Redbooks
ibm.com/redbooks

IBM FlashSystem Best Practices and Performance Guidelines for IBM Spectrum Virtualize Version 8.4.2

Antonio Rainero
Carlton Beatty
David Green
Hartmut Lonzer
Jonathan Wilkie
Kendall Williams
Konrad Trojok
Mandy Stevens
Nezih Boyacıoglu
Nils Olsson
Renato Santos
Rene Oehme
Sergey Kubin
Thales Noivo Ferreira
Uwe Schreiber
Vasfi Gucer

Storage

IBM

Redbooks

IBM Redbooks

IBM FlashSystem Best Practices and Performance Guidelines for IBM Spectrum Virtualize Version 8.4.2

February 2022

SG24-8508-00

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

First Edition (February 2022)

This edition applies to IBM Spectrum Virtualize Version 8.4.2.

Contents

Notices

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

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

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

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

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

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

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

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

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

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

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

COPYRIGHT LICENSE:

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

Trademarks

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

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

AIX®
Db2®
DB2®
DS8000®
Easy Tier®
FlashCopy®
Global Technology Services®
HyperSwap®
IBM®
IBM Cloud®
IBM FlashCore®
IBM FlashSystem®
IBM Garage™
IBM Research®
IBM Security™
IBM Spectrum®
Insight®
Interconnect®
MicroLatency®
Orchestrate®
POWER8®
POWER9™
PowerHA®
PowerVM®
ProtecTIER®
Redbooks®
Redbooks (logo) ®
Service Request Manager®
Storwize®
SystemMirror®
Tivoli®
XIV®
z/OS®

The following terms are trademarks of other companies:

Intel, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

ITIL is a Registered Trade Mark of AXELOS Limited.

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

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

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

Ansible, OpenShift, Red Hat, are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

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

VMware, VMware vSphere, and the VMware logo are registered trademarks or trademarks of VMware, Inc. or its subsidiaries in the United States and/or other jurisdictions.

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

Preface

This IBM® Redbooks® publication captures several of the preferred practices and describes the performance gains that can be achieved by implementing the IBM FlashSystem® products that are powered by IBM Spectrum® Virtualize Version 8.4.2. These practices are based on field experience.

This book highlights configuration guidelines and preferred practices for the storage area network (SAN) topology, clustered system, back-end storage, storage pools and managed disks, volumes, Remote Copy services, and hosts.

It explains how you can optimize disk performance with the IBM System Storage Easy Tier® function. It also provides preferred practices for monitoring, maintaining, and troubleshooting.

This book is intended for experienced storage, SAN, IBM FlashSystem, SAN Volume Controller, and IBM Storwize® administrators and technicians. Understanding this book requires advanced knowledge of these environments.

Authors

This book was produced by a team of specialists from around the world.

Antonio Rainero is an Infrastructure Architect working for Kyndryl Italy. Before joining Kyndryl, Antonio was Executive Technical Specialist for the IBM Global Technology Services® organization in IBM Italy. He has more than 20 years of experience in the delivery of storage services for Open Systems and IBM z/OS® clients. His areas of expertise include storage systems implementation, SANs, storage virtualization, performance analysis, disaster recovery, and high availability solutions. He has co-authored several IBM Redbooks publications. Antonio holds a degree in Computer Science from University of Udine, Italy.

Carlton Beatty is an Senior Staff Subject Matter Expert working with the IBM Spectrum Virtualize and IBM FlashSystem® Storage Family. He works under the IBM Systems Group primarily supporting and providing education for storage across the Americas Group Region. He joined IBM in 2015 and has more than 15 years of experience with IT, infrastructure, networking, DevOps, systems management, and Dev support. Carlton graduated from the Georgia Institute of Technology in 2013 with a Bachelors in Computer Engineering. His areas of expertise include storage systems metrics monitoring and cluster performance data analytics.

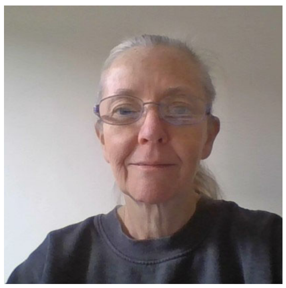
David Green works with the IBM SAN Central team troubleshooting performance and other problems on storage networks. He has authored, or contributed to, several IBM Redbooks publications. He is a regular speaker at IBM Technical University. You can find his blog at Inside IBM Storage Networking where he writes about all things related to Storage Networking and IBM Storage Insights.

Hartmut Lonzer is the IBM FlashSystem Territory Account Manager and SAN Offering Manager in DACH. Before this position, he was OEM Alliance Manager for Lenovo in IBM Germany. He works at the IBM Germany headquarters in Ehningen. His main focus is on the IBM FlashSystem Family and the IBM SAN Volume Controller. His experience with the IBM SAN Volume Controller and IBM FlashSystem (formerly IBM Storwize®) products goes back to the beginning of these products. Hartmut has been with IBM in various technical and sales roles for 43 years.

Jonathan Wilkie is an Advanced Subject Matter Expert/L3 support representative for IBM Spectrum Virtualize and IBM FlashSystem. He has more than 20 years of experience in IBM storage technical support. Over his career, he has provided technical support for Shark, DS4000, DS6000, and IBM DS8000® products. He has been supporting IBM Spectrum Virtualize-based products since 2010.

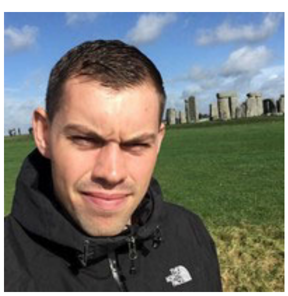
Kendall Williams is a SME and Project Field Engineer working with the IBM Spectrum Virtualize Storage Support Family. He holds a Bachelor of Science degree in Information Technology, with a concentration in DB Management Systems and Architecture from Florida State University. His areas of expertise include complex client performance analysis and copy services support for production environments. Kendall joined IBM in 2012, and has since become an advocate for some premier IBM clients.

Konrad Trojok has been responsible for the technical team lead for the IBM Storage team at SVA for the last 9 years. The role includes an active part in the daily IBM storage business, such as design, implementation, and care of storage solutions. It also includes strategic advisory regarding storage solutions. The beginning of his IT career included IBM Power solutions around SP systems and SSA storage. Konrad switched his technical focus with the emerging of SAN and SAN storage.



Mandy Stevens is a Technical Advisor for Storage, covering various IBM storage products with a focus on IBM FlashSystems. Mandy has been with IBM for 32 years and in that period has held a variety of positions within storage systems. She has been a chip designer for ASICs that are used in these subsystems, a tester, RAS team lead, and a first-line manager, all as part of the Hursley storage systems development team. She currently supports customers in UKI and Sweden.

Nezih Boyacıoglu has 20 years of experience as an SAN Storage specialist and currently works for IBM Premier Business Partner Istanbul Pazarlama in Turkey. He has over 20 years in the IT arena. His IBM storage experience started with Tivoli® Storage Manager (Spectrum Protect) and tape systems and his main focus for last 10 years has been on IBM Spectrum Virtualize family (IBM SAN Volume Controller, Storwize, FlashSystem) and storage area networks. He is an IBM Certified Specialist for Enterprise Storage Technical Support, Flash Technical Solutions, Virtualized Storage, and Spectrum Storage software.

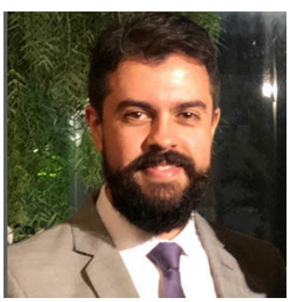


Nils Olsson is a Subject Matter Expert (previously Product Engineering Professional - Level 2 Support) working in the EMEA Storage Competence Center (ESCC) in Kelsterbach, IBM Germany. He provides remote technical support for Spectrum Virtualize Products (IBM SAN Volume Controller, Storwize, and FlashSystem) in Europe and worldwide.
Nils joined IBM in 2008 and is skilled in SAN, storage, and storage virtualization. He has over 10 years of experience with troubleshooting IBM Spectrum Virtualize and is certified on the IBM Storwize portfolio. In his current role, he analyzes complex field issues and provides technical expertise during critical situations, collaborating closely with development.

Renato Santos is a senior storage Technical Advisor for the USA Southeast team and is the SME for IBM A9000/R and XIV® Systems. As a formed Team Leader for Latin America Technical Advisor group, he advises customers on a portfolio of all IBM storage products. He started at IBM Brazil in 1995 and has been an IBM Level 1 Storage Support Specialist, XIV certified administrator, IBM ProtecTIER® certified Level 3 Support Specialist, Certified Flash System Administrator, and Storage Customer Support Manager. Renato has been with IBM for 26 years and is a certified IBM Customer Advocate and holds an MBA in Executive Management from Fundação Getulio Vargas (FGV) – Rio de Janeiro Brazil. and a bachelor's degree in Computer Science Rio de Janeiro, Brazil.

Rene Oehme is an IT Specialist working in Advanced Technical Skills (ATS) for the IBM Systems. He is located in the EMEA Storage Competence Center (ESCC) in Kelsterbach, Germany. René has 20 years of experience in IT support, performing various roles in Onsite and Remote Technical Support focusing on disk storage, virtualization solutions, SAN and open systems, and mainframe infrastructure. He holds a degree in Information Technology.

Sergey Kubin is an Advanced Subject Matter Expert (ASME) for IBM Spectrum Virtualize support team. He holds an Electronics Engineer degree from Ural Federal University in Russia and has more than 15 years of experience in IT. In IBM, he provides support and guidance for customers in Europe, Middle East and Russia. His expertise includes file and block storage, and storage are networks. He is an IBM Certified Specialist for FlashSystem Family Technical Solutions.



Thales Noivo Ferreira is a SAN Admin working for Kyndryl in Brazil. He has extensive experience with Netapp, IBM Storwize Family, DS8K Family, XiV, Cisco Switches and Brocade Switches. Currently, he is working on environments with IBM SAN Volume Controller and FlashSystem family.

Uwe Schreiber is a Solution Architect and System Engineer at SVA System Vertrieb Alexander GmbH. He has been working with IBM Spectrum Virtualize and IBM SAN Volume Controller since 2002 (until 2011 as customer, since 2012 as Business Partner employee). Uwe is an experienced professional providing technical pre- and post-sales solutions for IBM server and storage systems since 1995. He holds an engineering diploma in Computer Sciences from the University of Applied Science in Darmstadt, Germany.

Vasfi Gucer is an IBM Technical Content Services Project Leader with IBM Garage™ for Systems. He has more than 25 years of experience in the areas of systems management, networking hardware, and software. He writes extensively and teaches IBM classes worldwide about IBM products. His focus has been primarily on storage, cloud computing, and cloud storage technologies for the last 8 years. Vasfi is also an IBM Certified Senior IT Specialist, Project Management Professional (PMP), IT Infrastructure Library (ITIL) V2 Manager, and ITIL V3 Expert.

Thanks to the following for their contributions that made this book possible:

Evelyn Perez, Suri Polisetti
IBM Hursley, UK

Barry Whyte
IBM Australia

Angelo Bernasconi
IBM Italy

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 IBM Redbooks® residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

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

`ibm.com/redbooks/residencies.html`

Comments welcome

Your comments are important to us!

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

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

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, IBM Redbooks
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- 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

1

Introduction to IBM FlashSystem

This chapter introduces the IBM FlashSystem storage subsystem range that is supported by the new IBM Spectrum Virtualize software v8.4. It describes all of the relevant models, their key features, benefits, and technology.

This chapter includes the following topics:

1.1 IBM FlashSystem supported product range

This section describes the IBM FlashSystem products that are supported by the new IBM Spectrum Virtualize software v8.4 (see Figure 1-1).

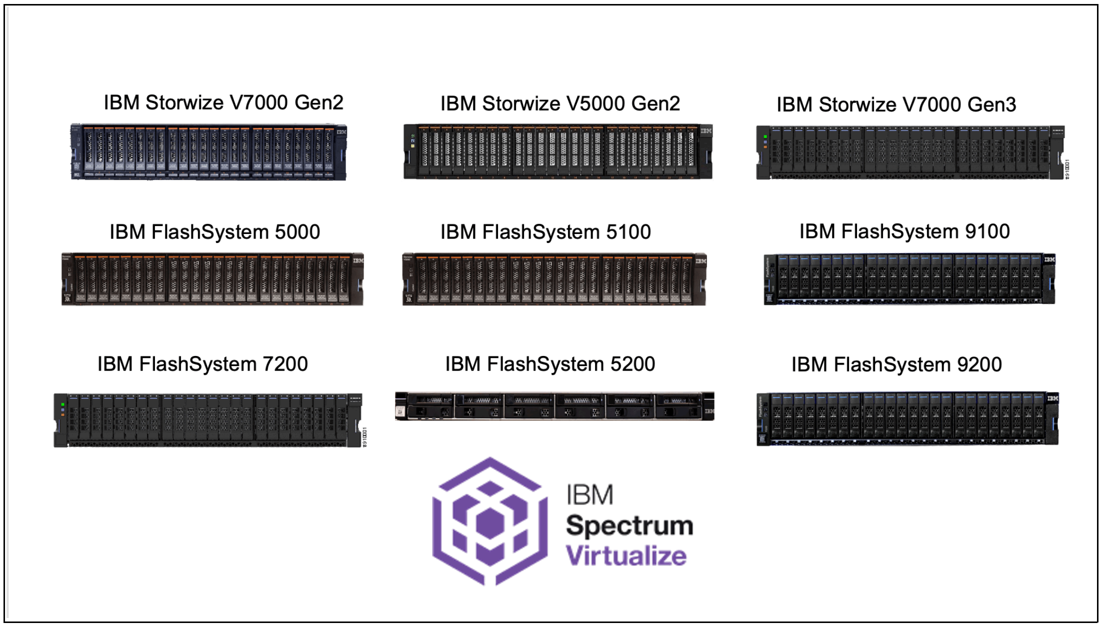


Figure 1-1 Products that support IBM Spectrum Virtualize software v8.4

Note: The IBM Storwize models that are shown in Figure 1-1 support IBM Spectrum Virtualize V 8.4.0.x, but not 8.4.1.x or 8.4.2.x.

1.1.1 New in V8.4

IBM Spectrum Virtualize 8.4 provides more features and updates to the IBM Spectrum Virtualize family of products of which IBM FlashSystem is part.

Software changes in version 8.4.2

The following software changes are featured in IBM Spectrum Virtualize version 8.4.2:

- Support for:
  - Increased number of volumes on the system:
    - Max volumes changed from 10.000 to 15.864.
    - Only applies to systems that have 10 K volumes (for example, V7000, FS7200, FS9100, FS9200, and IBM SAN Volume Controller).
  - Expanding and shrinking volumes in FlashCopy mappings:

    Volumes that are associated with User-Defined FlashCopy mappings can now be expanded. The process for expanding such volumes includes the following basic rules:
    - Source or Target volume can be expanded at any time.
    - For incremental FlashCopy maps, the Target VDisk must be expanded before the Source volume can be expanded.

    - Source and Target must be same size when mapping is prepared or started.
    - Source can be shrunk, but only to the size of the largest copying-or-stopping Target.
    - Target volume cannot be shrunk.
  - Safeguarded Copy function

    IBM FlashSystem Safeguarded Copy feature prevents point-in-time copies of data from being modified or deleted because of user errors, malicious destruction, or ransomware attacks.
  - Multiple IP partnerships + multiple IP addresses and VLANs:
    - Enhanced Spectrum Virtualize Ethernet support with more than 1 IPv4 and 1 IPv6 address to be defined per port for use by Ethernet host attach protocols, such as iSCSI, iSER, and NVMeF (in the future).
    - VLAN separation for individual IP address or as wanted.
    - New Portset based configuration model for Ethernet and IP connectivity.
    - For iSCSI, iSERHost Attach, and IP Replication. Extensible to NVMeF and Fibre Channel in the future.
    - OBAC based per-tenant administration and partitioning model for multi tenant cloud environments.
    - New CLI model for Ethernet network configuration.
  - Non-disruptive system migration:
    - Nondisruptive volume migration between independent clusters. Enables nondisruptive migration between non-clustering platforms for example, IBM SAN Volume Controller to FS9200.
    - Can migrate volumes away from a cluster that is, for example, reaching max limits.
    - Uses enhancements to SCSI (ALUA) path states. Migration is based upon Remote Copy (Metro Mirror) functions.
  - Throttling on child pools

    Up to the version 8.4.2.0, creating a throttle for a child pool was blocked. Now, the throttle can be created. As in other throttling types, an I/O obeys the most restrictive throttling that applies to it, and each I/O is counted against all the throttling that applies to it. That is, the throttling is hierarchical: An I/O to a VDisk in a child pool counts against parent pool and child pool throttle.
  - Downloading code through eSupport

    Allows the migration of the code download function to use the Call Home Using Rest API (CHURA) infrastructure, with the addition of the optional ability to use a HTTP proxy to download packages from esupport.ibm.com. The main value here is the ability to download selected code bundles (including prerequisites, drive firmware, or ifixes) without the need for the firewall hole to FixCentral and more importantly, by way of a HTTP proxy.
- RESTful API improvements.
- Improved web forwarding with remote support assistance.
- Improved HyperSwap scalability

  A HyperSwap volume can now be expanded by using the expand volume command when the volume copies are part of user-defined FlashCopy® mappings.
- Support for 32GB Cavium Fibre Channel adapter.

Version 8.4 major software changes

The following software changes were made in Spectrum Virtualize version 8.4:

- Data Reduction Pool (DRP) improvements:
  - A DRP allows for more flexibility, such as multi-tenancy.
  - FlashCopy with redirect-on-write support: Uses DRP's internal deduplication referencing capabilities to reduce overhead by creating references instead of copying the data. Redirect-on-write (RoW) is an alternative to the copy-on-write (CoW) capabilities.

    > **Note:** At the time of writing, this capability might be used only for volumes with supported deduplication without mirroring relationships and within the same pool and I/O group. The mode selection (RoW/CoW) is automatic based on these conditions.

  - Compresstimator always on; This improvement allows the systems to sample each volume at regular intervals, providing the ability to display the compressibility of the data in the GUI and IBM Storage Insights at any time.
  - RAID Reconstruct Read; This improvement increases reliability and availability by reducing chances of DRP going offline because of fixable array issues, leveraging RAID capabilities, DRP asks for a specific data block reconstruction when detecting a potential corruption.
- Distributed RAID 1 (DRAID 1) support provides the ability to extend distributed RAID advantages to smaller pools of drives. This improves performance over traditional RAID 1 implementations, allowing a better use of flash technology. These distributed arrays can support as few as two drives, with no rebuild area, and 3 - 16 drives, with a single rebuild area.

  > **Note:** At the time of writing, DRAID 1 is supported on IBM FlashSystems 5015, 5035, 5200, 7200, and 9200 only. It is *not* available for FlashCore Modules (FCM-XL) of 38.4 TB.

- With 8.4 FlashSystem 5100, 5200, 7200, and 9200, systems can support up to 12 Storage Class Memory (SCM) devices per enclosure with no slot restrictions. Previously, the limit for all SCM drives was four per enclosure in drive slots 21 - 24.

  > **Note:** With 8.4 code FlashSystem 5100, 5200, 7200, and 9200, systems can now support up to 12 zSSD SCM drives or Optane SCM drives.

- Expansion of mirrored volumes (also known as vDisks) allows the volumes capacity to be expanded or reduced online, without requiring an offline format and sync. This improves the availability of the volume because the new capacity is available immediately.
- Three-site replication with HyperSwap® support provides improved availability for data in three-site implementations. This improvement expands on the disaster recovery capabilities that are inherent in this topology.
- Host attachment support with FC-NVMe in HyperSwap systems.
- DNS support for LDAP and NTP with full DNS length (that is, 256 characters).
- Updates to maximum configuration limits. This doubles FlashCopy mapping from 5,000 to10,000 and increases HyperSwap volumes limit from 1,250 to 2,000.

- Password and login changes on the IBM FlashSystem v8.4 GUI to meet today's extra regulatory compliance with expiry and security enhancements.
- Support for internal proxy servers (also known as customer web proxy) uses IBM Call Home with cloud services and log upload features.

1.1.2 Supported products

In this section, the IBM FlashSystem and IBM Storwize products are supported to run the IBM Spectrum Virtualize software v8.4 software are described. Listed are the IBM Storwize and IBM FlashSystem series name and then, the hardware machine type and model for extra clarity.

Storwize products

The following Storwize products are supported:

- IBM Storwize V5100:
  - 2077-AF4
  - 2077-424
  - 2077-U5B
  - 2078-AF4
  - 2078-424
  - 2078-U5B
- IBM Storwize V7000:
  - 2076-724
  - 2076-U7B
  - 2076-U7A
  - 2076-AF6
  - 2076-624
  - 2076-524

> **Note:** Models 2076-AF6, 2076-624, and 2076-524 support 8.4.0.x, but not 8.4.1.x or 8.4.2.x.

FlashSystem products

The following FlashSystem products are supported:

- IBM FlashSystem 5000:
  - IBM FlashSystem 5010 and IBM FlashSystem 5030 (formerly known as IBM Storwize V5010E and Storwize V5030E):
    - 2072-2H2
    - 2072-U12
    - 2072-2H4
    - 2072-U24
      2072-3H2
    - 2072-V12
    - 2072-3H4
    - 2072-V24
  - IBM FlashSystem 5015 and IBM FlashSystem 5035:
    - 2072-2N2
    - 2072-U12

- 2072-2N4
- 2072-U24
- 2072-3N2
- 2072-V12
- 2072-3N4
- 2072-V24

- IBM FlashSystem 5100:
  - 2077-4H4
  - 2078-4H4
  - 2078-UHB
- IBM FlashSystem 5200:
  - 4662-6H2
  - 4662-UH6
- IBM FlashSystem 7200:
  - 2076-824
  - 2076-U7C
  - 4664-824
  - 4664-U7C
- IBM FlashSystem 9100:
  - 9846-AF7
  - 9848-AF7
  - 9848-UF7
  - 9846-AF8
  - 9848-AF8
  - 9848-UF8
- IBM FlashSystem 9200:
  - 9846-AG8
  - 9848-AG8
  - 9848-UG8
  - 4666-AG8
  - 4666-UG8

1.2 IBM FlashSystem high-level features

This IBM Redbooks publication describes and focuses on the best practices and options to gain the optimum performance from the product, including the set of software-defined storage features.

It also describes data-reduction techniques, including deduplication, compression, dynamic tiering, thin provisioning, snapshots, cloning, replication, data copy services, Safeguarded Copy and IBM HyperSwap for high availability.

Note: The detailed technical explanations, and theory of operations, of these features are not covered in this publication. For more information about this area, see the following IBM Redbooks publications:

- *Implementing the IBM FlashSystem with IBM Spectrum Virtualize Version 8.4.2*, SG24-8506
- *Implementing the IBM SAN Volume Controller with IBM Spectrum Virtualize Version 8.4.2*, SG24-8507
- *IBM FlashSystem 9200 Product Guide*, REDP-5586
- *IBM FlashSystem 9100 Product Guide*, REDP-5524
- *IBM FlashSystem 7200 Product Guide*, REDP-5587
- *IBM FlashSystem 5200 Product Guide,* REDP-5617
- *IBM FlashSystem 5000 and 5100 for Mid-Market*, REDP-5594
- *IBM Flashsystem 5000 and 5200 for Mid-Market,* REDP-5630
- *IBM Spectrum Virtualize HyperSwap SAN Implementation and Design Best Practices*, REDP-5597
- *IBM DS8870 Easy Tier Heat Map Transfer*, REDP-5015

The following two types of enclosures are part of the IBM FlashSystem products that run Spectrum Virtualize:

- A *control enclosure* manages your storage systems, communicates with the host, and manages interfaces. In addition, it can also house up to 24 drives. These drives can be Storage Class Memory (SCM), industry-standard NVMe type, the exclusive IBM NVMe FlashCore Modules (FCM), standard flash (SSD) Serial Attached SCSI (SAS) type drives, or hard disk drives (HDD), depending on which model of control enclosure is ordered.

  Each control enclosure is either a standard 2U high, or 1U high for the IBM FlashSystem 5200, 19”rack-mounted unit.

- An *expansion enclosure* enables you to increase the available capacity of the IBM FlashSystem cluster communicates with the control enclosure by way of a pair of 12 Gbps SAS connections. These expansion enclosures can house many flash (SSD) SAS type drives or hard disk drives (HDD), depending on which model of expansion enclosure is ordered.

  Expansion enclosures are generally of the three following types:

  – Large form factor (LFF). Figure 1-2 shows the LFF expansion enclosure that can hold 12 3.5-inch drives and is 2U high.

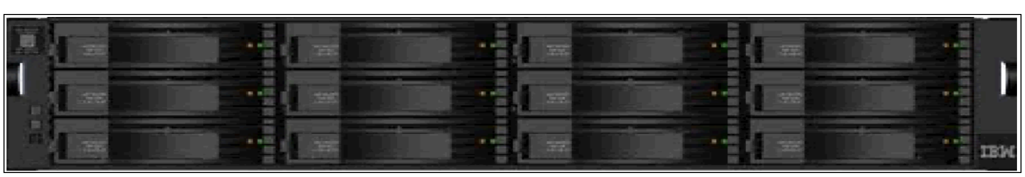


Figure 1-2 LFF expansion enclosure

– Small form factor (SFF). Figure 1-3 shows the SFF expansion enclosure that can hold 24 2.5-inch drives and is 2U high.

Figure 1-3 SFF expansion enclosure

– Large form factor high density (LFF HD). Figure 1-4 shows the large form factor high density (LFF HD) expansion enclosure that can hold 92 3.5-inch drives (or 92 2.5-inch drives in carriers) and is 5U high.

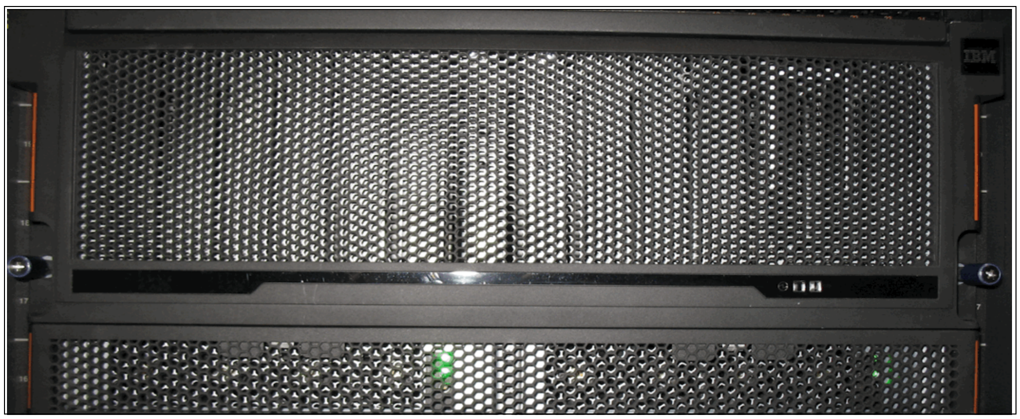


Figure 1-4 LFF HD expansion enclosure

The type and models of expansion enclosures that can attach to the relevant control enclosure is model-dependent. For more information, see 1.3, “IBM Storwize and IBM FlashSystem product range”.

1.3 IBM Storwize and IBM FlashSystem product range

In this section, we describe the various IBM Storwize and IBM FlashSystem products that are supported in IBM Spectrum Virtualize software v8.4. It includes in-depth information about each product, its capabilities, features, and functions.

Also supplied for each product range are links to information about the configuration limits and restrictions. The customer can research information or values that are needed for optimum performance and adhere to the best practices.

Note: Spectrum Virtualize V 8.4 or later does not support Storwize V5015, V5020, or V5030 systems.

IBM Storwize V5100

The IBM Storwize 5100 is a virtualized, software-defined storage system that is composed of hardware components and a requisite licensed software product, IBM Spectrum Virtualize Software. All functional capabilities for the IBM Storwize 5100 are provided through IBM Spectrum Virtualize software v8.4.

Figure 1-5 shows the front view of the IBM Storwize 5100 control enclosure.

Figure 1-5 IBM Storwize V5100 front view

IBM IBM Storwize 5100 models AF 4 and 424 are designed to meet modern high-performance storage requirements, including ultra-low latency, cost-effectiveness, operational efficiency, and mission-critical reliability. It is built on a flash-optimized design, with an end-to-end NVMe strategy to bring extremely low latencies to organizations of all sizes:

- IBM Storwize 5100 model AF4
  - All-flash storage system that supports Storage Class Memory drives (up to 12), NVMe FlashCore Modules, and industry-standard NVMe flash drives in the control enclosure.
  - Model AF4 attaches to expansion enclosure models AFF and A9F that support SAS Flash drives.
- IBM Storwize 5100 model 424
  - Hybrid storage system that supports Storage Class Memory drives (up to 12), NVMe FlashCore Modules, and industry-standard flash NVMe drives in the control enclosure.
  - Model 424 attaches to expansion enclosure models 12F, 24F, and 92F that support SAS Flash drives and SAS HDD Drives.

- IBM Storwize 5100 model U5B

  IBM Storwize 5100 hardware component to be used in the Storage Utility Offering space. It is physically and functionally identical to V5100 models 424 and AF4 except for target configurations and variable capacity billing. The variable capacity billing uses IBM Storage Insights to monitor the system usage, which enables allocated storage use that exceeds a base subscription rate to be billed per terabyte, per month.

Machine types 2077 and 2078 expansion enclosure models 12F, 24F, and 92F can be attached to a IBM Storwize 5100 model 424. Models AFF and A9F can be attached to a IBM Storwize 5100 model AF4.

IBM Storwize 5100 systems can be clustered with another IBM Storwize 5100. All systems within a cluster must use the same version of IBM Storwize 5100 software.

For a comprehensive list of supported environments, devices, and configurations, see IBM System Storage Interoperation Center (SSIC).

Note: Consider the following points:

- The IBM Storwize 5100 systems are withdrawn from Marketing (WFM) since April 2020, and as such are no longer available to purchase from IBM. They are included in this book for completeness because they support running the Spectrum Virtualize software v8.4.
- The IBM Storwize V5100 supports Safeguarded Copy (the FlashCopy License is required).

Table 1-1 lists the IBM IBM Storwize 5100 host, drive capacity, and functions.

Table 1-1 IBM Storwize 5100 host, drive capacity, and functions summary

Feature/Function	Description
Host interface	► 10 Gbps Ethernet (iSCSI) ► 25 Gbps Ethernet (iSCSI, iWARP, RoCE) ► 16 Gbps Fibre Channel (FC, FC-NVMe) ► 32 Gbps Fibre Channel (FC-NVMe
Control Enclosure Supported drives	► 2.5-inch NVMe FCMs: 4.8 TB, 9.6 TB, and 19.2 TB compressing FCMs ► 2.5-inch NVMe flash drives: 800 GB, 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB

Feature/Function	Description
SAS expansion enclosures Maximum Drives: 760 per control enclosure; 1,520 per clustered system	► Control enclosure model 424 supports the following expansions: – Model 12F/24F 2U 12 or 24 drive – Model 92F 5U 92 drive – 2.5-inch Storage Class Memory: 375 GB, 750 GB, 800GB and 1.6 TB – 2.5-inch flash drives supported: 800 GB, 1.6 TB, 1.92 TB, 3.84 TB, 7.68 TB, 15.36 TB, and 30.72 TB – 2.5-inch disk drives supported: • 600 GB and 900 GB 15k SAS drive • 600 GB, 900 GB, 1.2 TB, 1.8 TB, and 2.4 TB 10k SAS disk • 2 TB 7.2k nearline SAS disk – 3.5-inch disk drives supported: 4 TB, 6 TB, 8 TB, 10 TB, 12 TB, and 14 TB 7.2k nearline SAS disk ► Control enclosure model AF4 supports the following expansions: – Model AFF 2U 24 drive – Model A9F 5U 92 drive – 2.5-inch Storage Class Memory: 375 GB, 750 GB, 800GB and 1.6 TB – 2.5-inch flash drives supported: 800 GB, 1.6 TB, 1.92 TB, 3.84 TB, 7.68 TB, 15.36 TB, and 30.72 TB
RAID levels	► DRAID 5 (CLI-only) and 6, TRAID 10
Advanced features included with each system	► Virtualization of internal storage ► Data migration ► DRPs with thin provisioning ► UNMAP ► Compression and deduplication ► Metro Mirror (synchronous) and Global Mirror (asynchronous)
Additional available advanced features	► Remote mirroring ► Easy Tier compression ► External virtualization ► Encryption ► FlashCopy ► IBM Spectrum Control ► IBM Spectrum Protect Snapshot

For more information, see this IBM Support web page.

IBM Storwize V7000

IBM Storwize V7000 is a virtualized storage system to complement virtualized server environments that provides unmatched performance, availability, advanced functions, and highly scalable capacity never seen before in midrange disk systems.

IBM Storwize V7000 is a powerful midrange disk system that was designed to be easy to use and enable rapid deployment without extra resources. IBM Storwize V7000 is virtual storage that offers greater efficiency and flexibility through built-in SSD optimization and thin provisioning technologies.

The following three generations of systems in the IBM Storwize V7000 are supported by Spectrum Virtualize software v8.4.0:

- IBM Storwize V7000 Generation 2 (2076-524)
- IBM Storwize V7000 Generation 2+ (2076-624 and 2076-U7A)
- IBM Storwize V7000 Generation 3 (2075-724 and 2076 -U7B)

Important: Consider the following points:

- The V7000 systems are Withdrawn from Marketing (WFM) since October 2020, and no longer are available to purchase from IBM. They are included in this book for completeness because they support running the Spectrum Virtualize software v8.4.
- Spectrum Virtualize software v8.4.2 supports only IBM Storwize V7000 Generation 3 (2075-724 and 2076 -U7B.

  V8.4.2, or later, does not support V7000 Generation 2 or Generation 2+ systems. Customers with V7000 Generation 2 or Generation 2+ I/O groups cannot upgrade to v8.4.2.x or later.
- IBM Storwize V7000 Generation 3 (2075-724 and 2076 -U7B) supports Safeguarded Copy function, which requires a FlashCopy license.

Figure 1-6 shows the IBM Storwize V7000 Generation 2 (2076-524) and IBM V7000 Generation 2+ (2076-624) front view.

Figure 1-6 IBM Storwize V7000 Generation 2 (2076-524) and Generation2+ (2076-624) SFF

Figure 1-7 shows the IBM Storwize V7000 Generation 3 (2076-724) front view.

Figure 1-7 IBM Storwize V7000 Generation 3 (2076-724)

Figure 1-8 shows the rear view of the IBM Storwize V7000 Generation 3 (2076-724).

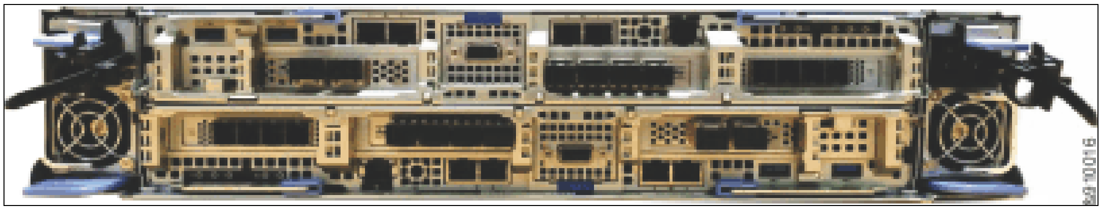


Figure 1-8 Rear view of the IBM Storwize V7000 Generation 3 (2076-724)

The IBM 2076 SFF Control Enclosure Model 524 and 624 includes the following features:

- Two node canisters and up to 256 GB cache (system total) in a 2U, 19-inch rack mount enclosure.
- 1 Gb iSCSI connectivity is standard, with options for 16 Gb FC, 10 Gb iSCSI/FCoE, and 25 Gb iSCSI connectivity.
- It holds up to twenty-four 2.5-inch SAS flash drives and supports the attachment of up to 20 Storwize V7000 expansion enclosures.

The IBM 2076 Model 724 SFF NVMe Control Enclosure includes the following features:

- Two node canisters and up to 1 TB cache (system total) in a 2U, 19-inch rack mount enclosure.
- 10 Gb iSCSI connectivity is standard, with options for 32 Gb FC, 16 Gb FC, and 25 Gb iSCSI connectivity.
- It holds up to twenty-four 2.5-inch NVMe FlashCore Modules or industry-standard flash drives and supports the attachment of up to 20 Storwize V7000 expansion enclosures.

The IBM 2076 Models U7A and U7B are the Storwize V7000 hardware components to be utilized in the Storage Utility Offering space. They are physically and functionally identical to the V7000 model 624 and 724 respectively, except for target configurations and variable capacity billing. The variable capacity billing uses IBM Spectrum Control Storage Insights to monitor the system usage, allowing allocated storage usage that exceeds a base subscription rate to be billed per TB, per month.

Table 1-2 lists the host connections, drive capacities, features, and standard options that are available with Spectrum Virtualize that are available on the IBM Storwize V7000.

Table 1-2 IBM Storwize V7000 host, drive capacity and functions summary

Feature/Function	Description
Host interface	► Gen 2 and Gen 2+ 1 Gb iSCSI connectivity is standard, with options for 16 Gb FC, 10 Gb iSCSI/FCoE, and 25 Gb iSCSI connectivity ► Gen 3 10 Gb iSCSI connectivity is standard, with options for 16 Gb FC and 25 Gb iSCSI connectivity
Control enclosure supported drives	► Gen 2 / Gen2+ – NL SAS 7.2K RPM: 2 TB, 4 TB, 6 TB, 8 TB, and 10 TB – Flash (SSD) drives: 400 GB, 800 GB, 1.6 TB, 1.92 TB, 3.2 TB, 3.84 TB, 7.68 TB, and 15.36 TB – Enterprise class disk drives: • 15K RPM - 300 GB, 600 GB, and 900 GB 15,000 rpm • 10K RPM _ 900 GB, 1.2 TB, 1.8 TB, and 2.4 TB ► Gen 3: – 2.5-inch SFF NVMe FlashCore Modules (FCM): • 4.8 TB, 9.6 TB, and 19.2 TB • 2.5-inch SFF NVMe industry-standard drives • 800 GB 2.5-inch 3DWPD – NVMe flash drive: 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB
SAS expansion enclosures: ► Storwize V7000 LFF Expansion Enclosure Model 12F: 12 slots for 3.5-inch SAS drives ► Storwize V7000 SFF Expansion Enclosure Model 24F: 24 slots for 2.5-inch SAS drives ► Storwize V7000 HD LFF Expansion Enclosure Model 92F: 92 slots for SAS drives in a 3.5-inch carrier	► NL SAS 7.2K RPM – 2 TB, 4 TB, 6 TB, 8 TB, and 10 TB ► Flash (SSD) drives: 400 GB, 800 GB, 1.6 TB, 1.92 TB, 3.2 TB, 3.84 TB, 7.68 TB, and 15.36 TB ► Enterprise class disk drives: – 15K RPM - 300 GB, 600 GB, and 900 GB 15,000 rpm – 10K RPM _ 900 GB, 1.2 TB, 1.8 TB and 2.4 TB
RAID levels	► Gen 2 and 2+: RAID 0, 1, 5, 6, and 10 ► Gen 3: – DRAID5, DRAID6, TRAID0, TRAID1, and TRAID10. – For compressed drives only, DRAID5 and DRAID6 are supported.
Advanced features included with each system	► IBM System Storage Easy Tier ► IBM FlashCopy ► Thin provisioning
Other available advanced features	► Remote Mirroring ► External Virtualization ► IBM FlashCopy Manager

Important: V8.4.2, or later, does not support V7000 Generation 2 or Generation 2+ systems. Customers with V7000 Generation 2 or Generation 2+ I/O groups *cannot* upgrade to v8.4.2.x or later.

For more information about the V8.4.2.x Configuration Limits and Restrictions for IBM Storwize V7000 Generation 2 and 2+ and IBM Storwize V7000 Generation 3, see this IBM Support web page.

IBM FlashSystem 5000

The IBM FlashSystem 5000 is a member of the IBM FlashSystem family of storage solutions. The IBM FlashSystem 5000 delivers increased performance and new levels of storage efficiency with superior ease of use. This entry storage solution enables organizations to overcome their storage challenges.

The solution includes technologies to complement and enhance virtual environments, which deliver a simpler, more scalable, and cost-efficient IT infrastructure. The IBM FlashSystem 5000 features two node canisters in a compact, 2U 19-inch rack mount enclosure.

Note: At the time of writing, the IBM FlashSystem 5010 and IBM FlashSystem 5030 are End of Marketing (EOM) and have been replaced by the IBM FlashSystem 5015 and IBM FlashSystem 5035 respectively. The IBM FlashSystem 5015/5035 offer superior CPU power and memory options, but the features and functions remain the same. The IBM FlashSystem 5015/5035 charts are included only as a reference.

The new IBM FlashSystem 5015 and IBM FlashSystem 5035 are similar to the older 5010 and 5030 models, but with higher-specification CPU and memory options. The new models also include all-flash and hybrid-flash solutions that provide enterprise-grade functions without compromising afford ability or performance, and are built with the rich features of IBM Spectrum Virtualize 8.4. The IBM FlashSystem 5000 helps make modern technologies, such as artificial intelligence, accessible to enterprises of all sizes.

IBM FlashSystem 5015

IBM FlashSystem 5015 is an entry-level solution that is focused on afford ability and ease of deployment and operation, with powerful scale-up features. It includes many IBM Spectrum Virtualize features and offers multiple flash and disk drive storage media and expansion options.

Figure 1-9 shows IBM FlashSystem 5015 and 5035 SFF control enclosure front view.

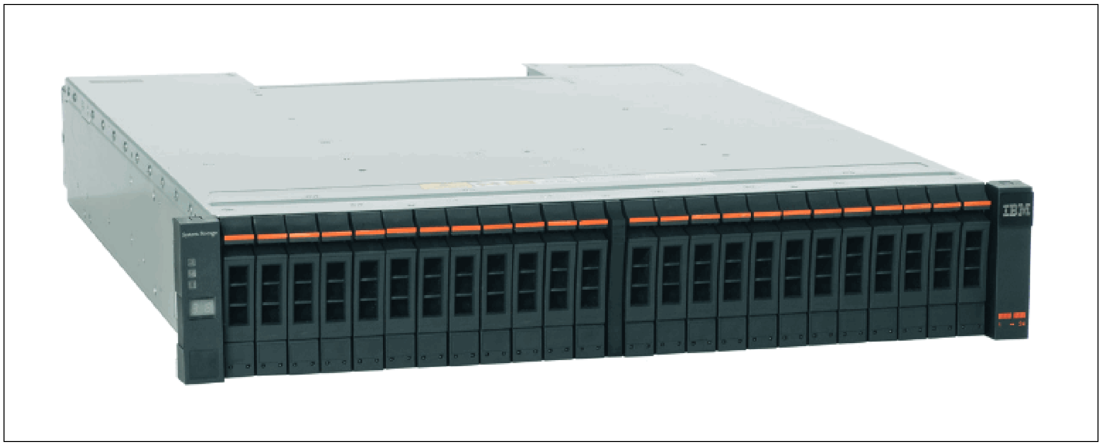


Figure 1-9 IBM FlashSystem 5015 and 5035 SFF control enclosure front view

Figure 1-10 shows IBM FlashSystem 5015 and 5035 LFF control enclosure front view.

Figure 1-10 IBM FlashSystem 5015 and 5035 LFF control enclosure front view

This next section provides hardware information about the IBM FlashSystem 5000 models and the feature set of each one.

Table 1-3 lists the model comparison chart for the IBM FlashSystem 5000 range.

Table 1-3 Machine type and model comparison for the IBM FlashSystem 5000

MTM	Full name
2072-2N2	IBM FlashSystem 5015 LFF Control Enclosure
2072-2N4	IBM FlashSystem 5015 SFF Control Enclosure
2072-3N2	IBM FlashSystem 5035 LFF Control Enclosure
2072-3N4	IBM FlashSystem 5035 SFF Control Enclosure
2072-12G/F12	IBM FlashSystem 5000 LFF Expansion Enclosure
2072-24G/F24	IBM FlashSystem 5000 SFF Expansion Enclosure
2072-92G/F92	IBM FlashSystem 5000 High-Density LFF Expansion Enclosure

Table 1-4 lists the host connections, drive capacities, features, and standard options with Spectrum Virtualize that are available on the IBM FlashSystem 5015.

Table 1-4 IBM FlashSystem 5015 host, drive capacity and functions summary

Feature / Function	Description
Host interface	► 1 Gb iSCSI (On the motherboard ► 16 Gbps Fibre Channel ► 12 Gbps SAS ► 25 Gbps iSCSI (iWARP or RoCE) ► 10 Gbps iSCSI
Control Enclosure and SAS expansion enclosures supported drives	► For SFF enclosures, see Table 1-5 on page 17 ► For LFF enclosures, see Table 1-6 on page 17
Cache per control enclosure/ clustered system	32 GB or 64 GB
RAID levels	distributed DRAID 1, 5 and 6
Maximum expansion enclosure capacity	► Up to 10 standard expansion enclosures per controller ► Up to 4 high-density expansion enclosures per controller
Advanced functions included with each system	► Virtualization of internal storage ► DRPs with thin provisioning and UNMAP ► One-way data migration
Additional available advanced features	► Easy Tier ► FlashCopy ► Remote mirroring

Table 1-5 lists the capacity for the 2.5-inch supported drives for the IBM FlashSystem 5000.

Table 1-5 2.5 inch supported drives for the IBM FlashSystem 5000

2.5-inch (SFF)	Capacity					
Tier 1 Flash	800 GB	1.9 TB	3.84 TB	7.68 TB	15.36 TB	30.72 TB
High Performance, Enterprise Disk Drives (10k rpm)	900 GB	1.2 TB	1.8 TB	2.4 TB		
High Capacity Nearline Disk Drives (7.2k rpm)	2 TB					

Table 1-6 lists the speed and capacity for the 3.5-inch supported drives for the IBM FlashSystem 5000.

Table 1-6 3.5 inch supported drives for the IBM FlashSystem 5000

3.5-inch (LFF)	Speed	Capacity							
High-Performance, Enterprise class Disk Drives	10,000 RPM	900 GB	1.2 TB	1.8 TB	2.4 TB				
High Capacity, Archival class Nearline Disk Drives	7,200 RPM	4 TB	6 TB	8 TB	10 TB	12 TB	14 TB	16 TB	18 TB

IBM FlashSystem 5035

IBM FlashSystem 5035 provides greater functionality, including powerful encryption capabilities and DRPs with compression, deduplication, thin provisioning, and the ability to cluster for scale-up and scale-out.

Available with the IBM FlashSystem 5035 model, DRPs help transform the economics of data storage. When applied to new or existing storage, they can significantly increase usable capacity, while maintaining consistent application performance. This can help eliminate or drastically reduce costs for storage acquisition, rack space, power, and cooling, and can extend the useful life of existing storage assets. Capabilities include:

- Block deduplication that works across all the storage in a DRP to minimize the number of identical blocks
- New compression technology that provides guaranteed consistent 2:1 or better reduction performance across a wide range of application workload patterns
- SCSI UNMAP support that de-allocates physical storage when operating systems delete logical storage constructs such as files in a file system

Table 1-7 lists the host connections, drive capacities, features, and standard options with Spectrum Virtualize that are available on the IBM FlashSystem 5035.

Table 1-7 IBM FlashSystem 5035 host, drive capacity and functions summary

Feature / Function	Description
Host interface	► 10 Gb iSCSI (On the motherboard ► 16 Gbps Fibre Channel ► 12 Gbps SAS ► 25 Gbps iSCSI (iWARP or RoCE) ► 10 Gbps iSCSI
Control Enclosure and SAS expansion enclosures Supported drives	► For SFF enclosures see Table 1-5 ► For LFF enclosures see Table 1-6
Cache per control enclosure / clustered system	32 GB or 64 GB / 64 GB or 128 GB
RAID levels	Distributed DRAID 1, 5 (CLI Only), and 6
Maximum expansion enclosure capacity	► Up to 20 standard expansion enclosures per controller ► Up to eight high-density expansion enclosures per controller
Advanced functions included with each system	► Virtualization of internal storage ► DRPs with thin provisioning ► UNMAP, compression, and deduplication ► One-way data migration ► Dual-system clustering
Additional available advanced features	► Easy Tier ► FlashCopy ► Remote mirroring ► Encryption

For more information about the V8.4.2.x Configuration Limits and Restrictions for IBM FlashSystem 5015 and 5035, see this VIBM Support web page.

IBM FlashSystem 5100

The IBM FlashSystem 5100 supports the following enhanced features, capacities, and enclosures:

- NVMe-accelerated flash arrays with control enclosures that are 100 percent, end-to-end NVMe-enabled, and Storage Class Memory (SCM)-capable.
- The systems offer industry-leading performance and scalability and support physical, virtual, and Docker environments.
- Hybrid-flash array enabled with multiple expansion enclosure options based on 12 Gbs SAS that support both SSDs and hard disk drives.
- AI-enhanced with the IBM Storage Insights analytics, resource management, and support platform. Also, IBM Spectrum Virtualize functions include AI-based data placement for optimal data center performance and zero-downtime data migration.
- Hybrid-cloud ready and can support private, hybrid, or public cloud deployments. The solutions include ready-to-use, proven, validated "cloud blueprints" with support for cloud API automation, replication, and secondary data orchestration software.
- Safeguarded Copy function supports the ability to create cyber-resilient point-in-time copies of volumes that cannot be changed or deleted through user errors, malicious actions, or ransomware attacks. The system integrates with IBM Copy Services Manager to provide automated backup copies and data recovery. This integration requires FlashCopy license.

The IBM FlashSystem 5100 control enclosure supports up to 24 2.5" NVMe capable flash drives in a 2U high form factor.

The IBM FlashSystem 5100 has two standard models (2077-4H4 and 2078-4H4) and one utility model (2078-UHB).

Figure 1-11 shows the IBM FlashSystem 5100 control enclosure front view.

Figure 1-11 IBM FlashSystem 5100 front view showing the 24 NVMe drives installed

Table 1-8 lists the host connections, drive capacities, features, and standard options with Spectrum Virtualize that are available on the IBM FlashSystem 5100.

Table 1-8 IBM FlashSystem 5100 host, drive capacity, and functions summary

Feature/Function	Description
Host interface	► 10 Gbps Ethernet (iSCSI) ► 25 Gbps Ethernet (iSCSI, iSER - iWARP, RoCE) ► 16 Gbps Fibre Channel (FC, FC-NVMe) ► 32 Gbps Fibre Channel (FC, FC-NVMe)
Control enclosure supported drives	► 2.5-inch Storage Class Memory: 375 GB, 750 GB, 800GB and 1.6 TB ► 2.5-inch NVMe self-compressing FCM 2.5-inch: 4.8 TB, 9.6 TB, 19.2 TB, and 38.4 TB ► NVMe flash drives: 800 GB, 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB
SAS expansion enclosures 760 per control enclosure; 1520 per clustered system Model 12G 2U 12 drives Model 24G 2U 24 drives Model 92G 5U 92 drives	► 2.5-inch flash drives supported: 800 GB, 1.6 TB, 1.92 TB, 3.84 TB, 7.68 TB, 15.36 TB, and 30.72 TB ► 2.5-inch disk drives supported: – 600 GB, 900 GB, 1.2 TB, 1.8 TB, and 2.4 TB 10k SAS disk – 2 TB 7.2k nearline SAS disk ► 3.5-inch disk drives supported: 4 TB, 6 TB, 8 TB, 10 TB, 12 TB, 14 TB, 16 TB, and 18 TB 7.2k nearline SAS disk
RAID levels	Distributed RAID 5 and 6, TRAID 1 and 10
Advanced features included with each system	► Virtualization of internal storage ► Data migration ► DRPs with thin provisioning ► UNMAP ► Compression and deduplication ► Metro Mirror (synchronous) and Global Mirror (asynchronous)
Additional available advanced features	► Remote mirroring ► Easy Tier compression ► External virtualization ► Encryption ► FlashCopy ► IBM Spectrum Control ► IBM Spectrum Protect Snapshot

For more information about the V8.4.2.x Configuration Limits and Restrictions for IBM FlashSystem 5100, see this IBM Support web page.

IBM FlashSystem 5200

IBM FlashSystem 5200 allows you to be ready for the technology transformation without sacrificing performance, quality, or security while simplifying your data management. This powerful and compact solution is focused on afford ability with a wide range of enterprise-grade features of IBM Spectrum Virtualize that can easily evolve and extend as businesses grows.

This system also has the flexibility and performance of flash and Non-Volatile Memory Express (NVMe) end-to-end, the innovation of IBM FlashCore technology, and Storage Class Memory (SCM) to help accelerate your business execution.

The innovative FlashSystem family is based on a common storage software platform, IBM Spectrum Virtualize, that provides powerful all-flash and hybrid-flash solutions, offering feature-rich, cost effective, enterprise-grade storage solutions.

The industry-leading capabilities of IBM Spectrum Virtualize include a wide range of data services that can be extended to more than 500 heterogeneous storage systems, as shown in the following examples:

- Automated data movement.
- Synchronous and asynchronous copy services on-premises or to the public cloud.
- High availability configurations.
- Storage automated tiering.
- Data reduction technologies, including deduplication.
- Safeguarded Copy, which supports the ability to create cyber-resilient point-in-time copies of volumes that cannot be changed or deleted through user errors, malicious actions, or ransomware attacks. The system integrates with IBM Copy Services Manager to provide automated backup copies and data recovery.

Available on IBM Cloud® and AWS, IBM Spectrum Virtualize for Public Cloud works with IBM FlashSystem 5200 to deliver consistent data management between on-premises storage and public cloud: for example:

- Moving data and applications between on-premises and public cloud
- Implementing new DevOps strategies
- Using public cloud for disaster recovery without the cost of a second data center
- Improving cyber resiliency with “air gap” cloud snapshots

IBM FlashSystem 5200 offers world-class customer support, product upgrade, and guarantee programs:

- The IBM Storage Expert Care service and support is simple. You can easily select the level of support and period that best fits your needs with predictable and up-front pricing that is a fixed percentage of the system cost.
- The IBM Data Reduction Guarantee helps reduce planning risks and lower storage costs with baseline levels of data compression effectiveness in IBM Spectrum Virtualize-based offerings.
- The IBM Controller Upgrade Program enables customers of designated all-flash IBM storage systems to reduce costs while maintaining leading-edge controller technology for essentially the cost of ongoing system maintenance.

The IBM FlashSystem 5200 control enclosure supports up to 12 x 2.5-inch NVMe capable flash drives or 12 x SCM drives in a 1U high form factor.

The IBM FlashSystem 5200 has one standard model (4662-6H2) and one utility model (4662-UH6).

Figure 1-12 shows the IBM FlashSystem 5200 control enclosure front view with 12 NVMe drives and a 3/4 ISO view.

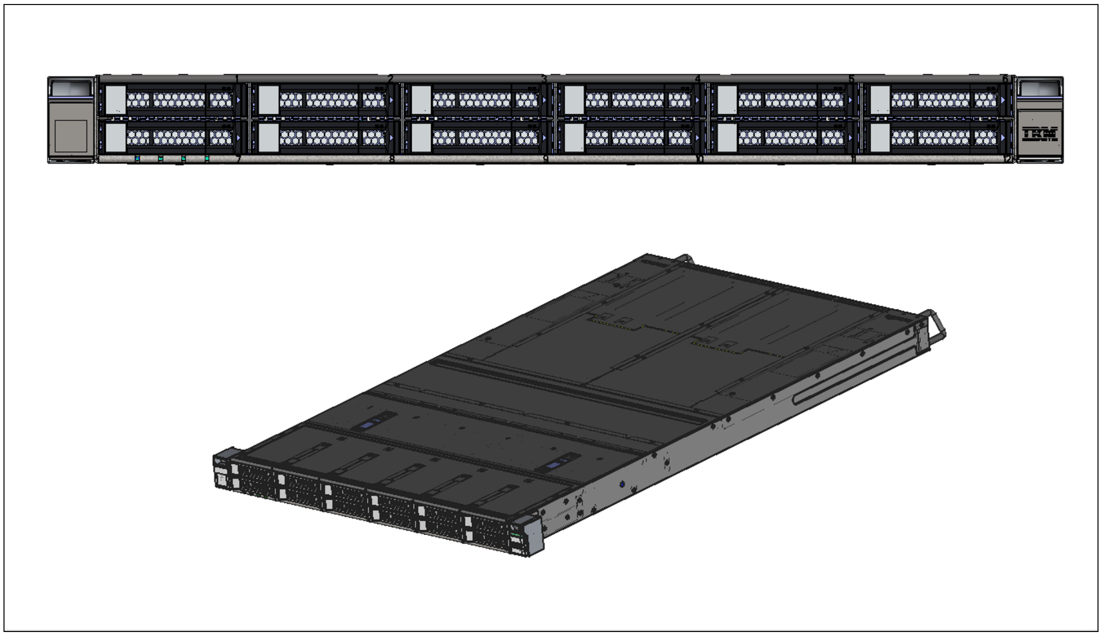


Figure 1-12 IBM FlashSystem 5200 control enclosure front and 3/4 ISO view

Table 1-9 lists the host connections, drive capacities, features, and standard options with Spectrum Virtualize that are available on the IBM FlashSystem 5200.

Table 1-9 IBM FlashSystem 5200 host, drive capacity, and functions summary

Feature/Function	Description
Host interface	► 10 Gbps Ethernet (iSCSI) ► 25 Gbps Ethernet (iSCSI, iSER - iWARP, RoCE) ► 16 Gbps Fibre Channel (FC, FC-NVMe) ► 32 Gbps Fibre Channel (FC, FC-NVMe)
Control Enclosure Supported drives (12 maximum)	► 2.5-inch Storage Class Memory: 375 GB, 750 GB, 800GB and 1.6 TB ► 2.5-inch NVMe self-compressing FCMs: 4.8 TB, 9.6 TB, 19.2 TB, and 38.4 TB ► NVMe flash drives: 800 GB, 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB
SAS expansion enclosures 760 per control enclosure, 1,520 per clustered system Model 12G 2U 12 drives Model 24G 2U 24 drives Model 92G 5U 92 drives	► 2.5-inch flash drives supported: 800 GB, 1.6 TB, 1.92 TB, 3.84 TB, 7.68 TB, 15.36 TB, and 30.72 TB ► 2.5-inch disk drives supported – 600 GB, 900 GB, 1.2 TB, 1.8 TB, and 2.4 TB 10k SAS disk – 2 TB 7.2k nearline SAS disk ► 3.5-inch disk drives supported: 4 TB, 6 TB, 8 TB, 10 TB, 12 TB, 14 TB, 16 TB, and 18 TB 7.2k nearline SAS disk
RAID levels	Distributed RAID 5 and 6, TRAID 1 and 10

Feature/Function	Description
Advanced features included with each system	► Virtualization of internal storage ► Data migration ► DRPs with thin provisioning ► UNMAP ► Compression and deduplication ► Metro Mirror (synchronous) and Global Mirror (asynchronous)
Additional available advanced features	► Remote mirroring ► Easy Tier compression ► External virtualization ► Encryption ► FlashCopy ► IBM Spectrum Control ► IBM Spectrum Protect Snapshot

For more information about the V8.4.2.x Configuration Limits and Restrictions for IBM FlashSystem 5200, see this IBM Support web page.

IBM FlashSystem 7200

Enterprises around the globe are rapidly moving to modernize older IT infrastructures to take advantage of artificial intelligence (AI)-enhanced applications, real-time big data analytics, and cloud architectures that require higher levels of system performance and storage capacity.

For many organizations, staff resources and expertise are not abundant, and cost-efficiency is a top priority. These organizations made important investments in their infrastructure that they want to maximize. They need enterprise-grade solutions that optimize cost-efficiency while simplifying the pathway to modernization.

The following new IBM FlashSystem 7200 models are designed specifically for these requirements and use cases:

- FlashSystem 7200 Models: 2076-824 or 4667-824
- FlashSystem 7200 Utility Models: 2076-U7C or 4664-U7C

The IBM FlashSystem 7200 includes the following highlights:

- Deploy enterprise-grade functions.
- Use NVMe performance in one cost efficient system.
- Build easy-to-manage, high-performance hybrid cloud environments.
- Extend data services across more than 500 heterogeneous systems.
- Transform data economics by using sophisticated data reduction.
- Use AI to optimize storage management and streamline issue resolution.
- Deploy leading-edge storage solutions with confidence by using IBM FlashWatch.
- Increase cost-efficiency with IBM Storage Utility programs.
- Safeguarded Copy function supports the ability to create cyber-resilient point-in-time copies of volumes that cannot be changed or deleted through user errors, malicious actions, or ransomware attacks. The system integrates with IBM Copy Services Manager to provide automated backup copies and data recovery.

Figure 1-13 shows the IBM FlashSystem 7200 control enclosure front view.

Figure 1-13 IBM FlashSystem 7200 control enclosure front view

Table 1-10 lists the host connections, drive capacities, features, and standard options with Spectrum Virtualize that are available on the IBM FlashSystem 7200.

Table 1-10 IBM FlashSystem 7200 host, drive capacity and functions summary

Feature/Function	**Description**
Host interface	Per control enclosure ► Up to 24 x 16 Gbps Fibre Channel (FC, NVMeoF) ► Up to 24 x 32 Gbps Fibre Channel (FC, NVMeoF) ► 8 x 10 Gbps Ethernet (iSCSI) ► Up to 12 x 25 Gbps Ethernet (iSCSI, iSER - iWARP, RoCE)
Maximum drives supported	► 24 x 2.5-inch NVMe drives per control enclosure ► 12 x 3.5-inch SAS drives per 12G expansion enclosure ► 24 x 2.5-inch SAS drives per 24G expansion enclosure ► 92 x 2.5-inch or 3.5-inch SAS drives per 92G expansion enclosure ► Up to a maximum of 760 SAS drives in expansion enclosures per control enclosure
Supported NVMe drives	► Storage Class Memory (SCM): 375 GB, 750 GB, 800 GB, 1.6 TB ► FlashCore Modules (FCM): 4.8 TB, 9.6 TB, 19.2 TB, and 38.4 TB with hardware compression ► Industry Standard NVMe: 800 GB, 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB
Supported SAS drives	► Supported SAS drives 2.5-inch SAS SSD: 800 GB, 1.6 TB, 1.92 TB, 3.84 TB, 7.68 TB, 15.36 TB, and 30.72 TB ► 2.5-inch SAS HDD: – 1.2 TB, 1.8 TB, and 2.4 TB 10k SAS – 2 TB 7.2k nearline SAS ► 3.5-inch disk drives supported: 4 TB, 6 TB, 8 TB, 10 TB, 12 TB, 14 TB, 16 TB, and 18 TB 7.2k nearline SAS
RAID levels	DRAID 1, 5, and 6 with dynamic DRAID expansion and TRAID 1 and 10
Maximum IOPS (4K read hit)	2.3 million
Minimum latency (4K read hit)	<70 µs

Feature/Function	Description
Maximum IOPS (4K read miss)	700 k
Maximum bandwidth (256Kb read miss)	35 GB/s
Advanced features included with each system	► Virtualization of internal storage ► Data migration ► DRPs with thin provisioning ► UNMAP ► Compression and deduplication ► Metro Mirror (synchronous) and Global Mirror (asynchronous)
Additional available advanced features	► Remote mirroring ► Easy Tier compression ► External virtualization ► Encryption ► FlashCopy ► IBM Spectrum Control ► IBM Spectrum Protect Snapshot

For more information, see this IBM Support web page.

IBM FlashSystem 9100

Figure 1-14 shows the IBM FlashSystem 9100 Control Enclosure with one of the IBM NVMe drives partially removed.

Figure 1-14 IBM FlashSystem 9100 control enclosure with one NVMe drives partially removed

The IBM FlashSystem 9100 control enclosure supports up to 24 NVMe capable flash drives in a 2U high form factor.

Two standard models of The IBM FlashSystem 9100 are available: 9110-AF7 and 9150-AF8.

These numbers are the sales models, and each one is available as a one-year (hardware machine type 9846), or a three-year (hardware machine type 9848) warranty product.

The IBM FlashSystem 9100 also has two utility models: 9110-UF7 and 9150-UF8.

Note: Consider the following points

- The IBM 9110-UF7 and 9150-UF8 are the IBM FlashSystem 9100 with a three-year warranty only. These models are physically and functionally identical to the IBM FlashSystem 9848-AF7 and AF8, except for target configurations and variable capacity billing.

  The variable capacity billing uses IBM Spectrum Control Storage Insights to monitor the system usage, which allows allocated storage usage that exceeds a base subscription rate to be billed per TB (written is considered used). For thick provisioning, total allocated volume space is considered used.

- The following IBM FS 9100 systems are now withdrawn form Marketing (WFM) since October 2020. As such, they are no longer available to purchase from IBM:
  - FS9110 9848-AF7
  - FS9150 9848-AF8
  - FS9110 Utility Model 9848-UF7
  - FS9150 Utility Model9848-UF8

  These systems are included in this book for completeness only because they support running the Spectrum Virtualize software v8.4.

Table 1-11 lists the host connections, drive capacities, features, and standard options with Spectrum Virtualize that are available on the IBM FlashSystem 9100.

Table 1-11 IBM FlashSystem 9100 host, drive capacity, and functions summary

Feature/Function	Description
Host interface	► 24 ports 16 Gb or 32 Gb Fibre Channel (FC, FC-NVMe) ► 8 ports 10 GbE iSCSI ► 12 ports 25 GbE iWARP or RoCE
Maximum drives supported	► 2.5-inch NVMe FCMs: 4.8 TB, 9.6 TB, and 19.2 TB compressing FCMs ► 2.5-inch NVMe flash drives: 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB
Control enclosure supported NVMe drives	► 2.5-inch NVMe FCMs: 4.8 TB, 9.6 TB, and 19.2 TB compressing FCMs ► 2.5-inch NVMe flash drives: 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB
Expansion enclosure supported SAS drives	► Model AFF 2U 24 drive ► Model A9F 5U 92 drive ► 2.5-inch flash drives supported: 1.92 TB, 3.2 TB, 3.84 TB, 7.68 TB, 15.36 TB, and 30.72 TB
RAID levels	► FCM drives: Distributed RAID 6 (recommended),: Distributed RAID5 (supported) ► NVMe flash drives: Traditional RAID 10 and Distributed RAID 6 (recommended), Distributed RAID 5 (supported)
Maximum IOPS (4K read hit)	3,800,000
Minimum write latency	120 µs
Maximum IOPS (4K read miss with hardware compression)	1,200,000

Feature/Function	Description
Maximum bandwidth (256Kb read miss)	34 GBps
Advanced features included with each system	► Virtualization of internal storage ► Data migration ► DRPs with thin provisioning ► UNMAP ► Compression and deduplication ► Metro Mirror (synchronous) and Global Mirror (asynchronous
Additional available advanced features	► Remote mirroring ► Easy Tier compression ► External virtualization ► Encryption ► FlashCopy ► IBM Spectrum Control ► IBM Spectrum Protect Snapshot

For more information, see this IBM Support web page.

IBM FlashSystem 9200

Some applications exist that are foundational to the operations and success of an enterprise. These applications might function as prime revenue generators, might guide or control important tasks, or might provide crucial business intelligence, among many other jobs. Whatever their purpose, they are mission-critical to the organization. They demand the highest levels of performance, functionality, security, and availability.

To support mission-critical applications, enterprises of all types and sizes turn to IBM FlashSystem 9200:

- FlashSystem 9200 Models: 9848-AG8 or 4666-AFF/4666-AG8/4666-A9F
- FlashSystem 9200 Utility Models: 9848-UG8 or 4664-U7C

The 46XX Models include Expert Care Support.

IBM FlashSystem 9200 combines the performance of flash and a Non-Volatile Memory Express (NVMe)-optimized architecture with the reliability and innovation of IBM FlashCore® technology and the rich feature set and high availability of IBM Spectrum Virtualize. This powerful new storage platform provides:

- The option to use large capacity IBM FlashCore modules (FCM) with inline-hardware compression, data protection, and innovative flash-management features; industry standard NVMe drives; or Storage Class Memory (SCM) drives.
- The software-defined storage functionality of IBM Spectrum Virtualize with a full range of industry-leading data services such as dynamic tiering,
- IBM FlashCopy management, data mobility, and high-performance data encryption.
- Innovative DRP (DRP) technology that includes deduplication and hardware-accelerated compression technology, with SCSI UNMAP support and all the thin provisioning, copy management, and efficiency you'd expect from IBM Spectrum Virtualize-based storage.
- Safeguarded Copy function supports the ability to create cyber-resilient point-in-time copies of volumes that cannot be changed or deleted through user errors, malicious actions, or ransomware attacks. The system integrates with IBM Copy Services Manager to provide automated backup copies and data recovery.

Figure 1-15 shows the IBM FlashSystem 9200 Control Enclosure with 24 NVMe FCM type drives installed.

Figure 1-15 IBM FlashSystem 9200 control enclosure

The IBM FlashSystem 9200 solutions provide a single enterprise class platform to address the full spectrum of 21st-century data storage requirements. IBM FlashSystem 9200 is designed to simplify storage and accelerate business productivity, with the following benefits:

- NVMe-powered all-flash performance and IBM FlashCore reliability
- Easy integration and almost unlimited scalability
- Data services that can transform and modernize existing systems

Table 1-12 lists the host connections, drive capacities, features, and standard options with Spectrum Virtualize that are available on the IBM FlashSystem 9200.

Table 1-12 IBM FlashSystem 9200 host, drive capacity, and functions summary

Feature/Function	Description
Host interface	► Up to 24 x 16 Gbps Fibre Channel (FC, NVMeoF) ► Up to 24 x 32 Gbps Fibre Channel (FC, NVMeoF) ► 8 x 10 Gbps Ethernet (iSCSI) ► Up to 12 x 25 Gbps (iSCSI, iSER - iWARP, RoCE)
Maximum drives supported	► 24 NVMe drives per control enclosure ► 24 2.5-inch SAS drives per AFF expansion enclosure ► 92 2.5-inch SAS drives per A9F expansion enclosure ► Up to a maximum of 760 SAS drives in expansion enclosures per control enclosure
Supported NVMe drives	► Storage Class Memory (SCM: 375 GB, 750 GB, 800 GB, 1.6 TB ► FlashCore Modules: 4.8 TB, 9.6 TB, 19.2 TB, and 38.4 TB with hardware compression ► Industry-standard NVMe: 800 GB, 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB
Supported SAS drives	► 2.5-Inch SAS SSD: 1.6 TB, 1.92 TB, 3,84 TB, 7.68 TB, 15.36 TB, and 30.72 TB
RAID levels	DRAID 1, 5, and 6 with dynamic DRAID expansion and TRAID 1 and 10
Maximum IOPS (4K read hit)	4.5 million
Minimum latency (4K read hit)	<70 µs

Feature/Function	Description
Maximum IOPS (4K read miss)	1.2 million
Maximum bandwidth (256Kb readmiss)	45 GBps
Advanced features	► Data reduction by way of thin provisioning ► UNMAP ► Compression and deduplication ► Data-at-rest AES-XTS 256 encryption ► Easy Tier ► Data migration ► External virtualization
Replication features	► FlashCopy ► Metro Mirror (synchronous) ► Global Mirror (asynchronous) ► Global Mirror with change volumes ► Three sites replication ► IBM HyperSwap (high availability)
Additional available advanced features	► IBM Storage Insights Pro ► IBM Spectrum Virtualize for Public Cloud ► IBM Spectrum Control ► IBM Spectrum Protect Snapshot

For more information, see IBM Support web page.

1.3.1 Clustering rules and upgrades

The IBM Storwize and IBM FlashSystem products can be clustered with up to four control enclosures and uses four I/O groups. The lower-level products feature some restrictions because of hardware and memory constraints, but most systems can cluster.

All of the options and rules for clustering control nodes are shown in the figures that are presented in this section. The numbers in the cells indicate the maximum control enclosures that is possible in the cluster.

IBM Storwize V5030 and IBM FlashSystem 503x clustering

Figure 1-16 shows the combinations of IBM Storwize V5030 and IBM FlashSystem 503x that can be clustered.

Product	MTM	V5030	V5030E	FS5030	FS5035
		2078-312 2078-324	2072-312 2072-324	2072-3H2 2072-3H4	2072-3N2 2072-3N4
V5030	2078-312/324	2			
V5030E	2072-312/324		2	2	2
FS5030	2072-3H2/3H4		2	2	2
FS5035	2072-3N2/3N4		2	2	2

Figure 1-16 IBM Storwize V5030 and FlashSystem 503x Clustering

IBM Storwize V5100 and IBM FlashSystem 5x00 clustering

Figure 1-17 shows the combinations of IBM Storwize V5100 and IBM FlashSystem 5x00 that can be clustered.

Product		V5100	FS5100	FS5200
	MTM	2078-424 2078-AF4	2077-4H4	4662-6H2
V5100	2078-424/AF4	2	2	
FS5100	2077-4H4	2	2	
FS5200	4662-6H2			4

Figure 1-17 IBM Storwize V5100 and FlashSystem 5x00 Clustering

IBM Storwize V7000, IBM FlashSystem 7200, and IBM FlashSystem 9xx0 clustering

Figure 1-18 shows the combinations of IBM Storwize V7000, IBM FlashSystem 7200, and IBM FlashSystem 9xx0 that can be clustered.

It also indicates the final or terminal code level for some systems and the minimum supported code level for other combinations.

Product		V7000	V7000 Gen2	V7000 Gen2+	V7000 Gen3	FS7200	FS9110	FS9150	FS9200
	MTM	2076-112/124 2076-312/324	2076-524	2076-624	2076-724	2076-824 4664-824	9848-AF7 9848-AF8	9848-AF8	9848-AG8 4666-AG8
	Terminal Level	7.8.1	8.4.0	8.4.0					
	Minimum Supported Level			7.8.1	8.2.0	8.3.1	8.2.0	8.2.0	8.3.1
V7000	2076-112/124 2076-312/324	4	4	4					
V7000 Gen2	2076-524	4	4	4	4	4	4	4	4
V7000 Gen2+	2076-624	4	4	4	4	4	4	4	4
V7000 Gen3	2076-724		4	4	4	4	4	4	4
FS7200	2076-824 4664-824		4	4	4	4	4	4	4
FS9110	9848-AF7/AF8		4	4	4	4	4	4	4
FS9150	9848-AF8		4	4	4	4	4	4	4
FS9200	9846-AG8 4666-AG8		4	4	4	4	4	4	4

Figure 1-18 IBM Storwize V7000, FlashSystem 7200 and FlashSystem 9xx0 Clustering

For more information, see this IBM Support web page.

1.3.2 Mixed clustering rules and licensing

From Spectrum Virtualize software version 8.2.0 onwards, when you cluster dissimilar models in a system, the resulting licensing scheme that is used for the system is overwritten by the licensing scheme of the most capable system. For example, V7000 is over-ridden by FS7200, which is over-ridden by FS9200.

The extended rule is that the new or highest system overrules anything else in the cluster.

The following priority order is used:

FS9200 → FS9100 → FS7200 → V7000 → FS5200

Consider the example of adding an FS7200 I/O group to an FS7200. If you then add an FS9100, the resulting cluster is an FS9100. If you then add an FS9200, the cluster reports as an FS9200.

Consider the following points:

- All systems must have the same level of IBM Spectrum Virtualize software installed to be able to cluster.
- To cluster the Storwize V7000 systems, it must have an all-inclusive license.
- Migration must be done through extra I/O groups.
- The default layer is storage, but a replication layer also is supported for clustering.
- The systems that are listed in Table 1-13 on page 32 cannot be clustered with the IBM FlashSystem V9000 or the IBM Storage Virtualization Controller.

1.3.3 IBM FlashSystem 9200R rack solution overview

IBM FlashSystem 9200R is a pre-cabled and preconfigured rack solution that includes many IBM FlashSystem 9200 Control Enclosures. It uses IBM Spectrum Virtualize to linearly scale the performance and capacity through clustering. For more information about this product, see *IBM FlashSystem 9200R Rack Solution Product Guide*, REDP-5593.

The IBM FlashSystem 9200R Rack Solution system features a dedicated FC network for clustering and optional expansion enclosures, which are delivered ready-assembled in a rack. Available with two, three, or four clustered IBM FlashSystem 9200 systems and up to four expansion enclosures, it can be ordered as a IBM FlashSystem 9202R, IBM FlashSystem 9203R, or IBM FlashSystem 9204R system with the last number denoting the number of AG8 controller enclosures in the rack.

The final configuration occurs on-site following the delivery of the systems. More components can be added to the rack after delivery to meet the growing needs of the business.

Note: Other than the IBM FlashSystem 9200 control enclosure and its expansion enclosures, the additional components of this solution are not covered under Enterprise Class Support (ECS). Instead, the components have their own warranty, maintenance terms, and conditions.

Rack rules

The IBM FlashSystem 9200R Rack Solution product represents a limited set of possible configurations. Each IBM FlashSystem 9200R Rack Solution order must contain the following components:

- Two, three, or four 9848 Model AG8 Control Enclosures.
- Two IBM SAN24B-6 or two IBM SAN32C-6 FC switches.
- Optionally, 0 - 4 9848 Model AFF Expansion Enclosures, with no more than one expansion enclosure per Model AG8 Control Enclosure and no mixing with the 9848 Model A9F Expansion Enclosure.
- Optionally, 0 - 2 9848 Model A9F Expansion Enclosures, with no more than one expansion enclosure per Model AG8 Control Enclosure and no mixing with 9848 Model A9F Expansion Enclosure.

- One 7965-S42 rack with the appropriate power distribution units (PDUs) that are required to power components within the rack.
- All components in the rack much include feature codes #FSRS and #4651.
- For Model AG8, AFF, and A9F Control Enclosures, the first and largest capacity enclosure includes feature code #AL01, with subsequent enclosures that use #AL02, #AL03, and #AL04 in capacity order. The 9848 Model AG8 Control Enclosure with #AL01 must also have #AL0R included.

Following the initial order, each 9848 Model AG8 Control Enclosures can be upgraded through MES.

More components can be ordered separately and added to the rack within the configuration limitations of the IBM FlashSystem 9200 system. Clients must ensure that the space, power, and cooling requirements are met. If assistance is needed with the installation of these additional components beyond the service that is provided by your IBM System Services Representative (IBM SSR), IBM Lab Services are available.

Table 1-13 lists the IBM FlashSystem 9200R Rack Solution combinations, the MTMs, and their associated feature codes.

Table 1-13 IBM FlashSystem 9200R Rack Solution combinations

Machine type and model	Description	Quantity
7965-S42	IBM Enterprise Slim Rack	1
8960-F24	IBM SAN24B-6 Fibre Channel switch (Brocade)	2[a]
8977-T32	IBM SAN32C-6 Fibre Channel switch (Cisco)	2[a]
9848-AFF	IBM FlashSystem 9000 2U SFF Expansion Enclosure with 3-year Warranty and ECS	0 - 4[b]
9848-AG8	IBM FlashSystem 9200 Control Enclosure with 3-year Warranty and ECS	2, 3, or 4
9848-A9F	IBM FlashSystem 9000 5U LFF high-density Expansion Enclosure with 3-year Warranty and ECS	0 - 2[b]

a. For the FC switch, choose two of machine type (MT) 8977 or two of MT 8960.
b. For extra expansion enclosures, choose model AFF, model A9F, or none. You cannot use both.

For more information about the FlashSystem 9200R solution, see the following IBM Redbooks publications:

- *IBM FlashSystem 9200R Rack Solution Product Guide*, REDP-5593
- *Implementing the IBM FlashSystem with IBM Spectrum Virtualize V8.4*, SG24-8467

1.4 Advanced functions for data reduction

The IBM FlashSystem range can function as a feature-rich, software-defined storage layer that virtualize and extends the functionality of all managed storage, including data reduction, dynamic tiering, copy services, and high-availability configurations. In this capacity, the IBM FlashSystem acts as the virtualization layer between the host and other external storage systems, providing flexibility, and extending functionality to the virtualized external storage capacity.

The IBM FlashSystem 5100, 5200, 7200, 9100, and 9200 all use several features to assist with data reduction and the ability to increase their effective capacity.

1.4.1 FlashCore Modules

All IBM FlashSystems 5100, 5200, 7200, 9100, and 9200 have the option to be supplied with FlashCore Modules (FCMs) or industry-standard NVMe drives. If the FCM option is chosen, the user can use the built-in hardware compression, which automatically attempts to compress the stored data when written to the drives. These FCMs can be used with standard pools or DRPs.

1.4.2 Data reduction pools

DRPs represent a significant enhancement to the storage pool concept. This is because the virtualization layer is primarily a simple layer that executes the task of lookups between virtual and physical extents. Now with the introduction of data reduction technology, compression, and deduplication, it is more of a requirement to have an uncomplicated way to stay “thin”. The pools enable you to automatically de-allocate (not to be confused with deduplicate) and reclaim capacity of thin-provisioned volumes containing deleted data.

1.4.3 Deduplication

Deduplication can be configured with thin-provisioned and compressed volumes in DRPs for added capacity savings. The deduplication process identifies unique chunks of data, or byte patterns, and stores a signature of the chunk for reference when writing new data chunks. If the signature of the new chunk matches an existing signature, the new chunk is replaced with a small reference that points to the stored chunk. The same byte pattern can occur many times, resulting in a sizable reduction of the amount of data that must be stored.

1.4.4 Thin provisioning

In a shared storage environment, thin provisioning is a method for optimizing the use of available storage. It relies on allocation of blocks of data on demand versus the traditional method of allocating all of the blocks up front.

This methodology eliminates almost all white space, which helps avoid the poor usage rates (often as low as 10%) that occur in the traditional storage allocation method. Traditionally, large pools of storage capacity are allocated to individual servers, but remain unused (not written to).

1.4.5 Thin-provisioned FlashCopy snapshots

Thin-provisioned IBM FlashCopy (or snapshot function in the GUI) uses disk space only when updates are made to the source or target data, and not for the entire capacity of a volume copy.

1.5 Advanced software features

The IBM FlashSystem 5100, 5200, 7200, 9100, and 9200 include the following advanced software features:

- Data Migration
- Copy services:
  - Metro Mirror
  - FlashCopy
  - 3-site replication
  - Safeguarded Copy
- Easy Tier
- External Virtualization
- HyperSwap
- Safeguarded Copy with CSM

1.5.1 Data migration

The IBM FlashSystem range provides online volume migration while applications are running, which is possibly the greatest single benefit for storage virtualization. This capability enables data to be migrated on and between the underlying storage subsystems without any effect on the servers and applications. In fact, this migration is performed without the knowledge of the servers and applications that it even occurred. The IBM FlashSystem deliver these functions in a homogeneous way on a scalable and highly available platform over any attached storage and to any attached server.

1.5.2 Copy services

Advanced copy services are a class of functionality within storage arrays and storage devices that enable various forms of block-level data duplication locally or remotely. By using advanced copy services, you can make mirror images of part or all of your data eventually between distant sites. Copy services functions are implemented within an IBM FlashSystem (FlashCopy and Image Mode Migration), or between one IBM FlashSystem and another IBM FlashSystem, or any other member of the IBM Spectrum Virtualize family, in three different modes:

- *Metro Mirror* is the IBM branded term for synchronous Remote Copy function.
- *Global Mirror* is the IBM branded term for the asynchronous Remote Copy function.
- *Global Mirror with Change Volumes* is the IBM branded term for the asynchronous Remote Copy of a locally and remotely created FlashCopy.

Remote replication can be implemented using both Fibre Channel and Internet Protocol (IP) network methodologies.

For more, see Chapter 6, “IBM FlashCopy services capabilities” on page 243.

FlashCopy

FlashCopy is the IBM branded name for point-in-time copy, which is sometimes called time-zero (T0) copy. This function makes a copy of the blocks on a source volume and can duplicate them on 1 - 256 target volumes.

Remote mirroring

The three remote mirroring modes are implemented at the volume layer within the IBM FlashSystem family. They are collectively referred to as Remote Copy capabilities. In general, the purpose of these functions is to maintain two copies of data.

Often, but not necessarily, the two copies are separated by distance. The Remote Copy can be maintained in one of two modes: synchronous or asynchronous, with a third asynchronous variant:

- Metro Mirror
- Global Mirror
- Global Mirror with Change Volumes

1.5.3 Easy Tier

Easy Tier is a performance function that automatically migrates or moves extents of a volume from one storage tier to another storage tier. With IBM FlashSystem, Easy Tier supports four kinds of storage tiers.

Consider the following information about Easy Tier:

- Easy Tier monitors the host volume I/O activity as extents are read, and migrates the most active extents to higher performing tiers.
- The monitoring function of Easy Tier is continual but, in general, extents are migrated over a 24-hour period. As extent activity cools, Easy Tier moves extents to slower performing tiers.
- Easy Tier creates a migration plan that organizes its activity to decide how to move extents. This plan can also be used to predict how extents will be migrated.

1.5.4 External virtualization

The IBM FlashSystem range includes data virtualization technology to help insulate hosts, hypervisors, and applications from physical storage. This enables them to run without disruption, even when changes are made to the underlying storage infrastructure. The IBM FlashSystem functions benefit all virtualized storage.

For example, Easy Tier and DRPs with compression help improve performance and increase effective capacity, where high-performance thin provisioning helps automate provisioning. These benefits can help extend the useful life of existing storage assets, reducing costs. Additionally, because these functions are integrated into the IBM FlashSystem 5100, 7200, 9100, and 9200, they can operate smoothly together, reducing management effort.

1.5.5 IBM HyperSwap

HyperSwap capability enables each volume to be presented by two IBM FlashSystem family I/O groups. The configuration tolerates combinations of node and site failures, using host multipathing driver based on the one that is available for the IBM FlashSystem family. The IBM FlashSystem provides GUI and CLI management of the HyperSwap function.

A more information about the HyperSwap function, see Appendix A, “IBM i considerations” on page 601.

For more information, see this IBM Doumentation web page.

1.5.6 Safeguarded Copy function

Safeguarded Copy function supports the ability to create cyber-resilient point-in-time copies of volumes that cannot be changed or deleted through user errors, malicious actions, or ransomware attacks. The system integrates with IBM Copy Services Manager to provide automated backup copies and data recovery.

The system supports IBM Copy Services Manager as an external scheduling application. IBM Copy Services Manager coordinates and automates Safeguarded Copy function across multiple systems.

IBM Copy Services Manager uses a Safeguarded policy to configure FlashCopy mapping and consistency groups automatically to create backup copies. When Safeguarded backups are created, IBM Copy Services Manager uses the retention time for the Safeguarded backups that are based on the settings in the Safeguarded policy. After copies expire, the IBM Spectrum Virtualize software deletes the expired copies from the Safeguarded backup location.

For more information, see *Implementation Guide for SpecV/FlashSystem Safeguarded Copy*, REDP-5654.

1.5.7 Licensing

The base license that is provided with the system includes the use of its basic functions. However, extra licenses can be purchased to expand the capabilities of the system. Administrators are responsible for purchasing extra licenses and configuring the systems within the license agreement, which includes configuring the settings of each licensed function on the system.

For a more information about the licensing on the IBM FlashSystem 5100, 5200, 7200, 9100, and 9200, see the “Licensing and Features” chapter in *Implementing the IBM FlashSystem with IBM Spectrum Virtualize Version 8.4.2*, SG24-8506.

2

Connecting IBM Spectrum Virtualize and IBM Storwize in storage area networks

The storage area network (SAN) is one of the most important aspects when implementing and configuring IBM Spectrum Virtualize and IBM FlashSystem.

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

> **Note:** Important: This chapter was written specifically for IBM FlashSystem 9200, however most of the general principles apply to the IBM FlashSystem 9100.
>
> If you are in doubt as to whether the principles are applicable to the FlashSystem 9100, contact your local IBM representative.
>
> For more information, see *IBM FlashSystem 9200 Product Guide*, REDP-5586.

This chapter includes the following sections:

2.1 SAN topology general guidelines

The SAN topology requirements for IBM FlashSystem do not differ too much from any other SAN. Remember that a well-sized and designed SAN allows you to build a redundant and failure-proof environment, and minimizing performance issues and bottlenecks. Therefore, before installing any of the products covered by this book, ensure that your environment follows an actual SAN design and architecture, with vendor recommended SAN devices and code levels.

For more SAN design and preferred practices, see the Brocade publication *SAN Fabric Administration Best Practices Guide Support Perspective*.

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

2.1.1 SAN performance and scalability

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

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

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

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

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

If new ports need to be added to the SAN, you might need to drastically modify the SAN to accommodate a larger-than-expected number of hosts or storage. Sometimes these changes increase the number of hops on the SAN, and so cause performance and ISL congestion issues. For additional information, see 2.1.2, “ISL considerations” on page 39.

Consider the use of SAN director-class switches. They reduce the number of switches in a SAN and provide the best scalability available. Most of the SAN equipment vendors provide high port density switching devices.

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

2.1.2 ISL considerations

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

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

- Beware of the ISL oversubscription ratio

  The standard recommendation is up to 7:1 (seven hosts using a single ISL). However, it can vary according to your SAN behavior. Most successful SAN designs are planned with an oversubscription ratio of 7:1 and some extra ports are reserved to support a 3:1 ratio. However, high-performance SANs start at a 3:1 ratio.

  Exceeding the standard 7:1 oversubscription ratio requires you to implement fabric bandwidth threshold alerts. If your ISLs exceed 70%, schedule fabric changes to distribute the load further.

- Avoid unnecessary ISL traffic

  If you plan to use external virtualized storages, connect all FlashSystem canister ports in a clustered system to the same SAN switches or Directors as all of the storage devices with which the clustered system of FlashSystem is expected to communicate. Conversely, storage traffic and internode traffic must *never* cross an ISL, except during migration scenarios.

  Keep high-bandwidth utilization servers and I/O Intensive application on the same SAN switches as the FlashSystem host ports. Placing these servers on a separate switch can cause unexpected ISL congestion problems. Also, placing a high-bandwidth server on an edge switch wastes ISL capacity.

- Properly size the ISLs on your SAN. They must have adequate bandwidth and buffer credits to avoid traffic or frames congestion. A congested inter-switch link can affect the overall fabric performance.
- Always deploy redundant ISLs on your SAN. Using an extra ISL avoids congestion if an ISL fails because of certain issues, such as a SAN switch line card or port blade failure.
- Use the link aggregation features, such as Brocade Trunking or Cisco Port Channel, to obtain better performance and resiliency.
- Avoid exceeding two hops between the FlashSystem and the hosts. More than two hops are supported. However, when ISLs are not sized properly, more than two hops can lead to ISL performance issues and buffer credit starvation (SAN congestion).

When sizing over two hops, consider that all the ISLs going to the switch where the Flash System is connected will also handle the traffic coming from the switches on the edges, as shown in Figure 2-1.

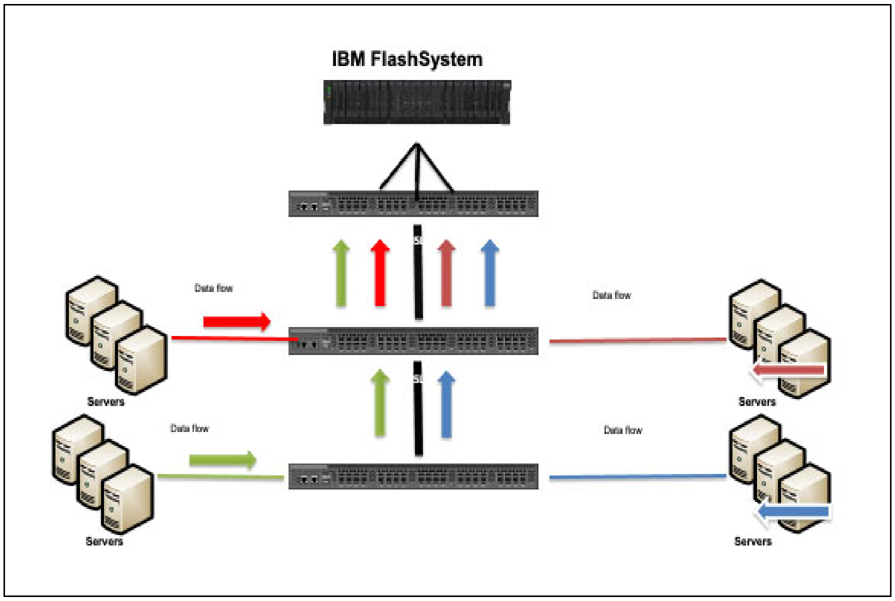


Figure 2-1 ISL data flow

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

2.2 SAN topology-specific guidelines

Some preferred practices apply to all SANs, as described in 2.1, "SAN topology general guidelines" on page 38. However, each SAN topology has its own specific preferred practices requirements. In this section, we discuss the difference between the different types of topology and highlight the specific considerations for each.

This section covers the following topologies:

- Single switch fabric
- Core-edge fabric
- Edge-core-edge
- Full MeSH

2.2.1 Single switch SANs

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

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

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

The preferred practice is to use a multi-slot director-class single switch rather than setting up a core-edge fabric that is made up solely of lower-end switches, as described in 2.1.1, "SAN performance and scalability" on page 38.

The single switch topology, as shown in Figure 2-2, has only two switches; therefore, the FlashSystem ports must be equally distributed on both fabrics.

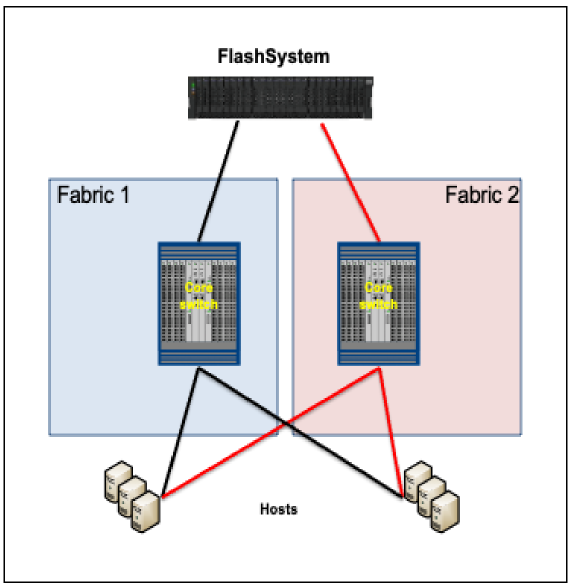


Figure 2-2 Singe Switch topology

Note: To correctly size your network, always calculate the short-term and mid-term growth to avoid lack of ports (on this topology, the limit of ports is based on the switch size). If other switches are added to the network, doing so automatically changes the topology type.

2.2.2 Basic core-edge topology

The core-edge topology (as shown in Figure 2-3) is easily recognized by most SAN architects. This topology consists of a switch in the center (usually, a director-class switch), which is surrounded by other switches. The *core switch* contains all FlashSystem and high-bandwidth hosts. It is connected by using ISLs to the edge switches. The edge switches can be of any size from 24 port switches up to multi-slot directors.

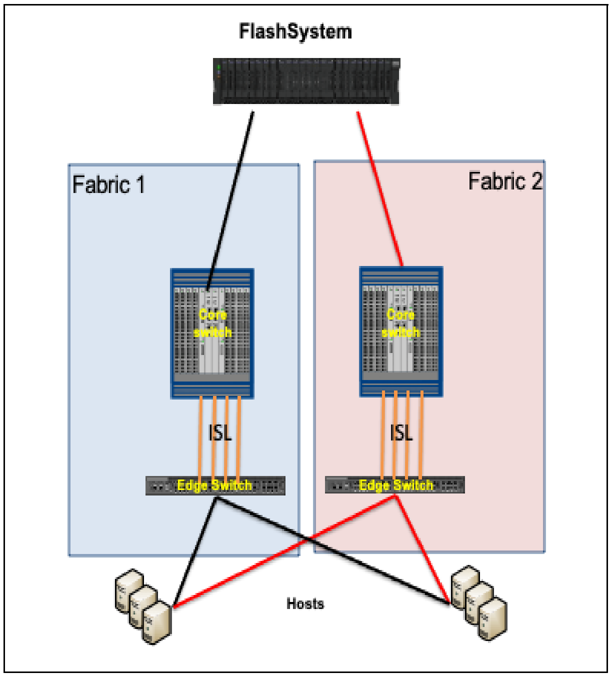


Figure 2-3 Core/Edge Topology

When the FlashSystem and servers are connected to different switches, the hop count for this topology is one.

Note: This topology is commonly used to easily growth your SAN network by adding edge switches to the core switch. Consider the ISL ratio and use of physical ports from the core switch when adding new edge switches to your network.

2.2.3 Edge-core-edge topology

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

Figure 2-4 shows an edge-core-edge topology with two different edges, one of which is exclusive for the FlashSystem and high-bandwidth servers. The other pair is exclusively for servers.

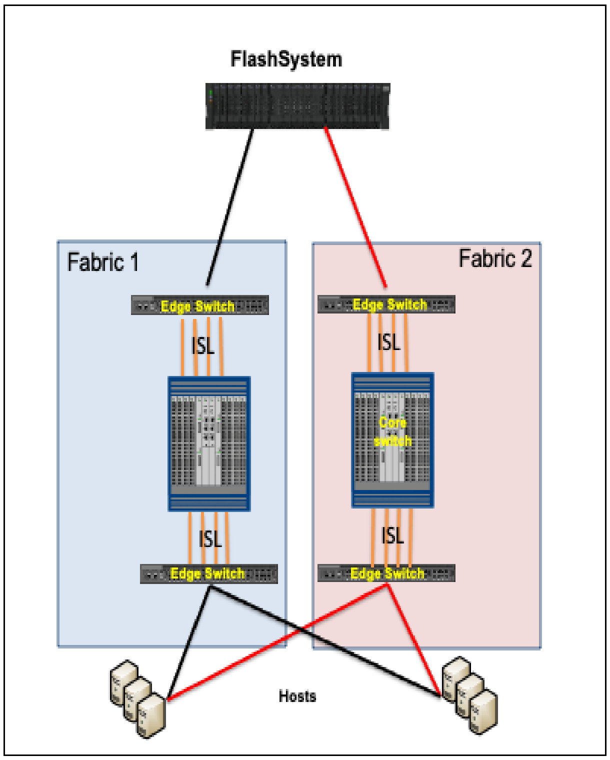


Figure 2-4 Edge-Core-Edge topology

Performance can be slightly affected if the number of hops increases, depending on the total number of switches and the distance between host and IBM FlashSystem.

Edge-core-edge fabrics allow better isolation between tiers. For more information, see 2.2.6, “Device placement” on page 46.

2.2.4 Full-mesh topology

In a full-mesh topology (see Figure 2-5), all switches are interconnected to all other switches on the same fabric. Therefore, the server and storage placement is not a concern if the number of hops is not more than one.

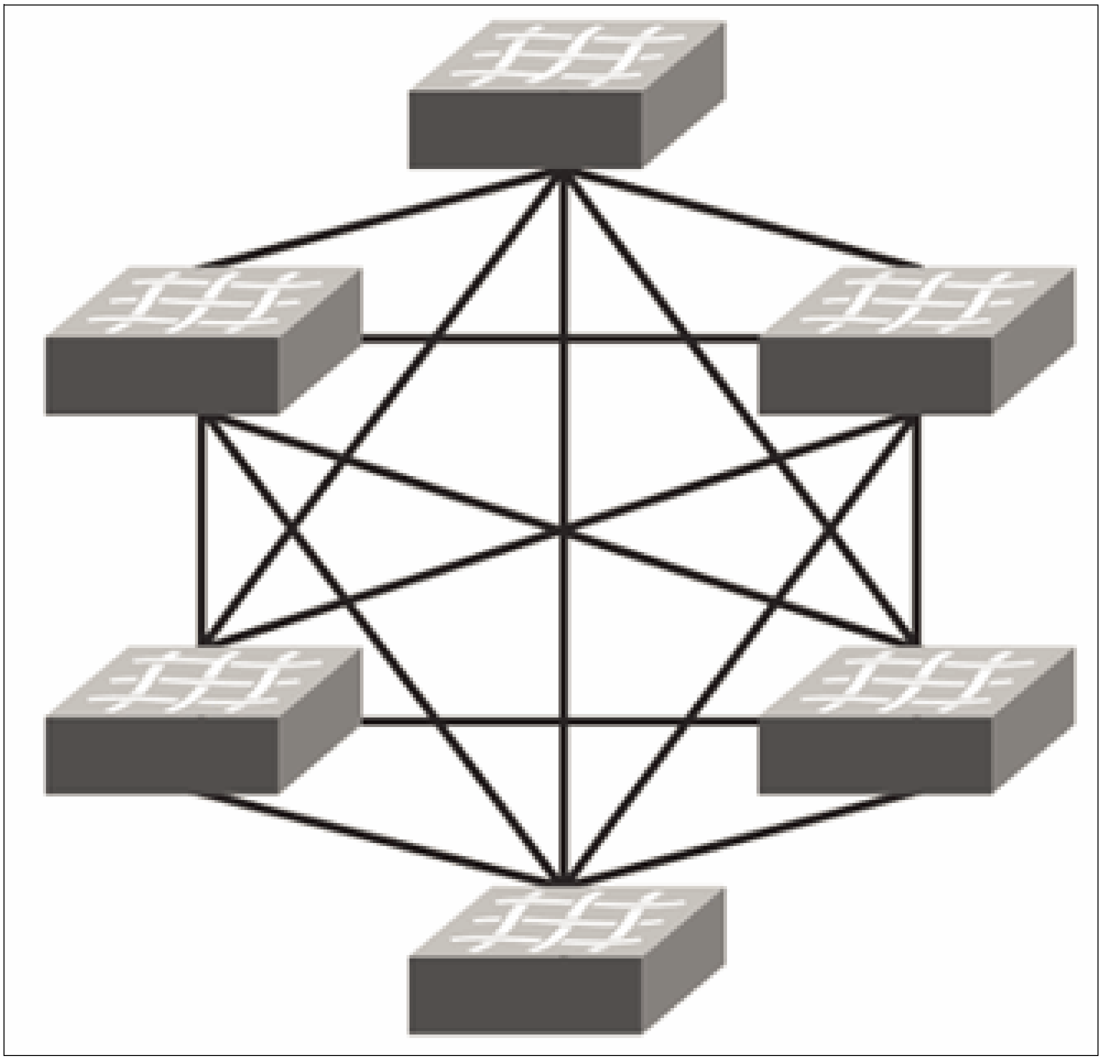


Figure 2-5 Full-mesh topology

> **Note:** Each ISL use one physical port. Depending on the total number of switches, you can ran out of ports easily based on the switch size, it is recommended that the port usage is calculated *before* a full-mesh SAN network is configured.

2.2.5 IBM FlashSystem as a multi SAN storage

IBM FlashSystem now features a maximum of 24 ports. In addition to the increased throughput capacity, this number of ports enables new possibilities and allows different kinds of topologies and migration scenarios.

One of these topologies is the use of a FlashSystem as a multi SAN storage between two isolated SANs. This configuration is useful for storage migration or sharing resources between SAN environments without merging them.

To use an external storage with IBM FlashSystem, this external storage must be attached to IBM FlashSystem through a zoning configuration and set up it as virtualized storage. This feature can be used for storage migration and decommission processes and speed up host migration. In some cases, based on the external storage configuration, virtualizing external storage with IBM FlashSystem can increase performance based on the cache capacity and processing.

Figure 2-6 shows an example of an FlashSystem as a multi SAN storage.

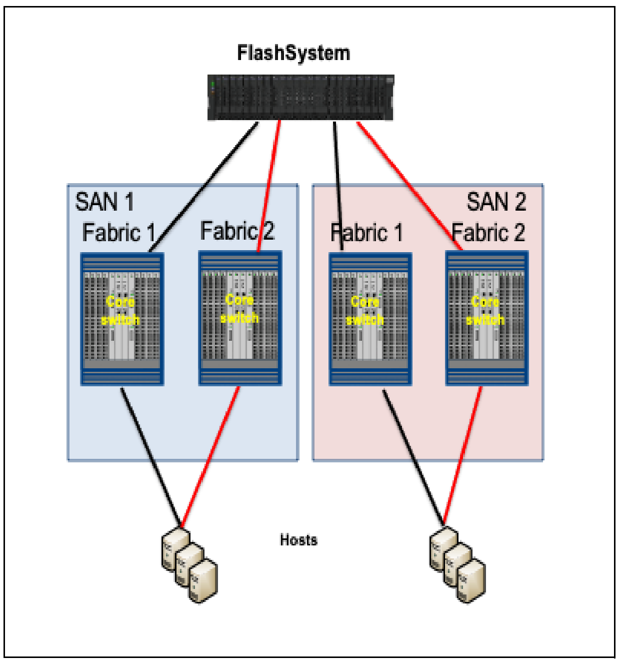


Figure 2-6 FlashSystem as a multi SAN storage.

Notice in Figure 2-6 that both SANs (shown in blue and red) are isolated. When connected to both SAN networks, IBM FlashSystem can allocate storage to hosts on both SAN networks and can virtualize storages from each SAN networks. This way, you can have older storage on the red SAN that is attached to IBM FlashSystem and providing disks to hosts on the blue network. This configuration is commonly used for migration purposes or in cases where the older storage has a lower performance compared to IBM FlashSystem.

2.2.6 Device placement

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

Keeping the traffic local to the fabric is a strategy to minimize the traffic between switches (and ISLs) by keeping storages and hosts attached to the same SAN switch, as shown in Figure 2-7.

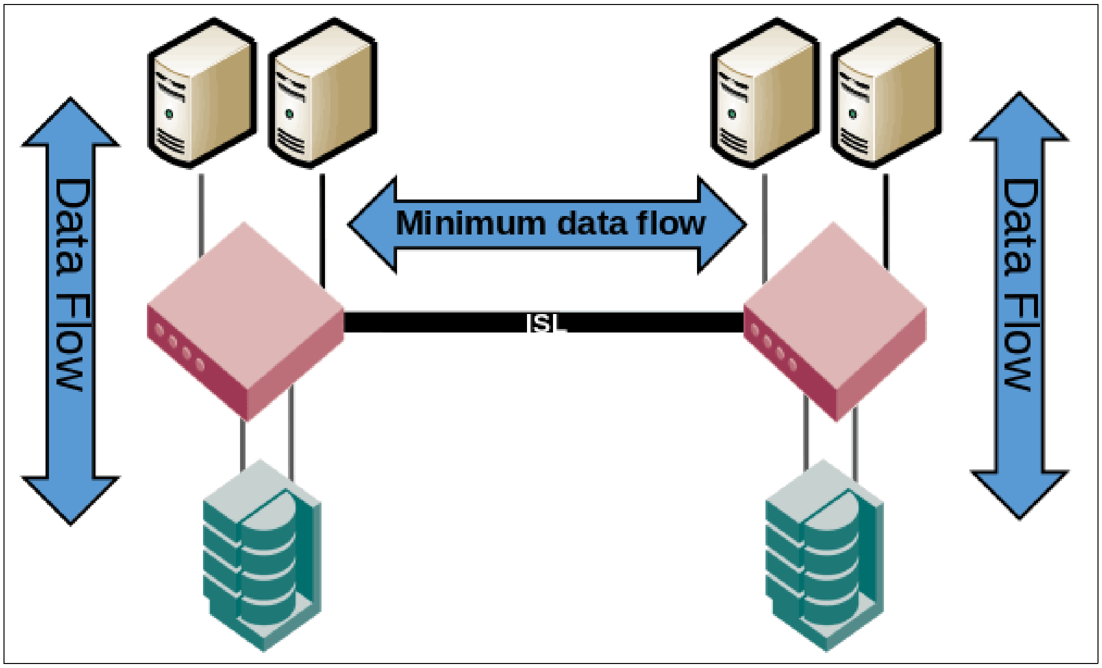


Figure 2-7 Storage and hosts attached to the same SAN switch

This solution can fit perfectly in small and medium SANs. However, it is not as scalable as other available topologies. As stated in 2.2.3, “Edge-core-edge topology” on page 43, the most scalable SAN topology is the edge-core-edge. Besides scalability, this topology provides various resources to isolate the traffic and reduce possible SAN bottlenecks.

Figure 2-8 shows an example of traffic segregation on the SAN using edge-core-edge topology.

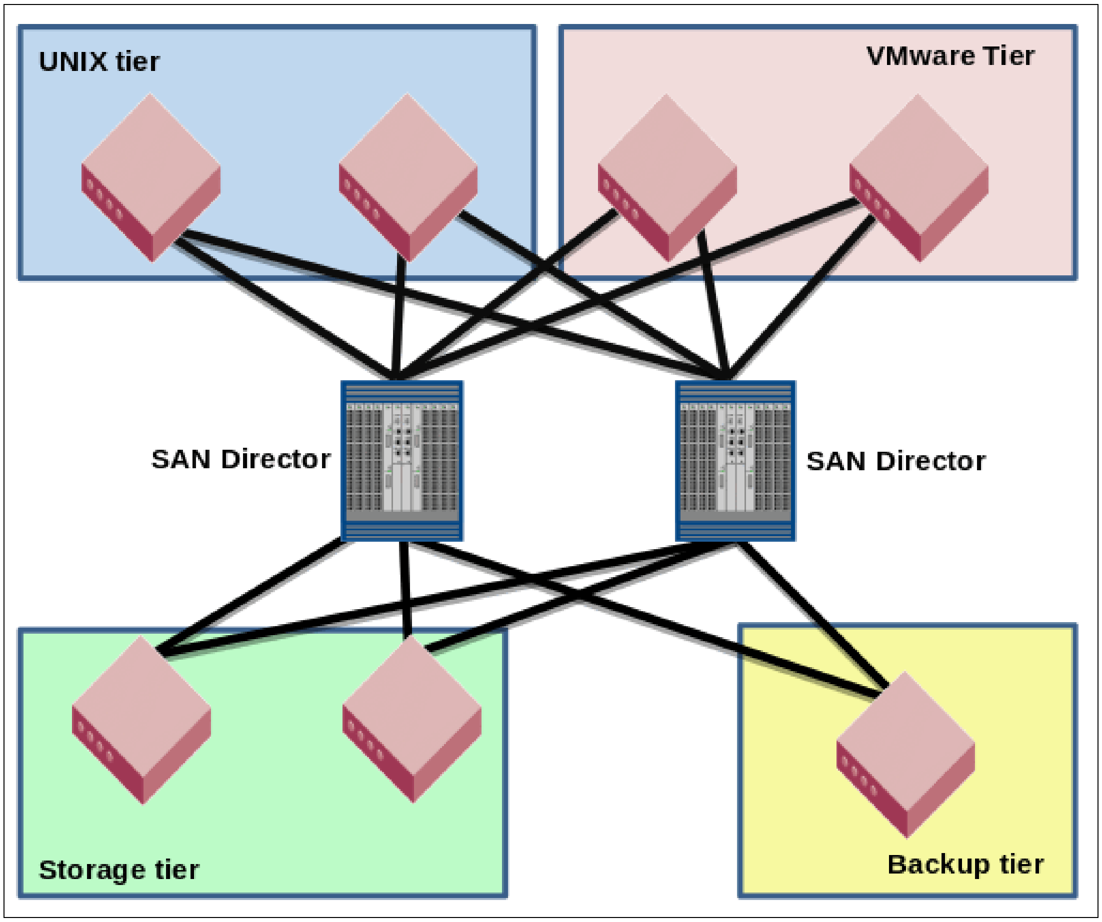


Figure 2-8 Segregation using edge-core-edge

Even when sharing the same core switches, it is possible to use virtual switches (see 2.2.7, "SAN partitioning" on page 47 for details) to isolate one tier from the other. This configuration helps avoid traffic congestion caused by slow drain devices that are connected to the backup tier switch.

2.2.7 SAN partitioning

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

Hardware-level fabric isolation is accomplished through the concept of switch virtualization, which allows you to partition physical switch ports into one or more "virtual switches." Virtual switches are then connected to form virtual fabrics.

As the number of available ports on a switch continues to grow, partitioning switches allow storage administrators to take advantage of high port density switches by dividing physical switches into different virtual switches. From a device perspective, SAN partitioning is completely transparent; therefore, the same guidelines and practices that apply to physical switches apply also to the virtual guidelines and practices.

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

Note: When director-class switches are used, use ports from different blades to seek load balance and avoid single point of failure.

For more information about IBM FlashSystem business continuity solutions, see Chapter 7, “Ensuring business continuity” on page 357.

2.3 IBM FlashSystem 9200 controller ports

Port connectivity options are significantly increased with IBM FlashSystem 9200 hardware. Models 9846-AG8 and 9848-AG8 deliver up to 12x16 Gb or 12x32GbFC ports per node canister, as listed in Table 2-1.

Table 2-1 FlashSystem 9200

Feature	FlashSystem 9200
Fibre Channel HBA	3x Quad 16 Gb or 3x Quad 32Gb
Ethernet I/O	2x Dual 25Gb iWARP/RoCE for iSCSI or iSER
Built in ports	4x 10 Gb for internet small computer systems interface (iSCSI)
Serial attached SCSI (SAS) expansion ports	1x Quad 12 Gb SAS (2 ports active)

Note: FlashSystem 9200 node canisters feature three peripheral component interconnect express (PCIe) slots that you can combine the cards as needed. If expansions are used, one of the slots must have the SAS expansion card. Then, two ports are left for Fibre Channel Host Bus Adapter (HBA) cards, Internet Wide-area RDMA Protocol (iWARP) or RDMA over Converged Ethernet (RoCE) Ethernet cards.

For more information, see this IBM Documentation web page.

This section describes some preferred practices and use cases that show how to connect a FlashSystem on the SAN to use this increased capacity.

2.3.1 Slots and ports identification

The IBM FlashSystem 9200 can have up to three quad Fibre Channel (FC) HBA cards (12 FC ports) per node canister. Figure 2-9 shows the port location in the rear view of the FlashSystem 9200 node canister.

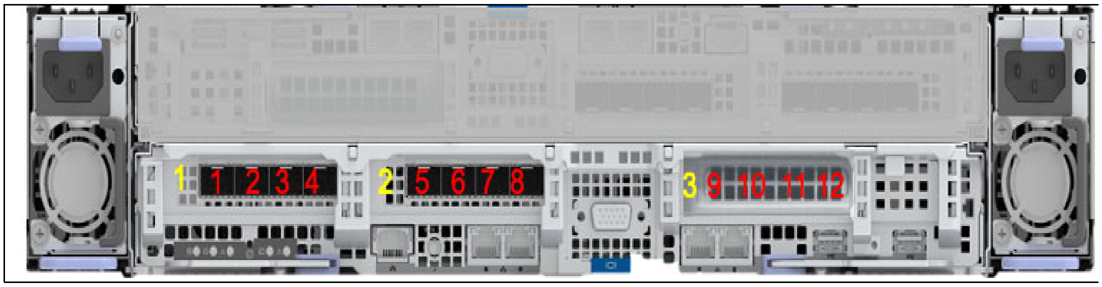


Figure 2-9 Port location in FlashSystem 9200 rear view

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

2.3.2 Port naming and distribution

In the field, fabric naming conventions vary. However, it is common to find fabrics with names such as `PROD_SAN_1` and `PROD_SAN_2`, or `PROD_SAN_A` and `PROD_SAN_B`. This type of naming convention is used to simplify the management and troubleshooting, after their denomination followed by *1* and *2* or *A* and *B*, which specifies that the devices connected to those fabrics contains the redundant paths of the same servers and SAN devices.

To simplify the SAN connection identification and troubleshooting, keep all odd ports on the odd fabrics, or “A” fabrics and the even ports on the even fabric or “B” fabrics, as shown in Figure 2-10.

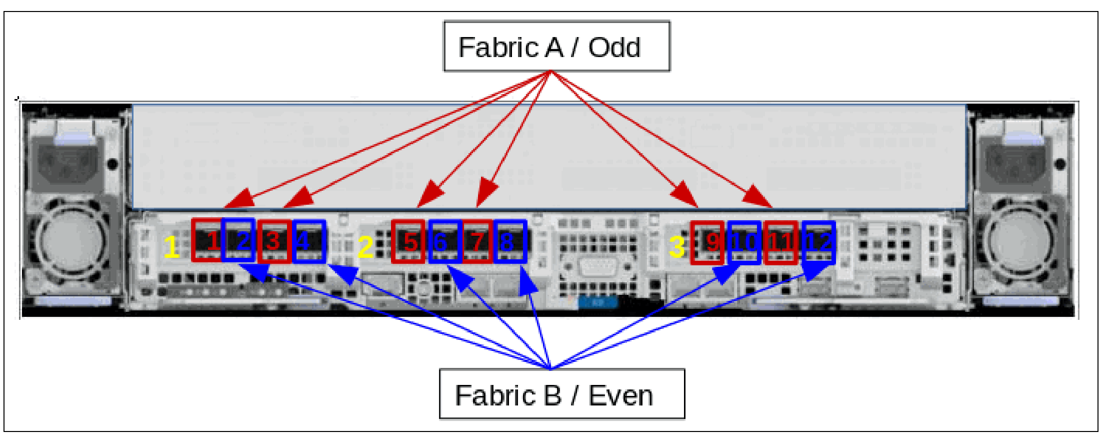


Figure 2-10 FlashSystem 9200 port distribution

As a preferred practice, assign specific uses to specific FlashSystem 9200 ports. This technique helps to optimize the port utilization by aligning the internal allocation of hardware CPU cores and software I/O threads to those ports.

Figure 2-11 shows the specific port use guidelines for the FlashSystem 9200.

Card / Port	4 ports	8 ports	12 ports
Card 1 Port 1	Host/Storage/Inter-node	Host/Storage	Host/Storage
Card 1 Port 2	Host/Storage/Inter-node	Host/Storage	Host/Storage
Card 1 Port 3	Host/Storage/Replication*	Inter-node	Inter-node
Card 1 Port 4	Host/Storage/Replication*	Inter-node	Inter-node
Card 2 Port 1		Host/Storage	Host/Storage
Card 2 Port 2		Host/Storage	Host/Storage
Card 2 Port 3		Host/Storage/Replication*	Host/Storage/Replication*
Card 2 Port 4		Host/Storage/Replication*	Host/Storage/Replication*
Card 3 Port 1			Host/Storage
Card 3 Port 2			Host/Storage
Card 3 Port 3			Host/Storage
Card 3 Port 4			Host/Storage
localfcportmask	0011	00001100	000000001100
partnerfcportmask	1100	11000000	000011000000

* Use for host/storage in case no replication is in place.
** Do not use the same port for replication and inter-node traffic.
*** For HyperSwap, dedicate ports for inter-node traffic

Figure 2-11 Port masking configuration on FlashSystem 9200

Note: Consider the following points:

- If you use an IBM FlashSystem 9200 system with a single I/O group, the system keeps `localfcportmask` set to `111111111111` and does not allow you to change this setting. However, this issue is not a problem because the inter-node traffic occurs on the internal PCI mid-plane link. The port-masking recommendations that are shown in Figure 2-11 apply to systems with more than one I/O group.
- Depending on the workload or number of I/O groups, you can reserve ports 1 and 2 from card 3 for inter-node traffic. In this case, ports are available for inter-node traffic, and the `localfcportmask` is set to `001100001100`.

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

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

Buffer credits

FlashSystem 9200 has a predefined number of buffer credits. The number of buffer credits determines the available throughput over distances: 4-port 16 Gbps adapters have 40 credits available per port, saturating links at up to 5 km (3.1 miles) at 16 Gbps.

Notes: Consider the following points:

- For stretched cluster and IBM HyperSwap configurations that do note use ISLs for the internode communication, set the switch port buffer credits to match the IBM FlashSystem 9200 port.
- Balance your bandwidth and make sure you have enough incoming bandwidth to saturate the back-end bandwidth. Attempt to balance your I/O across the ports.

2.4 Zoning

This section describes the zoning recommendations for FlashSystem 9200. For more information about external storage virtualization zoning, see *IBM System Storage SAN Volume Controller, IBM Storwize V7000,* and *IBM FlashSystem 7200 Best Practices and Performance Guidelines*, SG24-7521 (the recommendations are the same).

Important: Errors that are caused by incorrect FlashSystem 9200 zoning often are difficult to isolate and the steps to fix them can affect the SAN environment. Therefore, create your zoning configuration carefully.

The initial configuration for FlashSystem 9200 requires the following three zone types:

- Internode and intra-cluster zones
- Replication zones (if using replication)
- Host to FlashSystem 9200 zoning

Each zone type has its own guidelines, which are described in 2.4.1, “Types of zoning” on page 51.

Note: Although an internode or intra-cluster zone is not necessary for non-clustered FlashSystem 9200 family, it is generally preferred that one zone is used.

2.4.1 Types of zoning

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

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

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

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

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

Cisco Smart Zoning

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

For more information about Smart Zoning, see this Cisco web page.

For more information about implementation, see this IBM Support web page.

Brocade Peer Zoning

Brocade Peer Zoning is a feature that provides a similar functionality of restricting what devices can see other devices within the same zone. However, Peer Zoning is implemented such that some devices in the zone are designated as principal devices. The non-principal devices can only communicate with the principal device, not with each other.

As with Cisco, the communication is enforced in the fabric routing table. You can see more information about Peer Zoning on chapter 4.2.3 of *Modernizing Your IT Infrastructure with IBM b-type Gen 6 Storage Networking and IBM Spectrum Storage Products*, SG24-8415.

> **Note:** Use Smart and Peer zoning for the host zoning only. For intra-cluster, back-end, and replication zoning, use traditional zoning instead.

Simple zone for small environments

As an option for small environments, the IBM FlashSystem-based system supports a simple set of zoning rules that enable a small set of host zones to be created for different environments. For systems with fewer than 64 hosts that are attached, zones that contain host HBAs must contain no more than 40 initiators, including the ports that acts as initiators, such as the IBM Spectrum Virtualize-based system ports that are target + initiator.

Therefore, a valid zone can be 32 host ports plus 8 IBM FlashSystem based system ports. Include only one port from each node in the I/O groups that are associated with this host.

> **Note:** Do not place more than one HBA port from the same host in the same zone. Also, do not place dissimilar hosts in the same zone. Dissimilar hosts are hosts that are running different operating systems or are different hardware products.

For more information, see this IBM Documentation web page.

2.4.2 Pre-zoning tips and shortcuts

Several tips and shortcuts are available for FlashSystem 9200 zoning, as described in this section.

Naming convention and zoning scheme

When you create and maintain a FlashSystem 9200 zoning configuration, you must have a defined naming convention and zoning scheme. If you do not define a naming convention and zoning scheme, your zoning configuration can be difficult to understand and maintain.

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

For more information about FlashSystem 9200 naming convention, see 10.14.1, “Naming conventions” on page 506.

Aliases

Use zoning aliases when you create your FlashSystem 9200 zones if they are available on your specific type of SAN switch. Zoning aliases makes your zoning easier to configure and understand, and causes fewer possibilities for errors (see Table 2-2).

Table 2-2 Alias names examples

Port/WWPN	Use	Alias
Card 1 Port 1 physical WWPN	External Storage back-end	FS9200_N1P1_STORAGE
Card 1 Port 1 NPIV WWPN	Host attachment	FS9200_N1P1_HOST_NPIV
Card 1 Port 2 physical WWPN	External Storage back-end	FS9200_N1P2_STORAGE
Card 1 Port 2 NPIV WWPN	Host attachment	FS9200_N1P2_HOST_NPIV
Card 1 Port 3 physical WWPN	Inter-node traffic	FS9200_N1P3_CLUSTER
Card 1 Port 3 NPIV WWPN	No use	No alias
Card 1 Port 4 physical WWPN	Inter-node traffic	FS9200_N1P4_CLUSTER
Card 1 Port 4 NPIV WWPN	No use	No alias
Card 2 Port 3 physical WWPN	Replication traffic	FS9200_N1P7_REPLICATION
Card 2 Port 3 NPIV WWPN	No use	No alias
Card 2 Port 4 physical WWPN	Replication traffic	FS9200_N1P8_REPLICATION
Card 2 Port 4 NPIV WWPN	No use	No alias

Note: In Table 2-2, not all ports are used as example for aliases. Remember that NPIV ports can be used for host attachment only. If you are using external virtualized back-ends, use the physical port WWPN. For replication and inter-node, also use the physical WWPN. On the alias examples that are listed in Table 2-2, the *N* is for node, and all examples are from node 1. An N2 example is `FS9200_N2P4_CLUSTER`.

One approach is to create template zones from the host to the FlashSystem 9200. The zoning should contain one alias from the host. This alias must contain one initiator, and one alias from each node canister from the FlashSystem 9200 (preferably the same port).

As an example, create the following zone aliases:

- One zone alias for each FlashSystem 9200 port
- One alias for each host initiator
- One host initiator alias to FlashSystem 9200 port 1 from node 1, and to port 1 from node 2. Then, name this zone `HOST1_HBA1_T1_FS92009200`.
- (Optional) A second host initiator alias to FlashSystem 9200 port 3 from node 1 and to port 3 from node 2. Then, name this zone `HOST2_HBA1_T2_FS9200`.

By creating template zones, you keep the number of paths on the host side to four for each volume and a good workload balance among the FlashSystem 9200 ports. Table 2-3 shows how the aliases are distributed if you create template zones described in the example.

Table 2-3 Template examples

Template	FS9200 ports on Fabric A	FS9200 ports on Fabric B
T1	Node 1 port 1 Node 2 port 1	Node 1 port 2 Node 2 port 2
T2	Node 1 port 3 Node 2 port 3	Node 1 port 4 Node 2 port 4
T3	Node 1 port 5 Node 2 port 5	Node 1 port 6 Node 2 port 6
T4	Node 1 port 7 Node 2 port 7	Node 1 port 8 Node 2 port 8

Note: The number of templates varies depending on how many fiber ports are in your system, and how many are dedicated to host access. The port numbers are examples; you can use different ports depending on the number of HBA cards. Plan accordingly.

2.4.3 IBM FlashSystem 9200 internode communications zones

Internode (or intra-cluster) communication is critical to the stable operation of the cluster. The ports that carry internode traffic are used for mirroring write cache and metadata exchange between node canisters.

In the FlashSystem family, internode communication occurs primarily through the internal PCI connectivity between the two canisters of a control enclosure. However, for the clustered IBM FlashSystem, the internode communication requirements are similar to those requirements of the SAN Volume Controller.

To establish efficient, redundant, and resilient intra-cluster communication, the intra-cluster zone must contain at least two ports from each node/canister.

For FlashSystem 9200 clusters with two I/O groups or more with eight ports, isolate the intra-cluster traffic by dedicating node ports specifically to inter-node communication. The ports to be used for intra-cluster communication varies according to the port count. See Figure 2-12 on page 55 for port assignment recommendations.

Note: On NPIV-enabled configurations, use the physical WWPN for the intra-cluster zoning.

Only 16 port logins are allowed from one node to another node in a SAN fabric. Ensure that you apply the proper port masking to restrict the number of port logins. For FlashSystem 9200 clusters with two or more I/O groups, without port masking, any FlashSystem 9200 port and any member of the same zone can be used for intra-cluster communication. This includes the port members of FlashSystem 9200 that connect to host and external virtualized back-ends.

2.4.4 IBM FlashSystem 9200 host zones

The preferred practice to connect a host into a FlashSystem 9200 is to create a single zone to each host port. This zone must contain the host port and *one* port from each FlashSystem 9200 node canister that the host must access, as shown in Figure 2-12.

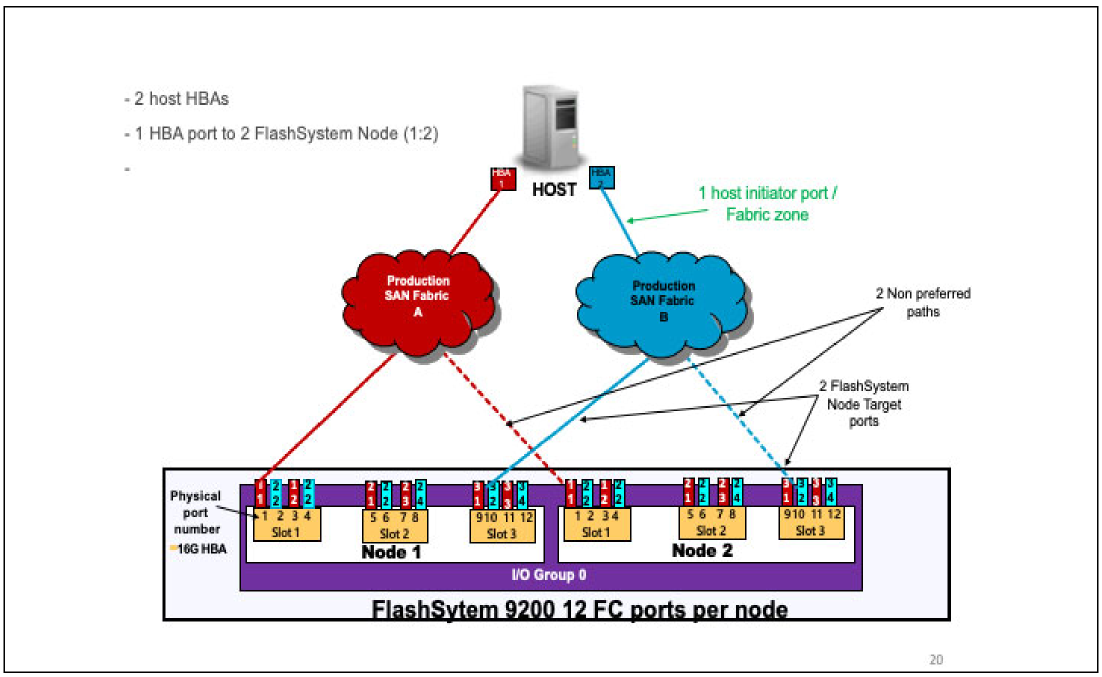


Figure 2-12 Typical host to FlashSystem 9200 zoning

This configuration provides four paths to each volume: two preferred paths (one per fabric) and two non-preferred paths. Multipathing software (such as AIXPCM, SDDDSM, VMWare NMP, and the FlashSystem 9200) is optimized to work such with four paths per volume.

NPIV consideration: The recommendations in this section also apply to NPIV-enabled configurations. For a list of the systems supported by the NPIV, see this IBM Support web page.

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

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

To obtain the best overall performance of the system and to prevent overloading, the workload to each FlashSystem 9200 port must be equal. Having the same amount of workload typically involves zoning approximately the same number of host FC ports to each FlashSystem 9200 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. Since eight paths is not an optimum number, configure your FlashSystem 9200 host definitions (and zoning) as though the single host is two separate hosts. During volume assignment, you alternate which volume was assigned to one of the “pseudo hosts.”

The reason for not assigning one HBA to each path is because the FlashSystem 9200 I/O group works as a cluster. When a volume is created, one node is assigned as preferred and the other node solely serves as a backup node for that specific volume. It means that using one HBA to each path will never balance the workload for that particular volume. Therefore, it is better to balance the load by I/O group instead so that the volume is assigned to nodes automatically.

Figure 2-13 shows an example of a four port host zoning.

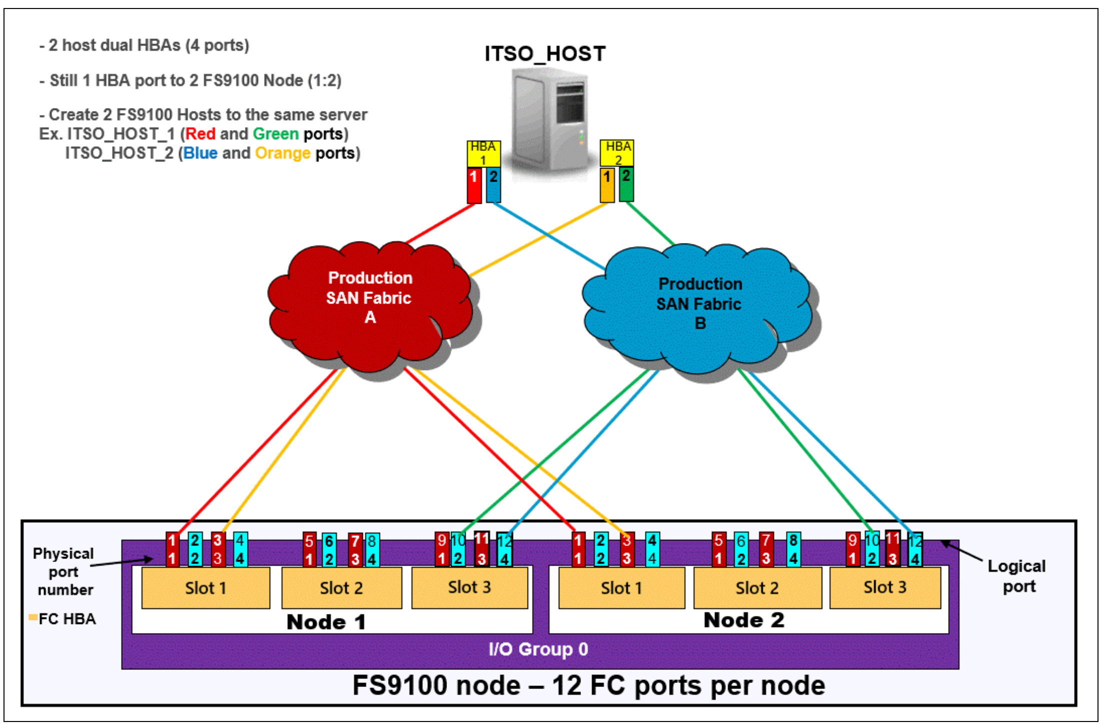


Figure 2-13 Four port host zoning

Because the optimal number of volume paths is four, you must create two or more hosts on FlashSystem 9200. During volume assignment, alternate which volume is assigned to each of the “pseudo-hosts” in a round-robin fashion.

Note: Pseudo-hosts is not a defined function or feature of SAN Volume Controller/ Storwize. To create a pseudo-host, you simply need to add another host ID to the SAN Volume Controller and Storwize host configuration. Instead of creating one host ID with four WWPNs, you define two hosts with two WWPNs, therefore you need to pay extra attention to the scsi ids assigned to each of the pseudo-hosts to avoid having 2 different volumes from the same storage subsystem with the same scsi id.

ESX Cluster zoning

For ESX Clusters, you must create separate zones for each host node in the ESX Cluster as shown in Figure 2-14.

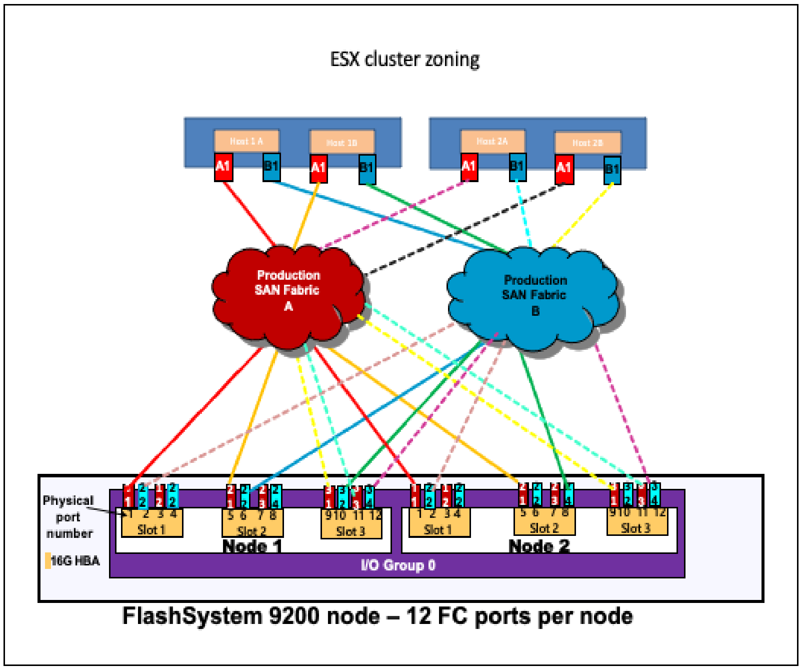


Figure 2-14 ESX Cluster zoning

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

- Zone a single ESX cluster in a manner that avoids ISL I/O traversing.
- Spread multiple host clusters evenly across the FlashSystem 9200 node ports and I/O groups.
- Create one host entity for each host node in FlashSystem 9200 and group them in a *hostcluster* entity.
- Create separate zones for each host node in FlashSystem 9200 and on the ESX cluster.

When you allocate a LUN or volume to a clustered system, use a host cluster on FlashSystem 9200. By doing so, your hosts have the same SCSI ID for every volume, which avoids outages that are caused by SCSI mismatch.

AIX VIOs: LPM zoning

When zoning IBM AIX® VIOs to IBM FlashSystem 9200, you must plan carefully. Because of its complexity, it is common to create more than four paths to each Volume or not provide for proper redundancy. The following preferred practices can help you to have a non-degraded path error on IBM Spectrum Virtualize/Storwize with four paths per volume:

- Create two separate and isolated zones on each fabric for each LPAR.
- Do not put the active and inactive LPAR WWPNs in the same zone or same IBM FlashSystem 9200 host definition.
- Map LUNs to the virtual host FC HBA port WWPNs, not the physical host FCA adapter WWPN.
- When using NPIV, generally make no more than a ratio of one physical adapter to eight Virtual ports. This configuration avoids I/O bandwidth oversubscription to the physical adapters.
- Create a pseudo host in IBM Spectrum Virtualize/FlashSystem host definitions that contain only two virtual WWPNs, one from each fabric as shown in Figure 2-15.

Figure 2-15 shows a correct SAN connection and zoning for LPARs.

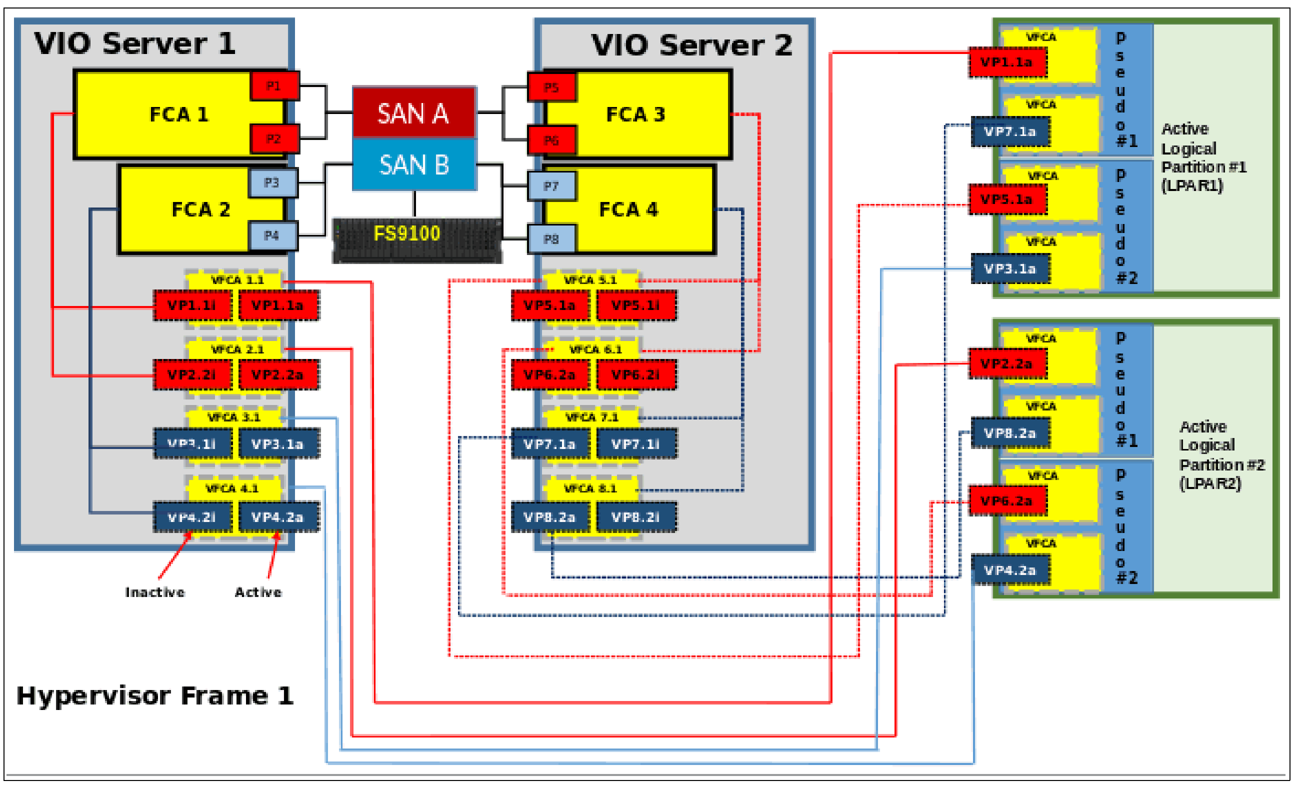


Figure 2-15 LPARs SAN connections

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

Figure 2-16 shows a Live partition migration from the hypervisor frame to another frame.

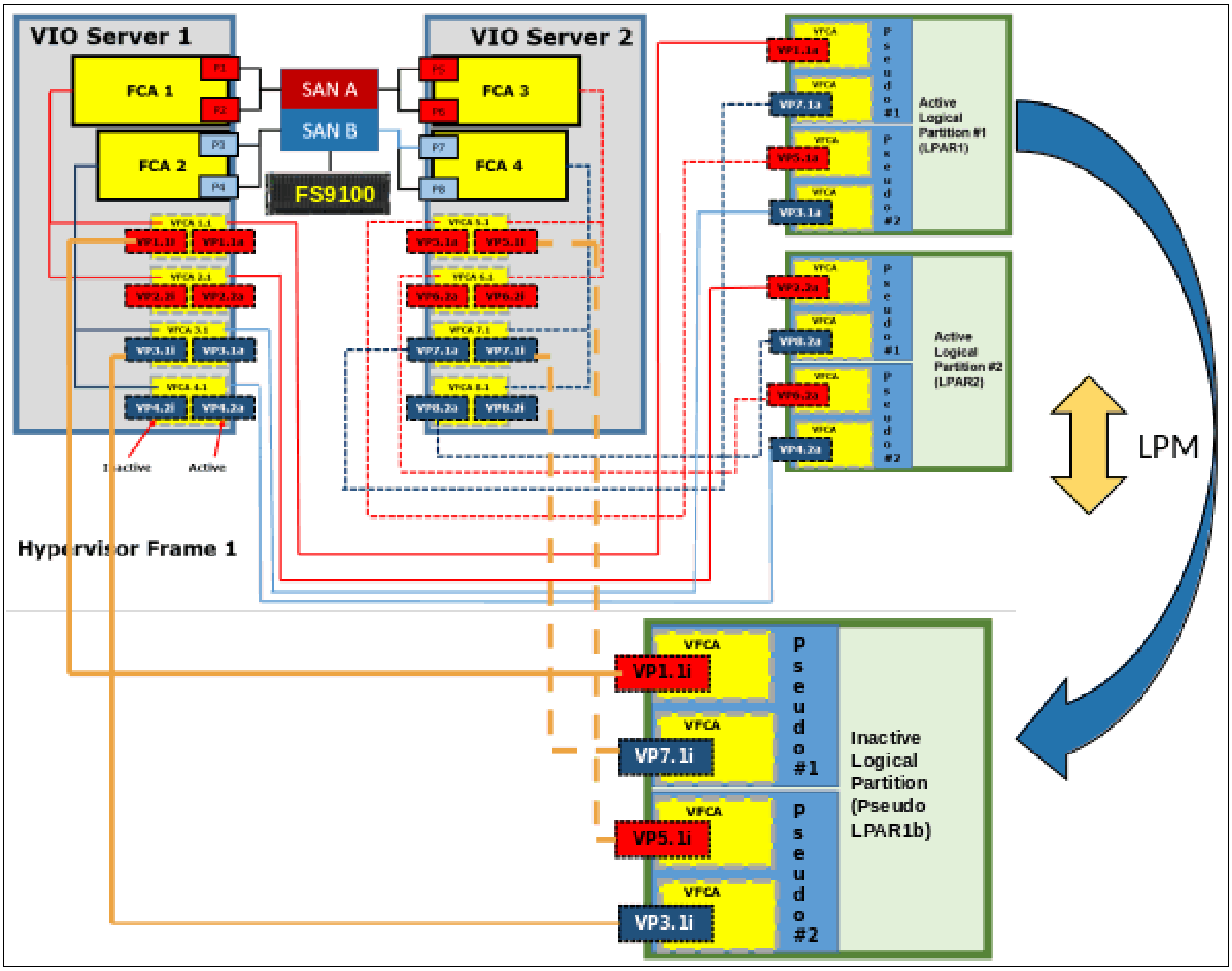


Figure 2-16 Live partition migration

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

2.5 Distance extension for Remote Copy services

To implement Remote Copy services over distance, the following options 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 Spectrum Virtualize 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.

2.5.1 Optical multiplexors

Optical multiplexors can extend your SAN up to hundreds of kilometers at high speeds. For this reason, they are the preferred method for long-distance expansion. When you are deploying optical multiplexing, make sure that the optical multiplexor is certified to work with your SAN switch model. The FlashSystem 9200 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 small form-factor pluggable (SFP) or Ten Gigabit Small Form Factor Pluggable (XFP) transceivers, so it is unlikely that Cisco SFPs will 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. FCIP is a technology that allows FC routing to be implemented over long distances by using the TCP/IP protocol. In most cases, FCIP is implemented in Disaster Recovery scenarios with some kind of data replication between the primary and secondary site.

FCIP is a tunneling technology, which means FC frames are encapsulated in the TCP/IP packets. As such, it is not apparent to devices that are connected through the FCIP link. To use FCIP, you need some kind of tunneling device on both sides of the TCP/IP link that integrates FC and Ethernet connectivity. Most of the SAN vendors offer FCIP capability through stand-alone devices (Multiprotocol routers) or by using blades that are integrated in the director class product. FlashSystem 9200 supports 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 is the only type of traffic. The design of FC is sensitive to congestion. You do not want a spyware problem or a DDOS attack on an IP network to disrupt your FlashSystem 9200.

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 approximately 15 MBps to your FlashSystem 9200. If you include the safety margins, this link is not as fast as you might hope; therefore, ensure that the terminology is correct.

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

- For redundancy purposes use as many TCP/IP links between sites as you have fabrics in each site that you want to connect. In most cases, there are two SAN FC fabrics in each site, so you need two TCP/IP connections between sites.
- Try to dedicate TCP/IP links only for storage interconnection. Separate them from other LAN/WAN traffic.
- Make sure that you have a service level agreement (SLA) with your TCP/IP link vendor that meets your needs and expectations.
- If you do not use Global Mirror with Change Volumes (GMCV), make sure that you have sized your TCP/IP link to sustain peak workloads.
- The use of FlashSystem 9200 internal Global Mirror (GM) simulation options can help you test your applications before production implementation. You can simulate the GM environment within one FlashSystem 9200 system without partnership with another. Use the **`chsystem`** command with the following parameters to perform GM testing:
  - **`gminterdelaysimulation`**
  - **`gmintradelaysimulation`**

  For more information about GM planning, see Chapter 6, "IBM FlashCopy services capabilities" on page 243.
- 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 FlashSystem 9200 to a value *lower* than your TCP/IP link capacity. Failing to set this rate can cause an unstable TCP/IP tunnel, which can lead to stopping all your Remote Copy relations that use that tunnel.
- The best case is to use GMCV when replication is done over long distances.
- Use compression on corresponding FCIP devices.
- Use at least two ISLs from your local FC switch to local FCIP router.
- On a Brocade SAN, use the Integrated Routing feature to avoid merging fabrics from both sites.

For more information about FCIP, see *IBM/Cisco Multiprotocol Routing: An Introduction and Implementation,* SG24-7543.

2.5.4 SAN extension with Business Continuity configurations

FlashSystem 9200 HyperSwap technology provides Business Continuity solutions over metropolitan areas with distances up to 300 km. Usually these solutions are achieved using SAN extension over WDM technology. Furthermore, multiple WDMs and physical links are implemented to avoid single points of failure. When implementing these solutions, particular attention must be paid in the intercluster connectivity set up.

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

Two possible configurations to interconnect the Private SANs are as follows:

- In Configuration 1, shown in Figure 2-17, one ISL per fabric is attached to each DWDM. In this case, the physical paths Path A and Path B are used to extend both fabrics.

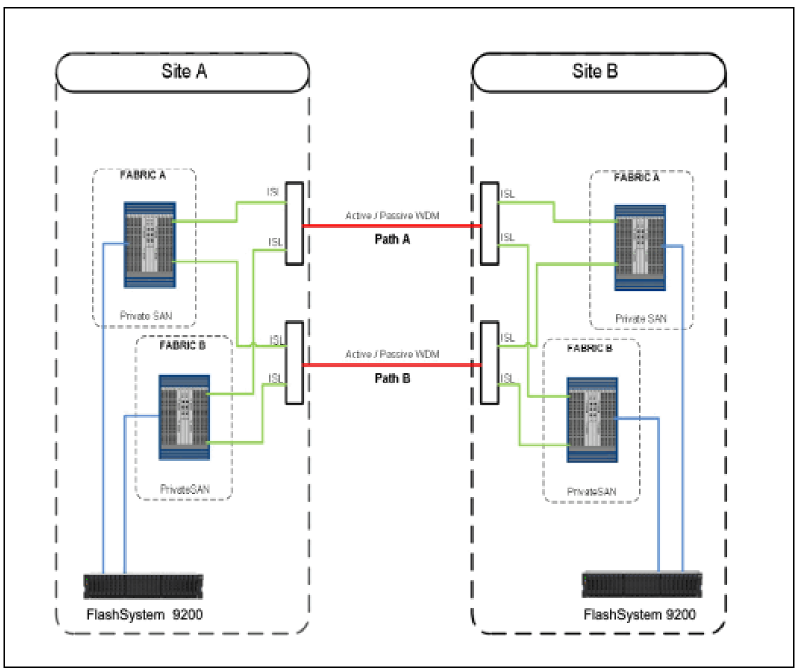


Figure 2-17 Configuration 1: Physical Paths Shared Among Fabrics

- In Configuration 2, shown in Figure 2-18, ISLs of fabric A are attached only to Path A, while ISLs of fabric B are attached only to Path B. In this case, the physical paths are not shared between the fabrics.

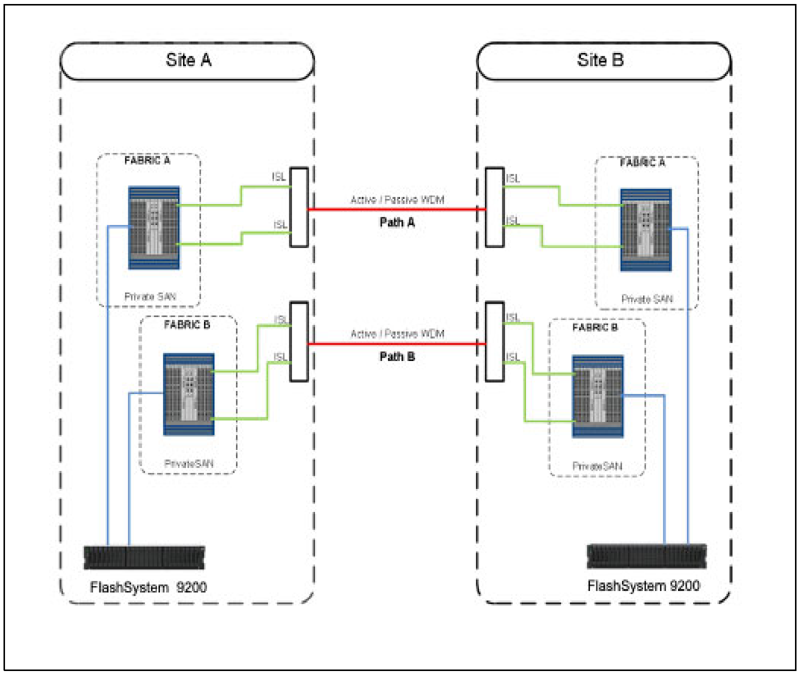


Figure 2-18 Configuration 2: physical paths not shared among the fabrics

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

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

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

2.5.5 Native IP replication

To enable native IP replication, FlashSystem 9200 implements the Bridgeworks SANSlide network optimization technology. For more information about this solution, see *IBM SAN Volume Controller and Storwize Family Native IP Replication*, REDP-5103.

It is possible to implement native IP-based replication on the FlashSystem 9200. *Native* means the FlashSystem 9200 does not need any FCIP routers to create a replication partnership. This partnership is based on the Internet Protocol network (IP) instead of the Fibre Channel (FC) network.

The main design point for the initial SANSlide implementation and subsequent enhancements, including the addition of replication compression is to reduce link utilization to allow the links to run closer to their respective line speed at distance and over poor quality links. IP replication compression does not significantly increase the effective bandwidth of the links beyond the physical line speed of the links.

If bandwidths are required that exceed the line speed of the physical links, alternative technologies that should be considered (such as FCIP), where compression is done in the tunnel and often yields an increase in effective bandwidth of 2:1 or more.

It is important to understand that the effective bandwidth of an IP link is highly dependent on latency and the quality of the link in terms of the rate of packet loss. Even a small amount of packet loss and resulting retransmits will significantly degrade the bandwidth of the link.

Figure 2-19 shows the effects distance and packet loss have on the effective bandwidth of the links in MBps. Numbers reflect pre-compression data rate with compression on and 50% compressible data. These numbers are as tested and can vary depending on specific link and data characteristics.

1G

	0ms	20ms	40ms	60ms	80ms
0%	122	108	61	41	30
0.1%	80	66	41	29	25
0.2%	59	44	30	24	20
0.5%	42	28	23	18	14
1%	31	21	17	14	13

10G

	0ms	1ms	2ms	5ms	10ms
0%	632	683	712	459	266
0.1%	120	120	115	117	87
0.2%	76	81	72	74	59
0.5%	49	48	41	41	35
1%	33	33	31	28	26

Figure 2-19 Effect of distance on packet loss

Notes: Consider the following points:

- The maximum bandwidth for a typical IP replication configuration that consists of two 1-Gb links is approximately 244 MBps at zero latency and zero packet loss.
- When two links are used, replication performs at twice the speed of the lower performing link. For example, the maximum combined data rate for two 1 Gb IP links at 0 latency and 0% packet loss on link A and 0.1% packet loss on link B is 160 MBps.
- 10 Gb links should not be used with latencies beyond 10 ms. Beyond 10 ms, a 1 Gb link begins to outperform a 10 Gb link.
- The FlashSystem 9200 supports volume compression. However, replication runs above volume compression in the IBM Spectrum Virtualize software stack, which means volumes are replicated at their full uncompressed capacity. This behavior differs from some storage systems, such as the IBM XIV, where replication runs below volume compression and therefore replicate the compressed capacity of the volumes. This difference must be considered when sizing workloads that are moving from one storage system technology to another.

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 FlashSystem 9200 SAN. However, do not put tape and disk traffic on the same FC HBA.

To avoid affecting ISL links and congestion on you SAN, do not put tape ports and backup servers on different switches; modern tape devices have high-bandwidth requirements.

During your backup SAN configuration, use the switch virtualization to separate the traffic type. Backup processes have different frames than production and can affect performance.

Backup requests often use all network resources available to finish writing on its destination target. Until the request is finished, the bandwidth is occupied and does not allow other frames access the network.

Figure 2-20 shows the difference between this two types of frames.

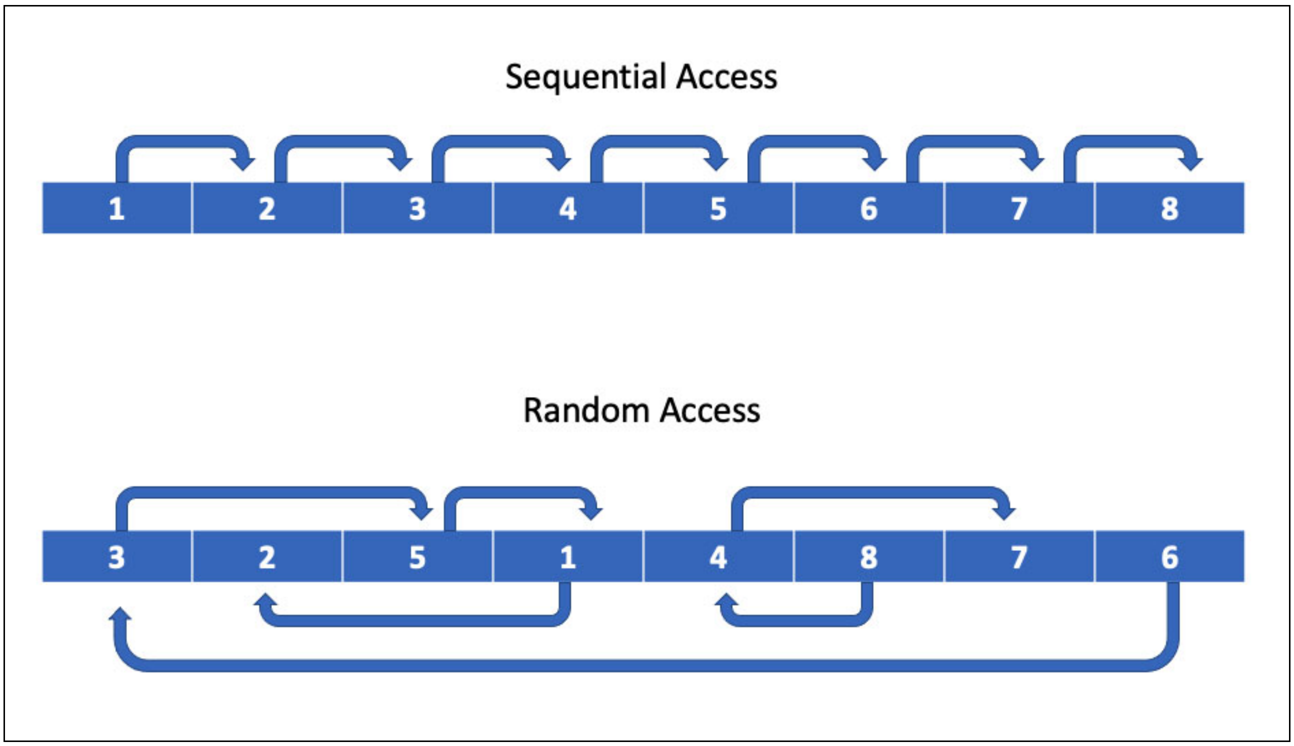


Figure 2-20 Access methods - FC Frames

Backup frames use the sequential method to write data (it releases only the path after it is done writing), while production frames write and read data randomly. Writing and reading is constantly occurring with the same physical path. If backup and production are setup on the same environment, production frames (read and write) can run only tasks when backup frames are complete, which causes latency on your production SAN network.

Figure 2-21 shows one example of a backup and production SAN configuration to avoid congestion that is caused by high bandwidth use from the backup process.

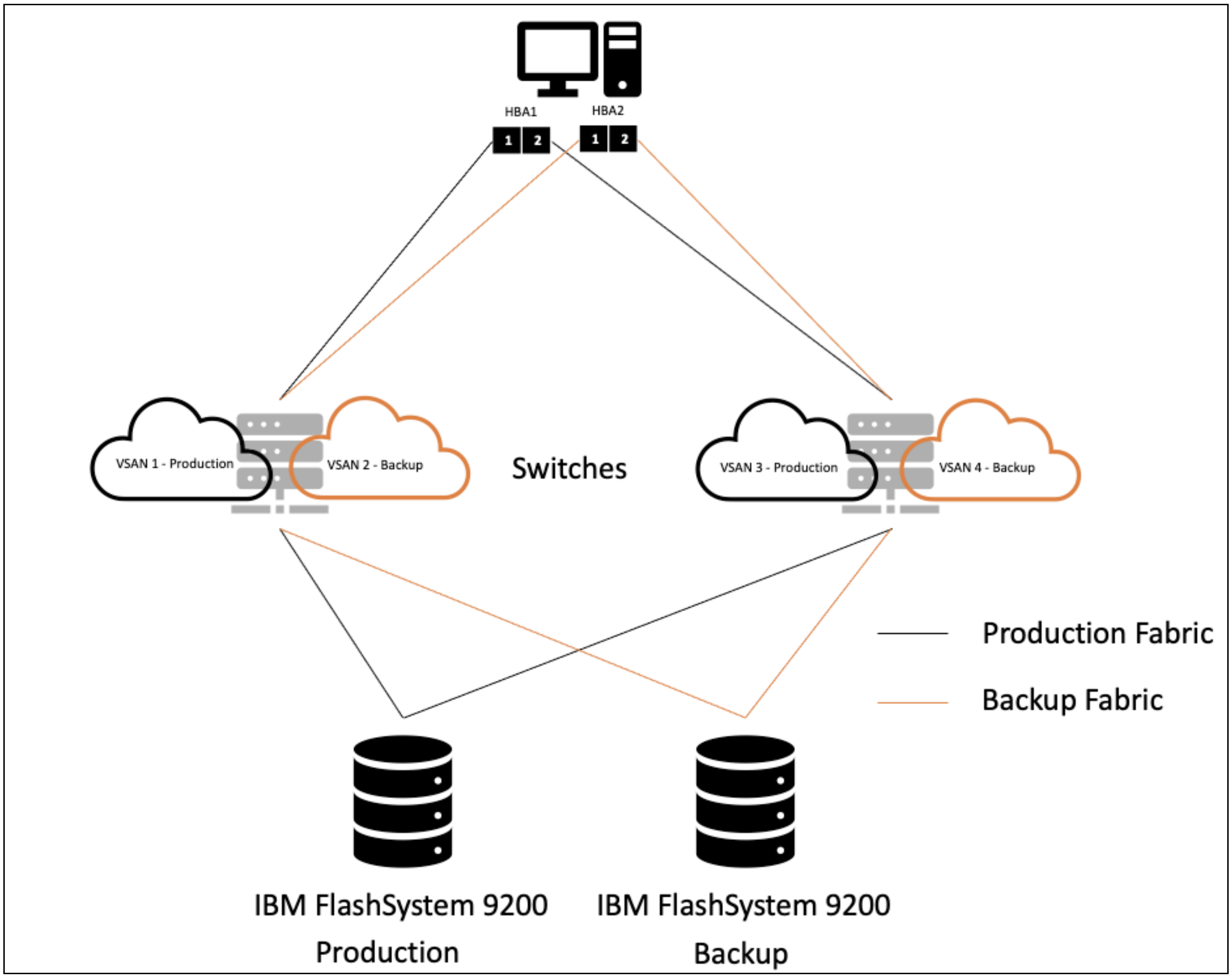


Figure 2-21 Production and Backup Fabric

2.7 Switch interoperability

> **Note:** For more information about interoperability, see IBM System Storage Interoperation Center (SSIC).

FlashSystem 9200 is flexible regarding switch vendors. All of the node canister connections on a specific FlashSystem 9200 single or 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.

FlashSystem 9200 supports combinations of SANs that are made up of switches from multiple vendors in the same SAN. However, this approach is not preferred in practice. Despite years of effort, interoperability among switch vendors is less than ideal because FC standards are not rigorously enforced. Interoperability problems between switch vendors are notoriously difficult and disruptive to isolate.

Also, it can take a long time to obtain a fix. For these reasons, run only multiple switch vendors in the same SAN long enough to migrate from one vendor to another vendor, if this setup is possible with your hardware.

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

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

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

3

Planning, configuring, and managing storage backend

This chapter describes the aspects and practices to consider when the internal and external back-end storage for a system is planned, configured, and managed:

- Internal storage consists of flash and disk drives that are installed in the control and expansion enclosures of the system.
- External storage is acquired by Spectrum Virtualize by virtualizing a separate IBM or third-party storage system, attached with Fibre Channel (FC) or iSCSI.

This chapter also provides information about traditional quorum disks. For information about IP Quorum, see Chapter 7, “Ensuring business continuity” on page 357.

This chapter contains the following sections:

3.1 Internal storage types

A system supports the following three types of devices attached with Non-Volatile Memory express (NVMe) protocol:

- Storage-class memory drives
- Industry-standard NVMe flash drives
- IBM FlashCore Modules

With a serial attached SCSI (SAS) attachment, flash (solid state) and spinning disk drives are supported. The set of supported drives depends on the platform.

3.1.1 NVMe storage

FlashSystem 5100, FlashSystem 7200, FlashSystem 9100, and FlashSystem 9200 control enclosures feature 24 x 2.5-inch slots to populate with NVMe storage. FlashSystem 5200 features 12 x 2.5-inch NVMe slots.

NVMe protocol

NVMe is an optimized, high-performance scalable host controller interface designed to address the needs of systems that utilize Peripheral Component Interconnect® Express (PCIe)-based solid-state storage. The NVMe protocol is an interface specification for communicating with storage devices. It is functionally analogous to other protocols, such as SAS. However, the NVMe interface was designed for extremely fast storage media, such as flash-based solid-state drives (SSDs) and low-latency non-volatile storage technologies.

NVMe storage devices are typically directly attached to a host system over a PCIe bus. That is, the NVMe controller is contained in the storage device itself, alleviating the need for an additional I/O controller between the CPU and the storage device. This architecture results in lower latency, throughput scalability, and simpler system designs.

NVMe protocol supports multiple I/O queues, versus older SAS and Serial Advanced Technology Attachment (SATA) protocols, which use only a single queue.

NVMe as a protocol, is similar to SCSI. It allows for discovery, error recovery, and read and write operations. However, NVMe uses RDMA over new or existing physical transport layers such as PCIe, Fibre Channel, or Ethernet. The major advantage of an NVMe-drive attachment is that this is usually by way of PCIe connectivity, thus the drives are physically connected to the CPU by way of a high-bandwidth PCIe connection, rather than using a “middle man”, such as a SAS controller chip which will limit total bandwidth to that available to the PCIe connection into the SAS controller. Where a SAS controller might have used 8 or 16 PCIe lanes in total, each NVMe drive has its own dedicated pair of PCIe lanes. This means a single drive can achieve data rates in excess of multiple GiB/s rather than hundreds of MiB/s when compared with SAS.

Overall latency can be improved by the adoption of larger parallelism and the modern device drivers used to control NVMe interfaces. For example, NVMe over Fibre Channel versus SCSI over Fibre Channel are both bound by the same Fibre Channel network speeds and bandwidths. However, the overhead on older SCSI device drivers (for example, reliance on kernel-based interrupt drivers) means that the software functionality in the device driver might limit its capability when compared with an NVMe driver. This is because an NVMe driver typically uses a polling loop interface, rather than an interrupt driven interface.

A polling interface is more efficient because the device itself looks for work to do and typically runs in user space (rather than kernel space). Therefore it has direct access to the hardware.

An interrupt-driven interface is less efficient because the hardware tells the software when it work must be done by pulling an interrupt line, which the kernel must process and then hand control of the hardware to the software. Interrupt-driven kernel drivers therefore waste time in switching between kernel and user space. As a result, all useful work is prevented from occurring on the CPU while the interrupt is handled. This adds latency and reduces total throughput, and therefore the amount of work done is bound by the work that a single CPU core can handle. Typically, a single hardware interrupt is owned by just one core.

All Spectrum Virtualize Fibre Channel and SAS drivers have always been implemented as polling drivers. Thus, on the storage side, almost no latency is saved when you switch from SCSI to NVMe as a protocol. However the above bandwidth increases are seen when a SAS controller is switched to a to PCIe-attached drive.

The majority of the advantages of using an end-to-end NVMe solution, when attaching to a Spectrum Virtualize based system, are seen as a reduction in the CPU cycles that are needed to handle the interrupts on the host server where the Fibre Channel HBA resides. Most SCSI device drivers remain interrupt driven, therefore switching to NVMe over Fibre Channel will result in the same latency reduction. CPU cycle reduction and general parallelism improvements have been enjoyed inside Spectrum Virtualize products since 2003.

Industry-standard NVMe drives

FlashSystem 5100, FlashSystem 5200, FlashSystem 7200, FlashSystem 9100, and FlashSystem 9200 control enclosures provide an option to use self-encrypting industry-standard (IS) NVMe flash drives, which are available with 800 GB - 15.36 TB capacity.

Supported IS NVMe SSD drives are built-in 2.5-inch form factor (SFF) and use a dual-port PCIe Gen3 interface to connect to the mid-plane.

Industry-standard NVMe drives start at a smaller capacity point than FCM drives, which allows for a smaller system.

NVMe FlashCore modules

At the heart of the IBM FlashSystem system is IBM FlashCore technology. IBM FlashCore Modules (FCMs) is a family of high-performance flash drives, that provide performance-neutral, hardware-based data compression and self-encryption.

FlashCore modules introduce the following features:

- Hardware-accelerated architecture that is engineered for flash, with a hardware-only data path
- Modified dynamic GZIP algorithm for data compression and decompression, implemented completely in drive hardware
- Dynamic SLC cache for reduced latency
- Cognitive algorithms for wear leveling and heat segregation

Variable stripe redundant array of independent disks (RAID) (VSR) stripes data across more granular, sub-chip levels. This allows for failing areas of a chip to be identified and isolated without failing the entire chip. Asymmetric wear-leveling understands the health of blocks within the chips and tries to place “hot” data within the healthiest blocks to prevent the weaker blocks from wearing out prematurely.

Bit errors caused by electrical interference are continually scanned for, and if any are found will be corrected by an enhanced Error Correcting Code (ECC) algorithm. If an error cannot be corrected, then the FlashSystem DRAID layer is used to rebuild the data.

NVMe FlashCore Modules use inline hardware compression to reduce the amount of physical space required. Compression cannot be disabled (and there is no reason to do that). If the written data cannot be compressed further, or compressing the data causes it to grow in size, the uncompressed data will be written. In either case, because the FCM compression is done in the hardware there will be no performance impact.

FlashSystem FlashCore Modules are not interchangeable with the flash modules that are used in FlashSystem 900 storage enclosures, as they have a different form factor and interface.

Modules that are used in FlashSystem 5100, 5200, 7200, 9100, and 9200 are built-in 2.5-inch U.2 dual-port form factor.

FCMs are available with physical, or usable capacity of 4.8, 9.6, 19.2, and 38.4 TB. The *usable capacity* is a factor of how many bytes the flash chips can hold.

They also have a maximum effective capacity (or virtual capacity), beyond which they cannot be filled. *Effective capacity* is the total amount of user data that can be stored on a module, assuming the compression ratio of the data is at least equal to (or higher than the ratio of effective capacity to usable capacity. Each FCM contains a fixed amount of space for metadata, and the maximum effective capacity is the amount of data it takes to fill the metadata space.

Module capacities are listed in Table 3-1.

Table 3-1 FlashCore module capacities

Usable capacity	Compression ratio at maximum effective capacity	Maximum effective capacity
4.8 TB	4.5: 1	21.99 TB
9.6 TB	2.3: 1	21.99 TB
19.2 TB	2.3: 1	43.98 TB
38.4 TB	2.3: 1	87.96 TB

A 4.8 TB FCM has a higher compression ratio because it has the same amount of metadata space as the 9.6 TB.

For more information about usable and effective capacities, see 3.1.3, “Internal storage considerations” on page 75.

As of this writing, IBM offers the second generation of IBM FlashCore modules, FCM2, which provides better performance and lower latency than FlashCore Module gen1. FCMs can be intermixed between generations within one system and within a single array. If needed, FCM1 can be replaced with an FCM2 of the same capacity.

Note: An array with intermixed FCM1 and FCM2 drives will perform like an FCM1 array.

Storage-class memory drives

Storage Class Memory (SCM) is a term that is used to describe non-volatile memory devices that perform faster (~10µs) than traditional NAND SSDs (100µs), but slower than DRAM (100ns).

IBM FlashSystem supports SCM drives that are built on two different technologies:

- 3D XPoint technology from Intel, developed by Intel and Micron (Intel Optane drives)
- zNAND technology from Samsung (Samsung zSSD)

Available SCM drive capacities are listed in Table 3-2.

Table 3-2 Supported SCM drive capacities

Technology	Small capacity	Large capacity
3D XPoint	350 GB	750 GB
zNAND	800 GB	1.6 TB

SCM drives have their own technology type and drive class in Spectrum Virtualize configuration. They cannot intermix in the same array with standard NVMe or SAS drives.

Due to their speed, SCM drives are placed in a new top tier, which is ranked higher than existing tier0_flash that is used for NVMe NAND drives.

A maximum of 12 Storage Class Memory drives can be installed per control enclosure.

Note: The maximum number can be limited to a lower number, depending on the platform and IBM Spectrum Virtualize code version.

3.1.2 SAS drives

IBM FlashSystem 5100, 5200, 7200, 9100, and 9200 control enclosures feature only NVMe drive slots, but systems can be scaled up by attaching SAS expansion enclosures with SAS drives.

IBM FlashSystem 5015 and 5035 control enclosures have 12 3.5-inch LFF or 24 2.5-inch SFF SAS drive slots. They also can be scaled up by connecting SAS expansion enclosures.

A single FlashSystem 5100, 7200, 9100, and 9200 control enclosure supports the attachment of up to 20 expansion enclosures with a maximum of 760 drives (748 drives for FlashSystem 5200), including NVMe drives in the control enclosure. By clustering control enclosures, the size of the system can be increased to a maximum of 1520 drives for FlashSystem 5100, 2992 drives for FlashSystem 5200, 3040 drives for FlashSystem 7200 and 9x00.

FlashSystem 5035 control enclosure supports up to 20 expansion enclosures with a maximum of 504 drives (including drives in the control enclosure). With two-way clustering, which is available for FlashSystem 5035, up to 1008 drives per system are allowed.

FlashSystem 5015 control enclosure supports up to 10 expansions and 392 drives maximum.

Expansion enclosures are designed to be dynamically added without downtime, helping to quickly and seamlessly respond to growing capacity demands.

The following types of SAS-attached expansion enclosures are available for the IBM FlashSystem family:

- 2U, 19-inch rack mount SFF expansion with 24 slots for 2.5-inch drives
- 2U, 19-inch rack mount LFF expansion with 12 slots for 3.5-inch drives (not available for FlashSystem 9x00)
- 5U, 19-inch rack mount LFF high density expansion enclosure with 92 slots for 3.5-inch drives.

Different expansion enclosure types can be attached to a single control enclosure and intermixed with each other.

Note: Intermixing expansion enclosures with MT 2077 and MT 2078 is not allowed.

IBM FlashSystem 5035, 5100, 5200, 7200, 9100 and 9200 control enclosures have two SAS chains for attaching expansion enclosures. Aim to keep both SAS chains equally loaded. For example, when attaching ten 2U enclosures, connect half of them to chain 1 and the other half to chain 2.

IBM FlashSystem 5015 has only a single SAS chain.

The number of drive slots per SAS chain is limited to 368. To achieve this, you need four 5U high-density enclosures. Table 3-3 lists th maximum number of drives that are allowed when different enclosures are attached and intermixed. For example, if three 5U enclosures are attached to a chain, you cannot connect more than two 2U enclosures to the same chain, and get 324 drive slots as the result.

Table 3-3 Maximum number of drive slots per SAS expansion chain

5U expansions	**2U expansions**										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	0	24	48	72	96	120	144	168	192	216	240
1	92	116	140	164	188	212	236	260	--	--	--
2	184	208	232	256	280	304	--	--	--	--	--
3	276	300	324	--	--	--	--	--	--	--	--
4	368	--	--	--	--	--	--	--	--	--	--

IBM FlashSystem 5015 and 5035 node canisters have on-board SAS ports for expansions. IBM FlashSystem 5100, 5200, 7200, 9100, and 9200 need a 12 GB SAS interface card to be installed in both nodes of a control enclosure to attach SAS expansions.

Expansion enclosures can be populated with spinning drives (high-performance enterprise-class disk drives or high-capacity nearline disk drives) or with solid state (flash) drives.

A set of allowed drive types depends on the system:

- FlashSystem 9x00 is all-flash.
- Other members of the family can be configured as all-flash or hybrid. In hybrid configurations, different drive types can be intermixed inside a single enclosure.

Drive capacities vary from less than 1 TB to more than 30 TB.

3.1.3 Internal storage considerations

In this section, we discuss the practices that must be considered when planning and managing IBM FlashSystem internal storage.

Planning for performance and capacity

With IBM FlashSystem 5100, 5200, 7200, 9100 and 9200, SAS enclosures are used to scale capacity within the performance envelope of a single controller enclosure. Clustering multiple control enclosures scales performance with the extra NVMe storage.

For best performance results, plan to operate your storage system with 85% or less physical capacity used. Flash drives depend on free pages being available to process new write operations and to quickly process garbage collection.

Without some level of free space, the internal operations to maintain drive health and host requests might over-work the drive, which causes the software to proactively fail the drive, or a hard failure might occur in the form of the drive becoming write-protected (zero free space left).

> **Note:** For more information about physical flash provisioning, see this IBM Support web page.

Intermix rules

Drives of the same form factor and connector type can be intermixed within an expansion enclosure.

For systems that support NVMe drives, NVMe and SAS drives can be intermixed in the same system. However, NVMe drives can exist only in the control enclosure, and SAS drives can exist only in SAS expansion enclosures.

Within a NVMe control enclosure, NVMe drives of different types and capacities can be intermixed: industry-standard NVMe drives and SCMs can be intermixed with FlashCore modules.

For more information about rules for mixing different drives in a single DRAID array, see “Drive intermix rules” on page 82.

Formatting

Drives and FlashCore modules must be formatted before they can be used. Format is important because when array is created, and its members must have zero used capacity. Drives automatically format when being changed to candidate.

An FCM is expected to format in under 70 seconds. Formatting an SCM drive takes much longer than an FCM or IS NVMe drive. On Intel Optane, drive formatting can take 15 minutes.

While a drive is formatting, it appears as an offline candidate. If you attempt to create an array before formatting is complete, the create command is delayed until all formatting is done. Aft3r formatting is done, the command completes.

If a drive fails to format, it goes offline. In this case, a manual format is required to bring it back online. The command line interface (CLI) scenario is shown in Example 3-1.

Example 3-1 Manual FCM format

```
IBM_FlashSystem:FS9100-ITSO:superuser>lsdrive | grep offline
id status  error_sequence_number use       tech_type ....
13 offline 118                   candidate tier0_flash ....
IBM_FlashSystem:FS9100-ITSO:superuser>chdrive -task format 13
```

Securely erasing

All SCM, FCM, and IS NVMe drives that are used in the system are self-encrypting. For SAS drives, encryption is performed by SAS chip in control enclosure.

For IS NVMe drives, SCMs and FCMs, formatting the drive completes a cryptographic erase of the drive. After the erasure, the original data on that device becomes inaccessible and cannot be reconstructed.

To securely erase SAS or NVMe drive, use the **chdrive -task erase <*drive_id*>** command.

The methods and commands that are used to securely delete data from drives enable the system to be used in compliance with European Regulation EU2019/424.

Monitoring FCM capacity

The IBM FlashSystem GUI (as shown in Figure 3-1) and CLI (as shown in Example 3-2) allows you to monitor effective and physical capacity for each FCM.

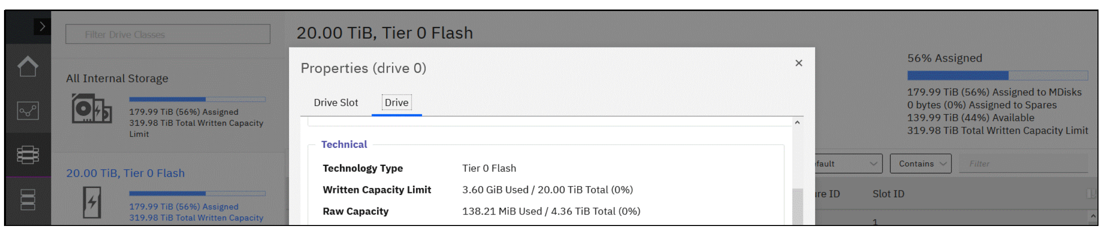


Figure 3-1 FCM capacity monitoring with GUI

Example 3-2 FCM capacity monitoring with CLI

```
IBM_FlashSystem:FS9100-ITSO:superuser>lsdrive 0
id 0
...
tech_type tier0_flash
capacity 20.0TB
...
write_endurance_used 0
write_endurance_usage_rate
replacement_date
transport_protocol nvme
compressed yes
physical_capacity 4.36TB
physical_used_capacity 138.22MB
effective_used_capacity 3.60GB
```

Both examples show same 4.8 TB FCM with maximum effective capacity of 20 TiB (or 21.99 TB).

To calculate actual compression ratio, divide the effective used capacity by the physical used capacity. Here, we have 3.60/0.134 = 26.7; therefore, written data is compressed 26.7:1 (highly compressible).

Physical used capacity is expected to be nearly the same on all modules in one array.

When FCMs are used, data compression ratios should be thoroughly planned and monitored.

If highly compressible data is written to an FCM, it still becomes full when it reaches the maximum effective capacity. Any spare data space remaining at this point is used to improve the performance of the module and extend the wear.

Example: A total of 20 TiB of data that is compressible 10:1 is written to a 4.8 TB module.

The maximum effective capacity of the module is 21.99 TB, which equals 20 TiB.

The usable capacity of the module is 4.8 TB = 4.36 TiB.

After 20 TiB of data is written, the module is 100% full for the array because it has no free effective (logical) capacity. At the same time, the data uses only 2 TiB of the physical capacity. The remaining 2.36 TiB cannot be used for host writes, only for drive internal tasks and to improve the module’s performance.

If non-compressible or low-compressible data is written, the module fills until the maximum physical capacity is reached.

Example: A total of 20 TiB of data that is compressible 1.2:1 is written to a 19.2 TB module.

The module’s maximum effective capacity is 43.99 TB, which equals 40 TiB. The module’s usable capacity is 19.2 TB = 17.46 TiB.

After 20 TiB is written, only 50% of effective capacity is used. With 1.2:1 compression, it occupies 16.7 TiB of physical capacity, which makes the module physically 95% full, and potentially affects the module’s performance.

Pool-level and array-level warnings can be set to alert and prevent compressed drive overfill.

DWPD

Drive Writes Per Day (DWPD) is a term that is used to express the number of times that the total capacity of a drive can be written per day within its warranty period. This metric shows drive write endurance.

If the drive write workload is continuously higher than the specified DWPD, the system will alert that the drive is wearing faster than expected. As DWPD is taken into account during system sizing, it usually means that workload differs from what was expected on the given array and it needs to be revised.

DWPD numbers are important with SSD drives of smaller sizes. With drive capacities below 1 TB, it is possible to write the total capacity of a drive several times a day. When a single SSD provides tens of TBs, it is unlikely that you can overrun the DWPD measurement.

Therefore, the DWPD measurement is less relevant for FCMs and large SSDs.

Consider the following points:

- SAS-attached Tier1 flash drives support up to 1 DWPD, which means that full drive capacity can be written on it every day and it lasts the five-year lifecycle.

  > **Example:** A total of 3.84 TB RI SAS drive is rated for 1 DWPD, which means 3840000 MB of data can be written on it each day. Each day has 24x60x60 = 86400 seconds; therefore, 3840000/86400 = 44.4 MBps of average daily write workload is required to reach 1 DWPD.
  >
  > Total cumulative writes over a 5-year period are 3.84 x 1 DWPD x 365 x 5 = 6.8 PB.

- FCM2 drives are rated with two DWPD over five years, which is measured in usable capacity. Therefore, if data is compressible (for example, 2:1), the DWPD doubles.

  > **Example:** A total of 19.2 TB FCM is rated for 2 DWPD. Its effective capacity is nearly 44 TB = 40 TiB, so considering 2.3:1 compression, to reach DWPD limit average daily workload over 5 years must be around 1 GBps. Total cumulative writes over a 5-year period are more than 140 PB.

- SCM drives are rated with 30 DWPD over five years.

System monitors amount of writes for each drive that supports DWPD parameter, and it logs a waning event if this amount is above than DWPD for the specific drive type.

It is acceptable to see `Write endurance usage rate is high` warnings, which indicate that write data rate exceeds expected for the drive type, during the initial phase of system implementation or during stress-testing. Afterwards, when system's workload is stabilized, the system recalculates usage rate and removes the warnings. Calculation is based on a long-run average; therefore, it can take up to one month for them to be automatically cleared.

3.2 Arrays

To use internal IBM FlashSystem drives in storage pools and provision their capacity to hosts, the drives must be joined into RAID arrays to form array-type MDisks.

3.2.1 Supported RAID types

RAID provides the following key design goals:

- Increased data reliability
- Increased I/O performance

The IBM FlashSystem supports the following RAID types:

- Traditional RAID (TRAID)

  In a traditional RAID approach, data is spread amongst drives in an array. However, the spare space is constituted by spare drives, which sit outside of the array. Spare drives are idling and do not share I/O load that comes to an array.

  When one of the drives within the array fails, all data is read from the mirrored copy (for RAID 10), or is calculated from remaining data stripes and parity (for RAID 5 or RAID 6), and then, written to a single spare drive.

- Distributed RAID (DRAID)

  With distributed RAID (DRAID), spare capacity is used instead of the idle spare drives from a traditional RAID. The spare capacity is spread across the disk drives. Because no drives are idling, all drives contribute to array performance.

  If a drive fails, the rebuild load is distributed across multiple drives. By spreading this load, DRAID addresses two main disadvantages of a traditional RAID approach: it reduces rebuild times by eliminating the bottleneck of one drive, and increases array performance by increasing the number of drives that are sharing the workload.

IBM FlashSystem implementation of DRAID allows effectively spread workload across multiple node canister CPU cores, which provides significant performance improvement over single-threaded traditional RAID arrays.

Table 3-4 lists RAID type and level support on different FlashSystem platforms.

Table 3-4 Supported RAID levels on FlashSystem platforms

	Non-distributed arrays (traditional RAID)				**Distributed arrays (DRAID)**		
	RAID 0	RAID 1 / 10	RAID 5	RAID 6	DRAID 1 / 10	DRAID 5	DRAID 6
FlashSystem 5010 / 5030	Yes	Yes	-	-	-	Yes	Yes
FlashSystem 5015 / 5035	-	-	-	-	Yes	Yes	Yes
FlashSystem 5100	Yes	Yes	-	-	-	Yes	Yes
FlashSystem 5200	-	-	-	-	Yes	Yes	Yes
FlashSystem 7200	Yes	Yes	-	-	Yes	Yes	Yes
FlashSystem 9100	Yes	Yes	-	-	-	Yes	Yes
FlashSystem 9200	Yes	Yes	-	-	Yes	Yes	Yes

Some drive types feature limitations and cannot be used in an array supported by a platform.

NVMe FlashCore Modules that are installed in IBM FlashSystem can be aggregated into DRAID 6, DRAID 5, or DRAID 1. All traditional RAID levels are not supported on FCMs.

SCM drives support DRAID levels 6 and 5, and DRAID 1, and TRAID 0.

Some limited RAID configurations do not allow large drives. For example, DRAID5 cannot be created with any drive type if drives capacities are equal or above 8 TB. Creating such arrays is blocked intentionally to prevent long rebuild times.

Table 3-5 lists the supported drives, array types, and RAID levels.

Table 3-5 Supported RAID levels with different drive types

Supported drives	**Non-distributed arrays (traditional RAID)**				**Distributed arrays (DRAID)**		
	RAID 0	RAID 1 /10	RAID 5	RAID 6	DRAID 1 / 10	DRAID 5	DRAID 6
SAS spinning drives	Yes	Yes	-	-	Yes *	Yes**	Yes
SAS flash drives	Yes	Yes	-	-	Yes	Yes**	Yes

Supported drives	Non-distributed arrays (traditional RAID)				Distributed arrays (DRAID)		
NVMe drives	Yes	Yes	-	-	Yes	Yes**	Yes
FlashCore Modules	-	-	-	-	Yes ***	Yes	Yes
SCM drives	Yes	Yes	-	-	Yes	Yes	Yes

* 3 or more spinning drives are required for DRAID 1, array with 2 members cannot be created with spinning drive type (while it can be created with other types). Also spinning drives larger than 8 TiB are not supported for DRAID.
** Drives with capacities of 8 TiB or more are not supported for DRAID 5
*** XL (38.4 TB) FCMs are not supported by DRAID 1

3.2.2 Array considerations

In this section, we discuss practices that must be considered when planning and managing drive arrays in IBM FlashSystem environment.

RAID level

Consider the following points when determining which RAID level to use:

- DRAID 6 is strongly recommended for all arrays with more than 6 drives.

  Traditional RAID levels 5 and 6 are not supported on the current generation of IBM FlashSystem, as DRAID is superior to them in all aspects.

  For most use cases, DRAID5 has no performance advantage compared to DRAID6. At the same time, DRAID6 offers protection from the second drive failure, which is vital as rebuild times are increasing together with the drive size. As DRAID6 offers the same performance level but provides more data protection, it is the top recommendation.

- On platforms that support DRAID 1, DRAID 1 is the recommended RAID level for arrays that consist of two or three drives.

  DRAID 1 has a mirrored geometry: it consists of mirrors of two strips, which are exact copies of each other. These mirrors are distributed across all array members.

- For arrays with four or five members, it is recommended use DRAID 1 or DRAID 5, which gives preference to DRAID 1 where it is available.

  DRAID 5 provides a capacity advantage over DRAID 1 with same number of drives, at the cost of performance. Particularly during rebuild, the performance of a DRAID 5 array is worse than that of a DRAID 1 array with the same number of drives.

- For arrays with six members, the choice is between DRAID 1 and DRAID 6.
- On platforms that support DRAID 1, do not use traditional RAID 1 or RAID 10 because they do not perform as well as distributed RAID type.
- On platforms that do not support DRAID 1, the recommended RAID level for NVMe SCM drives is TRAID10 for arrays of 2two drives, and DRAID 5 for arrays of four or five drives.
- RAID configurations that differ from the recommendations that are listed here are not available with the system GUI. If the wanted configuration is supported but differs from the these recommendations, arrays of required RAID levels can be created with the system CLI.

> **Note:** DRAID 1 arrays are supported only for pools with extent size of 1024 MiB or greater.

RAID geometry

Consider the following points when determining your RAID geometry:

- Data, parity, and spare space need to be striped across the number of devices available. The higher the number of devices, the lower the percentage of overall capacity the spare and parity devices will consume, and the more bandwidth that will be available during rebuild operations.

  Fewer devices are acceptable for smaller capacity systems that do not have a high-performance requirement, but solutions with a small number of large drives should be avoided. Sizing tools must be used to understand performance and capacity requirements.
- DRAID code makes full use of the multi-core environment, so splitting the same number of drives into multiple DRAID arrays does not bring performance benefits comparing to a single DRAID array with the same number of drives. Maximum system performance can be achieved from a single DRAID array. Recommendations that were given for traditional RAID, for example, to create four or eight arrays to spread load across multiple CPU threads, do not apply to DRAID.
- Consider the following guidelines to achieve best rebuild performance in a DRAID array:
  - For FCMs and IS NVMe drives, optimal number of drives in an array is 16 - 24. This limit ensures a balance between performance, rebuild times, and usable capacity. Array of NVMe drives cannot have more than 24 members.
  - For SAS HDDs, configure at least 40 drives to the array rather than create a large number of DRAID arrays with much less drives in each to achieve best rebuild times. A typical best benefit is approximately 48 - 64 HDD drives in a single DRAID6.
  - For SAS SSD drives, optimal array size is 24 - 36 drives per DRAID 6 array.
  - For SCM, maximum number of drives in an array is 12.
- Distributed spare capacity, or rebuild areas, are configured with the following guidelines:
  - DRAID 1 with two members: it is the only DRAID type, which is allowed without spare capacity (zero rebuild areas);
  - DRAID 1 with 3-16 members: array must have one rebuild area, it cannot have zero and cannot have more;
  - DRAID 5 or 6: minimum recommendation is one rebuild area per every 36 drives, optimal is one rebuild area per 24 drives.
  - Arrays with FCM drives cannot have more than one rebuild area per array.
- DRAID stripe width is set during array creation and indicates the width of a single unit of redundancy within a distributed set of drives. Note that reducing the stripe width will not enable the array to tolerate more failed drives. DRAID 6 will not get more redundancy than determined for level 6, independently of the width of a single redundancy unit.

  Reduced width increases capacity overhead, but also increases rebuild speed, as there is a smaller amount of data that RAID needs to read to reconstruct the missing data. For example, rebuild on DRAID with 14+P+Q geometry (width = 16) would be slower, or have a higher write penalty, than rebuild on DRAID with the same number of drives but 3+P+Q geometry (width = 5). In return, usable capacity for an array with width = 5 will be smaller than for an array with width = 16.

  Default stripe width settings (12 for DRAID6) provide an optimal balance between those parameters.
- Array strip size must be 256 KiB. With Spectrum Virtualize code releases before 8.4.x, it was possible to choose between 128 KiB and 256 KiB if DRAID member drive size was below 4 TB. From 8.4.x and above, you can only create arrays with 256 KiB strip size.

Arrays that were created on previous code levels, with strip size 128 KiB, are still fully supported.

- Stripe width and strip size (both) determine Full Stride Write (FSW) size. With FSW, data does not need to be read in a stride; therefore, the RAID I/O penalty is greatly reduced.

  For better performance, it is a good idea to set the host file system block size to the same value as the FSW size or a multiple of the FSW stripe size. However, IBM FlashSystem cache is designed to perform FSW whenever possible; therefore, no difference is noticed in the performance of the host in most scenarios.

  For fine-tuning for maximum performance, adjust the stripe width or host file system block size to match each other. For example, for a 2 MiB host file system block size, the best performance is achieved with 8+P+Q DRAID6 array (8 data disks x 256 KiB stripe size, array stripe width = 10).

Drive intermix rules

Consider the following points when intermixing drives:

- Compressing drives (FCMs) and non-compressing drives (SAS or NVMe) cannot be mixed in an array.
- SCM drives cannot be mixed in the same array with other types of NVMe or SAS devices.
- Physical and logical capacity:
  - For all types of NVMe drives: Members of an array must have the same physical and logical capacity. It is not possible to replace an NVMe drive with a "superior" NVMe drive (that is, one with a greater capacity),
  - For SAS drives: Members of an array do not need to have the same physical capacity. When creating an array on SAS drives, you can allow "superior" drives, or drives that have better characteristics than selected drive type. However, array capacity and performance in this case is determined by the base drive type.

    For example, if you have four 1.6 TB SSD drives and two 3.2 TB SSD drives, you still can create DRAID6 out of those six drives. Although 3.2 TB drives are members of this array, they use only half of their capacity (1.6 TB); the remaining capacity is never used.
- Mixing devices from different enclosures:
  - For NVMe devices, you cannot mix NVMe devices from different control enclosures in a system into one array.
  - For SAS drives, you can mix SAS drives from different control or expansion enclosures in a system into one array. One DRAID6 can span across multiple enclosures.

Drive failure and replacement

When a drive fails in a DRAID, arrays recover redundancy by rebuilding to spare capacity, which is distributed between all array members. After a failed drive is replaced, the array performs copyback, which uses the replaced drive and frees up the rebuild area.

DRAID distinguishes non-critical and critical rebuilds. If a single drive fails in a DRAID6 array, the array still has redundancy, and rebuild is performed with limited throughput to minimize the effect of the rebuild workload to an array's performance.

If an array has no more redundancy (which resulted from a single drive failure in DRAID5 or double drive failure in DRAID6), critical rebuild is performed. The goal of critical rebuild is to recover redundancy as fast as possible. Critical rebuild is expected to perform nearly twice as fast as non-critical.

When a failed drive that was an array member is replaced, the system includes it back to an array. For this, the drive must be formatted first, which might take some time for an FCM or SCM.

If the drive was encrypted in another array, it comes up as failed because this system does not have the required keys. The drive must be manually formatted to make it a candidate.

Note: An FCM drive that is a member of a RAID array must not be reseated unless directly advised to do so by IBM Support. Reseating FCM drives that are still in use by an array can cause unwanted consequences.

RAID expansion

Consider the following points for RAID expansion:

- You can expand distributed arrays to increase the available capacity. As part of the expansion, the system automatically migrates data for optimal performance for the new expanded configuration. Expansion is non-disruptive and compatible with other functions, such as IBM Easy Tier and data migrations.
- New drives are integrated and data is re-striped to maintain the algorithm placement of strips across the existing and new components. Each stripe is handled in turn, that is, the data in the existing stripe is redistributed to ensure the DRAID protection across the new larger set of component drives.
- Only the number of member drives and rebuild areas can be increased. RAID level and RAID stripe width stay as it was set during array creation.
- RAID-member count cannot be decreased: it is not possible to shrink an array.
- DRAID 5, DRAID 6 and DRAID 1 can be expanded. Traditional RAID arrays do not support expansion.
- Only one expansion process can run on array at one time. During a single expansion, up to 12 drives can be added.

  Only one expansion per storage pool is allowed, with a maximum of four per system.
- Once expansion is started, it cannot be canceled. You can only wait for it to complete or delete an array.
- As the array capacity increases, it becomes available to the pool as expansion progresses. There is no need to wait for expansion to be 100% complete, as added capacity can be used while expansion is still in progress.

  When you expand an FCM array, the physical capacity is not immediately available, and the availability of new physical capacity does not track with logical expansion progress.
- Array expansion is a process designed to run in background and can take significant time.

  It can affect host performance and latency, especially when expanding an array of spinning drives. Do not expand an array when the array has over 50% load. If you do not reduce host I/O load, the amount of time that is needed to complete the expansion increases greatly.
- Array expansion is not possible when an array is in write-protected mode because of being full (out of physical capacity). Any capacity issues must be resolved first.

- Creating a separate array can be an alternative for DRAID expansion.

  For example, if you have a DRAID6 array of 40 NLSAS drives and you have 24 new drives of the same type, the following options are available:

  – Perform two DRAID expansions by adding 12 drives in one turn. With this approach, the configuration is one array of 64 drives; however, the expansion process might take few weeks for large capacity drives, During that time, the host workload must be limited, which can be unacceptable.

  – Create a separate 24-drive DRAID6 array and add it to the same pool as a 40-drive array. The result is that you get two DRAID6 arrays with different performance capabilities, which is suboptimal. However, back-end performance aware cache and Easy Tier balancing can compensate for this flaw.

RAID capacity

Consider the following points when determining RAID capacity:

- If you are planning only your configuration, use the IBM Storage Modeller tool, which is available for IBM Business Partners.
- If your system is deployed, you can use the **`lspotentialarraysize`** CLI command to determine the capacity of a potential array for a specified drive count, drive class, and RAID level in the specified pool.
- To get an approximate amount of available space in DRAID6 array, use the following formula:

```
Array Capacity = D / ((W * 256) + 16) * ((N - S) * (W - 2) * 256)
Where:
D - Drive capacity
N - Drive count
S - Rebuild areas (spare count)
W - Stripe width
```

Example #1: Capacity of DRAID6 array out of 16 x 9.6 TB FlashCore modules, where:

- D = 9.6 TB = 8.7 TiB
- N = 16
- S = 1
- W = 12

Array capacity = 8.7 TiB / ((12*256)+16) * ((16-1) * (12-2) * 256) = 8.7 TiB / 3088 * 38400 = 108.2 TiB

3.2.3 Compressed array monitoring

DRAID arrays on FlashCore modules need to be carefully monitored and well-planned, as they are over-provisioned, which means they are susceptible to an out-of-space condition.

To minimize the risk of an out of space condition, ensure the following:

- The data compression ratio is known and taken into account when planning for array physical and effective capacity.
- Monitor array free space and avoid filling it up more than 85% of physical capacity.

  To monitor arrays, use IBM Spectrum Control or IBM Storage Insights with configurable alerts. For more information, see Chapter 9, “Implementing a storage monitoring system” on page 387.

IBM FlashSystem GUI and CLI will also display used and available effective and physical capacities. For examples, see Figure 3-2 and Example 3-3.

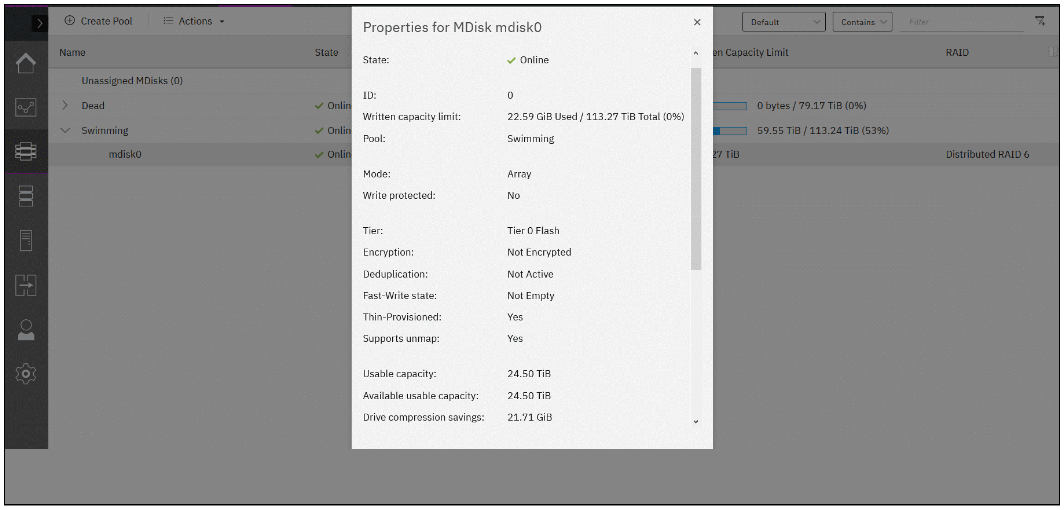


Figure 3-2 Array capacity monitoring with GUI

Example 3-3 Array capacity monitoring with CLI

```
IBM_FlashSystem:FS9100-ITSO:superuser>lsarray 0
mdisk_id 0
mdisk_name mdisk0
capacity 113.3TB
...
physical_capacity 24.50TB
physical_free_capacity 24.50TB
write_protected no
allocated_capacity 58.57TB
effective_used_capacity 22.59GB
```

- If the used physical capacity of the array reaches 99%, IBM FlashSystem raises event ID 1241: `1% physical space left for compressed array`. This event is a call for immediate action.

  To prevent running out of space, one or a combination of the following corrective actions must be taken:

  – Add storage to the pool and wait while data is balanced between arrays by Easy Tier.

  – Migrate volumes with extents on the managed disk that is running low on physical space to another storage pool or migrate extents from the array that is running low on physical space to other managed disks that have sufficient extents.

  – Delete or migrate data from the volumes using a host that supports UNMAP commands. IBM FlashSystem will issue UNMAP to the array and space will be released.

  For more information about out-of-space recovery, see this IBM Support web page.

- Arrays are most in danger of running out of space during a rebuild or when they are degraded. DRAID spare capacity, which is distributed across array drives, remains free during normal DRAID operation, thus reducing overall drive fullness. This means that if array capacity is 85% full, each array FCM is used for less than that due to spare space reserve. When DRAID is rebuilding this space becomes used.

  After the rebuild is complete, the extra space is filled up and the drives can be truly full, resulting in high levels of write amplification and degraded performance. In the worst case (for example, if the array is more than 99% full before rebuild starts), there is a chance that the rebuild might cause a physical out-of-space condition.

3.3 General external storage considerations

IBM FlashSystem can virtualize external storage and make it available to the system. External back-end storage systems (or `controllers` in Spectrum Virtualize terminology) provide their logical volumes (LUs), which are detected by IBM FlashSystem as MDisks and can be used in storage pools.

This section covers aspects of planning and managing external storage virtualized by IBM FlashSystem.

External back-end storage can be connected to IBM FlashSystem with FC (SCSI) or iSCSI. NVMe-FC back-end attachment is *not* supported because it provides no performance benefits for IBM FlashSystem. For more information, see “NVMe protocol” on page 70.

On IBM FlashSystem 5010/5030 and FlashSystem 5015/5035, virtualization is allowed only for data migration. Therefore, this systems can be used to externally virtualize storage as an image mode device for the purposes of data migration, *not* for long term virtualization.

3.3.1 Storage controller path selection

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

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

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

When the IBM FlashSystem node canisters select a set of ports to access the storage system, the two types of path selection described in the next sections are supported to access the MDisks. A type of path selection is determined by external system type and cannot be changed. To determine which algorithm is used for a specific back-end system, see System Storage Interoperaton Center (SSIC), as shown in Figure 3-3 on page 87.

Figure 3-3 SSIC example

Round-robin path algorithm

With the round-robin path algorithm, each MDisk uses one path per target port per IBM FlashSystem node. This means that in cases of storage systems without a preferred controller such as XIV or DS8000, each MDisk uses all of the available FC ports of that storage controller.

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, uses all available IBM FlashSystem ports.

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

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

> **Preferred practice:** If you have a storage system that supports the round-robin path algorithm, you should zone as many FC ports as possible from the back-end storage controller. IBM FlashSystem supports up to 16 FC ports per storage controller. See your storage system documentation for FC port connection and zoning guidelines.

Example 3-4 shows a storage controller that supports round-robin path selection.

Example 3-4 Round robin enabled storage controller

```
IBM_FlashSystem:FS9100-ITSO:superuser>lsmdisk 4
id 4
name mdisk4
...
preferred_WWPN 20010002AA0244DA
active_WWPN many                           <<< Round Robin Enabled
```

MDisk group balanced and controller balanced

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

With storage subsystems that use active-passive type systems, IBM FlashSystem accesses an MDisk LU through one of the ports on the preferred controller. To best use the back-end storage, it is important to make sure that the number of LUs that is created is a multiple of the connected FC ports and aggregate all LUs to a single MDisk group.

Example 3-5 shows a storage controller that supports MDisk group balanced path selection.

Example 3-5 MDisk group balanced path selection (no round robin enabled) storage controller

```
IBM_FlashSystem:FS9100-ITSO:superuser>lsmdisk 5
id 5
name mdisk5
...
preferred_WWPN
active_WWPN 20110002AC00C202           <<< indicates Mdisk group balancing
```

3.3.2 Guidelines for creating optimal backend configuration

Most of the backend controllers aggregate spinning or solid state drives into RAID arrays, then join arrays into pools. Logical volumes are created on those pools and provided to hosts. When connected to external backend storage, IBM FlashSystem acts as a host. It is important to create backend controller configuration that provides performance and resiliency, as IBM FlashSystem will rely on back-end storage when serving I/O to attached host systems.

If your back-end system has homogeneous storage, create the required number of RAID arrays (usually RAID 6 or RAID 10 are recommended) with equal number of drives. The type and geometry of an array depends on the back-end controller vendor's recommendations. If your back-end controller can spread the load stripe across multiple arrays in a resource pool (for example, by striping), create a single pool and add all arrays there.

On back-end systems with mixed drives, create a separate resource pool for each drive technology. Keep the drive-technology type in mind, as you will need to assign the correct tier for an MDisk when it is used by IBM FlashSystem.

Create a set of fully allocated logical volumes from the back-end system storage pool (or pools). Each volume is detected as MDisk on IBM FlashSystem. The number of logical volumes to create depends the type of drives, used by your back-end controller.

Back-end controller with spinning drives

If your back-end is using spinning drives, volume number calculation must be based on a queue depth. *Queue depth* is the number of outstanding I/O requests of a device.

For optimal performance, spinning drives need 8-10 concurrent I/O at the device, and this doesn't change with drive rotation speed. Make sure in a highly loaded system, that any given IBM FlashSystem MDisk can queue up approximately 8 I/O per back-end system drive.

IBM FlashSystem queue depth per MDisk is approximately 60. The exact maximum seen on a real system might vary depending on the circumstances. However, for the purpose of this calculation it does not matter.

The queue depth per MDisk number leads to the *HDD Rule of 8*. According to this rule, to achieve 8 I/O per drive and with queue depth 60 per MDisk from IBM FlashSystem, a back-end array with *60/8 = 7.5* that is approximately equal to eight physical drives is optimal, or we need one logical volume per every eight drives in an array.

> **Example:** The back-end controller to be virtualized is IBM Storwize V5030 with 64 NL-SAS 8 TB drives.
>
> The system is homogeneous. According to recommendations that are described in 3.2.2, "Array considerations" on page 80, create a single DRAID6 array at IBM Storwize and include it in a storage pool. By using the HDD rule of 8, we want 64/8 = 8 MDisks; therefore, create eight volumes from a pool to present to IBM FlashSystem and assign them to the nearline tier.

All-flash back-end controllers

For All-flash controllers, the considerations are more of I/O distribution across IBM FlashSystem ports and processing threads, than of queue depth per drive. Since most All-flash arrays that are put behind virtualizer have high I/O capabilities, make sure that IBM FlashSystem is given the optimal chance to spread the load and evenly make use of its internal resources, so queue depths are of less a concern here (because of the lower latency per I/O).

For all-flash back-end arrays, IBM recommends creating 32 logical volumes from the array capacity, as it allows to keep the queue depths high enough and spreads the work across the virtualizer resources. For smaller setups with a low number of solid state drives this number can be reduced to 16 logical volumes (which results in 16 MDisks) or even 8 volumes.

> **Example:** Back-end controllers to be virtualized are IBM FlashSystem 5035 with 24 Tier1 7.6 TB drives and IBM FlashSystem 9200. The virtualizer needs a pool with two storage tiers:
>
> - On IBM FlashSystem 5035, create a single DRAID6 array and add it to a storage pool. Using all-flash rule, we must create 32 volumes to present as MDisks. However, because it is small setup, we can reduce the number of volumes to 16.
> - On IBM FlashSystem 9200, join all micro-latency modules into a RAID5 array and add it to a storage pool. Because FlashSystem 9200 is Tier0 solution, use all-flash rule, and create 32 volumes to present as MDisks.
> - On virtualizer, add 16 MDisks from IBM FlashSystem 5035 as Tier1 flash, and 32 MDisks as Tier0 flash, to a single multi-tier pool.

Large setup considerations

For controllers like IBM DS8000 and XIV, you can use all-flash rule of 32. However, with installations involving this type of back-end controllers, it might be necessary to consider a maximum queue depth per back-end controller port, which is set to 1000 for most supported high-end storage systems.

With high-end controllers, queue depth per MDisk can be calculated by using the following formula:

```
Q = ((P x C) / N) / M
```

Where:

Q Calculated queue depth for each MDisk.

P Number of back-end controller host ports (unique WWPNs) that are zoned to IBM FlashSystem (minimum is 2 and maximum is 16).

C Maximum queue depth per WWPN, which is 1000 for controllers, such as XIV or DS8000.

N Number of nodes in the IBM FlashSystem cluster (2, 4, 6, or 8).

M Number of volumes that are presented by back-end controller and detected as MDisks.

For a result of `Q = 60`, calculate the number of volumes that is needed to create as `M = (P x C) / (N x Q)`, which can be simplified to `M = (16 x P) / N`.

> **Example:** A 4-node IBM FlashSystem 9200 is used with 12 host ports on the IBM XIV System.
>
> By using the previous formula, we must create `M = (16 x 12) / 4 = 48` volumes on IBM XIV to obtain a balanced high-performing configuration.

3.3.3 Considerations for compressing and deduplicating back-end

IBM FlashSystem supports over-provisioning on selected back-end controllers. This means that if back-end storage performs data deduplication or data compression on LUs provisioned from it, the LUs still can be used as external MDisks on IBM FlashSystem.

The implementation steps for thin-provisioned MDisks are the same as for fully allocated storage controllers. Extreme caution should be used when planning capacity for such configurations.

The IBM FlashSystem detects:

- If the MDisk is thin-provisioned.
- The total physical capacity of the MDisk.
- The used and remaining physical capacity of the MDisk.
- Whether **`unmap`** commands are supported by the back-end. By sending SCSI **`unmap`** commands to thin-provisioned MDisks, the system marks data that is no longer in use. Then, the garbage-collection processes on the back-end can free unused capacity and reallocate it to free space.

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

The nominal capacity from a compression and deduplication enabled storage system is not fixed and it varies based on the nature of the data. Always use a conservative data reduction ratio for the initial configuration.

Using the suitable ratio for capacity assignment can cause an out of space situation. If the MDisks do not provide enough capacity, IBM FlashSystem disables access to all the volumes in the storage pool.

Example: This example includes the following assumptions:

- Assumption 1: Sizing is performed with an optimistic 5:1 rate
- Assumption 2: Real rate is 3:1

Therefore:

- Physical Capacity: 20 TB
- Calculated capacity: 20 TB x 5 = 100 TB
- The volume that is assigned from the compression- or deduplication-enabled storage subsystem to the IBM SAN Volume Controller or IBM Storwize is 100 TB
- Real usable capacity: 20 TB x 3 = 60 TB

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

Thin-provisioned back-end storage must be carefully monitored. It is necessary to set up capacity alerts to be aware of the real remaining physical capacity.

Also, the best practice is to have an emergency plan and know the steps to recover from an "Out Of Physical Space" situation on the back-end controller. The plan must be prepared during the initial implementation phase.

3.4 Controller-specific considerations

This section discusses implementation-specific information that is related to different supported back-end systems. For more information about general requirements, see this IBM Documentation web page.

3.4.1 Considerations for DS8000 series

In this section, we discuss considerations for the DS800 series.

Interaction between DS8000 and IBM FlashSystem

It is important to know DS8000 drive virtualization process, which is the process of preparing physical drives for storing data that belongs to a volume that is used by a host. In this case, the host is the IBM FlashSystem.

In this regard, the basis for virtualization begins with the physical drives of DS8000, which are mounted in storage enclosures. Virtualization builds upon the physical drives as a series of layers:

- Array sites
- Arrays
- Ranks
- Extent pools
- Logical volumes
- Logical subsystems

Array sites are the building blocks that are used to define arrays, which are data storage systems for block-based, file-based, or object based storage. Instead of storing data on a server, storage arrays use multiple drives that are managed by a central management and can store a huge amount of data.

In general terms, eight identical drives that have the same capacity, speed, and drive class comprise the array site. When an array is created, the RAID level, array type, and array configuration are defined. RAID 5, RAID 6, and RAID 10 levels are supported.

> **Important:** Normally the RAID 6 is highly preferred, and is the default while using the Data Storage Graphical Interface (DS GUI). As with large drives in particular, the RAID rebuild times (after one drive failure) become larger. Using RAID 6 reduces the danger of data loss due to a double-RAID failure. For more information, see this IBM Documentation web page.

A rank, which is a logical representation for the physical array, is relevant for IBM FlashSystem because of the creation of a fixed block (FB) pool for each array that you want to virtualize. Ranks in DS8000 are defined in a one-to-one relationship to arrays. It is for this reason that a rank is defined as using only one array.

A fixed-block rank features one of the following extent sizes:

- 1 GiB, which is a large extent
- 16 MiB, which is a small extent

An *extent pool* or storage pool in DS8000 is a logical construct to add the extents from a set of ranks, forming a domain for extent allocation to a logical volume.

In synthesis, a *logical volume* consists of a set of extents from one extent pool or storage pool. DS8900F supports up to 65,280 logical volumes.

A logical volume that is composed of fix block extents is called logical unit number (LUN). A fixed-block LUN consists of one or more 1 GiB (large) extents, or one or more 16 MiB (small) extents from one FB extent pool. A LUN is not allowed to cross extent pools. However, a LUN can have extents from multiple ranks within the same extent pool.

> **Important:** DS8000 Copy Services does not support FB logical volumes larger than 2 TiB. Therefore, you cannot create a LUN that is larger than 2 TiB if you want to use Copy Services for the LUN, unless the LUN is integrated as Managed Disks in an IBM FlashSystem. Use IBM Spectrum Virtualize Copy Services instead. Based on the considerations, the maximum LUN sizes to create at DS8900F and present to IBM FlashSystem are as follows:
>
> - 16 TB LUN with large extents (1 GiB)
> - 16 TB LUN with small extent (16 MiB) for DS8880F with version or edition R8.5 or later, and for DS8900F R9.0 or later.

Logical subsystems (or LSS) are another logical construct, and mostly used in conjunction with fixed-block volumes. Thus, a maximum of 255 LSSs can exist on DS8900F. For more information, see this IBM Documentation web page.

The concepts of virtualization of DS8900F for IBM FlashSystem are schematically shown in Figure 3-4.

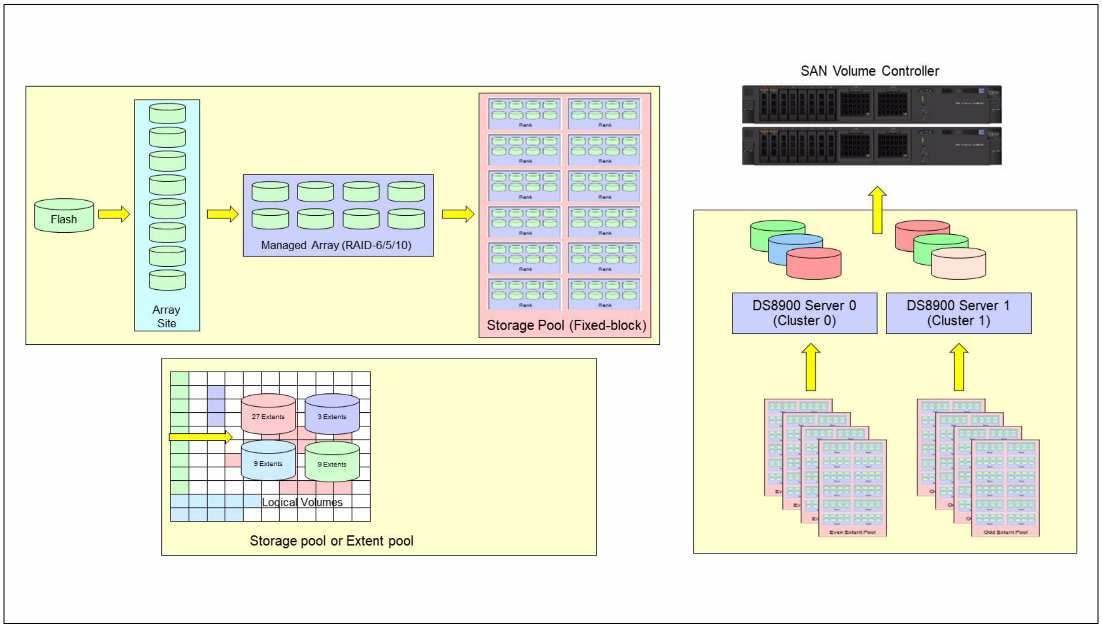


Figure 3-4 DS8900 virtualization concepts focus to IBM FlashSystem

Connectivity considerations

The number of DS8000 ports to be used is at least eight. With large and workload intensive configurations, consider using more ports, up to 16, which is the maximum supported by IBM FlashSystem.

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

For more information about SAN preferred practices and connectivity, see Chapter 2, “Connecting IBM Spectrum Virtualize and IBM Storwize in storage area networks” on page 37.

Defining storage

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

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

Balancing workload across DS8000 series controllers

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

The DS8000 series controllers assign server (controller) affinity to ranks when they are added to an extent pool. Ranks that belong to an even-numbered extent pool have an affinity to Server 0, and ranks that belong to an odd-numbered extent pool have an affinity to Server 1.

Figure 3-5 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 Server 0. Therefore, the adapter in Server 1 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.

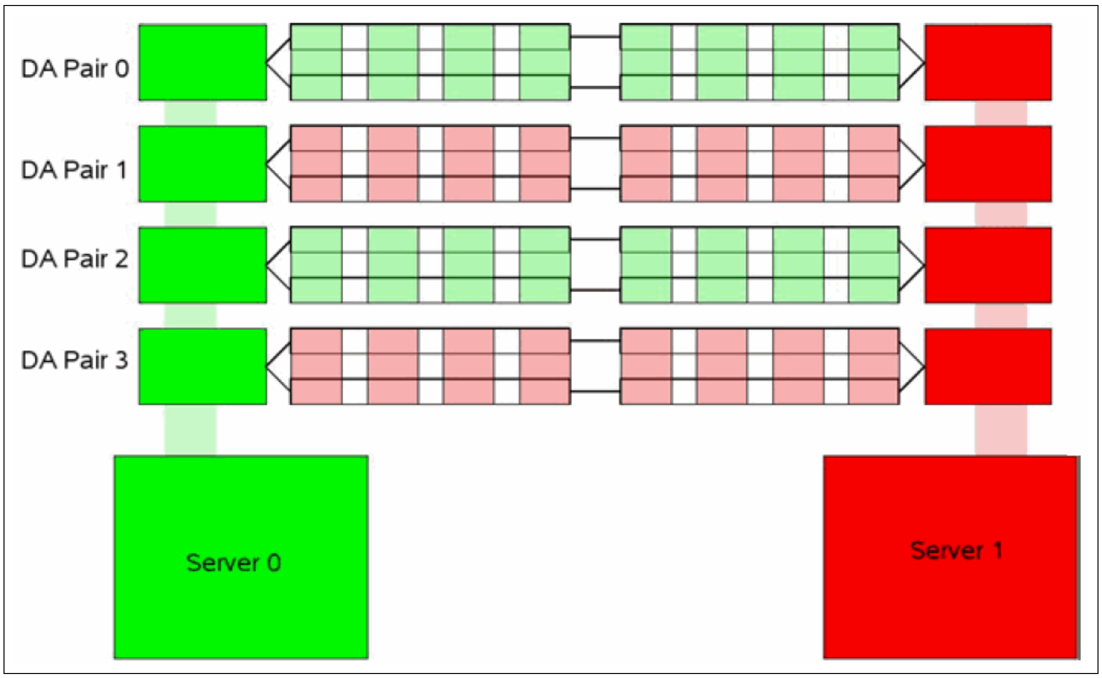


Figure 3-5 DA pair reduced bandwidth configuration

Example 3-6 on page 95 shows the invalid configuration, as depicted in 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, Server 0 is represented by Group 0, 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 Server 0 (represented by Group 0), which leaves the DA in Server 1 idle.

Example 3-6 Command output for the lsarray and lsrank commands

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

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

Figure 3-6 shows a configuration that balances the workload across all four DA pairs.

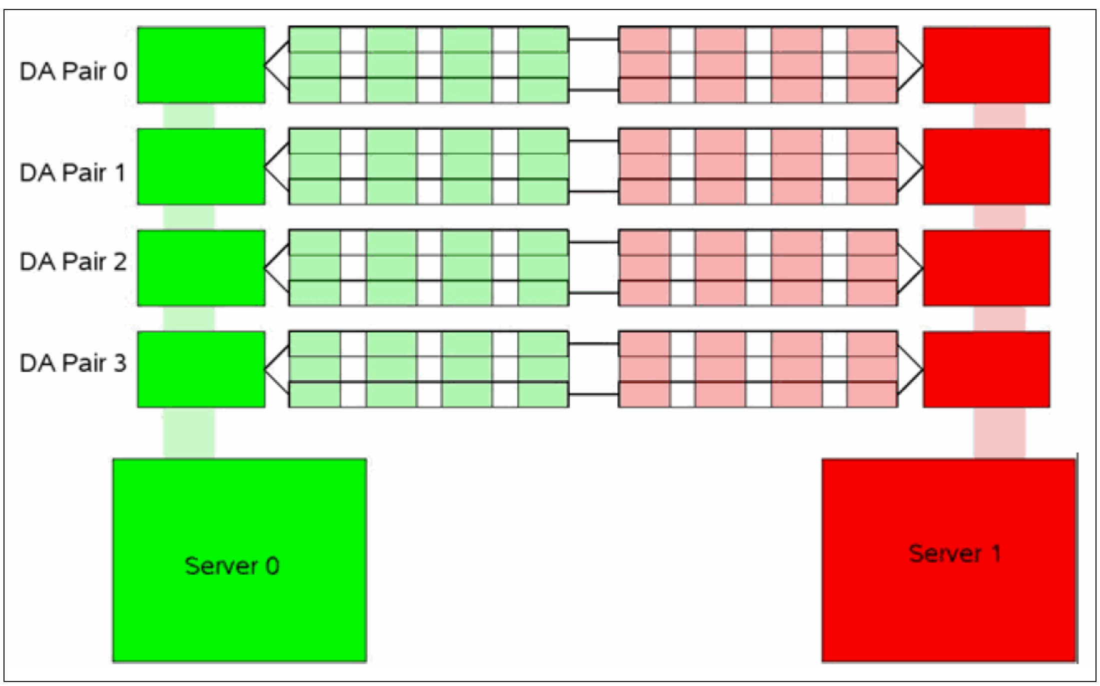


Figure 3-6 DA pair correct configuration

Figure 3-7 shows a correct configuration, as depicted in the CLI output of the **lsarray** and **lsrank** commands. Notice that the output shows that this configuration balances the workload across all four DA pairs with an even balance between odd and even extent pools. The arrays that are on the same DA pair are split between groups 0 and 1.

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

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

Figure 3-7 The lsarray and lsrank command output

DS8000 series ranks to extent pools mapping

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

- One-to-one approach: One rank per extent pool configuration.

  With the one-to-one approach, DS8000 is formatted in 1:1 assignment between ranks and extent pools. This configuration disables any DS8000 storage-pool striping or auto-rebalancing activity, if they were enabled. You can create one or two volumes in each extent pool exclusively on one rank only and put all of those volumes into one IBM FlashSystem storage pool. IBM FlashSystem stripes across all of these volumes and balances the load across the RAID ranks by that method. No more than two volumes per rank are needed with this approach. So, the rank size determines the volume size.

  Often, systems are configured with at least two storage pools:

  – One (or two) that contain MDisks of all the 6+P RAID 5 ranks of the DS8000 storage system.

  – One (or more) that contain the slightly larger 7+P RAID 5 ranks.

  This approach maintains equal load balancing across all ranks when the IBM FlashSystem striping occurs because each MDisk in a storage pool is the same size.

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

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

  – A natural skew, such as a small file of a few hundred KiB that is heavily accessed.

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

  An advantage of this approach is that it delivers more options for fault isolation and control over where a certain volume and extent are located.

- Many-to-one approach: Multi-rank extent pool configuration.

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

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

  For more information about choosing the level on which to run Easy Tier, see "Monitoring Easy Tier using the GUI" on page 191.

  You need only one volume size with this multi-rank approach because plenty of space is available in each large DS8000 extent pool. As mentioned previously, the maximum number of back-end storage ports to be presented to the IBM FlashSystem is 16. Each port represents a path to the IBM FlashSystem. Therefore, when sizing the number of LUN/MDisks to be presented to the IBM FlashSystem, the suggestion is to present least between two and four volumes per path. So using the maximum of 16 paths, create 32, 48, or 64 DS8000 volumes, and for this configuration IBM FlashSystem maintains a good queue depth.

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

LUN masking

For a storage controller, all IBM FlashSystem nodes must detect the same set of LUs from all target ports that logged in. If target ports are visible to the nodes or canisters that do not have the same set of LUs assigned, IBM FlashSystem 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 IBM FlashSystem.

The DS8000 series controllers perform LUN masking that is based on the volume group. Example 3-7 shows the output of the **showvolgrp** command for volume group (`V0`), which contains 16 LUNs that are being presented to a two-node IBM FlashSystem cluster.

Example 3-7 Output of the showvolgrp command

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

Example 3-8 shows output for the **lshostconnect** command from the DS8000 series. In this example, four ports of the two-node cluster are assigned to the same volume group (V0) and, therefore, are assigned to the same four LUNs.

Example 3-8 Output for the lshostconnect command

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

From Example 3-8 you can see that only the IBM FlashSystem WWPNs are assigned to V0.

> **Attention:** Data corruption can occur if the same LUN is assigned to IBM FlashSystem nodes and other devices, such as hosts attached to DS8000.

Next, you see how the IBM FlashSystem 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 IBM FlashSystem cluster by that specific controller.

Example 3-9 shows the general details of the output storage controller by using the system CLI.

Example 3-9 Output of the lscontroller command

```
IBM_FlashSystem:FS9100-ITSO:superuser>svcinfo lscontroller DS8K75FPX81
id 1
controller_name DS8K75FPX81
WWNN 5005076305FFC74C
mdisk_link_count 16
max_mdisk_link_count 16
degraded no
vendor_id IBM
product_id_low 2107900
...
WWPN 500507630500C74C
path_count 16
max_path_count 16
WWPN 500507630508C74C
path_count 16
max_path_count 16
```

IBM FlashSystem MDisks and storage pool considerations

Recommended practice is to create a single IBM FlashSystem storage pool per DS8900F system. This provides simplicity of management, and best overall performance.

An example of the preferred configuration is shown in Figure 3-8 on page 99. Four storage pools or extent pools (one even and one odd) of DS8900F are joined into one IBM FlashSystem storage pool.

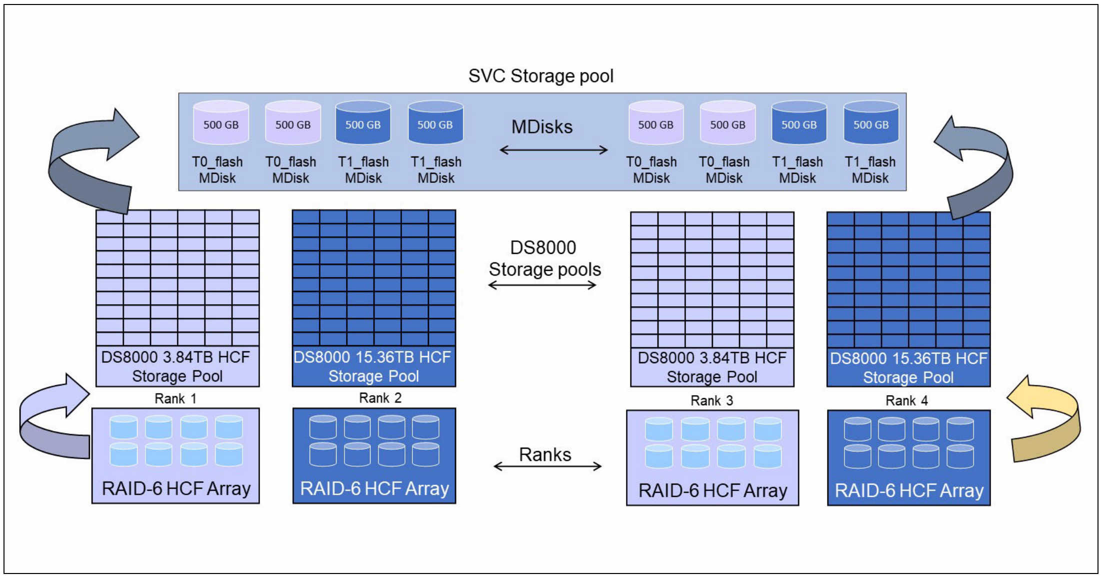


Figure 3-8 Four DS8900F extent pools as one IBM FlashSystem storage pool

To determine how many logical volumes must be created to present to IBM FlashSystem as MDisks, see 3.3.2, “Guidelines for creating optimal backend configuration” on page 88.

3.4.2 Considerations for IBM XIV Storage System

The XIV Gen3 volumes can be provisioned to IBM FlashSystem by way of iSCSI and FC. However, it is preferred that you implement FC attachment for performance and stability considerations, unless a dedicated IP infrastructure for storage is available.

Host options and settings for XIV systems

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

From an IBM FlashSystem perspective, an XIV type 281x controller can consist of more than one WWPN. However, all are placed under one worldwide node number (WWNN) that identifies the entire XIV system.

Creating a host object for IBM FlashSystem for an IBM XIV

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

When implemented in this manner, statistical metrics are more effective because performance can be collected and analyzed on IBM FlashSystem node level.

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

Volume considerations

As modular storage, XIV storage can be available in a minimum of six modules and up to a maximum of 15 modules in a configuration. Each additional module added to the configuration increases the XIV capacity, CPU, memory, and connectivity.

As of this writing, the XIV system supports the following configurations:

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

Figure 3-9 shows how XIV configurations vary according to the number of modules that are on the system.

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

Figure 3-9 XIV rack configuration: 281x-214

Although XIV has its own queue depth characteristics for direct host attachment, the best practices that are described in 3.3.2, “Guidelines for creating optimal backend configuration” on page 88 are preferred when you virtualize XIV with IBM Spectrum Virtualize.

Table 3-6 lists s the suggested volume sizes and quantities for IBM FlashSystem on the XIV systems with different drive capacities.

Table 3-6 XIV minimum volume size and quantity recommendations

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

Other considerations

Consider the following restrictions when using the XIV system as back-end storage for the IBM FlashSystem:

- Volume mapping

  When mapping a volume, you must use the same LUN ID to all IBM FlashSystem nodes. Therefore, map the volumes to the cluster, not to individual nodes.

- XIV Storage pools

  When creating an XIV storage pool, define the Snapshot Size as zero (0). Snapshot space does not need to be reserved, because it is not recommended that you use XIV snapshots on LUNs mapped as MDisks. The snapshot functions should be used on IBM FlashSystem level.

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

- Thin provisioning

  XIV thin provisioning pools are not supported by IBM FlashSystem. Instead, you must use a regular pool.

- Copy functions for XIV models

  You cannot use advanced copy functions, such as taking a snapshot and remote mirroring, for XIV models with disks that are managed by the IBM FlashSystem.

For more information about configuration of XIV behind IBM FlashSystem, see *IBM XIV Gen3 with IBM System Storage SAN Volume Controller and Storwize V7000*, REDP-5063.

3.4.3 Considerations for IBM FlashSystem A9000/A9000R

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

Deduplication can be implemented on the IBM FlashSystem by attaching an IBM FlashSystem A9000/A9000R as external storage instead of using IBM Spectrum Virtualize DRP-level deduplication.

There are several considerations when you are attaching an IBM FlashSystem A9000/A9000R system as a back-end controller.

Volume considerations

IBM FlashSystem A9000/A9000R designates resources to data reduction, and as this designation is always on, it is strongly advised that data reduction be done only in the IBM FlashSystem A9000/A9000R and not in the Spectrum Virtualize cluster. Otherwise, as IBM FlashSystem A9000/A9000R tries to reduce the data, unnecessary additional latency occurs.

Estimated data reduction is important because that helps determine volume size. Always try to use a conservative data-reduction ratio when attaching A9000/A9000R, because the storage pool will go offline if the back-end storage runs out of capacity.

To determine the controller volume size, complete the following tasks:

- Calculate effective capacity: reduce measured-data reduction ratio (for example, if the data reduction estimation tool provides a ratio of 4:1, use 3.5:1 for calculations) and multiply it to determine physical capacity.
- Determine the number of connected FC ports by using Table 3-7 and Table 3-8.
- Volume size is equal to effective capacity that is divided by the number of ports taken twice (effective capacity/path*2).

The remaining usable capacity can be added to the storage pool after the system reaches a stable date reduction ratio.

Table 3-7 Host connections for A9000

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

Table 3-8 Host connections for A9000R

Grid element	Number of controllers	Total FC ports available	Total ports that are connected to IBM SAN Volume Controller	Connected ports
2	4	16	8	All controllers, ports 1 and 3
3	6	24	12	All controllers, ports 1 and 3
4	8	32	8	Controllers 1 - 4, port 1 Controllers 5 - 8, port 3

Grid element	Number of controllers	Total FC ports available	Total ports that are connected to IBM SAN Volume Controller	Connected ports
5	10	40	10	Controllers 1 - 5, port 1 Controllers 6 - 10, port 3
6	12	48	12	Controllers 1 - 6, port 1 Controllers 7 - 12, port 3

It is important not to run out of hard capacity on the back-end storage because doing so takes the storage pool offline. It is important to closely monitor the FlashSystem A9000/A9000R. If you start to run out of space, you can use the migration functions of Spectrum Virtualize to move data to another storage system, as described in the examples next.

Examples: Consider the following examples:

- FlashSystem A9000 with 57 TB of usable capacity, or 300 TB of effective capacity, at the standard 5.26:1 data efficiency ratio.

  We ran the data reduction tool on a good representative sample of the volumes that we will be virtualizing. We know that we have a data reduction ratio of 4.2:1 and for extra safety, use 4:1 for further calculations. The use of 4 x 57 gives you 228 TB. Divide this result by 12 (six paths x 2), and you get 19 TB per volume.

- A five-grid element FlashSystem A9000R that uses 29 TB Flash enclosures has a total usable capacity of 145 TB.

  We are use 10 paths and have not run any of the estimation tools on the data. However, we know that the host is not compressing the data. We assume a compression ratio of 2:1, 2 x 145 gives 290, and divided by 20 gives 14.5 TB per volume. In this case, if we see that we are getting a much better data reduction ratio than we planned for, we can always create volumes and make them available to IBM Spectrum Virtualize.

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

Other considerations

Spectrum Virtualize is able to detect that the IBM FlashSystem A9000 controller is using deduplication technology and show that the `Deduplication` attribute of the managed disk is `Active`.

Deduplication status is important because it allows IBM Spectrum Virtualize to enforce the following restrictions:

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

3.4.4 Considerations for FlashSystem 5000, 5100, 5200, 7200, 9100, and 9200

Recommendations that are discussed in this section apply to a solution with IBM FlashSystem family or IBM Storwize family system that is virtualized by another IBM FlashSystem family system.

Connectivity considerations

It is expected that NPIV is enabled on both systems: the system that is virtualizing storage, and the system that works as a back-end. Zone "host" or "virtual" WWPNs of the back-end system to physical WWPNs of the front-end, or virtualizing system.

For more information about SAN and zoning preferred practices, see Chapter 2, "Connecting IBM Spectrum Virtualize and IBM Storwize in storage area networks" on page 37.

System layers

Spectrum Virtualize systems have a concept of system layers. There are two layers - *storage* and *replication*. Systems that are configured into storage layer can work as a back-end storage. Systems that are configured into replication layer, can virtualize another IBM FlashSystem clusters and use them as back-end controllers.

Systems that are configured with the same layer can be replication partners. Systems in the different layers cannot.

By default, IBM FlashSystem is configured to storage layer. The system layer on IBM FlashSystem can be switched. IBM SAN Volume Controller is configured to replication layer, and it cannot be changed on IBM SAN Volume Controller.

For more information and instructions and limitations, see this IBM Documentation web page.

Automatic configuration

IBM FlashSystem family systems that run code version 8.3x and above can be automatically configured for optimal performance as a back-end storage behind IBM SAN Volume Controller.

An automatic configuration wizard must be used on a system where volumes, pools, and host objects are not configured. Wizard will configure internal storage devices, create volumes and map the to the host object, representing IBM SAN Volume Controller.

Array and disk pool considerations

The back-end IBM FlashSystem family system can have a hybrid configuration, containing FlashCore Modules and SSD drives, or SSDs and spinning drives.

Internal storage attached to the back-end system needs to be joined into RAID arrays. You might need one or more DRAID6 arrays, depending on the number and the type of available drives. For RAID recommendations, see 3.2.2, "Array considerations" on page 80.

Consider creating a separate disk pool for each type (tier) of storage and use the Easy Tier function on a front-end system. Front-end FlashSystem family systems cannot monitor Easy Tier activity of the back-end storage.

If Easy Tier is enabled on front- and back-end systems, they independently rebalance the hot areas according to their own heat map. This process causes a rebalance over a rebalance. Such a situation can eliminate the performance benefits of extent reallocation. For this reason, Easy Tier must be enabled only on one level (preferably the front-end).

For more information about recommendations about Easy Tier with external storage, see Chapter 4, “Planning storage pools” on page 111.

For most use cases, standard pools are preferred to data-reduction pools on the back-end storage. If planned, the front-end will perform reduction. Data reduction on both levels is not recommended as it adds processing overhead and does not result in capacity savings.

If Easy Tier is disabled on the backend, as advised here, back-end FlashSystem pool extent size is not a performance concern.

SCSI Unmap considerations

When virtualized, IBM FlashSystem treats “virtualizer” system as a host. By default, host SCSI UNMAP support is enabled on IBM FlashSystem 9100 and FlashSystem 9200, and disabled on other platforms.

Consider enabling host Unmap support to achieve better capacity management if the system that is going to be virtualized meets the following qualifications:

- Contains FCMs
- Is Flash-only (no spinning drives)

Consider leaving host Unmap disabled to protect virtualized system from being over-loaded if you are going to virtualize a hybrid system, and storage to be virtualized uses spinning disks.

To switch host Unmap support on or off, use the **`chssystem`** CLI command. For more information, see this IBM Documentation web page.

Volume considerations

Volumes in IBM FlashSystem can be created as *striped* or *sequential*. The general rule is to create striped volumes. Volumes on back-end system must be fully allocated.

To determine a number of volumes to create on back-end IBM FlashSystem to provide a virtualizer as MDisks, see the general rules provided in 3.3.2, “Guidelines for creating optimal backend configuration” on page 88. When virtualizing back-end with spinning drives, perform queue depth calculations. For all flash solutions, create 32 volumes from the available pool capacity, which can be reduced to 16 or even 8 for small arrays (for example, if you have 16 or less flash drives in a back-end pool). For FCM arrays, the number of volumes is also governed by load distribution. 32 volumes out of a pool with an FCM array are recommended.

When choosing volume size, consider which system (front-end or back-end) performs the compression. If data is compressed and deduplicated on the front-end IBM FlashSystem (or “virtualizer”), FCMs cannot compress it further, which results in a 1:1 compression ratio.

Therefore, the back-end volume size is calculated from the pool physical capacity that is divided by the number of volumes (16 or more).

Example: FlashSystem 9200 with 24 x 19.2 TB modules.

This configuration provides raw disk capacity of 460 TB, with 10+P+Q DRAID6 and one distributed spare, and the physical array capacity is 365 TB or 332 TiB.

Because it is not recommended to provision more than 85% of a physical flash, we have 282 TiB. Because we do not expect any compression on FCM (the backend is getting data that is compressed by upper levels), we provision storage to upper level assuming a 1:1 compression, which means we create 32 volumes 282 TiB / 32 = 8.8 TiB each.

If the front-end system is not compressing data, space savings are achieved with FCM hardware compression. Use compression-estimation tools to determine the expected compression ratio and use a smaller ratio for further calculations (for example, if you expect 4.5:1 compression, use 4.3:1). Determine the volume size by using the calculated effective pool capacity.

Example: IBM FlashSystem 7200 with 12 x 9.6 TB modules.

This configuration provides raw disk capacity of 115 TB, with 9+P+Q DRAID6 and one distributed spare. The physical capacity is 85 TB or 78 TiB.

Because it is not recommended to provision more than 85% of a physical flash, we have 66 TiB. Compresstimator shows that we can achieve 3.2:1 compression ratio, decreasing in and assuming 3:1, we have 66 TiB x 3 = 198 TiB of effective capacity.

Create 16 volumes, 198TiB / 16 = 12.4 TiB each. If the compression ratio is higher than expected, we can create and provision more volumes to the frontend.

3.4.5 IBM FlashSystem 900 considerations

The main advantage of integrating FlashSystem 900 with IBM Spectrum Virtualize is to combine the extreme performance of IBM FlashSystem 900 with the Spectrum Virtualize enterprise-class solution such as tiering, volume mirroring, deduplication, and copy services.

When you configure the IBM FlashSystem 900 as a backend for Spectrum Virtualize family systems, you must remember the considerations that are described in this section.

Defining storage

IBM FlashSystem 900 supports up to 12 IBM MicroLatency® modules. IBM MicroLatency modules are installed in the IBM FlashSystem 900 based on the following configuration guidelines:

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

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

Volume considerations

To fully use all Spectrum Virtualize system resources, create 32 volumes (or 16 volumes if FlashSystem 900 is not fully populated). This way, all CPU cores, nodes, and FC ports of the virtualizer are fully used.

However, one important factor must be considered when volumes are created from a pure FlashSystem 900 MDisks storage pool. FlashSystem 900 can process I/Os much faster than traditional storage. Sometimes they are even faster than cache operations, because with cache all I/Os to the volume must be mirrored to another node in I/O group.

This operation can take as much as 1 millisecond while I/Os that are issued directly (which means without cache) to the FlashSystem 900 can take 100 - 200 microseconds. So, in some rare use-case, it might be recommended to disable Spectrum Virtualize cache to optimize for maximum IOPS.

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

- If volumes from FlashSystem 900 pool are compressed
- If volumes from FlashSystem 900 pool are in a Metro/Global Mirror relationship
- If volumes from FlashSystem 900 pool are in a FlashCopy relationship (either source or target)
- If the same pool has MDisks from FlashSystem 900 contains also MDisks from other back-end controllers.

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

3.4.6 Path considerations for third-party storage with EMC VMAX and Hitachi Data Systems

Many third-party storage options are available and supported, This section describes the multipathing considerations for EMC VMAX and Hitachi Data Systems (HDS).

Most storage controllers, when presented to the IBM FlashSystem, are recognized as a single WWNN per controller. However, for some EMC VMAX and HDS storage controller types, the system recognizes each port as a different WWNN. For this reason, each storage port, when zoned to an IBM FlashSystem, appears as a different external storage controller.

IBM Spectrum Virtualize supports a maximum of 16 WWNNs per storage system, so it is preferred to connect up to 16 storage ports.

To determine a number of logical volumes or LUNs to be configured on third-party storage, see 3.3.2, “Guidelines for creating optimal backend configuration” on page 88.

3.5 Quorum disks

Note: This section does not cover IP-attached quorum. For information about IP-attached quorums, see Chapter 7, “Ensuring business continuity” on page 357.

A system uses a quorum disk for two purposes:

- To break a tie when a SAN fault occurs, when exactly half of the nodes that were previously a member of the system are present
- To hold a copy of important system configuration data

After internal drives are prepared to be added to an array, or external MDisks become managed, a small portion of its capacity is reserved for quorum data. Its size is less than 0.5 GiB for a drive and not less than one pool extent for an MDisk.

Three devices from all available internal drives and managed MDisks are selected for the *quorum disk* role. They store system metadata which is used for cluster recovery after a disaster. Despite only three devices that are actually designated as quorums, capacity for quorum data is reserved on each of them, as the designation might change (for example, if quorum disk has a physical failure).

Only one of those disks is selected as the active quorum disk. It is used as a tie-breaker. If, as a result of a failure, the cluster is split in half and both parts lose sight of each other (for example, the inter-site link has failed in a HyperSwap cluster with two I/O groups), they appeal to the tie-breaker, active quorum device. The half of the cluster nodes that were able to reach and reserve the quorum disk after the split occurs, lock the disk and continue to operate. The other half stops its operation. This design prevents both sides from becoming inconsistent with each other.

The storage device must match following criteria to be considered a quorum candidate:

- Internal drive or module must follow these rules:
  - Be a member of an array or a “Candidate”.
  - Be in “Unused” state cannot be quorums.
  - MDisk must be in “Managed” state. “Unmanaged” or “Image” MDisks cannot be quorums.
- External MDisks can be provisioned over only FC and not iSCSI.
- An MDisk must be presented by a disk subsystem, LUNs from which are supported to be quorum disks.

The system uses the following rules when selecting quorum devices:

- Fully connected candidates are preferred over partially connected candidates.

  This means that in a multiple enclosure environment, MDisks will be preferred over drives.
- Drives are preferred over MDisks.

  If there is only one control enclosure and no external storage in the cluster, drives are considered first.
- Drives from a different control enclosure are to be preferred over a second drive from the same enclosure.

  If IBM FlashSystem contains more than one I/O group, at least one of the candidates from each group is selected.
- NVMe drives are preferred over SAS drives.

  NVMe drive in control enclosure will be chosen rather than SAS expansion drive.

To become an active quorum device (tie-break device), it must be visible to all nodes in a cluster.

In practice, these rules mean:

- For IBM FlashSystem with a single control enclosure, quorums including active quorum disk are assigned out of its internal drives automatically. No actions required.
- For IBM FlashSystem with two or more I/O groups and with external storage virtualized, the active quorum will be assigned to an external MDisk. None of the internal drives can become the active quorum, because they are connected to a single control enclosure and visible only by one pair of nodes.

- For IBM FlashSystem with two or more I/O groups and without external storage, there will be no active quorum selected automatically. However, a standard topology cluster in most use cases will operate without any issues. For HyperSwap topology, IP quorum or FC-attached quorum needs to be deployed on the third site.

To list IBM FlashSystem quorum devices, run the **lsquorum** command as shown in Example 3-10.

Example 3-10 The lsquorum command

```
IBM_FlashSystem:FS9100-ITSO:superuser>lsquorum
quorum_index status id name controller_id controller_name active object_type
0            online 4                                      no     drive
1            online 1                                      yes    drive
2            online 2                                      no     drive
```

To move quorum assignment, use the **chquorum** command. Note that it is not supported on NVMe drives, so you can move it only *from* the NVMe drive, but not *to* the NVMe drive.

4

Planning storage pools

This chapter describes considerations for planning storage pools for an IBM FlashSystem implementation. It explains various pool configuration options, including Easy Tier and data reduction pools (DRPs). It provides and provides best practices on implementation and an overview of some typical operations with MDisks.

This chapter includes the following topics:

4.1 Introduction to pools

In general, a storage pool or pool, sometimes referred to as *managed disk group*, is a grouping of storage capacity that is used to provision volumes and logical units (LUNs) that can subsequently be made visible to hosts.

IBM FlashSystem supports the following types of pools:

- Standard pools: Parent pools and child pools
- DRPs: Parent pools and quotaless child pools

Standard pools were available since the initial release of IBM Spectrum Virtualize in 2003 and can include fully allocated or thin-provisioned volumes.

Real-time Compression (RtC) is allowed only with standard pools on some older IBM SAN Volume Controller hardware models and should not be implemented in new configurations.

> **Note:** The latest node hardware does not support RtC.
>
> SA2 and SV2 IBM SAN Volume Controller node hardware do not support the use of RtC volumes. To migrate a system to use these node types, all RtC volumes must be removed (migrated) to uncompressed standard pool volumes, or into a DRP.

IBM FlashSystems that use standard pools cannot be configured by using RtC.

DRPs represent a significant enhancement to the storage pool concept because the virtualization layer is primarily a simple layer that runs the task of lookups between virtual and physical extents. With the introduction of data reduction technology, compression, and deduplication, it has become more of a requirement to have an uncomplicated way to stay thin.

DRPs increase infrastructure capacity usage by employing new efficiency functions and reducing storage costs. The pools enable you to automatically de-allocate (not to be confused with deduplicate) and reclaim capacity of thin-provisioned volumes containing deleted data. In addition, for the first time, the pools enable this reclaimed capacity to be reused by other volumes.

Either pool type can be made up of different tiers. A tier defines a performance characteristic of that subset of capacity in the pool. Often, no more than three tier types are be defined in a pool (fastest, average, and slowest). The tiers and their usage are managed automatically by the Easy Tier function.

4.1.1 Standard pool

Standard pools (also referred to as traditional storage pools), provide a way of providing storage in IBM FlashSystem. They use a fixed allocation unit of an extent. Standard pools are still a valid method to providing capacity to hosts. For more information about guidelines for implementing standard pools, see 4.2, “Storage pool planning considerations” on page 132.

IBM FlashSystem can define parent and child pools. A *parent* pool has all the capabilities and functions of a normal IBM FlashSystem pool. A *child* pool is a logical subdivision of a storage pool or managed disk group. Like a parent pool, a child pool supports volume creation and migration.

When you create a child pool in a standard parent pool the user must specify a capacity limit for the child pool. This limit allows for a quota of capacity to be allocated to the child pool. This capacity is reserved for the child pool and detracts from the available capacity in the parent pool. This process is different than the method with which child pools are implemented in a DRP. For more information, see “Quotaless data reduction child pool” on page 118.

A child pool inherits its tier setting from the parent pool. Changes to a parent’s tier setting are inherited by child pools.

A child pool supports the Easy Tier function if Easy Tier is enabled on the parent pool. The child pool also inherits Easy Tier status, pool status, capacity information, and back-end storage information. The I/O activity of parent pool is the sum of the I/O activity of itself and the child pools.

Parent pools

Parent pools receive their capacity from MDisks. To track the space that is available on an MDisk, the system divides each MDisk into chunks of equal size. These chunks are called *extents* and are indexed internally. The choice of extent size affects the total amount of storage that is managed by the system. The extent size remains constant throughout the lifetime of the parent pool.

All MDisks in a pool are split into extents of the same size. Volumes are created from the extents that are available in the pool. You can add MDisks to a pool at any time to increase the number of extents that are available for new volume copies or to expand volume copies. The system automatically balances volume extents between the MDisks to provide the best performance to the volumes.

You cannot use the volume migration functions to migrate volumes between parent pools that feature different extent sizes. However, you can use volume mirroring to move data to a parent pool that has a different extent size.

Choose extent size wisely according to your future needs. A small extent size limit your overall usable capacity, but a larger extent size can waste storage. For example, if you select an extent size of 8 GiB, but then only create a 6 GiB volume, one entire extent is allocated to this volume (8 GiB) and hence 2 GiB are unused.

When you create or manage a parent pool, consider the following general guidelines:

- Ensure that all MDisks that are allocated to the same tier of a parent pool are the same RAID type. This configuration ensures that the same resiliency is maintained across that tier. Similarly, for performance reasons, do not mix RAID types within a tier. The performance of all volumes is reduced to the lowest achiever in the tier and a mismatch of tier members can result in I/O convoying effects where everything is waiting on the slowest member.
- An MDisk can be associated with only one parent pool.
- You should specify a warning capacity for a pool. A warning event is generated when the amount of space that is used in the pool exceeds the warning capacity. The warning threshold is especially useful with thin-provisioned volumes that are configured to automatically use space from the pool.
- Volumes are associated with just one pool, except for the duration of any migration between parent pools.
- Volumes that are allocated from a parent pool are by default striped across all the storage that is placed into that parent pool. Wide striping can provide performance benefits.

- You can only add MDisks that are in unmanaged mode to a parent pool. When MDisks are added to a parent pool, their mode changes from unmanaged to managed.
- You can delete MDisks from a parent pool under the following conditions:
  - Volumes are not using any of the extents that are on the MDisk.
  - Enough free extents are available elsewhere in the pool to move extents that are in use from this MDisk.
  - The system ensures that all extents that are used by volumes in the child pool are migrated to other MDisks in the parent pool to ensure that data is not lost.

> **Important:** Before you remove MDisks from a parent pool, ensure that the parent pool has enough capacity for child pools that are associated with the parent pool.

- If the parent pool is deleted, you cannot recover the mapping that existed between extents that are in the pool or the extents that the volumes use. If the parent pool includes associated child pools, you must delete the child pools first and return its extents to the parent pool. After the child pools are deleted, you can delete the parent pool. The MDisks that were in the parent pool are returned to unmanaged mode and can be added to other parent pools. Because the deletion of a parent pool can cause a loss of data, you must force the deletion if volumes are associated with it.

> **Note:** Deleting a child or parent pool is unrecoverable.
>
> If you force-delete a pool, all volumes in that pool are deleted, even if they are mapped to a host and are still in use. Use extreme caution when force-deleting pool objects because volume-to-extent mapping cannot be recovered after the delete is processed.
>
> Force-deleting a storage pool is possible only with the command line tools. See the **rmmdiskgrp** command-help for details.

- When you delete a pool with mirrored volumes, consider the following points:
  - if the volume is mirrored and the synchronized copies of the volume are all in the same pool, the mirrored volume is destroyed when the storage pool is deleted.
  - If the volume is mirrored and a synchronized copy exists in a different pool, the volume remains after the pool is deleted.

You might not be able to delete a pool or child pool if Volume Delete Protection is enabled. In code versions 8.3.1 and later, Volume Delete Protection is enabled by default. However, the granularity of protection is improved; you can now specify Volume Delete Protection to be enabled or disabled on a per-pool basis, rather than on a system basis as was previously the case.

Child pools

Instead of being created directly from MDisks, child pools are created from existing capacity that is allocated to a parent pool. As with parent pools, volumes can be created that specifically use the capacity that is allocated to the child pool. Child pools are similar to parent pools with similar properties and can be used for volume copy operation.

Child pools are created with fully-allocated physical capacity; that is, the physical capacity that is applied to the child pool is reserved from the parent pool, as though you created a fully-allocated volume of the same size in the parent pool.

The allocated capacity of the child pool must be smaller than the free capacity that is available to the parent pool. The allocated capacity of the child pool is no longer reported as the *free* space of its parent pool. Instead, the parent pool reports the entire child pool as *used* capacity. You must monitor the used capacity (instead of the free capacity) of the child pool instead.

When you create or work with a child pool, consider the following general guidelines:

- Child pools are created automatically by IBM Spectrum Connect VASA client to implement VMware vVols.
- As with parent pools, you can specify a warning threshold that alerts you when the capacity of the child pool is reaching its upper limit. Use this threshold to ensure that access is not lost when the capacity of the child pool is close to its allocated capacity.
- On systems with encryption enabled, child pools can be created to migrate existing volumes in a non-encrypted pool to encrypted child pools. When you create a child pool after encryption is enabled, an encryption key is created for the child pool even when the parent pool is not encrypted. You can then use volume mirroring to migrate the volumes from the non-encrypted parent pool to the encrypted child pool.
- Ensure that any child pools that are associated with a parent pool have enough capacity for the volumes that are in the child pool before removing MDisks from a parent pool. The system automatically migrates all extents that are used by volumes to other MDisks in the parent pool to ensure data is not lost.
- You cannot shrink the capacity of a child pool to less than its real capacity. The system uses reserved extents from the parent pool that use multiple extents. The system also resets the warning level when the child pool is shrunk, and issues a warning if the level is reached when the capacity is shrunk.
- The system supports migrating a copy of volumes between child pools within the same parent pool or migrating a copy of a volume between a child pool and its parent pool. Migrations between a source and target child pool with different parent pools are not supported. However, you can migrate a copy of the volume from the source child pool to its parent pool. The volume copy can then be migrated from the parent pool to the parent pool of the target child pool. Finally, the volume copy can be migrated from the target parent pool to the target child pool.
- Migrating a volume between parent and child pool (with the same encryption key or no encryption) results in a *nocopy migration*. That is, the data does not move. Instead, the extents are re-allocated to the child or parent pool and the accounting of the used space is corrected. That is, the free extents are reallocated to the child or parent to ensure the total capacity allocated to the child pool remains unchanged.
- A special form of *quotaless* data reduction child pool can be created from a data reduction parent pool. For more information, see “Quotaless data reduction child pool” on page 118

Small Computer System Interface unmap in a standard pool

A standard pool can use Small Computer System Interface (SCSI) unmap space reclamation, but not as efficiently as a DRP.

When a host submits a SCSI **`unmap`** command to a volume in a standard pool, the system changes the **`unmap`** command into a **`write_same`** command of zeros. This **`unmap`** command becomes an internal special command and can be handled accordingly by different layers in the system.

For example, the cache does not mirror the data; instead, it passes the special reference to zeros. The RtC functions reclaim those areas (assuming 32 KB or larger) and shrink the volume allocation.

The back-end layers also submit the **`write_same`** command of zeros to the internal or external MDisk devices. For a Flash or SSD-based MDisk this process results in the device freeing the capacity back to its available space. Therefore, it shrinks the used capacity on Flash or SSD, which helps to improve the efficiency of *garbage collection* on the device and performance. The process of reclaiming space is called *garbage collection*.

For Nearline SAS drives, the **`write_same`** of zeros commands can overload the drives themselves, this can result in performance problems.

> **Important:** A standard pool does shrink its used space as the result of a SCSI **`unmap`** command. The backend capacity might shrink its used space, but the pool used capacity does not shrink.
>
> The exception is with RtC volumes where the reused capacity of the volume might shrink; however, the pool allocation to that RtC volume remains unchanged. It means that an RtC volume can reuse that unmapped space first before requesting more capacity from the thin provisioning code.

Thin-provisioned volumes in a standard pool

A thin-provisioned volume presents a different capacity to mapped hosts than the capacity that the volume uses in the storage pool. IBM FlashSystem supports thin-provisioned volumes in standard pools.

> **Note:** While DRPs fundamentally support thin-provisioned volumes, they are used in conjunction with compression and deduplication. With DRPs, you should avoid the use of thin-provisioned volumes without additional data reduction.

In standard pools, thin-provisioned volumes are created as a specific volume type; that is, based on capacity-savings criteria. These properties are managed at the volume level. The virtual capacity of a thin-provisioned volume is typically significantly larger than its real capacity. Each system uses the real capacity to store data that is written to the volume, and metadata that describes the thin-provisioned configuration of the volume. As more information is written to the volume, more of the real capacity is used.

The system identifies read operations to unwritten parts of the virtual capacity and returns zeros to the server without the use of any real capacity. For more information about storage system, pool, and volume capacity metrics, see Chapter 9, “Implementing a storage monitoring system” on page 387.

Thin-provisioned volumes can also help simplify server administration. Instead of assigning a volume with some capacity to an application and increasing that capacity as the needs of the application change, you can configure a volume with a large virtual capacity for the application. You can then increase or shrink the real capacity as the application needs change, without disrupting the application or server.

It is important to monitor physical capacity if you want to provide more space to your hosts than is physically available in your IBM FlashSystem. For more information about monitoring the physical capacity of your storage, and an explanation of the difference between thin provisioning and over-allocation, see 9.5, “Creating alerts for IBM Spectrum Control and IBM Storage Insights” on page 425.

Thin provisioning on top of Flash Core Modules

If you use the compression functions that are provided by the IBM Flash Core Modules (FCMs) in your FlashSystem as a mechanism to add data reduction to a standard pool while maintaining the maximum performance, take care to understand the capacity reporting, in particular if you want to thin provision on top of the FCMs.

The FCM RAID array reports its theoretical maximum capacity, which can be as large as 4:1. This capacity is the maximum that can be stored on the FCM array. However, it might not reflect the compression savings that you achieve with your data.

It is recommended that you start conservatively, especially if you are allocating this capacity to IBM SAN Volume Controller or another IBM FlashSystem (the virtualizer).

You must first understand your expected compression ratio. In an initial deployment, allocate approximately 50% fewer savings. You can easily add “volumes” to the back-end storage system to present as new external “MDisk” capacity to the virtualizer later if your compression ratio is met or bettered.

For example, you have 100 TiB of physical usable capacity in an FCM RAID array before compression. Your comprestimator results show savings of approximately 2:1, which suggests that you can write 200 TiB of volume data to this RAID array.

Start at 150 TiB of volumes that are mapped to as external MDisks to the virtualizer. Monitor the real compression rates and usage and over time add in the other 50 TiB of MDisk capacity to the same virtualizer pool. Be sure to leave spare space for unexpected growth, and consider the guidelines that are outlined in 3.2, “Arrays” on page 78

If you often over-provision your hosts at much higher rates, you can use a standard pool and create thin-provisioned volumes in that pool. However, be careful that you do not run out of space. You now need to monitor the backend controller pool usage and the virtualizer pool usage in terms of volume thin provisioning over-allocation. In essence, you are double accounting with the thin provisioning; that is, expecting 2:1 on the FCM compression, and then whatever level you over-provision at the volumes.

If you know that your hosts rarely grow to use the provisioned capacity, this process can be safely done; however, the risk comes from run-away applications (writing large amounts of capacity) or an administrator suddenly enabling application encryption and writing to fill the entire capacity of the thin-provisioned volume.

4.1.2 Data reduction pools

IBM FlashSystem uses innovative DRPs that incorporate deduplication and hardware-accelerated compression technology, plus SCSI **`unmap`** support. It also uses all of the thin provisioning and data efficiency features that you expect from IBM Spectrum Virtualize-based storage to potentially reduce your CAPEX and OPEX. Also, all of these benefits extend to over 500 heterogeneous storage arrays from multiple vendors.

DRPs were designed with space reclamation being a fundamental consideration. DRPs provide the following benefits:

- Log Structured Array allocation (re-direct on all overwrites)
- Garbage collection to free whole extents
- Fine-grained (8 KB) chunk allocation/de-allocation within an extent.
- SCSI `unmap` and write same (Host) with automatic space reclamation
- Support for “back-end” unmap and write same
- Support for compression

- Support for deduplication
- Support for traditional fully allocated volumes

Data reduction can increase storage efficiency and reduce storage costs, especially for flash storage. Data reduction reduces the amount of data that is stored on external storage systems and internal drives by compressing and deduplicating capacity and providing the ability to reclaim capacity that is no longer in use.

The potential capacity savings that compression alone can provide are shown directly in the GUI interfaces by way of the included “comprestimator” functions. Since version 8.4 of the Spectrum Virtualize software, comprestimator is always on and you can see the overall expected savings in the dashboard summary view. The specific savings per volume in the volumes views also are available.

To estimate potential total capacity savings that data reduction technologies (compression and deduplication) can provide on the system, use the Data Reduction Estimation Tool (DRET). This tool is a command line, host-based utility that analyzes user workloads that are to be migrated to a new system. The tool scans target workloads on all attached storage arrays, consolidates these results, and generates an estimate of potential data reduction savings for the entire system.

You download DRET and its `readme` file to a Windows client and follow the installation instructions in the `readme`. The `readme` file also describes how to use DRET on a variety of host servers.

The DRET can be downloaded from this IBM Support web page.

To use data reduction technologies on the system, you must create a DRP, and create compressed or compressed and deduplicated volumes.

For more information, see 4.1.4, “Data reduction estimation tools” on page 125.

Quotaless data reduction child pool

From version 8.4, DRP added support for a special type of child pool, known as a *quotaless child pool*.

The concepts and high-level description of parent-child pools are the same as for standard pools with a few major exceptions.

- You cannot define a capacity or quota for a DRP child pool.
- A DRP child pool shares the same encryption key as its parent.
- Capacity warning levels cannot be set on a DRP child pool. Instead, you must rely on the warning levels of the DRP parent pool.
- A DRP child pool consumes space from the DRP parent pool as volumes are written to it.
- Child and parent pools share the same data volume; therefore, data is de-duplicated between parent and child volumes.
- A DRP child pool can use 100% of the capacity of the parent pool.
- The **`migratevdisk`** commands can now be used between parent and child pools. Because they share the encryption key, this operation becomes a “nocopy” operation.
- From code level 8.4.2.0 throttling is supported on DRP child pools.

To create a DRP child pool, use the new pool type of `child_quotaless`.

Because a DRP share capacity between volumes (when deduplication is used), it is virtually impossible to attribute capacity ownership of a specific grain to a specific volume because it might be used by two more volumes, which the is value proposition of deduplication. This process results in the differences between standard and DRP child pools.

Object-based access control (OBAC) or multi-tenancy can now be applied to DRP child pools or volumes as OBAC requires a child pool to function.

VMware vVols for DRP is not yet supported or certified at the time of this writing.

SCSI unmap

DRPs support end-to-end unmap functionality. Space that is freed from the hosts by means of a SCSI **unmap** command results in the reduction of the used space in the volume and pool.

For example, a user deletes a small file on a host, which the operating system turns into a SCSI **unmap** for the blocks that made up the file. Similarly, a large amount of capacity can be freed if the user deletes (or Vmotions) a volume that is part of a data store on a host. This process might result in many contiguous blocks being freed. Each of these contiguous blocks results in a SCSI **unmap** command being sent to the storage device.

n a DRP, when the IBM FlashSystem receives a SCSI **unmap** command, the result is that the capacity is freed that is allocated within that contiguous chunk. The deletion is asynchronous, and the unmapped capacity is first added to the "reclaimable" capacity, which is later physically freed by the garbage collection code. For more information, see 4.1.5, "Understanding capacity use in a data reduction pool" on page 130.

Similarly, deleting a volume at the DRP level frees all of the capacity back to the pool. The DRP also marks those blocks as "reclaimable" capacity, which the garbage collector later frees back to unused space. After the garbage collection frees an entire extent, a new SCSI **unmap** command is issued to the backend MDisk device.

Unmapping can help ensure good MDisk performance; for example, Flash drives can reuse the space for wear-leveling and to maintain a healthy capacity of "pre-erased" (ready to be used) blocks.

Virtualization devices like IBM FlashSystem with external storage can also forward unmap information (such as when extents are deleted or migrated) to other storage systems.

Enabling, monitoring, throttling, and disabling SCSI unmap

By default, host-based unmap support is disabled on all product other than the FlashSystem 9000 series. Backend unmap is enabled by default on all products.

To enable or disable host-based unmap, run the following command:

```
chsystem -hostunmap on|off
```

To enable or disable backend unmap run the following command:

```
chsystem -backendunmap on|off
```

You can check how much SCSI unmap processing is occurring on a per volume or per-pool basis by using the performance statistics. This information can be viewed with Spectrum Control or Storage Insights.

> **Note:** SCSI **`unmap`** might add more workload to the backend storage.
>
> Performance monitoring helps to notice possible effects and if SCSI unmap workload is affecting performance, consider taking the necessary steps and consider the data rates that are observed. It might be expected to see GiBps of unmap if you just deleted many volumes.

You can throttle the amount of “offload” operations (such as the SCSI **`unmap`** command) using the per-node settings for offload throttle. For example:

```
mkthrottle -type offload -bandwidth 500
```

This setting limits each node to 500MiBps of offload commands.

You can also stop the IBM FlashSystem from processing SCSI **`unmap`** operations for one or more host systems. You might find an over-zealous host, or not have the ability to configure the settings on some of your hosts. To modify a host to disable **`unmap`**, change the host type:

```
chhost -type generic_no_unmap <host_id_or_name>
```

If you experience severe performance problems as a result of SCSI **`unmap`** operations, you can disable SCSI **`unmap`** on the entire IBM FlashSystem for the front end (host), backend, or both.

Fully allocated volumes in a DRP

It is possible to create fully allocated volumes in a DRP.

A fully allocated volume uses the entire capacity of the volume. That is, when created that space is reserved (used) from the DRP and is not available for other volumes in the DRP.

Data will not be deduplicated or compressed in a fully allocated volume. Similarly, because it does not use the internal fine-grained allocation functions, the allocation and performance are the same or better than a fully allocated volume in a standard pool.

Compressed and deduplicated volumes in a DRP

It is possible to create compressed only volumes in a DRP.

A compressed volume is by its nature thin-provisioned. A compressed volume uses only its compressed data size in the pool. The volume grows only as you write data to it.

It is possible, but *not* recommended that you create a deduplicated-only volume in a DRP. A deduplicated volume is thin-provisioned in nature. The additional processing that is required to also compress the de-duplicated block is minimal; therefore, it is recommended that you create a compressed and de-duplicated volume rather than only a de-duplicated volume.

The DRP will first look for deduplication matches; then, it compress the data before writing to the storage.

Thin-provisioned only volumes in a DRP

It is not recommended that you create a thin-provisioned only volume in a DRP.

Thin-provisioned volumes use the fine-grained allocation functions of DRP. The main benefit of DRP is in the data reduction functions (compression and deduplication). Therefore, if you want to create a thin-provisioned volume in a DRP, create a compressed volume.

> **Note:** In some cases, when the backend storage is thin-provisioned or data reduced, the GUI might not offer the option to create only thin-provisioned volumes in a DRP. This issue occurs because it is highly recommended that you do not use this option because it can cause extreme capacity-monitoring problems with a high probability of running out of space.

DRP internal details

DRPs consists of various internal metadata volumes and it is important to understand how these metadata volumes are used and mapped to user volumes. Each user volume has a corresponding journal, forward lookup, and directory volume.

The internal layout of a DRP is different from a standard pool. A standard pool creates volume objects within the pool. Some fine grained internal metadata is stored within a thin-provisioned or real-time-compressed volume in a standard pool. Overall, the pool contains volume objects.

A DRP reports volumes to the user in the same way as a standard pool. However, internally it defines a Directory Volume for each user volume that is created within the pool. The directory points to grains of data that are stored in the Customer Data Volume. All volumes in a single DRP use the same Customer Data Volume to actually store their data. Therefore, deduplication is possible across volumes in a single DRP.

Other internal volumes are created, one per DRP. There is one Journal Volume per I/O group that can be used for recovery purposes, to replay metadata updates if needed. There is one Reverse Lookup Volume per I/O group that is used by garbage collection.

Figure 4-1 shows the difference between DRP volumes and volumes in standard pools.

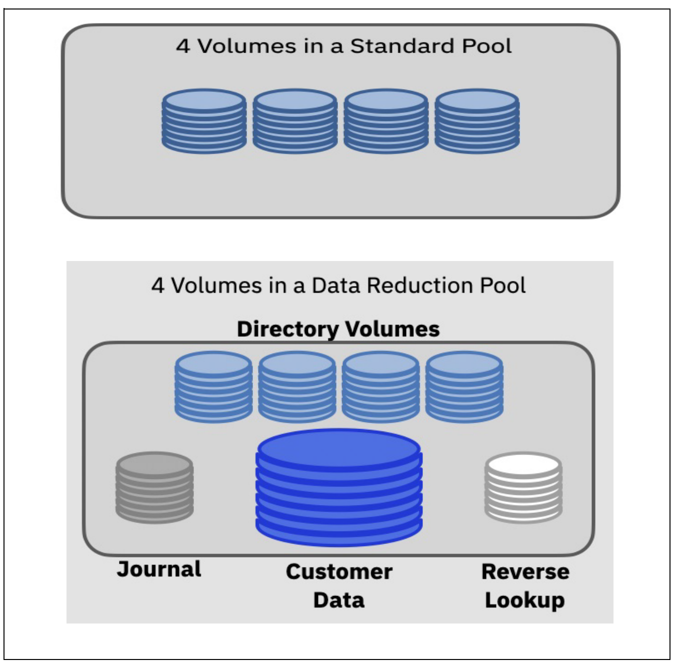


Figure 4-1 Standard and data reduction pool - volumes

The Customer Data Volume uses greater than 97% of pool capacity. The I/O pattern is a large sequential write pattern (256 KB) that is coalesced into full stride writes, and you typically see a short random read pattern.

Directory Volumes occupy approximately 1% of pool capacity. They typically have a short 4 KB random read and write I/O. The Journal Volume occupies approximately 1% of pool capacity, and shows large sequential write I/O (256 KB typically).

Journal Volumes are only read for recovery scenarios (for example, T3 recovery). Reverse Lookup Volumes are used by the garbage-collection process and occupy less than 1% of pool capacity. Reverse Lookup Volumes have a short, semi-random read/write pattern.

The primary task of garbage collection (see Figure 4-2) is to reclaim space; that is, to track all of the regions that were invalidated, and to make this capacity usable for new writes. As a result of compression and deduplication, when you overwrite a host-write, the new data does not always use the same amount of space that the previous data. This issue leads to the writes always occupying new space on back-end storage while the old data is still in its original location.

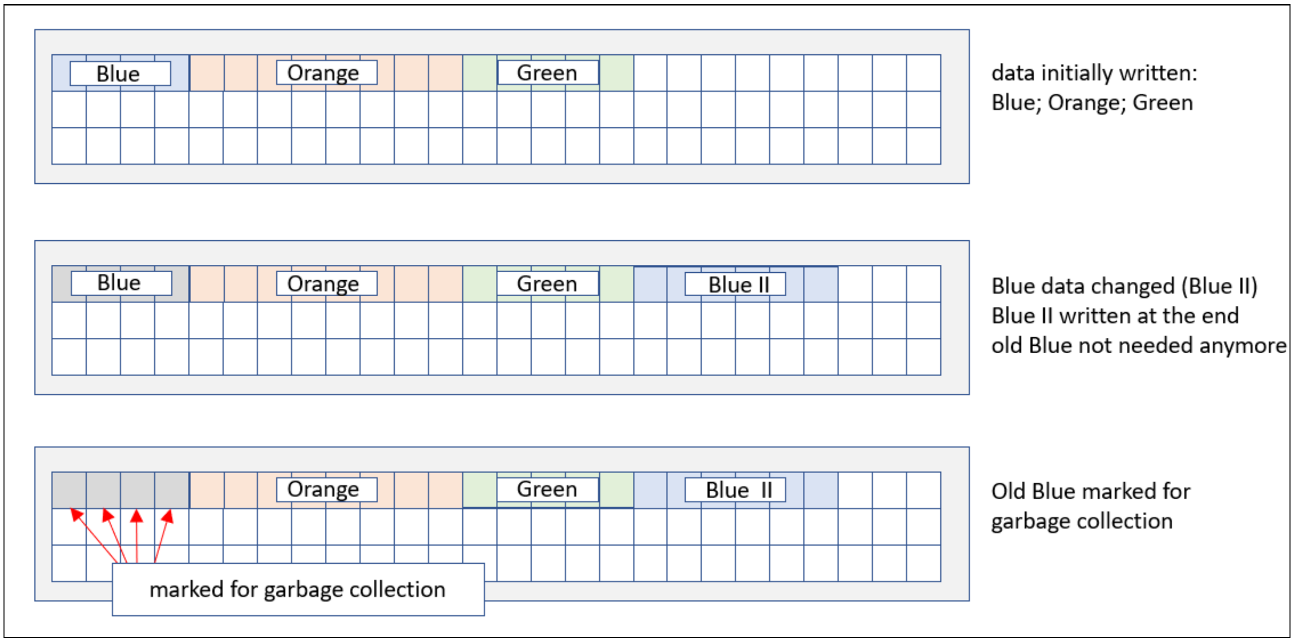


Figure 4-2 Garbage Collection principle

For garbage collection, stored data is divided into regions. As data is overwritten, a record is kept of which areas of those regions have been invalidated. Regions that have many invalidated parts are potential candidates for garbage collection. When the majority of a region has invalidated data, it is fairly inexpensive to move the remaining data to another location, therefore freeing the whole region.

DRPs include built-in services to enable garbage collection of unused blocks. Therefore, many smaller unmaps end up enabling a much larger chunk (extent) to be freed back to the pool. Trying to fill small holes is inefficient because too many I/Os are needed to keep reading and rewriting the directory. Therefore, garbage collection waits until an extent has many small holes and moves the remaining data into the extent, compacts the data, and rewrites the data. When there is an empty extent, it can be freed back to the virtualization layer (and back-end with unmap) or start writing into the extent with new data (or rewrites).

The reverse lookup metadata volume tracks the extent usage, or more importantly the holes created by overwrites or unmaps. garbage collection looks for extents with the most unused space. After a whole extent has had all valid data moved elsewhere, it can be freed back to the set of unused extents in that pool, or it can be reused for new written data.

Because garbage collection needs to move data to free regions, it is suggested that you size pools to keep a specific amount of free capacity available. This practice ensures that some free space for garbage collection. For more information, see 4.1.5, “Understanding capacity use in a data reduction pool” on page 130.

4.1.3 Standard pools versus data reduction pools

When it comes to designing pools during the planning of an IBM FlashSystem project, it is important to know all requirements, and to understand the upcoming workload of the environment. The IBM FlashSystem is flexible in creating and using pools. This section describes how to figure out which types of pool or setup you can use.

Some of the information that you should be aware of in the planned environment is as follows:

- Is your data compressible?
- Is your data deduplicable?
- What are the workload and performance requirements?
  - Read/write ratio
  - Block size
  - Input/Output Operations per Second (IOPS), MBps, and response time
- Flexibility for the future
- Thin provisioning

Determine if your data is compressible

Compression is one option of DRPs. The deduplication algorithm is used to reduce the on-disk footprint of data that is written-to by thin provisioning. In IBM FlashSystem, this compression is an inline compression or a deduplication approach rather than an attempt to compress data as a background task. DRP provides unmap support at the pool and volume level. Out-of-space situations can be managed at the DRP pool level.

Compression can be enabled in DRPs on a per-volume basis, and thin provisioning is a prerequisite. The input IO is split into a fixed 8 KiB block for internal handling, and compression is performed on each 8 K block. These compressed blocks are then consolidated into 256 K chunks of compressed data for consistent write performance by allowing the cache to build full stride writes enabling the most efficient RAID throughput.

Data compression techniques depend on the type of data that must be compressed and on the desired performance. Effective compression savings generally rely on the accuracy of your planning and the understanding if the specific data is compressible or not. Several methods are available to help you decide whether your data is compressible, including the following examples:

- General assumptions
- Tools

General assumptions

IBM FlashSystem compression is lossless; that is, data is compressed without losing any of the data. The original data can be recovered after the compress or expend cycle. Good compression savings might be achieved in the following environments (and others):

- Virtualized Infrastructure
- Database and Data Warehouse
- Home Directory, Shares, and shared project data
- CAD/CAM

- Oil and Gas data
- Log data
- SW development
- Text and some picture files

However, if the data is compressed in some cases, the savings are less, or even negative. Pictures (for example, GIF, JPG, and PNG), audio (MP3 and WMA) and video or audio (AVI and MPG) and even compressed databases data might not be good candidates for compression.

Table 4-1 lists the compression ratio of common data types and applications that provide high compression ratios

Table 4-1 Compression ratios of common data types

Data Types/Applications	Compression Ratio
Databases	Up to 80%
Server or Desktop Virtualization	Up to 75%
Engineering Data	Up to 70%
Email	Up to 80%

Also, do not compress encrypted data (for example, compression on host or application). Compressing already encrypted data does not result in many savings, because the data contains pseudo random data. The compression algorithm relies on patterns in order to gain efficient size reduction. Because encryption destroys such patterns, the compression algorithm would be unable to provide much data reduction.

For more information about compression, see 4.1.4, “Data reduction estimation tools” on page 125.

Note: Saving assumptions that are based on the type of data are imprecise. Therefore, you should determine compression savings with the proper tools.

Determine if your data is a deduplication candidate

Deduplication is done by using hash tables to identify previously written copies of data. If duplicate data is found, instead of writing the data to disk, the algorithm references the previously found data.

- Deduplication uses 8 KiB deduplication grains and an SHA-1 hashing algorithm.
- Data reduction pools build 256 KiB chunks of data consisting of multiple de-duplicated and compressed 8 KiB grains.
- Data reduction pools will write contiguous 256 KiB chunks allowing for efficient write streaming with the capability for cache and RAID to operate on full stride writes.
- Data reduction pools provide deduplication then compress capability.
- The scope of deduplication is within a DRP within an I/O Group.

General assumptions

Some environments have data with high deduplication savings, and are therefore candidates for deduplication.

Good deduplication savings can be achieved in several environments, such as virtual desktop and some virtual machine environments. Therefore, these environments might be good candidates for deduplication.

IBM provides the Data Reduction Estimate Tool (DRET) to help determine the deduplication capacity-saving benefits.

4.1.4 Data reduction estimation tools

IBM provides two tools to estimate the savings when you use data reduction technologies.

- Comprestimator

  This tool is built into the IBM FlashSystem. It reports the expected compression savings on a per-volume basis in the GUI and command line.

- Data Reduction Estimation Tool (DRET)

  The DRET tool must be installed on and used to scan the volumes that are mapped to a host and is primarily used to assess the deduplication savings. The DRET tool is the most accurate way to determine the estimated savings. However, it must scan all of your volumes to provide an accurate summary.

Comprestimator

Comprestimator is provided in the following ways:

- As a stand-alone, host-based command-line utility. It can be used to estimate the expected compression for block volumes where you do not have an IBM Spectrum Virtualize product providing those volumes.
- Integrated into the IBM FlashSystem. In software versions before 8.4, triggering a volume sampling (or all volumes) was done manually.
- Integrated into the IBM FlashSystem and always on, in versions 8.4 and later.

Host-based Comprestimator

The tool can be downloaded from this IBM Support web page.

IBM FlashSystem Comprestimator is a command-line and host-based utility that can be used to estimate an expected compression rate for block devices.

Integrated Comprestimator for software levels before 8.4.0

IBM FlashSystem also features an integrated Comprestimator tool that is available through the management GUI and CLI. If you are considering to apply compression on existing non-compressed volumes in an IBM FlashSystem, you can use this tool to evaluate if compression will generate capacity savings.

To access the Comprestimator tool in management GUI, select **Volumes** → **Volumes**.

If you want to analyze all the volumes in the system, click **Actions** → **Capacity Savings** → **Estimate Compression Savings**.

If you want to select a list of volumes and click **Actions** → **Capacity Savings** → **Analyze** to evaluate only the capacity savings of the selected volumes, as shown in Figure 4-3.

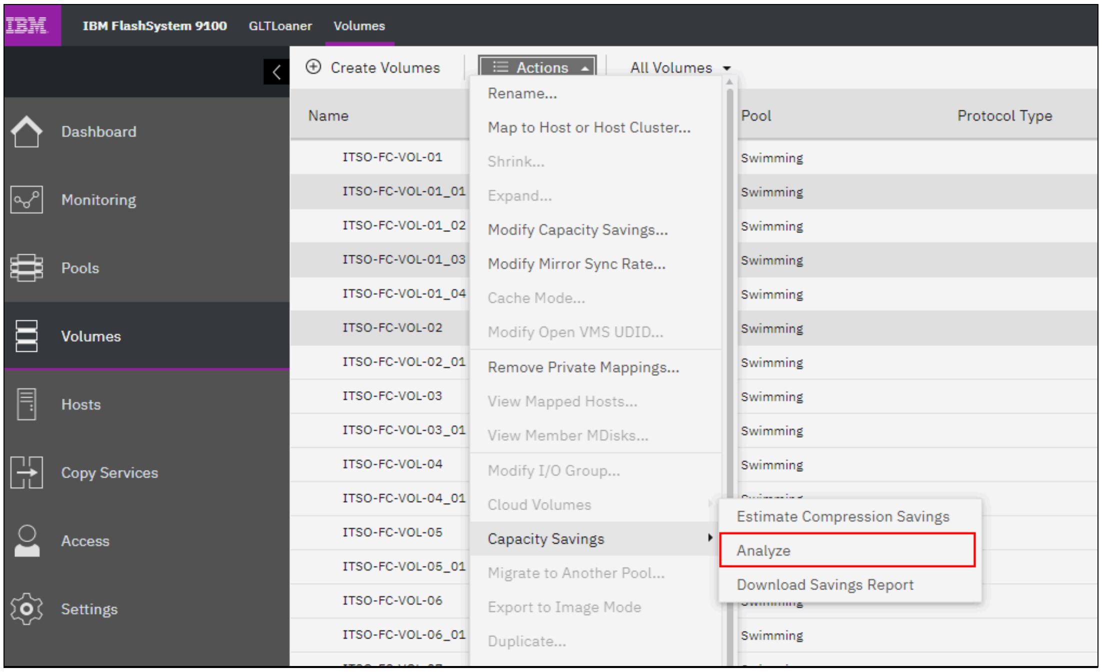


Figure 4-3 Capacity savings analysis

To display the results of the capacity savings analysis, click **Actions** → **Capacity Savings** → **Download Savings Report**, as shown in Figure 4-3, or enter the command **`lsvdiskanalysis`** in the command line, as shown in Example 4-1.

Example 4-1 Results of capacity savings analysis

```
IBM_FlashSystem:superuser>lsvdiskanalysis TESTVOL01
id 64
name TESTVOL01
state estimated
started_time 201127094952
analysis_time 201127094952
capacity 600.00GB
thin_size 47.20GB
thin_savings 552.80GB
thin_savings_ratio 92.13
compressed_size 21.96GB
compression_savings 25.24GB
compression_savings_ratio 53.47
total_savings 578.04GB
total_savings_ratio 96.33
margin_of_error 4.97
IBM_FlashSystem:superuser>
```

The following actions are preferred practices:

- After you run Comprestimator, consider applying compression only on those volumes that show greater than or equal to 25% capacity savings. For volumes that show less than 25% savings, the trade-off between space saving and hardware resource consumption to compress your data might not make sense. With DRPs, the penalty for the data that cannot be compressed is no longer seen. However, the DRP includes overhead in grain management.
- After you compress your selected volumes, review which volumes have the most space-saving benefits from thin provisioning rather than compression. Consider moving these volumes to thin provisioning only. This configuration requires some effort, but saves hardware resources that are then available to give better performance to those volumes, which achieves more benefit from compression than thin provisioning.

  You can customize the Volume view to view the metrics you might need to help make your decision, as shown in Figure 4-4.

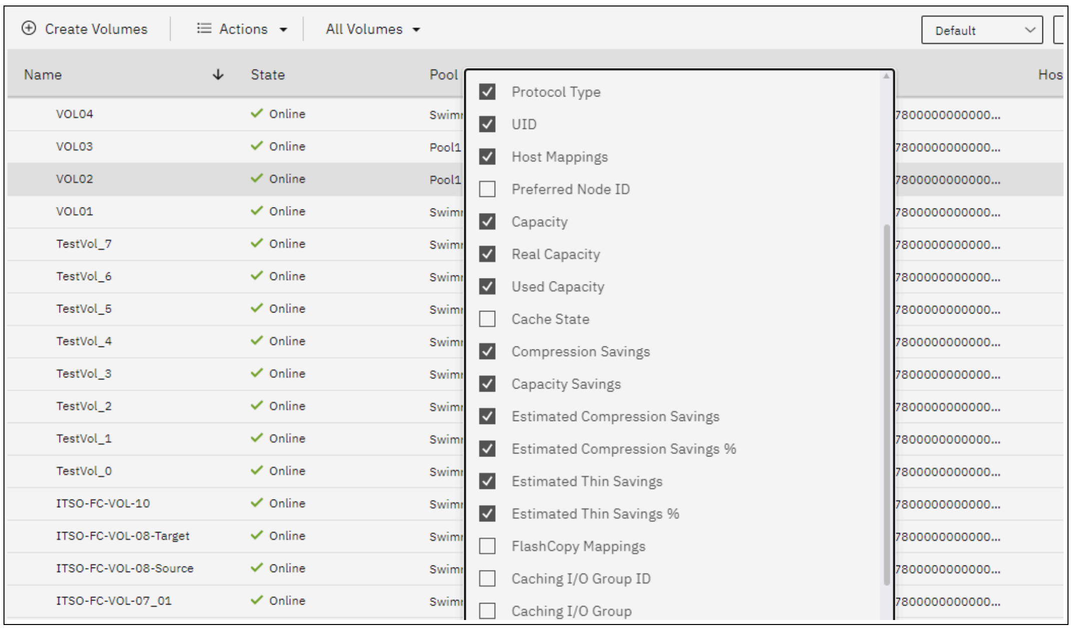


Figure 4-4 Customized view

Integrated comprestimator for software version 8.4 and onwards

Because the newer code levels include an always-on comprestimator, you can view the expected capacity savings in the main dashboard view, pool views. volume views. You do not need to first trigger the “estimate” or “analyze” tasks; these are performed automatically as background tasks.

Data Reduction Estimation Tool

IBM provides the Data Reduction Estimation Tool (DRET) to support both deduplication and compression. The host-based CLI tool scans target workloads on various older storage arrays (from IBM or another company), merges all scan results, and then provides an integrated system-level data reduction estimate for your IBM FlashSystem planning.

The DRET uses advanced mathematical and statistical algorithms to perform an analysis with a low memory “footprint”. The utility runs on a host that can access the devices to be analyzed. It performs only read operations, so it has no effect on the data stored on the device. Depending on the configuration of the environment, in many cases the DRET is used on more than one host to analyze additional data types.

It is important to understand block device behavior, when analyzing traditional (fully allocated) volumes. Traditional volumes that were created without initially zeroing the device might contain traces of old data on the block device level. Such data is not accessible or viewable on the file system level. When the DRET is used to analyze such volumes, the expected reduction results reflect the savings rate to be achieved for all the data on the block device level, including traces of old data.

Regardless of the block device type being scanned, it is also important to understand a few principles of common file system space management. When files are deleted from a file system, the space they occupied before the deletion becomes free and available to the file system. The freeing of space occurs even though the data on disk was not actually removed, but rather the file system index and pointers were updated to reflect this change.

When the DRET is used to analyze a block device used by a file system, all underlying data in the device is analyzed, regardless of whether this data belongs to files that were already deleted from the file system. For example, you can fill a 100 GB file system and use 100% of the file system, then delete all the files in the file system making it 0% used. When scanning the block device used for storing the file system in this example, the DRET (or any other utility) can access the data that belongs to the files that are deleted.

To reduce the impact of the block device and file system behavior, it is recommended that you use the DRET to analyze volumes that contain as much active data as possible rather than volumes that are mostly empty of data. The use increases the accuracy level and reduces the risk of analyzing old data that is deleted, but might still have traces on the device.

The DRET can be downloaded from this IBM Support web page.

Example 4-2 shows the DRET command line.

Example 4-2 DRET command line

```
Data-Reduction-Estimator –d <device> [-x Max MBps] [-o result data filename] [-s
Update interval] [--command scan|merge|load|partialscan]     [--mergefiles Files
to merge] [--loglevel Log Level] [--batchfile batch file to process] [-h]
```

The DRET can be used on the following client operating systems:

- Windows 2008 Server, Windows 2012
- Red Hat Enterprise Linux Version 5.x, 6.x, 7.x (64-bit)
- UBUNTU 12.04
- ESX 5.0, 5.5, 6.0
- AIX 6.1, 7.1
- Solaris 10

Note: According to the results of the DRET, use DRPs to use the available data deduplication savings, unless performance requirements exceed what DRP can deliver.

Do not enable deduplication if the data set is not expected to provide deduplication savings.

Determining the workload and performance requirements

An important factor of sizing and planning for an IBM FlashSystem environment is the knowledge of the workload characteristics of that specific environment.

Sizing and performance is affected by the following workloads, among others:

- Read/Write ratio

  Read/Write (%) ratio will affect performance because higher writes cause more IOPS to the DRP. To effectively size an environment, the Read/Write ratio should be considered. During a write I/O, when data is written to the DRP, it is stored on the data disk, the forward lookup structure is updated, and the I/O is completed.

  DRPs use metadata. Even when volumes are not in the pool, some of the space in the pool is used to store the metadata. The space that is allocated to metadata is relatively small. Regardless of the type of volumes that the pool contains, metadata is always stored separately from customer data.

  In DRPs, the maintenance of the metadata results in I/O amplification. I/O amplification occurs when a single host-generated read or write I/O results in more than one back-end storage I/O request because of advanced functions. A read request from the host results in two I/O requests, a directory lookup and a data read. A write request from the host results in three I/O requests, a directory lookup, a directory update, and a data write. Therefore, keep in mind that DRPs create *more IOPS* on the FCMs or drives.

- Block size

  The concept of a block size is simple and the impact on storage performance might be distinct. Block size effects might have an impact on overall performance. Therefore, consider that larger blocks affect performance more than smaller blocks. Understanding and considering for block sizes in the design, optimization, and operation of the storage system-sizing leads to more predictable behavior of the entire environment.

  > **Note:** Where possible limit the maximum transfer size sent to the IBM FlashSystem to no more than 256 KiB. This limitation is general best practice and not specific to only DRP.

- IOPS, MBps, and response time

  Storage constraints are IOPS, throughput, and latency, and it is crucial to correctly design the solution or plan for a setup for speed and bandwidth. Suitable sizing requires knowledge about the expected requirements.

- Capacity

  During the planning of an IBM FlashSystem environment, capacity (physical) must be sized accordingly. Compression and deduplication might save space, but metadata uses little space. For optimal performance, our recommendation is to use the DRP to a maximum of 85%.

Before planning a new environment, consider monitoring the storage infrastructure requirements with monitoring or management software (such as IBM Spectrum Control or IBM Storage Insights). At busy times, the peak workload (such as IOPS or MBps) and peak response time provide you with an understanding of the required workload plus expected growth. Also, consider allowing enough room for the performance that is required during planned and unplanned events (such as, upgrades and possible defects or failures).

It is important to understand the relevance of application response time rather than internal response time with required IOPS or throughput. Typical OLTP applications require IOPS and low latency.

Do not place capacity over performance while designing or planning a storage solution. Even if capacity might be sufficient, the environment can suffer from low performance.

Deduplication and compression might satisfy capacity needs, but aim for performance and robust application performance.

To size an IBM FlashSystem environment, your IBM account team or IBM Business Partner must access IBM Storage Modeller (StorM). The tool can be used to determine if DRPs can provide suitable bandwidth and latency. If the data does not deduplicate (according to the DRET), the volume can be either fully allocated or compressed only.

Flexibility for the future

During the planning and configuration of storage pools, you must decide which pools to create. Because the IBM FlashSystem enables you to create standard pools or DRPs, you must decide which type best fits the requirements.

Verify whether performance requirements meet the capabilities of the specific pool type. For more information, see "Determining the workload and performance requirements" on page 128.

For more information about the dependencies with child pools regarding vVols, see 4.3.3, "Data reduction pool configuration limits" on page 142, and "DRP restrictions" on page 143.

If other important factors do not lead you to choose standard pools, then DRPs are the right choice. Using DRPs can increase storage efficiency and reduce costs because it reduces the amount of data that is stored on hardware and reclaims previously used storage resources that are no longer needed by host systems.

DRPs provide great flexibility for future use because they add the ability of compression and deduplication of data at the volume level in a specific pool, even if these features are initially not used at creation time.

Note that it is not possible to convert a pool. If you must change the pool type (from standard pool to DRP, or vice versa), it will be an offline process and you will have to migrate your data as described in 4.3.6, "Data migration with DRP" on page 145.

> **Note:** We recommend the use of DRPs pools with fully allocated volumes if the restrictions and capacity do not affect your environment. For more information about the restrictions, see "DRP restrictions" on page 143.

4.1.5 Understanding capacity use in a data reduction pool

This section describes capacity terms associated with DRPs.

After a reasonable period of time, the DRP will have approximately 15-20% of overall free space. The garbage collection algorithm must balance the need to free space with the overhead of performing garbage collection. Therefore, the incoming write/overwrite rates and any unmap operations will dictate how much "reclaimable space" is present at any given time. The capacity in a DRP consists of the components that are listed in Table 4-2 on page 131.

Table 4-2 DRP capacity uses

Use	Description
Reduced Customer Data	The data that is written to the DRP, in compressed and de-duplicated form.
Fully Allocated Data	The amount of capacity allocated to fully allocated volumes (assumed to be 100% written)
Free	The amount of free space, not in use by any volume
Reclaimable Data	The amount of garbage in the pool. This is either old (overwritten) yet to be freed data or data that has is unmapped but not yet freed or associated with recently deleted volumes
Metadata	Approximately 1 - 3% overhead for DRP metadata volumes

Balancing how much garbage collection is done versus how much free space is available dictates how much reclaimable space is present at any time. The system dynamically adjusts the target rate of garbage collection to maintain a suitable amount of free space.

Figure shows an example of steady state DRP.

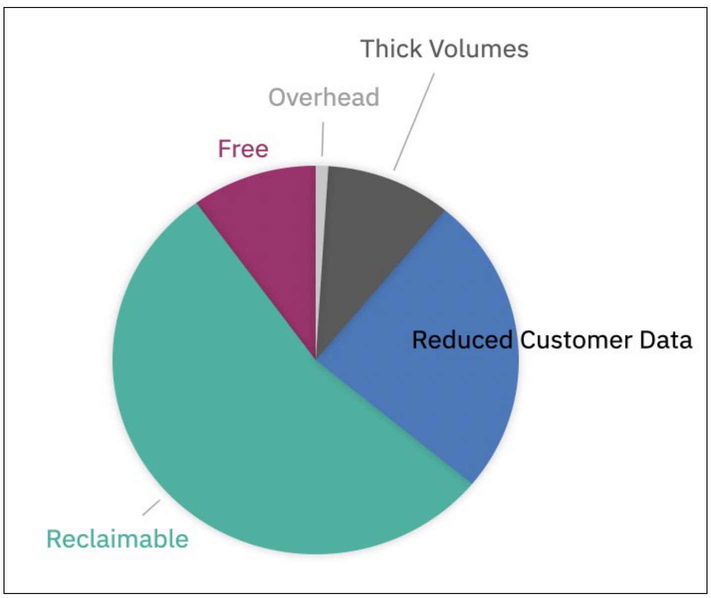


Figure 4-5 Data reduction pool capacity use example

Consider the following points:

- If you create a large capacity of fully allocated volumes in a DRP, you are taking this capacity directly from free space only. This could result in triggering heavy garbage collection if there is little free space remaining and a large amount of reclaimable space, as shown in Figure .
- If you create a large number of fully allocated volumes and experience degraded performance due to garbage collection, you can reduce the required work by temporarily deleting unused fully-allocated volumes.
- When deleting a fully-allocated volume, the capacity is returned directly to free space.

- When deleting a thin-provisioned volume (compressed or deduplicated), the following is a two-phase approach can be used:
  a. The grain must be inspected to determine if this was the last volume that referenced this grain (deduplicated):
     - If so, the grains can be freed.
     - If not, the grain references need to be updated and the grain might need to be re-homed to belong to one of the remaining volumes that still require this grain.
  b. When all grains that are to be deleted are identified, these grains are returned to the "reclaimable" capacity. It is the responsibility of garbage collection to convert them to free space.
  c. The garbage-collection process runs in the background, attempting to maintain a sensible amount of free space. If there is little free space and you delete a large number of volumes, the garbage-collection code might trigger a large amount of backend data movement and could result in performance issues.
- Deleting a volume might not immediately create free space.
- If you are at risk of running out of space, but a lot of reclaimable space exists, you can force garbage collection to work harder by creating a temporary fully allocated volume to reduce the amount of real free space and trigger more garbage collection.

Important: Use extreme caution when using up all or most of the free space with fully-allocated volumes. Garbage collection requires free space to coalesce data blocks into whole extents and hence free capacity. If little free space is available, the garbage collector must to work harder to free space.

- It might be worth creating some "get out of jail free" fully-allocated volumes in a DRP. This type of volume reserves some space that you can quickly return to the free space resources if you reach a point where you are almost out of space, or when garbage collection is struggling to free capacity in an efficient manner.

  Consider these points:
  - This type of volume should not be mapped to hosts.
  - This type of volume should be labeled accordingly. For example, "RESERVED_CAPACITY_DO_NOT_USE"

4.2 Storage pool planning considerations

The implementation of storage pools in an IBM FlashSystem requires an holistic approach that involves application availability and performance considerations. Usually a trade-off between these two aspects must be taken into account.

The main best practices in the storage pool planning activity are described in this section. Most of these practices apply to both standard and DRP pools, except where otherwise specified. For additional specific best practices for DRPs, see 4.6, "Easy Tier, tiered and balanced storage pools" on page 169. For more information, see specific practices for high-availability solutions.

4.2.1 Planning for availability

By design, IBM Spectrum Virtualize based storage systems take the entire storage pool offline if a single MDisk in that storage pool goes offline. This means that the storage pool's quantity and size define the failure domain. Reducing the hardware failure domain for back-end storage is only part of your considerations. When you are determining the storage pool layout, you must also consider application boundaries and dependencies to identify any availability benefits that one configuration might have over another.

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

Finally, increasing the number of pools to reduce the failure domain is not always a viable option. For instance, in IBM FlashSystems configurations that do not include expansion enclosures, the number of physical drives is limited (up to 24), and creating more arrays reduces the usable space because of spare and protection capacity.

Consider, for instance, a single I/O group FlashSystem configuration with 24 7.68 TB NVMe drives. In a case of a single array DRAID 6 creation, the available physical capacity would be 146.3 TB, while creating two arrays DRAID 6 would provide 137.2 TB of available physical capacity with a reduction of 9.1 TB.

When virtualizing external storage, remember that the failure domain is defined by the external storage itself, rather than by the pool definition on the front-end system. For instance, if you provide 20 MDisks from external storage and all of these MDisks are using the same physical arrays, the failure domain becomes the total capacity of these MDisks, no matter how many pools you have distributed them across.

The following actions are the starting preferred practices when planning storage pools for availability:

- Create separate pools for internal storage and external storage, unless you are creating a hybrid pool managed by Easy Tier (see 4.2.5, "External pools" on page 138).
- Create a storage pool for each external virtualized storage subsystem, unless you are creating a hybrid pool managed by Easy Tier (see 4.2.5, "External pools" on page 138).

> **Note:** If capacity from different external storage is shared across multiple pools, provisioning groups are created.
>
> IBM SAN Volume Controller detects that resources (MDisks) share physical storage and monitors provisioning group capacity; however, monitoring physical capacity must still be done. MDisks in a single provisioning group should not be shared between storage pools because capacity consumption on one pool can affect free capacity on other pools. IBM SAN Volume Controller detects this condition and shows that the pool contains shared resources.

- Use dedicated pools for image mode volumes.

> **Limitation:** Image Mode volumes are not supported with DRPs.

- For Easy Tier-enabled storage pools, always allow free capacity for Easy Tier to deliver better performance.

- Consider implementing child pools when you must have a logical division of your volumes for each application set. Cases often exist where you want to subdivide a storage pool but maintain a larger number of MDisks in that pool. Child pools are logically similar to storage pools, but allow you to specify one or more subdivided child pools. Thresholds and throttles can be set independently per child pool.

> **Note:** Throttling is supported on DRP child pools in code versions 8.4.2.0 and later.

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

4.2.2 Planning for performance

When planning storage pools for performance the capability to stripe across disk arrays is one of the most important advantages IBM Spectrum Virtualize provides. To implement performance-oriented pools, create large pools with many arrays rather than more pools with few arrays. This approach usually works better for performance than spreading the application workload across many smaller pools, because typically the workload is not evenly distributed across the volumes, and then across the pools.

Adding more arrays to a pool, rather than creating a new pool, can be a way to improve the overall performance if the added arrays have the same or better performance characteristics than the existing ones.

Note that in IBM FlashSystem configurations arrays built from FCM and SAS SSD drives have different characteristic, both in terms of performance and data reduction capabilities. Therefore, when using FCM and SAS SSD arrays in the same pool, follow these recommendations:

- Enable the Easy Tier function (see 4.6, "Easy Tier, tiered and balanced storage pools" on page 169). The Easy Tier treats the two-array technologies as different tiers (`tier0_flash` for FCM arrays and `tier1_flash` for SAS-SSD arrays), so the resulting pool is a multi-tiered pool with inter-tier balancing enabled.
- Strictly monitor the FCM physical usage. As Easy Tier moves the data between the tiers, the compression ratio can vary frequently and an out-of-space condition can be reached without changing the data contents.

The number of arrays that are required in terms of performance must be defined in the pre-sales or solution design phase, but when sizing the environment remember that adding too many arrays to a single storage pool increases the failure domain, and therefore it is important to find the trade-off between the performance, availability, and scalability cost of the solution.

Using the following external virtualization capabilities, you can boost the performance of the back-end storage systems:

- Using wide-striping across multiple arrays
- Adding additional read/write cache capability

It is typically understood that wide-striping can add approximately 10% additional Input/Output Processor (IOP) performance to the backend-system by using these mechanisms.

Another factor is the ability of the virtualized-storage subsystems to be scaled up or scaled out. For example, IBM System Storage DS8000 series is a scale-up architecture that delivers the best performance per unit, and the IBM FlashSystem series can be scaled out with enough units to deliver the same performance.

With a virtualized system, there is debate as to whether to scale out back-end system, or add them as individual systems behind IBM FlashSystem. Either case is valid. However, adding individual controllers is likely to allow IBM FlashSystem to generate more I/O, based on queuing and port-usage algorithms. It is recommended that you add each controller (I/O Group) of an IBM FlashSystem back-end as its own controller; that is, do not cluster the IBM FlashSystem when it acts as an external storage controller behind another Spectrum Virtualize product, such as IBM SAN Volume Controller. Adding each controller (I/O Group) of an IBM FlashSystem backend as its own controller adds additional management IP addresses and configuration. However, it provides the best scalability in terms of IBM FlashSystem performance.

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

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

The following actions are the starting preferred practices when planning storage pools for performance:

- Create a dedicated storage pool with dedicated resources if there is a specific performance application request.
- When using external storage in an Easy Tier enabled pool, do not intermix MDisks in the same tier with different performance characteristics.
- In a FlashSystem clustered environment, create storage pools with IOgrp or Control Enclosure affinity. That means you have to use only arrays or MDisks supplied by the internal storage that is directly connected to one IOgrp SAS chain only. This configuration avoids unnecessary IOgrp-to-IOgrp communication traversing the SAN and consuming Fibre Channel bandwidth.
- Use dedicated pools for image mode volumes.

> **Limitation:** Image Mode volumes are not supported with DRPs.

- For Easy Tier-enabled storage pools, always allow free capacity for Easy Tier to deliver better performance.
- Consider implementing child pools when you must have a logical division of your volumes for each application set. Cases often exist where you want to subdivide a storage pool but maintain a larger number of MDisks in that pool. Child pools are logically similar to storage pools, but allow you to specify one or more subdivided child pools. Thresholds and throttles can be set independently per child pool.

> **Note:** Before code version 8.4.2.0, throttling is not supported on DRP child pools.

Cache partitioning

The system automatically defines a logical cache partition per storage pool. Child pools do not count towards cache partitioning. The cache partition number matches the storage pool ID.

A cache partition is a logical threshold that stops a single partition from consuming the entire cache resource. This partition is provided as a protection mechanism and does not affect performance in normal operations. Only when a storage pool becomes overloaded, does the partitioning kick in and essentially slow down write operations in the pool to the same speed that the backend can handle. *Overloaded* means that the front-end write throughput is greater than back-end storage that the pool can sustain. This situation should be avoided.

In recent versions of IBM Spectrum Control, the fullness of the cache partition is reported and can be monitored. You should not see partitions reaching 100% full. If you do, then it suggest the corresponding storage pool is in an overload situation and workload should be moved from that pool, or additional storage capability should be added to that pool.

4.2.3 Planning for capacity

Capacity planning is never an easy task. Capacity monitoring has become more complex with the advent of data reduction. It is important to understand the terminology used to report usable, used, and free capacity.

The terminology and its reporting in the GUI changed in recent versions and is listed Table 4-3.

Table 4-3 Capacity terminology in 8.4.0

Old term	New term	Meaning
Physical Capacity	Usable Capacity	The amount of capacity that is available for storing data on a system, pool, array, or MDisk after formatting and RAID techniques are applied.
Volume Capacity	Provisioned Capacity	The total capacity of all volumes in the system.
N/A	Written Capacity	The total capacity that is written to the volumes in the system. This is shown as a percentage of the provisioned capacity and is reported before any data reduction.

The *usable capacity* describes the amount of capacity that can be written-to on the system and includes any backend data reduction (that is, the “virtual” capacity is reported to the system).

> **Note:** In DRP, the `rsize` parameter, used capacity, and tier capacity are not reported per volume. These items are reported only at the parent pool level because of the complexities of deduplication capacity reporting.

An example of the dashboard capacity view is shown in Figure 4-6.

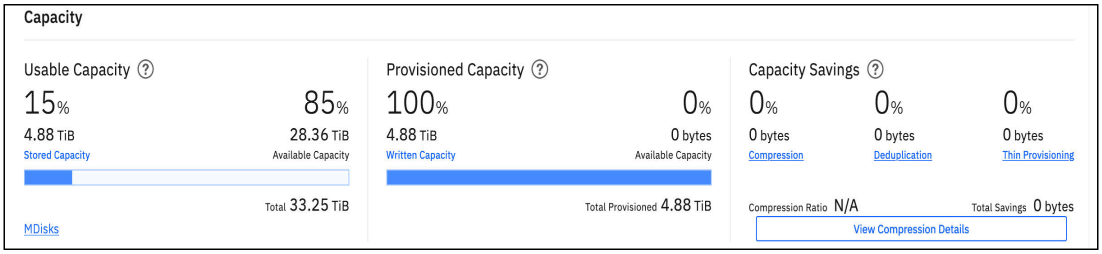


Figure 4-6 Example dashboard capacity view

For FlashCore Modules (FCM), this will be the maximum capacity that can be written to the system. However, for the smaller capacity drives (4.8 TB), this will report 20 TiB as usable. The actual usable capacity might be lower because of the actual data reduction achieved from the FCM compression.

Plan to achieve the default 2:1 compression, which is approximately an average of 10 TiB of usable space. Careful monitoring of the actual data reduction should be considered if you plan to provision to the maximum stated usable capacity when the small capacity FCMs are used.

The larger FCMs, 9.6 TB and above, report just over 2:1 usable capacity. Therefore 22, 44, and 88 for the 9.6, 19.2, and 38.4 TB modules respectively.

The *provisioned capacity* shows the total provisioned capacity in terms of the volume allocations. This is the “virtual” capacity that is allocated to fully-allocated, and thin-provisioned volumes. Therefore, it is in theory that the capacity could be written to if all volumes were filled 100% by the using system.

The *written capacity* is the actual amount of data that has been written into the provisioned capacity.

- For fully-allocated volumes, the written capacity is always 100% of the provisioned capacity.
- For thin-provisioned (including data reduced volumes), the written capacity is the actual amount of data the host writes to the volumes.

The final set of capacity numbers relates to the data reduction. This is reported in two ways:

- As the savings from DRP (compression and deduplication) provided at the DRP level, as shown in Figure 4-7 on page 138.
- As the FCM compression.

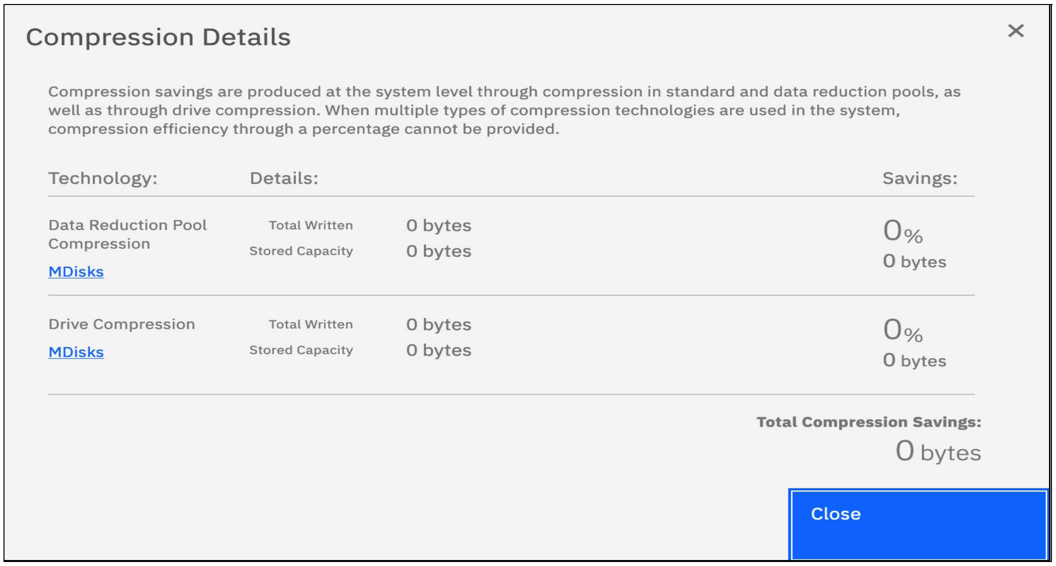


Figure 4-7 Compression Savings dashboard report

4.2.4 Extent size considerations

When adding MDisks to a pool they are logically divided into chunks of equal size. These chunks are called *extents* and are indexed internally. Extent sizes can be 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, or 8192 MB. IBM Spectrum Virtualize architecture can manage 2^22 extents for a system, and therefore the choice of extent size affects the total amount of storage that can be addressed. For the capacity limits per extent, see V8.4.2.x Configuration Limits and Restrictions for IBM FlashSystem 9200.

When planning for the extent size of a pool, remember that you cannot change the extent size later, it must remain constant throughout the lifetime of the pool.

For pool-extent size planning, consider the following recommendations:

- For standard pools, usually 1 GB is suitable.
- For DRPs, use 4 GB (see 4.6, “Easy Tier, tiered and balanced storage pools” on page 169 for further considerations on extent size on DRP).
- With Easy Tier enabled hybrid pools, consider smaller extent sizes to better utilize the higher tier resources and therefore provide better performance.
- Keep the same extent size for all pools if possible. The extent-based migration function is not supported between pools with different extent sizes. However, you can use volume mirroring to create copies between storage pools with different extent sizes.

> **Limitation:** Extent-based migrations from standard pools to DRPs are not supported unless the volume is fully allocated.

4.2.5 External pools

IBM FlashSystem-based storage systems have the ability to virtualize external storage systems. This section describes special considerations when configuring storage pools with external storage.

Availability considerations

IBM FlashSystem external storage virtualization feature provides many advantages through consolidation of storage. You must understand the availability implications that storage component failures can have on availability domains within the IBM FlashSystem cluster.

IBM Spectrum Virtualize offers significant performance benefits through its ability to stripe across back-end storage volumes. However, consider the effects that various configurations have on availability.

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

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

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

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

- It is recommended that each storage pool must contain only MDisks from a single storage subsystem. An exception exists when you are working with Easy Tier hybrid pools. For more information, see 4.6, “Easy Tier, tiered and balanced storage pools” on page 169.
- It is suggested that each storage pool contains only MDisks from a single storage tier (SSD or Flash, Enterprise, or NL_SAS) unless you are working with Easy Tier hybrid pools. For more information, see 4.6, “Easy Tier, tiered and balanced storage pools” on page 169.

IBM Spectrum Virtualize does not provide any physical-level data redundancy for virtualized external storages. The availability characteristics of the storage subsystems’ controllers have the most impact on the overall availability of the data that is virtualized by IBM Spectrum Virtualize.

Performance considerations

Performance is a determining factor, where adding IBM FlashSystem as a front-end results in considerable gains. Another factor is the ability of your virtualized storage subsystems to be scaled up or scaled out. For example:

- IBM System Storage DS8000 series is a scale-up architecture that delivers the best performance per unit.
- IBM FlashSystem series can be scaled out with enough units to deliver the same performance.

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

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

Number of MDisks per pool

The number of MDisks per pool also can effect availability and performance.

The backend storage access is controlled through MDisks where the IBM FlashSystem acts like a host to the backend controller systems. Just as you have to consider volume queue depths when accessing storage from a host, these systems must calculate queue depths to maintain high throughput capability while ensuring the lowest possible latency.

For more information about the queue depth algorithm, and the rules about how many MDisks to present for an external pool, see “Volume considerations” on page 100.

This section describes how many volumes to create on the backend controller (that are seen as MDisks by the virtualizing controller) based on the type and number of drives (such as HDD and SSD).

4.3 Data reduction pools best practices

This section describes the DRP planning and implementation best practices.

For information about estimating the deduplication ratio for a specific workload, see “Determine if your data is a deduplication candidate” on page 124.

4.3.1 Data reduction pools with IBM FlashSystem NVMe attached drives

> **Important:** If you plan to use DRP with deduplication and compression enabled with FCM storage, assume zero extra compression from the FCMs. That is, use the reported physical or usable capacity from the RAID array as the usable capacity in the pool and ignore the above maximum effective capacity.
>
> The reason for assuming zero extra compression from the FCMs is because the DRP function is sending compressed data to the FCMs, which cannot be further compressed. Therefore, the data reduction (effective) capacity savings are reported at the front-end pool level and the backend pool capacity is almost 1:1 for the physical capacity.
>
> Some small amount of other compression savings might be seen because of the compression of the DRP metadata on the FCMs.

When providing industry standard NVMe-attached flash drives capacity for the DRP, some considerations must be addressed.

The main point to consider is whether the data is deduplicable.Tools are available to provide estimation of the deduplication ratio. For more information, see “Determine if your data is a deduplication candidate” on page 124.

Consider DRP configurations with IBM FCM drives:

- Data is deduplicable. In this case, the recommendation is to use compressed and deduplicated volume type. The double compression, first from DRP and then from FCMs, will not affect the performance and the overall compression ratio.
- Data is not deduplicable. In this case, you might use standard pools (instead of DRP with FCM), and let the FCM hardware do the compression, because the overall achievable throughput will be higher.

With standard off-the-shelf NVMe drives, which do not support inline compression, similar considerations apply:

- Data is deduplicable. In this case, the recommendation is to use a compressed and deduplicated volume type. The DRP compression technology has more than enough compression bandwidth for these purposes, so compression should always be done.
- Data is not deduplicable. In this case, the recommendation is to use only a compressed volume type. The internal compression technology provides enough compression bandwidth.

Note: In general, avoid creating DRP volumes that are only deduplicated. When using DRP volumes, they should be either fully allocated, or deduplicated and compressed.

Various configuration items affect the performance of compression on the system. To attain high compression ratios and performance on your system, ensure that the following guidelines are met:

- Use FCM compression, unless your data deduplicates well with IBM FlashSystem Family products that support FCMs
- With SSD and HDD, use DRP and deduplicate if applicable with the IBM FlashSystem 5100, 7000, and 9000 family.
- Use of a small amount (1-3%) of SCM capacity in a DRP will significantly improve DRP metadata performance. As the directory data is the most frequently accessed data in the DRP and the design of DRP maintains directory data on the same extents, Easy Tier will very quickly promote the metadata extents to the fastest available tier.
- Never create a DRP with only Nearline (NL) SAS capacity. If you want to use predominantly NL SAS drives, ensure that you have a small amount of Flash or SCM capacity for the metadata.
- In general, DRP is avoided on FlashSystem 5030 unless you have few performance expectations or requirements. The FlashSystem 5030 does not have extra offload hardware and uses the internal CPU to provide the compression and decompression engine. This has limited throughput capability and is only suitable for extremely low throughput workloads. Latency also is adversely affected in most cases.
- Do not compress encrypted data. That is, if the application or operating system provides encryption, do not attempt to use DRP volumes. Data at rest encryption, which is provided by IBM FlashSystem, is still possible because the encryption is performed after the data is reduced. If host-based encryption is unavoidable, assume data reduction is not possible. That is, ensure there is a 1:1 mapping of physical-to-effective capacity.
- Although DRP and FCM do not have performance penalties if data cannot be compressed (that is, you can attempt to compress all data), the extra overhead of managing DRP volumes can be avoided by using standard pools or fully allocated volumes if no data reduction benefits are realized.
- You can use tools that estimate the compressible data, or use commonly-known ratios for common applications and data types. Storing these data types on compressed volumes saves disk capacity and improves the benefit of using compression on your system. See “Determine if your data is compressible” on page 123 for more details.
- Avoid the use of any client, file system, or application based-compression with the system compression. If this is not possible, use a standard pool for these volumes.
- Never use DRP on the IBM FlashSystem and virtualized external storage at same time (DRP over DRP). In all cases, use DRP at the virtualizer level rather than the backend storage as this simplifies capacity management and reporting.

4.3.2 DRP and external storage considerations

Avoid configurations that attempt to perform data reduction at two levels.

The recommended configuration is to run DRP at only the IBM FlashSystem that is acting as the virtualizer. For storage behind the virtualizer, you should provision fully-allocated volumes to the virtualizer.

By running in this configuration, you ensure that:

- The virtualizer understands the real physical capacity available and can warn and avoid out-of-space situations (where access is lost due to no space).
- Capacity monitoring can be wholly performed on the virtualizer level as it sees the true physical and effective capacity usage.
- The virtualizer performs efficient data reduction on previously unreduced data. Generally, the virtualizer has offload hardware and more CPU resource than the backend storage systems as it does not need to deal with RAID and so forth.

If you cannot avoid backend data reduction (for example the backend storage controller cannot disable its data reduction features), ensure that:

- You do not excessively over-provision the physical capacity on the backend.
  - For example, you have 100 TiB of real capacity. Start by presenting just 100 TiB of volumes to the IBM FlashSystem. Monitor the actual data reduction on the backend controller. If your data is reducing well over time, increase the capacity that is provisioned to the IBM FlashSystem.
  - This ensures you can monitor and validate your data reduction rates and avoids panic if you do not achieve the expected rates and have presented too much capacity to IBM FlashSystem.
- Do not run DRP on top of the backend device. Since the backend device is going to attempt to reduce the data, use a standard pool or fully-allocated volumes in the IBM FlashSystem DRP.
- Understand that IBM FlashSystem does not know the real capacity usage. You have to monitor and watch for out-of-space at the backend storage controller and the IBM FlashSystem.

Important: Never run DRP on top of DRP. This is wasteful and causes performance problems without additional capacity savings.

4.3.3 Data reduction pool configuration limits

For more information about the limitations of DRPs (IBM FlashSystem version 8.4.2) at the time of this writing, see IBM Support web page.

Since version 8.2.0, the software does not support 2145-CG8 or earlier node types. Only 2145-DH8 or later nodes support versions since 8.2.0.

For more information, see this IBM Support web page.

4.3.4 DRP provisioning considerations

This section describes practices to consider during DRP implementation.

DRP restrictions

Consider the following important restrictions when planning for a DRP implementation:

- Maximum number of supported DRPs is four.
- vVols is not currently supported in DRP.
- Volume shrinking is not supported in DRP with thin/compressed volumes.
- Non-Disruptive Volume Move (NDVM) is not supported with DRP volumes.
- The volume copy split of a Volume Mirror in a different I/O Group is not supported for DRP thinly-provisioned or compressed volumes.
- Image and Sequential mode VDisk are not supported in DRP.
- Extent level migration is not allowed between DRP unless volumes are fully allocated.
- Volume migrate for any volume type is permitted between a quotaless child and its parent DRP pool.
- A maximum of 128 K extents per `Customer Data Volume` per I/O group.
  - Therefore, the pool extent size dictates the maximum physical capacity in a pool, after data reduction.
  - Use 4 GB extent size or above.
- Recommended pool size is at least 20 TB.
- Lower than 1 PB per I/O group.
- Your pool should be no more than 85% occupied.

In addition, the following considerations apply to DRP:

- The real, used, free, and tier capacity are not reported per volume for DRP volumes. Instead, only information a pool level is available.
- Cache mode is always read/write on compressed or deduplicated volumes.
- Autoexpand is always on.
- No ability to place specific volume capacity on specific MDisks.

Extent size considerations

With DRP, the number of extents available per pool is limited by the internal structure of the pool and specifically by the size of the data volume. For more information, see 4.1.2, “Data reduction pools” on page 117.

As of this writing, the maximum number of extents supported for a data volume is 128 K. As shown in Figure 4-1 on page 121, one data volume is available per pool.

Table 4-4 lists the maximum size per pool, by extent size and I/O group number.

Table 4-4 Pool size by extent size and IO group number

Extent Size	Max size with one I/O group	Max size with two I/O groups	Max size with three I/O group	Max size with four I/O group
1024	128 TB	256 TB	384 TB	512 TB
2048	256 TB	512 TB	768 TB	1024 TB
4096	512 TB	1024 TB	1536 TB	2048 TB
8192	1024 TB	2048 TB	3072 TB	4096 TB

Considering that the extent size cannot be changed after the pool is created, it is recommended that you carefully plan the extent size according to the environment capacity requirements. For most of the configurations, an extent size of 4 GB is recommended for DRP.

Pool capacity requirements

A minimum capacity must be provisioned in a DRP to provide capacity for the internal metadata structures. Table 4-5 shows the minimum capacity that is required by extent size and I/O group number.

Table 4-5 Minimum recommended pool size by extent size and IO group number

Extent Size	Min size with one I/O group	Min size with two I/O group	Min size with three I/O group	Min size with four I/O group
1024	255 GB	516 GB	780 GB	1052 GB
2048	510 GB	1032 GB	1560 GB	2104 GB
4096	1020 GB	2064 GB	3120 GB	4208 GB
8192	2040 GB	4128 GB	6240 GB	8416 GB

Note that the values reported in Table 4-5 represent the minimum required capacity for a DRP to create a single volume.

When sizing a DRP, it is important to remember that the garbage collection process is constantly running to reclaim the unused space, which optimizes the extents usage. For more information on the garbage collection process, see “DRP internal details” on page 121.

This garbage-collection process then requires a certain amount of free space to work efficiently. For this reason it is recommended to keep approximately 15% free space in a DRP pool. For more information, see this IBM Support web page.

4.3.5 Standard and DRP pools coexistence

While homogeneous configurations in terms of pool type are preferable, there is no technical reason to avoid using standard and DRP pools in the same system. In some circumstances, this coexistence is unavoidable. Consider the following scenarios:

- IBM FlashSystem installation that require VMware vVols support and data reduction capabilities for other environments. This scenario requires the definition of both standard and DRP pools because of the restriction of DRP regarding the vVols. For more information, see “DRP restrictions” on page 143.

- In this case, the standard pool will be used for vVols environments only, while the DRP will be used for the other environments. Note that some data-reduction capability can be achieved for the vVols standard pool by using the inline data compression provided by the IBM FCMs on FlashSystem.
- IBM FlashSystem installations that require an external pool for image mode volumes and data reduction capabilities for other environments. Also, this scenario requires the definition of standard pools and DRPs because of the restriction of DRPs regarding the Image mode volumes. For more information, see "DRP restrictions" on page 143.

  In this case, the standard pool will be used for Image mode volumes only, optionally with the write cache disabled if needed for the back-end native copy services usage. For more information, see Chapter 6, "IBM FlashCopy services capabilities" on page 243. DRP is used for all the other environments.
- IBM FlashSystem installation that includes a FlashSystem system with DRP capabilities as an external pool. In this scenario, the external pool must be a standard pool, as recommended in 4.3.2, "DRP and external storage considerations" on page 142. In this case, the internal storage can be defined in a separate DRP enabling the data reduction capabilities if needed.
- IBM FlashSystem installation that requires more than four pools.

4.3.6 Data migration with DRP

As mentioned in "DRP restrictions" on page 143, extent-level migration to and from a DRP (such as migrate-volume or migrate-extent functions) is not supported. For an existing IBM FlashSystem configuration, where you plan to move data to or from a DRP and use of data reduced volumes, there are two options: host-based migrations and volume mirroring based migrations.

Host-based migration

Host-based migration uses operating-system features or software tools that run on the hosts to concurrently move data to the normal host operations. VMware vMotion and AIX Logical Volume Mirroring are two examples of these features. When you use this approach, a specific amount of capacity on the target pool is required to provide the migration target volumes.

The process includes the following steps:

1. Create the target volumes of the migration in the target pool. Note that, depending on the migration technique, the size and the amount of the volumes can be different from the original ones. For example, you can migrate two 2 TB VMware datastore volumes in a single 4 TB datastore volume.
2. Map the target volumes to the host.
3. Rescan the HBAs to attach the new volumes to the host.
4. Activate the data move or mirroring feature from the old volumes to the new ones.
5. Wait until the copy is complete.
6. Detach the old volumes from the host.
7. Unmap and remove the old volumes from the IBM FlashSystem.

When migrating data to a DRP, consider the following options:

- Migrate directly to compressed or deduplicated volumes. With this option, the migration duration mainly depends on the host-migration throughput capabilities. Consider that the target volumes are subject to high write-workload, which can use many resources because of the compression and deduplication tasks.

  To avoid a potential effect on performance on the workload, try to limit the migration throughput at the host level. If this limit cannot be used, implement the throttling function at the volume level.

- Migrate first to fully-allocated volumes and then convert them to compressed or deduplicated volumes. Also, with this option, the migration duration mainly depends on the host capabilities, but usually more throughput can be sustained because there is no overhead for compression and deduplication. The space-saving conversion can be done using the volume mirroring feature.

Volume mirroring based migration

The volume mirroring feature can be used to migrate data from a pool to another pool and at the same time, change the space saving characteristics of a volume. Like host-based migration, volume mirroring-based migration requires free capacity on the target pool, but it is not needed to create volumes manually.

Volume mirroring migration is a three-step process:

- 1. Add a volume copy on the DRP and specify the wanted data reduction features.
- 2. Wait until the copies are synchronized.
- 3. Remove the original copy.

With volume mirroring, the throughput of the migration activity can be adjusted at a volume level by specifying the Mirror Sync Rate parameter. Therefore, if performance is affected, the migration speed can be lowered or even suspended.

> **Note:** Volume mirroring supports only two copies of a volume. If a configuration uses both copies, one of the copies must be removed first before you start the migration. The volume copy split of a Volume Mirror in a different I/O Group is not supported for DRP thin-provisioned or compressed volumes.

4.4 Operations with storage pools

In the following section we describe some guidelines for the typical operation with pools, which apply both to standard and DRP pool type.

4.4.1 Creating data reduction pools

This section describes how to create DRPs.

Using the management GUI

To create DRPs by using the management GUI, complete the following steps:

1. Create a DRP, as shown in Figure 4-8:
   a. In the management GUI, select **Pools** → **Pools**.
   b. On the **Pools** page, click **Create**.
   c. On the **Create Pool** page, enter a name for the pool and select **Data Reduction**.
   d. Click **Create**.

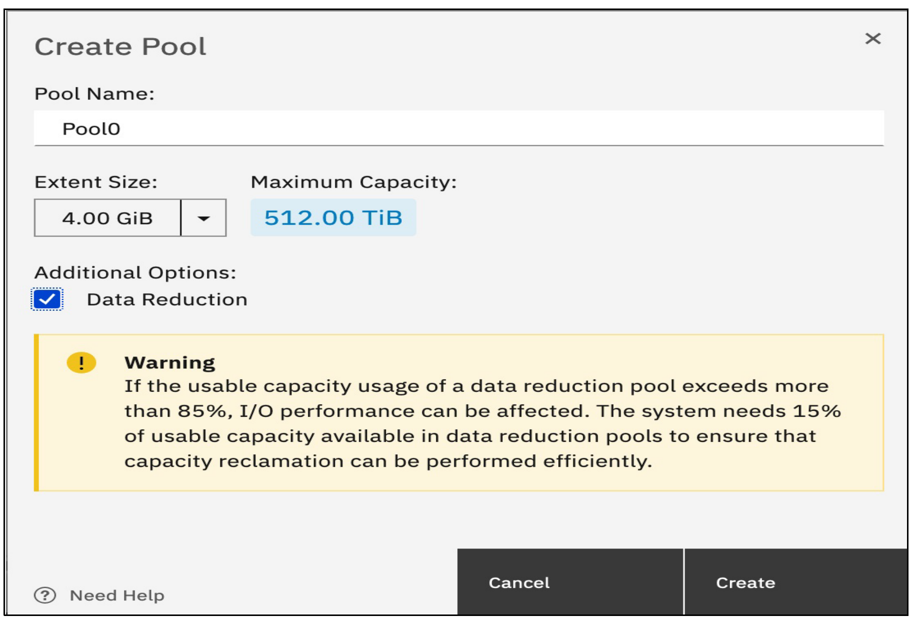


Figure 4-8 Create pool page

2. Create a Data Reduction child pool, as shown in Figure 4-10 on page 148:
   a. In the management GUI, select **Pools** → **Pools**.
   b. Right-click the parent pool you want to create the child pool in, as shown in Figure 4-9 on page 148.

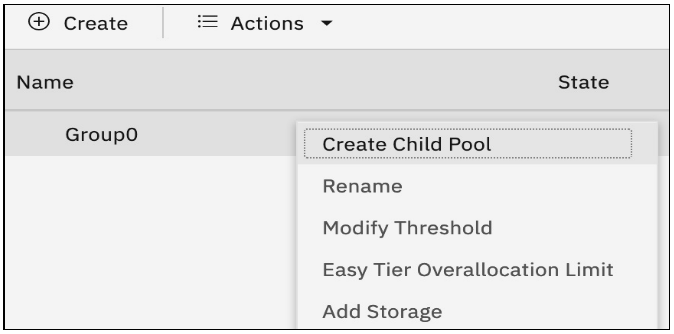


Figure 4-9 Right-click parent pool actions menu

   c. Select **Create Child Pool**.
   d. Enter a name for the child pool, as shown in Figure 4-10.
   e. Click **Create**.

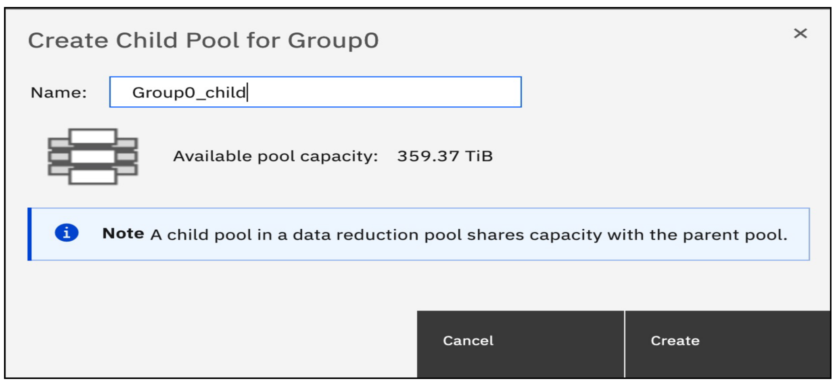


Figure 4-10 Create child pool page

3. Add storage to a parent DRP by completing these steps:
   a. In the management GUI, select **Pools** → **Pools**.
   b. Right-click the DRP that you created and select **Add Storage**.
   c. Select from the available storage and allocate capacity to the pool. Click **Assign**.
4. Create fully-allocated, compressed, deduplicated, or a combination of compressed and deduplicated volumes in the DRP and map them to hosts by completing the following steps:
   a. In the management GUI, select **Volumes** → **Volumes**.
   b. On the **Volumes** page, click **Create Volumes**.
   c. On the **Create Volume** page, select the type of volume that you want to create.
   d. Enter the following information for the volume:
      - Pool

        Select a DRP from the list. Compressed, thin-provisioned, and deduplicated volumes, and copies, must be in DRPs.
      - Volume details

        Enter the quantity, capacity, and name for the volume or volumes that you are creating.

- Capacity savings

  Select **None** (fully-allocated), or **Compressed**. When compressed is selected, you also can select to use deduplication for the volume that you create.

  > **Note:** If your system contains self-compressed drives, ensure that the volume is created with Compression enabled. If not, the system cannot calculate accurate available physical capacity.

e. Click **Create and Map**, as shown in Figure 4-11.

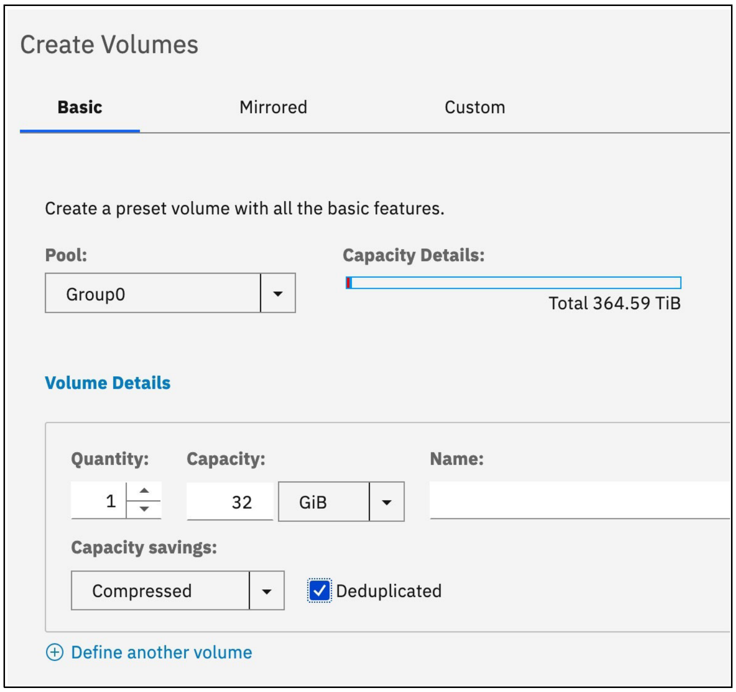


Figure 4-11 Create Volume page

> **Note:** Select **Create** to create the volumes in the DRP without mapping to hosts. If you want to map volumes to hosts later, select **Hosts** → **Hosts** → **Add Hosts**.

f. On the **Create Mapping** page, select **Host** to display all hosts that are available for mapping. Hosts must support SCSI **`unmap`** commands. Verify that the selected host type supports SCSI **`unmap`** commands. Click **Next**.

g. Starting with version 8.3.1 the system will try to map the SCSI LUN ID the same on all Host clusters, if you want to assign specific IDs then select the **Self Assign** checkbox.

h. Verify the volume, and then click **Map Volumes** (see Figure 4-12.

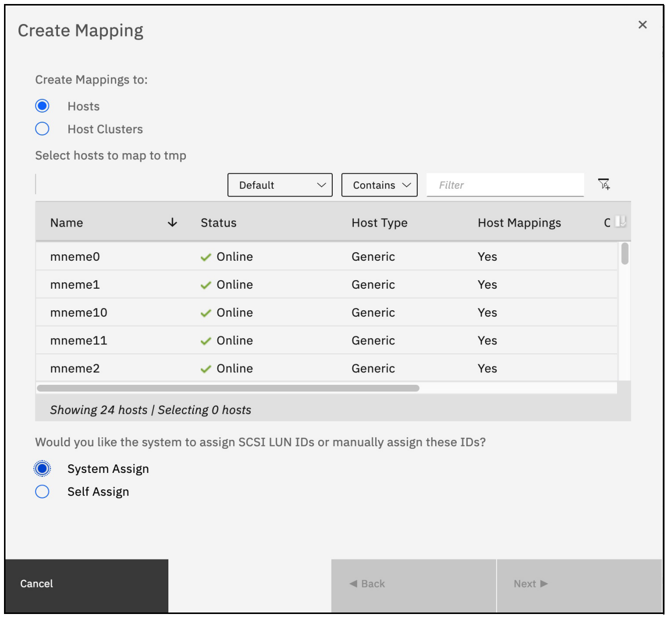


Figure 4-12 The Create Mapping page

Using the command line interface

To create DRPs by using the CLI, complete the following steps:

1. To create a DRP, enter the following command:

   ```
   mkmdiskgrp -name pool_name -ext extent_size -datareduction yes
   ```

   Where ***pool_name*** is the name of the pool and ***extent_size*** is the extent size of the pool. You can create DRPs only as parent pools, not child pools.

2. To create a compressed volume within a DRP, enter the following command:

   ```
   mkvolume -name name -pool storage_pool_name -size disk_size -compressed
   ```

3. Where ***name*** is the name of the new volume, ***storage_pool_name*** is the name of the DRP, and ***disk_size*** is the capacity of the volume.

4. To map the volume to a host, enter the following command:

   ```
   mkvdiskhostmap -host host_name vdisk_name
   ```

   Where ***host_name*** is the name of the host and ***vdisk_name*** is the name of the volume.

For more information, see this IBM Documentation web page.

Monitor the physical capacity of DRPs in the management GUI by selecting **Pools** → **Pools**. In the command line interface, use the **lsmdiskgrp** command to display the physical capacity of a DRP.

4.4.2 Adding external MDisks to existing storage pools

If MDisks are being added to an IBM FlashSystem cluster, it is likely because you want to provide more capacity. In Easy Tier enabled pools, the storage-pool balancing feature guarantees that the newly added MDisks are automatically populated with extents that come from the other MDisks. Therefore, manual intervention is not required to rebalance the capacity across the available MDisks.

> **Important:** When adding external MDisks, the system does not know to which tier the MDisk belongs. You must ensure that you specify or change the tier type to match the tier type of the MDisk.
>
> This specification is vital to ensure that Easy Tier keeps a pool as a single tier pool and balances across all MDisks, or Easy Tier adds the MDisk to the correct tier in a multitier pool.
>
> Failure to set the correct tier type creates a performance problem that might be difficult to diagnose in the future.

The tier_type can be changed using the CLI:

```
chmdisk -tier <new_tier> <mdisk>
```

For more information see 4.6.9, “Easy Tier settings” on page 189.

Adding MDisks to storage pools is a simple task, but it is suggested that you perform some checks in advance especially when adding external MDisks.

Checking access to new MDisks

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

In IBM Spectrum Virtualize, there is a feature that tests an MDisk automatically for reliable read/write access before it is added to a storage pool. Therefore, user action is not required. The test fails under the following conditions:

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

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

Persistent reserve

A common condition where external MDisks can be configured by IBM FlashSystem, but cannot perform read/write, is when a persistent reserve is left on a LUN from a previously attached host.

In this condition, rezone the back-end storage and map them back to the host that is holding the reserve. Alternatively, map them to another host that can remove the reserve by using a utility such as the Microsoft Windows SDD Persistent Reserve Tool.

4.4.3 Renaming MDisks

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

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

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

4.4.4 Removing MDisks from storage pools

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

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

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

Migrating extents from the MDisk to be deleted

If an MDisk contains volume extents, you must move these extents to the remaining MDisks in the storage pool. Example 4-3 shows how to list the volumes that have extents on an MDisk by using the CLI.

Example 4-3 Listing of volumes that have extents on an MDisk to be deleted

```
IBM_2145:itsosvccl1:admin>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
```

DRP restriction: The **`lsmdiskextent`** command does not provide accurate extent usage for thin-provisioned or compressed volumes on DRPs.

Specify the **-force** flag on the **rmmdisk** command, or select the corresponding option in the GUI. Both actions cause IBM FlashSystem 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.

Verifying the identity of an MDisk before removal

External MDisks must appear to the IBM FlashSystem cluster as unmanaged before their controller LUN mapping is removed. Unmapping LUNs from IBM FlashSystem 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 using the preferred practices, the correct LUNs are easier to identify. However, ensure that the identification of LUNs that are being unmapped from the controller match the associated MDisk on IBM FlashSystem by using the Controller LUN Number field and the unique identifier (UID) field.

The UID is unique across all MDisks on all controllers. However, the controller LUN is unique only within a specified controller and for a certain host. Therefore, when you use the controller LUN, check that you are managing the correct storage controller and that you are looking at the mappings for the correct IBM FlashSystem 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 "Correlating the back-end volume with the MDisk" on page 153.

Correlating the back-end volume with the MDisk

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

DS8000 LUN

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

To get the WWNN of the DS8000 controller, take the first 16 digits of the MDisk UID and change the first digit from `6` to `5`, such as `6005076305ffc74c` to `5005076305ffc74c`. When detected as IBM FlashSystem `ctrl_LUN_#`, the DS8000 LUN is decoded as `40XX40YY00000000`, where `XX` is the logical subsystem (LSS) and `YY` is the LUN within the LSS. As detected by the DS8000, the LUN ID is the four digits starting from the 29th digit, as shown in the Example 4-4.

Example 4-4 DS8000 UID example

```
6005076305ffc74c000000000000100700000000000000000000000000000000
```

In Example 4-4, you can identify the MDisk supplied by the DS8000, which is LUN ID `1007`.

XIV system volumes

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

- Serial number: `897`
- LUN: `2`

Complete the following steps:

1. To identify the volume serial number, right-click a volume and select **Properties**. Example 4-5 on page 154 shows the Volume Properties dialog box that opens.
2. To identify your LUN, in the volumes by Hosts view, expand your IBM FlashSystem host group and then review the LUN column, as shown in Example 4-5.
3. The MDisk UID field consists of part of the controller WWNN from bits 2 - 13. You might check those bits by using the **`lscontroller`** command, as shown in Example 4-5.

   Example 4-5 The lscontroller command

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

4. The correlation can now be performed by taking the first 16 bits from the MDisk UID field:
   - Bits 1 - 13 refer to the controller WWNN, as shown in Example 4-5.
   - Bits 14 - 16 are the XIV volume serial number (897) in hexadecimal format (resulting in 381 hex).
   - The translation is `0017380002860381000000000000000000000000000000000000000000000000`,

     where:

     - The controller WWNN (bits 2 - 13) is `0017380002860`
     - The XIV volume serial number that is converted in hex is `381`
5. To correlate the IBM FlashSystem **`ctrl_LUN_#`**:

   a. Convert the XIV volume number in hexadecimal format.

   b. Check the last three bits from the IBM FlashSystem ctrl_LUN_#.

   In this example, the number is `0000000000000002`, as shown in Figure 4-13 on page 155.

FlashSystem volumes

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

Complete the following steps to correlate the IBM FlashSystem volumes with the external MDisks that are seen by the virtualizer:

1. From the back-end IBM FlashSystem side, check the Volume UID field for the volume that was presented to the virtualizer, as shown in Figure 4-13 on page 155.

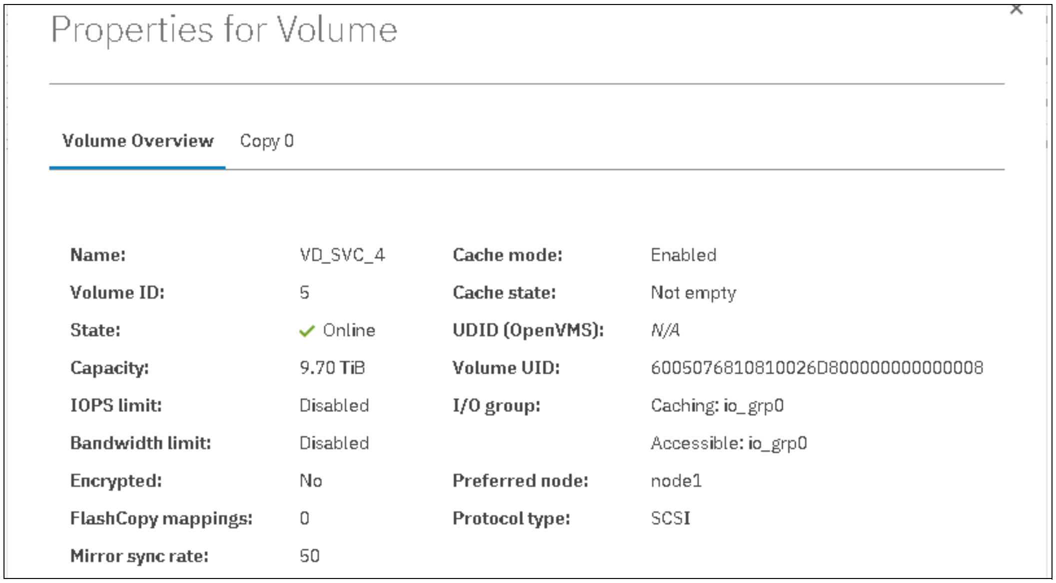


Figure 4-13 FlashSystem volume details

2. On the Host Maps tab, check the SCSI ID number for the specific volume, as shown in Figure 4-14. This value is used to match the virtualizer `ctrl_LUN_#` (in hexadecimal format).

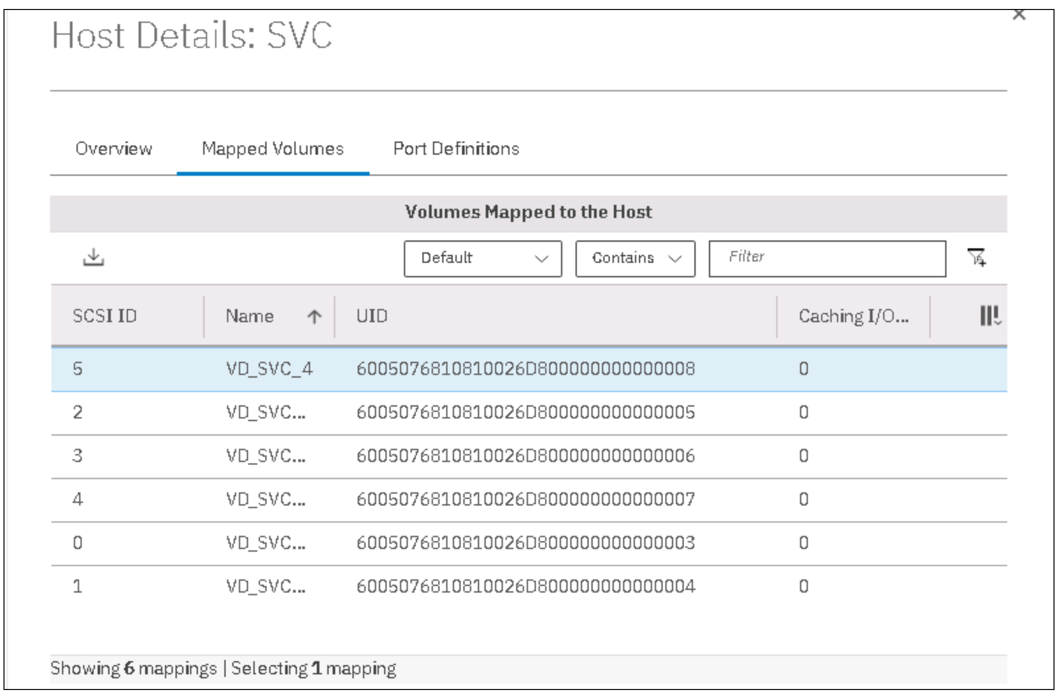


Figure 4-14 FlashSystem volume details for host maps

3. At the virtualizer, review the MDisk details and compare the MDisk UID field with the FlashSystem Volume UID, as shown in Figure 4-15. The first 32 bits should be the same.

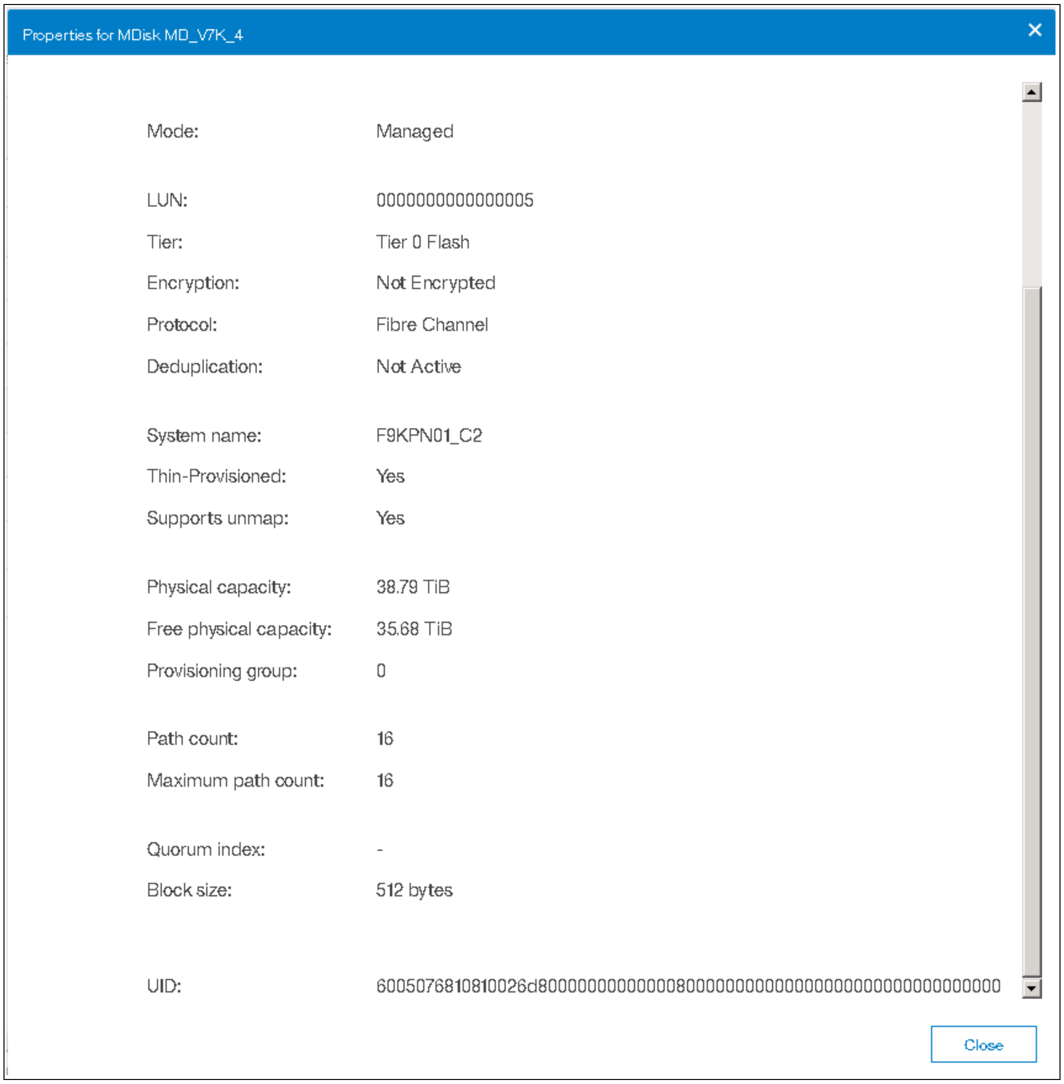


Figure 4-15 IBM SAN Volume Controller MDisk details for IBM FlashSystem volumes

4. Double-check that the virtualizer ctrl_LUN_# is the IBM FlashSystem SCSI ID number in hexadecimal format. In this example, the number is `0000000000000005`.

4.4.5 Remapping managed MDisks

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

If the LUN is mapped back with different attributes, IBM FlashSystem recognizes this MDisk as a new MDisk. In this case, 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.

4.4.6 Controlling extent allocation order for volume creation

When creating a new volume on a standard pool, the allocation of extents is performed using a round-robin algorithm, taking one extent from each MDisk in the pool in turn.

The first MDisk to allocate an extent from is chosen in a pseudo-random way rather than always starting from the same MDisk. The pseudo-random algorithm avoids the situation where the “striping effect” inherent in a round-robin algorithm places the first extent for many volumes on the same MDisk.

Placing the first extent of a number of volumes on the same MDisk might lead to poor performance for workloads that place a large I/O load on the first extent of each volume or that create multiple sequential streams.

However, this allocation pattern is unlikely to remain for long because Easy Tier balancing begins to move the extents to balance the load evenly across all MDisk in the tier. The hot and cold extents also are moved between tiers.

In a multi-tier pool, the middle tier is used by default for new volume creation. If free space is not available in the middle tier, the cold tier will be used if it exists. If the cold tier does not exist, the hot tier will be used. For more information on Easy Tier, see 4.6, “Easy Tier, tiered and balanced storage pools” on page 169.

> **DRP restriction:** With compressed and deduplicated volumes on DRP, the extent distribution cannot be checked across the MDisks. Initially, only a minimal number of extents are allocated to the volume, based on the `rsize` parameter.

4.5 Considerations when using encryption

IBM SAN Volume Controller (since 2145-DH8) and all IBM FlashSystem support optional encryption of data at rest. This support protects against the potential exposure of sensitive user data and user metadata that is stored on discarded, lost, or stolen storage devices. To use encryption on the system, an encryption license is required for each IBM FlashSystem I/O Group that support encryption.

> **Note:** Consider the following points:
>
> - Check if you have the required IBM Security™ Key Lifecycle Manager licenses on hand. Consider redundancy and high-availability regarding Key Lifecycle Manager servers.
> - In IBM Spectrum Virtualize code level V8.2.1 and later, Gemalto Safenet KeySecure also is supported. In code level V8.4.1 and later, Thales CipherTrust Manager is supported. For more information about the supported key servers see, this IBM Support web page.

4.5.1 General considerations

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

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

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

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

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

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

Several factors must be considered when planning for encryption:

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

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

For configuration details about IBM FlashSystem encryption, see the following publications:

- *Implementing the IBM FlashSystem with IBM Spectrum Virtualize V8.4.2*, SG24-8506
- *Implementing the IBM SAN Volume Controller with IBM Spectrum Virtualize V8.4.2,* SG24-8507

4.5.2 Hardware and software encryption

Encryption can be performed in IBM FlashSystem devices by using one of two methods: hardware encryption and software encryption.

Both methods protect against the potential exposure of sensitive user data that are stored on discarded, lost, or stolen media. Both methods also can facilitate the warranty return or disposal of hardware. The method to be used for encryption is chosen automatically by the system based on the placement of the data.

Figure 4-16 shows encryption placement in the lower layers of the IBM FlashSystem software stack.

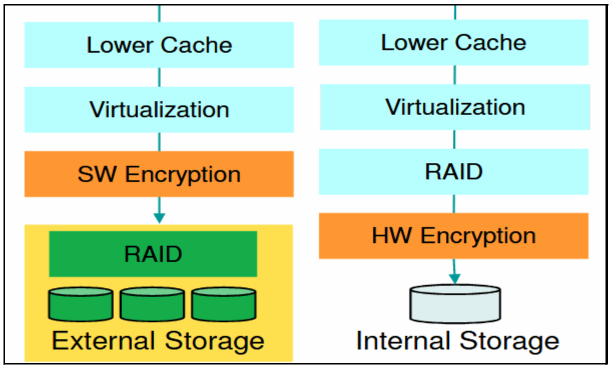


Figure 4-16 Encryption placement in lower layers of the IBM FlashSystem software stack

Hardware encryption only storage pool

Hardware encryption features the following characteristics:

- Algorithm is built in SAS chip for all SAS attached drives, or built into the drive itself for NVMe attached drives (FCM, IS NVMe and SCM).
- No system overhead.
- Only available to direct attached SAS disks.
- Can only be enabled when you create internal arrays.
- Child pools cannot be encrypted if the parent storage pool is not encrypted.
- Child pools are automatically encrypted if the parent storage pool is encrypted, but can have different encryption keys.
- DRP child pools can only use the same encryption key as their parent.

Software encryption only storage pool

Software encryption features the following characteristics:

- The algorithm is running at the interface device driver.
- Uses special CPU instruction set and engines (AES_NI).
- Allows encryption for virtualized external storage controllers, which are not capable of self-encryption.
- Less than 1% system overhead.
- Only available to virtualized external storage.
- Can only be enabled when you create storage pools and child pools made up of virtualized external storage.
- Child pools can be encrypted even if the parent storage pool is not encrypted.

Mixed encryption in a storage pool

It is possible to mix hardware and software encryption in a storage pool, as shown in Figure 4-17.

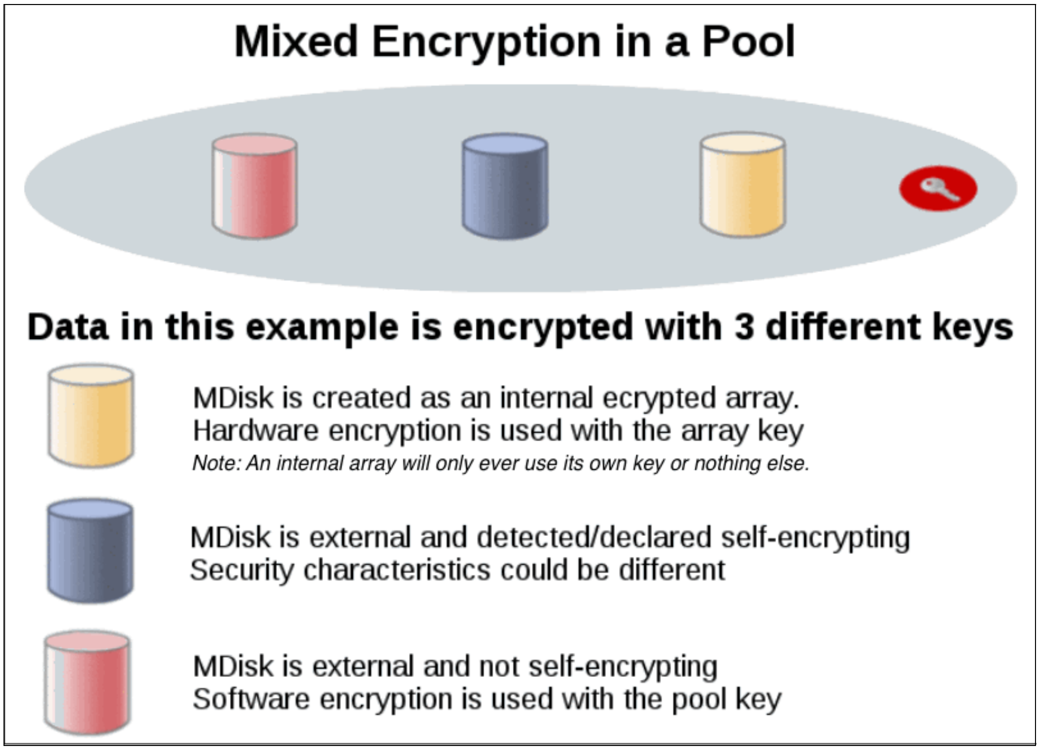


Figure 4-17 Mixed encryption in a storage pool

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

- The parent pool must not contain any unencrypted internal arrays.
- All IBM FlashSystem nodes in the system must support software encryption and have an activated encryption license.

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

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

> **Note:** Always implement encryption on the self-encryption capable back-end storage, such as IBM FlashSystem, IBM Storwize, IBM XIV, IBM FlashSystem A9000, and IBM DS8000, to avoid potential system overhead.

Declare or identify the self-encrypted virtualized external MDisks as encrypted on IBM FlashSystem by setting the **-encrypt** option to **yes** in the **chmdisk** command, as shown in Example 4-6. This configuration is important to avoid IBM FlashSystem trying to encrypt them again.

Example 4-6 Command to declare or identify a self-encrypted MDisk from a virtualized external storage

```
IBM_2145:ITSO_DH8_A:superuser>chmdisk -encrypt yes mdisk0
```

Note: It is important to declare or identify the self-encrypted MDisks from a virtualized external storage before creating an encrypted storage pool or child pool on IBM FlashSystem.

4.5.3 Encryption at rest with USB keys

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

USB flash drives have the following characteristics:

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

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

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

  If you want the system to restart automatically, a USB flash drive must be left inserted in all the nodes on the system. When you power on, all nodes then have access to the encryption key. This method requires that the physical environment where the system is located is secure. If the location is secure, it prevents an unauthorized person from making copies of the encryption keys, stealing the system, or accessing data that is stored on the system.
- USB flash drives are not left inserted into the system except as required

  For the most secure operation, do not keep the USB flash drives inserted into the nodes on the system. However, this method requires that you manually insert the USB flash drives that contain copies of the encryption key in the nodes during operations that the system requires an encryption key to be present. USB flash drives that contain the keys must be stored securely to prevent theft or loss.

4.5.4 Encryption at rest with key servers

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

Key servers

Key servers have the following characteristics:

- Physical access to the system is not required to process a rekeying operation.
- Support for businesses that have security requirements not to use USB ports.

- Strong key generation.
- Key self-replication and automatic backups.
- Implementations follow an open standard that aids in interoperability.
- Audit detail.
- Ability to administer access to data separately from storage devices.

Encryption key servers create and manage encryption keys that are used by the system. In environments with a large number of systems, key servers distribute keys remotely without requiring physical access to the systems. A key server is a centralized system that generates, stores, and sends encryption keys to the system. If the key server provider supports replication of keys among multiple key servers, you can specify up to 4 key servers (one master and three clones) that connect to the system over both a public network or a separate private network.

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

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

Replication must complete before keys can be used on the system. You can either schedule automatic replication or complete it manually with IBM Security Key Lifecycle Manager. During replication, key servers are not available to distribute keys or accept new keys. The time a replication completes on the IBM Security Key Lifecycle Manager depends on the number of key servers that are configured as clones, and the amount of key and certificate information that is being replicated.

The IBM Security Key Lifecycle Manager issues a completion message when the replication completes. Verify that all key servers contain replicated key and certificate information before keys are used on the system.

Recommendations for key server configuration

The following section provides some essential recommendations for key server configuration with IBM FlashSystem.

Transport Layer Security

Define the IBM Security Key Lifecycle Manager to use Transport Layer Security version 2 (TLSv2).

The default setting on IBM Security Key Lifecycle Manager since version 3.0.1 is TLSv1.2, but the IBM FlashSystem only supports version 2. On the IBM Security Key Lifecycle Manager, set the value to `SSL_TLSv2`, which is a set of protocols that includes TLSv1.2.

For more information about the protocols, see this IBM Documentation web page.

Example 4-7 shows the example of a `SKLMConfig.properties` configuration file. The default path on a Linux based server is `/opt/IBM/WebSphere/AppServer/products/sklm/config/SKLMConfig.properties`.

Example 4-7 Example of a SKLMConfig.properties configuration file

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

Self-signed certificate type and validity period

The default certificate type on IBM Security Key Lifecycle Manager server and IBM FlashSystem is RSA. If you use different certificate type, make sure you match the certificate type on both end. The default certificate validity period is 1095 days on IBM Security Key Lifecycle Manager server and 5475 days on IBM FlashSystem.

You can adjust the validity period to comply with specific security policies and always match the certificate validity period on IBM FlashSystem and IBM Security Key Lifecycle Manager server. A mismatch will cause certificate authorization error and lead to unnecessary certificate exchange.

Figure 4-18 shows the default certificate type and validity period on IBM FlashSystem.

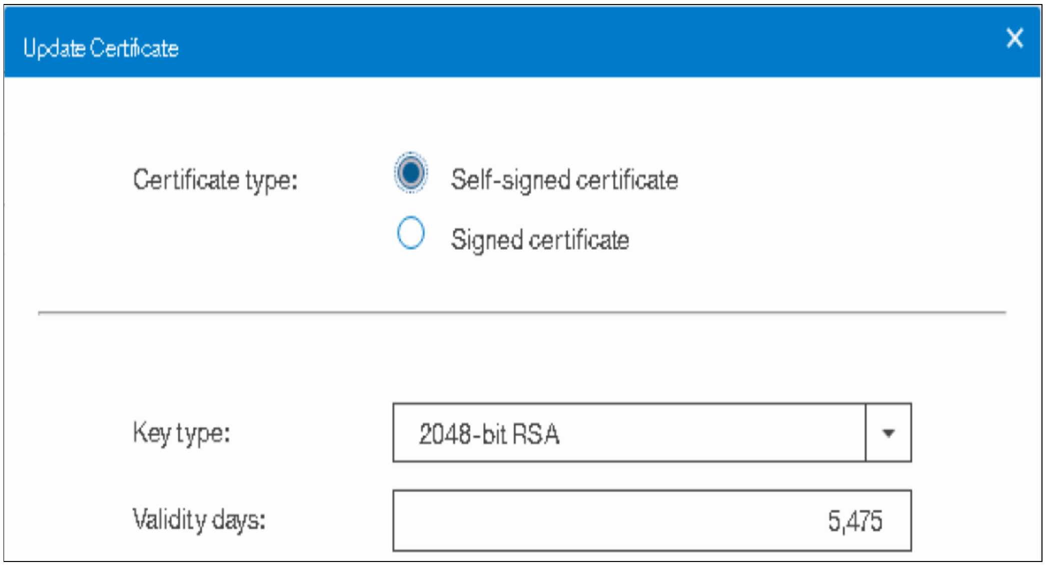


Figure 4-18 Update certificate on IBM FlashSystem

Figure 4-19 shows the default certificate type and validity period on IBM Security Key Lifecycle Manager server.

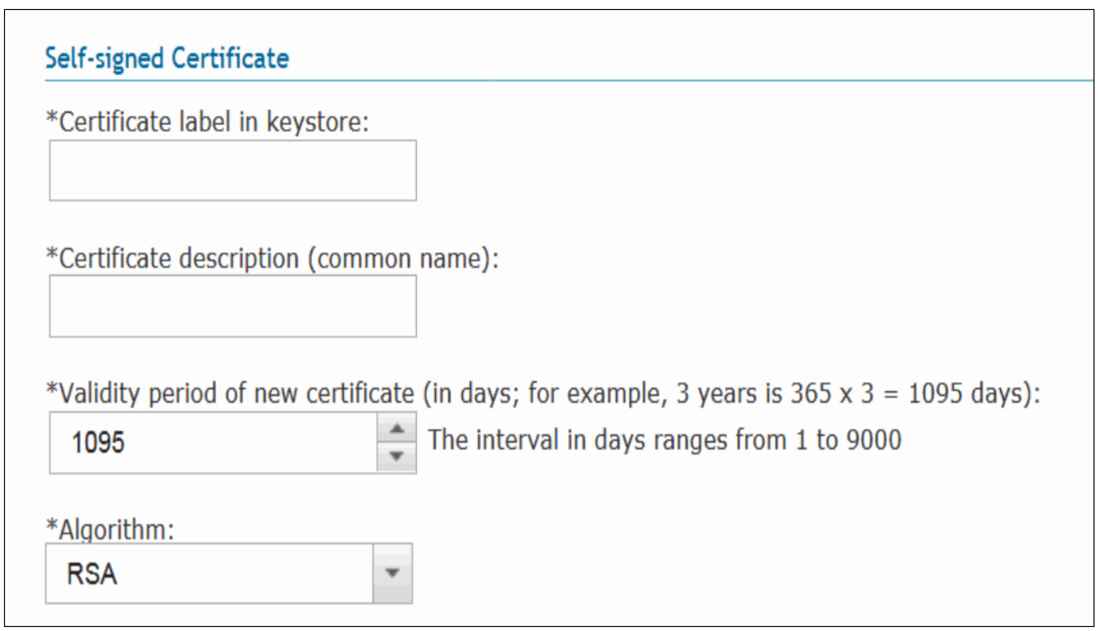


Figure 4-19 Create self-signed certificate on IBM Security Key Lifecycle Manager server

Device group configuration

The `SPECTRUM_VIRT` device group is not predefined on IBM Security Key Lifecycle Manager, it must be created based on a GPFS device family as shown in Figure 4-20.

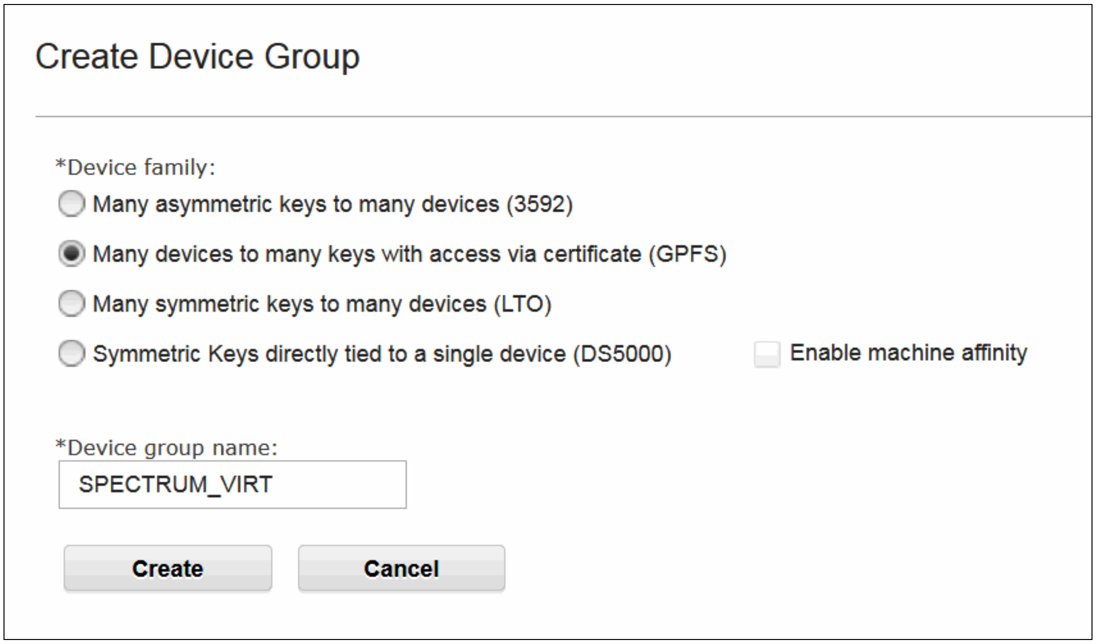


Figure 4-20 Create device group for IBM FlashSystem

By default, IBM FlashSystem the `SPECTRUM_VIRT` group name is predefined in the encryption configuration wizard. `SPECTRUM_VIRT` contains all the keys for the managed IBM FlashSystem. However, It is possible to use different device groups as long as they are GPFS device family based. For example, one device group for each environment (Production or Disaster Recovery (DR)). Each device group maintains its own key database, and this approach allows more granular key management.

Clone servers configuration management

The minimum replication interval on IBM Security Key Lifecycle Manager is one hour, as shown in Figure 4-21. It is more practical to perform backup and restore or manual replication for the initial configuration to speed up the configuration synchronization.

Also, the rekey process creates a new configuration on the IBM Security Key Lifecycle Manager server, and it is important not to wait for the next replication window but to manually synchronize the configuration to the additional key servers (clones). Otherwise, an error message is generated by the IBM FlashSystem system, which indicates that the key is missing on the clones.

Figure 4-21 shows the replication interval.

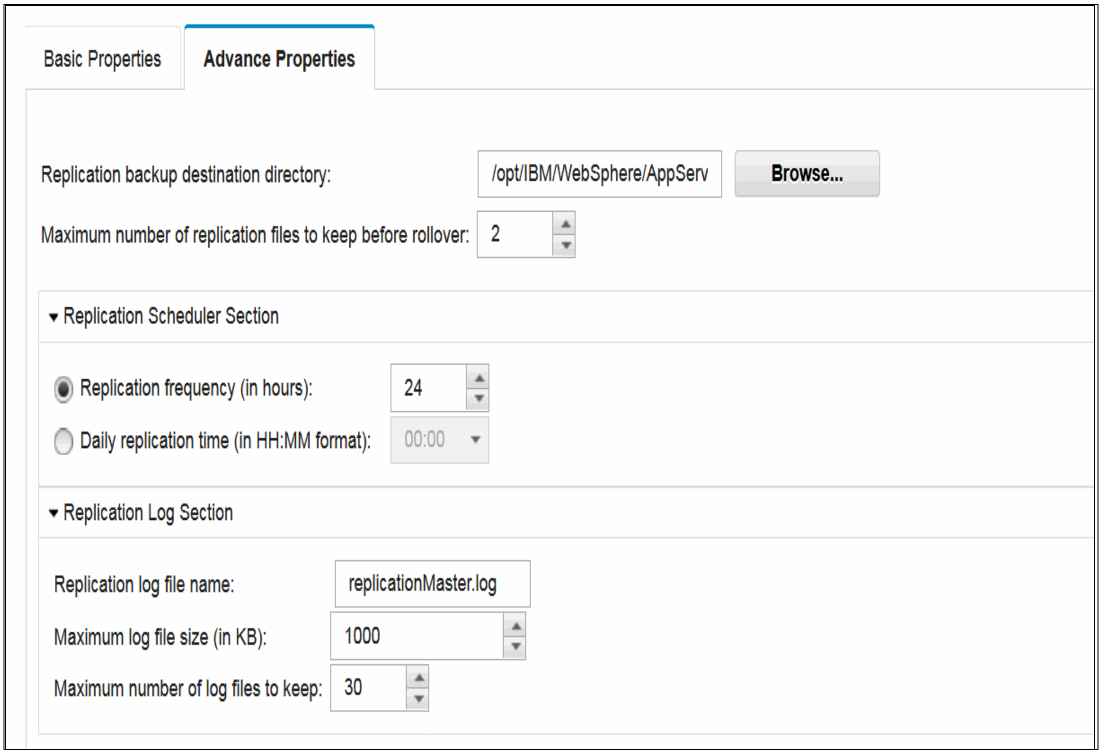


Figure 4-21 SKLM Replication Schedule

Example 4-8 shows an example of manually triggered replication.

Example 4-8 Manually triggered replication

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

Encryption key management

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

Figure 4-22 on page 167 shows the keys associated with a device group. In this example, the `SG247933_REDBOOK` device group contains one encryption-enabled IBM FlashSystem, and it has three associated keys. Only one of the keys is activated, and the other two were deactivated after the rekey process.

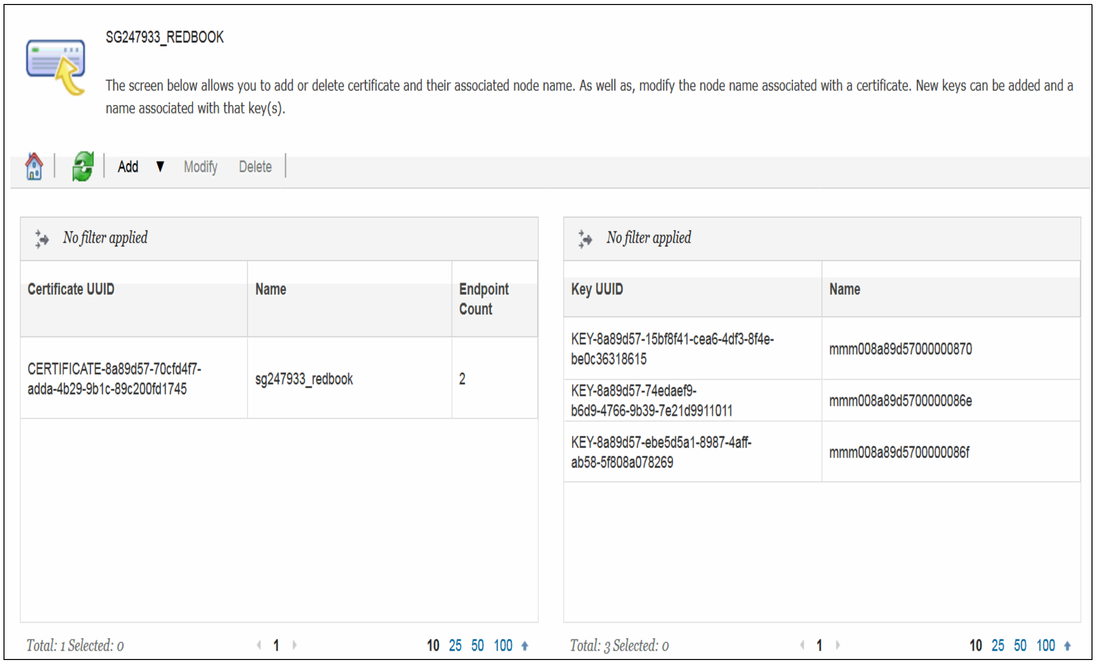


Figure 4-22 Keys associated to a device group

Example 4-9 shows an example to check the state of the keys.

Example 4-9 Verify key state

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

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

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

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

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

```
CTGKM0001I Command succeeded.

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

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

If encryption was enabled on a pre-V7.8.0 code level system and the system is updated to V7.8.x or above, you must run a USB rekey operation to enable key server encryption. Run the **chencryption** command before you enable key server encryption. To perform a rekey operation, run the commands that are shown in Example 4-10.

Example 4-10 Commands to enable key server encryption option on a system upgraded from pre-7.8.0

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

For more information about Encryption with Key Server, see this IBM Documentation web page.

4.6 Easy Tier, tiered and balanced storage pools

Easy Tier was originally developed to provide the maximum performance benefit from a few SSDs or flash drives. Because of their low response times, high throughput, and IOPS-energy-efficient characteristics, SSDs and flash arrays were a welcome addition to the storage system, but initially their acquisition cost per Gigabyte (GB) was more than for HDDs.

By implementing an evolving almost AI-like algorithm, Easy Tier moved the most frequently accessed blocks of data to the lowest latency device. Therefore, it provides an exponential improvement in performance when compared to a small investment in SSD and flash capacity.

The industry moved on in the more than 10 years since Easy Tier was first introduced. The cost of SSD and flash-based technology meant that more users can deploy all-flash environments.

HDD-based large capacity NL-SAS drives are still the most cost-effective online storage devices. Although SSD and flash ended the 15 K RPM and 10 K RPM drive market, it has yet to reach a price point that competes with NL-SAS for lower performing workloads. The use cases for Easy Tier changed, and most deployments now use “flash and trash” approaches, with 50% or more flash capacity and the remainder using NL-SAS.

Easy Tier also provides balancing within a tier. This configuration ensures that no one single component within a tier of the same capabilities is more heavily loaded than another. It does so to maintain an even latency across the tier and help to provide consistent and predictable performance.

As the industry strives to develop technologies that can enable higher throughput and lower latency than even flash, Easy Tier continues to provide user benefits. For example, Storage Class Memory (SCM) technologies, which were introduced to FlashSystem in 2020, now provide lower latency than even flash, but as with flash when first introduced, at a considerably higher cost of acquisition per GB.

Choosing the correct mix of drives and the data placement is critical to achieve optimal performance at the lowest cost. Maximum value can be derived by placing “hot” data with high I/O density and low response time requirements on the highest tier, while targeting lower tiers for “cooler” data, which is accessed more sequentially and at lower rates.

Easy Tier dynamically automates the ongoing placement of data among different storage tiers. It also can be enabled for internal and external storage to achieve optimal performance.

Also, the Easy Tier feature that is called *storage pool balancing* automatically moves extents within the same storage tier from overloaded to less loaded MDisks. Storage pool balancing ensures that your data is optimally placed among all disks within storage pools.

Storage pool balancing is designed to balance extents between tiers in the same pool to improve overall system performance and to avoid overloading a single MDisk in the pool.

However, considers only performance, it does *not* consider capacity. Therefore, if two FCM arrays are in a pool and one of them is nearly out of space and the other is empty, Easy Tier does not attempt to move extents between the arrays.

For this reason, it is recommended that if you must increase the capacity on an MDisk, increase the size of the array rather than add an FCM array.

4.6.1 Easy Tier concepts

IBM FlashSystem products implement Easy Tier enterprise storage functions, which were originally designed in conjunction with the development of Easy Tier on IBM DS8000 enterprise class storage systems. It enables automated subvolume data placement throughout different or within the same storage tiers. This feature intelligently aligns the system with current workload requirements and optimizes the usage of high-performance storage, such as SSD, flash and SCM.

Easy Tier reduces the I/O latency for hot spots, but it does not replace storage cache. Both Easy Tier and storage cache solve a similar access latency workload problem. However, these two methods weigh differently in the algorithmic construction that is based on *locality of reference*, recency, and frequency. Because Easy Tier monitors I/O performance from the device end (after cache), it can pick up the performance issues that cache cannot solve, and complement the overall storage system performance.

The primary benefit of Easy Tier is to reduce latency for hot spots; however, this feature also includes an added benefit where the remaining “medium” (that is, not cold) data has less contention for its resources and performs better as a result (that is, lower latency).

In addition, Easy Tier can be used in a single tier pool to balance the workload across storage MDisks. Is also ensures an even load on all MDisks in a tier or pool. Therefore, bottlenecks and convoying effects are removed when striped volumes are used. In a multitier pool, each tier is balanced.

In general, the storage environment’s I/O is monitored at a volume level, and the entire volume is always placed inside one suitable storage tier. Determining the amount of I/O, moving part of the underlying volume to an appropriate storage tier, and reacting to workload changes is too complex for manual operation. It is in this situation that the Easy Tier feature can be used.

Easy Tier is a performance optimization function that automatically migrates extents that belong to a volume between different storage tiers (see Figure 4-23 on page 171) or the same storage tier (see Figure 4-25 on page 178). Because this migration works at the extent level, it is often referred to as sublogical unit number (LUN) migration. Movement of the extents is dynamic, nondisruptive, and is not visible from the host perspective. As a result of extent movement, the volume no longer has all its data in one tier; rather, it is in two or three tiers, or is balanced between MDisks in the same tier.

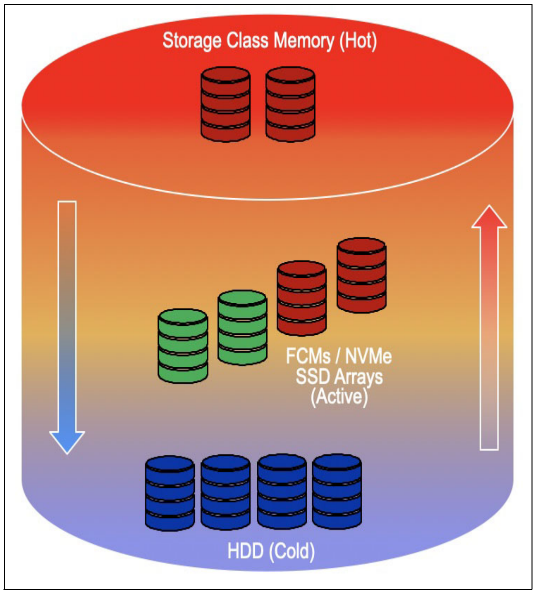


Figure 4-23 Easy Tier single volume, multiple tiers

You can enable Easy Tier on a per volume basis, except for non-fully allocated volumes in a DRP where Easy Tier is always enabled. It monitors the I/O activity and latency of the extents on all Easy Tier enabled volumes.

Based on the performance characteristics, Easy Tier creates an extent migration plan and dynamically moves (promotes) high activity or hot extents to a higher disk tier within the same storage pool. Generally, a new migration plan is generated on a stable system once every 24 hours. Instances might occur when Easy Tier reacts within 5 minutes; for example, when detecting an overload situation.

It also moves (demotes) extents whose activity dropped off, or cooled, from higher disk tier MDisks back to a lower tier MDisk. When Easy Tier runs in a storage pool rebalance mode, it moves extents from busy MDisks to less busy MDisks of the same type.

Note: Image mode and sequential volumes are not candidates for Easy Tier automatic data placement since all extents for those types of volumes must be on one specific MDisk, and cannot be moved.

4.6.2 Easy Tier definitions

Easy Tier measures and classifies each extent into one of its three tiers. It performs this classification process by looking for extents that are the outliers in any system:

1. It looks for the hottest extents in the pool. These extents contain the most frequently accessed data of a suitable workload type (less than 64 KiB I/O). Easy Tier plans to migrate these extents into whatever set of extents that come from MDisks that are designated as the hot tier.
2. It looks for coldest extents in the pool, which are classed as having done < 1 I/O in the measurement period. These extents are planned to be migrated onto extents that come from the MDisks that are designated as the cold tier.It is not necessary for Easy Tier to look for extents to place in the middle tier. By definition, if something is not designated as “hot” or “cold”, it stays or is moved to extents that come from MDisks in the middle tier.

With these three tier classifications, an Easy Tier pool can be optimized.

Internal processing

The Easy Tier function includes the following four main processes:

- I/O Monitoring

  This process operates continuously and monitors volumes for host I/O activity. It collects performance statistics for each extent, and derives averages for a rolling 24-hour period of I/O activity.

  Easy Tier makes allowances for large block I/Os; therefore, it considers only I/Os of up to 64 kilobytes (KiB) as migration candidates.

  This process is efficient and adds negligible processing resource to the IBM FlashSystem nodes.

- Data Placement Advisor (DPA)

  The DPA uses workload statistics to make a cost-benefit decision as to which extents are to be candidates for migration to a higher performance tier.

  This process also identifies extents that can be migrated back to a lower tier.

- Data Migration Planner (DMP)

  By using the extents that were previously identified, the DMP builds the extent migration plans for the storage pool. The DMP builds two plans:

  - The Automatic Data Relocation (ADR mode) plan to migrate extents across adjacent tiers.
  - The Rebalance (RB mode) plan to migrate extents within the same tier.

- Data migrator

  This process involves the actual movement or migration of the volume's extents up to, or down from, the higher disk tier. The extent migration rate is capped so that a maximum of up to 12 GiB every five minutes is migrated, which equates to approximately 3.4TiB per day that is migrated between disk tiers.

Note: You can increase the target migration rate to 48 GiB every five minutes by temporarily enabling accelerated mode. See "Easy Tier acceleration" on page 190.

When active, Easy Tier performs the following actions across the tiers:

- Promote

  Moves the hotter extents to a higher performance tier with available capacity. Promote occurs within adjacent tiers.

- Demote

  Demotes colder extents from a higher tier to a lower tier. Demote occurs within adjacent tiers.

- Swap

  Exchanges cold extent in an upper tier with hot extent in a lower tier.

- Warm demote

  Prevents performance overload of a tier by demoting a warm extent to a lower tier. This process is triggered when bandwidth or IOPS exceeds predefined threshold. If you see these operations, it is a trigger to suggest you should add additional capacity to the higher tier.

- Warm promote

  This feature addresses the situation where a lower tier suddenly becomes very active. Instead of waiting for the next migration plan, Easy Tier can react immediately. Warm promote acts in a similar way to warm demote. If the 5-minute average performance shows that a layer is overloaded, Easy Tier immediately starts to promote extents until the condition is relieved. This is often referred to as “overload protection”.

- Cold demote

  Demotes inactive (or cold) extents that are on a higher performance tier to its adjacent lower-cost tier. In that way Easy Tier automatically frees extents on the higher storage tier before the extents on the lower tier become hot. Only supported between HDD tiers.

- Expanded cold demote

  Demotes appropriate sequential workloads to the lowest tier to better use nearline disk bandwidth.

- Auto rebalance

  Redistributes extents within a tier to balance usage across MDisks for maximum performance. This process moves hot extents from high used MDisks to low used MDisks, and exchanges extents between high-use MDisks and low-use MDisks.

- Space reservation demote

  Introduced in code version 8.4.0, to prevent out-of-space conditions, EasyTier stops the migration of new data into a tier, and if necessary, migrates extents to a lower tier.

Easy Tier attempts to migrate the most active volume extents up to SSD first.

If a new migration plan is generated before the completion of the previous plan, the previous migration plan and queued extents that are not yet relocated are abandoned. However, migrations that are still applicable are included in the new plan.

> **Note:** Extent migration occurs only between adjacent tiers. For instance, in a three-tiered storage pool, Easy Tier will not move extents from the flash tier directly to the nearline tier and vice versa without moving them first to the enterprise tier.

Easy Tier extent migration types are shown in Figure 4-24.

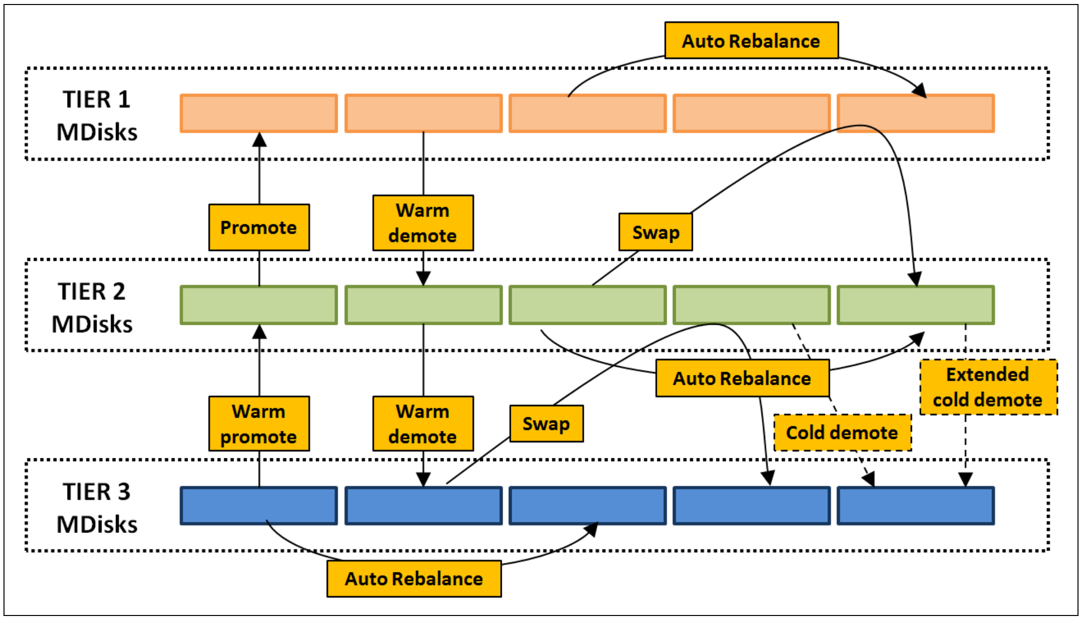


Figure 4-24 Easy Tier extent migration types

4.6.3 Easy Tier operating modes

Easy Tier includes the following main operating modes:

- Off
- On
- Automatic
- Measure

Easy Tier is a licensed feature on some FlashSystem 50x0 systems. If the license is not present and Easy Tier is set to Auto or On, the system runs in Measure mode.

Options: The Easy Tier function can be turned on or off at the storage pool level *and* at the volume level, except for non fully-allocated volumes in a DRP where Easy Tier is always enabled.

Easy Tier off mode

With Easy Tier turned off, statistics are not recorded, and cross-tier extent-migration does not occur.

Measure mode

Easy Tier can be run in an evaluation or measurement-only mode and collects usage statistics for each extent in a storage pool where the Easy Tier value is set to measure.

This collection is typically done for a single-tier pool, so that the benefits of adding additional performance tiers to the pool can be evaluated before any major hardware acquisition.

The heat and activity of each extent can be viewed in the GUI by clicking **Monitoring** → **Easy Tier Reports**. For more information, see 4.6.10, “Monitoring Easy Tier using the GUI” on page 191.

Automatic mode

In Automatic mode, the storage pool parameter `-easytier auto` must be set, and the volumes in the pool must have `-easytier` set to `on`.

The behavior of Easy Tier depends on the pool configuration. Consider the following points:

- If the pool only contains MDisks with a single tier type, the pool is in balancing mode.
- If the pool contains MDisks with more than one tier type, the pool runs automatic data placement and migration in addition to balancing within each tier.

Dynamic data movement is transparent to the host server and application users of the data, other than providing improved performance. Extents are automatically migrated, as explained in “Implementation rules” on page 184.

There might be situations where the Easy Tier setting is “auto” however the system is running in monitoring mode only. For example, with unsupported tier types or if you have not enabled the Easy Tier license. See Table 4-8 on page 180

The GUI provides the same reports as available in measuring mode and, in addition, provide the data movement report that shows the breakdown of the actual migration events triggered by Easy Tier. These migrations are reported in terms of the migration types, as described in “Internal processing” on page 172.

Easy Tier on mode

This mode forces Easy Tier to perform the tasks as in Automatic mode.

For example, when Easy Tier detects an unsupported set of tier types in a pool, as outlined in Table 4-8 on page 180, using On mode will force Easy Tier to the active state and it will perform to the best of its ability. The system raises an alert and there is an associated Directed Maintenance Procedure that guides you to fix the unsupported tier types.

> **Important:** Avoid creating a pool with more than three tiers. Although the system attempts to create generic hot, medium, and cold “buckets”, you might end up with Easy Tier that is running in measure mode only.
>
> These configurations are unsupported because they can cause a performance problem in the longer term; for example, disparate performance within a single tier.
>
> The ability to override the automatic mode is provided to enable temporary migration from an older set of tiers to new tiers and must be rectified as soon as possible.

Storage pool balancing

This feature assesses the extents that are written in a pool, and balances them automatically across all MDisks within the pool. This process works with Easy Tier when multiple classes of disks exist in a single pool. In this case, Easy Tier moves extents between the different tiers, and storage pool balancing moves extents within the same tier, to enable a balance in terms of workload across all MDisks that belong to a given tier.

Balancing is when you maintain equivalent latency across all MDisks in a given tier, this can result in different capacity usage across the MDisks. However, performance balancing is preferred over capacity-balancing in most cases.

The process automatically balances existing data when new MDisks are added into an existing pool, even if the pool only contains a single type of drive.

Balancing is automatically active on all storage pools, no matter the Easy Tier setting. For a single tier pool the Easy Tier state will report as balancing.

Note: Storage pool balancing can be used to balance extents when mixing different size disks of the same performance tier. For example, when adding larger capacity drives to a pool with smaller capacity drives of the same class, storage pool balancing redistributes the extents to take advantage of the additional performance of the new MDisks.

Easy Tier mode settings

The Easy Tier setting can be changed on a storage pool and volume level. Depending on the Easy Tier setting and the number of tiers in the storage pool, Easy Tier services might function in a different way. Table 4-6 lists possible combinations of Easy Tier setting.

Table 4-6 Easy Tier 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 or more	off	inactive (see note 2)
		on	inactive (see note 2)
Measure	One or More	off	measured (see note 3)
		on	measured (see note 3)
Auto	One	off	measured (see note 3)
		on	balanced (see note 4)
	Two - four	off	measured (see note 3)
		on	active (see note 5 & 6)
	Five	any	measured (see note 3)
On	One	off	measured (see note 3)
		on	balanced (see note 4)
	Two - four	off	measured (see note 3)
		on	active (see note 5)
	Five	off	measured (see note 3)
		on	active (see note 6)

Notes: The following notes apply to Table 4-6 on page 176:

1. If the volume copy is in image or sequential mode, or is being migrated, the volume copy Easy Tier status is measured rather than active.
2. When the volume copy status is inactive, no Easy Tier functions are enabled for that volume copy.
3. When the volume copy status is measured, the Easy Tier function collects usage statistics for the volume, but automatic data placement is not active.
4. When the volume copy status is balanced, the Easy Tier function enables performance-based pool balancing for that volume copy.
5. When the volume copy status is active, the Easy Tier function operates in automatic data placement mode for that volume.
6. When five tiers (or some four-tier) configurations are used and Easy Tier is in the **`On`** state, Easy Tier is forced to operate but might not behave exactly as expected. See Table 4-8 on page 180

The default Easy Tier setting for a storage pool is `Auto`, and the default Easy Tier setting for a volume copy is `On`. Therefore, Easy Tier functions, except pool performance balancing, are disabled for storage pools with a single tier. Automatic data placement mode is enabled by default for all striped volume copies in a storage pool with two or more tiers.

4.6.4 MDisk tier types

The three Easy Tier tier types (“hot”, “medium”, and “cold”) are generic “buckets” that Easy Tier uses to build a set of extents that belong to each tier. You must tell Easy Tier which MDisks belong to which bucket.

The type of disk and RAID geometry used by internal or external MDisks defines their expected performance characteristics. These characteristics are used to help define a *tier type* for each MDisk in the system.

Five tier types can be assigned. The tables in this section use the numbers from this list as a shorthand for the tier name:

1. `tier_scm` that represents Storage Class Memory MDisks
2. `tier0_flash` that represents enterprise flash technology, including FCM
3. `tier1_flash` that represents lower performing tier1 flash technology (lower DWPD)
4. `tier_enterprise` that represents enterprise HDD technology (both 10 K and 15 K RPM)
5. `tier_nearline` that represents nearline HDD technology (7.2 K RPM)

Consider the following points:.

- Easy Tier is designed to operate with up to 3 tiers of storage, “hot”, “medium”, “cold”
- An MDisk can only belong to one tier type.
- Today, five MDisk tier-types exist.
- Internal MDisks have their tier type set automatically.
- External MDisks default to the “enterprise” tier and might need to be changed by the user.
- The number of MDisk tier-types found in a pool will determine if the pool is a single-tier pool or a multi-tier pool.

Attention: As described in 4.6.5, “Changing the tier type of an MDisk” on page 181, IBM FlashSystem do not automatically detect the type of external MDisks. Instead, all external MDisks are initially put into the enterprise tier by default. The administrator must then manually change the MDisks tier and add them to storage pools.

Single-tier storage pools

Figure 4-25 shows a scenario in which a single storage pool is populated with MDisks that are presented by an external storage controller. In this solution, the striped volumes can be measured by Easy Tier, and can benefit from *storage pool balancing* mode, which moves extents between MDisks of the same type.

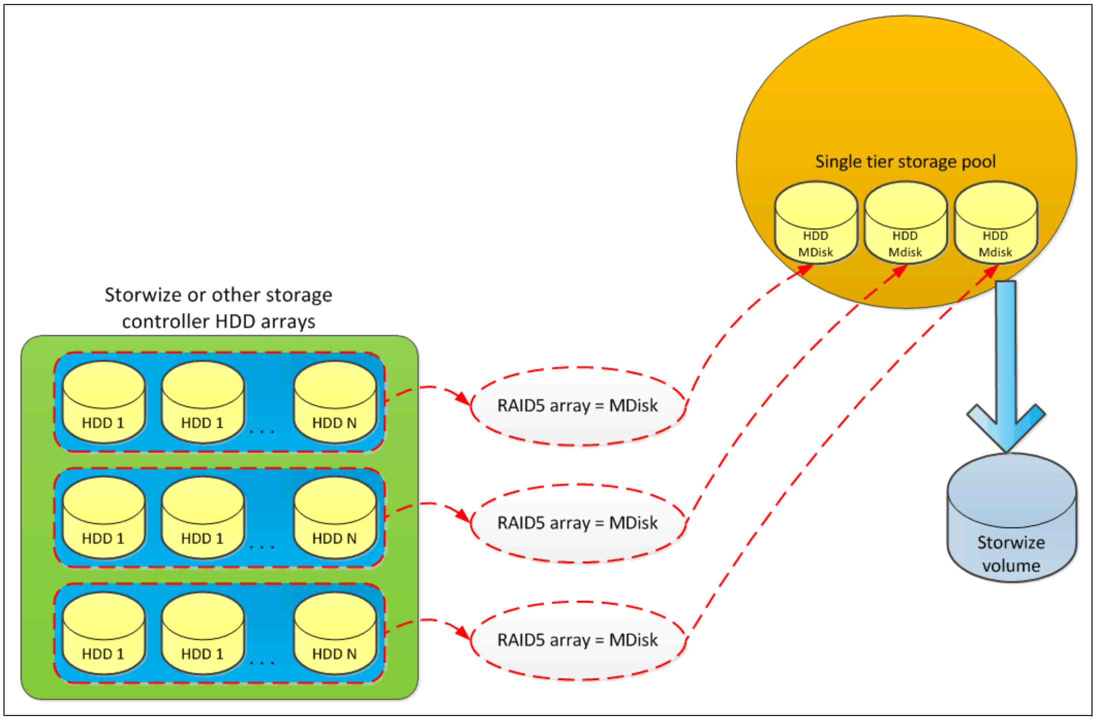


Figure 4-25 Single tier storage pool with striped volume

MDisks that are used in a single-tier storage pool should have the same hardware characteristics. These characteristics include the same RAID type, RAID array size, disk type, disk RPM, and controller performance characteristics.

For external MDisks, attempt to create all MDisks with the same RAID geometry (number of disks). If this is not possible, you can modify the Easy Tier load setting to manually balance the workload; however, care must be taken. For more information, see “MDisk Easy Tier load” on page 190.

For internal MDisks, the system can cope with different geometries as the number of drives will be reported to Easy Tier, which then uses the Overload Protection information to balance the workload appropriately. See 4.6.6, “Easy Tier overload protection” on page 183.

Multi-tier storage pools

A multi-tier storage pool has a mix of MDisks with more than one type of MDisk tier attribute. This pool can be, for example, a storage pool that contains a mix of enterprise and SSD MDisks or enterprise and NL-SAS MDisks.

Figure 4-26 shows a scenario in which a storage pool is populated with three different MDisk types:

- One belonging to an SSD array
- One belonging to an SAS HDD array
- One belonging to an NL-SAS HDD array)

Although Figure 4-26 shows RAID 5 arrays, other RAID types also can be used.

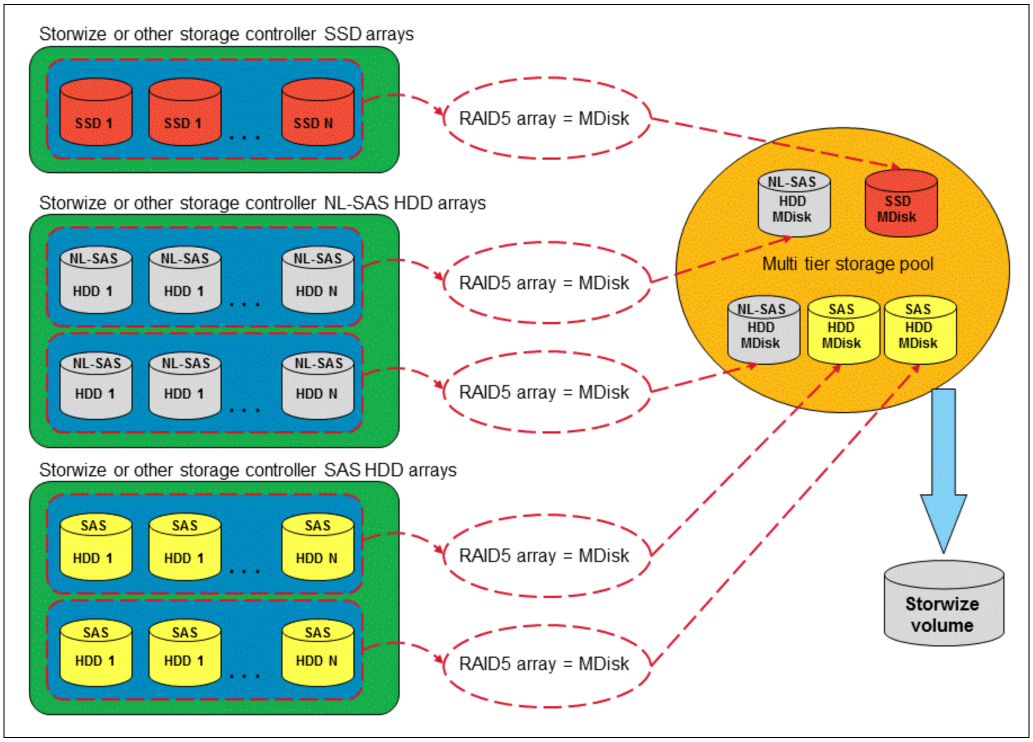


Figure 4-26 Multitier storage pool with striped volume

Note: If you add MDisks to a pool and they have (or you assign) more than three tier types, Easy Tier will try to group two or more of the tier types together into a single “bucket” and use them both as either the “middle” or “cold” tier. The groupings are described in table Table 4-8 on page 180

However, overload protection and pool balancing might result in a bias on the load being placed on those MDisks despite them being in the same “bucket”

Easy Tier mapping to MDisk tier types

The five MDisk tier-types are mapped to the three Easy Tier tiers depending on the pool configuration, as shown in Table 4-7.

Table 4-7 Recommended 3-tier Easy Tier mapping policy

Tier mix	1+2, 1+3, 1+4, 1+5	2+3, 2+4, 2+5	3+4, 3+5	4+5	1+2+3, 1+2+4, 1+2+5	1+3+4, 1+3+5	1+4+5, 2+4+5, 3+4+5	2+3+4, 2+3+5
Hot Tier	1	2			1	1	1 or 2 or 3	2
Middle Tier	2 or 3 or 4 or	3 or 4 or	3	4	2	3	4	3
Cold Tier	5	5	4 or 5	5	3 or 4 or 5	4 or 5	5	4 or 5

Four- and Five-Tier pools

In general, Easy Tier will try to place `tier_enterprise (4)`- and `tier1_flash (3)`-based tiers into the one bucket to reduce the number of tiers defined in a pool to 3. See Table 4-8.

Table 4-8 4 and 5 Tier mapping policy4 and 5 Tier mapping policy

Tier Mix	1+2+3+4, 1+2+3+5, 1+2+4+5	1+3+4+5, 2+3+4+5	1+2+3+4+5
Hot Tier	Not supported: measure only	1 or 2	Not supported: measure only
Middle Tier		3 & 4	
Cold Tier		5	

If you create a pool with all five tiers or one of the unsupported four-tier pools and Easy Tier is set to “auto” mode, Easy Tier enters “measure” mode and measures the statistics but does not move any extents. To return to a supported tier configuration, remove one or more MDisks.

Important: Avoid creating a pool with more than three tiers. Although the system attempts to create “buckets”, the result might be that Easy Tier runs in measure mode only.

Temporary unsupported 4 or 5 tier mapping

If you need to temporarily define four or five in a pool, and you end up with one of the unsupported configurations, you can force Easy Tier to migrate data by setting the Easy Tier mode to “on”.

Attention: Extreme caution should be deployed and a full understanding of the implications should be made before forcing Easy Tier to run in this mode.

This setting is provided to allow temporary migrations where it is unavoidable to create one of these unsupported configurations. The implications are that long-term use in this mode can cause performance issues due to the grouping of unlike MDisks within a single Easy Tier tier.

For these configurations, the following mapping in Table 4-9 will be used by Easy Tier.

Table 4-9 Unsupported temporary 4 and 5 Tier mapping policy

Tier mix	**1+2+3+4, 1+2+3+5**	**1+2+4+5**	**1+2+3+4+5**
Hot Tier	1	1	1
Middle Tier	2 & 3	2	2 & 3
Cold Tier	4 or 5	4 & 5	4 & 5
comment	See Note 1	See Note 2	See Note 1 & 2

Note: The following notes apply to Table 4-9:

- Note 1: In these configurations, Enterprise HDD and Nearline HDD are placed into the cold tier. These two drive types feature different latency characteristics and the difference can skew the metrics that are measured by Easy Tier for the cold tier.
- Note 2: In these configurations, Tier0 and Tier 1 flash devices are placed in the middle tier. The different drive writes per day does not make the most efficient use of the Tier0 flash.

4.6.5 Changing the tier type of an MDisk

By default, IBM FlashSystem adds external MDisks to a pool with the tier type "enterprise". This addition is made because it cannot determine the technology type of the MDisk without further information.

Attention: When adding external MDisks to a pool, be sure to validate the **`tier_type`** setting is correct. Incorrect **`tier_type`** settings can cause performance problems; for example, if you inadvertently create a multi-tier pool.

IBM FlashSystem internal MDisks should automatically be created with the correct **`tier_type`** as the IBM FlashSystem is aware of the drives that are used to create the RAID array and so can set the correct **`tier_type`** automatically.

The **`tier_type`** can be set when adding an MDisk to a pool, or subsequently change the tier of an MDisk by using the CLI, use the **`chmdisk`** command as in Example 4-11.

Example 4-11 Changing MDisk tier

```
IBM_2145:SVC_ESC:superuser>lsmdisk -delim " "
id name status mode mdisk_grp_id mdisk_grp_name capacity ctrl_LUN_#
controller_name UID tier encrypt site_id site_name distributed dedupe
1 mdisk1 online managed 1 POOL_V7K_SITEB 250.0GB 0000000000000001 V7K_SITEB_C2
6005076802880102c00000000000002000000000000000000000000000000000 tier_enterprise
no 2 SITE_B no no
2 mdisk2 online managed 1 POOL_V7K_SITEB 250.0GB 0000000000000002 V7K_SITEB_C2
6005076802880102c00000000000002100000000000000000000000000000000 tier_enterprise
no 2 SITE_B no no

IBM_2145:SVC_ESC:superuser>chmdisk -tier tier_nearline 1

IBM_2145:SVC_ESC:superuser>lsmdisk -delim " "
```

```
id name status mode mdisk_grp_id mdisk_grp_name capacity ctrl_LUN_#
controller_name UID tier encrypt site_id site_name distributed dedupe
1 mdisk1 online managed 1 POOL_V7K_SITEB 250.0GB 0000000000000001 V7K_SITEB_C2
6005076802880102c00000000000002000000000000000000000000000000000 tier_nearline no
2 SITE_B no no
2 mdisk2 online managed 1 POOL_V7K_SITEB 250.0GB 0000000000000002 V7K_SITEB_C2
6005076802880102c00000000000002100000000000000000000000000000000 tier_enterprise
no 2 SITE_B no no
```

It is also possible to change the MDisk tier from the GUI, but this applies only to external MDisks. To change the tier, complete the following steps:

1. Click **Pools** → **External Storage** and click the **Plus** sign (**+**) next to the controller that owns the MDisks for which you want to change the tier.
2. Right-click the wanted MDisk and select **Modify Tier** (see Figure 4-27).

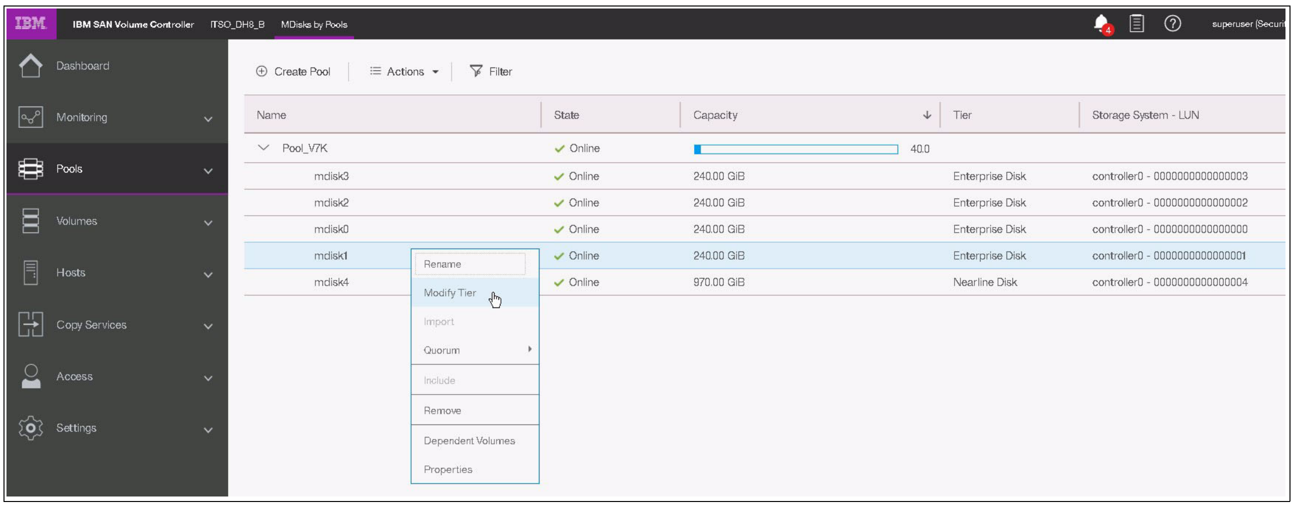


Figure 4-27 Change the MDisk tier

3. The new window opens with options to change the tier (see Figure 4-28).

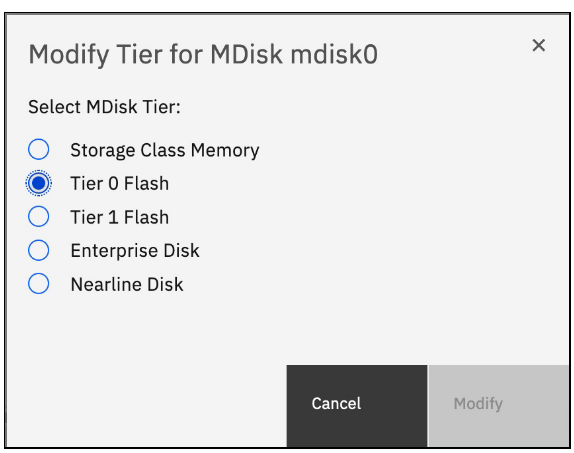


Figure 4-28 Select wanted MDisk tier

This change happens online and has no effect on hosts or availability of the volumes.

4. If you do not see the *Tier* column, right-click the blue title row and select the **Tier** check box, as shown in Figure 4-29.

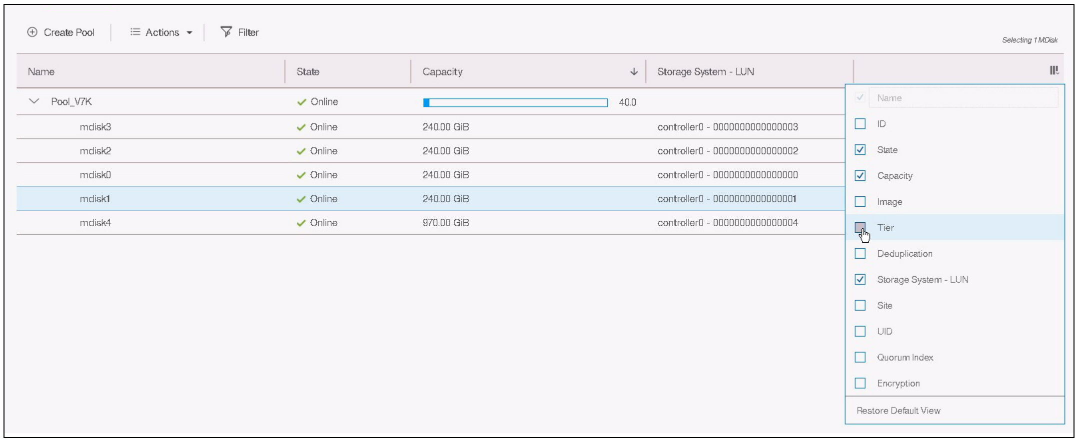


Figure 4-29 Customizing the title row to show the tier column

4.6.6 Easy Tier overload protection

Easy Tier is defined as a “greedy” algorithm. If overload protection is not used, Easy Tier attempts to use every extent on the hot tier. In some cases, this issue leads to overloading the hot tier MDisks and creates a performance problem.

Therefore, Easy Tier implements overload protection to ensure that it does not move too much workload onto the hot tier. If this protection is triggered, no other extents are moved onto that tier while the overload is detected. Extents can still be swapped; therefore, if one extent becomes colder and another hotter, they can be swapped.

To implement overload protection, Easy Tier must understand the capabilities of an MDisk. For internal MDisks, this understanding is handled automatically because the system can instruct Easy Tier as to the type of drive and RAID geometry (for example, 8+P+Q); therefore, the system can calculate the expected performance ceiling for any internal MDisk.

With external MDisks, the only measure or details we have is the storage controller type. Therefore, we know if the controller is an Enterprise, Midrange, or Entry level system and can make some assumptions about the load it can handle.

However, external MDisks cannot automatically have their MDisk tier type or “Easy Tier Load” defined. You must set the tier type manually and (if wanted), modify the load setting. For more information about Easy Tier load, see “MDisk Easy Tier load” on page 190.

Overload Protection is also used by the “warm promote” functionality. If Easy Tier detects a sudden change on a cold tier in which a workload is causing overloading of the cold tier MDisks, it can quickly react and recommend migration of the extents to the middle tier. This feature is useful when provisioning new volumes that overrun the capacity of the middle tier, or when no middle tier is present; for example, with Flash and Nearline only configurations.

4.6.7 Removing an MDisk from an Easy Tier pool

When you remove an MDisk from a pool that still includes defined volumes, and that pool is an Easy Tier pool, the extents that are still in use on the MDisk that you are removing are migrated to other free extents in the pool.

Easy Tier attempts to migrate the extents to another extent within the same tier. However, if there is not enough space in the same tier, Easy Tier picks the highest-priority tier with free capacity. Table 4-10 describes the migration target-tier priorities.

Table 4-10 Migration target tier priorities

Tier of MDisk being removed	**Target Tier Priority (pick highest with free capacity)**				
	1	**2**	**3**	**4**	**5**
tier_scm	tier_scm	tier0_flash	tier1_flash	tier_enterprise	tier_nearline
tier0_flash	tier0_flash	tier_scm	tier1_flash	tier_enterprise	tier_nearline
tier1_flash	tier1_flash	tier0_flash	tier_scm	tier_enterprise	tier_nearline
tier_enterprise	tier_enterprise	tier1_flash	tier_nearline	tier0_flash	tier_scm
tier_nearline	tier_nearline	tier_enterprise	tier1_flash	tier0_flash	tier_scm

The tiers are chosen to optimize for the typical migration cases, for example replacing the Enterprise HDD tier with Tier1 Flash arrays or replacing Nearline HDD with Tier1 Flash arrays.

4.6.8 Easy Tier implementation considerations

Easy Tier is part of the IBM Spectrum Virtualize code. For Easy Tier to migrate extents between different tier disks, storage that offers different tiers must be available (for example, a mix of Flash and HDD). With single tier (homogeneous) pools, Easy Tier uses storage pool balancing only.

> **Important:** Easy Tier uses the extent migration capabilities of IBM Spectrum Virtualize. These migrations require free capacity, as an extent is first cloned to a new extent, before the old extent is returned to the free capacity in the relevant tier.
>
> It is recommended that a minimum of 16 extents are needed for Easy Tier to operate. However, if only 16 extents are available, Easy Tier can move at most 16 extents at a time.
>
> Easy Tier and storage pool balancing will not function if you allocate 100% of the storage pool to volumes.

Implementation rules

Remember the following implementation and operational rules when you use Easy Tier:

- Easy Tier automatic data placement is not supported on image mode or sequential volumes. I/O monitoring for such volumes is supported, but you cannot migrate extents on these volumes unless you convert image or sequential volume copies to striped volumes.
- Automatic data placement and extent I/O activity monitors are supported on each copy of a mirrored volume. Easy Tier works with each copy independently of the other copy.

  > **Volume mirroring consideration:** Volume mirroring can have different workload characteristics on each copy of the data because reads are normally directed to the primary copy and writes occur to both copies. Therefore, the number of extents that Easy Tier migrates between the tiers might be different for each copy.

- If possible, the IBM FlashSystem system creates volumes or expands volumes by using extents from MDisks from the HDD tier. However, if necessary, it uses extents from MDisks from the SSD tier.
- Do not provision 100% of an Easy Tier enabled pool capacity. Reserve at least 16 extents for each tier for the Easy Tier movement operations.

When a volume is migrated out of a storage pool that is managed with Easy Tier, Easy Tier automatic data placement mode is no longer active on that volume. Automatic data placement is also turned off while a volume is being migrated, even when it is between pools that both have Easy Tier automatic data placement enabled. Automatic data placement for the volume is reenabled when the migration is complete.

Limitations

When you use Easy Tier on the IBM FlashSystem system, consider 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 MDisks in the same tier as the MDisk that is being removed, if possible. If insufficient extents exist in that tier, extents from another tier are used.

- Migrating extents

  When Easy Tier automatic data placement is enabled for a volume, you cannot use the **`migrateexts`** CLI command on that volume.

- Migrating a volume to another storage pool

  When IBM FlashSystem system migrates a volume to a new storage pool, Easy Tier automatic data-placement between the two tiers is temporarily suspended. After the volume is migrated to its new storage pool, Easy Tier automatic data placement between resumes for the moved volume, if appropriate.

  When the system migrates a volume from one storage pool to another, it attempts to migrate each extent to an extent in the new storage pool from the same tier as the original extent. In several cases, such as where a target tier is unavailable, another tier is used based on the same priority rules outlined in 4.6.7, “Removing an MDisk from an Easy Tier pool” on page 183.

- Migrating a volume to an image mode copy

  Easy Tier automatic data-placement does not support image mode. When a volume with active Easy Tier automatic data placement mode is migrated to an image mode volume, Easy Tier automatic data placement mode is no longer active on that volume.

- Image mode and sequential volumes cannot be candidates for automatic data placement. However, Easy Tier supports evaluation mode for image mode volumes.

Extent size considerations

The *extent size* determines the granularity level at which Easy Tier operates, which is the size of the chunk of data that Easy Tier moves across the tiers. By definition, a *hot extent* refers to an extent that has more I/O workload compared to other extents in the same pool and in the same tier.

It is unlikely that all the data that is contained in an extent features the same I/O workload, and as a result, the same temperature. Therefore, moving a hot extent likely also moves data that is not hot. The overall Easy Tier efficiency to put hot data in the correct tier is then inversely proportional to the extent size.

Consider the following points:

- Easy Tier efficiency is affecting the storage solution cost-benefit ratio. It is more effective for Easy Tier to place hot data in the top tier. In this case, less capacity can be provided for the relatively more expensive Easy Tier top tier.
- The extent size determines the bandwidth requirements for Easy Tier background process. The smaller the extent size, the lower that the bandwidth consumption.

However, Easy Tier efficiency is not the only factor considered when choosing the extent size. Manageability and capacity requirement considerations also must be taken into account.

As a general rule, use the default 1 GB (standard pool) or 4 GB (DRP) extent size for Easy Tier enabled configurations.

External controller tiering considerations

IBM Easy Tier is an algorithm that was developed by IBM Almaden Research and made available to many members of the IBM storage family, such as the DS8000, IBM SAN Volume Controller, and FlashSystem products. The DS8000 is the most advanced in Easy Tier implementation and currently provides features that are not yet available for IBM FlashSystem technology, such as Easy Tier Application, Easy Tier Heat Map Transfer, and Easy Tier Control.

In general, the use of Easy Tier at the highest level is recommended; that is, the virtualizer and any back-end controller tiering functions must be disabled.

> **Important:** Never run tiering at two levels. Doing so causes thrashing and unexpected heat and cold jumps at both levels.

Consider the following two options:

- Easy Tier is done at the virtualizer level

  In this case, complete the following steps at the backend level:

  i. Set up homogeneous pools according to the tier technology available.
  ii. Create volumes to present to the virtualizer from the homogeneous pool.
  iii. Disable tiering functions.

  At a virtualizer level, complete the following steps:

  i. Discover the MDisks provided by the backend storage and set the tier properly.
  ii. Create hybrid pools that aggregate the MDisks.
  iii. Enable the Easy Tier function.

- Easy Tier is done at the backend level

  In this case, complete these actions at the back-end level:

  i. Set up hybrid pools according to the tier technology available.
  ii. Create volumes to present to the virtualizer from the hybrid pools.
  iii. Enable the tiering functions.

  At virtualizer level, you need to complete the following actions:

  i. Discover the MDisks provided by the backend storage and set the same tier for all.
  ii. Create standard pools that aggregate the MDisks.
  iii. Disable the Easy Tier function.

Although both of these options provide benefits in term of performance, they have different characteristics.

Option 1 provides the following advantages compared to Option 2:

- With option 1, Easy Tier can be enabled or disabled at volume level. This feature allows users to decide which volumes benefit from Easy Tier and which do not.

  With option 2, this goal cannot be achieved.

- With option 1, the volume heat map matches directly to the host workload profile by using the volumes. This option also allows you to use Easy Tier across different storage controllers, which uses lower performance and cost systems to implement the middle or cold tiers.

  With option 2, the volume heat map on the backend storage is based on the IBM FlashSystem workload. Therefore, it does not represent the hosts workload profile because of the effects of the IBM FlashSystem caching.

- With option 1, the extent size can be changed to improve the overall Easy Tier efficiency (as described in “Monitoring Easy Tier using the GUI” on page 191).

Option 2, especially with DS8000 as the backend, offers some advantages when compared to option 1. For example, when external storage is used, the virtualizer uses generic performance profiles to evaluate the workload that can be placed on a specific MDisk, as described in “MDisk Easy Tier load” on page 190. These profiles might not match the back-end capabilities, which can lead to a resource usage that is not optimized.

However, this problem rarely occurs with option 2 because the performance profiles are based on the real back-end configuration.

Easy Tier and thin-provisioned backend considerations

When a data reduction-capable backend is used in Easy Tier-enabled pools, it is important to note that the data-reduction ratio on the physical backend might vary over time because of Easy Tier data-moving.

Easy Tier continuously moves extents across the tiers (and within the same tier) and attempts to optimize performance. As result, the amount of data that is written to the backend (and therefore the compression ratio) can unpredictably fluctuate over time, even though the data is not modified by the user.

> **Note:** It is not recommended to intermix data reduction capable and non-data reduction-capable storage in the same tier of a pool with Easy Tier enabled.

Easy Tier and Remote Copy considerations

When Easy Tier is enabled, the workloads that are monitored on the primary and secondary systems can differ. Easy Tier at the primary system sees a normal workload; it sees only the write workloads at the secondary system.

This situation means that the optimized extent distribution on the primary system can differ considerably from the one on the secondary system. The optimized extent reallocation that is based on the workload learning on the primary system is not sent to the secondary system at this time to allow the same extent optimization on both systems based on the primary workload pattern.

In a DR situation with a failover from the primary site to a secondary site, the extent distribution of the volumes on the secondary system is not optimized to match the primary workload. Easy Tier relearns the production I/O profile and builds a new extent migration plan on the secondary system to adapt to the new production workload.

It eventually achieves the same optimization and level of performance as on the primary system. This task takes a little time, so the production workload on the secondary system might not run at its optimum performance during that period. The Easy Tier acceleration feature can be used to mitigate this situation. For more information, see “Easy Tier acceleration” on page 190.

IBM FlashSystem Remote Copy configurations that use NearLine tier at the secondary system must be carefully planned, especially when practicing DR by using FlashCopy. In these scenarios, FlashCopy often is started just before the beginning of the DR test. It is likely that the FlashCopy target volumes are in the NearLine tier because of prolonged inactivity.

When the FlashCopy is started, an intensive workload often is added to the FlashCopy target volumes because of both the background and foreground I/Os. This situation can easily lead to overloading, and then possibly performance degradation of the NearLine storage tier if it is not correctly sized in terms of resources.

Easy Tier on DRP and interaction with garbage collection

DRPs use Log Structured Array (LSA) structures that need garbage collection activity to be done regularly. An LSA always appends new writes to the end of the allocated space. For more information, see “DRP internal details” on page 121.

Even if data exists and the write is an overwrite, the new data is not written in that place. Instead, the new write is appended at the end and the old data is marked as needing garbage collection. This process provides the following advantages:

- Writes to a DRP volume always are treated as sequential. Therefore, all the 8 KB chunks can be built into a larger 256 KB chunk and destage the writes from cache as full stripe writes or as large as a 256 KB sequential stream of smaller writes.
- Easy Tier with DRP gives the best performance both in terms of RAID on back-end systems and on Flash, where it becomes easier for the Flash device to perform its internal garbage collection on a larger boundary.

To improve the Easy Tier efficiency with this write workload profile, you can start to record metadata about how frequently certain areas of a volume are overwritten. The Easy Tier algorithm was modified so that we can then bin-sort the chunks into a heat map in terms of rewrite activity, and then, group commonly rewritten data onto a single extent. This method ensures that Easy Tier operates correctly for read data and write data when data reduction is in use.

Before DRP, write operations to compressed volumes held lower value to the Easy Tier algorithms because writes were always to a new extent; therefore, the previous heat was lost. Now, we can maintain the heat over time and ensure that frequently rewritten data is grouped. This process also aids the garbage collection process where it is likely that large contiguous areas are garbage collected together.

Tier sizing considerations

Tier sizing is a complex task that always requires an environment workload analysis to match the performance and costs expectations.

Consider the following sample configurations that address some or most common customer requirements. The same benefits can be achieved by adding Storage Class Memory (SCM) to the configuration. In these examples, the top Flash tier can be replaced with an SCM tier, or SCM can be added as the hot tier and the corresponding medium and cold tiers be shifted down to drop the coldest tier:

- 50% Flash, 50% Nearline

  This configuration provides a mix of storage for latency-sensitive and capacity-driven workloads.

- 10 - 20% Flash, 80 - 90% Enterprise

  This configuration provides Flash-like performance with reduced costs.

- 5% Tier 0 Flash, 15% Tier 1 Flash, 80% Nearline

  This configuration provides Flash-like performance with reduced costs.

- 3 - 5% Flash, 95 - 97% Enterprise

  This configuration provides improved performance compared to a single tier solution. All data is guaranteed to have at least enterprise performance. It also removes the requirement for over provisioning for high access density environments.

- 3 - 5% Flash, 25 - 50% Enterprise, 40 - 70% Nearline

  This configuration provides improved performance and density compared to a single tier solution. It also provides significant reduction in environmental costs.

- 20 - 50% Enterprise, 50 - 80% Nearline

  This configuration provides reduced costs and comparable performance to a single-tier Enterprise solution.

4.6.9 Easy Tier settings

The Easy Tier setting for storage pools and volumes can only be changed from the CLI. All the changes are done online without any effect on hosts or data availability.

Turning Easy Tier on and off

Use the **chvdisk** command to turn off or turn on Easy Tier on selected volumes. Use the **chmdiskgrp** command to change status of Easy Tier on selected storage pools, as shown in Example 4-12.

Example 4-12 Changing Easy Tier setting

```
IBM_FlashSystem:ITSO:superuser>chvdisk -easytier on test_vol_2
IBM_FlashSystem:ITSO:superuser>chmdiskgrp -easytier auto test_pool_1
```

Tuning Easy Tier

It is also possible to change more advanced parameters of Easy Tier. These parameters should be used with caution because changing the default values can affect system performance.

Easy Tier acceleration

The first setting is called *Easy Tier acceleration*. This is a system-wide setting, and is disabled by default. Turning on this setting makes Easy Tier move extents up to four times faster than when in default setting. In accelerate mode, Easy Tier can move up to 48 GiB every five minutes, while in normal mode it moves up to 12 GiB. Enabling Easy Tier acceleration is advised only during periods of low system activity. The following use cases for acceleration are the most likely:

- When installing a new system, accelerating Easy Tier will quickly reach a steady state and reduce the time needed to reach an optimal configuration. This applies to single-tier and multi-tier pools alike. In a single-tier pool this will allow balancing to spread the workload quickly and in a multi-tier pool it will allow both inter-tier movement and balancing within each tier.
- When adding capacity to the pool, accelerating Easy Tier can quickly spread existing volumes onto the new MDisks by way of pool balancing. It can also help if you added more capacity to stop warm demote operations. In this case, Easy Tier knows that certain extents are hot and were only demoted due to lack of space, or because Overload Protection was triggered.
- When migrating the volumes between the storage pools in cases where the target storage pool has more tiers than the source storage pool, accelerating Easy Tier can quickly promote or demote extents in the target pool.

This setting can be changed online, without any effect on host or data availability. To turn Easy Tier acceleration mode on or off, run the following command:

```
chsystem -easytieracceleration <on/off>
```

> **Important:** Do not leave accelerated mode on indefinitely. It is recommended to run in accelerated mode only for a few days to weeks to enable Easy Tier to reach a steady state quickly. After the system is performing fewer migration operations, disable accelerated mode to ensure Easy Tier does not affect system performance.

MDisk Easy Tier load

The second setting is called *MDisk Easy Tier load*. This setting is set on an individual MDisk basis, and indicates how much load Easy Tier can put on that particular MDisk. This setting was introduced to handle situations where Easy Tier is either underutilizing or overutilizing an external MDisk.

This setting cannot be changed for internal MDisks (array) because the system is able to determine the exact load that an internal MDisk can handle, based on the drive technology type, the number of drives, and type of RAID in use per MDisk.

For an external MDisk, Easy Tier uses specific performance profiles based on the characteristics of the external controller and on the tier assigned to the MDisk. These performance profiles are generic, which means that they do not take into account the actual backend configuration. For instance, the same performance profile is used for a DS8000 with 300 GB 15 K RPM and 1.8 TB 10 K RPM.

This feature is provided for advanced users to change the Easy Tier load setting to better align it with a specific external controller configuration.

> **Note:** The load setting is used with the MDisk tier type setting to calculate the number concurrent I/O and expected latency from the MDisk. Setting this value incorrectly, or using the wrong MDisk tier type, can have a detrimental effect on overall pool performance.

The following values can be set to each MDisk for the Easy Tier load:

- Default
- Low
- Medium
- High
- Very high

The system uses a default setting based on controller performance profile and the MDisk tier setting of the presented MDisks.

Change the default setting to any other value only when you are certain that a particular MDisk is underutilized and can handle more load, or that the MDisk is overutilized and the load should be lowered. Change this setting to `very high` only for SDD and Flash MDisks.

This setting can be changed online, without any effect on the hosts or data availability.

To change this setting, run the following command:

```
chmdisk -easytierload high mdisk0
```

Important: Consider the following points:

- When IBM SAN Volume Controller is used with FlashSystem backend storage, it is recommended to set the Easy Tier load to “very high” for FlashSystem MDisks other than FlashSystem 50x0 where the default is recommended.

  The same would be recommended for modern high-performance all-flash storage controllers from other vendors.
- After changing the load setting, make a note of the old and new settings and record the date and time of the change. Use Storage Insights to review the performance of the pool in the coming days to ensure that you have not inadvertently degraded the performance of the pool.

  You can also gradually increase the load setting and validate at each change that you are seeing an increase in throughput without a corresponding detrimental increase in latency (and vice versa if you are decreasing the load setting).

4.6.10 Monitoring Easy Tier using the GUI

Since software version 8.3.1, the GUI includes various reports and statistical analysis that can be used to understand which Easy Tier movement, activity, and skew is present in a storage pool. These panels replace the old IBM Storage Tier Advisor Tool (STAT) and STAT Charting Tool.

Unlike previous versions, where you were required to download the necessary log files from the system and upload to the STAT tool, from version 8.3.1 onwards, the system continually reports the Easy Tier information. Therefore, the GUI always displays the most up-to-date information.

Accessing Easy Tier reports

In the GUI, select **Monitoring** → **Easy Tier Reports** to show the Easy Tier Reports page.

If the system or Easy Tier has been running for less than 24 hours there might not be any data to display.

The reports page has three views, which can be accessed via the tabs at the top of the page:

- Data Movement
- Tier Composition
- Workload Skew

Data movement report

The data movement report shows the amount of data that has been moved in a given time period. You can change the time period using the drop-down selection on the right side (see Figure 4-30).

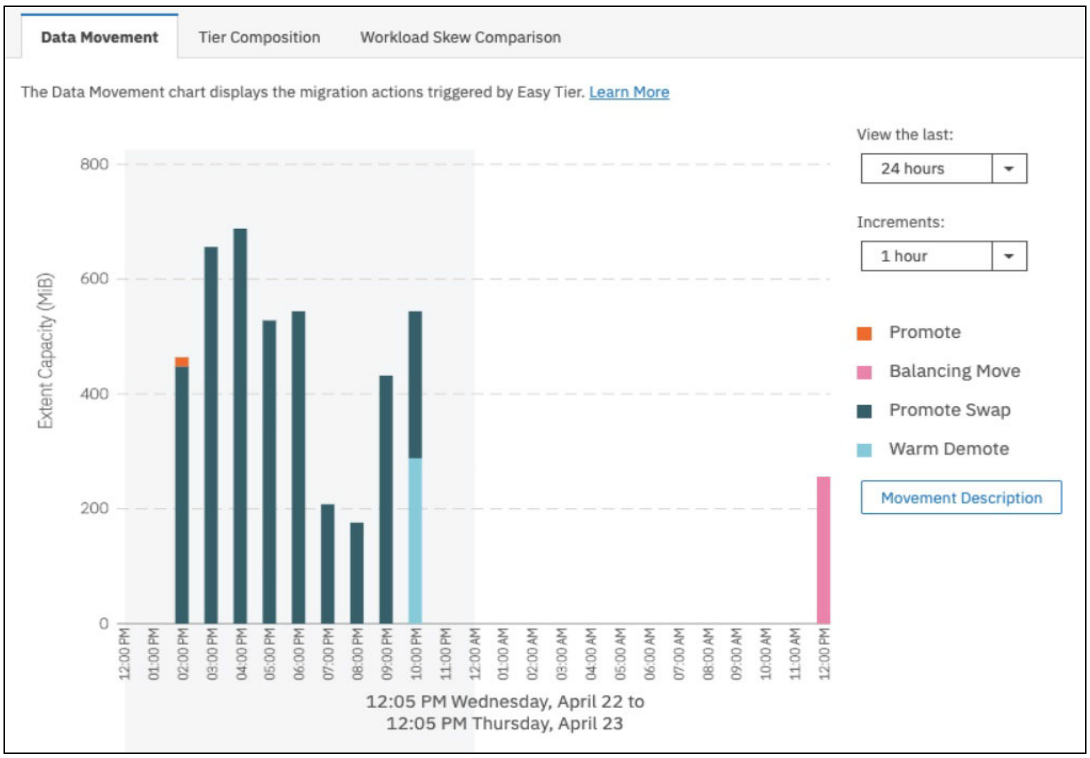


Figure 4-30 Easy Tier Data Movement page

The report breaks down the type of movement and these are described in terms of the internal Easy Tier extent movement types, as detailed in 4.6.2, “Easy Tier definitions” on page 171.

To aid your understanding and remind you of the definitions, click the **Movement Description** button to view the information panel, as shown in Figure 4-31.

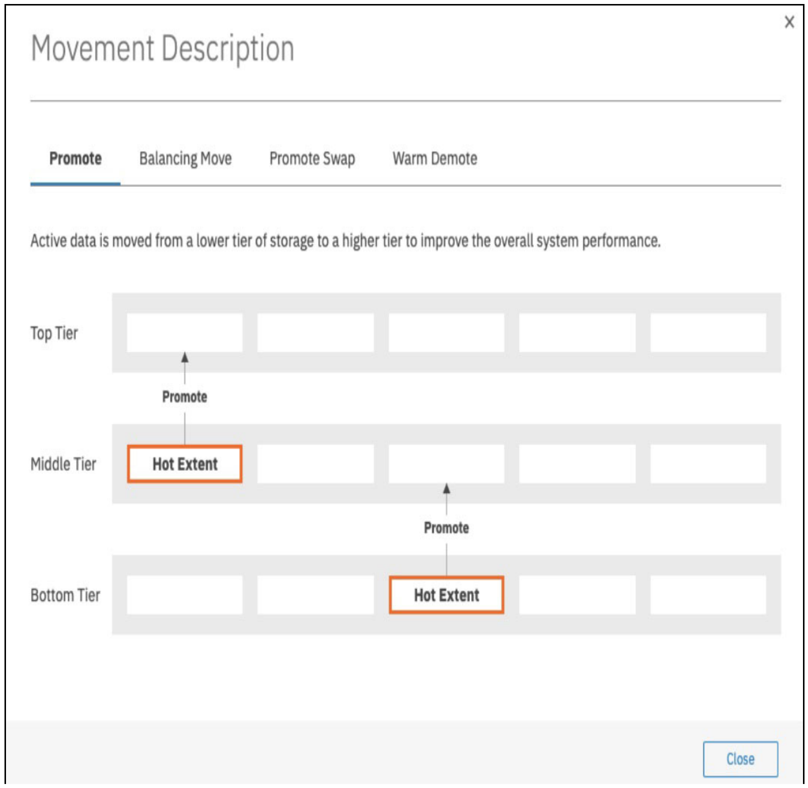


Figure 4-31 Easy Tier Movement description page

> **Important:** If you regularly see *warm demote* in the movement data, you should consider increasing the amount of hot tier that you have. A warm demote suggests that an extent is hot, but there is either not enough capacity or Overload Protection has been triggered in the hot tier.

Tier composition report

The tier composition window shows how much data in each tier is active versus inactive, as shown in Figure 4-32 on page 194. In an ideal case the majority of your active data should reside in the hot tier alone. In most cases the active data set will not be able to fit in only the hot tier. Therefore, you would expect to see active data in the middle tier also.

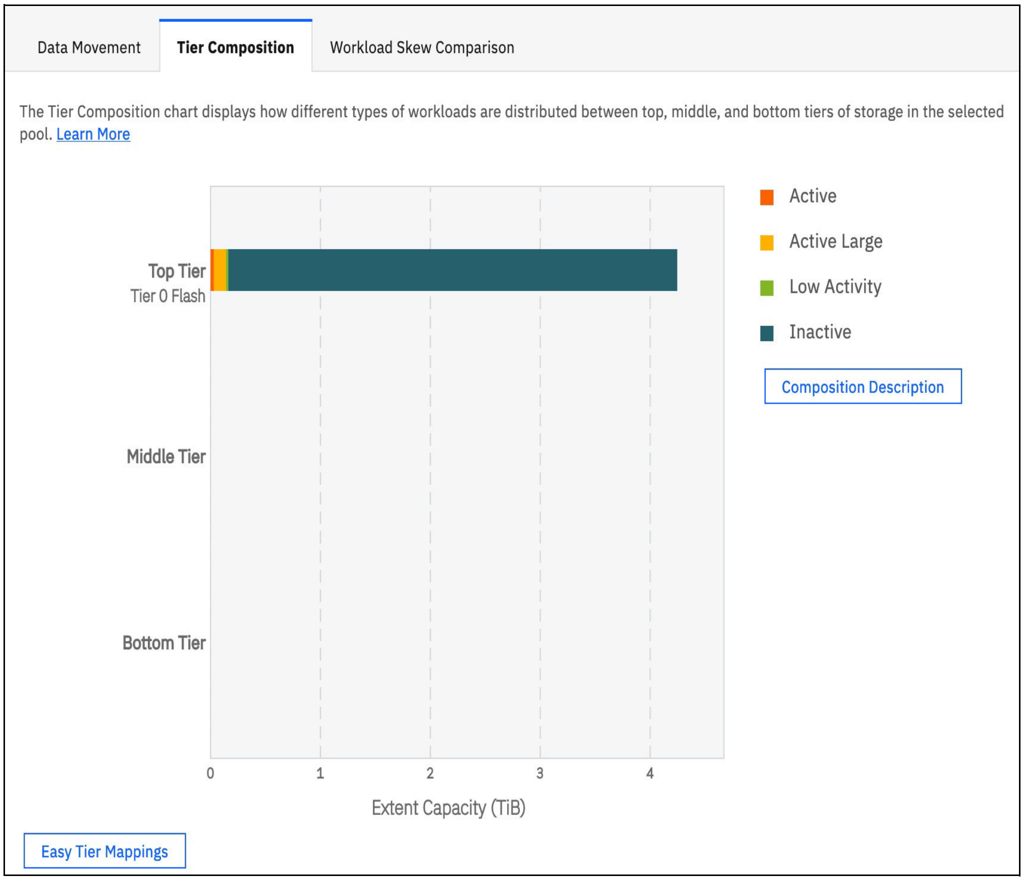


Figure 4-32 Easy Tier - single tier pool - composition report page

If all active data can fit in the hot tier alone, you see the best possible performance from the system. *Active large* is data that is active but is being accessed at block sizes larger than the 84 KiB for which Easy Tier is optimized. This data is still monitored and can contribute to “expanded cold demote” operations.

The presence of active data in the cold tier (regularly) suggests that you must increase the capacity or performance in the hot or middle tiers.

In the same way as with the movement page, you can click the **Composition Description** to view the information regarding each composition type (see Figure 4-33).

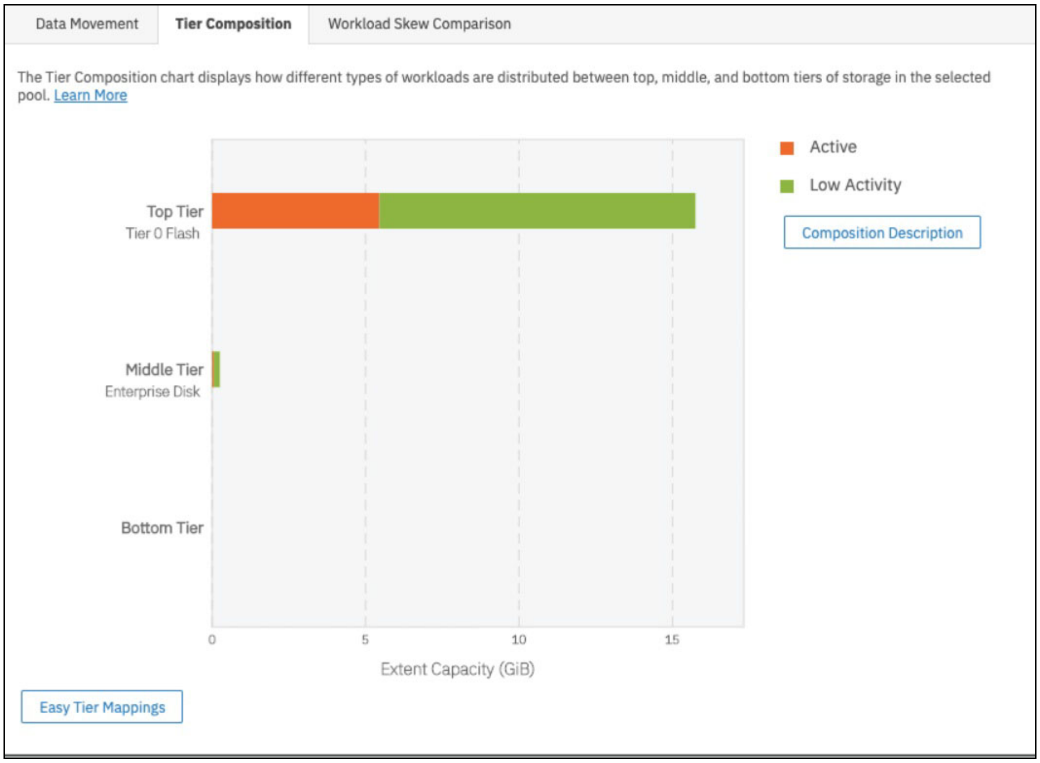


Figure 4-33 Easy tier - multi-tier pool - composition page

Workload skew comparison report

The workload skew comparison report plots the percentage of the workload against the percentage of capacity. The skew shows a good estimate for how much capacity is required in the top tier to have the most optimal configuration based on your workload.

> **Tip:** The skew can be viewed when the system is in measuring mode with a single-tier pool to help guide the recommended capacity to purchase that can be added to the pool in a hot tier.

A highly-skewed workload (the line on the graph rises sharply within the first few percentages of capacity) means that a smaller proportional capacity of hot tier is required. A low-skewed workload (the line on the graph rises slowly and covers a large percentage of the capacity) requires more hot-tier capacity, and consideration to a good performing middle tier when you cannot configure enough hot tier capacity (see Figure 4-34 on page 196).

Figure 4-34 Workload skew - single tier pool

In the first example, shown in Figure 4-34, you can clearly see that the workload is highly-skewed. This single-tier pool uses less than 5% of the capacity, but is performing 99% of the workload, both in terms of IOPS and MBps.

This result is a prime example of adding a small amount of faster storage to create a “hot” tier and improve overall pool performance, as shown in Figure 4-35.

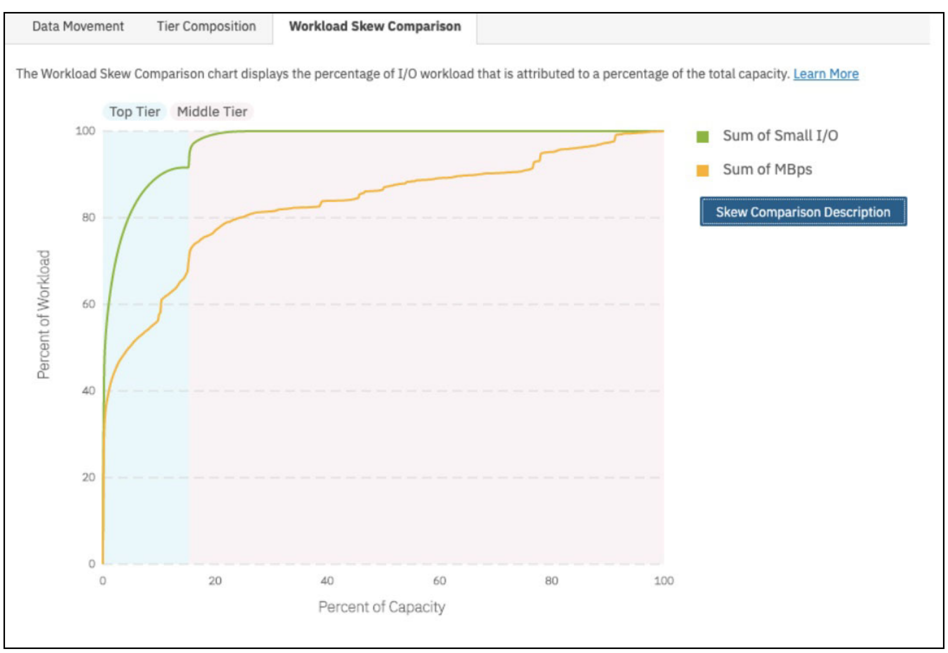


Figure 4-35 Workload skew - multi-tier configuration

In the second example, shown in Figure 4-35 on page 196, the system is already configured as a multi-tier pool and Easy Tier has been optimizing the data placement for some time. This workload is less skewed than in the first example with almost 20% of the capacity performing up to 99% of the workload.

Here again it might be worth considering increasing the amount of capacity in the top tier as about 10% of the IOP workload is coming from the middle tier and could be further optimized to reduce latency.

This graph that is shown in Figure 4-35 on page 196 also shows the split between IOPS and MBps. Although the middle tier is not doing much of the IOP workload it is providing a reasonably large proportion of the MBps workload.

In these cases, ensure that the middle tier can manage large-block throughput. A case might be made for further improving the performance by adding some higher-throughput devices as a new middle tier, and demoting the current middle tier to the cold tier. However, this change depends on the types of storage used to provide the existing tiers.

Any new configuration with three tiers would need to comply with the configuration rules regarding the different types of storage supported in three tier configurations as discussed in "Easy Tier mapping to MDisk tier types" on page 180.

If you implemented a new system and the majority of the workload is coming from a middle or cold tier, it might take a day or two for Easy Tier to complete the migrations after it has initially analyzed the system.

If after a few days, a distinct bias to the lower tiers still exists, you might want to consider enabling "Accelerated Mode" for a week or so. However, remember to turn it back off once the system reaches a steady state (see "Easy Tier acceleration" on page 190).

5

Volume types

In IBM FlashSystem, a *volume* is a logical disk that the system presents to attached hosts. This chapter describes the various types of volumes and it provides guidance for managing their properties.

This chapter includes the following topics:

5.1 Volumes overview

A volume can have one or two volume copies on the local storage system. A volume can also be replicated to a remote storage system. A *basic volume* has one local copy. A *mirrored volume* has two local copies. Each volume copy can be in different pools and have different capacity reduction attributes.

For best performance, spread host workload over multiple volumes.

Volumes can be created with the following attributes:

- Standard provisioned volumes

  Volumes with no special attributes, which also are referred to as *fully allocated volumes*.

- Thin-provisioned volumes

  Volumes that present a larger capacity to the host than their real capacity.

- Compressed volumes

  Volumes whose data is compressed.

- Deduplicated volumes

  Volumes whose data is deduplicated with other volumes in a data reduction pool (DRP).

- Mirrored volumes

  A volume can contain a duplicate copy of the data in another volume. Two copies are called a *mirrored volume*.

- HyperSwap volumes

  Volumes that participate in a HyperSwap relationship.

- VMware Virtual Volumes (VVols)

  Volumes that are managed remotely by VMware vCenter.

- Cloud volumes

  Volumes that are enabled for transparent cloud tiering.

Each volume in IBM FlashSystem also can feature the following attributes that affect where the extents are allocated;

- Striped

  A volume that is striped at the extent level. The extents are allocated from each MDisk that is in the storage pool. This volume type is the most frequently used because each I/O to the volume is spread across many external storage MDisks or internal disk drives compared to a sequential volume.

- Sequential

  A volume on which extents are allocated sequentially from one MDisk. This type of volume is rarely used because striped volume is better suited to most of the cases.

- Image

  Type of volume that has a direct relationship with one MDisk. The extents on the volume are directly mapped to the extents on the MDisk. This image is commonly used for data migration from a storage subsystem to an IBM FlashSystem.

5.2 Guidance for creating volumes

When creating volumes, consider the following guidelines:

- Consider the naming rules before you create volumes. It is easier to assign the correct name when the volume is created than to modify it later.
- Choose which type of volume that you want to create. First, decide whether fully allocated (standard volumes) or thin-provisioned. If you decide to create a thin-provisioned volume, analyze whether you need compression and deduplication enabled. Volume capacity reduction options are independent of any reduction done that is by the back-end controller.
- A fully allocated volume is automatically formatted, which can be a time-consuming process. However, this background process does not impede the immediate use of the volume. During the format, extents are over written with zeros and SCSI **unmap** commands are sent to the backend storage if supported.

  Actions, such as moving, expanding, shrinking, or adding a volume copy, are unavailable when the specified volume is formatting. Although it is unlikely that you must perform one of these actions after the volume is created, you can disable the format option in the Custom tab of the volume creation window by clearing the **Format volumes** option, as shown in Figure 5-1.

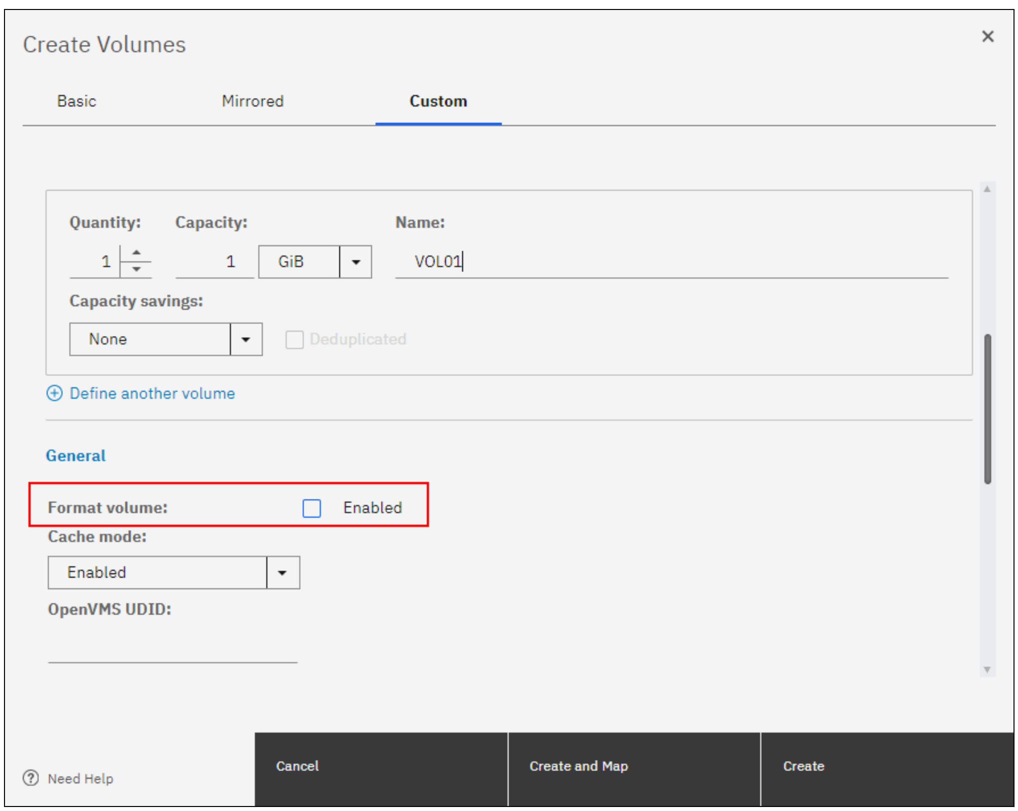


Figure 5-1 Volumes format option

You also can create volumes by using the CLI. Example 5-1 shows the command to disable the auto formatting option with the **-nofmtdisk** parameter.

Example 5-1 Volume creation without auto formatting option

```
superuser>mkvdisk -name VOL01 -mdiskgrp 0 -size 1 -unit gb -vtype striped -iogrp
io_grp0 -nofmtdisk
Virtual Disk, id [52], successfully created
superuser>lsvdisk VOL01
id 52
name VOL01
IO_group_id 0
IO_group_name io_grp0
status online
mdisk_grp_id 0
mdisk_grp_name Swimming
capacity 1.00GB
type striped
formatted no
formatting no
.
lines removed for brevity
```

When you create a volume, it takes some time to completely format it (depending on the volume size). The **syncrate** parameter of the volume specifies the volume copy synchronization rate and can be modified to accelerate the completion of the format process.

For example, the initialization of a 1 TB volume can take more than 120 hours to complete with the default syncrate value 50, or approximately 4 hours if you manually set the syncrate to 100. If you increase the syncrate to accelerate the volume initialization, remember to reduce it again to avoid issues the next time you use volume mirroring to perform a data migration of that volume.

For more information about creating a thin-provisioned volume, see 5.3, "Thin-provisioned volumes" on page 203.

- Each volume includes an I/O group and an associated preferred node. When creating a volume, consider balancing volumes across the I/O groups to balance the load across the cluster.

  In configurations where it is not possible to zone a host to multiple I/O groups so that the host can access only one I/O group, the volume must be created in the I/O group to which the host can access.

  Also, it is possible to define a list of I/O groups in which a volume can be accessible to hosts. It is recommended that a volume is accessible to hosts by the caching only I/O group. You can have more than one I/O group in the access list of a volume in some scenarios with specific requirements, such as when a volume is migrated to another I/O group.

  > **Tip:** Migrating volumes across I/O groups can be a disruptive action. Therefore, specify the correct I/O group at the time the volume is created.

- By default, the *preferred node*, which owns a volume within an I/O group, is selected in a load-balancing basis. Although it is not easy to estimate the workload when the volume is created, distribute the workload evenly on each node within an I/O group.

- Except in a few cases, the cache mode of a volume is set to read/write. For more information, see 5.12, “Volume cache mode” on page 236.
- A volume occupies an integer number of extents, but its length does not need to be an integer multiple of the extent size. Also, the length does need to be an integer multiple of the block size. Any space that is left over between the last logical block in the volume and the end of the last extent in the volume is unused.
- The maximum number of volumes per I/O group and system is listed in the “Configurations Limits and Restrictions” section for your system’s code level on the following IBM Support web pages:
  - Support Information for FlashSystem 9200
  - Support Information for FlashSystem 7200
  - Support Information for FlashSystem 5000

5.3 Thin-provisioned volumes

A thin-provisioned volume presents a different capacity to mapped hosts than the capacity that the volume uses in the storage pool. The system supports thin-provisioned volumes in standard pools and DRPs.

Note: We do not recommend the use of thin-provisioned volumes in a DRP with FlashCore Modules.

Figure 5-2 shows the basic concept of a thin-provisioned volume.

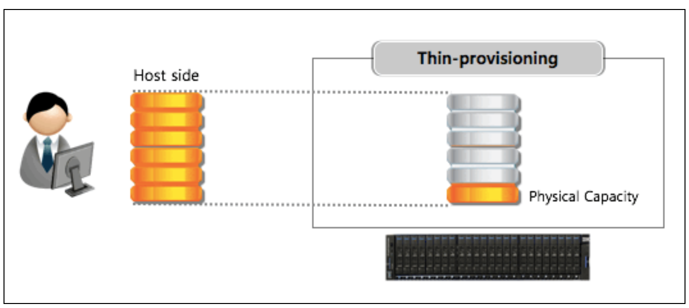


Figure 5-2 Thin-provisioned volume

The different types of volumes in a DRP are shown in Figure 5-3.

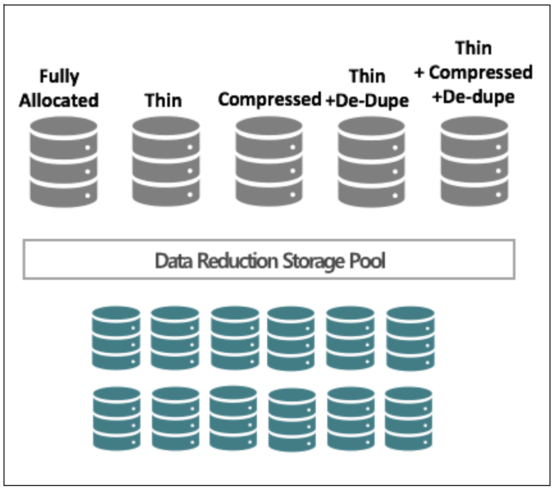


Figure 5-3 Different kinds of volumes in DRP

In standard pools, thin-provisioned volumes are created based on capacity savings criteria. These properties are managed at the volume level. However, in DRPs, all the benefits of thin-provisioning are available to all the volumes that are assigned to the pool. For the thin-provisioned volumes in DRPs, you can configure compression and data deduplication on these volumes, which increases the capacity savings for the entire pool.

You can enhance capacity efficiency for thin-provisioned volumes by monitoring the hosts' use of capacity. When the host indicates that the capacity is no longer needed, the space is released and can be reclaimed by the DRP. Standard pools do not have these functions.

Figure 5-4 shows the concepts of thin-provisioned volumes. These concepts are described next.

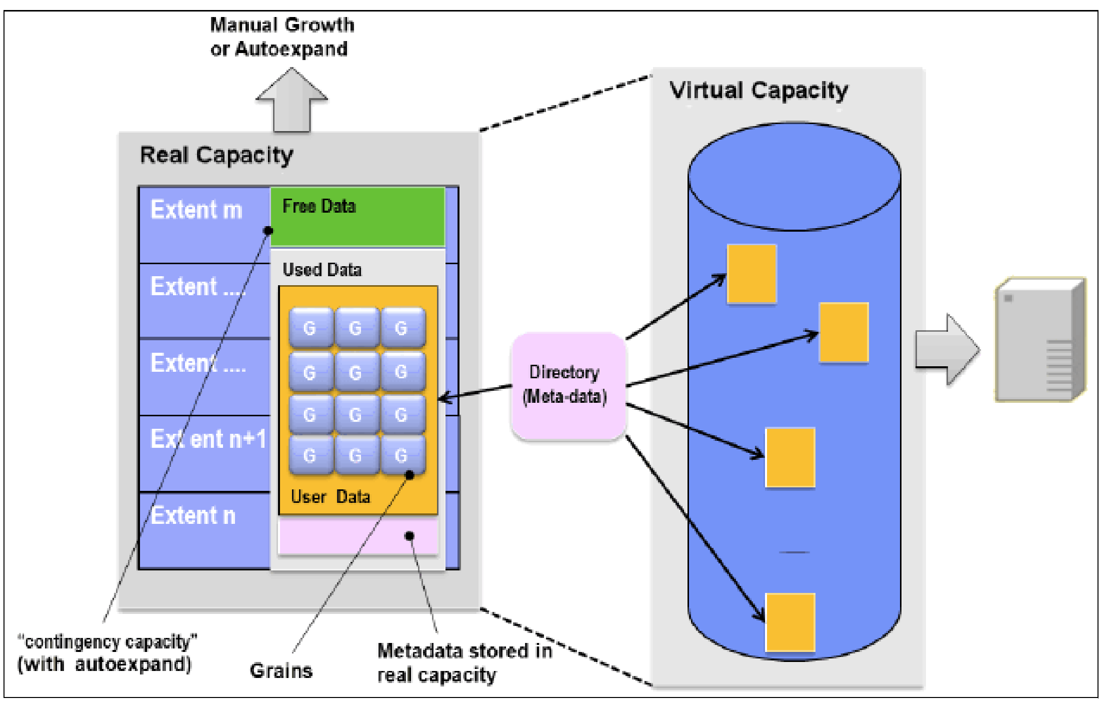


Figure 5-4 Thin-provisioned volume concepts

Real capacity defines how much disk space from a pool is allocated to a volume. *Virtual capacity* is the capacity of the volume that is reported to the hosts. A volume's virtual capacity is larger than its real capacity.

Each system uses the real capacity to store data that is written to the volume and metadata that describes the thin-provisioned configuration of the volume. As more information is written to the volume, more of the real capacity is used. The system identifies read operations to unwritten parts of the virtual capacity and returns zeros to the server without the use of any real capacity.

Thin-provisioned volumes in a standard pool are available in two operating modes: autoexpand and noautoexpand. You can switch the mode at any time. Thin-provisioned volumes in a DRP always have autoexpand enabled.

If you select the autoexpand feature, IBM FlashSystem automatically adds a fixed amount of 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. We recommend the use of autoexpand by default to avoid volume offline issues.

This amount extra real capacity is known as the *contingency capacity*. The contingency capacity is initially set to the real capacity that is assigned when the volume is created. If the user modifies the real capacity, the contingency capacity is reset to be the difference between the used capacity and real capacity.

A volume that is created *without* the **`autoexpand`** feature (and therefore has a zero contingency capacity) goes offline when the real capacity is used. In this case, it must be expanded.

When creating a thin-provisioned volume with compression and deduplication that is enabled, you must be careful of out-of-space issues in the volume and pool where the volume is created. Set the warning threshold in the pools that contain thin-provisioned volumes, and in the volume.

> **Warning threshold:** When you are working with thin-provisioned volumes, enable the warning threshold (by using email or an SNMP trap) in the storage pool. If the **`autoexpand`** feature is not used, you also must enable the warning threshold on the volume level. If the pool or volume runs out of space, the volume goes offline, which results in a loss of access.

If you do not want to be concerned with monitoring volume capacity, it is highly recommended that the **`autoexpand`** option is enabled. Also, when you create a thin-provisioned volume, you must specify the space that is initially allocated to it (**`-rsize`** option in the CLI) and the grain size.

By default, **`rsize`** (or real capacity) is set to 2% of the volume virtual capacity, and grain size is 256 KiB. These default values, with autoexpand enabled and warning disabled options, work in most scenarios. Some instances exist in which you might consider the use of different values to suit your environment.

Example 5-2 shows the command to create a volume with the suitable parameters.

Example 5-2 Creating thin-provisioned volume

```
superuser>mkvdisk -name VOL02 -mdiskgrp Pool1 -size 100 -unit gb -vtype striped
-iogrp io_grp0 -rsize 2% -autoexpand -warning 0 -grainsize 256
Virtual Disk, id [53], successfully created
superuser>lsvdisk VOL02
id 53
```

```
name VOL02
.
lines removed for brevity
.
capacity 100.00GB
.
lines removed for brevity
.
used_capacity 0.75MB
real_capacity 2.02GB
free_capacity 2.01GB
overallocation 4961
autoexpand on
warning 0
grainsize 256
se_copy yes
.
lines removed for brevity
```

A thin-provisioned volume can be converted nondisruptively to a fully allocated volume, or vice versa. Figure 5-5 shows how to modify capacity savings of a volume. You can right-click the volume and select **Modify Capacity Savings**.

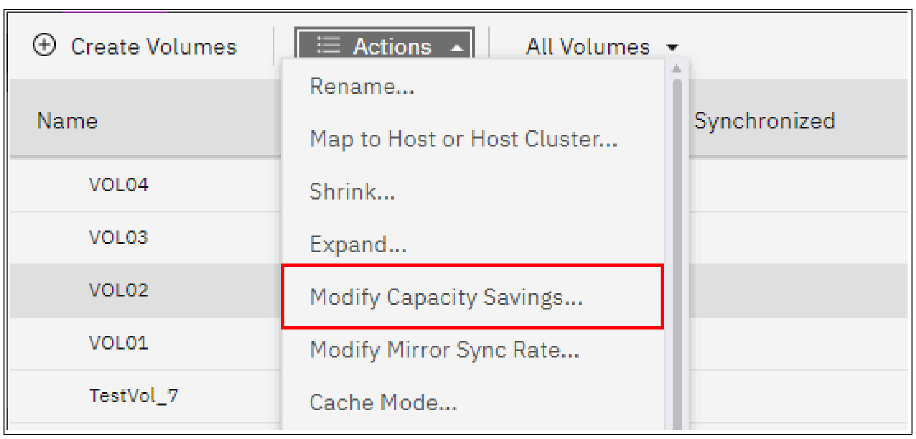


Figure 5-5 Modifying capacity savings of a volume nondisruptively

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.

5.3.1 Compressed volumes

When you create volumes, you can specify compression as a method to save capacity for the volume. With compressed volumes, data is compressed as it is written to disk, which saves more space. When data is read to hosts, the data is decompressed.

> **Note:** The volume compression attribute is independent of any compression that is performed by FlashCore Modules or any other compressing backend.

IBM FlashSystem support compressed volumes in a DRP only. DRPs also reclaims capacity that is not used by hosts if the host supports SCSI **unmap** commands. When these hosts issue SCSI **unmap** commands, a DRP reclaims the released capacity.

Compressed volumes in DRPs do not display their individual compression ratio. The pool used capacity *before* reduction indicates the total amount of data that is written to volume copies in the storage pool before data reduction occurs. The pool used capacity *after* reduction is the space that is used after thin provisioning, compression, and deduplication. This compression solution provides nondisruptive conversion between compressed and decompressed volumes.

If you are planning to virtualize volumes that are connected to your hosts directly from any storage subsystems, and you want an estimate of the space saving that likely is to be achieved, run the IBM Data Reduction Estimator Tool (DRET).

DRET tool is a command-line and host-based utility that can be used to estimate an expected compression rate for block devices. This tool also can evaluate capacity savings by using deduplication. For more information, see this IBM Support web page.

IBM FlashSystem also includes an integrated Comprestimator tool, which is available through the management GUI and command-line interface (CLI). If you are considering applying compression on noncompressed volumes in an IBM FlashSystem, you can use this tool to evaluate if compression generates enough capacity savings.

For more information, see 4.1.4, “Data reduction estimation tools” on page 125.

As shown in Figure 5-6, customize the Volume view to see the compression savings for a compressed volume, and estimated compression savings for a noncompressed volume that you are planning to migrate.

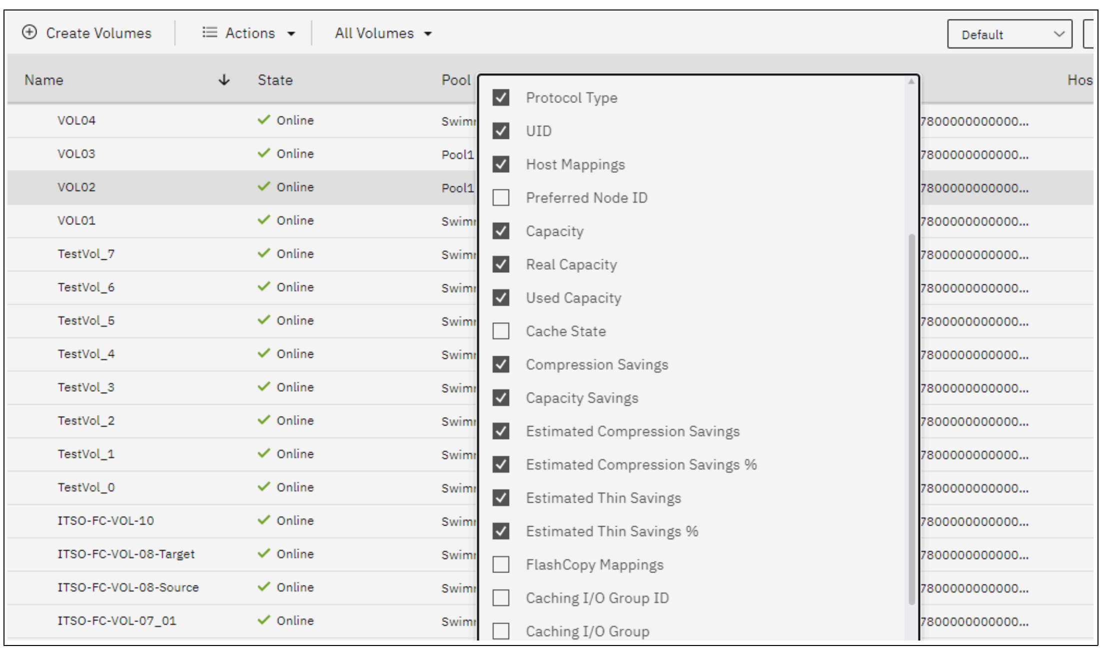


Figure 5-6 Customized view

5.3.2 Deduplicated volumes

Deduplication is a data reduction technique for eliminating duplicate copies of data. It can be configured with thin-provisioned and compressed volumes in a DRP for saving capacity.

The deduplication process identifies unique chunks of data (or byte patterns) and stores a signature of the chunk for reference when writing new data chunks. If the new chunk's signature matches an existing signature, the new chunk is replaced with a small reference that points to the stored chunk. The same byte pattern can occur many times, which results in the amount of data that must be stored being greatly reduced.

If a volume is configured with deduplication and compression, data is deduplicated first and then, compressed. Therefore, deduplication references are created on the compressed data that is stored on the physical domain.

The scope of deduplication is all deduplicated volumes in the same pool, regardless of the volume's preferred node or IO group.

Figure 5-7 shows the settings to create a compressed and deduplicated volume.

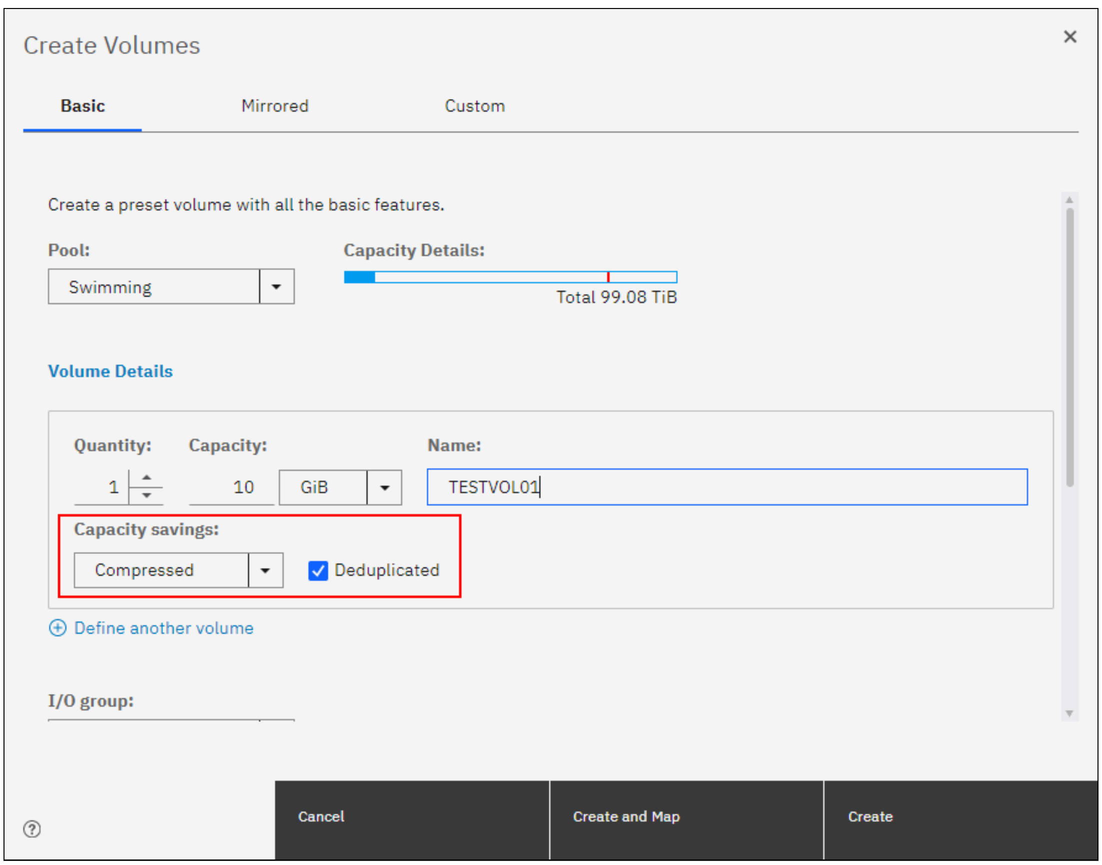


Figure 5-7 Creating deduplicated volumes

To create a thin-provisioned volume that uses deduplication, enter the command in the CLI that is shown in Example 5-3.

Example 5-3 Creating thin-provisioned volume with deduplication option

```
superuser>mkvolume -name dedup_test_01 -size 10 -unit gb -pool 0 -thin
-deduplicated
Volume, id [55], successfully created
```

To create a compressed volume that uses deduplication, enter the command that is shown in Example 5-4.

Example 5-4 Creating compressed volume with deduplication option

```
superuser>mkvolume -name dedup_test_02 -size 10 -unit gb -pool 0 -compressed
-deduplicated
Volume, id [56], successfully created
```

To maximize the space that is available for the deduplication database, the system distributes it between all nodes in the I/O groups that contain deduplicated volumes. Each node holds a distinct portion of the records that are stored in the database.

Depending on the data type stored on the volume, the capacity savings can be significant. Examples of use cases that typically benefit from deduplication are virtual environments with multiple virtual machines running the same operating system and backup servers.

In both cases, it is expected that multiple copies of identical files exist, such as components of the standard operating system or applications that are used in the organization. Conversely, data that is encrypted or compressed at the file system level does not benefit from deduplication because these operations were removed for redundancy.

If you want to evaluate if savings are realized by migrating a set of volumes to deduplicated volumes, you can use DRET. For more information about DRET, see 4.1.4, “Data reduction estimation tools” on page 125.

5.3.3 Thin provisioning considerations

Thin provisioning works best if the host limits writes to areas of the volume that store data. For example, if the host performs a low-level format of the entire volume, the host writes to the entire volume and no advantage is gained by using a thin provisioning volume over a fully allocated volume.

Consider the following properties of thin-provisioned volumes that are useful to understand for the rest of the section:

- When the used capacity first exceeds the volume *warning threshold*, an event is raised, which indicates that real capacity is required. The default warning threshold value is 80% of the volume capacity. To disable warnings, specify 0%.
- Compressed volumes include an attribute called *decompressed used capacity* (for standard pools) and *used capacity before reduction* (for a DRP). These volumes are the used capacities before compression or data reduction. They are used to calculate the compression ratio.

Thin provisioning and over-allocation

Because thin-provisioned volumes do not store the zero blocks, a storage pool is over-allocated only after the sum of all volume capacities exceeds the size of the storage pool.

Storage administrators likely think about the out-of-space problem. If enough capacity exists on disk to store fully allocated volumes, and you convert them to thin-provisioned volumes, enough space exists to store data (even if the servers writes to every byte of virtual capacity). Therefore, this issue is not going to be a problem for the short term, and you have time to monitor your system and understand how your capacity grows.

Monitoring capacity with thin-provisioned volumes

> **Note:** It is critical that capacity be monitored when thin-provisioned or compressed volumes are used. Be sure to add capacity *before* running out of space.

If you run out of space on a volume or storage pool, the host that uses the affected volumes cannot perform new write operations to these volumes. Therefore, an application or database that is running on this host becomes unavailable.

In a storage pool with only fully allocated volumes, the storage administrator can easily manage the used and available capacity in the storage pool as its used capacity grows when volumes are created or expanded.

However, in a pool with thin-provisioned volumes, the used capacity can increase at any time if the host file system grows. For this reason, the storage administrator must consider capacity planning carefully. It is critical to put in place volume and pool capacity monitoring.

Tools, such as IBM Spectrum Control and Storage Insights, can display the capacity of a storage pool in real time and graph how it is growing over time. These tools are important because they are used to predict when the pool will run out of space.

IBM FlashSystem also alerts you by including an event in the event log when the storage pool reaches the configured threshold, which is called the *warning level*. The GUI sets this threshold to 80% of the capacity of the storage pool by default.

By using enhanced Call Home and Storage Insights, IBM now has the ability to monitor and flag systems that have low capacity. This ability can result in a support ticket being generated and the client being contacted.

What to do if you run out of space in a storage pool

You can use one or a combination of the following options that are available if a storage pool runs out of space:

- Contingency capacity on thin-provisioned volumes

  If the storage pool runs out of space, each volume has its own contingency capacity, which is an amount of storage that is reserved by the volume and is sizable. Contingency capacity is defined by the *real capacity* parameter that is specified when the volume is created, which has a default value of 2%.

  The contingency capacity protects the volume from going offline when its storage pool runs out of space by having the storage pool use this reserved space first. Therefore, you have some time to repair things before everything starts going offline.

If you want more safety, you might implement a policy of creating volumes with 10% of *real capacity*. Also, remember that you do not need to have the same contingency capacity for every volume.

> **Note:** This protection likely solves most immediate problems. However, after you are informed that you ran out of space, a limited amount of time exists to react. You need a plan in place and well-understood about what to do next.

- Have unallocated storage on standby

  You can always have spare drives or managed disks ready to be added to whichever storage pool runs out of space within only a few minutes. This capacity gives you some breathing room while you take other actions. The more drives or MDisks you have, the more time you must solve the problem.

- Sacrificial emergency space volume

  Consider the use of a fully allocated sacrificial emergency space volume in each pool. If the storage pool is running out of space, you can delete or shrink this volume to quickly provide more available space in the pool.

- Move volumes

  You can migrate volumes to other pools to free up space. However, data migration on IBM FlashSystem is designed to move slowly to avoid performance problems. Therefore, it might be impossible to complete this migration before your applications go offline.

- Policy-based solutions

  No policy is going to solve the problem if you run out of space; however, you can use policies to reduce the likelihood of that ever happening to the point where you feel comfortable doing less of the other options.

  You can use these types of policies for thin provisioning:

  > **Note:** The following policies use arbitrary numbers. These numbers are designed to make the suggested policies more readable. We do not provide any recommended numbers to insert into these policies because they are determined by business risk, and this consideration is different for every client.

  – Manage free space such that enough free capacity always is available for your 10 largest volumes to reach 100% full without running out of free space.

  – Never over-allocate more than 200%. That is, if you have 100 TB of capacity in the storage pool, the sum of the volume capacities in the same pool must not exceed 200 TB.

  – Always start the process of adding capacity when the storage pool reaches 70% full.

Grain size

The *grain size* is defined when the thin-provisioned volume is created and can be set to 32 KB, 64 KB, 128 KB, or 256 KB (default). The grain size cannot be changed after the thin-provisioned volume is created.

Smaller granularities can save more space, but they have larger directories. If you select 32 KB for the grain size, the volume size cannot exceed 260,000 GB. Therefore, if you are not going to use the thin-provisioned volume as a FlashCopy source or target volume, use 256 KB by default to maximize performance.

Thin-provisioned volume copies in DRPs feature a grain size of 8 KB. This predefined value cannot be set or changed.

If you are planning to use thin provisioning with FlashCopy, remember that grain size for FlashCopy volumes can be only 64 KB or 256 KB. In addition, to achieve best performance, the grain size for the thin-provisioned volume and FlashCopy mapping must be same. For this reason, it is not recommended to use thin-provisioned volume in DRPS as a FlashCopy source or target volume.

Note: The use of thin-provisioned volumes in a DRP for FlashCopy is not recommended.

5.4 Mirrored volumes

By using volume mirroring, a volume can have two copies. Each volume copy can belong to a different pool and have different capacity reduction attributes. Each copy features the same virtual capacity as the volume. In the management GUI, an asterisk (*) indicates the primary copy of the mirrored volume. The primary copy indicates the preferred volume for read requests.

When a server writes to a mirrored volume, the system writes the data to both copies. When a server reads a mirrored volume, the system picks one of the copies to read. If one of the mirrored volume copies is temporarily unavailable (for example, because the storage system that provides the pool is unavailable), the volume remains accessible to servers. The system remembers which areas of the volume are written and resynchronizes these areas when both copies are available.

You can create a volume with one or two copies, and you can convert a non-mirrored volume into a mirrored volume by adding a copy. When a copy is added in this way, the system synchronizes the new copy so that it is the same as the existing volume. Servers can access the volume during this synchronization process.

You can convert a mirrored volume into a nonmirrored volume by deleting one copy or by splitting one copy to create a non-mirrored volume.

The volume copy can be any type: image, striped, or sequential. The volume copy can use thin-provisioning or compression to save capacity. If the copies are in DRPs, you can also use deduplication to the volume copies to increase the capacity savings.

If you are creating a volume, the two copies can use different capacity reduction attributes. However, to create a mirrored volume, both copies must be in a DRP. You can add a deduplicated volume copy in a DRP to a volume with a copy in a standard pool. You can use this method to migrate volume copies to data migration pools.

You can use mirrored volumes for the following reasons:

- Improving availability of volumes by protecting them from a single storage system failure.
- Providing concurrent maintenance of a storage system that does not natively support concurrent maintenance.

- Providing an alternative method of data migration with better availability characteristics. While a volume is migrated by using the data migration feature, it is vulnerable to failures on the source and target pool. Volume mirroring provides an alternative because you can start with a non-mirrored volume in the source pool, and then, add a copy to that volume in the destination pool.

  When the volume is synchronized, you can delete the original copy that is in the source pool. During the synchronization process, the volume remains available, even if a problem occurs with the destination pool.
- Converting fully allocated volumes to use data reduction technologies, such as thin-provisioning, compression, or deduplication.
- Converting compressed or thin-provisioned volumes in standard pools to DRPs to improve capacity savings.

When a volume mirror is synchronized, a mirrored copy can become unsynchronized if it goes offline and write I/O requests must be processed, or if a mirror fast failover occurs. The fast failover isolates the host systems from temporarily slow-performing mirrored copies, which affect the system with a short interruption to redundancy.

> **Note:** In standard-provisioned volumes, the primary volume formats before synchronizing to the volume copies. The **`-syncrate`** parameter on the **`mkvdisk`** command controls the format and synchronization speed.

You can create a mirrored volume by using the Mirrored option in the Create Volume window, as showing in Figure 5-8.

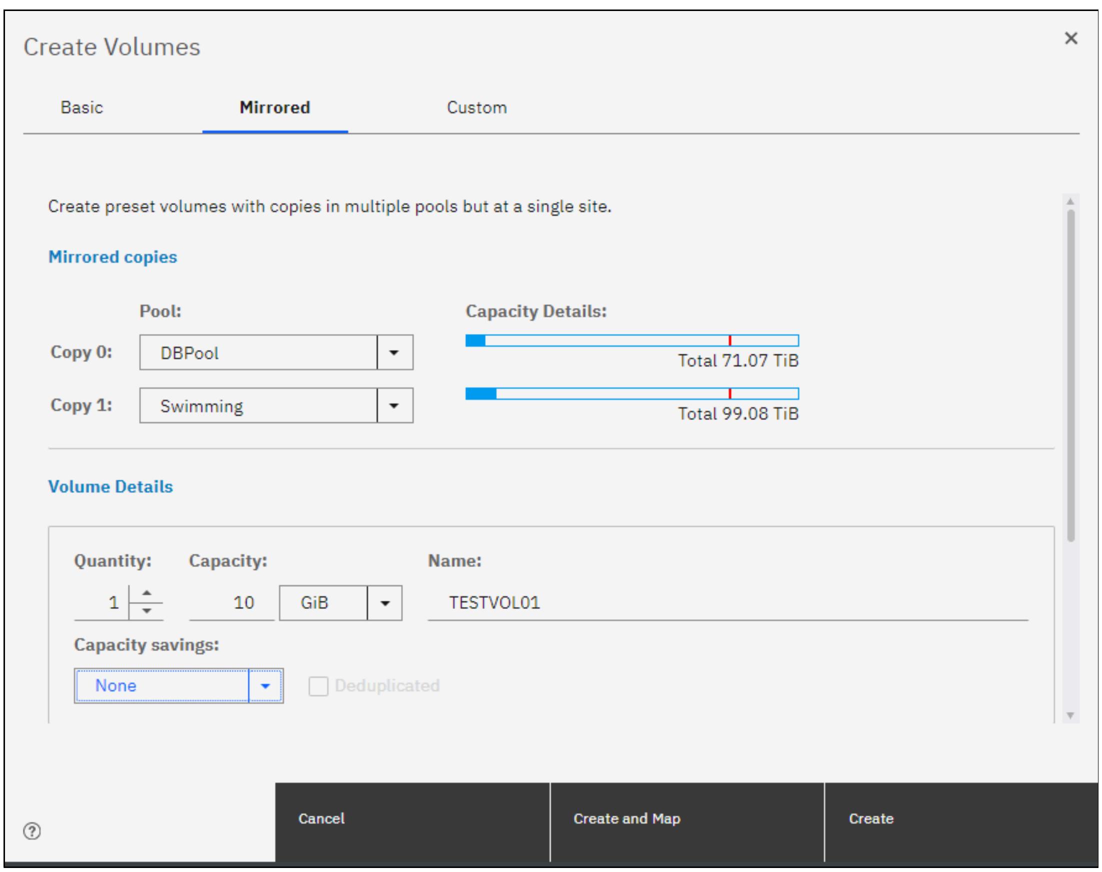


Figure 5-8 Mirrored volume creation

You can convert a non-mirrored volume into a mirrored volume by adding a copy, as shown in Figure 5-9.

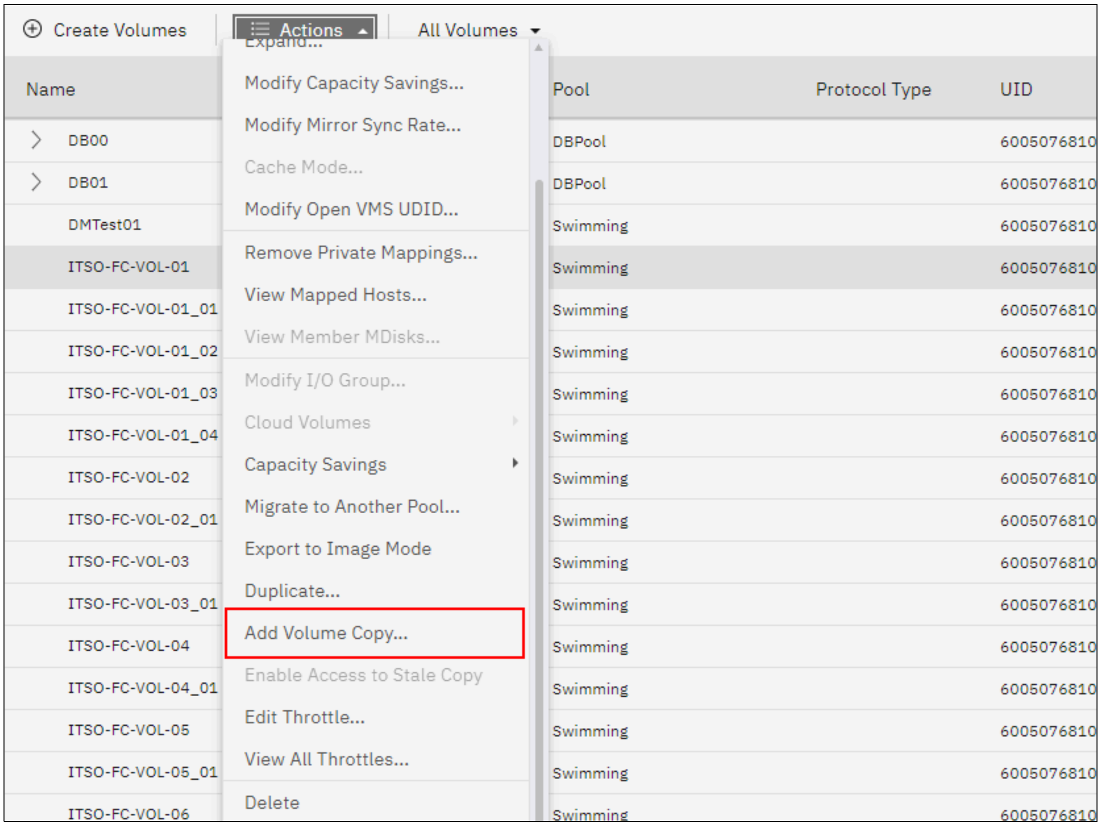


Figure 5-9 Adding a volume copy

5.4.1 Write fast failovers

With write fast failovers, the system submits writes to both copies during processing of host write I/O. If one write succeeds and the other write takes longer than 10 seconds, the slower request times-out and ends. The duration of the ending sequence for the slow copy I/O depends on the backend from which the mirror copy is configured. For example, if the I/O occurs over the Fibre Channel network, the I/O ending sequence typically completes in 10 - 20 seconds.

However, in rare cases, the sequence can take more than 20 seconds to complete. When the I/O ending sequence completes, the volume mirror configuration is updated to record that the slow copy is now no longer synchronized. When the configuration updates finish, the write I/O can be completed on the host system.

The volume mirror stops by using the slow copy for 4 - 6 minutes; subsequent I/O requests are satisfied by the remaining synchronized copy. During this time, synchronization is suspended. Also, the volume's synchronization progress shows less than 100% and decreases if the volume receives more host writes. After the copy suspension completes, volume mirroring synchronization resumes and the slow copy starts synchronizing.

If another I/O request times out on the unsynchronized copy during the synchronization, volume mirroring again stops by using that copy for 4 - 6 minutes. If a copy is always slow, volume mirroring attempts to synchronize the copy again every 4 - 6 minutes and another I/O timeout occurs.

The copy is not used for another 4 - 6 minutes and becomes progressively unsynchronized. Synchronization progress gradually decreases as more regions of the volume are written.

If write fast failovers occur regularly, an underlying performance problem can exist within the storage system that is processing I/O data for the mirrored copy that became unsynchronized. If one copy is slow because of storage system performance, multiple copies on different volumes are affected. The copies might be configured from the storage pool that is associated with one or more storage systems. This situation indicates possible overloading or other back-end performance problems.

When you run the **mkvdisk** command to create a volume, the **mirror_write_priority** parameter is set to latency by default. Fast failover is enabled. However, fast failover can be controlled by changing the value of the **mirror_write_priority** parameter on the **chvdisk** command. If the **mirror_write_priority** is set to **redundancy**, fast failover is disabled.

The system applies a full SCSI initiator-layer error recovery procedure (ERP) for all mirrored write I/O. If one copy is slow, the ERP can take up to 5 minutes. If the write operation is still unsuccessful, the copy is taken offline. Carefully consider whether maintaining redundancy or fast failover and host response time (at the expense of a temporary loss of redundancy) is more important.

Note: 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. To protect against mirrored volumes being taken offline, follow the guidelines for setting up quorum disks.

5.4.2 Read fast failovers

Read fast failovers affect how the system processes read I/O requests. A read fast failover determines which copy of a volume the system tries first for a read operation. The primary-for-read copy is the copy that the system tries first for read I/O.

The system submits a host read I/O request to one copy of a volume at a time. If that request succeeds, the system returns the data. If it is not successful, the system retries the request to the other copy volume.

With read fast failovers, when the primary-for-read copy goes slow for read I/O, the system fails over to the other copy. Therefore, the system tries the other copy first for read I/O during the following 4 - 6 minutes. After that attempt, the system reverts to read the original primary-for-read copy.

During this period, if read I/O to the other copy also is slow, the system reverts immediately. Also, if the primary-for-read copy changes, the system reverts to try the new primary-for-read copy. This issue can occur when the system topology changes or when the primary or local copy changes. For example, in a standard topology, the system normally tries to read the primary copy first. If you change the volume’s primary copy during a read fast failover period, the system reverts to read the newly set primary copy immediately.

The read fast failover function is always enabled on the system. During this process, the system does not suspend the volumes or make the copies out of sync.

5.4.3 Maintaining data integrity of mirrored volumes

Volume mirroring improves data availability by allowing hosts to continue I/O to a volume, even if one of the back-end storage systems fails. However, this mirroring does not affect data integrity. If either of the back-end storage systems corrupts the data, the host is at risk of reading that corrupted data in the same way as for any other volume.

Therefore, before you perform maintenance on a storage system that might affect the data integrity of one copy, it is important to check that both volume copies are synchronized. Then, remove that volume copy before you begin the maintenance.

5.5 HyperSwap volumes

HyperSwap volumes create copies on two separate sites for systems that are configured with HyperSwap topology. Data that is written to a HyperSwap volume is automatically sent to both copies so that either site can provide access to the volume if the other site becomes unavailable.

HyperSwap is a system topology that enables Disaster Recovery (DR) and high availability between I/O groups at different locations. Before you configure HyperSwap volumes, the system topology must be configured for HyperSwap and sites must be defined. Figure 5-10 shows an overall view of IBM FlashSystem HyperSwap that is configured with two sites.

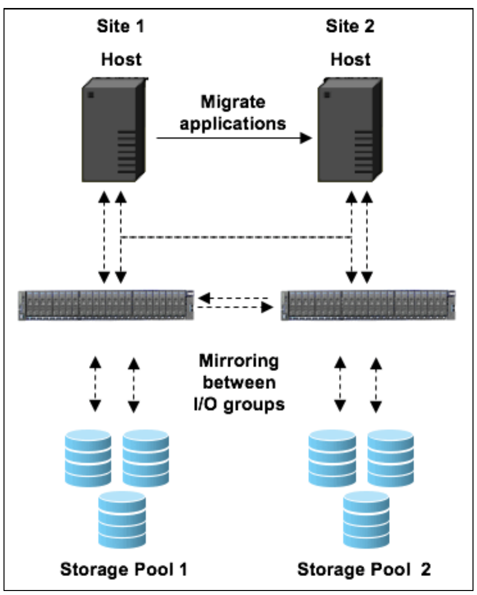


Figure 5-10 Overall HyperSwap diagram

In the management GUI, HyperSwap volumes are configured by specifying volume details, such as quantity, capacity, name, and the method for saving capacity. As with basic volumes, you can choose compression or thin-provisioning to save capacity on volumes.

For thin-provisioning or compression, you can also select to use deduplication for the volume that you create. For example, you can create a compressed volume that also uses deduplication to remove duplicated data.

The method for capacity savings applies to all HyperSwap volumes and copies that are created. The volume location displays the site where copies are located, based on the configured sites for the HyperSwap system topology. For each site, specify a pool and I/O group that are used by the volume copies that are created on each site. If you select to deduplicate volume data, the volume copies must be in DRPs on both sites.

The management GUI creates an HyperSwap relationship and change volumes automatically. HyperSwap relationships manage the synchronous replication of data between HyperSwap volume copies at the two sites.

If your HyperSwap system supports self-encrypting drives and the base volume is fully allocated in a DRP, the corresponding change volume is created with compression enabled. If the base volume is in a standard pool, the change volume is created as a thin-provisioned volume.

You can specify a consistency group that contains multiple active-active relationships to simplify management of replication and provide consistency across multiple volumes. A consistency group is commonly used when an application spans multiple volumes. Change volumes maintain a consistent copy of data during resynchronization. Change volumes allow an older copy to be used for DR if a failure occurred on the up-to-date copy before resynchronization completes.

You can also use the **mkvolume** command line to create a HyperSwap volume. The command also defines pools and sites for HyperSwap volume copies and creates the active-active relationship and change volumes automatically. If your HyperSwap system supports self-encrypting drives and the base volume is fully allocated in a DRP, the corresponding change volume is created with compression enabled. If the base volume is in a standard pool, the change volume is created as a thin-provisioned volume.

You can see the relationship between the Master and Auxiliary volume in a 2-site HyperSwap topology in Figure 5-11.

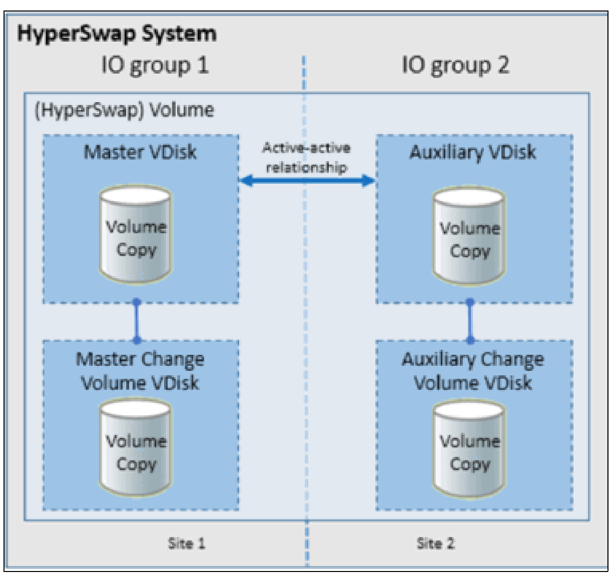


Figure 5-11 Master and Auxiliary volumes

For more information about HyperSwap volumes, see 7.3, “HyperSwap Volumes” on page 364.

5.6 VMware virtual volumes

The IBM FlashSystem supports VMware vSphere Virtual Volumes, sometimes referred to as *VVols*, which allows VMware vCenter to automate the management of system objects, such as volumes and pools.

You can assign ownership of Virtual Volumes to IBM Spectrum Connect by creating a user with the VASA Provider security role. IBM Spectrum Connect provides communication between the VMware vSphere infrastructure and the system.

Although you can complete specific actions on volumes and pools that are owned by the VASA Provider security role, IBM Spectrum Connect retains management responsibility for Virtual Volumes.

When virtual volumes are enabled on the system, a utility volume is created to store metadata for the VMware vCenter applications. You can select a pool to provide capacity for the utility volume. With each new volume that is created by the VASA provider, VMware vCenter defines a few kilobytes of metadata that are stored on the utility volume.

The utility volume can be mirrored to a second storage pool to ensure that the failure of a storage pool does not result in loss of access to the metadata. Utility volumes are exclusively used by the VASA provider and cannot be deleted or mapped to other host objects.

Note: The utility volume cannot be created in a DRP.

Figure 5-12 provides a high-level overview of the key components that enable the VVols management framework.

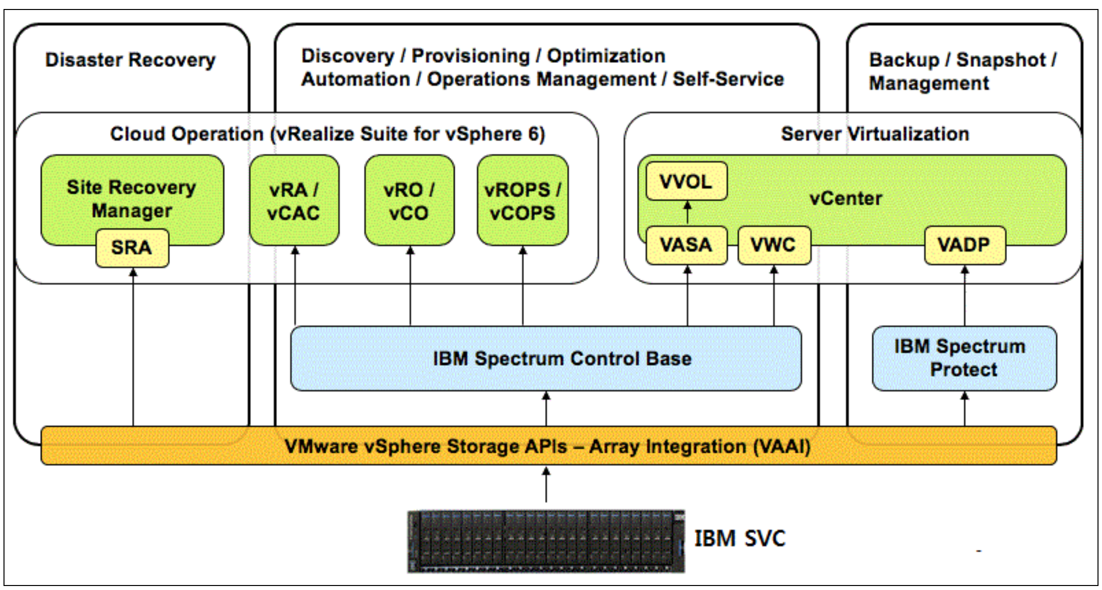


Figure 5-12 Overview of the key components of VMware environment

You can also use data copy through VMware vSphere Storage APIs Array Integration (VAAI), as shown in Figure 5-13.

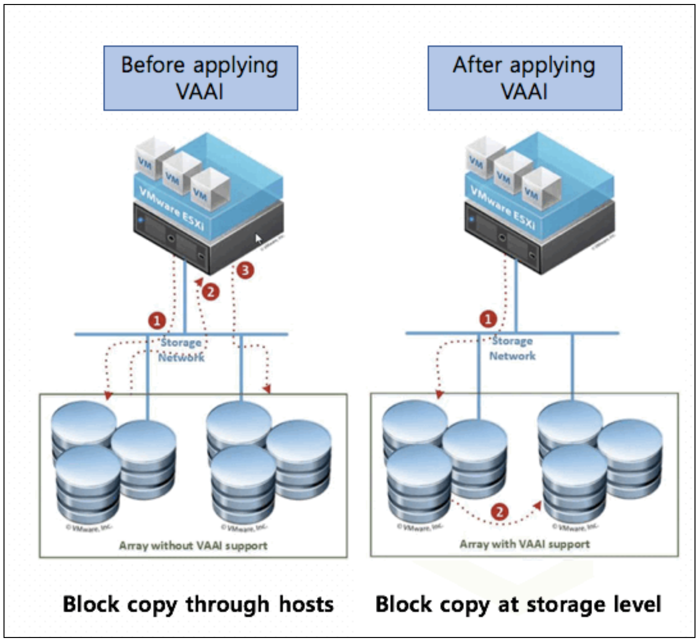


Figure 5-13 VMware vSphere Storage APIs Array Integration (VAAI)

The following are prerequisites must be met before configuring Virtual Volumes:

- An IBM Spectrum Connect must be set up.
- VMware vSphere ESXi hosts and vCenter running version 6.0 or later.
- The Network Time Protocol (NTP) server is configured on IBM FlashSystem and IBM Spectrum Connect.

To start using Virtual Volumes, complete the following steps on the IBM FlashSystem before you configure any settings within the IBM Spectrum Connect server:

1. Enable Virtual Volumes on the IBM FlashSystem:
   a. In the management GUI, click **Settings** → **System** → **VVOL** and select **On**.
   b. Select the pool to where the utility volume is stored. If possible, store a mirrored copy of the utility volume in a second storage pool that is in a separate failure domain. The utility volume cannot be created in a DRP.
   c. Create a user for IBM Spectrum Connect to communicate with the IBM FlashSystem, as shown in Figure 5-14 on page 220.

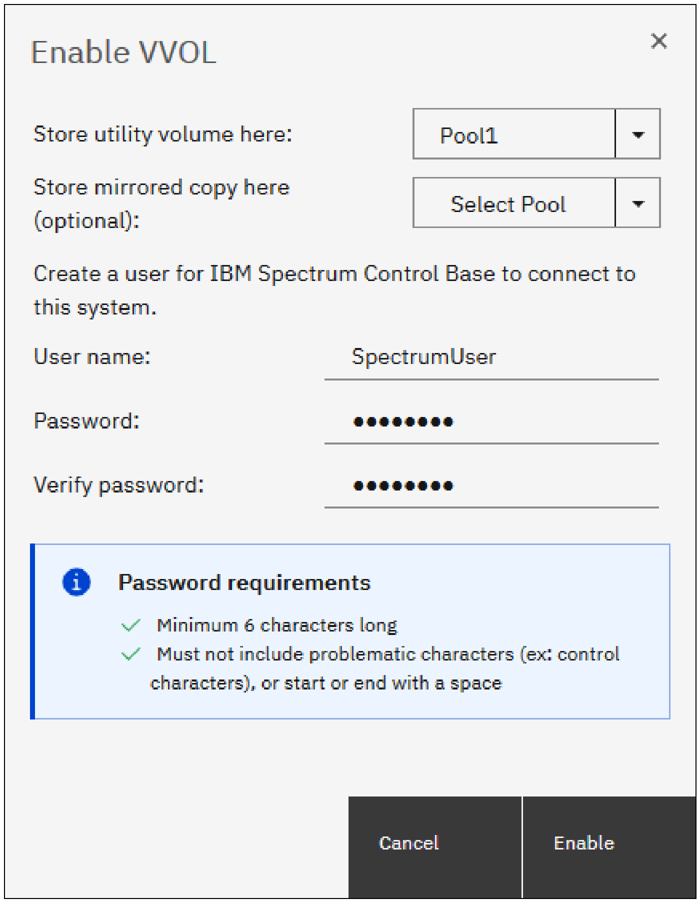


Figure 5-14 Enable VVOL window

2. Create the user account for the IBM Spectrum Connect and the user group with VMware vSphere API for Storage Awareness (VASA) provider role, if they were not set in the previous step:

   a. Create a user group by clicking **Access** → **Users by Group** → **Create User Group**. Enter the user group name, select **VASA Provider** for the role and click **Create**.

   b. Create the user account by clicking **Access** → **Users by Group**, select the user group that was created in the previous step, and click **Create User**. Enter the name of the user account, select the user group with VASA Provider role, enter a valid password for the user, and click **Create**.

3. For each ESXi host server to use Virtual Volumes, create a host object:

   a. In the management GUI, select **Hosts** → **Hosts** → **Add Host**.

   b. Enter the name of the ESXi host server, enter connection information, select **VVOL** for the host type and then, click **Add Host**.

   c. If the ESXi host was previously configured, the host type can be changed by modifying the ESXi host type.

Note: The user account with VASA Provider role is used by only the IBM Spectrum Connect server to access the IBM FlashSystem and to run the automated tasks that are required for Virtual Volumes. Users must not directly log in to the management GUI or CLI with this type of account and complete system tasks, unless they are directed to by IBM Support.

5.7 Cloud volumes

A *cloud volume* is any volume that is enabled for transparent cloud tiering. After transparent cloud tiering is enabled on a volume, point-in-time copies or snapshots can be created and copied to cloud storage that is provided by a cloud service provider. These snapshots can be restored to the system for DR purposes. Before you create cloud volumes, a valid connection to a supported cloud service provider must be configured.

With transparent cloud tiering, the system supports connections to cloud service providers and the creation of cloud snapshots of any volume or volume group on the system. Cloud snapshots are point-in-time copies of volumes that are created and transferred to cloud storage that is managed by a cloud service provider.

A cloud account defines the connection between the system and a supported cloud service provider. It also must be configured before data can be transferred to or restored from the cloud storage. After a cloud account is configured with the cloud service provider, you determine which volumes you want to create cloud snapshots of and enable transparent cloud tiering on those volumes.

Figure 5-15 shows an example of IBM Transparent Cloud Tiering.

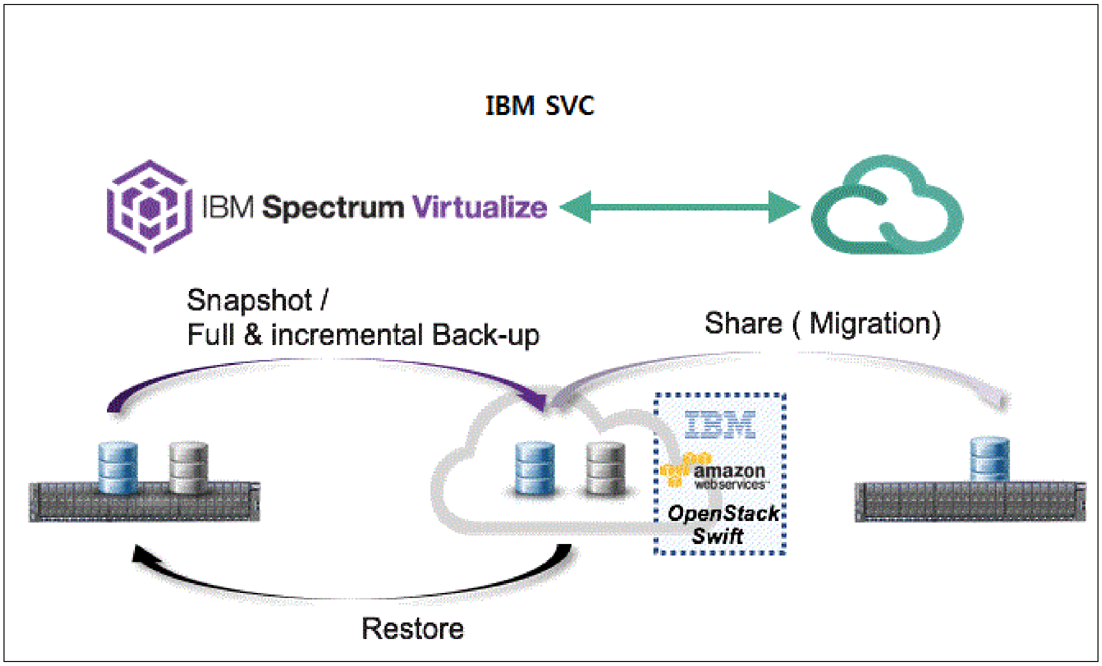


Figure 5-15 Transparent Cloud Tiering example

A cloud account is an object on the system that represents a connection to a cloud service provider by using a particular set of credentials. These credentials differ depending on the type of cloud service provider that is being specified. Most cloud service providers require the host name of the cloud service provider and an associated password. Some cloud service providers also require certificates to authenticate users of the cloud storage. Public clouds use certificates that are signed by well-known certificate authorities.

Private cloud service providers can use a self-signed certificate or a certificate that is signed by a trusted certificate authority. These credentials are defined on the cloud service provider and passed to the system through the administrators of the cloud service provider.

A cloud account defines whether the system can successfully communicate and authenticate with the cloud service provider by using the account credentials.

If the system is authenticated, it can access cloud storage to copy data to the cloud storage or restore data that is copied to cloud storage back to the system. The system supports one cloud account to a single cloud service provider. Migration between providers is *not* supported.

The system supports IBM Cloud, OpenStack Swift and Amazon S3 cloud service providers.

5.7.1 Transparent cloud tiering configuration limitations and rules

Consider the following limitations and rules regarding transparent cloud tiering:

- One cloud account per system.
- A maximum of 1024 volumes can have cloud-snapshot enabled volumes.
- The maximum number of active snapshots per volume is 256.
- The maximum number of volume groups is 512.
- Cloud volumes cannot be expanded or shrunk.
- A volume cannot be configured for a cloud snapshot if any of the following conditions exist:
  - The volume is part of a Remote Copy relationship (Metro Mirror, Global Mirror, or active-active) master, auxiliary, or change volume. This configuration prevents the cloud snapshot from being used with HyperSwap volumes.
  - The volume is a VMware vSphere Virtual Volumes volume, including IBM FlashCopy owned volumes that are used internally for Virtual Volumes restoration functions.
  - The volume is:
    - A file system volume
    - Associated with any user-owned FlashCopy maps
    - A mirrored volume with copies in different storage pools
    - Being migrated between storage pools
- A volume cannot be enabled for cloud snapshots if the cloud storage is set to import mode.
- A volume cannot be enabled for cloud snapshots if the maximum number of cloud volumes exists. The maximum number of cloud volumes on the system is 1024. If the system exceeds this limit, you can disable cloud snapshots on a cloud volume and delete its associated snapshots from the cloud storage to accommodate snapshots on new cloud volumes.
- A volume cannot be used for a restore operation if it meets any of the following criteria:
  - A Virtual Volume, including FlashCopy volumes that are used internally for Virtual Volumes restoration functions
  - A file system volume
  - Part of a Remote Copy relationship (Metro Mirror, Global Mirror, or active-active) master, auxiliary, or change volume
- A volume that is configured for backup or is being used for restoration cannot be moved between I/O groups.
- Only one operation (cloud snapshot, restore, or snapshot deletion) is allowed at a time on a cloud volume.
- Cloud volume traffic is allowed only through management interfaces (1 G or 10 G).

5.7.2 Restoring to the production volume

This process is used to restore a snapshot version to the production volume, which is the original volume from which the snapshots were created. After the restore operation completes, the snapshot version completely replaces the current data that exists on production volume. During the restore operation, the production volume goes offline until it completes. Data is not fully restored to the production volume until the changes are committed.

5.7.3 Restore to a new volume

If you do not want to have the production volume offline for the restore, you can restore a cloud snapshot to a new volume. The production volume remains online and host operations are not disrupted.

When the snapshot version is restored to a new volume, you can use the restored data independently of the original volume from which the snapshot was created. If the new volume exists on the system, the restore operation uses the unique identifier (UID) of the new volume.

If the new volume does not exist on the system, you must choose whether to use the UID from the original volume or create a UID. If you plan to use the new volume on the same system, use the UID that is associated with the snapshot version that is being restored.

5.8 Volume migration

Migrating an image mode volume to managed mode volume or vice versa is done by migrating a volume from one storage pool to another. A non-image mode volume also can be migrated to a different storage pool.

When migrating from image to managed or vice versa, the command varies, as listed in Table 5-1.

Table 5-1 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 that uses the volume. Depending on the workload of IBM FlashSystem, performance might be slightly affected.

The migration of a volume from one storage pool to another storage pool by using **`migratevdisk`** command is allowed only if both storage pools feature the same extent site. Volume mirroring can be used if a volume must be migrated from one storage pool to another storage pool with different extent sizes.

5.8.1 Image-type to striped-type volume migration

When you are migrating storage into the IBM FlashSystem, the storage is brought in as *image-type volumes*, which means that the volume is based on a single MDisk. The CLI command that can be used is **migratevdisk**.

Example 5-5 shows the **migratevdisk** command that can be used to migrate an image-type volume to a striped-type volume. The command also can be used to migrate a striped-type volume to a striped-type volume.

Example 5-5 The migratevdisk command

```
superuser> 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.

You can monitor the migration process by using the **lsmigrate** command, as shown in Example 5-6.

Example 5-6 Monitoring the migration process

```
superuser> lsmigrate
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
```

5.8.2 Migrating to image-type volume

An *image-type volume* is a direct, “straight-through” mapping to one image mode MDisk. If a volume is migrated to another MDisk, the volume is represented as being in managed mode during the migration (because it is striped on two MDisks).

It is represented only as an image-type volume after it reaches the state where it is a straight-through mapping. An image-type volume cannot be expanded.

Image-type disks are used to migrate data to an IBM FlashSystem and migrate data out of virtualization. In general, the reason for migrating a volume to an image type volume is to move the data on the disk to a non-virtualized environment.

If the migration is interrupted by a cluster recovery, the migration resumes after the recovery completes.

The **migratetoimage** command migrates the data of a user-specified volume by consolidating its extents (which might be on one or more MDisks) onto the extents of the target MDisk that you specify. After migration is complete, the volume is classified as an image type volume, and the corresponding MDisk is classified as an image mode MDisk.

The MDisk that is specified as the target must be in an *unmanaged* state at the time that the command is run. Running this command results in the inclusion of the MDisk into the user-specified storage pool.

> **Remember:** This command cannot be used if the source volume copy is in a child pool or if the target MDisk group that is specified is a child pool. This command does not work if the volume is fast formatting.

The `migratetoimage` command fails if the target or source volume is offline. Correct the offline condition before attempting to migrate the volume.

If the volume (or volume copy) is a target of a FlashCopy mapping with a source volume in an active-active relationship, the new managed disk group must be in the same site as the source volume. If the volume is in an active-active relationship, the new managed disk group must be in the same site as the source volume. Also, the site information for the MDisk being added must be well defined and match the site information for other MDisks in the storage pool.

> **Note:** You cannot migrate a volume or volume image between storage pools if cloud snapshot is enabled on the volume.

An encryption key cannot be used when migrating an image mode MDisk. To use encryption (when the MDisk has an encryption key), the MDisk must be self-encrypting before configuring storage pool.

The `migratetoimage` command is useful when you want to use your system as a data mover. For more information about the requirements and specifications for the `migratetoimage` command, see this IBM Documentation web page.

5.8.3 Migrating with volume mirroring

Volume mirroring also offers the ability to migrate volumes between storage pools with different extent sizes.

Complete the following steps to migrate volumes between storage pools:

1. Add a copy to the target storage pool.
2. Wait until the synchronization is complete.
3. Remove the copy in the source storage pool.

To migrate from a thin-provisioned volume to a fully allocated volume, the process is similar:

1. Add a target fully allocated copy.
2. Wait for synchronization to complete.
3. Remove the source thin-provisioned copy.

In both cases, if you set the `autodelete` option to `yes` when creating the volume copy, the source copy is automatically deleted, and you can skip the third step in both processes. The preferred practice on this type of migration is to try not to overload the systems with a high syncrate or with too many migrations at the same time.

The `syncrate` parameter specifies the copy synchronization rate. A value of zero prevents synchronization. The default value is 50. The supported `-syncrate` values and their corresponding rates are listed in Table 5-2.

Table 5-2 Sample syncrate values

User-specified syncrate attribute value	Data copied per second
1 - 10	128 KB
11 - 20	256 KB
21 - 30	512 KB
31 - 40	1 MB
41 - 50	2 MB
51 - 60	4 MB
61 - 70	8 MB
71 - 80	16 MB
81 - 90	32 MB
91 - 100	64 MB
101 - 110	128 MB
111 - 120	256 MB
121 - 130	512 MB
131 - 140	1 GB
141 - 150	2 GB

We recommend modifying syncrate after monitoring overall bandwidth and latency. Then, if the performance is not affected on migration, increase the syncrate to complete within the allotted time.

You can also use volume mirroring when you migrate a volume from a non-virtualized storage device to IBM FlashSystem. As you can see in Figure 5-16, you first must attach the storage to IBM FlashSystem by using the virtualization solution, which requires some downtime because hosts start to access the volumes through IBM FlashSystem.

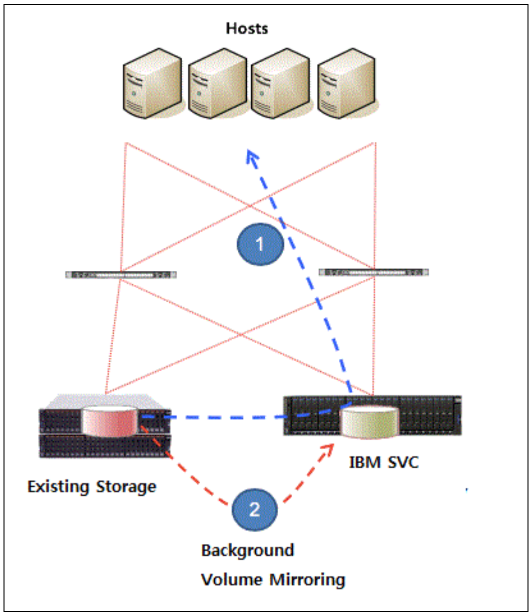


Figure 5-16 Migration with volume mirroring

After the storage is correctly attached to IBM FlashSystem, map the image-type volumes to the hosts so the host recognizes volumes as though they were accessed through the non-virtualized storage device. Then, you can restart applications.

After that process completed, you can use volume mirroring to migrate the volumes to a storage pool with managed MDisks, which creates striped-type copies of each volume in this target pool. Data synchronization in the volume copies then starts in the background.

For more information, see this IBM Documentation web page.

5.8.4 Migration from standard pools to Data Reduction Pools

If you want to migrate volumes to DRP, you can move them with volume mirroring between a standard pool and DRP. Hosts I/O operations are not disrupted during migration.

Figure 5-17 shows two examples of how you can use volume mirroring to convert volumes to a different type or migrate volumes to a different type of pool.

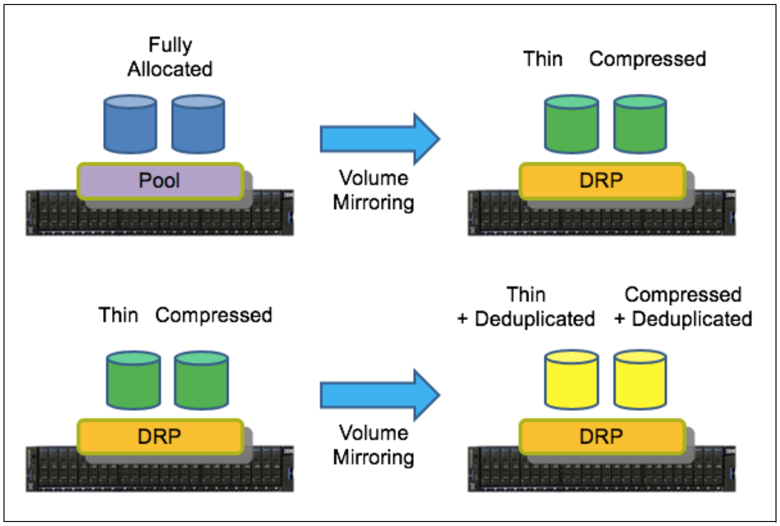


Figure 5-17 Converting volumes with volume mirroring

You can also move compressed or thin-provisioned volumes in standard pools to DRPs to simplify management of reclaimed capacity. The DRP tracks the unmap operations of the hosts and reallocates capacity automatically. The system supports volume mirroring to create a copy of the volume in a new DRP. This method creates a copy of the volume in a new DRP and does not disrupt host operations.

Deleting a volume copy in a DRP is a background task and can take a significant amount of time. During the deletion process, the deleting copy is still associated with the volume and a new volume mirror cannot be created until the deletion is complete. If you want to use volume mirroring again on the same volume without waiting for the delete, split the copy to be deleted to a new volume before deleting it.

5.8.5 Migrating a volume between systems nondisruptively

With nondisruptive system migration, storage administrators can migrate volumes from one IBM Spectrum Virtualize system to another without any application downtime. This function supports several use cases. For example, you can use this function to balance the load between multiple systems or to update and decommission hardware.

You can also migrate data between node-based systems and enclosure-based systems. Unlike replication remote-copy types, nondisruptive system migration does not require a Remote Mirroring license before you can configure a remote-copy relationship that is used for migration.

There are some configuration and host operating system restrictions that are documented in the *Configuration Limits and Restrictions* document under Volume Mobility.

Prerequisites

The following prerequisites must be met for nondisruptive system migration:

- Both systems are running 8.4.2 or later.
- A Fibre Channel or IP partnership exists between the two systems that you want to migrate volumes between. The maximum supported round-trip time (RTT) between the two systems is 3 milliseconds. Ensure that the partnership has sufficient bandwidth to support the write throughput for all the volumes you are migrating. For more information, see `mkfcpartnership` command for creating Fibre Channel partnerships and `mkippartnership` command for creating IP partnerships.
- Any hosts that are mapped to volumes that you are migrating are correctly zoned to both systems. Hosts must appear in an online state on both systems.

Using the management GUI

Complete the following steps to configure volume migration by using the GUI:

1. On the source system, select **Volumes** → **Volumes**. On the Volumes page, identify the volumes that you want to migrate and record of the volume name and capacity.
2. On the target system, select **Volumes** → **Volumes** and select **Create Volume**. Create the target volume within the appropriate storage tier with the same capacity as the source volume.
3. On the source system, select **Copy Services** → **Remote Copy**.
4. Select **Independent Relationship**.
5. Select **Create Relationship**.
6. On the Create Relationship page, select **Non-disruptive System Migration**.
7. Ensure that the auxiliary volume location specifies the system that you want to migrate to, select **Next**.
8. Select the **Master** and **Auxiliary** volumes to use in the relationship.

   > **Note:** The volumes must be the same size. If the GUI window does not show the expected auxiliary volume, check the size by running the `lsvdisk -unit b <volume name or id>` command.

9. Select **Yes** to start the copy. Click **Finish.**
10. In the management GUI, select **Copy Services** → **Remote Copy** → **Independent Relationship**. Wait until the migration relationship that you created displays the **Consistent Synchronized** state.

    > **Note:** Data is copied to the target system at the lowest of partnership background copy rate or relationship bandwidth. The relationship bandwidth default is 25 MBps per relationship and can be increased by running the `chsystem -relationshipbandwidthlimit <new value in MB>` command if needed.

11. Create host mappings to the auxiliary volumes on the remote system. Ensure that all auxiliary volumes are mapped to the same hosts that were previously mapped to the master volumes on the older system.
12. Ensure that the HBAs in all hosts that are mapped to the volume are rescanned to ensure that all new paths are detected to the auxiliary volumes. Record the current path states on any connected hosts. Identify the WWPNs used for the active and standby (ghost) paths.

13. In the management GUI, select **Copy Services** → **Remote Copy** → **Independent Relationship**. Right-click the migration relationship and select **Switch Direction.** This action reverses the copy direction of the relationship and switches the active and standby paths, which result in all host I/O being directed to the new volume.
14. Validate that all hosts use the new paths to the volume by verifying that the paths that were reporting as standby (or ghost) are now reporting active. Verify that all previously active paths are now reporting standby (or ghost).

> **Note:** Do not proceed if the added standby paths are not visible on the host. Standby paths might be listed under a different name on the host, such as "ghost" paths. Data access can be affected if all standby paths are not visible to the hosts when the direction is switched on the relationship.

15. Validate that all hosts use the target volume for I/O and verify that no issues exist.
16. On the original source system that was used in the migration, select the **Hosts** → **Hosts.** Right-click the hosts and select **Unmap Volumes.** Verify the number of volumes that are being unmapped and then, select **Unmap**.
17. On the original source system, select **Volumes** → **Volumes.** Right-click the volumes and select **Delete**. Verify the number of volumes that are being deleted and select **Continue.**

    The volume migration process is complete.

Using the command-line interface

Perform the following steps to configure volume migration by using the CLI:

1. On the source system, enter the **`lsvdisk`** command to determine all the volumes that you want to migrate to the target system.

   In the results, record the name, ID, and capacity for each volume that you want to migrate to the target system.
2. On the target system, create volumes for each volume that you want to migrate, ensuring that you create the volume with the same capacity as the source volume; for example, **`mkvolume -pool 0 -size 1000 -unit gb`**.
3. On the source system, enter the following command to create a relationship for migration: **`mkrcrelationship -master sourcevolume -aux targetvolume -cluster system2 -migration -name migrationrc`**, where *`sourcevolume`* is the name or ID of the master volume on the source system and *`targetvolume`* is the name or ID of the auxiliary volume that you created on the target system.

   The **`-migration`** flag indicates that the remote copy relationship can be used to migrate data between the two systems defined in the partnership only.

   Optionally, you can specify a name with the **`-name`** parameter. In this example, **`migrationrc`** is the name of the relationship. If no name is specified, an identifier is assigned to the relationship.
4. On the source system, start the relationship by entering the running the **`startrcrelationship migrationrc`** command, where **`migrationrc`** is the name of the relationship.
5. Verify that the state of the relationship is `consistent_synchronized`, by entering the **`lsrcrelationship migrationrc`** command, where **`migrationrc`** is the name of the relationship. In the results that display, ensure that the state is `consistent_synchronized`.

Attention: Do not proceed until the relationship is in the `consistent_synchronized` state.

Depending on the amount of data that is being migrated, the process can take some time.

Note: Data is copied to the target system at the lowest of partnership background copy rate or relationship bandwidth. The relationship bandwidth default is 25 MBps per relationship and can be increased by running the **`chsystem -relationshipbandwidthlimit <new value in MB>`** command, if needed.

6. After the relationship is in the `consistent_synchronized` state, create host mappings to the auxiliary volumes on the target system by entering the **`mkvdiskhostmap -host host1 targetvolume`** command, where `targetvolume` is the name of the auxiliary volume on the target system. Ensure that all auxiliary volumes are mapped to the same hosts that were previously mapped to the master volumes on the source system.
7. On all hosts, ensure that the HBAs are mapped to the volume are rescanned to ensure that all new paths are detected to the auxiliary volumes. Record the current path states on any connected hosts. Identify the WWPNs used for the active and standby (ghost) paths.

Attention: Do not proceed if the added standby paths are not visible on the host. Standby paths might be listed under a different name on the host, such as "ghost" paths. Data access can be affected if all standby paths are not visible to the hosts when the direction is switched on the relationship.

8. Switch the direction of the relationship so the auxiliary volume on the target system becomes the primary source for host I/O operations be entering the **`switchrcrelationship -primary aux migrationrc`** command, where **`migrationrc`** indicates the name of the relationship. This command reverses the copy direction of the relationship and switches the active and standby paths, which result in all host I/O being directed to the auxiliary volume.
9. Validate that all hosts use the new paths to the volume by verifying that the paths previously reporting as standby (or ghost) are now reporting active.
10. Verify that all previously active paths are now reporting standby (or ghost).
11. Validate that all hosts use the target volume for I/O and verify that no issues exist.
12. On the original source system, unmap hosts from the original volumes by entering the **`rmvdiskhostmap -host host1 sourcevolume`** command, where `sourcevolume` is the name of the original volume that was migrated.
13. On the original source system, delete the original source volumes by entering the **`rmvolume sourcevolume`** command, where `sourcevolume` is the name of the original volume that was migrated.

The migration process is now complete.

5.9 Preferred paths to a volume

When a volume is created, it is assigned to an I/O group and assigned a preferred node. The *preferred node* is the node that normally processes I/Os for the volume. The *primary purposes* of a preferred node are load balancing and determine which node destages writes to the backend storage.

Preferred node assignment is normally automatic. The system selects the node in the I/O group that includes the fewest volumes. However, the preferred node can be specified or changed, if needed.

All modern multipathing drivers support Asymmetric Logical Unit Access (ALUA). This access allows the storage to mark certain paths as preferred (paths to the preferred node). ALUA multipathing drivers honor preferred pathing and send I/O to only the other node if the preferred node is not accessible.

Figure 5-18 shows write operations from a host to two volumes with different preferred nodes.

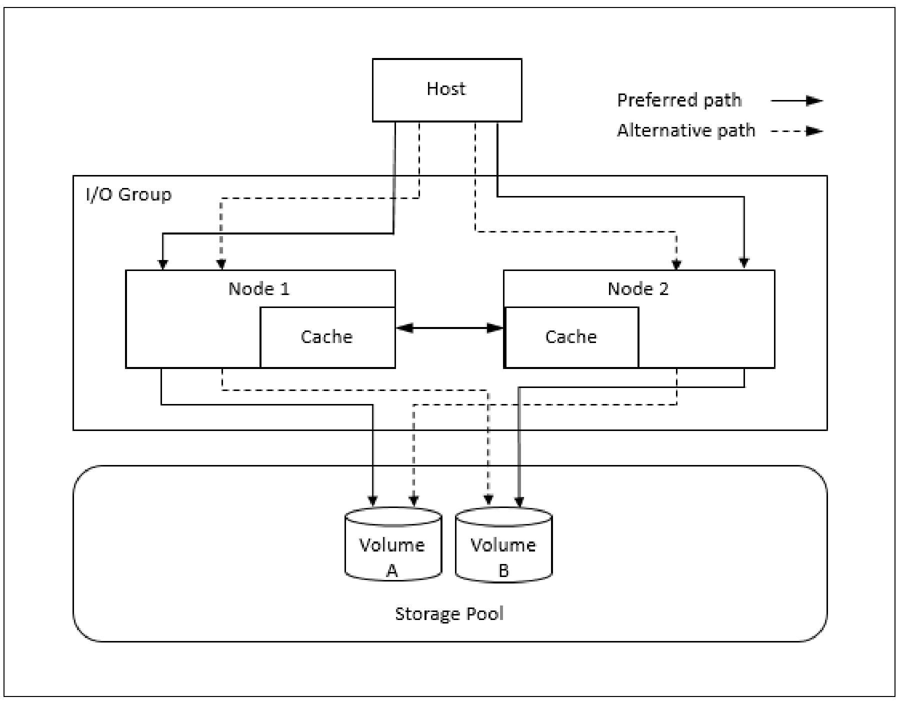


Figure 5-18 Write operations from a host

When debugging performance problems, it can be useful to review the `Non-Preferred Node Usage Percentage` metric in IBM Spectrum Control or IBM Storage Insights. I/O to the nonpreferred node might cause performance problems for the I/O group and can be identified on these tools.

For more information about this performance metric and more in IBM Spectrum Control, see this IBM Documentation web page.

5.10 Moving a volume between I/O groups and nodes

To balance the workload across I/O groups and nodes, you can move volumes between I/O groups and nodes.

The change of preferred node of a volume within an I/O group or to another I/O group is a nondisruptive process.

5.10.1 Changing the preferred node of a volume within an I/O group

Changing the preferred node within an I/O group can be done with concurrent I/O. However, it can lead to some delay in performance and in the case of some specific operating systems or applications, they might detect some timeouts.

This operation can be done by using the CLI and GUI; however, if you have only one I/O group, this operation is not possible by using the GUI. To change the preferred node within an I/O group by using CLI, use the `movevdisk -node <node_id or node_name> <vdisk_id or vdisk_name>` command.

5.10.2 Moving a volume between I/O groups

When moving a volume between I/O groups, it is recommended that the system chooses the volume preferred node in the new I/O group. However, it is possible to manually set the preferred node during this operation by using the GUI and CLI.

Some limitations exist in to moving a volume across I/O groups, which is named Non-Disruptive Volume Move (NDVM). These limitations are mostly in Host Cluster environments, and you can check the compatibility at the IBM SSIC website.

> **Note:** These migration tasks can be nondisruptive if performed correctly and the hosts that are mapped to the volume support NDVM. The cached data that is held within the system must first be written to disk before the allocation of the volume can be changed.

Modifying the I/O group that services the volume can be done concurrently with I/O operations if the host supports 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 rescan can be done in the situation where one pair of nodes becomes over-used.

If any host mappings are available for the volume, the hosts must be members of the target I/O group or the migration fails.

Verify that you created 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 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 used. These steps must be completed on all hosts to which the selected volumes are currently mapped.

> **Note:** If the selected volume is performing quick initialization, this wizard is unavailable until quick initialization is complete.

5.11 Volume throttling

Volume throttling effectively throttles the number of I/O operations per second (IOPS) or bandwidth (MBps) that can be achieved to and from a specific volume. You might want to use I/O throttling if you have a volume that has an access pattern that adversely affects the performance of other volumes.

For example, volumes that are used for backup or archive operations can have I/O intensive workloads, potentially taking bandwidth from production volumes. Volume throttle can be used to limit I/Os for these types of volumes so that I/O operations for production volumes are not affected.

Figure 5-19 shows the example of volume throttling.

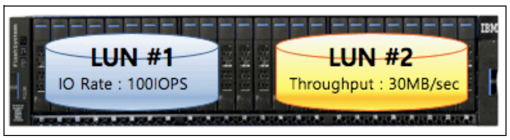


Figure 5-19 Volume throttling for each LUNs

When deciding between using IOPS or bandwidth as the I/O governing throttle, consider the disk access pattern of the application. Database applications often issue large amounts of I/O, but they transfer only a relatively small amount of data. In this case, setting an I/O governing throttle that is based on MBps does not achieve the expected result. Therefore, it is better to set an IOPS limit.

However, a streaming video application often issues a small amount of I/O, but it transfers large amounts of data. In contrast to the database example, defining an I/O throttle based in IOPS does not achieve a good result. For a streaming video application, it is better to set an MBps limit.

You can edit the throttling value in the menu, as shown in Figure 5-20.

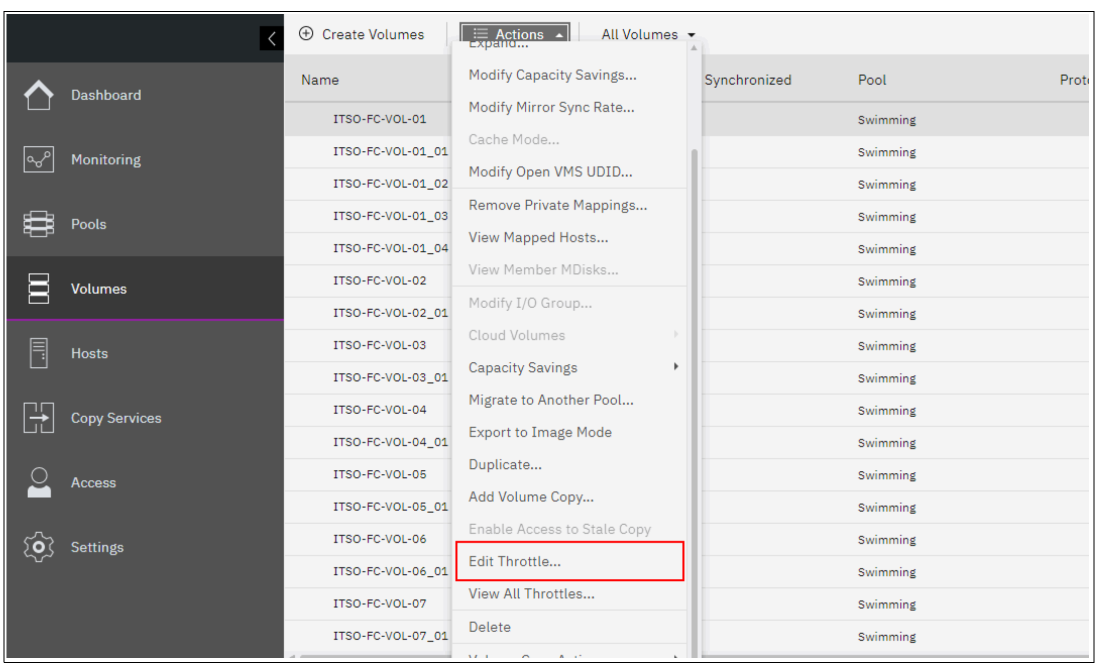


Figure 5-20 Volume throttling

Figure 5-21 shows both bandwidth and IOPS parameter that can be set.

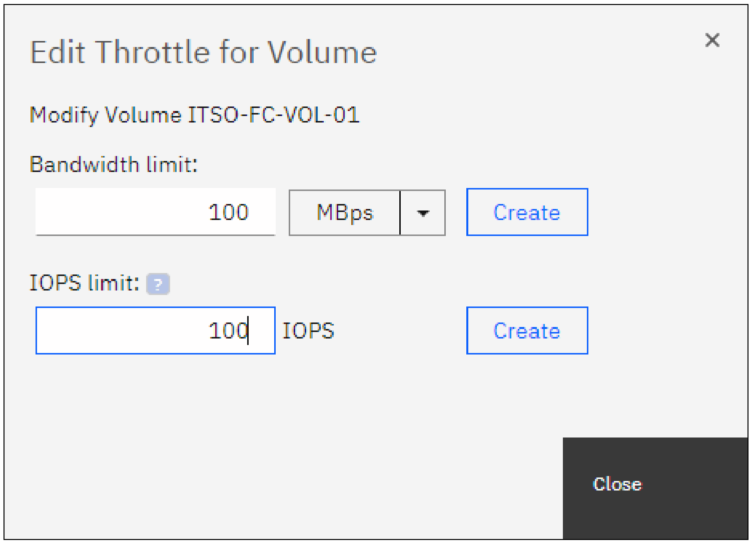


Figure 5-21 Edit bandwidth and IOPS limit

Throttling at a volume level can be set by using the following commands:

- **mkthrottle**

  This command is used to set I/O throttles for volumes that use this command. It must be used with **-type vdisk** parameter, followed by **-bandwidth bandwidth_limit_in_mbdisk** or**-iops iops_limit** to define MBps and IOPS limits.

- **chvdisk**

  When used with **-rate throttle_rate** parameter, this command specifies the IOPS and MBps limits. The default **throttle_rate** units are I/Os. To change the **throttle_rate** units to megabits per second (MBps), specify the **-unitmb** parameter. If **throttle_rate** value is zero, the throttle rate is disabled. By default, the **throttle_rate** parameter is disabled.

> **Note:** The **mkthrottle** command can be used to create throttles for volumes, hosts, host clusters, pools, or system offload commands.

When the IOPS limit is configured on a volume and it is smaller than 100 IOPS, the throttling logic rounds it to 100 IOPS. Even if throttle is set to a value smaller than 100 IOPS, the throttling occurs at 100 IOPS.

After any of the commands that were described thus far are used to set volume throttling, a throttle object is created. Then, you can list your created throttle objects by using the **lsthrottle** command and change their parameters with the **chthrottle** command. Example 5-7 shows some command examples.

Example 5-7 Throttle commands example

```
superuser>mkthrottle -type vdisk -bandwidth 100 -vdisk Vol01
Throttle, id [0], successfully created.
superuser>lsthrottle
throttle_id throttle_name object_id object_name throttle_type IOPs_limit
bandwidth_limit_MB
0           throttle0     52        Vol01       vdisk                   100
```

```
superuser>chthrottle -iops 1000 throttle0
superuser>lsthrottle
throttle_id throttle_name object_id object_name throttle_type IOPs_limit
bandwidth_limit_MB
0           throttle0     52        Vol01       vdisk         1000       100

superuser>lsthrottle throttle0
id 0
throttle_name throttle0
object_id 52
object_name Vol01
throttle_type vdisk
IOPs_limit 1000
bandwidth_limit_MB 100
```

Note: The throttle reduces IOPS or bandwidth by adding latency as a resource approaches a defined throttle. This increased response time is observable in performance monitoring tools.

For more information and the procedure to set volume throttling, see IBM Documentation.

5.12 Volume cache mode

Cache mode in IBM FlashSystem determines whether read and write operations are stored in cache. For each volume, one of the following cache modes can be used:

- `readwrite` (enabled)

  All read and write I/O operations that are performed by the volume are stored in cache. This default cache mode is used for all volumes. A volume or volume copy that is created from a DRP must have a cache mode of readwrite.

  When you create a thin provisioned volume, set the cache mode to `readwrite` to maximize performance. If you set the mode to `none`, the system cannot cache the thin-provisioned metadata and performance is decreased. In a DRP, a thin-provisioned or compressed volume copy setting cannot be created for the cache mode that is different than `readwrite`.

- `readonly`

  All read I/O operations that are performed by the volume are stored in cache.

- `none` (disabled)

  All read and write I/O operations that are performed by the volume are not stored in cache.

By default, when a volume is created, the cache mode is set to `readwrite`. Disabling cache can affect performance and increase read and write response time.

Figure 5-22 shows write operation behavior when volume cache is activated (`readwrite`).

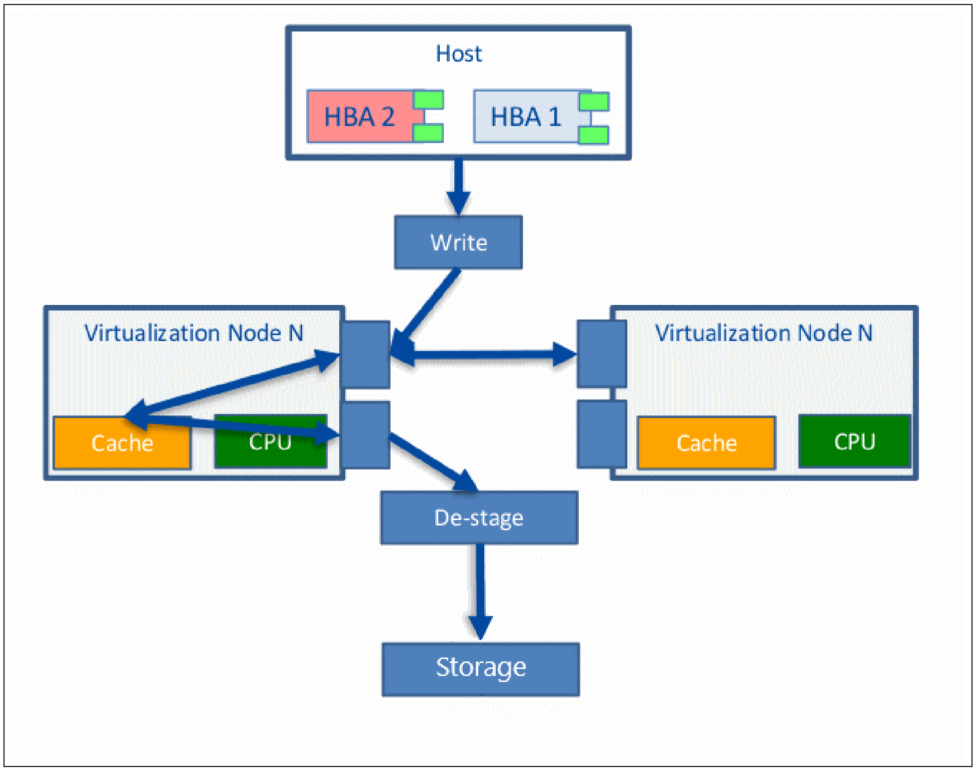


Figure 5-22 Cache activated

Figure 5-23 shows a write operation behavior when volume cache is deactivated (`none`).

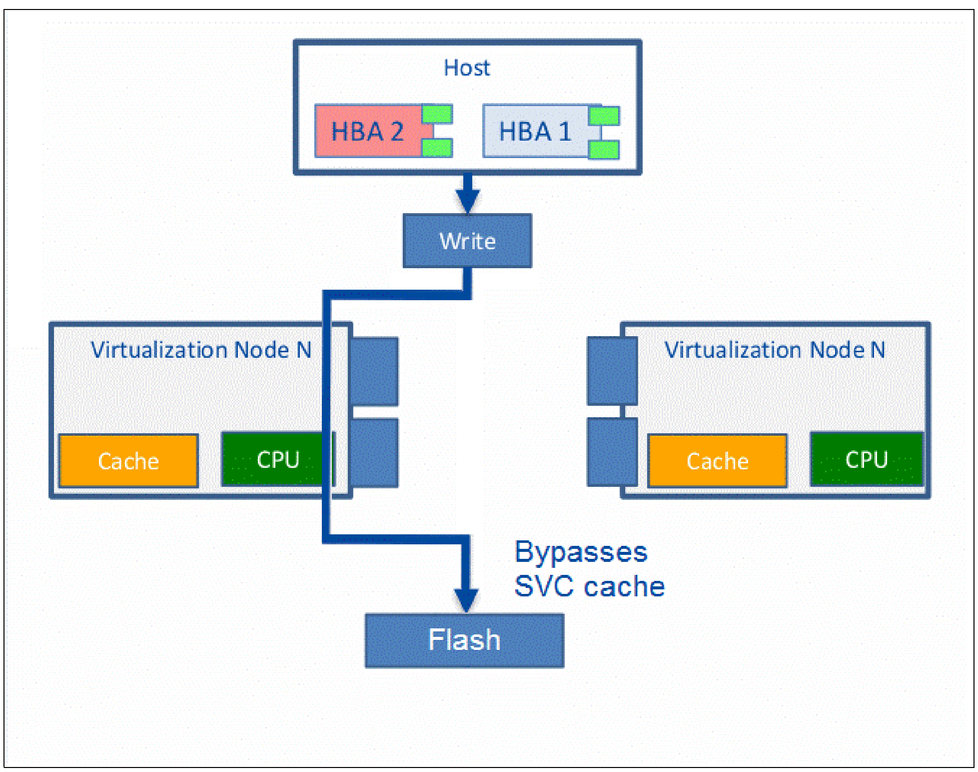


Figure 5-23 Cache deactivated

In most cases, the volume with `readwrite` cache mode is recommended because disabling cache for a volume can result in performance issues to the host. However, some specific scenarios exist in which it is recommended to disable the `readwrite` cache.

You might use cache-disabled (`none`*)* volumes when you have Remote Copy or FlashCopy in a backend storage controller, and these volumes are virtualized in IBM FlashSystem devices as image VDisks. Another possible use of a cache-disabled volume is when intellectual capital is in 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 specific volumes, more cache resources are available to cache I/Os to other volumes in the same I/O group. An example of this use is a non-critical application that uses volumes in MDisks from all-flash storage.

Note: Volumes with `readwrite` cache enabled is recommended.

By default, volumes are created with cache mode enabled (read/write); however, you can specify the cache mode when the volume is created by using the **`-cache`** option.

The cache mode of a volume can be concurrently changed (with I/O) by using the **`chvdisk`** command or GUI, and selecting **Volumes** → **Volumes** → **Actions** → **Cache Mode**.
Figure 5-24 shows editing cache mode for a volume.

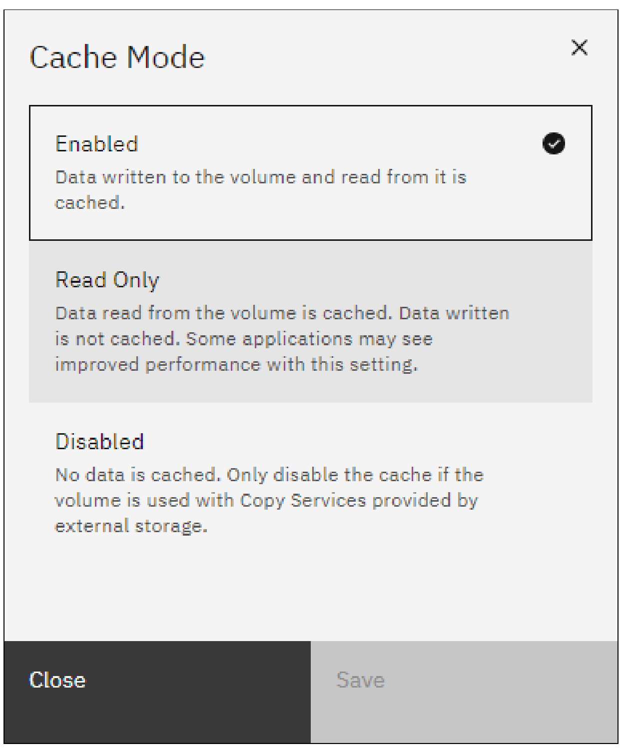


Figure 5-24 Edit cache mode

The command line does 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 does not result in a corrupted volume. Therefore, the cache must be flushed and you must discard cache data if the user disables cache on a volume.

Example 5-8 shows an image volume `VDISK_IMAGE_1` that changed the cache parameter after it was created.

Example 5-8 Changing the cache mode of a volume

```
superuser>mkvdisk -name VDISK_IMAGE_1 -iogrp 0 -mdiskgrp IMAGE_Test -vtype image
-mdisk D8K_L3331_1108
Virtual Disk, id [9], successfully created
superuser>lsvdisk VDISK_IMAGE_1
id 9
.
lines removed for brevity
.
fast_write_state empty
cache readwrite
.
lines removed for brevity

superuser>chvdisk -cache none VDISK_IMAGE_1
superuser>lsvdisk VDISK_IMAGE_1
id 9
.
lines removed for brevity
.
cache none
.
lines removed for brevity
```

In an environment with Copy Services (FlashCopy, Metro Mirror, Global Mirror, and volume mirroring) and typical workloads, disabling IBM FlashSystem cache is detrimental to overall performance.

Attention: Carefully evaluate the effect to the entire system with quantitative analysis before and after making this change.

5.13 Other considerations

This section describes other considerations regarding volumes.

5.13.1 Volume protection

You can protect volumes to prevent active volumes or host mappings from being deleted. IBM FlashSystem features a global setting enabled by default that prevents these objects from being deleted if the system detects recent I/O activity. You can set this value to apply to all volumes that are configured on your system, or control whether the system-level volume protection is enabled or disabled on specific pools.

To prevent an active volume from being deleted unintentionally, administrators must enable volume protection. They can also specify a period that the volume must be idle before it can be deleted. If volume protection is enabled and the period is not expired, the volume deletion fails, even if the **-force** parameter is used.

When you delete a volume, the system verifies whether it is a part of a host mapping, FlashCopy mapping, or remote-copy relationship. In these cases, the system fails to delete the volume, unless the **-force** parameter is specified. However, if volume protection is enabled, the **-force** parameter does not delete a volume if it has I/O activity in the last minutes defined in the protection duration time in volume protection.

> **Note:** The **-force** parameter overrides the volume dependencies, not the volume protection setting. Volume protection must be disabled to permit a volume or host-mapping deletion if the volume had recent I/O activity.

Consider enabling volume protection by using **chsystem vdiskprotectionenabled yes -vdiskprotectiontime <value_in_minutes>**.

If you want to have volume protection enabled in your system, but disabled in a specific storage pool, you can use the **chmdiskgrp -vdiskprotectionenabled no <pool_name_or_ID>** command.

You can also manage volume protection in GUI by selecting **Settings** → **System** → **Volume Protection**, as shown in Figure 5-25.

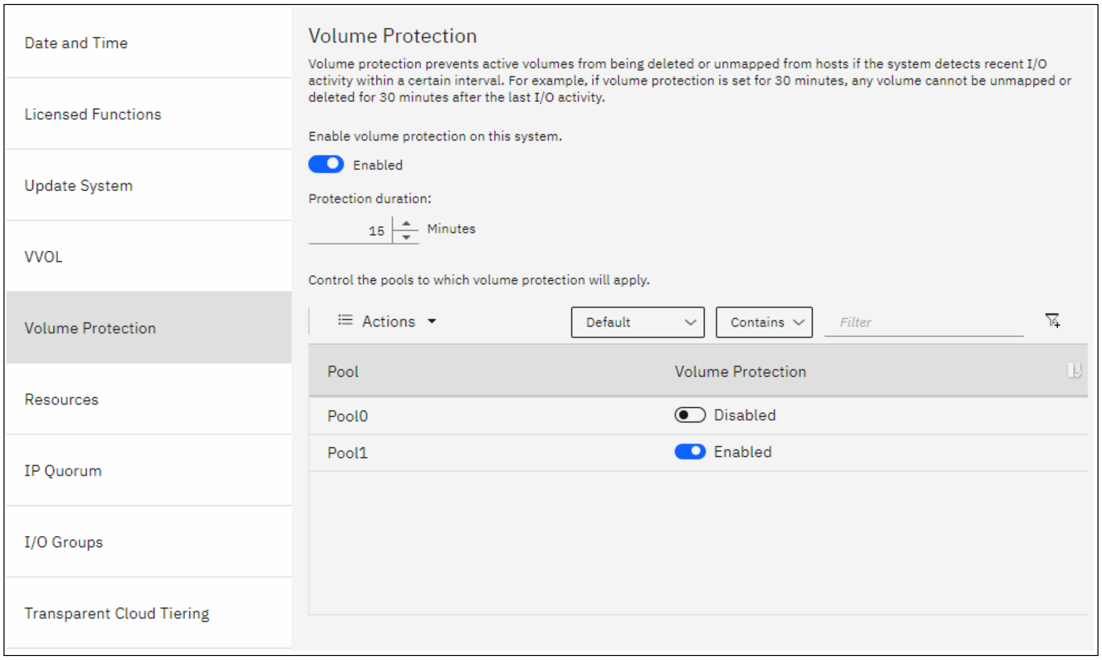


Figure 5-25   Volume Protection

5.13.2 Volume resizing

Fully allocated and thin-provisioned volumes can have their sizes increased or decreased. A volume can be expanded with concurrent I/Os for some operating systems. However, *never* attempt to shrink a volume that is in use that contains data because volume capacity is removed from the end of the disk, whether that capacity is in use by a server. A volume cannot be expanded or shrunk during its quick initialization process.

Expanding a volume

You can expand volumes for the following reasons:

- To increase the available capacity on a specific volume that is mapped to a host.
- To increase the size of a volume to make it match the size of the source or master volume so that it can be used in a FlashCopy mapping or Metro Mirror relationship.

Figure 5-26 shows the Expand Volume window.

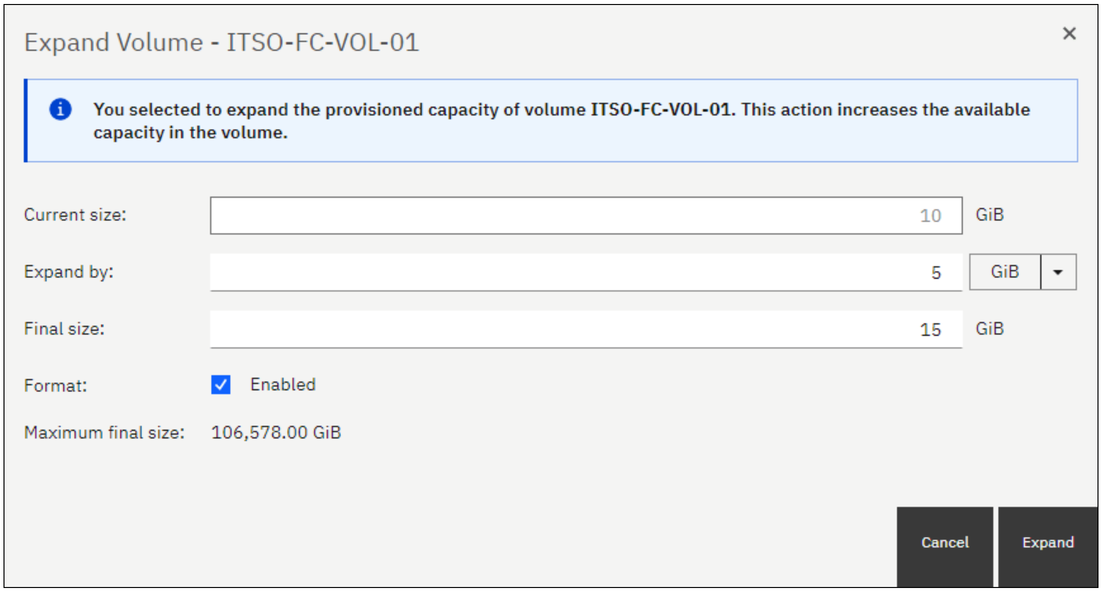


Figure 5-26 Expand Volumes

Shrinking a volume

Volumes can be reduced in size if necessary. If a volume does not contain any data, it is unlikely that you encounter any issues when shrinking its size. However, if a volume is in use and contains data, do not shrink its size because IBM Spectrum Virtualize is unaware if it is removing used or non-used capacity.

Attention: When you shrink a volume, capacity is removed from the end of the disk, whether that capacity is in use. Even if a volume includes free capacity, do not assume that only unused capacity is removed when you shrink a volume.

Figure 5-27 shows shrinking volumes.

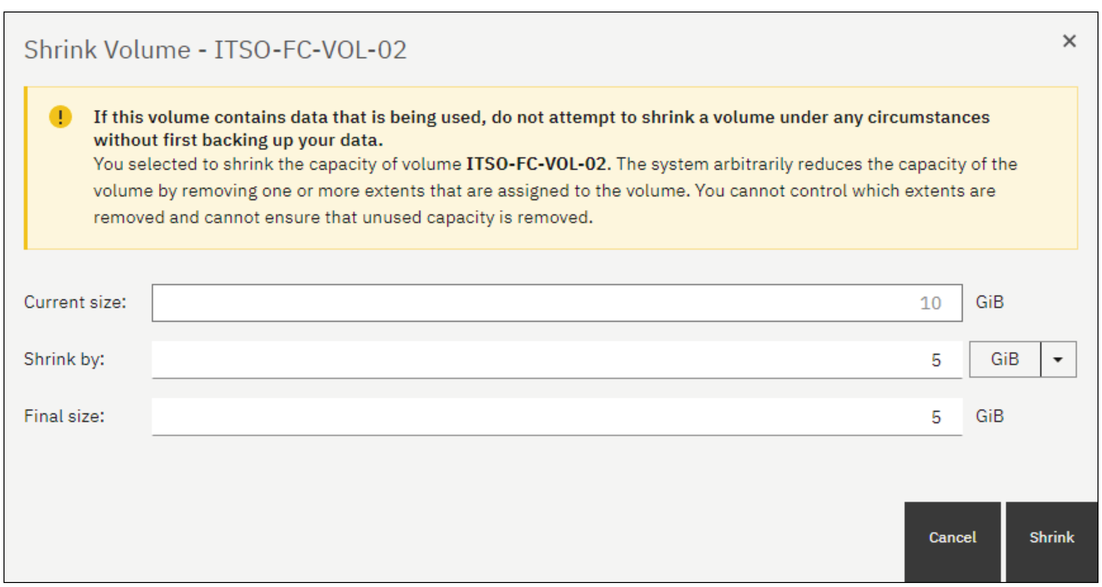


Figure 5-27 Shrink volumes

6

IBM FlashCopy services capabilities

Copy services are a collection of functions that provide capabilities for disaster recovery, data migration, and data duplication solutions.

This chapter provides an overview and the preferred practices of IBM FlashSystem copy services capabilities, including FlashCopy, Metro Mirror and Global Mirror, and volume mirroring.

This chapter includes the following sections:

6.1 Introduction to copy services

IBM Spectrum Virtualize based systems, including the IBM FlashSystem family, offer a complete set of copy services functions that provide capabilities for disaster recovery, business continuity, data movement, and data duplication solutions.

6.1.1 FlashCopy

FlashCopy is a function that allows you to create a point-in-time copy of one of your volumes. This function might be helpful when performing backups or application testing. These copies can be cascaded on one another, read from, written to, and even reversed. These copies are able to conserve storage, if needed, by being space-efficient copies that only record items that have changed from the originals instead of full copies.

6.1.2 Metro Mirror and Global Mirror

Metro Mirror and Global Mirror are technologies that enable you to keep a real-time copy of a volume at a remote site that contains another IBM Spectrum Virtualize based system. Consider the following points:

- Metro Mirror makes synchronous copies of your volumes. This means that the original writes are not considered complete until the write to the destination volume has been confirmed. The distance between your two sites is usually determined by the amount of latency your applications can handle.
- Global Mirror makes *asynchronous* copies of your volumes. This means that the write is considered complete after it is complete at the local volume. It does not wait for the write to be confirmed at the remote system as Metro Mirror does. This requirement greatly reduces the latency experienced by your applications if the other system is far away. However, it also means that during a failure, the data on the Remote Copy might not have the most recent changes committed to the local volume.

  IBM Spectrum Virtualize provides two types of asynchronous mirroring technology:

  - The standard Global Mirror (referred to as Global Mirror)
  - The Global Mirror with Change Volume (GMCV)

6.1.3 Volume mirroring

Volume mirroring is a function that is designed to increase high availability of the storage infrastructure. It provides the ability to create up to two local copies of a volume. Volume mirroring can use space from two storage pools, and preferably from two separate back-end disk subsystems.

Primarily, you use this function to insulate hosts from the failure of a storage pool and also from the failure of a back-end disk subsystem. During a storage pool failure, the system continues to provide service for the volume from the other copy on the other storage pool, with no disruption to the host.

You can also use Volume mirroring to change the capacity saving of a volume, and to migrate data between storage pools of different extent sizes and characteristics.

6.2 IBM FlashCopy

By using the IBM FlashCopy function of the IBM FlashSystem, you can perform a *point-in-time copy* of one or more volumes. This section describes the inner workings of FlashCopy, and provides some preferred practices for its use.

You can use FlashCopy to help you solve critical and challenging business needs that require duplication of data of your source volume. Volumes can remain online and active while you create consistent copies of the data sets. Because the copy is performed at the block level, it operates below the host operating system and its cache. Therefore, the copy is not apparent to the host.

> **Important:** Because FlashCopy operates at the block level below the host operating system and cache, those levels do need to be flushed for consistent FlashCopies.

While the FlashCopy operation is performed, the source volume is stopped briefly to initialize the FlashCopy bitmap, and then input/output (I/O) can resume. Although several FlashCopy options require the data to be copied from the source to the target in the background, which can take time to complete, the resulting data on the target volume is presented so that the copy appears to complete immediately.

This process is performed by using a bitmap (or bit array) that tracks changes to the data after the FlashCopy is started, and an indirection layer that enables data to be read from the source volume transparently.

6.2.1 FlashCopy use cases

When you are deciding whether FlashCopy addresses your needs, you must adopt a combined business and technical view of the problems that you want to solve. First, determine the needs from a business perspective. Then, determine whether FlashCopy can address the technical needs of those business requirements.

The business applications for FlashCopy are wide-ranging. In the following sections, a short description of the most common use cases is provided.

Backup improvements with FlashCopy

FlashCopy does not reduce the time that it takes to perform a backup to traditional backup infrastructure. However, it can be used to minimize and, under certain conditions, eliminate application downtime that is associated with performing backups. FlashCopy can also transfer the resource usage of performing intensive backups from production systems.

After the FlashCopy is performed, the resulting image of the data can be backed up to tape as though it were the source system. After the copy to tape is complete, the image data is redundant and the target volumes can be discarded. For time-limited applications, such as these examples, “no copy” or incremental FlashCopy is used most often. The use of these methods puts less load on your infrastructure.

When FlashCopy is used for backup purposes, the target data usually is managed as read-only at the operating system level. This approach provides extra security by ensuring that your target data was not modified and remains true to the source.

Restore with FlashCopy

FlashCopy can perform a restore from any existing FlashCopy mapping. Therefore, you can restore (or copy) from the target to the source of your regular FlashCopy relationships. It might be easier to think of this method as reversing the direction of the FlashCopy mappings. This capability has the following benefits:

- There is no need to worry about pairing mistakes because you trigger a restore.
- The process appears instantaneous.
- You can maintain a pristine image of your data while you are restoring what was the primary data.

This approach can be used for various applications, such as recovering your production database application after an errant batch process that caused extensive damage.

> **Preferred practices:** Although restoring from a FlashCopy is quicker than a traditional tape media restore, do not use restoring from a FlashCopy as a substitute for good archiving practices. Instead, keep one to several iterations of your FlashCopies so that you can near-instantly recover your data from the most recent history. Keep your long-term archive as appropriate for your business.

In addition to the restore option, which copies the original blocks from the target volume to modified blocks on the source volume, the target can be used to perform a restore of individual files. To do that, you must make the target available on a host. Do not make the target available to the source host, because seeing duplicates of disks causes problems for most host operating systems. Copy the files to the source by using the normal host data copy methods for your environment.

Moving and migrating data with FlashCopy

FlashCopy can be used to facilitate the movement or migration of data between hosts while minimizing downtime for applications. By using FlashCopy, application data can be copied from source volumes to new target volumes while applications remain online. After the volumes are fully copied and synchronized, the application can be brought down and then immediately brought back up on the new server that is accessing the new FlashCopy target volumes.

> **Use case:** FlashCopy can be used to migrate volumes from and to DRPs, which do not support extent based migrations.

This method differs from the other migration methods, which are described later in this chapter. Common uses for this capability are host and back-end storage hardware refreshes.

Application testing with FlashCopy

It is often important to test a new version of an application or operating system that is using actual production data. This testing ensures the highest quality possible for your environment. FlashCopy makes this type of testing easy to accomplish without putting the production data at risk or requiring downtime to create a constant copy.

Create a FlashCopy of your source and use that for your testing. This copy is a duplicate of your production data down to the block level so that even physical disk identifiers are copied. Therefore, it is impossible for your applications to tell the difference.

Cyber Resiliency

FlashCopy is the foundation of the Spectrum Virtualize *Safeguarded Copy* function that supports the ability to create cyber-resilient point-in-time copies of volumes that cannot be changed or deleted through user errors, malicious actions, or ransomware attacks.

The Safeguarded Copy function supports creating cyber-resilient copies of your important data by implementing the following features:

- Separation of duties provides more security capabilities to prevent non-privileged users from compromising production data. Operations that are related to Safeguarded backups are restricted to only a subset of users with specific roles on the system (Administrator, Security Administrator, Superuser).
- Protected Copies provides capabilities to regularly create Safeguarded backups. Safeguarded backups cannot be mapped directly to hosts to prevent any application from changing these copies.
- Automation manages safeguarded backups and restores and recovers data with the integration of IBM Copy Services Manager. IBM Copy Services Manager automates the creation of Safeguarded backups according to the schedule that is defined in a Safeguarded policy. IBM Copy Services Manager supports testing, restoring, and recovering operations with Safeguarded backups.

  For more information about Safeguarded Copy see *Implementation Guide for SpecV/FlashSystem Safeguarded Copy*, REDP-5654.

6.2.2 FlashCopy capabilities overview

FlashCopy occurs between a source volume and a target volume in the same storage system. The minimum granularity that IBM FlashSystem systems support for FlashCopy is an entire volume. It is not possible to use FlashCopy to copy only part of a volume.

To start a FlashCopy operation, a relationship between the source and the target volume must be defined. This relationship is called *FlashCopy Mapping*.

FlashCopy mappings can be stand-alone or a member of a Consistency Group. You can perform the actions of preparing, starting, or stopping FlashCopy on either a stand-alone mapping or a Consistency Group.

Figure 6-1 shows the concept of FlashCopy mapping.

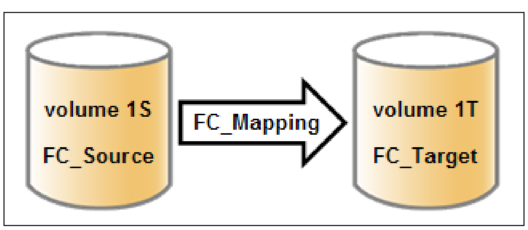


Figure 6-1 FlashCopy mapping

A FlashCopy mapping has a set of attributes and settings that define the characteristics and the capabilities of the FlashCopy.

These characteristics are explained in more detail in the following sections.

Background copy

The *background copy rate* is a property of a FlashCopy mapping that allows to specify whether a background physical copy of the source volume to the corresponding target volume occurs. A value of 0 disables the background copy. If the FlashCopy background copy is disabled, only data that has changed on the source volume is copied to the target volume. A FlashCopy with background copy disabled is also known as *No-Copy* FlashCopy.

The benefit of using a FlashCopy mapping with background copy enabled is that the target volume becomes a real clone (independent from the source volume) of the FlashCopy mapping source volume after the copy is complete. When the background copy function is not performed, the target volume remains a valid copy of the source data while the FlashCopy mapping remains in place.

Valid values for the background copy rate are 0 - 150. The background copy rate can be defined and changed dynamically for individual FlashCopy mappings.

Table 6-1 lists the relationship of the background copy rate value to the attempted amount of data to be copied per second.

Table 6-1 Relationship between the rate and data rate per second

Value	Data copied per second
1 - 10	128 KB
11 - 20	256 KB
21 - 30	512 KB
31 - 40	1 MB
41 - 50	2 MB
51 - 60	4 MB
61 - 70	8 MB
71 - 80	16 MB
81 - 90	32 MB
91 - 100	64 MB
101-110	128 MB
111-120	256 MB
121-130	512 MB
131-140	1024 MB
141-150	2048 MB

Note: To ensure optimal performance of all IBM Spectrum Virtualize features, it is advised not to exceed a copyrate value of 130.

FlashCopy Consistency Groups

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

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

Figure 6-2 shows a Consistency Group that consists of two volume mappings.

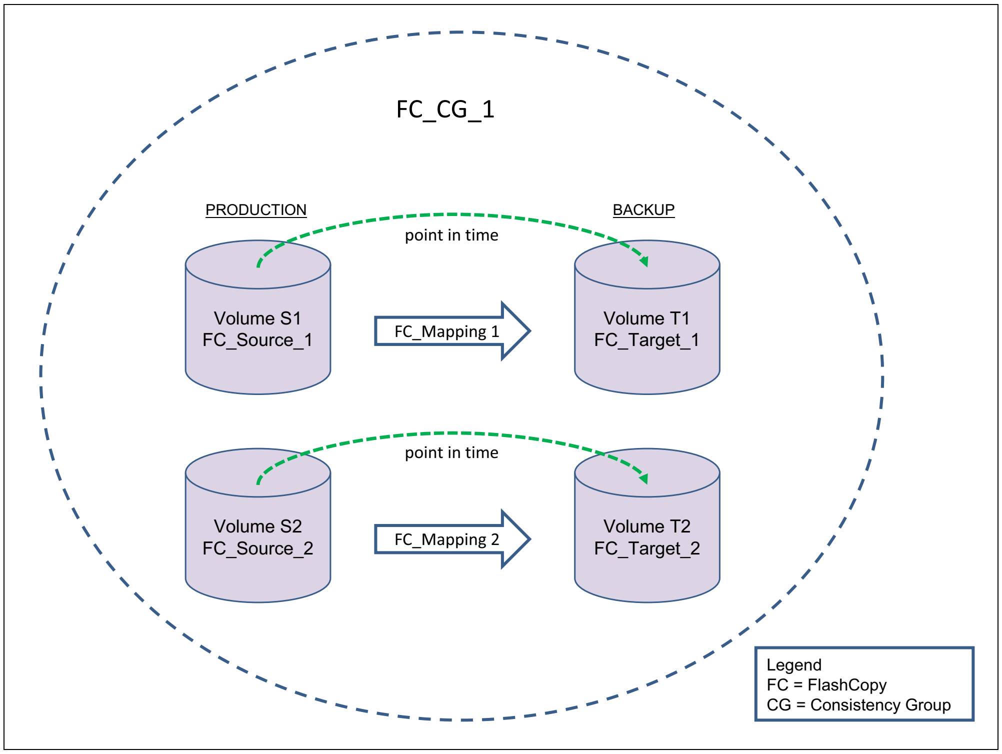


Figure 6-2 Multiple volumes mapping in a Consistency Group

> **FlashCopy mapping considerations:** If the FlashCopy mapping has been added to a Consistency Group, it can only be managed as part of the group. This limitation means that FlashCopy operations are no longer allowed on the individual FlashCopy mappings.

Incremental FlashCopy

By using Incremental FlashCopy, you can reduce the required time of copy. Also, because less data must be copied, the workload put on the system and the back-end storage is reduced.

Incremental FlashCopy does not require that you copy an entire disk source volume whenever the FlashCopy mapping is started. Instead, only the changed regions on source volumes are copied to target volumes, as shown in Figure 6-3.

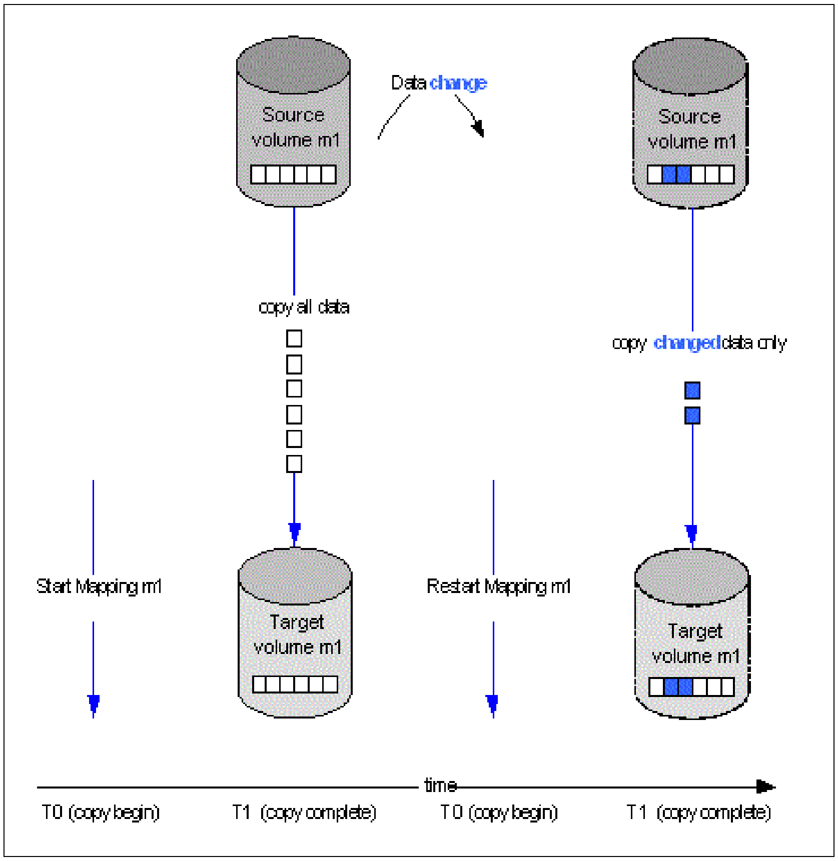


Figure 6-3 Incremental FlashCopy

If the FlashCopy mapping was stopped before the background copy completed, then when the mapping is restarted, the data that was copied before the mapping was stopped will not be copied again. For example, if an incremental mapping reaches 10 percent progress when it is stopped and then it is restarted, that 10 percent of data will not be recopied when the mapping is restarted, assuming that it was not changed.

Stopping an incremental FlashCopy mapping: If you are planning to stop an incremental FlashCopy mapping, make sure that the copied data on the source volume will not be changed, if possible. Otherwise, you might have an inconsistent point-in-time copy.

A *difference* value is provided in the query of a mapping, which makes it possible to know how much data has changed. This data must be copied when the Incremental FlashCopy mapping is restarted. The difference value is the percentage (0-100 percent) of data that has been changed. This data must be copied to the target volume to get a fully independent copy of the source volume.

An incremental FlashCopy can be defined setting the *incremental* attribute in the FlashCopy mapping.

Multiple Target FlashCopy

In Multiple Target FlashCopy, a source volume can be used in multiple FlashCopy mappings, while the target is a different volume, as shown in Figure 6-4.

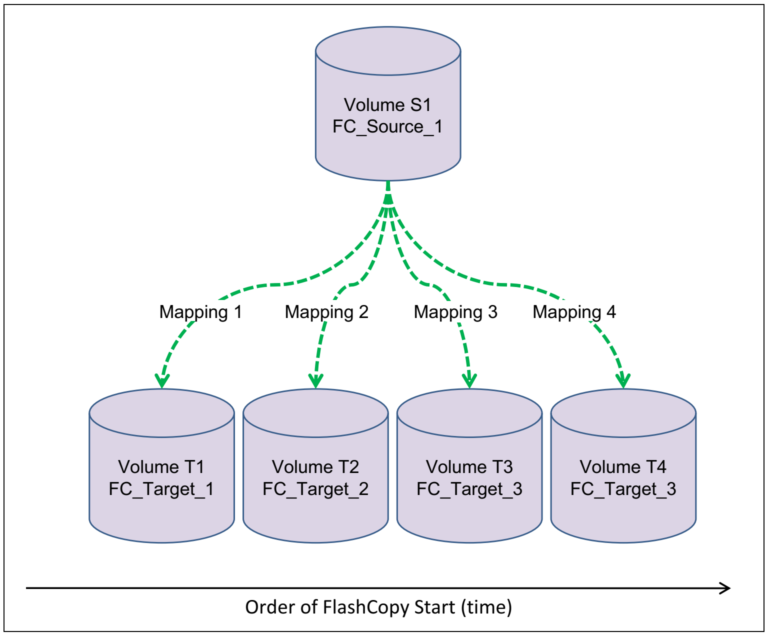


Figure 6-4 Multiple Target FlashCopy

Up to 256 different mappings are possible for each source volume. These mappings are independently controllable from each other. Multiple Target FlashCopy mappings can be members of the same or different Consistency Groups. In cases where all the mappings are in the same Consistency Group, the result of starting the Consistency Group will be to FlashCopy to multiple identical target volumes.

Cascaded FlashCopy

With Cascaded FlashCopy, you can have a source volume for one FlashCopy mapping and as the target for another FlashCopy mapping; this is referred to as a *Cascaded FlashCopy*. This function is illustrated in Figure 6-5.

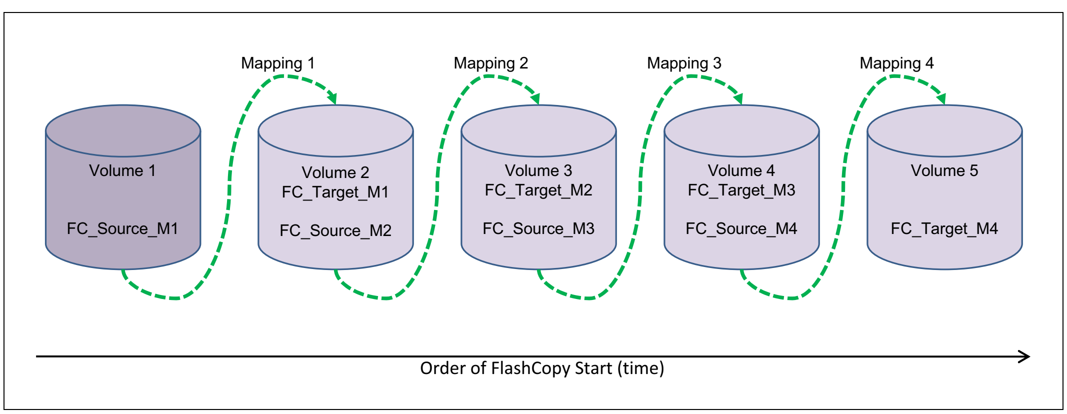


Figure 6-5 Cascaded FlashCopy

A total of 255 mappings are possible for each cascade.

Reverse FlashCopy

Reverse FlashCopy enables FlashCopy targets to become restore points for the source without breaking the FlashCopy relationship, and without having to wait for the original copy operation to complete. It can be used in combination with the Multiple Target Flashcopy to create multiple rollback points.

A key advantage of the Multiple Target Reverse FlashCopy function is that the reverse FlashCopy does not destroy the original target. This feature enables processes that are using the target, such as a tape backup, to continue uninterrupted. IBM FlashSystem systems also allow you to create an optional copy of the source volume to be made before the reverse copy operation starts. This ability to restore back to the original source data can be useful for diagnostic purposes.

Thin-provisioned FlashCopy

When a new volume is created, you can designate it as a *thin-provisioned volume,* and it has a virtual capacity and a real capacity.

Virtual capacity is the volume storage capacity that is available to a host. *Real capacity* is the storage capacity that is allocated to a volume copy from a storage pool. In a fully allocated volume, the virtual capacity and real capacity are the same. However, in a thin-provisioned volume, the virtual capacity can be much larger than the real capacity.

The virtual capacity of a thin-provisioned volume is typically larger than its real capacity. On IBM Spectrum Virtualize based systems, the real capacity is used to store data that is written to the volume, and metadata that describes the thin-provisioned configuration of the volume. As more information is written to the volume, more of the real capacity is used.

Thin-provisioned volumes can also help to simplify server administration. Instead of assigning a volume with some capacity to an application and increasing that capacity following the needs of the application if those needs change, you can configure a volume with a large virtual capacity for the application. You can then increase or shrink the real capacity as the application needs change, without disrupting the application or server.

When you configure a thin-provisioned volume, you can use the warning level attribute to generate a warning event when the used real capacity exceeds a specified amount or percentage of the total real capacity. For example, if you have a volume with 10 GB of total capacity and you set the warning to 80 percent, an event is registered in the event log when you use 80 percent of the total capacity. This technique is useful when you need to control how much of the volume is used.

If a thin-provisioned volume does not have enough real capacity for a write operation, the volume is taken offline and an error is logged (error code 1865, event ID 060001). Access to the thin-provisioned volume is restored by either increasing the real capacity of the volume or increasing the size of the storage pool on which it is allocated.

You can use thin volumes for cascaded FlashCopy and multiple target FlashCopy. It is also possible to mix thin-provisioned with normal volumes. It can be used for incremental FlashCopy too, but using thin-provisioned volumes for incremental FlashCopy only makes sense if the source and target are thin-provisioned.

When using thin provisioned volumes on Data Reduction Pools (DRPs), consider also implementing compression because it provides several benefits:

- Reduced amount of I/O operation to the back-end as the amount of data to be actually written to the back-end reduces with compressed data. This is particularly relevant with a poorly performing back-end, but less of an issue with the high performing back-end on IBM FlashSystem systems.
- Space efficiency as the compressed data provides more capacity savings.
- Better back-end capacity monitoring, as DRP pools with thin provisioned uncompressed volumes do not provide physical allocation information.

Therefore, the recommendation is to always enable compression on DRP thin provisioned volumes.

Thin-provisioned incremental FlashCopy

The implementation of thin-provisioned volumes does not preclude the use of incremental FlashCopy on the same volumes. It does not make sense to have a fully allocated source volume and then use incremental FlashCopy, which is always a full copy at first, to copy this fully allocated source volume to a thin-provisioned target volume. However, this action is not prohibited.

Consider this optional configuration:

- A thin-provisioned source volume can be copied incrementally by using FlashCopy to a thin-provisioned target volume. Whenever the FlashCopy is performed, only data that has been modified is recopied to the target. Note that if space is allocated on the target because of I/O to the target volume, this space will not be reclaimed with subsequent FlashCopy operations.
- A fully allocated source volume can be copied incrementally using FlashCopy to another fully allocated volume at the same time as it is being copied to multiple thin-provisioned targets (taken at separate points in time). This combination allows a single full backup to be kept for recovery purposes, and separates the backup workload from the production workload. At the same time, it allows older thin-provisioned backups to be retained.

6.2.3 FlashCopy functional overview

Understanding how FlashCopy works internally helps you to configure it and enables you to obtain more benefits from it.

FlashCopy mapping states

A FlashCopy mapping defines the relationship that copies data between a source volume and a target volume. FlashCopy mappings can be either stand-alone or a member of a Consistency Group. You can perform the actions of preparing, starting, or stopping FlashCopy on either a stand-alone mapping or a Consistency Group.

A FlashCopy mapping has an attribute that represents the state of the mapping. The FlashCopy states are the following:

Idle_or_copied

Read and write caching is enabled for both the source and the target. A FlashCopy mapping exists between the source and target, but the source and target behave as independent volumes in this state.

Copying

The FlashCopy indirection layer (see “Indirection layer” on page 256) governs all I/O to the source and target volumes while the background copy is running. The background copy process is copying *grains* from the source to the target. Reads and writes are executed on the target as though the contents of the source were instantaneously copied to the target during the **startfcmaporstartfcconsistgrp** command. The source and target can be independently updated. Internally, the target depends on the source for certain tracks. Read and write caching is enabled on the source and the target.

Stopped

The FlashCopy was stopped either by a user command or by an I/O error. When a FlashCopy mapping is stopped, the integrity of the data on the target volume is lost. Therefore, while the FlashCopy mapping is in this state, the target volume is in the Offline state. To regain access to the target, the mapping must be started again (the previous point-in-time will be lost) or the FlashCopy mapping must be deleted. The source volume is accessible, and read and write caching is enabled for the source. In the Stopped state, a mapping can either be prepared again or deleted.

Stopping

The mapping is in the process of transferring data to a dependent mapping. The behavior of the target volume depends on whether the background copy process had completed while the mapping was in the Copying state. If the copy process had completed, the target volume remains online while the stopping copy process completes. If the copy process had not completed, data in the cache is discarded for the target volume. The target volume is taken offline, and the stopping copy process runs. After the data has been copied, a stop complete asynchronous event notification is issued. The mapping will move to the Idle/Copied state if the background copy has completed or to the Stopped state if the background copy has not completed. The source volume remains accessible for I/O.

Suspended

The FlashCopy was in the Copying or Stopping state when access to the metadata was lost. As a result, both the source and target volumes are offline and the background copy process has been halted. When the metadata becomes available again, the FlashCopy mapping will return to the Copying or Stopping state. Access to the source and target volumes will be restored, and the background copy or stopping process will resume. Unflushed data that was written to the source or target before the FlashCopy was suspended is pinned in cache until the FlashCopy mapping leaves the Suspended state.

Preparing

The FlashCopy is in the process of preparing the mapping. While in this state, data from cache is destaged to disk and a consistent copy of the source exists on disk. At this time, cache is operating in write-through mode and therefore writes to the source volume will experience additional latency. The target volume is reported as online, but it will not perform reads or writes. These reads and writes are failed by the SCSI front end. Before starting the FlashCopy mapping, it is important that any cache at the host level, for example, buffers on the host operating system or application, are also instructed to flush any outstanding writes to the source volume. Performing the cache flush that is required as part of the **startfcmap** or **startfcconsistgrp** command causes I/Os to be delayed waiting for the cache flush to complete. To overcome this problem, FlashCopy supports the **prestartfcmap** or **prestartfcconsistgrp** commands. These commands prepare for a FlashCopy start while still allowing I/Os to continue to the source volume.

In the Preparing state, the FlashCopy mapping is prepared by the following steps:

1. Flush any modified write data associated with the source volume from the cache. Read data for the source is left in the cache.
2. Place the cache for the source volume into write-through mode so that subsequent writes wait until data is written to disk before completing the write command that is received from the host.
3. Discard any read or write data that is associated with the target volume from the cache.

Prepared

While in the Prepared state, the FlashCopy mapping is ready to perform a start. While the FlashCopy mapping is in this state, the target volume is in the Offline state. In the Prepared state, writes to the source volume experience additional latency, because the cache is operating in write-through mode.

Figure 6-6 represent the FlashCopy mapping state diagram. It illustrates the states in which a mapping can exist, and which events are responsible for a state change.

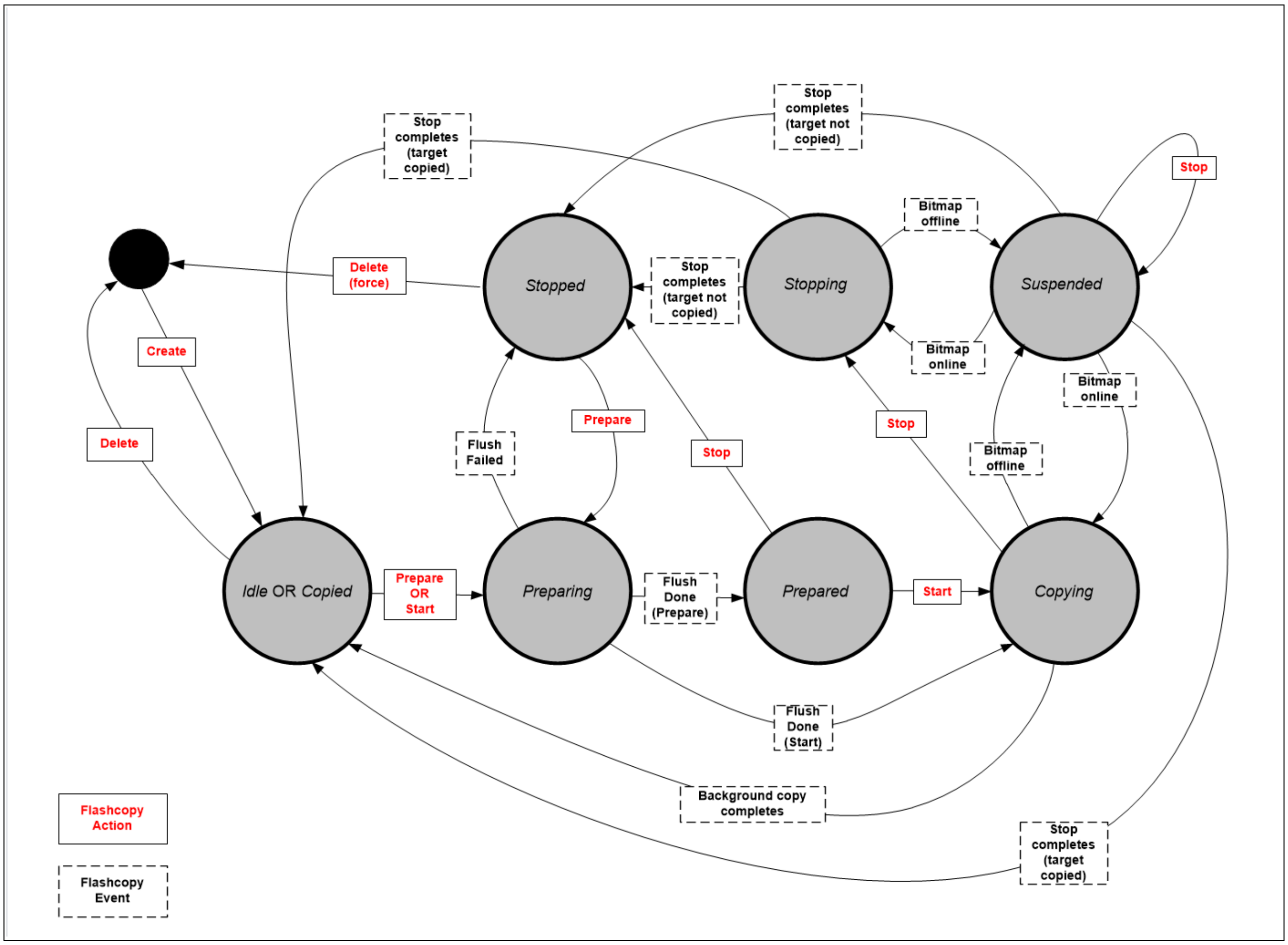


Figure 6-6 FlashCopy mapping states diagram

FlashCopy bitmaps and grains

A *bitmap* is an internal data structure stored in a particular I/O Group that is used to track which data in FlashCopy mappings has been copied from the source volume to the target volume. *Grains* are units of data grouped together to optimize the use of the bitmap. One bit in each bitmap represents the state of one grain. FlashCopy grain can be either 64 KB or 256 KB.

A FlashCopy bitmap takes up the bitmap space in the memory of the I/O group that must be shared with bitmaps of other features (such as Remote Copy bitmaps, volume mirroring bitmaps, and RAID bitmaps).

Indirection layer

The *FlashCopy indirection layer* governs the I/O to the source and target volumes when a FlashCopy mapping is started. This process is done by using a FlashCopy bitmap. The purpose of the FlashCopy indirection layer is to enable both the source and target volumes for read and write I/O immediately after FlashCopy starts.

The following description illustrates how the FlashCopy indirection layer works when a FlashCopy mapping is prepared and then started.

When a FlashCopy mapping is prepared and started, the following sequence is applied:

1. Flush the write cache to the source volume or volumes that are part of a Consistency Group.
2. Put the cache into write-through mode on the source volumes.
3. Discard the cache for the target volumes.
4. Establish a sync point on all of the source volumes in the Consistency Group (creating the FlashCopy bitmap).
5. Ensure that the indirection layer governs all of the I/O to the source volumes and target.
6. Enable the cache on source volumes and target volumes.

FlashCopy provides the semantics of a point-in-time copy that uses the indirection layer, which intercepts I/O that is directed at either the source or target volumes. The act of starting a FlashCopy mapping causes this indirection layer to become active in the I/O path, which occurs automatically across all FlashCopy mappings in the Consistency Group. The indirection layer then determines how each of the I/O is to be routed based on the following factors:

- The volume and the logical block address (LBA) to which the I/O is addressed
- Its direction (read or write)

The indirection layer allows the I/O to go through the underlying volume preserving the point-in-time copy. In order to do that, the Spectrum Virtualize code uses two mechanisms:

- Copy-on-Write (CoW). With this mechanism, when a write operation occurs in the source volume, a portion of data (grain) containing the data to be modified is copied to the target volume before the operation completion.
- Redirect-on-Write (RoW). With this mechanism, when a write operation occurs in the source volume, the data to be modified is written in another area leaving the original data unmodified to be used by the target volume.

Spectrum Virtualize implements CoW and RoW logics transparently to the user with the aim to optimize the performance and capacity. By using the RoW mechanism, the performance can improve by reducing the number of physical IOs for the write operations, while a significant capacity-saving can be achieved by improving the overall deduplication ratio.

The RoW was introduced with IBM Spectrum Virtualize version 8.4 and is used in the following conditions:

- Source and target volumes in the same pool.
- Source and target volumes in the same IO group.
- The pool that contains the source and target volumes must be a DRP.
- Source and target volumes do not participate in a volume mirroring relationship.
- Source and target volumes are not fully allocated.

In all the cases in which the RoW in not applicable, the CoW is used.

Table 6-2 lists the indirection layer algorithm in case of CoW.

Table 6-2 Summary table of the FlashCopy indirection layer algorithm

Volume being accessed	Has the grain been copied?	Host I/O operation	
		Read	**Write**
Source	No	Read from the source volume.	Copy grain to the most recently started target for this source, then write to the source.
	Yes	Read from the source volume.	Write to the source volume.
Target	No	If any newer targets exist for this source in which this grain has already been copied, read from the oldest of these targets. Otherwise, read from the source.	Hold the write. Check the dependency target volumes to see whether the grain has been copied. If the grain is not already copied to the next oldest target for this source, copy the grain to the next oldest target. Then, write to the target.
	Yes	Read from the target volume.	Write to the target volume.

Interaction with cache

The Spectrum Virtualize technology provides a two-layer cache, as follows:

- *Upper cache* serves mostly as write cache and hides the write latency from the hosts and application.
- *Lower cache* is a read/write cache and optimizes I/O to and from disks.

Figure 6-7 shows the IBM Spectrum Virtualize cache architecture.

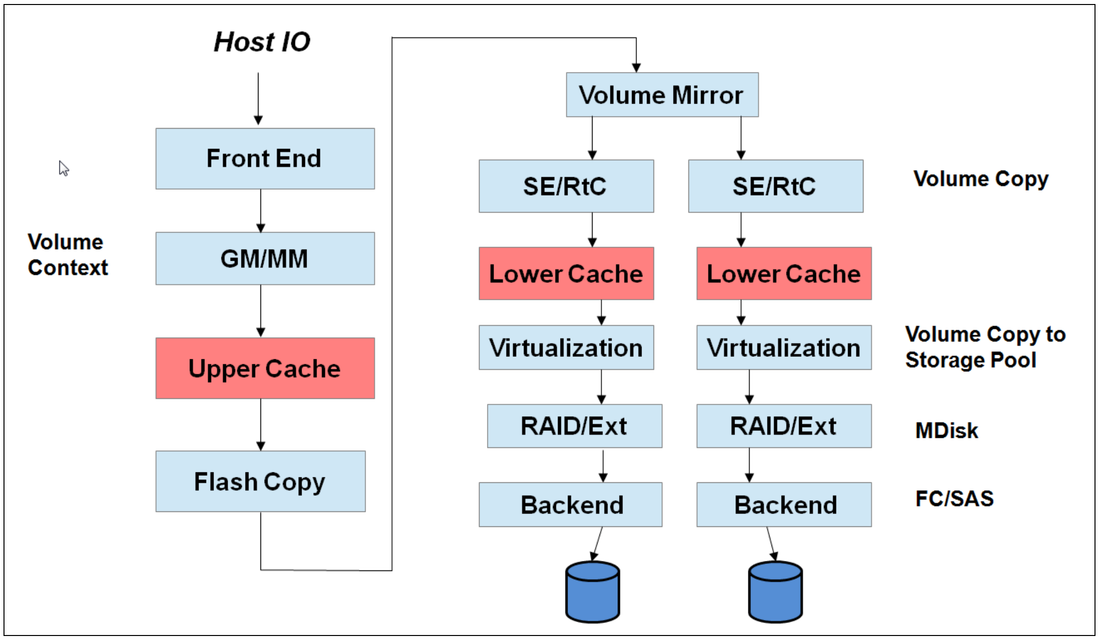


Figure 6-7 New cache architecture

The CoW process might introduce significant latency into write operations. To isolate the active application from this additional latency, the FlashCopy indirection layer is placed logically between the upper and lower cache. Therefore, the additional latency that is introduced by the CoW process is encountered only by the internal cache operations, and not by the application.

The logical placement of the FlashCopy indirection layer is shown in Figure 6-8.

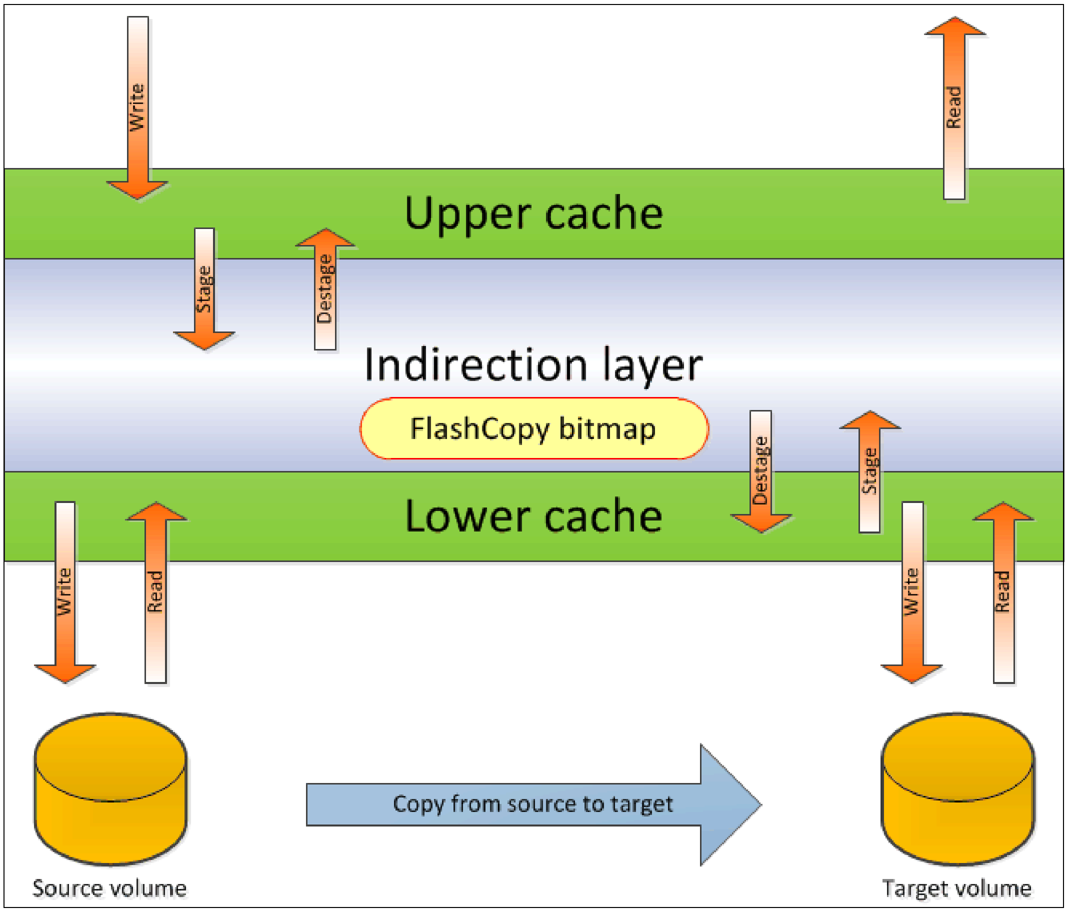


Figure 6-8 Logical placement of the FlashCopy indirection layer

The two-level cache architecture provides performance benefits to the FlashCopy mechanism. Because the FlashCopy layer is above the lower cache in the IBM Spectrum Virtualize software stack, it can benefit from read prefetching and coalescing writes to back-end storage.

Also, preparing FlashCopy is fast because upper cache write data does not have to go directly to back-end storage, but to the lower cache layer only.

Interaction and dependency between Multiple Target FlashCopy mappings

Figure 6-9 on page 260 shows a set of three FlashCopy mappings that share a common source. The FlashCopy mappings target volumes Target 1, Target 2, and Target 3.

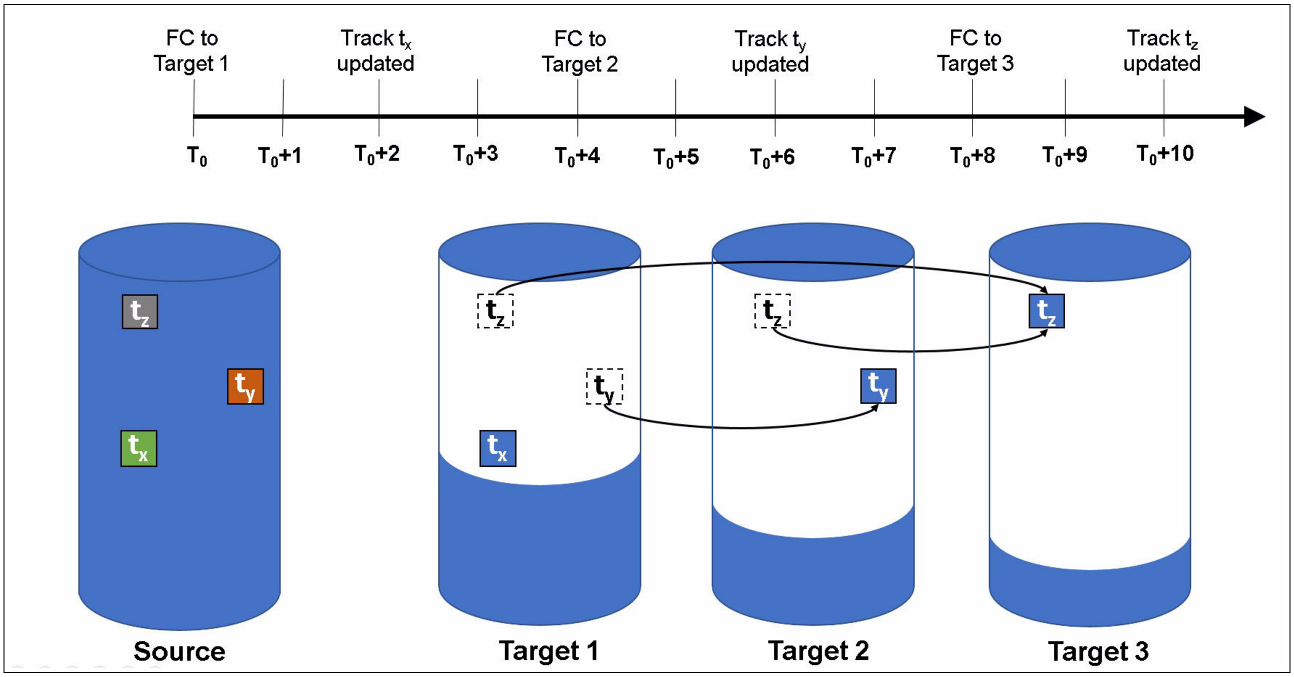


Figure 6-9 Interaction between Multiple Target FlashCopy mappings

Consider the following events timeline:

- At time T_0 a FlashCopy mapping is started between the source and the Target 1.
- At time T_0+2 the track t_x is updated in the source. Since this track has not yet been copied in background on Target 1, the copy-on-write process copies this track to the Target 1 before being updated on the source.
- At time T_0+4 a FlashCopy mapping is started between the source and the Target 2.
- At time T_0+6 the track t_y is updated in the source. Because this track has not yet been copied in background on Target 2, the copy-on-write process copies this track to the Target 2 only before being updated on the source.
- At time T_0+8 a FlashCopy mapping is started between the source and the Target 3.
- At time T_0+10 the track t_z is updated in the source. Because this track has not yet been copied in background on Target 3, the copy-on-write process copies this track to the Target 3 only before being updated on the source.

As a result of this sequence of events, the configuration in Figure 6-9 has the following characteristics:

- Target 1 is dependent upon Target 2 and Target 3. It remains dependent until all of Target 1 has been copied. No target depends on Target 1, so the mapping can be stopped without need to copy any data to maintain the consistency in the other targets.
- Target 2 depends on Target 3, and will remain dependent until all of Target 2 has been copied. Target 1 depends on Target 2, so if this mapping is stopped, the cleanup process is started to copy all data that is uniquely held on this mapping (that is t_y) to Target 1.
- Target 3 is not dependent on any target, but it has Target 1 and Target 2 depending on it, so if this mapping is stopped the cleanup process is started to copy all data that is uniquely held on this mapping (that is t_z) to Target 2.

Target writes with Multiple Target FlashCopy

A write to an intermediate or newest target volume must consider the state of the grain within its own mapping, and the state of the grain of the next oldest mapping:

- If the grain of the next oldest mapping has not been copied yet, it must be copied before the write is allowed to proceed to preserve the contents of the next oldest mapping. The data that is written to the next oldest mapping comes from a target or source.
- If the grain in the target being written has not yet been copied, the grain is copied from the oldest already copied grain in the mappings that are newer than the target, or the source if none are already copied. After this copy is done, the write can be applied to the target.

Target reads with Multiple Target FlashCopy

If the grain being read has already been copied from the source to the target, the read simply returns data from the target being read. If the grain has not been copied, each of the newer mappings is examined in turn and the read is performed from the first copy found. If none are found, the read is performed from the source.

6.2.4 FlashCopy planning considerations

The FlashCopy function, like all the advanced IBM FlashSystem products features, offers useful capabilities. However, some basic planning considerations are to be followed for a successful implementation.

FlashCopy configurations limits

To plan for and implement FlashCopy, you must check the configuration limits and adhere to them. Table 6-3 lists the system limits that apply to the latest version as of this writing.

Table 6-3 FlashCopy properties and maximum configurations

FlashCopy property	Maximum	Comment
FlashCopy targets per source	256	This maximum is the maximum number of FlashCopy mappings that can exist with the same source volume.
FlashCopy mappings per system	15864	This maximum is the maximum number of FlashCopy mappings per system.
FlashCopy Consistency Groups per system	500	This maximum is an arbitrary limit that is policed by the software.
FlashCopy volume space per I/O Group	4096 TB	This maximum is a limit on the quantity of FlashCopy mappings by using bitmap space from one I/O Group.
FlashCopy mappings per Consistency Group	512	This limit is due to the time that is taken to prepare a Consistency Group with many mappings.

Configuration limits: The configuration limits always change with the introduction of new hardware and software capabilities. For more information about the latest configuration limits, see this IBM Support web page.

The total amount of cache memory reserved for the FlashCopy bitmaps limits the amount of capacity that can be used as a FlashCopy target. Table 6-4 shows the relationship of bitmap space to FlashCopy address space, depending on the size of the grain and the kind of FlashCopy service being used.

Table 6-4 Relationship of bitmap space to FlashCopy address space for the specified I/O Group

Copy service	Grain size (KB)	1 MB of memory provides the following volume capacity for the specified I/O Group
FlashCopy	256	2 TB of target volume capacity
FlashCopy	64	512 GB of target volume capacity
Incremental FlashCopy	256	1 TB of target volume capacity
Incremental FlashCopy	64	256 GB of target volume capacity

Mapping consideration: For multiple FlashCopy targets, you must consider the number of mappings. For example, for a mapping with a 256 KB grain size, 8 KB of memory allows one mapping between a 16 GB source volume and a 16 GB target volume. Alternatively, for a mapping with a 256 KB grain size, 8 KB of memory allows two mappings between one 8 GB source volume and two 8 GB target volumes.

When you create a FlashCopy mapping, if you specify an I/O Group other than the I/O Group of the source volume, the memory accounting goes towards the specified I/O Group, not towards the I/O Group of the source volume.

The default amount of memory for FlashCopy is 20 MB. This value can be increased or decreased by using the `chiogrp` command or through the GUI. The maximum amount of memory that can be specified for FlashCopy is 2048 MB (512 MB for 32-bit systems). The maximum combined amount of memory across all copy services features is 2600 MB (552 MB for 32-bit systems).

Bitmap allocation: When creating a FlashCopy mapping, you can optionally specify the I/O group where the bitmap is allocated. If you specify an I/O Group other than the I/O Group of the source volume, the memory accounting goes towards the specified I/O Group, not towards the I/O Group of the source volume. This option can be useful when an I/O group is exhausting the memory that is allocated to the FlashCopy bitmaps and no more free memory is available in the I/O group.

FlashCopy general restrictions

The following implementation restrictions apply to FlashCopy:

- The size of source and target volumes must be the same when creating a FlashCopy mapping.
- Multiple FlashCopy mappings that use the same target volume can be defined, but only one of these mappings can be started at a time. This limitation means that no multiple FlashCopy can be active to the same target volume.

- The following restrictions apply when expanding or shrinking volumes that are defined in a FlashCopy mapping:
  - Target volumes cannot be shrunk
  - Source volume can be shrunk, but only to the largest starting size of a target volume (in a multiple target or cascading mappings) when in copying or stopping state.
  - Source and target volumes must be same size when the mapping is prepared or started.
  - Source and target volumes can be expanded in any order except in case of incremental FlashCopy where the target volume must be expanded before the source volume can be expanded.

  > **Note:** Expanding or shrinking volumes that are participating in a FlashCopy map is allowed with code level 8.4.2 or later.

- In a cascading FlashCopy, the grain size of all the FlashCopy mappings that participate must be the same.
- In a multi-target FlashCopy, the grain size of all the FlashCopy mappings that participate must be the same.
- In a reverse FlashCopy, the grain size of all the FlashCopy mappings that participate must be the same.
- No FlashCopy mapping can be added to a consistency group while the FlashCopy mapping status is `Copying`.
- No FlashCopy mapping can be added to a consistency group while the consistency group status is `Copying`.
- The use of Consistency Groups is restricted when using Cascading FlashCopy. A Consistency Group serves the purpose of starting FlashCopy mappings at the same point in time. Within the *same* Consistency Group, it is not possible to have mappings with these conditions:
  - The source volume of one mapping is the target of another mapping.
  - The target volume of one mapping is the source volume for another mapping.

  These combinations are not useful because within a Consistency Group, mappings cannot be established in a certain order. This limitation renders the content of the target volume undefined. For instance, it is not possible to determine whether the first mapping was established before the target volume of the first mapping that acts as a source volume for the second mapping.

  Even if it were possible to ensure the order in which the mappings are established within a Consistency Group, the result is equal to Multi Target FlashCopy (two volumes holding the same target data for one source volume). In other words, a cascade is useful for copying volumes in a certain order (and copying the changed content targets of FlashCopies), rather than at the same time in an undefined order (from within one single Consistency Group).
- Source and target volumes can be used as primary in a Remote Copy relationship. For more information about the FlashCopy and the Remote Copy possible interactions, see “Interaction between Remote Copy and FlashCopy” on page 304.

FlashCopy presets

The IBM FlashSystem GUI interface provides three FlashCopy presets (Snapshot, Clone, and Backup) to simplify the more common FlashCopy operations. Figure 6-10 shows the preset selection window in the GUI.

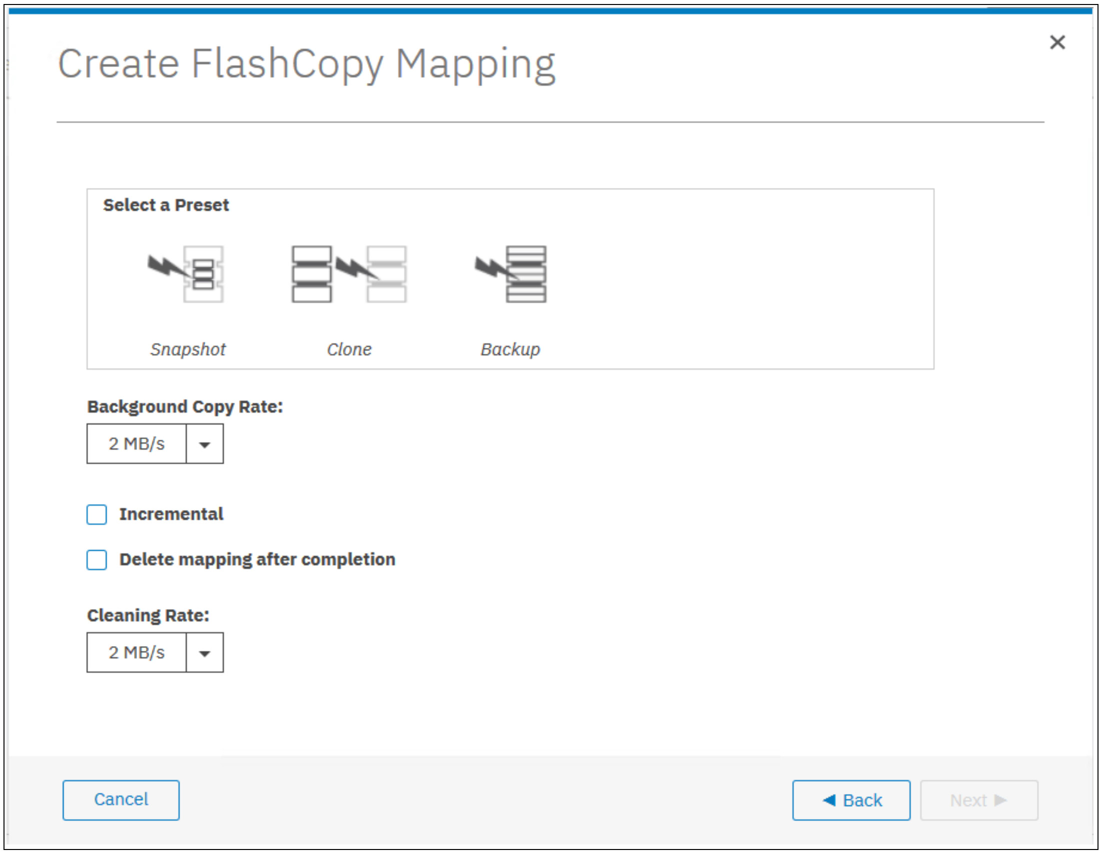


Figure 6-10 GUI FlashCopy Presets

Although these presets meet most FlashCopy requirements, they do not provide support for all possible FlashCopy options. If more specialized options are required that are not supported by the presets, the options must be performed by using CLI commands.

This section describes the three preset options and their use cases.

Snapshot

This preset creates a copy-on-write point-in-time copy. The snapshot is not intended to be an independent copy. Instead, the copy is used to maintain a view of the production data at the time that the snapshot is created. Therefore, the snapshot holds only the data from regions of the production volume that have changed since the snapshot was created. Because the snapshot preset uses thin provisioning, only the capacity that is required for the changes is used.

Snapshot uses the following preset parameters:

- ► Background copy: `None`
- ► Incremental: `No`
- ► Delete after completion: `No`
- ► Cleaning rate: `No`
- ► Primary copy source pool: `Target pool`

A typical use case for the Snapshot is when the user wants to produce a copy of a volume without affecting the availability of the volume. The user does not anticipate many changes to be made to the source or target volume. A significant proportion of the volumes remains unchanged.

By ensuring that only changes require a copy of data to be made, the total amount of disk space that is required for the copy is reduced. Therefore, many Snapshot copies can be used in the environment.

Snapshots are useful for providing protection against corruption or similar issues with the validity of the data. However, they do not provide protection from physical controller failures. Snapshots can also provide a vehicle for performing repeatable testing (including "what-if" modeling that is based on production data) without requiring a full copy of the data to be provisioned.

Clone

The clone preset creates a replica of the volume, which can then be changed without affecting the original volume. After the copy completes, the mapping that was created by the preset is automatically deleted.

Clone uses the following preset parameters:

- Background copy rate: `50`
- Incremental: `No`
- Delete after completion: `Yes`
- Cleaning rate: `50`
- Primary copy source pool: `Target pool`

A typical use case for the Snapshot is when users want a copy of the volume that they can modify without affecting the original volume. After the clone is established, there is no expectation that it is refreshed or that there is any further need to reference the original production data again. If the source is thin-provisioned, the target is thin-provisioned for the auto-create target.

Backup

The backup preset creates a point-in-time replica of the production data. After the copy completes, the backup view can be refreshed from the production data, with minimal copying of data from the production volume to the backup volume.

Backup uses the following preset parameters:

- Background Copy rate: `50`
- Incremental: `Yes`
- Delete after completion: `No`
- Cleaning rate: `50`
- Primary copy source pool: `Target pool`

The Backup preset can be used when the user wants to create a copy of the volume that can be used as a backup if the source becomes unavailable. This unavailability can happen during loss of the underlying physical controller. The user plans to periodically update the secondary copy, and does not want to suffer from the resource demands of creating a new copy each time.

Incremental FlashCopy times are faster than full copy, which helps to reduce the window where the new backup is not yet fully effective. If the source is thin-provisioned, the target is also thin-provisioned in this option for the auto-create target.

Another use case, which is not supported by the name, is to create and maintain (periodically refresh) an independent image. This image can be subjected to intensive I/O (for example, data mining) without affecting the source volume's performance.

Thin provisioning considerations

When creating FlashCopy with thin provisioned target volumes, the no-copy option often is used. The real size of a thin provisioned volume is an attribute that defines how much physical capacity is reserved for the volume. The real size can vary 0 - 100% of the virtual capacity.

When thin provisioned volumes are used as FlashCopy targets, it is important to provide a non-zero real size. This size is required because when the FlashCopy is started, the copy-on-write process requires to allocate capacity on the target volumes. If some capacity is not yet allocated, the write IO can be delayed until the capacity is made available (as with thin provisioned volumes with zero real size). Often, the write caching hides this effect, but in the case of heavy write workloads, the performance could be affected.

Sizing consideration

When Thin Provisioned FlashCopy is used, an estimation of the physical capacity consumption is required. Consider that while a FlashCopy is active, the thin provisioned target volume allocates physical capacity whenever a grain is modified for the first time on source or target volume.

The following factors must be considered that so that an accurate sizing can be completed:

- The FlashCopy duration in terms of seconds (D).
- The write operation per second (W).
- The grain size in terms of KB (G).
- The rewrite factor. This factor represents the average chance that a write operation reoccurs in the same grain (R) in percentage

While the first three factors are easy to assess, the rewrite factor can be only roughly estimated as it is dependent on the workload type and the FlashCopy duration. The used capacity (CC) of a thin provisioned target volume of C size while the FlashCopy is active can be estimated by using the following equation:

```
CC = min{(W - W x R) x G x D,C}
```

For example, consider a 100 GB volume that has a FlashCopy active for 3 hours (10.800 seconds) with a grain size of 64 K. Consider also a write workload of 100 IOPS with a rewrite factor of 85% (85% of writes occur on the same grains). In this case, the estimation of the used capacity is:

```
CC = (100 - 85) x 64 x 10.800 = 10.368.000 KB = 9,88 GB
```

Important: Consider the following points:

- The recommendation with thin provisioned target volumes is to assign at least 2 GB of real capacity.
- Thin provisioned FlashCopy can greatly benefit from the Redirect-on-Write capability that was introduced with Spectrum Virtualize version 8.4. For more information, see "Indirection layer" on page 256.

Grain size considerations

When creating a mapping a grain size of 64 KB can be specified as compared to the default 256 KB. This smaller grain size has been introduced specifically for the incremental FlashCopy, even though its use is not restricted to the incremental mappings.

In an incremental FlashCopy, the modified data is identified by using the bitmaps. The amount of data to be copied when refreshing the mapping depends on the grain size. If the grain size is 64 KB, as compared to 256 KB, there might be less data to copy to get a fully independent copy of the source again.

The following are the preferred settings for thin-provisioned FlashCopy:

- Thin-provisioned volume grain size should be equal to the FlashCopy grain size. Anyway if the 256 KB thin-provisioned volume grain size is chosen, it is still beneficial to limit the FlashCopy grain size to 64 KB. It is possible to minimize the performance impact to the source volume, even though this size increases the I/O workload on the target volume.
- Thin-provisioned volume grain size must be 64 KB for the best performance and the best space efficiency.

The exception is where the thin target volume is going to become a production volume (and is likely to be subjected to ongoing heavy I/O). In this case, the 256 KB thin-provisioned grain size is preferable because it provides better long-term I/O performance at the expense of a slower initial copy.

FlashCopy limitation: Configurations with large numbers of FlashCopy/Remote Copy relationships might be forced to choose a 256 KB grain size for FlashCopy to avoid constraints on the amount of bitmap memory.

Cascading FlashCopy and Multi Target FlashCopy require that all the mappings that are participating in the FlashCopy chain feature the same grain size. For more information, see “FlashCopy general restrictions” on page 262.

Volume placement considerations

The source and target volumes placement among the pools and the I/O groups must be planned to minimize the effect of the underlying FlashCopy processes. In normal condition (that is with all the canisters fully operative), the FlashCopy background copy workload distribution follows this schema:

- The preferred node of the *source* volume is responsible for the background copy *read* operations.
- The preferred node of the *target* volume is responsible for the background copy *write* operations.

Table 6-5 shows how the back-end I/O operations are distributed across the nodes.

Table 6-5 Workload distribution for back-end I/O operations

	Read from source	**Read from target**	**Write to source**	**Write to target**
Node that performs the back-end I/O if the grain is copied	Preferred node in source volume's I/O group	Preferred node in target volume's I/O group	Preferred node in source volume's I/O group	Preferred node in target volume's I/O group
Node that performs the back-end I/O if the grain is not yet copied	Preferred node in source volume's I/O group	Preferred node in source volume's I/O group	The preferred node in source volume's I/O group will read and write, and the preferred node in target volume's I/O group will write	The preferred node in source volume's I/O group will read, and the preferred node in target volume's I/O group will write

The data transfer among the source and the target volume's preferred nodes occurs through the node-to-node connectivity. Consider the following volume placement alternatives:

- Source and target volumes use the same preferred node.

  In this scenario, the node that is acting as preferred for source and target volume manages all the read and write FlashCopy operations. Only resources from this node are used for the FlashCopy operations, and no node-to-node bandwidth is used.

- Source and target volumes use the different preferred node.

  In this scenario, both nodes that are acting as preferred nodes manage read and write FlashCopy operations according to the previously described scenarios. The data that is transferred between the two preferred nodes goes through the node-to-node network.

Both alternatives described have advantages and disadvantages, but in general option 1 (source and target volumes use the same preferred node) is preferred. Consider the following exceptions:

- A clustered IBM FlashSystem system with multiple I/O groups in HyperSwap, where the source volumes are evenly spread across all the nodes.

  In this case the preferred node placement should follow the location of the source and target volumes on the back-end storage. For example, if the source volume is on site A and the target volume is on site B, then the target volumes preferred node must be in site B. Placing the target volumes preferred node in site A will cause the re-direction of the FlashCopy write operation through the node-to-node network.

- A clustered IBM FlashSystem system with multiple control enclosures, where the source volumes are evenly spread across all the canisters.

  In this case the preferred node placement should follow the location of source and target volumes on the internal storage. For example, if the source volume is on the internal storage attached to control enclosure A and the target volume is on internal storage attached to control enclosure B, then the target volumes preferred node must be in one canister of control enclosure B. Placing the target volumes preferred node on control enclosure A will cause the re-direction of the FlashCopy write operation through the node-to-node network.

Placement on the back-end storage is mainly driven by the availability requirements. Generally, use different back-end storage controllers or arrays for the source and target volumes.

DRP optimized snapshots: To use the Redirect-on-Write capability that was introduced with Spectrum Virtualize version 8.4, check the volume placement restrictions that are described in "Indirection layer" on page 256.

Background copy considerations

The background copy process uses internal resources, such as CPU, memory, and bandwidth. This copy process tries to reach the target copy data rate for every volume according to the background copy rate parameter setting (see Table 6-1 on page 248).

If the copy process is unable to achieve these goals, it starts contending resources to the foreground I/O (that is the I/O coming from the hosts). As result, both background copy and foreground I/O will tend to see an increase in latency and therefore reduction in throughput compared to the situation when the bandwidth not been limited. Degradation is graceful. Both background copy and foreground I/O continue to make progress, and will not stop, hang, or cause the node to fail.

To avoid any impact on the foreground I/O, that is in the hosts response time, carefully plan the background copy activity, taking in account the overall workload running in the systems. The background copy basically reads and writes data to managed disks. Usually, the most affected component is the back-end storage. CPU and memory are not normally significantly affected by the copy activity.

The theoretical added workload due to the background copy is easily estimable. For instance, starting 20 FlashCopy with a background copy rate of 70 each adds a maximum throughput of 160 MBps for the reads and 160 MBps for the writes.

The source and target volumes distribution on the back-end storage determines where this workload is going to be added. The duration of the background copy depends on the amount of data to be copied. This amount is the total size of volumes for full background copy or the amount of data that is modified for incremental copy refresh.

Performance monitoring tools like IBM Spectrum Control can be used to evaluate the existing workload on the back-end storage in a specific time window. By adding this workload to the foreseen background copy workload, you can estimate the overall workload running toward the back-end storage. Disk performance simulation tools, like Disk Magic or StorM, can be used to estimate the effect, if any, of the added back-end workload to the host service time during the background copy window. The outcomes of this analysis can provide useful hints for the background copy rate settings.

When performance monitoring and simulation tools are not available, use a conservative and progressive approach. Consider that the background copy setting can be modified at any time, even when the FlashCopy is already started. The background copy process can even be completely stopped by setting the background copy rate to `0`.

Initially set the background copy rate value to add a limited workload to the back-end (for example less than 100 MBps). If no effects on hosts are noticed, the background copy rate value can be increased. Do this process until you see negative effects. Note that the background copy rate setting follows an exponential scale, so changing, for instance, from 50 to 60 doubles the data rate goal from 2 MBps to 4 MBps.

Cleaning process and Cleaning Rate

The Cleaning Rate is the rate at which the data is copied among dependent FlashCopies, such as Cascaded and Multi Target FlashCopy. The Cleaning process aims to release the dependency of a mapping in such a way that it can be stopped immediately (without going to the `stopping` state). The typical use case for setting the Cleaning Rate is when it is required to stop a Cascaded or Multi Target FlashCopy that is not the oldest in the FlashCopy chain. In this case to avoid the stopping state lasting for a long time, the cleaning rate can be adjusted accordingly.

An interaction occurs between the background copy rate and the Cleaning Rate settings:

- Background copy = 0 and Cleaning Rate = 0

  No background copy or cleaning take place. When the mapping is stopped, it goes into stopping state and a cleaning process starts with the default cleaning rate, which is 50 or 2 MBps.

- Background copy > 0 and Cleaning Rate = 0

  The background copy takes place at the background copy rate but no cleaning process is started. When the mapping is stopped, it goes into stopping state and a cleaning process starts with the default cleaning rate (50 or 2 MBps).

- Background copy = 0 and Cleaning Rate > 0

  No background copy takes place, but the cleaning process runs at the cleaning rate. When the mapping is stopped, the cleaning completes (if not yet completed) at the cleaning rate.

- Background copy > 0 and Cleaning Rate > 0

  The background copy takes place at the background copy rate but no cleaning process is started. When the mapping is stopped, it goes into stopping state and a cleaning process starts with the specified cleaning rate.

Regarding the workload considerations for the cleaning process, the same guidelines as for background copy apply.

Host and application considerations to ensure FlashCopy integrity

Because FlashCopy is at the block level, it is necessary to understand the interaction between your application and the host operating system. From a logical standpoint, it is easiest to think of these objects as “layers” that sit on top of one another. The application is the topmost layer, and beneath it is the operating system layer.

Both of these layers have various levels and methods of caching data to provide better speed. Because IBM FlashSystem systems, and therefore FlashCopy, sit below these layers, they are unaware of the cache at the application or operating system layers.

To ensure the integrity of the copy that is made, it is necessary to flush the host operating system and application cache for any outstanding reads or writes before the FlashCopy operation is performed. Failing to flush the host operating system and application cache produces what is referred to as a *crash consistent* copy.

The resulting copy requires the same type of recovery procedure, such as log replay and file system checks, that is required following a host crash. FlashCopies that are crash consistent often can be used following file system and application recovery procedures.

> **Note:** Although the best way to perform FlashCopy is to flush host cache first, some companies, such as Oracle, support using snapshots without it, as stated in Metalink note 604683.1.

Various operating systems and applications provide facilities to stop I/O operations and ensure that all data is flushed from host cache. If these facilities are available, they can be used to prepare for a FlashCopy operation. When this type of facility is not available, the host cache must be flushed manually by quiescing the application and unmounting the file system or drives.

> **Preferred practice:** From a practical standpoint, when you have an application that is backed by a database and you want to make a FlashCopy of that application's data, it is sufficient in most cases to use the write-suspend method that is available in most modern databases. You can use this method because the database maintains strict control over I/O.
>
> This method is as opposed to flushing data from both the application and the backing database, which is always the suggested method because it is safer. However, this method can be used when facilities do not exist or your environment includes time sensitivity.

6.3 Remote Copy services

IBM FlashSystem technology offers various Remote Copy services functions that address Disaster Recovery and Business Continuity needs.

Metro Mirror is designed for metropolitan distances with a zero recovery point objective (RPO), which is zero data loss. This objective is achieved with a synchronous copy of volumes. Writes are not acknowledged until they are committed to both storage systems. By definition, any vendors' synchronous replication makes the host wait for write I/Os to complete at both the local and remote storage systems, and includes round-trip network latencies. Metro Mirror has the following characteristics:

- Zero RPO
- Synchronous
- Production application performance that is affected by round-trip latency

Global Mirror technologies are designed to minimize the network latency effects by replicating asynchronously. Spectrum Virtualize provides two types of asynchronous mirroring technology:

- The standard Global Mirror (referred to as Global Mirror)
- The Global Mirror with Change Volume (GMCV)

With the Global Mirror, writes are acknowledged as soon as they can be committed to the local storage system, sequence-tagged, and passed on to the replication network. This technique allows Global Mirror to be used over longer distances. By definition, any vendors' asynchronous replication results in an RPO greater than zero. However, for Global Mirror, the RPO is quite small, typically anywhere from several milliseconds to some number of seconds.

Although Global Mirror is asynchronous, the RPO is still small, and thus the network and the remote storage system must both still be able to cope with peaks in traffic. Global Mirror has the following characteristics:

- Near-zero RPO
- Asynchronous
- Production application performance that is affected by I/O sequencing preparation time

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

- Larger RPO
- Point-in-time copies
- Asynchronous
- Possible system performance effect because point-in-time copies are created locally

Successful implementation of Remote Copy depends on taking a holistic approach in which you consider all components and their associated properties. The components and properties include host application sensitivity, local and remote SAN configurations, local and remote system and storage configuration, and the inter-system network.

6.3.1 Remote Copy use cases

Data replication techniques are the foundations of Disaster Recovery and Business Continuity solutions. Besides these common use cases, Remote Copy technologies can be used in other data movement scenarios, as described in the following sections.

Storage systems renewal

Remote Copy functions can be used to facilitate the migration of data between storage systems while minimizing downtime for applications. By using remote copy, application data can be copied from an IBM Spectrum Virtualize-based system to another, while applications remain online. After the volumes are fully copied and synchronized, the application can be stopped and then immediately started on the new storage system.

Starting with IBM Spectrum Virtualize version 8.4.2, the Nondisruptive Volume Migration capability was introduced. This feature uses the Remote Copy capabilities to transparently move host volumes between IBM Spectrum Virtualize based systems.

For more information, see 5.8, “Volume migration” on page 223.

Data center moving

Remote Copy functions can be used to move data between Spectrum Virtualize-based systems to facilitate data centers moving operations. By using remote copy, application data can be copied from volumes in a source data center to volumes in another data center while applications remain online. After the volumes are fully copied and synchronized, the applications can be stopped and then immediately started in the target data center.

6.3.2 Remote Copy functional overview

This section presents the terminology and the basic functional aspects of the Remote Copy services.

Common terminology and definitions

When such a breadth of technology areas is covered, the same technology component can have multiple terms and definitions. This document uses the following definitions:

- *Local system* or *master system*

  The system on which the foreground applications run.

- *Local hosts*

  Hosts that run on the foreground applications.

- *Master volume* or *source volume*

  The local volume that is being mirrored. The volume has nonrestricted access. Mapped hosts can read and write to the volume.

- *inter-system link* or *inter-system network*

  The network that provides connectivity between the local and the remote site. It can be a Fibre Channel (FC) network (SAN), an IP network, or a combination of the two.

- *Remote system* or *auxiliary system*

  The system that holds the remote mirrored copy.

- *Auxiliary volume* or *target volume*

  The remote volume that holds the mirrored copy. It is read-access only.

- *Remote copy*

  A generic term that is used to describe a Metro Mirror or Global Mirror relationship in which data on the source volume is mirrored to an identical copy on a target volume. Often the two copies are separated by some distance, which is why the term *remote* is used to describe the copies. However, having remote copies is not a prerequisite. A Remote Copy relationship includes the following states:

  – Consistent relationship

    A Remote Copy relationship where the data set on the target volume represents a data set on the source volumes at a certain point.

  – Synchronized relationship

    A relationship is *synchronized* if it is consistent *and* the point that the target volume represents is the current point. The target volume contains identical data as the source volume.

- *Synchronous Remote Copy*

- Writes to the source and target volumes that are committed in the foreground before confirmation is sent about completion to the local host application. Metro Mirror is a synchronous Remote Copy type.

- *Asynchronous remote copy*

  A foreground write I/O is acknowledged as complete to the local host application before the mirrored foreground write I/O is cached at the remote system. Mirrored foreground writes are processed asynchronously at the remote system, but in way that a consistent copy is always present in the remote system. Global Mirror and GMCV are asynchronous Remote Copy types.

- The *background copy* process manages the initial synchronization or resynchronization processes between source volumes to target mirrored volumes on a remote system.
- *Foreground I/O* reads and writes I/O on a local SAN, which generates a mirrored foreground write I/O that is across the inter-system network and remote SAN.

Figure 6-11 shows some of the concepts of remote copy.

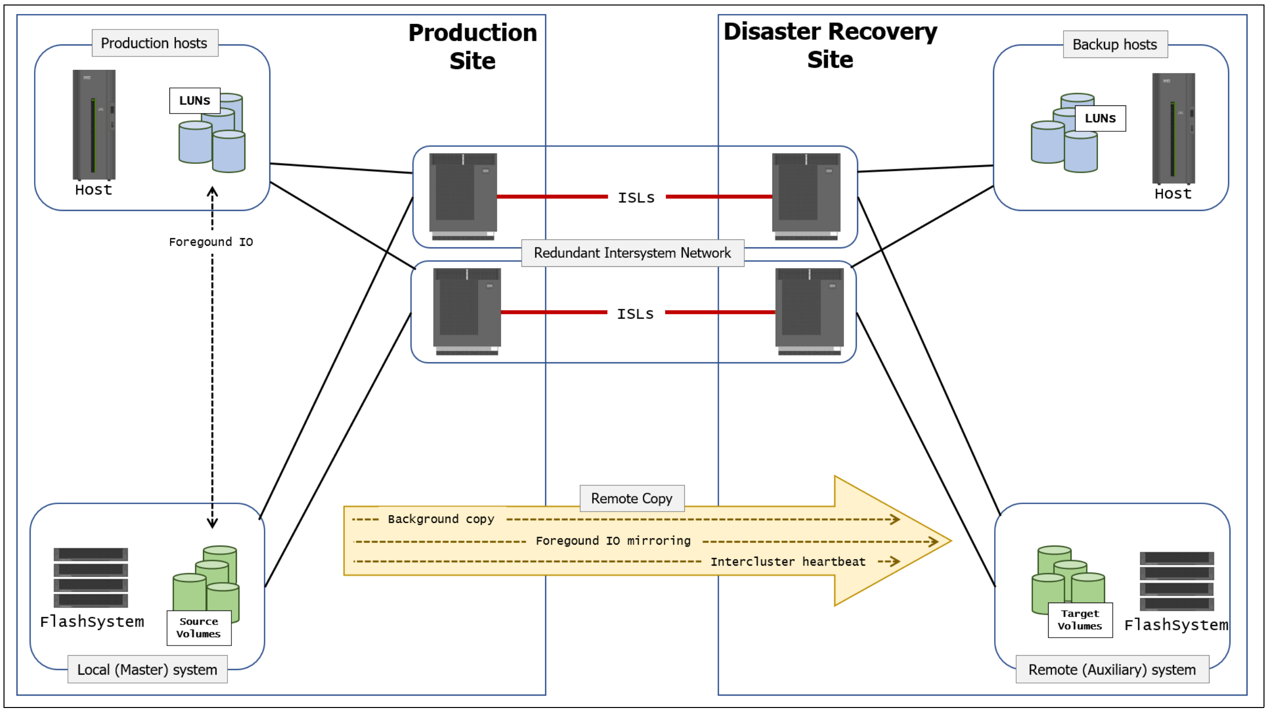


Figure 6-11 Remote Copy components and applications

A successful implementation of inter-system Remote Copy services significantly depends on the quality and configuration of the inter-system network.

Remote Copy partnerships and relationships

A Remote Copy *partnership* is a partnership that is established between a master (local) system and an auxiliary (remote) system, as shown in Figure 6-12.

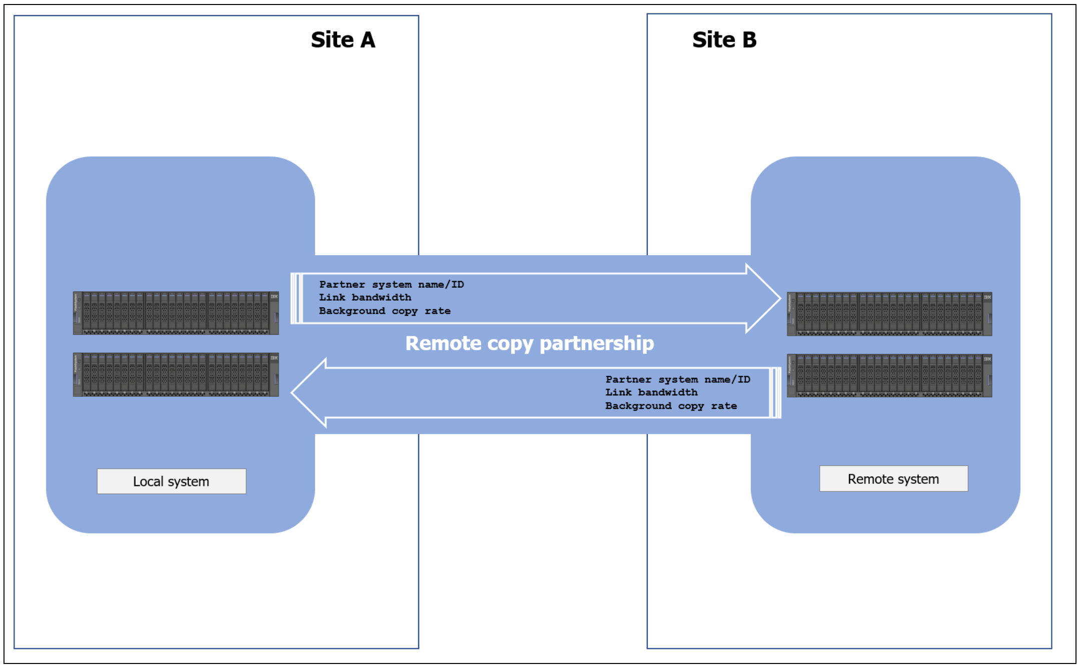


Figure 6-12 Remote Copy partnership

Partnerships are established between two systems by issuing the **mkfcpartnership** or **mkippartnership** command once from each end of the partnership. The following parameters must be specified:

- The remote system name (or ID).
- The link bandwidth (in Mbps).
- The background copy rate as a percentage of the link bandwidth.
- The background copy parameter that determines the maximum speed of the initial synchronization and resynchronization of the relationships.

> **Tip:** To establish a fully functional Metro Mirror or Global Mirror partnership, issue the **mkfcpartnership** or **mkippartnership** command from both systems.

In addition to the background copy rate setting, the initial synchronization can be adjusted at relationship level with the `relationship_bandwidth_limit` parameter. The `relationship_bandwidth_limit` is a system-wide parameter that sets the maximum bandwidth that can be used to initially synchronize a single relationship.

After background synchronization or resynchronization is complete, a Remote Copy relationship provides and maintains a consistent mirrored copy of a source volume to a target volume.

Copy directions and default roles

When a Remote Copy relationship is created, the source volume is assigned the role of the *master*, and the target volume is assigned the role of the *auxiliary*. This design implies that the initial copy direction of mirrored foreground writes and background resynchronization writes (if applicable) is from master to auxiliary. When a Remote Copy relationship is initially started, the master volume assumes the role of *primary* volume, while the auxiliary volume became *secondary* volumes.

After the initial synchronization is complete, you can change the copy direction (see Figure 6-13) by switching the roles of primary and secondary. The ability to change roles is used to facilitate disaster recovery.

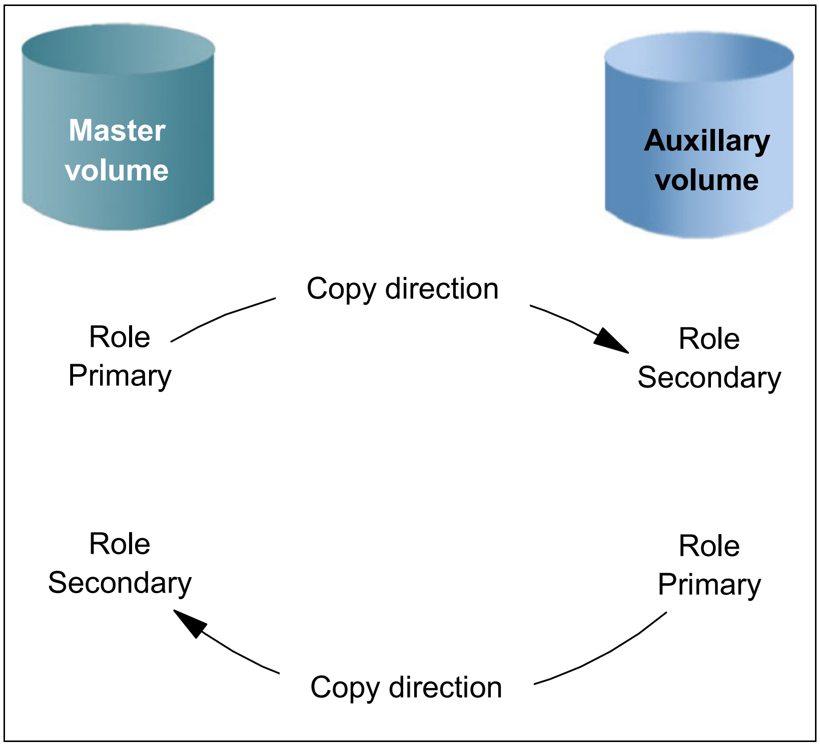


Figure 6-13 Role and direction changes

> **Attention:** When the direction of the relationship is changed, the roles primary/secondary of the volumes are altered. A consequence is that the read/write properties are also changed, meaning that the master volume takes on a secondary role and becomes read-only.

Consistency Groups

A Consistency Group (CG) is a collection of relationships that can be treated as one entity. This technique is used to preserve write order consistency across a group of volumes that pertain to one application, for example, a database volume and a database log file volume.

After a Remote Copy relationship is added into a Consistency Group, you cannot manage the relationship in isolation from the Consistency Group. So, for example, issuing a **`stoprcrelationship`** command on the stand-alone volume would fail because the system knows that the relationship is part of a Consistency Group.

Similarly to the Remote Copy relationships, also a Consistency Group, when created, assigns the role of *master* to the source storage system and *auxiliary* to the target storage system.

Consider the following points regarding Consistency Groups:

- Each volume relationship can belong to only one Consistency Group.
- Volume relationships can also be stand-alone; that is, not in any Consistency Group.
- Consistency Groups can also be created and left empty or can contain one or many relationships.
- You can create up to 256 Consistency Groups on a system.
- All volume relationships in a Consistency Group must have matching primary and secondary systems, but they do not need to share I/O groups.
- All relationships in a Consistency Group have the same copy direction and state.
- Each Consistency Group is either for Metro Mirror or for Global Mirror relationships, but not both. This choice is determined by the first volume relationship that is added to the Consistency Group.

Consistency Group consideration: A Consistency Group relationship does not have to be in a directly matching I/O group number at each site. A Consistency Group owned by I/O group 1 at the local site does not have to be owned by I/O group 1 at the remote site. If you have more than one I/O group at either site, you can create the relationship between any two I/O groups. This technique spreads the workload, for example, from local I/O group 1 to remote I/O group 2.

Streams

Consistency Groups can also be used as a way to spread replication workload across multiple streams within a partnership.

The Metro or Global Mirror partnership architecture allocates traffic from each Consistency Group in a round-robin fashion across 16 streams. That is, cg0 traffic goes into stream0, and cg1 traffic goes into stream1.

Any volume that is *not* in a Consistency Group also goes into stream0. You might want to consider creating an empty Consistency Group 0 so that stand-alone volumes do not share a stream with active Consistency Group volumes.

It can also pay to optimize your streams by creating more Consistency Groups. Within each stream, each batch of writes must be processed in tag sequence order and any delays in processing any particular write also delays the writes behind it in the stream. Having more streams (up to 16) reduces this kind of potential congestion.

Each stream is sequence-tag-processed by one node, so generally you would want to create at least as many Consistency Groups as you have IBM FlashSystem canisters, and, ideally, perfect multiples of the node count.

Layer concept

The *layer* is an attribute of Spectrum Virtualize-based systems which allows you to create partnerships among different Spectrum Virtualize products. The key points concerning layers are listed here:

- IBM SAN Volume Controller is always in the *Replication* layer.
- By default, IBM FlashSystem products are in the *Storage* layer.

- A system can only form partnerships with systems in the same layer.
- An IBM SAN Volume Controller can virtualize an IBM FlashSystem system only if the FlashSystem is in Storage layer.
- An IBM FlashSystem system in the Replication layer can virtualize an IBM FlashSystem system in the Storage layer.

Figure 6-14 shows the concept of layers.

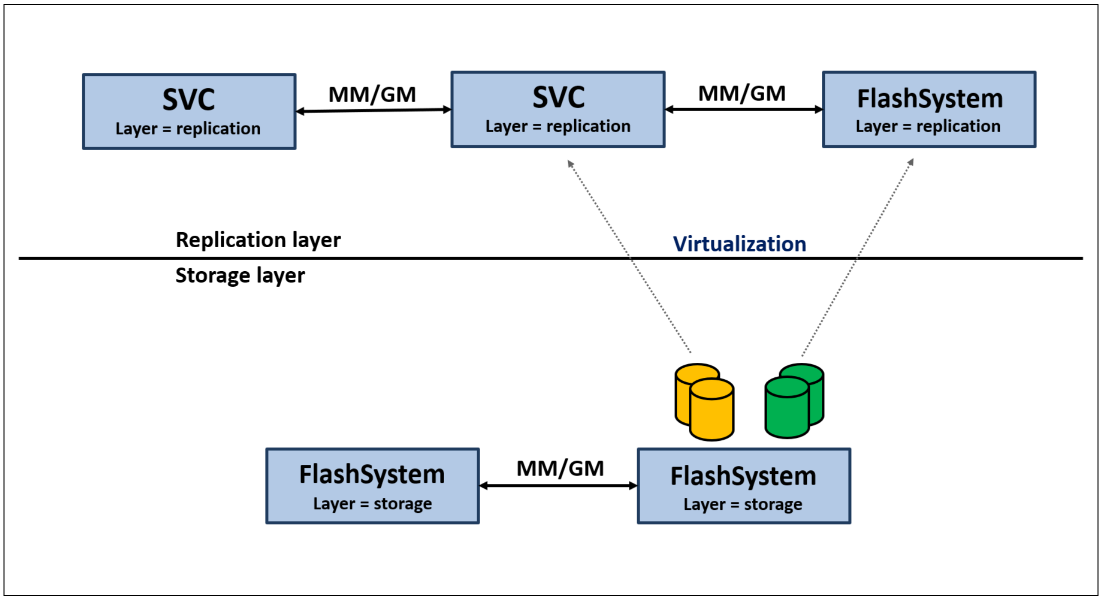


Figure 6-14 Conceptualization of layers

Generally, changing the layer is only performed at initial setup time or as part of a major reconfiguration. To change the layer of an IBM FlashSystem system, the system must meet the following preconditions:

- The IBM FlashSystem system must not have IBM Spectrum Virtualize, Storwize, or FlashSystem host objects defined, and must not be virtualizing any other IBM FlashSystem/Storwize controllers.
- The IBM FlashSystem system must not be visible to any other IBM Spectrum Virtualize, Storwize, or FlashSystem system in the SAN fabric, which might require SAN zoning changes.
- The IBM FlashSystem system must not have any system partnerships defined. If it is already using Metro Mirror or Global Mirror, the existing partnerships and relationships must be removed first.

Changing an IBM FlashSystem system from Storage layer to Replication layer can only be performed by using the CLI. After you are certain that all of the preconditions have been met, issue the following command:

```
chsystem -layer replication
```

Partnership topologies

IBM Spectrum Virtualize allows various partnership topologies, as shown in Figure 6-15. Each box represents an IBM Spectrum Virtualize based system.

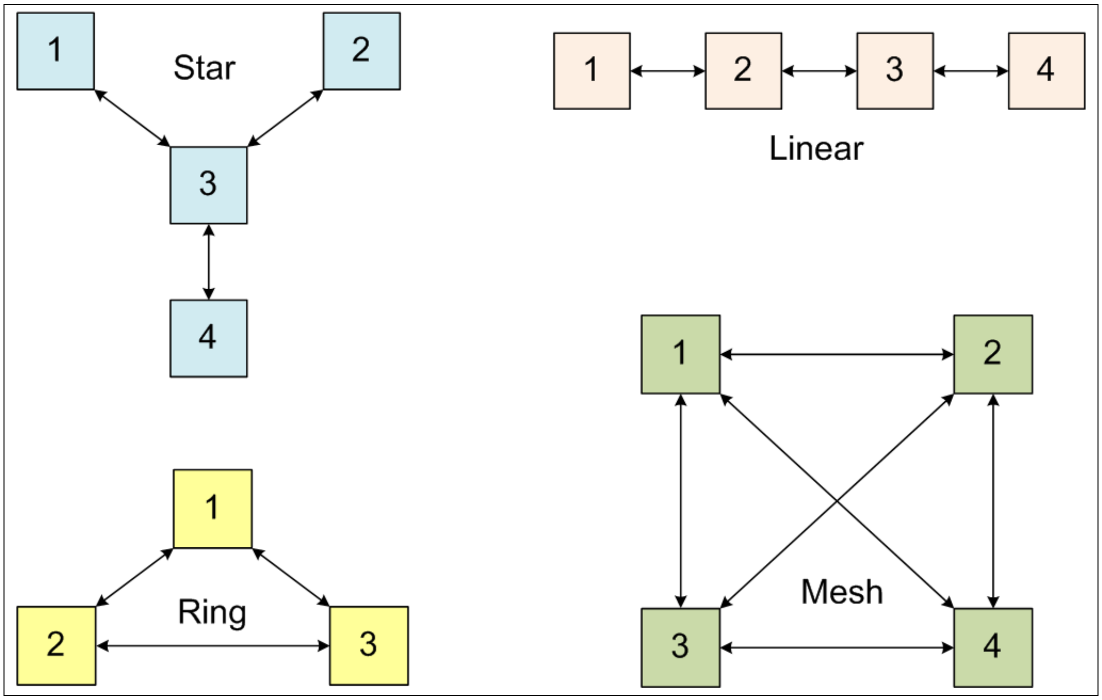


Figure 6-15 Supported topologies for Remote Copy partnerships

The set of systems directly or indirectly connected form the *connected set*. A system can be partnered with up to three remote systems. No more than four systems can be in the same connected set is allowed.

Star topology

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

Ring topology

A ring topology (3 or more systems) can be used to establish a one-in, one-out implementation. For example, the implementation can be 1 → 2, 2 → 3, 3 → 1 to spread replication loads evenly among three systems.

Linear topology

A linear topology of two or more sites is also possible. However, it would generally be simpler to create partnerships between system 1 and system 2, and separately between system 3 and system 4.

Mesh topology

A fully connected mesh topology is where every system has a partnership to each of the three other systems. This topology allows flexibility in that volumes can be replicated between any two systems.

Topology considerations: Consider the following points:

- Although systems can have up to three partnerships, any one volume can be part of only a single relationship. That is, you cannot establish a multi-target Remote Copy relationship for a specific volume. However, three-site replication is possible with the introduction of the Spectrum Virtualize 3-site replication. For more information, see *IBM Spectrum Virtualize 3-Site Replication*, SG24-8504.
- Although various topologies are supported, it is advisable to keep your partnerships as simple as possible, which in most cases mean system pairs or a star.

Intrasystem Remote Copy

The Intrasystem Remote Copy feature allows Remote Copy relationships to be created within the same Spectrum Virtualize system. A preconfigured *local parthership* is created by default in the system for the intrasystem Remote Copy.

Considering that within a single system a Remote Copy does not protect data in a disaster scenarios, this capability has no practical use except for functional testing. For this reason, intrasystem Remote Copy is not officially supported for production data.

Metro Mirror functional overview

Metro Mirror provides synchronous replication. It is designed to ensure that updates are committed to both the primary and secondary volumes before sending an acknowledgment (Ack) of the completion to the server.

If the primary volume fails completely for any reason, Metro Mirror is designed to ensure that the secondary volume holds the same data as the primary did immediately before the failure.

Metro Mirror provides the simplest way to maintain an identical copy on both the primary and secondary volumes. However, as with any synchronous copy over long distance, there can be a performance impact to host applications due to network latency.

Metro Mirror supports relationships between volumes that are up to 300 kilometers (km) apart. Latency is an important consideration for any Metro Mirror network. With typical fiber optic round-trip latencies of 1 millisecond (ms) per 100 km, you can expect a minimum of 3 ms extra latency, due to the network alone, on each I/O if you are running across the 300 km separation.

Figure 6-16 shows the order of Metro Mirror write operations.

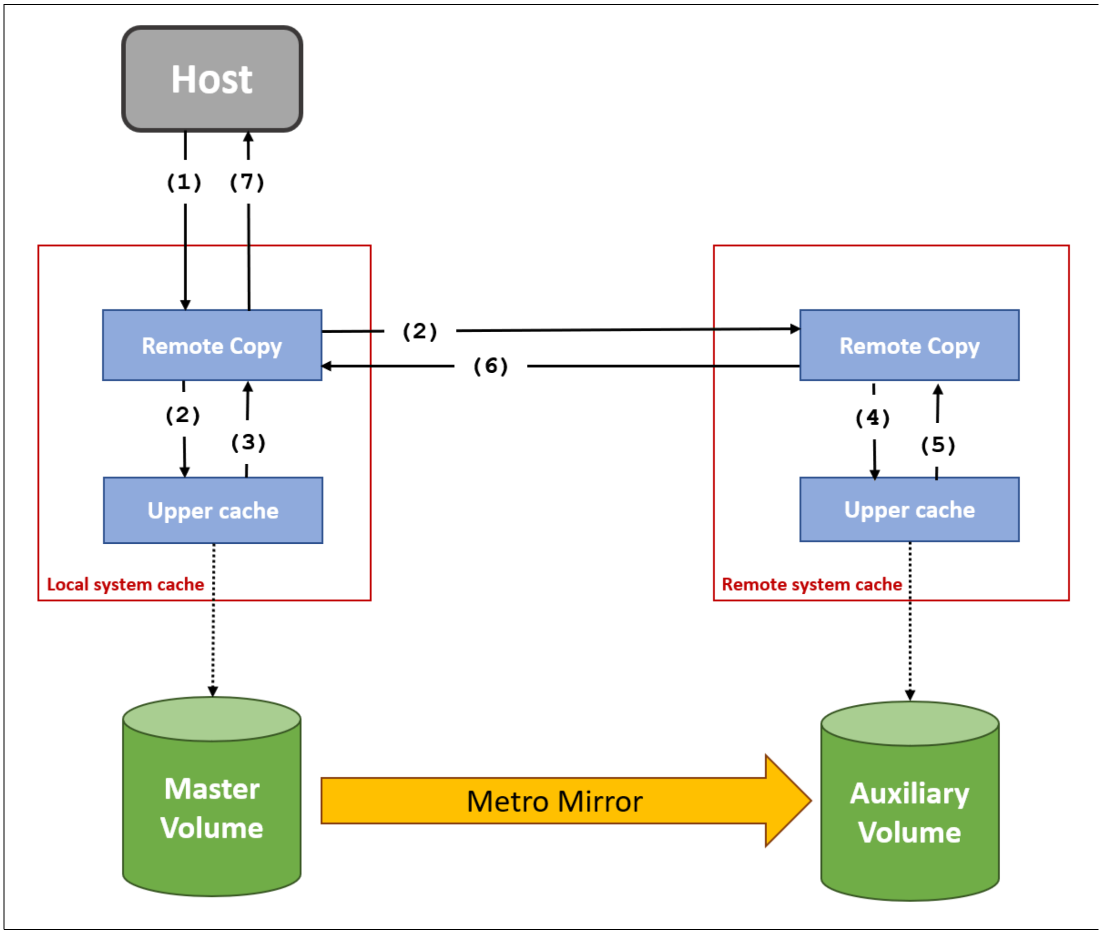


Figure 6-16 Metro Mirror write sequence

The write operation sequence includes the following steps:

1. The write operation is initiated by the host and intercepted by the Remote Copy component of the local system cache.
2. The write operation is simultaneously written in the upper cache component and sent to the remote system.
3. The write operation on local system upper cache is acknowledged back to Remote Copy component on local system.
4. The write operation is written in the upper cache component of the remote system. This operation is initiated as soon as the data arrives from the local system and do not depend on operation ongoing in the local system.
5. The write operation on remote system upper cache is acknowledged back to Remote Copy component on remote system.
6. The remote write operation is acknowledged back to Remote Copy component on local system.
7. The write operation is acknowledged back to the host.

For a write to be considered as committed, it is required that the data is written in both local and remote systems cache. De-staging to disk is a natural part of I/O management, but it is not generally in the critical path for a Metro Mirror write acknowledgment.

Global Mirror functional overview

Global Mirror provides asynchronous replication. It is designed to reduce the dependency on round-trip network latency by acknowledging the primary write in parallel with sending the write to the secondary volume.

If the primary volume fails completely for any reason, Global Mirror is designed to ensure that the secondary volume holds the same data as the primary did at a point a short time before the failure. That short period of data loss is typically 10 ms - 10 seconds, but varies according to individual circumstances.

Global Mirror provides a way to maintain a write-order-consistent copy of data at a secondary site only slightly behind the primary. Global Mirror has minimal impact on the performance of the primary volume.

Although Global Mirror is an asynchronous Remote Copy technique, foreground writes at the local system and mirrored foreground writes at the remote system are not wholly independent of one another. IBM Spectrum Virtualize implementation of Global Mirror uses algorithms to maintain a consistent image at the target volume always.

They achieve this image by identifying sets of I/Os that are active concurrently at the source, assigning an order to those sets, and applying these sets of I/Os in the assigned order at the target. The multiple I/Os within a single set are applied concurrently.

The process that marshals the sequential sets of I/Os operates at the remote system, and therefore is not subject to the latency of the long-distance link.

Figure 6-17 on page 283 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.

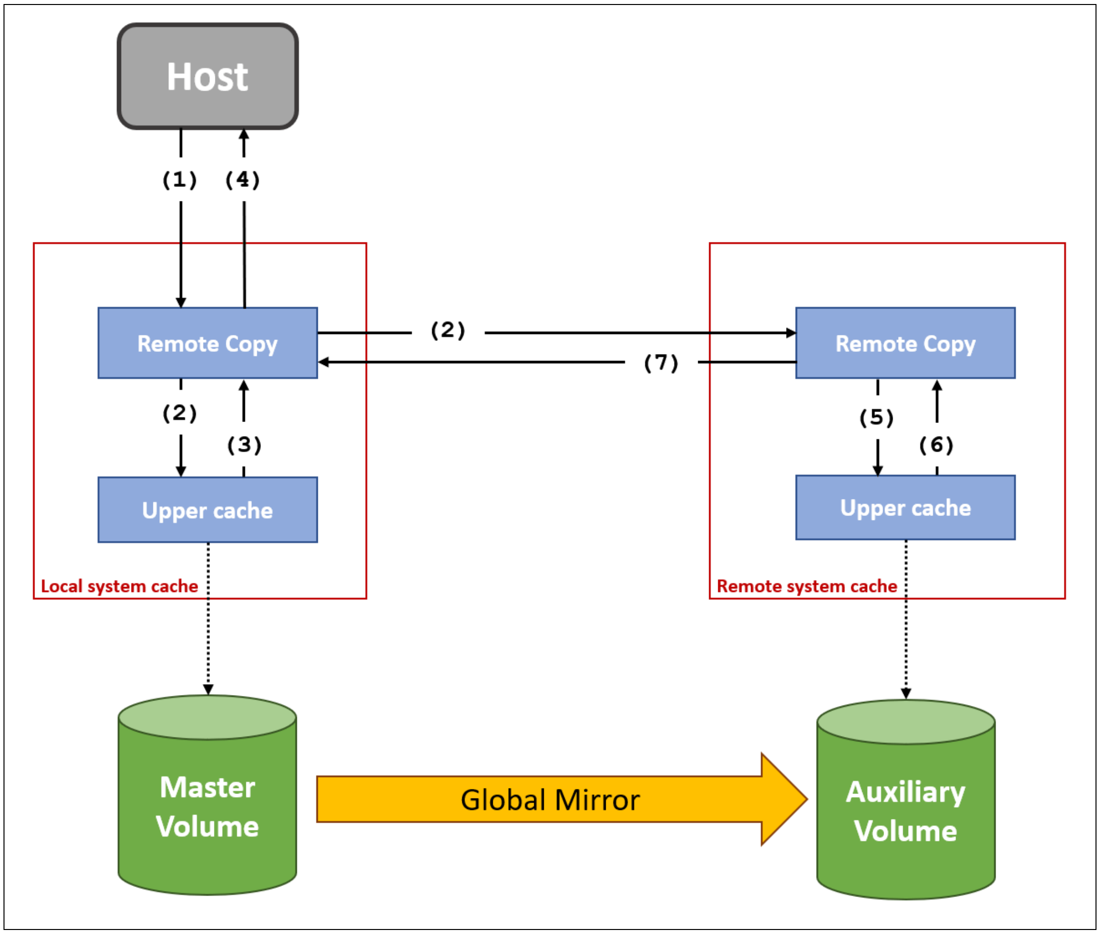


Figure 6-17 Global Mirror relationship write operation

The write operation sequence includes the following steps:

1. The write operation is initiated by the host and intercepted by the Remote Copy component of the local system cache.
2. The Remote Copy component on local system completes the sequence tagging and the write operation is simultaneously written in the upper cache component and sent to the remote system (along with the sequence number).
3. The write operation on local system upper cache is acknowledged back to Remote Copy component on local system.
4. The write operation is acknowledged back to the host.
5. The Remote Copy component on remote system initiates the write operation to the upper cache component according with the sequence number. This operation is initiated as soon as the data arrives from the local system and do not depend on operation ongoing in the local system.
6. The write operation on remote system upper cache is acknowledged back to Remote Copy component on remote system.
7. The remote write operation is acknowledged back to Remote Copy component on local system.

With Global Mirror, a confirmation is sent to the host server before the host receives a confirmation of the completion at the auxiliary volume. The Global Mirror function identifies sets of write I/Os that are active concurrently at the primary volume. It then assigns an order to those sets and applies these sets of I/Os in the assigned order at the auxiliary volume.

Further writes might be received from a host when the secondary write is still active for the same block. In this case, although the primary write might complete, the new host write on the auxiliary volume is delayed until the previous write is completed. Finally, note that any delay in step 2 is reflected in write-delay on primary volume.

Write ordering

Many applications that use block storage are required to survive failures, such as a loss of power or a software crash. They are also required to not lose data that existed before the failure. Because many applications must perform many update operations in parallel to that storage block, maintaining write ordering is key to ensuring the correct operation of applications after a disruption.

An application that performs a high volume of database updates is often designed with the concept of dependent writes. Dependent writes ensure that an earlier write completes before a later write starts. Reversing the order of dependent writes can undermine the algorithms of the application and can lead to problems, such as detected or undetected data corruption.

Colliding writes

Colliding writes are defined as new write I/Os that overlap existing active write I/Os.

The original Global Mirror algorithm required only a single write to be active on any 512-byte LBA of a volume. If another write was received from a host while the auxiliary write was still active, the new host write was delayed until the auxiliary write was complete (although the master write might complete). This restriction was needed if a series of writes to the auxiliary must be retried (which is known as *reconstruction*). Conceptually, the data for reconstruction comes from the master volume.

If multiple writes were allowed to be applied to the master for a sector, only the most recent write had the correct data during reconstruction. If reconstruction was interrupted for any reason, the intermediate state of the auxiliary was inconsistent.

Applications that deliver such write activity do not achieve the performance that Global Mirror is intended to support. A volume statistic is maintained about the frequency of these collisions. The original Global Mirror implementation has been modified to allow multiple writes to a single location to be outstanding in the Global Mirror algorithm.

A need still exists for master writes to be serialized. The intermediate states of the master data must be kept in a non-volatile journal while the writes are outstanding to maintain the correct write ordering during reconstruction. Reconstruction must never overwrite data on the auxiliary with an earlier version. The colliding writes of volume statistic monitoring are now limited to those writes that are not affected by this change.

Figure 6-18 shows a colliding write sequence.

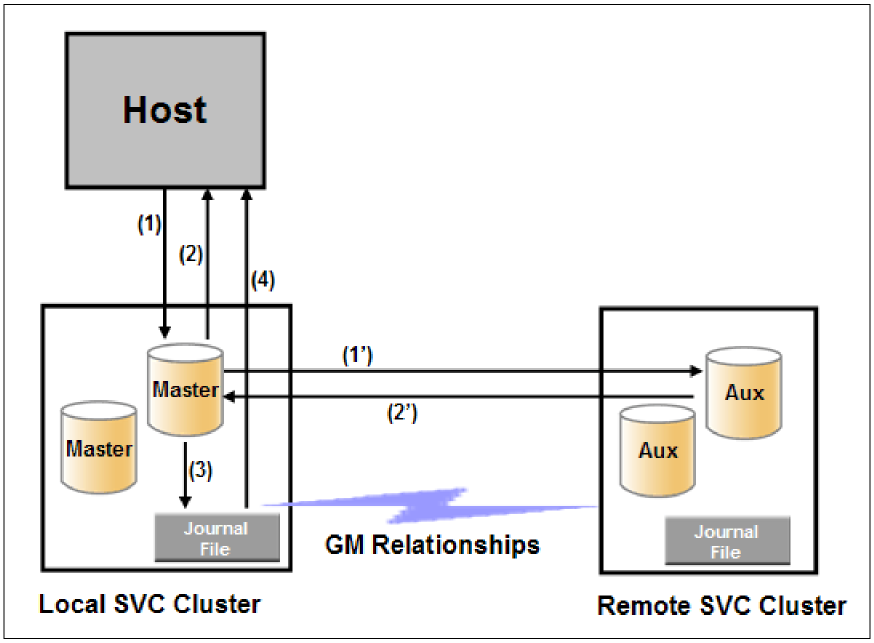


Figure 6-18 Colliding writes

The following numbers correspond to the numbers that are shown in Figure 6-18:

1. A first write is performed from the host to LBA X.
2. A host is provided acknowledgment that the write is complete, even though the mirrored write to the auxiliary volume is not yet completed.

   The first two actions (1 and 2) occur asynchronously with the first write.
3. A second write is performed from the host to LBA X. If this write occurs before the host receives acknowledgment (2), the write is written to the journal file.
4. A host is provided acknowledgment that the second write is complete.

Global Mirror Change Volumes functional overview

Global Mirror with Change Volumes (GMCV) provides asynchronous replication based on point-in-time copies of data. It is designed to allow for effective replication over lower bandwidth networks and to reduce any impact on production hosts.

Metro Mirror and Global Mirror both require the bandwidth to be sized to meet the peak workload. Global Mirror with Change Volumes must only be sized to meet the average workload across a cycle period.

Figure 6-19 shows a high-level conceptual view of Global Mirror with Change Volumes. GMCV uses FlashCopy to maintain image consistency and to isolate host volumes from the replication process.

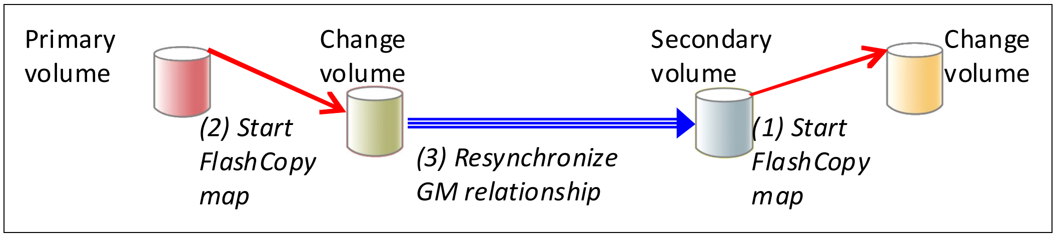


Figure 6-19 Global Mirror with Change Volumes

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

If the primary volume fails completely for any reason, GMCV is designed to ensure that the secondary volume holds the same data as the primary did at a specific point in time. That period of data loss is typically between 5 minutes and 24 hours, but varies according to the design choices that you make.

Change Volumes hold point-in-time copies of 256 KB grains. If any of the disk blocks in a grain change, that grain is copied to the change volume to preserve its contents. Change Volumes are also maintained at the secondary site so that a consistent copy of the volume is always available even when the secondary volume is being updated.

Primary and Change Volumes are always in the same I/O group and the Change Volumes are always thin-provisioned. Change Volumes cannot be mapped to hosts and used for host I/O, and they cannot be used as a source for any other FlashCopy or Global Mirror operations.

Figure 6-20 on page 287 shows how a Change Volume is used to preserve a point-in-time data set, which is then replicated to a secondary site. The data at the secondary site is in turn preserved by a Change Volume until the next replication cycle has completed.

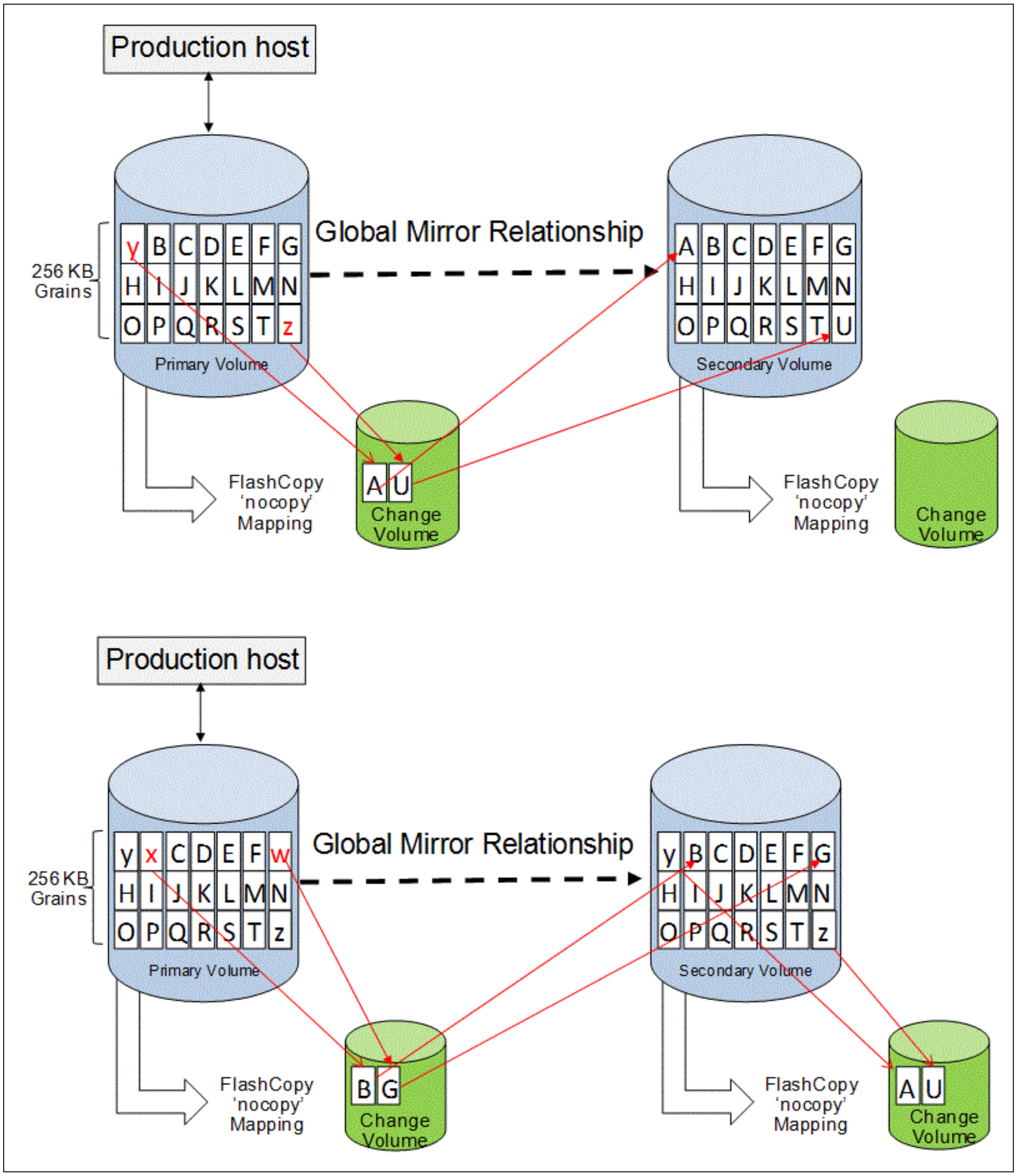


Figure 6-20 Global Mirror with Change Volumes uses FlashCopy point-in-time copy technology

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

The options for **`-cyclingmode`** are **`none`** and **`multi`**.

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

Specifying **`multi`** means that Global Mirror starts cycling based on the cycle period, which defaults to 300 seconds. The valid range is from 60 seconds to 24*60*60 seconds (86,400 seconds = one day).

If all of the changed grains cannot be copied to the secondary site within the specified time, then the replication is designed to take as long as it needs and to start the next replication as soon as the earlier one completes. You can choose to implement this approach by deliberately setting the cycle period to a short amount of time, which is a perfectly valid approach. However, remember that the shorter the cycle period, the less opportunity there is for peak write I/O smoothing, and the more bandwidth you need.

The **`-cyclingmode`** setting can only be changed when the Global Mirror relationship is in a stopped state.

Recovery point objective using Change Volumes

RPO is the maximum tolerable period in which data might be lost if you switch over to your secondary volume.

If a cycle completes within the specified cycle period, then the RPO is not more than 2x cycle long. However, if it does not complete within the cycle period, then the RPO is not more than the sum of the last two cycle times.

The current RPO can be determined by looking at the **`lsrcrelationship`** freeze time attribute. The freeze time is the time stamp of the last primary Change Volume that has completed copying to the secondary site. Note the following example:

1. The cycle period is the default of 5 minutes and a cycle is triggered at 6:00 AM. At 6:03 AM, the cycle completes. The freeze time would be 6:00 AM, and the RPO is 3 minutes.
2. The cycle starts again at 6:05 AM. The RPO now is 5 minutes. The cycle is still running at 6:12 AM, and the RPO is now up to 12 minutes because 6:00 AM is still the freeze time of the last complete cycle.
3. At 6:13 AM, the cycle completes and the RPO now is 8 minutes because 6:05 AM is the freeze time of the last complete cycle.
4. Because the cycle period has been exceeded, the cycle immediately starts again.

6.3.3 Remote Copy network planning

Remote Copy partnerships and relationships do not work reliably if the connectivity on which they are running is configured incorrectly. This section focuses on the inter-system network, giving an overview of the remote system connectivity options.

Terminology

The inter-system network is specified in terms of *latency* and *bandwidth*. These parameters define the capabilities of the link regarding the traffic that is on it. They be must be chosen so that they support all forms of traffic, including mirrored foreground writes, background copy writes, and inter-system 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 latency measures the time spent to send the data and to receive the acknowledgment back (Round Trip Time - RTT).

Link bandwidth is the network capacity to move data as measured in millions of bits per second (Mbps) or billions of bits per second (Gbps).

The term *bandwidth* is also used in the following context:

- Storage bandwidth: The ability of the back-end storage to process I/O. Measures the amount of data (in bytes) that can be sent in a specified amount of time.
- Remote Copy partnership bandwidth (parameter): The rate at which background write synchronization is attempted (unit of MBps).

inter-system connectivity supports mirrored foreground and background I/O. A portion of the link is also used to carry traffic that is associated with the exchange of low-level messaging between the nodes of the local and remote systems. A *dedicated amount* of the link bandwidth is required for the exchange of heartbeat messages and the initial configuration of inter-system partnerships.

Fibre Channel connectivity is the standard connectivity that is used for the Remote Copy inter-system networks. It uses the FC protocol and SAN infrastructures to interconnect the systems.

Native IP connectivity is a connectivity option based on standard TPC/IP infrastructures provided by IBM Spectrum Virtualize technology.

Standard SCSI operations and latency

A single SCSI read operation over a Fibre Channel network is shown in Figure 6-21.

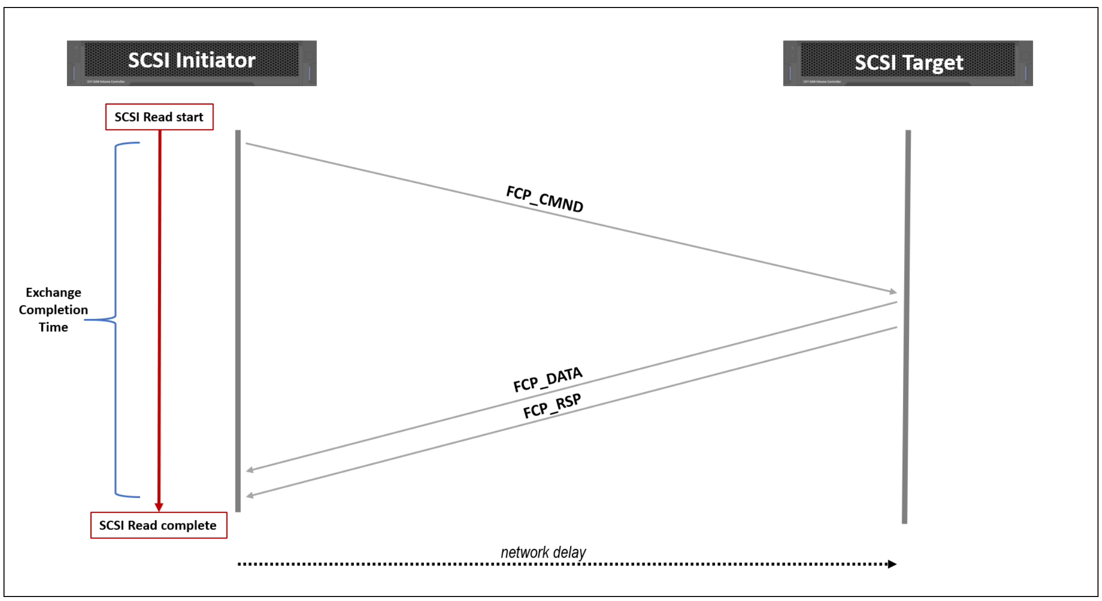


Figure 6-21 Standard SCSI read operation

The initiator starts by sending a read command (FCP_CMND) across the network to the target. The target is responsible to retrieve the data and to respond sending the data (FCP_DATA_OUT) to the initiator. Finally, the target completes the operation sending the command completed response (FCP_RSP). Note that FCP_DATA_OUT and FCP_RSP are sent to the initiator in sequence. Overall, one round trip is required to complete the read; therefore, the read takes at least one RTT, plus the time for the data out.

Typical SCSI behavior for a write is shown in Figure 6-22.

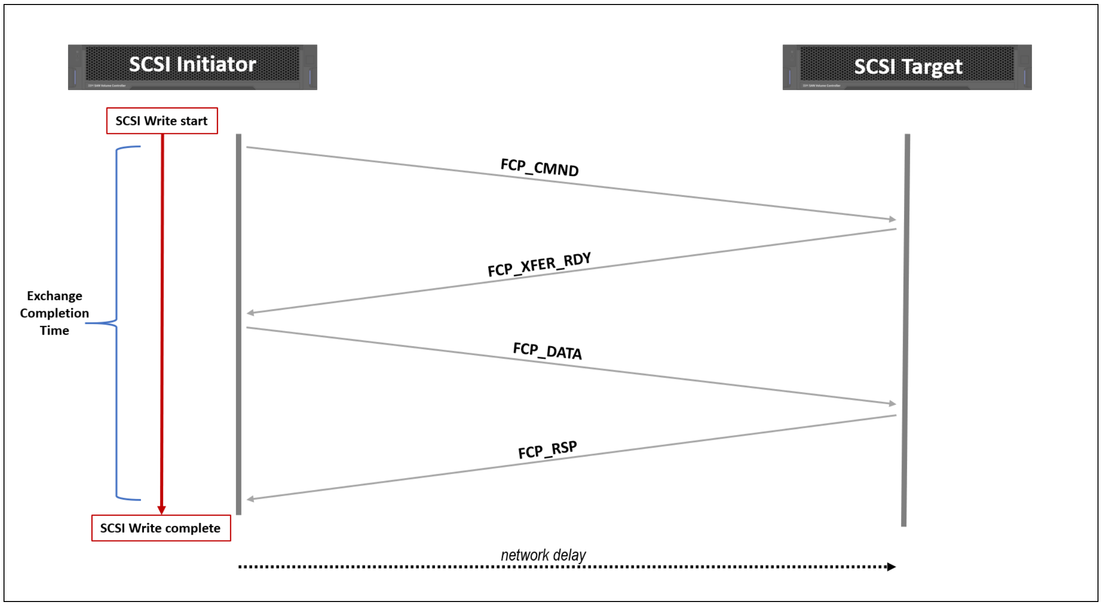


Figure 6-22 Standard SCSI write operation

A standard-based SCSI write is a two-step process:

1. The write command (**FCP_CMND**) is sent across the network to the target. The first round trip is essentially asking transfer permission from the target. The target responds with an acceptance (**FCP_XFR_RDY**).

   The initiator waits until it receives a response from the target before starting the second step; that is, sending the data (**FCP_DATA_OUT**).

2. The target completes the operation y sending the command completed response (**FCP_RSP**). Overall, two round trips are required to complete the write; therefore, the write takes at least 2 × RTT, plus the time for the data out.

Within the confines of a data center, where the latencies are measured in microseconds (µsec), no issues exist. However, across a geographical network where the latencies are measured in milliseconds (ms), the overall service time can be significantly affected.

Considering that the network delay over fiber optics per kilometer (km) is approximately 5 µsec (10 µsec RTT), the resulting minimum service time per every km of distance for a SCSI operation is 10 µsec and 20 µsec for reads and writes respectively; for example, a SCSI write over 50 km has a minimum service time of 1000 µsec (that is, 1ms).

IBM Spectrum Virtualize remote write operations

With the standard SCSI operations, the writes are especially affected by the latency. IBM Spectrum Virtualize implements a proprietary protocol to mitigate the effects of the latency in the write operations over a FC network.

Figure 6-23 shows how a remote copy write operation is performed over a Fibre Channel network.

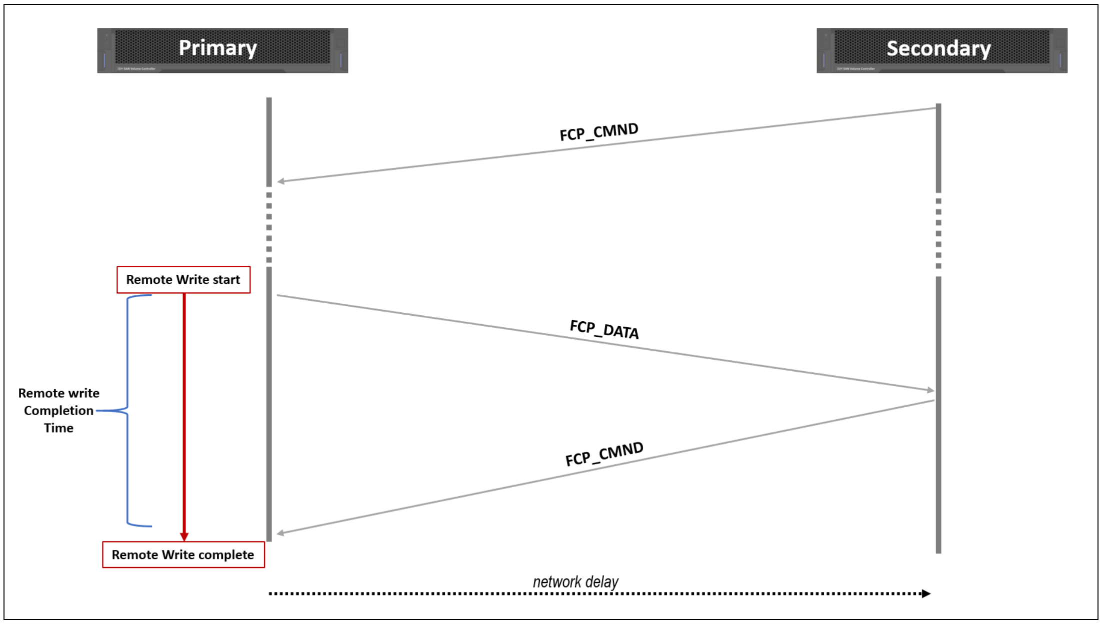


Figure 6-23 Spectrum Virtualize remote copy write

When the remote copy is initialized, the target system (secondary system) sends a dummy read command (**FCP_CMND**) to the initiator (primary system). This command waits on the initiator until a write operation is requested.

When a write operation is started, the data is sent to the target as response of the dummy read command (**FCP_DATA_OUT**). Finally, the target completes the operation by sending a new dummy read command (**FCP_CMND**).

Overall, one round trip is required to complete the remote write using this protocol; therefore, replicating a write takes at least one RTT plus the time for the data out.

Network latency considerations

The maximum supported round-trip latency between sites depends on the type of partnership between systems. Table 6-6 lists the maximum round-trip latency. This restriction applies to all variants of remote mirroring.

Table 6-6 Maximum round trip

Partnership		
FC	**1 Gbps IP**	**10 Gbps IP**
250 ms	80 ms	10 ms

More configuration requirements and guidelines apply to systems that perform remote mirroring over extended distances, where the round-trip time is greater than 80 ms. If you use remote mirroring between systems with 80 - 250 ms round-trip latency, you must meet the following additional requirements:

- The RC buffer size setting must be 512 MB on each system in the partnership. This setting can be accomplished by running the **chsystem -rcbuffersize 512** command on each system.

> **Important:** Changing this setting is disruptive to Metro Mirror and Global Mirror operations. Use this command only before partnerships are created between systems, or when all partnerships with the system are stopped.

- Two FC ports on each node that will be used for replication must be dedicated for replication traffic. This configuration can be achieved by using SAN zoning and port masking.
- SAN zoning should be applied to provide separate intrasystem zones for each local-remote I/O group pair that is used for replication. For more information about zoning guidelines, see “Remote system ports and zoning considerations” on page 298.

Link bandwidth that is used by internode communication

IBM Spectrum Virtualize uses part of the bandwidth for its internal inter-system heartbeat. The amount of traffic depends on how many nodes are in each of the local and remote systems. Table 6-7 shows the amount of traffic (in megabits per second) that is generated by different sizes of systems.

Table 6-7 IBM Spectrum Virtualize inter-system heartbeat traffic (megabits per second)

Local or remote system	Two nodes	Four nodes	Six nodes	Eight nodes
Two nodes	5	6	6	6
Four nodes	6	10	11	12
Six nodes	6	11	16	17
Eight nodes	6	12	17	21

These numbers represent the total traffic between the two systems when *no* I/O is occurring to a mirrored volume on the remote system. Half of the data is sent by one system, and half of the data is sent by the other system. The traffic is divided evenly over all available connections. Therefore, if you have two redundant links, half of this traffic is sent over each link during fault-free operation.

If the link between the sites is configured with redundancy to tolerate single failures, size the link so that the bandwidth and latency statements continue to be accurate even during single failure conditions.

Network sizing considerations

Proper network sizing is essential for the Remote Copy services operations. Failing to estimate the network sizing requirements can lead to poor performance in Remote Copy services and the production workload.

Consider that inter-system bandwidth should be capable of supporting the combined traffic of the following items:

- Mirrored foreground writes, as generated by your server applications at peak times
- Background write synchronization, as defined by the Global Mirror bandwidth parameter
- inter-system communication (*heartbeat messaging)*

Calculating the required bandwidth is essentially a question of mathematics based on your current workloads, so you should start by assessing your current workloads.

Metro Mirror and Global Mirror network sizing

With the Metro Mirror, due to its synchronous nature, the amount of replication bandwidth required to mirror a given foreground write-data throughput is not less than the foreground write-data throughput.

The Global Mirror, not having write buffering resources, tends to mirror the foreground write as soon as it is committed in cache and therefore the bandwidth requirements are similar to the Metro Mirror.

For a proper bandwidth sizing with Metro or Global Mirror, you must know your peak write workload to at least a five-minute interval. This information can be easily gained from tools like IBM Spectrum Control. Finally, you need to allow for the background copy, intercluster communication traffic, and a safe margin for unexpected peaks and workload growth.

> **Recommendation:** Do not compromise on bandwidth or network quality when planning a Metro or Global Mirror deployment. If bandwidth is likely to be an issue in your environment, consider GMCV.

As an example, consider a business with the following I/O profile:

- The average write size is 8 KB (= 8 x 8 bits/1024 = 0.0625 Mb).
- For most of the day between 8 AM and 8 PM, the write activity is approximately 1500 writes per second.
- Twice a day (once in the morning and once in the afternoon), the system bursts up to 4500 writes per second for up to 10 minutes.

This example represents a general traffic pattern that might be common in many medium-sized sites. Furthermore, 20% of bandwidth must be left available for the background synchronization.

Metro Mirror or Global Mirror require bandwidth on the instantaneous peak of 4500 writes per second as follows:

```
4500 x 0.0625 = 282 Mbps + 20% resync allowance + 5 Mbps heartbeat = 343 Mbps
dedicated plus any safety margin plus growth
```

GMCV network sizing

The GMCV is typically less demanding in terms of bandwidth requirements for a number of reasons.

First, by using its journaling capabilities, the GMCV provides a way to maintain point-in-time copies of data at a secondary site where insufficient bandwidth is available to replicate the peak workloads in real time.

Another factor that can reduce the bandwidth that is required for GMCV is that it only sends one copy of a changed grain, which might have been rewritten many times within the cycle period.

The GMCV network sizing is basically a trade-off between RPO, journal capacity, and network bandwidth. A direct relationship exists between the RPO and the physical occupancy of the change volumes: the lower the RPO, the less capacity is used by change volumes. However, higher RPO requires usually less network bandwidth.

For a proper bandwidth sizing with GMCV, you need to know your average write workload during the cycle time. This information can be easily gained from tools like IBM Spectrum Control. Finally, you need to allow for the background resync workload, intercluster communication traffic, and a safe margin for unexpected peaks and workload growth.

Consider the following sizing exercises:

- GMCV peak 30-minute cycle time

  If we look at this time broken into 10-minute periods, the peak 30-minute period is made up of one 10-minute period of 4500 writes per second, and two 10-minute periods of 1500 writes per second. The average write rate for the 30-minute cycle period can then be expressed mathematically as follows:

  ```
  (4500 + 1500 + 1500) / 3 = 2500 writes/sec for a 30-minute cycle period
  ```

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

  ```
  2500 x 0.0625 = 157 Mbps + 20% resync allowance + 5 Mbps heartbeat = 195 Mbps
  dedicated plus any safety margin plus growth
  ```

- GMCV peak 60-minute cycle time

  For a cycle period of 60 minutes, the peak 60-minute period is made up of one 10-minute period of 4500 writes per second, and five 10-minute periods of 1500 writes per second. The average write for the 60-minute cycle period can be expressed as follows:

  ```
  (4500 + 5 x 1500) replicating until 8 AM at the latest would probably require at
  least the following bandwidth:
  ```

  ```
  (9000 + 70 x 1500) / 72 = 1584 x 0.0625 = 99 Mbps + 100% + 5 Mbps heartbeat = 203
  Mbps at night plus any safety margin plus growth, non-dedicated, time-shared
  with daytime traffic
  ```

The central principle of sizing is that you need to know your write workload:

- For Metro Mirror and Global Mirror, you need to know the peak write workload.
- For GMCV, you need to know the average write workload.

GMCV bandwidth: In the above samples, the bandwidth estimation for the GMCV is based on the assumption that the write operations occurs in such a way that a change volume grain (that has a size of 256 KB) is completely changed before it is transferred to the remote site. In the real life, this situation is unlikely to occur.

Usually only a portion of a grain is changed during a GMCV cycle, but the transfer process always copies the whole grain to the remote site. This behavior can lead to an unforeseen processor burden in the transfer bandwidth that, in the edge case, can be even higher than the one required for a standard Global Mirror.

Global Mirror and GMCV coexistence considerations

Global Mirror and GMCV relationships can be defined in the same system. With these configurations, particular attention must be paid to bandwidth sizing and the partnership settings.

The two Global Mirror technologies, as previously described, use the available bandwidth in different ways:

- Regular Global Mirror uses the amount of bandwidth needed to sustain the write workload of the replication set.
- The GMCV uses the fixed amount of bandwidth as defined in the partnership as background copy.

For this reason, during GMCV cycle-creation, a fixed part of the bandwidth is allocated for the background copy and only the remaining part of the bandwidth is available for Global Mirror. To avoid bandwidth contention, which can lead to a 1920 error (see 6.3.6, “1920 error” on page 316) or delayed GMCV cycle creation, the bandwidth must be sized to consider both requirements.

Ideally, in these cases the bandwidth should be enough to accommodate the peak write workload for the Global Mirror replication set plus the estimated bandwidth needed to fulfill the RPO of GMCV. If these requirements cannot be met due to bandwidth restrictions, the least impacting option is to increase the GMCV cycle period and then reduce the background copy rate to minimize the chance of a 1920 error.

Note that these considerations also apply to configurations where multiple IBM Spectrum Virtualize based systems are sharing the same bandwidth resources.

Fibre Channel connectivity

When you use FC technology for the inter-system network, consider the following items:

- Redundancy
- Basic topology and problems
- Distance extensions options
- Hops
- Buffer credits
- Remote system ports and zoning considerations

Redundancy

The inter-system network must adopt the same policy toward redundancy as for the local and remote systems to which it is connecting. The ISLs must have redundancy, and the individual ISLs must provide the necessary bandwidth in isolation.

Basic topology and problems

Because of the nature of FC, you must avoid ISL congestion whether within individual SANs or across the inter-system network. Although FC (and IBM FlashSystem system) can handle an overloaded host or storage array, the mechanisms in FC are ineffective for dealing with congestion in the fabric in most circumstances. The problems that are caused by fabric congestion can range from dramatically slow response time to storage access loss. These issues are common with all high-bandwidth SAN devices and are inherent to FC. They are not unique to the IBM Spectrum Virtualize products.

When an FC network becomes congested, the FC switches stop accepting more frames until the congestion clears. They can also drop frames. Congestion can quickly move upstream in the fabric and clog the end devices from communicating anywhere.

This behavior is referred to as *head-of-line blocking*. Although modern SAN switches internally have a nonblocking architecture, head-of-line-blocking still exists as a SAN fabric problem. Head-of-line blocking can result in IBM FlashSystem canisters that cannot mirror their write caches because you have a single congested link that leads to an edge switch.

Distance extensions options

To implement remote mirroring over a distance by using the FC, you have the following choices:

- *Optical multiplexors*, such as dense wavelength division multiplexing (DWDM) or coarse wavelength division multiplexing (CWDM) devices.

  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 extensions, closely monitor your physical link error counts in your switches. Optical communication devices are high-precision units. When they shift out of calibration, you will start to see errors in your frames.

- Long-distance Small Form-factor Pluggable (SFP) transceivers and XFPs.

  Long-distance optical transceivers have the advantage of extreme simplicity. You do not need expensive equipment, and only a few configuration steps need to be performed. However, ensure that you use only transceivers that are designed for your particular SAN switch.

- Fibre Channel-to-IP conversion boxes. FC over IP (FCIP) is, by far, the most common and least expensive form of distance extension. It is also complicated to configure. Relatively subtle errors can have severe performance implications.

  With IP-based distance extension, you must dedicate bandwidth to your FCIP traffic if the link is shared with other IP traffic. Do not assume that because the link between two sites has low traffic or is used only for email, this type of traffic is always the case. FC is far more sensitive to congestion than most IP applications.

  Also, when you are communicating with the networking architects for your organization, make sure to distinguish between *megabytes per second* as opposed to *megabits per second*. In the storage world, bandwidth is often specified in megabytes per second (MBps), and network engineers specify bandwidth in megabits per second (Mbps).

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

For more information about supported SAN routers and FC extenders, see this IBM Documentation web page.

Hops

The hop count is not increased by the intersite connection architecture. For example, if you have a SAN extension that is based on DWDM, the DWDM components are not apparent to the number of hops. The hop count limit within a fabric is set by the fabric devices (switch or director) operating system. It is used to derive a frame hold time value for each fabric device.

This hold time value is the maximum amount of time that a frame can be held in a switch before it is dropped, or `the fabric is busy` condition is returned. For example, a frame might be held if its destination port is unavailable. The hold time is derived from a formula that uses the error detect timeout value and the resource allocation timeout value. It is considered that every extra hop adds about 1.2 microseconds of latency to the transmission.

Currently, IBM FlashSystem copy services support three hops when protocol conversion exists. Therefore, if you have DWDM extended between primary and secondary sites, three SAN directors or switches can exist between the primary and secondary systems.

Buffer credits

SAN device ports need memory to temporarily store frames as they arrive, assemble them in sequence, and deliver them to the upper layer protocol. The number of frames that a port can hold is called its *buffer credit*. FC architecture is based on a flow control that ensures a constant stream of data to fill the available pipe.

When two FC ports begin a conversation, they exchange information about their buffer capacities. An FC port sends only the number of buffer frames for which the receiving port gives credit. This method avoids overruns and provides a way to maintain performance over distance by filling the pipe with in-flight frames or buffers.

The following types of transmission credits are available:

- Buffer_to_Buffer Credit

  During login, N_Ports and F_Ports at both ends of a link establish its Buffer to Buffer Credit (BB_Credit).

- End_to_End Credit

  In the same way during login, all N_Ports establish End-to-End Credit (EE_Credit) with each other. During data transmission, a port must not send more frames than the buffer of the receiving port can handle before you receive an indication from the receiving port that it processed a previously sent frame. Two counters are used: BB_Credit_CNT and EE_Credit_CNT. Both counters are initialized to zero during login.

> **FC Flow Control:** Each time that a port sends a frame, it increments BB_Credit_CNT and EE_Credit_CNT by one. When it receives R_RDY from the adjacent port, it decrements BB_Credit_CNT by one. When it receives ACK from the destination port, it decrements EE_Credit_CNT by one.
>
> At any time, if BB_Credit_CNT becomes equal to the BB_Credit, or EE_Credit_CNT becomes equal to the EE_Credit of the receiving port, the transmitting port stops sending frames until the respective count is decremented.

The previous statements are true for Class 2 service. Class 1 is a dedicated connection. Therefore, BB_Credit is not important, and only EE_Credit is used (EE Flow Control). However, Class 3 is an unacknowledged service. Therefore, it uses only BB_Credit (BB Flow Control), but the mechanism is the same in all cases.

The number of buffers is an important factor in overall performance. You need enough buffers to ensure that the transmitting port can continue to send frames without stopping to use the full bandwidth, which is true with distance. The total amount of buffer credit needed to optimize the throughput depends on the link speed and the average frame size.

For example, consider an 8 Gbps link connecting two switches that are 100 km apart. At 8 Gbps, a full frame (2148 bytes) occupies about 0.51 km of fiber. In a 100 km link, you can send 198 frames before the first one reaches its destination. You need an ACK to go back to the start to fill EE_Credit again. You can send another 198 frames before you receive the first ACK.

You need at least 396 buffers to allow for nonstop transmission at 100 km distance. The maximum distance that can be achieved at full performance depends on the capabilities of the FC node that is attached at either end of the link extenders, which are vendor-specific. A match should occur between the buffer credit capability of the nodes at either end of the extenders.

Remote system ports and zoning considerations

Ports and zoning requirements for the remote system partnership have changed over time.

The current preferred configuration is based on the information that is available at this IBM Support web page.

The preferred practice for the IBM FlashSystem systems is to provision dedicated node ports for local node-to-node traffic (by using port masking) and isolate Global Mirror node-to-node traffic between the local nodes from other local SAN traffic.

> **Remote port masking:** To isolate the node-to-node traffic from the Remote Copy traffic, the local and remote port masking implementation is preferable.

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

- The minimum requirement to establish a Remote Copy partnership is to connect at least one node per system. When remote connectivity among all the nodes of both systems is not available, the nodes of the local system not participating to the remote partnership will use the node/nodes defined in the partnership as a bridge to transfer the replication data to the remote system.

  This replication data transfer occurs through the node-to-node connectivity. Note that this configuration, even though supported, allows the replication traffic to go through the node-to-node connectivity and this is not recommended.
- Partnered systems should use the same number of nodes in each system for replication.
- For maximum throughput, all nodes in each system should be used for replication, both in terms of balancing the preferred node assignment for volumes and for providing inter-system FC connectivity.
- Where possible, use the minimum number of partnerships between systems. For example, assume site A contains systems A1 and A2, and site B contains systems B1 and B2. In this scenario, creating separate partnerships between pairs of systems (such as A1-B1 and A2-B2) offers greater performance for Global Mirror replication between sites than a configuration with partnerships defined between all four systems.

For zoning, the following rules for the remote system partnership apply:

- For Remote Copy configurations where the round-trip latency between systems is less than 80 milliseconds, zone two FC ports on each node in the local system to two FC ports on each node in the remote system.

- For Remote Copy configurations where the round-trip latency between systems is more than 80 milliseconds, apply SAN zoning to provide separate intrasystem zones for each local-remote I/O group pair that is used for replication, as shown in Figure 6-24.

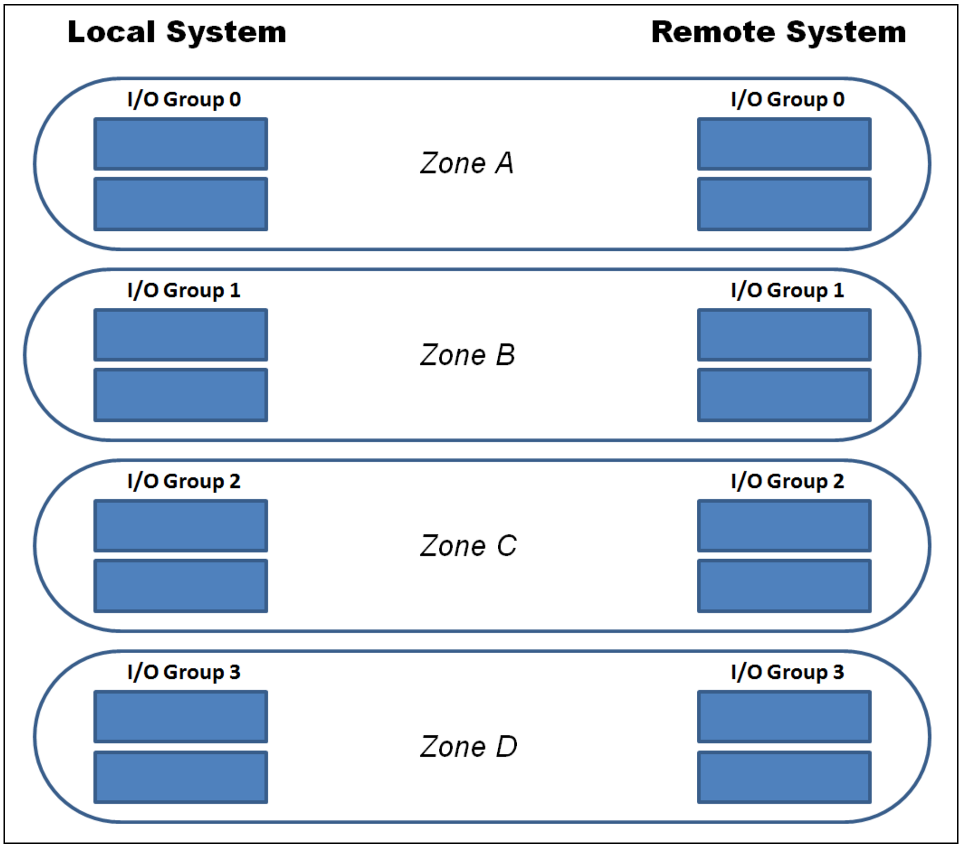


Figure 6-24 Zoning scheme for >80 ms Remote Copy partnerships

NPIV: IBM FlashSystem systems with the NPIV feature enabled provide virtual WWPN for the host zoning. Those WWPNs are intended for host zoning only and cannot be used for the Remote Copy partnership.

SAN Extension design considerations

Disaster Recovery solutions based on Remote Copy technologies require reliable SAN extensions over geographical links. In order to avoid single points of failure, multiple physical links are usually implemented. When implementing these solutions, particular attention must be paid in the Remote Copy network connectivity set up.

Consider a typical implementation of a Remote Copy connectivity using ISLs, as shown in Figure 6-25 on page 300.

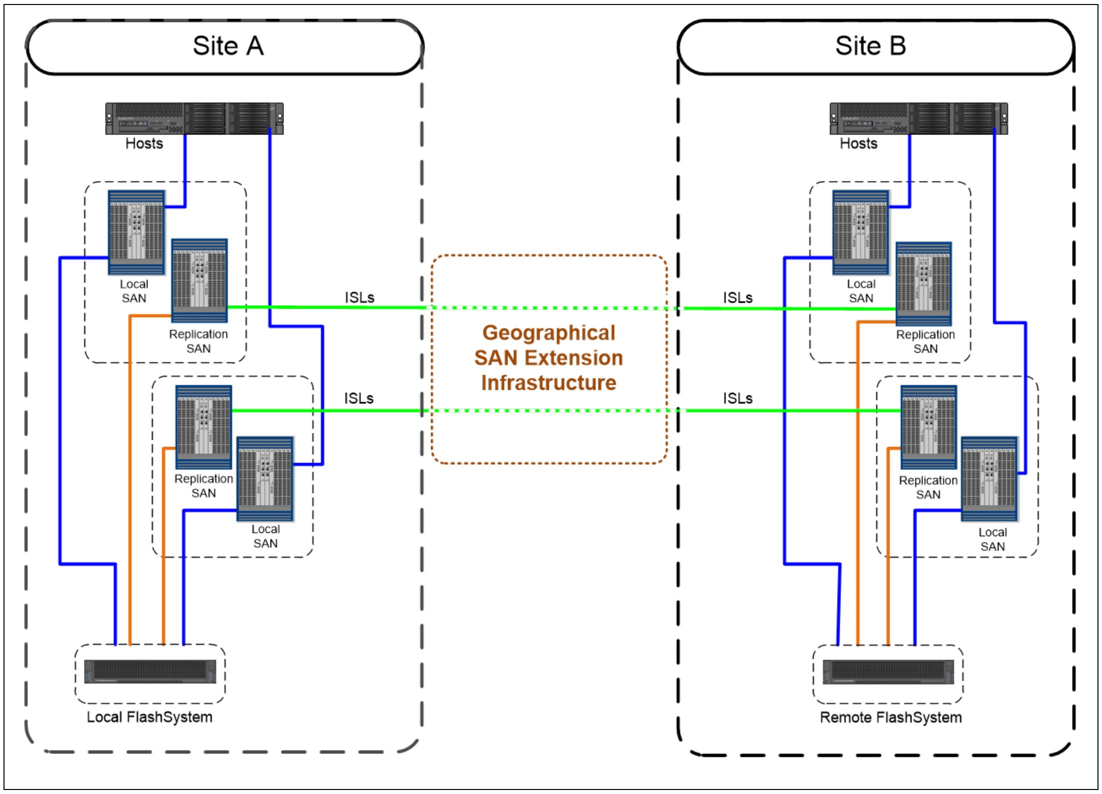


Figure 6-25 Typical Remote Copy network configuration

In the configuration that is shown in Figure 6-25, the Remote Copy network is isolated in a Replication SAN that interconnects Site A and Site B through a SAN extension infrastructure using of two physical links. Assume that, for redundancy reasons, two ISLs are used for each fabric for the Replication SAN extension.

There are two possible configurations to interconnect the Replication SANs. In Configuration 1, as shown in Figure 6-26, one ISL per fabric is attached to each physical link through xWDM or FCIP routers. In this case, the physical paths Path A and Path B are used to extend both fabrics.

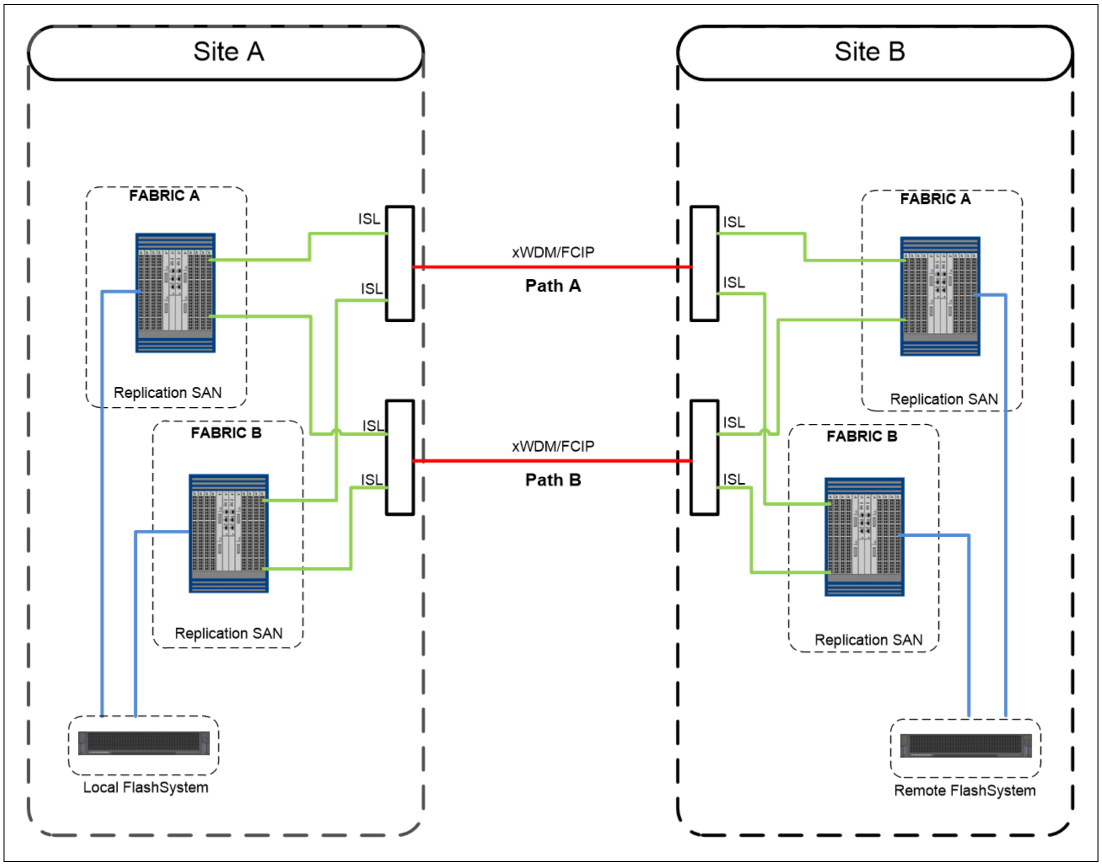


Figure 6-26 Configuration 1: physical paths shared among the fabrics

In Configuration 2, shown in Figure 6-27, ISLs of fabric A are attached only to Path A, while ISLs of fabric B are attached only to Path B. In this case the physical paths are not shared between the fabrics.

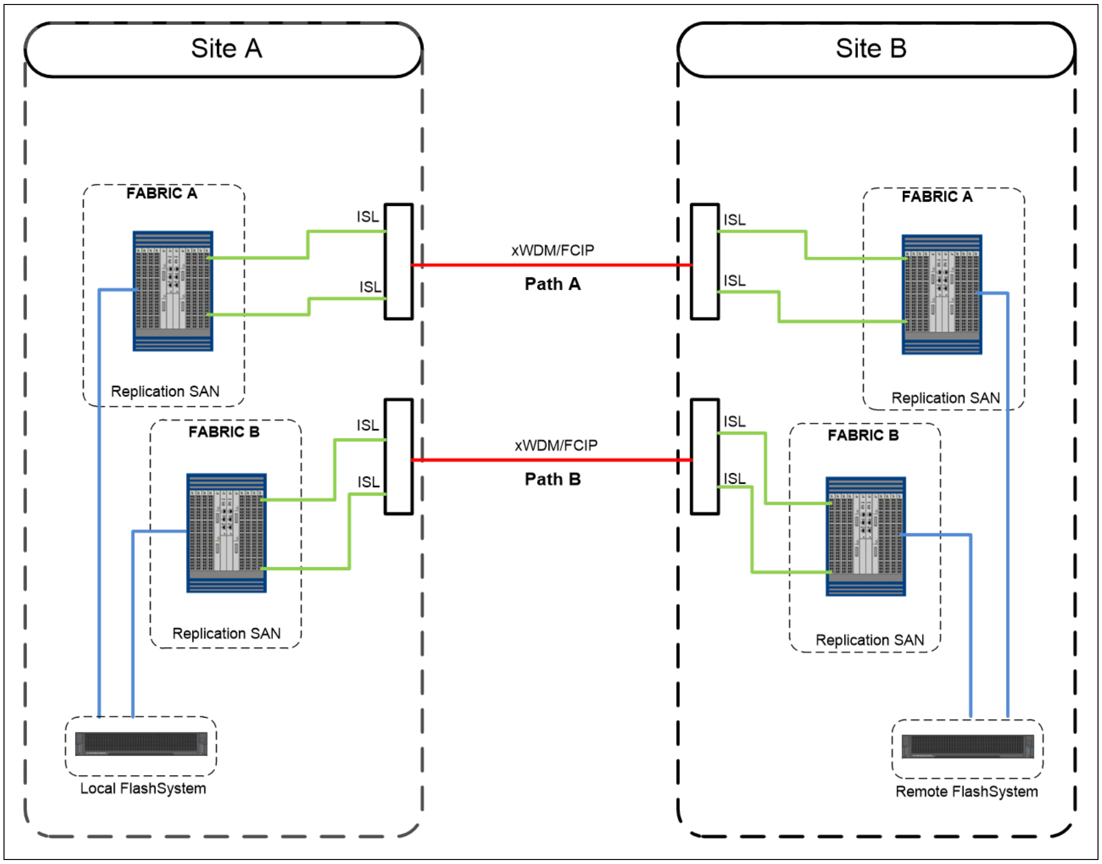


Figure 6-27   Configuration 2: physical paths not shared among the fabrics

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

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

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

6.3.4 Remote Copy services planning

When you plan for Remote Copy services, you must keep in mind the considerations that are outlined in the following sections.

Remote Copy configurations limits

To plan for and implement Remote Copy services, you must check the configuration limits and adhere to them. Table 6-8 shows the limits for a system that currently apply to IBM FlashSystem V8.4. Check the online documentation as these limits can change over time.

Table 6-8 Remote Copy maximum limits

Remote copy property	Maximum	Comment
Remote Copy (Metro Mirror and Global Mirror) relationships per system	10000	This configuration can be any mix of Metro Mirror and Global Mirror relationships.
Active-Active Relationships	2000	This is the limit for the number of HyperSwap volumes in a system.
Remote Copy relationships per consistency group	None	No limit is imposed beyond the Remote Copy relationships per system limit. Apply to Global Mirror and Metro Mirror.
GMCV relationships per consistency group	200	
Remote Copy consistency groups per system	256	
Total Metro Mirror and Global Mirror volume capacity per I/O group	2048 TB	This limit is the total capacity for all master and auxiliary volumes in the I/O group.
Total number of Global Mirror with Change Volumes relationships per system	256	60s cycle time.
	2500	300s cycle time.

Similar to FlashCopy, the Remote Copy services require memory to allocate the bitmap structures used to track the updates while volume are suspended or synchronizing. The default amount of memory for Remote Copy services is 20 MB. This value can be increased or decreased by using the `chiogrp` command. The maximum amount of memory that can be specified for Remote Copy services is 512 MB. The grain size for the Remote Copy services is 256 KB.

Remote Copy general restrictions

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

- You must have the same size for source and target volume when defining a Remote Copy relationship. However, the target volume can be a different type (image, striped, or sequential mode) or have different cache settings (cache-enabled or cache-disabled).
- You cannot move Remote Copy source or target volumes to different I/O groups.
- Remote Copy volumes can be resized with the following restrictions:
  - Resizing applies to Metro Mirror and Global Mirror only. GMCV is not supported.
  - The Remote Copy Consistency Protection feature is not allowed and must be removed before resizing the volumes.
  - The Remote Copy relationship must be in synchronized status.
  - The resize order must ensure that the target volume always is larger than the source volume.

Note: The volume expansion for Metro and Global Mirror volumes was introduced with Spectrum Virtualize version 7.8.1 with some restrictions:

- In the first implementation (up to version 8.2.1), only thin provisioned or compressed volumes were supported.
- With version 8.2.1 also non-mirrored fully allocated volumes were supported.
- With version 8.4 all the restrictions on volume type have been removed.

- You can mirror intrasystem Metro Mirror or Global Mirror only between volumes in the same I/O group.

Intrasystem remote copy: The intrasystem Global Mirror is not supported on IBM Spectrum Virtualize based systems running version 6 or later.

- Global Mirror is not recommended for cache-disabled volumes that are participating in a Global Mirror relationship.

Changing the Remote Copy type

Changing the Remote Copy type for an existing relationship is an easy task. It is enough to stop the relationship, if it is active, and change the properties to set the new Remote Copy type. Remember to create the change volumes in case of change from Metro or Global Mirror to GMCVs.

Interaction between Remote Copy and FlashCopy

Remote Copy functions can be used in conjunction with the FlashCopy function so that you can have both operating concurrently on the same volume. The possible combinations between Remote Copy and FlashCopy follow:

- Remote Copy source:
  - A Remote Copy source can be a FlashCopy source.
  - A Remote Copy source can be a FlashCopy target with the following restrictions:
    - A FlashCopy target volume cannot be updated while it is the source volume of a Metro or Global Mirror relationship that is actively mirroring. A FlashCopy mapping cannot be started while the target volume is in an active Remote Copy relationship.
    - The I/O group for the FlashCopy mappings must be the same as the I/O group for the FlashCopy target volume (that is the I/O group of the Remote Copy source).
- Remote Copy target:
  - A Remote Copy target can be a FlashCopy source.
  - A Remote Copy target can be a FlashCopy target with the following restriction: A FlashCopy mapping must be in the `idle_copied` state when its target volume is the target volume of an active Metro Mirror or Global Mirror relationship.

When implementing FlashCopy functions for volumes in GMCV relationships, FlashCopy multi-target mappings are created. As described in "Interaction and dependency between Multiple Target FlashCopy mappings" on page 259, this results in dependent mappings that can affect the cycle formation due to the cleaning process. For more information, see "**Cleaning process and Cleaning Rate**" on page 270.

With such configurations, it is recommended to set the Cleaning Rate accordingly. This recommendation applies also to Consistency Protection volumes and HyperSwap configurations.

Native back-end controller copy functions considerations

As previously discussed, the IBM FlashSystem technology provides a widespread set of copy services functions that cover most of the clients requirements.

However, some storage controllers can provide specific copy services capabilities not available with the current version of IBM Spectrum Virtualize software. The IBM FlashSystem technology addresses these situations by using cache-disabled image mode volumes that virtualize LUN participating to the native back-end controller's copy services relationships.

Keeping the cache disabled guarantees data consistency throughout the I/O stack, from the host to the back-end controller. Otherwise, by leaving the cache enabled on a volume, the underlying controller does not receive any write I/Os as the host writes them. IBM FlashSystem caches them and processes them later. This process can have more ramifications if a target host depends on the write I/Os from the source host as they are written.

> **Note:** Native copy services are not supported on all storage controllers. For more information about the known limitations, see this IBM Support web page.

As part of its copy services function, the storage controller might take a LUN offline or suspend reads or writes. IBM FlashSystem does not recognize why this happens. Therefore, it might log errors when these events occur. For this reason, if the IBM FlashSystem must detect the LUN, ensure that the LUN remains in the `unmanaged` state until full access is granted.

Native back-end controller copy services can also be used for LUNs that are not managed by the IBM FlashSystem. Note that accidental incorrect configurations of the back-end controller copy services involving IBM FlashSystem attached LUN can produce unpredictable results.

For example, if you accidentally use a LUN with IBM FlashSystem data on it as a point-in-time target LUN, you can corrupt that data. Moreover, if that LUN was a managed disk in a managed-disk group with striped or sequential volumes on it, the managed disk group might be brought offline. This situation, in turn, makes all the volumes that belong to that group go offline, leading to a widespread host access disruption.

Remote Copy and code upgrade considerations

When you upgrade system software where the system participates in one or more inter-system relationships, upgrade only one cluster at a time. That is, do not upgrade the systems concurrently.

> **Attention:** Upgrading both systems concurrently is not monitored by the software upgrade process.

Allow the software upgrade to complete one system before it is started on the other system. Upgrading both systems concurrently can lead to a loss of synchronization. In stress situations, it can further lead to a loss of availability.

Usually, pre-existing Remote Copy relationships are unaffected by a software upgrade that is performed correctly. However, always check in the target code release notes for special considerations on the copy services.

Although it is not a best practice, a Remote Copy partnership can be established, with some restrictions, among systems with different IBM Spectrum Virtualize versions. For more information, see this IBM Support web page.

Volume placement considerations

You can optimize the distribution of volumes within I/O groups at the local and remote systems to maximize performance.

Although defined at a system level, the partnership bandwidth and consequently the background copy rate) is evenly divided among the cluster's I/O groups. The available bandwidth for the background copy can be used by either canister or shared by both canisters within the I/O Group.

This bandwidth allocation is independent from the number of volumes for which a canister is responsible. Each node, in turn, divides its bandwidth evenly between the (multiple) Remote Copy relationships with which it associates volumes that are performing a background copy.

Volume preferred node

Conceptually, a connection (path) goes between each node on the primary system to each node on the remote system. Write I/O, which is associated with remote copying, travels along this path. Each node-to-node connection is assigned a finite amount of Remote Copy resource and can sustain only in-flight write I/O to this limit.

The node-to-node in-flight write limit is determined by the number of nodes in the remote system. The more nodes that exist at the remote system, the lower the limit is for the in-flight write I/Os from a local node to a remote node. That is, less data can be outstanding from any one local node to any other remote node. Therefore, to optimize performance, Global Mirror volumes must have their preferred nodes distributed evenly between the nodes of the systems.

The preferred node property of a volume helps to balance the I/O load between nodes in that I/O group. This property is also used by Remote Copy to route I/O between systems.

The IBM FlashSystem canister that receives a write for a volume is normally the preferred node of the volume. For volumes in a Remote Copy relationship, that node is also responsible for sending that write to the preferred node of the target volume. The primary preferred node is also responsible for sending any writes that relate to the background copy. Again, these writes are sent to the preferred node of the target volume.

Each node of the remote system has a fixed pool of Remote Copy system resources for *each node* of the primary system. That is, each remote node has a separate queue for I/O from each of the primary nodes. This queue is a fixed size and is the same size for every node. If preferred nodes for the volumes of the remote system are set so that every combination of primary node and secondary node is used, Remote Copy performance is maximized.

Figure 6-28 on page 307 shows an example of Remote Copy resources that are not optimized. Volumes from the local system are replicated to the remote system. All volumes with a preferred node of Node 1 are replicated to the remote system, where the target volumes also have a preferred node of Node 1.

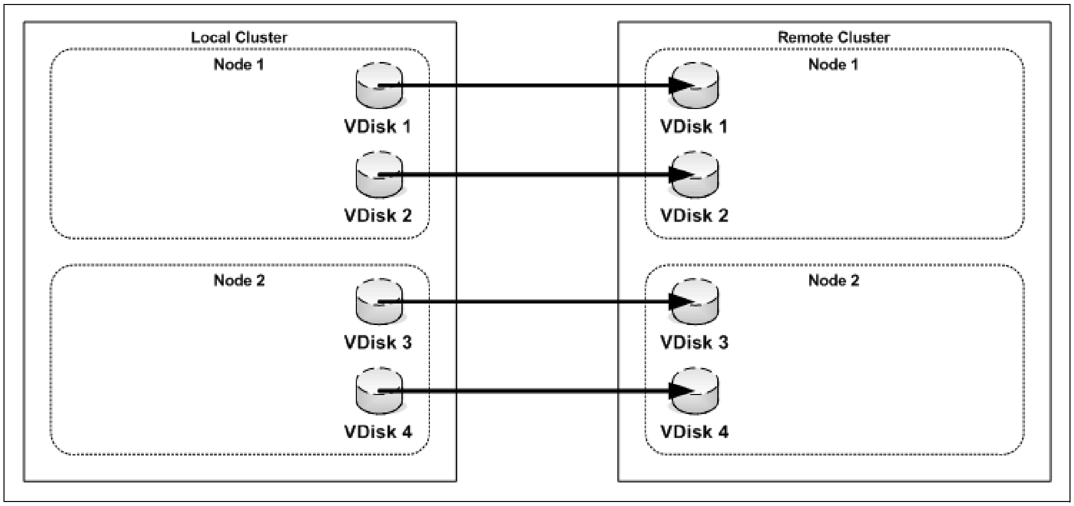


Figure 6-28 Remote Copy resources that are not optimized

With the configuration that is shown in Figure 6-28, the resources for remote system Node 1 that are reserved for local system Node 2 are not used. Also, the resources for local system Node 1 that are reserved for remote system Node 2 are not used.

If the configuration that is shown in Figure 6-28 changes to the configuration that is shown in Figure 6-29, all Remote Copy resources for each node are used, and Remote Copy operates with better performance.

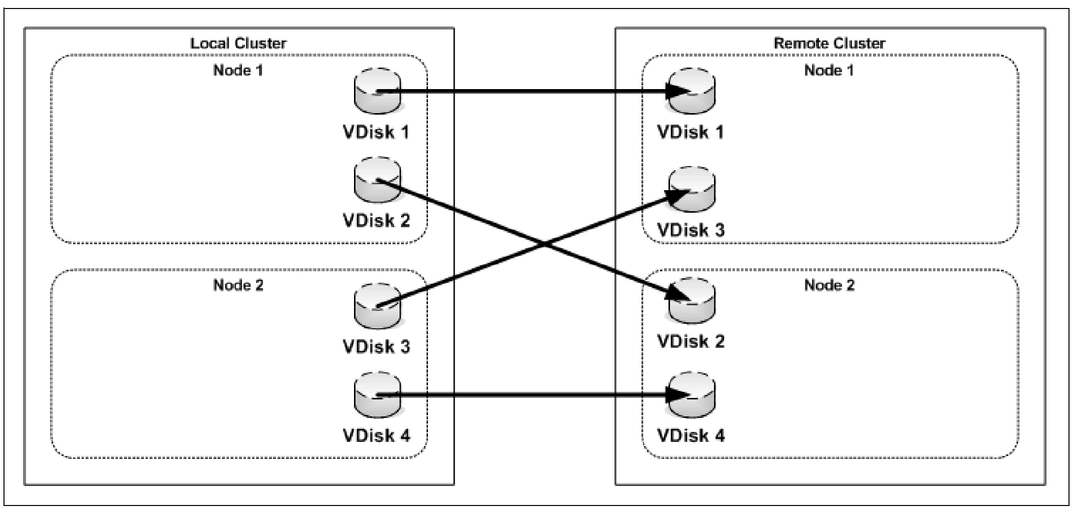


Figure 6-29 Optimized Global Mirror resources

GMCV Change Volumes placement considerations

The Change Volumes in a GMCV configuration are thin-provisioned volumes that are used as FlashCopy targets. For this reason the same considerations apply that are described in “Volume placement considerations” on page 267. The Change Volumes can be compressed to reduce the amount of space used; however, it is important to note that the Change Volumes might be subject to heavy write workload both in the primary and secondary system.

Therefore, the placement on the back-end is critical to provide adequate performances. Consider to use DRP for the change volumes only if beneficial in terms of space savings.

Trick: The internal FlashCopy used by the GMCV is with 256KB grain size. However, it is possible to force a 64KB grain size by creating a FlashCopy with 64KB grain size from the GMCV volume and a dummy target volume before assigning the change volume to the relationship. This can be done to both source and target volumes. After the CV assignment is done, the dummy FlashCopy can be deleted.

Background copy considerations

The Remote Copy partnership bandwidth parameter *explicitly* defines the rate at which the background copy is attempted, but also *implicitly* affects foreground I/O. Background copy bandwidth can affect foreground I/O latency in one of the following ways:

- Increasing latency of foreground I/O

  If the Remote Copy partnership bandwidth parameter is set too high for the actual inter-system network capability, the background copy resynchronization writes use too much of the inter-system network. It starves the link of the ability to service synchronous or asynchronous mirrored foreground writes. Delays in processing the mirrored foreground writes increase the latency of the foreground I/O as perceived by the applications.

- Read I/O overload of primary storage

  If the Remote Copy partnership background copy rate is set too high, the added read I/Os that are associated with background copy writes can overload the storage at the primary site and delay foreground (read and write) I/Os.

- Write I/O overload of auxiliary storage

  If the Remote Copy partnership background copy rate is set too high for the storage at the secondary site, the background copy writes overload the auxiliary storage. Again, they delay the synchronous and asynchronous mirrored foreground write I/Os.

  Important: An increase in the peak foreground workload can have a detrimental effect on foreground I/O. It does so by pushing more mirrored foreground write traffic along the inter-system network, which might not have the bandwidth to sustain it. It can also overload the primary storage.

To set the background copy bandwidth optimally, consider all aspects of your environments, starting with the following biggest contributing resources:

- Primary storage
- inter-system network bandwidth
- Auxiliary storage

Provision the most restrictive of these three resources between the background copy bandwidth and the peak foreground I/O workload. Perform this provisioning by calculation or by determining experimentally how much background copy can be allowed before the foreground I/O latency becomes unacceptable.

Then, reduce the background copy to accommodate peaks in workload. In cases where the available network bandwidth cannot sustain an acceptable background copy rate, consider alternatives to the initial copy as described in “Initial synchronization options and Offline Synchronization” on page 309.

Changes in the environment, or loading of it, can affect the foreground I/O. IBM FlashSystem technology provides a means to monitor, and a parameter to control, how foreground I/O is affected by running Remote Copy processes. IBM Spectrum Virtualize software monitors the delivery of the mirrored foreground writes.

If latency or performance of these writes extends beyond a (predefined or client-defined) limit for a period, the Remote Copy relationship is suspended. For more information, see 6.3.6, “1920 error” on page 316.

Finally, note that with Global Mirror Change Volume, the cycling process that transfers the data from the local to the remote system is a background copy task. For more information, see “Global Mirror and GMCV coexistence considerations” on page 294. For this reason, the background copy rate, and the `relationship_bandwidth_limit` setting, affects the available bandwidth not only during the initial synchronization, but also during the normal cycling process.

Background copy bandwidth allocation: As described in “Volume placement considerations” on page 306, the available bandwidth of a Remote Copy partnership is evenly divided among the cluster’s I/O Groups. In a case of unbalanced distribution of the remote copies among the I/O groups, the partnership bandwidth should be adjusted accordingly to reach the desired background copy rate.

Consider, for example, a 4-I/O groups cluster that has a partnership bandwidth of 4,000 Mbps and a background copy percentage of 50. The expected maximum background copy rate for this partnership is then 250 MBps.

Because the available bandwidth is evenly divided among the I/O groups, every I/O group in this cluster can theoretically synchronize data at a maximum rate of approximately 62 MBps (50% of 1,000 Mbps). Now, in an edge case where only volumes from one I/O group are replicated, the partnership bandwidth should be adjusted to 16000 Mbps to reach the full background copy rate (250 MBps).

Initial synchronization options and Offline Synchronization

When creating a Remote Copy relationship, two options regarding the initial synchronization process are available:

- The `not synchronized` option is the default. With this option, when a Remote Copy relationship is started, a full data synchronization at the background copy rate occurs between the source and target volumes. It is the simplest approach because apart from issuing the necessary IBM FlashSystem commands, other administrative activity is not required. However, in some environments, the available bandwidth makes this option unsuitable.
- The `already synchronized` option does not force any data synchronization when the relationship is started. The administrator must ensure that the source and target volumes contain identical data before a relationship is created. The administrator can perform this check in one of the following ways:
  - Create both volumes with the security delete feature to change all data to zero.
  - Copy a complete tape image (or other method of moving data) from one disk to the other.

  In either technique, write I/O must not take place to the source and target volume before the relationship is established. The administrator must then complete the following actions:
  - Create the relationship with the synchronized settings (**`-sync`** option).
  - Start the relationship.

  Attention: If you do not perform these steps correctly, the Remote Copy reports the relationship as being *consistent*, when it is not. This setting is likely to cause auxiliary volumes to be useless.

By understanding the methods to start a Metro Mirror and Global Mirror relationship, you can use one of them as a means to implement the Remote Copy relationship saving bandwidth.

Consider a situation where you have a large source volume (or many source volumes) containing already-active data and that you want to replicate to a remote site. Your planning shows that the mirror initial-sync time takes too long (or is too costly if you pay for the traffic that you use). In this case, you can set up the sync by using another medium that is less expensive. This synchronization method is called *Offline Synchronization*.

This example uses tape media as the source for the initial sync for the Metro Mirror relationship or the Global Mirror relationship target before it uses Remote Copy services to maintain the Metro Mirror or Global Mirror. This example does not require downtime for the hosts that use the source volumes.

Before you set up Global Mirror relationships and save bandwidth, complete the following steps:

1. Ensure that the hosts are up and running and are using their volumes normally. The Metro Mirror relationship nor Global Mirror relationship is not yet defined.

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

2. Establish the Remote Copy partnership with the target IBM Spectrum Virtualize-based system.

To set up Global Mirror relationships and save bandwidth, complete the following steps:

1. Define a Metro Mirror relationship or a Global Mirror relationship for each source disk. When you define the relationship, ensure that you use the `-sync` option, which stops the system from performing an initial sync.

   > **Attention:** If you do not use the `-sync` option, all of these steps are redundant because the IBM Spectrum Virtualize system will perform a full initial synchronization.

2. Stop each mirror relationship by using the `-access` option, which enables write access to the target volumes. You need write access later.
3. Copy the source volume to the alternative media by using the `dd` command to copy the contents of the volume to tape. Another option is to use your backup tool (for example, IBM Spectrum Protect) to make an image backup of the volume.

   > **Change tracking:** Although the source is being modified while you are copying the image, the IBM FlashSystem is tracking those changes. The image that you create might have some of the changes and is likely to also miss some of the changes.
   >
   > When the relationship is restarted, the IBM FlashSystem applies all of the changes that occurred since the relationship stopped in step 2. After all the changes are applied, you have a consistent target image.

4. Ship your media to the remote site and apply the contents to the targets of the Metro Mirror or Global Mirror relationship. You can mount the Metro Mirror and Global Mirror target volumes to a UNIX server and use the `dd` command to copy the contents of the tape to the target volume.

   If you used your backup tool to make an image of the volume, follow the instructions for your tool to restore the image to the target volume. Remember to remove the mount if the host is temporary.

> **Tip:** It does not matter how long it takes to get your media to the remote site to perform this step. However, the faster you can get the media to the remote site and load it, the quicker IBM FlashSystem system starts running and maintaining the Metro Mirror and Global Mirror.

5. Unmount the target volumes from your host. When you start the Metro Mirror and Global Mirror relationship later, the IBM FlashSystem stops write-access to the volume while the mirror relationship is running.
6. Start your Metro Mirror and Global Mirror relationships. The relationships must be started with the **`-clean`** parameter. This way, changes that are made on the secondary volume are ignored. Only changes made on the clean primary volume are considered when synchronizing the primary and secondary volumes.
7. While the mirror relationship catches up, the target volume is not usable at all. When it reaches `ConsistentSynchnonized` status, your remote volume is ready for use in a disaster.

Back-end storage considerations

To reduce the overall solution costs, it is a common practice to provide the remote systems with lower performance characteristics compared to the local system, especially when using asynchronous Remote Copy technologies. This attitude can be risky especially when using the Global Mirror technology where the application performances at the primary system can indeed be limited by the performance of the remote system.

The preferred practice is to perform an accurate back-end resource sizing for the remote system to fulfill the following capabilities:

- The peak application workload to the Global Mirror or Metro Mirror volumes
- The defined level of background copy
- Any other I/O that is performed at the remote site

Remote Copy tunable parameters

Several commands and parameters help to control Remote Copy and its default settings. You can display the properties and features of the systems by using the **`lssystem`** command. Also, you can change the features of systems by using the **`chsystem`** command.

relationshipbandwidthlimit

The **`relationshipbandwidthlimit`** is an optional parameter that specifies the new background copy bandwidth in the range 1 - 1000 MBps. The default is 25 MBps. This parameter operates system-wide, and defines the maximum background copy bandwidth that any relationship can adopt. The existing background copy bandwidth settings that are defined on a partnership continue to operate, with the lower of the partnership and volume rates attempted.

> **Important:** Do not set this value higher than the default without establishing that the higher bandwidth can be sustained.

The **`relationshipbandwidthlimit`** also applies to Metro Mirror relationships.

gmlinktolerance and *gmmaxhostdelay*

The **gmlinktolerance** and **gmmaxhostdelay** parameters are critical in the system for deciding internally whether to terminate a relationship due to a performance problem. In most cases, these two parameters need to be considered in tandem. The defaults would not normally be changed unless you had a specific reason to do so.

The **gmlinktolerance** parameter can be thought of as how long you allow the host delay to go on being significant before you decide to terminate a Global Mirror volume relationship. This parameter accepts values of 20 - 86,400 seconds in increments of 10 seconds. The default is 300 seconds. You can disable the link tolerance by entering a value of zero for this parameter.

The **gmmaxhostdelay** parameter can be thought of as the maximum host I/O impact that is due to Global Mirror. That is, how long would that local I/O take with Global Mirror turned off, and how long does it take with Global Mirror turned on. The difference is the host delay due to Global Mirror tag and forward processing.

Although the default settings are adequate for most situations, increasing one parameter while reducing another might deliver a tuned performance environment for a particular circumstance.

Example 6-1 shows how to change **gmlinktolerance** and the **gmmaxhostdelay** parameters using the **chsystem** command.

Example 6-1 Changing gmlinktolerance to 30 and gmmaxhostdelay to 100

```
chsystem -gmlinktolerance 30
chsystem -gmmaxhostdelay 100
```

> **Test and monitor:** To reiterate, thoroughly test and carefully monitor the host impact of any changes such as these before putting them into a live production environment.

For more information about settings considerations for the **gmlinktolerance** and the **gmmaxhostdelay** parameters, see 6.3.6, “1920 error” on page 316.

rcbuffersize

The **rcbuffersize** parameter was introduced to manage workloads with intense and bursty write I/O that do not fill the internal buffer while Global Mirror writes were undergoing sequence tagging.

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

Example 6-2 shows how to change **rcbuffersize** to 64 MB by using the **chsystem** command. The default value for **rcbuffersize** is 48 MB and the maximum is 512 MB.

Example 6-2 Changing rcbuffersize to 64 MB

```
chsystem -rcbuffersize 64
```

Remember that any extra buffers you allocate are taken away from the general cache.

maxreplicationdelay and *partnershipexclusionthreshold*

`maxreplicationdelay` is a system-wide parameter that defines a maximum latency (in seconds) for individual writes that pass through the Global Mirror logic. If a write is hung for the specified amount of time, for example due to a rebuilding array on the secondary system, Global Mirror stops the relationship (and any containing consistency group), which triggers a 1920 error.

The `partnershipexclusionthreshold` parameter was introduced to allow users to set the timeout for an I/O that triggers a temporarily dropping of the link to the remote cluster. The value must be a number from 30 - 315.

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

For more information about settings considerations for the `maxreplicationdelay` parameter, see 6.3.6, “1920 error” on page 316.

Link delay simulation parameters

Even though Global Mirror is an asynchronous replication method, there can be an impact to server applications due to Global Mirror managing transactions and maintaining write order consistency over a network. To mitigate this impact, as a testing and planning feature, Global Mirror allows you to simulate the effect of the round-trip delay between sites by using the following parameters:

- `gminterclusterdelaysimulation`

  This optional parameter specifies the inter-system delay simulation, which simulates the Global Mirror round-trip delay between two systems in milliseconds. The default is 0. The valid range is 0 - 100 milliseconds.

- `gmintraclusterdelaysimulation`

  This optional parameter specifies the intrasystem delay simulation, which simulates the Global Mirror round-trip delay in milliseconds. The default is 0. The valid range is 0 - 100 milliseconds.

6.3.5 Multiple site remote copy

The most common use cases for the Remote Copy functions are obviously Disaster Recovery solutions. Code level 8.3.1 introduced further Disaster Recovery capabilities such as the *Spectrum Virtualize 3-site replication* that provides a solution for coordinated disaster recovery across three sites in various topologies. A complete discussion about the Disaster Recovery solutions based on IBM Spectrum Virtualize technology is beyond the intended scope for this book. For an overview of the Disaster Recovery solutions with the IBM Spectrum Virtualize copy services, see *IBM System Storage SAN Volume Controller and Storwize V7000 Replication Family Services*, SG24-7574. For a deepening of the 3-site replication, see *IBM Spectrum Virtualize 3-Site Replication*, SG24-8504.

Another typical Remote Copy use-case is the data movement among distant locations as required, for instance, for data center relocation and consolidation projects. In these scenarios, the IBM Spectrum Virtualize Remote Copy technology is particularly effective when combined with the image copy feature that allows data movement among storage systems of different technology or vendor.

Mirroring scenarios that involve multiple sites can be implemented using a combination of Spectrum Virtualize capabilities as described in the following sections.

Performing cascading copy service functions

Cascading copy service functions that use IBM FlashSystem are not directly supported. However, you might require a three-way (or more) replication by using copy service functions (synchronous or asynchronous mirroring). You can address this requirement both by using IBM FlashSystem copy services and by combining IBM FlashSystem copy services (with image mode cache-disabled volumes) and native storage controller copy services.

DRP limitation: Currently, the image mode VDisk is not supported with DRP.

Cascading with native storage controller copy services

Figure 6-30 describes the configuration for 3-site cascading by using the native storage controller copy services in combination with IBM FlashSystem Remote Copy functions.

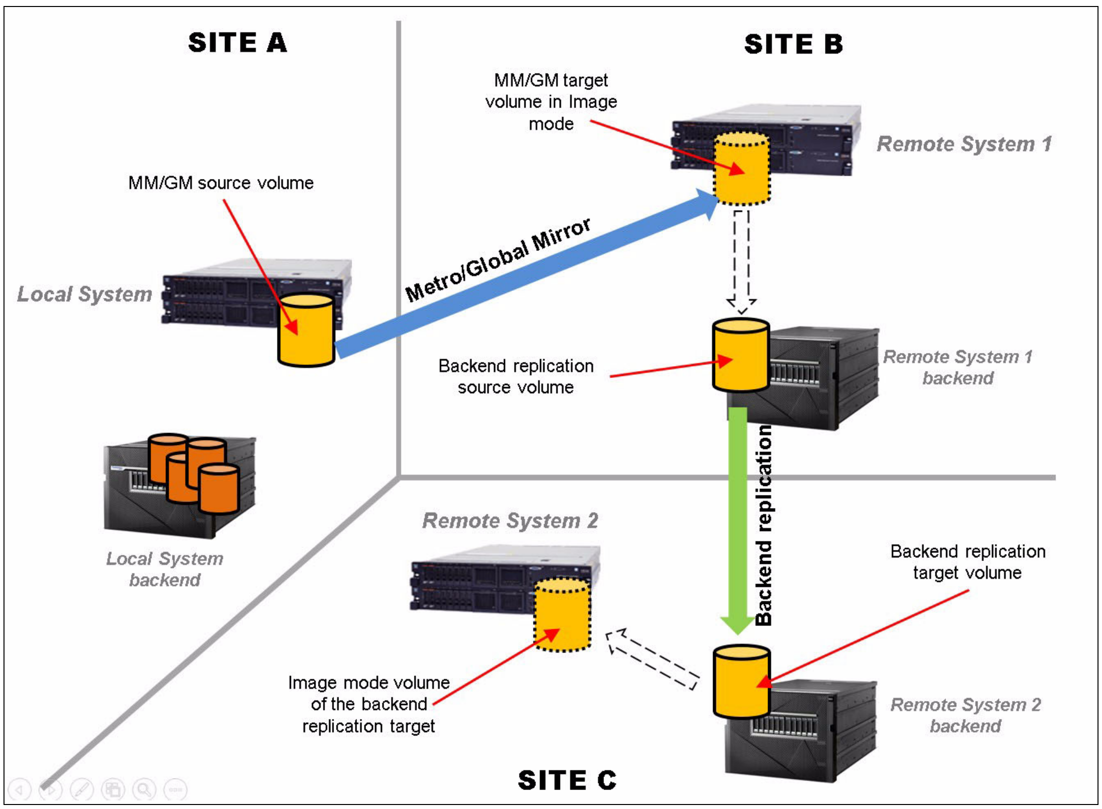


Figure 6-30 Using three-way copy services

In Figure 6-30, the primary site uses IBM FlashSystem Remote Copy functions (Global Mirror or Metro Mirror) at the secondary site. Therefore, if a disaster occurs at the primary site, the storage administrator enables access to the target volume (from the secondary site) and the business application continues processing.

While the business continues processing at the secondary site, the storage controller copy services replicate to the third site. This configuration is allowed under the following conditions:

- The back-end controller native copy services must be supported by IBM FlashSystem. For more information, see “Native back-end controller copy functions considerations” on page 305.

- The source and target volumes used by the back-end controller native copy services must be imported to the IBM FlashSystem system as image-mode volumes with the cache disabled.

Cascading with IBM FlashSystem systems copy services

Remote Copy services cascading is allowed with the Spectrum Virtualize 3-site replication capability. See *IBM Spectrum Virtualize 3-Site Replication*, SG24-8504. However, a cascading-like solution is also possible by combining the IBM FlashSystem copy services. These Remote Copy services implementations are useful in 3-site disaster recovery solutions and data center moving scenarios.

In the configuration described in Figure 6-31, a Global Mirror (Metro Mirror can also be used) solution is implemented between the Local System in Site A, the production site, and the Remote System 1 located in Site B, the primary disaster recovery site. A third system, Remote System 2, is located in Site C, the secondary disaster recovery site. Connectivity is provided between Site A and Site B, between Site B and Site C, and optionally between Site A and Site C.

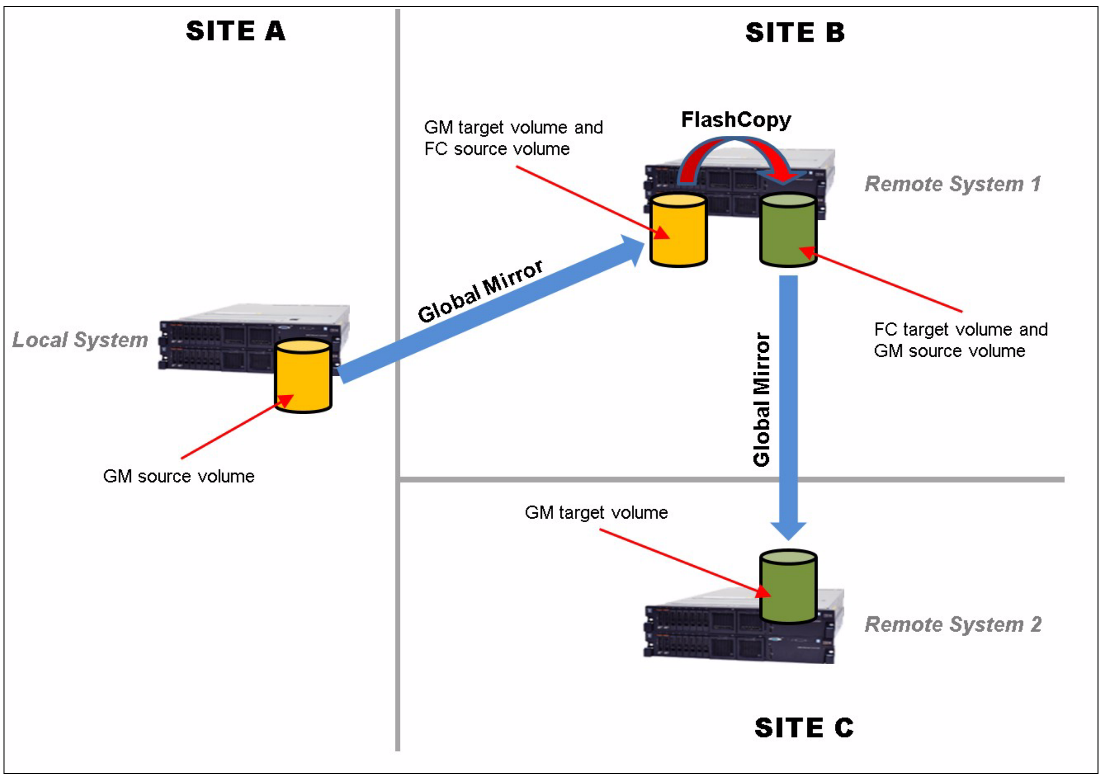


Figure 6-31 Cascading-like infrastructure

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

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

After the Global Mirror is in `ConsistentSynchronized` state, you are ready to create the cascading.

2. Consistency point creation phase. The following actions must be performed every time a consistency point creation in the Site C is required.

   a. Check whether the Global Mirror between Remote System 1 and Remote System 2 is in `stopped` or `idle` status, if it is not, stop the Global Mirror.

   b. Stop the Global Mirror between the Local System to Remote System 1.

   c. Start the FlashCopy in Remote Site 1.

   d. Resume the Global Mirror between the Local System and Remote System 1.

   e. Start/resume the Global Mirror between Remote System 1and Remote System 2.

   The first time that these operations are performed, a full copy between Remote System 1 and Remote System 2 occurs. Later executions of these operations perform incremental resynchronization instead. After the Global Mirror between Remote System 1 and Remote System 2 is in `Consistent Synchronized` state, the consistency point in Site C is created. The Global Mirror between Remote System 1 and Remote System 2 can now be stopped to be ready for the next consistency point creation.

6.3.6 1920 error

An IBM Spectrum Virtualize based system generates a 1920 error message whenever a Metro Mirror or Global Mirror relationship stops because of adverse conditions. The adverse conditions, if left unresolved, might affect performance of foreground I/O.

A 1920 error can result for many reasons. The condition might be the result of a temporary failure, such as maintenance on the inter-system connectivity, unexpectedly higher foreground host I/O workload, or a permanent error because of a hardware failure. It is also possible that not all relationships are affected and that multiple 1920 errors can be posted.

The 1920 error can be triggered for Metro Mirror *and* Global Mirror relationships. However, in Metro Mirror configurations, the 1920 error is associated with only a permanent I/O error condition. For this reason, the main focus of this section is 1920 errors in a Global Mirror configuration.

Internal Global Mirror control policy and raising 1920 errors

Although Global Mirror is an asynchronous Remote Copy service, the local and remote sites have some interplay. When data comes into a local volume, work must be done to ensure that the remote copies are consistent. This work can add a delay to the local write. Normally, this delay is low.

To mitigate the effects of the Global Mirror to the foreground I/Os, the IBM SAN Volume Controller code implements different control mechanisms for Slow I/O and Hung I/O conditions. The *Slow I/O condition* is a persistent performance degradation on write operations introduced by the Remote Copy logic; the *Hung I/O condition* is a long delay (seconds) on write operations.

Slow IO condition: gmmaxhostdelay and gmlinktolerance

The **`gmmaxhostdelay`** and **`gmlinktolerance`** parameters help 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.

In terms of nodes and back-end characteristics, the system configuration must be provisioned so that when combined, they can support the maximum throughput that is delivered by the applications at the primary that uses Global Mirror.

If the capabilities of the system configuration are exceeded, the system becomes backlogged and the hosts receive higher latencies on their write I/O. Remote Copy in Global Mirror implements a protection mechanism to detect this condition and halts mirrored foreground write and background copy I/O. Suspension of this type of I/O traffic ensures that misconfiguration or hardware problems (or both) do not affect host application availability.

Global Mirror attempts to detect and differentiate between backlogs that occur because of the operation of the Global Mirror protocol. It does not examine the general delays in the system when it is heavily loaded, where a host might see high latency even if Global Mirror were disabled.

Global Mirror uses the **gmmaxhostdelay** and **gmlinktolerance** parameters to monitor Global Mirror protocol backlogs in the following ways:

- Users set the **gmmaxhostdelay** and **gmlinktolerance** parameters to control how software responds to these delays. The **gmmaxhostdelay** parameter is a value in milliseconds with a maximum value of 100.
- Every 10 seconds, Global Mirror samples all of the Global Mirror writes and determines how much of a delay it added. If the delay added by at least a third of these writes is greater than the **gmmaxhostdelay** setting, that sample period is marked as *bad*.
- Software keeps a running count of *bad periods*. Each time that a bad period occurs, this count goes up by one. Each time a good period occurs, this count goes down by 1, to a minimum value of 0. That is for instance, ten bad periods, followed by five good periods, followed by ten bad periods, results in a bad period count of 15.
- The **gmlinktolerance** parameter is defined in seconds. Since bad periods are assessed at intervals of ten seconds, the maximum bad period count is the **gmlinktolerance** parameter value that is divided by 10. For instance, with a **gmlinktolerance** value of 300, the maximum bad period count is 30. When maximum bad period count is reached, a 1920 error is reported.

In this case, the 1920 error is identified with the specific event ID `985003` that is associated to the GM relationship, which in the last ten seconds period had the greatest accumulated time spent on delays. This event ID is generated with the text: `Remote Copy retry timeout`.

An edge case is achieved by setting the **gmmaxhostdelay** 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, ten-second sample periods. This single bad I/O theoretically can be counted as twice the number of bad periods and triggers a 1920 error.

A higher **gmlinktolerance** value, **gmmaxhostdelay** setting, or I/O load, might reduce the risk of encountering this edge case.

Hung IO condition: maxreplicationdelay and partnershipexclusionthreshold

The system-wide **maxreplicationdelay** attribute configures how long a single write can be outstanding from the host before the relationship is stopped, which triggers a 1920 error. It can protect the hosts from seeing timeouts because of secondary hung I/Os.

This parameter is mainly intended to protect from secondary system issues. It does not help with ongoing performance issues, but can be used to limit the exposure of hosts to long write response times that can cause application errors. For instance, setting **maxreplicationdelay** to 30 means that if a write operation for a volume in a Remote Copy relationship does not complete within 30 seconds, the relationship is stopped, triggering a 1920 error. This happens even if the cause of the write delay is not related to the remote copy. For this reason the **maxreplicationdelay** settings can lead to false positive1920 error triggering.

In addition to the 1920 error, the specific event ID `985004` is generated with the text: `Maximum replication delay exceeded`.

The **maxreplicationdelay** values can be 0 - 360 seconds. Setting **maxreplicationdelay** to 0 disables the feature.

The **partnershipexclusionthreshold** is a system-wide parameter that sets the timeout for an I/O that triggers a temporarily dropping of the link to the remote system. Similar to **maxreplicationdelay**, the **partnershipexclusionthreshold** attribute provides some flexibility in a part of replication that tries to shield a production system from hung I/Os on a secondary system.

To better understand the **partnershipexclusionthreshold** parameter, consider the following scenario. By default in an IBM FlashSystem 9200, a node assert (restart with a 2030 error) occurs if any individual I/O takes longer than 6 minutes. To avoid this situation, some actions are attempted to clean up anything that might be hanging I/O before the I/O reaches 6 minutes.

One of these actions is temporarily dropping (for 15 minutes) the link between systems if any I/O takes longer than 5 minutes 15 seconds (315 seconds). This action often removes hang conditions that are caused by replication problems. The **partnershipexclusionthreshold** parameter introduced the ability to set this value to a time lower than 315 seconds to respond to hung I/O more swiftly. The **partnershipexclusionthreshold** value must be a number in the range 30 - 315.

If an I/O takes longer the **partnershipexclusionthreshold** value, a 1720 error is triggered (with an event ID `987301`) and any regular Global Mirror or Metro Mirror relationships stop on the next write to the primary volume.

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

To set the **maxreplicationdelay** and **partnershipexclusionthreshold** parameters, the **chsystem** command must be used, as shown in Example 6-3.

Example 6-3 maxreplicationdelay and partnershipexclusionthreshold setting

```
IBM_FlashSystem:ITSO:superuser>chsystem -maxreplicationdelay 30
IBM_FlashSystem:ITSO:superuser>chsystem -partnershipexclusionthreshold 180
```

The **maxreplicationdelay** and **partnershipexclusionthreshold** parameters do not interact with the **gmlinktolerance** and **gmmaxhostdelay** parameters.

Troubleshooting 1920 errors

When you are troubleshooting 1920 errors that are posted across multiple relationships, you must diagnose the cause of the earliest error first. You must also consider whether other higher priority system errors exist and fix these errors because they might be the underlying cause of the 1920 error.

The diagnosis of a 1920 error is assisted by SAN performance statistics. To gather this information, you can use IBM Spectrum Control with a statistics monitoring interval of 1 or 5 minutes. Also, turn on the internal statistics gathering function, **IOstats**, in IBM FlashSystem. Although not as powerful as IBM Spectrum Control, **IOstats** can provide valuable debug information if the **snap** command gathers system configuration data close to the time of failure.

The following main performance statistics must be investigated for the 1920 error:

- Write I/O Rate and Write Data Rate

  For volumes that are primary volumes in relationships, these statistics are the total amount of write operations submitted per second by hosts on average over the sample period, and the bandwidth of those writes. For secondary volumes in relationships, this is the average number of replicated writes that are received per second, and the bandwidth that these writes consume. Summing the rate over the volumes you intend to replicate gives a coarse estimate of the replication link bandwidth required.

- Write Response Time and Peak Write Response Time

  On primary volumes, these are the average time (in milliseconds) and peak time between a write request being received from a host, and the completion message being returned. The write response time is the best way to show what kind of write performance that the host is seeing.

  If a user complains that an application is slow, and the stats show the write response time leap from 1 ms to 20 ms, the two are most likely linked. However, some applications with high queue depths and low to moderate workloads will not be affected by increased response times. Note that this being high is an effect of some other problem. The peak is less useful, as it is sensitive to individual glitches in performance, but it can show more detail of the distribution of write response times.

  On secondary volumes, these statistics describe the time for the write to be submitted from the replication feature into the system cache, and should normally be of a similar magnitude to those on the primary volume. Generally, the write response time should be below 1 ms for a fast-performing system.

- Global Mirror Write I/O Rate

  This statistic shows the number of writes per second, the (regular) replication feature is processing for this volume. It applies to both types of Global Mirror and to Metro Mirror, but in each case only for the secondary volume. Because writes are always separated into 32 KB or smaller tracks before replication, this setting might be different from the Write I/O Rate on the primary volume (magnified further because the samples on the two systems will not be aligned, so they will capture a different set of writes).

- Global Mirror Overlapping Write I/O Rate

  This statistic monitors the amount of overlapping I/O that the Global Mirror feature is handling for regular Global Mirror relationships. That is where an LBA is written again after the primary volume has been updated, but before the secondary volume has been updated for an earlier write to that LBA. To mitigate the effects of the overlapping I/Os, a journaling feature was implemented, as discussed in “Colliding writes” on page 284.

- Global Mirror secondary write lag

  This statistic is valid for regular Global Mirror primary and secondary volumes. For primary volumes, it tracks the length of time in milliseconds that replication writes are outstanding from the primary system. This amount includes the time to send the data to the remote system, consistently apply it to the secondary non-volatile cache, and send an acknowledgment back to the primary system.

For secondary volumes, this statistic records only the time that is taken to consistently apply it to the system cache, which is normally up to 20 ms. Most of that time is spent coordinating consistency across many nodes and volumes. Primary and secondary volumes for a relationship tend to record times that differ by the round-trip time between systems. If this statistic is high on the secondary system, look for congestion on the secondary system's fabrics, saturated auxiliary storage, or high CPU utilization on the secondary system.

- Write-cache Delay I/O Rate

  These statistics show how many writes could not be instantly accepted into the system cache because cache was full. It is a good indication that the write rate is faster than the storage can cope with. If this amount starts to increase on auxiliary storage while primary volumes suffer from increased Write Response Time, it is possible that the auxiliary storage is not fast enough for the replicated workload.

- Port to Local Node Send Response Time

  The time in milliseconds that it takes this node to send a message to other nodes in the same system (which will mainly be the other node in the same I/O group) and get an acknowledgment back. This amount should be well below 1 ms, with values below 0.3 ms being essential for regular Global Mirror to provide a Write Response Time below 1 ms.

  This requirement is necessary because up to three round-trip messages within the local system will happen before a write completes to the host. If this number is higher than you want, look at fabric congestion (Zero Buffer Credit Percentage) and CPU Utilization of all nodes in the system.

- Port to Remote Node Send Response Time

  This value is the time in milliseconds that it takes to send a message to nodes in other systems and get an acknowledgment back. This amount is not separated out by remote system, but for environments that have replication to only one remote system. This amount should be close to the low-level ping time between your sites. If this starts going significantly higher, it is likely that the link between your systems is saturated, which usually causes high Zero Buffer Credit Percentage as well.

- Sum of Port-to-local node send response time and Port-to-local node send queue time

  The time must be less than 1 ms for the primary system. A number in excess of 1 ms might indicate that an I/O group is reaching its I/O throughput limit, which can limit performance.

- System CPU Utilization

  These values show how heavily loaded the nodes in the system are. If any core has high utilization (say, over 90%) and there is an increase in write response time, it is possible that the workload is being CPU limited. You can resolve this by upgrading to faster hardware, or spreading out some of the workload to other nodes and systems.

- Zero Buffer Credit Percentage or Port Send Delay IO Percentage

  This is the fraction of messages that this node attempted to send through FC ports that had to be delayed because the port ran out of buffer credits. If you have a long link from the node to the switch it is attached to, there might be benefit in getting the switch to grant more buffer credits on its port.

  It is more likely to be the result of congestion on the fabric, because running out of buffer credits is how FC performs flow control. Normally, this value is well under 1%. From 1 - 10% is a concerning level of congestion, but you might find the performance acceptable. Over 10% indicates severe congestion. This amount is also called out on a port-by-port basis in the port-level statistics, which gives finer granularity about where any congestion might be.

When looking at the port-level statistics, high values on ports used for messages to nodes in the same system are much more concerning than those on ports that are used for messages to nodes in other systems.

- Back-end Write Response Time

  This value is the average response time in milliseconds for write operations to the back-end storage. This time might include several physical I/O operations, depending on the type of RAID architecture.

  Poor back-end performance on secondary system is a frequent cause of 1920 errors, while it is not so common for primary systems. Exact values to watch out for depend on the storage technology, but usually the response time should be less than 50 ms. A longer response time can indicate that the storage controller is overloaded. If the response time for a specific storage controller is outside of its specified operating range, investigate for the same reason.

Focus areas for 1920 errors

The causes of 1920 errors might be numerous. To fully understand the underlying reasons for posting this error, consider the following components that are related to the Remote Copy relationship:

- The inter-system connectivity network
- Primary storage and remote storage
- IBM FlashSystem node canisters
- Storage area network

Data collection for diagnostic purposes

A successful diagnosis depends on the collection of the following data at both systems:

- The **snap** command with **livedump** (triggered at the point of failure)
- I/O Stats that are running at operating system level (if possible)
- IBM Spectrum Control performance statistics data (if possible)
- The following information and logs from other components:
  - Inter-system network and switch details:
    - Technology
    - Bandwidth
    - Typical measured latency on the inter-system network
    - Distance on all links (which can take multiple paths for redundancy)
    - Whether trunking is enabled
    - How the link interfaces with the two SANs
    - Whether compression is enabled on the link
    - Whether the link dedicated or shared; if so, the resource and amount of those resources they use
    - Switch Write Acceleration to check with IBM for compatibility or known limitations
    - Switch Compression, which should be transparent but complicates the ability to predict bandwidth
  - Storage and application:

- Specific workloads at the time of 1920 errors, which might not be relevant, depending upon the occurrence of the 1920 errors and the volumes that are involved
- RAID rebuilds
- Whether 1920 errors are associated with Workload Peaks or Scheduled Backup

Inter-system network

For diagnostic purposes, ask the following questions about the inter-system network:

- Was network maintenance being performed?

  Consider the hardware or software maintenance that is associated with inter-system network, such as updating firmware or adding more capacity.

- Is the inter-system network overloaded?

  You can find indications of this situation by using statistical analysis with the help of I/O stats, IBM Spectrum Control, or both. Examine the internode communications, storage controller performance, or both. By using IBM Spectrum Control, you can check the storage metrics for the Global Mirror relationships were stopped, which can be tens of minutes depending on the **gmlinktolerance** and **maxreplicationdelay** parameters.

  Diagnose the overloaded link by using the following methods:

  – Look at the statistics generated by the routers or switches near your most bandwidth-constrained link between the systems

    Exactly what is provided, and how to analyze it varies depending on the equipment used.

  – Look at the port statistics for high response time in the internode communication

    An overloaded long-distance link causes high response times in the internode messages (the *Port to remote node send response time* statistic) that are sent by IBM Spectrum Virtualize. If delays persist, the messaging protocols exhaust their tolerance elasticity and the Global Mirror protocol is forced to delay handling new foreground writes while waiting for resources to free up.

  – Look at the port statistics for buffer credit starvation

    The *Zero Buffer Credit Percentage* and *Port Send Delay IO Percentage* statistic can be useful here, because you normally have a high value here as the link saturates. Only look at ports that are replicating to the remote system.

  – Look at the volume statistics (before the 1920 error is posted):

    - Target volume write throughput approaches the link bandwidth.

      If the write throughput on the target volume is equal to your link bandwidth, your link is likely overloaded. Check what is driving this situation. For example, does peak foreground write activity exceed the bandwidth, or does a combination of this peak I/O and the background copy exceed the link capacity?

    - Source volume write throughput approaches the link bandwidth.

      This write throughput represents only the I/O that is performed by the application hosts. If this number approaches the link bandwidth, you might need to upgrade the link's bandwidth. Alternatively, reduce the foreground write I/O that the application is attempting to perform, or reduce the number of Remote Copy relationships.

   - Target volume write throughput is greater than the source volume write throughput.

     If this condition exists, the situation suggests a high level of background copy and mirrored foreground write I/O. In these circumstances, decrease the background copy rate parameter of the Global Mirror partnership to bring the combined mirrored foreground I/O and background copy I/O rates back within the remote links bandwidth.

 – Look at the volume statistics (after the 1920 error is posted):

   - Source volume write throughput after the Global Mirror relationships were stopped.

     If write throughput increases greatly (by 30% or more) after the Global Mirror relationships are stopped, the application host was attempting to perform more I/O than the remote link can sustain.

     When the Global Mirror relationships are active, the overloaded remote link causes higher response times to the application host. This overload, in turn, decreases the throughput of application host I/O at the source volume. After the Global Mirror relationships stop, the application host I/O sees a lower response time, and the true write throughput returns.

     To resolve this issue, increase the remote link bandwidth, reduce the application host I/O, or reduce the number of Global Mirror relationships.

Storage controllers

Investigate the primary and remote storage controllers, starting at the remote site. If the back-end storage at the secondary system is overloaded, or another problem is affecting the cache there, the Global Mirror protocol fails to keep up. Similarly, the problem exhausts the (**gmlinktolerance**) elasticity and has a similar effect at the primary system.

In this situation, ask the following questions:

- Are the storage controllers at the remote system overloaded (performing slowly)?

  Use IBM Spectrum Control to obtain the back-end write response time for each MDisk at the remote system. A response time for any individual MDisk that exhibits a sudden increase of 50 ms or more, or that is higher than 100 ms, generally indicates a problem with the back-end. In case of 1920 error triggered by the “max replication delay exceeded” condition, check the peek back-end write response time to see if it has exceeded the **maxreplicationdelay** value around the 1920 occurrence.

  Check whether an error condition is on the internal 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 cache settings of the controller and checking that the cache batteries are healthy, if applicable.

- 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. In case of 1920 error triggered by the “max replication delay exceeded” condition, check the peek back-end write response time to see if it has exceeded the **maxreplicationdelay** value around the 1920 occurrence.

  However, if bad performance continues for a prolonged period, a false 1920 error might be flagged.

Node canister

For IBM FlashSystem node canister hardware, the possible cause of the 1920 errors might be from a heavily loaded secondary or primary system. If this condition persists, a 1920 error might be posted.

Global Mirror needs to synchronize its I/O processing across all nodes in the system to ensure data consistency. If any node is running out of CPU, it can affect all relationships. So check the CPU cores usage statistic. If it looks higher when there is a performance problem, then running out of CPU bandwidth might be causing the problem. Of course, CPU usage goes up when the IOPS going through a node goes up, so if the workload increases, you would expect to see CPU usage increase.

If there is an increase in CPU usage on the secondary system but no increase in IOPS, and volume write latency increases too, it is likely that the increase in CPU usage has caused the increased volume write latency. In that case, try to work out what might have caused the increase in CPU usage (for example, starting many FlashCopy mappings). Consider moving that activity to a time with less workload. If there is an increase in both CPU usage and IOPS, and the CPU usage is close to 100%, then that node might be overloaded. A *Port-to-local node send queue time* value higher than 0.2 ms often denotes CPU cores overloading.

In a primary system, if it is sufficiently busy, the write ordering detection in Global Mirror can delay writes enough to reach a latency of **gmmaxhostdelay** and cause a 1920 error. Stopping replication potentially lowers CPU usage, and also lowers the opportunities for each I/O to be delayed by slow scheduling on a busy system.

Solve overloaded nodes by upgrading them to newer, faster hardware if possible, or by adding more I/O groups/control enclosures (or systems) to spread the workload over more resources.

Storage area network

Issues and congestions both in local and remote SANs can lead to 1920 errors. The *Port to local node send response time* is the key statistic to investigate on. It captures the round-trip time between nodes in the same system. Anything over 1.0 ms is surprisingly high, and will cause high secondary volume write response time. Values greater than 1 ms on primary system will cause an impact on write latency to Global Mirror primary volumes of 3 ms or more.

If you have checked CPU core utilization on all the nodes, and it has not gotten near 100%, a high *Port to local node send response time* means that there is fabric congestion or a slow-draining FC device.

A good indicator of SAN congestion is the Zero Buffer Credit Percentage and Port Send Delay IO Percentage on the port statistics. For more information about buffer credit, see “Buffer credits” on page 297.

If a port has more than 10% zero buffer credits, that will definitely cause a problem for all I/O, not just Global Mirror writes. Values from 1 - 10% are moderately high and might contribute to performance issues.

For both primary and secondary systems, congestion on the fabric from other slow-draining devices becomes much less of an issue when only dedicated ports are used for node-to-node traffic within the system. However, this only really becomes an option on systems with more than four ports per node. Use port masking to segment your ports.

FlashCopy considerations

Check that FlashCopy mappings are in the *prepared* state. Check whether the Global Mirror target volumes are the sources of a FlashCopy mapping and whether that mapping was in the *prepared* state for an extended time.

Volumes in the prepared state are cache disabled, so their performance is impacted. To resolve this problem, start the FlashCopy mapping, which re-enables the cache and improves the performance of the volume and of the Global Mirror relationship.

Consider also that FlashCopy can add significant workload to the back-end storage, especially when the background copy is active (see “Background copy considerations” on page 269). In cases where the remote system is used to create golden or practice copies for Disaster Recovery testing, the workload added by the FlashCopy background processes can overload the system. This overload can lead to poor Remote Copy performances and then to a 1920 error, even though with IBM FlashSystem this is less of an issue because of high-performing flash back-end.

Careful planning of the back-end resources is particularly important with these kinds of scenarios. Reducing the FlashCopy background copy rate can also help to mitigate this situation. Furthermore, note that the FlashCopy copy-on-write process adds some latency by delaying the write operations on the primary volumes until the data is written to the FlashCopy target.

This process does mot directly affect the Remote Copy operations because it is logically placed below the Remote Copy processing in the I/O stack, as shown in Figure 6-7 on page 258. Nevertheless, in some circumstances, especially with write intensive environments, the copy-on-write process tends to stress some of the internal resources of the system, such as CPU and memory. This condition can also affect the remote copy that competes for the same resources, eventually leading to 1920 errors.

FCIP considerations

When you get a 1920 error, always check the latency first. The FCIP routing layer can introduce latency if it is not properly configured. If your network provider reports a much lower latency, you might have a problem at your FCIP routing layer. Most FCIP routing devices have built-in tools to enable you to check the RTT. When you are checking latency, remember that TCP/IP routing devices (including FCIP routers) report RTT by using standard 64-byte ping packets.

In Figure 6-32 on page 326, you can see why the effective transit time must be measured only by using packets that are large enough to hold an FC frame, or 2148 bytes (2112 bytes of payload and 36 bytes of header). Allow estimated resource requirements to be a safe amount because various switch vendors have optional features that might increase this size. After you verify your latency by using the proper packet size, proceed with normal hardware troubleshooting.

Look at the second largest component of your RTT, which is *serialization delay*. Serialization delay is the amount of time that is required to move a packet of data of a specific size across a network link of a certain bandwidth. The required time to move a specific amount of data decreases as the data transmission rate increases.

Figure 6-32 shows the orders of magnitude of difference between the link bandwidths. It is easy to see how `1920` errors can arise when your bandwidth is insufficient. Never use a TCP/IP ping to measure RTT for FCIP traffic.

Packet Size	Link Size	Serialization Delay (Time Required to Send Data)	Unit
64	256 Kbps	2.0E+03	microseconds
64	1.5 Mbps	3.4E+02	microseconds
64	100 Mbps	5.1E+00	microseconds
64	155 Mbps	3.3E+00	microseconds
64	622 Mbps	8.2E-01	microseconds
64	1 Gbps	5.1E-04	microseconds
64	10 Gbps	5.1E-05	microseconds
1500	256 Kbps	4.7E+04	microseconds
1500	1.5 Mbps	8.0E+03	microseconds
1500	100 Mbps	1.2E+02	microseconds
1500	155 Mbps	7.7E+01	microseconds
1500	622 Mbps	1.9E+01	microseconds
1500	1 Gbps	1.2E+01	microseconds
1500	10 Gbps	1.2E+00	microseconds
2148	256 Kbps	6.7E+04	microseconds
2148	1.5 Mbps	1.1E+04	microseconds
2148	100 Mbps	1.7E+02	microseconds
2148	155 Mbps	1.1E+02	microseconds
2148	622 Mbps	2.8E+01	microseconds
2148	1 Gbps	1.7E+01	microseconds
2148	10 Gbps	1.7E-03	microseconds

Figure 6-32 Effect of packet size (in bytes) versus the link size

In Figure 6-32, the amount of time in microseconds that is required to transmit a packet across network links of varying bandwidth capacity is compared. The following packet sizes are used:

- 64 bytes: The size of the common ping packet
- 1500 bytes: The size of the standard TCP/IP packet
- 2148 bytes: The size of an FC frame

Finally, your path maximum transmission unit (MTU) affects the delay that is incurred to get a packet from one location to another location. An MTU might cause fragmentation, or be too large and cause too many retransmits when a packet is lost.

Hung I/O

An hung I/O condition is reached when a write operation is delayed in the IBM Spectrum Virtualize stack for a significant time (typically seconds). This condition is monitored by IBM Spectrum Virtualize, which eventually leads to a 1920 error if the delay is higher than maxreplicationdelay settings.

Hung I/Os can be caused by many factors, such as back-end performance, cache fullness, internal resource starvation, and remote copy issues. When the `maxreplicationdelay` setting triggers a 1920 error, the following areas must be investigated:

- Inter-site network disconnections. This kind of event generates partnership instability, which leads the mirrored write operations to be delayed until the condition is resolved.

- Secondary system poor performance. In the case of bad performance, the secondary system can become virtually unresponsive, which delays the replica of the write operations.
- Primary or secondary system node warmstarts. During a node warmstart, the system freezes all the I/Os for few seconds to get a consistent state of the cluster resources. These events often are not directly related to the remote copy operations.

> **Note:** The `maxreplicationdelay` trigger can occur, even if the cause of the write delay is not related to the remote copy. In this case, the replication suspension does not resolve the hung I/O condition.
>
> To exclude the remote copy as cause of the hung I/O, the duration of the delay (peek write response time) can be checked by using tools, such as Spectrum Control. If the measured delay is greater than the `maxreplicationdelay` settings, it is unlikely that the remote copy is responsible.

Recovery after 1920 errors

After a 1920 error occurs, the Global Mirror auxiliary volumes are no longer in a `Consistent Synchronized` state. You must establish the cause of the problem and fix it before you restart the relationship.

When the relationship is restarted, you must resynchronize it. During this period, the data on the Metro Mirror or Global Mirror auxiliary volumes on the secondary system is inconsistent, and your applications cannot use the volumes as backup disks. To address this data consistency exposure on the secondary system, a FlashCopy of the auxiliary volumes can be created to maintain a consistent image until the Global Mirror (or the Metro Mirror) relationships are synchronized again and back in a consistent state.

IBM Spectrum Virtualize provides the Remote Copy *Consistency Protection* feature that automates this process. When Consistency Protection is configured, the relationship between the primary and secondary volumes does not go in to the `Inconsistent copying` status after it is restarted. Instead, the system uses a secondary *change volume* to automatically copy the previous consistent state of the secondary volume.

The relationship automatically moves to the `Consistent copying` status as the system resynchronizes and protects the consistency of the data. The relationship status changes to `Consistent synchronized` when the resynchronization process completes.

For more information about the Consistency Protection feature, see *Implementing the IBM SAN Volume Controller with IBM Spectrum Virtualize Version 8.4.2*, SG24-8507.

To ensure that the system can handle the background copy load, delay restarting the Metro Mirror or Global Mirror relationship until a quiet period occurs. If the required link capacity is unavailable, you might experience another 1920 error, and the Metro Mirror or Global Mirror relationship might stop in an inconsistent state.

Copy services tools, like IBM Copy Services Manager (CSM), or manual scripts can be used to automatize the relationships to restart after a 1920 error. CSM implements a logic to avoid recurring restart operations in case of a persistent problem. CSM attempts an automatic restart for every occurrence of 1720/1920 error a certain number of times (determined by the **`gmlinktolerance`** value) within a 30 minute time period.

If the number of allowable automatic restarts is exceeded within the time period, CSM will not automatically restart GM on the next 1720/1920 error. Furthermore, with CSM it is possible to specify the amount of time, in seconds, in which the tool will wait after an 1720/1920 error before automatically restarting the GM. For more information, see this IBM Documentation web page.

> **Tip:** When implementing automatic restart functions, it is advised to preserve the data consistency on GM target volumes during the resychronization by using features, such as FlashCopy or Consistency Protection.

Adjusting the Global Mirror settings

Although the default values are valid in most configurations, the settings of the **gmlinktolerance** and **gmmaxhostdelay** can be adjusted to accommodate particular environment or workload conditions.

For example, Global Mirror is designed to look at average delays. However, some hosts such as VMware ESX might not tolerate a single I/O getting old, for example, 45 seconds, before it decides to reboot. Given that it is better to terminate a Global Mirror relationship than it is to reboot a host, you might want to set **gmlinktolerance** to something like 30 seconds and then compensate so that you do not get too many relationship terminations by setting **gmmaxhostdelay** to something larger, such as 100 ms.

If you compare the two approaches, the default (**gmlinktolerance 300**, **gmmaxhostdelay 5**) is a rule that “If more than one third of the I/Os are slow and that happens repeatedly for 5 minutes, then terminate the busiest relationship in that stream.” In contrast, the example of **gmlinktolerance 30**, **gmmaxhostdelay 100** is a rule that “If more than one third of the I/Os are extremely slow and that happens repeatedly for 30 seconds, then terminate the busiest relationship in the stream.”

So one approach is designed to pick up general slowness, and the other approach is designed to pick up shorter bursts of extreme slowness that might disrupt your server environment. The general recommendation is to change the **gmlinktolerance** and **gmmaxhostdelay** values progressively and evaluate the overall impact to find an acceptable compromise between performances and Global Mirror stability.

You can even disable the **gmlinktolerance** feature by setting the **gmlinktolerance** value to 0. However, the **gmlinktolerance** parameter cannot protect applications from extended response times if it is disabled. You might consider disabling the **gmlinktolerance** feature in the following circumstances:

- During SAN maintenance windows, where degraded performance is expected from SAN components and application hosts can withstand extended response times from Global Mirror volumes.
- During periods when application hosts can tolerate extended response times and it is expected that the **gmlinktolerance** feature might stop the Global Mirror relationships. For example, you are testing usage of an I/O generator that is configured to stress the back-end storage. Then, the **gmlinktolerance** feature might detect high latency and stop the Global Mirror relationships. Disabling the **gmlinktolerance** parameter stops the Global Mirror relationships at the risk of exposing the test host to extended response times.

Another tunable parameter that interacts with the GM is the **maxreplicationdelay**. Note that the **maxreplicationdelay** settings do not mitigate the 1920 error occurrence because it actually adds a trigger to the 1920 error itself. However, the **maxreplicationdelay** provides users with a fine granularity mechanism to manage the hung I/Os condition and it can be used in combination with **gmlinktolerance** and **gmmaxhostdelay** settings to better address particular environment conditions.

In the above VMware example, an alternative option is to set the **maxreplicationdelay** to 30 seconds and leave the **gmlinktolerance** and **gmmaxhostdelay** settings to their default. With these settings, the **maxreplicationdelay** timeout effectively handles the hung I/Os conditions, while the **gmlinktolerance** and **gmmaxhostdelay** settings still provide an adequate mechanism to protect from ongoing performance issues.

6.4 Native IP replication

The native IP replication feature enables replication between any IBM Spectrum Virtualize products by using the built-in networking ports or optional 1/10 Gb adapter.

Native IP replication uses SANslide technology developed by Bridgeworks Limited of Christchurch, UK. They specialize in products that can bridge storage protocols and accelerate data transfer over long distances. Adding this technology at each end of a wide area network (WAN) TCP/IP link significantly improves the utilization of the link.

This technology improves the link utilization by applying patented artificial intelligence (AI) to hide latency that is normally associated with WANs. Doing so can greatly improve the performance of mirroring services, in particular GMCV over long distances.

6.4.1 Native IP replication technology

Remote Mirroring over IP communication is supported on the IBM FlashSystem systems by using Ethernet communication links. The IBM Spectrum Virtualize Software IP replication uses innovative *Bridgeworks SANSlide* technology to optimize network bandwidth and utilization. This new function enables the use of a lower-speed and lower-cost networking infrastructure for data replication.

Bridgeworks' SANSlide technology, which is integrated into the IBM Spectrum Virtualize Software, uses artificial intelligence to help optimize network bandwidth use and adapt to changing workload and network conditions. This technology can improve remote mirroring network bandwidth usage up to three times. It can enable clients to deploy a less costly network infrastructure, or speed up remote replication cycles to enhance disaster recovery effectiveness.

With an Ethernet network data flow, the data transfer can slow down over time. This condition occurs because of the latency that is caused by waiting for the acknowledgment of each set of packets that are sent. The next packet set cannot be sent until the previous packet is acknowledged, as shown in Figure 6-33 on page 330.

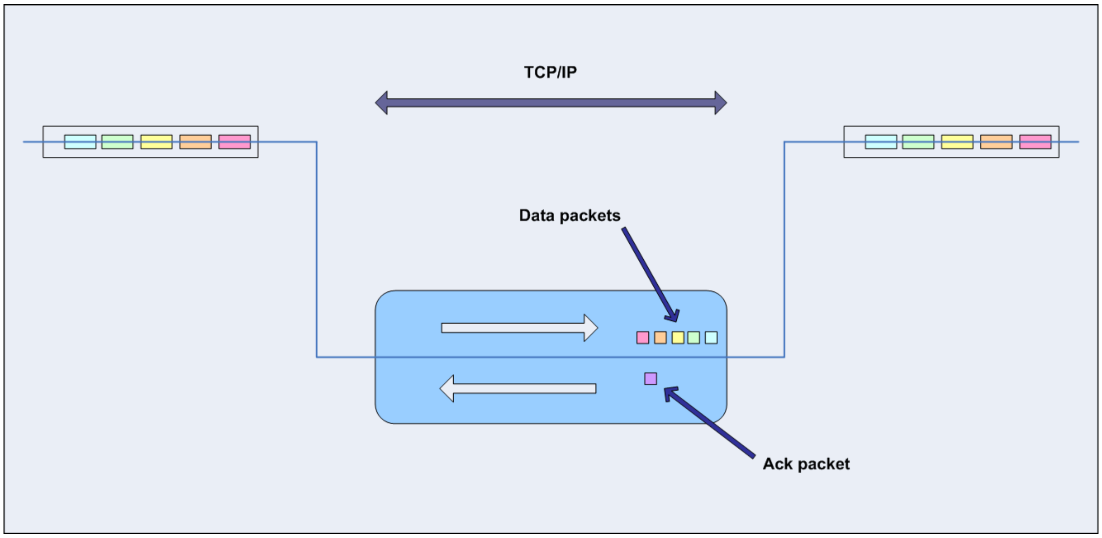


Figure 6-33 Typical Ethernet network data flow

However, by using the embedded IP replication, this behavior can be eliminated with the enhanced parallelism of the data flow. This parallelism uses multiple virtual connections (VCs) that share IP links and addresses.

The artificial intelligence engine can dynamically adjust the number of VCs, receive window size, and packet size as appropriate to maintain optimum performance. While the engine is waiting for one VC's ACK, it sends more packets across other VCs. If packets are lost from any VC, data is automatically retransmitted, as shown in Figure 6-34.

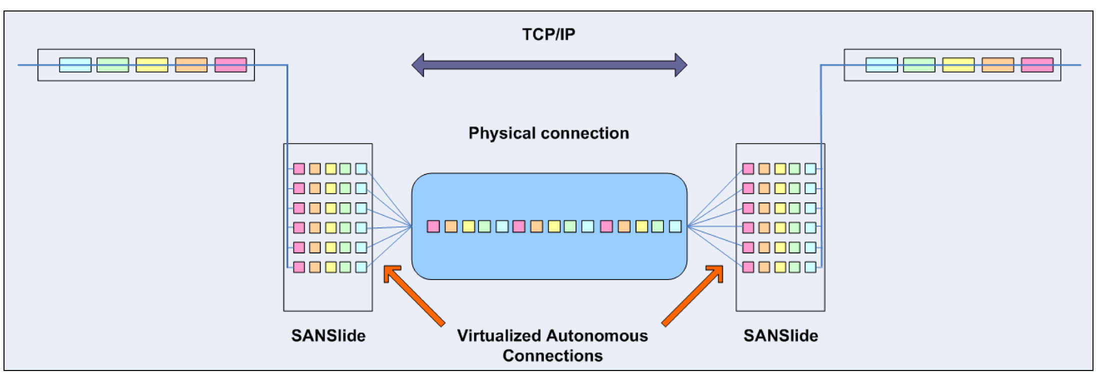


Figure 6-34 Optimized network data flow by using Bridgeworks SANSlide technology

For more information about this technology, see *IBM SAN Volume Controller and Storwize Family Native IP Replication*, REDP-5103.

Metro Mirror, Global Mirror, and Global Mirror Change Volume are supported with native IP partnership.

6.4.2 IP partnership limitations

The following prerequisites and assumptions must be considered before IP partnership between two IBM Spectrum Virtualize based systems can be established:

- The systems have 7.2 or later code levels.
- The systems have the necessary licenses that enable Remote Copy partnerships to be configured between two systems. A separate license is not required to enable IP partnership.
- The storage SANs are configured correctly and the correct infrastructure to support the systems in Remote Copy partnerships over IP links is in place.
- The two systems must be able to ping each other and perform the discovery.
- The maximum number of partnerships between the local and remote systems, including IP *and* FC partnerships, is limited to the current maximum that is supported, which is three partnerships (four systems total).

  > **Note:** With code versions earlier than 8.4.2, only a single partnership over IP is supported.

- A system can have simultaneous partnerships over FC and IP, but with separate systems. The FC zones between two systems must be removed before an IP partnership is configured.
- The use of WAN-optimization devices, such as Riverbed, is not supported in IP partnership configurations containing SAN Volume Controller.
- IP partnerships are supported with 25, 10, and 1 Gbps links. However, the intermix /on a single link is not supported.
- The maximum supported round-trip time is 80 ms for 1 Gbps links.
- The maximum supported round-trip time is 10 ms for 25 and 10 Gbps links.
- The minimum supported link bandwidth is 10 Mbps.
- The inter-cluster heartbeat traffic uses 1 Mbps per link.
- Migrations of Remote Copy relationships directly from FC-based partnerships to IP partnerships are not supported.
- IP partnerships between the two systems can be over either IPv4 or IPv6, but not both.
- Virtual LAN (VLAN) tagging of the IP addresses that are configured for Remote Copy is supported.
- Management IP and Internet SCSI (iSCSI) IP on the same port can be in a different network.
- An added layer of security is provided by using Challenge Handshake Authentication Protocol (CHAP) authentication.
- Direct attached systems configurations are supported with the following restrictions:
  - Only two direct attach link are allowed.
  - The direct attach links must be on the same I/O group.
  - Use two portsets, where a portset contains only the two ports that are directly linked.
- Transmission Control Protocol (TCP) ports 3260 and 3265 are used for IP partnership communications. Therefore, these ports must be open in firewalls between the systems.
- Network address translation (NAT) between systems that are being configured in an IP Partnership group is not supported.

- Only one Remote Copy data session per portset can be established. It is intended that only one connection (for sending or receiving Remote Copy data) is made for each independent physical link between the systems.

  > **Note:** A physical link is the physical IP link between the two sites, A (local) and B (remote). Multiple IP addresses on local system A can be connected (by Ethernet switches) to this physical link. Similarly, multiple IP addresses on remote system B can be connected (by Ethernet switches) to the same physical link. At any point, only a single IP address on cluster A can form an RC data session with an IP address on cluster B.

- The maximum throughput is restricted based on the use of 1 Gbps or 10 Gbps Ethernet ports. The output varies based on distance (for example, round-trip latency) and quality of communication link (for example, packet loss). The following maximum throughputs are achievable:
  - One 1 Gbps port can transfer up to 120 MB
  - One 10 Gbps port can transfer up to 600 MB

Table 6-9 lists the IP replication limits.

Table 6-9 IP replication limits

Remote Copy property	Maximum	Apply to	Comment
Inter-system IP partnerships per system	3 MBps	All models	A system can be partnered with up to three remote systems.
Inter-site links per IP partnership	2 MBps	All models	A maximum of two inter-site links can be used between two IP partnership sites.
Ports per node	1 MBps	All models	A maximum of one port per node can be used for IP partnership
IP partnership Software Compression Limit	140 MBps	All models	

6.4.3 VLAN support

VLAN tagging is supported for iSCSI host attachment and IP replication. Hosts and remote-copy operations can connect to the system through Ethernet ports. Each traffic type has different bandwidth requirements, which can interfere with each other if they share a port.

VLAN tagging creates two separate connections on the same IP network for different types of traffic. The system supports VLAN configuration on both IPv4 and IPv6 connections.

When the VLAN ID is configured for the IP addresses that are used for iSCSI host attach or IP replication, the suitable VLAN settings on the Ethernet network and servers must be configured correctly to avoid connectivity issues. After the VLANs are configured, changes to the VLAN settings disrupt iSCSI and IP replication traffic to and from the partnerships.

During the VLAN configuration for each IP address, the VLAN settings for the local and failover ports on two nodes of an I/O Group can differ. To avoid any service disruption, switches must be configured so that the failover VLANs are configured on the local switch ports and the failover of IP addresses from a failing node to a surviving node succeeds.

If failover VLANs are not configured on the local switch ports, no paths are available to the Spectrum Virtualize system during a node failure and the replication fails.

Consider the following requirements and procedures when implementing VLAN tagging:

- VLAN tagging is supported for IP partnership traffic between two systems.
- VLAN provides network traffic separation at the layer 2 level for Ethernet transport.
- VLAN tagging by default is disabled for any IP address of a node port. You can use the CLI or GUI to set the VLAN ID for port IPs on both systems in the IP partnership.
- When a VLAN ID is configured for the port IP addresses that are used in Remote Copy port groups, appropriate VLAN settings on the Ethernet network must also be properly configured to prevent connectivity issues.

Setting VLAN tags for a port is disruptive. Therefore, VLAN tagging requires that you stop the partnership first before you configure VLAN tags. Then, restart again when the configuration is complete.

6.4.4 IP compression

IBM FlashSystem can leverage the IP compression capability to speed up replication cycles or to reduce bandwidth utilization.

This feature reduces the volume of data that must be transmitted during Remote Copy operations by using compression capabilities similar to those experienced with existing Real-time Compression implementations.

> **No License:** The IP compression feature does not require an RtC software license.

The data compression is made within the IP replication component of the IBM Spectrum Virtualize code. It can be used with all the Remote Copy technology (Metro Mirror, Global Mirror, and GMCV). The IP compression feature provides two kinds of compression mechanisms: the hardware compression and software compression.

The IP compression can be enabled on hardware configurations that support hardware-assisted compression acceleration engines. The hardware compression is active when compression accelerator engines are available, otherwise software compression is used.

Hardware compression makes use of currently underused compression resources. The internal resources are shared between data and IP compression. Software compression uses the system CPU and might have an impact on heavily used systems.

To evaluate the benefits of the IP compression, the Comprestimator tool can be used to estimate the compression ratio of the data to be replicated. The IP compression can be enabled and disabled without stopping the Remote Copy relationship by using the **`mkippartnership`** and **`chpartnership`** commands with the **`-compress`** parameter. Furthermore, in systems with replication enabled in both directions, the IP compression can be enabled in only one direction. IP compression is supported for IPv4 and IPv6 partnerships.

6.4.5 Replication portsets

This section describes the replication portsets and different ways to configure the links between the two remote systems. Two systems can be connected over one link or, at most, two links. To address the requirement to enable the systems to know about the physical links between the two sites, the concept of a portset is used.

Portsets are groupings of logical addresses that are associated with the specific traffic types. IBM Spectrum Virtualize supports portsets for host attachment (iSCSI or iSER), back-end storage connectivity (iSCSI only), and IP replication. Each physical Ethernet Port can have maximum 64 IP addresses with each IP on unique portset.

A *portset object* is a system-wide object that might contain IP addresses from every I/O group. Figure 6-35 shows a sample of portsets definition across the canister ports in a two IO group clustered FlashSystem.

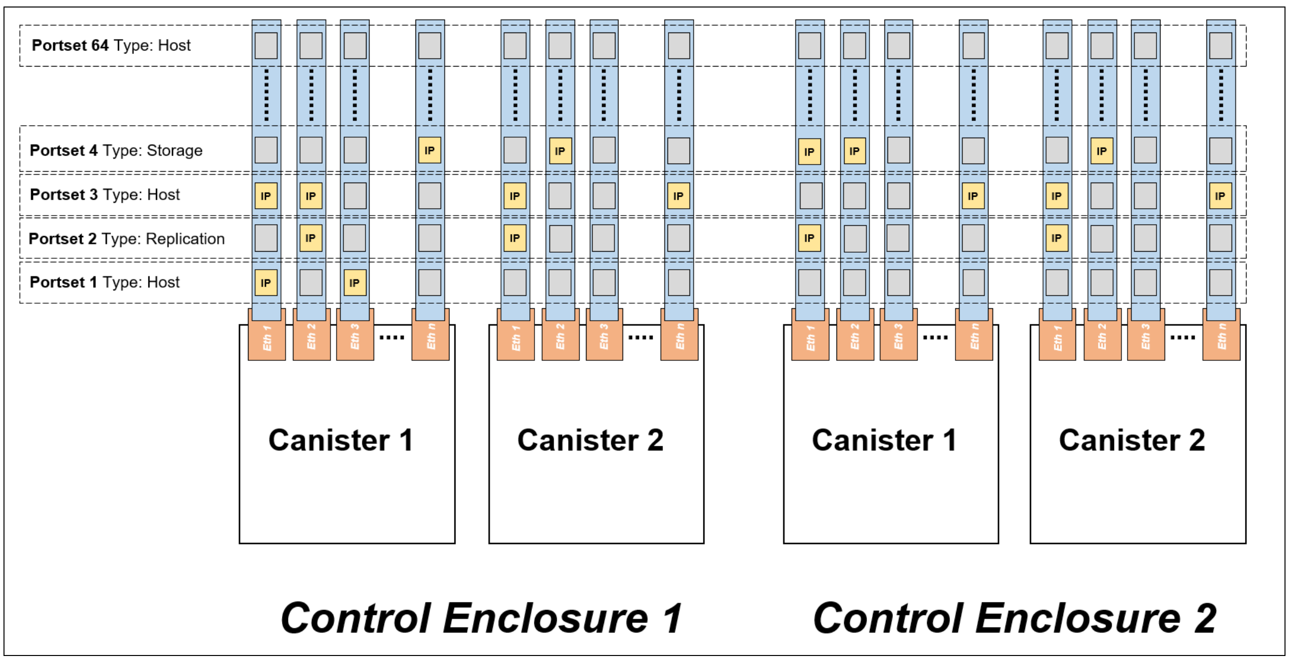


Figure 6-35 Portsets

Complete the following steps to establish an IP partnership between two systems:

1. Identify the Ethernet ports to be used for the IP replication.
2. Define a replication type portset.
3. Set the IP addresses to the identified ports and add them to the portset.
4. Create the IP partnership from both systems specifying the portset to be used.

Multiple FlashSystem canisters can be connected to the same physical long-distance link by setting IP addresses in the same portset. Samples of supported configurations are described in 6.4.6, “Supported configurations examples” on page 336.

In scenarios with two physical links between the local and remote clusters, two separate replication portset must be used to designate which IP addresses are connected to which physical link. The relationship between the physical links and the replication portsets is not monitored by the IBM Spectrum Virtualize code. Therefore, two different replication portsets can be used with a single physical link and vice versa.

All IP addresses in a replication portset must be IPv4 *or* IPv6 addresses (IP types cannot be mixed). IP addresses can be shared among replication and host type portsets, although it is not recommended.

> **Note:** The concept of a portset was introduced in IBM Spectrum Virtualize version 8.4.2 and the IP Multi-tenancy feature. Versions before 8.4.2 use the Remote Copy Port Groups concept to tag the IP addresses to associate with an IP partnership. For more information about the Remote Copy Port Group configuration, see this IBM Documentation web page.
>
> When upgrading to version 8.4.2, an automatic process occurs to convert the Remote Copy Port Groups configuration to an equivalent replication portset configuration.

Failover operations within and between portsets

Within one portset, only one IP from each system is selected for sending and receiving Remote Copy data at any one time. Therefore, on each system, at most one IP for each portset group is reported as `used`.

If the IP partnership cannot continue over an IP, the system fails over to another IP within that portset. Some reasons this issue might occur include the switch to which it is connected fails, the node goes offline, or the cable that is connected to the port is unplugged.

For the IP partnership to continue during a failover, multiple ports must be configured within the portset. If only one link is configured between the two systems, configure at least two IPs (one per node) within the portset. You can configure these two IPs on two nodes within the same I/O group or within separate I/O groups.

While failover is in progress, no connections in that portset exist between the two systems in the IP partnership for a short time. Typically, failover completes within 30 seconds to 1 minute. If the systems are configured with two portsets, the failover process within each portset continues independently of each other.

The disadvantage of configuring only one link between two systems is that, during a failover, a discovery is initiated. When the discovery succeeds, the IP partnership is reestablished. As a result, the relationships might stop, in which case a manual restart is required. To configure two inter-system links, you must configure two replication type portsets.

When a node fails in this scenario, the IP partnership can continue over the other link until the node failure is rectified. Failback then occurs when both links are again active and available to the IP partnership. The discovery is triggered so that the active IP partnership data path is made available from the new IP address.

In a two-node system, or more than one I/O Group exists and the node in the other I/O group has IP addresses within the replication portset, the discovery is triggered. The discovery makes the active IP partnership data path available from the new IP address.

6.4.6 Supported configurations examples

Different IP replication topologies are available depending on the number of physical links, the number of nodes and the number of IP partnerships. In the following sections, some typical configurations are described.

Single partnership configurations

In this section, some single partnership configurations are described.

Single inter-site link configurations

Consider single control enclosure systems in IP partnership over a single inter-site link (with failover ports configured), as shown in Figure 6-36.

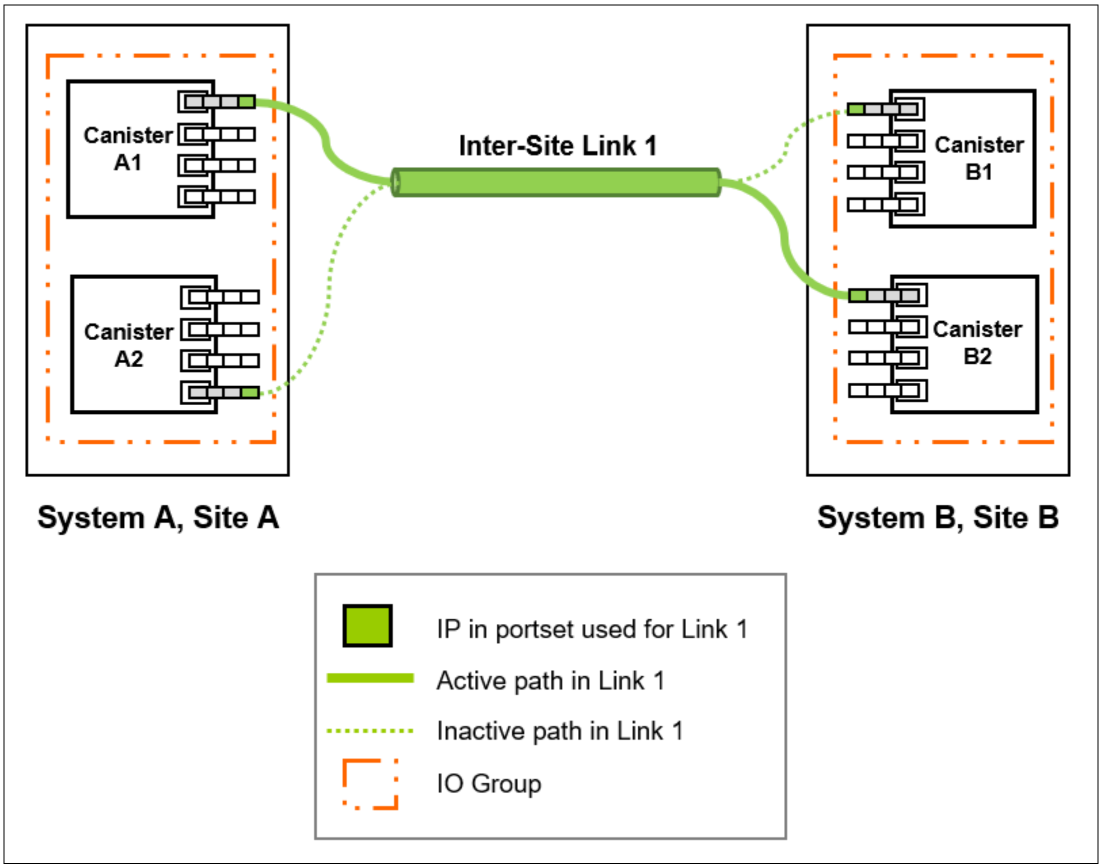


Figure 6-36 Only one link on each system and canister with failover ports configured

Figure 6-36 shows two systems: System A and System B. A single portset is used with IP addressees on two Ethernet ports, one each on Canister A1 and Canister A2 on System A. Similarly, a single portset is configured on two Ethernet ports on Canister B1 and Canister B2 on System B.

Although two ports on each system are configured in the portset, only one Ethernet port in each system actively participates in the IP partnership process. This selection is determined by a path configuration algorithm that is designed to choose data paths between the two systems to optimize performance.

The other port on the partner canister in the control enclosure behaves as a standby port that is used during a canister failure. If Canister A1 fails in System A, IP partnership continues servicing replication I/O from Ethernet Port 2 because a failover port is configured on Canister A2 on Ethernet Port 2.

However, it might take some time for discovery and path configuration logic to reestablish paths post failover. This delay can cause partnerships to change to `Not_Present` for that time. The details of the particular IP port that is actively participating in IP partnership is provided in the **`lspartnership`** output (reported as `link1_ip_id` and `link2_ip_id`).

This configuration has the following characteristics:

- Each canister in the control enclosure has ports with IP addresses defined in the same replication type portset. However, only one path is active at any time at each system.
- If Canister A1 in System A or Canister B2 in System B fails in the respective systems, IP partnerships rediscovery is triggered and continues servicing the I/O from the failover port.
- The discovery mechanism that is triggered because of failover might introduce a delay where the partnerships momentarily change to the `Not_Present` state and recover.

A four control enclosures system in IP partnership with two control enclosure system over single inter-site link is shown in Figure 6-37.

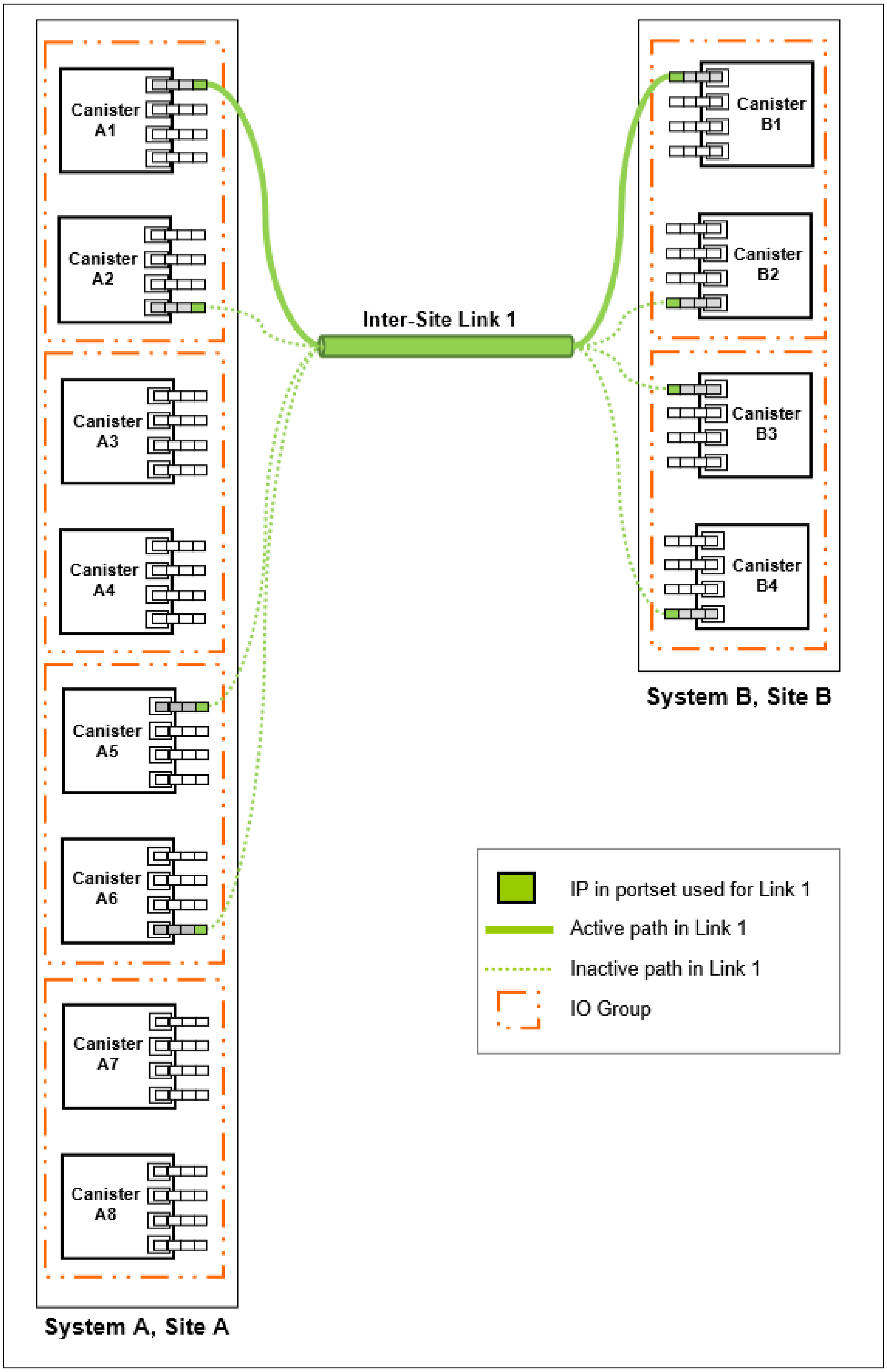


Figure 6-37 Configuration 2: clustered systems single inter-site link with only one link

Figure 6-37 on page 338 shows a four control enclosure system (System A in Site A) and a two control enclosure system (System B in Site B). A single replication portset is used on canisters A1, A2, A5, and A6 on System A at Site A. Similarly, a single portset is used on canisters B1, B2, B3, and B4 on System B.

Although four control enclosures are in System A, only two control enclosure are configured for IP partnerships. Port selection is determined by a path configuration algorithm. The other ports play the role of standby ports.

If Canister A1 fails in System A, IP partnership continues using one of the ports that is configured in the portset from any of the canisters from either of the two control enclosures in System A.

However, it might take some time for discovery and path configuration logic to reestablish paths post-failover. This delay might cause partnerships to change to the `Not_Present` state. This process can lead to Remote Copy relationships stopping. The administrator must manually start them if the relationships do not auto-recover.

The details of which particular IP port is actively participating in IP partnership process is provided in **`lspartnership`** output (reported as `link1_ip_id` and `link2_ip_id`).

This configuration includes the following characteristics:

- The replication portset that is used contains IPs from canisters of all the control enclosures. However, only one path is active at any time at each system.
- If the Canister A1 in System A or the Canister B2 in System B fails in the system, the IP partnerships trigger discovery and continue servicing the I/O from the failover ports.
- The discovery mechanism that is triggered because of failover might introduce a delay where the partnerships momentarily change to the `Not_Present` state and then recover.
- The bandwidth of the single link is used completely.

Two inter-site link configurations

A single control enclosure system with two inter-site links configuration is shown in Figure 6-38.

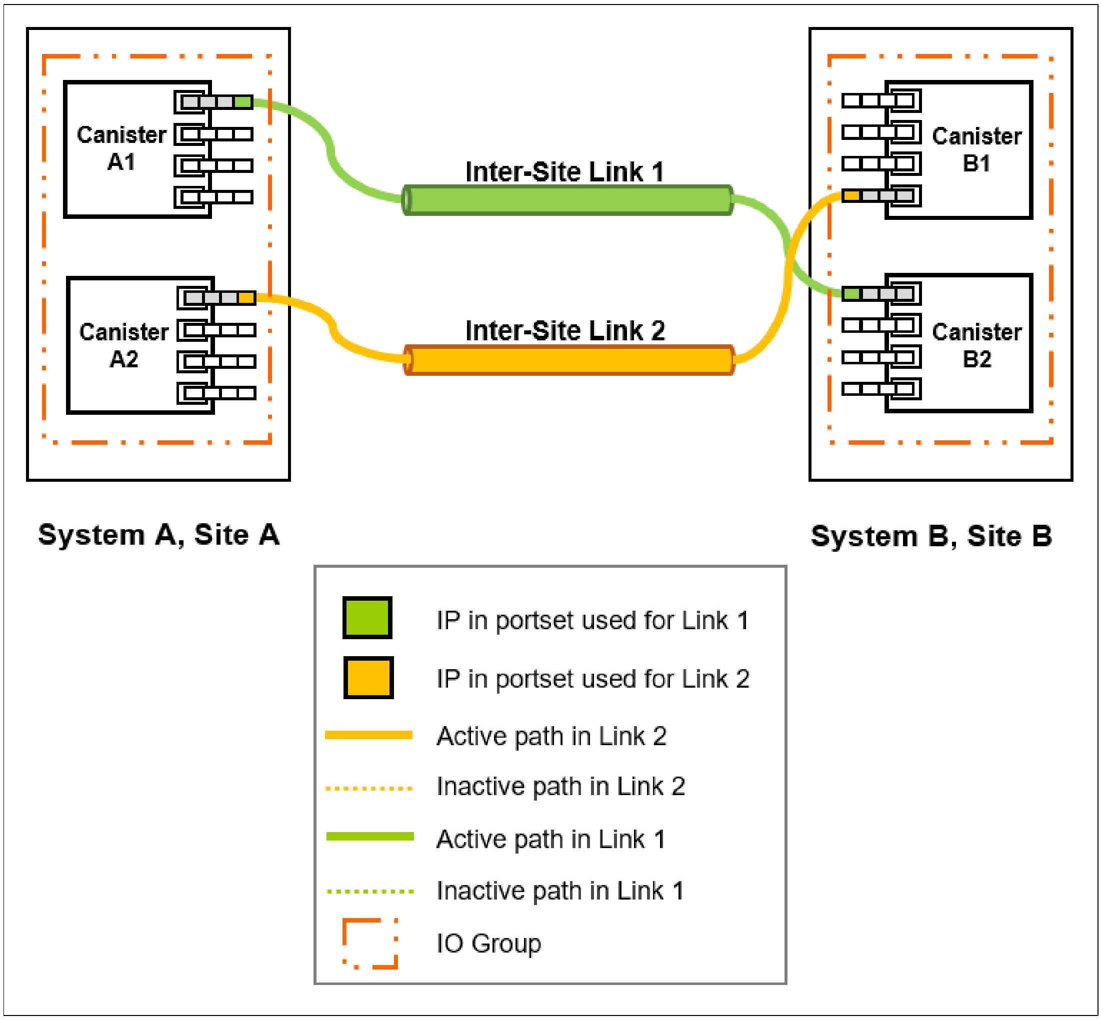


Figure 6-38 Dual links with two replication portset on each system configured

As shown in Figure 6-38, two replication portsets are configured on System A and System B because two inter-site links are available. In this configuration, the failover ports are not configured on partner canisters in the control enclosure. Rather, the ports are maintained in different portsets on both of the canisters. They can remain active and participate in IP partnership by using both of the links. Failover ports cannot be used with this configuration because only one active path per canister per partnership is allowed.

However, if either of the canisters in the control enclosure fail (that is, if Canister A1 on System A fails), the IP partnership continues from only the available IP that is configured in portset associated to link 2. Therefore, the effective bandwidth of the two links is reduced to 50% because only the bandwidth of a single link is available until the failure is resolved.

This configuration includes the following characteristics:

- Two inter-site links exist, and two replication portset are used.
- Each node has only one IP in each replication portset.

- Both IP in the two portsets participate simultaneously in IP partnerships. Therefore, both of the links are used.
- During canister failure or link failure, the IP partnership traffic continues from the other available link. Therefore, if two links of 10 Mbps each are available and you have 20 Mbps of effective link bandwidth, bandwidth is reduced to 10 Mbps only during a failure.
- After the canister failure or link failure is resolved and failback happens, the entire bandwidth of both of the links is available as before.

A four control enclosures clustered system in IP partnership with a two control enclosure clustered system over dual inter-site links is shown in Figure 6-39.

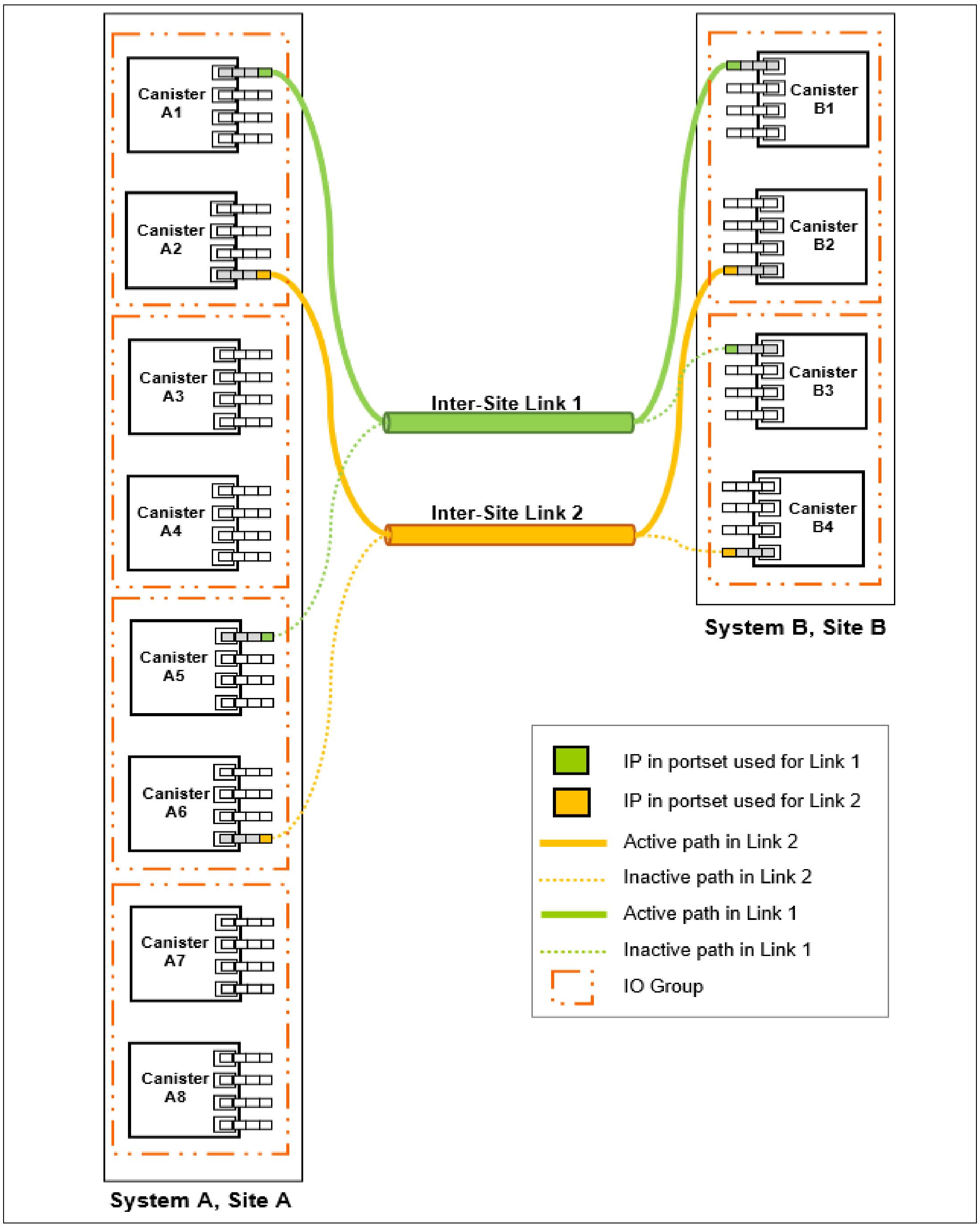


Figure 6-39 Clustered systems with dual inter-site links between the two systems

Figure 6-39 on page 341 shows a four control enclosure System A in Site A and a two control enclosure System B in Site B. Canisters from only two control enclosures are configured with replication portsets in System A.

In this configuration, two links and two control enclosures are configured with replication portsets. However, path selection logic is managed by an internal algorithm. Therefore, this configuration depends on the pathing algorithm to decide which of the canisters actively participate in IP partnership. Even if Canister A5 and Canister A6 have IPs configured within replication portsets properly, active IP partnership traffic on both of the links can be driven from Canister A1 and Canister A2 only.

If Canister A1 fails in System A, IP partnership traffic continues from Canister A2 (that is, link 2). The failover also causes IP partnership traffic to continue from Canister A5 on which a portset associated to link 1 is configured. The details of the specific IP port actively participating in IP partnership process is provided in the **`lspartnership`** output (reported as `link1_ip_id` and `link2_ip_id`).

This configuration includes the following characteristics:

- Two control enclosure have IPs configured in two replication portsets because two inter-site links for participating in IP partnership are used. However, only one IP per system in a particular portset remains active and participates in IP partnership.
- One IP per system from each replication portset participates in IP partnership simultaneously. Therefore, both of the links are used.
- If a canister or port on the canister that is actively participating in IP partnership fails, the Remote Copy data path is established from that port because another IP is available on an alternative canister in the system within the replication portset.
- The path selection algorithm starts discovery of available IPs in the affected portset in the alternative I/O groups and paths are reestablished. This process restores the total bandwidth across both links.

Multiple partnerships configurations

In this section some multiple partnerships configurations are described.

Figure 6-40 on page 343 shows a two control enclosures System A in Site A, a two control enclosures System B in Site B and a two control enclosures System C in Site C.

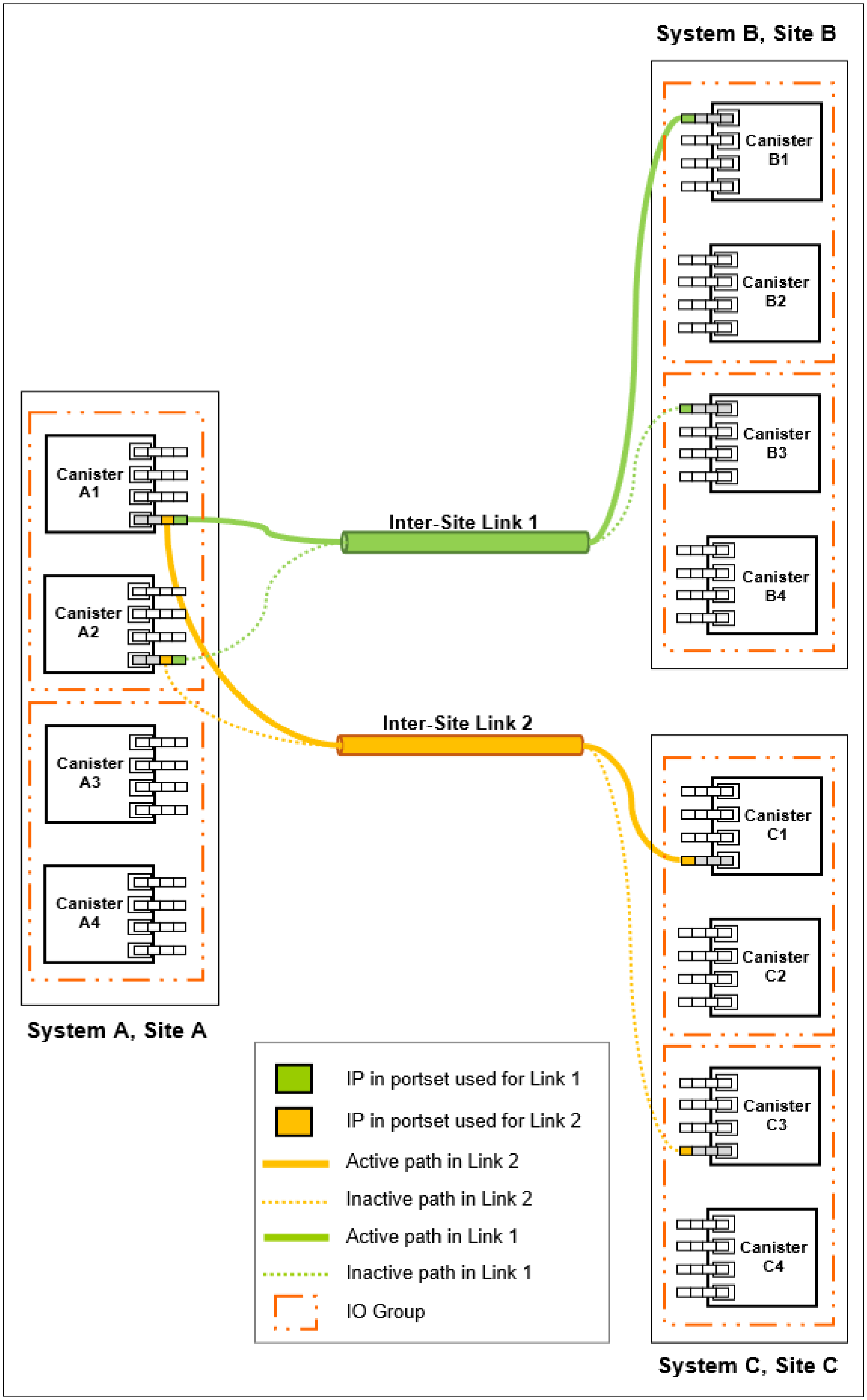


Figure 6-40 Multiple IP partnerships with two links and only one IO group

In this configuration, two links and only one control enclosure are configured with replication portsets in System A. Both replication portsets use the same Ethernet ports in canister A1 and A2. System B uses a replication portset associated to link 1, while System C uses a replication portset associated to link 2. System B and System C have configured portsets across both control enclosures.

However, path selection logic is managed by an internal algorithm. Therefore, this configuration depends on the pathing algorithm to decide which of the canisters actively participate in a IP partnerships. In this example, the active paths go from Canister A1 to Canister B1 and Canister A1 to Canister C1, respectively. In this configuration, multiple paths are allowed for a single canister because they are used for different IP partnerships.

If Canister A1 fails in System A, IP partnerships continues servicing replication I/O from Canister A2 because a failover port is configured on that node.

However, it might take some time for discovery and path configuration logic to reestablish paths post failover. This delay can cause partnerships to change to `Not_Present` for that time and this can lead to a replication stopping. The details of the specific IP port that is actively participating in IP partnership is provided in the **`lspartnership`** output (reported as `link1_ip_id` and `link2_ip_id`).

This configuration includes the following characteristics:

- One IP per system from each replication portset participates in IP partnership simultaneously. Therefore, both of the links are used.
- Replication portsets on System A for both links are defined in the same physical ports
- If a canister or port on the canister that is actively participating in IP partnership fails, the Remote Copy data path is established from that port because another IP is available on an alternative canister in the system within the replication portset.
- The path selection algorithm starts discovery of available IPs in the affected portset in the alternative control enclosures and paths are reestablished. This process restores the total bandwidth across both links.

Finally, an alternative partnership layout configuration is shown in Figure 6-41 on page 345.

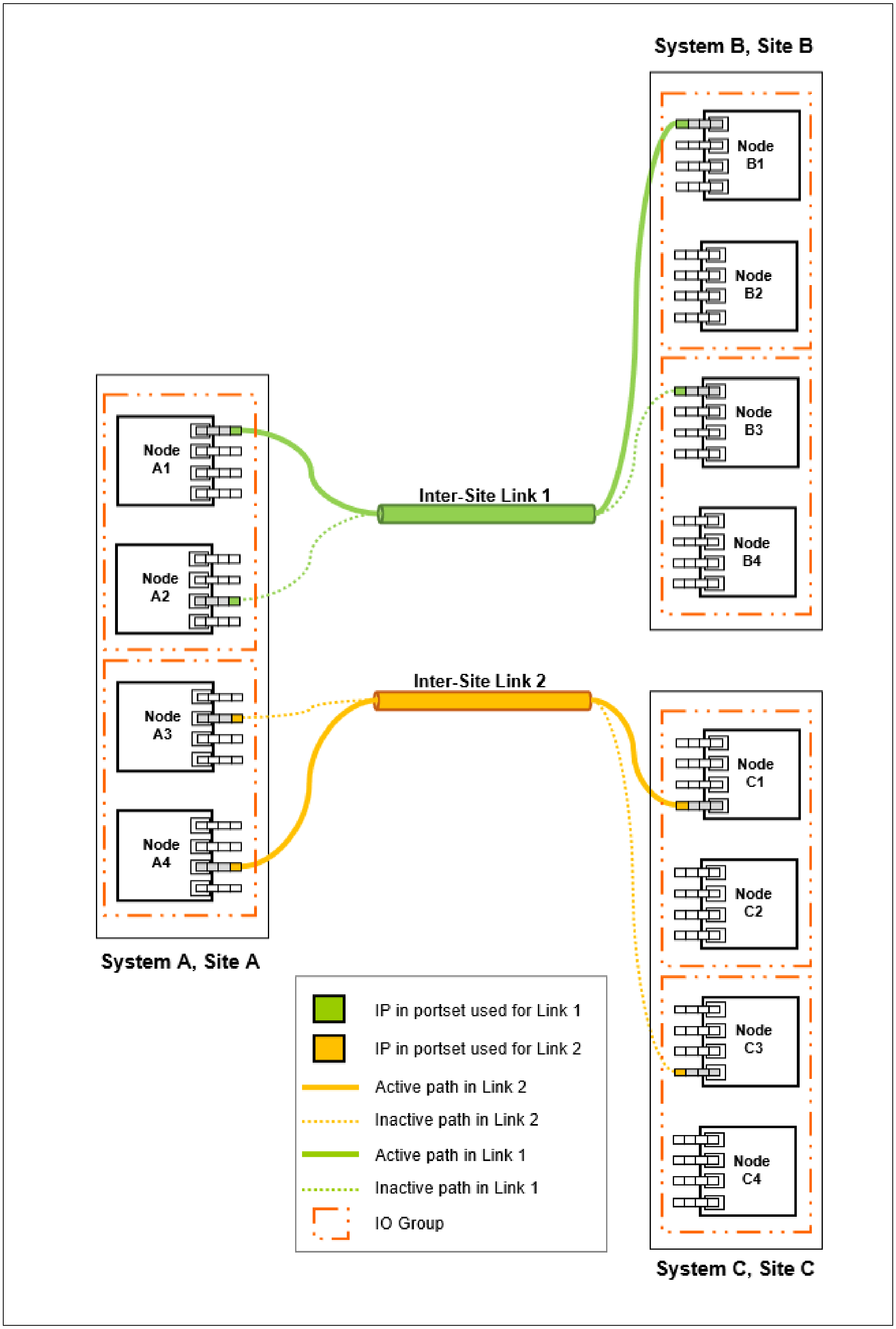


Figure 6-41 Multiple IP partnerships with two links

In this configuration, two links and two control enclosures are configured with replication portsets in System A. On System A control enclosure 0 (Canister A1 and Canister A2) uses IPs on the replication portset associated to link 1, while control enclosure 1 (Canister A3 and Canister A4) uses IPs on the replication portset associated to link 2. System B uses a replication portset associated to link 1, while System C uses a replication portset associated to link 2. System B and System C have configured portsets across both control enclosures.

However, path selection logic is managed by an internal algorithm. Therefore, this configuration depends on the pathing algorithm to decide which of the nodes actively participate in IP partnerships. In this example, the active paths go from Canister A1 to Canister B1 and Canister A4 to Canister C1 for System A to System B and System A to System C, respectively.

If Canister A1 fails in System A, IP partnership for System A to System B continues servicing replication I/O from Canister A2 because a failover port is configured on that node.

However, it might take some time for discovery and path configuration logic to reestablish paths post failover. This delay can cause partnerships to change to `Not_Present` for that time, which can lead to a replication stopping. The partnership for System A to System C remains unaffected. The details of the specific IP port that is actively participating in IP partnership is provided in the **`lspartnership`** output (reported as `link1_ip_id` and `link2_ip_id`).

This configuration includes the following characteristics:

- One IP per system from each replication portset participates in IP partnership simultaneously. Therefore, both of the links are used.
- Replication portsets on System A for the two links are defined in different physical ports
- If a canister or port on the canister that is actively participating in IP partnership fails, the Remote Copy data path is established from that port because another IP is available on an alternative canister in the system within the replication portset.
- The path selection algorithm starts discovery of available IPs in the affected portset in the alternative canisters and paths are reestablished. This process restores the total bandwidth across both links.
- If a canister or link failure occurs, only one partnership is affected.

> **Replication portsets:** As described in these sections, configuring two replication portsets provides more bandwidth and resilient configurations in case of a link failure. Two replication portsets also can be configured with a single physical link. This configuration make sense only if the total link bandwidth exceeds the aggregate bandwidth of two replication portsets together. The use of two portsets when the link bandwidth does not provide the aggregate throughput can lead to network resources contention and bad link performance.

6.4.7 Native IP replication performance consideration

A number of factors affect the performance of an IP partnership. Some of these factors are latency, link speed, number of intersite links, host I/O, MDisk latency, and hardware. Since the introduction, many improvements have been made to make the IP replication better performing and more reliable.

Nevertheless, in presence of poor quality networks that have significant packet loss and high latency, the actual usable bandwidth might decrease considerably.

Figure 6-42 shows the throughput trend for a 1 Gbps port in respect of the packet loss ratio and the latency.

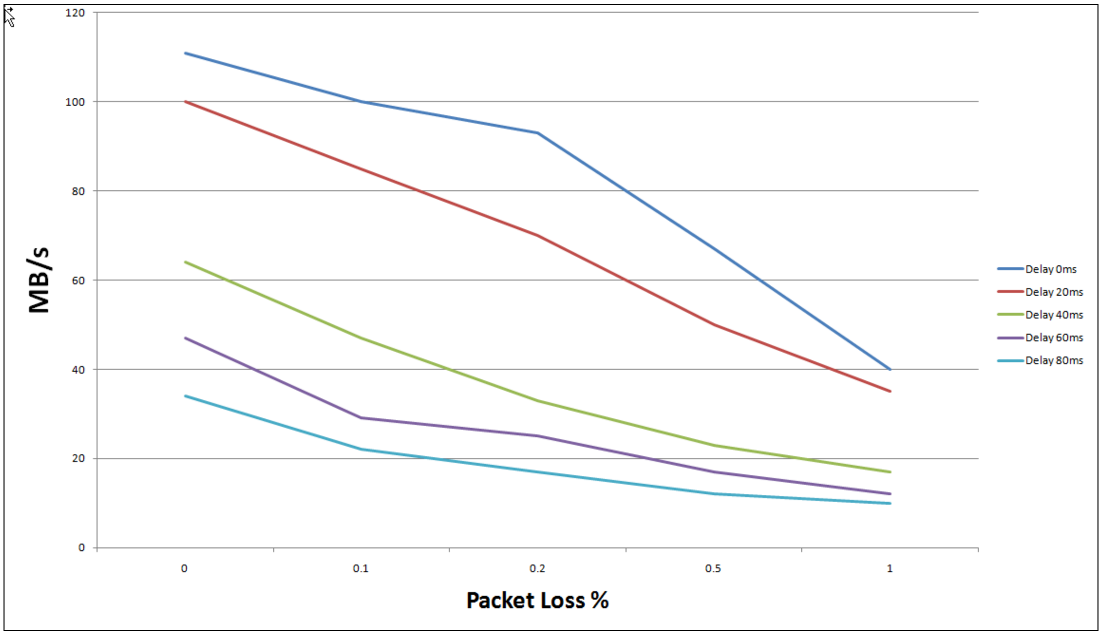


Figure 6-42 1 Gbps port throughput trend

The chart in Figure 6-42 shows how the combined effect of the packet loss and the latency can lead to a throughput reduction of more than 85%. For these reasons, the IP replication option should be considered only for the replication configuration that is not affected by poor quality and poor performing networks. Due to its characteristic of low-bandwidth requirement, the GMCV is the preferred solution with the IP replication.

To improve the performance when using compression and IP partnership in the same system, it is advised that you use a different port for iSCSI host I/O and IP partnership traffic. Also, use a different VLAN ID for iSCSI host I/O and IP partnership traffic.

6.5 Volume mirroring

By using volume mirroring, you can have two physical copies of a volume that provide a basic RAID-1 function. These copies can be in the same storage pool or in different storage pools, with different extent sizes of the storage pool. Typically. the two copies are allocated in different storage pools.

The first storage pool contains the original (primary volume copy). If one storage controller or storage pool fails, a volume copy is not affected if it has been placed on a different storage controller or in a different storage pool.

If a volume is created with two copies, both copies use the same virtualization policy. However, you can have two copies of a volume with different virtualization policies. In combination with *thin-provisioning*, each mirror of a volume can be thin-provisioned, compressed or fully allocated, and in striped, sequential, or image mode.

A mirrored (secondary) volume has all of the capabilities of the primary volume copy. It also has the same restrictions (for example, a mirrored volume is owned by an I/O Group, just as any other volume). This feature also provides a *point-in-time copy* function that is achieved by “splitting” a copy from the volume. However, the mirrored volume does not address other forms of mirroring based on Remote Copy (Global or Metro Mirror functions), which mirrors volumes across I/O Groups or clustered systems.

One copy is the primary copy, and the other copy is the secondary copy. Initially, the first volume copy is the primary copy. You can change the primary copy to the secondary copy if required.

Figure 6-43 shows an overview of volume mirroring.

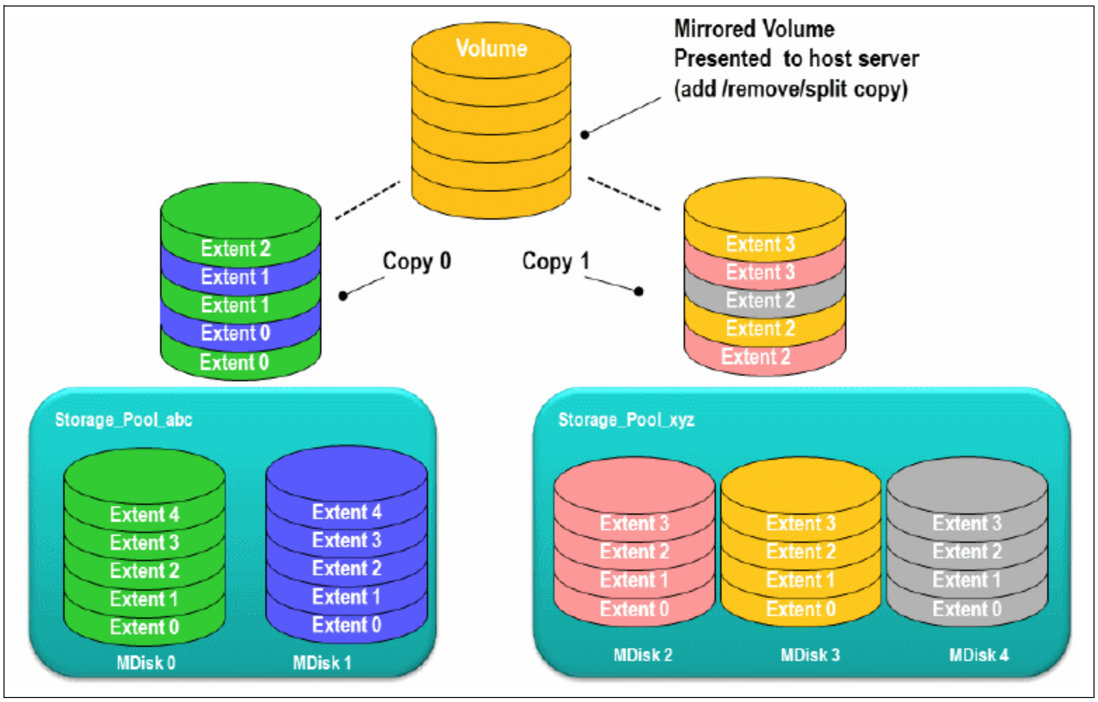


Figure 6-43 Volume mirroring overview

6.5.1 Read and write operations

Read and write operations behavior depends on the status of the copies and on other environment settings. During the initial synchronization or a resynchronization, only one of the copies is in synchronized status, and all the reads are directed to this copy. The write operations are directed to both copies.

When both copies are synchronized, the write operations are again directed to both copies. The read operations usually are directed to the primary copy, unless the system is configured in Enhanced Stretched Cluster topology, which applies to SAN Volume Controller system types only.

During back-end storage failure, note the following points:

- If one of the mirrored volume copies is temporarily unavailable, the volume remains accessible to servers.
- The system remembers which areas of the volume are written and resynchronizes these areas when both copies are available.

- The remaining copy can service read I/O when the failing one is offline, without user intervention.

6.5.2 Volume mirroring use cases

Volume mirroring offers the capability to provide extra copies of the data that can be used for High Availability solutions and data migration scenarios. You can convert a non-mirrored volume into a mirrored volume by adding a copy. When a copy is added using this method, the cluster system synchronizes the new copy so that it is the same as the existing volume. You can convert a mirrored volume into a non-mirrored volume by deleting one copy or by splitting one copy to create a new non-mirrored volume.

Access: Servers can access the volume during the synchronization processes described.

You can use mirrored volumes to provide extra protection for your environment or to perform a migration. This solution offers several options:

- Export to Image mode

  This option allows you to move storage from *managed mode* to *image mode*. This option is useful if you are using IBM FlashSystem as a migration device. For example, suppose vendor A's product cannot communicate with vendor B's product, but you need to migrate existing data from vendor A to vendor B.

  Using *Export to image mode* allows you to migrate data by using the Copy Services functions and then return control to the native array, while maintaining access to the hosts.

- Import to Image mode

  This option allows you to import an existing storage MDisk or logical unit number (LUN) with its existing data from an external storage system, without putting metadata on it. The existing data remains intact. After you import it, the volume mirroring function can be used to migrate the storage to the other locations, while the data remains accessible to your hosts.

- Volume cloning by using volume mirroring and then by using the Split into New Volume option

  This option allows any volume to be cloned without any interruption to the host access. You have to create two mirrored copies of data and then break the mirroring with the split option to make two independent copies of data. This option doesn't apply to already mirrored volumes.

- Volume pool migration using the volume mirroring option

  This option allows any volume to be moved between storage pools without any interruption to the host access. You might use this option to move volumes as an alternative to Migrate to Another Pool function.

  Compared to the Migrate to Another Pool function, volume mirroring provides more manageability because it can be suspended and resumed anytime and also it allows you to move volumes among pools with different extent sizes. This option doesn't apply to already mirrored volumes.

  Use case: Volume mirroring can be used to migrate volumes from and to DRPs which do not support extent based migrations. For more information, see 4.3.6, "Data migration with DRP" on page 145.

- Volume capacity saving change

  This option allows you to modify the capacity saving characteristics of any volume from standard to thin provisioned or compressed and vice versa, without any interruption to host access. This option works the same as the volume pool migration but specifying a different capacity saving for the newly created copy. This option doesn't apply to already mirrored volumes.

When you use volume mirroring, consider how quorum candidate disks are allocated. Volume mirroring maintains some state data on the quorum disks. If a quorum disk is not accessible and volume mirroring cannot update the state information, a mirrored volume might need to be taken offline to maintain data integrity. To ensure the high availability of the system, ensure that multiple quorum candidate disks, which are allocated on different storage systems, are configured.

Quorum disk consideration: Mirrored volumes can be taken offline if there is no quorum disk available. This behavior occurs because synchronization status for mirrored volumes is recorded on the quorum disk. To protect against mirrored volumes being taken offline, follow the guidelines for setting up quorum disks.

The following are other volume mirroring usage cases and characteristics:

- Creating a mirrored volume:
  - The maximum number of copies is two.
  - Both copies are created with the same virtualization policy, by default.

    To have a volume mirrored using different policies, you need to add a volume copy with a different policy to a volume that has only one copy.
  - Both copies can be located in different storage pools. The first storage pool that is specified contains the primary copy.
  - It is not possible to create a volume with two copies when specifying a set of MDisks.
- Add a volume copy to an existing volume:
  - The volume copy to be added can have a different space allocation policy.
  - Two existing volumes with one copy each cannot be merged into a single mirrored volume with two copies.
- Remove a volume copy from a mirrored volume:
  - The volume remains with only one copy.
  - It is not possible to remove the last copy from a volume.
- Split a volume copy from a mirrored volume and create a new volume with the split copy:
  - This function is only allowed when the volume copies are synchronized. Otherwise, use the `-force` command.
  - It is not possible to recombine the two volumes after they have been split.
  - Adding and splitting in one workflow enables migrations that are not currently allowed.
  - The split volume copy can be used as a means for creating a point-in-time copy (clone).

- Repair or validate volume copies, by comparing them and performing the following three functions:
  - Report the first difference found. It can iterate by starting at a specific LBA by using the **`-startlba`** parameter.
  - Create virtual medium errors where there are differences. This is useful in case of back-end data corruption.
  - Correct the differences that are found (reads from primary copy and writes to secondary copy).
- View to list volumes affected by a back-end disk subsystem being offline:
  - Assumes that a standard use is for mirror between disk subsystems.
  - Verifies that mirrored volumes remain accessible if a disk system is being shut down.
  - Reports an error in case a quorum disk is on the back-end disk subsystem.
- Expand or shrink a volume:
  - This function works on both of the volume copies at once.
  - All volume copies always have the same size.
  - All copies must be synchronized before expanding or shrinking them.

  DRP limitation: DRPs do not support thin/compressed volumes shrinking.

- Delete a volume. When a volume gets deleted, all copies get deleted.
- Migration commands apply to a specific volume copy.
- Out-of-sync bitmaps share the bitmap space with FlashCopy and Metro Mirror/Global Mirror. Creating, expanding, and changing I/O groups might fail if there is insufficient memory.
- GUI views contain volume copy identifiers.

6.5.3 Mirrored volume components

Note the following points regarding mirrored volume components:

- A mirrored volume is always composed of two copies (copy 0 and copy1).
- A volume that is not mirrored consists of a single copy (which for reference might be copy 0 or copy 1).

A mirrored volume looks the same to upper-layer clients as a non-mirrored volume. That is, upper layers within the cluster software, such as FlashCopy and Metro Mirror/Global Mirror, and storage clients, do not know whether a volume is mirrored. They all continue to handle the volume as they did before without being aware of whether the volume is mirrored.

6.5.4 Volume mirroring synchronization options

As soon as a volume is created with two copies, copies are in the *out-of-sync* state. The primary volume copy (located in the first specified storage pool) is defined as in sync and the secondary volume copy as out of sync. The secondary copy is synchronized through the synchronization process.

This process runs at the default synchronization rate of 50 (as shown in Table 6-10 on page 353), or at the defined rate while creating or modifying the volume. For more information on the effect of the copy rate setting, see 6.5.5, “Volume mirroring performance considerations” on page 352. When the synchronization process is completed, the volume mirroring copies are *in-sync* state.

By default, when a mirrored volume is created a format process is also initiated. This process guarantees that the volume data is zeroed, avoiding access to data that is still present on reused extents.

This format process runs in background at the defined synchronization rate, as shown in Table 6-10 on page 353. Before Spectrum Virtualize version 8.4, the format processing overwrite with zeros only the Copy 0 and then synchronize the Copy 1. With version 8.4 or later, the format process is initiated concurrently to both volume mirroring copies and thus avoiding the second synchronization step.

You can specify that a volume is synchronized (**`-createsync`** parameter), even if it is not. Using this parameter can cause data corruption if the primary copy fails and leaves an unsynchronized secondary copy to provide data. Using this parameter can cause loss of read stability in unwritten areas if the primary copy fails, data is read from the primary copy, and then different data is read from the secondary copy. To avoid data loss or read stability loss, use this parameter only for a primary copy that has been formatted and not written to. When using the **`-createsync`** setting, the initial formatting is skipped.

Another example use case for **`-createsync`** is for a newly created mirrored volume where both copies are thin provisioned or compressed because no data has been written to disk and unwritten areas return zeros (0). If the synchronization between the volume copies has been lost, the resynchronization process is incremental. This term means that only grains that have been written to need to be copied, and then get synchronized volume copies again.

The progress of the volume mirror synchronization can be obtained from the GUI or by using the **`lsvdisksyncprogress`** command.

6.5.5 Volume mirroring performance considerations

Because the writes of mirrored volumes always occur to both copies, mirrored volumes put more workload on the cluster, the back-end disk subsystems, and the connectivity infrastructure. The mirroring is symmetrical, and writes are only acknowledged when the write to the last copy completes. The result is that if the volumes copies are on storage pools with different performance characteristics, the slowest storage pool determines the performance of writes to the volume. This performance applies when writes must be destaged to disk.

> **Tip:** Locate volume copies of one volume on storage pools of the same or similar characteristics. Usually, if only good read performance is required, you can place the primary copy of a volume in a storage pool with better performance. Because the data is always only read from one volume copy, reads are not faster than without volume mirroring.
>
> However, be aware that this is only true when both copies are synchronized. If the primary is out of sync, then reads are submitted to the other copy.

Synchronization between volume copies has a similar impact on the cluster and the back-end disk subsystems as FlashCopy or data migration. The synchronization rate is a property of a volume that is expressed as a value of 0 - 150. A value of 0 disables synchronization.

Table 6-10 shows the relationship between the rate value and the data copied per second.

Table 6-10 Relationship between the rate value and the data copied per second

User-specified rate attribute value per volume	Data copied per second
0	Synchronization is disabled
1 - 10	128 KB
11 - 20	256 KB
21 - 30	512 KB
31 - 40	1 MB
41 - 50	2 MB (50% is the default value)
51 - 60	4 MB
61 - 70	8 MB
71 - 80	6 MB
81 - 90	32 MB
91 - 100	64 MB
101 - 110	128 MB
111 - 120	256 MB
121 - 130	512 MB
131 - 140	1024 MB
141 - 150	2048 MB

Rate attribute value: The rate attribute is configured on each volume that you want to mirror. The default value of a new volume mirror is 50%.

In large, IBM FlashSystem configurations, the settings of the copy rate can considerably affect the performance in scenarios where a back-end storage failure occurs. For instance, consider a scenario where a failure of a back-end storage controller is affecting one copy of 300 mirrored volumes. The host continues the operations by using the remaining copy.

When the failed controller comes back online, the resynchronization process for all the 300 mirrored volumes starts at the same time. With a copy rate of 100 for each volume, this process would add a theoretical workload of 18.75 GBps, which will considerably overload the system.

The general suggestion for the copy rate settings is then to evaluate the impact of massive resynchronization and set the parameter accordingly. Consider setting the copy rate to high values for initial synchronization only, and with a limited number of volumes at a time. Alternatively, consider defining a volume provisioning process that allows the safe creation of already synchronized mirrored volumes, as described in 6.5.4, “Volume mirroring synchronization options” on page 351.

Volume mirroring I/O time-out configuration

A mirrored volume has pointers to the two copies of data, usually in different storage pools, and each write completes on both copies before the host receives I/O completion status. For a synchronized mirrored volume, if a write I/O to a copy has failed or a long timeout has expired, then system has completed all available controller level Error Recovery Procedures (ERPs). In this case, that copy is taken offline and goes out of sync. The volume remains online and continues to service I/O requests from the remaining copy.

The *Fast Failover* feature isolates hosts from temporarily poorly-performing back-end storage of one Copy at the expense of a short interruption to redundancy. The fast failover feature behavior is that during normal processing of host write I/O, the system submits writes to both copies with a timeout of 10 seconds (20 seconds for stretched volumes). If one write succeeds and the other write takes longer than 5 seconds, then the slow write is stopped. The FC `abort` sequence can take around 25 seconds.

When the stop is completed, one copy is marked as out of sync and the host write I/O completed. The overall fast failover ERP aims to complete the host I/O in approximately 30 seconds (or 40 seconds for stretched volumes).

The fast failover can be set for *each* mirrored volume by using the **`chvdisk`** command and the **`mirror_write_priority`** attribute settings:

- *Latency* (default value): A short timeout prioritizing low host latency. This option enables the fast failover feature.
- *Redundancy*: A long timeout prioritizing redundancy. This option indicates a copy that is slow to respond to a write I/O can use the full ERP time. The response to the I/O is delayed until it completes to keep the copy in sync if possible. This option disables the fast failover feature.

Volume mirroring ceases to use the slow copy for 4 - 6 minutes, and subsequent I/O data is not affected by a slow copy. Synchronization is suspended during this period. After the copy suspension completes, volume mirroring resumes, which allows I/O data and synchronization operations to the slow copy that often quickly completes the synchronization.

If another I/O times out during the synchronization, then the system stops using that copy again for 4 - 6 minutes. If one copy is always slow, then the system tries it every 4 - 6 minutes and the copy gets progressively more out of sync as more grains are written. If fast failovers are occurring regularly, there is probably an underlying performance problem with the copy's back-end storage.

The preferred **`mirror_write_priority`** setting for the Enhanced Stretched Cluster configurations is `latency`.

6.5.6 Bitmap space for out-of-sync volume copies

The grain size for the synchronization of volume copies is 256 KB. One grain takes up one bit of bitmap space. 20 MB of bitmap space supports 40 TB of mirrored volumes. This relationship is the same as the relationship for copy services (Global and Metro Mirror) and standard FlashCopy with a grain size of 256 KB (see Table 6-11).

Table 6-11 Relationship of bitmap space to volume mirroring address space

Function	Grain size in KB	1 byte of bitmap space gives a total of	4 KB of bitmap space gives a total of	1 MB of bitmap space gives a total of	20 MB of bitmap space gives a total of	512 MB of bitmap space gives a total of
Volume mirroring	256	2 MB of volume capacity	8 GB of volume capacity	2 TB of volume capacity	40 TB of volume capacity	1024 TB of volume capacity

Shared bitmap space: This bitmap space on one I/O group is shared between Metro Mirror, Global Mirror, FlashCopy, and volume mirroring.

The command to create Mirrored Volumes can fail if there is not enough space to allocate bitmaps in the target I/O Group. To verify and change the space allocated and available on each I/O Group with the CLI, see the Example 6-4.

Example 6-4 A lsiogrp and chiogrp command example

```
IBM_FlashSystem:ITSO:superuser>lsiogrp
id name            node_count vdisk_count host_count site_id site_name
0  io_grp0         2          9           0
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_FlashSystem:ITSO:superuser>lsiogrp io_grp0|grep _memory
flash_copy_total_memory 20.0MB
flash_copy_free_memory 20.0MB
remote_copy_total_memory 20.0MB
remote_copy_free_memory 20.0MB
mirroring_total_memory 20.0MB
mirroring_free_memory 20.0MB
raid_total_memory 40.0MB
raid_free_memory 40.0MB
flash_copy_maximum_memory 2048.0MB
compression_total_memory 0.0MB
.
IBM_FlashSystem:ITSO:superuser>chiogrp -feature mirror -size 64 io_grp0
IBM_FlashSystem:ITSO:superuser>lsiogrp io_grp0|grep _memory
flash_copy_total_memory 20.0MB
flash_copy_free_memory 20.0MB
remote_copy_total_memory 20.0MB
remote_copy_free_memory 20.0MB
mirroring_total_memory 64.0MB
mirroring_free_memory 64.0MB
raid_total_memory 40.0MB
raid_free_memory 40.0MB
flash_copy_maximum_memory 2048.0MB
compression_total_memory 0.0MB
```

7

Ensuring business continuity

Business continuity and continuous application availability are among the most important requirements for many organizations. Advances in virtualization, storage, and networking made enhanced business continuity possible.

Information technology solutions can now manage planned and unplanned outages, and provide the flexibility and cost efficiencies that are available from cloud-computing models.

This chapter briefly describes the HyperSwap solutions for IBM Spectrum Virtualize systems. Technical details or implementation guidelines are not presented in this chapter because they are described in separate publications.

> **Important:** This book was written specifically for IBM FlashSystems products. Therefore, it does not cover Stretched Cluster and Enhanced Stretched Cluster topologies. For more information about IBM SAN Volume Controller, see *IBM SAN Volume Controller Best Practices and Performance Guidelines*, SG24-8502.
>
> This book does not cover the 3-site replication solutions, which are available with the IBM Spectrum Virtualize code version 8.3.1 or later. For more information, see *IBM Spectrum Virtualize 3-Site Replication*, SG24-8504.

This chapter includes the following topics:

7.1 Business continuity with HyperSwap

The *HyperSwap* high-availability feature in the IBM Spectrum Virtualize and FlashSystems products enables business continuity during a hardware failure, power outage, connectivity problem, or other disasters, such as fire or flooding.

It provides highly available volumes accessible through two sites located at up to 300 kilometers (km) apart. A fully independent copy of the data is maintained at each site. When data is written by hosts at either site, both copies are synchronously updated before the write operation is completed. HyperSwap automatically optimizes itself to minimize data that is transmitted between sites, and to minimize host read and write latency.

For more information about the optimization algorithm, see 7.3, "HyperSwap Volumes" on page 364.

HyperSwap includes the following key features:

- Works with all IBM Spectrum Virtualize products except for IBM FlashSystem 5010.
- Uses intra-cluster synchronous Remote Copy (Active-Active Metro Mirror) capability, with change volumes and access I/O group technologies.
- Makes a host's volumes accessible across two IBM Spectrum Virtualize I/O groups in a clustered system by using the Active-Active Metro Mirror relationship. The volumes are presented as a single volume to the host.
- Works with the standard multipathing drivers that are available on various host types. Additional host support is not required to access the highly available volumes.

The IBM Spectrum Virtualize HyperSwap configuration requires that at least one control enclosure is implemented in each location. Therefore, a minimum of two control enclosures for each cluster is needed to implement HyperSwap. Configurations with three or four control enclosures also are supported for the HyperSwap.

The typical IBM FlashSystems HyperSwap implementation is shown in Figure 7-1.

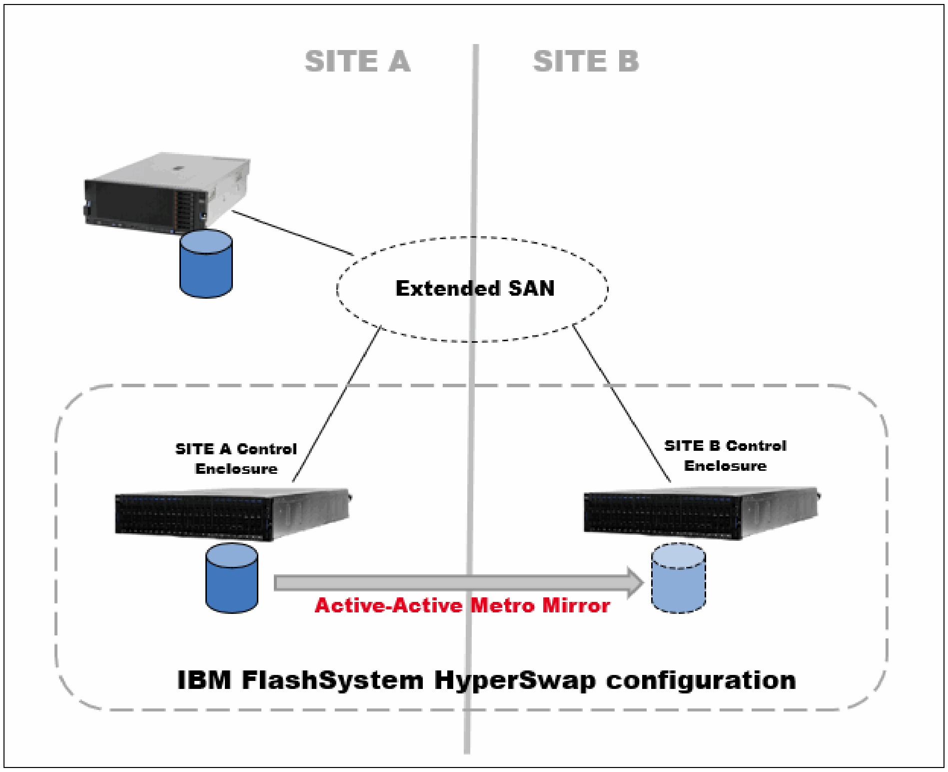


Figure 7-1 Typical HyperSwap configuration with IBM FlashSystem

With a copy of the data that is stored at each location, HyperSwap configurations can handle different failure scenarios.

Figure 7-2 shows how HyperSwap operates in a storage failure in one location. In this case, after the storage failure was detected in Site A, the HyperSwap function provides access to the data through the copy in the surviving site.

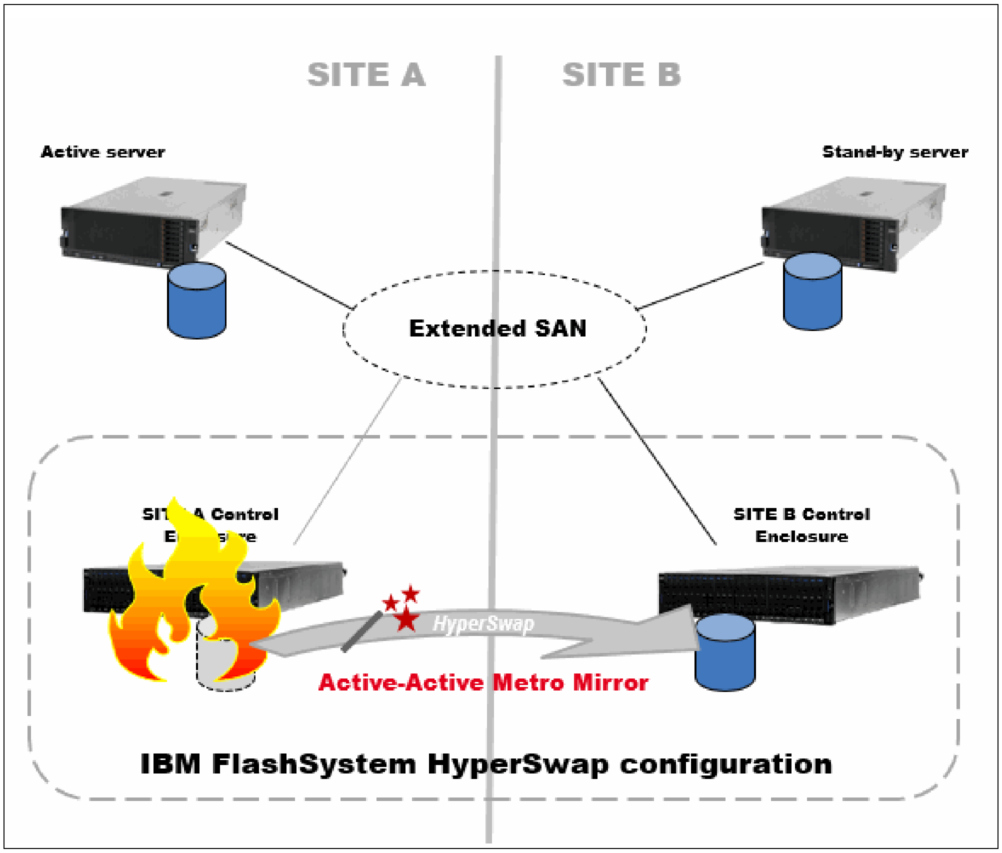


Figure 7-2 IBM FlashSystem HyperSwap in a storage failure scenario

You can lose an entire location, and access to the disks remains available at the alternate location. The use of this behavior requires clustering software at the application and server layer to fail over to a server at the alternate location and resume access to the disks.

The active-active synchronous mirroring feature, as shown in Figure 7-3, provides the capability to keep both copies of the storage in synchronization. Therefore, the loss of one location causes no disruption to the alternate location.

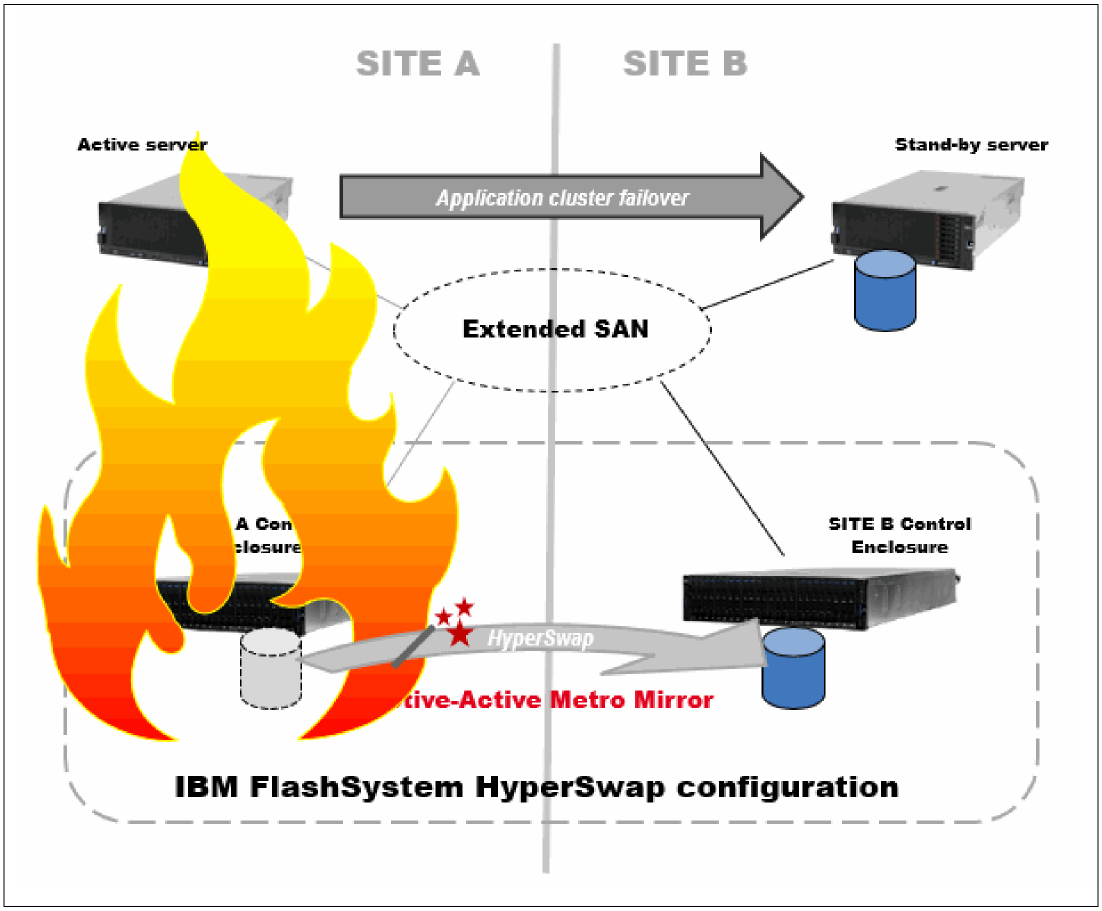


Figure 7-3 IBM FlashSystem HyperSwap in a site failure scenario

In addition to the active-active Metro Mirror feature, the HyperSwap feature also introduced the *site awareness* concept for node canisters, internal and external storage, and hosts. Finally, with the HyperSwap *DR feature*, you can manage rolling-disaster scenarios effectively.

7.2 Third site and IP quorum

In HyperSwap configurations, you can use a third, independent site to house a quorum device to act as the tie-breaker in case of split-brain scenarios. The quorum device can also hold a backup copy of the cluster metadata to be used in certain situations that might require a full cluster recovery.

To use a quorum disk as the quorum device, this third site must have Fibre Channel or iSCSI connectivity between an external storage system and the IBM Spectrum Virtualize cluster. Sometimes, this third site quorum disk requirement turns out to be expensive in terms of infrastructure and network costs. For this reason, a less demanding solution based on a Java application, known as the IP quorum application, is introduced with the release V7.6.

Initially, IP quorum was used only as a tie-breaker solution. However, with the release V8.2.1, it was expanded to be able to store cluster configuration metadata, fully serving as an alternative for quorum disk devices. To use an IP quorum application as the quorum device for the third site, Fibre Channel connectivity is not used. An IP quorum application can be run on any host at the third site, as shown in Figure 7-4.

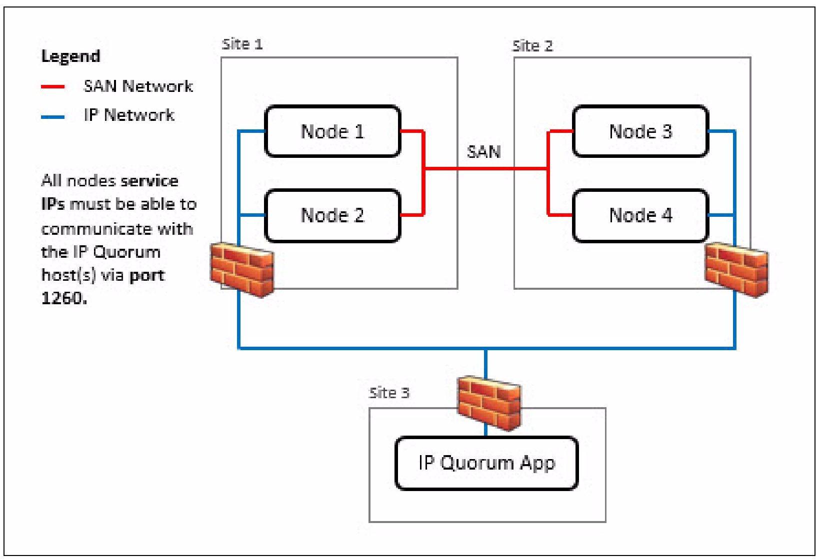


Figure 7-4 IP Quorum network layout

However, the following strict requirements must be met on the IP network when an IP quorum application is used:

- Connectivity from the servers that are running an IP quorum application to the service IP addresses of all nodes or node canisters. The network must also handle the possible security implications of exposing the service IP addresses, as this connectivity can also be used to access the service assistant interface if the IP network security is configured incorrectly.
- On each server that runs an IP quorum application, ensure that only authorized users can access the directory that contains the IP quorum application. Metadata is stored in the directory in a readable format, so ensure access to the IP quorum application and the metadata is restricted to only authorized users.
- Port 1260 is used by the IP quorum application to communicate from the hosts to all nodes or enclosures.

- The maximum round-trip delay must not exceed 80 milliseconds (ms), which means 40 ms each direction.
- If you are configuring the IP quorum application without a quorum disk for metadata, a minimum bandwidth of two megabytes per second is guaranteed for traffic between the system and the quorum application. If your system is using an IP quorum application with quorum disk for metadata, a minimum bandwidth of 64 megabytes per second is guaranteed for traffic between the system and the quorum application.
- Ensure that the directory that stores an IP quorum application with metadata contains at least 250 megabytes of available capacity.

Quorum devices are also required at Site 1 and Site 2, and can be either disk-based quorum devices or IP quorum applications. A maximum number of five IP quorum applications can be deployed.

> **Important:** *Do not* host the quorum disk devices or IP quorum applications on storage provided by the system it is protecting, as during a tie-break situation this storage is paused for I/O.

For more information about IP Quorum requirements and installation, including supported Operating Systems and Java runtime environments (JREs), see this IBM Documentation web page.

For more information about quorum disk devices, see 3.5, “Quorum disks” on page 107.

> **Note:** The IP Quorum configuration process has been integrated into the IBM Spectrum Virtualize GUI and can be found at **Settings** → **Systems** → **IP Quorum**.

7.2.1 Quorum modes

Quorum mode is a new configuration option that was added to the IP Quorum functionality with the release of IBM Spectrum Virtualize V8.3. By default, the IP quorum mode is set to **Standard**. In HyperSwap clusters, this mode can be changed to **Preferred** or **Winner**.

This configuration allows you to specify which site will resume I/O after a disruption, based on the applications that run on each site or other factors. For example, you can specify whether a selected site is the preferred for resuming I/O, or if the site automatically “wins” in tie-break scenarios.

Preferred Mode

If only one site runs critical applications, you can configure this site as *preferred*. During a split-brain situation, the system delays processing tie-break operations on other sites that are not specified as “preferred”. That is, the designated preferred site has a timed advantage when a split-brain situation is detected, and starts racing for the quorum device a few seconds before the non-preferred sites.

Therefore, the likelihood of reaching the quorum device first is higher. If the preferred site is damaged or cannot reach the quorum device, the other sites have the chance to win the tie breaker and continue I/O.

Winner Mode

This configuration is recommended for use when a third site is not available for a quorum device to be installed. In this case, when a split-brain situation is detected, the site configured as the winner will always be the one to continue processing I/O, regardless of the failure and its condition. The nodes at the non-winner site always loses the tie-break and stops processing I/O requests until the fault is fixed.

7.3 HyperSwap Volumes

HyperSwap Volumes is one type of volume. It consists of a Master Volume and a Master Change Volume (CV) in one system site, and an Auxiliary Volume and Auxiliary Change Volume (CV) in the other system site. An active-active synchronous mirroring relationship exists between the two sites. As with a regular Metro Mirror relationship, the active-active relationship keeps the Master Volume and Auxiliary Volume synchronized.

The relationship uses the CVs as journaling volumes during any resynchronization process. The Master CV must be in the same I/O Group as the Master Volume, and it is recommended that it is in the same pool as the Master Volume. A similar practice applies to the Auxiliary CV and the Auxiliary Volume.

For more information about the Change Volume, see “Global Mirror Change Volumes functional overview” on page 285.

The HyperSwap Volume always uses the unique identifier (UID) of the Master Volume. The HyperSwap Volume is assigned to the host by mapping only the Master Volume even though access to the Auxiliary Volume is ensured by the HyperSwap function.

Figure 7-5 shows how the HyperSwap Volume is implemented.

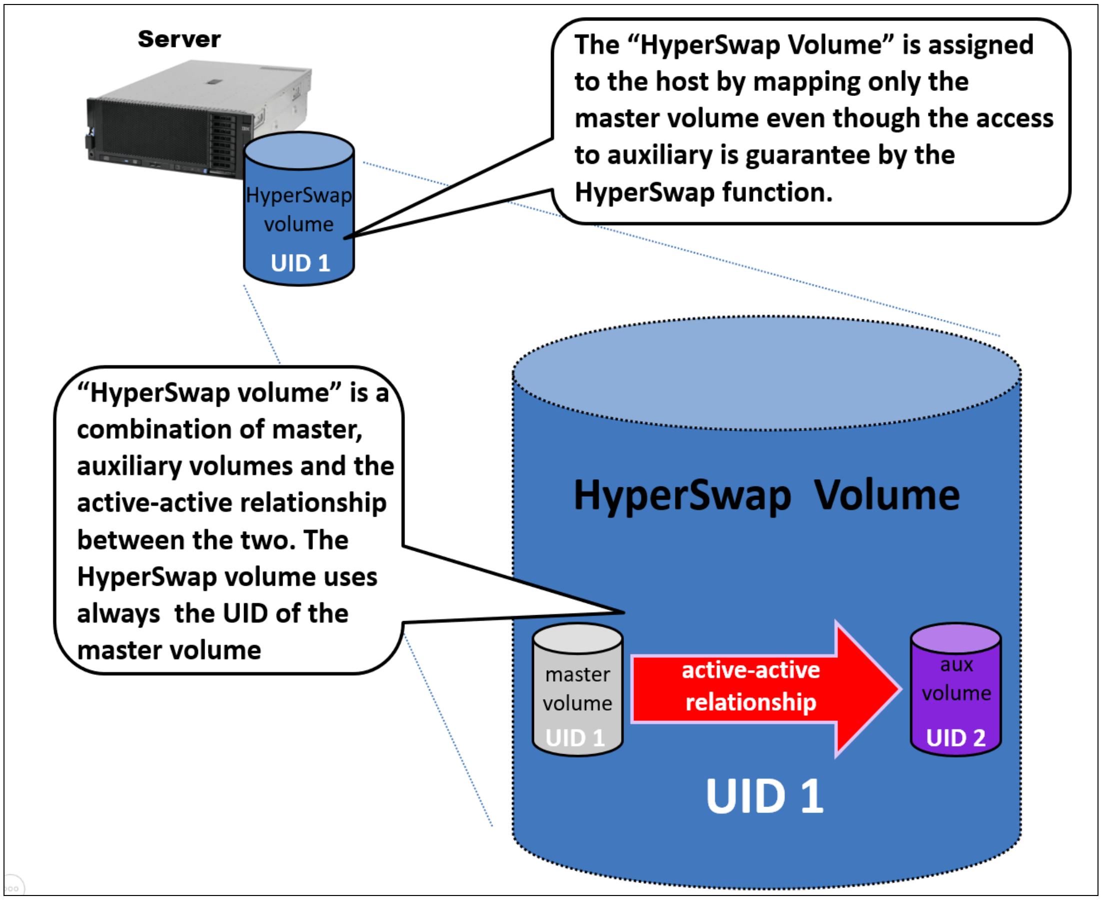


Figure 7-5 HyperSwap Volume

The active-active synchronous replication workload traverses the SAN by using the node-to-node communication. Master and Auxiliary Volumes also have a specific role of Primary or Secondary. Master or Auxiliary Volumes are Primary or Secondary based on the Metro Mirror active-active relationship direction.

Starting with the IBM Spectrum Virtualize 8.3.1 code level, reads are always done in the local copy of the volume. Write operations are always routed to the Primary copy. Therefore, hosts that access the Secondary copy for writes might experience an increased latency in the I/O operations. As a mitigation of this behavior, if sustained workload (that is, more than 75% of I/O operations for at least 20 minutes) is running over Secondary volumes, the HyperSwap function switches the direction of the active-active relationships, swapping the Secondary volume to Primary and vice versa.

Note: Frequent or continuous primary to secondary volume swap can lead to performance degradation. Avoid constantly switching the workload between sites at the host level.

7.4 Other considerations and general recommendations

Business continuity solutions implementation requires special considerations in the infrastructure and network setup. In HyperSwap topologies, the communication between the IBM Spectrum Virtualize controllers must be optimal and free of errors for best performance, as the internode messaging and cache mirroring is done across the sites. Have a dedicated private SAN for internode communication so that it is not impacted by regular SAN activities.

One other important recommendation is to review the site attribute of all the components, to make sure they are accurate. With the site awareness algorithm present in the IBM Spectrum Virtualize code, optimizations are done to reduce the cross-site workload. If this attribute is missing or not accurate, there might be an increased unnecessary cross-site traffic, which might lead to higher response time to the applications.

The HyperSwap feature requires implementing the storage network to ensure that the inter-node communication on the Fibre Channel ports on the control enclosures between the sites is on dedicated fabrics. No other traffic (hosts, back-end controllers, if any) or traffic that is unrelated to the HyperSwapped cluster can be allowed on this fabric. Two fabrics are used: one private for the inter-node communication, and one public for all other data.

A few SAN designs are available that can achieve this separation and some incorrect SAN designs can result in some potential problems that can occur with incorrect SAN design and implementation.

Find more information about design options and some common problems, see the following resources:

- *SAN and Fabric Resiliency Best Practices for IBM b-type Products*, REDP-4722
- This IBM Support web page.

For more information about a step-by-step configuration, see this IBM Documentation web page.

8

Configuring hosts

This chapter provides general guidelines and best practices for configuring host systems. The primary reference for host configuration is available at this IBM Documentation web pae.

For more information about host attachment, see this IBM Documentation web page.

For more information about hosts that are connected by using Fibre Channel, see Chapter 2, SAN Design. Host connectivity is a key consideration in overall SAN design.

Before attaching a new host, confirm that the host is supported by the IBM Spectrum Virtualize storage. For more information, see IBM System Storage Interoperation Center (SSIC).

The host configuration guidelines apply equally to all IBM Spectrum Virtualize systems. As such, the product name often is referred to as an *IBM Spectrum Virtualize* system.

This chapter includes the following topics:

8.1 General configuration guidelines

In this section, we discuss some general configuration guidelines. The information that is presented here complements the content in Chapter 2, “Connecting IBM Spectrum Virtualize and IBM Storwize in storage area networks” on page 37.

8.1.1 Number of paths

It is generally recommended that the total number of Fibre Channel paths per volume be limited to four paths. For HyperSwap and Stretch Cluster configurations, eight paths per volume is recommended. Adding paths does not significantly increase redundancy and it tends to bog down the host with path management. Also, too many paths might increase failover time.

8.1.2 Host ports

Each host uses two ports from two different host bus adapters (HBAs). These ports should go to separate SAN fabrics and be zoned to one target port of each node or node canister. When the volumes are created, they are assigned to an I/O group and the resulting path count between the volume and the host should be four.

Preferred practice: Keep Fibre Channel tape (including Virtual Tape Libraries) and Fibre Channel disks on separate HBAs. These devices have two different data patterns when operating in their optimum mode. Switching between them can cause unwanted processor usage and performance slowdown for the applications.

8.1.3 Port masking

In general, Fibre Channel ports should be dedicated to specific functions. Hosts must be zoned to only ports that are designated for host I/O.

For more information about port masking, see Chapter 2, “Connecting IBM Spectrum Virtualize and IBM Storwize in storage area networks” on page 37.

8.1.4 N-port ID virtualization

IBM Spectrum Virtualize now uses N-port ID virtualization (NPIV) by default. This use reduces failover time and allows for features, such as hot spare nodes.

For more information about configuring NPIV, see Chapter 2, “Connecting IBM Spectrum Virtualize and IBM Storwize in storage area networks” on page 37.

8.1.5 Host to I/O group mapping

An *I/O group* consists of two nodes or node canisters that share the management of volumes within the cluster. Use a single I/O group (`iogrp`) for all volumes that are allocated to a specific host. This guideline results in the following benefits:

- Minimizes port fan-outs within the SAN fabric.
- Maximizes the potential host attachments to IBM Spectrum Virtualize because maximums are based on I/O groups.
- Reduces the number of target ports that must be managed within the host.

8.1.6 Volume size versus 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

Host mapping is the process of controlling which hosts can access specific volumes within the system. IBM Spectrum Virtualize always presents a specific volume with the same SCSI ID on all host ports. When a volume is mapped, IBM Spectrum Virtualize software automatically assigns the next available SCSI ID if none is specified. In addition, a unique identifier, called the *UID*, is on each volume.

You can allocate the operating system volume of the SAN boot as the lowest SCSI ID (zero for most hosts), and then allocate the various data disks. If you share a volume among multiple hosts, consider controlling the SCSI ID so that the IDs are identical across the hosts. This consistency ensures ease of management at the host level and prevents potential issues during IBM Spectrum Virtualize updates and even node reboots, mostly for ESX operating systems.

If you are using image mode to migrate a host to IBM Spectrum Virtualize, allocate the volumes in the same order that they were originally assigned on the host from the back-end storage.

The **`lshostvdiskmap`** command displays a list of VDisk (volumes) that are mapped to a host. These volumes are recognized by the specified host. Example 8-1 shows the syntax of the **`lshostvdiskmap`** command that is used to determine the SCSI ID and the UID of volumes.

Example 8-1 The lshostvdiskmap command

```
svcinfo lshostvdiskmap -delim
```

Example 8-2 shows the results of using the **lshostvdiskmap** command.

Example 8-2 Output of using the lshostvdiskmap command

```
svcinfo lshostvdiskmap -delim : HG-ESX6
id:name:SCSI_id:vdisk_id:vdisk_name:vdisk_UID:IO_group_id:IO_group_name:mapping_type:host_cluste
r_id:host_cluster_name:protocol
3:HG-ESX6:0:5:DB_Volume:60050768108104A2F000000000000037:0:io_grp0:private:::scsi
3:HG-ESX6:1:15:Infra_Volume:60050768108104A2F000000000000041:0:io_grp0:private:::scsi
3:HG-ESX6:2:43:onprem_volume_Ansible:60050768108104A2F000000000000081:0:io_grp0:private:::scsi
3:HG-ESX6:3:14:Volume IP Replication:60050768108104A2F000000000000040:0:io_grp0:private:::scsi
3:HG-ESX6:4:48:ansible:60050768108104A2F000000000000086:0:io_grp0:private:::scsi
3:HG-ESX6:5:49:ansible2:60050768108104A2F000000000000087:0:io_grp0:private:::scsi
3:HG-ESX6:6:34:Onprem_Demo_Ansible_Vol:60050768108104A2F00000000000009F:0:io_grp0:private:::scsi
3:HG-ESX6:7:50:vol_HG-ESX6_1:60050768108104A2F0000000000000A5:0:io_grp0:private:::scsi
3:HG-ESX6:8:51:vol_HG-ESX6_10:60050768108104A2F0000000000000A8:0:io_grp0:private:::scsi
```

Example 8-3 shows the **lsvdiskhostmap** command.

Example 8-3 The lsvdiskhostmap command

```
svcinfo lsvdiskhostmap -delim
```

Example 8-4 shows the results of using the **lsvdiskhostmap** command.

Example 8-4 Output of using the lsvdiskhostmap command

```
svcinfo lsvdiskhostmap -delim : EEXCLS_HBin01
id:name:SCSI_id:host_id:host_name:wwpn:vdisk_UID
950:EEXCLS_HBin01:14:109:HDMCENTEX1N1:10000000C938CFDF:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:109:HDMCENTEX1N1:10000000C938D01F:600507680191011D4800000000000466
950:EEXCLS_HBin01:13:110:HDMCENTEX1N2:10000000C938D65B:600507680191011D4800000000000466
950:EEXCLS_HBin01:13:110:HDMCENTEX1N2:10000000C938D3D3:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:111:HDMCENTEX1N3:10000000C938D615:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:111:HDMCENTEX1N3:10000000C938D612:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:112:HDMCENTEX1N4:10000000C938CFBD:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:112:HDMCENTEX1N4:10000000C938CE29:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:113:HDMCENTEX1N5:10000000C92EE1D8:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:113:HDMCENTEX1N5:10000000C92EDFFE:600507680191011D4800000000000466
```

8.1.8 Server adapter layout

> **Note:** Example 8-4 shows the same volume that is mapped to five different hosts, but host 110 features a different SCSI ID than the other four hosts. This example is a non-recommended practice that can lead to loss of access in some situations because of SCSI ID mismatch.

If your host system includes multiple internal I/O buses, place the two adapters that are used for IBM Spectrum Virtualize cluster access on two different I/O buses to maximize the availability and performance. When purchasing a server, always have two cards instead of one. For example, two dual-port HBA cards are preferred over one quad-port HBA card because you can spread the I/O and add redundancy.

8.1.9 Host status improvements

IBM Spectrum Virtualize provides an alternative for reporting host status.

Previously, a host was marked as *degraded* if one of the host ports logged off the fabric. However, examples exist in which this marking might be normal and can cause confusion.

At the host level, a new **`status_policy`** setting is available that includes the following settings:

- The **`complete`** setting uses the original host status definitions.
- By using the **`redundant`** setting, a host is not reported as *degraded* unless not enough ports are available for redundancy.

8.1.10 NVMe over Fibre Channel host attachments considerations

IBM Spectrum Virtualize now supports a single host initiator port that uses SCSI and NVMe connections to the storage.

Asymmetric Namespace Access was added to the FC-NVMe protocol standard, which gives it functions that are similar to Asymmetric Logical Unit Access (ALUA). As a result, FC-NVMe can now be used in stretched clusters.

IBM Spectrum Virtualize code 8.4.2 allows a maximum of 64 NVMe hosts per system and 16 hosts per I/O group, if no other types of hosts are attached. IBM Spectrum Virtualize code does not monitor or enforce these limits.

For more information if you are planning to use NVMe hosts with IBM FlashSystem, see this IBM Support web page.

> **Note:** Do not map the same volumes to SCSI and NVMe hosts concurrently. Also, take care not to add NVMe hosts and SCSI hosts in the same host cluster.

8.1.11 iSER host attachment considerations

On the IBM FlashSystem, iSCSI Extensions for RDMA (iSER) hosts with different operating systems can be attached to the system. iSER is a network protocol that extends the iSCSI to use Remote Direct Memory Access (RDMA).

If you are planning to use iSER hosts in your IBM FlashSystem, see following links as you plan your environment:

- IBM FlashSystem 9200 8.4.2 Documentation - iSER Ethernet host attachment
- IBM Support: V8.4.2.x Configuration Limits and Restrictions for IBM FlashSystem 9200

8.2 IP multitenancy

IP support for all IBM Spectrum Virtualize products previously allowed only a single IPv4 and IPv6 address per port for use with Ethernet connectivity protocols (iSCSI, iSER).

As of 8.4.2, IBM Spectrum Virtualize removed that limitation and supports an increased per port limit to 64 IP addresses (IPv4, IPv6, or both). The scaling of the IP definition also scaled the VLAN limitation, which can be done per IP address or as wanted.

The OBAC model (that is, OBAC-based per tenant administration and partitioned for multitenant cloud environments), also was added to the Ethernet configuration management. For cloud platforms and environments, each port supports a maximum of two IP addresses and VLANs for multiple clients or tenants that share storage resources.

IBM Spectrum Virtualize code 8.4.2 with it is new IP object model introduced a new feature named the *portset*. The *portset object* is a group of logical addresses that represents a typical IP function and traffic type. Portsets can be used for traffic types, such as host attachment, backend storage connectivity (iSCSI only), or IP replication.

The following commands can be used to manage IP/Ethernet configuration:

- `lsportset`
- `mkportset`
- `chportset`
- `rmportset`
- `lsip (lsportip deprecated)`
- `mkip (cfgportip deprecated)`
- `rmkip (rmportip deprecated)`
- `lsportethernet (lsportip deprecated)`
- `chportethernet (cfgportip deprecated)`
- `mkhost (with parameter -portset to bind the host to portset)`
- `chost (with parameter -portset to bind the host to portset)`

A host can access storage through the IP addresses that are included in the portset that is mapped to the host. The process to bind a host to a portset includes the following steps:

1. Create portset.
2. Configure IPs with the portset.
3. Create a host object.
4. Bind the host to the portset.
5. Discover and login from the host.

IP portsets can be added by using the management GUI or the command-line interface (CLI). You can configure portsets by using the GUI and selecting **Settings** → **Network** → **Portsets**.

Example 8-6 shows the results of the use of the **`lsportset`** command.

Example 8-5 Output of using the lsportset command

```
svcinfo lsportset
id name      type        port_count host_count lossless owner_id owner_name
0  portset0  host        2          8          no
1  portset1  replication 2          0
2  portset2  replication 0          0
3  portset3  storage     0          0
4  myportset host        0          0
```

After the portsets are created, IP addresses can be assigned by using the management GUI or the CLI. You can configure portsets by using the GUI and selecting **Settings** → **Network** → **Ethernet Ports**.

Example 8-6 shows the results of the use of the **lsip** command.

Example 8-6 Output of using the lsip command

```
svcinfo lsip
id node_id node_name port_id portset_id portset_name IP_address   prefix vlan gateway
owner_id owner_name
0  1       node1     1       0          portset0     10.0.240.110 24          10.0.240.9
1  1       node1     1       1          portset1     10.0.240.110 24          10.0.240.9
2  2       node2     1       0          portset0     10.0.240.111 24          10.0.240.9
3  2       node2     1       1          portset1     10.0.240.111 24          10.0.240.9
```

8.2.1 Considerations and limitations

Consider the following points about IP multitenancy:

- Multiple hosts can be mapped to a single portset.
- A single host cannot be mapped to multiple portsets.
- IP addresses can belong to multiple portsets.
- Port masking is used to enable or disable each port per feature for specific traffic types (host, storage, and replication).
- Portset 0, Portset 3, and the replication portset are predefined.
- When an IP address or host is configured, a portset must be specified.
- Portset 0 is the default portset that is automatically configured when the system is updated or created and cannot be deleted.
- Portset 0 allows administrators to continue to use an original configuration that does not require multi-tenancy.
- After an update, all configured host objects are automatically mapped to Portset 0.
- Portset 3 is used for iSCSI back-end storage virtualization.
- Unconfigured logins are rejected upon discovery.
- The iSNS function registers IP addresses in Portset 0 only with the iSNS server.
- Each port can be configured with only one unique routable IP address (gateway specified).

8.3 CSI Block Driver

Container Storage Interface (CSI) enables Container Orchestrators Platform to perform actions on storage systems. The CSI Block Driver connects Kubernetes (K8S) and Red Hat OpenShift Container Platform (OCP) to IBM Block storage devices (Spectrum Virtualize, FlashSystem, DS8K). This process is done through the use of persistent volumes (PVs) to dynamically provision for block storage with stateful containers. Provisioning can be fully automated to scale, deploy, and manage containerized applications. The CSI driver allows hybrid multicloud environments for modern infrastructures.

To use IBM block storage CSI driver, complete the following steps:

1. Create an array secret.
2. Create a storage class.
3. Create a PersistentVolumeClaim (PVC) that is 1 Gb.
4. Display the PVC and the created persistent volume (PV).

5. Create a StatefulSet.

For more information about installing, configuring, and the use of CSI Block Driver, see this IBM Documentation web page.

8.4 Host pathing

Each host mapping associates a volume with a host object and allows all HBA ports in the host object to access the volume. You can map a volume to multiple host objects.

When a mapping is created, multiple paths normally exist across the SAN fabric from the hosts to the IBM Spectrum Virtualize system. Most operating systems present each path as a separate storage device. Therefore, multipathing software is required on the host. The multipathing software manages the paths that are available to the volume, presents a single storage device to the operating system, and provides failover if a path is lost.

If your IBM Spectrum Virtualize system uses NPIV, path failures that occur because of an offline node are masked from host multipathing.

8.4.1 Path selection

I/O for a specific volume is handled exclusively by the nodes in a single I/O group. Although both nodes in the I/O group can service the I/O for the volume, the system prefers to use a consistent node, which is called the *preferred node*. The primary purposes of the use of a preferred node are to have load balancing and to determine which node destages writes to the backend storage.

When a volume is created, an I/O group and preferred node are defined and optionally can be set by the administrator. The owner node for a volume is the preferred node when both nodes are available.

IBM Spectrum Virtualize uses Asymmetric Logical Unit Access (ALUA), as do most multipathing drivers. Therefore, the multipathing driver gives preference to paths to the preferred node. Most modern storage systems use ALUA.

> **Note:** Some competitors claim that ALUA means that IBM Spectrum Virtualize is effectively an active-passive cluster. This claim is not true. Both nodes in IBM Spectrum Virtualize can and do service I/O concurrently.

In the small chance that an I/O goes to the non-preferred node, that node services the I/O without issue.

8.5 I/O queues

Host operating system and HBA 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 use unique methods to control the I/O queue.

The I/O queue can be controlled by using one of the following unique methods:

- Host adapter-based
- Memory and thread resources-based
- Based on the number of commands that are outstanding for a device

8.5.1 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.

Queue depths apply at various levels of the system:

- Disk or flash
- Storage controller
- Per volume and HBA on the host

For example, each IBM Spectrum Virtualize node has a queue depth of 10,000. A typical disk drive operates efficiently at a queue depth of 8. Most host volume queue depth defaults are approximately around 32.

Guidance for limiting queue depths in large SANs that was described in previous documentation was replaced with calculations for overall I/O group based queue depth considerations.

No set rule is available for setting a queue-depth value per host HBA or per volume. The requirements for your environment is driven by the intensity of each workload.

Ensure that one application or host cannot run away and use the entire controller queue. However, if you have a specific host application that requires the lowest latency and highest throughput, consider giving it a proportionally larger share than others.

Consider the following points:

- A single IBM Spectrum Virtualize Fibre Channel port accepts a maximum concurrent queue depth of 2048.
- A single IBM Spectrum Virtualize node accepts a maximum concurrent queue depth of 10,000. After this depth is reached, it reports a full status for the queue.
- Host HBA queue depths must be set to the maximum (typically, 1024).
- Host queue depth must be controlled through the per volume value:
  - A typical random workload volume must use approximately 32
  - To limit the workload of a volume use 4 or less
  - To maximize throughput and give a higher share to a volume, use 64

The total workload capability can be calculated by multiplying the number of volumes by their respective queue depths and summing. With low latency storage, a workload of over 1 million IOPs can be achieved with a concurrency on a single IO Group of 1000.

For more information about queue depths, see the following IBM Documentation web pages:

- FC hosts
- iSCSI hosts
- iSER hosts

8.6 Host clusters

IBM Spectrum Virtualize supports host clusters. This feature allows multiple hosts to access the same set of volumes.

Volumes that are mapped to that host cluster are assigned to all members of the host cluster with the same SCSI ID. A typical use case is to define a host cluster that contains all the WWPNs that belong to the hosts that are participating in a host operating system-based cluster, such as IBM PowerHA®, Microsoft Cluster Server (MSCS), or VMware ESXi clusters.

The following commands can be used to manage host clusters:

- `lshostcluster`
- `lshostclustermember`
- `lshostclustervolumemap`
- `addhostclustermember`
- `chhostcluster`
- `mkhost` (with parameter -`hostcluster` to create the host in one cluster)
- `mkhostcluster`
- `mkvolumehostclustermap`
- `rmhostclustermember`
- `rmhostcluster`
- `rmvolumehostclustermap`

Host clusters can be added by using the GUI. By using the GUI, the system assigns the SCSI IDs for the volumes (you also can manually assign them). For ease of management purposes, it is suggested to use separate ranges of SCSI IDs for hosts and host clusters.

For example, you can use SCSI IDs 0 - 99 for non-cluster host volumes, and greater than 100 for the cluster host volumes. When you choose the **System Assign** option, the system automatically assigns the SCSI IDs starting from the first available in the sequence.

If you choose **Self Assign**, the system enables you to select the SCSI IDs manually for each volume. On the right side of the window, the SCSI IDs are shown that are used by the selected host or host cluster (see Figure 8-1 on page 377).

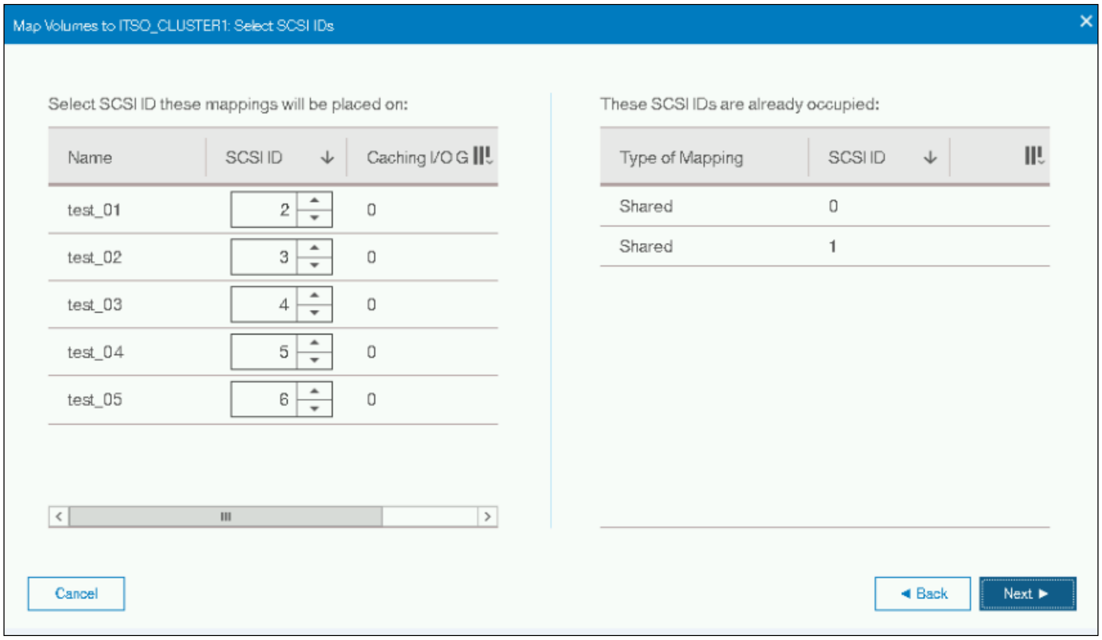


Figure 8-1   SCSI ID assignment on volume mappings

Note: Although extra care is always recommended when dealing with hosts, IBM Spectrum Virtualize does not allow you to join a host into a host cluster if it includes a volume mapping with a SCSI ID that also exists in the host cluster:

```
IBM_2145:ITSO-SVCLab:superuser>addhostclustermember -host ITSO_HOST3
ITSO_CLUSTER1

CMMVC9068E Hosts in the host cluster have conflicting SCSI ID's for their
private mappings.

IBM_2145:ITSO-SVCLab:superuser>
```

8.6.1 Persistent reservations

To prevent hosts from sharing storage inadvertently, establish a storage reservation mechanism. The mechanisms for restricting access to IBM Spectrum Virtualize volumes use the SCSI-3 persistent reserve commands or the SCSI-2 reserve and release commands.

The host software uses several methods to implement host clusters. These methods require sharing the volumes on IBM Spectrum Virtualize between hosts. To share storage between hosts, maintain control over accessing the volumes. Some clustering software uses software locking methods.

You can choose other methods of control by the clustering software or by the device drivers to use the SCSI architecture reserve or release mechanisms. The multipathing software can change the type of reserve that is used from an earlier reserve to persistent reserve, or remove the reserve.

Persistent reserve refers to a set of SCSI-3 standard commands and command options that provide SCSI initiators with the ability to establish, preempt, query, and reset a reservation policy with a specified target device. The functions that are provided by the persistent reserve commands are a superset of the original reserve or release commands.

The persistent reserve commands are incompatible with the earlier reserve or release mechanism. Also, target devices can support only reservations from the earlier mechanism or the new mechanism. Attempting to mix persistent reserve commands with earlier reserve or release commands results in the target device returning a reservation conflict error.

Earlier reserve and release mechanisms (SCSI-2) reserved the entire LUN (volume) for exclusive use down a single path. This approach prevents access from any other host or even access from the same host that uses a different host adapter. The persistent reserve design establishes a method and interface through a reserve policy attribute for SCSI disks. This design specifies the type of reservation (if any) that the operating system device driver establishes before it accesses data on the disk.

The following possible values are supported for the reserve policy:

- `No_reserve`: No reservations are used on the disk.
- `Single_path`: Earlier reserve or release commands are used on the disk.
- `PR_exclusive`: Persistent reservation is used to establish *exclusive host access* to the disk.
- `PR_shared`: Persistent reservation is used to establish *shared host access* to the disk.

When a device is opened (for example, when the AIX **`varyonvg`** command opens the underlying hdisks), the device driver checks the object data manager (ODM) for a `reserve_policy` and a `PR_key_value`. The driver then opens the device. For persistent reserve, each host that is attached to the shared disk must use a unique registration key value.

8.6.2 Clearing reserves

It is possible to accidentally leave a reserve on the IBM Spectrum Virtualize volume or on the IBM Spectrum Virtualize MDisk during migration into IBM Spectrum Virtualize, or when disks are reused for another purpose. Several tools are available from the hosts to clear these reserves.

Instances exist in which a host image mode migration appears to succeed; however, problems occur when the volume is opened for read or write I/O. The problems can result from not removing the reserve on the MDisk before image mode migration is used in IBM Spectrum Virtualize.

You cannot clear a leftover reserve on an IBM Spectrum Virtualize MDisk from IBM Spectrum Virtualize. You must clear the reserve by mapping the MDisk back to the owning host and clearing it through host commands, or through back-end storage commands as advised by IBM technical support.

8.7 AIX hosts

This section discusses support and considerations for AIX hosts.

For more information about configuring AIX hosts, see this IBM Documentation web page.

8.7.1 Multipathing support

Subsystem Device Driver Path Control Module (SDD PCM) is no longer supported. Use the default AIX PCM. For more information, see this IBM Support web page.

8.7.2 AIX configuration recommendations

These device settings can be changed by using the **chdev** AIX command:

```
reserve_policy=no_reserve
```

The default reserve policy is **single_path** (SCSI-2 reserve). Unless a specific need exists for reservations, use **no_reserve**:

```
algorithm=shortest_queue
```

If coming from SDD PCM, AIX defaults to **fail_over**. You cannot set the algorithm to **shortest_queue** unless the reservation policy is **no_reserve**:

```
queue_depth=32
```

The default queue depth is `20`. IBM recommends `32`:

```
rw_timeout=30
```

The default for SDD PCM is `60`; for AIX PCM, the default is `30`. IBM recommends `30`.

For more information about configuration best practices, see AIX Multi Path Best Practices.

8.8 Virtual I/O server hosts

This section discusses support and considerations for virtual I/O server hosts.

For more information about configuring VIOS hosts, see this IBM Documentation web page.

8.8.1 Multipathing support

Subsystem Device Driver Path Control Module (SDDPCM) is no longer supported. Use the default AIX PCM.

For more information, see this IBM Support web page. Where Virtual I/O Server SAN Boot or dual Virtual I/O Server configurations are required, see IBM System Storage Interoperation Center (SSIC).

For more information about VIOS, see this IBM Documentation web page.

8.8.2 VIOS configuration recommendations

These device settings can be changed by using the **chdev** AIX command:

`reserve_policy=single_path`

The default reserve policy is **single_path** (SCSI-2 reserve).

`algorithm=fail_over`

If coming from SDD PCM, AIX defaults to **fail_over**.

`queue_depth=32`

The default queue depth is `20`. IBM recommends `32`:

`rw_timeout=30`

The default for SDD PCM is `60`; for AIX PCM, the default is `30`. IBM recommends `30`.

8.8.3 Physical and logical volumes

Virtual SCSI (VSCSI) is based on a client/server relationship. The Virtual I/O Server (VIOS) owns the physical resources and acts as the server or target device.

Physical storage with attached disks (in this case, volumes on IBM Spectrum Virtualize) on the VIOS partition can be shared by one or more client logical partitions. These client logical partitions contain a virtual SCSI client adapter (scsi initiator) that detects these virtual devices (virtual scsi targets) 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. An LV VSCSI hdisk cannot be served up from multiple VIOSs.

LV VSCSI hdisks are in LVM volume groups on the VIOS and must not span PVs in that volume group or be striped LVs. Because of these restrictions, use PV VSCSI hdisks.

8.8.4 Identifying 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)
- Physical volume identifier (PVID)
- IEEE volume identifier

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. For more information about how to determine your disks identifiers, see this IBM Documentation web page.

8.9 Microsoft Windows hosts

This section discusses support and considerations for Microsoft Windows hosts, including Microsoft Hyper-V.

For more information about configuring Windows hosts, see this IBM Documentation web page.

8.9.1 Multipathing support

For multi-pathing support, use Microsoft MPIO with Microsoft Device Specific Module (MS DSM), which is included in the Windows Server operating system. The older Subsystem Device Driver Device Specific Module (SDDDSM) is no longer supported. For more information, see this SIBM Support web page.

The Windows multipathing software supports the following maximum configuration:

- Up to 8 paths to each volume
- Up to 2048 volumes per windows server/host
- Up to 512 volumes per Hyper-V host

8.9.2 Windows and Hyper-V configuration recommendations

Ensure the following components are configured:

- Operating system service packs and patches and clustered-system software
- HBAs and HBA device drivers
- Multipathing drivers (MSDSM)

Regarding Disk Timeout for Windows Servers, change the disk I/O timeout value to 60 in the Windows registry.

8.10 Linux hosts

This section discusses support and considerations for Linux hosts.

For more information about configuring Linux hosts, see this IBM Documentation web page.

8.10.1 Multipathing support

IBM Spectrum Virtualize supports Linux hosts that use native Device Mapper-Multipathing (DM-MP) and native multipathing support.

> **Note:** Occasionally, we see storage administrators modify parameters in the `multipath.conf` file to address some perceived shortcoming in the DM-MP configuration. These modifications can create unintended and unexpected behaviors. The recommendations that are provided in IBM Documentation are optimal for most configurations.

8.10.2 Linux configuration recommendations

Consider the following points about configuration settings for Linux:

- Settings and udev rules can be edited in `/etc/multipath.conf`.
- Some Linux levels require `polling_interval` to be under the defaults section. If `polling_interval` is under the device section, comment it out by using the # key, as shown in the following example:

  ```
  # polling_interval
  ```

- Use default values as described at this IBM Documentation web page.
- The `dev_loss_tmo settings` control how long to wait for device/paths to be pruned. If the inquiry is too short, it might timeout before paths are available. IBM recommends 120 seconds for this setting.

> **Preferred practice:** The `scsi_mod.inq_timeout` should be set to 70. If this timeout is set incorrectly, it can cause paths to not be rediscovered after a node is restarted.
>
> For more information and this setting and other attachment requirements, see this IBM Documentation web page.

8.11 Oracle Solaris hosts support

This section discusses support and considerations for Oracle hosts. SAN boot and clustering support is available for Oracle hosts.

For more information about configuring Solaris hosts, see this IBM Documentation web page.

8.11.1 Multipathing support

IBM Spectrum Virtualize supports multipathing for Oracle Solaris hosts through Oracle Solaris MPxIO, Symantec Veritas Volume Manager Dynamic Multipathing (DMP), and the Native Multipathing Plug-in (NMP). Specific configurations are dependent on file system requirements, HBA, and operating system level.

> **Note:** The Native Multipathing Plug-In (NMP) does not support the Solaris operating system in a clustered-system environment. For more information about your supported configuration, see IBM System Storage Interoperation Center (SSIC).

8.11.2 Solaris MPxIO configuration recommendations

IBM Spectrum Virtualize software supports load balancing of the MPxIO software. Ensure the host object is configured 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 `hpux`, `tpgs`, `generic`, `openvms`, `adminlun`, and `hide_secondary`. The `tpgs` host type enables extra target port unit attentions required by Solaris hosts.

Complete your configuration by using the following process:

1. Configure host objects with host type `tpgs`.
2. Install the latest Solaris host patches.
3. Copy the `/kernel/drv/scsi_vhci.conf` file to the `/etc/driver/drv/scsi_vhci.conf` file.
4. Set the `load-balance="round-robin"` parameter.
5. Set the `auto-failback="enable"` parameter.
6. Comment out the `device-type-scsi-options-list = "IBM     2145", "symmetric-option"` parameter.
7. Comment out the `symmetric-option = 0x1000000` parameter.
8. Reboot hosts or run **`stmsboot -u`** based on the host level.
9. Verify changes by running **`luxadm display /dev/rdsk/cXtYdZs2`** where **`cXtYdZs2`** is your storage device.
10. Check that preferred node paths are primary and online and non-preferred node paths are secondary and online.

8.11.3 Symantec Veritas DMP configuration recommendations

When you are managing IBM Spectrum Virtualize storage in Symantec volume manager products, you must install an Array Support Library (ASL) on the host so that the volume manager is aware of the storage subsystem properties (active/active or active/passive).

If the suitable ASL is not installed, the volume manager did not claim the LUNs. Use of the ASL is required to enable the special failover or failback multipathing that IBM Spectrum Virtualize requires for error recovery.

Use the commands that are shown in Example 8-7 to determine the basic configuration of a Symantec Veritas server.

Example 8-7 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 ctlr=cX (lists paths by controller)
vxdmpadm getsubpaths dmpnodename=cxtxdxs2’ (lists paths by LUN)
```

The commands that are shown in Example 8-8 on page 384 and Example 8-9 on page 384 determine whether the IBM Spectrum Virtualize is correctly connected. They also show which ASL is used: native Dynamic Multi-Pathing (DMP), ASL, or SDD ASL.

Example 8-8 on page 384 shows what you see when Symantec Volume Manager correctly accesses IBM Spectrum Virtualize by using the SDD pass-through mode ASL.

Example 8-8 Symantec Volume Manager that uses SDD pass-through mode ASL

```
# vxdmpadm list enclosure all
ENCLR_NAME ENCLR_TYPE ENCLR_SNO STATUS
============================================================
OTHER_DISKS OTHER_DISKS OTHER_DISKS CONNECTED
VPATH_SANVCO VPATH_SANVC 0200628002faXX00 CONNECTED
```

Example 8-9 shows what you see when IBM Spectrum Virtualize is configured by using native DMP ASL.

Example 8-9 IBM Spectrum Virtualize that is configured by using native ASL

```
# vxdmpadm list enclosure all
ENCLR_NAME ENCLR_TYPE ENCLR_SNO STATUS
============================================================
OTHER_DISKS OTHER_DSKSI OTHER_DISKS CONNECTED
SAN_VC0 SAN_VC 0200628002faXX00 CONNECTED
```

For more information about the latest ASL levels to use native DMP, see the array-specific module table that is available at this Veritas web page.

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 **vxdctl enable** command. To list the ASLs that are active, enter the following command:

```
vxddladm listsupport
```

8.12 HP 9000 and HP Integrity hosts

This section discusses support and considerations for Linux hosts. SAN boot is supported for all HP-UX 11.3x releases on both HP 9000 and HP Integrity servers.

For more information about configuring Linux hosts, see this IBM Documentation web page.

8.12.1 Multipathing support

IBM Spectrum Virtualize supports multipathing for HP-UX hosts through HP PVLinks and the Native Multipathing Plug-in (NMP). Dynamic multipathing is available when you add paths to a volume or when you present a new volume to a host.

To use PVLinks while NMP is installed, ensure that NMP did not configure a vpath for the specified volume.

For more information about a list of configuration maximums, see this IBM Documentation web page.

8.12.2 HP configuration recommendations

Consider the following configuration recommendations for HP:

- HP-UX versions 11.31 September 2007 and later 0803 releases are supported.
- HP-UX version 11.31 contains Native Multipathing as part of the mass storage stack feature.
- Native Multipathing Plug-in supports only HP-UX 11iv1 and HP-UX 11iv2 operating systems in a clustered-system environment.
- SCSI targets that will use more than 8 LUNs must have type attribute hpux set to host object.
- Ensure the host object is configured with the type attribute set to hpux as shown in the following example:

  ```
  svctask mkhost -name new_name_arg -hbawwpn wwpn_list -type hpux
  ```

- Configure the Physical Volume timeout for NMP for 90 seconds.
- Configure the Physical Volume timeout for PVLinks for 60 seconds (the default is 4 minutes).

8.13 VMware ESXi server hosts

This section discusses considerations for VMware hosts.

For more information about configuring VMware hosts, see this IBM Documentation web page.

To determine the various VMware ESXi levels that are supported, see the IBM System Storage Interoperation Center (SSIC).

8.13.1 Multipathing support

VMware features a built-in multipathing driver that supports IBM Spectrum Virtualize ALUA-preferred path algorithms.

The VMware multipathing software supports the following maximum configuration:

- A total of 256 SCSI devices
- Up to 32 paths to each volume
- Up to 4096 paths per server

> **Tip:** Each path to a volume equates to a single SCSI device.

For more information about a complete list of maximums, see VMware Configuration Maximums.

8.13.2 VMware configuration recommendations

For more information about specific configuration best practices for VMware, see this CIBM Documentation wbe page.

Consider and verify the following settings:

- The storage array type plug-in should be ALUA (VMW_SATP_ALUA).
- Path selection policy should be RoundRobin (VMW_PSP_RR).
- The Round Robin IOPS should be changed from 1000 to 1 so that I/Os are evenly distributed across as many ports on the system as possible. For more information about how to change this setting, see this web page.
- If preferred, all VMware I/O paths (active optimized and non-optimized) can be placed in use by issuing the following `esxcli` command:

  ```
  (esxcli storage nmp psp roundrobin deviceconfig set --useano=1 -d <naa of the device>)
  ```

For more information about active optimized and active non-optimized paths, see this IBM Documentation web page.

> **Note:** For more information about IBM i-related considerations, see Appendix A, “IBM i considerations” on page 601.

9

Implementing a storage monitoring system

Monitoring in a storage environment is crucial and it is part of what usually is called *storage governance*.

With a robust and reliable storage monitoring system, you can save significant money and minimize pain in your operation, by monitoring and predicting utilization bottlenecks in your storage environment.

This chapter provides suggestions and the basic concepts of how to implement a storage monitoring system for IBM FlashSystem, using specific functions or external IBM Tools.

This chapter includes the following topics:

9.1 Generic monitoring

With IBM FlashSystem, you can implement generic monitoring using IBM FlashSystem-specific functions that are integrated with the product without adding external tools or cost.

9.1.1 Monitoring by using the GUI

The management GUI is the primary tool that is used to service your system. Regularly monitor the status of the system by using the management GUI. If you suspect a problem, use the management GUI first to diagnose and resolve the problem.

Use the views that are available in the management GUI to verify the status of the system, the hardware devices, the physical storage and the available volumes. The Events window provides access to all problems that exist on the system. Use the Recommended Actions filter to display the most important events that need to be resolved.

If a service error code exists for the alert, you can run a fix procedure that assists you in resolving the problem. These fix procedures analyze the system and provide more information about the problem. These actions also ensure that the required changes do not cause volumes to be inaccessible to the hosts and automatically perform configuration changes that are required to return the system to its optimum state.

If any interaction is required, they suggest actions to take and guide you through those actions that automatically manage the system where necessary. If the problem is fixed, the alert is excluded.

Call Home

Call Home connects your system to service representatives who can monitor issues and respond to problems efficiently and quickly to keep your system up and running. The Call Home feature transmits operational and event-related data to you and IBM through a Simple Mail Transfer Protocol (SMTP) server or cloud services connection through Representational State Transfer (RESTful) APIs.

The SMTP sends notifications through an email server to support to fix errors, to internal users or services that monitor activity on the system.

Multiple email addresses can be added to receive notifications from the storage. You also can set notification options for each email box that you added with different sets of information (see Figure 9-1).

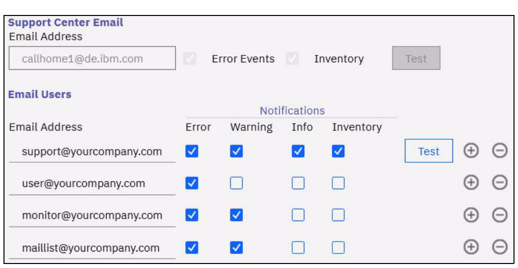


Figure 9-1 Email users

Representational State Transfer (RESTful) APIs transmitting data through web services. You can also specify an internal proxy server to manage outbound connections with the support center (see Figure 9-2).

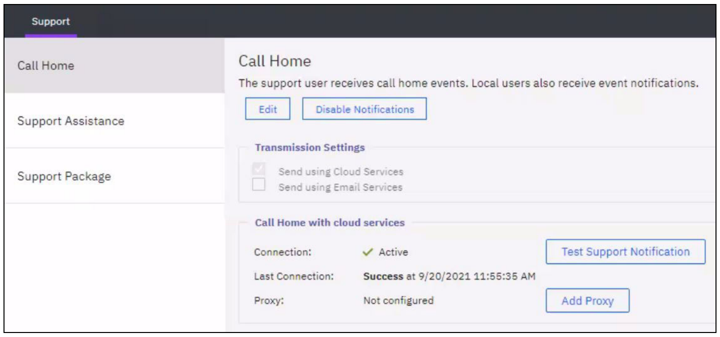


Figure 9-2 Call Home with cloud services

From a monitoring perspective, email notification is one of the most common and important tools that you can use and set up. From the notification events, you can validate if your system is running under normal status or needs attention.

SNMP notification

Simple Network Management Protocol (SNMP) is a standard protocol for managing networks and exchanging messages. The system can send SNMP messages that notify personnel about an event. You can use an SNMP manager to view the SNMP messages that are sent by the IBM FlashSystem system.

The Management Information Base (MIB) file describes the format of the SNMP messages that are sent by IBM FlashSystem. Use the MIB file to configure a network management program to receive SNMP event notifications that are sent from an IBM FlashSystem system. This MIB file is suitable for use with SNMP messages from all versions of IBM FlashSystem.

For more information about IBM FlashSystem MIB file for 8.4.2 FlashSystem devices, see this IBM Support web page.

Syslog notification

The syslog protocol is a standard protocol for forwarding log messages from a sender to a receiver on an IP network. The IP network can be IPv4 or IPv6. The system can send Syslog messages that notify personnel about an event. You can configure a syslog server to receive log messages from various systems and store them in a central repository.

Figure 9-3 shows the new syslog grid layout from the IBM FlashSystem GUI. You can configure multiple syslog servers and monitor the communication between IBM FlashSystem to the syslog server from the syslog window.

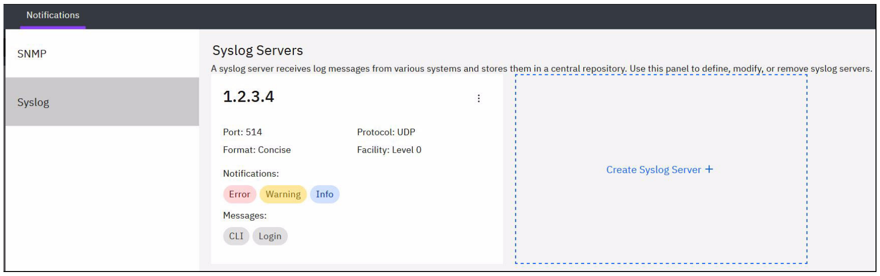


Figure 9-3 Syslog layout

> **Note:** Starting with version 8.4, it is possible to use FQDN for services, such as Syslog, LDAP, and NTP.

9.1.2 Monitoring using quotas and alert

In an IBM Spectrum Virtualize system, the space use of storage pools and thin provisioned or compressed VDisks can be monitored by setting some specific quota alerts. These quota alerts can be defined in the management GUI and through the CLI.

Storage pool

At the storage pool level, an integer defines a threshold at which a warning is generated. The warning is generated the first time that the threshold is exceeded by the used disk capacity in the storage pool. The threshold can be specified with a percentage (see Figure 9-4) or size (see Example 9-1) value.

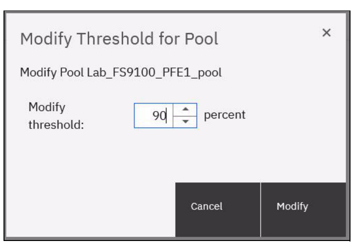


Figure 9-4 Pool threshold

Example 9-1 Threshold specified as a size

```
IBM_FlashSystem:FS9200:superuser>svctask chmdiskgrp -warning 1 -unit tb 3
```

VDisk

At the VDisk level, a warning is generated when the used disk capacity on the thin-provisioned or compressed copy first exceeds the specified threshold. The threshold can be specified with a percentage (see Figure 9-5) or size (see Example 9-2) value.

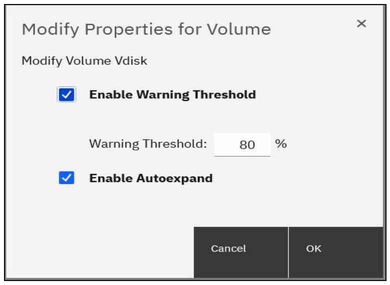


Figure 9-5 VDisk threshold

Example 9-2 Threshold specified as a value

```
IBM_FlashSystem:FS9200:superuser>svctask chvdisk -copy 0 -warning 1 -unit gb 0
```

Note: You can specify a `disk_size` integer, which defaults to megabytes (MB) unless the **`-unit`** parameter is specified. Or, you can specify a `disk_size%`, which is a percentage of the storage pool size. To disable warnings, specify 0 or `0%`. The default value is `0`.

9.2 Performance monitoring

The ability to collect historical performance metrics is essential to properly monitor and manage storage subsystems and IBM FlashSystem. During troubleshooting and performance tuning, the historical data can be used as a parameter for changes and fixes.

The next sections show which performance analysis tools are integrated with IBM FlashSystem and which IBM external tools are available to collect performance statistics for historical retention.

Remember that performance statistics are useful to debug or prevent some potential bottlenecks, and make capacity planning for future growth easier.

9.2.1 On-board performance monitoring

In IBM FlashSystem, real-time performance statistics provide short-term status information for your systems. The statistics are shown as graphs in the management GUI.

You can use system statistics to monitor the bandwidth of all the volumes, interfaces, and MDisks that are used on your system. You can also monitor the overall CPUs usage for the system. These statistics also summarize the overall performance health of the system.

You can monitor changes to stable values or differences between related statistics, such as the latency between volumes and MDisks. These differences can then be further evaluated by performance diagnostic tools.

With system-level statistics, you also can quickly view bandwidth of volumes, interfaces, and MDisks. Each of these graphs displays the current bandwidth in megabytes per second and a view of bandwidth over time.

Each data point can be accessed to determine its individual bandwidth use and to evaluate whether a specific data point might represent performance impacts. For example, you can monitor the interfaces, such as for Fibre Channel or SAS interfaces, to determine whether the host data-transfer rate is different from the expected rate.

You can also select node-level statistics, which can help you determine the performance effect of a specific node. As with system statistics, node statistics help you to evaluate whether the node is operating within normal performance metrics.

The CPU utilization graph shows the current percentage of CPU usage and specific data points on the graph that show peaks in utilization. If compression is being used, you can monitor the amount of CPU resources that are being used for compression and the amount that is available to the rest of the system.

The Interfaces graph displays data points for Fibre Channel (FC), iSCSI, serial-attached SCSI (SAS), and IP Remote Copy interfaces. You can use this information to help determine connectivity issues that might affect performance.

The Volumes and MDisks graphs in the Performance window show four metrics: Read, Write, Read latency, and Write latency. You can use these metrics to help determine the overall performance health of the volumes and MDisks on your system. Consistent unexpected results can indicate errors in configuration, system faults, or connectivity issues.

Each graph represents 5 minutes of collected statistics, which are updated every 5 seconds. They also provide a means of assessing the overall performance of your system, as shown in Figure 9-6.

Figure 9-6 Monitoring/Performance overview

You can then choose the metrics that you want to be displayed, as shown in Figure 9-7.

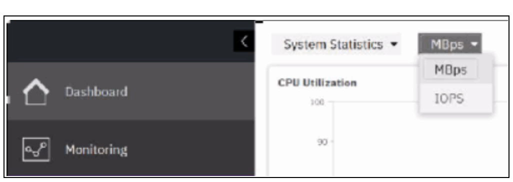


Figure 9-7 Metrics

You can also obtain a quick overview by using the GUI option by selecting **System** → **Dashboard** (see Figure 9-8).

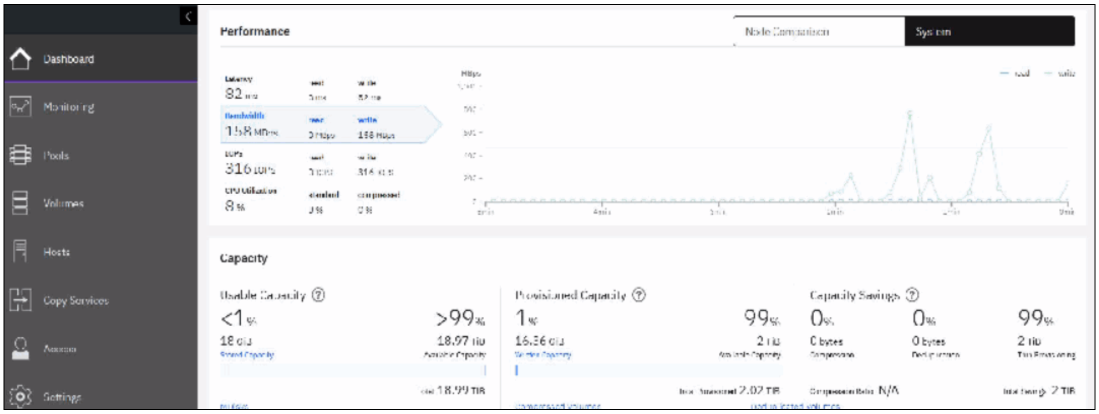


Figure 9-8 Management GUI Dashboard

9.2.2 Performance monitoring with IBM Spectrum Control

IBM Spectrum Control is an on-premises storage management, monitoring, and reporting solution. It uses the metadata that it collects about vendors' storage devices to provide services such as custom alerting, analytics, and replication management. Both IBM Spectrum Control and IBM Storage Insights monitor storage systems, but IBM Spectrum Control also monitors hypervisors, fabrics, and switches to provide you with unique analytics and insights into the topology of your storage network.

It also provides more granular collection of performance data, with 1-minute intervals rather than the 5-minute intervals in IBM Storage Insights or IBM Storage Insights Pro.

Because IBM Spectrum Control is an on-premises tool, it does not send the metadata about monitored devices offsite, which is ideal for dark shops and sites that do not want to open ports to the cloud.

For more information about the capabilities of IBM Spectrum Control, see this IBM Documentation web page.

For more information about pricing and purchasing, see this web page.

For more information about IBM Storage Insights, see *Getting Started with IBM Storage Insights*.

> **Note:** If you use IBM Spectrum Control or manage IBM block storage systems, you can access the free of charge version of IBM Storage Insights.

IBM Spectrum Control offers several reports that you can use to monitor IBM FlashSystem systems to identify performance problems. IBM Spectrum Control provides improvements to the web-based user interface that is designed to offer easy access to your storage environment.

IBM Spectrum Control provides a large amount of detailed information about IBM FlashSystem. The next sections provide basic suggestions about the metrics that need to be monitored and analyzed to debug potential bottleneck problems.

In addition, which alerts need to be set to be notified when some specific metrics exceed limits that are considered important for this specific environment.

For more information about the installation, configuration, and administration of IBM Spectrum Control (including how to add a storage system), see this IBM Documentation web page.

Note: IBM Spectrum Control 5.3.0 or higher is recommended for monitoring IBM FlashSystem Version 8.4.2.

IBM Spectrum Control dashboard

The IBM Spectrum Control dashboard provides a status overview of all monitored resources and identifies potential problem areas in a storage environment:

- Condition and usage of resources.
- Entities that consume storage on those resources.
- Number and status of unacknowledged alert conditions that are detected on the monitored resources.
- Most active storage systems in your environment.

Figure 9-9 shows the Spectrum Control dashboard.

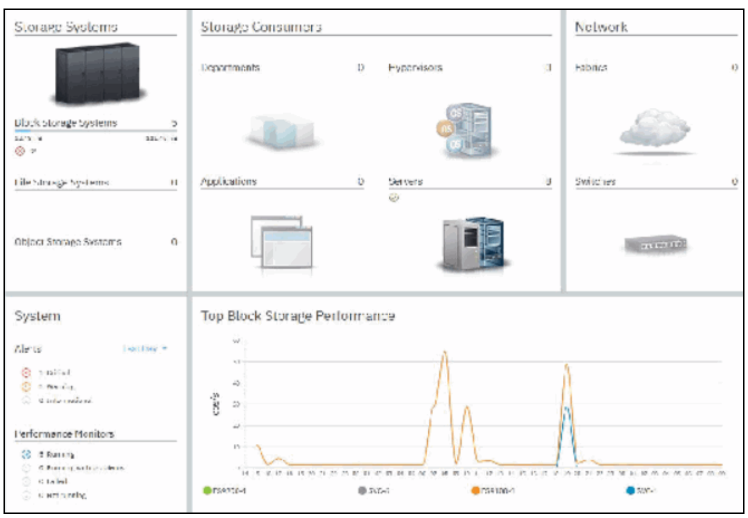


Figure 9-9 Spectrum Control Dashboard

Key Performance Indicators

IBM Spectrum Control provides Key Performance Indicators (in earlier releases, Best Practice Performance Guidelines) for the critical monitoring metrics. These guidelines do not represent the maximum operating limits of the related components. Instead, they suggest limits that are selected with an emphasis on maintaining a stable and predictable performance profile.

The Key Performance Indicators web interface of IBM Spectrum Control (see Figure 9-10 on page 395) displays by default the last 24 hours from the active viewing time and date. Selecting an element from the chart overlays the corresponding 24 hours for the previous day and seven days prior. This display allows for an immediate historical comparison of the respective metric. The day of reference can also be changed to allow historical comparison of previous days.

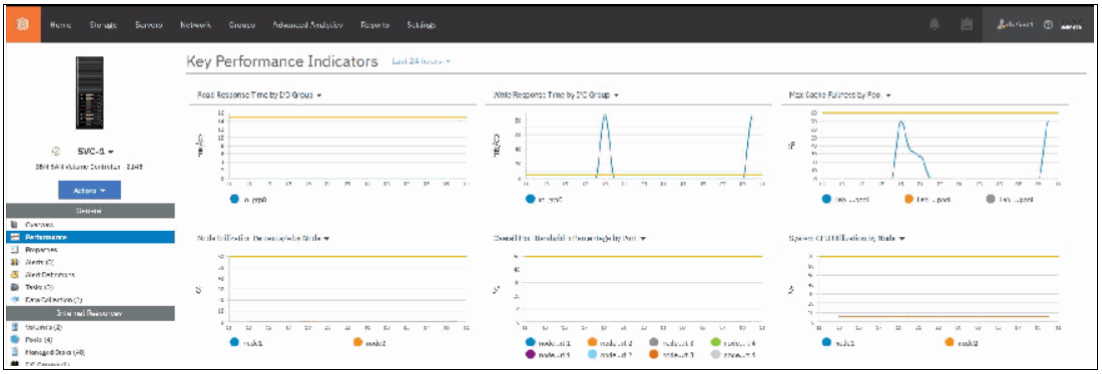


Figure 9-10 Key Performance Indicators

Note: The window was renamed to *Key Performance Indicators*.

The yellow line indicates the best practice value for the metric. These guidelines are established as the levels that allow for a diverse set of workload characteristics while maintaining a stable performance profile. The other lines on each chart represent the measured values for the metric for the resources on your storage system: I/O groups, ports, or nodes.

You can use the lines to compare how close your resources are to potentially becoming overloaded. If your storage system is responding poorly and the charts indicate overloaded resources, the workload might need to be better balanced. You can balance the workload between the hardware of the cluster, which can add hardware to the cluster, or move some workload to other storage systems.

The charts show the hourly performance data that is measured for each resource on the selected day. Use the following charts to compare the workloads on your storage system with the best practice guidelines:

- Node Utilization Percentage by Node

  Compare the guideline value for this metric (for example, 60% utilization) with the measured value from your system. The average of the bandwidth percentages of those ports in the node that are actively used for host and MDisk send and receive operations. The average is weighted by port speed and adjusted according to the technology limitations of the node hardware. This chart is empty for clusters without FC ports (or when no host I/O is occurring).

- Overall Port Bandwidth Percentage by Port

  Compare the guideline value for this metric (for example, 50%) with the measured value from your system. Because a cluster can have many ports, the chart shows only the eight ports with the highest average bandwidth over the selected day.

- Port-to-Local Node Send Response Time by Node

  Compare the guideline value for this metric (for example, 0.6 ms/op) with the measured value from your system.

- Port-to-Remote Node Send Response Time by Node

  Because latencies for copy-services operations can vary widely, a guideline is not established for this metric. Use this chart to identify any discrepancies between the data rates of different nodes.

- Read Response Time by I/O Group:

  Compare the guideline value for this metric (for example, 15 ms/op) with the measured value from your system. When this metric is constantly breached, an issue with the hardware might exist.

- System CPU Utilization by Node

  Compare the guideline value for this metric (for example, 70% utilization) with the measured value from your system.

- Total Data Rate by I/O Group

  Because data rates can vary widely, a guideline is not established for this metric. Use this chart to identify any significant discrepancies between the data rates of different I/O groups because these discrepancies indicate that the workload is not balanced.

- Write Response Time by I/O Group

  Compare the guideline value for this metric (for example, 5 ms/op) with the measured value from your system.

- Zero Buffer Credit Percentage by Node

  Compare the guideline value for this metric (for example, 20%) with the measured value from your system.

Figure 9-11 shows an example of the Write Response Time by I/O Group, which exceeded the best practice limit (yellow line). The drop-down menu provides more options.

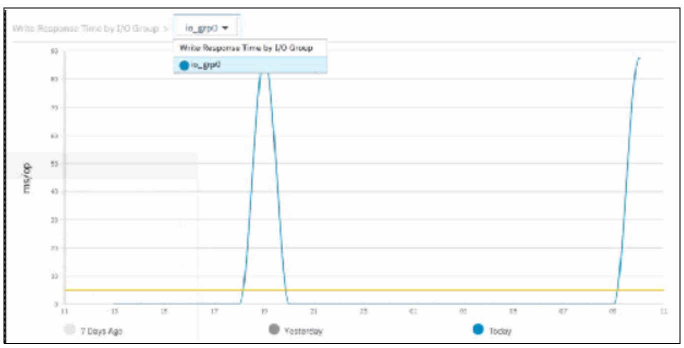


Figure 9-11 Write Response Time by I/O Group

> **Note:** The guidelines are not thresholds and they are not related to the alerting feature in IBM Spectrum Control. To create performance alerts that use the guidelines as thresholds, open a resource detail window in the web-based GUI, click **Alerts** in the General section and then, click **Definitions**.

9.2.3 Performance monitoring with IBM Storage Insights

IBM Storage Insights (ISI) is an off-premises, IBM Cloud service that provides cognitive support capabilities, monitoring, and reporting for storage systems. Because it is an IBM Cloud service, getting started is simple and upgrades are handled automatically.

By using the IBM Cloud infrastructure, IBM Support can monitor your storage environment to help minimize the time to resolution of problems and collect diagnostic packages without requiring you to manually upload them. This wraparound support experience, from environment to instance, is unique to IBM Storage Insights and transforms how and when you get help.

IBM Storage Insights is a SaaS (Software as a Service) offering with its core running over IBM Cloud. IBM Storage Insights provides an unparalleled level of visibility across your storage environment to help you manage complex storage infrastructures and make cost-saving decisions. It combines proven IBM data management leadership with IBM analytics leadership from IBM Research® and a rich history of storage management expertise with a cloud delivery model, enabling you to take control of your storage environment.

As a cloud-based service, it enables you to deploy quickly and save storage administration time while optimizing your storage. It also helps automate aspects of the support process to enable faster resolution of issues. ISI optimizes storage infrastructure using cloud-based storage management and support platform with predictive analytics.

It allows you to optimize performance and to tier your data and storage systems for the right combination of speed, capacity, and economy. IBM Storage Insights provides comprehensive storage management, helps to keep costs low, and can prevent downtime and loss of data or revenue. IBM Storage Insights Key features are:

- Rapid results when you need them
- Single-pane view across your storage environment
- Performance analyses at your fingertips
- Valuable insight from predictive analytics
- Two editions that meet your needs
- Simplified, comprehensive, and proactive product support

Figure 9-12 shows an IBM Storage Insight® example window.

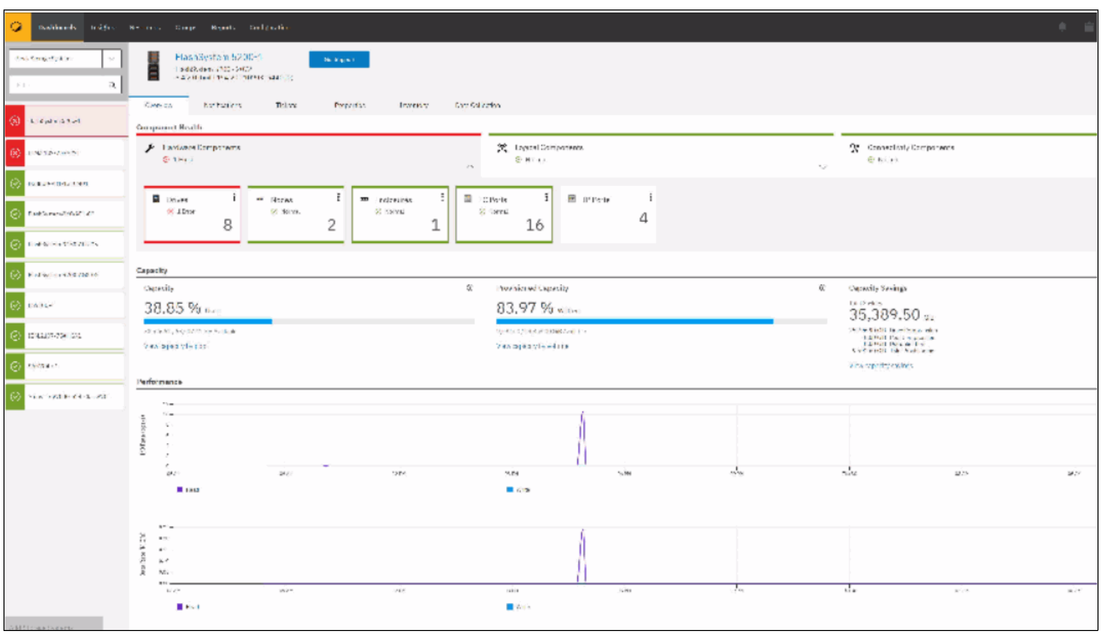


Figure 9-12 Storage Insights dashboard

Understanding the security and data collection features of IBM Storage Insights Pro and IBM Storage Insights can help address the concerns of administrators and IT professionals who deploy the products in their environments and want to learn more about security and data collection. For more information, see this IBM Documentation web page.

Note: IBM strongly recommends the use of IBM Storage Insights or IBM Spectrum Control for a better user experience. IBM Storage Insights requires the use of data collectors. The method of data collection has recently changed to improve security and ease of management. It is no longer required that you have a user with admin privileges for data collectors. A simple monitor user can get status information from the management node.

Licensing and editions of IBM Storage Insights

Several editions of IBM Storage Insights enable you to select the capabilities that serve your needs best. Licensing is implemented through different subscription levels.

- The free of charge version is called *IBM Storage Insights* and provides a unified view of a storage environment with a diagnostic events feed, an integrated support experience, and key capacity and performance metrics. IBM Storage Insights is available at no cost to IBM Storage Insights Pro subscribers and owners of IBM block storage systems who sign up. IBM Storage Insights provides an environment overview, integration in support processes, and shows you IBM analysis results.
- The capacity-based, subscription version is called *IBM Storage Insights Pro* and includes all of the features of IBM Storage Insights plus a more comprehensive view of the performance, capacity, and health of storage resources. It also helps you reduce storage costs and optimize your data center by providing features like intelligent capacity planning, storage reclamation, storage tiering, and advanced performance metrics.

  The storage systems that you can monitor are expanded to include IBM file, object, software-defined storage (SDS) systems, and non-IBM block and file storage systems, such as EMC storage systems.

  In both versions, when problems occur on your storage, you can get help to identify and resolve those problems and minimize potential downtime, where and when you need it.

Table 9-1 lists the different features of both versions.

Table 9-1 Features in IBM Storage Insights and IBM Storage Insights Pro

Resource Management	**Functions**	**IBM Storage Insights (free)**	**IBM Storage Insights Pro (subscription)**
Monitoring	Inventory management	IBM block storage	IBM and non-IBM block storage, file storage, and object storage
	Logical configuration	Basic	Advanced
	Health	Call Home events	Call Home events
	Performance	Basic: ► Three storage system metrics: I/O rate, data rate, and response times aggregated for storage systems ► Four switches metrics: port saturation, port congestion, port hardware errors, and port logical errors	Advanced: ► 100+ metrics for storage systems and their components ► 40+ metrics for switches and related components
	Capacity	Basic Four metrics: allocated space, available space, total space, and compression savings aggregated for storage systems	Advanced (25+ metrics for storage systems and their components)
	Drill down performance workflows to enable deep troubleshooting		✓
	Explore virtualization relationships		✓
	Explore replication relationships		✓
	Retention of configuration and capacity data	Only the last 24 hours is shown	2 years
	Retention of performance data	Only the last 24 hours is shown	1 year
	Reporting		✓

Resource Management	Functions	IBM Storage Insights (free)	IBM Storage Insights Pro (subscription)
Service	Filter events to quickly isolate trouble spots	✓	✓
	Hassle-free log collection	✓*	✓
	Simplified ticketing	✓	✓
	Show active PMRs and ticket history	✓*	✓
Reporting	Inventory, capacity, performance, and storage consumption reports	► Capacity reports for block storage systems and pools ► Inventory reports for block storage systems	All reports
Analytics and optimization	Predictive Alerts	✓	✓
	Customizable, multi-conditional alerting, including alert policies		✓
	Performance planning		✓
	Capacity planning		✓
	Business impact analysis (applications, departments, and groups)		✓
	Optimize data placement with tiering		✓
	Optimize capacity with reclamation		✓
Security	ISO/IEC 27001 Information Security Management standards certified	✓	✓
Entitlements		Free	Capacity-based subscription

Restriction: If you can access IBM Storage Insights but are not an IBM Storage Insights Pro subscriber, you must have a warranty or maintenance agreement for an IBM block storage system to open tickets and send log packages.

IBM Storage Insights for IBM Spectrum Control

IBM Storage Insights for IBM Spectrum Control is an IBM Cloud service that can help you predict and prevent storage problems before they impact your business. It is complementary to IBM Spectrum Control and is available at no additional cost if you have an active license with a current subscription and support agreement for IBM Virtual Storage Center, IBM Spectrum Storage Suite, or any edition of IBM Spectrum Control.

As an on-premises application, IBM Spectrum Control doesn't send the metadata about monitored devices offsite, which is ideal for dark shops and sites that do not want to open ports to the cloud. However, if your organization allows for communication between its network and the cloud, you can use IBM Storage Insights for IBM Spectrum Control to transform your support experience for IBM block storage.

IBM Storage Insights for IBM Spectrum Control and IBM Spectrum Control work hand in hand to monitor your storage environment. IBM Storage Insights for IBM Spectrum Control can transform your monitoring and support experiences:

- Open, update, and track IBM Support tickets easily for your IBM block storage devices.
- Get hassle-free log collection by allowing IBM Support to collect diagnostic packages for devices so you do not have to collect such information.
- Use Call Home to monitor devices, get best practice recommendations, and filter events to quickly isolate trouble spots.
- Use IBM Support's ability to view the current and historical performance of your storage systems and help reduce the time-to-resolution of problems.

You can use IBM Storage Insights for IBM Spectrum Control for as long as you have an active license with a current subscription and support agreement for IBM Spectrum Control license. If your subscription and support lapses, you're no longer eligible for IBM Storage Insights for IBM Spectrum Control. To continue using IBM Storage Insights for IBM Spectrum Control, simply renew your IBM Spectrum Control license. You can also choose to subscribe to IBM Storage Insights Pro.

IBM Spectrum Control and IBM Storage Insights for IBM Spectrum Control feature comparison

To understand the usability of IBM Spectrum Control and IBM Storage Insights for IBM Spectrum Control for your environment, we compare the features of IBM Spectrum Control and IBM Storage Insights for IBM Spectrum Control.

Table 9-2 lists the features in IBM Spectrum Control and IBM Storage Insights for IBM Spectrum Control.

Table 9-2 IBM Spectrum Control and IBM Storage Insights for IBM Spectrum Control comparison

Resource Management	**Features**	**IBM Spectrum Control (Advanced edition)**	**IBM Storage Insights for IBM Spectrum Control**
Monitoring	Inventory	IBM and non-IBM block storage, file storage, object storage, hypervisors, fabrics, switches	IBM and non-IBM block storage, file storage, and object storage
	Call Home events		✓
	Performance	✓ (1-minute intervals)	✓ (5-minute intervals)
	Capacity	✓	✓
	Drill down performance workflow to troubleshoot bottlenecks	✓	✓
	Explore virtualization relationships		
	Explore replication relationships	✓	✓
	Retain performance data		
Service	Deployment method		
	Filter Call Home events to quickly isolate trouble spots		✓
	Hassle-free log collection		✓
	Simplified ticketing		✓
	Show active PMRs and ticket history		✓
	Active directory and LDAP integration for managing users	✓	
Reporting	Inventory, capacity, performance, and storage consumption reports	✓	✓
	Rollup reporting	✓	
	REST API	✓	

Resource Management	Features	IBM Spectrum Control (Advanced edition)	IBM Storage Insights for IBM Spectrum Control
Alerting	Predictive Alerts	✓	✓
	Customizable, multi-conditional alerting, including alert policies	✓	✓
Analytics	Performance planning	✓	✓
	Capacity planning	✓	✓
	Business impact analysis (applications, departments, and groups)	✓	✓
	Provisioning with service classes and capacity pools	✓	
	Balance workload across pools	✓	
	Optimize data placement with tiering	✓	✓
	Optimize capacity with reclamation	✓	✓
	Transform and convert volumes	✓	
Pricing		On-premises licensing	No charge for IBM Spectrum Control customers

You can upgrade IBM Storage Insights to IBM Storage Insights for IBM Spectrum Control if you have an active license of IBM Spectrum Control. For more information, see this web page. At this web page, choose the option for IBM Spectrum Control (log in required).

IBM Storage Insights for IBM Spectrum Control does not include the service level agreement for IBM Storage Insights Pro. Terms and conditions for IBM Storage Insights for IBM Spectrum Control are available at this IBM Support web page.

IBM Storage Insights, IBM Storage Insights Pro, and IBM Storage Insights for IBM Spectrum Control show some similarities, but there are differences:

- *IBM Storage Insights* is an off-premises, IBM Cloud service that is available free of charge if you own IBM block storage systems. It provides a unified dashboard for IBM block storage systems with a diagnostic events feed, a streamlined support experience, and key capacity and performance information.
- *IBM Storage Insights Pro* is an off-premises, IBM Cloud service that is available on subscription and expands the capabilities of IBM Storage Insights. You can monitor IBM file, object, and software-defined storage (SDS) systems, and non-IBM block and file storage systems such as Dell/EMC storage systems.

It also includes configurable alerts and predictive analytics that help you to reduce costs, plan capacity, and detect and investigate performance issues. You get recommendations for reclaiming unused storage, recommendations for optimizing the placement of tiered data, capacity planning analytics, and performance troubleshooting tools.

- *IBM Storage Insights for IBM Spectrum Control* is similar to IBM Storage Insights Pro in capability and is available for no additional cost if you have an active license with a current subscription and support agreement for IBM Virtual Storage Center, IBM Spectrum Storage Suite, or any edition of IBM Spectrum Control.

IBM Spectrum Storage Suite

IBM Spectrum Storage Suite gives you unlimited access to the IBM Spectrum Storage software family and IBM Cloud Object Storage software with licensing on a flat, cost-per-TB basis to make pricing easy to understand and predictable as capacity grows.

Structured specifically to meet changing storage needs, the suite is ideal for organizations just starting out with software-defined storage, and for those with established infrastructures who need to expand their capabilities. The IBM Spectrum Storage Suite consists of the following products:

- IBM Spectrum Control: Analytics-driven hybrid cloud data management to reduce costs
- IBM Spectrum Protect: Optimized hybrid cloud data protection to reduce backup costs
- IBM Spectrum Protect Plus: Complete VM protection and availability that is easy to set up and manage yet scalable for the enterprise
- IBM Spectrum Archive: Fast data retention that reduces total cost of ownership for active archive data
- IBM Spectrum Virtualize: Virtualization of mixed block environments to increase data storage
- IBM Spectrum Accelerate: Enterprise block storage for hybrid cloud
- IBM Spectrum Scale: High-performance, highly scalable hybrid cloud storage for unstructured data driving cognitive applications
- IBM Cloud Object Storage: Flexible, scalable, and simple object storage with geo dispersed enterprise availability and security for hybrid cloud workloads

Because IBM Spectrum Storage Suite includes IBM Spectrum Control, you can deploy IBM Storage Insight for IBM Spectrum Control.

> **Tip:** Alerts are a good way to be notified of conditions and potential problems that are detected on your storage. If you use IBM Spectrum Control and IBM Storage Insights for IBM Spectrum Control together to enhance your monitoring capabilities, it is recommended that you define alerts in one of the offerings, but not both.
>
> By defining all your alerts in one offering, you can avoid receiving duplicate or conflicting notifications when alert conditions are detected.

Implementation and set up of IBM Storage Insights

To use IBM Storage Insights with the IBM Spectrum Virtualize, you must sign up. For more information, see this web page.

Sign up process

Consider the following points about the sign-up process:

- For the sign-up process, you need an IBM ID. If you do not have such an ID, create your IBM account and complete the short form.
- When you register, specify an owner for IBM Storage Insights. The owner manages access for other users and acts as the main contact.
- You receive a Welcome email when IBM Storage Insights is ready. The email contains a direct link to your dashboard.

Figure 9-13 shows the IBM Storage Insight registration window.

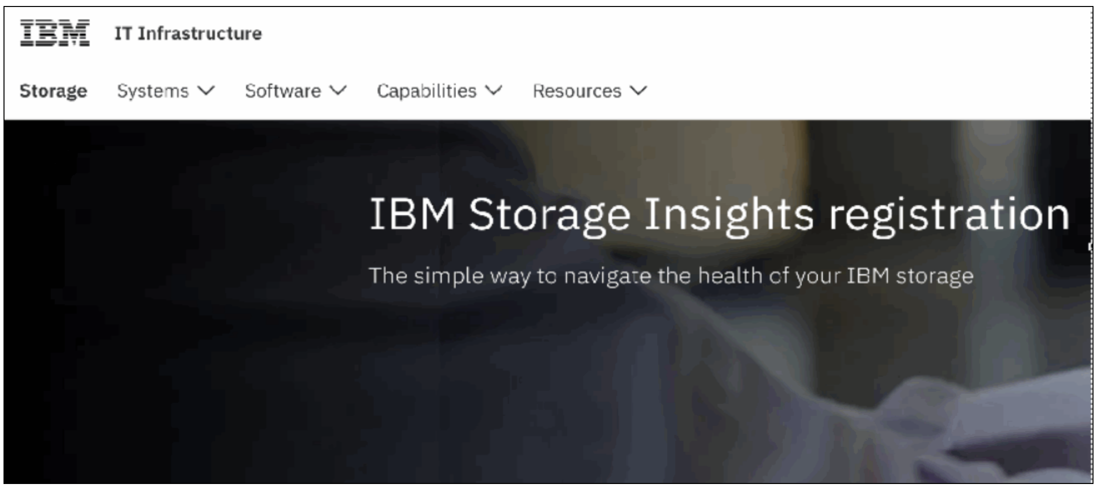


Figure 9-13 IBM Storage Insight registration window

Figure 9-14 shows the registration website when you scroll down. You can select whether you want to register for IBM Storage Insights or IBM Storage Insights for Spectrum Controls.

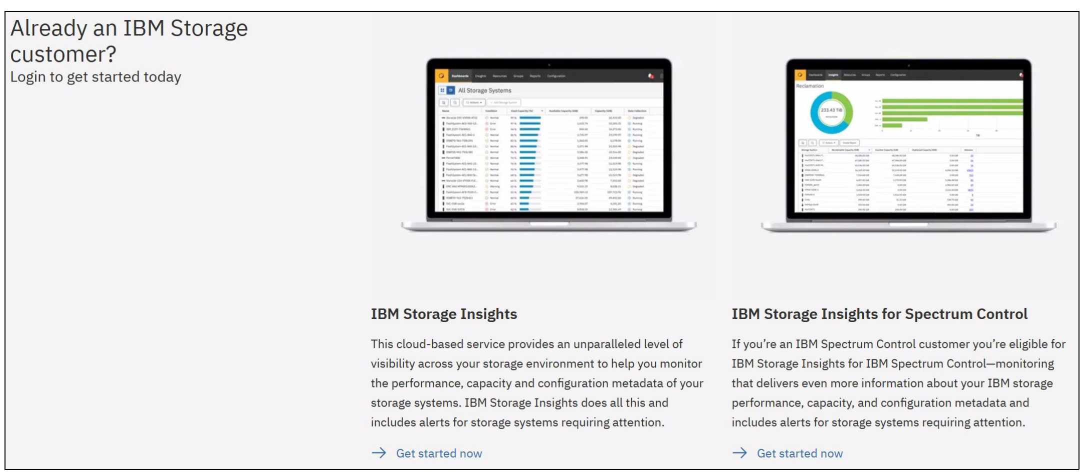


Figure 9-14 Choose IBM Storage Insights or IBM Storage Insights for Spectrum Control

Figure 9-15 shows the log in window in the registration process. If you have your credentials, enter your ID and click **Continue** to proceed. If you do not have an ID, click **Create an IBMid**.

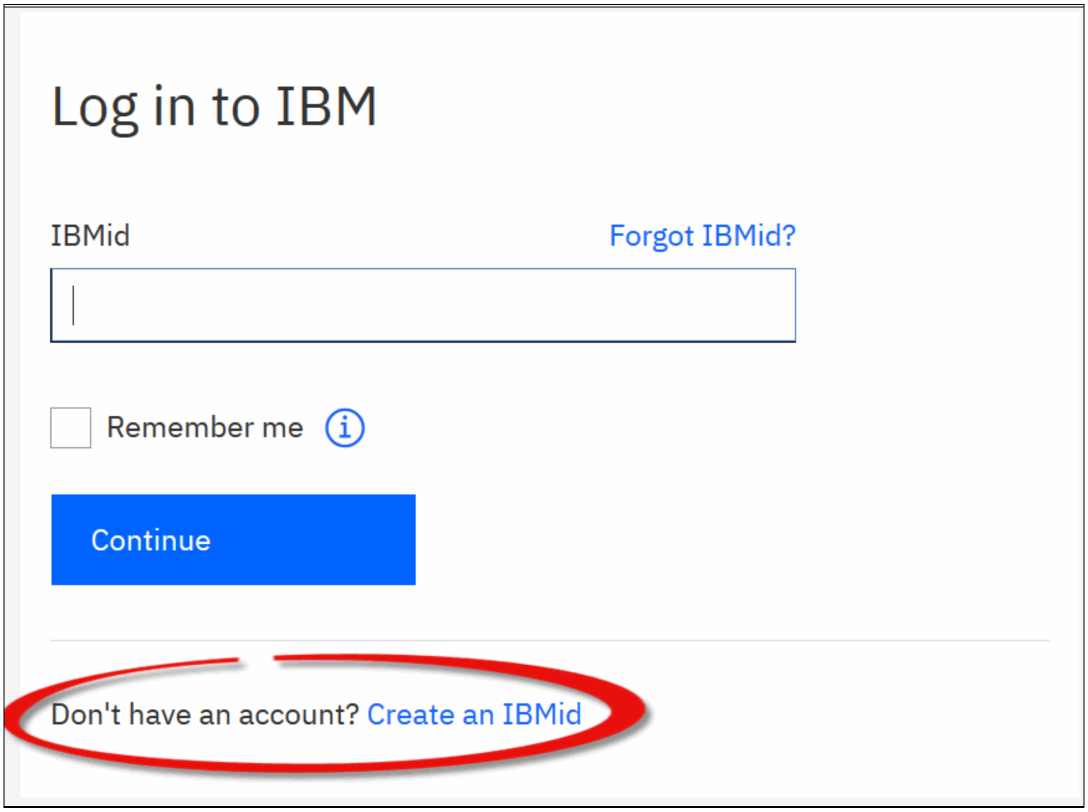


Figure 9-15 Registration login window

If you want to create an IBMid, provide the following information (see Figure 9-16 on page 407):

- Email
- First name
- Last name
- Country or region
- Password

Select the option if you want to receive Information from IBM to keep you informed of products, services, and offerings. You can withdraw your marketing consent at any time by sending an email to `netsupp@us.ibm.com`. Also, you can unsubscribe from receiving marketing emails by clicking the unsubscribe link in an email.

For more information, see IBM Privacy Statement.

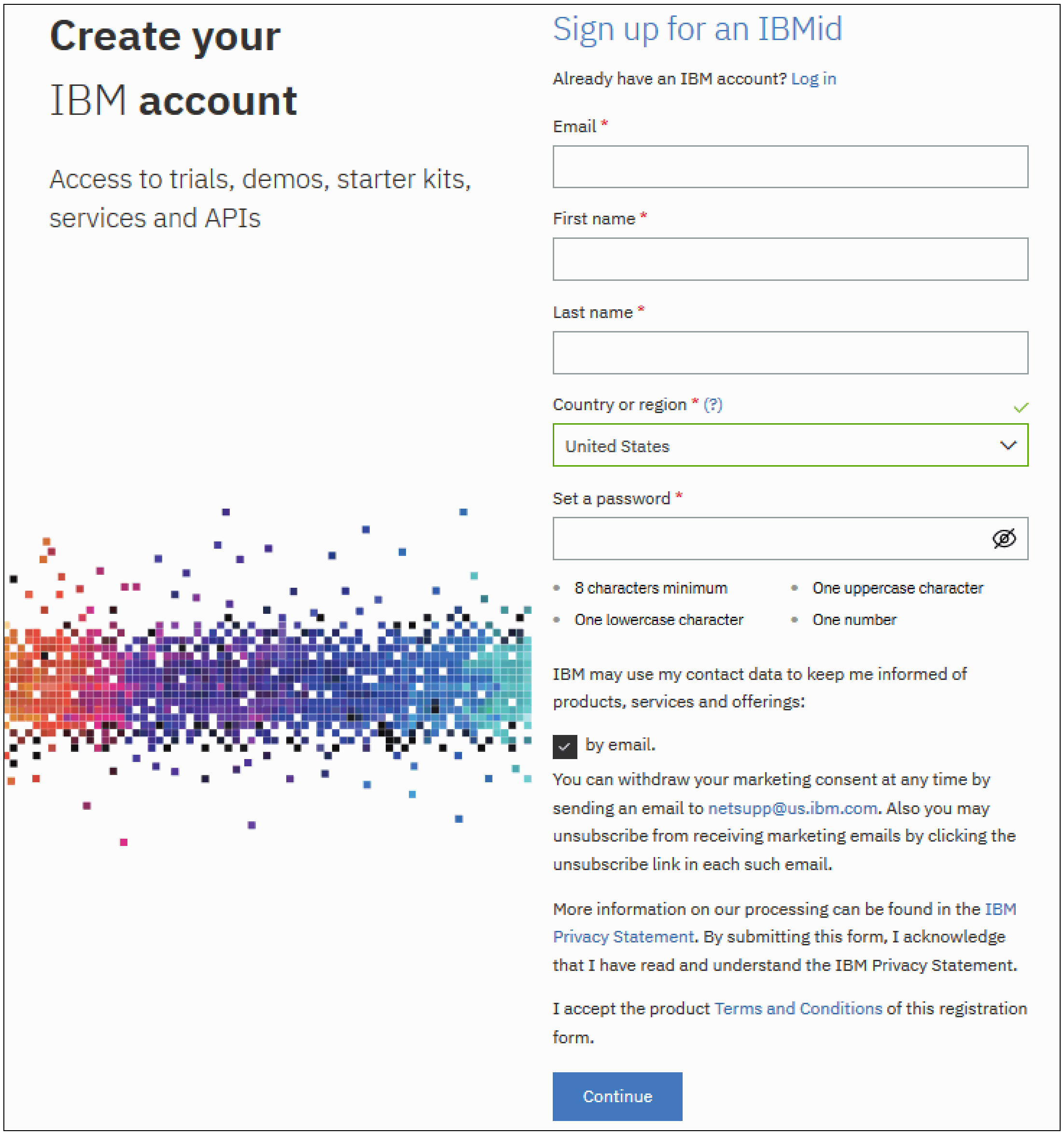


Figure 9-16 Creating an IBM account

In the next window, sign in with your IBM Account and password.

Complete the following information in the IBM Storage Insights registration form (see Figure 9-17 on page 408):

- Company name (must be unique).
- You might complete other identifying features, such as a location or department:
  - Owner details
  - Email address / ID
  - The person who registered for IBM Storage Insights
  - Access granted for storage trends, health of storage and access to support
  - First and last name

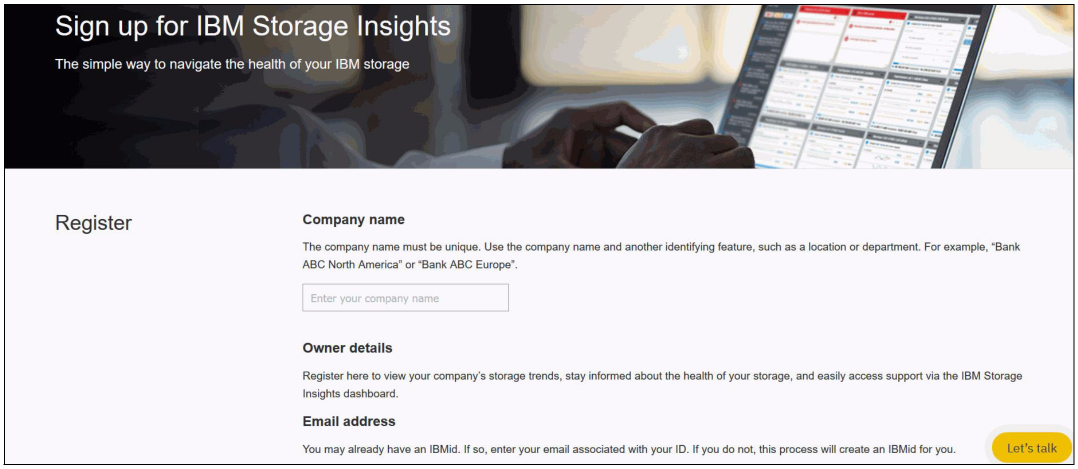


Figure 9-17 IBM Storage Insights registration form

After registration for Storage Insights is complete, download and install the data collector for your system. Extract the data collector, run the data collector installer script, and ensure that your server (or virtual machine) can access the `host_name:port` that is specific to your instance of Storage Insights. After the data collector is installed on the system, you can add your storage devices to a Storage Insights dashboard.

> **Note:** To connect to your instance of Storage Insights, you must configure your firewall to allow outbound communication on the default HTTPS port 443 using Transmission Control Protocol (TCP). User Datagram Protocol (UDP) is not supported.

Deploy a data collector

To deploy a lightweight data collector in your data center to stream performance, capacity, and configuration metadata to IBM Storage Insights:

1. Log in to IBM Storage Insights (the link is in your Welcome email).
2. From the Configuration > Data Collector page, download the data collector for your operating system (Windows, Linux, or AIX).
3. Extract the contents of the data collector file on the virtual machine or physical server where you want it to run.
4. For Windows, run `installDataCollectorService.bat`.

   For Linux or AIX, run `installDataCollectorService.sh`.

After the data collector is deployed, it attempts to establish a connection to IBM Storage Insights. When the connection is complete, you are ready to start adding your storage systems for monitoring.

> **Requirements:** The following requirements must be met:
>
> - 1 GB RAM
> - 1 GB disk space
> - Windows, AIX, or Linux (x86-64 systems only)

For more information, see this IBM Documentation web page.

Note: To avoid potential problems, ensure that the operating system on the server or virtual machine where you install the data collector has general or extended support for maintenance and security.

Storage system metadata is sent to IBM Storage Insights, such as the following examples:

- Information about the configuration of the storage system, such as name, firmware, and capacity.
- Information about the internal resources of the storage system, such as volumes, pools, nodes, ports, and disks. This includes the names and the configuration and capacity information for each internal resource.
- Information about the performance of storage system resources and internal resources such as pools and volumes.

For more information on how the metadata is collected and used, see the following resources:

- IBM Storage Insights Fact Sheet
- IBM Storage Insights Security Guide

Adding a storage system

Complete the following steps to connect IBM Storage Insights to the storage systems that you want to monitor:

1. On the Operations dashboard in IBM Storage Insights, look for the button to add storage systems.
2. Click **Add Storage Systems** and follow the prompts. You can add one or more storage systems at a time.

For more information, see this IBM Documentation web page.

Dashboard

The operations dashboard provides a full view of your storage inventory and metadata. It also includes a diagnostic feed that tells you which storage systems require attention.

The dashboard includes the following key elements:

- Storage systems that are being monitored.
- A dynamic diagnostic feed that tells you which storage systems require attention.
- Key capacity metrics so you know whether you have enough capacity to meet your storage demands.
- Key performance metrics so that you know whether the performance of your storage systems meets operational requirements.

For more information, see this IBM Documentation web page.

Enable Call Home

Get the most out of IBM Storage Insights by enabling Call Home on your IBM block storage systems. With Call Home, your dashboard includes a diagnostic feed of events and notifications about their health and status.

Stay informed so you can act quickly to resolve incidents before they affect critical storage operations.

For more information, see this IBM Documentation web page.

Adding users to your dashboard

Users, such as other storage administrators, IBM Technical Advisors, and IBM Business Partners, can be added at any time so that they can access your IBM Storage Insights dashboard. Complete the following steps:

1. In IBM Storage Insights, click your username in the upper-right of the dashboard.
2. Click **Manage Users**.
3. On your MYIBM page, ensure that **IBM Storage Insights** is selected.
4. Click **Add new user**.

For more information, see this IBM Documentation web page.

9.3 Capacity metrics for block storage systems

Effective and exact capacity management is based on fundamental knowledge of capacity metrics in the IBM SAN Volume Controller system. Data reduction pools (DRPs), thin provisioning, compression, and deduplication add many metrics to the IBM SAN Volume Controller management GUI, IBM Spectrum Control, and IBM Storage Insights.

This section describes capacity monitoring by using:

- Management GUI
- IBM Spectrum Control
- IBM Storage Insights

Figure 9-18 shows how to interpret the capacity and savings in a storage environment.

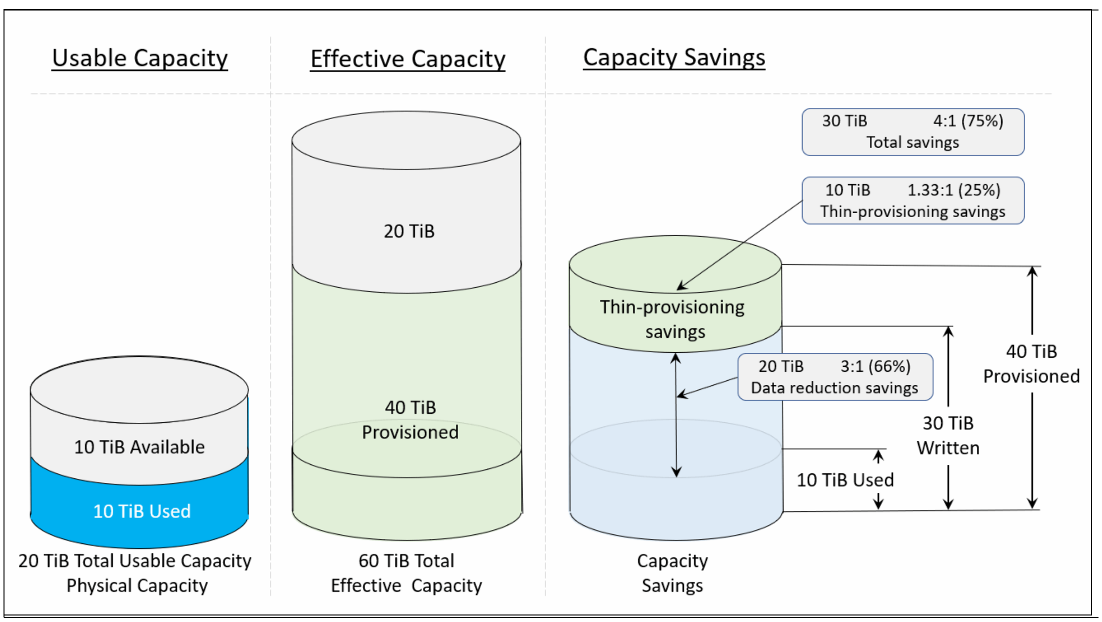


Figure 9-18 Understanding capacity information

9.3.1 Capacity monitoring by using the management GUI

The Capacity section of the Dashboard provides an overall view of system capacity. This section displays usable capacity, provisioned capacity, and capacity savings.

Usable capacity

Usable capacity (see Figure 9-19) indicates the total capacity in all storage on the system and includes all of the storage the system that can be virtualized and assigned to pools. Usable capacity is displayed in a bar graph and is divided into three categories: Stored Capacity, Available Capacity, and Total.

Figure 9-19 Usable Capacity display

> **Note:** If a DRP on IBM SAN Volume Controller level is using back-end storage that also compresses data (FCM drives or storage that uses FCMs) it is important to note that data sent to the back end from DRP is already compressed and the back end will not be able to compress it further. This makes it important to not allocate more than the total capacity of the back-end device to a DRP.

Stored capacity

Stored capacity (see Figure 9-19) indicates the amount of capacity that is used on the system after capacity savings. The system calculates the stored capacity by subtracting the available capacity and any reclaimable capacity from the total capacity that is allocated to MDisks. To calculate the percentage, the stored capacity is divided by the total capacity that is allocated to MDisks. On the left side of the bar graph, the stored capacity is displayed in the total capacity and as a percentage.

Available capacity

The total A*vailable capacity* (see Figure 9-19) displays on the right side of the bar graph. Available capacity is calculated by adding the available capacity and the total reclaimable capacity. To calculate the percentage of available capacity on the system, the available capacity is divided by the total amount of capacity that is allocated to MDisks.

Total capacity

The *Total capacity* (see Figure 9-19) displays on the right under the bar graph and shows all the capacity available on the system. The bar graph is a visual representation of capacity usage and availability and can be used to determine whether storage must be added to the system. Select MDisks to view more information about the usable capacity of the system on the MDisks by Pools page. You can also select Compressed Volumes, Deduplicated Volumes, or Thin-Provisioned Volumes.

If you use the command-line interface to determine usable capacity on your system, several parameter values are used from the **lssystem** command to calculate stored, available, and total capacities.

Stored capacity is calculated with the values in the `total_mdisk_capacity`, `total_free_space`, `total_reclaimable_capacity` by using the following formula:

```
Total stored capacity =
total_mdisk_capacity - total_free_space - total_reclaimable_capacity
```

To calculate the available capacity, use the values in `total_free_space` and `total_reclaimable_capacity`, as shown in the following formula:

```
Total available capacity = total_free_space + total_reclaimable_capacity
```

An example of determining the usable capacity is shown in Example 9-3.

Example 9-3 Determining usable capacity

```
IBM_FlashSystem:FS9200:superuser>lssystem |grep total_mdisk
total_mdisk_capacity 5.3TB
IBM_FlashSystem:FS9200:superuser>lssystem |grep total_free
total_free_space 5.3TB
IBM_FlashSystem:FS9200:superuser>lssystem |grep total_reclaim
total_reclaimable_capacity 0.00MB
IBM_FlashSystem:FS9200:superuser>
```

Provisioned capacity

Provisioned capacity (see Figure 9-20) is the total capacity of all virtualized storage on the system. Provisioned capacity is displayed in a bar graph and is divided into two categories: Written Capacity and Available Capacity.

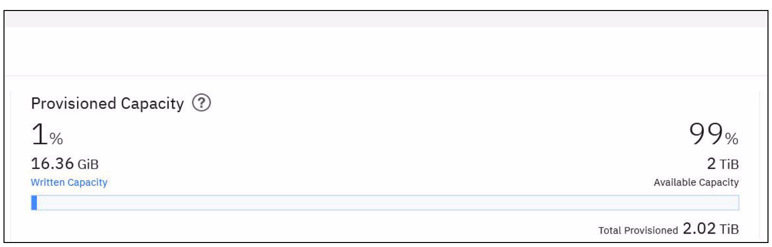


Figure 9-20 Provisioned Capacity

Written Capacity (see Figure 9-20) displays on the left side of the bar graph and indicates the amount of capacity that has data that is written to all the configured volumes on the system. The system calculates the written capacity for volumes by adding the stored capacity to capacity savings. The percentage of written capacity for volumes is calculated by dividing the written capacity by the total provisioned capacity for volumes on the system.

The *Available Capacity* (see Figure 9-20) displays on the right side of the bar graph and indicates the capacity on all configured volumes that is available to write new data. The available capacity is calculated by subtracting the written capacity for volumes from the total amount of capacity that is provisioned for volumes.

The percentage of available capacity is calculated by dividing the available capacity for volumes by the total amount of capacity that is provisioned to volumes on the system.

The *Total Provisioned capacity* displays under the Available Capacity and indicates the total amount of capacity that is allocated to volumes. The Provisioned Capacity also displays the percentage for over-provisioned volumes. The Overprovisioned value indicates the percentage of provisioned capacity that is increased because of capacity savings.

Capacity Savings (see Figure 9-21 and Example 9-4) indicates the amount of capacity that is saved on the system by using compression, deduplication, and thin-provisioning. The percentage value for each of these capacity savings methods compares the stored capacity before capacity savings are applied to the stored capacity after capacity savings is applied.

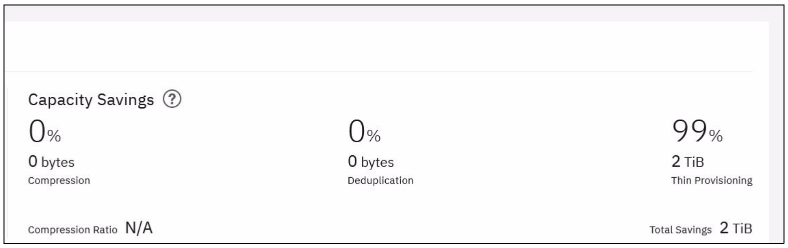


Figure 9-21 Capacity Savings

Compression shows the total capacity savings gained from the use of compression on the system. Deduplication indicates the total capacity savings that the system is saved from all deduplicated volumes.

Thin-Provisioning displays the total capacity savings for all thin-provisioned volumes on the system. You can view all of the volumes that use each of these technologies. Different system models can have more requirements to use compression or deduplication. Verify all system requirements before these functions are used.

Example 9-4 Capacity savings

```
IBM_FlashSystem:FS9200:superuser>lssystem |grep deduplication
deduplication_capacity_saving 0.00MB
IBM_FlashSystem:FS9200:superuser>lssystem |grep compression
compression_active no
compression_virtual_capacity 0.00MB
compression_compressed_capacity 0.00MB
compression_uncompressed_capacity 0.00MB
compression_destage_mode off
IBM_FlashSystem:FS9200:superuser>lssystem |grep reduction
used_capacity_before_reduction 0.00MB
used_capacity_after_reduction 0.00MB
```

9.4 Capacity monitoring by using IBM Spectrum Control or IBM Storage Insights

The Capacity section of IBM Spectrum Control and IBM Storage Insights provides an overall view of system capacity. This section displays usable capacity, provisioned capacity, and capacity savings.

The Capacity chart (see Figure 9-22) of IBM Spectrum Control at the top of the Overview page (select **Spectrum Control GUI** → **Storage** → **Block Storage Systems** and then, double-click the device) shows how much capacity is used and how much capacity is available for storing data.

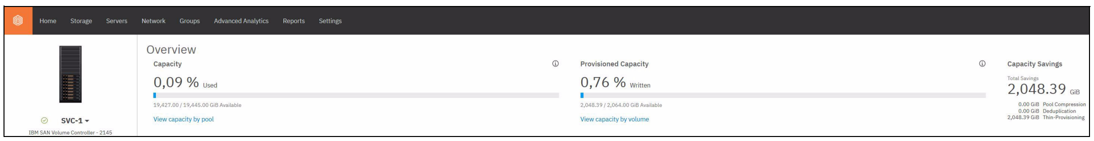


Figure 9-22 Spectrum Control overview page

In IBM Storage Insights, the Capacity chart shows the capacity usage (see Figure 9-23) on the Dashboards page (select **Storage Insights GUI** → **Dashboards** → and then, click the device).

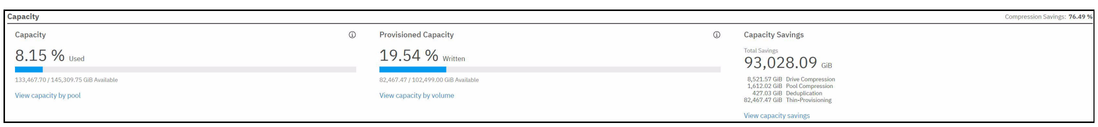


Figure 9-23 Storage Insights overview page

The Provisioned Capacity chart shows the written capacity values in relation to the total provisioned capacity values before data reduction techniques are applied. The following values are shown:

- The capacity of the data that is written to the volumes as a percentage of the total provisioned capacity of the volumes.
- The amount of capacity that is still available for writing data to the thin-provisioned volumes in relation to the total provisioned capacity of the volumes. Available capacity is the difference between the provisioned capacity and the written capacity, which is the thin-provisioning savings.
- A breakdown of the total capacity savings that are achieved when the written capacity is stored on the thin-provisioned volumes.

In the capacity overview chart, a horizontal bar is shown when a capacity limit is set for the storage system. Mouse over the chart to determine the capacity limit and how much capacity is left before the limit is reached.

For a breakdown of the capacity usage by pool or volume, click the links on the page.

Capacity view and their metrics

In this section, we describe the metrics of the Capacity View of IBM Spectrum Control and IBM Storage Insights for block storage systems.

To open the Capacity View in IBM Spectrum Control, you can start from the Storage menu and click **Block Storage Systems**. Then, right-click one or more storage systems and click **View Capacity** (Figure 9-24).

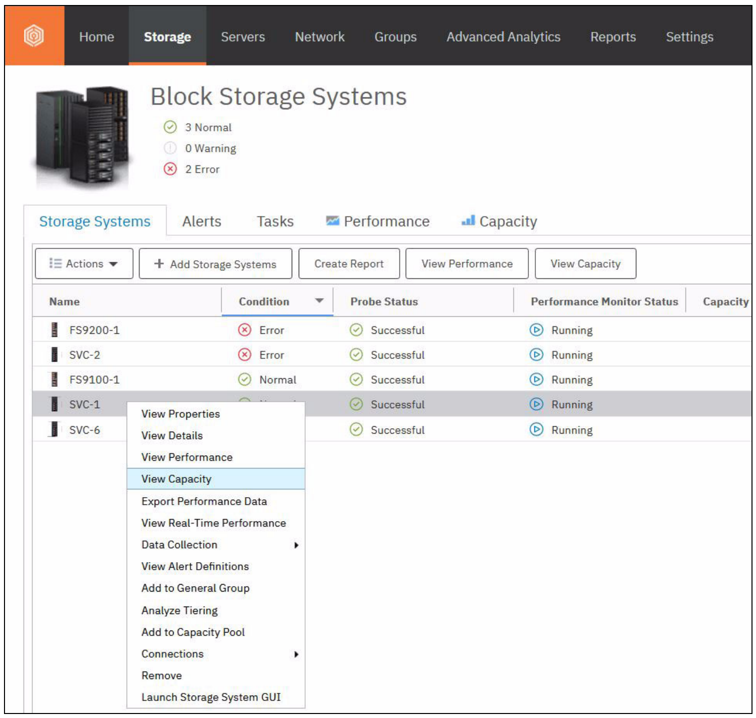


Figure 9-24 IBM Spectrum Control and IBM Storage Insights Block Storage Systems overview

You also can click **Actions** → **View Capacity** (see Figure 9-25) for each device.

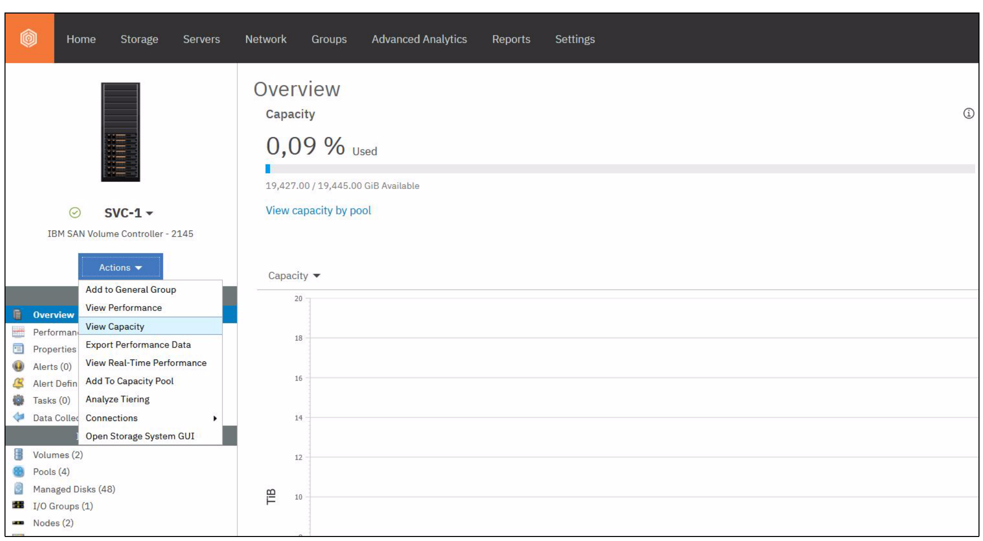


Figure 9-25 Overview Storage System

To open the Capacity View in IBM Storage Insights, click **Resources** → **Block Storage Systems**. Then, right-click one or more storage systems and click **View Capacity** (see Figure).

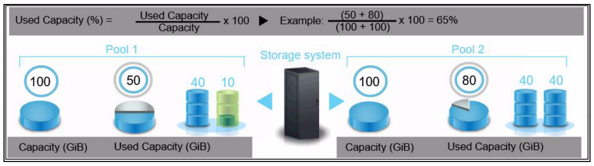


Figure 9-26 Used Capacity

> **Note:** Used Capacity (%) was previously known as Physical Allocation.

Storage system, Pool capacity, and Volume capacity metrics

Used Capacity (%) shows the percentage of physical capacity in the pools that is used by the standard-provisioned volumes, thin-provisioned volumes, and volumes in child pools. Check the value for used capacity percentage to see the following information:

- Whether the physical capacity of the pools is fully allocated; that is, the value for used capacity is 100%.
- Whether you have sufficient capacity to:
  - Provision new volumes with storage
  - Allocate to the compressed and thin-provisioned volumes in the pools

The following formula (see the example that is shown in Figure) is used to calculate Used Capacity (%):

```
[(Used Capacity ÷ Capacity)*100]
```

`Used Capacity (GiB)` shows the amount of space that is used by the standard- and thin-provisioned volumes in the pools. If the pool is a parent pool, the amount of space that is used by the volumes in the child pools is also calculated.

The capacity that is used by for thin-provisioned volumes is less than their provisioned capacity, which is shown in the Provisioned Capacity (GiB) column. If a pool does not have thin-provisioned volumes, the value for used capacity is the same as the value for provisioned capacity.

> **Note:** Used Capacity (GiB) was previously known as Allocated Space.

Adjusted Used Capacity (%) shows the amount of capacity that can be used without exceeding the capacity limit.

The following formula is used to calculate Adjusted Used Capacity (%):

```
[(Used Capacity in GiB ÷ Capacity Limit in GiB)*100]
```

For example, if the capacity is 100 GiB, the used capacity is 40 GiB, and the capacity limit is 80% or 80 GiB, the value for Adjusted Used Capacity (%) is (40 GiB/80 GiB)* 100 or 50%.

In this example, you can use 30% or 40 GiB of the usable capacity of the resource before you reach the capacity limit (see Figure 9-27).

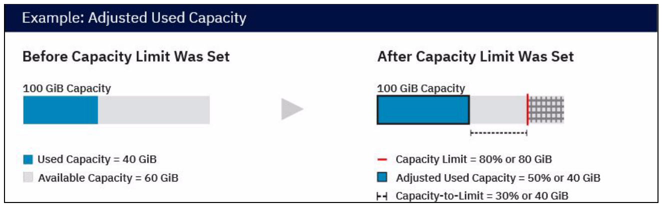


Figure 9-27 Adjusted Used Capacity example

If the used capacity exceeds the capacity limit, the value for Adjusted Used Capacity (%) is over 100%.

To add the Adjusted Used Capacity (%) column, right-click any column heading on the Block Storage Systems page.

Available Capacity (GiB) shows the total amount of the space in the pools that is not used by the volumes in the pools. To calculate available capacity, the following formula is used:

```
[pool capacity - used capacity]
```

Note: Available Capacity was previously known as Available Pool Space.

Available Volume Capacity (GiB) shows the total amount of remaining space that can be used by the volumes in the pools. The following formula is used to calculate this value:

```
[provisioned capacity · used capacity]
```

The capacity that is used by thin-provisioned volumes is typically less than their provisioned capacity. Therefore, the available capacity represents the difference between the provisioned capacity and the used capacity for all the volumes in the pools. For Hitachi VSP non-thin provisioned pool capacity, the available capacity is always zero.

Note: Available Volume Capacity (GiB) is previously known as Effective Unallocated Volume Space.

Capacity (GiB) shows the total amount of storage space in the pools. For XIV systems and IBM Spectrum Accelerate, capacity represents the physical (“hard”) capacity of the pool, not the provisioned (“soft”) capacity. Pools that are allocated from other pools are not included in the total pool space.

Note: Capacity is previously known as Pool Capacity.

Capacity Limit (%) and *Capacity Limit (GiB)* can be set on the capacity that is used by your storage systems. For example, the policy of your company is to keep 20% of the usable capacity of your storage systems in reserve. Therefore, you log in to the GUI as Administrator and set the capacity limit to 80% (see Figure 9-28).

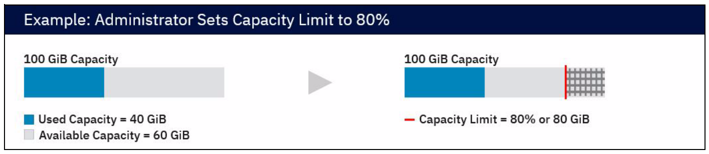


Figure 9-28 Capacity limit example

Capacity-to-Limit (GiB) shows the amount of capacity that is available before the capacity limit is reached.

The formula for calculating Capacity-to-Limit (GiB) is:

```
Capacity Limit in GiB - Used Capacity in GiB)
```

For example, if the capacity limit is 80% or 80 GiB and the used capacity is 40 GiB, the value for Capacity-to-Limit (GiB) is (80 GiB - 40 GiB or 80% - 50%) which is 30% or 40 GiB (see Figure 9-29).

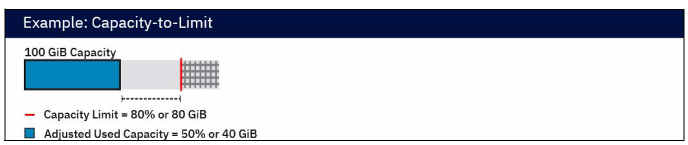


Figure 9-29 Capacity-to-Limit

> **Note:** This metric is not available for all storage systems, such as FlashSystem A9000, FlashSystem A9000R, and Dell EMC VMAX.

Compression Savings (%) are the estimated amount and percentage of capacity that is saved by using data compression across all pools on the storage system. The percentage is calculated across all compressed volumes in the pools and does not include the capacity of noncompressed volumes.

For storage systems with drives that use inline data compression technology, the Compression Savings does not include the capacity savings that are achieved at the drive level.

The following formula is used to calculate the amount of storage space that is saved:

```
[written capacity · compressed size]
```

The following formula is used to calculate the percentage of capacity that is saved:

```
[(written capacity · compressed size) ÷ written capacity] × 100
```

For example, the written capacity, which is the amount of data that is written to the volumes before compression, is 40 GiB. The compressed size, which reflects the size of compressed data that is written to disk, is only 10 GiB. Therefore, the compression savings percentage across all compressed volumes is 75%.

Note: Compression Savings (%) metric is available for FlashSystem A9000 and FlashSystem A9000R, IBM Spectrum Accelerate, XIV storage systems with firmware version 11.6 or later, and resources that run IBM Spectrum Virtualize.

For FlashSystem A9000 and FlashSystem A9000R, all volumes in the pools are compressed.

The exception is for compressed volumes that are also deduplicated on storage systems that run IBM Spectrum Virtualize. This column is blank.

Deduplication Savings (%) shows the estimated amount and percentage of capacity that is saved by using data deduplication across all DRPs on the storage system. The percentage is calculated across all deduplicated volumes in the pools and does not include the capacity of volumes that are not deduplicated.

The following formula is used to calculate the amount of storage space that is saved:

```
[written capacity · deduplicated size]
```

The following formula is used to calculate the percentage of capacity that is saved:

```
[(written capacity · deduplicated size) ÷ written capacity] × 100
```

For example, the written capacity, which is the amount of data that is written to the volumes before deduplication, is 40 GiB. The deduplicated size, which reflects the size of deduplicated data that is written to disk, is just 10 GB. Therefore, data deduplication reduced the size of the data that is written by 75%.

Note: Deduplication Savings (%) metric is available for FlashSystem A9000, FlashSystem A9000R, and resources that run IBM Spectrum Virtualize version 8.1.3 or later.

Drive Compression Savings (%) shows amount and percentage of capacity that is saved with drives that use inline data compression technology. The percentage is calculated across all compressed drives in the pools.

The amount of storage space that is saved is the sum of drive compression savings.

The following formula is used to calculate the percentage of capacity that is saved:

```
[(used written capacity · compressed size) ÷ used written capacity] × 100
```

Note: Drive Compression Savings (%) metric is available for Storage systems that contain IBM FlashCore Modules with hardware compression.

Mapped Capacity (GiB) shows the total volume space in the storage system that is mapped or assigned to host systems, including child pool capacity.

Note: Mapped Capacity (GiB) is previously known as Assigned Volume Space.

Overprovisioned Capacity (GiB) shows the capacity that cannot be used by volumes because the physical capacity of the pools cannot meet the demands for provisioned capacity. The following formula is used to calculate this value:

```
[Provisioned Capacity · Capacity]
```

> **Note:** Overprovisioned Capacity (GiB) is previously known as Unallocatable Volume Space.

Shortfall (%) shows the difference between the remaining unused volume capacity and the available capacity of the associated pool, expressed as a percentage of the remaining unused volume capacity. The shortfall represents the relative risk of running out of space for overallocated thin-provisioned volumes.

If the pool has sufficient available capacity to satisfy the remaining unused volume capacity, no shortfall exists. As the remaining unused volume capacity grows, or as the available pool capacity decreases, the shortfall increases and the risk of running out of space becomes higher. If the available capacity of the pool is exhausted, the shortfall is 100% and any volumes that are not yet fully allocated run out of space.

If the pool is not thin-provisioned, the shortfall percentage equals zero. If shortfall percentage isn't calculated for the storage system, the field is left blank.

The following formula is used to calculate this value:

```
[Overprovisioned Capacity ÷ Committed but Unused Capacity]
```

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 provisioned capacity, the pool does not have enough space to fulfill the commitment to provisioned capacity. This value represents the percentage of the committed provisioned capacity 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.

Next, we consider Shortfall (%) based on example values. The remaining physical capacity of a pool is 70 GiB, but 150 GiB of provisioned capacity was committed to thin-provisioned volumes. If the volumes use 50 GiB, 100 GiB is still committed to the volumes (150 GiB • 50 GiB) with a shortfall of 30 GiB (70 GiB remaining pool space • 100 GiB remaining commitment of volume space to the volumes). Because the volumes are overcommitted by 30 GiB based on the available capacity in the pool, the shortfall is 30% when the following calculation is used:

```
[(100 GiB unused volume capacity · 70 GiB remaining pool capacity) ÷ 100 GiB
unused volume capacity] × 100
```

For more information, see this video.

> **Note:** Shortfall (%) is available for DS8000, Hitachi Virtual Storage Platform, and storage systems that run IBM Spectrum Virtualize.
>
> This value is not available for FlashSystem A9000 and FlashSystem A9000R.

Provisioned Capacity (%) shows the percentage of the physical capacity that is committed to the provisioned capacity of the volumes in the pools. If the value exceeds 100%, the physical capacity does not meet the demands for provisioned capacity. To calculate provisioned capacity percentage, the following formula is used:

```
[(provisioned capacity ÷ pool capacity) × 100]
```

For example, if the provisioned capacity percentage is 200% for a storage pool with a physical capacity of 15 GiB, the provisioned capacity that is committed to the volumes in the pools is 30 GiB. Twice as much space is committed to the pools than is physically available to the pools.

If the provisioned capacity percentage is 100% and the physical capacity is 15 GiB, the provisioned capacity that is committed to the pools is 15 GiB. The total physical capacity that is available to the pools is used by the volumes in the pools.

A provisioned capacity percentage that is higher than 100% is considered to be aggressive because insufficient physical capacity is available to the pools to satisfy the allocation of the committed space to the compressed and thin-provisioned volumes in the pools. In such cases, you can check the Shortfall (%) value to determine how critical the shortage of space is for the storage system pools.

Note: Provisioned Capacity (%) is previously known as Virtual Allocation.

Provisioned Capacity (GiB) shows the total amount of provisioned capacity of volumes within the pool. If the pool is a parent pool, it also includes the storage space that can be made available to the volumes in the child pools.

Note: Provisioned Capacity (GiB) is previously known as Total Volume Capacity.

Safeguarded Capacity (GiB) shows the total amount of capacity that is used to store volume backups that are created by the Safeguarded Copy feature in DS8000.

Total Capacity Savings (%) shows the estimated amount and percentage of capacity that is saved by using data deduplication, pool compression, thin provisioning, and drive compression, across all volumes in the pool.

The following formula is used to calculate the amount of storage space that is saved:

```
[Provisioned Capacity · Used Capacity]
```

The following formula is used to calculate the percentage of capacity that is saved:

```
[(Provisioned Capacity · Used Capacity) ÷ Provisioned Capacity] × 100
```

Note: Total Capacity Savings (%) is previously known as Total Data Reduction Savings and is available for: FlashSystem A9000 and FlashSystem A9000R, IBM Spectrum Accelerate, XIV storage systems with firmware version 11.6 or later, and resources that run IBM Spectrum Virtualize.

Unmapped Capacity (GiB) shows the total amount of space in the volumes that are not assigned to hosts.

Note: Unmapped Capacity (GiB) is previously known as Unassigned Volume Space.

In the Zero Capacity column (see Figure 9-30) on the Pools page, you can see the date, which is based on the storage usage trends for the pool, for when the pool runs out of available capacity.

Zero Capacity: The capacity information that is collected over 180 days is analyzed to determine, based on historical storage consumption, when the pools are to run out of capacity. The pools ran out of capacity are marked as `Depleted`. For the other pools, a date is provided so that you know when the pools are projected to run out of capacity. If sufficient information is not collected to analyze the storage usage of the pool, `None` is shown as the value for zero capacity.

If a capacity limit is set for the pool, the date that is shown in the Zero Capacity column is the date when the available capacity based on the capacity limit is to be depleted. For example, if the capacity limit for a 100 GiB pool is 80%, it is the date when the available capacity of the pool is less than 20 GiB. `Depleted` is shown in the column when the capacity limit is reached.

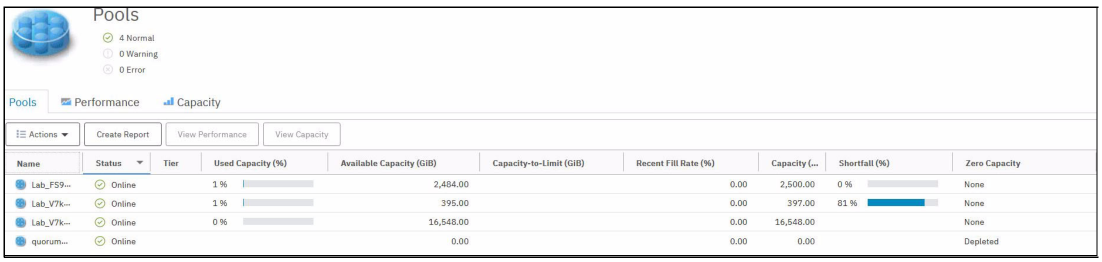


Figure 9-30 Zero Capacity

The following values can be shown in the Zero Capacity column:

- A date

  The data that is based on space usage trends for the pool when the capacity runs out (projected).

- None

  Based on the current trend, no date can be calculated for when the pool is to be filled (for example, if the trend is negative) as data is moved out of the pool.

- Depleted

  The pool is full.

The metrics that are described next can be added to capacity charts for storage systems within capacity planning. Use the charts to detect capacity shortages and space usage trends.

Available Repository Capacity (GiB) shows the available, unallocated storage space in the repository for Track Space-Efficient (TSE) thin-provisioning.

Note: Available for: DS8000 thin-provisioned pools.

Soft Capacity (GiB) shows the amount of virtual storage space that is configured for the pool.

Note: Soft Capacity (GiB) is available for XIV systems and IBM Spectrum Accelerate storage systems.

Available Soft Capacity (GiB) shows the amount of virtual storage space that is available to allocate to volumes in a storage pool.

Note: Available for: XIV systems, and IBM Spectrum Accelerate storage systems.

Written Capacity (GiB) shows the amount of data that is written from the assigned hosts to the volume before compression or data deduplication are used to reduce the size of the data. For example, the written capacity for a volume is 40 GiB. After compression, the volume used space, which reflects the size of compressed data that is written to disk, is only 10 GiB.

Note: Written Capacity (GiB) is previously known as Written Space.

Available Written Capacity (GiB) shows the amount of capacity that can be written to the pools before inline compression is applied. If the pools are not compressed, this value is the same as Available Capacity.

Note: Available Written Capacity (GiB) is previously known as Effective Used Capacity.

Because data compression is efficient, a pool can run out of Available Written Capacity while physical capacity is still available. To stay aware of your capacity needs, monitor this value and Available Capacity.

Enterprise HDD Available Capacity (GiB) shows the amount of storage space that is available on the Enterprise hard disk drives that can be used by Easy Tier for retiering the volume extents in the pool.

Note: Enterprise HDD Available Capacity (GiB) is available for DS8000 and storage systems that run IBM Spectrum Virtualize.

Enterprise HDD Capacity (GiB) shows the total amount of storage space on the Enterprise hard disk drives that can be used by Easy Tier for retiering the volume extents in the pool.

Note: Enterprise HDD Capacity (GiB) is available for DS8000 and storage systems that run IBM Spectrum Virtualize.

Nearline HDD Available Capacity (GiB) shows the amount of storage space that is available on the Nearline hard disk drives that can be used by Easy Tier for retiering the volume extents in the pool.

Note: Nearline HDD Available Capacity (GiB) is available for DS8000 and storage systems that run IBM Spectrum Virtualize.

Nearline HDD Capacity (GiB) shows the total amount of storage space on the Nearline hard disk drives that can be used by Easy Tier for retiering the volume extents in the pool.

Note: Nearline HDD Available Capacity (GiB) is available for DS8000 and storage systems that run IBM Spectrum Virtualize.

Repository Capacity (GiB) shows the total storage capacity of the repository for Track Space-Efficient (TSE) thin-provisioning.

Note: Repository Capacity (GiB) is available for DS8000 thin-provisioned pools.

Reserved Volume Capacity shows the amount of pool capacity that is reserved but has not been used yet to store data on the thin-provisioned volume.

Note: Reserved Volume Capacity is previously known as Unused Space and is available for resources that run IBM Spectrum Virtualize.

SCM Available Capacity (GiB) shows the available capacity on Storage Class Memory (SCM) drives in the pool. Easy Tier can use these drives to retier the volume extents in the pool.

Note: SCM Available Capacity (GiB) is available for IBM Spectrum Virtualize systems, such as FlashSystem 9200, FlashSystem 7200, and Storwize family storage systems that are configured with block storage.

SCM Capacity (GiB) shows the total capacity on Storage Class Memory (SCM) drives in the pool. Easy Tier can use these drives to retier the volume extents in the pool.

Note: SCM Capacity (GiB) is available for IBM Spectrum Virtualize systems, such as FlashSystem 9200, FlashSystem 7200, and Storwize family storage systems that are configured with block storage.

Tier 0 Flash Available Capacity (GiB) shows the amount of storage space that is available on the Tier 0 flash solid-state drives that can be used by Easy Tier for retiering the volume extents in the pool.

Note: Tier 0 Flash Available Capacity (GiB) is available for DS8000 and storage systems that run IBM Spectrum Virtualize.

Tier 0 Flash Capacity (GiB) shows the total amount of storage space on the Tier 0 flash solid-state drives that can be used by Easy Tier for retiering the volume extents in the pool.

Note: Tier 0 Flash Capacity (GiB) is available for DS8000 and storage systems that run IBM Spectrum Virtualize.

Tier 1 Flash Available Capacity (GiB) shows the amount of storage space that is available on the Tier 1 flash, read-intensive solid-state drives that can be used by Easy Tier for retiering the volume extents in the pool.

Note: Tier 1 Flash Available Capacity (GiB) is available for DS8000 and storage systems that run IBM Spectrum Virtualize.

Tier 1 Flash Capacity (GiB) shows the total amount of storage space on the Tier 1 flash, read-intensive solid-state drives that can be used by Easy Tier for retiering the volume extents in the pool.

Note: Tier 1Flash Capacity (GiB) is available for DS8000 and storage systems that run IBM Spectrum Virtualize.

Tier 2 Flash Available Capacity (GiB) shows the available capacity on Tier 2 flash, high-capacity drives in the pool. Easy Tier can use these drives to retier the volume extents in the pool.

Note: Tier 2 Flash Available Capacity (GiB) is available for: DS8000 storage systems.

Tier 2 Flash Capacity (GiB) shows the total capacity on Tier 2 flash, high-capacity drives in the pool. Easy Tier can use these drives to retier the volume extents in the pool.

Note: Tier 2 Flash Capacity (GiB) is available for DS8000 storage systems.

9.5 Creating alerts for IBM Spectrum Control and IBM Storage Insights

In this section, we provide information about alerting with IBM Spectrum Control and IBM Storage Insights. The free version of IBM Storage Insights does *not* support alerting.

New data reduction technologies add intelligence and capacity savings to your environment. If you use data reduction on different layers, such as hardware compression in the IBM FlashSystem 9200 Flash Core Modules (if an FS9200 is virtualized by the IBM SAN Volume Controller) and in the DRPs, you must pay more attention in preventing insufficient space remaining in the back-end storage device.

First, it is important to distinguish between thin provisioning and over-allocation (over-provisioning). Thin provisioning is a method for optimizing the use of available storage. It relies on allocation of blocks of data on-demand versus the traditional method of allocating all of the blocks up front. This method eliminates almost all white space, which helps avoid the poor usage rates (often as low as 10%) that occur in the traditional storage allocation method. Traditionally, large pools of storage capacity are allocated to individual servers, but remain unused (not written to).

Over provisioning means that more space is being assigned and promised to the hosts in total. They can try to store more data on the storage subsystem as physical capacity is available. However, an out-of-space condition can result.

Remember: You must constantly monitor your environment to avoid over-provisioning situations that can be harmful to the environment and can cause data loss.

It is also important to keep at least 15% free space for Garbage Collection in the background.

Data reduction technologies return back some space. If the space that is used for the data can be reduced, the saved up space can be used for other data. However, depending on the type of data, deleting might not result in freeing up much space.

Consider the example in which if you have three identical or almost identical files on a file system that were deduplicated. This issue resulted in getting a good compression ratio (three files, but stored only once). If you now delete one file, you do not gain more space because the deduplicated data must stay on the storage (because two other versions refer to the data). Similar results can be seen when several FlashCopies of one source are used.

9.5.1 Alert examples

Table 9-3 lists alerts for IBM FlashSystem systems that are based on Array or Pool level.

Table 9-3 Event examples for IBM FlashSystem

System	Entity	Resource Type	Event
FlashSystem with FCM	Array	Usable capacity	Available Physical Space <= nn% (* Example shown in 9.5.2, “Alert to monitor back-end capacity: Available Physical Space (%)” on page 426)
	Pool	Efficient Capacity	Physical allocation >= nn%
FlashSystem other media	Pool	Usable Capacity	Physical allocation >= nn%

Other alerts are possible as well; percentage alerts often are best suited because the alert definition applies to all pool in a storage system.

9.5.2 Alert to monitor back-end capacity: Available Physical Space (%)

In this section, we show how to deploy IBM Spectrum Control or IBM Storage Insights to monitor storage capacity and set up thresholds to notify and prevent us from running out of space.

The following example shows how to create an alert to get status Information about the remaining physical space on an IBM FlashSystem.

First, assign a severity to an alert. Assigning a severity can help you quickly identify and address the critical conditions that are detected on resources. The severity that you assign depends on the guidelines and procedures within your organization. Default assignments are provided for each alert.

Table 9-4 lists the possible alert severities.

Table 9-4 Alert severities

Option	Description
Critical	Alert is critical and needs to be resolved. For example, alerts that notify you when the amount of available space on a file system falls below a specified threshold.
Warning	Alerts that are not critical but represent potential problems. For example, alerts that notify you when the status of a data collection job is not normal.
Informational	Alerts that might not require any action to resolve and are primarily for informational purposes. For example, alerts that are generated when a new pool is added to a storage system

In this example, we created three thresholds:

- Critical (15% Available Physical Capacity (%) left)
- Warning (20% Available Physical Capacity (%) left)
- Information (30% Available Physical Capacity (%) left)

Adjust the percentage levels to the required levels as needed. The process to extend storage might take some time (ordering, installation, provisioning, and so on).

The advantage of this way to set up an Alert Policy is that you can add various IBM FlashSystem to this customized alert.

Figure 9-31 shows how to start creating an Alert Policy in Spectrum Control.

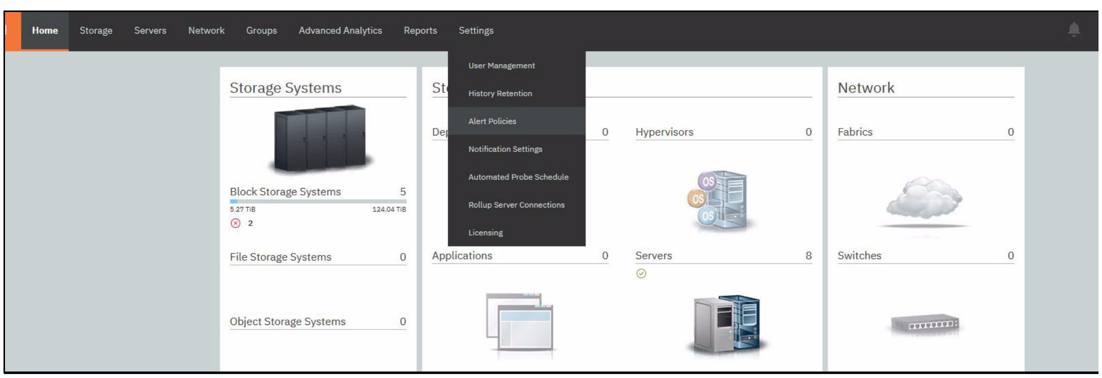


Figure 9-31 Spectrum Control Alert Policies

For Storage Insights, Figure 9-32 shows how to start creating an Alert Policy.

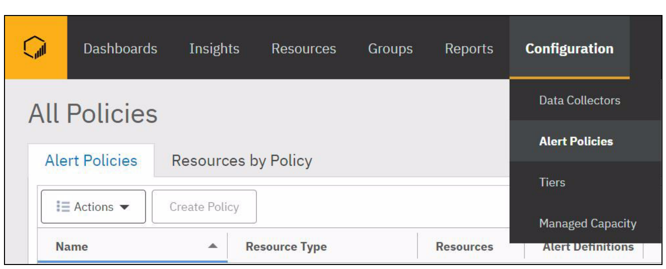


Figure 9-32 Storage Insights Alert Policies

The following example shows how to create an Alert Policy by copying the existing policy. You might need to change an existing Alert Policy (in our example the Default Policy) as well. Consider that a storage subsystem can be active in only one Alert Policy.

Note: No difference exists from IBM Spectrum Control to IBM Storage Insights for the steps that are presented next.

Figure 9-33 shows the Default Policy of IBM SAN Volume Controller in IBM Spectrum Control.

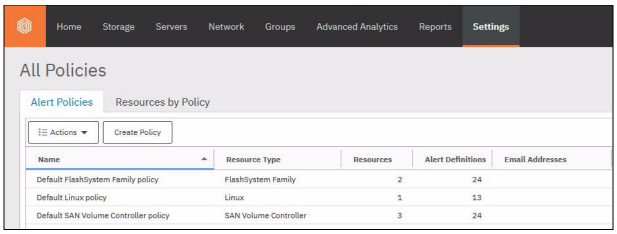


Figure 9-33 All Policies in Spectrum Control

Figure 9-34 shows how to copy a policy to create a policy. Mouse over the policy that you want to copy and then, left-click and choose **Copy Policy**.

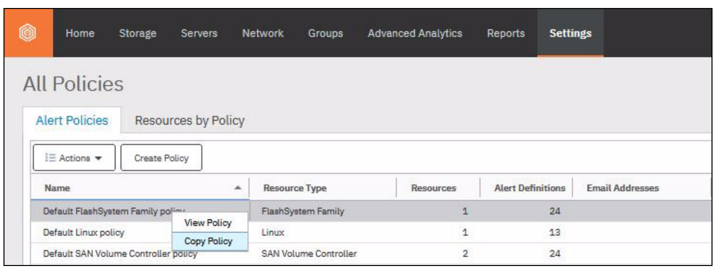


Figure 9-34 Copying a Policy in Spectrum Control

Figure 9-35 shows how to rename the previously copied policy. The new policy is stored as another policy. One IBM FlashSystem system can be added to a single policy only. You can add the system later on if you are not sure now (optional: Select **Resource** and select the option).

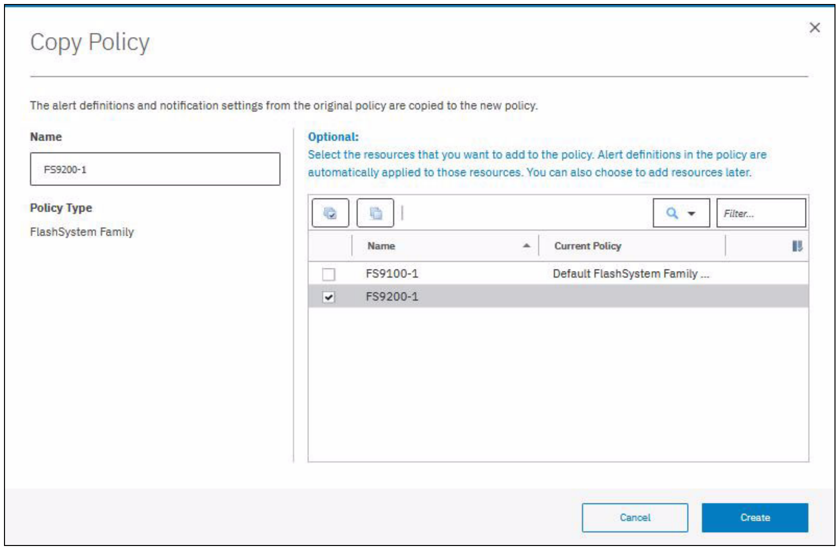


Figure 9-35 Copy Policy process

Figure 9-36 shows the newly created Alert Policy `FS9200-1` with all alerts that were inherited from the Default Policy.

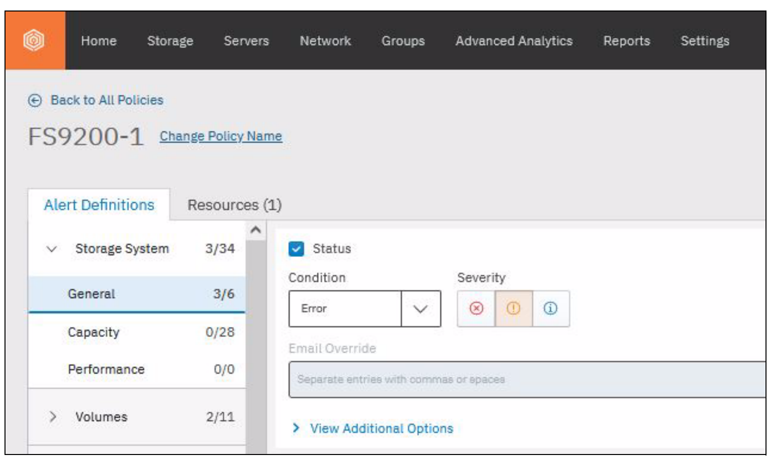


Figure 9-36 Spectrum Control - New Policy with inherited Alert Definitions

Figure 9-37 shows how to choose the required Alert Definitions by clicking **Pool** → **Capacity**.

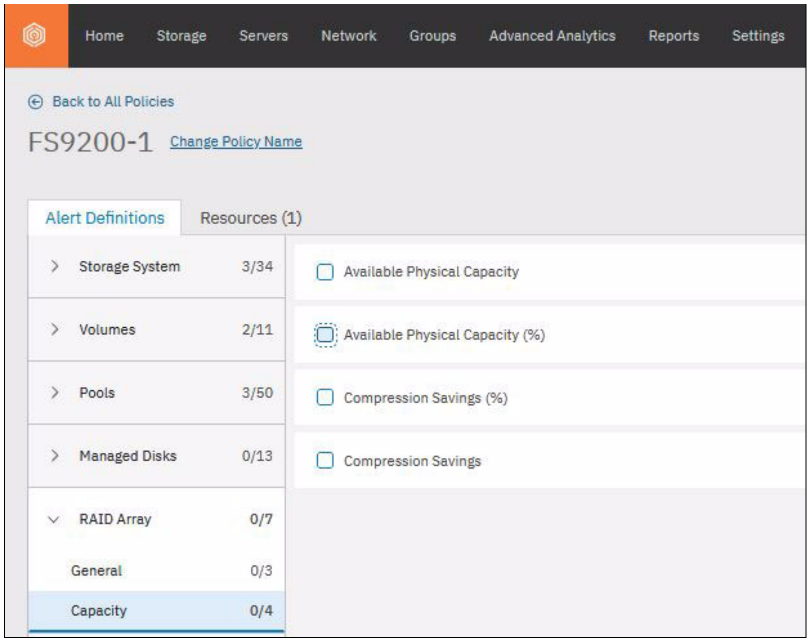


Figure 9-37 Alert Definitions

Figure 9-38 shows the tasks for setting up the Critical definition by monitoring the Used Capacity (%) and releasing Policy Notifications at 15%.

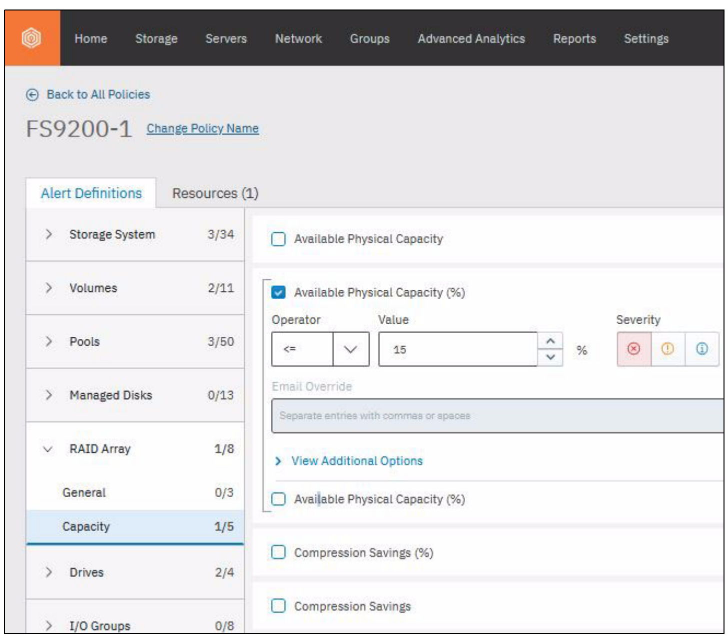


Figure 9-38 Spectrum Control - Alert Definition 15% Available Physical Capacity (%) - Critical

Predefined methods can be one of the following options:

- Email Addresses
- SNMP
- IBM Netcool/OMNIbus
- Windows Event Log or UNIX syslog

These methods must be defined before you can choose them. If your environment does not have predefined methods (see Figure 9-38 on page 430).

Figure 9-39 shows how to change the Frequency of the notification. You can choose here to get more frequent notification for the Critical Threshold “15% Available Physical Capacity (%)”. In this example, we choose to set the frequency to **Send every 1 day**.

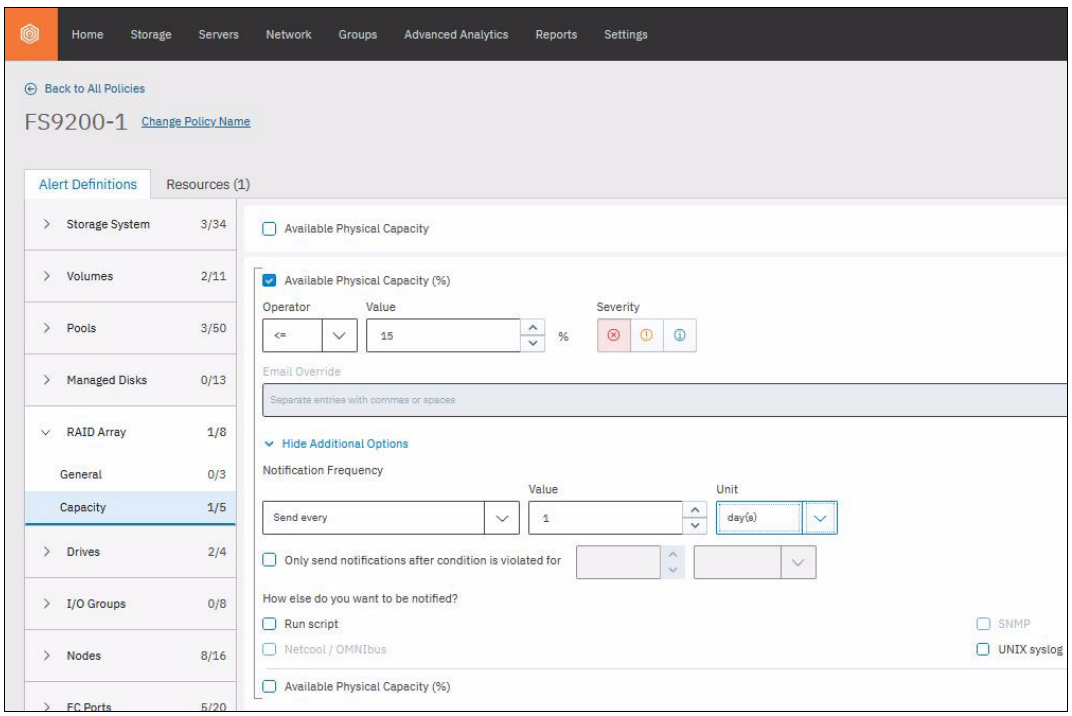


Figure 9-39 Spectrum Control - Alert Definition 15% Available Physical Capacity (%) - Frequency

Figure 9-40 shows how to set up the Warning level at 30% for Available Physical Capacity (%). To proceed, choose the plus sign at the previously defined Definition (Critical) and complete the fields in the window, as shown in Figure 9-40; that is, Operator: “<=”, Value: “30%”, and Severity “Warning”.

Figure 9-40 Spectrum Control - Alert Definition 30% Available Physical Capacity (%) - Warning

Figure 9-41 shows how to set up the Notification Threshold at 30%.; that is, Operator: “<=”, Value: “45%”, and Severity “Notification”.

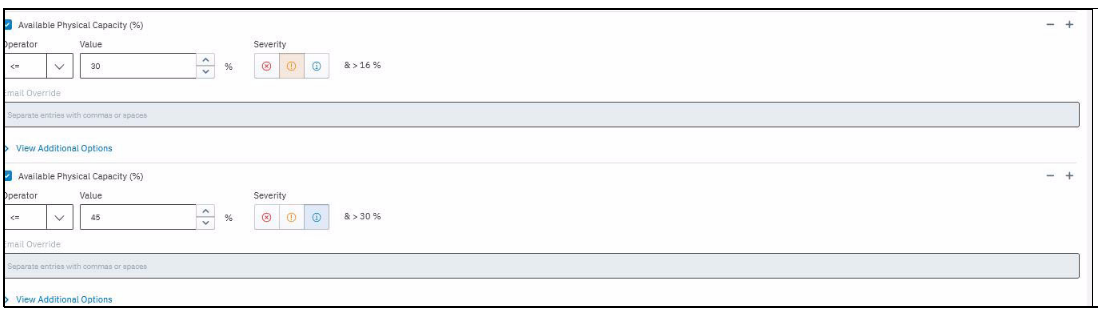


Figure 9-41 Spectrum Control - Alert Definition 45% Available Physical Capacity (%) - Notification

Figure 9-42 shows how to open the Notification Settings in Spectrum Control.

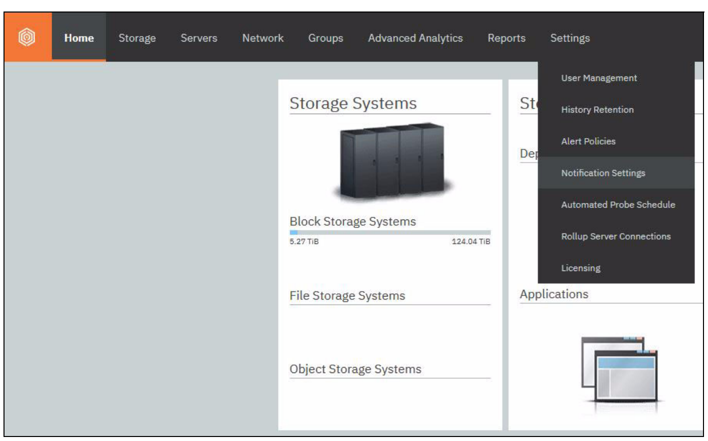


Figure 9-42 Spectrum Control - Notification Settings

Note: With IBM Storage Insights, you can send emails only.

9.6 Error condition example

This section discusses an error condition example.

9.6.1 Offline Fibre Channel port in the management GUI

This section shows an example of an offline Fibre Channel port that is analyzed through the management GUI. It also shows how you can identify the error and drill down into the details.

By using the management GUI dashboard, you can detect errors in the System Health tile.

Each tile contains one type of component, but it can contain multiple items of the same type. For example, a Fibre Channel port is a component; therefore, it is contained in a tile. However, the Fibre Channel tile can contain multiple Fibre Channel ports.

Tiles with errors and warnings are displayed first so that components that require attention have higher visibility. Healthy tiles are sorted in order of importance in day-to-day use.

The following categories of tiles are available:

- *Hardware components* display the health of all components that are specific to the physical hardware.
- *Logical components* display the health of all logical and virtual components in the management GUI.
- *Connectivity components* display the health of all components that are related to the system's connectivity and the relationship between other components or systems.

For more information, see this IBM Documentation web page.

An example of the System Health tile is shown in Figure 9-43.

Figure 9-43 System Health state FlashSystem management GUI

More information about the system health state can be displayed by expanding the components, as shown in Figure 9-44

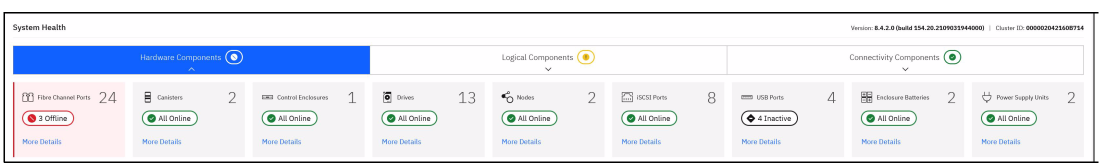


Figure 9-44 Expanded components of the system health tiles in the FlashSystem management GUI

The More Details feature provides an overview overall Fibre Channel ports. Figure 9-45 shows an example of the management GUI that represents the physical and virtualized offline and online Fibre Channel ports.

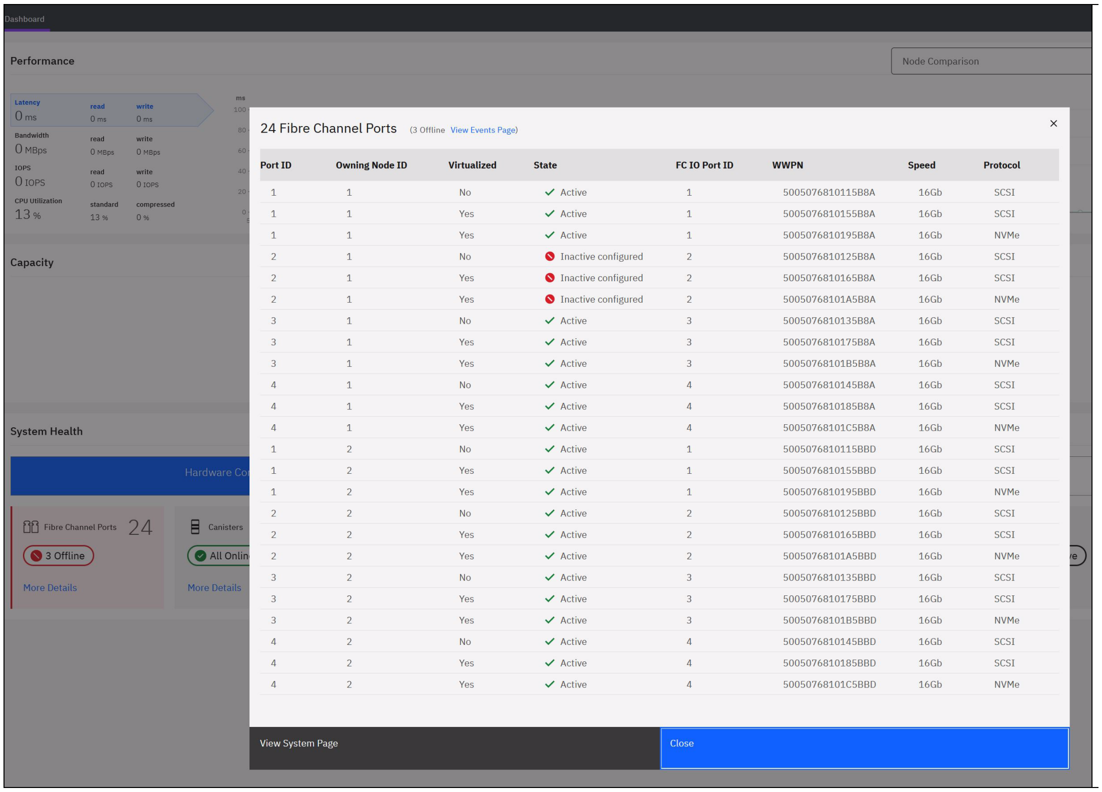


Figure 9-45 More information about expanded system health tiles in FlashSystem management GUI

Figure 9-46 shows an example of cluster error code 1061/1450: `Fibre Channel ports or Fibre Channel I/O ports not operational FlashSystem management GUI.`

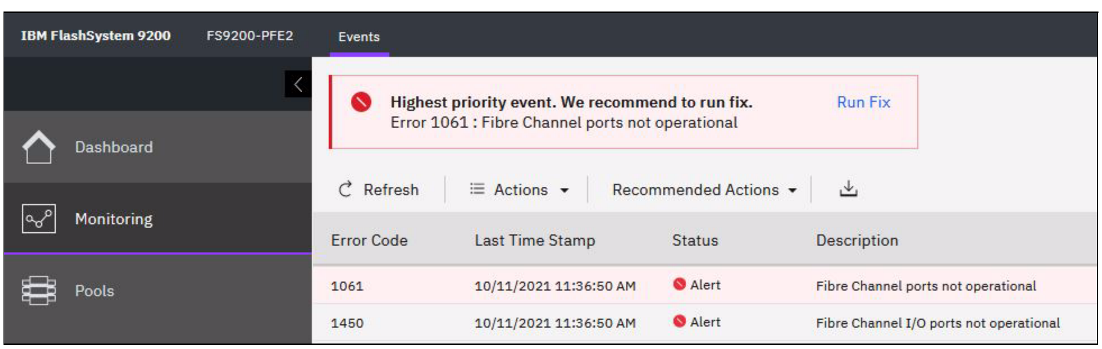


Figure 9-46 Cluster error code 1061/1450

With all these details of the affected Fibre Channel port, the physical components (SFP of the FlashSystem, cable, and the SFP of the SAN Switch) must be checked.

9.6.2 Offline FC-Ports in IBM Spectrum Control and IBM Storage Insights

In this section, we present an example of an IBM FlashSystem with offline FC-Ports in IBM Spectrum Control and IBM Storage Insights. It represents how you can spot the error and drill down into the details.

Figure 9-47 shows the dashboard in which three errors (overall added Block Storage Systems) are detected in IBM Spectrum Control. The Block Storage Systems dashboard also shows the product in an `Error` condition by highlighting it with a red X in the Condition column.

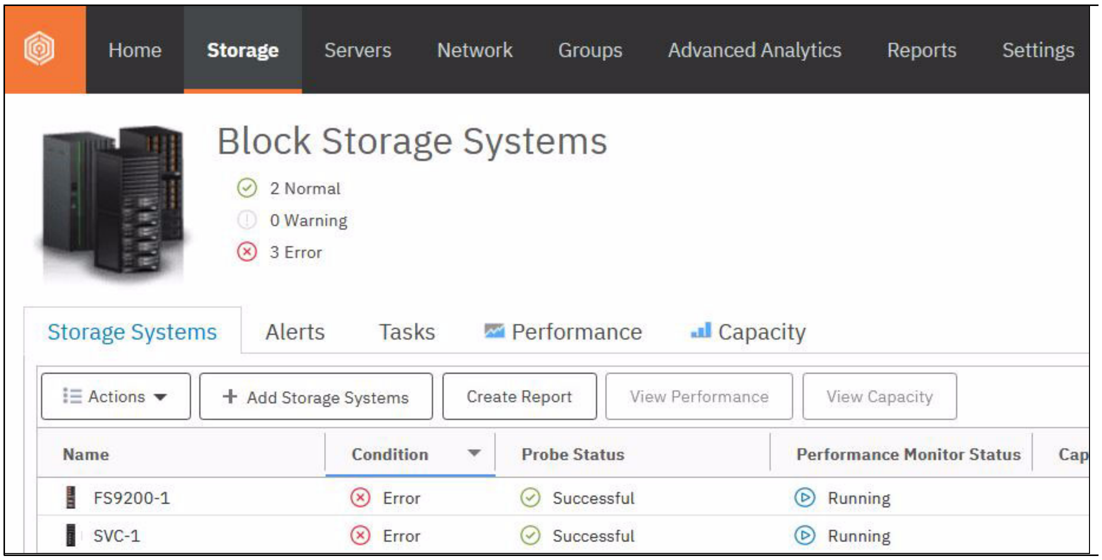


Figure 9-47 Error condition in Spectrum Control

> **Note:** Unless otherwise noted, there is no difference from IBM Spectrum Control to IBM Storage Insights for the steps that are shown next.

The overview page of the IBM FlashSystem product (double-click the device as shown in Figure 9-47) gives you more information about the error condition and the affected layer.

Figure 9-48 shows the error condition of FS9200-1 and denotes the error reported for FC Ports.

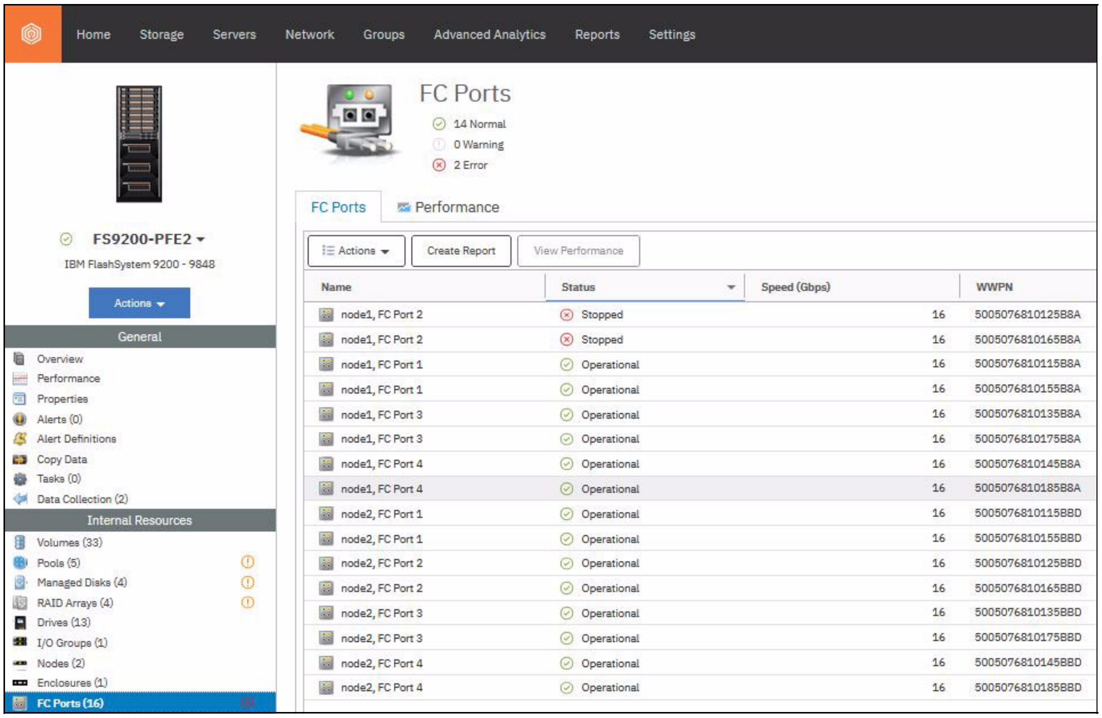


Figure 9-48 FC Ports of “FS9200-1” in Spectrum Control

Figure 9-48 also shows details of the storage subsystem and which entity is affected. In this case it is related to internal resources: ports. Two ports were stopped and caused this condition in the environment.

Figure 9-49 shows the details of one of the stopped ports.

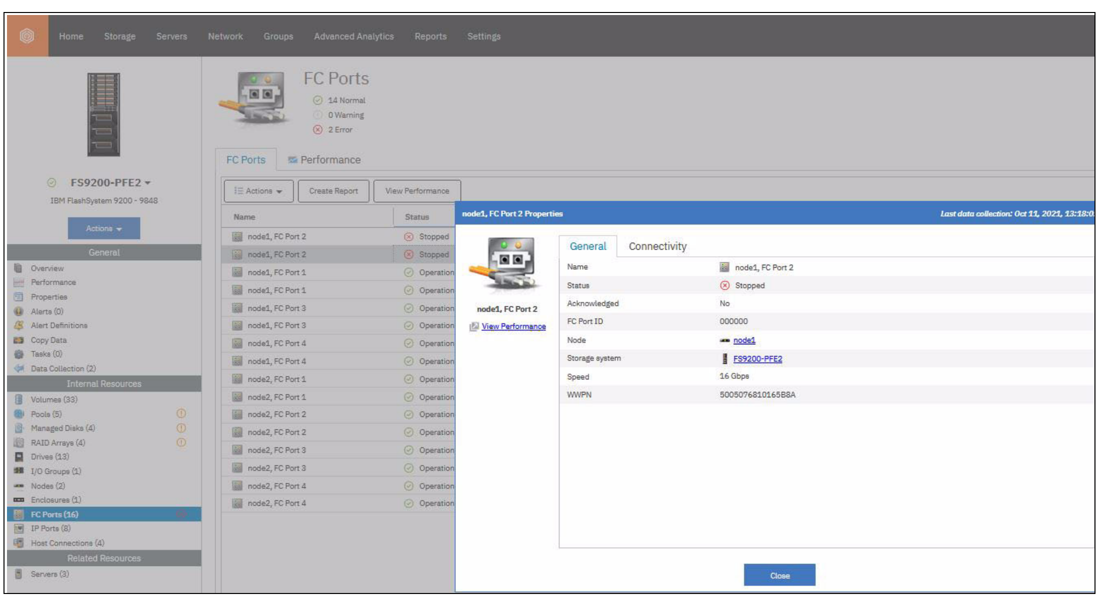


Figure 9-49 FC Ports properties of “FS9200-1” in Spectrum Control

The ports are likely stopped for a reason; therefore, from the window that is shown in Figure 9-48 on page 436 select both, click **Actions** and then, select **Mark Status as Acknowledged** (see Figure 9-50) or right-click the affected part and **Mark Status as Acknowledged** (Figure 9-51). The ports are still shown with the red icon, but now the icon is overlaid with a checkmark. After a short time, this change is propagated so that the storage system is shown as being in a green status again.

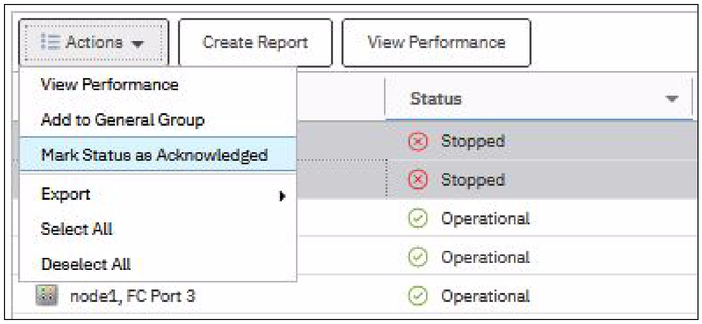


Figure 9-50 Mark Status as Acknowledged per Actions drop down menu in Spectrum Control

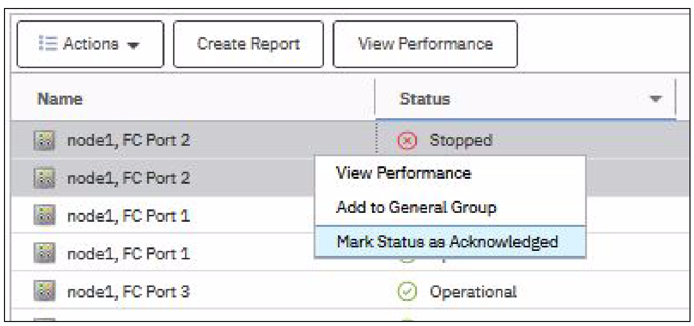


Figure 9-51 Mark Status as Acknowledged per right click in Spectrum Control

Note: If you have Spectrum Control and Storage Insights running, the acknowledgment needs to be done in both instances.

In other cases, you might have to replace hardware after you open a ticket in your internal system with the vendor. You should still acknowledge the status so that any other errors make the storage system go from green to red again and you see that a second event occurred.

9.7 Important metrics

The following metrics are some of the most important metrics that must be analyzed to understand a performance problem in IBM FlashSystem systems. Those metrics are valid to analyze the front end (by node, by host, or by volume) or the back-end (by MDisk or by Storage Pool):

Terminology: R/W stands for Read and Write operations.

- I/O Rate R/W: The term I/O is used to describe any program, operation, or device that transfers data to or from a computer, and to or from a peripheral device. Every transfer is an output from one device and an input into another. Typically measured in IOPS.
- Data Rate R/W: The data transfer rate (DTR) is the amount of digital data that is moved from one place to another in a specific time. In case of Disk or Storage Subsystem, this metric is the amount of data that is moved from a host to a specific storage device. Typically measured in MB per second.
- Response time R/W: This is the time taken for a circuit or measuring device, when subjected to a change in input signal, to change its state by a specified fraction of its total response to that change. In case of Disk or Storage Subsystem, this is the time that is used to complete an I/O operation. Typically measured in ms.
- Cache Hit R/W: This is the percentage of times that read data or write data can be found in cache or can find cache free space that it can be written to.
- Average Data Block Size R/W: The block size is the unit of work for the file system. Every read and write is done in full multiples of the block size. The block size is also the smallest size on disk that a file can have.
- Port-to-Local Node Queue Time (Send): The average time in milliseconds that a send operation spends in the queue before the operation is processed. This value represents the queue time for send operations that are issued to other nodes that are in the local cluster. A good scenario has less than 1 ms on average.
- Port Protocol Errors (Zero Buffer Credit Percentage): The amount of time, as a percentage, that the port was not able to send frames between ports because of insufficient buffer-to-buffer credit. The amount of time value is measured from the last time that the node was reset. In Fibre Channel technology, buffer-to-buffer credit is used to control the flow of frames between ports. In our experience less is better than more. However, in the real life this metric can be from 5% on average up to 20% peak without affecting performance.
- Port data rate (send and receive): The average amount of data in MBps for operations in which the port receives or sends data.
- Port Protocol Errors (Zero Buffer Credit Timer): The number of microseconds that the port is not able to send frames between ports because there is insufficient buffer-to-buffer credit. In Fibre Channel technology, buffer-to-buffer credit is used to control the flow of frames between ports. Buffer-to-buffer credit is measured from the last time that the node was reset. This value is related to the data collection sample interval.
- Port Congestion Index: The estimated degree to which frame transmission was delayed due to a lack of buffer credits. This value is generally 0 - 100. The value 0 means that there was no congestion. The value can exceed 100 if the buffer credit exhaustion persisted for an extended amount of time. When you troubleshoot a SAN, use this metric to help identify port conditions that might slow the performance of the resources to which those ports are connected.

- Global Mirror (Overlapping Write Percentage): The percentage of overlapping write operations that are issued by the Global Mirror primary site. Some overlapping writes are processed in parallel, and so they are excluded from this value.
- Global Mirror (Write I/O Rate): The average number of write operations per second that are issued to the Global Mirror secondary site. Keep in mind that IBM FlashSystem systems have a limited number of GM I/Os that can be delivered.
- Global Mirror (Secondary Write Lag): The average number of extra milliseconds that it takes to service each secondary write operation for Global Mirror. This value does not include the time to service the primary write operations. Monitor the value of Global Mirror Secondary Write Lag to identify delays that occurred during the process of writing data to the secondary site.

> **Note:** The host attributed response time also is an important metric, which should be used with IBM Spectrum Control V5.3.3 or higher. Previous versions featured a calculation error.
>
> Version 5.2.x is *not* supported as of September 30, 2019.

Many others metrics are supplied to IBM Spectrum Control from IBM SAN Volume Controller. For more information about all metrics, see this IBM Documentation web page.

9.8 Performance diagnostic information

If you experience performance issues on your system at any level (Host, Volume, Nodes, Pools, and so on), consult IBM Support, who require detailed performance data about the IBM Spectrum Virtualize system to diagnose the problem. Generate a performance support package with detailed data by collecting a Snap by using IBM Spectrum Control or IBM Storage Insights.

9.8.1 Performance diagnostic information included in a Snap

During the process of generating a snap (click **Settings** → **Support Package** → **Download Support Package**), all performance diagnostic statistics of each node also are captured.

A maximum of 16 files are stored in a directory at any one time for each statistics file type.

Depending on the configured `startstats` interval, the performance statistics are captured frequently.

Use the **`startstats`** command (see Example 9-5) to modify the interval at which per-node statistics for volumes, managed disks (MDisks), and nodes are collected.

Example 9-5 startstats command

```
IBM_IBM_FlashSystem:FS9200-1:superuser>lssystem |grep frequency
statistics_frequency 1
IBM_IBM_FlashSystem:FS9200-1:superuser>startstats -interval 5
IBM_IBM_FlashSystem:FS9200-1:superuser>lssystem |grep frequency
statistics_frequency 5
```

If an interval of 5 minutes (default value) is configured, a time frame of 80 minutes (5min*16 = 80 minutes) in the past is covered by a snap.

Note: The lower the value for the interval is set, the shorter is the time frame that is covered in the performance statistics of the snap. However, the statistic values are much more precise. With a large interval value, a larger time frame is covered, but the performance statistic values are to some way imprecise and some peaks might not be visible.

9.8.2 Performance diagnostic information exported from IBM Spectrum Control

You can export performance diagnostic data for a managed resource. You might be asked to send this data if you contact IBM Support to help you analyze a performance problem with storage systems or fabrics.

The performance data might be large, especially if the data is for storage systems that have many volumes, or the performance monitors are running with a 1-minute sampling frequency. If the time range for the data is greater than 12 hours, volume data and 1-minute sample data is automatically excluded from the performance data, even if it is available.

When you export performance data, you can specify a time range to export performance data for. The time range cannot exceed the history retention limit for sample performance data. By default, this history retention limit is *two weeks*.

To export hourly or daily performance data, use the `exportPerformanceData` script. However, the time range still cannot exceed the history retention limits for the type of performance data.

Complete the following steps:

1. In the menu bar, select the type of storage system.

   For example, to create a compressed file for a block storage system, select **Storage** → **Block** → **Storage Systems**.

2. To create a compressed file for a fabric, click **Network** → **Fabrics**.
3. Right-click the storage resource, and then, click **Export Performance Data** (see Figure 9-52 on page 441).

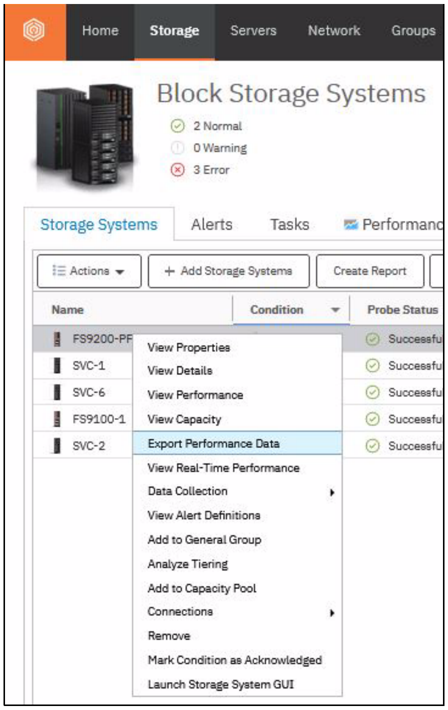


Figure 9-52 Spectrum Control - Export Performance Data

To include volume data and 1-minute sample data, select the **Advanced package** (see Figure 9-53) option when you export performance data.

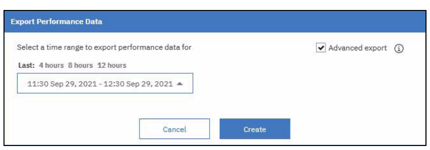
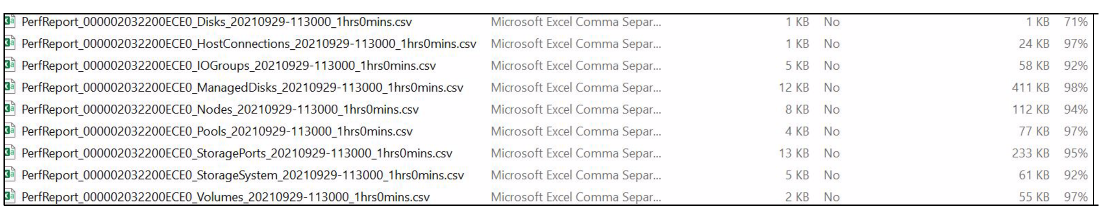


Figure 9-53 Spectrum Control - Export Performance Data - Advanced Export

After the package is created, the `.zip` file can be downloaded through the browser. The package includes different reports in `.csv` format, as shown in Figure 9-54.



PerfReport_000002032200ECE0_Disks_20210929-113000_1hrs0mins.csv	Microsoft Excel Comma Separ...	1 KB	No	1 KB	71%
PerfReport_000002032200ECE0_HostConnections_20210929-113000_1hrs0mins.csv	Microsoft Excel Comma Separ...	1 KB	No	24 KB	97%
PerfReport_000002032200ECE0_IOGroups_20210929-113000_1hrs0mins.csv	Microsoft Excel Comma Separ...	5 KB	No	58 KB	92%
PerfReport_000002032200ECE0_ManagedDisks_20210929-113000_1hrs0mins.csv	Microsoft Excel Comma Separ...	12 KB	No	411 KB	98%
PerfReport_000002032200ECE0_Nodes_20210929-113000_1hrs0mins.csv	Microsoft Excel Comma Separ...	8 KB	No	112 KB	94%
PerfReport_000002032200ECE0_Pools_20210929-113000_1hrs0mins.csv	Microsoft Excel Comma Separ...	4 KB	No	77 KB	97%
PerfReport_000002032200ECE0_StoragePorts_20210929-113000_1hrs0mins.csv	Microsoft Excel Comma Separ...	13 KB	No	233 KB	95%
PerfReport_000002032200ECE0_StorageSystem_20210929-113000_1hrs0mins.csv	Microsoft Excel Comma Separ...	5 KB	No	61 KB	92%
PerfReport_000002032200ECE0_Volumes_20210929-113000_1hrs0mins.csv	Microsoft Excel Comma Separ...	2 KB	No	55 KB	97%

Figure 9-54 Spectrum Control - Package files example

For more information about how to create a performance support package, see this IBM Documentation web page.

9.8.3 Performance diagnostic information exported from IBM Storage Insights

To help resolve performance issues with storage systems, you can also export performance data for the resource to a compressed file from IBM Storage Insights. Complete the following steps:

1. To export the performance data, select the type of storage system in the menu bar.

   For example, to create a compressed file for a block storage system, select **Resources** → **Block Storage Systems** (see Figure 9-55).

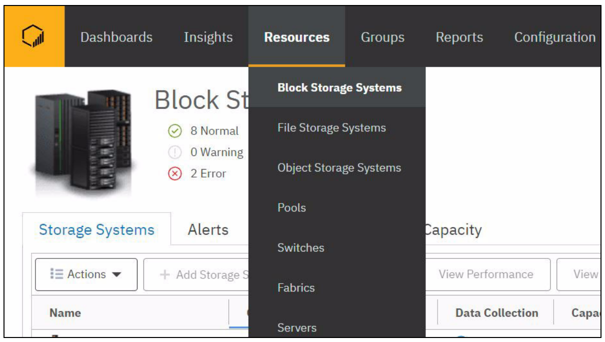


Figure 9-55 Selecting Block Storage Systems

2. Right-click the storage system and select **Export Performance Data** (see Figure 9-56).

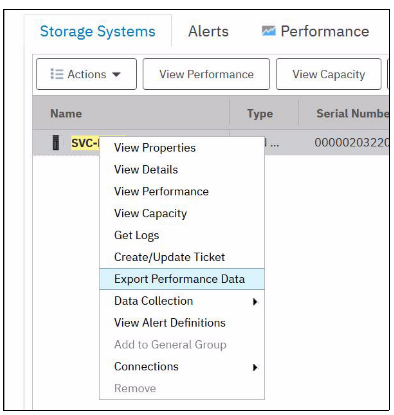


Figure 9-56 Export Performance Data

3. Select the time range of the performance data that you want to export. You can select a time range of the previous 4, 8, or 12 hours, or specify an earlier time range by clicking the time and date.

   Because the amount of performance data might be large (especially for storage systems that have many volumes), volume data is exported only if the time range is less than 12 hours. For time ranges of 12 or more hours, click **Advanced export** to include volume data.

4. Click **Create**.

   A task is started and shown in the Running tasks icon in the menu bar.

5. When the task is complete, click the **Download** icon in the running tasks list in the task to save the file locally.

For more information about how to create a performance support package, see this IBM Documentation web page.

9.9 Metro and Global Mirror monitoring with IBM Copy Services Manager and scripts

Copy Services Manager is part of IBM Spectrum Control and controls copy services in storage environments. Copy services are features that are used by storage systems, such as IBM FlashSystem systems, to configure, manage, and monitor data-copy functions. Copy services include IBM FlashCopy, Metro Mirror, Global Mirror, and Global Mirror Change Volumes (GMCV).

You can use Copy Services Manager to complete the following data replication tasks and help reduce the downtime of critical applications:

- Plan for replication when you are provisioning storage
- Keep data on multiple related volumes consistent across storage systems if there is a planned or unplanned outage
- Monitor and track replication operations
- Automate the mapping of source volumes to target volumes

One of the most important events that needs to be monitored when IBM FlashSystem systems are implemented in a disaster recovery (DR) solution with Metro Mirror (MM) or Global Mirror (GM) functions, is to check whether MM or GM has been suspended because of a 1920 or 1720 error.

With IBM FlashSystem systems, you can suspend the MM or GM relationship to protect the performance on the primary site when MM or GM starts to affect write response time. That suspension can be caused by several factors.

IBM FlashSystem systems do not restart MM or GM automatically. They must be restarted manually.

IBM FlashSystem systems alert monitoring is explained in 9.1.1, “Monitoring by using the GUI” on page 388. When MM or GM is managed by IBM CSM and if a cluster error code 1920 occurs, IBM CSM can automatically restart MM or GM sessions, and can set the delay time on the automatic restart option. This delay allows some time for the situation to correct itself.

Alternatively, if you have several sessions, you can stagger them so that they do not all restart at the same time, which can affect system performance. Choose the set delay time feature to define a time, in seconds, for the delay between when Copy Services Manager processes the 1720/1920 event and when the automatic restart is issued.

CSM is also able to automatically restart unexpected suspends. When you select this option, the Copy Services Manager server automatically restarts the session when it unexpectedly suspends due to reason code 1720 or 1920. An automatic restart is attempted for every suspend with reason code 1720 or 1920 up to a predefined number of times within a 30-minute time period.

The number of times that a restart is attempted is determined by the storage server `gmlinktolerance` value. If the number of allowable automatic restarts is exceeded within the time period, the session does not restart automatically on the next unexpected suspend.

Issue a `Start` command to restart the session, clear the automatic restart counters, and enable automatic restarts.

> **Warning:** When you enable this option, the session is automatically restarted by the server. When this situation occurs, the secondary site is not consistent until the relationships are fully resynched.

You can specify the amount of time (in seconds) that the copy services management server waits after an unexpected suspend before automatically restarting the session. The range of possible values is 0 - 43200. The default is 0, which specifies that the session is restarted immediately following an unexpected suspend.

9.9.1 Monitoring MM and GM with scripts

The IBM FlashSystem system provides a complete command-line interface (CLI), which you can use to interact with your systems by using scripts. The scripts can run in the IBM FlashSystem shell, but with a limited script command set available, or they can run out of the shell using any preferred scripting language.

An example of script usage is one to check at a specific interval time whether MM or GM are still active, if any 1920 errors have occurred, or to react to an SNMP or email alert received. The script can then start some specific recovery action based on your recovery plan and environment.

Customers who do not use IBM Copy Service Manager have created their own scripts. These scripts are sometimes supported by IBM as part of ITS professional services or IBM System Lab services. Tell your IBM representative what kind of monitoring you want to implement with scripts, and together try to find if one exists in the IBM Intellectual Capital Management repository that can be reused.

9.10 Monitoring Tier1 solid-state drives

Monitoring Tier1 SSDs requires that special attention must be paid to the endurance events that can be triggered. For monitoring purposes, make note of the new fields that are listed in Table 9-5.

Table 9-5 Field changes to drive and array devices

Field	Description
write_endurance_used	Metric pulled from within drive (SAS spec) relating to the amount of data that is written across the life of the drive that is divided by the anticipated amount (2.42 PB for the 15.36 TB drive) Starts at 0, and can continue > 100
write_endurance_usage_rate	Measuring/Low/Marginal/High Takes 160 Days to get initial measurement; Low: Approximately 5.5 Years or more Marginal: Approximately 4.5 – 5.5 Years High: Approximately < 4.5 years High triggers event `SS_EID_VL_ER_SSD_WRITE_ENDURANCE_USAGE_RATE_HIGH`
replacement_date	The Current Date + Endurance Rate * Remaining Endurance Triggers event `SS_EID_VL_ER_SSD_DRIVE_WRITE_ENDURANCE_LIMITED` at 6 Months before limit

If you see either of these triggered events, contact your IBM service representative to put an action plan in place:

```
SS_EID_VL_ER_SSD_WRITE_ENDURANCE_USAGE_RATE_HI4GH
SS_EID_VL_ER_SSD_DRIVE_WRITE_ENDURANCE_LIMITED
```

10

Maintaining storage infrastructure

As an IT environment grows and is renewed, so must the storage infrastructure. One of the many benefits that the IBM FlashSystem family software (IBM Spectrum Virtualize) provides, is to greatly simplify the storage management tasks that system administrators need to perform.

This chapter highlights guidance for the maintenance activities of storage administration by using the IBM FlashSystem family software that is installed on the product. This guidance can help you to maintain your storage infrastructure with the levels of availability, reliability, and resiliency that are demanded by today's applications, and to keep up with storage growth needs.

This chapter concentrates on the most important topics to consider in IBM FlashSystem administration so that you can use it as a checklist. It also provides best practice tips and guidance. To simplify the Storage Area Network (SAN) storage administration tasks that you use often, such as adding new users, storage allocation and removal, or adding and removing a host from the SAN, create step-by-step, standard procedures for them.

The discussion in this chapter focuses on the IBM FlashSystem 9200 for the sake of simplicity, using screen captures and command outputs from this model. The recommendations and practices that are discussed in this chapter are applicable to the following models:

- IBM FlashSystem 5010
- IBM FlashSystem 5015
- IBM FlashSystem 5030
- IBM FlashSystem 5035
- IBM FlashSystem 5100
- IBM FlashSystem 5200
- IBM FlashSystem 7200
- IBM FlashSystem 9100
- IBM FlashSystem 9200

Note: The practices that are described in this chapter were effective in many installations of different models of the IBM FlashSystem family. These installations were performed in various business sectors for various international organizations. They all had one common need: to manage their storage environment easily, effectively, and reliably.

This chapter includes the following topics:

10.1 User interfaces

The IBM FlashSystem family provides several user interfaces to allow you to maintain your system. The interfaces provide different sets of facilities to help resolve situations that you might encounter. The interfaces for servicing your system connect through the 1 Gbps Ethernet ports that are accessible from port 1 of each canister.

- Use the management graphical user interface (GUI) to monitor and maintain the configuration of storage that is associated with your clustered systems.
- Use the service assistant tool GUI to complete service procedures.
- Use the command-line interface (CLI) to manage your system.

The best practice recommendation is to use the interface most appropriate to the task you are attempting to complete. For example, a manual software update is best performed using the service assistant GUI or the CLI. Running fix procedures to resolve problems or configuring expansion enclosures can only be performed using the management GUI. The creation of many volumes with customized names is best performed using the CLI using a script. To ensure efficient storage administration, become familiar with all available user interfaces.

10.1.1 Management GUI

The management GUI is the primary tool that is used to service your system. Regularly monitor the status of the system using the management GUI. If you suspect a problem, use the management GUI first to diagnose and resolve the problem. Use the views that are available in the management GUI to verify the status of the system, the hardware devices, the physical storage, and the available volumes.

To access the Management GUI, start a supported web browser and go to `https://<flashsystem_ip_address>`, where the `<flashsystem_ip_address>` is the management IP address set when the clustered system is created.

For more information about the task menus and functions of the Management GUI, see Chapter 4 of *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize Version 8.4*, SG24-8491.

10.1.2 Service assistant tool GUI

The service assistant interface is a browser-based GUI that can be used to service individual node canisters in the control enclosures.

> **Important:** If used incorrectly, the service actions that are available through the service assistant can cause loss of access to data or even data loss.

You connect to the service assistant on one node canister by entering the service IP address. If there is a working communications path between the node canisters, you can view status information and perform service tasks on the other node canister by making the other node canister the current node. You do not have to reconnect to the other node. On the system itself, you can also access the service assistant interface by using the technician port.

The service assistant only provides facilities to help you service control enclosures. Always service the expansion enclosures by using the management GUI.

You can also complete the following actions using the service assistant:

- Collect logs to create and download a package of files to send to support personnel.
- Provide detailed status and error summaries.
- Remove the data for the system from a node.
- Recover a system if it fails.
- Install a code package from the support site or rescue the code from another node.
- Update code on node canisters manually.
- Configure a control enclosure chassis after replacement.
- Change the service IP address that is assigned to Ethernet port 1 for the current node canister.
- Install a temporary SSH key if a key is not installed and CLI access is required.
- Restart the services used by the system.

To access the Service Assistant Tool GUI, start a supported web browser and go to: https://<flashsystem_ip_address>/service, where <flashsystem_ip_address> is the service IP address for the node canister or the management IP address for the system on which you want work.

10.1.3 Command-line interface

The system CLI is intended for use by advanced users who are confident using a CLI. Up to 32 simultaneous interactive Secure Shell (SSH) sessions to the management IP address are supported.

Nearly all the functions that is offered by the CLI is available through the management GUI. However, the CLI does not provide the fix procedures that are available in the management GUI. Alternatively, use the CLI when you require a configuration setting that is unavailable in the management GUI.

Entering `help` in a CLI displays a list of all available commands. You have access to a few other UNIX commands in the restricted shell, such as **`grep`** and **`more`**, which are useful in formatting output from the CLI commands. Reverse-i-search (Ctrl+R) is also available. Table 10-1 shows a list of UNIX commands:

Table 10-1 UNIX commands available in the CLI

UNIX command	Description
grep	Filter output by keywords
more	Moves through output one page at a time
sed	Filters output
sort	Sorts output
cut	Removes individual columns from output
head	Display only first lines
less	Moves through the output one page at a time
tail	Display only last lines
uniq	Hides any duplicates in the output

UNIX command	Description
tr	Translates characters
wc	Counts lines, words and characters in the output
history	Display command history
scp	Secure copy protocol

For more information about command reference and syntax, see the following resources:

- This IBM Documentation web page
- *IBM Spectrum Virtualize for SAN Volume Controller, FlashSystem, and Storwize Family - Command-Line Interface User's Guide*

Service command-line interface

Service CLI commands also can be run on a specific node. To run such a command in this way, log in to the service IP address of the node that requires servicing.

For more information about the use of the service command line, see this IBM Documentation web page.

USB command interface

When a USB flash drive is inserted into one of the USB ports on a node, the software searches for a control file (`satask.txt`) on the USB flash drive and runs the command that is specified in the file. Using the USB flash drive is required in the following situations:

- When you cannot connect to a node canister in a control enclosure using the service assistant and you want to see the status of the node.
- When you do not know, or cannot use, the service IP address for the node canister in the control enclosure and must set the address.
- When you have forgotten the superuser password and must reset the password.

For more information about the use of the USB port, see this IBM Documentation web page.

Technician port

The technician port is an Ethernet port on the back window of the IBM FlashSystem product that you can use to configure the node. You can use the technician port to do most of the system configuration operations, which includes the following tasks:

- Defining a management IP address
- Initializing a new system
- Servicing the system

For more information about the use of the Technician port, see this IBM Documentation web page.

10.2 Users and groups

Almost all organizations have IT security policies that enforce the use of password-protected user IDs when their IT assets and tools are used. However, some storage administrators still use generic shared IDs, such as `superuser`, `admin` or `root`, in their management consoles to perform their tasks. They might even use a factory-set default password. Their justification might be a lack of time, forgetfulness, or the fact that their SAN equipment does not support the organization's authentication tool.

SAN storage equipment management consoles often do not provide direct access to stored data, but one can easily shut down (accidentally or deliberately) 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 auditing of changes if you must analyze your logs.

IBM FlashSystem supports the following authentication methods:

- Local authentication using a password
- Local authentication using SSH keys
- Remote authentication using Lightweight Directory Access Protocol (LDAP) (Microsoft Active Directory or IBM Security Directory Server)

Local authentication is appropriate for small, single enclosure environments. Larger environments with multiple clusters and enclosures benefit from the ease of maintenance that is achieved by using single sign-on (SSO) that uses remote authentication using LDAP, for example.

By default, the following user groups are defined:

- Monitor: Users with this role can view objects but cannot manage the system or its resources. Support personnel can be assigned this role to monitor the system and to determine the cause of problems. This role must be assigned to the IBM Storage Insights user. For more information about IBM Storage Insights, see Chapter 9, "Implementing a storage monitoring system" on page 387.
- Copy Operator: Users with this role have monitor role privileges and can create, change, and manage all Copy Services functions.
- Service: Users can set the time and date on the system, delete dump files, add and delete nodes, apply service, and shut down the system. Users can also perform the same tasks as users in the monitor role.
- Administrator: Users with this role can access all functions on the system except those that deal with managing users, user groups, and authentication.
- Security Administrator: Users with this role can access all functions on the system, including managing users, user groups, and user authentication.
- Restricted Administrator: Users with this role can complete some tasks, but are restricted from deleting certain objects. Support personnel can be assigned this role to solve problems.
- 3-Site Administrator: Users with this role can configure, manage, and monitor 3-site replication configurations through certain command operations only available on the 3-Site Orchestrator.
- vStorage Application Programming Interface (API) for Storage Awareness (VASA) Provider: Users with this role can manage virtual volumes (vVols) that are used by VMware vSphere and managed through IBM Spectrum Control software.

- FlashCopy Administrator: Use the FlashCopy commands to work with FlashCopy system methods and functions. For more information, see this IBM Documentation web page.

In addition to standard groups, you can also configure ownership groups to manage access to resources on the system. An ownership group defines a subset of users and objects within the system. You can create ownership groups to further restrict access to specific resources that are defined in the ownership group.

Users within an ownership group can view or change only resources within the ownership group in which they belong. For example, you can create an ownership group for database administrators to provide monitor-role access to a single pool used by their databases. Their views and privileges in the management GUI are automatically restricted, as shown in Figure 10-1.

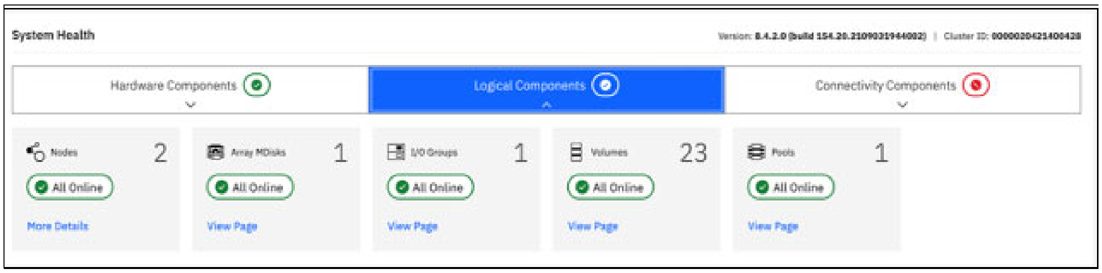


Figure 10-1 System Health Logical Components view

Figure 10-2 shows the Dashboard System Health hardware components view.

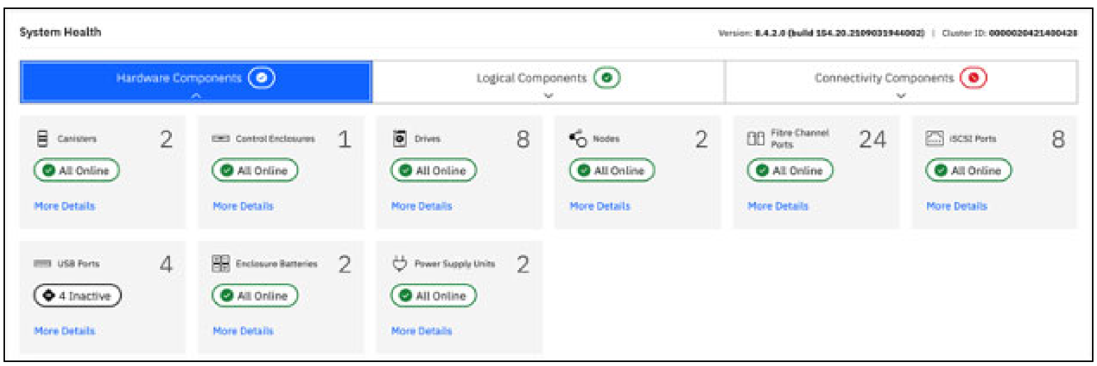


Figure 10-2 System Health Hardware Components view

Regardless of the authentication method you choose, complete the following tasks:

- Create individual user IDs for your Storage Administration staff. Choose user IDs that easily identify the user and meet your organization's security standards.
- Include each individual user ID into the UserGroup with only enough privileges to perform the required tasks. For example, your first-level support staff probably only require Monitor group access to perform their daily tasks, whereas second-level support might require Restricted Administrator access. Consider using Ownership groups to further restrict privileges.
- If required, create generic user IDs for your batch tasks, such as Copy Services or Monitoring. Include them in a Copy Operator or Monitor UserGroup. Never use generic user IDs with the SecurityAdmin privilege in batch tasks.
- Create unique SSH public and private keys for each administrator requiring local access.

- Store your `superuser` password in a safe location in accordance to your organization's security guidelines and use it only in emergencies.

10.3 Volumes

A volume is a logical disk that is presented to a host by an I/O group (pair of nodes), and within that group a preferred node serves I/O requests to the volume.

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 disk path errors that you find in your server before you proceed.
- When you plan for future growth of space efficient volumes (VDisks), determine whether your server's operating system supports the particular volume to be extended online. AIX V6.1 TL2 and lower, for example, do not support online expansion of rootvg logical unit numbers (LUNs). Test the procedure in a nonproduction server first.
- Always cross-check the host LUN ID information with the vdisk_UID of the IBM FlashSystem. 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 IBM FlashSystem.
- Ensure that you delete any volume or LUN definition in the server *before* you unmap it in the IBM FlashSystem. For example, in AIX, remove the hdisk from the volume group (`reducevg`) and delete the associated hdisk device (`rmdev`).
- Consider enabling volume protection by using **`chsystem vdiskprotectionenabled yes -vdiskprotectiontime <value_in_minutes>`**. Volume protection ensures that some CLI actions (most of those that either explicitly or implicitly remove host-volume mappings or delete volumes) are policed to prevent the removal of mappings to volumes or deletion of volumes that are considered *active*; the system detected I/O activity to the volume from any host within a specified time period (15 - 1440 minutes).

  > **Note:** Volume protection cannot be overridden using the **`-force`** flag in the affected CLI commands. Volume protection must be disabled to carry on an activity that is blocked.

- Ensure that you explicitly remove a volume from any volume-to-host mappings and any copy services relationship to which it belongs *before* you delete it.

  > **Attention:** You must avoid the use of the **`-force`** parameter in **`rmvdisk`**.

- If you issue the **`svctask rmvdisk`** command and it still has pending mappings, the IBM FlashSystem prompts you to confirm the action and this is a hint that you might have done something incorrectly.
- When you are deallocating volumes, plan for an interval between unmapping them to hosts (**`rmvdiskhostmap`**) and deleting them (**`rmvdisk`**). The IBM internal Storage Technical Quality Review Process (STQRP) asks for a minimum of a 48-hour period, and having at least a one business day interval so that you can perform a quick backout if you later realize you still need some data on that volume.

For more information about volumes, see Chapter 5, "Volume types" on page 199.

10.4 Hosts

A host is a computer that is connected to the SAN switch through Fibre Channel (FC), iSCSI, and other protocols.

When you add and remove hosts in the IBM FlashSystem, consider the following guidelines:

- Before you map new servers to the IBM FlashSystem, verify that they are all error free. Fix errors that you find in your server and IBM FlashSystem before you proceed. In the IBM FlashSystem, pay special attention to anything inactive in the **lsfabric** command.
- Plan for an interval between updating the zoning in each of your redundant SAN fabrics, such as at least 30 minutes. This interval allows for failover to occur and stabilize, and for you to be notified if unexpected errors occur.
- After you perform the SAN zoning from one server's host bus adapter (HBA) to the IBM FlashSystem, you should list its worldwide port name (WWPN) by using the **lshbaportcandidate** command. Use the **lsfabric** command to certify that it was detected by the IBM FlashSystem nodes and ports that you expected. When you create the host definition in the IBM FlashSystem (**mkhost**), try to avoid the **-force** parameter. If you do not see the host's WWPNs, it might be necessary to scan fabric from the host. For example, use the **cfgmgr** command in AIX.

For more information about hosts, see Chapter 8, "Configuring hosts" on page 367.

10.5 Software updates

Because the IBM FlashSystem might be at the core of your disk and SAN storage environment, the software update procedure requires planning, preparation, and verification. However, with the appropriate precautions, an update can be conducted easily and transparently to your servers and applications. This section highlights applicable guidelines for the IBM FlashSystem update.

Most of the following sections explain how to prepare for the software update. These sections also present version-independent guidelines on how to update the IBM FlashSystem family systems and flash drives.

Before you update the system, ensure that the following requirements are met:

- The latest update test utility is downloaded from IBM Fix Central to your management workstation. For more information, see this IBM Fix Central web page.
- The latest system update package is downloaded from IBM Fix Central to your management workstation.
- All node canisters are online.
- All errors in the system event log are addressed and marked as fixed.
- There are no volumes, MDisks, or storage systems with Degraded or Offline status.
- The service assistant IP is configured on every node in the system.
- The system superuser password is known.
- The system configuration is backed up and saved (preferably off-site), as shown in Example 10-11 on page 512.
- You can physically access the hardware.

The following actions are not required, but are recommended to reduce unnecessary load on the system during the update:

- Stop all Metro Mirror, Global Mirror, or HyperSwap operations.
- Avoid running FlashCopy operations.
- Avoid migrating or formatting volumes.
- Stop collecting IBM Spectrum Control performance data for the system.
- Stop automated jobs that access the system.
- Ensure that no other processes are running on the system.
- If you want to update without host I/O, then shut down all hosts.

> **Note:** For customers who purchased the IBM FlashSystem 9200 with a three-year warranty (9848 Models AG8 and UG8), with Enterprise Class Support (ECS) is included. This support entitles the customer to two code upgrades per year, which are performed by IBM (for total of six across the three-year warranty).
>
> These upgrades are done by the IBM dedicated Remote Code Load (RCL) team or, where remote support is not allowed or enabled, by an onsite Systems Service Representative (SSR). A similar optional service is available for the IBM FlashSystem 7200.
>
> For more information about ECS, see IBM FlashSystem 9200 8.4.2 Documentation - Enterprise Class Support (ECS).

10.5.1 Deciding the target software level

The first step is to determine your current and your target IBM FlashSystem software level.

Using the example of an IBM FlashSystem 9200, log in to the web-based GUI and find the current version. Either from the right side of the top menu drop-down line, click the question mark symbol (**?**) and select **About IBM FlashSystem 9200** to display the current version or select **Settings** → **System** → **Update System** to display both current and target levels.

Figure 10-3 shows the Update System output window and displays the code levels. In this example, the software level is 8.4.2.0.

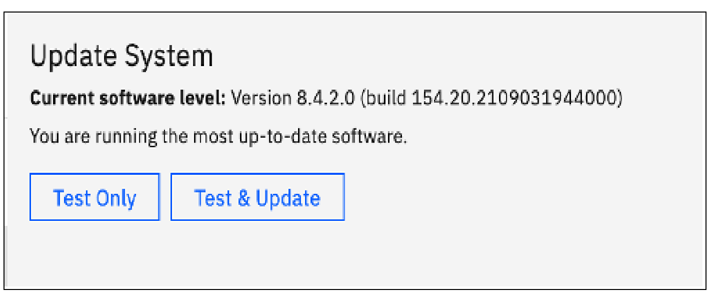


Figure 10-3 Update System output window

Alternatively, if you use the CLI, run the **svcinfo lssystem** command. Example 10-1 shows the output of the **lssystem** CLI command and where the code level output can be found.

Example 10-1 lssystem command

```
IBM_FlashSystem:IBM Redbook FS:superuser>lssystem|grep code
Version 8.4.2.0 (build 154.20.2109031944000)
```

IBM FlashSystem software levels are specified by four digits in the following format (in our example V.R.M.F = 8.4.2.0):

- V is the major version number
- R is the release level
- M is the modification level
- F is the fix level

Use the latest IBM FlashSystem release unless you have a specific reason not to update, such as the following examples:

- The specific version of an application or other component of your SAN Storage environment has a known problem or limitation.
- The latest IBM FlashSystem software 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 release minus 1” or requiring “seasoning” in the field before implementation in a production environment.
- For more information, see Spectrum Virtualize Family of Products Upgrade Planning.

Obtaining the software packages

To obtain a new release of software for a system update, see IBM Fix Central and follow these steps:

1. From the **Product selector** list, type **IBM FlashSystem 9200** (or whatever model is appropriate in your environment).
2. From the Installed Version list, select the current software version level that was determined in 10.5.1, “Deciding the target software level” on page 456.
3. Select **Continue**.

4. In the **Product Software** section, select the three items that are shown in Figure 10-4.

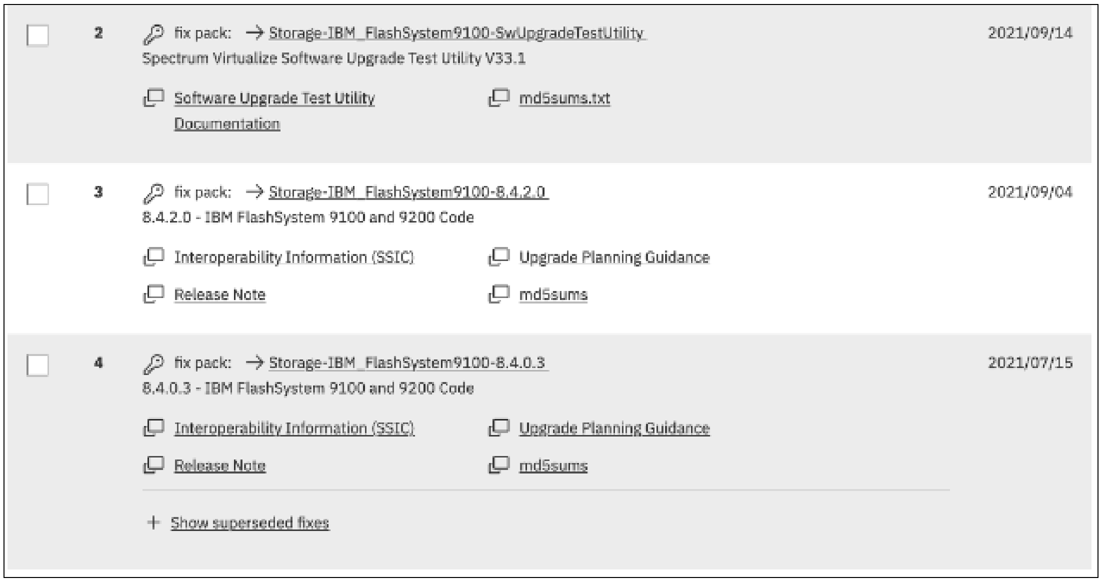


Figure 10-4 Fix Central software packages

5. Select **Continue.**
6. Click the option button for your preferred download options and click **Continue.**
7. Enter your machine type and serial number.
8. Select **Continue.**
9. Read the terms and conditions and then, select **I Agree**.
10. Select **Download Now** and save the three files onto your management computer.

10.5.2 Hardware considerations

Before you start the update process, always check whether your IBM FlashSystem hardware and target code level are compatible.

If part or all your current hardware is not supported at the target code level that you want to update to, replace the unsupported hardware with newer models before you update to the target code level.

Conversely, if you plan to add or replace hardware with new models to an existing cluster, you might have to update your IBM FlashSystem code first.

10.5.3 Update sequence

Check the compatibility of your target IBM FlashSystem code level with all components of your SAN storage environment (SAN switches, storage controllers, server HBAs) and its attached servers (operating systems and eventually, applications).

Applications often certify only the operating system that they run under and leave to the operating system provider the task of certifying its compatibility with attached components (such as SAN storage). However, various applications might use special hardware features or raw devices and certify the attached SAN storage. If you have this situation, consult the compatibility matrix for your application to certify that your IBM FlashSystem target code level is compatible.

The IBM FlashSystem Supported Hardware List provides the complete information for using your IBM FlashSystem SAN storage environment components with the current and target code level. For links to the Supported Hardware List, Device Driver, Firmware, and Recommended Software Levels for different products and different code levels, see this IBM Support web page.

By cross-checking the version of IBM FlashSystem is compatible with the versions of your SAN environment components, you can determine which one to update first. By checking a component's update path, you can determine whether that component requires a multistep update.

If you are not making major version or multi-step updates in any components, the following update order is recommended to avoid eventual problems:

1. SAN switches or directors
2. Storage controllers
3. Servers HBAs microcode and multipath software
4. IBM FlashSystem
5. IBM FlashSystem internal Non-Volatile Memory express (NVMe) drives
6. IBM FlashSystem Serial Attached SCSI (SAS) attached solid-state drive (SSD)

Attention: Do *not* update two components of your IBM FlashSystem SAN storage environment simultaneously, such as an IBM FlashSystem 9200 and one storage controller. This caution is true even if you intend to do it with your system offline. An update of this type can lead to unpredictable results, and an unexpected problem is more difficult to debug.

10.5.4 SAN fabrics preparation

If you are using symmetrical, redundant, independent SAN fabrics, preparing these fabrics for an IBM FlashSystem update can be safer than hosts or storage controllers. This statement is true assuming that you follow the guideline of a 30-minute minimum interval between the modifications that you perform in one fabric to the next. Even if an unexpected error brings down your entire SAN fabric, the IBM FlashSystem environment continues working through the other fabric and your applications remain unaffected.

Because you are updating your IBM FlashSystem, also update your SAN switches code to the latest supported level. Start with your principal core switch or director, continue by updating the other core switches, and update the edge switches last. Update one entire fabric (all switches) before you move to the next one so that a problem you might encounter affects only the first fabric. Begin your other fabric update only after you verify that the first fabric update has no problems.

If you are not running symmetrical, redundant, independent SAN fabrics, fix this problem as a high priority because it represents a single point of failure.

10.5.5 Storage controllers preparation

As critical as with the attached hosts, the attached storage controllers must correctly handle the failover of MDisk paths. Therefore, they must be running supported microcode versions and their own SAN paths to IBM FlashSystem must be free of errors.

10.5.6 Hosts preparation

If the appropriate precautions are taken, the IBM FlashSystem update is not apparent to the attached servers and their applications. The automated update procedure updates one IBM FlashSystem node at a time, while the other node in the I/O group covers for its designated volumes.

However, to ensure that this feature works, the failover capability of your multipath software must be working correctly. This capability can be mitigated by enabling NPIV if your current code level supports this function. For more information about N_Port ID Virtualization (NPIV), see Chapter 8, “Configuring hosts” on page 367.

Before you start IBM FlashSystem update preparation, check the following items for every server that is attached to IBM FlashSystem that you update:

- The operating system type, version, and maintenance or fix level
- The make, model, and microcode version of the HBAs
- The multipath software type, version, and error log

For information about troubleshooting, see this IBM Documentation web page.

Fix every problem or “suspect” that you find with the disk path failover capability. Because a typical IBM FlashSystem environment can have hundreds of servers that are attached to it, a spreadsheet might help you with the Attached Hosts Preparation tracking process. If you have some host virtualization, such as VMware ESX, AIX Logical Partitions (LPARs), IBM Virtual I/O Server (VIOS), or Solaris containers in your environment, verify the redundancy and failover capability in these virtualization layers.

10.5.7 Copy services considerations

When you update an IBM FlashSystem family product that participates in an intercluster Copy Services relationship, do *not* update both clusters in the relationship simultaneously. This situation is not verified or monitored by the automatic update process and might lead to a loss of synchronization and unavailability.

You must successfully finish the update in one cluster before you start the next one. Try to update the next cluster as soon as possible to the same code level as the first one. Avoid running them with different code levels for extended periods.

10.5.8 Running the Upgrade Test Utility

It is a requirement that you install and run the latest IBM FlashSystem Upgrade Test Utility before you update the IBM FlashSystem software. For more information, see Software Upgrade Test Utility.

This tool verifies the health of your IBM FlashSystem storage array for the update process. It also checks for unfixed errors, degraded MDisks, inactive fabric connections, configuration conflicts, hardware compatibility, drive firmware, and many other issues that might otherwise require cross-checking a series of command outputs.

> **Note:** The Upgrade Test Utility does not log in to storage controllers or SAN switches. Instead, it reports the status of the connections of the IBM FlashSystem to these devices. It is the users' responsibility to check these components for internal errors.

You can use the management GUI or the CLI to install and run the Upgrade Test Utility.

Using the management GUI

To test the software on the system, complete these steps:

1. In the management GUI, select **Settings** → **System** → **Update System**.
2. Click **Test Only**.
3. Select the test utility that you downloaded from the Fix Central support site. Upload the Test utility file and enter the code level that you are planning to update to. Figure 10-5 shows the IBM FlashSystem management GUI window that is used to install and run the Upgrade Test Utility.

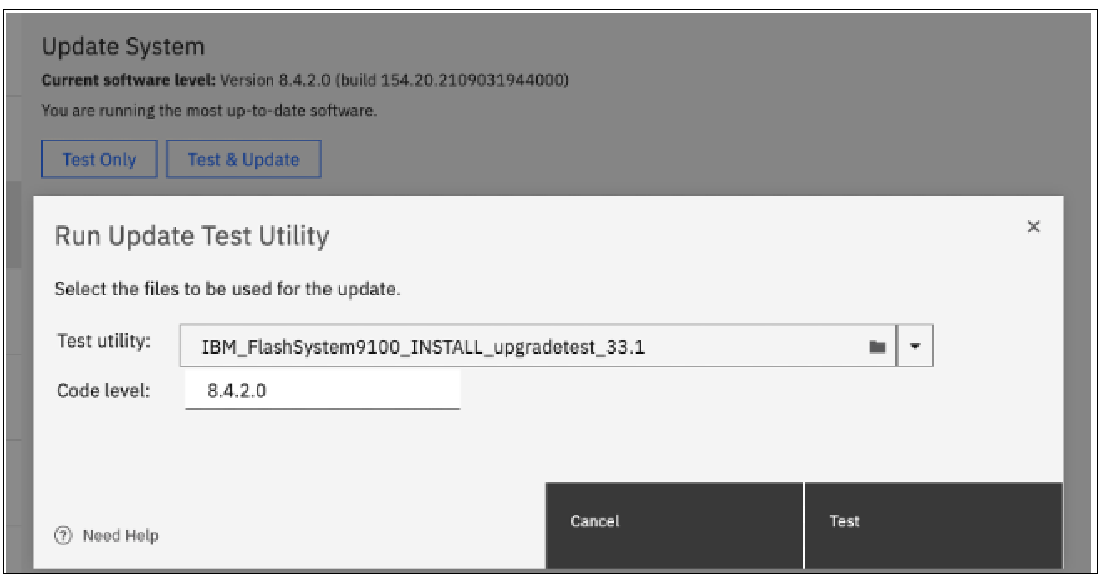


Figure 10-5 IBM FlashSystem Upgrade Test Utility using the GUI

4. Click **Test.** The test utility verifies that the system is ready to be updated. After the Update Test Utility completes, you are presented with the results. The results state that no warnings or problems were found, or directs you to more information about known issues that were discovered on the system.

Figure 10-6 shows a successful completion of the update test utility.

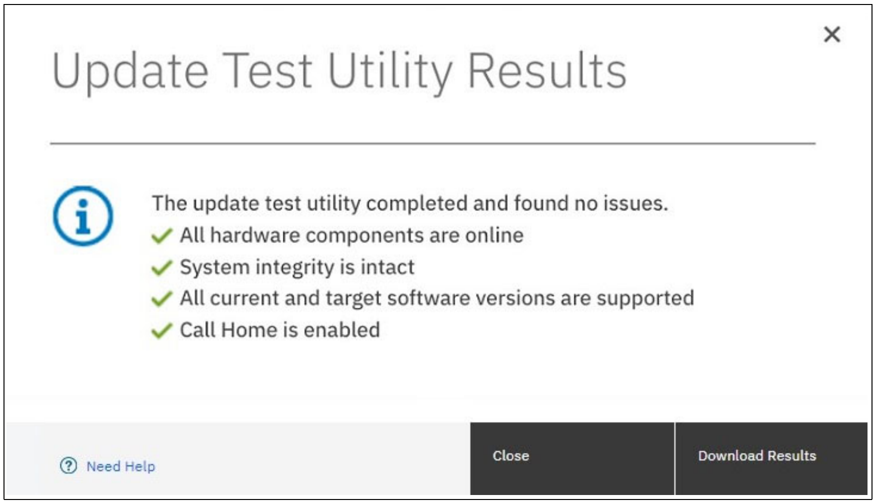


Figure 10-6 IBM FlashSystem Upgrade Test Utility completion panel

5. Click **Download Results** to save the results to a file.
6. Click **Close**.

Using the command line

To test the software on the system, complete these steps:

1. Using OpenSSH scp or PuTTY pscp, copy the software update file and the Software Update Test Utility package to the `/home/admin/upgrade` directory by using the management IP address of the IBM FlashSystem. Documentation and online help might refer to the `/home/admin/update` directory, which points to the same location on the system.

   An example for the IBM FlashSystem 9200 is shown in Example 10-2.

Example 10-2 Copying the upgrade test utility to IBM FlashSystem 9200

```
C:\>pscp -v -P 22 IBM_FlashSystem9200_INSTALL_upgradetest_33.1
superuser@9.10.11.12:/home/admin/upgrade
Looking up host "9.10.11.12" for SSH connection
Connecting to 9.10.11.12 port 22
We claim version: SSH-2.0-PuTTY_Release_0.74
Remote version: SSH-2.0-OpenSSH_8.0
Using SSH protocol version 2
No GSSAPI security context available
Doing ECDH key exchange with curve Curve25519 and hash SHA-256 (unaccelerated)
Server also has ssh-rsa host key, but we don't know it
Host key fingerprint is:
ecdsa-sha2-nistp521 521 a8:f0:de:cf:eb:fd:b4:74:9e:95:c7:bd:5c:f1:3b:b5
Initialised AES-256 SDCTR (AES-NI accelerated) outbound encryption
Initialised HMAC-SHA-256 (unaccelerated) outbound MAC algorithm
Initialised AES-256 SDCTR (AES-NI accelerated) inbound encryption
Initialised HMAC-SHA-256 (unaccelerated) inbound MAC algorithm
Using username "superuser".
Attempting keyboard-interactive authentication
Keyboard-interactive authentication prompts from server:
| Password:
```

```
End of keyboard-interactive prompts from server
Access granted
Opening main session channel
Opened main channel
Primary command failed; attempting fallback
Started a shell/command
Using SCP1
Connected to 9.10.11.12
Sending file IBM_FlashSystem9200_INSTALL_upgradetest_33.1, size=333865
Sink: C0644 333865 IBM_FlashSystem9100_INSTALL_upgradetest_33.1
IBM_FlashSystem9200_INSTA | 326 kB | 326.0 kB/s | ETA: 00:00:00 | 100%
Session sent command exit status 0
Main session channel closed
All channels closed
C:\>
```

2. Ensure that the update file was successfully copied as shown by the `exit status 0` return code or you can use the **`lsdumps -prefix /home/admin/upgrade`** command.
3. Install and run Upgrade Test Utility in the CLI, as shown in Example 10-3. In this case, the Upgrade Test Utility found no errors and completed successfully.

Example 10-3 Upgrade test using the CLI

```
IBM_FlashSystem:IBM Redbook FS:superuser>svctask applysoftware -file
IBM_FlashSystem9200_INSTALL_upgradetest_33.1

CMMVC9001I The package installed successfully.

IBM_FlashSystem:IBM Redbook FS:superuser>svcupgradetest -v 8.4.2.0

svcupgradetest version 33.1

Please wait, the test may take several minutes to complete.

Results of running svcupgradetest:
==================================

The tool has found 0 errors and 0 warnings.
The tool has not found any problems with the cluster.
```

Note: The return code for the **`applysoftware`** command always is 1, whether the installation succeeded or failed. However, the message that is returned when the command completes reports the correct installation result.

Review the output to check whether there have been any problems found by the utility. The output from the command either states that there have been no problems found, or directs you to details about known issues that have been discovered on the system.

10.5.9 Updating the software

FlashSystem software can be updated by using one of the following methods:

- GUI: During a standard update procedure in the management GUI, the system updates each of the nodes systematically. This is the recommended method for updating software on nodes.
- CLI: The command-line interface gives you more control over the automatic upgrade process. You have the ability to resolve multipathing issues when nodes go offline for updates. You can also override the default 30-minute mid-point delay, pause an update, and resume a stalled update.
- Manual: To provide even more flexibility in the update process, you can manually update each node individually using the Service Assistant Tool GUI. When upgrading the software manually, you remove a node from the system, update the software on the node, and return the node to the system. You repeat this process for the remaining nodes until the last node is removed from the system. At this point, the remaining nodes switch to running the new software. When the last node is returned to the system, it updates and runs the new level of software. This action cannot be performed on an active node. To update software manually, the nodes must either be candidate nodes (a candidate node is a node that is not in use by the system and cannot process I/O) or in a service state. During this procedure, every node must be updated to the same software level and the node becomes unavailable during the update.

Whichever method (GUI, CLI, or manual) that you choose to perform the update, make sure you adhere to the following guidelines for your IBM FlashSystem software update:

- Schedule the IBM FlashSystem software update for a low I/O activity time. The update process puts one node at a time offline. It also disables the write cache in the I/O group that node belongs to until both nodes are updated. Therefore, with lower I/O, you are less likely to notice performance degradation during the update.
- Never power off, reboot, or reset an IBM FlashSystem node during software update unless you are instructed to do so by IBM Support. Typically, if the update process encounters a problem and fails, it backs out. Bear in mind that the update process can take one hour per node with a further, optional, 30-minute mid-point delay.
- If you are planning for a major IBM FlashSystem version update, update your current version to its latest fix level *before* you run the major update.
- Check whether you are running a web browser type and version that is supported by the IBM FlashSystem target software level on every computer that you intend to use to manage your IBM FlashSystem.

This section describes the steps required to update the software.

Using the management GUI

To update the software on the system automatically, complete these steps:

1. In the management GUI, select **Settings** → **System** → **Update System**.
2. Click **Test & Update**.
3. Select the test utility and the software package that you downloaded from the Fix Central support site. The test utility verifies (again) that the system is ready to be updated.
4. Click **Next**. Select **Automatic update**.

5. Select whether you want to create intermittent pauses in the update to verify the process. Select one of the following options.
   - Fully automatic update without pauses (recommended)
   - Pausing the update after half of the nodes are updated
   - Pausing the update before each node updates
6. Click **Finish**. As the canisters on the system are updated, the management GUI displays the progress for each canister.
7. Monitor the update information in the management GUI to determine when the process is complete.

Using the command line

To update the software on the system automatically, complete these steps:

1. You must run the latest version of the test utility to verify that no issues exist with the current system. See Example 10-3 on page 463.
2. Copy the software package to the IBM FlashSystem using the same method as described in Example 10-2 on page 462.
3. Before you begin the update, you must be aware of the following situations:
   - The installation process fails under the following conditions:
     - If the software that is installed on the remote system is not compatible with the new software or if an inter-system communication error does not allow the system to check that the code is compatible.
     - If any node in the system has a hardware type that is not supported by the new software.
     - If the system determines that one or more volumes in the system would be taken offline by rebooting the nodes as part of the update process. You can find details about which volumes would be affected by using the **`lsdependentvdisks`** command. If you are prepared to lose access to data during the update, you can use the force flag to override this restriction.
   - The update is distributed to all the nodes in the system by using internal connections between the nodes.
   - Nodes are updated one at a time.
   - Nodes run the new software concurrently with normal system activity.
   - While the node is updated, it does not participate in I/O activity in the I/O group. As a result, all I/O activity for the volumes in the I/O group is directed to the other node in the I/O group by the host multipathing software.
   - There is a thirty-minute delay between node updates. The delay allows time for the host multipathing software to rediscover paths to the nodes that are updated. There is no loss of access when another node in the I/O group is updated.
   - The update is not committed until all nodes in the system are successfully updated to the new software level. If all nodes are successfully restarted with the new software level, the new level is committed. When the new level is committed, the system vital product data (VPD) is updated to reflect the new software level.
   - Wait until all member nodes are updated and the update is committed before you invoke the new functions of the updated software.

– Because the update process takes some time, the installation command completes as soon as the software level is verified by the system. To determine when the update is completed, you must either display the software level in the system VPD or look for the Software update complete event in the error/event log. If any node fails to restart with the new software level or fails at any other time during the process, the software level is backed off.

– During an update, the version number of each node is updated when the software is installed and the node is restarted. The system software version number is updated when the new software level is committed.

– When the update starts, an entry is made in the error or event log and another entry is made when the update completes or fails.

4. Issue the following CLI command to start the update process:

```
applysoftware -file <software_update_file>
```

where <software_update_file> is the filename of the software update file. If the system identifies any volumes that would go offline as a result of rebooting the nodes as part of the system update, the software update does not start. An optional force parameter can be used to indicate that the update continues regardless of the problem identified. If you use the force parameter, you are prompted to confirm that you want to continue.

5. Issue the following CLI command to check the status of the update process:

```
lsupdate
```

This command displays success when the update is complete.

6. To verify that the update has successfully completed, issue the **lsnodecanistervpd** command for each node in the system. The code_level field displays the new code level for each node.

10.6 Drive firmware updates

The updating of drive firmware is concurrent process that can be performed online while the drive is in use, whether it is NVMe or SCM drives in the control enclosure or the SSD drives in any SAS-attached expansion enclosures.

When used on an array member drive the update checks for volumes that are dependent on the drive and refuses to run if any are found. Drive dependent volumes are usually caused by non-redundant or degraded RAID arrays. Where possible you should restore redundancy to the system by replacing any failed drives before upgrading drive firmware. When this is not possible, you can either add redundancy to the volume by adding a second copy in another pool or use the **-force** parameter to bypass the dependent volume check. Use **-force** only if you are willing to accept the risk of data loss on dependent volumes (if the drive fails during the firmware update).

> **Note:** Due to some system constraints, it is not possible to produce a single NVMe firmware package that works on all NVMe drives on all Spectrum Virtualize code levels. Therefore, you find three different NVMe firmware files available for download depending on the size of the drives you have installed.

Using the management GUI

To update the drive firmware automatically, complete the following steps:

1. Select **Pools** → **Internal Storage** → **Actions** → **Upgrade All**.
2. As shown in Figure 10-7, in the Upgrade Package text box, browse to the drive firmware package you downloaded as described in “Obtaining the software packages” on page 457.

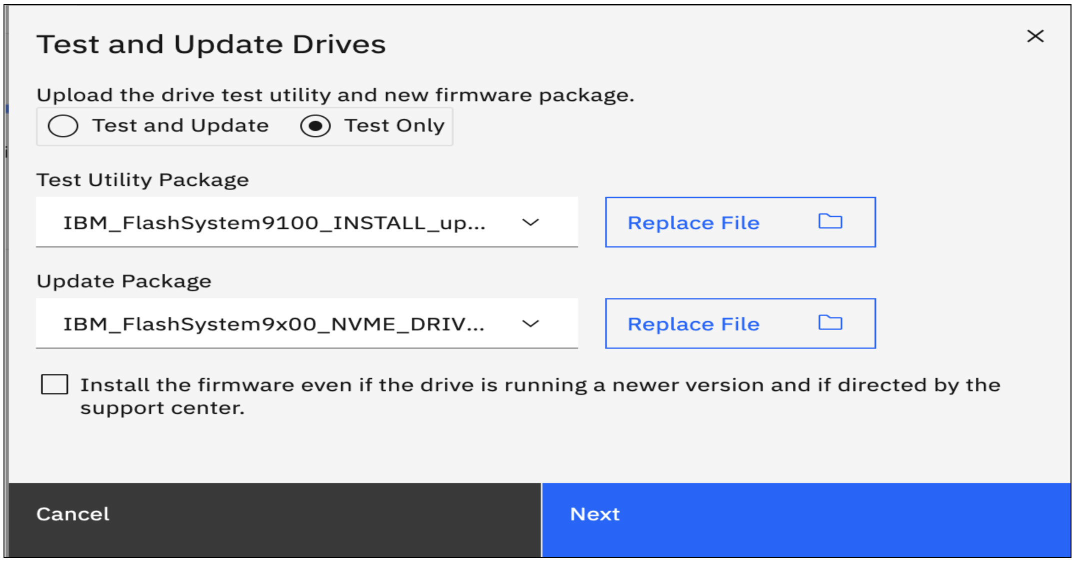


Figure 10-7 Drive firmware upgrade

3. Click **Upgrade.** Each drive upgrade takes approximately 6 minutes to complete.
4. You can also update individual drives by right-clicking a single drive and selecting **Upgrade**.
5. To monitor the progress of the upgrade, select **Monitoring** → **Background Tasks**.

Using the command line

To update the software on the system manually, complete these steps:

1. Copy the drive firmware package to the IBM FlashSystem by using the same method as described in Example 10-2 on page 462.
2. Issue the following CLI command to start the update process for all drives:

```
applydrivesoftware -file <software_update_file> -type firmware -all
```

   where <software_update_file> is the filename of the software update file. The use of the `-all` option updates firmware on all eligible drives including quorum drives, which is a slight risk. To avoid this risk, use the `-drive` option instead and make sure the quorum is moved using the `lsquorum` and `chquorum` commands in between `applydrivesoftware` invocations.

> **Note:** The maximum number of drive IDs that can be specified on a command line using the **-drive** option is 128. If you have more than 128 drives, use the **-all** option or run multiple invocations of applydrivesoftware to complete the update.

3. Issue the following CLI command to check the status of the update process:

```
lsdriveupgradeprogress
```

   This command displays success when the update is complete.

4. To verify that the update has successfully completed, issue the **lsdrive** command for each drive in the system. The firmware_level field displays the new code level for each drive. Example 10-4 demonstrates how to list the firmware level for four specific drives:

Example 10-4 List firmware level for drives 0,1, 2 and 3

```
IBM_FlashSystem:GLTL-FS9K:superuser>for i in 0 1 2 3; do echo "Drive $i = `lsdrive $i|grep firmware`"; done
Drive 0 = firmware_level 1_2_11
Drive 1 = firmware_level 1_2_11
Drive 2 = firmware_level 1_2_11
Drive 3 = firmware_level 1_2_11
```

For more information, see this IBM Documentation web page.

10.7 Remote Code Load

Remote Code Load (RCL) is a service offering provided by IBM, which allows code updates to be performed by remote support engineers, as opposed to an onsite Support Services Representative (SSR).

IBM Assist on-site (AOS) or remote support center or Secure Remote Access (SRA) including Call Home enablement are required to enable Remote Code Load. With the Assist on-site enabled, the live remote-assistance tool, a member of IBM support team can view your desktop and share control of your mouse and keyboard to get you on your way to a solution. Rather than the RCL the tool can also speed up problem determination, collection of data, and ultimately, your problem solution.

For more information about configuring support assistance, see this IBM Documentation web page.

Before the Assist On-site application is used, you can test your connectivity to the Assist On-site network by downloading the IBM connectivity testing tool. For more information, see this IBM Support web page.

To request the RCL for your system, see this IBM Support web page and select your product type. Then, complete the following steps:

1. At the IBM Remote Code Load web page, select **Product type** → **Book Now - FlashSystem 9200 and 7200 Remote Code Load**.

Figure 10-8 shows the RCL Schedule Service page.

Figure 10-8 FlashSystem RCL Schedule Service page

2. Click **Schedule Service** to start scheduling the service.
3. Next is the Product type selection for RCL. Go to FlashSystem 9200 option and click **Select** (see Figure 10-9).

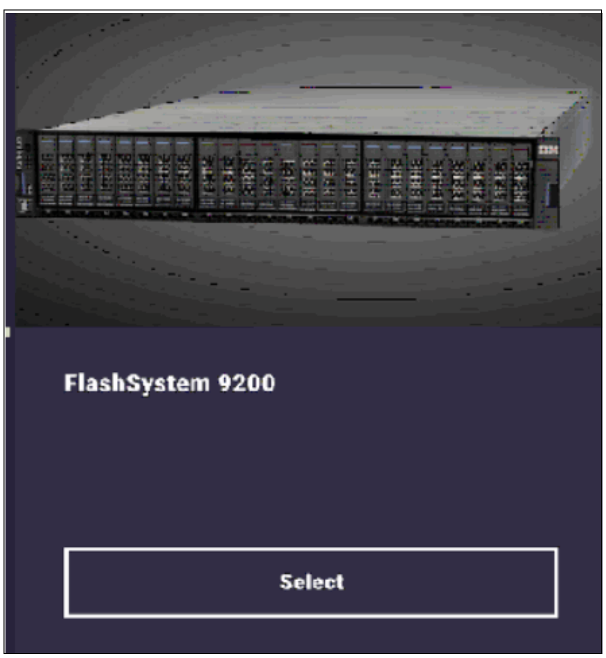


Figure 10-9 RCL Product type page

4. In the RCL time frame option, select the date (see Figure 10-10) and time frame (see Figure 10-11).

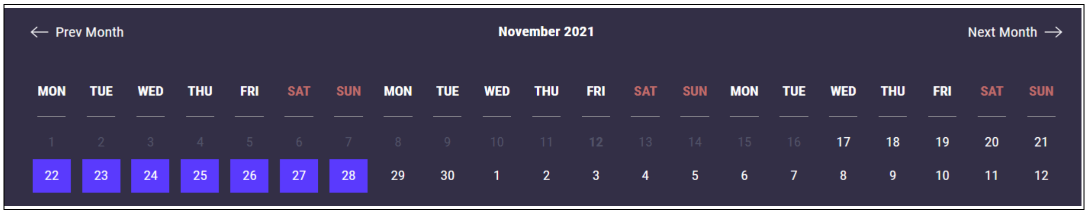


Figure 10-10 Time Frame selection page

Figure 10-11 RCL Time selection page

5. Enter your booking details in the RCL booking information form (see Figure 10-12).

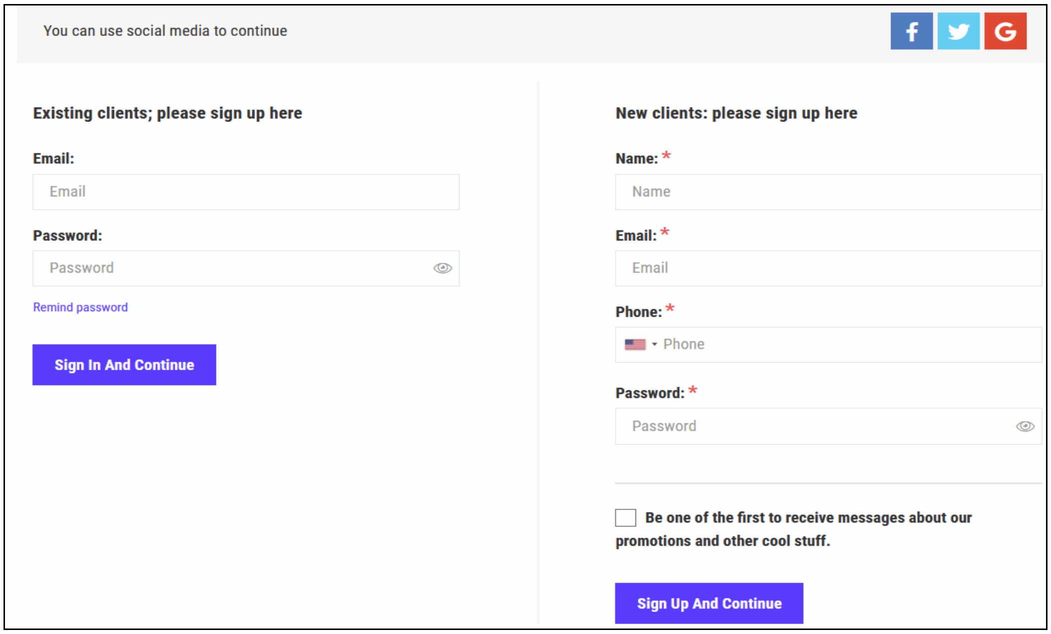


Figure 10-12 RCL Booking information page

10.8 Replacing Flash Core Module

Replacing a Flash Core Module (FCM) in your IBM FlashSystem requires special attention to avoid out-of-scope procedures that can damage your system. Before you start the IBM FlashSystem FCM replacement procedure, review the following items to prevent any damage to your system and FCM or to your system data:

- Do not replace, re-seat, or run any task on your FlashSystem if you are not sure or not comfortable with the procedure, Always engage IBM support Level 1/2 in case of issues or any problem during any procedure you are running.
- Call IBM Support to ensure that the logs were verified and a correct action plan was provided to replace the failed FCM.
- Confirm with the IBM System Service Representative (SSR) if the correct FRU number was shipped to be replaced and that the action plan that was provided by IBM Support was revised.
- Confirm that your system does not have any other FCM failure or error messages in the error log tab *before* conducting the replacement procedure.

If you have the IBM Expert Care coverage feature for your FlashSystem, make sure that your Technical Account Manager (TAM) is aware of the procedure and engaged with the service team to proceed with the replacement.

For more information, see this IBM Documentation web page.

Note: Re-seating an FCM can reformat the module in specific instances. All FCM drive failure alerts must be addressed before any re-seat or replacement procedure is done. Upon receiving any error message for the FCM drives, it is recommended to escalate the problem to IBM Support.

10.9 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 suitable precautions were not taken.

Human error can include some the following examples:

- Disrupting or disabling the working disk paths of a server while trying to fix failed ones.
- Disrupting a neighbor SAN switch port while inserting or removing an FC cable or small form-factor pluggable (SFP).
- Disabling or removing the working part in a redundant set instead of the failed one.
- Making modifications that affect both parts of a redundant set without an interval that allows for automatic failover during unexpected problems.

Adhere to the following guidelines to perform these actions with assurance:

- Uniquely and correctly identify the components of your SAN.
- Use the proper failover commands to disable only the failed parts.
- Understand which modifications are necessarily disruptive, and which can be performed online with little or no performance degradation.

10.9.1 Cross-referencing WWPN

With the WWPN of an HBA, you can uniquely identify one server in the SAN. If the name of the server is changed at the operating system level and not at the IBM FlashSystem host definitions, it continues to access its mapped volumes exactly because the WWPN of the HBA did not change.

Alternatively, if the HBA of a server is removed and installed in a second server and the SAN zones for the first server and the IBM FlashSystem host definitions are not updated, the second server can access volumes that it probably should not access.

Complete the following steps to cross-reference HBA WWPNs:

1. In your server, verify the WWPNs of the HBAs that are used for disk access. Typically, you can complete this task by using the SAN disk multipath software of your server.

   If you are using SDDPCM, run the **pcmpath query WWPN** command to see output similar to what is shown in Example 10-5.

   Example 10-5 Output of the pcmpath query WWPN command

   ```
   [root@Server127]> pcmpath query wwpn
   Adapter Name PortWWN
   fscsi0       10000090FA021A13
   fscsi1       10000090FA021A12
   ```

   If you are using server virtualization, verify the World Wide Port Names (WWPNs) in the server that is attached to the SAN, such as AIX Virtual Input/Output (VIO) or VMware ESX. Cross-reference with the output of the IBM FlashSystem **lshost <hostname>** command, as shown in Example 10-6 on page 473.

Example 10-6 Output of the lshost <hostname> command

```
IBM_FlashSystem:IBM Redbook FS:superuser>svcinfo lshost Server127
id 0
name Server127
port_count 2
type generic
mask 1111111111111111111111111111111111111111111111111111111111111111
iogrp_count 4
status active
site_id
site_name
host_cluster_id
host_cluster_name
protocol scsi
WWPN 10000090FA021A13
node_logged_in_count 1
state active
WWPN 10000090FA021A12
node_logged_in_count 1
state active
```

2. If necessary, cross-reference information with your SAN switches, as shown in Example 10-7. In Brocade switches use the **nodefind <WWPN>** command.

Example 10-7 Cross-referencing information with SAN switches

```
blg32sw1_B64:admin> nodefind 10:00:00:90:FA:02:1A:13
Local:
 Type Pid    COS     PortName                NodeName                 SCR
 N    401000;    2,3;10:00:00:90:FA:02:1A:13;20:00:00:90:FA:02:1A:13; 3
    Fabric Port Name: 20:10:00:05:1e:04:16:a9
    Permanent Port Name: 10:00:00:90:FA:02:1A:13
    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 Has 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 requests with the servers' WWPNs that are listed in the Description and Implementation sections.

10.9.2 Cross-referencing LUN ID

Always cross-reference the IBM FlashSystem `vdisk_UID` with the server logical unit number (LUN) ID before you perform any modifications that involve IBM FlashSystem volumes. Example 10-8 shows an AIX server that is running Subsystem Device Driver Path Control Module (SDDPCM). The IBM FlashSystem `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 10-8 Results of running the lshostvdiskmap command

```
IBM_FlashSystem:IBM Redbook FS:superuser>lshostvdiskmap NYBIXTDB03
id name       SCSI_id vdisk_id vdisk_name     vdisk_UID
0  NYBIXTDB03 0       0        NYBIXTDB03_T01 60050768018205E12000000000000000


root@nybixtdb03::/> pcmpath query device
Total Dual Active and Active/Asymmetric Devices : 1
DEV#:   4  DEVICE NAME: hdisk4  TYPE: 2145  ALGORITHM:  Load Balance
SERIAL: 60050768018205E12000000000000000
==========================================================================
Path#      Adapter/Path Name          State     Mode     Select     Errors
    0*          fscsi0/path0           OPEN   NORMAL          7          0
    1           fscsi0/path1           OPEN   NORMAL       5597          0
    2*          fscsi2/path2           OPEN   NORMAL          8          0
    3           fscsi2/path3           OPEN   NORMAL       5890          0
```

If your organization uses a change management tracking tool, include the `vdisk_UID` and LUN ID information in every change request that performs SAN storage allocation or reclaim.

> **Note:** Because a host can have many volumes with the same `scsi_id`, always cross-reference the IBM FlashSystem volume UID with the host volume UID and record the `scsi_id` and LUN ID of that volume.

10.10 Server HBA replacement

Replacing a failed HBA in a server is a fairly trivial and safe operation if it is performed correctly. However, more precautions are required if your server has multiple, redundant HBAs on different SAN fabrics and the server hardware permits you to “hot” replace it (with the server still running).

Complete the following steps to replace a failed HBA and retain the working HBA:

1. In your server, identify the failed HBA and record its WWPNs. (For more information, see 10.9.1, “Cross-referencing WWPN” on page 472.) Then, place this HBA and its associated paths offline (gracefully if possible). This approach is important so that the multipath software stops attempting to recover it. Your server might even show a degraded performance while you perform this task.
2. Some HBAs have an external label that shows the WWPNs. If you have this type of label, record the WWPNs before you install the new HBA in the server.

3. If your server does not support HBA hot-swap, power off your system, replace the HBA, connect the used FC cable into the new HBA, and power on the system.

   If your server does support hot-swap, follow the appropriate procedures to perform a "hot" replace of the HBA. Do *not* disable or disrupt the working HBA in the process.

4. Verify that the new HBA successfully logged in to the SAN switch. If it logged in successfully, you can see its WWPNs logged in to the SAN switch port. Otherwise, fix this issue before you continue to the next step.

   Cross-check the WWPNs that you see in the SAN switch with the one you noted in step 1, and make sure that you did not record the wrong WWNN.

5. In your SAN zoning configuration tool, replace the old HBA WWPNs for the new ones in every alias and zone to which they belong. Do *not* touch the other SAN fabric (the one with the working HBA) while you perform this task.

   Only one alias should use each WWPN, and zones must reference this alias.

   If you are using SAN port zoning (though you should not be) and you did not move the new HBA FC cable to another SAN switch port, you do not need to reconfigure zoning.

6. Verify that the new HBA's WWPNs appear in the IBM FlashSystem by using the **`lsfcportcandidate`** command.

   If the WWPNs of the new HBA do not appear, troubleshoot your SAN connections and zoning.

7. Add the WWPNs of this new HBA in the IBM FlashSystem host definition by using the **`addhostport`** command. It is important that you do not remove the old one yet. Run the **`lshost <servername>`** command. Then, verify that the working HBA shows as `active`, while the failed HBA should show as `inactive` or `offline`.

8. Use software to recognize the new HBA and its associated SAN disk paths. Certify that all SAN LUNs have redundant disk paths through the working HBA and the new HBA.

9. Return to the IBM FlashSystem and verify again (by using the **`lshost <servername>`** command) that both the working and the new HBA's WWPNs are active. In this case, you can remove the old HBA WWPNs from the host definition by using the **`rmhostport`** command.

10. Do not remove any HBA WWPNs from the host definition until you ensure that you have at least two active ones that are working correctly.

By following these steps, you avoid removing your only working HBA by mistake.

10.11 Hardware upgrades

The IBM FlashSystem scalability features allow significant flexibility in its configuration. As discussed in previous chapters, the IBM FlashSystem family has two different types of enclosures: control enclosures and expansion enclosures.

- Control Enclosures manage your storage systems, communicate with the host, and manage interfaces. In addition, they can also house up to 24 NVMe-capable flash drives.
- Expansion Enclosures increase the available capacity of an IBM FlashSystem cluster. They communicate with the control enclosure through a dual pair of 12 Gbps serial-attached SCSI (SAS) connections. These expansion enclosures can house many of flash (solid-state drive (SSD)) SAS type drives,

A basic configuration of an IBM FlashSystem storage platform consists of one IBM FlashSystem control enclosure. For a balanced increase of performance and scale, up to four (depending on model) IBM FlashSystem control enclosures can be clustered into a single storage system. Similarly, to increase capacity, up to two chains (depending on model) of expansion enclosures can be added per control enclosure. Therefore, several scenarios are possible for its growth.

These processes are described next.

10.11.1 Adding control enclosures

If your IBM FlashSystem cluster is below the maximum I/O groups limit for your specific product and you intend to upgrade it, you can install another control enclosure. It is also feasible that you might have a cluster of IBM Storwize V7000 nodes that you want to add the IBM FlashSystem enclosures to because the latter are more powerful than your existing ones. Therefore, your cluster can include different node models in different I/O groups.

To install these control enclosures, determine whether you need to upgrade your IBM FlashSystem first (or Storwize V7000 code level if you are merging an existing Storwize V7000 Gen2 cluster with a IBM FlashSystem 9200 for example).

For more information, see 10.5.2, “Hardware considerations” on page 458.

> **Note:** If exactly two control enclosures are in a system, you must set up a quorum disk or application outside of the system. If the two control enclosures lose communication with each other, the quorum disk prevents both I/O groups from going offline.

IBM FlashSystem 9200

To add a control enclosure to an existing FlashSystem 9200 system, the IBM SSR engineer must first install the new control enclosure in the rack and cable it to SAN or Ethernet switches or directly to the existing control enclosure. You are then able to add it to the system using the management GUI where it should automatically appear if cabled correctly. For more information, see this IBM Documentation web page.

IBM FlashSystem 9100

To add a control enclosure to an existing FlashSystem 9100 system, the IBM SSR engineer must first install the new control enclosure in the rack and cable it to SAN or Ethernet switches or directly to the existing control enclosure. You are then able to add it to the system using the management GUI where it should automatically appear if cabled correctly. For more information, see this IBM Documentation web page.

IBM FlashSystem 7200

To add a control enclosure to an existing FlashSystem 7200 system, you must first install it in the rack. Then, you must connect it to the system through a zone in the SAN or by using RDMA over Ethernet. Finally, you can add it to the system using the management GUI where it should automatically appear if cabled correctly. For more information, see this IBM Documentation web page.

IBM FlashSystem 5x00

To add a control enclosure to an existing FlashSystem 5100 system, you must first install it in the rack. Then, you must connect it to the system through a zone in the SAN or by using RDMA over Ethernet. Finally, you can add it to the system using the management GUI where it should automatically appear if cabled correctly. For more information, see this IBM Documentation web page.

After you install the new 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). Although this process can be done without stopping the host, careful planning and preparation is advised. For more information about supported operating systems, see this IBM Documentation web page.

  Note: You cannot move a volume that is in a type of Remote Copy relationship.

- If each of your servers is zoned to only one I/O group, modify your SAN zoning configuration as you move its volumes to another I/O group. As best you can, balance the distribution of your servers across I/O groups according to I/O workload.
- Use the **-iogrp** parameter in the **mkhost** command to define which I/O groups of the IBM FlashSystem that the new servers use. Otherwise, IBM FlashSystem maps by default the host to all I/O groups, even if they do not exist and regardless of your zoning configuration.

  Example 10-9 shows this scenario and how to resolve it by using the **rmhostiogrp** and **addhostiogrp** commands.

Example 10-9 Mapping the host to I/O groups

```
IBM_FlashSystem:IBM Redbook FS:superuser>lshost
id name        port_count iogrp_count status site_id site_name host_cluster_id
host_cluster_name protocol
0  Win2012srv1 2          4           online
scsi
1  linuxsrv3   1          4           online
scsi

IBM_FlashSystem:IBM Redbook FS:superuser>lshost Win2012srv1
id 0
name Win2012srv1
port_count 2
type generic
mask 1111111111111111111111111111111111111111111111111111111111111111
iogrp_count 4
status online
site_id
site_name
host_cluster_id
```

```
host_cluster_name
protocol scsi
WWPN 10000090FAB386A3
node_logged_in_count 2
state inactive
WWPN 10000090FAB386A2
node_logged_in_count 2
state inactive

IBM_FlashSystem:IBM Redbook FS:superuser>lsiogrp
id name           node_count vdisk_count host_count site_id site_name
0  io_grp0        2          11          2
1  io_grp1        0          0           2
2  io_grp2        0          0           2
3  io_grp3        0          0           2
4  recovery_io_grp 0         0           0

?IBM_FlashSystem:IBM Redbook FS:superuser>lshostiogrp Win2012srv1
id name
0  io_grp0
1  io_grp1
2  io_grp2
3  io_grp3

IBM_FlashSystem:IBM Redbook FS:superuser>rmhostiogrp -iogrp 3 Win2012srv1
IBM_FlashSystem:IBM Redbook FS:superuser>

IBM_FlashSystem:IBM Redbook FS:superuser>lshostiogrp Win2012srv1
id name
0  io_grp0
1  io_grp1
2  io_grp2
IBM_FlashSystem:IBM Redbook FS:superuser>

IBM_FlashSystem:IBM Redbook FS:superuser>addhostiogrp -iogrp io_grp3
Win2012srv1
IBM_FlashSystem:IBM Redbook FS:superuser>

IBM_FlashSystem:IBM Redbook FS:superuser>lshostiogrp Win2012srv1
id name
0  io_grp0
1  io_grp1
2  io_grp2
3  io_grp3

IBM_FlashSystem:IBM Redbook FS:superuser>lsiogrp
id name           node_count vdisk_count host_count site_id site_name
0  io_grp0        2          11          2
1  io_grp1        0          0           2
2  io_grp2        0          0           2
3  io_grp3        0          0           2
4  recovery_io_grp 0         0           0
```

- If possible, avoid setting a server to use volumes from different I/O groups that have different node types for extended periods of time. Otherwise, as this server's storage capacity grows, you might experience a performance difference between volumes from different I/O groups. This mismatch makes it difficult to identify and resolve eventual performance problems.

10.11.2 Upgrading nodes in an existing cluster

If you want to upgrade the nodes or canisters of your existing IBM FlashSystem, there is the option to increase the cache memory size or the adapters in each node. This can be done, one node at a time, and so as to be nondisruptive to the systems operations. For more information, see this IBM Documentation web page.

When evaluating cache memory upgrades, consider the following points:

- As your working set and total capacity increases, you should consider increasing your cache memory size. A *working set* is the most accessed workloads, excluding snapshots and backups. *Total capacity* implies more or larger workloads and a larger working set.
- If you are consolidating from multiple controllers, consider at least matching the amount of cache memory across those controllers.
- When externally virtualizing controllers (such as switched virtual circuit), a large cache can accelerate older controllers with smaller caches.
- If you are using DRP, then maximize the cache size and consider adding SCM drives with Easy Tier for the best performance.
- If you are making heavy use of copy services, consider increasing the cache beyond just your working set requirements.
- A truly random working set might not benefit greatly from the cache.

Important: Do not power on a node that is shown as offline in the management GUI, if you powered off the node to add memory to increase total memory. Before you increase memory, you must remove a node from the system so that it is not showing in the management GUI or in the output from the `svcinfo lsnode` command.

Do not power on a node that is still in the system and showing as offline with more memory than the node had when it powered off. Such a node can cause an immediate outage or an outage when you update the system software.

When evaluating adapter upgrades, consider the following points:

- A single 32 Gb Fibre Channel port can deliver over 3 GBps (allowing for overheads).
- A 32 Gb FC card in each canister, with 8 ports can deliver more than 24 GBps.
- An FCM NVMe device can perform at over 1 GBps.
- A single 32 Gb Fibre Channel port can deliver 80,000 to 125,000 IOPS with a 4k block size.
- A 32 Gb FC card in each canister, with 8 ports can deliver up to 1,000,000 IOPS.
- A FlashSystem 9200 can deliver 1,200,000 4k read miss input/output operations per second (IOPS) and up to 4,500,000 4k read hit IOPS.
- If you have more than 12 NVMe devices, consider 2 Fibre Channel cards per canister. A third Fibre Channel card allows you to achieve up to 45 GBps.

- If you want to achieve more than 800,000 IOPS, use at least 2 Fibre Channel cards per canister.
- If the FlashSystem is performing Remote Copy or clustering, consider using separate ports to ensure that there is no conflict with host traffic.
- iSER by way of 25 GbE ports has similar capabilities as 16 Gb FC ports, but with less overall ports available. If you are planning on using 10 Gb iSCSI, ensure it can service your expected workloads.

Real-time performance statistics are available in the management GUI from the **Monitoring** → **Performance** menu, as shown in Figure 10-13.

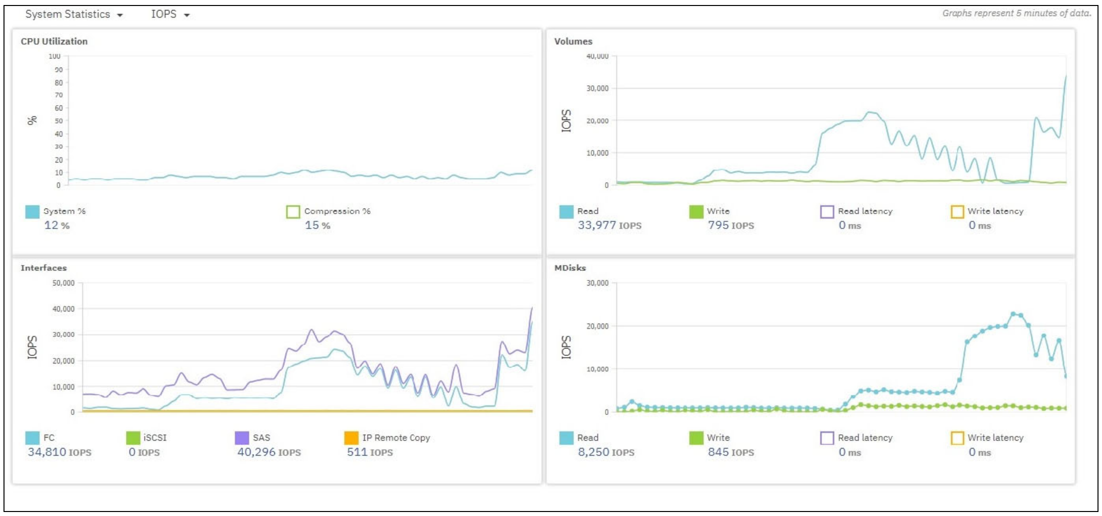


Figure 10-13 IBM FlashSystem performance statistics (IOPS)

Memory options for an IBM FlashSystem 9200 control enclosure

A CPU processor has six memory channels, which are labeled A-F. Each memory channel has two dual inline memory module (DIMM) slots, numbered 0-1. For example, DIMM slots A0 and A1 are in memory channel A.

On the system board, the DIMM slots are labeled according to their memory channel and slot. They are associated with the CPU nearest to their DIMM slots. You can install three distinct memory configurations in those 24 DIMM slots in each node canister.

Important: The memory in both node canisters must be configured identically to create the total enclosure memory size

Table 10-2 shows the available memory configuration for each FlashSystem 9200 control enclosure. Each column gives the valid configuration for each total enclosure memory size. DIMM slots are listed in the same order that they appear in the node canister.

To ensure proper cooling and a steady flow of air from the fan modules in each node canister, blank DIMMs must be inserted in any slot that does not contain a memory module.

Table 10-2 Available memory configuration for one node in a control enclosure

DIMM Slot	Total enclosure memory 256 GB	Total enclosure memory 768 GB	Total enclosure memory 1536 GB
F0 (CPU0)	Blank	32 GB	32 GB
F1 (CPU0)	Blank	Blank	32 GB
E0 (CPU0)	Blank	32 GB	32 GB
E1 (CPU0)	Blank	Blank	32 GB
D0 (CPU0)	32 GB	32 GB	32 GB
D1 (CPU0)	Blank	Blank	32 GB
A1 (CPU0)	Blank	Blank	32 GB
A0 (CPU0)	32 GB	32 GB	32 GB
B1 (CPU0)	Blank	Blank	32 GB
B0 (CPU0)	Blank	32 GB	32 GB
C1 (CPU0)	Blank	Blank	32 GB
C0 (CPU0)	Blank	32 GB	32 GB
C0 (CPU1)	Blank	32 GB	32 GB
C1 (CPU1)	Blank	Blank	32 GB
B0 (CPU1)	Blank	32 GB	32 GB
B1 (CPU1)	Blank	Blank	32 GB
A0 (CPU1)	32 GB	32 GB	32 GB
A1 (CPU1)	Blank	Blank	32 GB
D1 (CPU1)	Blank	Blank	32 GB
D0 (CPU1)	32 GB	32 GB	32 GB
E1 (CPU1)	Blank	Blank	32 GB
E0 (CPU1)	Blank	32 GB	32 GB
F1 (CPU1)	Blank	Blank	32 GB
F0 (CPU1)	Blank	32 GB	32 GB

Memory options for an IBM FlashSystem 9100 control enclosure

Each of the six memory channels in each CPU has two DIMM slots, for a total of 12 DIMM slots per CPU, which means 24 DIMM slots per node canister and 48 DIMM slots per enclosure. You can install six distinct memory configurations in those 24 DIMM slots in each node canister. (Each canister must have the same amount of memory and the same configuration).

Initially, each control enclosure ships with one of the following features, depending on what has been ordered, as shown in Table 10-3.

Table 10-3 Base memory features

Feature	Memory per enclosure	Maximum per enclosure
ACG0	128 GB base cache memory (eight 16 GB DIMMs - 2 per CPU)	1
ACG1	768 GB base cache memory (twenty-four 32 GB DIMMs - 6 per CPU)	1

You can order the features that are listed in Table 10-4 to upgrade to more memory at any time.

Table 10-4 Memory features

Feature	Memory per enclosure	Maximum per enclosure
ACGA	128 GB memory upgrade (eight 16 GB DIMMs)	3
ACGB	768 GB memory upgrade (twenty-four 32 GB DIMMs)	2

Memory options for an IBM FlashSystem 7200 control enclosure

Table 10-5 lists the various memory options available for the IBM FlashSystem 7200 by feature code.

Table 10-5 IBM FlashSystem 7200 memory options

Base memory (GB)	Field Upgrade ACGJ (GB)	Field Upgrade ACGB (GB)	Total memory (GB)
256	N/A	N/A	256
256	512	N/A	768
256	512	768	1536

Memory options for an IBM FlashSystem 5000 control enclosure

The IBM FlashSystem 5000 family consists of different models, and each model type provides a different set of features. Table 10-6 lists the memory features of the FlashSystem 5000 and 5100 models.

Table 10-6 Memory options

Platform	FS5010	FS5015	FS5030	FS5035	FS5100
Option 1 per node	1 x 8 GB	1 x 16 GB	1 x 16 GB	1 x 16 GB	2 x 16 GB
Option 2 per node	1 x 16 GB	2 x 16 GB	2 x 16 GB	2 x 16 GB	6 x 16 GB
Option 3 per node	2 x 16 GB	N/A	N/A	N/A	6 x 16 GB + 6 x 32 GB
Maximum per I/O Group	64 GB	64 GB	64 GB	64 GB	576 GB

Memory options for an IBM FlashSystem 5200 control enclosure

Table 10-7 lists the FlashSystem 5200 memory options.

Table 10-7 FS 5200 memory options

Feature code	Existing memory size per canister (GB)	Memory upgrade (GB)	Total memory per enclosure (GB)	Comments
ALG0	32 GB	N/A	64 GB	Ships two 32 GB DIMMs (DDR4) that are installed with the base system
ALG1	256 GB	N/A	512 GB	Ships eight 64 GB memory DIMMS (DDR4) that are installed with the base system
ALGC	32 GB	192 GB Cache Upgrade	256 GB	Ships six 32 GB memory DIMMS to add to the system
ALGD	32 GB	512 GB Cache Upgrade	512 GB	Ships eight 64 GB memory DIMMS to replace all existing 32 GB DIMMS

Adapter options for an IBM FlashSystem 9200 control enclosure

You can also add new adapters to the IBM FlashSystem 9200 nodes. These adapters are added as a pair (one card in each node). Six PCIe slots are available for port expansions in the IBM FlashSystem 9200 control enclosure. Each canister has three PCIe adapter slots and both canisters must have the same configuration. The PCIe adapter feature codes offer a pair of adapters to ensure that they are supplied symmetrically in each canister.

The control enclosure can be configured with three I/O adapter features to provide up to 24 16 Gb FC ports or up to 12 25 Gb Ethernet (iSCSI or iSCSI Extensions for Remote Direct Memory Access (RDMA) (iSER) capable) ports. The control enclosure also includes eight 10 Gb Ethernet ports as standard for iSCSI connectivity and two 1 Gb Ethernet ports for system management. A feature code also is available to include the SAS Expansion card if the user wants to use optional expansion enclosures. The options for the features available are shown in Table 10-8.

Table 10-8 IBM FlashSystem 9200 control enclosure adapter options

Number of control enclosures	16 Gbps/ 32 Gbps FC	On-board ISCSI	25 Gbps iSCSI (RoCE)	25 Gbps iSCSI (iWARP)
1	24	8	12	12
2	48	16	24	24
3	72	24	36	36
4	96	32	48	48

For more information about the feature codes, memory options, and functions of each adapter, see *IBM FlashSystem 9200 Product Guide*, REDP-5586.

Adapter options for an IBM FlashSystem 9100 control enclosure

You can also add new adapters to the IBM FlashSystem 9100 nodes. These adapters are added as a pair (one card in each node) and the options for the features available are shown in Table 10-9.

Table 10-9 IBM FlashSystem 9100 control enclosure adapter options

Number of cards	Ports	Protocol	Possible slots	Comments
0 - 3	4	16 Gb Fibre Channel	1, 2, 3	
0 - 3	2	25 Gb Ethernet (iWarp)	1, 2, 3	
0 - 3	2	25 Gb Ethernet (RoCE)	1, 2, 3	
0 - 1	2 - see comment	12 Gb SAS Expansion	1, 2, 3	Card is 4 port with only 2 ports active (ports 1 and 3)

For more information about the feature codes, memory options, and functions of each adapter ca, see Planning chapter of *IBM FlashSystem 9100 Architecture, Performance, and Implementation*, SG24-8425.

Adapter options for an IBM FlashSystem 7200 control enclosure

Six PCIe slots are available for port expansions in the IBM FlashSystem 7200 control enclosure. Each canister has three PCIe adapter slots and both canisters must have the same configuration. The PCIe adapter feature codes offer a pair of adapters to ensure that they are supplied symmetrically in each canister.

The IBM FlashSystem 7200 control enclosure can be configured with three I/O adapter features to provide up to twenty-four 16 Gb FC ports or up to twelve 25 Gb Ethernet (iSCSI or iSER-capable) ports. The control enclosure also includes eight 10 Gb Ethernet ports as standard for iSCSI connectivity and two 1 Gb Ethernet ports for system management. A feature code also is available to include the SAS Expansion adapter if the user wants to implement the optional expansion enclosures.

Adapter options for an IBM FlashSystem 5x00 control enclosure

All of the FlashSystem 5000 control enclosures include 1 Gb Ethernet (GbE) or 10 GbE ports as standard for iSCSI connectivity. The standard connectivity can be extended with extra ports or enhanced with more connectivity through an optional I/O adapter feature.
Table 10-10 lists which configurations are available for the FlashSystem 5000 and 5100.

Table 10-10 IBM FlashSystem 5000 family configurations

Platform	FS5010	FS5030	FS5100
iSCSI	1 x 1 GbE tech port + iSCSI 1 x 1 GbE iSCSI only	1 x 1 GbE dedicated tech port	1 x 1 GbE dedicated tech port
iSCSI	N/A	2 x 10 GbE (iSCSI only)	4 x 10 GbE (iSCSI only)
SAS	1 x 12 Gb SAS expansion	2 x 12 Gb SAS expansion	N/A

Table 10-11 lists the possible adapter installation for the FlashSystem 5000 and 5100. Only one interface card can be installed per canister and the interface card must be the same in both canisters.

Table 10-11 IBM FlashSystem 5000 family adapters

Platform	FS5010	FS5030	FS5100
FC	4-port 16 Gb Fibre Channel or	4-port 16 Gb Fibre Channel or	4-port 16 Gb Fibre Channel, FC NVMeoF or
iSCSI	4-port 10 GbE iSCSI or	4-port 10 GbE iSCSI or	2-port 25 GbE ROCE ISER, iSCSI or
iSCSI	2-port 25 GbE iSCSI or	2-port 25 GbE iSCSI or	2-port 25 GbE iWARP ISER, iSCSI and
SAS	4-port 12 Gb SAS host attach	4-port 12 Gb SAS host attach	2-port 12 Gb SAS to allow SAS expansions

IBM FlashSystem 5015 and 5035 control enclosures include 1 Gb Ethernet (GbE) or 10 GbE ports as standard for iSCSI connectivity. The standard connectivity can be extended by using more ports or enhanced with more connectivity through an optional I/O adapter feature. For more information, see this web page.

IBM FlashSystem 5015 control enclosure models 2N2 and 2N4

IBM FlashSystem 5015 control enclosure models 2N2 and 2N4 include the following features:

- Two node canisters, each with a two-core processor.
- 32 GB cache (16 GB per canister) with optional 64 GB cache (32 GB per canister).
- 1 Gb iSCSI connectivity standard with optional 16 Gb FC, 12 Gb SAS, 10 Gb iSCSI (optical), 25 Gb iSCSI (optical), and 1 Gb iSCSI connectivity.
- 12 Gb SAS ports for expansion enclosure attachment.
- Twelve slots for 3.5-inch LFF SAS drives (Model 2N2) and 24 slots for 2.5-inch SFF SAS drives (Model 2N4).
- 2U, 19-inch rack mount enclosure with 100 - 240 V AC or -48 V DC power supplies.

IBM FlashSystem 5035 control enclosure models 3N2 and 3N4

IBM FlashSystem 5035 control enclosure models 3N2 and 3N4 include the following features:

- Two node canisters, each with a six-core processor.
- 32 GB cache (16 GB per canister) with optional 64 GB cache (32 GB per canister).
- 10 Gb iSCSI (copper) connectivity standard with optional 16 Gb FC, 12 Gb SAS, 10 Gb iSCSI (optical), and 25 Gb iSCSI (optical) connectivity.
- 12 Gb SAS ports for expansion enclosure attachment.
- Twelve slots for 3.5-inch LFF SAS drives (Model 3N2) and 24 slots for 2.5-inch SFF SAS drives (Model 3N4).
- 2U, 19-inch rack mount enclosure with 100 - 240 V AC or -48 V DC power supplies.

IBM FlashSystem 5200 Host I/O connectivity and expansion enclosure adapters

Table 10-12 lists the maximum host count per control enclosure.

Table 10-12 Maximum host count per control enclosure

Number of control enclosures	16 Gb FC	32 Gb FC	On-board iSCSI	10 Gb Ethernet	25 Gb iSCSI
One enclosure	16	8	4	16	8
Two enclosures	32	16	8	32	16
Three enclosures	48	24	12	48	24
Four enclosures	64	32	16	64	32

IBM FlashSystem 5200 supported expansion enclosure and interface components

Table 10-13 lists the supported expansion enclosure MTMs 4662-6H2 and 4662-UH6.

Table 10-13 Supported expansion enclosure MTMs 4662-6H2 and 4662-UH6

Item	Feature code	Description	Ports
16 Gb FC 4 Port Adapter Pair	#ALBJ	This feature provides two I/O adapters. It is used to add 16 Gb FC connectivity.	Each adapter has four 16 Gb FC ports and shortwave SFP transceivers.
32 Gb FC 4 Port Adapter Pair	#ALBK	This feature provides two I/O adapters. It is used to add 32 Gb FC connectivity.	Each adapter has two 32 Gb FC ports and shortwave SFP transceivers.
10 Gb Ethernet Adapter Pair	#ALBL	This feature provides two I/O adapters. It is used to add 10 Gb Ethernet connectivity.	Each adapter has four 10 Gb Ethernet ports and SFP+ transceivers.
25 GbE (RoCE) Adapter Pair	#ALBM	This feature provides two I/O adapters. It is used to add 25 Gb Ethernet connectivity. Supports RoCE V2.	Each adapter has two 25 Gb Ethernet ports and SFP28 transceivers.
25 GbE (iWARP) Adapter Pair	#ALBN	This feature provides two I/O adapters. It is used to add 25 Gb Ethernet connectivity. Supports RDMA with iWARP.	Each adapter has two 25 Gb Ethernet ports and SFP28 transceivers.
12 Gb SAS expansion enclosure Attach Card (Pair)	#ALBP	This feature provides two 4-port 12 Gb SAS expansion enclosure attachment adapters. This feature is used to attach up to 20 expansion enclosures.	Each adapter only has two active SAS ports per card.
12 Gb SAS Host Adapters (Pair)	#ALBQ	This feature provides two 4-port 12 Gb SAS Host attachment adapters.	Two cards with 4 ports and mini-SAS HD connectors for host attachment.
16 Gb FC LW SFP Transceivers (Pair)	#ACHU	This feature provides two 16 Gb longwave SFP transceivers for use with 16 Gb FC I/O ports.	#ALBJ is a prerequisite. The maximum that is allowed is four for each instance of #ALBJ.
32 Gb FC LW SFP Transceivers (Pair)	#ACHV	This feature provides two 32 Gb longwave SFP transceivers for use with 32 Gb FC I/O ports.	#ALBK is a prerequisite. The maximum that is allowed is four for each instance of #ALBK.

10.11.3 Upgrading NVMe drives

To provide ultra-low latency for performance sensitive but less cache-friendly workloads, storage-class memory (SCM) drives from Intel and Samsung are available as a persistent storage tier for IBM FlashSystem family. SCM is a substantial step forward in memory technology, offering nonvolatile, ultra low latency memory for a fraction of the cost of traditional memory chips.

IBM FlashSystem products support SCM drives over NVMe to improve overall storage performance, or offer a higher performance storage pool. This means SCM drives can be used for small workloads that need exceptional levels of performance at the lowest latencies, or they can be combined with other NVMe drives using Easy Tier to accelerate much larger workloads. Like the FlashCore Modules, SCM drives are also available as upgrades for the previous generation of all flash arrays.

Spectrum Virtualize V8.4 supports up to 12 SCM drives in a control enclosure for IBM FlashSystem 9000, 7000, 5000, 5100 and 5200 families.

For more information, see the following resources:

- IBM FlashSystem 8.4.2 Documentation 9x00 - Updating drive
- IBM FlashSystem 8.4..2 Documentation 5x00 - Updating Drive Firmware

10.11.4 Moving to a new IBM FlashSystem cluster

You might have a highly populated, intensively used IBM FlashSystem cluster that you want to upgrade. You might also want to use the opportunity to refresh your IBM FlashSystem and SAN storage environment.

Complete the following steps to replace your cluster entirely with a newer, more powerful one:

1. Install your new IBM FlashSystem cluster.
2. Create a replica of your data in your new cluster.
3. Migrate your servers to the new IBM FlashSystem cluster when convenient.

If your servers can tolerate a brief, scheduled outage to switch from one IBM FlashSystem cluster to another, you can use the IBM FlashSystem Remote Copy services (Metro Mirror or Global Mirror) to create your data replicas, following these steps:

1. Select a host that you want to move to the new IBM FlashSystem cluster and find all the old volumes you must move.
2. Zone your host to the new IBM FlashSystem cluster.
3. Create Remote Copy relationships from the old volumes in the old IBM FlashSystem cluster to new volumes in the new IBM FlashSystem cluster.
4. Map the new volumes from the new IBM FlashSystem cluster to the host.
5. Discover new volumes on the host.
6. Stop all I/O from the host to the old volumes from the old IBM FlashSystem cluster.
7. Disconnect and remove the old volumes on the host from the old IBM FlashSystem cluster.
8. Unmap the old volumes from the old IBM FlashSystem cluster to the host.
9. Make sure Remote Copy relationships between old and new volumes in the old and new IBM FlashSystem cluster are synced.
10. Stop and remove Remote Copy relationships between old and new volumes so that the target volumes in the new IBM FlashSystem cluster receive read/write access.
11. Import data from the new volumes and start your applications on the host.

If you must migrate a server online, instead, you must use host-based mirroring by completing the following steps:

1. Select a host that you want to move to the new IBM FlashSystem cluster and find all the old volumes that you must move.
2. Zone your host to the new IBM FlashSystem cluster.
3. Create volumes in the new IBM FlashSystem cluster of the same size as the old volumes in the old IBM FlashSystem cluster.
4. Map the new volumes from the new IBM FlashSystem cluster to the host.
5. Discover new volumes on the host.
6. For each old volume, use host-based mirroring (such as AIX **`mirrorvg`**) to move your data to the corresponding new volume.
7. For each old volume, after the mirroring is complete, remove the old volume from the mirroring group.
8. Disconnect and remove the old volumes on the host from the old IBM FlashSystem cluster.
9. Unmap the old volumes from the old IBM FlashSystem cluster to the host.

This approach uses the server's computing resources (CPU, memory, and I/O) to replicate the data. It can be done online if properly planned. Before you begin, make sure it has enough spare resources.

The biggest benefit to using either approach is that they easily accommodate (if necessary) the replacement of your SAN switches or your back-end storage controllers. You can upgrade the capacity of your back-end storage controllers or replace them entirely, as you can replace your SAN switches with bigger or faster ones. However, you do need to have spare resources, such as floor space, power, cables, and storage capacity, available during the migration.

10.11.5 Splitting an IBM FlashSystem cluster

Splitting an IBM FlashSystem cluster might become a necessity if you have one or more of the following requirements:

- To grow the environment beyond the maximum number of I/O groups that a clustered system can support.
- To grow the environment beyond the maximum number of attachable subsystem storage controllers.
- To grow the environment beyond any other maximum system limit.
- To achieve new levels of data redundancy and availability.

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

After safely removing enclosures from the existing cluster and creating a second IBM FlashSystem cluster, choose from the following approaches to balance the two systems:

- Attach new storage subsystems and hosts to the new system and start adding only new workload on the new system.
- Migrate the workload onto the new system by using the approach described in 10.11.4, "Moving to a new IBM FlashSystem cluster" on page 488.

10.11.6 Adding expansion enclosures

As time passes and your environment grows, you must add storage to your system. Depending on the IBM FlashSystem family product and the code level that you installed, you can add different numbers of expansion enclosures to your system. Before you add an enclosure to a system, check that the licensed functions of the system support the extra enclosure.

Because all IBM FlashSystem models were designed to make managing and maintaining them as simple as possible, adding an expansion enclosure is an easy task. For more information, see this IBM Documentation web page.

IBM FlashSystem 9200

Currently, IBM offers the following SAS expansion enclosures that can be attached to the IBM FlashSystem 9200. Each node can support 10 SAS Connections thus a control enclosure can support up to 20 expansion enclosures.

> **Note:** To support SAS expansion enclosures, an AHBA - SAS Enclosure Attach adapter must be installed in each node canister of the IBM FlashSystem 9200 control enclosure.

The following types of expansion enclosures are available:

- IBM FlashSystem 9000 LFF Expansion Enclosure Model A9F
- IBM FlashSystem 9000 SFF Expansion Enclosure Model AFF

The new IBM FlashSystem 9200 SFF expansion enclosure Model AFF offers new tiering options with solid-state drive (SSD flash drives). Up to 480 drives of serial-attached SCSI (SAS) expansions are supported per IBM FlashSystem 9200 control enclosure. The expansion enclosure is 2U high.

The new IBM FlashSystem 9200 LFF expansion enclosure Model A9F offers new tiering options with solid-state drive (SSD flash drives). Up to 736 drives of serial-attached SCSI (SAS) expansions are supported per IBM FlashSystem 9200 control enclosure. The expansion enclosure is 5U high.

The best practice recommendation is to balance equally the expansion enclosures between chains. So, if you have two extra expansion enclosures one should be installed on the first SAS chain and one on the second SAS chain. In addition, when you add a single expansion enclosure to an existing system, it is preferable to add the enclosure directly below the control enclosure. When you add a second expansion enclosure, it is preferable to add the enclosure directly above the control enclosure. As more expansion enclosures are added, alternative adding them above and below.

The IBM FlashSystem 9200 system supports up to four control enclosures and up to two chains of SAS expansion enclosures per control enclosure. To limit contention for bandwidth on a chain of SAS enclosures, no more than 10 expansion enclosures can be chained to SAS port 1 of a node canister and no more than 10 expansion enclosures can be chained to SAS port 3 of a node canister. On each SAS chain, the systems can support up to a SAS chain weight of ten, where:

- Each 9846-A9F or 9848-A9F expansion enclosure adds a value of 2.5 to the SAS chain weight.
- Each 9846-AFF or 9848-AFF expansion enclosure adds a value of 1 to the SAS chain weight.

For example, each of the following expansion enclosure configurations has a total SAS weight of ten:

- Four 9848-A9F enclosures per SAS chain
- Two 9846-A9F enclosures and five 9848-AFF enclosures per SAS chain

Figure 10-14 shows the cabling for adding two A9F expansion enclosures and two AFF expansion enclosures to a single control enclosure (in the center of Figure 10-14).

For more information, see this IBM Documentation web page.

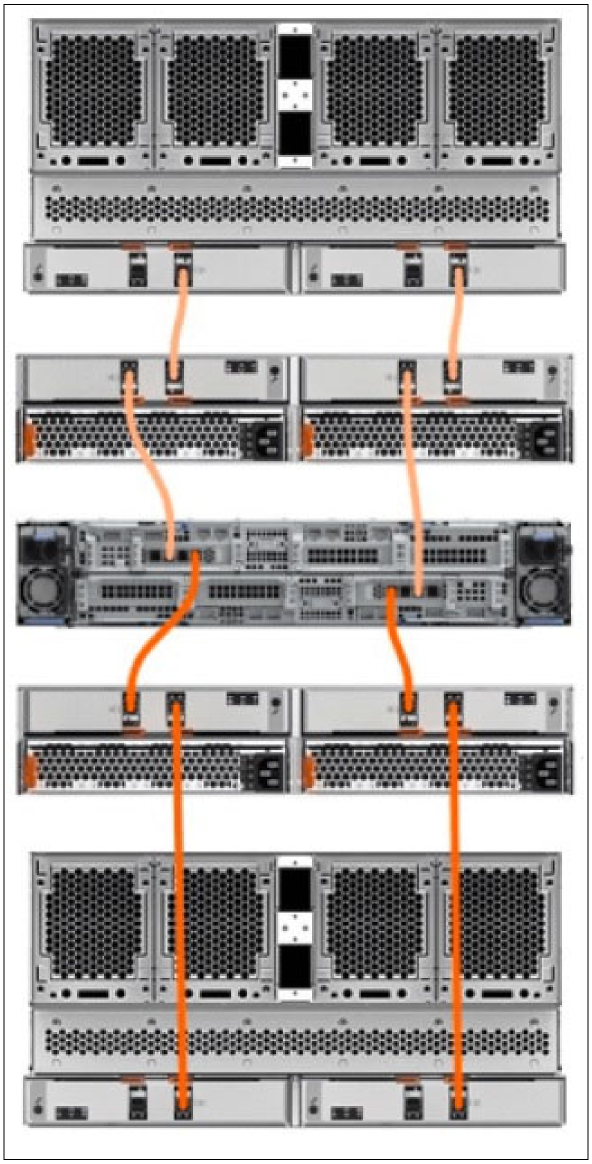


Figure 10-14 Cabling for adding four expansion enclosures in two SAS chains

Adding expansion enclosures is simplified because IBM FlashSystem 9200 can automatically discover new expansion enclosures after the SAS cables are connected. It is possible to manage and use the new drives without managing the new expansion enclosures. However, unmanaged expansion enclosures are not monitored properly. This issue can lead to more difficult troubleshooting and can make problem resolution take longer.

To avoid this situation, always manage newly added expansion enclosures and follow these guidelines:

- FlashSystem 9200 systems support 4-port SAS interface adapters. However, only ports 1 and 3 are used for SAS connections.
- Connect SAS port 1 of the upper node canister in the control enclosure to SAS port 1 of the left expansion canister in the first expansion enclosure.
- Connect SAS port 1 of the lower node canister in the control enclosure to SAS port 1 of the right expansion canister in the first expansion enclosure.
- In general, the SAS interface adapter must be installed in PCIe slot 3 of the node canister.
- No cable can be connected between a port on a left canister and a port on a right canister.
- A cable must not be connected between ports in the same enclosure.
- A connected port on the node canister must connect to a single port on an expansion canister. Cables that split the connector out into separate physical connections are not supported.
- Attach cables serially between enclosures; do not skip an enclosure.
- The last expansion enclosure in a chain must not have cables in port 2 of canister 1 or port 2 of canister 2.
- Ensure that cables are installed in an orderly way to reduce the risk of cable damage when replaceable units are removed or inserted.

IBM FlashSystem 9100

The procedure for adding expansion enclosures to an IBM FlashSystem 9100 control enclosure is similar to that described in section “IBM FlashSystem 9200” on page 490.

For more information, see this IBM Documentation web page.

IBM FlashSystem 7200

The following types of expansion enclosures are available:

- IBM FlashSystem 7200 LFF Expansion Enclosure Model 12G
- IBM FlashSystem 7200 SFF Expansion Enclosure Model 24G
- IBM FlashSystem 7200 Dense Expansion Enclosure Model 92G

When attaching expansion enclosures to the control enclosure, you are not limited by the type of the enclosure (if it meets all generation level restrictions). The only limitation for each SAS chain is its chain weight. Each type of enclosure defines its own chain weight, as follows:

- Enclosures 12G and 24G have a chain weight of 1.
- Enclosure 92G has a chain weight of 2.5.
- The maximum chain weight for any SAS chain is 10.
- The maximum number of SAS chains per control enclosure is 2.

For example, you can combine seven 24G and one 92G expansion enclosures (7x1 + 1x2.5 = 9.5 chain weight), or two 92G enclosures, one 12G, and four 24G (2x2.5 + 1x1 + 4x1 = 10 chain weight).

You can use either the `addcontrolenclosure` command or the Add Enclosure wizard in the management GUI to add the new expansion enclosure to the system.

To access the Add Enclosure wizard, select **Monitoring** → **System Hardware**. On the **System Hardware - Overview** page, select **Add Enclosure** to start the wizard. Complete the wizard and verify the new enclosure.

If **Add Enclosure** is not displayed, it indicates a potential cabling issue. Check the installation information to ensure that the enclosure was cabled correctly.

Complete the following steps to add an enclosure to the system by using the CLI:

1. Using the **`sainfo lsservicestatus`** command (on the service CLI of the new enclosure), record the WWNN of the new enclosure.
2. Record the serial number of the enclosure, which is needed in later steps.
3. Enter the following command to verify that the enclosure is detected on the fabric:

   `svcinfo lscontrolenclosurecandidate`

4. Enter the **`lsiogrp`** command to determine the I/O group, where the enclosure must be added:
5. Record the name or ID of the first I/O group that has a node count of zero. You need the ID for the next step.
6. Enter the following command to add the enclosure to the system:

   `addcontrolenclosure  -iogrp iogrp_name | iogrp_id -sernum enclosureserialnumber`

   where:

   – **`iogrp_name | iogrp_id`** is the name or ID of the I/O group
   – **`enclosureserialnumber`** is the serial number of the enclosure

7. Record this information for future reference:

   – Serial number.
   – Worldwide node name of both node canisters.
   – All of the worldwide port names.
   – The name or ID of the I/O group that contains the enclosure.

8. Enter the **`lsnodecanister`** command to verify that the node canisters in the enclosure are online.

For more information, see this IBM Documentation web page.

IBM FlashSystem 5x00

Similar to the IBM FlashSystem 7200, the following types of expansion enclosures are available for the 5x00 family:

- FlashSystem 5x00 family supports the attachment of FlashSystem 5x00 expansion enclosure models 12G, 24G, and 92G.
- IBM FlashSystem Model 12G (LFF) supports up to 12 3.5-inch drives.
- IBM FlashSystem Model 24G (SFF) supports up to 24 2.5-inch drives.
- IBM FlashSystem Model 92G holds 92 3.5-inch SAS drives.

High-performance disk drives, high-capacity nearline disk drives, and flash (solid state) drives are supported. Drives of the same form factor can be intermixed within an enclosure, and LFF and SFF expansion enclosures can be intermixed within a FlashSystem 5x00 system. The IBM FlashSystem 5010 supports only one control enclosure and only one SAS expansion chain.

The procedure for adding expansion enclosures to an IBM FlashSystem 5x00 control enclosure is similar to that described in “IBM FlashSystem 7200” on page 492.

For more information, see this IBM Documentation web page.

10.11.7 Removing expansion enclosures

As storage environments change and grow it is sometimes necessary to move expansion enclosures between control enclosures. Removing an expansion enclosure is a straightforward task.

To remove an expansion enclosure from a control enclosure, complete the following steps:

1. If the expansion enclosure that you want to move is not at the end of a SAS chain, you might need a longer pair of SAS cables to complete the procedure. In that case, ensure that you have two SAS cables of suitable length before you start this procedure.
2. Delete any volumes that are no longer needed and that depend on the enclosure that you plan to remove.
3. Delete any remaining arrays that are formed from drives in the expansion enclosure. Any data in those arrays is automatically migrated to other managed disks in the pool if there is enough capacity.
4. Wait for data migration to complete.
5. Mark all the drives (including any configured as spare or candidate drives) in the enclosures to be removed as unused.
6. Unmanage and remove the expansion enclosure by using the management GUI. Select **Monitoring** → **System Hardware**. On the **System Hardware - Overview** page, select the directional arrow next to the enclosure that you are removing to open the Enclosure Details page. Select **Enclosure Actions** → **Remove**.

> **Important:** Do not proceed until the enclosure removal process completes successfully.

7. On the I/O group that contains the expansion enclosure that you want to remove, enter the following command to put the I/O group into maintenance mode:

   ```
   chiogrp -maintenance yes <iogroup_name_or_id>
   ```

8. If the expansion enclosure that you want to move is at the end of a SAS chain, complete the following steps to remove the enclosure from the SAS chain:
   a. Disconnect the SAS cable from port 1 of canister 1 and canister 2. The enclosure is now disconnected from the system.
   b. Disconnect the other ends of the SAS cables from the previous enclosure in the SAS chain. The previous enclosure is now the end of the SAS chain. Proceed to step 10.
9. If the expansion enclosure is not at the end of a SAS chain, complete the following steps to remove the enclosure from the SAS chain:
   a. Disconnect the SAS cable from port 2 of canister 1 of the expansion enclosure that you want to move.
   b. Disconnect the other end of the same SAS cable from port 1 of canister 1 of the next expansion enclosure in the SAS chain.
   c. Disconnect the SAS cable from port 1 of canister 1 of the expansion enclosure that you want to move.
   d. Reroute the cable that was disconnected in the previous step and connect it to port 1 of canister 1 of the next expansion enclosure in the SAS chain.

> **Important:** Do not continue until you complete this cable connection step.

e. Disconnect the SAS cable from port 2 of canister 2 of the expansion enclosure that you want to move.
f. Disconnect the other end of the same SAS cable from port 1 of canister 2 of the next expansion enclosure in the SAS chain.
g. Disconnect the SAS cable from port 1 of canister 2 of the expansion enclosure that you want to move.
h. Reroute the cable that was disconnected in the previous step and connect it to port 1 of canister 2 of the next expansion enclosure in the SAS chain.

10. Take the I/O group out of maintenance mode by entering the following command:

```
chiogrp -maintenance no <iogroup_name_or_id>
```

11. Check the event log for any errors and fix those errors as needed.
12. Disconnect the power from the expansion enclosure that you want to remove.
13. Remove the expansion enclosure from the rack along with its two power cables and two SAS cables.

Note that the IBM FlashSystem products provide methods to securely erase data from a drive when an enclosure is decommissioned or before a drive is removed from the system during a repair activity.

For more information about the CLI commands that are used to run this secure erase function, see this IBM Documentation web page.

10.11.8 IBM FlashWatch

Driven by the concept of “Storage Made Simple,” IBM FlashWatch is a suite of programs that enhances your experience of owning IBM FlashSystem storage. Bringing together programs which span the acquisition, operation and migration phases, this suite aims to reduce deployment and operational risks to improve your support experience, and to offer a fully flexible, commitment-free hardware refresh. For more information, see *What is IBM FlashWatch? Peace of Mind Made Simple*.

IBM FlashWatch is an offering from IBM to complement the purchase of the IBM FlashSystem product. It provides the following features that are included in the purchase of the product:

- IBM Flash Momentum

  Flash Momentum is a storage upgrade program which allows you to replace your controller and storage every 3 years with full flexibility. Before the expiration of the agreement period, you decide whether to keep your FlashSystem, refresh it or simply walk away. You can refresh your FlashSystem for the same monthly price or less, or upsize or downsize your system to meet your needs.

- High Availability guarantee

  Robust Spectrum Virtualize software has a measured availability of 99.9999% and IBM offers an optional 100% availability commitment when HyperSwap is also used.

- Data Reduction Guarantee:

  A 2:1 data reduction is ensured and you must self-certify that the data you are writing can be reduced (for example, not encrypted, not compressed). Up to 5:1 data reduction can be ensured with more detailed profiling of your workload.

- All-inclusive Licensing

  All storage functions available are included in the licensing cost for internal storage.

- Comprehensive Care

  Up to seven years of 24x7 support, with three years of IBM Technical Advisor support, enhanced response times of 30 minutes for severity 1 incidents, and six managed code upgrades over three years. However, this feature is not available for all IBM FlashSystem models (see Table 10-14).

- Storage Insights

  Storage Insights is included at no extra cost to proactively manage your environment.

- Flash Endurance Guarantee

  Flash media is covered for all workloads while under warranty or maintenance.

- IBM Storage Utility pricing

  The IBM Storage Utility pricing solution delivers three years of your planned capacity needs on day one. To predict and control your future needs, IBM uses IBM Storage Insights to help you easily meet your capacity needs without interrupting your data center. The IBM FlashSystem 9200 (9848-UG8) is leased through IBM Global Finance on a three-year lease, which entitles the customer to use approximately 30 - 40% of the total system capacity at no extra cost. If storage needs to increase beyond that initial capacity, usage is billed on a quarterly basis based on the average daily provisioned capacity per terabyte per month.

- No Cost Migration

  For a 90-day period, from the date of installation, you can migrate data from over 500 older storage systems (IBM and non-IBM) to your FlashSystem product using an approach of your choice, without having to pay any additional external licensing.

Table 10-14 provides a summary product matrix for IBM FlashSystem products.

Table 10-14 IBM FlashWatch product matrix for IBM FlashSystem products

IBM FlashWatch feature	FS5000	FS5200	FS7200	FS9200	FS9200R
High Availability guarantee	FS5035 only	Yes	Yes	Yes	Yes
Data reduction guarantee	FS5035 only	Yes	Yes	Yes	Yes
All-inclusive licensing (Excluding external virtualization, encryption)	N/A	Yes	Yes	Yes	Yes
Expert Care	Alternative optional services available; 9x5 NBD warranty	Yes (4662-6H2, UH6, 12G, 24G, 92G)	Yes (4664-824, U7C, 12G, 24G, 92G)	Yes (4666-AG8, UG8, AFF, A9F)	Yes (4666-AG8, UG8, AFF, A9F)

IBM FlashWatch feature	FS5000	FS5200	FS7200	FS9200	FS9200R
Cloud analytics with Storage Insights	Yes	Yes, with IBM Storage Expert Care	Yes	Yes	Yes
Flash Endurance guarantee	Yes	Yes	Yes	Yes	Yes
IBM Flash momentum Storage Upgrade Program	Yes (2072-2N2, 2N4, 3N2, 3N4, 12G, 24G, 92G)	Yes (4662-6H2, UH6, 12G, 24G, 92G)	Yes (4664, 2076-824, U7C, 12G, 24G, 92G)	Yes (4666, 9848-AG8, UG8, AFF, A9F)	Yes (4666, 9848-AG8)
Cloud-like pricing (Storage Utility)	N/A	Yes (4662-UH6)	Yes (4664, 2076-U7C)	Yes (4666, 9848-UG8)	N/A
No Cost Migration	Yes	Yes	Yes	Yes	Yes

For more information about the IBM FlashWatch offering, see *IBM FlashWatch FAQ*.

10.12 I/O throttling

I/O throttling is a mechanism that allows you to limit the volume of I/O processed by the storage controller at various levels to achieve quality of service (QoS). If a throttle is defined, the system either processes the I/O, or delays the processing of the I/O to free resources for more critical I/O. Throttling is a way to achieve a better distribution of storage controller resources.

IBM FlashSystem V8.3 and later code brings the possibility to set the throttling at a volume level, host, host cluster, storage pool, and offload throttling by using the GUI. This section describes some details of I/O throttling and shows how to configure the feature in your system.

10.12.1 General information about I/O throttling

I/O throttling features the following characteristics:

- IOPS and bandwidth throttle limits can be set.
- It is an upper bound QoS mechanism.
- No minimum performance is guaranteed.
- Volumes, hosts, host clusters, and managed disk groups can be throttled.
- Queuing occurs at microsecond granularity.
- Internal I/O operations (FlashCopy, cluster traffic, and so on) are not throttled.
- Reduces I/O bursts and smooths the I/O flow with variable delay in throttled I/Os.
- Throttle limit is a per node value.

10.12.2 I/O throttling on front-end I/O control

You can use throttling for a better front-end I/O control at the volume, host, host cluster, and offload levels:

- In a multi-tenant environment, hosts can have their own defined limits.

  You can use this to allow restricted I/Os from a data mining server and a higher limit for an application server.
- An aggressive host consuming bandwidth of the controller can be limited by a throttle.

  For example, a video streaming application can have a limit set to avoid consuming too much of the bandwidth.
- Restrict a group of hosts by their throttles.

  For example, Department A gets more bandwidth than Department B.
- Each volume can have a defined throttle.

  For example, a volume that is used for backups can be configured to use less bandwidth than a volume used for a production database.
- When performing migrations in a production environment consider using host or volume level throttles.
- Offloaded I/Os.

  Offload commands, such as UNMAP and XCOPY, free hosts and speed the copy process by offloading the operations of certain types of hosts to a storage system. These commands are used by hosts to format new file systems or copy volumes without the host needing to read and then write data. Throttles can be used to delay processing for offloads to free bandwidth for other more critical operations, which can improve performance but limits the rate at which host features, such as VMware VMotion, can copy data.

10.12.3 I/O throttling on back-end I/O control

You can also use throttling to control the back-end I/O by throttling the storage pools, which can be useful in the following scenarios:

- Each storage pool can have a defined throttle.
- Allows control of back-end I/Os from the IBM FlashSystem.
- Useful to avoid overwhelming any external back-end storage.
- Useful in case of VVOLS since a VVOL gets created in a child pool. A child pool (**`mdiskgrp`**) throttle can control I/Os coming from that VVOL.
- Only parent pools support throttles because only parent pools contain MDisks from internal or external back-end storage. For volumes in child pools, the throttle of the parent pool is applied.
- If more than one throttle applies to an I/O operation, the lowest and most stringent throttle is used. For example, if a throttle of 100 MBps is defined on a pool and a throttle of 200 MBps is defined on a volume of that pool, the I/O operations are limited to 100 MBps.

10.12.4 Overall benefits of using I/O throttling

The overall benefit of the use of I/O throttling is a better distribution all system resources that includes the following benefits:

- Avoids overwhelming the controller objects.
- Avoids starving the external entities, like *hosts*, from their share of controller.
- Creates scheme of distribution of controller resources that, in turn, results in better utilization of external resources such as host capacities.

With throttling not enabled, a scenario exists in where Host1 dominates the bandwidth, and after enabling the throttle, a much better distribution of the bandwidth among the hosts results, as shown in Figure 10-15.

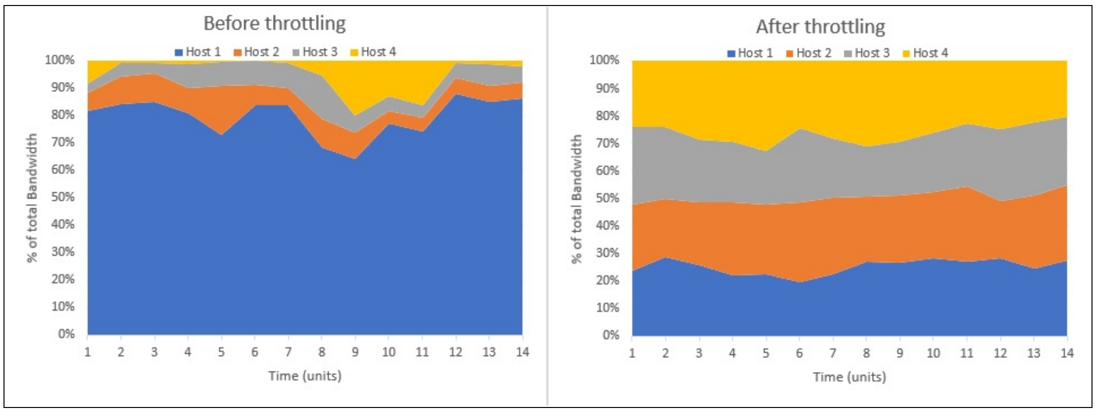


Figure 10-15 Distribution of controller resources before and after I/O throttling

10.12.5 Considerations for I/O throttling

Consider the following points when you are planning to use I/O throttling:

- The throttle cannot be defined for the host if it is part of a host cluster that has a host cluster throttle.
- If the host cluster does not have a throttle that is defined, its member hosts can have their individual host throttles defined.
- If a volume has multiple copies, then throttling would be done for the storage pool serving the primary copy. The throttling is not applicable on the secondary pool for mirrored volumes and stretched cluster implementations.
- A host cannot be added to a host cluster if both have their individual throttles defined. If just one of the host or host cluster throttles is present, the command succeeds.
- A seeding host that is used for creating a host cluster cannot have a host throttle that is defined for it.

> **Note:** Throttling is only applicable at the I/Os that an IBM FlashSystem receives from hosts and host clusters. The I/Os generated internally, such as mirrored volume I/Os, cannot be throttled.

10.12.6 Configuring I/O throttling using the CLI

To create a throttle using the CLI, use the **mkthrottle** command, as shown in Example 10-10. The bandwidth limit is the maximum amount of bandwidth the system can process before the system delays I/O processing. Similarly, the iops_limit is the maximum amount of IOPS the system can process before the system delays I/O processing.

Example 10-10 Creating a throttle using the mkthrottle command in the CLI

```
Syntax:

mkthrottle -type [offload | vdisk | host | hostcluster | mdiskgrp]
          [-bandwidth bandwidth_limit_in_mb]
          [-iops iops_limit]
          [-name throttle_name]
          [-vdisk vdisk_id_or_name]
          [-host host_id or name]
          [-hostcluster hostcluster_id or name]
          [-mdiskgrp mdiskgrp_id or name]

Usage examples:
IBM_FlashSystem:IBM Redbook FS:superuser>mkthrottle -type host -bandwidth 100
-host ITSO_HOST3
IBM_FlashSystem:IBM Redbook FS:superuser>mkthrottle -type host cluster -iops 30000
-hostcluster ITSO_HOSTCLUSTER1
IBM_FlashSystem:IBM Redbook FS:superuser>mkthrottle -type mdiskgrp -iops 40000
-mdiskgrp 0
IBM_FlashSystem:IBM Redbook FS:superuser>mkthrottle -type offload -bandwidth 50
IBM_FlashSystem:IBM Redbook FS:superuser>mkthrottle -type vdisk -bandwidth 25
-vdisk volume1

IBM_FlashSystem:IBM Redbook FS:superuser>lsthrottle
throttle_id throttle_name object_id object_name      throttle_type IOPs_limit
bandwidth_limit_MB
0        throttle0    2       ITSO_HOST3      host              100
1          throttle1     0         ITSO_HOSTCLUSTER1 host cluster
30000
2          throttle2     0         Pool0            mdiskgrp
40000
3        throttle3                         offload             50
4          throttle4     10        volume1          vdisk                       25
```

Note: You can change a throttle parameter by using the **chthrottle** command.

10.12.7 Configuring I/O throttling using the GUI

The following sections show how to configure the throttle by using the management GUI.

10.12.8 Creating a volume throttle

To create a volume throttle, go to **Volumes** → **Volumes**, then select the wanted volume, right-click it and chose **Edit Throttle**, as shown in Figure 10-16. The bandwidth can be set from 1 MBps - 256 TBps and IOPS can be set from 1 - 33,254,432.

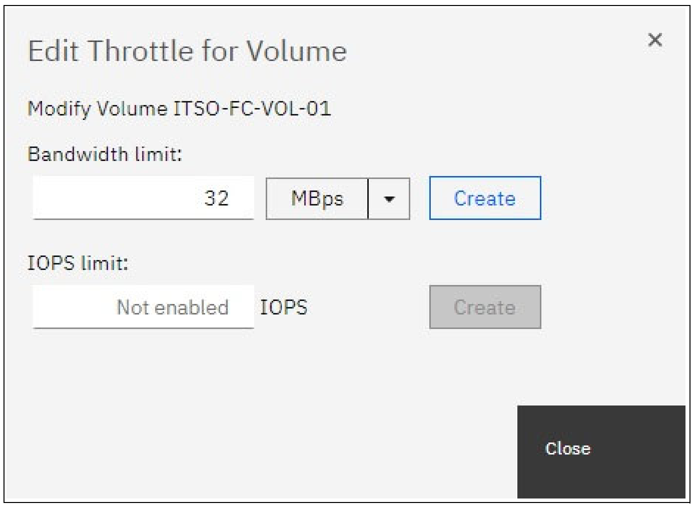


Figure 10-16 Creating a volume throttle in the GUI

If a throttle already exists, the dialog box that is shown in Figure 10-16 on page 501 also shows a **Remove** button that is used to delete the throttle.

10.12.9 Creating a host throttle

To create a host throttle, go to **Hosts** → **Hosts**, select the wanted host, then right-click it and chose **Edit Throttle**, as shown in Figure 10-17.

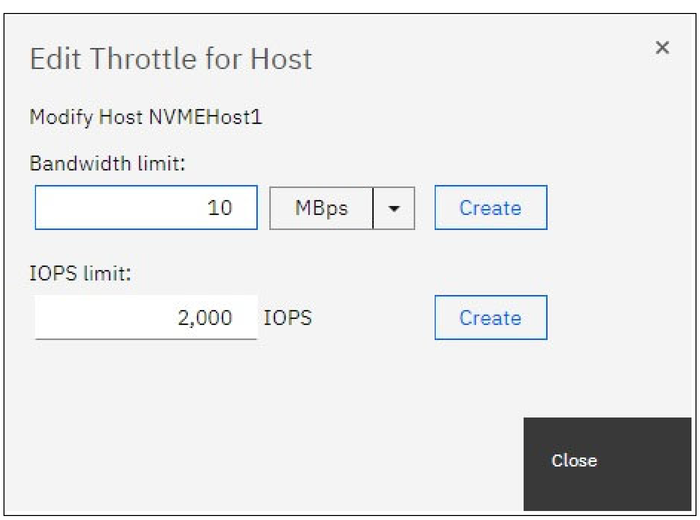


Figure 10-17 Creating a host throttle in the GUI

10.12.10 Creating a host cluster throttle

To create a host cluster throttle, go to **Hosts** → **Host Clusters**, select the wanted host cluster, then right-click it and chose **Edit Throttle**, as shown in Figure 10-18.

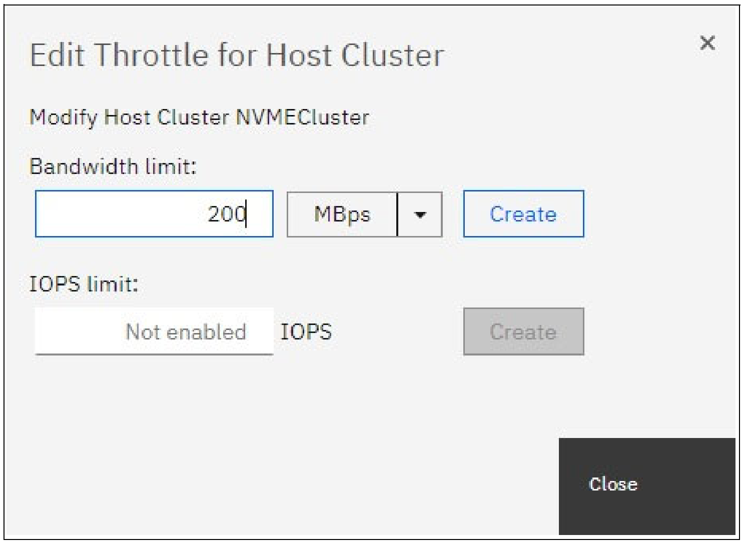


Figure 10-18 Creating a host cluster throttle in the GUI

10.12.11 Creating a storage pool throttle

To create a storage pool throttle, go to **Pools** → **Pools**, select the wanted storage pool, then right-click it and choose **Edit Throttle**, as shown in Figure 10-19.

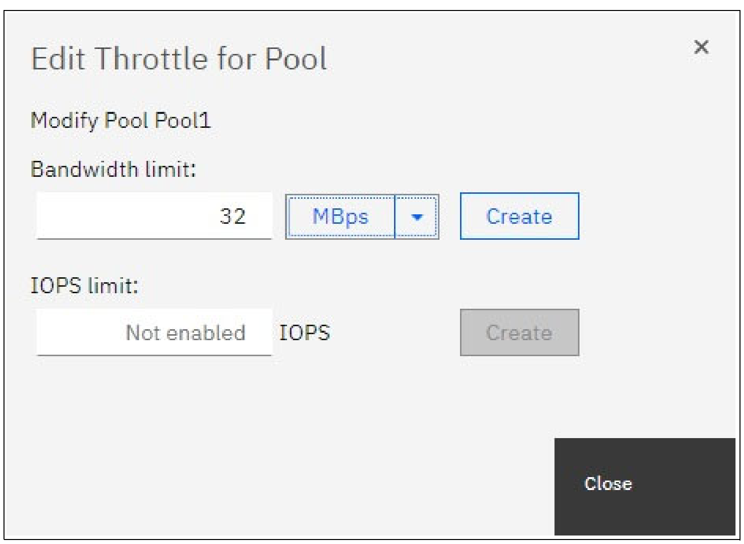


Figure 10-19 Creating a storage pool throttle in the GUI

10.12.12 Creating an offload throttle

To create an offload throttle, go to **Monitoring** → **System Hardware** → **Actions**, then select **Edit System Offload Throttle**, as shown in Figure 10-20.

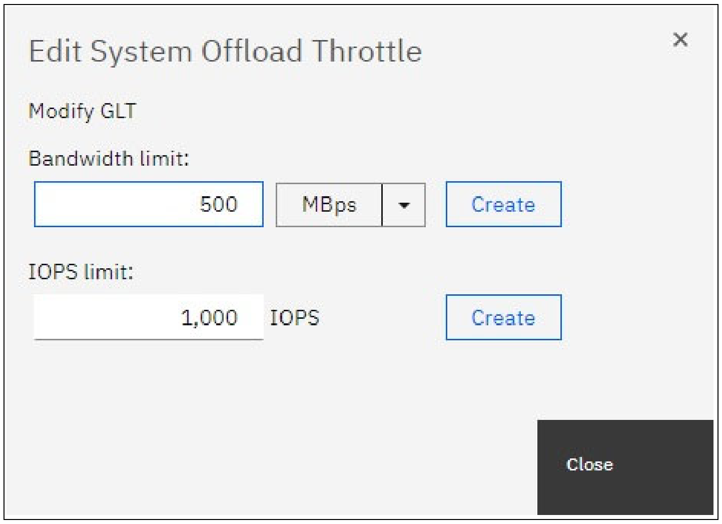


Figure 10-20 Creating system offload throttle in the GUI

10.13 Automation

Automation has become a priority for maintaining today's busy storage environments. Automation software allows the creation of repeatable sets of instructions and processes to reduce the need for human interaction with computer systems. Red Hat Ansible and other third-party automation tools are becoming increasingly used across the enterprise IT environments and it is not unexpected that their use in storage environments is becoming more popular.

10.13.1 Red Hat Ansible

IBM FlashSystem family includes integration with Red Hat Ansible Automation Platform, allowing IT to create an Ansible playbook that automates repetitive tasks across an organization in a consistent way, helping improve outcomes and reduce errors.

Ansible is an agentless automation management tool that uses the SSH protocol. Currently, Ansible can be run from any machine with Python 2 (version 2.7) or Python 3 (versions 3.5 and higher) installed. This includes Red Hat, Debian, CentOS, macOS, any of the BSDs. Windows is not supported for the Ansible control node.

IBM is a Red Hat certified support module vendor, providing simple management for the following commands used in the IBM Spectrum Virtualize Ansible Collection:

- Collect facts: Collect basic information including hosts, host groups, snapshots, consistency groups, and volumes.
- Manage hosts: Create, delete, or modify hosts.
- Manage volumes: Create, delete, or extend the capacity of volumes.
- Manage MDisk: Create or delete a managed disk.

- Manage pool: Create or delete a pool (managed disk group).
- Manage volume map: Create or delete a volume map.
- Manage consistency group snapshot: Create or delete consistency group snapshots.
- Manage snapshot: Create or delete snapshots.
- Manage volume clones: Create or delete volume clones.

This collection provides a series of Ansible modules and plug-ins for interacting with the IBM Spectrum Virtualize Family storage systems. The modules in the IBM Spectrum Virtualize Ansible collection use the Representational State Transfer (REST) application programming interface (API) to connect to the IBM Spectrum Virtualize storage system. These storage systems include the IBM SAN Volume Controller, IBM FlashSystem family including FlashSystem 5010, 5030, 5100, 7200, 9100, 9200, 9200R, and IBM Spectrum Virtualize for Public Cloud.

More information can be found in the following IBM Redpaper *Automate and Orchestrate® Your IBM FlashSystem Hybrid Cloud with Red Hat Ansible*, REDP-5598.

For IBM Spectrum Virtualize modules, Ansible version 2.9 or higher is required. For more information about IBM Spectrum Virtualize modules, see Ansible Collections for IBM Spectrum Virtualize.

10.13.2 RESTful API

The Spectrum Virtualize REST model API consists of command targets that are used to retrieve system information and to create, modify, and delete system resources. These command targets allow command parameters to pass through unedited to the Spectrum Virtualize command-line interface, which handles parsing parameter specifications for validity and error reporting. Hypertext Transfer Protocol Secure (HTTPS) is used to communicate with the RESTful API server.

To interact with the storage system by using the RESTful API, use the curl utility (see https://curl.se) to make an HTTPS command request with a valid configuration node URL destination. Open TCP port 7443 and include the keyword `rest` followed by the Spectrum Virtualize target command you want to run:

Each **`curl`** command takes the following form:

```
curl –k –X POST –H <header_1> –H <header_2> ... -d <JSON input>
https://<flashsystem_ip_address>:7443/rest/<target>
```

Where:

- POST is the only HTTPS method that the Spectrum Virtualize RESTful API supports.
- Headers <**`header_1`**> and <**`header_2`**> are individually specified HTTP headers (for example, Content-Type and X-AuthUsername).
- **`-d`** is followed by the JSON input; for example, ‘{“raid_level”: “raid5”}’.
- <**`flashsystem_ip_address`**> is the IP address of the IBM FlashSystem that you are sending requests to.
- <**`target`**> is the target object of commands, which includes any object IDs, names, and parameters.

Authentication

Aside from data encryption, the HTTPS server requires authentication of a valid username and password for each API session. Use two authentication header fields to specify your credentials: X-Auth-Username and X-Auth-Password.

Initial authentication requires that you POST the authentication target (/auth) with the username and password. The RESTful API server returns a hexadecimal token. A single session lasts a maximum of two active hours or thirty inactive minutes, whichever occurs first. When your session ends due to inactivity, or if you reach the maximum time that is allotted, error code 403 indicates the loss of authorization. Use the /auth command target to reauthenticate with the user name and password.

The following is an example of the correct procedure for authenticating. You authenticate by first producing an authentication token and then using that token in all future commands until the session ends.

For example, the following command passes the authentication command to IBM FlashSystem node IP address 192.168.10.20 at port 7443:

```
curl –k –X POST –H ‘Content-Type: application/json’ –H ‘X-Auth-Username: superuser’ –H ‘X-Auth-Password: passw0rd’ https://192.168.10.20:7443/rest/auth
```

> **Note:** Make sure you format the request correctly using spaces after each colon in each header otherwise the command fails.

This request yields an authentication token, which can be used for all subsequent commands. For example:

```
{"token": "38823f60c758dca26f3eaac0ffee42aadc4664964905a6f058ae2ec92e0f0b63"}
```

Example command

Most actions must be taken only after authentication. The following example of creating an array demonstrates the use of the previously generated token in place of the authentication headers used in the authentication process.

```
curl –k –X POST –H ‘Content-Type: application/json’ –H ‘X-Auth-Token: 38823f60c758dca26f3eaac0ffee42aadc4664964905a6f058ae2ec92e0f0b63’
–d ‘{"level": "raid5", "drive": "6:7:8:9:10", "raid6grp"}’
https://192.168.10.20:7443/rest/mkarray
```

For more information about other examples, see the following web pages:

- IBM Spectrum Virtualize Interfacing Using the RESTful API
- Implementing a RESTful API to the IBM Storwize Family
- Tips and tricks using the Spectrum Virtualize REST API

For more information about RESTful API, see Chapter 12., “Automation and scripting” on page 567.

10.14 Documenting IBM FlashSystem and SAN environment

This section focuses on the challenge of automating the documentation that is needed for an IBM FlashSystem 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.
- Remember to also include photographs in the documentation where appropriate.

> **Storing documentation:** Avoid storing IBM FlashSystem and SAN environment documentation only in the SAN. If your organization has a disaster recovery plan, include this storage documentation in it. Follow its guidelines about how to update and store this data. If no disaster recovery plan exists and you have the proper security authorization, it might be helpful to store an updated copy offsite.

In theory, this IBM FlashSystem and SAN environment documentation should be written at a level sufficient for any system administrator who has average skills in the products to understand. Make a copy that includes all 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-built documentation") as you install your solution. If you completed forms to help plan the installation of your IBM FlashSystem solution, use these forms to help you document how your IBM FlashSystem solution was first configured. Minimum documentation is needed for an IBM FlashSystem 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.

10.14.1 Naming conventions

Whether you are creating your IBM FlashSystem and SAN environment documentation, or you are updating what is already in place, first evaluate whether you have a good naming convention in place. With a good naming convention, you can quickly and uniquely identify the components of your IBM FlashSystem and SAN environment. System administrators can then determine whether a name belongs to a volume, storage pool, MDisk, host, or HBA by looking at it.

Because error messages often point to the device that generated an error, a good naming convention quickly highlights where to start investigating when an error occurs. Typical IBM FlashSystem and SAN component names limit the number and type of characters you can use. For example, IBM FlashSystem names are limited to 63 characters, which make creating a naming convention easier

Many names in IBM FlashSystem and SAN environment can be modified online. Therefore, you do not need to worry about planning outages to implement your new naming convention. The naming examples that are used in the following sections are effective in most cases but might not be fully adequate for your environment or needs. The naming convention to use is your choice, but you must implement it in the whole environment.

Enclosures, node canisters, and external storage controllers

IBM FlashSystem names its internal canisters or nodes as `nodeX`, with `X` being a sequential decimal number. These range 2 - 8, in a four IBM FlashSystem 9200 system cluster.

If multiple external controllers are attached to your IBM FlashSystem solution, these are detected as `controllerX` so you might need to 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 IBM FlashSystem to know which storage controller to check.

> **Note:** An IBM FlashSystem detects external controllers based on their worldwide node name (WWNN). If you have an external storage controller that has one WWNN for each worldwide port name (WWPN), this configuration might lead to many `controllerX` names pointing to the same physical box. In this case, prepare a naming convention to cover this situation.

MDisks and storage pools

When an IBM FlashSystem detects new MDisks, it names them by default as `mdiskXX`, where `XX` is a sequential number. You should change the `XX` value to something more meaningful. MDisks are either arrays (DRAID) from internal storage or volumes from an external storage system. Ultimately, it comes down to personal preference and what works in your environment. The main “convention” you should follow is avoid the use of special characters in names, apart from the underscore, the hyphen and the period, which are permitted and spaces (which can make scripting difficult).

For example, you can change it to include the following information:

- For internal MDisks refer to the IBM FlashSystem system or cluster name
- A reference to the external 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:

- `FS9200CL01-MD03`, where FS9200CL01 is the system or cluster name, and MD03 is the MDisk name.
- `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 if external, 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:

- `FS9200-POOL01`: where FS9200 is the system or cluster name, and POOL01 is the pool.
- `P05XYZ1_3GR5`: Pool 05 from serial 75VXYZ1, LUNs with 300 GB FC DDMs and RAID 5.
- `P16XYZ1_EX01`: Pool 16 from serial 75VXYZ1, pool 01 dedicated to Exchange Mail servers.
- `XIV01_F9H02_ET`: Pool with disks from XIV named XIV01 and FlashSystem 900 F9H02, both managed by Easy Tier.

Volumes

Volume names should include the following information:

- The host or cluster to which the volume is mapped.
- A single letter that indicates its usage by the host, as shown in the following examples:
  - B: For a boot disk, or R for a rootvg disk (if the server boots from SAN)
  - D: For a regular data disk
  - Q: For a cluster quorum disk (do not confuse with IBM FlashSystem quorum disks)
  - L: For a database log disk
  - T: For a database table disk
- A few sequential digits, for uniqueness.
- Sessions standard for VMware data stores:
  - `esx01-sessions-001`: For a data store composed of a single volume
  - `esx01-sessions-001a` and `esx01-sessions-001b`: For a data store composed of two volumes

For example, `ERPNY01-T03` indicates a volume that is mapped to server `ERPNY01` and database table disk `03`.

Hosts

In today's environment, administrators deal with large networks, the internet, and cloud computing. Use good server naming conventions so that they can quickly identify a server and determine the following information:

- Where it is (to know how to access it).
- What kind it is (to determine the vendor and support group in charge).
- What it does (to engage the proper application support and notify its owner).
- Its importance (to determine the severity if problems occur).

Changing a server's name in IBM FlashSystem is as simple as changing any other IBM FlashSystem object name. However, changing the name on the operating system of a server might have implications for application configuration, DNS and can require a server restart. Therefore, you might want to prepare a detailed plan if you decide to rename several servers in your network. The following example is for a server naming convention of `LLAATRFFNN` where:

- LL is the location, which might designate a city, data center, building floor, or room.
- AA is a major application, for example, billing, ERP, and Data Warehouse.
- T is the type, for example, UNIX, Windows, and VMware.
- R is the role, for example, Production, Test, Q&A, and Development.
- FF is the function, for example, DB server, application server, web server, and file server.
- NN is numeric.

SAN aliases and zones

SAN aliases often need to reflect only the device and port that is associated to it. Including information about where one particular device port is physically attached on the SAN might lead to inconsistencies if you make a change or perform maintenance and then forget to update the alias. Create one alias for each device port WWPN in your SAN and use these aliases in your zoning configuration. Consider the following examples:

- `AIX_NYBIXTDB02_FC2: Interface fcs2 of AIX server NYBIXTDB02.`
- `LIN-POKBIXAP01-FC1: Interface fcs1 of Linux Server POKBIXAP01.`
- `WIN_EXCHSRV01_HBA1: Interface HBA1 of physical Windows server EXCHSRV01.`
- `ESX_NYVMCLUSTER01_VMHBA2: Interface vmhba2 of ESX server NYVMCLUSTER01.`
- `IBM-NYFS9200-N1P1_HOST: Port 1 of Node 1 from FS9200 Cluster NYFS9200 dedicated for hosts.`
- `IBM-NYFS9200-N1P5_INTRA: Port 5 of Node 1 from FS9200 Cluster NYFS9200 dedicated to intracluster traffic.`
- `IBM-NYFS9200-N1P7_REPL: Port 7 of Node 1 from FS9200 Cluster NYFS9200 dedicated to replication.`

  Be mindful of the IBM FlashSystem 9200 port aliases. There are mappings between the last digits of the port WWPN and the node FC port.

- `IBM_D88870_75XY131_I0301: DS8870 serial number75XY131, port I0301.`
- `TS4500-TD06: TS4500 tape library, tape drive 06.`
- `EMC_VNX7500_01_SPA2: EMC VNX7500 hostname VNX7500_01, SP A, port 2.`

If your SAN does not support aliases (for example, in heterogeneous fabrics with switches in some interoperation modes), use WWPNs in your zones. However, remember to update every zone that uses a WWPN if you change it.

Your SAN zone name should reflect the devices in the SAN it includes (normally in a one-to-one relationship), as shown in the following examples:

- `SERVERALIAS_T1_FS9200CLUSTERNAME` (from a server to the IBM FlashSystem 9200, where you use T1 as an identifier to zones that uses, for example, node ports P1 on Fabric A, and P2 on Fabric B).
- `SERVERALIAS_T2_FS9200CLUSTERNAME` (from a server to the IBM FlashSystem 9200, where you use T2 as an identifier to zones that uses, for example, node ports P3 on Fabric A, and P4 on Fabric B).
- `IBM_DS8870_75XY131_FS9200CLUSTERNAME` (zone between an external back-end storage and the IBM FlashSystem 9200).
- `NYC_FS9200_POK_FS9200_REPLICATION` (for Remote Copy services).

10.14.2 SAN fabric documentation

The most basic piece of SAN documentation is a SAN diagram. It is likely to be one of the first pieces of information you need if you ever seek support from your SAN switches vendor. Also, a good spreadsheet with ports and zoning information eases the task of searching for detailed information, which if included in the diagram, makes the diagram easier to use.

Brocade SAN Health

The *Brocade SAN Health Diagnostics Capture tool* is a no-cost, automated tool that can help you retain this documentation. SAN Health consists of a data collection tool that logs in to the SAN switches that you indicate and collects data by using standard SAN switch commands. The tool then creates a compressed file with the data collection. This file is sent to a Brocade automated machine for processing by secure web or email.

After some time (typically a few hours), you receive 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 Brocade SAN Health.

The first time that you use the SAN Health Diagnostics Capture tool, explore the options that are provided to learn how to create a well-organized and useful diagram.

Figure 10-21 shows an example of a poorly formatted diagram.

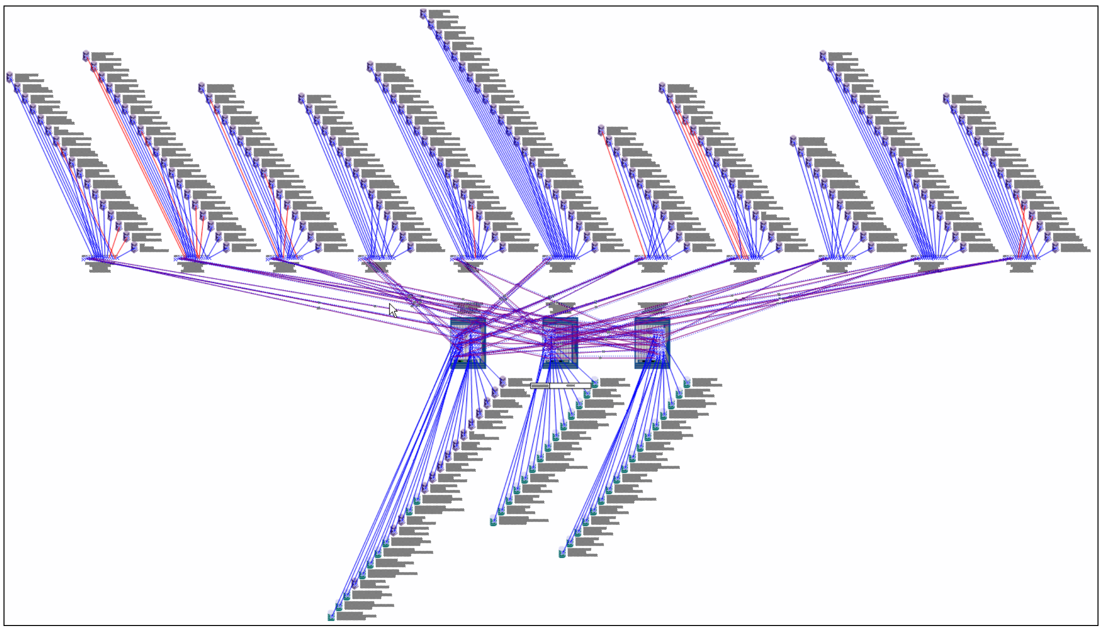


Figure 10-21 Poorly formatted SAN diagram

Figure 10-22 shows a tab of the SAN Health Options window in which you can choose the format of SAN diagram that best suits your needs. Depending on the topology and size of your SAN fabrics, you might want to manipulate the options in the Diagram Format or Report Format tabs.

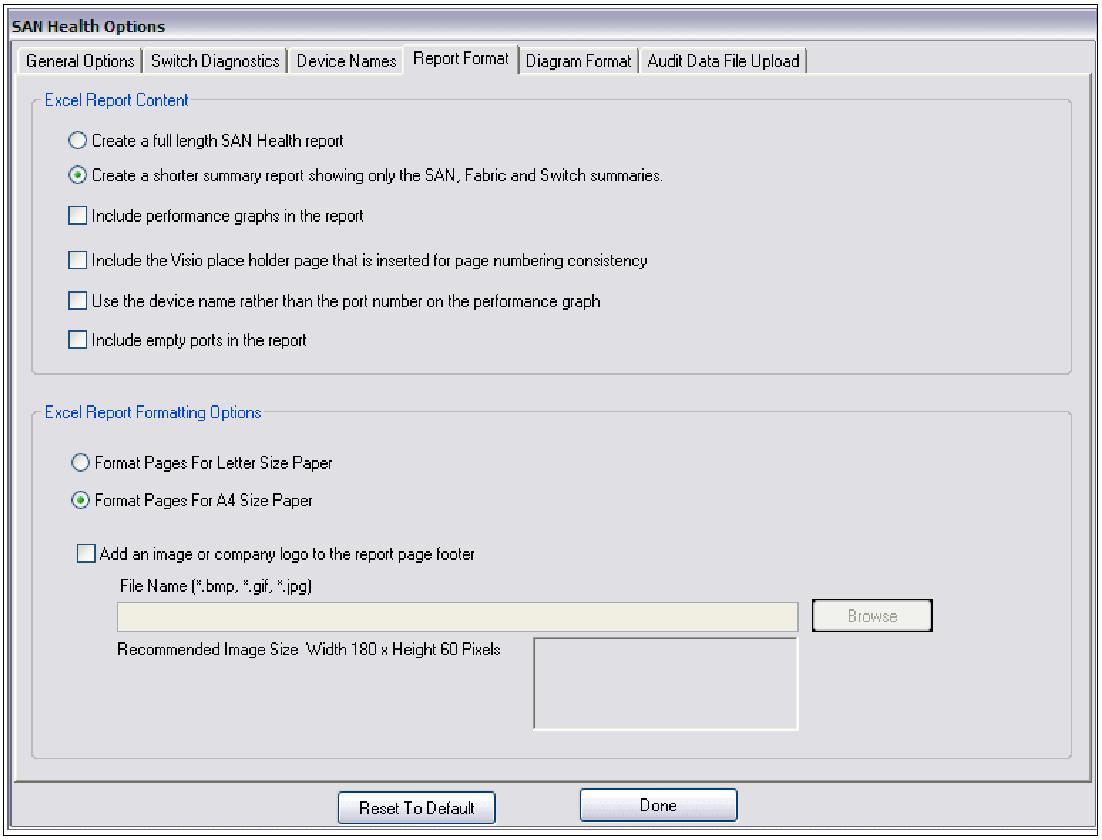


Figure 10-22 Brocade SAN Health Options window

SAN Health supports switches from manufacturers other than Brocade, such as Cisco. Both the data collection tool download and the processing of files are available at no cost. You can download Microsoft Visio and Excel viewers at no cost from the Microsoft website.

Another tool, which is known as *SAN Health Professional*, is also available for download at no cost. With this tool, you can audit the reports in detail by using advanced search functions and inventory tracking. You can configure the SAN Health Diagnostics Capture tool as a Windows scheduled task. To download of the SAN Health Diagnostics Capture tool, see this Broadcom web page.

> **Tip:** Regardless of the method that is used, generate a fresh report at least once a month or after any major changes. Keep previous versions so that you can track the evolution of your SAN.

IBM Spectrum Control reporting

If you have IBM Spectrum Control running in your environment, you can use it to generate reports on your SAN. For more information about how to configure and schedule IBM Spectrum Control reports, see this IBM Documentation web page.

For more information about how to configure and set up Spectrum Control, see Chapter 9, “Implementing a storage monitoring system” on page 387.

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.

10.14.3 IBM FlashSystem documentation

You can back up the configuration data for an IBM FlashSystem system after preliminary tasks are completed. Configuration data for the system provides information about your system and the objects that are defined in it. It contains the configuration data of arrays, pools, volumes, and so on. The backup does not contain any data from the volumes themselves.

Before you back up your configuration data, the following prerequisites must be met:

- Independent operations that change the configuration for the system cannot be running while the **backup** command is running.
- Object names cannot begin with an underscore character (_).

> **Note:** The system automatically creates a backup of the configuration data each day at 1 AM. This backup is known as a *cron backup* and on the configuration node is copied to `/dumps/svc.config.cron.xml_<serial#>`.

Complete the following steps to generate a manual backup at any time:

1. Issue the **svcconfig backup** command to back up your configuration. The command displays messages similar to the ones in Example 10-11 on page 512.

   Example 10-11 Sample svcconfig backup command output

   ```
   IBM_FlashSystem:IBM Redbook FS:superuser>svcconfig backup
   ..............................................................................
   ...
   ..............................................................................
   ...
   ..............................................................................
   CMMVC6155I SVCCONFIG processing completed successfully
   ```

   The **svcconfig backup** command creates three files that provide information about the backup process and the configuration. These files are created in the `/tmp` directory and copied to the `/dumps` directory of the configuration node. You can use the **lsdumps** command to list them. Table 10-15 describes the three files that are created by the backup process.

Table 10-15 Files created by the backup process

File name	Description
svc.config.backup.xml_<serial#>	Contains your configuration data.
svc.config.backup.sh_<serial#>	Contains the names of the commands that were issued to create the backup of the system.
svc.config.backup.log_<serial#>	Contains details about the backup, including any reported errors or warnings.

2. Check that the **svcconfig backup** command completes successfully and examine the command output for any warnings or errors. The following output is an example of the message that is displayed when the backup process is successful:

   ```
   CMMVC6155I SVCCONFIG processing completed successfully
   ```

3. If the process fails, resolve the errors and run the command again.
4. Keep backup copies of the files outside the system to protect them against a system hardware failure. With Microsoft Windows, use the PuTTY **pscp** utility. With UNIX or Linux, you can use the standard **scp** utility. By using the **-unsafe** option, you can use a wildcard to download all the `svc.config.backup` files with a single command. Example 10-12 shows the output of the **pscp** command.

Example 10-12 Saving the configuration backup files to your workstation

```
C:\>
pscp -unsafe superuser@9.10.11.12:/dumps/svc.config.backup.* C:\
Using keyboard-interactive authentication.
Password:
svc.config.backup.log_78E | 33 kB | 33.6 kB/s | ETA: 00:00:00 | 100%
svc.config.backup.sh_78E0 | 13 kB | 13.9 kB/s | ETA: 00:00:00 | 100%
svc.config.backup.xml_78E | 312 kB | 62.5 kB/s | ETA: 00:00:00 | 100%
C:\>
```

The configuration backup file is in Extensible Markup Language (XML) format and can be inserted as an object into your IBM FlashSystem documentation spreadsheet. The configuration backup file might be large. For example, it contains information about each internal storage drive that is installed in the system.

Note: Directly importing the file into your IBM FlashSystem documentation spreadsheet might make the file unreadable.

Also, consider collecting the output of specific commands. At a minimum, you should collect the output of the following commands:

- `svcinfo lsfabric`
- `svcinfo lssystem`
- `svcinfo lsmdisk`
- `svcinfo lsmdiskgrp`
- `svcinfo lsvdisk`
- `svcinfo lshost`
- `svcinfo lshostvdiskmap`

Note: Most CLI commands that are shown here work without the `svcinfo` prefix; however, some commands might not work with only the short name and therefore require the `svcinfo` prefix to be added.

Import the commands into the master spreadsheet, preferably with the output from each command on a separate sheet.

One way to automate either task is to first create a batch file (Windows), shell script (UNIX or Linux) or playbook (Ansible) that collects and stores this information. Then, use spreadsheet macros to import the collected data into your IBM FlashSystem documentation spreadsheet.

When you are gathering IBM FlashSystem information, consider the following preferred practices:

- If you are collecting the output of specific commands, use the **-delim** option of these commands to make their output delimited by a character other than tab, such as comma, colon, or exclamation mark. You can import the temporary files into your spreadsheet in comma-separated values (CSV) format, specifying the same delimiter.

> **Note:** It is important to use a delimiter that is not already part of the output of the command. Commas can be used if the output is a particular type of list. Colons might be used for special fields, such as IPv6 addresses, WWPNs, or ISCSI names.

- If you are collecting the output of specific commands, save the output to temporary files. To make your spreadsheet macros simpler, you might want to preprocess the temporary files and remove any “garbage” or undesired lines or columns. With UNIX or Linux, you can use commands such as **`grep`**, **`sed`**, and **`awk`**. Freeware software is available for Windows with the same commands, or you can use any batch text editor tool.

The objective is to fully automate this procedure so you can schedule it to regularly run automatically. Make the resulting spreadsheet easy to consult and have it contain only the information that you use frequently. The automated collection and storage of configuration and support data (which is typically more extensive and difficult to use) are described in 10.14.7, “Automated support data collection” on page 516.

10.14.4 Storage documentation

You must generate documentation of your back-end storage controllers after configuration. Then, you can update the documentation when these controllers receive hardware or code updates. As such, there is little point to automating this back-end storage controller documentation. The same applies to the IBM FlashSystem internal drives and enclosures.

Any portion of your external storage controllers that is used outside the IBM FlashSystem solution might have its configuration changed frequently. In this case, see your back-end storage controller documentation for more information about how to gather and store the information that you need.

Fully allocate all of the available space in any of the optional external storage controllers that you might use as more back-end to the IBM FlashSystem solution. This way, you can perform all your disk storage management tasks by using the IBM FlashSystem user interface.

10.14.5 Technical support information

If you must open a technical support incident for your storage and SAN components, create and keep available a spreadsheet with all relevant information for all storage administrators. This spreadsheet should include the following information:

- Hardware information:
  - Vendor, machine and model number, serial number (example: `IBM 9848-AF8 S/N 7812345`)
  - Configuration, if applicable
  - Current code level
- Physical location:
  - Data center, including the complete street address and phone number
  - Equipment physical location (room number, floor, tile location, and rack number)
  - Vendor’s security access information or procedure, if applicable
  - Onsite person’s contact name and phone or page number

- Support contract information:
  - Vendor contact phone numbers and website
  - Customer's contact name and phone or page number
  - User ID to the support website, if applicable
  - Do not store the password in the spreadsheet under any circumstances.
  - Support contract number and expiration date

By keeping this data on a spreadsheet, storage administrators have all the information that they need to complete a web support request form or to provide to a vendor's call support representative. Typically, you are asked first for a brief description of the problem and then asked later for a detailed description and support data collection.

10.14.6 Tracking incident and change tickets

If your organization uses an incident and change management and tracking tool (such as IBM Tivoli Service Request Manager®), you or the storage administration team might need to develop proficiency in its use for several reasons:

- If your storage and SAN equipment are not configured to send SNMP traps to this incident management tool, you should manually open incidents whenever an error is detected.
- The IBM FlashSystem has the ability to be managed by the IBM Storage Insights (SI) tool that is available free of charge to owners of IBM storage systems. The SI tool allows you to monitor all the IBM storage devices information on SI. For more information, see Chapter 9, "Implementing a storage monitoring system" on page 387.
- Disk storage allocation and deallocation and SAN zoning configuration modifications should be handled under properly submitted and approved change requests.
- If you are handling a problem yourself, or calling your vendor's technical support desk, you might need to produce a list of the changes that you recently implemented in your SAN or that occurred since the documentation reports were last produced or updated.

When you use incident and change management tracking tools, adhere to the following guidelines for IBM FlashSystem 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 mobile phones or pagers of the storage administrators on duty or on call.
- Discuss within your organization the risk classification that a storage allocation or deallocation change request 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 requests, such as storage allocation or SAN zoning modification, to facilitate and speed up their submission.
- Do not open change requests in advance to replace failed, redundant, hot-pluggable parts, such as disk drive modules (DDMs) in storage controllers with hot spares, or SFPs in SAN switches or servers with path redundancy.

Typically, these fixes do not change anything in your SAN storage topology or configuration, and do not cause any more service disruption or degradation than you already had when the part failed. Handle these fixes within the associated incident ticket because it might take longer to replace the part if you need to submit, schedule, and approve a nonemergency change request.

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.

10.14.7 Automated support data collection

In addition to the easier-to-use documentation of your IBM FlashSystem and SAN Storage environment, collect and store for some time the configuration files and technical support data collection for all your SAN equipment.

For IBM FlashSystem, this information includes `snap` data. For other equipment, see the related documentation for more information about how to gather and store the support data that you might need.

You can create procedures that automatically create and store this data on scheduled dates, delete old data, or transfer the data to tape.

There is also the possibility to use IBM Storage Insights to create support tickets and then attach the snap data to this record from within the SI GUI. For more information, see Chapter 11, “Troubleshooting and diagnostics” on page 519.

10.14.8 Subscribing to IBM FlashSystem support

Subscribing to IBM FlashSystem 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 this web page.

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.

Sign up and tailor the requests and alerts you want to receive. For example, type **IBM FlashSystem 9200** in the Product lookup text box and then click **Subscribe** to subscribe to FlashSystem 9200 notifications, as shown in Figure 10-23.

Figure 10-23 Creating a subscription to IBM FlashSystem 9200 notifications

11

Troubleshooting and diagnostics

This chapter provides information to start troubleshooting common problems that can occur in IBM FlashSystem environment. It describes situations that are related to IBM FlashSystem, the SAN environment, optional external storage subsystems, and hosts. It also explains how to collect the necessary problem determination data.

This chapter includes the following topics:

11.1 Starting troubleshooting

Troubleshooting is a systematic approach to solving a problem. The goal of troubleshooting or problem determination is to understand why something does not work as expected and find a resolution. Hence, the first step is to describe the problem as accurately as possible, then perform log collection from all the involved products of the solution as soon as the problem is reported. An effective problem report ideally should describe the expected behavior, the actual behavior, and, if possible, how to reproduce the behavior.

The following questions help define the problem for effective troubleshooting.

- What are the symptoms of the problem?
  - What is reporting the problem?
  - What are the error codes and messages?
  - What is the business impact of the problem?
  - Where does the problem occur?
  - Is the problem specific to one or multiple hosts, one or both nodes?
  - Is the current environment and configuration supported?
- When does the problem occur?
  - Does the problem happen only at a certain time of day or night?
  - How often does the problem happen?
  - What sequence of events leads up to the time that the problem is reported?
  - Does the problem happen after an environment change, such as upgrading or installing software or hardware?
- Under which conditions does the problem occur?
  - Does the problem always occur when the same task is being performed?
  - Does a certain sequence of events need to occur for the problem to surface?
  - Do any other applications fail at the same time?
- Can the problem be reproduced?
  - Can the problem be re-created on a test system?
  - Are multiple users or applications encountering the same type of problem?
  - Can the problem be re-created by running a single command, a set of commands, or a particular application, or a stand-alone application?

Collecting log files close to the time of the incident and providing an accurate timeline are critical for effective troubleshooting.

11.1.1 Using the GUI

IBM FlashSystem graphical user interface (GUI) is a good starting point for your troubleshooting. It has two icons at the top that can be accessed from any window of the GUI.

As shown in Figure 11-1, the first icon shows IBM FlashSystem events, such as an error or a warning, and the second icon shows suggested tasks and background tasks that are running, or that were recently completed.

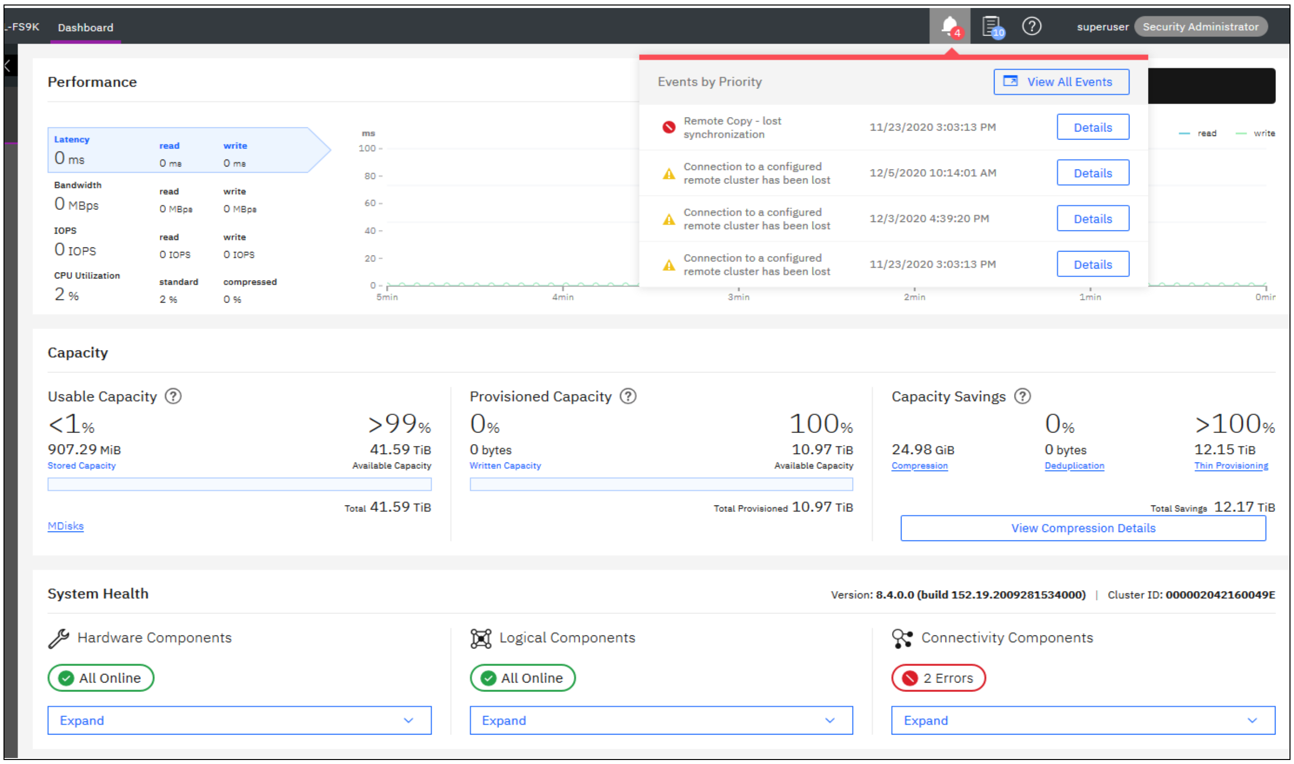


Figure 11-1 Events icon in GUI

The Dashboard provides an at-a-glance look into the condition of the system and notification of any critical issues that require immediate action. It contains sections for performance, capacity, and system health that provide an overall understanding of what is happening on the system.

Figure 11-2 shows the Dashboard window that displays the system health windows.

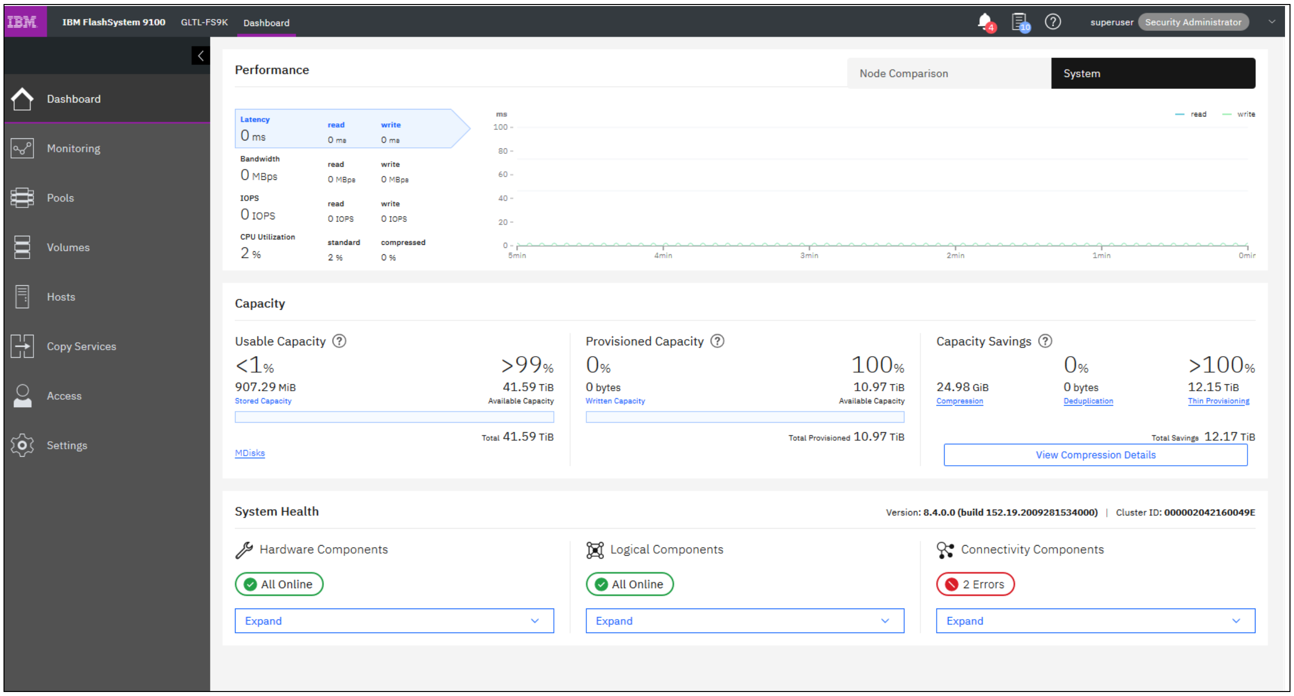


Figure 11-2 Dashboard showing system health

The System Health section in the bottom part of the Dashboard provides information about the health status of hardware, and logical and connectivity components.

If you click **Expand** in each of these categories, the status of individual components is shown (see Figure 11-3). You can also click **More Details**, which takes you to the window that is related to that specific component, or shows you more information about it.

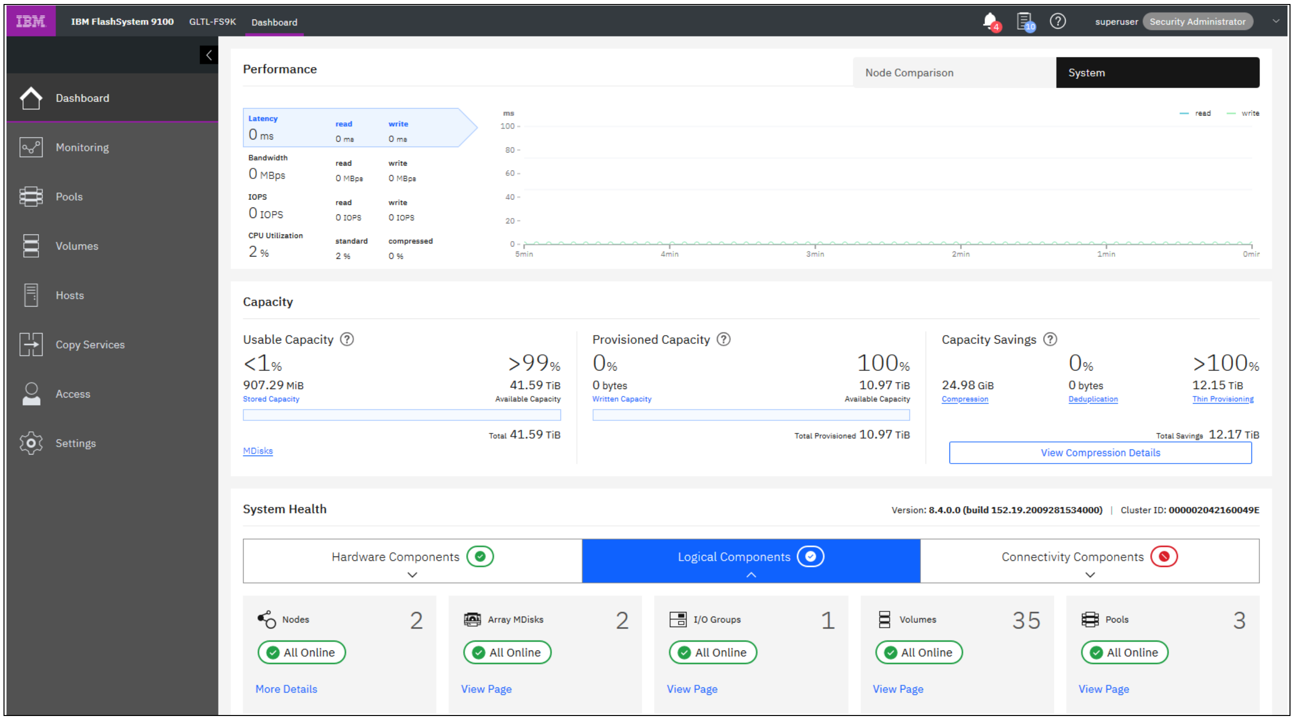


Figure 11-3 System Health expanded section in dashboard

For more information about the components in each category and IBM FlashSystem troubleshooting, see this IBM Documentation web page.

11.1.2 Recommended actions and fix procedure

The fix procedures were carefully designed to assist users to fix the problem without doing harm. When many unfixed error codes are in the event log, the management GUI provides a way to run the next recommended fix procedure. Therefore, the first step in troubleshooting is to run the fix procedures on the error codes in the event log.

These messages and codes provide reference information about configuration events and error event codes when a service action is required. The Cluster Error Code (CEC) is visible in the cluster event log; Node Error Code (NEC) is visible in node status in the service assistant GUI.

A cluster might encounter the following types of failure recoveries because of various conditions:

- Node assert (warmstart or Tier1/T1 recovery) is reported as CEC 2030.
- Cluster recovery (Tier2/T2 recovery) is reported as CEC 1001.
- System recovery (Tier3/T3 recovery) is required when all nodes of the clustered system report NEC 550/578.
- System restore (Tier4/T4 recovery) is to restore the cluster to a point where it can be used to restore from an off-cluster backup (be used by IBM Support only).

For more information about messages and codes, see this IBM Documentation web page.

The **Monitoring** → **Events** window shows information messages, warnings, and issues about the IBM FlashSystem. Therefore, this area is a good place to check the problems in the system.

Use the **Recommended Actions** filter to display the most important events that must be fixed.

If an important issue that must be fixed, the **Run Fix** button is available in the upper left with an error message that indicates which event must be fixed as soon as possible. This fix procedure assists you to resolve problems in IBM FlashSystem. It analyzes the system, provides more information about the problem, suggests actions to be taken with steps to be followed, and finally checks to see whether the problem is resolved.

Always use the fix procedures to resolve errors that are reported by the system, such as system configuration problems or hardware failures.

Note: IBM FlashSystem detects and reports error messages; however, many events can be triggered by external storage subsystems or the SAN.

Figure 11-4 shows the window with the Run Fix button enabled.

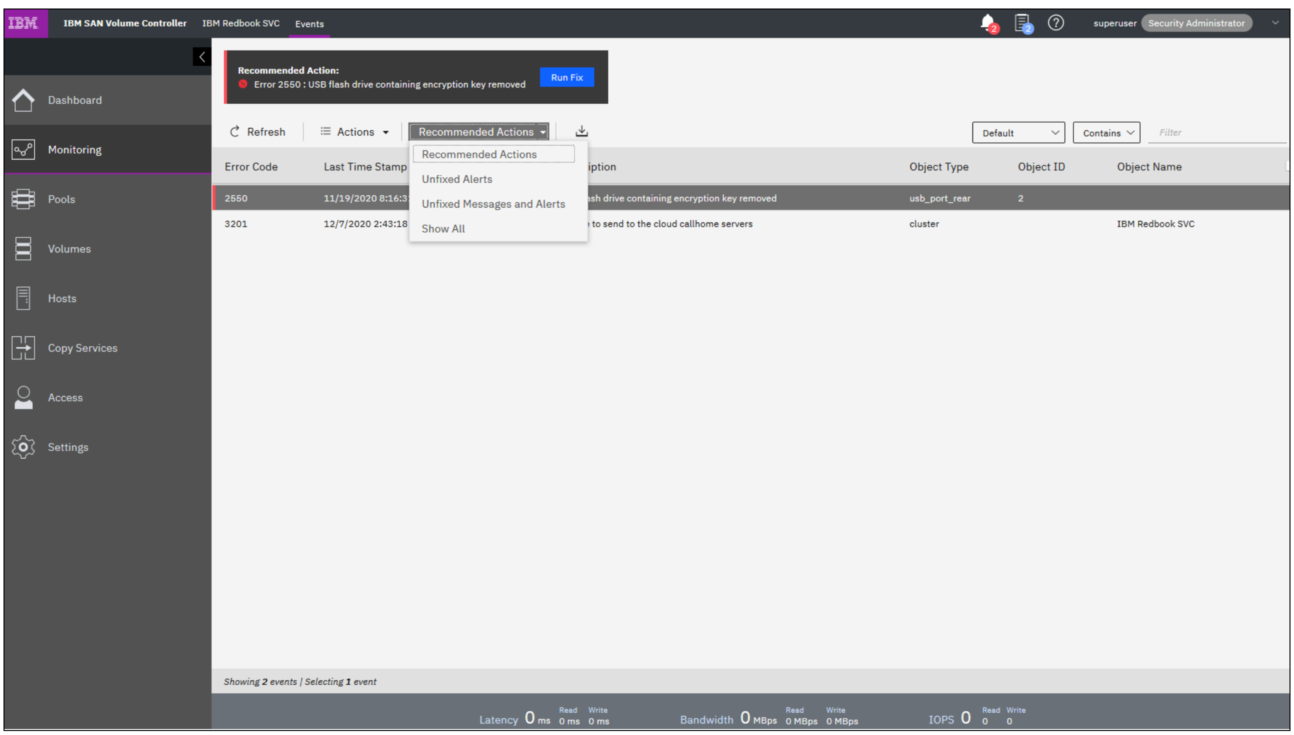


Figure 11-4 Events window

Resolve alerts in a timely manner: When an issue or a potential issue is reported, resolve it as quickly as possible to minimize its impact and potentially avoid more serious problems with your system.

To obtain more information about any event, select an event in the table, and click **Actions** → **Properties**. You can also access **Run Fix Procedure** and properties by right-clicking an event.

More information about it is displayed in the Properties and Sense Data window for the specific event, as shown in Figure 11-5. You can review and click **Run Fix** to run the fix procedure.

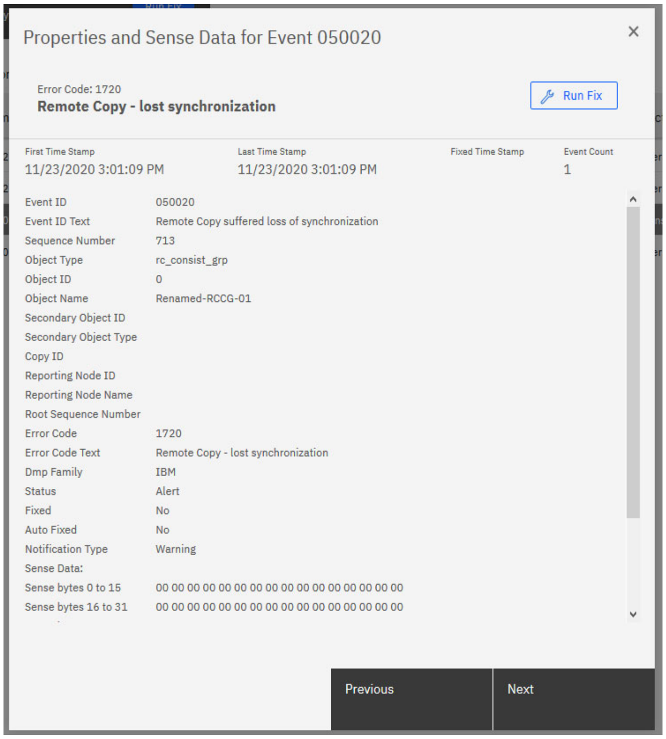


Figure 11-5 Properties and sense data for event window

Tip: In the Properties and Sense Data for Event window, use the **Previous** and **Next** buttons to move between events.

Another common practice is to use the IBM Spectrum Virtualize command-line interface (CLI) to find issues and resolve them. You also can use the 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 the commands when 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 the commands when any type of change that is related to the communication between IBM Spectrum Virtualize and back-end storage subsystem occurs (such as back-end storage is configured or a zoning change occurs). This process ensures that IBM Spectrum Virtualize recognizes the changes.

Common error-recovery involves the following IBM Spectrum Virtualize CLI commands:

- **`lscontroller` and `lsmdisk`**

  Provides status of all controllers and MDisks.

- **`detectmdisk`**

  Discovers the changes in the back-end.

- `lscontroller <controller_id_or_name>`

  Checks the controller that was causing the issue and verifies that all the WWPNs are listed as you expect. It also checks that the `path_counts` are distributed evenly across the WWPNs.

- `lsmdisk`

  Determines whether all MDisks are online.

> **Note:** When an issue is resolved by using the CLI, verify that the error disappears from **Monitoring** → **Events** window. If not, make sure that the error was fixed, and if so, manually mark the error as fixed.

11.2 Collecting diagnostic data

Data collection and problem isolation in an IT environment are sometimes difficult tasks. In the following section, the essential steps that are needed to collect debug data to find and isolate problems in an IBM FlashSystem environment are described.

11.2.1 IBM FlashSystem data collection

When a problem occurs with an IBM FlashSystem and you must open a case with IBM support, you must provide the support packages for the device. The support packages are collected and uploaded to the IBM Support center automatically by using IBM FlashSystem, or downloading the package from the device and manually upload to IBM.

The easiest method is to automatically upload the support packages from IBM FlashSystem by using the GUI or CLI.

You can also use the new IBM Storage Insights application to upload the log data, as described in 11.6.4, “Updating a support ticket” on page 562.

Collecting data by using the GUI

Complete the following steps to collect data by using the GUI:

1. Click **Settings** → **Support** → **Support Package**. Both options to collect and upload support packages are available
2. To automatically the support packages, click **Upload Support Package**.
3. In the pop-up window, enter the IBM Salesforce case number (`TS00`*`xxxxx`*) and the type of support package to upload to the IBM Support center. The Snap Type 4 can be used to collect standard logs and generate a statesave on each node of the system.

The Upload Support Package window is shown in Figure 11-6.

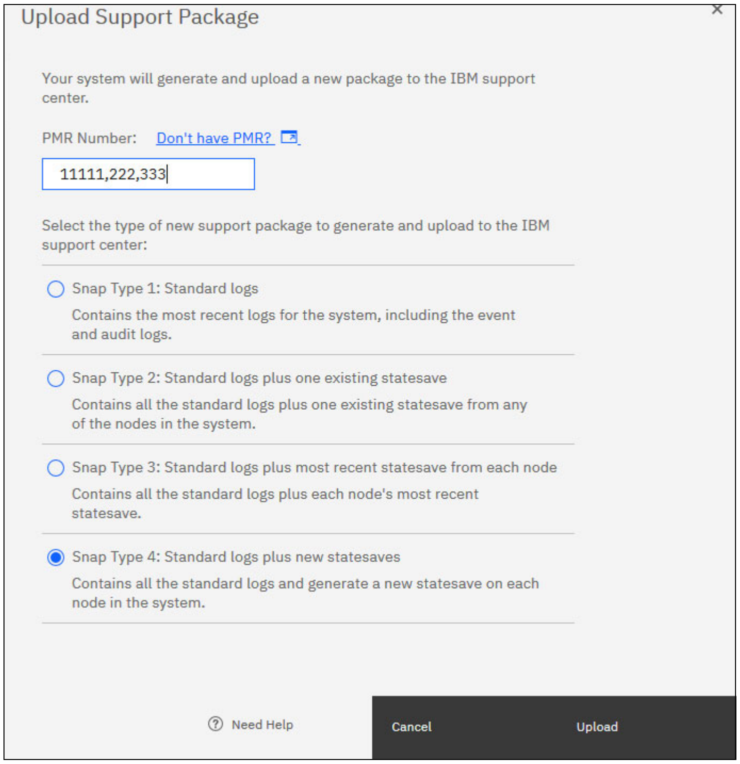


Figure 11-6 Upload Support Package window

For more information about the required support package that is most suitable to diagnose different type of issues, see this IBM Documentation web page.

Consider the following points:

- For issues that are related to interoperability with hosts or storage, use Snap Type 4.
- For critical performance issues, collect Option 1 and then, collect Snap Type 4.
- For general performance issues, collect Snap Type 4.
- For issues that are related to replication (including 1920 errors), collect Snap Type 4 from both systems.
- For issues that are related to compressed volumes, collect Snap Type 4.
- For 2030, 1196 or 1195 errors collect Snap Type 4.
- For all other issues, collect Snap Type 4.

Collecting data by using the CLI

To use the CLI to collect the same type of support packages, a `livedump` of the system must be generated by using the **`svc_livedump`** command. Then, the log files and newly generated dumps are uploaded by using the **`svc_snap`** command, as shown in Example 11-1 on page 528. To verify whether the support package was successfully uploaded, use the **`sainfo lscmdstatus`** command (`XXXXX,YYY,ZZZ` is the PMR number).

Example 11-1 The svc_livedump command

```
IBM_FlashSystem:FLASHPFE95:superuser>svc_livedump -nodes all -yes
Livedump - Fetching Node Configuration
Livedump - Checking for dependent vdisks
Livedump - Check Node status
Livedump - Prepare specified nodes  - this may take some time...
Livedump - Prepare node 1
Livedump - Prepare node 2
Livedump - Trigger specified nodes
Livedump - Triggering livedump on node 1
Livedump - Triggering livedump on node 2
Livedump - Waiting for livedumps to complete dumping on nodes 1,2
Livedump - Waiting for livedumps to complete dumping on nodes 2
Livedump - Successfully captured livedumps on nodes 1,2

IBM_FlashSystem:FLASHPFE95:superuser>svc_snap upload pmr=TS00XXXXX gui3
Collecting data
Packaging files
Snap data collected in /dumps/snap.ABCDEFG.171128.223133.tgz

IBM_FlashSystem:FLASHPFE95:superuser>sainfo lscmdstatus
last_command satask supportupload -pmr TS00xxxxx -filename
/dumps/snap.ABCDEFG.171128.223133.tgz
last_command_status CMMVC8044E Command completed successfully.
T3_status
T3_status_data
cpfiles_status Complete
cpfiles_status_data Copied 1 of 1
snap_status Complete
snap_filename /dumps/snap.ABCDEFG.171128.223133.tgz
installcanistersoftware_status
supportupload_status Complete
supportupload_status_data [PMR=TS00xxxxx] Upload complete
supportupload_progress_percent 0
supportupload_throughput_KBps 0
supportupload_filename /dumps/snap.ABCDEFG.171128.223133.tgz
downloadsoftware_status
downloadsoftware_status_data
downloadsoftware_progress_percent 0
downloadsoftware_throughput_KBps 0
downloadsoftware_size
IBM_FlashSystem:FLASHPFE95:superuser>
```

11.2.2 Host multipath software data collection

If a problem occurs that is related to host communication with IBM FlashSystem, collecting data from hosts and multipath software is useful.

If the systems still use SDD multipathing modules, note SDDPCM for AIX provides the **sddpcmgetdata** script to collect information that is used for problem determination. This script creates a `.tar` file in the directory with the current date and time as a part of the file name.

When you suspect an issue with SDDPCM exists, it is essential to run this script and send this `.tar` file to IBM Support.

SDDDSM for Windows hosts also include a utility to collect information for problem determination. The **sddgetdata.bat** tool creates a CAB file in the installation directory with the current date and time as part of the file name. The CAB file includes the following information:

- SystemInfo
- HKLM \SYSTEM\CurrentControlSet, HKLM\HARDWARE\DEVICEMAP, and HKLM\Cluster output from the registry
- SDDDSM directory contents
- HBA details
- Datapath outputs
- Pathtest trace
- SDDSRV logs
- Cluster logs
- System disks and paths

The execution and output of **sddgetdata.bat** tool is shown in Example 11-2.

Example 11-2 The sddgetdata.bat tool

```
C:\Program Files\IBM\SDDDSM>sddgetdata.bat
Collecting SDD trace Data

Flushing SDD kernel logs

SDD logs flushed

Collecting datapath command outputs

Collecting System Information

Collecting SDD and SDDSrv logs

Collecting Most current driver trace

Please wait for 30 secs... Writing DETAILED driver trace to trace.out

Generating a CAB file for all the Logs

sdddata_WIN-IWG6VLJN3U3_20171129_151423.cab file generated

C:\Program Files\IBM\SDDDSM>
```

For more information about diagnostics for IBM SDD, see this IBM Support web page.

11.2.3 More data collection

Data collection methods vary by storage platform, SAN switch, and operating system.

For an issue in a SAN environment when it is not clear where the problem is occurring, you might need to collect data from several devices in the SAN.

The following basic information must be collected for each type of device:

- Hosts:
  - Operating system: Version and level
  - HBA: Driver and firmware level
  - Multipathing driver level
- SAN switches:
  - Hardware model
  - Software version
- Storage subsystems:
  - Hardware model
  - Software version

11.3 Common problems and isolation techniques

SANs, storage subsystems, and host systems can be complicated. They often consist of hundreds or thousands of disks, multiple redundant subsystem controllers, virtualization engines, and different types of SAN switches. All of these components must be configured, monitored, and managed correctly. If issues occur, administrators must know what to look for and where to look.

IBM FlashSystem features useful error logging mechanisms. It tracks its internal events and informs the user about issues in the SAN or storage subsystem. It also helps to isolate problems with the attached host systems. Therefore, by using these functions, administrators can easily locate any issue areas and take the necessary steps to fix any events.

In many cases, IBM FlashSystem and its service and maintenance features guide administrators directly, provide help, and suggest remedial action. Furthermore, IBM FlashSystem determines whether the problem still persists.

Another feature that helps administrators to isolate and identify issues that might be related to IBM FlashSystem is the ability of their nodes to maintain a database of other devices that communicate with the IBM FlashSystem device. Devices, such as hosts and optional back-end storages, are added or removed from the database as they start or stop communicating to IBM FlashSystem.

Although IBM FlashSystem node hardware and software events can be verified in the GUI or CLI, external events, such as failures in the SAN zoning configuration, hosts, and back-end storages are common. They also must have troubleshooting performed outside of IBM FlashSystem.

For example, a misconfiguration in the SAN zoning might lead to the IBM FlashSystem cluster not working correctly. This problem occurs because the IBM FlashSystem cluster nodes communicate with each other by using the Fibre Channel SAN fabrics.

In this case, check the following areas from an IBM FlashSystem perspective:

- The attached hosts. For more information, see 11.3.1, “Host problems” on page 531.
- The SAN. For more information, see 11.3.2, “SAN problems” on page 535.
- The optional attached storage subsystem. For more information, see 11.3.3, “Storage subsystem problems” on page 537.
- The local FC port masking. For more information, see 8.1.3, “Port masking” on page 368.

11.3.1 Host problems

From the host perspective, you can experience various situations that range from performance degradation to inaccessible disks. The first step in troubleshooting such issues is to check whether any potential interoperability issues exist.

Interoperability

When you experience events in the IBM FlashSystem environment, ensure that all components that comprise the storage infrastructure are interoperable. In an IBM FlashSystem environment, the IBM FlashSystem support matrix is the main source for this information. For the latest IBM FlashSystem support matrix, see IBM System Storage Interoperations Center (SSIC).

Although the latest IBM FlashSystem code level is supported to run on older host bus adapters (HBAs), storage subsystem drivers, and code levels, use the latest tested levels for best results.

After interoperability is verified, check the configuration of the host on the IBM FlashSystem side. The *Hosts* window in the GUI or the following CLI commands can be used to start a verification in host-related issues:

- **lshost**

  Checks the host’s status. If the status is online, the host ports are online in both nodes of an I/O group. If the status is offline, the host ports are offline in both nodes of an I/O group. If the status is inactive, it means that the host has volumes that are mapped to it, but all of its ports did not receive Small Computer System Interface (SCSI) commands in the last 5 minutes. Also, if the status is degraded, it means at least one (but not all) of the host ports are not online in at least one node of an I/O group. Example 11-3 shows the **lshost** command output.

  Example 11-3 lshost command

```
IBM_FlashSystem:FLASHPFE95:superuser>lshost
0  Win2K8    2          4           degraded
1  ESX_62_B  2          4           online
2  ESX_62_A  2          1           offline
3  Server127 2          1           degraded
```

- **lshost <host_id_or_name>**

  Shows more information about a specific host. It often is used when you must identify which host port is not online in an IBM FlashSystem node. Example 11-4 shows the **lshost <host_id_or_name>** command output.

Example 11-4 lshost <host_id_or_name> command

```
IBM_FlashSystem:FLASHPFE95:superuser>lshost Win2K8
id 0
name Win2K8
port_count 2
type generic
mask 1111111111111111111111111111111111111111111111111111111111111111
iogrp_count 4
status degraded
site_id
site_name
host_cluster_id
host_cluster_name
WWPN 100000051E0F81CD
node_logged_in_count 2
state active
WWPN 100000051E0F81CC
node_logged_in_count 0
state offline
```

- **lshostvdiskmap**

  Check that all volumes are mapped to the correct hosts. If a volume is not mapped correctly, create the necessary host mapping.

- **lsfabric -host <host_id_or_name>**

  Use this command with parameter **-host <host_id_or_name>** to display Fibre Channel connectivity between nodes and hosts. Example 11-5 shows the **lsfabric -host <host_id_or_name>** command output.

Example 11-5 lsfabric -host <host_id_or_name> command

```
IBM_FlashSystem:FLASHPFE95:superuser>lsfabric -host Win2K8
 remote_wwpn      remote_nportid id node_name local_wwpn       local_port
 local_nportid state    name        cluster_name type
 10000090FAB386A3 502100       3 node1    5005076810120230 2        540200
 inactive Win2K8 host
 10000090FAB386A3 502100       1 node2    5005076810120242 2        540000
 inactive Win2K8 host
```

To perform troubleshooting on the host side, check the following areas:

- Special software that you use
- Recent changes in the operating system (patching the operating system, an upgrade, and so on)
- Operating system version and maintenance or service pack level
- Multipathing type and driver level

- Host bus adapter model, firmware, and driver level
- Host bus adapter connectivity issues

Based on this list, the host administrator must check and correct any problems.

Hosts with higher queue depth can potentially overload shared storage ports. Therefore, it is recommended that you verify that the total of the queue depth of all hosts that are sharing a single target Fibre Channel port is limited to 2048. If any of the hosts have a queue depth of more than 128, that depth must be reviewed because queue-full conditions can lead to I/O errors and extended error recoveries.

For more information about managing hosts on IBM FlashSystem, see Chapter 8, “Configuring hosts” on page 367.

Apart from hardware-related situations, problems can exist in such areas as the operating system or the software that is used on the host. These problems normally are handled by the host administrator or the service provider of the host system. However, the multipathing driver that is installed on the host and its features can help to determine possible issues.

For example, a volume path issue is reported by SDD output on the host by using the **`datapath query adapter`** and **`datapath query device`** commands. The adapter in degraded state means that specific HBA on the server side cannot reach all the nodes in the I/O group to which the volumes are associated.

> **Note:** SDDDSM and SDDPCM reached EOS. Therefore, migrate SDDDSM to MSDSM on Windows platform and SDDPCM to AIXPCM on AIX/VIOS platforms, respectively.

For more information, see this IBM Support web page.

Faulty paths can be caused by hardware and software problems, such as the following examples:

- Hardware:
  - Faulty Small Form-factor Pluggable transceiver (SFP) on the host or SAN switch
  - Faulty fiber optic cables
  - Faulty HBAs
  - Faulty physical SAN ports within a switch can lead to replacement of entire switch
  - Contaminated SFPs/cable connectors
- Software:
  - A back-level multipathing driver
  - Obsolete HBA firmware or driver
  - Wrong zoning
  - Incorrect host-to-VDisk mapping

Based on field experience, it is recommended that you complete the following hardware checks first:

- Whether connection error indicators are lit on the host or SAN switch.
- Whether all of the parts are seated correctly. For example, cables are securely plugged in to the SFPs and the SFPs are plugged all the way into the switch port sockets.
- Ensure that fiber optic cables are not 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:

- Whether the HBA driver level and firmware level are at the preferred and supported levels.
- The multipathing driver level, and make sure that it is at the preferred and supported level.
- 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.
- The general SAN switch status and health for all switches in the fabric.

iSCSI/iSER configuration and performance issues

This section describes the Internet Small Computer Systems Interface (iSCSI) and iSCSI Extensions for RDMA (iSER) configuration and performance issues.

Link issues

If the Ethernet port link does not come online, check whether the SFP/cables and the port support auto-negotiation with the switch. This issue is especially true for SFPs, which support 25 G and higher because a mismatch might exist in Forward Error Correction (FEC) that might prevent a port to auto-negotiate.

Longer cables are exposed to more noise or interference (high Bit Error Ratio [BER]); therefore, they require more powerful error correction codes.

Two IEEE 802.3 FEC specifications are important. For an auto-negotiation issue, verify whether a compatibility issue exists with SFPs at both end points:

- Clause 74: Fire Code (FC-FEC) or BASE-R (BR-FEC) (16.4 dB loss specification).
- Clause 91: Reed-Solomon; that is, RS-FEC (22.4 dB loss specification)

Priority flow control

Priority flow control (PFC) is an Ethernet protocol that supports the ability to assign priorities to different types of traffic within the network. On most Data Center Bridging Capability Exchange protocol (DCBX) supported switches, verify whether Link Layer Discovery Protocol (LLDP) is enabled. The presence of a virtual local area network (VLAN) is a prerequisite for the configuration of PFC. It is recommended to set the priority tag 0 - 7.

A DCBX-enabled switch and a storage adapter exchange parameter that describe traffic classes and PFC capabilities.

In the IBM FlashSystem, Ethernet traffic is divided into the following Classes of Service based on feature use case:

- Host attachment (iSCSI/iSER)
- Back-end Storage (iSCSI)
- Node-to-node communication (Remote Direct Memory Access (RDMA) clustering)

If challenges occur as the PFC is configured, verify the following attributes to determine the issue:

- Configure IP/VLAN by using **`cfgportip`**.
- Configure class of service (COS) by using **`chsytsemethernet`**.
- Ensure that the priority tag is enabled on the switch.
- Ensure that **`lsportip`** output shows: `dcbx_state, pfc_enabled_tags`.
- Enhanced Transmission Selection (ETS) settings is recommended if a port is shared.

Standard network connectivity check

Verify that the required TCP/UDP ports are allowed in the network firewall. The following ports can be used for various host attachments:

- Software iSCSI requires TCP Port 3260.
- iSER/RoCE host requires 3260.
- iSER/iWRAP host requires TCP Port 860.

Verify that the IP addresses are reachable and the TCP ports are open.

iSCSI performance issues

In specific situations, the TCP/IP layer might attempt to combine several ACK responses into a single response to improve performance. However, that combination can negatively affect iSCSI read performance as the storage target waits for the response to arrive. This issue is observed when the application is single-threaded and has a low queue depth.

It is recommended to disable the TCPDelayedAck parameter on the host platforms to improve overall storage I/O performance. If the host platform does not provide a mechanism to disable TCPDelayedAck, verify whether a smaller "Max I/O Transfer Size" with more concurrency (queue depth >16) improves overall latency and bandwidth use for the specific host workload. In most Linux distributions, this Max I/O Transfer Size is controlled by the `max_sectors_kb` parameter with a suggested transfer size of 32 KB.

In addition, review network switch diagnostic data to evaluate packet drop or retransmission in the network. It is advisable to enable flow control or PFC to enhance the reliability of the network delivery system to avoid packet loss, which enhances storage performance.

11.3.2 SAN problems

It is not a difficult task to introduce IBM FlashSystem into your SAN environment and to use its virtualization functions. However, before you can use IBM FlashSystem in your environment, you must follow some basic rules. These rules are not complicated, but you can make mistakes that lead to accessibility issues or a reduction in the performance experienced.

Two types of SAN zones are needed to run IBM FlashSystem in your environment: A *host zone* and a *storage zone* for optional external-attached storage. In addition, you must have an IBM FlashSystem zone that contains all the IBM FlashSystem node ports of the IBM FlashSystem cluster. This IBM FlashSystem zone enables intra-cluster communication.

For more information and important points about setting up IBM FlashSystem in a SAN fabric environment, see Chapter 2, "Connecting IBM Spectrum Virtualize and IBM Storwize in storage area networks" on page 37.

Because IBM FlashSystem is a major component of the SAN and connects the host to the storage subsystem, check and monitor the SAN fabrics.

Some situations of performance degradation and buffer-to-buffer credit exhaustion can be caused by incorrect local FC port masking and remote FC port masking. To ensure healthy operation of your IBM FlashSystem, configure your local FC port masking and your remote FC port masking.

The ports that are intended to have only intra-cluster/node-to-node communication traffic must not have replication data or host or back-end data running on it. The ports that are intended to have only replication traffic must not have intra-cluster or node-to-node communication data or host or back-end data running on it.

Some situations can cause issues in the SAN fabric and SAN switches. Problems can be related to a hardware fault or to a software problem on the switch. The following hardware defects are normally the easiest problems to find:

- Switch power, fan, or cooling units
- Installed SFP modules
- Fiber optic cables

Software failures are more difficult to analyze. In most cases, you must collect data and involve IBM Support. However, before you take any other steps, check the installed code level for any known issues. Also, check whether a new code level is available that resolves the problem that you are experiencing.

The most common SAN issues often are related to zoning. For example, perhaps you chose the wrong WWPN for a host zone, such as when two IBM FlashSystem node ports must be zoned to one HBA with one port from each IBM FlashSystem node. As shown in Example 11-6, two ports are zoned that belong to the same node. Therefore, the result is that the host and its multipathing driver do not see all of the necessary paths.

Example 11-6 Incorrect WWPN zoning

```
zone:  Senegal_Win2k3_itsosvccl1_iogrp0_Zone
                50:05:07:68:10:20:37:dc
                50:05:07:68:10:40:37:dc
                20:00:00:e0:8b:89:cc:c2
```

The correct zoning must look like the zoning that is shown in Example 11-7.

Example 11-7 Correct WWPN zoning

```
zone:  Senegal_Win2k3_itsosvccl1_iogrp0_Zone
                50:05:07:68:10:40:37:e5
                50:05:07:68:10:40:37:dc
                20:00:00:e0:8b:89:cc:c2
```

The following IBM FlashSystem error codes are related to the SAN environment:

- Error 1060 - `Fibre Channel ports are not operational`
- Error 1220 - `A remote port is excluded`

A bottleneck is another common issue that is related to SAN switches. The bottleneck can be present in a port where a host, storage subsystem, or IBM Spectrum Virtualize device is connected, or in Inter-Switch Link (ISL) ports. The bottleneck can occur in some cases, such as when a device that is connected to the fabric is slow to process received frames or if a SAN switch port cannot transmit frames at a rate that is required by a device that is connected to the fabric.

These cases can slow down communication between devices in your SAN. To resolve this type of issue, refer to the SAN switch documentation or open a case with the vendor to investigate and identify what is causing the bottleneck and how fix it.

If you cannot fix the issue with these actions, use the method that is described in 11.2, “Collecting diagnostic data” on page 526, collect the SAN switch debugging data, and then, contact the vendor for assistance.

11.3.3 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 depending on the manufacturer. To support a stable environment, all subsystems must be correctly configured, following the respective preferred practices and with no existing issues.

Check the following areas if you experience a storage-subsystem-related issue:

- Storage subsystem configuration: Ensure that a valid configuration and preferred practices are applied to the subsystem.
- Storage subsystem node controllers: Check the health and configurable settings on the node controllers.
- Storage subsystem array: Check the state of the hardware, such as an FCM or SSD failures or enclosure alerts.
- Storage volumes: Ensure that the logical unit number (LUN) masking is correct.
- Host attachment ports: Check the status, configuration and connectivity to SAN switches.
- Layout and size of RAID arrays and LUNs: Performance and redundancy are contributing factors.

IBM FlashSystem features several CLI commands that you can use to check the status of the system and attached optional 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 IBM FlashSystem perspective:

- `lsmdisk`

  Check that all MDisks are online (not degraded or offline).
- `lsmdisk <MDisk_id_or_name>`

  Check several of the MDisks from each storage subsystem controller. Are they online? An example of the output from this command is shown in Example 11-8.

Example 11-8 Issuing a lsmdisk command

```
IBM_FlashSystem:FLASHPFE95:superuser>lsmdisk 0
id 0
name MDisk0
status online
mode array
MDisk_grp_id 0
MDisk_grp_name Pool0
capacity 198.2TB
quorum_index
block_size
controller_name
ctrl_type
ctrl_WWNN
controller_id
path_count
max_path_count
ctrl_LUN_#
UID
preferred_WWPN
active_WWPN
fast_write_state empty
```

```
raid_status online
raid_level raid6
redundancy 2
strip_size 256
spare_goal
spare_protection_min
balanced exact
tier tier0_flash
slow_write_priority latency
fabric_type
site_id
site_name
easy_tier_load
encrypt no
distributed yes
drive_class_id 0
drive_count 8
stripe_width 7
rebuild_areas_total 1
rebuild_areas_available 1
rebuild_areas_goal 1
dedupe no
preferred_iscsi_port_id
active_iscsi_port_id
replacement_date
over_provisioned yes
supports_unmap yes
provisioning_group_id 0
physical_capacity 85.87TB
physical_free_capacity 78.72TB
write_protected no
allocated_capacity 155.06TB
effective_used_capacity 16.58TB.

IBM_FlashSystem:FLASHPFE95:superuser>lsmdisk 1
id 1
name flash9h01_itsosvccl1_0
status online
mode managed
MDisk_grp_id 1
MDisk_grp_name Pool1
capacity 51.6TB
quorum_index
block_size 512
controller_name itsoflash9h01
ctrl_type 6
ctrl_WWNN 500507605E852080
controller_id 1
path_count 16
max_path_count 16
ctrl_LUN_# 0000000000000000
UID 6005076441b53004400000000000000100000000000000000000000000000000
preferred_WWPN
active_WWPN many
```

```
NOTE: lines removed for brevity
```

Example 11-8 on page 537 shows that for MDisk 1, the external storage controller has eight ports that are zoned to IBM FlashSystem, and IBM FlashSystem has two nodes (8 x 2= 16).

- **`lsvdisk`**

  Check that all volumes are online (not degraded or offline). If the volumes are degraded, are there stopped FlashCopy jobs? Restart stopped FlashCopy jobs or seek IBM FlashSystem support guidance.

- **`lsfabric`**

  Use this command with the various options, such as **`-controller`** ***`controllerid`***. Also, check different parts of the IBM FlashSystem configuration to ensure that multiple paths are available from each IBM FlashSystem node port to an attached host or controller. Confirm that IBM FlashSystem node port WWPNs are also consistently connected to the optional external back-end storage.

Determining the number of paths to an external storage subsystem

By using IBM FlashSystem CLI commands, the total number of paths to an optional external storage subsystem can be determined. To determine the value of the available paths, use the following formula:

```
Number of MDisks x Number of FlashSystem nodes per Cluster = Number of paths
MDisk_link_count x Number of FlashSystem nodes per Cluster = Sum of path_count
```

Example 11-9 shows how to obtain this information by using the **`lscontroller <controllerid>`** and **`svcinfo lsnode`** commands.

Example 11-9 Output of the svcinfo lscontroller command

```
IBM_FlashSystem:FLASHPFE95:superuser>lscontroller 1
id 1
controller_name itsof9h01
WWNN 500507605E852080
MDisk_link_count 16
max_MDisk_link_count 16
degraded no
vendor_id IBM
product_id_low FlashSys
product_id_high tem-9840
product_revision 1430
ctrl_s/n 01106d4c0110-0000-0
allow_quorum yes
fabric_type fc
site_id
site_name
WWPN 500507605E8520B1
path_count 32
max_path_count 32
WWPN 500507605E8520A1
path_count 32
max_path_count 64
WWPN 500507605E852081
path_count 32
max_path_count 64
WWPN 500507605E852091
```

```
path_count 32
max_path_count 64
WWPN 500507605E8520B2
path_count 32
max_path_count 64
WWPN 500507605E8520A2
path_count 32
max_path_count 64
WWPN 500507605E852082
path_count 32
max_path_count 64
WWPN 500507605E852092
path_count 32
max_path_count 64


IBM_FlashSystem:FLASHPFE95:superuser>svcinfo lsnode

id name UPS_serial_number WWNN             status IO_group_id IO_group_name
config_node UPS_unique_id hardware iscsi_name
iscsi_alias panel_name enclosure_id canister_id enclosure_serial_number site_id
site_name

1 node1              500507681000000A online 0        io_grp0     no
AF8     iqn.1986-03.com.ibm:2145.flashpfe95.node1           01-2      1
2          F313150

2 node2              5005076810000009 online 0        io_grp0     yes
AF8     iqn.1986-03.com.ibm:2145.flashpfe95.node2           01-1      1
1          F313150
IBM_FlashSystem:FLASHPFE95:superuser>
```

Example 11-9 on page 539 shows that 16 MDisks are present for the external storage subsystem controller with ID 1, and two IBM FlashSystem nodes are in the cluster. In this example, the `path_count` is 16 x 2 = 32.

IBM FlashSystem features useful tools for finding and analyzing optional back-end storage subsystem issues because it includes a monitoring and logging mechanism.

Typical events for storage subsystem controllers include incorrect configuration, which results in a `1625 - Incorrect disk controller configuration` error code. Other issues that are related to the storage subsystem include failures that point 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 specific reason for being issued. Therefore, you must check many areas for issues, not just the storage subsystem.

To determine the root cause of a problem, complete the following steps:

1. Check the Recommended Actions window by clicking **Monitoring** → **Events**.
2. Check the attached storage subsystem for misconfigurations or failures:
   a. Independent of the type of storage subsystem, first check whether the system includes any unfixed errors. Use the service or maintenance features that are provided with the storage subsystem to fix these issues.

b. Check whether the volume mapping is correct. The storage subsystem LUNs must be mapped to a host object with IBM FlashSystem ports. For more information about the IBM FlashSystem restrictions for optional back-end storage subsystems, see this IBM Support web page.

   If you need to identify which of the externally attached MDisks has which corresponding LUN ID, run the IBM FlashSystem **lsmdisk** CLI command, as shown in Example 11-10. This command also shows to which storage subsystem a specific MDisk belongs (the controller ID).

   Example 11-10 Determining the ID for the MDisk

```
IBM_FlashSystem:FLASHPFE95:superuser>lsmdisk
id name   status mode  MDisk_grp_id MDisk_grp_name capacity ctrl_LUN_#
controller_name UID tier       encrypt site_id site_name distributed dedupe
over_provisioned supports_unmap
0  MDisk0 online array 0         Pool0       198.2TB
tier0_flash no                         yes         no     yes
yes
0          MDisk1 online      managed     0         MDG-1
600.0GB        0000000000000000 controller0
600a0b800017423300000059469cf84500000000000000000000000000000000
2          MDisk2 online      managed     0         MDG-1
70.9GB         0000000000000002 controller0
600a0b800017443100000096469cf0e800000000000000000000000000000000
```

3. Check the SAN environment for switch problems or zoning failures.

   Make sure the zones are correctly configured, and that the zoneset is activated. The zones that allow communication between the storage subsystem and the IBM FlashSystem device must contain WWPNs of the storage subsystem and WWPNs of IBM FlashSystem.

4. Collect all support data and contact IBM Support.

   Collect the support data for the involved SAN, IBM FlashSystem, or optional external storage systems as described in 11.2, “Collecting diagnostic data” on page 526.

11.3.4 Native IP replication problems

The native IP replication feature uses the following TCP/IP ports for remote cluster path discovery and data transfer:

- IP Partnership management IP communication: TCP Port 3260
- IP Partnership data path connections: TCP Port 3265

If a connectivity issue exists between the cluster in the management communication path, the cluster reports error code `2021: Partner cluster IP address unreachable`. However, when a connectivity issue exists in the data path, the cluster reports error code `2020: IP Remote Copy link unavailable`.

If the IP addresses are reachable and TCP ports are open, verify whether the end-to-end network supports a Maximum Transmission Unit (MTU) of 1500 bytes without packet fragmentation. When an external host-based ping utility is used to validate end-to-end MTU support, use the “do not fragment” qualifier.

Fix the network path so that traffic can flow correctly. After the connection is made, the error auto-corrects.

The network quality of service largely influences the effective bandwidth use of the dedicated link between the cluster. Bandwidth use is inversely proportional to round-trip time (RTT) and rate of packet drop or retransmission in the network.

For standard block traffic, a packet drop or retransmission of 0.5% or more can lead to unacceptable use of the available bandwidth. Work with the network team to investigate over-subscription or other quality-of-service of the link, with an objective of having the lowest possible (less than 0.1%) packet-drop percentage.

11.3.5 Remote Direct Memory Access-based clustering

RDMA technology supports zero-copy networking, which makes it possible to read data directly from the main memory of one computer and write that data directly to the main memory of another computer. This technology bypasses the CPU intervention while processing the I/O leading to lower latency and a faster rate of data transfer.

IBM FlashSystem Cluster can be formed by using RDMA-capable NICs that use RoCE or iWARP technology. Consider the following points:

- Inter-node Ethernet connectivity can be done over identical ports only; such ports must be connected within the same switching fabric.
- If the cluster is to be created without any ISL (up to 300 meters [984 feet]), deploy Independent (isolated) switches.
- If the cluster is to be created on short-distance ISL (up to 10 km [6.2 miles]), provision as many ISLs between switches as RDMA-capable cluster ports.
- For long-distance ISL (up to 100 km [62 miles]) DWDM and CWDM methods are applicable for L2 networks. Packet switched or VXLAN methods are deployed for L3 network as this equipment comes with deeper buffer “pockets”.

Following Ports must be opened in the firewall for IP-based RDMA clustering

- TCP 4791, 21451, 21452, and 21455
- UDP 4791, 21451, 21452, and 21455

The first step to review whether the node IP address is reachable and verify the required TCP/UDP ports are accessible in both directions. The following CLI output can be helpful to find the reason for connectivity error:

```
sainfo lsnodeipconnectivity
```

11.3.6 Advanced Copy services problems

Performance of a specific storage feature or overall storage subsystem is generally interlinked, meaning that a bottleneck in one software or hardware layer can propagate to other layers. Therefore, problem isolation is a critical part of performance analysis.

The first thing to check is whether any unfixed events exist that require attention. After the fix procedure is followed to correct the alerts, the next step is to check the audit log to determine whether any activity exists that can trigger the performance issue. If that information correlates, more analysis can be done to check whether that specific feature is used.

The most common root causes for performance issues are SAN congestion, configuration changes, incorrect sizing/estimation of advanced copy services (replication, FlashCopy, volume mirroring), or I/O load change, because of hardware component failure.

Remote Copy

Disturbances in the SAN or wide area network (WAN) can cause congestion and packet drop, which can affect Metro Mirror (MM) or Global Mirror (GM) traffic. Because host I/O latency depends on MM or GM I/O completion to the remote cluster, a host can experience high latency. Based on various parameters, replication can be operatively stopped to protect host.

The following conditions can affect GM/MM:

- Network congestion or fluctuation. Fix the network. Also, verify that port masking is enabled so that the congestion in replication ports does not affect clustering or host or storage ports.
- Overload of secondary/primary cluster. Monitor and throttle the host that causes the condition.
- High background copy rate, which leaves less bandwidth to replicate foreground host I/O. Adjust the background copy rate so that the link does not become oversubscribed.
- A large Global Mirror with Change Volumes (GMCV) consistency group can introduce hundreds of milliseconds of pause when the replication cycle starts. Reduce the number of relationships in a consistency group if the observed I/O pause is unacceptable.

HyperSwap

Verify that the link between the sites is stable and has enough bandwidth to replicate the peak workload. Also, check whether a volume must frequently change the replication direction from one site to other. This issue occurs when a specific volume is being written by hosts from both the sites. Evaluate if this issue can be avoided to reduce frequent direction change. Ignore it if the solution is designed consider active/active access.

If a single volume resynchronization between the sites takes a very long time, review the partnership `link_bandwidth_mbits` and `per relationship_bandwidth_limit` parameters.

FlashCopy

Consider the following points for FlashCopy troubleshooting:

- Verify that the preferred node of FlashCopy source and target volumes is the same to avoid excessive internode communications.
- High background copy rate and clean rate of FlashCopy relations can cause back-end overload.
- Port saturation or node saturation. Review if the values are correctly sized.
- Check the number of Fibre Channel relationships in any FlashCopy consistency group. The larger the number of relationships, the higher the I/O pause time (Peak I/O Latency) when the CG starts.
- If the host I/O pattern is small and random, evaluate whether reducing the FlashCopy grain size to 64 KB provides any improvement in latency compared to the default grain size of 256 KB.

Compression

Compress a volume if the data is compressible. No benefit is gained by compressing a volume where compression saving is less than 25% because that can reduce the overall performance of the Random Access Compression Engine (RACE).

If the I/O access pattern is sequential, that volume might not be a suitable candidate for RACE. Use the Comprestimator or Data Reduction Estimation Tool to size the workload.

Volume mirroring

Write-performance of the mirrored volumes is dictated by the slowest copy. Reads are served from the Copy0 of the volume (in the case of a stretched cluster topology, both the copies can serve reads, which are dictated by the host site attribute). Therefore, size the solution accordingly.

The mirroring layer maintains a bitmap copy on the quorum device; therefore, any unavailability of the quorum takes the mirroring volumes offline. Similarly, slow access to the quorum also can affect the performance of mirroring volumes.

Data reduction pools

Data reduction pools (DRPs) internally implement a log structured array (LSA), which means that writes (new or over-writes or updates) always allocate newer storage blocks. The older blocks (with invalid data) are marked for garbage collection later.

The garbage collection process is designed to defer the work as much as possible because the more it is deferred, the higher the chance of having to move only a small amount of valid data from the block to make that block available it to the free pool. However, when the pool reaches more than 85% of its allocated capacity, garbage collection must speed up and to move valid data more aggressively to make space available sooner. This issue might lead to increased latency because of increased CPU use and load on the back-end. Therefore, it is recommended to manage storage provisioning to avoid such scenarios.

Users are encouraged to pay close attention to any GUI notifications and use best practices for managing physical space. Use data reduction only at one layer (at the virtualization layer or the back-end storage or drives) because no benefit is realized by compressing and deduplicating the same data twice.

Because encrypted data cannot be compressed, data reduction must be done before the data is encrypted. Correct sizing is important to get the best of performance from data reduction; therefore, use data reduction estimation tools to evaluate system performance and space saving.

IBM FlashSystem uses the following types of data reduction techniques:

- IBM FlashSystem using the FCM NVMe drives have built-in hardware compression.
- IBM FlashSystem using industry-standard NVMe drives rely on the Spectrum Virtualize software and DRP pools to deliver data reduction.

For more information about DRPs, see *Introduction and Implementation of Data Reduction Pools and Deduplication*, SG24-8430.

11.3.7 Health status during upgrade

It is important to understand that during the software upgrade process, alerts that indicate the system is not healthy are reported. This issue is a normal behavior because the IBM FlashSystem node canisters go offline during this process; therefore, the system triggers these alerts.

While trying to upgrade an IBM FlashSystem, you might also receive a message, such as an error in verifying the signature of the update package.

This message does not mean that an issue exists in your system. Sometimes, this issue occurs because not enough space is available on the system to copy the file, or the package is incomplete or contains errors. In this case, open a Salesforce case with IBM Support and follow their instructions.

11.3.8 Managing physical capacity of over provisioned storage controllers

Drives and back-end controllers exist that include built-in hardware compression and other data reduction technologies that allow capacity to be provisioned over and above the available real physical capacity. Different data sets lead to different capacity savings and some data, such as encrypted data or compressed data, do not even compress. When the physical capacity savings do not match the expected or provisioned capacity, the storage can run out of physical space, which leads to a write-protected drive or array.

To avoid running out of space on the system, the usable capacity must be monitored carefully by using the GUI of the IBM FlashSystem. The IBM FlashSystem GUI is the only capacity dashboard that shows the physical capacity.

Monitoring is especially important when migrating substantial amounts of data onto the IBM FlashSystem, which typically occurs during the first part of the workload life cycle as data is on-boarded, or initially populated into the storage system.

IBM strongly encourages users to configure Call Home on the IBM FlashSystem. Call Home monitors the physical free space on the system and automatically opens a service call for systems that reach 99% of their usable capacity.

IBM Storage Insights also can monitor and report on any potential out-of-space conditions and the new Advisor function warns when the IBM FlashSystem almost at full capacity. For more information, see 11.6.5, “IBM Storage Insights Advisor” on page 564.

When IBM FlashSystem reaches an out-of-space condition, the device drops into a read-only state. An assessment of the data compression ratio and the re-planned capacity estimation should be done to determine how much outstanding storage demand might exist. This extra capacity must be prepared and presented to the host so that recovery can begin.

The approaches that can be taken to reclaim space on the IBM FlashSystem in this scenario vary by the capabilities of the system, optional external back-end controllers, the system configuration, and planned capacity overhead needs.

In general, the following options are available:

- Add capacity to the IBM FlashSystem. Customers should have a plan that allows them to add capacity to the system when needed.
- Reserve a set of space in the IBM FlashSystem that makes it “seem” fuller than it really is, and that you can free up in an emergency situation. IBM FlashSystem can create a volume that is not compressed, de-duped, or thin provisioned (a fully allocated volume).

Create some of these volumes to reserve an amount of physical space. You can name (for example, “emergency buffer space”). If you are reaching the limits for physical capacity, you can delete one or more of these volumes to give yourself a temporary reprieve.

Important: Running out of space can be a serious situation. Recovery can be time-consuming. For this reason, it is imperative that suitable planning and monitoring be done to avoid reaching this condition.

Next, we describe the process for recovering from an out-of-space condition.

Reclaiming and unlocking

After you assessed and accounted for storage capacity, the first step is to contact IBM Support who can help in unlocking the read-only mode and restoring operations. The reclamation task can take a long time to run, and larger flash arrays take longer to recover than smaller ones.

Freeing up space

You can reduce the amount of used space after the IBM FlashSystem is unlocked by IBM support by using several methods.

To recover from Out of Space conditions on Standard Pools, complete the following steps:

1. Add storage to the system, if possible.
2. Migrate extents from the write-protected array to other nonwrite protected MDisks with enough extents, such as an external back-end storage array.
3. Migrate volumes with extents on the write-protected array to another pool. If possible, moving volumes from the IBM FlashSystem pool to another external pool can free up space in the IBM FlashSystem pool to allow for space reclamation.

   As this volume moves into the new pool, its previously occupied flash extends are freed up (by using SCSI **`unmap`**), which then provides more free space to the IBM FlashSystem enclosure to be configured to a proper provisioning to support the compression ratio.
4. Delete dispensable volumes to free-up space. If possible, within the pool (managed disk group) on the IBM FlashSystem, delete unnecessary volumes. The IBM FlashSystem supports SCSI **`unmap`** so deleting volumes results in space reclamation benefits by using this method.
5. Bring the volumes in the pool back online by using a Directed Maintenance Procedure.

For more information about the types of recovery, see this IBM Support web page.

11.3.9 Replacing a failed flash drive

When IBM FlashSystem detects a failed NVMe drive or optional external attached flash drive, it automatically generates an error in the Events window. To replace the failed drive, always run Fix Procedure for this event in the **Monitoring** → **Events** window.

The Fix Procedure helps you to identify the enclosure and slot where the bad drive is located, and guides you to the correct steps to follow to replace it. When a flash drive fails, it is removed from the array. If a suitable spare drive is available, it is taken into the array and the rebuild process starts on this drive.

After the failed flash drive is replaced and the system detects the replacement, it reconfigures the new drive as spare. Therefore, the failed flash drive is removed from the configuration, and the new drive is used to fulfill the array membership goals of the system.

11.3.10 Recovering from common events

You can recover from several of the more common events that you might encounter. In all cases, you must read and understand the current product limitations to verify the configuration and to determine whether you must upgrade any components or install the latest fixes or patches.

To obtain support for any IBM product, see the IBM Support Homepage.

If the problem is caused by IBM Spectrum Virtualize and you cannot fix it by using the Recommended Action feature or by examining the event log, collect the IBM Spectrum Virtualize support package as described in 11.2.1, “IBM FlashSystem data collection” on page 526.

11.4 Remote Support Assistance

Remote Support Assistance (RSA) enables IBM support to access the IBM FlashSystem device to perform troubleshooting and maintenance tasks. Support assistance can be configured to support personnel work onsite only, or to access the system both onsite and remotely. Both methods use secure connections to protect data in the communication between support center and system. Also, you can audit all actions that support personnel conduct on the system.

To set up the remote support options in the GUI, select **Settings** → **Support** → **Support Assistance** → **Reconfigure Settings**, as shown in Figure 11-7.

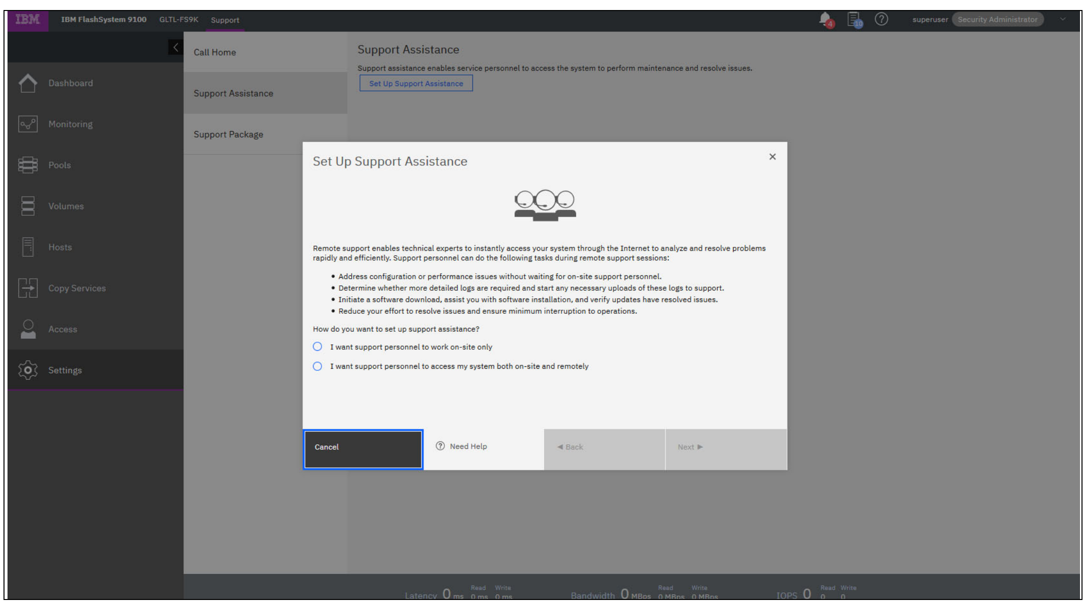


Figure 11-7 Remote Support options

You can use local support assistance if you have security restrictions that do not allow support to connect remotely to your systems. With RSA, support personnel can work onsite and remotely by using a secure connection from the support center.

They can perform troubleshooting, upload support packages, and download software to the system with your permission. When you configure remote support assistance in the GUI, local support assistance also is enabled.

The following access types are in the remote support assistance method:

- At any time

  Support center can start remote support sessions at any time.

- By permission only

  Support center can start a remote support session only if permitted by an administrator. A time limit can be configured for the session.

> **Note:** Systems that are purchased with a three-year warranty include Enterprise Class Support (ECS) and are entitled to IBM Support by using RSA to quickly connect and diagnose problems. However, IBM Support might choose to use this feature on non-ECS systems at their discretion; therefore, we recommend configuring and testing the connection on all systems.

To configure remote support assistance, the following prerequisites must be met:

- Call Home is configured with a valid email server.
- A valid service IP address is configured on each node on the system.
- If your system is behind a firewall or if you want to route traffic from multiple storage systems to the same place, a Remote Support Proxy server must be configured. Before you configure remote support assistance, the proxy server must be installed and configured separately. The IP address and the port number for the proxy server must be set up on when enabling remote support centers.

  For more information about setting up the Remote Proxy Server, see this IBM Documentation web page.

- If you do not have firewall restrictions and the storage nodes are directly connected to the Internet, request your network administrator to allow connections to `129.33.206.139` and `204.146.30.139` on port 22.
- Uploading support packages and downloading software require direct connections to the Internet. A DNS server must be defined on your system for both of these functions to work. The Remote Proxy Server cannot be used to download files.
- To ensure that support packages are uploaded correctly, configure the firewall to allow connections to the following IP addresses on port 443: `129.42.56.189`, `129.42.54.189`, and `129.42.60.189`.
- To ensure that software is downloaded correctly, configure the firewall to allow connections to the following IP addresses on port 22:
  - `170.225.15.105`
  - `170.225.15.104`
  - `170.225.15.107`
  - `129.35.224.105`
  - `129.35.224.104`
  - `129.35.224.107`

RSA can be configured by using the GUI and CLI. For more information about configuring RSA, see *IBM FlashSystem 9100 Architecture, Performance, and Implementation*, SG24-8425.

11.5 Call Home Connect Cloud and Health Checker feature

Formerly known as *Call Home Web*, the new Call Home Connect Cloud is a cloud-based version with improved features to view Call Home information on the web.

Call Home is available in several IBM systems, including IBM FlashSystem, which allows them to automatically report problems and status to IBM.

Call Home Connect Cloud provides the following information about IBM systems:

- Automated tickets
- Combined Ticket View
- Warranty and contract status
- Health check alerts and recommendations
- System connectivity heartbeat
- Recommended software levels
- Inventory
- Security bulletins

For more information about Call Home Connect Cloud (Call Home Web), see the IBM Support website.

at the IBM Support website, Call Home Web is available at **My support** → **Call Home Web** as shown in Figure 11-8.

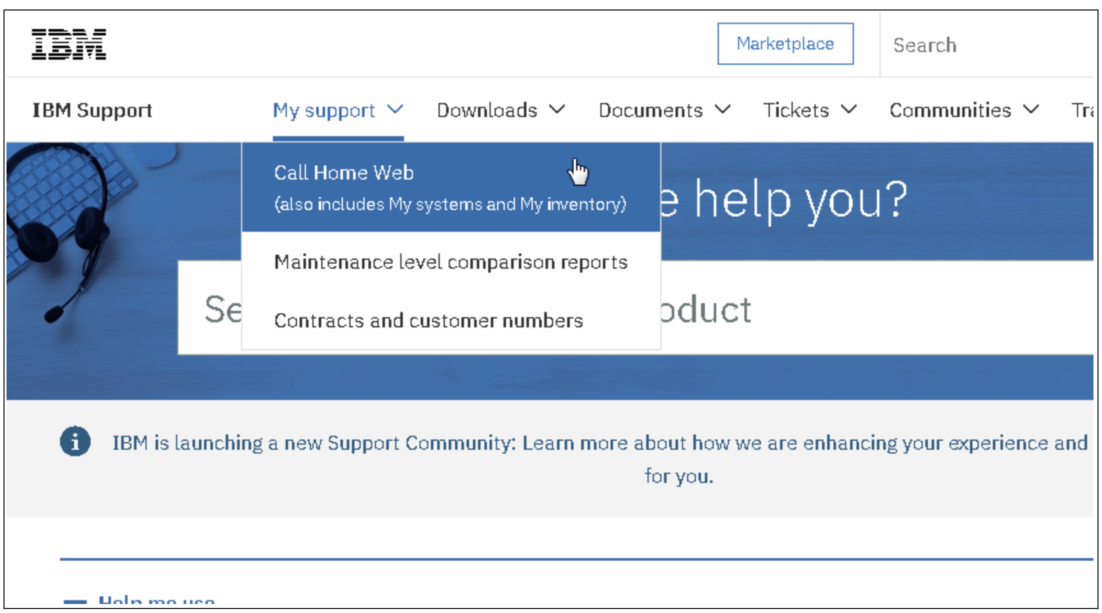


Figure 11-8 Call Home Web

Call Home Web was replaced by the new Call Home Connect Cloud web application. The new cloud-based application provides an enhanced, live view of your assets. This view includes the status of cases, warranties, maintenance contracts, service levels, end-of-service information, and other online tools.

To allow Call Home Connect Cloud analyze data of IBM FlashSystem systems and provide useful information about them, the devices must be added to the tool. The machine type, model, and serial number are required to register the product in Call Home Web. Also, it is required that IBM FlashSystem have Call Home and inventory notification enabled and operational.

Figure 11-9 shows the Call Home Connect Cloud details window of IBM FlashSystem.

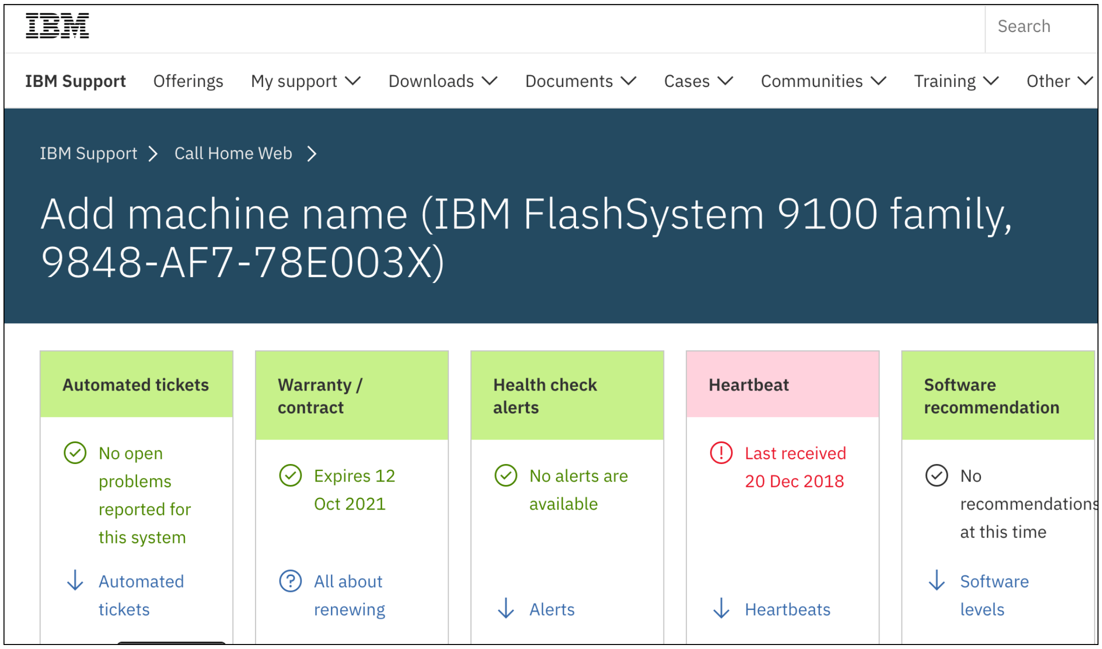


Figure 11-9 Call Home Web details window

For more information about how to set up and use Call Home Connect Cloud, see Introducing Call Home Connect Cloud.

11.5.1 Health Checker

A new feature of Call Home Connect Cloud is the Health Checker, which is a tool that runs in the IBM Cloud.

It analyzes Call Home and inventory data of systems that are registered in Call Home Connect Cloud and validates their configuration. Then, it displays alerts and provides recommendations in the Call Home Connect Cloud tool.

> **Note:** Use Call Home Connect Cloud because it provides useful information about your systems. The Health Checker feature helps you to monitor the system, and operatively provides alerts and creates recommendations that are related to them.

Some of the functions of the IBM Call Home Connect Cloud and Health Checker were ported to IBM Storage Insights, as described in 11.6, “IBM Storage Insights” on page 551.

11.6 IBM Storage Insights

IBM Storage Insights is an important part of the monitoring and ensuring continued availability of the IBM FlashSystem.

Available at no charge, cloud-based IBM Storage Insights provides a single dashboard that gives you a clear view of all of your IBM block storage. You can make better decisions by seeing trends in performance and capacity. Storage health information enables you to focus on areas that need attention.

In addition, when IBM Support is needed, IBM Storage Insights simplifies uploading logs, speeds resolution with online configuration data, and provides an overview of open tickets all in one place.

The following features are included:

- A unified view of IBM systems provides:
  - Single window to see all of your system's characteristics
  - List of all of your IBM storage inventory
  - Live event feed so that you know, up to the second, what is going on with your storage, which enables you to act fast.
- IBM Storage Insight collects telemetry data and Call Home data, and provides up-to-the-second system reporting of capacity and performance.
- Overall storage monitoring:
  - The overall health of the system
  - The configuration to see whether it meets the best practices
  - System resource management: determine whether the system is being overly taxed and provide proactive recommendations to fix it
- Storage Insights provides advanced customer service with an event filter that enables the following functions:
  - The ability for you and IBM Support to view, open, and close support tickets, and track trends.
  - Auto log collection capability to enable you to collect the logs and send them to IBM before support starts looking into the problem. This feature can save as much as 50% of the time to resolve the case.

In addition to the free Storage Insights, option of Storage Insights Pro is available, which is a subscription service that provides longer historical views of data, offers more reporting and optimization options, and supports IBM file and block storage with EMC VNX and VMAX.

Figure 11-10 shows the comparison of Storage Insights and Storage Insights Pro.

Product Comparison

	Capability	IBM Storage Insights (Free)	IBM Storage Insights Pro (Subscription)
Monitoring	Health, Performance and Capacity	✓	✓
	Filter events to quickly isolate trouble spots	✓	✓
	Drill down performance workflows to enable deep troubleshooting		✓
	Application / server storage performance troubleshooting		✓
	Customizable multi-conditional alerting		✓
Support Services	Simplified ticketing / log workflows and ticket history	✓	✓
	Proactive notification of risks (select systems)	✓	✓
Device Analytics	Part failure prediction	✓	✓
	Configuration best practice	✓	✓
	Customized upgrade recommendation	✓	✓
TCO Analytics	Capacity planning		✓
	Performance planning		✓
	Application / server storage consumption		✓
	Capacity optimization with reclamation planning		✓
	Data optimization with tier planning		✓

Figure 11-10 Storage Insights versus Storage Insights Pro comparison

Storage Insights provides a lightweight data collector that is deployed on a customer-supplied server. This server can be a Linux, Windows, or AIX server, or a guest in a virtual machine (for example, a VMware guest).

The data collector streams performance, capacity, asset, and configuration metadata to your IBM Cloud instance.

The metadata flows in one direction: from your data center to IBM Cloud over HTTPS. In the IBM Cloud, your metadata is protected by physical, organizational, access, and security controls. IBM Storage Insights is ISO/IEC 27001 Information Security Management certified.

Collected metadata

The following metadata about the configuration and operations of storage resources is collected:

- Name, model, firmware, and type of storage system.
- Inventory and configuration metadata for the storage system’s resources, such as volumes, pools, disks, and ports.
- Capacity values, such as capacity, unassigned space, used space, and the compression ratio.
- Performance metrics, such as read and write data rates, I/O rates, and response times.
- The application data that is stored on the storage systems cannot be accessed by the data collector

Accessing the metadata

Access to the metadata that is collected is restricted to the following users:

- The customer who owns the dashboard
- The administrators who are authorized to access the dashboard, such as the customer's operations team
- The IBM Cloud team that is responsible for the day-to-day operation and maintenance of IBM Cloud instances
- IBM Support for investigating and closing service tickets

11.6.1 Storage Insights Customer Dashboard

Figure 11-11 shows a view of the IBM Storage Insights main dashboard and the systems that it is monitoring.

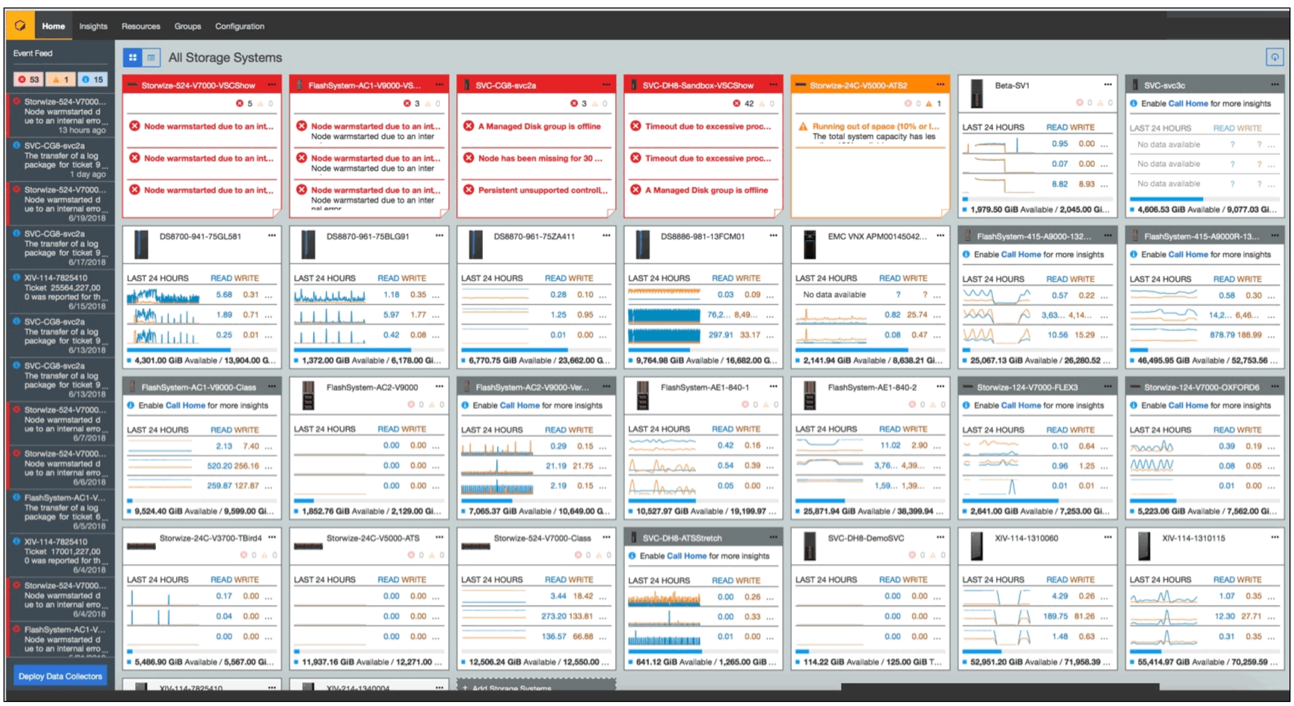


Figure 11-11 Storage Insights main dashboard

11.6.2 Customized dashboards to monitor your storage

With the latest release of IBM Storage Insights, you can customize the dashboard to show only a subset of the systems that are monitored. This feature is useful for customers that might be Cloud Service Providers (CSP) and want only a specific user to see those machines for which they are paying.

For more information about setting up the customized dashboard, see this IBM Documentation web page.

11.6.3 Creating a support ticket

The IBM Storage Insights Dashboard GUI can be used to create a support ticket for any of the systems that IBM Storage Insights reports about.

Complete the following steps:

1. In the IBM Storage Insights dashboard, choose the system for which you want to create the ticket. Then, select **Actions** → **Create/Update Ticket** (see Figure 11-12).

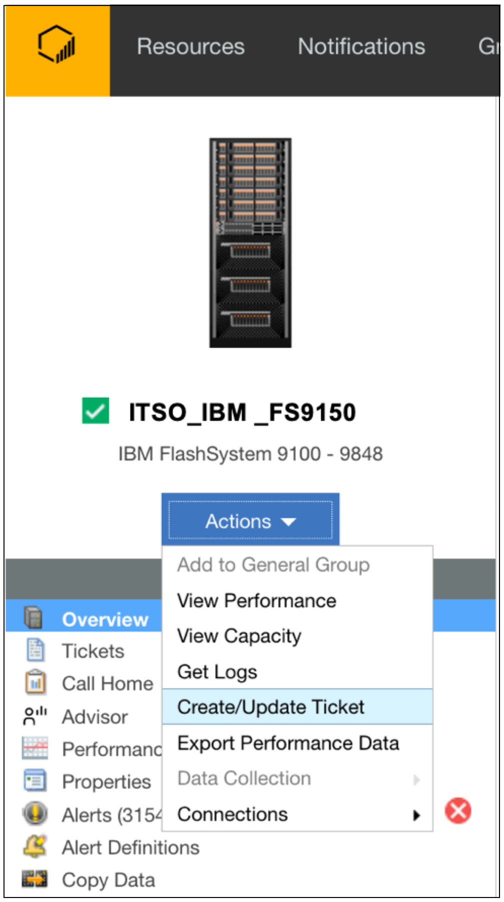


Figure 11-12 IBM Storage Insights Create/Update Ticket option

2. Select **Create Ticket** (see Figure 11-13). Several windows open in which you enter information about the machine, a problem description, and the option to upload logs.

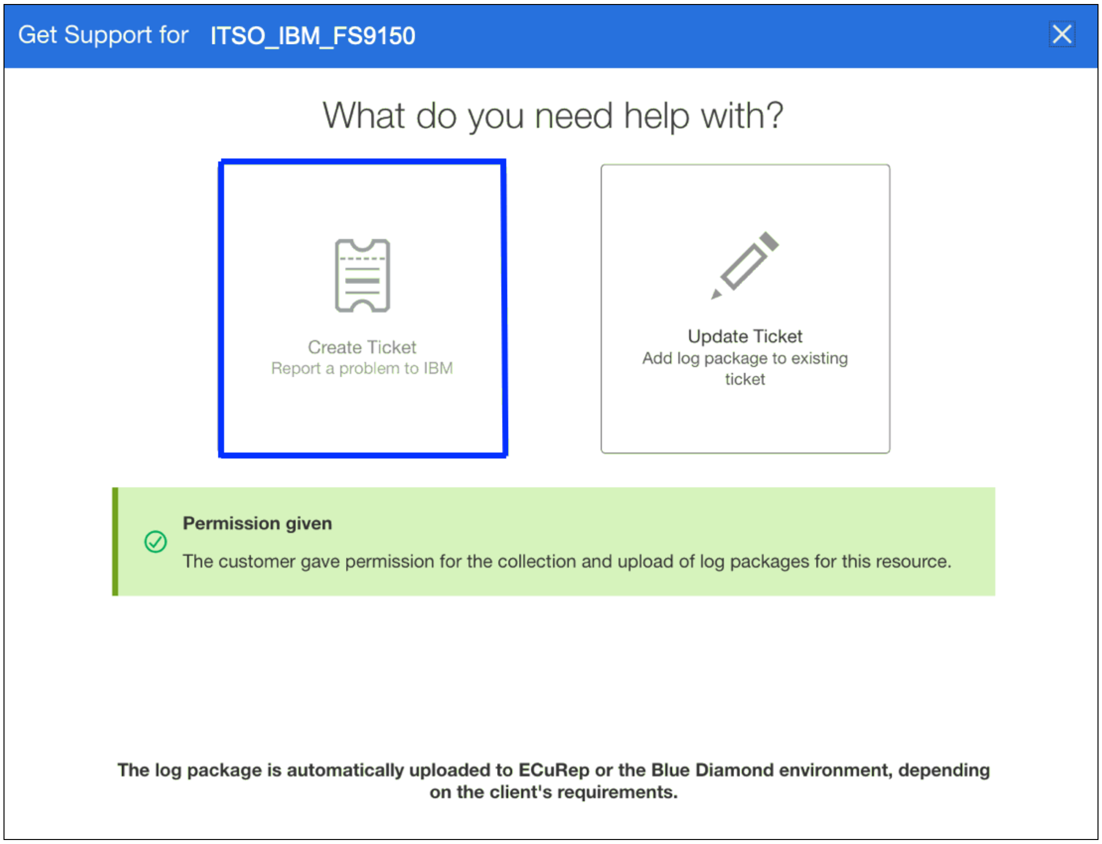


Figure 11-13 Create ticket option

> **Note:** The Permission given information box (see Figure 11-13) is an option that the customer must enable in the IBM FlashSystem GUI. For more information, see 11.4, “Remote Support Assistance” on page 547.

Figure 11-14 shows the ticket data collection that is done by the IBM Storage Insights application.

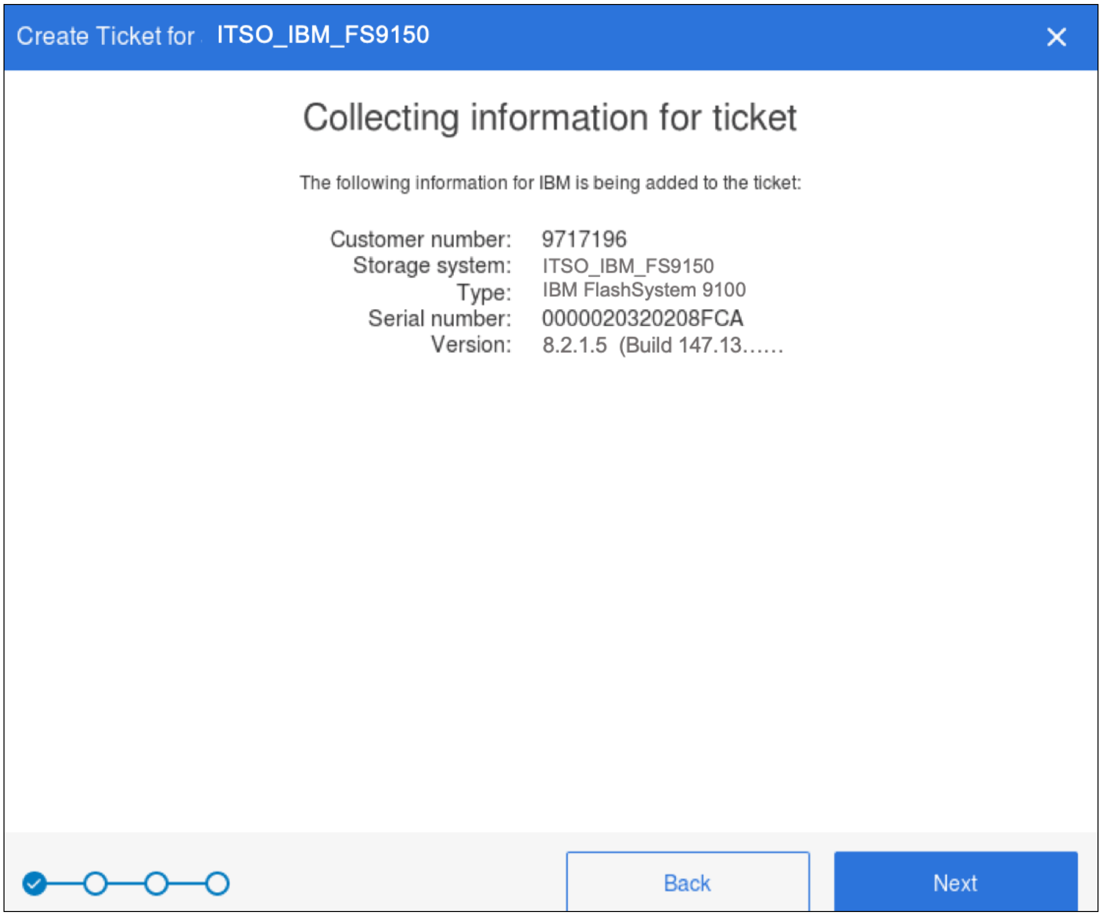


Figure 11-14 Collecting ticket information

3. As shown in Figure 11-15, you can add a problem description and attach other files to support the ticket, such as error logs or window captures of error messages.

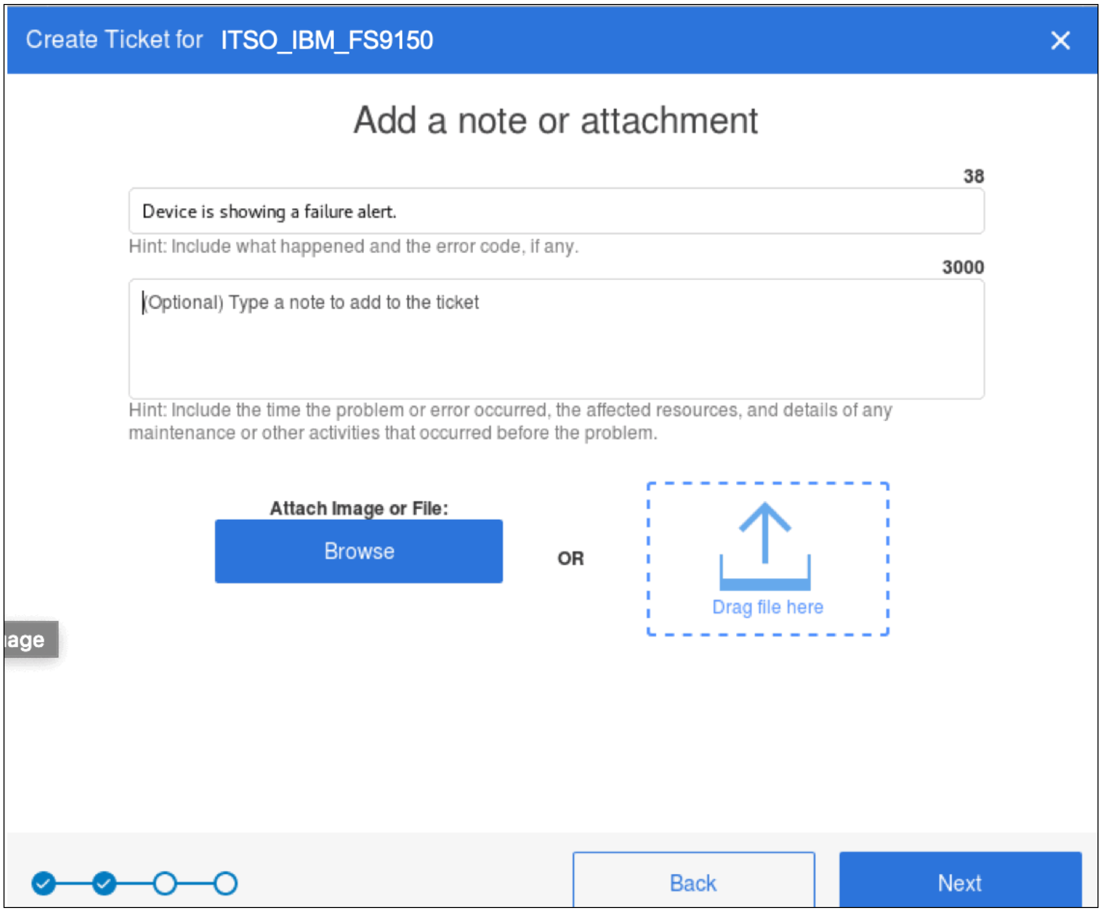


Figure 11-15 Adding problem description and any other information

4. You are then prompted to set a severity level for the ticket, as shown in Figure 11-16. Severity levels range from Severity 1(for a system that is down or extreme business impact) to Severity 4 (noncritical issue).

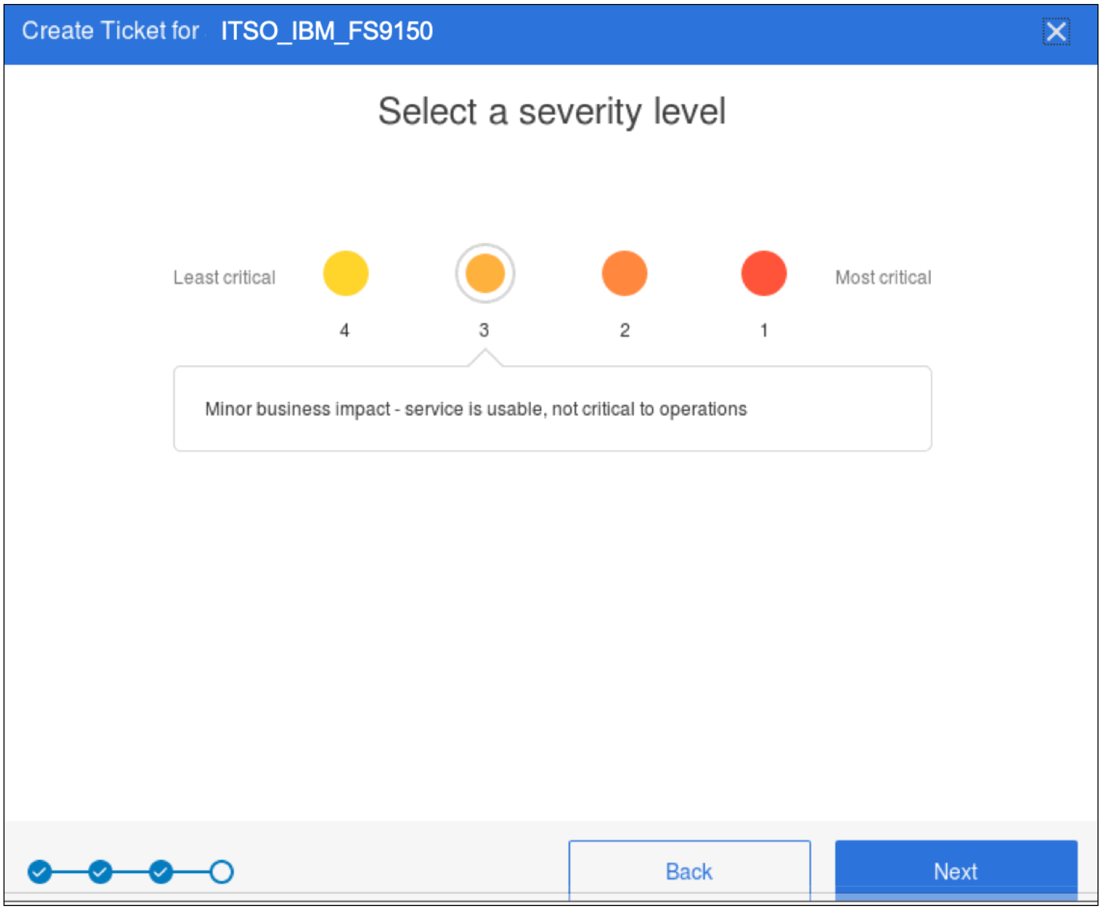


Figure 11-16 Set severity level

A summary of the data that is used to create the ticket is shown (see Figure 11-17).

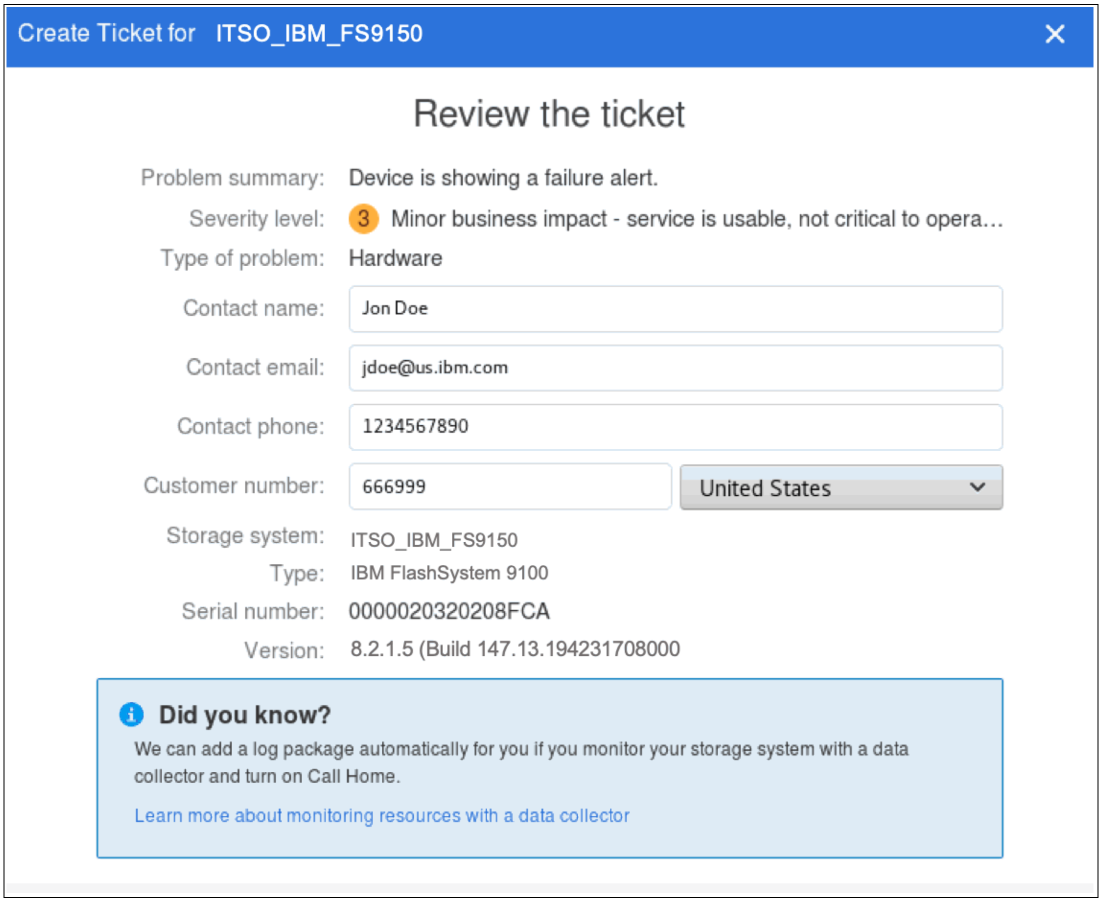


Figure 11-17 Review the ticket window

5. The final summary window (see Figure 11-18) includes the option to add logs to the ticket. When completed, click **Create Ticket** to create the support ticket and send it to IBM. The ticket number is created by the IBM Support system and returned to your IBM Storage Insights instance.

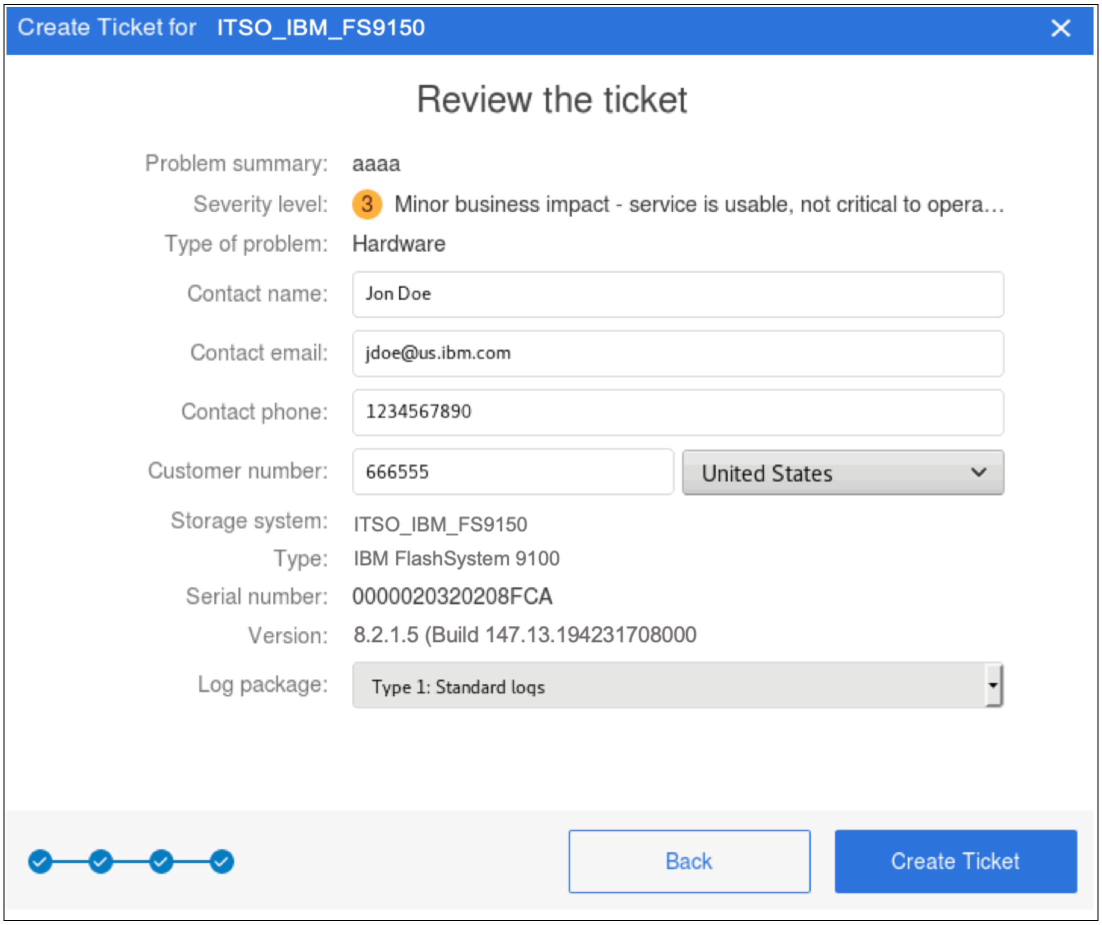


Figure 11-18 Final summary before ticket creation

6. Figure 11-19 shows how to view the summary of the open and closed ticket numbers for the system that is selected by using the **Action** menu option.

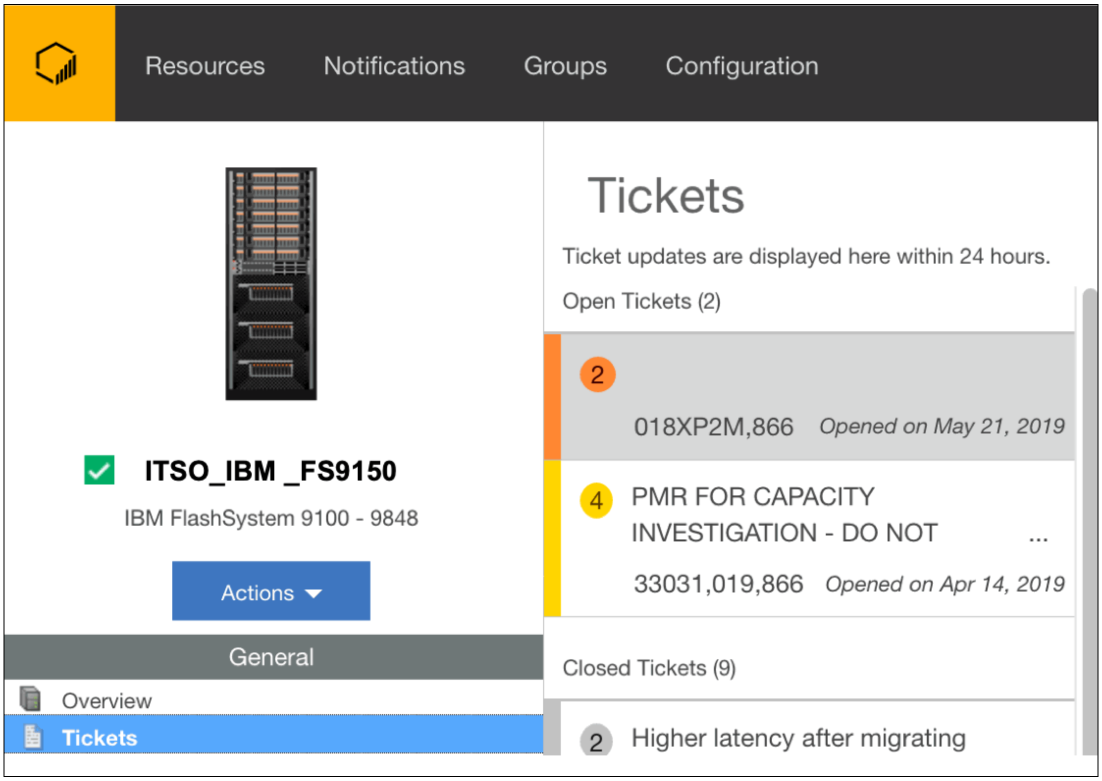


Figure 11-19 Ticket summary

11.6.4 Updating a support ticket

After a support ticket is created, the IBM Storage Insights Dashboard GUI can be used update tickets.

Complete the following steps:

1. In IBM Storage Insights dashboard, select **Update Ticket** (see Figure 11-20).

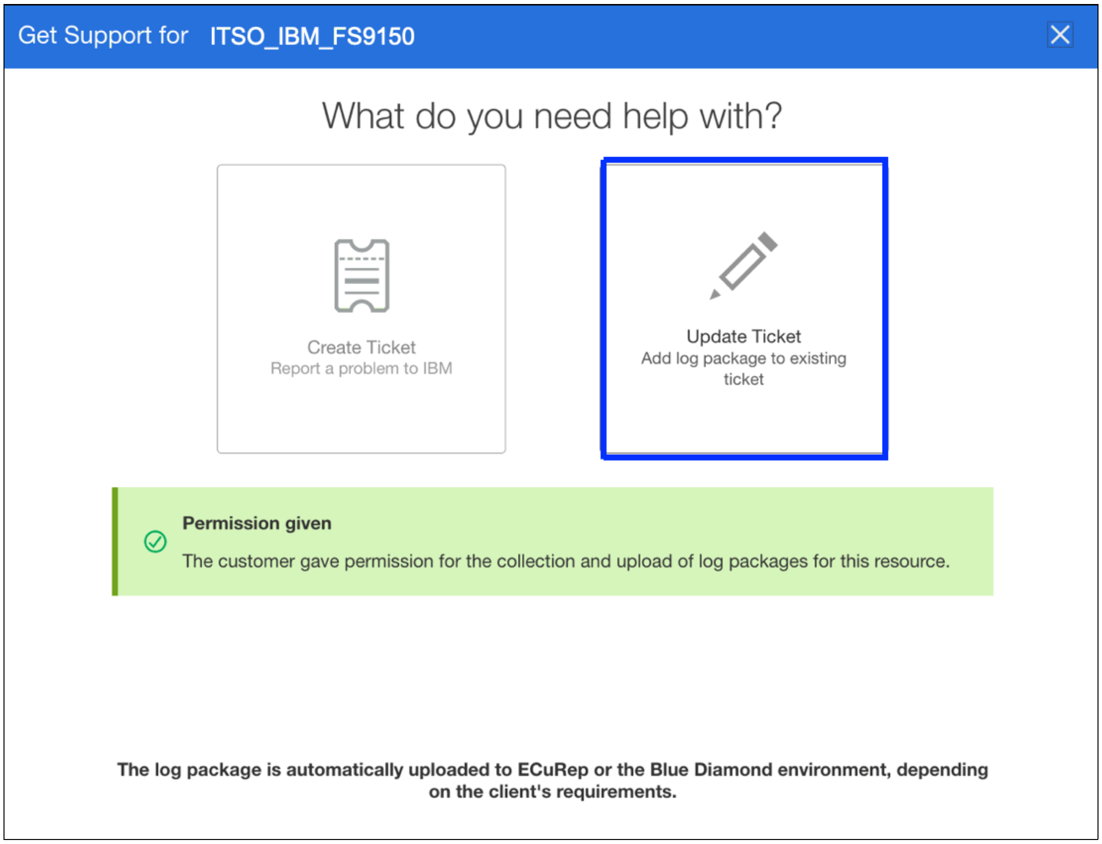


Figure 11-20 IBM Storage Insights Update Ticket

2. Enter the IBM Salesforce case number and then, click **Next** (see Figure 11-21). The IBM Salesforce case number uses the following format:

   `TS000XXXXX`

   These details are supplied when you created the ticket or by IBM Support if the PMR was created by a problem Call Home event (assuming that Call Home is enabled).

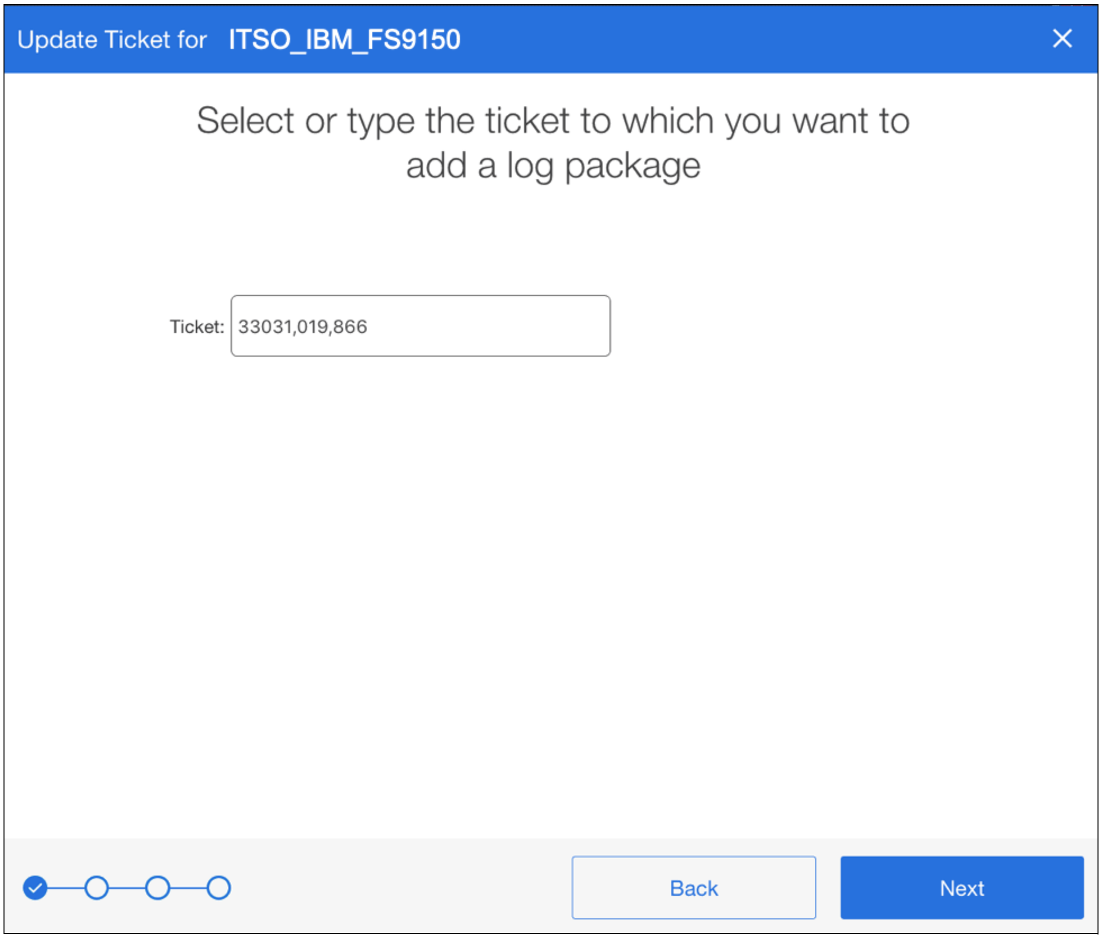


Figure 11-21 Entering the Salesforce/PMR case number

   A window opens in which you can choose the log type to upload. The window and the available options are shown in Figure 11-22 on page 564.

   The following options are available:

   – Type 1 - Standard logs

     For general problems, including simple hardware and simple performance problems

   – Type 2 - Standard logs and the most recent statesave log

   – Type 3 - Standard logs and the most recent statesave log from each node

     For 1195 and 1196 node errors and 2030 software restart errors

   – Type 4 - Standard logs and new statesave logs

     For complex performance problems, and problems with interoperability of hosts or storage systems, compressed volumes, and Remote Copy operations, including 1920 errors

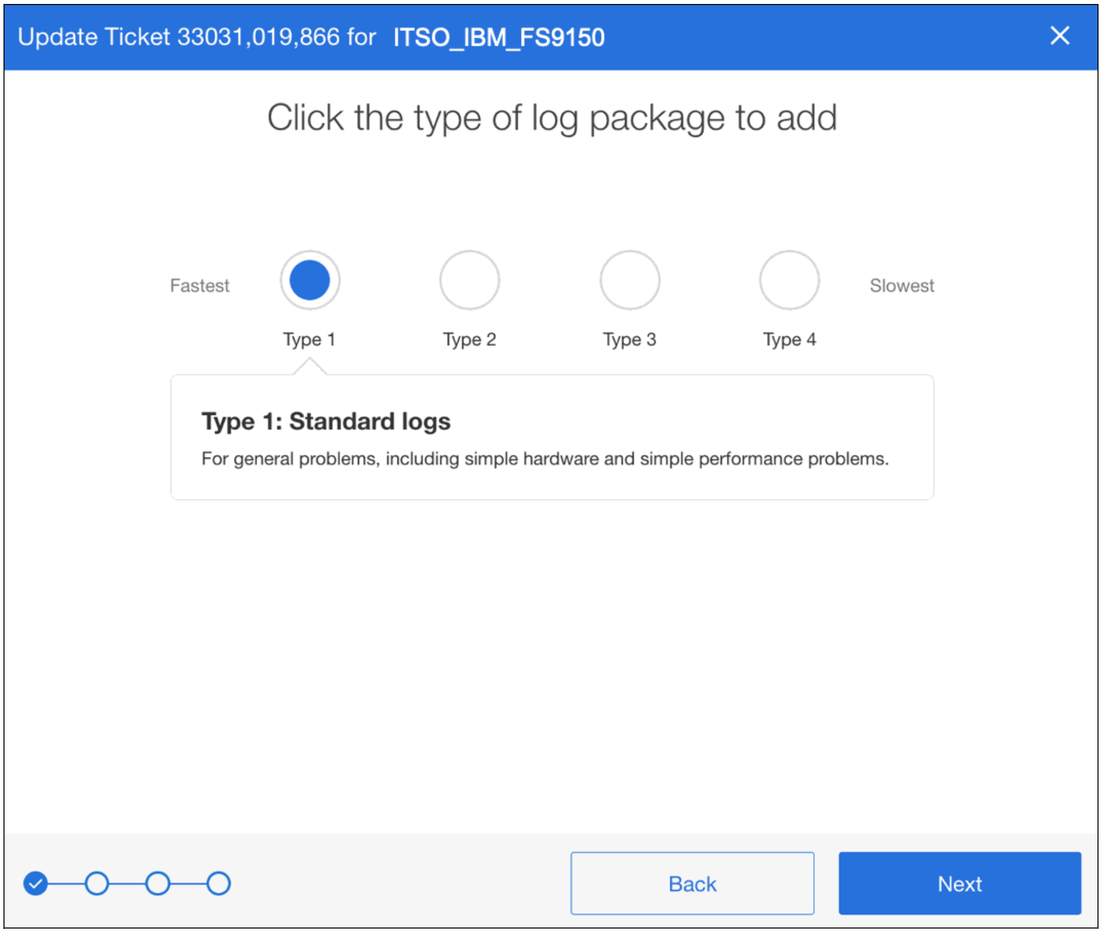


Figure 11-22 Log type selection

If you are unsure about which log type to upload, contact IBM Support for guidance. The most common type to use is type 1, which is the default type. The other types are more detailed logs and for issues in order of complexity.

3. After the type of log is selected, click **Next**. The log collection and upload process starts. When completed, the log completion window is displayed.

11.6.5 IBM Storage Insights Advisor

IBM Storage Insights continually evolves and the latest addition is a new option from the action menu called *Advisor*.

IBM Storage Insights analyzes your device data to identify violations of best practice guidelines and other risks, and to provide recommendations about how to address these potential problems.

Select the system from the dashboard and then, click the **Advisor** option to view these recommendations. To see more information about a recommendation or to acknowledge it, double-click the recommendation.

Figure 11-23 shows the initial IBM Storage Insights Advisor menu.

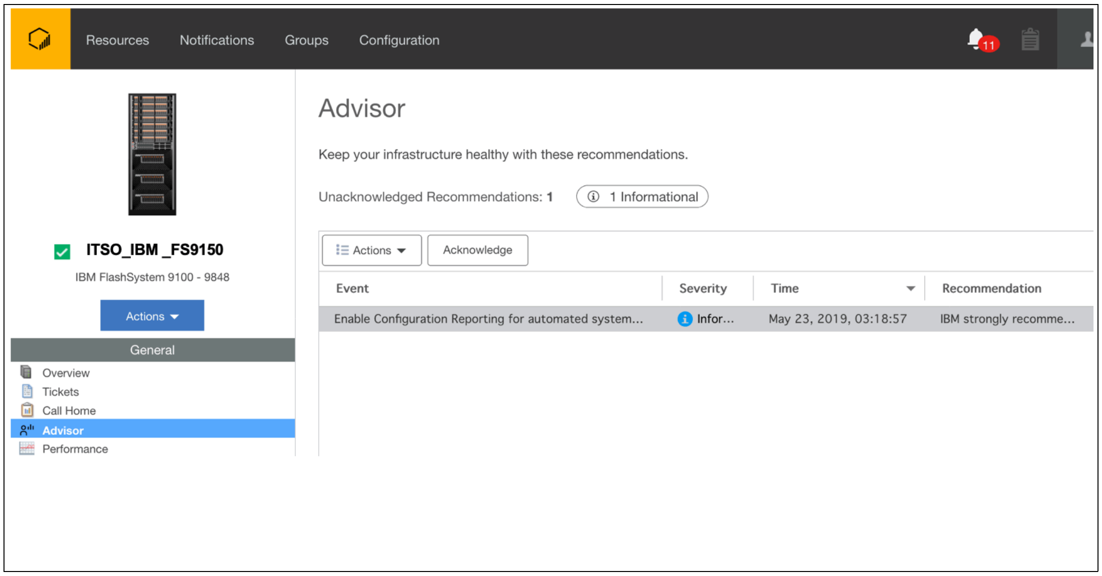


Figure 11-23 IBM Storage Insights Advisor menu

Figure 11-24 shows an example of the detailed IBM Storage Insights Advisor recommendations.

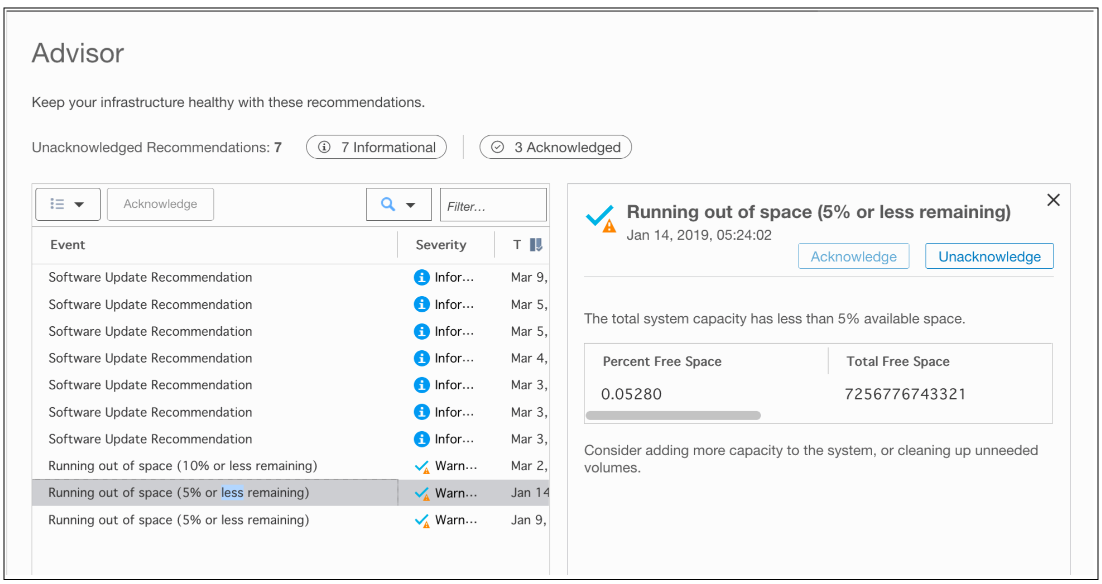


Figure 11-24 Advisor detailed summary of recommendations

As shown in Figure 11-24, the details of a `Running out of space` recommendation is shown the Advisor page. In this scenario, the user clicked the **Warning** tag to focus on only the recommendations that feature a severity of `Warning`.

For more information about setting and configuring the Advisor options, see this IBM Documentation web page.

12

Automation and scripting

This chapter provides information about scripting and automation tasks that can occur in an IBM Spectrum Virtualize environment that uses Ansible.

This chapter includes the following topics:

12.1 REST API on IBM Spectrum Virtualize

The IBM Spectrum Virtualize Representational State Transfer (REST) application programming interface (API) consists of command targets that are used to retrieve system information and to create, modify, and delete system resources. These command targets allow command parameters to pass through unedited to the IBM Spectrum Virtualize command-line interface (CLI), which handles parsing parameter specifications for validity and error reporting. Hypertext Transfer Protocol Secure (HTTPS) is used to communicate with the REST API server.

The easiest way to interact with the storage system by using the REST API is the `curl` utility (for more information, see this website) to make an HTTPS command request with a valid configuration node URL destination. Open TCP port 7443 and include the API version in combination with the keyword `rest` followed by the IBM Spectrum Virtualize target command that you want to run.

Each curl command uses the following format:

```
curl –k –X POST –H <header_1> –H <header_2> ... -d <JSON input>
https://SVC_ip_address:7443/rest/<api_version>/target
```

Where the following definitions apply:

- POST is the only HTTPS method that the Spectrum Virtualize REST API supports.
- Headers `<header_1>` and `<header_2>` are individually specified HTTP headers (for example, Content-Type and X-Auth-Username).
- The use of the `-d` parameter followed by the input in Java script Object Notation (JSON); for example, '{"raid_level": "raid5"}' to provide required more configuration information.
- `<SVC_ip_address>` is the IP address of the IBM SAN Volume Controller to which you are sending requests.
- `<target>` is the target object of commands, which includes any object IDs, names, and parameters.
- `<api_version>` specifies which version of the API should get used. The latest API version is v1.

> **Note:** Compatibility with an earlier version is implemented by auto redirection of nonversioned requests to v0. It is recommended to use versioned endpoints for guaranteed behavior.

Consider the following examples:

- `https://SVC_ip_address:7443/rest/target` uses API v0.
- `https://SVC_ip_address:7443/v0/rest/target` uses API v0.
- `https://SVC_ip_address:7443/v1/rest/target` uses API v1.

Authentication

Aside from data encryption, the HTTPS server requires authentication of a valid username and password for each API session. Two authentication header fields must be specified for your credentials: `X-Auth-Username` and `X-Auth-Password`.

Initial authentication requires that you POST the authentication target (**`/auth`**) with the username and password. The REST API server returns a hexadecimal token. A single session lasts a maximum of two active hours or 30 inactive minutes, whichever occurs first.

Note: The **chsecurity** command configures the amount of time (in minutes) before a token expires. The allowed range is 10 - 120 minutes. The default value is 60 minutes.

When your session ends because of inactivity (or if you reach the maximum time that is allotted), error code 403 indicates the loss of authorization. Use the **/auth** command target to re-authenticate by using the username and password.

The following example shows the correct procedure for authenticating. You authenticate by first producing an authentication token and then, use that token in all future commands until the session ends.

For example, the following command passes the authentication command to IBM SAN Volume Controller node IP address 192.168.10.20 at port 7443 by using API version v1:

```
curl –k –X POST –H ‘Content-Type: application/json’ –H ‘X-Auth-Username: myuser’ –
H ‘X-Auth-Password: mypassw0rd’ https://192.168.10.20:7443/rest/v1/auth
```

Note: Make sure that you format the request correctly by using spaces after each colon in each header; otherwise, the command fails.

This request yields an authentication token, which can be used for all subsequent commands, as shown in the following example:

```
{"token": "38823f60c758dca26f3eaac0ffee42aadc4664964905a6f058ae2ec92e0f0b63"}
```

The X-Auth-Token header in combination with the authentication token replaces the username and password for all further actions. The token is good for one session, but the session times out after two hours of activity or 30 minutes of inactivity. Repeat the authentication for creating another token.

Example commands

In this section, we discuss some of example commands.

Using the authentication token

Most actions can be taken only after authentication. The following example of displaying the system information demonstrates the use of the previously generated token in place of the authentication headers that are used in the authentication process:

```
curl –k –X POST –H ‘Content-Type: application/json’ –H ‘X-Auth-Token:
38823f60c758dca26f3eaac0ffee42aadc4664964905a6f058ae2ec92e0f0b63’
https://192.168.10.20:7443/rest/v1/lssystem
```

Note: If you use curl, you do not receive the HTTPS error code that is displayed if you do not specify the **-f** option.

Specifying more parameters

Although querying any information does not require that any other parameters are required within the REST call, this mandatory requirement exists for any action that intends to modify an object or create an object; for example, host, hostcluster, array, or VDisk.

The following example demonstrates the use of the **-d** parameter to specify the parameters and associated values that are required for creating a mirrored volume:

```
curl –k –X POST –H ‘Content-Type: application/json’ –H ‘X-Auth-Token: 38823f60c758dca26f3eaac0ffee42aadc4664964905a6f058ae2ec92e0f0b63’
–d ’{"name":"myVDisk1", "copies":"2", "mdiskgrp":"mdiskgrp0:mdiskgrp1",
"size":"200", "vtype":"striped", "unit":"gb", "rsize":"5%" }’
https://192.168.10.20:7443/rest/v1/mkvdisk
```

This REST call is equivalent to running the following **mkvdisk** command (it produces the same output):

```
mkvdisk -name myVDisk1 -copies 2 -mdiskgrp mdiskgrp0:mdiskgrp1 -size 200 -vtype striped -unit gb -rsize 5%
```

The parameters and values that can or must be specified when a REST target is used, such as **mkvdisk** or **mkhost**, are identical to those within the CLI.

Note: Parameters (keys) and values must be specified in JSON notation.

JSON is a lightweight data-interchange text format that is language-independent, but uses conventions that are familiar to programmers of the C-family of languages, including C, C++, C#, Java, JavaScript, Perl, and Python.

JSON data is written as key/value pairs. A key/value pair consists of a field key, followed by a colon, followed by a value, whereby key and value must be placed in double quotes.

One or more key/value pairs build an object, which begins with a left brace ({) and ends with a right brace (}), as shown in Example 12-1).

Example 12-1 JSON notation for creating a thin provisioned mirrored VDisk

```
{
  "name": "myVDisk1",
  "copies": "2",
  "mdiskgrp": "mdiskgrp0:mdiskgrp1",
  "size": "200",
  "vtype": "striped",
  "unit": "gb",
  "rsize": "5%"
}
```

For more information about JSON, see this website.

Rate limiting

Rate limiting (see Example 12-1) helps with security and the prevention of an attack, such as a denial of service in which unlimited work is sent to the system. The rate limiting is implemented at millisecond granularity and creates a return code (`429 - too many requests`) when a violation occurs.

Table 12-1 REST API rate limits

Limit	Type	Value
Maximum active connections per cluster	REST API	4
Maximum requests per second to the `/auth` endpoint	REST API	3 per second
Maximum requests per second to the `/non-auth` endpoint	REST API	10 per second
Number of simultaneous CLIs in progress	System	1

REST API Explorer

REST API documentation is available at this IBM Documentation web page. Support also can be found directly on the system within the REST API Explorer.

The REST API Explorer is based on the Swagger UI and runs within a browser. It offers an easy way to familiarize yourself with the API and to test the commands that it contains.

To access the REST API Explorer, enter the following URL in a browser:

```
https://<SVC_ip_address | FQDN>:7443/rest/explorer
```

Figure 12-1 shows the grouping of available actions within the RET API Explorer.

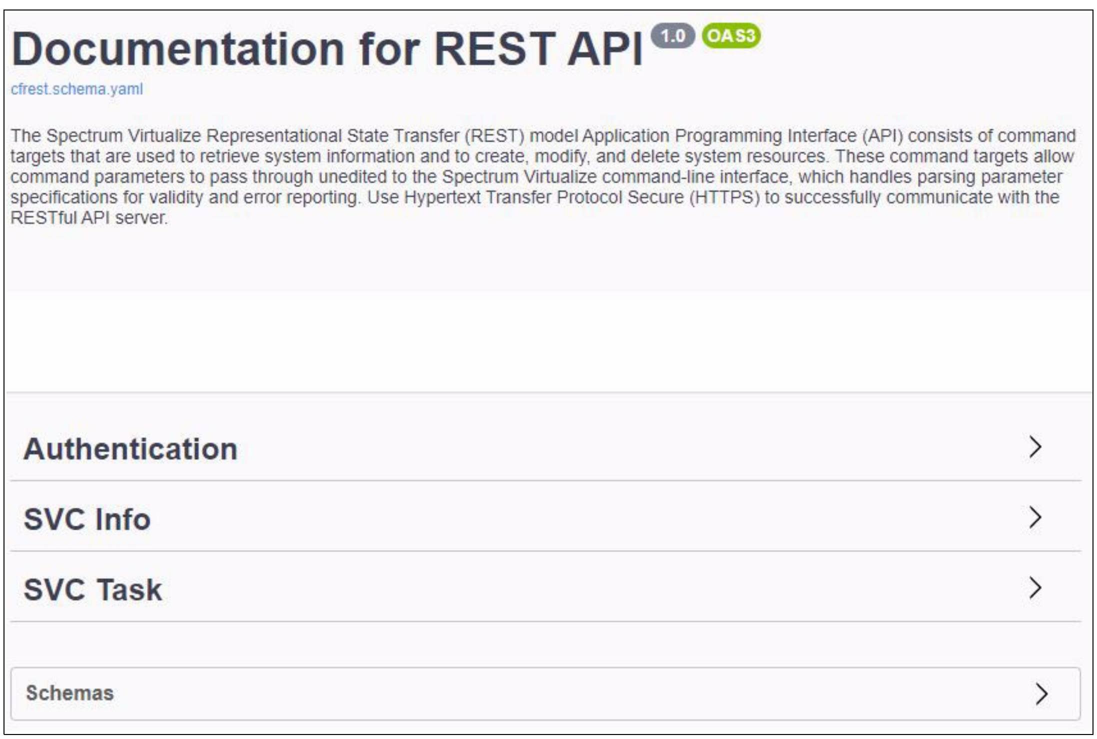


Figure 12-1 REST API Explorer actions

The use of the REST API Explorer also requires the generation of a token, which can be used for all further actions. Figure 12-2 shows how to create an authentication token for within the IBM SAN Volume Controller Info and IBM SAN Volume Controller Task actions.

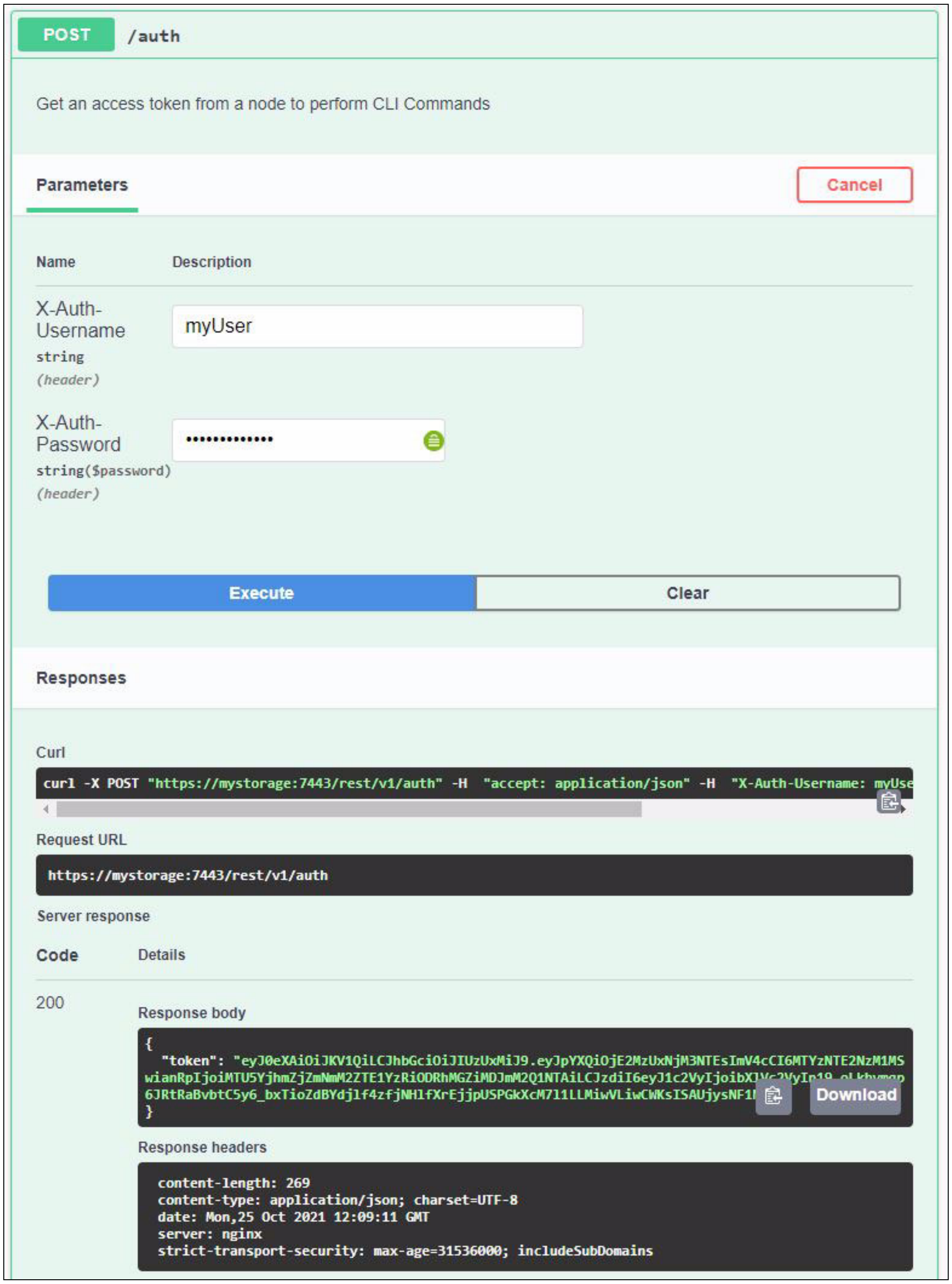


Figure 12-2 REST API Explorer authentication

Figure 12-2 also shows the following outputs after successful authentication:

- The **`curl`** command that is required to carry out the action
- Request URL, which was addressed during the execution of the action
- Server response in from of the Response body and the response header

The token is displayed in JSON notation in the response body.

By using the generated authentication token, more actions can be completed in the REST API Explorer.

Figure 12-3 shows the **mkvdisk** task in the REST API Explorer. All accepted parameters are listed in the request body. These parameters can then be adapted or deleted according to the requirements for creating the VDisk.

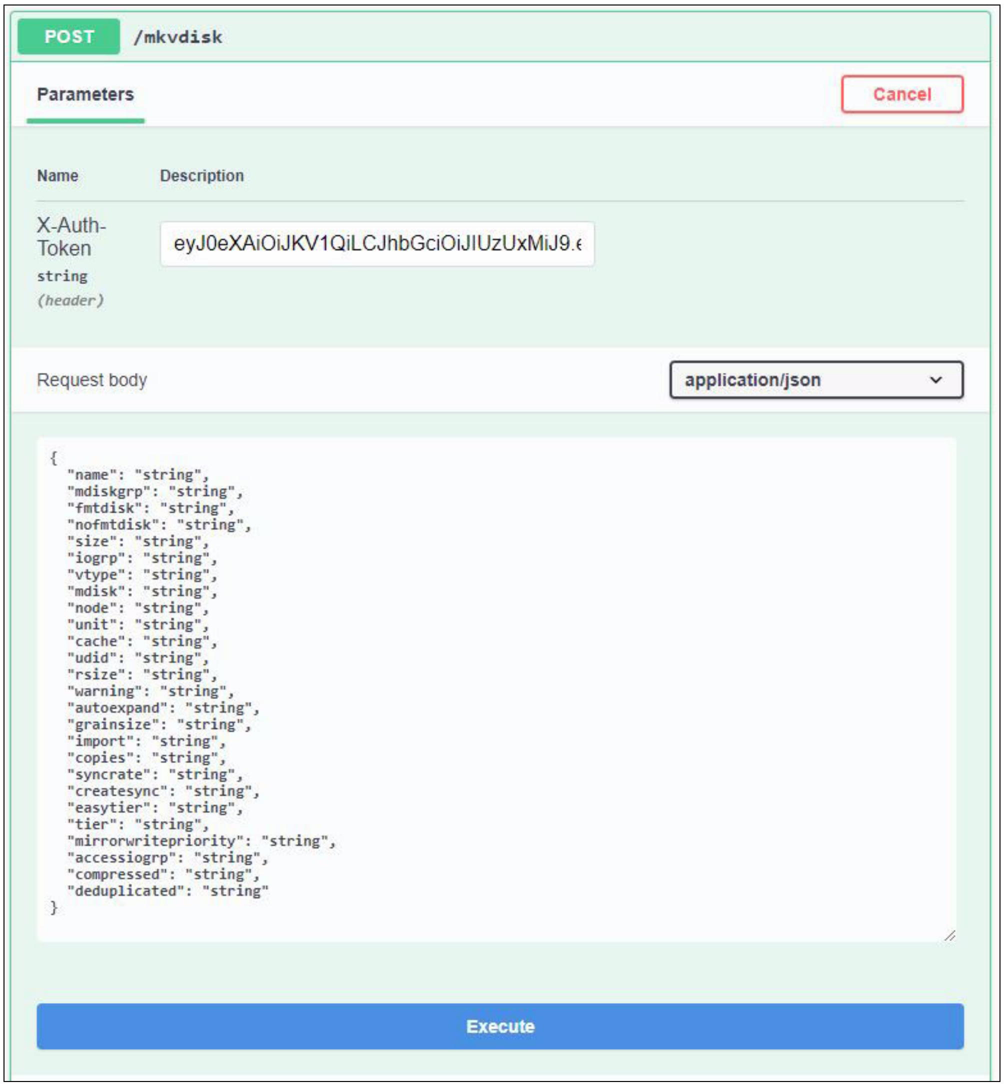


Figure 12-3 REST API Explorer /mkvdisk

Figure 12-4 shows an example for the Request body to create a mirrored VDisk and the output within the Response body.

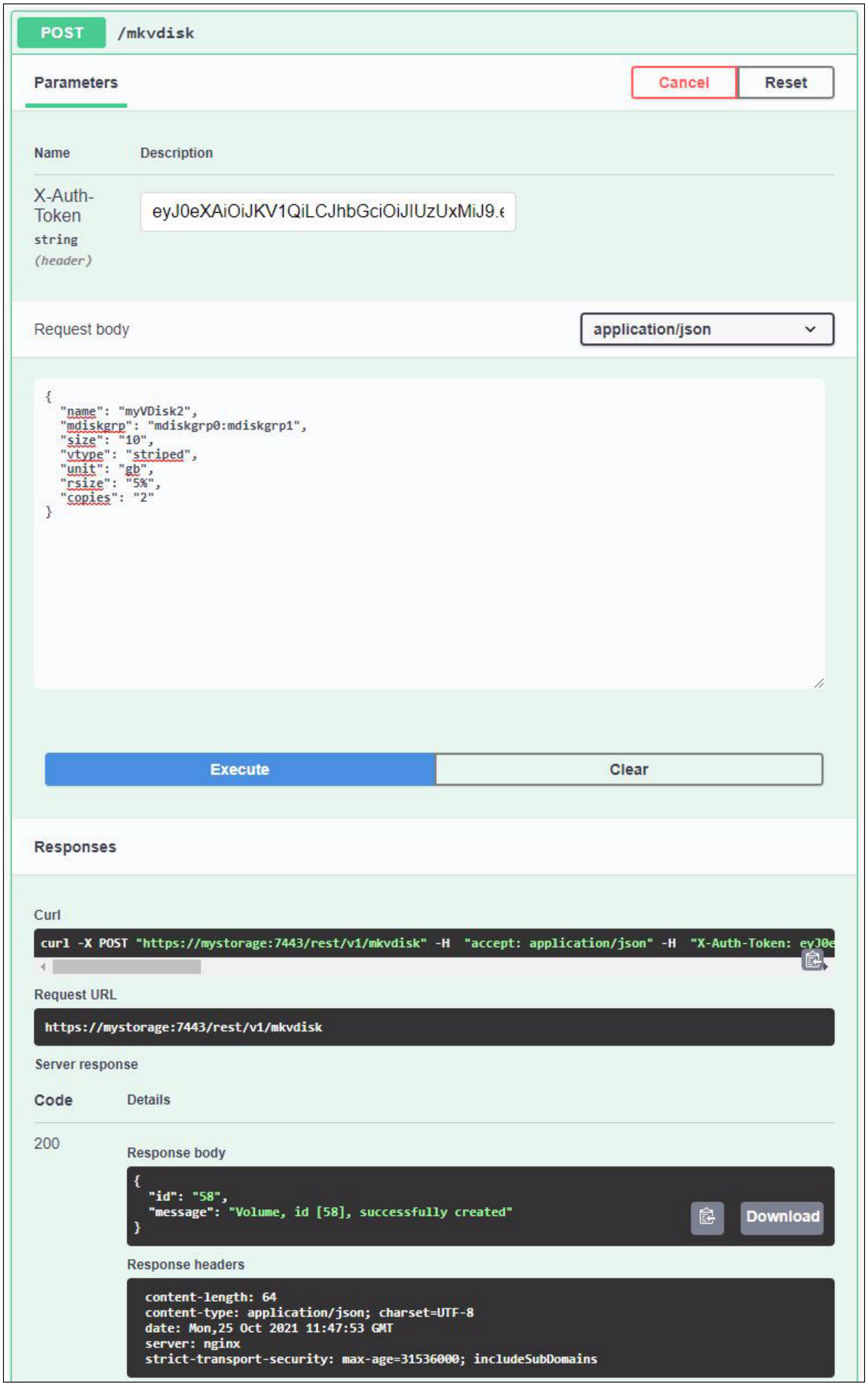


Figure 12-4 REST API Explorer /mkvdisk output

REST API HTTP messages

When an issue exists with the information that you provided, an error message appears.

Different types of error messages can appear, depending on the issue. The only error messages that are described in this document are HTTP errors. Other error messages are explained in other product documentation, which is available at this IBM Documentation web page.

The following HTTP error codes are returned to the user by the REST API in response to a problem with the request:

- `400: bad request`

  The command did not specify a required parameter or gave a parameter value that did not pass the REST API checks.

- `401: unauthorized`

  The command requires a successful authentication.

- `403: forbidden`

  The user did not send a valid authentication token to interact with the specified URL.

- `404: not found`

  The command attempted to issue a request to a URL that does not exist.

- `405: method not allowed`

  The command attempted to use an HTTP method that is invalid for the specified URL.

- `409: conflict`

  The sent request conflicts with the current state of the system.

- `429: too many requests`

  Too many requests violate the rate limiting.

- `500: something went wrong on the server`

  A Spectrum Virtualize command error was forwarded from the REST API.

- `502: bad gateway`

  The API received an invalid response from the upstream system.

For more information about the use of the REST API, see this IBM Documentation web page.

For more information about other examples, see the following web pages:

- IBM Spectrum Virtualize Interfacing Using the RESTful API
- Tips and tricks using the Spectrum Virtualize REST API

Audit logging

Commands that are started by the REST API are auditable, such as actions that are started by the CLI or GUI. The Origin field within the output of the **`catauditlog`** CLI command shows the source interface of the executed command.

12.2 Scripting

This section describes some methods that can be used to access the IBM Spectrum Virtualize Controller family by using scripts. These methods can be used for configuration, reporting, and administration tasks.

IBM Spectrum Virtualize Controller family supports the following methods or protocols for running configuration tasks and monitoring, in addition to the traditional web-based graphical user interface (GUI):

- Secure Shell (SSH)
- SMI-S
- HTTPS and REST API on IBM Spectrum Virtualize
- HTTPS and REST API on IBM Spectrum Control

12.2.1 Scripting principles

The following practices are recommended for scripting:

- Always use secure protocols, such as SSH and HTTPS.
- Use SSH-keys for authentication if possible and protect the SSH-keys.
- Use dedicated users for monitoring and configuring and administering purposes.
- Assign only the required permissions according to the purpose of the configured user.
- Implement error handling in the scripts.

12.2.2 Secure Shell

SSH is a network protocol that enables secure communication and operation over an insecure network.

All members of the IBM Spectrum Virtualize storage products feature a CLI, which is accessible by using the SSH protocol.

> **Note:** The SSH protocol authenticates users by using passwords or SSH keys (that i, asymmetric cryptography methods). For security reasons, it is recommended to use and protect the configured SSH keys.

The system supports up to 32 interactive SSH sessions on the management IP address simultaneously.

> **Note:** After an SSH interactive session times out, the session is automatically closed. The session timeout limit is set to 15 minutes by default. The limit value can be changed by using the **chsecurity** command. For more information, see this IBM Documentation web page.

To connect to the system, the SSH client requires a user login name and an SSH password (or if you require command-line access without entering a password, the key pair). Authenticate to the system by using a management username and password.

When you use an SSH client to access a system, you must use your username and password. The system uses the password (and if not a password, the SSH key pair) to authorize the user who is accessing the system.

General tips

Consider the following general tips when SSH is used:

- Use of **`svcinfo`** and **`svctask`**

  Some small differences exist between the CLIs of the different products; for example, **`lsnodecanister`** is used on IBM FlashSystem Controllers, and **`lsnode`** is used on IBM Spectrum Virtualize controllers.

- Use of **`-delim`** **parameter** on **`ls-commands`**

  Parsing the output of a **`ls-command`** becomes much easier because it inserts a single, selectable character between each field instead of several spaces. The colon (:) is a good choice for a delimiter for all output that does not contain any IPv6 addresses.

- Use of **`-nohdr`** on **`ls-commands`**

  The use of the **`-nohdr`** parameter suppresses the output of the header so that the required code for skipping the first line of the output is bypassed.

Using SSH in bash/ksh

The use of an SSH client within a shell is a common way of running a specified command, but, not a login shell on a remote system. Instead of opening an interactive session, SSH runs a command on the remote system, forwards the output to the local computer, and then, exits the session.

Running commands remotely by way of SSH provides a way to write and use advanced scripts to collect data from an IBM Spectrum Virtualize system. It also continues processing that data on a local computer in combination with other available tools and utilities.

Running commands remotely by way of SSH allows SSH to be used in a shell and piping the output to any external program for parsing, filtering, and data processing.

Example 12-2 shows how to use SSH to run a command (**`svcinfo lssystem`**) on the IBM Spectrum Virtualize system (`mystorage`) with the privileges of `myuser` and piping the output to filter only for lines containing `unmap`.

Example 12-2 Using ssh and grabbing selected information

```
ssh myuser@mystorage ‘svcinfo lssystem | grep unmap’
```

Using SSH in Windows command line

For the use of the SSH in combination with the Windows operating system, the PuTTY Plink utility or the optional available OpenSSH client feature (which integrates into the standard command line) must be installed.

PuTTY Plink (see Example 12-3) enables authentication by using SSH keys by using or configuring the PuTTY authentication agent (Pageant).

Example 12-3 Using PuTTY PLink

```
cd C:\Program Files\PuTTY\
plink.exe myuser@mystorage ‘svcinfo lssystem | findstr unmap’
```

To enable the Windows OpenSSH client for authentication by using SSH keys, the keys must be placed within the following directory structure `C:\User\<username>\.ssh\`.

Using SSH with Python

Python requires the use of another external module to connect to IBM Spectrum Virtualize by using the SSH protocol.

The `paramiko` open source module is popular and can be installed by using `pip`.

Example 12-4 shows the simple use of `paramiko` within a Python script by connecting to `myStorage` by using the `myUser` user and running the commands to display the configured host and controller objects.

Example 12-4 Using paramiko within Python

```
#!/usr/bin/python

import paramiko
mystorage = 'myStorage'
myuser = 'myUser'

ssh = paramiko.SSHClient()
ssh.set_missing_host_key_policy(paramiko.AutoAddPolicy())
ssh.connect(hostname=mystorage, username=myuser)

command1 = 'lshost -delim :'
command2 = 'lscontroller -delim :'

stdin, stdout, stderr = ssh.exec_command(command1)
data = stdout.read()

errors = stderr.read()
if data:
        print(data)
if errors:
        print(errors)

print "-----------------------------------------------------------------------------\n";

stdin, stdout, stderr = ssh.exec_command(command2)
data = stdout.read()

errors = stderr.read()
if data:
        print(data)
if errors:
        print(errors)

ssh.close()
```

The `set_missing_host_key_policy(paramiko.AutoAddPolicy())` method defines how to proceed if the remote system SSH fingerprint is not known locally.

The `connect(hostname=target, username=user)` method connects with the storage system. If keys-based authentication is configured correctly, keys are checked automatically and a session with SSH server is established.

Several options are available with `client.connect()`. For example, certificates can be specified by using `pkey` or `key_filename` arguments, and set the user password with password argument if better authentication methods cannot be used.

The following example shows how to specify user and password when creating a connection. This type is the most insecure because the password is saved in plain text in the script. Therefore, this approach is not recommended:

```
client.connect(hostname='myStorage', username='myUser', password='myPassword')
```

> **Note:** If you do not want to manage session handling and `paramiko` methods, you can use the IBM Spectrum Virtualize Python Client (`pysvc`), which is available for download at this GitHub web page.

Using SSH with Perl

Perl requires the usage of an extra external module to connect to IBM Spectrum Virtualize by using the SSH protocol.

`Net::OpenSSH` is a Secure Shell client package that is implemented on the OpenSSH binary client, which is installed by using CPAN.

Example 12-5 shows the use of `Net::OpenSSH` within a Perl script by connecting to `myStorage` by using the user `myUser` and running the commands to display the configured host and controller objects.

Example 12-5 Using Net::OpenSSH within Perl

```
#!/usr/bin/perl

use strict;
use Net::OpenSSH;

my $host = "myStorage";
my $user = "myUser";

my $command1 = "lshost -delim :";
my $command2 = "lscontroller -delim :";

my $ssh = Net::OpenSSH->new("$user\@$host", forward_agent => 1);
$ssh->error and die "SSH connection failed: " . $ssh->error;
print "Connected to $host\n";

my @vDisk = $ssh->capture($command1) or die "Unable to run command";
my @controller = $ssh->capture($command2) or die "Unable to run command";

print @vDisk;
print
"-----------------------------------------------------------------------------------\n";
print @controller;

$ssh->disconnect();
```

The `forwared_agent=>1` option defines the use of the `ssh-agent` authentication agent for the SSH key-based authentication.

12.2.3 SMI-S

The *Storage Management Initiative Specification* (SMI-S) is a common standard that was developed and maintained by the Storage Network Industry Association (SNIA). SMI-S also was ratified as an ISO standard.

The main objective of SMI-S is the management of heterogeneous storage systems across different vendors.

Because SMI-S was available before the REST API was introduced, several products, such as IBM Spectrum Protect Snapshot, still use this interface.

SMI-S consists of the following three main components:

- Common Information Model (CIM)

  The CIM is an open standard that defines how managed elements are represented as a set of objects and their relationships in an IT environment.
- Web-Based Enterprise Management standards (WBEM)

  WBEM is a set of standards that enable computers and other network devices to be managed by using a standard web browser.
- Service Location Protocol (SLP)

  The SLP is a service discovery protocol that allows computers and other devices to find services in a LAN without prior configuration.

Python requires the use of an extra external module `pywbem` to connect to IBM Spectrum Virtualize by using the SMI-S interface.

The script that is shown in Example 12-6 shows the basic use of `pywbem`.

Example 12-6 Basic use of pywbem

```
#!/usr/bin/python

import pywbem
import getpass

mystorage = 'myStorage'
url = 'https://' + mystorage

username = 'myUser'
password = getpass.getpass()

wbemc = pywbem.WBEMConnection(url, (username, password), 'root/ibm', no_verification=True)
cluster = wbemc.EnumerateInstances('IBMTSSVC_Cluster')
print(cluster[0].items())
```

In this example, `WBEMConnection()` establishes HTTPS connection with WBEM services of IBM Spectrum Virtualize controller. Here, target storage system URL is specified by the URL argument. The username and password and the CIM namespace (`root/ibm`) to query also are provided in the next lines.

Note: The `getpass` module is not necessary to work with SMI-S because its purpose is to securely read passwords from standard input with the terminal echo function switched off to hide what is entered.

The `no_verification=True` argument disables SSL certificate verification. That is, it forces the script to trust any certificate that is provided by the WBEM server.

After the connection is successfully established, instances of a specific CIM class can be enumerated by using the `EnumerateInstances()` method, which returns a complex data structure (a list of `CIMInstance()` classes). As shown in Example 12-6 on page 580, it is done over the `IBMTSSVC_Cluster` class, which represents system-level information that is comparable with the results of running the **`lssystem`** command.

Different CIM classes are available for comprehensive management of the IBM SAN Volume Controller system, including the following examples:

- `IBMTSSVC_Cluster`: System level information
- `IBMTSSVC_Node`: Information about nodes
- `BMTSSVC_ConcreteStoragePool`: MDisk groups
- `IBMTSSVC_BackendVolume`: MDisks
- `IBMTSSVC_StorageVolume`: VDisk information

This section gives a brief overview of these CIM classes to illustrate SMI-S capabilities, but it does not provide full list of these classes or their descriptions. For more information about IBM SAN Volume Controller WBEM/CIM classes, their purposes, and relationship diagrams, see *IBM Spectrum Virtualize: Interfacing Using the RESTful API*.

The last line of the script parses and prints the data. But it is not the only way to complete the job. Python is a flexible language and it performs work in different ways. Several approaches of processing the data that is acquired by `EnumerateInstances()` for several CIM classes are listed in Example 12-7.

Example 12-7 Parsing EnumerateInstances() output for classes cluster, nodes, and storage pools

```
print('Cluster information')
cluster = wbemc.EnumerateInstances('IBMTSSVC_Cluster')
print(cluster[0]['ElementName'])

for c_prop in cluster[0]:
    print('\t{prop}: "{val}"'.format(prop=c_prop, val=cluster[0].properties[c_prop].value))
print('Nodes information')

nodes = wbemc.EnumerateInstances('IBMTSSVC_Node')
for node in nodes:
    print(node['ElementName'])
    for n_prop in node:
        print('\t{prop}: "{val}"'.format(prop=n_prop, val=node[n_prop]))
print('Pools information')

pools = wbemc.EnumerateInstances('IBMTSSVC_ConcreteStoragePool')
print('PoolID', 'NumberOfBackendVolumes', 'ExtentSize', 'UsedCapacity',
      'RealCapacity', 'VirtualCapacity', 'TotalManagedSpace', sep=',')

for pool in pools:
    print(
        pool['ElementName'], pool['NumberOfBackendVolumes'], pool['ExtentSize'],
        pool['UsedCapacity'], pool['RealCapacity'], pool['VirtualCapacity'],
        pool['TotalManagedSpace'], sep=','
    )
```

Using similar, yet different approaches, `Cluster information` and `Nodes information` sections of the example parse data in key/value pairs to show all acquired data. However, the `Pools information` part filters data to print selected fields only. It wastefully ignores all other fields.

For some classes, such as `IBMTSSVC_StorageVolume`, full enumeration of all the instances can be slow and can generate several megabytes of unnecessary data. This data must be prepared by the storage system, passed over the network, and then, parsed by the script. Fortunately, it is possible to significantly reduce such data flows by requesting limited amount of necessary information only.

As shown in Example 12-8, by using the `ExecQuery()` method, the WBEM server can be requested in a convenient query language, which is similar to SQL.

Example 12-8 Querying only required data using the ExecQuery() method

```
print('Vdisks:')
vdisks = wbemc.ExecQuery(
    'DMTF:CQL',
    "SELECT VolumeId, VolumeName, NumberOfBlocks FROM IBMTSSVC_StorageVolume"
    " WHERE VolumeName LIKE 'vdisk.'"
)
for vd in vdisks:
    print(vd['VolumeId'], vd['VolumeName'], vd['NumberOfBlocks'], sep=',')
```

Two dialects (CIM Query Language [DMTF:CQL] and WBEM Query Language [WQL]) are recognized by PyWBEM and both can be used with IBM Spectrum Virtualize. However, we use the DMTF:CQL syntax in the examples in this chapter. The DMTF specification (DSP0202) for CQL can be found in *CIM Query Language Specification*.

One of the advantages of SMI-S on IBM SAN Volume Controller is its capability to collect performance data of various storage system components by using “Statistic” family CIM classes, as shown in the following examples:

- `IBMTSSVC_BackendVolumeStatistics`
- `IBMTSSVC_FCPortStatistics`
- `IBMTSSVC_NodeStatistics`
- `IBMTSSVC_StorageVolumeStatistics`

A detailed, with commentaries, example of performance data collecting, and processing script is shown in Example 12-9. It works with `IBMTSSVC_StorageVolumeStatistics` to retrieve VDisks statistics, as shown in Example 12-9.

Example 12-9 Accessing performance metrics by using the PyWBEM module

```
import pywbem
import getpass
import time

mystorage = 'myStorage'
myuser = 'myUser'
mypassword = getpass.getpass()
url = 'https://' + mystorage

ofs = ',' # Output field separator
header = ['InstanceID', 'ReadIOs', 'WriteIOs', 'TotalIOs',
   'KBytesRead', 'KBytesWritten', 'KBytesTransferred']
frequency = 5 # Performance collection interval in minutes
def vdisks_perf(wbem_connection, hdr):
```

```
    """Get performance statistics for vdisks"""
    # Form "select" request string
    request = "SELECT " + ','.join(hdr) + " FROM IBMTSSVC_StorageVolumeStatistics"
    result = []
    # Request WBEM
    vd_stats = wbem_connection.ExecQuery('DMTF:CQL', request)
    # parse reply and form a table
    for vds in vd_stats:
        # Handle 'InstanceID' in a specific way
        vde = [int(vds.properties[hdr[0]].value.split()[1])]
        # Collect the rest of numeric performance fields
        for fld in header[1:]:
            vde.append(int(vds.properties[fld].value))
        result.append(vde)
    return result

def count_perf(new, old, interval):
"""Calculate performance delta divided by interval to get per second values"""
    result = []
    for r in range(0, len(new)):
        row = [new[r][0]]                     # InstanceID
        for c in range(1, len(new[0])):
            row.append(round(float(new[r][c] - old[r][c]) / interval, 2))
result.append(row)
    return result
def print_perf(stats, hdr):
    """Printout performance data matrix"""
    # Print header
    print(ofs.join(str(fld) for fld in hdr))
    # Print performance table
    for ln in stats:
        print('{}{}{}'.format(ln[0], ofs, ofs.join(str(fld) for fld in ln[1:])))

# Connect with WBEM/CIM services
wbemc = pywbem.WBEMConnection(url, (myuser, mypassword), 'root/ibm', no_verification=True)

# Infinite performance processing loop
new_perf = vdisks_perf(wbemc, header)
while True:
    old_perf = new_perf
    new_perf = vdisks_perf(wbemc, header)
    delta_perf = count_perf(new_perf, old_perf, frequency * 60)
    print_perf(delta_perf, header)
    time.sleep(frequency * 60)
```

12.2.4 HTTPS and REST API on IBM Spectrum Virtualize

In this section, we discuss various ways in which the REST API can be used by using Curl, Python, and Perl.

We do not provide a recommendation for which programming language is to be used regarding the REST API.

Although Curl offers to test individual REST API calls quickly, Python and Perl are suitable for more complex tasks in which several REST API calls are to be run depending on each other.

Curl

Table 12-2 shows the `curl` command options.

Table 12-2 Options of the curl command

Command option	Description	Notes
`curl`	This is the executable that is sending the request to the server.	
`-k`	By default, every SSL connection curl makes is verified to be secure. This option allows curl to proceed and operate, even for server connections otherwise considered insecure.	If you are using a signed SSL certificate, you do not need this option.
`-H 'Key:Value'`	Send the information in the quote as a header. Key is the name of the header - describing what specific header is being sent. Value is the value for the key.	
`X-Auth-Username`	The username that you use to log in.	Only used for initial authentication.
`X-Auth-Password`	The password that you use to log in.	Only used for initial authentication.
`X-Auth-Token`	The authentication token that is used to authenticate the REST calls after authentication is complete.	Only used for running commands, not for the authentication.
`Content-Type:application/json`	Tells the server to send the result back in JSON format.	
`https://{{cluster IP or DNS name}}:7443/rest/v1/auth`	The URI that you send an authentication request to.	
`https://{{cluster IP or DNS name}}:7443/v1/{{cli command}}`	The URI to which you send a CLI command.	
`-d '{{DATA}}'`	The `-d` flag is used to send the CLI options, encoded in JSON.	

Creating an authentication token

Example 12-10 shows how to authenticate at the REST API endpoint. The successful authentication creates an authentication token for further use with the REST API.

Example 12-10 Creating a JSON Web Token (JWT)

```
curl -k -X POST -H 'Content-Type:application/json' -H 'X-Auth-Username: MyUser' -H
'X-Auth-Password: MyPassword' https://myStorage:7443/rest/v1/auth

{"token": "4d8916c21058db218d623df51c33f5f01cefeafc988ed7af78c1c51b4a104212"}
```

Query for all configured MDisks

Example 12-11 shows how to use the REST API to get a list of all MDisks by using the formerly generated authentication token.

Example 12-11 Get all M0Disks

```
curl -k -X POST -H 'Content-Type:application/json' -H 'X-Auth-Token:
4d8916c21058db218d623df51c33f5f01cefeafc988ed7af78c1c51b4a104212'
https://myStorage:7443/rest/v1/lsmdisk

[{ "id": "0", "name": "mdisk0", "status": "online", "mode": "array", "mdisk_grp_id": "0",
"mdisk_grp_name": "Pool0", "capacity": "21.7TB", "ctrl_LUN_#": "", "controller_name"
: "", "UID": "", "tier": "tier1_flash", "encrypt": "no", "site_id": "", "site_name": "",
"distributed": "yes", "dedupe": "no", "over_provisioned": "no", "supports_unmap": "ye
s" }]
```

Because this output is difficult to read, add "| `python -m json.tool`" to get a better readable output (see Example 12-12).

Example 12-12 Piping the output to python for getting better readable JSON output

```
curl -k -X POST -H 'Content-Type:application/json' -H 'X-Auth-Token:
4d8916c21058db218d623df51c33f5f01cefeafc988ed7af78c1c51b4a104212'
https://10.1.1.10:7443/rest/v1/lsmdisk | python -m json.tool

[
    {
        "UID": "",
        "capacity": "21.7TB",
        "controller_name": "",
        "ctrl_LUN_#": "",
        "dedupe": "no",
        "distributed": "yes",
        "encrypt": "no",
        "id": "0",
        "mdisk_grp_id": "0",
        "mdisk_grp_name": "Pool0",
        "mode": "array",
        "name": "mdisk0",
        "over_provisioned": "no",
        "site_id": "",
        "site_name": "",
        "status": "online",
        "supports_unmap": "yes",
        "tier": "tier1_flash"
    }
]
```

Python

The script that is shown in Example 12-13 shows an example of the authentication and creation of an access token for further use in the context of querying all available MDisks.

The output of the script provides the following information about each MDisk:

- Name
- Name of the providing controller
- Name of the MDisk group
- Capacity
- Status

The script uses the `getpass` module to prompt for the password and prevent the storage of the credentials in clear text within the script.

Example 12-13 Using the REST API by using Python

```
#!/usr/bin/python

import json
import requests
import getpass

myStorage = 'myStorage'
myUser = 'myUser'
myPassword = getpass.getpass()

### disable SSL verification
ssl_verify = False

### ignore warning for SSL not being used
from requests.packages.urllib3.exceptions import InsecureRequestWarning
requests.packages.urllib3.disable_warnings(InsecureRequestWarning)

### get session token
tokenRequest = requests.post('https://' + myStorage + ':7443/rest/v1/auth',
        headers={
                'Content-type':         'application/json',
                'X-Auth-Username':      myUser,
                'X-Auth-Password':      myPassword
                },
        params="", data="", verify=ssl_verify)

### convert to JSON
_token = json.loads(tokenRequest.text)
token = _token['token']

### get mdisks
mdiskRequest = requests.post('https://' + myStorage + ':7443/rest/v1/lsmdisk',
        headers={
                'Content-type':         'application/json',
                'X-Auth-token':         token
                },
        params="", data="", verify=ssl_verify)

_mdisks = json.loads(mdiskRequest.text)

print( '{:32.32s} {:20.20s} {:32.32s} {:8.8s} {:10.10s}' \
        .format("name","controller_name","mdisk_grp_name","capacity","status") )
```

```
for mdisk in _mdisks:
        print( '{:32.32s} {:20.20s} {:32.32s} {:8.8s} {:10.10s}' \
.format(mdisk['name'],mdisk['controller_name'],mdisk['mdisk_grp_name'],mdisk['capacity'],md
isk['status']) )
```

The use of the `verify=False` option allows insecure SSL connections. By default, every SSL connection is verified to be secure. This option allows the request to get proceed; otherwise, the connection is considered insecure. If you use a signed SSL certificate, you do *not* need this option.

Perl

The script that is shown in Example 12-14 shows an example of the authentication and creation of an access token for further use in the context of querying all available MDisks.

The output of the script provides the following information about each MDisk:

- Name
- Name of the providing controller
- Name of the MDisk group
- Capacity
- Status

The script uses the `IO::Prompter` module to prompt for the password and prevent the storage of the credentials in clear text within the script.

Example 12-14 Using the REST API by using Perl

```
#!/usr/bin/perl

use strict;
use JSON;
use REST::Client;
use IO::Prompter;

my $myStorage = 'myStorage';
my $myUser = 'myUser';

my $myPassword = prompt 'Please enter your password:', -echo=>"*";

my $restURL = 'https://' . $myStorage . ':7443/rest/v1/';

### get the session token
my $tokenRequest = REST::Client->new();
$tokenRequest->addHeader('Content-type', 'application/json');
$tokenRequest->addHeader('X-Auth-Username' , $myUser);
$tokenRequest->addHeader('X-Auth-Password', $myPassword);
$tokenRequest->getUseragent()->ssl_opts('verify_hostname' => 0);
$tokenRequest->POST($restURL . '/auth');
my $token = decode_json($tokenRequest->responseContent())->{'token'};

### get the list of mdisks
my $mdiskRequest = REST::Client->new();
$mdiskRequest->addHeader('Content-type', 'application/json');
$mdiskRequest->addHeader('X-Auth-Token', $token);
$mdiskRequest->getUseragent()->ssl_opts('verify_hostname' => 0);
$mdiskRequest->POST($restURL . '/lsmdisk');

my $mdiskList = $mdiskRequest->responseContent();
my @mdiskListJSON = @{decode_json($mdiskList)};
```

```
for my $key (@mdiskListJSON) {
        printf "%32s %20s %32s %8s %10s\n",
                $key->{'name'},
                $key->{'controller_name'},
                $key->{'mdisk_grp_name'},
                $key->{'capacity'},
                $key->{'status'};
}
```

The use of the `getUseragent()->ssl_opts('verify_hostname' => 0)` method allows insecure SSL connections. By default, every SSL connection is verified to be secure. This option allows the request to proceed; otherwise, the connection is considered insecure. If you use a signed SSL certificate, you do *not* need this option.

12.2.5 HTTPS and REST API on IBM Spectrum Control

You can use the Representational State Transfer (REST) API for IBM Spectrum Control to access information about resources and to generate custom capacity, configuration, and performance reports.

The main advantage of this method is that it allows to get information about the entire SAN and storage infrastructure that is managed by the IBM Spectrum Control server (see Example 12-15).

Example 12-15 Using the IBM Spectrum Control REST API by using Python

```
#!/usr/bin/python

import requests
import getpass
username = 'myUser'
password = getpass.getpass()

url = 'https://spectrumcontrol-server:9569/srm/'

sesssion = requests.Session()
sesssion.verify = False

response = session.post(url + 'j_security_check',
                      data={'j_username': username, 'j_password': password})
response.raise_for_status()

response = session.get(url + 'REST/api/v1/' + 'StorageSystems')
response.raise_for_status()

print(response.json())
```

12.3 Automation with Red Hat Ansible

Automation is a priority for maintaining today's busy storage environments. Automation software allows the creation of repeatable sets of instructions. It also reduces the need for human interaction with computer systems.

Red Hat Ansible and other third-party automation tools are becoming increasingly used across the enterprise IT environments. It is not unexpected that their use in storage environments will become more popular.

12.3.1 Red Hat Ansible

The IBM SAN Spectrum Virtualize Controller family includes integration with Red Hat Ansible Automation Platform. This integration allows IT to create an Ansible playbook that automates repetitive tasks across an organization in a consistent way, which helps improve outcomes and reduces errors.

Ansible is an agentless automation management tool that uses the SSH protocol. As of this writing, Ansible can be run from any machine with Python 2 (version 2.7) or Python 3 (versions 3.5 and higher) installed. Supported platforms for Ansible include Red Hat, SUSE, Debian, CentOS, macOS, and any of the Berkeley Software Distribution (BSD) versions.

> **Note:** Windows is not supported for the Ansible control node.

12.3.2 Red Hat Ansible Editions

The following Red Hat Ansible Editions are available:

- Ansible Core

  Ansible Core is the command-line tool that is installed from community repositories or the official Red Hat repositories for Ansible.

- Ansible Tower

  Ansible Tower is the GUI tool that is used to run Ansible tasks. Tower requires a license that is based on the number of systems Ansible Tower is to manage. Ansible Tower is available as Standard or Premium Edition, whereby the difference is the 24x7 support that is included in the Premium Edition.

12.3.3 Requirements

Ansible server (Control Node) features the following requirements:

- Python 2 (version 2.7) or Python 3 (versions 3.5 and higher)

  > **Note:** Some plug-ins that run on the control node include other requirements. These requirements are listed in the plug-in documentation.

- Host requirements:
  - Although you do not need a daemon on your managed nodes, you need a way for Ansible to communicate with them.
  - For most managed nodes, Ansible makes a connection over SSH and transfers modules by using SFTP. If SSH works but SFTP is not available on some of your managed nodes, you can switch to SCP in `ansible.cfg`.
  - For any machine or device that can run Python, you also need Python 2 (version 2.6 or later) or Python 3 (version 3.5 or later).

Note: Some modules feature more requirements that must be met on the 'target' machine (the managed node). These requirements are listed in the module documentation.

12.3.4 Essential terminology in an Ansible environment

Ansible environment features the following essential terminology:

- Ansible Galaxy: A hub for finding and sharing Ansible content.
- Ansible server: The machine with Ansible installed, which runs all tasks and playbooks.
- Playbook: A framework where Ansible automation tasks are defined (written in YAML).
- Task: A section that contains a single procedure you want to be run.
- Tag: A name that you can assign to a task.
- Play: The execution of a playbook.
- Hosts: The devices that you manage with Ansible.
- Modules: A command or set of commands that are made for execution on the client side.
- Handler: A task that is called only if a notifier is present.
- Notifier: A section that is assigned to a task that calls a handler if the output is changed.
- Inventory: A file that contains Ansible client/server data.
- Fact: Information that is retrieved from the client from global variables by using the `gather-facts` operation.
- Roles: A structured way of grouping tasks, handlers, variables, and other properties.
- Container: Ansible Container uses Ansible roles to build images, initialize projects, and add services to projects.

12.3.5 Automating IBM Storage with Ansible

IBM data storage provides simple storage solutions that address modern data requirements and provides a solution to your hybrid multicloud strategy.

With the speed, scale, and complexity of hybrid multicloud and even traditional on-premises environments, automation became a priority.

IBM FlashSystem family for hybrid multicloud includes integration with Red Hat Ansible Automation Platform. It allows IT to create an Ansible playbook that automates the tasks that are repeated across an organization in a consistent way, which helps improve outcomes and reduces risk.

It also standardizes how IT and application owners interact together and features the following benefits:

- With Red Hat Ansible Automation Platform and IBM Storage, customers can easily automate tasks, such as configuration management, provisioning, workflow orchestration, application deployment, and life-cycle management.
- By using Red Hat Ansible Automation Platform and IBM Storage, customers can reduce system inconsistencies with the automation modules.
- Red Hat Ansible Automation Platform can also be used to configure end-to-end infrastructure in an orchestrated fashion.
- Ansible provides a single pane of glass visibility to multi cluster, multicloud environments, which allows lines of business to use playbooks to accomplish their goals without needing to understand the details of how the work is being done.

IBM is a Red Hat-certified support module vendor that provides simple management for the following commands that are used in the IBM Spectrum Virtualize Ansible Collection:

- Collect facts: Collect basic information, including hosts, host groups, snapshots, consistency groups, and volumes
- Manage hosts: Create, delete, or modify hosts
- Manage volumes: Create, delete, or extend the capacity of volumes
- Manage MDisk: Create or delete a managed disk
- Manage pool: Create or delete a pool (managed disk group)
- Manage volume map: Create or delete a volume map
- Manage consistency group snapshot: Create or delete consistency group snapshots
- Manage snapshot: Create or delete snapshots
- Manage volume clones: Create or delete volume clones

This collection provides a series of Ansible modules and plug-ins for interacting with the IBM Spectrum Virtualize family storage products. The modules in the IBM Spectrum Virtualize Ansible collection use the REST API to connect to the IBM Spectrum Virtualize storage system. These products include:

- IBM SAN Volume Controller
- IBM FlashSystem family members that are built with IBM Spectrum Virtualize (FlashSystem 5000, 5100, 5200, 7200, 9100, 9200, 9200R, and V9000
- IBM Storwize family
- IBM Spectrum Virtualize for Public Cloud

For more information, see *Automate and Orchestrate Your IBM FlashSystem Hybrid Cloud with Red Hat Ansible*, REDP-5598.

For IBM Spectrum Virtualize modules, Ansible version 2.9 or higher is required. For more information about IBM Spectrum Virtualize modules, see this web page.

12.3.6 Getting started

The Ansible Collection (`ibm.spectrum_virtualize`) provides a series of Ansible modules and plug-ins for interacting with the IBM Spectrum Virtualize family storage products,.

As of this writing, the Ansible collection for IBM Spectrum Virtualize is available in version 1.6.0.

All information in this section are based on this version.

Prerequisites for using the modules

`Paramiko` must be installed to use i**`bm_svctask_command`** and **`ibm_svcinfo_command`** modules.

`Paramiko` is a Python (2.7, 3.4+) implementation of the SSHv2 protocol, and provides client and server functions.

Although `Paramiko` is a Python C extension for low-level cryptography, it is a pure Python interface around SSH networking concepts.

Current limitations

The modules in the IBM Spectrum Virtualize Ansible collection use the REST API to connect to the IBM Spectrum Virtualize storage system.

This collection has the following limitations:

- Using the REST API to list more than 2000 objects might create a loss of service from the API side because it automatically restarts because of memory constraints.
- The Ansible collection can run on all IBM Spectrum Virtualize storage versions that are 8.1.3, except versions 8.3.1.3 and 8.3.1.4.
- It is not possible to access the REST API by using a cluster IPv6 address.

Prerequisites

Ensure that the following prerequisites are met:

- Ansible and is installed and configured on a controller node
- Ansible Galaxy Collection `ibm.spectrum_virtualize` is installed on the same controller node.
- Network access is available from the controller node to Spectrum Virtualize Management IP.
- A user with the necessary is available permissions to create or delete objects on IBM Spectrum Virtualize.
- IBM Spectrum Virtualize operates at version 8.1.3 or higher.

Installing or upgrading Ansible Galaxy Collection ibm.spectrum_virtualize

To install the IBM Spectrum Virtualize collection that is hosted in Galaxy, use the following command:

```
ansible-galaxy collection install ibm.spectrum_virtualize
```

To upgrade to the latest version of the IBM Spectrum Virtualize collection, use the following command:

```
ansible-galaxy collection install ibm.spectrum_virtualize --force
```

Functions provided by IBM Spectrum Virtualize Ansible modules

The `ibm.spectrum_virtualize` collection provides the following modules:

- `ibm_svc_auth`: Generates an authentication token for a user on the IBM Spectrum Virtualize family storage system.
- `ibm_svc_host`: Manages hosts that are on IBM Spectrum Virtualize system.
- `ibm_svc_hostcluster`: Manages the host cluster that is on IBM Spectrum Virtualize system.
- `ibm_svc_info`: Collects information about the IBM Spectrum Virtualize system.
- `ibm_svc_manage_consistgrp_flashcopy`: Manages the FlashCopy consistency groups that are on IBM Spectrum Virtualize system.
- `ibm_svc_manage_cv`: Manages the change volume in remote copy replication that is on the IBM Spectrum Virtualize system.
- `ibm_svc_manage_flashcopy`: Manages the FlashCopy mappings that are on IBM Spectrum Virtualize system.
- `ibm_svc_manage_mirrored_volume`: Manages the mirrored volumes that are on the IBM Spectrum Virtualize system.
- `ibm_svc_manage_migration`: Manages the volume migration between clusters that are on the IBM Spectrum Virtualize system.
- `ibm_svc_manage_replication`: Manages the remote copy replication that is on the IBM Spectrum Virtualize system.
- `ibm_svc_manage_replicationgroup`: Manages the remote copy consistency group on IBM Spectrum Virtualize system.
- `ibm_svc_manage_volume`: Manages the standard volumes on IBM Spectrum Virtualize system.
- `ibm_svc_manage_volumegroup`: Manages the volume groups that are on IBM Spectrum Virtualize system.
- `ibm_svc_mdisk`: Manages the MDisks for IBM Spectrum Virtualize system.
- `ibm_svc_mdiskgrp`: Manages pools for IBM Spectrum Virtualize system.
- `ibm_svc_start_stop_flashcopy`: Starts or stops the FlashCopy mapping and consistency groups that are on IBM Spectrum Virtualize system.
- `ibm_svc_start_stop_replication`: Starts or stops the remote copy relationship or group on IBM Spectrum Virtualize system.
- `ibm_svc_vol_map`: Manages the volume mapping for IBM Spectrum Virtualize system.
- `ibm_svcinfo_command`: Runs the **`svcinfo`** CLI command on the IBM Spectrum Virtualize system over an SSH session.
- `ibm_svctask_command`: Runs the **`svctask`** CLI commands on the IBM Spectrum Virtualize system over and SSH session.

> **Note:** Beginning with version 1.6.0, the `ibm_svc_vdisk` module is considered a deprecated feature. A new module (`ibm_svc_manage_volume`) was introduced to manage standard volumes.

Getting help for IBM Spectrum Virtualize Ansible modules

To get the online documentation for a specific module that is displayed, use the following command:

```
ansible-doc <collection-name>.<module-name>
```

The output of the help includes all permissible options and some examples of how to use the module (see Example 12-16).

Example 12-16 Example displaying online help

```
ansible-doc ibm.spectrum_virtualize.ibm_svc_manage_volume

> IBM_SVC_MANAGE_VOLUME
Ansible interface to manage 'mkvolume', 'rmvolume', and 'chvdisk' volume commands.

  * This module is maintained by The Ansible Community
OPTIONS (= is mandatory):

- buffersize
        Specifies the pool capacity that the volume will reserve as a buffer for
thin-provissioned and compressed volumes.
        Parameter 'thin' or 'compressed' must be specified to use this parameter.
        The default buffer size is 2%.
        `thin' or `compressed' is required when using `buffersize'.
        Valid when `state=present', to create a volume.
        [Default: (null)]
        type: str

= clustername
        The hostname or management IP of the Spectrum Virtualize storage system.

        type: str
:
:
:
```

12.3.7 Securing credentials in Ansible

While working with Ansible, you can create several playbooks, inventory files, variable files, and so on. Some of the files might contain sensitive data, such as access credentials. To protect this kind of data, Ansible provides the Ansible Vault, which helps to prevent this data from being exposed. Sensitive data and passwords are kept in an encrypted file rather than in plain text files.

12.3.8 Creating an Ansible playbook

The playbook automates in a consistent manner the tasks that are repeated across an organization, which improves outcomes and reduces risk. It also standardizes how IT and application owners interact.

In this section, we discuss creating an Ansible playbook. The creation of the playbook is based on the use case that is used here.

For a new VMware ESX cluster, consisting of two new servers, two VDisks are to be created and mapped to the hostcluster object.

> **Note:** `Idempotence` is a property that might be included in a mathematics or computer science operation. It roughly means that an operation can be carried out multiple times without changing the result.
>
> The IBM Spectrum Virtualize Ansible modules provide `idempotency` in Ansible playbooks.
>
> The IBM Spectrum Virtualize Ansible modules check whether the object to be created exists in the defined state and does not attempt to create it again.

Table 12-3 lists the variable parameters and their values for the example playbook.

Table 12-3 Variable parameters and their values for the example playbook

Attribute	Value
Name of new host cluster	ESX-Cluster-1
Name of new host 1	ESX-Host-1
WWPNs of new host 1	100000109C400798, 1000001AB0440446
Name of new host 2	ESX-Host-2
WWPNs of new host 2	100000109B600424, 1000001BC0660146
Name of VDisk 1	Datastore1
Name of VDisk 2	Datastore2

Step 1: Authentication

Example 12-17 shows the required YAML notation for the part of the playbook to authenticate at the IBM Spectrum Virtualize REST API to obtain a token for further use. To avoid storing the password in clear text within the playbook, the password was encrypted in a vault.

Example 12-17 YAML notation for obtaining an authentication token

```
vars:
    clustername: <Cluster management ip | hostname>
    domain: <FQDN>
    username: myuser
    password: !vault |
              $ANSIBLE_VAULT;1.1;AES256
62653531313434393266646438306537396264306433653638343439643136333238383139616561
65303734306362653166396262343763363066303433333640a3263326265646562333233336333239
39633132656631353030386430663736363631656438343364346235653534316333333233333531
316634326362653836 0a6336646162643261336433393363333636383232323739623938393566 37
              6138
tasks:
    - name: Obtain an authentication token
      register: result
      ibm_svc_auth:
        clustername: "{{ clustername }}"
        domain: "{{ domain }}"
        username: "{{ username }}"
        password: "{{ password }}"
```

For more informations about how to work with Ansible vaults, see this Ansible Documentation web page.

Step 2: Creating the host cluster object

Example 12-18 shows the required YAML notation for the part of the playbook to create an empty host cluster object.

Example 12-18 YAML notation for creating an empty host cluster

```
- name: Define a new host cluster
      ibm_svc_hostcluster:
        clustername: "{{ clustername }}"
        domain: "{{ domain }}"
         token: "{{ result.token }}"
         log_path: "{{ log_path }}"
        name: <hostcluster_name>
        state: present
```

Step 3: Creating an FC host

Example 12-19 shows the required YAML notation for the part of the playbook to create an FC host object.

Example 12-19 YAML notation for creating a new FC host object

```
- name: Define a new FC host
      ibm_svc_host:
        clustername: "{{ clustername }}"
        domain: "{{ domain }}"
        token: "{{ result.token }}"
        log_path: "{{ log_path }}"
        name: "{{ hostname }}"
        state: present
        fcwwpn: "{{ fcwwpn(s) }}"
        iogrp: 0:1:2:3
        protocol: scsi
        type: generic
         hostcluster: "{{ hostcluster_name }}"
```

Step 4: Creating a thin-provisioned volume

Example 12-20 shows the required YAML notation for the part of the playbook to create a thin-provisioned volume.

Example 12-20 YAML notation to create a thin-provisioned volume

```
- name: Create a thin-provisioned volume
      ibm_svc_manage_volume:
        clustername: "{{ clustername }}"
        domain: "{{ domain }}"
        token: "{{ result.token }}"
        log_path: "{{ log_path }}"
        name: "volume_name"
        state: "present"
        pool: "<pool_name>"
        size: "<size>"
```

```
        unit: "<size_unit>"
        thin: true
        buffersize: 10%
```

Step 5: Mapping the new volume to the host cluster object

Example 12-21 shows the required YAML notation for the part of the playbook to map the new volume to the hostcluster.

Example 12-21 YAML notation to map a volume to the hostcluster

```
    - name: Map a volume to a hostcluster
      ibm_svc_vol_map:
        clustername: "{{ clustername }}"
        domain: "{{ domain }}"
        token: "{{ result.token }}"
        log_path: "{{ log_path }}"
        volname: <volume_name>
        hostcluster: <hostcluste_-name>
        state: present
```

If a SCSI-Id must be specified, use the scsi: `<scsi-id>` parameter.

Putting it all together

Example 12-22 shows the combined required tasks for the use of the IBM Spectrum Virtualize collection to create hostcluster volumes (this use case is used in this chapter) to be used as a playbook by Ansible. All customized lines of the playbook are highlighted in bold in the example.

Example 12-22 Complete playbook for specified use-case

```
- name: Using Spectrum Virtualize collection to create hostcluster - hosts -
volumes
  hosts: localhost
  collections:
    - ibm.spectrum_virtualize
  gather_facts: no
  connection: local

# definition of global variables
  vars:
    clustername: mySVC
    domain: mydomain.com
    username: myuser
    password: !vault |
              $ANSIBLE_VAULT;1.1;AES256
626535313134343932666464383065373962643064336536383434396431363332383831396165661
65303734306362653166396262343763363066303433336 40a326332626564656233323336333239
3963313265666313530303864306637363636316564383433643462356535343163333332333333531
316634326362653836 0a633664616264326133643339336333363638323232373962393839356637
6138
    log_path: /tmp/redbook-example.log

# define variables for running the playbook
```

```
# hostcluster
    hostcluster_name: ESX-Cluster-1

# host 1
    host1_name: ESX-Host-1
    host1_fcwwpn: 100000109C400798{{":"}}1000001AB0440446

# host 2
    host2_name: ESX-Host-2
    host2_fcwwpn: 100000109B600424{{":"}}1000001BC0660146

# pools to use for volume mirror
    pool1_name: pool1
    pool2_name: pool2

#  volume 1
    volume1_name: Datastore1
    volume1_size: '10'
    volume1_size_unit: tb

#  volume 2
    volume2_name: Datastore2
    volume2_size: '10'
    volume2_size_unit: tb

tasks:
# creating an authentication token for further usage within the playbook
    - name: Obtain an authentication token
      register: result
      ibm_svc_auth:
        clustername: "{{  clustername  }}"
        domain: "{{  domain }}"
        username: "{{  username }}"
        password: "{{  password }}"
        log_path: "{{  log_path  }}"

# create the hostcluster object
    - name: Create the hostcluster
      ibm_svc_hostcluster:
        clustername: "{{  clustername  }}"
        domain: "{{  domain  }}"
        token: "{{  result.token  }}"
        log_path: "{{  log_path  }}"
        name: "{{  hostcluster_name  }}"
        state: present

# create first host object
    - name: Define first FC host
      ibm_svc_host:
        clustername: "{{ clustername }}"
        domain: "{{ domain }}"
        token: "{{ result.token }}"
        log_path: "{{ log_path }}"
        name: "{{ host1_name }}"
        state: present
```

```
        fcwwpn: "{{ host1_fcwwpn }}"
        iogrp: 0:1:2:3
        protocol: scsi
        type: generic
        hostcluster: "{{ hostcluster_name }}"

# create second host object
    - name: Define second FC host
      ibm_svc_host:
        clustername: "{{ clustername }}"
        domain: "{{ domain }}"
        token: "{{ result.token}}"
        log_path: "{{ log_path }}"
        name: "{{ host2_name }}"
        state: present
        fcwwpn: "{{ host2_fcwwpn }}"
        iogrp: 0:1:2:3
        protocol: scsi
        type: generic
        hostcluster: "{{ hostcluster_name }}"

# create first mirrored thin-provisioned volume
    - name: Create first thin-provisioned volume
      ibm_svc_manage_volume:
        clustername: "{{ clustername }}"
        domain: "{{ domain }}"
        token: "{{ result.token }}"
        log_path: "{{ log_path }}"
        name: "{{ volume1_name }}"
        state: "present"
        pool: "{{  pool1_name }}:{{ pool2_name }}"
        size: "{{ volume1_size }}"
        unit: "{{ volume1_size_unit }}"
        thin: true
        buffersize: 10%

# create second mirrored thin-provisioned volume
    - name: Create second thin-provisioned volume
      ibm_svc_manage_volume:
        clustername: "{{ clustername }}"
        domain: "{{ domain }}"
        token: "{{ result.token }}"
        log_path: "{{ log_path }}"
        name: "{{ volume2_name }}"
        state: "present"
        pool: "{{ pool1_name }}:{{ pool2_name }}"
        size: "{{ volume2_size }}"
        unit: "{{ volume2_size_unit }}"
        thin: true
        buffersize: 10%

# mapping of first volume to the hostcluster
    - name: Map first volume to the hostcluster
      ibm_svc_vol_map:
        clustername: "{{ clustername }}"
```

```
        domain: "{{ domain }}"
        token: "{{ result.token }}"
        log_path: "{{ log_path }}"
        volname: "{{ volume1_name }}"
        hostcluster: "{{ hostcluster_name }}"
        state: present

# mapping of second volume to the hostcluster
    - name: Map second volume to the hostcluster
      ibm_svc_vol_map:
        clustername: "{{ clustername }}"
        domain: "{{ domain }}"
        token: "{{ result.token }}"
        log_path: "{{ log_path }}"
        volname: "{{ volume2_name }}"
        hostcluster: "{{ hostcluster_name }}"
        state: present
```

12.3.9 More automation

The use case that is described in this chapter can be extended by completing the following steps:

1. Create the required FC zoning.
2. Scan the HBA for the newly created volumes.
3. Create a VMFS data store on the discovered volumes.
4. Create one or more virtual machines (VMs).

For more information about the Brocade FOS FC collection on Ansible Galaxy, see this Ansible web page.

For more information about the `community.vmware` Ansible Collection on Ansible Galaxy, see this Ansible web page.

A

IBM i considerations

The IBM Spectrum Virtualize family of block storage systems including the IBM Flash System 5000 series, IBM FlashSystem 7200, and IBM FlashSystem 9200/9200R provides a broad range of flexible and scalable SAN storage solutions. These solutions can meet demands of IBM i customers for entry to high-end storage infrastructure solutions.

All family members that are based on IBM Spectrum Virtualize software use a common management interface. They also provide a comprehensive set of advanced functions and technologies, such as advanced Copy Services functions, encryption, compression, storage tiering, NVMe flash, and storage class memory (SCM) devices, and external storage virtualization. Many of these advanced functions and technologies also are of interest to IBM i customers who are looking for a flexible, high-performing, and highly available SAN storage solution.

This appendix provides important considerations and guidelines for successfully implementing the IBM Spectrum Virtualize family and its advanced functions with IBM i. Unless otherwise stated, the considerations also apply to previous generations of products like the IBM Storwize family, the IBM Flash System 9100 series, and IBM Flash System V9000.

This appendix includes the following topics:

IBM i Storage management

Because of the unique IBM i storage architecture, special considerations for planning and implementing a SAN storage solution are required (also with IBM Spectrum Virtualize-based storage). This section describes how IBM i storage management manages its available disk storage.

Many host systems require the user to take responsibility for how information is stored and retrieved from the disk units. An administrator must also manage the environment to balance disk usage, enable disk protection, and maintain balanced data to be spread for optimum performance.

The IBM i architecture is different in a way that the system takes over many of the storage management functions, which on other platforms are the responsibility of a system administrator.

IBM i, with its Technology Independent Machine Interface (TIMI), largely abstracts the underlying hardware layer from the IBM i operating system and its users and manages its system and user data 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 (see Figure A-1).

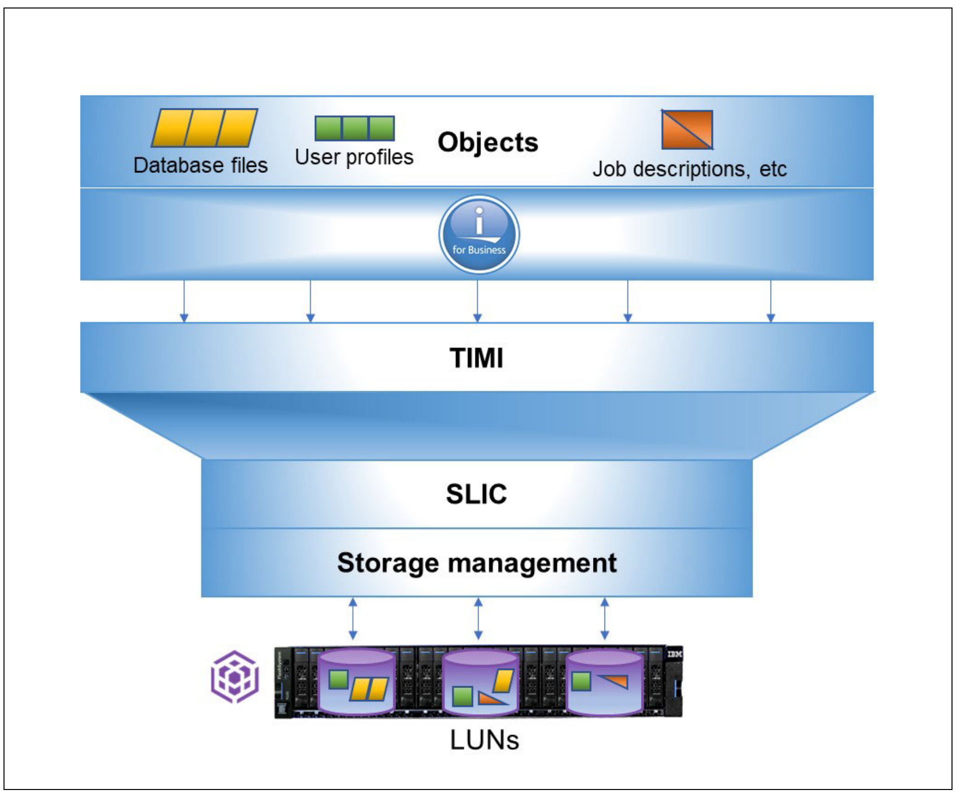


Figure A-1 IBM i storage management spreads objects across LUNs

Note: When a program presents instructions to the machine interface for execution, the interface appears to the program as the system hardware, but it is not. The instructions that are presented to TIMI pass through a layer of microcode before they are understood by the hardware.

As a component of the IBM i System Licensed Internal Code (SLIC), IBM i storage management normally spreads the data in the file across multiple disk units (LUNs when external storage is used). When you add 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 or main storage as one address space. It also uses the same set of virtual addresses to cover main memory and disk space. Paging the objects in this virtual address space is performed in 4 KB pages, as shown in Figure A-2. After a page gets written to disk, it is stored with metadata, including its unique virtual address. For this purpose, IBM i originally used a proprietary 520 bytes per sector disk format.

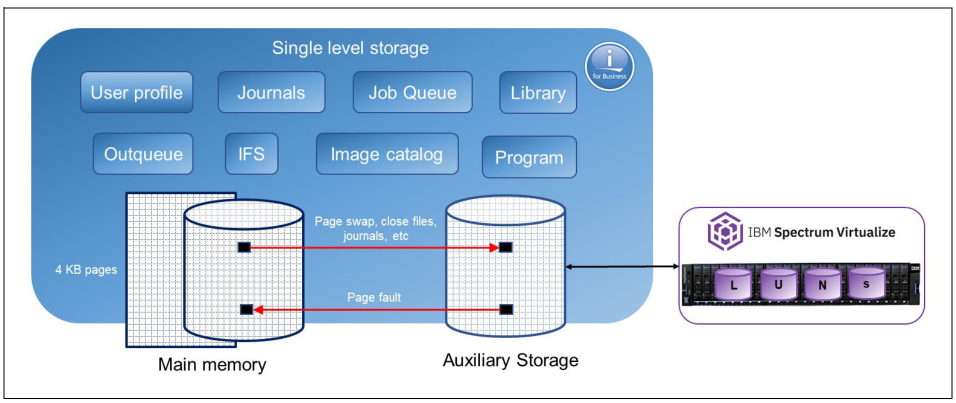


Figure A-2 Virtual address space.

Note: The system storage that is conformed with main storage or main memory and auxiliary storage is addressed in the same way. This single, device-independent addressing mechanism means that objects are referred to by name or name and library, but *never* by disk location. The virtual addressing of IBM i is independent of the physical location of the object, type, capacity, and the number of disks units or LUNs on the system.

The IBM i disk storage space is managed by using auxiliary storage pools. Each IBM i system has a system ASP (ASP 1), which includes the load source (also known as *boot volume* on other systems) as disk unit 1, and optional user ASPs (ASP 2-33). The system ASP and the user ASPs are designated as SYSBAS and constitute the system database.

The single-level storage with its unique virtual addresses also implies that the disk storage that is configured in SYSBAS of an IBM i system must be available in its entirety for the system to remain operational. It cannot be shared for simultaneous access by other IBM i systems.

To allow for sharing of IBM i disk storage space between multiple IBM i systems in a cluster, switchable *independent auxiliary storage pools* (IASPs) can be configured. The IBM i auxiliary storage pools architecture is shown in Figure A-3.

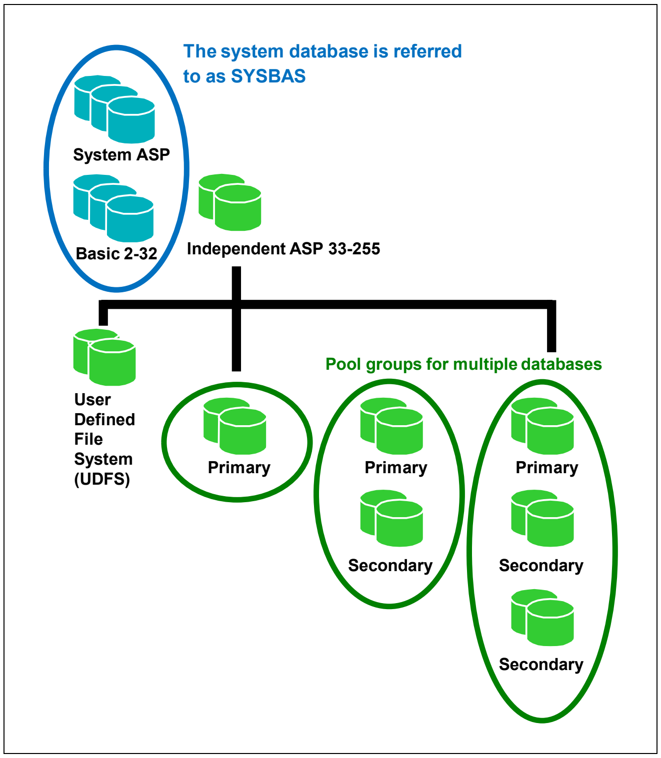


Figure A-3 IBM i auxiliary storage pools architecture

Single-level storage makes main memory work as a large cache. Reads are done from pages in main memory, and requests to disk are done only when the needed page is not there yet.

Writes are done to main memory or main storage, and write operations to disk are performed as a result of swap, file close, or forced write. Application response time depends on disk response time and many other factors.

Other storage-related factors include the IBM i storage pool configuration for the application, how frequently the application closes files, and whether it uses journaling. An example is shown in Figure A-4.

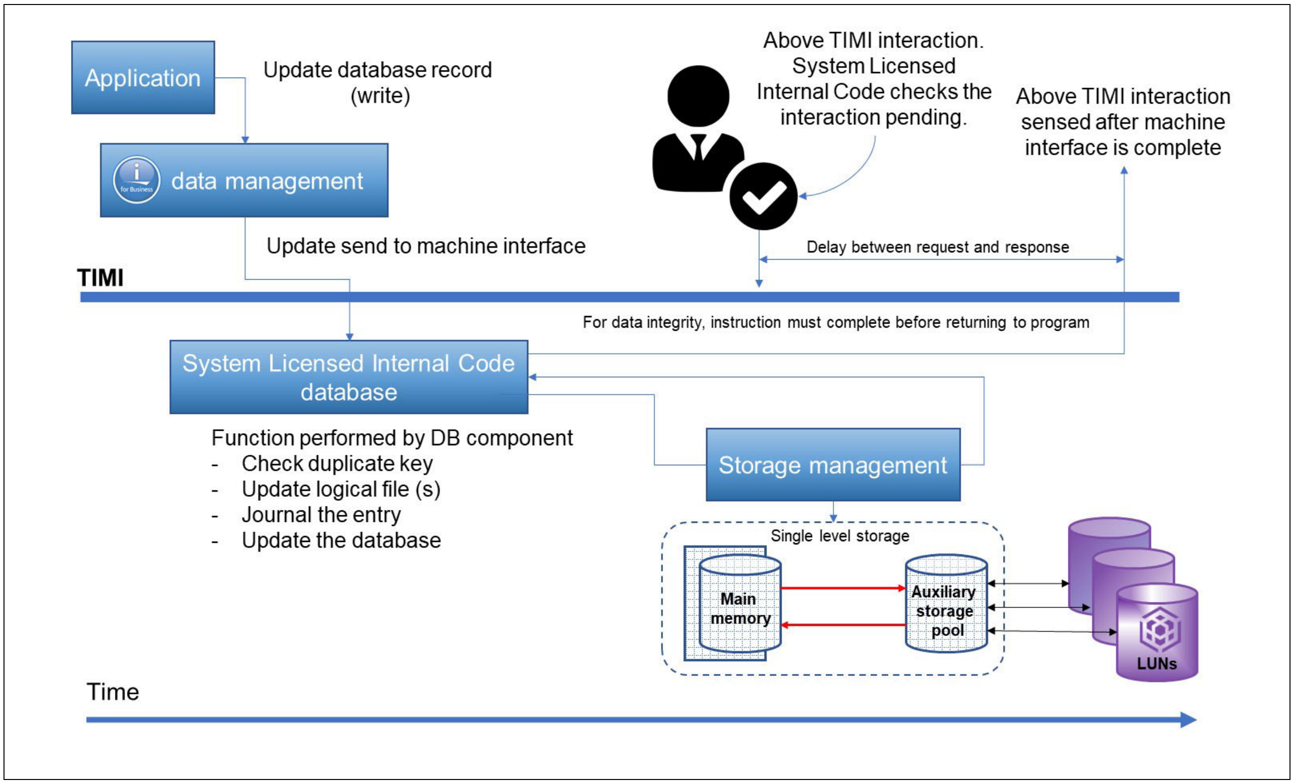


Figure A-4 TIMI atomicity

Note: In Figure A-4 auxiliary storage pool (ASP) is conformed and assigned LUNs from IBM Spectrum Virtualize to the IBM i. It also shows an application requests and updates a database record. Throughout the time that the TIMI task is in progress, an interaction above TIMI can ensue. The interaction does not carry out upon until TIMI task concludes.

IBM i response time

IBM i customers often are concerned about the following types of performance:

- Application response time: The response time of an application transaction. This time is usually critical for the customer.
- Duration of batch job: Batch jobs are usually run during the night or other off-peak periods. The duration of a batch job is critical for the customer because it must be finished before regular daily transactions start.
- Disk response time: Disk response time is the time that is needed for a disk I/O operation to complete. It includes the service time for I/O processing and the wait time for potential I/O queuing on the IBM i host.

  Disk response time can significantly influence application response time and the duration of a batch job. Because the performance of the disk subsystem has a significant effect on overall system performance, this issue is discussed next.

Disk performance considerations

Performance of the disk subsystem significantly affects overall IBM i system performance. especially in a commercial data processing environment in which a large volume of data must proceed. Disk drives or LUNs' response times contribute to a major portion of the overall response time (OLTP) or runtime (batch).

The performance of a disk subsystem is affected by the type of protection (RAID, DRAID, or mirroring).

The amount of free space (GB) on the drives and the extent of fragmentation also has an effect. The reason for the effect is the need to find suitable contiguous space on the disks to create objects or extend objects. Disk space often is allocated in extents of 32 KB. If a 32 KB contiguous extent is not available, two extents of 16 KB are used.

Next, we discuss the following important disk performance considerations:

- Disk I/O requests
- Disk subsystems
- Disk operation
- Asynchronous I/O wait
- Disk protection
- Logical Database I/O versus physical disk I/O

Disk I/O requests

Greater sources of disk request are faults that arise from a request for information not being satisfied by what is in memory. Request to bring information in to memory also result in disk I/O. Memory pages also can be purged from periodically, which results disk I/O activity.

> **Note:** The Set Object Access (SETOBJACC) command on IBM i temporarily changes the speed of access to an object by bringing the object into a main storage pool or purging it from all main storage pools. An object can be kept in main storage resident by selecting a pool for the object that has available space and does not have jobs that are associated with it.
>
> For more information see, this IBM Documentation web page.

Disk subsystems

Typically, an external disk subsystem (storage system) connects a server through a SAN, as shown in Figure .

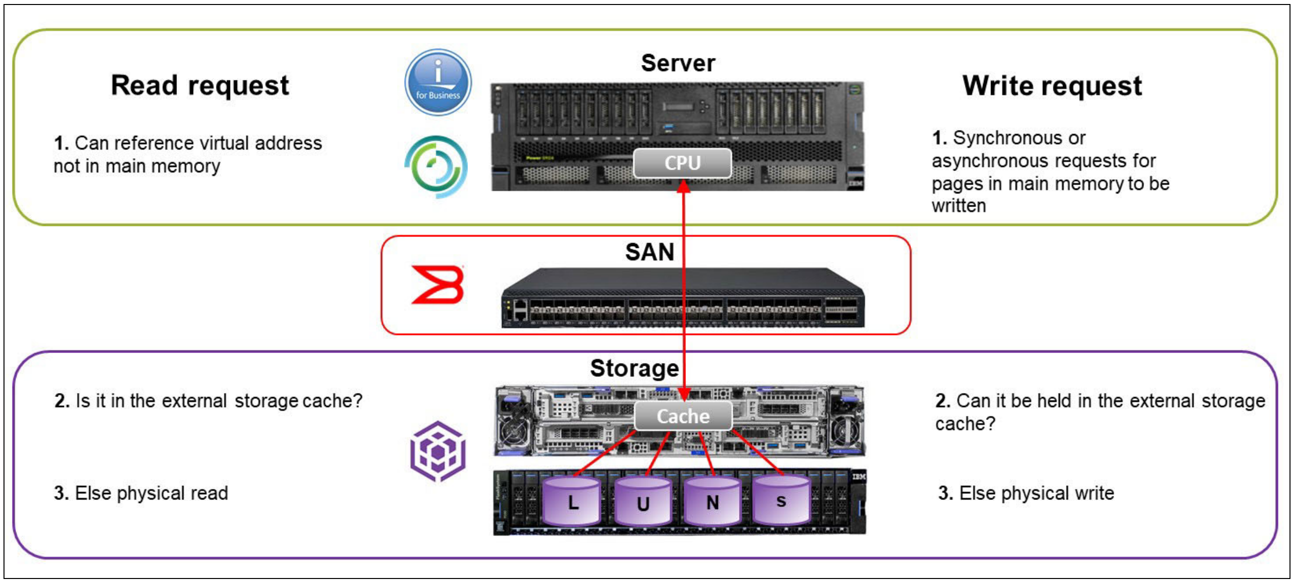


Figure A-5 Disk subsystem

A request information (data or instructions) from the CPU based on user interactions are submitted to the disk subsystem if it cannot be satisfied from the contents of memory. Whether the request can be satisfied from the disk subsystem cache, it responds or forwards the request to the disk drives or LUNs.

Similarly, a write request is retained in memory, unless the operating system determines that it must be written to the disk subsystem. The operating system attempts to satisfy the request by writing to the controller cache.

Note: The QAPMDISKRB from collections services data files in IBM i includes disk file response bucket entries. It also contains one record for each device resource name. It is intended to be used with the QAPMDISK file. For more information, see this IBM Documentation web page.

Disk operation

On IBM i, physical disk I/O requests are categorized as database (physical or logical files) or nondatabase I/Os, as shown in Figure A-6.

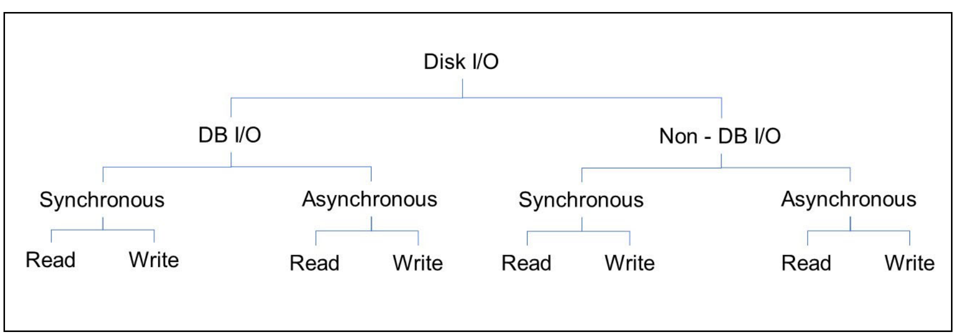


Figure A-6 Disk I/O on IBM i

The time that is taken to respond to synchronous disk I/Os contributes to the Online transaction processing (OLTP) response time or batch runtime. With asynchronous I/O, the progress of a request does not wait for the completion of I/O.

Often, write requests are asynchronous, including journal deposits with commitment control. However, the writes become synchronous if journaling is active without commitment control.

Asynchronous I/O wait

On IBM i, at times jobs might have to wait for *asynchronous* I/O requests to complete. The job issues a request but requires the data sooner than it can be made available by the disk subsystem. When a job waits for *asynchronous* I/O portion of the operation, it becomes synchronous. The time is recorded as *asynchronous* disk I/O wait in the QAPMJOBL file.

JBWIO is the number of times that the process specifically waited for outstanding *asynchronous* I/O operations to complete. For more information, see this IBM Documentation web page.

This issue might be caused by faster processors running with relatively poor disk subsystems performance. Disk subsystem performance can be affected by busy or slow disk, small I/O cache.

Disk protection

For more information about external storage consideration to set up your RAID protection, see Chapter 3, “Planning, configuring, and managing storage backend” on page 69.

> **Note:** If you need high I/O performance on your IBM i workload, you can create a DRAID 1 on your supported storage system, such as IBM Flash System 7200 and 9200 with IBM Spectrum Virtualize 8.4. In this configuration, the rebuild area is distributed over all member drives. The minimum extent size for this type of DRAID is 1024 MB.

Logical database I/O versus physical disk I/O

Information that is in partition buffers memory is available for use by any job or thread. Commonly, information is available in the partition buffer as a block of data rather than individual records. Data in a job buffer is available for use by the job only.

When an application program requests data, storage management checks whether they are available in memory. If so, it is moved to the open data path in the job buffer. If the data is not in memory, the request is submitted to the disk subsystem as a read command.

In that context, *logical Database I/O* information is moved between the open data path of user program and the partition buffer. This information is a count of the number of buffer movements and not a reflection of the records that were processed.

For more information, see the following IBM Documentation web pages:

- Sharing an Open Data Path
- Searching for a Perspective: Metric Finder

Physical disk I/O occurs when information is read or written as a block of data to or from the disk. It involves the movement of data between the disk and the partition buffer in memory. For more information, see this IBM Documentation web page.

Planning for IBM i storage capacity

To correctly plan the storage capacity that is provided by IBM Spectrum Virtualize family systems for IBM i, you must be aware of IBM i block translation for external storage that is formatted in 512-byte blocks. IBM i internal disks use a block size of 520 or 4160 bytes.

Because IBM Spectrum Virtualize storage for hosts is formatted with 512-byte blocks, a translation or mapping is required to attach it to IBM i. IBM i changes the data layout to support 512-byte blocks (sectors) in external storage by using an extra ninth sector to store the headers for every page.

The eight 8-byte headers from each 520-byte sector of a page are stored in the ninth sector, which is different than 520-byte sector storage where the 8 bytes are stored continuously with the 512 bytes of data to form the 520-byte sector.

The data that was stored in eight sectors is now stored by using nine sectors; therefore, the required disk capacity on IBM Spectrum Virtualize based systems is 9/8 of the IBM i usable capacity. Similarly, the usable capacity in IBM i is 8/9 of the allocated capacity in these storage systems.

When attaching IBM Spectrum Virtualize family storage to IBM i, plan for the extra capacity on the storage system so that the 8/9ths of the effective storage capacity that is available to IBM i covers the capacity requirements for the IBM i workload.

The performance effect of block translation in IBM i is small or negligible.

Figure A-7 shows byte sectors for IBM i.

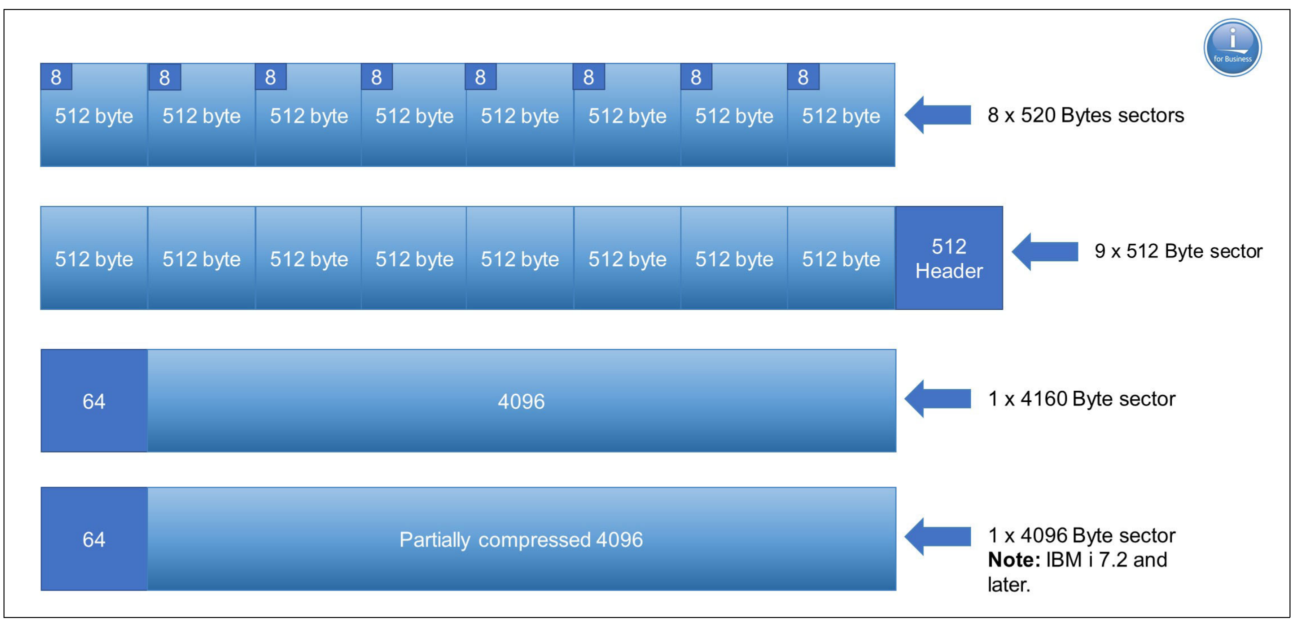


Figure A-7 IBM i with different sector sizes

Storage connection to IBM i

IBM Spectrum Virtualize storage can be attached to IBM i in the following ways:

- Native connection without the use of the IBM PowerVM Virtual I/O Server (VIOS)
- Connection with VIOS in:
  - NPIV mode
  - Virtual SCSI mode

The decision for IBM i native storage attachment or a VIOS attachment is based on the customer's requirements. Native attachment has its strength in terms of simplicity and can be a preferred option for static and smaller IBM i environments with only a few partitions. It does not require extra administration and configuration of a VIOS environment. However, it also provides the least flexibility and cannot be used with IBM PowerVM advanced functions, such as Live Partition Mobility or remote restart.

Table A-1 lists key criteria to help you with the decision for selecting an IBM i storage attachment method.

Table A-1 Comparing IBM i native and Virtual I/O Server attachment

Criteria	Native attachment	VIOS attachment
Simplicity (Configuration, maintenance, failure analysis)	ü	More complex
Performance	ü	ü (with NPIV)
Consolidation (Storage or network adapters)	More limited	ü
PowerVM advanced functions (Partition mobility, suspend, or resume, remote restart, private cloud deployment)	Not available	ü
Hardware support (storage or network adapters, entry level servers)	More limited	ü

Next, we discuss the guidelines and preferred practices for each type of connection.

Note: For more information about the current requirements, see the following web pages:

- IBM System Storage Interoperation Center (SSIC)
- IBM i POWER External Storage Support Matrix Summary

Native attachment

Native connection support for IBM i with IBM Spectrum Virtualize storage is available with IBM Power Systems POWER7 or later server technology. It requires IBM i 7.1, Technology Refresh (TR) 7 or later for POWER7, and IBM i 7.1 TR 8 or later for POWER8.

Native connection *with* SAN switches can be done by using the following adapters:

- 32 Gb PCIe3 2-port FC adapters feature number #EN1A or #EN1B (IBM POWER9™ only)
- 16 Gb PCIe3 4-port FC adapters feature number #EN1C or #EN1D (POWER9 only)
- 16 Gb PCIe3 2-port FC adapters feature number #EN0A or #EN0B
- 8 Gb PCIe 2-port FC adapters feature number #5735 or #5273
- 4 Gb PCIe 2-port Fibre Channel (FC) adapters feature number #5774 or #5276

Direct native connection *without* SAN switches can be done by using the following adapters:

- 16 Gb adapters in IBM i connected to 16 Gb adapters in IBM Spectrum Virtualize V7.5 or later based storage with non-NPIV target ports
- 4 Gb FC adapters in IBM i connected to 8 Gb adapters in IBM Spectrum Virtualize based storage with non-NPIV target ports

For resiliency and performance reasons, connect IBM Spectrum Virtualize storage to IBM i with multipath by using two or more FC adapters. Consider the following points:

- You can define a maximum of 127 LUNs (up to 127 active + 127 passive paths) to a 16 or 32 Gb port in IBM i, with IBM i 7.2 Technology Refresh (TR) 7 or later, and with IBM i 7.3 TR3 or later.
- You can define a maximum of 64 LUNs (up to 64 active + 64 passive paths) to a 16 or 32 Gb port with IBM i release and TR lower than i 7.2 TR7 and i 7.3 TR3.
- You can define a maximum of 64 LUNs (up to 64 active + 64 passive paths) to a 4 or 8 Gb port, regardless of the IBM i level.

The LUNs report in IBM i as disk units with type 2145.

IBM i enables SCSI command tag queuing in the LUNs from natively connected IBM Spectrum Virtualize storage. The IBM i queue depth per LUN and path with this type of connection is 16.

VIOS attachment

The following FC adapters are supported for VIOS attachment of IBM i to IBM Spectrum Virtualize storage:

- 32 Gb PCIe3 2-port FC adapter feature number #EN1A or #EN1B (POWER9 only)
- 16 Gb PCIe3 4-port FC adapter feature number #EN1C or #EN1D (POWER9 only)
- 16 Gb PCIe3 2-port FC adapter feature number #EN0A or #EN0B
- 8 Gb PCIe 2-port FC adapter feature number #5735 or #5273
- 8 Gb PCIe2 2-port FC adapter feature number #EN0G or #EN0F
- 8 Gb PCIe2 4-port FC adapter feature number #5729
- 8 Gb PCIe2 4-port FC adapter feature number #EN12 or #EN0Y

Important: For more information about the current requirements, see the following web pages:

- IBM System Storage Interoperation Center (SSIC)
- IBM i POWER External Storage Support Matrix Summary

Connection with VIOS NPIV

IBM i storage attachment support that uses IBM PowerVM® Virtual I/O Server N_Port ID Virtualization (NPIV) was introduced with POWER6 server technology. With NPIV, volumes (LUNs) from the IBM Spectrum Virtualize storage system are directly mapped to the IBM i server. VIOS does not see NPIV connected LUNs; instead, it is an FC pass-through.

The storage LUNs are presented to IBM i with their native device type of 2145 for IBM Spectrum Virtualize-based storage. NPIV attachment requires 8 Gb or newer generation FC adapter technology and SAN switches that must be NPIV enabled, as shown in Figure A-8.

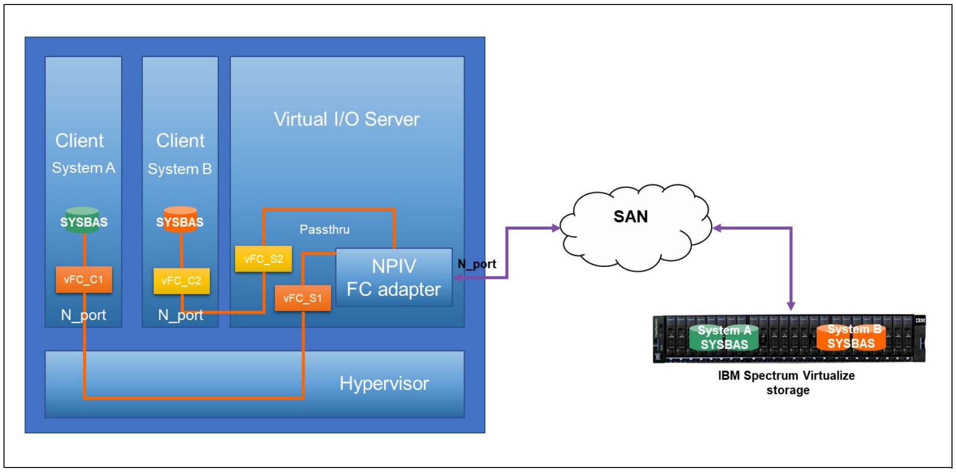


Figure A-8 IBM i SAN access using NPIV

Redundant VIOS with NPIV

For resiliency and performance reasons, connect IBM Spectrum Virtualize storage to IBM i by using multipathing across two or more VIOS servers.

Observe the following rules for mapping IBM i server virtual FC client adapters to the physical FC ports in VIOS when implementing NPIV connection:

- You can map up to 64 virtual FC adapters to the same physical FC adapter port in VIOS. With VIOS 3.1 and later, this limit was increased for support of mapping of up to 255 virtual FC adapters to a 32 Gb physical FC adapter port.
- Mapping of more than one NPIV client virtual FC adapter from the *same* IBM i system to a VIOS physical FC adapter port is supported since IBM i 7.2 TR7 and i 7.3 TR3; however, when PowerVM partition mobility is used, only a single virtual FC adapter can be mapped from the *same* IBM i system to a VIOS physical FC adapter port.
- You can use the same port in VIOS for NPIV mapping and connecting with VIOS virtual SCSI (VSCSI).
- If PowerHA solutions with IBM i independent auxiliary storage pools (IASPs) are implemented, you must use different virtual FC adapters for attaching the IASP LUNs, and not share adapters between SYSBAS and IASP LUNs.

You can configure a maximum of 127 LUNs (up to 127 active + 127 passive paths) to a virtual FC adapter with IBM i 7.2 TR7 or later, and with IBM i 7.3 TR3 or later.

You can configure a maximum of 64 LUNs (up to 64 active + 64 passive paths) to a virtual FC adapter with IBM i release and TR lower than i 7.2 TR7 and i 7.3 TR3.

IBM i enables SCSI command tag queuing for LUNs from VIOS NPIV connected to IBM Spectrum Virtualize storage. The IBM i queue depth per LUN and path with this type of connection is 16.

Note: If you encounter issues with NPIV/Virtual FC of IBM i that is attached to an IBM Spectrum Virtualize, such as missing paths and missing disk units, consider the following suggestions:

- Use System Snapshot (SYSSNAP) and be sure to include LIC LOGs, QHST, and PALs. Change the date range to include the date range of the problem. For more information, see this IBM Support web page.
- VIOS SNAPs can be collected from the VIOS partitions as part of the SYSSNAP or separately. For more information, see this IBM Support web page.
- Collect switch logs as close as possible to the time of the problem.
- Collect the applicable statesave, snap, and so on, from the IBM Spectrum Virtualize at the time the problem is occurring. This information is needed by the IBM Storage Support team.

If you experience a performance problem with poor disk response time and the IBM Spectrum Virtualize is connected with NPIV, see this IBM Support web page.

NPIV acceleration

Virtual I/O Server version 3.1.2 or later strengthened FC N_Port ID Virtualization (NPIV) to provide multi-queue support. This enhanced performance, including more throughput, reduced latency, and higher IOPS, spreads the I/O workload across multiple work queues.

The following FC adapter feature codes are supported:

- 32 Gb PCIe3 2-port FC adapters feature number #EN1A or #EN1B (POWER9 only)
- 16 Gb PCIe3 4-port FC adapters feature number #EN1C or #EN1D (POWER9 only)
- 16 Gb PCIe3 2-port FC adapters feature number #EN2A or #EN2B

Note: NPIV acceleration is supported by IBM i 7.2 or later, and by the firmware level for IBM Power Systems 9 is FW940 or later.

Connection with VIOS virtual SCSI

IBM i storage attachment by way of the IBM PowerVM Virtual I/O Server Connection that uses virtual SCSI was introduced with IBM Power Systems POWER6 technology.

When deciding on an IBM PowerVM Virtual I/O Server storage attachment for IBM i, NPIV attachment is often preferred over virtual SCSI attachment for the following reasons:

- With virtual SCSI, an emulation of generic SCSI devices is performed by VIOS for its client partitions, such as IBM i, which requires extra processing and adds a small delay to I/O response times.
- Virtual SCSI provides much lower scalability in terms of maximum supported LUNs per virtual adapter than NPIV, and requires more storage management, such as multipath configuration and customization at the VIOS layer, which adds complexity.
- Because of the virtual SCSI emulation unique device characteristics of the storage device, such as device type or in case of tape devices media type and other device attributes are no longer presented to the IBM i client.

Virtual SCSI attachment is not supported for PowerHA LUN level switching technology, which is required for IASP HyperSwap solutions with IBM Spectrum Virtualize. Similar considerations for NPIV apply with regard to the use of IBM i multipathing across two or more VIOS to improve resiliency and performance. However, because multipathing is implemented at the VIOS layer with virtual SCSI, the following considerations apply:

- IBM i multipathing is performed with two or more VSCSI client adapters, each of them assigned to a VSCSI server adapter in different VIOS. With virtual SCSI, volumes (LUNs) from the IBM Spectrum Virtualize storage system are not mapped directly to an IBM i host but instead to the two or more VIOS servers. These LUNs are detected as hdisks on each VIOS and must be mapped as a virtual target device to the relevant VSCSI server adapters to be used by the IBM i client.
- In addition to IBM i multipathing across multiple VIOS servers, with virtual SCSI, multipathing must be implemented at the VIOS server layer to provide further I/O parallelism and resiliency by using multiple physical FC adapters and SAN fabric paths from each VIOS server to its storage.

  The IBM recommended multipath driver for IBM Spectrum Virtualize based storage running microcode V7.6.1 or later is the VIOS built-in AIXPCM multipath driver, which replaces the previously recommended SDDPCM multipath driver.

  For more information, see this IBM Support web page.

IBM i storage attachment by using the IBM PowerVM Virtual I/O Server Connection that uses virtual SCSI was introduced with IBM Power Systems POWER6 technology.

When deciding on an IBM PowerVM Virtual I/O Server storage attachment for IBM i, NPIV attachment is often preferred over virtual SCSI attachment for the following reasons:

- With virtual SCSI, an emulation of generic SCSI devices is performed by VIOS for its client partitions, such as IBM i, which requires extra processing and adds a small delay to I/O response times.
- Virtual SCSI provides much lower scalability in terms of the maximum supported LUNs per virtual adapter than NPIV. It also requires more storage management, such as multipath configuration and customization at the VIOS layer, which adds complexity.
- Because of the virtual SCSI emulation unique device characteristics of the storage device, such as device type (or in the case of tape devices), media type and other device attributes are no longer presented to the IBM i client.
- Virtual SCSI attachment is not supported for PowerHA LUN level switching technology, which is required for IASP HyperSwap solutions with IBM Spectrum Virtualize.

Similar considerations for NPIV apply with regard to the use of IBM i multipathing across two or more VIOS to improve resiliency and performance. However, because multipathing is also implemented at the VIOS layer with virtual SCSI, the following considerations apply:

- IBM i multipathing is performed with two or more VSCSI client adapters, each is assigned to a VSCSI server adapter in different VIOS. With virtual SCSI, volumes (LUNs) from the IBM Spectrum Virtualize storage system are not mapped directly to an IBM i host but to the two or more VIOS servers. These LUNs that are detected as HDDs on each VIOS must be mapped as a virtual target device to the relevant VSCSI server adapters to be used by the IBM i client.
- In addition to IBM i multipathing across multiple VIOS servers, with virtual SCSI, multipathing is implemented at the VIOS server layer to provide further I/O parallelism and resiliency by using multiple physical FC adapters and SAN fabric paths from each VIOS server to its storage.

- The IBM recommended multipath driver for IBM Spectrum Virtualize-based storage running microcode V7.6.1 or later is the VIOS built-in AIXPCM multipath driver, which replaces the previously recommended SDDPCM multipath driver.
- For more information, see this IBM Support web page.

Up to 4095 LUNs can be connected per target, and up to 510 targets per port in a physical adapter in VIOS. With IBM i 7.2 and later, you can attach a maximum of 32 disk LUNs to a virtual SCSI adapter in IBM i. With IBM i releases before i 7.2, a maximum of 16 disk LUNs can be attached to a virtual SCSI adapter in IBM i. The LUNs are reported in IBM i as generic SCSI disk units of type 6B22.

IBM i enables SCSI command tag queuing in the LUNs from VIOS VSCSI connected to IBM Spectrum Virtualize storage. The queue depth on a LUN with this type of connection is 32.

Setting attributes in VIOS

This section describes the values of specific device attributes in VIOS that must be configured for resiliency and performance.

FC adapter attributes

With VIOS virtual SCSI connection or NPIV connection, use the VIOS **`chdev`** command to specify the following attributes for each SCSI I/O Controller Protocol Device (fscsi) device that connects an IBM Spectrum Virtualize storage LUN for IBM i:

- The attribute `fc_err_recov` must be set to `fast_fail`
- The attribute `dyntrk` must be set to `yes`

The specified values for the two attributes specify how the VIOS FC adapter driver or VIOS disk driver handle specific types of fabric-related failures and dynamic configuration changes. Without setting these values for the two attributes, the way these events are handled is different and causes unnecessary retries or manual actions.

Note: These attributes also are set to the recommended values when applying the *default rules set* that is available with VIOS 2.2.4.x or later.

Disk device attributes

With VIOS virtual SCSI connection, use the VIOS **`chdev`** command to specify the following attributes for each hdisk device that represents an IBM Spectrum Virtualize storage LUN connected to IBM i:

- If IBM i multipathing across two or more VIOS servers is used, the attribute `reserve_policy` must set to `no_reserve`.
- The attribute `queue_depth` must be set to `32`.
- The attribute `algorithm` must be set to `shortest_queue`.

Setting `reserve_policy` to `no_reserve` is required to be set in each VIOS if multipath with two or more VIOS is implemented to prevent SCSI reservations on the hdisk device.

Set `queue_depth` to `32` for performance reasons. Setting this value ensures that the maximum number of I/O requests that can be outstanding on an HDD in the VIOS at a time matches the maximum number of 32 I/O operations that IBM i operating system allows at a time to one VIOS VSCSI-connected LUN.

Set `algorithm` to `shortest_queue` for performance reasons. Setting this value allows the AIXPCM driver in VIOS to use a dynamic load balancing instead of the default path failover algorithm for distributing the I/O across the available paths to IBM Spectrum Virtualize storage.

Setting a physical volume identifier (PVID) for HHD devices that are used for virtual SCSI attachment of IBM i client partitions is not recommended because it makes those devices ineligible for a possible later migration to NPIV or native attachment.

Important: While working with SCSI and NPIV, you cannot mix both regarding the paths to the same LUN. However, VIOS supports NPIV and SCSI concomitantly; that is, some LUNs can be attached to the virtual WWPNs of the NPIV FC adapter. At the same time, the VIOS also can provide access to LUNs that are mapped to virtual target devices and exported as VSCSI devices.

You can have one or more Virtual I/O Servers providing the pass-through function for NPIV. Also, you can have one or more Virtual I/O Servers hosting VSCSI storage. Therefore, the physical HBA in the Virtual I/O Server supports NPIV and VSCSI traffic.

Guidelines for Virtual I/O Server resources

Be aware of the memory requirements of the hypervisor when determining the overall memory of the system. Above and beyond the wanted memory for each partition, you must add memory for virtual resources (VSCSI and Virtual FC) and hardware page tables to support the maximum memory value for each partition.

The suggestion is to use the IBM Workload Estimator tool to estimate the needed Virtual I/O Server resources. However, as a starting point in context of CPU and memory for Virtual I/O Server. For more information, see this IBM Support web page.

Disk drives for IBM i

This section describes how to implement internal disk drives in IBM Spectrum Virtualize storage or externally virtualized back-end storage for an IBM i host. These suggestions are based on the characteristics of a typical IBM i workload, such as a relatively high write ratio, a relatively high access density, and a small degree of I/O skew because of the spreading of data by IBM i storage management.

Considering these characteristics and typical IBM i customer expectations for low I/O response times, we expect that many SAN storage configurations for IBM i are based on an all-flash storage configuration.

If a multitier storage configuration that uses enterprise class (tier0_flash) and high-capacity (tier1_flash) flash drives or even enterprise hard disk drives (tier2_HDD) is preferred for less demanding workloads or for commercial reasons, ensure that a sufficiently large part of disk capacity is on flash drives. As a best practice for a multitier configuration considering the typically low IBM i I/O skew, at least 20% of IBM i capacity should be based on the higher tier flash storage technology.

Even if specific parts of IBM i capacity are on flash drives, it is important that you provide enough HDDs with high rotation speed for a hybrid configuration with flash drives and HDDs. Preferably, use 15 K RPM HDDs of 300 GB or 600 GB capacity, along with flash technology.

IBM i transaction workload usually achieves the best performance when disk capacity is used entirely from enterprise class flash (tier0_flash) storage. High capacity or read-intensive flash drives are typically not the best choice for IBM i performance critical workload. This issue is especially for the top storage tier, considering a usually high IBM i write percentage of often 50% and higher, the disk write amplification by using RAID 6, and the significant lower random write performance of tier1 compared to tier0 flash drives.

The use of a multitier storage configuration by IBM Spectrum Virtualize storage is achieved by using Easy Tier. For more information, see *Implementing the IBM SAN Volume Controller with IBM Spectrum Virtualize Version 8.4.2*, SG24-8507.

Even if you do not plan to install multitier storage configuration, or have no multitier storage configuration that is installed, you can still use Easy Tier for intra-tier rebalancing. You also can evaluate your workload with its I/O skew, which provides information about the benefit you might gain by adding flash technology in the future.

Compression considerations

If compression is wanted, the preferred choice for using compression at the IBM Spectrum Virtualize storage system layer for performance critical IBM i workload is by using IBM FlashCore module (FCM) hardware compression technology at the disk drive level within IBM Spectrum Virtualize standard pools or data reduction pools (DRPs) with fully allocated volumes. These configuration options do not affect performance compared to other compression technologies, such as DRP compressed volumes or Real-Time Compression at the storage subsystem level.

Important: Data reduction or deduplication can be used with IBM i, which affects performance positively.

Nevertheless, the performance is affected and different whenever something is touched, such as 30 minutes taking 3 - 18 hours. The data is affected whenever something is created, changed, or used. The integrity of the objects is maintained.

However, if a physical page on disk is corrupted, potentially hundreds or thousands of objects become corrupted instead of only one. Another consideration is the amount of wear that occurs on the drives from so much read/write activity.

If you plan to use deduplication for archival or test purposes, deduplication might be a viable solution for saving huge amounts of storage. If the deduplication solution is planned for a production or development environment, we strongly recommend that you test thoroughly before committing.

Storage sizing and performance modeling

IBM provides tools, such as IBM Storage Modeller (StorM) and IntelliMagic Disk Magic for IBM representatives and Business Partners, which are recommended to be used for performance modeling and sizing before implementing a wanted IBM Spectrum Virtualize storage configuration for IBM i. These tools allow the user to enter the performance data of the current IBM i workload manually or by using file import from IBM i (5770-PT1 reports or PDI data) or from IBM Spectrum Control performance data. Enter the current storage configuration and model the wanted configuration.

When modeling Easy Tier, specify the lowest skew level for IBM i workload or import an existing I/O skew curve from available Easy Tier reports. The steps that are taken for sizing and modeling IBM i are shown in Figure A-9.

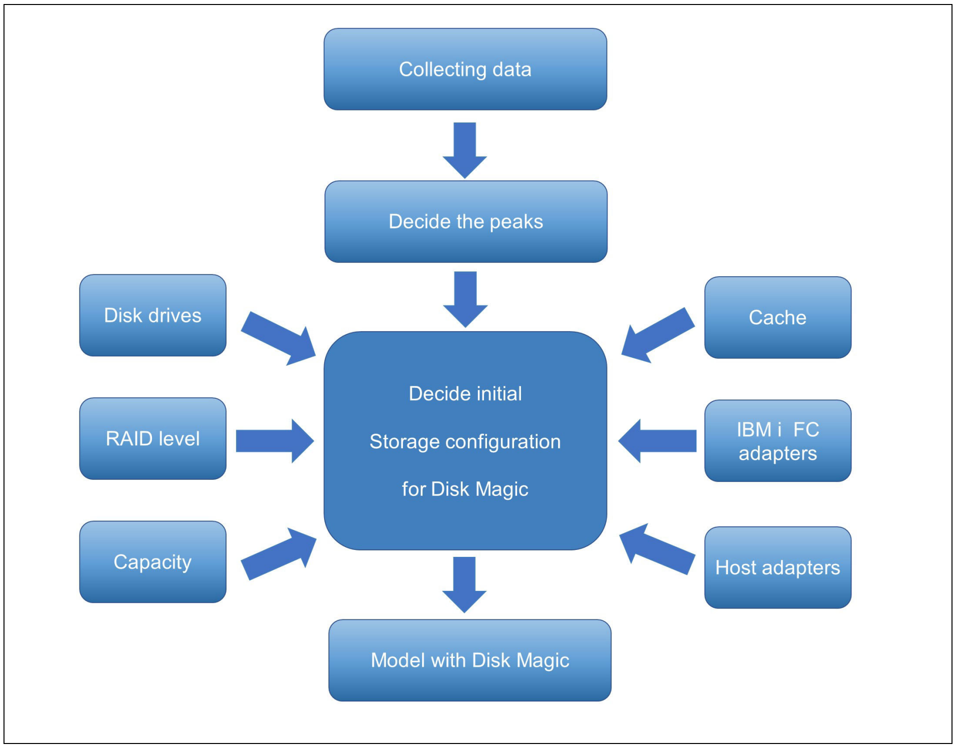


Figure A-9 Diagram of sizing and modeling for IBM i using Disk Magic

The modeling helps assure an adequate solution sizing by providing predictions for the modeled IBM Spectrum Virtualize storage resource of system usage, predicted disk response time for IBM i, and usage and response times at workload growth.

Note: Contact your IBM representative or IBM Business Partner to discuss a performance modeling and sizing for a planned IBM Spectrum Virtualize storage solution for IBM i.

IBM i Unmap support

To better use IBM Spectrum Virtualize storage flash technology with an efficient storage space allocation and deallocation, IBM i supports the space of storage system unmap capabilities by corresponding host unmap functions.

Initially, IBM i unmap support that is implemented by way of the SCSI Write Same command was introduced with i 7.2 TR8 and i 7.3 TR4 for LUN initialization only; that is, for the add disk units to ASP function.

With i 7.3 TR9 and i 7.3 TR5, runtime support was added, which also supports synchronous unmap for scenarios, such as object deletion and journal clearance. The runtime unmap algorithm was further enhanced supported by i 7.3 TR7 and i 7.4 TR1, which implements an asynchronous periodic free-space cleaning.

IBM Spectrum Virtualize V8.1.1 and later storage systems can use the **`unmap`** function for efficiently deallocate space, such as for volume deletion, on their back-end storage by sending SCSI **`unmap`** commands to specific supported internal SSDs and FCMs, and selected virtualized external flash storage.

Space reclamation that is triggered by host **unmap** commands is supported by IBM Spectrum Virtualize V8.1.2 and later for DRP thin provisioned volumes, which can increase the free capacity in the storage pool so that it becomes available also for use by other volumes in the pool.

For more information about IBM Spectrum Virtualize storage SCSI unmap support, see 4.1.2, "Data reduction pools" on page 117, and this IBM Support web page.

Defining LUNs for IBM i

LUNs for an IBM i host are defined from IBM Spectrum Virtualize block-based storage. They are created from available extents within a storage pool, the same way as for open system hosts.

Even though IBM i supports a usable LUN size of up to 2 TB (1 byte for IBM Spectrum Virtualize storage), the use of only a few large-size LUNs for IBM i is *not* recommended for performance reasons.

In general, the more LUNs that are available to IBM i, the better the performance for the following the reasons:

- If more LUNs are attached to IBM i, storage management uses more threads and therefore enables better performance.
- More LUNs provide a higher I/O concurrency, which reduces the likelihood of I/O queuing and therefore the wait time component of the disk response time, which results in lower latency of disk I/O operations.

For planning, consider that a higher number of LUNs might also require more physical or virtual FC adapters on IBM i based on the maximum number of LUNs that is supported by IBM i per FC adapter port.

The sizing process helps to determine a reasonable number of LUNs that are required to access the needed capacity, while meeting performance objectives. Regarding both of these aspects and the preferred practices, consider the following guidelines:

- For any IBM i disk pool (ASP), define all the LUNs as the same size.
- 40 GB is the preferred minimum LUN size.
- You should not define LUNs larger than approximately 200 GB.

  > **Note:** This rule is not a fixed rule because it is important that enough LUNs are configured, with which this guideline helps. Selecting a larger LUN size should not lead to configurations, such as storage migrations, with a significantly fewer number of LUNs being configured with possibly detrimental effects on performance.

- A minimum of 8 LUNs for each ASP is preferred for small IBM i partitions and typically a couple of dozen LUNs for medium and up to a few hundreds for large systems.

When defining LUNs for IBM i, consider the following required minimum capacities for the load source (boot disk) LUN:

- With IBM i release 7.1, the minimum capacity is 20 GB
- With IBM i release 7.2 before TR1, the minimum capacity is 80 GB in IBM i
- With IBM i release 7.2 TR1 and later, the minimum capacity is 40 GB in IBM i

IBM Spectrum Virtualize dynamic volume expansion is supported for IBM i with IBM i 7.3 TR4 and later. An IBM i IPL is required to use the extra volume capacity.

Tip: For more information about cross-referencing IBM i disks units with IBM Spectrum Virtualize LUNs by using N-Port ID Virtualization (NPIV), see this IBM Support web page.

Disk arms and maximum LUN size

Selected limits that are related to disk arms and LUNs sizes were increased in IBM i 7.4, as listed in Table A-2.

Table A-2 Limits increased for Max Disk Arms and LUN size

System Limits	IBM i 7.2	IBM i 7.3	IBM i 7.4
Disk arms in all basic auxiliary storage pools (ASPs 1 - 32), per LPAR	2047	2047	3999
Disk arms in all independent auxiliary storage pools (IASPs 33 - 255) in all nodes in a cluster	2047	2047	5999
Maximum combined number of disk arms and redundant connections to disk units	35.600	35.600	35.600
512 byte block size LUNs [a]	2 TB	2 TB	2 TB
4096 byte block size LUNs [b]	2 TB	2 TB	16 TB

a. Actual limit is one block short of the maximum that is shown in Table A-2. For all 512-block LUNs, the maximum is still up to 2 TB, including IBM Storwize LUNs and SAN Volume Controller LUNs.
b. Includes IBM FlashSystems LUNs, and 4 K block SAS disks (VSCSI attached).

Note: For more information about these limits, and others, see IBM Documentation web page.

Data layout

Spreading workloads across all IBM Spectrum Virtualize storage components maximizes the use of the hardware resources in the storage subsystem. I/O activity must be balanced between the two nodes or controllers of the IBM Spectrum Virtualize storage system I/O group, which is usually taken care of by the alternating preferred node volume assignments at LUN creation.

However, especially with improper sizing or unanticipated workload increases, performance problems might arise when sharing resources because of resource contention. Some isolation of workloads, at least regarding a shared back-end storage, can be accomplished by a configuration in which each IBM i ASP or LPAR has its own managed storage pool.

Such a configuration with dedicated storage pools results in a trade-off between accomplishing savings from storage consolidation and isolating workloads for performance protection. This is because a dedicated storage pool configuration likely requires more back-end storage hardware resources because it cannot use the averaging effect of multiple workloads that typically show their peaks at different intervals.

Consider the following data layout:

- For all-flash storage configurations (assuming a correctly sized storage backend), no reason often exists for not sharing the disk pool among multiple IBM i workloads.
- For hybrid configurations with Easy Tier on mixed HDD and flash disks, the storage pool can also be shared among IBM i workloads. Only very large performance critical workloads must be configured in isolated disk pools.
- For HDD-only pools, ensure that you isolate performance critical IBM i workloads in separate storage pools.
- Avoid mixing IBM i LUNs and non-IBM i LUNs in the same disk pool.

Apart from the use of Easy Tier on IBM Spectrum Virtualize for managing a multitier storage pool, an option also exists to create a separate storage pool for different storage tiers on IBM Spectrum Virtualize storage and create different IBM i ASPs for each tier. IBM i applications that have their data in an ASP of a higher storage tier experience a performance boost compared to those applications that use an ASP with a lower storage tier.

IBM i internal data relocation methods, such as the ASP balancer hierarchical storage management function and IBM DB2 media preference, are not available to use with IBM Spectrum Virtualize flash storage.

Fibre Channel adapters in IBM i and VIOS

When you size the number of FC adapters for an IBM i workload for native or VIOS attachment, consider the maximum I/O rate (IOPS) and data rate (MBps) that a port in a particular adapter can sustain at 70% utilization. Also, consider the I/O rate and data rate of the IBM i workload.

If multiple IBM i partitions connect through the same FC port in VIOS, consider the maximum rate of the port at 70% utilization and the sum of I/O rates and data rates of all connected LPARs.

For sizing, you might consider the throughput that is listed in Table A-3 that shows the throughput of a port in a specific adapter at 70% utilization.

Table A-3 Throughput of Fibre Channel adapters

Maximal I/O rate per port	16 Gb 2-port adapter	8 Gb 2-port adapter
IOPS per port	52,500 IOPS	23,100 IOPS
Sequential throughput per port	1,330 MBps	770 MBps
Transaction throughput per port	840 MBps	371 MBps

Make sure to plan for the use of separate FC adapters for IBM i disk and tape attachment. This separation is recommended because of the required IBM i virtual I/O processor (IOP) reset for tape configuration changes and for workload performance isolation.

Zoning SAN switches

With IBM i native attachment or VIOS NPIV attachment, zone the SAN switches so that one IBM i FC initiator port is in a zone with two FC ports from the IBM Spectrum Virtualize storage target, each port from one node canister of the I/O group, as shown in Figure A-10. This configuration provides resiliency for the I/O to and from a LUN that is assigned to the IBM i FC initiator port. If the preferred node for that LUN fails, the I/O continues using the nonpreferred node.

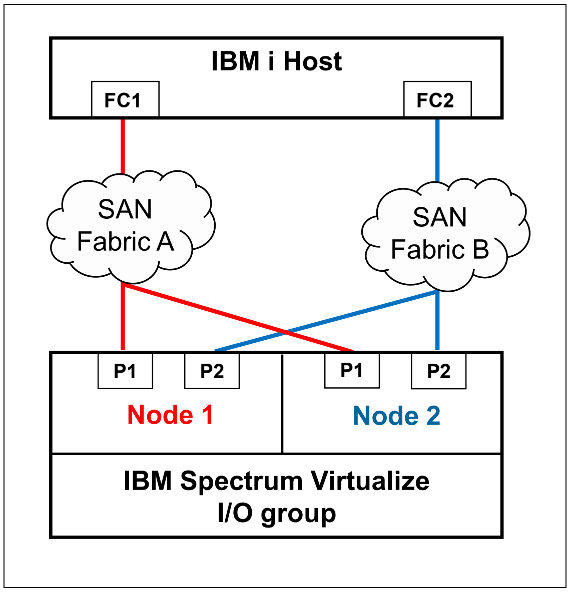


Figure A-10 SAN switch zoning for IBM i with IBM Spectrum Virtualize storage

For VIOS virtual SCSI attachment, zone one physical port in VIOS with one or more available FC ports from each of both node canisters of the IBM Spectrum Virtualize storage I/O group. IBM SAN Volume Controller or IBM Storwize ports that are zoned with one VIOS port must be evenly spread between both node canisters. A maximum of eight host paths is supported from VIOS to IBM Spectrum Virtualize storage.

IBM i Multipath

Multipath provides greater resiliency for SAN-attached storage. IBM i supports up to eight active paths and up to eight passive 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.

Typically, two active paths to a LUN are a good balance of price and performance. The scenario that is shown in Figure A-10 results in two active and two passive paths to each LUN for IBM i. However, you can implement more than two active paths for workloads where high I/O rates are expected to the LUNs where a high I/O access density is expected.

It is important to understand that IBM i multipath for a LUN is achieved by connecting the LUN to two or more FC ports that belong to different adapters in an IBM i partition. Adding more than one FC port from the same IBM Spectrum Virtualize storage node canister to a SAN switch zone with an IBM i FC initiator port does not provide more active paths because an IBM i FC initiator port, by design, logs in to one target port of a node only.

With IBM i native attachment, the ports for multipath must be from different physical FC adapters in IBM i. With VIOS NPIV, the virtual Fibre Channel adapters for multipath must be assigned to different VIOS for redundancy. However, if more than two active paths are used, you can use two VIOS and split the paths among them. With VIOS virtual SCSI attachment, the virtual SCSI adapters for IBM i multipath must be assigned to different VIOS.

IBM Spectrum Virtualize storage uses a redundant dual active controller design that implements SCSI asymmetrical logical unit access (ALUA). That is, some of the paths to a LUN are presented to the host as optimized and others as nonoptimized.

With an ALUA aware host, such as IBM i, the I/O traffic to and from a specific LUN normally goes through only the optimized paths, which often are associated with a specific LUN of preferred node. The nonoptimized paths, which often are associated with the nonpreferred node, are not actively used.

In the case of an IBM Spectrum Virtualize storage topology, such as HyperSwap or IBM SAN Volume Controller Enhanced Stretched Cluster that implements host site awareness, the optimized paths are not necessarily associated with a preferred node of a LUN but with the node of the I/O group that includes the same site attribute as the host.

If the node with the optimized paths fails, the other node of the I/O group takes over the I/O processing. With IBM i multipath, all of the optimized paths to a LUN are reported as *active* on IBM i, while the nonoptimized paths are reported as *passive*. IBM i multipath employs its load balancing among the active paths to a LUN and starts the use of the passive paths if all active paths failed.

Boot from SAN

All IBM i storage attachment options that are native (VIOS NPIV, and VIOS virtual SCSI), support IBM i boot from SAN. The IBM i load source is on an IBM Spectrum Virtualize storage LUN that is connected in the same manner as the other LUNs.

Apart from the required minimum size, the load source LUN does include any special requirements. The FC or SCSI I/O adapter for the load source must be *tagged* (that is, specified) by the user in the IBM i partition profile on the IBM Power Systems Hardware Management Console (HMC). When installing the IBM SLIC with disk capacity on IBM Spectrum Virtualize storage, the installation prompts you to select one of the available LUNs for the load source.

IBM i mirroring

Some clients prefer to use IBM i mirroring functions for resiliency. For example, they use IBM i mirroring between two IBM Spectrum Virtualize storage systems, each connected with one VIOS.

When setting up IBM i mirroring with VIOS connected IBM Spectrum Virtualize storage, 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 the LUNs.
2. After mirroring is started for those LUNs, add the LUNs from another two new virtual adapters, each adapter connecting one to-be mirrored half, and so on. This way, you ensure that IBM i mirroring is started between the two IBM Spectrum Virtualize storage systems and not among the LUNs from the same storage system.

Copy services considerations

This section describes IBM Spectrum Virtualize Copy Services considerations for usage with IBM i.

Remote replication

The IBM Spectrum Virtualize family products support Metro Mirror synchronous remote replication and Global Mirror asynchronous remote replication.

Two options are available for Global Mirror: *Standard* Global Mirror, and Global Mirror with *change volumes*, which allows for a flexible and configurable recovery point objective (RPO) that allows data replication to be maintained during peak periods of bandwidth constraints, and data consistency at the remote site to be maintained and also during resynchronization.

Regarding the IBM Spectrum Virtualize Copy Services functions, the IBM i single-level storage architecture requires that the disk storage of an IBM i system needs to be treated as a single entity; that is, the scope of copying or replicating an IBM i disk space must include SYSBAS, which is referred to as *full system replication*, or an IASP (referred to *IASP replication*).

Full system replication is used for disaster recovery (DR) purposes where an IBM i standby server is used at the DR site, as shown in Figure A-11 on page 626. When a planned or unplanned outage occurs on the IBM i production server, the IBM i standby server can be started (IPLed) from the replicated SYSBAS volumes after they are switched from IBM Spectrum Virtualize to a primary role to become accessible for the IBM i standby host.

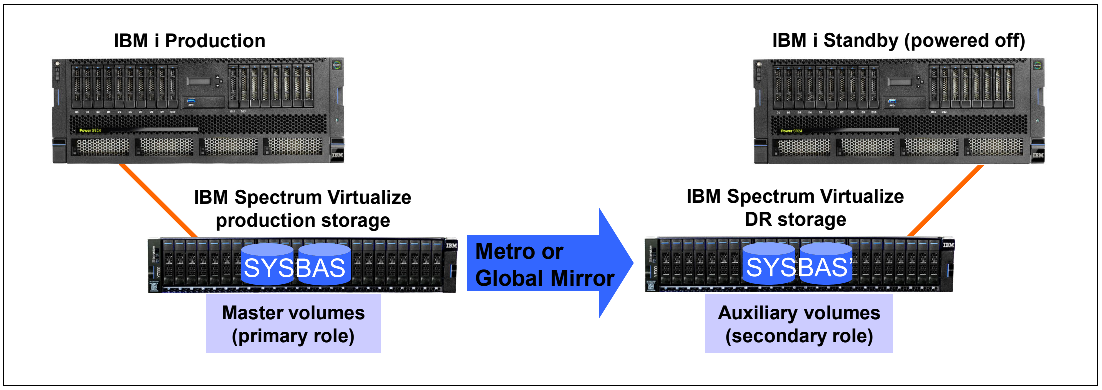


Figure A-11 IBM i full system replication with IBM Spectrum Virtualize

IASP-based replication for IBM i is used for a high availability (HA) solution in which an IBM i production and an IBM i backup node are configured in an IBM i cluster and the IASP that is replicated by IBM Spectrum Virtualize remote replication is switchable between the two cluster nodes, as shown in Figure A-12.

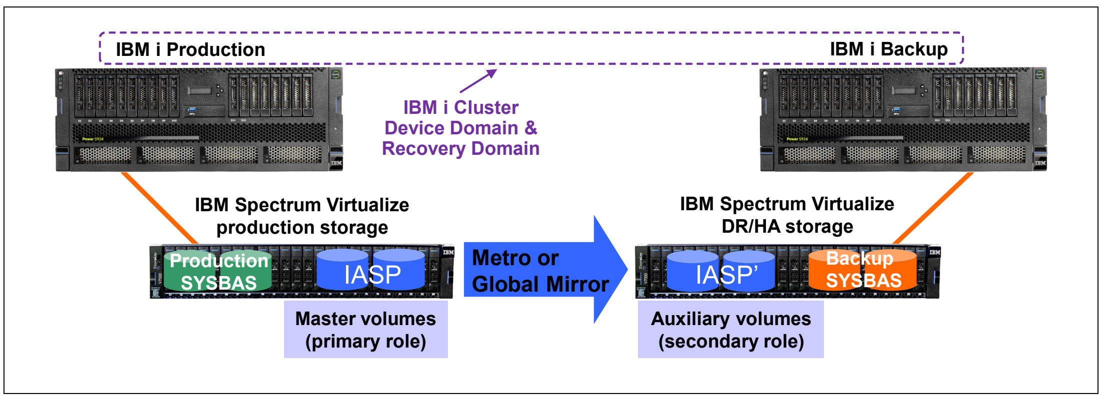


Figure A-12 IBM i IASP replication with IBM Spectrum Virtualize

In this scenario, the IBM i production system and the IBM i backup system each have their own non-replicated SYSBAS volumes and only the IASP volumes are replicated. This solution requires IBM PowerHA SystemMirror® for i Enterprise Edition (5770-HAS *BASE and option 1) for managing the IBM i cluster node switch and failovers and the IBM Spectrum Virtualize storage remote replication switching.

In this scenario, the IBM i production system and the IBM i backup system each have their own nonreplicated SYSBAS volumes and only the IASP volumes are replicated. This solution requires IBM PowerHA SystemMirror for i Enterprise Edition (5770-HAS *BASE and option 1) for managing the IBM i cluster node switch and fail overs and the IBM Spectrum Virtualize storage remote replication switching.

For more information about IBM i high availability solutions with IBM Spectrum Virtualize Copy Services see *PowerHA SystemMirror for IBM i Cookbook*, SG24-7994.

The sizing of the required replication link bandwidth for Metro Mirror or Global Mirror must be based on the peak write data rate of the IBM i workload to avoid affecting production performance. For more information, see “SAN Extension design considerations” on page 299.

For more information about IBM Spectrum Virtualize storage zoning guidelines, see 2.3.2, “Port naming and distribution” on page 49.

For environments that use remote replication, a minimum of two FC ports is suggested on each IBM Spectrum Virtualize storage node that is used for remote mirroring. The remaining ports on the node should not have any visibility to any other IBM Spectrum Virtualize cluster. Following these zoning guidelines helps to avoid configuration-related performance issues.

FlashCopy

When planning for FlashCopy with IBM i, make sure that enough disk drives are available to the FlashCopy target LUNs to maintain a good performance of the IBM i production workload while FlashCopy relationships are active. This guideline is valid for FlashCopy with and without background copy.

When FlashCopy is used with thinly provisioned target LUNs, ensure that sufficient capacity is available in the storage pool to be dynamically allocated when needed for the copy-on-write operations. The required thin target LUN capacity depends on the amount of write operations to the source and target LUNs, the locality of the writes, and the duration of the FlashCopy relationship.

FlashCopy temperature and considerations for IBM i

FlashCopy temperature indicates the amount of disruption to source system and quality of the FlashCopy target. FlashCopy copies what was sent to disk. Updates that are sitting in memory on the IBM i are not known to the storage system.

FlashCopy cold

The following considerations apply to FlashCopy cold:

- All memory is flushed to disk.
- Source IASP must be varied off before performing a FlashCopy.
- This method is the only method to ensure all writes are sent out to disk and included.

FlashCopy warm

The following considerations apply to FlashCopy warm:

- Memory is not flushed to disk.
- Writes in memory are excluded from the FlashCopy target.
- Zero disruption to IBM i source system.

FlashCopy quiesced

IBM i provides a quiesce function that can suspend database transactions and database and Integrated File System (IFS) file change operations for the system and configured basic auxiliary storage pools (ASPs) or independent ASPs (IASPs).

Consider the following points about FlashCopy quiesced:

- Some memory is flushed to disk.
- Attempts to flush writes to disk and suspends DB I/O and to reach commitment control boundaries.
- Minimal disruption to source, is the preferred practice, and better quality than warm.

HyperSwap

IBM Spectrum Virtualize storage HyperSwap as an active-active remote replication solution is supported for IBM i full system replication with IBM i 7.2 TR3 or later. It is supported for native and for VIOS NPIV attachment.

HyperSwap for IBM i IASP replication is supported by IBM i 7.2 TR5 or later and by IBM i 7.3 TR1 or later. With this solution, you must install IBM PowerHA SystemMirror for i Standard Edition (5770-HAS *BASE and option 2) that enables LUN level switching to site 2. It is supported for native and VIOS NPIV attachment.

IBM Spectrum Virtualize HyperSwap relies on the SCSI ALUA aware IBM i host multipath driver to manage the paths to the local and remote IBM Spectrum Virtualize storage systems that are logically configured as a single clustered system.

From a SAN switch zoning perspective, HyperSwap requires that the IBM i host is zoned with IBM Spectrum Virtualize nodes of the I/O group on each site. For a balanced configuration, the SAN switch from a dual fabric configuration is evenly used.

An example of the SAN fabric connections for IBM i HyperSwap with VIOS NPIV attachment is shown in Figure A-13. This configuration example results in four active paths and 12 passive paths that are presented on IBM i for each HyperSwap LUN.

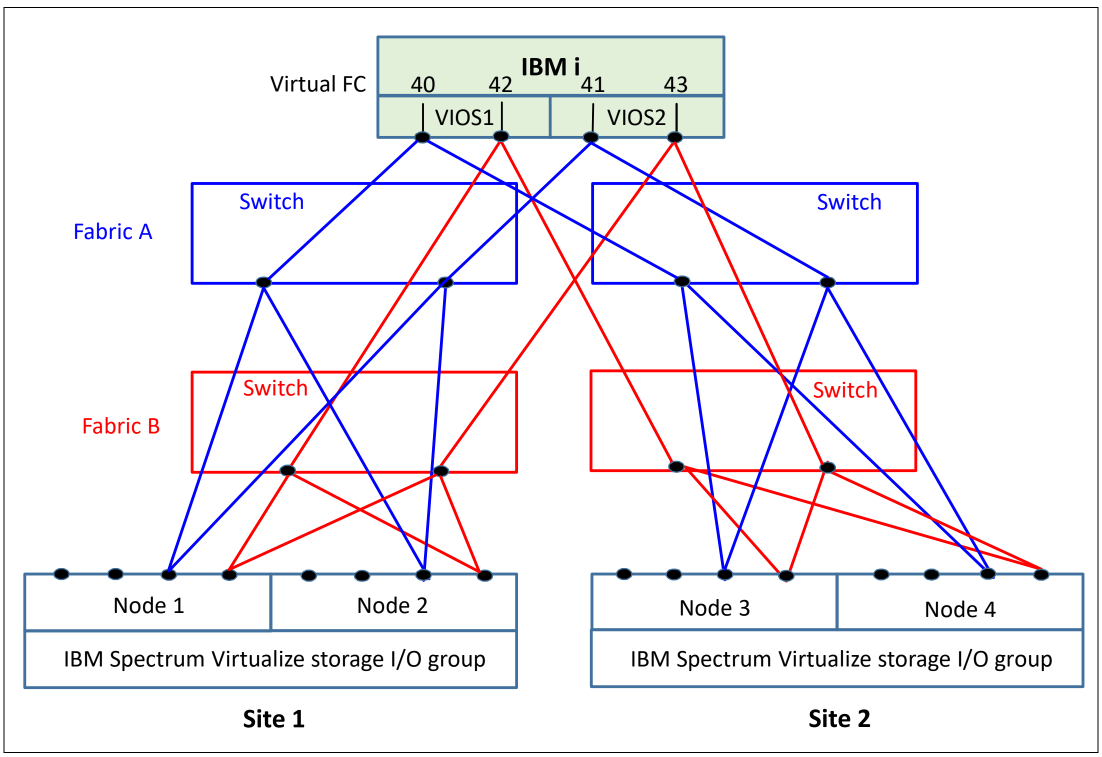


Figure A-13   IBM i HyperSwap SAN fabric connection example

Next, we briefly describe some high availability scenarios that use HyperSwap for IBM i.

Outage of Spectrum Virtualize I/O group at site 1

In this scenario, the entire IBM i storage capacity is on HyperSwap LUNs.

After the outage of I/O group at site 1 occurs, the I/O rate automatically transfers to the IBM Spectrum Virtualize nodes at site 2. The IBM i workload keeps running, and no relevant messages exist in IBM i message queues.

When the outage finishes, the IBM i I/O rate automatically transfers to nodes on site 1. The IBM i workload keeps running without interruption.

Disaster at site 1 with full system HyperSwap

In this scenario, we use a prepared IBM i standby system at site 2. The entire IBM i storage capacity is on HyperSwap LUNs. Two hosts are defined in the IBM Spectrum Virtualize storage cluster: one host with the WWPNs of IBM i at site 1, and one with WWPNs of site 2.

After a failure of site 1, including a failure of the IBM i production system and the storage at site 1, the IBM i LUNs are still available from the IBM Spectrum Virtualize nodes at site 2.

In the HyperSwap cluster, we manually unmap the HyperSwap LUNs from the IBM i production host at site 1, map the LUNs to the IBM i standby host at site 2, and IPL the IBM i standby host at site 2. After the IPL completes, we can resume the workload on site 2.

After the outage of site 1 is finished, power-down IBM i at site 2, unmap the IBM i LUNs from the host at site 2 and then, map them to the host at site 1. IPL IBM i at site 1 and resume the workload. The I/O rate is transferred to the IBM Spectrum Virtualize storage nodes at site 1.

Disaster at site 1 with IASP HyperSwap

This scenario requires IBM PowerHA SystemMirror for i software to be installed. It also needs the corresponding IBM i setup that consists of two IBM i partitions in a cluster and a switchable IASP on IBM i at site 1, a PowerHA cluster resource group, and PowerHA copy description. The workload is running in the IASP.

For more information about PowerHA for i setup, see *IBM PowerHA SystemMirror for i: Preparation (Volume 1 of 4)*, SG24-8400.

In this scenario, ensure that all IBM i LUNs, not just the IASP LUNs, are HyperSwap volumes.

If a disaster occurs at site 1, PowerHA automatically switches the IASP to the system at site 2, and the workload can be resumed at site 2.

After the failure at site 1 is fixed, use PowerHA to switch the IASP back to site 1 and resume the workload at this site.

Planned outage with Live Partition Mobility

IBM PowerVM Live Partition Mobility (LPM) allows you to move a running logical partition (including its operating system and running applications) from one system to another without any shutdown or disruption to operation of that logical partition.

In this scenario, LPM is combined with HyperSwap to transfer the workload onto site 2 during a planned outage of site 1. This combination requires VIOS NPIV attachment and all IBM i LUNs configured as HyperSwap LUNs.

For more information about LPM and its requirements, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

To use LPM, you must define the IBM i host in IBM Spectrum Virtualize with the WWPNs of the second port of the virtual FC adapters. We recommend creating a separate host object definition for the secondary ports to specify site 2 for this host object. Then, you enable the I/O rate to be transferred to the nodes at site 2 after migrating the IBM i partition with LPM.

After the outage is completed, you can use LPM again to transfer the IBM i partition back to site 1. After the migration, the I/O rate automatically moves to the nodes at site 1.

Important: LPM now supports multiple client virtual FC (vFC) adapter ports being mapped to a single physical FC port. Each client virtual FC must be mapped to a separate physical port in advance, whether LPM with FC N_Port ID Virtualization is used. That restriction was removed for the use of Virtual I/O Server version 3.1.2.10 or later and IBM i 7.2 or later. Therefore, the same physical port can be double-mapped to the same IBM i client partition. This configuration allows for better use of the adapter.

IBM Db2 mirroring for IBM i

The Db2® Mirror base configuration consists of two systems that are in the same data center. This configuration cannot span locations because it is active-active read/write, which mean that by definition, all write operations are synchronous (by using Remote Direct Memory Access [RDMA] over Converged Ethernet [RoCE] network) to the application state. The write operations between two systems necessitate that the distances between the systems are limited to not affect performance.

The following two broad approaches are used to deploy active-active solutions:

- Distributed lock management, in which multiple application servers can access the common or shared database but are prevented from performing simultaneous updates by the distributed lock management, which locks out the other users while an update is done.
- The replication approach, in which each update of any type is synchronous to the application state. When an application has an update, it does not proceed to the next application step until the current write operations completed on the primary and secondary objects. This replication approach is referred to as a two-phase commit between two systems.

Note: Applications can be deployed in an active-active manner, in which each application server has simultaneous access to the database on both systems in the two-node active-active complex. If one of the database servers fails, the application servers continue performing I/O operations to the other system in the mirrored pair. This configuration includes the added benefit of enabling workload balancing.

However, applications also can be deployed in an active-passive manner, where application servers conduct write operations to one of the two systems in the two-system complex and, if the primary is removed, the application groups are switched to the secondary system.

The active-active case necessitates that the application servers be hosted separately from the database servers and be connected through a client/server construct, such as Java Database Connectivity (JDBC).

Note: IBM i JDBC drivers now contain alternative server fail-over support to automatically transition the JDBC request between systems when one connection is no longer available. For many IBM i application workloads, deployment is through the traditional 5250 emulation window and contained in the same LPAR as the operating system and database. In this case, if the primary fails, the database is continuously replicated to the secondary system synchronously and is immediately available. The application must be restarted on the secondary system *before* the workload processing is resumed.

When one of the systems in the Db2 Mirror configuration is unavailable, Db2 Mirror tracks all update, change, and delete operations to the database table and all other mirror-eligible objects. When the pair is reconnected, changes are synchronized between the systems, including databases that are in an Independent Auxiliary Storage Pool (IASP) or are part of the base system storage.

Db2 Mirror is compatible with IASPs and uses IASPs for IFS support within the Db2 Mirror configuration. For non-IFS objects, IASPs can be used, but are not required.

In addition, Db2 Mirror supports applications that use traditional record-level access or SQL-based database access. Support for IFS and IFS journals is accomplished through deployment into an IASP, which can be configured as a switchable LUN or in a mirrored pair of IASPs through storage replication.

This solution requires the following software:

- IBM Power Systems POWER8® or later
- IBM i 7.4 with IBM Db2 Mirror for i V7.4 (5770-DBM)
- IBM i Option 48 and Db2 Data Mirroring are required for Db2 Mirror for i. Therefore, entitlement for Option 48 is automatically included with Db2 Mirror for i orders. Make sure that IBM i Option 48 is installed and a key is applied with the Db2 Mirror for i Licensed Program Product.

For more information about the software requirements for Db2 Mirror see this IBM Documentation web page.

DR can be performed with various options, such as:

- The IBM PowerHA SystemMirror for i Enterprise Edition
- Full system replication
- Logical replication

Important: Consider the following points when Db2 Mirror local continuous availability is combined with HA) and DR replication technologies:

- Remote replication for DR can be implemented by storage-based replication; that is, using the Copy Services of IBM Spectrum Virtualize software.
- Integrated File System (IFS) IASP must remain switchable between local Db2 Mirror nodes by choosing a DR topology that is supported by IBM PowerHA SystemMirror for i.
- DB IASP is available on both local nodes (no switch between local nodes).

  A DB IASP is not required for local Db2 Mirror database replication, but might be preferred for implementing a remote replication solution with shorter recovery times compared to SYSBAS replication.
- For a complete business continuity solution at the DR site, a remote DB2 Mirror node pair can be configured for a four-node Db2 Mirror PowerHA Cluster configuration. IFS IASPs and DB IASPs must be registered with the remote DB2 Mirror pair (by using the SHADOW option for the DB IASP to maintain its Db2 Mirror configuration data, such as default inclusion state and RCL).

For more information, see *IBM Db2 Mirror for i Getting Started*, REDP-5575.

Overview of the setup process

The following node types are part of the setup and configuration of Db2 Mirror:

- Managing
- Setup source
- Setup copy mode

For more information about the nodes, setup, and configuration, see this IBM Documentation web page.

Db2 Mirror is initially configured on a single partition, which is the setup source node. During the setup and configuration process, the setup source node is cloned to create the second node of the Db2 Mirror pair, which is the setup copy node. The setup copy node is configured and initialized automatically by Db2 Mirror during its first IPL.

The Db2 Mirror configuration and setup process supports external and internal storage. For the external storage that is used during the cloning process, IBM storage systems are recommended rather that non-IBM external storage because Db2 Mirror automates the cloning for IBM Spectrum Virtualize family.

The following cloning technologies are used for IBM storage systems:

- FlashCopy (cold and warm) is used when both Db2 Mirror nodes connect to the same IBM Spectrum Virtualize storage system. Consider the following points:
  - A cold clone requires the setup source node to be shut down during the cloning portion of the setup process.
  - A warm clone allows the setup source node to remain active during the entire Db2 Mirror setup and configuration process.
- Remote Copy is used when the Db2 Mirror nodes are connected to different IBM Spectrum Virtualize storage.

For more information, see this IBM Documentation web page.

Note: Although volume mirroring that is supported in IBM FlashSystem 9200 and IBM SAN Volume Controller is a valid cloning method for DB2 Mirror in the manual copy category, it is not automated, unlike the use of FlashCopy, Metro Mirror, or Global Mirror, where mirroring *is* automated.

IBM Spectrum Virtualize and Db2 Mirror

IBM Spectrum Virtualize storage systems establish communication with Db2 Mirror by using Secure Shell (SSH) to manage Copy Services functions. By IBM Spectrum Virtualize user IDs must have the administrator role. In that context, the following products are mandatory for managing nodes:

- 5733SC1 *BASE IBM Portable Utilities for i
- 5733SC1 Option 1 OpenSSH, OpenSSL, zlib
- 5770SS1 Option 33 Portable Application Solutions Environment

Note: For more information about creating an SSH key pair, see this IBM Documentation web page. After an SSH key pair is created, attach the SSH public key to a use on the IBM Spectrum Virtualize storage system. The corresponding private key file must be uploaded to the managing node so it can be used during DB2 Mirror setup.

Virtual I/O Server and native attachment

The Db2 Mirror storage cloning process for IBM Spectrum Virtualize requires Fibre Channel adapters with native attachment or attachment with Virtual I/O Server N_Port ID Virtualization.

Host object definition and volume planning

Before you set up Db2 Mirror, you must define the host object and assign volumes to the hosts to be used by the setup copy node. The following prerequisites must be met:

- The same number of host objects and volumes are available
- The same size volumes are defined for the setup source node and setup copy node

The Db2 Mirror cloning process pairs storage volumes between the setup source node and setup copy node. The cloning process applies to SYSBAS and IASPs:

- The setup source node and setup copy node must have the same number and sizes of LUNs or disks in SYSBAS.
- The host object and volumes for any database IASPs must be predefined for the setup copy node before a database IASP is added to DB2 Mirror.

Remote Copy cloning

Db2 Mirror Remote Copy cloning uses the following IBM Spectrum Virtualize Copy Services operations to copy the setup source node volumes to the setup copy nodes volumes:

- Global Mirror for cold clone
- GMCV for warm clone

Regardless of whether you plan to perform the Remote Copy during a planned outage window, you must ensure that your bandwidth between storage systems is sufficient to complete the Remote Copy during that period. The Db2 Mirror cloning process does not provide the capability of pausing the cloning and then resuming it later. Therefore, you must plan enough time for the Remote Copy to complete.

Important: For IBM Spectrum Virtualize, the Copy Services partnership between storage systems must be manually created before Db2 Mirror is configured.

Architectures and considerations for DB2 Mirror

Because of the synchronous design of Db2 Mirror, the distance between the nodes is limited to within a data center in most cases. Multiple configurations are supported for a data center Db2 Mirror implementation and the addition of a DR solution.

Several options are discussed in this section as examples with IBM Spectrum Virtualize storage systems. A specific implementation depends on your business resilience requirement.

Note: Db2 Mirror supports IBM SAN Volume Controller topologies, such as Enhanced Stretched Cluster or HyperSwap.

Db2 Mirror environment with one IBM Spectrum Virtualize storage system

In this example, one IBM Spectrum Virtualize storage system is used as a basic configuration for the use of Db2 Mirror. This configuration features some key advantages.

By using one storage system, you can take advantage of FlashCopy to set up your configuration rapidly. You can consider this solution as a DR strategy to provide storage resiliency.

Figure A-14 shows two IBM Power System servers are used (at least one RoCE adapter per server). However, you can reduce this scenario in terms of cost of decreased resiliency by implementing Db2 Mirror across two IBM i LPARs on the same IBM Power Systems. For this example, a SYSBAS is cloned; however, IASP also can be added by using another set of volumes.

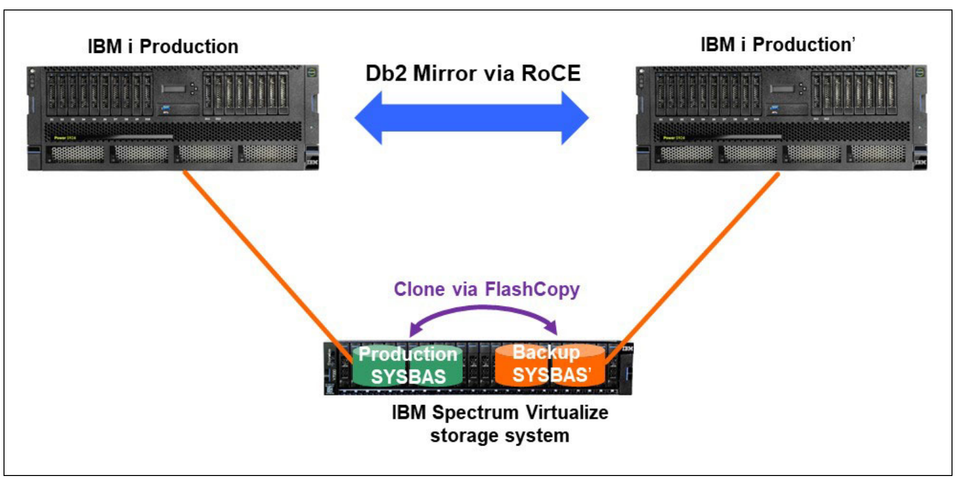


Figure A-14 Db2 Mirror environment with one IBM Spectrum Virtualize storage

DB2 Mirror environment with two IBM Spectrum Virtualize storage systems

The use of two IBM Spectrum Virtualize storage systems provides further redundancy by helping to ensure that the active node remains running and available during a storage outage. In this example, two IBM Power Systems servers and IBM Spectrum Virtualize storage systems are used. Also, Remote Copy is used to set up DB2 Mirror.

As shown in Figure A-15, the set of volumes for SYSBAS and the set of volumes for IASP are replicated. Global Mirror can also be used.

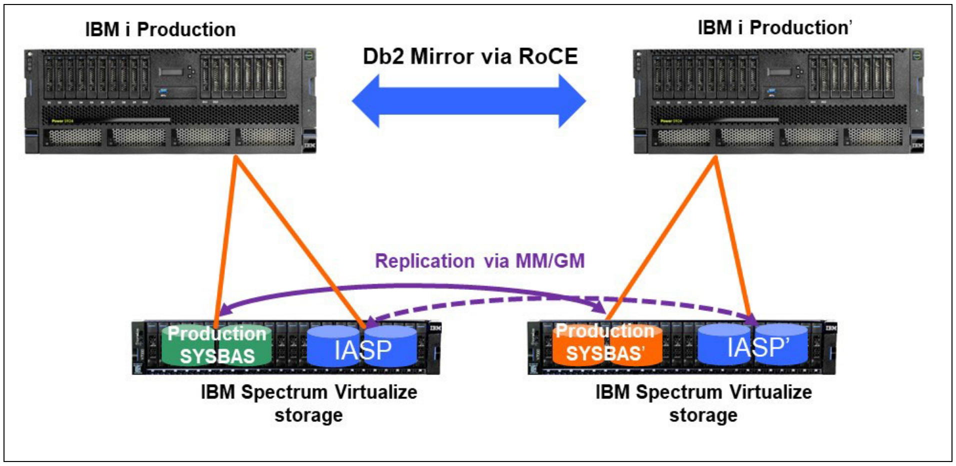


Figure A-15 Db2 Mirror environment with two IBM Spectrum Virtualize storages

Db2 Mirror and DR considerations

Db2 Mirror is a continuous availability solution, but it is *not* considered a DR solution. However, Db2 Mirror can be used within your DR strategy to improve your availability, even within a disaster situation.

The Db2 Mirror nodes must be close to each other because the maximum distance between IBM Power Systems servers is 200 meters (656 feet). Consider the following points:

- At Site 1 (the continuous-availability location), Db2 Mirror nodes are used.
- At Site 2 (the DR location), you can have a single server or multiple servers with Db2 Mirror nodes can b, and a unique or multiple IBM Spectrum Virtualize storage systems.

The communication between the continuous availability at Site 1 and the DR at Site 2 can be achieved by using specific technology, such as the following examples:

- IBM PowerHA SystemMirror for i using Metro Mirror
- Global Mirror with IASPs
- Full system replication
- Logical replication from third-party vendor

Db2 Mirror and full system replication

The use of a mirrored pair within the disaster site provides extra protection if you are required to role-swap to the DR location. With this scenario, a continuously available environment exists in DR.

A topology with multiple IBM Spectrum Virtualize storage systems and multiple IBM Power Systems servers is shown in Figure A-16.

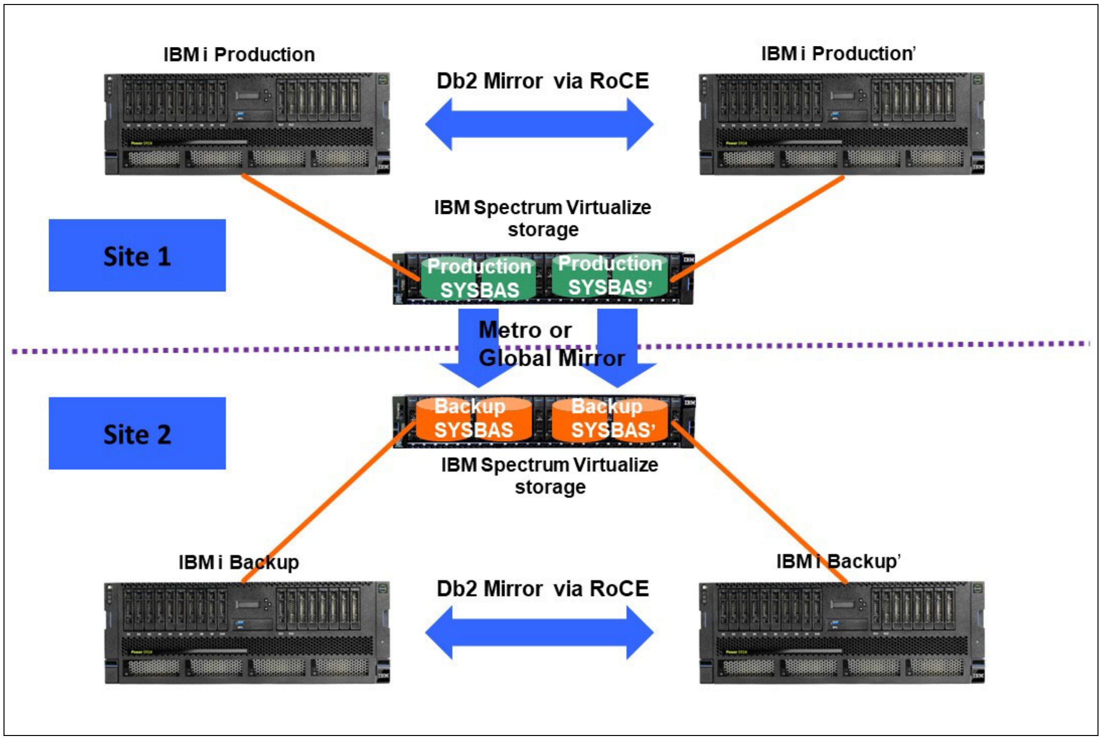


Figure A-16   DB2 Mirror and full system replication

Full system replication is fully supported. If IASP is not used, this type of replication can be done for IBM i at the IBM Spectrum Virtualize storage level.

In Figure A-16, the following sites are configured:

- Site 1: Db2 Mirror with DR and Db2 Mirror production

  At Site 1, an active side exists because of full system replication.

- Site 2: Db2 Mirror with DR

  At Site 2, the IBM i systems are powered off, and the replication is active across sites.

Two copies are at a DR location because if one side fails, the other side must continue replicating. If only three nodes are replicating, you cannot predict which side fails and does not have a valid copy of storage to switch.

Related publications

The publications that are listed in this section are considered particularly suitable for a more detailed discussion of the topics that are bcovered in this book.

IBM Redbooks

The following IBM Redbooks publications provide more information about the topic in this document. Note that some publications that are referenced in this list might be available in softcopy only:

- *Implementing the IBM FlashSystem with IBM Spectrum Virtualize Version 8.4.2*, SG24-8506
- *Implementing the IBM SAN Volume Controller with IBM Spectrum Virtualize Version 8.4.2*, SG24-8507
- IBM SAN Volume Controller Best Practices and Performance Guidelines for IBM Spectrum Virtualize V8.4.2, SG24-8509
- *IBM DS8900F Architecture and Implementation*, SG24-8456.
- *IBM Spectrum Virtualize 3-Site Replication*, SG24-8474
- *IBM Spectrum Virtualize: Hot Spare Node and NPIV Target Ports*, REDP-5477
- *Implementation Guide for SpecV/FlashSystem Safeguarded Copy*, REDP-5654
- *IBM Spectrum Virtualize HyperSwap SAN Implementation and Design Best Practices*, REDP-5597
- *Automate and Orchestrate Your IBM FlashSystem Hybrid Cloud with Red Hat Ansible*, REDP-5598
- *IBM FlashSystem 9200 Product Guide*, REDP-5586
- *IBM FlashSystem 9100 Product Guide*, REDP-5524
- *IBM FlashSystem 7200 Product Guide*, REDP-5587
- *IBM FlashSystem 5200 Product Guide,* REDP-5617
- *IBM FlashSystem 5000 and 5100 for Mid-Market*, REDP-5594
- *IBM Flashsystem 5000 and 5200 for Mid-Market,* REDP-5630
- *IBM DS8870 Easy Tier Heat Map Transfer*, REDP-5015

You can search for, view, download, or order these documents and other Redbooks, Redpapers, Web Docs, draft, and other materials at the following website:

ibm.com/redbooks

Redbooks

IBM FlashSystem Best Practices and Performance Guidelines for

SG24-8508-00

ISBN 0738460311

IBM

(1.0” spine)
0.875”<->1.498”
460 <-> 788 pages

SG24-8508-00

ISBN 0738460311

Printed in U.S.A.

Get connected

ibm.com/redbooks