ibm.com/redbooks

IBM FlashSystem 9200 and 9100 Best Practices and Performance Guidelines

Tiago Moreira Candelaria Bastos

Jon Herd

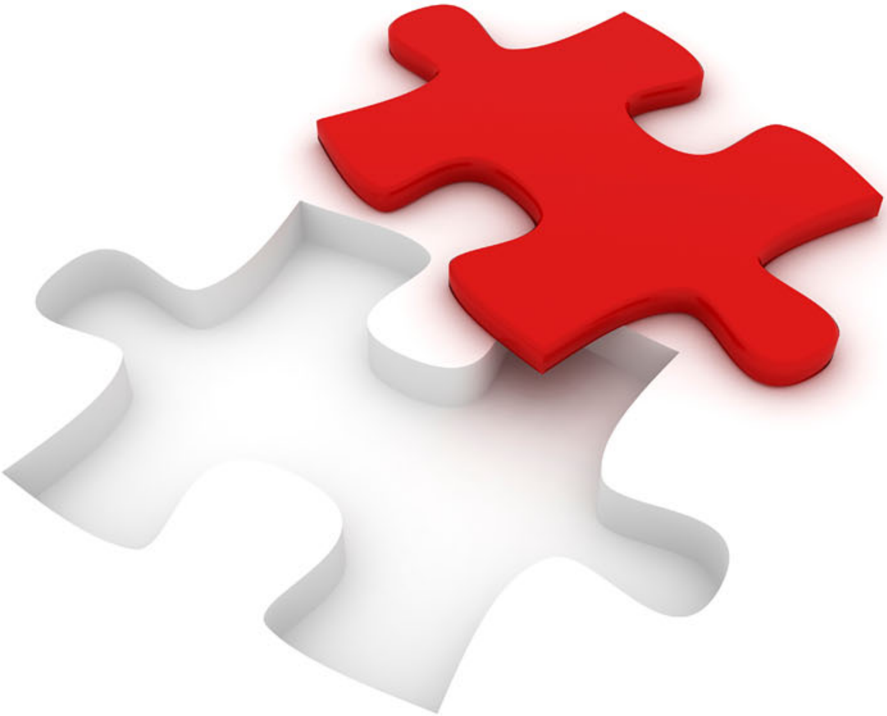
Sang-hyun Kim

Sergey Kubin

Dirk Peitzmann

Antonio Rainero

Jon Tate



IBM Redbooks

IBM FlashSystem 9200 and 9100 Best Practices and Performance Guidelines

February 2020

SG24-8448-00

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

First Edition (February 2020)

This edition applies only to the hardware and software products and features described and documented in this book. On 11th February 2020 IBM announced the arrival of the IBM FlashSystem 9200 to the family. This book was written specifically for IBM FlashSystem 9100, however most of the general principles will apply to the IBM FlashSystem 9200. If you are in any doubt as to their applicability to the FlashSystem 9200 then you should work with your local IBM representative.

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®
developerWorks®
DS8000®
Easy Tier®
FlashCopy®
Global Technology Services®
HyperSwap®
IBM®
IBM Cloud™
IBM FlashCore®
IBM FlashSystem®
IBM Research™
IBM Spectrum®
IBM Spectrum Storage™
Insight®
Netcool®
POWER®
PowerHA®
Redbooks®
Redbooks (logo) ®
Service Request Manager®
Storwize®
System Storage™
Tivoli®
XIV®
z/OS®

The following terms are trademarks of other companies:

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.

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® 9200 and 9100 (FlashSystem). 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 (SVC), and IBM Storwize® administrators and technicians. Understanding this book requires advanced knowledge of these environments.

Important, IBM FlashSystem 9200: On 11th February 2020 IBM announced the arrival of the IBM FlashSystem 9200 to the family. This book was written specifically for IBM FlashSystem 9100, however most of the general principles will apply to the IBM FlashSystem 9200.

If you are in any doubt as to their applicability to the FlashSystem 9200 then you should work with your local IBM representative.

This book will be updated to include FlashSystem 9200 in due course.

The Flashsystem 9200 product guide is available at:

IBM FlashSystem 9200 Product Guide

Authors

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

Tiago Moreira Candelaria Bastos is a SAN and Storage Disk specialist at IBM Brazil. He has over 20 years of experience in the IT arena, and is an IBM Certified Master IT Specialist. Certified for Storwize, he works on Storage as a Service implementation projects and his areas of expertise include planning, configuring, and troubleshooting IBM DS8000®, Storwize V5000 and V7000, FlashSystem 900, SVC and IBM XIV®, lifecycle management, and copy services.

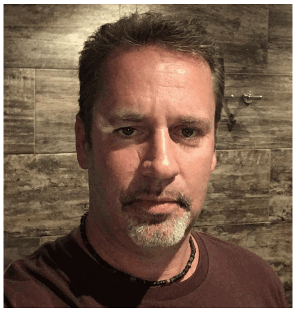
Jon Herd is an IBM Storage Technical Advisor working for the ESCC, Germany. He covers the United Kingdom, Ireland, and Sweden, advising customers on a portfolio of IBM storage products, including the FlashSystem products. Jon has been with IBM for more than 40 years, and has held various technical roles, including Europe, Middle East, and Africa (EMEA) level support on mainframe servers and technical education development. He has written many Redbooks publications about the FlashSystem products and is a Redbooks Platinum level author. He holds IBM certifications in Supporting IT Solutions at an expert level, and Technical IT Specialist at an experienced level. He also chairs the UKI PS profession certification board.

He is a certified Chartered Member of the British Computer Society (MBCS - CITP), and a Certified Member of the Institution of Engineering and Technology (MIET).

Sang-hyun Kim, at the time of writing, was a Storage Client Technical Specialist for IBM Systems group based in South Korea. He was with IBM for more than 11 years, and worked on IBM storage, including IBM FlashSystem, IBM Storwize, IBM DS8000, and Software Defined Storage software. Before joining the Technical Sales team in 2014, he built up 6 years of experience working with IBM storage and POWER® servers, providing technical support to customers and helping clients to improve their infrastructure environments. Sang-hyun has since left IBM.

Sergey Kubin is a subject matter expert (SME) for IBM Storage and SAN technical support. 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 works for IBM Technology Support Services, providing support and guidance on IBM Spectrum® Virtualize family systems for customers in Europe, Middle East and Russia.

His expertise also includes SAN, IBM block-level DS family storage, and file-level storage systems. He is an IBM Certified Specialist for Storwize Family Technical Solutions.



Dirk Peitzmann is a Leading Technical Sales Professional with IBM Systems Sales located in Munich, Germany. Dirk is an experienced professional providing technical pre-sales and post-sales solutions for IBM server and storage systems since 1995. His areas of expertise include designing virtualization infrastructures and disk solutions, as well as carrying out performance analysis and the sizing of SAN and NAS solutions. He holds an engineering diploma in Computer Sciences from the University of Applied Science in Isny, Germany.

Antonio Rainero is an Executive Technical Specialist working for the IBM Global Technology Services® organization in IBM Italy. He joined IBM in 1998, and 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.

Jon Tate is a Project Manager for IBM System Storage SAN Solutions at the ITSO, San Jose Center. Before joining the ITSO in 1999, he worked in the IBM Technical Support Center, providing Level 2/3 support for IBM mainframe storage products. Jon has 33 years of experience in storage software and management, services, and support. He is an IBM Certified IT Specialist, an IBM SAN Certified Specialist, and is Project Management Professional (PMP) certified. He is also the UK Chairman of the Storage Networking Industry Association (SNIA).

Special thanks to the following people for their participation and contribution to this project:

Matt Smith
IBM Systems, UK

Frank Enders
IBM GTS, Germany

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

Christopher Bulmer
Ben Sasson
Djihed Afifi
Alastair Cooper
Alex Ainscow
Carlos Fuente
Paul Marris
IBM Systems, Hursley, UK

Barry Whyte
IBM Systems, New Zealand

Ian MacQuarrie
IBM Systems, US

Markus Standau
IBM Systems, Germany

Joshua Martin
IBM Systems, US

Dave Gimpl
IBM Systems, US

Duane Bolland
IBM Systems, US

Jim Olson
IBM GTS, US

Bruno Galle
IBM Systems, Brazil

Dongsung Yun
IBM Global Markets, Korea

Sudharsan Vangal
IBM GTS, India

Jeff Irving
IBM Systems, US

Kavita Shah
IBM Systems, US

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

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

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

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

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

- Stay current on recent Redbooks publications with RSS Feeds:

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

1

IBM FlashSystem 9100 introduction

This chapter introduces the IBM FlashSystem 9100 (FS9100) storage system and its key features, benefits, and technology.

Important, IBM FlashSystem 9200: On 11th February 2020 IBM announced the arrival of the IBM FlashSystem 9200 to the family. This book was written specifically for IBM FlashSystem 9100, however most of the general principles will apply to the IBM FlashSystem 9200.

If you are in any doubt as to their applicability to the FlashSystem 9200 then you should work with your local IBM representative.

This book will be updated to include FlashSystem 9200 in due course.

The Flashsystem 9200 product guide is available at:

IBM FlashSystem 9200 Product Guide

This chapter includes the following topics:

- IBM FlashSystem 9100 high-level features
- Clustering rules and upgrades
- Advanced data reduction features
- Advanced software features
- IBM HyperSwap
- Licensing

1.1 IBM FlashSystem 9100 high-level features

This IBM Redbooks publication describes the IBM FlashSystem 9100, which is a comprehensive all-flash, NVMe enabled, enterprise storage solution that delivers the full capabilities of IBM FlashCore® technology. In addition, this Redbook focuses on the best practises 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, and IBM HyperSwap® for high availability.

> **Note:** The detailed technical explanations, and theory of operations, of these features are not covered in this publication. If you need extra information in this area, see *IBM FlashSystem 9100 Architecture, Performance, and Implementation*, SG24-8425.

With the release of IBM FlashSystem 9100 software V8.2, support for the new and more powerful NVMe based IBM FlashCore Modules (FCM) within the control enclosure is included. Additonal software features added also include GUI enhancements, a new dashboard, remote support assistance, data deduplication and Storage Insights configuration.

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

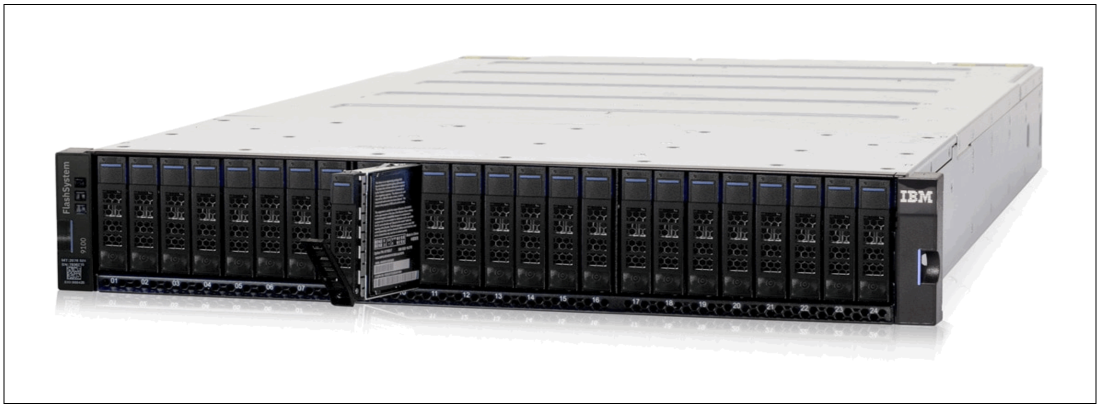


Figure 1-1 BM FlashSystem 9100 control enclosure with one NVMe drives partially removed

The IBM FlashSystem 9100 system has two different types of enclosures: control enclosures and expansion enclosures.

- A *control enclosure* manages your storage systems, communicates with the host, and manages interfaces. In addition, it can also house up to 24 NVMe capable flash drives. These drives can be either industry-standard NVMe type or the exclusive IBM NVMe FlashCore Modules (FCM).
- An *expansion enclosure* enables you to increase the available capacity of the IBM FlashSystem 9100 cluster. it communicates with the control enclosure via a pair of 12 Gbps SAS connections. These expansion enclosures can house a large number of flash (SSD) SAS type drives, depending on which model of enclosure is ordered.

1.1.1 Control enclosures

Each control enclosure can have multiple attached expansion enclosures, which expands the available capacity of the whole system. The IBM FlashSystem 9100 system supports up to four control enclosures and up to two chains of SAS expansion enclosures per control enclosure.

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

There are two standard models of IBM FlashSystem 9100: 9110-AF7 and 9150-AF8.

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

There are also two utility models of the IBM FlashSystem 9100: the 9110-UF7 and 9150-UF8.

> **Note:** 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 respectively, with the exception of 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 above a base subscription rate to be billed per TB, per month. Allocated storage is identified as storage that is allocated to a specific host (and unusable to other hosts), whether data is written or not. For thin-provisioning, the data that is actually written is considered used. For thick provisioning, total allocated volume space is considered used.

1.1.2 Expansion enclosures

New SAS-based small form factor (SFF) and large form factor (LFF) expansion enclosures support flash-only MDisks in a storage pool, which can be used for IBM Easy Tier:

- The new IBM FlashSystem 9100 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 9100 control enclosure. The expansion enclosure is 2U high.
- The new IBM FlashSystem 9100 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 9100 control enclosure. The expansion enclosure is 5U high.

The IBM FlashSystem 9100 control enclosure can be recognized by the nomenclature IBM FlashSystem 9100 on the left side of the bezel cover, which covers the rack mounting screws.

Figure 1-2 shows the IBM FlashSystem 9100 bezel and NVMe drive description.

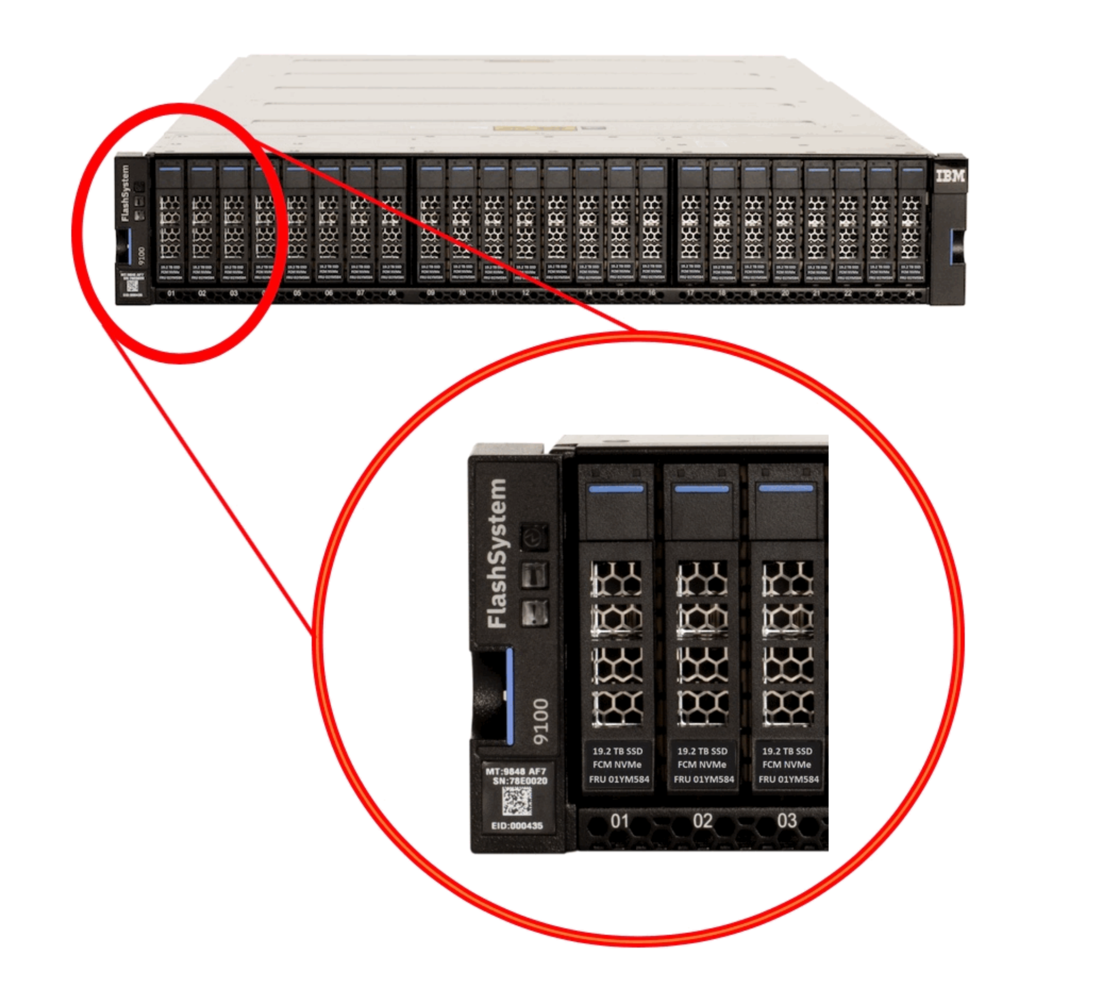


Figure 1-2 IBM FlashSystem 9100 bezel and NVMe drive description

Labeling on the NVMe drive itself gives the drive type, capacity, the type of drive, and the FRU number. The example shown in Figure 1-2 is the IBM 19.2 TB NVMe FlashCore Module type.

The FS9100 Model 9110 has a total of 32 cores (16 per canister) while the 9150 has 56 cores (28 per canister).

The FS9100 supports six different memory configurations, as shown in Table 1-1.

Table 1-1 FS9100 memory configurations

Memory per Canister	Memory per Control Enclosure
64 GB	128 GB
128 GB	256 GB
192 GB	384 GB
384 GB	768 GB
576 GB	1152 GB
768 GB	1536 GB

Note: FS9100 refers to both the FS9110 (Model AF7) and the FS9150 (model AF8). If a feature or function is specific to one of the models, then FS9110 or FS9150 will be used.

The FS9100 supports NVMe attached flash drives, both the IBM Flash Core Modules (FCM) and commercial off the shelf (COTS) SSDs. The IBM FCMs support hardware compression at line data rates. IBM offers the FCMs in three capacities: 4.8 TB, 9.6 TB, and 19.2 TB. Standard NVMe SSDs are offered in four capacities, 1.92 TB, 3.84 TB, 7.68 TB, and 15.36 TB.

The FS9100 also supports additional capacity in serial-attached SCSI (SAS) expansion enclosures. Up to 20 2U enclosures (Model AFF) or up to 8 5U enclosures (Model AF9) can be attached. Only SAS SSD drives are supported in the AFF and AF9 enclosures.

Host interface support includes 8 gigabit (Gb) and 16 Gb Fibre Channel (FC), and 10 Gb Fibre Channel over Ethernet (FCoE) or Internet Small Computer System Interface (iSCSI). Advanced Encryption Standard (AES) 256 hardware-based encryption adds to the rich feature set.

The IBM FlashSystem 9100 includes a single, easy-to-use management graphical user interface (GUI) to help you monitor, manage, and configure your system.

1.1.3 FlashSystem 9100 utility models UF7 and UF8

IBM FlashSystem 9100 utility models UF7 and UF8 provide a variable capacity storage offering. These models offer a fixed capacity, with a base subscription of 35% of the total capacity. IBM Storage Insights (free edition or pro) is used to monitor system usage, and capacity used beyond the base 35% is billed on a per-month, per-terabyte basis. This enables you to grow or shrink usage, and only pay for the configured capacity.

IBM FlashSystem utility models are provided for customers who can benefit from a variable capacity system, where billing is based only on actually provisioned space. The hardware is leased through IBM Global Finance on a three-year lease, which entitles the customer to utilize up to 35% of the total system capacity at no additional cost. If storage needs increase beyond that 35% capacity, usage is billed based on the average daily provisioned capacity per terabyte, per month, on a quarterly basis.

1.2 Clustering rules and upgrades

The IBM FlashSystem 9100 can be clustered with up to four FS9100 enclosures, using four I/O groups.

The FS9100 also allows for clustering with the IBM Storwize V7000 with a maximum of four enclosures total, if done following these guidelines:

- Both systems must have the same level of V8.2 code installed to be able to cluster.
- To cluster the Storwize V7000, it must have an all-inclusive license.
- When clustered, the clustered system presents itself as a FlashSystem 9100.
- Migration must be done through additional I/O groups.
- The default layer is storage, but a replication layer is supported.
- The FS9100 cannot be clustered with the IBM FlashSystem V9000 or the IBM Storage Virtualization Controller (SVC).

Note: At the time of writing, the supported release for clustering was V8.2.0 or later.

1.3 Advanced functions for data reduction

The IBM FlashSystem FS9100 can function as a feature-rich, software-defined storage layer that virtualizes and extends the functionality of all managed storage. These include data reduction, dynamic tiering, copy services, and high-availability configurations. In this capacity, it 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 FS9100 employs several features to assist with data reduction and the ability to increase its effective capacity.

1.3.1 FlashCore Modules (FCM)

The IBM FlashSystem FS9100 has the option to be supplied with either FCMs or industry standard NVMe drives. If the FCM option is chosen, then the user can take advantage of the built-in hardware compression, which will automatically try to compress the stored data when written to the drives. These FCM’s can either be used with standard pools or DRP pools.

1.3.2 Data reduction pools

Data reduction pools (DRP) 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 has become more of a requirement to have an uncomplicated way to stay “thin”.

1.3.3 Deduplication

Deduplication can be configured with thin-provisioned and compressed volumes in data reduction pools 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 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, resulting in the amount of data that must be stored being greatly reduced.

1.3.4 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).

1.3.5 Thin-provisioned flash copies

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.

1.4 Advanced software features

These are some of the advanced software features of FS9100.

- Data Migration
- Copy Services
  - Metro Mirror
  - Global Mirror
  - Global Mirror with Change Volumes
  - Flashcopy
  - Remote Mirroring
- External Virtualization
- EasyTier

1.4.1 Data migration

The IBM FlashSystem 9100 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 9100 delivers these functions in a homogeneous way on a scalable and highly available platform over any attached storage and to any attached server.

1.4.2 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 9100 (FlashCopy and Image Mode Migration), or between one IBM FlashSystem 9100 and another IBM FlashSystem 9100, or any other member of the IBM Spectrum Virtualize family, in three different modes:

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

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

Further detailed information see Chapter 6, “Copy services” on page 141.

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 9100. 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* 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.

1.4.3 External virtualization

The IBM FlashSystem 9100 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 9100 functions benefit all virtualized storage.

For example, Easy Tier and Data Reduction Pools 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 9100, they can operate smoothly together, reducing management effort.

1.4.4 Easy Tier

Easy Tier is a performance function that automatically migrates or moves extents of a volume to or from one storage tier to another storage tier. With IBM FlashSystem 9100, 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 IBM HyperSwap

HyperSwap capability enables each volume to be presented by two IBM FlashSystem 9100 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 9100. IBM FlashSystem 9100 provides GUI management of the HyperSwap function.

There is a more detailed overview and explanation of the HyperSwap function in Appendix A, “Business continuity” on page 401.

For more information on the HyperSwap function see
https://www.ibm.com/support/knowledgecenter/STSLR9_8.2.1/com.ibm.fs9100_821.doc/svc_hyperswapovr.html

1.6 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 more detailed overview and explanation of the licensing on the FS9100, see the chapter “Licensing and Features” in:

IBM FlashSystem 9100 Architecture, Performance, and Implementation, SG24-8425

2

Storage area network

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 to connect IBM Spectrum Virtualize and Storwize in an existing SAN to achieve a stable, redundant, resilient, scalable, and performance-likely environment. However, you can take the principles here into account when building your SAN.

> **Important, IBM FlashSystem 9200:** On 11th February 2020 IBM announced the arrival of the IBM FlashSystem 9200 to the family. This book was written specifically for IBM FlashSystem 9100, however most of the general principles will apply to the IBM FlashSystem 9200.
>
> If you are in any doubt as to their applicability to the FlashSystem 9200 then you should work with your local IBM representative.
>
> This book will be updated to include FlashSystem 9200 in due course.
>
> The Flashsystem 9200 product guide is available at:
>
> IBM FlashSystem 9200 Product Guide

This chapter includes the following sections:

- SAN topology general guidelines
- SAN topology-specific guidelines
- IBM FlashSystem 9100 controller ports
- Zoning
- Distance extension for remote copy services
- Tape and disk traffic that share the SAN
- Switch interoperability

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, as well as minimizing performance issues and bottlenecks. Therefore, before installing any of the products covered by this book, ensure that your environment follows an actual SAN design and architecture, with vendor recommended SAN devices and code levels.

For more SAN design and preferred practices, see the white paper at:

SAN Fabric Administration Best Practices Guide Support Perspective

A topology is described in terms of how the switches are interconnected. There are several different SAN topologies, such as core-edge, edge-core-edge, or full mesh. Each topology has its 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 13.

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/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 of the ISLs going to the switch where the Flash System 9200 and 9100 is connected will also handle the traffic coming from the switches on the edges, as shown in Figure 2-1.

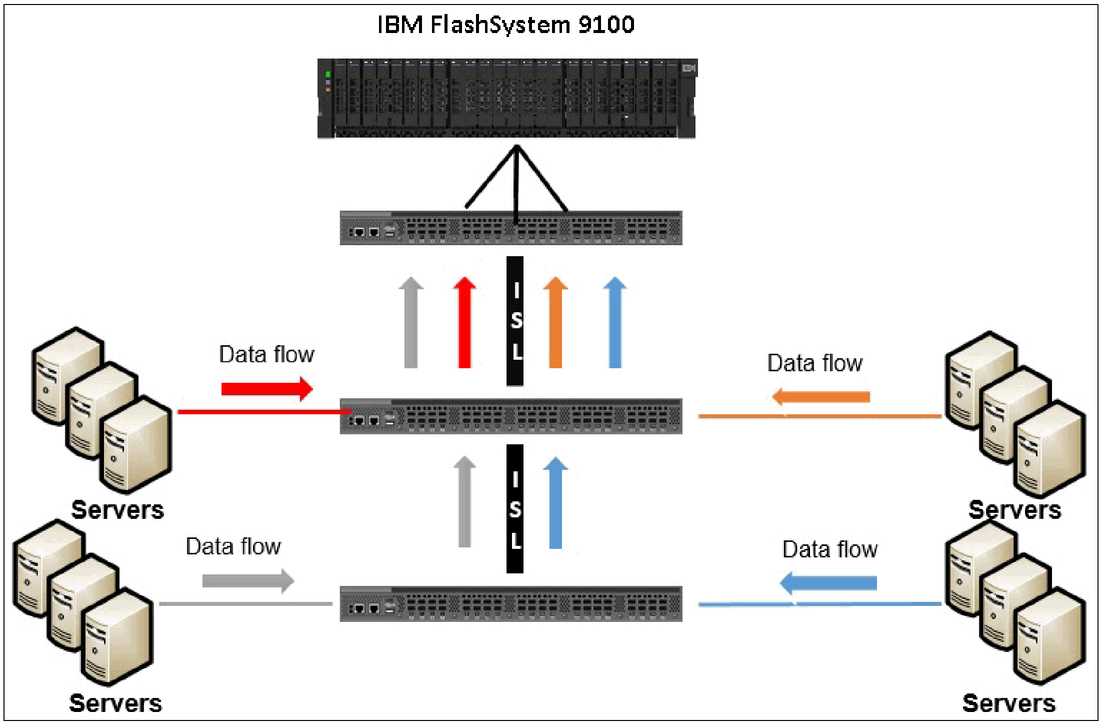


Figure 2-1 ISL data flow

- 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 interconnecting SAN directors through ISL, spread the ISL cables across different directors blades. In a situation where an entire blade fails, the ISL will still be redundant through the links connected to other blades.
- Plan for the peak load, not for the average load.

2.2 SAN topology-specific guidelines

Some preferred practices, as mentioned in 2.1, “SAN topology general guidelines” on page 12, apply to all SANs. However, there are specific preferred practices requirements to each SAN topology available. The following topic shows the difference between the different kinds of topology and highlights the specific considerations for each of them.

This section covers the following topologies:

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

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 is a low-cost design solution that 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 directors, with one per different fabric.

The preferred practice is to use a multislot director-class single switch over setting up a core-edge fabric that is made up solely of lower-end switches, as described in 2.1.1, “SAN performance and scalability” on page 12.

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

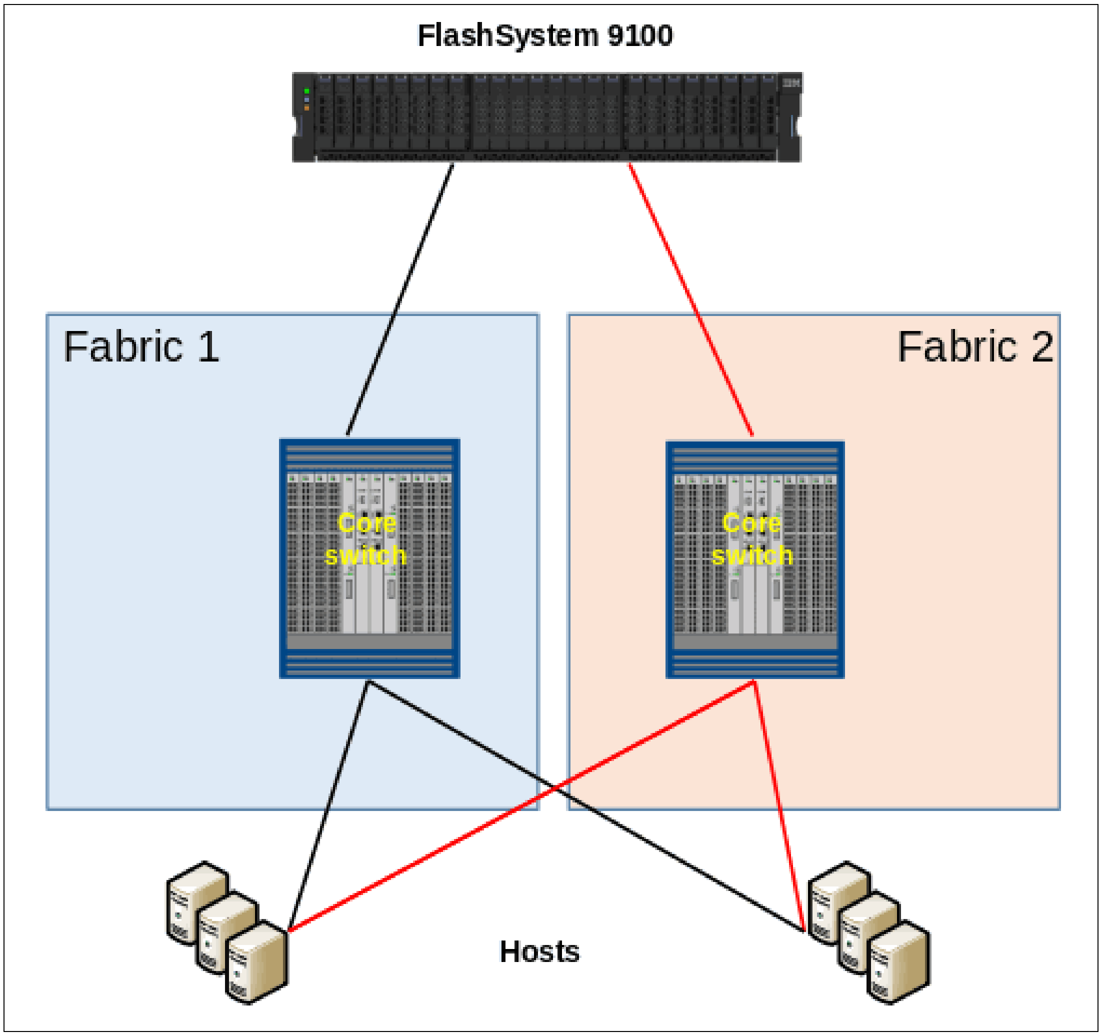


Figure 2-2 Single switch topology

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.

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

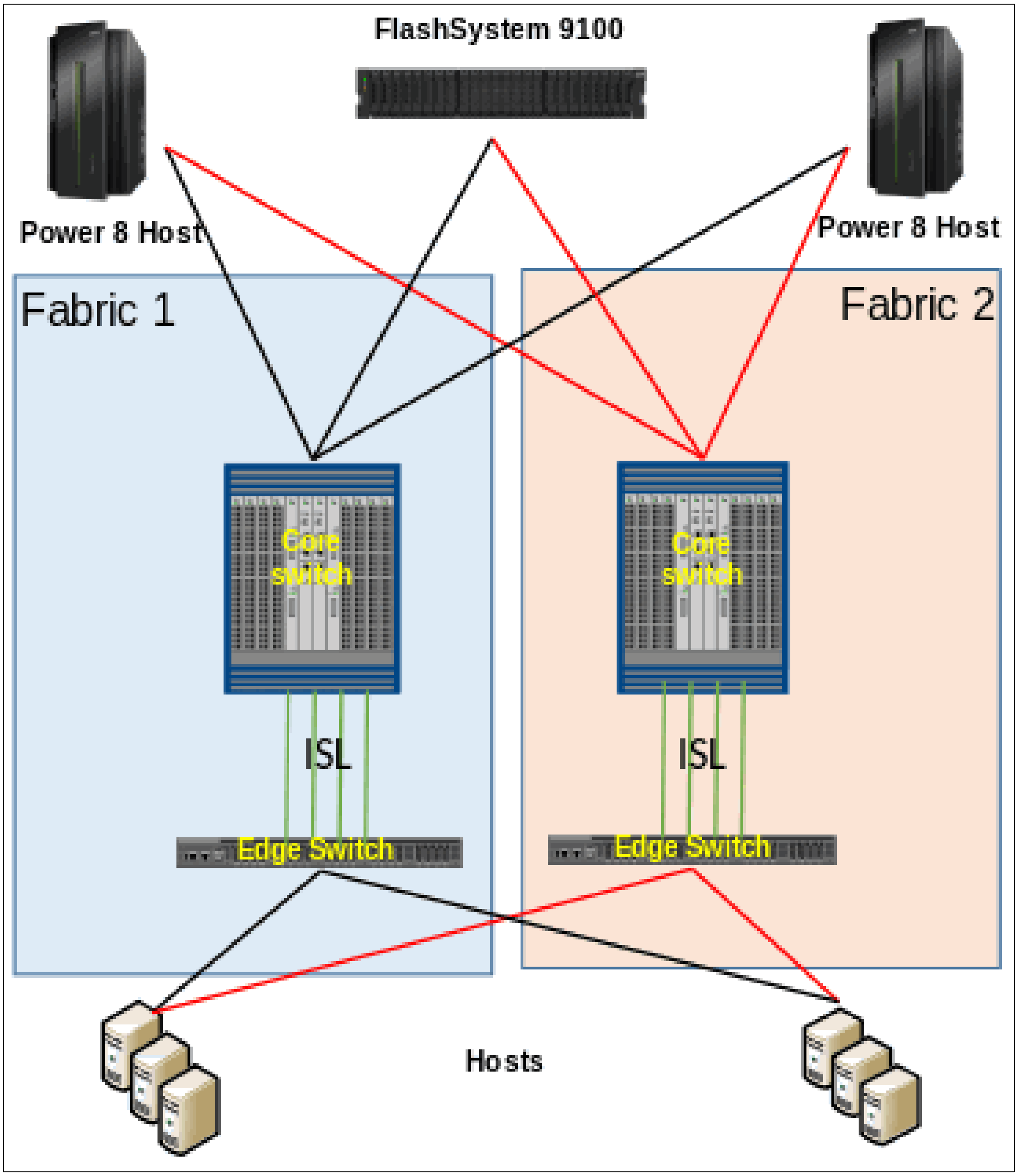


Figure 2-3 Core-edge topology

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

Figure 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.

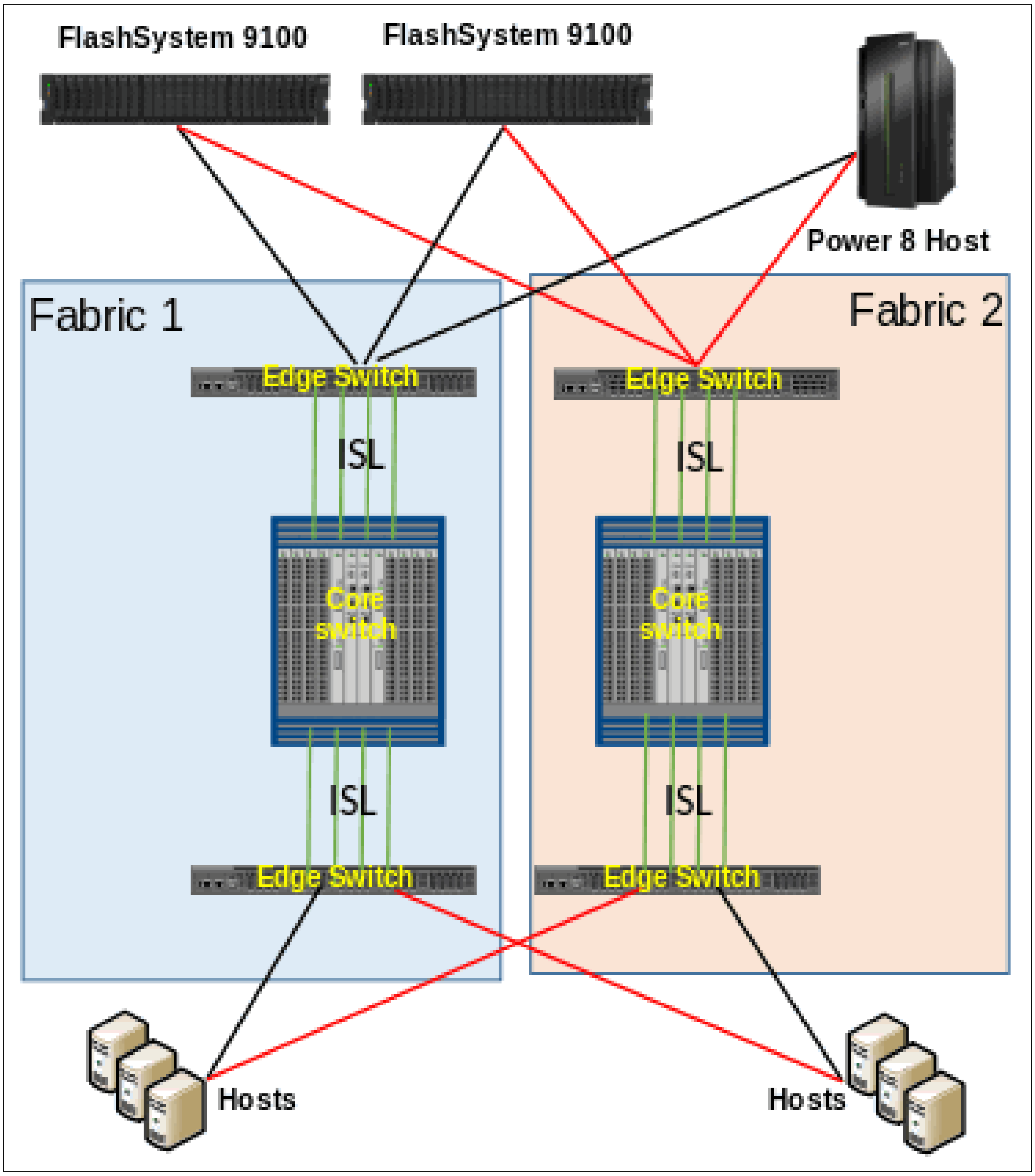


Figure 2-4 Edge-core-edge topology

Edge-core-edge fabrics allow better isolation between tiers. For additional information, see 2.2.6, “Device placement” on page 19.

2.2.4 Full mesh topology

In a full mesh topology, all switches are interconnected to all other switches on the same fabric. So the server and storage placement is not a concern after the number of hops is no more than one hop. Figure 2-5 shows a full mesh topology.

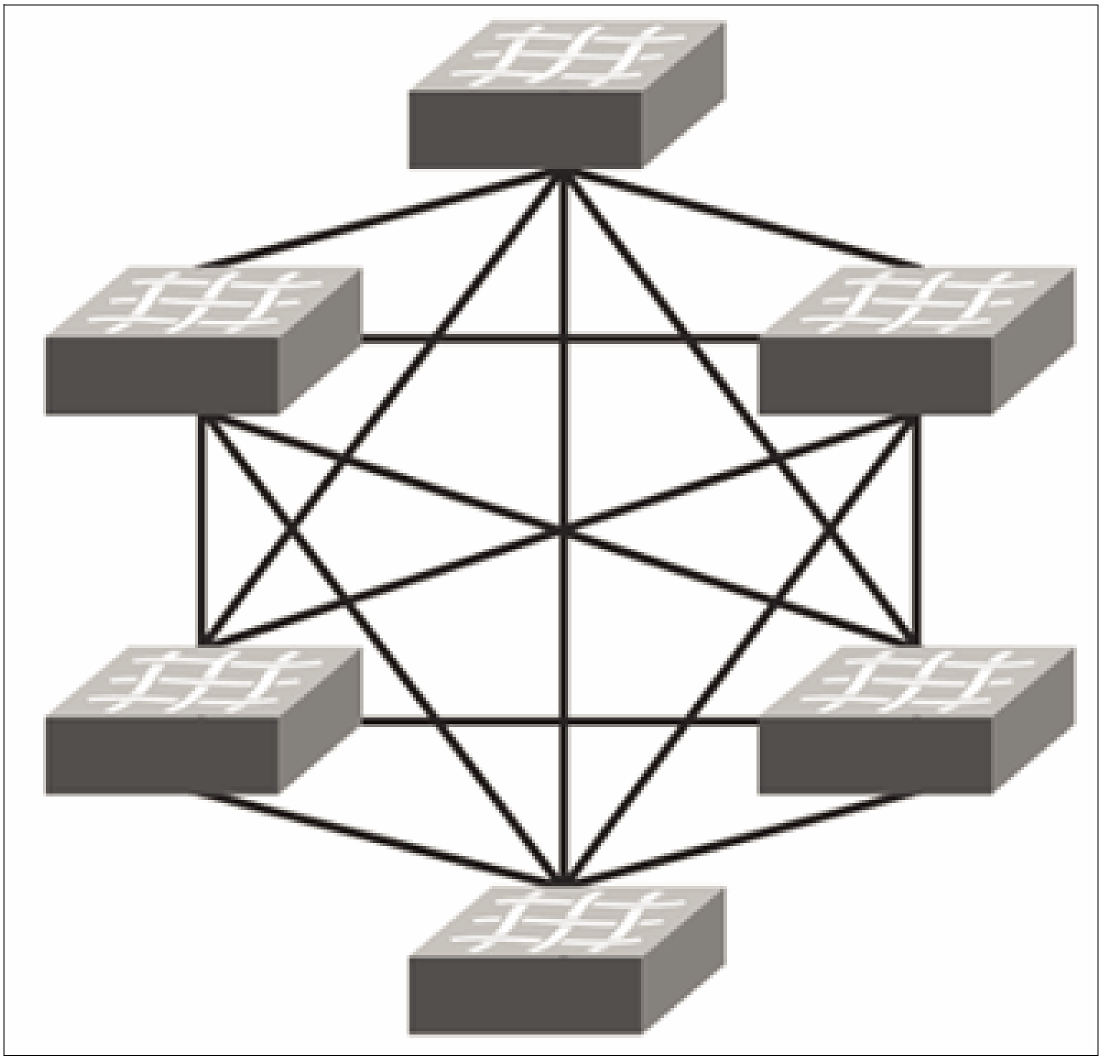


Figure 2-5 Full mesh topology

2.2.5 IBM FlashSystem as a SAN bridge

IBM FlashSystem now has 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 bridge between two isolated SANs. This configuration is useful for storage migration or sharing resources between SAN environments without merging them. Another use is if you have devices with different SAN requirements in your installation.

Figure 2-6 has an example of an FlashSystem as a SAN bridge.

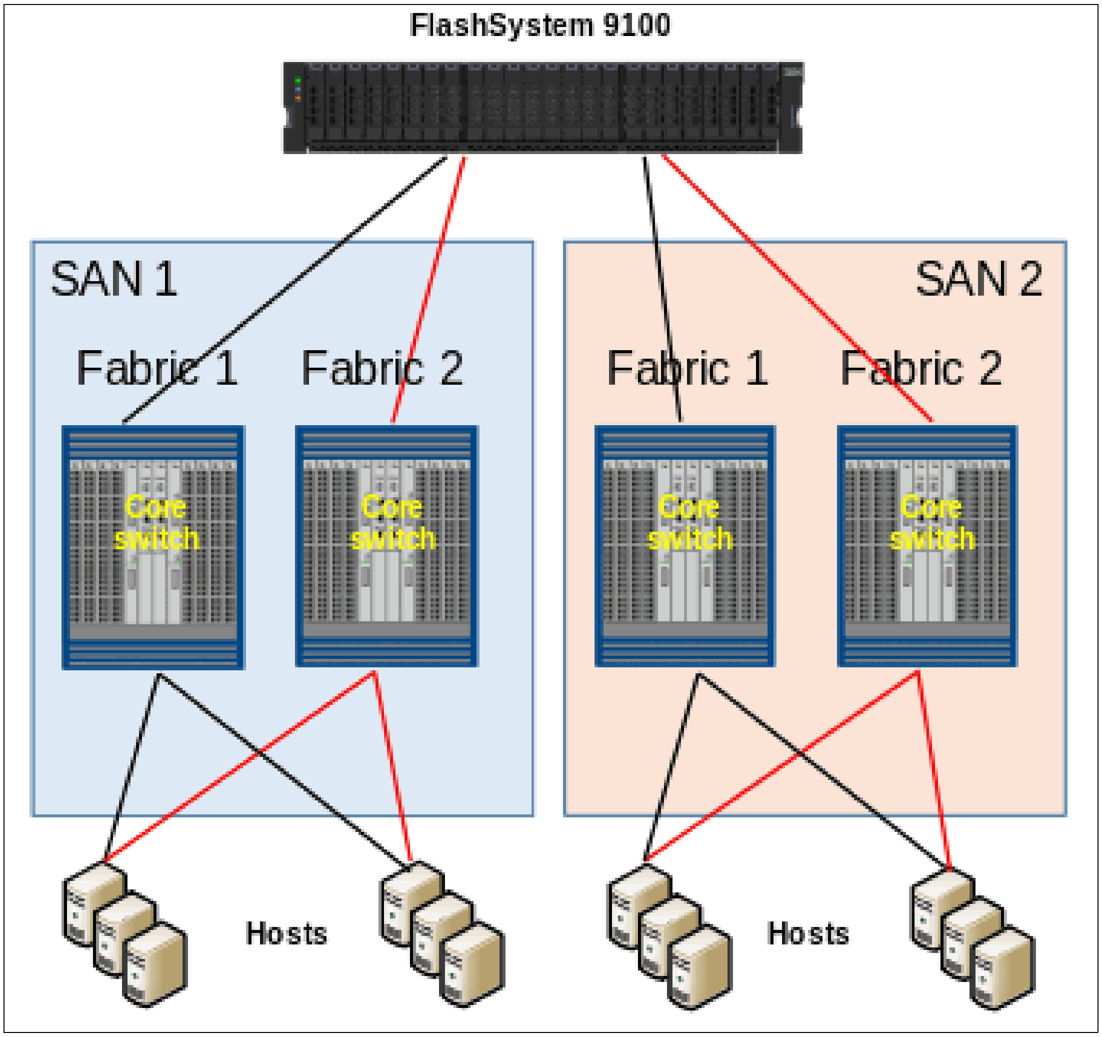


Figure 2-6 FlashSystem as a SAN bridge

Notice in Figure 2-6 that both SANs (Blue and Pink) are isolated and there is no communication through ISLs. When connected to both fabrics, FlashSystem is able to serve hosts and virtualize storages from either fabrics. They can provide disks to hosts from both SAN and when it has external virtualized storage, it can provide disks from storage on the Pink SAN (right), for example, to hosts on blue SAN (left).

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.

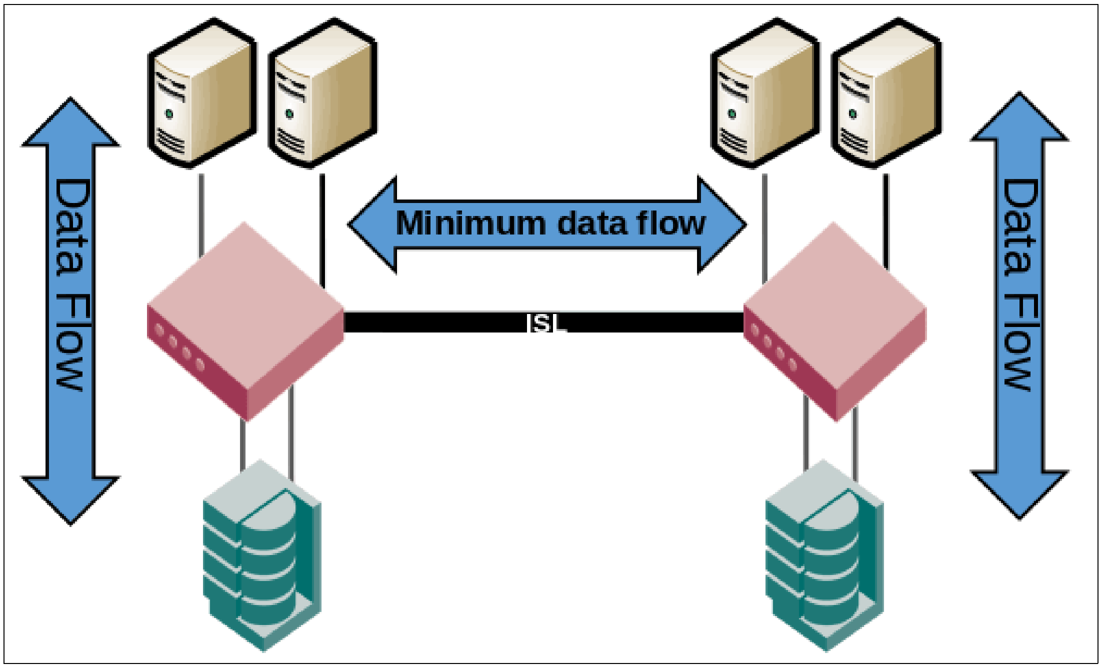


Figure 2-7 Storage and hosts attached to the same SAN switch

This solution can fit perfectly in small and medium SANs. However, it is not as scalable as other topologies available. As stated in 2.2.3, “Edge-core-edge topology” on page 17, the most scalable SAN topology is the edge-core-edge. Besides scalability, this topology provides different resources to isolate the traffic and reduce possible SAN bottlenecks.

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

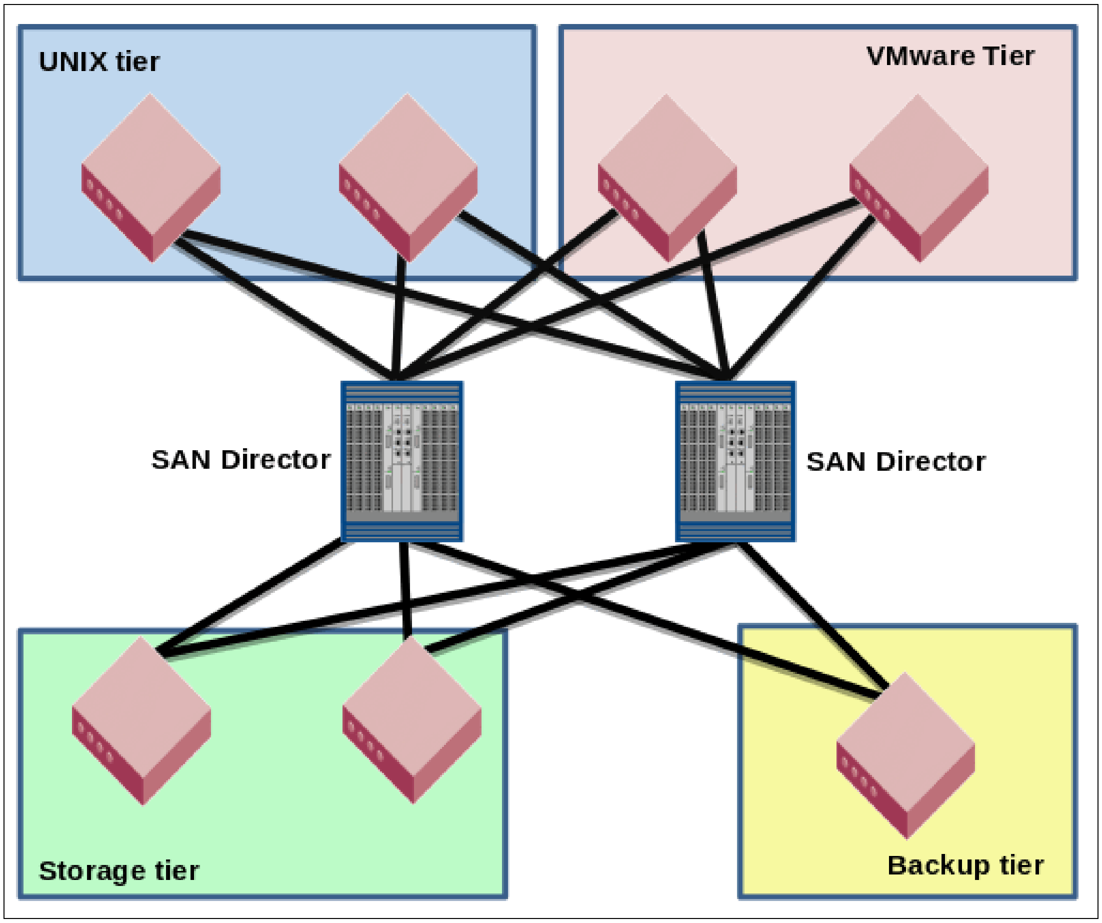


Figure 2-8 Segregation using edge-core-edge

Even when sharing the same core switches, it is possible to use virtual switches (see 2.2.7, “SAN partitioning” 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 and so the same guidelines and practices that apply to physical switches apply also to the virtual ones.

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

For a description of the IBM FlashSystem business continuity solutions, see Appendix A, “Business continuity” on page 401.

2.3 IBM FlashSystem 9100 controller ports

IBM FlashSystem 9100 hardware has significantly increased port connectivity options. Models AF7 and AF8 deliver up to 12x 16 Gb FC ports per node canister as shown in Table 2-1.

Table 2-1 FlashSystem 9100

Feature	FlashSystem 9100
Fibre Channel HBA	3x Quad 16 Gb
Ethernet I/O	3x Dual 25Gb iWARP for iSCSI or iSER 3x Dual 25Gb RoCE for iSCSI or iSER
Built in ports	4x 10 Gb for iSCSI
SAS expansion ports	1x Quad 12 Gb SAS (2 ports active)

Note: FlashSystem 9100 node canisters have 3 PCIe slots which you can combine the cards as needed. If expansions will be used, one of the slots must have the SAS expansion card. Then 2 ports will be left for fiber channel HBA cards, iWARP or RoCE ethernet cards. For more information see `IBM Knowledge Center.`

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 9100 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 9100 node canister.

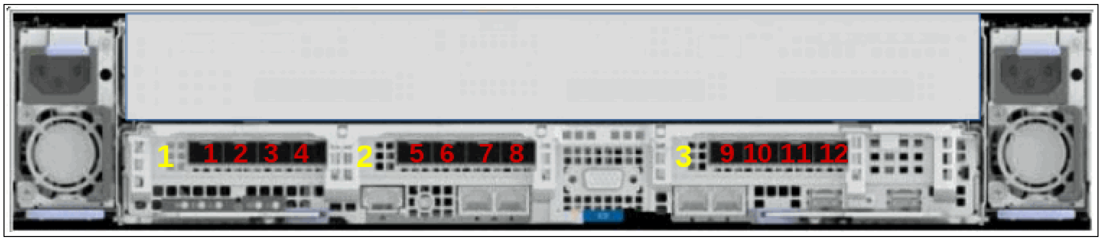


Figure 2-9 Port location in FlashSystem 9100 rear view

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 that are named, for example, `PROD_SAN_1` and `PROD_SAN_2`, or `PROD_SAN_A` and `PROD_SAN_B`. This type of naming convention is used to simplify the 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.

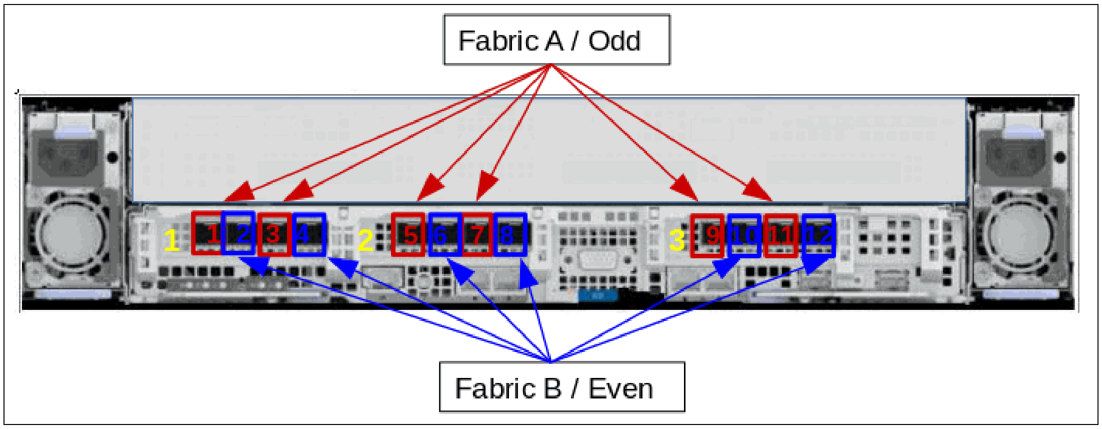
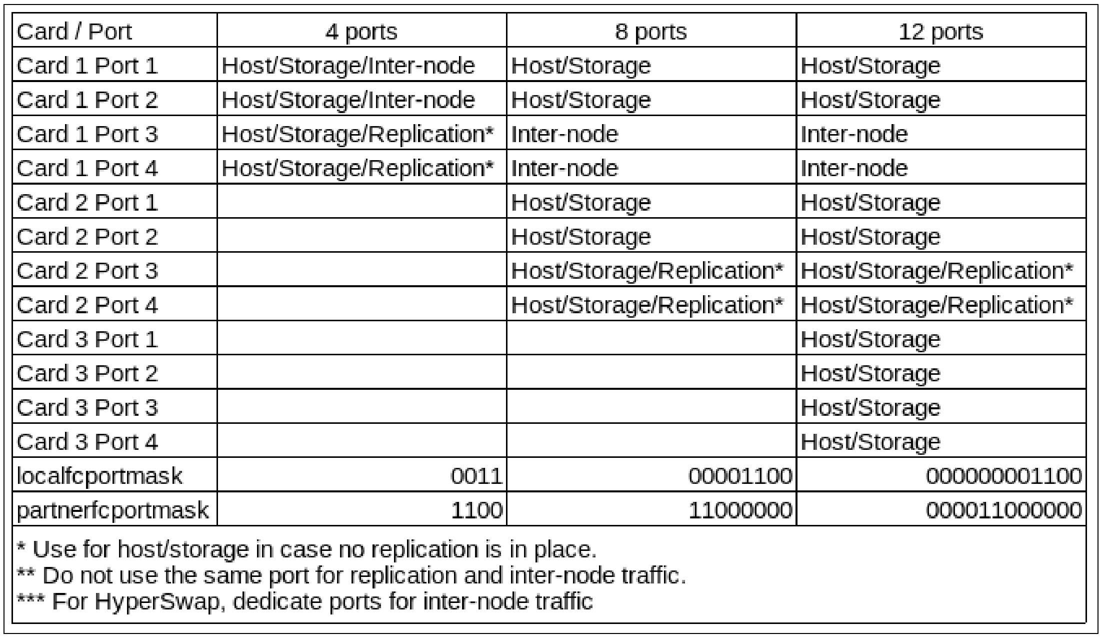


Figure 2-10 FlashSystem 9100 port distribution

As a preferred practice, assign specific uses to specific FlashSystem 9100 ports. This technique helps optimize the port utilization by aligning the internal allocation of hardware CPU cores and software I/O threads to those ports.

Figure 2-11 shows the specific port use guidelines for the FlashSystem 9100.



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 9100

Note: If you are using a FlashSystem 9100 system with a single I/O group the system will keep `localfcportmask` as 111111111111 and will not allow you to change this. However there is no reason to be concerned as the inter-node traffic will happen on the internal PCI midplane link. The portmasking recommendations above applies to systems with more than one I/O group.

Note 2: Depending on the workload and/or number of I/O groups, you can reserve ports 1 and 2 from card 3 for inter-node traffic. In this case you will have four ports for inter-node traffic, and the `localfcportmask` will be **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 9100 has a predefined number of buffer credits. The amount of buffer credits determines the available throughput over distances as follows:

- 4-port 16 Gbps adapters have 40 credits available per port, saturating links at up to 5 km at 16 Gbps

Switch port buffer credit: For stretched cluster and IBM HyperSwap configurations not using ISLs for the internode communication, it is advised to set the switch port buffer credits to match the IBM FlashSystem 9100 port.

Performance recommendation: Balance your bandwidth and make sure you have enough incoming bandwidth to saturate your backend bandwidth. Follow the simple guidance on where to plug them in, and try to balance your io across the ports.

2.4 Zoning

In this topic we will be covering the zoning recommendations for FlashSystem 9100. For external storage virtualization zoning information, please see *IBM System Storage SAN Volume Controller and Storwize V7000 Best Practices and Performance Guidelines*, SG24-7521 as the recommendations are the same.

Important: Errors that are caused by improper FlashSystem 9100 zoning are often difficult to isolate and the steps to fix them can impact the SAN environment. Therefore, create your zoning configuration carefully.

The initial configuration for FlashSystem 9100 requires the following different zones:

- Internode and intra-cluster zones
- Replication zones (if using replication)
- Host to FlashSystem 9100 zoning

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

> **Note:** Although internode/intra-cluster zone is not necessary for non clustered FlashSystem 9100 family, it is generally preferred to have one.

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 that they are zoned to. The singe initiator approach removes the possibility of a a misbehaving initiator affecting other initiators.

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

Cisco and 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.

You can read more about Smart Zoning here:

`https://ibm.biz/Bdjuu2`

Other relevant implementation information can be found here:

`https://ibm.biz/Bdjuuq`

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 intracluster, back-end, and replication zoning, use traditional zoning instead.

Simple zone for small environments

As an option for small environments, IBM Spectrum Virtualize based systems support 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, like the IBM Spectrum Virtualize based system ports which are target + initiator.

So, a valid zone can be 32 host ports plus 8 IBM Spectrum Virtualize based system ports. Include exactly 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.

2.4.2 Pre-zoning tips and shortcuts

Several tips and shortcuts are available for FlashSystem 9100 zoning.

Naming convention and zoning scheme

When you create and have to maintain a FlashSystem 9100 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 9100 naming convention, see 9.1.1, “Naming conventions” on page 321.

Aliases

Use zoning aliases when you create your FlashSystem 9100 zones if they are available on your particular type of SAN switch. Zoning aliases makes your zoning easier to configure and understand, and causes fewer possibilities for errors, see Table 2-2.

Table 2-2 Alias names examples

Port/WWPN	Use	Alias
Card 1 Port 1 physical WWPN	External Storage back-end	FS9100_N1P1_STORAGE
Card 1 Port 1 NPIV WWPN	Host attachment	FS9100_N1P1_HOST_NPIV
Card 1 Port 2 physical WWPN	External Storage back-end	FS9100_N1P2_STORAGE
Card 1 Port 2 NPIV WWPN	Host attachment	FS9100_N1P2_HOST_NPIV
Card 1 Port 3 physical WWPN	Inter-node traffic	FS9100_N1P3_CLUSTER
Card 1 Port 3 NPIV WWPN	No use	No alias
Card 1 Port 4 physical WWPN	Inter-node traffic	FS9100_N1P4_CLUSTER
Card 1 Port 4 NPIV WWPN	No use	No alias
Card 2 Port 3 physical WWPN	Replication traffic	FS9100_N1P7_REPLICATION
Card 2 Port 3 NPIV WWPN	No use	No alias
Card 2 Port 4 physical WWPN	Replication traffic	FS9100_N1P8_REPLICATION
Card 2 Port 4 NPIV WWPN	No use	No alias

Note: In the example above just some ports were 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 above, the N stands for node, and all examples are from node 1. An N2 example is *FS9100_N2P4_CLUSTER*.

One approach is to create template zones from the host to the FlashSystem 9100, The zoning should contain one alias from the host, and this alias must contain one initiator, and one alias from each node canister from the FlashSystem 9100, preferably the same port. As an example, create the following zone aliases:

- One zone alias for each FlashSystem 9100 port
- One alias for each host initiator
- Zone one host initiator alias to FlashSystem 9100 port 1 from node 1, and port 1from node 2. Then name this zone *HOST1_HBA1_T1_FS9100*
- If you are zoning a second host, zone one host initiator alias to FlashSystem 9100 port 3 from node 1 and port 3 from node 2. Then name this zone *HOST2_HBA1_T2_FS9100*

This way you keep a the number of paths on the host side to four for each volume and a good workload balance among the FlashSystem 9100 ports. Following these templates you would have the aliases distributed as in Table 2-3.

Table 2-3 Template examples.

Template	FS9100 ports on Fabric A	FS9100 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

Template	FS9100 ports on Fabric A	FS9100 ports on Fabric B
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 will vary depending on how many fiber ports you have in your system, and how many are dedicated to host access.The port numbers are examples, you can use different ports depending on how many HBA cards you have.Plan accordingly.

2.4.3 IBM FlashSystem 9100 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 FlashSystem 9100, internode communication primarily take place through the internal PCI connectivity between the two canisters of a control enclosure. However for the clustered FlashSystem 9100, the internode communication requirements are very similar to the SAN Volume Controller ones.

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

For FlashSystem 9100 clusters with two I/O groups or more with eight ports, we recommend to isolate the intracluster traffic by dedicating node ports specifically to inter-node communication. The ports to be used for intracluster communication varies according to the port count. See Figure 2-11 on page 23 for port assignment recommendations.

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

Only 16 port logins are allowed from one node to any other node in a SAN fabric. Ensure that you apply the proper port masking to restrict the number of port logins. In FlashSystem 9100 clusters with two or more I/O groups, without port masking any FlashSystem 9100 port and any member of the same zone can be used for intracluster communication, even the port members of FlashSystem 9100 to host and external virtualized back-ends.

2.4.4 IBM FlashSystem 9100 host zones

The preferred practice to connect a host into a FlashSystem 9100 is creating a single zone to each host port. This zone must contain the host port and *one* port from each FlashSystem 9100 node canister that the host must access as shown in Figure 2-12.

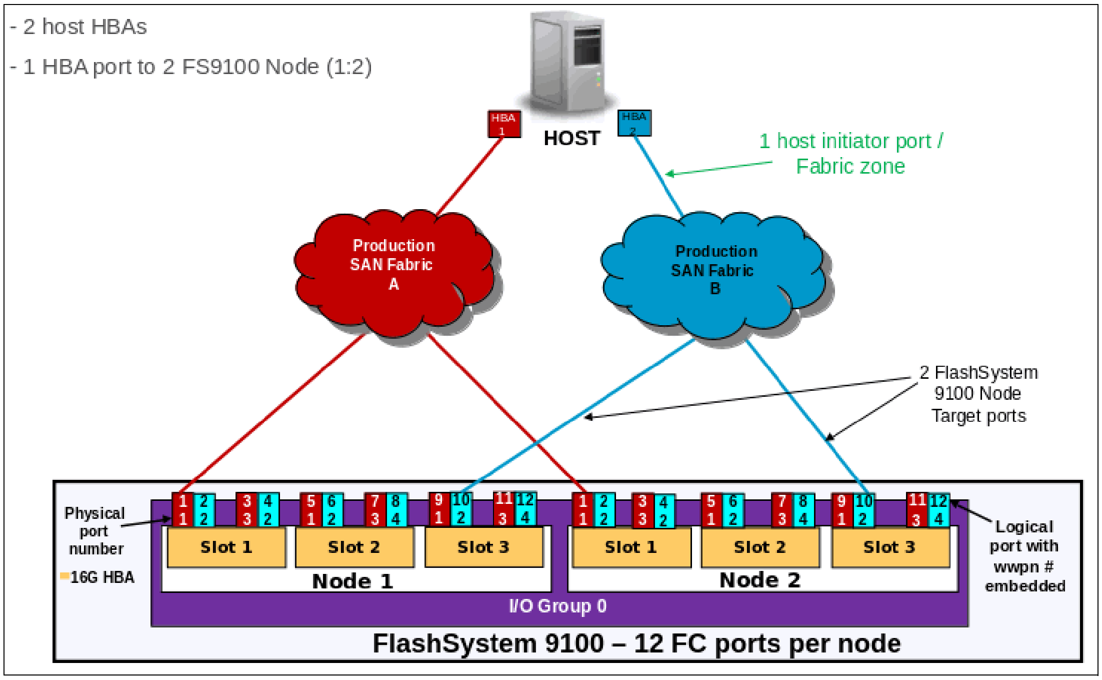

- 2 host HBAs
- 1 HBA port to 2 FS9100 Node (1:2)
HBA 1
HOST
HBA 2
1 host initiator port / Fabric zone
Production SAN Fabric A
Production SAN Fabric B
2 FlashSystem 9100 Node Target ports
Physical port number
16G HBA
Slot 1
Slot 2
Slot 3
Node 1
Slot 1
Slot 2
Slot 3
Node 2
Logical port with wwpn # embedded
I/O Group 0
FlashSystem 9100 – 12 FC ports per node

Figure 2-12 Typical host to FlashSystem 9100 zoning

This configuration provides four paths to each volume, being two preferred paths (one per fabric) and two non-preferred paths. Four paths is the number of paths, per volume, for which multipathing software such as AIXPCM, SDDDSM, NMP and the FlashSystem 9100, are optimized to work with.

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

`V8.2.1.x Configuration Limits and Restrictions for IBM FlashSystem 9100 family`

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

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

To obtain the best overall performance of the system and to prevent overloading, the workload to each FlashSystem 9100 port must be equal. Having the same amount of workload typically involves zoning approximately the same number of host FC ports to each FlashSystem 9100 FC port.

Hosts with four or more host bus adapters

If you have four HBAs in your host instead of two HBAs, more planning is required. Because eight paths is not an optimum number, configure your FlashSystem 9100 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 9100 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.

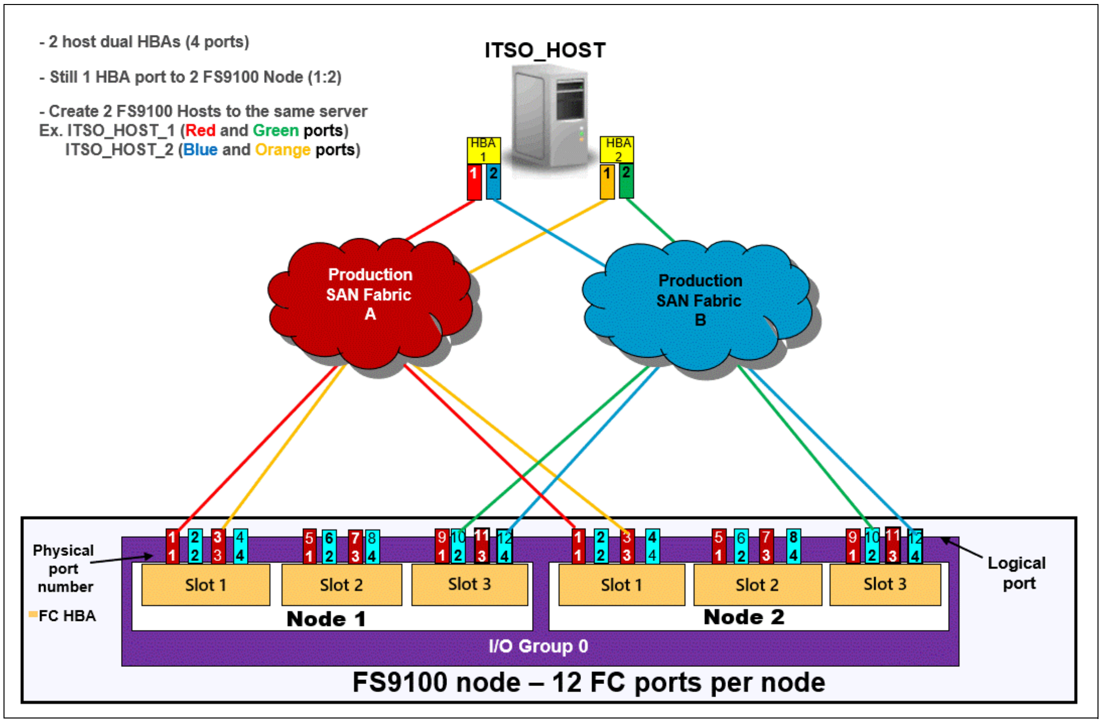

- 2 host dual HBAs (4 ports)
- Still 1 HBA port to 2 FS9100 Node (1:2)
- Create 2 FS9100 Hosts to the same server
Ex. ITSO_HOST_1 (Red and Green ports)
ITSO_HOST_2 (Blue and Orange ports)
ITSO_HOST
HBA 1
HBA 2
1
2
Production SAN Fabric A
Production SAN Fabric B
Physical port number
FC HBA
Logical port
Slot 1
Slot 2
Slot 3
Node 1
Node 2
I/O Group 0
FS9100 node – 12 FC ports per node

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 9100. 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.

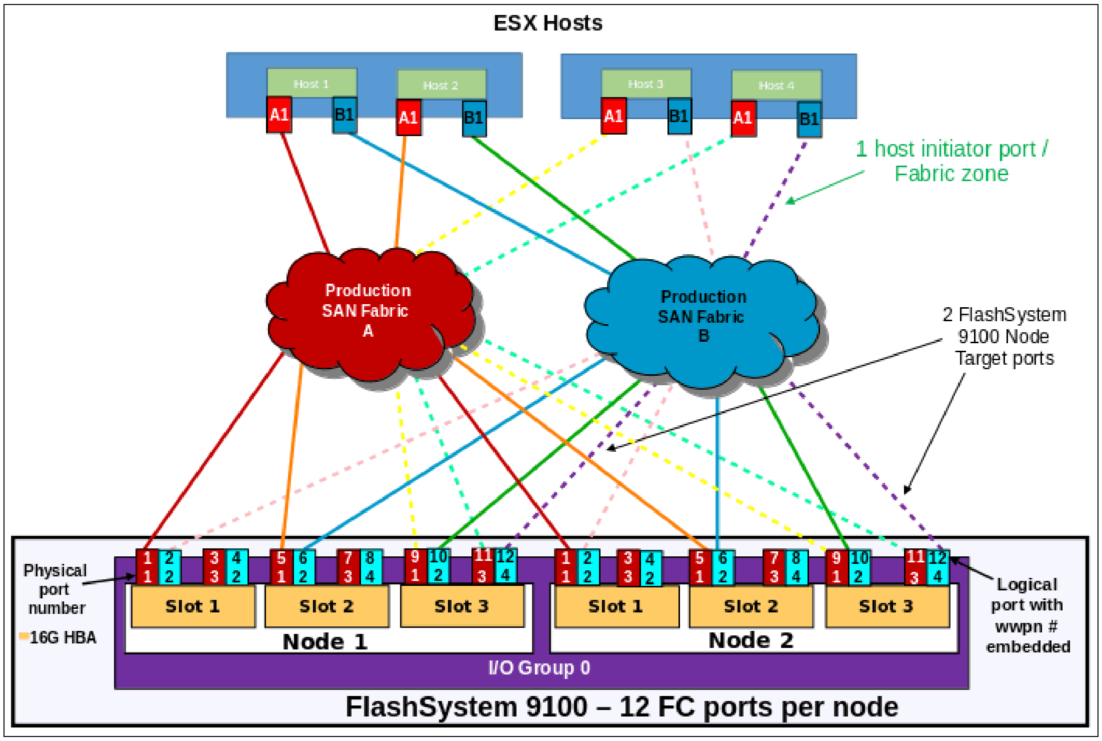


Figure 2-14 ESX Cluster zoning

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 9100 node ports and I/O Groups.
- Create one host entity for each host node in FlashSystem 9100 and group them in a *hostcluster* entity.
- Create separate zones for each host node in FlashSystem 9100 and on the ESX cluster.

When allocating a LUN/volume to a clustered system, it is highly recommended to use host cluster on FlashSystem 9100, this way you will have your hosts with the same scsi id for every volume, which will avoid outages due to scsi mismatch.

AIX VIOs: LPM zoning

When zoning IBM AIX® VIOs to IBM FlashSystem 9100, 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 both the active and inactive LPAR WWPNs in either the same zone or same IBM FlashSystem 9100 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/Storwize 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.

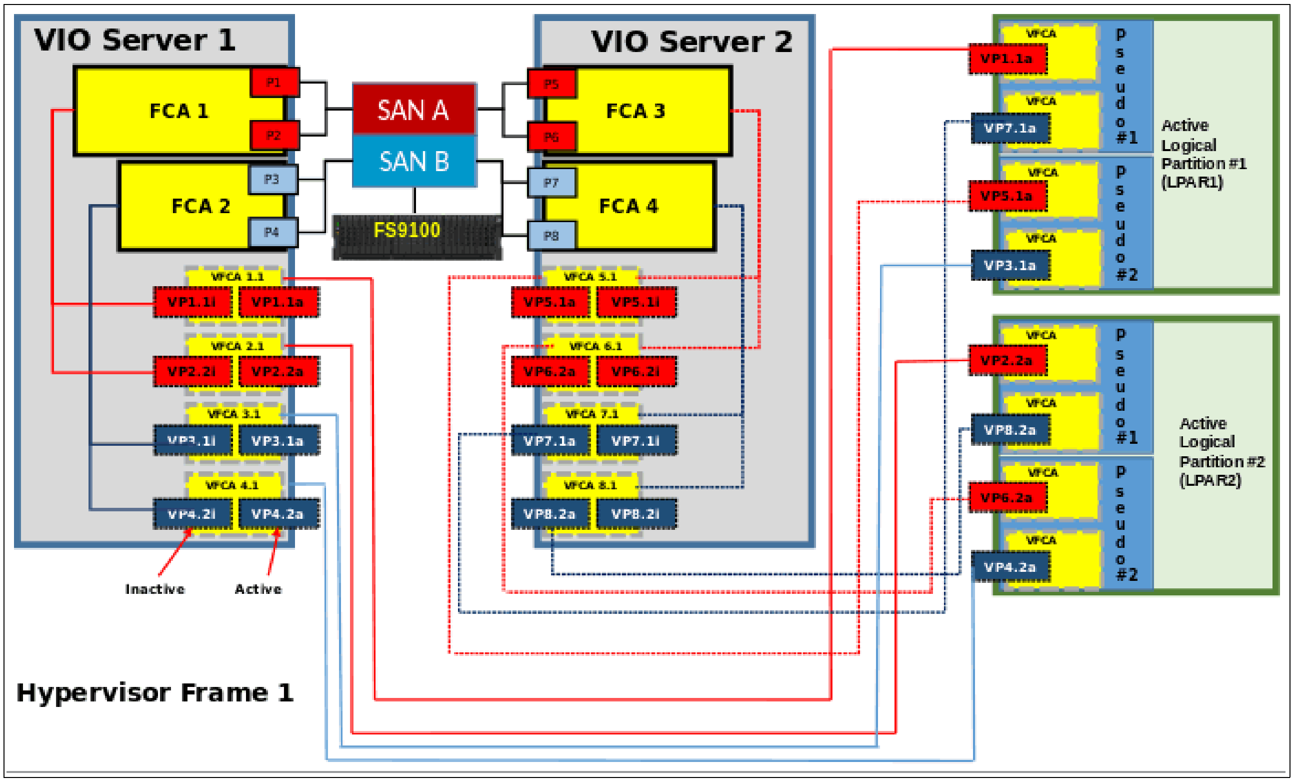


Figure 2-15 LPARs SAN connections

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.

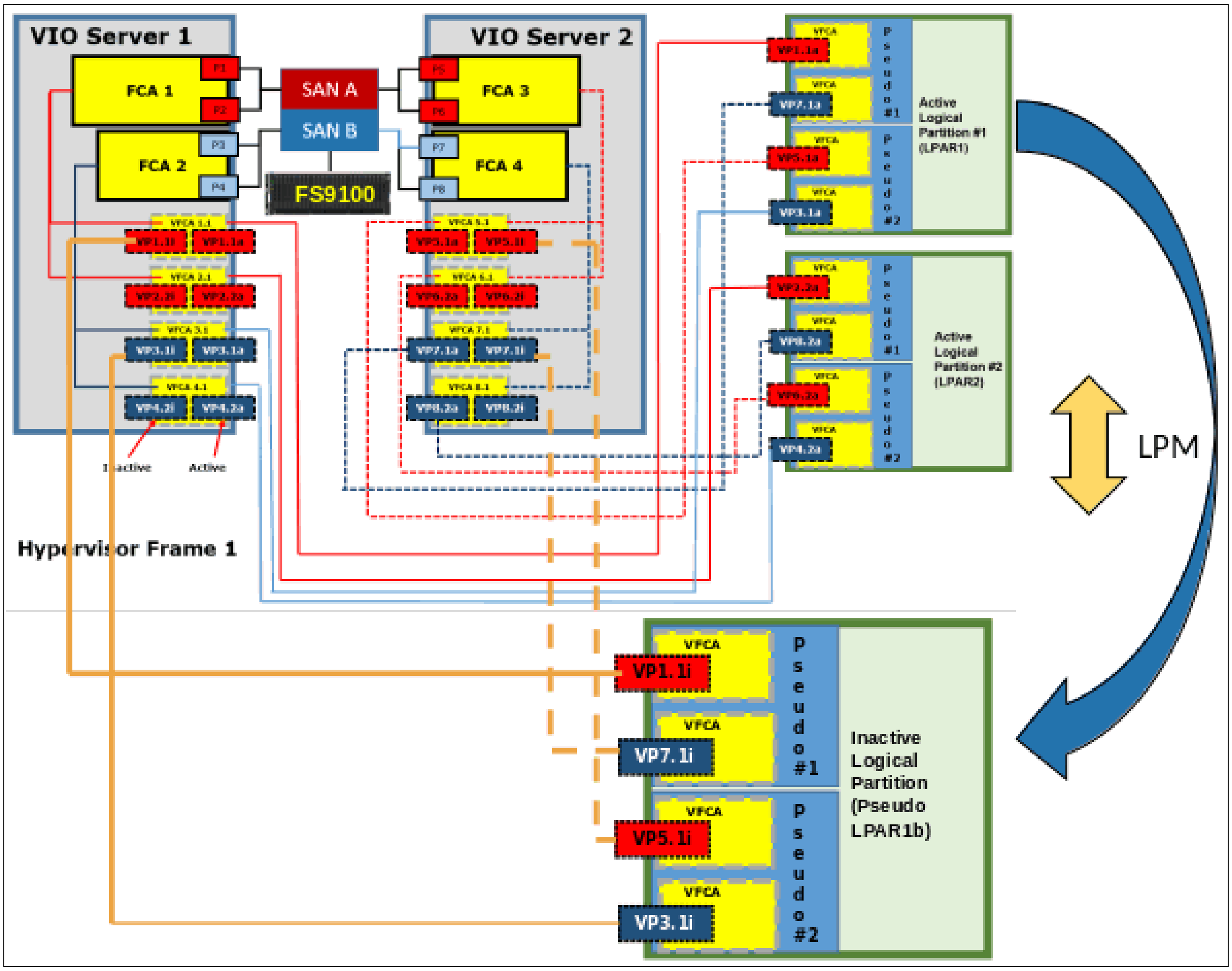


Figure 2-16 Live partition migration

Note: During LPM, the number of paths doubles from 4 to 8. Starting with eight paths per LUN/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 choices are available:

- Optical multiplexors, such as Dense Wavelength Division Multiplexing (DWDM) or Coarse Wavelength Division Multiplexing (CWDM) devices
- Long-distance SFPs and XFPs
- FC-to-IP conversion boxes
- Native IP-based replication with 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 9100 has no allegiance to a particular model of optical multiplexor.

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

2.5.2 Long-distance SFPs or XFPs

Long-distance optical transceivers have the advantage of extreme simplicity. Although no expensive equipment is required, a few configuration steps are necessary. Ensure that you use transceivers that are designed for your particular SAN switch *only*. Each switch vendor supports only a specific set of SFP or XFP transceivers, so it is unlikely that Cisco SFPs will work in an 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 either through stand-alone devices (Multiprotocol routers) or using blades integrated in the director class product. FlashSystem 9100 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 e-mail, do not assume that this type of traffic is always the case. The design of FC is sensitive to congestion and you do not want a spyware problem or a DDOS attack on an IP network to disrupt your FlashSystem 9100.

Also, when you are communicating with your organization’s networking architects, distinguish between megabytes per second (MBps) and megabits per second (Mbps). In the storage world, bandwidth often is specified in MBps, but network engineers specify bandwidth in Mbps. If you fail to specify MB, you can end up with an impressive-sounding 155 Mbps OC-3 link, which supplies only 15 MBps or so to your FlashSystem 9100. If you include the safety margins, this link is not as fast as you might hope, so ensure that the terminology is correct.

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

- For redundancy purposes use as many TCP/IP links between sites as you have fabrics in each site that you want to connect. In most cases, there are two SAN FC fabrics in each site, so you need two TCP/IP connections between sites.
- Try to dedicate TCP/IP links only for storage interconnection. Separate them from other LAN/WAN traffic.
- Make sure that you have a service level agreement (SLA) with your TCP/IP link vendor that meets your needs and expectations.
- If you do not use Global Mirror with Change Volumes (GMCV), make sure that you have sized your TCP/IP link to sustain peak workloads.
- The use of FlashSystem 9100 internal Global Mirror (GM) simulation options can help you test your applications before production implementation. You can simulate the GM environment within one FlashSystem 9100 system without partnership with another. Use the **chsystem** command with the following parameters to perform GM testing:

  – **gminterdelaysimulation**
  – **gmintradelaysimulation**

  Further details on GM planning are described in Chapter 6, “Copy services” on page 141.
- 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 9100 to a value *lower* than your TCP/IP link capacity. Failing to do that can cause an unstable TCP/IP tunnel, which can lead to stopping all your remote copy relations that use that tunnel.
- The best case is to use GMCV when replication is done over long distances.
- Use compression on corresponding FCIP devices.
- Use at least two ISLs from your local FC switch to local FCIP router.
- On a Brocade SAN, use the Integrated Routing feature to avoid merging fabrics from both sites.

For more information about FCIP, see the following publications:

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

2.5.4 SAN extension with Business Continuity configurations

FlashSystem 9100 HyperSwap technology provide Business Continuity solutions over metropolitan areas with distances up to 300 km. Usually this is achieved using SAN extension over WDM technology. Furthermore, in order to avoid single points of failure, multiple WDMs and physical links are implemented. When implementing these solutions, particular attention must be paid in the intercluster connectivity set up.

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

There are basically two possible configurations to interconnect the Private SANs. In the Configuration 1, shown in Figure 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.

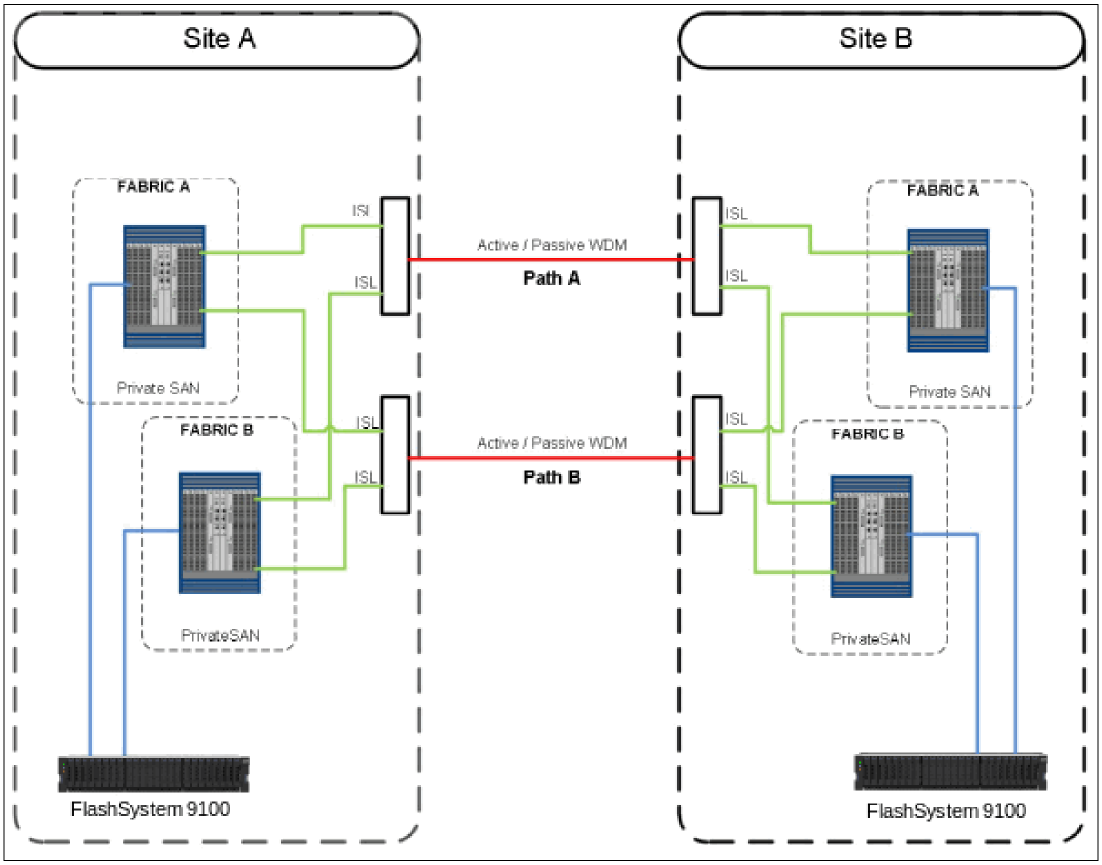


Figure 2-17 Configuration 1: physical paths shared among the fabrics

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.

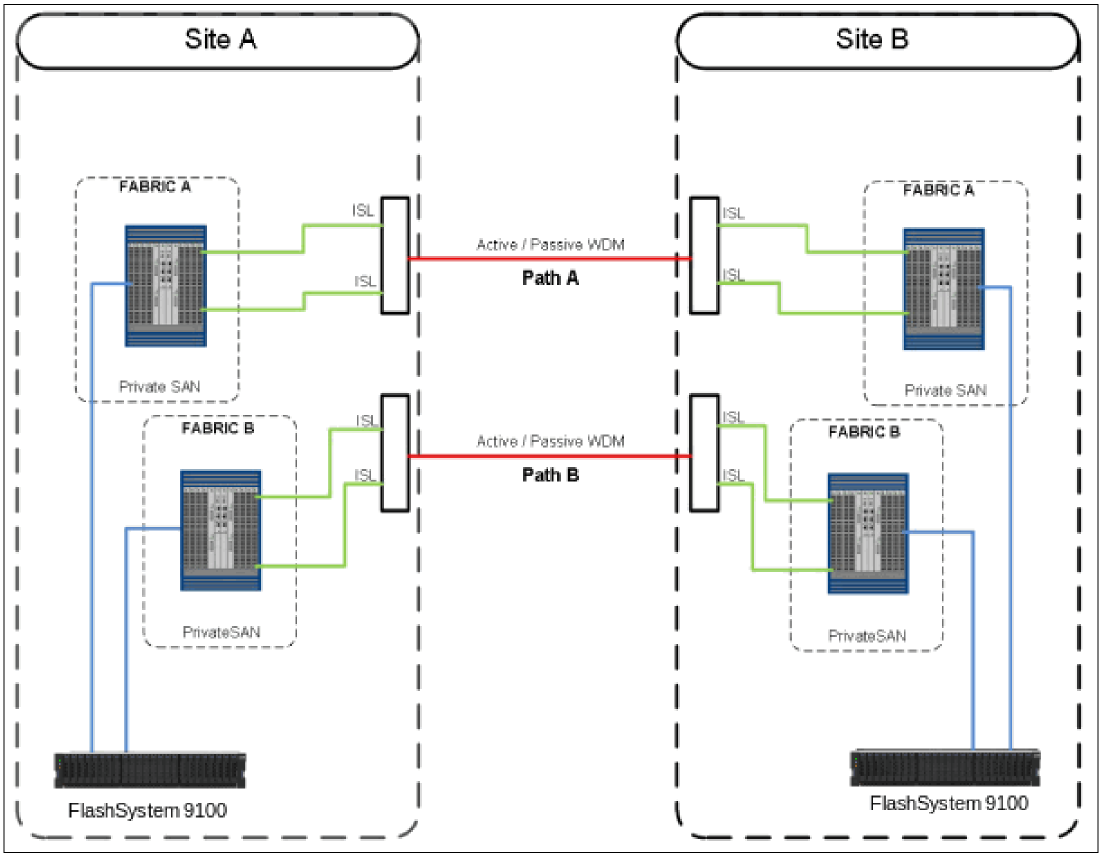


Figure 2-18 Configuration 2: physical paths not shared among the fabrics

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 intracluster 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 intracluster communication would be guaranteed through the unaffected fabric.

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

2.5.5 Native IP replication

To enable native IP replication, FlashSystem 9100 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 9100. *Native* means the FlashSystem 9100 does not need any FCIP routers to create a replication partnership. This partnership is based on the Internet Protocol network (IP) as opposed to the Fibre Channel (FC) network.

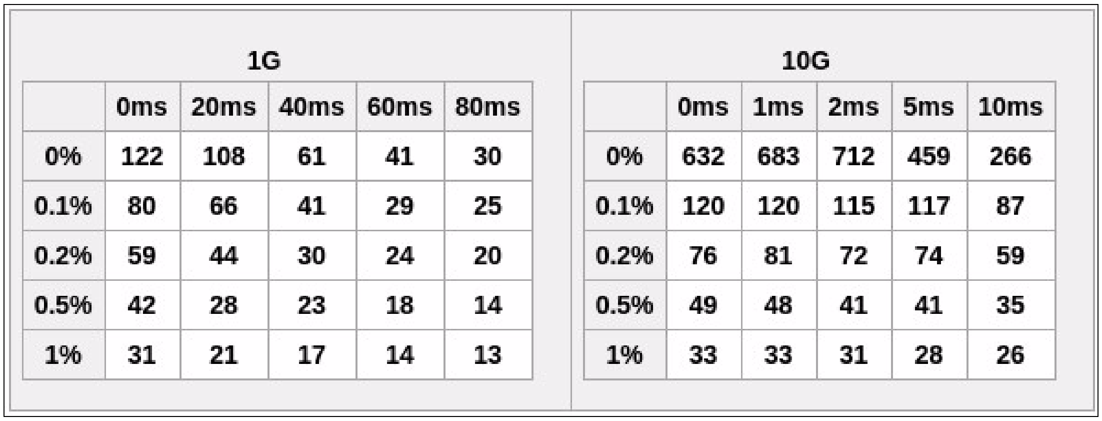
Native IP replication was added to the IBM SAN Volume Controller and Storwize Family in the V7.2 release of microcode via the implementation of the Bridgeworks SANSlide network optimization technology. Starting in the V7.7 release, further enhancements were made with the introduction of replication compression.

The main design point for the initial SANSlide implementation as well as subsequent enhancements including the addition of replication compression is to reduce link utilization in order to allow the links to run closer to their respective line speed at distance and over poor quality links. IP replication compression will 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 there are alternative technologies that should be considered such as FCIP where compression is done in the tunnel and will often yield an increase in effective bandwidth of 2:1 or more.

It’s 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 MB/s. 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

Note 1: The maximum bandwidth for a typical IP replication configuration consisting of two 1 Gb links is approximately 244 MBps at zero latency and zero packet loss.

Note 2: When using two links replication will perform 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 will be 160 MBps.

Note 3: 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.

Note 4: The FlashSystem 9100 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 differs from some storage systems such as the XIV and A9000 where replication runs below volume compression and therefore replicate the compressed capacity of the volumes. This difference needs to be taken into account when sizing workloads that are moving from one storage system technology to another.

For more information regarding native IP replication, see *Implementing the IBM Storwize V7000 with IBM Spectrum Virtualize V8.2.1*, Section 11.8.

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 9100 SAN. However, do not put tape and disk traffic on the same FC HBA.

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

2.7 Switch interoperability

FlashSystem 9100 is flexible as far as switch vendors are concerned. All of the node canister connections on a particular FlashSystem 9100 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 9100 supports some combinations of SANs that are made up of switches from multiple vendors in the same SAN. However, this approach is not preferred in practice. Despite years of effort, interoperability among switch vendors is less than ideal because FC standards are not rigorously enforced. Interoperability problems between switch vendors are notoriously difficult and disruptive to isolate. Also, it can take a long time to obtain a fix. For these reasons, run only multiple switch vendors in the same SAN long enough to migrate from one vendor to another vendor, if this setup is possible with your hardware.

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

Interoperability between Cisco switches and Brocade switches 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

Drives and arrays

This chapter describes the aspects and practices to consider when FlashSystem internal storage is planed and managed. Internal storage consists of NVMe FlashCore modules, NVMe drives or SAS flash drives expansion enclosures.

FlashSystem supports attachment and virtualization of external back-end storage devices. This aspect was left out of the scope of this book intentionally. For best practices for external storage, see the following publication: *IBM System Storage SAN Volume Controller and Storwize V7000 Best Practices and Performance Guidelines*, SG24-7521.

> **Important, IBM FlashSystem 9200:** On 11th February 2020 IBM announced the arrival of the IBM FlashSystem 9200 to the family. This book was written specifically for IBM FlashSystem 9100, however most of the general principles will apply to the IBM FlashSystem 9200.
>
> If you are in any doubt as to their applicability to the FlashSystem 9200 then you should work with your local IBM representative.
>
> This book will be updated to include FlashSystem 9200 in due course.
>
> The Flashsystem 9200 product guide is available at:
>
> IBM FlashSystem 9200 Product Guide

This chapter contains the following sections:

- Modules and drives
- Flash arrays
- Quorum disks

3.1 Modules and drives

IBM FlashSystem can contain three types of internal storage: NVMe FlashCore Modules (FCM), NVMe flash drives, and SAS flash drives. The system is all-flash, spinning drives are not supported and can't be installed to its control or expansion enclosures. However, FS9100 can be clustered with Storwize V7000 family machines, which support them.

3.1.1 NVMe storage

FlashSystem 9100 control enclosure has 24 x 2.5" slots to populate with NVMe storage.

NVMe protocol

NVM Express (NVMe) is an optimized, high-performance scalable host controller interface designed to address the needs of systems that utilize PCI Express-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 PCI Express (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 legacy SAS and SATA protocols that use only a single queue.

According to Fibre Channel Industry Association (FCIA) publications as a result of the simplicity, parallelism and efficiently of NVMe, it delivers significant performance gains versus SCSI, such as:

- For 100% random reads, NVMe has 3x better IOPS than 12 Gbps SAS
- For 100% random writes, NVMe has 1.5x better IOPS than 12 Gbps SAS
- For 100% sequential reads: NVMe has 2x higher throughput than 12 Gbps SAS
- For 100% sequential writes: NVMe has 2.5x higher throughput than 12 Gbps SAS

In addition to just IOPS and throughput, the efficiencies of the command structure of NVMe also reduce CPU cycles by half, as well as reduce latency by more than 200 microseconds compared to 12 Gbps SAS.

Industry-standard NVMe drives

FS9100 provides an option to use self-encrypting industry-standard NVMe flash drives, which are available with capacity from 1.92 TB to 15.36 TB.

Supported NVMe SSD drives are built in 2.5-inch form factor (SFF) and use a dual-port PCIe Gen3 x4 interface to connect to the midplane.

NVMe FlashCore modules

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

The current generation of FCMs is built with 64-layer 3D TLC NAND flash technology and Everspin MRAM cache in 2.5-inch U.2 form factor drive.

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;
- IBM Advanced Flash Management, which improves 3D TLC flash endurance over standard implementations without sacrificing latency;
- Cognitive algorithms for wear leveling and heat segregation.

Variable stripe 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 levelling understands the health of blocks within the chips and tries to place "hot" data within the most healthy 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 ECC (Error Correcting Code) algorithm. If an error cannot be corrected, then the FlashSystem DRAID layer will be used to rebuild the data.

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

NVMe Flashcore Modules use inline hardware compression to reduce the amount of physical space required. Compression can't 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.

FCMs are available in 4.8, 9.6 and 19.2TB sizes, which is their physical capacity, or usable capacity. 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 could 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 contain a fixed amount of space for metadata. The maximum effective capacity is the amount of data it takes to fill the metadata space.

Module capacities are shown in Table 3-1.

Table 3-1 FS9100 FlashCore module capacities

Usable capacity	Compression ratio at maximum eff. 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

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

Usable and effective capacities as discussed later in this chapter, see 3.1.3, "Internal storage considerations" on page 44.

3.1.2 SAS-attached drives

IBM FlashSystem supports attachment of SFF expansion enclosure with 24 slots for 2.5-inch SAS flash drives and LFF high density expansion enclosure with 92 slots for SAS flash drives in a 3.5-inch carrier.

A single FlashSystem 9100 control enclosure can support up to twenty FlashSystem 9100 SFF expansion enclosures with a maximum of 504 drives per system (including NVMe) or up to eight FlashSystem 9100 LFF HD expansion enclosures with a maximum of 760 drives per system.

With four-way system clustering, the size of the system can be increased to a maximum of 3,040 drives.

The 12 GB SAS expansion card needs to be installed in both nodes of a control enclosure to attach expansions. The card has four ports with only two ports active (ports 1 and 3). It can be used only to connect expansion enclosures, SAS host connection is not supported.

SAS flash drives are available with capacities of 1.92 TB, 3.84 TB, 7.68 TB and 15.36 TB.

All SAS SSDs attached to FlashSystem 9100 are mapped to Easy Tier tier1, while NVMe flash drives and NVMe FCMs are mapped to tier0.

3.1.3 Internal storage considerations

The following practices should be considered when planning and managing FS9100 internal storage.

- SAS enclosures are used to scale capacity within the performance envelope of a single controller enclosure. Clustering (up to four control enclosures) scales performance with the additional NVMe storage.

  Intermixing of expansion enclosures in a system is supported: SFF and LFF HD expansion enclosures can be intermixed behind the SFF control enclosure. Expansion enclosures are designed to be dynamically added with virtually no downtime, helping to quickly and seamlessly respond to growing capacity demands.

  Drives of the same form factor and connector type can be intermixed within an expansion enclosure.
- NVMe and SAS drives in the same system can be mixed, but NVMe drives can only exist in the control enclosure, and SAS drives can only exist in SAS expansion enclosures.
- Industry-standard NVMe drives start at a smaller capacity point, allowing for a smaller system. Industry-standard NVMe drives can encrypt data but don’t compress it.
- Within a control enclosure, NVMe drives of different capacities can be intermixed, also industry-standard NVMe drives can be intermixed with FlashCore modules.

  However, within a DRAID array NVMe drives must all be the same size. Having mixed drive types and sizes would mean that you’d need to purchase the drives in blocks of six or more of each kind, and manage them as a few independent RAID arrays as opposed to one big one, which will cost more capacity needed for RAID overhead.
- FlashCore modules need to be formatted before they can be used. Format is important since we want members to have zero used capacity when an array is created. Drives will automatically format when being changed to candidate. While they are formatting, they will appear as offline candidates. If you attempt to create an array before format is complete, the create command is delayed until all formatting is done. Once this happens the command will complete.

If a drive fails to format, it should go offline with its format marked as degraded and will require a manual format to bring it back online. The CLI scenario is shown in Example 3-1.

Example 3-1 Manual FCM format

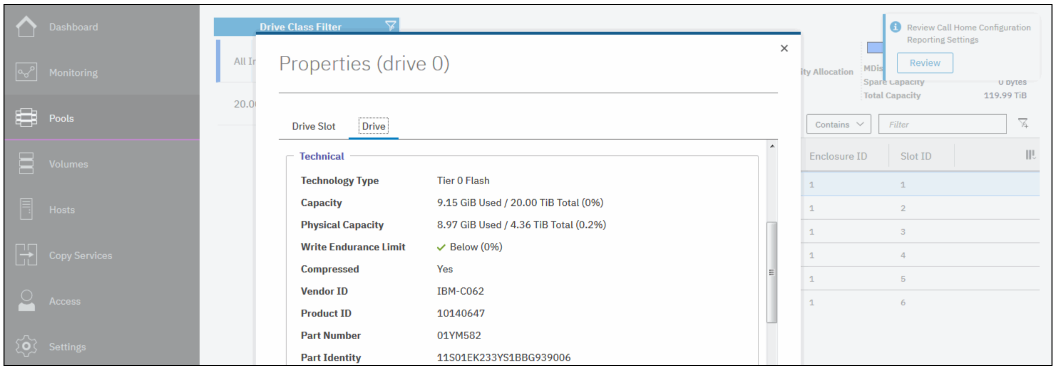
```
IBM_FlashSystem:FS9100-ITSO:superuser>lsdrive
id status  error_sequence_number use       tech_type ....
....
13 offline 118                   candidate tier0_flash ....
IBM_FlashSystem:FS9100-ITSO:superuser>chdrive -task format 13
IBM_FlashSystem:FS9100-ITSO:superuser>
```

An FCM is expected to format in under 70 seconds.

- When a drive is replaced in the same slot, the system tries to take the drive as a replacement for the array member. The drive will have to format first, so this might take some time for an FCM.

  If the drive was previously encrypted in another array, it will come up as failed since this system won't have the required keys. The drive must be manually formatted to make it a candidate.

- For NVMe drives and FCMs, formatting the drive completes a cryptographic erase of the drive.
- The FlashSystem 9100 system 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.



Properties (drive 0)

Technical

Technology Type	Tier 0 Flash
Capacity	9.15 GiB Used / 20.00 TiB Total (0%)
Physical Capacity	8.97 GiB Used / 4.36 TiB Total (0.2%)
Write Endurance Limit	✓ Below (0%)
Compressed	Yes
Vendor ID	IBM-C062
Product ID	10140647
Part Number	01YM582
Part Identity	11S01EK233YS1BBG939006

Spare Capacity	0 bytes
Total Capacity	119.99 TiB

Enclosure ID	Slot ID
1	1
1	2
1	3
1	4
1	5
1	6

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
...
compressed yes
physical_capacity 4.36TB
physical_used_capacity 8.97GB
effective_used_capacity 9.15GB
```

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 effective used capacity to physical used capacity. Here we have 9.15 / 8.97 = 1.02, so written data is compressed only 1.02:1 (nearly incompressible).

- Physical used capacity is expected to be nearly the same on all modules in one array.
- Plan and 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 be able 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 causing 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:** IBM has published a `flash alert` that warns against 100% of physical flash provisioning.

- When using FCMs, data compression ratios should be thoroughly planned and monitored.

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

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

  The module’s maximum effective capacity is 21.99 TB which equals to 20 TiB, and the module’s usable capacity is 4.8 TB = 4.36 TiB.

  After 20TiB of data is written, the module will be 100% full for the system. At the same time, data will use only 2 TiB of the physical space. 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 will fill up until the maximum physical capacity is reached.

  Example #2: 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 to 40 TiB, and the module’s usable capacity is 19.2 TB = 17.46 TiB.

  After 20 TiB is written, only 50% of effective capacity will be used. After 1.2 compression, it will occupy 16.7 TiB of physical capacity, making the module 95% full, and potentially impacting the module’s performance.

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

- When using SAS-attached Read Intensive (tier1) flash drives, the DWPD metric needs to be considered. DWPD stands for Drive Writes Per Day and shows drive write endurance.

  RI SSDs can support up to 1 DWPD, which means that full drive capacity can be written on it every day and it will last the 5 years warranty.

  ```
  Example: 3.84TB RI SAS drive is rated for 1 DWPD which means 3840000 MB of data
  may be written on it every day. Each day has 24x60x60 = 86400 seconds, so
  3840000/86400 = 44.4 MB/s of average daily write workload is required to reach
  1 DWPD.
  ```

  If the drive write workload is continuously higher than allowed for DWPD set for it, 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.

It is acceptable to see “write endurance usage rate is high” warnings during the initial phase of system implementation or during continuous testing. Afterwards the system’s workload will reduce to what is sized for it, and the system recalculates the usage rate and removes the warnings. Because the calculation is based on a long-run averages, it can take time (up to one month) for them to be cleared.

3.2 Flash arrays

In order to use internal IBM FlashSystem 9100 drives in storage pools and provision their capacity to hosts, they need to be joined into RAID arrays to form array-type MDisks.

3.2.1 Supported RAID types

RAID provides two key design goals:

- Increased data reliability
- Increased input/output (I/O) performance

The IBM FlashSystem 9100 system supports two RAID types: traditional RAID and Distributed RAID (DRAID).

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 RAID10), or is calculated from remaining data stripes and parity (for RAID5 or RAID6), and written to a single spare drive.

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. In the case of a drive failure, the rebuild load is distributed across multiple drives. By doing this, 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 sharing the workload.

NVMe FlashCore Modules installed in FlashSystem 9100 can be aggregated into DRAID6 and DRAID5 arrays, DRAID6 is recommended. Traditional RAID levels 5 and 6, as well as RAID 0, 1 and 10, are not supported on FCMs.

Industry-standard NVMe drives and SAS drives in expansion enclosures can be aggregated into DRAID6 and DRAID5 arrays, and also can form RAID 1 and RAID 10 arrays. TRAID 5 and 6 are not supported.

Table 3-2 summarizes the supported drives, array types, and RAID levels.

Table 3-2 Supported RAID levels

Supported drives	**Non-distributed arrays (traditional RAID)**					**Distributed arrays (DRAID)**	
	RAID 0	RAID 1	RAID 5	RAID 6	RAID 10	RAID 5	RAID 6
SAS flash drives	Yes	Yes	No	No	Yes	Yes	**Yes**
NVMe drives	Yes	Yes	No	No	Yes	Yes	**Yes**
FlashCore Modules	No	No	No	No	No	Yes	**Yes**

3.2.2 Array considerations

The following practices should be considered when planning and managing flash drive arrays in an FS9100 environment.

- DRAID6 is strongly recommended for all drive configurations. A minimum 6 drives or modules are required for it.

  The only exception from this rule applies to an array that has to be constructed out of four or five members. DRAID5 is the preferred option for it (and the only available option for an array of four to five FCMs).

  Traditional RAID levels 5 and 6 are not supported on FS9100, as DRAID is superior to them in all aspects.

  Even DRAID 5 requires less I/O and CPU overhead compared to level 6, this benefit is compensated by operation parallelism implemented with DRAID6. 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 compared to DRAID5, it is the only recommended setting for NVMe drives.

- When creating FCM arrays with the GUI, the system will stick to the recommended configuration, which is DRAID6. Other configurations are possible by using the CLI only.
- FlashSystem 9100 is optimized for arrays made of 16 to 24 NVMe devices, balancing performance, rebuild times and usable capacity. For SAS SSD drives in expansions, optimal array size is 24-32 drives per DRAID6 array.

  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 don’t 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 the number of DRAID arrays doesn’t matter. 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. Maximum system performance can be achieved from a single DRAID array.
- At the time of writing, it is not possible to add modules to an existing array. To expand a storage pool another array needs to be created and added to a pool.
- At the time of writing, only one DRAID rebuild area is allowed for a DRAID array of FlashCore modules.
- All NVMe drives in an array must have the same physical and logical capacity. It is not possible to replace a drive with a “superior” - one with a greater capacity - as it is allowed for SAS drives.
- Compressing (FCM) and non-compressing (industry-standard NVMe) drives can’t be mixed in an array.
- It is not recommended to mix NVMe devices from different control enclosures in a system into one array. For SAS drives it is allowed: one DRAID6 or TRAID 10 can span across multiple expansion enclosures.
- It is not recommended to mix arrays of FCMs and arrays of NVMe drives in a single storage pool.

- 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 stripe width will not make the array able to tolerate more failed drives. DRAID6 will not get more redundancy than determined for level 6, independently of the width of a single redundancy unit.

  Reduced width will increase capacity overhead, but also increase rebuild speed, as there is a smaller amount of data that RAID needs to read in order 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.

  Stripe width together with strip size (which is fixed to 256KiB for FCMs) determine Full Stride Write (FSW) size. With full stride write, there is no need to read existing data in a stride, so the RAID I/O penalty is massively reduced. FlashSystem 9100 cache is designed to perform FSW whenever possible, so in most scenarios the host will not get any noticeable difference in performance. However, for fine-tuning for maximum performance, stripe width can be adjusted to match host file system block size. For example, 1MiB host transfer size will get best results on 4+P+Q array (4 data disks x 256 KiB stripe size, array width = 6).

To calculate the approximate amount of usable space that is available when creating a 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) which is 1 for all FCM arrays
W - Stripe width

Example #1:
Capacity of DRAID6 array out of 16 x 9.6 TB FlashCore modules.
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
Example #2:
Capacity of DRAID6 array out of 6 x 4.8 TB FlashCore modules.
D = 4.8 TB = 4.36TiB
N = 6
S = 1
W = 5
Array capacity = 4.36 TiB / ((5 * 256) + 16) * ((6 - 1) * (5 - 2) * 256 =
                                           = 4.36 TiB / 1296 * 3840 = 12.8 TiB
```

Note: This formula gives only a close estimate. It is not intended to give exact result. For exact results, use capacity estimation tool or `lspotentialarraysize` CLI command.

3.2.3 Compressed array monitoring

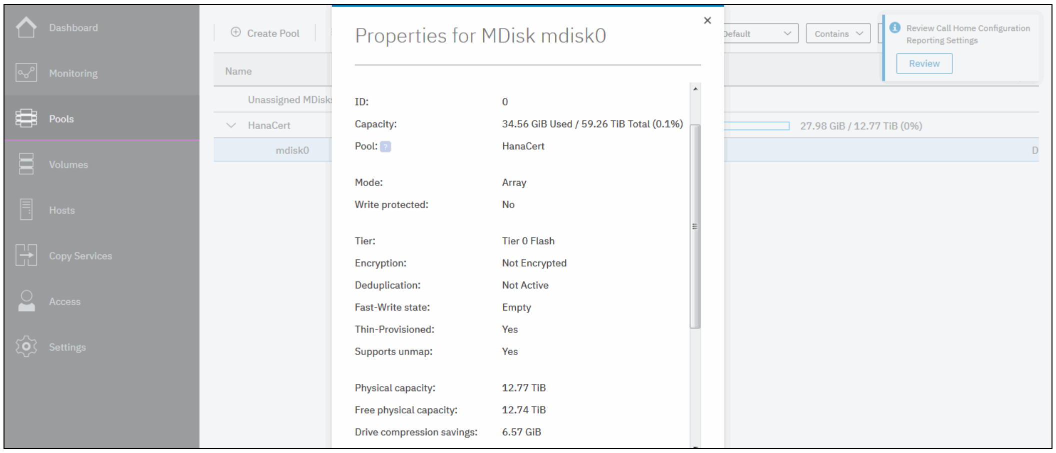
DRAID arrays on FlashCore modules need to be carefully monitored and well planned, as they are over-provisioned which means they are susceptible to an out-of-space condition.

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

- That 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 details see Chapter 8, “Monitoring” on page 267.

  FlashSystem 9100 GUI and CLI will also display used and available effective and physical capacities, for examples see Figure 3-2 and Example 3-3.



Properties for MDisk mdisk0

ID: 0
Capacity: 34.56 GiB Used / 59.26 TiB Total (0.1%)
Pool: HanaCert

Mode: Array
Write protected: No

Tier: Tier 0 Flash
Encryption: Not Encrypted
Deduplication: Not Active
Fast-Write state: Empty
Thin-Provisioned: Yes
Supports unmap: Yes

Physical capacity: 12.77 TiB
Free physical capacity: 12.74 TiB
Drive compression savings: 6.57 GiB

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 59.3TB
...
physical_capacity 12.77TB
physical_free_capacity 12.74TB
allocated_capacity 1.50TB
effective_used_capacity 34.56GB
```

- If the array used physical capacity reaches 99%, FlashSystem 9100 raises event ID 1241 “1% physical space left for compressed array”. This is a call for immediate action.

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

  - Add more 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, FlashSystem 9100 will issue UNMAP to the array and space will be released.

  You can also see IBM Flash Alert dedicated to out-of-space recovery.

- Arrays are most in danger of going out of space during a rebuild or when degraded. During normal DRAID operation, DRAID spare capacity which is distributed across array drives, remains free, 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 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 it might cause a physical out-of-space condition.

3.3 Quorum disks

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 IOgroups), 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 should be a member of an array or a “Candidate”, drives in “Unused” state can’t be quorums. MDisk must be in “Managed” state, “Unmanaged” or “Image” MDisks can’t be quorums.
- External MDisks can’t be provisioned over iSCSI, only FC.
- An MDisk must be presented by a disk subsystem that is supported to provide IBM FlashSystem 9100 quorum disks.

The system uses the following rules when selecting quorum devices:

- Fully connected candidates are preferred over partially connected candidates.

  It means that in a multiple enclosure environment MDisks will be preferred over drives.

- Drives are preferred over MDisks.

  If there is only one enclosure 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 FlashSystem 9100 contains more than one IOgroup, 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 FlashSystem 9100 with a single control enclosure, quorums including active quorum disk are assigned out of its internal drives automatically. No actions required.
- For FlashSystem 9100 with two or more IOgroups 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 FlashSystem 9100 with two or more IOgroups and without external storage, there will be no active quorum selected automatically. IP quorum or FC-attached quorum needs to be deployed.

To list FlashSystem 9100 quorum devices, run the **lsquorum** command as shown in Example 3-4.

Example 3-4 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* NVMe drive, but not *to* NVMe drive.

3.3.1 IP Quorum

The IP quorum application is a Java application that runs on a separate server or host.

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

Before IP quorum support was introduced, the third site had to be connected using Fibre Channel, and maintaining this third site and storage controller over FC makes the system costly for site recovery implementation of multi-IOgroup FlashSystem 9100.

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

IBM FlashSystem 9100 supports up to five IP Quorums at the same time, however, multiple instances of the IP quorum application cannot be run on the same host or server. The host might be physical or deployed as a virtual machine. Note that it must not depend on storage that is presented by the system. Dependent host can result in a situation where the nodes need to detect the IP quorum application to process I/O, but cannot because the IP quorum application cannot access storage.

In a HyperSwap implementation, it is suggested to have at least IP Quorums. They should be accessible to nodes on both sides independently of the inter-site link state. So in case the inter-site link fails, nodes on both sites must be able to communicate to Ethernet-attached IP quorum.

Note that with FlashSystem code levels before V8.3.0 it was impossible to be certain which site will be the winner during inter-site link failure. Because it is a race condition, its result cannot be predicted. With V8.3.0, for HyperSwap topology systems it is possible to choose the preferred site. 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.

If only one site runs critical applications, you can configure this site as preferred. During a disruption, the system delays processing tie-break operations on other sites that are not specified as preferred. The designated preferred site becomes more apt to resume I/O and critical applications remain online. If the preferred site is the site that is disrupted, the other site continues to win the tie-breaks and continues I/O.

> **Note:** If you have a multi-site configuration with IP Quorum as the active quorum device and quorum disks at site 1 and site 2, and you lose all IP Quorum Apps, you will not have any tie-break protection. If you have no active Quorum device, then the node with the lowest node ID (as shown by `lsnodecanister`) is used to resolve the tie-break, and a site containing this node will survive and continue serving I/O.
>
> This ID can change any time. For example, it happens if a node is removed from a cluster and added back.

4

Storage Pools

This chapter highlights considerations when you are planning storage pools for an IBM FlashSystem 9100 implementation. It explains various pool configuration options, including Data Reduction Pools, and provides best practices on the implementation as well as an overview of the process of adding and removing MDisks from existing storage pools.

> **Important, IBM FlashSystem 9200:** On 11th February 2020 IBM announced the arrival of the IBM FlashSystem 9200 to the family. This book was written specifically for IBM FlashSystem 9100, however most of the general principles will apply to the IBM FlashSystem 9200.
>
> If you are in any doubt as to their applicability to the FlashSystem 9200 then you should work with your local IBM representative.
>
> This book will be updated to include FlashSystem 9200 in due course.
>
> The Flashsystem 9200 product guide is available at:
>
> IBM FlashSystem 9200 Product Guide

This chapter includes the following sections:

- Introduction to Pools
- General considerations for Data Reduction Pools
- Storage pools planning considerations
- Tiered storage pool and Easy Tier
- Operations with Pools
- Considerations when using encryption
- Data Reduction Pools Best Practices

4.1 Introduction to Pools

In general, a Storage Pool or Pool is a grouping of storage visible to hosts, that consists of volumes, logical unit numbers (LUNs), or addresses that share a common set of administrative characteristics.

The IBM FlashSystem 9100 supports different types of Pools:

- Standard Pools (Parent Pools and Child Pools)
- Data Reduction Pools

Data Reduction Pools (DRP) represent a significant enhancement to the storage pool concept. This is 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.

Data Reduction Pools increase existing 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. This Chapter is dedicated to Data Reduction Pools.

4.1.1 Standard Pools

Standard Pools, sometimes also called Traditional or Legacy Storage Pools, is a way of providing storage in an IBM FlashSystem 9100 system and they use a fixed allocation unit of an extent. Standard Pools are still a valid method to providing capacity to hosts. See 4.3, “Storage pools planning considerations” on page 71 for guidelines regarding standard pools implementation.

IBM FlashSystem 9100 have the capacity to define Parent and Child Pools. A Parent Pool has all the capabilities and functions of a normal IBM FlashSystem 9100. A Child Pool is a logical sub division of a storage pool or managed disk group. Like a storage pool, a Child Pool supports volume creation and migration, but the user can specify the capacity of the Child Pool at creation.

A Child Pool inherits its tier setting from the Parent Pool. Changes to a parent’s tier setting are inherited by Child Pools. Changes to a Child Pool are applied to the Parent Pool and inherited by other siblings.

A Child Pool supports the Easy Tier function if the Parent Pool has Easy Tier enabled. The Child Pool also inherits Easy Tier status, pool status, capacity information, solid state status, and back-end storage information. The activity of Parent Pool and Child Pool are the same since the volumes from the Child Pool reside on the Parent Pool.

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 and it must remain 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 either to increase the number of extents that are available for new volume copies or to expand existing volume copies. The system automatically balances volume extents between the MDisks to provide the best performance to the volumes.

You cannot use the data migration function to migrate volumes between Parent Pools that have different extent sizes. However, you can use volume mirroring to move data to a Parent Pool that has a different extent size. Consider choosing extent size wisely according to you future needs. Small extents limit your overall usable capacity as well as using a larger extent size can waste storage.

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. Allocating MDisks within the same tier ensures that a single failure of a physical disk does not take the entire pool offline. For example, if you have three RAID-5 arrays in one pool and add a non-RAID disk to this pool, you lose access to all the data that is striped across the pool if the non-RAID disk fails. Similarly, for performance reasons, you must not mix RAID types. The performance of all volumes is reduced to the lowest achiever in the tier.
- An MDisk can be associated with just one Parent Pool.
- You can 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 when you migrate between Parent Pools.
- Volumes that are allocated from a Parent Pool are striped across all the storage that is placed into that Parent Pool. This also enables nondisruptive migration of data from one storage system to another storage system and helps simplify the decommissioning process if you want to decommission a storage system later.
- You can only add MDisks that are in un-managed mode. When MDisks are added to a Parent Pool, their mode changes from un-managed 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 any 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.
- You can delete an array MDisk from a Parent Pool when:
  - Volumes are not using any of the extents that are on the MDisk.
  - Enough free extents are available elsewhere in the Parent Pool to move any extents that are in use from this MDisk.

  Before you remove MDisks from a Parent Pool, ensure that the Parent Pool has enough capacity for any 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 has associated Child Pools, then 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 un-managed 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.

- If the volume is mirrored and the synchronized copies of the volume are all in one pool, the mirrored volume is destroyed when the storage pool is deleted. If the volume is mirrored and there is a synchronized copy in another pool, the volume remains after the pool is deleted.

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. The 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.

When you create or work with a Child Pool, consider the following general guidelines:

- Child Pools can be created and changed with the command-line interface or through IBM Spectrum Control when creating VMware vSphere Virtual Volumes. You can use the management GUI to view Child Pools and their properties.
- 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 below 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
- A Child Pool cannot be created from a Data Reduction Pool.

4.1.2 Data Reduction Pools

The IBM FlashSystem 9100 arrays leverages innovative new Data Reduction Pools (DRP) that incorporate deduplication and hardware-accelerated compression technology, plus SCSI UNMAP support and all the thin provisioning and data efficiency features you'd expect from IBM Spectrum Virtualize based storage to potentially reduce your CAPEX and OPEX. Additionally, all these benefits extend to over 200 heterogeneous storage arrays from multiple vendors.

Data Reduction Pools were built from the ground up with space reclamation in mind. Data Protection Pools provide

- Log Structured Array allocation
- Garbage collection to free whole extents
- Fine grained (8KB) chunk allocation within an extent.
- SCSI unmap and write same (Host) with automatic space reclamation
- Support for 'back-end' unmap and write same
- Support Compression
- Support Deduplication

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 reclaiming previously used storage resources that are no longer needed by host systems. To estimate potential capacity savings that data reduction technologies can provide on the system, use the Data Reduction Estimation Tool (DRET).

This tool analyzes existing user workloads which are being 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.

The Data Reduction Estimation Tool (DRET) is a command-line, host-based utility for estimating the data reduction savings on block storage devices. To help with the profiling and analysis of existing user workloads that must be migrated to a new system, IBM provides the highly accurate DRET to support both deduplication and compression. The tool scans target workloads on various legacy storage arrays (from IBM or another company), merges all scan results, then provides an integrated system-level data reduction estimate.

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.

Go here to search under FlashSystem 9100 to find the tool and its readme.

To use data reduction technologies on the system, you need to create a Data Reduction Pool, create thin-provisioned or compressed volumes, and map these volumes to hosts that support SCSI unmap commands.

SCSI UNMAP

DRPs support end-to-end unmap functionality. Space that is freed from the hosts is a process called unmap. A host can issue a small file unmap (or a large chunk of unmap space if you are deleting a volume that is part of a data store on a host), and either of these results in the freeing of all the capacity allocated within that unmap. Similarly, deleting a volume at the DRP level frees all of the capacity back to the pool.

When a Data Reduction Pool is created, the system monitors the pool for reclaimable capacity from host unmap operations. This capacity can be reclaimed by the system and redistributed into the pool. Create volumes that use thin provisioning or compression within the Data Reduction Pool to maximize space within the pool.

Unmap can also save work on the storage system. For example, RAID arrays don't need to rebuild unmapped capacity. Flash drives can re-use the space for wear levelling or extra 'pre-erased' blocks.

Virtualization devices like IBM FlashSystem 9100 with external storage can also pass on Unmap information to other storage systems - for example when extents are deleted or migrated.

> **Note:** Standard Pool also support SCSI UNMAP.
> Hosts can issue SCSI Unmap commands to storage controllers, to indicate that an LBA range on a disk can be freed. This might happen, for example, when formatting a new volume, or deleting files in a filesystem.
>
> The IBM FlashSystem 9100 receives a SCSI Unmap command, it overwrites the relevant region of the volume with all-zero data and can reclaim physical capacity through garbage collection (FCM level).
>
> This helps prevent a thin-provisioning storage controller from running out of free capacity for write I/O requests. Host Unmap commands will not increase the free capacity reported by the standard storage pool.
>
> **Note:** SCSI UNMAP might add more workload to the back-end storage.
>
> Performance monitoring helps to notice possible impact and if SCSI Unmap workload is impacting performance, consider taking appropriate steps.

DRP internal details

Data Reduction Pools can also contain fully allocated (or thick) volumes. The space for these volumes is claimed at creation time. Thin-provisioned volumes, on the other hand, delay allocating storage until it is first written. Therefore, unused space in volumes presented to host servers does not actually use any physical storage capacity. To create a thin-provisioned volume, the user specifies both the real and virtual capacity of the volume when it is created. The real capacity defines how much physical storage is initially associated with the volume, and the virtual capacity defines how large the volume will appear to the host server.

Data Reduction Pools consists of various number of volumes and it is important to understand the approach how it is implemented. Every user volume has corresponding journal, forward lookup, and directory volumes.

The structure of how Data Reduction Pool volumes are used, is important for a inter-volume deduplication, and as well amortizes space allocation overheads. The journal volume per I/O group, used for recovery purposes, and a reverse lookup volume per I/O group, used by garbage collection.

Figure 4-1 denotes the difference between Data Reduction Pool volumes and volumes in Standard Pools.

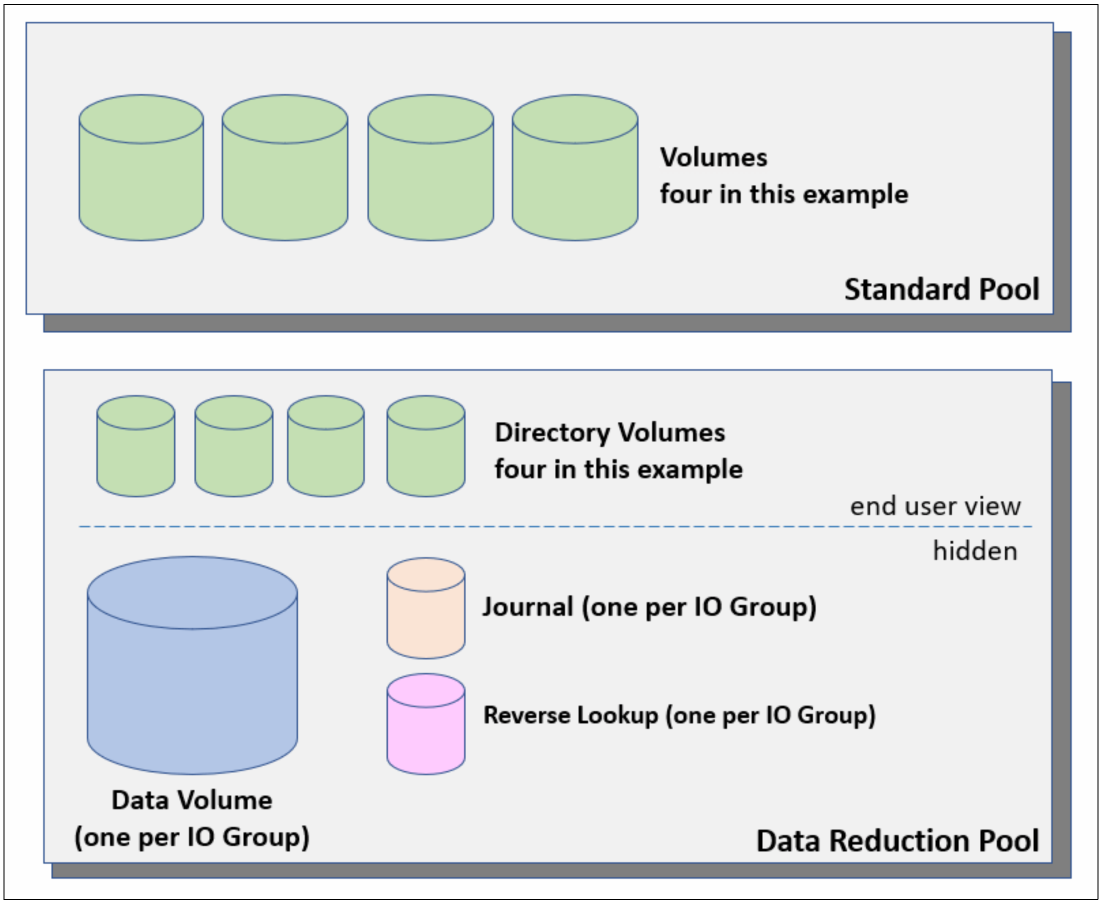


Figure 4-1 Standard and Data Reduction Pool - volumes

The Data Volume uses >97% of Pool capacity. The I/O pattern is a large sequential write pattern (256 KB), coalesced into full stride writes and you typically see a short random read pattern. Directory Volumes occupy approximately 1% of pool capacity. They have a short 4 KB random read and write I/O. Journal Volumes occupy approximately 1% of pool capacity, and show large sequential write I/O (256k typically).

Journal Volumes are only read for recovery scenarios (for example, T3 recovery and so on). Reverse Lookup Volumes are used by the Garbage Collection process and occupy less 1% of pool capacity. Reverse Lookup Volumes have a short, semi-random read write pattern.

The process of reclaiming space is called Garbage Collection. As a result of compression and deduplication, overwriting host writes does not always use the same amount of space that the previous data was using. This leads to these writes always having to occupy new space on back-end storage while the old data is still in its original location. The primary task of Garbage Collection is to track all of the regions that have been invalidated, and to make this capacity usable for new writes.

Figure 4-2 shows the Garbage Collection process.

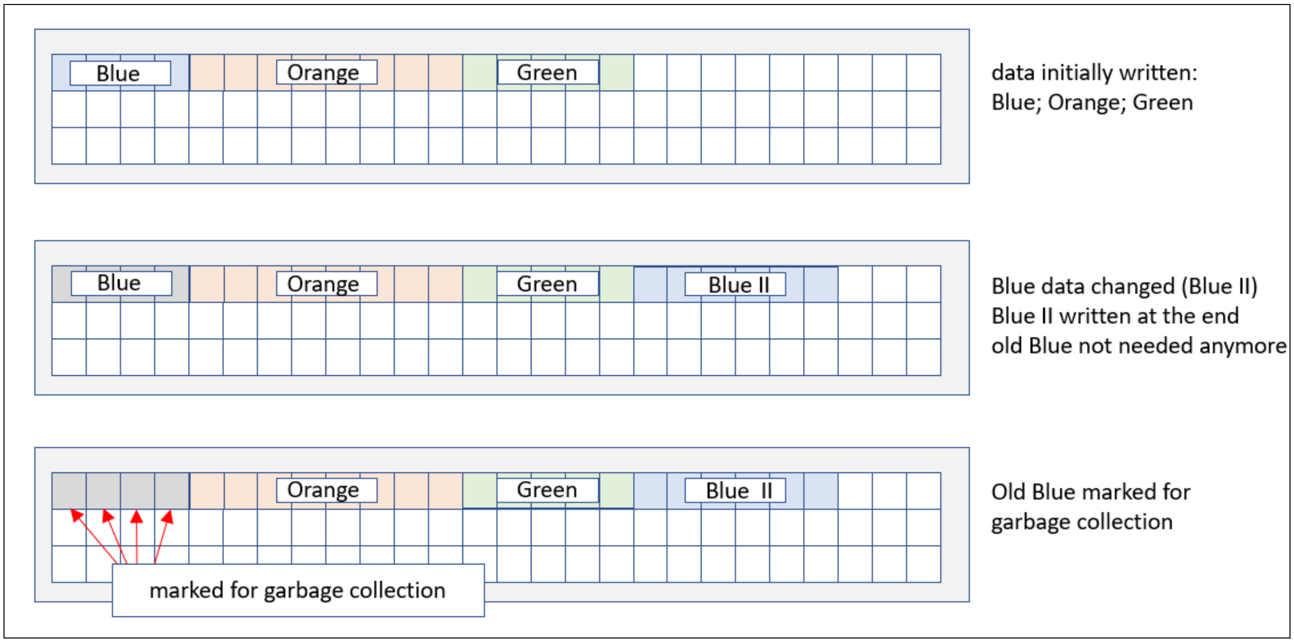


Figure 4-2 Garbage Collection principle

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.

Data Reduction Pools have built in services to enable Garbage Collection (GC) of unused blocks. This means that lots of smaller unmaps end up enabling a much larger chunk (extent) to be freed back to the pool. Trying to fill small holes is very inefficient: too many I/Os would be needed to keep reading and rewriting the directory. So, GC waits until an extent has many small holes.

Move the remaining data in the extent, compact, and rewrite. When we have 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 in order to free regions, it is suggested that you size pools in order to keep a certain amount of free capacity available. This practice ensures that there will always be some free space for Garbage Collection. For more information, see Chapter 4.7.3, “DRP provisioning considerations” on page 96.

4.1.3 Standard Pools versus Data Reduction Pools

When it comes to pools during the planning of a IBM FlashSystem 9100 project, it is important to know all requirements and to understand the upcoming workload of the environment. As the IBM FlashSystem 9100 is very flexible in creating and using Pools, we will discuss in this section how to figure out which types of Pool or setup you can use.

Among the things you should be aware of the planned environment, so you have to know details about:

- Is your data compressible?
- Is your data deduplicable?
- What are the workload and performance requirements
  - Read / Write ratio
  - Blocksize
  - IOps, MB/sec and response time
- Flexibility for the future
- Thin Provisioning
- Child Pools

Is your data compressible?

Compression is one option of Data Reduction Pools, and the deduplication algorithm is used to reduce the on-disk footprint of data written to by thin provisioning. In IBM FlashSystem 9100 this is an in-line deduplication approach rather than attempting to compress as a background task. DRP provides unmap support at pool and volume level and out-of-space situations can be managed at the DRP pool level.

Compression can be enabled in Data Reduction Pools on a per Volume basis and Thin Provisioning is a pre-requisite. The input size has changed to a fixed 8 KB. Compression is suited to Flash workloads (IOPS) and a typical 2:1 compression ratio will result in ~4 KB operations and Streaming 256 KB chunks of 8 KB blocks for consistent write performance.

Data compression techniques depend on the type of data that has to 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. There are several ways to decide if your data is compressible or not:

- General assumptions
- Tools

General assumptions

IBM FlashSystem 9100 compression is lossless and, as the name implies, it involves no loss of information. It can be losslesly compressed and the original data can be exactly recovered after the compress/expend cycle. Good compression savings can be achieved in various environments:

- Virtualized Infrastructure
- Database and Data Warehouse
- Home Directory, Shares, as well as shared project data
- CAD/CAM
- Oil and Gas data
- Log data
- SW development
- Text, some picture files

However, you should take care if the data is already compressed. The savings will be less or could even be negative. Pictures (for example, GIF, JPG, PNG, and so on), Audio (MP3, WMA, and so on) and Video or Audio (AVI, MPG, and so on) and even compressed databases data might not be a good candidate for compression.

Table 4-1 describes 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 encrypted data will not show much savings, as it 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.

See 4.7.1, “IBM FlashSystem 9100 and DRP” on page 93 for further considerations on compression.

Note: Saving assumptions based on this kind of data are very imprecise, and we recommend to determine compression savings with the proper tools.

Tools

Using compression saving tools or functions is the most accurate way to determine the estimated savings. IBM provides some techniques which can help you to figure out if data compresses reasonably:

- Comprestimator
- Comprestimator embedded in IBM storage systems
- Data Reduction Estimator Tool (DRET)

  IBM provides the highly accurate DRET to support both deduplication and compression. The host based cli tool scans target workloads on various legacy storage arrays (from IBM or another company), merges all scan results, then provides an integrated system-level data reduction estimate.

  The Data Reduction Estimator utility uses advanced mathematical and statistical algorithms to perform an analysis with very low memory footprint. The utility runs on a host that has access to 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 environment configuration, in many cases Data Reduction Estimator will be used on more than one host, in order 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 initial 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 using Data Reduction Estimator to analyze such volumes, the expected reduction results reflect the saving 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 prior to the deletion becomes free and available to the file system. This happens 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 using Data Reduction Estimator 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 make it 100% used, 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, Data Reduction Estimator (or any other utility for that matter) will access the data that belongs to the files that are already deleted.

In order to reduce the impact of block device and file system behavior mentioned above, it is recommended to use Data Reduction Estimator to analyze volumes that contain as much active data as possible rather than volumes that are mostly empty of data. This increases accuracy level and reduces the risk of analyzing old data that is already deleted, but might still have traces on the device.

The Tool can be downloaded from FixCentral.

Example 4-1 shows the CLI usage of the Data Reduction Estimator Tool.

Example 4-1 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]
```

- Data Reduction Estimator 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
  - HP-UX 11.31

Is your data a deduplication candidate?

Deduplication is done by using hash tables in order to identify previously written copies of data. If duplicate data is found, instead of writing the data to disk, the algorithm will reference to the previously found data.

Data Reduction Pools are designed to support reclamation, compression and deduplication from the start. Deduplication uses 8KB deduplication chunks and an SHA-1 hashing algorithm and 256KB streaming capability. Data Reduction Pools provide deduplicate 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 environments such as virtual desktop and some virtual machine environments. So they might be good candidates for deduplication.

Tools

Data Reduction Estimation Tool (DRET) can be used to assess the deduplication savings it is the most accurate way to determine the estimated savings.

Data Reduction Estimation Tool (DRET)

IBM provides DRET to support both deduplication and compression. The host based cli tool scans target workloads on various legacy storage arrays (from IBM or another company), merges all scan results, then provides an integrated system-level data reduction estimate for your IBM FlashSystem 9100I planning.

The Data Reduction Estimator utility uses advanced mathematical and statistical algorithms to perform an analysis with very low memory footprint. The utility runs on a host that has access to 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 environment configuration, in many cases Data Reduction Estimator will be used on more than one host, in order 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 initial zeroing the device can contain traces of old data on the block device level. Such data is not accessible or viewable on the file system level. When using Data Reduction Estimator to analyze such volumes, the expected reduction results reflect the saving rate to be achieved for all the data on the block device level, including the 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 prior to the deletion becomes free and available to the file system. This happens 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 using Data Reduction Estimator 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 100GB file system and make it 100% used, 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, Data Reduction Estimator (or any other utility for that matter) will access the data that belongs to the files that are already deleted.

In order to reduce the impact of block device and file system behavior mentioned above, it is recommended to use Data Reduction Estimator to analyze volumes that contain as much active data as possible rather than volumes that are mostly empty of data. This increases accuracy level and reduces the risk of analyzing old data that is already deleted, but might still have traces on the device.

The Tool can be downloaded on FixCentral.

Example 4-2 shows CLI usage of the Data Reduction Estimation Tool.

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]
```

Data Reduction Estimator 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 tool, use Data Reduction Pools to exploit data deduplication savings that are available, unless performance requirements exceed what DRP can deliver.

Use Data Reduction Pools to exploit data deduplication savings that are available, unless performance requirements exceed what DRP can deliver.

Do not enable deduplication if the data set is not expected to provide deduplication savings.

What are the workload and performance requirements

An important factor of sizing and planning for a IBM FlashSystem 9100 environment is the knowledge of the workload characteristic of a specific environment.

Workloads which impact the sizing and performance are, among others:

- Read/Write ratio

  Read/Write (%) ratio will affect performance as higher write cause more IOps to the Data Reduction Pool. To effectively size an environment, the Read/Write ratio should being taken into consideration. During a write I/O, when data is written to the Data Reduction Pool, it will be stored on the data disk and the forward lookup structure is getting updated and the I/O will be completed.

  Data Reduction Pools are making use of metadata. Even when there are no volumes in the pool, some of the space in the pool is used to store the metadata. The space allocated to metadata is relatively small. Regardless of the type of volumes that the pool contains, metadata are always stored separately from customer data.

  In Data Reduction Pools, 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 due to 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. So keep in mind that Data Reduction Pools create **more IOps** on the FCMs or drives.

- Blocksize

  The concept of a block size is quite simple and the impact on storage performance might be distinct. Blocksize affects might have an impact on overall performance as well, so consider larger blocks to affect performance more then smaller blocks. Understanding and considering for block sizes in the design, optimization and operation of the IBM FlashSystem 9100 sizing, leads to a more predictable behavior of the entire environment.

- IOps, MBps, and response time

  Storage constraints are IOps, Throughput and latency and it's crucial to correctly design the solution or plan for a setup for speed and bandwidth. Proper sizing requires knowledge on the expected requirements.

- Capacity

  During the planning of an IBM FlashSystem 9100 environment, capacity (physical) has to be sized accordingly. Compression and deduplication might save space, but metadata will consume a little bit of space and for an optimal performance our recommendation is to utilize the Data Reduction Pool to a maximum of 85%.

Consider monitoring storage infrastructure requirements with monitoring and/or management software such as IBM Spectrum Control or IBM Storage Insights before planning a new environment. The peak workload, like IOps, MB/sec, peak response time, and so on, at busy times, will give you an understanding of the required workload plus expected growth. Also take into account to allow enough headroom regarding performance required during planned and unplanned events (upgrades and possible defects or failures).

It's important to understand the relevance of Application response time rather than internal response time in conjunction with required IOps or throughput. Typical OLTP applications require IOps and low latency as well. Do not place capacity over performance while designing or planning a storage solution. Even if capacity might be sufficient, the environment might suffer from low performance. Deduplication and compression might satisfy capacity needs, but aim on performance as well for robust application performance.

In order to size an IBM FlashSystem 9100 environment accordingly, you can use Disk Magic. The tool can be used to determine, that Data Reduction Pools are going to give suitable bandwidth and latency. If the data won't deduplicate (according to the DRET tool) the Volume can also be fully allocated or compressed only.

4.1.4 Flexibility for the future

During the planning and configuration of Storage Pools, the decision has to been taken which pools to create. As the IBM FlashSystem 9100 allows you to create Standard Pools or Data Reduction Pools you have to decide which type fits the requirements best.

We have already discussed requirements, such as workload and performance, and if the data is compressible or if the data will effectively deduplicate. Verify if the performance requirements meet the capabilities of the specific Pool type. Therefore you can see the section "What are the workload and performance requirements" on page 67 for more details. Later we will cover the dependencies with Child Pools and VVols in "Child Pools and VVols" and "DRP restrictions" on page 96.

If other important factors do not lead to using Standard Pools, then Data Reduction Pools are the right choice. Usage of Data Reduction Pools 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.

Data Reduction Pools provide great flexibility for future use. They add the flexibility of compression and deduplication of data at the volume level in a specific pool even if these features are initially not used at the time of creation.

Keep in mind that it is not possible to convert a Pool. If you have to change the Pool Type (Standard Pool to a Data Reduction Pool or vice versa) it will be an offline process and you have to migrate the your data as described in 4.7.5, "Data migration with DRP" on page 99.

Note: We recommend to use Data Reduction Pools with fully allocated volumes as long as the restrictions and capacity do not affect your environment. For more details on the restrictions please see "DRP restrictions" on page 96.

4.1.5 Thin Provisioning

A thin-provisioned volume presents a different capacity to mapped hosts than the capacity that the volume consumes in the storage pool. The IBM FlashSystem 9100 system supports thin-provisioned volumes in both standard pools and Data Reduction Pools.

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. With Data Reduction Pools, all the benefits of thin-provisioning are available to all the volumes that are assigned to the pool. Only fully allocated volumes do not gain these benefits. For the thin-provisioned volumes in data reduction pools, you can also configure compression and data deduplication on these volumes, increasing the capacity savings for the entire pool.

Data Reduction Pools 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 Data Reduction Pool to be redistributed automatically (UNMAP). Standard pools do not have these functions.

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 using any real capacity.

For more details about storage system, pool, and volume capacity metric, see Chapter 8, “Monitoring” on page 267.

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 the you have physically available in you IBM FlashSystem 9100 storage subsystem. Details on monitoring physical capacity of your storage and an explanation of the difference between Thin provisioning and over allocation can be found at fount at 8.4, “Creating Alerts for IBM Spectrum Control and IBM Storage Insights” on page 302.

Child Pools and VVols

At the time of writing, Data Reduction Pools are created as parent pools only, so no Child Pools are available for Data Reduction Pools.

If you have to plan for VMware VVols, then Standard/Traditional Pools is the only option.

> **Note:** Child Pools are only available in conjunction with Standard Pools.

4.2 General considerations for Data Reduction Pools

This section discusses limitations with Standard Pool and Data Reduction Pools.

4.2.1 Data Reduction Pool limitations

Table 4-2 describes the limitations of Data Reduction Pools (IBM FlashSystem 9100 V8.2.1) at the time of writing.

Table 4-2 Data Reduction Pool limitations

Property	**Maximum Number**	**Comments**
Data Reduction Pools per system	4	
MDisks per Data Reduction Pool	128	
Volumes per Data Reduction Pool	10000 - (Number of Data Reduction Pools x 12)	
Extents per I/O group per Data Reduction Pool	128K	
Compressed volume copies in data reduction pools per system	-	No limit is imposed here beyond the volume copy limit per data reduction pool
Compressed volume copies in data reduction pools per I/O group	-	No limit is imposed here beyond the volume copy limit per data reduction pool
Deduplicated volume copies in data reduction pools per system	-	No limit is imposed here beyond the volume copy limit per data reduction pool
Deduplicated volume copies in data reduction pools per I/O group	-	No limit is imposed here beyond the volume copy limit per data reduction pool
Fully-allocated volume capacity	256 TB	Maximum size for an individual fully-allocated volume. Maximum size is dependent on the extent size of the Storage Pool. Comparison Table: Maximum Volume, MDisk and System capacity for each `extents size.`

Property	Maximum Number	Comments
Thin-provisioned (space-efficient) per-volume capacity for volumes copies in regular and data reduction pools	256 TB	Maximum size for an individual thin-provisioned volume. Maximum size is dependent on the extent size of the Storage Pool. Comparison Table: Maximum Volume, MDisk and System capacity for each extent size. Comparison Table: Maximum Volume, MDisk and System capacity for each `extent size.`
Compressed volume capacity in regular pools	96 TB	Maximum size for an individual compressed volume Maximum size is dependent on the extent size of the Storage Pool. Comparison Table: Maximum Volume, MDisk and System capacity for each `Extent size`

For detailed Information see the current Support Information for the FlashSystem 9100 family.

4.3 Storage pools planning considerations

The implementation of storage pools in a IBM FlashSystem 9100 requires a holistic approach that involves application availability and performance considerations. Usually a trade-off between these two aspects must be taken into account.

In this section the main best practices in the storage pool planning activity are described. Most of these practices apply to both standard pools and DRP except where otherwise specified. For additional specific best practices for Data Reduction Pools (DRP) see 4.7, “Data Reduction Pools Best Practices” on page 93. See 4.7, “Data Reduction Pools Best Practices” on page 93.

4.3.1 Planning for availability

By design, IBM Spectrum Virtualize based storage systems takes the entire storage pool offline if a single MDisk in that storage pool goes offline. This means that the storage pools quantity and size define the failure domain.

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

Sometimes, reducing the hardware failure domain, such as placing the volumes of an application into a single storage pool, is not always an advantage from the application perspective.

Alternatively, splitting the volumes of an application across multiple storage pools increases the chances of having an application outage if one of the storage pools that is associated with that application goes offline.

Finally, increasing the number of pools, to reduce the failure domain, is not always a viable option. For instance in IBM FlashSystem 9100 systems with configurations without expansion enclosures, the number of physical drives is limited (up to 24) and creating more arrays would end up reducing usable space due to spare and protection capacity. Consider for instance a single I/O group IBM FlashSystem 9100 configuration with 24 7.68 TB NVMe drives. In case of a single array DRAID6 creation the available physical capacity would be 146.3 TB, while creating two arrays DRAID6 would provide 137.2 TB of available physical capacity with a reduction of 9.1 TB.

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 an hybrid pool managed by Easy Tier (see 4.4, “Tiered storage pool and Easy Tier” on page 75)
- Create a storage pool for each external virtualized storage subsystem, unless you are creating a hybrid pool managed by Easy Tier (see 4.4, “Tiered storage pool and Easy Tier” on page 75)
- Use dedicated pools for image mode volumes.

> **Limitation:** Image Mode volumes are not supported with Data Reduction Pools.

4.3.2 Planning for performance

When planning storage pools for performance the capability to stripe across disk arrays is one of the most important advantages of 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.

Also adding more arrays to a pool, rather than creating a new one, can be a way to improve the overall performance as long as the added arrays have the same or better performance characteristics of the existing ones.

Note that in IBM FlashSystem 9100 configurations arrays built from FCM and SAS SSD drives have different characteristic, both in terms of performance and data reduction capabilities. For this reason when using FCM and SAS SSD arrays in the same pool the recommendations are:

- Enable the Easy Tier function (see 4.4, “Tiered storage pool and Easy Tier” on page 75). The Easy Tier treats the two array technologies as different tier (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 even 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. However, when sizing the environment keep in mind 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.

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.
- In a IBM FlashSystem 9100 clustered environment, create storage pools with IOgrp or Control Enclosure affinity. That means you have to use only arrays/MDisks supplied by the internal storage that is directly connected to one IOgrp SAS chain only. This configuration avoids unnecessary IOgrp-to-IOgrp communication traversing the SAN and consuming Fibre Channel bandwidth.
- Use dedicated pools for image mode volumes.

  > **Limitation:** Image Mode volumes are not supported with Data Reduction Pools.

- Try to limit the number of storage pools to avoid excessive cache partitioning.
- For those Easy Tier enabled storage pools, always allow some free capacity for Easy Tier to deliver better performance.
- Consider implementing child pools when you need to have a logical division of your volumes for each application set. There are often cases where you want to subdivide a storage pool but maintain a larger number of MDisks in that pool. Child pools are logically similar to storage pools, but allow you to specify one or more subdivided child pools. Thresholds and throttles can be set independently per child pool.

  > **Limitation:** Child pools (VVols) are not supported with Data Reduction Pools

4.3.3 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. The IBM Spectrum Virtualize architecture can manages 2^22 extents for a system and therefore the choice of extent size affects the total amount of storage that can be addressed. Check the link for the capacity limits per extent size.

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 the pool extent size planning, consider the following recommendations:

- For standard pools usually 1 GB is suitable
- For Data Reduction Pools use 4 GB (see 4.7, “Data Reduction Pools Best Practices” on page 93 for further considerations on extent size for 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 Data Reduction Pools is currently not supported.

4.3.4 External pools

All the IBM Spectrum Virtualize based storage systems, including the IBM FlashSystem 9100, have the ability to virtualize external storage systems. There are special considerations when configuring storage pools with external storage.

Availability considerations

Although the IBM FlashSystem 9100 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 IIBM FlashSystem 9100 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 9100 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 Easy Tier hybrid pools. For more information, see 4.4, “Tiered storage pool and Easy Tier” on page 75.
- 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 4.4, “Tiered storage pool and Easy Tier” on page 75.

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.

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 also a determining factor, where adding IBM FlashSystem 9100 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, and the IBM System Storwize V7000 series can be scaled out with enough units to deliver the same performance.

A significant consideration when you compare native performance characteristics between storage subsystem types is the amount of scaling that is required to meet the performance objectives. Although lower-performing subsystems can typically be scaled to meet performance objectives, the additional hardware that is required lowers the availability characteristics of the IBM FlashSystem 9100 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.

Other important consideration regarding the external virtualization pools can be found in *IBM System Storage SAN Volume Controller and Storwize V7000 Best Practices and Performance Guidelines*, SG24-7521 because they equally apply to the FS9100.

4.4 Tiered storage pool and Easy Tier

IBM Spectrum Virtualize makes it easy to configure multiple tiers of storage within the same cluster by using the Easy Tier function. Easy Tier automates the placement of data among different storage tiers, and it can be enabled for internal and external storage. This includes the ability to automatically and non-disruptively relocate data (at the extent level) from one tier to another tier, or even within the same tier, in either direction. This process achieves the best available storage performance for your workload in your environment. Easy Tier reduces the I/O latency for hot spots, but it does not replace storage cache.

You might have single-tiered pools, multi-tiered storage pools, or both.

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

- They have the same hardware characteristics. For example, they need the same RAID type, RAID array size, disk type, and disk revolutions per minute (RPM).
- It is recommended to keep size and performance characteristics in a *single-tiered storage pool*. If this consistency is not possible because of storage pool space, upgrades at different times, and disks with the same size would no longer be available, use disks with sizes closer to the original one.

  This configuration would not have a serious side effect on the performance, because Easy Tier has introduced *storage pool balancing* (or *Intra-Tier* balancing) that balances the workload on different MDisks (in this case with different drive sizes) based on I/O density and response time.
- The MDisks that are configured must have the same size whenever possible. If this requirement is not feasible, IBM Spectrum Virtualize Easy Tier with Intra-Tier balancing can balance the workload on different MDisks (in this case with different drive size) based on I/O density and response time.

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

- **Flash or SSD**

  Specifies a flash (or SSD drive) drive or an external MDisk for the newly discovered or external volume. These kind of technologies can be mapped in two different tier:

  - *tier0_flash* that represents enterprise flash technology
  - *tier1_flash* that represents Read Intensive flash technology or SAS-SSD technology included in the IBM FlashSystem 9100 systems expansion enclosures

- **Enterprise HDD**

  Specifies an enterprise hard disk drive or an **external** MDisk for the newly discovered or external volume. These technologies are mapped to the *tier2_hdd* tier.

- **Nearline HDD**

  Specifies a nearline hard disk drive or an **external** MDisk for the newly discovered or external volume. These technologies are mapped to the *tier3_nearline* tier.

In multi-tiered pools Easy Tier enables the data movement at extent level across the available tiers with the aim to optimize the performance for the specific workload running on the pool's volumes.

When enabled, Easy Tier performs the following actions across the tiers:

- Promote

  Moves the hotter extents to a higher performance tier with available capacity. Promote occurs within adjacent tiers.

- Demote

  Demotes colder extents from a higher tier to a lower tier. Demote occurs within adjacent tiers.

- Swap

  Exchanges cold extent in an upper tier with hot extent in a lower tier.

- Warm Demote

  Prevents performance overload of a tier by demoting a warm extent to a lower tier. This process is triggered when bandwidth or IOPS exceeds predefined threshold.

- Warm Promote

  This feature addresses the situation where a lower tier suddenly becomes very active. Instead of waiting for the next migration plan, Easy Tier can react immediately. Warm promote acts in a similar way to warm demote. If the 5-minute average performance shows that a layer is overloaded, Easy Tier immediately starts to promote extents until the condition is relieved.

- Cold Demote

  Demotes inactive (or cold) extents that are on a higher performance tier to its adjacent lower-cost tier. In that way Easy Tier automatically frees extents on the higher storage tier before the extents on the lower tier become hot. Only supported between HDD tiers.

- Expanded Cold Demote

  Demotes appropriate sequential workloads to the lowest tier to better use nearline disk bandwidth.

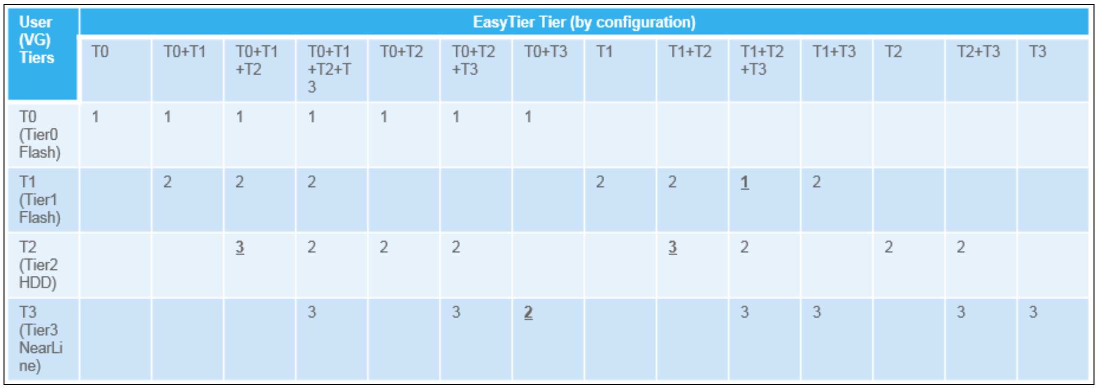
- Storage Pool Balancing

  Redistributes extents within a tier to balance usage across MDisks for maximum performance. This process moves hot extents from high used MDisks to low used MDisks, and exchanges extents between high used MDisks and low used MDisks.

- Easy Tier attempts to migrate the most active volume extents up to SSD first.
- A previous migration plan and any queued extents that are not yet relocated are abandoned.

> **Note:** Extent migration occurs only between adjacent tiers. For instance, in a three-tiered storage pool, Easy Tier will not move extents from the flash tier directly to the nearline tier and vice versa without moving them first to the enterprise tier.

An internal level Easy Tier is a three tier storage architecture, while from a user perspective, four tiers (or `Tech Types`) are available. These user tiers are mapped to Easy Tier tiers depending on the pool configuration. Figure 4-3 shows the possible combinations for the pool configuration of the four user tiers.



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

Figure 4-3 Tier and tech types combination

Therefore, a multitiered storage pool contains MDisks with various characteristics, as opposed to a single-tier storage pool. However, each tier must try to follow the same rules applied for the *single-tiered storage pool*.

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

Note that in IBM FlashSystem 9100 systems the Tech Type of the **internal** drives varies on the drive type, and specifically:

- NVMe drives have the *tier0_flash* Tech Type and therefore when creating an array from these drives the resulting MDisk will have *tier0_flash* tier
- SAS attached SSD drives have the *tier1_flash* Tech Type and therefore when creating an array from these drives the resulting MDisk will have *tier1_flash* tier

According to the table in Figure 4-3, in a IBM FlashSystem 9100 system a pool containing a mix of NVMe and SAS based arrays, is treated as a 2 tiers pool with the NVMe MDisks as tier 1 and the SAS MDisks as tier 2.

> **Important:** In a multi-tiered pool containing a mix of IBM FCM and SAS based arrays, pay attention to the physical FCMs usage as Easy Tier inter-tier balancing could affect the compression ratio and an out of space condition can be reached even without changing the data contents.

If you present an external controller to IBM FlashSystem 9100, a specific Easy Tier `easytierload` profile is assigned. It can be low, medium, high, or very_high. It specifies the Easy Tier load (amount) to place on a non-array MDisk within its tier.

If you present an external MDisk to IBM FlashSystem 9100, it becomes *tier2_hdd* by default, even if that external MDisk was built by using SSD drives or a flash memory system. You must change the MDisk tier only for MDisks that are presented from external storage systems accordingly with the owning storage controller by using the **`chmdisk`** command.

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

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

Effectively, you have four tiers within a 3-Tier Mapping. When you create a volume, the initial capacity is always allocated from *tier1_flash* or *tier2_hdd* by default.

This default behavior can be manipulated by specifying an MDisk list with the **`mkvdisk`** command to avoid exhausting the capacity from *tier1_flash* or *tier2_hdd*.

Additionally, the Easy Tier feature 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.

4.4.1 Interaction between Easy Tier and garbage collection in DRP

DRP makes use of Log Structured Array (LSA) structures which need garbage collection activity to be done regularly. An LSA always appends new writes to the end of the allocated space (see “DRP internal details” on page 60). Even if data already 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 collected. This process provides the following advantages:

- Writes to a DRP volume are always treated as sequential: so we can build all the 8 KB chunks into a larger 256 KB chunk and destage the writes from cache, either as full stripe writes, or as large as a 256 KB sequential stream of smaller writes.
- This should give the best performance both in terms of RAID on back-end systems, and also on Flash, where it becomes easier for the Flash device to perform its internal garbage collection on a larger boundary.

We can start to record metadata about how frequently certain areas of a volume are overwritten. 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 is so that Easy Tier will operate correctly for not only read data, but write data, when data reduction is in use.

Previous writes to compressed volumes held lower value to the Easy Tier algorithms, because writes were always to a new extent, so the previous heat was lost. Now, we can maintain the heat over time and ensure that frequently rewritten data gets grouped together. This also aids the garbage collection process where it is likely that large contiguous areas will end up being garbage collected together.

4.5 Operations with Pools

In the following section we describe some guidelines for the typical operation with pools, which apply both to standard and DRP type.

4.5.1 Creating Data Reduction Pools

This section describes how to create Data Reduction Pools.

Using the management GUI

To create a Data Reduction Pool on the system, complete these steps:

1. Create a data reduction pool by completing these steps:
   a. In the management GUI, select **Pools** → **Pools**.
   b. On the **Pools** page, click **Create**.
   c. On the **Create Pool** page, enter a name of the pool and select **Data Reduction**.
   d. Click **Create**. Data Reduction Pools are created as parent pools only, not child pools.
2. Add storage to the data reduction pool by completing these steps:
   a. In the management GUI, select **Pools** → **Pools**.
   b. Right-click the Data Reduction Pool that you created and select **Add Storage**.
   c. Select from the available storage and allocate capacity to the pool. Click **Assign**.
3. Create compressed, thin-provisioned, deduplicated, or a combination of these volumes in the Data Reduction Pool and map them to hosts by completing these 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 Data Reduction Pool from the list. Compressed, thin-provisioned, and deduplicated volumes and copies must be in Data Reduction Pools.
   - **Volume details**
     Enter the quantity, capacity, and name for the volume or volumes that you are creating.
   - **Capacity savings**
     Select either **None**, **Thin-provisioning**, or **Compression**. For any of these options 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 duplicate data.

   > **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.

4. Click **Create and Map**:

   > **Note:** Select **Create** to create the volumes in the Data Reduction Pool without mapping to hosts. If you want to map volumes to hosts later, select **Hosts** → **Hosts** → **Add Hosts**.

a. 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**.

b. Verify the volume, and then click **Map Volumes**.

Using the command-line interface

To create data reduction on the system, complete these steps:

1. To create a Data Reduction Pool, 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. Data Reduction Pools are created as parent pools only, not child pools.
2. To create a compressed or thin-provisioned volume within a data reduction pool, enter the following command:
   ```
   mkvolume -name name -pool storage_pool_name -size disk_size -compressed
   mkvolume -name name -pool storage_pool_name -size disk_size -thin
   ```
   Where *name* is the name of the new volume, *storage_pool_name* is the name of the data reduction pool, and *disk_size* is the capacity of the volume.
3. 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.

Monitor the physical capacity of Data Reduction Pools in the management GUI by selecting **Pools** → **Pools**. In the command-line interface, use the `lsmdiskgrp` command to display the physical capacity of a data reduction pool.

4.5.2 Adding external MDisks to existing storage pools

If MDisks are being added to the IBM FlashSystem 9100 cluster, it is probably because you you want to provide more capacity. In the Easy Tier enabled pools the storage pool balancing feature guarantees that the newly added MDisks are automatically populated with extents from the other MDisks, so no manual intervention is required to rebalance the capacity across the available MDisks.

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

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

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

Persistent reserve

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

In this condition, rezone the 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 **`lquerypr`** (which is included with SDD and SDDPCM) or the Microsoft Windows SDD Persistent Reserve Tool.

4.5.3 Renaming MDisks

After you discover MDisks, rename them from their IBM FlashSystem 9100 default name. This will help during problem isolation and avoid confusion that can lead to an administrative error by using a naming convention for MDisks that associates the MDisk with the controller and array.

When multiple tiers of storage are on the same IBM FlashSystem 9100 cluster, you might also want to indicate the storage tier in the name. For example, you can use R5 and R10 to differentiate RAID levels, or you can use T1, T2, and so on, to indicate the defined tiers.

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

4.5.4 Removing MDisks from existing storage pools

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

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

When you remove the MDisk made of internal disk drives from the storage pool on IBM FlashSystem 9100 systems, this MDisk is deleted. This process also deletes the array on which this MDisk was built, and converts all drives that were included in this array to *candidate* state. You can now use those disk drives to create another array of different size and raid type, or you can use them as hot spares.

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 a MDisk by using the CLI.

Example 4-3 Listing of volumes that have extents on an MDisk to be deleted

```
IBM_FlashSystem:ITSO:superuser>svcinfo lsmdiskextent mdisk14
id              number_of_extents copy_id
5               16                0
3               16                0
6               16                0
8               13                1
9               23                0
8               25                0
```

DRP restriction: **`lsmdiskextent`** command doesn't provide accurate extent usage for thin provisioned/compressed volumes on DRPs

Specify the **`-force`** flag on the **`svctask rmmdisk`** command, or select the corresponding option in the GUI. Both actions cause IBM FlashSystem 9100 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.

4.5.5 Controlling extent allocation order for volume creation

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

Placing the first extent of a number of volumes on the same MDisk might lead to poor performance for workloads that place a large I/O load on the first extent of each volume or that create multiple sequential streams.

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

DRP restriction: With thin/compressed volumes on DRP it is not possible to check the extent distribution across the MDisks.

4.6 Considerations when using encryption

IBM FlashSystem 9100 systems support optional encryption of data at rest. This support protects against the potential exposure of sensitive user data and user metadata that is stored on discarded, lost, or stolen storage devices. To use encryption on the system, an encryption license is required for every IBM FlashSystem 9100 control enclosure that supports encryption.

> **Note:** Check if you have the required IBM Security Key Lifecycle Manager licenses to hand. Consider redundancy/high availability regarding Key Lifecycle Manager servers.

4.6.1 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 9100 control enclosures 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.

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

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

Several factors must be considered when planning for encryption:

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

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

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

- *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize V8.2.1*, SG24-7933
- *Implementing the IBM Storwize V7000 with IBM Spectrum Virtualize V8.2.1*, SG24-7938

4.6.2 Hardware and software encryption

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

Figure 4-4 shows the encryption placement in the IBM Spectrum Virtualize software stack.

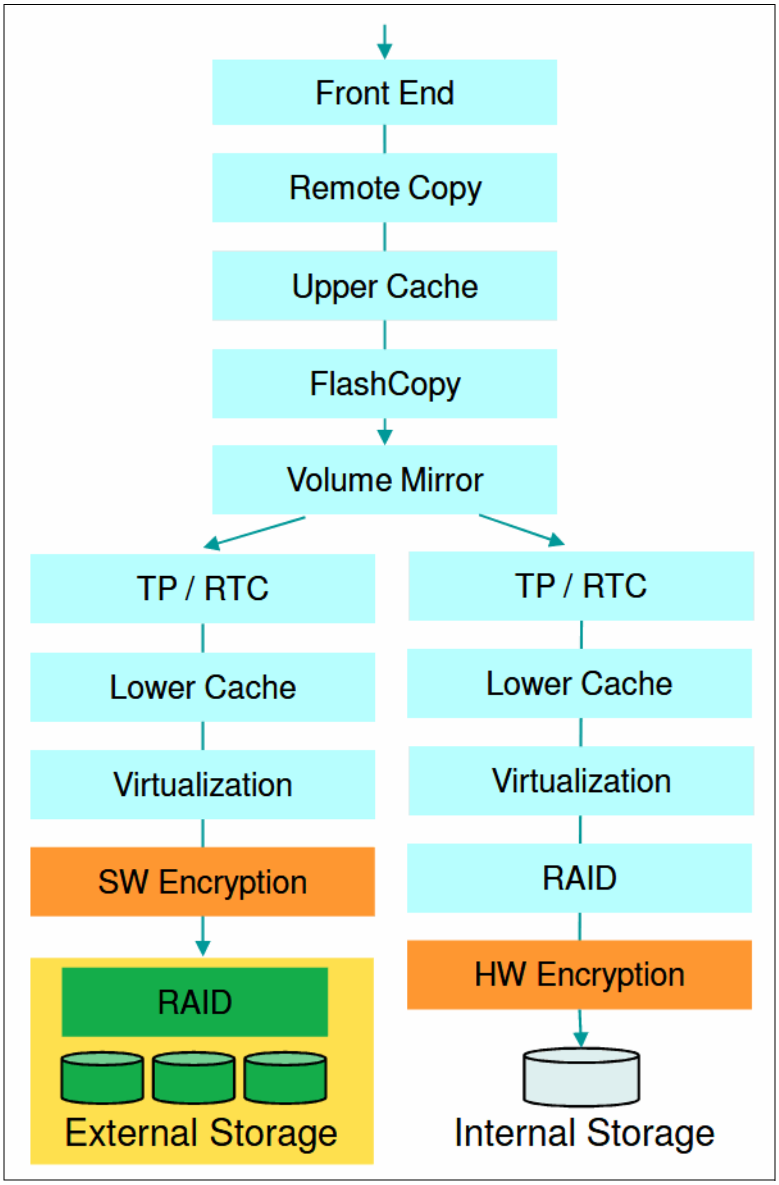


Figure 4-4 Encryption placement in the IBM Spectrum Virtualize and Storwize software stack

Hardware encryption only storage pool

Hardware encryption has the following characteristics:

- Algorithm is built in SAS chip
- No system overhead
- Only available to direct attached SAS disks

- Can only be enabled when you create internal arrays
- Child pools can not be encrypted if the parent storage pool is not encrypted
- Child pools are automatically encrypted if the parent storage pool is encrypted

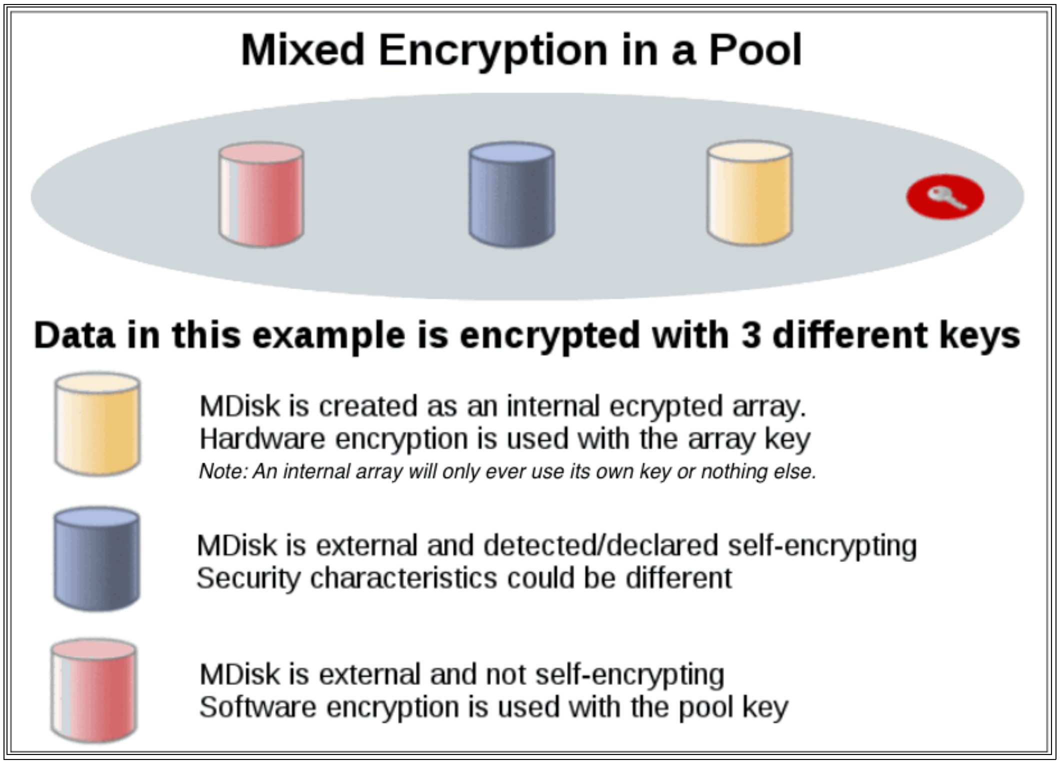
Software encryption only storage pool

Software encryption has the following characteristics:

- Algorithm is running on IBM FlashSystem 9100 canisters
- Uses special CPU instruction set and engines (AES_NI)
- Allows encryption for virtualized external storages, which are not capable of self-encryption
- Potential system overhead
- Only available to virtualized external storages
- Can only be enabled when you create storage pools and child pools made up of virtualized external storages
- Child pools can be encrypted even if the parent storage pool is not encrypted

Mixed encryption in a storage pool

It is possible to mix hardware and software encryption in a storage pool as shown in Figure 4-5.



Mixed Encryption in a Pool

Data in this example is encrypted with 3 different keys

MDisk is created as an internal ecrypted array.
Hardware encryption is used with the array key
Note: An internal array will only ever use its own key or nothing else.

MDisk is external and detected/declared self-encrypting
Security characteristics could be different

MDisk is external and not self-encrypting
Software encryption is used with the pool key

Figure 4-5 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 9100 canisters in the system must support software encryption and have encryption license activated

Note: An encrypted child pool created from an unencrypted parent storage pool reports as unencrypted if the parent pool contains any unencrypted internal arrays. Remove these arrays to ensure that the child pool is fully encrypted. (Child Pools available on Standard Pools only)

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

Note: Always implement encryption on the self-encryption capable back-end storage such as IBM Storwize V7000/V5000, IBM XIV, IBM FlashSystem 900/9100/A9000/A9000R and IBM DS8000 to avoid potential system overhead.

Declare/identify the self-encrypted virtualized external MDisks as encrypted on IBM FlashSystem 9100 by specifying the **`-encrypt`** `option to` **`yes`** with the **`chmdisk`** command as shown in Example 4-4. This configuration is important to avoid IBM Spectrum Virtualize or Storwize trying to encrypt them again.

Example 4-4 Command to declare/identify a self-encrypted MDisk from a virtualized external storage

```
IBM_FlashSystem:Cluster_ITSO:superuser>chmdisk -encrypt yes mdisk0
```

Note: It is important to declare/identify the self-encrypted MDisks from a virtualized external storage before create a encrypted storage pool or child pool on IBM FlashSystem 9100.

4.6.3 Encryption at rest with USB keys

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

USB flash drives have the following characteristics:

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

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

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

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

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

  For the most secure operation, do not keep the USB flash drives inserted into the nodes on the system. However, this method requires that you manually insert the USB flash drives that contain copies of the encryption key in the nodes during operations that the system requires an encryption key to be present.

  USB flash drives that contain the keys must be stored securely to prevent theft or loss. During operations that the system requires an encryption key to be present, the USB flash drives must be inserted manually into each node so data can be accessed. After the system completes unlocking the drives, the USB flash drives must be removed and stored securely to prevent theft or loss.

4.6.4 Encryption at rest with key servers

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

Key servers

Key servers have the following characteristics:

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

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

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

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

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

Replication must complete before keys can be used on the system. You can either schedule automatic replication or complete it manually with IBM Security Key Lifecycle Manager. During replication, key servers are not available to distribute keys or accept new keys. The time a replication completes on the IBM Security Key Lifecycle Manager depends on the number of key servers that are configured as clones and the amount of key and certificate information that is being replicated.

The IBM Security Key Lifecycle Manager issues a completion message when the replication completes. Verify that all key servers contain replicated key and certificate information before keys are used on the system.

Recommendations for key server configuration

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

Transport Layer Security

Define the IBM Security Key Lifecycle Manager to use Transport Layer Security version 2 (TLSv2). The default setting on IBM Security Key Lifecycle Manager is TLSv1, but the IBM Spectrum Virtualize and Storwize only support version 2. On the IBM Security Key Lifecycle Manager set the value to `SSL_TLSv2`, which is a set of protocols that includes TLSv1.2. Example 4-5 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`.

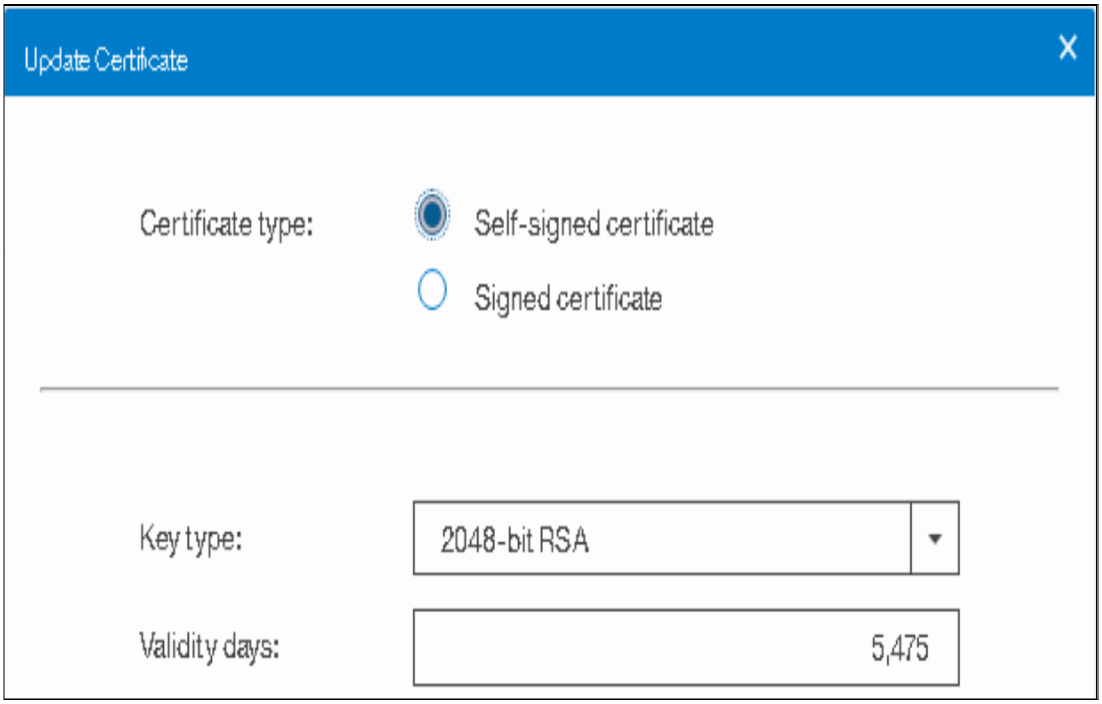
Example 4-5 Example of a SKLMConfig.properties configuration file

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

You can adjust the validity period to comply with specific security policies and always match the certificate validity period on IBM Spectrum Virtualize and IBM Security Key Lifecycle Manager server. A mismatch will cause certificate authorization error and lead to unnecessary certificate exchange. Figure 4-6 shows the default certificate type and validity period on IBM Spectrum Virtualize.



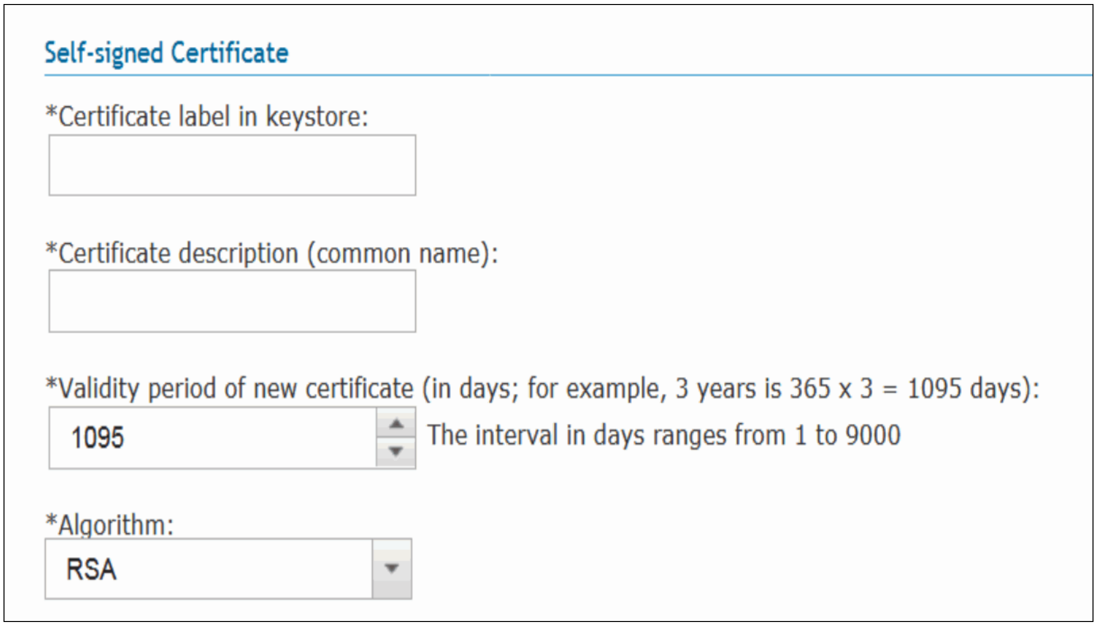
Certificate type: ● Self-signed certificate ○ Signed certificate

Key type: 2048-bit RSA

Validity days: 5,475

Figure 4-6 Update certificate on IBM Spectrum Virtualize and Storwize

Figure 4-7 shows the default certificate type and validity period on IBM Security Key Lifecycle Manager server.



Self-signed Certificate

*Certificate label in keystore:

*Certificate description (common name):

*Validity period of new certificate (in days; for example, 3 years is 365 x 3 = 1095 days):

1095 The interval in days ranges from 1 to 9000

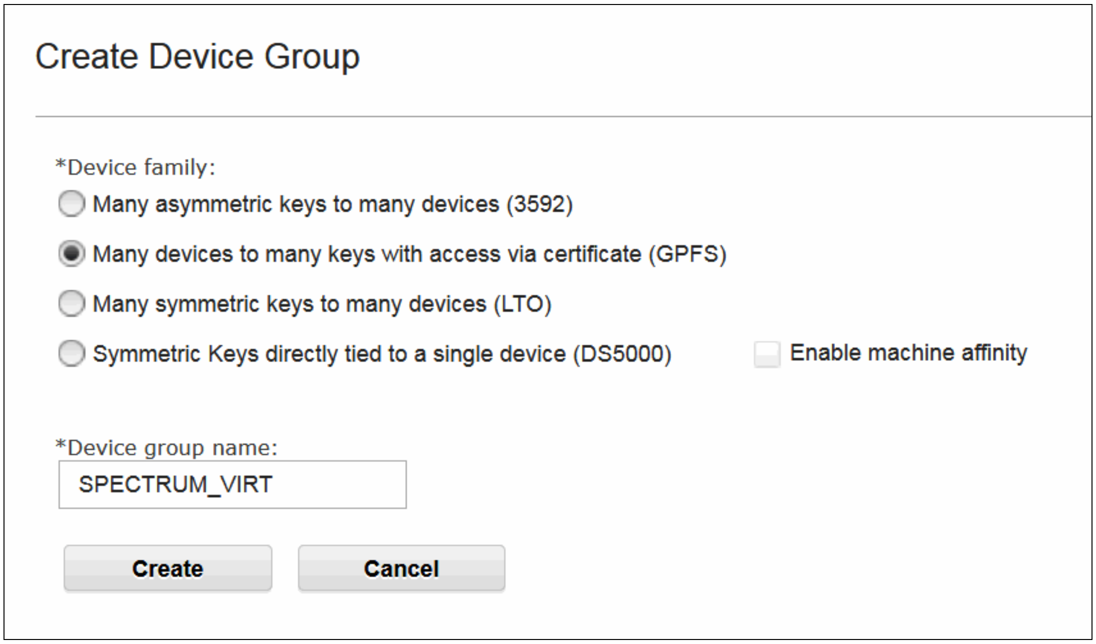
*Algorithm:

RSA

Figure 4-7 Create self-signed certificate on IBM Security Key Lifecycle Manager server

Device group configuration

The SPECTRUM_VIRT device group is not pre-defined on IBM Security Key Lifecycle Manager, it must be created based on an IBM GPFS device family as shown in Figure 4-8.



Create Device Group

*Device family:

○ Many asymmetric keys to many devices (3592)

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

○ Many symmetric keys to many devices (LTO)

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

*Device group name:

SPECTRUM_VIRT

Figure 4-8 Create device group for IBM Spectrum Virtualize or Storwize

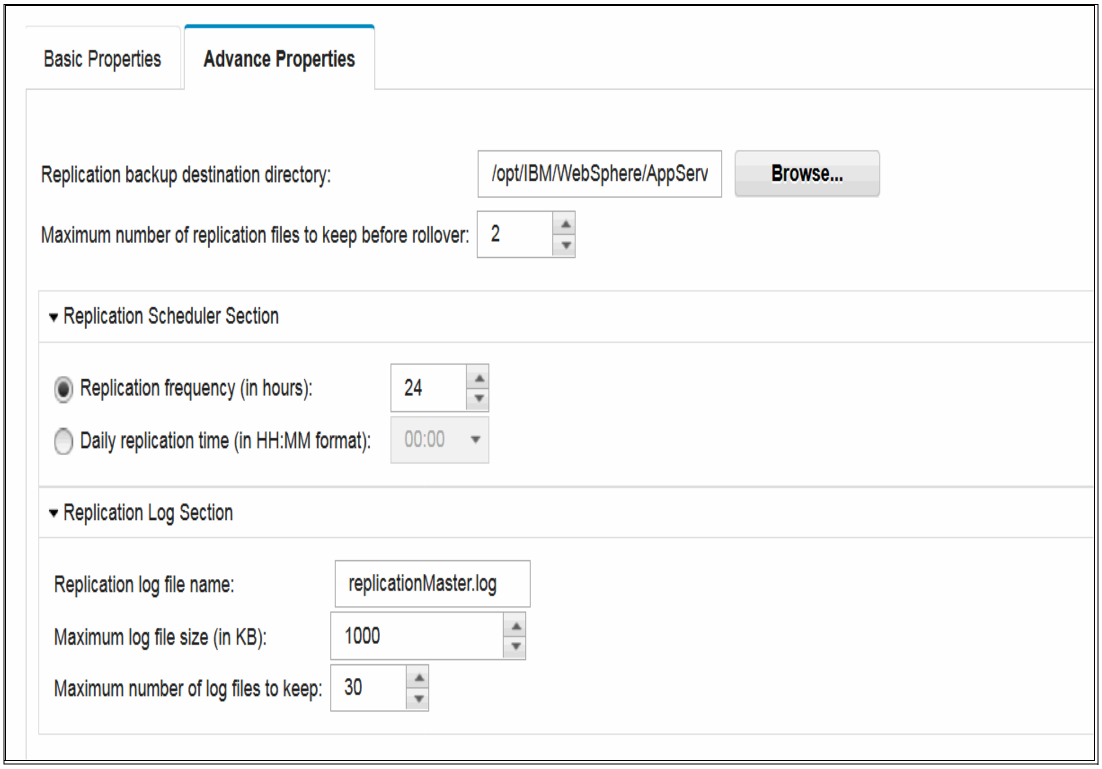
By default, IBM Spectrum Virtualize and Storwize has the `SPECTRUM_VIRT` pre-defined in the encryption configuration wizard, and `SPECTRUM_VIRT` contains all of the keys for the managed IBM FlashSystem 9100 systems. However, It is possible to use different device groups as long as they are GPFS device family based. For example, one device group for each environment (Production or DR). Each device group maintains its own key database, and this approach allows more granular key management.

Clone servers configuration management

The minimum replication interval on IBM Security Key Lifecycle Manager is one hour, as shown in Figure 4-9 on page 91. It is more practical to perform backup and restore or manual replication for the initial configuration to speed up the configuration synchronization.

Also, the rekey process creates a new configuration on the IBM Security Key Lifecycle Manager server, and it is important not to wait for the next replication window but to manually synchronize the configuration to the additional key servers (clones), otherwise, an error message will be generated by the IBM FlashSystem 9100 system indicating the key is missing on the clones.

Figure 4-9 shows the replication interval.



Replication backup destination directory: /opt/IBM/WebSphere/AppServ

Maximum number of replication files to keep before rollover: 2

● Replication frequency (in hours): 24

○ Daily replication time (in HH:MM format): 00:00

Replication log file name: replicationMaster.log

Maximum log file size (in KB): 1000

Maximum number of log files to keep: 30

Figure 4-9 SKLM Replication Schedule

Example 4-6 shows an example of manually triggered replication.

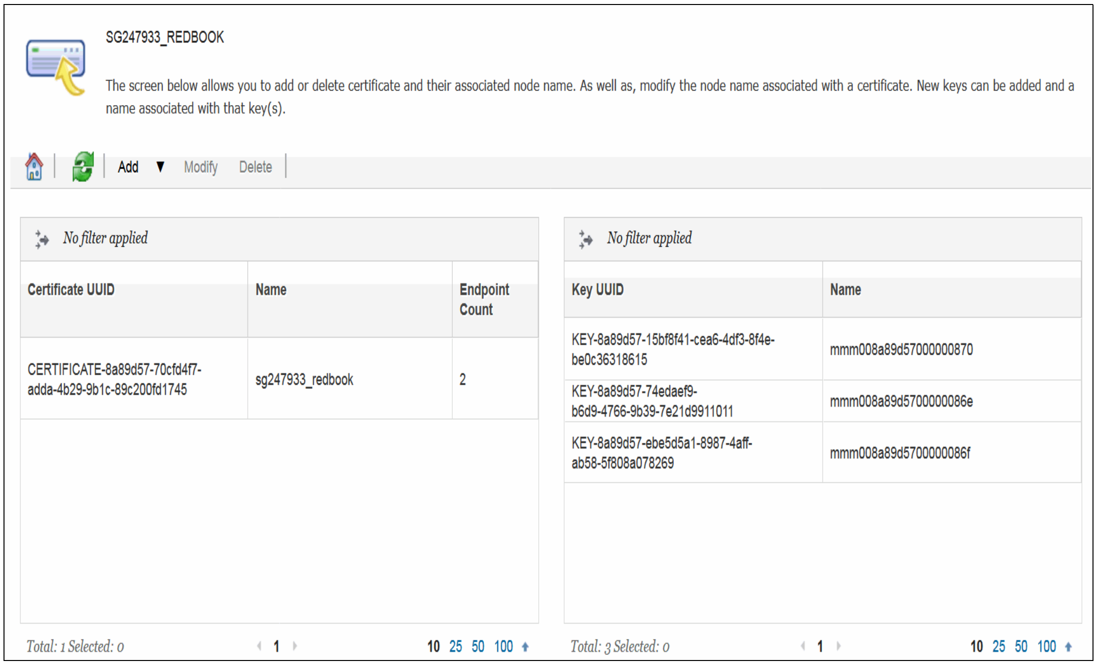
Example 4-6 Manually triggered replication

```
/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 9100 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 4-10 shows the keys associated with a device group. In this example, the SG247933_REDBOOK device group contains one encryption enabled Storwize V7000 system, and it has three associated keys. Only one of the keys is activated, and the other two were deactivated after the rekey process.



SG247933_REDBOOK

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

No filter applied

Certificate UUID	Name	Endpoint Count
CERTIFICATE-8a89d57-70cfd4f7-adda-4b29-9b1c-89c200fd1745	sg247933_redbook	2

Total: 1 Selected: 0

No filter applied

Key UUID	Name
KEY-8a89d57-15bf8f41-cea6-4df3-8f4e-be0c36318615	mmm008a89d57000000870
KEY-8a89d57-74edaef9-b6d9-4766-9b39-7e21d9911011	mmm008a89d5700000086e
KEY-8a89d57-ebe5d5a1-8987-4aff-ab58-5f808a078269	mmm008a89d5700000086f

Total: 3 Selected: 0

Figure 4-10 Keys associated to a device group

Example 4-7 shows an example to check the state of the keys.

Example 4-7 Verify key state

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

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

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

uuid = KEY-8a89d57-74edaef9-b6d9-4766-9b39-7e21d9911011
```

```
alias = mmm008a89d5700000086e
key algorithm = AES
key store name = defaultKeyStore
key state = DEACTIVATED
creation date = 17/11/2017, 20:07:19 Greenwich Mean Time
expiration date = 17/11/2017, 23:18:37 Greenwich Mean Time

wsadmin>print AdminTask.tklmKeyList('[-uuid
KEY-8a89d57-ebe5d5a1-8987-4aff-ab58-5f808a078269]')
CTGKM0001I Command succeeded.

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

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

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

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

4.7 Data Reduction Pools Best Practices

In this section we describe the DRP planning and implementation best practices relating to IBM FlashSystem 9100 systems.

4.7.1 IBM FlashSystem 9100 and DRP

The IBM FlashSystem 9100 has 24 x 2.5” slots in the control enclosure to populate with NVMe storage. Both IBM Flash Core Modules (FCM) and off-the-shelf NVMe drives are supported. IBM FCM drives use inline hardware compression to reduce the amount of physical space required, and these are available in 4.8 TB, 9.6 TB, and 19.2 TB sizes. IBM FCM drives contain a fixed amount of space for compression metadata. The maximum effective capacity is the amount of data it takes to fill the metadata space. Table 4-3 shows the maximum effective capacity and the resulting data compression ratio per module by module size.

Table 4-3 IBM FCM data compression ratio and effective capacity

FCM module size	Max effective capacity	Max compression ratio
4.8TB	21.99TB	4.5:1
9.6TB	21.99TB	2.3:1
19.2TB	43.98TB	2.3:1

Note that the compression ratio reported in Table 4-3 is the maximum achievable considering the effective capacity available. Depending on the pool filling, the actual compression ratio can be higher. To get an estimation of the compression ratio for a specific workload, see “Tools” on page 64.

When providing NVMe attached flash drives capacity for DRP there are considerations that must be taken into account.

The main point to consider is whether the data is a good candidate for deduplication or not. Tools are available to provide some estimation of deduplication ratio (see “Tools” on page 64). Let’s consider first DRP configurations with IBM **FCM drives**.

- **Data is a good candidate for deduplication**. 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.
- **If you are not doing deduplication** (as its not a good candidate for dedupe), you might use Traditional Pools (instead of DRP with FCM), and let the FCM hardware do the compression, as the performance will be twice as fast. Cache miss of a volume in a Traditional Pool is faster then cache hit of a DRP volume.

With standard off-the-shelf **NVMe drives** that do not support in-line compression, similar considerations apply.

- **Data is a good candidate for deduplication**. In this case the recommendation is to use compressed and deduplicated volume type. The DRP compression technology has more than enough compression bandwidth for these purposes, so we might as well always compress.
- **Data is not a good candidate for deduplication**. In this case the recommendation is to use compressed volume type only. Again the internal compression technology provides enough compression bandwidth for these purposes.

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 Storwize Family (V7000 and V5100) and IBM FlashSystem 9100.
- With SSD and HDD, use DRP, deduplicate if applicable with IBM Storwize Family (V7000 and V5100) and IBM FlashSystem 9100.
- With Storwize V5030 you can use compression if not expecting high number of IOps, SW compression might limit performance benefits of SSD.
- Do not compress encrypted data
- Identify and use compressible data only. Different data types have different compression ratios, and it is important to determine the compressible data currently on your system. 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 “Is your data compressible?” on page 63 for more details.
- Avoid using any client, file system, or application based-compression with the system compression.
- Avoid using compression on IBM SAN Volume Controller and Storwize systems and virtualized external storage at same time (DRP over DRP).

4.7.2 IBM FlashSystem 9100 as external storage

IBM FlashSystem 9100 offers the same external storage virtualization capability as the IBM SAN Volume Controller and the Storwize family. However, the IBM FlashSystem 9100 can be configured as an external storage controller to a SAN Volume Controller cluster as well.

When planning the configuration of the IBM FlashSystem 9100 that will be virtualized by SAN Volume Controller, the following recommendations apply:

- Set up your RAID protection as Distributed RAID 6 (DRAID6) for greater protection and performance.
- Configure a minimum of 16 volumes that are provisioned to the SAN Volume Controller, but more is better. Approximately 30 volumes unlock the maximum performance.

  On SVC use a Spectrum Virtualize code level that manages the reporting of the back-end physical space. The minimum Spectrum Virtualize level that supports this feature is V8.1.1. For the supported and recommended level of Spectrum Virtualize on SVC managing IBM FlashSystem 9100 as back-end, refers to IBM System Storage Interoperation Center (SSIC) at the following link.

Regarding the DRP usage, many options are available, as shown in Figure 4-11.

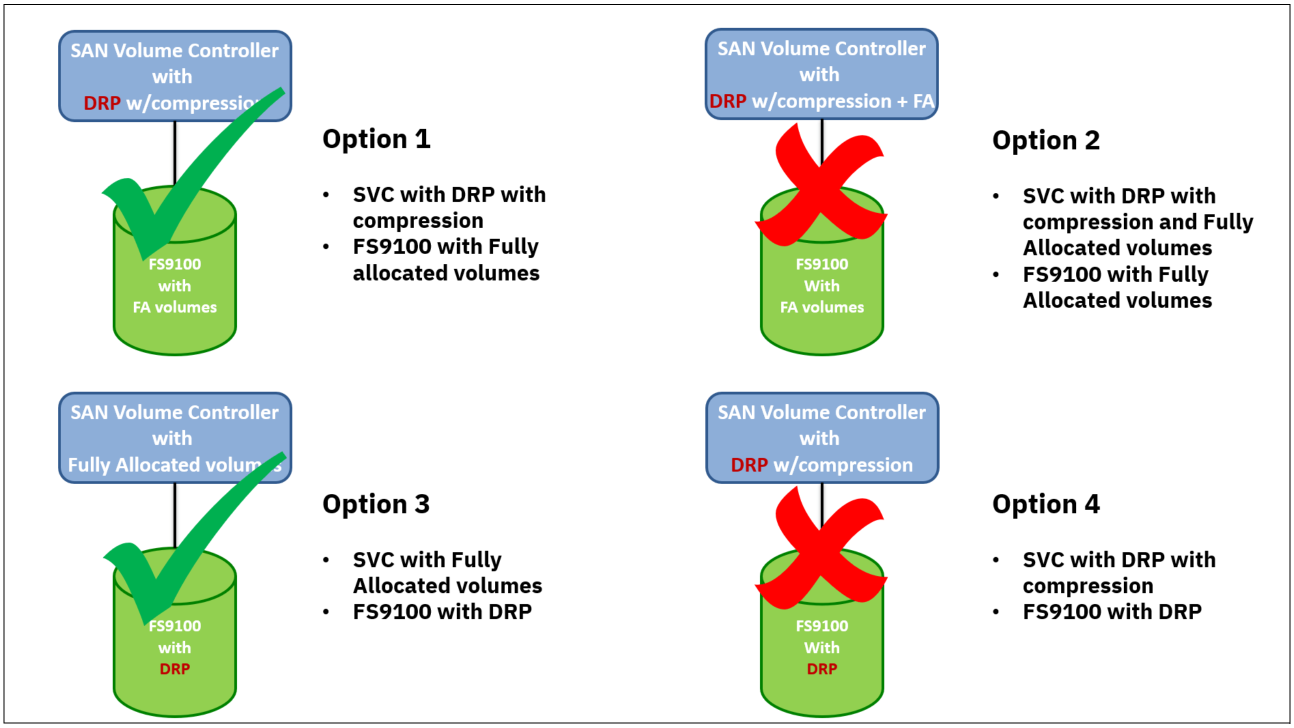

SAN Volume Controller with DRP w/compression
FS9100 with FA volumes
Option 1
• SVC with DRP with compression
• FS9100 with Fully allocated volumes
SAN Volume Controller with DRP w/compression + FA
FS9100 With FA volumes
Option 2
• SVC with DRP with compression and Fully Allocated volumes
• FS9100 with Fully Allocated volumes
SAN Volume Controller with Fully Allocated volumes
FS9100 with DRP
Option 3
• SVC with Fully Allocated volumes
• FS9100 with DRP
SAN Volume Controller with DRP w/compression
FS9100 With DRP
Option 4
• SVC with DRP with compression
• FS9100 with DRP

Figure 4-11 DRP options with SVC and IBM FlashSystem 9100

Some of these options are recommended, others have to be avoided. In particular:

- Option 1: SVC uses DRP with compression/deduplication enabled and back-end IBM FlashSystem 9100 uses fully allocated volumes to provide capacity to SVC. This is the recommended configuration. With this option, DRP at the SVC level can be used to plan for deduplication and capacity optimization. Furthermore, DRP at the top level provides best application capacity reporting (volume written capacity). In configurations where IBM FlashSystem 9100 volumes are inline compressed by FCMs, do not overprovision the physical space (assume 1:1 compression in the back-end).

- Option 2: SVC uses DRP with both compression enabled volumes and allocated volumes, while the IBM FlashSystem 9100 uses fully allocated volumes to provide capacity to SVC. This is not a recommended configuration. The problem with this configuration is that it is very difficult to understand the physical capacity use of the fully allocated volumes.
- Option 3: SVC uses fully allocated volumes while the IBM FlashSystem 9100 uses DRP with compressed and deduplicated volumes to provide capacity to SVC. This is a valid configuration, even not the preferred. The problem with this configuration is that SVC can report physical use but does not manage to avoid out-of-space. This option requires careful monitoring of capacity to not run out of space.
- Option 4: SVC uses DRP with compression/deduplication enabled and the IBM FlashSystem 9100 uses DRP with compressed and deduplicated volumes to provide capacity to SVC. This is a configuration to be avoided since it creates two levels of I/O amplification on metadata. Furthermore, DRP at the bottom layer provides no benefit.

Note: The usage of DRP with compressed/deduplicated volumes as SVC back-end requires substantial resources in terms of cache and CPU. To avoid performance issues with entry level Storwize systems like Storwize V5010 and Storwize V5010E, consider using fully allocated volumes unless data reduction is strictly required.

4.7.3 DRP provisioning considerations

Consider the following practices when planning for DRP implementation.

DRP restrictions

There are some important restrictions that must be taken into consideration when planning for a DRP implementation:

- Maximum number of supported DRP is 4
- Child Pools are not supported in DRP
- VVols are not 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 thin/compressed volumes
- Image and Sequential mode VDisk are not supported in DRP
- No extent level migration is allowed
- A maximum of 128,000 extents per Customer Data Volume - per I/O Group
- Thus the Pool extent size dictates the maximum physical capacity in a pool - after data reduction.
- Recommended Pool size is at least 20 TB
- Lower than 1 PB per I/O group
- Use 4 GB extent size or above
- Your Pool should be no more than 85% occupied

Further, the following considerations apply to DRP:

- The real/used/free/tier capacity is not reported per volume for DRP volumes, only information at a pool level is available
- Cache mode is always read-write on thin/compressed 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 (see 4.1.2, "Data Reduction Pools" on page 59) and specifically by the size of the data volume. Currently the maximum number of extent supported for a data volume is 128 KB. Note that according to Figure 4-1 on page 61 there is one data volume per pool. Table 4-4 shows the maximum size per pool by extent size and I/O group number.

Table 4-4 Pool size by extent size and I/O 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

Be aware that it cannot be changed after the pool is created, so it is recommended to carefully plan the extent size according to the environment capacity requirements. For most of the configurations we recommend 4 GB extent size for DRP as a general rule of thumb.

Pool capacity requirements

A minimum capacity must also be provisioned in a DRP to provide capacity for the internal metadata structures. Table 4-5 shows the minimum capacity required by extent size and I/O group number.

Table 4-5 Minimum recommended pool size by extent size and I/O 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 (GC) process runs all the time to reclaim the unused space to optimize the extents usage.

The GC process main steps are summarized in Figure 4-12.

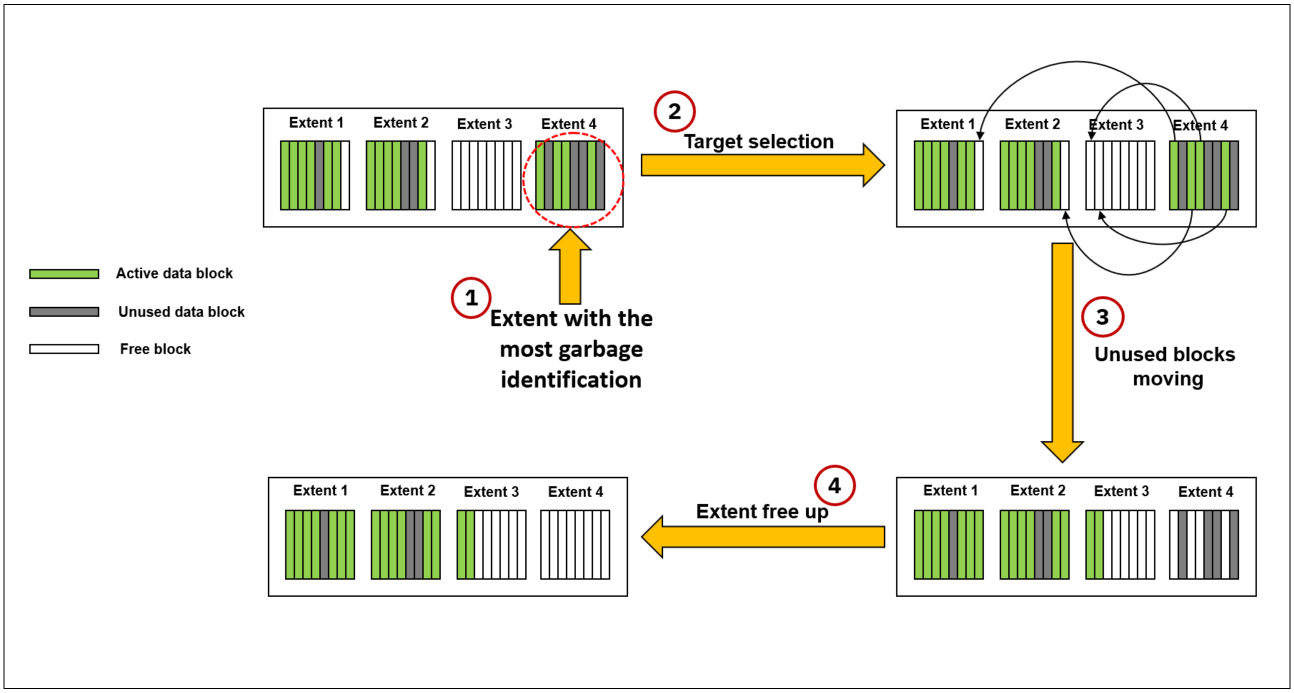


Figure 4-12 Garbage Collection main steps

Where:

1. Extent with the most garbage identification. A table containing the active/unused/free blocks information at extent level is used to identify the extent with the most unused block (or less active data).
2. Target selection. The targets for active blocks are identified among the available extents.
3. Unused blocks moving. The active blocks are moved leaving only unused blocks in the source extent.
4. Extent free-up. The unused block in the extent are discarded and the extent is set to free.

As the amount of unused data tends to increase over time, the amount of active data the GC process needs to move to free-up extents decreases accordingly. For this reason GC delays the moving activity to try to increase the chance of having extents full of garbage and therefore minimizing the active data movement. This behavior, while on the one hand optimizes the bandwidth usage for the active data movement, on the other hand it leaves unused capacity unavailable for new allocations waiting to be reclaimed. If the amount of free space in the DRP decreases below a certain threshold, GC starts aggressively to move active data in order to free-up extents as much as possible.

This process then requires a certain amount of free space to work efficiently. For this reason it is recommended to keep at least 15% of free space in a DRP pool. Consider leaving 20% free space if being virtualized behind SVC or another virtualization engine. See the support document for more details.

4.7.4 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 for instance the following scenarios:

- IBM FlashSystem 9100 installation requiring 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 (see “DRP restrictions” on page 96). In this case the standard pool will be used for VVols environments only, while DRP will be used for the other environments. Note that some data reduction capability could be achieved also for the VVols standard pool by using the in-line data compression provided by the IBM FCMs.
- IBM FlashSystem 9100 installations requiring an external pool for image mode volumes and data reduction capabilities for other environments. Also, this scenario requires the definition of both standard and DRP pools because of the restriction of DRP regarding image mode volumes (see “DRP restrictions” on page 96). In this case the standard pool will be used for image mode volumes only, and optionally with write cache disabled if needed for the back-end native copy services usage (see Chapter 6, “Copy services” on page 141). DRP will be used for all the other environments.
- IBM FlashSystem 9100 installations using as an external pool a Storwize system that uses DRP capabilities. In this scenario the external pool must be a standard pool, as recommended in 4.7.2, “IBM FlashSystem 9100 as external storage” on page 95. In this case the internal storage can be defined in a separate DRP enabling the data reduction capabilities if needed.
- IBM FlashSystem 9100 installation requiring more than 4 pools.

4.7.5 Data migration with DRP

As mentioned in “DRP restrictions” on page 96 it is not supported to make an extent level migration, such as migrate volume or migrate extent functions, from and to a DRP. For an existing IBM FlashSystem 9100 configuration that wants to move data from or to a DRP, basically two options are available: host based migrations, Volume Mirroring based migrations.

Host based migration

The host based migrations exploit operating system features or software tools running on the hosts to move data concurrently with normal host operations. VMware vMotion and AIX Logical Volume Mirroring are just two examples of such features. When using this approach a certain amount of capacity on the target pool is required to provide the migration target volumes. The process can be summarized as follows:

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 instance we can migrate two 2TB VMware datastore volumes in one single 4TB 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/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 9100.

When migrating data to a DRP consider the following options:

- Migrate directly to a compressed/deduplicated volumes. With this option, the migration duration mainly depends on the host migration throughput capabilities. Consider that the target volumes are subject to very high write workload that can consume a lot of resources due to the compression and deduplication tasks. To avoid any potential performance impact on the existing workload, try to limit the migration throughput at host level, or, if this is not possible, implement the throttling function at volume level.
- Migrate first to fully allocated volumes and then convert them to compressed/deduplicated volumes. Also with this option, the migration duration mainly depends on the host capabilities, but usually more throughput can be sustained because no compression and deduplication overhead. 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 new volumes manually. The Volume Mirroring migration is basically a three step activity:

1. Add a new volume copy on the DRP specifying the data reduction features desired.
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 volume level specifying the Mirror Sync Rate parameter, and therefore, in case of performance impacts, the migration speed can be lowered or even suspended.

Note: Volume Mirroring supports only two copies of a volume. In case of configuration already using both copies, one of the copies must be removed first before to start the migration.

Note: The volume copy split of a Volume Mirror in a different I/O Group is not supported for DRP thin/compressed volumes.

5

Volumes

In IBM FlashSystem 9100, a volume is logical disk provisioned out of a storage pool and is recognized by a host with a unique identifier (UID) field. This chapter describes the several types of volumes and guidance for managing the properties.

> **Important, IBM FlashSystem 9200:** On 11th February 2020 IBM announced the arrival of the IBM FlashSystem 9200 to the family. This book was written specifically for IBM FlashSystem 9100, however most of the general principles will apply to the IBM FlashSystem 9200.
>
> If you are in any doubt as to their applicability to the FlashSystem 9200 then you should work with your local IBM representative.
>
> This book will be updated to include FlashSystem 9200 in due course.
>
> The Flashsystem 9200 product guide is available at:
>
> IBM FlashSystem 9200 Product Guide

This chapter includes the following sections:

- Overview of volumes
- Guidance for creating volumes
- Thin-provisioned volumes
- Mirrored volumes
- HyperSwap volumes
- VMware Virtual volumes
- Cloud volumes
- Volume migration
- Preferred paths to a volume
- Changing the preferred node moving a volume between I/O groups
- Volume throttling
- Volume cache mode
- Additional considerations

5.1 Overview of volumes

A volume is a logical disk presented to a host by an I/O Group, and within that group there is a preferred node that is a node which, by default, will serve I/O requests to the volume.

There are different types of volumes:

- Standard volumes
- Thin-provisioned volumes (deduplication is optional)
- Compressed volumes (deduplication is optional)
- Mirrored volumes
- HyperSwap volumes
- VMware Virtual volumes (VVols)
- Cloud volumes

In this chapter, details on each volume type will be described.

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 duplicate data, and save more capacity.

The two types of standard volumes are determined by how the extents are allocated from the storage pool. The *striped-mode volume* has extents that are allocated from each managed disk (MDisk) in the storage pool in a round-robin way. With a *sequential-mode volume*, extents are allocated sequentially from an MDisk.

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

A Cloud volume is any volume that is enabled for transparent cloud tiering. When transparent cloud tiering is enabled on a volume, point-in-time copies, or snapshots, can be created and copied to cloud storage that is provided by a cloud service provider.

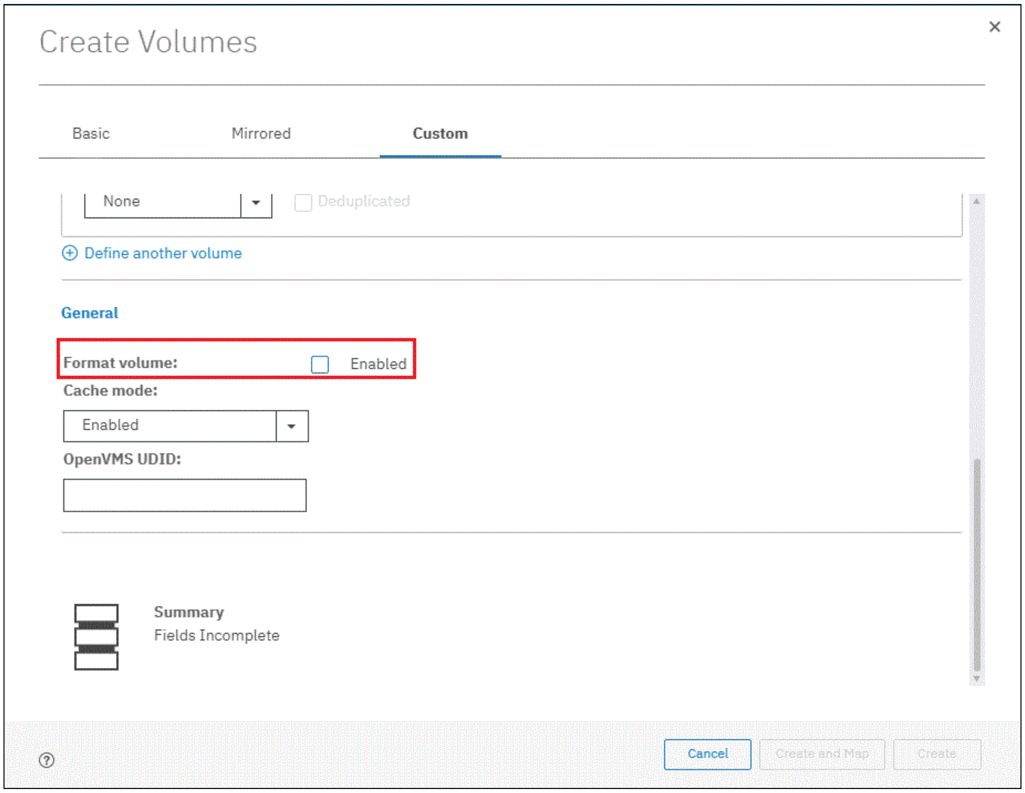
5.2 Guidance for creating volumes

When you create volumes, see the following guidelines. More information about each of these items can be found in the next section of this chapter:

- Consider the naming rules before you create volumes. We recommend that it is easier to assign the correct names when the volume is created than to modify them afterwards.
- Choose the type of volume that you will create. First, decide whether fully allocated (standard volumes) or thin-provisioned. And then, if you decide to create a thin-provisioned volume, choose compression and deduplication options.
  - If you decide to create a fully allocated volume, you have to take into consideration that volumes are automatically formatted through the quick initialization process right after their creation.

    This quick initialization is a background process that writes zeros to all the blocks of a fully allocated volume. This does not affect the usage of the volume. It is possible to use volumes immediately.

If you don't want the volume to be automatically formatted after its creation, you can disable the format option on the Custom tab of the volume creation window by clearing the **Format volumes** box shown in Figure 5-1.



Format volume: □ Enabled

Cache mode: Enabled

OpenVMS UDID:

Figure 5-1 Volumes format option

And you can also create volumes by using the CLI. Example 5-1 shows the command to disable auto formatting option with the **-nofmtdisk** parameter.

Example 5-1 Volume creation without auto formatting option

```
superuser>mkvdisk -name TESTVOL01 -mdiskgrp Pool0 -size 1 -unit gb -vtype
striped -iogrp io_grp0 -nofmtdisk
Virtual Disk, id [38], successfully created
superuser>lsvdisk TESTVOL01
id 10
name TESTVOL01
IO_group_id 0
IO_group_name io_grp0
status online
mdisk_grp_id 0
mdisk_grp_name Pool0
capacity 1.00GB
type striped
formatted no
formatting no
.
lines removed for brevity
```

- Remember that when you create a volume, depending on the volume size, it takes a long time to completely format volumes, and those actions like moving, expanding, shrinking, or adding a volume copy are disabled when the specified volume is initializing. For reference, the initialization of a 1 TB volume can take more than 120 hours to complete with the default syncrate value 50, or around 4 hours if you manually set the syncrate to 100.
- If you decide to create a thin-provisioned volume, with compression and deduplication enabled or not, you have to be careful for out of space in the volume and pool where the volume is created. You should set the warning threshold in the pools containing thin-provisioned volumes, and you can set it in the volume too.

  If you do not want to worry about monitoring volume capacity, it is highly recommended to enable the **autoexpand** option. Also, when you create a thin-provisioned volume, you must specify the space which is initially allocated to it (**-rsize** option in the CLI) and the grain size.

  By default, **rsize** (or real capacity) is set to 2% of the volume virtual capacity, and grain size is 256 KiB. These default values, with autoexpand enabled and warning disabled options will work in most scenarios. There are some cases that you might consider using different values to suit your environment.

  Example 5-2 shows the command to create a volume with the parameters mentioned previously.

  Example 5-2 Thin-provisioned volume creation

  ```
  superuser>mkvdisk -name TESTVOL02 -mdiskgrp Pool0 -size 100 -unit gb -vtype
  striped -iogrp io_grp0 -rsize 2% -autoexpand -warning 0 -grainsize 256
  Virtual Disk, id [40], successfully created
  superuser>lsvdisk TESTVOL02
  id 11
  name TESTVO02
  .
  lines removed for brevity
  .
  capacity 100.00GB
  .
  lines removed for brevity
  .
  used_capacity 0.75MB
  real_capacity 2.02GB
  free_capacity 2.01GB
  overallocation 4961
  autoexpand on
  warning 0
  grainsize 256
  se_copy yes
  .
  lines removed for brevity
  ```

- Each volume has an I/O group and preferred node that balances the load between nodes in the I/O group. Therefore, you should consider balancing volumes across the I/O groups in the cluster.

  In configurations with many attached hosts, where it is not possible to zone a host to multiple I/O groups, you might not be able to choose the specific I/O group to attach the volumes. The volume must be created in the I/O group to which its host belongs.

Tip: Migrating volumes across I/O groups can be a disruptive action. Therefore, specify the correct I/O group at the time the volume is created.

Also, when a volume is being created, it is possible to define a list of I/O groups in which a volume can be accessible to hosts. It is recommended that a volume is accessible to hosts only by the caching I/O group. You can have more than one I/O group in the access list of a volume in some scenarios with specific requirements, like when a volume is being migrated to another I/O group.

- By default, the *preferred node*, which owns a volume within an I/O group, is selected for a *load balancing* basis. Even though it is not easy to expect the workload when the volume is created, you should distribute the workload evenly on each node within an I/O group.
- With the exception of a few cases, cache mode of a volume should be set to readwrite. More details can be found in section 5.12, “Volume cache mode” on page 135.
- The maximum number of volumes per I/O group and system is described in Table 5-1.

Table 5-1 Maximum number of volumes in IBM FlashSystem 9100

Volume type	Maximum number	Comments
Volumes (VDisks) per system	10,000	Each basic volume uses 1 VDisk, each with one copy.
HyperSwap volumes per system	1,250	Each HyperSwap volume uses 4 VDisks, each with one copy, 1 active-active remote copy relationship and 4 FlashCopy mappings.
Volumes per I/O group (volumes per caching I/O group)	10,000	
Volumes accessible per I/O group	10,000	
Thin-provisioned (space-efficient) volume copies in regular pools per system	8,192	
Compressed volume copies in regular pools per system	2,048	Maximum requires a system containing four control enclosures; see the Compressed volume copies in regular pools per I/O group limit below
Compressed volume copies in regular pools per I/O group	512	With 32 GB Cache upgrade and 2nd Compression Accelerator card installed.
Compressed volume copies in data reduction pools per system	-	No limit is imposed here beyond the volume copy limit per data reduction pool
Compressed volume copies in data reduction pools per I/O group	-	No limit is imposed here beyond the volume copy limit per data reduction pool
Deduplicated volume copies in data reduction pools per system	-	No limit is imposed here beyond the volume copy limit per data reduction pool
Deduplicated volume copies in data reduction pools per I/O group	-	No limit is imposed here beyond the volume copy limit per data reduction pool

Volume type	Maximum number	Comments
Volumes per storage pool	-	No limit is imposed beyond the volumes per system limit
Fully-allocated volume capacity	256 TB	Maximum size for an individual fully-allocated volume. Maximum size is dependent on the extent size of the Storage Pool. Comparison Table: Maximum Volume, MDisk and System capacity for each extent size.
Thin-provisioned (space-efficient) per-volume capacity for volumes copies in regular and data reduction pools	256 TB	Maximum size for an individual thin-provisioned volume. Maximum size is dependent on the extent size of the Storage Pool. Comparison Table: Maximum Volume, MDisk and System capacity for each extent size.
Compressed volume capacity in regular pools	96 TB	Maximum size for an individual compressed volume Maximum size is dependent on the extent size of the Storage Pool. Comparison Table: Maximum Volume, MDisk and System capacity for each extent size.

- The pool extent size does not affect the overall storage performance. A volume occupies an integer number of extents, but its length does not need to be an integer multiple of the extent size. And the length does need to be an integer multiple of the block size. Any space left over between the last logical block in the volume and the end of the last extent in the volume is unused.

  A small extent size is used to minimize this unused space, and also to have a finer granularity of the volume space that is occupied on the underlying storage controller. On the other hand, you have to consider the best extent size for your storage pools considering the back-end storage.

  > **Important:** Volume migration (using the `migratevdisk` command) between storage pools requires that both (source and destination) pools have the same extent size.

5.3 Thin-provisioned volumes

A thin-provisioned volume presents a different capacity to mapped hosts than the capacity that the volume consumes in the storage pool. The system supports thin-provisioned volumes in both *standard pools* and *Data Reduction Pools (DRP).*

Figure 5-2 shows the basic concept of a Thin-provisioned volume.

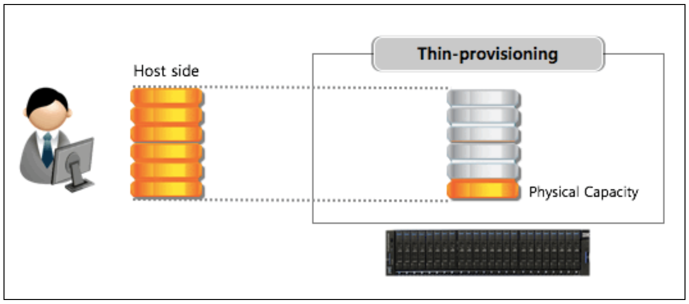


Figure 5-2 Thin-provisioned volume

Also, you can see the different types of volumes in DRP described in Figure 5-3.

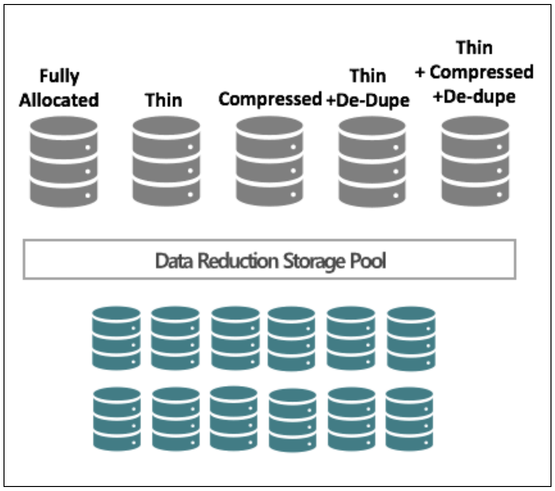


Figure 5-3 Different volume types in DRP

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. However, in DRP, all the benefits of thin-provisioning are available to all the volumes that are assigned to the pool. For the thin-provisioned volumes in DRP, you can configure compression and data deduplication on these volumes, increasing 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. It is redistributed automatically. Standard pools do not have these functions.

Before describing the different volume types in detail, we will describe the concept of thin-provisioned volumes again.

Figure 5-4 shows a conceptual diagram of thin-provisioned volumes.

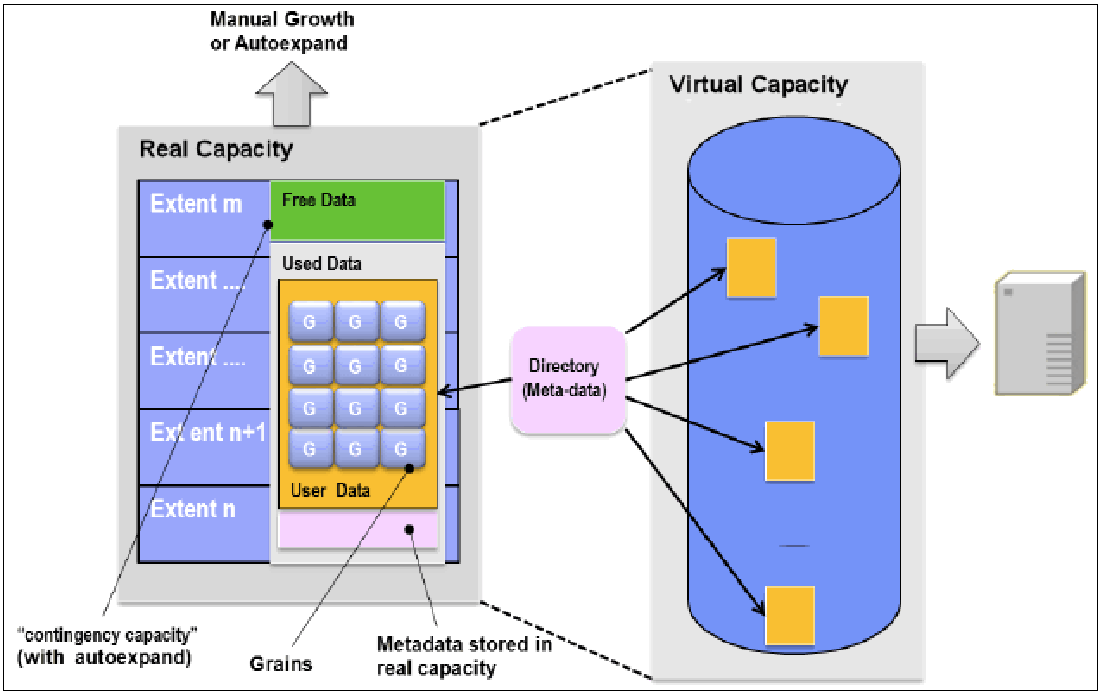


Figure 5-4 Conceptual diagram of thin-provisioned volume

Real capacity defines how much disk space from a pool is allocated to a volume. *Virtual capacity* is the capacity of the volume that is reported to the hosts. Virtual capacity is typically 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 using any real capacity.

Thin-provisioned volumes are available in two operating modes: *Autoexpand* and *nonautoexpand*. You can switch the mode at any time. If you select the autoexpand feature, IBM FlashSystem 9100 automatically adds a fixed amount of extra real capacity to the thin volume as required. Therefore, the autoexpand feature attempts to maintain a fixed amount of unused real capacity for the volume. We recommend to use Autoexpand by default to avoid volume offline issues.

This amount is known as the *contingency capacity*. The contingency capacity is initially set to the real capacity that is assigned when the volume is created. If the user modifies the real capacity, the contingency capacity is reset to be the difference between the used capacity and real capacity.

A volume that is created *without* the **autoexpand** feature, and therefore has a zero contingency capacity, goes offline when the real capacity is used. In this case, it must be expanded.

Warning threshold: When you are working with thin-provisioned volumes, enable the warning threshold (by using email or an SNMP trap) in the storage pool. If you are not using the **autoexpand** feature, you must enable the warning threshold on the volume side too. If the pool or volume runs out of space, the volume will go offline resulting in a loss of access situation.

A thin-provisioned volume can be converted nondisruptively to a fully allocated volume, or vice versa, by using the Modify Capacity Savings functions. Figure 5-5 shows how to convert volume type. You can right click on the volume and select Modify Capacity Savings.

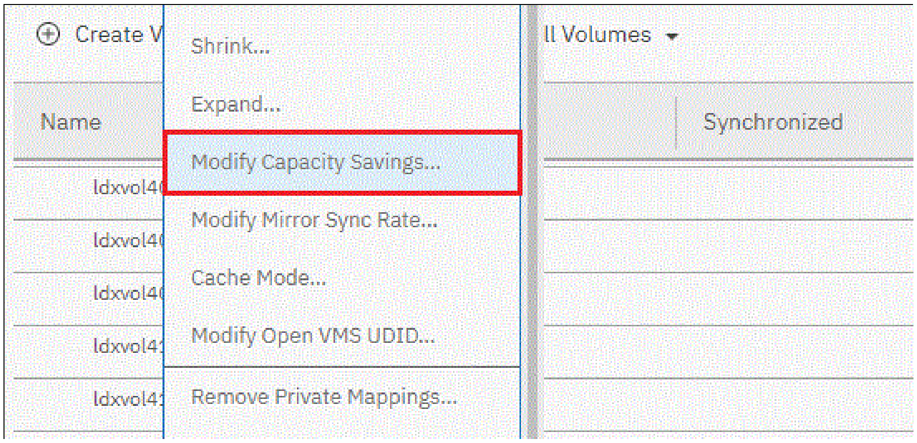


Figure 5-5 Converting volume types nondisruptively

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.

Tip: Consider the use of thin-provisioned volumes as targets in FlashCopy relationships.

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, saving more space. When data is read to hosts, the data is decompressed.

Compression is available through data reduction support as part of the system. If you want volumes to use compression as part of data reduction support, compressed volumes must belong to data reduction pools. Data reduction pools also support reclaiming unused capacity automatically after mapped hosts no longer need the capacity for operations.

These host issue SCSI unmap commands and the released capacity is reclaimed by the data reduction pool for redistribution. For compressed volumes in data reduction pools, the used capacity before compression indicates the total amount of data that is written to volume copies in the storage pool before data reduction occurs.

This compression solution provides nondisruptive conversion between compressed and uncompressed volumes and eliminates the need for special procedures to deal with compressed volumes.

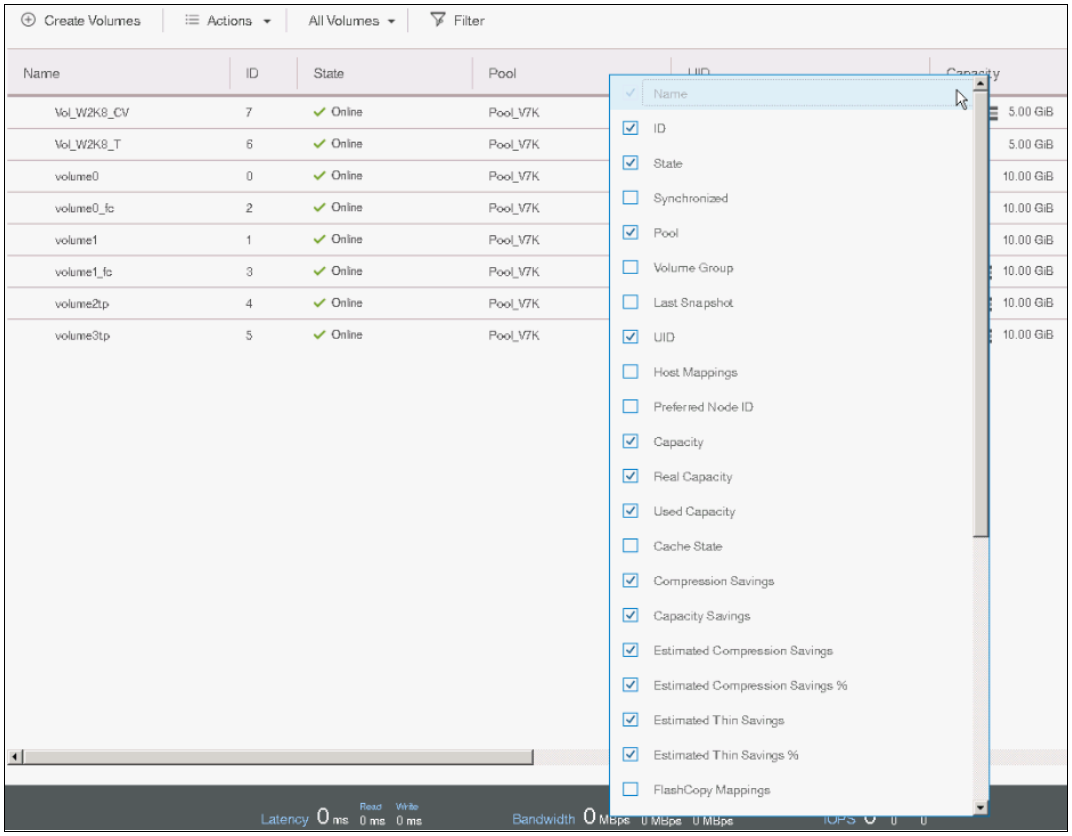
If you are planning to virtualize volumes that are connected to your hosts directly from any storage subsystems, and you want to know what the space saving you will achieve run the Comprestimator Utility.

Comprestimator is a command-line, host-based utility that can be used to estimate an expected compression rate for block devices. The previous link provides all of the instructions needed.

The following actions are preferred practices:

- After you run Comprestimator, consider applying compression only on those volumes that show greater than or equal to 40% capacity savings. For volumes that show less than 40% savings, the trade-off between space saving and hardware resource consumption to compress your data might not make sense.
- After you compress your selected volumes, look at which volumes have the most space saving benefits from thin provisioning rather than RtC. Consider moving these volumes to thin provisioning only. This configuration requires some effort, but saves hardware resources that are then available to give better performance to those volumes, which achieves more benefit from RtC than thin provisioning.

  As shown in Figure 5-6, customize the Volume view to get all the metrics you might need to help make your decision.


Create Volumes
Actions
All Volumes
Filter
Latency 0 ms Read 0 ms Write 0 ms
Bandwidth 0 MBps 0 MBps 0 MBps
IOPS 0 0 0

Name	ID	State	Pool	Capacity
Vol_W2K8_CV	7	✓ Online	Pool_V7K	5.00 GiB
Vol_W2K8_T	6	✓ Online	Pool_V7K	5.00 GiB
volume0	0	✓ Online	Pool_V7K	10.00 GiB
volume0_fc	2	✓ Online	Pool_V7K	10.00 GiB
volume1	1	✓ Online	Pool_V7K	10.00 GiB
volume1_fc	3	✓ Online	Pool_V7K	10.00 GiB
volume2tp	4	✓ Online	Pool_V7K	10.00 GiB
volume3tp	5	✓ Online	Pool_V7K	10.00 GiB

✓ Name
☑ ID
☑ State
□ Synchronized
☑ Pool
□ Volume Group
□ Last Snapshot
☑ UID
□ Host Mappings
□ Preferred Node ID
☑ Capacity
☑ Real Capacity
☑ Used Capacity
□ Cache State
☑ Compression Savings
☑ Capacity Savings
☑ Estimated Compression Savings
☑ Estimated Compression Savings %
☑ Estimated Thin Savings
☑ Estimated Thin Savings %
□ FlashCopy Mappings

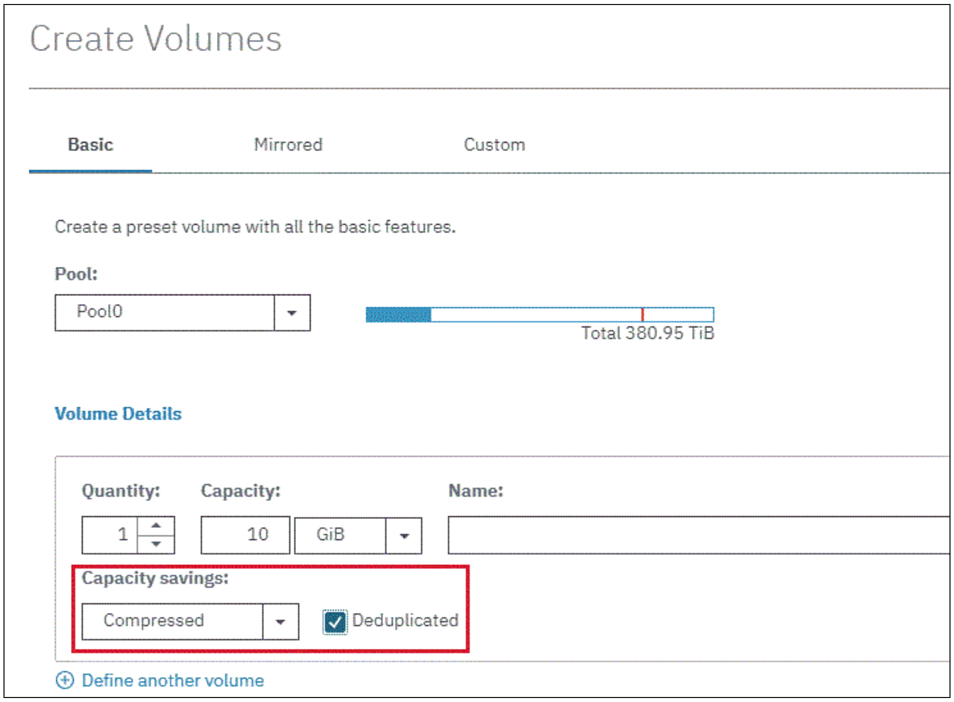
Figure 5-6 Customized view

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 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, resulting in the amount of data that must be stored being greatly reduced. In IBM FlashSystem 9100, you can enable deduplication for thin provisioned and compressed volumes.

You can see this in Figure 5-7.



Create a preset volume with all the basic features.

Pool: Pool0

Total 380.95 TiB

Volume Details

Quantity: 1

Capacity: 10 GiB

Name:

Capacity savings: Compressed ☑ Deduplicated

Figure 5-7 Creating deduplicated volumes

To create a thin-provisioned volume that uses deduplication, enter the following command in the CLI, see Example 5-3.

Example 5-3 Creating thin-provisioned volume with deduplication option

```
superuser>mkvolume -name dedup_test_01 -size 10 -unit gb -pool Pool0 -thin
-deduplicated
Volume, id [42], successfully created
```

To create a compressed volume that uses deduplication, enter the following command in Example 5-4.

Example 5-4 Creating compressed volume with deduplication option

```
superuser>mkvolume -name dedup_test_02 -size 10 -unit gb -pool Pool0 -compressed
-deduplicated
Volume, id [43], successfully created
```

As stated before, deduplication works by identifying repeating chunks in the data written to the storage system. IBM FlashSystem 9100 calculates a signature for each data chunk and checks if that signature is already present in the deduplication data base. If a signature match is found, the data chunk is replaced by a reference to an already stored chunk, which reduces storage space required for storing the data.

To maximize the space that is available for the deduplication data base, 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 VMs running the same operating system and backup servers.

In both cases it is expected that there will be multiple copies of identical files, such as components of the standard operating system or applications used in the given organization. Conversely, data encrypted and/or compressed at the file system level, will not benefit from deduplication as these operations have already removed redundancy.

While deduplication, like other features of IBM FlashSystem 9100, is designed to be transparent to end users and applications, it needs to be planned for and understood before implementation, as it can reduce redundancy of a given solution. For example, for applications designed to store two copies of a file for redundancy, you should ensure that both copies are not on the same volume.

Note: You can use the management GUI or the command-line interface to run the built-in compression estimation tool, Comprestimator. This tool can be used to determine the capacity savings that are possible for existing data on the system by using compression.

5.3.3 Capacity reclamation

File deletion in FlashSystem is realized by updating metadata and marking the physical storage space used by the removed file as unused. The data of the removed file is not overwritten. This improves file system performance by reducing the number of I/O operations on physical storage required to perform file deletion. However, this approach impacts management of real capacity of volumes with enabled capacity savings – file system deletion frees space at the FlashSystem level, but physical data blocks allocated by the storage for the given file still take up the real capacity of a volume.

To address this issue FlashSystem added support for the SCSI unmap command, which can be issued after file deletion, and which informs the storage system, that physical blocks used by the removed file should be marked as no longer in use and can be freed up. Modern operating systems issue SCSI unmap commands only to storage that advertises support for this feature.

Note: For volumes outside data reduction pools the complete stack from the operating system down to back-end storage controller has to support unmap in order to enable the capacity reclamation. SCSI unmap is passed only to specific back-end storage controllers.

Before enabling SCSI unmap read the IBM Support article:

IBM Support article

5.3.4 Space allocation

When a thin-provisioned volume is created, a small amount of the real capacity is used for initial metadata. Write I/O to the thin-provisioned volume cause grains of the real capacity to store metadata and user data. Write I/O updates the grain where data was previously written.

Grain definition: The grain value is defined when the volume is created and can be 32 KB, 64 KB, 128 KB, or 256 KB (default). The grain size cannot be changed after the thin-provisioned volume is created.

Smaller granularities can save more space, but they have larger directories. If you select 32 KB for the grain size, the volume size cannot exceed 260,000 GB. Therefore, if you are not going to use the thin-provisioned volume as a FlashCopy source or target volume, use 256 KB by default to maximize performance.

Now, if you are planning to use Thin Provisioning with FlashCopy, you should 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.

5.3.5 Thin Provisioning considerations

Thin Provisioning is a well-understood technology and it saves capacity only if the host server does not write to whole volumes. Even if the thin-provisioned volume works well, it depends on how the file system has allocated the space.

A volume that is thin-provisioned by IBM FlashSystem 9100 is a volume where the large chunk of binary zeros are not stored in the storage pool. So, if you did not write data to the volume yet, you do not need to use valuable resources storing data that does not exist yet in the form of zeros.

There are a number of different properties of thin-provisioned volumes that are useful to understand for the rest of the chapter:

- When the used capacity first exceeds the volume *warning threshold*, an event is raised, indicating additional real capacity is required. The default warning threshold value is 80% of the volume capacity. To disable warnings, specify 0%.
- For a Compressed Volume only (because Compressed Volumes are based on thin-provisioned volumes), there is the amount of uncompressed user data that has been written to the volume. This is called the *uncompressed used capacity*. In other words, the uncompressed used capacity is the used capacity if data had not be compressed. It is used to calculate the compression ratio:

  ```
  ((uncompressed used capacity - used capacity) / uncompressed used capacity =
  compression ratio)
  ```

Thin Provisioning and overallocation

Since thin provisioned volumes do not store the zero blocks, a storage pool is only overallocated after the sum of all volume capacities exceeds the size of the storage pool.

Storage administrators probably think about the “out of space” problem. If you already have enough capacity on disk to store fully allocated volumes, and then you convert them to thin provisioned volumes, then you will have enough space to store data even if the server writes to every byte of virtual capacity. So, this is not going to be a problem for the short term, and you will have time to monitor your system and understand how your capacity grows.

How to monitor Thin Provisioning

Capacity planning for Thin Provisioning or compressed volumes are no different from fully allocated volumes. The administrator needs to monitor the amount of capacity being used in the storage pool. Make sure that you add more capacity before it runs out.

In case of fully allocated volumes, the used capacity increases during working hours because the administrator can increase volume size or create more volumes. However, in case of a thin-provisioned volume, the used capacity can be increased at any time as long as the file system grows. Thus you need to consider capacity planning carefully.

To avoid unpleasant situations where some volumes can go offline due to lack of space, the storage administrator needs to monitor the real capacity rather than the volume capacity. They also need to monitor it more regularly because the real capacity can increase at any time of day for any reason.

Tools like IBM Spectrum Control can capture the real capacity of a storage pool and enable you to create a graph as to how it is growing over time. Having a tool to show how the real capacity is growing over time is an important requirement to predict when the space will run out.

IBM FlashSystem 9100 also alerts you by putting an event into the event log when the storage pool reaches a configurable threshold, called the warning level. The GUI sets this threshold to 80% of the capacity of the storage pool by default.

What to do if you run out of space

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

The following mechanisms and processes can help you deal with this situation:

- Automatic out of space protection provided by the product

  If the storage pool runs out of space, each volume has its own emergency capacity. That emergency capacity is normally sizable (2% is the default). The emergency capacity that is dedicated to a volume could allow that volume to stay online. This feature means that when you run out of space, you do have some time to repair things before everything starts going offline.

  So you might implement a policy of 10% emergency capacity per volume if you want more safety. Also, remember that you do not need to have the same contingency capacity for every volume.

  > **Note:** This automatic protection will probably solve most immediate problems, but remember that after you are informed that you have run out of space, you have a limited amount of time to react. You need a plan already 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 FCM or SSD you have, the more time you have to solve the problem.

- Move volumes

  You can migrate volumes to other pools to free up space. However, data migration on IBM FlashSystem 9100 is designed to go 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, but you can use policies to reduce the likelihood of that ever happening to the point where you feel comfortable doing less of the other options.

  You can use these types of policies for Thin Provisioning:

Note: The following policies use arbitrary numbers. These arbitrary numbers are designed to make the suggested policies more readable. We do not give any recommended numbers to insert into these policies because they are determined by business risk, and this consideration is different for every client.

- Manage free space such that there is always enough free capacity for your 10 biggest volumes to reach 100% full without running out of free space.
- Never overallocate more than 200%. In other words, if you have 100 TB of capacity in the storage pool, then the sum of the volume capacities in the same pool must not exceed 200 TB.
- Always start the process of adding more capacity when the storage pool reaches more than 70% full.

5.3.6 Limits on virtual capacity of thin-provisioned volumes

The extent and grain size factors limit the virtual capacity of thin-provisioned volumes beyond the factors that limit the capacity of regular volumes. Table 5-2 shows the maximum volume for each extent size.

Table 5-2 Maximum thin volume virtual capacities for an extent size

Extent size in MB	Maximum non thin-provisioned volume capacity	Maximum thin-provisioned volume capacity in GB (for regular pool)	Maximum compressed volume size (for regular pool)	Maximum thin-provisioned and compressed volume size in DRP	Maximum total thin-provisioned and compressed capacity for all volumes in a single DRP per IOgroup
16	2 TB	2,000	2 TB	2 TB	2 TB
32	4 TB	4,000	4 TB	4 TB	4 TB
64	8 TB	8,000	8 TB	8 TB	8 TB
128	16 TB	16,000	16 TB	16 TB	16 TB
256	32 TB	32,000	32 TB	32 TB	32 TB
512	64 TB	65,000	64 TB	64 TB	64 TB
1024	128 TB	130,000	96 TB	128 TB	128 TB
2048	256 TB	260,000	96 TB	256 TB	256 TB
4096	256 TB	262,144	96 TB	256 TB	512 TB
8192	262,144	262,144	96 TB	256 TB	1,024 TB

When creating a compressed volume, the size of the compressed volume should never be greater than 96 TiB of user data. Any compressed volume that is larger than 96 TiB might experience a loss of access to the volume. This limitation applies to compressed volumes in regular pools only. Compressed volumes in data reduction pools are not affected. And this limitation does not affect data integrity.

If you have compressed volumes with a capacity of greater than 96 TiB, create multiple smaller compressed volumes and use host techniques to migrate the data from the current volume onto the new volumes. Alternatively you can convert the large compressed volume into a thin provisioned volume by adding a thin provisioned copy to the volume and then deleting the compressed copy once the two copies are in sync.

5.4 Mirrored volumes

By using volume mirroring, a volume can have two physical copies. Each volume copy can belong to a different pool, and each copy has 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 non-mirrored volume by deleting one copy or by splitting one copy to create a new 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 located in data reduction pools, you can also use deduplication to the volume copies to increase the capacity savings. If you are creating a new volume, the two copies can be of different types, but to use deduplication, both copies must reside in a data reduction pool. You can add a deduplicated volume copy in a data reduction pool to an existing volume with a copy in a standard pool. You can use this method to migrate existing 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 both 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.

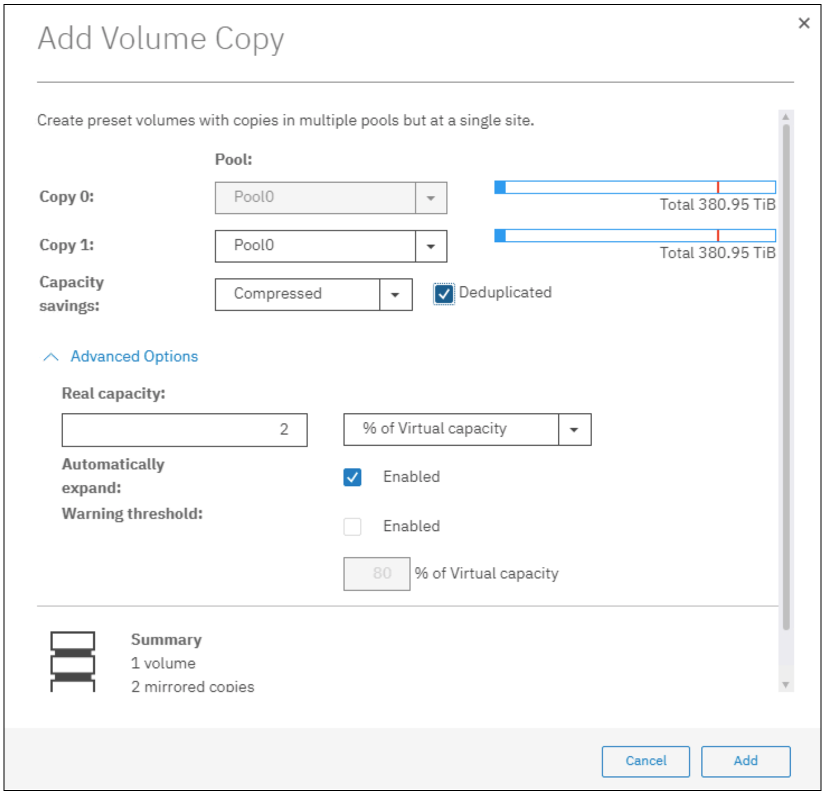
  When the volume is synchronized, you can delete the original copy that is in the source pool. During the synchronization process, the volume remains available even if there is a problem with the destination pool.
- Converting fully allocated volumes to use data reduction technologies, such as thin-provisioning, compression, or deduplication.
- Converting compressed or thin-provisioned volumes in standard pools to data reduction pools to improve capacity savings.

When you use volume mirroring, consider how quorum candidate disks are allocated. Volume mirroring maintains some state data on the quorum disks. If a quorum disk is not accessible and volume mirroring is unable to update the state information, a mirrored volume might need to be taken offline to maintain data integrity. To ensure the high availability of the system, ensure that multiple quorum candidate disks are allocated and configured on different storage systems.

When a volume mirror is synchronized, a mirrored copy can become unsynchronized if it goes offline and write I/O requests need to 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: If the capacity is fully allocated, 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 Add Volume Copy, as shown in Figure 5-8.



Add Volume Copy

Create preset volumes with copies in multiple pools but at a single site.

Pool:

Copy 0: Pool0 — Total 380.95 TiB

Copy 1: Pool0 — Total 380.95 TiB

Capacity savings: Compressed ☑ Deduplicated

Advanced Options

Real capacity: 2 % of Virtual capacity

Automatically expand: ☑ Enabled

Warning threshold: □ Enabled

80 % of Virtual capacity

Summary
1 volume
2 mirrored copies

Figure 5-8 Add Volume Copy options

5.4.1 Write fast failovers

With write fast failovers, during processing of host write I/O, the system submits writes to both copies. 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 back end 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 to 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 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. Additionally, 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 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, there can be an underlying performance problem 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 enter the **`mkvdisk`** command to create a new 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, then 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. This means that the system tries the other copy first for read I/O during the following 4 - 6 minutes. After that, the system reverts to read the original primary-for-read copy.

During this period, if read I/O to the other copy also goes 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 can happen 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 has failed. 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 and high availability between I/O groups at different locations. Before you configure HyperSwap volumes, the system topology needs to be configured for HyperSwap and sites must be defined. Figure 5-9 shows overall FlashSystem 9100 HyperSwap diagram.

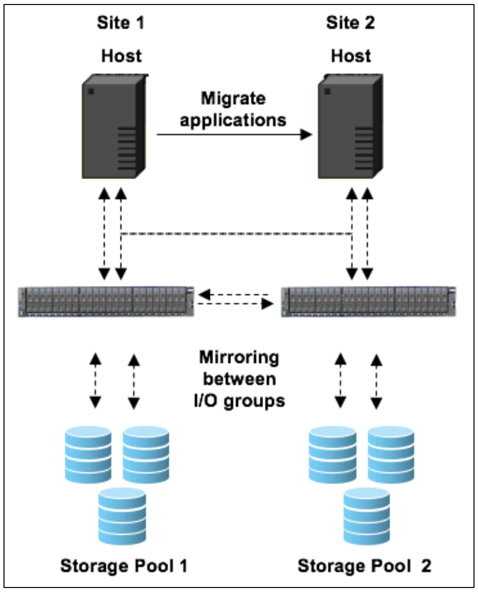


Figure 5-9 Overall HyperSwap diagram

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 either 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 will be 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 data reduction pools on both sites.

The management GUI creates an active-active relationship and change volumes automatically. Active-active 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 data reduction pool, then the corresponding change volume is created with compression enabled. If the base volume is in a standard pool, then 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 disaster recovery 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 data reduction pool, then the corresponding change volume is created with compression enabled. If the base volume is in a standard pool, then the change volume is created as a thin-provisioned volume.

You can see the relationship between Master and Auxiliary volume in the HyperSwap topology in Figure 5-10.

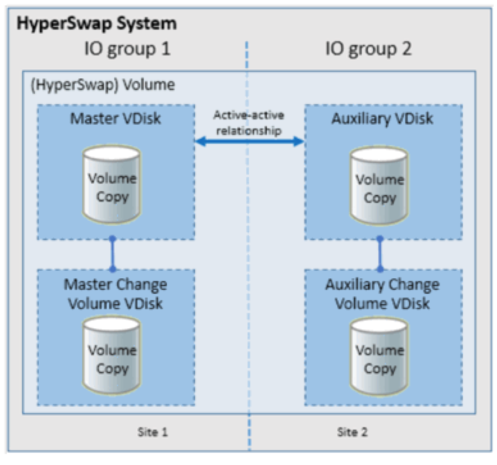


Figure 5-10 Master and Auxiliary vdisks

5.6 VMware virtual volumes

The IBM FlashSystem 9100 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 certain 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 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.

Figure 5-11 provides a high-level overview of the key components that enable the VVols management framework.

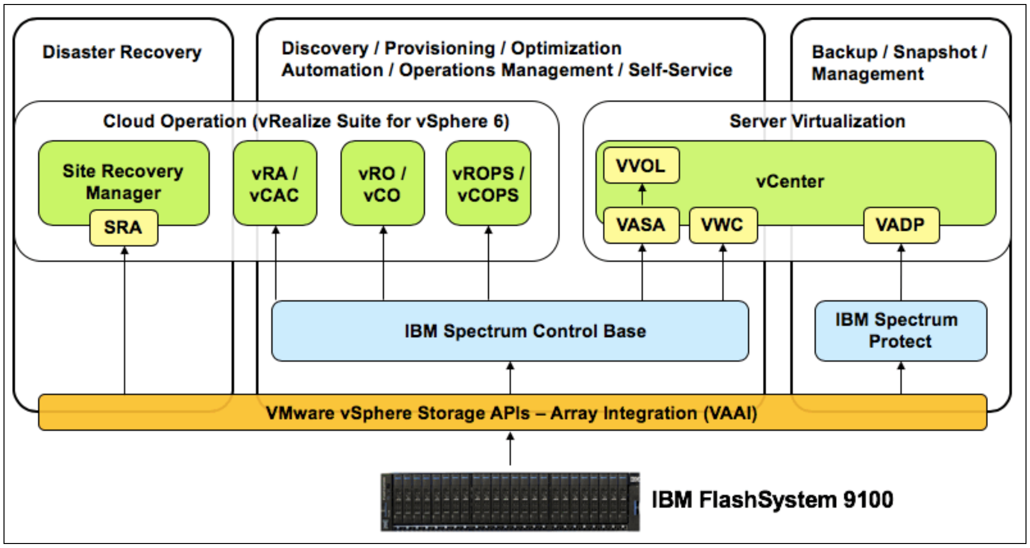

Disaster Recovery
Discovery / Provisioning / Optimization
Automation / Operations Management / Self-Service
Backup / Snapshot / Management
Cloud Operation (vRealize Suite for vSphere 6)
Server Virtualization
Site Recovery Manager
SRA
vRA / vCAC
vRO / vCO
vROPS / vCOPS
VVOL
vCenter
VASA
VWC
VADP
IBM Spectrum Control Base
IBM Spectrum Protect
VMware vSphere Storage APIs – Array Integration (VAAI)
IBM FlashSystem 9100

Figure 5-11 Overview of the key components of VMware environment

You can also use data copy through VMware vSphere Storage APIs Array Integration (VAAI) in Figure 5-12.

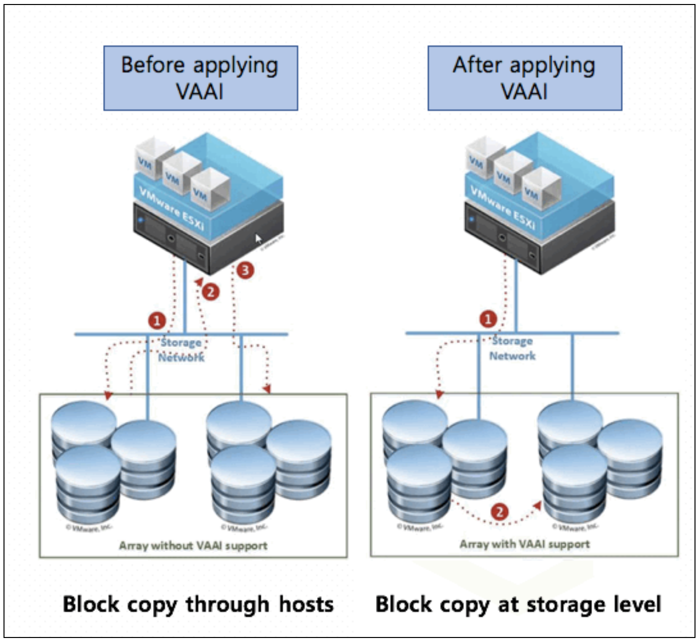


Figure 5-12 VMware vSphere Storage APIs Array Integration (VAAI)

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 disaster recovery 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 and 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 5-13 shows a sample diagram of IBM FlashSystem 9100 Transparent Cloud Tiering.

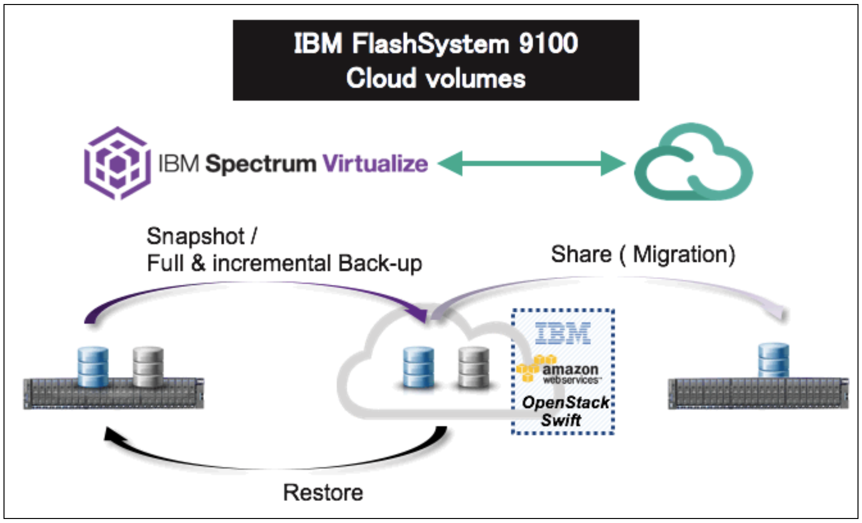


Figure 5-13 Cloud volumes - Transparent Cloud Tiering

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, and 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 either 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 then access cloud storage to either 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

There are certain limitations and rules:

- 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 are valid:
  - The volume is part of a remote copy relationship (Metro Mirror, Global Mirror, 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.
  - The volume is associated with any user-owned FlashCopy maps.
  - The volume is a mirrored volume with copies in different storage pools.
  - The volume is 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 an existing 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.
  - Part of a remote copy relationship (Metro Mirror, Global Mirror, 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 Restore to the production volume

The snapshot version is restored 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. 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.

5.7.3 Restore to a new volume

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, then the restore operation uses the unique identifier (UID) of the new volume. If the new volume does not exist on the system, you need to choose whether to use the UID from the original volume or create a new UID. If you plan on using 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

A volume can be migrated from one storage pool to IBM FlashSystem 9100 regardless of the virtualized volumes type. The command varies depending on the type of migration, as shown in Table 5-3.

Table 5-3 Migration types and associated commands

Storage pool-to-storage pool type	Command
Managed-to-managed or Image-to-managed	**migratevdisk**
Managed-to-image or Image-to-image	**migratetoimage**

Migrating a volume from one storage pool to another is nondisruptive to the host application using the volume. Depending on the workload of IBM FlashSystem 9100, there might be a slight performance impact. This section provides guidance for migrating volumes.

5.8.1 Image-type to striped-type migration

When you are migrating existing storage into the IBM FlashSystem 9100, the existing storage is brought in as *image-type volumes*, which means that the volume is based on a single MDisk. The CLI command that can be used is **migratevdisk**.

Example 5-5 shows the **migratevdisk** command that can be used to migrate an *image-type volume* to a *striped-type volume,* and can be used to migrate a *striped-type volume* to a *striped-type volume* as well.

Example 5-5 The migratevdisk command

```
superuser>svctask migratevdisk -mdiskgrp MDG1DS4K -threads 4 -vdisk Migrate_sample
```

This command migrates the volume `Migrate_sample` to the storage pool `MDG1DS4K`, and uses four threads when migrating. Instead of using the volume name, you can use its ID number.

You can monitor the migration process by using the **svcinfo lsmigrate** command, as shown in Example 5-6.

Example 5-6 Monitoring the migration process

```
superuser>svcinfo 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 only represented as an *image-type volume* after it reaches the state where it is a straight-through mapping. An image-type volume cannot be expanded.

Image-type disks are used to migrate existing data to an IBM FlashSystem 9100 and to migrate data out of virtualization. In general, the reason for migrating a volume to an image type volume is to move the data on the disk to a 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 managed disk that is specified as the target must be in an *unmanaged* state at the time that the command is run. Running this command results in the inclusion of the MDisk into the user-specified storage pool.

> **Remember:** This command cannot be used if the source volume copy is in a child pool or if the target MDisk group that is specified is a child pool. This command does not work if the volume is fast formatting.

The **`migratetoimage`** command fails if the target or source volume is offline. Correct the offline condition before attempting to migrate the volume.

If the volume (or volume copy) is a target of a FlashCopy mapping with a source volume in an active-active relationship, the new managed disk group must be in the same site as the source volume. If the volume is in an active-active relationship, the new managed disk group must be located in the same site as the source volume. Additionally, the site information for the MDisk being added must be well-defined and match the site information for other MDisks in the storage pool.

> **Note:** You cannot migrate 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*. To better understand all requirements and specifications for that command, see IBM Knowledge Center.

5.8.3 Migration from Standard Pool to Data Reduction Pool

If you want to migrate volumes to DRP, you can move them with volume mirroring between standard pool and DRP. Hosts I/O operations are not disrupted during migration. It is not supported to change from standard pool to DRP without disruption. Figure 5-14 on page 127 shows volume mirroring migration to change pool types and volume types.

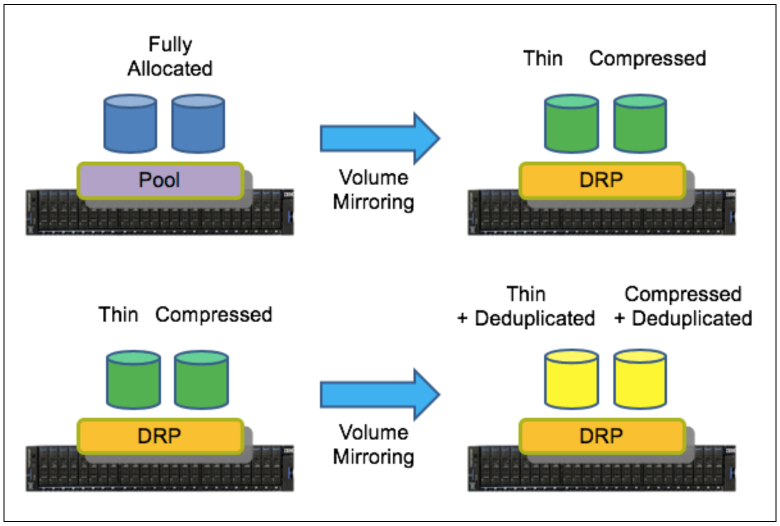


Figure 5-14 Volume Mirroring to change pools and volume types

You can also move compressed or thin-provisioned volumes in standard pools to data reduction pools to simplify management of reclaimed capacity. The data reduction pool 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 data reduction pool. This method creates a copy of the volume in a new data reduction pool and does not disrupt host operations.

5.8.4 Migrating from Fibre Channel connections to RDMA over Ethernet connections between nodes

The system supports node-to-node connections that use Ethernet protocols that support remote direct memory access (RDMA) technology, such as RDMA over Converged Ethernet (RoCE) or iWARP. To use these protocols, the system requires that an RDMA-capable adapter is installed on each node and dedicated RDMA-capable Ethernet ports are only configured for node-to-node communication. If your system currently uses Fibre Channel ports, you can migrate to RDMA-capable Ethernet ports for node-to-node communications.

RDMA technologies, like RoCE and iWARP, enable the RDMA-capable adapter to transfer data directly between nodes, bypassing CPU, and cache, making transfers faster. RDMA technologies provide faster connection and processing time than traditional iSCSI connections.

The following prerequisites are required for all RDMA-capable Ethernet ports that are used between nodes:

- All installation of the node hardware is complete.
- The 25-Gbps Ethernet adapter that supports RDMA technology is installed on each node. If you are using RDMA-technology for node-to-node communications, ensure that the RDMA-capable adapters use the same technology, such as RoCE or iWARP. These RDMA-capable adapters must be installed in the same slots across all the nodes of the system. These installation requirements ensure that port identifiers are same across all nodes in the system.
- Ethernet cables between each node are connected correctly.
- The protocol technology on the source and destination adapters is the same.

- The local and remote IP addresses can be reached.
- Each IP address for RDMA-capable Ethernet ports and their associated subnet masks are unique on each node.
- Router must not be placed between nodes that use RDMA-capable Ethernet ports for node-to-node communication.
- The negotiated speeds on the local and remote adapters are the same.
- The local and remote port virtual LAN identifiers are the same. Use virtual LAN to create physical separation of networks for unrelated systems, wherever possible. All the ports that are used for node-to node communication must be assigned with one VLAN ID and ports that are used for host attachment must have a different VLAN ID.

  If you plan to use VLAN to create this separation, you must configure VLAN support on the all the Ethernet switches in your network before you define the RDMA-capable Ethernet ports on nodes in the system. On each switch in your network, set VLAN to Trunk mode and specify the VLAN ID for the RDMA-ports that will be in the same VLAN.

  In addition, if VLAN settings for a RDMA-capable Ethernet port needs to be updated, these settings cannot be updated independently of other configuration settings. Before you update VLAN settings on specific RDMA-capable Ethernet ports, you must unconfigure the port, make any necessary changes to the switch configuration, then reconfigure RDMA-capable Ethernet ports on each of the nodes in the system.
- A minimum of two dedicated RDMA-capable Ethernet ports are required for node-to-node communications to ensure best performance and reliability. These ports must be configured for inter-node traffic only and must not be used for host attachment, virtualization of Ethernet-attached external storage, or IP replication traffic.
- A maximum of four RDMA-capable Ethernet ports per node are allowed for node-to-node communications.

5.8.5 Migrating with volume mirroring

You can also use volume mirroring when you migrate a volume from existing storage to IBM FlashSystem 9100. As you can see in Figure 5-15 on page 129, first of all, you need to attach existing storage to IBM FlashSystem 9100 by using a virtualization solution. In this case, you will need some downtime to attach storage to IBM FlashSystem 9100. Then, attach mirrored volumes to the hosts so that host can recognize volumes the same as existing ones. You can now restart applications. Volume mirroring will be implemented in the background after sync starts.

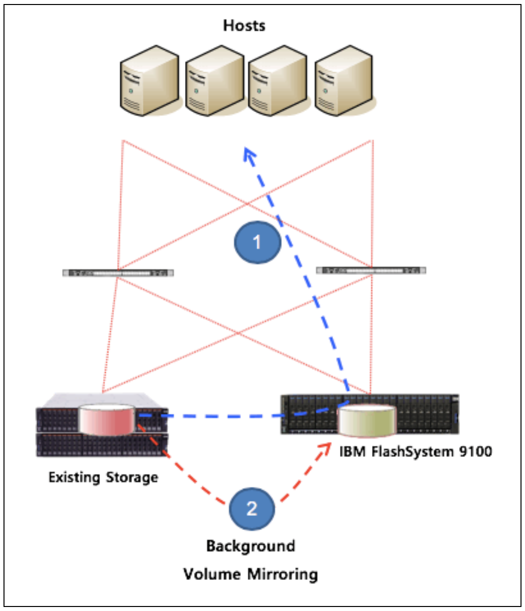


Figure 5-15 Migration with Volume Mirroring

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 following steps are similar:

1. Add a target fully allocated copy.
2. Wait for synchronization to complete.
3. Remove the source thin-provisioned copy.

In both cases, if you set the **`autodelete`** option to *yes* when creating the volume copy, the source copy is automatically deleted, and you can skip the steps 3 mentioned previously. The preferred practice on this type of migration is to try not to overload the systems with a high *syncrate,* and not overload the system with too many migrations at the same time.

Note: You cannot use the data migration function to move a volume between storage pools that have different extent sizes. Migration commands fail if the target or source volume is offline, there is no quorum disk defined, or the defined quorum disks are unavailable. Correct the offline or quorum disk condition and reissue the command.

The **syncrate** parameter specifies the copy synchronization rate. A value of zero (0) prevents synchronization. The default value is 50. See Table 5-4 for the supported **-syncrate** values and their corresponding rates. We recommend modify syncrate after monitoring overall bandwidth and latency. And then, if the performance is not impacted on migration, increase syncrate properly to complete within allotted time.

Table 5-4 Sample syncrate values

User-specified syncrate attribute value	Data copied/sec
1 - 10	128 KB
11 - 20	256 KB
21 - 30	512 KB
31 - 40	1 MB
41 - 50	2 MB
51 - 60	4 MB
61 - 70	8 MB
71 - 80	16 MB
81 - 90	32 MB
91 - 100	64 MB

For more information, see IBM Knowledge Center.

5.9 Preferred paths to a volume

The pair of nodes within a single enclosure is known as an input/output (I/O) group. When a write operation is performed to a volume, the node that processes the I/O duplicates the data onto the partner node that is in the I/O group.

After the data is protected on the partner node, the write operation to the host application is completed. The data is physically written to the disk later. Volumes are logical disks that are presented to the system by nodes. Volumes are also associated with the I/O group.

When you create a volume, you can specify a preferred node. Many of the multipathing driver implementations that the system supports use this information to direct I/O to the preferred node. The other node in the I/O group is used only if the preferred node is not accessible.

If you do not specify a preferred node for a volume, the system selects the node in the I/O group that has the fewest volumes to be the preferred node. After the preferred node is chosen, it can be changed only when the volume is moved to a different I/O group.

> **Note:** The management GUI provides a wizard that moves volumes between I/O groups without disrupting host I/O operations.

IBM FlashSystem 9100 implements the concept of each volume having a preferred owner node, which improves cache efficiency and cache usage. The cache component read/write algorithms depend on one node that owns all the blocks for a specific track. The preferred node is set at the time of volume creation manually by the user or automatically by IBM FlashSystem 9100.

Because read-miss performance is better when the host issues a read request to the owning node, you want the host to know which node owns a track. The SCSI command set provides a mechanism for determining a preferred path to a specific volume. Because a track is part of a volume, the cache component distributes ownership by volume. The preferred paths are then all the paths through the owning node. Therefore, a preferred path is any port on a preferred controller, assuming that the SAN zoning is correct.

Tip: Performance can be better if the access is made on the preferred node. The data can still be accessed by the partner node in the I/O group if a failure occurs.

By default, IBM FlashSystem 9100 assigns ownership of even-numbered volumes to one node of a caching pair and the ownership of odd-numbered volumes to the other node. It is possible for the ownership distribution in a caching pair to become unbalanced if volume sizes are different between the nodes or if the volume numbers that are assigned to the caching pair are predominantly even or odd.

To provide flexibility in making plans to avoid this problem, the ownership for a specific volume can be explicitly assigned to a specific node when the volume is created. If a node becomes overloaded, the preferred node of a volume can be changed to the other node in the same I/O group, or to a node in another I/O group. This procedure can be performed concurrently with I/O operations if the host supports non-disruptive volume move. Figure 5-16 shows a write operation from a host.

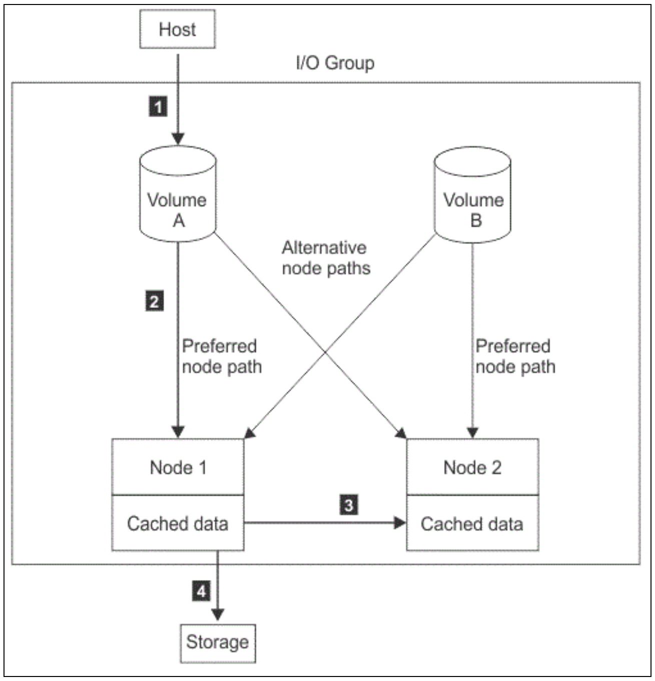


Figure 5-16 Write operation from a host

IBM multipathing software (SDDPCM or SDDDSM) on hosts are aware of the preferred paths that IBM FlashSystem 9100 sets per volume. They use algorithms to select paths and balance the load across them. In cases where all paths to preferred and non-preferred nodes are all available, the host will perform I/O operations using the paths to the preferred node. If all paths to preferred node become unavailable, the multipath software will make the host use the non-preferred paths. If all paths become unavailable, the host will set the device offline.

Sometimes when debugging performance problems, it can be useful to look at the `Non-Preferred Node Usage Percentage` metric in IBM Spectrum Control. I/O to the non-preferred node might cause performance problems for the I/O group. This metric identifies any usage of non-preferred nodes to the user.

For more information about this metric and more, see IBM Spectrum Control in IBM Knowledge Center.

5.10 Changing the preferred node moving a volume between I/O groups

The change of preferred node of a volume either within an I/O group or to another I/O group is a nondisruptive process.

Changing the preferred node within an I/O group can be done with concurrent I/O. However, it can lead to some delay in performance and, in case of some specific operating systems or applications, they could detect some time outs.

Changing the preferred node within an I/O group can be done by using both CLI and GUI, but if you have only one I/O group, this is not possible using the GUI. To change the preferred node within an I/O group using CLI, use the command **`movevdisk -node <node_id or node_name> <vdisk_id or vdisk_name>`**.

There are some limitations to change the preferred node across I/O groups, which is named Non-Disruptive Volume Move (NDVM). These limitations are mostly in Host Cluster environments, and you can 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 non-disruptive volume move. It also requires a rescan at the host level to ensure that the multipathing driver is notified that the allocation of the preferred node has changed and the ports by which the volume is accessed has changed. This can be done in the situation where one pair of nodes becomes over used.

If there are any host mappings for the volume, the hosts must be members of the target I/O group or the migration fails.

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 volumes so that I/O operations for production volumes are not affected. Figure 5-17 shows the example of volume throttling.

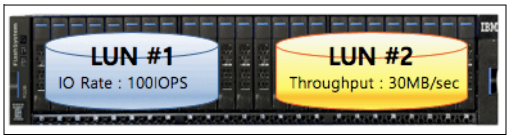


Figure 5-17 Volume throttling for each LUNs

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. It would be better to set an IOPS limit.

On the other hand, a streaming video application often issues a small amount of I/O, but it transfers large amounts of data. In contrast to the database example, defining an I/O throttle based in IOPS does not achieve a good result. For a streaming video application, it would be better to set an MBps limit.

You can edit the throttling value in this menu in Figure 5-18.

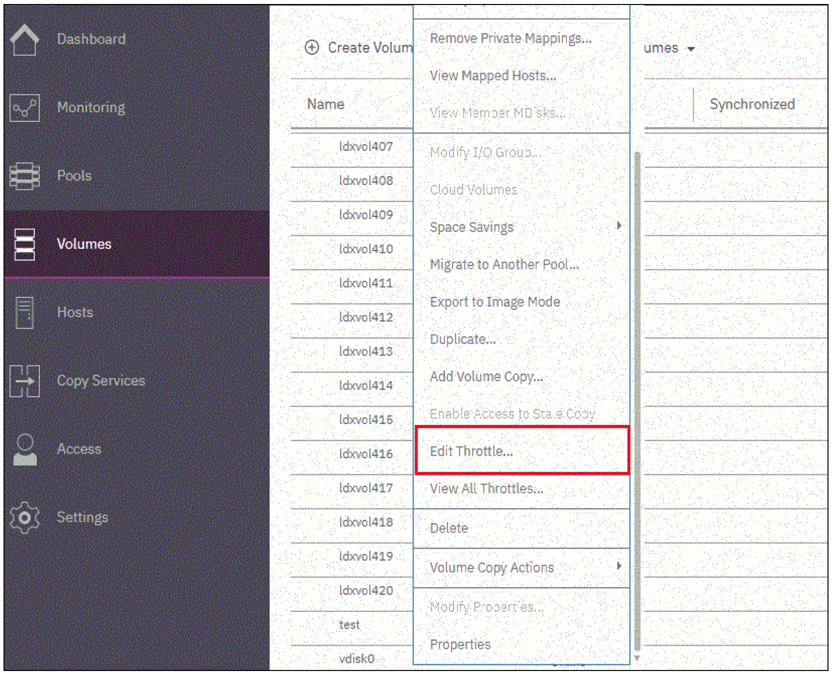
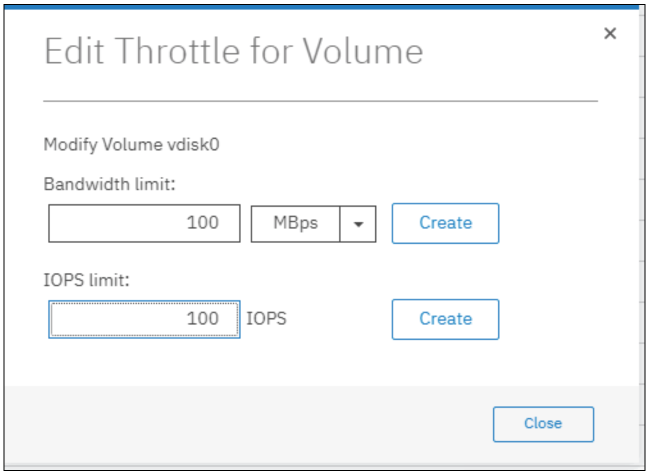


Figure 5-18 Volume Throttling

Figure 5-19 shows both bandwidth and IOPS parameter.



Edit Throttle for Volume

Modify Volume vdisk0

Bandwidth limit: 100 MBps

IOPS limit: 100 IOPS

Figure 5-19 Edit bandwidth and IOPS limit

As mentioned previously, with Volume throttling, the IOPS limit, the bandwidth limit, or both, can be set for a volume.

Throttling at a volume level can be set by using the two commands below:

- **`mkthrottle`**: To set I/O throttles for volumes using this command, it must be used with **`-type vdisk`** parameter, followed by **`-bandwidth bandwidth_limit_in_mbdisk`** and/or**`-iops iops_limit`** to define MBps and IOPS limits.

- **chvdisk**: This command used with **-rate throttle_rate** parameter specifies the IOPS and MBps limits. The default throttle_rate units are I/Os. To change the **throttle_rate** units to megabits per second (MBps), specify the **-unitmb** parameter. If **throttle_rate** value is zero, the throttle rate is disabled. By default, the **throttle_rate** parameter is disabled.

> **Note:** The command **mkthrottle** is used not to create throttles for volumes only, but also for hosts, host clusters, pools, and system offload.

When the IOPS limit is configured on a volume, and it is smaller than 100 IOPS, the throttling logic rounds it to 100 IOPS. Even if throttle is set to a value smaller than 100 IOPS, the actual throttling occurs at 100 IOPS.

After using any of the commands shown previously to set volume throttling, a throttle object is created. Then, you can list your created throttle objects by using the **lsthrottle** command, and change their parameters with the **chthrottle** command. Example 5-7 shows some command examples.

Example 5-7 Throttle command example

```
superuser>mkthrottle -type vdisk -bandwidth 100 -vdisk testvol10
Throttle, id [0], successfully created.
superuser>lsthrottle
throttle_id throttle_name object_id object_name throttle_type IOPs_limit
bandwidth_limit_MB
0           throttle0     25        testvol10   vdisk                    100


superuser>lsthrottle
throttle_id throttle_name object_id object_name throttle_type IOPs_limit
bandwidth_limit_MB
0           throttle0     25        testvol10   vdisk         1000       100

superuser>lsthrottle throttle0
id 0
throttle_name throttle0
object_id 25
object_name testvol10
throttle_type vdisk
IOPs_limit 1000
bandwidth_limit_MB 100
```

For more information, and the procedure to set volume throttling, see IBM Knowledge Center.

5.12 Volume cache mode

Cache in IBM FlashSystem 9100 can be set at a single volume granularity. For each volume, the cache can be *readwrite (enabled)*, *readonly*, or *none*. The meaning of each parameter is self-explanatory. By default, when a volume is created, the cache mode is set to *readwrite (Enabled)*.

readwrite (enabled)

All read and write I/O operations that are performed by the volume are stored in cache. This is the default cache mode for all volumes. A volume or volume copy created from a data reduction pool must have a cache mode of readwrite. If you try to create a thin provisioned or compressed volume copy from a data reduction pool and the volume cache mode is not readwrite, the operation fails.

readonly

All read I/O operations that are performed by the volume are stored in cache.

none (disabled)

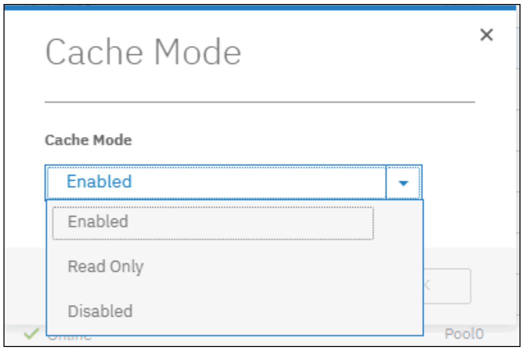
All read and write I/O operations that are performed by the volume are not stored in cache.

In most cases, the volume with readwrite cache mode is recommended, because disabling cache for a volume can result in performance issues to the host. But, there are some specific scenarios that it is recommended to disable the readwrite cache.

You use cache-disabled (*none)* volumes primarily when you have remote copy or FlashCopy in the underlying storage controller, and these volumes are virtualized in IBM FlashSystem 9100 devices as image vdisks. You might want to use cache-disabled volumes where intellectual capital is in existing copy services automation scripts. Keep the use of cache-disabled volumes to minimum for normal workloads.

You can also use cache-disabled volumes to control the allocation of cache resources. By disabling the cache for certain volumes, more cache resources are available to cache I/Os to other volumes in the same I/O group. Also, a case in which you can use cache-disabled volumes is in a scenario of an application which requires very low response time and uses volumes which are in mdisks from all-flash storage. If this application generates many IOPS and requires low response time, disabling the cache of the volumes in IBM FlashSystem 9100 would take advantage of the all-flash performance capabilities, and consume less resources.

The cache mode of a volume can be concurrently changed (with I/O) by using the **`chvdisk`** command or GUI. Figure 5-20 shows editing cache mode for a volume.



Cache Mode

Cache Mode: Enabled

Figure 5-20 Edit cache mode

The command line will not fail I/O to the user, and the command must be allowed to run on any volume. If used correctly without the **`-force`** flag, the command will not result in a corrupted volume. Therefore, the cache must be flushed and you must discard cache data if the user disables cache on a volume.

Example 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>svctask mkvdisk -name VDISK_IMAGE_1 -iogrp 0 -mdiskgrp IMAGE_Test -vtype
image -mdisk D8K_L3331_1108
Virtual Disk, id [9], successfully created
superuser>svcinfo lsvdisk VDISK_IMAGE_1
id 9
.
lines removed for brevity
.
fast_write_state empty
cache readwrite
.
lines removed for brevity

superuser>svctask chvdisk -cache none VDISK_IMAGE_1
superuser>svcinfo lsvdisk VDISK_IMAGE_1
id 9
.
lines removed for brevity
.
cache none
.
lines removed for brevity
```

Tip: By default, the volumes are created with cache mode enabled (read/write), but you can specify the cache mode when the volume is created by using the **`-cache`** option.

Figure 5-21 shows write operation behavior when volume cache is activated (*readwrite*).

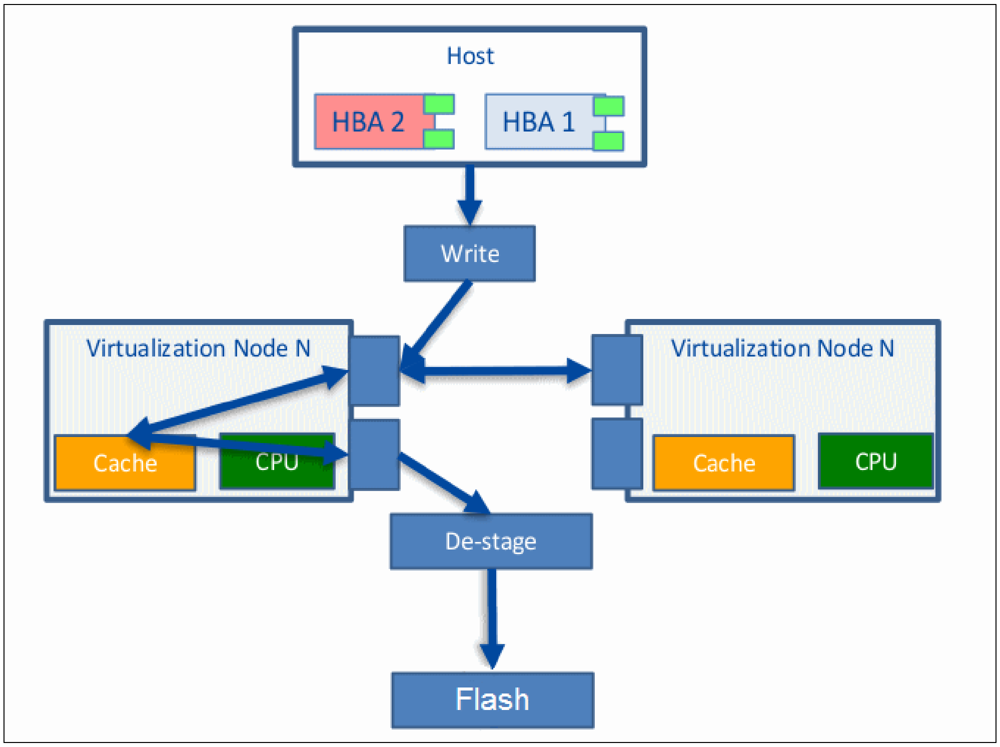


Figure 5-21 Cache activated

Figure 5-22 shows a write operation behavior when volume cache is deactivated (*none*).

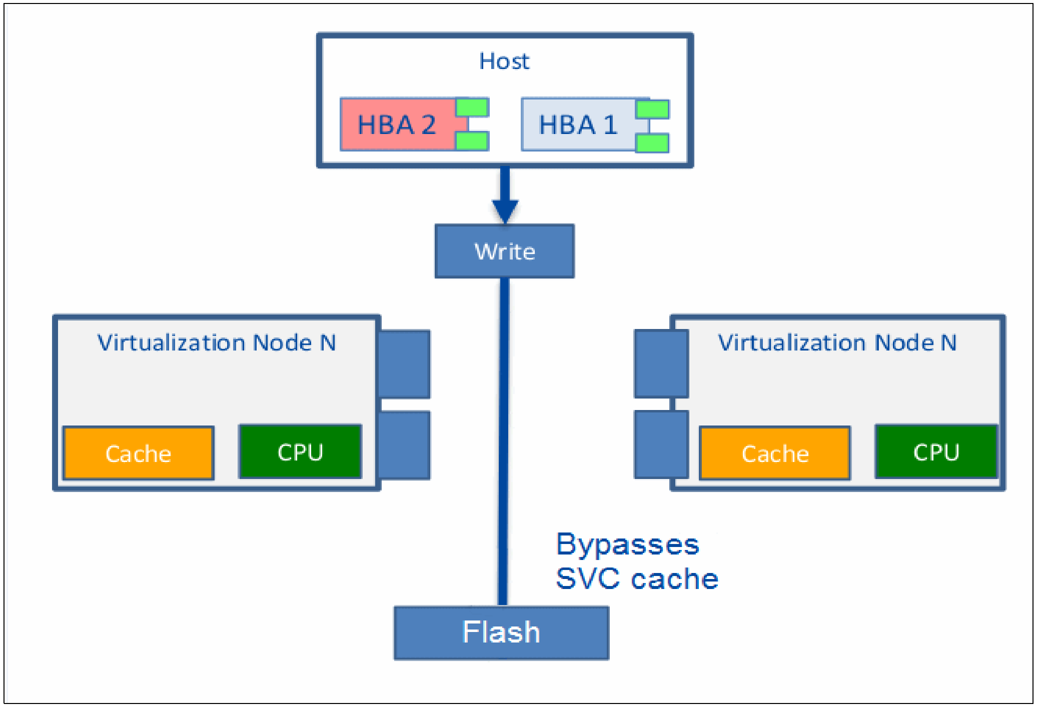


Figure 5-22 Cache deactivated

In this case, an environment with Copy Services (FlashCopy, Metro Mirror, Global Mirror, and Volume Mirroring) and typical workloads, disabling SVC cache is detrimental to overall performance. In cases where there are no advanced functions and extremely high IOPS rate is required, disabling the cache might help.

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

5.13 Additional considerations

The following section describes additional and brief considerations regarding volumes.

5.13.1 Volume protection

You can protect volumes to prevent active volumes or host mappings from being deleted. IBM FlashSystem 9100 supports a global setting that prevents these objects from being deleted if the system detects recent I/O activity.

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. Using the `-force` parameter can lead to unintentional deletions of volumes that are still active. Active means that the system detected recent I/O activity to the volume from any host.

To prevent an active volume from being deleted unintentionally, administrators can use a global system setting to enable volume protection. They can also specify a time period that the volume must be idle before it can be deleted. If volume protection is enabled and the time period is not expired, the volume deletion fails even if the -force parameter is used.

Consider enabling volume protection by using **`chsystem vdiskprotectionenabled yes -vdiskprotectiontime <value_in_minutes>`**.

> **Note:** Volume protection cannot be overridden by the use of the `-force` flag in the affected CLI commands. Volume protection must be disabled to permit an activity that is currently blocked.

5.13.2 Volume resize

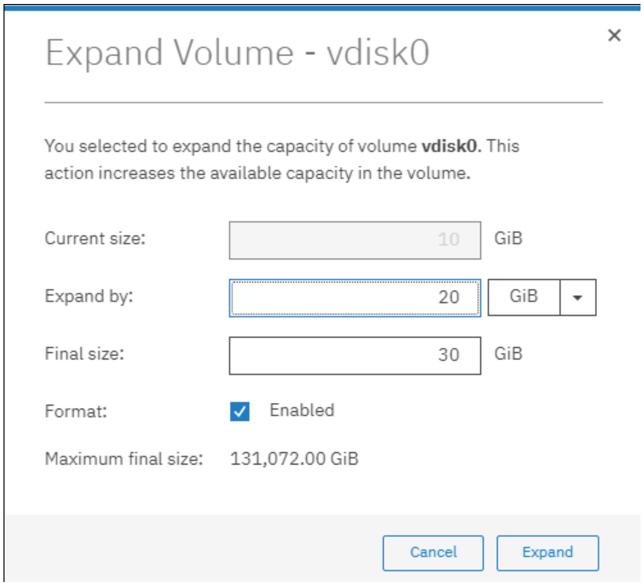
Fully allocated and thin-provisioned volumes can have their sizes increased or decreased. A volume can be expanded with concurrent I/Os for some operating systems, but you should never attempt to shrink a volume in use that contains data, because volume capacity is removed from the end of the disk, whether or not that capacity is in use by a server. Remember that a volume cannot be expanded or shrunk during its quick initialization process.

Expanding a volume

You can expand volumes for the following reasons:

- To increase the available capacity on a particular volume that is already mapped to a host.
- To increase the size of a volume aiming to make it match the size of the source or master volume so that it can be used in a FlashCopy mapping or Metro Mirror relationship.

Figure 5-23 shows the Expand Volume window.


Expand Volume - vdisk0

You selected to expand the capacity of volume **vdisk0**. This action increases the available capacity in the volume.

Current size: 10 GiB

Expand by: 20 GiB

Final size: 30 GiB

Format: ☑ Enabled

Maximum final size: 131,072.00 GiB

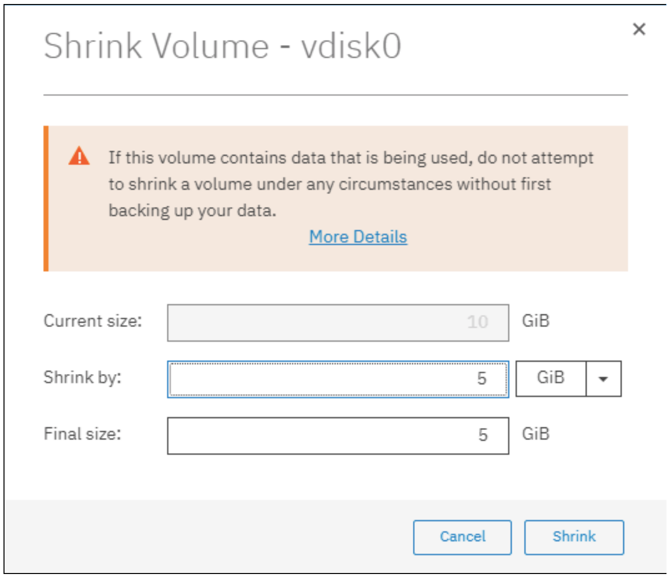
Figure 5-23 Expand Volume

Shrinking a volume

Volumes can be reduced in size if necessary. If a volume does not contain any data, there should be no issues to shrink its size. However, if a volume is in use and contains data, do not shrink its size, because IBM Spectrum Virtualize will not be aware if it is removing used or non-used capacity.

> **Attention:** When you shrink a volume, capacity is removed from the end of the disk, whether or not that capacity is in use. Even if a volume has free capacity, do not assume that only unused capacity is removed when you shrink a volume.

Figure 5-24 shows shrinking volumes.


Shrink Volume - vdisk0

If this volume contains data that is being used, do not attempt to shrink a volume under any circumstances without first backing up your data.

More Details

Current size: 10 GiB

Shrink by: 5 GiB

Final size: 5 GiB

Figure 5-24 Shrink Volume

6

Copy services

Copy services are a collection of functions that provide capabilities for disaster recovery, data migration, and data duplication solutions. This chapter provides an overview and the preferred practices of IBM FS9100 copy services capabilities, including FlashCopy, Metro Mirror and Global Mirror, and Volume Mirroring.

Important, IBM FlashSystem 9200: On 11th February 2020 IBM announced the arrival of the IBM FlashSystem 9200 to the family. This book was written specifically for IBM FlashSystem 9100, however most of the general principles will apply to the IBM FlashSystem 9200.

If you are in any doubt as to their applicability to the FlashSystem 9200 then you should work with your local IBM representative.

This book will be updated to include FlashSystem 9200 in due course.

The Flashsystem 9200 product guide is available at:

IBM FlashSystem 9200 Product Guide

This chapter includes the following sections:

- Introduction to copy services
- FlashCopy
- Remote Copy services
- Native IP replication
- Volume Mirroring

6.1 Introduction to copy services

IBM FlashSystem 9100, like the other IBM Spectrum Virtualize based products, offers 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:

- Metro Mirror makes *synchronous* copies, which means that the original writes are not considered complete until the write to the destination disk has been confirmed. The distance between your two sites is usually determined by how much latency your applications can handle.
- Global Mirror makes *asynchronous* copies of your disk. This fact means that the write is considered complete after it is complete at the local disk. It does not wait for the write to be confirmed at the remote system as Metro Mirror does. This requirement greatly reduces the latency experienced by your applications if the other system is far away. However, it also means that during a failure, the data on the remote copy might not have the most recent changes committed to the local disk.
- Global Mirror with Change Volumes also makes *asynchronous* copies of your disk.This function has been introduced in Spectrum Virtualize version 6.3 as Cycle-Mode Global Mirror. If you use this feature, the system takes periodic FlashCopies of a disk and write them to your remote destination.This feature completely isolates the local copy from wide area network (WAN) issues and from sudden spikes in workload that might occur. The drawback is that your remote copy might lag behind the original by a significant amount, depending on how you have set up the cycle time.

6.1.3 Volume Mirroring function

Volume Mirroring is a function that is designed to increase high availability of the storage infrastructure. It provides the ability to create up to two local copies of a volume. Volume Mirroring can use space from two storage pools, and preferably from two separate back-end disk subsystems.

Primarily, you use this function to insulate hosts from the failure of a storage pool and also from the failure of a back-end disk subsystem. During a storage pool failure, the system continues to provide service for the volume from the other copy on the other storage pool, with no disruption to the host.

You can also use Volume Mirroring to migrate from a thin-provisioned volume to a non-thin-provisioned volume, and to migrate data between storage pools of different extent sizes.

6.2 FlashCopy

By using the IBM FlashCopy function of the IBM Spectrum Virtualize based systems, you can perform a *point-in-time copy* of one or more volumes. This section describes the inner workings of FlashCopy, and provides some preferred practices for its use.

You can use FlashCopy to help you solve critical and challenging business needs that require duplication of data of your source volume. Volumes can remain online and active while you create consistent copies of the data sets. Because the copy is performed at the block level, it operates below the host operating system and its cache. Therefore, the copy is not apparent to the host.

Important: Because FlashCopy operates at the block level below the host operating system and cache, those levels do need to be flushed for consistent FlashCopies.

While the FlashCopy operation is performed, the source volume is stopped briefly to initialize the FlashCopy bitmap, and then input/output (I/O) can resume. Although several FlashCopy options require the data to be copied from the source to the target in the background, which can take time to complete, the resulting data on the target volume is presented so that the copy appears to complete immediately.

This process is performed by using a bitmap (or bit array) that tracks changes to the data after the FlashCopy is started, and an indirection layer that enables data to be read from the source volume transparently.

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 Data Reduction Pools 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.

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 Spectrum Virtualize based 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.

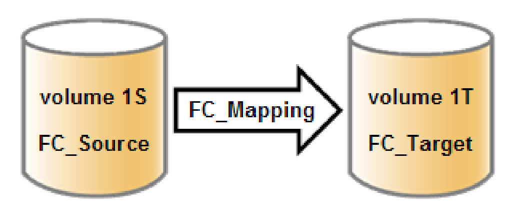


Figure 6-1 FlashCopy mapping

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 shows 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 illustrates a Consistency Group consisting of two volume mappings.

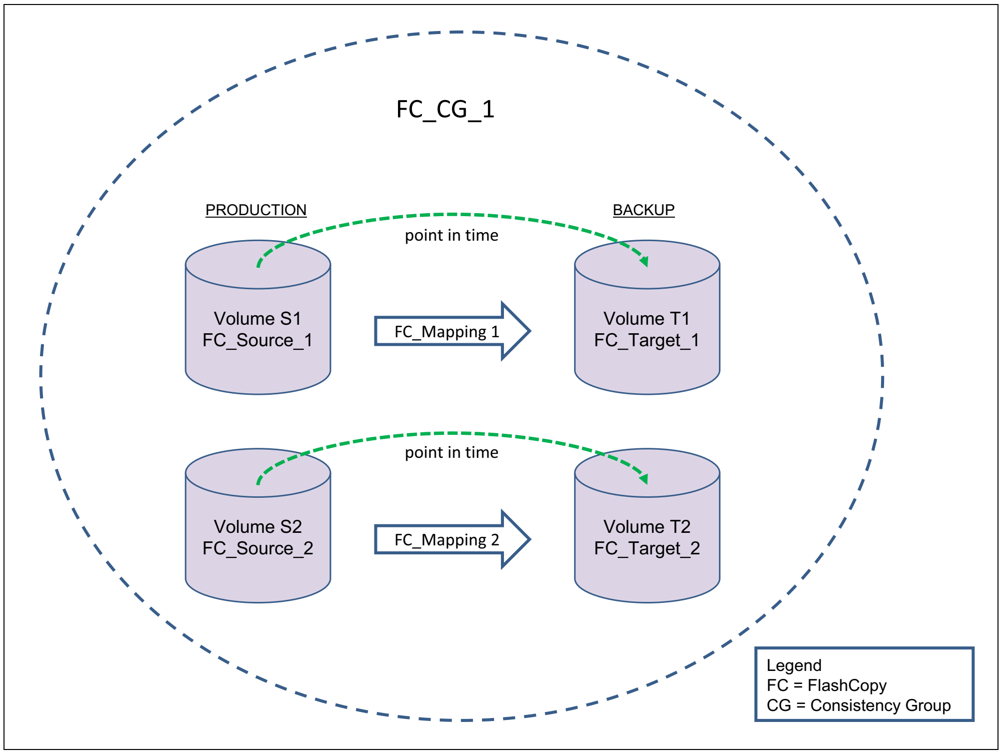


Figure 6-2 Multiple volumes mapping in a Consistency Group

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

Using Incremental FlashCopy, you can reduce the required time of copy. Also, because less data must be copied, the workload put on the system and the back-end storage is reduced.

Basically, Incremental FlashCopy does not require that you copy an entire disk source volume every time the FlashCopy mapping is started. It means that only the changed regions on source volumes are copied to target volumes, as shown in Figure 6-3.

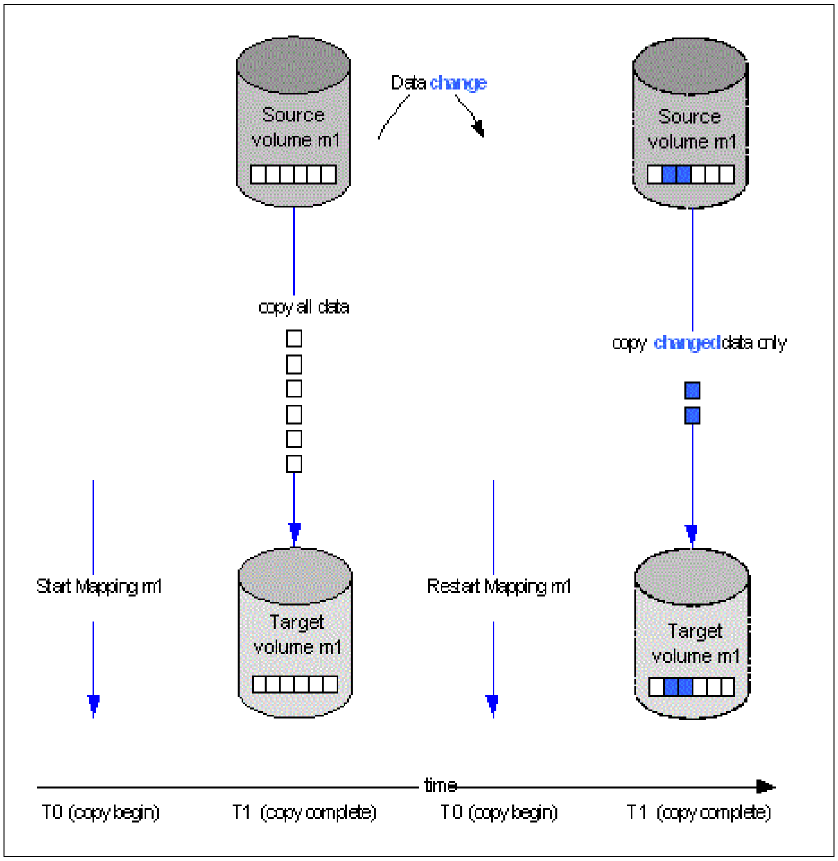


Figure 6-3   Incremental FlashCopy

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.

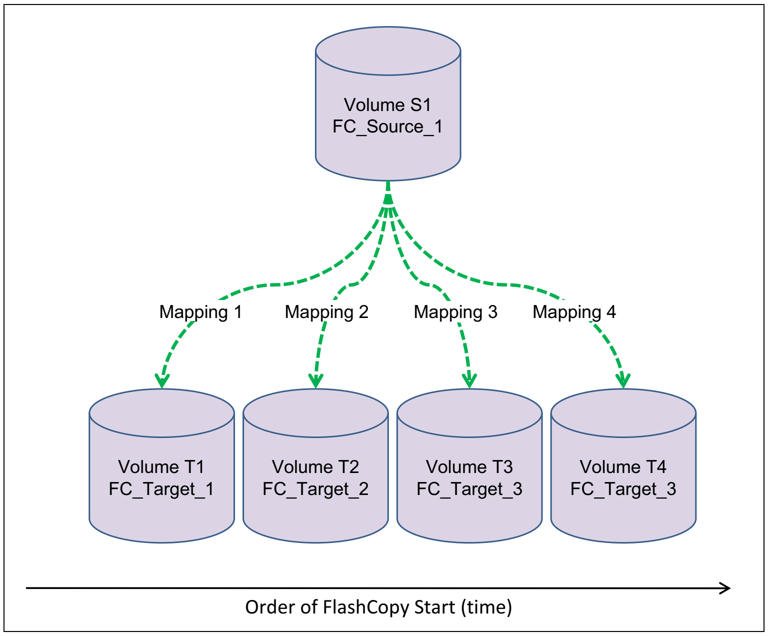


Figure 6-4 Multiple Target FlashCopy

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.

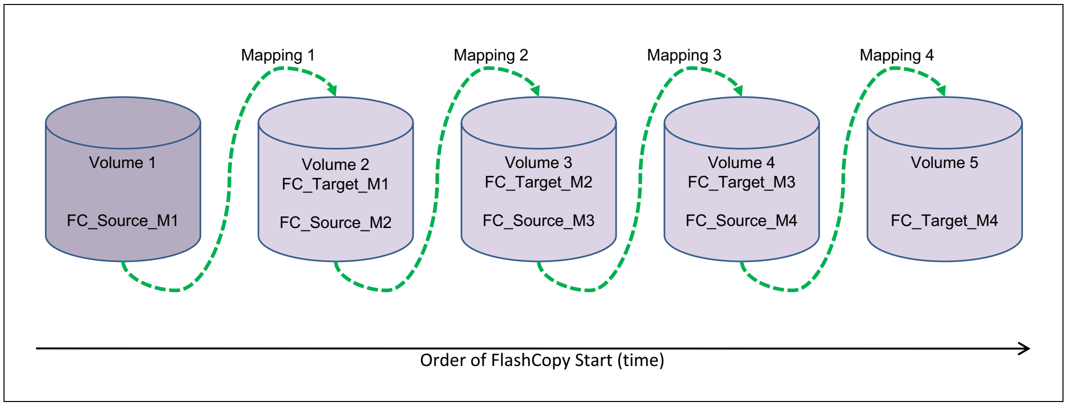


Figure 6-5 Cascaded FlashCopy

A total of 255 mappings are possible for each cascade.

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.

Data Reduction Pools (DRP) is a no charge feature that offers significant benefits on implementing data reduction techniques such as compression and data deduplication (see Chapter 4, “Storage Pools” on page 55). When using thin provisioned volumes on DRP also consider implementing compression as it provides benefits such as:

- 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 FS9100.
- Space efficiency as the compressed data provides more capacity savings.
- Better back-end capacity monitoring, as DRP pools with thin provisioned uncompressed volumes doesn’t 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.

Reverse FlashCopy

Reverse FlashCopy enables FlashCopy targets to become restore points for the source without breaking the FlashCopy relationship, and without having to wait for the original copy operation to complete. Therefore, it supports multiple targets (up to 256) and multiple rollback points.

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

IBM Spectrum Virtualize based 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.

6.2.3 FlashCopy functional overview

Understanding how FlashCopy works internally helps you to configure it in a way that you want and enables you to obtain more benefits from it.

FlashCopy bitmaps and grains

A *bitmap* is an internal data structure stored in a particular I/O Group that is used to track which data in FlashCopy mappings has been copied from the source volume to the target volume. *Grains* are units of data grouped together to optimize the use of the bitmap. One bit in each bitmap represents the state of one grain. FlashCopy grain can be either 64 KB or 256 KB.

A FlashCopy bitmap takes up the bitmap space in the memory of the I/O group that must be shared with other features’ bitmaps (such as Remote Copy bitmaps, Volume Mirroring bitmaps, and RAID bitmaps).

Indirection layer

The *FlashCopy indirection layer* governs the I/O to the source and target volumes when a FlashCopy mapping is started. This process is done by using a FlashCopy bitmap. The purpose of the FlashCopy indirection layer is to enable both the source and target volumes for read and write I/O immediately after FlashCopy starts.

The following description illustrates how the FlashCopy indirection layer works when a FlashCopy mapping is prepared and then started.

When a FlashCopy mapping is prepared and started, the following sequence is applied:

1. Flush the write cache to the source volume or volumes that are part of a Consistency Group.
2. Put the cache into write-through mode on the source volumes.
3. Discard the cache for the target volumes.
4. Establish a sync point on all of the source volumes in the Consistency Group (creating the FlashCopy bitmap).
5. Ensure that the indirection layer governs all of the I/O to the source volumes and target.
6. Enable the cache on source volumes and target volumes.

FlashCopy provides the semantics of a point-in-time copy that uses the indirection layer, which intercepts I/O that is directed at either the source or target volumes. The act of starting a FlashCopy mapping causes this indirection layer to become active in the I/O path, which occurs automatically across all FlashCopy mappings in the Consistency Group. The indirection layer then determines how each of the I/O is to be routed based on the following factors:

- The volume and the logical block address (LBA) to which the I/O is addressed
- Its direction (read or write)
- The state of an internal data structure, the FlashCopy bitmap

The indirection layer allows the I/O to go through the underlying volume. It redirects the I/O from the target volume to the source volume, or queues the I/O while it arranges for data to be copied from the source volume to the target volume. The process of queueing the write operations on the source volume while the indirection layer completes the grain copy on the target volume is called *copy-on-write*.

Table 6-2 summarizes the indirection layer algorithm.

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.

Volume being accessed	Has the grain been copied?	Host I/O operation	
		Read	Write
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

Cache is divided into upper and lower cache. Upper cache serves mostly as write cache and hides the write latency from the hosts and application. Lower cache is a read/write cache and optimizes I/O to and from disks.

Figure 6-6 shows the IBM Spectrum Virtualize cache architecture.

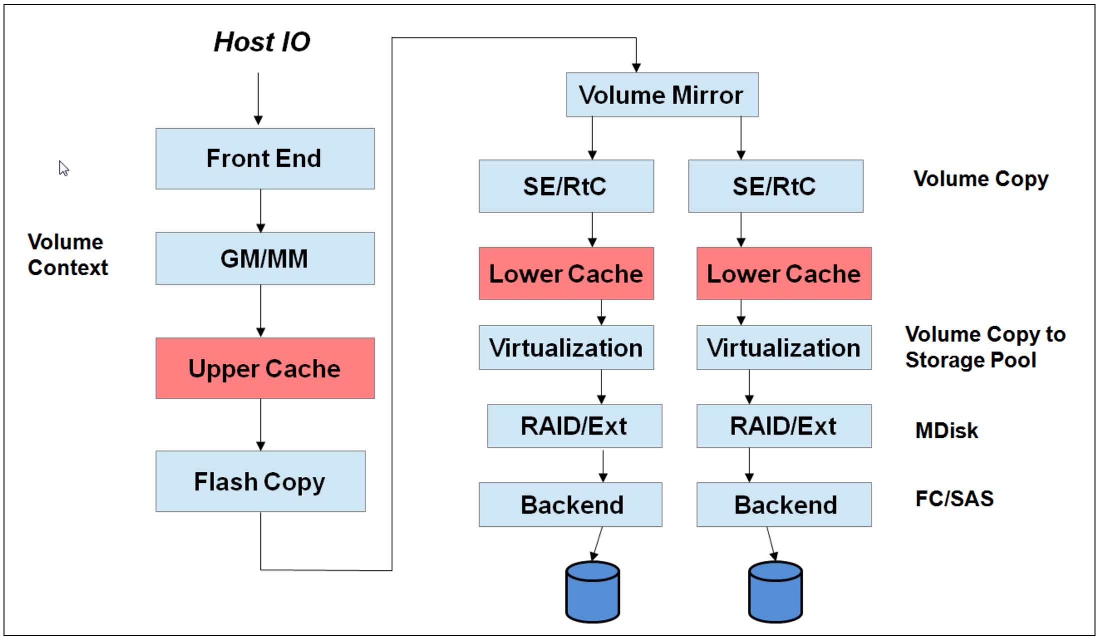


Figure 6-6 New cache architecture

The copy-on-write process introduces significant latency into write operations. To isolate the active application from this additional latency, the FlashCopy indirection layer is placed logically between the upper and lower cache. Therefore, the additional latency that is introduced by the copy-on-write process is encountered only by the internal cache operations, and not by the application.

The logical placement of the FlashCopy indirection layer is shown in Figure 6-7.

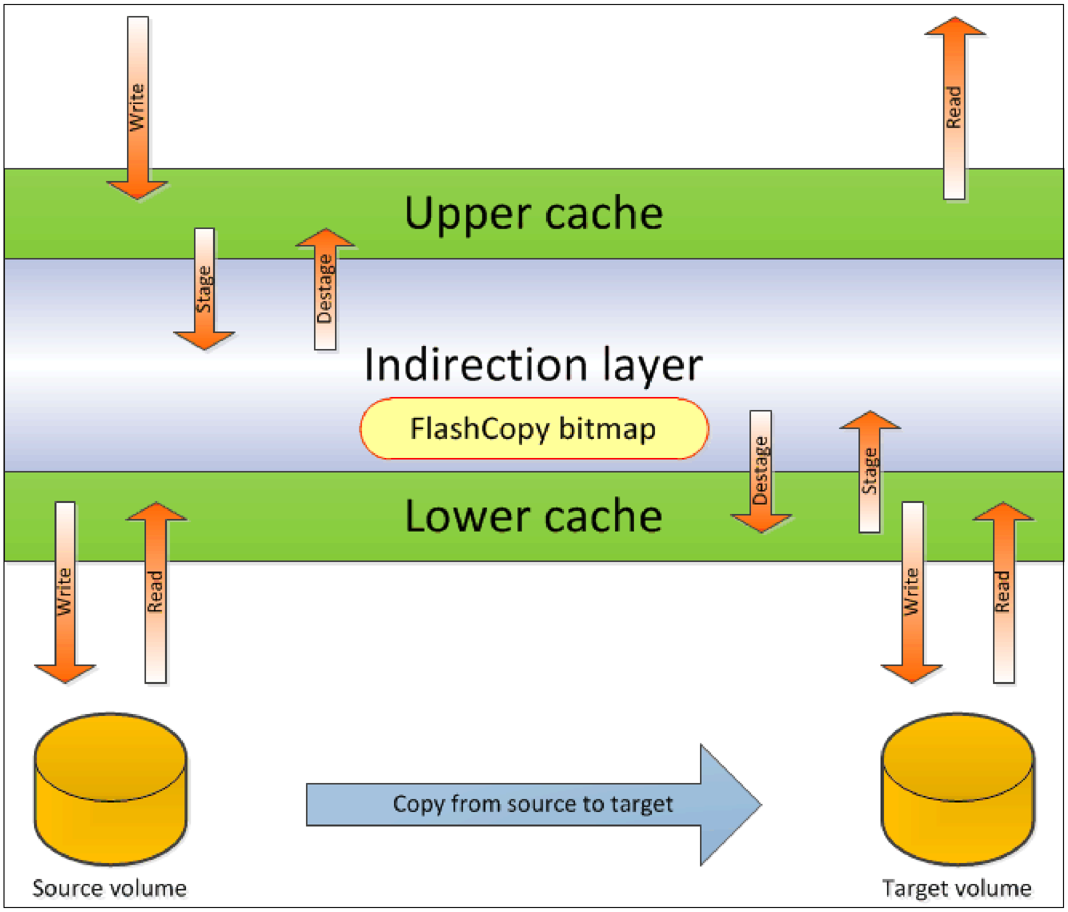


Figure 6-7 Logical placement of the FlashCopy indirection layer

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

Additionally, in multi-target FlashCopy, the target volumes of the same image share cache data. This design is opposite to previous IBM Spectrum Virtualize code versions, where each volume had its own copy of cached data.

Interaction and dependency between Multiple Target FlashCopy mappings

Figure 6-8 on page 155 represents a set of three FlashCopy mappings that share a common source. The FlashCopy mappings target volumes Target 1, Target 2, and Target 3.

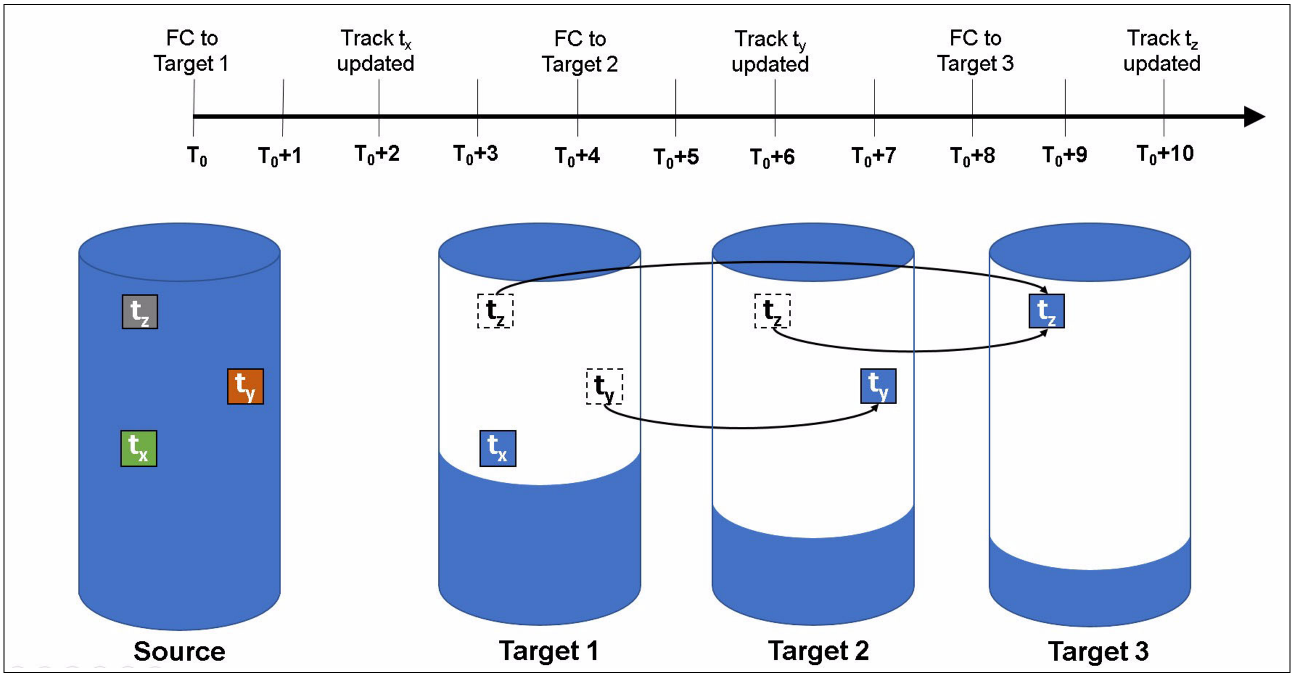


Figure 6-8 Interaction between Multiple Target FlashCopy mappings

Consider the following events timeline:

- At time T_0 a Flashcopy mapping is started between the source and the Target 1.
- At time T_0+2 the track t_x is updated in the source. Since the this track has not yet been copied in background on Target 1, the copy-on-write process copies this track to the Target 1 before being updated on the source.
- At time T_0+4 a Flashcopy mapping is started between the source and the Target 2.
- At time T_0+6 the track t_y is updated in the source. Because this track has not yet been copied in background on Target 2, the copy-on-write process copies this track to the Target 2 only before being updated on the source.
- At time T_0+8 a Flashcopy mapping is started between the source and the Target 3.
- At time T_0+10 the track t_z is updated in the source. Because this track has not yet been copied in background on Target 3, the copy-on-write process copies this track to the Target 3 only before being updated on the source.

As result of this sequence of events, the configuration in Figure 6-8 has these characteristics:

- Target 1 is dependent upon Target 2 and Target 3. It remains dependent until all of Target 1 has been copied. No target depends on Target 1, so the mapping can be stopped without need to copy any data to maintain the consistency in the other targets.
- Target 2 depends on Target 3, and will remain dependent until all of Target 2 has been copied. Target 1 depends on Target 2, so if this mapping is stopped, the cleanup process is started to copy all data that is uniquely held on this mapping (that is t_y) to Target 1.
- Target 3 is not dependent on any target, but it has Target 1 and Target 2 depending on it, so if this mapping is stopped the cleanup process is started to copy all data that is uniquely held on this mapping (that is t_z) to Target 2.

Target writes with Multiple Target FlashCopy

A write to an intermediate or newest target volume must consider the state of the grain within its own mapping, and the state of the grain of the next oldest mapping:

- If the grain of the next oldest mapping has not been copied yet, it must be copied before the write is allowed to proceed to preserve the contents of the next oldest mapping. The data that is written to the next oldest mapping comes from a target or source.
- If the grain in the target being written has not yet been copied, the grain is copied from the oldest already copied grain in the mappings that are newer than the target, or the source if none are already copied. After this copy is done, the write can be applied to the target.

Target reads with Multiple Target FlashCopy

If the grain being read has already been copied from the source to the target, the read simply returns data from the target being read. If the grain has not been copied, each of the newer mappings is examined in turn and the read is performed from the first copy found. If none are found, the read is performed from the source.

6.2.4 FlashCopy planning considerations

The FlashCopy function, like all the advanced IBM Spectrum Virtualize product features, offers useful capabilities. However, some basic planning considerations are to be followed for a successful implementation.

FlashCopy configurations limits

To plan for and implement FlashCopy, you must check the configuration limits and adhere to them. Table 6-3 shows the system limits that apply to the latest version at the time of writing this book.

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	5000	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. Check the IBM FS9100 online documentation for the latest configuration limits.

The total amount of cache memory reserved for the FlashCopy bitmaps limits the amount of capacity that can be used as a FlashCopy target. Table 6-4 illustrates the relationship of bitmap space to FlashCopy address space, depending on the size of the grain and the kind of FlashCopy service being used.

Table 6-4 Relationship of bitmap space to FlashCopy address space for the specified I/O Group

Copy Service	**Grain size in KB**	**1 MB of memory provides the following volume capacity for the specified I/O Group**
FlashCopy	256	2 TB of target volume capacity
FlashCopy	64	512 GB of target volume capacity
Incremental FlashCopy	256	1 TB of target volume capacity
Incremental FlashCopy	64	256 GB of target volume capacity

Mapping consideration: For multiple FlashCopy targets, you must consider the number of mappings. For example, for a mapping with a 256 KB grain size, 8 KB of memory allows one mapping between a 16 GB source volume and a 16 GB target volume. Alternatively, for a mapping with a 256 KB grain size, 8 KB of memory allows two mappings between one 8 GB source volume and two 8 GB target volumes.

When you create a FlashCopy mapping, if you specify an I/O Group other than the I/O Group of the source volume, the memory accounting goes towards the specified I/O Group, not towards the I/O Group of the source volume.

The default amount of memory for FlashCopy is 20 MB. This value can be increased or decreased by using the `chiogrp` command. The maximum amount of memory that can be specified for FlashCopy is 2048 MB (512 MB for 32-bit systems). The maximum combined amount of memory across all copy services features is 2600 MB (552 MB for 32-bit systems).

Bitmap allocation: When creating a FlashCopy mapping, you can optionally specify the I/O group where the bitmap is allocated. If you specify an I/O Group other than the I/O Group of the source volume, the memory accounting goes towards the specified I/O Group, not towards the I/O Group of the source volume. This option can be useful when an I/O group is exhausting the memory that is allocated to the FlashCopy bitmaps and no more free memory is available in the I/O group.

FlashCopy general restrictions

The following implementation restrictions apply to FlashCopy:

- The size of source and target volumes in a FlashCopy mapping must be the same.
- Multiple FlashCopy mappings that use the same target volume can be defined, but only one of these mappings can be started at a time. This limitation means that no multiple FlashCopy can be active to the same target volume.
- Expansion or shrinking of volumes defined in a FlashCopy mapping is not allowed. To modify the size of a source or target volume, first remove the FlashCopy mapping.
- In a cascading FlashCopy, the grain size of all the FlashCopy mappings that participate must be the same.
- In a multi-target FlashCopy, the grain size of all the FlashCopy mappings that participate must be the same.

- In a reverse FlashCopy, the grain size of all the FlashCopy mappings that participate must be the same.
- No FlashCopy mapping can be added to a consistency group while the FlashCopy mapping status is `Copying`.
- No FlashCopy mapping can be added to a consistency group while the consistency group status is `Copying`.
- The use of Consistency Groups is restricted when using Cascading FlashCopy. A Consistency Group serves the purpose of starting FlashCopy mappings at the same point in time. Within the *same* Consistency Group, it is not possible to have mappings with these conditions:

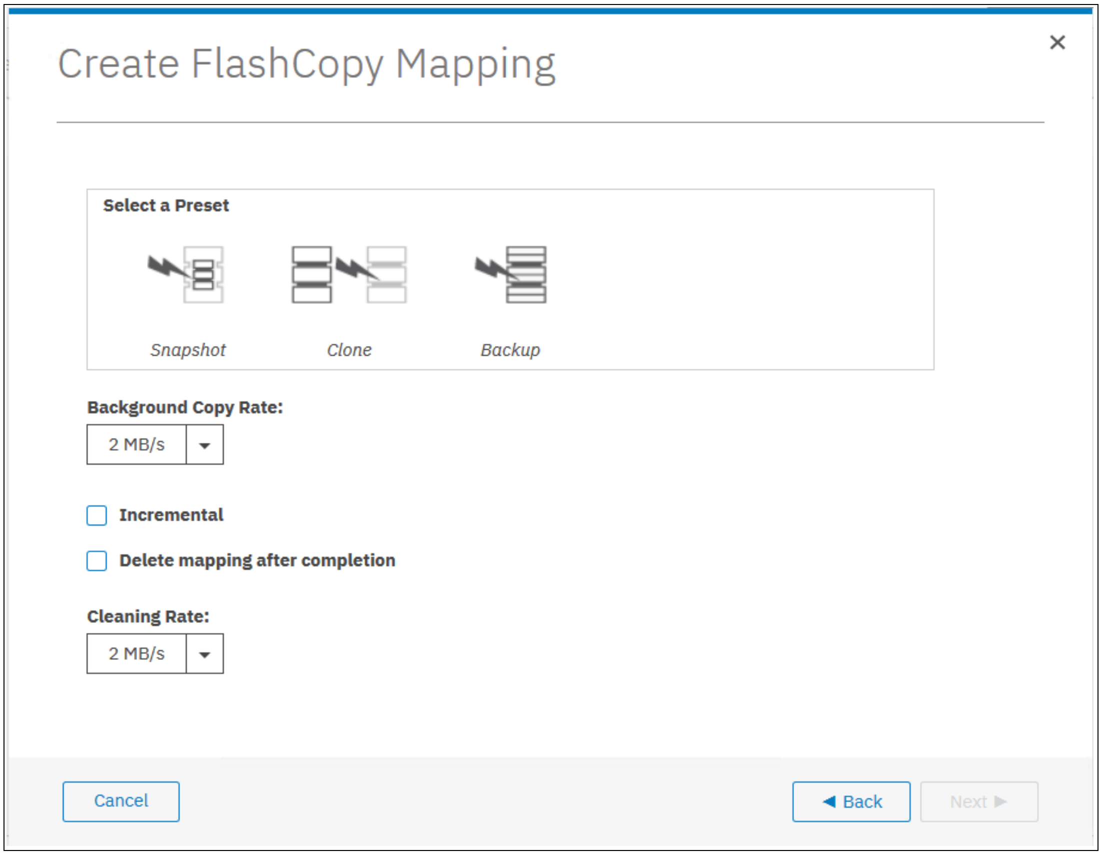
  - The source volume of one mapping is the target of another mapping.
  - The target volume of one mapping is the source volume for another mapping.

  These combinations are not useful because within a Consistency Group, mappings cannot be established in a certain order. This limitation renders the content of the target volume undefined. For instance, it is not possible to determine whether the first mapping was established before the target volume of the first mapping that acts as a source volume for the second mapping.

  Even if it were possible to ensure the order in which the mappings are established within a Consistency Group, the result is equal to Multi Target FlashCopy (two volumes holding the same target data for one source volume). In other words, a cascade is useful for copying volumes in a certain order (and copying the changed content targets of FlashCopies), rather than at the same time in an undefined order (from within one single Consistency Group).
- Both source and target volumes can be used as primary in a Remote Copy relationship. For more details about the FlashCopy and the Remote Copy possible interactions see “Interaction between Remote Copy and FlashCopy” on page 191.

FlashCopy presets

The IBM FS9100 GUI interface provides three FlashCopy presets (Snapshot, Clone, and Backup) to simplify the more common FlashCopy operations. Figure 6-9 on page 159 shows the preset selection panel in the GUI.



Create FlashCopy Mapping

Select a Preset

Background Copy Rate: 2 MB/s

□ Incremental

□ Delete mapping after completion

Cleaning Rate: 2 MB/s

Figure 6-9 GUI Flashcopy Presets

Although these presets meet most FlashCopy requirements, they do not provide support for all possible FlashCopy options. If more specialized options are required that are not supported by the presets, the options must be performed by using CLI commands.

This section describes the three preset options and their use cases.

Snapshot

This preset creates a copy-on-write point-in-time copy. The snapshot is not intended to be an independent copy. Instead, the copy is used to maintain a view of the production data at the time that the snapshot is created. Therefore, the snapshot holds only the data from regions of the production volume that have changed since the snapshot was created. Because the snapshot preset uses thin provisioning, only the capacity that is required for the changes is used.

Snapshot uses the following preset parameters:

- Background copy: `None`
- Incremental: `No`
- Delete after completion: `No`
- Cleaning rate: `No`
- Primary copy source pool: `Target pool`

A typical use case for the Snapshot is when the user wants to produce a copy of a volume without affecting the availability of the volume. The user does not anticipate many changes to be made to the source or target volume. A significant proportion of the volumes remains unchanged.

By ensuring that only changes require a copy of data to be made, the total amount of disk space that is required for the copy is reduced. Therefore, many Snapshot copies can be used in the environment.

Snapshots are useful for providing protection against corruption or similar issues with the validity of the data. However, they do not provide protection from physical controller failures. Snapshots can also provide a vehicle for performing repeatable testing (including “what-if” modeling that is based on production data) without requiring a full copy of the data to be provisioned.

Clone

The clone preset creates a replica of the volume, which can then be changed without affecting the original volume. After the copy completes, the mapping that was created by the preset is automatically deleted.

Clone uses the following preset parameters:

- Background copy rate: `50`
- Incremental: `No`
- Delete after completion: `Yes`
- Cleaning rate: `50`
- Primary copy source pool: `Target pool`

A typical use case for the Snapshot is when users want a copy of the volume that they can modify without affecting the original volume. After the clone is established, there is no expectation that it is refreshed or that there is any further need to reference the original production data again. If the source is thin-provisioned, the target is thin-provisioned for the auto-create target.

Backup

The backup preset creates a point-in-time replica of the production data. After the copy completes, the backup view can be refreshed from the production data, with minimal copying of data from the production volume to the backup volume.

Backup uses the following preset parameters:

- Background Copy rate: `50`
- Incremental: `Yes`
- Delete after completion: `No`
- Cleaning rate: `50`
- Primary copy source pool: `Target pool`

The Backup preset can be used when the user wants to create a copy of the volume that can be used as a backup if the source becomes unavailable. This unavailability can happen during loss of the underlying physical controller. The user plans to periodically update the secondary copy, and does not want to suffer from the resource demands of creating a new copy each time.

Incremental FlashCopy times are faster than full copy, which helps to reduce the window where the new backup is not yet fully effective. If the source is thin-provisioned, the target is also thin-provisioned in this option for the auto-create target.

Another use case, which is not supported by the name, is to create and maintain (periodically refresh) an independent image. This image can be subjected to intensive I/O (for example, data mining) without affecting the source volume’s performance.

Grain size considerations

When creating a mapping a grain size of 64 KB can be specified as compared to the default 256 KB. This smaller grain size has been introduced specifically for the incremental FlashCopy, even though its use is not restricted to the incremental mappings.

In an incremental FlashCopy, the modified data is identified by using the bitmaps. The amount of data to be copied when refreshing the mapping depends on the grain size. If the grain size is 64 KB, as compared to 256 KB, there might be less data to copy to get a fully independent copy of the source again.

> **Incremental FlashCopy:** For incremental FlashCopy, the 64 KB grain size is preferred.

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

The following are the preferred settings for thin-provisioned FlashCopy:

- Thin-provisioned volume grain size must be equal to the FlashCopy grain size.
- Thin-provisioned volume grain size must be 64 KB for the best performance and the best space efficiency.

The exception is where the thin target volume is going to become a production volume (and is likely to be subjected to ongoing heavy I/O). In this case, the 256 KB thin-provisioned grain size is preferable because it provides better long-term I/O performance at the expense of a slower initial copy.

> **FlashCopy grain size considerations:** Even if the 256 KB thin-provisioned volume grain size is chosen, it is still beneficial to limit the FlashCopy grain size to 64 KB. It is possible to minimize the performance impact to the source volume, even though this size increases the I/O workload on the target volume.
>
> However, clients with very large numbers of FlashCopy/Remote Copy relationships might still be forced to choose a 256 KB grain size for FlashCopy to avoid constraints on the amount of bitmap memory.

Volume placement considerations

The source and target volumes placement among the pools and the I/O groups must be planned to minimize the effect of the underlying FlashCopy processes. In normal condition (that is with all the nodes/canisters fully operative), the FlashCopy background copy workload distribution follows this schema:

- The preferred node of the source volume is responsible for the background copy read operations.
- The preferred node of the target volume is responsible for the background copy write operations.

Table 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

Note that the data transfer among the source and the target volume's preferred nodes occurs through the node-to-node connectivity. Consider the following volume placement alternatives:

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

   In this scenario, the node that is acting as preferred for both source and target volume manages all the read and write FlashCopy operations. Only resources from this node are consumed for the FlashCopy operations, and no node-to-node bandwidth is used.

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

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

Both alternatives described have advantages and disadvantages, but in general option 1 is preferred. Consider the following exceptions:

- A clustered IBM FS9100 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 FS9100 system with multiple control enclosures where the source volumes are evenly spread across all the canisters.

  In this case the preferred node placement should follow the location of source and target volumes on the internal storage. For example, if the source volume is on the internal storage attached to control enclosure A and the target volume is on internal storage attached to control enclosure B, then the target volumes preferred node must be in one canister of control enclosure B. Placing the target volumes preferred node on control enclosure A will cause the re-direction of the FlashCopy write operation through the node-to-node network.

Placement on the back-end storage is mainly driven by the availability requirements. Generally, use different back-end storage controllers or arrays for the source and target volumes.

Background copy considerations

The background copy process uses internal resources such as CPU, memory, and bandwidth. This copy process tries to reach the target copy data rate for every volume according to the background copy rate parameter setting (as reported in Table 6-1 on page 146).

If the copy process is unable to achieve these goals, it starts contending resources to the foreground I/O (that is the I/O coming from the hosts). As result, both background copy and foreground I/O will tend to see an increase in latency and therefore reduction in throughput compared to the situation when the bandwidth not been limited. Degradation is graceful. Both background copy and foreground I/O continue to make progress, and will not stop, hang, or cause the node to fail.

To avoid any impact on the foreground I/O, that is in the hosts response time, carefully plan the background copy activity, taking in account the overall workload running in the systems. The background copy basically reads and writes data to managed disks. Usually, the most affected component is the back-end storage. CPU and memory are not normally significantly affected by the copy activity.

The theoretical added workload due to the background copy is easily estimable. For instance, starting 20 FlashCopy with a background copy rate of 70 each adds a maximum throughput of 160 MBps for the reads and 160 MBps for the writes.

The source and target volumes distribution on the back-end storage determines where this workload is going to be added. The duration of the background copy depends on the amount of data to be copied. This amount is the total size of volumes for full background copy or the amount of data that is modified for incremental copy refresh.

Performance monitoring tools like IBM Spectrum Control can be used to evaluate the existing workload on the back-end storage in a specific time window. By adding this workload to the foreseen background copy workload, you can estimate the overall workload running toward the back-end storage. Disk performance simulation tools, like Disk Magic, can be used to estimate the effect, if any, of the added back-end workload to the host service time during the background copy window. The outcomes of this analysis can provide useful hints for the background copy rate settings.

When performance monitoring and simulation tools are not available, use a conservative and progressive approach. Consider that the background copy setting can be modified at any time, even when the FlashCopy is already started. The background copy process can even be completely stopped by setting the background copy rate to `0`.

Initially set the background copy rate value to add a limited workload to the 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.

There is an interaction between the background copy rate and the Cleaning Rate settings:

- Background copy = 0 and Cleaning Rate = 0

  No background copy or cleaning take place. When the mapping is stopped, it goes into stopping state and a cleaning process starts with the default cleaning rate, which is 50 or 2 MBps.

- Background copy > 0 and Cleaning Rate = 0

  The background copy takes place at the background copy rate but no cleaning process is started. When the mapping is stopped, it goes into stopping state and a cleaning process starts with the default cleaning rate (50 or 2 MBps).

- Background copy = 0 and Cleaning Rate > 0

  No background copy takes place, but the cleaning process runs at the cleaning rate. When the mapping is stopped, the cleaning completes (if not yet completed) at the cleaning rate.

- Background copy > 0 and Cleaning Rate > 0

  The background copy takes place at the background copy rate but no cleaning process is started. When the mapping is stopped, it goes into stopping state and a cleaning process starts with the specified cleaning rate.

Regarding the workload considerations for the cleaning process, the same guidelines as for background copy apply.

Host and application considerations to ensure FlashCopy integrity

Because FlashCopy is at the block level, it is necessary to understand the interaction between your application and the host operating system. From a logical standpoint, it is easiest to think of these objects as “layers” that sit on top of one another. The application is the topmost layer, and beneath it is the operating system layer.

Both of these layers have various levels and methods of caching data to provide better speed. Because IBM Spectrum Virtualize systems, and therefore FlashCopy, sit below these layers, they are unaware of the cache at the application or operating system layers.

To ensure the integrity of the copy that is made, it is necessary to flush the host operating system and application cache for any outstanding reads or writes before the FlashCopy operation is performed. Failing to flush the host operating system and application cache produces what is referred to as a *crash consistent* copy.

The resulting copy requires the same type of recovery procedure, such as log replay and file system checks, that is required following a host crash. FlashCopies that are crash consistent often can be used following file system and application recovery procedures.

> **Note:** Although the best way to perform FlashCopy is to flush host cache first, some companies, like Oracle, support using snapshots without it, as stated in Metalink note 604683.1.

Various operating systems and applications provide facilities to stop I/O operations and ensure that all data is flushed from host cache. If these facilities are available, they can be used to prepare for a FlashCopy operation. When this type of facility is not available, the host cache must be flushed manually by quiescing the application and unmounting the file system or drives.

Preferred practice: From a practical standpoint, when you have an application that is backed by a database and you want to make a FlashCopy of that application's data, it is sufficient in most cases to use the write-suspend method that is available in most modern databases. You can use this method because the database maintains strict control over I/O.

This method is as opposed to flushing data from both the application and the backing database, which is always the suggested method because it is safer. However, this method can be used when facilities do not exist or your environment includes time sensitivity.

6.3 Remote Copy services

IBM Spectrum Virtualize technology offers various remote copy services functions that address Disaster Recovery and Business Continuity needs.

Metro Mirror is designed for metropolitan distances with a zero recovery point objective (RPO), which is zero data loss. This objective is achieved with a synchronous copy of volumes. Writes are not acknowledged until they are committed to both storage systems. By definition, any vendors' synchronous replication makes the host wait for write I/Os to complete at both the local and remote storage systems, and includes round-trip network latencies. Metro Mirror has the following characteristics:

- Zero RPO
- Synchronous
- Production application performance that is affected by round-trip latency

Global Mirror is designed to minimize application performance impact by replicating asynchronously. That is, writes are acknowledged as soon as they can be committed to the local storage system, sequence-tagged, and passed on to the replication network. This technique allows Global Mirror to be used over longer distances. By definition, any vendors' asynchronous replication results in an RPO greater than zero. However, for Global Mirror, the RPO is quite small, typically anywhere from several milliseconds to some number of seconds.

Although Global Mirror is asynchronous, the RPO is still small, and thus the network and the remote storage system must both still be able to cope with peaks in traffic. Global Mirror has the following characteristics:

- Near-zero RPO
- Asynchronous
- Production application performance that is affected by I/O sequencing preparation time

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

- Larger RPO
- Point-in-time copies
- Asynchronous
- Possible system performance effect because point-in-time copies are created locally

Successful implementation depends on taking a holistic approach in which you consider all components and their associated properties. The components and properties include host application sensitivity, local and remote SAN configurations, local and remote system and storage configuration, and the intersystem network.

6.3.1 Remote copy functional overview

In this section, the terminology and the basic functional aspects of the remote copy services are presented.

Common terminology and definitions

When such a breadth of technology areas is covered, the same technology component can have multiple terms and definitions. This document uses the following definitions:

- *Local system* or *master system*

  The system on which the foreground applications run.

- *Local hosts*

  Hosts that run on the foreground applications.

- *Master volume* or *source volume*

  The local volume that is being mirrored. The volume has nonrestricted access. Mapped hosts can read and write to the volume.

- *Intersystem link* or *intersystem network*

  The network that provides connectivity between the local and the remote site. It can be a Fibre Channel network (SAN), an IP network, or a combination of the two.

- *Remote system* or *auxiliary system*

  The system that holds the remote mirrored copy.

- *Auxiliary volume* or *target volume*

  The remote volume that holds the mirrored copy. It is read-access only.

- *Remote copy*

  A generic term that is used to describe a Metro Mirror or Global Mirror relationship in which data on the source volume is mirrored to an identical copy on a target volume. Often the two copies are separated by some distance, which is why the term *remote* is used to describe the copies. However, having remote copies is not a prerequisite. A remote copy relationship includes the following states:

  - Consistent relationship

    A remote copy relationship where the data set on the target volume represents a data set on the source volumes at a certain point.

  - Synchronized relationship

    A relationship is *synchronized* if it is consistent *and* the point that the target volume represents is the current point. The target volume contains identical data as the source volume.

- *Synchronous remote copy* (Metro Mirror)

  Writes to the source and target volumes that are committed in the foreground before confirmation is sent about completion to the local host application.

- *Asynchronous remote copy* (Global Mirror)

  A foreground write I/O is acknowledged as complete to the local host application before the mirrored foreground write I/O is cached at the remote system. Mirrored foreground writes are processed asynchronously at the remote system, but in a committed sequential order as determined and managed by the Global Mirror remote copy process.

- *Global Mirror Change Volume*

  Holds earlier consistent revisions of data when changes are made. A change volume must be created for the master volume and the auxiliary volume of the relationship.

- The *background copy* process manages the initial synchronization or resynchronization processes between source volumes to target mirrored volumes on a remote system.
- *Foreground I/O* reads and writes I/O on a local SAN, which generates a mirrored foreground write I/O that is across the intersystem network and remote SAN.

Figure 6-10 shows some of the concepts of remote copy.

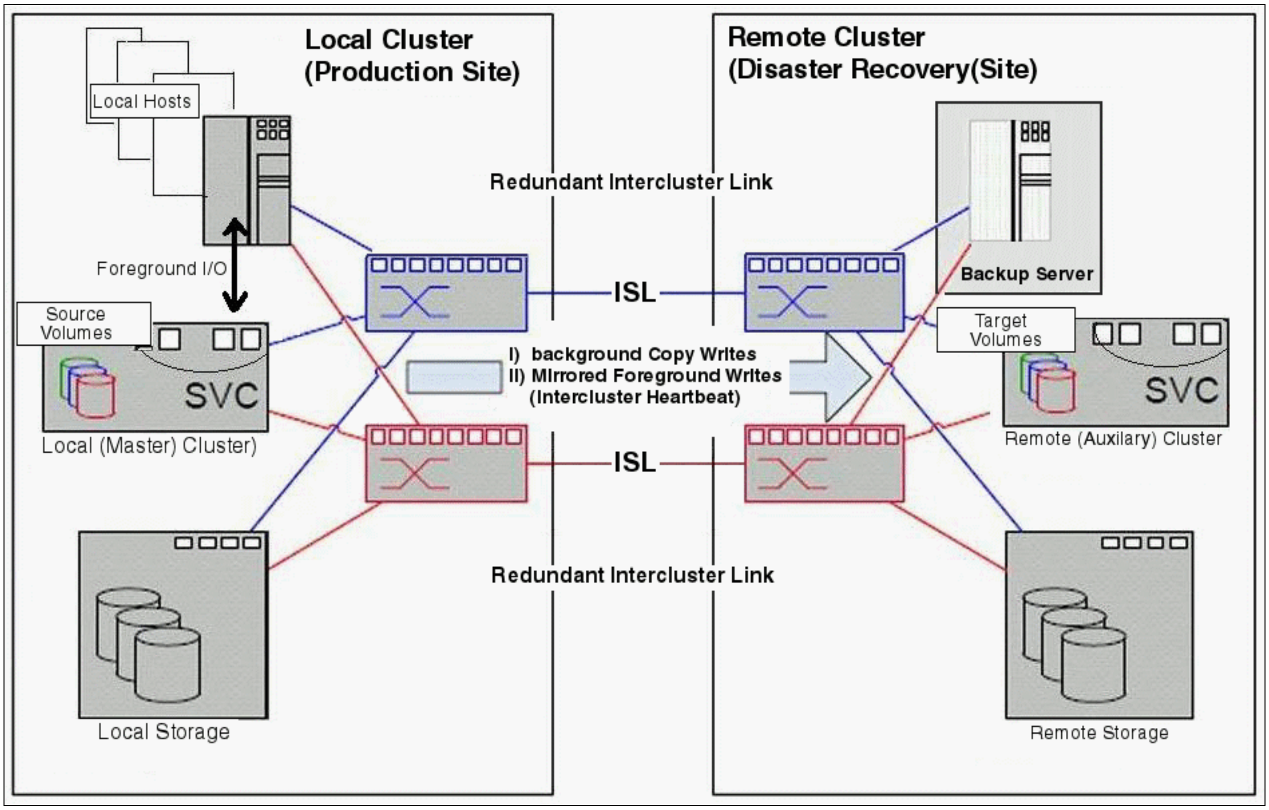


Figure 6-10 Remote copy components and applications

A successful implementation of intersystem remote copy services significantly depends on the quality and configuration of the intersystem network.

Remote Copy partnerships and relationships

A remote copy *partnership* is a partnership that is established between a master (local) system and an auxiliary (remote) system, as shown in Figure 6-11.

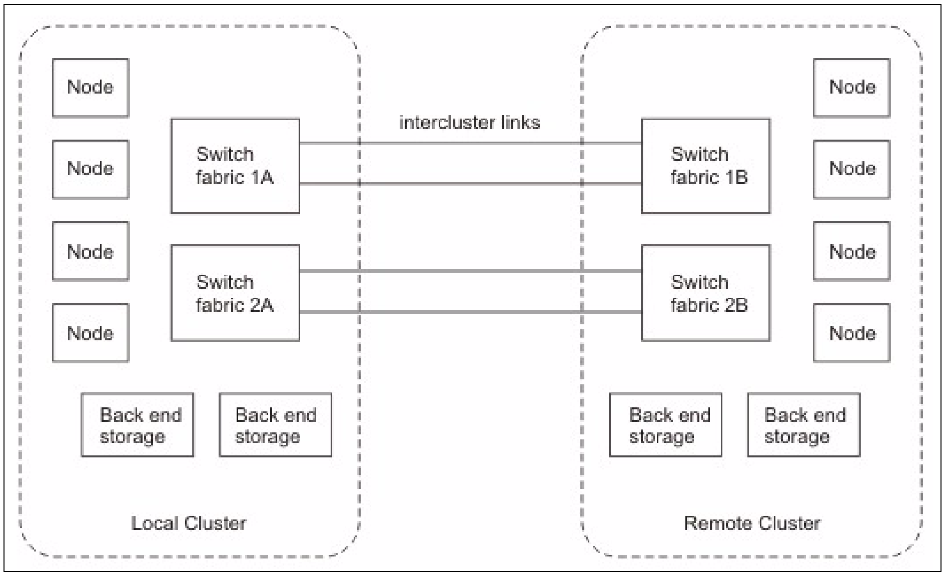


Figure 6-11 Remote copy partnership

Partnerships are established between two systems by issuing the **`mkfcpartnership`** or **`mkippartnership`** command once from each end of the partnership. The parameters that need to be specified are the remote system name (or ID), the available bandwidth (in Mbps), and the maximum background copy rate as a percentage of the available bandwidth. The background copy parameter determines the maximum speed of the initial synchronization and resynchronization of the relationships.

> **Tip:** To establish a fully functional Metro Mirror or Global Mirror partnership, issue the **`mkfcpartnership`** or **`mkippartnership`** command from both systems.

A remote copy *relationship* is a relationship that is established between a source (primary) volume in the local system and a target (secondary) volume in the remote system. Usually when a remote copy relationship is started, a background copy process that copies the data from source to target volumes is started as well.

In addition to the background copy rate setting, the initial synchronization can be adjusted at relationship level with the `relationship_bandwidth_limit` parameter. The `relationship_bandwidth_limit` is a system-wide parameter that sets the maximum bandwidth that can be used to initially synchronize a single relationship.

After background synchronization or resynchronization is complete, a Global Mirror relationship provides and maintains a consistent mirrored copy of a source volume to a target volume.

Copy directions and default roles

When you create a remote copy relationship, the source or master volume is initially assigned the role of the master, and the target auxiliary volume is initially assigned the role of the auxiliary. This design implies that the initial copy direction of mirrored foreground writes and background resynchronization writes (if applicable) is from master to auxiliary.

After the initial synchronization is complete, you can change the copy direction (see Figure 6-12). The ability to change roles is used to facilitate disaster recovery.

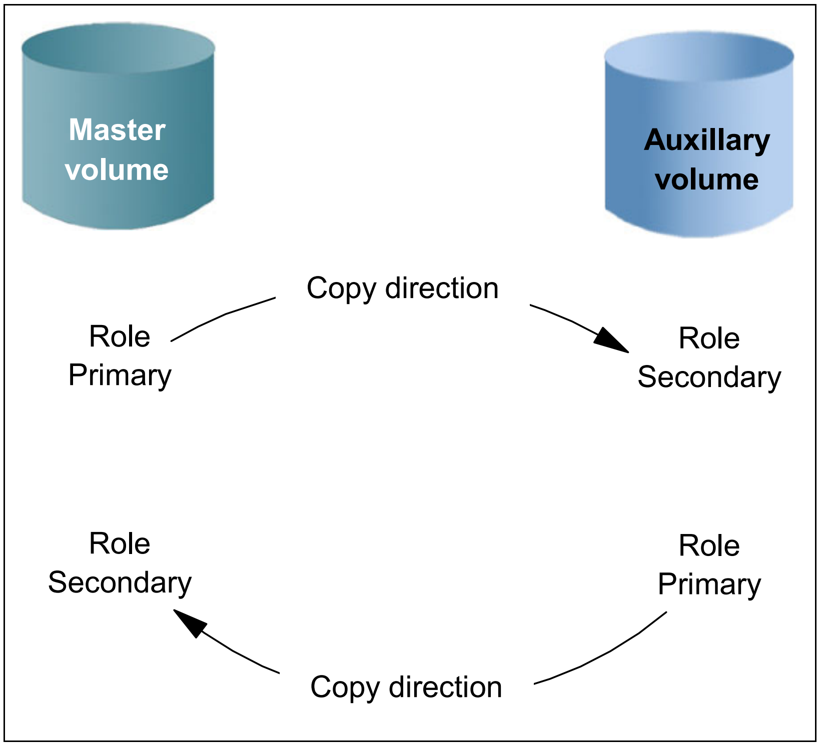


Figure 6-12 Role and direction changes

> **Attention:** When the direction of the relationship is changed, the roles of the volumes are altered. A consequence is that the read/write properties are also changed, meaning that the master volume takes on a secondary role and becomes read-only.

Consistency Groups

A Consistency Group (CG) is a collection of relationships that can be treated as one entity. This technique is used to preserve write order consistency across a group of volumes that pertain to one application, for example, a database volume and a database log file volume.

After a remote copy relationship is added into a Consistency Group, you cannot manage the relationship in isolation from the Consistency Group. So, for example, issuing a **`stoprcrelationship`** command on the stand-alone volume would fail because the system knows that the relationship is part of a Consistency Group.

Note the following points regarding Consistency Groups:

- Each volume relationship can belong to only one Consistency Group.
- Volume relationships can also be stand-alone, that is, not in any Consistency Group.
- Consistency Groups can also be created and left empty, or can contain one or many relationships.
- You can create up to 256 Consistency Groups on a system.
- All volume relationships in a Consistency Group must have matching primary and secondary systems, but they do not need to share I/O groups.

- All relationships in a Consistency Group have the same copy direction and state.
- Each Consistency Group is either for Metro Mirror or for Global Mirror relationships, but not both. This choice is determined by the first volume relationship that is added to the Consistency Group.

Consistency Group consideration: A Consistency Group relationship does not have to be in a directly matching I/O group number at each site. A Consistency Group owned by I/O group 1 at the local site does not have to be owned by I/O group 1 at the remote site. If you have more than one I/O group at either site, you can create the relationship between any two I/O groups. This technique spreads the workload, for example, from local I/O group 1 to remote I/O group 2.

Streams

Consistency Groups can also be used as a way to spread replication workload across multiple streams within a partnership.

The Metro or Global Mirror partnership architecture allocates traffic from each Consistency Group in a round-robin fashion across 16 streams. That is, cg0 traffic goes into stream0, and cg1 traffic goes into stream1.

Any volume that is *not* in a Consistency Group also goes into stream0. You might want to consider creating an empty Consistency Group 0 so that stand-alone volumes do not share a stream with active Consistency Group volumes.

It can also pay to optimize your streams by creating more Consistency Groups. Within each stream, each batch of writes must be processed in tag sequence order and any delays in processing any particular write also delays the writes behind it in the stream. Having more streams (up to 16) reduces this kind of potential congestion.

Each stream is sequence-tag-processed by one node, so generally you would want to create at least as many Consistency Groups as you have IBM FS9100 canisters, and, ideally, perfect multiples of the node count.

Layer concept

IBM FS9100 has the concept of *layer*, which allows you to create partnerships among IBM Spectrum Virtualize based products. The key points concerning layers are listed here:

- IBM SAN Volume Controller is always in the *Replication* layer.
- By default, FS9100 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 a FS9100 system only if the FS9100 is in Storage layer.
- An IBM FS9100 system in the Replication layer can virtualize a FS9100/Storwize system in the Storage layer.

Figure 6-13 illustrates the concept of layers.

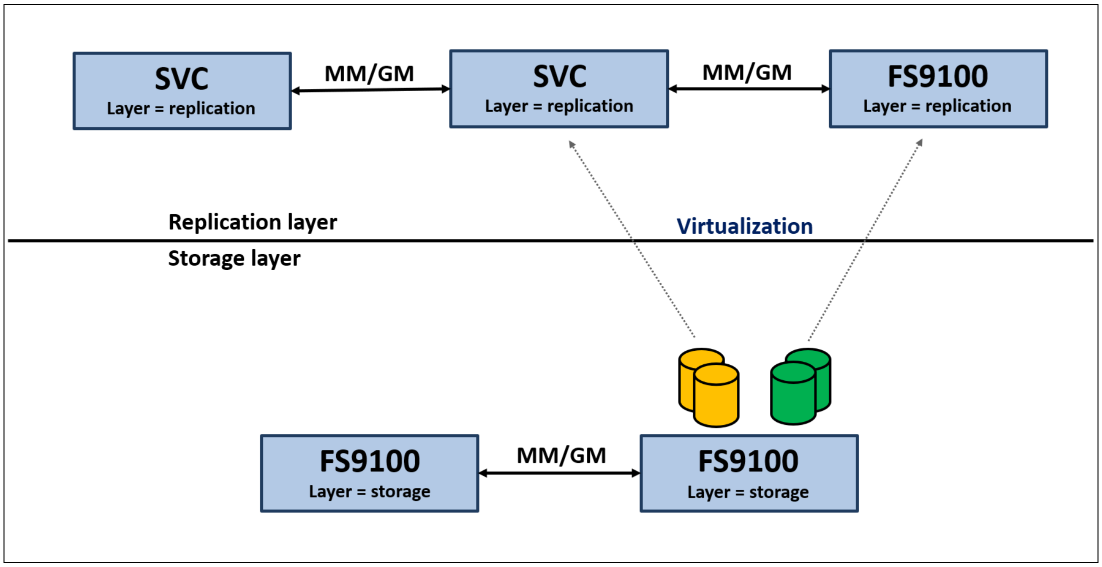


Figure 6-13 Conceptualization of layers

Generally, changing the layer is only performed at initial setup time or as part of a major reconfiguration. To change the layer of a FS9100 system, the system must meet the following pre-conditions:

- The FS9100 system must not have any IBM Spectrum Virtualize, Storwize or FS9100 host objects defined, and must not be virtualizing any other FS9100/Storwize controllers.
- The FS9100 system must not be visible to any other IBM Spectrum Virtualize, Storwize or FS9100 system in the SAN fabric, which might require SAN zoning changes.
- The FS9100 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 FS9100 system from Storage layer to Replication layer can only be performed by using the CLI. After you are certain that all of the pre-conditions have been met, issue the following command:

```
chsystem -layer replication
```

Partnership topologies

Each system can be connected to a maximum of three other systems for the purposes of Metro or Global Mirror.

Figure 6-14 shows examples of the principal supported topologies for Metro and Global Mirror partnerships. Each box represents an IBM Spectrum Virtualize based system.

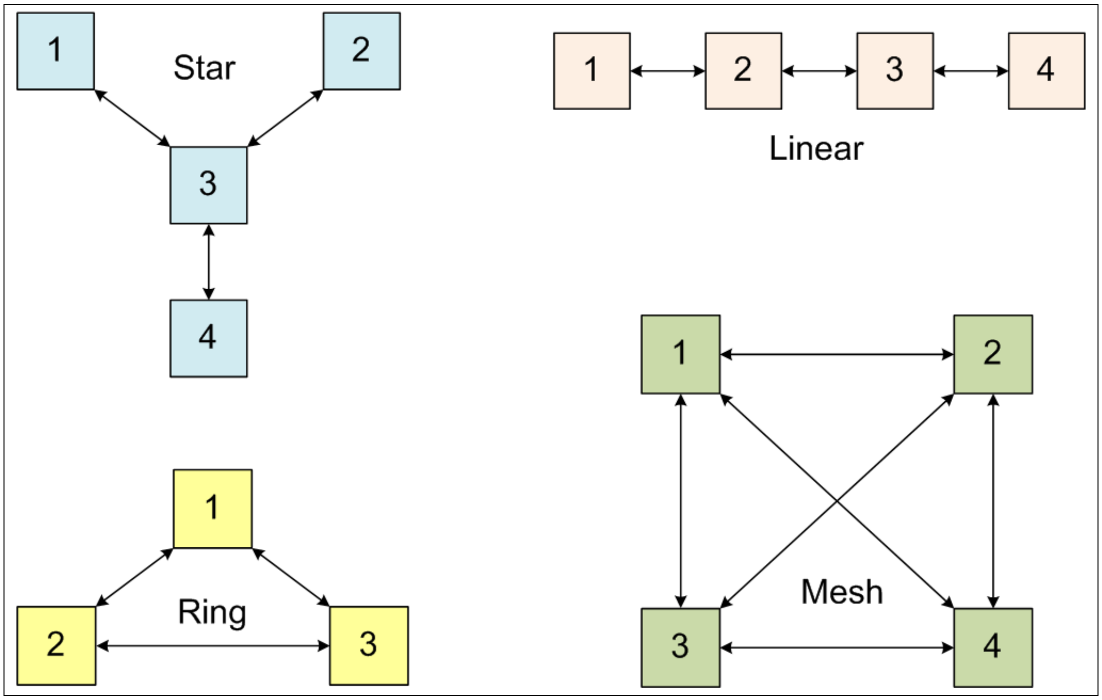


Figure 6-14 Supported topologies for Metro and Global Mirror

Star topology

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

Ring topology

A ring topology (3 or more systems) can be used to establish a one-in, one-out implementation. For example, the implementation can be 1 → 2, 2 → 3, 3 → 1 to spread replication loads evenly among three systems.

Linear topology

A linear topology of two or more sites is also possible. However, it would generally be simpler to create partnerships between system 1 and system 2, and separately between system 3 and system 4.

Mesh topology

A fully connected mesh topology is where every system has a partnership to each of the three other systems. This topology allows flexibility in that volumes can be replicated between any two systems.

Topology considerations:

- Although systems can have up to three partnerships, any one volume can be part of only a single relationship. That is, you cannot replicate any given volume to multiple remote sites.
- Although various topologies are supported, it is advisable to keep your partnerships as simple as possible, which in most cases means system pairs or a star.

Intrasystem versus intersystem

Although remote copy services are available for intrasystem, it has no functional value for production use. Intrasystem Metro Mirror provides the same capability with less overhead. However, leaving this function in place simplifies testing and allows for experimentation and testing. For example, you can validate server failover on a single test system.

> **Intrasystem remote copy:** Intrasystem Global Mirror is not supported on IBM Spectrum Virtualize based systems that run V6 or later.

Metro Mirror functional overview

Metro Mirror provides synchronous replication. It is designed to ensure that updates are committed to both the primary and secondary volumes before sending an acknowledgment (Ack) of the completion to the server.

If the primary volume fails completely for any reason, Metro Mirror is designed to ensure that the secondary volume holds the same data as the primary did immediately before the failure.

Metro Mirror provides the simplest way to maintain an identical copy on both the primary and secondary volumes. However, as with any synchronous copy over long distance, there can be a performance impact to host applications due to network latency.

Metro Mirror supports relationships between volumes that are up to 300 km apart. Latency is an important consideration for any Metro Mirror network. With typical fiber optic round-trip latencies of 1 ms per 100 km, you can expect a minimum of 3 ms extra latency, due to the network alone, on each I/O if you are running across the 300 km separation.

Figure 6-15 shows the order of Metro Mirror write operations.

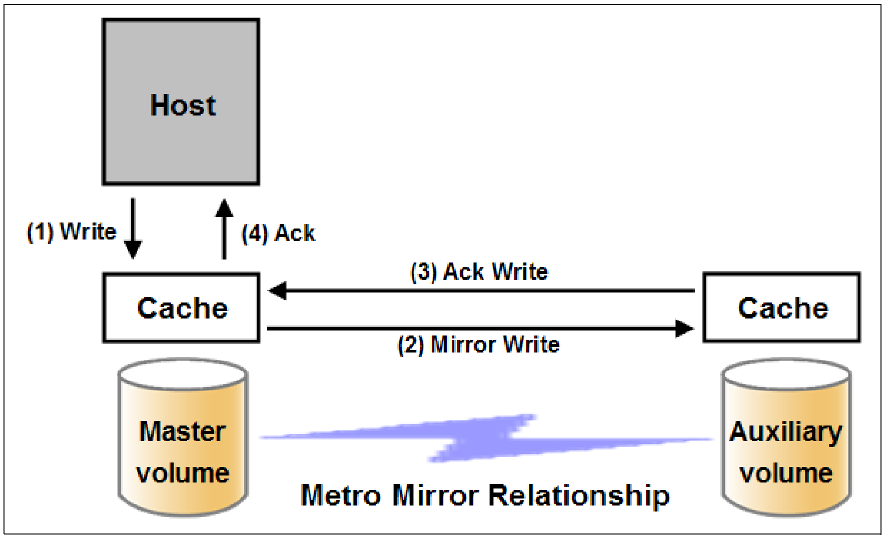


Figure 6-15 Metro Mirror write sequence

A write into mirrored cache on an IBM Spectrum Virtualize based system is all that is required for the write to be considered as committed. De-staging to disk is a natural part of I/O management, but it is not generally in the critical path for a Metro Mirror write acknowledgment.

Global Mirror functional overview

Global Mirror provides asynchronous replication. It is designed to reduce the dependency on round-trip network latency by acknowledging the primary write in parallel with sending the write to the secondary volume.

If the primary volume fails completely for any reason, Global Mirror is designed to ensure that the secondary volume holds the same data as the primary did at a point a short time before the failure. That short period of data loss is typically between 10 ms and 10 seconds, but varies according to individual circumstances.

Global Mirror provides a way to maintain a write-order-consistent copy of data at a secondary site only slightly behind the primary. Global Mirror has minimal impact on the performance of the primary volume.

Although Global Mirror is an asynchronous remote copy technique, foreground writes at the local system and mirrored foreground writes at the remote system are not wholly independent of one another. IBM Spectrum Virtualize implementation of asynchronous remote copy uses algorithms to maintain a consistent image at the target volume always.

They achieve this image by identifying sets of I/Os that are active concurrently at the source, assigning an order to those sets, and applying these sets of I/Os in the assigned order at the target. The multiple I/Os within a single set are applied concurrently.

The process that marshals the sequential sets of I/Os operates at the remote system, and therefore is not subject to the latency of the long-distance link.

Figure 6-16 shows that a write operation to the master volume is acknowledged back to the host that issues the write before the write operation is mirrored to the cache for the auxiliary volume.

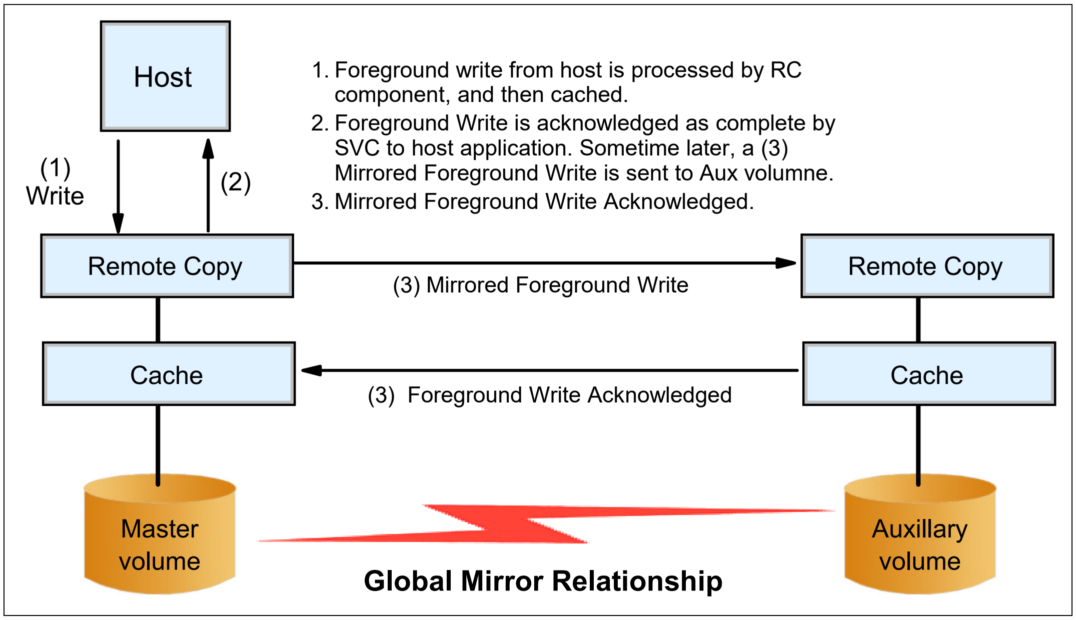

Host
1. Foreground write from host is processed by RC component, and then cached.
2. Foreground Write is acknowledged as complete by SVC to host application. Sometime later, a (3) Mirrored Foreground Write is sent to Aux volumne.
3. Mirrored Foreground Write Acknowledged.
(1) Write
(2)
Remote Copy
(3) Mirrored Foreground Write
Remote Copy
Cache
(3) Foreground Write Acknowledged
Cache
Master volume
Auxillary volume
Global Mirror Relationship

Figure 6-16 Global Mirror relationship write operation

With Global Mirror, a confirmation is sent to the host server before the host receives a confirmation of the completion at the auxiliary volume. The GM function identifies sets of write I/Os that are active concurrently at the primary volume. It then assigns an order to those sets, and applies these sets of I/Os in the assigned order at the auxiliary volume.

Further writes might be received from a host when the secondary write is still active for the same block. In this case, although the primary write might complete, the new host write on the auxiliary volume is delayed until the previous write is completed.

Write ordering

Many applications that use block storage are required to survive failures, such as a loss of power or a software crash. They are also required to not lose data that existed before the failure. Because many applications must perform many update operations in parallel to that storage block, maintaining write ordering is key to ensuring the correct operation of applications after a disruption.

An application that performs a high volume of database updates is often designed with the concept of dependent writes. Dependent writes ensure that an earlier write completes before a later write starts. Reversing the order of dependent writes can undermine the algorithms of the application and can lead to problems, such as detected or undetected data corruption.

Colliding writes

Colliding writes are defined as new write I/Os that overlap existing active write I/Os.

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 6-17 shows a colliding write sequence.

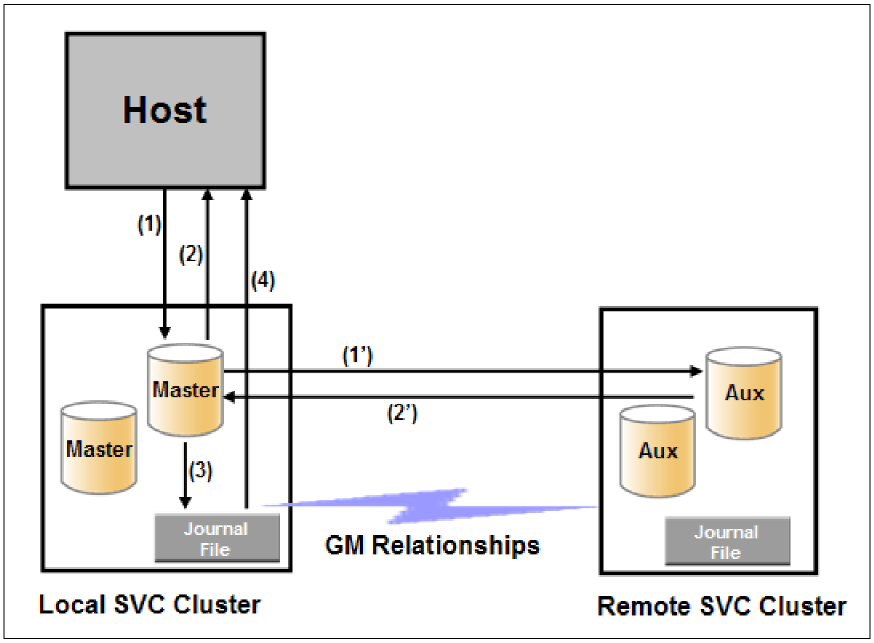


Figure 6-17 Colliding writes

The following numbers correspond to the numbers that are shown in Figure 6-17:

1. A first write is performed from the host to LBA X.
2. A host is provided acknowledgment that the write is complete, even though the mirrored write to the auxiliary volume is not yet completed.

   The first two actions (1 and 2) occur asynchronously with the first write.
3. A second write is performed from the host to LBA X. If this write occurs before the host receives acknowledgment (2), the write is written to the journal file.
4. A host is provided acknowledgment that the second write is complete.

Global Mirror Change Volumes functional overview

Global Mirror with Change Volumes (GM/CV) provides asynchronous replication based on point-in-time copies of data. It is designed to allow for effective replication over lower bandwidth networks and to reduce any impact on production hosts.

Metro Mirror and Global Mirror both require the bandwidth to be sized to meet the peak workload. Global Mirror with Change Volumes must only be sized to meet the average workload across a cycle period.

Figure 6-18 shows a high-level conceptual view of Global Mirror with Change Volumes. GM/CV uses FlashCopy to maintain image consistency and to isolate host volumes from the replication process.

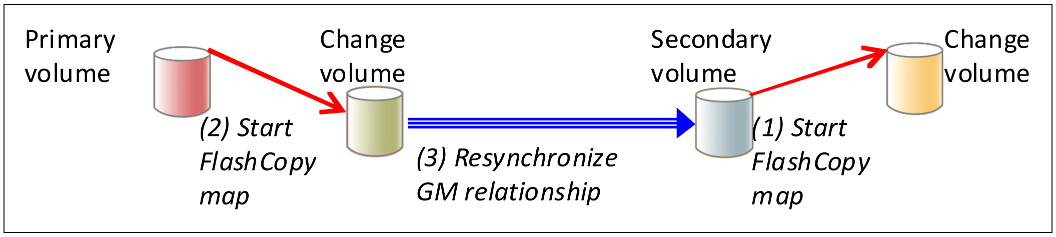


Figure 6-18 Global Mirror with Change Volumes

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

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

Change Volumes hold point-in-time copies of 256 KB grains. If any of the disk blocks in a grain change, that grain is copied to the change volume to preserve its contents. Change Volumes are also maintained at the secondary site so that a consistent copy of the volume is always available even when the secondary volume is being updated.

Primary and Change Volumes are always in the same I/O group and the Change Volumes are always thin-provisioned. Change Volumes cannot be mapped to hosts and used for host I/O, and they cannot be used as a source for any other FlashCopy or Global Mirror operations.

Figure 6-19 shows how a Change Volume is used to preserve a point-in-time data set, which is then replicated to a secondary site. The data at the secondary site is in turn preserved by a Change Volume until the next replication cycle has completed.

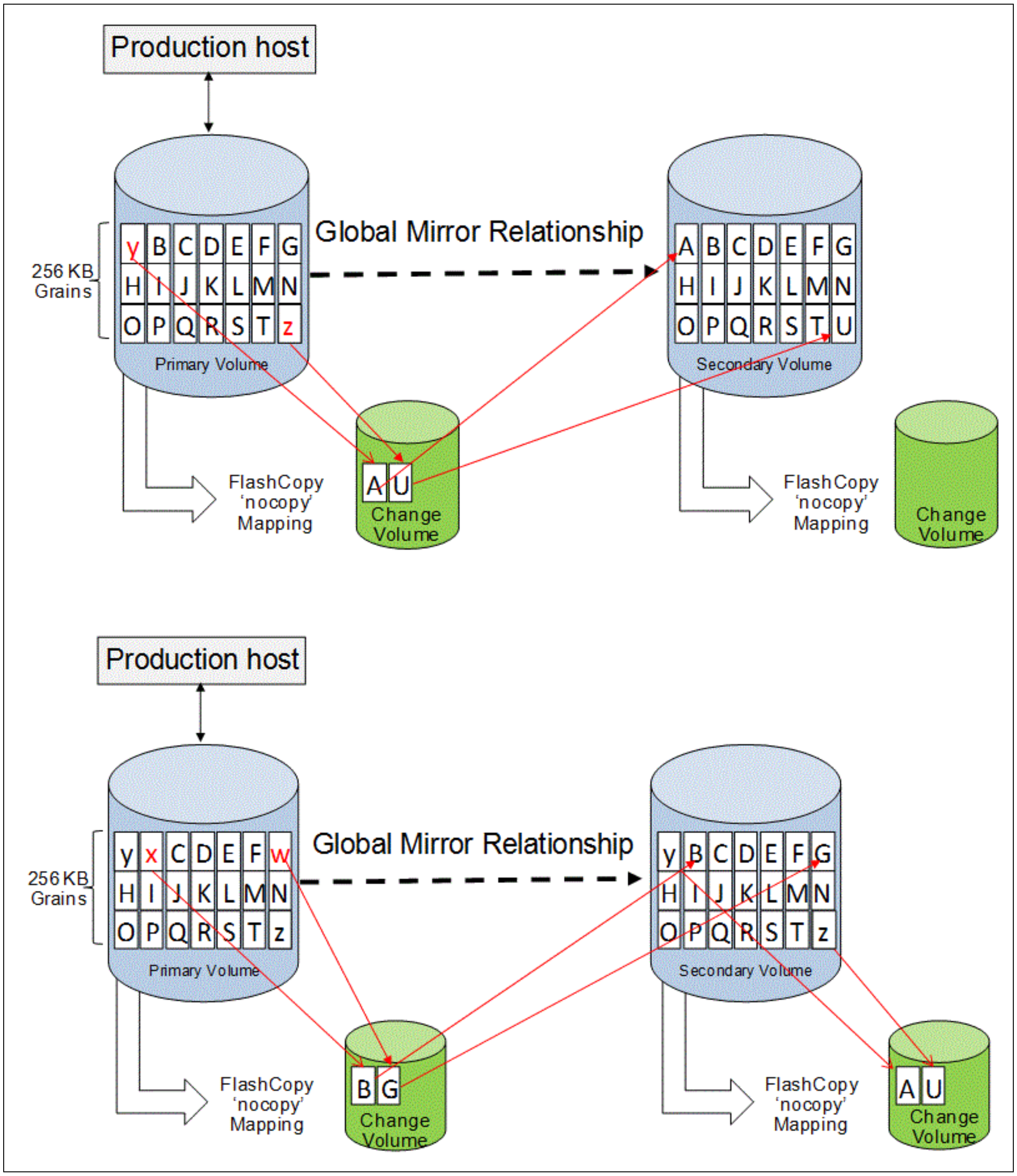


Figure 6-19 Global Mirror with Change Volumes uses FlashCopy point-in-time copy technology

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

The options for **`-cyclingmode`** are **`none`** and **`multi`**.

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

Specifying **multi** means that Global Mirror starts cycling based on the cycle period, which defaults to 300 seconds. The valid range is from 60 seconds to 24*60*60 seconds (86,400 seconds = one day).

If all of the changed grains cannot be copied to the secondary site within the specified time, then the replication is designed to take as long as it needs and to start the next replication as soon as the earlier one completes. You can choose to implement this approach by deliberately setting the cycle period to a short amount of time, which is a perfectly valid approach. However, remember that the shorter the cycle period, the less opportunity there is for peak write I/O smoothing, and the more bandwidth you need.

The **-cyclingmode** setting can only be changed when the Global Mirror relationship is in a stopped state.

Recovery point objective using Change Volumes

RPO is the maximum tolerable period in which data might be lost if you switch over to your secondary volume.

If a cycle completes within the specified cycle period, then the RPO is not more than 2x cycle long. However, if it does not complete within the cycle period, then the RPO is not more than the sum of the last two cycle times.

The current RPO can be determined by looking at the **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.

6.3.2 Remote Copy network planning

Remote copy partnerships and relationships do not work reliably if the connectivity on which they are running is configured incorrectly. This section focuses on the intersystem network, giving an overview of the remote system connectivity options.

Terminology

The intersystem network is specified in terms of *latency* and *bandwidth*. These parameters define the capabilities of the link regarding the traffic that is on it. They be must be chosen so that they support all forms of traffic, including mirrored foreground writes, background copy writes, and intersystem heartbeat messaging (node-to-node communication).

Link latency is the time that is taken by data to move across a network from one location to another and is measured in milliseconds. The longer the time, the greater the performance impact.

Tip: SCSI write over FC requires two round trips per I/O operation, as shown in the following example:

```
2 (round trips) x 2 (operations) x 5 microsec/km = 20 microsec/km
```

At 50 km, you have another latency, as shown in the following example:

```
20 microsec/km x 50 km = 1000 microsec = 1 msec (msec represents millisecond)
```

Each SCSI I/O has 1 ms of more service time. At 100 km, it becomes 2 ms for more service time.

Link bandwidth is the network capacity to move data as measured in millions of bits per second (Mbps) or billions of bits per second (Gbps).

The term *bandwidth* is also used in the following context:

- Storage bandwidth: The ability of the back-end storage to process I/O. Measures the amount of data (in bytes) that can be sent in a specified amount of time.
- Remote copy partnership bandwidth (parameter): The rate at which background write synchronization is attempted (unit of MBps).

Intersystem connectivity supports mirrored foreground and background I/O. A portion of the link is also used to carry traffic that is associated with the exchange of low-level messaging between the nodes of the local and remote systems. A *dedicated amount* of the link bandwidth is required for the exchange of heartbeat messages and the initial configuration of intersystem partnerships.

Interlink bandwidth must support the following traffic:

- Mirrored foreground writes, as generated by foreground processes at peak times
- Background write synchronization, as defined by the Global Mirror bandwidth parameter
- Intersystem communication (*heartbeat messaging)*

Fibre Channel connectivity is the standard connectivity that is used for the remote copy intersystem networks. It uses the Fibre Channel protocol and SAN infrastructures to interconnect the systems.

Native IP connectivity has been introduced with IBM Spectrum Virtualize version 7.2 to implement intersystem networks by using standard TPC/IP infrastructures.

Network latency considerations

The maximum supported round-trip latency between sites depends on the type of partnership between systems. Table 6-6 on page 180 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 Fibre Channel ports on each node that will be used for replication must be dedicated for replication traffic. This configuration can be achieved by using SAN zoning and port masking.
- SAN zoning should be applied to provide separate intrasystem zones for each local-remote I/O group pair that is used for replication. See "Remote system ports and zoning considerations" on page 187 for further zoning guidelines.

Link bandwidth that is used by internode communication

IBM Spectrum Virtualize uses part of the bandwidth for its internal intersystem heartbeat. The amount of traffic depends on how many nodes are in each of the local and remote systems. Table 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 intersystem heartbeat traffic (megabits per second)

Local or remote system	Two nodes	Four nodes	Six nodes	Eight nodes
Two nodes	5	6	6	6
Four nodes	6	10	11	12
Six nodes	6	11	16	17
Eight nodes	6	12	17	21

These numbers represent the total traffic between the two systems when *no* I/O is occurring to a mirrored volume on the remote system. Half of the data is sent by one system, and half of the data is sent by the other system. The traffic is divided evenly over all available connections. Therefore, if you have two redundant links, half of this traffic is sent over each link during fault-free operation.

If the link between the sites is configured with redundancy to tolerate single failures, size the link so that the bandwidth and latency statements continue to be accurate even during single failure conditions.

Network sizing considerations

Proper network sizing is essential for the remote copy services operations. Failing to estimate the network sizing requirements can lead to poor performance in remote copy services and the production workload.

Consider that intersystem bandwidth should be capable of supporting the combined traffic of the following items:

- Mirrored foreground writes, as generated by your server applications at peak times
- Background resynchronization, for example, after a link outage
- Inter-system heartbeat

Calculating the required bandwidth is essentially a question of mathematics based on your current workloads, so it is advisable to start by assessing your current workloads.

For Metro or Global Mirror, you need to know your peak write rates and I/O sizes down to at least a 5-minute interval. This information can be easily gained from tools like IBM Spectrum Control. Finally, you need to allow for unexpected peaks.

There are also unsupported tools to help with sizing available from IBM:

Quick performance overview

Do not compromise on bandwidth or network quality when planning a Metro or Global Mirror deployment. If bandwidth is likely to be an issue in your environment, consider Global Mirror with Change Volumes.

Bandwidth sizing examples

As an example, consider a business with the following I/O profile:

- Average write size 8 KB (= 8 x 8 bits/1024 = 0.0625 Mb).
- For most of the day between 8 AM and 8 PM, the write activity is around 1500 writes per second.
- Twice a day (once in the morning and once in the afternoon), the system bursts up to 4500 writes per second for up to 10 minutes.
- Outside of the 8 AM to 8 PM window, there is little or no I/O write activity.

This example is intended to represent a general traffic pattern that might be common in many medium-sized sites. Futhermore, 20% of bandwidth must be left available for the background synchronization.

Here we consider options for Metro Mirror, Global Mirror, and for Global Mirror with Change Volumes based on a cycle period of 30 minutes and 60 minutes.

Metro Mirror or Global Mirror require bandwidth on the instantaneous peak of 4500 writes per second as follows:

```
4500 x 0.0625 = 282 Mbps + 20% resync allowance + 5 Mbps heartbeat = 343 Mbps
dedicated plus any safety margin plus growth
```

In the following two examples, the bandwidth for GM/CV needs to be able to handle the peak 30-minute period, or the peak 60-minute period.

GMCV peak 30-minute period example

If we look at this time broken into 10-minute periods, the peak 30-minute period is made up of one 10-minute period of 4500 writes per second, and two 10-minute periods of 1500 writes per second. The average write rate for the 30-minute cycle period can then be expressed mathematically as follows:

```
(4500 + 1500 + 1500) / 3 = 2500 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 period example

For a cycle period of 60 minutes, the peak 60-minute period is made up of one 10-minute period of 4500 writes per second, and five 10-minute periods of 1500 writes per second. The average write for the 60-minute cycle period can be expressed as follows:

```
(4500 + 5 x 1500) / 6 = 2000 writes/sec for a 60-minute cycle period
```

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

```
2000 x 0.0625 = 125 Mbps + 20% resync allowance + 5 Mbps heartbeat = 155 Mbps
dedicated plus any safety margin plus growth
```

Now consider whether the business does not have aggressive RPO requirements and does not want to provide dedicated bandwidth for Global Mirror. But the network is available and unused at night, so Global Mirror can use that. There is an element of risk here, which is if the network is unavailable for any reason, GM/CV cannot keep running during the day until it catches up. Therefore, you would need to allow a much higher resync allowance in your replication window, for example, 100 percent.

A GM/CV replication based on daily point-in-time copies at 8 PM each night, and replicating until 8 AM at the latest would probably require at least the following bandwidth:

```
(9000 + 70 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
```

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

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

Remember that these are examples. The central principle of sizing is that you need to know your data write rate, which is the number of write I/Os and the average size of those I/Os. For Metro Mirror and Global Mirror, you need to know the peak write I/O rates. For GM/CV, you need to know the average write I/O rates.

> **GM/CV bandwidth:** In the above samples, the bandwidth estimation for the GM/CV is based on the assumption that the write operations occurs in such a way that a change volume grain (that has a size of 256 KB) is completely changed before it is transferred to the remote site. In the real life, this situation is unlikely to occur.
>
> Usually only a portion of a grain is changed during a GMCV cycle, but the transfer process always copies the whole grain to the remote site. This behavior can lead to an unforeseen processor burden in the transfer bandwidth that, in the edge case, can be even higher than the one required for a standard Global Mirror.

Fibre Channel connectivity

You must remember several considerations when you use Fibre Channel technology for the intersystem network:

- Redundancy
- Basic topology and problems
- Switches and ISL oversubscription
- Distance extensions options

- Optical multiplexors
- Long-distance SFPs and XFPs
- Fibre Channel over IP
- Hops
- Buffer credits
- Remote system ports and zoning considerations

Redundancy

The intersystem network must adopt the same policy toward redundancy as for the local and remote systems to which it is connecting. The ISLs must have redundancy, and the individual ISLs must provide the necessary bandwidth in isolation.

Basic topology and problems

Because of the nature of Fibre Channel, you must avoid ISL congestion whether within individual SANs or across the intersystem network. Although FC (and IBM Spectrum Virtualize) can handle an overloaded host or storage array, the mechanisms in FC are ineffective for dealing with congestion in the fabric in most circumstances. The problems that are caused by fabric congestion can range from dramatically slow response time to storage access loss. These issues are common with all high-bandwidth SAN devices and are inherent to FC. They are not unique to the IBM Spectrum Virtualize based 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 FS9100 nodes that cannot communicate each others because you have a single congested link that leads to an edge switch.

Switches and ISL oversubscription

As specified in Chapter 3, “Drives and arrays” on page 41, the suggested maximum host port to ISL ratio is 7:1. With modern 8 Gbps or 16 Gbps SAN switches, this ratio implies an average bandwidth (in one direction) per host port of approximately 230 MBps (16 Gbps).

You must take peak loads (not average loads) into consideration. For example, while a database server might use only 20 MBps during regular production workloads, it might perform a backup at higher data rates.

Congestion to one switch in a large fabric can cause performance issues throughout the entire fabric, including traffic between IBM FS9100 nodes and external virtualized storage subsystems, even if they are not directly attached to the congested switch. The reasons for these issues are inherent to FC flow control mechanisms, which are not designed to handle fabric congestion. Therefore, any estimates for required bandwidth before implementation must have a safety factor that is built into the estimate.

On top of the safety factor for traffic expansion, implement a spare ISL or ISL trunk. The spare ISL or ISL trunk can provide a fail-safe that avoids congestion if an ISL fails because of issues, such as a SAN switch line card or port blade failure.

Exceeding the standard 7:1 oversubscription ration requires you to implement fabric bandwidth threshold alerts. When one of your ISLs exceeds 70%, you must schedule fabric changes to distribute the load further.

You must also consider the bandwidth consequences of a complete fabric outage. Although a complete fabric outage is a fairly rare event, insufficient bandwidth can turn a single-SAN outage into a total access loss event.

Take the bandwidth of the links into account. It is common to have ISLs run faster than host ports, which reduces the number of required ISLs.

Distance extensions options

To implement remote mirroring over a distance by using the Fibre Channel, you have the following choices:

- Optical multiplexors, such as dense wavelength division multiplexing (DWDM) or coarse wavelength division multiplexing (CWDM) devices
- Long-distance Small Form-factor Pluggable (SFP) transceivers and XFPs
- Fibre Channel-to-IP conversion boxes

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

Check the list of supported SAN routers and FC extenders on the online documentation at: IBM FlashSystem 9100 family.

Optical multiplexors

Optical multiplexors can extend a SAN up to hundreds of kilometers (or miles) at high speeds. For this reason, they are the preferred method for long-distance expansion. If you use multiplexor-based distance extension, closely monitor your physical link error counts in your switches. Optical communication devices are high-precision units. When they shift out of calibration, you will start to see errors in your frames.

Long-distance SFPs and XFPs

Long-distance optical transceivers have the advantage of extreme simplicity. You do not need any expensive equipment, and you have only a few configuration steps to perform. However, ensure that you only use transceivers that are designed for your particular SAN switch.

Fibre Channel over IP

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

With IP-based distance extension, you must dedicate bandwidth to your FCIP traffic if the link is shared with other IP traffic. Do not assume that because the link between two sites has low traffic or is used only for email, this type of traffic is always the case. FC is far more sensitive to congestion than most IP applications.

Also, when you are communicating with the networking architects for your organization, make sure to distinguish between *megabytes per second* as opposed to *megabits per second.* In the storage world, bandwidth often is specified in megabytes per second (MBps), and network engineers specify bandwidth in megabits per second (Mbps).

Hops

The hop count is not increased by the intersite connection architecture. For example, if you have a SAN extension that is based on DWDM, the DWDM components are not apparent to the number of hops. The hop count limit within a fabric is set by the fabric devices (switch or director) operating system. It is used to derive a frame hold time value for each fabric device.

This hold time value is the maximum amount of time that a frame can be held in a switch before it is dropped or `the fabric is busy` condition is returned. For example, a frame might be held if its destination port is unavailable. The hold time is derived from a formula that uses the error detect timeout value and the resource allocation timeout value. It is considered that every extra hop adds about 1.2 microseconds of latency to the transmission.

Currently, IBM Spectrum Virtualize remote copy services support three hops when protocol conversion exists. Therefore, if you have DWDM extended between primary and secondary sites, three SAN directors or switches can exist between the primary and secondary systems.

Buffer credits

SAN device ports need memory to temporarily store frames as they arrive, assemble them in sequence, and deliver them to the upper layer protocol. The number of frames that a port can hold is called its *buffer credit*. Fibre Channel architecture is based on a flow control that ensures a constant stream of data to fill the available pipe.

When two FC ports begin a conversation, they exchange information about their buffer capacities. An FC port sends only the number of buffer frames for which the receiving port gives credit. This method avoids overruns and provides a way to maintain performance over distance by filling the pipe with in-flight frames or buffers.

The following types of transmission credits are available:

- Buffer_to_Buffer Credit

  During login, N_Ports and F_Ports at both ends of a link establish its Buffer to Buffer Credit (BB_Credit).

- End_to_End Credit

  In the same way during login, all N_Ports establish End-to-End Credit (EE_Credit) with each other. During data transmission, a port must not send more frames than the buffer of the receiving port can handle before you receive an indication from the receiving port that it processed a previously sent frame. Two counters are used: BB_Credit_CNT and EE_Credit_CNT. Both counters are initialized to zero during login.

> **FC Flow Control:** Each time that a port sends a frame, it increments BB_Credit_CNT and EE_Credit_CNT by one. When it receives R_RDY from the adjacent port, it decrements BB_Credit_CNT by one. When it receives ACK from the destination port, it decrements EE_Credit_CNT by one.
>
> At any time, if BB_Credit_CNT becomes equal to the BB_Credit, or EE_Credit_CNT becomes equal to the EE_Credit of the receiving port, the transmitting port stops sending frames until the respective count is decremented.

The previous statements are true for Class 2 service. Class 1 is a dedicated connection. Therefore, BB_Credit is not important, and only EE_Credit is used (EE Flow Control). However, Class 3 is an unacknowledged service. Therefore, it uses only BB_Credit (BB Flow Control), but the mechanism is the same in all cases.

Here, you see the importance that the number of buffers has in overall performance. You need enough buffers to ensure that the transmitting port can continue to send frames without stopping to use the full bandwidth, which is true with distance. The total amount of buffer credit needed to optimize the throughput depends on the link speed and the average frame size.

For example, consider an 8 Gbps link connecting two switches that are 100 km apart. At 8 Gbps, a full frame (2148 bytes) occupies about 0.51 km of fiber. In a 100 km link, you can send 198 frames before the first one reaches its destination. You need an ACK to go back to the start to fill EE_Credit again. You can send another 198 frames before you receive the first ACK.

You need at least 396 buffers to allow for nonstop transmission at 100 km distance. The maximum distance that can be achieved at full performance depends on the capabilities of the FC node that is attached at either end of the link extenders, which are vendor-specific. A match should occur between the buffer credit capability of the nodes at either end of the extenders.

Remote system ports and zoning considerations

Ports and zoning requirements for the remote system partnership have changed over time. The current preferred configuration is based on the following:

Flash Alert

The preferred practice for the IBM Spectrum Virtualize based 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 FS9100 systems, where traffic between node canisters in an I/O group is serviced by the dedicated PCI inter-canister link in the enclosure. The following guidelines apply to the remote system connectivity:

- The minimum requirement to establish a remote copy partnership is to connect at least one node per system. When remote connectivity among all the nodes of both systems is not available, the nodes of the local system not participating to the remote partnership will use the node/nodes defined in the partnership as a bridge to transfer the replication data to the remote system.

  This replication data transfer occurs through the node-to-node connectivity. Note that this configuration, even though supported, allows the replication traffic to go through the node-to-node connectivity and this is not recommended.
- Partnered systems should use the same number of nodes in each system for replication.
- For maximum throughput, all nodes in each system should be used for replication, both in terms of balancing the preferred node assignment for volumes and for providing intersystem Fibre Channel connectivity.
- Where possible, use the minimum number of partnerships between systems. For example, assume site A contains systems A1 and A2, and site B contains systems B1 and B2. In this scenario, creating separate partnerships between pairs of systems (such as A1-B1 and A2-B2) offers greater performance for Global Mirror replication between sites than a configuration with partnerships defined between all four systems.

For zoning, the following rules for the remote system partnership apply:

- For Metro Mirror and Global Mirror configurations where the round-trip latency between systems is less than 80 milliseconds, zone two Fibre Channel ports on each node in the local system to two Fibre Channel ports on each node in the remote system.
- For Metro Mirror and Global Mirror configurations where the round-trip latency between systems is more than 80 milliseconds, apply SAN zoning to provide separate intrasystem zones for each local-remote I/O group pair that is used for replication, as shown in Figure 6-20.

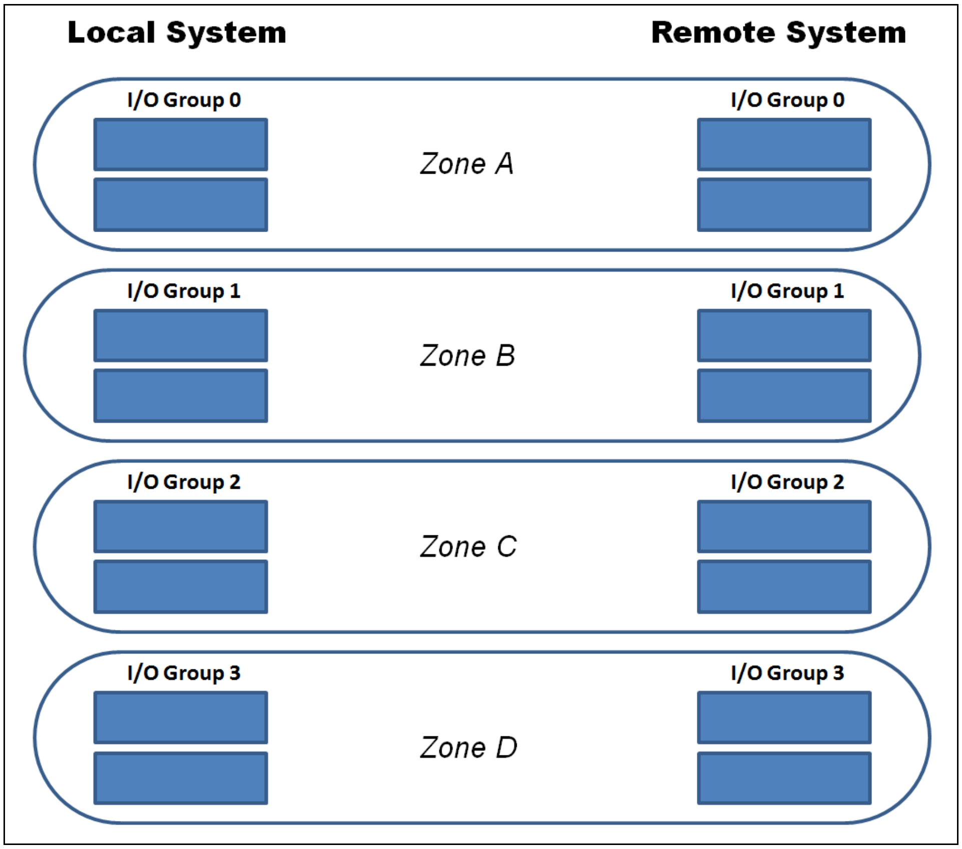


Figure 6-20 Zoning scheme for >80 ms remote copy partnerships

NPIV: IBM Spectrum Virtualize based systems with the NPIV feature enabled provide virtual WWPN for the host zoning. Those WWPNs are intended for host zoning only and can not be used for the remote copy partnership.

6.3.3 Remote Copy services planning

When you plan for remote copy services, you must keep in mind the considerations that are outlined in the following sections.

Remote Copy configurations limits

To plan for and implement remote copy services, you must check the configuration limits and adhere to them. Table 6-8 on page 189 shows the limits for a system that currently apply to IBM Spectrum Virtualize V8.2. 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	1250	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	1024 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
Inter-system IP partnerships per system	1	A system can be partnered with up to three remote systems. A maximum of one of those can be IP and the other two FC.
I/O groups per system in IP partnerships	2	The nodes from a maximum of two I/O groups per system can be used for IP partnership.
Inter site links per IP partnership	2	A maximum of two inter site links can be used between two IP partnership sites.
Ports per node	1	A maximum of one port per node can be used for IP partnership.
IP partnership Software Compression Limit	140 MBps	

Similar to FlashCopy, the remote copy services require memory to allocate the bitmap structures used to track the updates while volume are suspended or synchronizing. The default amount of memory for remote copy services is 20 MB. This value can be increased or decreased by using the `chiogrp` command. The maximum amount of memory that can be specified for remote copy services is 512 MB. The grain size for the remote copy services is 256 KB.

Remote Copy general restrictions

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

- You must have the same size for source and target volume when defining a remote copy relationship. However, the target volume can be a different type (image, striped, or sequential mode) or have different cache settings (cache-enabled or cache-disabled).
- You cannot move Metro Mirror or Global Mirror source or target volumes to different I/O groups.
- Metro Mirror and Global Mirror volumes can be resized with the following restrictions:
  - The volumes must be thin-provisioned or compressed.
  - Apply to Metro Mirror and Global Mirror only, GM/CV is not supported.

   - The Remote Copy Consistency Protection feature is not allowed and must be removed before resizing the volumes.
   - No active FlashCopy allowed.
   - The remote copy relationship must be in synchronized status.
   - The resize order must guarantee the target volume to be always larger than the source volume.
- You can mirror intrasystem Metro Mirror or Global Mirror only between volumes in the same I/O group.

> **Intrasystem remote copy:** The intrasystem 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 quite an easy task. It is enough to stop the relationship, if it is active, and change the properties to set the new remote copy type. Do not forget to create the change volumes in case of change from Metro or Global Mirror to Global Mirror Change Volumes.

Global Mirror and GM/CV coexistence considerations

Global Mirror and Global Mirror Change Volumes relationships can be defined in the same system. With these configurations, particular attention must be paid on the bandwidth sizing and the partnership settings. The two Global Mirror technologies use the available bandwidth in different ways. As described in “Network sizing considerations” on page 181, while the regular Global Mirror uses the amount of bandwidth needed to sustain the write workload of the replication set, the GM/CV uses the fixed amount of bandwidth as defined in the partnership as background copy.

For this reason, during the GM/CV 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 the Global Mirror. To avoid contention on the bandwidth, which could lead to 1920 error (see 6.3.5, “1920 error” on page 203) or delayed GM/CV cycle creation, the bandwidth must be sized taking in account both requirements.

Ideally in these cases the bandwidth should be enough to accommodate the peek write workload for the Global Mirror replication set plus the estimated bandwidth needed to fulfill the RPO of the GM/CV. If these requirement can not be met due to bandwidth restrictions, the less impacting option is to increase the GM/CV cycle period and then reduce the background copy rate minimizing in such way the chance of 1920 error occurrence.

Note that these considerations apply also to configurations where multiple IBM FS9100 or IBM Spectrum Virtualize based clusters are sharing the same bandwidth resources.

Interaction between Remote Copy and FlashCopy

Remote Copy functions can be used in conjunction with the Flash Copy function so that you can have both operating concurrently on the same volume. The possible combinations between Remote Copy and FlashCopy follow:

- Remote copy source:
  - A remote copy source can be a FlashCopy source.
  - A remote copy source can be a FlashCopy target with the following restrictions:
    - A FlashCopy target volume cannot be updated while it is the source volume of a Metro or Global Mirror relationship that is actively mirroring. A FlashCopy mapping cannot be started while the target volume is in an active remote copy relationship.
    - The I/O group for the FlashCopy mappings must be the same as the I/O group for the FlashCopy target volume (that is the I/O group of the Remote copy source).
- Remote copy target:
  - A remote copy target can be a FlashCopy source.
  - A remote copy target can be a FlashCopy target with the following restrictions:
    - A FlashCopy mapping must be in the `idle_copied` state when its target volume is the target volume of an active Metro Mirror or Global Mirror relationship.

When implementing Flashcopy functions for volumes in GM/CV relationships remember that Flashcopy multi-target mappings will be created. As described in "Interaction and dependency between Multiple Target FlashCopy mappings" on page 154, this results in dependent mappings that can affect the cycle formation due to the cleaning process (see "**Cleaning process and Cleaning Rate**" on page 163).

With such configurations, it is recommended to set the Cleaning Rate accordingly. This recommendation apply also to Consistency Protection volumes and HyperSwap configurations.

Native back-end controller copy functions considerations

As discussed in the above sections, the IBM FS9100 technology provides a widespread set of copy services functions that cover most of the clients requirements.

However, some storage controllers can provide specific copy services capabilities not available with the current version of IBM Spectrum Virtualize software. The IBM Spectrum Virtualize 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 Spectrum Virtualize caches them and processes them later. This process can have more ramifications if a target host depends on the write I/Os from the source host as they are written.

> **Note:** Native copy services are not supported on all storage controllers. For more information about the known limitations, see *Using Native Controller Copy Services*, S1002852, at this website:
>
> `Using Native Controller Copy Services`

As part of its copy services function, the storage controller might take a LUN offline or suspend reads or writes. As IBM FS9100 does not recognize why this happens; therefore, it might log errors when these events occur. For this reason, if the IBM FS9100 must detect the LUN, ensure to keep that LUN in the `unmanaged` state until full access is granted.

Native back-end controller copy services can also be used for LUNs not managed by the IBM FS9100. Note that accidental incorrect configurations of the back-end controller copy services involving IBM FS9100 attached LUN can produce unpredictable results.

For example, if you accidentally use a LUN with IBM FS9100 data on it as a point-in-time target LUN, you can corrupt that data. Moreover, if that LUN was a managed disk in a managed disk group with striped or sequential volumes on it, the managed disk group might be brought offline. This situation, in turn, makes all of the volumes that belong to that group go offline, leading to a widespread host access disruption.

Remote Copy and code upgrade considerations

When you upgrade system software where the system participates in one or more intersystem relationships, upgrade only one cluster at a time. That is, do not upgrade the systems concurrently.

Attention: Upgrading both systems concurrently is not monitored by the software upgrade process.

Allow the software upgrade to complete one system before it is started on the other system. Upgrading both systems concurrently can lead to a loss of synchronization. In stress situations, it can further lead to a loss of availability.

Usually, pre-existing remote copy relationships are unaffected by a software upgrade that is performed correctly. However, always check in the target code release notes for special considerations on the copy services.

Even if it is not a best practice, a remote copy partnership can be established, with some restriction, among systems with different IBM Spectrum Virtualize versions. For more information about a compatibility table for intersystem Metro Mirror and Global Mirror relationships between IBM Spectrum Virtualize code levels, see *Spectrum Virtualize Family of Products Inter-System Metro Mirror and Global Mirror Compatibility Cross Reference:*

Spectrum Virtualize Family of Products Inter-System Metro Mirror and Global Mirror Compatibility Cross Reference

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 FS9100 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 6-21 shows an example of remote copy resources that are not optimized. Volumes from the local system are replicated to the remote system. All volumes with a preferred node of node 1 are replicated to the remote system, where the target volumes also have a preferred node of node 1.

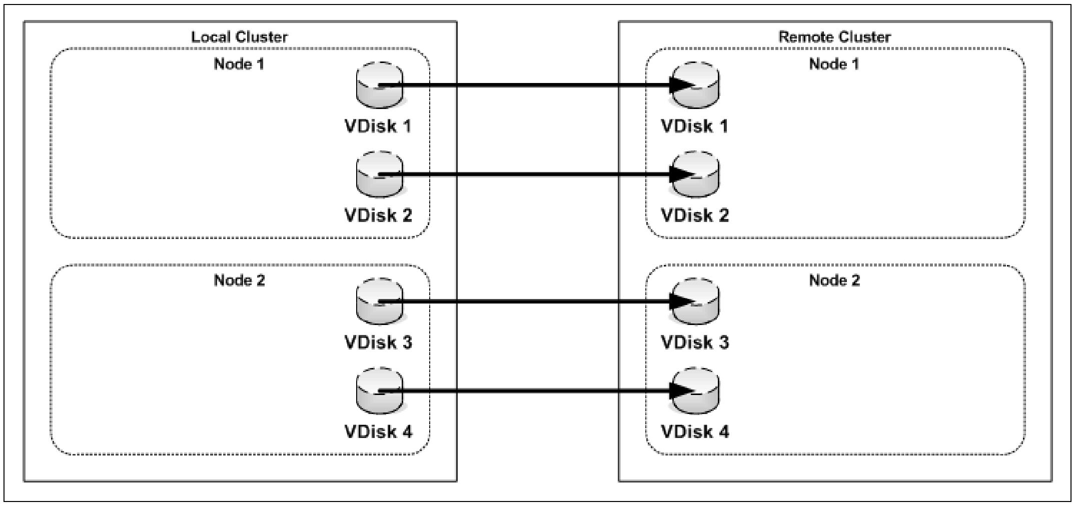


Figure 6-21 Remote copy resources that are not optimized

With this configuration, the resources for remote system node 1 that are reserved for local system node 2 are not used. The resources for local system node 1 that are used for remote system node 2 also are not used.

If the configuration changes to the configuration that is shown in Figure 6-22, all remote copy resources for each node are used, and remote copy operates with better performance.

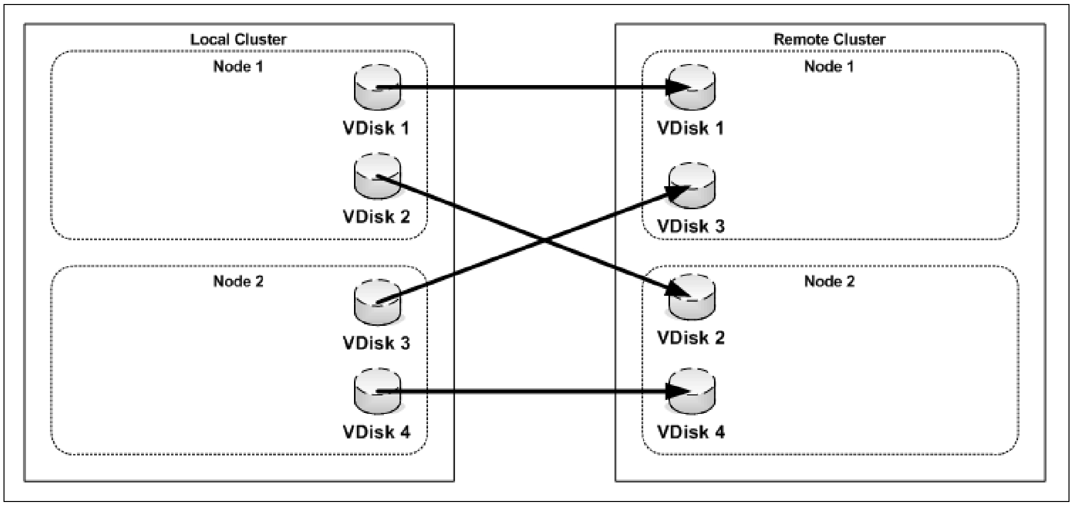


Figure 6-22 Optimized Global Mirror resources

GM/CV Change Volumes placement considerations

The Change Volumes in a GM/CV configuration are basically thin provisioned volumes used as FlashCopy targets. For this reason the same considerations described in "Volume placement considerations" on page 161 apply. 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 very beneficial in terms of space savings.

Background copy considerations

The remote copy partnership bandwidth parameter *explicitly* defines the rate at which the background copy is attempted, but also *implicitly* affects foreground I/O. Background copy bandwidth can affect foreground I/O latency in one of the following ways:

- Increasing latency of foreground I/O

  If the remote copy partnership bandwidth parameter is set too high for the actual intersystem network capability, the background copy resynchronization writes use too much of the intersystem network. It starves the link of the ability to service synchronous or asynchronous mirrored foreground writes. Delays in processing the mirrored foreground writes increase the latency of the foreground I/O as perceived by the applications.

- Read I/O overload of primary storage

  If the remote copy partnership background copy rate is set too high, the added read I/Os that are associated with background copy writes can overload the storage at the primary site and delay foreground (read and write) I/Os.

- Write I/O overload of auxiliary storage

  If the remote copy partnership background copy rate is set too high for the storage at the secondary site, the background copy writes overload the auxiliary storage. Again, they delay the synchronous and asynchronous mirrored foreground write I/Os.

Important: An increase in the peak foreground workload can have a detrimental effect on foreground I/O. It does so by pushing more mirrored foreground write traffic along the intersystem network, which might not have the bandwidth to sustain it. It can also overload the primary storage.

To set the background copy bandwidth optimally, consider all aspects of your environments, starting with the following biggest contributing resources:

- Primary storage
- Intersystem network bandwidth
- Auxiliary storage

Provision the most restrictive of these three resources between the background copy bandwidth and the peak foreground I/O workload. Perform this provisioning by calculation or by determining experimentally how much background copy can be allowed before the foreground I/O latency becomes unacceptable.

Then, reduce the background copy to accommodate peaks in workload. In cases where the available network bandwidth is not able to sustain an acceptable background copy rate, consider alternatives to the initial copy as described in “Initial synchronization options and Offline Synchronization” on page 196.

Changes in the environment, or loading of it, can affect the foreground I/O. IBM Spectrum Virtualize technology provides a means to monitor, and a parameter to control, how foreground I/O is affected by running remote copy processes. IBM Spectrum Virtualize software monitors the delivery of the mirrored foreground writes. If latency or performance of these writes extends beyond a (predefined or client-defined) limit for a period, the remote copy relationship is suspended (see 6.3.5, “1920 error” on page 203).

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

Background copy bandwidth allocation: As already mentioned in “Volume placement considerations” on page 192, the available bandwidth of a remote copy partnership is evenly divided among the cluster’s I/O Groups. In a case of unbalanced distribution of the remote copies among the I/O groups, the partnership bandwidth should be adjusted accordingly to reach the desired background copy rate.

Consider, for example, a 4-I/O groups cluster that has a partnership bandwidth of 4,000 Mbps and a background copy percentage of 50. The expected maximum background copy rate for this partnership is then 250MB/s. Having the available bandwidth evenly divided among the I/O groups, every I/O group in this cluster can theoretically synchronize data at a maximum rate of about 62 MBps (50% of 1,000 Mbps). Now in an edge case where only volumes from one I/O group are being replicated, in order to reach the full background copy rate (250 MBps) the partnership bandwidth should be adjusted to 16000 Mbps.

Initial synchronization options and Offline Synchronization

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

- The `not synchronized` option is the default. With this option, when a remote copy relationship is started, a full data synchronization at the background copy rate occurs between the source and target volumes. It is the simplest approach in that it requires no other administrative activity apart from issuing the necessary IBM Spectrum Virtualize commands. However, in some environments, the available bandwidth makes this option unsuitable.
- The `already synchronized` option does not force any data synchronization when the relationship is started. The administrator must ensure that the source and target volumes contain identical data before a relationship is created. The administrator can perform this check in one of the following ways:
  - Create both volumes with the security delete feature to change all data to zero.
  - Copy a complete tape image (or other method of moving data) from one disk to the other.

  In either technique, no write I/O must take place to the source and target volume before the relationship is established. The administrator must then complete the following actions:

  - Create the relationship with the already synchronized settings (**`-sync`** option).
  - Start the relationship.

  > **Attention:** If you do not perform these steps correctly, the remote copy reports the relationship as being *consistent*, when it is not. This setting is likely to make any auxiliary volume useless.

By understanding the methods to start a Metro Mirror and Global Mirror relationship, you can use one of them as a means to implement the remote copy relationship, save bandwidth, and resize the Global Mirror volumes.

Consider a situation where you have a large source volume (or many source volumes) containing already active data and that you want to replicate to a remote site. Your planning shows that the mirror initial sync time takes too long (or is too costly if you pay for the traffic that you use). In this case, you can set up the sync by using another medium that is less expensive. This synchronization method is called *Offline Synchronization*.

Another reason that you might want to use this method is if you want to increase the size of the volume that is in a Metro Mirror relationship or in a Global Mirror relationship. Increasing the size of these volumes might require a deletion and redefinition of the current mirror relationships when the requirements described in “Remote Copy general restrictions” on page 189 are not met.

This example uses tape media as the source for the initial sync for the Metro Mirror relationship or the Global Mirror relationship target before it uses remote copy services to maintain the Metro Mirror or Global Mirror. This example does not require downtime for the hosts that use the source volumes.

Before you set up Global Mirror relationships, save bandwidth, and resize volumes, complete the following steps:

1. Ensure that the hosts are up and running and are using their volumes normally. No Metro Mirror relationship nor Global Mirror relationship is defined yet.

2. Identify all of the volumes that become the source volumes in a Metro Mirror relationship or in a Global Mirror relationship.
3. Establish the IBM Spectrum Virtualize based system partnership with the target IBM Spectrum Virtualize based system.

To set up Global Mirror relationships, save bandwidth, and resize volumes, complete the following steps:

1. Define a Metro Mirror relationship or a Global Mirror relationship for each source disk. When you define the relationship, ensure that you use the `-sync` option, which stops the system from performing an initial sync.

> **Attention:** If you do not use the `-sync` option, all of these steps are redundant because the IBM FS9100 system performs a full initial synchronization anyway.

2. Stop each mirror relationship by using the `-access` option, which enables write access to the target volumes. You need this write access later.
3. Copy the source volume to the alternative media by using the `dd` command to copy the contents of the volume to tape. Another option is to use your backup tool (for example, IBM Spectrum Protect) to make an image backup of the volume.

> **Change tracking:** Although the source is being modified while you are copying the image, the IBM Spectrum Virtualize based system is tracking those changes. The image that you create might have some of the changes and is likely to also miss some of the changes. When the relationship is restarted, the IBM Spectrum Virtualize based system applies all of the changes that occurred since the relationship stopped in step 2. After all the changes are applied, you have a consistent target image.

4. Ship your media to the remote site and apply the contents to the targets of the Metro Mirror or Global Mirror relationship. You can mount the Metro Mirror and Global Mirror target volumes to a UNIX server and use the `dd` command to copy the contents of the tape to the target volume.

   If you used your backup tool to make an image of the volume, follow the instructions for your tool to restore the image to the target volume. Remember to remove the mount if the host is temporary.

> **Tip:** It does not matter how long it takes to get your media to the remote site and perform this step. However, the faster you can get the media to the remote site and load it, the quicker IBM FS9100 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 FS9100 system stops write access to the volume while the mirror relationship is running.
6. Start your Metro Mirror and Global Mirror relationships. The relationships must be started with the `-clean` parameter. In this way, any changes that are made on the secondary volume are ignored, and only changes made on the clean primary volume are considered when synchronizing the primary and secondary volumes.
7. While the mirror relationship catches up, the target volume is not usable at all. When it reaches `ConsistentSynchnonized` status, your remote volume is ready for use in a disaster.

Back-end storage considerations

To reduce the overall solution costs, it is a common practice to provide the remote systems with lower performance characteristics compared to the local system, especially when using asynchronous remote copy technologies. This attitude can be risky especially when using the Global Mirror technology where the application performances at the primary system can indeed be limited by the performance of the remote system.

The preferred practice is to perform an accurate back-end resource sizing for the remote system to fulfill the following capabilities:

- The peak application workload to the Global Mirror or Metro Mirror volumes
- The defined level of background copy
- Any other I/O that is performed at the remote site

Remote Copy tunable parameters

Several commands and parameters help to control remote copy and its default settings. You can display the properties and features of the systems by using the **`lssystem`** command. Also, you can change the features of systems by using the **`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 like these before putting them into a live production environment.

A detailed description and settings considerations about the **gmlinktolerance** and the **gmmaxhostdelay** parameters are described in 6.3.5, “1920 error” on page 203.

rcbuffersize

rcbuffersize was introduced with the version 6.2 code level so that systems with intense and bursty write I/O would not fill the internal buffer while Global Mirror writes were undergoing sequence tagging.

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

Example 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 additional buffers you allocate are taken away from the general cache.

maxreplicationdelay and partnershipexclusionthreshold

IBM Spectrum Virtualize uses two parameters, **maxreplicationdelay** and **partnershipexclusionthreshold**, for remote copy advanced tuning.

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

The **partnershipexclusionthreshold** parameter was introduced to allow users to set the timeout for an I/O that triggers a temporarily dropping of the link to the remote cluster. The value must be a number from 30 - 315.

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

A detailed description and settings considerations about the **maxreplicationdelay** parameter are described in 6.3.5, “1920 error” on page 203.

Link delay simulation parameters

Even though Global Mirror is an asynchronous replication method, there can be an impact to server applications due to Global Mirror managing transactions and maintaining write order consistency over a network. To mitigate this impact, as a testing and planning feature, Global Mirror allows you to simulate the effect of the round-trip delay between sites by using the following parameters:

- The **`gminterclusterdelaysimulation`** parameter

  This optional parameter specifies the intersystem delay simulation, which simulates the Global Mirror round-trip delay between two systems in milliseconds. The default is 0. The valid range is 0 - 100 milliseconds.

- The **`gmintraclusterdelaysimulation`** parameter

  This optional parameter specifies the intrasystem delay simulation, which simulates the Global Mirror round-trip delay in milliseconds. The default is 0. The valid range is 0 - 100 milliseconds.

6.3.4 Remote Copy special use cases

The most common use cases for the remote copy functions are obviously Disaster Recovery solutions. A complete discussion about the Disaster Recovery solutions based on IBM Spectrum Virtualize technology is beyond the intended scope for this book. For an overview of the Disaster Recovery solutions with the IBM Spectrum Virtualize copy services see *IBM System Storage SAN Volume Controller and Storwize V7000 Replication Family Services*, SG24-7574.

Another typical remote copy use case is the data movement among distant locations as required, for instance, for data center relocation and consolidation projects. In these scenarios, the IBM FS9100 remote copy technology, in conjuction with the virualization features, is particularly effective when combined with the image copy feature that allows data movement among storage systems of different technology or vendor.

DRP limitation: Currently the image mode VDisk is not supported with DRP.

Performing cascading copy service functions

Cascading copy service functions that use IBM FS9100 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 FS9100 copy services and by combining IBM FS9100 copy services (with image mode cache-disabled volumes) and native storage controller copy services.

Cascading with native storage controller copy services

Figure 6-23 describes the configuration for three-site cascading by using the native storage controller copy services in combination with IBM FS9100 remote copy functions.

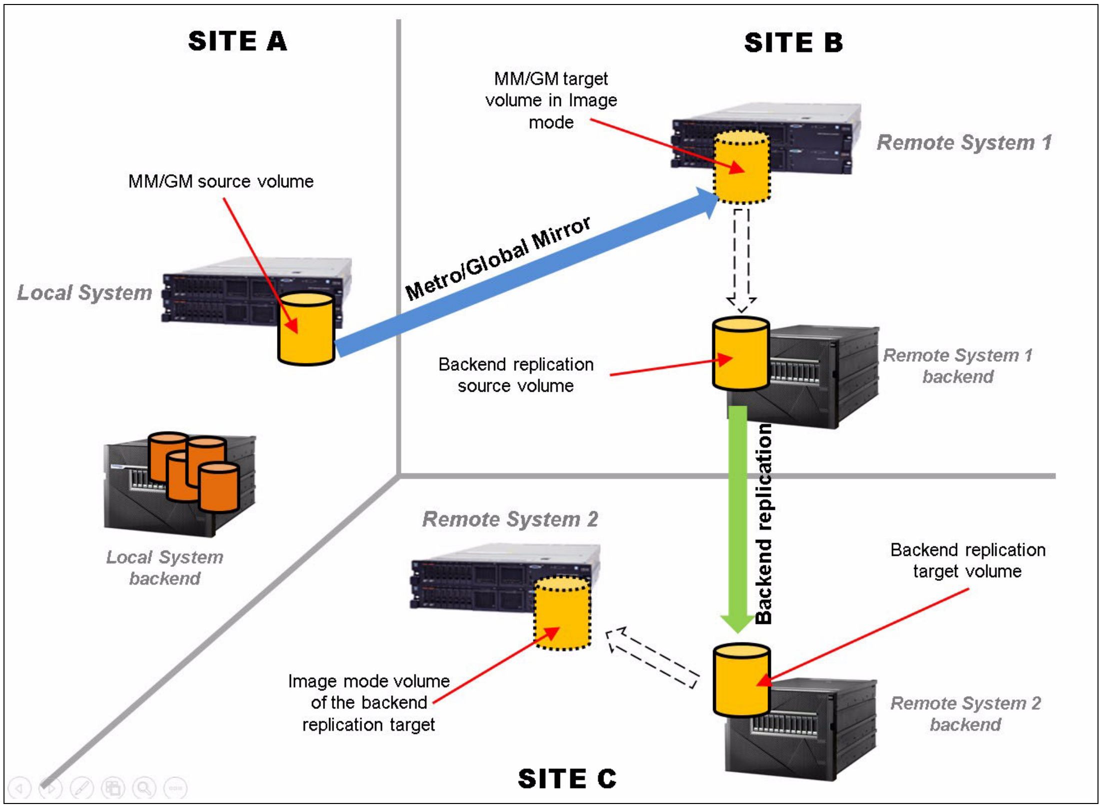


Figure 6-23 Using three-way copy services

In Figure 6-23, the primary site uses IBM Spectrum Virtualize 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 Spectrum Virtualize (see “Native back-end controller copy functions considerations” on page 191).
- The source and target volumes used by the back-end controller native copy services must be imported to the IBM FS9100 system as image-mode volumes with the cache disabled.

Cascading with IBM FS9100 systems copy services

A cascading-like solution is also possible by combining the IBM FS9100 copy services. These remote copy services implementations are useful in three site disaster recovery solutions and data center moving scenarios.

In the configuration described in Figure 6-24, a Global Mirror (Metro Mirror can also be used) solution is implemented between the Local System in Site A, the production site, and the Remote System 1 located in Site B, the primary disaster recover site. A third system, Remote System 2, is located in Site C, the secondary disaster recover site. Connectivity is provided between Site A and Site B, between Site B and Site C, and optionally between Site A and Site C.

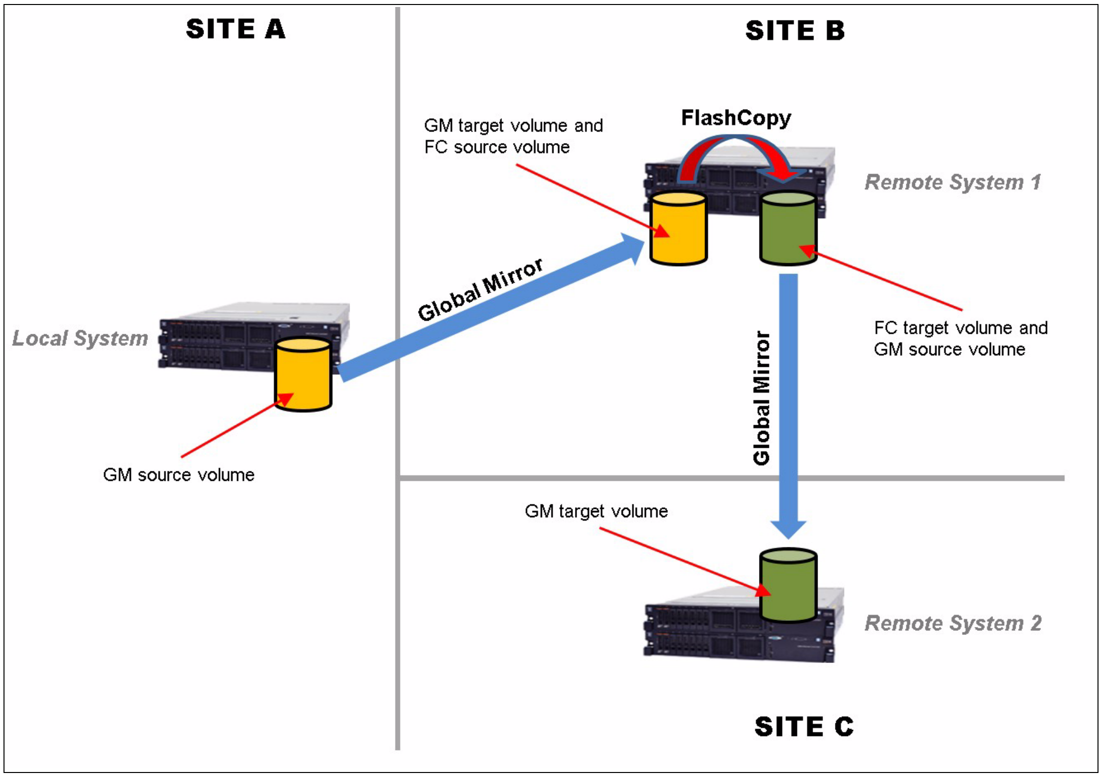


Figure 6-24 Cascading-like infrastructure

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:

   c. Create the Global Mirror relationships between the Local System and Remote System 1.

   d. Create the FlashCopy mappings in the Remote System 1 using the target Global Mirror volumes as FlashCopy source volumes. The FlashCopy must be incremental.

   e. Create the Global Mirror relationships between Remote System 1 and Remote System 2 using the FlashCopy target volumes as Global Mirror source volumes.

   f. 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 `ConsistenSynchronized` 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.

6.3.5 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 intersystem connectivity, unexpectedly higher foreground host I/O workload, or a permanent error because of a hardware failure. It is also possible that not all relationships are affected and that multiple 1920 errors can be posted.

The 1920 error could be triggered both for Metro Mirror and Global Mirror relationships. However, in Metro Mirror configurations the 1920 error is associated only with a permanent I/O error condition. For this reason, the main focus of this section is 1920 errors in a Global Mirror configuration.

Internal Global Mirror control policy and raising 1920 errors

Although Global Mirror is an asynchronous remote copy service, the local and remote sites have some interplay. When data comes into a local volume, work must be done to ensure that the remote copies are consistent. This work can add a delay to the local write. Normally, this delay is low. The IBM Spectrum Virtualize code implements many control mechanisms that mitigate the impacts of the Global Mirror to the foreground I/Os.

gmmaxhostdelay and gmlinktolerance

The **`gmlinktolerance`** parameter helps to ensure that hosts do not perceive the latency of the long-distance link, regardless of the bandwidth of the hardware that maintains the link or the storage at the secondary site. The hardware and storage must be provisioned so that, when combined, they can support the maximum throughput that is delivered by the applications at the primary that is using Global Mirror.

If the capabilities of this hardware are exceeded, the system becomes backlogged and the hosts receive higher latencies on their write I/O. Remote copy in Global Mirror implements a protection mechanism to detect this condition and halts mirrored foreground write and background copy I/O. Suspension of this type of I/O traffic ensures that misconfiguration or hardware problems (or both) do not affect host application availability.

Global Mirror attempts to detect and differentiate between backlogs that occur because of the operation of the Global Mirror protocol. It does not examine the general delays in the system when it is heavily loaded, where a host might see high latency even if Global Mirror were disabled.

To detect these specific scenarios, Global Mirror measures the time that is taken to perform the messaging to assign and record the sequence number for a write I/O. If this process exceeds the expected value over a period of 10 seconds, this period is treated as being overloaded (*bad period*).

Global Mirror uses the **gmmaxhostdelay** and **gmlinktolerance** parameters to monitor Global Mirror protocol backlogs in the following ways:

- Users set the **gmmaxhostdelay** and **gmlinktolerance** parameters to control how software responds to these delays. The **gmmaxhostdelay** parameter is a value in milliseconds that can go up to 100.
- Every 10 seconds, Global Mirror samples all of the Global Mirror writes and determines how much of a delay it added. If at least a third of these writes are greater than the **gmmaxhostdelay** setting, that sample period is marked as *bad*.
- Software keeps a running count of *bad periods*. Each time that a bad period occurs, this count goes up by one. Each time a good period occurs, this count goes down by 1, to a minimum value of 0.

The **gmlinktolerance** parameter is defined in seconds. Bad periods are assessed at intervals of 10 seconds. The maximum bad period count is the **gmlinktolerance** parameter value that is divided by 10. 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.

Bad periods do not need to be consecutive, and the bad period count increments or decrements at intervals of 10. That is, 10 bad periods, followed by five good periods, followed by 10 bad periods, results in a bad period count of 15.

Within each sample period, Global Mirror writes are assessed. If in a write operation, the delay added by the Global Mirror protocol exceeds the **gmmaxhostdelay** value, the operation is counted as a bad write. Otherwise, a good write is counted. The proportion of bad writes to good writes is calculated. If at least one third of writes are identified as bad, the sample period is defined as a bad period.

A consequence is that, under a light I/O load, a single bad write can become significant. For example, if only one write I/O is performed for every 10 and this write is considered slow, the bad period count increments.

An edge case is achieved by setting the **gmmaxhostdelay** and **gmlinktolerance** parameters to their minimum settings (1 ms and 20 s). With these settings, you need only two consecutive bad sample periods before a 1920 error condition is reported. Consider a foreground write I/O that has a light I/O load.

For example, a single I/O happens in the 20 s. With unlucky timing, a single bad I/O results (that is, a write I/O that took over 1 ms in remote copy), and it spans the boundary of two, 10-second sample periods. This single bad I/O theoretically can be counted as 2 x the bad periods and trigger a 1920 error.

A higher **gmlinktolerance** value, **gmmaxhostdelay** setting, or I/O load might reduce the risk of encountering this edge case.

maxreplicationdelay and partnershipexclusionthreshold

IBM FS9100 uses the **maxreplicationdelay** and **partnershipexclusionthreshold** parameters to provide further performance protection mechanisms when remote copy services (Metro Mirror and Global Mirror) are used.

maxreplicationdelay is a system-wide attribute that configures how long a single write can be outstanding from the host before the relationship is stopped, triggering a 1920 error. It can protect the hosts from seeing timeouts due to secondary hung I/Os.

This parameter is mainly intended to protect from secondary system issues. It does not help with ongoing performance issues, but can be used to limit the exposure of hosts to long write response times that can cause application errors. For instance, setting **maxreplicationdelay** to 30 means that if a write operation for a volume in a remote copy relationship does not complete within 30 seconds, the relationship is stopped, triggering a 1920 error. 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.

In an IBM FS9100 system, a node assert (restart with a 2030 error) occurs if any individual I/O takes longer than 6 minutes. To avoid this situation, some actions are attempted to clean up anything that might be hanging I/O before the I/O gets to 6 minutes.

One of these actions is temporarily dropping (for 15 minutes) the link between systems if any I/O takes longer than 5 minutes 15 seconds (315 seconds). This action often removes hang conditions caused by replication problems. The **partnershipexclusionthreshold** parameter introduced the ability to set this value to a time lower than 315 seconds to respond to hung I/O more swiftly. The **partnershipexclusionthreshold** value must be a number in the range 30 - 315.

If an I/O takes longer the **partnershipexclusionthreshold** value, a 1720 error is triggered (with an event ID 987301) and any regular Global Mirror or Metro Mirror relationships stop on the next write to the primary volume.

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

To set the **maxreplicationdelay** and **partnershipexclusionthreshold** parameters, the **chsystem** command must be used, as shown in Example 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 Spectrum Virtualize. Although not as powerful as IBM Spectrum Control, **IOstats** can provide valuable debug information if the **snap** command gathers system configuration data close to the time of failure.

The following are the main performance statistics to investigate for the 1920 error:

- *Write I/O Rate and Write Data Rate*

  For volumes that are primary volumes in relationships, these statistics are the total amount of write operations submitted per second by hosts on average over the sample period, and the bandwidth of those writes. For secondary volumes in relationships, this is the average number of replicated writes that are received per second, and the bandwidth that these writes consume. Summing the rate over the volumes you intend to replicate gives a coarse estimate of the replication link bandwidth required.

- *Write Response Time and Peak Write Response Time*

  On primary volumes, these are the average time (in milliseconds) and peak time between a write request being received from a host, and the completion message being returned. The write response time is the best way to show what kind of write performance that the host is seeing.

  If a user complains that an application is slow, and the stats show the write response time leap from 1 ms to 20 ms, the two are most likely linked. However, some applications with high queue depths and low to moderate workloads will not be affected by increased response times. Note that this being high is an effect of some other problem. The peak is less useful, as it is very sensitive to individual glitches in performance, but it can show more detail of the distribution of write response times.

  On secondary volumes, these statistics describe the time for the write to be submitted from the replication feature into the system cache, and should normally be of a similar magnitude to those on the primary volume. Generally, the write response time should be below 1 ms for a fast-performing system.

- *Global Mirror Write I/O Rate*

  This statistic shows the number of writes per second, the (regular) replication feature is processing for this volume. It applies to both types of Global Mirror and to Metro Mirror, but in each case only for the secondary volume. Because writes are always separated into 32 KB or smaller tracks before replication, this setting might be different from the Write I/O Rate on the primary volume (magnified further because the samples on the two systems will not be aligned, so they will capture a different set of writes).

- *Global Mirror Overlapping Write I/O Rate*

  This statistic monitors the amount of overlapping I/O that the Global Mirror feature is handling for regular Global Mirror relationships. That is where an LBA is written again after the primary volume has been updated, but before the secondary volume has been updated for an earlier write to that LBA. To mitigate the effects of the overlapping I/Os, a journaling feature has been implemented, as discussed in “Colliding writes” on page 175.

- *Global Mirror secondary write lag*

  This statistic is valid for regular Global Mirror primary and secondary volumes. For primary volumes, it tracks the length of time in milliseconds that replication writes are outstanding from the primary system. This amount includes the time to send the data to the remote system, consistently apply it to the secondary non-volatile cache, and send an acknowledgment back to the primary system.

For secondary volumes, this statistic records only the time that is taken to consistently apply it to the system cache, which is normally up to 20 ms. Most of that time is spent coordinating consistency across many nodes and volumes. Primary and secondary volumes for a relationship tend to record times that differ by the round-trip time between systems. If this statistic is high on the secondary system, look for congestion on the secondary system's fabrics, saturated auxiliary storage, or high CPU utilization on the secondary system.

- *Write-cache Delay I/O Rate*

  These statistics show how many writes could not be instantly accepted into the system cache because cache was full. It is a good indication that the write rate is faster than the storage can cope with. If this amount starts to increase on auxiliary storage while primary volumes suffer from increased Write Response Time, it is possible that the auxiliary storage is not fast enough for the replicated workload.

- *Port to Local Node Send Response Time*

  The time in milliseconds that it takes this node to send a message to other nodes in the same system (which will mainly be the other node in the same I/O group) and get an acknowledgment back. This amount should be well below 1 ms, with values below 0.3 ms being essential for regular Global Mirror to provide a Write Response Time below 1 ms.

  This requirement is necessary because up to three round-trip messages within the local system will happen before a write completes to the host. If this number is higher than you want, look at fabric congestion (Zero Buffer Credit Percentage) and CPU Utilization of all nodes in the system.

- *Port to Remote Node Send Response Time*

  This value is the time in milliseconds that it takes to send a message to nodes in other systems and get an acknowledgment back. This amount is not separated out by remote system, but for environments that have replication to only one remote system. This amount should be very close to the low-level ping time between your sites. If this starts going significantly higher, it is likely that the link between your systems is saturated, which usually causes high Zero Buffer Credit Percentage as well.

- Sum of *Port-to-local node send response time* and *Port-to-local node send queue time*

  The time must be less than 1 ms for the primary system. A number in excess of 1 ms might indicate that an I/O group is reaching its I/O throughput limit, which can limit performance.

- *System CPU Utilization* (Core 1-8)

  These values show how heavily loaded the nodes in the system are. If any core has high utilization (say, over 90%) and there is an increase in write response time, it is possible that the workload is being CPU limited. You can resolve this by upgrading to faster hardware, or spreading out some of the workload to other nodes and systems.

- *Zero Buffer Credit Percentage*

  This is the fraction of messages that this node attempted to send through Fibre Channel ports that had to be delayed because the port ran out of buffer credits. If you have a long link from the node to the switch it is attached to, there might be benefit in getting the switch to grant more buffer credits on its port.

  It is more likely to be the result of congestion on the fabric, because running out of buffer credits is how Fibre Channel performs flow control. Normally, this value is well under 1%. From 1 - 10% is a concerning level of congestion, but you might find the performance acceptable. Over 10% indicates severe congestion. This amount is also called out on a port-by-port basis in the port-level statistics, which gives finer granularity about where any congestion might be.

When looking at the port-level statistics, high values on ports used for messages to nodes in the same system are much more concerning than those on ports that are used for messages to nodes in other systems.

- *Back-end Write Response Time*

  This value is the average response time in milliseconds for write operations to the back-end storage. This time might include several physical I/O operations, depending on the type of RAID architecture.

  Poor back-end performances on secondary system is a frequent cause of 1920 errors, while it is not so common for primary systems. Exact values to watch out for depend on the storage technology, but usually the response time should be less than 50 ms. A longer response time can indicate that the storage controller is overloaded. If the response time for a specific storage controller is outside of its specified operating range, investigate for the same reason.

Focus areas for 1920 errors

The causes of 1920 errors might be numerous. To fully understand the underlying reasons for posting this error, consider the following components that are related to the remote copy relationship:

- The intersystem connectivity network
- Primary storage and remote storage
- IBM FS9100 node canisters
- Storage area network

Data collection for diagnostic purposes

A successful diagnosis depends on the collection of the following data at both systems:

- The `snap` command with `livedump` (triggered at the point of failure)
- I/O Stats running (if possible)
- IBM Spectrum Control performance statistics data (if possible)
- The following information and logs from other components:
  - Intersystem network and switch details:
    - Technology
    - Bandwidth
    - Typical measured latency on the Intersystem network
    - Distance on all links (which can take multiple paths for redundancy)
    - Whether trunking is enabled
    - How the link interfaces with the two SANs
    - Whether compression is enabled on the link
    - Whether the link dedicated or shared; if so, the resource and amount of those resources they use
    - Switch Write Acceleration to check with IBM for compatibility or known limitations
    - Switch Compression, which should be transparent but complicates the ability to predict bandwidth
  - Storage and application:
    - Specific workloads at the time of 1920 errors, which might not be relevant, depending upon the occurrence of the 1920 errors and the volumes that are involved

    - RAID rebuilds
    - Whether 1920 errors are associated with Workload Peaks or Scheduled Backup

Intersystem network

For diagnostic purposes, ask the following questions about the intersystem network:

- Was network maintenance being performed?

  Consider the hardware or software maintenance that is associated with intersystem network, such as updating firmware or adding more capacity.

- Is the intersystem network overloaded?

  You can find indications of this situation by using statistical analysis with the help of I/O stats, IBM Spectrum Control, or both. Examine the internode communications, storage controller performance, or both. By using IBM Spectrum Control, you can check the storage metrics for the Global Mirror relationships were stopped, which can be tens of minutes depending on the **`gmlinktolerance`** and **`maxreplicationdelay`** parameters.

  Diagnose the overloaded link by using the following methods:

  - Look at the statistics generated by the routers or switches near your most bandwidth-constrained link between the systems

    Exactly what is provided, and how to analyze it varies depending on the equipment used.

  - Look at the port statistics for high response time in the internode communication

    An overloaded long-distance link causes high response times in the internode messages (the *Port to remote node send response time* statistic) that are sent by IBM Spectrum Virtualize. If delays persist, the messaging protocols exhaust their tolerance elasticity and the Global Mirror protocol is forced to delay handling new foreground writes while waiting for resources to free up.

  - Look at the port statistics for buffer credit starvation

    The *Zero Buffer Credit Percentage* statistic can be useful here too, because you normally have a high value here as the link saturates. Only look at ports that are replicating to the remote system.

  - Look at the volume statistics (before the 1920 error is posted):

    - Target volume write throughput approaches the link bandwidth.

      If the write throughput on the target volume is equal to your link bandwidth, your link is likely overloaded. Check what is driving this situation. For example, does peak foreground write activity exceed the bandwidth, or does a combination of this peak I/O and the background copy exceed the link capacity?

    - Source volume write throughput approaches the link bandwidth.

      This write throughput represents only the I/O that is performed by the application hosts. If this number approaches the link bandwidth, you might need to upgrade the link's bandwidth. Alternatively, reduce the foreground write I/O that the application is attempting to perform, or reduce the number of remote copy relationships.

    - Target volume write throughput is greater than the source volume write throughput.

      If this condition exists, the situation suggests a high level of background copy and mirrored foreground write I/O. In these circumstances, decrease the background copy rate parameter of the Global Mirror partnership to bring the combined mirrored foreground I/O and background copy I/O rates back within the remote links bandwidth.

  - Look at the volume statistics (after the 1920 error is posted):
    - Source volume write throughput after the Global Mirror relationships were stopped.

      If write throughput increases greatly (by 30% or more) after the Global Mirror relationships are stopped, the application host was attempting to perform more I/O than the remote link can sustain.

      When the Global Mirror relationships are active, the overloaded remote link causes higher response times to the application host. This overload, in turn, decreases the throughput of application host I/O at the source volume. After the Global Mirror relationships stop, the application host I/O sees a lower response time, and the true write throughput returns.

      To resolve this issue, increase the remote link bandwidth, reduce the application host I/O, or reduce the number of Global Mirror relationships.

Storage controllers

Investigate the primary and remote storage controllers, starting at the remote site. If the back-end storage at the secondary system is overloaded, or another problem is affecting the cache there, the Global Mirror protocol fails to keep up. Similarly, the problem exhausts the (**`gmlinktolerance`**) elasticity and has a similar effect at the primary system.

In this situation, ask the following questions:

- Are the storage controllers at the remote system overloaded (performing slowly)?

  Use IBM Spectrum Control to obtain the back-end write response time for each MDisk at the remote system. A response time for any individual MDisk that exhibits a sudden increase of 50 ms or more, or that is higher than 100 ms, generally indicates a problem with the back end. 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 FS9100 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 utilization on all the nodes, and it has not gotten near 100%, a high *Port to local node send response time* means that there is fabric congestion or a slow-draining Fibre Channel device.

A good indicator of SAN congestion is the *Zero Buffer Credit Percentage* on the port statistics (see “Buffer credits” on page 186 for more information on Buffer Credit). If any port is seeing over 10% zero buffer credits, that is definitely going to cause a problem for all I/O, not just Global Mirror writes. Values from 1 - 10% are moderately high and might contribute to performance issues.

For both primary and secondary systems, congestion on the fabric from other slow-draining devices becomes much less of an issue when only dedicated ports are used for node-to-node traffic within the system. However, this only really becomes an option on systems with more than four ports per node. Use port masking to segment your ports.

FlashCopy considerations

Check that FlashCopy mappings are in the *prepared* state. Check whether the Global Mirror target volumes are the sources of a FlashCopy mapping and whether that mapping was in the *prepared* state for an extended time.

Volumes in the prepared state are cache disabled, so their performance is impacted. To resolve this problem, start the FlashCopy mapping, which 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 163). 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 FS9100 this is less of an issue having 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 doesn’t affect directly the remote copy operations since it is logically placed below the remote copy processing in the I/O stack, as shown in Figure 6-6 on page 153. Nevertheless, in some circumstances, especially with write intensive environments, the copy-on-write process tends to stress some systems’s internal resources, like CPU and memory, and this condition can also affect the remote copy, that competes for the same resources, leading eventually to 1920 errors.

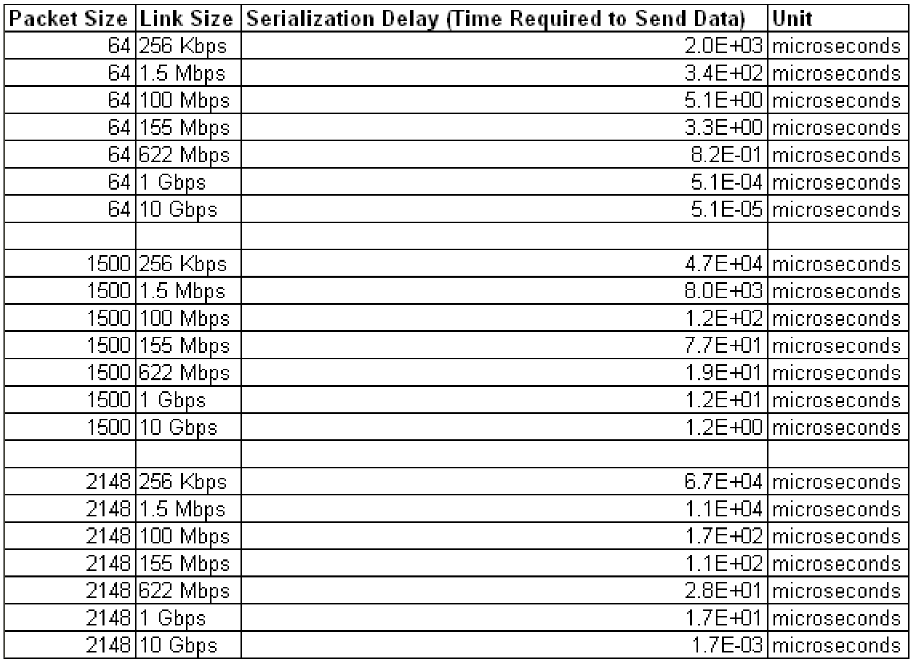
FCIP considerations

When you get a 1920 error, always check the latency first. The FCIP routing layer can introduce latency if it is not properly configured. If your network provider reports a much lower latency, you might have a problem at your FCIP routing layer. Most FCIP routing devices have built-in tools to enable you to check the RTT. When you are checking latency, remember that TCP/IP routing devices (including FCIP routers) report RTT by using standard 64-byte ping packets.

In Figure 6-25 on page 213, 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 6-25 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 6-25 Effect of packet size (in bytes) versus the link size

In Figure 6-25, the amount of time in microseconds that is required to transmit a packet across network links of varying bandwidth capacity is compared. The following packet sizes are used:

- 64 bytes: The size of the common ping packet
- 1500 bytes: The size of the standard TCP/IP packet
- 2148 bytes: The size of an FC frame

Finally, your path maximum transmission unit (MTU) affects the delay that is incurred to get a packet from one location to another location. An MTU might cause fragmentation, or be too large and cause too many retransmits when a packet is lost.

Recovery after 1920 errors

After a 1920 error occurs, the Global Mirror auxiliary volumes are no longer in a Consistent Synchronized state. You must establish the cause of the problem and fix it before you restart the relationship.

When the relationship is restarted, you must resynchronize it. During this period, the data on the Metro Mirror or Global Mirror auxiliary volumes on the secondary system is inconsistent, and your applications cannot use the volumes as backup disks. To address this data consistency exposure on the secondary system, a FlashCopy of the auxiliary volumes can be created to maintain a consistent image until the Global Mirror (or the Metro Mirror) relationships are synchronized again and back in a consistent state.

IBM Spectrum Virtualize 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 once restarted. Instead, the system uses a secondary *change volume* to automatically copy the previous consistent state of the secondary volume.

The relationship automatically moves to the `Consistent copying` status as the system resynchronizes and protects the consistency of the data. The relationship status changes to `Consistent synchronized` when the resynchronization process completes. For further details about the Consistency Protection feature, see *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize V8.2.1*, SG24-7933.

To ensure that the system can handle the background copy load, delay restarting the Metro Mirror or Global Mirror relationship until a quiet period occurs. If the required link capacity is unavailable, you might experience another 1920 error, and the Metro Mirror or Global Mirror relationship 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. Further details about IBM Copy Services Manager can be found at:

`IBM Copy Services Manager`

> **Tip:** When implementing automatic restart functions, it is advised to preserve the data consistency on GM target volumes during the resychronization using features like Flashcopy or Consistency Protection.

Adjusting the Global Mirror settings

Although the default values are valid in most configurations, the settings of the **`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 based products running code V7.2 or higher. It does so by using the built-in networking ports or optional 1/10 Gb adapter.

Following a recent partnership with IBM, native IP replication uses SANslide technology developed by Bridgeworks Limited of Christchurch, UK. They specialize in products that can bridge storage protocols and accelerate data transfer over long distances. Adding this technology at each end of a wide area network (WAN) TCP/IP link significantly improves the utilization of the link.

It does this by applying patented artificial intelligence (AI) to hide latency that is normally associated with WANs. Doing so can greatly improve the performance of mirroring services, in particular Global Mirror with Change Volumes (GM/CV) over long distances.

6.4.1 Native IP replication technology

Remote Mirroring over IP communication is supported on the IBM Spectrum Virtualize based 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 6-26.

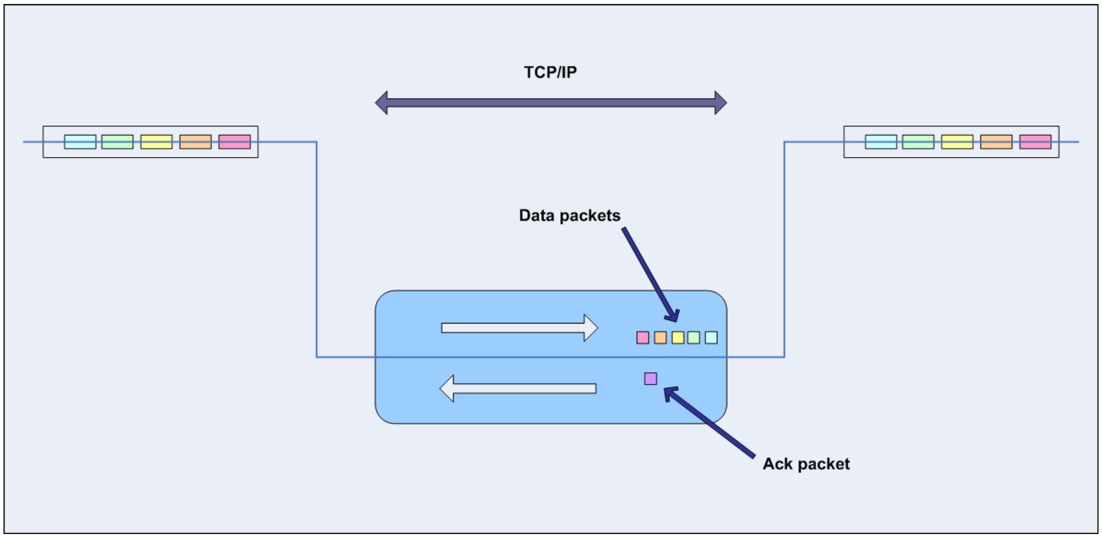


Figure 6-26 Typical Ethernet network data flow

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 6-27.

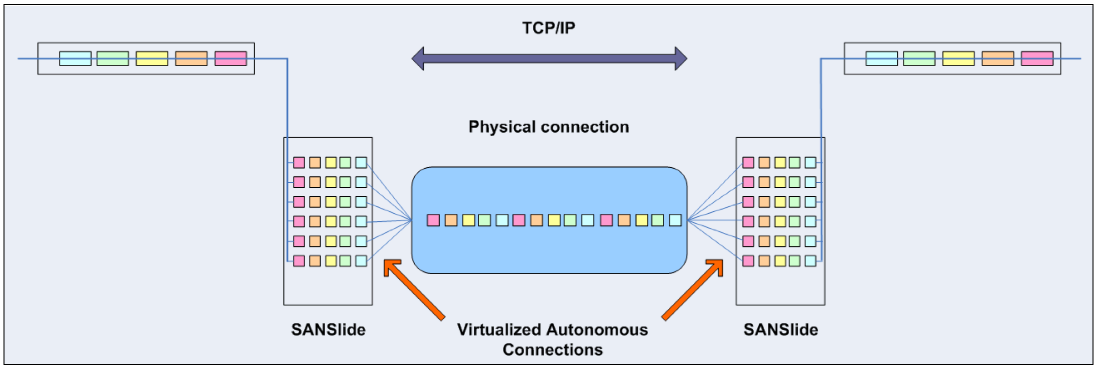


Figure 6-27 Optimized network data flow by using Bridgeworks SANSlide technology

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 involved in the partnership are successfully installed with V7.2 or later code levels.
- The systems have the necessary licenses that enable remote copy partnerships to be configured between two systems. No separate license is required to enable IP partnership.
- The storage SANs are configured correctly and the correct infrastructure to support the systems in remote copy partnerships over IP links is in place.
- The two systems must be able to ping each other and perform the discovery.
- The maximum number of partnerships between the local and remote systems, including both IP and Fibre Channel (FC) partnerships, is limited to the current maximum that is supported, which is three partnerships (four systems total).
- Only a single partnership over IP is supported.
- A system can have simultaneous partnerships over FC and IP, but with separate systems. The FC zones between two systems must be removed before an IP partnership is configured.
- IP partnerships are supported on both 10 gigabits per second (Gbps) links and 1 Gbps links. However, the intermix of both on a single link is not supported.
- The maximum supported round-trip time is 80 milliseconds (ms) for 1 Gbps links.
- The maximum supported round-trip time is 10 ms for 10 Gbps links.
- The minimum supported link bandwidth is 10 Mbps.
- The inter-cluster heartbeat traffic uses 1 Mbps per link.
- Only nodes from two I/O Groups can have ports that are configured for an IP partnership.
- Migrations of remote copy relationships directly from FC-based partnerships to IP partnerships are not supported.
- IP partnerships between the two systems can be over IPv4 or IPv6 only, but not both.
- An added layer of security is provided by using Challenge Handshake Authentication Protocol (CHAP) authentication.
- Direct attached systems configurations are supported with the following restrictions:
  - Only two direct attach link are allowed.
  - The direct attach links must be on the same I/O group.
  - Use two port groups, where a port group contains only the two ports that are directly linked.
- Transmission Control Protocol (TCP) ports 3260 and 3265 are used for IP partnership communications. Therefore, these ports must be open in firewalls between the systems.
- Network address translation (NAT) between systems that are being configured in an IP Partnership group is not supported.

- Only a single Remote Copy data session per physical link can be established. It is intended that only one connection (for sending/receiving Remote Copy data) is made for each independent physical link between the systems.

> **Note:** A physical link is the physical IP link between the two sites, A (local) and B (remote). Multiple IP addresses on local system A can be connected (by Ethernet switches) to this physical link. Similarly, multiple IP addresses on remote system B can be connected (by Ethernet switches) to the same physical link. At any point, only a single IP address on cluster A can form an RC data session with an IP address on cluster B.

- The maximum throughput is restricted based on the use of 1 Gbps or 10 Gbps Ethernet ports. The output varies based on distance (for example, round-trip latency) and quality of communication link (for example, packet loss). The maximum achievable throughput is the following:
  - One 1 Gbps port can transfer up to 110 MB
  - One 10 Gbps port can transfer up to 500 MB

6.4.3 VLAN support

VLAN tagging is supported for both iSCSI host attachment and IP replication. Hosts and remote-copy operations can connect to the system through Ethernet ports. Each traffic type has different bandwidth requirements, which can interfere with each other if they share IP connections. VLAN tagging creates two separate connections on the same IP network for different types of traffic. The system supports VLAN configuration on both IPv4 and IPv6 connections.

When the VLAN ID is configured for the IP addresses that are used for either iSCSI host attach or IP replication, the appropriate VLAN settings on the Ethernet network and servers must be configured correctly to avoid connectivity issues. After the VLANs are configured, changes to the VLAN settings disrupt iSCSI and IP replication traffic to and from the partnerships.

During the VLAN configuration for each IP address, the VLAN settings for the local and failover ports on two nodes of an I/O Group can differ. To avoid any service disruption, switches must be configured so the failover VLANs are configured on the local switch ports and the failover of IP addresses from a failing node to a surviving node succeeds. If failover VLANs are not configured on the local switch ports, there are no paths to IBM FS9100 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 FS9100 can average the IP compression capability to speed up replication cycles or to reduce the 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: IP compression feature does not require an RtC software license.

The data compression is made within the IP replication component of the IBM Spectrum Virtualize code. It can be used with all the remote copy technology (Metro Mirror, Global Mirror, and Global Mirror Change Volume). The IP compression is supported, with some restrictions, in the following systems:

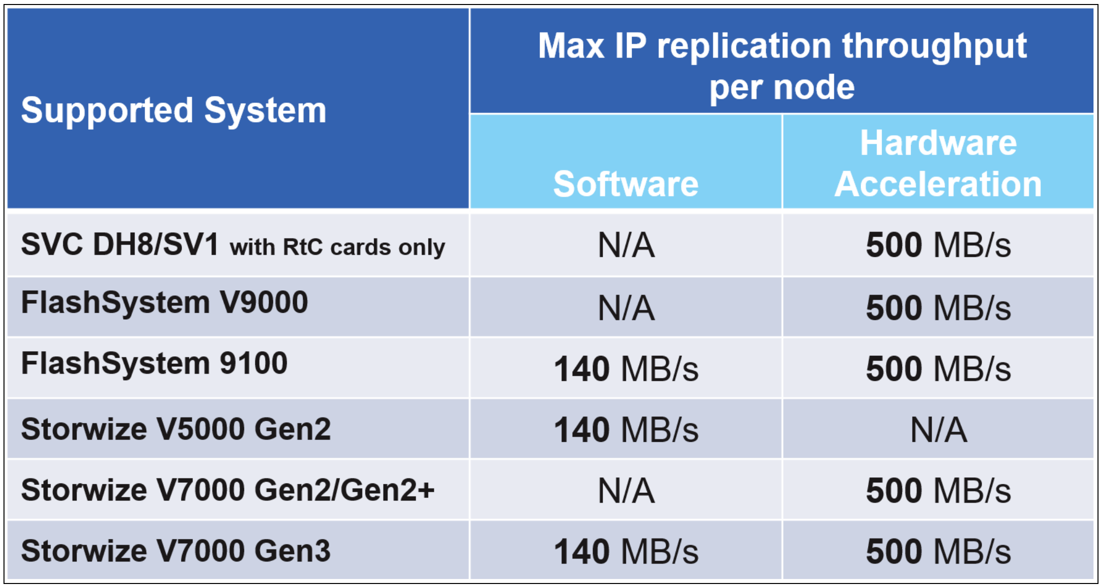
- SAN Volume controller with DH8 nodes
- SAN Volume controller with SV1nodes
- FlashSystem V9000
- Storwize V7000 Gen1
- Storwize V7000 Gen2, Gen2+, Gen3
- Storwize V5000 Gen2
- Storwize V5100, V500E
- FlashSystem 7200 / 9100 / 9200

The IP compression can be enabled on hardware configurations that support hardware-assisted compression acceleration engines only (RtC or integrated onboard compression acceleration engine). The IP compression feature provides two kinds of compression mechanisms: The hardware compression and software compression. The hardware compression is active whencompression accelerator engines are available, otherwise software compression is used.

Hardware compression makes use of currently underused cards. The internal resources are shared between RACE and IP compression. Software compression uses the system CPU and might have an impact on heavily used systems.

To evaluate the benefits of the IP compression, the Comprestimator tool can be used to estimate the compression ratio of the data to be replicated. The IP compression can be enabled and disabled without stopping the remote copy relationship by using the **`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.

Figure 6-28 reports the current compression limits by system type and compression mechanism.



Supported System	Max IP replication throughput per node	
	Software	Hardware Acceleration
SVC DH8/SV1 **with RtC cards only**	N/A	**500** MB/s
FlashSystem V9000	N/A	**500** MB/s
FlashSystem 9100	**140** MB/s	**500** MB/s
Storwize V5000 Gen2	**140** MB/s	N/A
Storwize V7000 Gen2/Gen2+	N/A	**500** MB/s
Storwize V7000 Gen3	**140** MB/s	**500** MB/s

Figure 6-28 IP compression limits by systems and compression types

6.4.5 Remote copy groups

This section describes remote copy groups (or remote copy port groups) and different ways to configure the links between the two remote systems. The two systems can be connected to each other over one link or, at most, two links. To address the requirement to enable the systems to know about the physical links between the two sites, the concept of remote copy port groups was introduced.

Remote copy port group ID is a numerical tag that is associated with an IP port of system to indicate which physical IP link it is connected to. Multiple IBM FS9100 canisters can be connected to the same physical long-distance link, and must therefore share a remote copy port group ID.

In scenarios with two physical links between the local and remote clusters, two remote copy port group IDs must be used to designate which IP addresses are connected to which physical link. This configuration must be done by the system administrator by using the GUI or the **`cfgportip`** CLI command. Note that the relationship between the physical links and the remote copy group IDs is not policed by the IBM Spectrum Virtualize code. This means that two different remote copy group can be used with a single physical link and vice versa.

Remember: IP ports on both partners must have been configured with identical remote copy port group IDs for the partnership to be established correctly.

The system IP addresses that are connected to the same physical link should be designated with identical remote copy port groups. The IBM Spectrum Virtualize based systems supports three remote copy groups: 0, 1, and 2.

The IP addresses are, by default, in remote copy port group 0. Ports in port group 0 are not considered for creating remote copy data paths between two systems. For partnerships to be established over IP links directly, IP ports must be configured in remote copy group 1 if a single inter-site link exists, or in remote copy groups 1 and 2 if two inter-site links exist.

You can assign one IPv4 address and one IPv6 address to each Ethernet port on the IBM FS9100 systems. Each of these IP addresses can be shared between iSCSI host attach and the IP partnership. The user must configure the required IP address (IPv4 or IPv6) on an Ethernet port with a remote copy port group.

The administrator might want to use IPv6 addresses for remote copy operations and use IPv4 addresses on that same port for iSCSI host attach. This configuration also implies that for two systems to establish an IP partnership, both systems must have IPv6 addresses that are configured.

Administrators can choose to dedicate an Ethernet port for IP partnership only. In that case, host access must be explicitly disabled for that IP address and any other IP address that is configured on that Ethernet port.

Note: To establish an IP partnership, each FS9100 canister must have only a single remote copy port group that is configured 1 or 2. The remaining IP addresses must be in remote copy port group 0.

Failover operations within and between port groups

Within one remote-copy port group, only one port from each system is selected for sending and receiving remote copy data at any one time. Therefore, on each system, at most one port for each remote-copy port group is reported as `used`.

If the IP partnership becomes unable to continue over an IP port, the system fails over to another port within that remote-copy port group. Some reasons this might occur are the switch to which it is connected fails, the node goes offline, or the cable that is connected to the port is unplugged.

For the IP partnership to continue during a failover, multiple ports must be configured within the remote-copy port group. If only one link is configured between the two systems, configure two ports (one per node) within the remote-copy port group. You can configure these two ports on two nodes within the same I/O group or within separate I/O groups. Configurations 4, 5, and 6 in IP partnership requirements are the supported dual-link configurations.

While failover is in progress, no connections in that remote-copy port group exist between the two systems in the IP partnership for a short time. Typically, failover completes within 30 seconds to 1 minute. If the systems are configured with two remote-copy port groups, the failover process within each port group continues independently of each other.

The disadvantage of configuring only one link between two systems is that, during a failover, a discovery is initiated. When the discovery succeeds, the IP partnership is reestablished. As a result, the relationships might stop, in which case a manual restart is required. To configure two intersystem links, you must configure two remote-copy port groups.

When a node fails in this scenario, the IP partnership can continue over the other link until the node failure is rectified. Failback then happens when both links are again active and available to the IP partnership. The discovery is triggered so that the active IP partnership data path is made available from the new IP address.

In a two-node system, or if there is more than one I/O Group and the node in the other I/O group has IP ports pre-configured within the remote-copy port group, the discovery is triggered. The discovery makes the active IP partnership data path available from the new IP address.

6.4.6 Supported configurations examples

Multiple IP partnership configurations are available depending on the number of physical links and the number of nodes. In the following sections, some example configurations are described.

Single inter-site link configurations

Consider two 2-node systems in IP partnership over a single inter-site link (with failover ports configured), as shown in Figure 6-29.

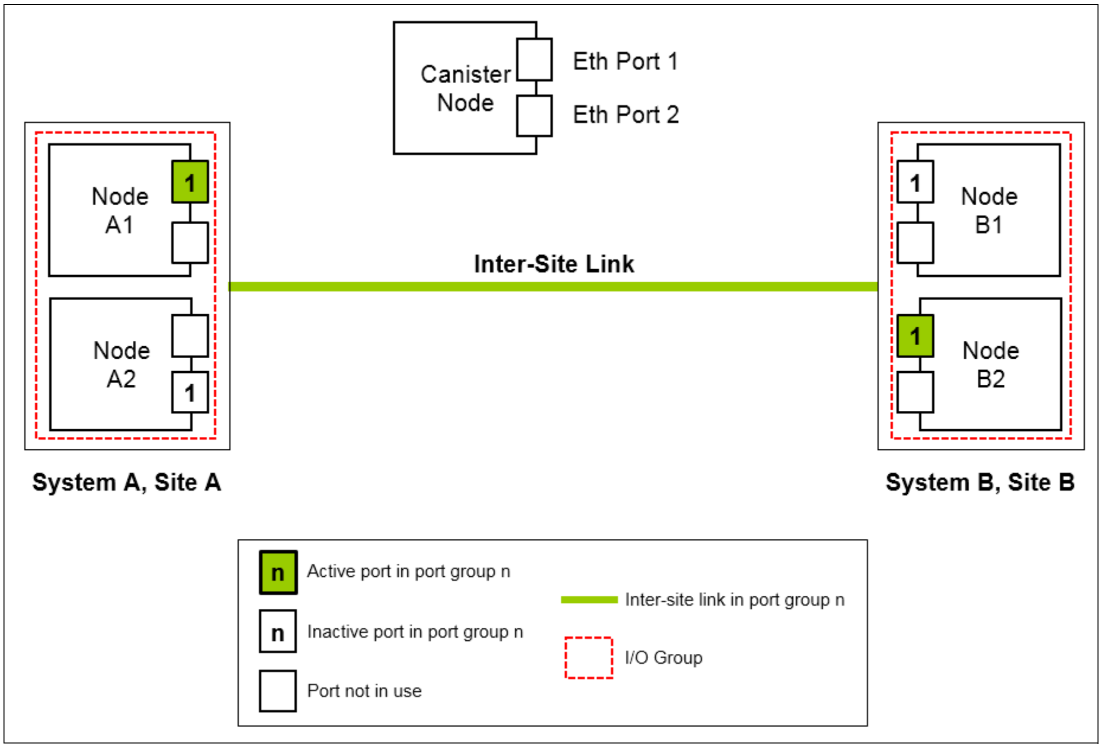


Figure 6-29 Only one remote copy group on each system and nodes with failover ports configured

Figure 6-29 shows two systems: System A and System B. A single remote copy port group 1 is configured on two Ethernet ports, one each on Node A1 and Node A2 on System A. Similarly, a single remote copy port group is configured on two Ethernet ports on Node B1 and Node B2 on System B.

Although two ports on each system are configured for remote copy port group 1, only one Ethernet port in each system actively participates in the IP partnership process. This selection is determined by a path configuration algorithm that is designed to choose data paths between the two systems to optimize performance.

The other port on the partner node in the I/O Group behaves as a standby port that is used during a node failure. If Node A1 fails in System A, IP partnership continues servicing replication I/O from Ethernet Port 2 because a failover port is configured on Node A2 on Ethernet Port 2.

However, it might take some time for discovery and path configuration logic to reestablish paths post failover. This delay can cause partnerships to change to `Not_Present` for that time. The details of the particular IP port that is actively participating in IP partnership is provided in the **`lsportip`** output (reported as `used`).

This configuration has the following characteristics:

- Each node in the I/O group has the same remote copy port group that is configured. However, only one port in that remote copy port group is active at any time at each system.
- If Node A1 in System A or Node B2 in System B fails in the respective systems, IP partnerships rediscovery is triggered and continues servicing the I/O from the failover port.
- The discovery mechanism that is triggered because of failover might introduce a delay where the partnerships momentarily change to the `Not_Present` state and recover.

Figure 6-30 shows a configuration with two 4-node systems in IP partnership over a single inter-site link (with failover ports configured).

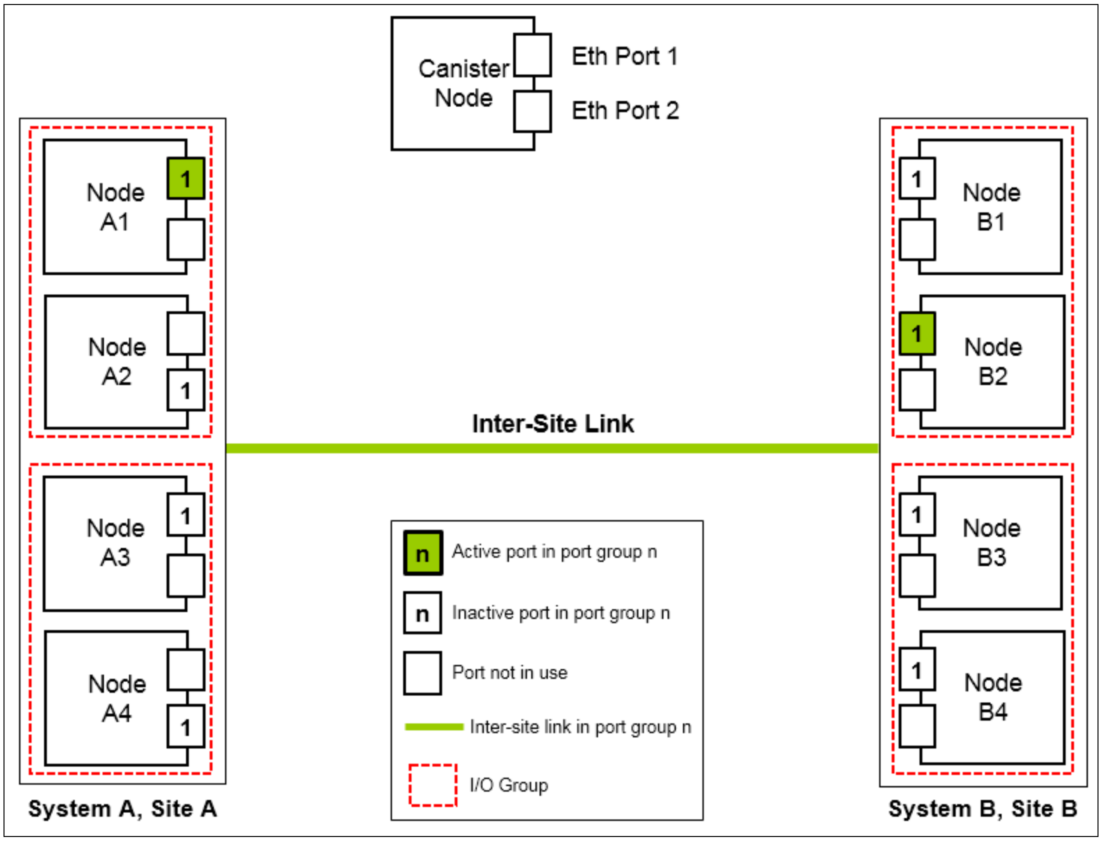


Figure 6-30 Multinode systems single inter-site link with only one remote copy port group

Figure 6-30 shows two 4-node systems: System A and System B. A single remote copy port group 1 is configured on nodes A1, A2, A3, and A4 on System A, Site A, and on nodes B1, B2, B3, and B4 on System B, Site B.

Although four ports are configured for remote copy group 1, only one Ethernet port in each remote copy port group on each system actively participates in the IP partnership process. Port selection is determined by a path configuration algorithm. The other ports play the role of standby ports.

If Node A1 fails in System A, the IP partnership selects one of the remaining ports that is configured with remote copy port group 1 from any of the nodes from either of the two I/O groups in System A. However, it might take some time (generally seconds) for discovery and path configuration logic to reestablish paths post failover. This process can cause partnerships to change to the `Not_Present` state.

This result causes remote copy relationships to stop. The administrator might need to manually verify the issues in the event log and start the relationships or remote copy consistency groups, if they do not automatically recover. The details of the particular IP port actively participating in the IP partnership process is provided in the **lsportip** view (reported as `used`). This configuration has the following characteristics:

- Each node has the remote copy port group that is configured in both I/O groups. However, only one port in that remote copy port group remains active and participates in IP partnership on each system.
- If Node A1 in System A or Node B2 in System B encounter some failure in the system, IP partnerships discovery is triggered and continues servicing the I/O from the failover port.
- The discovery mechanism that is triggered because of failover might introduce a delay where the partnerships momentarily change to the `Not_Present` state and then recover.
- The bandwidth of the single link is used completely.

An eight-node system in IP partnership with four-node system over single inter-site link is shown in Figure 6-31.

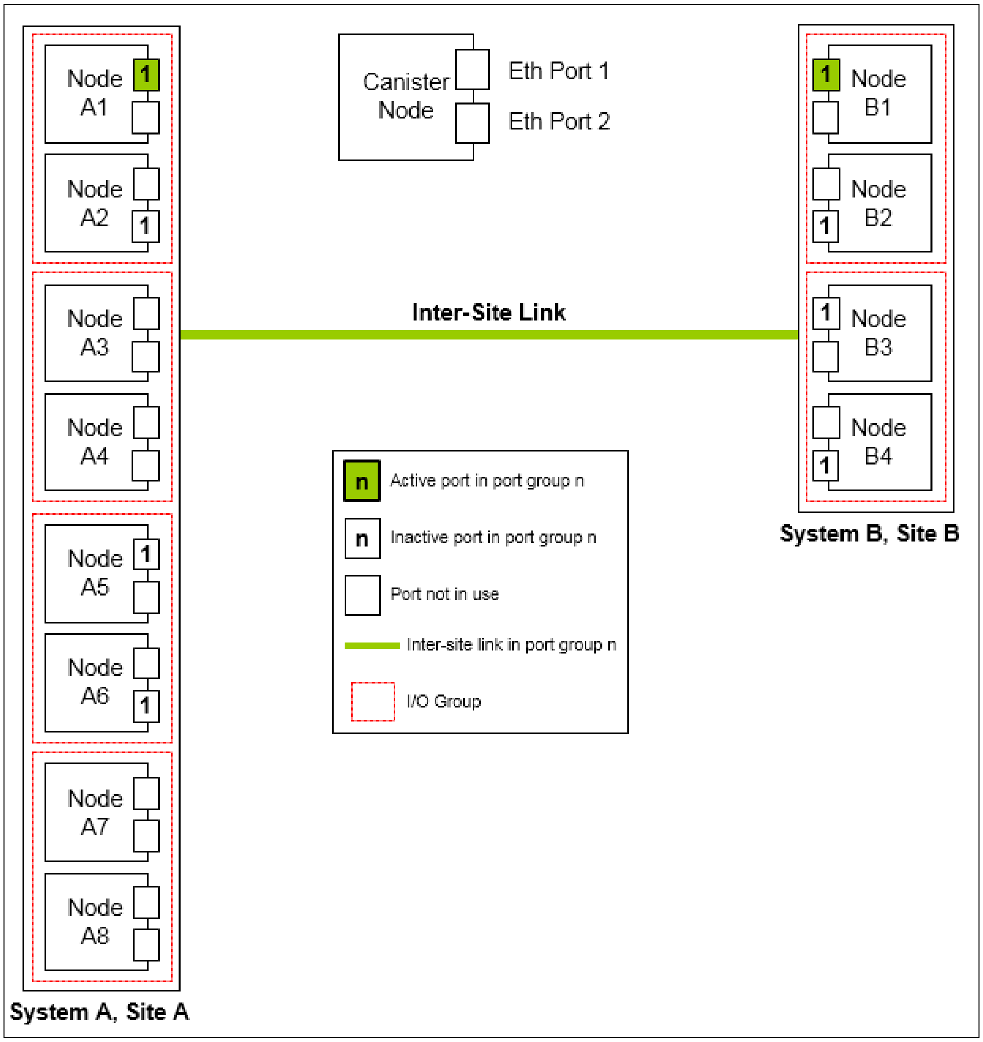


Figure 6-31 Multinode systems single inter-site link with only one remote copy port group

Figure 6-31 shows an eight-node system (System A in Site A) and a four-node system (System B in Site B). A single remote copy port group 1 is configured on nodes A1, A2, A5, and A6 on System A at Site A. Similarly, a single remote copy port group 1 is configured on nodes B1, B2, B3, and B4 on System B.

Although there are four I/O groups (eight nodes) in System A, any two I/O groups at maximum are supported to be configured for IP partnerships. If Node A1 fails in System A, IP partnership continues using one of the ports that is configured in remote copy port group from any of the nodes from either of the two I/O groups in System A.

However, it might take some time for discovery and path configuration logic to reestablish paths post-failover. This delay might cause partnerships to change to the `Not_Present` state.

This process can lead to remote copy relationships stopping. The administrator must manually start them if the relationships do not auto-recover. The details of which particular IP port is actively participating in IP partnership process is provided in **`lsportip`** output (reported as `used`).

This configuration has the following characteristics:

- Each node has the remote copy port group that is configured in both the I/O groups that are identified for participating in IP Replication. However, only one port in that remote copy port group remains active on each system and participates in IP Replication.
- If the Node A1 in System A or the Node B2 in System B fails in the system, the IP partnerships trigger discovery and continue servicing the I/O from the failover ports.
- The discovery mechanism that is triggered because of failover might introduce a delay where the partnerships momentarily change to the `Not_Present` state and then recover.
- The bandwidth of the single link is used completely.

Two inter-site link configurations

A two 2-node systems with two inter-site links configuration is depicted in Figure 6-32.

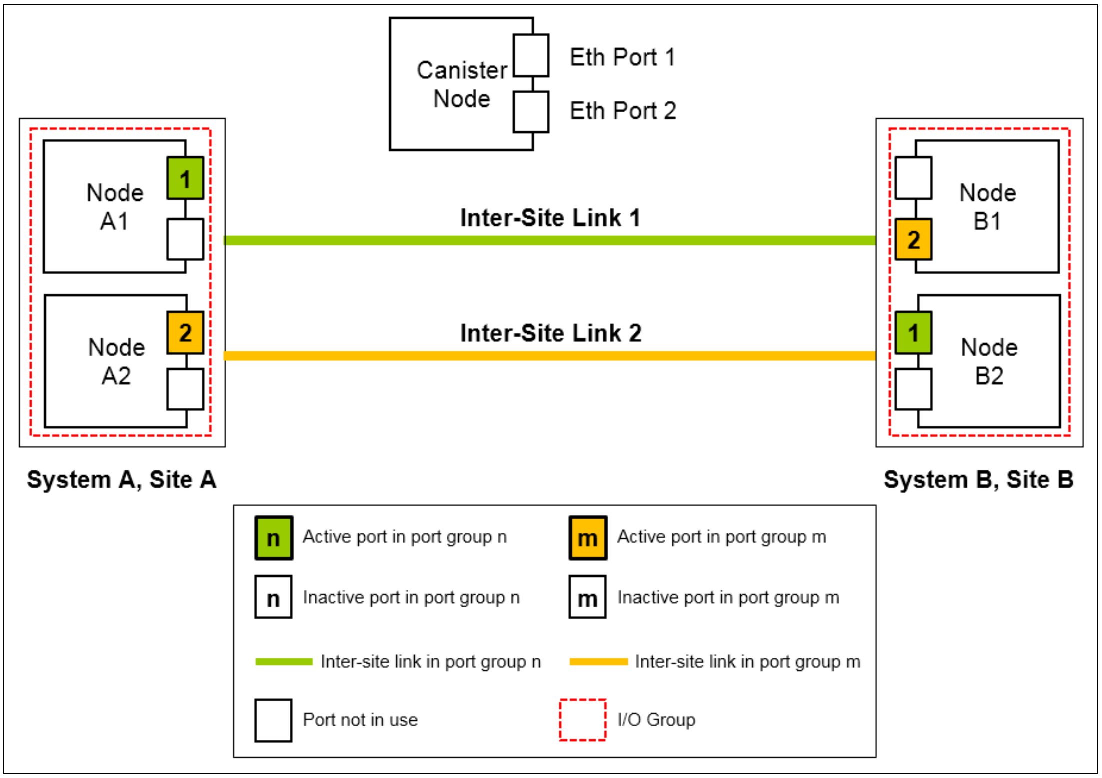

Canister Node
Eth Port 1
Eth Port 2
Node A1
1
Node A2
2
Inter-Site Link 1
Inter-Site Link 2
Node B1
2
Node B2
1
System A, Site A
System B, Site B
n
Active port in port group n
m
Active port in port group m
n
Inactive port in port group n
m
Inactive port in port group m
Inter-site link in port group n
Inter-site link in port group m
Port not in use
I/O Group

Figure 6-32 Dual links with two remote copy groups on each system configured

As shown in Figure 6-32, remote copy port groups 1 and 2 are configured on the nodes in System A and System B because two inter-site links are available. In this configuration, the failover ports are not configured on partner nodes in the I/O group. Rather, the ports are maintained in different remote copy port groups on both of the nodes. They can remain active and participate in IP partnership by using both of the links.

However, if either of the nodes in the I/O group fail (that is, if Node A1 on System A fails), the IP partnership continues only from the available IP port that is configured in remote copy port group 2. Therefore, the effective bandwidth of the two links is reduced to 50% because only the bandwidth of a single link is available until the failure is resolved.

This configuration has the following characteristics:

- There are two inter-site links, and two remote copy port groups are configured.
- Each node has only one IP port in remote copy port group 1 or 2.
- Both the IP ports in the two remote copy port groups participate simultaneously in IP partnerships. Therefore, both of the links are used.
- During node failure or link failure, the IP partnership traffic continues from the other available link and the port group. Therefore, if two links of 10 Mbps each are available and you have 20 Mbps of effective link bandwidth, bandwidth is reduced to 10 Mbps only during a failure.
- After the node failure or link failure is resolved and failback happens, the entire bandwidth of both of the links is available as before.

A configuration with two 4-node systems in IP partnership with dual inter-site links is shown in Figure 6-33.

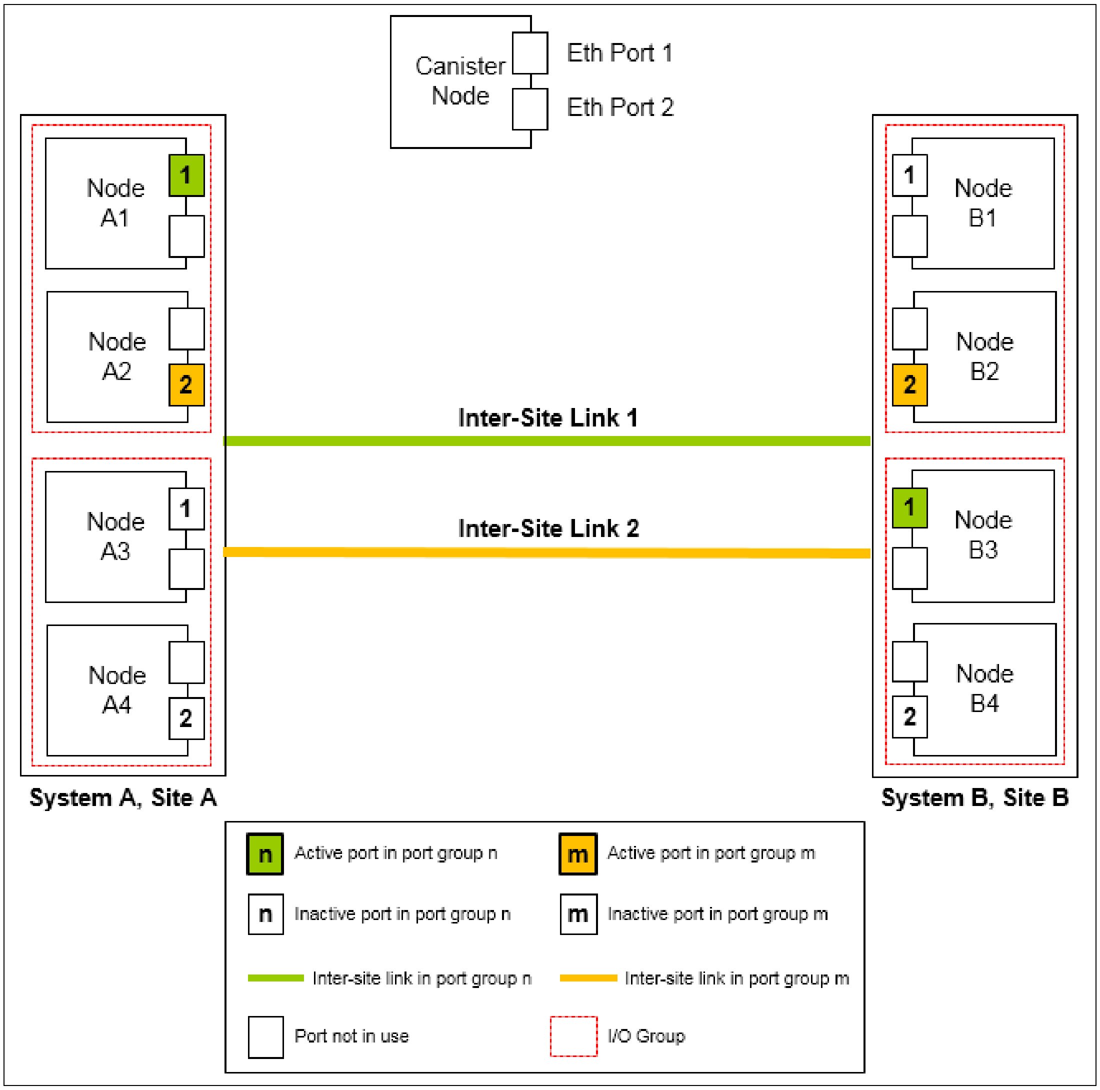


Figure 6-33 Multinode systems with dual inter-site links between the two systems

Figure 6-33 shows two 4-node systems: System A and System B. This configuration is an extension of Configuration 5 to a multinode multi-I/O group environment.

As seen in this configuration, there are two I/O groups. Each node in the I/O group has a single port that is configured in remote copy port groups 1 or 2.

Although two ports are configured in remote copy port groups 1 and 2 on each system, only one IP port in each remote copy port group on each system actively participates in IP partnership. The other ports that are configured in the same remote copy port group act as standby ports during a failure. Which port in a configured remote copy port group participates in IP partnership at any moment is determined by a path configuration algorithm.

In this configuration, if Node A1 fails in System A, IP partnership traffic continues from Node A2 (that is, remote copy port group 2). At the same time, the failover also causes discovery in remote copy port group 1. Therefore, the IP partnership traffic continues from Node A3 on which remote copy port group 1 is configured. The details of the particular IP port that is actively participating in IP partnership process is provided in the **`lsportip`** output (reported as `used`).

This configuration has the following characteristics:

- Each node has the remote copy port group that is configured in the I/O groups 1 or 2. However, only one port per system in both remote copy port groups remains active and participates in IP partnership.
- Only a single port per system from each configured remote copy port group participates simultaneously in IP partnership. Therefore, both of the links are used.
- During node failure or port failure of a node that is actively participating in IP partnership, IP partnership continues from the alternative port because another port is in the system in the same remote copy port group, but in a different I/O Group.
- The pathing algorithm can start discovery of available port in the affected remote copy port group in the second I/O group and pathing is reestablished. This process restores the total bandwidth, so both of the links are available to support IP partnership.

Finally, an eight-node system in IP partnership with a four-node system over dual inter-site links is depicted in Figure 6-34.

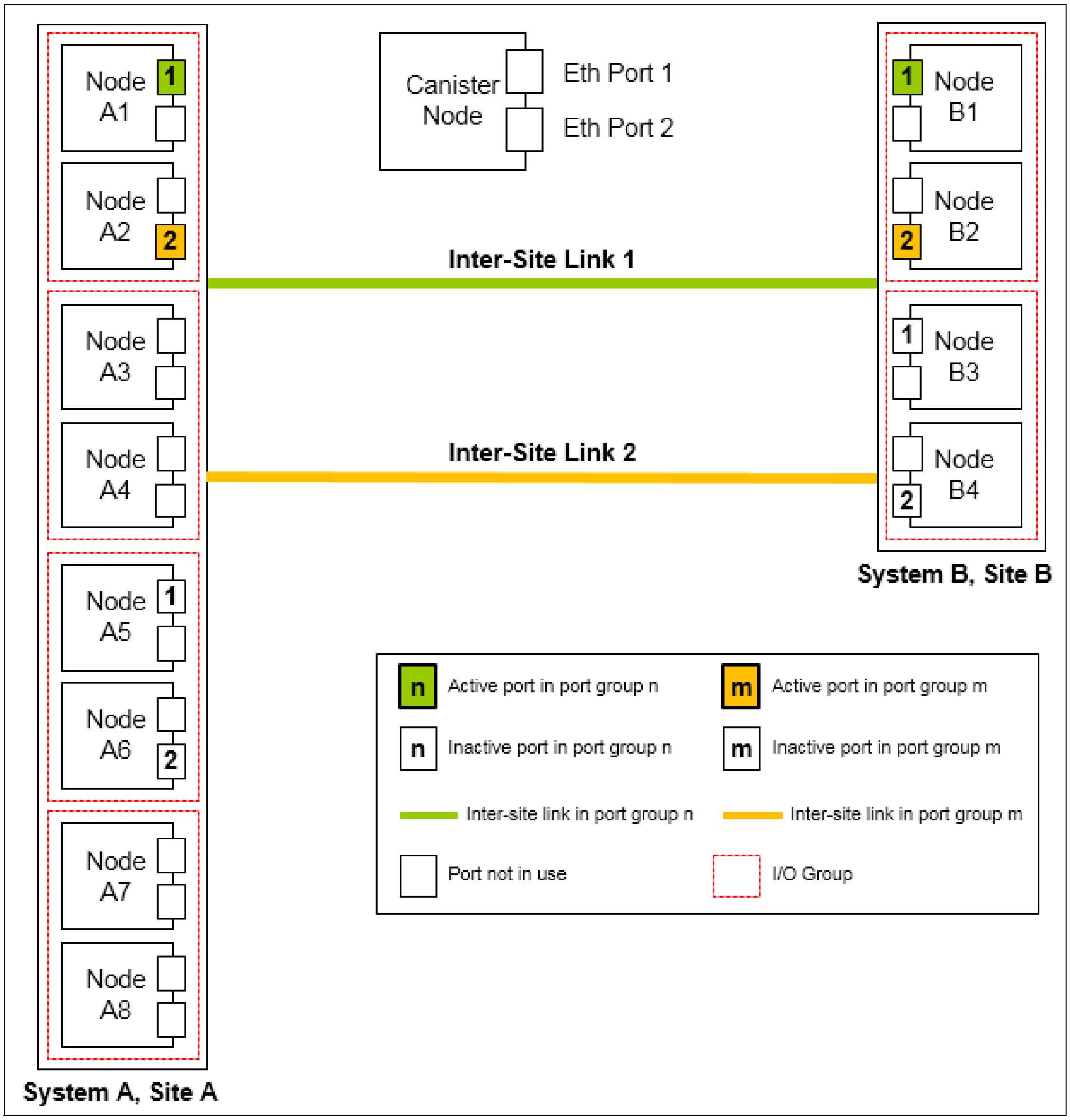


Figure 6-34 Multinode systems with dual inter-site links between the two systems

Figure 6-34 shows an eight-node System A in Site A and a four-node System B in Site B. Because a maximum of two I/O groups in IP partnership is supported in a system, although there are four I/O groups (eight nodes), nodes from only two I/O groups' are configured with remote copy port groups in System A. The remaining or all of the I/O groups can be configured to be remote copy partnerships over FC.

In this configuration, there are two links and two I/O groups that are configured with remote copy port groups 1 and 2. However, path selection logic is managed by an internal algorithm. Therefore, this configuration depends on the pathing algorithm to decide which of the nodes actively participate in IP partnership. Even if Node A5 and Node A6 are configured with remote copy port groups properly, active IP partnership traffic on both of the links can be driven from Node A1 and Node A2 only.

If Node A1 fails in System A, IP partnership traffic continues from Node A2 (that is, remote copy port group 2). The failover also causes IP partnership traffic to continue from Node A5 on which remote copy port group 1 is configured. The details of the particular IP port actively participating in IP partnership process is provided in the **`lsportip`** output (reported as `used`).

This configuration has the following characteristics:

- There are two I/O Groups with nodes in those I/O groups that are configured in two remote copy port groups because there are two inter-site links for participating in IP partnership. However, only one port per system in a particular remote copy port group remains active and participates in IP partnership.
- One port per system from each remote copy port group participates in IP partnership simultaneously. Therefore, both of the links are used.
- If a node or port on the node that is actively participating in IP partnership fails, the remote copy (RC) data path is established from that port because another port is available on an alternative node in the system with the same remote copy port group.
- The path selection algorithm starts discovery of available ports in the affected remote copy port group in the alternative I/O groups and paths are reestablished. This process restores the total bandwidth across both links.
- The remaining or all of the I/O groups can be in remote copy partnerships with other systems.

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 6-35 shows the throughput trend for a 1 Gbps port in respect of the packet loss ratio and the latency.

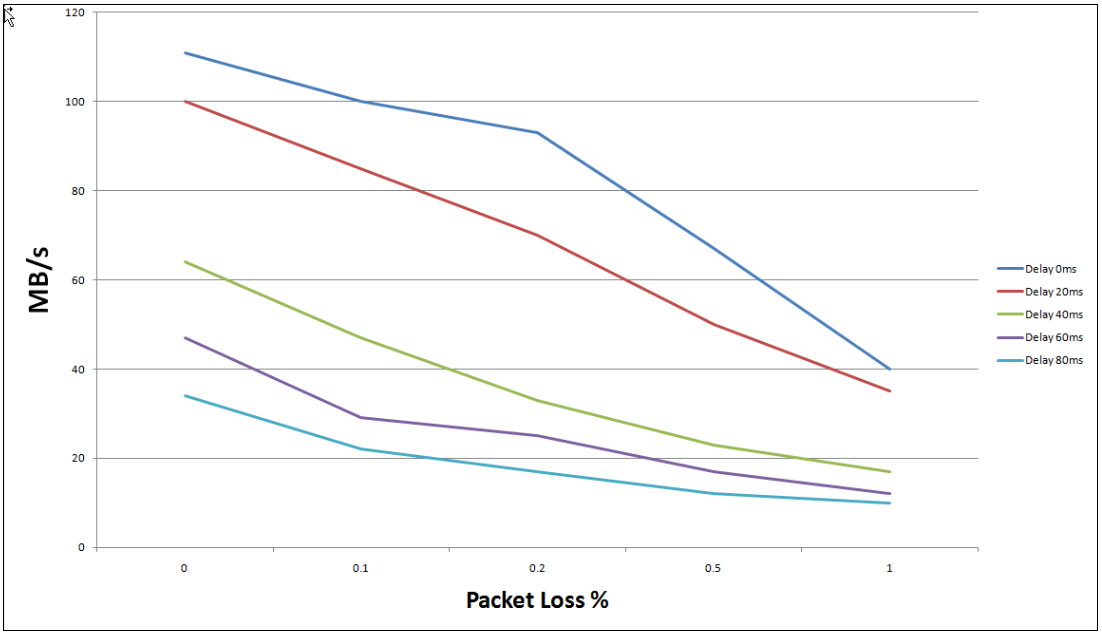


Figure 6-35 1 Gbps port throughput trend

The chart shows how the combined effect of the packet loss and the latency can lead to a throughput reduction of more than 85%. For these reasons, the IP replication option should only be considered for the replication configuration not affected by poor quality and performing networks. Due to its characteristic of low-bandwidth requirement, the Global Mirror Change Volume is the preferred solution with the IP replication.

The following recommendations might help improve this performance when using compression and IP partnership in the same system:

- To use the IP partnership on a multiple I/O group system that has nodes older than SAN Volume Controller 2145-CG8 and compressed volumes, configure ports for the IP partnership in I/O groups that do not contain compressed volumes.
- To use the IP partnership on Storwize Family product that has compressed volumes, configure ports for the IP partnership in I/O groups that do not contain compressed volumes.
- 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 6-36 provides an overview of Volume Mirroring.

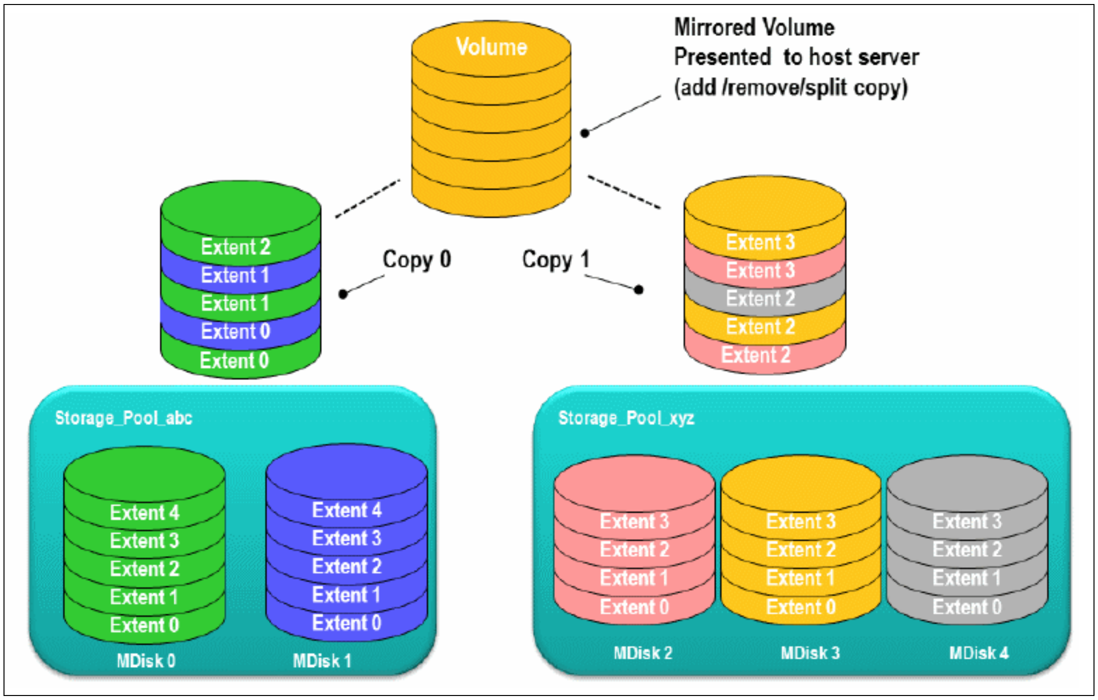


Figure 6-36 Volume Mirroring overview

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.

Important: For best performance, keep consistency between Hosts, Nodes, and Storage Controller site affinity as long as possible.

During back-end storage failure, note the following points:

- If one of the mirrored volume copies is temporarily unavailable, the volume remains accessible to servers.
- The system remembers which areas of the volume are written and resynchronizes these areas when both copies are available.
- The remaining copy can service read I/O when the failing one is offline, without user intervention.

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 FS9100 as a migration device. For example, suppose vendor A’s product cannot communicate with vendor B’s product, but you need to migrate existing data from vendor A to vendor B.

  Using *Export to image mode* allows you to migrate data by using the Copy Services functions and then return control to the native array, while maintaining access to the hosts.

- Import to Image mode

  This option allows you to import an existing storage MDisk or logical unit number (LUN) with its existing data from an external storage system, without putting metadata on it. The existing data remains intact. After you import it, the volume mirroring function can be used to migrate the storage to the other locations, while the data remains accessible to your hosts.

- Volume cloning using Volume Mirroring and then using the Split into New Volume option

  This option allows any volume to be cloned without any interruption to the host access. You have to create two mirrored copies of data and then break the mirroring with the split option to make two independent copies of data. This option doesn’t apply to already mirrored volumes.

- Volume pool migration using the volume mirroring option

  This option allows any volume to be moved between storage pools without any interruption to the host access. You might use this option to move volumes as an alternative to Migrate to Another Pool function.

Compared to the Migrate to Another Pool function, volume mirroring provides more manageability because it can be suspended and resumed anytime and also it allows you to move volumes among pools with different extent sizes. This option doesn't apply to already mirrored volumes.

> **Use Case:** Volume Mirroring can be used to migrate volumes from and to Data Reduction Pools which do not support extent based migrations. See 4.7.5, "Data migration with DRP" on page 99.

- Volume capacity saving change

  This option allows you to modify the capacity saving characteristics of any volume from standard to thin provisioned or compressed and vice versa, without any interruption to host access. This option works the same as the volume pool migration but specifying a different capacity saving for the newly created copy. This option doesn't apply to already mirrored volumes.

When you use Volume Mirroring, consider how quorum candidate disks are allocated. Volume Mirroring maintains some state data on the quorum disks. If a quorum disk is not accessible and Volume Mirroring is unable to update the state information, a mirrored volume might need to be taken offline to maintain data integrity. To ensure the high availability of the system, ensure that multiple quorum candidate disks, which are allocated on different storage systems, are configured.

> **Quorum disk consideration:** Mirrored volumes can be taken offline if there is no quorum disk available. This behavior occurs because synchronization status for mirrored volumes is recorded on the quorum disk. To protect against mirrored volumes being taken offline, follow the guidelines for setting up quorum disks.

The following are other Volume Mirroring usage cases and characteristics:

- Creating a mirrored volume:
  - The maximum number of copies is two.
  - Both copies are created with the same virtualization policy.

    To have a volume mirrored using different policies, you need to add a volume copy with a different policy to a volume that has only one copy.
  - Both copies can be located in different storage pools. The first storage pool that is specified contains the primary copy.
  - It is not possible to create a volume with two copies when specifying a set of 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/validate in three ways. This compares volume copies and performs these functions:
  - Reports the first difference found. It can iterate by starting at a specific LBA by using the `-startlba` parameter.
  - Creates virtual medium errors where there are differences.
  - Corrects the differences that are found (reads from primary copy and writes to secondary copy).
- View to list volumes affected by a back-end disk subsystem being offline:
  - Assumes that a standard use is for mirror between disk subsystems.
  - Verifies that mirrored volumes remain accessible if a disk system is being shut down.
  - Reports an error in case a quorum disk is on the back-end disk subsystem.
- Expand or shrink a volume:
  - This function works on both of the volume copies at once.
  - All volume copies always have the same size.
  - All copies must be synchronized before expanding or shrinking them.

> **DRP limitation:** Data Reduction Pools 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 (Table 6-9 on page 237), or at the defined rate while creating or modifying the volume (see 6.5.5, "Volume Mirroring performance considerations" on page 236 for the effect of the copy rate setting). Once the synchronization process is completed, the volume mirroring copies are *in-sync* state.

By default when a mirrored volume is created with **`mkvdisk`** command the Copy 0 is overwritten with zeros. After this formatting, the volume starts the copies synchronization process, making the Copy 1 zeroed as well. The format and sychronization processes are initiated by default at the time of the volume creation without keeping the volume offline.

You can specify that a volume is synchronized (**`-createsync`** parameter), even if it is not. Using this parameter can cause data corruption if the primary copy fails and leaves an unsynchronized secondary copy to provide data. Using this parameter can cause loss of read stability in unwritten areas if the primary copy fails, data is read from the primary copy, and then different data is read from the secondary copy. To avoid data loss or read stability loss, use this parameter only for a primary copy that has been formatted and not written to.

Another example use case for **`-createsync`** is for a newly created mirrored volume where both copies are thin provisioned or compressed because no data has been written to disk and unwritten areas return zeros (0). If the synchronization between the volume copies has been lost, the resynchronization process is incremental. This term means that only grains that have been written to need to be copied, and then get synchronized volume copies again.

The progress of the volume mirror synchronization can be obtained from the GUI or by using the **`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 - 100. A value of 0 disables synchronization.

Table 6-9 shows the relationship between the *rate value* and the *data copied per second*.

Table 6-9 Relationship between the rate value and the data copied per second

User-specified rate attribute value per volume	Data copied/sec
0	Synchronization is disabled
1 - 10	128 KB
11 - 20	256 KB
21 - 30	512 KB
31 - 40	1 MB
41 - 50	2 MB **** 50% is the default value**
51 - 60	4 MB
61 - 70	8 MB
71 - 80	16 MB
81 - 90	32 MB
91 - 100	64 MB

Rate attribute value: The rate attribute is configured on each volume that you want to mirror. The default value of a new volume mirror is 50%.

In large, IBM FS9100 system configurations, the settings of the copy rate can considerably affect the performance in scenarios where a back-end storage failure occurs. For instance, consider a scenario where a failure of a back-end storage controller is affecting one copy of 300 mirrored volumes. The host continues the operations by using the remaining copy.

When the failed controller comes back online, the resynchronization process for all the 300 mirrored volumes starts at the same time. With a copy rate of 100 for each volume, this process would add a theoretical workload of 18.75 GB/s, which will 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 236.

Volume mirroring I/O Time-out configuration

A mirrored volume has pointers to the two copies of data, usually in different storage pools, and each write completes on both copies before the host receives I/O completion status. For a synchronized mirrored volume, if a write I/O to a copy has failed or a long timeout has expired, then system has completed all available controller level Error Recovery Procedures (ERPs). In this case, that copy is taken offline and goes out of sync. The volume remains online and continues to service I/O requests from the remaining copy.

The *Fast Failover* feature isolates hosts from temporarily poorly-performing back-end storage of one Copy at the expense of a short interruption to redundancy. The fast failover feature behavior is that during normal processing of host write I/O, the system submits writes to both copies with a timeout of 10 seconds (20 seconds for stretched volumes). If one write succeeds and the other write takes longer than 5 seconds, then the slow write is stopped. The Fibre Channel `abort` sequence can take around 25 seconds.

When the stop is completed, one copy is marked as out of sync and the host write I/O completed. The overall fast failover ERP aims to complete the host I/O in around 30 seconds (40 seconds for stretched volumes).

In v6.3.x and later, the fast failover can be set for *each* mirrored volume by using the **`chvdisk`** command and the **`mirror_write_priority`** attribute settings:

- *Latency* (default value): A short timeout prioritizing low host latency. This option enables the fast failover feature.
- *Redundancy*: A long timeout prioritizing redundancy. This option indicates a copy that is slow to respond to a write I/O can use the full ERP time. The response to the I/O is delayed until it completes to keep the copy in sync if possible. This option disables the fast failover feature.

Volume Mirroring ceases to use the slow copy for 4 - 6 minutes, and subsequent I/O data is not affected by a slow copy. Synchronization is suspended during this period. After the copy suspension completes, Volume Mirroring resumes, which allows I/O data and synchronization operations to the slow copy that will, typically, quickly complete the synchronization.

If another I/O times out during the synchronization, then the system stops using that copy again for 4 - 6 minutes. If one copy is always slow, then the system tries it every 4 - 6 minutes and the copy gets progressively more out of sync as more grains are written. If fast failovers are occurring regularly, there is probably an underlying performance problem with the copy's back-end storage.

The preferred **`mirror_write_priority`** setting for the Enhanced Stretched Cluster configurations is *latency*.

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 (Table 6-10).

Table 6-10 Relationship of bitmap space to Volume Mirroring address space

Function	Grain size in KB	1 byte of bitmap space gives a total of	4 KB of bitmap space gives a total of	1 MB of bitmap space gives a total of	20 MB of bitmap space gives a total of	512 MB of bitmap space gives a total of
Volume Mirroring	256	2 MB of volume capacity	8 GB of volume capacity	2 TB of volume capacity	40 TB of volume capacity	1024 TB of volume capacity

Shared bitmap space: This bitmap space on one I/O group is shared between Metro Mirror, Global Mirror, FlashCopy, and Volume Mirroring.

The command to create Mirrored Volumes can fail if there is not enough space to allocate bitmaps in the target I/O Group. To verify and change the space allocated and available on each I/O Group with the CLI, see the Example 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
id 0
name io_grp0
node_count 2
vdisk_count 9
host_count 0
flash_copy_total_memory 20.0MB
flash_copy_free_memory 19.9MB
remote_copy_total_memory 20.0MB
remote_copy_free_memory 19.9MB
mirroring_total_memory 20.0MB
mirroring_free_memory 20.0MB
raid_total_memory 40.0MB
raid_free_memory 40.0MB
.
lines removed for brevity
.
IBM_FlashSystem:ITSO:superuser>chiogrp -feature mirror -size 64 io_grp0
IBM_FlashSystem:ITSO:superuser>lsiogrp io_grp0
id 0
name io_grp0
node_count 2
vdisk_count 9
host_count 0
flash_copy_total_memory 20.0MB
flash_copy_free_memory 19.9MB
remote_copy_total_memory 20.0MB
remote_copy_free_memory 19.9MB
mirroring_total_memory 64.0MB
mirroring_free_memory 64.0MB
.
lines removed for brevity
.
```

7

Hosts

This chapter describes the guidelines on how to configure host systems on IBM FlashSystem 9100 by following several preferred practices. A *host system* is a computer that is connected to the switch through a Fibre Channel (FC), iSCSI and iSCSI Extensions over RDMA (iSER).

Characteristics of the host play an important role in tuning, troubleshooting and performance of your IBM FlashSystem 9100. Consider the following areas for performance:

- The use of multipathing and bandwidth (physical capability of SAN)
- Understanding how your host performs I/O and the types of I/O
- The use of measurement and test tools to determine host performance and for tuning

This chapter supplements the IBM System Storage Spectrum Virtualize V8.2 documentation at Knowledge Center, which is available at:

IBM Knowledge Center - FlashSystem 9100 documentation

Important, IBM FlashSystem 9200: On 11th February 2020 IBM announced the arrival of the IBM FlashSystem 9200 to the family. This book was written specifically for IBM FlashSystem 9100, however most of the general principles will apply to the IBM FlashSystem 9200.

If you are in any doubt as to their applicability to the FlashSystem 9200 then you should work with your local IBM representative.

This book will be updated to include FlashSystem 9200 in due course.

The Flashsystem 9200 product guide is available at:

IBM FlashSystem 9200 Product Guide

This chapter includes the following sections:

- Configuration guidelines
- N-Port ID Virtualization
- Host pathing
- I/O queues
- Host clustering and reserves

- AIX hosts
- Virtual I/O Server
- Windows hosts
- Linux hosts
- Solaris hosts
- VMware server

7.1 Configuration guidelines

When IBM FlashSystem 9100 is used to provide storage to any host, you must follow basic configuration guidelines. These guidelines pertain to these considerations:

- The number of paths through the fabric that are allocated to the host
- The number of host ports to use
- Logical unit number (LUN) mapping
- The correct size of virtual disks (volumes) to use

7.1.1 Host levels and host object name

Before attaching a new host, confirm in SSIC that the host operating system, drivers, firmware and HBAs are supported by the storage. See *IBM System Storage Interoperation Center (SSIC)*, available at the following website:

`IBM SSIC`

When you are creating the host, use the host name from the host as the host object name in IBM FlashSystem 9100 to aid in configuration updates or problem determination in the future.

7.1.2 Host cluster

IBM FlashSystem 9100 runs IBM Spectrum Virtualize software, which supports host clusters. The host cluster allows a user to create a group of hosts to form a cluster, which is treated as one single entity. This technique allows multiple hosts to have access to the same set of volumes.

Volumes that are mapped to that host cluster are assigned to all members of the host cluster with the same SCSI ID.

A typical use-case is to define a host cluster that contains all the WWPNs belonging to the hosts participating in a host operating system based cluster, such as IBM PowerHA®, Microsoft Cluster Server (MSCS) or VMware ESXi clusters.

The following commands can be used to deal with host clusters:

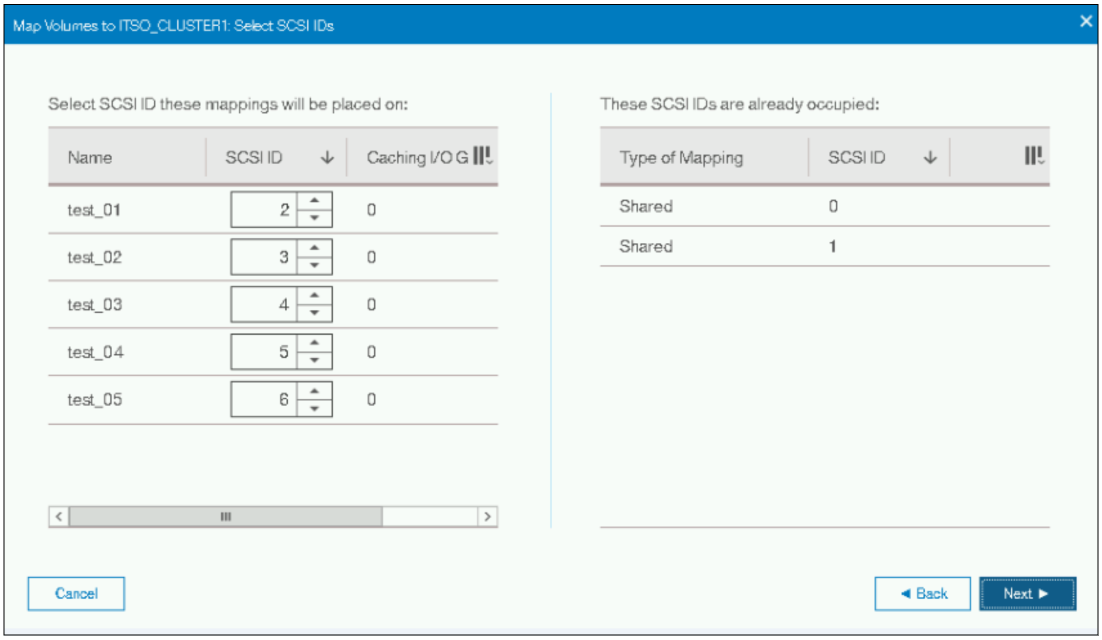
- **`lshostcluster`**
- **`lshostclustermember`**
- **`lshostclustervolumemap`**
- **`addhostclustermember`**
- **`chhostcluster`**
- **`mkhost`** (with parameter -**`hostcluster`** to create the host in one existing cluster)
- **`rmhostclustermember`**
- **`rmhostcluster`**
- **`rmvolumehostclustermap`**

Starting with IBM Spectrum Virtualize 8.1, host clusters can be added by using the GUI. For more information, see *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize V8.2.1*, SG24-7933.

For GUI users, when you use the *Map to Host or Host Cluster* function, it now allows you to let the system assign the SCSI ID for the volume or to manually assign the SCSI ID. For ease of management purposes, it is suggested to use separate ranges of SCSI IDs for hosts and host clusters.

For example, you can use SCSI IDs 0 - 99 to non-cluster host volumes, and above 100 for the cluster host volumes. When you choose the option **System Assign** the system automatically assigns the SCSI IDs starting from the first available in the sequence.

If you choose **Self Assign**, the system enables you to select the SCSI IDs manually for each volume, and on the right part of the screen it shows the SCSI IDs that are already used by the selected host/host cluster, as shown in Figure 7-1.



Map Volumes to ITSO_CLUSTER1: Select SCSI IDs

Select SCSI ID these mappings will be placed on:

Name	SCSI ID	Caching I/O G
test_01	2	0
test_02	3	0
test_03	4	0
test_04	5	0
test_05	6	0

These SCSI IDs are already occupied:

Type of Mapping	SCSI ID
Shared	0
Shared	1

Figure 7-1 SCSI ID assignment on volume mappings

Note: Although extra care is always recommended when dealing with hosts, IBM Spectrum Virtualize does not allow you to join a host into a host cluster if it already has a volume mapping with a SCSI ID that also exists in the host cluster, as shown below:

```
IBM_FlashSystem:ITSO-FS9100:superuser>addhostclustermember -host ITSO_HOST3
ITSO_CLUSTER1

CMMVC9068E Hosts in the host cluster have conflicting SCSI ID's for their
private mappings.

IBM_FlashSystem:ITSO-FS9100:superuser>
```

7.1.3 Number of paths

Based on our general experience, it is generally recommended that the total number of paths from each host to the IBM FlashSystem 9100 be limited to four paths, even though the maximum supported is eight paths. This solves many issues with high port fan-outs, fabric state changes, and host memory management, and it improves performance.

For more information about maximum host configurations and restrictions, see *V8.2.0.x Configuration Limits and Restrictions for IBM FlashSystem 9100 family* which is available at:

FlashSystem 9100 V8.2.1 Configuration Limits

The most important reason to limit the number of paths that are available to a host from IBM Spectrum Virtualize is for error recovery, failover, and failback purposes. The overall time for handling errors by a host is reduced. In addition, resources within the host are greatly reduced when you remove a path from the multipathing management.

Two path configurations have only one path to each node, which is a supported configuration but not preferred for most configurations. For IBM FlashSystem 9100 V8.2, this information is now consolidated into the IBM FlashSystem 9100 Knowledge Center, which is available at:

IBM FlashSystem 9100 Knowledge Center

7.1.4 Host ports

When you are using host ports that are connected to IBM Spectrum Virtualize, limit the number of physical ports to two ports on two different physical adapters. Each port is zoned to one target port in each IBM Spectrum Virtualize node, which limits the number of total paths to four, preferably on separate redundant SAN fabrics.

When working with clustered hosts, the preferred practice is to use the host cluster object on IBM Spectrum Virtualize. Previously, the advice was to use a single host and register all the initiators under it. A single host object can have a maximum of 32 initiators, while a host cluster object can have 128 hosts, therefore, our recommendation is to use the host cluster, instead of a single host object.

Preferred practice: Keep Fibre Channel tape (including Virtual Tape Libraries) and Fibre Channel disks on separate HBAs. These devices have two different data patterns when operating in their optimum mode. The switching between them can cause unwanted processor usage and performance slowdown for the applications.

7.1.5 Port masking

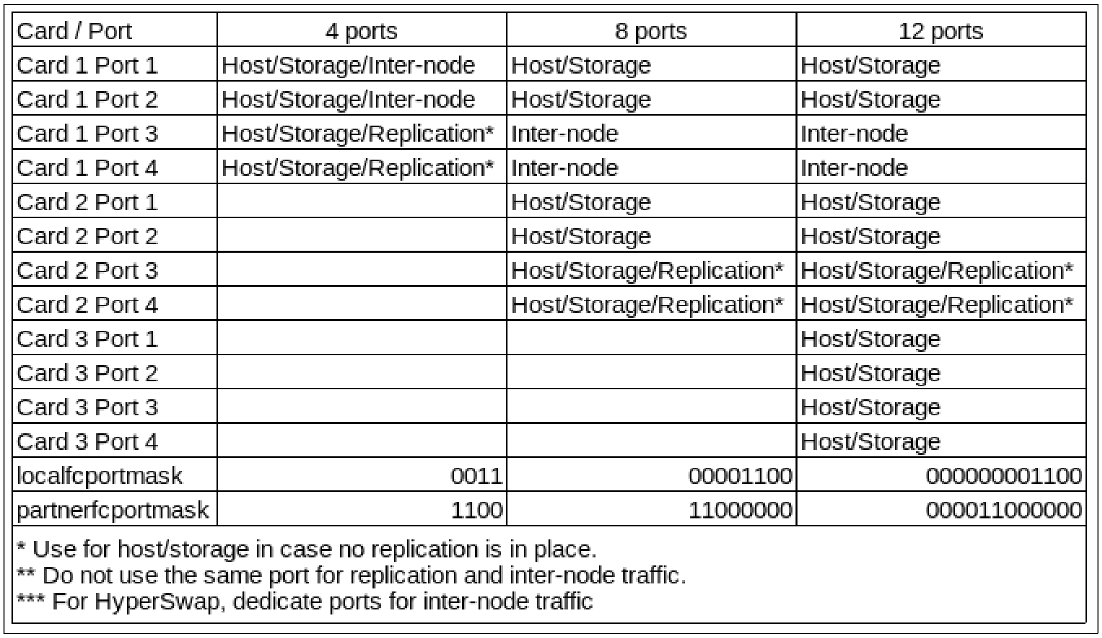
You can use a port mask to control the node target ports that a host can access. Using local FC port masking, you can set which ports can be used for node-to-node/intracluster communication. By using remote FC port masking, you can set which ports can be used for replication communication.

Using port masking is preferable because mixed traffic of host, back-end, intracluster, and replication might cause congestion and buffer-to-buffer credit exhaustion. This kind of traffic could otherwise result in heavy degradation of performance in your IBM Spectrum Virtualize environment.

The port mask is a 64-bit field that applies to all nodes in the cluster. In the local FC port masking, you can set a port to be dedicated to node-to-node/intracluster traffic by setting a 1 to that port. Also, by using remote FC port masking, you can set which ports can be used for replication traffic by setting 1 to that port. For IBM FlashSystem 9100 with single I/O group configurations, node-to-node/intracluster dedicated ports are not needed, because the intracluster traffic is performed by the internal PCI midplane on the FS9100 canister.

If a port has a 0 in the mask, it means no traffic of that type is allowed. So, in a local FC port map, a 0 means no node-to-node traffic happens, and a 0 on the remote FC port masking means no replication traffic happens on that port. Therefore, if a port has a 0 on both local and remote FC port masking, only host/back-end traffic is allowed on it. The port mask can vary depending on the number of ports that your IBM FlashSystem 9100 has.

For an example of portmask on FlashSystem 9100, see Figure 7-2.



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 7-2 Port masking on IBM FlashSystem 9100

> **Note:** Port masking for *localfcportmask* is only required on systems with more than one I/O group. In single I/O group systems the intracluster traffic is performed on the internal PCI midplane.

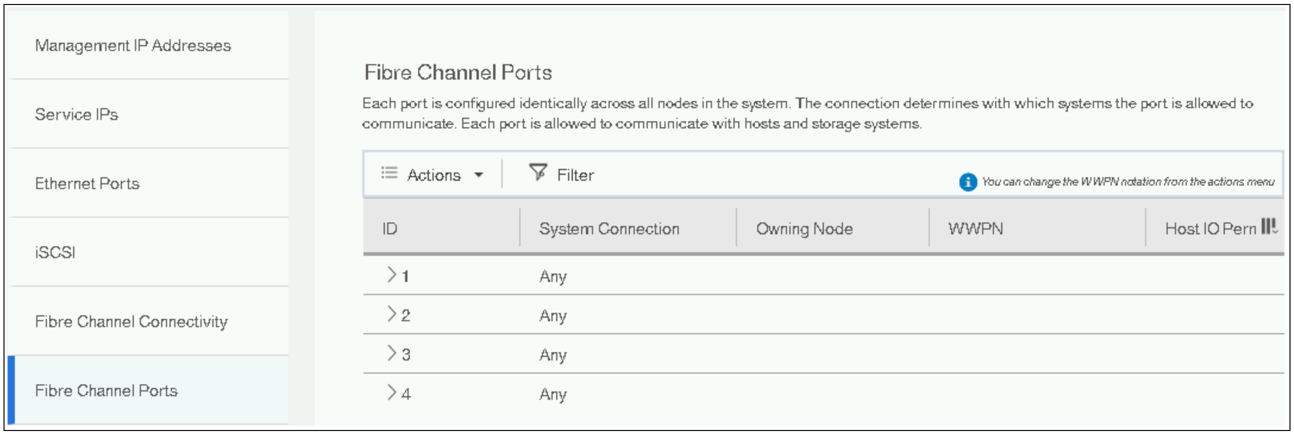
How to set a port mask using the CLI and GUI

The command to apply a local FC port mask on CLI is **`chsystem -localfcportmask mask`**. The command to apply a remote FC port mask is **`chsystem -partnerfcportmask mask`**.

If you are using the GUI, click **Settings** → **Network** → **Fibre Channel Ports**. Then you can select the use of a port from these options:

- Setting `none` means no node-to-node and no replication traffic is allowed. Only host and storage traffic is allowed.
- Setting `local` means only node-to-node traffic is allowed.
- Setting `remote` means that only replication traffic is allowed.

Figure 7-3 shows the port mask in the GUI.



Fibre Channel Ports

Each port is configured identically across all nodes in the system. The connection determines with which systems the port is allowed to communicate. Each port is allowed to communicate with hosts and storage systems.

ID	System Connection	Owning Node	WWPN	Host IO Pern
1	Any			
2	Any			
3	Any			
4	Any			

Figure 7-3 Fiber Channel Ports menu

7.1.6 Host to I/O group mapping

An *I/O group* consists of two IBM FlashSystem 9100 node canisters that share management of volumes within a cluster. Use a single I/O group (iogrp) for all volumes that are allocated to a particular host. This guideline has the following benefits:

- Minimizes port fan-outs within the SAN fabric
- Maximizes the potential host attachments to IBM Spectrum Virtualize because maximums are based on I/O groups
- Fewer target ports to manage within the host

The number of host ports and host objects that are allowed per I/O group depends on the switch fabric type. For more information about the maximum configurations, see V8.2 Configuration Limits and Restrictions:

`FlashSystem 9100 V8.2.1 Configuration Limits`

7.1.7 Volume size as opposed to quantity

In general, host resources, such as memory and processing time, are used up by each storage LUN that is mapped to the host. For each extra path, more memory can be used, and a portion of more processing time is also required. The user can control this effect by using fewer larger LUNs rather than many small LUNs.

However, you might need to tune queue depths and I/O buffers to support controlling the memory and processing time efficiently. If a host does not have tunable parameters, such as on the Windows operating system, the host does not benefit from large volume sizes. AIX greatly benefits from larger volumes with a smaller number of volumes and paths that are presented to it.

7.1.8 Host volume mapping

Host Mapping is the process of controlling which hosts or host clusters have access to specific volumes within the system. IBM FlashSystem 9100 always present a specific volume with the same Small Computer System Interface ID (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 7-1 shows the syntax of the **lshostvdiskmap** command that is used to determine the SCSI ID and the UID of volumes.

Example 7-1 The lshostvdiskmap command

```
svcinfo lshostvdiskmap -delim
```

Example 7-2 shows the results of using the **lshostvdiskmap** command.

Example 7-2 Output of using the lshostvdiskmap command

```
svcinfo lsvdiskhostmap -delim : EEXCLS_HBin01
id:name:SCSI_id:host_id:host_name:wwpn:vdisk_UID
950:EEXCLS_HBin01:14:109:HDMCENTEX1N1:10000000C938CFDF:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:109:HDMCENTEX1N1:10000000C938D01F:600507680191011D4800000000000466
950:EEXCLS_HBin01:13:110:HDMCENTEX1N2:10000000C938D65B:600507680191011D4800000000000466
950:EEXCLS_HBin01:13:110:HDMCENTEX1N2:10000000C938D3D3:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:111:HDMCENTEX1N3:10000000C938D615:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:111:HDMCENTEX1N3:10000000C938D612:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:112:HDMCENTEX1N4:10000000C938CFBD:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:112:HDMCENTEX1N4:10000000C938CE29:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:113:HDMCENTEX1N5:10000000C92EE1D8:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:113:HDMCENTEX1N5:10000000C92EDFFE:600507680191011D4800000000000466
```

Note: Example 7-2 shows the same volume mapped to five different hosts, but host 110 has a different SCSI ID from the other four hosts. This is an example of a non-recommended practice which can lead to loss of access in some situations due to SCSI ID mismatch.

In this example, `VDisk 10` has a unique device identifier (UID, which is represented by the UID field) of `6005076801958001500000000000000A` (see Example 7-3), but the SCSI_ id that host2 uses for access is `0`.

Example 7-3 VDisk 10 with a UID

```
id:name:SCSI_id:vdisk_id:vdisk_name:wwpn:vdisk_UID
2:host2:0:10:vdisk10:0000000000000ACA:6005076801958001500000000000000A
2:host2:1:11:vdisk11:0000000000000ACA:6005076801958001500000000000000B
2:host2:2:12:vdisk12:0000000000000ACA:6005076801958001500000000000000C
```

```
2:host2:3:13:vdisk13:0000000000000ACA:6005076801958001500000000000000D
2:host2:4:14:vdisk14:0000000000000ACA:6005076801958001500000000000000E
```

If you are using IBM multipathing software (Subsystem Device Driver Device Specific Module (SDDDSM)), the **datapath query device** command shows the `vdisk_UID` (unique identifier), which enables easier management of volumes. The equivalent command for Subsystem Device Driver Path Control Module (SDDPCM) is the **pcmpath query device** command.

Host mapping from more than one I/O group

The SCSI ID field in the host mapping might not be unique for a volume for a host because it does not completely define the uniqueness of the LUN. The target port is also used as part of the identification. If two I/O groups of volumes are assigned to a host port, one set starts with SCSI ID 0 and then increments (by default). The SCSI ID for the second I/O group also starts at zero and then increments by default.

Example 7-4 shows the **datapath query device** output of this Windows host. The order of the volumes of the two I/O groups is reversed from the hostmap. Volume `s-1-8-2` is first, followed by the rest of the LUNs from the second I/O group, then volume `s-0-6-4`, and the rest of the LUNs from the first I/O group. Most likely, Windows discovered the second set of LUNS first. However, the relative order within an I/O group is maintained.

Example 7-4 Using datapath query device for the hostmap

```
C:\Program Files\IBM\Subsystem Device Driver>datapath query device

Total Devices : 12

DEV#:   0  DEVICE NAME: Disk1 Part0  TYPE: 2145       POLICY: OPTIMIZED
SERIAL: 60050768018101BF28000000000000B5
============================================================================
Path#            Adapter/Hard Disk        State  Mode       Select     Errors
    0     Scsi Port2 Bus0/Disk1 Part0     OPEN   NORMAL          0          0
    1     Scsi Port2 Bus0/Disk1 Part0     OPEN   NORMAL       1342          0
    2     Scsi Port3 Bus0/Disk1 Part0     OPEN   NORMAL          0          0
    3     Scsi Port3 Bus0/Disk1 Part0     OPEN   NORMAL       1444          0

DEV#:   1  DEVICE NAME: Disk2 Part0  TYPE: 2145       POLICY: OPTIMIZED
SERIAL: 60050768018101BF28000000000000B1
============================================================================
Path#            Adapter/Hard Disk        State  Mode       Select     Errors
    0     Scsi Port2 Bus0/Disk2 Part0     OPEN   NORMAL       1405          0
    1     Scsi Port2 Bus0/Disk2 Part0     OPEN   NORMAL          0          0
    2     Scsi Port3 Bus0/Disk2 Part0     OPEN   NORMAL       1387          0
    3     Scsi Port3 Bus0/Disk2 Part0     OPEN   NORMAL          0          0

DEV#:   2  DEVICE NAME: Disk3 Part0  TYPE: 2145       POLICY: OPTIMIZED
SERIAL: 60050768018101BF28000000000000B2
============================================================================
Path#            Adapter/Hard Disk        State  Mode       Select     Errors
    0     Scsi Port2 Bus0/Disk3 Part0     OPEN   NORMAL       1398          0
    1     Scsi Port2 Bus0/Disk3 Part0     OPEN   NORMAL          0          0
    2     Scsi Port3 Bus0/Disk3 Part0     OPEN   NORMAL       1407          0
    3     Scsi Port3 Bus0/Disk3 Part0     OPEN   NORMAL          0          0

DEV#:   3  DEVICE NAME: Disk4 Part0  TYPE: 2145       POLICY: OPTIMIZED
```

```
SERIAL: 60050768018101BF28000000000000B3
============================================================================
Path#            Adapter/Hard Disk        State  Mode       Select     Errors
    0     Scsi Port2 Bus0/Disk4 Part0     OPEN   NORMAL       1504          0
    1     Scsi Port2 Bus0/Disk4 Part0     OPEN   NORMAL          0          0
    2     Scsi Port3 Bus0/Disk4 Part0     OPEN   NORMAL       1281          0
    3     Scsi Port3 Bus0/Disk4 Part0     OPEN   NORMAL          0          0

DEV#:   4  DEVICE NAME: Disk5 Part0  TYPE: 2145       POLICY: OPTIMIZED
SERIAL: 60050768018101BF28000000000000B4
============================================================================
Path#            Adapter/Hard Disk        State  Mode       Select     Errors
    0     Scsi Port2 Bus0/Disk5 Part0     OPEN   NORMAL          0          0
    1     Scsi Port2 Bus0/Disk5 Part0     OPEN   NORMAL       1399          0
    2     Scsi Port3 Bus0/Disk5 Part0     OPEN   NORMAL          0          0
    3     Scsi Port3 Bus0/Disk5 Part0     OPEN   NORMAL       1391          0

DEV#:   5  DEVICE NAME: Disk6 Part0  TYPE: 2145       POLICY: OPTIMIZED
SERIAL: 60050768018101BF28000000000000A8
============================================================================
Path#            Adapter/Hard Disk        State  Mode       Select     Errors
    0     Scsi Port2 Bus0/Disk6 Part0     OPEN   NORMAL       1400          0
    1     Scsi Port2 Bus0/Disk6 Part0     OPEN   NORMAL          0          0
    2     Scsi Port3 Bus0/Disk6 Part0     OPEN   NORMAL       1390          0
    3     Scsi Port3 Bus0/Disk6 Part0     OPEN   NORMAL          0          0

DEV#:   6  DEVICE NAME: Disk7 Part0  TYPE: 2145       POLICY: OPTIMIZED
SERIAL: 60050768018101BF28000000000000A9
============================================================================
Path#            Adapter/Hard Disk        State  Mode       Select     Errors
    0     Scsi Port2 Bus0/Disk7 Part0     OPEN   NORMAL       1379          0
    1     Scsi Port2 Bus0/Disk7 Part0     OPEN   NORMAL          0          0
    2     Scsi Port3 Bus0/Disk7 Part0     OPEN   NORMAL       1412          0
    3     Scsi Port3 Bus0/Disk7 Part0     OPEN   NORMAL          0          0

DEV#:   7  DEVICE NAME: Disk8 Part0  TYPE: 2145       POLICY: OPTIMIZED
SERIAL: 60050768018101BF28000000000000AA
============================================================================
Path#            Adapter/Hard Disk        State  Mode       Select     Errors
    0     Scsi Port2 Bus0/Disk8 Part0     OPEN   NORMAL          0          0
    1     Scsi Port2 Bus0/Disk8 Part0     OPEN   NORMAL       1417          0
    2     Scsi Port3 Bus0/Disk8 Part0     OPEN   NORMAL          0          0
    3     Scsi Port3 Bus0/Disk8 Part0     OPEN   NORMAL       1381          0

DEV#:   8  DEVICE NAME: Disk9 Part0  TYPE: 2145       POLICY: OPTIMIZED
SERIAL: 60050768018101BF28000000000000AB
============================================================================
Path#            Adapter/Hard Disk        State  Mode       Select     Errors
    0     Scsi Port2 Bus0/Disk9 Part0     OPEN   NORMAL          0          0
    1     Scsi Port2 Bus0/Disk9 Part0     OPEN   NORMAL       1388          0
    2     Scsi Port3 Bus0/Disk9 Part0     OPEN   NORMAL          0          0
    3     Scsi Port3 Bus0/Disk9 Part0     OPEN   NORMAL       1413          0

DEV#:   9  DEVICE NAME: Disk10 Part0  TYPE: 2145       POLICY: OPTIMIZED
SERIAL: 60050768018101BF28000000000000A7
```

```
=============================================================================
Path#            Adapter/Hard Disk        State  Mode       Select     Errors
    0    Scsi Port2 Bus0/Disk10 Part0     OPEN   NORMAL       1293          0
    1    Scsi Port2 Bus0/Disk10 Part0     OPEN   NORMAL          0          0
    2    Scsi Port3 Bus0/Disk10 Part0     OPEN   NORMAL       1477          0
    3    Scsi Port3 Bus0/Disk10 Part0     OPEN   NORMAL          0          0

DEV#:  10  DEVICE NAME: Disk11 Part0  TYPE: 2145       POLICY: OPTIMIZED
SERIAL: 60050768018101BF28000000000000B9
=============================================================================
Path#            Adapter/Hard Disk        State  Mode       Select     Errors
    0    Scsi Port2 Bus0/Disk11 Part0     OPEN   NORMAL          0          0
    1    Scsi Port2 Bus0/Disk11 Part0     OPEN   NORMAL      59981          0
    2    Scsi Port3 Bus0/Disk11 Part0     OPEN   NORMAL          0          0
    3    Scsi Port3 Bus0/Disk11 Part0     OPEN   NORMAL      60179          0

DEV#:  11  DEVICE NAME: Disk12 Part0  TYPE: 2145       POLICY: OPTIMIZED
SERIAL: 60050768018101BF28000000000000BA
=============================================================================
Path#            Adapter/Hard Disk        State  Mode       Select     Errors
    0    Scsi Port2 Bus0/Disk12 Part0     OPEN   NORMAL      28324          0
    1    Scsi Port2 Bus0/Disk12 Part0     OPEN   NORMAL          0          0
    2    Scsi Port3 Bus0/Disk12 Part0     OPEN   NORMAL      27111          0
    3    Scsi Port3 Bus0/Disk12 Part0     OPEN   NORMAL          0          0
```

Sometimes, a host might discover everything correctly at the initial configuration, but it does not keep up with the dynamic changes in the configuration. Therefore, the SCSI ID is important.

7.1.9 Server adapter layout

If your host system has multiple internal I/O busses, place the two adapters that are used for IBM Spectrum Virtualize cluster access on two different I/O busses to maximize the availability and performance. When purchasing a server, always have two cards instead of one. For example, have two dual port HBA cards instead of one quad port HBA card because you can spread the I/O and keep the redundancy.

7.1.10 Considerations for NVMe over Fibre Channel host attachments

At the time of this writing, Spectrum Virtualize code 8.2.1 code allows a maximum of 6 NVMe hosts, and if no other types of hosts are attached. Spectrum Virtualize code does not police these limits. If you are planning to use NVMe hosts with your IBM FlashSystem 9100, please see the following links:

IBM FlashSystem 9100 Knowledge Center - FC-NVMe limitations

IBM FlashSystem 9100 V8.2.1 Configuration Limits

IBM SSIC

7.1.11 Considerations for iSER host attachments

On IBM FlashSystem 9100, iSER (iSCSI Extensions over RDMA) hosts with different operating systems can be attached to the system. iSER is a network protocol that extends the Internet Small Computer System Interface (iSCSI) to use Remote Direct Memory Access (RDMA).

If you are planning to use iSER hosts in your IBM FlashSystem 9100, please see the following links when you are planning your environment:

`IBM FlashSystem 9100 Knowledge Center - iSER Ethernet host attachment`

`IBM FlashSystem 9100 V8.2.1 Configuration Limits`

`IBM SSIC`

7.2 N-Port ID Virtualization

The usage model for all IBM Spectrum Virtualize products is based around two-way active/active node models. That is, a pair of nodes that share active/active access for a volume. These nodes each have their own Fibre Channel WWNN, so all ports presented from each node have a set of WWPNs that are presented to the fabric.

Traditionally, if one node fails or is removed for some reason, the paths presented for volumes from that node go offline. It is up to the native OS multipathing software to fail over from using both sets of WWPN to just those that remain online.

N-Port ID Virtualization (NPIV) on IBM Spectrum Virtualize is a feature that was released in V7.7. The NPIV feature aims to provide an availability improvement for the hosts that are connected to the IBM FlashSystem 9100 node canisters. It creates a virtual WWPN that is available only for host connection. During a node assert, failure, reboot, or service mode, the virtual WWPN from that node is transferred to the other node in the iogrp, to the same port.

That process ensures that, instead of having the host lose the connection to the IBM FlashSystem 9100 node canister WWPN, the connection remains active. The multipath software does not have to handle the path failures, mitigating in this case the occurrence of problems of hosts not recovering from path failure and an alerting storm from servers, for instance, in a code upgrade situation on IBM Spectrum Virtualize.

NPIV works in a symmetric way, which means the NPIV port from node 1 port 1 has the failover on node 2 port 1. For NPIV to work properly, you must have a symmetric cabling of IBM Spectrum Virtualize in your switch, which means you must have the ports that perform the failover in the same SAN Fabric.

NPIV is available only for the hosts. The back-end storage must still be zoned to the physical WWPN address. No intracluster or replication zone is allowed on the NPIV WWPN as well, because the NPIV ports are target only, as shown in Example 7-5.

Example 7-5 NPIV ports

```
ITSO_SAN_01:root> nodefind 50:05:07:68:0c:25:45:28
Local:
 Type Pid    COS     PortName                NodeName                 SCR
 N    0a4c01;    2,3;50:05:07:68:0c:25:45:28;50:05:07:68:0c:00:45:28; 0x00000003
    FC4s: FCP
    Fabric Port Name: 20:4c:50:eb:1a:a9:8f:b8
    Permanent Port Name: 50:05:07:68:0c:21:45:28
    Device type: NPIV Target
    Port Index: 76
    Share Area: No
    Device Shared in Other AD: No
    Redirect: No
    Partial: No
    LSAN: No
    Aliases: FS9100_ITSO_LAB_N1_P3_HOST_NPIV
```

NPIV is native on new deployments from V7.8 and above. You can disable the NPIV feature at installation, although generally we recommend that you do not do this, because it is a new deployment, so no extra effort needs to take place in order for your hosts to greatly benefit from this feature.

When NPIV is enabled on IBM FlashSystem 9100, each physical WWPN reports up to five virtual WWPNs, as shown in Table 7-1.

Table 7-1 IBM Spectrum Virtualize NPIV Ports

NPIV port	Port description
Primary NPIV Port	This is the WWPN that communicates with back-end storage only.
Primary Host Attach Port	This is the WWPN that communicates with hosts. It is a target port only, and this is the primary port that represents this local nodes WWNN.
Failover Host Attach Port	This is a standby WWPN that communicates with hosts and is only brought online on this node if the partner node in this I/O Group goes offline. This is the same as the Primary Host Attach WWPN on the partner node.
Primary nvme Host Attach Port	This is the WWPN that communicates with hosts. It is a target port only. This is the primary port, so it is based on this local node’s WWNN.
Failover nvme Host Attach Port	This is a standby WWPN that communicates with hosts and is only brought online if the partner node within the I/O Group goes offline. This is the same as the Primary Host Attach WWPN of the partner node.

Then, when NPIV effectively goes into action, you can see a situation such as that illustrated in Example 7-6.

Example 7-6 NPIV failover example

```
itso-sansw01:admin> portshow 4/12
portIndex:  60
portName: slot4 port12
portHealth: HEALTHY

Authentication: None
portDisableReason: None
portCFlags: 0x1
```

```
portFlags: 0x24b03 PRESENT ACTIVE F_PORT G_PORT U_PORT NPIV LOGICAL_ONLINE LOGIN
NOELP LED ACCEPT FLOGI
LocalSwcFlags: 0x0
portType:  24.0
portState: 1Online
Protocol: FC
portPhys:  6In_Sync  portScn:   32F_Port
port generation number:    2164
state transition count:    18

portId:    0a3c00
portIfId:    4342080b
portWwn:   20:3c:50:eb:1a:a9:8f:b8
portWwn of device(s) connected:
   50:05:07:68:0c:15:45:28
   50:05:07:68:0c:13:45:28
   50:05:07:68:0c:11:45:28
Distance:  normal
portSpeed: N16Gbps

itso-sansw01:admin> nodefind 50:05:07:68:0c:15:45:28
Local:
 Type Pid    COS     PortName                NodeName                 SCR
 N    0a3c01;    2,3;50:05:07:68:0c:15:45:28;50:05:07:68:0c:00:45:28; 0x00000003
    FC4s: FCP
    Fabric Port Name: 20:3c:50:eb:1a:a9:8f:b8
    Permanent Port Name: 50:05:07:68:0c:11:45:28
    Device type: NPIV Target
    Port Index: 60
    Share Area: No
    Device Shared in Other AD: No
    Redirect: No
    Partial: No
    LSAN: No
    Aliases: ITSO_FS9100LAB01_NODE1_NP1
itso-sansw01:admin>

Then we took the node offline

itso-sansw01:admin> nodefind 50:05:07:68:0c:15:45:28
Local:
 Type Pid    COS     PortName                NodeName                 SCR
 N    0a2502;    2,3;50:05:07:68:0c:15:45:28;50:05:07:68:0c:00:45:28; 0x00000003
    FC4s: FCP
    Fabric Port Name: 20:25:50:eb:1a:a9:8f:b8
    Permanent Port Name: 50:05:07:68:0c:11:46:fc
    Device type: NPIV Target
    Port Index: 37
    Share Area: No
    Device Shared in Other AD: No
    Redirect: No
    Partial: No
    LSAN: No
    Aliases: ITSO_FS9100LAB01_NODE1_NP1
itso-sansw01:admin>
```

```
itso-sansw01:admin> portshow 3/5
portIndex:  37
portName: slot3 port5
portHealth: HEALTHY

Authentication: None
portDisableReason: None
portCFlags: 0x1
portFlags: 0x24b03 PRESENT ACTIVE F_PORT G_PORT U_PORT NPIV LOGICAL_ONLINE LOGIN
NOELP LED ACCEPT FLOGI
LocalSwcFlags: 0x0
portType:  24.0
portState: 1Online
Protocol: FC
portPhys:  6In_Sync  portScn:   32F_Port
port generation number:    2130
state transition count:    14

portId:    0a2500
portIfId:    4332001a
portWwn:   20:25:50:eb:1a:a9:8f:b8
portWwn of device(s) connected:
   50:05:07:68:0c:15:46:fc
   50:05:07:68:0c:13:46:fc
   50:05:07:68:0c:11:46:fc
   50:05:07:68:0c:15:45:28
   50:05:07:68:0c:13:45:28
```

7.3 Host pathing

Each host mapping associates a volume with a host object and allows all HBA ports on the host object to access the volume. You can map a volume to multiple host objects.

When a mapping is created, multiple paths will normally exist across the SAN fabric from the hosts to the IBM FlashSystem 9100 nodes that present the volume. Most operating systems present each path to a volume as a separate storage device. Therefore IBM Spectrum Virtualize requires that multipathing software runs on the host. The multipathing software manages the many paths that are available to the volume and presents a single storage device to the operating system and provide failover in the case of a lost path. If your IBM Spectrum Virtualize system uses NPIV, the multipathing driver will not need to do the failover.

7.3.1 Multipathing software

IBM Spectrum Virtualize requires the use of multipathing software on hosts that are connected. For the latest levels for each host operating system and multipathing software package, see *IBM System Storage Interoperation Center (SSIC)*:

IBM SSIC

7.3.2 Preferred path algorithm

I/O traffic for a particular volume is managed exclusively by the nodes in a single I/O group. Although both nodes in the I/O group can handle the traffic for the volume, the system prefers to use a consistent node to improve performance, this is the *preferred node*.

When a volume is created, an I/O group and preferred node are defined and can optionally be set by the administrator. The owner node for a volume is the preferred node when both nodes are available.

IBM FlashSystem 9100 is a storage subsystem ALUA compliant. Most of the modern multipath software is also ALUA compliant, it means the server administrator should not be concerned with anything but installing the multipath software. Restrictions might apply for specific applications, which requires specific multipath settings.

7.3.3 Path selection

IBM FlashSystem 9100 is storage subsystem ALUA compliant. As most modern multipath software is ALUA compliant, it means the multipath software applies the load balance multipath policy only in the paths that belong to the preferred node, providing a better performance.

When a read or write I/O comes through a non-preferred node, FlashSystem 9100 sends the data by using the intracluster/node-to-node traffic. This process enables the operation to be run by the preferred node.

7.3.4 Non-disruptive volume migration between I/O groups

> **Attention:** These migration tasks can be non-disruptive if they are performed correctly and hosts that are mapped to the volume support non-disruptive volume move. The cached data that is held within the system must first be written to disk before the allocation of the volume can be changed.

Modifying the I/O group that services the volume can be done concurrently with I/O operations if the host supports non-disruptive volume move. It also requires a rescan at the host level to ensure that the multipathing driver is notified that the allocation of the preferred node changed and the ports by which the volume is accessed changed. This process can be performed in the situation where one pair of nodes becomes over-used.

If there are any host mappings for the volume, the hosts must be members of the target I/O group or the migration fails. Make sure that you create paths to I/O groups on the host system. After the system successfully adds the new I/O group to the volume's access set and you move the selected volumes to another I/O group, detect the new paths to the volumes on the host.

The commands and actions on the host vary depending on the type of host and the connection method that is used. This process must be completed on all hosts to which the selected volumes are currently mapped.

You can also use the management GUI to move volumes between I/O groups non-disruptively. In the management GUI, click **Volumes** → **Volumes**. In the Volumes window, select the volume that you want to move and click **Actions** → **Move to Another I/O Group**.

The wizard guides you through the steps for moving a volume to another I/O group, including any changes to hosts that are required. For more information, click **Need Help** in the associated management GUI windows.

Additional information regarding volume migrations between I/O groups can be found at IBM Knowledge Center:

IBM FlashSystem 9100 Knowledge Center - Managing Volumes

7.4 I/O queues

Host operating system and host bus adapter software must have a way to fairly prioritize I/O to the storage. The host bus might run faster than the I/O bus or external storage. Therefore, you must have a way to queue I/O to the devices. Each operating system and host adapter have unique methods to control the I/O queue. The unique method to control I/O queue can be host adapter-based or memory and thread resources-based, or based on the number of commands that are outstanding for a device.

You have several configuration parameters available to control the I/O queue for your configuration. The storage adapters (volumes on IBM FlashSystem 9100) have host adapter parameters and queue depth parameters. Algorithms are also available within multipathing software, such as the `qdepth_enable` attribute.

7.4.1 Queue depths

Queue depth is used to control the number of concurrent operations that occur on different storage resources. Queue depth is the number of I/O operations that can be run in parallel on a device.

Guidance about limiting queue depths in large SANs as described in previous IBM documentation was replaced with a calculation for homogeneous and nonhomogeneous FC, iSCSI and iSER hosts. This calculation is for an overall queue depth per I/O group. You can use this number to reduce queue depths that are lower than the recommendations or defaults for individual host adapters.

For more information, see Chapter 3, “Drives and arrays” on page 41and *Queue depth in Fibre Channel hosts* topic in the V8.2 Documentation site:

- FlashSystem 9100 FC hosts Queue Depth

  IBM FlashSystem 9100 Knowledge Center - FC Queue Depth
- FlashSystem 9100 iSCSI hosts Queue Depth

  IBM FlashSystem 9100 Knowledge Center - iSCSI Queue Depth
- FlashSystem 9100 iSER hosts Queue Depth

  IBM FlashSystem 9100 Knowledge Center - iSER Queue Depth

7.5 Host clustering and reserves

To prevent hosts from sharing storage inadvertently, establish a storage reservation mechanism. The mechanisms for restricting access to IBM Spectrum Virtualize volumes use the SCSI-3 persistent reserve commands or the SCSI-2 reserve and release commands.

The host software uses several methods to implement host clusters. These methods require sharing the volumes on IBM Spectrum Virtualize between hosts. To share storage between hosts, maintain control over accessing the volumes. Some clustering software use software locking methods.

You can choose other methods of control by the clustering software or by the device drivers to use the SCSI architecture reserve or release mechanisms. The multipathing software can change the type of reserve that is used from an earlier reserve to persistent reserve, or remove the reserve.

Persistent reserve refers to a set of SCSI-3 standard commands and command options that provide SCSI initiators with the ability to establish, preempt, query, and reset a reservation policy with a specified target device. The functions that are provided by the persistent reserve commands are a superset of the original reserve or release commands.

The persistent reserve commands are incompatible with the earlier reserve or release mechanism. Also, target devices can support only reservations from the earlier mechanism or the new mechanism. Attempting to mix persistent reserve commands with earlier reserve or release commands results in the target device returning a reservation conflict error.

Earlier reserve and release mechanisms (SCSI-2) reserved the entire LUN (volume) for exclusive use down a single path. This approach prevents access from any other host or even access from the same host that uses a different host adapter. The persistent reserve design establishes a method and interface through a reserve policy attribute for SCSI disks. This design specifies the type of reservation (if any) that the operating system device driver establishes before it accesses data on the disk.

The following possible values are supported for the reserve policy:

- No_reserve: No reservations are used on the disk.
- Single_path: Earlier reserve or release commands are used on the disk.
- PR_exclusive: Persistent reservation is used to establish *exclusive host access* to the disk.
- PR_shared: Persistent reservation is used to establish *shared host access* to the disk.

When a device is opened (for example, when the AIX **varyonvg** command opens the underlying hdisks), the device driver checks the object data manager (ODM) for a reserve_policy and a PR_key_value. The driver then opens the device. For persistent reserve, each host that is attached to the shared disk must use a unique registration key value.

7.5.1 Clearing reserves

It is possible to accidentally leave a reserve on the IBM Spectrum Virtualize volume or on the IBM Spectrum Virtualize MDisk during migration into IBM Spectrum Virtualize or when disks are reused for another purpose. Several tools are available from the hosts to clear these reserves. The easiest tools to use are the **pcmquerypr** (AIX SDDPCM host) commands. Another tool is a menu-driven Windows SDDDSM tool.

The Windows Persistent Reserve Tool is called `PRTool.exe` and is installed automatically when SDDDSM is installed in the `C:\Program Files\IBM\Subsystem Device Driver\PRTool.exe` directory. You can clear the IBM Spectrum Virtualize volume reserves by removing all the host mappings.

Example 7-7 shows a failing **pcmquerypr** command to clear the reserve and the error.

Example 7-7 Output of the pcmquerypr command

```
# pcmquerypr -ph /dev/hdisk232 -V
connection type: fscsi0
open dev: /dev/hdisk232
couldn't open /dev/hdisk232, errno=16
```

Use the AIX `errno.h` include file to determine what error number 16 indicates. This error indicates a busy condition, which can indicate a legacy reserve or a persistent reserve from another host (or that this host is from a different adapter). However, some AIX technology levels have a diagnostic open issue that prevents the **pcmquerypr** command from opening the device to display the status or to clear a reserve.

For more information about older AIX technology levels that break the **pcmquerypr** command, see *IBM Support*, which is available at this website:

IBM Support

7.5.2 IBM Spectrum Virtualize MDisk reserves

There are instances in which a host image mode migration appears to succeed, but problems occur when the volume is opened for read or write I/O. The problems can result from not removing the reserve on the MDisk before image mode migration is used in IBM Spectrum Virtualize. You cannot clear a leftover reserve on an IBM Spectrum Virtualize MDisk from IBM Spectrum Virtualize. You must clear the reserve by mapping the MDisk back to the owning host and clearing it through host commands, or through back-end storage commands as advised by IBM technical support.

7.6 AIX hosts

This section describes various topics that are specific to AIX.

7.6.1 AIX Multipathing

Since IBM Spectrum Virtualize version 7.6, IBM recommends the use of native AIXPCM for AIX systems instead of SDDPCM, however, SDDPCM is still supported. In case you are deploying a new AIX system, use AIXPCM. If you are migrating a existing AIX which runs SDDPCM to the FlashSystem 9100, our recommendation is to check your system and migrate from SDDPCM to AIXPCM whenever it is possible.

For more information on how to migrate SDDPCM to AIXPCM, please see the following links:

The Recommended Multi-path Driver to use on IBM AIX and VIOS:

IBM Support - Multipath Driver for IBM AIX and VIOS

How To Migrate SDDPCM to AIXPCM:

IBM Support - SDDPCM to AIXPCM migration

7.6.2 HBA parameters for performance tuning

You can use the example settings in this section to start your configuration in the specific workload environment. These settings are a guideline, and are not guaranteed to be the answer to all configurations. Always try to set up a test of your data with your configuration to see whether further tuning can help. For best results, it helps to have knowledge about your specific data I/O pattern.

The settings in the following sections can affect performance on an AIX host. These sections examine these settings in relation to how they affect the two workload types.

Transaction-based settings

The host attachment script sets the default values of attributes for IBM Spectrum Virtualize hdisks: `devices.fcp.disk.IBM.rte` or `devices.fcp.disk.IBM.mpio.rte`. You can modify these values as a starting point. In addition, you can use several HBA parameters to set higher performance or large numbers of hdisk configurations.

You can change all attribute values that are changeable by using the **chdev** command for AIX.

AIX settings that can directly affect transaction performance are the `queue_depth` hdisk attribute and `num_cmd_elem` attribute in the HBA attributes.

The queue_depth hdisk attribute

For the logical drive (which is known as the hdisk in AIX), the setting is the attribute `queue_depth`, as shown in the following example:

```
# chdev -l hdiskX -a queue_depth=Y -P
```

In this example, *X* is the hdisk number, and *Y* is the value to which you are setting *X* for `queue_depth`.

For a high-volume transaction workload of small random transfers, try a `queue_depth` value of `25` or more. For large sequential workloads, performance is better with shallow queue depths, such as a value of `4`.

The num_cmd_elem attribute

For the HBA settings, the `num_cmd_elem` attribute for the fcs device represents the number of commands that can be queued to the adapter, as shown in the following example:

```
chdev -l fcsX -a num_cmd_elems=2048 -P
```

The default value is 200, but the following maximum values can be used, depending on your HBA vendor.

The AIX settings that can directly affect throughput performance with large I/O block size are the `lg_term_dma` and `max_xfer_size` parameters for the fcs device.

Throughput-based settings

In the throughput-based environment, you might want to decrease the queue-depth setting to a smaller value than the default from the host attach. In a mixed application environment, you do not want to lower the `num_cmd_elem` setting because other logical drives might need this higher value to perform. In a purely high throughput workload, this value has no effect.

Start values: For high throughput sequential I/O environments, use the start values `lg_term_dma` = `0x400000` or `0x800000` (depending on the adapter type) and `max_xfr_size` = `0x200000`.

First, test your host with the default settings. Then, make these possible tuning changes to the host parameters to verify whether these suggested changes enhance performance for your specific host configuration and workload.

The lg_term_dma attribute

The lg_term_dma AIX Fibre Channel adapter attribute controls the direct memory access (DMA) memory resource that an adapter driver can use. The default value of `lg_term_dma` is `0x200000`, and the maximum value is `0x8000000`.

One change is to increase the value of `lg_term_dma` to `0x400000`. If you still experience poor I/O performance after changing the value to `0x400000`, you can increase the value of this attribute again. If you have a dual-port Fibre Channel adapter, the maximum value of the `lg_term_dma` attribute is divided between the two adapter ports. Therefore, never increase the value of the `lg_term_dma` attribute to the maximum value for a dual-port Fibre Channel adapter because this value causes the configuration of the second adapter port to fail.

The max_xfer_size attribute

The `max_xfer_size` AIX Fibre Channel adapter attribute controls the maximum transfer size of the Fibre Channel adapter. Its default value is `100,000`, and the maximum value is `1,000,000`. You can increase this attribute to improve performance.

Setting the `max_xfer_size` attribute affects the size of the memory area that is used for data transfer by the adapter. With the default value of `max_xfer_size=0x100000`, the area is 16 MB, and for other allowable values of the `max_xfer_size` attribute, the memory area is 128 MB.

7.7 Virtual I/O Server

Virtual SCSI is based on a client/server relationship. The VIOS owns the physical resources and acts as the server or target device. Physical adapters with attached disks (in this case, volumes on IBM Spectrum Virtualize) on the VIOS partition can be shared by one or more partitions. These partitions contain a virtual SCSI client adapter that detects these virtual devices as standard SCSI-compliant devices and LUNs.

You can create the following types of volumes on a VIOS:

- Physical volume (PV) VSCSI hdisks
- Logical volume (LV) VSCSI hdisks

PV VSCSI hdisks are entire LUNs from the VIOS perspective. If you are concerned about failure of a VIOS and have configured redundant VIOSs for that reason, you must use PV VSCSI hdisks. Therefore, PV VSCSI hdisks are entire LUNs that are volumes from the virtual I/O client perspective. An LV VSCSI hdisk cannot be served up from multiple VIOSs. LV VSCSI hdisks are in LVM volume groups on the VIOS, and cannot span PVs in that volume group or be striped LVs. Because of these restrictions, use PV VSCSI hdisks.

Multipath support for IBM Spectrum Virtualize attachment to Virtual I/O Server is provided by MPIO with SDDPCM and AIXPCM. Where Virtual I/O Server SAN Boot or dual Virtual I/O Server configurations are required, see *IBM System Storage Interoperation Center (SSIC)*:

`IBM SSIC`

For more information about VIOS, see this website:

`IBM Knowledge Center - VIOS overview`

7.7.1 Methods to identify a disk for use as a virtual SCSI disk

The VIOS uses the following methods to uniquely identify a disk for use as a virtual SCSI disk:

- Unique device identifier (UDID)
- IEEE volume identifier
- Physical volume identifier (PVID)

Each of these methods can result in different data formats on the disk. The preferred disk identification method for volumes is the use of UDIDs.

7.8 Windows hosts

To release new enhancements more quickly, the newer hardware architectures are tested only on the SDDDSM code stream. Therefore, only SDDDSM packages are available.

For Microsoft Windows 2016, and Microsoft Windows 2012, download the latest version of SDDDSM from this website:

`IBM Support - SDDDSM`

7.8.1 Clustering and reserves

Windows SDDDSM uses the persistent reserve functions to implement Windows clustering. A stand-alone Windows host does not use reserves.

When SDDDSM is installed, the reserve and release functions are converted into the appropriate persistent reserve and release equivalents to allow load balancing and multipathing from each host.

7.8.2 Tunable parameters

With Windows operating systems, the queue-depth settings are the responsibility of the host adapters. They are configured through the BIOS setting. Configuring the queue-depth settings varies from vendor to vendor. For more information about configuring your specific cards, see “Hosts running the Microsoft Windows Server operating system” in IBM FlashSystem 9100 Knowledge Center

`IBM FlashSystem 9100 Knowledge Center - Windows Hosts`

Queue depth is also controlled by the Windows application program. The application program controls the number of I/O commands that it allows to be outstanding before waiting for completion. You might have to adjust the queue depth that is based on the overall I/O group queue depth calculation, as described in Chapter 5, “Volumes” on page 101.

7.9 Linux hosts

IBM Spectrum Virtualize multipathing supports Linux native DM-MPIO multipathing. Veritas DMP is also available for certain kernels.

For more information about which versions of each Linux kernel require DM-MPIO support, see *IBM System Storage Interoperation Center (SSIC*) available:

`IBM SSIC`

Certain types of clustering are now supported. However, the multipathing software choice is tied to the type of cluster and HBA driver. For example, Veritas Storage Foundation is supported for certain hardware and kernel combinations, but it also requires Veritas DMP multipathing. Contact IBM marketing for SCORE/RPQ support if you need Linux clustering in your specific environment and it is not listed.

For further reference on RHEL 6 and 7 operating systems, you can check the following sites:

- Red Hat Enterprise Linux 6 DM Multipath Configuration and Administration

  `Red Hat RHEL 6 documentation`

- Red Hat Enterprise Linux 7 DM Multipath Configuration and Administration

  `Red Hat RHEL 7 documentation`

7.9.1 Tunable parameters

Linux performance is influenced by HBA parameter settings and queue depth. The overall calculation for queue depth for the I/O group is described in Chapter 3, “Drives and arrays” on page 41. In addition, the IBM FlashSystem 9100 Knowledge Center provides maximums per HBA adapter or type. For more information, see:

`IBM FlashSystem 9100 Knowledge Center - Host Attachment`

For more information about the settings for each specific HBA type and general Linux OS tunable parameters, see the *Attaching to a host running the Linux operating system* topic in the IBM Spectrum Virtualize IBM Knowledge Center:

`IBM FlashSystem 9100 Knowledge Center - Linux Hosts`

In addition to the I/O and operating system parameters, Linux has tunable file system parameters. You can use the **`tune2fs`** command to increase file system performance that is based on your specific configuration. You can change the journal mode and size, and index the directories. For more information, see “Learn Linux, 101: Maintain the integrity of filesystems” in IBM developerWorks® at this website:

`https://www.ibm.com/developerworks/library/l-lpic1-104-2/`

7.10 Solaris hosts

Two options are available for multipathing support on Solaris hosts: Symantec Veritas Volume Manager and Solaris MPxIO. The option that you choose depends on your file system requirements and the operating system levels in the latest interoperability matrix. For more information, see *IBM System Storage Interoperation Center (SSIC)*:

`IBM SSIC`

IBM SDD is no longer supported because its features are now available natively in the multipathing driver for Solaris MPxIO. If SDD support is still needed, contact your IBM marketing representative to request an RPQ for your specific configuration.

From Solaris 10 and later, Oracle has released a combined file system and logical volume manager called ZFS, designed by Sun Microsystems. It uses MPxIO and is inbound to the Solaris 11, being a native option to Veritas Volume Manager. For more information about Oracle ZFS, see the following links:

http://www.oracle.com/technetwork/systems/hands-on-labs/s11-intro-zfs-1408637.html
https://docs.oracle.com/cd/E19253-01/819-5461/819-5461.pdf
https://docs.oracle.com/cd/E23824_01/pdf/821-1448.pdf

For further reference on host attachment for Solaris hosts, check *IBM Knowledge Center* at the following website:

IBM FlashSystem 9100 Knowledge Center - Solaris Hosts

7.10.1 Solaris MPxIO

SAN boot and clustering support is available for V5.9, V5.10, and 5.11, depending on the multipathing driver and HBA choices. Support for load balancing of the MPxIO software is included in IBM FlashSystem 9100 running Spectrum Virtualize 8.2. If you want to run MPxIO on your Sun SPARC host, configure your IBM Spectrum Virtualize host object with the type attribute set to `tpgs`, as shown in the following example:

```
svctask mkhost -name new_name_arg -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` option enables an extra target port unit. The default is `generic`.

7.10.2 Symantec Veritas Volume Manager

When you are managing IBM Spectrum Virtualize storage in Symantec volume manager products, you must install an ASL on the host so that the volume manager is aware of the storage subsystem properties (active/active or active/passive). If the appropriate Array Support Library (ASL) is not installed, the volume manager did not claim the LUNs. Usage of the ASL is required to enable the special failover or failback multipathing that IBM Spectrum Virtualize requires for error recovery. Use the commands that are shown in Example 7-8 to determine the basic configuration of a Symantec Veritas server.

Example 7-8 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 7-9 and Example 7-10 determine whether the IBM Spectrum Virtualize is properly connected. They show at a glance which ASL is used (native DMP ASL or SDD ASL).

Example 7-9 shows what you see when Symantec Volume Manager correctly accesses IBM Spectrum Virtualize by using the SDD pass-through mode ASL.

Example 7-9 Symantec Volume Manager using SDD pass-through mode ASL

```
# vxdmpadm list enclosure all
ENCLR_NAME ENCLR_TYPE ENCLR_SNO STATUS
============================================================
OTHER_DISKS OTHER_DISKS OTHER_DISKS CONNECTED
VPATH_SANVC0 VPATH_SANVC 0200628002faXX00 CONNECTED
```

Example 7-10 shows what you see when IBM Spectrum Virtualize is configured by using native DMP ASL.

Example 7-10 IBM Spectrum Virtualize that is configured by using native ASL

```
# vxdmpadm listenclosure all
ENCLR_NAME ENCLR_TYPE ENCLR_SNO STATUS
============================================================
OTHER_DISKS OTHER_DSKSI OTHER_DISKS CONNECTED
SAN_VC0 SAN_VC 0200628002faXX00 CONNECTED
```

7.10.3 DMP multipathing

For the latest ASL levels to use native DMP, see the array-specific module table at this website:

https://sort.symantec.com/asl

For the latest Veritas Patch levels, see the patch table at this website:

https://sort.symantec.com/patch/matrix

To check the installed Symantec Veritas version, enter the following command:

```
showrev -p |grep vxvm
```

To check which IBM ASLs are configured into the Volume Manager, enter the following command:

```
vxddladm listsupport |grep -i ibm
```

After you install a new ASL by using the **pkgadd** command, restart your system or run the **vxdctl enable** command. To list the ASLs that are active, enter the following command:

```
vxddladm listsupport
```

7.11 VMware server

To determine the various VMware ESXi levels that are supported, see the IBM System Storage Interoperation Center (SSIC) available at:

IBM SSIC

On this website you can also find information about the available support in FlashSystem 9100 for VMware vStorage APIs for Array Integration (VAAI).

IBM FlashSystem 9100 supports VMware vStorage APIs. IBM Spectrum Virtualize implemented new storage-related tasks that were previously performed by VMware, which helps improve efficiency and frees server resources for more mission-critical tasks. The new functions include full copy, block zeroing, and hardware-assisted locking.

The minimum supported VMware level is V6.0. If earlier versions are required, contact your IBM marketing representative and ask about the submission of an RPQ for support. The required patches and procedures are supplied after the specific configuration is reviewed and approved.

For more information about host attachment recommendations, see the *Attachment requirements for hosts running VMware operating systems* topic in the *IBM FlashSystem 9100 Knowledge Center* at:

`IBM FlashSystem 9100 Knowledge Center - VMware Hosts`

7.11.1 Multipathing solutions supported

Multipathing is supported at VMware ESX level 2.5.x and later. Therefore, installing multipathing software is not required. The following multipathing algorithms are available on Native Multipathing (NMP):

- Fixed-path
- Round-robin
- Most recently used (MRU)

VMware multipathing was improved to use the IBM Spectrum Virtualize preferred node algorithms starting with V4.0. Preferred paths are ignored in VMware versions before V4.0. VMware multipathing software performs static load balancing for I/O, which defines the fixed path for a volume.

The round-robin algorithm rotates path selection for a volume through all paths. For any volume that uses the fixed-path policy, the first discovered preferred node path is chosen. The VMW_PSP_MRU policy selects the first working path, discovered at system boot time. If this path becomes unavailable, the ESXi/ESX host switches to an alternative path and continues to use the new path while it is available.

All these algorithms were modified with V4.0 and later to honor the IBM Spectrum Virtualize preferred node that is discovered by using the **TPGS** command. Path failover is automatic in all cases. If the round-robin algorithm is used, path failback might not return to a preferred node path. Therefore, manually check pathing after any maintenance or problems occur.

Update: From vSphere version 5.5 and later, VMware multipath driver fully supports IBM Spectrum Virtualize/Storwize V7000 ALUA preferred path algorithms. VMware administrators should select Round Robin and validate that `VMW_SATP_ALUA` is displayed. This configuration reduces operational burden and improves cache hit rate by sending the I/O to the preferred node.

Our recommendation for NMP Round Robin configuration is to change the default value of IOs before changing a path from the default 1,000 to 1. This will better spread the workload among the paths, avoiding some high latency situations that are not visible at a SAN level.

For more information on how to change this setting, please see the following *VMware KB*:

https://kb.vmware.com/s/article/2069356

7.11.2 Multipathing configuration maximums

The VMware multipathing software supports the following maximum configuration:

- A total of 256 SCSI devices
- Up to 32 paths to each volume
- Up to 4096 paths per server

Tip: Each path to a volume equates to a single SCSI device.

See the following VMware website for a complete list of maximums:

https://configmax.vmware.com/#

8

Monitoring

Monitoring in a storage environment is crucial and it is part of what usually is called *storage governance*.

With a robust and reliable storage monitoring system, you can save significant money and minimize pain in your operation, by monitoring and predicting utilization bottlenecks in your storage environment.

This chapter provides suggestions and the basic concepts of how to implement a storage monitoring system for IBM FlashSystem 9100 using specific functions or external IBM Tools.

> **Important, IBM FlashSystem 9200:** On 11th February 2020 IBM announced the arrival of the IBM FlashSystem 9200 to the family. This book was written specifically for IBM FlashSystem 9100, however most of the general principles will apply to the IBM FlashSystem 9200.
>
> If you are in any doubt as to their applicability to the FlashSystem 9200 then you should work with your local IBM representative.
>
> This book will be updated to include FlashSystem 9200 in due course.
>
> The Flashsystem 9200 product guide is available at:
>
> IBM FlashSystem 9200 Product Guide

This chapter includes the following sections:

- Generic monitoring
- Performance monitoring
- Capacity metrics for block storage systems
- Creating Alerts for IBM Spectrum Control and IBM Storage Insights
- Error condition example with IBM Spectrum Control: FC port
- Important metrics
- Performance support package
- Metro and Global Mirror monitoring with IBM Copy Services Manager and scripts
- Monitoring Tier1 SSD

8.1 Generic monitoring

With IBM FlashSystem 9100, you can implement generic monitoring using IBM FlashSystem 9100 specific functions that are integrated with the product itself without adding any external tools or cost.

8.1.1 Monitoring with the GUI

The management GUI is the primary tool that is used to service your system. Regularly monitor the status of the system by using the management GUI. If you suspect a problem, use the management GUI first to diagnose and resolve the problem.

Use the views that are available in the management GUI to verify the status of the system, the hardware devices, the physical storage, and the available volumes. The **Monitoring** → **Events** window provides access to all problems that exist on the system. Use the **Recommended Actions** filter to display the most important events that need to be resolved.

If there is a service error code for the alert, you can run a fix procedure that assists you in resolving the problem. These fix procedures analyze the system and provide more information about the problem. They suggest actions to take and step you through the actions that automatically manage the system where necessary. Finally, they check that the problem is resolved.

If an error is reported, always use the fix procedures within the management GUI to resolve the problem. Always use the fix procedures for both system configuration problems and hardware failures. The fix procedures analyze the system to ensure that the required changes do not cause volumes to be inaccessible to the hosts. The fix procedures automatically perform configuration changes that are required to return the system to its optimum state.

Email notification

The Call Home feature transmits operational and event-related data to you and IBM through a Simple Mail Transfer Protocol (SMTP) server connection in an event notification email. When configured, this function alerts IBM service personnel about hardware failures and potentially serious configuration or environmental issues.

SNMP notification

Simple Network Management Protocol (SNMP) is a standard protocol for managing networks and exchanging messages. The system can send SNMP messages that notify personnel about an event. You can use an SNMP manager to view the SNMP messages that are sent by the IBM FlashSystem 9100 system.

The MIB file describes the format of the SNMP messages that are sent by IBM FlashSystem 9100. Use this MIB file to configure a network management program to receive SNMP event notifications that are sent from an IBM FlashSystem 9100 system. This MIB file is suitable for use with SNMP messages from all versions of IBM FlashSystem 9100.

IBM FlashSystem 9100 MIB file can be downloaded at:

`IBM FS9100 MIB File using FTP`

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.

8.1.2 Monitoring using quotas and alert

In an IBM FlashSystem 9100 system, the space usage of storage pools and Thin Provisioned or Compressed Volumes can be monitored by setting some specific quota alerts.

Storage Pool

During storage pool configuration, you can set a warning such that when the pool capacity reaches this quota setting, an alert is issued. This setting generates a warning when the used capacity in the storage pool first exceeds the specified threshold. You can specify a `disk_size` integer, which defaults to megabytes (MB) unless the **`-unit`** parameter is specified. Or you can specify a `disk_size%`, which is a percentage of the storage pool size. To disable warnings, specify 0 or `0%`. The default value is `0`.

Volumes

Thin Provisioned and compressed volumes near their size limits are monitored at specified thresholds to preserve data integrity. If a volume can be shrunk to below the recommended new limit, you are advised to do so. If volume capacity cannot be reduced to meet the recommended limit, you are advised to create a non-compressed mirror of the data (if one does not exist) and delete the primary copy.

8.2 Performance monitoring

Monitoring performance and the ability to collect historical performance metrics statistics is almost compulsory for any storage subsystem, and is for IBM FlashSystem 9100 as well.

The next sections show what performance analysis tools are integrated with IBM FlashSystem 9100 systems, and what IBM external tools are available to collect performance statistics to allow historical retention as well.

Remember that performance statistics are useful not only to debug or prevent some potential bottlenecks, but also to make capacity planning for future growth easier, as shown in Figure 8-1 on page 271.

8.2.1 Performance monitoring with the GUI

In IBM FlashSystem 9100, real-time performance statistics provide short-term status information for your systems. The statistics are shown as graphs in the management GUI.

You can use system statistics to monitor the bandwidth of all the volumes, interfaces, and MDisks that are being used on your system. You can also monitor the overall CPUs utilization for the system. These statistics summarize the overall performance health of the system and can be used to monitor trends in bandwidth and CPU utilization.

You can monitor changes to stable values or differences between related statistics, such as the latency between volumes and MDisks. These differences can then be further evaluated by performance diagnostic tools.

Additionally, with system-level statistics, you can quickly view bandwidth of volumes, interfaces, and MDisks. Each of these graphs displays the current bandwidth in megabytes per second and a view of bandwidth over time.

Each data point can be accessed to determine its individual bandwidth use and to evaluate whether a specific data point might represent performance impacts. For example, you can monitor the interfaces, such as for Fibre Channel or SAS interfaces, to determine whether the host data-transfer rate is different from the expected rate.

You can also select canister-level statistics, which can help you determine the performance impact of a specific canister. As with system statistics, canister statistics help you to evaluate whether the canister is operating within normal performance metrics.

The CPU utilization graph shows the current percentage of CPU usage and specific data points on the graph that show peaks in utilization. If compression is being used, you can monitor the amount of CPU resources that are being used for compression and the amount that is available to the rest of the system.

The Interfaces graph displays data points for Fibre Channel (FC), iSCSI, serial-attached SCSI (SAS), and IP Remote Copy interfaces. You can use this information to help determine connectivity issues that might affect performance.

The Volumes and MDisks graphs on the performance window show four metrics: Read, Write, Read latency, and Write latency. You can use these metrics to help determine the overall performance health of the volumes and MDisks on your system. Consistent unexpected results can indicate errors in configuration, system faults, or connectivity issues.

Each graph represents 5 minutes of collected statistics, updated every 5 seconds, and provides a means of assessing the overall performance of your system, as shown in Figure 8-1 on page 271.

Figure 8-1 denotes a Monitoring example of the IBM FlashSystem 9100 subsystem GUI.

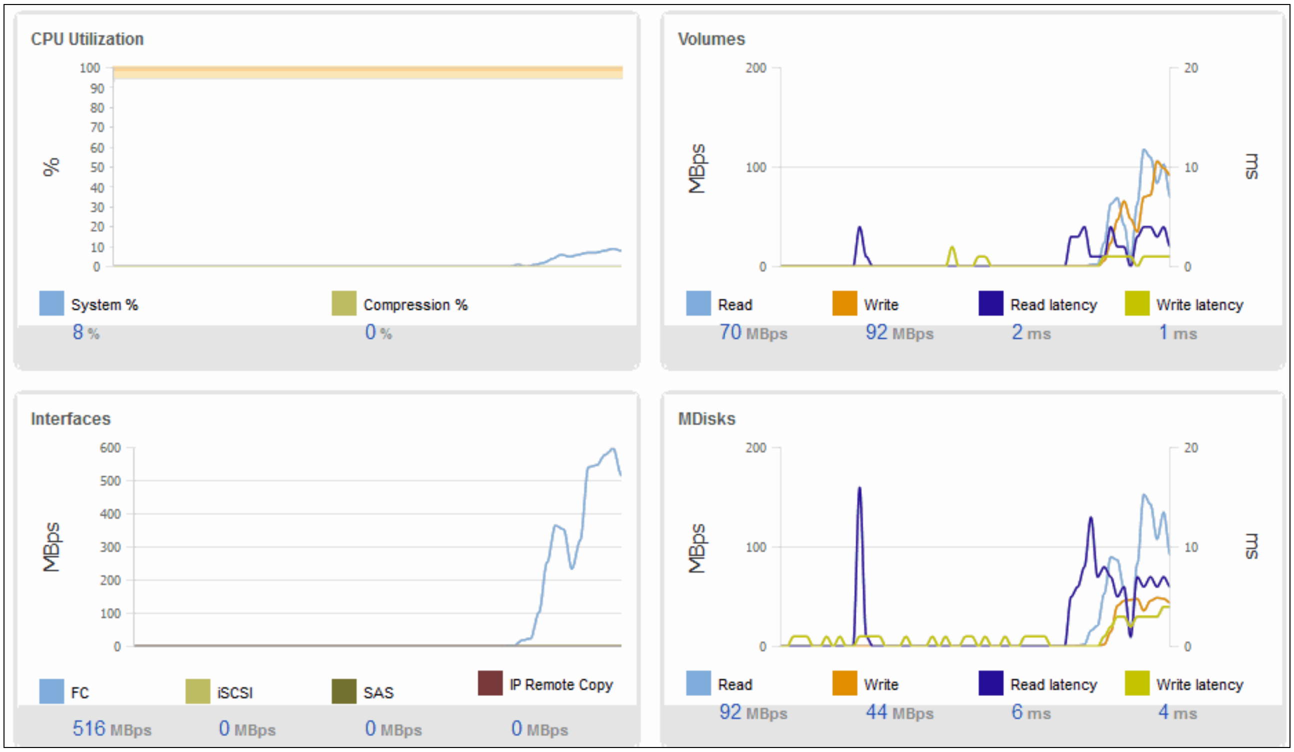


Figure 8-1 Monitoring GUI example

You can then choose the metrics that you want to be displayed, as shown in Figure 8-2.

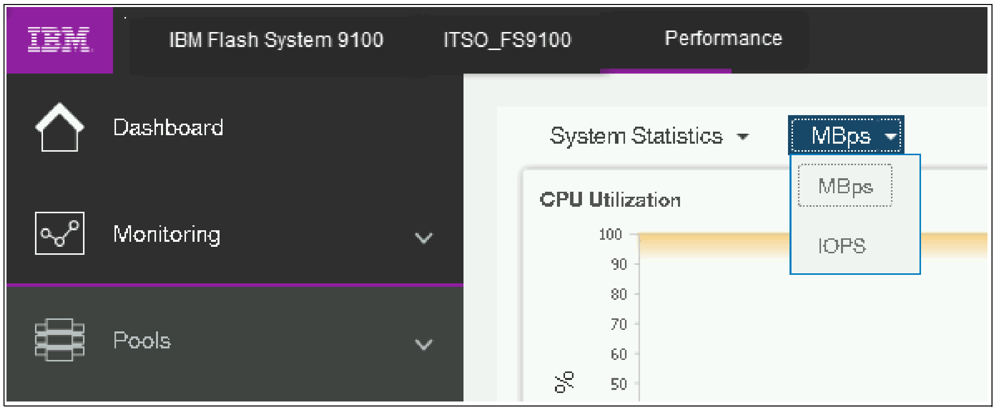
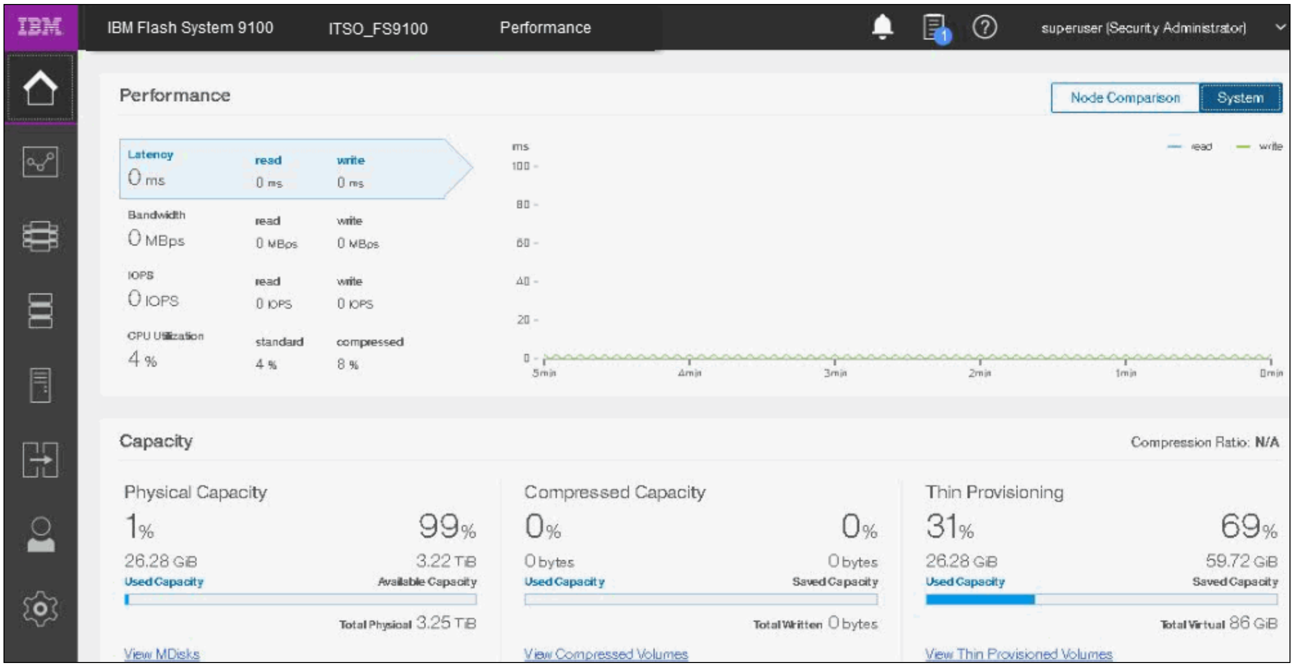


Figure 8-2 Selecting metrics

You can also obtain a quick overview by using the GUI option **System** → **Dashboard**, as shown in Figure 8-3.



Performance

Latency 0 ms — read 0 ms — write 0 ms

Bandwidth 0 MBps — read 0 MBps — write 0 MBps

IOPS 0 IOPS — read 0 IOPS — write 0 IOPS

CPU Utilization 4 % — standard 4 % — compressed 8 %

Capacity

Compression Ratio: N/A

Physical Capacity: 1% 26.28 GiB Used Capacity; 99% 3.22 TiB Available Capacity; Total Physical 3.25 TiB

Compressed Capacity: 0% 0 bytes Used Capacity; 0% 0 bytes Saved Capacity; Total Written 0 bytes

Thin Provisioning: 31% 26.28 GiB Used Capacity; 69% 59.72 GiB Saved Capacity; Total Virtual 86 GiB

Figure 8-3 System -> Dashboard

8.2.2 Performance monitoring with IBM Spectrum Control

IBM Spectrum Control is an on-premises storage management, monitoring, and reporting solution. It leverages 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. For more information on IBM Storage Insights, please see 8.2.3, "Performance monitoring with IBM Storage Insights" on page 276

Because IBM Spectrum Control is an on-premise tool, it does not send the metadata about monitored devices offsite, which is ideal for dark shops and sites that don't want to open ports to the cloud.

If you want to learn more about the capabilities of IBM Spectrum Control, check out the dedicated knowledge center for detailed information at `IBM Spectrum Control documentation`.

For pricing and other purchasing information, go to:

`IBM Spectrum Control`

> **Note:** If you currently have IBM Spectrum Control or manage IBM block storage systems, you already have access to IBM Storage Insights (free version). Go `here` to learn how to get started.

IBM Spectrum Control offers several reports that you can use to monitor IBM FlashSystem 9100 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 9100 systems. The next sections provides some basic suggestions about what metrics need to be monitored and analyzed to debug potential bottleneck problems. In addition, which alerts need to be set to be notified when some specific metrics exceed limits that are considered important for this specific environment.

For more information about the installation, configuration, and administration of IBM Spectrum Control (including how to add a storage system), see these websites:

- IBM Spectrum Control 5.3.1 Limitations and known issues
- Installing IBM Spectrum Control 5.3.1

> **Note:** IBM Spectrum Control 5.3.0 or higher is recommended for monitoring IBM FlashSystem 9100.

IBM Spectrum Control dashboard

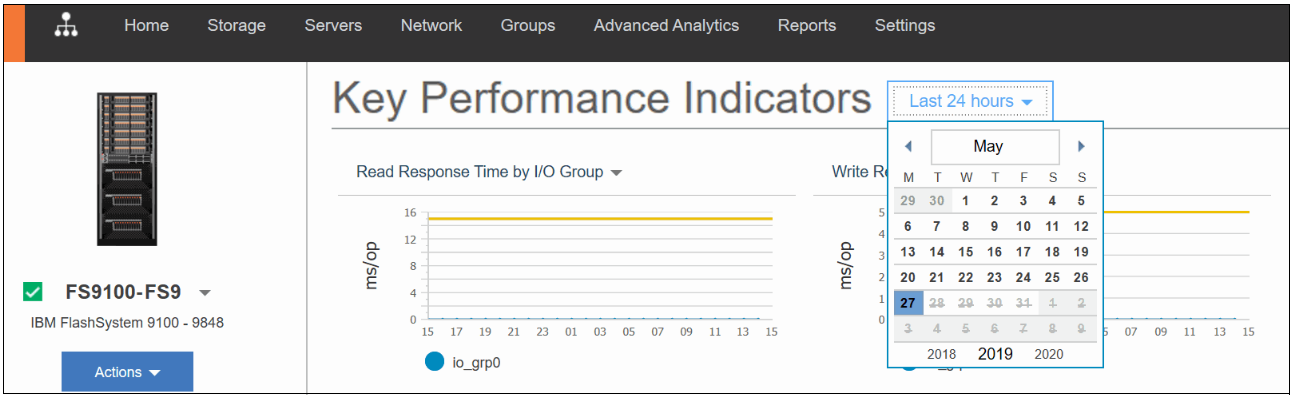
The performance dashboard provides Key Performance Indicators (in prior releases Best Practice Performance Guidelines) for the critical monitoring metrics. These guidelines do not represent the maximum operating limits of the related components, but are rather suggested limits that are selected with an emphasis on maintaining a stable and predictable performance profile.

The dashboard displays the *Last 24 hours* from the active viewing time and date. Selecting an individual element from the chart overlays the corresponding 24 hours for the previous day and seven days prior. This display allows for an immediate historical comparison of the respective metric. The day of reference can also be changed to allow historical comparison of previous days.

These dashboards provide two critical functions:

- Provides an “at-a-glance” view of all the critical FS9100 monitoring metrics.
- Provides a historical comparison of the current metric profile with previous days that enables rapid detection of anomalous workloads and behaviors.

Figure 8-4 shows how to change the day of reference.



M	T	W	T	F	S	S
29	30	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	~~28~~	~~29~~	~~30~~	~~31~~	~~1~~	~~2~~
~~3~~	~~4~~	~~5~~	~~6~~	~~7~~	~~8~~	~~9~~

2018 2019 2020

Figure 8-4 Change day of reference

Figure 8-5 shows a metric that is exceeding the best practice limit (orange line).

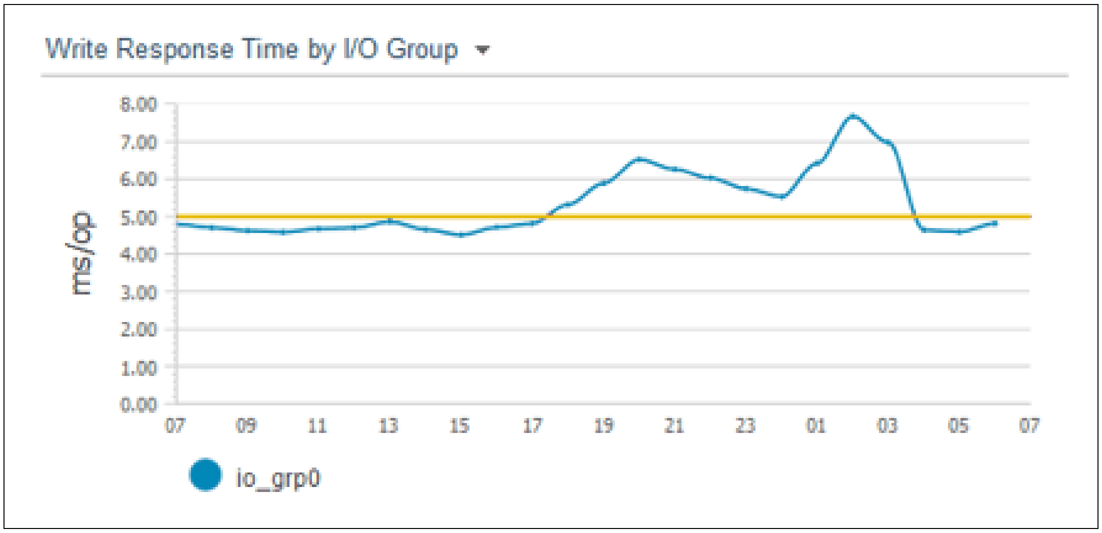


Figure 8-5 Metric exceeding best practice

Figure 8-6 shows the same chart as in Figure 8-5 with `io_grp0` selected, which overlays the previous day and 7 days prior.

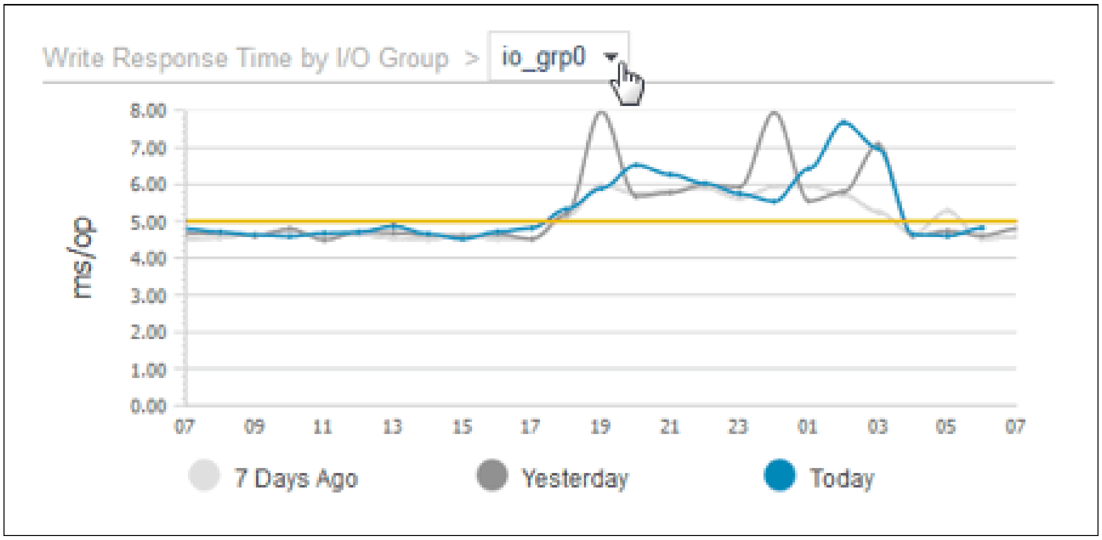


Figure 8-6 Changed chart due to iogrp selection

From this information, you can quickly conclude that this exception occurs every day at this same time, and is not a new phenomenon.

The line in yellow line is if a component is constantly breaching the limit, then this component might be overly utilized, an occasional peak doesn’t matter: this is more to help understand how the hardware is going and is not a service indicator (response times > 10 ms are hardly acceptable these days).

Note: The Best Practices Guidelines panel has recently been renamed *Key Performance Indicators*.

Key Performance Indictors (Best Practice Performance Guidelines)

You can view the key metrics that are outside of a standard range for storage systems that run IBM FlashSystem 9100 by using the performance guidelines. The guidelines were established by a historical analysis of storage environments.

Most of the performance charts show an orange line that indicates the best practice value for the metric. These guidelines are established as the levels that allow for a diverse set of workload characteristics while maintaining a stable performance profile. The other lines on each chart represent the measured values for the metric for the resources on your storage system: I/O groups, ports, or canisters.

You can use the lines to compare how close to potentially becoming overloaded your resources are. If your storage system is responding poorly and the charts indicate overloaded resources, you might have to better balance the workload. You can balance the workload between the canisters of the cluster, potentially adding more canisters to the cluster, or move some workload to other storage systems.

Figure 8-7 shows the Key Performance Indicators in the Dashboard.

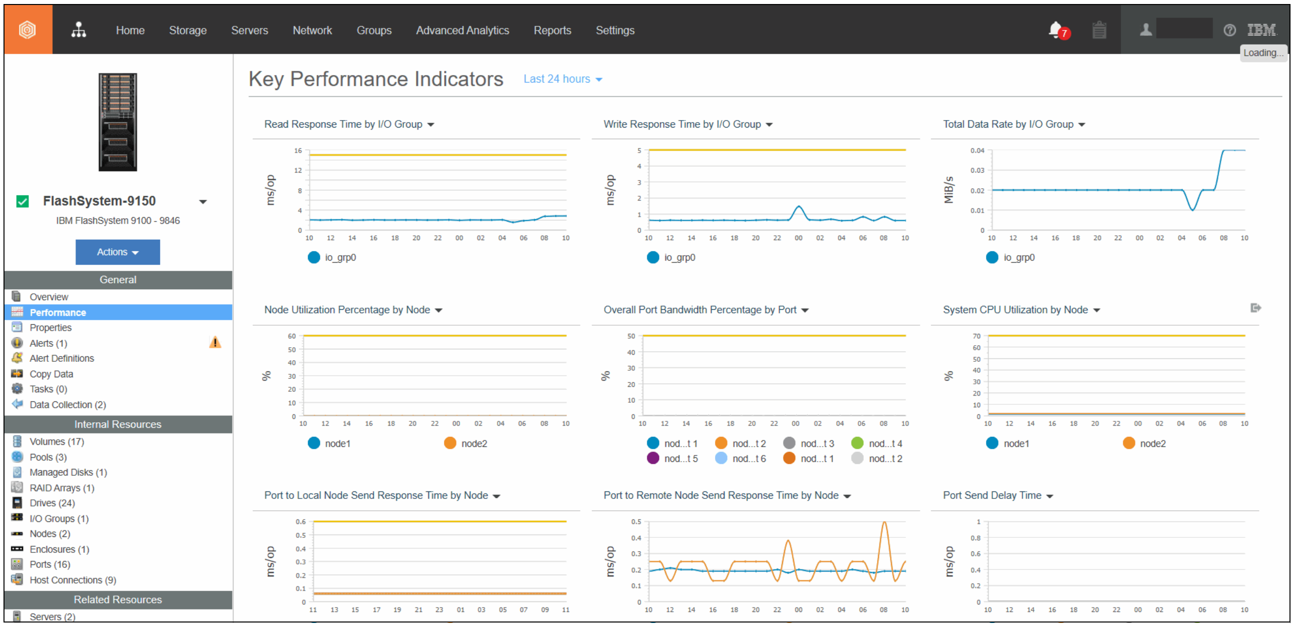


Figure 8-7 DashBoard- Key Performance Indicators

The charts show the hourly performance data measured for each resource on the selected day. Use the following charts to compare the workloads on your storage system with the best practice guidelines:

- **Node Utilization Percentage by Node**: Compare the guideline value for this metric, for example, 60% utilization, with the measured value from your system.
  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.
  For clusters without FC ports this chart is empty (or when no host I/O is going on).
- **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.
  This is a very important metric for a good performing cluster.
- **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.
  It means, when you see this constantly being exceeded, something might be wrong with the hardware.
- **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.
  Keep in mind, that this will only work with 8 Gbps adapters, for 16 Gbps or 32 Gbps adapters using the port delay metrics (not in this overview).

Note: The guidelines are not thresholds, and they are not related to the alerting feature in IBM Spectrum Control. To create performance alerts that use the guidelines as thresholds, go to a resource detail window in the web-based GUI, click **Alerts** in the General section, and then click **Definitions**.

8.2.3 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’s an IBM Cloud service, getting started is simple and upgrades are handled automatically.

By leveraging 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 and 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 8-8 shows an IBM Storage Insight® example screen.

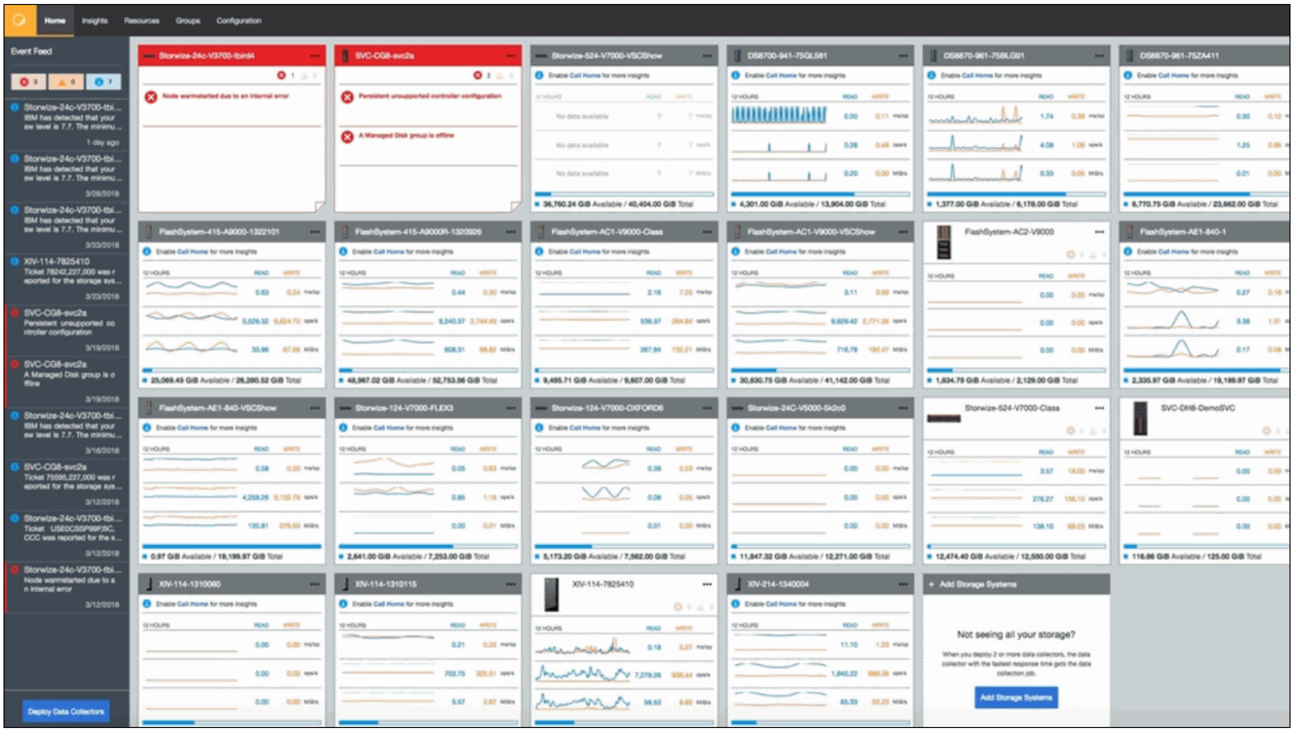


Figure 8-8 IBM Storage Insight

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.

`IBM Storage Insights Security`

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** 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 will you show IBM analysis results.

- The **capacity-based**, subscription version is called **IBM Storage Insights Pro** and includes **all** 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 8-1 shows the different features of both versions.

Table 8-1 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 (3 metrics: I/O rate, data rate, and response times aggregated for storage systems)	Advanced (100+ metrics for storage systems and their components)
	Capacity	Basic (4 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	✓*	✓
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 have access to IBM Storage Insights but are not an IBM Storage Insights Pro subscriber, you must have a current warranty or maintenance agreement for an IBM block storage system to open tickets and send log packages.

The IBM FlashSystem 9100 (V8.2) is supported in conjunction with IBM Storage Insights and IBM Storage Insights Pro.

Note: The reporting feature is not available in IBM Storage Insights (free). In order to use the reporting feature, you must subscribe to IBM Storage Insights Pro or you can use IBM Spectrum Control.

For information about how to try to buy the IBM Storage Insights Pro version, go to IBM Support.

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 don't 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. Here's how IBM Storage Insights for IBM Spectrum Control can transform your monitoring and support experience:

- 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 don't have to.
- Use Call Home to monitor devices, get best practice recommendations, and filter events to quickly isolate trouble spots.
- Leverage 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.

Feature comparison of IBM Spectrum Control and IBM Storage Insights for IBM Spectrum Control

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 8-2 on page 281 shows the features in IBM Spectrum Control and IBM Storage Insights for IBM Spectrum Control.

Table 8-2 Feature 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	✓	
Alerting	Predictive Alerts	✓	✓
	Customizable, multi-conditional alerting, including alert policies	✓	✓

Resource Management	Features	IBM Spectrum Control (Advanced edition)	IBM Storage Insights for IBM Spectrum Control
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. Details can be found at: `Storage Insights Registration,` choose the option for IBM Spectrum Control, and follow the prompts.

IBM Storage Insights for IBM Spectrum Control doesn't include the service level agreement for IBM Storage Insights Pro. Terms and conditions for IBM Storage Insights for IBM Spectrum Control are available at `Cloud Services Terms.`

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 as well as those with established infrastructures who need to expand their capabilities.

- 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’s 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.

As IBM Spectrum Storage Suite it contains 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’s recommended that you define alerts in one of the offerings and 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 setup of IBM Storage Insights

The following sections describe the steps to implement and set up IBM Storage Insights.

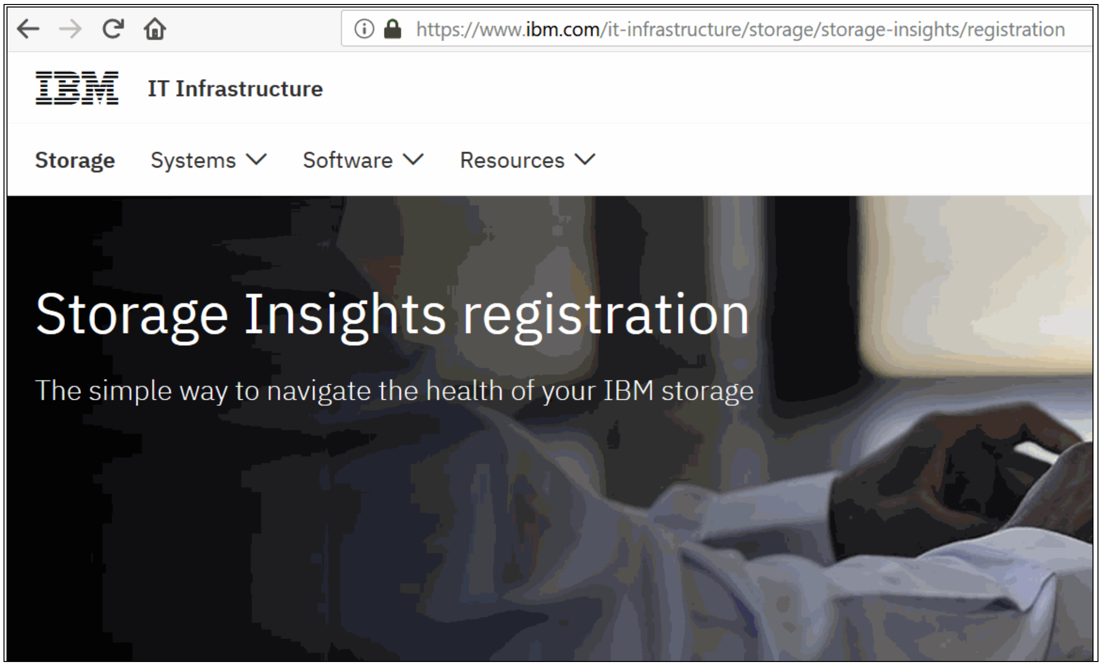
Sign up process

In order to use IBM Storage Insights with the IBM FlashSystem 9100 you have first to sign up:

`Storage Insights Registration`

- For the sign-up process, you’ll need an IBM ID. If you don’t have one, please 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’ll receive a Welcome email when IBM Storage Insights is ready. The email contains a direct link to your dashboard.

Figure 8-9 shows the IBM Storage Insight registration screen.



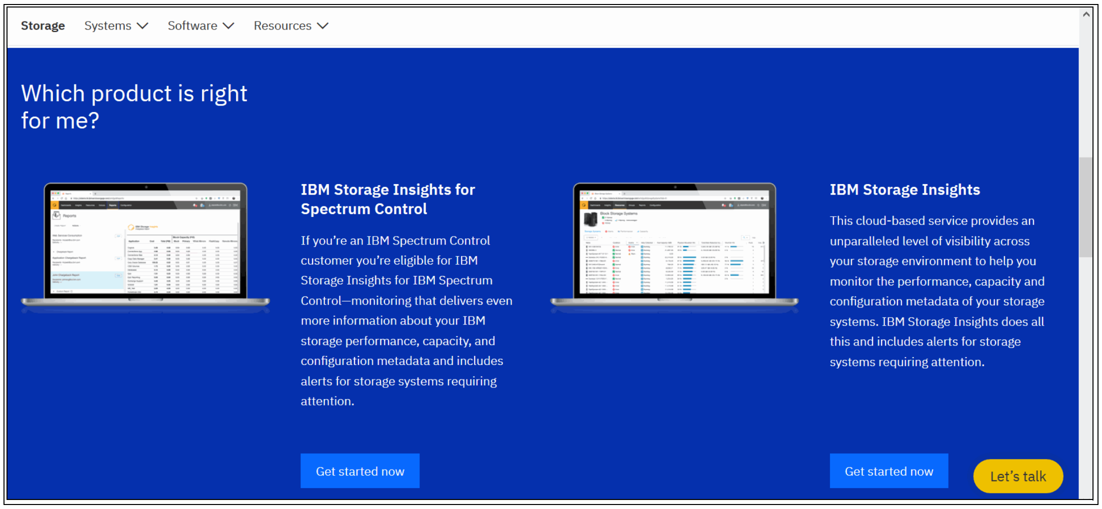
https://www.ibm.com/it-infrastructure/storage/storage-insights/registration
IBM IT Infrastructure
Storage
Systems
Software
Resources

Storage Insights registration

The simple way to navigate the health of your IBM storage

Figure 8-9 IBM Storage Insight registration screen

Figure 8-10 denotes the registration website when you scroll down. You can select here whether you want to register for IBM Storage Insights or IBM Storage Insights for Spectrum Control. For more details on differences of the IBM Storage Insights software, see "Licensing and editions of IBM Storage Insights" on page 277.



Which product is right for me?

IBM Storage Insights for Spectrum Control

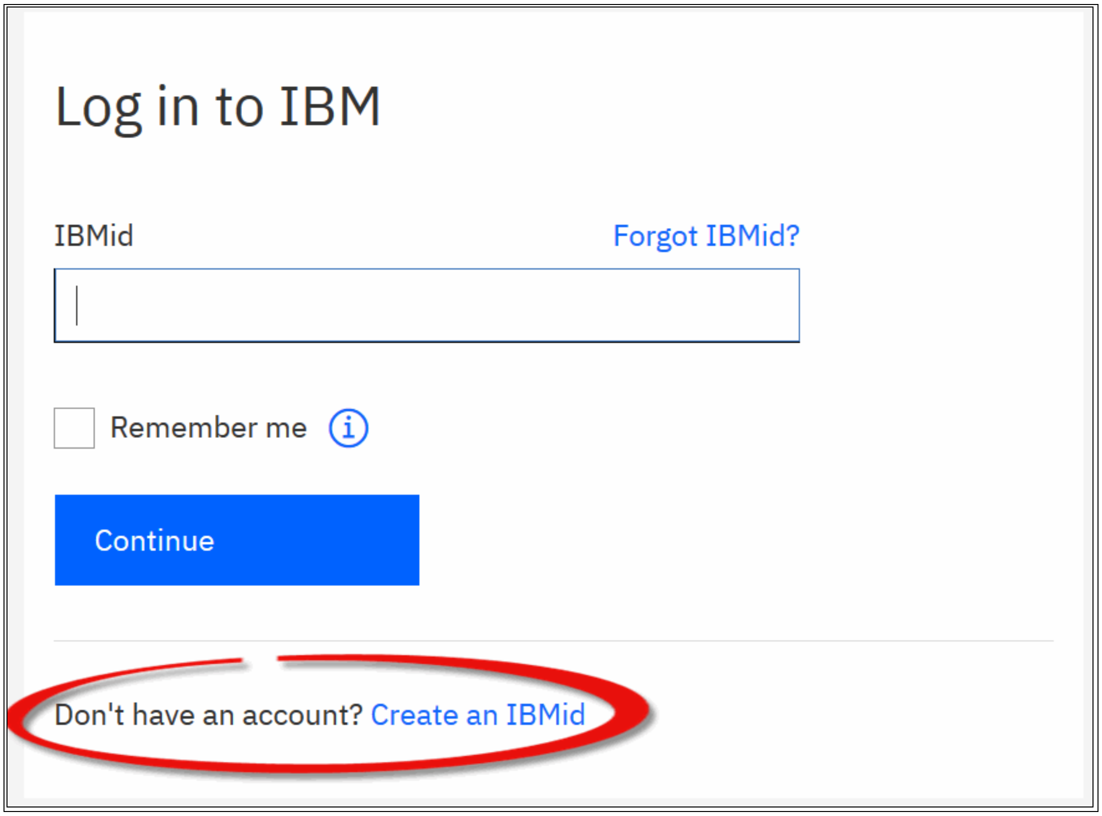
If you're an IBM Spectrum Control customer you're eligible for IBM Storage Insights for IBM Spectrum Control—monitoring that delivers even more information about your IBM storage performance, capacity, and configuration metadata and includes alerts for storage systems requiring attention.

IBM Storage Insights

This cloud-based service provides an unparalleled level of visibility across your storage environment to help you monitor the performance, capacity and configuration metadata of your storage systems. IBM Storage Insights does all this and includes alerts for storage systems requiring attention.

Figure 8-10 Choose IBM Storage Insights or IBM Storage Insights for Spectrum Control

Figure 8-11 shows the Log-in screen in the registration process. If you already have your credentials, type in you ID and proceed to the next screen by clicking on Continue. If you have no ID, proceed to **Create an IBMid**.



Log in to IBM

Don't have an account? Create an IBMid

Figure 8-11 Registration log-in screen

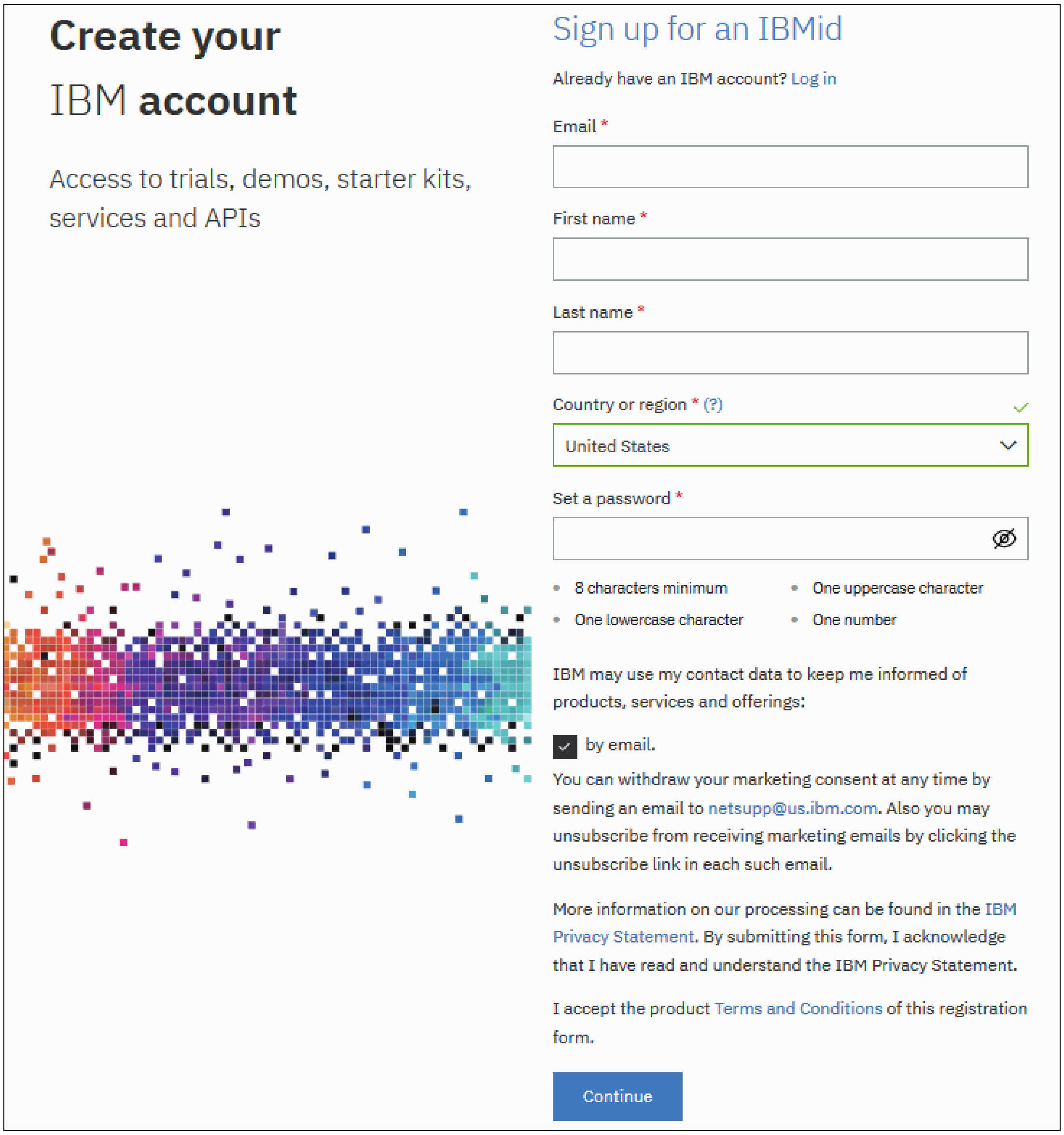
Figure 8-13 on page 286 shows the log-in screen prompt for ID & password.

If you want to create an IBMid, see Figure 8-12 on page 286 and provide the following information

- Email
- First name
- Last name
- Country or region
- Password

Select the box 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 any email.

More information on our processing can be found in the `IBM Privacy Statement`.



Create your

IBM **account**

Access to trials, demos, starter kits, services and APIs

Sign up for an IBMid

Already have an IBM account? Log in

Email *

First name *

Last name *

Country or region * (?)

United States

Set a password *

- 8 characters minimum
- One uppercase character
- One lowercase character
- One number

IBM may use my contact data to keep me informed of products, services and offerings:

☑ by email.

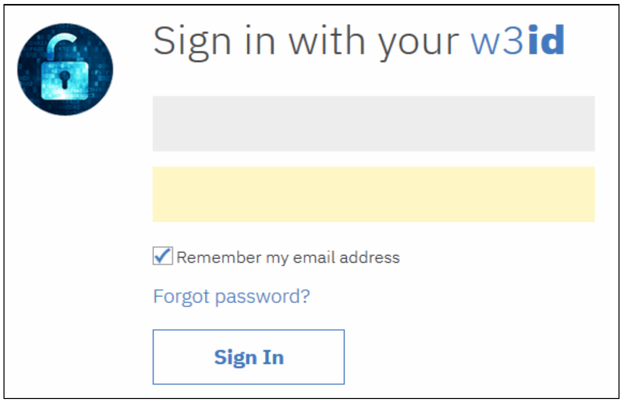
You can withdraw your marketing consent at any time by sending an email to netsupp@us.ibm.com. Also you may unsubscribe from receiving marketing emails by clicking the unsubscribe link in each such email.

More information on our processing can be found in the IBM Privacy Statement. By submitting this form, I acknowledge that I have read and understand the IBM Privacy Statement.

I accept the product Terms and Conditions of this registration form.

Figure 8-12 Create an IBM account

Figure 8-13 shows the log-in screen prompt for ID and password.



Sign in with your w3**id**

☑ Remember my email address

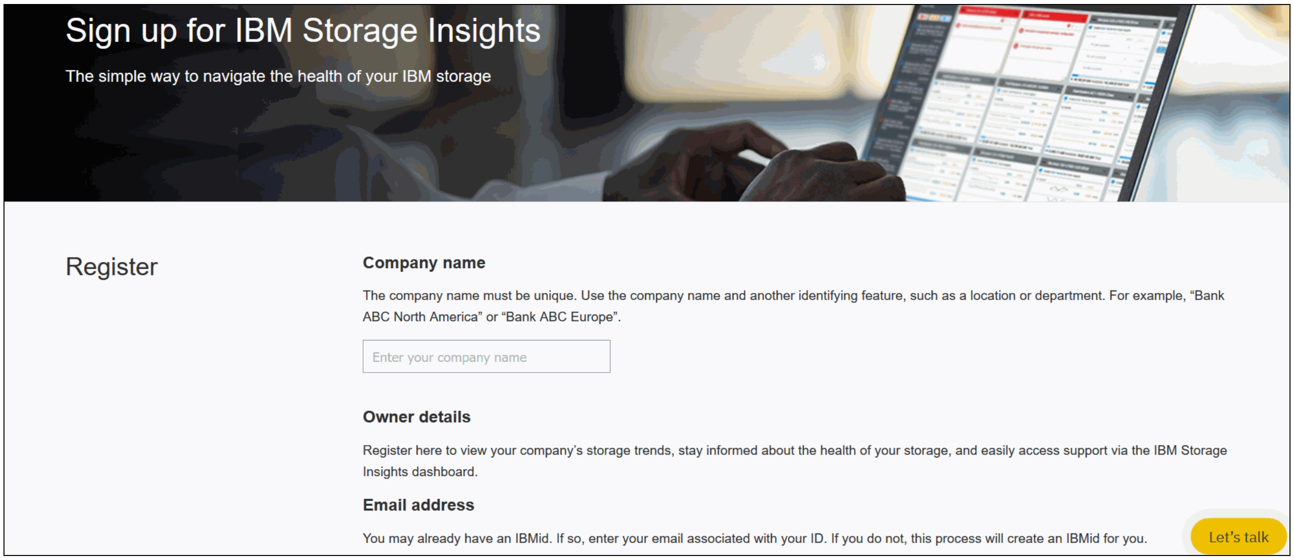
Figure 8-13 Registration - ID and password

Figure 8-14 shows the registration form. Complete the necessary information:

- Company name

  The name must be unique

- You might consider other identifying features, such as a location or department.
  - Owner details
  - The person who registered for IBM Storage Insights
  - Access granted for storage trends, health of storage and access to support
  - Email address / ID
  - First and last name



Sign up for IBM Storage Insights

The simple way to navigate the health of your IBM storage

Register

Company name

The company name must be unique. Use the company name and another identifying feature, such as a location or department. For example, “Bank ABC North America” or “Bank ABC Europe”.

Enter your company name

Owner details

Register here to view your company’s storage trends, stay informed about the health of your storage, and easily access support via the IBM Storage Insights dashboard.

Email address

You may already have an IBMid. If so, enter your email associated with your ID. If you do not, this process will create an IBMid for you.

Figure 8-14 IBM storage insights registration form

After your 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're ready to start adding your storage systems for monitoring.

Requirements: 1 GB RAM, 1 GB disk space, and Windows, AIX, or Linux (x86-64 systems only).

Learn more at Downloading and installing data collectors.

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:

- 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 about the metadata that is collected and how it's used see:

IBM Storage InsightsFact Sheet PDF Download and IBM Storage Insights Security Guide PDF Download.

Add storage system

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.

Read more at Adding storage systems.

View your dashboard

On the Operations dashboard, view:

- 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've got enough capacity to meet your storage demands.
- Key performance metrics so you know whether the performance of your storage systems meets operational requirements.

Read more at Operations dashboard.

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.

Read more at Monitoring resources with Call Home.

Add users to your dashboard

Optional: Add users, such as other storage administrators, IBM Technical Advisors, and IBM Business Partners, at any time so that they can access your IBM Storage Insights dashboard.

1. In IBM Storage Insights, click your user name in the upper-right corner of the dashboard.
2. Click Manage Users.
3. On your MYIBM page, ensure that IBM Storage Insights is selected.
4. Click Add new user.

Read more at Managing users.

8.3 Capacity metrics for block storage systems

Effective and exact capacity management is based on fundamental knowledge of capacity metrics in the IBM FlashSystem 9100 system. Data Reduction pool, Thin Provisioning, compression and deduplication are adding a lot of metrics to the IBM FlashSystem 9100 GUI, IBM Spectrum Control as well as IBM Storage Insights. The capacity metrics in this section are based on IBM Spectrum Control V5.3.3.

The Capacity section on the Dashboard provides an overall view of system capacity. This section displays physical capacity, volume capacity, and capacity savings.

Physical capacity indicates the total capacity in all storage on the system. Physical capacity includes all the storage the system can virtualize and assign to pools. Physical capacity is displayed in a bar graph and divided into three categories: Stored Capacity, Available Capacity, and Total Physical. If the system supports self-compressing drives, certain system configurations make determining accurate physical capacity on the system difficult.

For example, if the system contains self-compressed drives and data reduction pools without compression enabled, the system cannot determine the accurate amount of physical capacity that is used on the system. In this case, over provisioning and losing access to write operations is possible. If this condition is detected by the system, the Physical Capacity section of the Dashboard page displays a message instead of capacity information.

To recover from this condition, you need to ensure that all thin-provisioned volumes and thin-provisioned volumes that are deduplicated are migrated to volumes with compression enabled in the data reduction pools. Alternatively, you can migrate the volumes to fully allocated volumes and use the drive compression to save capacity.

We will discuss the different values and how they help to determine the capacity utilization, trends in capacity and space usage, and more importantly can prevent an out of space situation in the environment. In order to not run out of space you have to understand the different level of components, such as arrays, pools, system, and so on.

You should understand which limits exist for each of them so that you understand while one pool might be fine, another pool(s) could run out of storage and just monitoring at the system level is not appropriate if you have two or more pools.

Figure 8-15 shows how to interpret the capacity and savings in a storage environment.

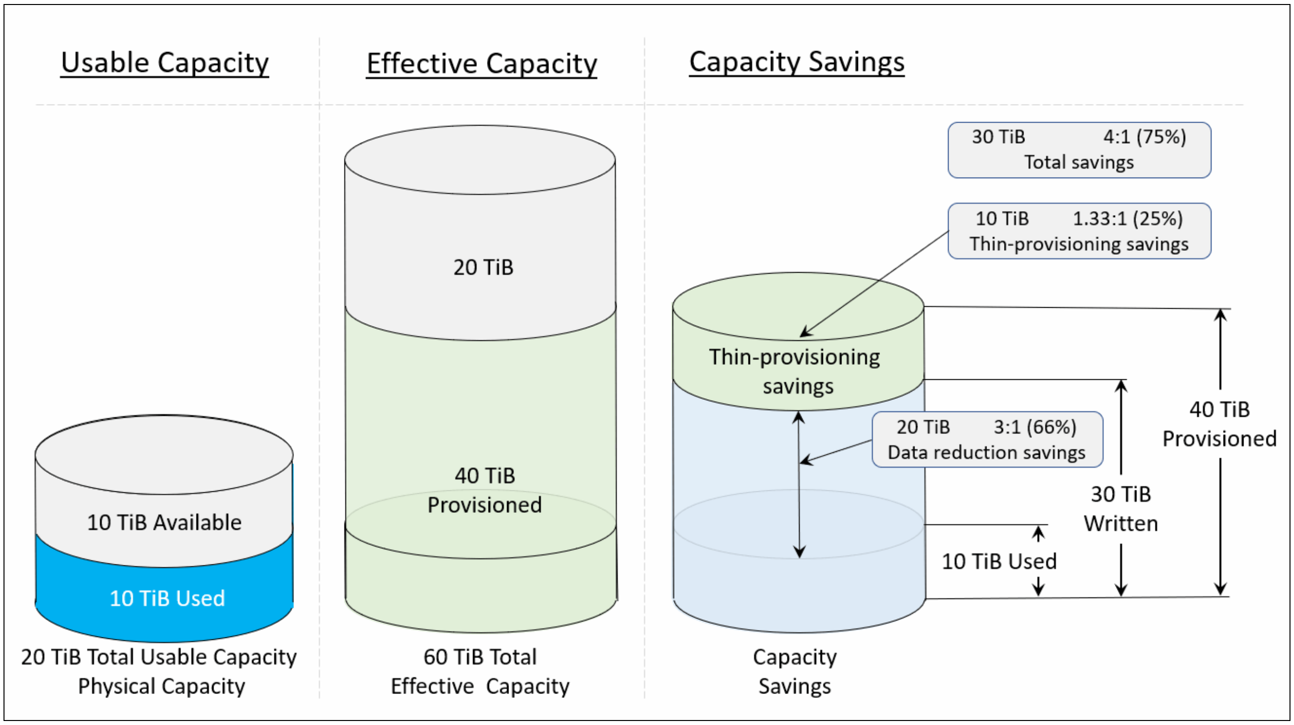


Figure 8-15 Understanding capacity information

The following metrics can be added to capacity charts for storage systems within capacity planning. Use the charts to detect capacity shortages and space usage trends.

Alphabetical lists of the capacity and space usage metrics that you can add to charts are provided in the following sections:

- Storage system capacity metrics
- Pool capacity metrics
- Volume capacity metrics

8.3.1 Storage system capacity metrics

The following are the system capacity metrics.

Allocated Space (GiB)

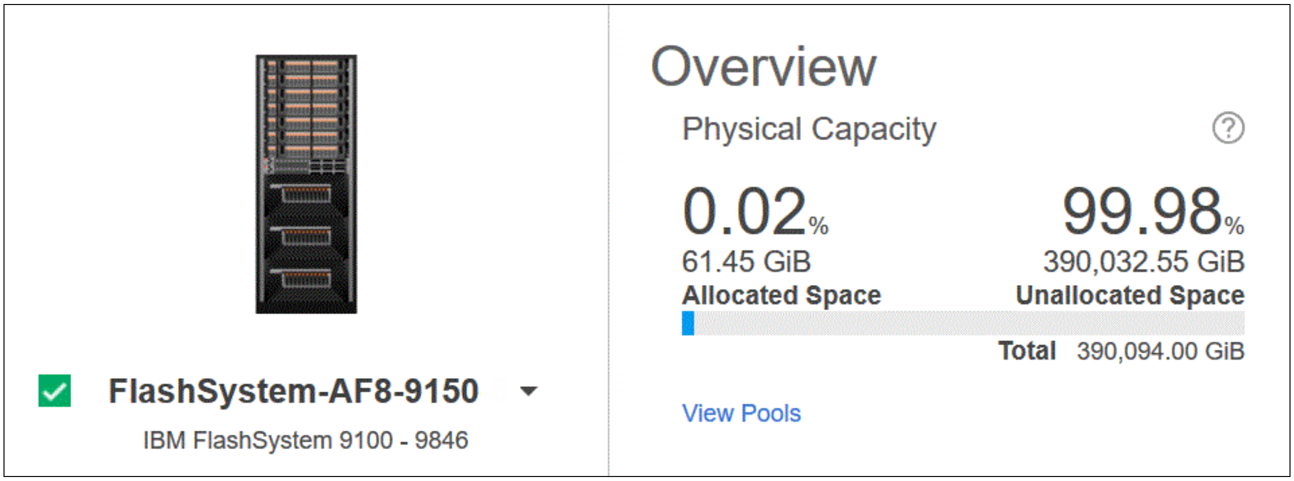
The amount of space that is allocated to the regular and thin-provisioned volumes in the pools. If the pool is a parent pool, the amount of space that is allocated to the volumes in the child pools is also included.

The space that is allocated for thin-provisioned volumes is less than their virtual capacity, which is shown in the **Total Volume Capacity (GiB)** column. If a pool doesn't have thin-provisioned volumes, the value for allocated space is the same as the value for total volume capacity.

Allocated space is the same as used space on all storage systems with the following exceptions:

- IBM FlashSystem 9100 / SAN Volume Controller / Storwize that are thin provisioned

Figure 8-16 shows the Allocated Space (61.45 GiB) of an IBM FlashSystem 9100.



☑ FlashSystem-AF8-9150

IBM FlashSystem 9100 - 9846

Overview

Physical Capacity

0.02% 61.45 GiB Allocated Space

99.98% 390,032.55 GiB Unallocated Space

Total 390,094.00 GiB

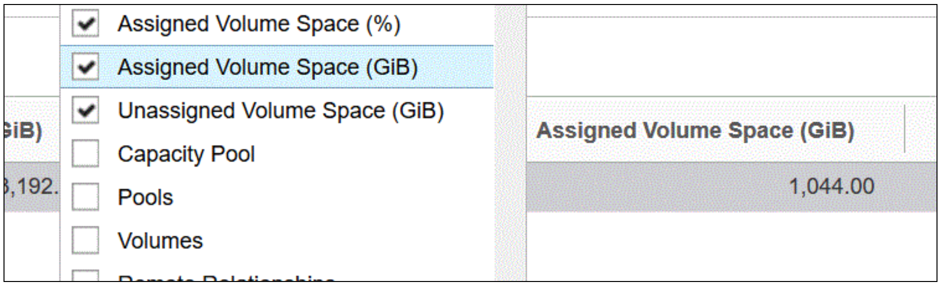
Figure 8-16 Allocated Space

Assigned Volume Space (GiB)

The total volume space in the storage system that is mapped or assigned to host systems, including child pool capacity. It is the sum of all volumes virtual size, so this value can exceed the physical capacity of your system, when you over provision storage. Complete the following steps:

1. Select **IBM Spectrum Control** → **choose your FS9100** (for example, FlashSystem AF8-9150).
2. Select **Actions** → **View Capacity** and scroll down.
3. Verify whether the **Assigned Volume Space** check box has been selected (right-click and determine if the **Assigned Volume Space** is marked).

Figure 8-17 shows the Assigned Volume Space (GiB) in an IBM FlashSystem 9100. The figure shows the total volume space for this system.

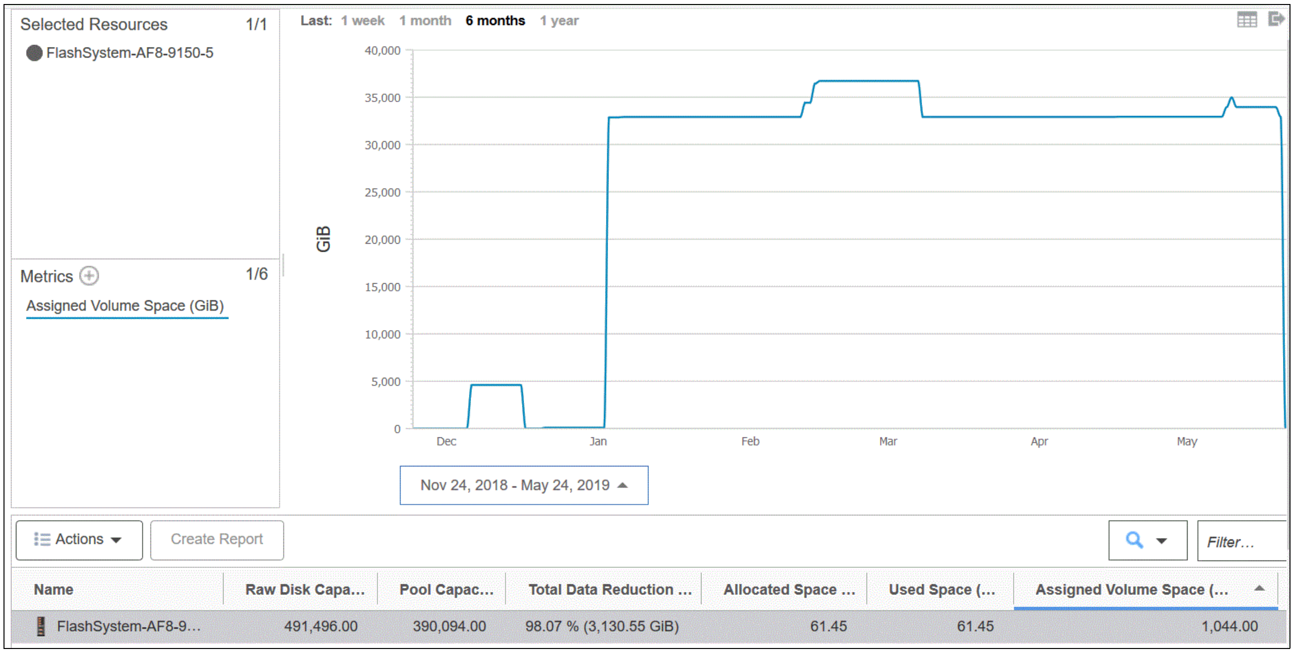


☑ Assigned Volume Space (%)
☑ Assigned Volume Space (GiB)
☑ Unassigned Volume Space (GiB)
□ Capacity Pool
□ Pools
□ Volumes

Assigned Volume Space (GiB)
1,044.00

Figure 8-17 Assigned Volume Space (GiB)

Figure 8-18 shows a view of the actual and historic Assigned Volume Space in an IBM FlashSystem 9100.



Name	Raw Disk Capa…	Pool Capac…	Total Data Reduction …	Allocated Space …	Used Space (…	Assigned Volume Space (…
FlashSystem-AF8-9…	491,496.00	390,094.00	98.07 % (3,130.55 GiB)	61.45	61.45	1,044.00

Figure 8-18   Assigned Volume Space - graph

To view other information, click the plus sign next to metrics on the left.

Available Pool Space (GiB)

The total amount of the space in the pools that is not allocated to the volumes in the pools. To calculate available space, the following formula is used:

```
(pool capacity - allocated space)
```

For some storage systems the pool space is limited by the physical capacity after data reduction, so the more you can compress data the more you can store in such a pool, For other systems there is a limit in the address space before compression, which means that even if you can compress the data extremely highly, you might not be able to use all the physical space, because the address range before compression is exceeded. See Table 4-3, “IBM FCM data compression ratio and effective capacity” on page 93 for details.

Compression Savings (%)

The estimated amount and percentage of capacity that is saved by using data compression, across all pools in the storage system. The percentage is calculated across all compressed volumes in the pools and does not include the capacity of non-compressed 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. Drive level compression occurs after striping across MDisk and the RAID distribution on the disks/FCM level. There is no information available about which volume a block of data belongs to that was just compressed, so the information about compression within the drives is only available at the array level.

The following formula is used to calculate the amount of storage space that is saved:

```
(written space · compressed size)
```

The following formula is used to calculate the percentage of capacity that is saved:

```
((written space · compressed size) ÷ written space) × 100
```

For example, the written space, 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 just 10 GiB. Therefore, the compression savings percentage across all compressed volumes is 75%.

Deduplication Savings (%)

The estimated amount and percentage of capacity that is saved by using data deduplication, across all data reduction pools 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 space · deduplicated size)
```

The following formula is used to calculate the percentage of capacity that is saved:

```
((written space · deduplicated size) ÷ written space) × 100
```

For example, the written space, 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%.

Physical Allocation (%)

The percentage of physical capacity in the pools that is allocated to the regular volumes, the thin-provisioned volumes, and the volumes in child pools. Check the value for physical allocation to see:

- Whether the physical capacity of the pools is fully allocated. That is, the value for physical allocation is 100%.
- Whether you have sufficient capacity to provision new volumes with storage
- Whether you have sufficient capacity to allocate to the compressed and thin-provisioned

Figure 8-19 shows a brief description of Physical Allocation (%).

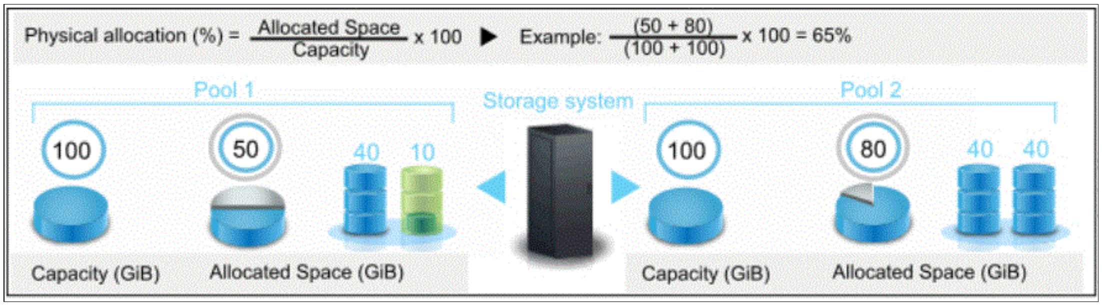


Figure 8-19 Physical Allocation (%)

Pool Capacity (GiB)

The total amount of storage space in the pools, which might include overhead space if the disks for the pools aren't formatted.

Pool Shortfall (%)

The percentage of space that is over-committed to the pools with thin-provisioned volumes. For example, you commit 100 GiB of space to a thin-provisioned volume in a pool with a capacity of 50 GiB. As the space is allocated to the thin-provisioned volume in increments of 10 GiB, the space available for allocation decreases and the shortfall in capacity becomes more acute.

To calculate the shortfall, the following formula is used:

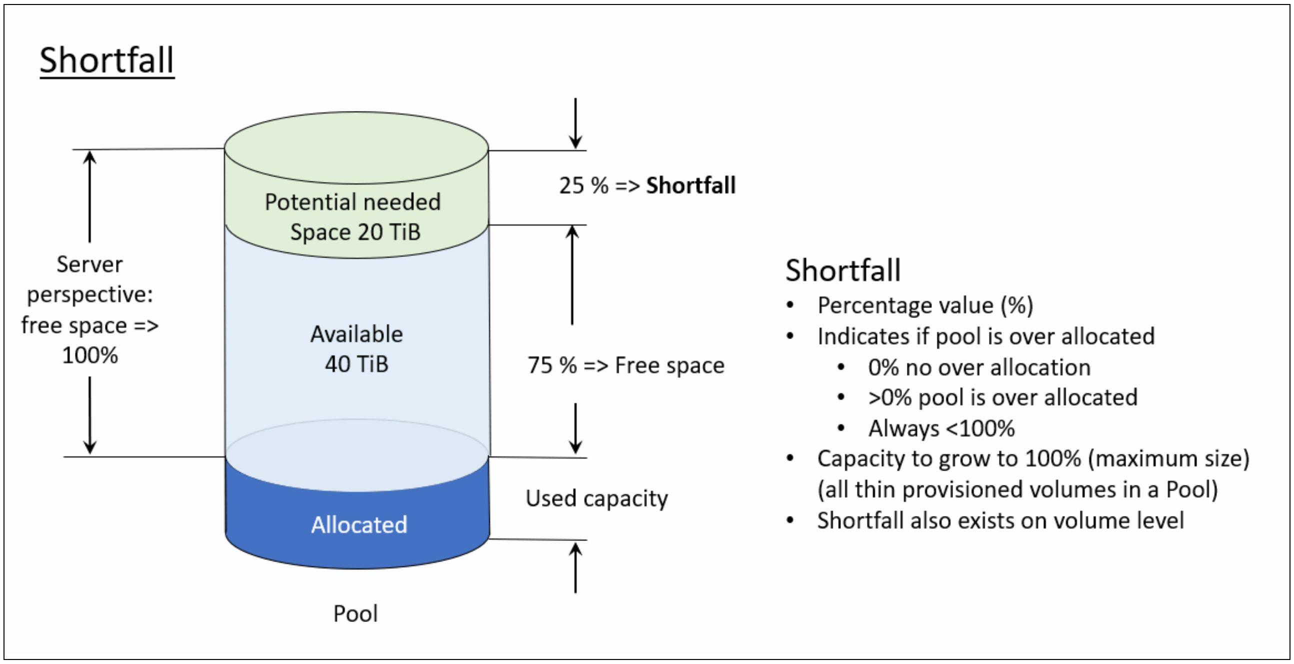
```
[(unallocatable space ÷ committed but unallocated space) × 100]
```

A pool shortfall occurs when you commit more space to the volumes in the pools than is physically available to the pools. If the physical space available to the pools is less than the committed virtual space, then the pools do not have enough space to fulfill the commitment to the virtual space.

For example, the physical capacity of the pools is 70 GiB, but 150 GiB of virtual space was committed to the thin-provisioned volumes. If the volumes are using 50 GiB, then 100 GiB is still committed to those volumes (150 GiB • 50 GiB) with only 20 GiB of available pool space (70 GiB • 50 GiB). Because only 20 GiB of the pool space is available, 80 GiB of the committed space cannot be allocated (100 GiB - 20 GiB). In this case, the percentage of committed space that is unavailable is 80% [(80 GiB ÷ 100 GiB) × 100].

The advantage of using shortfall rather than a simple overprovisioning factor, is that shortfall always has values between 0 and 100% and therefore is well suited for simple alerting that can be applied to multiple pools.

Figure 8-20 explains the shortfall.



Shortfall
- Percentage value (%)
- Indicates if pool is over allocated
  - 0% no over allocation
  - >0% pool is over allocated
  - Always <100%
- Capacity to grow to 100% (maximum size) (all thin provisioned volumes in a Pool)
- Shortfall also exists on volume level

Figure 8-20 Shortfall

Total Data Reduction Savings (%)

The estimated amount and percentage of capacity that is saved by using data deduplication, data compression, and thin provisioning.

The following formula is used to calculate the amount of storage space that is saved:

```
(Total Volume Capacity · Allocated Space)
```

The following formula is used to calculate the percentage of capacity that is saved:

```
((Total Volume Capacity · Allocated Space) ÷ Total Volume Capacity) × 100
```

Total Volume Capacity (GiB)

The total amount of storage space that can be made available to the regular and thin-provisioned volumes in the pools. 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. In other words this is the capacity that a server will see.

Unallocatable Volume Space (GiB)

The amount of space that cannot be allocated to volumes because the physical capacity of the pools cannot meet the demands for virtual space. The following formula is used to calculate this value:

```
[Total Volume Capacity · Pool Capacity]
```

Unallocated Volume Space (GiB)

The total amount of remaining space that can be allocated to the volumes in the pools. The following formula is used to calculate this value:

```
[Total Volume Capacity - Allocated Space]
```

The space that is allocated for thin-provisioned volumes is typically less than their virtual capacity. Therefore, the unallocated space represents the difference between the virtual capacity and the allocated space for all the volumes in the pools.

Virtual Allocation (%)

The percentage of the physical capacity that is committed to the virtual capacity of the volumes in the pool. If the value exceeds 100%, the physical capacity doesn’t meet the demands for virtual capacity. The following formula is used to calculate this value:

```
[(Total Volume Capacity ÷ Pool Capacity) × 100]
```

Example: If the allocation percentage is 200% for a total storage pool size of 15 GiB, then the virtual capacity that is committed to the volumes in the pool is 30 GiB. This configuration means that twice as much space is committed than is physically contained in the pool. If the allocation percentage is 100% for the same pool, then the virtual capacity that is committed to the pool is 15 GiB. This configuration means that all the physical capacity of the pool is already allocated to volumes.

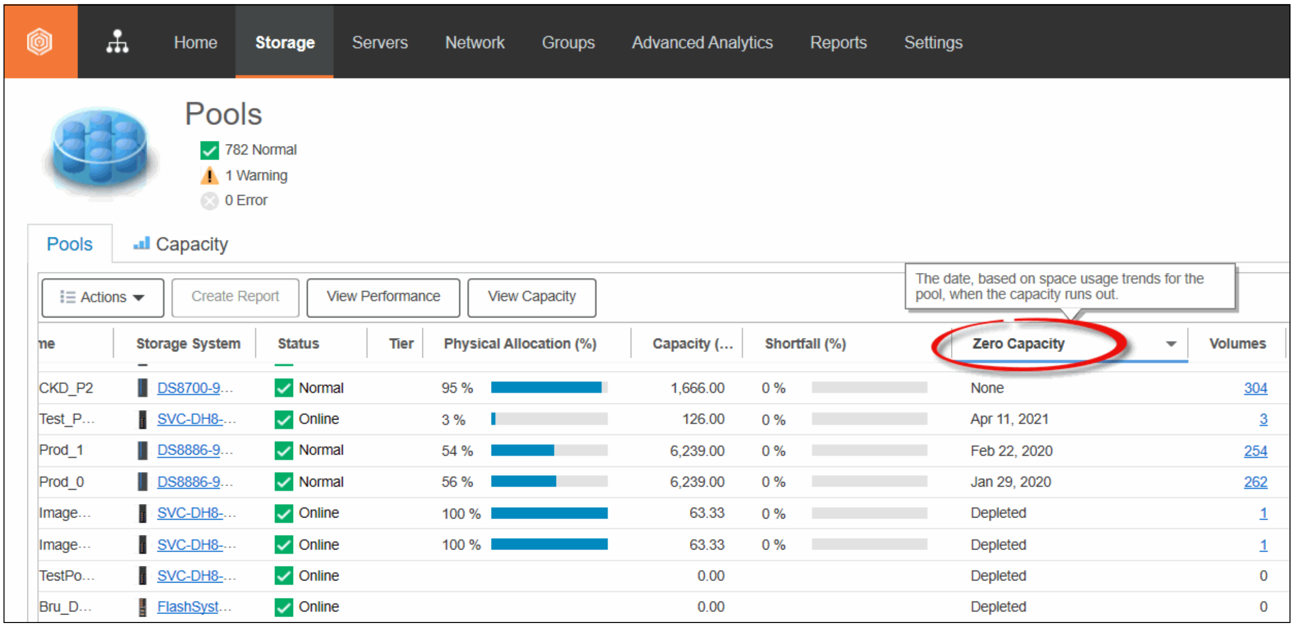
An allocation percentage that is higher than 100% is considered aggressive because insufficient physical capacity is available in the pool to satisfy the maximum allocation for all the thin-provisioned volumes in the pool. In such cases, you can use the value for Shortfall (%) to estimate how critical the shortage of space is for a pool.

8.3.2 Pool capacity metrics

If sufficient data is collected about the pools in your data center, you can view charts that compare the capacity of the pools with the space that is allocated to the pools and the space that is still available in the pools. In the **Zero Capacity** column on the Pools page, you can see the date, based on the space usage trends for the pool, when the pool runs out of available space.

> **Tip:** To order the pools in the table by the amount of space available to the pools, click **Filter by column**, and then click **Zero Capacity**.

Figure 8-21 shows an example of Zero Capacity trend.



Pools

✓ 782 Normal
⚠ 1 Warning
⊗ 0 Error

The date, based on space usage trends for the pool, when the capacity runs out.

ne	Storage System	Status	Tier	Physical Allocation (%)	Capacity (...	Shortfall (%)	Zero Capacity	Volumes
CKD_P2	DS8700-9...	✓ Normal		95 %	1,666.00	0 %	None	304
Test_P...	SVC-DH8-...	✓ Online		3 %	126.00	0 %	Apr 11, 2021	3
Prod_1	DS8886-9...	✓ Normal		54 %	6,239.00	0 %	Feb 22, 2020	254
Prod_0	DS8886-9...	✓ Normal		56 %	6,239.00	0 %	Jan 29, 2020	262
Image...	SVC-DH8-...	✓ Online		100 %	63.33	0 %	Depleted	1
Image...	SVC-DH8-...	✓ Online		100 %	63.33	0 %	Depleted	1
TestPo...	SVC-DH8-...	✓ Online			0.00		Depleted	0
Bru_D...	FlashSyst...	✓ Online			0.00		Depleted	0

Figure 8-21 Zero Capacity trend

The values that can be shown in the Zero Capacity column:

- **A date**
  The data 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 when the pool will be filled, for example if the trend is negative, as data is moved out of the pool
- **Depleted**
  the pool is already full

The following metrics can be added to capacity charts for storage systems within capacity planning. Use the charts to detect capacity shortages and space usage trends.

Allocated Space (GiB)

The amount of space that is allocated to the regular and thin-provisioned volumes in the pool. If the pool is a parent pool, the amount of space that is allocated to the volumes in the child pools is also included.

The space that is allocated for thin-provisioned volumes is less than their virtual capacity, which is shown in the **Total Volume Capacity** column. If a pool does not contain thin-provisioned volumes, this value is the same as Total Volume Capacity.

Allocated space is the same as used space on all storage systems, except for resources that run IBM Spectrum Virtualize. These resources might have more allocated space than used space if the storage administrator pre-allocated some space for thin-provisioned volumes when the volumes were created.

Assigned Volume Space (GiB)

The space on all of the volumes in a pool that are mapped or assigned to host systems. For a thin-provisioning pool, this value includes the virtual capacity of thin-provisioned volumes, which might exceed the total space in the pool.

Available Pool Space (GiB)

The amount of space that is available to create new volumes in the pool. If the pool is a parent pool, the amount of space that is allocated to the volumes in the child pools is also included.

Available Soft Space (GiB)

The amount of virtual storage space that is available to allocate to volumes in a storage pool.

Capacity (GiB)

The total amount of storage space in the pool, which might include overhead space if the disks for the pool aren’t formatted.

Compression Savings (%)

The estimated amount and percentage of capacity that is saved by using data compression. The percentage is calculated across all compressed volumes in the pool and does not include the capacity of non-compressed 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 space · compressed size)
```

The following formula is used to calculate the percentage of capacity that is saved:

```
((written space · compressed size) ÷ written space) × 100
```

For example, the written space, 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 just 10 GiB. Therefore, the compression savings percentage across all compressed volumes is 75%.

Deduplication Savings (%)

The estimated amount and percentage of capacity that is saved by using data deduplication. The percentage is calculated across all deduplicated volumes in the pool 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 space · deduplicated size)
```

The following formula is used to calculate the percentage of capacity that is saved:

```
((written space · deduplicated size) ÷ written space) × 100
```

For example, the written space, 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%.

Enterprise HDD Available Space (GiB)

The amount of storage space that is available on the Enterprise hard disk drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Enterprise HDD Capacity (GiB)

The total amount of storage space on the Enterprise hard disk drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Nearline HDD Available Space (GiB)

The amount of storage space that is available on the Nearline hard disk drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Nearline HDD Capacity (GiB)

The total amount of storage space on the Nearline hard disk drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Physical Allocation (%)

The percentage of physical capacity in the pool that is allocated to the regular volumes, the thin-provisioned volumes, and the volumes in child pools. This value is always less than or equal to 100% because you cannot allocate more physical space than is available in a pool. Check the value for physical allocation to see:

- Whether the physical capacity of the pool is fully allocated. That is, the value for physical allocation is 100%.
- Whether you have sufficient capacity to provision new volumes with storage.
- Whether you have sufficient capacity to allocate to the compressed and thin-provisioned volumes in the pool.

Figure 8-22 shows the Physical Allocation.

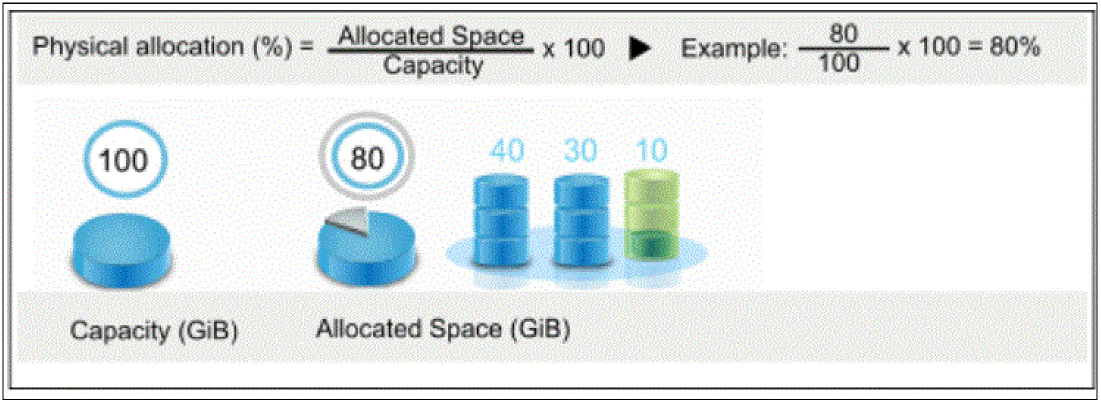

Physical allocation (%) = Allocated Space / Capacity x 100 ▶ Example: 80 / 100 x 100 = 80%
100
80
40
30
10
Capacity (GiB)
Allocated Space (GiB)

Figure 8-22 Physical Allocation

Shortfall (%)

The percentage of space that is over committed to pools with thin-provisioned volumes. For example, you commit 100 GiB of space to a thin-provisioned volume in a pool with a capacity of 50 GiB. As the space is allocated to the thin-provisioned volume in increments of 10 GiB, the space available for allocation decreases and the shortfall in capacity becomes more acute.

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.

To calculate shortfall, the following formula is used:

```
[(Unallocatable Space ÷ Committed but Unallocated Space) × 100]
```

You can use this percentage to determine when the amount of over-committed space in a pool is at a critically high level. Specifically, if the physical space in a pool is less than the committed virtual space, then the pool does not have enough space to fulfill the commitment to virtual space. This value represents the percentage of the committed virtual space that is not available in a pool. As more space is used over time by volumes while the pool capacity remains the same, this percentage increases.

Example: The remaining physical capacity of a pool is 70 GiB, but 150 GiB of virtual space was committed to thin-provisioned volumes. If the volumes are using 50 GiB, then 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 space in the pool, the shortfall is 30% when the following calculation is used:

```
[(100 GiB unallocated volume space · 70 GiB remaining pool space)
 ÷ 100 GiB unallocated volume space] × 100
```

Soft Space (GiB)

The amount of virtual storage space that is configured for the pool.

SSD Available Space (GiB)

The amount of storage space that is available on the solid-state drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

SSD Capacity (GiB)

The total amount of storage space on the solid-state drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Tier 0 Flash Available Space (GiB)

The amount of storage space that is available on the Tier 0 flash solid-state drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Tier 0 Flash Capacity (GiB)

The total amount of storage space on the Tier 0 flash solid-state drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Tier 1 Flash Available Space (GiB)

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 re-tiering the volume extents in the pool.

Tier 1 Flash Capacity (GiB)

The total amount of storage space on the Tier 1 flash, read-intensive solid-state drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Total Data Reduction Savings (%)

The estimated amount and percentage of capacity that is saved by using data deduplication, data compression, and thin provisioning, across all volumes in the pool.

The following formula is used to calculate the amount of storage space that is saved:

```
Total Volume Capacity · Allocated Space
```

The following formula is used to calculate the percentage of capacity that is saved:

```
((Total Volume Capacity · Allocated Space) ÷ Total Volume Capacity) × 100
```

Total Volume Capacity (GiB)

The total amount of storage space that can be made available to the regular and thin-provisioned volumes in 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.

Unallocatable Volume Space (GiB)

The amount of space that cannot be allocated to volumes because the physical capacity of the pools cannot meet the demands for virtual space. The following formula is used to calculate this value:

```
[Total Volume Capacity · Pool Capacity]
```

Unallocated Volume Space (GiB)

The total amount of remaining space that can be allocated to the volumes in the pools. The following formula is used to calculate this value:

```
[Total Volume Capacity - Allocated Space]
```

The space that is allocated for thin-provisioned volumes is typically less than their virtual capacity. Therefore, the unallocated space represents the difference between the virtual capacity and the allocated space for all the volumes in the pools.

Unassigned Volume Space (GiB)

The total amount of space in the volumes that are not assigned to hosts.

Virtual Allocation (%)

The percentage of the physical capacity that is committed to the virtual capacity of the volumes in the pool. If the value exceeds 100%, the physical capacity doesn't meet the demands for virtual capacity. The following formula is used to calculate this value:

```
[(Total Volume Capacity ÷ Pool Capacity) × 100]
```

Example: If the allocation percentage is 200% for a total storage pool size of 15 GiB, then the virtual capacity that is committed to the volumes in the pool is 30 GiB. This configuration means that twice as much space is committed than is physically contained in the pool. If the allocation percentage is 100% for the same pool, then the virtual capacity that is committed to the pool is 15 GiB. This configuration means that all the physical capacity of the pool is already allocated to volumes.

An allocation percentage that is higher than 100% is considered aggressive because insufficient physical capacity is available in the pool to satisfy the maximum allocation for all the thin-provisioned volumes in the pool. In such cases, you can use the value for Shortfall (%) to estimate how critical the shortage of space is for a pool.

8.3.3 Volume capacity metrics

You use the capacity chart to detect capacity shortages for the following types of volumes:

- Space-efficient volumes, such as compressed volumes and thin-provisioned volumes
- Regular volumes that use Easy Tier to re-tier volume extents

You can review the allocation of space to space-efficient volumes to detect capacity shortfalls. You can also review the space usage of volumes that use Easy Tier to distribute volume extents across Enterprise HDD, Nearline HDD, and SSD storage.

The following metrics can be added to capacity charts for storage systems within capacity planning. Use the charts to detect capacity shortages and space usage trends.

Allocated Space (GiB)

The amount of space that is allocated to the compressed, 5-provisioned, or the Easy Tier volume. Typically, the space that is allocated to the compressed or thin-provisioned volume is less than the capacity of the volume. For Easy Tier volumes, allocated space is the space that is allocated to the volume’s extents on the Enterprise HDD, Nearline HDD, or SSD drives.

Capacity (GiB)

The capacity of the compressed or the thin-provisioned volume, which comprises the sum of the allocated and the unallocated space. If the disks for the pool aren’t formatted, the capacity of the volume might include the overhead space.

Compression Savings (%)

The estimated amount and percentage of capacity that is saved by using data compression. the following formula is used to calculate the amount of storage space that is saved:

```
(written space · compressed size)
```

The following formula is used to calculate the percentage of capacity that is saved:

```
((written space · compressed size) ÷ written space) × 100
```

Exception: For compressed volumes that are also deduplicated, on storage systems that run IBM Spectrum Virtualize (for example IBM FlashSystem 9100), this column is blank.

Enterprise HDD Capacity (GiB)

The total amount of storage space on the Enterprise hard disk drive that the Easy Tier volume uses for re-tiering the volume extents.

Nearline HDD Capacity (GiB)

The total amount of storage space on the Nearline hard disk drive that the Easy Tier volume uses for re-tiering the volume extents.

SSD Capacity (GiB)

The total amount of storage space on the solid-state drive that the Easy Tier volume uses for re-tiering the volume extents.

Tier 0 Flash Capacity (GiB)

The total amount of storage space on the Tier 0 flash solid-state drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Tier 1 Flash Capacity (GiB)

The total amount of storage space on the Tier 1 flash, read-intensive solid-state drives that can be used by Easy Tier for re-tiering the volume extents in the pool.

Used Space (GiB)

The amount of allocated space that is used by the compressed, thin-provisioned, or Easy Tier volume.

For compressed and thin-provisioned volumes, used space might not be the same as allocated space (for example, IBM FlashSystem 9100, Storwize, IBM SAN Volume Controller).

For thin-provisioned volumes, used space might not be the same as allocated space because you can preallocate space to the thin-provisioned volumes when the volumes are created. For compressed volumes, used space might not be the same as allocated space because more space is used to read the data than is used to write the data to the disk. For regular volumes that use Easy Tier on the storage systems that are listed, used space is the same as allocated space.

Written Space (GiB)

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 space 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 just 10 GiB.

8.4 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 Pro. Keep in mind that the free version of Storage Insights does not support alerting.

New data reduction technologies add more intelligence and capacity savings to your environment. If you use data reduction on different layers, such as hardware compression in the IBM FlashSystem 9100 Flash Core Modules and additionally in the Data Reduction Pools, or if you virtualize an IBM FlashSystem 9100 system) you will have to pay more attention in preventing insufficient space remaining in the back-end storage device.

First it’s 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 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).

Over provisioning means, that in total more space is being assigned and promised to the hosts. They can possibly try to store more data on the storage subsystem, as physical capacity is available. This will result in an out-of-space condition.

> **Note:** It is extremely important to avoid this situation and to monitoring is very important!
>
> Its also important to keep free space for Garbage collection in the background , "DRP internal details" on page 60 and Chapter 4, "Storage Pools" on page 55.

Data Reduction technologies will give back some space. If the space that's used for the data can be reduced, the saved up space can be used for other data. But keep in mind that, depending on the type of data, deleting might not result in freeing up much space.

Just imagine if you have three identical/almost identical files on a file system which have been deduplicated. This resulted in getting a quite good compression ratio (3 files - but stored only once). If you now delete one file, you would not gain more space, because the deduplicated data has to stay on the storage, because two other versions see the data. Similar results might be seen when using several FlashCopies of one source.

8.4.1 Alert examples

Table 8-3 shows alerts for IBM FlashSystem 9100 systems based on Array or Pool level.

Table 8-3 Event examples for IBM Flash System 9100

System	Entity	Resource Type	Event
FS9100 with FCM	Array	Usable capacity	Available Physical Space <= nn% (* Example shown in "Alert to monitor back-end capacity: Available Physical Space (%)")
	Pool	Efficient Capacity	Physical allocation >= nn%
FS9100 other media	Pool	Usable Capacity	Physical allocation >= nn%

Other alerts are possible as well, but generally % alerts are best suited, as the alert definition applies to all pool in a storage system.

8.4.2 Alert to monitor back-end capacity: Available Physical Space (%)

In this part of the book we show how to deploy IBM Spectrum Control or IBM Storage Insights Pro in order 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 9100.

Firstly, assign a severity to an alert. Assigning a severity can help you more 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 8-4 shows the possible alert severities.

Table 8-4 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% space in the RAID Array left)
- Warning (20% space in the RAID Array left)
- Information (30% space in the RAID Array left)

Adjust the percentage levels to the required levels as needed. Keep in mind that the process to extend storage might take its time (ordering process, installation, provisioning, and so on).

The advantage of this way to setup an **Alert Policy** is, that you can add various IBM FlashSystem 9100 (or even other storage subsystem such as the IBM FlashSystem 900) to this customized alert.

Figure 8-23 shows how to start creating a new Alert Policy which monitors the remaining free capacity in the RAID Array. The customized Alert is not tied to the Data Reduction Pool, as it works regardless of the pool type.

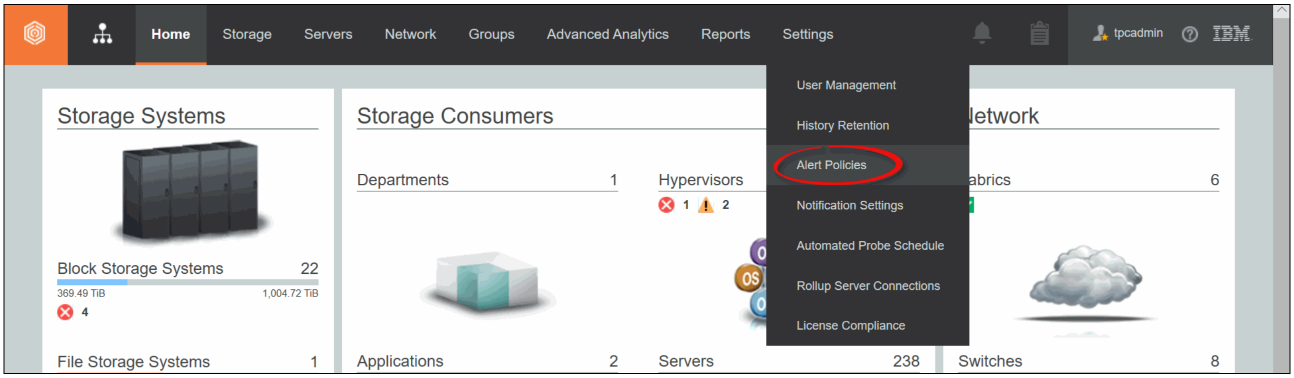
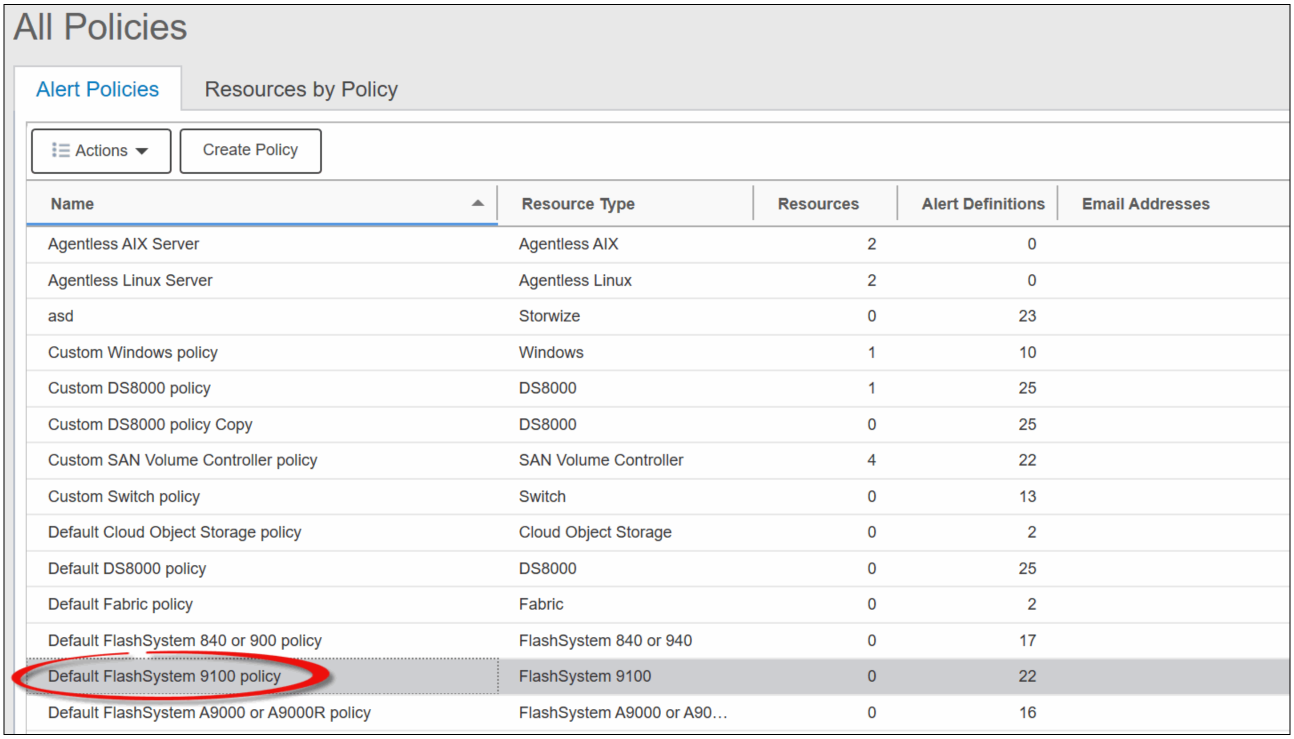


Figure 8-23 Create new Alert Policy

In the following example we create a new Alert Policy by copying the existing one. You might change an existing Alert Policy (in our example the Default Policy) as well. Keep in mind, a storage subsystem can be active in only one Alert Policy.

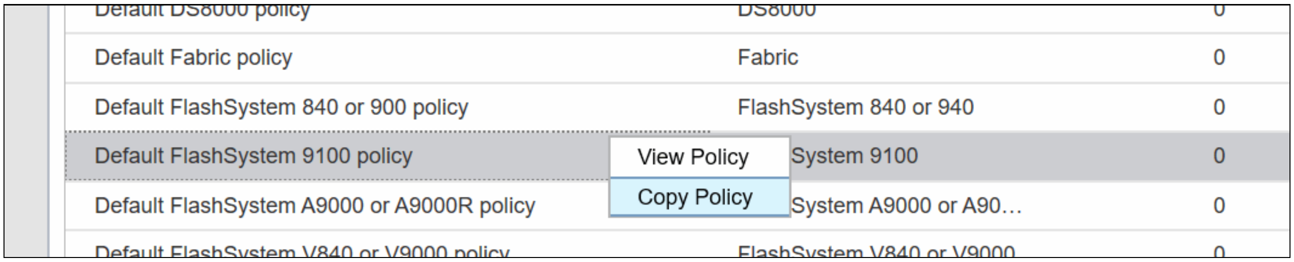
Figure 8-24 shows the Default Policy of the IBM FlashSystem 9100.



Name	Resource Type	Resources	Alert Definitions	Email Addresses
Agentless AIX Server	Agentless AIX	2	0	
Agentless Linux Server	Agentless Linux	2	0	
asd	Storwize	0	23	
Custom Windows policy	Windows	1	10	
Custom DS8000 policy	DS8000	1	25	
Custom DS8000 policy Copy	DS8000	0	25	
Custom SAN Volume Controller policy	SAN Volume Controller	4	22	
Custom Switch policy	Switch	0	13	
Default Cloud Object Storage policy	Cloud Object Storage	0	2	
Default DS8000 policy	DS8000	0	25	
Default Fabric policy	Fabric	0	2	
Default FlashSystem 840 or 900 policy	FlashSystem 840 or 940	0	17	
Default FlashSystem 9100 policy	FlashSystem 9100	0	22	
Default FlashSystem A9000 or A9000R policy	FlashSystem A9000 or A90…	0	16	

Figure 8-24 Default Alert Policy

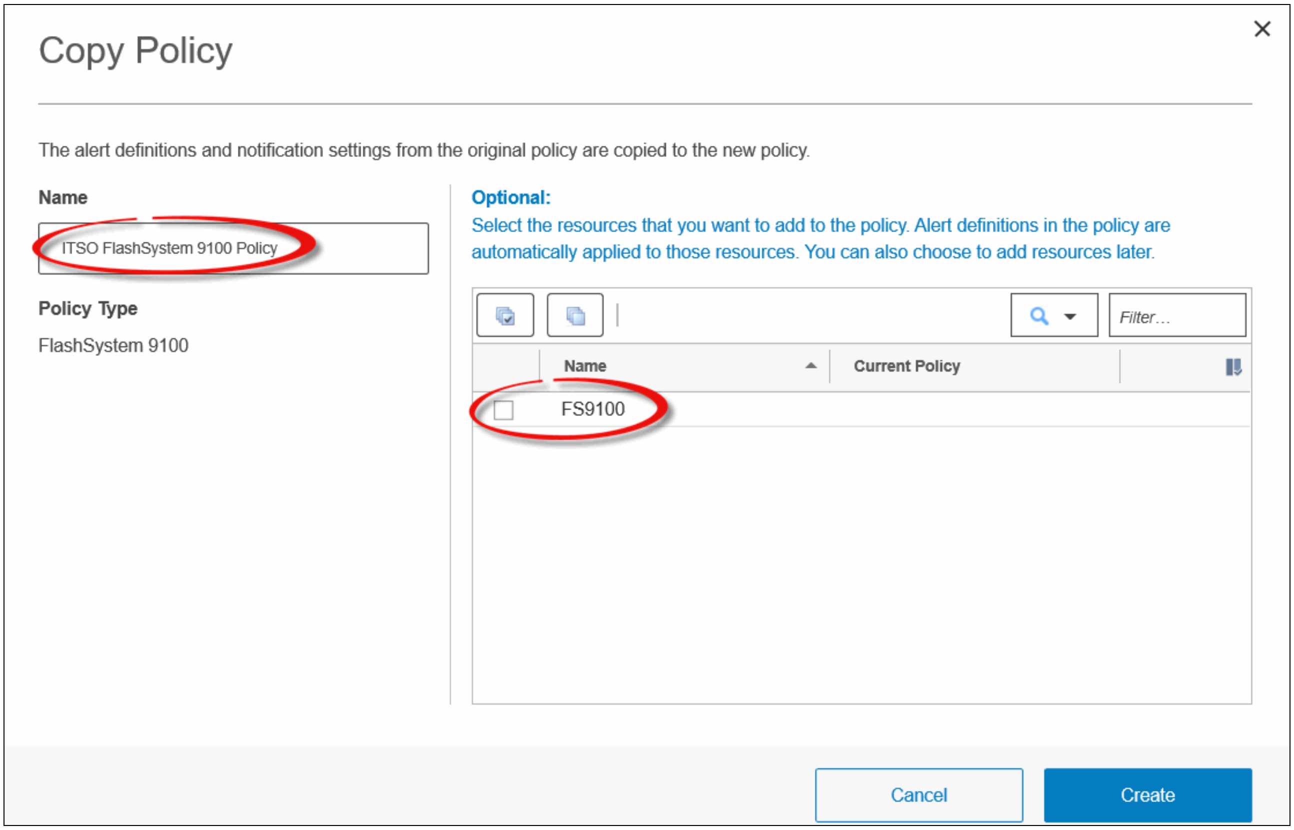
Figure 8-25 describes how to copy an existing Policy in order to create a new one. Hover the mouse pointer over the existing Policy that you want to copy, then click it and choose **Copy Policy**.



Default DS8000 policy	DS8000	0
Default Fabric policy	Fabric	0
Default FlashSystem 840 or 900 policy	FlashSystem 840 or 940	0
Default FlashSystem 9100 policy	System 9100	0
Default FlashSystem A9000 or A9000R policy	System A9000 or A90…	0
Default FlashSystem V840 or V9000 policy	FlashSystem V840 or V9000	0

Figure 8-25 Copy existing Policy and create a new one

Figure 8-26 on page 306 shows how to rename the previously copied Policy. The new Policy will be stored as an additional Policy. Keep in mind that one IBM FlashSystem 9100 can only be added to a single policy. You can add the system (choose **Optional**, select the resource, and select the check box) later if you are not sure at this time.



Copy Policy

The alert definitions and notification settings from the original policy are copied to the new policy.

Name

ITSO FlashSystem 9100 Policy

Policy Type

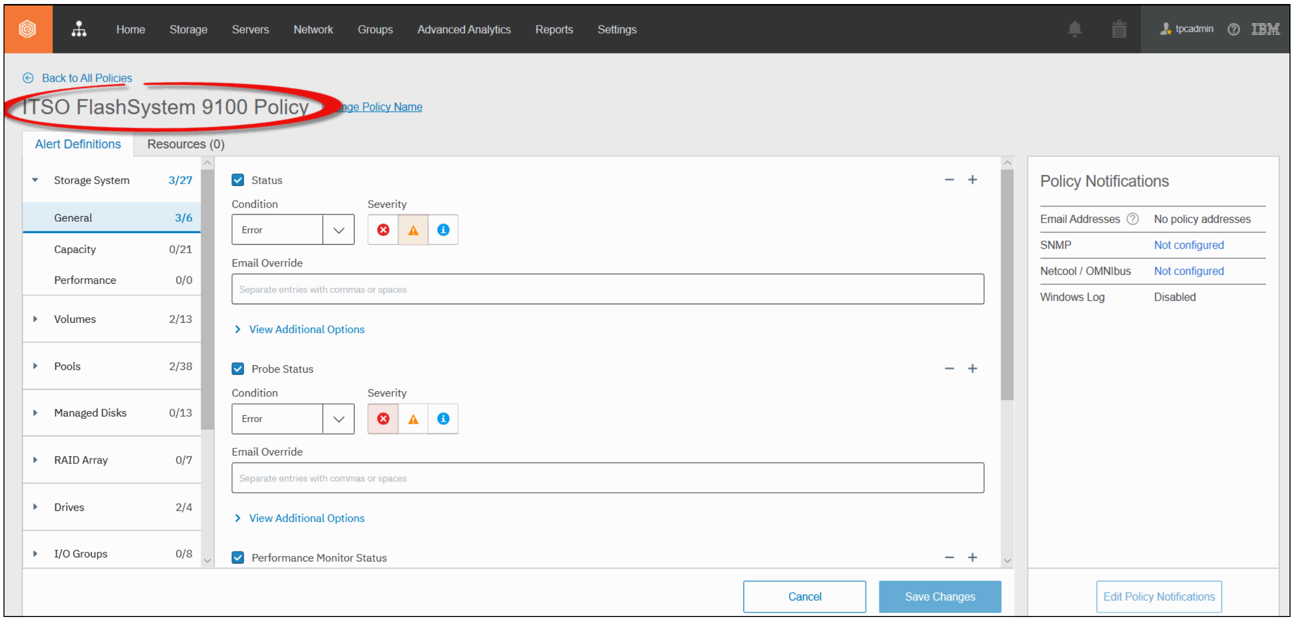
FlashSystem 9100

Optional:
Select the resources that you want to add to the policy. Alert definitions in the policy are automatically applied to those resources. You can also choose to add resources later.

	Name	Current Policy
□	FS9100	

Figure 8-26 Store copied policy

Figure 8-27 shows the newly created Alert Policy "ITSO FlashSystem 9100 Policy" with all the existing alerts inherited Alert Definitions from the Default Policy.



ITSO FlashSystem 9100 Policy nge Policy Name

Alert Definitions | Resources (0)

Storage System	3/27
General	3/6
Capacity	0/21
Performance	0/0
Volumes	2/13
Pools	2/38
Managed Disks	0/13
RAID Array	0/7
Drives	2/4
I/O Groups	0/8

☑ Status
Condition: Error
Severity
Email Override: Separate entries with commas or spaces
View Additional Options

☑ Probe Status
Condition: Error
Severity
Email Override: Separate entries with commas or spaces
View Additional Options

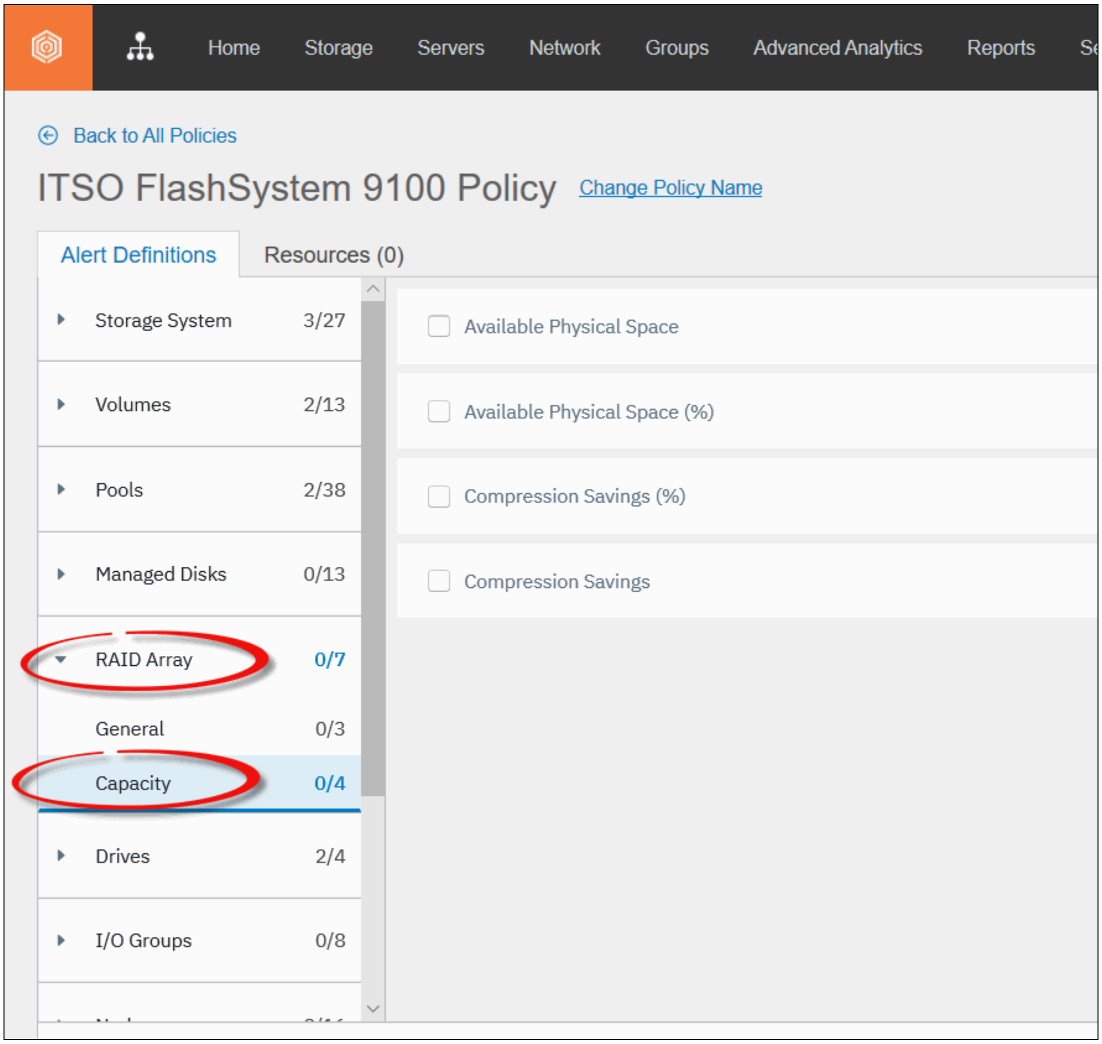
☑ Performance Monitor Status

Policy Notifications

Email Addresses	No policy addresses
SNMP	Not configured
Netcool / OMNIbus	Not configured
Windows Log	Disabled

Figure 8-27 New Policy with inherited Alert Definitions

Figure 8-28 shows how to choose the required Alert Definitions **RAID Array** → **Capacity** in the screen.



ITSO FlashSystem 9100 Policy

Storage System	3/27
Volumes	2/13
Pools	2/38
Managed Disks	0/13
RAID Array	0/7
General	0/3
Capacity	0/4
Drives	2/4
I/O Groups	0/8

☐ Available Physical Space

☐ Available Physical Space (%)

☐ Compression Savings (%)

☐ Compression Savings

Figure 8-28 Alert Definition RAID Array > Capacity

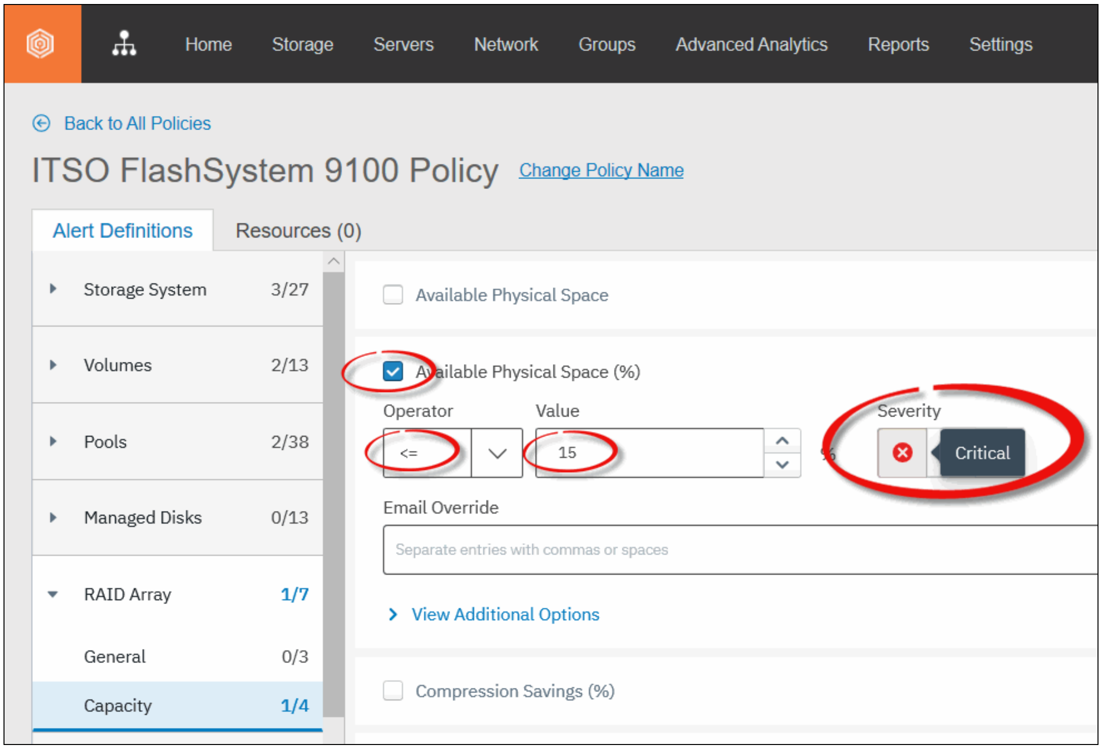
Figure 8-29 on page 308 denotes the tasks for setting up the Critical definition by monitoring the Available Physical Space (%) and releasing Policy Notifications at 15%. This example implies that when 85% or more physical space is taken a critical Notification via the predefined method will be sent.

Predefined methods can be:

- Email Addresses
- SNMP
- IBM Netcool® / OMNIbus
- Windows Event Log or UNIX syslog

These methods have to be defined before you can choose them. If your environment does not have predefined methods see Figure 8-29.

Note: With IBM Storage Insights, you can only send emails.



ITSO FlashSystem 9100 Policy

Storage System	3/27
Volumes	2/13
Pools	2/38
Managed Disks	0/13
RAID Array	1/7
General	0/3
Capacity	1/4

□ Available Physical Space

☑ Available Physical Space (%)

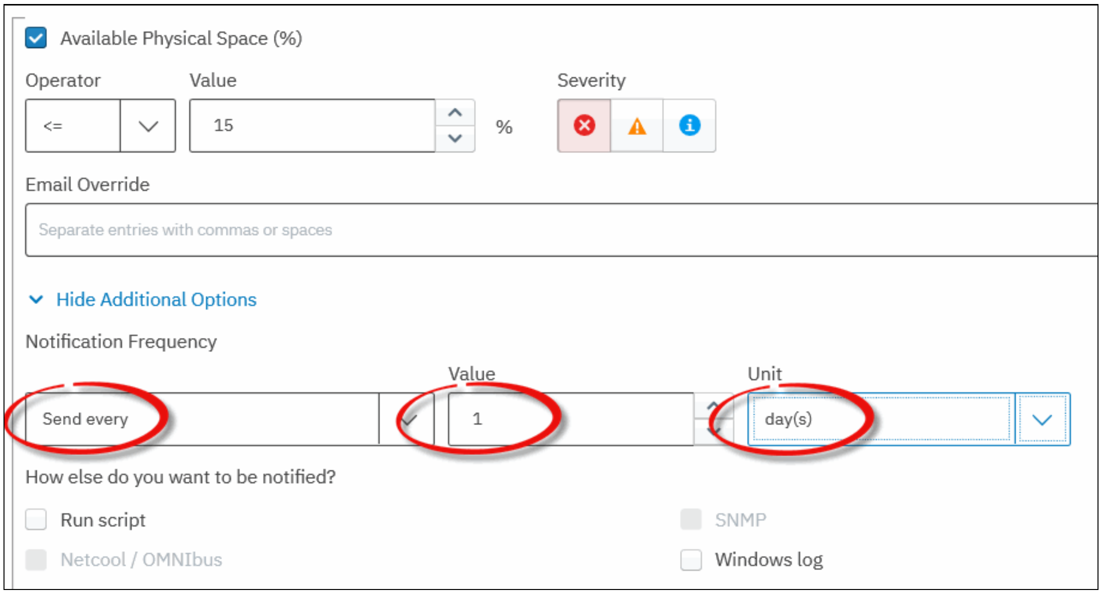
Operator: <=
Value: 15 %
Severity: Critical

Email Override: Separate entries with commas or spaces

□ Compression Savings (%)

Figure 8-29 Alert Definition 15% or less Available Physical Space - Critical

Figure 8-30 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 Space”. In this example we choose to set the frequency to “every day”.



☑ Available Physical Space (%)

Operator: <=
Value: 15 %
Severity

Email Override: Separate entries with commas or spaces

Notification Frequency

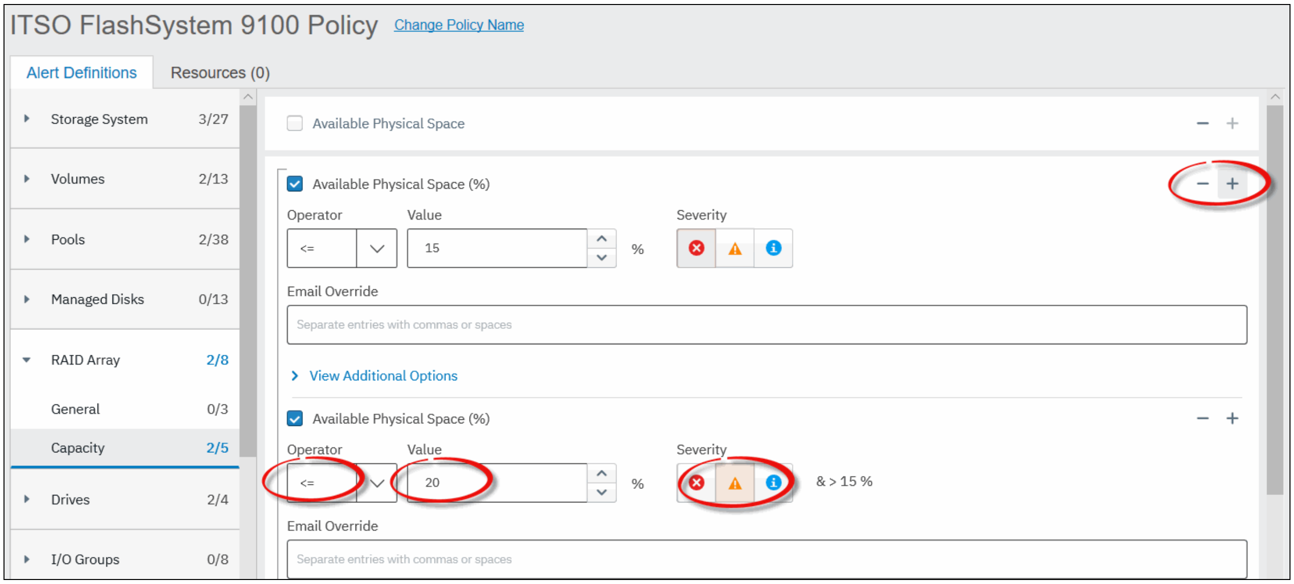
Send every
Value: 1
Unit: day(s)

How else do you want to be notified?

□ Run script
□ SNMP
□ Netcool / OMNIbus
□ Windows log

Figure 8-30 Alert Definition - Critical - change frequency

Figure 8-31 shows how to set up the Warning level at 20% or less Available Physical Space. To proceed, choose the plus sign at the previously defined Definition (Critical) and make the following selections, (Operator: **<=**, Value: **20%**, and Severity: **Warning**).



ITSO FlashSystem 9100 Policy Change Policy Name

Alert Definitions | Resources (0)

Storage System	3/27
Volumes	2/13
Pools	2/38
Managed Disks	0/13
RAID Array	2/8
General	0/3
Capacity	2/5
Drives	2/4
I/O Groups	0/8

□ Available Physical Space

☑ Available Physical Space (%)

Operator: <= Value: 15 % Severity: Critical

Email Override: Separate entries with commas or spaces

View Additional Options

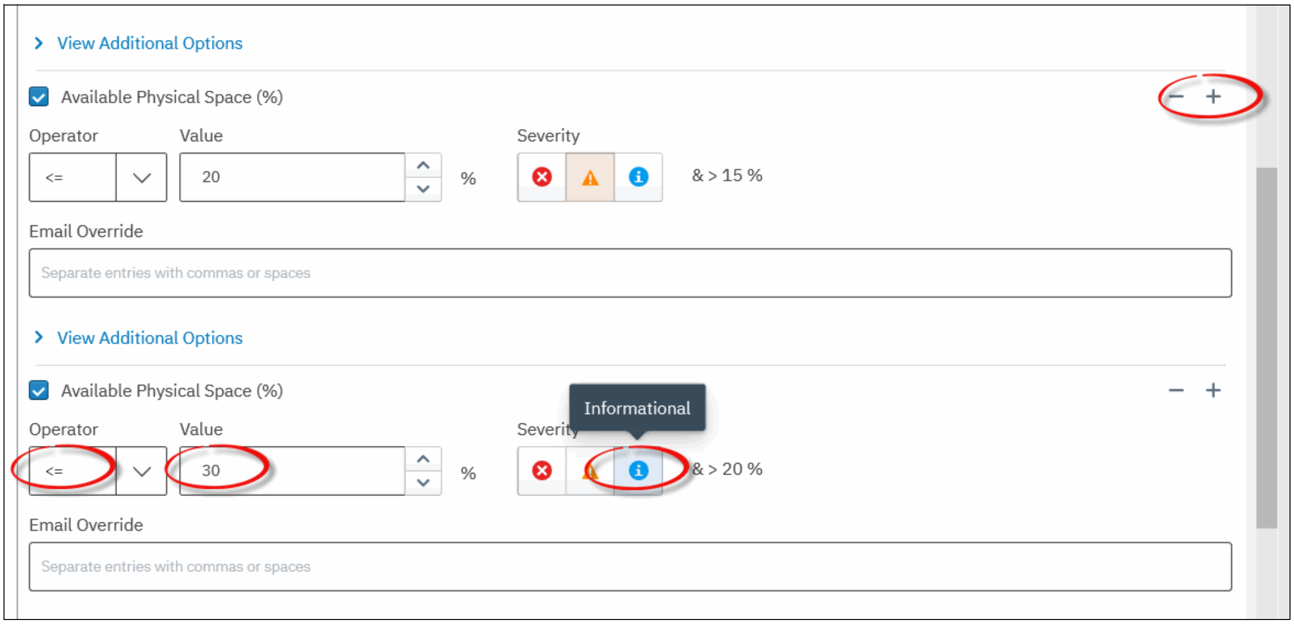
☑ Available Physical Space (%)

Operator: <= Value: 20 % Severity: Warning & > 15 %

Email Override: Separate entries with commas or spaces

Figure 8-31 Alert Definition 20% or less Available Physical Space - Warning

Figure 8-32 depicts how to set up the Notification Threshold at 30%.



View Additional Options

☑ Available Physical Space (%)

Operator: <= Value: 20 % Severity: Warning & > 15 %

Email Override: Separate entries with commas or spaces

View Additional Options

☑ Available Physical Space (%)

Operator: <= Value: 30 % Severity: Informational & > 20 %

Email Override: Separate entries with commas or spaces

Figure 8-32 Alert Definition 30% or less Available Physical Space - Notification

Figure 8-33 shows how to configure the Notification Settings in your monitoring environment.

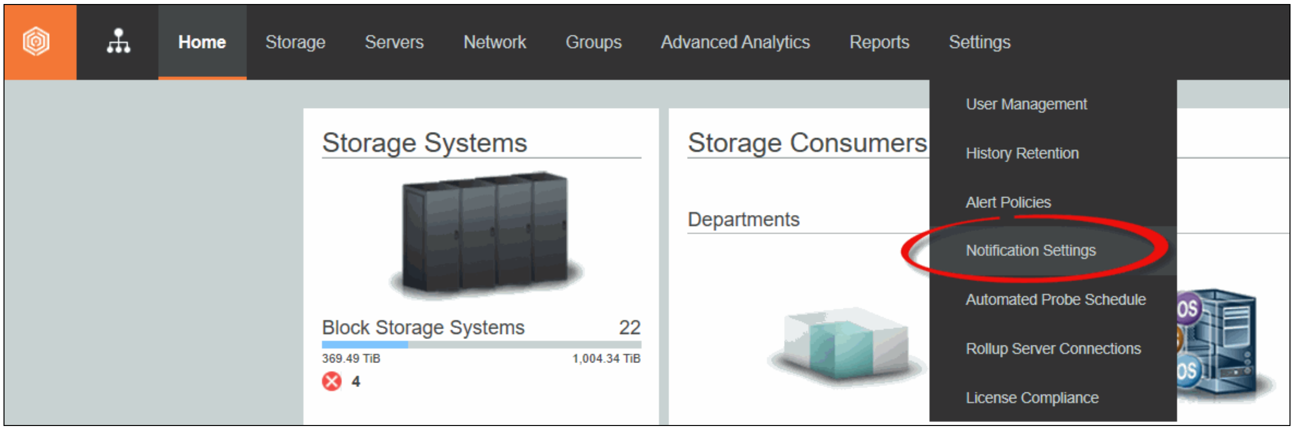
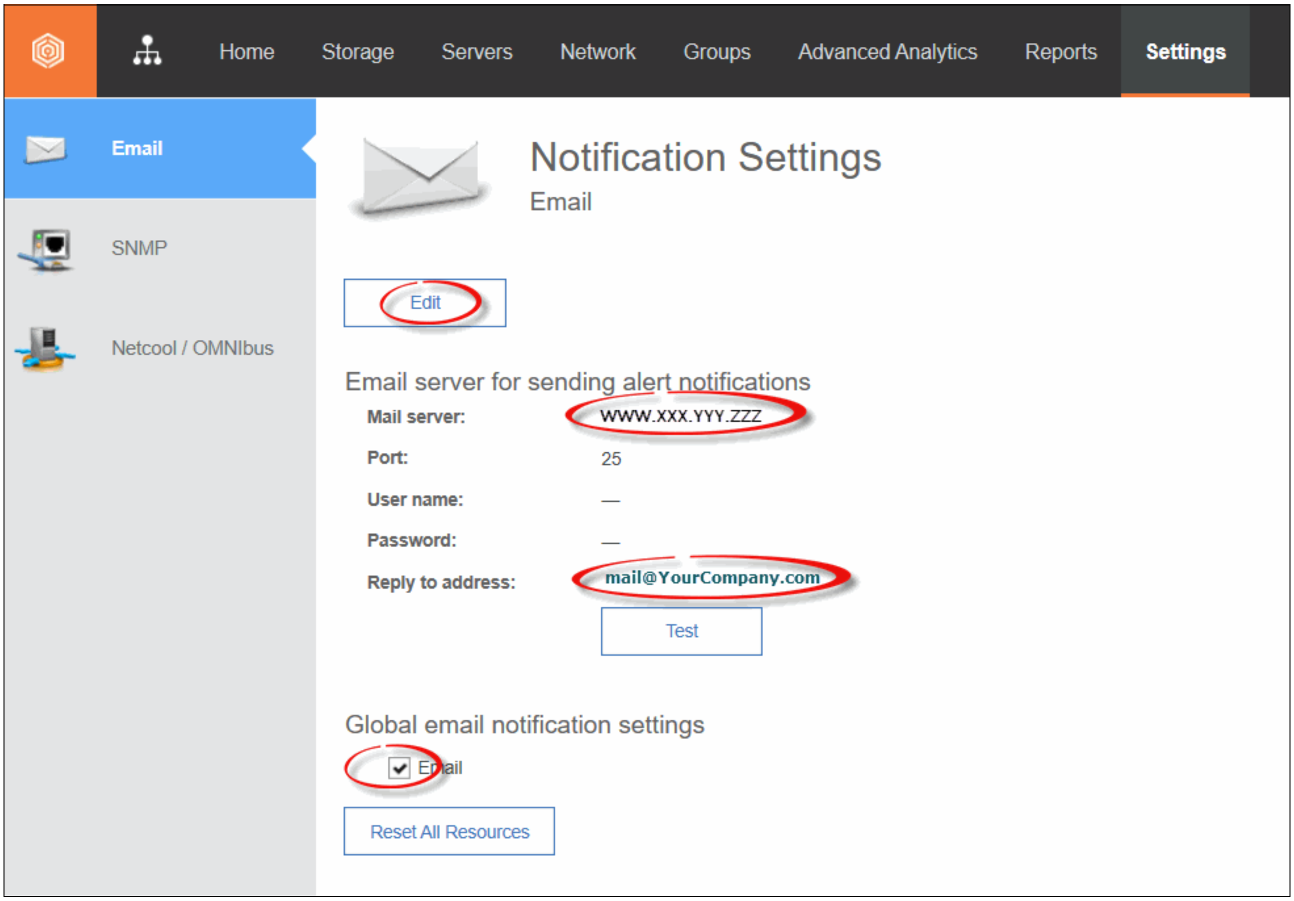


Figure 8-33 Change Notification Settings

Figure 8-34 shows how to set up the Notification Settings.



Email server for sending alert notifications

Mail server: www.xxx.yyy.zzz

Port: 25

User name: —

Password: —

Reply to address: mail@YourCompany.com

Global email notification settings

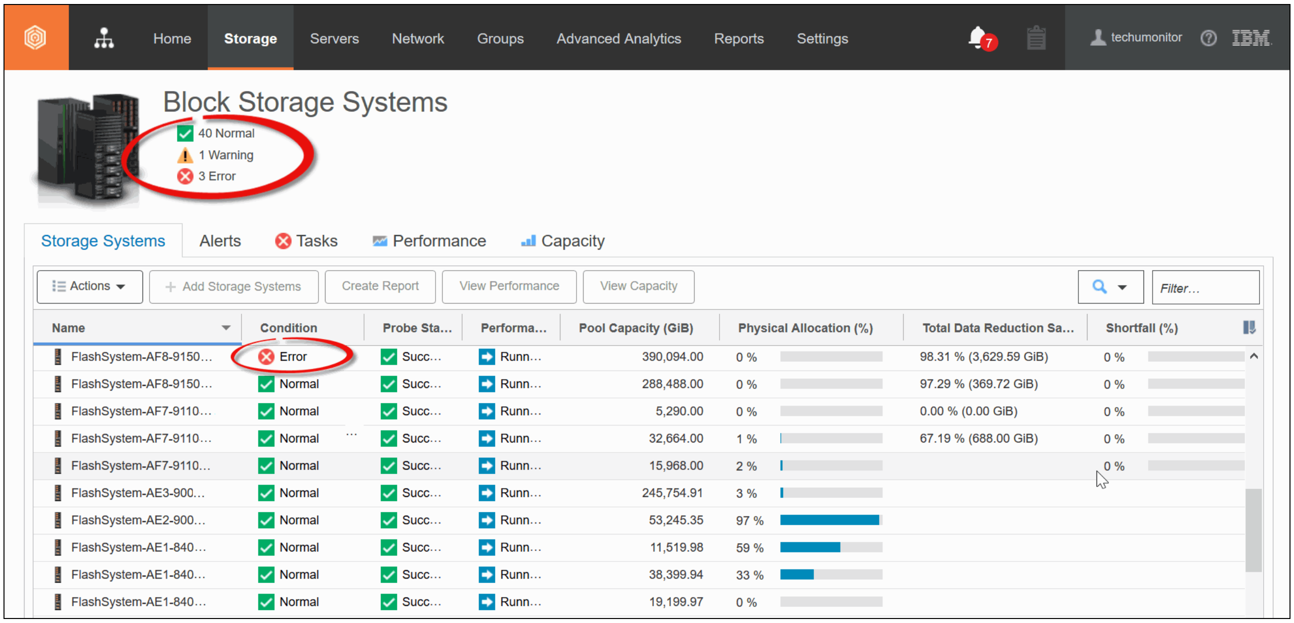
☑ Email

Figure 8-34 Notification Settings: Details

8.5 Error condition example with IBM Spectrum Control: FC port

The following guidance shows an example of an FC port problem. It shows how you can spot an error with IBM Spectrum Control and shows how to drill down into the details.

Figure 8-35 denotes a testing environment with several storage subsystems. Three errors can be spotted in the GUI. In this example, we show how you can get more details about the first highlighted error.



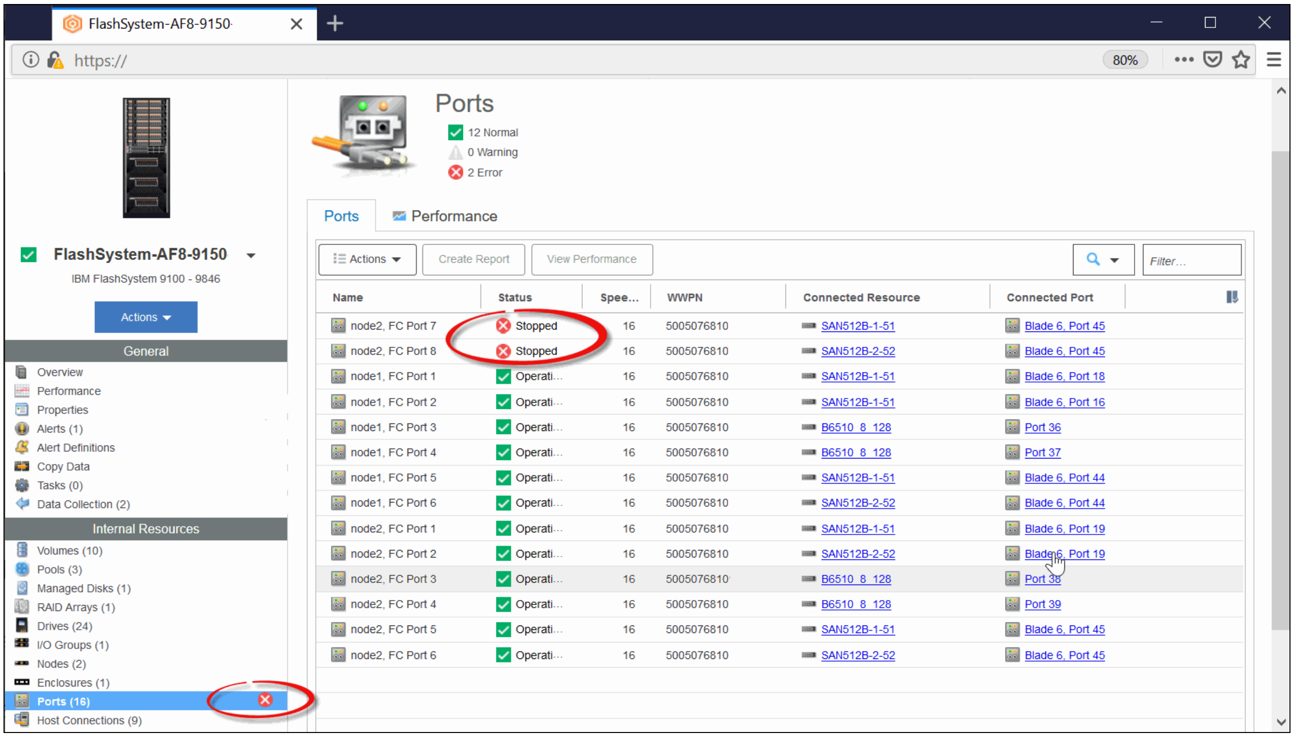
Block Storage Systems

- ✓ 40 Normal
- ⚠ 1 Warning
- ✖ 3 Error

Name	Condition	Probe Sta...	Performa...	Pool Capacity (GiB)	Physical Allocation (%)	Total Data Reduction Sa...	Shortfall (%)
FlashSystem-AF8-9150...	✖ Error	✓ Succ...	Runn...	390,094.00	0 %	98.31 % (3,629.59 GiB)	0 %
FlashSystem-AF8-9150...	✓ Normal	✓ Succ...	Runn...	288,488.00	0 %	97.29 % (369.72 GiB)	0 %
FlashSystem-AF7-9110...	✓ Normal	✓ Succ...	Runn...	5,290.00	0 %	0.00 % (0.00 GiB)	0 %
FlashSystem-AF7-9110...	✓ Normal	✓ Succ...	Runn...	32,664.00	1 %	67.19 % (688.00 GiB)	0 %
FlashSystem-AF7-9110...	✓ Normal	✓ Succ...	Runn...	15,968.00	2 %		0 %
FlashSystem-AE3-900...	✓ Normal	✓ Succ...	Runn...	245,754.91	3 %		
FlashSystem-AE2-900...	✓ Normal	✓ Succ...	Runn...	53,245.35	97 %		
FlashSystem-AE1-840...	✓ Normal	✓ Succ...	Runn...	11,519.98	59 %		
FlashSystem-AE1-840...	✓ Normal	✓ Succ...	Runn...	38,399.94	33 %		
FlashSystem-AE1-840...	✓ Normal	✓ Succ...	Runn...	19,199.97	0 %		

Figure 8-35 Error condition spotted on a storage subsystem with IBM Spectrum Control

Figure 8-36 shows details of the storage subsystem and which entity is affected. In this case it is related to internal resources: ports. Two ports have been stopped and have caused this condition in the environment.



FlashSystem-AF8-9150

https://

FlashSystem-AF8-9150

IBM FlashSystem 9100 - 9846

Actions

General

- Overview
- Performance
- Properties
- Alerts (1)
- Alert Definitions
- Copy Data
- Tasks (0)
- Data Collection (2)

Internal Resources

- Volumes (10)
- Pools (3)
- Managed Disks (1)
- RAID Arrays (1)
- Drives (24)
- I/O Groups (1)
- Nodes (2)
- Enclosures (1)
- Ports (16) ☒
- Host Connections (9)

Ports

- ✓ 12 Normal
- 0 Warning
- ☒ 2 Error

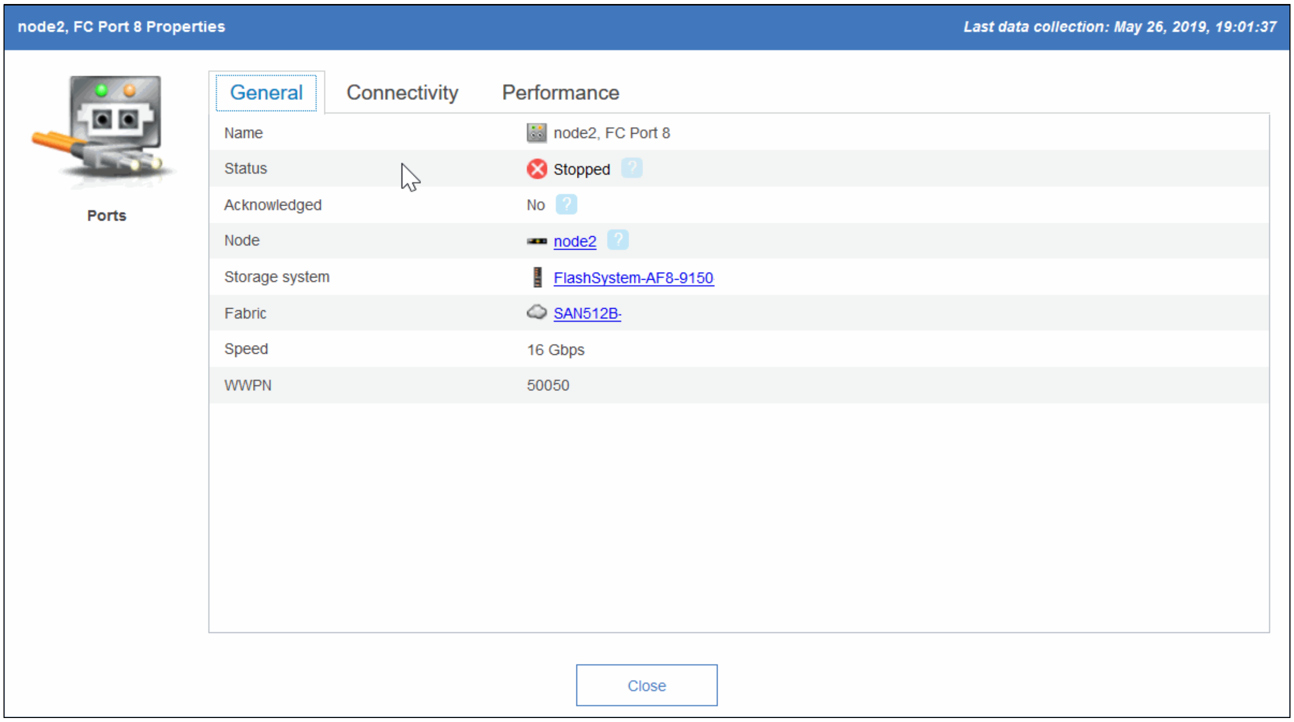
Ports | Performance

Actions | Create Report | View Performance | Filter...

Name	Status	Spee...	WWPN	Connected Resource	Connected Port
node2, FC Port 7	☒ Stopped	16	5005076810	SAN512B-1-51	Blade 6, Port 45
node2, FC Port 8	☒ Stopped	16	5005076810	SAN512B-2-52	Blade 6, Port 45
node1, FC Port 1	✓ Operati...	16	5005076810	SAN512B-1-51	Blade 6, Port 18
node1, FC Port 2	✓ Operati...	16	5005076810	SAN512B-1-51	Blade 6, Port 16
node1, FC Port 3	✓ Operati...	16	5005076810	B6510_8_128	Port 36
node1, FC Port 4	✓ Operati...	16	5005076810	B6510_8_128	Port 37
node1, FC Port 5	✓ Operati...	16	5005076810	SAN512B-1-51	Blade 6, Port 44
node1, FC Port 6	✓ Operati...	16	5005076810	SAN512B-2-52	Blade 6, Port 44
node2, FC Port 1	✓ Operati...	16	5005076810	SAN512B-1-51	Blade 6, Port 19
node2, FC Port 2	✓ Operati...	16	5005076810	SAN512B-2-52	Blade 6, Port 19
node2, FC Port 3	✓ Operati...	16	5005076810	B6510_8_128	Port 38
node2, FC Port 4	✓ Operati...	16	5005076810	B6510_8_128	Port 39
node2, FC Port 5	✓ Operati...	16	5005076810	SAN512B-1-51	Blade 6, Port 45
node2, FC Port 6	✓ Operati...	16	5005076810	SAN512B-2-52	Blade 6, Port 45

Figure 8-36   FlashSystem error condition - internal resources - ports

Figure 8-37 shows the details of one of the stopped ports.



node2, FC Port 8 Properties

Last data collection: May 26, 2019, 19:01:37

Ports

General | Connectivity | Performance

Name	node2, FC Port 8
Status	☒ Stopped
Acknowledged	No
Node	node2
Storage system	FlashSystem-AF8-9150
Fabric	SAN512B-
Speed	16 Gbps
WWPN	50050

Close

Figure 8-37   FC ports stopped: detail view

The ports are probably stopped for a reason, so from the panel in Figure 8-36 on page 312 select both, click **Actions** and select **Acknowledge Status**. The ports will still be shown with the red icon, but now that is overlaid by a check mark, and after a short time this change will be propagated so that the storage system is shown as being in a green status again.

In other cases you might have to replace hardware, and after you have opened a ticket in your internal system with the vendor, you should still acknowledge the status, so that any other errors will make the storage system go from green to red again and you see that a second event has happened.

8.6 Important metrics

The following metrics are some of the most important metrics that need to be analyzed to understand a performance problem in IBM FlashSystem 9100 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 moved from a host to a specific storage device. Typically measured in MB per second.
- **Response time R/W**: This is the time taken for a circuit or measuring device, when subjected to a change in input signal, to change its state by a specified fraction of its total response to that change. In case of Disk or Storage Subsystem, this is the time used to complete an I/O operation. Typically measured in ms.
- **Cache Hit R/W**: This is the percentage of times where 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 number of data in MBps for operations in which the port receives or sends data.

- **Port Protocol Errors (Zero Buffer Credit Timer)**: The number of microseconds that the port is not able to send frames between ports because there is insufficient buffer-to-buffer credit. In Fibre Channel technology, buffer-to-buffer credit is used to control the flow of frames between ports. Buffer-to-buffer credit is measured from the last time that the node was reset. This value is related to the data collection sample interval.
- **Port Congestion Index:** The estimated degree to which frame transmission was delayed due to a lack of buffer credits. This value is generally 0 - 100. The value 0 means there was no congestion. The value can exceed 100 if the buffer credit exhaustion persisted for an extended amount of time. When you troubleshoot a SAN, use this metric to help identify port conditions that might slow the performance of the resources to which those ports are connected.
- **Global Mirror (Overlapping Write Percentage)**: The percentage of overlapping write operations that are issued by the Global Mirror primary site. Some overlapping writes are processed in parallel, and so they are excluded from this value.
- **Global Mirror (Write I/O Rate)**: The average number of write operations per second that are issued to the Global Mirror secondary site. Keep in mind that IBM FlashSystem 9100 systems have 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 is also a very important metric, which should be used in conjunction with IBM Spectrum Control V5.3.3 or higher. Previous versions had a calculation error.
>
> V5.2.x version is not supported after September 30th 2019.

Many others metrics are supplied to IBM Spectrum Control from IBM FlashSystem 9100. For more information about all metrics, see:

Performance metrics for resources that run IBM Spectrum Virtualize

8.7 Performance support package

If you have performance issues on your system at any level (Host, Volume, Node, Pools, and so on), consult IBM Support, who require detailed performance data about the IBM FlashSystem 9100 system to diagnose the problem. Generate a performance support package with detailed data by using IBM Spectrum Control.

In this scenario, you export performance data for a IBM FlashSystem 9100 to a compressed package, and you then send the package to IBM Support, as shown in Figure 8-38.

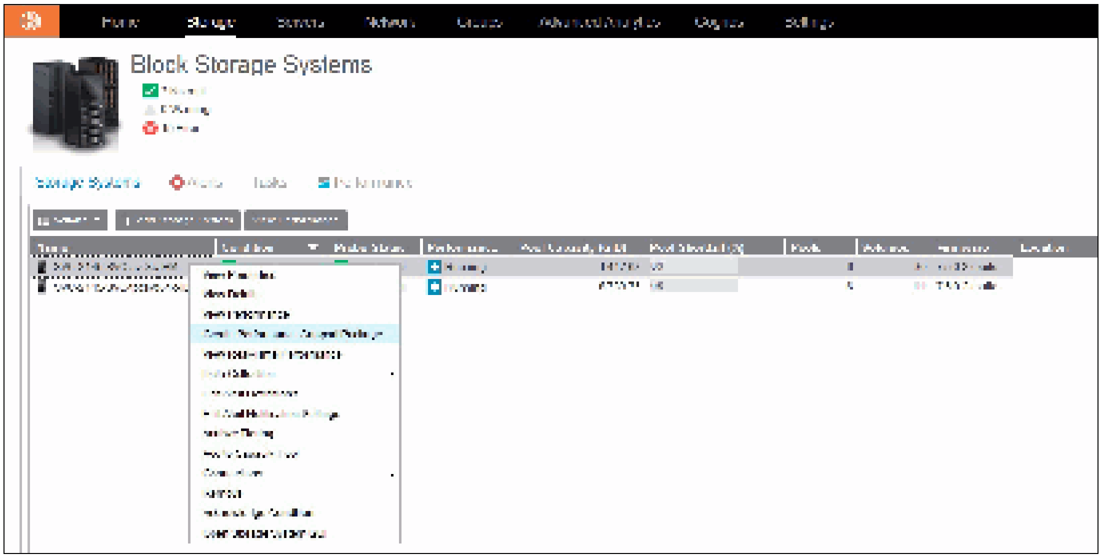
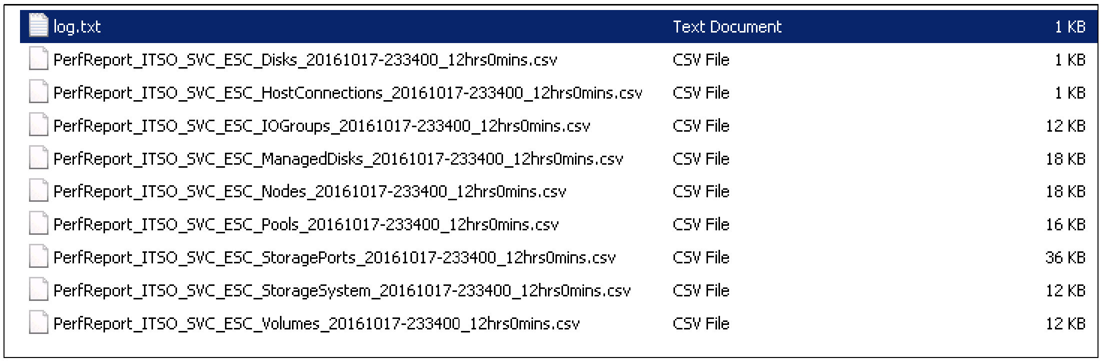


Figure 8-38 Performance support package creation

When the package has been created, you are requested to download it in `.zip` format. The package includes different reports in `.csv` format, as shown in Figure 8-39.



log.txt	Text Document	1 KB
PerfReport_ITSO_SVC_ESC_Disks_20161017-233400_12hrs0mins.csv	CSV File	1 KB
PerfReport_ITSO_SVC_ESC_HostConnections_20161017-233400_12hrs0mins.csv	CSV File	1 KB
PerfReport_ITSO_SVC_ESC_IOGroups_20161017-233400_12hrs0mins.csv	CSV File	12 KB
PerfReport_ITSO_SVC_ESC_ManagedDisks_20161017-233400_12hrs0mins.csv	CSV File	18 KB
PerfReport_ITSO_SVC_ESC_Nodes_20161017-233400_12hrs0mins.csv	CSV File	18 KB
PerfReport_ITSO_SVC_ESC_Pools_20161017-233400_12hrs0mins.csv	CSV File	16 KB
PerfReport_ITSO_SVC_ESC_StoragePorts_20161017-233400_12hrs0mins.csv	CSV File	36 KB
PerfReport_ITSO_SVC_ESC_StorageSystem_20161017-233400_12hrs0mins.csv	CSV File	12 KB
PerfReport_ITSO_SVC_ESC_Volumes_20161017-233400_12hrs0mins.csv	CSV File	12 KB

Figure 8-39 Package files example

For more information about how to create a performance support package, see:

`Exporting performance data for a SAN Volume Controller system`

Note: The performance data might be large, especially if the data is for storage systems that have many volumes, or the performance monitors are running with a 1-minute sampling frequency. If the time range for the data is greater than 12 hours, volume data and 1-minute sample data is automatically excluded from the performance data. To include volume data and 1-minute sample data, select the **Advanced package** option on the Create Performance Support Package wizard.

8.8 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 9100 systems to configure, manage, and monitor data-copy functions. Copy services include IBM FlashCopy, Metro Mirror, Global Mirror, and Global Mirror Change Volumes.

You can use Copy Services Manager to complete the following data replication tasks and help reduce the downtime of critical applications:

- Plan for replication when you are provisioning storage
- Keep data on multiple related volumes consistent across storage systems if there is a planned or unplanned outage
- Monitor and track replication operations
- Automate the mapping of source volumes to target volumes

One of the most important events that needs to be monitored when IBM FlashSystem 9100 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 9100 systems are able to suspend the MM or GM relationship to protect the performance on the primary site when MM or GM starts to affect write response time. That suspension can be caused by several factors. IBM FlashSystem 9100 systems *do not restart MM or GM automatically*. They must be restarted manually.

Setting IBM FlashSystem 9100 systems alert monitoring is explained in 8.1.1, “Monitoring with the GUI” on page 268. When MM or GM is managed by IBM CSM and if a 1920 error occurs, IBM CSM can automatically restart MM or GM sessions, and can set the delay time on the automatic restart option. This delay allows some time for the situation to correct itself.

Alternatively, if you have several sessions, you can stagger them so that they do not all restart at the same time, which can affect 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.

For more information about IBM Copy Service Manager, see:

`Metro Mirror and Global Mirror Failover/Failback with Practice session properties`

8.8.1 Monitoring MM and GM with scripts

IBM FlashSystem 9100 system provides a complete command-line interface (CLI), which allows you to interact with your systems by using scripts. Those scripts can run in the IBM FlashSystem 9100 shell, but with a limited script command set available, or they can run out of the shell using any scripting language that you prefer.

An example of script usage is one to check at a specific interval time whether MM or GM are still active, if any 1920 errors have occurred, or to react to an SNMP or email alert received. The script can then start some specific recovery action based on your recovery plan and environment.

Customers who do not use IBM Copy Service Manager have created their own scripts. These scripts are sometimes supported by IBM as part of ITS professional services or IBM System Lab services. Tell your IBM representative what kind of monitoring you want to implement with scripts, and together try to find if one exists in the IBM Intellectual Capital Management repository that can be reused.

8.9 Monitoring Tier1 SSD

s paid to the endurance events that can be triggered. For monitoring purposes, stay alert to the new fields listed in Table 8-5.

Table 8-5 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 written across the life of the drive divided by the anticipated amount (2.42 PB for the 15.36 TB drive) Starts at 0, and can continue > 100
write_endurance_usage_rate	Measuring / Low / Marginal / High Takes 160 Days to get initial measurement; Low: Approximately 5.5 Years or more Marginal: Approximately 4.5 – 5.5 Years High: Approximately < 4.5 years High triggers event `SS_EID_VL_ER_SSD_WRITE_ENDURANCE_USAGE_RATE_HIGH`
replacement_date	The Current Date + Endurance Rate * Remaining Endurance Triggers event `SS_EID_VL_ER_SSD_DRIVE_WRITE_ENDURANCE_LIMITED` at 6 Months before limit

If you see either of these triggered events, contact your IBM service representative to put an action plan in place:

```
SS_EID_VL_ER_SSD_WRITE_ENDURANCE_USAGE_RATE_HI4GH
SS_EID_VL_ER_SSD_DRIVE_WRITE_ENDURANCE_LIMITED
```

9

Maintenance

Among the many benefits that the IBM FlashSystem 9100 software provides is to greatly simplify the storage management tasks that system administrators need to perform. However, as the IT environment grows and gets renewed, so does the storage infrastructure.

This chapter highlights guidance for the daily activities of storage administration by using the IBM FlashSystem 9100 software installed on the product. This guidance can help you to maintain your storage infrastructure with the levels of availability, reliability, and resiliency demanded by today's applications, and to keep up with storage growth needs.

This chapter focuses on the most important topics to consider in IBM Flashsystem 9100 administration so that you can use it as a checklist. It also provides tips and guidance.

> **Important:** The practices that are described here were effective in many IBM Flashsystem 9100 installations worldwide for organizations in several areas. They all had one common need, which was to easily, effectively, and reliably manage their SAN storage environment. Nevertheless, whenever you have a choice between two possible implementations or configurations, if you look deep enough, you *always* have *both* advantages and disadvantages over one another.
>
> Do not take these practices as absolute truth, but rather use them as a guide. The choice of which approach to use is ultimately yours.

Important, IBM FlashSystem 9200: On 11th February 2020 IBM announced the arrival of the IBM FlashSystem 9200 to the family. This book was written specifically for IBM FlashSystem 9100, however most of the general principles will apply to the IBM FlashSystem 9200.

If you are in any doubt as to their applicability to the FlashSystem 9200 then you should work with your local IBM representative.

This book will be updated to include FlashSystem 9200 in due course.

The Flashsystem 9200 product guide is available at:

IBM FlashSystem 9200 Product Guide

This chapter includes the following sections:

- Documenting IBM Flashsystem 9100 and SAN environment
- Storage management users
- Standard operating procedures
- IBM FlashSystem 9100 code update
- SAN modifications
- Hardware upgrades for IBM FlashSystem 9100
- Adding expansion enclosures
- I/O Throttling

9.1 Documenting IBM FlashSystem 9100 and SAN environment

This section focuses on the challenge of automating the documentation that is needed for an IBM FlashSystem 9100 solution. Consider the following points:

- Several methods and tools are available to automate the task of creating and updating the documentation. Therefore, the IT infrastructure might handle this task.
- Planning is key to maintaining sustained and organized growth. Accurate documentation of your storage environment is the blueprint with which you plan your approach to short-term and long-term storage growth.
- Your storage documentation must be conveniently available and easy to consult when needed. For example, you might need to determine how to replace your core SAN directors with newer ones, or how to fix the disk path problems of a single server. The relevant documentation might consist of a few spreadsheets and a diagram.

Storing documentation: Avoid storing IBM FlashSystem 9100 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 9100 and SAN environment documentation should be sufficient for any system administrator who has average skills in the products that are included. Make a copy that includes all of your configuration information.

Use the copy to create a functionally equivalent copy of the environment by using similar hardware without any configuration, off-the-shelf media, and configuration backup files. You might need the copy if you ever face a disaster recovery scenario, which is also why it is so important to run periodic disaster recovery tests.

Create the first version of this documentation as you install your solution. If you completed forms to help plan the installation of your IBM FlashSystem 9100 solution, use these forms to help you document how your IBM FlashSystem 9100 solution was first configured. Minimum documentation is needed for an IBM FlashSystem 9100 solution. Because you might have more business requirements that require other data to be tracked, remember that the following sections do not address every situation.

9.1.1 Naming conventions

Whether you are creating your IBM FlashSystem 9100 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 9100 and SAN environment. System administrators can then determine whether a name belongs to a volume, storage pool, MDisk, host, or host bus adapter (HBA) by looking at it.

Because error messages often point to the device that generated an error, a good naming convention quickly highlights where to start investigating when an error occurs. Typical IBM FlashSystem 9100 and SAN component names limit the number and type of characters you can use. For example, IBM FlashSystem 9100 names are limited to 63 characters, which makes creating a naming convention a bit easier

Many names in IBM FlashSystem 9100 and SAN environment can be modified online. Therefore, you do not need to worry about planning outages to implement your new naming convention. The naming examples that are used in the following sections are effective in most cases, but might not be fully adequate for your particular environment or needs. The naming convention to use is your choice, but you must implement it in the whole environment.

Enclosures, node canisters and external storage controllers,

IBM FlashSystem 9100 names its internal canisters or nodes as `nodeX`, with `X` being a sequential decimal number. These will range from two, and up to eight, in a four IBM FlashSystem 9100 system cluster.

If multiple additional external controllers are attached to your IBM FlashSystem 9100 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 9100 to know which storage controller to check.

> **Note:** IBM FlashSystem 9100 detects external controllers based on their 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 IBM FlashSystem 9100 detects new MDisks, it names them by default as `mdiskXX`, where `XX` is a sequential number. Change the `XX` value to something more meaningful. MDisks are either arrays (DRAID) from internal storage or volumes from an external storage system.

For example, you can change it to include the following information:

- For internal MDisks make reference to the IBM FlashSystem 9100 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:

- `FS9100CL01_MD03`, where FS9100CL01 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:

- `FS9100CL01_POOL01`, where FS9100CL01 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 (formerly VDisks)

Volume names should include the following information:

- The hosts or cluster to which the volume is mapped.
- A single letter that indicates its usage by the host, as shown in the following examples:
  - B: For a boot disk, or R for a rootvg disk (if the server boots from SAN)
  - D: For a regular data disk
  - Q: For a cluster quorum disk (do not confuse with IBM FlashSystem 9100 quorum disks)
  - L: For a database logs disks
  - T: For a database table disk
- A few sequential digits, for uniqueness.

For example, `ERPNY01_T03` indicates a volume that is mapped to server ERPNY01 and database table disk 03.

- Sessions standard for VMware datastores:
  - esx01-sessions-001: For a datastore composed of a single volume.
  - esx01-sessions-001a and esx01-sessions-001b: For a datastore composed by 2 volumes.

Hosts

In today's environment, administrators deal with large networks, the internet, and Cloud Computing. Use good server naming conventions so that they can quickly identify a server and determine the following information:

- Where it is (to know how to access it).
- What kind it is (to determine the vendor and support group in charge).
- What it does (to engage the proper application support and notify its owner).
- Its importance (to determine the severity if problems occur).

Changing a server's name in IBM FlashSystem 9100 is as simple as changing any other IBM FlashSystem 9100 object name. However, changing the name on the operating system of a server might have implications for application configuration and require a server reboot. Therefore, you might want to prepare a detailed plan if you decide to rename several servers in your network. The following example is for server name conventions for `LLAATRFFNN`:

- 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_NYFS9100_N1_P1_HOST: Port 1 of Node 1 from FS9100 Cluster NYFS9100 dedicated for hosts`
- `IBM_NYFS9100_N1_P5_INTRACLUSTER: Port 5 of Node 1 from FS9100 Cluster NYFS9100 dedicated to intracluster traffic`
- `IBM_NYFS9100_N1_P7_REPLICATION: Port 7 of Node 1 from FS9100 Cluster NYFS9100 dedicated to replication`

  Be mindful of the IBM FlashSystem 9100 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`
- `IBM_TL01_TD06: Tape library 01, tape drive 06`
- `EMC_VNX7500_01_SPAP2: EMC VNX7500 hostname VNX7500_01, SP A, port 2`

If your SAN does not support aliases, for example, in heterogeneous fabrics with switches in some interoperations modes, use WWPNs in your zones. However, remember to update every zone that uses a WWPN if you ever change it.

Have your SAN zone name reflect the devices in the SAN it includes (normally in a one-to-one relationship) as shown in the following examples:

- `SERVERALIAS_T1_FS9100CLUSTERNAME` (from a server to the IBM FlashSystem 9100, 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_FS9100CLUSTERNAME` (from a server to the IBM FlashSystem 9100, 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_FS9100CLUSTERNAME` (zone between a external back-end storage and the IBM FlashSystem 9100).
- `NYC_FS9100_POK_FS9100_REPLICATION` (for remote copy services).

9.1.2 SAN fabric documentation

The most basic piece of SAN documentation is a SAN diagram. It is likely to be one of the first pieces of information you need if you ever seek support from your SAN switches vendor. Also, a good spreadsheet with ports and zoning information eases the task of searching for detailed information, which, if included in the diagram, makes the diagram 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 e-mail.

After some time (typically a few hours), the user receives an e-mail with instructions about how to download the report. The report includes a Visio Diagram of your SAN and an organized Microsoft Excel spreadsheet that contains all of your SAN information. For more information and to download the tool, see this website:

Brocade Healthcheck Tool

The first time that you use the SAN Health Diagnostics Capture tool, explore the options provided to learn how to create a well-organized and useful diagram.

Figure 9-1 on page 325 shows an example of a poorly formatted diagram.

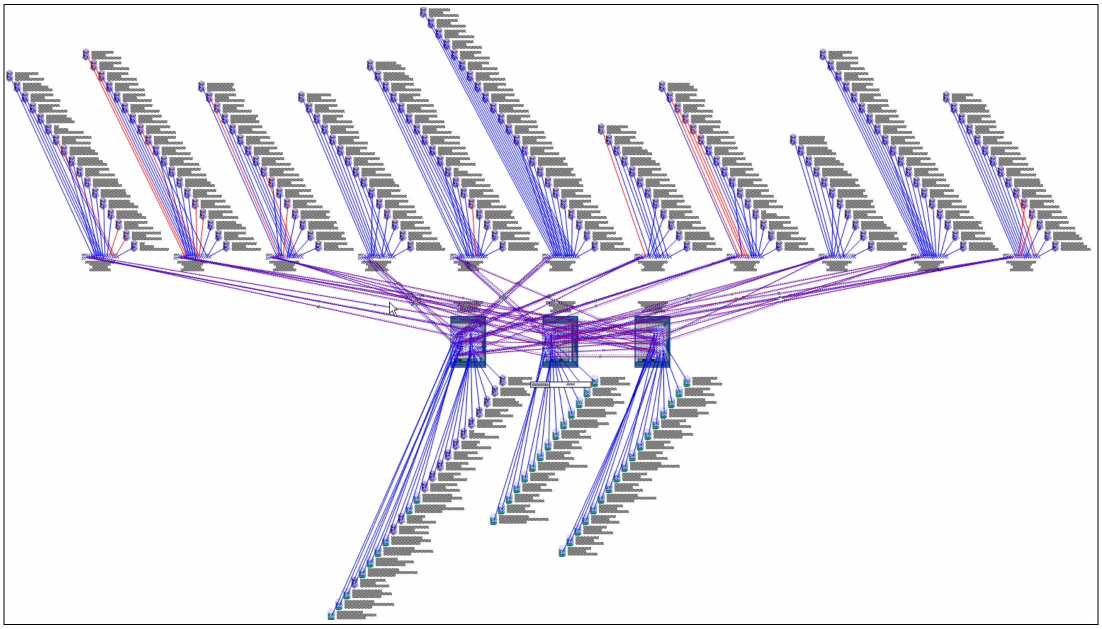
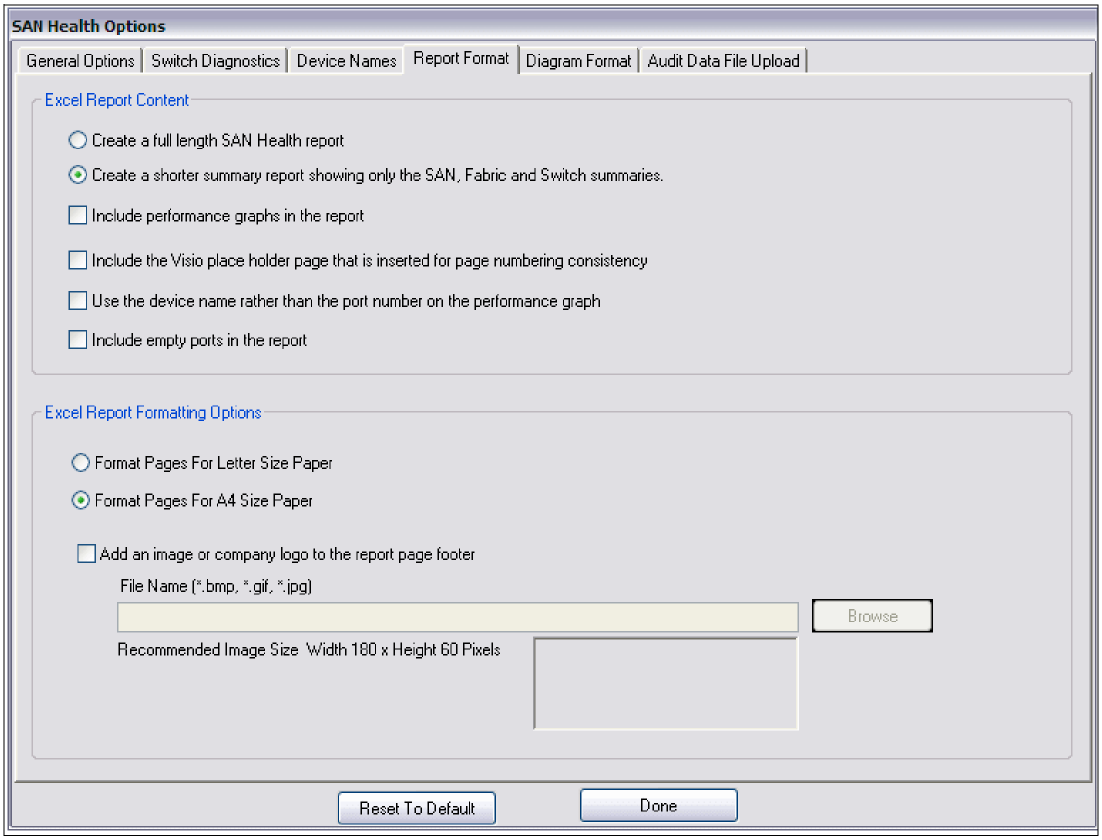


Figure 9-1   A poorly formatted SAN diagram

Figure 9-2 shows a tab of the SAN Health Options window in which you can choose the format of SAN diagram that best suits your needs. Depending on the topology and size of your SAN fabrics, you might want to manipulate the options in the Diagram Format or Report Format tabs.



Excel Report Content

○ Create a full length SAN Health report

● Create a shorter summary report showing only the SAN, Fabric and Switch summaries.

□ Include performance graphs in the report

□ Include the Visio place holder page that is inserted for page numbering consistency

□ Use the device name rather than the port number on the performance graph

□ Include empty ports in the report

Excel Report Formatting Options

○ Format Pages For Letter Size Paper

● Format Pages For A4 Size Paper

□ Add an image or company logo to the report page footer

File Name (*.bmp, *.gif, *.jpg)

Recommended Image Size Width 180 x Height 60 Pixels

Figure 9-2   Brocade SAN Health Options window

SAN Health supports switches from 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:

Broadcom SAN Health Professional Download

> **Tip:** Regardless of the method that is used, generate a fresh report at least once a month. Keep previous versions so that you can track the evolution of your SAN.

IBM Spectrum Control reporting

If you have IBM Spectrum Control running in your environment, you can use it to generate reports on your SAN. For more information about how to configure and schedule IBM Spectrum Control reports, see the IBM Spectrum Control Documentation.

Also see Chapter 8, “Monitoring” on page 267 on how to configure and set-up Spectrum Control.

Ensure that the reports that you generate include all the information that you need. Schedule the reports with a period that you can use to backtrack any changes that you make.

9.1.3 IBM FlashSystem 9100 documentation

You can back up the configuration data for an IBM FlashSystem 9100 system after preliminary tasks are completed. Configuration data for the system provides information about your system and the objects that are defined in it.

Before you back up your configuration data, the following prerequisites must be met:

- No independent operations that change the configuration for the system can be running while the **backup** command is running.
- No object name can begin with an underscore character (_).

> **Note:** The system automatically creates a backup of the configuration data each day at 1 AM. This backup is known as a *cron backup* and is written on the configuration node to `/dumps/svc.config.cron.xml_<serial#>`.

Use these instructions to generate a manual backup at any time:

1. Issue the **svcconfig backup** command to back up your configuration:

   The command displays messages similar to the ones in Example 9-1.

Example 9-1 Sample svcconfig backup command output

```
CMMVC6112W io_grp io_grp1 has a default name
CMMVC6112W io_grp io_grp2 has a default name
CMMVC6112W mdisk mdisk14 ...
CMMVC6112W node node1 ...
CMMVC6112W node node2 ...
....................................................
```

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 of the configuration node. Table 9-1 describes the three files that are created by the backup process.

Table 9-1 Files created by the backup process

File name	Description
svc.config.backup.xml_<serial#>	Contains your configuration data.
svc.config.backup.sh_<serial#>	Contains the names of the commands that were issued to create the backup of the system.
svc.config.backup.log_<serial#>	Contains details about the backup, including any reported errors or warnings.

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. Copy the backup file from the configuration node. With MS Windows, use the PuTTY **pscp** utility. With UNIX or Linux, you can use the standard **scp** utility.

The configuration backup file is in Extensible Markup Language (XML) format and can be imported into your IBM FlashSystem 9100 documentation spreadsheet. The configuration backup file might contain too much data; for example, it contains information about each internal storage drive that is installed in the system. Importing the file into your IBM FlashSystem 9100 documentation spreadsheet might make it unreadable.

In this case, consider collecting the output of specific commands. At a minimum, you should collect the output of the following commands:

- **svcinfo lsfabric**
- **svcinfo lssystem**
- **svcinfo lsmdisk**
- **svcinfo lsmdiskgrp**
- **svcinfo lsvdisk**
- **svcinfo lshost**
- **svcinfo lshostvdiskmap**

> **Note:** Most CLI commands shown above will work without the *svcinfo* prefix, however there might be some that do not work with just the short-name, and so require the *svcinfo* prefix to be added.

Import the commands into a spreadsheet, preferably with each command output on a separate sheet.

One way to automate either task is to first create a batch file (Windows) or shell script (UNIX or Linux) that collects and stores this information. For more information, see 9.8, “I/O Throttling” on page 349. Then, use spreadsheet macros to import the collected data into your IBM FlashSystem 9100 documentation spreadsheet.

When you are gathering IBM FlashSystem 9100 information, consider the following preferred practices:

- If you are collecting the output of specific commands, use the **`-delim`** option of these commands to make their output delimited by a character other than tab, such as comma, colon, or exclamation mark. You can import the temporary files into your spreadsheet in comma-separated values (CSV) format, specifying the same delimiter.

  > **Note:** It is important to use a delimiter that is not already part of the output of the command. Commas can be used if the output is a particular type of list. Colons might be used for special fields, such as IPv6 addresses, WWPNs, or ISCSI names.

- If you are collecting the output of specific commands, save the output to temporary files. To make your spreadsheet macros simpler, you might want to preprocess the temporary files and remove any “garbage” or undesired lines or columns. With UNIX or Linux, you can use text edition commands such as **`grep`**, **`sed`**, and **`awk`**. Freeware software is available for Windows with the same commands, or you can use any batch text editor tool.

The objective is to fully automate this procedure so you can schedule it to run automatically on a regular basis. Make the resulting spreadsheet easy to consult and have it contain only the information that you use frequently. The automated collection and storage of configuration and support data (which is typically more extensive and difficult to use) are described in 9.1.7, “Automated support data collection” on page 330.

9.1.4 Storage documentation

You must generate only documentation of your back-end storage controllers manually once 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 9100 internal disk drives and enclosures.

Any portion of your external storage controllers that is used outside the IBM FlashSystem 9100 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 additional back-end to the IBM FlashSystem 9100 solution. This way, you can perform all your Disk Storage Management tasks by using IBM FlashSystem 9100.

9.1.5 Technical support information

If you must open a technical support incident for your storage and SAN components, create and keep available a spreadsheet with all relevant information for all storage administrators. This spreadsheet should include the following information:

- Hardware information:
  - Vendor, machine and model number, serial number (example: `IBM 9848-AF8 S/N 7812345`)
  - Configuration, if applicable
  - Current code level

- Physical location:
  - Datacenter, including the complete street address and phone number
  - Equipment physical location (room number, floor, tile location, and rack number)
  - Vendor’s security access information or procedure, if applicable
  - Onsite person’s contact name and phone or page number
- Support contract information:
  - Vendor contact phone numbers and website
  - Customer’s contact name and phone or page number
  - User ID to the support website, if applicable

    Do not store the password in the spreadsheet under any circumstances
  - Support contract number and expiration date

By keeping this data on a spreadsheet, storage administrators have all the information that they need to complete a web support request form or to provide to a vendor’s call support representative. Typically, you are asked first for a brief description of the problem and then asked later for a detailed description and support data collection.

9.1.6 Tracking incident and change tickets

If your organization uses an incident and change management and tracking tool (such as IBM Tivoli® Service Request Manager®), you or the storage administration team might need to develop proficiency in its use for several reasons:

IBM Tivoli Service Request Manager Support Page

- If your storage and SAN equipment are not configured to send SNMP traps to this incident management tool, manually open incidents whenever an error is detected.
- The IBM FlashSystem 9100 has the ability to be managed by the new 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, see Chapter 8, “Monitoring” on page 267.
- Disk storage allocation and deallocation and SAN zoning configuration modifications should be handled under properly submitted and approved change tickets.
- If you are handling a problem yourself, or calling your vendor’s technical support desk, you might need to produce a list of the changes that you recently implemented in your SAN or that occurred since the documentation reports were last produced or updated.

When you use incident and change management tracking tools, adhere to the following guidelines for IBM FlashSystem 9100 and SAN Storage Administration:

- Whenever possible, configure your storage and SAN equipment to send SNMP traps to the incident monitoring tool so that an incident ticket is automatically opened and the proper alert notifications are sent. If you do not use a monitoring tool in your environment, you might want to configure e-mail alerts that are automatically sent to the mobile phones or pagers of the storage administrators on duty or on call.
- Discuss within your organization the risk classification that a storage allocation or deallocation change ticket is to have. These activities are typically safe and non disruptive to other services and applications when properly handled.

However, they have the potential to cause collateral damage if a human error or an unexpected failure occurs during implementation. Your organization might decide to assume more costs with overtime and limit such activities to off-business hours, weekends, or maintenance windows if they assess that the risks to other critical applications are too high.

- Use templates for your most common change tickets, such as storage allocation or SAN zoning modification, to facilitate and speed up their submission.
- Do not open change tickets in advance to replace failed, redundant, hot-pluggable parts, such as disk drive modules (DDMs) in storage controllers with hot spares, or SFPs in SAN switches or servers with path redundancy. Typically, these fixes do not change anything in your SAN storage topology or configuration, and do not cause any more service disruption or degradation than you already had when the part failed. Handle these fixes within the associated incident ticket because it might take longer to replace the part if you need to submit, schedule, and approve a non-emergency change ticket.

  An exception is if you must interrupt more servers or applications to replace the part. In this case, you must schedule the activity and coordinate support groups. Use good judgment and avoid unnecessary exposure and delays.
- Keep handy the procedures to generate reports of the latest incidents and implemented changes in your SAN Storage environment. Typically, you do not need to periodically generate these reports because your organization probably already has a Problem and Change Management group that runs such reports for trend analysis purposes.

9.1.7 Automated support data collection

In addition to the easier-to-use documentation of your IBM FlashSystem 9100 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 9100, 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 now the possibility to use IBM Storage Insights (SI) to create support tickets and then attach the snap data to this record from within the SI GUI. For further information see Chapter 10, “Troubleshooting and diagnostics” on page 357.

9.1.8 Subscribing to IBM FlashSystem 9100 support

Subscribing to IBM FlashSystem 9100 support is probably the most overlooked practice in IT administration, and yet it is the most efficient way to stay ahead of problems. With this subscription, you can receive notifications about potential threats before they can reach you and cause severe service outages.

To subscribe to this support and receive support alerts and notifications for your products, see the following site:

https://www.ibm.com/support/pages/node/718119

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 for My Notifications and tailor the requests and alerts you wants to receive

9.2 Storage management users

Almost all organizations have IT security policies that enforce the use of password-protected user IDs when their IT assets and tools are used. However, some storage administrators still use generic, shared IDs, such as `superuser`, `admin`, or `root`, in their management consoles to perform their tasks. They might even use a factory-set default password. Their justification might be a lack of time, forgetfulness, or the fact that their SAN equipment does not support the organization’s authentication tool.

SAN storage equipment management consoles often do not provide access to stored data, but one can easily shut down a shared storage controller and any number of critical applications along with it. Moreover, having individual user IDs set for your storage administrators allows much better backtracking of your modifications if you must analyze your logs.

IBM FlashSystem 9100 supports the following authentication methods:

- Local authentication by using password
- Local authentication by using SSH keys
- Remote authentication using LDAP
- Remote authentication using Tivoli

Regardless of the authentication method you choose, complete the following tasks:

- Create individual user IDs for your Storage Administration staff. Choose user IDs that easily identify the user. Use your organization’s security standards.
- Include each individual user ID into the UserGroup with only enough privileges to perform the required tasks.
- If required, create generic user IDs for your batch tasks, such as Copy Services or Monitoring. Include them in a Copy Operator or Monitor UserGroup. Do not use generic user IDs with the SecurityAdmin privilege in batch tasks.
- Create unique SSH public and private keys for each of your administrators.
- Store your `superuser` password in a safe location in accordance to your organization’s security guidelines and use it only in emergencies.

9.3 Standard operating procedures

To simplify the SAN storage administration tasks that you use most often (such as SAN storage allocation or removal, or adding or removing a host from the SAN), create step-by-step, predefined standard procedures for them. The following sections provide guidance for keeping your IBM FlashSystem 9100 environment working correctly and reliably.

9.3.1 Allocating and deallocating volumes to hosts

When you allocate and deallocate volumes to hosts, consider the following guidelines:

- Before you allocate new volumes to a server with redundant disk paths, verify that these paths are working well and that the multipath software is free of errors. Fix any disk path errors that you find in your server before you proceed.
- When you plan for future growth of space efficient volumes (VDisks), determine whether your server's operating system supports the particular volume to be extended online. Previous AIX releases, for example, do not support online expansion of rootvg LUNs. Test the procedure in a non-production server first.
- Always cross-check the host LUN ID information with the vdisk_UID of the IBM FlashSystem 9100. 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 9100.
- Ensure that you delete any volume or LUN definition in the server *before* you unmap it in IBM FlashSystem 9100. 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 has detected I/O activity within the specified time in minutes to the volume from any host).

> **Note:** Volume protection cannot be overridden by the use of the **`-force`** flag in the affected CLI commands. Volume protection must be disabled to carry on an activity that is currently blocked.

- Ensure that you explicitly remove a volume from any volume-to-host mappings and any copy services relationship to which it belongs *before* you delete it.

> **Attention:** You must avoid the use of the **`-force`** parameter in `rmvdisk`.

- If you issue the **`svctask rmvdisk`** command and it still has pending mappings, IBM FlashSystem 9100 prompts you to confirm and is a hint that you might have done something incorrectly.
- When you are deallocating volumes, plan for an interval between unmapping them to hosts (**`rmvdiskhostmap`**) and deleting them (**`rmvdisk`**). The IBM internal Storage Technical Quality Review Process (STQRP) asks for a minimum of a 48-hour period, and having at least a one business day interval so that you can perform a quick backout if you later realize you still need some data on that volume.

9.3.2 Adding and removing hosts

When you add and remove host (or hosts) in IBM FlashSystem 9100, consider the following guidelines:

- Before you map new servers to IBM FlashSystem 9100, verify that they are all error free. Fix any errors that you find in your server and IBM FlashSystem 9100 before you proceed. In IBM FlashSystem 9100, 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 HBA to IBM FlashSystem 9100, you should list its WWPN by using the **lshbaportcandidate** command. Use the **lsfabric** command to certify that it was detected by the IBM FlashSystem 9100 nodes and ports that you expected. When you create the host definition in IBM FlashSystem 9100 (**mkhost**), try to avoid the **-force** parameter. If you do not see the host's WWPNs, it might be necessary to scan fabric from the host. For example, use the **cfgmgr** command in AIX.

9.4 IBM FlashSystem 9100 code update

Because IBM FlashSystem 9100 might be at the core of your disk and SAN storage environment, its update requires planning, preparation, and verification. However, with the appropriate precautions, an update can be conducted easily and transparently to your servers and applications. This section highlights applicable guidelines for IBM FlashSystem 9100 update.

Most of the following sections explain how to prepare for the IBM FlashSystem 9100 update. The last two sections present version-independent guidelines to update the IBM FlashSystem 9100 system and flash drives.

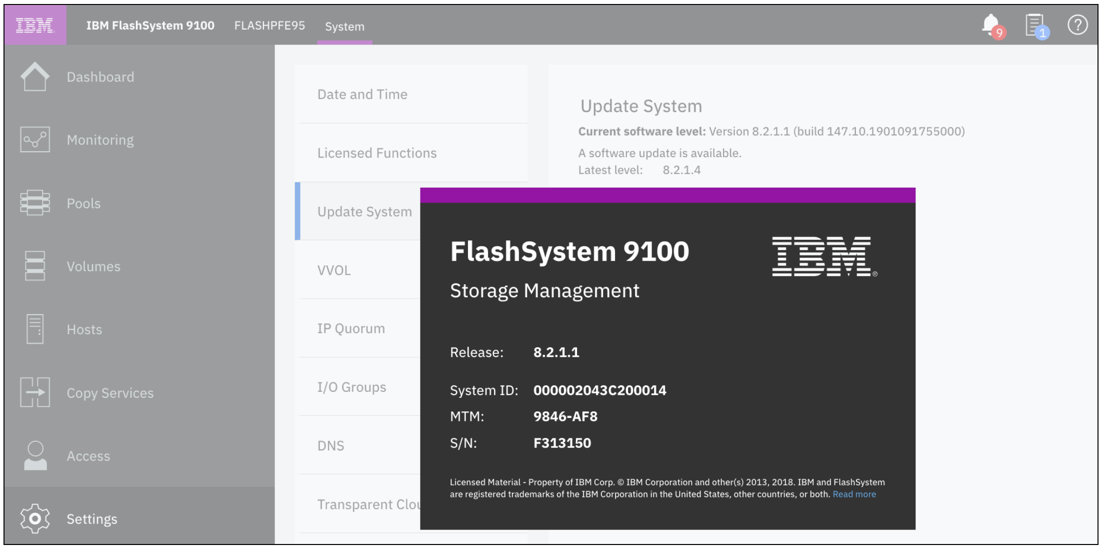
> **Note:** For customers who have puchased the IBM FlashSystem 9100 with 3 years warranty (9848 Models AF7 or AF8), this comes with Enterprise Class Support (ECS) and this entitles the customer to two code upgrades, done by IBM, per year (total of 6 across the 3 years of warranty). These upgrades are done by the IBM dedicated Remote Code Load (RCL) team or, where remote support is not allowed or enabled, by an onsite SSR.
>
> For further information on ECS, see the IBM Knowledge Center:
>
> IBM Knowledge Center for ECS

9.4.1 Current and target IBM FlashSystem 9100 code level

First, determine your current and target IBM FlashSystem 9100 code level. Log in to your IBM FlashSystem 9100 web-based GUI and find the current version.

On the right hand side of the top menu drop down line, select**?** → **About IBM FlashSystem 9100**.

Figure 9-3 on page 334 shows the About IBM FlashSystem 9100 output panel and displays the current code level. In this example the current code level is 8.2.1.1.



Update System

Current software level: Version 8.2.1.1 (build 147.10.1901091755000)

A software update is available.
Latest level: 8.2.1.4

FlashSystem 9100

Storage Management

Release: **8.2.1.1**

System ID: **000002043C200014**
MTM: **9846-AF8**
S/N: **F313150**

Licensed Material - Property of IBM Corp. © IBM Corporation and other(s) 2013, 2018. IBM and FlashSystem are registered trademarks of the IBM Corporation in the United States, other countries, or both. Read more

Figure 9-3 About IBM FlashSystem 9100 output panel

Alternatively, if you are using the CLI, run the **`svcinfo lssystem`** command.

Example 9-2 shows the output of the **`lssystem`** CLI command and where the code level statement can be found.

Example 9-2 lssystem command

```
IBM_FlashSystem:FLASHPFE95:superuser>lssystem
id 000002043C200014
name FLASHPFE95
location local
partnership
total_mdisk_capacity 198.0TB
space_in_mdisk_grps 198.0TB
space_allocated_to_vdisks 155.06TB
total_free_space 43.0TB
total_vdiskcopy_capacity 501.00TB
total_used_capacity 148.00TB
total_overallocation 252
total_vdisk_capacity 501.00TB
total_allocated_extent_capacity 155.06TB
statistics_status on
statistics_frequency 5
cluster_locale en_US
time_zone 292 Australia/Sydney
code_level 8.2.1.1 (build 147.10.1901091755000)

NOTE: Output reduced for convenience of example
```

IBM FlashSystem 9100 code levels are specified by four digits in the following format:

- In our example V.R.M.F = 8.2.1.1
  - V is the major version number
  - R is the release level
  - M is the modification level
  - F is the fix level

Use the recommended IBM FlashSystem 9100 release unless you have a specific reason not to update:

- The specific version of an application or other component of your SAN Storage environment has a known problem or limitation.
- The latest IBM FlashSystem 9100 code release is not yet cross-certified as compatible with another key component of your SAN storage environment.
- Your organization has mitigating internal policies, such as the use of the “latest minus 1” release, or prompting for “seasoning” in the field before implementation.

For more information, see the following website:

- Spectrum Virtualize Family of Products and FlashSystem 9100 Upgrade Planning

9.4.2 IBM FlashSystem 9100 Upgrade Test Utility

Install and run the latest IBM FlashSystem 9100 Upgrade Test Utility before you update the IBM FlashSystem 9100 code. To download the Upgrade Test Utility, see this website:

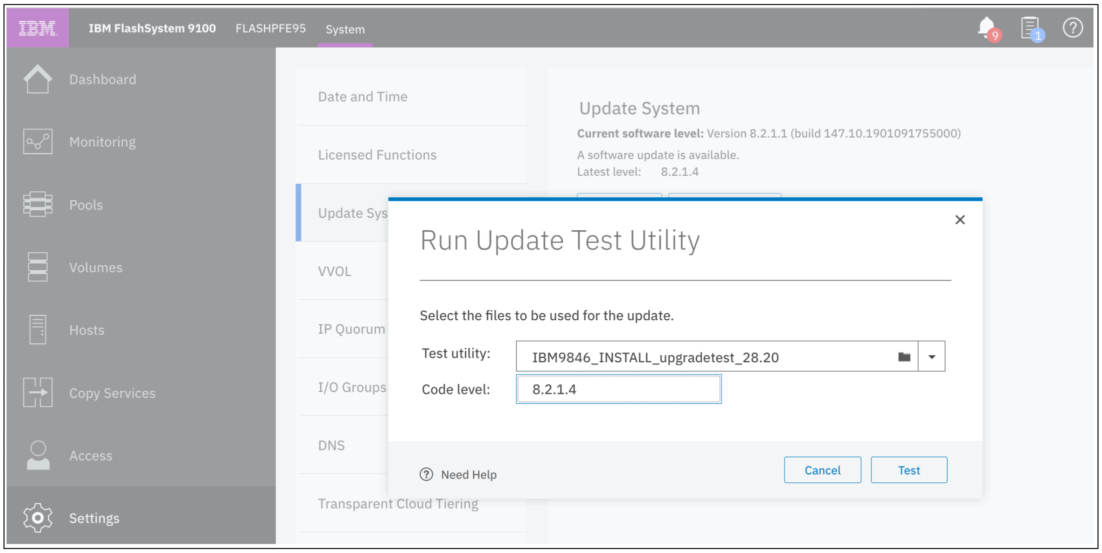
Software Upgrade Test Utility

This tool verifies the health of your IBM FlashSystem 9100 solution for the update process. It also checks for unfixed errors, degraded MDisks, inactive fabric connections, configuration conflicts, hardware compatibility, disk drives firmware, and many other issues that might otherwise require cross-checking a series of command outputs.

> **Note:** The Upgrade Test Utility does not log in storage controllers or SAN switches. Instead, it reports the status of the connections of IBM FlashSystem 9100 to these devices. It is the users’ responsibility to check these components for internal errors.

You can use the GUI or the CLI to install and run the Upgrade Test Utility.

Figure 9-4 shows the IBM FlashSystem 9100 version 8.2.1.1 GUI window, that is used to install and run the Upgrade Test Utility. It is uploaded and installed like any other software update.



Update System

Current software level: Version 8.2.1.1 (build 147.10.1901091755000)

A software update is available.

Latest level: 8.2.1.4

Run Update Test Utility

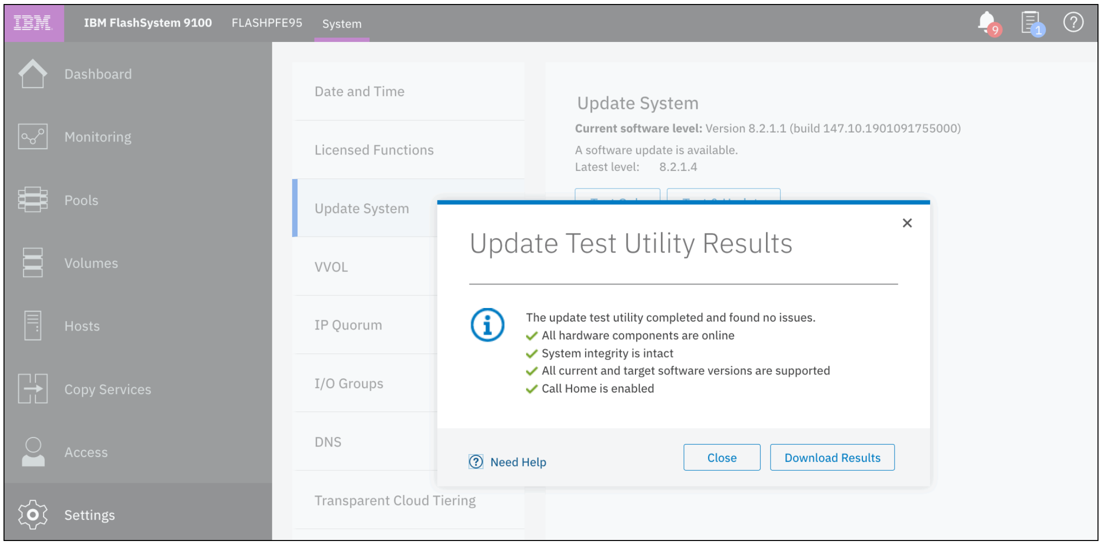
Select the files to be used for the update.

Test utility: IBM9846_INSTALL_upgradetest_28.20

Code level: 8.2.1.4

Figure 9-4 IBM FlashSystem 9100 Upgrade Test Utility installation using the GUI

Figure 9-5 shows a successful completion of the update test utility.



Update System

Current software level: Version 8.2.1.1 (build 147.10.1901091755000)

A software update is available.

Latest level: 8.2.1.4

Update Test Utility Results

The update test utility completed and found no issues.

✓ All hardware components are online

✓ System integrity is intact

✓ All current and target software versions are supported

✓ Call Home is enabled

Figure 9-5 IBM FlashSystem 9100 Upgrade Test Utility completion panel

Example 9-3 on page 336 shows how to install and run Upgrade Test Utility in the CLI. In this case, the Upgrade Test Utility found no errors and completed successfully.

> **Note:** If you use the CLI method to run the test utility, you will need to upload the test utility file manually to the machine, using a pscp or scp client. See this article in the IBM Knowledge Center for guidance on doing this.

Example 9-3 Upgrade test by using the CLI

```
IBM_FlashSystem:FLASHPFE95:superuser>svctask applysoftware -file
IBM9846_INSTALL_upgradetest_28.20

CMMVC9001I The package installed successfully.

IBM_FlashSystem:FLASHPFE95:superuser>svcupgradetest -v 8.2.1.4
svcupgradetest version 28.20

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.
```

9.4.3 IBM FlashSystem 9100 hardware considerations

Before you start the update process, always check whether your IBM FlashSystem 9100 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 9100 code first.

9.4.4 Attached hosts preparation

If the appropriate precautions are taken, the IBM FlashSystem 9100 update is not apparent to the attached servers and their applications. The automated update procedure updates one IBM FlashSystem 9100 node at a time, while the other node in the I/O group covers for its designated volumes.

However, to ensure that this feature works, the *failover capability* of your multipath software must be working properly. This capability can be mitigated by enabling NPIV if your current code level supports this function. For more information about NPIV, see Chapter 7, “Hosts” on page 241.

Before you start IBM FlashSystem 9100 update preparation, check the following items for every server that is attached to IBM FlashSystem 9100 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 website (require an IBM ID):

- The IBM Support page on IBM FlashSystem 9100 Troubleshooting

Fix every problem or “suspect” that you find with the disk path failover capability. Because a typical IBM FlashSystem 9100 environment has several dozens of servers to a few hundred servers attached to it, a spreadsheet might help you with the Attached Hosts Preparation tracking process. If you have some host virtualization, such as VMware ESX, AIX LPARs, IBM VIOS, or Solaris containers in your environment, verify the redundancy and failover capability in these virtualization layers.

9.4.5 Storage controllers preparation

As critical as with the attached hosts, the attached storage controllers must correctly handle the failover of MDisk paths. Therefore, they must be running supported microcode versions and their own SAN paths to IBM FlashSystem 9100 must be free of errors.

9.4.6 SAN fabrics preparation

If you are using symmetrical, redundant, independent SAN fabrics, preparing these fabrics for an IBM FlashSystem 9100 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 9100 environment must continue working through the other fabric and your applications must remain unaffected.

Because you are updating your IBM FlashSystem 9100, also update your SAN switches code to the latest supported level. Start with your principal core switch or director, continue by updating the other core switches, and update the edge switches last. Update one entire fabric (all switches) before you move to the next one so that any problem you might encounter affects only the first fabric. Begin your other fabric update only after you verify that the first fabric update has no problems.

If you are not running symmetrical, redundant independent SAN fabrics, fix this problem as a high priority because it represents a single point of failure (SPOF).

9.4.7 SAN components update sequence

Check the compatibility of your target IBM FlashSystem 9100 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 9100 target code level is compatible.

The IBM FlashSystem 9100 Supported Hardware List provides the complete information for using your IBM FlashSystem 9100 SAN storage environment components with the current and target code level. For links to the Supported Hardware List, Device Driver, Firmware, and Recommended Software Levels for different products and different code levels, see the following resource:

- Support Information for FlashSystem 9100 family

By cross-checking the version of IBM FlashSystem 9100 is compatible with the versions of your SAN environment components, you can determine which one to update first. By checking a component's update path, you can determine whether that component requires a multistep update.

If you are not making major version or multistep updates in any components, the following update order is less prone to eventual problems:

1. SAN switches or directors
2. Storage controllers
3. Servers HBAs microcodes and multipath software
4. IBM FlashSystem 9100 system
5. IBM FlashSystem 9100 internal NVMe drives
6. IBM FlashSystem 9100 SAS attached SSD drives

Attention: Do *not* update two components of your IBM FlashSystem 9100 SAN storage environment simultaneously, such as the IBM FlashSystem 9100 system and one storage controller. This caution is true even if you intend to do it with your system offline. An update of this type can lead to unpredictable results, and an unexpected problem is much more difficult to debug.

9.4.8 IBM FlashSystem 9100 participating in copy services relationship

When you update an IBM FlashSystem 9100 system that participates in an intercluster Copy Services relationship, do *not* update both clusters in the relationship simultaneously. This situation is not verified or monitored by the automatic update process, and might lead to a loss of synchronization and unavailability.

You must successfully finish the update in one cluster before you start the next one. Try to update the next cluster as soon as possible to the same code level as the first one. Avoid running them with different code levels for extended periods.

9.4.9 IBM FlashSystem 9100 update

Adhere to the following version-independent guidelines for your IBM FlashSystem 9100 code update:

- Schedule the IBM FlashSystem 9100 code update for a low I/O activity time. The update process puts one node at a time offline. It also disables the write cache in the I/O group that node belongs to until both nodes are updated. Therefore, with lower I/O, you are less likely to notice performance degradation during the update.
- Never power off, reboot, or reset an IBM FlashSystem 9100 node during code update unless you are instructed to do so by IBM Support. Typically, if the update process encounters a problem and fails, it backs out.
- Check whether you are running a web browser type and version that is supported by the IBM FlashSystem 9100 target code level on every computer that you intend to use to manage your IBM FlashSystem 9100.
- If you are planning for a major IBM FlashSystem 9100 version update, update your current version to its latest fix level before you run the major update.

9.4.10 IBM FlashSystem 9100 disk drive update

Updating of disk drive firmware is concurrent whether it is NVMe drives in the control enclosure or the SSD drives in any SAS attached expansion enclosures. This procedure updates firmware on a drive that is in the control enclosure or in one of the expansion enclosures. If the update would cause any volumes to go offline, the force option is required.

To update the drive firmware for all drives, by using the management GUI, select **Pools** → **Internal Storage** → **Actions** → **Update All**. You can also update individual drives. If you wish to update the drive firmware via the CLI, or need more detailed information, then see the IBM Knowledge Center.

> **Important:** It is not possible to update multiple drives of type NVMe by using the following `applydrivesoftware` command. Update these types of drives one at a time by specifying each drive ID in a separate command, or use the `utilitydriveupgrade` command.

9.5 SAN modifications

When you administer shared storage environments, human error can occur when a failure is fixed or a change is made that affects one or more servers or applications. That error can then affect other servers or applications because appropriate precautions were not taken.

Human error can include the following examples:

- Disrupting or disabling the working disk paths of a server while trying to fix failed ones
- Disrupting a neighbor SAN switch port while inserting or pulling out an FC cable or SFP
- Disabling or removing the working part in a redundant set instead of the failed one
- Making modifications that affect both parts of a redundant set without an interval that allows for automatic failover during unexpected problems

Adhere to the following guidelines to perform these actions with assurance:

- Uniquely and correctly identify the components of your SAN.
- Use the proper failover commands to disable only the failed parts.
- Understand which modifications are necessarily disruptive, and which can be performed online with little or no performance degradation.

9.5.1 Cross-referencing HBA WWPNs

With the WWPN of an HBA, you can uniquely identify one server in the SAN. If a server's name is changed at the operating system level and not at the IBM FlashSystem 9100 host definitions, it continues to access its previously mapped volumes exactly because the WWPN of the HBA did not change.

Alternatively, if the HBA of a server is removed and installed in a second server and the first server's SAN zones and IBM FlashSystem 9100 host definitions are not updated, the second server can access volumes that it probably should not access.

Complete the following steps to cross-reference HBA WWPNs:

1. In your server, verify the WWPNs of the HBAs that are used for disk access. Typically, you can complete this task by using the SAN disk multipath software of your server.

   If you are using SDDPCM, run the **pcmpath query WWPN** command to see output similar to what is shown in Example 9-4.

   Example 9-4 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 WWPNs in the server that is attached to the SAN, such as AIX VIO or VMware ESX.Cross-reference with the output of the IBM FlashSystem 9100 **lshost <hostname>** command, as shown in Example 9-5.

   Example 9-5 Output of the lshost <hostname> command

   ```
   IBM_FlashSystem:FLASHPFE95: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 9-6. (In Brocade, switches use `nodefind <WWPN>`.)

Example 9-6 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 HBAs for performance. You cannot assume that any new volume is supposed to be mapped to every WWPN that server logged in the SAN.

If your organization uses a change management tracking tool, perform all your SAN storage allocations under approved change tickets with the servers' WWPNs listed in the Description and Implementation sessions.

9.5.2 Cross-referencing LUN IDs

Always cross-reference the IBM FlashSystem 9100 `vdisk_UID` with the server LUN ID before you perform any modifications that involve IBM FlashSystem 9100 volumes. Example 9-7 shows an AIX server that is running SDDPCM. The IBM FlashSystem 9100 `vdisk_name` has no relation to the AIX device name. Also, the first SAN LUN mapped to the server (`SCSI_id=0`) shows up as `hdisk4` in the server because it had four internal disks (`hdisk0 - hdisk3`).

Example 9-7 Results of running the lshostvdiskmap command

```
IBM_FlashSystem:FLASHPFE95: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 ticket that performs SAN storage allocation or reclaim.

> **Note:** Because a host can have many volumes with the same scsi_id, always cross-reference the IBM FlashSystem 9100 volume UID with the host volume UID, and record the scsi_id and LUN ID of that volume.

9.5.3 HBA replacement

Replacing a failed HBA is a fairly trivial and safe operation if it is performed correctly. However, more precautions are required if your server has redundant HBAs and its hardware permits you to “hot” replace it (with the server still running).

Complete the following steps to replace a failed HBA and retain the good HBA:

1. In your server, using the multipath software, identify the failed HBA and record its WWPNs. For more information, see 9.5.1, “Cross-referencing HBA WWPNs” on page 340. Then, place this HBA and its associated paths offline, gracefully if possible. This approach is important so that the multipath software stops trying to recover it. Your server might even show a degraded performance while you perform this task.
2. Some HBAs have a label that shows the WWPNs. If you have this type of label, record the WWPNs before you install the new HBA in the server.
3. If your server does not support HBA hot-swap, power off your system, replace the HBA, connect the used FC cable into the new HBA, and power on the system.

   If your server does support hot-swap, follow the appropriate procedures to perform a “hot” replace of the HBA. Do *not* disable or disrupt the good HBA in the process.
4. Verify that the new HBA successfully logged in to the SAN switch. If it logged in successfully, you can see its WWPNs logged in to the SAN switch port.

   Otherwise, fix this issue before you continue to the next step.

   Cross-check the WWPNs that you see in the SAN switch with the one you noted in step 1, and make sure that you did not get the WWNN mistakenly.
5. In your SAN zoning configuration tool, replace the old HBA WWPNs for the new ones in every alias and zone to which they belong. Do *not* touch the other SAN fabric (the one with the good HBA) while you perform this task.

   Only one alias should use each WWPN, and zones must reference this alias.

   If you are using SAN port zoning (though you should not be) and you did not move the new HBA FC cable to another SAN switch port, you do not need to reconfigure zoning.
6. Verify that the new HBA’s WWPNs appear in the IBM FlashSystem 9100 system by using the **lsfcportcandidate** command.

   If the WWPNs of the new HBA do not appear, troubleshoot your SAN connections and zoning.

7. Add the WWPNs of this new HBA in the IBM FlashSystem 9100 host definition by using the **addhostport** command. Do not remove the old one yet. Run the **lshost <servername>** command. Then, verify that the good HBA shows as `active`, while the failed and old HBA should show as `inactive` or `offline`.
8. Use software to recognize the new HBA and its associated SAN disk paths. Certify that all SAN LUNs have redundant disk paths through the good and the new HBAs.
9. Return to the IBM FlashSystem 9100 system and verify again (by using the **lshost <servername>** command) that both the good and the new HBA's WWPNs are active. In this case, you can remove the old HBA WWPNs from the host definition by using the **rmhostport** command.
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 good HBA by mistake.

9.6 Hardware upgrades for IBM FlashSystem 9100

The IBM FlashSystem 9100 scalability features allow significant flexibility in its configuration. As a consequence, several scenarios are possible for its growth. The following sections describe these processes:

- Adding IBM FlashSystem 9100 enclosures to an existing cluster
- Upgrading IBM FlashSystem 9100 nodes in an existing cluster
- Upgrading NVMe drives in an existing enclosure
- Moving to a new IBM FlashSystem 9100 cluster
- Splitting an IBM FlashSystem 9100 cluster

9.6.1 Adding IBM FlashSystem 9100 enclosures to an existing cluster

If your existing IBM FlashSystem 9100 cluster is below the maximum I/O groups limit for your specific product and you intend to upgrade it, you might find yourself installing another IBM FlashSystem 9100 control enclosure. It is also feasible that you might have an exiting cluster of IBM Storwize V7000 nodes that you wish to add the IBM FlashSystem 9100 enclosures to, which are more powerful than your existing ones. Therefore, your cluster will have different node models in different I/O groups.

To install these newer enclosures, determine whether you need to upgrade your IBM FlashSystem 9100 first (or Storwize V7000 code level if you are merging an existing V7000 Gen2 cluster with a FS9100). For more information, see 9.4.3, "IBM FlashSystem 9100 hardware considerations" on page 336.

After you install the newer nodes, you might need to redistribute your servers across the I/O groups. Consider the following points:

- Moving a server's volume to different I/O groups can be done online because of a feature called Non-Disruptive Volume Movement (NDVM). Although this process can be done without stopping the host, careful planning and preparation is advised.

  Note: You cannot move a volume that is in any type of remote copy relationship.

- If each of your servers is zoned to only one I/O group, modify your SAN zoning configuration as you move its volumes to another I/O group. As best you can, balance the distribution of your servers across I/O groups according to I/O workload.
- Use the **-iogrp** parameter in the **mkhost** command to define which I/O groups of IBM FlashSystem 9100 that the new servers will use. Otherwise, IBM FlashSystem 9100 maps by default the host to all I/O groups, even if they do not exist and regardless of your zoning configuration. Example 9-8 shows this scenario and how to resolve it by using the **rmhostiogrp** and **addhostiogrp** commands.

Example 9-8 Mapping the host to I/O groups

```
IBM_FlashSystem:FS9100-PFE1: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:FS9100-PFE1: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:FS9100-PFE1: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:FS9100-PFE1:superuser>lshostiogrp Win2012srv1
id name
0  io_grp0
1  io_grp1
2  io_grp2
3  io_grp3

IBM_FlashSystem:FS9100-PFE1:superuser>rmhostiogrp -iogrp 3 Win2012srv1
IBM_FlashSystem:FS9100-PFE1:superuser>
```

```
IBM_FlashSystem:FS9100-PFE1:superuser>lshostiogrp Win2012srv1
id name
0  io_grp0
1  io_grp1
2  io_grp2
IBM_FlashSystem:FS9100-PFE1:superuser>

IBM_FlashSystem:FS9100-PFE1:superuser>addhostiogrp -iogrp io_grp3 Win2012srv1
IBM_FlashSystem:FS9100-PFE1:superuser>

IBM_FlashSystem:FS9100-PFE1:superuser>lshostiogrp Win2012srv1
id name
0  io_grp0
1  io_grp1
2  io_grp2
3  io_grp3

IBM_FlashSystem:FS9100-PFE1: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.

9.6.2 Upgrading IBM FlashSystem 9100 nodes in an existing cluster

If you are upgrading the nodes or canisters of your existing IBM FlashSystem 9100 cluster, there is the option to increase the cache memory size and or the adapter cards in each node and this can be done, one node at a time, and so be again be non disruptive to the systems operations, For further details see this section of the IBM Knowledge Center.

Memory options for a 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 9-2.

Table 9-2 Base memory features

Feature	Memory per enclosure	Maximum per enclosure
ACG0	128 GB base cache memory (eight 16GB DIMMs - 2 per CPU)	1
ACG1	768 GB base cache memory (twenty-four 32GB DIMMs - 6 per CPU)	1

You can order the following features to upgrade to more memory at any time. Table 9-3 shows the various options.

Table 9-3 Additonal memory features

Feature	Description	Maximum per enclosure	DIMM FRU part
ACGA	128 GB memory upgrade (eight 16GB DIMMs)	3	01EJ361
ACGB	768 GB memory upgrade (twenty-four 32GB DIMMs)	2	01LJ207

Adapter cards options for a FlashSystem 9100 control enclosure

You can also add new adapter cards 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 9-4.

Table 9-4 IBM FlashSystem 9100 control enclosure adapter card 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)

Further details of the feature codes, memory options and functions of each adapter can be found in the following IBM Redbook, see the Planning chapter:

IBM FlashSystem 9100 Architecture, Performance, and Implementation, SG24-8425

9.6.3 Moving to a new IBM FlashSystem 9100 cluster

You might have a highly populated, intensively used IBM FlashSystem 9100 cluster that you want to upgrade. You might also want to use the opportunity to overhaul your IBM FlashSystem 9100 and SAN storage environment.

Complete the following steps to replace your cluster entirely with a newer, bigger, and more powerful one:

1. Install your new IBM FlashSystem 9100 cluster.
2. Create a replica of your data in your new cluster.
3. Migrate your servers to the new IBM FlashSystem 9100 cluster when convenient.

If your servers can tolerate a brief, scheduled outage to switch from one IBM FlashSystem 9100 cluster to another, you can use the IBM FlashSystem 9100 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 9100 cluster and find all the old volumes you must move.
2. Zone your host to the new IBM FlashSystem 9100 cluster.
3. Create remote copy relationships from the old volumes in the old IBM FlashSystem 9100 cluster to new volumes in the new IBM FlashSystem 9100 cluster.
4. Map the new volumes from the new IBM FlashSystem 9100 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 9100 cluster.
7. Disconnect and remove the old volumes on the host from the old IBM FlashSystem 9100 cluster.
8. Unmap the old volumes from the old IBM FlashSystem 9100 cluster to the host.
9. Make sure remote copy relationships between old and new volumes in the old and new IBM FlashSystem 9100 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 9100 cluster receive read/write access.
11. Import data from the new volumes and start your applications on the host.

If you must migrate a server online, instead, you must use host-based mirroring by completing these steps:

1. Select a host that you want to move to the new IBM FlashSystem 9100 cluster and find all the old volumes that you must move.
2. Zone your host to the new IBM FlashSystem 9100 cluster.
3. Create volumes in the new IBM FlashSystem 9100 cluster of the same size as the old volumes in the old IBM FlashSystem 9100 cluster.
4. Map the new volumes from the new IBM FlashSystem 9100 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 9100 cluster.
9. Unmap the old volumes from the old IBM FlashSystem 9100 cluster to the host.

This approach uses the server's computing resources (CPU, memory, and I/O) to replicate the data. It can be done online if properly planned. Before you begin, make sure it has enough spare resources.

The biggest benefit to using either approach is that they easily accommodate (if necessary) the replacement of your SAN switches or your back-end storage controllers. You can upgrade the capacity of your back-end storage controllers or replace them entirely, as you can replace your SAN switches with bigger or faster ones. However, you do need to have spare resources, such as floor space, power, cables, and storage capacity, available during the migration.

9.6.4 Splitting an IBM FlashSystem 9100 cluster

Splitting an IBM FlashSystem 9100 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 9100 system that handles all I/O operations, hosts, and subsystem storage attachments. The goal is to create a second IBM FlashSystem 9100 system 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 9100 system, choose from the following approaches to balance the two systems:

- Attach new storage subsystems and hosts to the new system, and start putting only new workload on the new system.
- Migrate the workload onto the new system by using the approach described in 9.6.3, "Moving to a new IBM FlashSystem 9100 cluster" on page 347.

9.7 Adding expansion enclosures

If you plan well, you can buy an IBM FlashSystem 9100 product with enough internal storage to run your business for some time. But as time passes and your environment grows, you will need to add more storage to your system.

Depending on the IBM FlashSystem 9100 product and the code level that you have installed, you can add different numbers of expansion enclosures to your system. Because all IBM FlashSystem 9100 systems were designed to make managing and maintaining them as simple as possible, adding an expansion enclosure is an easy task.

Currently, IBM offers the following SAS expansion enclosures that can be attached to the IBM FlashSystem 9100. 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 card must be installed in each node canister of the IBM FlashSystem 9100 control enclosure.

New SAS-based small form factor (SFF) and large form factor (LFF) expansion enclosures support flash-only MDisks in a storage pool, which can be used for IBM Easy Tier:

- The new IBM FlashSystem 9100 SFF expansion enclosure Model AAF 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 9100 control enclosure. The expansion enclosure is 2U high.
- The new IBM FlashSystem 9100 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 9100 control enclosure. The expansion enclosure is 5U high.

The IBM FlashSystem 9100 system supports up to four control enclosures and up to two chains of SAS expansion enclosures per control enclosure. On each SAS chain, the systems can support up to a SAS chain weight of 10.

- Each 9846-A9F or 9848-A9F expansion enclosure adds a value of 2.5 to the SAS chain weight.
- Each 9846-A9F 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 10:

- Four 9848-A9F enclosures per SAS chain
- Two 9846-A9F enclosures and five 9848-AFF enclosures per SAS chain

Adding expansion enclosures is simplified because IBM FlashSystem 9100 can automatically discover new expansion enclosures after the SAS cables are connected. It is possible to manage and use the new disk drives without managing the new expansion enclosures. However, unmanaged expansion enclosures are not monitored properly. This issue can lead to more difficult troubleshooting and can make problem resolution take longer. To avoid this situation, always manage newly added expansion enclosures.

9.8 I/O Throttling

I/O Throttling is a mechanism that allows you to limit the volume of I/O processed by the storage controller at various levels to achieve QoS. The I/O rate is limited by queuing I/Os if it exceeds the preset limits. I/O Throttling is a way to achieve a better distribution of storage controller resources. IBM FlashSystem 9100 V8.2 code brings the possibility to set the throttling at a volume level, host, host cluster, storage pool, and then offload throttling by using the GUI. This section intends to describe some details of I/O throttling and show how to configure the feature in your system.

9.8.1 General information on I/O Throttling

This is a list of items to keep in mind when thinking about I/O Throttling:

- IOPS and BW throttles limits can be set
- Upper Bound QoS mechanism
- No minimum performance guaranteed
- Volumes, hosts, host clusters and managed disk groups can be throttled
- Queuing at microsecond granularity
- Internal I/Os are not throttled. (such as FlashCopy, cluster traffic, and so on)
- Reduces I/O bursts and smoothens I/O flow with variable delay in throttled I/Os
- Throttle limit is a per node value

9.8.2 I/O Throttling on front end I/O control

You can use throttling for a better front end I/O control, at volume, host, host cluster, and offload levels:

- In a multi tenant environment, hosts can have their own defined limits.

  You can use this to allow restricted I/Os from a data mining server and a higher limit for an application server.

- An aggressive host consuming bandwidth of the controller can be limited by a throttle.

  For example, a video streaming application can have a limit set to avoid consuming too much of the bandwidth.

- Restrict a group of hosts by their throttles.

  For example, Department A gets more bandwidth than Department.

- Each volume can have a throttle defined.

  For example, a backup volume can have less bandwidth than a production volume.

- Offloaded I/Os.

  [XCOPY/WRITESAME(VMware), ODX-WUT (HyperV)] can be confined by defined controller resources.

9.8.3 I/O Throttling on 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 throttle defined.
- Both parent and child pool throttles are supported (non DRP pools).
- Allows control of back-end I/Os from the IBM FlashSystem 9100.
- 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. At the time of writing child pools only applies to standard pools and not DRP pools.
- Parent and child pool throttles are independent of each other. A child pool can have higher throttle limits than its parent pool (non-DRP pools).

9.8.4 Overall benefits of using I/O Throttling

The overall benefits of using I/O Throttling is a better distribution all system resources:

- Avoids overwhelming the controller objects.
- Avoids starving the external entities, *like hosts*, from their share of controller.
- A scheme of distribution of controller resources that, in turn, results in better utilization of external resources such as host capacities.

With no throttling enabled, we have a scenario where Host 1 dominates the bandwidth, and after enabling the throttle, we see a much better distribution of the bandwidth among the hosts, as shown in Figure 9-6.

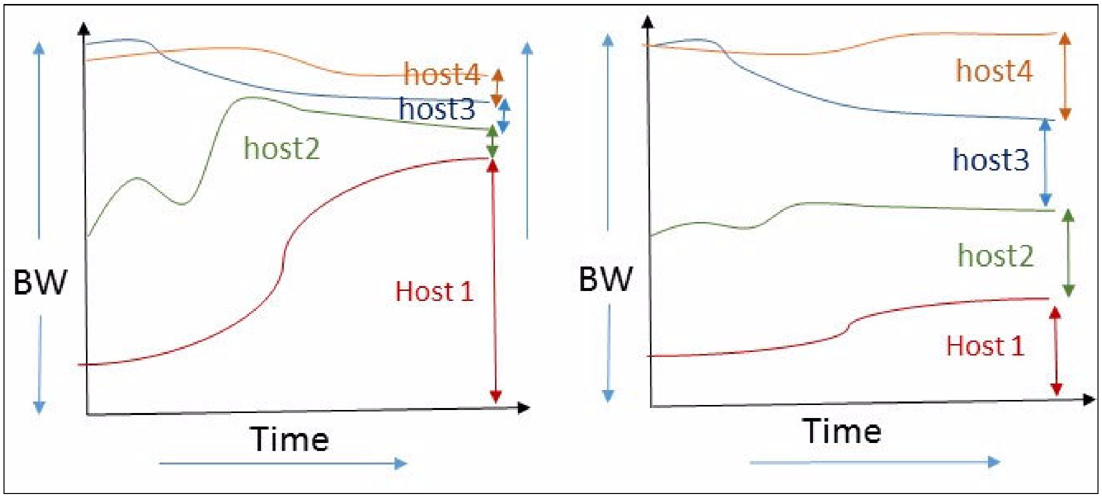


Figure 9-6 Distribution of controller resources after I/O Throttling

9.8.5 Considerations for I/O Throttling

When you are planning to use I/O Throttling there are a few points to be considered:

- The throttle cannot be defined for the host if it is part of a hostcluster which already has a hostcluster throttle.
- If the hostcluster does not have a throttle defined, its member hosts can have their individual host throttles defined.
- The storage pool throttles for child pool and parent pool work independently.
- If a volume has multiple copies then throttling would be done for the storage pool serving the primary copy. The throttling will not be applicable on the secondary pool for mirrored volumes and stretched cluster implementations.
- A host cannot be added to a hostcluster if both of them have their individual throttles defined. If just one of the host/hostcluster throttles is present, the command will succeed.
- A seeding host used for creating a hostcluster cannot have a host throttle defined for it.

> **Note:** Throttling is only applicable at the I/Os that an IBM FlashSystem 9100 receives from hosts and hostclusters. The I/Os generated by IBM FlashSystem 9100 internally, like mirrored volume I/Os, cannot be throttled.

9.8.6 Configuring I/O Throttling using the CLI

In order to create a throttle using the CLI, you use the **mkthrottle** command (Example 9-9).

Example 9-9 Creating a throttle using the mkthrottle command on the CLI

```
Syntax: (Highlighted are the new options on version 8.2)

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:FLASHPFE95:superuser>mkthrottle -type host -bandwidth 100 -host
ITSO_HOST3
IBM_FlashSystem:FLASHPFE95:superuser>mkthrottle -type hostcluster -iops 30000
-hostcluster ITSO_HOSTCLUSTER1
IBM_FlashSystem:FLASHPFE95:superuser>mkthrottle -type mdiskgrp -iops 40000
-mdiskgrp 0
IBM_FlashSystem:FLASHPFE95:superuser>mkthrottle -type offload -bandwidth 50
IBM_FlashSystem:FLASHPFE95:superuser>mkthrottle -type vdisk -bandwidth 25 -vdisk
volume1

IBM_FlashSystem:FLASHPFE95:superuser>lsthrottle
throttle_id throttle_name object_id object_name       throttle_type IOPs_limit
bandwidth_limit_MB
0         throttle0   2       ITSO_HOST3      host                 100
1           throttle1     0         ITSO_HOSTCLUSTER1 hostcluster
30000
2           throttle2     0         Pool0             mdiskgrp
40000
3         throttle3                          offload              50
4           throttle4     10        volume1           vdisk                    25
```

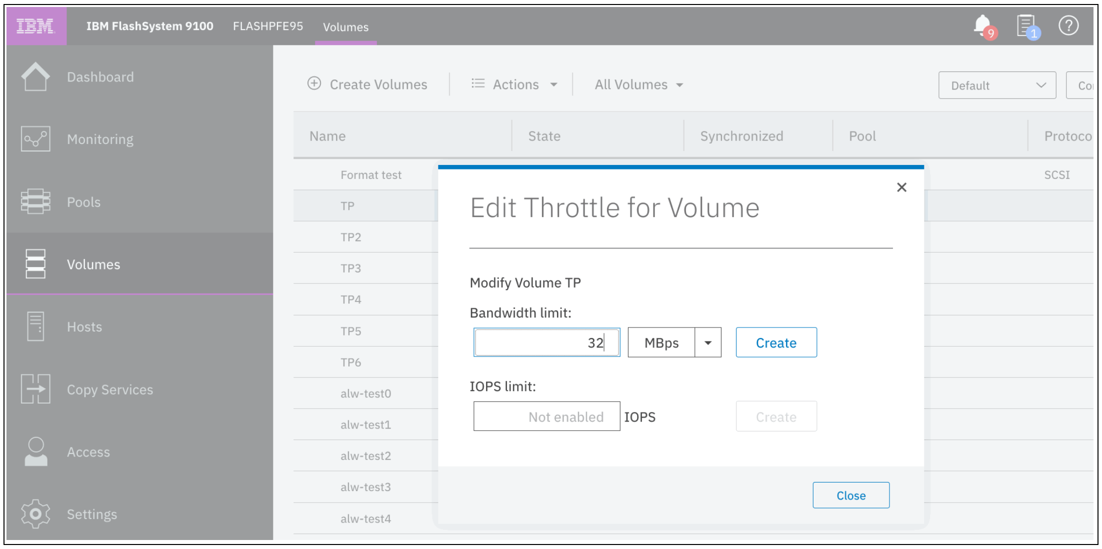
Note: You can change a throttle parameter by using the *chthrottle* command.

9.8.7 Configuring I/O Throttling using the GUI

The following pages shows how to configure the throttle by using the GUI.

Creating a volume throttle

To create a volume throttle, go to **Volumes** → **Volumes**, then select the desired volume, right click on it and chose *Edit Throttle*, as shown in Figure 9-7.



Edit Throttle for Volume

Modify Volume TP

Bandwidth limit: 32 MBps Create

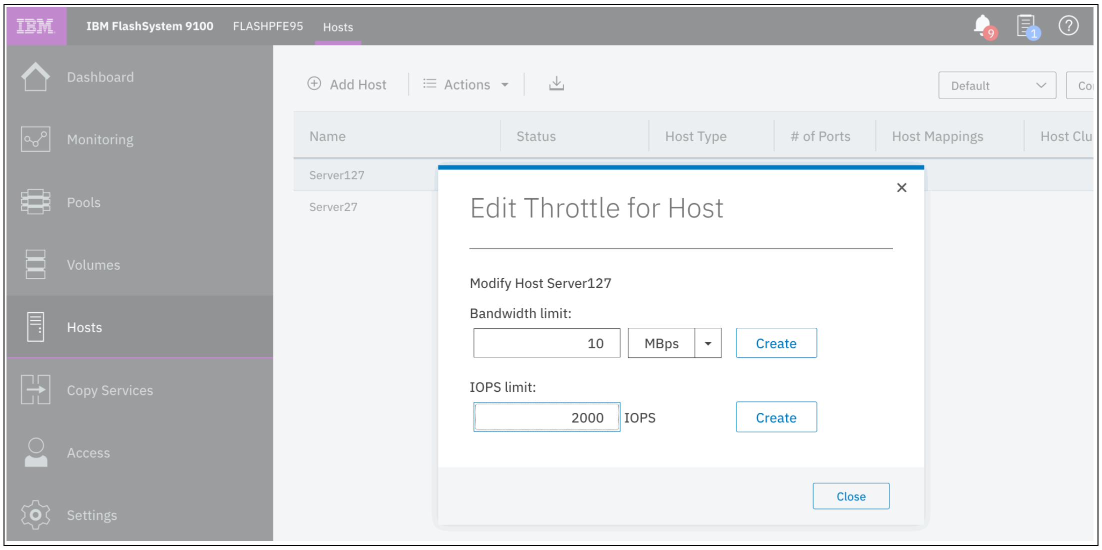
IOPS limit: Not enabled IOPS Create

Close

Figure 9-7 Creating a volume throttle on the GUI

Creating a host throttle

To create a host throttle, go to **Hosts** → **Hosts**, select the desired host, then right-click it and chose **Edit Throttle**, as shown in Figure 9-8.



Edit Throttle for Host

Modify Host Server127

Bandwidth limit: 10 MBps Create

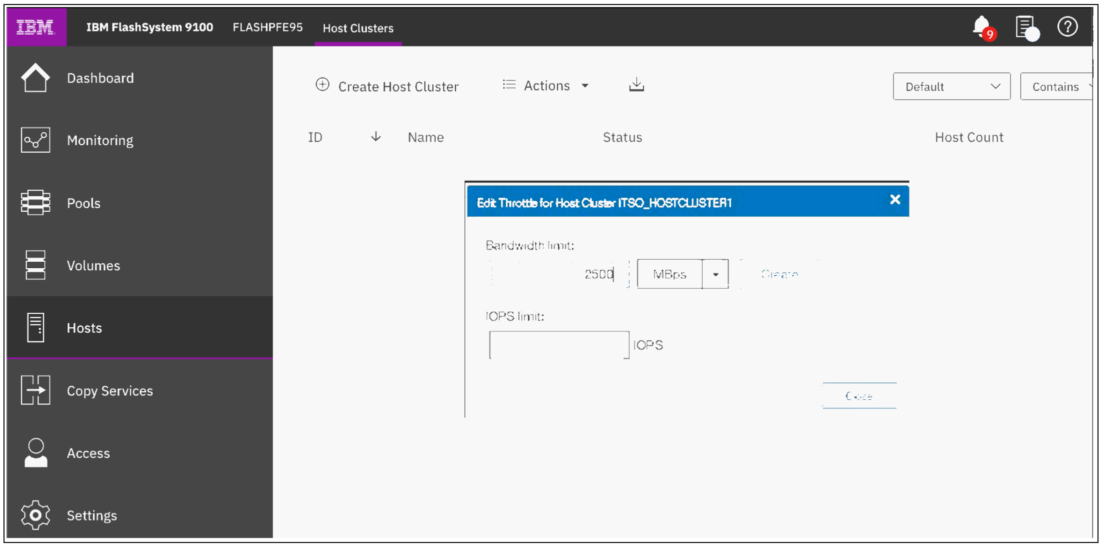
IOPS limit: 2000 IOPS Create

Close

Figure 9-8 Creating a host throttle on GUI

Creating a host cluster throttle

To create a host cluster throttle, go to **Hosts** → **Host Clusters**, select the desired host cluster, then right-click it and chose **Edit Throttle**, as shown in Figure 9-9.



Edit Throttle for Host Cluster ITSO_HOSTCLUSTER1

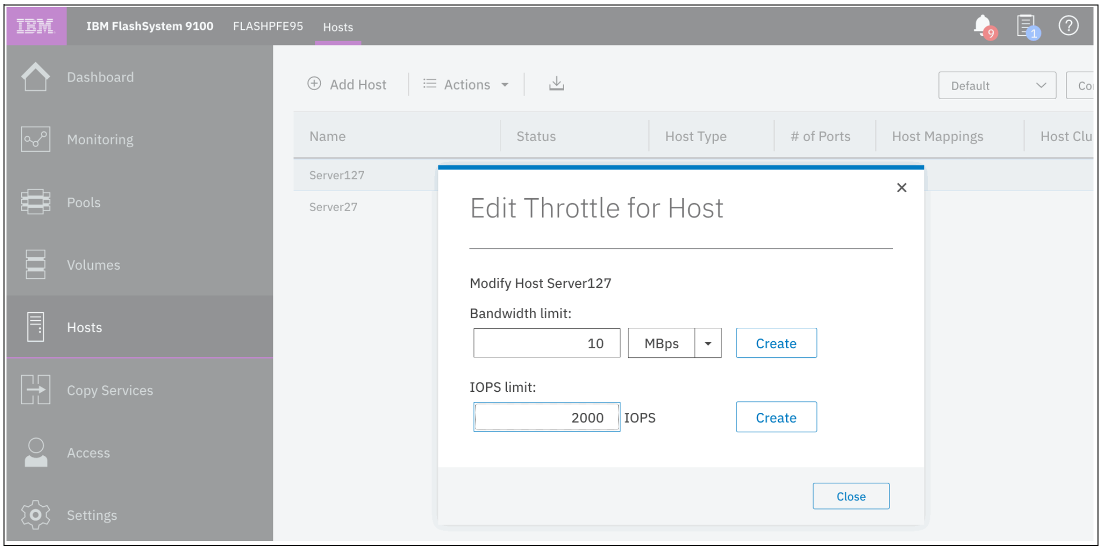
Bandwidth limit: 2500 MBps

IOPS limit: IOPS

Figure 9-9 Creating a host cluster throttle on GUI

Creating a storage pool throttle

To create a storage pool throttle, go to **Pools** → **Pools**, select the desired storage pool, then right click on it and chose *Edit Throttle*, as shown in Figure 9-10.



Edit Throttle for Host

Modify Host Server127

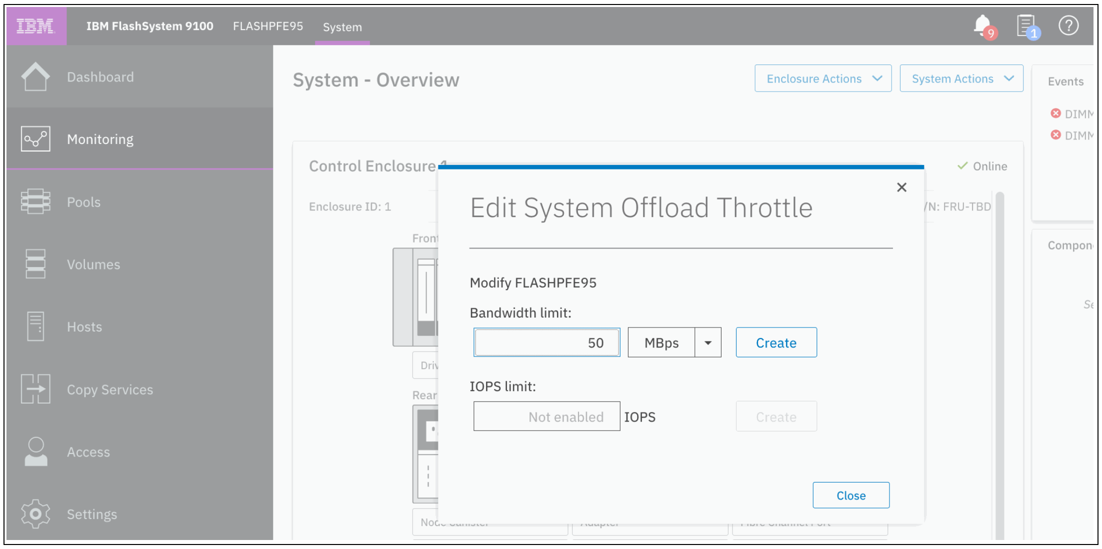
Bandwidth limit: 10 MBps

IOPS limit: 2000 IOPS

Figure 9-10 Creating a storage pool throttle on GUI

Creating an offload throttle

To create an offload throttle, go to **Monitoring** → **System** → **Actions**, then select **Edit System Offload Throttle**, as shown in Figure 9-11.



Edit System Offload Throttle

Modify FLASHPFE95

Bandwidth limit: 50 MBps

IOPS limit: Not enabled IOPS

Figure 9-11 Creating system offload throttle on GUI

10

Troubleshooting and diagnostics

IBM FlashSystem 9100 is a robust and reliable virtualization storage solution that demonstrates excellent availability in the field. However, today's storage area networks (SANs), storage subsystems, and host systems are external components that might cause some events.

This chapter provides useful information to start your troubleshooting and an overview of common events that can occur in your environment. It describes situations that are related to IBM FlashSystem 9100, the SAN environment, optional external storage subsystems, hosts, and multipathing drivers. It also explains how to collect the necessary problem determination data.

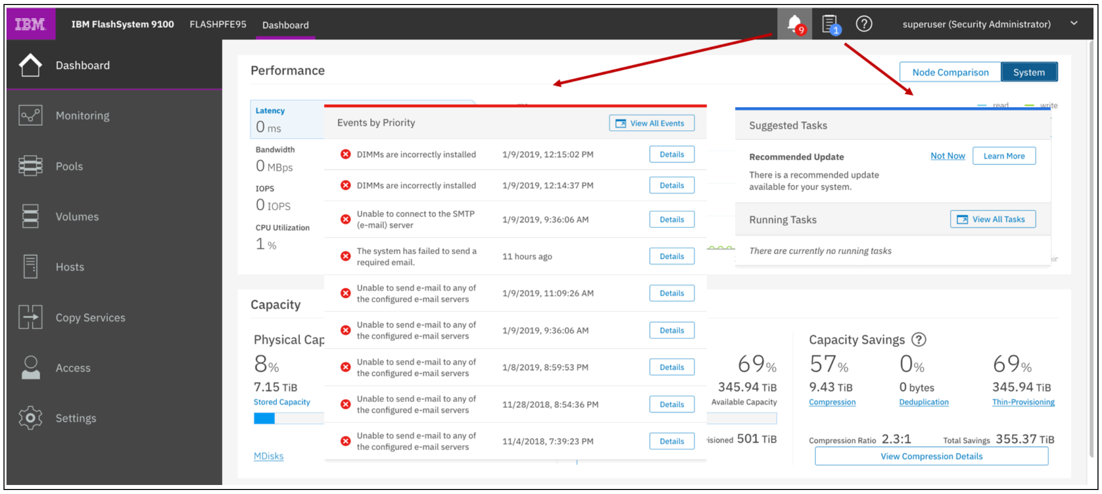
> **Important, IBM FlashSystem 9200:** On 11th February 2020 IBM announced the arrival of the IBM FlashSystem 9200 to the family. This book was written specifically for IBM FlashSystem 9100, however most of the general principles will apply to the IBM FlashSystem 9200.
>
> If you are in any doubt as to their applicability to the FlashSystem 9200 then you should work with your local IBM representative. This book will be updated to include FlashSystem 9200 in due course.
>
> The Flashsystem 9200 product guide is available at:
>
> `IBM FlashSystem 9200 Product Guide`

This chapter includes the following sections:

- Starting troubleshooting
- Remote Support Assistance
- Common issues
- Collecting data and isolating the problem
- Recovering from problems
- Health status during upgrade and known error
- Call Home Web and Health Checker feature
- IBM Storage Insights
- Out of Space monitoring and recovery

10.1 Starting troubleshooting

The Graphical User Interface (GUI) is a good start point for your troubleshooting. It has two icons at the top, which can be accessed from any panel of the GUI. As shown in Figure 10-1, the first icon shows IBM FlashSystem 9100 events, like an error or warning, and the second icon shows suggested tasks and background tasks that are running, or that were recently completed.



Events by Priority — View All Events

DIMMs are incorrectly installed	1/9/2019, 12:15:02 PM	Details
DIMMs are incorrectly installed	1/9/2019, 12:14:37 PM	Details
Unable to connect to the SMTP (e-mail) server	1/9/2019, 9:36:06 AM	Details
The system has failed to send a required email.	11 hours ago	Details
Unable to send e-mail to any of the configured e-mail servers	1/9/2019, 11:09:26 AM	Details
Unable to send e-mail to any of the configured e-mail servers	1/9/2019, 9:36:06 AM	Details
Unable to send e-mail to any of the configured e-mail servers	1/8/2019, 8:59:53 PM	Details
Unable to send e-mail to any of the configured e-mail servers	11/28/2018, 8:54:36 PM	Details
Unable to send e-mail to any of the configured e-mail servers	11/4/2018, 7:39:23 PM	Details

Suggested Tasks

Recommended Update — Not Now — Learn More

There is a recommended update available for your system.

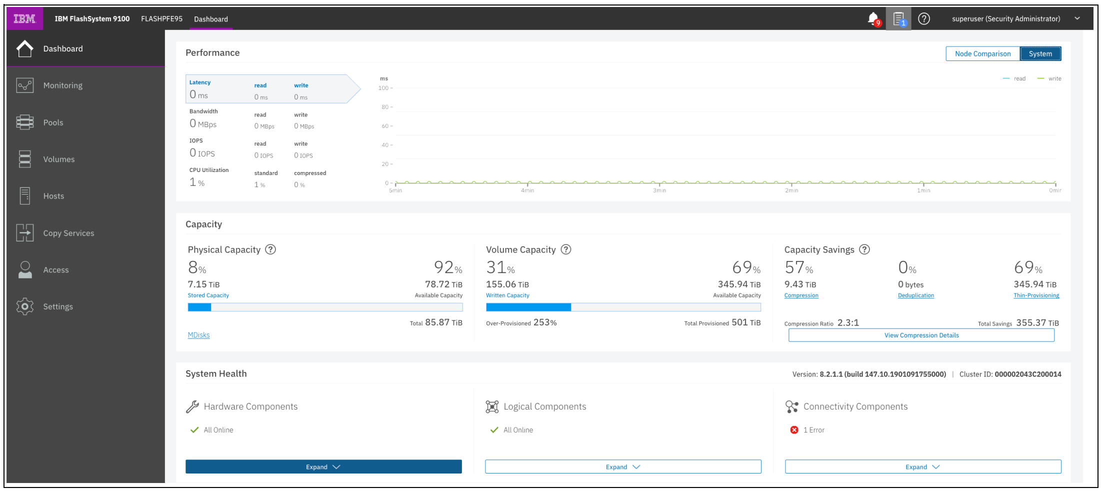
Running Tasks — View All Tasks

There are currently no running tasks

Figure 10-1 Events and tasks icons in GUI

The Dashboard provides an at-a-glance look into the condition of the system and notification of any critical issues that require immediate action. It contains sections for performance, capacity, and system health that provide an overall understanding of what is happening on the system.

Figure 10-2 shows the Dashboard panel displaying the system health panels as well.



Version: 8.2.1.1 (build 147.10.1901091755000) | Cluster ID: 000002043C200014

Hardware Components	Logical Components	Connectivity Components
✓ All Online	✓ All Online	☒ 1 Error

Figure 10-2 Dashboard showing system health

The System Health section in the bottom part of the Dashboard provides information on the health status of hardware, and logical and connectivity components. If you click **Expand** in each of these categories, the status of individual components is shown, as shown in the example in Figure 10-3. You can also go further and click **More Details**, which will take you to the panel related to that specific component, or will show you more information about it.

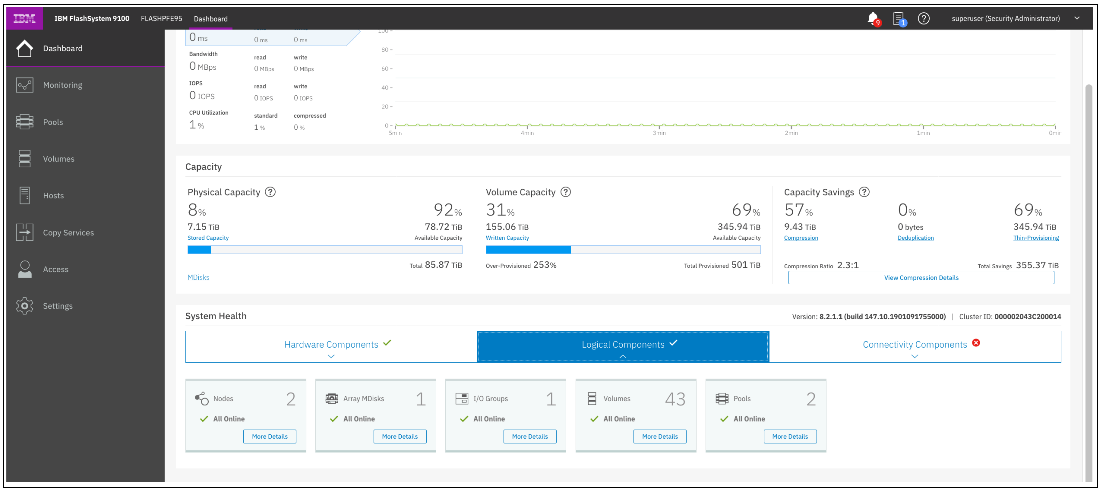


Figure 10-3 System Health expanded section in dashboard

The entire list of components in each category can be found in IBM Knowledge Center for:

IBM FlashSystem 9100 - System Health

More information about IBM FlashSystem 9100 troubleshooting can be found in IBM Knowledge Center:

IBM FlashSystem 9100 - Troubleshooting

10.1.1 Recommended actions and fix procedure

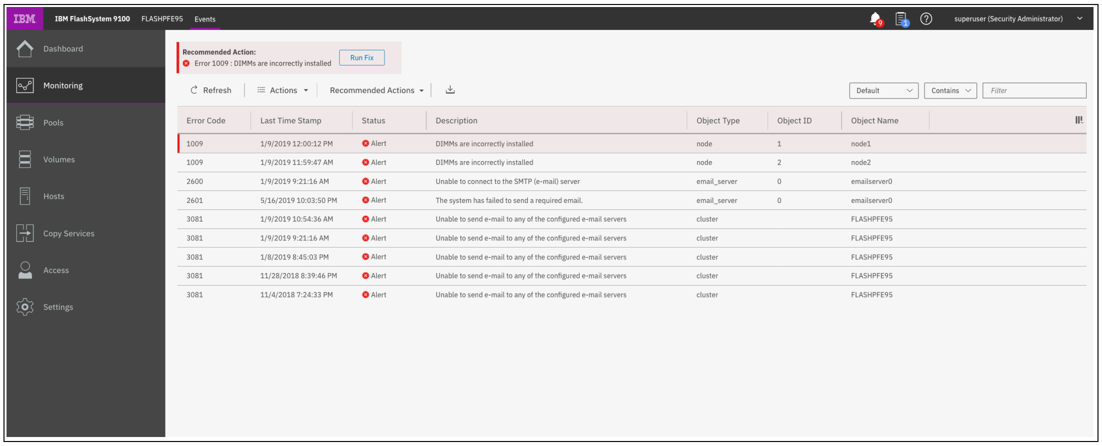
The **Monitoring** → **Events** panel shows information messages, warnings and issues on the IBM FlashSystem 9100. So, this is a good place to check the current problems in the system.

Using the **Recommended Actions** filter, the most important events that need to be fixed are displayed.

If there is an important issue that needs to be fixed, the **Run Fix** button will be available in the top-left corner with an error message, indicating which event should be fixed as soon as possible. This fix procedure assists you to resolve problems in IBM FlashSystem 9100. It analyzes the system, provides more information on the problem, suggest actions to be taken with steps to be followed, and finally checks to see if the problem is resolved.

Note: If any error is reported by the system, such as system configuration problems or hardware failures, always use the fix procedures to resolve it.

Figure 10-4 shows **Monitoring** → **Events** panel with the **Run Fix** button.



Recommended Action:
Error 1009 : DIMMs are incorrectly installed
Run Fix

Error Code	Last Time Stamp	Status	Description	Object Type	Object ID	Object Name
1009	1/9/2019 12:00:12 PM	Alert	DIMMs are incorrectly installed	node	1	node1
1009	1/9/2019 11:59:47 AM	Alert	DIMMs are incorrectly installed	node	2	node2
2600	1/9/2019 9:21:16 AM	Alert	Unable to connect to the SMTP (e-mail) server	email_server	0	emailserver0
2601	5/16/2019 10:03:50 PM	Alert	The system has failed to send a required email.	email_server	0	emailserver0
3081	1/9/2019 10:54:36 AM	Alert	Unable to send e-mail to any of the configured e-mail servers	cluster		FLASHPFE95
3081	1/9/2019 9:21:16 AM	Alert	Unable to send e-mail to any of the configured e-mail servers	cluster		FLASHPFE95
3081	1/8/2019 8:45:03 PM	Alert	Unable to send e-mail to any of the configured e-mail servers	cluster		FLASHPFE95
3081	11/28/2018 8:39:46 PM	Alert	Unable to send e-mail to any of the configured e-mail servers	cluster		FLASHPFE95
3081	11/4/2018 7:24:33 PM	Alert	Unable to send e-mail to any of the configured e-mail servers	cluster		FLASHPFE95

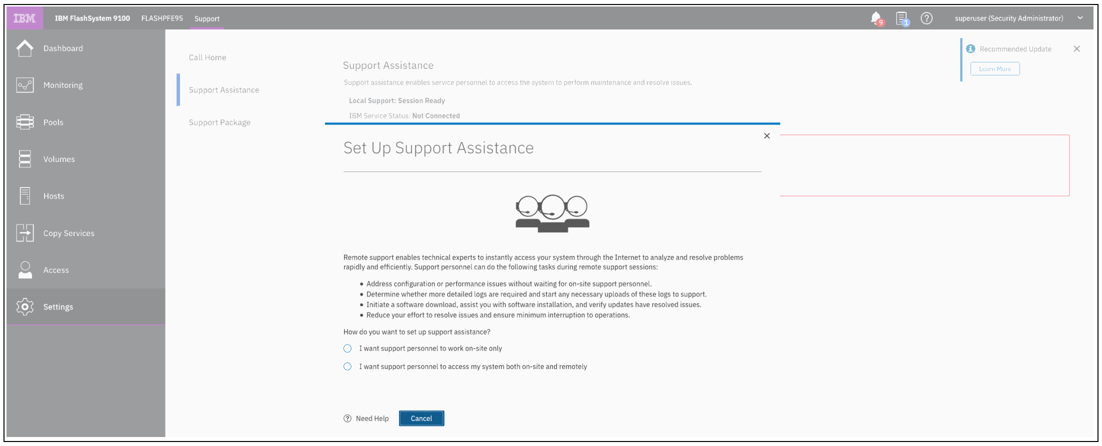
Figure 10-4 Monitoring > Events panel

Resolve alerts in a timely manner: When an issue or a potential issue is reported, resolve it as quick as possible to minimize its impact and potentially avoid more serious problems with your system.

10.2 Remote Support Assistance

Remote Support Assistance (RSA) enables IBM support to access the IBM FlashSystem 9100 device to perform troubleshooting and maintenance tasks. Support assistance can be configured to support personnel work on-site only, or to access the system both on-site and remotely. Both methods use secure connections to protect data in the communication between support center and system. Also, you can audit all actions that support personnel conduct on the system.

Figure 10-5 shows how to set up the remote support options in the GUI by selecting **Settings** → **Support** → **Support Assistance** → **Reconfigure Settings**.



Support Assistance

Support assistance enables service personnel to access the system to perform maintenance and resolve issues.

Local Support: Session Ready

IBM Service Status: Not Connected

Set Up Support Assistance

Remote support enables technical experts to instantly access your system through the Internet to analyze and resolve problems rapidly and efficiently. Support personnel can do the following tasks during remote support sessions:

- Address configuration or performance issues without waiting for on-site support personnel.
- Determine whether more detailed logs are required and start any necessary uploads of these logs to support.
- Initiate a software download, assist you with software installation, and verify updates have resolved issues.
- Reduce your effort to resolve issues and ensure minimum interruption to operations.

How do you want to set up support assistance?

○ I want support personnel to work on-site only

○ I want support personnel to access my system both on-site and remotely

Need Help
Cancel

Figure 10-5 Remote Support options

You can use just local support assistance if you have security restrictions that don’t allow support to connect remotely to your systems. With Remote Support Assistance, support personnel can work both on-site or remotely through a secure connection from the support center. They can perform troubleshooting, upload support packages and download software to the system with your permission. When you configure remote support assistance in the GUI, local support assistance is enabled too.

With the remote support assistance method, you have two access types:

- At any time

  Support center can start remote support sessions at any time.

- By permission only

  Support center can start a remote support session only if permitted by an administrator. A time limit can be configured for the session.

> **Note:** Systems purchased with 3 years warranty include Enterprise Class Support (ECS) and are entitled to IBM support using Remote Support Assistance to quickly connect and diagnose problems. However IBM support might choose to utilize this feature on non-ECS systems at their discretion, therefore we recommend configuring and testing the connection on all systems.

To configure remote support assistance, the following prerequisites should be met:

- Ensure that call home is configured with a valid email server.
- Ensure that a valid service IP address is configured on each node on the system.
- If your system is behind a firewall or if you want to route traffic from multiple storage systems to the same place, you must configure a Remote Support Proxy server. Before you configure remote support assistance, the proxy server must be installed and configured separately. The IP address and the port number for the proxy server needs to be set-up on when enabling remote support centers.
- For more information about setting up the Remote Proxy Server, see IBM Knowledge Center:
  - IBM Remote Proxy Server
- If you do not have firewall restrictions and the storage nodes are directly connected to the Internet, request your network administrator to allow connections to 129.33.206.139 and 204.146.30.139 on port 22.
- Both uploading support packages and downloading software require direct connections to the Internet. A DNS server must be defined on your system for both of these functions to work. 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, and 129.35.224.107.

Remote support assistance can be configured both using GUI and CLI. The detailed steps to configure it can be found in the following publication:

- *IBM FlashSystem 9100 Architecture, Performance, and Implementation*, SG24-8425

10.3 Common issues

SANs, storage subsystems and host systems can be complicated. They often consist of hundreds or thousands of disks, multiple redundant subsystem controllers, virtualization engines, and different types of SAN switches. All of these components must be configured, monitored, and managed properly. If issues occur, administrators must know what to look for and where to look.

IBM FlashSystem 9100 has useful error logging mechanisms. It keeps track of its internal events and informs the user about issues in the SAN or storage subsystem. It also helps to isolate problems with the attached host systems. So, with these functions, administrators can easily locate any issue areas and take the necessary steps to fix any events.

In many cases, IBM FlashSystem 9100 and its service and maintenance features guide administrators directly, provide help, and suggest remedial action. Furthermore, IBM FlashSystem 9100 determines whether the problem still persists or not.

Another feature that helps administrators to isolate and identify issues that might be related to IBM FlashSystem 9100 is the ability of their nodes to maintain a database of other devices that communicate with the IBM FlashSystem 9100 device. Devices, like hosts and optional back-end storages, are added or removed from the database as they start or stop communicating to IBM FlashSystem 9100.

Although IBM FlashSystem 9100 node hardware and software events can be verified in the GUI or CLI, external events like failures in the SAN zoning configuration, hosts, and back-end storages are common. They need to have a troubleshooting performed outside of IBM FlashSystem 9100, too. As an example, a misconfiguration in the SAN zoning might lead to the IBM FlashSystem 9100 cluster not working properly. This problem occurs because the IBM FlashSystem 9100 cluster nodes communicate with each other by using the Fibre Channel SAN fabrics.

In this case, check the following areas from an IBM FlashSystem 9100 perspective:

- The attached hosts. For more information, see 10.3.1, “Host problems” on page 362.
- The SAN. For more information, see 10.3.2, “SAN events” on page 364.
- The optional attached storage subsystem. For more information, see 10.3.3, “Storage subsystem issues” on page 364.
- The local FC port masking. For more information, see 10.3.4, “Port masking issues” on page 369.

10.3.1 Host problems

From the host perspective, you can experience various situations that range from performance degradation to inaccessible disks. To diagnose any host-related issue, you can start checking the hosts configuration on IBM FlashSystem 9100 side. The *Hosts* panel in the GUI or the following CLI commands should be used to start a verification in any possible hosts related issue:

- **`lshost`**

  Check the host’s status. If status is `online`, the host ports are online in both nodes of an I/O group. If status is `offline`, the host ports are offline in both nodes of an I/O group. If status is `inactive`, it means that the host has volumes mapped to it, but all of its ports have no SCSI commands in the last 5 minutes. Also, if status is `degraded`, it means at least one but not all of the host ports is not online in at least one node of an I/O group.

Example 10-1 shows the **lshost** command output.

Example 10-1 lshost command

```
IBM_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>**

  It shows more details about a specific host, and it is very often used in a case you need to identify which host port is not online in IBM FlashSystem 9100 node. Example 10-2 shows the **lshost <host_id_or_name>** command output.

Example 10-2 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 (FC) connectivity between nodes and hosts. Example 10-3 shows the **lsfabric -host <host_id_or_name>** command output.

Example 10-3 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:

- Any special software that you are using
- Any recent change in the OS, such as patching the OS, 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.

For more information about managing hosts on IBM FlashSystem 9100, see Chapter 7, “Hosts” on page 241.

10.3.2 SAN events

Introducing IBM FlashSystem 9100 into your SAN environment and the use of its virtualization functions are not difficult tasks. However, before you can use IBM FlashSystem 9100 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 9100 in your environment: A *host zone* and a *storage zone* for any optional external attached storage. In addition, you must have an IBM FlashSystem 9100 zone that contains all of the IBM FlashSystem 9100 node ports of the IBM FlashSystem 9100 cluster. This IBM FlashSystem 9100 zone enables intracluster communication. For more information and important points about setting up IBM FlashSystem 9100 in a SAN fabric environment, see Chapter 2, “Storage area network” on page 11.

Because IBM FlashSystem 9100 is in the middle of the SAN and connects the host to the storage subsystem, check and monitor the SAN fabrics.

10.3.3 Storage subsystem issues

Today, various heterogeneous storage subsystems are available. All of these subsystems have different management tools, different setup strategies, and possible problem areas depending on the manufacturer. To support a stable environment, all subsystems must be correctly configured, following the respective preferred practices and with no existing issues.

Check the following areas if you experience a storage-subsystem-related issue:

- Storage subsystem configuration. Ensure that a valid configuration and preferred practices are applied to the subsystem.
- Storage subsystem node controllers. Check the health and configurable settings on the node controllers.
- Storage subsystem array. Check the state of the hardware, such as a FCM’s, SSD’s 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 9100 has several CLI commands that you can use to check the status of the system and attached optional storage subsystems too. Before you start a complete data collection or problem isolation on the SAN or subsystem level, use the following commands first and check the status from the IBM FlashSystem 9100 perspective:

- **`lsMDisk`**

  Check that all MDisks are online (not degraded or offline).

- **`lsMDisk <MDiskid_id_or_name>`**

  Check several of the MDisks from each storage subsystem controller. Are they online? See Example 10-4 for an example of the output from this command.

Example 10-4 Issuing an 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 6005076441b530044000000000000001000000000000000000000000000000000
preferred_WWPN
active_WWPN many

NOTE: lines removed for brevity
```

Example 10-4 on page 365 shows that for MDisk 1, the external FlashSystem FS900 has eight ports zoned to IBM FlashSystem 9100, and IBM FlashSystem 9100 has two nodes, so 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 any stopped FlashCopy jobs or seek IBM FlashSystem 9100 support guidance.

- **lsfabric**

  Use this command with the various options, such as **-controller *controllerid***. Also, check different parts of the IBM FlashSystem 9100 configuration to ensure that multiple paths are available from each IBM FlashSystem 9100 node port to an attached host or controller. Confirm that all IBM FlashSystem 9100 node port WWPNs are also connected to any of the optional external back-end storage consistently.

Example 10-5 shows the output of the **lsfabric** command.

Example 10-5 Example output of lsfabric CLI command

```
IBM_FlashSystem:FLASHPFE95:superuser>lsfabric
remote_wwpn      remote_nportid id node_name local_wwpn       local_port
local_nportid state    name      cluster_name type
10000090FAB386E6 01DDC0       3 node1    5005076810110230 1        550200
inactive                          unknown
10000090FAB386E6 01DDC0       1 node2    5005076810110242 1        550000
inactive                          unknown
5005076810120242 540000       3 node1    5005076810120230 2        540200
active   node2 FLASHPFE95 node
10000090FAB386A3 502100       3 node1    5005076810120230 2        540200
inactive Win2012srv1              host
10000090FAB386A3 502100       1 node2    5005076810120242 2        540000
inactive Win2012srv1              host
5005076810110242 550000       3 node1    5005076810110230 1        550200
active   node2 FLASHPFE95 node
5005076810110230 550200       1 node2    5005076810110242 1        550000
active   node1 FLASHPFE95 node
10000090FAB386E7 502000       3 node1    5005076810120230 2        540200
inactive                          unknown
10000090FAB386E7 502000       1 node2    5005076810120242 2        540000
inactive                          unknown
10000090FAB386A2 010001       3 node1    5005076810110230 1        550200
inactive Win2012srv1              host
10000090FAB386A2 010001       1 node2    5005076810110242 1        550000
inactive Win2012srv1              host
5005076810120230 540200       1 node2    5005076810120242 2        540000
active   node1 FLASHPFE95 node
2100001B32157539 241B00       3 node1    5005076810110230 1        550200
inactive                          unknown
2100001B32157539 241B00       1 node2    5005076810110242 1        550000
inactive                          unknown
```

Determining the number of paths to an external storage subsystem

By using IBM FlashSystem 9100 CLI commands, it is possible to determine the total number of paths to an optional external storage subsystem. To determine the proper value of the available paths, use the following formula:

```
Number of MDisks x Number of FS9100 nodes per Cluster = Number of paths
MDisk_link_count x Number of FS9100 nodes per Cluster = Sum of path_count
```

Example 10-6 shows how to obtain this information by using the **lscontroller <controllerid>** and **svcinfo lsnode** commands.

Example 10-6 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 10-6 on page 367 shows that sixteen MDisks are present for the external storage subsystem controller with ID 1, and two IBM FlashSystem 9100 nodes are in the cluster. In this example, the `path_count` is 16 x 2 = 32.

Further information about FC and iSCSI configurations can be found in the IBM Knowledge Center here:

IBM FS9100 External Storage Reference

10.3.4 Port masking issues

Some situations of performance degradation and buffer-to-buffer credit exhaustion can be caused by incorrect local FC port masking and remote FC port masking. To ensure healthy operation of your IBM FlashSystem 9100, configure both your local FC port masking and your remote FC port masking accordingly.

The ports intended to have only intracluster/node to node communication traffic must not have replication data or host/back-end data running on it. The ports intended to have only replication traffic must not have intracluster/node to node communication data or host/back-end data running on it.

10.3.5 Interoperability

When you experience events in the IBM FlashSystem 9100 environment, ensure that all components that comprise the storage infrastructure are interoperable. In an IBM FlashSystem 9100 environment, the IBM FlashSystem 9100 support matrix is the main source for this information. For the latest IBM FlashSystem 9100 support matrix, see *IBM System Storage Interoperations Center (SSIC)* website:

`http://www.ibm.com/systems/support/storage/ssic/interoperability.wss`

Although the latest IBM FlashSystem 9100 code level is supported to run on older host bus adapters (HBAs), storage subsystem drivers, and code levels, use the latest tested levels for best results.

10.4 Collecting data and isolating the problem

Data collection and problem isolation in an IT environment are sometimes difficult tasks. In the following section, the essential steps that are needed to collect debug data to find and isolate problems in an IBM FlashSystem 9100 environment are described.

10.4.1 Collecting data from IBM FlashSystem 9100

When there is a problem with an IBM FlashSystem 9100 and you have to open a case with IBM support, you need to provide the support packages for the device. To collect and upload the support packages to IBM support center, you can do it automatically via IBM FlashSystem 9100, or download the package from the device and manually upload to IBM. The easiest way is automatically upload the support packages from IBM FlashSystem 9100. It can be done via both GUI and CLI.

You can also use the new IBM Storage Insights application to do the log data upload and this is described in 10.8.4, “Updating support tickets” on page 394.

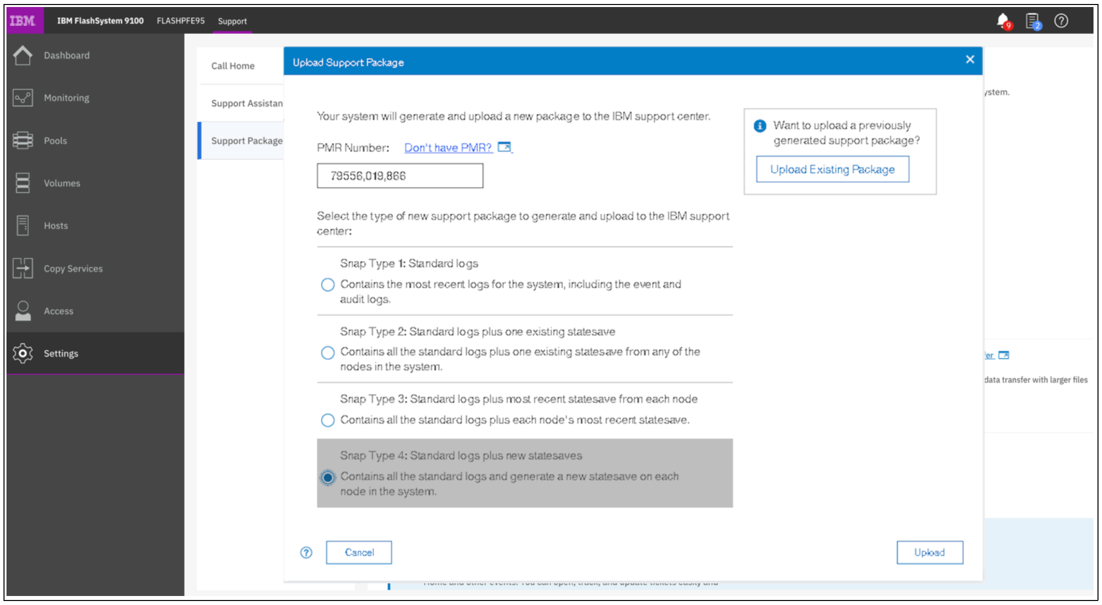
The next sections show how to automatically upload the support package to IBM support center. More details of this procedure can be found in:

- *IBM FlashSystem 9100 Architecture, Performance, and Implementation*, SG24-8425

Data collection using the GUI

To perform data collection using the GUI, complete the following steps:

1. In the panel **Settings** → **Support** → **Support Package**, both options to collect and upload support packages are available
2. To automatically upload them, click **Upload Support Package** button
3. In the pop-up screen, enter the PMR number and the type of support package to upload to the IBM support center. The **Snap Type 4** can be used to collect standard logs and generate a new statesave on each node of the system
4. The *Upload Support Package* panel is shown in Figure 10-6.



Your system will generate and upload a new package to the IBM support center.

PMR Number: Don't have PMR?

79556,019,866

Want to upload a previously generated support package?

Upload Existing Package

Select the type of new support package to generate and upload to the IBM support center:

Snap Type 1: Standard logs
○ Contains the most recent logs for the system, including the event and audit logs.

Snap Type 2: Standard logs plus one existing statesave
○ Contains all the standard logs plus one existing statesave from any of the nodes in the system.

Snap Type 3: Standard logs plus most recent statesave from each node
○ Contains all the standard logs plus each node's most recent statesave.

Snap Type 4: Standard logs plus new statesaves
● Contains all the standard logs and generate a new statesave on each node in the system.

Figure 10-6 Upload Support Package panel

Data collection using the CLI

To collect the same type of support packages mentioned above using the CLI, you have to first generate a new `livedump` of the system using the **`svc_livedump`** command, and then upload the log files and new generated dumps using the **`svc_snap`** command, as shown in Example 10-7. To verify if the support package was successfully uploaded, use the **`sainfo lscmdstatus`** command.

Example 10-7 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=ppppp,bbb,ccc 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 ppppp,bbb,ccc -filename
/dumps/snap.ABCDEFG.171128.223133.tgz
last_command_status CMMVC8044E Command completed successfully.
T3_status
T3_status_data
cpfiles_status Complete
cpfiles_status_data Copied 1 of 1
snap_status Complete
snap_filename /dumps/snap.ABCDEFG.171128.223133.tgz
installcanistersoftware_status
supportupload_status Complete
supportupload_status_data [PMR=ppppp,bbb,ccc] Upload complete
supportupload_progress_percent 0
supportupload_throughput_KBps 0
supportupload_filename /dumps/snap.ABCDEFG.171128.223133.tgz
downloadsoftware_status
downloadsoftware_status_data
downloadsoftware_progress_percent 0
downloadsoftware_throughput_KBps 0
downloadsoftware_size
IBM_FlashSystem:FLASHPFE95:superuser>
```

10.4.2 SDDPCM and SDDDSM data collection

If there is a problem related to host communication with IBM FlashSystem 9100, collecting data from hosts and multipath software is very useful.

SDDPCM for AIX provides the **sddpcmgetdata** script to collect information used for problem determination. This script creates a tar file at the current directory with the current date and time as a part of the file name. When you suspect you have an issue with SDDPCM, it is essential to run this script and send this tar file to IBM support.

SDDDSM for Windows hosts also contains a utility to collect information for problem determination. The **sddgetdata.bat** tool creates a CAB file in the installation directory with the current date and time as part of the file name. The CAB file includes the following information:

- SystemInfo
- HKLM \SYSTEM\CurrentControlSet, HKLM\HARDWARE\DEVICEMAP, and HKLM\Cluster output from registry
- SDDDSM directory contents
- HBA details
- Datapath outputs

- Pathtest trace
- SDDSRV logs
- Cluster logs
- System disks and paths

The execution of **`sddgetdata.bat`** tool is shown in Example 10-8.

Example 10-8 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>
```

More information about diagnostics for IBM SDD can be found in the latest *Multipath Subsystem Device Driver User's Guide*:

Multipath Subsystem Device User's Guide

10.4.3 Additional data collection

Data collection methods vary by storage platform, SAN switch and operating system.

When there is an issue in a SAN environment and it is not clear where the problem is occurring, you might have to collect data from several devices in the SAN.

Bellow you can find basic information that should be collected for each type of device:

- Hosts
  - Operating system: Version and level
  - HBA: Driver and firmware level
  - Multipathing driver level
- SAN switches
  - Hardware model
  - Software version

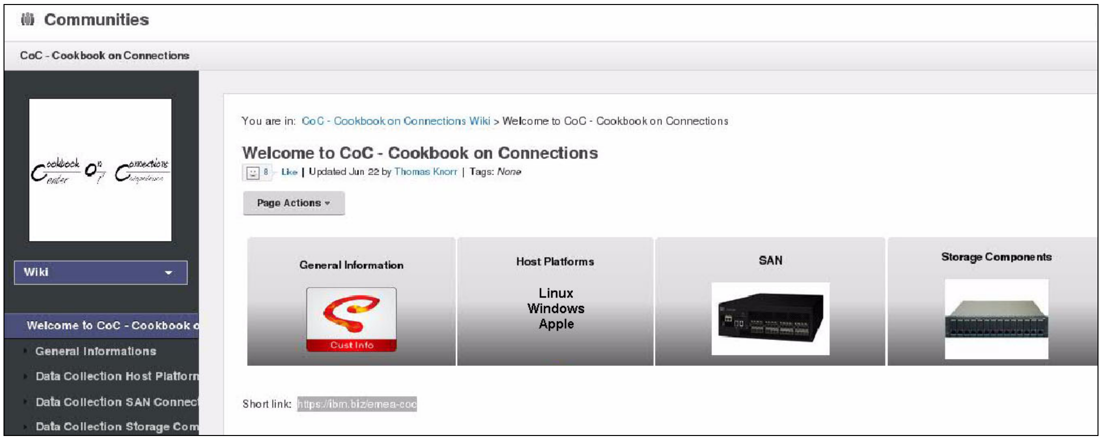
- Storage subsystems
  - Hardware model
  - Software version

Regarding host, storage and SAN data collection, due to the dynamic changes that occur over time, follow this IBM w3 Connections community (only available to IBMers):

IBM CoC - Cookbook on Connections

> **Note:** This community is IBM internal only. You must use your intranet ID and password. If you are not an IBM employee, contact your IBM representative or the vendor of your hardware and follow the specific procedures for data collection.

The w3 Connections community has up-to-date procedures for several kinds of devices, including hosts, storage, and SAN, as shown in Figure 10-7.


Communities
CoC - Cookbook on Connections
Cookbook Center Of Connections
Wiki
Welcome to CoC - Cookbook o
General Informations
Data Collection Host Platforn
Data Collection SAN Connect
Data Collection Storage Com
Page Actions
General Information
Cust Info
Host Platforms
Linux
Windows
Apple
SAN
Storage Components

You are in: CoC - Cookbook on Connections Wiki > Welcome to CoC - Cookbook on Connections

Welcome to CoC - Cookbook on Connections

8 Like | Updated Jun 22 by Thomas Knorr | Tags: *None*

Short link: https://ibm.biz/emea-coc

Figure 10-7 CoC - Cookbook on Connections internal wiki

10.5 Recovering from problems

You can recover from several of the more common events that you might encounter. In all cases, you must read and understand the current product limitations to verify the configuration and to determine whether you need to upgrade any components or install the latest fixes or patches.

To obtain support for any IBM product, see the following IBM Support website:

IBM Support Homepage

For more information about the latest flashes, concurrent code upgrades, code levels, and compatibility matrix, see the following IBM FlashSystem 9100 support website:

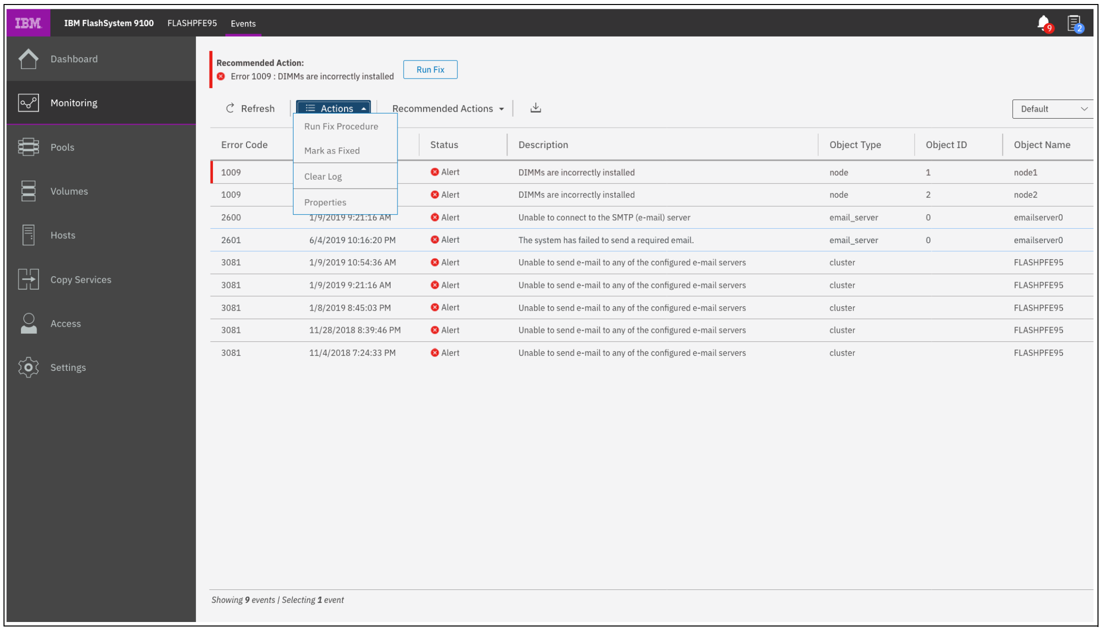
Support for the IBM FlashSystem 9100 family

10.5.1 Solving IBM FlashSystem 9100 events

For any events in the IBM FlashSystem 9100, before you try to fix the problem anywhere else, use the **Recommended Actions** functionality in **Monitoring** → **Events** panel. This is shown in 10.1.1, “Recommended actions and fix procedure” on page 359.

The Events panel with the **Recommended Actions** filter shows event conditions that require actions and the procedures to diagnose and fix them. The highest-priority event is indicated in the top-left corner and it will also appear highlighted in the table of events.

If for any reason you need to run the **Fix Procedure** in another event before fixing the highest-priority event*,* you must select the event, click in **Actions** menu and select Run Fix Procedure, as shown in the Figure 10-8.



Recommended Action:
Error 1009 : DIMMs are incorrectly installed

Error Code		Status	Description	Object Type	Object ID	Object Name
1009		Alert	DIMMs are incorrectly installed	node	1	node1
1009		Alert	DIMMs are incorrectly installed	node	2	node2
2600	1/9/2019 9:21:16 AM	Alert	Unable to connect to the SMTP (e-mail) server	email_server	0	emailserver0
2601	6/4/2019 10:16:20 PM	Alert	The system has failed to send a required email.	email_server	0	emailserver0
3081	1/9/2019 10:54:36 AM	Alert	Unable to send e-mail to any of the configured e-mail servers	cluster		FLASHPFE95
3081	1/9/2019 9:21:16 AM	Alert	Unable to send e-mail to any of the configured e-mail servers	cluster		FLASHPFE95
3081	1/8/2019 8:45:03 PM	Alert	Unable to send e-mail to any of the configured e-mail servers	cluster		FLASHPFE95
3081	11/28/2018 8:39:46 PM	Alert	Unable to send e-mail to any of the configured e-mail servers	cluster		FLASHPFE95
3081	11/4/2018 7:24:33 PM	Alert	Unable to send e-mail to any of the configured e-mail servers	cluster		FLASHPFE95

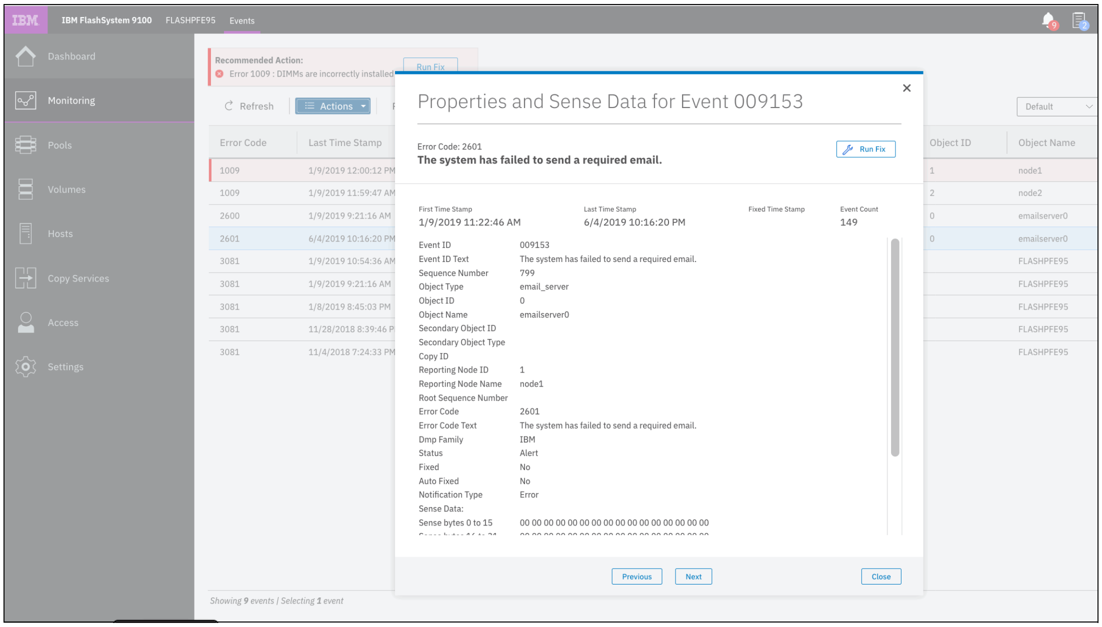
*Showing **9** events | Selecting **1** event*

Figure 10-8 Action menu in events table

To obtain more information about any event, select an event in the table, and click **Properties** in the **Actions** menu, as shown in Figure 10-8.

Tip: You can also get access to **Run Fix Procedure** and **Properties** by right-clicking an event.

In the *Properties and Sense Data* window for the specific event, as shown in Figure 10-9 on page 375, additional information about it is displayed. You can review and also click **Run Fix** to run the **Fix Procedure**.



Recommended Action:
Error 1009 : DIMMs are incorrectly installed

Error Code	Last Time Stamp	Object ID	Object Name
1009	1/9/2019 12:00:12 PM	1	node1
1009	1/9/2019 11:59:47 AM	2	node2
2600	1/9/2019 9:21:16 AM	0	emailserver0
2601	6/4/2019 10:16:20 PM	0	emailserver0
3081	1/9/2019 10:54:36 AM		FLASHPFE95
3081	1/9/2019 9:21:16 AM		FLASHPFE95
3081	1/8/2019 8:45:03 PM		FLASHPFE95
3081	11/28/2018 8:39:46 P		FLASHPFE95
3081	11/4/2018 7:24:33 PM		FLASHPFE95

Showing 9 events | Selecting 1 event

Properties and Sense Data for Event 009153

Error Code: 2601
The system has failed to send a required email.

First Time Stamp	Last Time Stamp	Fixed Time Stamp	Event Count
1/9/2019 11:22:46 AM	6/4/2019 10:16:20 PM		149

Event ID	009153
Event ID Text	The system has failed to send a required email.
Sequence Number	799
Object Type	email_server
Object ID	0
Object Name	emailserver0
Secondary Object ID	
Secondary Object Type	
Copy ID	
Reporting Node ID	1
Reporting Node Name	node1
Root Sequence Number	
Error Code	2601
Error Code Text	The system has failed to send a required email.
Dmp Family	IBM
Status	Alert
Fixed	No
Auto Fixed	No
Notification Type	Error
Sense Data:	
Sense bytes 0 to 15	00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00

Figure 10-9 Properties and sense data for event window

> **Tip:** From the Properties and Sense Data for Event Window, you can use the **Previous** and **Next** buttons to move between events.

Another common practice is to use the IBM FlashSystem 9100 CLI to find issues and resolve them. The following list of commands scan internal NVMe storage and optional back-end storage subsystems changes and provides information about the status of your environment:

- `lseventlog`

  Display a list of events and more detailed information about an event.

- `detectMDisk`

  Discovers changes in the internal IBM FlashSystem 9100 and optional back-end storage configuration.

- `lspartnership`

  Checks the IBM FlashSystem 9100 local and remote clusters status.

- `lssystem`

  Displays detailed information about IBM FlashSystem 9100 cluster.

- `svcinfo lsnode <node_id_or_name>`

  Checks the IBM FlashSystem 9100 nodes and port status.

- `lscontroller <controller_id_or_name>`

  Checks the optional back-end storage status.

- `lsMDisk <MDisk_id_or_name>`

  Provides a status of MDisks.

- `lsMDiskgrp <MDiskgrp_id_or_name>`

  Provides a status for storage pools.

- `lsvdisk <vdisk_id_or_name>`

  Checks whether volumes are online and working correctly.

Locating issues: Although IBM FlashSystem 9100 raises error messages, events might not be caused by IBM FlashSystem 9100. Some issues are introduced by the storage subsystems or the SAN.

If the problem is caused by IBM FlashSystem 9100 and you are unable to fix it by using the Recommended Action feature or the event log, collect the IIBM FlashSystem 9100 support package as described in 10.4.1, “Collecting data from IBM FlashSystem 9100” on page 369. To identify and fix other issues outside of IBM FlashSystem 9100, consider the guidance in the other sections in this chapter that are not related to IBM FlashSystem 9100.

Replacing a failed flash drive

When IBM FlashSystem 9100 detects a failed NVMe drive or optional external attached flash drive, it automatically generates an error in the *Events* panel. To replace the failed drive, always run **Fix Procedure** for this event in **Monitoring** → **Events** panel.

The **Fix Procedure** will help you to identify the enclosure and slot where the bad drive is located, and will guide you to the correct steps to follow in order to replace it. When a 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. So, the failed flash drive is removed from the configuration, and the new drive is then used to fulfill the array membership goals of the system.

10.5.2 Solving host problems

Apart from hardware-related situations, problems can exist in such areas as the operating system or the software that is used on the host. These problems normally are handled by the host administrator or the service provider of the host system. However, the multipathing driver that is installed on the host and its features can help to determine possible issues.

Example 10-9 shows a volume path issue reported by SDD output on the host by using the **`datapath query adapter`** and **`datapath query device`** commands. The adapter in degraded state means that specific HBA on the server side can’t reach all the nodes in the I/O group which the volumes are associated. Also we can notice in the **`datapath query device`** command output that each device (or volume) has only three paths, when it is expected to have four paths.

Example 10-9 SDD output on a host with faulty paths

```
C:\Program Files\IBM\SDDDSM>datapath query adapter

Active Adapters :2

Adpt#        Name      Interface      State        Mode       Select     Errors  Paths   Active
    0  Scsi Port2 Bus0        FC     DEGRAD      ACTIVE      1860589        293      4        4
    1  Scsi Port3 Bus0        FC     NORMAL      ACTIVE      1979793        259      8        8

C:\Program Files\IBM\SDDDSM>datapath query device

Total Devices : 8

DEV#:   0  DEVICE NAME: Disk2 Part0             TYPE: 2145      POLICY: LEAST I/O AND WEIGHT
SERIAL: 600507680C838020E80000000000001B        Reserved: No    LUN SIZE:  5.0GB
```

```
HOST INTERFACE: FC
PREFERRED PATH SET :None
===========================================================================================
Path#     Adapter/Hard Disk                      State         Mode       Select    Errors
    2     Scsi Port2 Bus0/Disk2 Part0             OPEN         NORMAL    2859569          0
    5 *   Scsi Port3 Bus0/Disk2 Part0             OPEN         NORMAL          0          0
    6     Scsi Port3 Bus0/Disk2 Part0             OPEN         NORMAL    2466689          0

DEV#:   1  DEVICE NAME: Disk3 Part0              TYPE: 2145      POLICY: LEAST I/O AND WEIGHT
SERIAL: 600507680C838020E8000000000000001E       Reserved: No    LUN SIZE: 10.0GB
HOST INTERFACE: FC
PREFERRED PATH SET :None
===========================================================================================
Path#     Adapter/Hard Disk                      State         Mode       Select    Errors
    2     Scsi Port2 Bus0/Disk3 Part0             OPEN         NORMAL         35          0
    5 *   Scsi Port3 Bus0/Disk3 Part0             OPEN         NORMAL          0          0
    6     Scsi Port3 Bus0/Disk3 Part0             OPEN         NORMAL         22          0
.
lines removed for brevity
```

Faulty paths can be caused by hardware and software problems, such as the following examples:

- Hardware
  - Faulty Small Form-factor Pluggable transceiver (SFP) on the host or SAN switch
  - Faulty fiber optic cables
  - Faulty HBAs
- Software
  - A back-level multipathing driver
  - Obsolete HBA firmware or driver
  - Wrong zoning
  - Incorrect host-to-VDisk mapping

Based on field experience, complete the following hardware checks first:

- Check whether any connection error indicators are lit on the host or SAN switch.
- Check whether all of the parts are seated correctly. For example, cables are securely plugged in to the SFPs and the SFPs are plugged all the way into the switch port sockets.
- Ensure that no fiber optic cables are broken. If possible, swap the cables with cables that are known to work.

After the hardware check, continue to check the following aspects of software setup:

- Check that the HBA driver level and firmware level are at the preferred and supported levels.
- Check the multipathing driver level, and make sure that it is at the preferred and supported level.
- Check for link layer errors that are reported by the host or the SAN switch, which can indicate a cabling or SFP failure.
- Verify your SAN zoning configuration.
- Check the general SAN switch status and health for all switches in the fabric.

The output of SDD commands also helps to troubleshoot possible connectivity issues to the IBM FlashSystem 9100 device. Example 10-10 shows that one of the HBAs reported errors, and the cause can be any of the hardware or software examples mentioned above.

Example 10-10 Output from datapath query adapter and datapath query device

```
C:\Program Files\IBM\SDDDSM>datapath query adapter

Active Adapters :2

Adpt#        Name      Interface      State        Mode       Select     Errors  Paths   Active
    0  Scsi Port2 Bus0        FC     NORMAL      ACTIVE      1755262         12      8        8
    1  Scsi Port3 Bus0        FC     NORMAL      ACTIVE      1658236          0      8        8

C:\Program Files\IBM\SDDDSM>datapath query device

Total Devices : 8


DEV#:   0  DEVICE NAME: Disk2 Part0            TYPE: 2145      POLICY: LEAST I/O AND WEIGHT
SERIAL: 600507680C838020E80000000000001B       Reserved: No    LUN SIZE:  5.0GB
HOST INTERFACE: FC
PREFERRED PATH SET :None
============================================================================================
Path#     Adapter/Hard Disk                    State         Mode       Select    Errors
    1 *   Scsi Port2 Bus0/Disk2 Part0           OPEN         NORMAL          0         0
    2     Scsi Port2 Bus0/Disk2 Part0           OPEN         NORMAL    1599542        10
    5 *   Scsi Port3 Bus0/Disk2 Part0           OPEN         NORMAL          0         0
    6     Scsi Port3 Bus0/Disk2 Part0           OPEN         NORMAL    1492830         0

DEV#:   1  DEVICE NAME: Disk3 Part0            TYPE: 2145      POLICY: LEAST I/O AND WEIGHT
SERIAL: 600507680C838020E80000000000001E       Reserved: No    LUN SIZE: 10.0GB
HOST INTERFACE: FC
PREFERRED PATH SET :None
============================================================================================
Path#     Adapter/Hard Disk                    State         Mode       Select    Errors
    1 *   Scsi Port2 Bus0/Disk3 Part0           OPEN         NORMAL          0         0
    2     Scsi Port2 Bus0/Disk3 Part0           OPEN         NORMAL          9         0
    5 *   Scsi Port3 Bus0/Disk3 Part0           OPEN         NORMAL          0         0
    6     Scsi Port3 Bus0/Disk3 Part0           OPEN         NORMAL         10         0
.
lines removed for brevity
```

10.5.3 Solving SAN issues

Some situations can cause issues in the SAN fabric and SAN switches. Problems can be related to a hardware fault or to a software problem on the switch. The following hardware defects are normally the easiest problems to find:

- Switch power, fan, or cooling units
- Installed SFP modules
- Fiber optic cables

Software failures are more difficult to analyze. In most cases, you must collect data and involve IBM Support. But before you take any other steps, check the installed code level for any known issues. Also, check whether a new code level is available that resolves the problem that you are experiencing.

The most common SAN issues often are related to zoning. For example, perhaps you chose the wrong WWPN for a host zone, such as when two IIBM FlashSystem 9100 node ports must be zoned to one HBA with one port from each IBM FlashSystem 9100 node. However, as shown in Example 10-11, two ports are zoned that belong to the same node. Therefore, the result is that the host and its multipathing driver do not see all of the necessary paths.

Example 10-11 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 10-12.

Example 10-12 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 9100 error codes are related to the SAN environment:

- Error 1060 - `Fibre Channel ports are not operational`
- Error 1220 - `A remote port is excluded`

Bottleneck is another common issue related to SAN switches. The bottleneck can be presented in a port where a host, storage subsystem or IBM Spectrum Virtualize device is connected, or in Inter-Switch Link (ISL) ports. The bottleneck can occur in some cases, like when a device connected to the fabric is slow to process received frames or if a SAN switch port is unable to transmit frames at a rate that is required by a device connected to the fabric.

These cases can slow down communication between devices in your SAN. To resolve this type of issue, you have to see the SAN switch documentation or open a case with the vendor to investigate and identify what is causing the bottleneck and fix it.

If you cannot fix the issue with these actions, use the method that is described in 10.4, “Collecting data and isolating the problem” on page 369, collect the SAN switch debugging data, and then contact the vendor for assistance.

10.5.4 Solving optional back-end storage issues

IBM FlashSystem 9100 has useful tools for finding and analyzing optional back-end storage subsystem issues because it has a monitoring and logging mechanism.

Typical events for storage subsystem controllers include incorrect configuration, which results in a *1625 - Incorrect disk controller configuration* error code. Other issues related to the storage subsystem include failures pointing to the managed disk I/O (error code 1310), disk media (error code 1320), and error recovery procedure (error code 1370).

However, all messages do not have only one explicit reason for being issued. Therefore, you must check multiple areas for issues, not just the storage subsystem.

To determine the root cause of a problem, complete the following tasks:

1. Check the Recommended Actions panel by clicking **Monitoring** → **Events** as shown in 10.5.1, “Solving IBM FlashSystem 9100 events” on page 373.
2. Check the attached storage subsystem for misconfigurations or failures:
   a. Independent of the type of storage subsystem, first check whether the system has any unfixed errors. Use the service or maintenance features that are provided with the storage subsystem to fix these issues.

   b. Check if volume mapping is correct. The storage subsystem LUNs should be mapped to a host object with IBM FlashSystem 9100 ports. Also, observe the IBM FlashSystem 9100 restrictions for optional back-end storage subsystems, which can be found at this website:

   `IBM FlashSystem 9100`

   If you need to identify which of the externally attached MDisks has which corresponding LUN ID, run the IBM FlashSystem 9100 **`lsMDisk`** CLI command as shown in Example 10-13. This command also shows to which storage subsystem a specific MDisk belongs (the controller ID).

   Example 10-13   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
   600a0b80001742330000005946 9cf8450000000000000000000000000000000000
   2          MDisk2 online      managed     0         MDG-1
   70.9GB         0000000000000002 controller0
   600a0b800017443100000096469cf0e80000000000000000000000000000000000
   ```

3. Check the SAN environment for switch problems or zoning failures.

   Make sure the zones are properly configured, and the zoneset activated. The zones that allow communication between the storage subsystem and the IBM FlashSystem 9100 device should contain WWPNs of the storage subsystem and WWPNs of IBM FlashSystem 9100.

4. Collect all support data and contact IBM Support.

   Collect the support data for the involved SAN, IBM FlashSystem 9100, or optional external storage systems as described in 10.4, “Collecting data and isolating the problem” on page 369.

10.5.5 Common error recovery using IBM FlashSystem 9100 CLI

For SAN or optional back-end storage issues, you can use the IBM FlashSystem 9100 CLI to perform common error recovery steps. Although the maintenance procedures perform these steps, it is sometimes faster to run these commands directly through the CLI.

Run these commands any time that you have the following issues:

- You experience an optional back-end storage issue (for example, error code 1370 or error code 1630).
- You performed maintenance on the optional back-end storage subsystems.

Important: Run these commands when back-end storage is just configured, a zoning change occurs or any other type of changes related to the communication between IIBM FlashSystem 9100 and any external back-end storage subsystem happens. This ensures that IBM FlashSystem 9100 has recognized the changes.

Common error recovery involves the following IBM FlashSystem 9100 CLI commands:

- **`lscontroller and lsMDisk`**

  Provides current status of all optional external controllers and MDisks.
- **`detectMDisk`**

  Discovers the changes in both the IBM FlashSystem 9100 and any optional external back end storage.
- **`lscontroller <controller_id_or_name>`**

  Checks the controller that was causing the issue and verifies that all the WWPNs are listed as you expect. It also checks that the `path_counts` are distributed evenly across the WWPNs.
- **`lsMDisk`**

  Determines whether all MDisks are now online.

Note: When an issue is resolved using the CLI, check if the error has cleared from the **Monitoring** → **Events** panel. If not, make sure the error has been really fixed, and if so, manually mark the error as fixed.

10.6 Health status during upgrade and known error

It’s important to understand that during the software upgrade process, alerts indicating the system is not healthy are reported. This is a normal behavior because the IBM FlashSystem 9100 node canisters go offline during this process, so the system triggers these alerts.

Known error

While trying to upgrade an IBM FlashSystem 9100, you might get a message such as `Error in verifying the signature of the update package`.

This message does not mean that you have an issue in your system. Sometimes this happens because there is not enough space on the system to copy the file, or the package is incomplete or contains errors. In this case, open a PMR with IBM support and follow their instructions.

10.7 Call Home Web and Health Checker feature

Call Home Web is an IBM tool to view Call Home information on the web.

Call Home is a functionality present in several IBM systems, including IBM FlashSystem 9100, which allow them to automatically report problems and status to IBM.

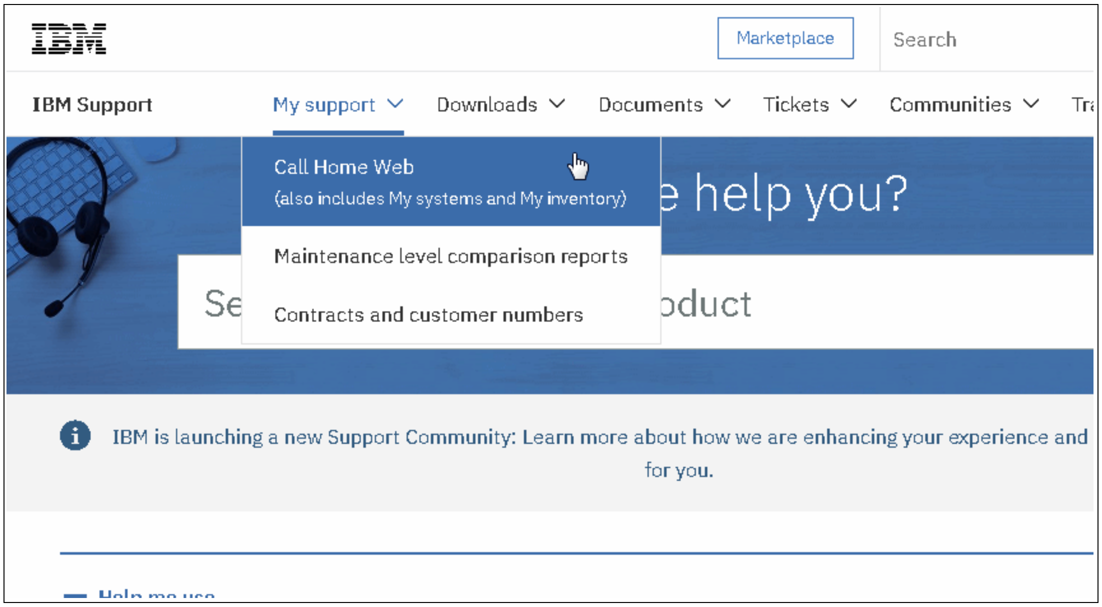
Call Home Web provides the following information about IBM systems:

- Automated tickets
- Warranty and contract status
- Health check alerts and recommendations
- System connectivity heartbeat
- Recommended software levels
- Inventory
- Security bulletins

To access the Call Home Web, go to IBM support website at:

IBM Support Homepage

In the IBM support website, Call Home Web is available at **My support** → **Call Home Web** as shown in Figure 10-10.

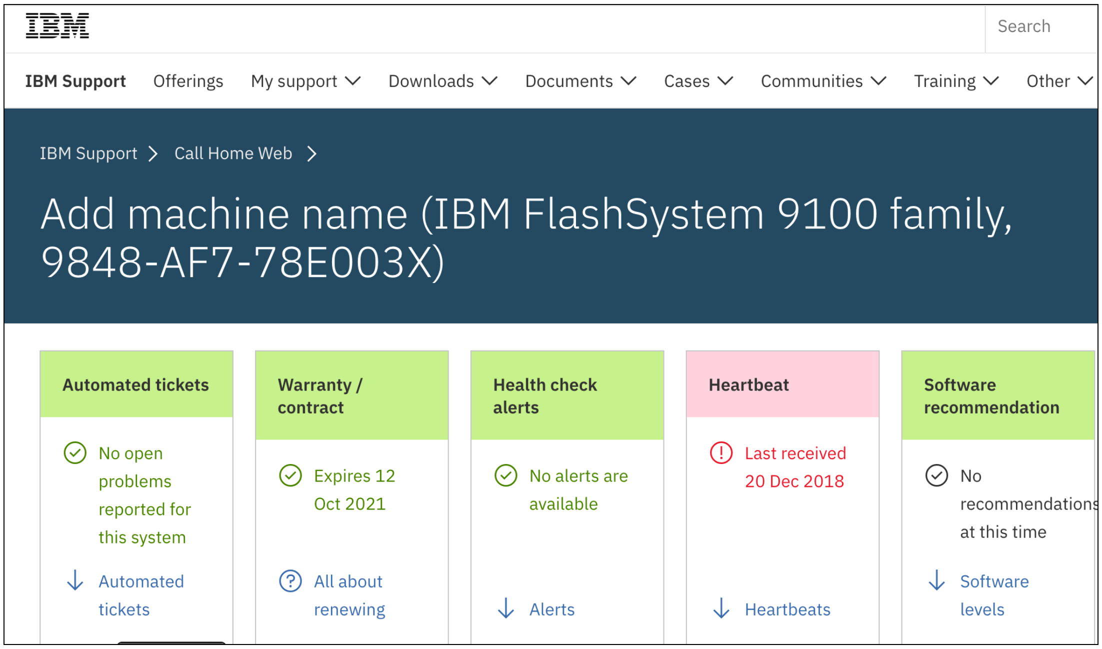


IBM is launching a new Support Community: Learn more about how we are enhancing your experience and for you.

Figure 10-10 Call Home Web

To allow Call Home Web analyze data of IBM FlashSystem 9100 systems and provide useful information about them, the devices need to be added to the tool. The machine type, model and serial number are required to register the product in Call Home Web. Also, it is required that IBM FlashSystem 9100 have call home and inventory notification enabled and operational.

Figure 10-11 on page 383 shows the Call Home Web details panel of IBM FlashSystem 9100 system.



Add machine name (IBM FlashSystem 9100 family, 9848-AF7-78E003X)

Automated tickets	Warranty / contract	Health check alerts	Heartbeat	Software recommendation
No open problems reported for this system	Expires 12 Oct 2021	No alerts are available	Last received 20 Dec 2018	No recommendations at this time
Automated tickets	All about renewing	Alerts	Heartbeats	Software levels

Figure 10-11 Call Home Web details panel

For a video guide on how to setup and use IBM Call Home Web see:

IBM Call Home Web Set-up Video Guide

10.7.1 Health Checker

A new feature of Call Home Web is the Health Checker, a tool that runs in the IBM Cloud.

It analyzes call home and inventory data of systems registered in Call Home Web and validates their configuration. Then, it displays alerts and provide recommendations in the Call Home Web tool.

> **Note:** Call Home Web should be used because it provides useful information about your systems, and with the Health Checker feature it helps you to monitor the system, and proactively provides alerts and creates recommendations related to them.

Some of the function of the IBM Call Home Web and Health Checker has been ported to IBM Storage Insights which is explained in detail in 10.8, “IBM Storage Insights”.

10.8 IBM Storage Insights

IBM Storage Insights is an integral part of the monitoring and ensuring continued availability of the IBM FlashSystem 9100.

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’ll be able to make better decisions by seeing trends in performance and capacity. Storage health information enables you to focus on areas needing attention.

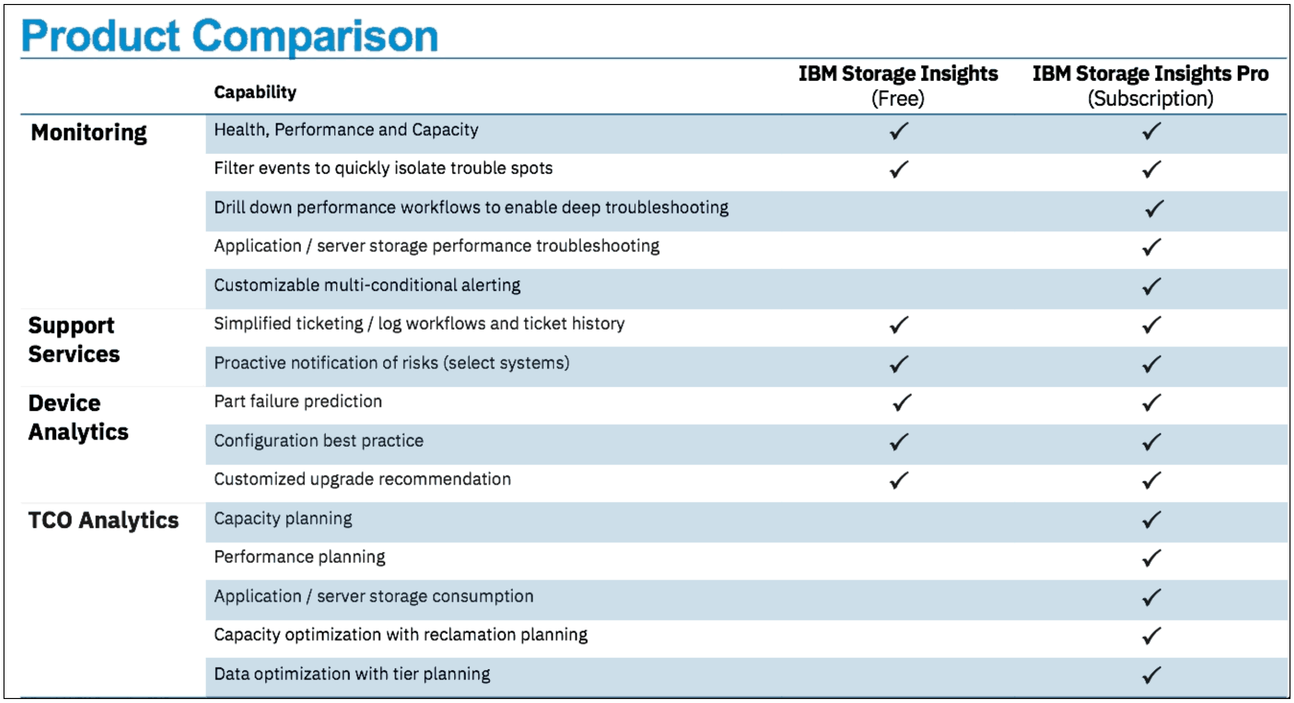
In addition, when IBM support is needed, Storage Insights simplifies uploading logs, speeds resolution with online configuration data, and provides an overview of open tickets all in one place.

The following features are some of those included:

- A unified view of IBM systems:
  - Provides a single pane to see all of your system's characteristics
  - See all of your IBM storage inventory
  - Provides a live event feed so you know, up to the second, what is going on with your storage and enables you to take action 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
  - Monitor the configuration to see if it meets the best practices
  - System resource management: determine if 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 support to view support tickets, open and close them, 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 can save as much as 50% of the time to resolve the case

In addition to the free Storage Insights, there is also the option of Storage Insights Pro, 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 together with EMC VNX and VMAX.

Figure 10-12 on page 385 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 10-12 Storage Insights versus Storage Insights Pro comparison

Storage Insights provides a very lightweight data collector that is deployed on a customer supplied server. This can be either 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.

What metadata is collected

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 actual application data that is stored on the storage systems can't be accessed by the data collector

Who can access 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

10.8.1 SI Customer Dashboard

Figure 10-13 shows a view of the Storage Insights main dashboard and the systems that it is monitoring.

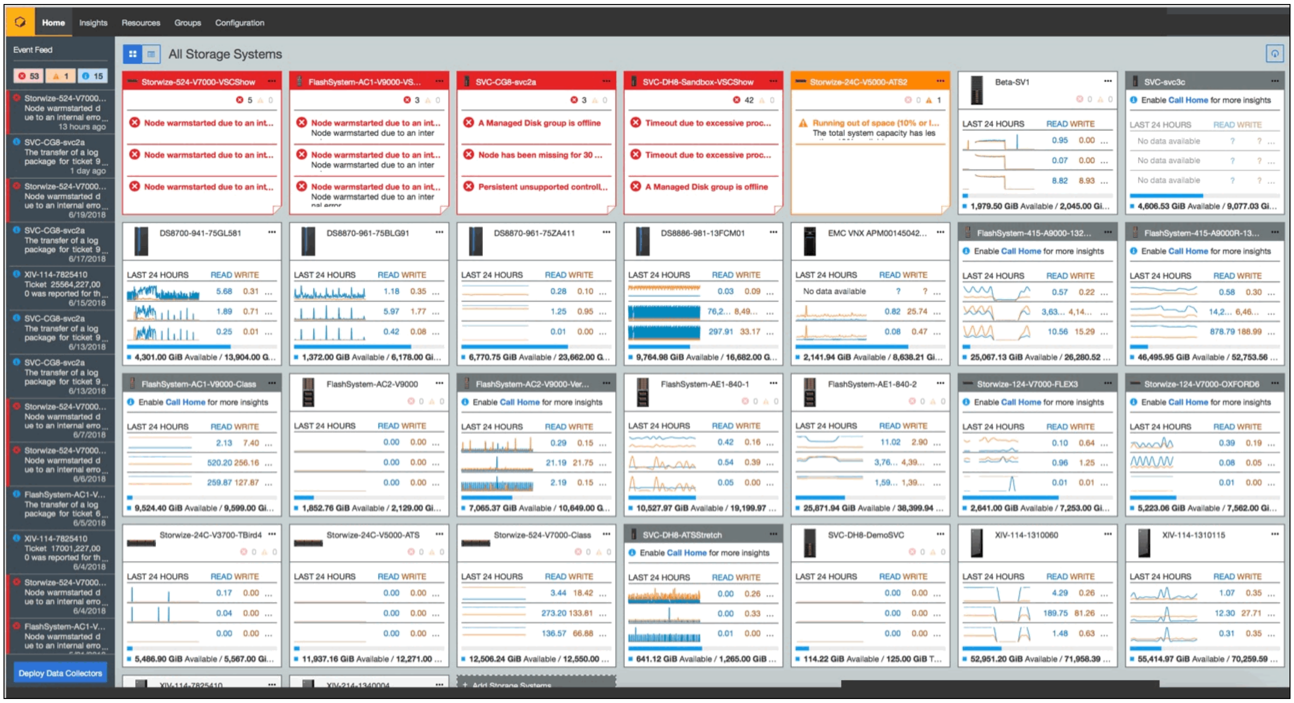


Figure 10-13 Storage Insights Main Dashboard

10.8.2 Customized dashboards to monitor your storage

With the latest release of IBM Storage Insights (SI) you are able to customize the dashboard to only show a subset of the systems monitored. This is particularly useful for customers that might be Cloud Service Providers (CSP) and only want a particular end user to see those machines that are paying for.

For further details on setting up the customized dashboard, please see *Customizing the Dashboard* in the IBM SI Knowledge Center.

IBM SI Customizing the Dashboard

10.8.3 Creating support tickets

IBM SI has the ability to create support tickets for one of the systems it reports about, from the Dashboard GUI. To do this go to the SI main dashboard and then chose the system you want to raise the ticket for. From this screen select **Action** → **Create/Update Ticket**.

Figure 10-14 shows how to create or update a support ticket from the SI dashboard.

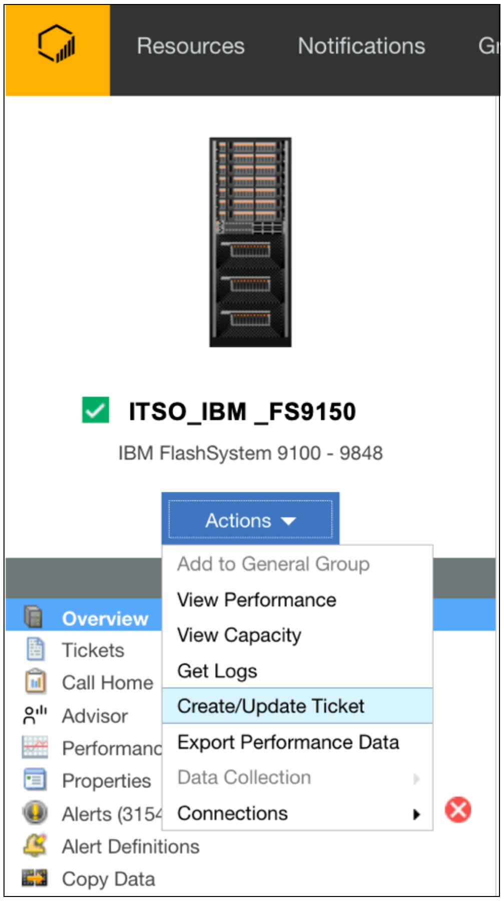
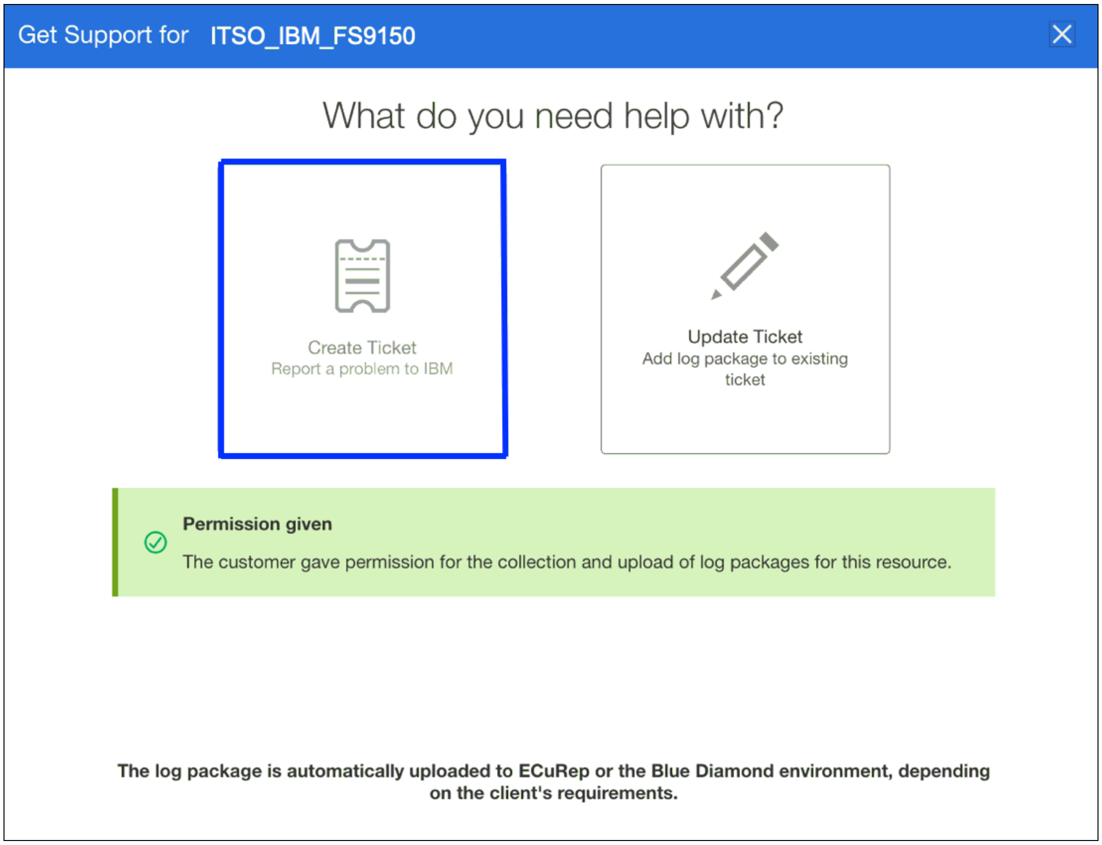


Figure 10-14 SI Create / Update a support Ticket

Figure 10-15 shows you the panel where you can either create a new ticket or update a previously created one.



Get Support for ITSO_IBM_FS9150

What do you need help with?

Permission given

The customer gave permission for the collection and upload of log packages for this resource.

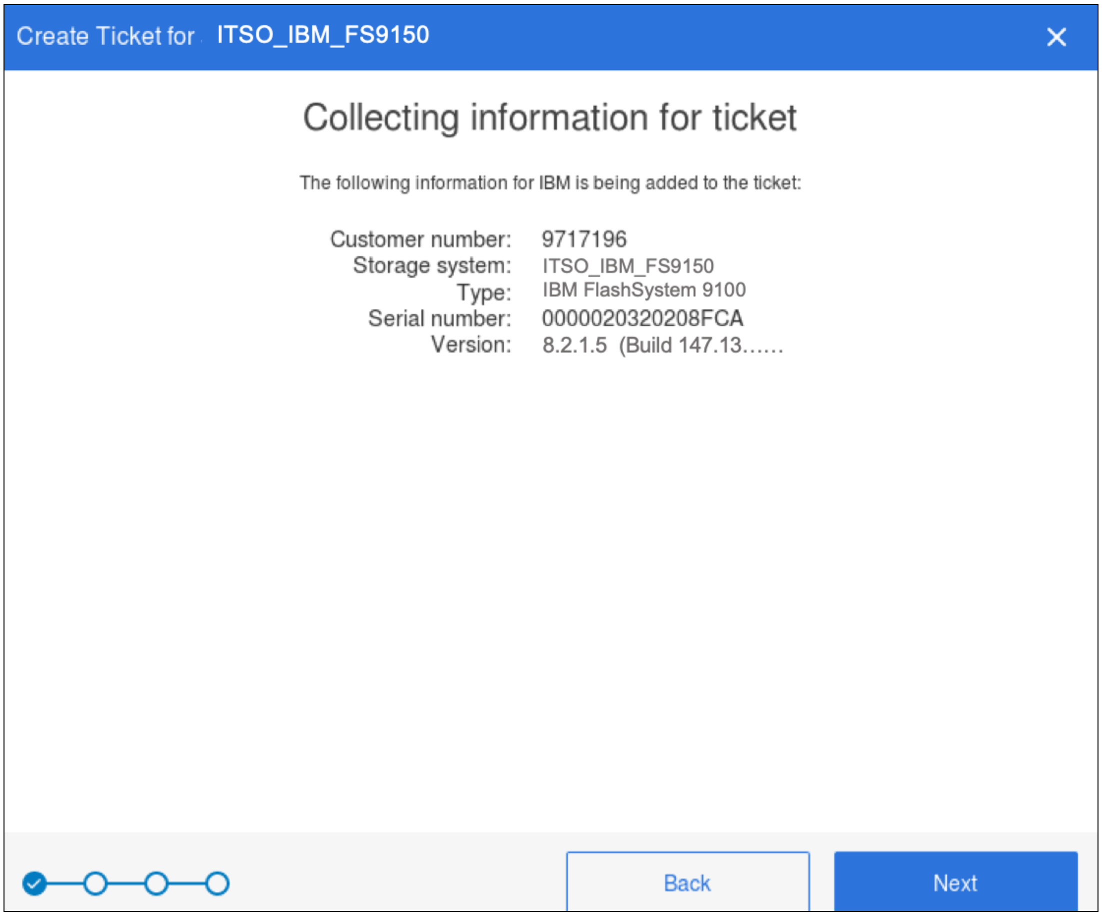
The log package is automatically uploaded to ECuRep or the Blue Diamond environment, depending on the client's requirements.

Figure 10-15 Create ticket

Note: The *Permission given* information box shown above, is an option the customer needs to enable on in the IBM FlashSystem 9100.GUI. See the 10.2, "Remote Support Assistance" on page 360 to enable this function.

Select the **Create Ticket** option and you will presented with the following screens to complete with the machine details, problem description and the option to upload logs.

Figure 10-16 shows the ticket data collection done by the SI application.



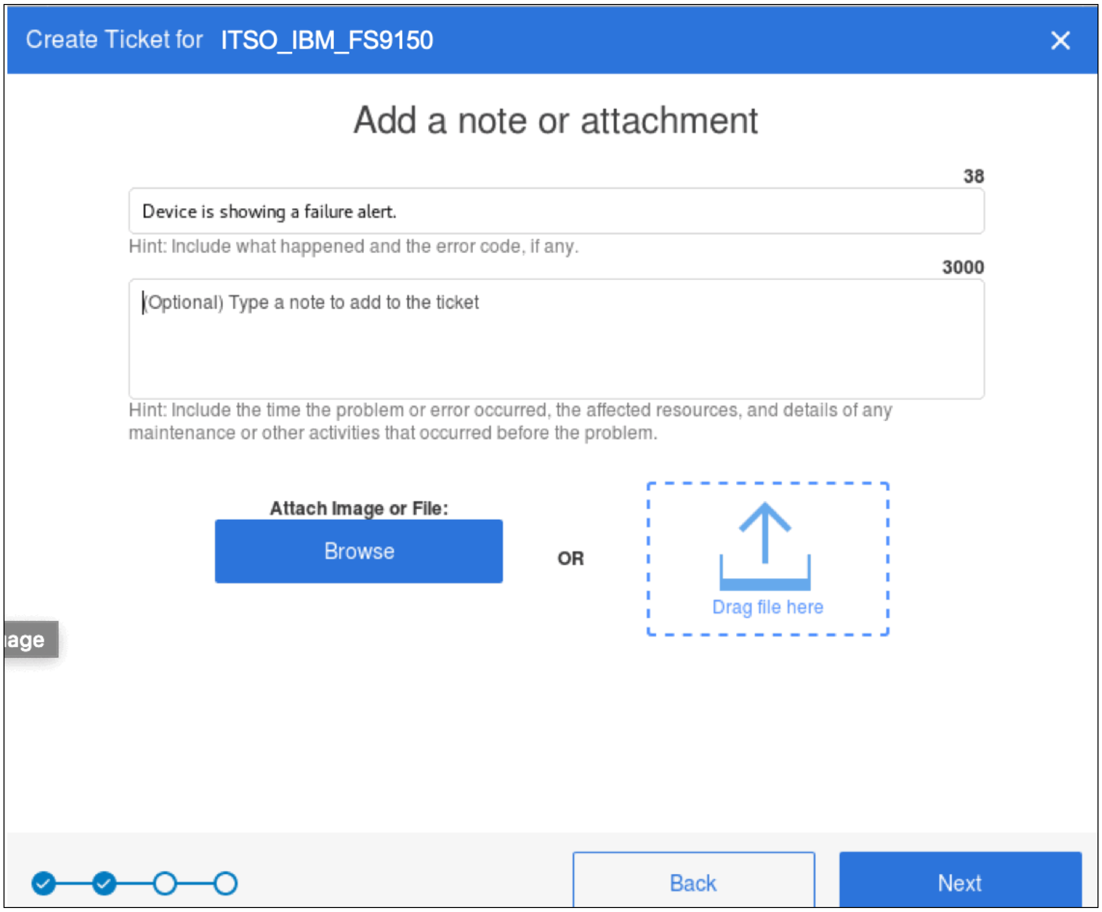
Collecting information for ticket

The following information for IBM is being added to the ticket:

Customer number:	9717196
Storage system:	ITSO_IBM_FS9150
Type:	IBM FlashSystem 9100
Serial number:	0000020320208FCA
Version:	8.2.1.5 (Build 147.13......

Figure 10-16 Collecting ticket information

Figure 10-17 allows you to add a problem description and also attach additional files, such as error logs or screen grabs of error messages, and so on.



Create Ticket for ITSO_IBM_FS9150

Add a note or attachment

38

Device is showing a failure alert.

Hint: Include what happened and the error code, if any.

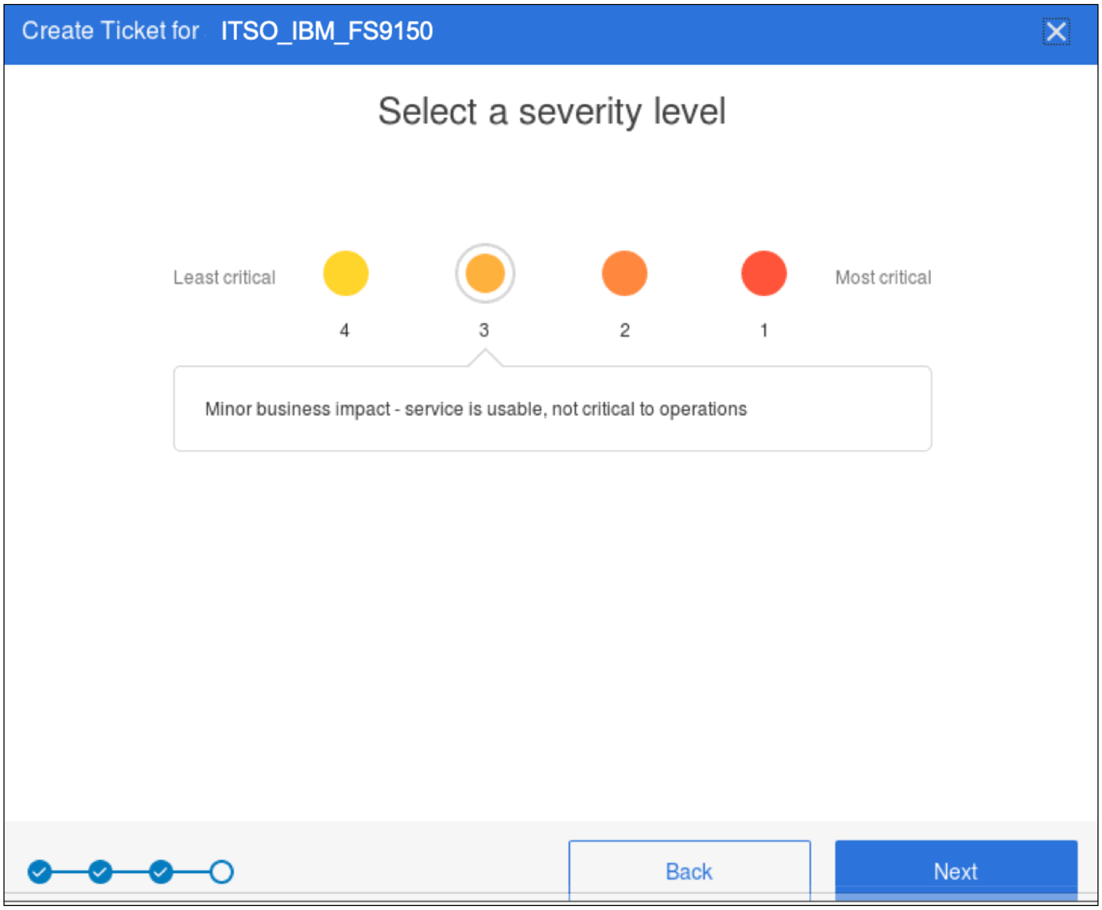
3000

(Optional) Type a note to add to the ticket

Hint: Include the time the problem or error occurred, the affected resources, and details of any maintenance or other activities that occurred before the problem.

Figure 10-17 Adding problem description and any additonal information

Figure 10-18 prompts you to set a severity level for the ticket, ranging from a severity 1 for a system down or extreme business impact, through to severity 4, which is for non critical issues and so on.



Create Ticket for ITSO_IBM_FS9150

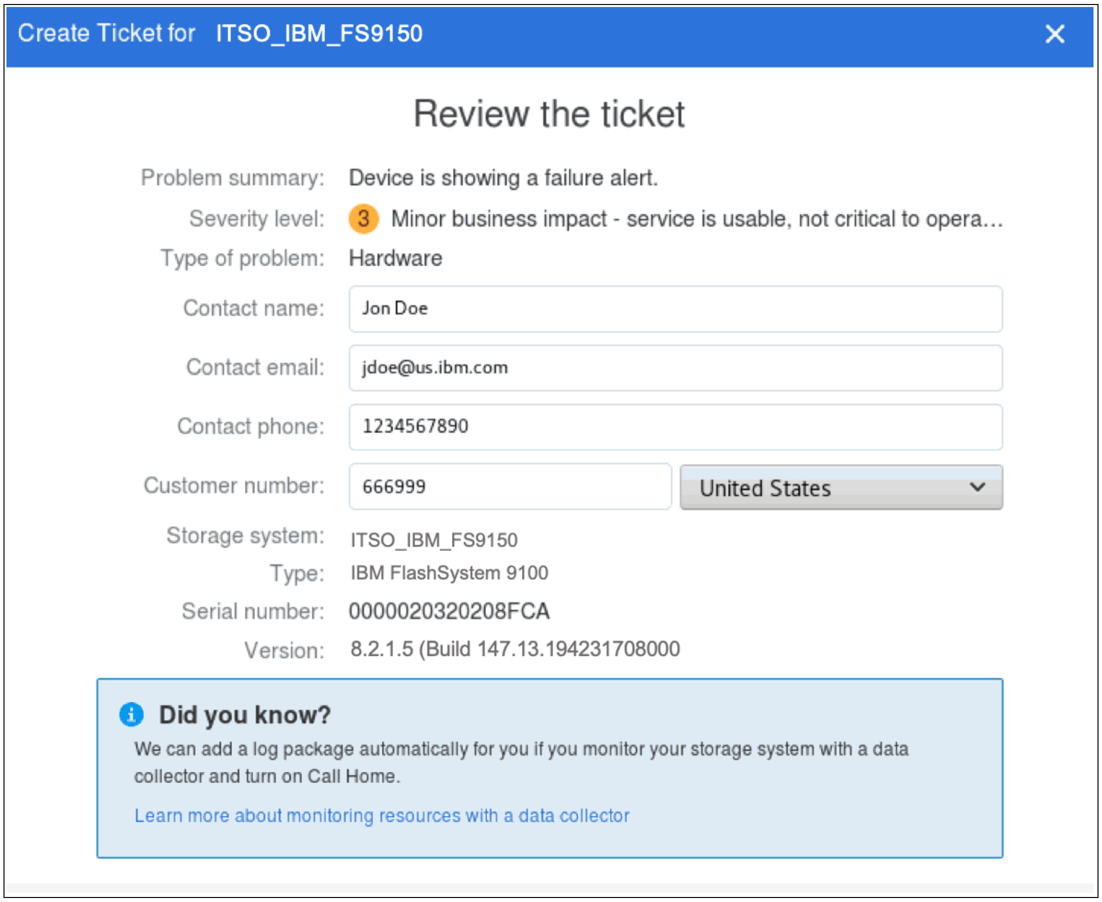
Select a severity level

Least critical ○ 4 ● 3 ○ 2 ○ 1 Most critical

Minor business impact - service is usable, not critical to operations

Figure 10-18 Set severity level

Figure 10-19 gives you a summary of the data that will be used to create the ticket.



Create Ticket for ITSO_IBM_FS9150

Review the ticket

Problem summary: Device is showing a failure alert.

Severity level: 3 Minor business impact - service is usable, not critical to opera…

Type of problem: Hardware

Contact name: Jon Doe

Contact email: jdoe@us.ibm.com

Contact phone: 1234567890

Customer number: 666999 United States

Storage system: ITSO_IBM_FS9150

Type: IBM FlashSystem 9100

Serial number: 0000020320208FCA

Version: 8.2.1.5 (Build 147.13.194231708000

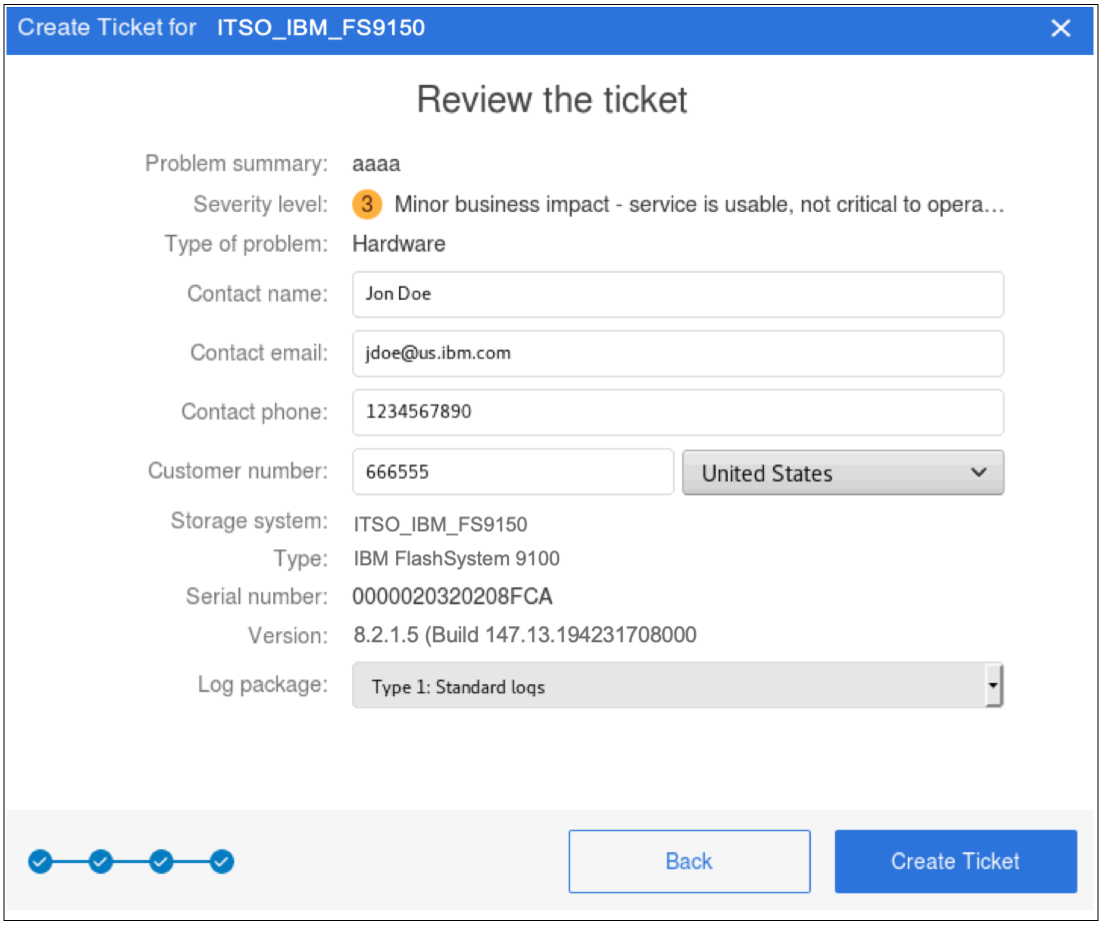
Did you know?

We can add a log package automatically for you if you monitor your storage system with a data collector and turn on Call Home.

Learn more about monitoring resources with a data collector

Figure 10-19 Review the ticket information

Figure 10-20 shows the final summary panel, and the option to add logs to the ticket. When completed, click the **Create Ticket** button to create the support ticket and send it to IBM. The ticket number is created by the IBM Support system and sent back to your SI instance.



Create Ticket for ITSO_IBM_FS9150

Review the ticket

Problem summary: aaaa

Severity level: 3 Minor business impact - service is usable, not critical to opera…

Type of problem: Hardware

Contact name: Jon Doe

Contact email: jdoe@us.ibm.com

Contact phone: 1234567890

Customer number: 666555 United States

Storage system: ITSO_IBM_FS9150

Type: IBM FlashSystem 9100

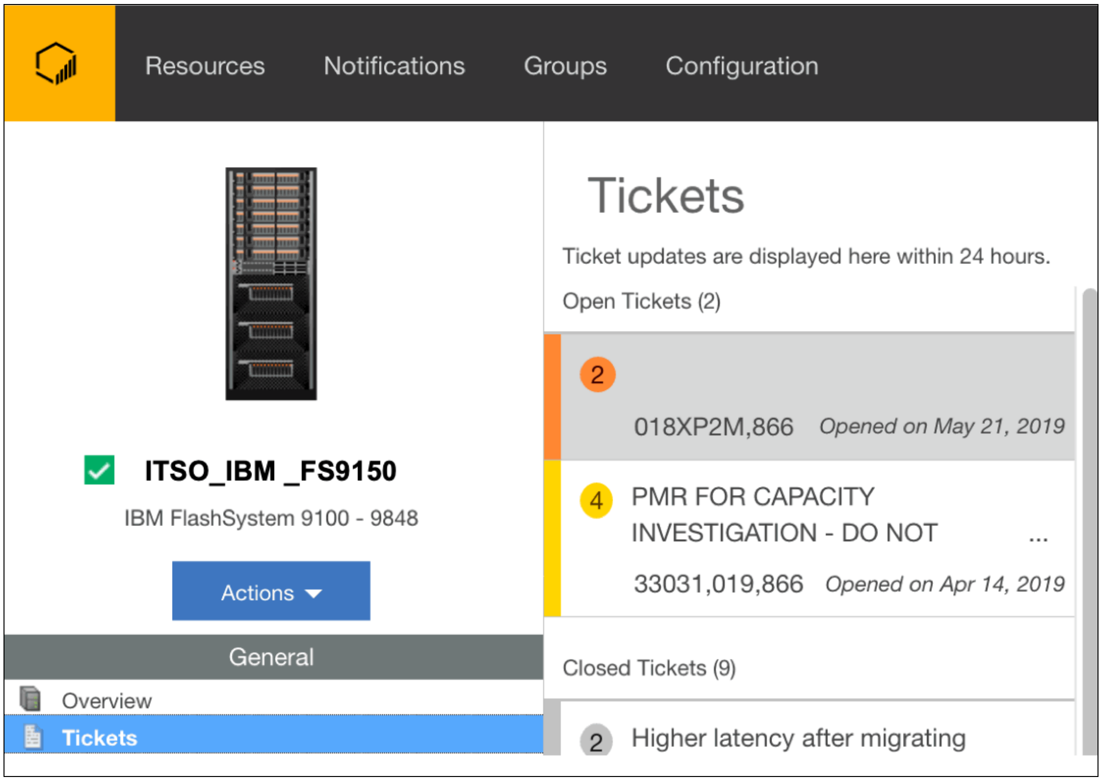
Serial number: 0000020320208FCA

Version: 8.2.1.5 (Build 147.13.194231708000

Log package: Type 1: Standard logs

Figure 10-20 Final summary before ticket creation

Figure 10-21 shows how to view the summary of the open and closed ticket numbers for the system selected, using the **Action** menu option.



Tickets

Ticket updates are displayed here within 24 hours.

Open Tickets (2)

2 018XP2M,866 *Opened on May 21, 2019*

4 PMR FOR CAPACITY INVESTIGATION - DO NOT ... 33031,019,866 *Opened on Apr 14, 2019*

Closed Tickets (9)

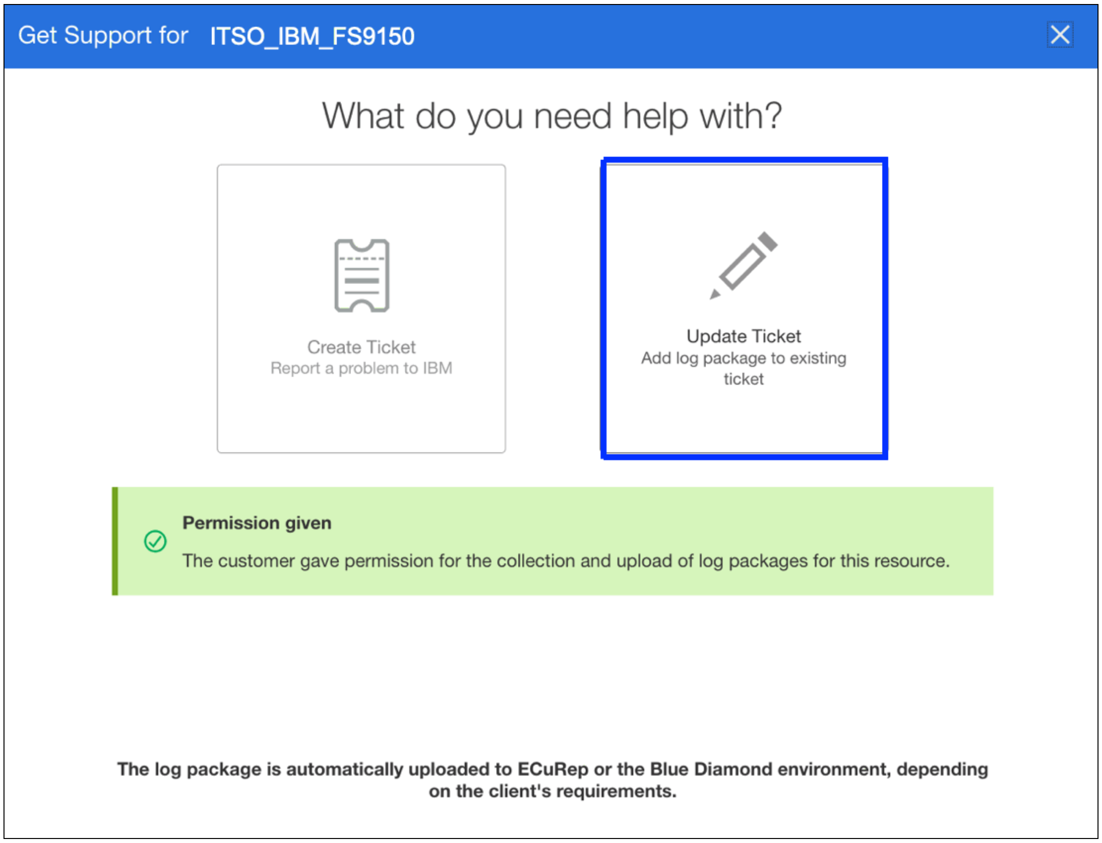
2 Higher latency after migrating

Figure 10-21 Ticket summary

10.8.4 Updating support tickets

IBM Storage Insights also has the ability to update support tickets, for any of the systems it reports about, also from the Dashboard GUI. To do this go to the SI dashboard and then chose the system you want to update the ticket for. From this screen select **Action** → **Create/Update Ticket**.

Figure 10-22 shows the initial menu to update an existing ticket. Select this option as shown here.


Get Support for ITSO_IBM_FS9150
What do you need help with?
Create Ticket
Report a problem to IBM
Update Ticket
Add log package to existing ticket

Permission given

The customer gave permission for the collection and upload of log packages for this resource.

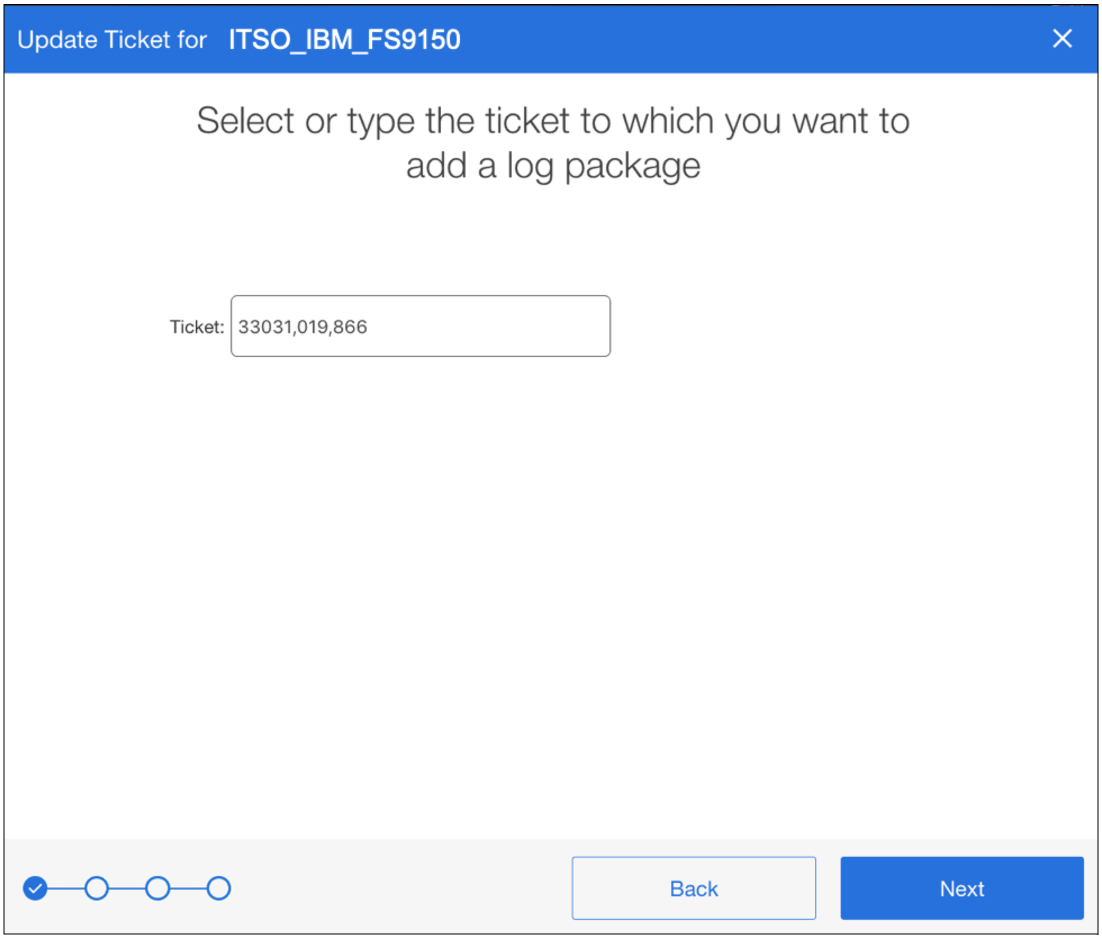
The log package is automatically uploaded to ECuRep or the Blue Diamond environment, depending on the client's requirements.

Figure 10-22 SI Update Ticket

Figure 10-23 on page 396 shows the next screen where you have to enter the PMR number then press Next, This PMR input is in the format **XXXXX,YYY,ZZZ**, where:

- **XXXXX** is the PMR record number
- **YYY** is the IBM Branch office number
- **ZZZ** in the IBM country number

These details would have been either supplied when you created the ticket or by IBM support in the event of the PMR being created by an problem call home event (assuming that call home is enabled).


Update Ticket for ITSO_IBM_FS9150

Select or type the ticket to which you want to add a log package

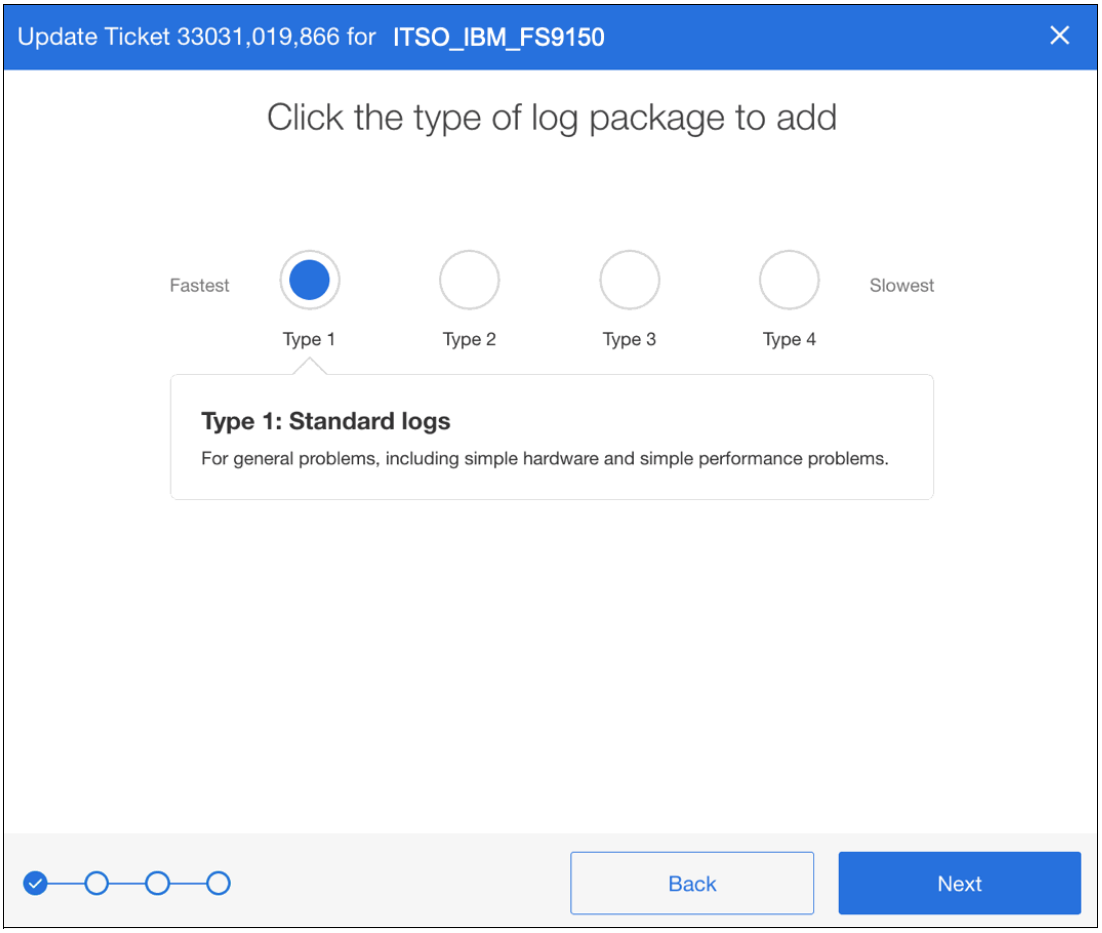
Ticket: 33031,019,866

Figure 10-23 Entering the PMR ticket number

Pressing **Next** displays the screen where you need to choose the log type to upload. Figure 10-24 on page 397 shows the log selection screen and the options.

The options are as follows:

- **Type 1 - Standard logs**
  - For general problems, including simple hardware and simple performance problems
- **Type 2 - Standard logs and the most recent state save log**
- **Type 3 - Standard logs and the most recent state save log from each node**
  - For 1195 and 1196 node errors and 2030 software restart errors
- **Type 4 - Standard logs and new state save logs**
  - For complex performance problems, and problems with interoperability of hosts or storage systems, compressed volumes, and remote copy operations including 1920 errors


Update Ticket 33031,019,866 for ITSO_IBM_FS9150

Click the type of log package to add

Type 1: Standard logs

For general problems, including simple hardware and simple performance problems.

Figure 10-24 Log type selection

If you are unsure which log type to upload, then please ask IBM Support for guidance. The most common type to use is type 1, so this is the default. The other types are more detailed logs and for issues in order of complexity.

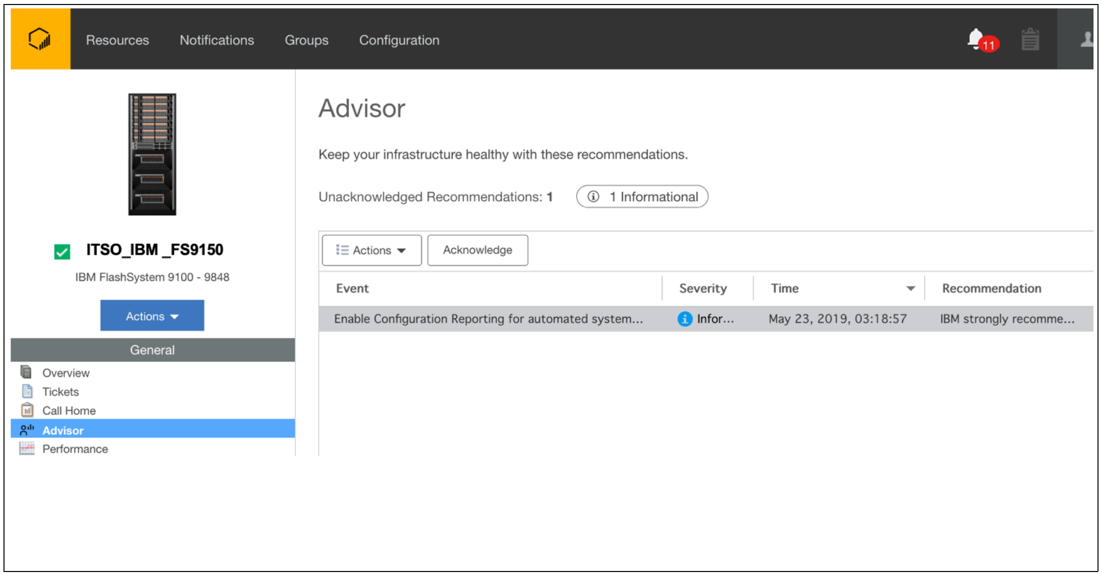
After selecting the type of logs and pressing Next, the log collection and upload will start. When completed you will be presented with the log completion screen.

10.8.5 SI 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 details of a recommendation or to acknowledge it, double-click the recommendation.

Figure 10-25 shows the initial SI advisor menu.



Advisor

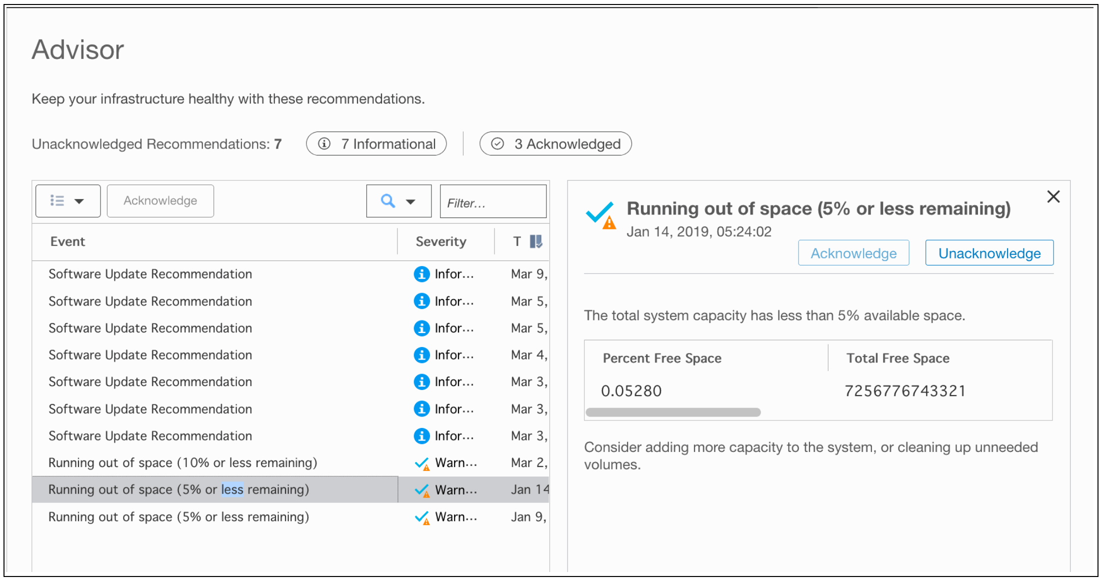
Keep your infrastructure healthy with these recommendations.

Unacknowledged Recommendations: **1** ⓘ 1 Informational

Event	Severity	Time	Recommendation
Enable Configuration Reporting for automated system...	Infor...	May 23, 2019, 03:18:57	IBM strongly recomme...

Figure 10-25 SI Advisor menu

Figure 10-26 shows an example of the detailed SI Advisor recommendations.



Advisor

Keep your infrastructure healthy with these recommendations.

Unacknowledged Recommendations: **7** ⓘ 7 Informational ✓ 3 Acknowledged

Event	Severity	T
Software Update Recommendation	Infor...	Mar 9,
Software Update Recommendation	Infor...	Mar 5,
Software Update Recommendation	Infor...	Mar 5,
Software Update Recommendation	Infor...	Mar 4,
Software Update Recommendation	Infor...	Mar 3,
Software Update Recommendation	Infor...	Mar 3,
Software Update Recommendation	Infor...	Mar 3,
Running out of space (10% or less remaining)	Warn...	Mar 2,
Running out of space (5% or less remaining)	Warn...	Jan 14
Running out of space (5% or less remaining)	Warn...	Jan 9,

Running out of space (5% or less remaining)
Jan 14, 2019, 05:24:02

Acknowledge Unacknowledge

The total system capacity has less than 5% available space.

Percent Free Space	Total Free Space
0.05280	7256776743321

Consider adding more capacity to the system, or cleaning up unneeded volumes.

Figure 10-26 Advisor detailed summary of recommendations

The image shows the details of a “Running out of space” recommendation on the Advisor page. In this scenario, the user clicked the **Warning** tag to focus only on recommendations that have a severity of “Warning.” For more information about setting and configuring the Advisor options, see this IBM SI Knowledge Center link:

IBM Storage Insights Advisor

10.9 Out of Space monitoring and recovery

Space efficient storage arrays and controllers, such as the IBM FlashSystem 9100 with FCM's, can suffer from running out of space. This can happen for a variety of reasons including:

- Poorer than expected capacity savings on the IBM FlashSystem 9100 system
- Under-sizing the capacity need for the solution

Also any geothermal controller that is presenting thin provisioned LUNs to the IBM FlashSystem 9100 cluster as well as other hosts could result in space being consumed faster than expected.

IBM FlashSystem 9100 makes use of two types of data reduction techniques

- IBM FlashSystem 9100 using the FCM NVMe drives have built-in hardware compression
- IBM FlashSystem 9100 using industry standard NVMe drives relies on the Spectrum Virtuatlize software and DRP pools to lever data reduction

Users are encouraged to pay attention to any GUI / notifications and employ best practices for managing physical space.

10.9.1 Monitoring

To avoid running out of space on the system, the usable capacity should be carefully monitored on the GUI of the IBM FlashSystem 9100. The IBM FlashSystem 9100 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 9100, which typically happen 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 9100. Call Home monitors the physical free space on the system and will automatically open a service call for systems that reach 99% of their usable capacity.

IBM Storage Insights also has the ability to monitor and report on any potential out of space conditions and the new Advisor function will warn when the IBM FlashSystem 9100 is getting near to full capacity. See the IBM Storage Insights section 10.8.5, “SI Advisor” on page 397.

When IBM FlashSystem 9100 reaches that last space condition, out of space, the device will drop into a read only state. Assessment of data compression ratio and the re-planned capacity estimation should be done to determine how much actual outstanding storage demand might exist. This additional capacity will likely need to 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 9100 in this scenario vary by the capabilities of the system, and any optional external back end controllers as well, as system configuration and pre-planned capacity overhead needs.

Generally speaking, the following options are available:

- Add additional capacity to the IBM FlashSystem 9100. Customers should have a plan that allows them to add additional capacity to the system when needed.
- Reserve a set of space in the IBM FlashSystem 9100 that makes it “seem” fuller than it really is, and that you can free up in an emergency situation

- IBM FlashSystem 9100 has the ability to create a volume which isn’t compressed, de-duped or thin provisioned (so a fully allocated volume). Simply create some of these volumes to reserve an amount of physical space, you probably want to name them something like “emergency buffer space”. Then if you are reaching the limits for physical capacity you can simply delete one or more of these volumes to give yourself a temporary reprieve.

Warning: Running completely out of space can be a very severe situation. Recovery can be extremely complicated and time-consuming. For this reason, it is imperative that proper planning and monitoring be done to avoid reaching this condition.

10.9.2 Recovery

The following sections describe the process for recovering from an out of space condition.

Reclaiming and unlocking

After you’ve assessed and accounted for storage capacity, the first step is to contact IBM Support who can aid in unlocking the read only mode and restoring the operation. The reclamation task can take a long time to run, and larger flash arrays will take longer to recover than smaller ones.

Freeing up space

There are several ways to reduce the amount of consumed space once the IBM FlashSystem 9100 has been unlocked by IBM support.

To recover from Out of Space conditions on Standard Pools, these are the steps for the user:

1. Add more storage to the system if possible.
2. Migrate extents from the write protected array to other non-write protected MDisks with enough extents. This could be an external back end storage array.
3. Migrate volumes with extents on the write protected array to another pool.
4. Deleting unrequired volumes to free space.
5. Bring the volumes in the pool back online using a Directed Maintenance Procedure.

Deleting unneeded volumes

If possible, within the pool (managed disk group) on the IBM FlashSystem 9100, delete any unnecessary volumes. The IBM FlashSystem 9100 v8.2 supports SCSI unmap so deleting volumes will have space reclamation benefits using this method.

Migrating volumes

If possible, moving volumes from the IBM FlashSystem 9100 pool to another external pool can free up space in the IBM FlashSystem 9100 pool to allow for space reclamation. As this volume moves into the new pool, its previously occupied flash extends will be freed up (via SCSI unmap), which then goes to provide more free space to the IBM FlashSystem 9100 enclosure to be configured to a proper provisioning to support the compression ratio.

Further information on types of recovery can be found in the IBM Support Technote here:

Out of Space Recovery

A

Business continuity

Business continuity (BC) and continuous application availability are among the most important requirements for many organizations. Advances in virtualization, storage, and networking have made enhanced business continuity possible. Information technology solutions can now manage both planned and unplanned outages, and provide the flexibility and cost efficiencies that are available from cloud-computing models. IBM Flash System 9100 implements Business Continuity capabilities through the HyperSwap technology.

This Appendix briefly describes the HyperSwap solution for IBM FS9100 providing some basic guidelines for implementation.

Important, IBM FlashSystem 9200: On 11th February 2020 IBM announced the arrival of the IBM FlashSystem 9200 to the family. This book was written specifically for IBM FlashSystem 9100, however most of the general principles will apply to the IBM FlashSystem 9200.

If you are in any doubt as to their applicability to the FlashSystem 9200 then you should work with your local IBM representative.

This book will be updated to include FlashSystem 9200 in due course.

The Flashsystem 9200 product guide is available at:

IBM FlashSystem 9200 Product Guide

Business continuity with HyperSwap

The *HyperSwap* high availability feature in IBM FS9100 enables business continuity during a hardware failure, power failure, connectivity failure, or disasters, such as fire or flooding. The HyperSwap feature is available on the IBM Spectrum Virtualize based products.

The HyperSwap feature provides highly available volumes accessible through two sites at up to 300 km apart. A fully independent copy of the data is maintained at each site. When data is written by hosts at either site, both copies are synchronously updated before the write operation is completed. The HyperSwap feature automatically optimizes itself to minimize data that is transmitted between sites, and to minimize host read and write latency.

HyperSwap has the following key features:

- Works with IBM Spectrum Virtualize and IBM Storwize V7000, V5000, and FS9100.
- Uses intra-cluster synchronous remote copy (named Active-Active Metro Mirror with change volumes) capabilities along with existing change volume and access I/O group technologies.
- Makes a host's volumes accessible across two FS9100 I/O groups in a clustered system by using the Active-Active Metro Mirror relationship. The volumes appear as a single volume to the host.
- Works with the standard multipathing drivers that are available on various host types, with no additional host support required to access the highly available volume.

The IBM FS9100 HyperSwap configuration requires that at least one control enclosure is implemented in each location. Therefore, a minimum of two control enclosures for each IBM FS9100 cluster are needed to implement the HyperSwap. Configuration with three or four control enclosures is also supported for the HyperSwap.

The typical IBM FS9100 implementation is depicted in Figure A-1.

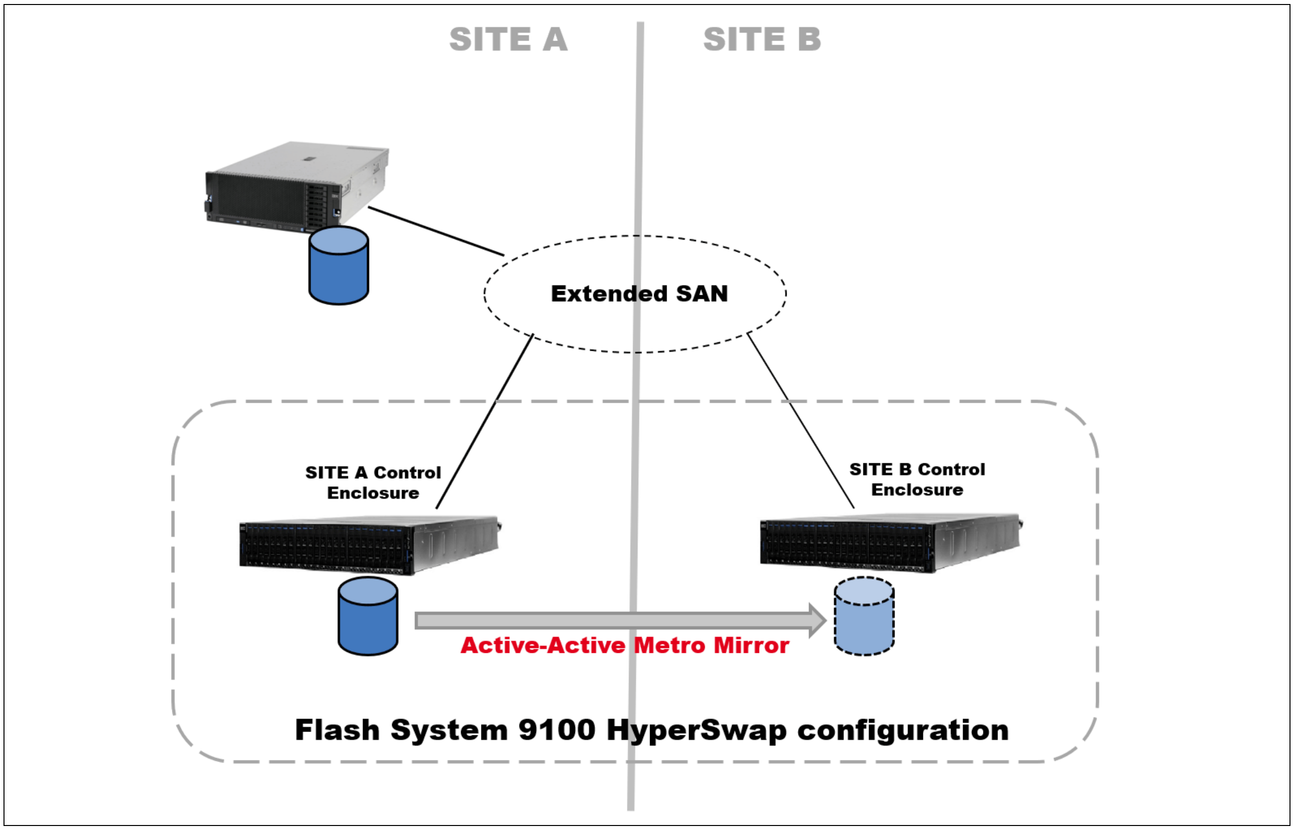


Figure A-1 Typical HyperSwap configuration with Flash System 9100

With a copy of the data that is stored at each location, HyperSwap configurations can handle different failure scenarios.

Figure A-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.

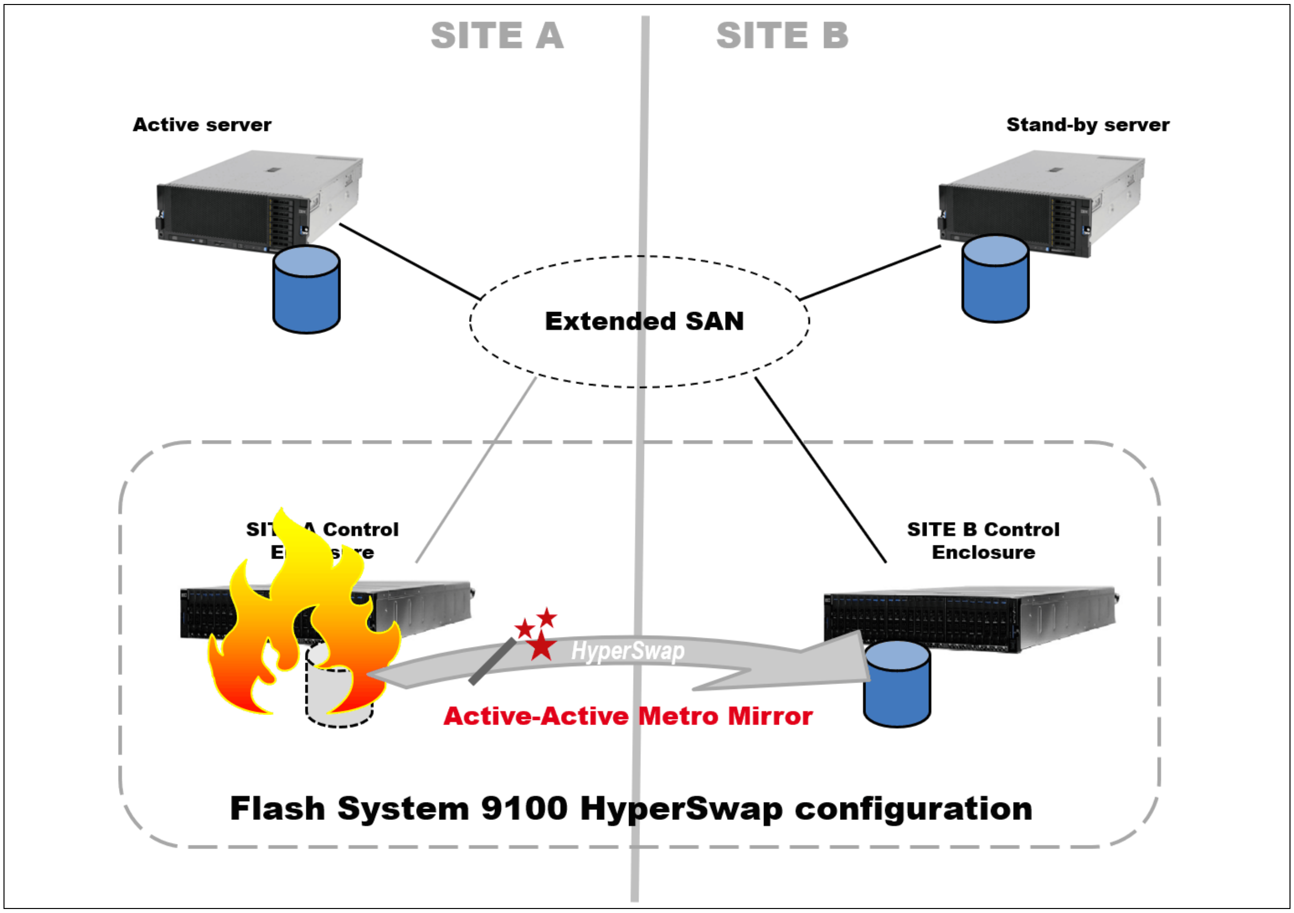


Figure A-2 Flash System 9100 HyperSwap in a storage failure scenario

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.

This scenario is depicted in Figure A-3. The active-active Metro Mirror feature 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.

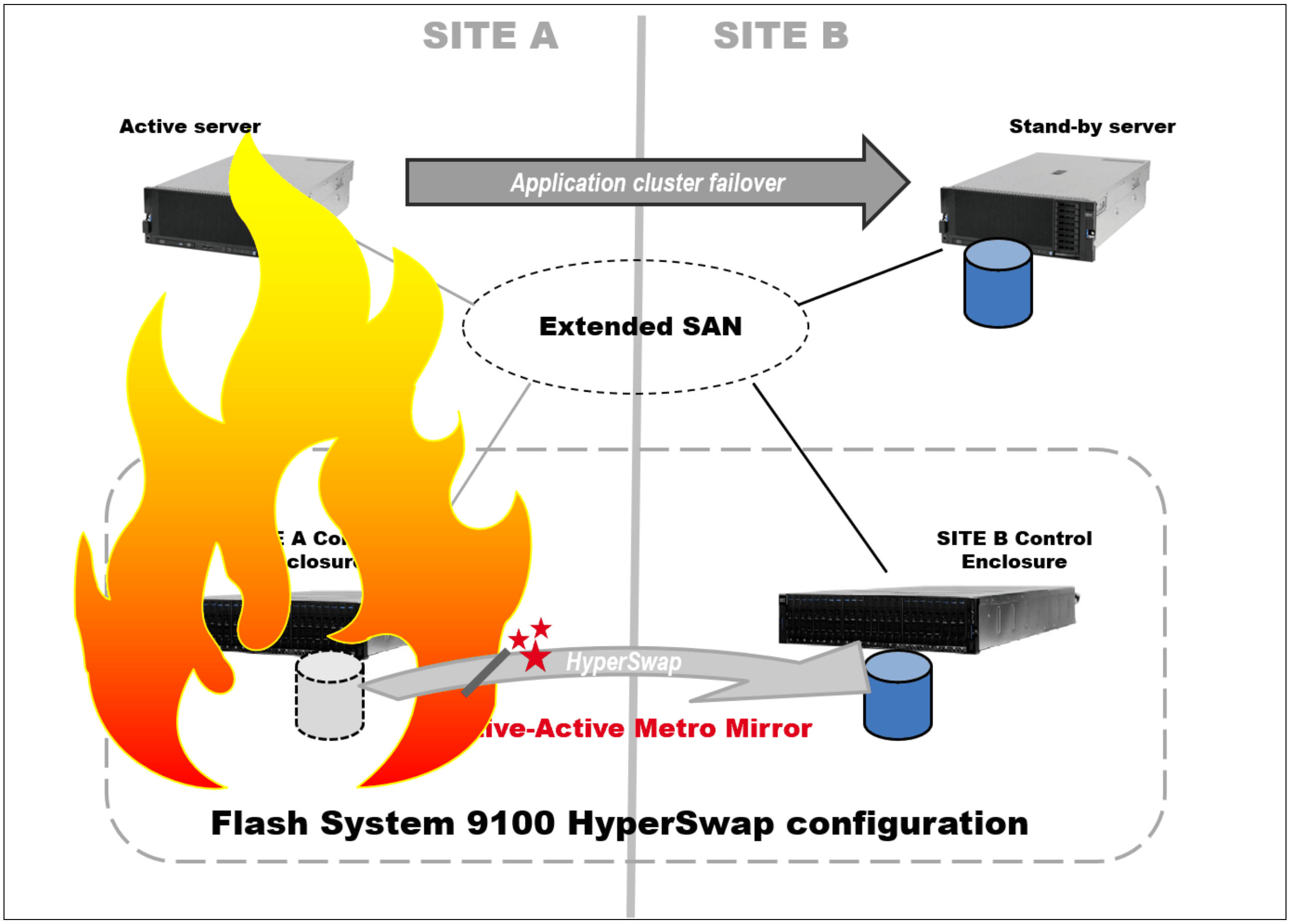


Figure A-3 Flash System 9100 HyperSwap in a site failure scenario

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, the HyperSwap *DR feature* allows us to manage rolling disaster scenarios effectively.

Quorum considerations

As with any clustering solution, avoiding a “split-brain” situation (where the control enclosures are no longer able to communicate with each other) requires a tie-break. The Flash System 9100 HyperSwap configuration is no exception. IBM Flash System 9100 software uses a tie-break mechanism that is facilitated through the implementation of quorum disks. It uses three quorum devices that are attached to the cluster.

The tie-break mechanism for the HyperSwap solution requires each location to provide one of quorum disks by using either internal or external storage capacity, and an additional third quorum device in a third location (*quorum site*) that acts as a tie-breaker (*Active* quorum). The Active quorum can be an external storage provided MDisk or a Java application, called *IP Quorum*, that acts as a witness.

To use an IP-based quorum application as the quorum device for the third site, no Fibre Channel connectivity is used. Java applications are run on hosts at the third site. However, there are strict requirements on the IP network, with using IP Quorum applications.

For stable quorum resolutions, an IP network must provide the following requirements:

- Connectivity from the hosts to the service IP addresses of all nodes. If IP quorum is configured incorrectly, the network must also deal with possible security implications of exposing the service IP addresses, because this connectivity can also be used to access the service GUI.
- Port 1260 is used by IP quorum applications to communicate from the hosts to all nodes.
- The maximum round-trip delay must not exceed 80 ms, which means 40 ms each direction.
- A minimum bandwidth of 2 MBps is ensured for node-to-quorum traffic.

Even with IP Quorum applications at the third site, quorum disks at site one and site two are required because they are used to store important cluster data. The maximum number of applications that can be deployed is five. For more information about IP Quorum requirements and installation, see the IP Quorum configuration section in IBM Knowledge Center:

IP Quorum requirements

HyperSwap Volumes

I type of volume is used that is called the *HyperSwap Volume*. These HyperSwap Volumes consist of a Master Volume and a Master Change Volume (CV) in one system site, and an Auxiliary Volume and an Auxiliary CV in the other system site. An active-active Metro Mirror relationship exists between the two sites. As with a regular Metro Mirror relationship, the active-active relationship attempts to keep 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 other considerations regarding the Change Volume, see “Global Mirror Change Volumes functional overview” on page 176.

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 guaranteed by the HyperSwap function.

Figure A-4 shows how the HyperSwap Volume is implemented.

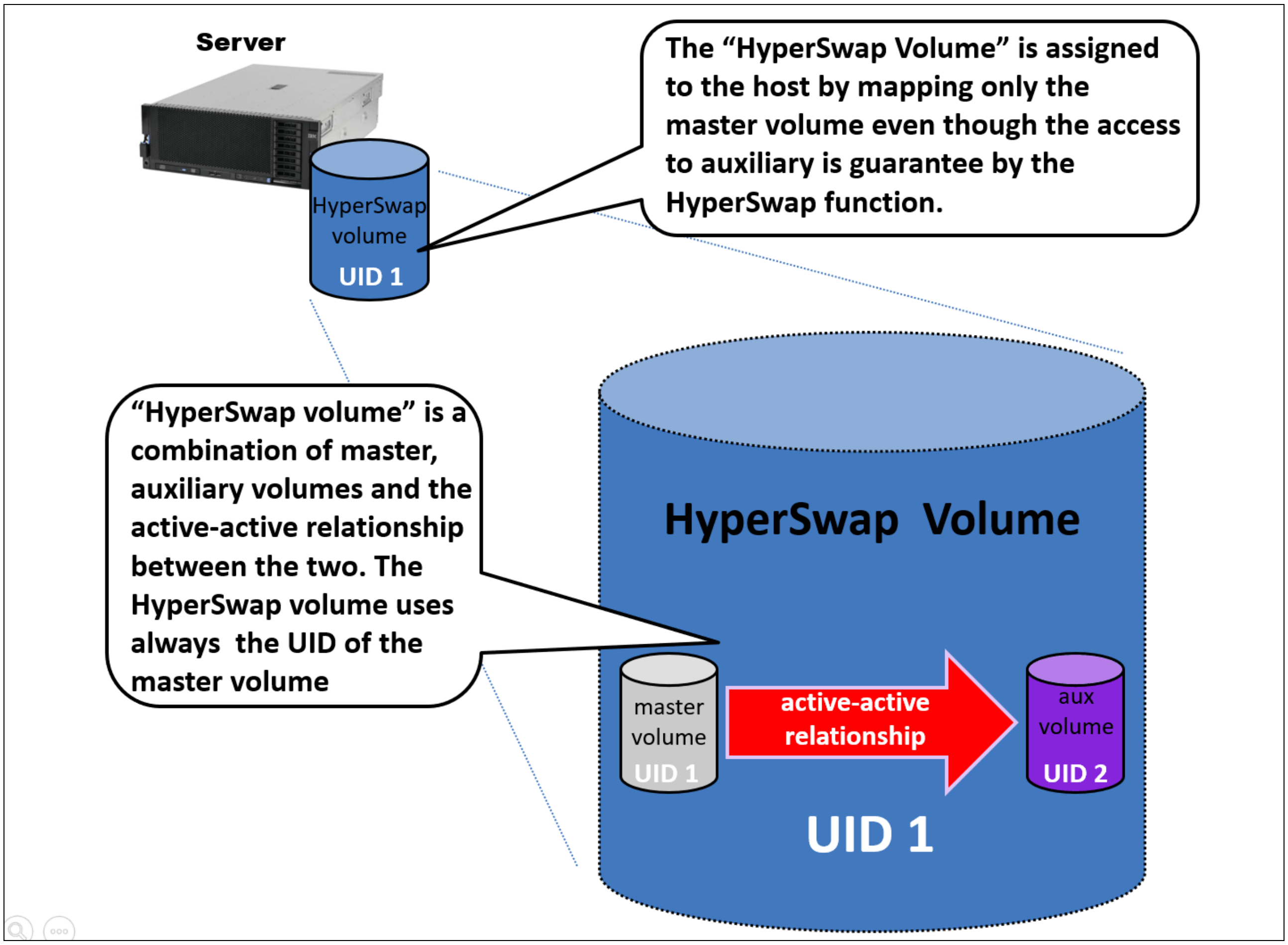

Server
HyperSwap volume
UID 1
The "HyperSwap Volume" is assigned to the host by mapping only the master volume even though the access to auxiliary is guarantee by the HyperSwap function.
"HyperSwap volume" is a combination of master, auxiliary volumes and the active-active relationship between the two. The HyperSwap volume uses always the UID of the master volume
HyperSwap Volume
master volume
UID 1
active-active relationship
aux volume
UID 2
UID 1

Figure A-4 HyperSwap Volume

The active-active Metro Mirror replication workload will traverse 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 will be Primary or Secondary based on the Metro Mirror active-active relationship direction.

In the current HyperSwap implementation, the read and write operations are always routed to the Primary copy. Therefore, hosts that access the Secondary copy will 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 will switch the direction of the active-active relationships, swapping the Secondary volume to Primary and vice versa.

Implementation guidelines

Configuring an IBM FS9100 HyperSwap solution requires to create a clustered IBM FS9100 across two locations or *Failure Domains*.

In IBM FS9100 HyperSwap, the minimum bandwidth for node-to-node communication between the failure domains is the peak write throughput from all hosts in both sites. This bandwidth is sufficient only if all volumes are accessed from hosts in one site.

To ensure the flawless operation of the HyperSwap cluster under all workload conditions, the bandwidth for node-to-node communication between the sites needs to be the sum of these numbers: for volumes that are accessed from hosts in the same site, the peak write throughput, and in addition for volumes that are accessed concurrently from hosts in different sites, the peak read and twice the peak write throughput.

To interconnect the IBM FS9100 to the two failure domains, two approaches are supported:

- Attach each IBM FS9100 node canister to the FC switches directly in the local and the remote failure domains. Therefore, all node-to-node traffic can be occur without traversing inter-switch links (ISLs). This approach is referred to as *HyperSwap No ISL configuration*.
- Attach each IBM FS9100 node canister only to local FC switches and configure ISLs between failure domains for node-to-node traffic. This approach is referred to as *HyperSwap ISL configuration*.

No ISL configuration

Figure A-5 shows the typical HyperSwap No ISL configuration.

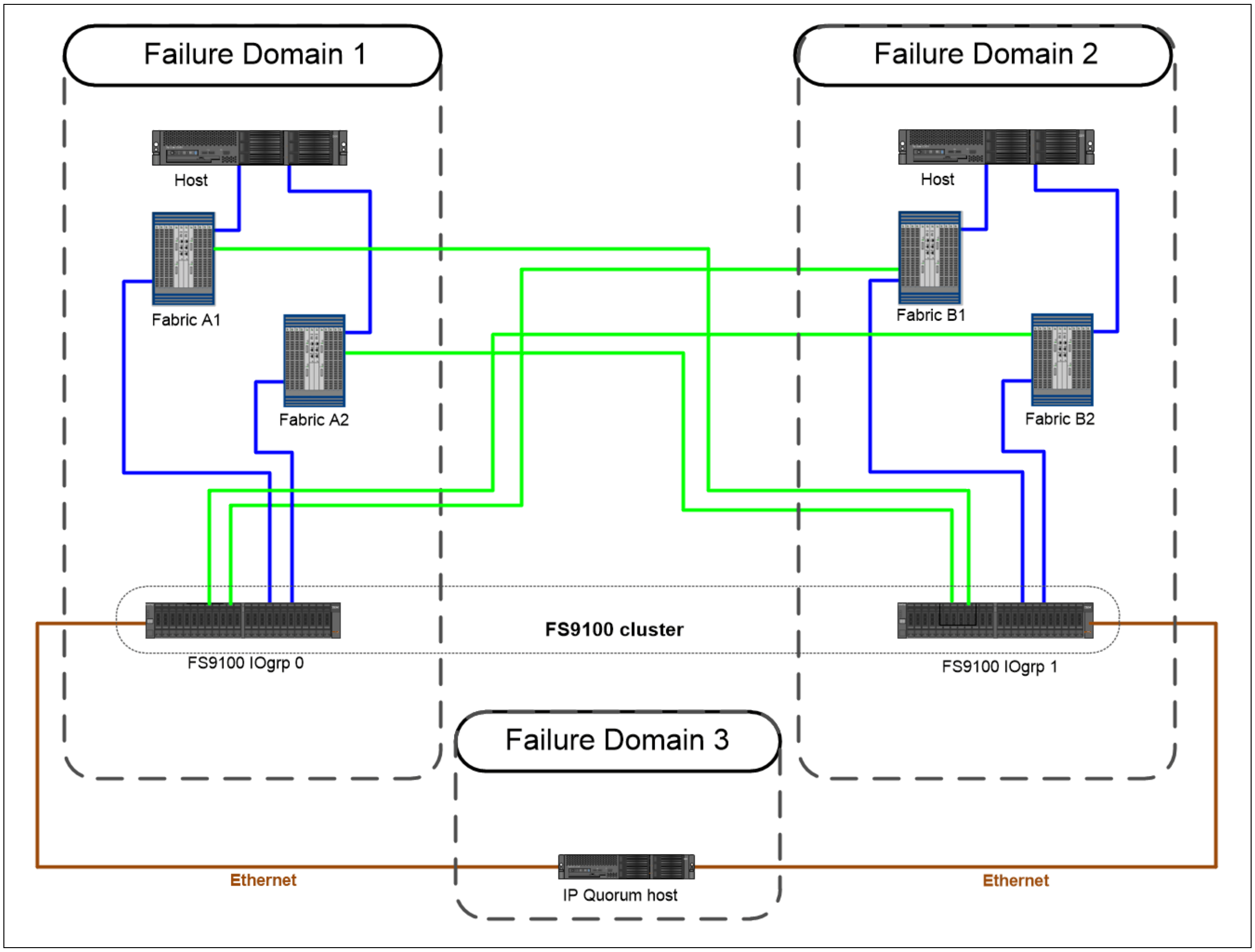


Figure A-5 FS9100 HyperSwap No ISL configuration

Note that the picture summarizes the connectivity schema only, the actual number of ports to be connected to local and remote site might vary on the specific environment and workload. The maximum distance of the inter site connectivity is limited by the IBM FS9100 ports buffer-to-buffer credits amount, that is 40.

With this value, we can achieve 20 km, 10 km, or 5 km with 4 Gb, 8 Gb, or 16 Gb port speed. Remember that to cover such distances, Single Mode Long Wave SFPs are required.

For this configuration, use the following guidelines:

- Use dedicated ports for the node-to-node connectivity. Considering the configuration in Figure A-5 on page 408 at least four per-node canisters are needed. Two of this ports' per-node canisters have to be attached to the switches at the remote site.
- Use the other FC ports for host. Attach these ports to the switches at the local and remote site.
- A separate zone is configured for node-to-node traffic in every switch.
- Hosts must be zoned with the IOgroup of both sites.
- On Flash System 9100, configure the **`localfcportmask`** value to the port that is dedicated to the node-to-node connectivity. Use switch zoning to ensure that the node-to-remote site node communication uses the FC ports that are dedicated to this purpose.
- The Flash System 9100 node canisters of the same I/O Group communicate across the controller enclosure internal PCI and FC I/O fabric.
- If you virtualize external storage, the zones must not contain multiple back-end disk systems.
- Set the BB credit to switch ports used by IBM FS9100 to 40.

ISL configuration

Figure A-6 shows the typical HyperSwap ISL configuration.

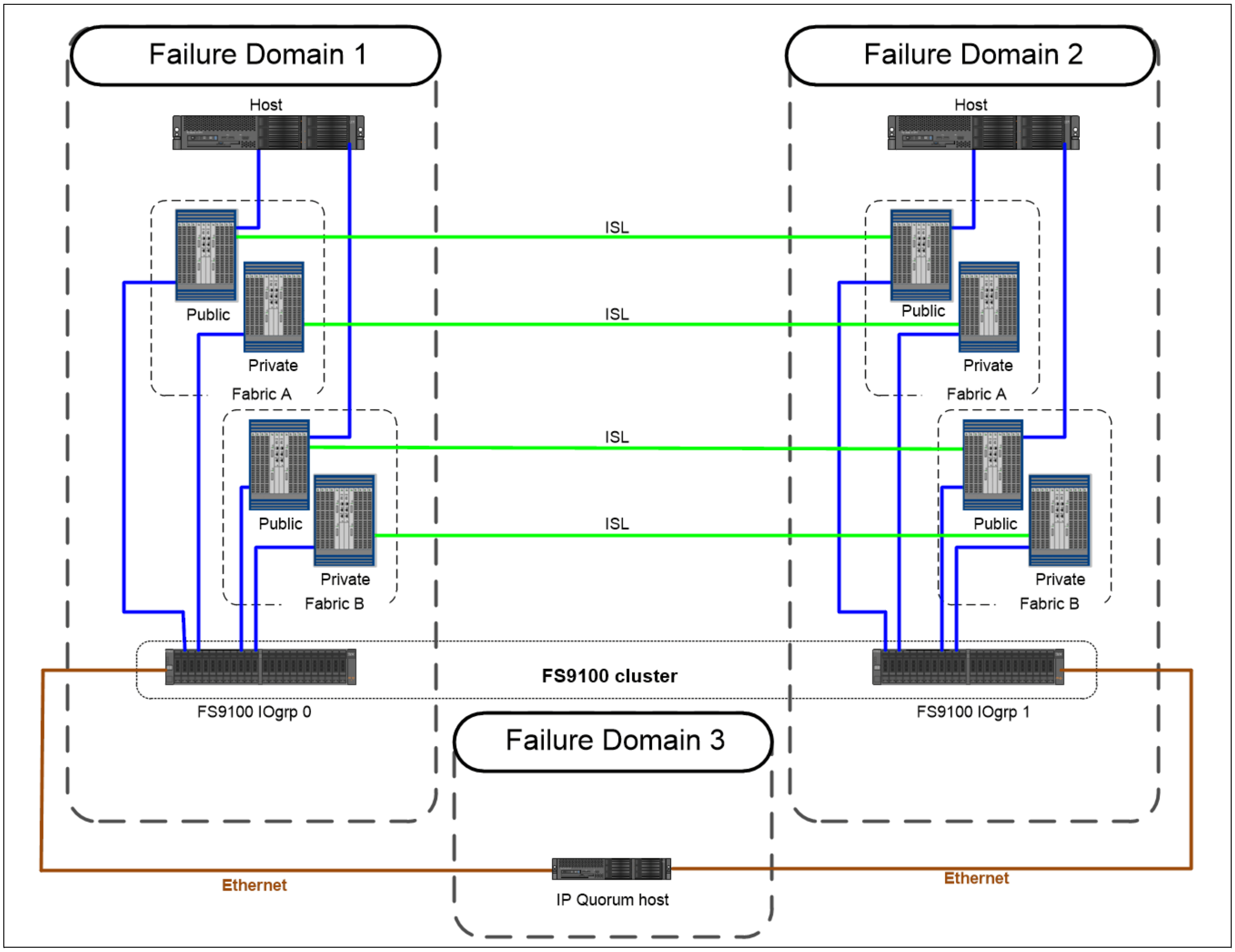


Figure A-6 FS9100 HyperSwap ISL configuration

This configuration has some requirements in terms of SAN design:

- One SAN is dedicated for IBM FS9100 node-to-node communication. This SAN is referred to as the *Private SAN*.
- One SAN is dedicated for host attachment. This SAN is referred to as the *Public SAN*.
- Each SAN must have at least one ISL for redundancy, and the bandwidth that is supplied by the ISL must be sized correctly.

Private and Public SANs can be implemented by using any of the following approaches:

- Dedicated FC switches for each SAN
- Switch partitioning features
- Virtual or logical fabrics

Similar configurations using WDM or FCIP links between the sites are also supported. Note that the picture summarizes the connectivity schema only, the actual number of ports to be connected to local switches and the number of ISLs might vary on the specific environment and workload.

The maximum distance of the inter-site connectivity supported with this configuration is 300 km.

For this configuration use the following guidelines:

- Use dedicated ports for the node-to-node connectivity, at least two per node canister, one per fabric, to be attached to the Private SAN.
- Use the other FC ports for host. Attach these ports to the Public SAN.
- A separate zone is configured for node-to-node traffic in each fabric.
- Hosts must be zoned with the IOgroup of both sites.
- A single trunk between switches is required for the Private SAN.
- External storage controller virtualized by IBM FS9100 must be attached the Public SANs.
- If you virtualize external storage, the zones must not contain multiple back-end disk systems.
- On Flash System 9100, configure the **`localfcportmask`** value to the port that is dedicated to the node-to-node connectivity. Use switch zoning to ensure that the node-to-remote site node communication uses the FC ports that are dedicated to this purpose.
- ISLs that belong to the Private SANs must not be shared with other traffic, and they must not be oversubscribed.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Some publications referenced in this list might be available in softcopy only:

- *Accelerate with IBM FlashSystem V840 Compression*, REDP-5147
- *Deploying IBM FlashSystem V840 Storage in a VMware and Cloud Environment*, REDP-5148
- *IBM FlashSystem 900 Model AE3 Product Guide*, REDP-5467
- *Implementing IBM FlashSystem 900 Model AE3*, SG24-8414
- *FlashSystem V9000 Product Guide*, REDP-5468
- *IBM FlashSystem V9000 and VMware Best Practices Guide*, REDP-5247
- *IBM FlashSystem V9000 in a VersaStack Environment*, REDP-5264
- *IBM FlashSystem V9000 Version 7.7 Product Guide*, REDP-5409
- *IBM System Storage SAN Volume Controller and Storwize V7000 Best Practices and Performance Guidelines*, SG24-7521
- *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize V8.1*, SG24-7933
- *Implementing the IBM Storwize V7000 with IBM Spectrum Virtualize V8.1*, SG24-7938
- *Introducing and Implementing IBM FlashSystem V9000*, SG24-8273

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Other publications and resources

These websites are also relevant as further information sources:

- IBM FlashSystem 9100

  https://www.ibm.com/uk-en/marketplace/flashsystem-9100/details
- IBM FlashSystem resources

  https://www.ibm.com/uk-en/marketplace/flashsystem-9100/resources
- IBM FlashSystem 9100 in IBM Knowledge Center

  https://www.ibm.com/support/knowledgecenter/STSLR9_8.2.0/com.ibm.fs9100_820.doc/fs9100_ichome.html

- IBM Storage Insights

  https://www.ibm.com/support/knowledgecenter/SSQRB8/com.ibm.spectrum.si.doc/tpch_saas_welcome.html
- IBM FlashSystem family

  https://ibm.biz/BdsaFH
- IBM Flash Storage

  https://www.ibm.com/it-infrastructure/storage/flash
- IBM System Storage Interoperation Center (SSIC)

  https://www.ibm.com/systems/support/storage/ssic/interoperability.wss

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Redbooks

IBM FlashSystem 9200 and 9100 Best Practices and Performance

SG24-8448-00

ISBN 0738458295

IBM

(1.0” spine)
0.875”<->1.498”
460 <-> 788 pages

SG24-8448-00

ISBN 0738458295

Printed in U.S.A.