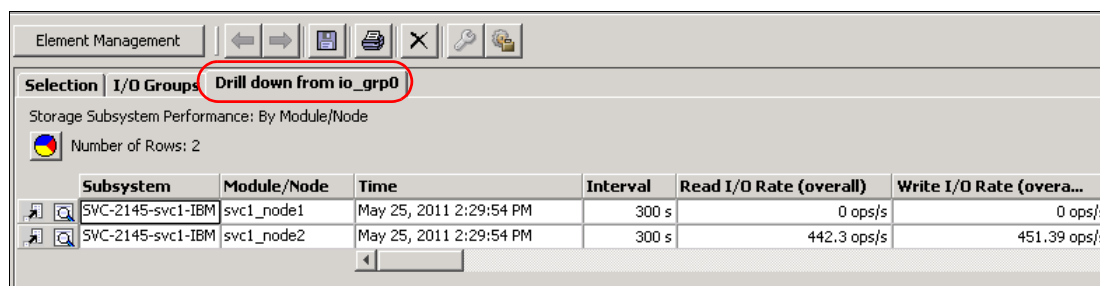


Subsystem	I/O Group	Time	Interval	CPU Utilization Percentage	Read I/O Rate (overall)	Write I/O Rate (overall)
Storwize V7000-2076-ford1_tbird-IBM	io_grp0	May 25, 2011 11:25:29 AM	300 s	14.29 %	3,154.02 ops/s	2,780.76 ops/s
SVC-2145-svc1-IBM	io_grp0	May 25, 2011 2:29:54 PM	300 s	5.89 %	442.3 ops/s	451.39 ops/s

Figure 13-20 I/O group performance

Important: The data that is displayed in a performance report is the last collected value at the time the report is generated. It is not an average of the last hours or days, but it shows the last data collected.

Click the magnifying glass icon (🔍) that is next to SAN Volume Controller io_grp0 entry to drill down and view the statistics by nodes within the selected I/O group. The drill down from io_grp0 tab is created, as shown in Figure 13-21. This tab contains the report for nodes within the SAN Volume Controller.



Subsystem	Module/Node	Time	Interval	Read I/O Rate (overall)	Write I/O Rate (overall)
SVC-2145-svc1-IBM	svc1_node1	May 25, 2011 2:29:54 PM	300 s	0 ops/s	0 ops/s
SVC-2145-svc1-IBM	svc1_node2	May 25, 2011 2:29:54 PM	300 s	442.3 ops/s	451.39 ops/s

Figure 13-21 Drill down from io_grp0 tab

To view a historical chart of one or more specific metrics for the resources, click the **pie chart** icon (📊). A list of metrics is displayed, as shown in Figure 13-22 on page 386. You can select one or more metrics that use the same measurement unit. If you select metrics that use different measurement units, you receive an error message.

CPU Utilization Percentage metric

The CPU Utilization reports indicate the degree to which the cluster nodes are used. To generate a graph of CPU utilization by node, select the **CPU Utilization Percentage** metric and then click **OK**, as shown in Figure 13-22.

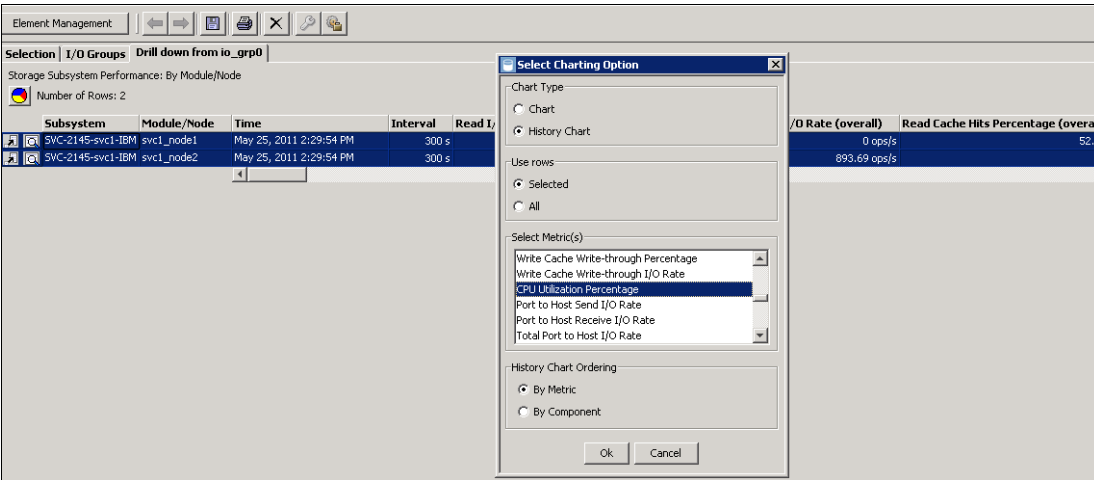


Figure 13-22 CPU utilization selection for SAN Volume Controller

You can change the reporting time range and click **Generate Chart** to regenerate the graph, as shown in Figure 13-23. A continual high Node CPU Utilization rate indicates a busy I/O group. In our environment, CPU utilization does not rise above 24%, which is a more than acceptable value.

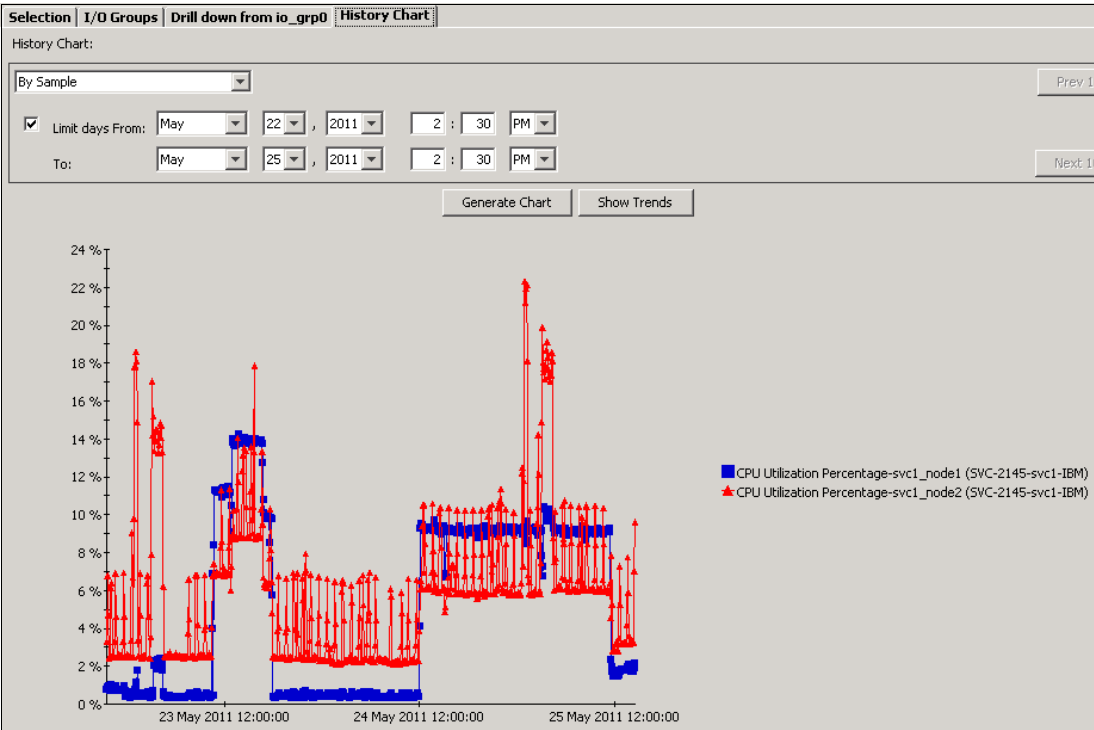


Figure 13-23 CPU utilization graph for SAN Volume Controller

CPU utilization guidelines for SAN Volume Controller only

If the CPU utilization for the SAN Volume Controller node remains constantly above 70%, it might be time to increase the number of I/O groups in the cluster. You can also redistribute workload to other I/O groups in the SAN Volume Controller cluster if it is available. You can add cluster I/O groups up to the maximum of four I/O groups per SAN Volume Controller cluster.

If four I/O groups are in a cluster (with the latest firmware installed) and you are still having high SAN Volume Controller node CPU utilization as indicated in the reports, build a cluster. Consider migrating some storage to the new cluster, or if existing SAN Volume Controller nodes are not of the 2145-CG8 version, upgrade them to the CG8 nodes.

Total I/O Rate (overall)

To view the overall total I/O rate (as shown in Figure 13-24), complete the following steps:

1. In the drill down from io_grp0 tab, which returns you to the performance statistics for the nodes in the SAN Volume Controller, click the pie chart icon (📊).
2. In the Select Charting Option window, select the **Total I/O Rate (overall)** metric, and then click **OK**.

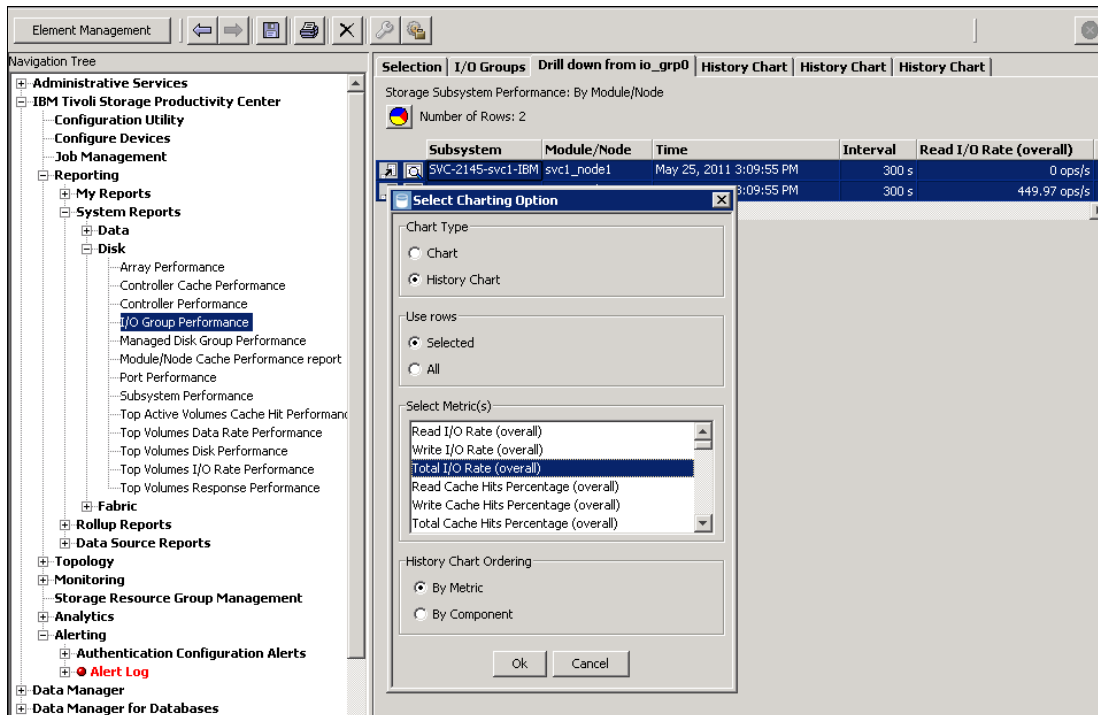


Figure 13-24 I/O rate

The I/Os are present only on Node 2. Therefore, in Figure 13-25 on page 388, you can see a configuration problem where the workload is not well-balanced, at least during this time frame.

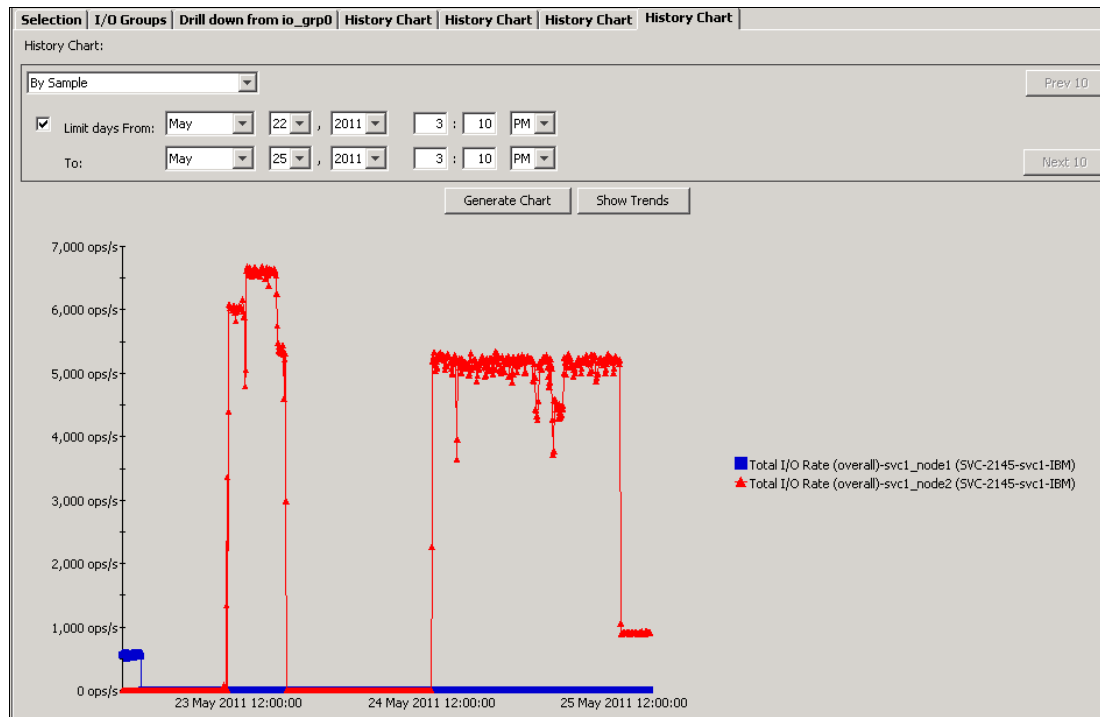


Figure 13-25 I/O rate graph

In the I/O rate graph (as shown in Figure 13-25), you can see a configuration problem.

Backend Response Time

To view the read and write response time at the node level, complete the following steps:

1. In the drill down from io_grp0 tab, which returns you to the performance statistics for the nodes within the SAN Volume Controller, click the pie chart icon (📊).
2. In the Select Charting Option window (as shown in Figure 13-26 on page 389), select the **Backend Read Response Time** and **Backend Write Response Time** metrics. Then, click **OK** to generate the report.

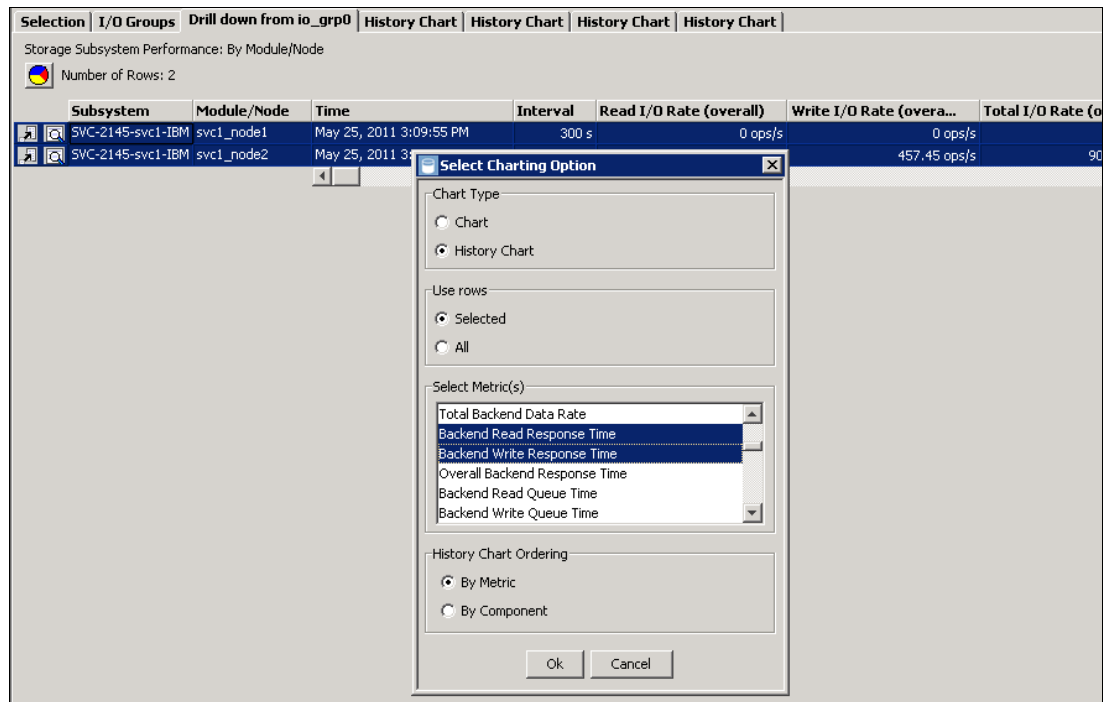


Figure 13-26 Response time selection for the SAN Volume Controller node

Figure 13-27 shows the report. The values are shown that might be accepted in the back-end response time for read and write operations. These values are consistent for both I/O groups.

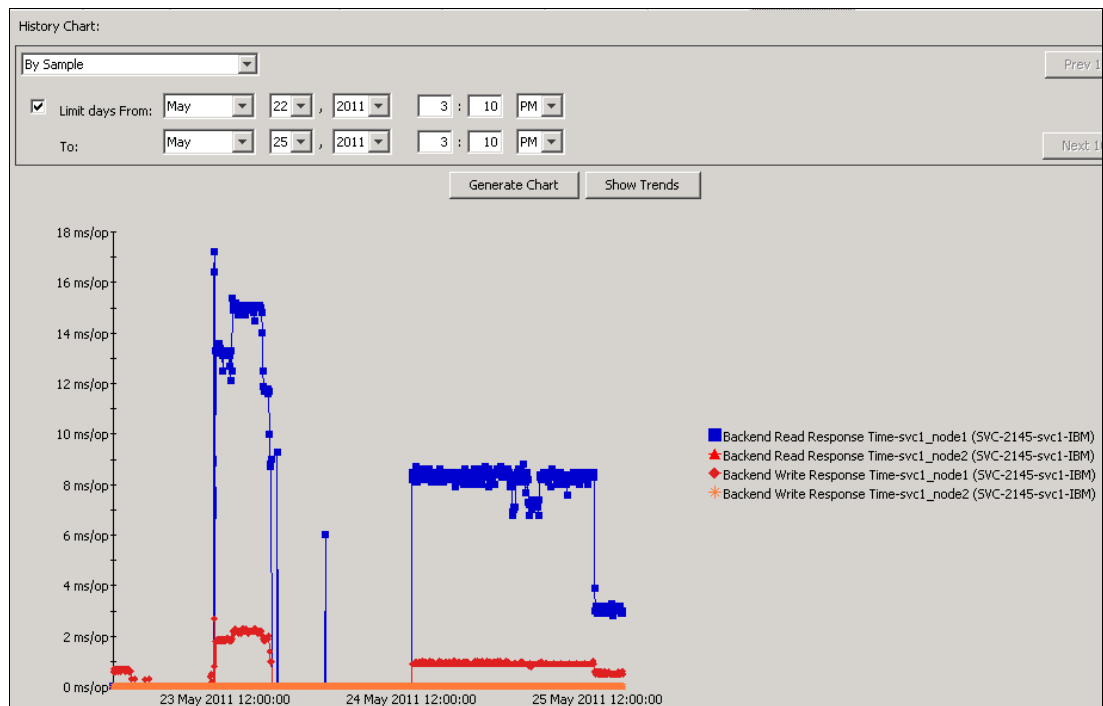


Figure 13-27 Response Time report for the SAN Volume Controller node

Guidelines for poor response times

For random read I/O, the back-end rank (disk) read response times should seldom exceed 25 ms, unless the read hit ratio is near 99%. Backend Write Response Times are higher because of RAID 5 (or RAID 10) algorithms, but should seldom exceed 80 ms. Some time intervals might exist when response times exceed these guidelines.

If you are experiencing poor response times, use all available information from the SAN Volume Controller and the back-end storage controller to investigate the times. The following possible causes for a significant change in response times from the back-end storage might be visible by using the storage controller management tool:

- ▶ A physical array drive failure that leads to an array rebuild. This failure drives more internal read/write workload of the back-end storage subsystem when the rebuild is in progress. If this situation causes poor latency, you might want to adjust the array rebuild priority to reduce the load. However, the array rebuild priority must be balanced with the increased risk of a second drive failure during the rebuild, which might cause data loss in a RAID 5 array.
- ▶ Cache battery failure that leads to the controller disabling the cache. You can usually resolve this situation by replacing the failed battery.

For more information about rules of thumb and how to interpret the values, see *SAN Storage Performance Management Using Tivoli Storage Productivity Center*, SG24-7364.

Data Rate

To view the Read Data rate, complete the following steps:

1. In the drill down from `io_grp0` tab, which returns you to the performance statistics for the nodes within the SAN Volume Controller, click the **pie chart** icon (📊).
2. Select the **Read Data Rate** metric. Press Shift and select **Write Data Rate** and **Total Data Rate**. Then, click **OK** to generate the chart, as shown in Figure 13-28.

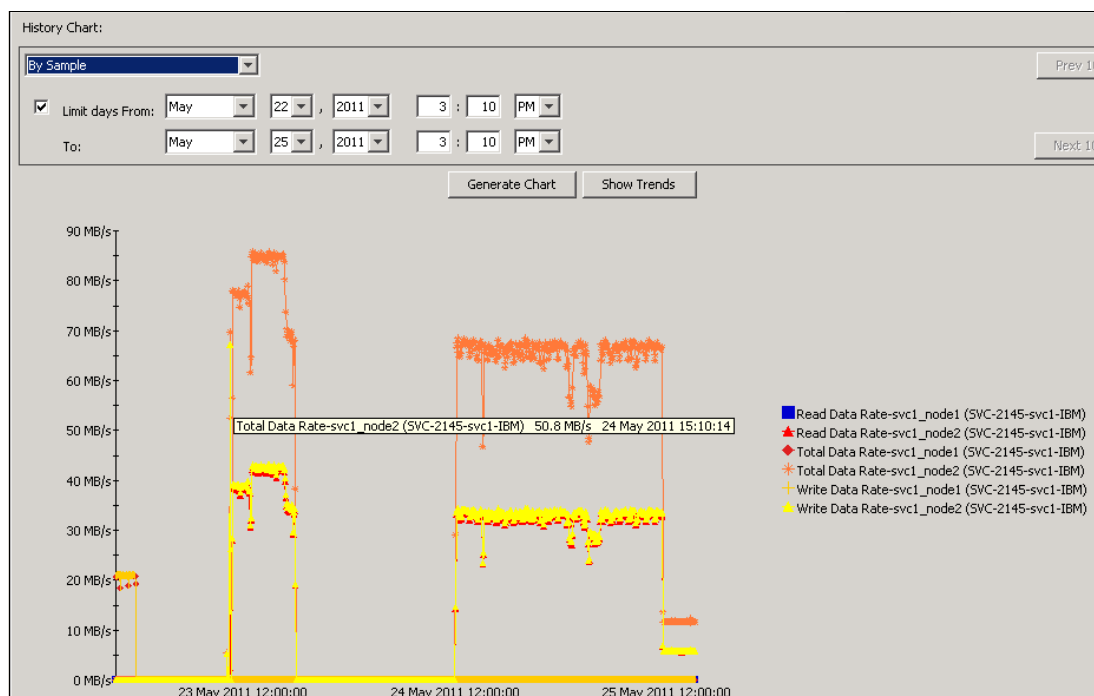


Figure 13-28 Data Rate graph for SAN Volume Controller

To interpret your performance results, always return to your baseline. For more information about creating a baseline, see *SAN Storage Performance Management Using Tivoli Storage Productivity Center*, SG24-7364.

The throughput benchmark (which is 7,084.44 SPC-2 MBPS) is the industry-leading throughput benchmark. For more information about this benchmark, see *SPC Benchmark 2 Executive Summary IBM System Storage SAN Volume Controller SPC-2 V1.2.1*, which is available at this website:

http://www.storageperformance.org/results/b00024_IBM-SVC4.2_SPC2_executive-summary.pdf

Accessing I/O Group Performance by using the web-based GUI

To view the performance of a SAN Volume Controller I/O Group, start by clicking **Volumes** under the Internal Resources section, as shown in Figure 13-29.



Figure 13-29 Accessing the I/O Groups view for a storage system

The I/O Groups window opens for this storage system, which displays the I/O groups for this system in a tabular format, with various details, as shown in Figure 13-30 on page 392. In our lab environment, data was collected for a SAN Volume Controller with a single I/O group.

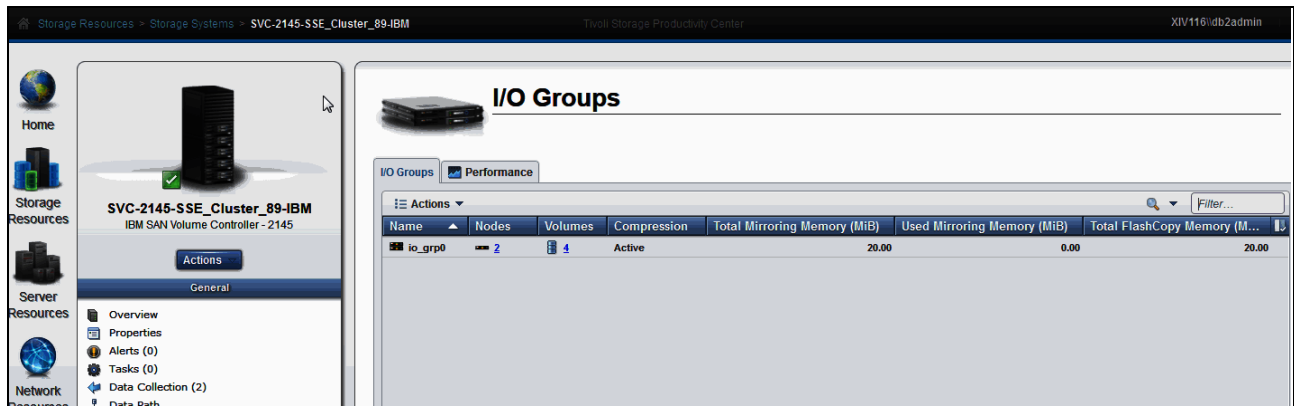


Figure 13-30 The I/O Groups page for a storage system

Tip: For a SAN Volume Controller with multiple I/O groups, a separate row is generated for every I/O group within that SAN Volume Controller.

To quickly view the performance of all the I/O Groups in this storage system, click the **Performance** tab. Alternatively, to view the performance of a single I/O Group, right-click anywhere in the corresponding row in the I/O Groups tab, as shown in Figure 13-31.

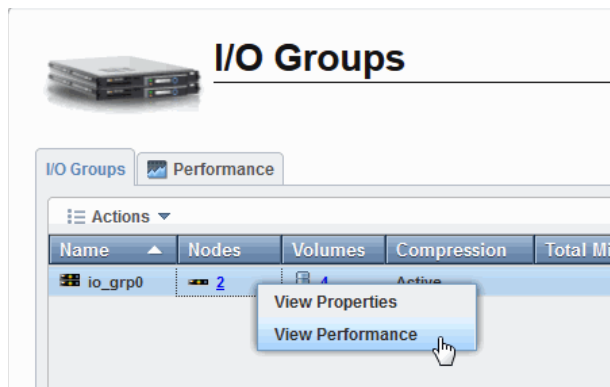


Figure 13-31 Accessing the performance view of an I/O group

A window opens that displays the performance of that I/O Group, as shown in Figure 13-32.

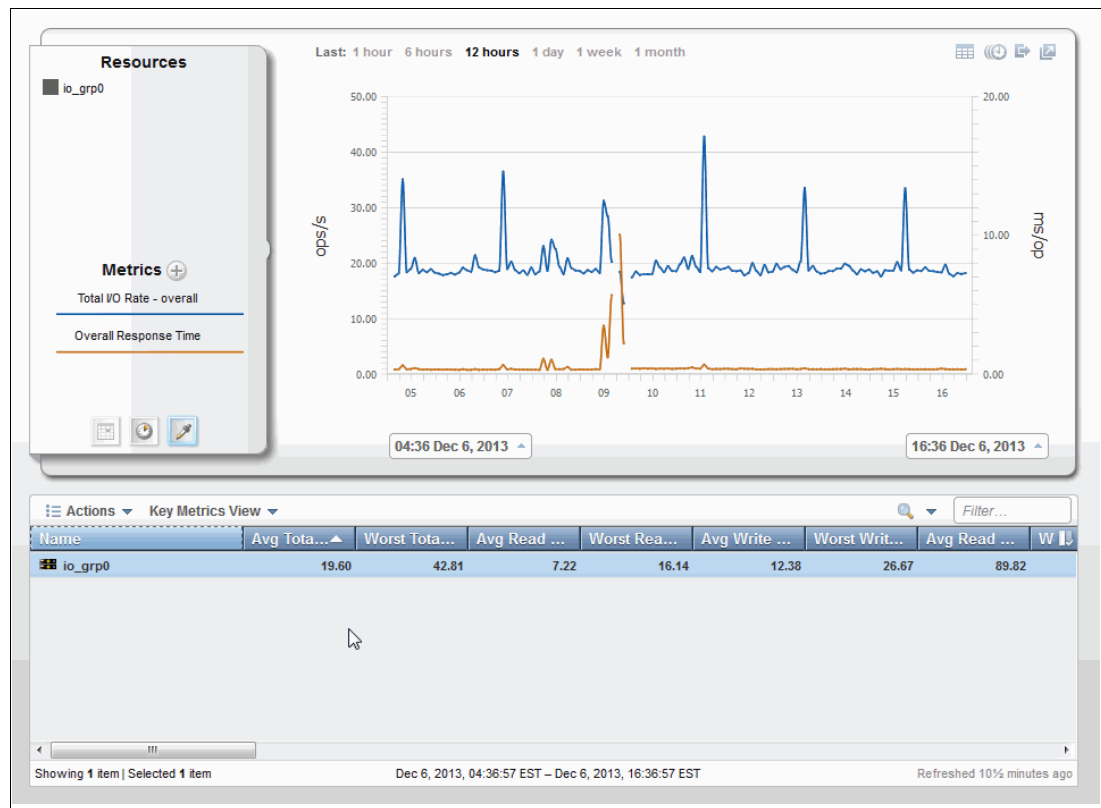


Figure 13-32 Viewing the performance of an I/O group

The default view in the I/O Group performance page shows a graph in the top, which depicts certain key metric values over time. It also includes a table at the bottom that contains averages of the data that is used to display the graph.

The performance graph can be easily customized to suit your specific needs in the following ways:

- The default time window that is displayed is for the last 12 hours. To display metrics for a different time frame, click one of the predefined time frame links at the top of the window, or enter specific start and end dates and times by using the drop-down fields at the bottom of the chart.
- To display the performance metrics for a related entity, such as the volumes of the I/O Group or the storage system to which the I/O group belongs, right-click anywhere in the row for the I/O Group and click a selection, as shown in Figure 13-33 on page 394. A browser window opens that displays the corresponding performance.

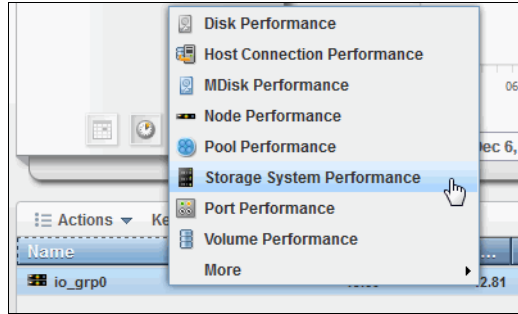


Figure 13-33 Accessing related performance metrics

- To display different metrics in the chart, click + next to the Metrics heading in the legend area to the left of the chart, as shown in Figure 13-34.

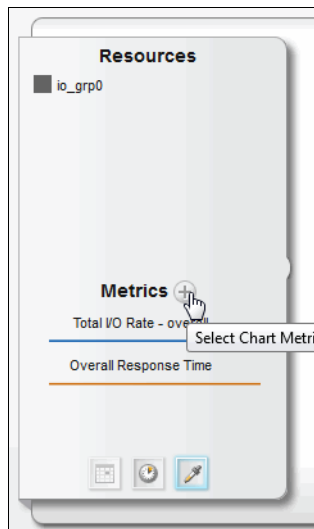


Figure 13-34 Modifying the performance metrics displayed

A window opens in which you can choose the metrics to display, as shown in Figure 13-35.

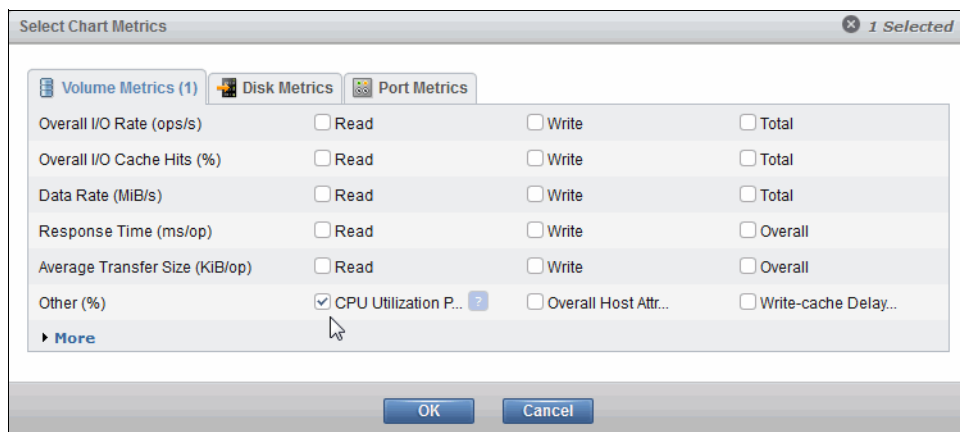


Figure 13-35 Performance metric selection

Total I/O Rate (overall)

The default performance view for an I/O group displays total I/O rate (overall) and overall response time, as shown in Figure 13-32 on page 393. This provides you with a total I/O rate and average response time (read and writes) for the two nodes that make up the I/O group.

To drill down into the nodes that make up the I/O group, right-click anywhere in the row that corresponds to the I/O group and select **Node Performance**, as shown in Figure 13-33 on page 394. This opens a window that displays the total I/O rate and CPU utilization for each of the nodes in the I/O group, as shown in Figure 13-36.

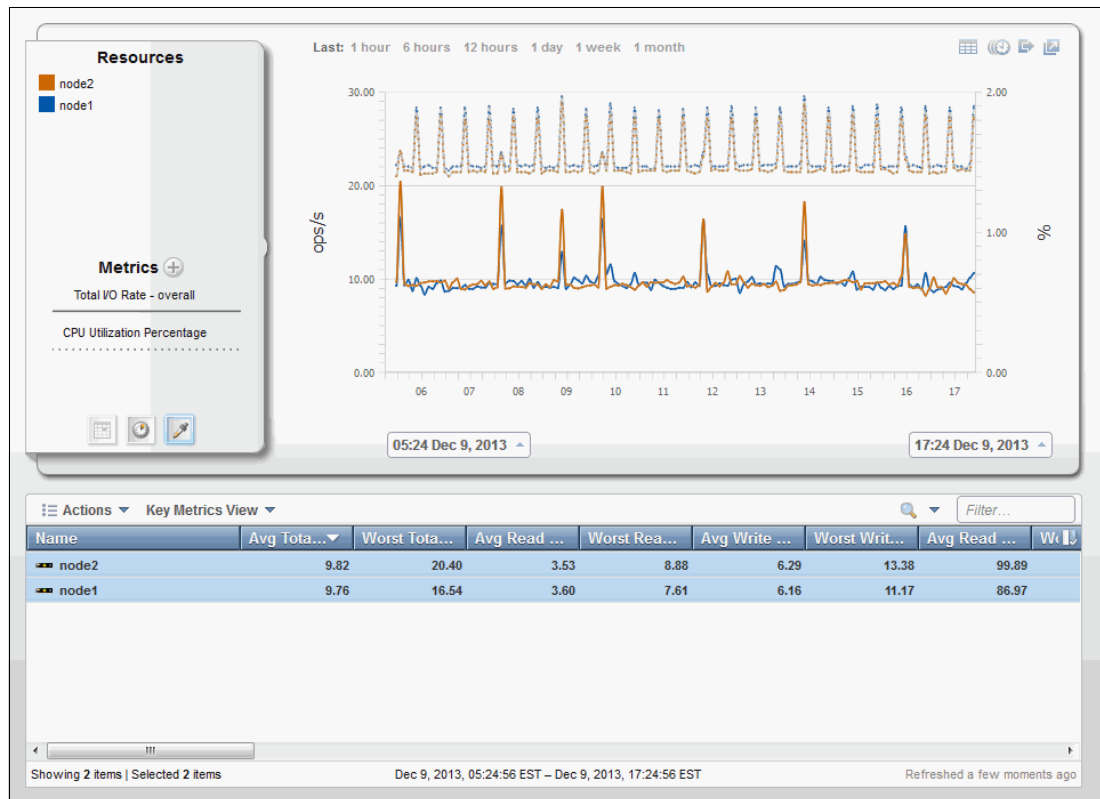


Figure 13-36 Viewing the performance of I/O group nodes

This report can be used to determine whether the I/O workload is evenly distributed across the nodes in the I/O group. An absence of I/O on a single node, or disproportionately large amount of I/O on a single node might indicate a configuration issue. In our example, I/Os are present on each node and appear to be evenly distributed.

CPU Utilization Percentage metric

I/O Group CPU Utilization indicates how busy the cluster nodes are. To generate a graph of CPU utilization, click the + next to the Metrics heading in the legend area to the left of the chart, as shown in Figure 13-34 on page 394, select the **CPU Utilization Percentage** metric and clear any other selected metrics, as shown in Figure 13-36, then click **OK**.

This chart updates to display the CPU Utilization Percentage over the selected time range, as shown in Figure 13-37 on page 396.

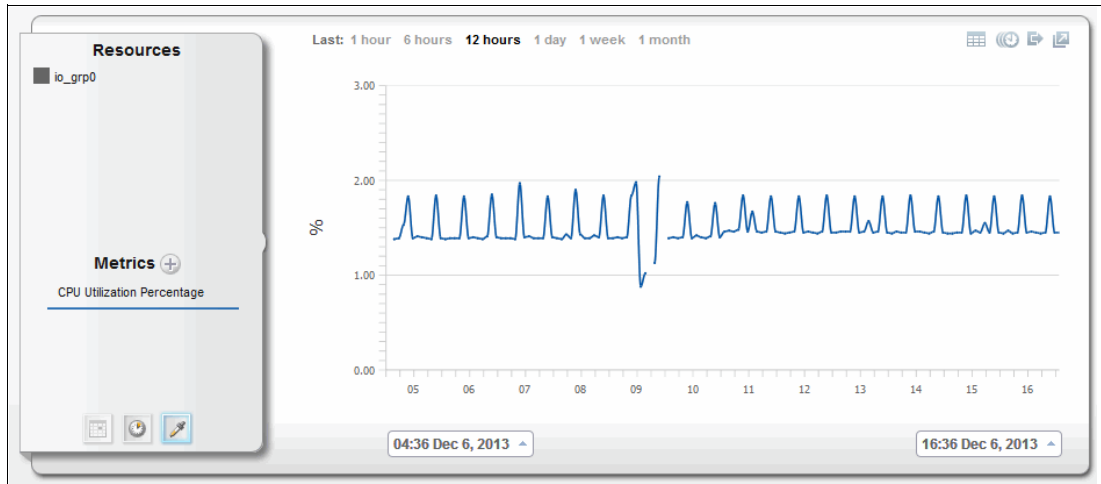


Figure 13-37 CPU utilization selection for SAN Volume Controller

A continual high CPU Utilization rate indicates a busy I/O group. In our environment, CPU utilization does not rise above 2%, which indicates that our I/O Group is mostly idle.

Tip: To view the CPU Utilization percentage of the individual nodes in the I/O Group, open a new performance window for the I/O Group nodes, as shown in Figure 13-33 on page 394. In that window, follow the same steps as described in this section.

CPU utilization guidelines for SAN Volume Controller only

If the CPU utilization for the SAN Volume Controller node remains constantly above 70%, it might be time to increase the number of I/O groups in the cluster. You can also redistribute workload to other I/O groups in the SAN Volume Controller cluster if it is available. You can add cluster I/O groups up to the maximum of four I/O groups per SAN Volume Controller cluster.

If four I/O groups are in a cluster (with the latest firmware installed) and you are still having high SAN Volume Controller node CPU utilization as indicated in the reports, build a cluster. Consider migrating some storage to the new cluster; or, if existing SAN Volume Controller nodes are not of the 2145-CG8 version, upgrade them to the CG8 nodes.

Backend Response Time

To view the disk read and write response time at the node level, click the + next to the Metrics heading in the legend area to the left of the chart (as shown in Figure 13-34 on page 394), clear any selected metrics, select the **Response Time (ms/op) Read** and **Write** metrics, as shown in Figure 13-38, and then click **OK**.

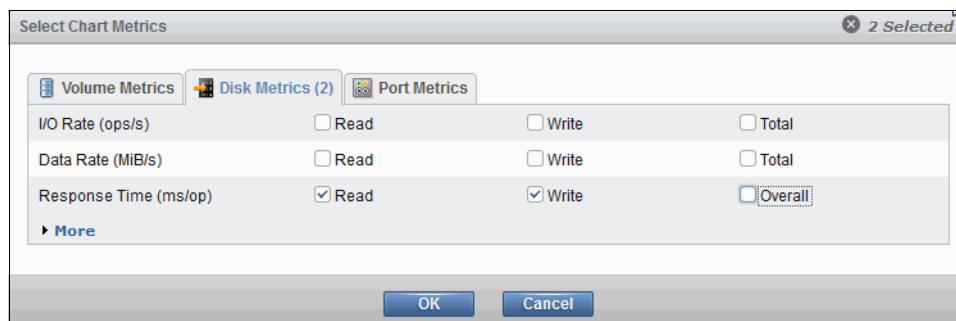


Figure 13-38 Disk metrics selection

The graph updates to display the average backend read and write response times for the disks that are servicing the nodes in the I/O group, as shown in Figure 13-39.

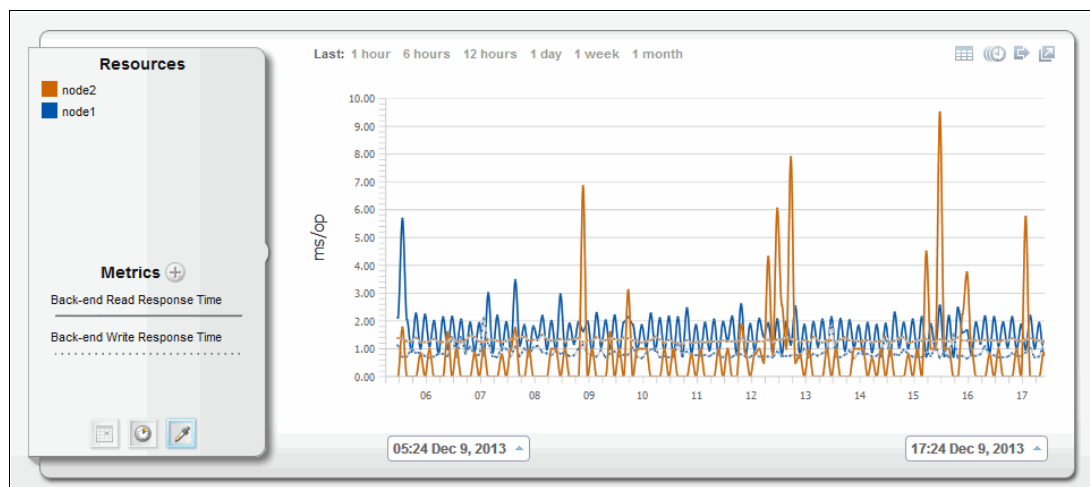


Figure 13-39 Backend response time for I/O group nodes

Guidelines for poor response times

For random read I/O, the back-end rank (disk) read response times should seldom exceed 25 ms, unless the read hit ratio is near 99%. Backend Write Response Times are higher because of RAID 5 (or RAID 10) algorithms, but should seldom exceed 80 ms. Some time intervals might exist when response times exceed these guidelines.

If you are experiencing poor response times, use all available information from the SAN Volume Controller and the back-end storage controller to investigate the times. The following possible causes for a significant change in response times from the back-end storage might be visible by using the storage controller management tool:

- ▶ A physical array drive failure that leads to an array rebuild. This failure drives more internal read/write workload of the back-end storage subsystem when the rebuild is in progress. If this situation causes poor latency, you might want to adjust the array rebuild priority to reduce the load. However, the array rebuild priority must be balanced with the increased risk of a second drive failure during the rebuild, which might cause data loss in a RAID 5 array.
- ▶ Cache battery failure that leads to the controller disabling the cache. You can usually resolve this situation by replacing the failed battery.

For more information about rules of thumb and how to interpret the values, see *SAN Storage Performance Management Using Tivoli Storage Productivity Center*, SG24-7364.

Data Rate

To view the data rates of the nodes in the I/O group, click the + next to the Metrics heading in the legend area to the left of the chart (as shown in Figure 13-37 on page 396) clear any selected metrics, select the **Data Rate (MiB/s) Read** and **Write** metrics (as shown in Figure 13-40), and then click **OK**.

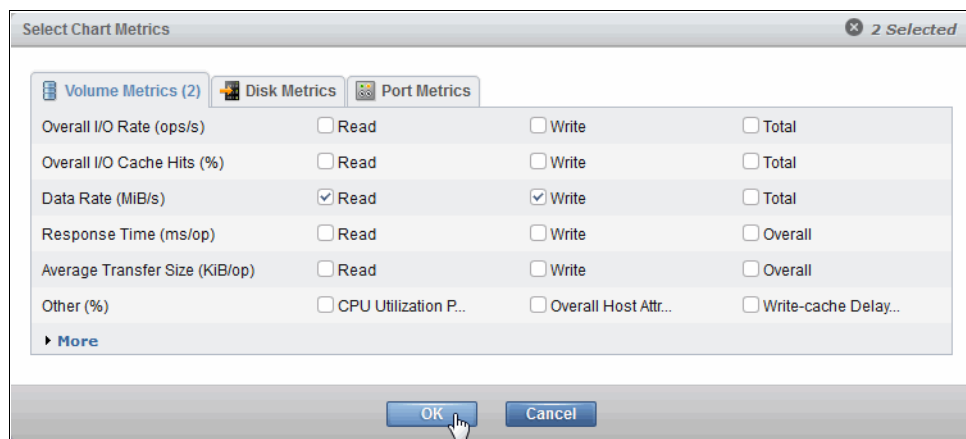


Figure 13-40 Data rate metrics selection

The graph updates to display the average read and write data rates for the nodes in the I/O group, as shown in Figure 13-41.

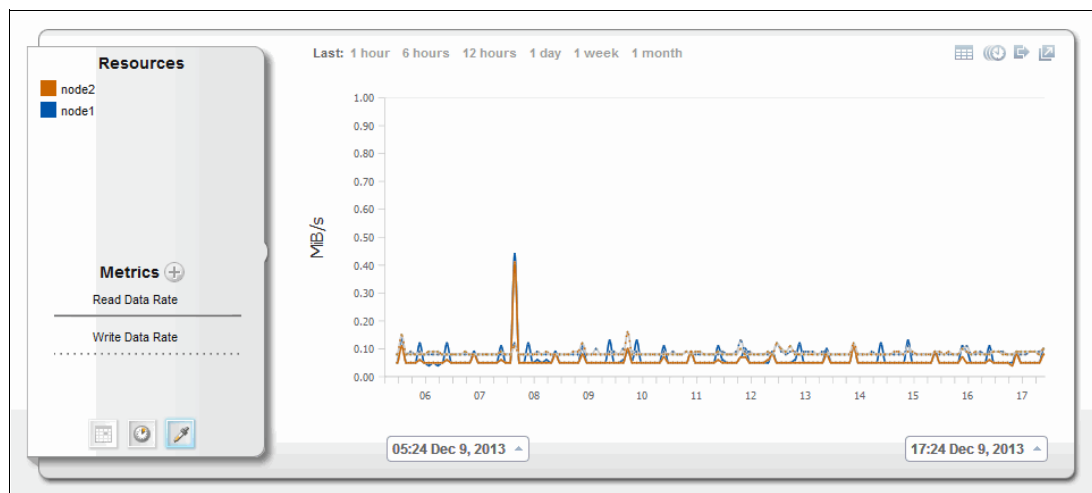


Figure 13-41 Node data rates

Understanding your performance results

To interpret your performance results, always return to your baseline. For more information about creating a baseline, see *SAN Storage Performance Management Using Tivoli Storage Productivity Center*, SG24-7364.

Some industry benchmarks for the SAN Volume Controller and Storwize V7000 are available. SAN Volume Controller V6.4 and the CG8 node brought a dramatic increase in performance as demonstrated by the results in the Storage Performance Council (SPC) Benchmarks, SPC-1, and SPC-2.

► *SPC Benchmark 1 Executive Summary: IBM System Storage SAN Volume Controller SPC-2 V1.3*, which is available at this website:

► *SPC Benchmark 2 Executive Summary: IBM System Storage SAN Volume Controller V6.4 with IBM Storwize V7000 Disk Storage SPC-2 V1.3 at:*

SPC-1 and SPC-2 Benchmarks were also performed for Storwize V7000. For more information, see the following publications:

http://www.storageperformance.org/benchmark_results_files/SPC-2/IBM_SPC-2/B00052_IBM_Storwize-V7000/b00052_IBM_Storwize-V7000_SPC2_executive-summary.pdf

- ▶ The specific SAN Volume Controller nodes in your configuration
- ▶ The type of managed disks (volumes) in the managed disk group
- ▶ The application I/O workloads that use the managed disk group
- ▶ The paths to the back-end storage

In reviewing the SPC benchmark (see Figure 13-42), the results for the I/O and Data Rate are different depending on the transfer block size used.

Max I/Os and MBps Per I/O Group – 70/30 Read/Write Miss

- **2145-8G4**
 - 4K Transfer Size
 - 122K – 500 MBps
 - 64K Transfer Size
 - 29K – 1.8 GBps
- **2145-8F4**
 - 4K Transfer Size
 - 72K – 300 MBps
 - 64K Transfer Size
 - 23K – 1.4 GBps
- **2145-4F2**
 - 4K Transfer Size
 - 38K – 156 MBps
 - 64K Transfer Size
 - 11K – 700 MBps
- **2145-8F2**
 - 4K Transfer Size
 - 72K – 300 MBps
 - 64K Transfer Size
 - 15K – 1 GBps

Chapter 13. Monitoring 399

Reviewing the two-node I/O group that is used, you might see 122,000 I/Os if all of the transfer blocks were 4 K. However, they rarely are 4 K in typical environments. If you go down to 64 K (or bigger) with anything over about 32 K, you might realize a result more typical of the 29,000 I/Os as noticed by the SPC benchmark.

13.3.2 Node Cache Performance for SAN Volume Controller and Storwize V7000

Node cache performance is described in this section.

Node Cache Performance by using the stand-alone GUI

Efficient use of cache can help enhance virtual disk I/O response time. The Node Cache Performance report displays a list of cache-related metrics, such as Read and Write Cache Hits percentage and Read Ahead percentage of cache hits.

The cache memory resource reports provide an understanding of the utilization of the SAN Volume Controller or Storwize V7000 cache. These reports provide an indication of whether the cache can service and buffer the current workload.

To access these reports, click **IBM Tivoli Storage Productivity Center**  **Reporting**  **System Reports** , and then select **Module/Node Cache performance report**. This report is generated at the SAN Volume Controller and Storwize V7000 node level (an entry that refers to an IBM XIV storage device), as shown in Figure 13-43.

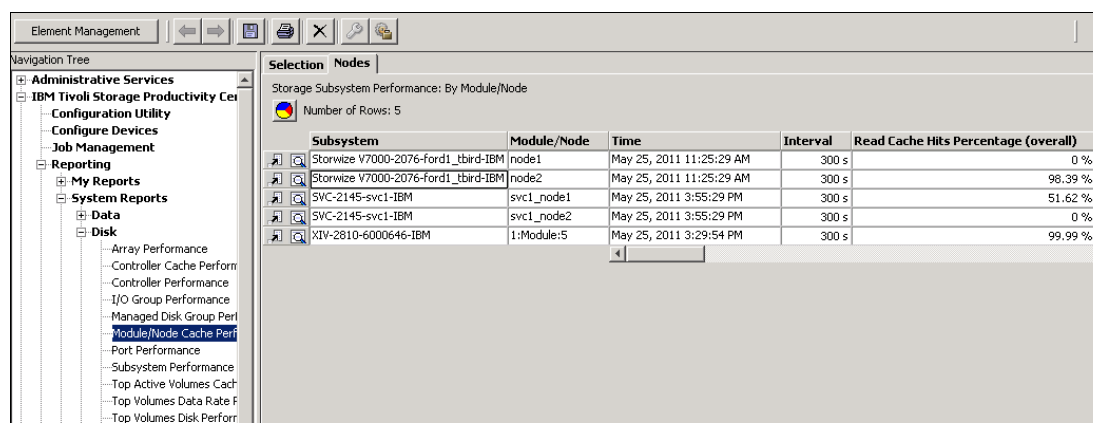


Figure 13-43 Module/Node Cache Performance report for SAN Volume Controller and Storwize V7000

Cache Hit percentage

Total Cache Hit percentage is the percentage of reads and writes that are handled by the cache without needing immediate access to the back-end disk arrays. Read Cache Hit percentage focuses on reads because writes are almost always recorded as cache hits. If the cache is full, a write might be delayed when some changed data is destaged to the disk arrays to make room for the new write data. The Read and Write Transfer Sizes are the average number of bytes that are transferred per I/O operation.

To review the read cache hits percentage for Storwize V7000 nodes, complete the following steps:

1. Select both nodes.
2. Click the pie chart icon ().

3. Select **Read Cache Hits percentage (overall)**, and then click **OK** to generate the chart, as shown in Figure 13-44.

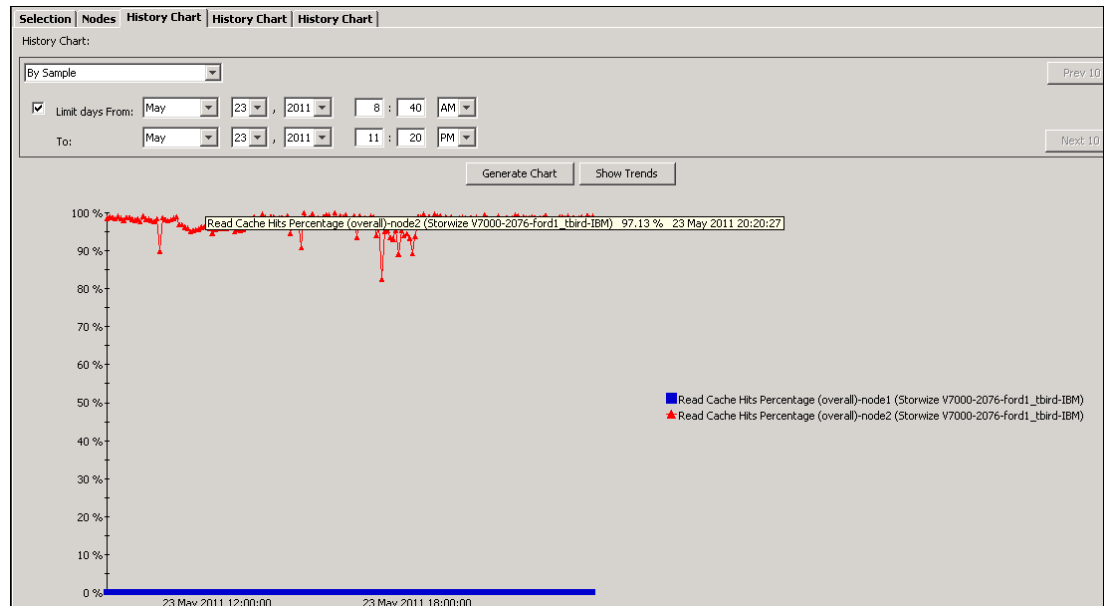


Figure 13-44 Storwize V7000 Cache Hits percentage that shows no traffic on node1

Important: The flat line for node 1 does not mean that the read request for that node cannot be handled by the cache. It means that no traffic is on that node, as shown in Figure 13-45 on page 402 and Figure 13-46 on page 402, where Read Cache Hit Percentage and Read I/O Rates are compared in the same time interval.

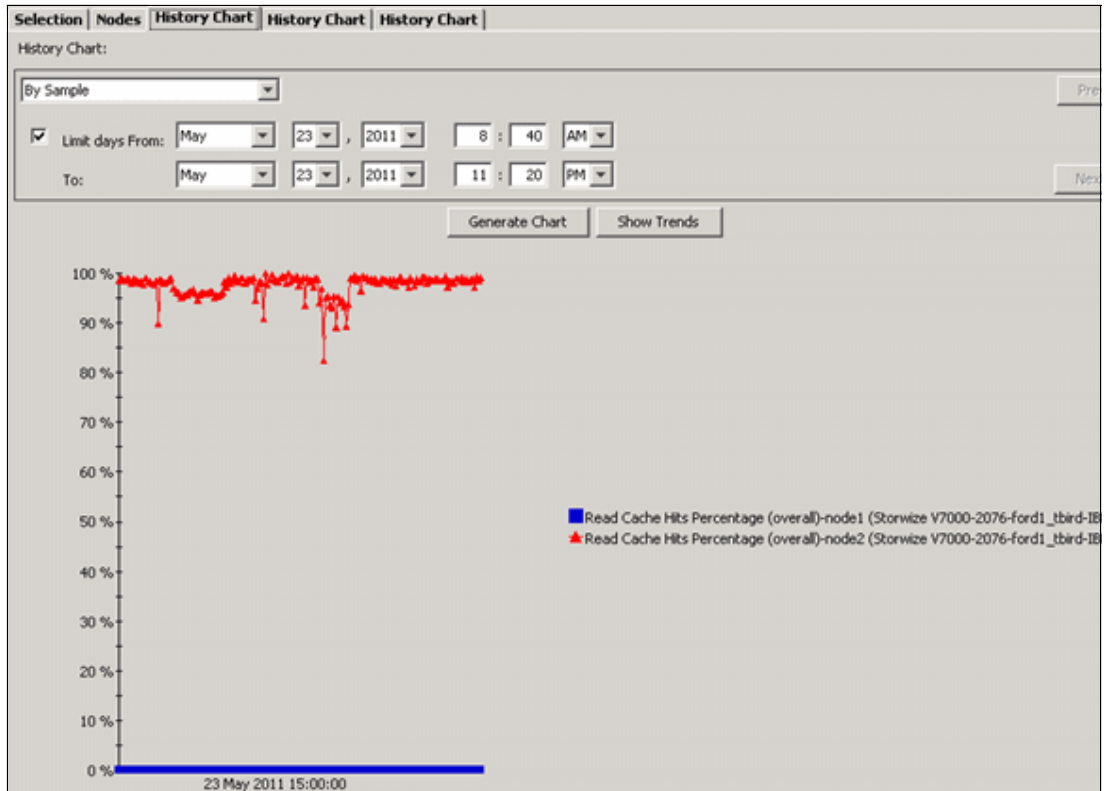


Figure 13-45 Storwize V7000 Read Cache Hit Percentage

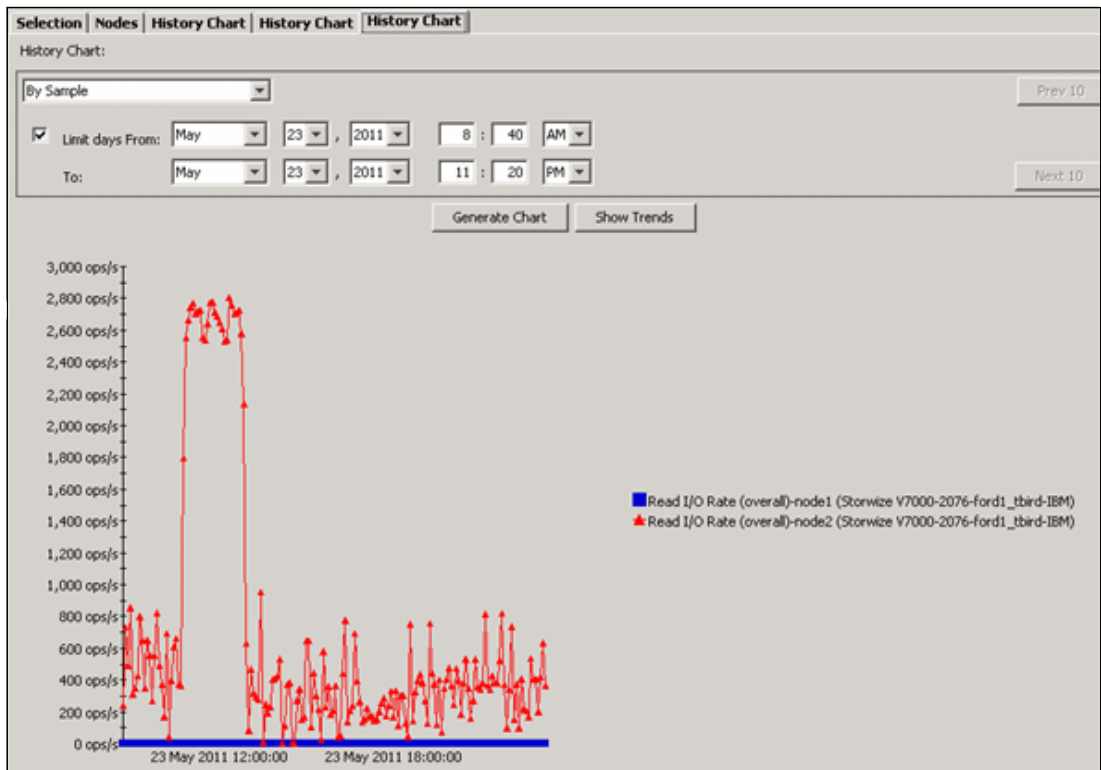


Figure 13-46 Storwize V7000 Read I/O Rate

This configuration might not be good because the two nodes are not balanced. In the lab environment for this book, the volumes that were defined on Storwize V7000 were all defined with node 2 as the preferred node.

After we moved the preferred node for the tpcb1ade3-7-ko volume from node 2 to node 1, we obtained the graph for Read Cache Hit percentage that is shown in Figure 13-47.

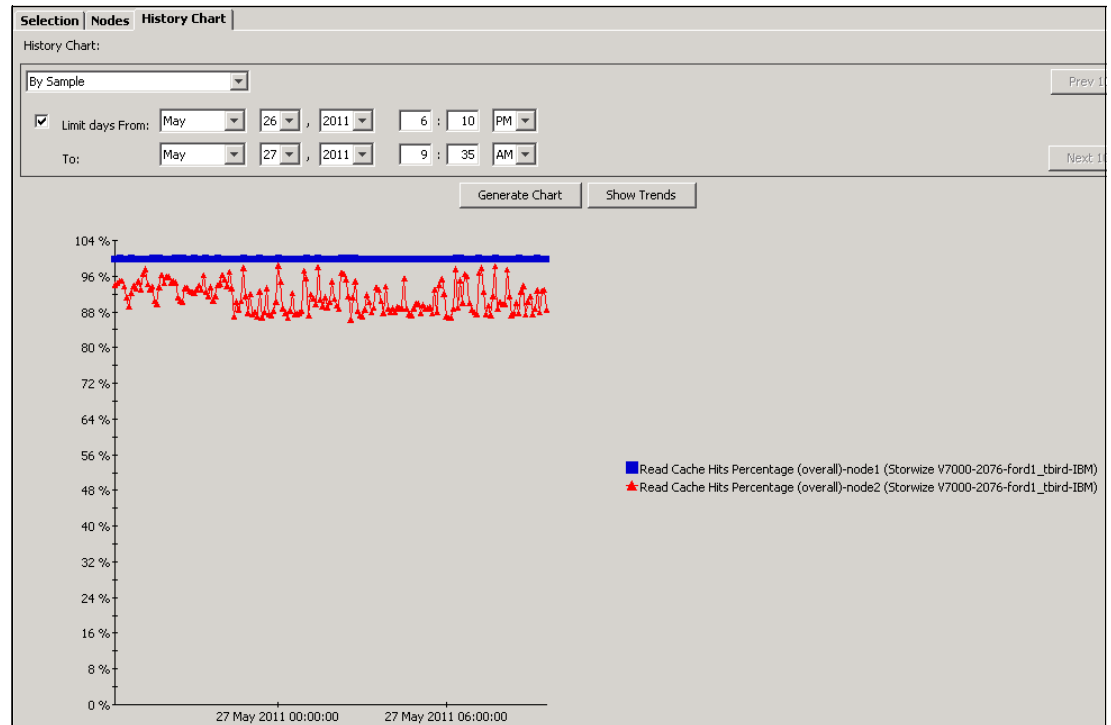


Figure 13-47 Cache Hit Percentage for Storwize V7000 after reassignment

We also obtained the graph for Read I/O Rates that is shown in Figure 13-48.

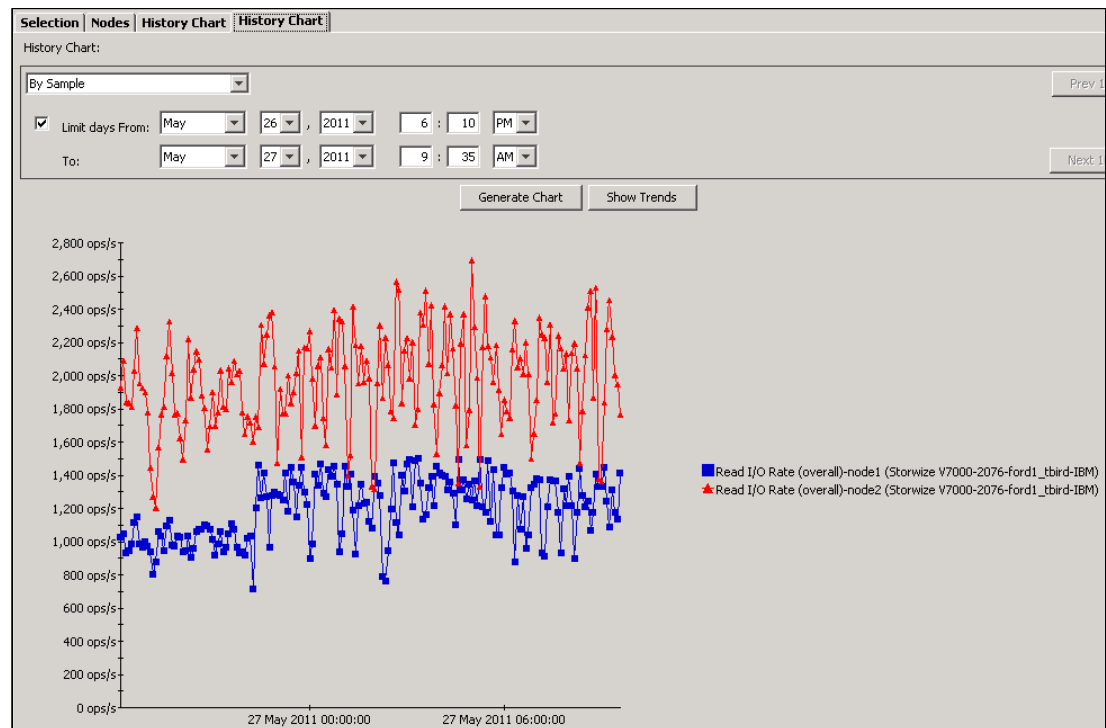


Figure 13-48 Read I/O rate for Storwize V7000 after reassignment

More analysis of Read Hit percentages

Read Hit percentages can vary from near 0% to near 100%. Any percentage below 50% is considered low, but many database applications show hit ratios below 30%. For low-hit ratios, you need many ranks that provide a good back-end response time. It is difficult to predict whether more cache improves the hit ratio for a particular application. Hit ratios depend more on the application design and the amount of data than on the size of cache (especially for Open System workloads). However, larger caches are always better than smaller caches. For high-hit ratios, the back-end ranks can be driven a little harder to higher utilizations.

If you must analyze further cache performances and to understand whether it is enough for your workload, you can run multiple metrics charts. The following metrics are available:

- ▶ CPU utilization percentage
The average utilization of the node controllers in this I/O group during the sample interval.
- ▶ Dirty Write percentage of Cache Hits
The percentage of write cache hits that modified only data that was already marked dirty (rewritten data) in the cache. This measurement is an obscure way to determine how effectively writes are coalesced before destaging.
- ▶ Read/Write/Total Cache Hits percentage (overall)
The percentage of reads/writes/total cache hits during the sample interval that are found in cache. This metric is important to monitor. The write cache hit percentage must be nearly 100%.
- ▶ Readahead percentage of Cache Hits
An obscure measurement of cache hits that involve data that was prestaged for one reason or another.

- ▶ **Write Cache Flush-through percentage**
For SAN Volume Controller and Storwize V7000, the percentage of write operations that were processed in Flush-through write mode during the sample interval.
- ▶ **Write Cache Overflow percentage**
For SAN Volume Controller and Storwize V7000, the percentage of write operations that were delayed because of a lack of write-cache space during the sample interval.
- ▶ **Write Cache Write-through percentage**
For SAN Volume Controller and Storwize V7000, the percentage of write operations that were processed in Write-through write mode during the sample interval.
- ▶ **Write Cache Delay percentage**
The percentage of all I/O operations that were delayed because of write-cache space constraints or other conditions during the sample interval. Only writes can be delayed, but the percentage is of all I/O.
- ▶ **Small Transfers I/O percentage**
Percentage of I/O operations over a specified interval. Applies to data transfer sizes that are less than or equal to 8 KB.
- ▶ **Small Transfers Data percentage**
Percentage of data that was transferred over a specified interval. Applies to I/O operations with data transfer sizes that are less than or equal to 8 KB.
- ▶ **Medium Transfers I/O percentage**
Percentage of I/O operations over a specified interval. Applies to data transfer sizes that are greater than 8 KB and less than or equal to 64 KB.
- ▶ **Medium Transfers Data percentage**
Percentage of data that was transferred over a specified interval. Applies to I/O operations with data transfer sizes that are greater than 8 KB and less than or equal to 64 KB.
- ▶ **Large Transfers I/O percentage**
Percentage of I/O operations over a specified interval. Applies to data transfer sizes that are greater than 64 KB and less than or equal to 512 KB.
- ▶ **Large Transfers Data percentage**
Percentage of data that was transferred over a specified interval. Applies to I/O operations with data transfer sizes that are greater than 64 KB and less than or equal to 512 KB.
- ▶ **Very Large Transfers I/O percentage**
Percentage of I/O operations over a specified interval. Applies to data transfer sizes that are greater than 512 KB.
- ▶ **Very Large Transfers Data percentage**
Percentage of data that was transferred over a specified interval. Applies to I/O operations with data transfer sizes that are greater than 512 KB.
- ▶ **Overall Host Attributed Response Time Percentage**
The percentage of the average response time (read response time and write response time) that can be attributed to delays from host systems. This metric is provided to help diagnose slow hosts and poorly performing fabrics. The value is based on the time it takes for hosts to respond to transfer-ready notifications from the SAN Volume Controller nodes (for read). The value is also based on the time it takes for hosts to send the write data after the node responded to a transfer-ready notification (for write).

The Global Mirror Overlapping Write Percentage metric is applicable only in a Global Mirror Session. This metric is the average percentage of write operations that are issued by the Global Mirror primary site and that were serialized overlapping writes for a component over a specified time interval. For SAN Volume Controller V4.3.1 and later, some overlapping writes are processed in parallel (not serialized) and are excluded. For earlier SAN Volume Controller versions, all overlapping writes were serialized.

Select the metrics that are named “percentage” because you can have multiple metrics, with the same unit type, in one chart. Complete the following steps:

1. In the Selection window (as shown in Figure 13-49), move the percentage metrics that you want include from the Available Column to the Included Column. Then, click **Selection** to select only the Storwize V7000 entries.
2. In the Select Resources window, select the node or nodes, and then click **OK**.

Figure 13-49 shows an example where several percentage metrics are chosen for Storwize V7000.

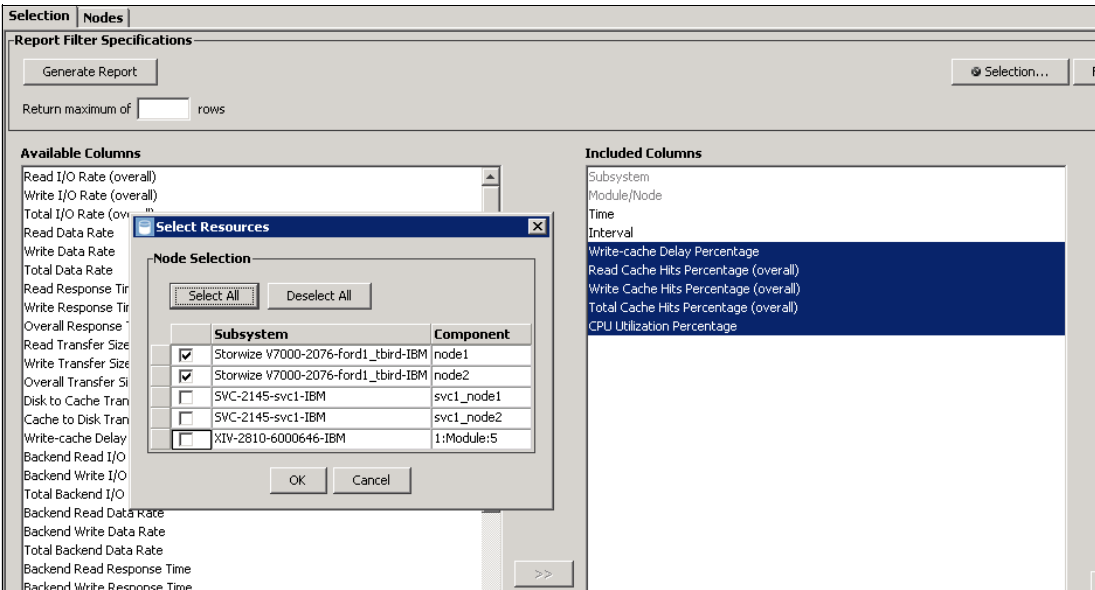


Figure 13-49 Storwize V7000 multiple metrics Cache selection

3. In the Select Charting Options window, select all the metrics, and then click **OK** to generate the chart.

As shown in Figure 13-50, we notice in our test a drop in the Cache Hits percentage. Even a drop that is not so dramatic can be considered as an example for further investigation of problems that arise.

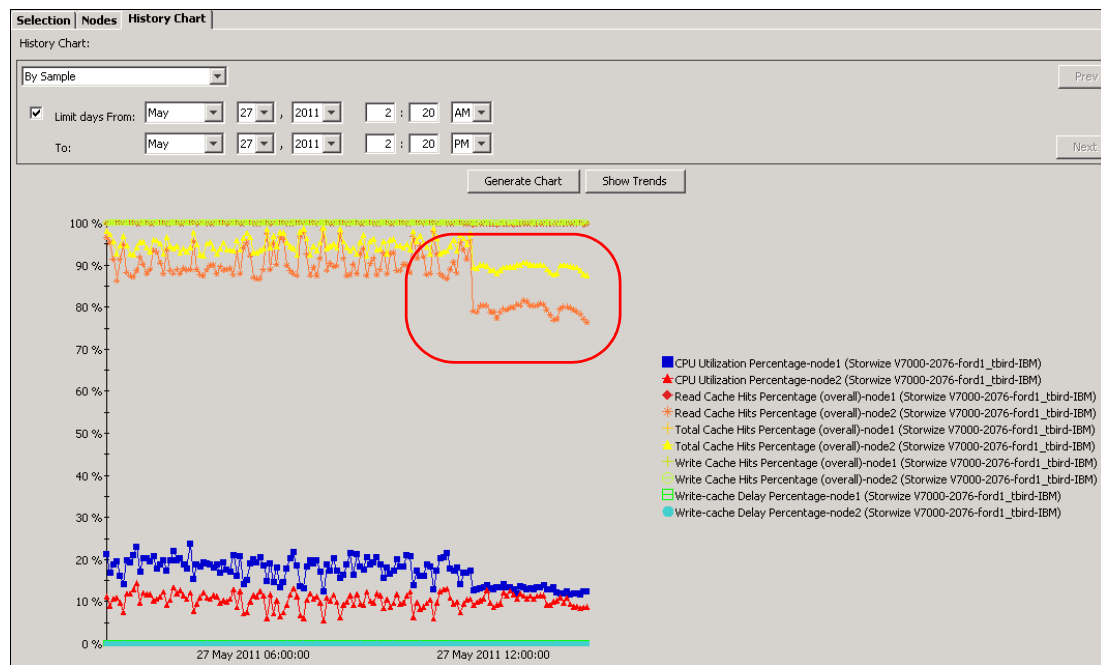


Figure 13-50 Resource performance metrics for multiple Storwize V7000 nodes

Changes in these performance metrics and an increase in back-end response time (see Figure 13-51) shows that the storage controller is heavily burdened with I/O, and the Storwize V7000 cache can become full of outstanding write I/Os.

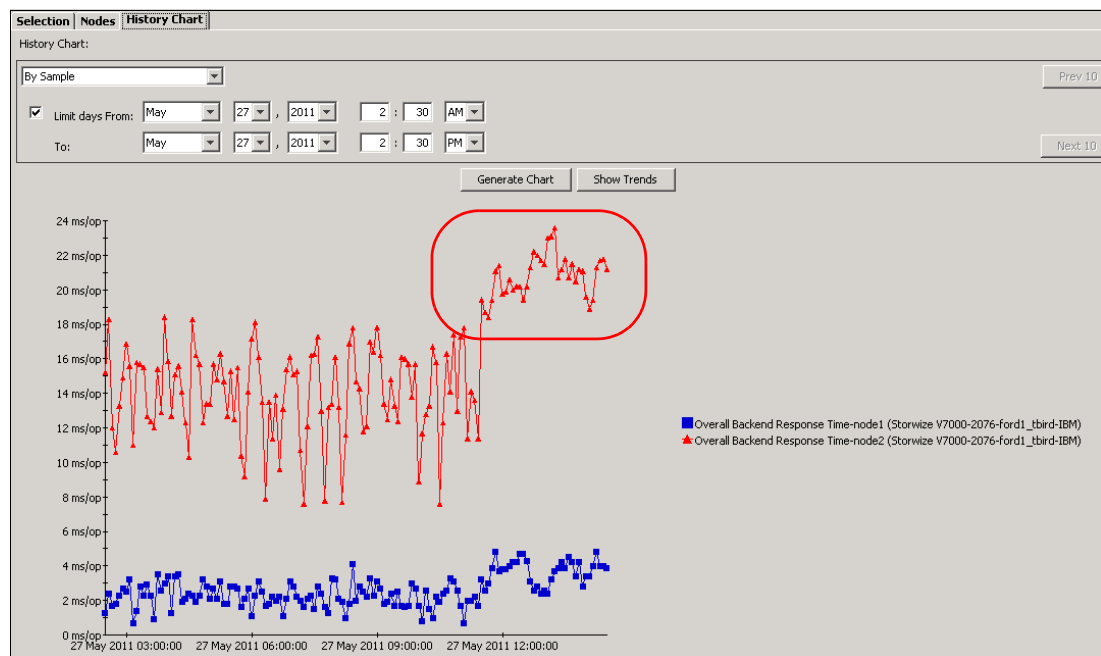


Figure 13-51 Increased overall back-end response time for Storwize V7000

Host I/O activity is affected with the backlog of data in the Storwize V7000 cache and with any other Storwize V7000 workload that is going on to the same MDisks.

I/O groups: If cache utilization is a problem, you can add cache to the cluster by adding an I/O group and moving volumes to the new I/O Group in SAN Volume Controller and Storwize V7000 V6.2. Also, adding an I/O group and moving a volume from one I/O group to another are still disruptive actions. Therefore, you must properly plan how to manage this disruption.

For more information about rules of thumb and how to interpret these values, see *SAN Storage Performance Management Using Tivoli Storage Productivity Center*, SG24-7364.

Efficient use of cache can help enhance virtual disk I/O response time. Node cache performance can be easily viewed in Tivoli Storage Productivity Center 5.2, with which you can monitor cache related metrics, such as Read and Write Cache Hits percentage and Read Ahead and Dirty Writes percentage of cache hits. These metrics can provide an indication of whether the cache can service and buffer the current workload.

Node Cache Performance by using the web-based GUI

To view the performance of a SAN Volume Controller Node, start by clicking **Nodes** under the Internal Resources section, as shown in Figure 13-52.



Figure 13-52 Accessing the Nodes view for a storage system

This loads the Nodes window for this storage system, which displays the nodes for this system in a tabular format with various details, as shown in Figure 13-53 on page 409. In our lab environment, data was collected for a SAN Volume Controller with two nodes.

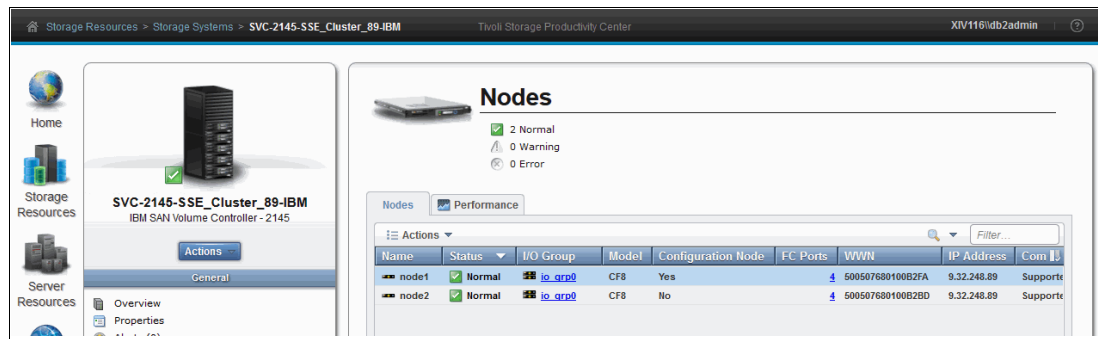


Figure 13-53 The Nodes page for a storage system

To quickly view the performance of all the Nodes in this storage system, click the **Performance** tab. Alternatively, to view the performance of a specific node or nodes (that is, the two nodes in an I/O group), highlight the row (or rows) of the wanted nodes by pressing Ctrl and then left-click, then right-click in any of the highlighted rows in the Nodes tab and click **View Performance**, as shown in Figure 13-54.

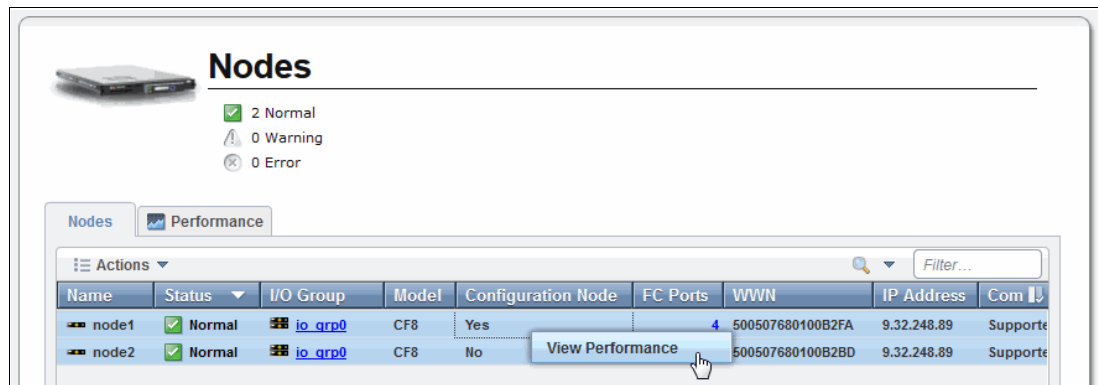


Figure 13-54 Accessing the performance view of nodes

Tip: When you analyze node cache performance, it is often useful to view the cache metrics of both nodes in an I/O group in the same chart. You can determine whether your workload is evenly balanced across the nodes in the I/O group.

A window opens that displays the performance of the selected nodes, as shown in Figure 13-55 on page 410.

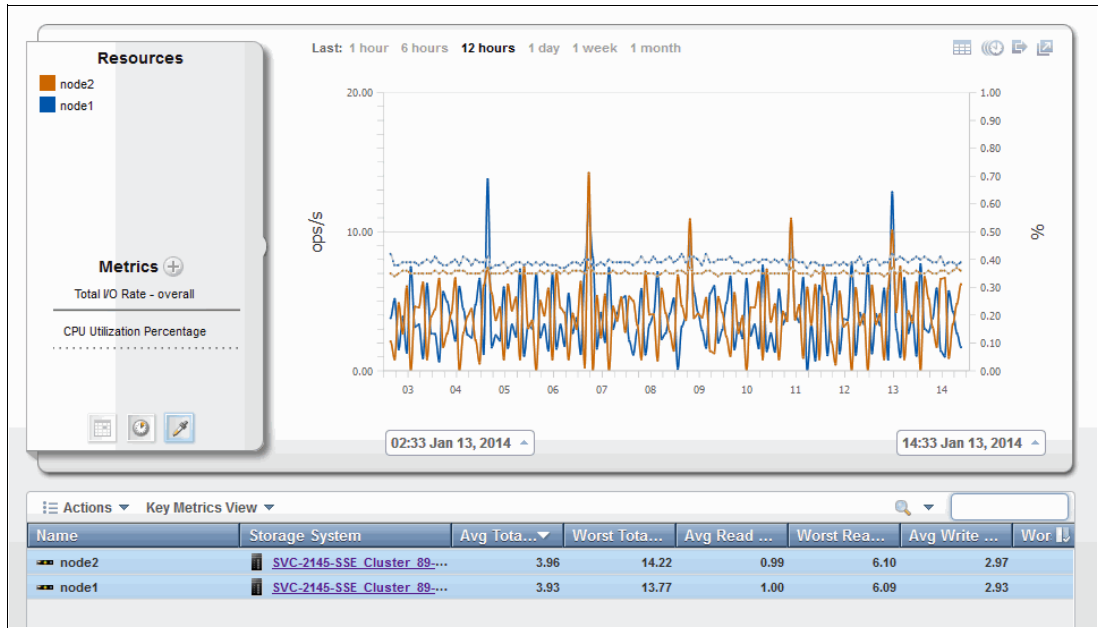


Figure 13-55 Viewing the performance of nodes

The default view in the nodes performance window is similar to that of the I/O group performance window. For more information about customizing the performance view, see 13.3.1, “I/O Group Performance for SAN Volume Controller and Storwize V7000” on page 384.

Cache Hit percentage

Total Cache Hit percentage is the percentage of reads and writes that are handled by the cache without needing immediate access to the back-end disk arrays. *Read Cache Hit* percentage focuses on reads because writes are almost always recorded as cache hits. If the cache is full, a write might be delayed when some changed data is destaged to the disk arrays to make room for the new write data.

To display the cache performance-related metrics in the chart, click the + that is next to the Metrics heading in the legend area to the left of the chart, as shown in Figure 13-56.

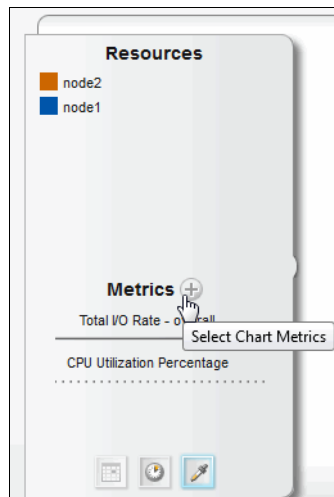


Figure 13-56 Modifying the nodes performance page to display cache related metrics

A window opens in which you can choose the metrics to display, as shown in Figure 13-57.

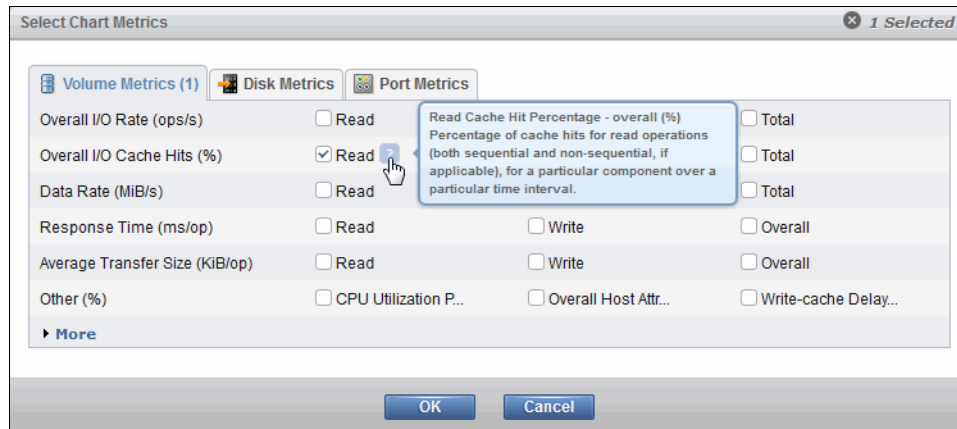


Figure 13-57 Selecting Read Cache Hits Percentage

In this example, we chose to display Read Cache Hit Percentage only for both nodes in the I/O group, as shown in Figure 13-58.

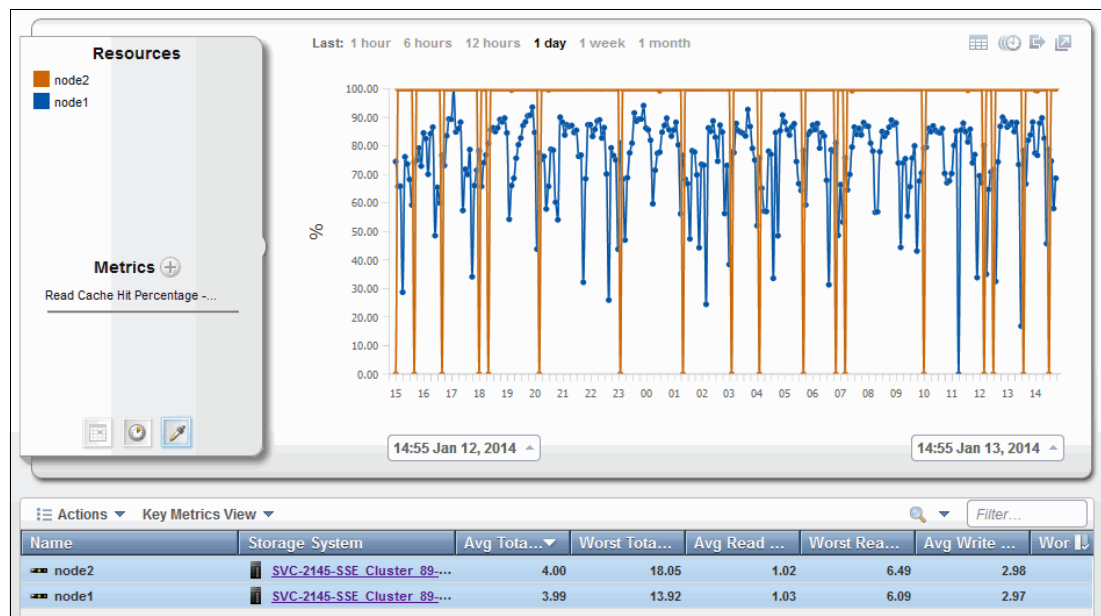


Figure 13-58 Node Read Cache Hit Percentage

Important: A Read Cache Hit Percentage at or close to zero does not necessarily mean that the read requests for that node cannot be handled by the cache. It might mean that no traffic is arriving at that node. You can verify this fact by also displaying the Read I/O Rate metric in the chart. If doing so verifies that there is little or no I/O on that node, consider whether to balance the workload more evenly across the nodes by modifying the preferred nodes settings for the volumes in the I/O group.

More analysis of Read Hit percentages

Read Hit percentages can vary from near 0% to near 100%. Any percentage below 50% is considered low, but many database applications show hit ratios below 30%. For low hit ratios, you need many ranks that provide a good back-end response time. It is difficult to predict whether more cache improves the hit ratio for a particular application. Hit ratios depend more on the application design and the amount of data than on the size of cache (especially for Open System workloads). However, larger caches are always better than smaller caches. For high-hit ratios, the back-end ranks can be driven a little harder to higher utilizations.

If you must analyze further cache performances and to understand whether it is enough for your workload, you can run multiple metrics charts. The following metrics are available:

- ▶ CPU utilization percentage
The average CPU utilization of the node during the sample interval.
- ▶ Dirty Write percentage of Cache Hits
The percentage of write cache hits that modified only data that was already marked dirty (rewritten data) in the cache. This measurement is an obscure way to determine how effectively writes are coalesced before destaging.
- ▶ Read/Write/Total Cache Hits percentage (overall)
The percentage of reads/writes/total cache hits during the sample interval that are found in cache. This metric is important to monitor. The write cache hit percentage should be nearly 100%.
- ▶ Readahead percentage of Cache Hits
An obscure measurement of cache hits that involve data that was prestaged for some reason.
- ▶ Write Cache Flush-through percentage
For SAN Volume Controller and Storwize V7000, the percentage of write operations that were processed in flush-through write mode during the sample interval.
- ▶ Write Cache Overflow percentage
For SAN Volume Controller and Storwize V7000, the percentage of write operations that were delayed because of a lack of write-cache space during the sample interval.
- ▶ Write Cache Write-through percentage
For SAN Volume Controller and Storwize V7000, the percentage of write operations that were processed in write-through write mode during the sample interval.
- ▶ Write Cache Delay percentage
The percentage of all I/O operations that were delayed because of write-cache space constraints or other conditions during the sample interval. Only writes can be delayed, but the percentage is of all I/O.
- ▶ Overall Host Attributed Response Time Percentage
The percentage of the average response time (read response time and write response time) that can be attributed to delays from host systems. This metric is provided to help diagnose slow hosts and poorly performing fabrics. The value is based on the time it takes for hosts to respond to transfer-ready notifications from the SAN Volume Controller nodes (for read). The value is also based on the time it takes for hosts to send the write data after the node responded to a transfer-ready notification (for write).

► Overall Host Attributed Response Time Percentage

The percentage of the average response time (read response time and write response time) that can be attributed to delays from host systems. This metric is provided to help diagnose slow hosts and poorly performing fabrics. The value is based on the time it takes for hosts to respond to transfer-ready notifications from the SAN Volume Controller nodes (for read). The value is also based on the time it takes for hosts to send the write data after the node responded to a transfer-ready notification (for write).

For more information about rules of thumb and how to interpret these values, see *SAN Storage Performance Management Using Tivoli Storage Productivity Center*, SG24-7364.

13.3.3 Viewing the Managed Disk Group Performance report for SAN Volume Controller by using the stand-alone GUI

The Managed Disk Group Performance report provides disk performance information at the managed disk group level. It summarizes the read and write transfer size and the back-end read, write, and total I/O rate. From this report, you can easily browse to see the statistics of virtual disks that are supported by a managed disk group or drill down to view the data for the individual MDisks that make up the managed disk group.

To access this report, click **IBM Tivoli Storage Productivity Center** ◊ **Reporting** ◊ **System Reports** ◊ **Disk**, and select **Managed Disk Group Performance**. A table is displayed (as shown in Figure 13-59) that lists all the known managed disk groups and their last collected statistics, which are based on the latest performance data collection.

Subsystem	Managed Disk Group	Time	Interval	Backend Read Data Rate	Backend Write Data Rate
SVC-2145-svc1-IBM	CET_DS8K1901mdg	May 26, 2011 12:00:06 AM	300 s	8.89 MB/s	0 MB/s
SVC-2145-svc1-IBM	codrin_MIDGroup	May 26, 2011 12:00:06 AM	300 s	0 MB/s	0 MB/s
SVC-2145-svc1-IBM	ds8k_1901	May 26, 2011 12:00:06 AM	300 s	0 MB/s	0 MB/s
SVC-2145-svc1-IBM	ds8k_1k_vd	May 26, 2011 12:00:06 AM	300 s	0 MB/s	0 MB/s
SVC-2145-svc1-IBM	Irina_mdg_xiv	May 26, 2011 12:00:06 AM	300 s	0 MB/s	0 MB/s
SVC-2145-svc1-IBM	tpc_r	May 26, 2011 12:00:06 AM	300 s	0 MB/s	0 MB/s

Figure 13-59 Managed Disk Group Performance report

One of the managed disk groups is named CET_DS8K1901mdg. When you click the magnifying glass icon (🔍) for the CET_DS8K1901mdg entry, a new page opens (as shown in Figure 13-60) that shows the managed disks in the managed disk group.

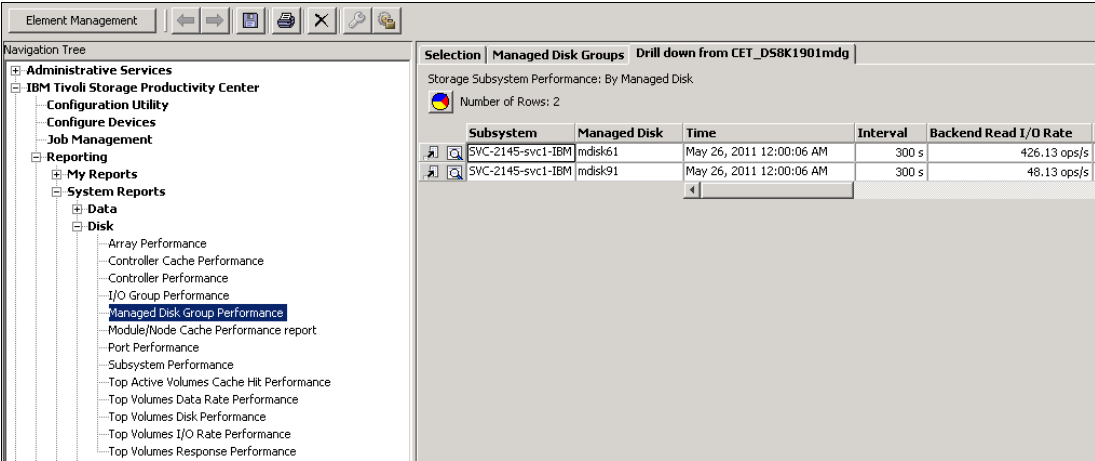


Figure 13-60 Drill down from Managed Disk Group Performance report

When you click the magnifying glass icon (🔍) for the mdisk61 entry, a new page (as shown in Figure 13-61) opens that shows the volumes in the managed disk.

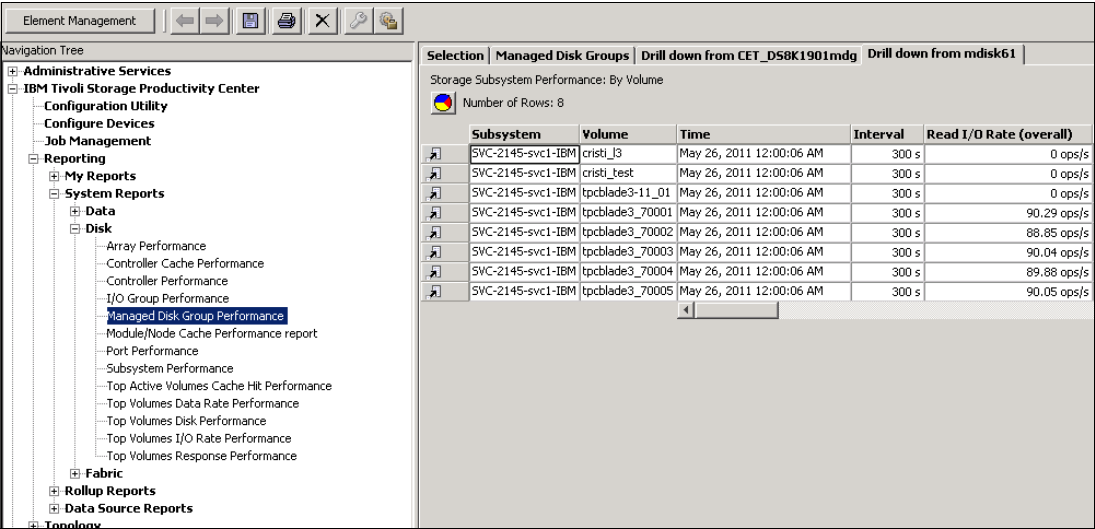


Figure 13-61 Drill down from Managed Disk Performance report

Back-end I/O Rate

Complete the following steps to analyze how the I/O workload is split between the managed disk groups to determine whether it is well-balanced.

1. On the Managed Disk Groups tab, select all managed disk groups and click the pie chart icon ().
2. In the Select Charting Option window (see Figure 13-62), select **Total Backend I/O Rate**. Then, click **OK**.

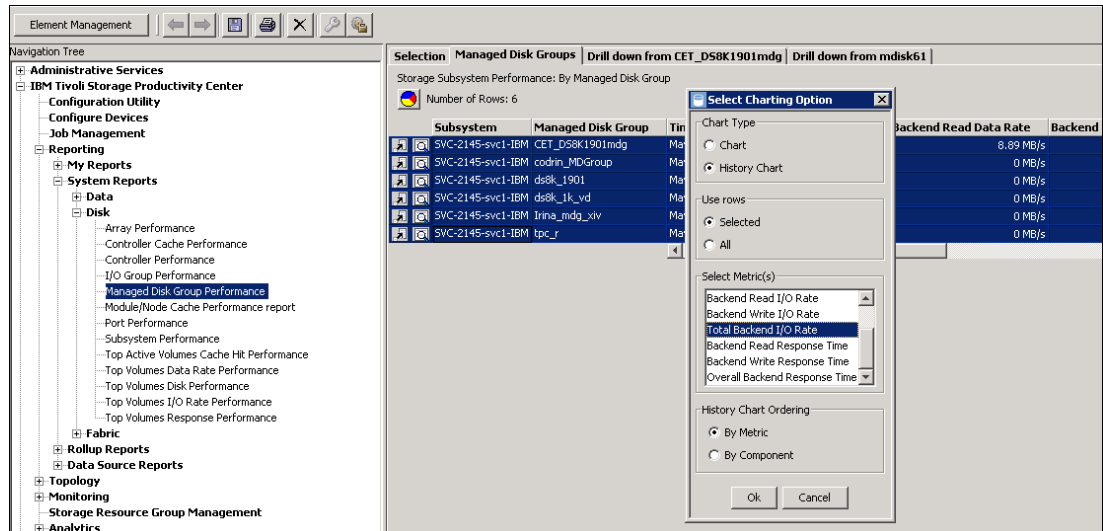


Figure 13-62 Managed disk group I/O rate selection for SAN Volume Controller

A chart is generated that is similar to the one that is shown in Figure 13-63.

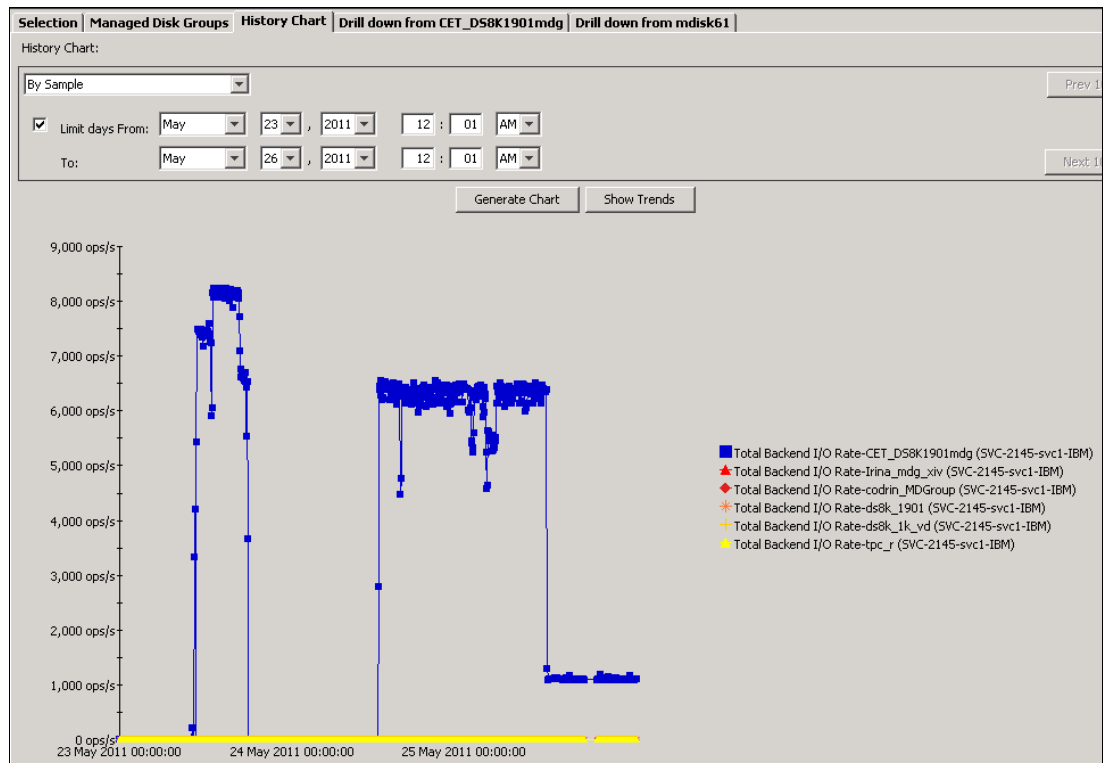


Figure 13-63 Managed Disk Group I/O rate report for SAN Volume Controller

When you review this general chart, you must understand that it reflects all I/O to the back-end storage from the MDisks that are included in this managed disk group. The key for this report is a general understanding of back-end I/O rate usage, not whether balance is outright. In this report, for the time frame that is specified at one point is a maximum of nearly 8200 IOPS.

Although the SAN Volume Controller and Storwize V7000, by default, stripe write and read I/Os across all MDisks, the striping is not through a RAID 0 type of stripe. Rather, because the VDisk is a concatenated volume, the striping that is injected by the SAN Volume Controller and Storwize V7000 is only in how you identify the extents to use when you create a VDisk. Until host I/O write actions fill up the first extent, the remaining extents in the block VDisk that is provided by SAN Volume Controller are not used. When you are looking at the Managed Disk Group Backend I/O report, that you might not see a balance of write activity, even for a single managed disk group.

Backend Response Time

Complete the following steps to return to the list of MDisks:

1. In the drill-down menu from the Select Charting Option click on Backend Read Response Time, as shown in Figure 13-64.

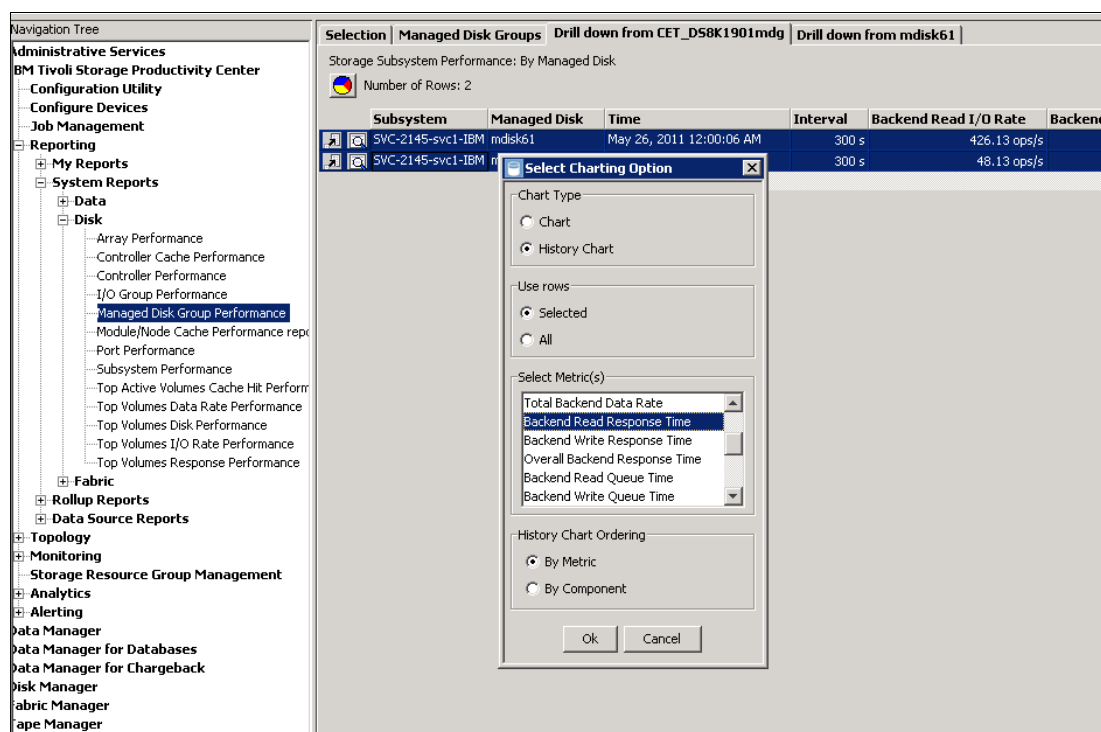


Figure 13-64 Backend Read Response Time for the managed disk

2. Select all of the managed disks entries and then click the pie chart icon (📊).
3. In the Select Charting Option window, select the **Backend Read Response time** metric. Then, click **OK**.

The chart that is shown in Figure 13-65 is generated.

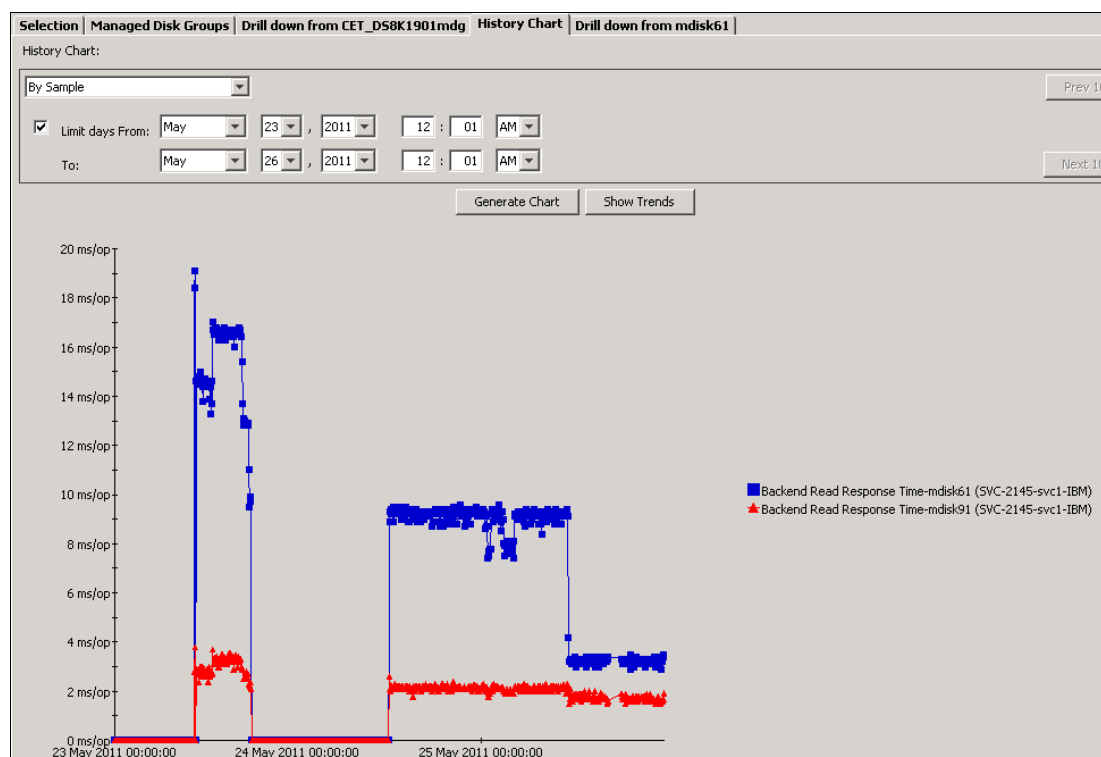


Figure 13-65 Backend response time

Guidelines for random read I/O

For random read I/O, the back-end rank (disk) read response time should seldom exceed 25 ms, unless the read hit ratio is near 99%. Backend Write Response Time is higher because of RAID 5 (or RAID 10) algorithms, but should seldom exceed 80 ms. Some time intervals exist when the response times exceed these guidelines.

Backend Data Rates

Back-end throughput and response time depend on the disk drive modules (DDMs) that used by the storage subsystem from which the LUN or volume was created. This time also depends on the specific RAID type in use. With this report, you can also check how MDisk workload is distributed.

Complete the following steps to obtain the back-end data rates:

1. In the drill down from CET_DS8K1901mdg tab, select all of the managed disks. Then, click the pie chart icon (📊).
2. In the Select Charting Option window (as shown in Figure 13-66 on page 418), select the **Backend Data Rates**. Then, click **OK**.

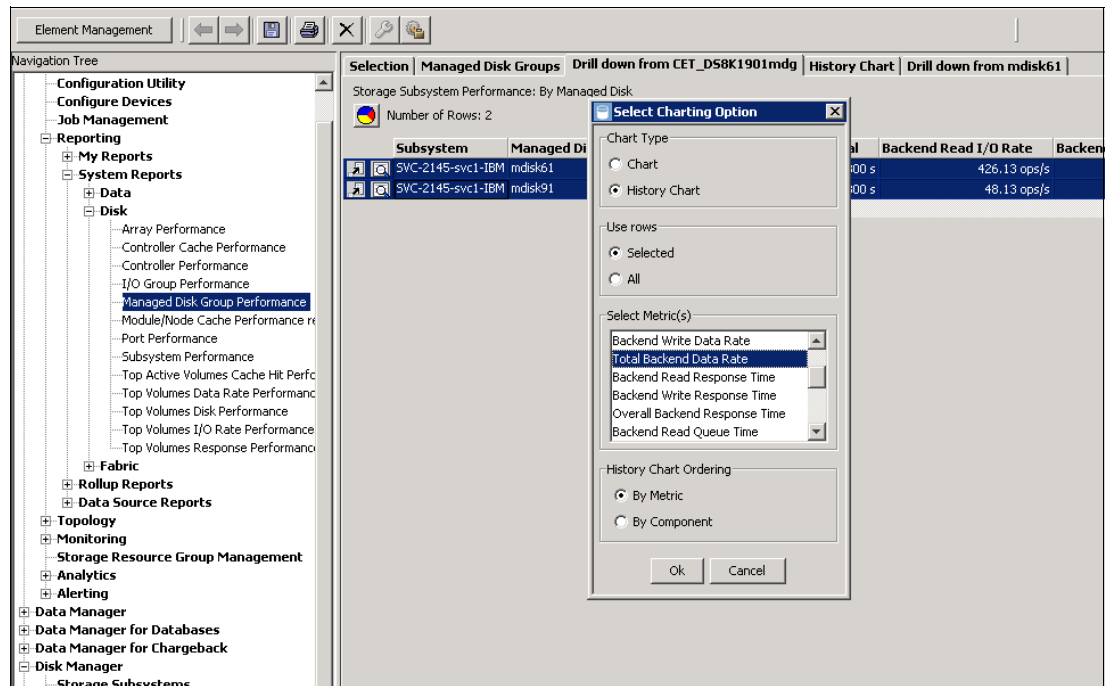


Figure 13-66 MDisk Backend Data Rates selection

Figure 13-67 shows the report that is generated, which in this case indicates that the workload is not balanced on MDisk.

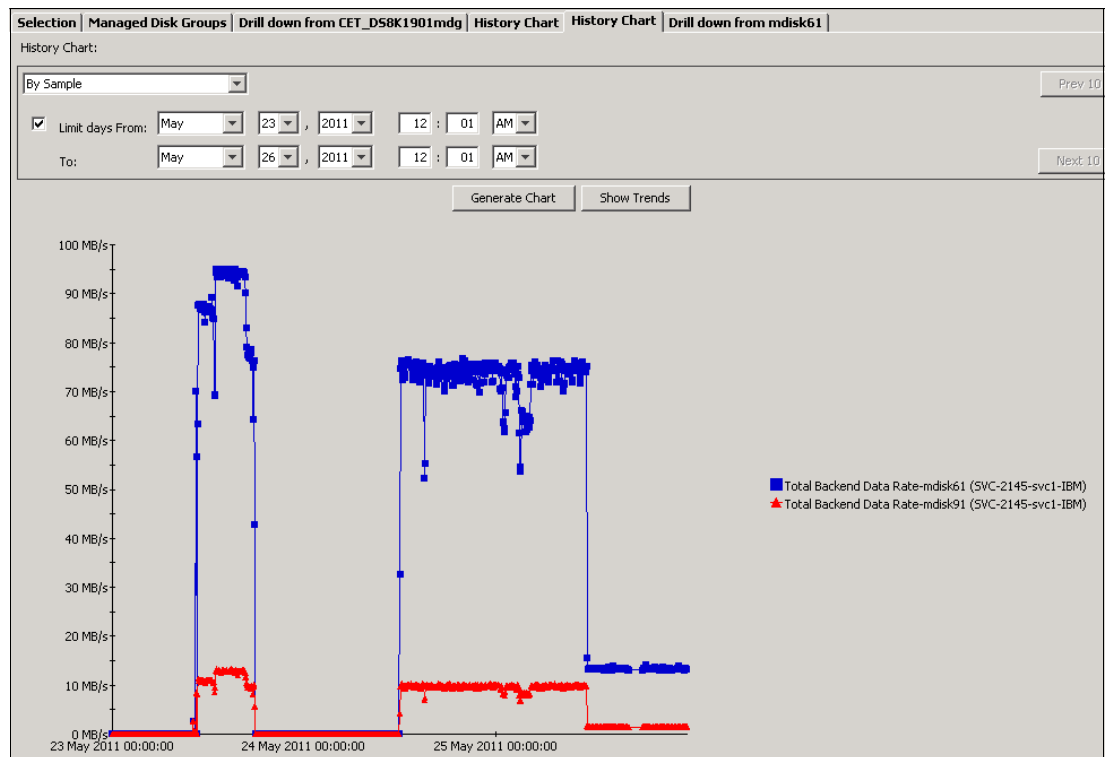


Figure 13-67 MDisk Backend Data Rates report

Viewing the Pool Performance report by using the web-based GUI

The performance of a storage system pool can be easily viewed in the Tivoli Storage Productivity Center 5.2 web-based GUI. To access the performance chart for the pools of a particular storage system, browse to that storage system, then click the **Pools** link that is under Internal Resources section, as shown in Figure 13-68.

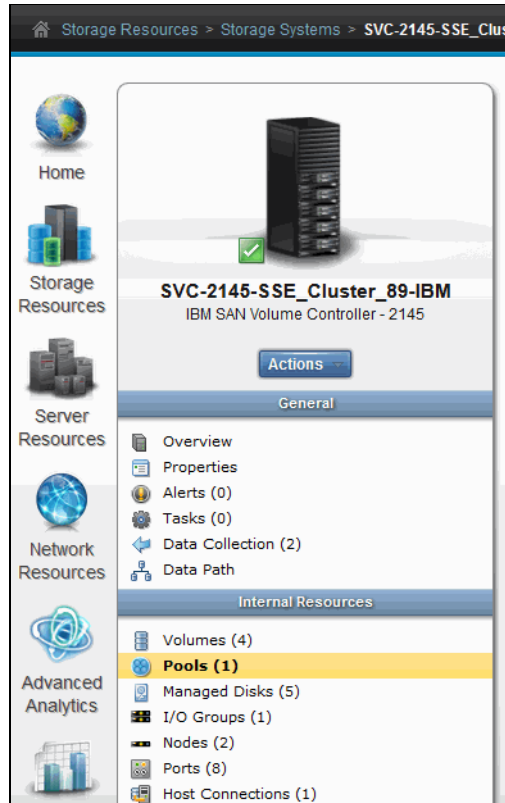


Figure 13-68 Selecting Pools

The Pools view for that storage system opens. Click the **Performance** tab to see a performance chart for these pools, as shown in Figure 13-69 on page 420.

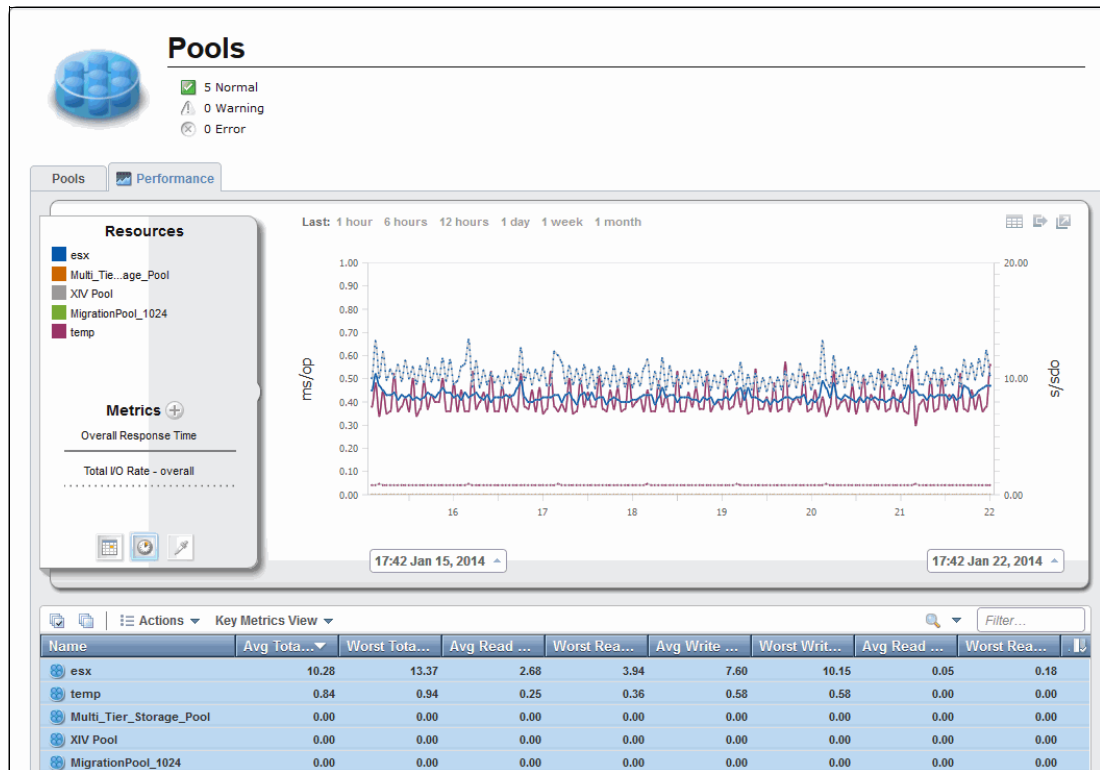


Figure 13-69 Pool performance view

The default performance view displays a chart of Total I/O Rate and Overall Response Time for all the pools. Performance charts are similar for all storage systems elements (that is, I/O Groups, Nodes, Volumes, and Pools.). For more information about how to modify the displayed metrics, the pools that are included in the chart, or the time frame for the chart, see “Accessing I/O Group Performance by using the web-based GUI” on page 391.

From this report, you can easily browse to see the statistics of the volumes that are supported by a pool or drill down to view the statistics for the individual MDisks that make up the pool by right-clicking the row for the wanted pool, as shown in Figure 13-70.

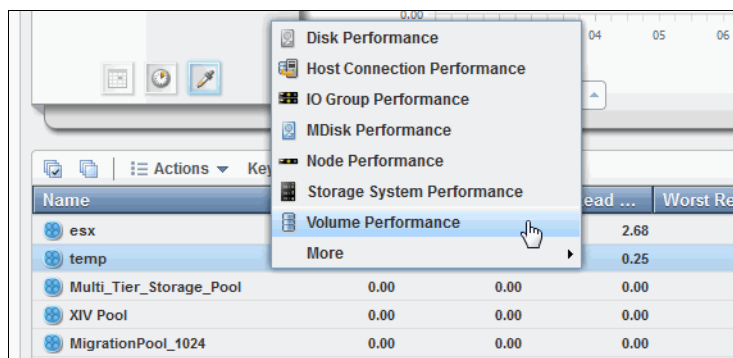


Figure 13-70 Accessing related performance information for a Pool

Back-end I/O Rate and Backend Response Time

As shown in Figure 13-69 on page 420, the default performance view displays Total I/O Rate and Overall Response Time, with which you can quickly analyze how the I/O workload is split between the managed disk groups to determine whether it is well-balanced. When you review this general chart, you must understand that it reflects all I/O to the back-end storage from the MDisks that are included in this managed disk group. The key for this report is a general understanding of back-end I/O rate usage, not whether balance is outright.

Although the SAN Volume Controller and Storwize V7000, by default, stripe write and read I/Os across all MDisks, the striping is not through a RAID 0 type of stripe. Rather, because the VDisk is a concatenated volume, the striping that is injected by the SAN Volume Controller and Storwize V7000 is only in how you identify the extents to use when you create a VDisk. Until host I/O write actions fill up the first extent, the remaining extents in the block VDisk that are provided by SAN Volume Controller are not used. When you are reviewing the Managed Disk Group Backend I/O report, you might not see a balance of write activity, even for a single managed disk group.

Guidelines for random read I/O

For random read I/O, the back-end rank (disk) read response time should seldom exceed 25 ms, unless the read hit ratio is near 99%. Backend Write Response Time is higher because of RAID 5 (or RAID 10) algorithms, but should seldom exceed 80 ms. Some time intervals exist when the response times exceed these guidelines.

Backend Data Rates

Back-end throughput and response time depend on the disk drive modules (DDMs) that are used by the storage subsystem from which the LUN or volume was created. This time also depends on the specific RAID type that is used. To display pool data rates in the chart, click the wanted metric from the available choices, as shown in Figure 13-71.

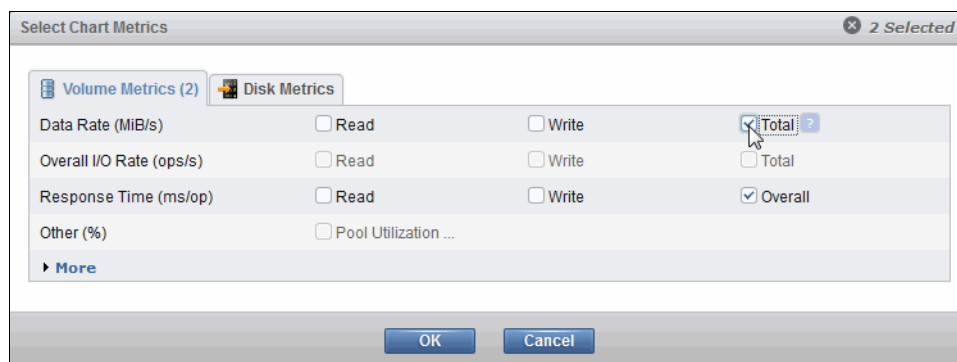


Figure 13-71 Displaying pool back-end data rates

The chart now shows the Total Data Rate instead of Overall I/O Rate, as shown in Figure 13-72.

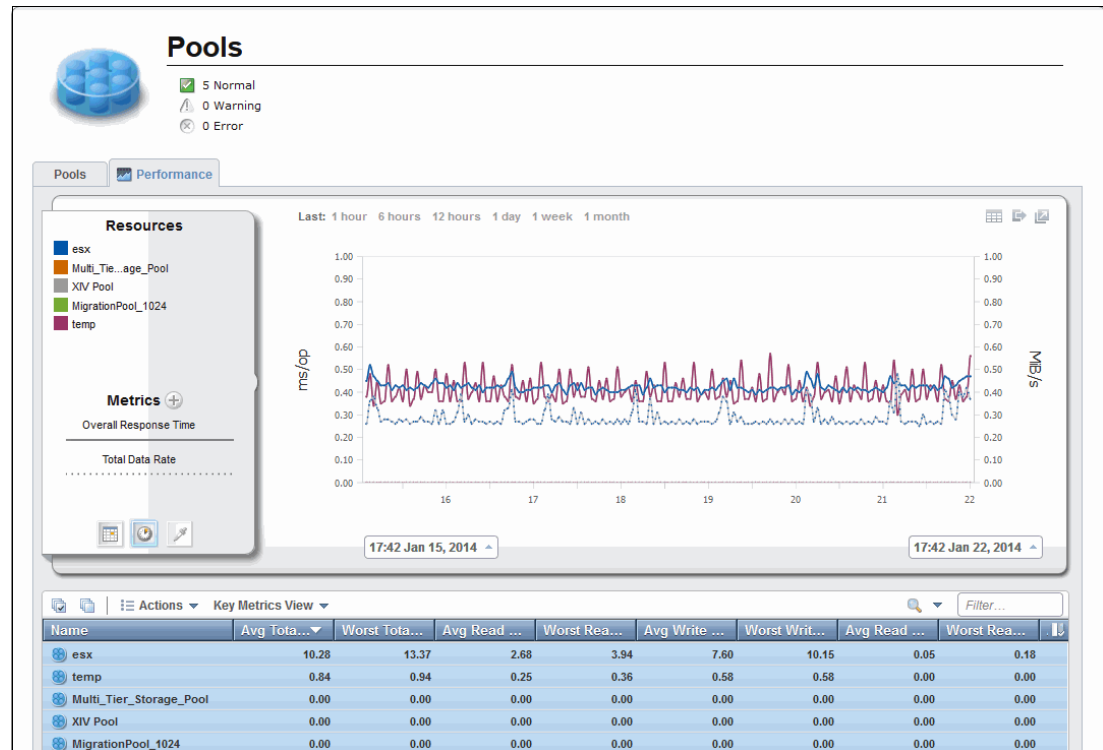


Figure 13-72 Pool total data rate

13.3.4 Top Volume Performance reports

By using the stand-alone GUI, Tivoli Storage Productivity Center provides the following reports on top volume performance for SAN Volume Controller and Storwize V7000:

- ▶ *Top Volume Cache Performance*, which is prioritized by the Total Cache Hits percentage (overall) metric.
- ▶ *Top Volumes Data Rate Performance*, which is prioritized by the Total Data Rate metric.
- ▶ *Top Volumes Disk Performance*, which is prioritized by the Disk to cache Transfer rate metric.
- ▶ *Top Volumes I/O Rate Performance*, which is prioritized by the Total I/O Rate (overall) metric.
- ▶ *Top Volume Response Performance*, which is prioritized by the Total Data Rate metric.

The volumes that are referred to in these reports correspond to the VDisks in SAN Volume Controller.

Important: The last collected performance data on volumes is used for the reports. The report creates a ranked list of volumes that are based on the metric that is used to prioritize the performance data. You can customize these reports according to the needs of your environment.

To limit these system reports to SAN Volume Controller subsystems, complete the following steps to specify a *filter*, as shown in Figure 13-73:

1. In the Selection tab, click **Filter**.
2. In the Edit Filter window, click **Add** to specify another condition to be met.

You must complete the filter process for all five reports.

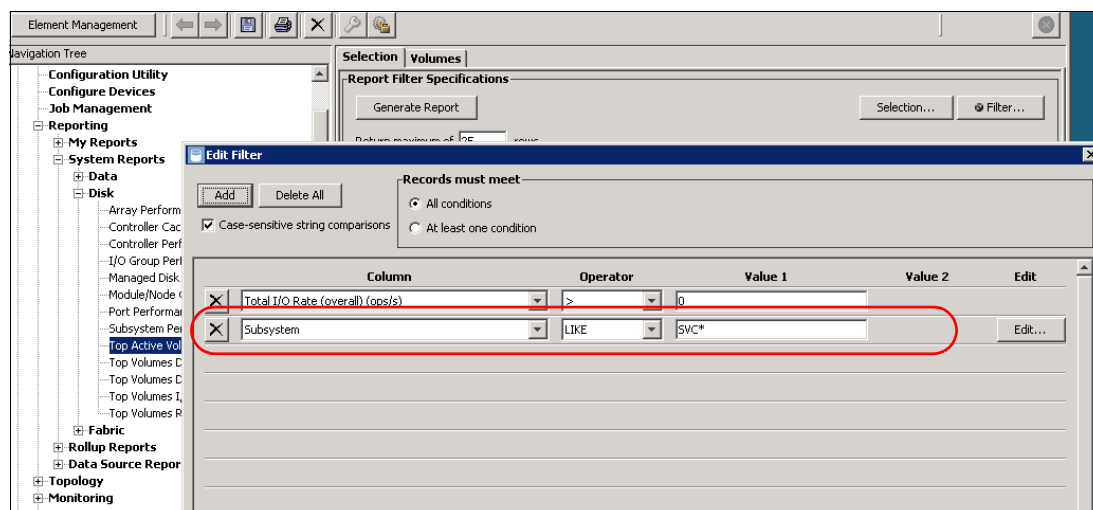


Figure 13-73 Specifying a filter for SAN Volume Controller Top Volume Performance Reports

Top Volumes Cache Performance report

The Top Volumes Cache Performance report shows the cache statistics for the top 25 volumes, which are prioritized by the Total Cache Hits percentage (overall) metric, as shown in Figure 13-74. This metric is the weighted average of read cache hits and write cache hits. The percentage of writes that is handled in cache should be 100% for most enterprise storage. An important metric is the percentage of reads during the sample interval that are found in cache.

The image shows a report titled 'Storage Subsystem Performance: By Volume' with 'Number of Rows: 5'. The report table is as follows:

Subsystem	Volume	Time	Interval	Read Cache Hits Percentage (overall)	Write
SVC-2145-svc1-IBM	tpcblade3_70002	May 26, 2011 12:00:06 AM	300 s	50.43 %	
SVC-2145-svc1-IBM	tpcblade3_70001	May 26, 2011 12:00:06 AM	300 s	50.25 %	
SVC-2145-svc1-IBM	tpcblade3_70004	May 26, 2011 12:00:06 AM	300 s	50.82 %	
SVC-2145-svc1-IBM	tpcblade3_70005	May 26, 2011 12:00:06 AM	300 s	52.02 %	
SVC-2145-svc1-IBM	tpcblade3_70003	May 26, 2011 12:00:06 AM	300 s	53.03 %	

Figure 13-74 Top Volumes Cache Hit performance report for SAN Volume Controller

More analysis of Read Hit percentages

Read Hit percentages can vary near 0% to near 100%. Any percentage below 50% is considered low, but many database applications show hit ratios below 30%. For low hit ratios, you need many ranks that provide a good back-end response time. It is difficult to predict whether more cache improves the hit ratio for a particular application. Hit ratios depend more on the application design and amount of data than on the size of cache (especially for Open System workloads). However, larger caches are always better than smaller caches. For high-hit ratios, the back-end ranks can be driven a little harder to higher utilizations.

Top Volumes Data Rate Performance

To determine the top five volumes with the highest total data rate during the last data collection time interval, click **IBM Tivoli Storage Productivity Center** **Reporting** **System Reports** **Disk**. Then, select **Top Volumes Data Rate Performance**.

By default, the scope of the report is not limited to a single storage subsystem. Tivoli Storage Productivity Center evaluates the data that is collected for all the storage subsystems that it has statistics for and creates the report with a list of 25 volumes that have the highest total data rate.

To limit the output, in the Selection tab (see Figure 13-75), enter 5 in the Return maximum of field. This figure is the maximum number of rows to be displayed on the report. Then, click **Generate Report**.

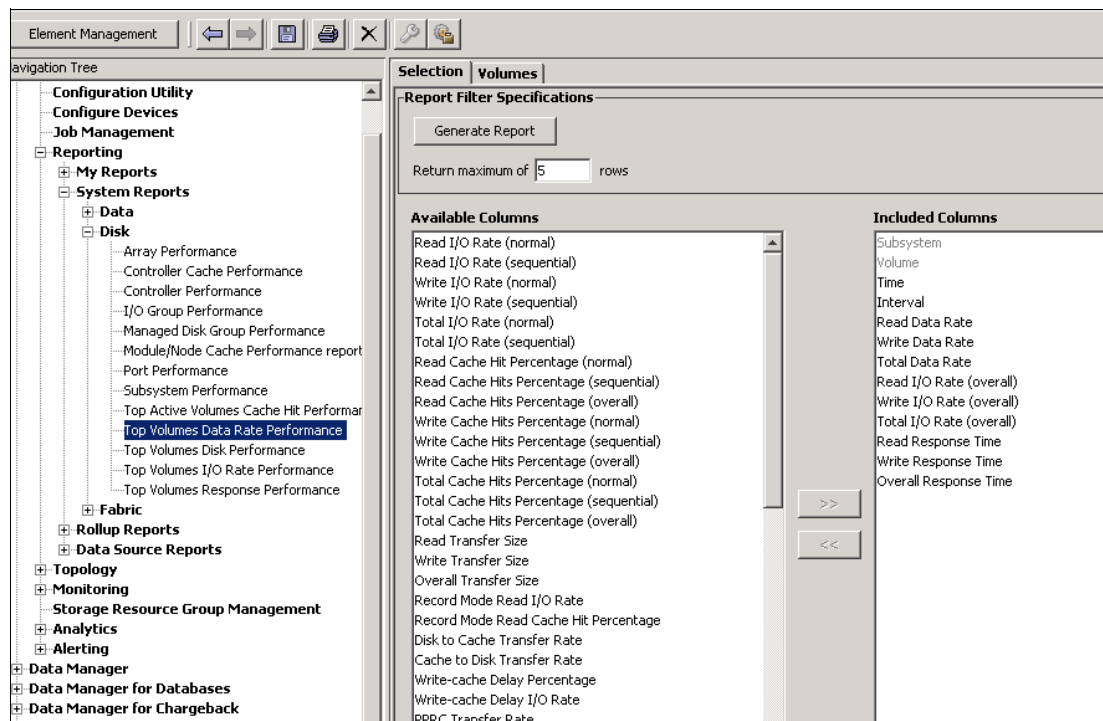


Figure 13-75 Top Volume Data Rate selection

Figure 13-76 shows the report that is generated. If this report is generated during the run time periods, the volumes have the highest total data rate and are listed on the report.

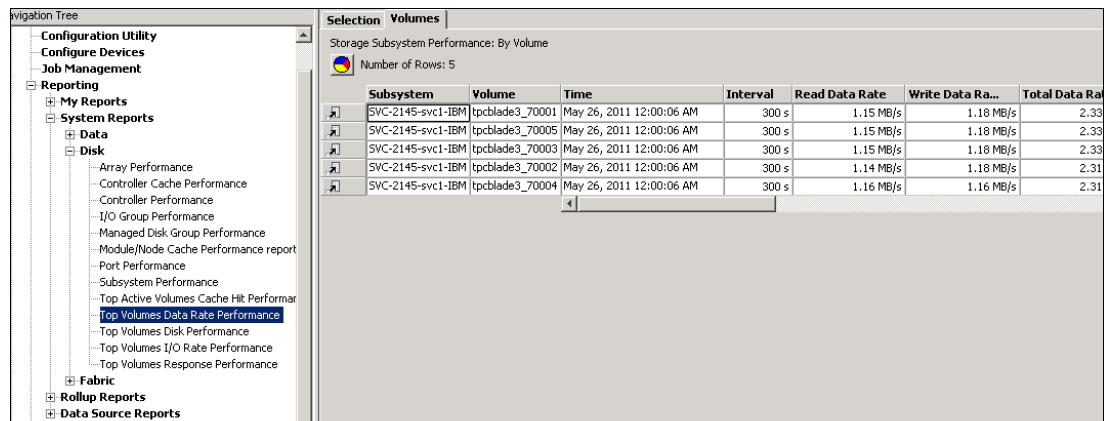


Figure 13-76 Top Volume Data Rate report for SAN Volume Controller

Top Volumes Disk Performance

The Top Volumes Disk Performance report includes many metrics about cache and volume-related information. Figure 13-77 shows the list of top 25 volumes that are prioritized by the Disk to Cache Transfer Rate metric. This metric indicates the average number of track transfers per second from disk to cache during the sample interval.

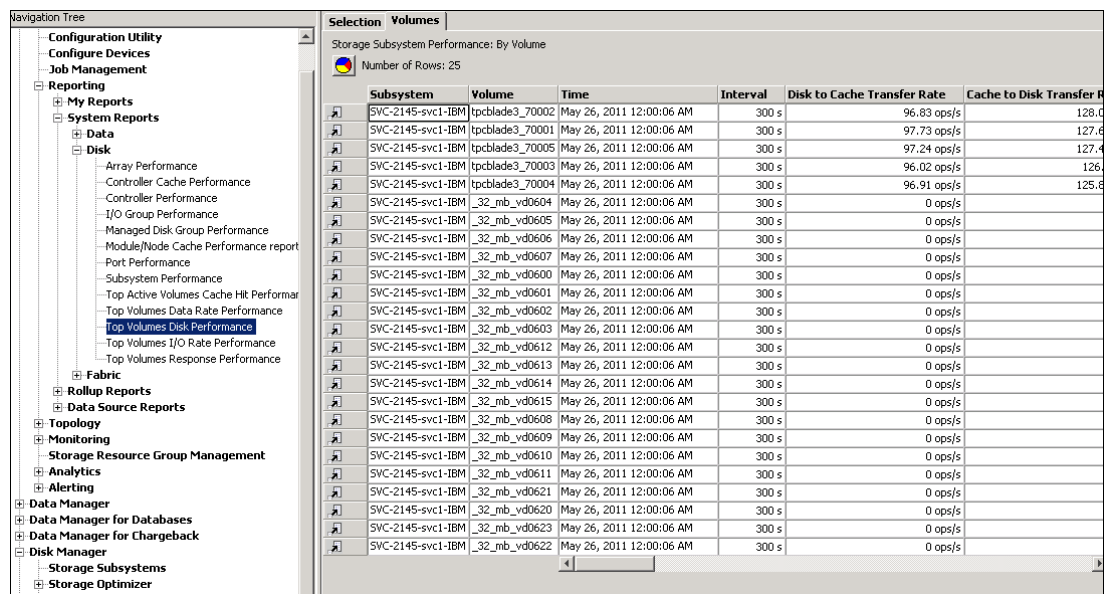


Figure 13-77 Top Volumes Disk Performance for SAN Volume Controller

Top Volumes I/O Rate Performance

The Top Volumes Data Rate Performance, Top Volumes I/O Rate Performance, and Top Volumes Response Performance reports include the same type of information. However, because of different sorting methods, other volumes might be included as the top volumes. Figure 13-78 on page 426 shows the top 25 volumes that are prioritized by the Total I/O Rate (overall) metrics.

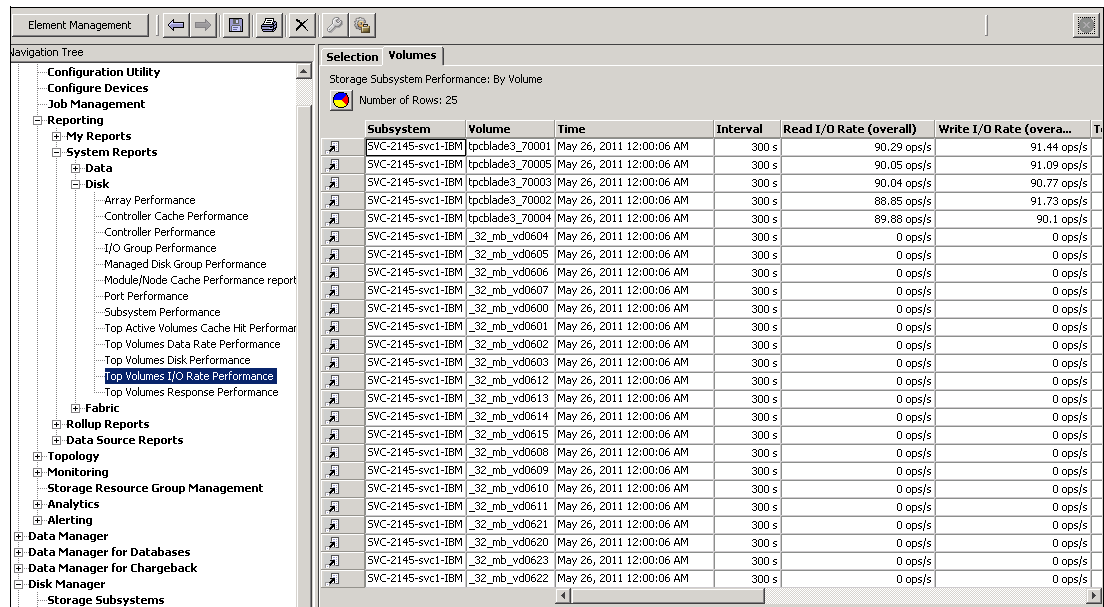


Figure 13-78 Top Volumes I/O Rate Performance for SAN Volume Controller

Guidelines for throughput storage

The throughput for storage volumes can range from fairly small numbers (1 - 10 I/O per second) to large values (more than 1000 I/O/second). The result depends on the nature of the application. I/O rates (throughput) that approach 1000 IOPS per volume occur because the volume is receiving good performance, usually from good cache behavior. Otherwise, it is not possible to perform so many IOPS to a volume.

Top Volumes Response Performance

The Top Volumes Data Rate Performance, Top Volumes I/O Rate Performance, and Top Volumes Response Performance reports include the same type of information. However, because of different sorting methods, other volumes might be included as the top volumes in this report. Figure 13-79 on page 427 shows the top 25 volumes that are prioritized by the Overall Response Time metrics.

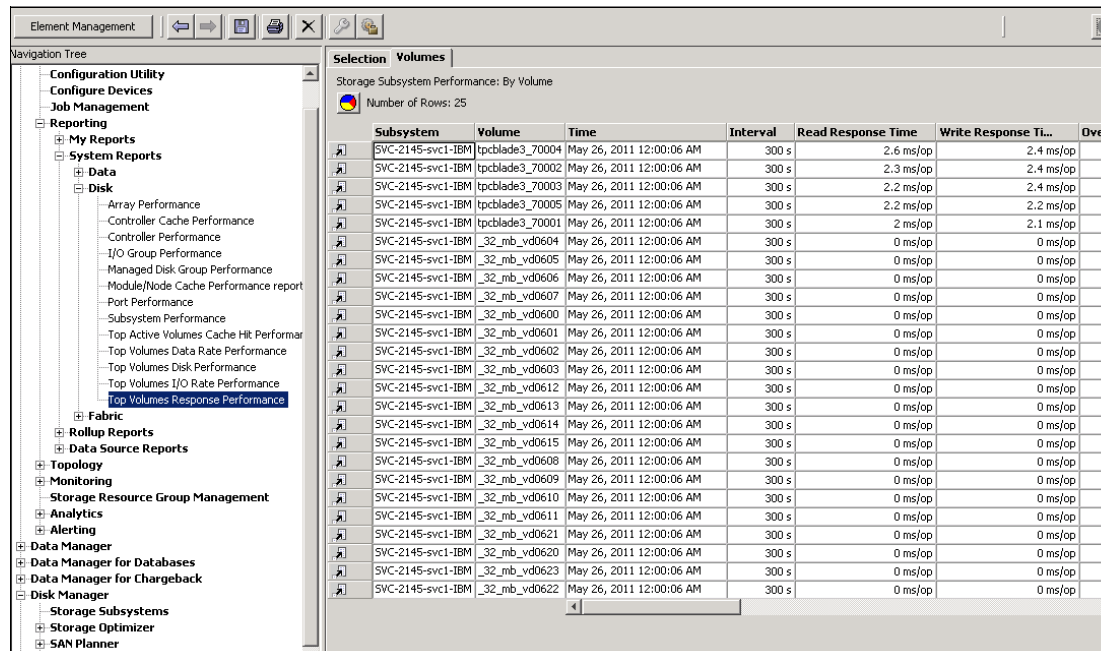


Figure 13-79 Top Volume Response Performance report for SAN Volume Controller

Guidelines about response times

Typical response time ranges are only slightly more predictable. In the absence of more information, you might often assume (and our performance models assume) that 10 ms is high. However, for a particular application, 10 ms might be too low or too high. Many OLTP environments require response times that are closer to 5 ms, where batch applications with large sequential transfers might accept a 20 ms response time. The appropriate value might also change between shifts or on the weekend. A response time of 5 ms might be required from 8 a.m. - 5 p.m., but 50 ms is acceptable near midnight. The result all depends on the customer and application.

Although the value of 10 ms is arbitrary, it is related to the nominal service time of current generation disk products. In crude terms, the service time of a disk is composed of a seek, a latency, and a data transfer. Current nominal seek times can range 4 - 8 ms, although in practice, many workloads do better than nominal. It is common for applications to experience one-third to one-half the nominal seek time. Latency is assumed to be half of the rotation time for the disk, and transfer time for typical applications is less than 1 ms. Therefore, it is reasonable to expect a service time of 5 - 7 ms for simple disk access. Under ordinary queuing assumptions, a disk operating at 50% utilization might have a wait time roughly equal to the service time. Therefore, a 10 - 14 ms response time for a disk is common and represents a reasonable goal for many applications.

For cached storage subsystems, you can expect to do as well or better than uncached disks, although that might be harder than you think. If many cache hits occur, the subsystem response time might be well below 5 ms. However, poor read hit ratios and busy disk arrays behind the cache drive up the average response time number.

With a high cache hit ratio, you can run the back-end storage ranks at higher utilizations than you might otherwise be satisfied with. Rather than 50% utilization of disks, you might push the disks in the ranks to 70% utilization, which might produce high rank response times that are averaged with the cache hits to produce acceptable average response times. Conversely, poor cache hit ratios require good response times from the back-end disk ranks to produce an acceptable overall average response time.

To simplify, you can assume that (front-end) response times probably need to be 5 - 15 ms. The rank (back-end) response times can usually operate at 20 - 25 ms, unless the hit ratio is poor. Back-end write response times can be even higher, generally up to 80 ms.

Important: All of these considerations are not valid for SSDs, where seek time and latency are not applicable. You can expect these disks to have much better performance and, therefore, a shorter response time (less than 4 ms).

For more information about creating a tailored report for your environment, see 13.5.3, “Top volumes response time and I/O rate performance reports” on page 455.

Creating Top Volume performance reports by using the web-based GUI

The Overview page for a Storage System in the Tivoli Storage Productivity Center 5.2 web-based GUI provides a quick method to view the top 10 most active volumes on that storage system as measured over the previous 24 hours. As shown in Figure 13-80, the right area of the Overview page is divided into four quadrants. Each of the quadrants automatically loads a chart displaying a specific storage system metric.



Figure 13-80 Overview page of a SAN Volume Controller

By default, the lower left quadrant loads a chart displaying the Most Active Volumes for this storage system. You can choose to have this chart displayed in any of the four quadrants by clicking the Title in that quadrant and selecting **Most Active Volumes**, as shown in Figure 13-79 on page 427.

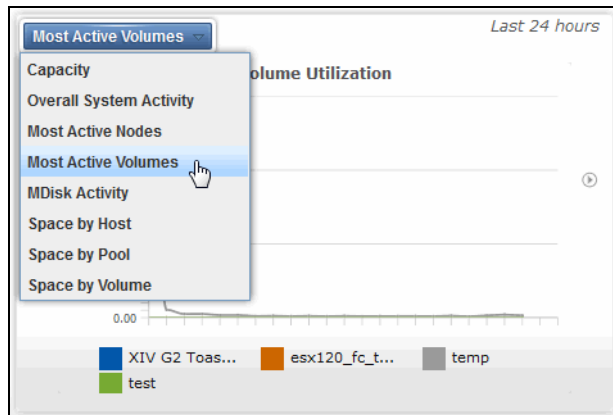


Figure 13-81 Changing the metrics that are displayed in the Storage System Overview page

In addition, you can change the metric by which the most active volumes are computed by toggling left or right by using the arrows, as shown in Figure 13-82.

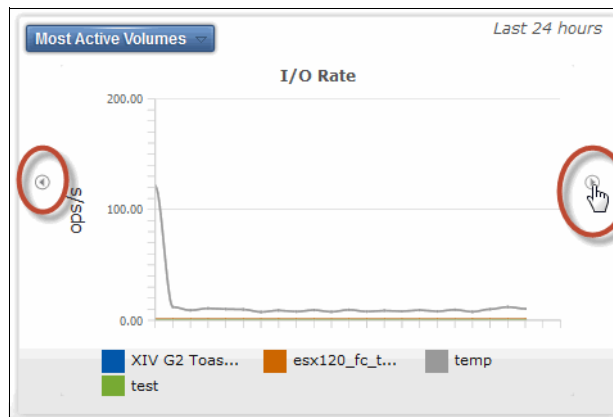


Figure 13-82 Changing the metric used to compute Most Active Volumes

You can quickly view the top 10 most active volumes in the storage system as measured by the following metrics:

- ▶ I/O Rate (ops/s)
- ▶ Data Rate (MB/s)
- ▶ Response Time (ms/op)
- ▶ Read Cache Hits (%)
- ▶ Volume Utilization (%)

Creating Most Active Volumes performance report by using the web-based GUI

The predefined Most Active Volumes report can be useful for identifying problem performance areas. You can quickly generate a report of the most active volumes within a system that is based on the criteria you choose, and allows for easy export of the report data.

To access the report, click **View predefined reports** in the Reporting section of the left-side navigation in the Tivoli Storage Productivity Center 5.2 web-based GUI, as shown in Figure 13-83 on page 430.



Figure 13-83 Accessing predefined reports

A window opens in which you can log in to the Cognos® reporting component of Tivoli Storage Productivity Center 5.2, as shown in Figure 13-84.

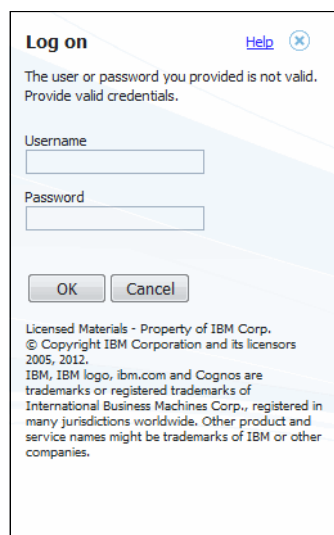
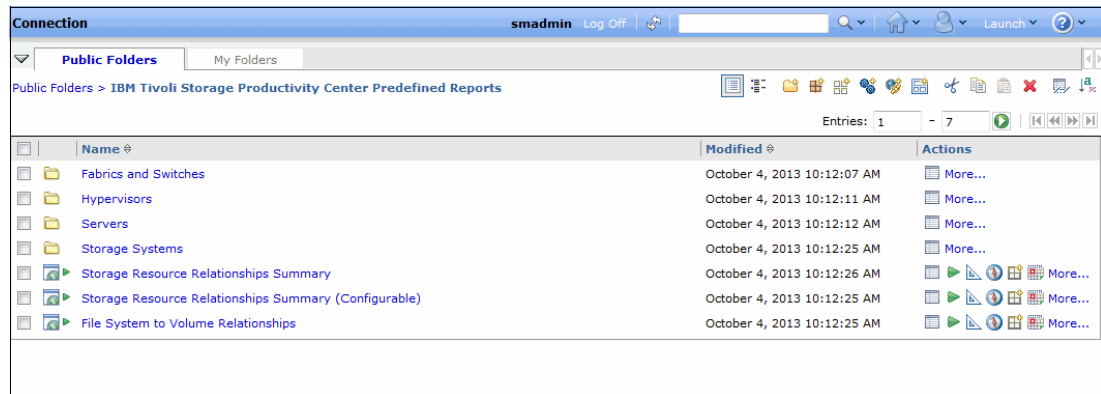


Figure 13-84 Logging in to Cognos Reporting

Note: Access to this Reporting functionality requires the installation of the necessary Tivoli Common Reporting/Cognos components.

After you are logged in, you see a list of available reports, as shown in Figure 13-85.

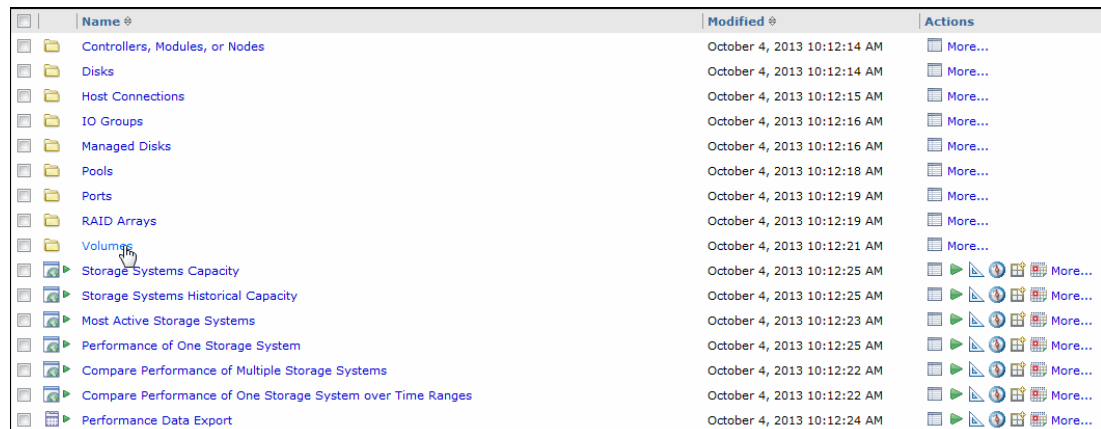


The screenshot shows the 'Public Folders' view of the Tivoli Storage Productivity Center. The breadcrumb path is 'Public Folders > IBM Tivoli Storage Productivity Center Predefined Reports'. A table lists the following reports:

Name	Modified	Actions
Fabrics and Switches	October 4, 2013 10:12:07 AM	More...
Hypervisors	October 4, 2013 10:12:11 AM	More...
Servers	October 4, 2013 10:12:12 AM	More...
Storage Systems	October 4, 2013 10:12:25 AM	More...
Storage Resource Relationships Summary	October 4, 2013 10:12:26 AM	More...
Storage Resource Relationships Summary (Configurable)	October 4, 2013 10:12:25 AM	More...
File System to Volume Relationships	October 4, 2013 10:12:25 AM	More...

Figure 13-85 Top-level view of the available predefined reports

To access the Most Active Volumes report, click **Storage Systems**. In the next window, click **Volumes**, as shown in Figure 13-86.

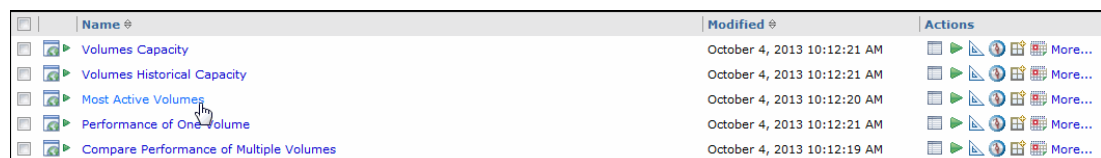


The screenshot shows the 'Storage Systems' view. The breadcrumb path is 'Public Folders > IBM Tivoli Storage Productivity Center Predefined Reports > Storage Systems'. A table lists the following reports:

Name	Modified	Actions
Controllers, Modules, or Nodes	October 4, 2013 10:12:14 AM	More...
Disks	October 4, 2013 10:12:14 AM	More...
Host Connections	October 4, 2013 10:12:15 AM	More...
IO Groups	October 4, 2013 10:12:16 AM	More...
Managed Disks	October 4, 2013 10:12:16 AM	More...
Pools	October 4, 2013 10:12:18 AM	More...
Ports	October 4, 2013 10:12:19 AM	More...
RAID Arrays	October 4, 2013 10:12:19 AM	More...
Volumes	October 4, 2013 10:12:21 AM	More...
Storage Systems Capacity	October 4, 2013 10:12:25 AM	More...
Storage Systems Historical Capacity	October 4, 2013 10:12:25 AM	More...
Most Active Storage Systems	October 4, 2013 10:12:23 AM	More...
Performance of One Storage System	October 4, 2013 10:12:25 AM	More...
Compare Performance of Multiple Storage Systems	October 4, 2013 10:12:22 AM	More...
Compare Performance of One Storage System over Time Ranges	October 4, 2013 10:12:22 AM	More...
Performance Data Export	October 4, 2013 10:12:24 AM	More...

Figure 13-86 View of the available predefined Storage System reports

The window then displays the available predefined reports for Volumes, as shown in Figure 13-87.



The screenshot shows the 'Volumes' view. The breadcrumb path is 'Public Folders > IBM Tivoli Storage Productivity Center Predefined Reports > Storage Systems > Volumes'. A table lists the following reports:

Name	Modified	Actions
Volumes Capacity	October 4, 2013 10:12:21 AM	More...
Volumes Historical Capacity	October 4, 2013 10:12:21 AM	More...
Most Active Volumes	October 4, 2013 10:12:20 AM	More...
Performance of One Volume	October 4, 2013 10:12:21 AM	More...
Compare Performance of Multiple Volumes	October 4, 2013 10:12:19 AM	More...

Figure 13-87 View of the available predefined Volume reports

Click the link for the **Most Active Volumes** report. You are prompted to select the report criteria, as shown in Figure 13-88 on page 432.

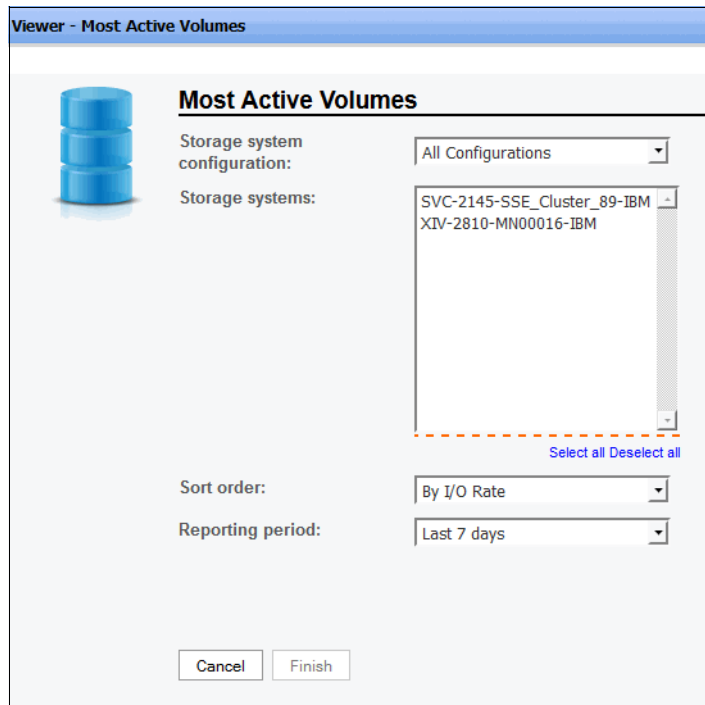


Figure 13-88 Selecting the criteria for the Most Active Volumes report

By using the report criteria window, you can select the following criteria:

- ▶ Wanted storage system
- ▶ Metric to use for sorting the report results. The following values are available:
 - I/O Rate (ops/s)
 - Data Rate (MB/s)
 - Response Time (ms/op)
 - Read Cache Hits (%)
 - Volume Utilization (%)
- ▶ The reporting period. You can select from various predefined time periods, or you can define a custom time period.

Select the criteria appropriate to your scenario and click **Finish** to generate the report. The report results are then loaded in the next window, as shown in Figure 13-89 on page 433.

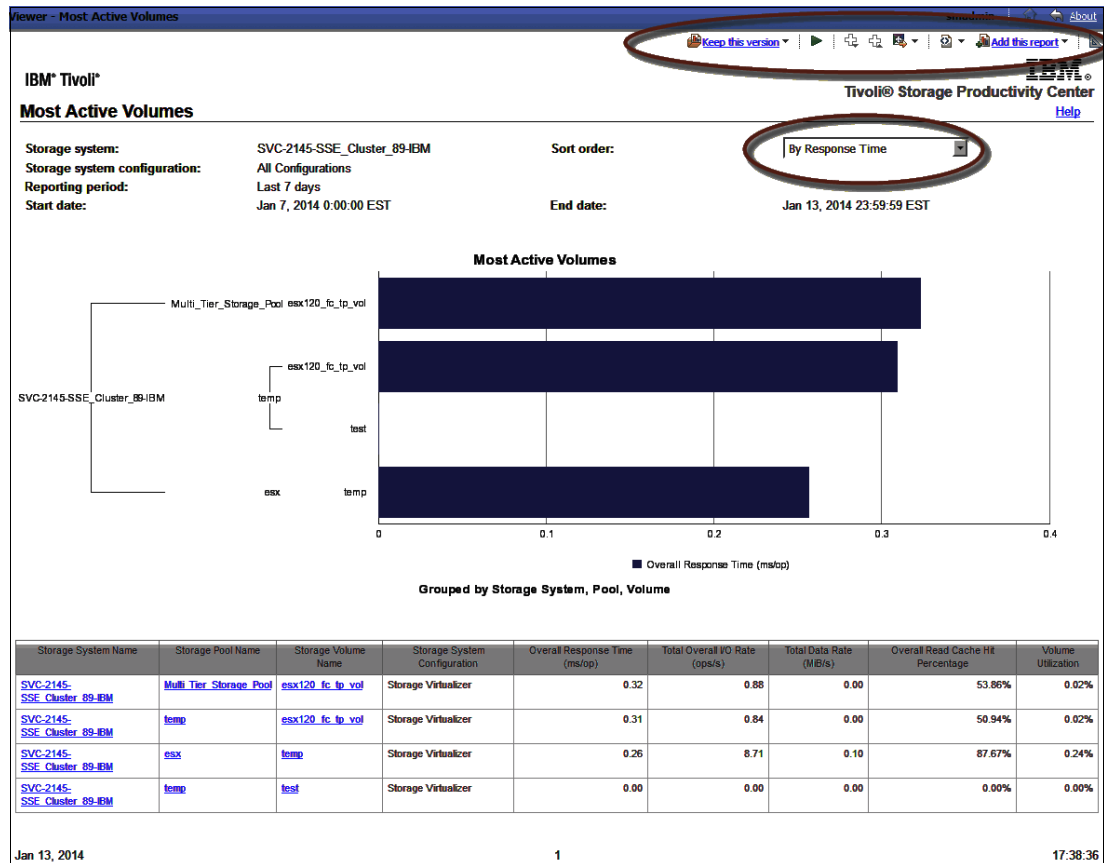


Figure 13-89 Report results

After the report is generated, you can select an alternative metric from the drop-down menu and the report automatically reloads. Additionally, you can use the menu options at the top of the page to save, email, or export the report.

13.3.5 Port Performance reports for SAN Volume Controller and Storwize V7000

The SAN Volume Controller and Storwize V7000 Port Performance reports help you understand the SAN Volume Controller and Storwize V7000 effect on the fabric. They also provide an indication of the following traffic:

- ▶ SAN Volume Controller (or Storwize V7000) and hosts that receive storage
- ▶ SAN Volume Controller (or Storwize V7000) and back-end storage
- ▶ Nodes in the SAN Volume Controller (or Storwize V7000) cluster

These reports can help you understand whether the fabric might be a performance bottleneck and whether upgrading the fabric can lead to performance improvement.

The Port Performance report summarizes the various send, receive, and total port I/O rates and data rates. To access this report, click **IBM Tivoli Storage Productivity Center** **My Reports** **System Reports** **Disk**, and select **Port Performance**. To display only SAN Volume Controller and Storwize V7000 ports, click **Filter**. Then, produce a report for all the volumes that belong to SAN Volume Controller or Storwize V7000 subsystems, as shown in Figure 13-90 on page 434.

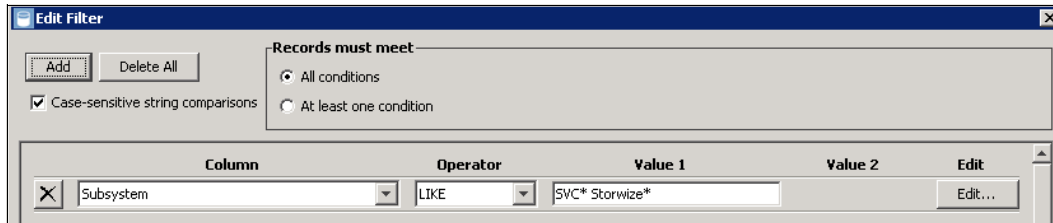


Figure 13-90 Subsystem filter for the Port Performance report

A separate row is generated for the ports of each subsystem. The information that is displayed in each row reflects that data that was last collected for the port.

The Time column (not shown in Figure 13-90) shows the last collection time, which might be different for the various subsystem ports. Not all of the metrics in the Port Performance report are applicable for all ports. For example, the Port Send Utilization percentage, Port Receive Utilization Percentage, and Overall Port Utilization percentage data are not available on SAN Volume Controller or Storwize V7000 ports.

The value N/A is displayed when data is not available, as shown in Figure 13-91. By clicking **Total Port I/O Rate**, you see a prioritized list by I/O rate.

Selection Ports		Storage Subsystem Performance: By Port					
 Number of Rows: 16							
Subsystem	Port	Port Send Utilization Percentage	Port Receive Utilization Percentage	Overall Port Utilization Percentage	Port Send I/O Rate	Port Receive I/O Rate	Total
 Storwize V7000-2076-ford1_tbird-IBM	5005076802100099	N/A	N/A	N/A	20,716.33 ops/s	22.26 ops/s	
 Storwize V7000-2076-ford1_tbird-IBM	500507680210009A	N/A	N/A	N/A	208,379.98 ops/s	1.58 ops/s	
 Storwize V7000-2076-ford1_tbird-IBM	5005076802200099	N/A	N/A	N/A	193,239.5 ops/s	1.62 ops/s	
 Storwize V7000-2076-ford1_tbird-IBM	500507680220009A	N/A	N/A	N/A	23,761.11 ops/s	22.32 ops/s	
 Storwize V7000-2076-ford1_tbird-IBM	5005076802300099	N/A	N/A	N/A	0 ops/s	0 ops/s	
 Storwize V7000-2076-ford1_tbird-IBM	500507680230009A	N/A	N/A	N/A	0 ops/s	0 ops/s	
 Storwize V7000-2076-ford1_tbird-IBM	5005076802400099	N/A	N/A	N/A	0 ops/s	0 ops/s	
 Storwize V7000-2076-ford1_tbird-IBM	500507680240009A	N/A	N/A	N/A	0 ops/s	0 ops/s	
 SVC-2145-svc1-IBM	5005076801105EB1	N/A	N/A	N/A	21.33 ops/s	13,743.65 ops/s	
 SVC-2145-svc1-IBM	5005076801106140	N/A	N/A	N/A	21.28 ops/s	21.44 ops/s	
 SVC-2145-svc1-IBM	5005076801205EB1	N/A	N/A	N/A	21.29 ops/s	13,743.56 ops/s	
 SVC-2145-svc1-IBM	5005076801206140	N/A	N/A	N/A	21.37 ops/s	21.33 ops/s	
 SVC-2145-svc1-IBM	5005076801305EB1	N/A	N/A	N/A	0 ops/s	0 ops/s	
 SVC-2145-svc1-IBM	5005076801306140	N/A	N/A	N/A	0 ops/s	0 ops/s	
 SVC-2145-svc1-IBM	5005076801405EB1	N/A	N/A	N/A	42.52 ops/s	13,765.63 ops/s	
 SVC-2145-svc1-IBM	5005076801406140	N/A	N/A	N/A	42.57 ops/s	31,550.39 ops/s	

Figure 13-91 Port Performance report

You can now verify whether the data rates to the back-end ports (as shown in the report) are beyond the normal rates that are expected for the speed of your fiber links, as shown in Figure 13-92 on page 435. This report often is generated to support problem determination, capacity management, or SLA reviews. Based on the 8 Gb per second fabric, these rates are well-below the throughput capability of this fabric. Therefore, the fabric is not a bottleneck here.

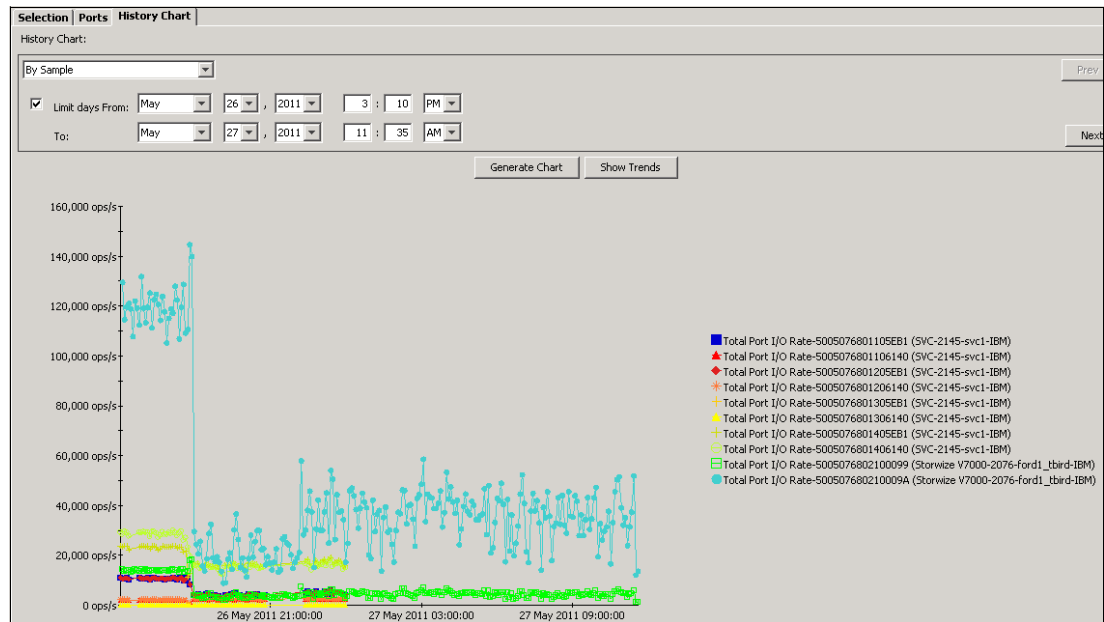


Figure 13-92 Port I/O Rate report for SAN Volume Controller and Storwize V7000

Next, select the **Port Send Data Rate** and **Port Receive Data Rate** metrics to generate another historical chart, as shown in Figure 13-93. This chart confirms the unbalanced workload for one port.



Figure 13-93 SAN Volume Controller and Storwize V7000 Port Data Rate report

To investigate further by using the Port Performance report, return to the I/O group performances report and complete the following steps:

1. Click **IBM Tivoli Storage Productivity Center** **My Reports** **System Reports** **Disk**. Select **I/O group Performance**.

- Click the magnifying glass icon (🔍) to drill down to the node level. As shown in Figure 13-94, we chose node 1 of the SAN Volume Controller subsystem. Click the pie chart icon (📊).

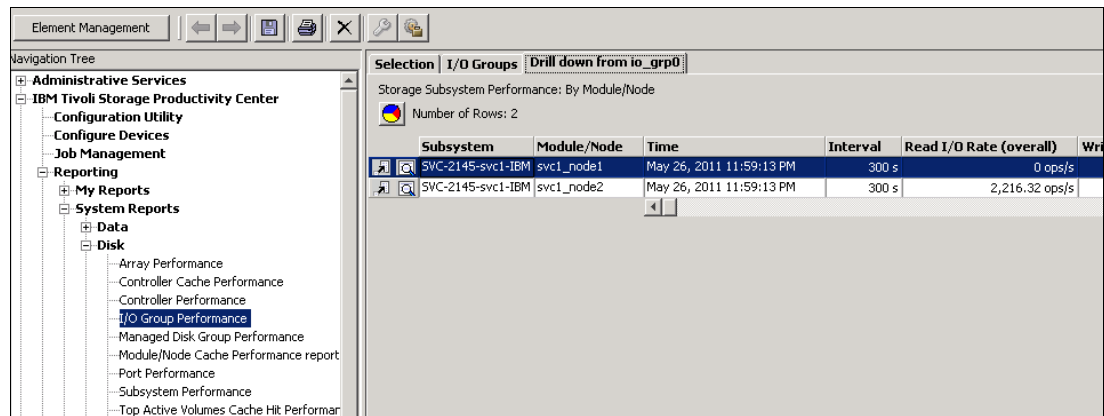


Figure 13-94 SAN Volume Controller node port selection

- In the Select Charting Option window (see Figure 13-95), select **Port to Local Node Send Queue Time**, **Port to Local Node Receive Queue Time**, **Port to Local Node Receive Response Time**, and **Port to Local Node Send Response Time**. Then, click **OK**.

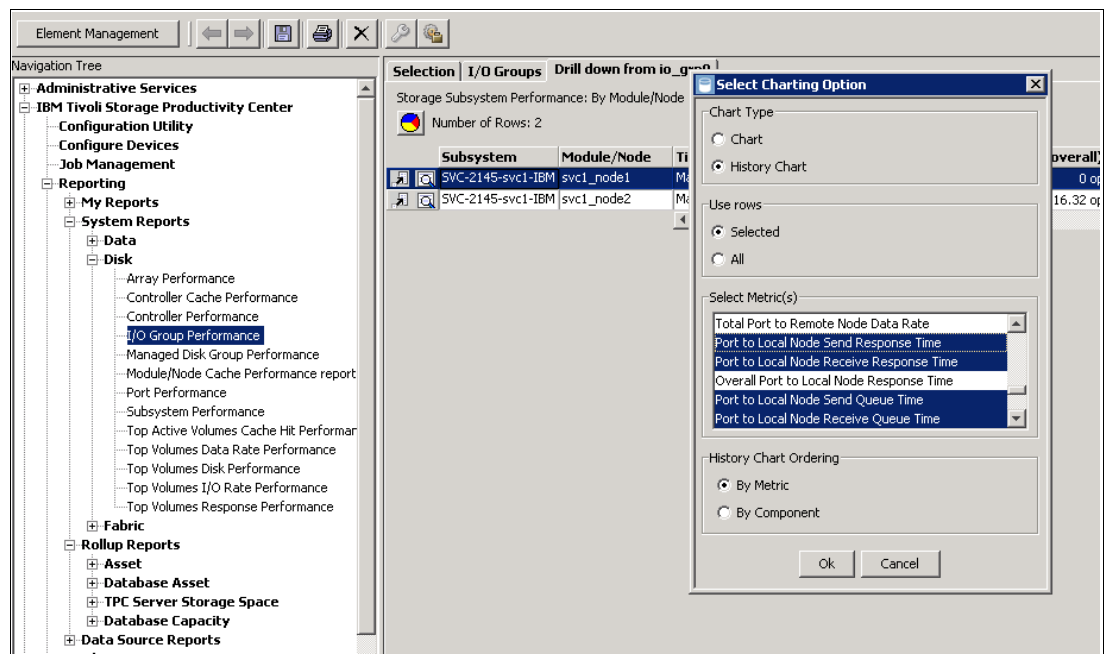


Figure 13-95 SAN Volume Controller Node port selection queue time

Review the port rates between SAN Volume Controller nodes, hosts, and disk storage controllers. Figure 13-96 on page 437 shows low queue and response times, which indicates that the nodes do not have a problem communicating with each other.

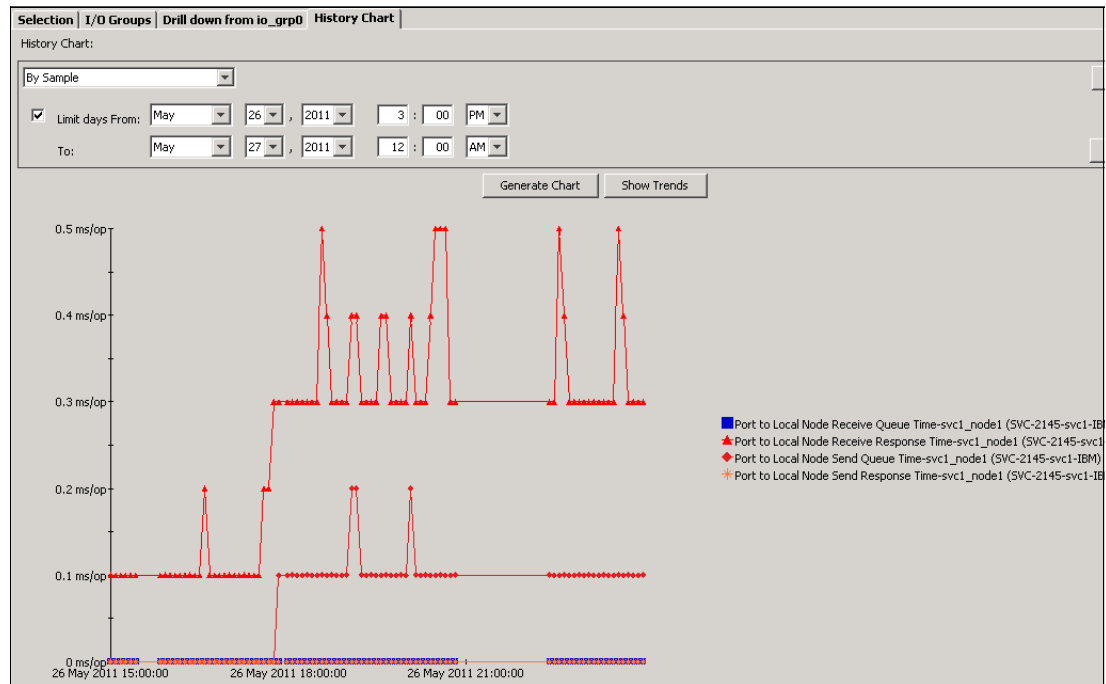


Figure 13-96 SAN Volume Controller Node ports report

If this report shows high queue and response times, the write activity is affected because each node communicates to each other node over the fabric.

Unusually high numbers in this report indicate the following issues:

- ▶ A SAN Volume Controller (or Storwize V7000) node or port problem (unlikely)
- ▶ Fabric switch congestion (more likely)
- ▶ Faulty fabric ports or cables (most likely)

Guidelines for the data range values

Based on the nominal speed of each FC port, which can be 4 Gb, 8 Gb or more, do not exceed a range of 50% - 60% of that value as the data rate. For example, an 8 Gb port can reach a maximum theoretical data rate of around 800 MBps. Therefore, you must generate an alert when it is more than 400 MBps.

Identifying overused ports

You can verify whether you have any host adapter or SAN Volume Controller (or Storwize V7000) ports that are heavily loaded when the workload is balanced between the specific ports of a subsystem that your application server is using. If you identify an imbalance, review whether the imbalance is a problem. If an imbalance occurs and the response times and data rate are acceptable, the only action that might be required is to note the effect.

If a problem occurs at the application level, review the volumes that are using these ports, and review their I/O and data rates to determine whether redistribution is required.

To support this review, you can generate a port chart. By using the date range, you can specify the specific time frame when you know the I/O and data was in place. Then, select the **Total Port I/O Rate** metric on all of SAN Volume Controller (or Storwize V7000) ports, or the specific Host Adapter ports in question. The graphical report that is shown in Figure 13-97 on page 438 refers to all the Storwize ports.



Figure 13-97 SAN Volume Controller Port I/O Send/Receive Rate

After you have the I/O rate review chart, generate a data rate chart for the same time frame to support a review of your high availability ports for this application.

Then, generate another historical chart by choosing the **Total Port Data Rate** metric (see Figure 13-98) that confirms the unbalanced workload for one port that is shown in the report in Figure 13-97.



Figure 13-98 Port Data Rate report

Guidelines for the data range values

According to the nominal speed of each FC port, which can be 4 Gb, 8 Gb, or more, do not exceed a range of 50% - 60% of that value as the data rate. For example, an 8 Gb port can reach a maximum theoretical data rate of around 800 MBps. Therefore, you must generate an alert when it is more than 400 MBps.

13.4 Reports for fabric and switches

Fabric and switches provide metrics that you cannot create in the top 10 reports list. Tivoli Storage Productivity Center provides the most important metrics to create reports against them. Figure 13-99 shows a list of system reports that are available for your Fabric.

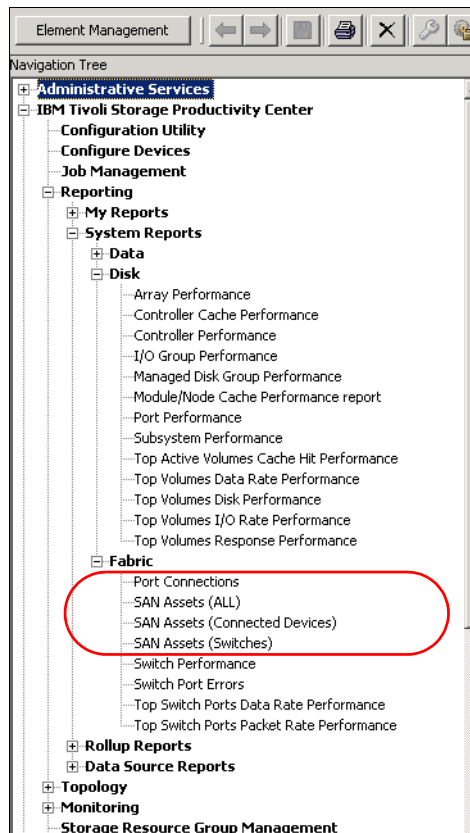


Figure 13-99 Fabric list of reports

13.4.1 Switches reports

The first four reports that are shown in Figure 13-99 provide Asset information in a tabular view. You can see the same information in a graphic view by using the Topology Viewer, which is the preferred method for viewing the information.

Tip: Rather than using a specific report to monitor Switch Port Errors, use the Constraint Violation report. By setting an alert for the number of errors at the switch port level, the Constraint Violation report becomes a direct tool to monitor the errors in your fabric. For more information about Constraint Violation reports, see *SAN Storage Performance Management Using Tivoli Storage Productivity Center*, SG24-7364.

13.4.2 Switch Port Data Rate Performance

For the Top report, analyze the Switch Ports Data Rate report. The Total Port Data Rate report shows the average number of megabytes (2^{20} bytes) per second that were transferred for send and receive operations for a particular port during the sample interval.

Complete the following steps to access this report:

1. Click **IBM Tivoli Storage Productivity Center** **Reporting** **System Reports** **Fabric**, and then select **Top Switch Ports Data Rates performance**.
2. Click the pie chart icon ().
3. In the Select Charting Option window (see Figure 13-100), select **Total Port Data Rate**, and then click **OK**.

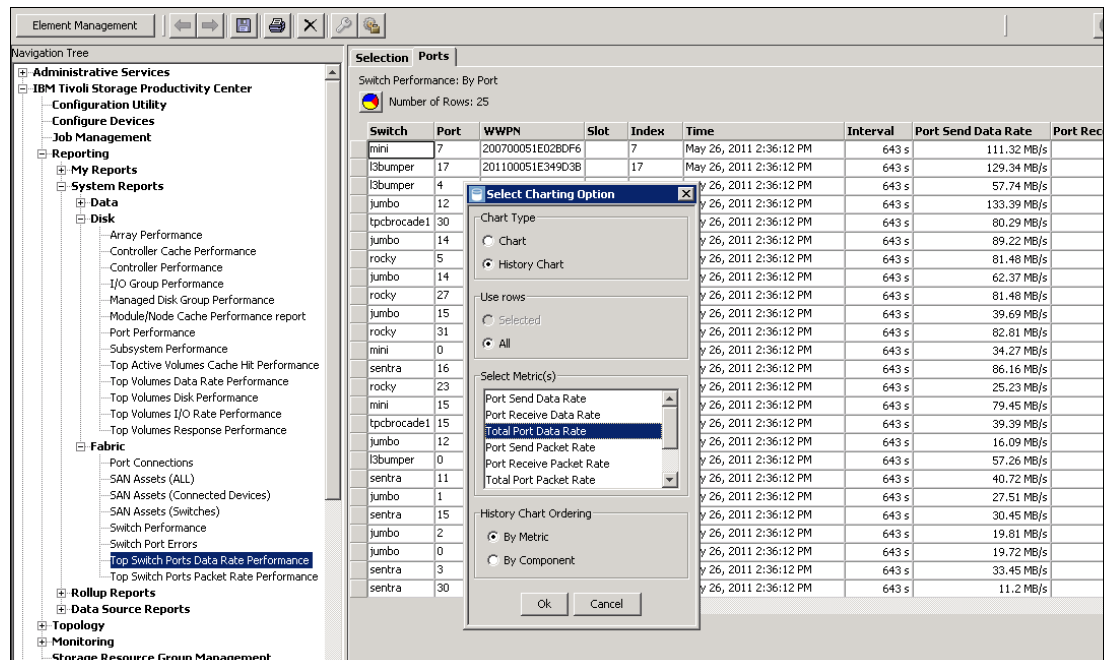


Figure 13-100 Port Data Rate selection for the Fabric report

You now see a chart similar to the example that is shown in Figure 13-101 on page 441. Port Data Rates do not reach a warning level, in this case, knowing that FC Port speed is 8 Gbps.



Figure 13-101 Port Data Rate report

Monitoring whether switch ports are overloaded

Use this report to monitor whether some switch ports are overloaded. According to the FC port nominal speed (2 Gb, 4 Gb, or more) as shown in Table 13-1, you must establish the maximum workload that a switch port can reach. Do not exceed 50% - 70%.

Table 13-1 Switch port data rates

FC port speed Gbps	FC port speed MBps	Port data rate threshold
1 Gbps	100 MBps	50 MBps
2 Gbps	200 MBps	100 MBps
4 Gbps	400 MBps	200 MBps
8 Gbps	800 MBps	400 MBps
10 Gbps	1000 MBps	500 MBps
16 Gbps	1600 MBps	800 MBps

13.5 Case studies

This section provides the following case studies that demonstrate how to use the reports to monitor SAN Volume Controller and Storwize V7000:

- ▶ Server performance problem
- ▶ Disk performance problem in a Storwize V7000 subsystem
- ▶ Top volumes response time and I/O rate performance reports
- ▶ Performance constraint alerts for SAN Volume Controller and Storwize V7000
- ▶ Monitoring and diagnosing performance problems for a fabric Viewing
- ▶ Verifying the SAN Volume Controller and Fabric configuration by using Topology Viewer

As appropriate, examples are provided that use the Tivoli Storage Productivity Center 5.2 stand-alone GUI, the Tivoli Storage Productivity Center 5.2 web-based GUI, and Tivoli Common Reporting/Cognos.

13.5.1 Server performance problem

Often, a problem is reported as a server that is suffering from poor performance. Usually the storage disk subsystem is the first suspect. This case study shows how Tivoli Storage Productivity Center can help you to debug this problem. With Tivoli Storage Productivity Center, you can verify whether it is a storage problem or an out-of-storage issue, provide volume mapping for this server, and identify which storage components are involved in the path.

Tivoli Storage Productivity Center provides reports that show the storage that is assigned to the computers within your environment. To display one of the reports, complete the following steps:

1. Click **Disk Manager** $\varnothing$ **Reporting** $\varnothing$ **Storage Subsystem** $\varnothing$ **Computer Views**, and select **By Computer**.
2. Click **Selection**.
3. In the Select Resources window (see Figure 13-102), select the particular available resources to be included in the report. In this example, we select the **tpcblade3-7** server. Then, click **OK**.

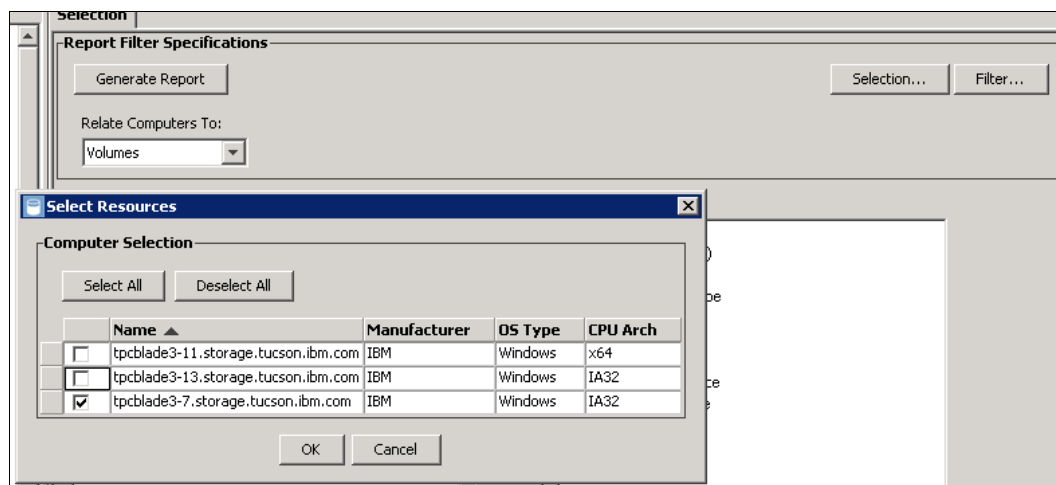


Figure 13-102 Selecting resources

4. Click **Generate Report**. You then see the output on the Computers tab, as shown in Figure 13-103 on page 443.

You can scroll to the right at the bottom of the table to view more information, such as the volume names, volume capacity, and allocated and deallocated volume spaces.

Selection Computers				
Storage Subsystem: By Computer				
Number of Rows: 25				
Computer	Device Path (computer)	Storage Subsystem	Storage Subsystem Type	Volume N
TOTAL				
tpcblade3-7.storage.tucson.ibm.com	Disk 43	Storwize V7000-2076-ford1_tbird-IBM	IBM Storwize V7000	tpcblade3-
tpcblade3-7.storage.tucson.ibm.com	Disk 35	Storwize V7000-2076-ford1_tbird-IBM	IBM Storwize V7000	tpcblade3-
tpcblade3-7.storage.tucson.ibm.com	Disk 34	Storwize V7000-2076-ford1_tbird-IBM	IBM Storwize V7000	tpcblade3-
tpcblade3-7.storage.tucson.ibm.com	Disk 33	Storwize V7000-2076-ford1_tbird-IBM	IBM Storwize V7000	tpcblade3-
tpcblade3-7.storage.tucson.ibm.com	Disk 36	XIV-2810-6000646-IBM	IBM XIV	*tpcblade3
tpcblade3-7.storage.tucson.ibm.com	Disk 29	D58000-2107-1302541-IBM	IBM D58000	tpcblade3-
tpcblade3-7.storage.tucson.ibm.com	Disk 28	D58000-2107-1302541-IBM	IBM D58000	tpcblade3-
tpcblade3-7.storage.tucson.ibm.com	Disk 42	SVC-2145-svc1-IBM	IBM SAN Volume Controller Cluster	tpcblade3
tpcblade3-7.storage.tucson.ibm.com	Disk 41	SVC-2145-svc1-IBM	IBM SAN Volume Controller Cluster	tpcblade3
tpcblade3-7.storage.tucson.ibm.com	Disk 40	SVC-2145-svc1-IBM	IBM SAN Volume Controller Cluster	tpcblade3
tpcblade3-7.storage.tucson.ibm.com	Disk 39	SVC-2145-svc1-IBM	IBM SAN Volume Controller Cluster	tpcblade3
tpcblade3-7.storage.tucson.ibm.com	Disk 38	SVC-2145-svc1-IBM	IBM SAN Volume Controller Cluster	tpcblade3
tpcblade3-7.storage.tucson.ibm.com	Disk 16	XIV-2810-6000646-IBM	IBM XIV	*tpcblade3
tpcblade3-7.storage.tucson.ibm.com	Disk 15	XIV-2810-6000646-IBM	IBM XIV	*tpcblade3
tpcblade3-7.storage.tucson.ibm.com	Disk 14	XIV-2810-6000646-IBM	IBM XIV	*tpcblade3
tpcblade3-7.storage.tucson.ibm.com	Disk 13	XIV-2810-6000646-IBM	IBM XIV	*tpcblade3
tpcblade3-7.storage.tucson.ibm.com	Disk 12	XIV-2810-6000646-IBM	IBM XIV	*tpcblade3
tpcblade3-7.storage.tucson.ibm.com	Disk 11	XIV-2810-6000646-IBM	IBM XIV	*tpcblade3
tpcblade3-7.storage.tucson.ibm.com	Disk 10	XIV-2810-6000646-IBM	IBM XIV	*tpcblade3
tpcblade3-7.storage.tucson.ibm.com	Disk 1	XIV-2810-6000646-IBM	IBM XIV	*tpcblade3
tpcblade3-7.storage.tucson.ibm.com	Disk 27	D58000-2107-1302541-IBM	IBM D58000	tpcbl3_7
tpcblade3-7.storage.tucson.ibm.com	Disk 26	D58000-2107-1302541-IBM	IBM D58000	tpcbl3_7
tpcblade3-7.storage.tucson.ibm.com	Disk 25	D58000-2107-1302541-IBM	IBM D58000	tpcblade3-
tpcblade3-7.storage.tucson.ibm.com	Disk 24	D58000-2107-1302541-IBM	IBM D58000	tpcblade3-
tpcblade3-7.storage.tucson.ibm.com	Disk 23	D58000-2107-1301901-IBM	IBM D58000	tpcblade3-

Figure 13-103 Volume list

5. (Optional) To export data from the report, select **File**  **Export Data** to a comma-delimited file, a comma-delimited file with headers, a formatted report file, and an HTML file.

From the list of this volume, you can start to analyze performance data and workload I/O rate. Tivoli Storage Productivity Center provides a report that shows volume to back-end volume assignments.

6. To display the report, complete the following steps:
 - a. Click **Disk Manager**  **Reporting**  **Storage Subsystem**  **Volume to Backend Volume Assignment**, and select **By Volume**.
 - b. Click **Filter** to limit the list of the volumes to those volumes that belong to the tpcblade3-7 server, as shown in Figure 13-104 on page 444.

[illegible]

Figure 13-104 Volume to back-end filter

- c. Click **Generate Report**.

You now see a list similar to the one that is shown in Figure 13-105.

Selection Volumes							
Volumes: By Volume							
Number of Rows: 24							
Storage Subsystem	Storage Subsystem Type	User-Defined Volume Name	Volume Name	Volume Capacity	Storage Pool	Disk	Disk Capacity
SVC-2145-ITSOSVC01-IBM	SVC	zmassimo61	60050768018600C470000000000000A3	500.00 MB	MDG_HADR	mdisk6	15.00
SVC-2145-ITSOSVC01-IBM	SVC	zmassimo0823	60050768018600C470000000000000A4	100.00 MB	MDG_HADR	mdisk6	15.00
SVC-2145-ITSOSVC01-IBM	SVC	zmassimo60	60050768018600C470000000000000A2	500.00 MB	MDG_HADR	mdisk6	15.00
SVC-2145-ITSOSVC01-IBM	SVC	zmassimo0826	60050768018600C470000000000000A5	376.00 MB	MDG_HADR	mdisk6	15.00
SVC-2145-ITSOSVC01-IBM	SVC	zmassimo0826	60050768018600C470000000000000A5	376.00 MB	MDG_HADR	mdisk8	15.00
SVC-2145-ITSOSVC01-IBM	SVC	zmassimo61	60050768018600C470000000000000A3	500.00 MB	MDG_HADR	mdisk8	15.00
SVC-2145-ITSOSVC01-IBM	SVC	zmassimo60	60050768018600C470000000000000A2	500.00 MB	MDG_HADR	mdisk8	15.00
SVC-2145-ITSOSVC01-IBM	SVC	zmassimo0826TPC	60050768018600C470000000000000A6	100.00 MB	MDG_HADR	mdisk8	15.00
SVC-2145-ITSOSVC01-IBM	SVC	ITSO_Perf_01	60050768018600C4700000000000009C	10.00 GB	MDG_FC	mdisk16	5.00
SVC-2145-ITSOSVC01-IBM	SVC	ITSO_PERF_02	60050768018600C470000000000000A1	4.97 GB	MDG_FC	mdisk16	5.00
SVC-2145-ITSOSVC01-IBM	SVC	ITSO_Perf_01	60050768018600C4700000000000009C	10.00 GB	MDG_FC	mdisk18	5.00
SVC-2145-ITSOSVC01-IBM	SVC	ITSO_PERF_02	60050768018600C470000000000000A1	4.97 GB	MDG_FC	mdisk18	5.00
SVC-2145-ITSOSVC01-IBM	SVC	zmassimo0826	60050768018600C470000000000000A5	376.00 MB	MDG_HADR	mdisk7	15.00
SVC-2145-ITSOSVC01-IBM	SVC	zmassimo61	60050768018600C470000000000000A3	500.00 MB	MDG_HADR	mdisk7	15.00
SVC-2145-ITSOSVC01-IBM	SVC	zmassimo0823	60050768018600C470000000000000A4	100.00 MB	MDG_HADR	mdisk7	15.00
SVC-2145-ITSOSVC01-IBM	SVC	zmassimo60	60050768018600C470000000000000A2	500.00 MB	MDG_HADR	mdisk7	15.00
SVC-2145-ITSOSVC01-IBM	SVC	zmassimo0826TPC	60050768018600C470000000000000A6	100.00 MB	MDG_HADR	mdisk7	15.00
SVC-2145-ITSOSVC01-IBM	SVC	zmassimo0826	60050768018600C470000000000000A5	376.00 MB	MDG_HADR	mdisk5	15.00
SVC-2145-ITSOSVC01-IBM	SVC	zmassimo61	60050768018600C470000000000000A3	500.00 MB	MDG_HADR	mdisk5	15.00
SVC-2145-ITSOSVC01-IBM	SVC	zmassimo60	60050768018600C470000000000000A2	500.00 MB	MDG_HADR	mdisk5	15.00
SVC-2145-ITSOSVC01-IBM	SVC	ITSO_Perf_01	60050768018600C4700000000000009C	10.00 GB	MDG_FC	mdisk19	5.00
SVC-2145-ITSOSVC01-IBM	SVC	ITSO_PERF_02	60050768018600C470000000000000A1	4.97 GB	MDG_FC	mdisk19	5.00
SVC-2145-ITSOSVC01-IBM	SVC	ITSO_Perf_01	60050768018600C4700000000000009C	10.00 GB	MDG_FC	mdisk17	5.00
SVC-2145-ITSOSVC01-IBM	SVC	ITSO_PERF_02	60050768018600C470000000000000A1	4.97 GB	MDG_FC	mdisk17	5.00

Figure 13-105 Volume to back-end list

- d. Scroll to the right to see the SAN Volume Controller managed disks and back-end volumes on the DS8000, as shown in Figure 13-106.

Back-end storage subsystem: The highlighted lines with the value N/A are related to a back-end storage subsystem that is not defined in our Tivoli Storage Productivity Center environment. To obtain the information about the back-end storage subsystem, we must add it in the Tivoli Storage Productivity Center environment with the corresponding probe job. See the first line in the report in Figure 13-106, where the back-end storage subsystem is part of our Tivoli Storage Productivity Center environment. Therefore, the volume is correctly shown in all details.

Selection Volumes							
Volumes: By Volume							
Number of Rows: 17							
Storage Subsystem	Available Disk Space	Backend Storage Subsystem	Backend Storage Subsystem Type	Backend Volume Name	Backend Volume Space	Co	
SVC-2145-svc1-IBM	0	DS8000-2107-1302541-IBM	DS8000	SVC00227(ID:f7e3)	1.00 GB	0	
Storwize V7000-2076-ford1_tbird-IBM	2.00 GB	DS8000-2107-1302541-IBM	DS8000	rad(ID:00a3)	30.00 GB	0	
Storwize V7000-2076-ford1_tbird-IBM	3.25 GB	DS8000-2107-1302541-IBM	DS8000	rad(ID:009f)	30.00 GB	0	
Storwize V7000-2076-ford1_tbird-IBM	3.25 GB	DS8000-2107-1302541-IBM	DS8000	rad(ID:009f)	30.00 GB	0	
SVC-2145-svc1-IBM	1.07 TB	N/A	N/A	N/A	N/A	0	
SVC-2145-svc1-IBM	1.07 TB	N/A	N/A	N/A	N/A	0	
SVC-2145-svc1-IBM	1.07 TB	N/A	N/A	N/A	N/A	0	
SVC-2145-svc1-IBM	1.07 TB	N/A	N/A	N/A	N/A	0	
SVC-2145-svc1-IBM	1.07 TB	N/A	N/A	N/A	N/A	0	
SVC-2145-svc1-IBM	1.07 TB	N/A	N/A	N/A	N/A	0	
SVC-2145-svc1-IBM	1.07 TB	N/A	N/A	N/A	N/A	0	
SVC-2145-svc1-IBM	1.07 TB	N/A	N/A	N/A	N/A	0	
SVC-2145-svc1-IBM	1.07 TB	N/A	N/A	N/A	N/A	0	
SVC-2145-svc1-IBM	1.07 TB	N/A	N/A	N/A	N/A	0	
SVC-2145-svc1-IBM	1.07 TB	N/A	N/A	N/A	N/A	0	
Storwize V7000-2076-ford1_tbird-IBM	3.50 GB	DS8000-2107-1302541-IBM	DS8000	rad(ID:00a0)	30.00 GB	0	
Storwize V7000-2076-ford1_tbird-IBM	3.50 GB	DS8000-2107-1302541-IBM	DS8000	rad(ID:00a0)	30.00 GB	0	

Figure 13-106 Back-end storage subsystems

With this information and the list of volumes that are mapped to this computer, you can start to run a Performance report to understand where the problem for this server might be.

13.5.2 Disk performance problem in a Storwize V7000 subsystem

This case study examines a problem that is reported by a customer. In this case, one disk volume has different and lower performance results during the last period. At times, it has a good response time, but at other times, the response time is unacceptable. Throughput is also changing. The customer specified that the name of the affected volume is tpcblade3-7-ko2, which is a VDisk in a Storwize V7000 subsystem.

Tip: Looking at disk performance problems, check the overall response time and its overall I/O rate. If both are high, a problem might exist. If the overall response time is high and the I/O rate is trivial, the effect of the high overall response time might be inconsequential.

To check the overall response time, complete the following steps:

1. Click **Disk Manager** $\varnothing$ **Reporting** $\varnothing$ **Storage Subsystem Performance**, and select **By Volume**.
2. In the Selection tab, click **Filter**.
3. Create a filter to produce a report for all the volumes that belong to the Storwize V7000 subsystems, as shown in Figure 13-107.

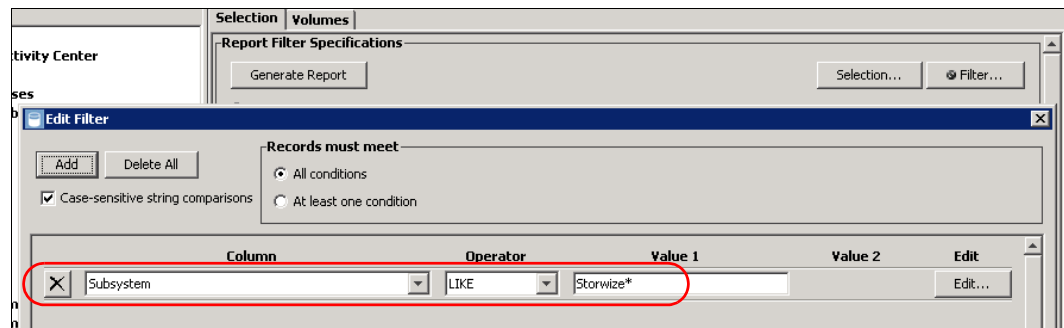


Figure 13-107 SAN Volume Controller performance report by Volume

4. In the Volumes tab, click the volume that you need to investigate and then click the pie chart icon (📊).
5. In the Select Charting Option window (see Figure 13-108 on page 447), select **Total I/O Rate (overall)**. Then, click **OK** to produce the graph.

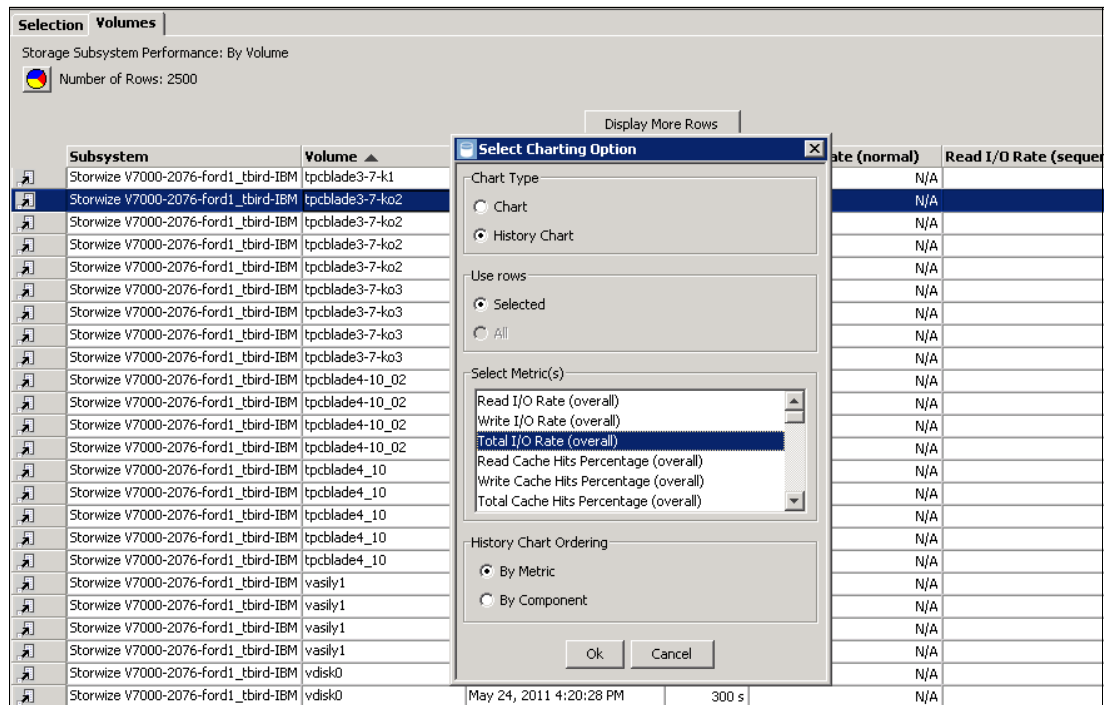


Figure 13-108 Storwize V7000 performance report: Volume selection

The history chart that is shown in Figure 13-109 shows that I/O rate was around 900 operations per second and suddenly declined to around 400 operations per second. Then, the rate goes back to 900 operations per second. In this case study, we limited the days to the time frame that was reported by the customer when the problem was noticed.

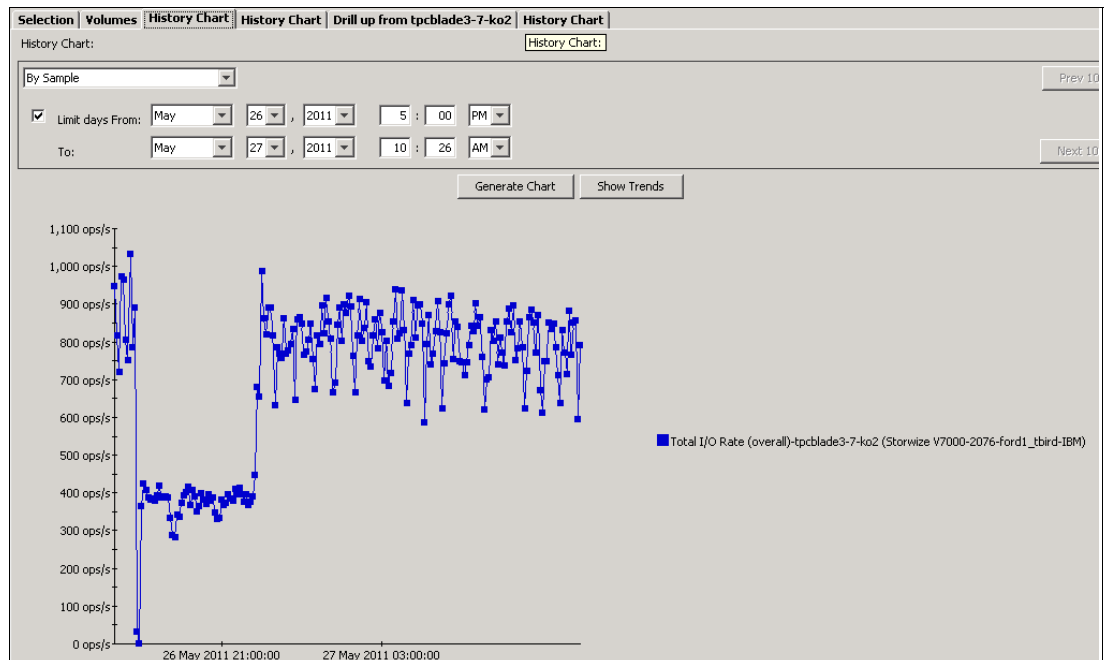


Figure 13-109 Total I/O rate chart for the Storwize V7000 volume

- In the Volumes tab, select the volume that you need to investigate, and then click the pie chart icon (📊).

7. In the Select Charting Option window (see Figure 13-110), scroll down and select **Overall Response Time**. Then, click **OK** to produce the chart.

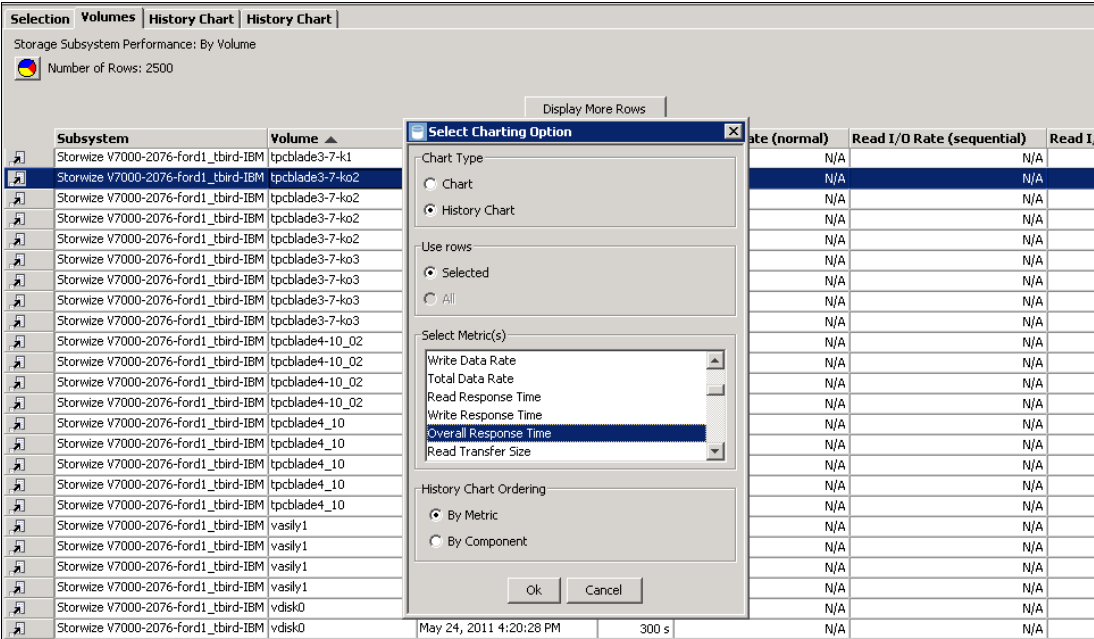


Figure 13-110 Volume selection for the Storwize V7000 performance report

The chart that is shown in Figure 13-111 indicates an increase in response time from a few milliseconds to around 30 ms. This information and the high I/O rate indicate the occurrence of a significant problem. Therefore, further investigation is appropriate.

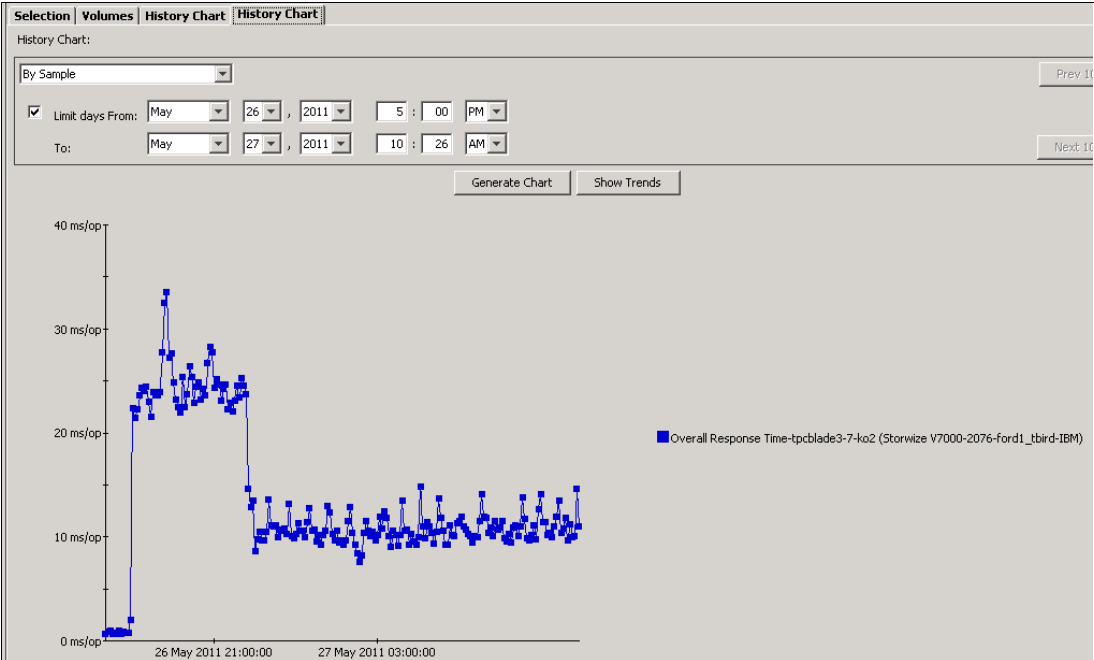


Figure 13-111 Response time for Storwize V7000 volume

8. Complete the following steps to review the performance of MDisks in the managed disk group:
 - a. To identify to which MDisk the tpcblade3-7-ko2 VDisk belongs, in the Volumes tab (see Figure 13-112), click the drill-up icon ().

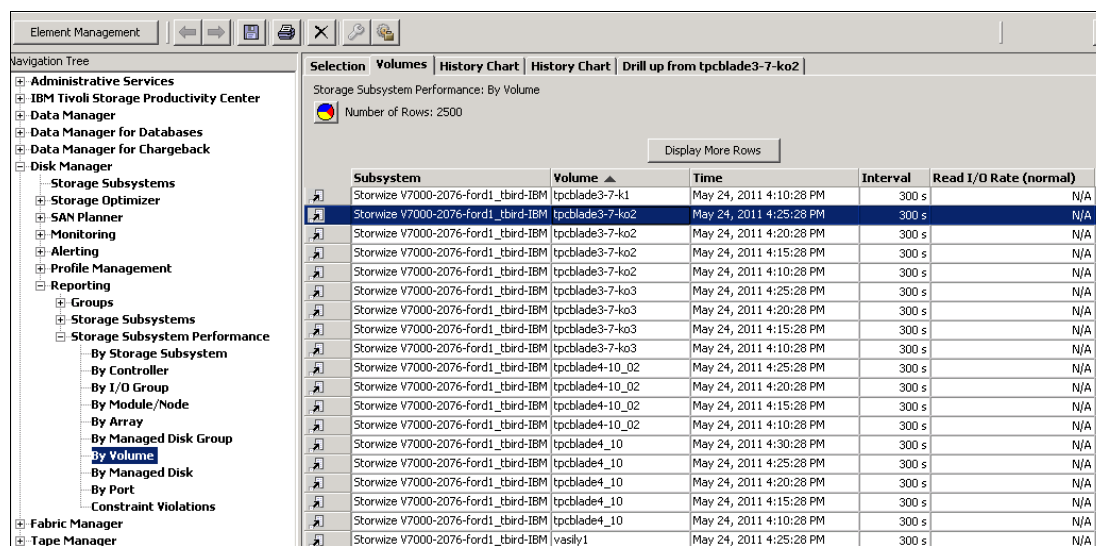


Figure 13-112 Drilling up to determine the MDisk

Figure 13-113 shows the MDisks where the tpcblade3-7-ko2 extents are installed.

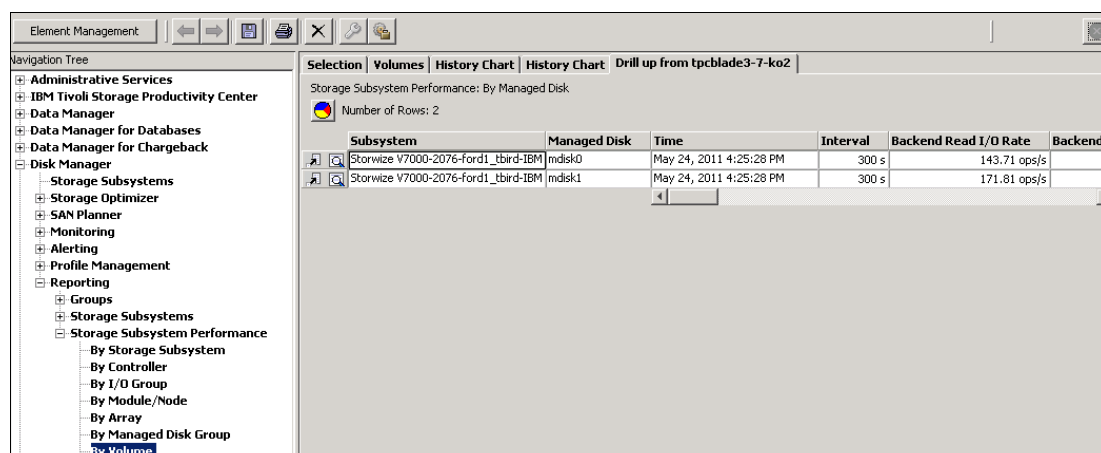


Figure 13-113 Storwize V7000 Volume and MDisk selection

- b. Select all the MDisks and click the pie chart icon ().
- c. In the Select Charting Option window (see Figure 13-114 on page 450), select **Overall Backend Response Time**, and then click **OK**.

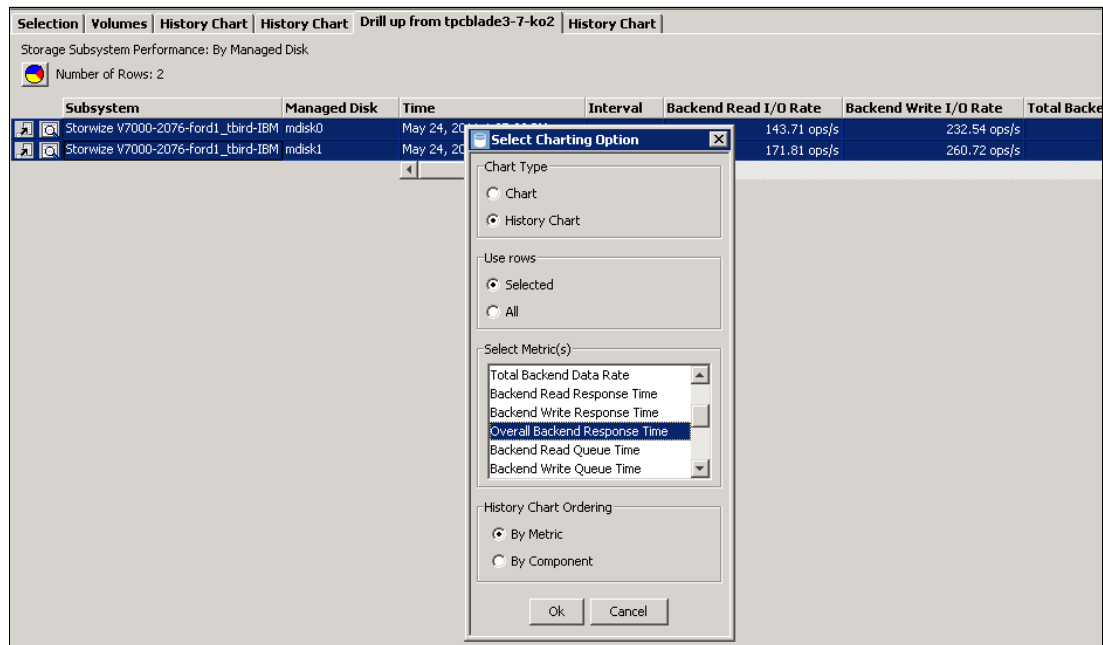


Figure 13-114 Storwize V7000 metric selection

Keep the charts that are generated relevant to this scenario by using the charting time range. You can see from the chart that is shown in Figure 13-115 that something happened on 26 May around 6:00 p.m. that likely caused the back-end response time for all MDisks to dramatically increase.

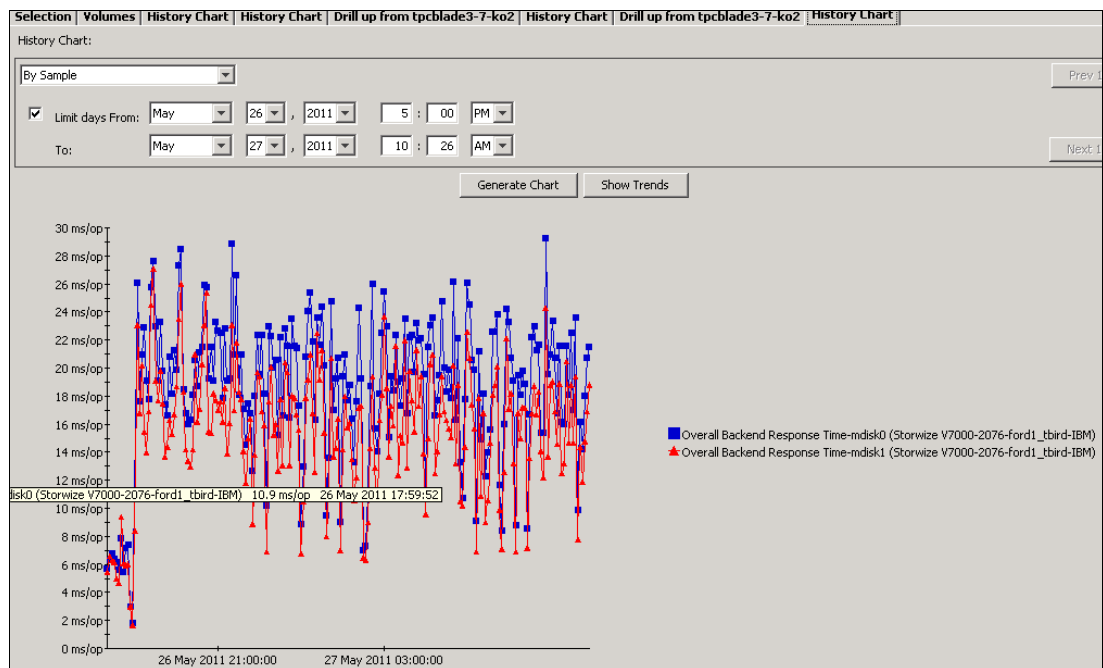


Figure 13-115 Overall Backend Response Time

If you review the chart for the Total Backend I/O Rate for these two MDisks during the same time period, you see that their I/O rates all remained in a similar overlapping pattern, even after the problem was introduced.

This result is as expected and might occur because tpcblade3-7-ko2 is evenly striped across the two MDisk. The I/O rate for these MDisk is only as high as the slowest MDisk, as shown in Figure 13-116.

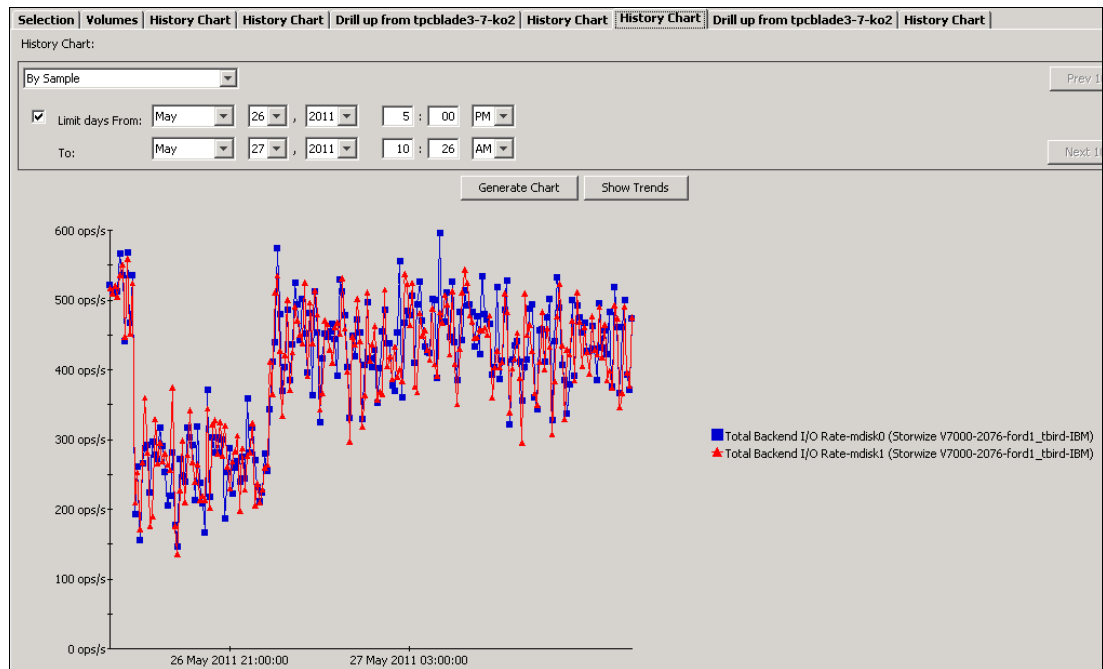


Figure 13-116 Backend I/O Rate

We identified that the response time for all MDisk dramatically increased.

9. Generate a report to show the volumes that have an overall I/O rate equal to or greater than 1000 Ops/ms. We also generate a chart to show which I/O rates of the volume changed around 5:30 p.m. on 20 August. Complete the following steps:
 - a. Click **Disk Manager** **Reporting** **Storage Subsystem Performance**, and select **By Volume**.
 - b. In the Selection tab, complete the following steps:
 - i. Click **Display historic performance data using absolute time**.
 - ii. Limit the time period to 1 hour before and 1 hour after the event that was reported, as shown in Figure 13-115 on page 450.
 - iii. Click **Filter** to limit to Storwize V7000 Subsystem.
 - c. In the Edit Filter window (see Figure 13-117 on page 452), complete the following steps:
 - i. Click **Add** to add a second filter.
 - ii. Select the **Total I/O Rate (overall)**, and set it to greater than 1000 (meaning a high I/O rate).
 - iii. Click **OK**.

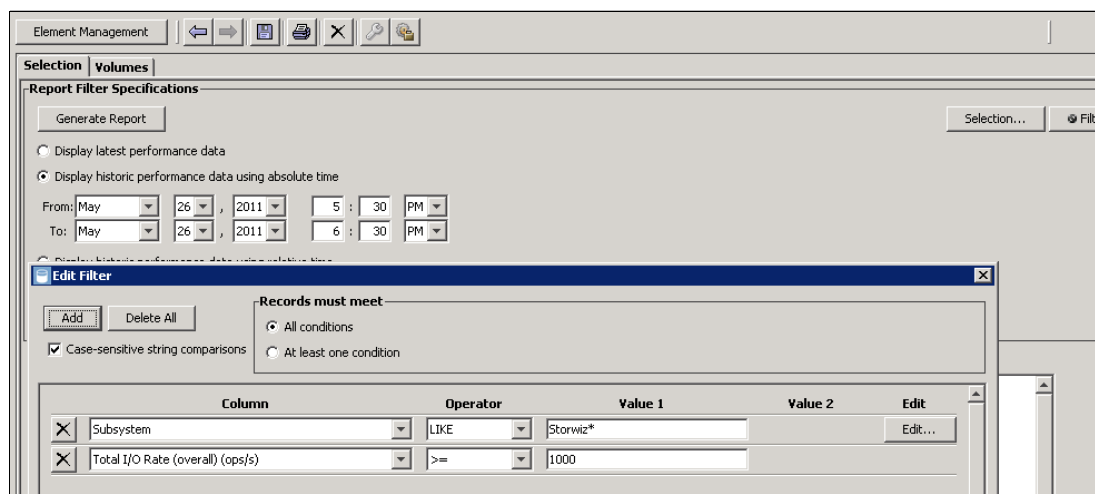


Figure 13-117 Displaying the historic performance data

The report that is shown in Figure 13-118 shows all of the performance records of the volumes that were filtered previously. In the Volume column, only three volumes meet these criteria: tpcblade3-7-ko2, tpcblade3-7-ko3, and tpcblade3ko4. Multiple rows are available for each volume because each performance data record has a row. Look for which volumes had an I/O rate change around 6:00 p.m. on 26 May. You can click the **Time** column to sort the data.

Selection Volumes		Storage Subsystem Performance: By Volume							
		Number of Rows: 20							
Subsystem	Volume	Time	Interval	Read I/O Rate (normal)	Read I/O Rate (sequential)	Read I/O Rate (overall)	Write I/O Rate (normal)	Write I/O Rate (sequential)	Write I/O Rate (overall)
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-7-ko2	May 26, 2011 5:34:52 PM	300 s	N/A	N/A	512.87 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-7-ko3	May 26, 2011 6:29:52 PM	300 s	N/A	N/A	1,123.95 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-7-ko3	May 26, 2011 6:24:52 PM	300 s	N/A	N/A	1,104.84 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-7-ko3	May 26, 2011 6:19:52 PM	300 s	N/A	N/A	1,081.52 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-7-ko3	May 26, 2011 6:14:52 PM	300 s	N/A	N/A	1,179.31 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-7-ko3	May 26, 2011 6:09:52 PM	300 s	N/A	N/A	1,199.63 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-7-ko3	May 26, 2011 6:04:52 PM	300 s	N/A	N/A	1,362.37 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-7-ko3	May 26, 2011 5:59:52 PM	300 s	N/A	N/A	842.77 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-7-ko3	May 26, 2011 5:34:52 PM	300 s	N/A	N/A	509.8 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-ko4	May 26, 2011 6:29:52 PM	300 s	N/A	N/A	989.61 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-ko4	May 26, 2011 6:24:52 PM	300 s	N/A	N/A	942.46 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-ko4	May 26, 2011 6:19:52 PM	300 s	N/A	N/A	932.98 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-ko4	May 26, 2011 6:14:52 PM	300 s	N/A	N/A	1,048.82 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-ko4	May 26, 2011 6:09:52 PM	300 s	N/A	N/A	1,025.83 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-ko4	May 26, 2011 6:04:52 PM	300 s	N/A	N/A	1,122.77 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-ko4	May 26, 2011 5:59:52 PM	300 s	N/A	N/A	780.44 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-ko4	May 26, 2011 5:44:52 PM	300 s	N/A	N/A	635.01 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-ko4	May 26, 2011 5:39:52 PM	300 s	N/A	N/A	584.95 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-ko4	May 26, 2011 5:34:52 PM	300 s	N/A	N/A	644.32 ops/s	N/A	N/A	N/A
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-ko4	May 26, 2011 5:29:52 PM	300 s	N/A	N/A	668.67 ops/s	N/A	N/A	N/A

Figure 13-118 I/O rate of the volume changed

10. Complete the following steps to compare the Total I/O rate (overall) metric for these volumes and the volume subject of the case study, tpcblade3-7-k02:

- Remove the filtering condition on the Total I/O Rate that is defined in Figure 13-117, and then generate the report again.
- Select one row for each of these volumes.

- c. In the Select Charting Option window (see Figure 13-119), select **Total I/O Rate (overall)**, and then click **OK** to generate the chart.

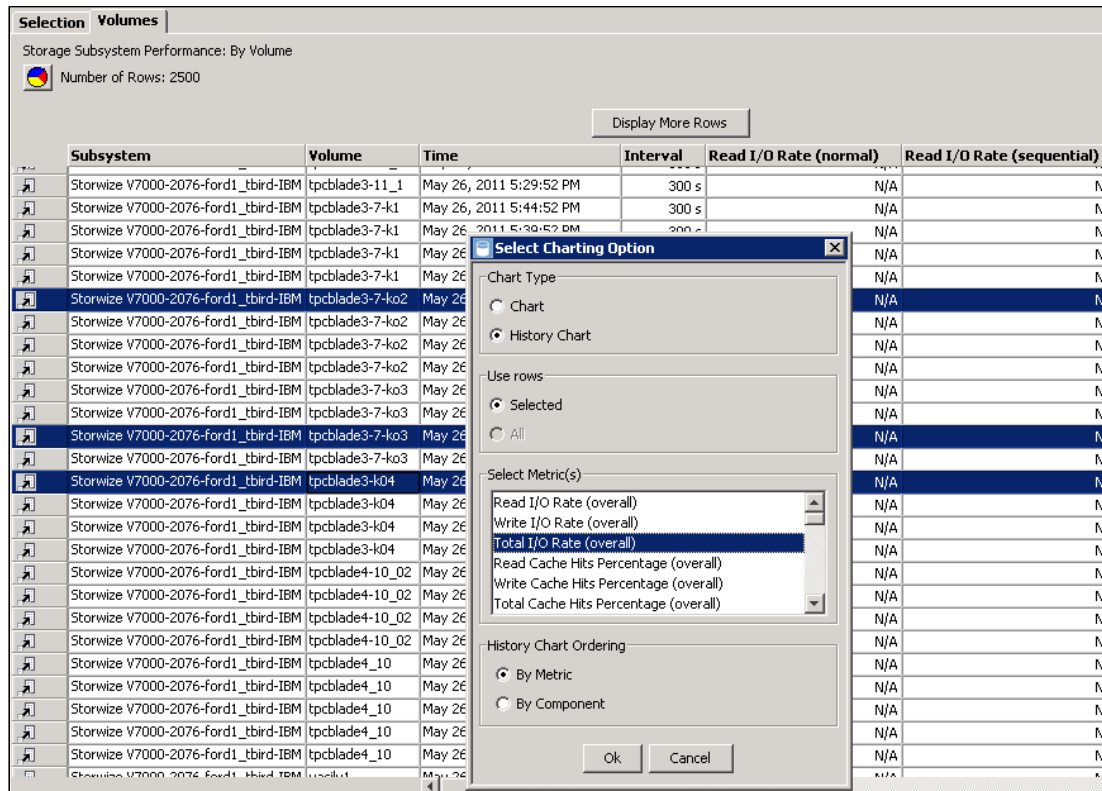


Figure 13-119 Total I/O rate selection for three volumes

- d. For Limit days From, insert the time frame that you are investigating.

Figure 13-120 on page 454 shows the root cause. The tpcblade3-7-ko2 volume (the blue line in Figure 13-120 on page 454) started around 5:00 p.m. and has a total I/O rate of around 1000 IOPS. When the new workloads (which were generated by the tpcblade3-7-ko3 and tpcblade3-ko4 volumes) started, the total I/O rate for the tpcblade3-7-ko2 volume fell from around 1000 IOPS to less than 500 I/Os. Then, it grew again to about 1000 I/Os when one of the two loads decreased. The hardware has physical limitations on the number of IOPS that it can handle. This limitation was reached at 6:00 p.m.

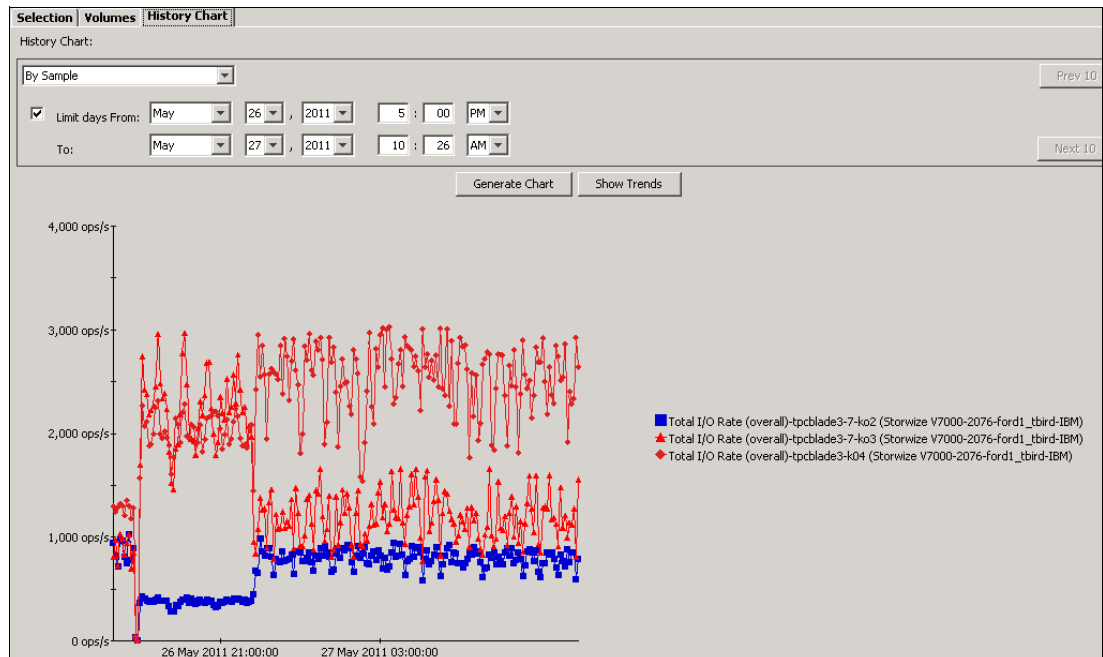


Figure 13-120 Total I/O rate chart for three volumes

To confirm this behavior, you can generate a chart by selecting **Response time**. The chart that is shown in Figure 13-121 confirms that, when the new workload started, the response time for the tpcblade3-7-ko2 volume becomes worse.

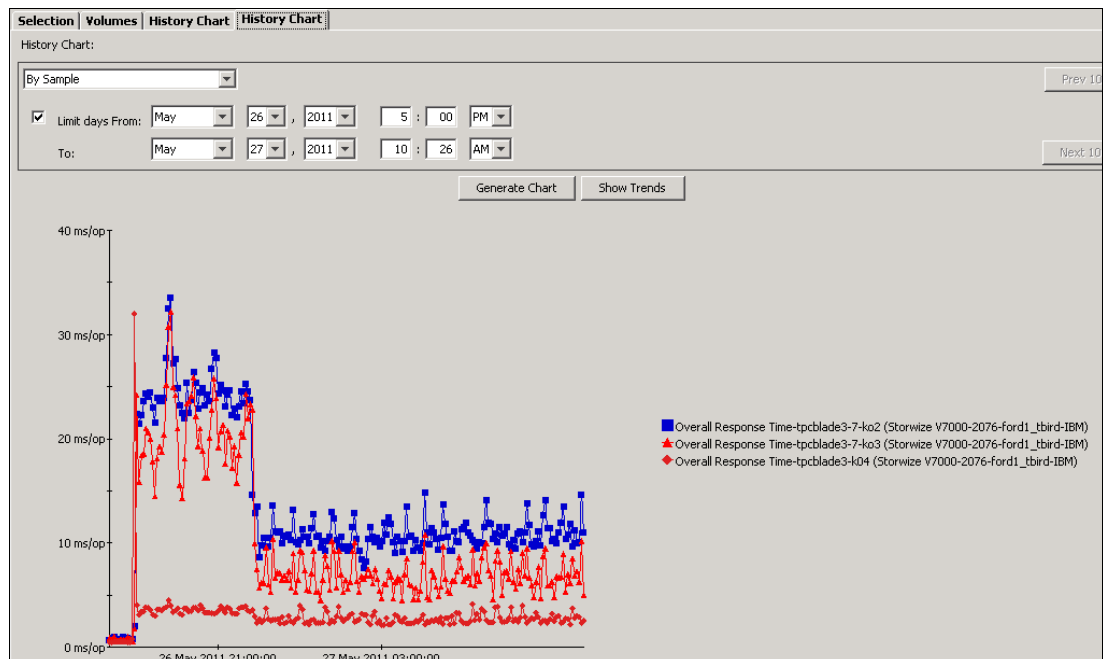


Figure 13-121 Response time chart for three volumes

The easy solution is to split this workload by moving one VDisk to another managed disk group.

13.5.3 Top volumes response time and I/O rate performance reports

Reports about the most active volumes in a SAN Volume Controller or Storwize family systems can be accessed via all three Tivoli Storage Productivity Center GUI interfaces, as described in 13.3.4, “Top Volume Performance reports” on page 422. In this case study, we use the stand-alone GUI.

In this section, we show how to tailor the Top Volumes Response Performance report (which is available in the Tivoli Storage Productivity Center 5.2 stand-alone GUI) to identify volumes with long response times and high I/O rates. You can tailor the report for your environment. You can also update your filters to exclude volumes or subsystems that you no longer want in this report.

Complete the following steps to tailor the Top Volumes Response Performance report:

1. Click **Disk Manager** **Reporting** **Storage Subsystem Performance**, and select **By Volume**, as shown in the left pane in Figure 13-122.

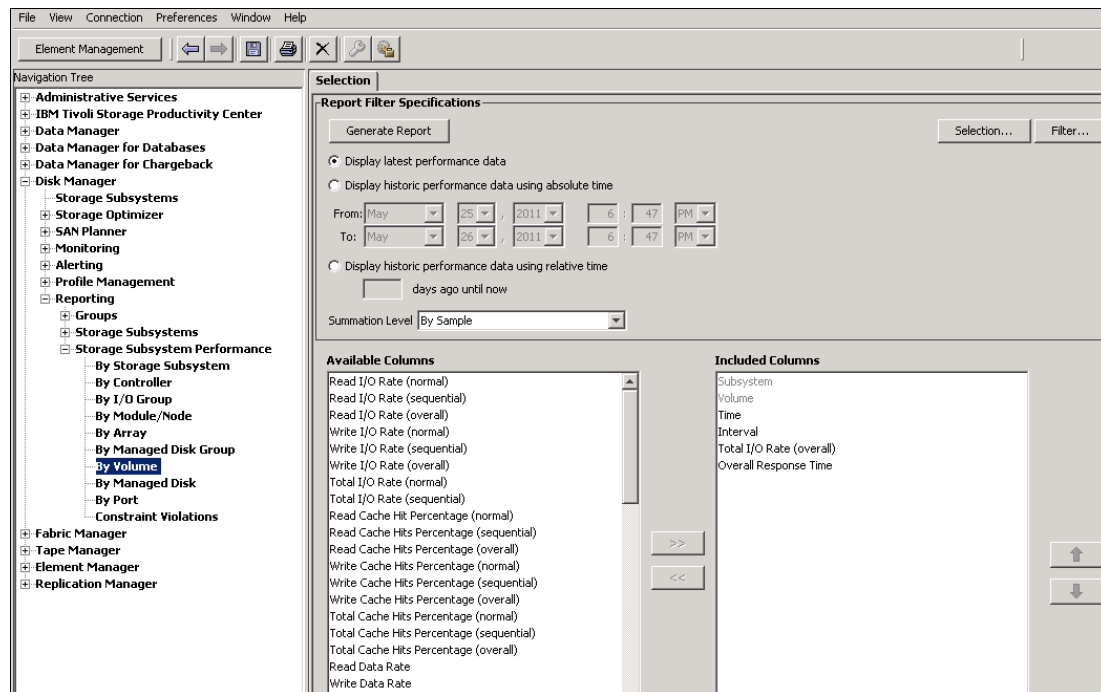


Figure 13-122 Metrics for tailored reports of top volumes

2. In the Selection tab (as shown in the right pane in Figure 13-122), keep only the wanted metrics in the Included Columns box and move all other metrics (by using the arrow buttons) to the Available Columns box.

You can save this report for future reference by clicking **IBM Tivoli Storage Productivity Center** **My Reports** **“your user” Reports**.

Click **Filter** to specify the filters to limit the report.

3. In the Edit Filter window (see Figure 13-123 on page 456), click **Add** to add the conditions. In this example, we limit the report to Subsystems SVC* and DS8*. We also limit the report to the volumes that have an I/O rate greater than 100 Ops/sec and a Response Time greater than 5 ms.

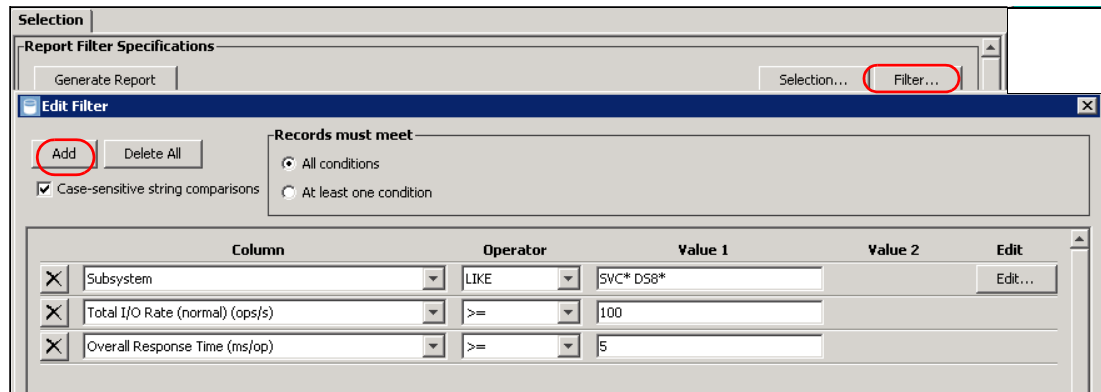


Figure 13-123 Filters for the top volumes tailored reports

4. Complete the following steps in the Selection tab, as shown in Figure 13-124:
 - a. Specify the date and time of the period for which you want to make the inquiry.

Important: Specifying large intervals might require intensive processing and a long time to complete.

- b. Click **Generate Report**.

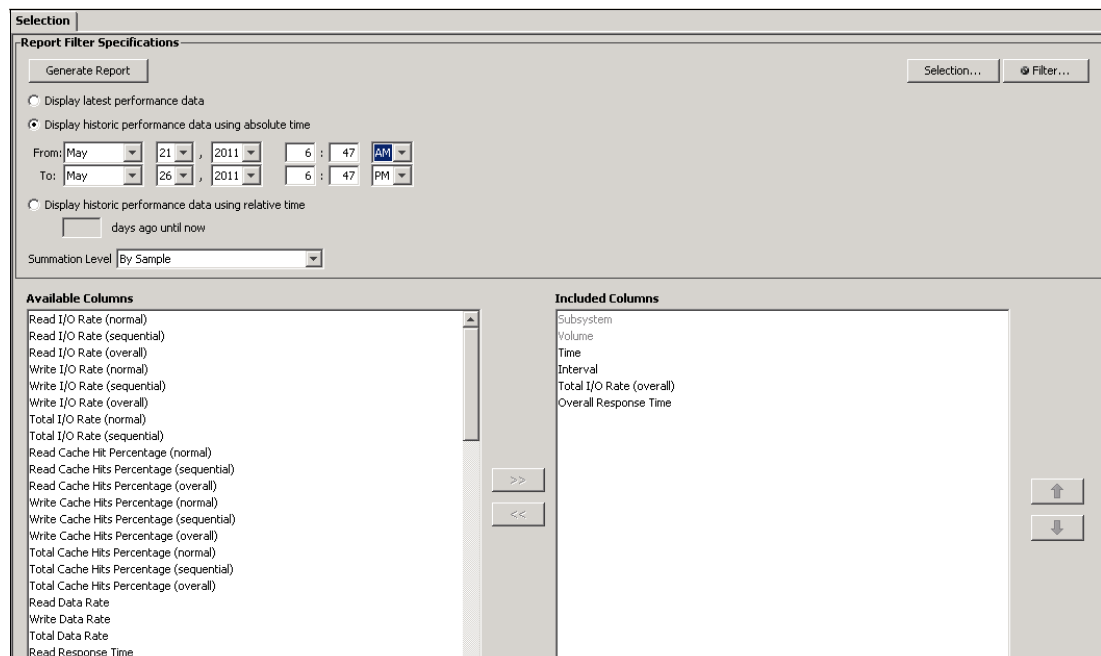


Figure 13-124 Limiting the days for the top volumes tailored report

Figure 13-125 on page 457 shows the resulting Volume list. By sorting by the Overall Response Time or I/O Rate columns (by clicking the column header), you can identify which entries have interesting total I/O rates and overall response times.

Selection		Volumes				
Storage Subsystem Performance: By Volume						
 Number of Rows: 1756						
	Subsystem	Volume	Time	Interval	Total I/O Rate (overall)	Overall Response Time
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 7:01:40 PM	298 s	270.52 ops/s	9.7 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 6:56:40 PM	303 s	264.73 ops/s	11 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 6:51:41 PM	299 s	286.45 ops/s	8.5 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 6:46:40 PM	301 s	267.39 ops/s	9.8 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 6:41:40 PM	300 s	279.13 ops/s	18.9 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 6:36:41 PM	299 s	291.74 ops/s	16.7 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 6:31:40 PM	301 s	314.06 ops/s	9 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 6:26:41 PM	299 s	190.29 ops/s	6 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 6:21:40 PM	301 s	171.8 ops/s	5.9 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 6:16:40 PM	300 s	294.28 ops/s	13.1 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 6:11:40 PM	300 s	334.57 ops/s	8.4 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 6:06:41 PM	299 s	290.78 ops/s	14.7 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 5:41:39 PM	299 s	458.17 ops/s	5.1 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 5:31:41 PM	297 s	445.55 ops/s	8 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 5:11:45 PM	301 s	516.68 ops/s	5.7 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 4:46:41 PM	300 s	486.53 ops/s	6.3 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 4:41:41 PM	299 s	493.62 ops/s	5.5 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 4:31:41 PM	298 s	546.46 ops/s	6.3 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 4:16:40 PM	301 s	548.43 ops/s	6 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 4:11:41 PM	298 s	462.37 ops/s	5.5 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 4:06:41 PM	300 s	462.92 ops/s	6.5 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 4:01:41 PM	301 s	583.32 ops/s	5.6 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 3:31:41 PM	298 s	610.27 ops/s	5.1 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 3:01:41 PM	299 s	504.57 ops/s	8.5 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 2:51:41 PM	297 s	435.12 ops/s	5.9 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 2:31:41 PM	300 s	513.22 ops/s	5.3 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 2:26:40 PM	302 s	515.81 ops/s	8.4 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 2:21:41 PM	298 s	536.55 ops/s	5.6 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 2:11:40 PM	300 s	515.08 ops/s	5.2 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 1:51:40 PM	299 s	460.17 ops/s	7.5 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 1:46:41 PM	299 s	498.92 ops/s	6.4 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 1:41:40 PM	302 s	510.44 ops/s	6.2 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 1:36:40 PM	300 s	527.82 ops/s	5.7 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 1:26:41 PM	301 s	494.09 ops/s	5.7 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 1:21:41 PM	300 s	519.38 ops/s	5.7 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 1:16:41 PM	298 s	456.72 ops/s	7.2 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 1:11:41 PM	301 s	543.03 ops/s	6.5 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 1:06:40 PM	301 s	439.14 ops/s	11.3 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 1:01:41 PM	299 s	476.71 ops/s	9.1 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 12:56:40 PM	303 s	466.55 ops/s	7.7 ms/op
	DS8000-2107-1302541-IBM	rad(ID:009f)	May 26, 2011 12:51:40 PM	298 s	462.38 ops/s	10.5 ms/op

Figure 13-125 Volumes list of the top volumes tailored report

Guidelines for total I/O rate and overall response time in a production environment

In a production environment, you initially might want to specify a total I/O rate overall of 1 - 100 Ops/sec and an overall response time (ms) that is greater than or equal to 15 ms. Then, adjust these values to suit your needs as you gain more experience.

13.5.4 Performance constraint alerts

In this section, we describe the alerts that are available to show performance constraint for SAN Volume Controller and Storwize V7000.

Using the Tivoli Storage Productivity Center 5.2 stand-alone GUI to create and view alerts

Along with reporting on SAN Volume Controller and Storwize V7000 performance, Tivoli Storage Productivity Center can generate alerts when performance thresholds are not met or exceed a defined threshold. Similar to most Tivoli Storage Productivity Center tasks, Tivoli Storage Productivity Center sends alerts to the following items:

- ▶ Simple Network Management Protocol (SNMP)

With an alert, you can send an SNMP trap to an upstream systems management application. The SNMP trap can then be used with other events that occur in the environment to help determine the root cause of an SNMP trap.

In this case, the SNMP trap was generated by the SAN Volume Controller. For example, if the SAN Volume Controller or Storwize V7000 reported to Tivoli Storage Productivity Center that a fiber port went offline, this problem might occur because of a failed switch.

By using by a systems management tool, the port failed trap and the switch offline trap can be analyzed as a switch problem, not a SAN Volume Controller (or Storwize V7000) problem.

- ▶ Tivoli Omnibus Event

Select this option to send a Tivoli Omnibus event.

- ▶ Login Notification

Select this option to send the alert to a Tivoli Storage Productivity Center user. The user receives the alert upon logging in to Tivoli Storage Productivity Center. In the Login ID field, enter the user ID.

- ▶ UNIX or Windows NT system event logger

Select this option to log to a UNIX or Windows BT system event logger.

- ▶ Script

By using the Script option, you can run a predefined set of commands that can help address the event, such as opening a ticket in your help-desk ticket system.

- ▶ Email

Tivoli Storage Productivity Center sends an email to each person who is listed in its email settings.

Tip: For Tivoli Storage Productivity Center to send email to a list of addresses, you must identify an email relay by selecting **Administrative Services** Ø **Configuration** Ø **Alert Disposition** and then selecting **E-mail** settings.

Consider setting the following alert events:

- ▶ CPU utilization threshold

The CPU utilization report alerts you when your SAN Volume Controller or Storwize V7000 nodes become too busy. If this alert is generated too often, you might need to upgrade your cluster with more resources.

For development reasons, use a setting of 75% to indicate a warning alert or a setting of 90% to indicate a critical alert. These settings are the default settings for Tivoli Storage Productivity Center V4.2.1. To enable this function, create an alert by selecting the CPU Utilization. Then, define the alert actions to be performed.

In the Storage Subsystem tab, select the SAN Volume Controller or Storwize V7000 cluster for which to set this alert.

- Overall port response time threshold

The port response times alert can inform you of when the SAN fabric is becoming a bottleneck. If the response times are consistently bad, perform more analysis of your SAN fabric.

- Overall back-end response time threshold

An increase in back-end response time might indicate that you are overloading your back-end storage for the following reasons:

- Because back-end response times can vary depending on which I/O workloads are in place. Before you set this value, capture 1 - 4 weeks of data to set a baseline for your environment. Then, set the response time values.
- Because you can select the storage subsystem for this alert. You can set different alerts that are based on the baselines that you captured. Start with your mission-critical Tier 1 storage subsystems.

Complete the following steps to create an alert:

1. Click **Disk Manager** **Alerting** **Storage Subsystem Alerts**. Right-click and select **Create a Storage Subsystems Alert**, as shown in the left pane in Figure 13-126.

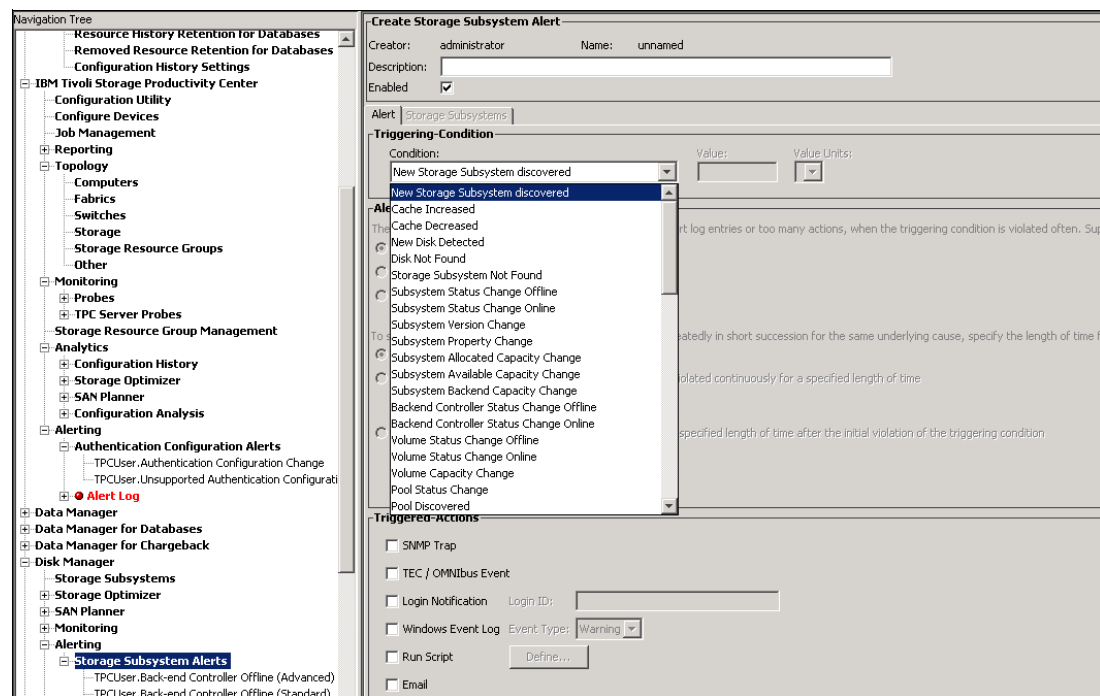


Figure 13-126 SAN Volume Controller constraints alert definition

2. In the right pane as shown in Figure 13-126, in the Triggering Condition box under Condition, select the alert that you want to set.

Tip: The best place to verify which thresholds are currently enabled (and at what values) is at the beginning of a Performance Collection job.

To schedule the Performance Collection job and verify the thresholds, complete the following steps:

1. Click **Tivoli Storage Productivity Center** **Job Management**, as shown in the left pane of Figure 13-127 on page 460.

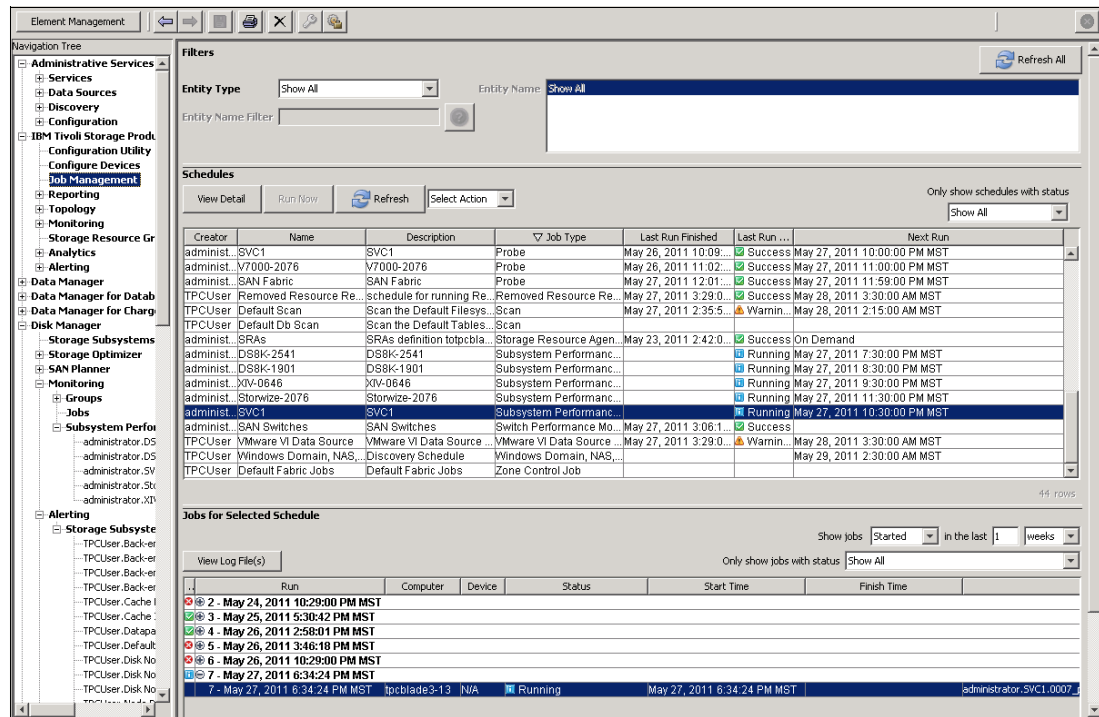


Figure 13-127 Job management panel and SAN Volume Controller performance job log selection

2. In the Schedules table (as shown in the upper part of the right pane), select the latest performance collection job that is running or that ran for your subsystem.
3. In the Job for Selected Schedule (as shown in the lower part of the right pane), expand the corresponding job, and select the instance.
4. To access to the corresponding log file, click **View Log File(s)**. Then, you can see the threshold that is defined, as shown in Figure 13-128 on page 461.

Tip: To return to the beginning of the log file, click **Top**.

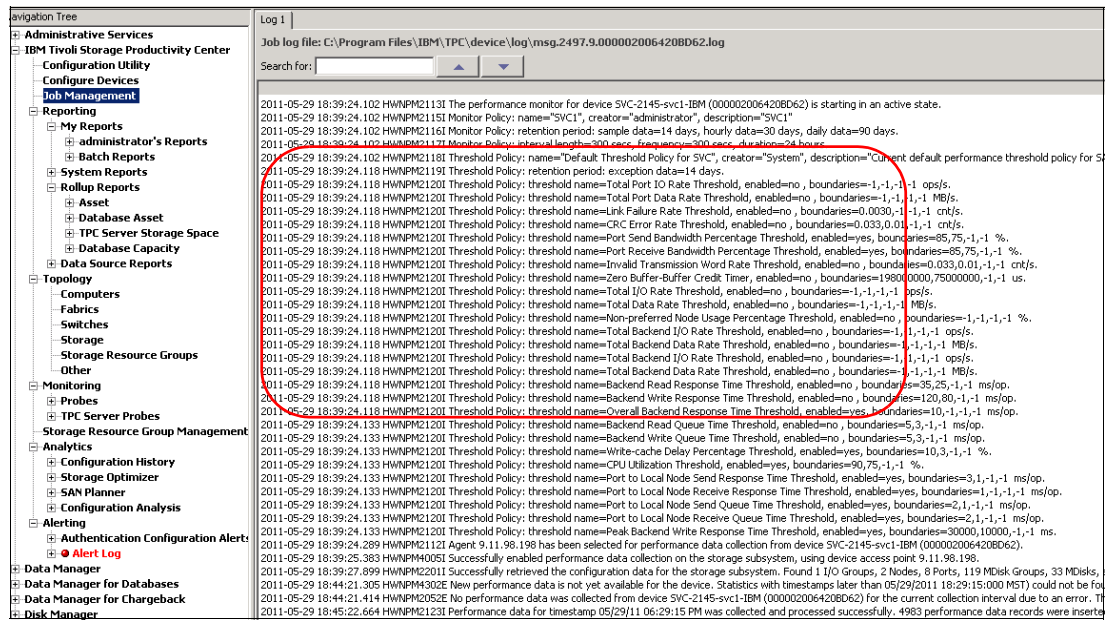


Figure 13-128 SAN Volume Controller constraint threshold enabled

To list all the alerts that occurred, complete the following steps:

1. Click **IBM Tivoli Storage Productivity Center** **Alerting** **Alert Log** **Storage Subsystem**.
2. Look for your SAN Volume Controller subsystem, as shown in Figure 13-129.

Element Management

Navigation Tree

Administrative Services

IBM Tivoli Storage Productivity Center

Configuration Utility

Configure Devices

Job Management

Reporting

Topology

Monitoring

Storage Resource Group Management

Analytics

Alerting

Authentication Configuration Alerts

Alert Log

All

Alerts Directed to administrator

Authentication Configuration

Replication

Storage Subsystem

Computer

Disk

Clear

Delete

Clear All

Delete All

Page Up

Page Down

Refresh

Storage Subsystem	Computer	Disk/Volume	Alert Type	First Triggered
SVC-2145-svc1-IBM	tpcblade3-13.storage.tucson.ibm.com	(N/A)	Overall Backend Response Time Threshold	May 30, 2011 9:20:30 AM
SVC-2145-svc1-IBM	tpcblade3-13.storage.tucson.ibm.com	(N/A)	Overall Backend Response Time Threshold	May 30, 2011 7:15:20 AM
SVC-2145-svc1-IBM	tpcblade3-13.storage.tucson.ibm.com	(N/A)	Overall Backend Response Time Threshold	May 30, 2011 6:15:24 AM
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-13.storage.tucson.ibm.com	(N/A)	Overall Backend Response Time Threshold	May 27, 2011 9:55:14 PM
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-13.storage.tucson.ibm.com	(N/A)	Overall Backend Response Time Threshold	May 30, 2011 1:10:29 AM
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-13.storage.tucson.ibm.com	(N/A)	Overall Backend Response Time Threshold	May 30, 2011 7:40:49 PM
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-13.storage.tucson.ibm.com	(N/A)	Overall Backend Response Time Threshold	May 30, 2011 3:05:59 AM
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-13.storage.tucson.ibm.com	(N/A)	Overall Backend Response Time Threshold	May 30, 2011 3:05:59 AM
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-13.storage.tucson.ibm.com	(N/A)	Overall Backend Response Time Threshold	May 27, 2011 8:15:49 PM
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-13.storage.tucson.ibm.com	(N/A)	Overall Backend Response Time Threshold	May 28, 2011 8:45:14 AM
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-13.storage.tucson.ibm.com	(N/A)	Overall Backend Response Time Threshold	May 30, 2011 3:00:44 AM
Storwize V7000-2076-ford1_tbird-IBM	tpcblade3-13.storage.tucson.ibm.com	(N/A)	Overall Backend Response Time Threshold	May 30, 2011 3:00:44 AM

Figure 13-129 SAN Volume Controller constraints alerts history

3. Click the magnifying glass icon (🔍) that is next to the alert for which you want to see detailed information, as shown in Figure 13-130.

Detail for Alert	
Computer	tpcblade3-13.storage.tucson.ibm.com
Alert Type	Overall Backend Response Time Threshold
State	Active
First Triggered	May 30, 2011 7:15:20 AM
Last Triggered	May 30, 2011 7:15:20 AM
Consecutive Triggers	1
Alert Creator	administrator
Alert Name	SVC and Storwize overall backend
Alert Text	ALR0512W: The Overall Backend Response Time of MDisk mdisk7 in storage subsystem SVC-2145-svc1-IBM was measured to be 10.0 defined critical-stress boundary value of 10 ms/op.

Figure 13-130 Alert details for SAN Volume Controller constraints

For more information about defining alerts, see *SAN Storage Performance Management Using Tivoli Storage Productivity Center*, SG24-7364.

Using the Tivoli Storage Productivity Center 5.2 web-based GUI to monitor Alerts and Threshold Violations

After thresholds or alerts are configured by using the Tivoli Storage Productivity Center 5.2 stand-alone GUI, the web-based GUI can be used for quick access monitoring of them. To access alerts, click **Home** $\varnothing$ **Alerts** on the left side navigation, as shown in Figure 13-131.



Figure 13-131 Accessing Alerts

The Alerts window opens, which displays all of the alerts for all the monitored resources, as shown in Figure 13-132. You can filter or sort the alerts that are displayed (as required) by using the column headings or the Filter field at the upper right of the table.

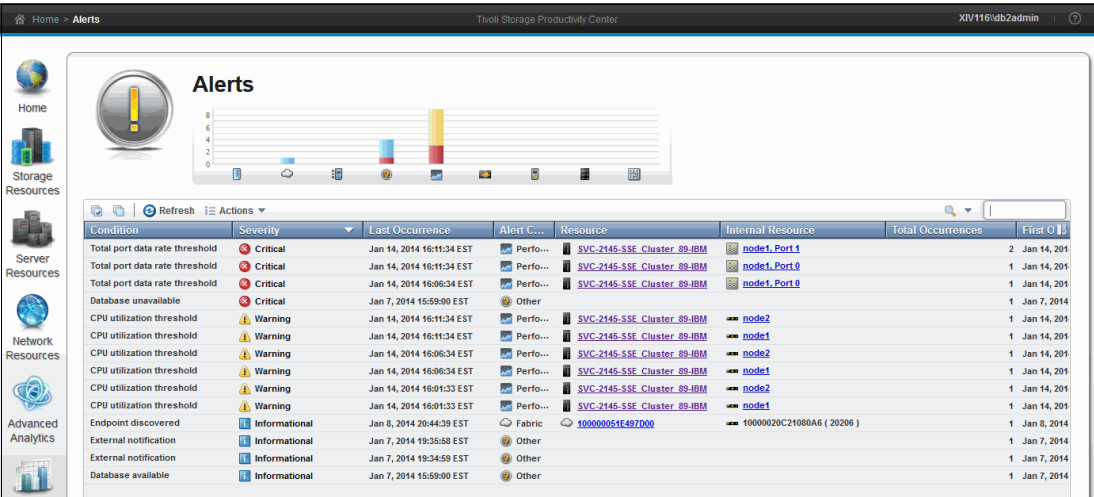


Figure 13-132 Alerts view

Right-click any alert to view, remove, or acknowledge the alert, as shown in Figure 13-133.

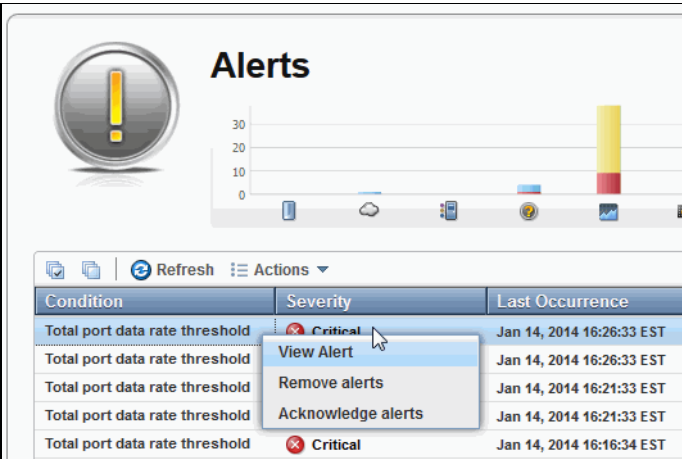


Figure 13-133 Accessing alert details

Select **View Alert** to display the details of the alert in a separate window, as shown in Figure 13-134.

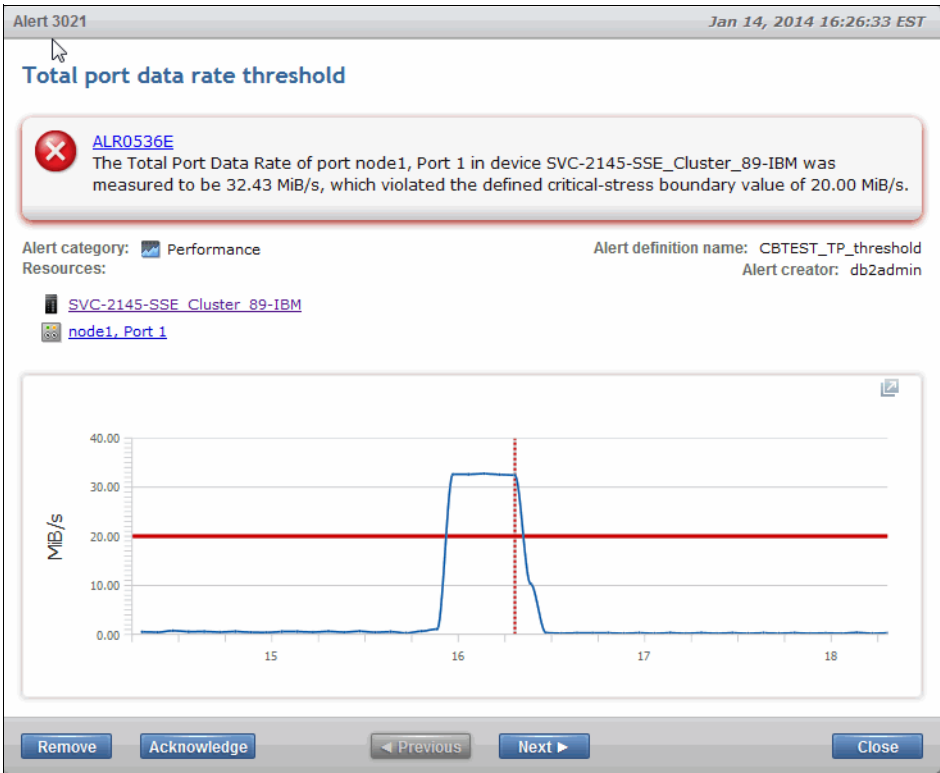


Figure 13-134 Viewing alert details

To monitor threshold violations in the Tivoli Storage Productivity Center 5.2 web GUI, click **Home**  **Performance Monitors** in the left side navigation, as shown Figure 13-135.

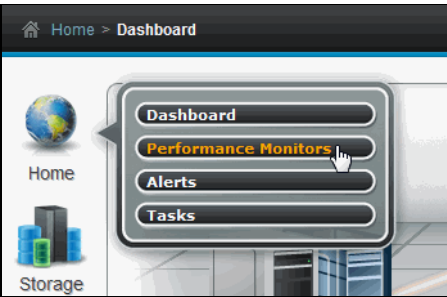


Figure 13-135 Accessing the Performance Monitors view

After the Performance Monitors window opens, click the **Threshold Violations** tab to display all threshold violations for the monitored resources, as shown in Figure 13-136.

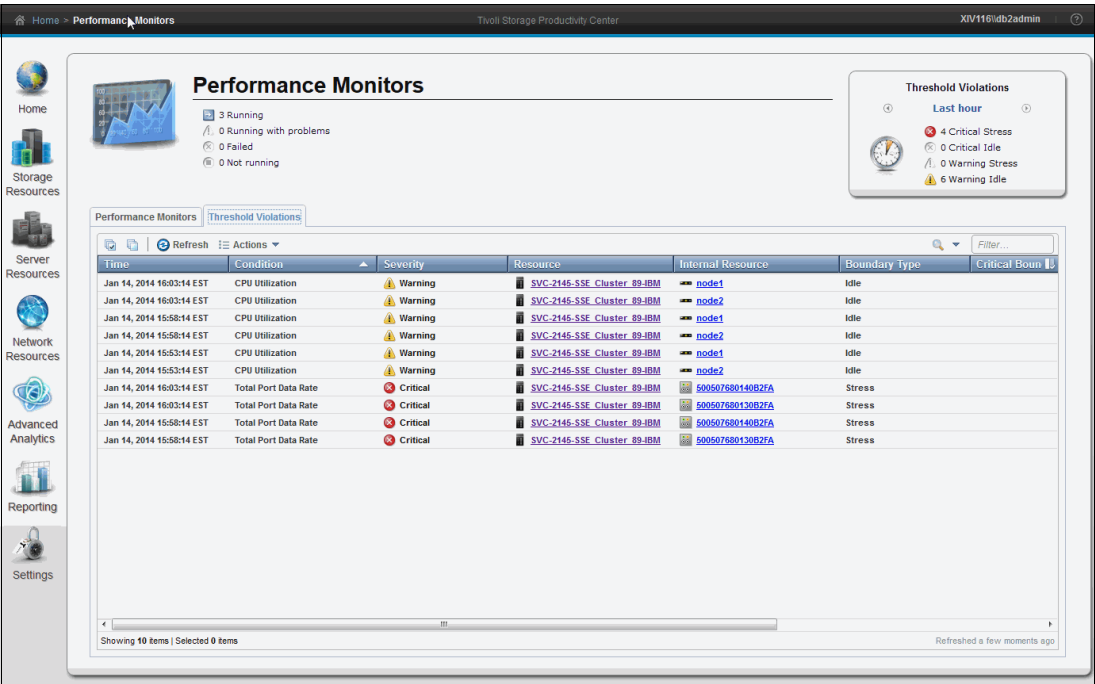


Figure 13-136 Accessing the Threshold Violations view

To view the details of a threshold violation, right-click the corresponding row and choose **View Threshold Violation**, as shown in Figure 13-137 on page 465.

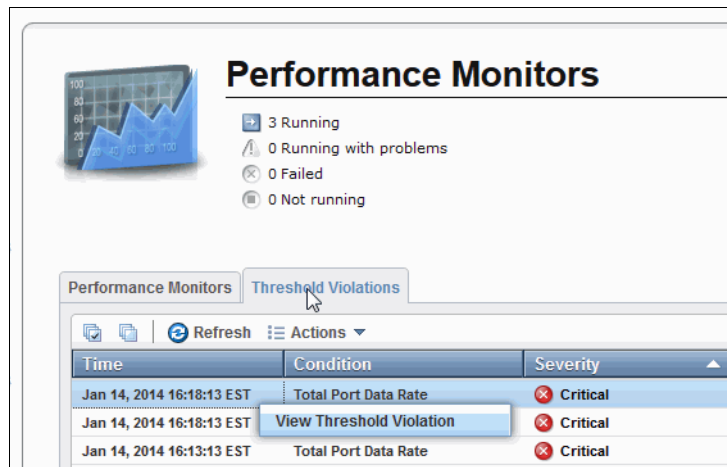


Figure 13-137 Access Threshold Violation details

A window opens that contains the details of the threshold violation, as shown in Figure 13-138.

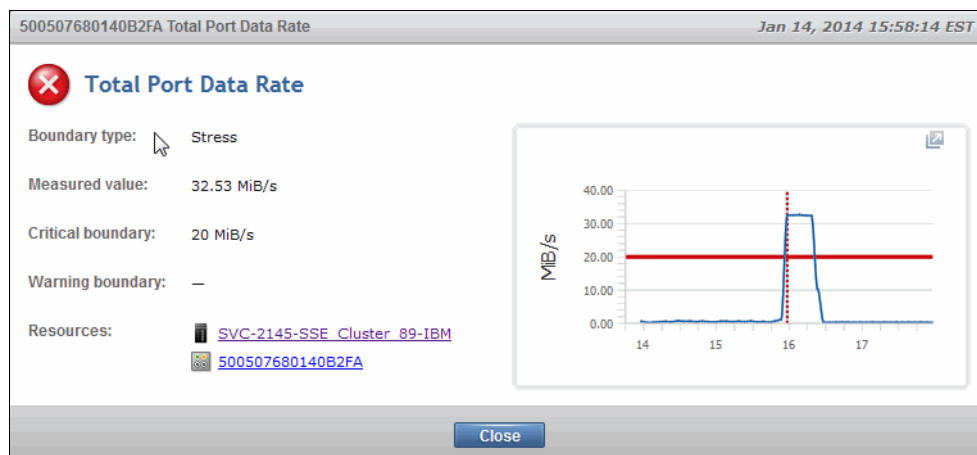


Figure 13-138 Viewing threshold violation details

13.5.5 Monitoring and diagnosing performance problems for a fabric Viewing

This case study tries to find a fabric port bottleneck that exceeds 50% port utilization. We use 50% for lab purposes only for this book.

Tip: In a production environment, a more realistic percentage to monitor is 80% of port utilization.

Ports on the switches in this SAN are 8 Gb. Therefore, a 50% utilization is approximately 400 MBps. To create a performance collection job by specifying filters, complete the following steps:

1. Complete the following steps to specify the filters:
 - a. Click **Fabric Manager** $\varnothing$ **Reporting** $\varnothing$ **Switch Performance** $\varnothing$ **By Port**.
 - b. In the Select tab, in the upper right corner, click **Filter**.

- c. In the Edit Filter window (see Figure 13-139), specify the conditions. In this case study, we specify the following conditions under Column:
- Port Send Data Rate
 - Port Receive Data Rate
 - Total Port Data Rate

Important: In the Records must meet box, you must select **At least one condition** so that the report identifies switch ports that satisfy either filter parameter.

Figure 13-139 Filter for fabric performance reports

2. After you generate this report, identify on the next page by using the Topology Viewer which device is being affected, and identify a possible solution. Figure 13-140 shows the result in our lab.

Switch	Port	WWP	Slot	Index	Time	Interval	Port Send Packet Rate	Port Receive Packet Rate	Total Port
mini	7	200700051E02BDF6	7	7	May 30, 2011 10:07:35 AM	326 s	169,275.56 pkts/s	150,028.48 pkts/s	31

Figure 13-140 Ports exceeding filters set for switch performance report

3. Click the pie chart icon (📊).
4. In the Select Charting Option window, hold down the Ctrl key and select **Port Send Data Rate**, **Port Receive Data Rate**, and **Total Port Data Rate**. Click **OK** to generate the chart.

As shown in Figure 13-141 on page 467, the chart shows a consistent throughput that is higher than 300 MBps in the selected time period. You can change the dates by extending the Limit days settings.

Tip: This chart shows how persistent high utilization is for this port. This consideration is important for establishing the significance and effect of this bottleneck.

Important: To get all the values in the selected interval, remove the defined filters in the Edit Filter window, as shown in Figure 13-139.



Figure 13-141 Data rate of the switch ports

5. Complete the following steps to identify which device is connected to port 7 on this switch:
 - a. Click **IBM Tivoli Storage Productivity Center** $\varnothing$ **Topology**. Right-click **Switches**, and select **Expand all Groups**, as shown in the left pane in Figure 13-142 on page 468.
 - b. Look for your switch, as shown in the right pane in Figure 13-142 on page 468.

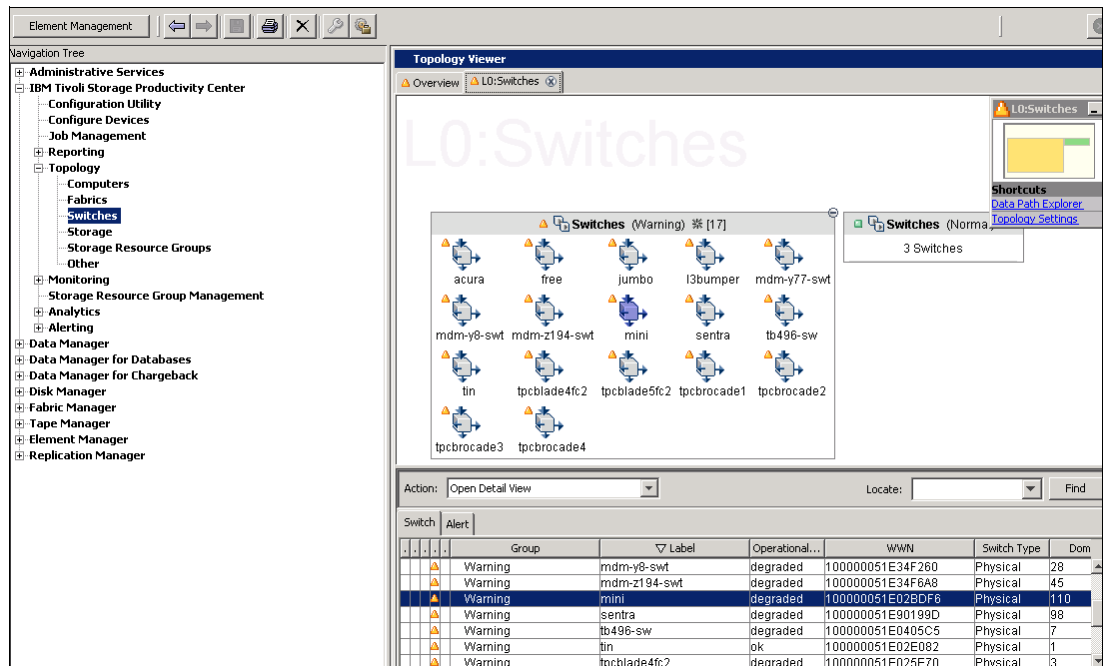


Figure 13-142 Topology Viewer for switches

Tip: To navigate in the Topology Viewer, press and hold the Alt key and the left mouse button to anchor your cursor. When you hold down these keys, you can use the mouse to drag the panel to quickly move to the information you need.

- Find and click **port 7**. The line shows that it is connected to the tpcblade3-7 computer, as shown in Figure 13-143 on page 469. You can see Port details in the tabular view at the bottom of the display. If you scroll to the right, you can also check the Port speed.

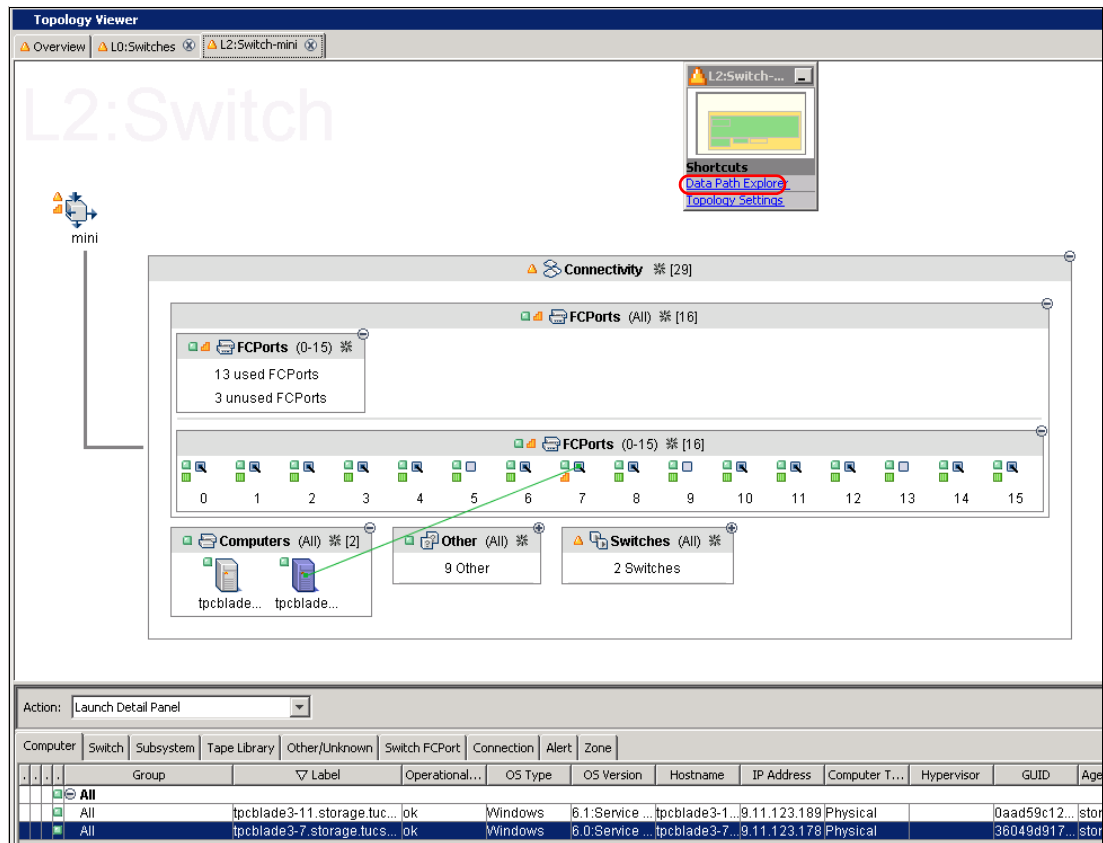


Figure 13-143 Switch port and computer

- d. Double-click the **tpcblade3-7** computer to highlight it. Then, click **Datapath Explorer** (under Shortcuts in the small box at the top of Figure 13-143) to see the paths between the servers and storage subsystems or between storage subsystems. For example, you can get SAN Volume Controller to back-end storage or server to storage subsystem.

The view consists of three panels (host information, fabric information, and subsystem information) that show the path through a fabric or set of fabrics for the endpoint devices, as shown in Figure 13-144 on page 470.

Tip: A possible scenario of using Data Path Explorer is an application on a host that is running slow. The system administrator wants to determine the health status for all associated I/O path components for this application. The system administrator determines whether all components along that path healthy. In addition, the system administrator sees whether there are any component-level performance problems that might be causing the slow application response.

Looking at the data paths for **tpcblade3-7** computer as shown in Figure 13-144 on page 470, you can see that it has a single port HBA connection to the SAN. A possible solution to improve the SAN performance for **tpcblade3-7** computer is to upgrade it to a dual port HBA.

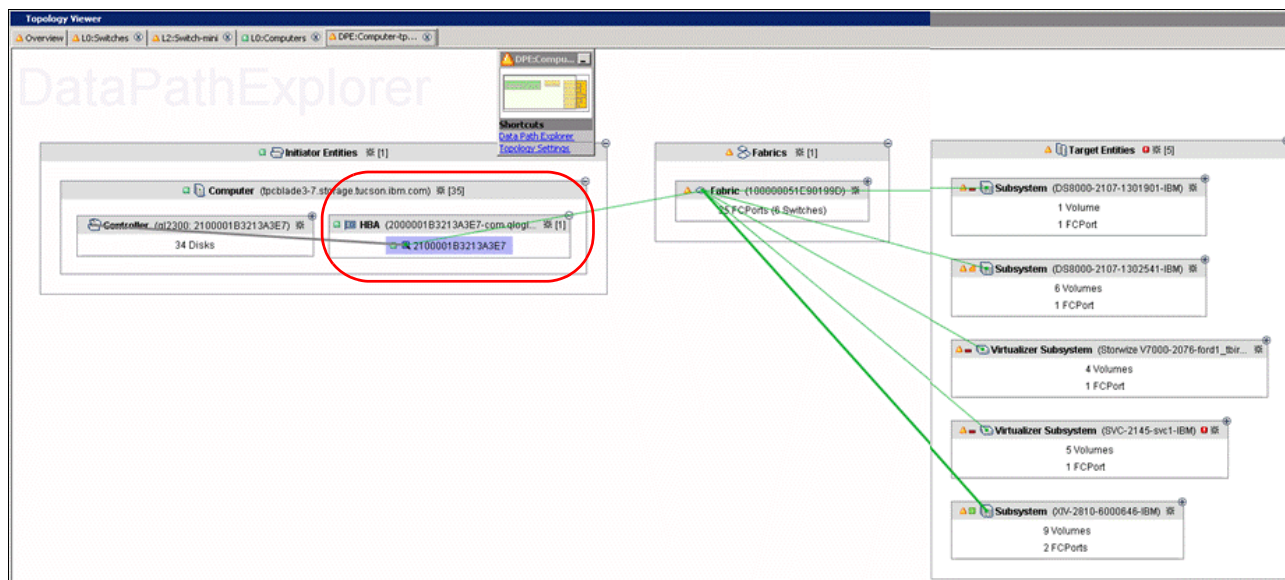


Figure 13-144 Data Path Explorer

13.5.6 Verifying the SAN Volume Controller and Fabric configuration by using Topology Viewer

After Tivoli Storage Productivity Center probes the SAN environment, it automatically builds a graphical display of the SAN environment by using the information from all the SAN components (switches, storage controllers, and hosts). This graphical display is available by using the Topology Viewer option on the Tivoli Storage Productivity Center Navigation Tree.

The information in the Topology Viewer panel is current as of the successful resolution of the last problem. By default, Tivoli Storage Productivity Center probes the environment daily. However, you can run an unplanned or immediate probe at any time.

Tip: If you are analyzing the environment for problem determination, run an ad hoc probe to ensure that you have the latest information about the SAN environment. Make sure that the probe completes successfully.

Ensuring that all SAN Volume Controller ports are online

Information in the Topology Viewer can also confirm the health and status of the SAN Volume Controller and the switch ports. When you look at the Topology Viewer, Tivoli Storage Productivity Center shows a Fibre Channel port with a box next to the worldwide port name (WWPN). If this box has a black line in it, the port is connected to another device. Table 13-2 shows an example of the ports with their connected status.

Table 13-2 Tivoli Storage Productivity Center port connection status

Port view	Status
	This is a port that is connected.
	This is a port that is not connected.

Figure 13-145 shows the SAN Volume Controller ports that are connected and the switch ports.

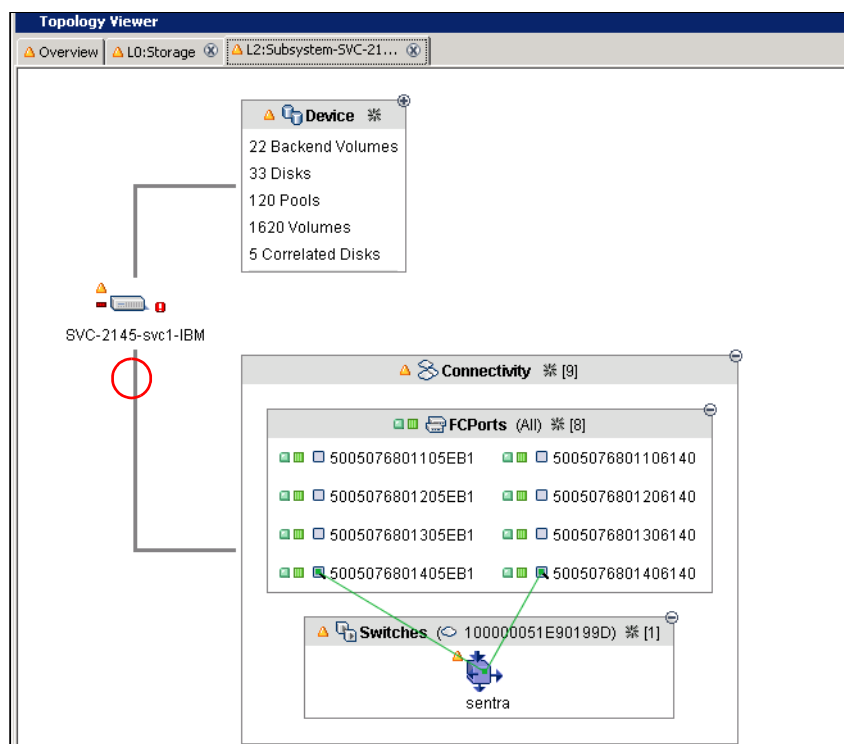


Figure 13-145 SAN Volume Controller connection

Important: Figure 13-145 shows an incorrect configuration for the SAN Volume Controller connections. This configuration is incorrect because it was implemented for lab purposes only. In real environments, each SAN Volume Controller (or Storwize V7000) node port is connected to two separate fabrics. If any SAN Volume Controller (or Storwize V7000) node port is not connected, each node in the cluster displays an error on LCD display. Tivoli Storage Productivity Center also shows the health of the cluster as a warning in the Topology Viewer, as shown in Figure 13-145.

Consider the following points:

- ▶ You have at least one port from each node in each fabric.
- ▶ You have an equal number of ports in each fabric from each node. That is, do not have three ports in Fabric 1 and only one port in Fabric 2 for a SAN Volume Controller (or Storwize V7000) node.

In this example, the connected SAN Volume Controller ports are both online. When a SAN Volume Controller port is not healthy, a black line is shown between the switch and the SAN Volume Controller node.

Tivoli Storage Productivity Center can detect to where the unhealthy ports were connected on a previous probe (which, therefore, were previously shown with a green line). Therefore, the probe discovered that these ports were no longer connected, which resulted in the green line becoming a black line.

If these ports are never connected to the switch, they do not have any lines.

Verifying SAN Volume Controller port zones

When Tivoli Storage Productivity Center probes the SAN environment to obtain information about SAN connectivity, it also collects information about the SAN zoning that is active. The SAN zoning information is also available in the Topology Viewer in the Zone tab.

By going to the Zone tab and clicking the switch and the zone configuration for the SAN Volume Controller, you can confirm that all of SAN Volume Controller node ports are correct in the Zone configuration.

Attention: By default, the Zone tab is not enabled. To enable the Zone tab, you must configure it and turn it on by using the Global Settings. Complete the following steps to access the Global Settings list:

1. Open the Topology Viewer window.
2. Right-click in any white space and select **Global Settings**.
3. In the Global Setting box, select **Show Zone Tab** so that you can see SAN Zoning details for your switch fabrics.

Figure 13-146 shows a SAN Volume Controller node zone that is called SVC_CL1_NODE in our FABRIC-2GBS. We defined this zone and correctly included all of the SAN Volume Controller node ports.

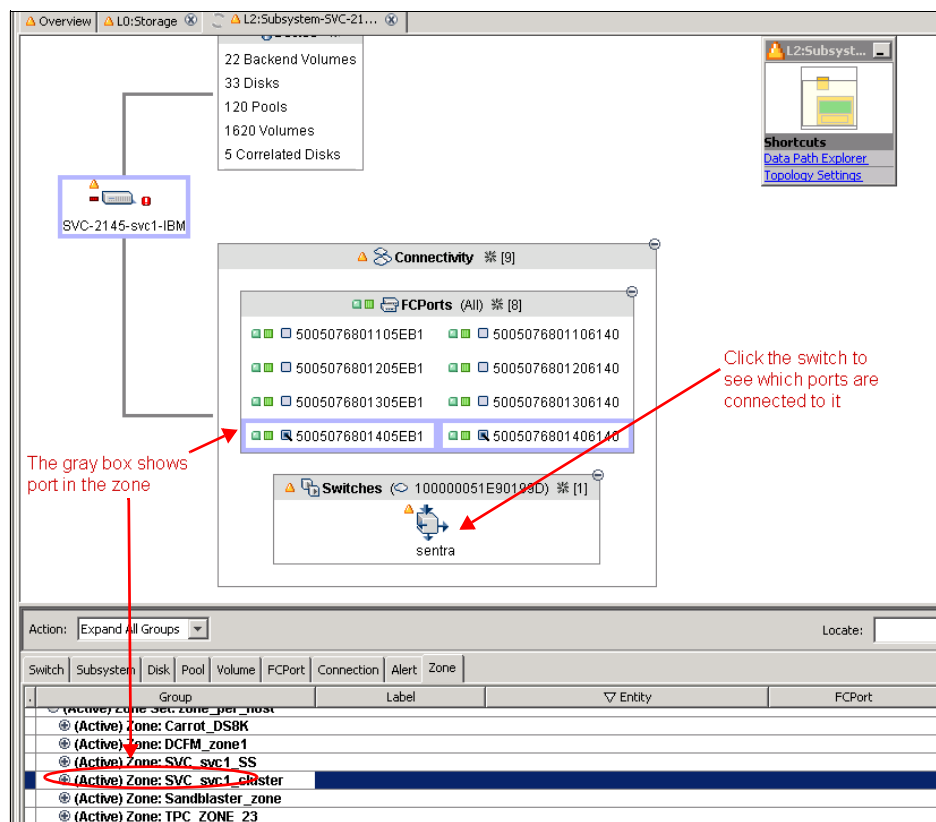


Figure 13-146 SAN Volume Controller zoning in the Topology Viewer

Verifying paths to storage

You can use the Data Path Explorer functions in the Topology Viewer to see the path between two objects. They also show the objects and the switch fabric in one view.

By using Data Path Explorer, you can see, for example, that mdisk1 in Storwize V7000-2076-ford1_tbird-IBM is available through two Storwize V7000 ports. You can trace that connectivity to its logical unit number (LUN) rad (ID:009f) as shown in Figure 13-147.



Figure 13-147 Topology Viewer: Data Path Explorer

In addition, you can hover over the MDisk, LUN, and switch ports (not shown in Figure 13-147) and get health and performance information about these components. This way, you can verify the status of each component to see how well it is performing.

Verifying the host paths to the Storwize V7000

By using the computer display in Tivoli Storage Productivity Center, you can see all of the fabric and storage information for the computer that you select. Figure 13-148 on page 474 shows the host tpcb1ade3-11, which has two HBAs, but only one is active and connected to the SAN. This host was configured to access some Storwize V7000 storage, as you can see in the upper-right part of the panel.

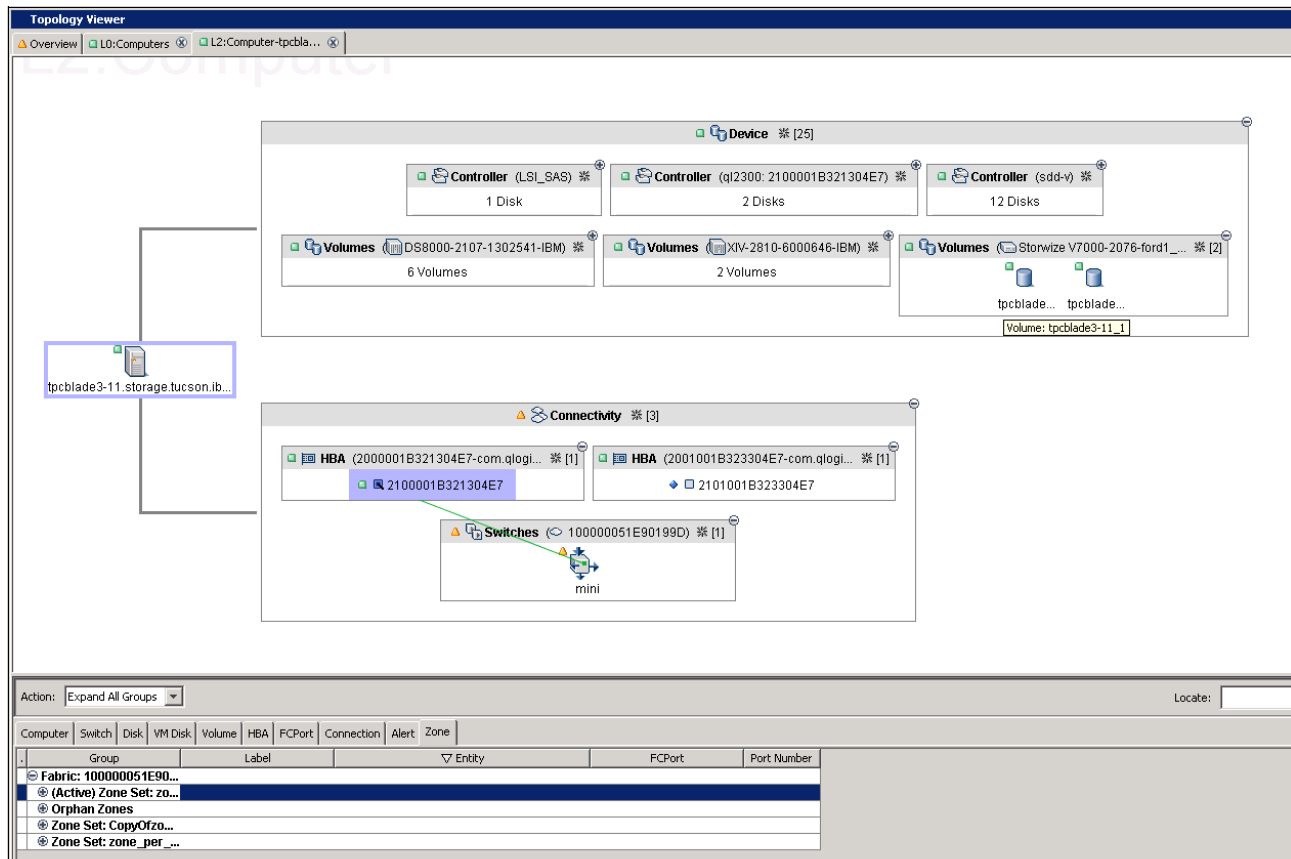


Figure 13-148 tpclblade3-11 with only one active HBA

The Topology Viewer shows that tpclblade3-11 is physically connected to a single fabric. By using the Zone tab, you can see the single zone configuration that is applied to tpclblade3-11 for the 100000051E90199D zone. Therefore, tpclblade3-11 does not have redundant paths, and if the mini switch goes offline, tpclblade3-11 loses access to its SAN storage.

By clicking the zone configuration, you can see which port is included in a zone configuration and which switch has the zone configuration. The port that has no zone configuration is not surrounded by a gray box.

You can also use the Data Path Viewer in Tivoli Storage Productivity Center to check and confirm path connectivity between a disk that an operating system detects and the VDisk that the Storwize V7000 provides. Figure 13-149 on page 475 shows the path information that relates to the tpclblade3-11 host and its VDisks. You can hover over each component to also get health and performance information (not shown), which might be useful when you perform problem determination and analysis.



Figure 13-149 Viewing host paths to the Storwize V7000

13.5.7 Verifying the SAN Volume Controller and Fabric configuration by using the Tivoli Storage Productivity Center 5.2 web-based GUI Data Path tools

The Tivoli Storage Productivity Center 5.2 web-based GUI includes a set of Data Path tools to help verify and monitor your environment. The tools include a Topology View, with which you can view your environment elements in Tree, Hierarchical, or customized views, and a Table view, with which you can view your environment elements in a tabular view. To access the Data Path tools, click **Data Path** under the General section, as shown in Figure 13-150.

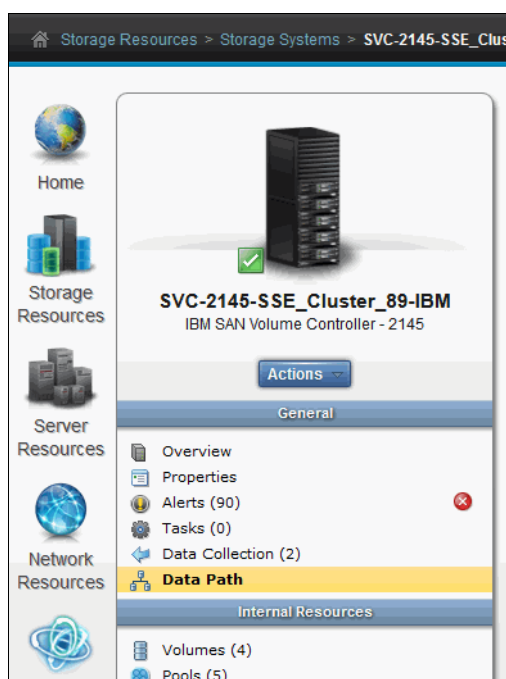


Figure 13-150 Accessing the Data Path tools for a storage system

Topology View window opens for this storage system, which displays the connected elements for this storage system in a tree view format, as shown in Figure 13-151 on page 476.

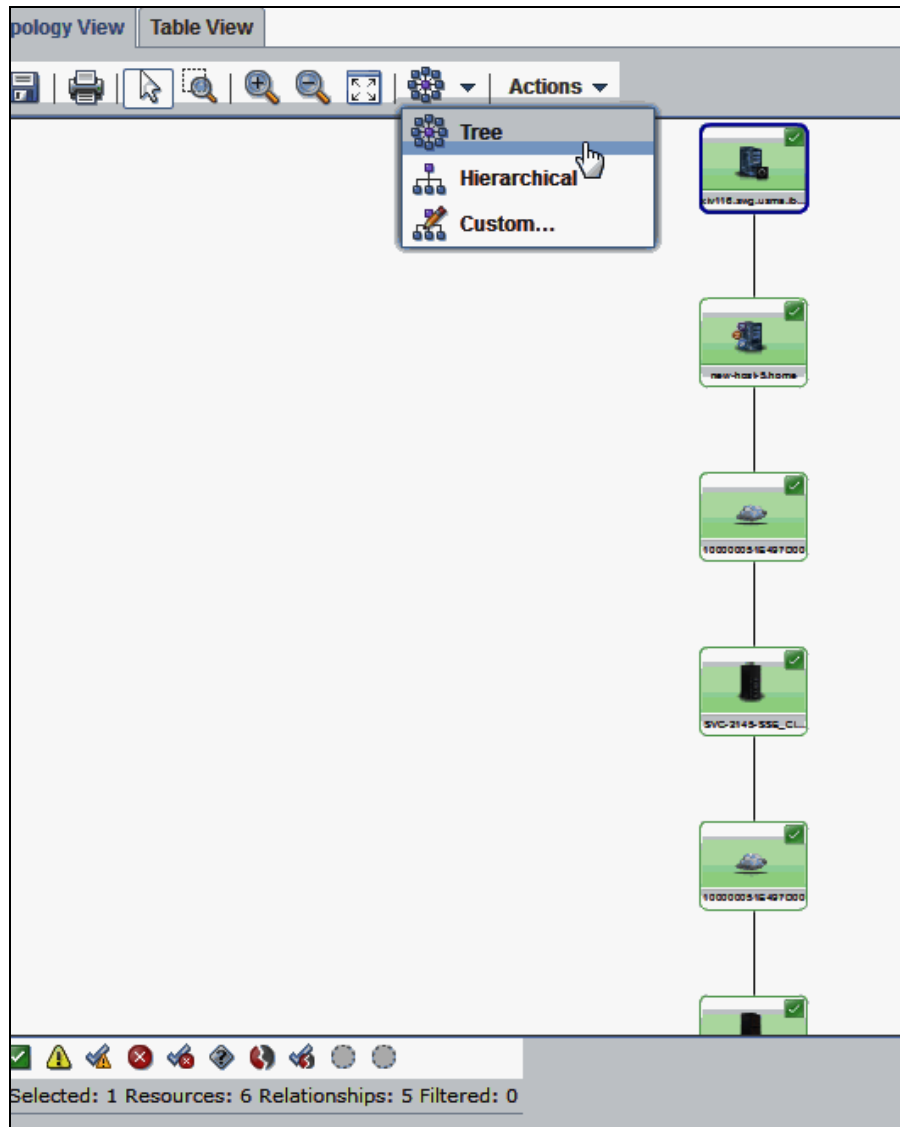


Figure 13-151 The Pools page for a storage system

The elements are displayed with icons that denote their status. With your mouse, point to the icons at the lower left of the window to see captions that describe each icon. To change the display method, use the drop-down menu, as shown in Figure 13-151. To view the details or properties of an element, right-click the element in the window, or select the element and use the Actions menu. To view the elements in a tabular format, click the **Table View** tab at the top of the window.

Note: In Tivoli Storage Productivity Center version 5.2, the Data Path tools provide a limited amount of information and functionality. For in-depth troubleshooting, the Topology Viewer that is described in 13.5.6, “Verifying the SAN Volume Controller and Fabric configuration by using Topology Viewer” on page 470, might provide greater capabilities.

13.6 Monitoring in real time by using the SAN Volume Controller or Storwize V7000 GUI

By using the SAN Volume Controller or Storwize V7000 GUI, you can monitor CPU usage, volume, interface, and MDisk bandwidth of your system and nodes. You can use system statistics to monitor the bandwidth and IOPS of all the volumes, interfaces, and MDisks that are being used on your system. You can also monitor the system and compression CPU utilization for the system.

By using the GUI, you also can view these metrics for a single node. 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. Complete the following steps to start the performance monitor:

1. Start your GUI session by entering the following address in a web browser:
`https://<system ip address>/`
2. Select **Monitoring** $\varnothing$ **Performance**, as shown in Figure 13-152.

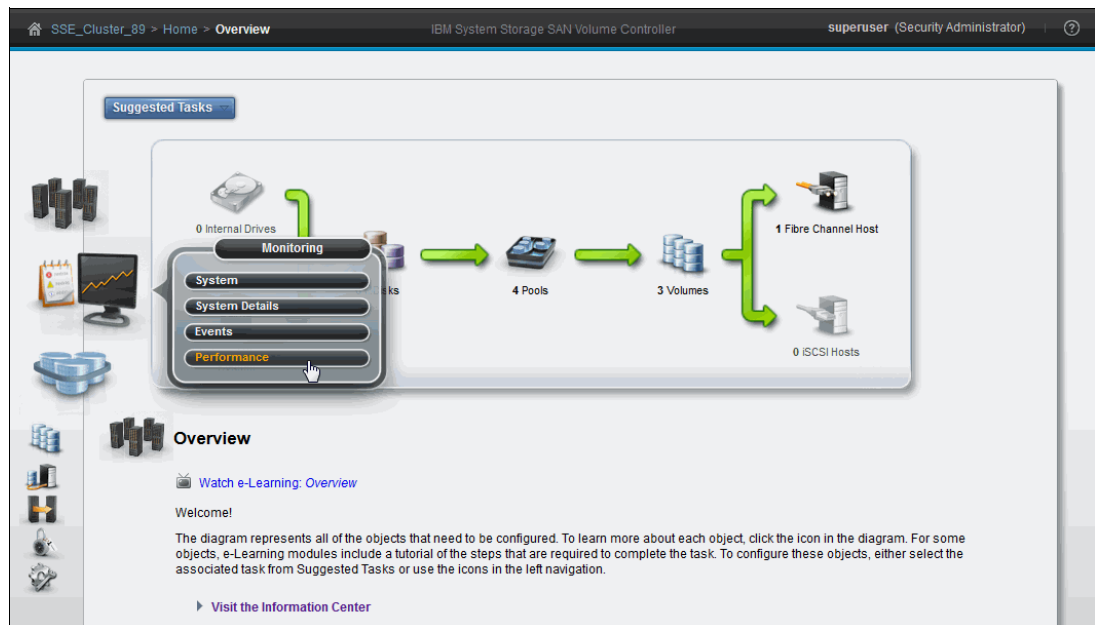


Figure 13-152 Starting the performance monitor panel

The performance monitor panel (as shown in Figure 13-153 on page 478) presents the graphs in the following four quadrants:

- ▶ The upper left quadrant is the CPU utilization percentage for system CPUs and compression CPUs.
- ▶ The upper right quadrant is volume throughput in MBps and current volume latency.
- ▶ The lower left quadrant is the interface throughput (FC, SAS, iSCSI, and IP Remote Copy).
- ▶ The lower right quadrant is MDisk throughput in MBps, current MDisk latency.

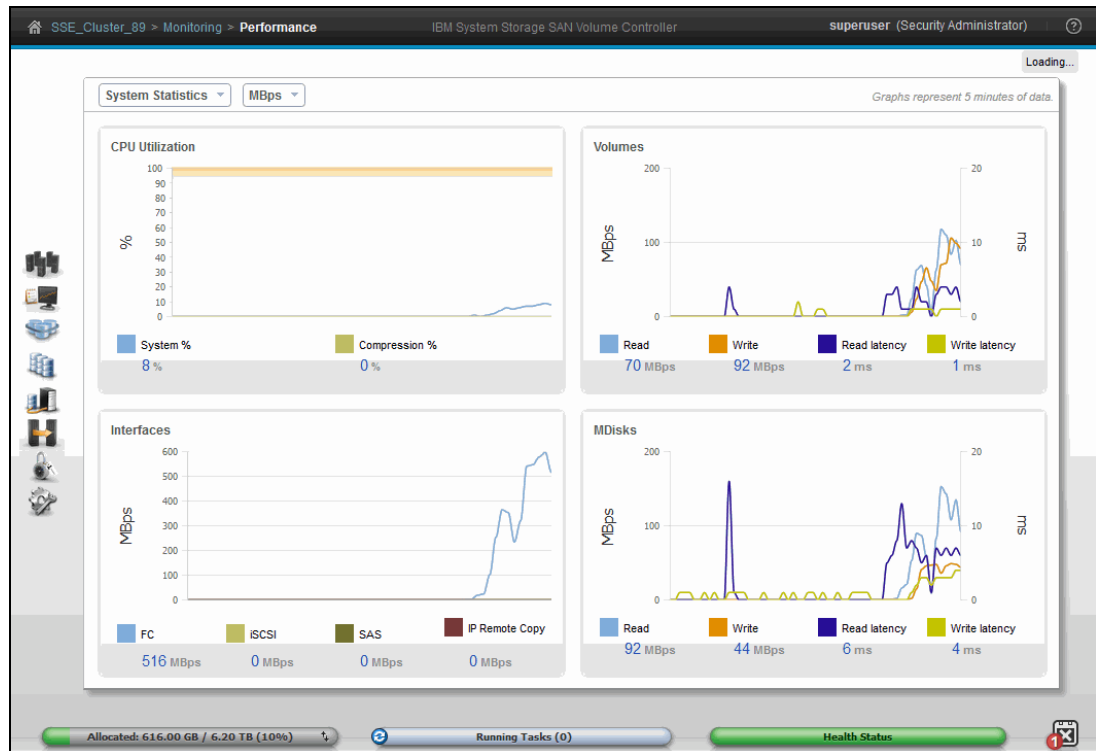


Figure 13-153 Performance monitor panel in the SAN Volume Controller 7.2 web GUI

To toggle the charts to show IOPS instead of throughput (MBps), select **IOPS** in the drop-down menu above the CPU Utilization chart, as shown in Figure 13-154.

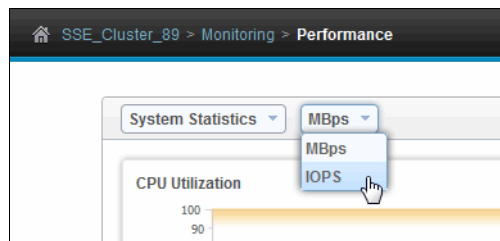


Figure 13-154 Toggling between MBps and IOPS in the performance monitor panel

Each graph represents 5 minutes of collected statistics and provides a means of assessing the overall performance of your system. For example, CPU utilization shows the current percentage of CPU usage and specific data points on the graph that show peaks in utilization.

With this real-time performance monitor, you can quickly view bandwidth or IOPS of volumes, interfaces, and MDisks. Each graph shows the current bandwidth or IOPS and a view of the value over the last 5 minutes. Each data point can be accessed to determine its individual bandwidth utilization or IOPS, and to evaluate whether a specific data point might represent performance impacts. For example, you can monitor the interfaces, such as Fibre Channel or SAS, to determine whether the host data-transfer rate is different from the expected rate. The volumes and MDisk graphs also show the IOPS and latency values.

To remove a specific metric from a graph, click the corresponding item below the graph, as shown in Figure 13-155 on page 479. Click it again to add it back to the graph. Displayed metrics are denoted by a solid color square. Hidden metrics are denoted by a hollow colored square.

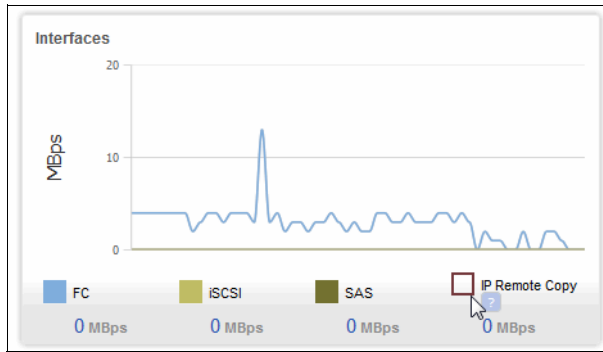


Figure 13-155 Toggling the display of graph metrics in the performance monitor panel

In the left drop-down menu above the CPU Utilization chart, you can switch from system statistic to statistics by node, and select a specific node to get its real-time performance graphs, as shown in Figure 13-156.

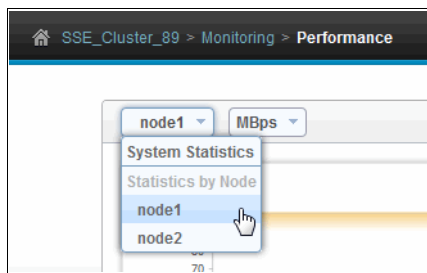


Figure 13-156 Displaying node level metrics in the performance monitor panel

Viewing node level performance statistics can help identify an unbalanced usage of your system nodes.

13.7 Manually gathering SAN Volume Controller statistics

SAN Volume Controller collects three types of statistics: MDisk, VDisk, and node statistics. The statistics are collected on a per-node basis, which means that the statistics for a VDisk are for its usage by using that particular node. In SAN Volume Controller V6 code, you do not need to start the statistics collection if it is already enabled by default.

The **lssystem** command shows the **statistics_status**. The default **statistic_frequency** is 15 minutes, which you can adjust by using the **startstats -interval <minutes>** command.

For each collection interval, the SAN Volume Controller creates the following statistics files:

- ▶ Nm_stat file for MDisk
- ▶ Nv_stats file for VDisks
- ▶ Nn_stats file for nodes

The files are written to the **/dumps/iostats** directory on each node. A maximum of 16 files of each type can be created for the node. When the 17th file is created, the oldest file for the node is overwritten.

To retrieve the statistics files from the nonconfiguration nodes, copy them beforehand onto the configuration node by using the following command:

```
cpdumps -prefix /dumps/iostats <non_config node id>
```

To retrieve the statistics files from the SAN Volume Controller, you can use the secure copy (**scp**) command, as shown in the following example:

```
scp -i <private key file> admin@clustername:/dumps/iostats/* <local destination dir>
```

If you do not use Tivoli Storage Productivity Center, you must retrieve and parse these XML files to analyze the long-term statistics. The counters on the files are posted as absolute values. Therefore, the application that processes the performance statistics must compare two samples to calculate the differences from the two files.

An easy way to gather and store the performance statistic data and generate graphs is to use the **svcmmon** command. This command collects SAN Volume Controller and Storwize V7000 performance data every 1 - 60 minutes. Then, it creates the spreadsheet files (in CSV format) and graph files (in GIF format). By using a database, the **svcmmon** command manages SAN Volume Controller and Storwize V7000 performance statistics from minutes to years.

For more information about the **svcmmon** command, see “SVC / Storwize V7000 Performance Monitor - svcmmon” in IBM developerWorks at this website:

<https://www.ibm.com/developerworks/mydeveloperworks/blogs/svcmmon>

Disclaimer: The **svcmmon** command is a set of Perl scripts that were designed and programmed by Yoshimichi Kosuge, personally. It is not an IBM product and it is provided without any warranty. Therefore, you use **svcmmon** at your own risk.

The **svcmmon** command works in online mode or stand-alone mode, which is described briefly in this section. The package is well-documented to run on Windows or Linux workstations. For other platforms, you must adjust the **svcmmon** scripts.

For a Windows workstation, you must install the ActivePerl, PostgreSQL, and the Command-Line Transformation Utility (**msxsl.exe**). PuTTY is required if you want to run in online mode. However, even in stand-alone mode, you might need it to secure copy the `/dumps/iostats/` files and the `/tmp/svc.config.backup.xml` files. You might also need it to access the SAN Volume Controller from a command line. Follow the installation guide about the **svcmmon** command on the following IBM developerWorks blog website:

<https://www.ibm.com/developerworks/mydeveloperworks/blogs/svcmmon>

To run **svcmmon** in stand-alone mode, you must convert the `.xml` configuration backup file into HTML format by using the **svconfig.pl** script. Then, you must copy the performance files to the **iostats** directory and create the **svcmmon** database by using **svcdb.pl --create** or populate the database by using **svcperf.pl --offline**. The last step is report generation, which you run with the **svcreport.pl** script.

The reporting functionality generates multiple GIF files per object (MDisk, VDisk, and node) with aggregated CSV files. By using the CSV files, we can generate customized charts that are based on spreadsheet functions, such as Pivot Tables or DataPilot and search (xLOOKUP) operations. The backup configuration file that is converted in HTML is a good source to create another spreadsheet tab to relate, for example, VDisks with their I/O group and preferred node.

Figure 13-157 shows a spreadsheet chart that was generated from the <system_name>__vdisk.csv file that was filtered for I/O group 2. The VDisks for this I/O group were selected by using a secondary spreadsheet tab that was populated with the VDisk section of the configuration backup HTML file.

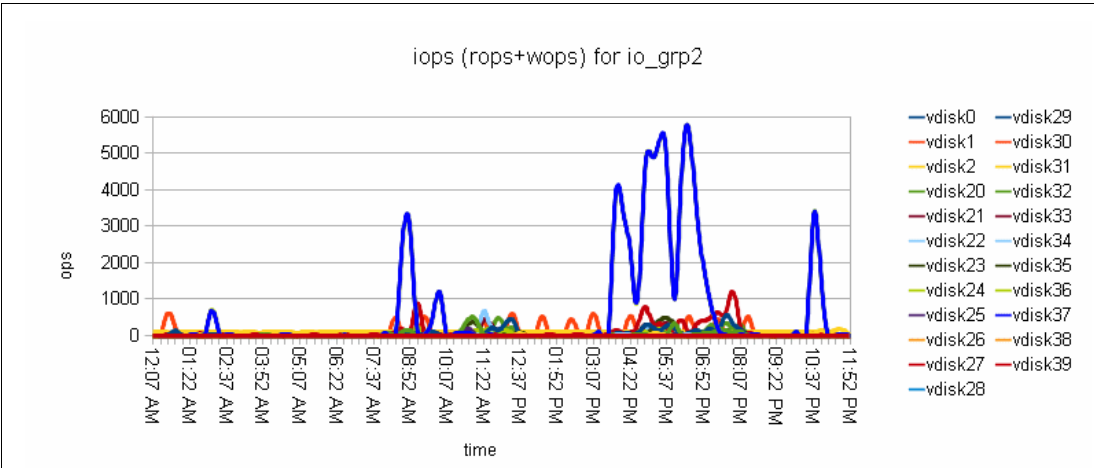


Figure 13-157 Total operations per VDisk for I/O group 2, where Vdisk37 is the busiest volume

By default, the **svcreport.pl** script generates GIF charts and CSV files with one hour of data. The CSV files aggregate a large amount of data, but the GIF charts are presented by VDisk, MDisk, and node as described in Table 13-3.

Table 13-3 Spreadsheets and GIF chart types that are produced by svcreport

Spreadsheets (CSV)	Charts per VDisk	Charts per MDisk	Charts per node
cache_node	cache.hits	mdisk.response.worst.resp	cache.usage.node
cache_vdisk	cache.stage	mdisk.response	cpu.usage.node
cpu	cache.throughput	mdisk.throughput	N/A
drive	cache.usage	mdisk.transaction	N/A
MDisk	vdisk.response.tx	N/A	N/A
node	vdisk.response.wr	N/A	N/A
VDisk	vdisk.throughput	N/A	N/A
N/A	vdisk.transaction	N/A	N/A

To generate a 24-hour chart, specify the **--for 1440** option. The **-for** option specifies the time range by minute for which you want to generate SAN Volume Controller/Storwize V7000 performance report files (CSV and GIF). The default value is 60 minutes.

Figure 13-158 on page 482 shows a chart that was automatically generated by the **svcperf.pl** script for the vdisk37. This chart, which is related to vdisk37, is shown if the chart in Figure 13-157 shows that VDisk is the one that reaches the highest IOPS values.

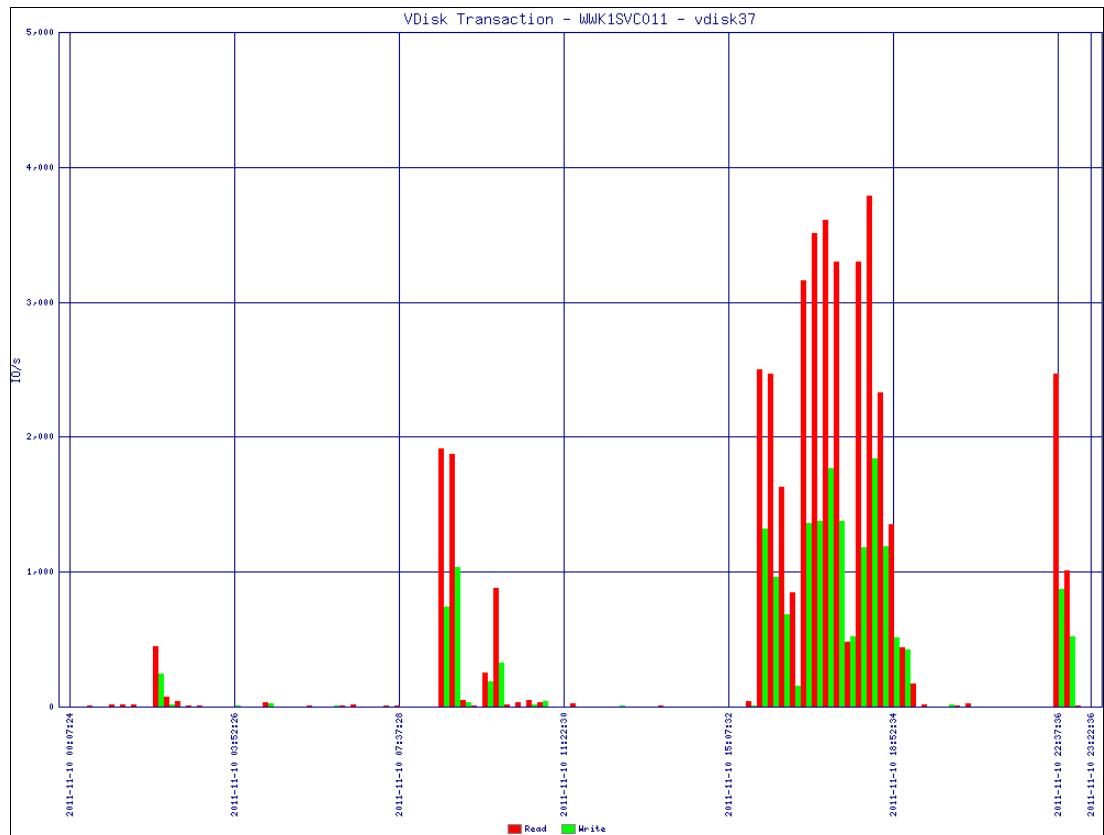


Figure 13-158 Number of read and write operations for vdisk37

The **svcmmon** command is not intended to replace Tivoli Storage Productivity Center. However, it helps when Tivoli Storage Productivity Center is not available, allowing an easy interpretation of the SAN Volume Controller performance XML data.

Figure 13-159 shows the read/write throughput for vdisk37 in bytes per second.

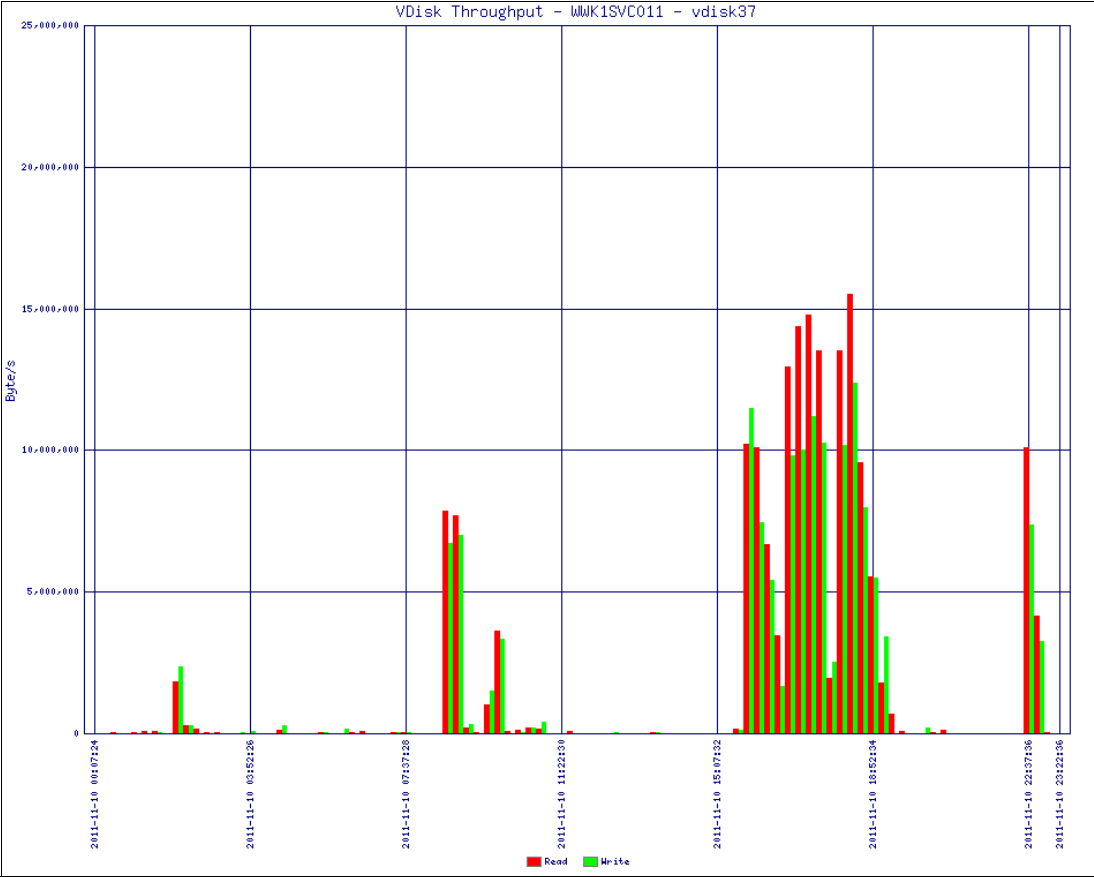


Figure 13-159 Read/write throughput for vdisk37 in bytes per second



Maintenance

Among the many benefits that the IBM System Storage SAN Volume Controller and Storwize family products provide 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 SAN Volume Controller/Storwize family products. 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 SAN Volume Controller and Storwize administration so that you can use this chapter as a checklist. It also provides and tips and guidance. For practical examples of the procedures that are described here, see Chapter 16, "SAN Volume Controller scenarios" on page 555.

Important: The practices that are described here were effective in many SAN Volume Controller/Storwize 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 disk 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 the other. 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:

- ▶ Automating the documentation for SAN Volume Controller/Storwize and SAN environment
- ▶ Storage management IDs
- ▶ Standard operating procedures
- ▶ SAN Volume Controller/Storwize code upgrade
- ▶ SAN modifications
- ▶ Hardware upgrades for SAN Volume Controller
- ▶ Adding expansion enclosures to Storwize family systems
- ▶ More information

14.1 Automating the documentation for SAN Volume Controller/Storwize and SAN environment

This section focuses on the challenge of automating the documentation that is needed for a SAN Volume Controller and Storwize 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 SAN Volume Controller/Storwize 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 SAN Volume Controller/Storwize 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 SAN Volume Controller/Storwize, use of these forms might also help you document how your SAN Volume Controller/Storwize was first configured.

Minimum documentation is needed for a SAN Volume Controller/Storwize 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.

14.1.1 Naming conventions

Whether you are creating your SAN and SAN Volume Controller/Storwize 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 SAN Volume Controller/Storwize 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 SAN and SAN Volume Controller component names limit the number and type of characters you can use. For example, SAN Volume Controller/Storwize names are limited to 63 characters, which makes creating a naming convention a bit easier than in previous versions of SAN Volume Controller/Storwize code.

Names: In previous versions of SAN Volume Controller/Storwize code, names were limited to 15 characters. Starting with version 7.x, the limit is 63 characters.

Many names in SAN storage and in the SAN Volume Controller/Storwize 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 488.) 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

SAN Volume Controller names the storage controllers `controllerX`, with `X` being a sequential decimal number. If multiple controllers are attached to your SAN Volume Controller, 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 SAN Volume Controller to know which storage controller to check.

Note: SAN Volume Controller/Storwize detects controllers that are based on their WWNN. If you have a storage controller that has one WWNN per one WWPN this might lead to many `controllerX` names pointing to same physical box. In this case, you should prepare a naming convention to cover this situation.

MDisks and storage pools

When the SAN Volume Controller 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 SAN Volume Controller 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 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 LLAATRFFNN, 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 5005076801PXXXXX).

Be mindful of the SAN Volume Controller port aliases. The 11th digit of the port WWPN (P) reflects the SAN Volume Controller node FC port, but not directly, as listed in Table 14-1.

Table 14-1 WWPNs for the SAN Volume Controller node ports

Value of P	SAN Volume Controller physical port
4	1
3	2
1	3
2	4
0	None: SAN Volume Controller Node WWNN

- ▶ *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)
- ▶ *svcccluster1_svcccluster2* (for remote copy services)

14.1.2 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 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 email.

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 data collection tool, you must explore the options that are provided to learn how to create a well-organized and useful diagram. Figure 14-1 shows an example of a poorly formatted diagram.

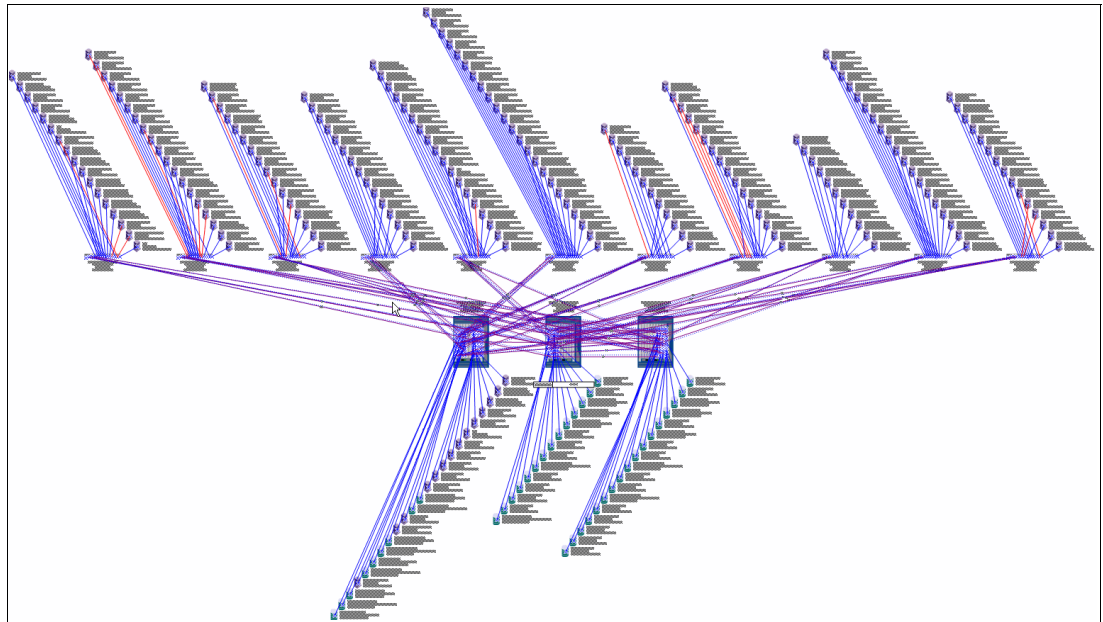


Figure 14-1 A poorly formatted SAN diagram

Figure 14-2 shows a 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.

The screenshot displays the 'SAN Health Options' window, specifically the 'Report Format' tab. The window has a tabbed interface with the following tabs: 'General Options', 'Switch Diagnostics', 'Device Names', 'Report Format' (selected), 'Diagram Format', and 'Audit Data File Upload'. The 'Report Format' tab is divided into two main sections: 'Excel Report Content' and 'Excel Report Formatting Options'.

Under 'Excel Report Content', there are several options:

- ☐ 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

Under 'Excel Report Formatting Options', there are:

- ☐ Format Pages For Letter Size Paper
- ☒ Format Pages For A4 Size Paper
- ☐ Add an image or company logo to the report page footer

Below these options, there is a text field for 'File Name (*.bmp, *.gif, *.jpg)' with a 'Browse' button next to it. Below the text field, it says 'Recommended Image Size Width 180 x Height 60 Pixels'. At the bottom of the window, there are two buttons: 'Reset To Default' and 'Done'.

Figure 14-2 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 data collection 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.

Tivoli Storage Productivity Center reporting

If you have Tivoli Storage Productivity Center running in your environment, you can use it to generate reports on your SAN. For more information about how to configure and schedule Tivoli Storage Productivity Center reports, see the Tivoli Storage Productivity Center documentation.

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.

14.1.3 SAN Volume Controller and Storwize family products

For the SAN Volume Controller and Storwize, periodically collect (at a minimum) the output of the following commands:

- ▶ `svcinfo lsfabric`
- ▶ `svcinfo lssystem`
- ▶ `svcinfo lsmdisk`
- ▶ `svcinfo lsmdiskgrp`
- ▶ `svcinfo lsvdisk`
- ▶ `svcinfo lshost`
- ▶ `svcinfo lshostvdiskmap X` (with *X* ranging to all defined host numbers in your SAN Volume Controller cluster)

Import the commands into a spreadsheet, preferably with each command output on a separate sheet.

You might also want to store the output of more commands; for example, if you have SAN Volume Controller/Storwize Copy Services that are configured or there are dedicated managed disk groups to specific applications or servers.

One way to automate this task is to first create a batch file (Windows) or shell script (UNIX or Linux) that runs these commands and stores their output in temporary files. Then, use spreadsheet macros to import these temporary files into your SAN Volume Controller/Storwize documentation spreadsheet.

When you are gathering SAN Volume Controller/Storwize information, consider following preferred practices:

- ▶ With MS Windows, use the PuTTY **p1nk** utility to create a batch session that runs these commands and stores their output. With UNIX or Linux, you can use the standard SSH utility.

- ▶ Create a SAN Volume Controller user with the Monitor privilege to run these batches. Do not grant it Administrator privilege. Create and configure an SSH key specifically for it.
- ▶ Use the `-delim` option of these commands to make their output delimited by a character other than Tab, such as comma or colon. By using a comma, you can import the temporary files into your spreadsheet in CSV format.
- ▶ To make your spreadsheet macros simpler, you might want to preprocess the temporary output 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 edition utility.

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 14.1.7, “Automated support data collection” on page 494.

14.1.4 Storage

Fully allocate all of the available space in the storage controllers that you use in its back end to the SAN Volume Controller/Storwize. This way, you can perform all your Disk Storage Management tasks by using SAN Volume Controller/Storwize. 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 upgrades. As such, there is little point to automating this back-end storage controller documentation. The same applies to the Storwize internal disk drives and enclosures.

However, if you use split controllers, this option might not be the best option. The portion of your storage controllers that is used outside SAN Volume Controller/Storwize 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.

14.1.5 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.

14.1.6 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 SAN Volume Controller/Storwize 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.

14.1.7 Automated support data collection

In addition to the easier-to-use documentation of your SAN Volume Controller/Storwize and SAN Storage environment, collect and store for some time the configuration files and technical support data collection for all your SAN equipment. Such information includes the following items:

- ▶ The **supportSave** and **configSave** files on Brocade switches
- ▶ Output of the **show tech-support details** command on Cisco switches
- ▶ Data collections on the IBM Network Advisor Software (formerly known as Data Center Fabric Manager)
- ▶ SAN Volume Controller **snap**
- ▶ DS4x00 subsystem profiles
- ▶ The following DS8x00 LUN inventory commands:
 - **lsfbvol**
 - **lshostconnect**
 - **lsarray**
 - **lsrank**
 - **lsioports**
 - **lsvolgrp**

You can create procedures that automatically create and store this data on scheduled dates, delete old data, or transfer the data to tape.

14.1.8 Subscribing to SAN Volume Controller/Storwize support

Subscribing to SAN Volume Controller/Storwize 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>

Select the products that you want to be notified about. You can use the same IBM ID that you created to access the Electronic Service Call web page (ESC+) at this website:

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

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.

14.2 Storage management IDs

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 reason might be because of a lack of time, forgetfulness, or because 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.

In addition to authenticating with private/public SSH key, SAN Volume Controller and Storwize family systems support the following authentication methods:

- ▶ Local authentication by using password
- ▶ 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 Reporting. 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.

Figure 14-3 on page 496 shows the SAN Volume Controller V7.2 GUI user ID creation window.

Figure 14-3 Creating user ID by using the GUI

14.3 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 SAN Volume Controller/Storwize environment healthy and reliable. For practical examples, see Chapter 16, “SAN Volume Controller scenarios” on page 555.

14.3.1 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 SAN Volume Controller/Storwize. For example, in AIX, remove the `hdisk` from the volume group (`reducevg`) and delete the associated `hdisk` device (`rmdev`).

- ▶ 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, the SAN Volume Controller/Storwize 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.

14.3.2 Adding and removing hosts in SAN Volume Controller/Storwize

When you add and remove host (or hosts) in SAN Volume Controller or Storwize, consider the following guidelines:

- ▶ Before you map new servers to SAN Volume Controller/Storwize, verify that they are all error free. Fix any errors that you find in your server and SAN Volume Controller/Storwize before you proceed. In SAN Volume Controller/Storwize, 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 SAN Volume Controller/Storwize, you should list its WWPN by using the **svcinfo lshbaportcandidate** command. Use the **svcinfo lsfabric** command to certify that it was detected by the SAN Volume Controller/Storwize nodes and ports that you expected. When you create the host definition in SAN Volume Controller/Storwize (**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.

14.4 SAN Volume Controller/Storwize code upgrade

Because the SAN Volume Controller is and Storwize might be in the core of your disk and SAN storage environment, its upgrade requires planning, preparation, and verification. However, with the appropriate precautions, an upgrade can be conducted easily and transparently to your servers and applications.

At the time of this writing, SAN Volume Controller V5.1 is approaching its End-of-Support date. Therefore, your SAN Volume Controller must already be at least at V6.1. This section highlights applicable guidelines for a SAN Volume Controller and Storwize upgrade. For more information how to upgrade from version 5.x to version 6.x, see the 14.4.2, "SAN Volume Controller upgrade from V5.1 to V6.2" on page 504.

Note: Upgrading from 5.1 to 6.1 does not apply to Storwize family products because Storwize V7000 was first introduced with version 6.1.

14.4.1 Preparing for the upgrade

This section explains how to upgrade your SAN Volume Controller/Storwize code.

Current and target SAN Volume Controller/Storwize code level

First, determine your current and target SAN Volume Controller/Storwize code level. Log in to your SAN Volume Controller/Storwize web-based GUI and see its version in the **Monitoring**  **System** tab. Alternatively, if you are using the CLI, run the `svcinfo lssystem` command.

SAN Volume Controller/Storwize code levels are specified by four digits in the format V.R.M.F, where:

- ▶ V is the major version number
- ▶ RT is the release level
- ▶ M is the modification level
- ▶ F is the fix level

Note: If you are running SAN Volume Controller V5.1 or earlier, check the SAN Volume Controller Console version. The version is displayed in the SAN Volume Controller Console Welcome panel, in the upper-right corner. It is also displayed in the Windows “Control Panel - Add or Remove Software” panel.

Set the SAN Volume Controller Target/Storwize code level to the latest Generally Available (GA) release unless you have a specific reason not to upgrade, such as the following reasons:

- ▶ The specific version of an application or other component of your SAN Storage environment has a known problem.
- ▶ The latest SAN Volume Controller/Storwize 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.

Check the compatibility of your target SAN Volume Controller/Storwize 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 SAN Volume Controller/Storwize target code level is compatible.

For more information, see the following websites:

- ▶ Storwize V3700 Concurrent Compatibility and Code Cross-Reference:
<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004172>
- ▶ Storwize V5000 Concurrent Compatibility and Code Cross-Reference:
<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004336>
- ▶ Storwize V7000 Concurrent Compatibility and Code Cross-Reference:
<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1003705>

- SAN Volume Controller Concurrent Compatibility and Code Cross-Reference:

<http://www.ibm.com/support/docview.wss?rs=591&uid=ssg1S1001707>

SAN Volume Controller/Storwize Upgrade Test Utility

Install and run the latest SAN Volume Controller/Storwize Upgrade Test Utility before you upgrade the SAN Volume Controller/Storwize code. To download the SAN Volume Controller/Storwize Upgrade Test Utility, see this website:

<https://www.ibm.com/support/docview.wss?uid=ssg1S4000585>

Figure 14-4 shows the Storwize V7.1 GUI window that is used to install the test utility. It is uploaded and installed like any other software upgrade. This tool verifies the health of your SAN Volume Controller/Storwize for the upgrade process. It also checks for unfixed errors, degraded MDisk, inactive fabric connections, configuration conflicts, hardware compatibility, Storwize disk drives firmware, and many other issues that might otherwise require cross-checking a series of command outputs.

How this utility works: The SAN Volume Controller/Storwize Upgrade Test Utility does not log in to the storage controllers or SAN switches that it uses to check for errors. Instead, it reports the status of its connections to these devices as it detects them.

Also, check these components for errors. Before you run the upgrade procedure, read the SAN Volume Controller code version Release Notes.

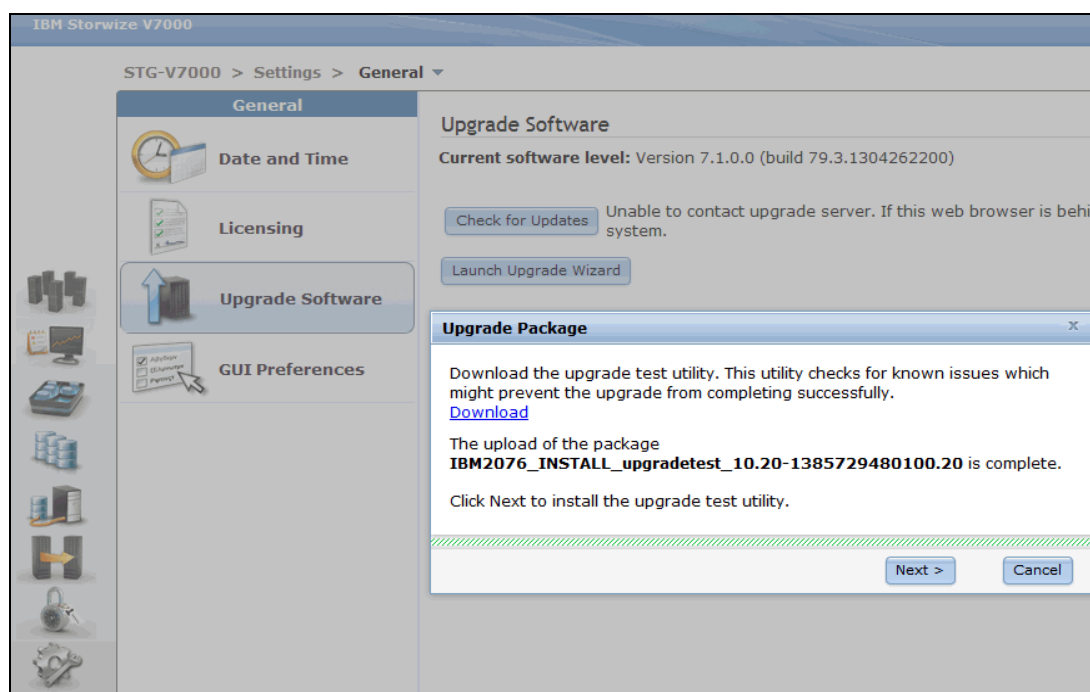


Figure 14-4 SAN Volume Controller/Storwize Upgrade Test Utility installation by using the GUI

You can use the GUI or the CLI to upload and install the SAN Volume Controller/Storwize Upgrade Test Utility. After installation, you can run Upgrade Test Utility from CLI or GUI. In Example 14-1 on page 500, we show the output of running Upgrade Test Utility in CLI.

Example 14-1 Upgrade test by using the CLI

```
IBM_Storwize:superuser>svcupgradetest -v 7.1.0.6 -d
svcupgradetest version 10.20
```

Please wait, the test may take several minutes to complete.

***** Warning found *****

This tool has found the internal disks of this system are not running the recommended firmware versions. Details follow:

Model	Latest FW	Current FW	Drive Info
ST91000640SS	BD2F	BD2E	Drive in slot 1 in enclosure 1 Drive in slot 2 in enclosure 1 Drive in slot 9 in enclosure 1 Drive in slot 8 in enclosure 1 Drive in slot 10 in enclosure 1 Drive in slot 11 in enclosure 1 Drive in slot 5 in enclosure 1 Drive in slot 7 in enclosure 1 Drive in slot 6 in enclosure 1 Drive in slot 4 in enclosure 1 Drive in slot 12 in enclosure 1 Drive in slot 3 in enclosure 1 Drive in slot 13 in enclosure 1 Drive in slot 14 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.

Results of running svcupgradetest:
=====

The tool has found warnings.
For each warning above, follow the instructions given.

The tool has found 0 errors and 1 warnings

The SAN Volume Controller/Storwize Upgrade Test Utility found some drives with an older firmware version and indicates recommended actions. The drive upgrade process is covered in the following sections.

SAN Volume Controller/Storwize hardware considerations

Before you start the upgrade process, always check whether your SAN Volume Controller/Storwize hardware and target code level are compatible.

Figure 14-5 shows the compatibility matrix between the latest SAN Volume Controller hardware node models and code versions. If your SAN Volume Controller cluster has nodes model 4F2, replace them with newer models before you upgrade their code. Conversely, if you plan to add or replace nodes with new models CF8 or CG8 to an existing cluster, upgrade your SAN Volume Controller code first.

SVC node model SVC code version	SVC node model	4F2	8F2	8F4	8G4	8A4	CF8	CG8
7.2		✗	✗	✗	✓	✓	✓	✓
7.1		✗	✓	✓	✓	✓	✓	✓
6.4		✗	✓	✓	✓	✓	✓	✓
6.3		✗	✓	✓	✓	✓	✓	✓
6.2		✗	✓	✓	✓	✓	✓	✓
6.1		✗	✓	✓	✓	✓	✓	✗
5.1		✗	✓	✓	✓	✓	✓	✗

Figure 14-5 SAN Volume Controller node models and code versions relationship

Figure 14-6 on page 502 shows the compatibility matrix between the Storwize family systems and code versions.

Storwize system Storwize code version	V3700	V5000	V7000
7.2	✓	✓	✓
7.1	✓	✓	✓
6.4	✓	✗	✓
6.3	✗	✗	✓
6.2	✗	✗	✓
6.1	✗	✗	✓
5.1	✗	✗	✗

Figure 14-6 Storwize family systems and code versions relationship

Attached hosts preparation

If the appropriate precautions are taken, the SAN Volume Controller/Storwize upgrade is not apparent to the attached servers and their applications. The automated upgrade procedure updates one SAN Volume Controller node at a time (or one Storwize node canister), 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.

Before you start SAN Volume Controller/Storwize upgrade preparation, check the following items for every server that is attached to the SAN Volume Controller/Storwize you upgrade:

- ▶ 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):
[http://www.ibm.com/support/entry/portal/Troubleshooting/Hardware/System_Storage/Storage_software/Storage_virtualization/SAN_Volume_Controller_\(2145\)](http://www.ibm.com/support/entry/portal/Troubleshooting/Hardware/System_Storage/Storage_software/Storage_virtualization/SAN_Volume_Controller_(2145))
- ▶ The IBM Support page on Storwize V7000 Flashes and Alerts (Troubleshooting):
https://www-947.ibm.com/support/entry/myportal/all_troubleshooting_links/system_storage/disk_systems/mid-range_disk_systems/ibm_storwize_v7000_%282076%29?productContext=-1546771614
- ▶ The IBM Support page on Storwize V5000 Flashes and Alerts (Troubleshooting):
https://www-947.ibm.com/support/entry/myportal/all_troubleshooting_links/system_storage/disk_systems/mid-range_disk_systems/ibm_storwize_v5000?productContext=-2033461677

- The IBM Support page on Storwize V3700 Flashes and Alerts (Troubleshooting):

https://www-947.ibm.com/support/entry/myportal/all_troubleshooting_links/system_storage/disk_systems/entry-level_disk_systems/ibm_storwize_v3700?productContext=-124971743

Fix every problem or “suspect” that you find with the disk path failover capability. Because a typical SAN Volume Controller/Storwize 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 and VIOS, or Solaris containers in your environment, verify the redundancy and failover capability in these virtualization layers.

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 the SAN Volume Controller/Storwize must be free of errors.

SAN fabrics preparation

If you are using symmetrical, redundant, independent SAN fabrics, preparing these fabrics for a SAN Volume Controller/Storwize upgrade 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 SAN Volume Controller/Storwize environment must continue working through the other fabric and your applications must remain unaffected.

Because you are upgrading your SAN Volume Controller/Storwize, also upgrade your SAN switches code to the latest supported level. Start with your principal core switch or director, continue by upgrading the other core switches, and upgrade the edge switches last. Upgrade 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 upgrade only after you verify that the first fabric upgrade has no problems.

If you are still not running symmetrical, redundant independent SAN fabrics, fix this problem as a high priority because it represents a single point of failure (SPOF).

Upgrade sequence

The SAN Volume Controller/Storwize Supported Hardware List provides the correct sequence for upgrading your SAN Volume Controller/Storwize SAN storage environment components. For V7.2 of this list, see the following resources:

- V7.2.x Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller:

http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004453#_Prev

- V7.2.x Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for IBM Storwize V7000:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004450>

- V7.2.x Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for IBM Storwize V5000:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004515>

- V7.2.x Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for IBM Storwize V3700:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004515>

By cross-checking the version of SAN Volume Controller/Storwize that are compatible with the versions of your SAN directors, you can determine which one to upgrade first. By checking a component's upgrade path, you can determine whether that component requires a multistep upgrade.

If you are not making major version or multistep upgrades in any components, the following upgrade order is less prone to eventual problems:

1. SAN switches or directors
2. Storage controllers
3. Servers HBAs microcodes and multipath software
4. SAN Volume Controller cluster/Storwize system
5. SAN Volume Controller/Storwize internal disk drives

Attention: Do *not* upgrade two components of your SAN Volume Controller SAN storage environment simultaneously, such as the SAN Volume Controller and one storage controller, even if you intend to do it with your system offline. An upgrade of this type can lead to unpredictable results, and an unexpected problem is much more difficult to debug.

14.4.2 SAN Volume Controller upgrade from V5.1 to V6.2

SAN Volume Controller incorporated several new features in V6 compared to the previous version. The most significant differences regarding the upgrade process concern the SAN Volume Controller Console, the new configuration, and the use of internal SSD disks with Easy Tier.

For a practical example of this upgrade, see Chapter 16, “SAN Volume Controller scenarios” on page 555.

SAN Volume Controller Console

With SAN Volume Controller V6.1, separate hardware with the specific function of the SAN Volume Controller Console is no longer required. The SAN Volume Controller Console software is incorporated in the nodes. To access the SAN Volume Controller Management GUI, use the cluster IP address.

If you purchased your SAN Volume Controller with a console or SSPC server, and you no longer have any SAN Volume Controller clusters that are running SAN Volume Controller V5.1 or earlier, you can remove the SAN Volume Controller Console software from this server. In fact, SAN Volume Controller Console V6.1 and V6.2 utilities remove the previous SAN Volume Controller Console GUI software and create desktop shortcuts to the new console GUI. For more information and to download the GUI, see *V6.x IBM System Storage SAN Volume Controller Console (SVCC) GUI*, S4000918, at this website:

<https://www.ibm.com/support/docview.wss?uid=ssg1S4000918>

Easy Tier with SAN Volume Controller internal SSDs

SAN Volume Controller V6.2 included support for Easy Tier by using SAN Volume Controller internal SSDs with node models CF8 and CG8. If you are using internal SSDs with a SAN Volume Controller release before V6.1, remove these SSDs from the managed disk group that they belong to and put it into the unmanaged state before you upgrade to release 6.2.

Example 14-2 The svcupgradetest command with SSDs in managed state

Example 14-3 Upgrade effect on SSDs

Previous configuration in SVC version 5.1:

```
IBM_2145:svccf8:admin>svcinfolscntroller
id      controller_name  ctrl_s/n      vendor_id      product_id_low  product_id_high
0       controller0          75L3001FFFF   IBM            2145            Internal
1       controller1          75L3001FFFF   IBM            2107900         Internal
2       controller2          75L3331FFFF   IBM            2107900         Internal
3       controller3          75L3331FFFF   IBM            2145            Internal
IBM_2145:svccf8:admin>
```

After upgrade SVC to version 6.2:

```
IBM_2145:svccf8:admin>lscntroller
id controller_name ctrl_s/n      vendor_id      product_id_low  product_id_high
1  DS8K75L300I    75L3001FFFF   IBM            2107900
2  DS8K75L333I    75L3331FFFF   IBM            2107900
IBM_2145:svccf8:admin>
IBM_2145:svccf8:admin>lssdrive
id status error_sequence_number use tech_type capacity mdisk_id mdisk_name member_id enclosure_id slot_id node_id node_name
0  online                unused sas_ssd  136.2GB                0      2      node2
1  online                unused sas_ssd  136.2GB                0      1      node1
IBM_2145:svccf8:admin>
```

You must decide which RAID level you configure in the new arrays with SSDs, depending on the purpose that you give them and the level of redundancy that is needed to protect your data if a hardware failure occurs. Table 14-2 lists the factors to consider in each case. By using your internal SSDs for Easy Tier, you can achieve a gain in overall performance in most cases.

Table 14-2 RAID levels for internal SSDs

RAID level (GUI Preset)	What you need	When to use it	For best performance
RAID 0 (Striped)	1 - 4 drives, all in a single node.	When VDisk Mirror is on external MDisks.	A pool should contain only arrays from a single I/O group.
RAID 1 (Easy Tier)	2 drives, one in each node of the I/O group.	When Easy Tier is used or both mirrors on SSDs	An Easy Tier pool should contain only arrays from a single I/O group. The external MDisks in this pool should be used only by the same I/O group.
RAID 10 (Mirrored)	4 - 8 drives, equally distributed among each node of the I/O group	When multiple drives for a VDisk are used	A pool should contain only arrays from a single I/O group. Preferred over VDisk Mirroring.

14.4.3 Upgrading SAN Volume Controller clusters/Storwize systems that are participating in Metro Mirror or Global Mirror

When you upgrade an SAN Volume Controller cluster that participates in an intercluster Copy Services relationship, do *not* upgrade both clusters in the relationship simultaneously. This situation is not verified or monitored by the Automatic Upgrade process and might lead to a loss of synchronization and unavailability. You must successfully finish the upgrade in one cluster before you start the next one. Try to upgrade 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.

If possible, stop all intercluster relationships during the upgrade, and then start them again after the upgrade is completed.

Note: Although stopping remote copy services is not necessary, it is the preferred practice to do so. One exception is when you are upgrading to version 7.2. You *must* stop all Global Mirror (GM) relationships before starting the upgrade process because of performance improvements in GM code in SAN Volume Controller/Storwize software version 7.2. Other remote copy relationships, such as Metro Mirror (MM) or Global Mirror with Change Volumes (GMCV), do not have to be stopped.

14.4.4 SAN Volume Controller/Storwize upgrade

Adhere to the following version-independent guidelines for your SAN Volume Controller code upgrade:

- ▶ Schedule the SAN Volume Controller/Storwize code upgrade for a low I/O activity time. The upgrade 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 upgraded. Therefore, with lower I/O, you are less likely to notice performance degradation during the upgrade.
- ▶ Never power off an SAN Volume Controller node during code upgrade unless you are instructed to do so by IBM Support. Typically, if the upgrade 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 SAN Volume Controller/Storwize target code level on every computer that you intend to use to manage your SAN Volume Controller/Storwize.
- ▶ If you are planning for a major SAN Volume Controller/Storwize version upgrade (such as V6 to V7), update your current version to its latest fix level before you run the major upgrade.

Note: If You are running code version older than 6.4, you must upgrade SVC/Storwize to 6.4 first before upgrading to 7.2. Direct upgrade to v7.2 from versions older than 6.4 is not supported.

14.4.5 Storwize family systems disk drive upgrade

With the introduction of Storwize V7000, a disk drive was added to the code upgrade list. Since the initial release of Storwize V7000, there always was a possibility to download a Drive Microcode Package from the Storwize support website and install it by using the CLI. Although it was possible to upgrade one disk at a time, the multiple drive upgrade was possible only with a special tool called `utilitydriveupgrade.pl`, which also is downloadable at no charge from the Storwize support website.

With the introduction of the 7.2 code version, there is no need to use this tool because multiple drive upgrade functionality is embedded in SAN Volume Controller/Storwize software. Additionally, 7.2 added the following enhancements:

- ▶ Command **`applydrivesoftware`** added to the CLI
- ▶ Checks for existing drives' firmware levels against package content and upgrades only if required
- ▶ Firmware upgrade progress monitoring possible by using the CLI
- ▶ Automatically stops all pending downloads for any error conditions or changes of configuration

Example 14-4 shows **applydrivesoftware** command syntax

Example 14-4 applydrivesoftware syntax

```
IBM_2145:superuser>applydrivesoftware -?
```

```
applydrivesoftware
```

```
Syntax
```

```
>>- applydrivesoftware -- -file --name----->

>--+-----+--+ -drive --drive_id--+----->
|           .-firmware-. | '- -all -----'
| '- -type --+fpga-----+-'

>--+-----+--+-----+--+-----+--><
| '- -force -' '- -allowreinstall -' '- -allowdowngrade -'

>>- applydrivesoftware -- -cancel -----><
```

For more details type 'help applydrivesoftware'.

Preferred practice: If you are running v7.2 of SVC/Storwize code, use the **applydrivefirmware** command instead of the `utilitydriveupgrade.pl` tool. This utility is not supported or tested beyond version 7.1 of SAN Volume Controller/Storwize code.

Upgrade of disk drive firmware is concurrent whether it is HDD or SSD. However, with SSD, the firmware level and FPGA level can be upgraded. Upgrade of FPGA is *not* concurrent, so all IOs to the SSDs must be stopped before the upgrade. 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 upgrade.

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 upgrading the FPGA on SSD in the 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.

14.5 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:

- ▶ Removing the mapping of a LUN (volume, or VDisk) that is still in use by a server.
- ▶ 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.
- ▶ Avoid unintended disruption of servers and applications.
- ▶ Dramatically increase the overall availability of your IT infrastructure.

14.5.1 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 SAN Volume Controller's 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 SAN Volume Controller/Storwize 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 14-5.

Example 14-5 Output of the datapath query WWPNS command

```
[root@nybixtdb02]> pcmpath query wwpn
Adapter Name PortWWN
fscsi0       10000000C925F5B0
fscsi1       10000000C9266FD1
```

If you are using server virtualization, verify the WWPNS in the server that is attached to the SAN, such as AIX VIO or VMware ESX.

2. Cross-reference with the output of the SAN Volume Controller/Storwize **lshost** **<hostname>** command, as shown in Example 14-6.

Example 14-6 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 14-7. (In Brocade, switches use **nodefind <WWPN>**.)

Example 14-7 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.

14.5.2 Cross-referencing LUN IDs

Always cross-reference the SAN Volume Controller/Storwize vdisk_UID with the server LUN ID before you perform any modifications that involve SAN Volume Controller/Storwize volumes. Example 14-8 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 14-8 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 LUN ID information in every change ticket that performs SAN storage allocation or reclaim.

Preferred practice: Because a host can have many volumes with the same `scsi_id`, always cross-reference SAN Volume Controller/Storwize volume UID with host volume UID and record the `scsi_id` of that volume. With AIX, the hdisk UID is presented as SERIAL in the SDDPCM and this number corresponds with Volume UID in the SAN Volume Controller/Storwize GUI (or Vdisk_UID in the CLI). This configuration is important in the various data migration scenarios.

14.5.3 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 replace it “in hot” (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 and by using the multipath software, identify the failed HBA and record its WWPN (for more information, see 14.5.1, “Cross-referencing HBA WWPNs” on page 509). 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 WWPN. If you have this type of label, record the WWPN 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 replace the HBA in hot. 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 WWPN logged in to the SAN switch port.

Otherwise, fix this issue before you continue to the next step.

Cross-check the WWPN 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 WWPN for the new one in every alias and zone to which it belongs. 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 this WWPN, and zones must reference this alias.

If you are using SAN port zoning (though you should not be) and you did not move the new HBA FC cable to another SAN switch port, you do not need to reconfigure zoning.

6. Verify that the new HBA's WWPN appears in the SAN Volume Controller/Storwize by using the **lsfcportcandidate** command.

If the WWPN of the new HBA does not appear, troubleshoot your SAN connections and zoning if you did not do so.

7. Add the WWPN of this new HBA in the SAN Volume Controller/Storwize 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 new HBAs 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 SAN Volume Controller/Storwize 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 WWPN from the host definition by using the **rmhostport** command.

Troubleshoot your SAN connections and zoning if you did not do so. 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.

14.6 Hardware upgrades for SAN Volume Controller

The SAN Volume Controller's scalability features allow significant flexibility in its configuration. As a consequence, several scenarios are possible for its growth. The following sections describe adding SAN Volume Controller nodes to an existing cluster, upgrading SAN Volume Controller nodes in an existing cluster, and moving to a new SAN Volume Controller cluster. It also includes suggested ways to deal with each scenario.

14.6.1 Adding SAN Volume Controller nodes to an existing cluster

If your existing SAN Volume Controller cluster is below four I/O groups and you intend to upgrade it, you might find yourself installing newer nodes that are more powerful than your existing ones. Therefore, your cluster has different node models in different I/O groups.

To install these newer nodes, determine whether you need to upgrade your SAN Volume Controller code level first. For more information, see “SAN Volume Controller/Storwize hardware considerations” on page 501.

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 SAN Volume Controller/Storwize firmware. Although this can be done without stopping the host, careful planning and preparation are advised. For more information about NDVM, see the IBM SAN Volume Controller Information Center at this website:

http://pic.dhe.ibm.com/infocenter/svc/ic/topic/com.ibm.storage.svc.console.doc/svc_migratevdiskiogrpcli_42lens.html

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 in the SAN Volume Controller which servers use which I/O groups. Otherwise, the SAN Volume Controller by default maps the host to all I/O groups, even if they do not exist and regardless of your zoning configuration. Example 14-9 shows this scenario and how to resolve it.

Example 14-9 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>

```

-
- If possible, avoid setting a server to use volumes from I/O groups by using different node types (as a permanent situation, in any case). Otherwise, as this server's storage capacity grows, you might experience a performance difference between volumes from different I/O groups, which makes it difficult to identify and resolve eventual performance problems.

14.6.2 Upgrading SAN Volume Controller 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 IBM SAN Volume Controller Information Center at this website:

http://pic.dhe.ibm.com/infocenter/svc/ic/index.jsp?topic=%2Fcom.ibm.storage.svc.console.doc%2Fsvc_replacingnodemodelsnondisrupttask_2r1r4z.html

Nondisruptive node replacement uses failover capabilities to replace one node in an I/O group at a time. If a new node has a newer version of SAN Volume Controller code, it upgrades an older node automatically during the node replacement procedure.

14.6.3 Moving to a new SAN Volume Controller cluster

You might have a highly populated, intensively used SAN Volume Controller cluster that you want to upgrade, and you also want to use the opportunity to overhaul your SAN Volume Controller and SAN storage environment.

The following scenario might make it easier is to replace your cluster entirely with a newer, bigger, and more powerful one:

1. Install your new SAN Volume Controller cluster.
2. Create a replica of your data in your new cluster.
3. Migrate your servers to the new SAN Volume Controller Cluster when convenient.

If your servers can tolerate a brief, scheduled outage to switch from one SAN Volume Controller to another, you can use SAN Volume Controller's remote copy services (Metro Mirror or Global Mirror) to create your data replicas. After all data is replicated to the new SAN Volume Controller, you must complete the following steps:

1. Stop all IO from the host you want to move to the new SAN Volume Controller.
2. From this host, disconnect and remove all volumes that are mapped from the old SAN Volume Controller.

3. Stop and remove remote copy relations with those volumes so that the target volumes on the new SAN Volume Controller receive read/write access.
4. Unmap the volumes from the old cluster.
5. Zone your host to the new SAN Volume Controller cluster.
6. Map the volumes from the new cluster to the host.
7. Discover new volumes on the host.
8. Import data from those volumes and start your applications.

If you must migrate a server online, modify its zoning so that it uses volumes from both SAN Volume Controller clusters. Also, use host-based mirroring (such as AIX `mirrorvg`) to move your data from the old SAN Volume Controller to the new one. 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 this approach is that it 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.

Chapter 16, "SAN Volume Controller scenarios" on page 555, describes a possible approach for this scenario that replaces the SAN Volume Controller, the switches, and the back-end storage.

14.7 Adding expansion enclosures to Storwize family systems

Usually, if planned well, when you buy one of Storwize family storage systems 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 Storwize system you have, you can add four (V3700), six (V5000), or nine (v7000) expansion enclosures to the Storwize control enclosure. Because all Storwize storage 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.

When you are adding an expansion enclosure, complete the following steps:

1. Place the expansion enclosure in the rack cabinet.
2. Connect power cables and power on the enclosure. Wait a few minutes until the expansion enclosure boots up and is in stable state.
3. Connect SAS cables to the control enclosure or other expansion enclosure.

In Storwize storage systems, all added enclosures form two chains. The first chain is formed by all enclosures that are connected to the SAS 1 ports in both control node canisters. The second chain is formed by enclosures that are connected to the SAS 2 ports in both control nodes canisters.

The control enclosure is the first enclosure in the second chain and because of that, you can add five enclosures to the first chain and four enclosures to the second chain in Storwize V7000. For Storwize V3700, you can add two enclosures to every chain; for Storwize V5000, three enclosures to every chain can be added. 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: Storwize can detect incorrect cabling and creates warning messages in the event log.

Correct cabling in a fully populated Storwize V7000 is shown in Figure 14-7.

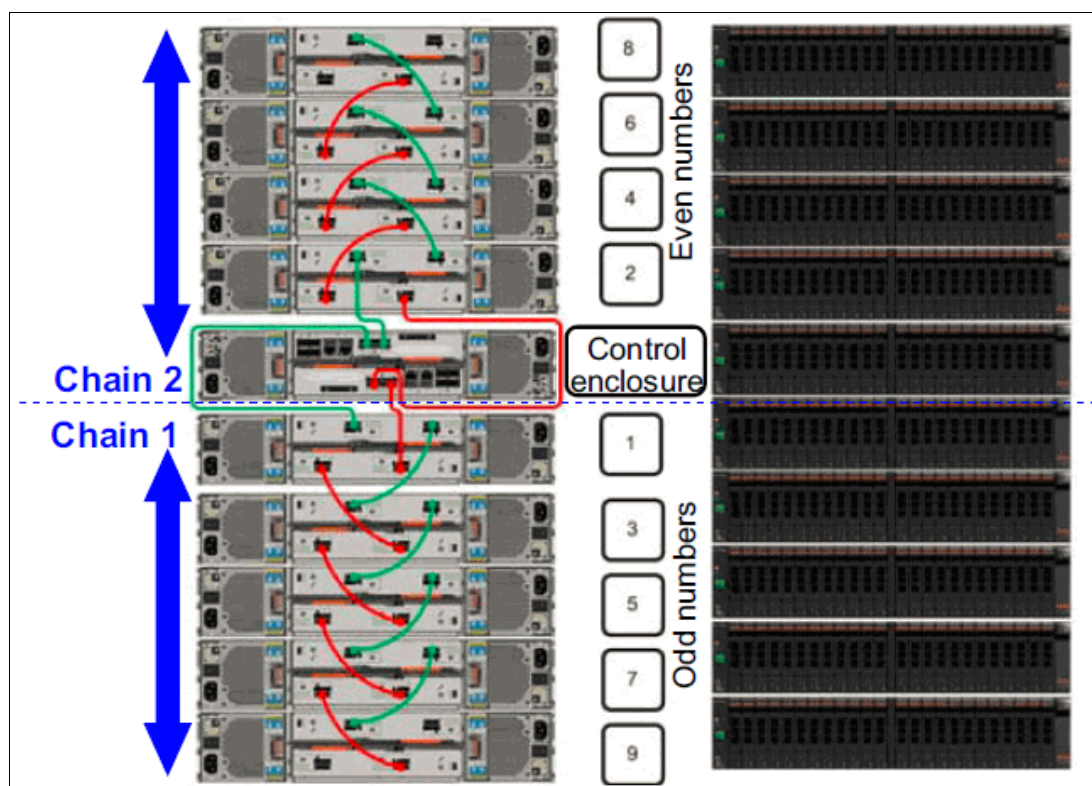


Figure 14-7 Storwize V7000 Control enclosure with maximum number of expansion enclosures

Preferred practice: If you plan for future growth, always leave space in the rack cabinet; 10 U under and 8 U over Storwize control enclosure for more expansion enclosures.

Adding expansion enclosures is simplified because Storwize can automatically discover new expansion enclosures after the SAS cables are connected and prompts for configuration of new disk drives. Expansion enclosures that are left in this state work properly; however, they are not monitored properly and because of that, there is a lack of expansion enclosure information in logs if there are any problems. This issue can lead to more difficult troubleshooting and can extend time of problem resolution. To avoid this situation, always run the Add Expansion Enclosure procedure in the GUI, as shown in Figure 14-8 on page 517.

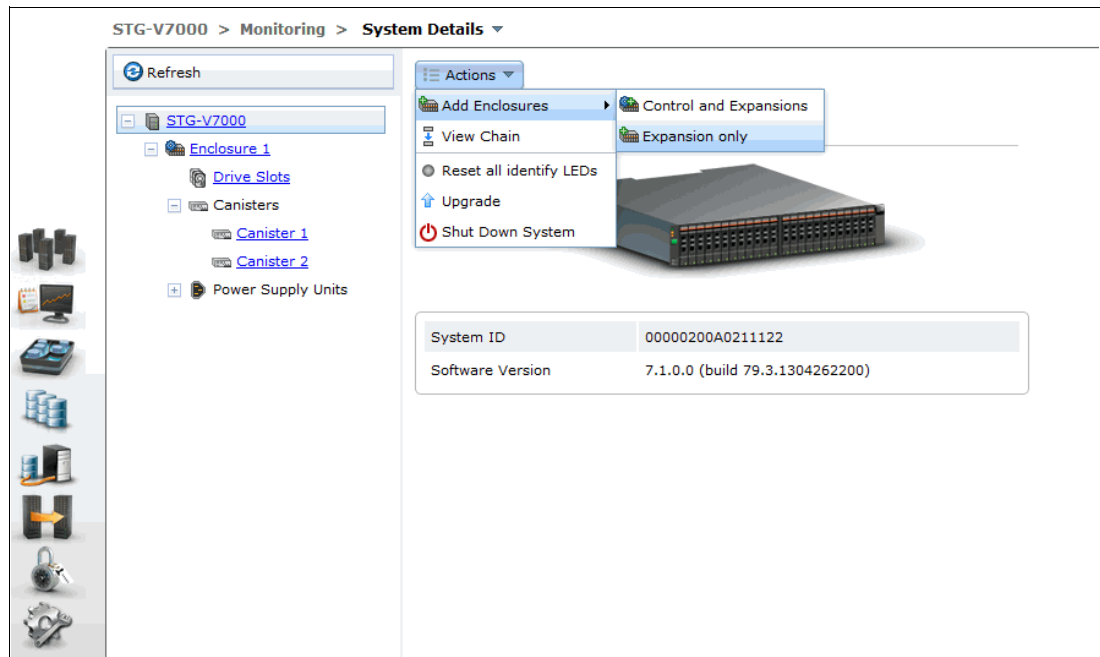


Figure 14-8 Adding expansion enclosures

Preferred practice: Because of Storwize 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 put 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.

For more information about adding control enclosures, see Chapter 3, “SAN Volume Controller and Storwize V7000 Cluster” on page 59.

14.8 More information

More practices can be applied to SAN storage environment management that can benefit its administrators and users. For more information about the practices that are covered here and others that you can use, see Chapter 16, “SAN Volume Controller scenarios” on page 555.



Troubleshooting and diagnostics

The SAN Volume Controller 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 complicated and problems can occur.

This chapter provides an overview of common problems that can occur in your environment. It describes problems that are related to the SAN Volume Controller, the SAN environment, storage subsystems, hosts, and multipathing drivers. It also describes how to collect the necessary problem determination data and how to overcome such issues.

This chapter includes the following sections:

- ▶ Common problems
- ▶ Collecting data and isolating the problem
- ▶ Recovering from problems
- ▶ Mapping physical LBAs to volume extents
- ▶ Medium error logging
- ▶ Replacing a bad disk
- ▶ Health status during upgrade

15.1 Common problems

SANs, storage subsystems, and host systems are 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 errors occur, administrators must know what to look for and where to look.

The SAN Volume Controller is a useful tool for isolating problems in the storage infrastructure. With functions that are found in the SAN Volume Controller, administrators can more easily locate any problem areas and take the necessary steps to fix the problems. In many cases, the SAN Volume Controller and its service and maintenance features guide administrators directly, provide help, and suggest remedial action. Furthermore, the SAN Volume Controller probes whether the problem still persists.

When you experience problems with the SAN Volume Controller environment, ensure that all components that comprise the storage infrastructure are interoperable. In a SAN Volume Controller environment, the SAN Volume Controller support matrix is the main source for this information. For the latest SAN Volume Controller V7.2 support matrix, see *V7.2 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller*, S1004453, which is available at this website:

<https://www-01.ibm.com/support/docview.wss?uid=ssg1S1004453>

Although the latest SAN Volume Controller code level is supported to run on older host bus adapters (HBAs), storage subsystem drivers, and code levels, use the latest tested levels.

15.1.1 Host problems

From the host perspective, you can experience various problems that range from performance degradation to inaccessible disks. To diagnose these issues, you can check a few items from the host before you drill down to the SAN, SAN Volume Controller, and storage subsystems.

Check the following areas on the host:

- ▶ Any special software that you are using
- ▶ Operating system version and maintenance or service pack level
- ▶ Multipathing type and driver level
- ▶ Host bus adapter model, firmware, and driver level
- ▶ Fibre Channel SAN connectivity

Based on this list, the host administrator must check and correct any problems.

For more information about managing hosts on the SAN Volume Controller, see Chapter 8, “Hosts” on page 225.

15.1.2 SAN Volume Controller problems

The SAN Volume Controller has useful error logging mechanisms. It keeps track of its internal problems and informs the user about problems in the SAN or storage subsystem. It also helps to isolate problems with the attached host systems. Every SAN Volume Controller 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 SAN Volume Controller 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 SAN Volume Controller nodes. The fast node reset function means that SAN Volume Controller software problems 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 SAN Volume Controller node. This node informs the other members of the cluster that it is resetting.

Other than SAN Volume Controller node hardware and software problems, failures in the SAN zoning configuration are a problem. A misconfiguration in the SAN zoning configuration might lead to the SAN Volume Controller cluster not working because the SAN Volume Controller cluster nodes communicate with each other by using the Fibre Channel SAN fabrics.

You must check the following areas from the SAN Volume Controller perspective:

- ▶ The attached hosts. For more information, see 15.1.1, “Host problems” on page 520.
- ▶ The SAN. For more information, see 15.1.3, “SAN problems” on page 522.
- ▶ The attached storage subsystem. For more information, see 15.1.4, “Storage subsystem problems” on page 522.

The SAN Volume Controller has several command-line interface (CLI) commands that you can use to check the status of the SAN Volume Controller 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 SAN Volume Controller 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 15.3.4, “Solving back-end storage problems” on page 545. The total of all *path_counts* must add up to the number of managed disks (MDisks) multiplied by the number of SAN Volume Controller 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 nodes?

- ▶ **svcinfo lsvdisk**

Check that all virtual disks (volumes) are online (not degraded or offline). If the volumes are degraded, are there stopped FlashCopy jobs? Restart these stopped FlashCopy jobs or delete the mappings.

- ▶ **svcinfo lshostvdiskmap**

Check that all volumes are mapped to the correct hosts. If a volume is not mapped correctly, create the necessary host mapping.

► **svcinfo lsfabric**

Use this command with the various options, such as **-controller**. Also, you can check different parts of the SAN Volume Controller configuration to ensure that multiple paths are available from each SAN Volume Controller node port to an attached host or controller. Confirm that all SAN Volume Controller node port WWPNs are connected to the back-end storage consistently.

15.1.3 SAN problems

Introducing the SAN Volume Controller into your SAN environment and the use of its virtualization functions are not difficult tasks. Before you can use the SAN Volume Controller in your environment, you must follow the basic rules. These rules are not complicated. However, you can make mistakes that lead to accessibility problems or a reduction in the performance experienced.

Two types of SAN zones are needed to run the SAN Volume Controller in your environment: a *host zone* and a *storage zone*. In addition, you must have a SAN Volume Controller zone that contains all of the SAN Volume Controller node ports of the SAN Volume Controller cluster. This SAN Volume Controller zone enables intracluster communication. For more information and important points about setting up the SAN Volume Controller in a SAN fabric environment, see Chapter 2, “SAN topology” on page 17.

Because the SAN Volume Controller is in the middle of the SAN and connects the host to the storage subsystem, check and monitor the SAN fabrics.

15.1.4 Storage subsystem problems

Today, various heterogeneous storage subsystems are available. All of these subsystems have different management tools, different setup strategies, and possible problem areas. To support a stable environment, all subsystems must be correctly configured and in good working order, without open problems.

Check the following areas if you experience a problem:

- Storage subsystem configuration. Ensure that a valid configuration is 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 problems.
- 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 important factors.

For more information about managing subsystems, see Chapter 4, “Back-end storage” on page 71.

Determining the correct number of paths to a storage subsystem

By using SAN Volume Controller 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 15-1 shows how to obtain this information by using the **svcinfo lscontroller controllerid** and **svcinfo lsnode** commands.

Example 15-1 The svcinfo lscontroller 0 command

```
IBM_2145:itsosvcc11:admin>svcinfo lscontroller 0
id 0
controller_name controller0
WWNN 200400A0B8174431
mdisk_link_count 2
max_mdisk_link_count 4
degraded no
vendor_id IBM
product_id_low 1742-900
product_id_high
product_revision 0520
ctrl_s/n
WWPN 200400A0B8174433
path_count 4
max_path_count 12
WWPN 200500A0B8174433
path_count 4
max_path_count 8

IBM_2145:itsosvcc11:admin>svcinfo lsnode
id      name      UPS_serial_number  WWNN      status
IO_group_id  IO_group_name  config_node  UPS_unique_id  hardware
6          Node1          1000739007    50050768010037E5 online
0          io_grp0        no            20400001C3240007 8G4
5          Node2          1000739004    50050768010037DC online
0          io_grp0        yes           20400001C3240004 8G4
4          Node3          100068A006    5005076801001D21 online
1          io_grp1        no            2040000188440006 8F4
8          Node4          100068A008    5005076801021D22 online
1          io_grp1        no            2040000188440008 8F4
```

Example 15-1 shows that two MDisks are present for the storage subsystem controller with ID 0, and four SAN Volume Controller nodes are in the SAN Volume Controller cluster. In this example, the following path_count is used:

$$2 \times 4 = 8$$

If possible, spread the paths across all storage subsystem controller ports, as is the case for Example 15-1 (four for each WWPN).

15.2 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 SAN Volume Controller environment are described.

Today, many approaches are available for monitoring the complete client environment. IBM offers the Tivoli Storage Productivity Center storage management software. Together with problem and performance reporting, Tivoli Storage Productivity Center for Replication offers a powerful alerting mechanism and a powerful Topology Viewer, which enables users to monitor the storage infrastructure. For more information about the Tivoli Storage Productivity Center Topology Viewer, see Chapter 13, “Monitoring” on page 357.

15.2.1 Host data collection

Data collection methods vary by operating system. You can collect the data for various major host operating systems.

First, collect the following information from the host:

- ▶ Operating system: Version and level
- ▶ HBA: Driver and firmware level
- ▶ Multipathing driver level

Then, collect the following operating system-specific information:

- ▶ IBM AIX

Collect the AIX system error log by collecting a **snap -gfiLGc** for each AIX host.

- ▶ For Microsoft Windows or Linux hosts

Use the IBM Dynamic System Analysis (DSA) tool to collect data for the host systems.

For more information about the DSA tool, see the following websites:

- IBM systems management solutions for System x:

<http://www.ibm.com/systems/management/dsa>

- IBM Dynamic System Analysis (DSA):

<http://www.ibm.com/support/entry/portal/docdisplay?brand=5000008&indocid=SERV-D5A>

If your server is based on hardware other than IBM, use the Microsoft problem reporting tool, such as MPSRPT_SETUPPerf.EXE or similar tool, which is available from this website:

<https://www.microsoft.com/en-us/download/default.aspx>

For more information, see this website:

[http://msdn.microsoft.com/en-us/library/bb219076\(v=office.12\).aspx#MicrosoftErrorReporting_MicrosoftErrorReportingInstalledFiles](http://msdn.microsoft.com/en-us/library/bb219076(v=office.12).aspx#MicrosoftErrorReporting_MicrosoftErrorReportingInstalledFiles)

For Linux hosts, another option is to run the **sysreport** tool.

- ▶ VMware ESX Server

Run the **/usr/bin/vm-support** script on the service console. This script collects all relevant ESX Server system and configuration information and ESX Server log files.

In most cases, it is also important to collect the multipathing driver that is used on the host system. Again, based on the host system, the multipathing drivers might be different.

If the driver is an IBM Subsystem Device Driver (SDD), SDDDSM, or SDDPCM host, use **datapath query device** or **pcmpath query device** to check the host multipathing. Ensure that paths go to the preferred and nonpreferred SAN Volume Controller nodes. For more information, see Chapter 8, “Hosts” on page 225.

Check that paths are open for both preferred paths (with select counts in high numbers) and nonpreferred paths (the * or nearly zero select counts). In Example 15-2, path 0 and path 2 are the preferred paths with a high select count. Path 1 and path 3 are the nonpreferred paths, which show an asterisk (*) and 0 select counts.

Example 15-2 Checking paths

```
C:\Program Files\IBM\Subsystem Device Driver>datapath query device -l
```

Total Devices : 1

```
DEV#: 0  DEVICE NAME: Disk1 Part0  TYPE: 2145          POLICY: OPTIMIZED
SERIAL: 60050768018101BF28000000000000037
LUN IDENTIFIER: 60050768018101BF28000000000000037
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	Scsi Port2 Bus0/Disk1 Part0	OPEN	NORMAL	1752399	0
1 *	Scsi Port3 Bus0/Disk1 Part0	OPEN	NORMAL	0	0
2	Scsi Port3 Bus0/Disk1 Part0	OPEN	NORMAL	1752371	0
3 *	Scsi Port2 Bus0/Disk1 Part0	OPEN	NORMAL	0	0

Multipathing driver data (SDD)

SDD was enhanced to collect SDD trace data periodically and to write the trace data to the system’s local hard disk drive. You collect the data by running the **sddgetdata** command. If this command is not found, collect the following four files, where SDD maintains its trace data:

- ▶ sdd.log
- ▶ sdd_bak.log
- ▶ sddsrv.log
- ▶ sddsrv_bak.log

These files can be found in one of the following directories:

- ▶ AIX: /var/adm/ras
- ▶ Hewlett-Packard UNIX: /var/adm
- ▶ Linux: /var/log
- ▶ Solaris: /var/adm
- ▶ Windows 2000 Server and Windows NT Server: \WINNT\system32
- ▶ Windows Server 2003: \Windows\system32

SDDPCM

SDDPCM was enhanced to collect SDDPCM trace data periodically and to write the trace data to the system’s local hard disk drive. SDDPCM maintains the following files for its trace data:

- ▶ pcm.log
- ▶ pcm_bak.log
- ▶ pcmsrv.log
- ▶ pcmsrv_bak.log

Starting with SDDPCM 2.1.0.8, the relevant data for debugging problems is collected by running the **sddpcmgetdata** script, as shown in Example 15-3.

Example 15-3 The sddpcmgetdata script (output shortened for clarity)

```
>sddpcmgetdata
>ls
sddpcmdata_confucius_20080814_012513.tar
```

The **sddpcmgetdata** script collects information that is used for problem determination. Then, it creates a .tar file in the current directory with the current date and time as a part of the file name, as shown in the following example:

```
sddpcmdata_hostname_yyyymmdd_hhmmss.tar
```

When you report an SDDPCM problem, you must run this script and send this .tar file to IBM Support for problem determination.

If the **sddpcmgetdata** command is not found, collect the following files:

- ▶ The pcm.log file
- ▶ The pcm_bak.log file
- ▶ The pcmsrv.log file
- ▶ The pcmsrv_bak.log file
- ▶ The output of the **pcmpath query adapter** command
- ▶ The output of the **pcmpath query device** command

You can find these files in the /var/adm/ras directory.

SDDDSM

SDDDSM also provides the **sddgetdata** script (see Example 15-4) to collect information to use for problem determination. The SDDGETDATA.BAT batch file generates the following information:

- ▶ The sddgetdata_%host%_%date%_%time%.cab file
- ▶ SDD\SDDSRV log files
- ▶ Datapath output
- ▶ Event log files
- ▶ Cluster log files
- ▶ SDD-specific registry entry
- ▶ HBA information

Example 15-4 The sddgetdata script for SDDDSM (output shortened for clarity)

```
C:\Program Files\IBM\SDDDSM>sddgetdata.bat
Collecting SDD trace Data

Collecting datapath command outputs

Collecting SDD and SDDSRV logs

Collecting Most current driver trace

Generating a CAB file for all the Logs

sdddata_DIOEDE_20080814_42211.cab file generated
```

```
C:\Program Files\IBM\SDDDSM>dir
Volume in drive C has no label.
Volume Serial Number is 0445-53F4
```

Directory of C:\Program Files\IBM\SDDDSM

```
06/29/2008  04:22 AM                574,130  sdddata_DIOMEDE_20080814_42211.cab
```

Data collection script for IBM AIX

Example 15-5 shows a script that collects all of the necessary data for an AIX host at one time (operating system and multipathing data). To start the script, complete the following steps:

1. Run `vi /tmp/datacollect.sh`
2. Cut and paste the script into the `/tmp/datacollect.sh` file, and save the file.
3. Run `chmod 755 /tmp/datacollect.sh`
4. Run `/tmp/datacollect.sh`

Example 15-5 Data collection script

```
#!/bin/ksh

export PATH=/bin:/usr/bin:/sbin

echo "y" | snap -r # Clean up old snaps

snap -gGfkLN # Collect new; don't package yet

cd /tmp/ibmsupt/other # Add supporting data
cp /var/adm/ras/sdd* .
cp /var/adm/ras/pcm* .
cp /etc/vpexclude .
datapath query device > sddpath_query_device.out
datapath query essmap > sddpath_query_essmap.out
pcmpath query device > pcmpath_query_device.out
pcmpath query essmap > pcmpath_query_essmap.out
sddgetdata
sddpcmgetdata
snap -c # Package snap and other data

echo "Please rename /tmp/ibmsupt/snap.pax.Z after the"
echo "PMR number and ftp to IBM."

exit 0
```

15.2.2 SAN Volume Controller data collection

Starting with v6.1.0.x, a SAN Volume Controller snap can come from the cluster (collecting information from all online nodes) by running the **svc_snap** command. Alternatively, it can come from a single node snap (in SA mode) by running the **satask snap** command.

You can collect SAN Volume Controller data by using the SAN Volume Controller Console GUI or by using the SAN Volume Controller CLI. You can also generate a SAN Volume Controller livedump.

Data collection for SAN Volume Controller by using the SAN Volume Controller Console GUI

From the support panel that is shown in Figure 15-1, you can download support packages that contain log files and information that can be sent to support personnel to help troubleshoot the system. You can download individual log files or download *statesaves*, which are dumps or livedumps of the system data.



Figure 15-1 Support panel

Complete the following steps to download the support package:

1. Click **Download Support Package**, as shown in Figure 15-2.

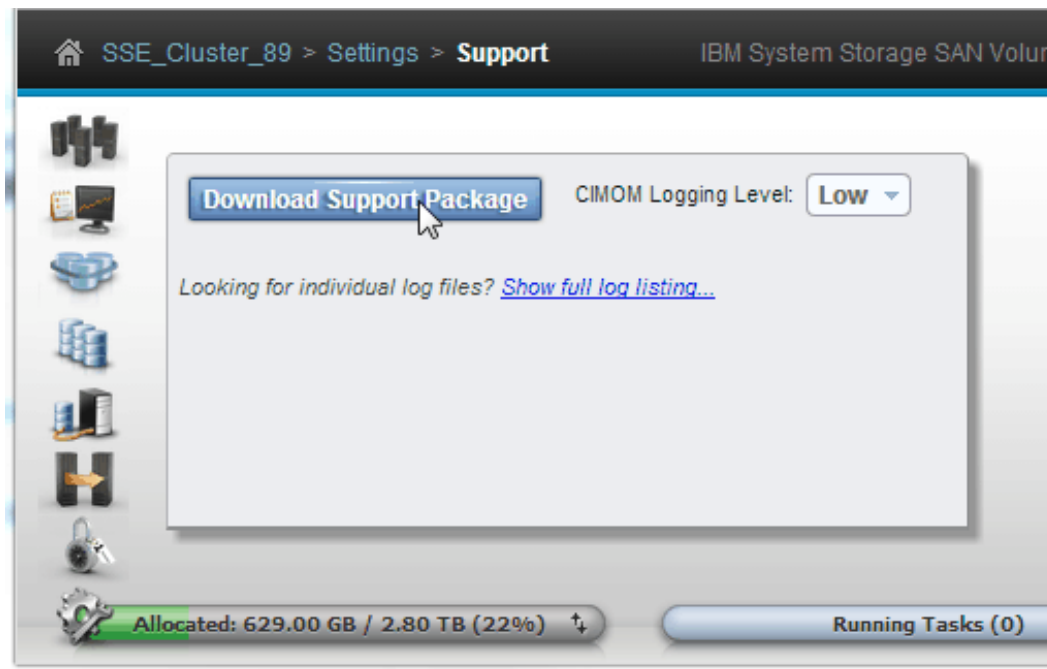


Figure 15-2 Download Support Package

2. In the Download Support Package window that opens (as shown in Figure 15-3), select the log types that you want to download. The following download types are available:
 - Standard logs, which contain the most recent logs that were collected for the system. These logs are the most commonly used by Support to diagnose and solve problems.
 - Standard logs plus one existing statesave, which contain the standard logs for the system and the most recent statesave from any of the nodes in the system. Statesaves are also known as *dumps* or *livedumps*.
 - Standard logs plus most recent statesave from each node, which contains the standard logs for the system and the most recent statesaves from each node in the system.
 - Standard logs plus new statesaves, which generate new statesaves (livedumps) for all nodes in the system, and package them with the most recent logs.

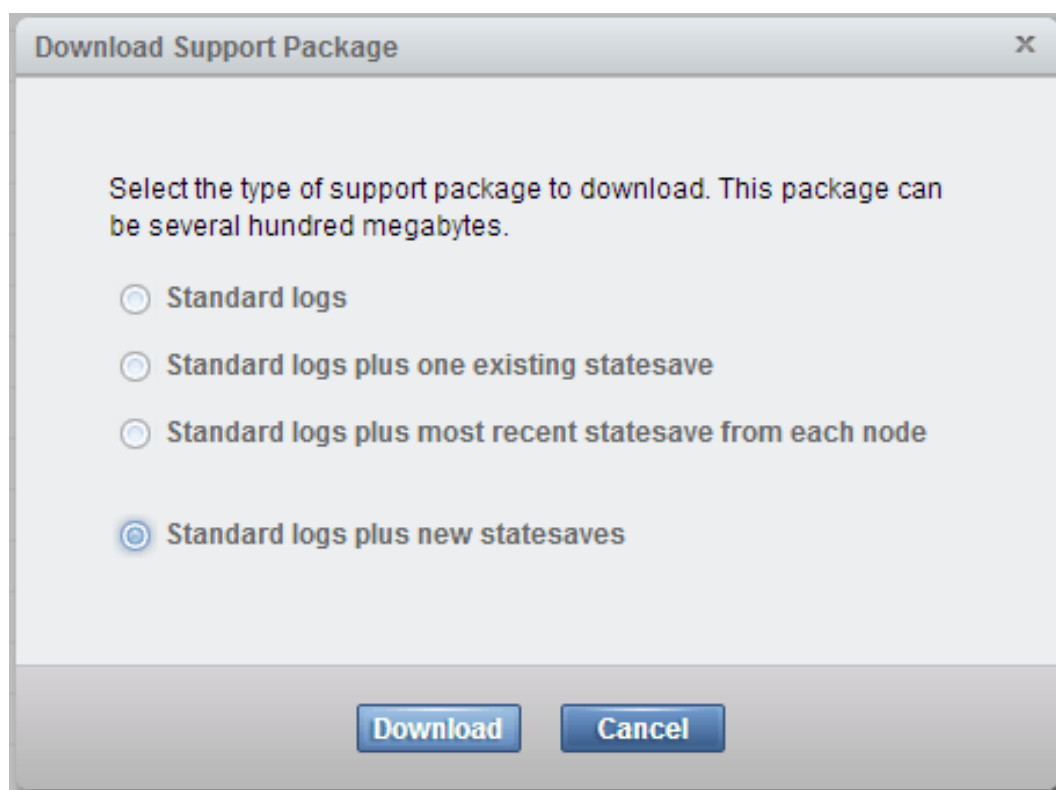


Figure 15-3 Download Support package window

When you are collecting support package for troubleshooting IBM Real-time compression-related issues, remember to note that RACE diagnostics information is only available in a statesave (also called *livedump*). Standard logs do not contain this information. Depending on the problem symptom, download the appropriate support package that contains the RACE diagnostics information. If the system observed a failure, select **Standard logs plus most recent statesave from each node** and for all other diagnostic purposes, select **Standard logs plus new statesaves**.

Then, click **Download**.

Action completion time: Depending on your choice, this action can take several minutes to complete.

3. Select where you want to save these logs, as shown in Figure 15-4. Then, click **OK**.

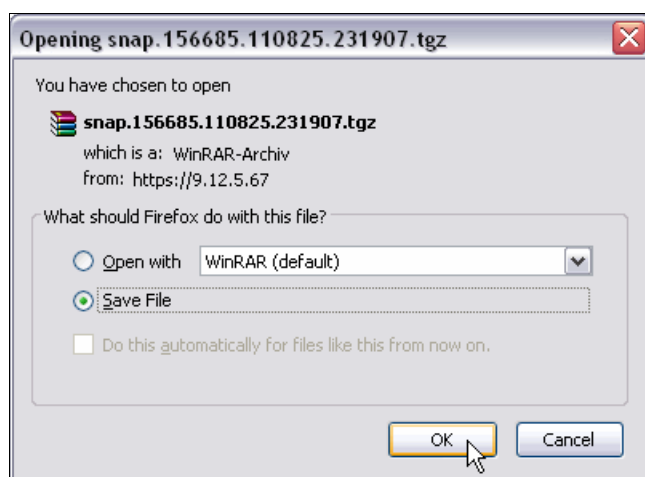


Figure 15-4 Saving the log file on your system

Performance statistics: Any option that is used in the GUI (1 - 4), in addition to using the CLI, collects the performance statistics files from all nodes in the cluster.

Data collection for SAN Volume Controller by using the SAN Volume Controller CLI 4.x or later

Because the config node is always the SAN Volume Controller node with which you communicate, you must copy all the data from the other nodes to the config node. To copy the files, first run the **svcinfolnode** command to determine the non-config nodes.

Example 15-6 shows the output of this command.

Example 15-6 Determine the non-config nodes (output shortened for clarity)

```
IBM_2145:itsosvcc11:admin>svcinfolnode
```

id	name	WWNN	status	IO_group_id	config_node
1	node1	50050768010037E5	online	0	no
2	node2	50050768010037DC	online	0	yes

The output in Example 15-6 shows that the node with ID 2 is the config node. Therefore, for all nodes *except* the config node, you must run the **svctask cpdumps** command. No feedback is provided for this command. Example 15-7 shows the command for the node with ID 1.

Example 15-7 Copying the dump files from the other nodes

```
IBM_2145:itsosvcc11:admin>svctask cpdumps -prefix /dumps 1
```

To collect all the files (including the config.backup file, trace file, errorlog file, and more), run the **svc_snap dumpall** command. This command collects all of the data, including the dump files. To ensure that a current backup is available on the SAN Volume Controller cluster configuration, run the **svcconfig backup** command before you run the **svc_snap dumpall** command, as shown in Example 15-8 on page 531).

There are instances in which it is better to use the **svc_snap** command and request the dumps individually. You can perform this task by omitting the **dumpall** parameter, which captures the data collection apart from the dump files.

Attention: Dump files are large. Collect them only if you really need them.

Example 15-8 The svc_snap dumpall command

```
IBM_2145:itsosvcc11:admin>svc_snap dumpall
Collecting system information...
Copying files, please wait...
Copying files, please wait...
Dumping error log...
Waiting for file copying to complete...
Waiting for file copying to complete...
Waiting for file copying to complete...
Waiting for file copying to complete...
Creating snap package...
Snap data collected in /dumps/snap.104603.080815.160321.tgz
```

After the data collection by using the **svc_snap dumpall** command is complete, verify that the new snap file appears in your 2145 dumps directory by using the **svcinfol s2145dumps** command, as shown in Example 15-9.

Example 15-9 The ls2145 dumps command (shortened for clarity)

```
IBM_2145:itsosvcc11:admin>svcinfol s2145dumps
id                2145_filename
0                 dump.104603.080801.161333
1                 svc.config.cron.bak_node2
.
.
23                104603.trc
24                snap.104603.080815.160321.tgz
```

To copy the file from the SAN Volume Controller cluster, use secure copy (SCP). For more information about the PuTTY SCP function, see *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933.

Livedump

SAN Volume Controller livedump is a procedure that IBM Support might ask clients to run for problem investigation. You can generate it for all nodes from the GUI, as described in “Data collection for SAN Volume Controller by using the SAN Volume Controller Console GUI” on page 528. Alternatively, you can trigger it from the CLI; for example, on only one node of the cluster.

Attention: Start the SAN Volume Controller livedump procedure *only* under the direction of IBM Support.

Sometimes, investigations require a livedump from the configuration node in the SAN Volume Controller cluster. A *livedump* is a lightweight dump from a node that can be taken without affecting host I/O. The only effect is a slight reduction in system performance (because of reduced memory that is available for the I/O cache) until the dump is finished.

Complete the following steps to perform a livedump:

1. Prepare the node for taking a livedump by running the following command:
svctask preplivedump <node id/name>

This command reserves the necessary system resources to take a livedump. The operation can take some time because the node might have to flush data from the cache. System performance might be slightly affected after you run this command because part of the memory that is normally available to the cache is not available while the node is prepared for a livedump.

After the command completes, the livedump is ready to be triggered, which you can see by examining the output from the following command:

```
svcinfolivedump <node id/name>
```

The status must be reported as prepared.

2. Run the following command to trigger the livedump:

```
svctask triggerlivedump <node id/name>
```

This command completes when the data capture is complete but before the dump file is written to disk.

3. Run the following command to query the status and copy the dump off when complete:

```
svcinfolivedump <nodeid/name>
```

The status is “dumping” when the file is being written to disk. The status is inactive after it is completed. After the status returns to the inactive state, you can find the livedump file in the /dumps folder on the node with a file name in the following format:

```
livedump.<panel_id>.<date>.<time>
```

You can then copy this file off the node as you copy a normal dump by using the GUI or SCP. Then, upload the dump to IBM Support for analysis.

15.2.3 SAN data collection

You can capture and collect switch support data. If problems exist that cannot be fixed by a simple maintenance task, such as exchanging hardware, an IBM Support representative asks you to collect the SAN data.

You can collect switch support data by using the IBM Network Adviser V11 for Brocade and McDATA SAN switches, and by using CLI commands to collect support data for a Brocade and a Cisco SAN switch.

IBM System Storage and IBM Network Adviser V11

You can use Technical Support to collect Support Save data (such as, RASLOG and TRACE) from Fabric OS devices.

Fabric OS level: The switch must be running Fabric OS 5.2.X or later to collect technical support data.

Complete the following steps:

1. Select **Monitor**  **Technical Support**  **Product/Host SupportSave**, as shown in Figure 15-5 on page 533.

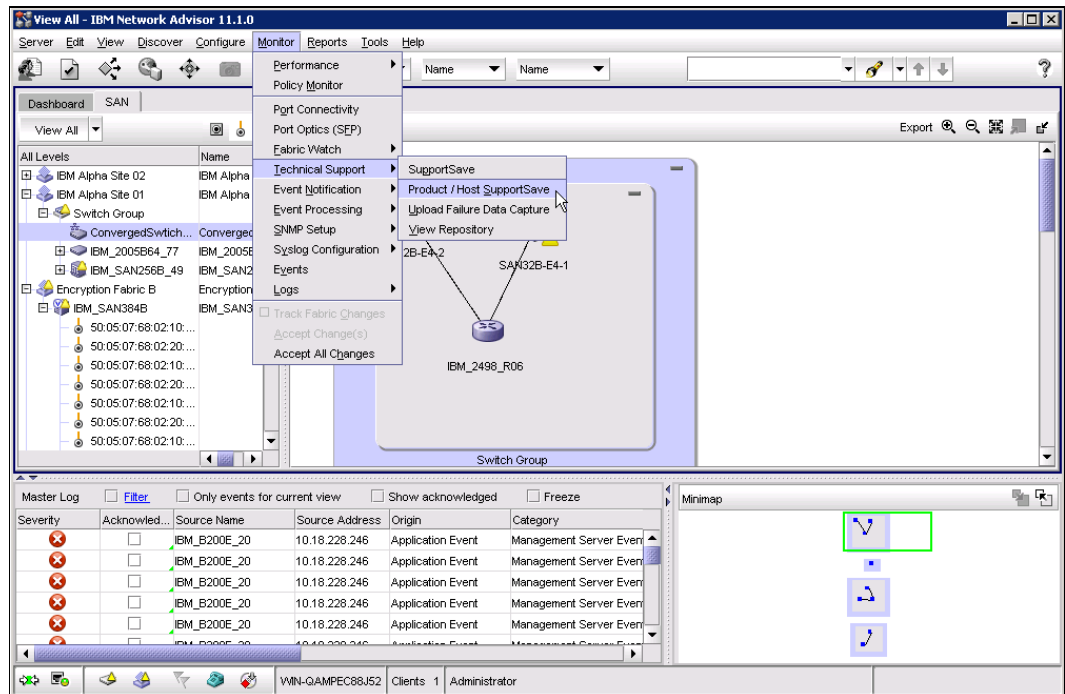


Figure 15-5 Product/Host SupportSave

2. In the Technical SupportSave dialog box (see Figure 15-6), select the switches that you want to collect data for in the Available SAN Products table. Click the right arrow to move them to the Selected Products and Hosts table. Then, click **OK**.

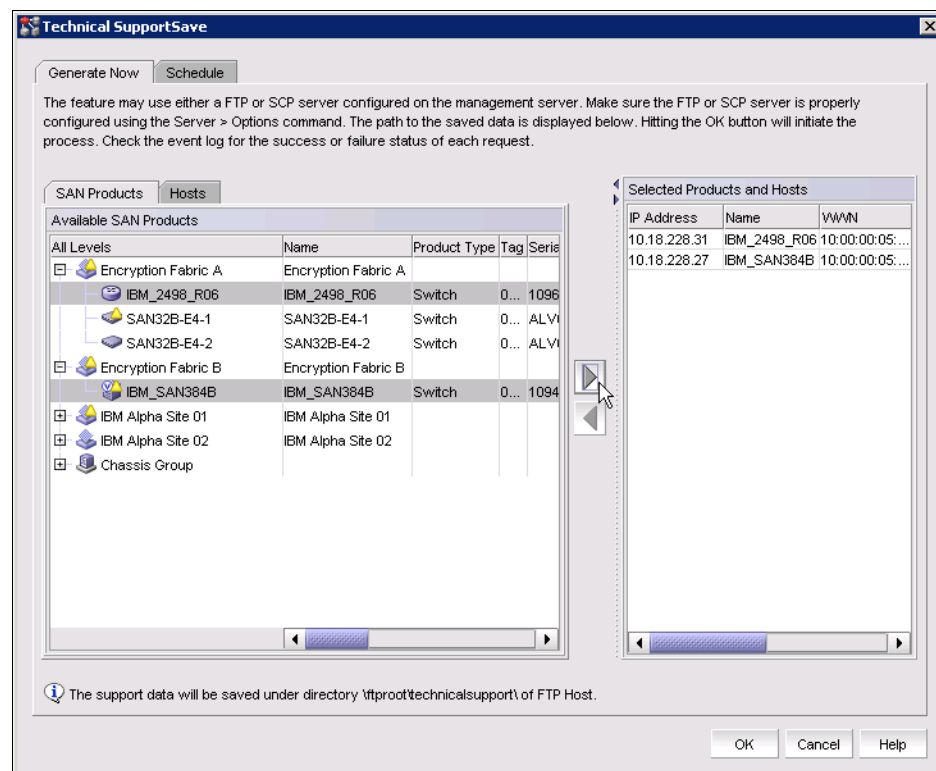


Figure 15-6 Technical SupportSave dialog box

The Technical SupportSave Status window opens, as shown in Figure 15-7.

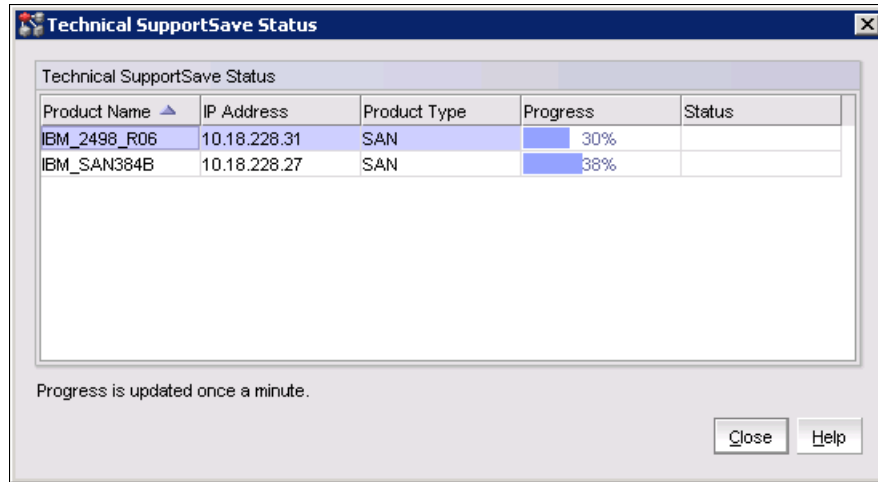


Figure 15-7 Technical SupportSave Status

Data collection can take 20 - 30 minutes for each selected switch. This estimate can increase depending on the number of switches selected.

3. To view and save the technical support information, click **Monitor** $\varnothing$ **Technical Support** $\varnothing$ **View Repository**, as shown in Figure 15-8.

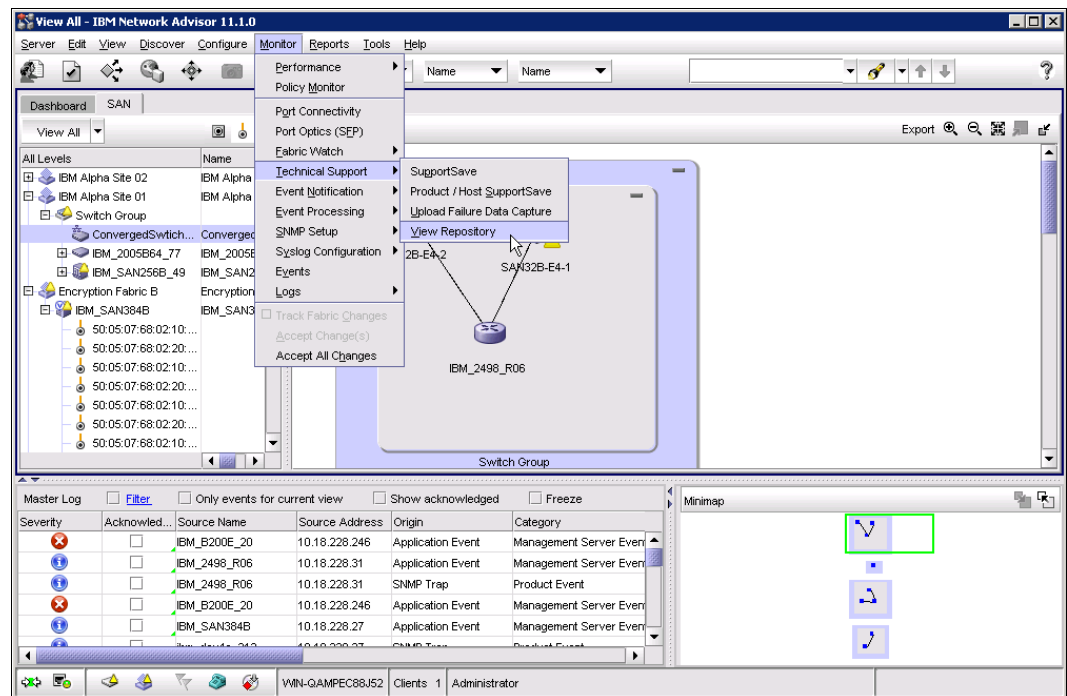


Figure 15-8 View Repository

4. In the Technical Support Repository display (see Figure 15-9), click **Save** to store the data on your system.

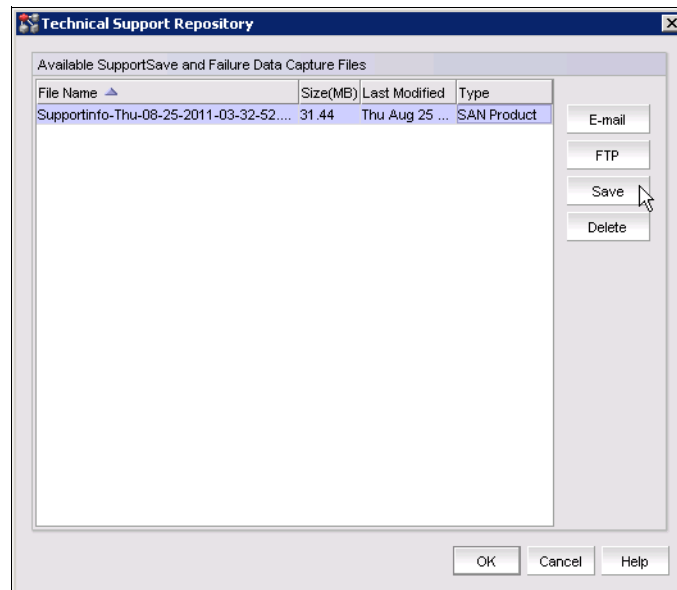


Figure 15-9 Technical Support Repository

You find a User Action Event in the Master Log when the download was successful, as shown in Figure 15-10.

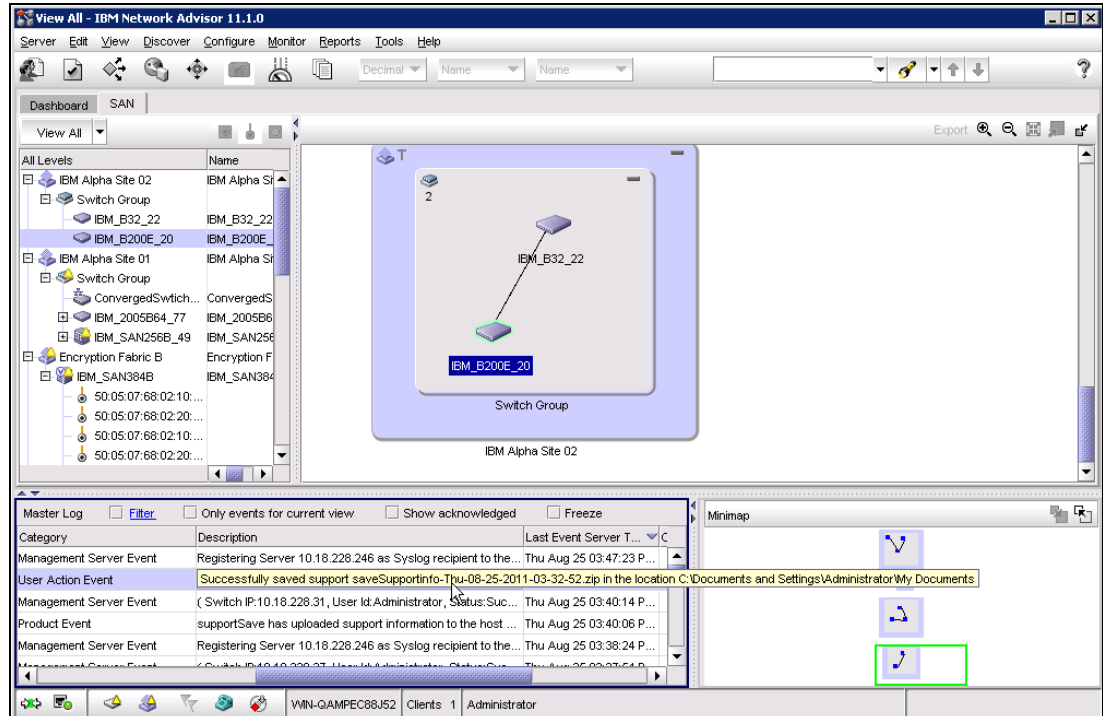


Figure 15-10 User Action Event

Gathering data: You can gather technical data for M-EOS (McDATA SAN switches) devices by using the Element Manager of the device.

IBM System Storage and Brocade SAN switches

For most of the current Brocade switches, run the **supportSave** command to collect the support data. Example 15-10 shows output from running the **supportSave** command (interactive mode) on an IBM System Storage SAN32B-3 (type 2005-B5K) SAN switch that is running Fabric OS v6.1.0c.

Example 15-10 The supportSave output from IBM SAN32B-3 switch (output shortened for clarity)

```
IBM_2005_B5K_1:admin> supportSave
This command will collect RASLOG, TRACE, supportShow, core file, FFDC data
and other support information and then transfer them to a FTP/SCP server
or a USB device. This operation can take several minutes.
NOTE: supportSave will transfer existing trace dump file first, then
automatically generate and transfer latest one. There will be two trace dump
files transfered after this command.
OK to proceed? (yes, y, no, n): [no] y

Host IP or Host Name: 9.43.86.133
User Name: fos
Password:
Protocol (ftp or scp): ftp
Remote Directory: /

Saving support information for switch:IBM_2005_B5K_1, module:CONSOLE0...
..._files/IBM_2005_B5K_1-S0-200808132042-CONSOLE0.gz: 5.77 kB 156.68 kB/s
Saving support information for switch:IBM_2005_B5K_1, module:RASLOG...
..._files/IBM_2005_B5K_1-S0-200808132042-RASLOG.ss.gz: 38.79 kB 0.99 MB/s
Saving support information for switch:IBM_2005_B5K_1, module:TRACE_OLD...
..._M_2005_B5K_1-S0-200808132042-old-tracedump.dmp.gz: 239.58 kB 3.66 MB/s
Saving support information for switch:IBM_2005_B5K_1, module:TRACE_NEW...
..._M_2005_B5K_1-S0-200808132042-new-tracedump.dmp.gz: 1.04 MB 1.81 MB/s
Saving support information for switch:IBM_2005_B5K_1, module:ZONE_LOG...
..._les/IBM_2005_B5K_1-S0-200808132042-ZONE_LOG.ss.gz: 51.84 kB 1.65 MB/s
Saving support information for switch:IBM_2005_B5K_1, module:RCS_LOG...
..._files/IBM_2005_B5K_1-S0-200808132044-CONSOLE1.gz: 5.77 kB 175.18 kB/s
Saving support information for switch:IBM_2005_B5K_1, module:SSAVELOG...
..._files/IBM_2005_B5K_1-S0-200808132044-sslog.ss.gz: 1.87 kB 55.14 kB/s
SupportSave completed
IBM_2005_B5K_1:admin>
```

IBM System Storage and Cisco SAN switches

Establish a terminal connection to the switch (Telnet, SSH, or serial), and collect the output from the following commands:

- ▶ terminal length 0
- ▶ show tech-support detail
- ▶ terminal length 24

15.2.4 Storage subsystem data collection

How you collect the data depends on the storage subsystem model. Here, you see only how to collect the support data for IBM System Storage subsystems.

IBM Storwize V7000

The management GUI and the service assistant have features to assist you in collecting the required information. The management GUI collects information from all the components in the system. The service assistant collects information from a single node canister. When the information that is collected is packaged together in a single file, the file is called a snap file.

Always follow the instructions that are provided by the support team to determine whether to collect the package by using the management GUI or by using the service assistant. Instruction is also provided for which package content option is required.

The use of the management GUI to collect the support data is similar to collecting the information about a SAN Volume Controller. For more information, see “Data collection for SAN Volume Controller by using the SAN Volume Controller Console GUI” on page 528.

If you choose the statesave option for the Support Package, you also receive Enclosure dumps for all the enclosures in the system.

IBM XIV Storage System

Complete the following steps to collect Support Logs from an IBM XIV Storage System:

1. Open the XIV GUI.
2. Click **Tools**  **Collect Support Logs**, as shown in Figure 15-11.

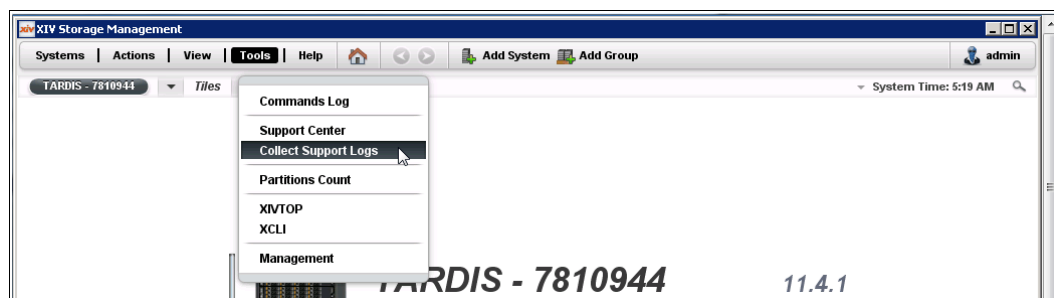


Figure 15-11 XIV Storage Management

3. In the Collect And Send Support Logs dialog box (see Figure 15-12 on page 538), click **Start** to auto collect and send the data and upload to the predefined FTP server. If the FTP server is not reachable, save the file locally and upload it manually to the Problem Management Report (PMR) system.

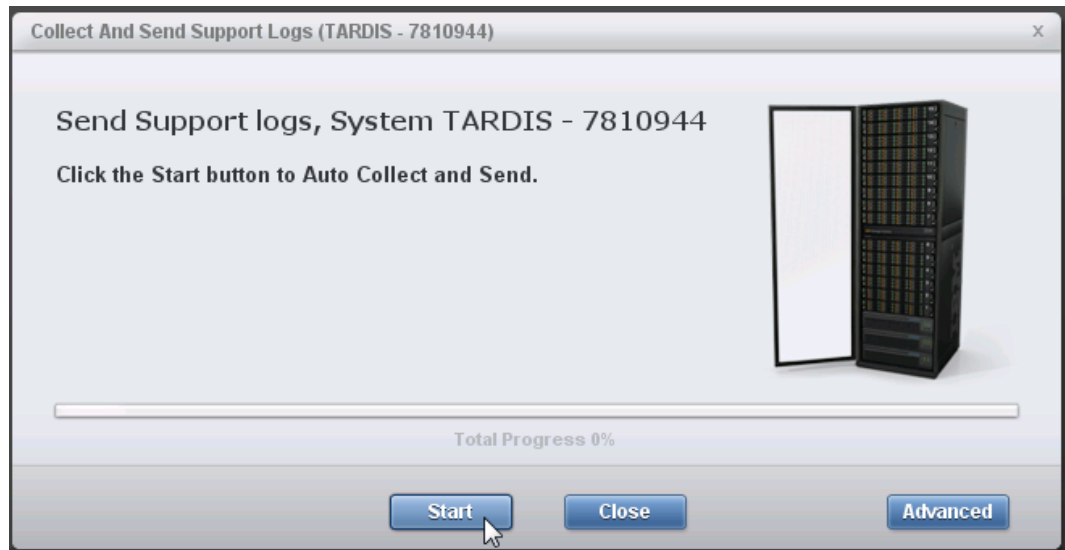


Figure 15-12 Collect the Support Logs

When the collecting process is complete, it shows up under System Log File Name panel, as shown in Figure 15-13.

4. Click **Advanced** $\emptyset$ **Select the log** $\emptyset$ **Get** to save the file on your system, as shown in Figure 15-13.

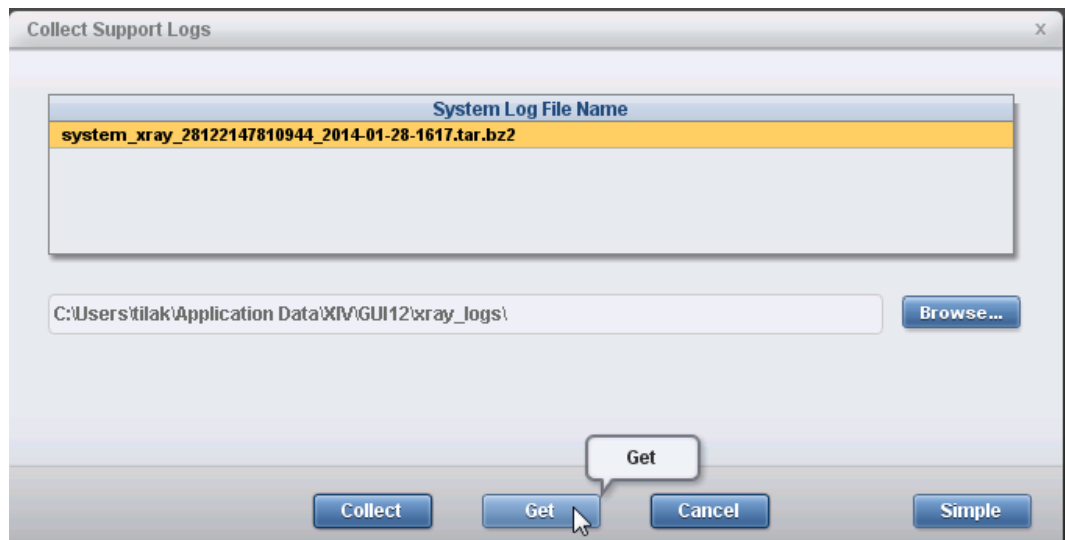


Figure 15-13 Getting the support logs

IBM System Storage DS4000 series

Storage Manager V9.1 and later include the Collect All Support Data feature. To collect the information, open the Storage Manager and click **Advanced** $\emptyset$ **Troubleshooting** $\emptyset$ **Collect All Support Data**, as shown in Figure 15-14 on page 539.

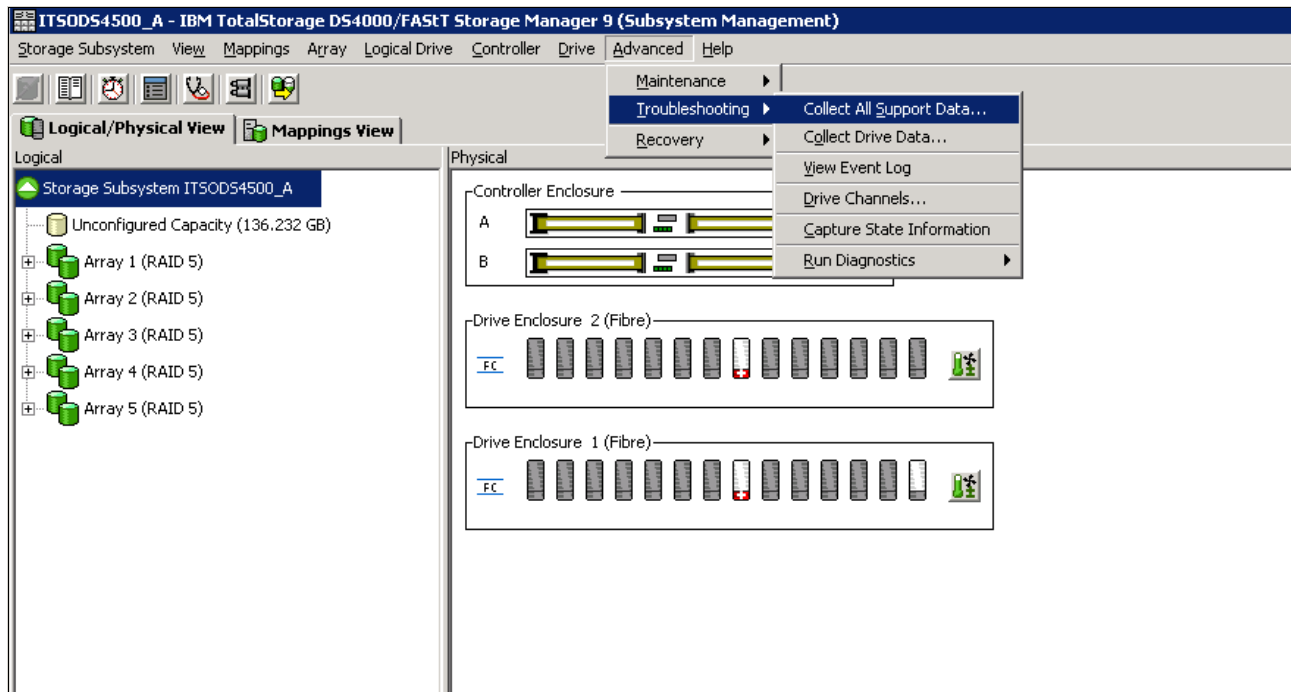


Figure 15-14 DS4000 data collection

IBM System Storage DS8000 and DS6000 series

Running the following series of commands gives you an overview of the current configuration of an IBM System Storage DS8000 or DS6000:

- ▶ `lss`
- ▶ `lsarray -l`
- ▶ `lsrank`
- ▶ `lsvolgrp`
- ▶ `lsfbvol`
- ▶ `lsioport -l`
- ▶ `lshostconnect`

The complete data collection task normally is performed by the IBM Service Support Representative (IBM SSR) or the IBM Support center. The IBM product engineering (PE) package includes all current configuration data and diagnostic data.

15.3 Recovering from problems

You can recover from several of the more common problems 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, see the following IBM Support website:

<http://www.ibm.com/support/entry/portal/Overview>

For more information about the latest flashes, concurrent code upgrades, code levels, and matrixes, see the following SAN Volume Controller website:

http://www-947.ibm.com/support/entry/portal/Overview/Hardware/System_Storage/Storage_software/Storage_virtualization/SAN_Volume_Controller_%282145%29

15.3.1 Solving host problems

Apart from hardware-related problems, 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 problems. Example 15-11 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 15-11 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 problems, 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
- ▶ Earlier HBA firmware
- ▶ Failures in the 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 15-12 shows that one of the HBAs was experiencing a link failure because of a fiber optic cable that bent over too far. After you change the cable, the missing paths reappeared.

Example 15-12 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	0

15.3.2 Solving SAN Volume Controller problems

For any problem in an environment that is implementing the SAN Volume Controller, use the Recommended Actions panel before you try to fix the problem anywhere else. Find the Recommended Actions panel by clicking **Monitoring** $\varnothing$ **Events** $\varnothing$ **Recommended Actions** in the SAN Volume Controller Console GUI, as shown in Figure 15-15.

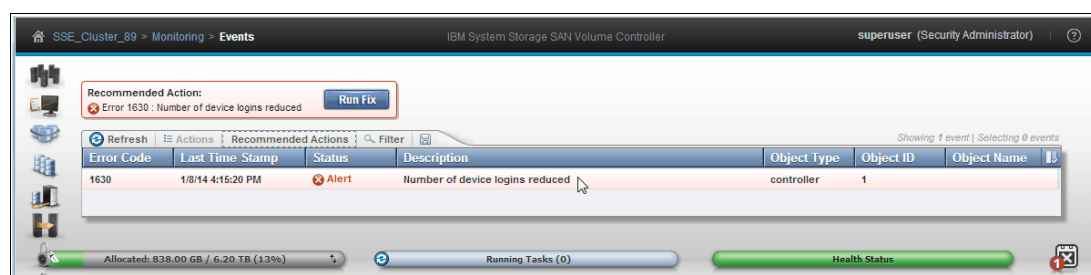


Figure 15-15 Recommended Action panel

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 15-16.

Tip: You can also obtain access to the Properties by right-clicking an event.

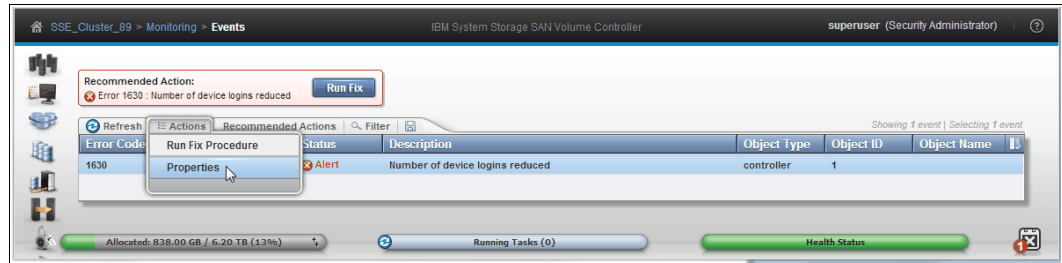


Figure 15-16 Event properties action

3. In the Properties and Sense Data for Event sequence_number window (see Figure 15-17, where sequence_number is the sequence number of the event that you selected in the previous step), review the information. Then, click **Close**.

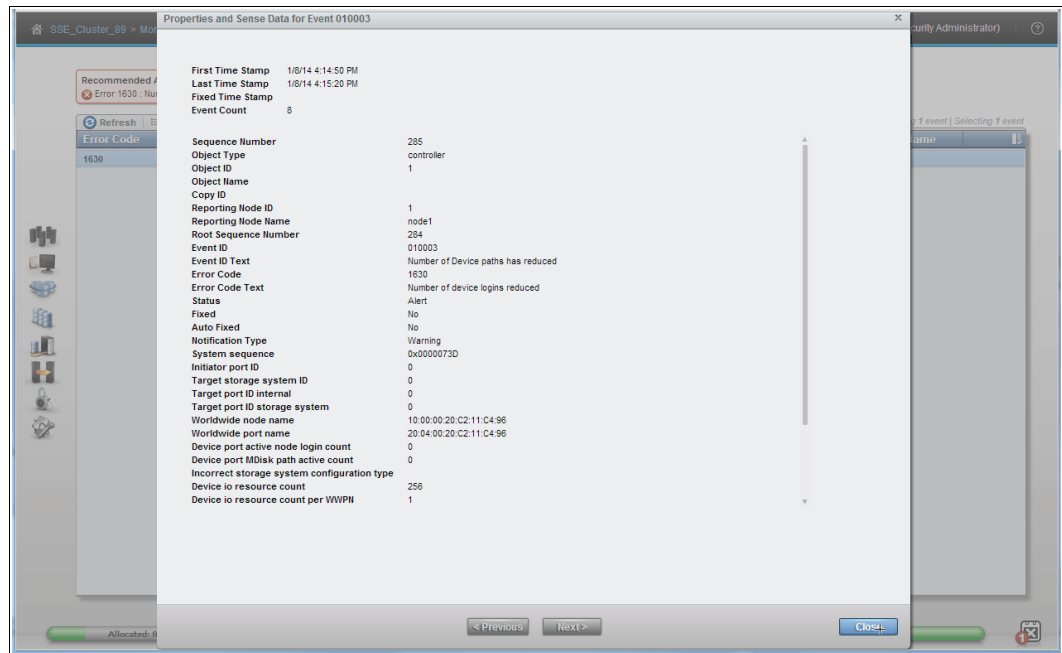


Figure 15-17 Properties and sense data for event window

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 SAN Volume Controller CLI to find problems. The following list of commands provides information about the status of your environment:

- ▶ `svctask detectmdisk`
Discovers changes in the back-end storage configuration.
- ▶ `svcinfolcluster clustername`
Checks the SAN Volume Controller cluster status.
- ▶ `svcinfolnode nodeid`
Checks the SAN Volume Controller nodes and port status.
- ▶ `svcinfolcontroller controllerid`
Checks the back-end storage status.
- ▶ `svcinfolsmdisk`
Provides a status of all the MDisk.
- ▶ `svcinfolsmdisk mdiskid`
Checks the status of a single MDisk.
- ▶ `svcinfolsmdiskgrp`
Provides a status of all the storage pools.
- ▶ `svcinfolsmdiskgrp mdiskgrpid`
Checks the status of a single storage pool.
- ▶ `svcinfolsvdisk`
Checks whether volumes are online.

Locating problems: Although the SAN Volume Controller raises error messages, most problems are not caused by the SAN Volume Controller. Most problems are introduced by the storage subsystems or the SAN.

If the problem is caused by the SAN Volume Controller and you are unable to fix it by using the Recommended Action panel or the event log, collect the SAN Volume Controller debug data as described in 15.2.2, “SAN Volume Controller data collection” on page 527.

To determine and fix other problems outside of SAN Volume Controller, consider the guidance in the other sections in this chapter that are not related to SAN Volume Controller.

Cluster upgrade checks

Before you perform an SAN Volume Controller cluster code load, complete the following prerequisite checks to confirm readiness:

- ▶ Check the back-end storage configurations for SCSI ID-to-LUN ID mappings. Normally, a 1625 error is detected if a problem occurs. However, you might also want to manually check these back-end storage configurations for SCSI ID-to-LUN ID mappings. Specifically, make sure that the SCSI ID-to-LUN ID is the same for each SAN Volume Controller node port.

You can use the following commands on the IBM Enterprise Storage Server® (ESS) to pull out the data to check ESS mapping:

```
esscli list port -d "ess=<ESS name>"
esscli list hostconnection -d "ess=<ESS name>"
esscli list volumeaccess -d "ess=<ESS name>"
```

Also, verify that the mapping is identical.

Use the following commands for an IBM System Storage DS8000 series storage subsystem to check the SCSI ID-to-LUN ID mappings:

```
lsioport -l  
lshostconnect -l  
showvolgrp -lunmap <volume group>  
lsfbvol -l -vol <SAN Volume Controller volume groups>
```

LUN mapping problems are unlikely on a storage subsystem that is based on DS800 because of the way that volume groups are allocated. However, it is still worthwhile to verify the configuration just before upgrades.

For the IBM System Storage DS4000 series, also verify that each SAN Volume Controller node port has an identical LUN mapping.

From the DS4000 Storage Manager, you can use the Mappings View to verify the mapping. You can also run the data collection for the DS4000 and use the subsystem profile to check the mapping.

- ▶ For storage subsystems from other vendors, use the corresponding steps to verify the correct mapping.
- ▶ Check the host multipathing to ensure path redundancy.
- ▶ Use the **svcinfo lsmdisk** and **svcinfo lscontroller** commands to check the SAN Volume Controller cluster to ensure the path redundancy to any back-end storage controllers.
- ▶ Use the “Run Maintenance Procedure” function or “Analyze Error Log” function in the SAN Volume Controller Console GUI to investigate any unfixed or investigated SAN Volume Controller errors.
- ▶ Download and run the following SAN Volume Controller Software Upgrade Test Utility:
<http://www.ibm.com/support/docview.wss?uid=ssg1S4000585>
- ▶ Review the latest flashes, hints, and tips before the cluster upgrade. The SAN Volume Controller code download page includes a list of directly applicable flashes, hints, and tips. Also, review the latest support flashes on the SAN Volume Controller support page.

15.3.3 Solving SAN problems

Various situations can cause problems 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 find:

- ▶ Switch power, fan, or cooling units
- ▶ Application-specific integrated circuit (ASIC)
- ▶ Installed SFP modules
- ▶ Fiber optic cables

Software failures are more difficult to analyze. In most cases, you must collect data and to involve IBM Support. But before you take any other steps, check the installed code level for any known problems. Also, check whether a new code level is available that resolves the problem that you are experiencing.

The most common SAN problems often are related to zoning. For example, perhaps you choose the wrong WWPN for a host zone, such as when two SAN Volume Controller node ports must be zoned to one HBA with one port from each SAN Volume Controller node. However, as shown in Example 15-13, 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 15-13.

Example 15-13 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 15-14.

Example 15-14 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 SAN Volume Controller 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 problem with these actions, use the method that is described in 15.2.3, “SAN data collection” on page 532, collect the SAN switch debugging data, and then contact IBM Support for assistance.

15.3.4 Solving back-end storage problems

The SAN Volume Controller is a useful tool for finding and analyzing back-end storage subsystem problems because it has a monitoring and logging mechanism.

However, it is not as helpful in finding problems from a host perspective, because the SAN Volume Controller is a SCSI target for the host and the SCSI protocol defines that errors are reported through the host.

Typical problems for storage subsystem controllers include incorrect configuration, which results in a 1625 error code. Other problems 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 problems and not the storage subsystem only. To determine the root cause of a problem, complete the following tasks:

- ▶ Check the Recommended Actions panel under SAN Volume Controller.
- ▶ 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** $\emptyset$ **Events** $\emptyset$ **Recommended Actions**, as shown in Figure 15-15 on page 541.

For more information about how to use the Recommended Actions panel, see the IBM System Storage SAN Volume Controller Information Center at this website:

<http://publib.boulder.ibm.com/infocenter/svc/ic/index.jsp>

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 open problems. Use the service or maintenance features that are provided with the storage subsystem to fix these problems.
 - b. Check whether the LUN masking is correct. When attached to the SAN Volume Controller, 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 SAN Volume Controller. Also, observe the SAN Volume Controller restrictions for back-end storage subsystems, which can be found at this website:
<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004510>
 - c. Run the **svcinfo lscontroller ID** command, and you see output similar to what you see in Example 15-15. As highlighted in the example, the MDisk and, therefore, the LUNs, are not equally allocated. In our example, the LUNs provided by the storage subsystem are visible only by one path, which is storage subsystem WWPN.

Example 15-15 The svcinfo lscontroller command output

```
IBM_2145:itsosvcc11:admin>svcinfo lscontroller 0
id 0
controller_name controller0
WWNN 200400A0B8174431
mdisk_link_count 2
max_mdisk_link_count 4
degraded no
vendor_id IBM
product_id_low 1742-900
product_id_high
product_revision 0520
ctrl_s/n
WWPN 200400A0B8174433
path_count 8
max_path_count 12
    WWPN 200500A0B8174433
path_count 0
max_path_count 8
```

This imbalance has the following possible causes:

- If the back-end storage subsystem implements a preferred controller design, perhaps the LUNs are all allocated to the same controller. This situation is likely with the IBM System Storage DS4000 series, and you can fix it by redistributing the LUNs evenly across the DS4000 controllers and then rediscovering the LUNs on the SAN Volume Controller.

Because a DS4500 storage subsystem (type 1742) was used in Example 15-15, you must check for this situation.

- Another possible cause is that the WWPN with zero count is not visible to all the SAN Volume Controller nodes through the SAN zoning or the LUN masking on the storage subsystem.

Use the SAN Volume Controller CLI command **svcinfo lsfabric 0** to confirm.

If you are unsure which of the attached MDisk has which corresponding LUN ID, run the SAN Volume Controller **svcinfo lsmdisk** CLI command (see Example 15-16). This command also shows to which storage subsystem a specific MDisk belongs (the controller ID).

Example 15-16 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				
600a0b800017423300000059469cf845		00000000000000000000000000000000		
2	mdisk2	online	managed	0
MDG-1		70.9GB	0000000000000000	
controller0				
600a0b800017443100000096469cf0e8		00000000000000000000000000000000		

In this case, the problem was with the LUN allocation across the DS4500 controllers. After you fix this allocation on the DS4500, a SAN Volume Controller MDisk rediscovery fixed the problem from the SAN Volume Controller perspective. Example 15-17 shows an equally distributed MDisk.

Example 15-17 Equally distributed MDisk on all available paths

```
IBM_2145:itsosvcc11:admin>svctask detectmdisk
```

```
IBM_2145:itsosvcc11:admin>svcinfo lscontroller 0
```

```
id 0
controller_name controller0
WWNN 200400A0B8174431
mdisk_link_count 2
max_mdisk_link_count 4
degraded no
vendor_id IBM
product_id_low 1742-900
product_id_high
product_revision 0520
ctrl_s/n
WWPN 200400A0B8174433
path_count 4
max_path_count 12
WWPN 200500A0B8174433
path_count 4
max_path_count 8
```

- d. In this example, the problem was solved by changing the LUN allocation. If step 2 does not solve the problem in your case, continue with step 3.

3. Check the SANs for switch problems or zoning failures.

Many situations can cause problems in the SAN. For more information, see 15.2.3, “SAN data collection” on page 532.

4. Collect all support data and involve IBM Support.

Collect the support data for the involved SAN, SAN Volume Controller, or storage systems as described in 15.2, “Collecting data and isolating the problem” on page 524.

Common error recovery steps by using the SAN Volume Controller CLI

For back-end SAN problems or storage problems, you can use the SAN Volume Controller CLI to perform common error recovery steps.

Although the maintenance procedures perform these steps, it is sometimes faster to run these commands directly through the CLI. Run these commands any time that you have the following issues:

- ▶ You experience a back-end storage issue (for example, error code 1370 or error code 1630).
- ▶ You performed maintenance on the back-end storage subsystems.

Important: Run these commands when back-end storage is configured or a zoning change occurs to ensure that the SAN Volume Controller follows the changes.

Common error recovery involves the following SAN Volume Controller CLI commands:

- ▶ `svctask detectmdisk`
Discovers the changes in the back end.
- ▶ `svcinfo lscontroller` and `svcinfo lsmdisk`
Provides overall status of all controllers and MDisks.
- ▶ `svcinfo lscontroller controllerid`
Checks the controller that was causing the problems and verifies that all the WWPNs are listed as you expect.
- ▶ `svctask includemdisk mdiskid`
For each degraded or offline MDisk.
- ▶ `svcinfo lsmdisk`
Determines whether all MDisks are now online.
- ▶ `svcinfo lscontroller controllerid`
Checks that the `path_counts` are distributed evenly across the WWPNs.

Finally, run the maintenance procedures on the SAN Volume Controller to fix every error.

15.4 Mapping physical LBAs to volume extents

SAN Volume Controller V4.3 provides new functions that makes it easy to find the volume extent to which a physical MDisk LBA maps, and to find the physical MDisk LBA to which the volume extent maps. This function might be useful in the following situations, among others:

- ▶ If a storage controller reports a medium error on a logical drive but the SAN Volume Controller is not yet taken MDisk offline, you might want to establish which volumes are affected by the medium error.
- ▶ When you investigate application interaction with thin-provisioned volumes (SEV), it can be useful to determine whether a volume LBA was allocated. If an LBA was allocated when it was not intentionally written to, it is possible that the application is not designed to work well with SEV.

Two new commands, **svctask lsmdisklba** and **svctask lsvdisklba**, are available. Their output varies depending on the type of volume (for example, thin-provisioned versus fully allocated) and type of MDisk (for example, quorum versus non-quorum). For more information, see the *IBM System Storage SAN Volume Controller V6.2.0 - Software Installation and Configuration Guide*, GC27-2286-01.

15.4.1 Investigating a medium error by using lsvdisklba

Assume that a medium error is reported by the storage controller at LBA 0x00172001 of MDisk 6. Example 15-18 shows the command to use to discover which volume is affected by this error.

Example 15-18 The lsvdisklba command to investigate the effect of an MDisk medium error

```
IBM_2145:itsosvcc11:admin>svcinfolsvdisklba -mdisk 6 -lba 0x00172001
```

vdisk_id	vdisk_name	copy_id	type	LBA	vdisk_start	vdisk_end
mdisk_start	mdisk_end					
0	diomede0	0	allocated	0x00102001	0x00100000	0x0010FFFF
0x00170000	0x0017FFFF					

This output shows the following information:

- ▶ This LBA maps to LBA 0x00102001 of volume 0.
- ▶ The LBA is within the extent that runs from 0x00100000 to 0x0010FFFF on the volume and from 0x00170000 to 0x0017FFFF on the MDisk. Therefore, the extent size of this storage pool is 32 MB.

Therefore, if the host performs I/O to this LBA, the MDisk goes offline.

15.4.2 Investigating thin-provisioned volume allocation by using lsmdisklba

After you use an application to perform I/O to a thin-provisioned volume, you might want to determine which extents were allocated real capacity, which you can check by using the **svcinfolsmdisklba** command. Example 15-19 on page 550 shows the difference in output between an allocated and an unallocated part of a volume.

Example 15-19 Using `lsmdisklba` to check whether an extent was allocated

```
IBM_2145:itsosvcc11:admin>svcinflba -vdisk 0 -lba 0x0
copy_id  mdisk_id  mdisk_name  type      LBA      mdisk_start  mdisk_end  vdisk_start  vdisk_end
0         6         mdisk6     allocated 0x00050000  0x00050000  0x0005FFFF  0x00000000  0x0000FFFF

IBM_2145:itsosvcc11:admin>svcinflba -vdisk 14 -lba 0x0
copy_id  mdisk_id  mdisk_name  type      LBA      mdisk_start  mdisk_end  vdisk_start  vdisk_end
0         6         mdisk6     unallocated 0x00000000  0x00000000  0x00000000  0x00000000  0x0000003F
```

Volume 0 is a fully allocated volume. Therefore, the MDisk LBA information is displayed as shown in Example 15-18 on page 549.

Volume 14 is a thin-provisioned volume to which the host has not yet performed any I/O. All of its extents are unallocated. Therefore, the only information that is shown by the `lsmdisklba` command is that it is unallocated and that this thin-provisioned grain starts at LBA 0x00 and ends at 0x3F (the grain size is 32 KB).

15.5 Medium error logging

Medium errors on back-end MDisks can be encountered by Host I/O and by SAN Volume Controller background functions, such as volume migration and FlashCopy. This section describes the detailed sense data for medium errors that are presented to the host and the SAN Volume Controller.

15.5.1 Host-encountered media errors

Data checks that are encountered on a volume from a host read request return a check condition status with Key/Code/Qualifier = 030000. Example 15-20 shows an example of the detailed sense data that is returned to an AIX host for an unrecoverable medium error.

Example 15-20 Sense data

```
LABEL:      SC_DISK_ERR2
IDENTIFIER:  B6267342

Date/Time:   Thu Aug  5 10:49:35 2008
Sequence Number: 4334
Machine Id:  00C91D3B4C00
Node Id:     testnode
Class:       H
Type:        PERM
Resource Name: hdisk34
Resource Class: disk
Resource Type: 2145
Location:     U7879.001.DQDFLVP-P1-C1-T1-W5005076801401FEF-L4000000000000
VPD:
    Manufacturer.....IBM
    Machine Type and Model.....2145
    ROS Level and ID.....0000
    Device Specific.(Z0).....0000043268101002
    Device Specific.(Z1).....0200604
    Serial Number.....60050768018100FF78000000000000F6

SENSE DATA
0A00 2800 001C ED00 0000 0104 0000 0000 0000 0000 0000 0000 0102 0000 F000 0300
```

```

0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000 0000 0000 0000 0800 0000 0000 0000 0000 0000 0000 0000 0000 0000
0000 0000 0000

```

The sense byte decode includes the following information:

- ▶ Byte 2 = SCSI Op Code (28 = 10-byte read)
- ▶ Bytes 4 - 7 = Logical block address for volume
- ▶ Byte 30 = Key
- ▶ Byte 40 = Code
- ▶ Byte 41 = Qualifier

15.5.2 SAN Volume Controller-encountered medium errors

Medium errors that are encountered by volume migration, FlashCopy, or volume Mirroring on the source disk are logically transferred to the corresponding destination disk for a maximum of 32 medium errors. If the 32 medium error limit is reached, the associated copy operation ends. Attempts to read destination error sites results in medium errors as though attempts were made to read the source media site.

Data checks that are encountered by SAN Volume Controller background functions are reported in the SAN Volume Controller error log as 1320 errors. The detailed sense data for these errors indicates a check condition status with Key, Code, and Qualifier = 03110B. Example 15-21 shows a SAN Volume Controller error log entry for an unrecoverable media error.

Example 15-21 Error log entry

```

Error Log Entry 1965
  Node Identifier      : Node7
  Object Type         : mdisk
  Object ID           : 48
  Sequence Number     : 7073
  Root Sequence Number : 7073
  First Error Timestamp : Thu Jul 24 17:44:13 2008
                        : Epoch + 1219599853
  Last Error Timestamp : Thu Jul 24 17:44:13 2008
                        : Epoch + 1219599853
  Error Count          : 21
  Error ID : 10025 : Amedia error has occurred during I/O to a Managed Disk
  Error Code           : 1320 : Disk I/O medium error
  Status Flag          : FIXED
  Type Flag            : TRANSIENT ERROR

```

```

40 11 40 02 00 00 00 00 00 00 00 02 28 00 58 59
6D 80 00 00 40 00 00 00 00 00 00 00 00 00 80 00
04 02 00 02 00 00 00 00 00 00 01 0A 00 00 80 00
02 03 11 0B 80 6D 59 58 00 00 00 00 08 00 C0 AA
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 0B 00 00 00 04 00 00 00 10 00 02 01

```

The sense byte is decoded as shown in the following information:

- ▶ Byte 12 = SCSI Op Code (28 = 10-byte read)
- ▶ Bytes 14 - 17 = Logical block address for MDisk
- ▶ Bytes 49 - 51 = Key, code, or qualifier

Locating medium errors: The storage pool can go offline as a result of error handling behavior in current levels of SAN Volume Controller microcode. This situation can occur when you attempt to locate medium errors on MDisks in the following examples:

- ▶ By scanning volumes with host applications, such as dd.
- ▶ By using SAN Volume Controller background functions, such as volume migrations and FlashCopy.

This behavior changes in future levels of SAN Volume Controller microcode. Check with IBM Support before you attempt to locate medium errors by any of these means.

Error code information: Consider the following points:

- ▶ Medium errors that are encountered on volumes log error code 1320 "Disks I/O Medium Error."
- ▶ If more than 32 medium errors are found when data is copied from one volume to another volume, the copy operation ends with log error code 1610 "Too many medium errors on Managed Disk."

15.5.3 Replacing a bad disk

Always run directed maintenance procedures (dmp) to replace a bad disk. SAN Volume Controller 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.

15.5.4 Health status during upgrade

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.



Part 4

Practical examples

This part shows practical examples of typical procedures that use the preferred practices that are highlighted in this IBM Redbooks publication. Some of the examples were taken from actual cases in production environment, and some examples were run in IBM Laboratories.



SAN Volume Controller scenarios

This chapter provides working scenarios to reinforce and demonstrate the information in this book. It includes the following sections:

- ▶ SAN Volume Controller upgrade with CF8 nodes and internal solid-state drives
- ▶ Handling Stuck SAN Volume Controller Code Upgrades
- ▶ Moving an AIX server
- ▶ Migrating to a new SAN Volume Controller by using Copy Services
- ▶ SAN Volume Controller scripting
- ▶ Migrating AIX cluster volumes off DS4700
- ▶ Easy Tier and FlashSystem planned outages
- ▶ Changing LUN ID presented to a VMware ESXi host

16.1 SAN Volume Controller upgrade with CF8 nodes and internal solid-state drives

This section describes a special case scenario upgrade. If you are upgrading SAN Volume Controller CF8 nodes from version 5.1.0.8 and you use the internal solid-state drives (SSDs), there are more steps in the upgrade procedure that you must perform. These steps are described in this section.

Note: Upgrading the nodes from 5.1 code directly to 7.x code is not supported. For more information about planning your upgrade path, see Concurrent Compatibility and Code Cross Reference for SAN Volume Controller, S1001707, which is available at this website:

<http://www.ibm.com/support/docview.wss?uid=ssg1S1001707>

Follow the procedure that is described when you are upgrading CF8 nodes with internal SSDs to the latest code levels.

You can upgrade a two-node, model CF8 SAN Volume Controller cluster with two internal SSDs (one per node) that were used in a separate managed disk group. This section describes how to perform the upgrade from version 5.1.0.8 to version 6.2.0.2. A GUI and a command-line interface (CLI) were used for SAN Volume Controller versions 5.1.0.8 and 6.2.0.2, but you can use only the CLI. Only the **svcupgradetest** utility can prevent you from performing this procedure entirely by using the GUI.

This scenario involves moving the current virtual disks (VDisks) by using the managed disk group of the existing SSDs into a managed disk group that uses regular MDisks from an IBM System Storage DS8000 for the upgrade process. As such, we can unconfigure the existing SSD managed disk group and place the SSD managed disks (MDisks) in unmanaged state before the upgrade. After the upgrade, we intend to include the same SSDs (now as a RAID array) into the same managed disk group (now storage pool) that received the volume disks by using IBM System Storage Easy Tier. Example 16-1 shows the existing configuration in preparation for the upgrade.

Example 16-1 SAN Volume Controller cluster existing managed disk groups, SSDs, and controllers in V5.1.0.8

[illegible]

Complete the following steps to upgrade the SAN Volume Controller code from V5 to V6.2:

1. Complete the steps that are described in 14.4.1, “Preparing for the upgrade” on page 498. Verify the attached servers, SAN switches, and storage controllers for errors. Define the current and target SAN Volume Controller code levels, which in this case are 5.1.0.8 and 6.2.0.2.
2. From IBM Storage Support website, download the following software:
 - SAN Volume Controller Console Software V6.1
 - SAN Volume Controller Upgrade Test Utility version 6.6 (latest)
 - SAN Volume Controller code release 5.1.0.10 (latest fix for current version)
 - SAN Volume Controller code release 6.2.0.2 (latest release)

IBM Storage Support is available at this website:

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

3. In the left pane of the IBM System Storage SAN Volume Controller window (see Figure 16-1), expand Service and Maintenance and select **Upgrade Software**.

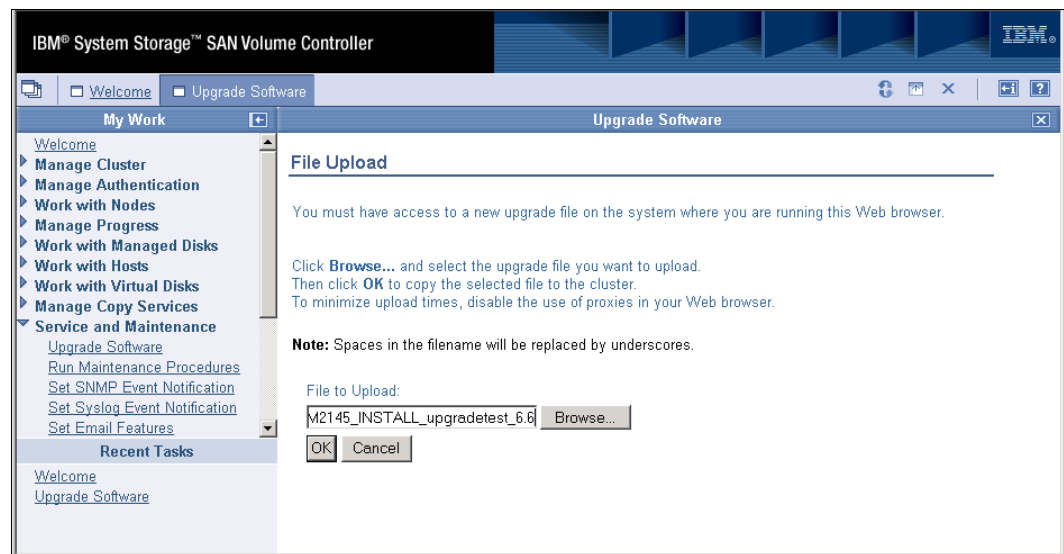


Figure 16-1 Upload SAN Volume Controller Upgrade Test Utility version 6.6

4. In the File to Upload field that is in the File Upload pane (on the right side of Figure 16-1), select the SAN Volume Controller Upgrade Test Utility. Click **OK** to copy the file to the cluster. Point the target version to SAN Volume Controller code release 5.1.0.10. Fix any errors that the Upgrade Test Utility finds before you proceed.

Important: Before you proceed, ensure that all servers that are attached to this SAN Volume Controller include compatible multipath software versions. You must also ensure that, for each server, the redundant disk paths are working error free. In addition, you must have a clean exit from the SAN Volume Controller Upgrade Test Utility.

5. Install SAN Volume Controller Code release 5.1.0.10 in the cluster.
6. In the Software Upgrade Status window (see Figure 16-2 on page 558), click **Check Upgrade Status** to monitor the upgrade progress.

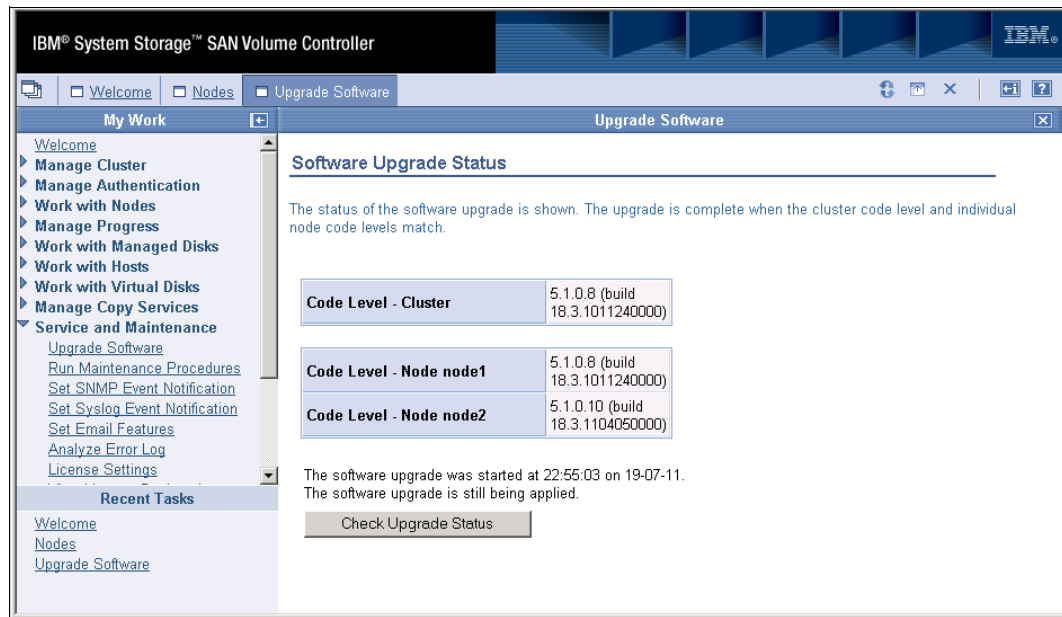


Figure 16-2 SAN Volume Controller Code upgrade status monitor by using the GUI

Example 16-2 shows how to monitor the upgrade by using the CLI.

Example 16-2 Monitoring the SAN Volume Controller code upgrade by using the CLI

```
IBM_2145:svccf8:admin>svcinfolsoftwareupgradestatus
status
upgrading
IBM_2145:svccf8:admin>
```

7. After the upgrade to SAN Volume Controller code release 5.1.0.10 is completed, check the SAN Volume Controller cluster again for any possible errors as a precaution.
8. Migrate the existing VDisks from the existing SSDs managed disk group. Example 16-3 shows a simple approach that uses the **migratevdisk** command.

Example 16-3 Migrating SAN Volume Controller VDisk by using the migratevdisk command

```
IBM_2145:svccf8:admin>svctask migratevdisk -mdiskgrp MDG4DS8KL3331 -vdisk
NYBIXTDB02_T03 -threads 2
IBM_2145:svccf8:admin>svcinfolsmigrate
migrate_type MDisk_Group_Migration
progress 5
migrate_source_vdisk_index 0
migrate_target_mdisk_grp 3
max_thread_count 2
migrate_source_vdisk_copy_id 0
IBM_2145:svccf8:admin>
```

Example 16-4 on page 559 shows another approach in which you add and then remove a VDisk mirror copy, which you can do even if the source and target managed disk groups have different extent sizes. Because this cluster does not use VDisk mirror copies before, you must first configure memory for the VDisk mirror bitmaps (**chiogrp**).

Use care with the **-syncrate** parameter to avoid any performance impact during the VDisk mirror copy synchronization. Changing this parameter from the default value of 50 to 55 as shown doubles the sync rate speed.

Example 16-4 SAN Volume Controller VDisk migration by using VDisk mirror copy

```
IBM_2145:svccf8:admin>svctask chiogrp -feature mirror -size 1 io_grp0
IBM_2145:svccf8:admin>svctask addvdiskcopy -mdiskgrp MDG4DS8KL3331 -syncrate 55 NYBIXTDB02_T03
Vdisk [0] copy [1] successfully created
IBM_2145:svccf8:admin>svcinfolsvdisk NYBIXTDB02_T03
id 0
name NYBIXTDB02_T03
IO_group_id 0
IO_group_name io_grp0
status online
mdisk_grp_id many
mdisk_grp_name many
capacity 20.00GB
type many
formatted no
mdisk_id many
mdisk_name many
FC_id
FC_name
RC_id
RC_name
vdisk_UID 60050768018205E12000000000000000
throttling 0
preferred_node_id 2
fast_write_state empty
cache readwrite
udid 0
fc_map_count 0
sync_rate 55
copy_count 2

copy_id 0
status online
sync yes
primary yes
mdisk_grp_id 2
mdisk_grp_name MDG3SVCCF8SSD
type striped
mdisk_id
mdisk_name
fast_write_state empty
used_capacity 20.00GB
real_capacity 20.00GB
free_capacity 0.00MB
overallocation 100
autoexpand
warning
grainsize

copy_id 1
status online
sync no
primary no
mdisk_grp_id 3
mdisk_grp_name MDG4DS8KL3331
type striped
mdisk_id
mdisk_name
fast_write_state empty
used_capacity 20.00GB
```

```

real_capacity 20.00GB
free_capacity 0.00MB
overallocation 100
autoexpand
warning
grainsize
IBM_2145:svccf8:admin>
IBM_2145:svccf8:admin>svctask addvdiskcopy -mdiskgrp MDG4DS8KL3331 -syncrate 75 NYBIXTDB02_T03
Vdisk [0] copy [1] successfully created
IBM_2145:svccf8:admin>svcinfolsvdiskcopy
vdisk_id  vdisk_name      copy_id  status  sync  primary  mdisk_grp_id  mdisk_grp_name  capacity  type
0          NYBIXTDB02_T03  0        online  yes   yes       2             MDG3SVCCF8SSD   20.00GB  striped
0          NYBIXTDB02_T03  1        online  no    no        3             MDG4DS8KL3331   20.00GB  striped
IBM_2145:svccf8:admin>
IBM_2145:svccf8:admin>svcinfolsvdiskcopy
vdisk_id  vdisk_name      copy_id  status  sync  primary  mdisk_grp_id  mdisk_grp_name  capacity  type
0          NYBIXTDB02_T03  0        online  yes   yes       2             MDG3SVCCF8SSD   20.00GB  striped
0          NYBIXTDB02_T03  1        online  yes   no        3             MDG4DS8KL3331   20.00GB  striped
IBM_2145:svccf8:admin>
IBM_2145:svccf8:admin>svctask rmvdiskcopy -copy 0 NYBIXTDB02_T03
IBM_2145:svccf8:admin>svcinfolsvdisk NYBIXTDB02_T03
id 0
name NYBIXTDB02_T03
IO_group_id 0
IO_group_name io_grp0
status online
mdisk_grp_id 3
mdisk_grp_name MDG4DS8KL3331
capacity 20.00GB
type striped
formatted no
mdisk_id
mdisk_name
FC_id
FC_name
RC_id
RC_name
vdisk_UID 60050768018205E12000000000000000
throttling 0
preferred_node_id 2
fast_write_state empty
cache readwrite
udid 0
fc_map_count 0
sync_rate 75
copy_count 1

copy_id 1
status online
sync yes
primary yes
mdisk_grp_id 3
mdisk_grp_name MDG4DS8KL3331
type striped
mdisk_id
mdisk_name
fast_write_state empty
used_capacity 20.00GB
real_capacity 20.00GB
free_capacity 0.00MB

```


[illegible]Chapter 16. SAN Volume Controller scenarios **561**

11. In the Software Upgrade Status window (see Figure 16-3), click **Check Upgrade Status** to monitor the upgrade progress. You notice the GUI changing its shape.

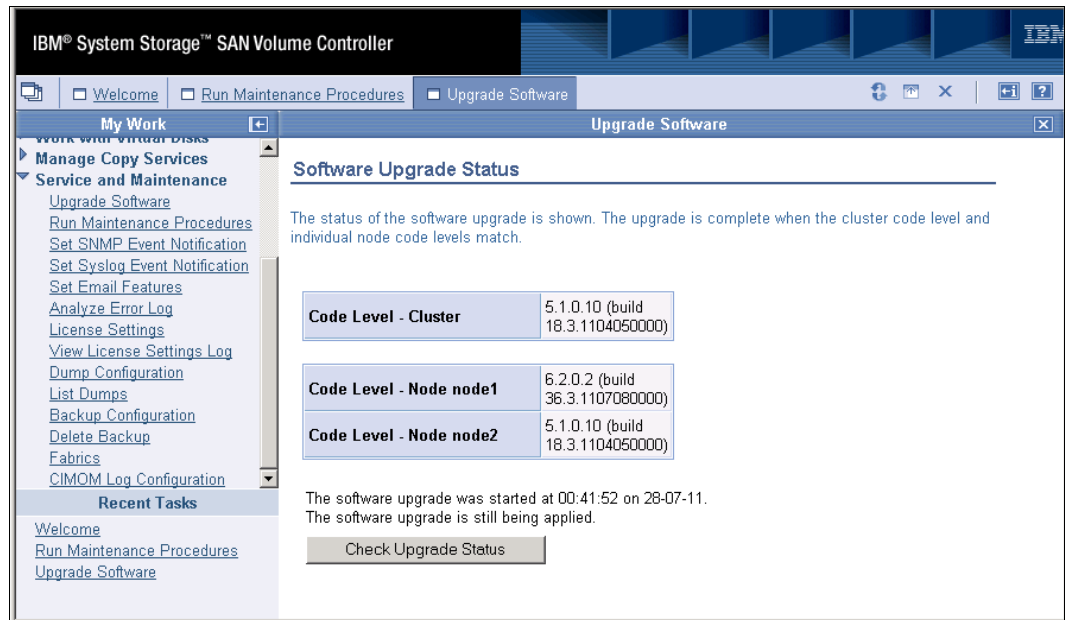


Figure 16-3 First SAN Volume Controller node being upgraded

Figure 16-4 shows that the second node is being upgraded.

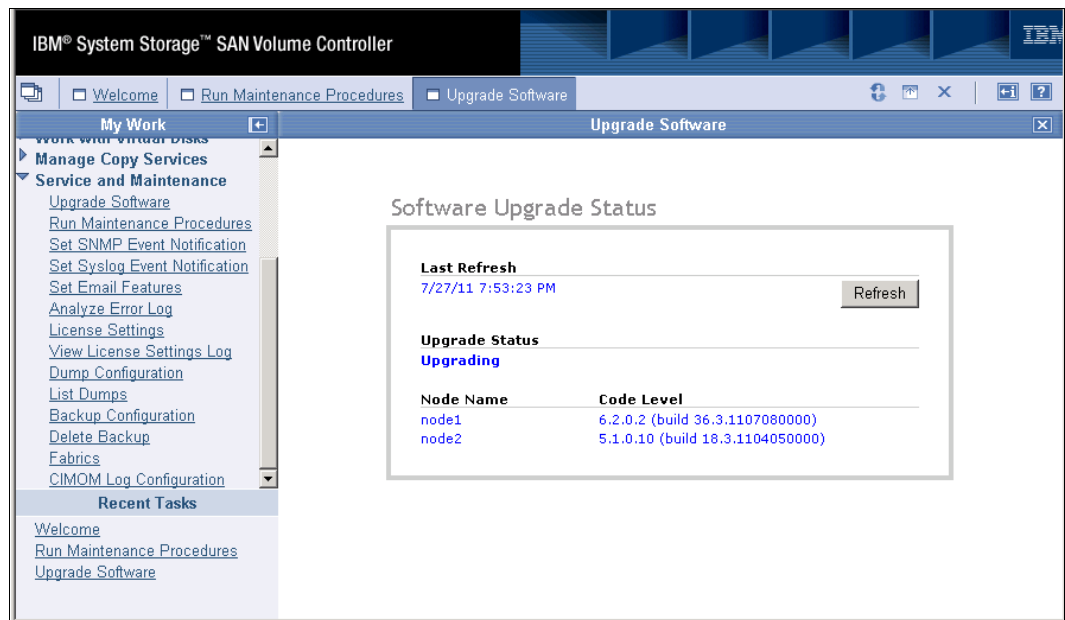


Figure 16-4 Second SAN Volume Controller node being upgraded

Figure 16-5 shows both nodes successfully upgraded.

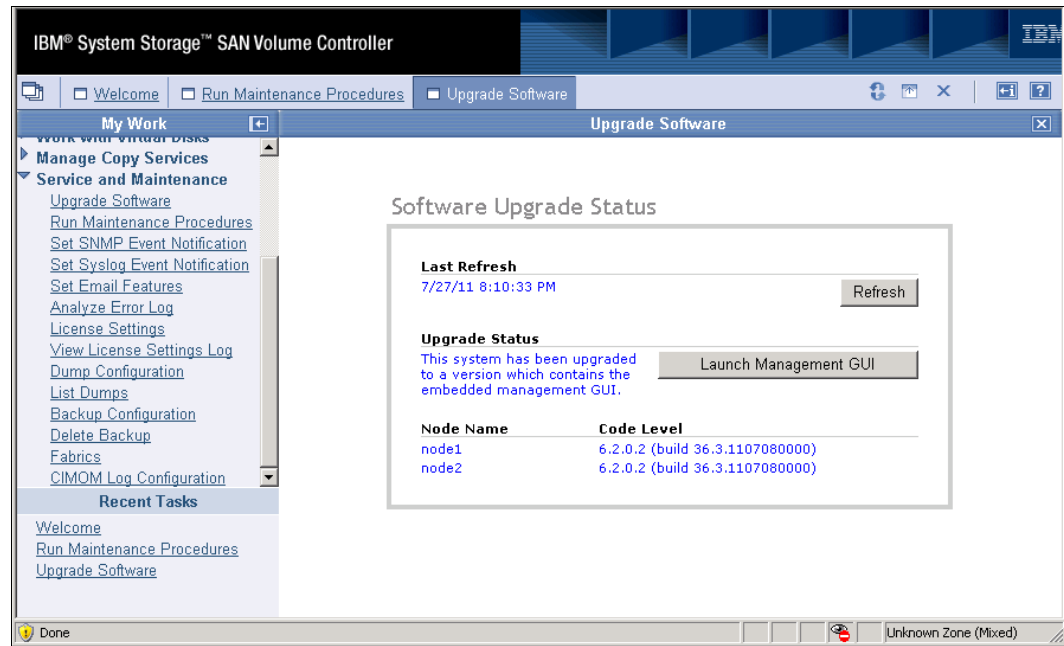


Figure 16-5 SAN Volume Controller cluster running SAN Volume Controller code release 6.2.0.2

12. After the upgrade is complete, click **Launch Management GUI** (see Figure 16-5) to restart the management GUI.

The management GUI now runs in one SAN Volume Controller node instead of the SAN Volume Controller console, as shown in Figure 16-6.

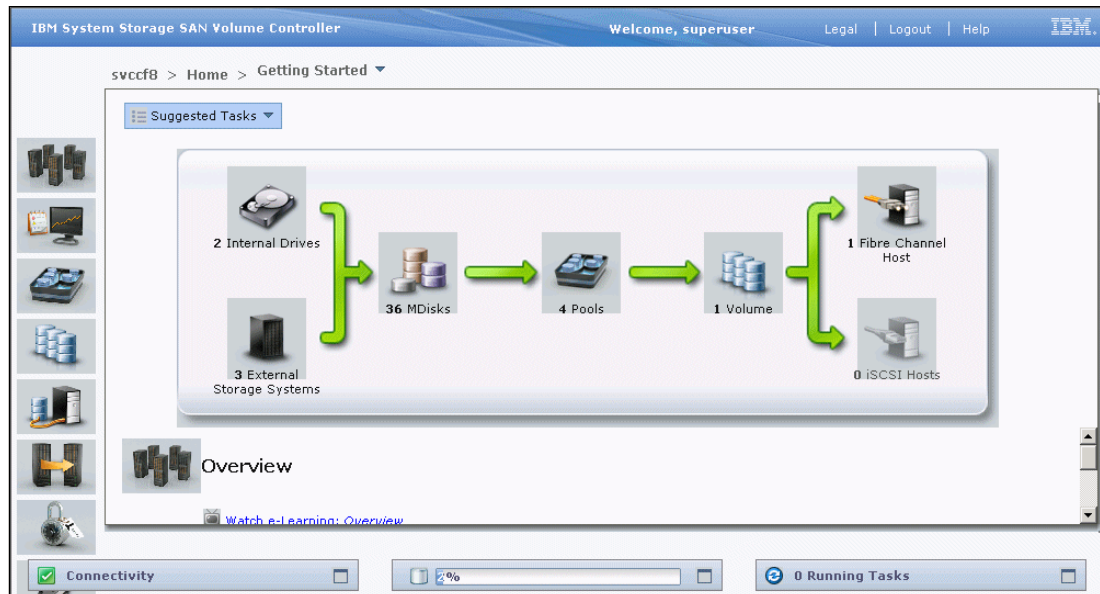


Figure 16-6 SAN Volume Controller V6.2.0.2 management GUI

13. As a precaution, check the SAN Volume Controller for errors.
14. Configure the internal SSDs that are used by the managed disk group that received the VDisks that were migrated in step 8 on page 558, but now use the Easy Tier function.

From the GUI home page (see Figure 16-7), click **Physical Storage** $\varnothing$ **Internal**. Then, on the Internal page, click **Configure Storage** in the upper left corner of the right pane.



Figure 16-7 Configure Storage button

15. Because two drives are unused, click **Yes** to continue when you are prompted about whether to include them in the configuration, as shown in Figure 16-8.

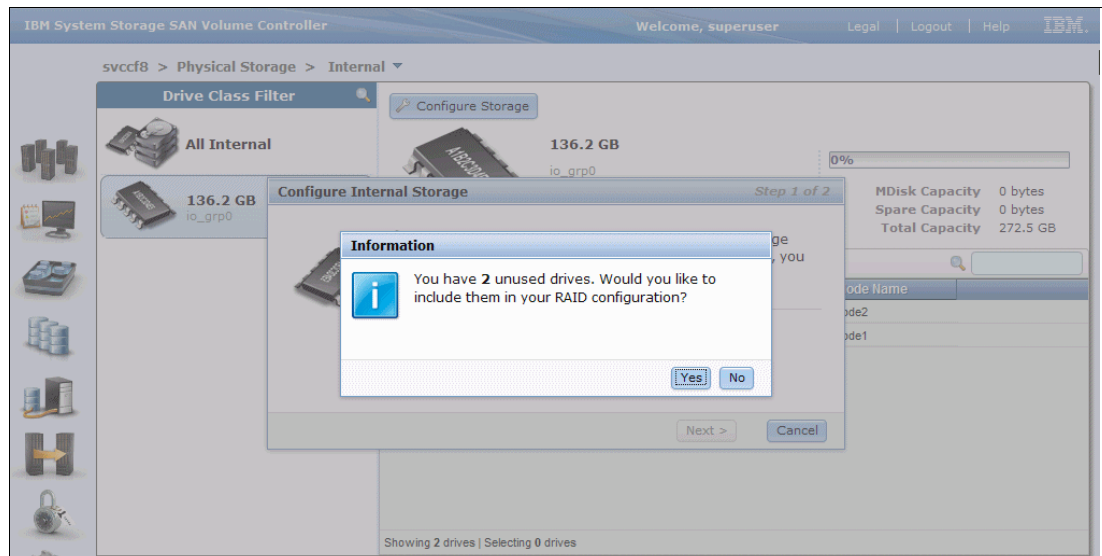


Figure 16-8 Confirming the number of SSDs to enable

Figure 16-9 shows the progress as the drives are marked as candidates.

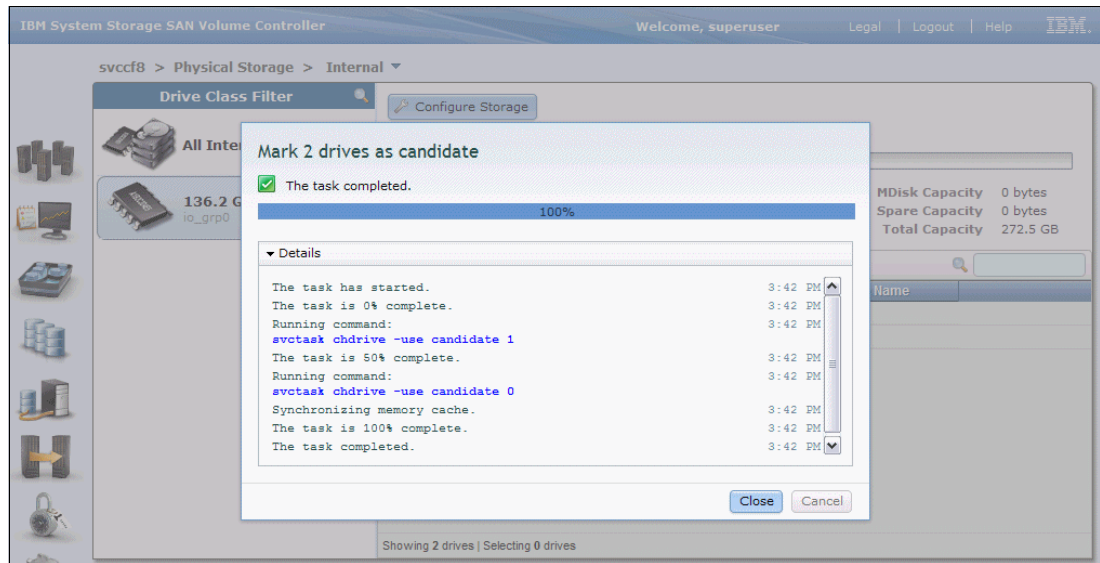


Figure 16-9 Enabling the SSDs as RAID candidates

16. Complete the following steps in the Configure Internal Storage window (see Figure 16-10):

- a. Select a RAID preset for the SSDs. For more information, see Table 14-2 on page 506.

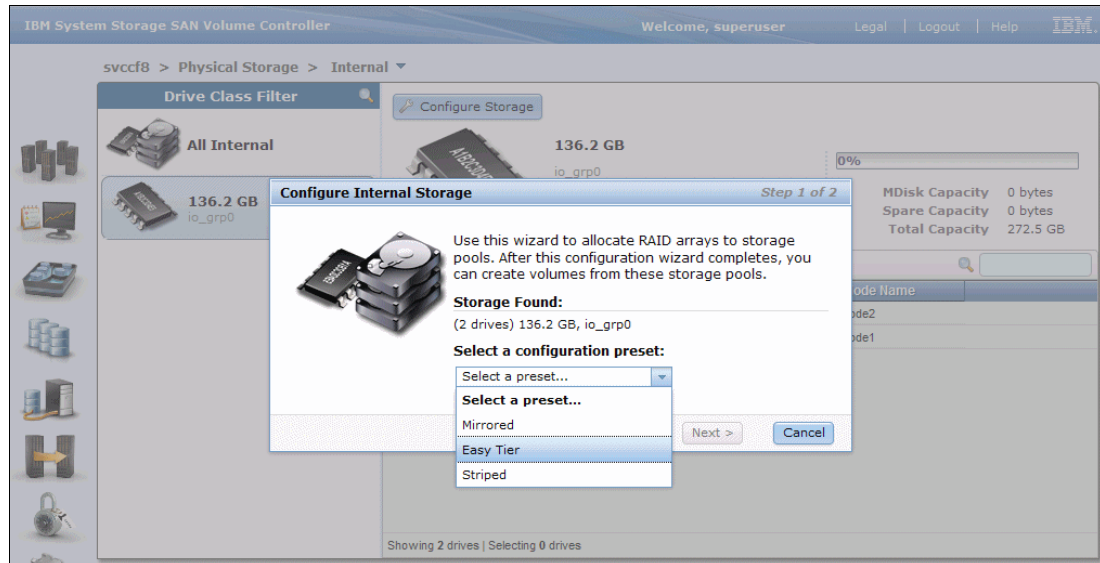


Figure 16-10 Selecting a RAID preset for the SSDs

- b. Confirm the number of SSDs (see Figure 16-11 on page 566) and the RAID preset.

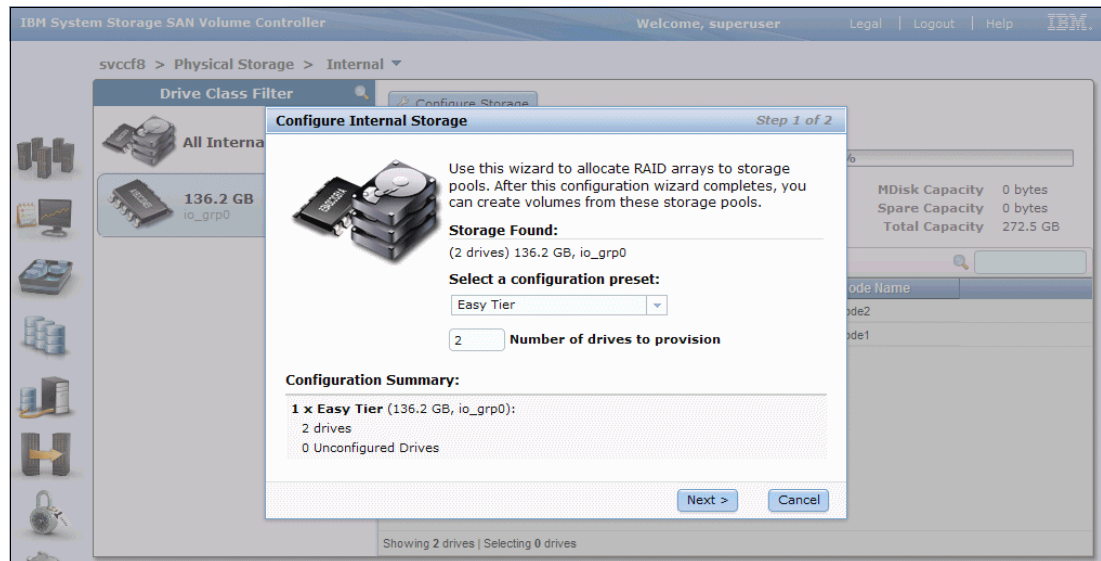


Figure 16-11 Configuration Wizard confirmation

c. Click **Next**.

17. Select the storage pool (former managed disk group) to include the SSDs, as shown in Figure 16-12. Click **Finish**.

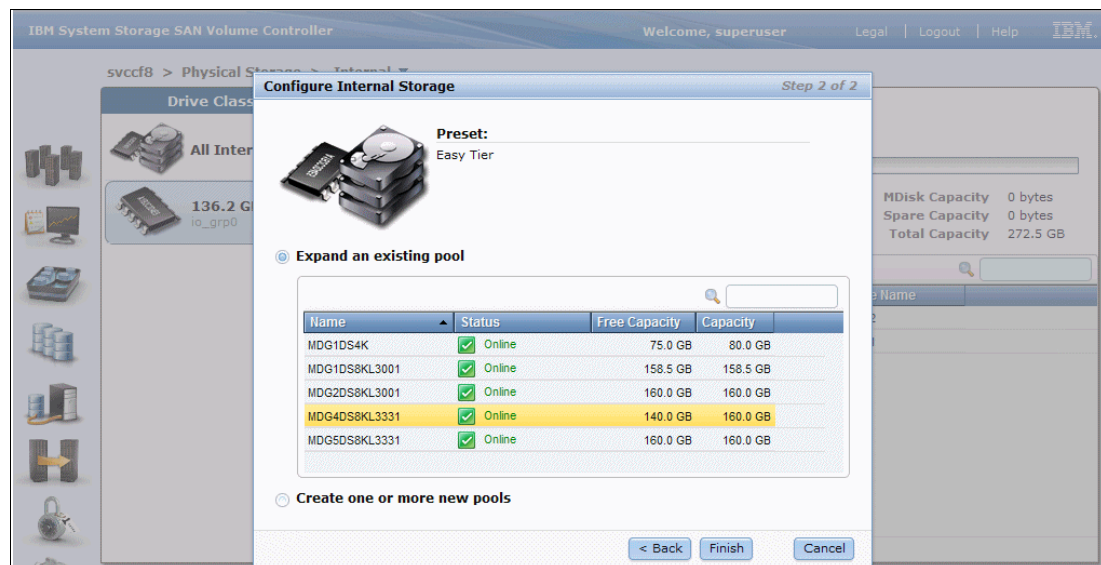


Figure 16-12 Selecting the storage pool for SSDs

18. In the Create RAID Arrays window (see Figure 16-13 on page 567), review the status. When the task is completed, click **Close**.

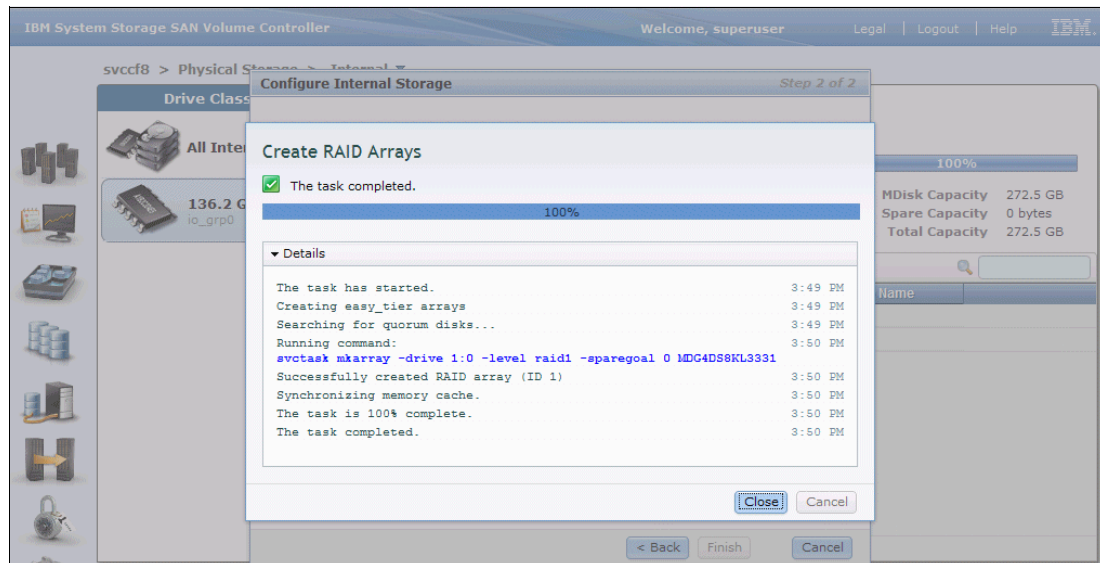


Figure 16-13 Create RAID Arrays dialog box

The SAN Volume Controller now continues the SSD array initialization process, but places the Easy Tier function of this pool in the Active state by collecting I/O data to determine which VDisk extents to migrate to the SSDs. You can monitor your array initialization progress in the lower right corner of the Tasks panel, as shown in Figure 16-14.

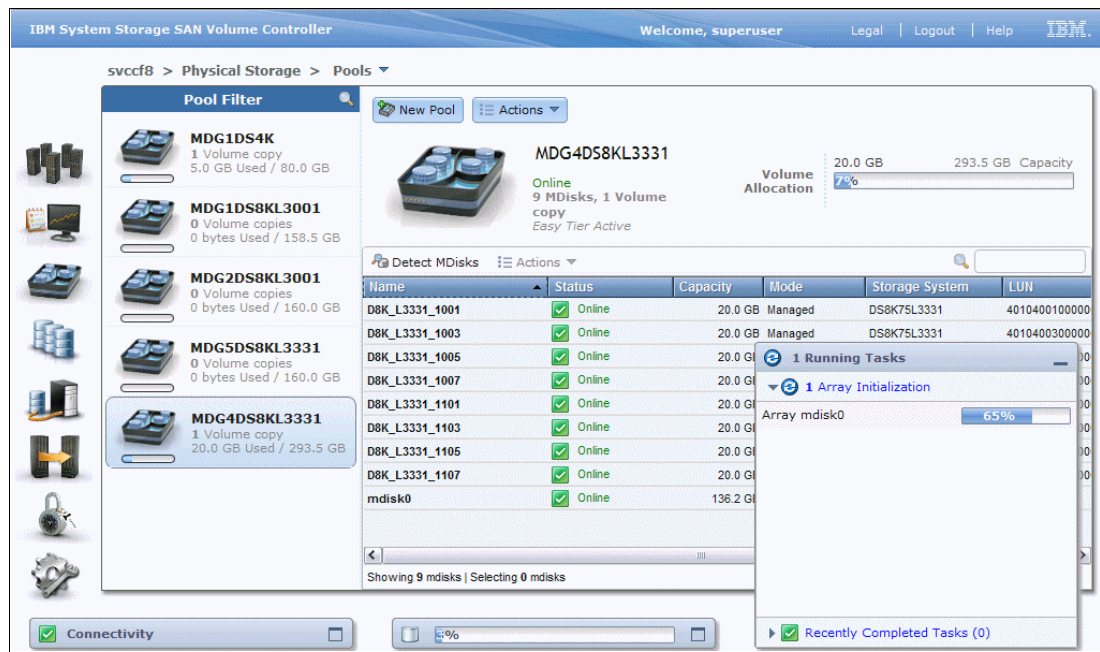


Figure 16-14 Monitoring the array initialization in the Tasks panel

The upgrade is finished. If you did not yet do so, plan your next steps into fine-tuning the Easy Tier function. If you do not have any other SAN Volume Controller clusters that are running SAN Volume Controller code V5.1 or earlier, you can install SAN Volume Controller Console code V6.

16.2 Handling Stuck SAN Volume Controller Code Upgrades

Upgrading the firmware of SAN Volume Controller/Storwize family device often is a straightforward process. However, although extremely rare, it is possible that an attempted firmware upgrade might get stuck. This means that the upgrade cannot proceed automatically any further and needs a service intervention to progress.

The first thing is to determine whether the upgrade is stuck because a normal upgrade can take a relatively long time. The upgrade process of an SAN Volume Controller/Storwize family storage device cluster has three phases. First, one node of each I/O group is upgraded, then the system waits for approximately 30 minutes for the hosts to settle their multipathing to the upgraded nodes, then the other half of the nodes is upgraded. Although a typical upgrade does not take more than approximately 60 minutes per node, certain combinations of hardware and firmware level can take longer. Also, a downgrade can last more than 60 minutes per node. Therefore, if you think that an upgrade is taking longer than it should, first, run the `svcinfo lssoftwareupgradestatus` command to verify the status of the cluster. If it does not indicate that the upgrade is stalled, wait for the progress of the upgrade.

However, an error can occur, which prevents the upgrade from completing. It does not have to be a hardware or software issue, it can be a configuration error. For example, if a SAN Volume Controller/Storwize family storage device node must be rebooted for the upgrade to progress but it is the only node with access to some back-end storage, the node refuses to reboot, which preserves access to the data.

If your code upgrade stalls or fails, do *not* take any recovery actions before you contact IBM Support. They can help you to complete the upgrade or to safely back off the upgrade attempt.

16.3 Moving an AIX server

In this case, an AIX server running in an IBM eServer pSeries logical partition (LPAR) is moved to another LPAR in a newer frame with a more powerful configuration. The server is brought down in a maintenance window. The SAN storage task is to switch over the SAN Volume Controller SAN LUNs that are used by the old LPAR to the new LPAR. Both LPARs use internal disks for their operating system rootvg volumes and have their own host bus adapters (HBAs) that are directly attached to the SAN. The scenario is identical if the source or the target LPAR uses NPIV technology to access the disks via Virtual I/O Server (VIOS). The SAN uses Brocade switches only.

You must replace only the HBAs worldwide port names (WWPNs) in the SAN aliases for both fabrics and in the SAN Volume Controller host definition.

Example 16-6 on page 569 shows the SAN Volume Controller and SAN commands. The procedure is the same regardless of the application and operating system.

In addition, the example includes the following information:

- ▶ Source (old) LPAR WWPNs: fcs0 - 10000000C9599F6C, fcs2 - 10000000C9594026
- ▶ Target (new) LPAR WWPNs: fcs0 - 10000000C99956DA, fcs2 - 10000000C9994E98
- ▶ The following SAN Volume Controller LUN IDs to be moved:
 - 60050768019001277000000000000030
 - 60050768019001277000000000000031
 - 600507680190012770000000000000146

- 60050768019001277000000000000147
- 60050768019001277000000000000148
- 60050768019001277000000000000149
- 6005076801900127700000000000014A
- 6005076801900127700000000000014B

Example 16-6 Commands to move the AIX server to another pSeries LPAR

```
###
### Verify that both old and new HBA WPNs are logged in both fabrics:
### Here an example in one fabric
###
b32sw1_B64:admin> nodefind 10:00:00:00:C9:59:9F:6C
Local:
Type Pid   COS      PortName      NodeName      SCR
N   401000;  2,3;10:00:00:00:c9:59:9f:6c;20:00:00:00:c9:59:9f:6c; 3
Fabric Port Name: 20:10:00:05:1e:04:16:a9
Permanent Port Name: 10:00:00:00:c9:59:9f:6c
Device type: Physical Unknown(initiator/target)
Port Index: 16
Share Area: No
Device Shared in Other AD: No
Redirect: No
Partial: No
Aliases: nybixpdb01_fcs0
b32sw1_B64:admin> nodefind 10:00:00:00:C9:99:56:DA
Remote:
Type Pid   COS      PortName      NodeName
N   4d2a00;  2,3;10:00:00:00:c9:99:56:da;20:00:00:00:c9:99:56:da;
Fabric Port Name: 20:2a:00:05:1e:06:d0:82
Permanent Port Name: 10:00:00:00:c9:99:56:da
Device type: Physical Unknown(initiator/target)
Port Index: 42
Share Area: No
Device Shared in Other AD: No
Redirect: No
Partial: No
Aliases:
b32sw1_B64:admin>

###
### Cross check SVC for HBAs WPNs and LUNid
###

IBM_2145:VIGSVC1:admin>
IBM_2145:VIGSVC1:admin>svcinflshost nybixpdb01
id 20
name nybixpdb01
port_count 2
type generic
mask 1111
iogrp_count 1
WWPN 10000000C9599F6C
node_logged_in_count 2
state active
WWPN 10000000C9594026
node_logged_in_count 2
state active
IBM_2145:VIGSVC1:admin>svcinflshostvdiskmap nybixpdb01
id      name      SCSI_id  vdisk_id  vdisk_name  wwpn      vdisk_UID
20      nybixpdb01  0        47        nybixpdb01_d01 10000000C9599F6C 60050768019001277000000000000030
20      nybixpdb01  1        48        nybixpdb01_d02 10000000C9599F6C 60050768019001277000000000000031
20      nybixpdb01  2        119       nybixpdb01_d03 10000000C9599F6C 60050768019001277000000000000146
20      nybixpdb01  3        118       nybixpdb01_d04 10000000C9599F6C 60050768019001277000000000000147
20      nybixpdb01  4        243       nybixpdb01_d05 10000000C9599F6C 60050768019001277000000000000148
20      nybixpdb01  5        244       nybixpdb01_d06 10000000C9599F6C 60050768019001277000000000000149
20      nybixpdb01  6        245       nybixpdb01_d07 10000000C9599F6C 6005076801900127700000000000014A
20      nybixpdb01  7        246       nybixpdb01_d08 10000000C9599F6C 6005076801900127700000000000014B
IBM_2145:VIGSVC1:admin>

###
### At this point both the old and new servers were brought down.
### As such, the HBAs would not be logged into the SAN fabrics, hence the use of the -force parameter.
### For the same reason, it makes no difference which update is made first - SAN zones or SVC host definitions
###

svctask addhostport -hbawwpn 10000000C99956DA -force nybixpdb01
svctask addhostport -hbawwpn 10000000C9994E98 -force nybixpdb01

svctask rmhostport -hbawwpn 10000000C9599F6C -force nybixpdb01
```

```

svctask rmhostport -hbawpn 10000000C9594026 -force nybixpdb01

### Alias WPN update in the first SAN fabric

aliadd "nybixpdb01_fcs0", "10:00:00:00:C9:99:56:DA"
aliremove "nybixpdb01_fcs0", "10:00:00:00:C9:59:9F:6C"
alishow nybixpdb01_fcs0

cfigsave
cfgenable "cr_BlueZone_FA"

### Alias WPN update in the second SAN fabric

aliadd "nybixpdb01_fcs2", "10:00:00:00:C9:99:4E:98"
aliremove "nybixpdb01_fcs2", "10:00:00:00:c9:59:40:26"
alishow nybixpdb01_fcs2

cfigsave
cfgenable "cr_BlueZone_FB"

### Back to SVC to monitor as the server is brought back up

IBM_2145:VIGSVC1:admin>svcinflshostvdiskmap nybixpdb01
id          name          SCSI_id      vdisk_id      vdisk_name      wwpn          vdisk_UID
20          nybixpdb01          0            47            nybixpdb01_d01  10000000C9994E98 60050768019001277000000000000030
20          nybixpdb01          1            48            nybixpdb01_d02  10000000C9994E98 60050768019001277000000000000031
20          nybixpdb01          2            119           nybixpdb01_d03  10000000C9994E98 60050768019001277000000000000146
20          nybixpdb01          3            118           nybixpdb01_d04  10000000C9994E98 60050768019001277000000000000147
20          nybixpdb01          4            243           nybixpdb01_d05  10000000C9994E98 60050768019001277000000000000148
20          nybixpdb01          5            244           nybixpdb01_d06  10000000C9994E98 60050768019001277000000000000149
20          nybixpdb01          6            245           nybixpdb01_d07  10000000C9994E98 6005076801900127700000000000014A
20          nybixpdb01          7            246           nybixpdb01_d08  10000000C9994E98 6005076801900127700000000000014B
IBM_2145:VIGSVC1:admin>svcinflshost nybixpdb01
id 20
name nybixpdb01
port_count 2
type generic
mask 1111
iogrp_count 1
WWPN 10000000C9994E98
node_logged_in_count 2
state inactive
WWPN 10000000C99956DA
node_logged_in_count 2
state inactive
IBM_2145:VIGSVC1:admin>

IBM_2145:VIGSVC1:admin>svcinflshost nybixpdb01
id 20
name nybixpdb01
port_count 2
type generic
mask 1111
iogrp_count 1
WWPN 10000000C9994E98
node_logged_in_count 2
state active
WWPN 10000000C99956DA
node_logged_in_count 2
state active
IBM_2145:VIGSVC1:admin>

```

After the new LPAR shows both its HBAs as active, you can confirm that it recognized all SAN disks that were previously assigned and that they all had healthy disk paths.

16.4 Migrating to a new SAN Volume Controller by using Copy Services

In this case, you migrate several servers from one SAN Volume Controller SAN storage infrastructure to another. Although the original case asked for this move for accounting reasons, you can use this scenario to renew your entire SAN storage infrastructure for SAN Volume Controller, as described in 14.6.3, “Moving to a new SAN Volume Controller cluster” on page 514.

The initial configuration was the typical SAN Volume Controller environment with a 2-node cluster, a DS8000 series as a back-end storage controller, and servers that are attached through redundant, independent SAN fabrics, as shown in Figure 16-15.

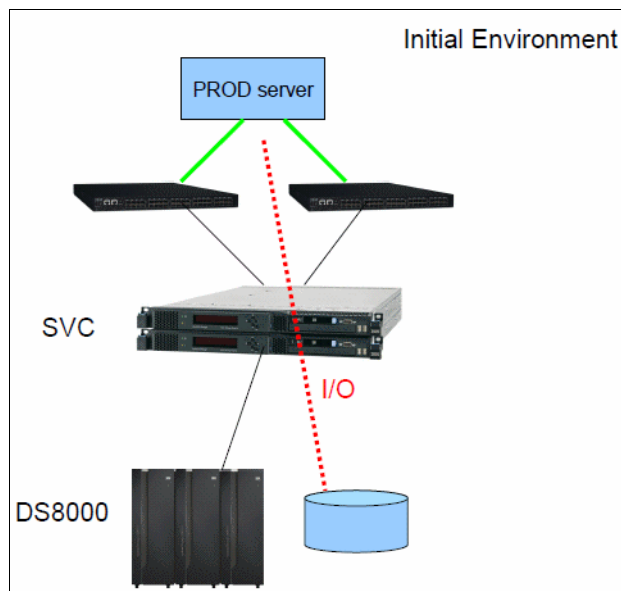


Figure 16-15 Initial SAN Volume Controller environment

By using SAN Volume Controller Copy Services to move the data from the old infrastructure to the new infrastructure, you can do so with the production servers and applications still running. You can also fine-tune the replication speed as you attempt to achieve the fastest possible migration without causing any noticeable performance degradation.

This scenario asks for a brief, planned outage to restart each server from one infrastructure to the other. Alternatives are possible to perform this move fully online. However, in our case, we had a pre-scheduled maintenance window every weekend and kept an integral copy of the server's data before the move, which allows a quick back out if required.

The new infrastructure is installed and configured with the new SAN switches that are attached to the existing SAN fabrics (preferably by using trunks for bandwidth) and the new SAN Volume Controller ready to use, as shown in Figure 16-16 on page 572.

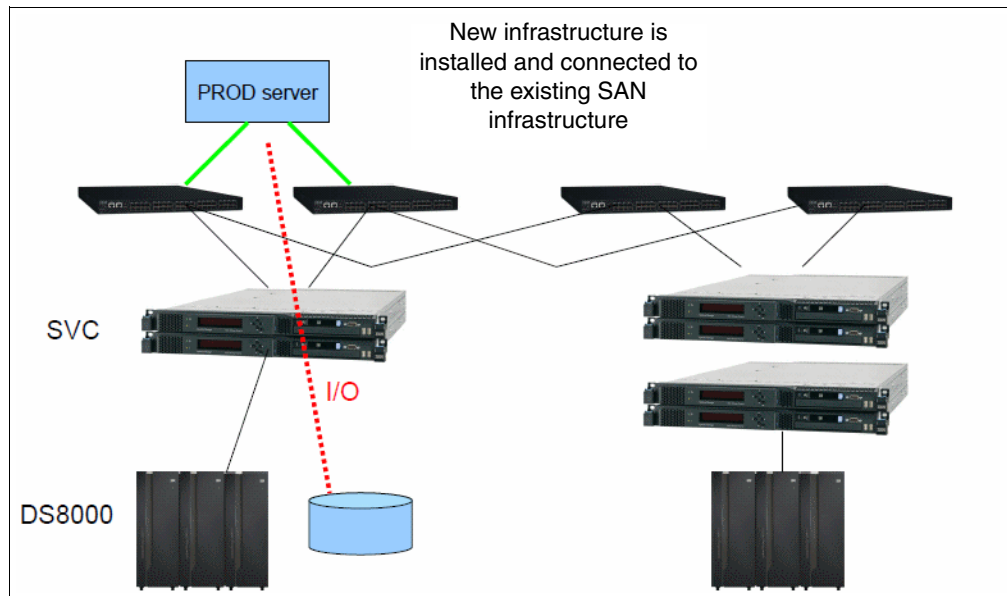


Figure 16-16 New SAN Volume Controller and SAN installed

Also, the necessary SAN zoning configuration is made between the initial and the new SAN Volume Controller clusters, and a remote copy partnership is established between them (notice the **-bandwidth** parameter). Then, for each VDisk in use by the production server, we created a target VDisk in the new environment with the same size and a remote copy relationship between these VDisks. We included this relationship in a consistency group.

The initial VDisks synchronization was started, which took some time for the copies to become synchronized considering the large amount of data and the bandwidth that stayed at its default value as a precaution.

Example 16-7 shows the SAN Volume Controller commands to set up the remote copy relationship.

Example 16-7 SAN Volume Controller commands to set up a remote copy relationship

```
SVC commands used in this phase:
# lscluster
# mkpartnership -bandwidth <bw> <svcpartnercluster>
# mkvdisk -mdiskgrp <mdg> -size <sz> -unit gb -iogrp <iogrp> -vtype striped -node <node> -name <targetvdisk> -easytier off
# mkrcconsistgrp -name <cgname> -cluster <svcpartnercluster>
# mkrcrelationship -master <sourcevdisk> -aux <targetvdisk> -name <rlname> -consistgrp <cgname> -cluster <svcpartnercluster>
# starttrconsistgrp -primary master <cgname>
# chpartnership -bandwidth <newbw> <svcpartnercluster>
```

Figure 16-17 shows the initial remote copy relationship setup that results from successful completion of the commands.

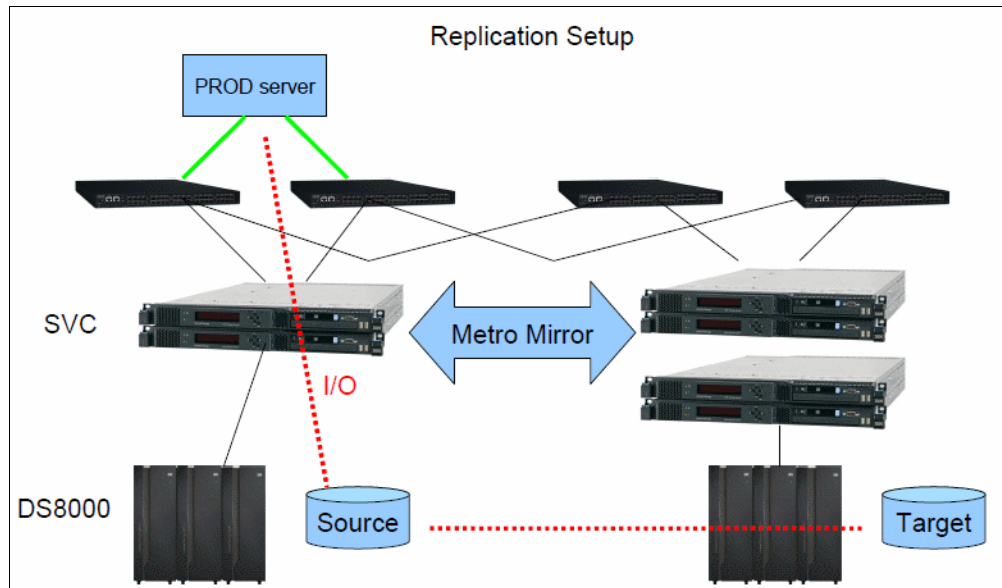


Figure 16-17 Initial SAN Volume Controller remote copy relationship setup

After the initial synchronization finished, a planned outage was scheduled to reconfigure the server to use the new SAN Volume Controller infrastructure. Figure 16-18 shows what happened in the planned outage. The I/O from the production server is quiesced and the replication session is stopped.

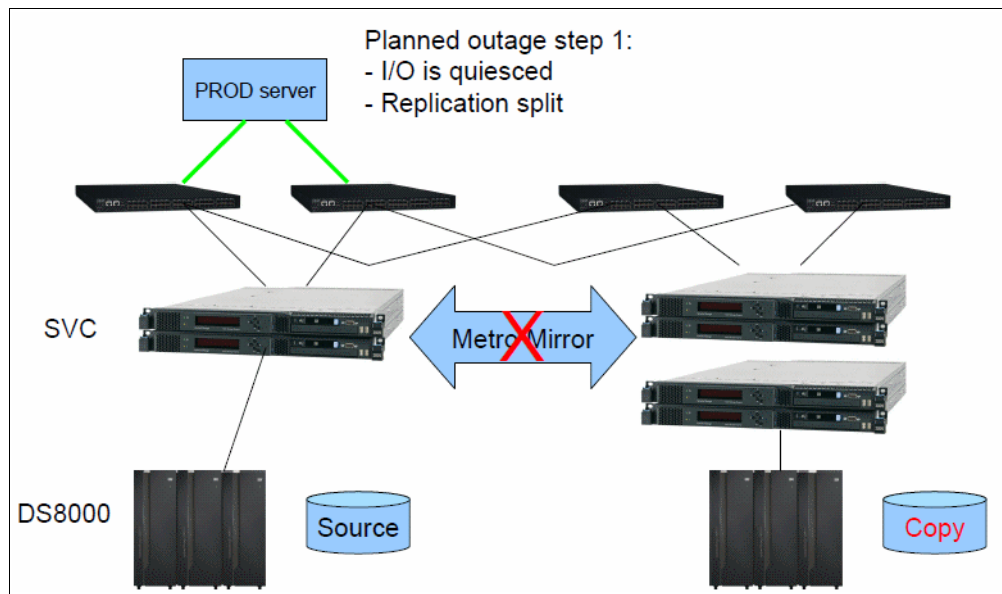


Figure 16-18 Planned outage to switch over to the new SAN Volume Controller

The next step is to move the fiber connections, as shown in Figure 16-19.

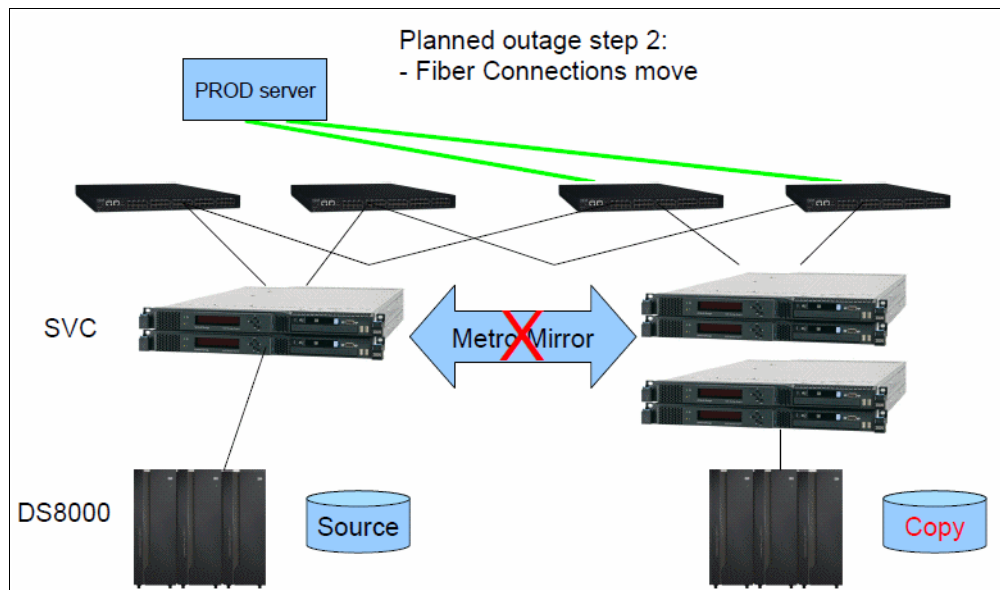


Figure 16-19 Moving the fiber connections to the new SAN

With the server reconfigured, the application is restarted, as shown in Figure 16-20.

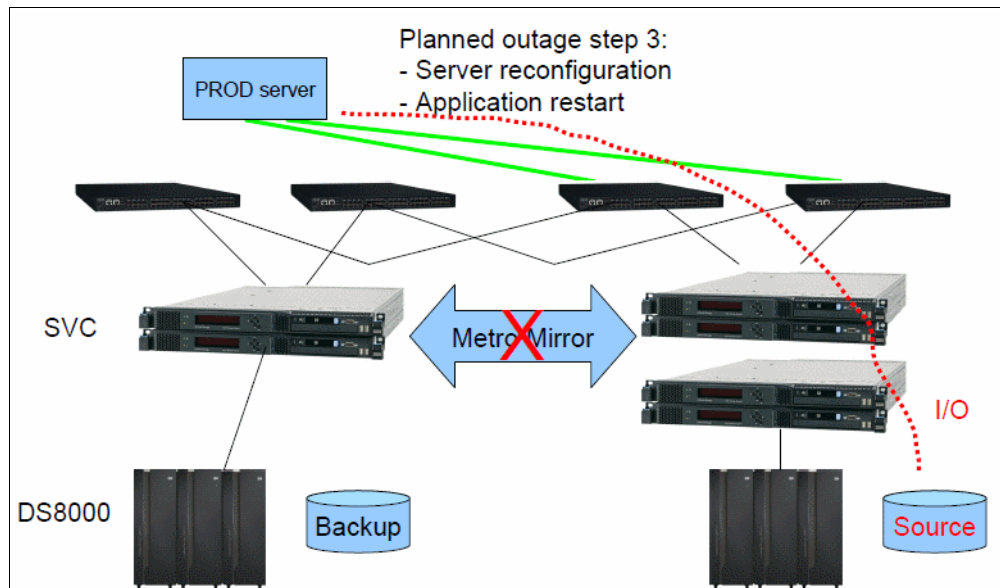


Figure 16-20 Server reconfiguration and application restart

After some time for testing, the remote copy session is removed and the move to the new environment is completed, as shown in Figure 16-21.

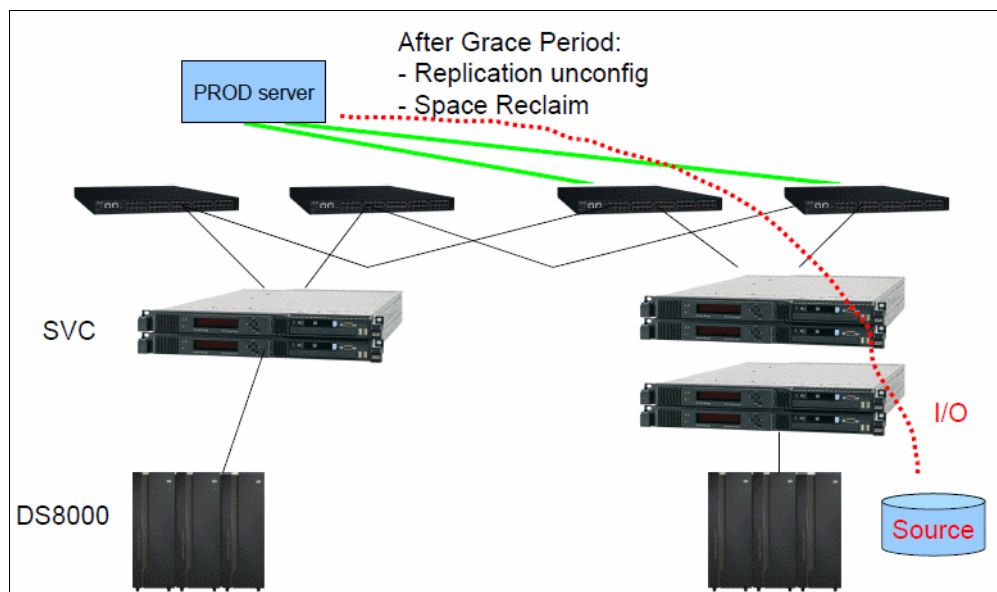


Figure 16-21 Removing remote copy relationships and reclaiming old space (backup copy)

16.5 SAN Volume Controller scripting

Although the SAN Volume Controller Console GUI is a tool (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 script the SAN Volume Controller CLI. The SAN Volume Controller CLI can be scripted by using any program that can pass text commands to the SAN Volume Controller cluster Secure Shell (SSH) connection.

On UNIX systems, you can use the `ssh` command to create an SSH connection with the SAN Volume Controller. On Windows systems, you can use the `plink.exe` utility (which is provided with the PuTTY tool) to create an SSH connection with the SAN Volume Controller. The examples in the following sections use the `plink.exe` utility to create the SSH connection to the SAN Volume Controller.

16.5.1 Connecting to SAN Volume Controller by using predefined SSH 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 your SAN Volume Controller administrator user name (for example, `admin`). To set this parameter, click **Connection** $\emptyset$ **Data** in the left pane of the PuTTY Configuration window, as shown in Figure 16-22 on page 576.

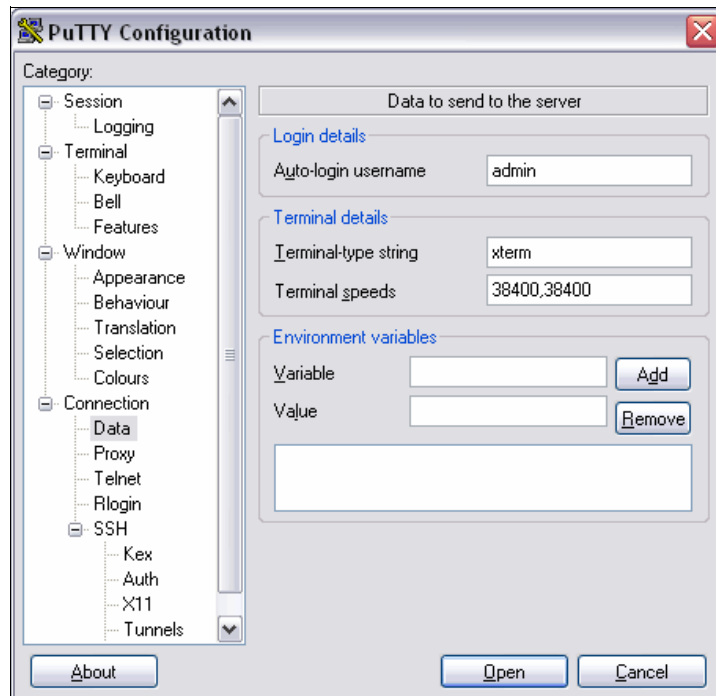


Figure 16-22 Configuring the auto-login user name

- 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 16-23.

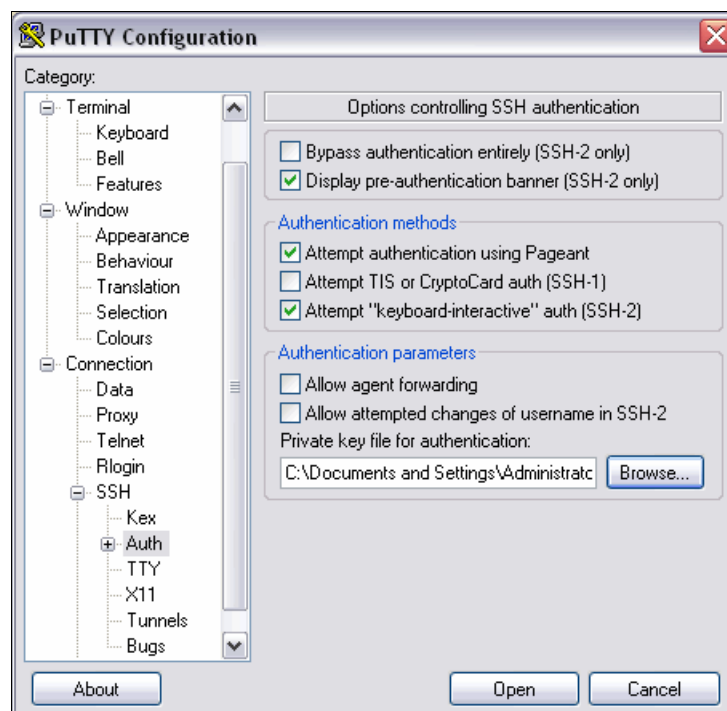


Figure 16-23 Configuring the SSH private key

- The IP address of the SAN Volume Controller cluster. To set this parameter, select **Session** at the top of the left pane of the PuTTY Configuration window, as shown in Figure 16-24.

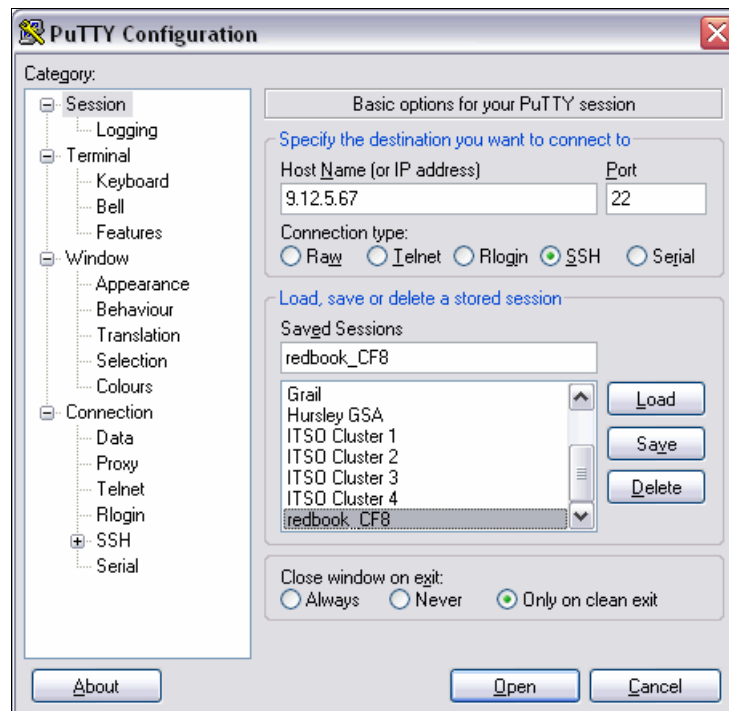


Figure 16-24 Specifying the IP address

When you are specifying the basic options for your PuTTY session, you need the following information:

- A session name, which in this example is `redbook_CF8`.
- The PuTTY version, which is 0.61.

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"
```

Example 16-8 show a script to restart Global Mirror relationships and groups.

Example 16-8 Restarting Global Mirror relationships and groups

```
svcinfolsrconsistgrp -filtervalue state=consistent_stopped -nohdr -delim : | while IFS=: read
id name mci mcn aci acn p state junk; do
    echo "Restarting group: $name ($id)"
    svctask startlrconsistgrp -force $name
    echo "Clearing errors..."
    svcinfo lserrlogbyrconsistgrp -unfixed $name | while read id type fixed snmp err_type node
seq_num junk; do
        if [ "$id" != "id" ]; then
            echo "Marking $seq_num as fixed"
            svctask cherrstate -sequencenumber $seq_num
        fi
    done
done
svcinfolsrrelationship -filtervalue state=consistent_stopped -nohdr -delim : | while IFS=:
read id name mci mcn mvi mvn aci acn avi avn p cg_id cg_name state junk; do
    if [ "$cg_id" == "" ]; then
        echo "Restarting relationship: $name ($id)"
        svctask startlrrelationship -force $name
        echo "Clearing errors..."
        svcinfo lserrlogbyrrelationship -unfixed $name | while read id type fixed snmp
err_type node seq_num junk; do
            if [ "$id" != "id" ]; then
                echo "Marking $seq_num as fixed"
                svctask cherrstate -sequencenumber $seq_num
            fi
        done
    fi
done
```

You can run various limited scripts directly in the SAN Volume Controller shell, as shown in Example 16-9, Example 16-10, and Example 16-11.

Example 16-9 shows a script to create 50 volumes.

Example 16-9 Creating 50 volumes

```
IBM_2145:svccf8:admin>for ((num=0;num<50;num++)); do svctask mkvdisk -mdiskgrp 2
-size 20 -unit gb -iogrp 0 -vtype striped -name Test_$num; echo Volumenname
Test_$num created; done
```

Example 16-10 shows a script to change the name for the 50 volumes that were created.

Example 16-10 Changing the name of the 50 volumes

```
IBM_2145:svccf8:admin>for ((num=0;num<50;num++)); do svctask chvdisk -name
ITS0_$num $num; done
```

Example 16-11 shows a script to remove the 50 volumes that were created.

Example 16-11 Removing all the created volumes

```
IBM_2145:svccf8:admin>for ((num=0;num<50;num++)); do svctask rmvdisk $num; done
```

16.5.2 Scripting toolkit

IBM engineers developed a scripting toolkit that helps to automate SAN Volume Controller 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>

The scripting toolkit includes a sample script that you can use to redistribute extents across existing MDisks in the pool. For more information about the redistribute extents script from the scripting toolkit, see 5.7, “Rebalancing extents across a storage pool” on page 107.

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.

16.6 Migrating AIX cluster volumes off DS4700

In this scenario, a two node AIX PowerHA cluster is migrated from a DS4700 to Storwize V7000. The nodes of the cluster use dedicated I/O cards. Such a migration can be combined with an operating system upgrade or PowerHA upgrade. It does not affect the steps that must be performed at the storage configuration level.

This scenario requires a relatively short downtime that is required to import image mode volumes that are presented to AIX systems by DS4700 storage and to present them to AIX hosts.

The migration is conceptually divided into the following phases:

1. Preparation.
2. Detachment of DS4700 from AIX systems and importing of its volumes in image mode on Storwize V7000.
3. Presentation of virtualized volumes to AIX systems and migration of storage.
4. Final configuration.

The schematic representation of storage configuration during consecutive phases of the migration is shown in Figure 16-25 on page 580. Two AIX systems (aix01 and aix02) are shown in Figure 16-25 on page 580: DS4700 storage system and a Storwize V7000 storage system. The lines that are connecting them show the SAN zoning that is required at a stage and the path for the data that is transmitted between the hosts and the storage systems.

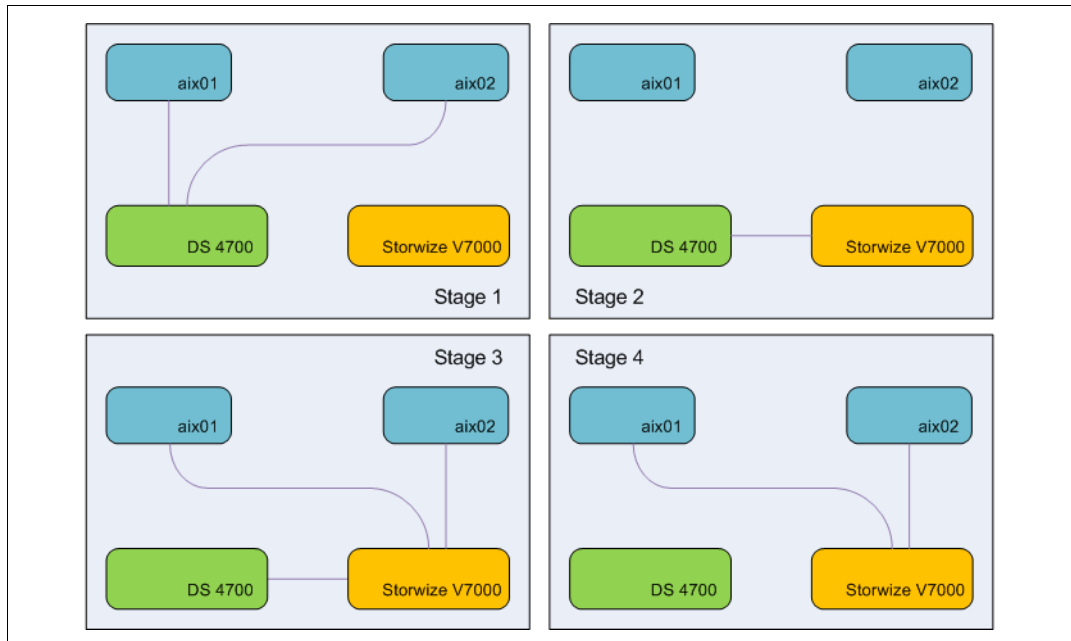


Figure 16-25 Four phases of migration

16.6.1 Preparation

In the initial configuration both AIX cluster nodes (aix01 and aix02), use only storage that is exported from DS4700 volumes. In this phase, you must document the environment to re-create it in the final configuration. Because some applications might be sensitive to device names and LUN mappings, this scenario preserves both device names and LUN IDs after the migration. This process reduces the possibility that no issues emerge after the storage is migrated from DS4700 system to Storwize V7000.

During this phase of the migration, you should gather and record the following information about each volume that is presented by the DS4700 system to each AIX host:

- ▶ HDisk number (for example, hdisk2)
- ▶ PVID of the hdisk
- ▶ Volume group to which this hdisk belongs (if any)
- ▶ Array name on DS 4700 on which this volume is defined
- ▶ Logical drive name on DS 4700
- ▶ LUN ID on node aix01
- ▶ LUN ID on node aix02
- ▶ Volume size
- ▶ Comments (for example, “this is a heartbeat hdisk”)

This information makes it easier for you to successfully re-create the configuration after the storage is migrated. A configuration might be asymmetrical; for example, a specific volume on DS4700 storage might appear as hdiskX on one host and as hdiskY on the second host.

Also, you should record the following information for the purposes of zoning definitions in the SAN environment:

- ▶ WWPNs of HBAs that are used on the AIX hosts and their aliases in the SAN configuration
- ▶ Name of zones that contain AIX hosts and DS4700 controller ports
- ▶ WWPNs of DS 4700 controller ports
- ▶ WWPNs of Storwize V7000 controller ports

Complete the following steps to define an empty storage pool on Storwize V7000 into which you import the image mode volumes:

1. In the Storwize V7000 GUI, click **Pools** $\varnothing$ **MDisks by Pools** and then click **New Pool**.
2. Enter the wanted storage pool name (for example, `aix_cluster_img`).
3. Click **Next** then, click **Create**.
4. Click **OK** to confirm that you want to create an empty storage pool.

If you plan to use an existing storage pool to host the volumes for the AIX cluster, make sure that it has sufficient free space. Alternatively, define a new storage pool into which the volumes are migrated.

If you do not have unused licensing capacity for external systems, in the Storwize V7000 GUI, click **Settings** $\varnothing$ **General** and then click **Licensing**. In the External Virtualization field, increase the value by 1. With this setting, you can attach the DS4700 storage to the Storwize V7000 system. Since version 6.2 of SAN Volume Controller code, you can exceed the licensed virtualization entitlement for 45 days from the installation date to migrate data into the new Storwize V7000 system.

On the AIX hosts, install SDDPCM drivers and the host attachment kit on the AIX servers. They are needed for the operating system to communicate with Storwize V7000 storage system.

Additionally, verify that your environment meets all the requirements, including versions of HBA firmware, FC switch firmware, OS, and driver levels. Update components as required to run a supported combination.

Before you proceed, ensure that you have the current and verified backups of your environment. Whenever you change your storage configuration, it is always a good practice to have a current backup that you know you can restore in case of emergency. You can also make a copy of the rootvg volume group by running the **alt_disk_copy** command, which provides you with a quick recovery path if there are unexpected problems.

To migrate volumes from DS 4700 to Storwize V7000, you must zone these systems. Follow the guidelines concerning external storage configuration requirements that are available at this website:

http://pic.dhe.ibm.com/infocenter/storwize/ic/index.jsp?topic=%2Fcom.ibm.storwize.v7000.doc%2Fsvc_configdiskcontrollersovr_22n9uf.html

For more information about zoning requirements, see this website:

http://pic.dhe.ibm.com/infocenter/storwize/ic/index.jsp?topic=%2Fcom.ibm.storwize.v7000.doc%2Fsvc_configrules_21rdwo.html

In the DS 4700 Storage Manager GUI, define a new storage partition and set the host type to IBM TS SAN VCE. Add Storwize V7000 controller WWPNs to the definition of the storage partition.

In the Storwize V7000 GUI, click **Hosts** $\varnothing$ **Hosts** $\varnothing$ **New Host**. Define AIX hosts by using HBA WWPNs that you gathered in 16.6.1, “Preparation” on page 580.

After DS 4700 and Storwize V7000 are zoned, open the Storwize V7000 GUI and click **Pools** $\varnothing$ **External Storage** $\varnothing$ **Detect MDisks**. You should see a new storage controller detected. At this time, it does not present any MDisks to the Storwize V7000.

All the tasks up to this point do not require any downtime and can be done in a working environment.

16.6.2 Importing image mode volumes

Complete the following steps to import image mode volumes:

1. On the AIX hosts, stop the cluster and bring down the resource groups.
2. Remove all of the disks that are migrated by using **rmdev -d1** command. If you do not perform this step, you have two sets of hdisk devices: old (which correspond to the volumes that are presented by the DS4700) in defined state), and the new set corresponding to the volumes that are presented to the system by the Storwize V7000 after you present the volumes from the Storwize V7000 system. This situation requires extra work to be done to achieve the goal of the new hdisk names for presented volumes that match the original names.
3. Shut down AIX systems. After you verify that the hosts are stopped, perform the following steps for each volume that is presented from DS4700 to AIX host:
 - a. In the DS 4700 Storage Manager GUI, note the LUN ID of the volume and change the mapping of the volume from the storage partition that contains AIX hosts to the storage partition you created for Storwize V7000. To make it easier for you to manage the change, you can preserve the LUN ID (that is, present the volume to Storwize V7000 by using the same LUN ID) because you used to present the volume to the AIX host.
 - b. In the Storwize V7000 GUI, click **Pools**  **External Storage** and then click **Detect MDisks**. You should see a new MDisk presented by the external storage controller. Right-click the volume and import it in image mode.

Note: If you import the MDisk as a managed disk, the data on the volume is lost. To make it easier for you to manage the change, you can use the same name for the imported volume as the volume name that is defined in DS4700 storage system. Alternatively, you can use this opportunity to introduce a common volume naming scheme.

Also, the access LUN that is presented by the DS4700 must not be mapped to the Storwize V7000 system.

After you complete these steps for all volumes that are presented by DS 4700 to the AIX hosts, you can proceed to the next stage of the migration.

16.6.3 Data migration

Complete the following steps to migrate data:

1. Remove the zoning between AIX hosts and DS 4700 storage system.
2. Create zoning between AIX hosts and Storwize V7000 system.
3. Power on your AIX hosts. They are accessing the data on the DS4700 through Storwize V7000, but we want to minimize the downtime of the systems. Complete the following steps:
 - a. In the Storwize V7000 GUI, click **Pools**  **Volumes by Pool**.
 - b. Right-click each of the volumes in the `aix_cluster_img` pool and map a single volume to the AIX host.

- c. After you present the volume to the AIX host, run **cfgmgr** command on the host. The operating system creates a single hdisk device at a time, which gives you full control over the mapping of volumes to hdisks. Perform this step for each volume you need to present to the host in the order that gives the wanted hdisk names. Make sure you use LUN IDs that you gathered in 16.6.1, “Preparation” on page 580 when volumes are presented to the hosts.

The data migration is done by creating a mirror copy of the imported image mode volume in the target storage pool. By using this approach, you can control the speed of data replication and therefore, prevent overloading the DS4700 storage system, which must handle more I/Os that are generated by the migration process.

After the volume mirroring process is complete, you delete the copy that is on the image mode volume that was imported from the DS4700 storage system.

Complete the following steps to create a mirrored copy of the volume:

1. In the Storwize V7000 GUI, click **Volumes** $\varnothing$ **Volumes** and then right-click an image mode volume that was imported from the DS4700 storage.
2. From the menu, choose **Volume Copy Actions** and then **Add Mirrored Copy**.

A window opens in which you can choose the destination storage pool. Choose the target pool and click **Add Copy**.

3. In the Running Tasks window, a new task appears that is called Volume Synchronization. If you click this task, you see the name of the volume that is copied, the name of the copy, and the estimated time until the replication operation completes. The name of the copy (for example, copy 1).

Repeat these steps for all image mode volumes that were imported from the DS 4700 system.

When the task completes, complete the following steps:

1. In the Storwize V7000 GUI, click **Pools** $\varnothing$ **Volumes by Pool**.
2. Right-click the newly created copy of the volume (for example, copy 1) and click **Make Primary**.
3. Right-click the copy of the volume that is on the image mode volume that was imported from the DS4700 and click **Delete this Copy**. Confirm your choice by clicking **Yes**.

The migration process is complete.

16.6.4 Final configuration

The last step is to unconfigure and remove the DS4700 system. After you deleted copies of volumes that were on the image mode volumes that were imported from the DS4700 system, these volumes are put in unmanaged mode. Verify that all volumes that are presented by DS4700 to Storwize V7000 are in unmanaged mode.

Complete the following steps:

1. In the DS4700 Storage Manager GUI, unmap the volumes that are mapped to the Storwize V7000 system.
2. In the Storwize V7000 GUI, click **Pools** $\varnothing$ **External Storage** and then click **Detect MDisks**. You should see a storage controller, but no MDisks are presented to the Storwize V7000 system.

You can now remove the zoning between the DS4700 and Storwize V7000 systems.

In the Storwize V7000 GUI, click **Pools** $\varnothing$ **External Storage** and click **Detect MDisks**. There should be no external controllers visible. The migration is completed.

If you increased the number of licensed External Virtualization units (as described in section 16.6.1, “Preparation” on page 580), remember to return it to the original value.

16.7 Easy Tier and FlashSystem planned outages

This section describes how to configure Easy Tier for use when an outage is required for an IBM FlashSystem.

If an outage is required on a FlashSystem storage system that it is being used as Easy Tier capacity for a storage pool, the extents that are on it must be relocated in advance of the outage. Complete the following steps:

1. Ensure that the hard disk drives in the storage pool contain enough free capacity to allow the FlashSystem extents to be migrated off the FlashSystem MDisks and on to the HDD MDisks.
2. If Easy Tier is set to auto for the storage pool, change the setting to on. By setting Easy Tier to on for the storage pool during this process, the heat maps are retained and the duration of the overall process is minimized.
3. Remove the FlashSystem MDisks from the storage pool to force the extents to be migrated to the HDD MDisks.
4. Ensure that the extent migration completes and that the FlashSystem MDisks are now “unmanaged” in the SAN Volume Controller. This step can take up to 48 hours to complete because the data must be migrated from the SSDs to the HDDs and the migration is performed at a controlled pace to avoid overloading the storage. The actual duration of this step depends, in part, on the size of the FlashSystem MDisks that were removed.
5. Perform the required FlashSystem maintenance.
6. Add the FlashSystem MDisks back into the storage pool. Within 24 hours, Easy Tier begins migrating hot extents to the FlashSystem MDisks. This step can take up to 48 hours to complete because Easy Tier works on a 24-hour sliding window and can move up to 2 TB per hour. The actual duration of this step depends, in part, on the size of the FlashSystem MDisks that are added to the pool.

Retaining heat data: In step 2 of this procedure, Easy Tier loses all knowledge about what data is hot if the setting is left on auto. This information must be relearned when the SSDs are added again. By changing the setting to on, this heat data is retained, which allows Easy Tier to make more intelligent and informed decisions when it starts to move data back on to the FlashSystem MDisks that was described in step 6.

Important: Verify that Easy Tier extents were migrated before you shut down and perform maintenance. Manually verify that all Easy Tier extents were placed back in their original location after the maintenance process. If there are other environmental workload changes, the extents might be placed back in the same location.

16.8 Changing LUN ID presented to a VMware ESXi host

VMware recommends that LUN IDs of volumes that are presented to VMware ESXi hosts in a cluster are consistent; that is, that a specific volume has the same LUN ID on all VMware ESXi hosts to which it is mapped. Consistent LUN IDs are also a requirement for VMware Site Recovery Manager (SRM). To bring the environment into conformance with requirements of VMware SRM, the LUN IDs of some of the volumes that are presented to a specific host of hosts might need to be changed. However, the installation of VMware SRM enables the datastore resignaturing option (`LVM.enableResignature`) on VMware ESXi hosts. Therefore, a proper procedure must be followed to avoid accidental resignature of the datastore or entering a resignature loop in the VMware cluster.

This section describes how to change the LUN ID of a volume that is presented to a VMware ESXi 5.x host, which is a part of VMware Infrastructure cluster. To change the LUN ID of the volume, the volume must be unrepresented from the host and then presented again with the correct LUN ID.

In this scenario, there are two VMware ESXi hosts: `esxi01` and `esxi02`. Two volumes are presented to these hosts: `esxi_datastore_0` and `esxi_datastore_1`. The `esxi_datastore_0` volume is consistently mapped at LUN ID 1 to both hosts, while `esxi_datastore_1` has LUN ID 3 on host `esxi01` and LUN ID 2 on host `esxi02`. The goal is for it to have LUN ID 2 on both hosts.

Figure 16-26 shows the inconsistent mapping of the `esxi_datastore_1` volume. This window was displayed in the Storwize family storage device GUI, in the Volumes ∅ Volumes by Pool view after right-clicking the **esxi_datastore_1** and clicking **View Mapped Hosts**.

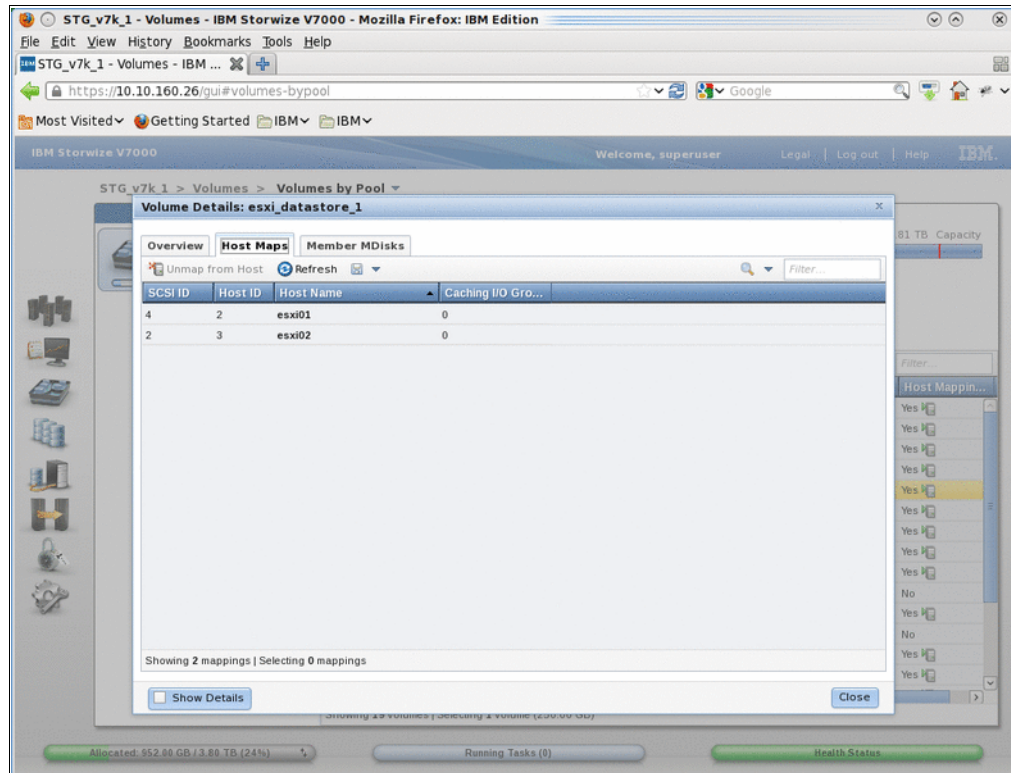


Figure 16-26 Inconsistent volume mapping

The volume must be unrepresented from the VMware ESXi hosts before its LUN ID can be changed. The following prerequisites that must be met before a LUN can be unrepresented from a VMware ESXi host:

- ▶ All objects (virtual machines, snapshots, templates, CD/DVD images, and so on) must be unregistered or removed from the datastore.
- ▶ The datastore cannot be used for vSphere HA heartbeat.
- ▶ The datastore cannot be a part of a datastore cluster.
- ▶ The datastore cannot be managed by Storage DRS.
- ▶ No process that is running on the ESXi host can access the LUN.
- ▶ The LUN cannot be used as the persistent scratch location for the host.
- ▶ If the LUN is used as RDM storage by a virtual machine, delete this LUN from the configuration of the virtual machine. This action removes the mapping of the LUN to the virtual machine but preserves the contents of the LUN.

Note: You must coordinate the change of the virtual machine configuration with the administrator of the operating system of the virtual machine.

To unrepresent a datastore from a host, no virtual machines can be registered on this datastore. However, it is important to understand this requirement because it means that no virtual machine that is running on the host from which the datastore is unrepresented is using this datastore. In our scenario, server_00 and server_01 have their virtual disks on datastore_1, which must be unrepresented from host esxi01 to have its LUN ID changed. However, server_01 is running on host esxi02, so no change in configuration is needed. Because server_00 is running on host esxi01, you must move it to another host by using VMware vMotion. This operation results in no downtime and server_00 can be moved back after the LUN ID change operation is completed.

Complete the following steps to check whether and what objects are on the datastore datastore_1 and to unmount the datastore:

1. In the vSphere Client, switch the view by clicking **Home** ∅ **Inventory** ∅ **Datastores and Datastore Clusters**. As shown in Figure 16-27 on page 587, there are two servers present on the datastore.

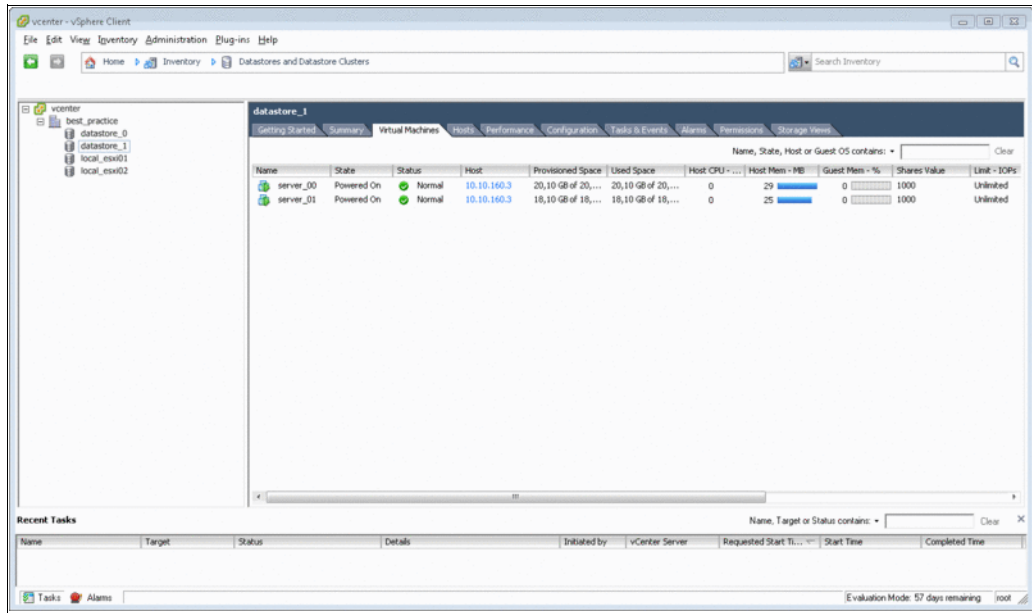


Figure 16-27 Objects present on datastore_1

2. Use VMware vMotion to migrate server_00 to host esxi02. Right-click the server_00 entry and choose **Migrate** (see Figure 16-28).

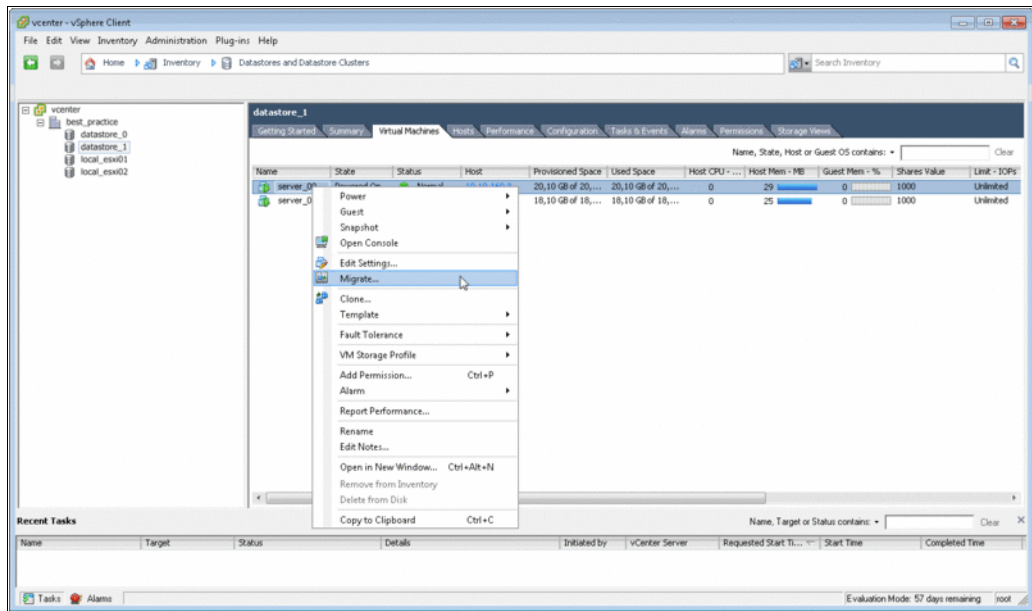


Figure 16-28 Migrating server_00 to host esxi02

3. Set host esxi02 as the destination host for the virtual machine. After the migration completes, the environment is ready for other operations.
4. In the vSphere Client, switch to the Configuration tab of the esxi01 host and click **Storage Adapters** entry.

In the lower right pane, you can see that LUN ID 3 is used to access a device with the identifier `naa.60050768028a02b9300000000000001c`. This is the datastore with the incorrectly assigned LUN ID (see Figure 16-29). This information is available if the Devices option is clicked, as indicated by the arrow in Figure 16-29.

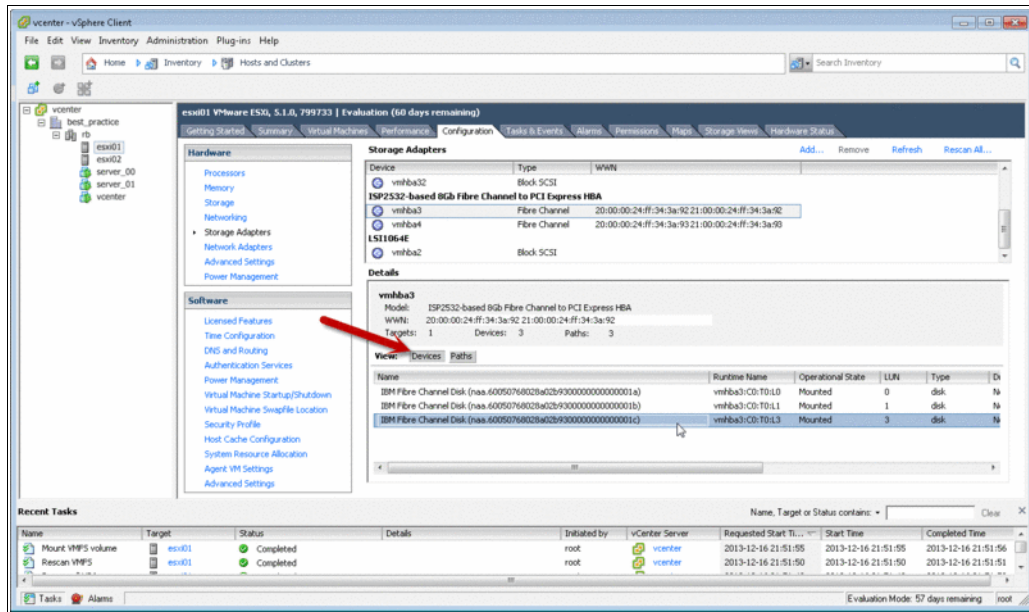


Figure 16-29 LUN ID of 3 assigned to the device with id `naa.60050768028a02b9300000000000001c`

5. Switch to the Storage view, right-click the **datastore_1** entry, and click **Unmount**, as shown in Figure 16-30.

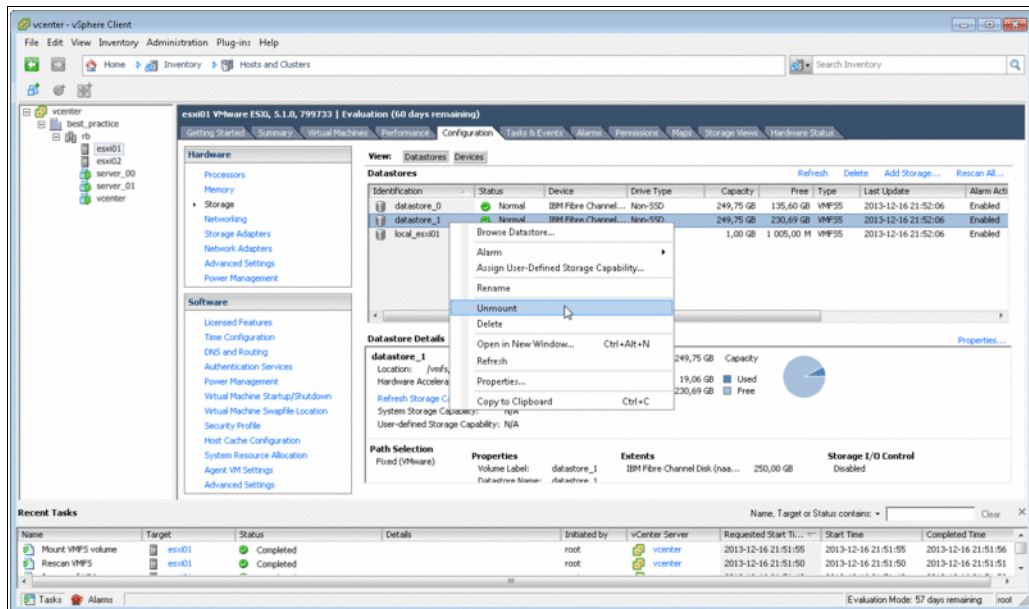


Figure 16-30 Unmounting the datastore

6. A window opens in which the unmount prerequisites are listed. If all of the prerequisites are met, click **OK** to confirm unmounting the datastore.

- Switch to the Storage view and in the lower right pane, click **Devices**. In this view, right-click the device identifier and click **Detach**, as shown Figure 16-31. A window opens in which the prerequisites are listed. If all the prerequisites are met, click **OK** to confirm detaching the device.

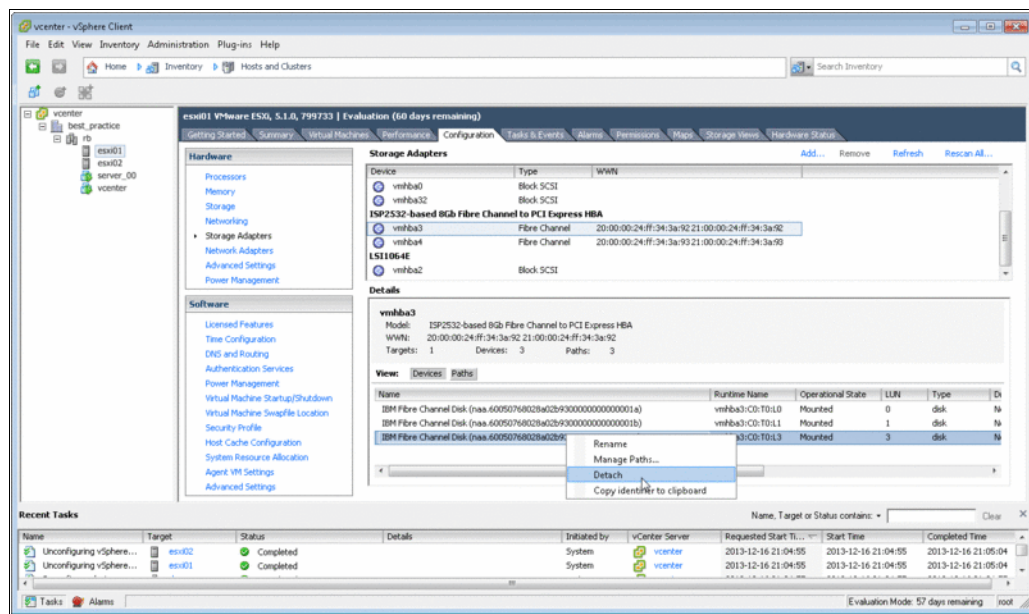


Figure 16-31 Detaching the device corresponding to the datastore

Complete the following steps to remove the mapping of the LUN on the Storwize family storage device:

- In the Storwize family storage device GUI, click **Volumes** $\emptyset$ **Volumes by Pool** and right-click **Map to Host**. Choose **esxi01** from the pull-down menu.
- In the right pane, select **esxi_datastore_1** and click **Unmap**, as shown in Figure 16-32 on page 590.

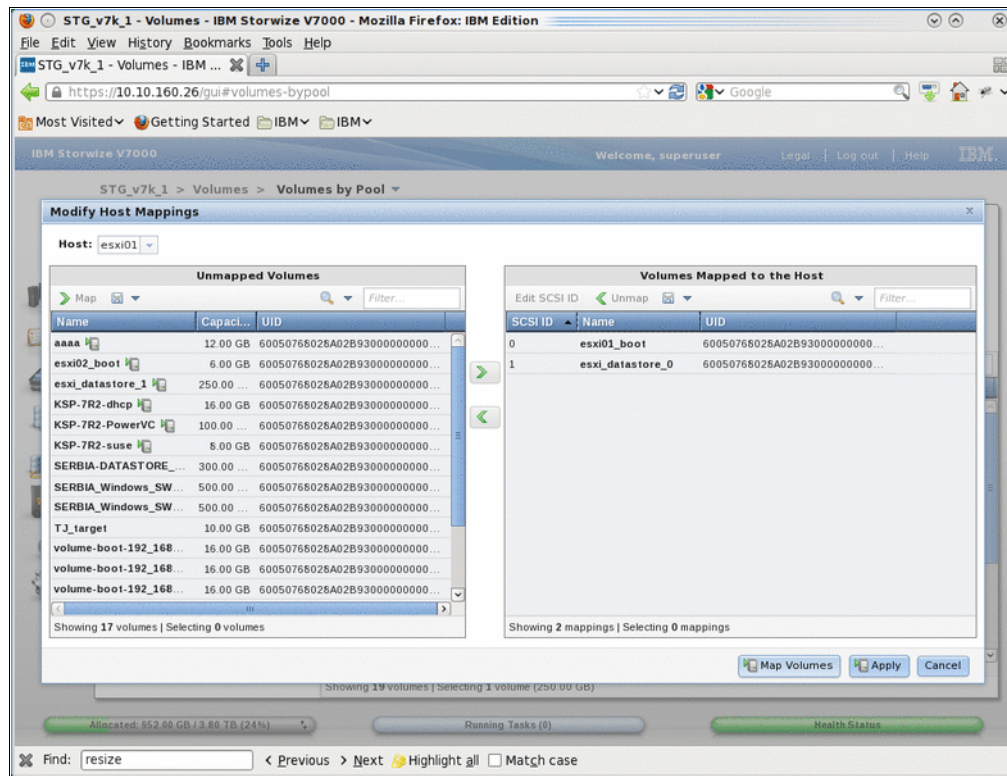


Figure 16-32 Volume unmapped from the esxi01 host

3. Click **Map Volumes**. The volume is unrepresented from the esxi01 host.
4. In the vSphere Client, switch to the Configuration tab of the ESXi host, click **Storage Adapters** and run the Rescan task on the storage adapters. The device is removed from the Storage Adapters view, as shown in Figure 16-33.

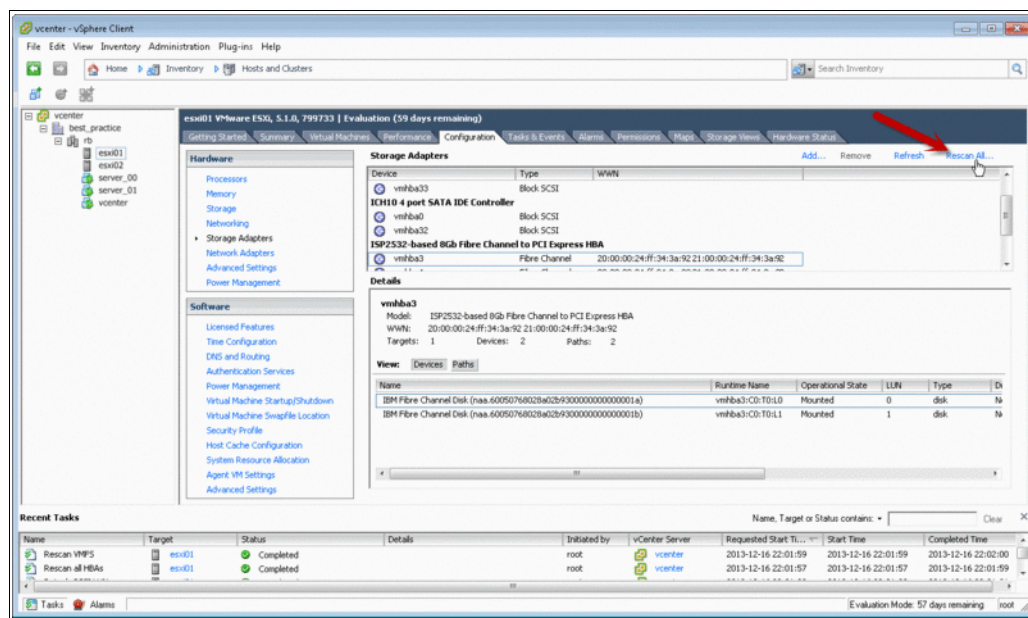


Figure 16-33 Rescan and removal of the unmapped volume

Complete the following steps to map the volume to the ESXi host with the correct LUN ID:

1. In the Storwize family storage device GUI, click **Volumes** $\otimes$ **Volumes by Pool** view and right-click **Map to Host**.
2. Choose **esxi01** from the pull-down menu. The volume is mapped to the host. Make sure that the correct LUN ID (ID 2) is set in the right pane (as shown in Figure 16-34) and click **Map Volumes**.

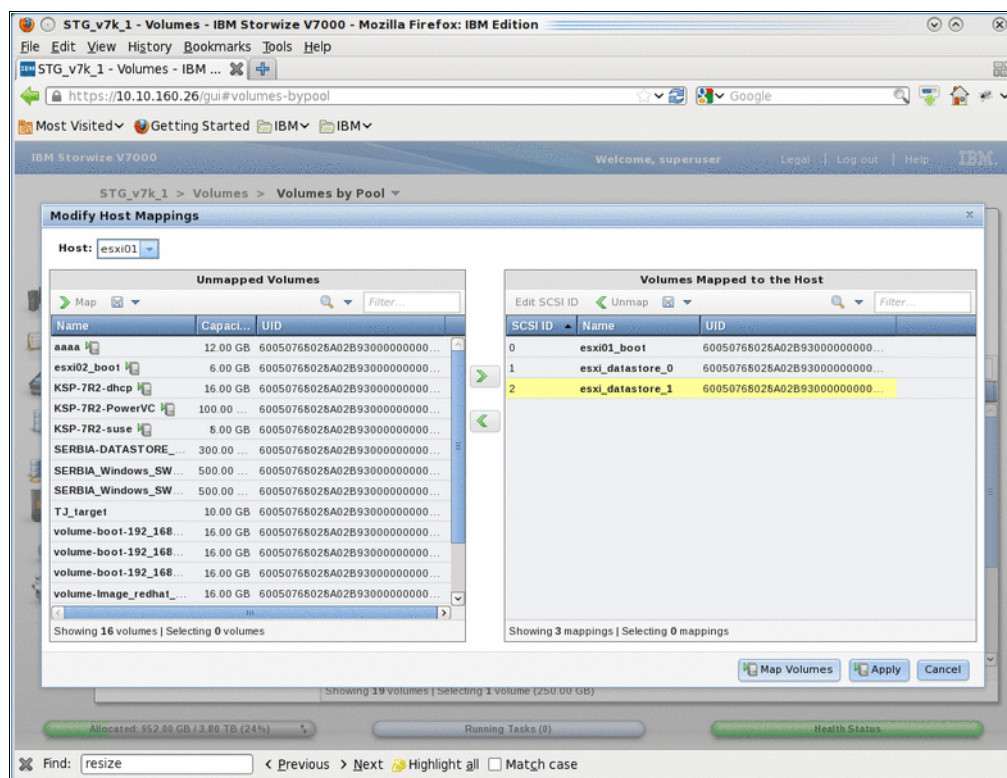


Figure 16-34 Correct mapping of the esxi_datastore_1 datastore

3. In the vSphere Client, switch to the Configuration tab of the esxi01 host, click **Storage Adapters** and run the Rescan task on the storage adapters. The device is added to the Storage Adapters view.

Because the detached state is persistent, you must right-click the device and click **Attach**. Similarly, the unmounted state of the datastore is persistent. You must switch to the Configuration tab of the esxi01 host, click **Storage** and then click **Rescan All** to detect the datastore.

4. Right-click the datastore that is added again and then click **Mount**, as shown in Figure 16-35.

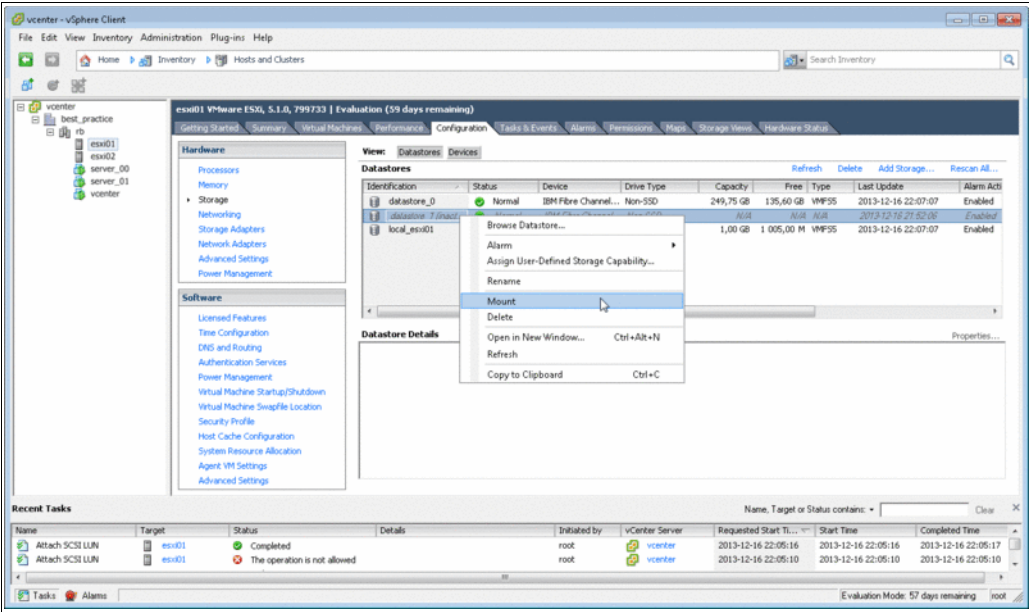


Figure 16-35 Remounting of the datastore after correcting the LUN ID

The procedure is completed. The LUN IDs of volumes that are presented to the hosts are consistent across all VMware ESXi hosts in the cluster.



IBM Real-time Compression

This chapter highlights the preferred practices for using IBM Real-time Compression with IBM System Storage SAN Volume Controller and Storwize V7000 systems. 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 this innovative IBM Real-time Compression technology offers.

This chapter includes the following sections:

- ▶ Overview
- ▶ What is new in version 7.2
- ▶ Evaluate data types for estimated compression savings by using the Comprestimator utility
- ▶ Evaluate workload by using Disk Magic sizing tool
- ▶ Configure a balanced system
- ▶ Verify available CPU resources
- ▶ Compressed and non-compressed volumes in the same MDisk group
- ▶ Application benchmark results
- ▶ Standard benchmark tools
- ▶ Compression with FlashCopy
- ▶ Compression with Easy Tier
- ▶ Compression on SAN Volume Controller with Storwize V700
- ▶ Related Publications

17.1 Overview

With the current trend of data growth in the IT industry and ongoing economic turmoil, there is an immediate need for technologies that optimize and reduce the amount of data that is written to disk storage and reduce the costs. Most available traditional methods of data optimization technologies involve a post-compression mechanism, which means that the optimization is done on data sets that are stored on the disks and are not effective in nature.

Contrary to conventional methods, IBM Real-time Compression is an inline data compression technology that performs real-time compression of active primary data *before* it is written to the disk storage without affecting performance. IBM Real-time Compression technology is embedded into IBM System Storage SAN Volume Controller and IBM Storwize V7000 Software stack, starting with SAN Volume Controller version 6.4 and is based on the proven Random-Access Compression Engine (RACE).

This chapter outlines the preferred practices to follow when IBM Real-time Compression is used with IBM System Storage SAN Volume Controller and Storwize V7000 systems, which enables customers to enjoy the compression savings IBM Real-time Compression technology offers.

There are many IBM Redbooks publications about IBM Real-time Compression in SAN Volume Controller and IBM Storwize V7000, including the following publications:

- ▶ *Real-time Compression in SAN Volume Controller and Storwize V7000*, REDP-4859
- ▶ *Implementing IBM Real-time Compression in SAN Volume Controller and IBM Storwize V7000*, TIPS1083

These books cover many aspects of implementing compression. This chapter complements those publications and provides details to reflect the compression enhancements in SAN Volume Controller version 7.2.

17.2 What is new in version 7.2

Version 7.2 includes the following highlights:

- ▶ New license structure for Real-time Compression on Storwize V7000. New licensing is based on per controller and the maximum number of IBM Real-time Compression licenses per control enclosure is capped at three. This means customers with four or more enclosures must purchase only a maximum of three licenses for that system. This includes externally attached enclosures when external virtualization is used. This represents significant savings with systems that contain more than three enclosures overall.
- ▶ IBM Real-time Compression Software in SAN Volume Controller version 7.2 has software enhancements that are based on the advanced LZ4 algorithm. This configuration delivers the following significant performance enhancements and makes much more efficient use of the system CPU resources:
 - Up to 3x higher sequential write throughput (VMware vMotion)
 - 35% higher throughput (IOPS) in OLTP workload (TPC-C)
 - 35% lower compression CPU usage for the same workload
- ▶ New threshold for volume compressibility is lowered to 40% from the previous figure of 45%. For more information, see 17.3, “Evaluate data types for estimated compression savings by using the Comprestimator utility” on page 595.

17.3 Evaluate data types for estimated compression savings by using the Comprestimator utility

Before you use IBM Real-time Compression Software, 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.

The best candidates for data often are the data types that are not compressed by nature, such as databases, server virtualization, CAD/CAM, software development systems, and vector data. To determine the compression savings you are likely to achieve for the workload type, IBM developed an easy-to-use utility called IBM Comprestimator. IBM Comprestimator is a host-based utility that can be used to estimate expected compression rate for block devices.

The Comprestimator 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 that is based on the formulas it uses. The utility runs on a host that has access to the devices that are analyzed and performs only read operations so it has no effect on the data that is stored on the device.

IBM recommends that Comprestimator is used. The latest version of the Comprestimator utility is available at this website:

<http://www-304.ibm.com/webapp/set2/sas/f/comprestimator/home.html>

The Comprestimator Quick Start Guide is part of the download package. After you download the utility package, you see detailed instructions in the `readme.txt` file about how to run the utility with example syntax, explanation of scan results, release notes information, common issues, and troubleshooting steps.

The latest comprestimator (version 1.4 at the time of this writing) adds support for analyzing expected compression rates in accordance with Storwize V7000/SAN Volume Controller/Flex storage systems software version 7.2.

For a complete list of supported client operating system versions to run Comprestimator, see the Comprestimator Quick Start Guide.

In summary, remember the following preferred practices regarding the use of Comprestimator:

- ▶ Run the Comprestimator utility before implementing the compression solution.
- ▶ Download the latest version of the utility from IBM.
- ▶ Use Comprestimator to analyze volumes that contain as much active data as possible rather than volumes that are mostly empty of data. 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 17-1 thresholds for volume compressibility to determine whether to compress a volume.

Table 17-1 To compress or not

Data Compression Rate	Recommendation
Higher than 40% compression savings	Use compression
Less than 40% compression savings	Evaluate workload with compression

17.4 Evaluate workload by using Disk Magic sizing tool

Proper initial sizing greatly helps to avoid future sizing problems. Disk Magic is one such tool that is used for sizing and modeling storage subsystems for various open systems environments and various IBM platforms. It provides accurate performance and capacity analysis and planning for IBM SAN Volume Controller, IBM Storwize V7000, IBM, 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 with IBM Real-time Compression.

Beginning with Disk Magic V9.9.0, support for compression in Storwize V7000 and SAN Volume Controller is added. At the time of this writing, the latest Disk Magic version is V9.10.5, which also includes support for IBM Real-time Compression model V7.2 that can be downloaded from this website:

<http://www.ibm.com/partnerworld/wps/servlet/ContentHandler/SSPQ048068H83479I86>

17.5 Configure a balanced system

In a system with more than one IO group, it is not a good practice to configure just a few compressed volumes in an IO group. Consider a four-node Storwize V7000 (two IO groups) with 200 volumes and the following configuration:

- iogrp0: nodes 1 and 2 with 180 compressed volumes
- iogrp1: nodes 3 and 4 with 20 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 20 volumes out of a total of 200 compressed volumes. The following preferred practices in this scenario should be used:

- ▶ Alternative 1: Migrate all compressed volumes from iogrp1 to iogrp0
- ▶ Alternative 2: Migrate compressed volumes from iogrp0 to iogrp1 and load balance across nodes. Table 17-2 shows the load distribution.

Table 17-2 Load distribution

	node1	node2	node3	node4
Original setup	90 compressed volumes X non-compressed volumes	90 compressed volumes X non-compressed volumes	10 compressed volumes X non-compressed volumes	10 compressed volumes X non-compressed volumes
Alternative 1	100 compressed volumes X non-compressed volumes	100 compressed volumes X non-compressed volumes	X non-compressed volumes	X non-compressed volumes
Alternative 2	50 compressed volumes X non-compressed volumes	50 compressed volumes X non-compressed volumes	50 compressed volumes X non-compressed volumes	50 compressed volumes X non-compressed volumes

17.6 Verify available CPU resources

Before compression is enabled on Storwize V7000 and SAN Volume Controller 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.

Note: IBM recommends to customers who are planning to use IBM Real-time Compression on 6-core SAN Volume Controller CG8 systems 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, this configuration eliminates the need to measure CPU usage before compression is enabled on the system. Information about upgrading to SAN Volume Controller CG8 dual CPU model is available with RPQ #8S1296.

- ▶ Consider the following points regarding Storwize V7000 systems, SAN Volume Controller CF8 nodes, and early SAN Volume Controller CG8 nodes (4-core systems):
 - If the CPU utilization of a node is persistently above 25% most of the time, this I/O group might not be suitable for compression because it is too busy.
 - If your system is supported for the Non-Disruptive Volume Move feature, this volume can be moved to another I/O group that has the resources that are required for compression.

- ▶ Consider the following points regarding SAN Volume Controller CG8 nodes (6-core systems):
 - If CPU utilization on the nodes in the I/O group is below 50%, this I/O group is suitable for using compression.
 - If the CPU utilization of a node is sustained above 50% most of the time, this I/O group might not be suitable for compression because it is too busy. If your system is supported for the Non-Disruptive Volume Move feature, this volume can be moved to another I/O group that has the resources that are required for compression.

Table 17-3 shows the preferred practice CPU resource recommendations. Compression is recommended for an I/O Group if the sustained CPU utilization is *below* the values that are listed.

Table 17-3 CPU resources recommendations

Per Node	SAN Volume Controller CF8 and CG8 (4 core)	SAN Volume Controller CG8 (6 core)	SAN Volume Controller CG8 (12 core)	Storwize V7000
CPU already close to or above	25%	50%	No consideration	25%

Add nodes if CPU utilization is consistently above the levels that are shown. Upgrade existing SAN Volume Controller CG8 to SAN Volume Controller-CG8-Dual-CPU-RPQ, as needed.

17.7 Compressed and non-compressed volumes in the same MDisk group

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.

Starting with version V6.4.1.5 and above, the following thresholds changed for 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.

17.8 Application benchmark results

The following section describes the results of performance benchmark testing that uses typical Storwize V7000 and SAN Volume Controller system configurations.

17.8.1 Synthetic workloads

The tests were run with code level 7.2.0.1. To demonstrate the variability between workloads that demonstrate zero and 100% temporal locality, you must modify traditional block benchmark tools to enable repeatable random workloads. This process results in best case and worst case raw block performance workloads when compressed volumes are used. These tests were run with a known compressibility of data blocks by using a 1 MB pattern file that has a 65% compression ratio (see Table 17-4).

Table 17-4 Comparisons of Storwize V7000 and SAN Volume Controller CG8

Workload	Storwize V7000 Worst Case	Storwize V7000 Best Case	SAN Volume Controller CG8 Worst Case	SAN Volume Controller CG8 Best Case	SAN Volume Controller CG8 Dual CPU Worst Case	SAN Volume Controller CG8 Dual CPU Best Case
Read Miss 4 KB Random IOPs	1,259	60,948	1,928	133,328	61,230	158,163
Write Miss 4 KB Random IOPs	1,165	12,455	2,155	77,612	24,312	98,573
70/30 Miss 4 KB Random IOPs	1.642	51,033	2,716	108,318	50,984	127,618

The following configuration was used for the tests:

- ▶ Storwize V7000
- ▶ 96 Disks
- ▶ RAID 5
- ▶ SAN Volume Controller CG8 model with IBM XIV Gen3 Full Storage System as back-end storage

From these purely synthetic workload tests, there is a wide variability in the compressed performance. This test shows the boundaries and in a real user scenario, the results are likely in the middle.

17.9 Standard benchmark tools

Traditional block and file-based benchmark tools (such as IOmeter, IOzone, dbench, and fio) that generate truly random 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.

17.10 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 V6.4.1.5 and V7.1.0.1 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.

17.11 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 MDisks. As a result, frequently read areas of the compressed volumes are serviced from SSDs. Easy Tier on non-compressed volumes operates as before and it is based on read and write operations that are smaller than 64 KB.

For more information about implementing IBM Easy Tier with IBM Real-time Compression, see *Implementing IBM Easy Tier with IBM Real-time Compression*, TIPS1072

17.12 Compression on SAN Volume Controller with Storwize V700

If you have an SAN Volume Controller system setup with Storwize V7000 as backend storage and if you plan to implement compression, it is recommended to configure compression volumes on the SAN Volume Controller system and not on the backend storage Storwize V7000 system.

17.13 Related Publications

For more information, see the following publications:

- ▶ Implementing IBM Real-time Compression in SAN Volume Controller and IBM Storwize V7000: *Implementing IBM Real-time Compression in SAN Volume Controller and IBM Storwize V7000*, TIPS1083
- ▶ Real-time Compression in SAN Volume Controller and Storwize V7000: *Real-time Compression in SAN Volume Controller and Storwize V7000*, REDP-4859
- ▶ Implementing IBM Easy Tier with IBM Real-time Compression: *Implementing IBM Easy Tier with IBM Real-time Compression*, TIPS1072



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.

Disk response time

The time that is needed for a disk I/O operation to complete includes the service time for actual I/O processing and the wait time for potential I/O queuing on the IBM i host.

Main memory

Single-level storage makes main memory work as a large cache. Reads are done from pages in main memory, and read requests to disk are done only when the needed page is not in main memory. 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 application closes files, and whether it uses journaling.

Planning for IBM i capacity

To correctly plan the disk capacity that is virtualized by SAN Volume Controller or Storwize V7000 for IBM i, you must consider that the IBM i block translation for external storage is formatted in 512-byte blocks.

IBM i disks have a block size of 520 bytes. The SAN Volume Controller and Storwize V7000 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 another ninth sector; it stores the 8-byte headers of the 520-byte sectors in the ninth sector, and therefore changes the previous 8* 520-byte blocks to 9* 512-byte blocks.

The data that was previously stored in 8 * sectors is now spread across 9 * sectors, so the required disk capacity on SAN Volume Controller or Storwize V7000 is 9/8 of the IBM i usable capacity. Also, the usable capacity in IBM i is 8/9 of the allocated capacity in these storage systems.

Therefore, when a SAN Volume Controller or Storwize V7000 is attached to IBM i, you should have the capacity overhead on the storage system to use only 8/9ths of the effective capacity.

The performance effect of block translation in IBM i is insignificant or negligible.

Connecting SAN Volume Controller or Storwize to IBM i

SAN Volume Controller or Storwize V7000 can be attached to IBM i 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.

Native connection

Native connection requires that IBM i logical partition (LPAR) is in POWER7 and that IBM i level V7.1, Technology Release 6, Resave 710-H or higher is installed.

Connection with SAN switches

If 8 Gb Fibre Channel (FC) adapters feature number #5735 or #5273 are implemented in IBM i, the connection to SAN Volume Controller or Storwize V7000 must use SAN switches.

Connection with or without SAN switches

If 4 Gb FC adapters feature number #5774 or #5276 are used in IBM i, the connection to SAN Volume Controller or Storwize V7000 can be made via SAN switches or directly attached without the use of the switches.

For resiliency and performance reasons, we recommend connecting SAN Volume Controller or Storwize V7000 to IBM i with Multipath by using two or more FC adapters.

IBM i enables SCSI command tag queuing in the LUNs from SAN Volume Controller or Storwize V7000; the queue depth on a LUN with this type of connection is 16.

Connection with VIOS_NPIV

Connection with VIOS_NPIV requires that IBM i partition is in POWER6 server, POWER7 processor-based server, Power Systems blade server, or in IBM PureFlex System. It also requires IBM i V7.1 Technology Release 6 or higher.

For resiliency and performance reasons, we recommend connecting SAN Volume Controller or Storwize V7000 to IBM i in Multipath by using two or more VIOS.

You must adhere to the following rules for mapping server virtual FC adapters to the ports in VIOS when an NPIV connection is implemented:

- ▶ 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 NPIV mapping and a connection with VIOS VSCSI.
- ▶ If PowerHA solutions of IBM i independent auxiliary storage pool (IASP) is implemented, you must map the virtual FC adapter of System disk pool to different port than virtual FC adapter of the IASP.
- ▶ The SCSI command tag queue depth on a LUN with this type of connection is 16.

Connection with VIOS virtual SCSI

A connection in VIOS virtual SCSI (VSCSI) mode requires the following components:

- ▶ IBM i partition is in POWER6 server, POWER7 processor-based server, Power Systems blade server, or in IBM PureFlex System
- ▶ IBM i V6.1.1 or higher
- ▶ IBM i V7.1 and PowerHA for i V7.1 are required for PowerHA support of copy services solutions with Storwize V7000 or SAN Volume Controller

The use of Multipath with two or more VIOS is recommended for resiliency and performance. When you are implementing Multipath with this type of connection, consider the following points:

- ▶ IBM i Multipath is performed with two or more virtual SCSI (VSCSI) adapters, with 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 SAN Volume Controller / Storwize LUN.
- ▶ In addition to IBM i Multipath, we implement Multipath in each VIOS by using one of the Multipath drives (preferably the SDDPCM driver). The paths that are connecting adapters in VIOS to the LUNs in SAN Volume Controller/Storwize are managed by VIOS Multipath driver.
- ▶ When you are planning the connection of IBM i with VIOS virtual SCSI, consider the queue depth on a LUN with this type of connection is 32.

Setting of attributes in VIOS

For FC adapter attributes with a VIOS Virtual SCSI connection or NPIV connection, we recommend specifying the following attributes for each SCSI I/O Controller Protocol Device (fscsi) device that connects a SAN Volume Controller or Storwize V7000 LUN for IBM i:

- ▶ The attribute `fc_err_recov` should be set to `fast_fail`.
- ▶ The attribute `dyntrk` should be set to `yes`.

Setting 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 is different, and it causes unnecessary retries.

For disk device attributes with a VIOS Virtual SCSI connection, specify the following attributes for each hdisk device that represents a SAN Volume Controller or Storwize LUN that is 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. In 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 maximal number of 32 I/O operations that IBM i operating system allows at a time to one VIOS VSCSI that is connected LUN.
- ▶ Setting `algorithm` to `load_balance` is recommended for performance reason. By setting this value, we ensure that SDDPCM driver in VIOS balances the I/O across available path to Storwize or SAN Volume Controller.

Preparing SAN Volume Controller or Storwize storage for IBM i

In this section, we describe preferred practices for setting up the Storwize V7000 or SAN Volume Controller storage for IBM i host. We include the recommendations about disk drives, disk pools, and LUNs.

Disk drives for IBM i

When Storwize V7000 is implemented with internal hard disk drives (HDDs) for IBM i, make sure that sufficient number of disk arms are provided to the IBM i workload.

Table A-1 shows recommended maximal number of I/O/sec per disk drive at different types of HDDs, different RAID levels, and different read/write ratios of IBM i workload.

Table A-1 Recommended maximal number of I/O per second

Maximum IO/sec per disk drive	70% Read	50% Read
15 K RPM Disk drives	IO/sec per Disk drive	IO/sec per Disk drive
RAID-1 or RAID-10	138	122
RAID-5	96	75
10 K RPM disk drives	IO/sec per Disk drive	IO/sec per Disk drive
RAID-1 or RAID-10	92	82
RAID-5	64	50

For example, the IBM i workload experiences 1500 I/O per second at its peak and the read/write ratio is approximately 50/50. We are planning Storwize V7000 with 15 K RPM disk drives in RAID-10. The following calculation is used for the needed number of disk drives:

$1500 \text{ (IBM i peak IO/sec)} / 122 \text{ (IO/sec per 15 K RPM disk drive in RAID-10 at Read / Write ratio 50/50)} = \text{ape } 12$

Therefore, we recommend implementing at least 12 disk drives for this IBM i workload.

With SAN Volume Controller or when the Storwize V7000 is implemented with background storage, you should still consider enough hard disk drives in the storage system that is connected to Storwize or SAN Volume Controller to accommodate IBM i workload in the peaks. You can take into account that the part of I/O behind Storwize or SAN Volume Controller is done to or from the cache in the background storage system. Therefore, slightly fewer disk arms often are sufficient when comparing the needed disk arms of the storage system that are connected to IBM i without Storwize or SAN Volume Controller.

Defining LUNs for IBM i

The LUNs for IBM i host are defined from block-based storage. We create them in the same way as for open hosts. The minimal size of an IBM i LUN is 180 MB, which provides 160 MB to IBM i because of block translation. The size of a LUN that is reporting in IBM i can be up to 2 TB, excluding 2 TB. This means that the size of an IBM i LUN (when looking to it in Storwize) must be smaller than $9 / 8 * 2 \text{ TB} \approx 2.25 \text{ TB}$. For more information about the relation between the size of a LUN on Storwize and the size of the same LUN on IBM i, see “Planning for IBM i capacity” on page 604.

In general, the more LUNs that are available to IBM i, the better the performance. The reasons for this are IBM i storage management handling of disk units (LUNs), and the wait time component of disk response time, which is lower when more LUNs are used.

However, more LUNs increases the requirement for more FC adapters on the IBM i because of the addressing restrictions of IBM i if you are using native attachment. With VIOS attached to IBM i, more LUN increases the complexity in implementing and management.

The sizing process determines the correct number of LUNs that are required to access the needed capacity while still meeting performance objectives. Considering these aspects and the preferred practices, we make the following recommendations:

- ▶ For any IBM i disk pool (ASP), define all of the LUNs to be the same size.
- ▶ The minimum LUN size should be 80 GB.
- ▶ You should not define LUNs bigger than 400 GB.
- ▶ A minimum of 6 * LUNs for each ASP or LPAR.
- ▶ To support future product enhancements, the IBM i boot device (LoadSource) that is created is at least 80 GB.

Data layout

Spreading workloads across all Storwize V7000 or SAN Volume Controller components maximizes the utilization of the hardware resources in the storage system. However, it is always possible when sharing resources that performance problems might arise because of contention on these resources. Isolation of workloads is most easily accomplished where each ASP or LPAR has its own managed storage pool. This configuration ensures that you can place data where you intend. I/O activity should be balanced between the two nodes or controllers on the SAN Volume Controller or Storwize V7000.

We make the following preferred practice recommendations for the layout:

- ▶ Isolate critical IBM i workloads.
- ▶ Use only IBM i LUNs on any storage pool rather than mixed with non-IBM i LUNs.

- ▶ If production and development workloads are mixed in storage pools, the customer must understand that this configuration can affect production performance.

Solid-state drives

The use of solid-state drives (SSDs) with the Storwize V7000 or SAN Volume Controller is done through Easy Tier. Even if you do not plan to install SSDs you can still use Easy Tier to evaluate your workload and provide information about the benefit you might gain by adding SSDs in the future.

Before SSDs are implemented, the performance improvement of SSD with Easy Tier to the host system can be estimated by IBM's Disk Magic modeling. In Disk Magic, you insert the planned configuration of SSDs and HDDs and select one of the predefined skew levels. With IBM i, select the skew level Very low because the degree of skew of an IBM i workload is often small because of spreading of the objects by IBM i storage management.

When Easy Tier automated management is used, it is important to allow Easy Tier some space to move data. You should not allocate 100% of the pool capacity; instead, leave some capacity deallocated to allow Easy Tier migrations. At a minimum, leave one extent free per tier in each storage pool. However, for optimum use of future functions, plan to leave 10 extents free total per pool.

There is also an option to create a disk pool of SSD in Storwize V7000 or SAN Volume Controller, and create an IBM i ASP that uses disk capacity from the SSD pool. The applications that are running in that ASP experience a performance boost.

Note: IBM i data relocation methods, such as ASP balancing and Media preference, are not available to use with SSDs in Storwize V7000 or SAN Volume Controller.

Sizing Fibre Channel adapters in IBM i and VIOS

The following Fibre Channel adapters are used in IBM i when Storwize or SAN Volume Controller are connected in native mode:

- ▶ 8 Gb PCIe Dual Port Fibre Channel Adapter feature number 5735, or feature number 5273 (Low Profile)
- ▶ 4 Gb PCIe Dual Port Fibre Channel Adapter feature number 5774, or feature number 5276 (Low Profile)
- ▶ For VIOS_NPIV connection, use the following FC adapters in VIOS:
 - 8 Gb PCIe Dual Port Fibre Channel Adapter feature number 5735 or feature number 5273 (Low Profile)
 - 8 Gb PCIe2 4-Port Fibre Channel Adapter feature number 5729

To determine the number of FC adapters that sustain a particular IBM i workload without performance bottlenecks, consider the measured throughput that is specified in Table A-2.

Table A-2 Throughput of Fibre Channel adapters

Maximum I/O rate per port	8 Gb 2-port adapter	4 Gb 2-port adapter
Maximum IO/sec per port	33000 I/O per second	15000 I/O per second
Maximum sequential throughput per port	1100 MBps	310 MBps
Maximum transaction throughput per port	530 MBps	250 MBps

Zoning SAN switches

With native connection and the connection with VIOS_NPIV, zone the switches so that one worldwide port name (WWPN) of one IBM i port is in zone with two ports of Storwize V7000 or SAN Volume Controller, each port from one node canister. By using this configuration, we ensure resiliency for the I/O to and from a LUN that is assigned to that WWPN. If the preferred node for that LUN fails, the I/O rate continues using the non-preferred node.

Note: In a SAN Volume Controller Stretched Cluster configuration, you might need to create two zones, each containing an IBM i port and one port from SAN Volume Controller. The two zones overlap on the IBM i port.

When you are connecting with VIOS virtual SCSI, zone one physical port in VIOS with all available ports in SAN Volume Controller or Storwize V7000, or with as many ports as possible to allow load balancing. There are a maximum of eight paths that are available from VIOS to SAN Volume Controller or Storwize V7000. SAN Volume Controller or Storwize V7000 ports that are zoned with one VIOS port should be evenly spread between the node canisters.

Boot from SAN

All connection options (Native, VIOS_NPIV, and VIOS Virtual SCSI) support IBM i Boot from SAN. IBM i boot disk (LoadSource) is on a Storwize V7000 or SAN Volume Controller LUN that is connected in the same way as the other LUNs. There are no special requirements for LoadSource connection.

When you are installing an IBM i operating system with disk capacity on Storwize or SAN Volume Controller, the installation prompts you to select one of the available LUNs for the LoadSource.

IBM i mirroring

Some customers prefer to have more resiliency with the IBM i mirroring function. For example, they use mirroring between two Storwize V7000 or SAN Volume Controller systems, each connected with one VIOS. When you are starting the mirroring process with VIOS connected to Storwize V7000 or SAN Volume Controller, you should add the LUNs to the mirrored ASP by completing the following steps:

1. Add the LUNs from two virtual adapters, with each adapter connecting one to-be mirrored half of LUNs.
2. After mirroring is started for those LUNs, add the LUNs from two new virtual adapters, with each adapter connecting one to-be mirrored half, and so on. This way, you ensure that the mirroring is started between the two SAN Volume Controller or Storwize V7000 and not among the LUNs in the same SAN Volume Controller.

IBM i Multipath

Multipath provides greater resiliency for SAN-attached storage. IBM i supports up to eight paths to each LUN. In addition to the availability considerations, lab performance testing shows that two or more paths provide performance improvements when compared to a single path. Often, two paths to a LUN are the ideal balance of price and performance. You might want to consider more than two paths for workloads in which there is high wait time, or where high I/O rates are expected to LUNs.

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 V7000 or SAN Volume Controller, 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. With VIOS VSCSI, connection the virtual SCSI adapters for Multipath must be assigned to different VIOS.

Every LUN in Storwize V7000 or SAN Volume Controller uses one node as the preferred node. The I/O traffic to and from the particular LUN normally goes through the preferred node. If that node fails, the I/O operations are transferred to the remaining node. With IBM i Multipath, all of the paths to a LUN through the preferred node are active and the path through the non-preferred node are passive. Multipath uses the load balancing among the paths to a LUN that go through the node, which is preferred for that LUN.

Copy services considerations

Storwize V7000 or SAN Volume Controller has two options for Global Mirror: the classic Global Mirror and the Global Mirror Change Volumes (GMCV) enhancement, which allows for a flexible and configurable RPO allowing Global Mirror to be maintained during peak periods of bandwidth constraint.

GMCV is not currently supported by PowerHA, so you must size the bandwidth to accommodate the peaks or risk affecting production performance. There is a limit of 256 GMCV relationships per system.

The current zoning guidelines for mirroring installations advise that a maximum of two ports on each SAN Volume Controller node or Storwize V7000 node canister are used for mirroring. The remaining two ports on the node or canister should not have any visibility to any other cluster. If you are experiencing performance issues when mirroring is in operation, implementing zoning in this fashion might help to alleviate this situation.

When you are planning for FlashCopy for IBM i, make sure that enough disk drives are available to the FlashCopy target LUNs to keep good performance of production IBM i while the FlashCopy relationships are active. This suggestion is valid for FlashCopy with background copying and without background copying. When you are 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.

Related publications

The publications that are listed in this section are considered particularly suitable for a more detailed discussion of the topics that are covered in this book.

IBM Redbooks publications

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>

Index

Numerics

- 10 Gb Ethernet adapter 7
- 1862 error 123
- 1920 error 193–194, 216
 - bad period count 217
 - troubleshooting 217
- 2145-4F2 node support 5
- 2145-CG8 7, 61

A

- access 19, 60, 72, 228
 - pattern 139
- Access LUN 78
- access option 200
- adapters 227, 315, 350
 - DS8000 296
- administrator 60, 93, 260, 343, 520
- ADT (Auto Logical Drive Transfer) 75
- aggregate workload 72, 98, 283
- AIX 227, 339, 550
 - host 244, 527
 - server migration 568
- alert 20, 26, 182
 - events
 - CPU utilization threshold 458
 - overall back-end response time threshold 459
 - overall port response time threshold 459
- algorithms 138
- alias 40, 47, 488
 - storage subsystem 48
- alignment 349
- amount of I/O 139, 197
- application 61, 184, 227, 270, 339
 - availability 350
 - database 139
 - performance 136, 340
 - streaming video 139
 - testing 151
- Application Specific Integrated Circuit (ASIC) 31
- architecture 72, 242, 264
- array 19, 60, 76, 96, 98–99, 136, 153, 197, 282, 314
 - considerations for storage pool 283
 - layout 282
 - midrange storage controllers 282
 - parameters 76
 - per storage pool 283
 - provisioning 282
 - site, spare 80
 - size, mixing in storage pool 317
- array support library (ASL) 254
- ASIC (Application Specific Integrated Circuit) 31
- ASL (array support library) 254
- asynchronous mirroring 205
- asynchronous mode 158

- asynchronous remote copy 142, 160, 174–177
- attributes 95
- Auto Logical Drive Transfer (ADT) 75
- autoexpand feature 126
- autoexpand option 128
- automatically discover 237
- automation scripts 141
- auxiliary cluster 159
- auxiliary VDisk 176
- auxiliary volume 159, 174
- availability 32, 96, 232
 - versus isolation 233
- average I/O per volume 146
- average I/O rate 146

B

- back-end I/O capacity 272
- back-end storage 269
 - controller 145, 197
- back-end striping 271
- back-end transfer size 285
- background copy 158
 - bandwidth 192
- background write synchronization 158
- backplane 31
- backup 19, 152, 250, 342, 530
 - node 46
 - sessions 349
- bad period count 217
- balance 46, 75, 130, 233, 343
 - workload 138
- bandwidth 19, 62, 153, 161, 179, 182, 225, 228, 341
 - parameter 172, 572
 - requirements 53
- batch workloads 270
- BIOS 69, 252
- blade 40, 42
- BladeCenter 53
- block 77, 131, 138, 146, 340
 - size 342
- boot 229
 - device 248
- bottleneck 340
 - detection feature 36
- boundary crossing 349
- bridge 23
- Brocade 536
 - Webtools GUI 39
- buffer 147, 154, 227, 351
 - credit 187
- bus 240

C

- cache 99, 135, 227, 282, 340–341, 343, 531

- battery failure 390, 397
- block size 315
- disabled 141, 148
- friendly workload 271
- influence 274
- management 63
- mode 143
- partitioning 290
- track size 287
- usage 271
- cache-disabled
 - image mode 206
 - state 207
 - VDisk 141–142
- cache-disabled settings 155
- cache-enabled settings 155
- caching 139, 141
 - algorithm 288
- capacity 25, 60, 99, 131, 347, 547
- cards 252
- case study
 - fabric performance 465
 - performance alerts 458
 - server performance 442
 - top volumes response performance 429
- cfgportip command 55
- change of state 211
- channel 245
- chcluster command 190
- chdev command 244
- chpartnership command 190
- chquorum command 29, 102
- Cisco 19, 494
- CLI 147
 - commands 294, 543
- client 249, 349
- cluster 19, 51, 61, 96, 130, 184, 227, 285, 521
 - affinity 72
- clustered systems advantage 65
- clustering software 242
- coalescing writes 288
- colliding writes 177, 188
- command
 - CreateRelationship 198
 - dd 200
 - prefix removal 8
- commit 149
- complexity 38
- conception 39
- concurrent code update 68
- configuration 17, 96, 226, 340, 521
 - data 237, 539
 - node 531
 - parameters 220, 240
- congestion 19
- connected state 215
- connections 25, 72, 243, 248
- connectivity 247, 520
- consistency 258
 - freeze 215

- consistent relationship 159
- ConsistentStopped state 213, 215
- ConsistentSynchronized state 213, 216
- consolidation 96
- containers 349
- contingency capacity 126
- control 60, 79, 121, 141, 228
- controller ports 284, 296
 - DS4000 317
- controller types, constant 285
- copy rate 148
- copy services 61, 67
 - relationship 506
- core switch 21, 25, 31
- core-edge ASIC 22
- corruption 82
- cost 184
- counters 259
- CPU utilization 63
- CreateRelationship command 198
- cross-bar architecture 31
- CWDM 50, 183

D

- daisy-chain topology 204
- data 20, 60, 196, 227, 340, 519
 - consistency 149
 - corruption, zone considerations 50
 - formats 250
 - integrity 135, 147
 - layout 131, 343
 - strategies 351
 - migration 153, 250
 - planner 323
 - mining 151
 - pattern 340
 - rate 224, 270
 - redundancy 96
 - traffic 26
- data collection, host 524
- data layout 343
- Data Path View 472
- Data Placement Advisor 323
- database 20, 139, 149, 236, 342, 520
 - applications 139
 - log 342
- Datapath Explorer 469
- dd command 200
- debug 524
- decibel 186
 - milliwatt 186
- dedicated ISLs 26
- degraded performance 196
- design 18, 60, 234
- destage 288
 - size 315
- DetectMDisks GUI option 75
- device 18, 229, 525
 - adapter 293
 - adapter loading 78

- data partitions 77, 315
- driver 207, 242
- diagnostic 244, 539
- direct-attached host 284
- director-class SAN switch 32
- disaster recovery 65, 214
 - solutions 201
- disconnected state 213
- discovery 75, 121
- discovery method 236
- disk 60, 131, 146, 236, 340, 550
 - access profile 139
 - latency 341
- Disk Magic 98, 270
- distance 50–51, 183–184
 - extension 184
 - limitations 51, 183
- domain 81, 96
 - ID 50
- download 544
- downtime 148
- drive loops 74
- driver 242, 520
- DS4000 72, 99, 106, 260, 283, 341, 538
 - controller ports 317
 - storage controllers 42
- DS4800 48, 315
- DS5000
 - array and cache parameters 314
 - availability 315
 - default values 315
 - storage controllers 42
 - Storage Manager 316
 - throughput parameters 314
- DS6000 72, 246, 283, 539
- DS8000 72, 99, 246, 282, 539
 - adapters 296
 - alias considerations 48
 - architecture 97
 - bandwidth 294
 - controller ports 296
 - LUN ID 114
- dual fabrics 41
- dual-redundant switch controllers 31
- DWDM 50, 183
 - components 186
- dynamic tracking 246

E

- Easy Tier 6, 266, 320
 - activate 324
 - check mode 334
 - check status 338
 - CLI 329
 - evaluation mode 330
 - GUI activate 335
 - manual operation 293
 - operation modes 323
 - processes 322
- edge switch 19–22, 31, 182

- efficiency 138
- egress port 31
- email 51, 184, 260
- EMC Symmetrix 92
- error 226, 520–521
 - handling 552
 - log 524, 551
 - logging 93, 550
- error code 551
 - 1625 81
- Ethernet ports 5
- event 20, 72, 138, 246
- exchange 149
- execution throttle 252
- expansion 20
- explicit sequential detect 288
- extended-unique identifier 54
- extenders 184
- extension 50, 183
- extent 79, 131, 294, 342
 - balancing script 108
 - size 131, 288, 293, 347
 - 8 GB 6
- extent pool 78–79
 - affinity 291
 - storage pool striping 293
 - striping 79, 291

F

- fabric 3, 17, 22, 153, 226–227, 520, 593
 - hop count limit 186
 - isolation 233
 - login 235
 - outage 20, 182
 - watch 35
- failover 139, 226, 521
 - logical drive 75
 - scenario 177
- failure boundary 97, 346, 352
- FASTT FC2-133 252
- fastwrite cache 280
- fault tolerant LUNs 100
- FC flow control mechanism 20, 182
- fcs adapter 245
- fcs device 245
- Fibre Channel 18, 182, 184, 225, 227, 235, 520
 - adapters 315
 - IP conversion 51, 184
 - port speed 439
 - ports 53, 72, 229, 545
 - router 37, 184
 - traffic 20, 182
- file system 147, 253
 - level 258
- firmware 219
- FlashCopy 8, 64, 93, 276, 287, 521, 550
 - applications 552
 - creation 149
 - I/O operations 277
 - incremental 278

- mapping 94, 134, 147
 - preparation 169
- preparation 148, 207
- relationship
 - target as Remote Copy source 167
 - thin provisioned volumes 281
- rules 154
- source 551
- storage pool 277
- target 143
- target, Remote Copy source 8
- thin provisioning 281
- flexibility 139, 242
- fmt disk security delete feature 212
- force flag 112, 135
- force parameter 497
- foreground I/O 158
 - latency 192
- free extents 138
- full stride writes 99, 288
- full synchronization 198
- fully allocated copy 138
- fully allocated VDisk 138
- fully connected mesh 204

G

- General Public License 260
- Global Mirror 158, 160, 188
 - 1920 errors 217
 - bandwidth parameter 192
 - bandwidth resource 190
 - change to Metro Mirror 201
 - features by release 165
 - parameters 172, 189
 - partnership 173
 - partnership bandwidth parameter 161
 - planning 195
 - planning rules 194
 - relationship 180, 200
 - restart script 222
 - switching direction 200
 - upgrade scenarios 208
 - writes 176
- gm_inter_cluster_delay_simulation parameter 189
- gm_intra_cluster_delay_simulation parameter 189
- gm_link_tolerance parameter 189
- gm_max_host_delay parameter 189
- gm_max_hostdelay parameter 172–173
- gmlinktolerance parameter 172–173, 192, 220
 - bad periods 193
 - disabling 195, 222
- GNU 260
- grain size 287
- granularity 131, 258
- graphs 241

H

- HACMP 247
- hardware

- redundancy 72
 - SVC node 521
- upgrade 67
- HBA 39, 53, 61, 188, 227–228, 233, 245, 520
 - parameters for performance tuning 244
 - replacement 511
 - zoning 46
- head-of-line blocking 20, 182
- health checker 248
- health, SAN switch 541
- heartbeat 185
 - messages 161
 - messaging 161
 - signal 104
- heterogeneous 60, 522
- high-bandwidth hosts 21, 31
- hop count 186
- hops 19
- host 19, 72, 130, 284, 339, 520
 - cluster implementation 242
 - configuration 45, 154, 226, 522
 - creation 47
 - data collection 524
 - definitions 236
 - HBA 46
 - I/O capacity 275
 - information 69, 525
 - mapping 521
 - port login 228
 - problems 520
 - system monitoring 225
 - systems 67, 225, 520
 - type 76
 - volume mapping 229
 - zone 44, 47, 130, 227, 522
- host-based mirroring 258
- hot extents 320

I

- I/O balancing 350
- I/O capacity 272
 - rule of thumb 280
- I/O collision 282
- I/O governing 139
 - rate 141
 - throttle 139
- I/O group 25, 46, 61, 64, 130, 138, 194, 233
 - host mapping 230
 - mirroring 26
 - performance 267
 - performance scalability 267
 - switch splitting 25
- I/O Monitoring Easy Tier 322
- I/O operations, FlashCopy 277
- I/O per volume 146
- I/O performance 245
- I/O rate calculation 146
- I/O rate setting 141
- I/O resources 270
- I/O service times 98

- I/O size of 256 KB 286
- I/O stats 217
- I/O throughput delay 148
- I/O workload 346
- ICL (intercluster link)
 - definition 159
 - distance extension 183
 - parameters 161
- identical data 212
- identification 230
- idling state 216
- IdlingDisconnected 216
- IEEE 250
- Ignorer Bandwidth parameter 195
- image 64, 131, 146, 229, 343
- image mode 67, 136, 205, 235
 - virtual disk 142
 - volumes 206
- image type VDisk 137
- import failed 123
- improvements 68, 248, 263, 284
- InconsistentCopying state 213, 215
- InconsistentStopped state 213–214
- incremental FC 278
- in-flight write limit 190
- infrastructure 141
 - tiering 271
- ingress port 31
- initiators 242
- installation 17, 22, 105, 153, 251
- insufficient bandwidth 20, 182
- integrated routing 37
- Integrated Routing licensed feature 37
- integrity 147
- intercluster communication 161
- intercluster link (ICL) 181
 - definition 159
 - distance extension 183
 - parameters 161
- intercluster link bandwidth 180
- interlink bandwidth 161
- internal SSD 8
- internode communications zone 41
- interoperability 8, 53
- interswitch link (ISL) 19–20, 22, 182
 - capacity 31
 - hop count 174
 - oversubscription 20
 - trunk 31, 182
- intracluster copying 197
- intracluster Global Mirror 188
- intracluster Metro Mirror 174
- iogrp 135, 228
- iometer 260
- IOPS 227, 340
- iostat tool 260
- IP traffic 51, 184
- iSCSI
 - driver 54
 - initiators 54

- protocol 4
 - limitations 56
- qualified name 54
- support 54
- target 54
- ISL (interswitch link) 19–20, 22, 182
 - capacity 31
 - hop count 174
 - oversubscription 20, 182
 - trunk 31, 182
- isolated SAN networks 296
- isolation versus availability 233

J

- journal 253

K

- kernel 252
- keys 235, 243

L

- last extent 131
- latency 31, 149, 179, 341
- LDAP directory 4
- lease expiry event 183
- lg_term_dma attribute 245
- lg_term_dma parameter 245
- licensing 8
- limitation 18, 206, 240, 342, 539
- limits 61, 240, 351
- lines of business (LOB) 346
- link 65, 182, 196
 - bandwidth 161, 185
 - latency 161, 185
 - speed 178
- Linux 252
- livedump 531
- load balance 139, 233
 - traffic 25
- load balancing 130, 248, 251
- LOB (lines of business) 346
- local cluster 159
- local hosts 159
- log 551
- logical block address 551
- logical drive 120, 244, 343
 - failover 75
 - mapping 78
- logical unit (LU) 64, 229
- logical unit number 206
- logical volumes 349
- login from host port 228
- logs 149, 342, 526
- long-distance link latency 185
- long-distance optical transceivers 51
- loops 316
- LPAR 250
- lquerypr utility 106

- lsarray command 294
- lscontroller command 117
- lsfabric command 497
- lshbaportcandidate command 497
- lshostconnect command 81
- lsmdisklba command 93
- lsmigrate command 109
- lsportip command 56
- lsquorum command 102
- lsrank command 294
- lsvdisklba command 93
- LU (logical unit) 64, 229
- LUN 67, 72, 99, 142, 206, 226, 228, 282
 - access 242
 - ID, DS8000 114
 - mapping 112, 229
 - masking 50, 81, 522
 - maximum 101
 - number 112
 - selection 100
 - size on XIV 303
- LVM 248
 - volume groups 349

M

- maintenance 69, 235, 520
 - procedures 545
- managed disk 545
 - group 95, 136, 346, 361
- Managed Disk Group Performance report 413
- managed mode 78, 137, 315
- management 23, 226, 343, 522
 - capability 228
 - port 228
 - software 230
- map 154, 230, 351, 546
- mapping 112, 134, 147, 226, 243, 521
 - rank to extent pools 291
 - VDisk 233
- masking 67, 81, 154, 228, 522
- master 69, 147
 - cluster 159
 - volume 159
- max_xfer_size attribute 245
- max_xfer_size parameter 245
- maxhostdelay parameter 193
- maximum I/O 349
- maximum transmission unit 55
- McDATA 536
- MDisk 95, 131, 233, 342
 - checking access 106
 - group 342
 - moving to cluster 122
 - performance 449
 - performance levels 99
 - removing reserve 244
 - selecting 97
 - transfer size 285
- media 220, 545
 - error 551

- medium errors 550
- members 47, 74, 521
- memory 146, 226, 342, 531
- messages 233
- metadata 127
 - corruption 123
- Metro Mirror 64, 158–159, 196
 - planning rules 194
 - relationship 149
 - change to Global Mirror 201
- microcode 552
- Microsoft Volume Shadow Copy Service 155
- migration 19, 67, 93, 135, 235, 550
 - data 137, 250
 - scenarios 25
- Mirror Copy activity 63
- mirrored copy 174
- mirrored data 258
- mirrored foreground write I/O 158
- mirrored VDisk 129
- mirroring 50, 183, 196, 248
 - considerations 258
 - relationship 50
- misalignment 349
- mkpartnership command 172, 174, 179
- mkrcrelationship command 174, 201
- mode 64, 78, 205, 227, 230, 315, 343, 525, 547
 - settings 155
- monitoring, host system 225
- MPIO 247
- multicenter installations 22
- multicenter mirroring 201
- multipath drivers 106
- multipath I/O 247
- multipath software 242
- multipathing 72, 226, 519
 - software 233–234
- multiple cluster mirroring 166
 - topologies 202
- multiple paths 139, 233, 522
- multiple vendors 53
- multitiered storage pool 105

N

- names 47, 130, 251
- naming convention 40, 105, 130, 486
- native copy services 206
- nest aliases 46
- no synchronization option 129
- NOCOPY 148
- node 19–20, 130–131, 182, 226, 228, 263, 270, 285, 520–521
 - adding 65
 - failure 138, 235
 - maximum 61
 - port 40, 138, 221, 227, 522
- Node Cache performance report 400
- Node level reports 388, 396
- num_cmd_elem attribute 244–245

O

- offline I/O group 135
- OLTP (online transaction processing) 342
- online transaction processing (OLTP) 342
- operating systems
 - alignment with device data partitions 349
 - data collection methods 524
 - host pathing 233
- optical distance extension 51
- optical multiplexors 51, 184
- optical transceivers 51
- Oracle 248, 347
- oversubscription, ISL 20, 182

P

- parameters 139, 220, 228, 341
- partitions 249
- partnership bandwidth parameter 189
- path 19, 24, 68, 72, 138, 226, 270, 351, 521–522
 - count connection 83
 - selection 247
- pcmquerypr command 243
- performance 20, 60, 96, 130, 182, 225, 263, 339, 520
 - advantage 98
 - striping 97
 - back-end storage 269
 - characteristics 131, 260
 - LUNs 99
 - tiering 271
 - degradation 99, 196, 282
 - degradation, number of extent pools 294
 - improvement 136
 - level, MDisk 99
 - loss 160
 - monitoring 223, 228
 - reports
 - Managed Disk Group 413
 - SVC port performance 433
 - requirements 68
 - scalability, I/O groups 267
 - statistics 8
 - storage pool 96
 - tuning, HBA parameters 244
- Perl packages 108
- persistent reserve 106
- physical link error 51
- physical volume 249, 351
- Plain Old Documentation 111
- plink.exe utility 575
- PLOGI 235
- point-in-time consistency 176
- point-in-time copy 151, 206
- policies 242, 248
- pool 60
- port 19, 64, 72, 221, 226, 284, 521–522
 - bandwidth 31
 - channel 37
 - density 31
 - mask 228

- naming convention in XIV 85
- types 92
- XIV 305
- zoning 39

- power 544
- preferred node 46, 130, 190, 233
- preferred owner node 138
- preferred path 72, 138–139, 233, 525
- prefetch logic 288
- prepared state 221
- prezoning tips 40
- primary considerations for LUN attributes 99
- primary environment 65
- problems 19, 69, 92, 154, 340, 519
- profile 77, 120, 139, 315, 544
- properties 254
- provisioning 105
 - LUNs 99
- pSeries 49, 260
- PuTTY session 575
- PVID 250–251

Q

- queue depth 240, 245, 252, 284, 351
- queue_depth hdisk attribute 244
- quick synchronization 198
- quiesce 134, 149
- quorum disk 102
 - considerations 104
 - placement 29

R

- RAID 78, 99, 136, 197, 316
 - array 220, 345–346
- RAID 5
 - algorithms 417, 421
 - storage pool 273
- random I/O performance 272
- random writes 273
- rank to extent pool mapping
 - additional ranks 293
 - considerations 292
- RAS capabilities 5
- RC management 196
- RDAC 72, 106
- read
 - cache 340
 - data rate 390
 - miss performance 138
 - stability 177
- real address space 126
- real capacity 549
- Real Time Performance Monitor 268
- rebalancing script, XIV 305
- reconstruction 178
- recovery 120, 137, 149, 226, 545
 - point 195
- redundancy 31, 185, 227, 522
- redundant paths 227

- redundant SAN 81, 298
- registry 236, 526
- relationship 72, 249
- relationship_bandwidth_limit parameter 172, 189
- reliability 45, 106
- remote cluster 159, 185
 - upgrade considerations 69
- Remote Copy
 - functions 4
 - parameters 172
 - relationship 158
 - increased number 167
 - service 158
- remote mirroring 50
 - distance 183
- repairsevdisk command 123
- reports 236, 357
 - Fabric and Switches 439
 - SVC 382
- Request for Price Quotation (RPQ) 20, 254
- reset 235, 521
- resources 60, 121, 141, 226, 532
- response time 427
- restart 154, 235
- restore 200, 250
- restricting access 242
- resynchronization support 188
- reverse FlashCopy 4, 60
- risk assessment 93
- rmhostport command 512
- rmmdisk command 112
- rmvdisk command 497
- roles 342, 345
- root 243, 509
- round-robin method 122
- router technologies 184
- routers 185
- routes 38
- routing 72
- RPQ (Request for Price Quotation) 20, 254
- rule of thumb for SVC response 427
- rules 146, 226, 522

S

- SameWWN.script 74
- SAN 17, 61, 153, 225, 350, 519–520
 - availability 233
 - bridge 23
 - configuration 17, 181
 - fabric 17, 153, 228, 233, 522
 - Health Professional 491
 - performance monitoring tool 195
 - zoning 138, 472
- SAN switch 31
 - director class 32
 - edge 31
 - models 31
- SAN Volume Controller 3, 18, 20, 38, 45, 59–60, 130, 158, 182, 225, 263, 342, 519
 - back-end read response time 416

- caching 77, 315
- CLI scripts 575
- cluster 19, 59, 96, 232, 531
 - copy services relationship 506
 - migration 514
- clustered system
 - growth 64
 - splitting 66
- code upgrade 507
- configuration 228
- Console code 6
- Entry Edition 5
- error log 123
- extent size 288
- features 61
- health 470
- installations 21, 270
- managed disk group information 361
- master console 150
- multipathing 252
- node 44, 521
- nodes 24, 67, 81, 228, 545
 - redundant 61
- performance 380
 - Top Volume I/O Rate 425
 - Top Volumes Data Rate 424
- performance benchmarks 398
- ports 471
- rebalancing script 305
- reports
 - cache performance 423
 - cache utilization 400
 - CPU utilization 386, 395
 - CPU utilization by node 386, 395
 - CPU utilization percentage 404, 412
 - Dirty Write percentage of Cache Hits 404, 412
 - I/O Group Performance reports 384
 - Managed Disk Group 413
 - MDisk performance 449
 - Node Cache performance 400
 - Node CPU Utilization rate 386
 - node statistics 385
 - overall I/O rate 387
 - overused ports 437
 - Port Performance reports 433
 - Read Cache Hit percentage 400, 410
 - Read Cache Hits percentage 404, 412
 - Read Data rate 390
 - Read Hit Percentages 404, 412
 - Readahead percentage of Cache Hits 404, 412
 - report metrics 385
 - response time 388, 396
 - Top Volume Cache performance 422
 - Top Volume Data Rate performances 422
 - Top Volume Disk performance 422, 425
 - Top Volume I/O Rate performances 422
 - Top Volume Performance reports 422
 - Top Volume Response performances 422
 - Total Cache Hit percentage 400, 410
 - Total Data Rate 390

- Write Cache Flush-through percentage 405, 412
- Write Cache Hits percentage 405, 412
- Write Cache Overflow percentage 405, 412
- Write Cache Write-through percentage 405, 412
- Write Data Rate 390
- Write-cache Delay Percentage 405, 412
- restrictions 65
- software 229, 521
- storage zone 48
- traffic 380
- V5.1 enhancements 4
- V7000 considerations 307
- XIV 5
 - considerations 83
 - port connections 306
 - zoning 38, 46
- SANHealth tool 489
- save capacity 127
- scalability 18, 59
- scaling 67
- scripting toolkit 579
- scripts 240
- SCSI 138, 235, 543, 551
 - commands 242, 544
 - disk 250
- SCSI-3 242
- SDD (Subsystem Device Driver) 8, 45, 72, 106, 135, 207, 227, 247, 525
 - Linux 253
- SDDDSM 230, 525
- sddgetdata script 526
- SDDPCM 247
 - features 248
- sddpcmgetdata script 526
- SE VDisks 126
- secondary site 65
- secondary SVC 65
- security 39, 248
 - delete feature 212
- segment size 77, 315
- sequential 131, 227, 341
- serial number 229–230
- server 19, 62, 149, 248–249, 251, 314, 350, 525
- service 69, 351, 520
 - assistant 6
- setquorum command 103
- settings 220, 244, 340, 522
- setup 244, 347, 522
- SEV 152
- SFP 51
- shortcuts 40
- showvolgrp command 81
- shutdown 153, 235
- single initiator zones 45
- single storage device 233
- single-member aliases 47
- single-tiered storage pool 104
- site 67, 142, 551
- slice 349
- slot number 49

- slots 74, 316
- slotted design 31
- snapshot 153
- software 18, 20, 38, 45, 182, 226, 237, 520–521
 - locking methods 242
- Solaris 254, 525
- solid state drive (SSD) 4, 6, 60
 - managed disks, quorum disks 29
 - mirror 8
 - quorum disk 102
 - redundancy 267
 - upgrade effect 505
- solution 18, 223, 340, 486
- source 38, 551
- source volume 159
- space 131
- space efficient 128
 - copy 138
- space-efficient function 281
- space-efficient VDisk 152, 549
 - performance 127
- spare 20, 296
- speed 32, 197
- split cluster
 - quorum disk 103
- split clustered system 27
- split clustered system configuration 27–28
- split SVC I/O group 4
- SSD (solid state drive) 4, 6, 60
 - managed disks, quorum disks 29
 - mirror 8
 - quorum disk 102
 - redundancy 267
 - upgrade effect 505
- SSPC 108
- standards 53
- star topology 203
- state 137, 226, 532
 - ConsistentStopped 215
 - ConsistentSynchronized 216
 - idling 216
 - IdlingDisconnected 216
 - InconsistentCopying 215
 - InconsistentStopped 214
 - overview 211
- statistics 259
 - summary file 323
- status 244, 521, 550
- storage 17, 60, 131, 225, 339, 520
 - administrator role 345
 - bandwidth 161
 - subsystem aliases 48
 - tier attribute 321
 - traffic 19
- Storage Advisor Tool 326
- storage controller 40–41, 61, 76, 96, 105, 142, 145, 206, 282, 487
 - LUN attributes 99
- Storage Manager 74, 538
- Storage Performance Council 260

- storage pool
 - array considerations 283
 - I/O capacity 273
 - performance 96
 - striping 79, 291
 - extent pools 293
- Storwize V7000 43, 87, 89, 283, 307
 - configuration 89
 - performance 381
 - traffic 381
- streaming 342
 - video application 139
- stride writes 272, 315
- stripe 96
 - across disk arrays 97
- striped mode 148, 343
 - VDisks 347
- striping 76, 349
 - DS5000 314
 - performance advantage 283
 - workload 98, 283
- sub-LUN migration 320
- subsystem cache influence 274
- Subsystem Device Driver (SDD) 8, 45, 72, 106, 135, 207, 227, 230, 247, 525
 - for Linux 253
- support 342
- support alerts 494
- svcinfo command 107, 134, 229, 521
- svcinfo lscluster command 189
- svcinfo lscontroller controllerid command 523
- svcinfo lsmigrate command 107
- svcinfo lsnode command 523
- svcmmon tool 62
- svctask chcluster
 - command 189
- svctask command 107, 153, 254, 530
- svctask detectmdisk command 75, 123
- svctask migratetoimage command 123
- svctask mkrcrelationship command 201
- svctask mkvdisk command 123
- svctask rmvdisk command 123
- SVCTools package 107
- switch 225, 520
 - fabric 19, 228
 - failure 20, 259
 - interoperability 53
 - port layout 31–32
 - ports 25, 470
 - splitting 25
- sync flag 201
- sync option 199
- synchronization 185
- synchronized relationship 159
- synchronized state 196
- synchronous mode 158
- synchronous remote copy 159
- system 146, 225, 341, 524
 - performance 131, 253, 531
 - statistics setting 268

T

- table space 342
- tape media 19, 199, 227
- target 92, 227, 545
 - port 81, 228
 - volume 159
- test 19, 225
- thin provisioning 279, 287
 - FlashCopy considerations 281
- thin-provisioned volume 126
 - FlashCopy 127
- thread 240
- three-way copy service functions 205
- threshold 20, 182, 196
- throttle 139, 252
 - setting 140
- throughput 234, 245, 270, 341–342
 - environment 342
 - RAID arrays 98
 - requirements 77
- throughput-based workload 340
- tiers 105, 270–271
- time 19, 72, 226, 270
- tips 40
- Tivoli Storage Manager 207, 342, 348
- Tivoli Storage Productivity Center 195, 359, 524
 - performance best practice 381
 - top 10 reports 382
 - volume performance reports 422
- tools 225, 522
- topology 18, 524
 - issues 24
 - problems 24
- Topology Viewer
 - Data Path Explorer 469
 - Data Path View 472
 - navigation 468
 - SAN Volume Controller and Fabric 470
 - SAN Volume Controller health 470
 - zone configuration 472
- Total Cache Hit percentage 400, 410
- traffic 19, 25, 233
 - congestion 20
 - Fibre Channel 53
 - isolation 25
 - threshold 26
- transaction 76, 149
 - environment 342
 - log 342
- transaction-based workloads 314, 340–341
- transceivers 184
- transfer 227, 340
- transit 19
- triangle topology 203
- troubleshooting 39, 225, 519
- tuning 225

U

- UID field 112, 547

- unique identifier 229
- UNIX 149, 260
- unmanaged MDisk 137, 206
- unsupported topology 205
- unused space 131
- upgrade 219, 235, 497, 543–544
 - code 556
 - scenarios 208
- Upgrade Test Utility 499
- user 31, 60, 235
 - data 127
 - interface 5
- utility 260

V

- V7000
 - ports 310
 - SAN Volume Controller considerations 307
 - solution 118
 - storage pool 312
 - volume 308
- VDisk 45, 228, 342, 521
 - creation 151
 - mapping 233
 - migration 136, 551
 - mirroring 129
 - size maximum 4
- VDisk deletion 134
- Veritas file sets 249
- VIOS 248–250, 350
 - clients 350
- virtual address space 126
- virtual capacity 128
- virtual disk 138, 250
- Virtual Disk Service 155
- virtual fabrics 36
- virtual SAN 37
- virtualization 59, 343, 519
 - layer 93
 - policy 129
- virtualizing 235
- VMware
 - multipathing 257
 - vStorage APIs 8, 257
- volume
 - group 81, 247
 - allocation 544
 - types 126
- volume mirroring 60, 99, 129, 325
- VSAN 19, 37–38
 - trunking 37
- VSCSI 249, 351

W

- warning threshold 126
- workload 20, 77, 98, 121, 182, 227, 244, 270, 340–341
 - throughput based 340
 - transaction based 340

- type 341
- worldwide node name (WWNN) 38–39, 92
 - setting 74
 - zoning 39
- worldwide port number (WWPN) 39, 66, 82, 227, 285, 521
 - debug 83
 - zoning 39
- write 227, 282, 342
 - ordering 177
 - penalty 272–273
 - performance 130
- write cache destage 273
- WWNN (worldwide node name) 38–39, 92
 - setting 74
- WWPN (worldwide port number) 39, 66, 82, 227, 285, 521
 - debug 83
 - zoning 39

X

- XFP 51
- XIV
 - LUN size 303
 - port naming conventions 85
 - ports 42, 305
 - storage pool layout 306
 - SVC considerations 83
 - zoning 42
- XIV Storage System 283

Z

- zone 38, 153, 228, 522
 - configuration 472
 - name 49
 - SAN Volume Controller 24
 - set 48, 546
 - share 50
- zoning 23, 38, 50, 82, 138, 228, 491
 - configuration 38
 - guideline 183
 - HBA 46
 - requirements 167
 - scheme 40
 - single host 45
 - Storwize V7000 43
 - XIV 42
- zSeries attach capability 97



Redbooks

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

(1.0" spine)

0.875" <-> 1.498"

460 <-> 788 pages



IBM System Storage SAN Volume Controller and Storwize V7000

Best Practices and Performance Guidelines



Redbooks®

Read about best practices learned from the field

Learn about SAN Volume Controller performance advantages

Fine-tune your SAN Volume Controller

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 V7.2.

This book begins with a look at the latest developments with SAN Volume Controller and Storwize V7000 and reviews the changes in the previous versions of the product. 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. Finally, this book highlights several scenarios that demonstrate the preferred practices and performance guidelines.

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.

**INTERNATIONAL
TECHNICAL
SUPPORT
ORGANIZATION**

**BUILDING TECHNICAL
INFORMATION BASED ON
PRACTICAL EXPERIENCE**

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-7521-03

ISBN 0738439762