ibm.com/redbooks

IBM DS8900F
Performance Best Practices and Monitoring

Peter Kimmel

Sherri Brunson

Lisa Martinez

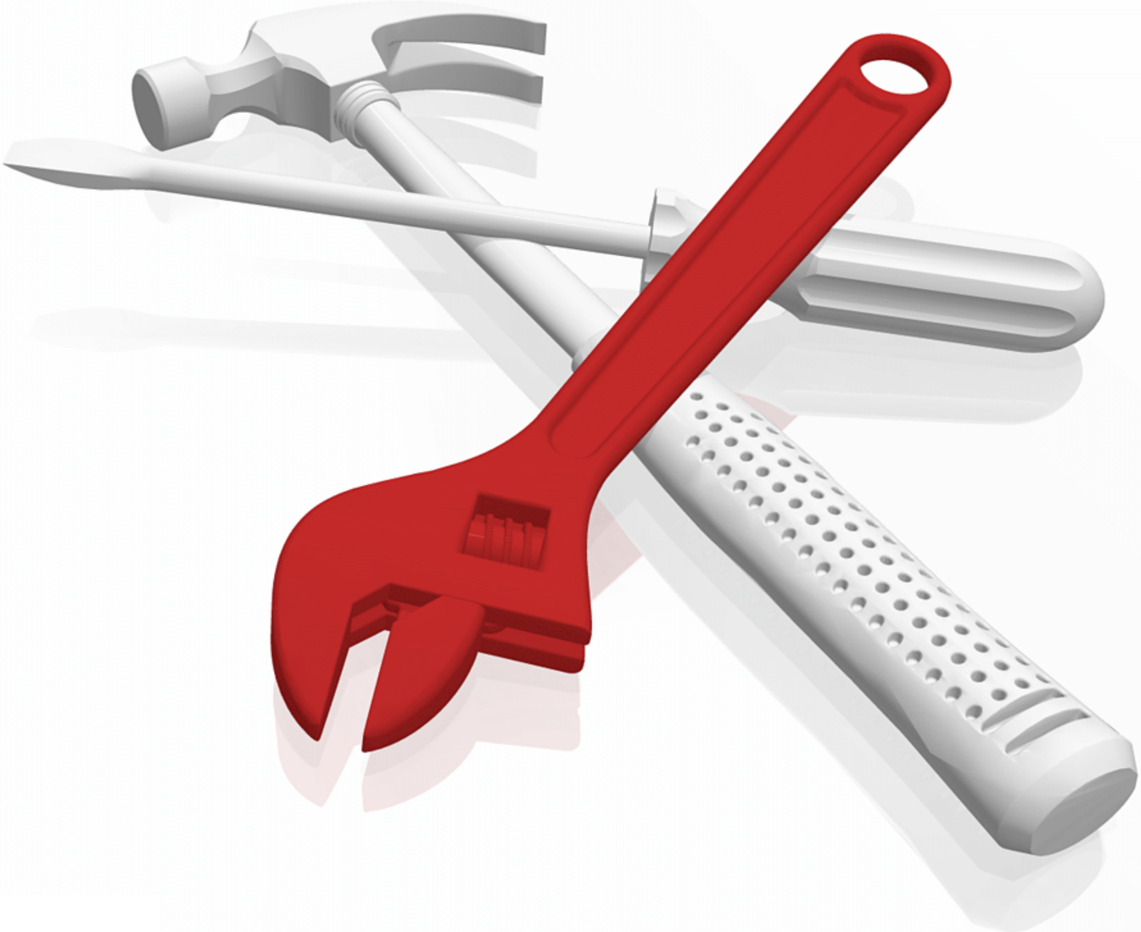
Luiz Moreira

Ewerson Palacio

Rick Pekosh

Ali Rizvi

Paul Smith



IBM Redbooks

IBM DS8900F Performance Best Practices and Monitoring

December 2021

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

First Edition (December 2021)

This edition applies to Version 9, Release 2, of the DS8000 family of storage systems.

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®
CICS®
Cognos®
Db2®
DB2®
DS8000®
Easy Tier®
Enterprise Storage Server®
FICON®
FlashCopy®
GDPS®
Global Technology Services®
HyperSwap®
IBM®
IBM Cloud®
IBM FlashSystem®
IBM Research®
IBM Spectrum®
IBM Z®
IBM z Systems®
IBM z14®
Interconnect®
Parallel Sysplex®
POWER®
POWER8®
POWER9™
PowerPC®
Redbooks®
Redbooks (logo) ®
Storwize®
System z®
WebSphere®
z Systems®
z/Architecture®
z/OS®
z/VM®
z/VSE®
z13®
z15™

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, 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 is intended for individuals who want to maximize the performance of their DS8900 storage systems and investigate the planning and monitoring tools that are available.

Performance: Any sample performance measurement data that is provided in this book is for comparative purposes only. The data was collected in controlled laboratory environments at a specific point by using the configurations, hardware, and firmware levels that were available then. The performance in real-world environments can vary. Actual throughput or performance that any user experiences also varies depending on considerations, such as the I/O access methods in the user's job, the I/O configuration, the storage configuration, and the workload processed. The data is intended only to help illustrate how different hardware technologies behave in relation to each other. Contact your IBM representative or IBM Business Partner if you have questions about the expected performance capability of IBM products in your environment.

Authors

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

Peter Kimmel is an IT Specialist and Advanced Technical Skills team lead of the Enterprise Storage Solutions team at the ESCC in Frankfurt, Germany. He joined IBM Storage in 1999, and since then has worked with various DS8000® generations, with a focus on architecture and performance. Peter co-authored several DS8000 IBM publications. He holds a Diploma (MSc) degree in physics from the University of Kaiserslautern.

Sherri Brunson joined IBM in March of 1985 and worked as a large-system IBM service representative before becoming a Top Gun in 1990. Sherry is a Top Gun in the Eastern US for all storage products, Power Systems servers, and IBM Z®. She has supported and implemented DS8000 and scaled-out network appliance storage products globally, and developed and taught educational classes. She also has taught IBM Z classes in the United States.

Lisa Martinez has been working as a client technical specialist within the Advanced Technology Group since January 2012. Her primary focus has been with pre-sales support for DS8000 and IBM Copy Services Manager. Her experience includes roles as a storage architect in the Specialty Services Area in IBM Global Technology Services® (IBM GTS) and a test architect in disk storage. Lisa holds degrees in computer science from New Mexico Highlands University and electrical engineering from the University of New Mexico.

Luiz Fernando Moreira is a Certified IT Specialist Thought Leader in the areas of Storage, Performance and Capacity. Luiz has been working in the Information Technology area for many years, always working with the z Architecture and more specifically with z/VM® systems, DS8000 and IBM Tape Libraries. Luiz is a former ITSC - International Technical System Center foreign assignee in Poughkeepsie, NY where he coordinated z/VM residencies and has written some IBM Redbooks and taught technical workshops. Luiz holds a a Diploma (MSc) degree in Engineering from the Instituto Militar de Engenharia (IME) in Rio de Janeiro Brazil.

Ewerson Palacio is an ITSO Redbooks Project Leader. He holds Bachelors's degree in Math and Computer Science. He worked for IBM Brazil for over 40 years and retired in late 2016 as an IBM Distinguished Engineer. Ewerson co-authored a large number of zSystems Redbooks, created and taught ITSO Technical Workshops and Seminars to Clients, IBMers, and Business Partners around the globe.

Rick Pekosh is a Certified IT Specialist working in the IBM Advanced Technology Group as a DS8000 subject matter expert specializing in High Availability, Disaster Recovery, and Logical Corruption Protection solutions. He works with customers, IBM Business Partners, and IBMers in North America. Rick began working with the DS8000 in early 2005 while working as a technical sales specialist and functioning as a regional designated specialist. He joined IBM in 2001 after spending 20 years in application development in various roles for the Bell System and as a consultant. Rick earned a BS degree in Computer Science from Northern Illinois University; and an MBA from DePaul University.

Ali Rizvi is a z/OS® Systems Programmer working on GTS managed accounts in Melbourne, Australia. He has 13 years of experience primarily in setting up Mainframe Hardware and z/OS. He has also worked in IBM Systems as a Client Technical Specialist in Canberra, Australia. Before Ali joined IBM, he worked as the Mainframe Hardware Team Lead for Department of Human Services, Australia. His areas of expertise include working with IODFs and using HCD. He holds a bachelor's degree in computing and information systems.

Paul Smith joined IBM in August 1994 and worked as a Mid-Range system and storage IBM service representative before joining Lab Services in the United Kingdom.
Paul has supported the IBM storage portfolio including DS8000 from inept through to SAN Volume Controller, Storwize® V7000 and IBM FlashSystem® 9200 along with associated SAN technology from Cisco and Brocade. He is also experienced with data migration of storage systems along with design and providing education classes to colleagues and clients.

Thanks to the authors of the previous editions of the DS8000 performance Redbooks.

- Authors of *DS8800 Performance Monitoring and Tuning*, SG24-8013, published in 2012, were:

  Gero Schmidt, Bertrand Dufrasne, Jana Jamsek, Peter Kimmel, Hiroaki Matsuno, Flavio Morais, Lindsay Oxenham, Antonio Rainero, Denis Senin

- Authors of I*BM System Storage DS8000 Performance Monitoring and Tuning*, SG24-8318, published in 2016, were:

  Axel Westphal, Bert Dufrasne, Wilhelm Gardt, Jana Jamsek, Peter Kimmel, Flavio Morais, Paulus Usong, Alexander Warmuth, Kenta Yuge.

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 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

Part 1

IBM System Storage DS8000 performance considerations

This part provides an overview of the DS8000 characteristics and logical configuration concepts to achieve optimum performance.

This part includes the following topics:

- IBM System Storage DS8900F family characteristics
- Hardware configuration
- Logical configuration concepts and terminology
- Logical configuration performance considerations

1

IBM System Storage DS8900F family characteristics

This chapter contains a high-level description and introduction to the storage server performance capabilities. It provides an overview of the model characteristics of the DS8900F family which allow the DS8000 series to meet a wide range of performance challenges. For some charts, when comparing performance differences between models, we include the previous generation DS8000 which is the DS8880 product family.

This chapter includes the following topics:

- The storage server challenge
- Meeting the challenge: The DS8000 product family
- Advanced caching algorithms

1.1 The storage server challenge

One of the primary criteria in judging a storage server is performance, that is, how fast it responds to a read or write request from an application server. How well a storage server accomplishes this task depends on the design of its hardware, its firmware, and its configuration.

Data continually flows from one component to another within a storage server. The objective of server design is hardware that has sufficient throughput to keep data flowing smoothly without waiting because a particular component is busy. When data stops flowing because a component is busy, a bottleneck forms. Obviously, it is preferable to minimize the frequency and severity of bottlenecks.

The ideal storage server is one in which all components are used effectively and efficiently with minimal bottlenecks. This scenario is the case if the following conditions are met:

- The storage system is designed well, with all hardware components in balance. To provide balance over a range of workloads, a storage server must allow a range of hardware component options.
- The storage system is appropriately sized for the client workload. Where options exist, the correct quantities of each option are chosen to maximize price/performance.
- The storage system is set up well. Where options exist in hardware installation and logical configuration, these options are chosen correctly.

Easy Tier® automatic rebalancing and tiering options can help achieve and maintain optimum performance even in an environment of ever-changing workload patterns, but they cannot replace a correct sizing of the storage system.

1.1.1 Performance numbers

Performance numbers alone provide only value with an explanation and making clear what exactly they mean. Raw performance numbers provide evidence that a particular storage server is better than the previous generation model or better than a competitive product. But, isolated performance numbers often do not match up with a production environment’s performance requirements. It is important to understand how raw performance numbers relate to the performance of the storage server when processing a particular production workload.

Throughput numbers are achieved in controlled tests that push as much data as possible through the storage server as a single component. At the point of maximum throughput, the system is so overloaded that response times are greatly extended. Trying to achieve such throughput numbers in a normal business environment brings protests from the users of the system because response times are poor.

To assure yourself that the DS8900F family offers the fastest technology, consider the performance numbers for the individual flash drives, host adapters, and other components of the various models, and for the storage system as a whole. The DS8900F family uses the most current technology available and using a rigorous approach when planning the DS8000 hardware configuration to meet the requirements of a specific environment.

1.1.2 Recommendations and rules

Hardware selection is sometimes based on general recommendations and rules. A general rule is a simple guideline for making a selection based on limited information. You can make a quick decision, with little effort, that provides a solution that may meet performance requirements most of the time. The disadvantage is that it may not satisfy performance requirements when it matters most - during peak on-line activity where latency (response time) typically matters most or during batch processing where throughput (MB/sec or GB/sec) typically matters most. That, in turn, may leave the IT staff unable to meet the business's expectations for a seamless customer experience nor its Service Level Agreements (SLAs). You can increase the chances of meeting those expectations and SLAs during those peaks by making it more conservative. However, that generally involves more hardware, which means a more expensive solution, without substantiating the need. Use general rules only when no information is available to make a more informed decision.

1.1.3 Modeling your workload

The better and recommended approach is to base hardware configuration decisions on empirical analyses. In other words, base the decision on performance models of the actual workload. IBM's performance modeling tool is Storage Modeller (StorM) which calculates the latency (response time) and peak throughput (MB/sec or GB/sec) of the proposed storage server configurations based on the workload modeled. By converting the models into mathematical formulas, StorM allows the results to be applied to a wide range of workloads. StorM allows many variables of hardware to be brought together. This way, the effect of each variable is integrated, producing a result that shows the overall performance of the storage server.

For more information about StorM, see Chapter 6.1, "IBM Storage Modeller" on page 126.

1.1.4 Allocating hardware components to workloads

Two contrasting methods are available to allocate the use of hardware components to workloads:

- The first method is spreading the workloads across components, which means that you try to share the use of hardware components across all, or at least many, workloads. The more hardware components are shared, the more effectively they are used, which reduces total cost of ownership (TCO). For example, to attach multiple UNIX servers, you can use the same host adapters (HAs) for all hosts instead of acquiring a separate set of HAs for each host. However, the more that components are shared, the more potential exists that one workload may dominate shared resources.
- The second method is isolating workloads to specific hardware components, which means that specific hardware components are used for one workload, others for different workloads. The downside of isolating workloads is that certain components are unused when their workload is not demanding service. On the upside, when that workload demands service, the component is available immediately. In other words, the workload does not contend with other workloads for that resource.

Spreading the workload maximizes the usage and performance of the storage server as a whole. Isolating a workload is a way to maximize the workload performance, making the workload run as fast as possible. Automatic I/O privatization can help avoid a situation in which less-important workloads dominate the mission-critical workloads in shared environments, and allow more shared environments.

If you expect growing loads, for example, when replacing one system with a new one that also has a much bigger capacity, add some contingency for this amount of foreseeable I/O growth.

For more information, see 4.2, “Configuration principles for optimal performance” on page 66.

1.2 Meeting the challenge: The DS8000 product family

The DS8910F, D8950F, and DS8980F make up the current generation of IBM DS8000 storage servers. They share the same architecture and offer the same set of rich features and advanced functions differing only in:

- POWER9™ processor
- Quantity of cores
- Available system memory which consists of cache and non-volatile storage (NVS) or persistent memory
- Number of host adapters
- Number of zHyperLinks
- Physical capacity

This means there is a wide range of possible hardware configurations available to fit a multitude of performance requirements in both workload type and size.

The DS8900Fs include and share many common features that improve performance including:

- Dual-clustered POWER9 servers with symmetrical multiprocessor (SMP) architecture
- Multi-threaded by design, simultaneous multithreading (SMT)
- Intelligent caching algorithms developed by IBM Research®
- PCIe 16 Gb/sec and 32 Gb/sec Fibre Channel/FICON® Host Adapters
- High-bandwidth, fault-tolerant internal interconnections with a switched PCIe Gen3 architecture
- High Performance Flash Enclosures (HPFE) Gen2
- Three distinct tiers of flash media
- Easy Tier which is a set of built-in advanced functions that autonomically rebalance backend resources to reduce or eliminate hot spots

In addition to this list of hardware and microcode based features, the following sections provide additional on those and other performance-enhancing features not listed above.

1.2.1 DS8900F family models and characteristics

The DS8900F, which is the ninth generation of the DS8000, is available in three models:

- **DS8910F** Flexibility Class model 993 which can be rack mounted in a customer provided rack, installed in a compatible IBM Z or LinuxONE rack; or, model 994 which is shipped racked as a standalone system just like the DS8950F & DS8980F are.
- **DS8950F** Agility Class
- **DS8980F** Analytic Class

Table 1-1 provides a high level overview of the DS8900F models for comparison. This is only a high level overview and comparison. Note that all three are available with either single phase or 3-phase power.

Table 1-1 DS8900F model overview & comparison

Component	**DS8910F Flexibility Class (rack mounted)**	**DS8910F Flexibility Class (racked)**	**DS8950F Agility Class**	**DS8980F Analytic Class**
Base frame model	993	994	996	998
Expansion frame model (0 / 1)	n/a	n/a	E96	E96
IBM POWER9 CPCs	MTM 9009-22A	MTM 9009-22A	MTM 9009-42A	MTM 9009-42A
Number of processor cores per system (Min / Max)	16 / 16	16 / 16	20 / 40	44 / 44
System memory (Min / Max)	192 GB / 512 GB	192 GB / 512 GB	512 GB / 3.4 TB	4.3 TB / 4.3 TB
System NVS memory (Min / Max)	8 GB / 32 GB	8 GB / 32 GB	32 GB / 128 GB	128 GB / 128 GB
Host Adapters (Min / Max) (16 Gb/sec or 32 Gb/sec)	2 / 8	2 / 16	2 / 32	2 / 32
Host Adapter ports (Min / Max)	8 / 32	8 / 64	8 / 128	8 / 128
zHyperLinks (Min / Max)	0 / 4	0 / 4	0 / 12	0 / 12
High-Performance Flash Enclosures (HPFE) Gen2 pairs (Min / Max)	1 / 2	1 / 4	1 / 8	1 / 8
Device Adapter pairs (Min / Max)	1 / 2	1 / 4	1 / 8	1 / 8
Flash drives (Min / Max)	16 / 96	16 / 192	16 / 384	16 / 384

The following sections provide a short description of the main hardware components.

POWER9 processor technology

The DS8000 series uses the IBM Power technology, which is the foundation of the storage system logical partitions (LPARs). The DS8900F family (99x) models use the dual 6-way up to dual 44-way processor complexes of the 64-bit microprocessors. Within the POWER9 servers, the DS8900F family offers up to 4.3 TB of cache.

The IBM POWER® processor architecture offers superior performance and availability features, compared to other conventional processor architectures on the market. POWER7+ allowed up to four intelligent simultaneous threads per core, which is twice of what larger x86 processors have. Like the DS8880s, which were based on POWER8® architecture, the DS8900Fs allow up to eight simultaneous threads per core.

Switched PCIe disk connection

The I/O bay interconnect for the members of the DS8900F family runs through the Peripheral Component Interconnect® Express Gen3 (PCI Express Gen3 or “PCIe Gen3”) standard. This is a major change to the DS8870, which previously had the GX++ I/O bus with the PCIe Gen2 interconnects. Because this supplemental I/O bridge component of the GX++ bus is gone in the DS8880 storage system, it also gives some additional latency advancements in the back end. Also, the HPFE connect directly through PCIe in all DS8880 models. In addition, here the number of lanes for the HPFE attachment are doubled, to potentially allow even higher throughputs.

Flash media

The DS8900F models are all-flash arrays offering three distinct tiers of flash drives with each having its own respective performance profile. DS8000 storage pools can have from one to three flash tiers. That means that there are a variety of possible configurations that can satisfy both performance and capacity requirements. The goal is to configure a storage system that meets the performance and capacity requirements while achieving a balanced price for performance configuration.

Flash tiers and flash media sizes:

- Tier 0:
  - High-Performance Flash (HPF)
  - 800 GB, 1.6 TB, and 3.2 TB
- Tier 1:
  - HCF
  - 3.84 TB
- Tier 2:
  - HCF
  - 1.92 TB, 7.68 TB, and 15.36 TB

Typically, DS8900Fs are configured with just one or two flash tiers. Depending on the particular workload, that may mean a single tier of high performance flash or high capacity flash or a combination of the two. That flexibility is a testament to the overall performance capabilities of DS8900Fs and the effectiveness of Easy Tier which autonomically rebalances backend resources by eliminating or greatly reducing hot spots.

For example, a workload with extremely high transaction rates, that demands the lowest latency possible, may require a single tier of high performance flash; a workload with high capacity requirements and low I/O rates may be satisfied with a single tier of high capacity flash; or, for smaller workloads, suitable for a DS8910F, a combination of 800 GB HPF and 1.92 TB HCF may be appropriate and cost effective; or, for some very large workloads, suitable for a DS8950F, a combination of 3.2 TB HPF and 7.68 TB HCF may be appropriate and cost effective. The point is, the configuration to consider is workload dependent and the goal is to align it with price for performance.

Host adapters

Host Adapters serve two purposes. One, they provide connectivity between the host server and the DS8000 storage server for I/O (input/output *or* reads and writes). Two, they provide connectivity between DS8000s to carry synchronous (Metro Mirror) and asynchronous (Global Mirror and Global Copy) replication traffic.

The DS8900F models offer enhanced connectivity with the availability of both 16 Gb/sec and 32 Gb/sec four-port Fibre Channel/IBM FICON Host Adapters. Each adapter can auto-negotiate two generations back. That is the fiber channel standard for backward compatibility and is known as N-2 compatibility. The 16 Gb/sec adapter can auto-negotiate to 8 Gb/sec and 4 Gb/sec. Similarly, the 32 Gb/sec adapter can auto-negotiate to 16 Gb/sec and 8 Gb/sec. Both are available with either short wave or long wave Small Form-factor Pluggable (SFP). With this flexibility, you can benefit from the higher performance while maintaining compatibility with existing host processors and SAN infrastructures. Also, you can configure the ports individually which enables Fibre Channel Protocol (FCP) and FICON intermix on the same adapter. That enables the potential for very cost effective and efficient use of host adapter resources to meet both performance and availability requirements.

1.2.2 Advanced caching algorithms

The DS8000 series benefits from very sophisticated and advanced caching algorithms, developed by IBM Research, which autonomically manage cache resources to ensure they are used effectively and efficiently. There is nothing for the storage administrator to tune; in fact, there are no buttons to push nor knobs to turn. Additionally, the DS8000 uses 4 KB cache segment sizes which are sometimes referred to as 4 KB pages. The DS8000 is the only enterprise storage server on the market that employs a 4 KB cache segment size which is the most efficient design in terms of cache usage efficiency, especially for small block OLTP workloads. Data Warehouse (DWH) workloads benefit from caching exactly what is needed. For larger I/Os, 4 KB segments are linked together, as needed. Details on the caching algorithms follow immediately below.

Sequential Prefetching in Adaptive Replacement Cache

Another performance enhancer is the use of the self-learning cache algorithms. The DS8000 caching technology improves cache efficiency and enhances cache-hit ratios, especially in environments that change dynamically. One of the patent-pending algorithms that is used in the DS8000 series is called Sequential Prefetching in Adaptive Replacement Cache (SARC).

SARC provides these advantages:

- Sophisticated, patented algorithms to determine what data to store in cache based on the recent access and frequency needs of the hosts.
- Prefetching, which anticipates data before a host request and loads it into cache.
- Self-learning algorithms to adapt and dynamically learn what data to store in cache based on the frequency needs of the hosts.

Adaptive Multi-Stream Prefetching

Adaptive Multi-Stream Prefetching (AMP) introduces an autonomic, workload-responsive, self-optimizing prefetching technology. This technology adapts both the amount and the timing of prefetch on a per-application basis to maximize the performance of the system.

AMP provides a provable, optimal sequential read performance and maximizes the sequential read throughputs of all RAID arrays where it is used, and therefore of the system.

Intelligent Write Caching

Intelligent Write Caching (IWC) improves performance through better write-cache management and destaging the order of writes. IWC can also double the throughput for random write workloads. Specifically, database workloads benefit from this new IWC cache algorithm.

SARC, AMP, and IWC play complementary roles. Although SARC is carefully dividing the cache between the RANDOM and the SEQ lists to maximize the overall hit ratio, AMP is managing the contents of the SEQ list to maximize the throughput that is obtained for the sequential workloads. IWC manages the write cache and decides the order and rate to destage to the flash media.

List Prefetch Optimizer

The Adaptive List Prefetch (ALP) algorithm, which is also known as the List Prefetch Optimizer (LPO), enables prefetch of a list of non-sequential tracks, providing improved performance for Db2® workloads. ALP, when working together with AMP, dramatically improves the performance of record identifier (RID) list scans. These features are an essential part of Data Warehouse and Business Intelligence workloads, and help analytics applications benefit from being run on DS8900F.

1.2.3 Performance with IBM Z

The DS8000 supports numerous IBM performance innovations for IBM Z environments. Here are the most significant ones:

- FICON extends the ability of the DS8000 storage system to deliver high-bandwidth potential to the logical volumes that need it when they need it. Working together with other DS8000 functions, FICON provides a high-speed pipe that supports a multiplexed operation.
- High Performance FICON for z (zHPF) takes advantage of the hardware that is available today, with enhancements that are designed to reduce the impact that is associated with supported commands. These enhancements can improve FICON I/O throughput on a single DS8000 port by 100%. Enhancements to the IBM z/Architecture® and the FICON interface architecture deliver improvements for online transaction processing (OLTP) workloads. When used by the FICON channel, the IBM z/OS operating system, and the control unit, zHPF is designed to help reduce impact and improve performance. Since the first implementation of zHPF, several generational enhancements were added.
- Parallel Access Volume (PAV) enables a single z Systems® server to process simultaneously multiple I/O operations to the same logical volume, which can help reduce device queue delays. This function is achieved by defining multiple addresses per volume. With Dynamic PAV, the assignment of addresses to volumes can be managed automatically to help the workload meet its performance objectives and reduce overall queuing.
- HyperPAV enables applications to achieve equal or better performance than with PAV alone while using fewer unit control blocks (UCBs) and eliminating the latency in targeting an alias to a base. With HyperPAV, the system can react immediately to changing I/O workloads.
- SuperPAV is an extension of HyperPAV and allows z/OS to use an *alias address* from another logical control unit (LCU).
- Multiple Allegiance expands the simultaneous logical volume access capability across multiple IBM Z servers. This function, along with PAV, enables the DS8000 to process more I/Os in parallel, helping to improve performance and enabling greater use of large volumes.
- Application and host integration. With the conversion of IBM Db2 I/Os to zHPF, massive performance gains for certain database tasks can be achieved. For example, the Db2 application can give cache prefetching hints into the storage system.
- zHyperWrite allows the I/O driver to schedule up to three parallel write I/Os and direct one or two of these writes to Metro Mirror (MM) target volumes to avoid the potential synchronous replication impact of Metro Mirror. Especially when considering Db2 logs, this integration of host-based mirroring into the DS8000 Metro Mirror with the zHyperWrite function can lead to significant response time reductions for these logs.
- zHyperLink provides a connectivity method that dramatically reduces latency by interconnecting the IBM Z system directly to the I/O bay of the DS8900F. Extremely low response times smaller than 20 µs are possible for qualifying I/Os. The current zHyperLink release supports both read and write I/Os.
- I/O priority queuing allows the DS8000 series to use I/O priority information that is provided by the z/OS Workload Manager to manage the processing sequence of I/O operations.
- Fabric I/O Priority provides end-to-end quality of service (QoS). It is a unique synergy feature between the IBM Z z/OS WLM, Brocade SAN Fabric, and the DS8900F system that is designed to manage QoS on a single I/O level.

- FICON Dynamic Routing (FIDR) is an IBM Z and DS8900F feature that supports the use of dynamic routing policies in the switch to balance loads across inter-switch links (ISLs) on a per I/O basis.

1.2.4 Easy Tier

Easy Tier is a built-in dynamic data relocation advanced function that allows host transparent movement of data among the DS8000 storage system's backend resources. It acts as an optimizer that autonomically eliminates or greatly reduces "hot spots" by redistributing volume (or logical unit number (LUN)) extents, based on their respective "heatmap," across those resources. Easy Tier's purpose is to help maximize performance by maintaining an even spread of "hot" and "cold" extents across those resources. Easy Tier is included in the base license and enabled by default. It does not require any software to be installed on the host server, does not use any host cycles, and works regardless of the attached host operating system. And, storage administrators need not do anything except monitor Easy Tier through the integrated reporting available in the GUI.

> **Note:** Without getting too deep into technical details, this note is a brief description of the backend resources that Easy Tier monitors and optimizes. Flash drivesets come in groups of sixteen like type drives with eight flash drives comprising an array or rank. That means for every flash driveset there are two ranks. Ranks are grouped into storage pools during logical configuration. Volumes are carved up from one or more "extents", which are internal units of storage, within a particular storage pool. For both Count Key Data (CKD) and Fixed Block (FB) based operating systems, the DS8000 offers a choice of using all small or all large extents within a given storage pool. (Our recommendation, with few exceptions, is to configure storage pools using small extents, rather than large extents, because they provide far more granularity; therefore, small extents enable Easy Tier to work more efficiently and effectively.) Flash drivesets physically reside in High-Performance Flash Enclosures which are connected to the CPCs (POWER9 servers) via Device Adapter (DAs) pairs. So, the backend resources that Easy Tier is monitoring and optimizing are the ranks and device adapters.
>
> For a deep dive on the DS8000's internal virtualization, see "IBM DS8900F Architecture and Implementation Release 9.2" at:
>
> https://www.redbooks.ibm.com/redbooks/pdfs/sg248456.pdf

While there are likely noticeable or predictable patterns over time (across a day, week, or month), different volumes, or portions of them, typically experience different periods of extremely busy activity while at other times they may be idle. Some volumes, or portions of them, may rarely be busy. All of that is common and normal. Easy Tier is constantly monitoring and evaluating backend activity to decide what to move and whether to move it based on a cost/benefit analysis. It maintains statistics on backend activity collected on five-minute intervals and kept for seven days with the most recent twenty-four hours being the most important and the last or oldest twenty-four the least. Easy Tier collects statistics on large block I/Os, small block I/Os, reads & writes, latency (response time), I/O rates, and amount of data transferred. As a new day's statistics are collected the oldest day's statistics roll off. Hot extents are those that are overly busy compared with other extents in the same storage pool; cold extents are those will little to no activity. So, hot extents are eligible for promotion and cold extents for demotion.

Easy Tier is able to move extents between storage tiers within a storage pool with two or three tiers; that is known as inter-tier rebalancing. Think of this as vertical movement between tiers. Extents only move between adjacent tiers; there is no double promotion nor double demotion in three-tier storage pools. Easy Tier builds the inter-tier heatmap once approximately every twenty-four hours; that heatmap build start time varies by design. Easy Tier is also able to move extents laterally within a tier; that is known as intra-tier rebalancing. Think of this as horizontal movement within a tier. Easy Tier builds the intra-tier heatmap every six hours. With both inter-tier and intra-tier rebalancing, Easy Tier begins moving extents immediately after the respective plan is built. It considers small block I/Os when prioritizing extents to be moved. It moves a measured amount of extents at regular intervals without negatively impacting performance. Additionally, Easy Tier evaluates whether ranks or device adapters in multi-tier storage pools suddenly become overloaded or saturated by immediately moving extents to a lower tier. This process is known as warm demote and takes the highest priority.

Easy Tier has two operating modes, automatic mode and manual mode. In automatic mode, Easy Tier promotes and demotes extents (sub volume) between different flash tiers in the same storage pool (inter-tier rebalancing) for multi-tier storage pools; performs intra-tier rebalancing within a given tier in a storage pool; and, rebalances extents when ranks are added to a storage pool. Manual mode enables things like dynamic volume relocation from like type storage pool to another; dynamic storage pool merge of two like type storage pools; rank depopulation; and restripe a volume within the same pool.

Easy Tier Advanced and Innovative Features

- Heat Map Transfer Utility (HMTU)

  This is the ability to propagate the primary storage system's heatmap to the other DS8000 storage systems in 2-site, 3-site, and 4-site solutions. The primary storage system is performing read and writes, but the other storage systems are only performing writes (replicated data). HMTU enables the other storage systems in the solution to have a heatmap similar to the primary. When production I/O is transferred to another storage system in the solution, perhaps due to a HyperSwap® between Metro Mirror pairs, a site swap to flip production and disaster recovery sites, or an actual disaster recovery event, it does not need to learn from scratch. HMTU is integrated into IBM replication management solutions such as GDPS® and Copy Services Manager or can be implemented on a standalone server running Linux or Windows.

- Dynamic Pool Merge

  This is the ability to dynamically and seamlessly merge storage pools to create a two tier or three tier storage pool. For instance, add a high performance flash tier into a high capacity tiered storage pool to cost effectively improve performance; or, add high capacity flash into a high performance flash storage pool to cost effectively add capacity.

- Dynamic Volume Relocation (move volumes to another pool)

  This is the ability to dynamically and seamlessly move volumes to another storage pool. For instance, move volumes with little or modest I/O from a high performance flash pool to a high capacity flash pool or move volumes with high I/O activity from a high capacity pool to a high performance pool.

- Assign and Unassign volumes to a particular tier in a multi-tier storage pool

  This is known as pinning and unpinning, respectively. This ability allows you to assign volumes to a particular tier within a multi-tier storage pool. For instance, keep highly active volumes in a high performance flash tier and less active volumes, such as reference data, in a high capacity flash tier. That can be done permanently or scheduled to move volumes up and down before and after a particular workload runs such as month-end or quarter-end.

- Pause and resume the Easy Tier learning process

  This is the ability to turn off and on Easy Tier learning before and after an unusual workload runs to prevent it from skewing the Easy Tier statistics.

- Erase all monitoring data, including migration plans

This ability allows Easy Tier to “forget” everything it knows and start from scratch.

Note: This discussion on Easy Tier is not meant to be exhaustive. For additional details on Easy Tier see, *IBM DS8000 Easy Tier*, REDP-4667 at:

https://www.redbooks.ibm.com/redpapers/pdfs/redp4667.pdf

Summary

Easy Tier is a built-in advanced function designed to help optimize performance with minimal to no involvement of a storage administrator. The DS8000s are really smart storage systems and our recommendation is let them do what they are designed to do:

- Leave Easy Tier automatic mode running 24 x 7 x 365. The only exception to that is if you want to pause and resume Easy Tier learning before and after, respectively, an unusual workload runs.
- With few exceptions, configure storage pools using small extents, rather than large extents, because they provide far more granularity and enable Easy Tier to work more efficiently and effectively.
- Implement the Heat Map Transfer Utility to propagate the primary’s heatmap to the other DS8000 storage systems in the solution.
- Monitor Easy Tier via the built-in reporting integrated into the GUI.

2

Hardware configuration

This chapter describes the IBM System Storage DS8900F hardware configuration and its relationship to the performance of the device.

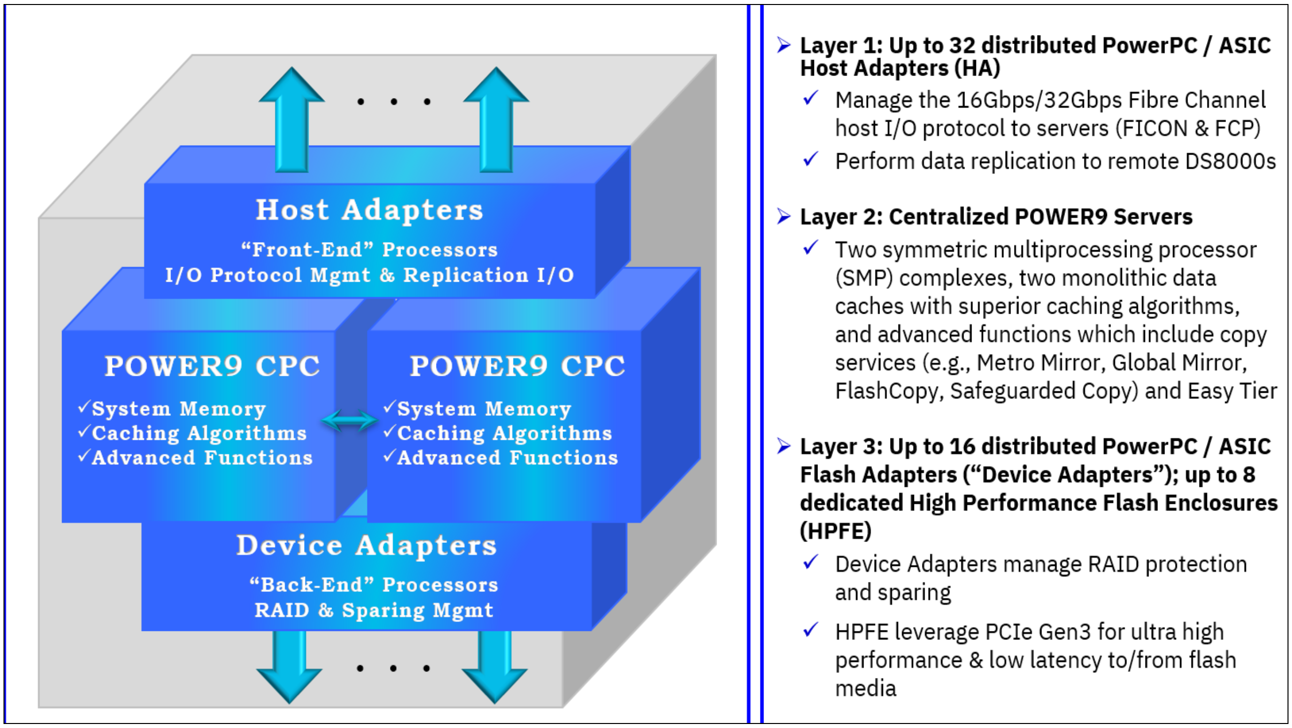
Understanding the hardware components, the functions they perform, and the DS8000 technology as a whole can help you select the appropriate DS8900F model and quantities of each component for the target workload(s). However, do not focus too much on any one component. Instead, focus on the storage system as a whole because they are designed to work as fully integrated systems. The ultimate criteria for storage system performance are response times (latency) and the total throughput (MBps or GBps).

This chapter includes the following topics:

- DS8900F model similarities
- Central Processor Complex and system memory
- DS8000 caching algorithms
- High-Performance Flash Enclosure - Gen2 (HPFE)
  - Device Adapters
  - Flash Tiers
- Host Adapters (HAs)
- zHyperLinks

2.1 DS8900F model similarities

The DS8910F, DS8950F, and DS8980F share the same “three layer” processor architecture (for reference see Figure 2-1), same set of caching algorithms, same high performance and high capacity flash drive options, and the same set of available advanced functions including copy services (such as Metro Mirror, Global Mirror, Global Copy, FlashCopy®, and Safeguarded Copy) and Easy Tier. All three models are capable of running the same exact types of workloads and delivering ultra high performance and ultra low latency. The differences between the models are: Scalability with respect to the POWER9 processor cores, system memory, host adapters, zHyperLinks, and physical capacity.



- **Layer 1: Up to 32 distributed PowerPC / ASIC Host Adapters (HA)**
  - ✓ Manage the 16Gbps/32Gbps Fibre Channel host I/O protocol to servers (FICON & FCP)
  - ✓ Perform data replication to remote DS8000s
- **Layer 2: Centralized POWER9 Servers**
  - ✓ Two symmetric multiprocessing processor (SMP) complexes, two monolithic data caches with superior caching algorithms, and advanced functions which include copy services (e.g., Metro Mirror, Global Mirror, FlashCopy, Safeguarded Copy) and Easy Tier
- **Layer 3: Up to 16 distributed PowerPC / ASIC Flash Adapters (“Device Adapters”); up to 8 dedicated High Performance Flash Enclosures (HPFE)**
  - ✓ Device Adapters manage RAID protection and sparing
  - ✓ HPFE leverage PCIe Gen3 for ultra high performance & low latency to/from flash media

Figure 2-1 DS8000 “Three Layer” Processor Architecture

With few exceptions, most of those components can be scaled independently from the others which enables very flexible custom configurations to address very specific performance requirements. For example, a DS8910F Flexibility Class, configured with sixteen POWER9 cores, 192 GB of system memory, and all 1.92 TB High Capacity Flash drives, may be appropriate for a moderately intensive workload; a DS8980F Analytic Class, which includes forty-four POWER9 cores and 4.3 TB of system memory, configured with the smallest available High Performance Flash drives (800 GB), may be required for a very I/O intensive workload requiring the lowest latency; or, a DS8950F Agility Class, configured with forty POWER9 cores, 1024 GB of system memory, and a mix of HPF and HCF may be appropriate other workloads.

The point is the DS8900F family offers configuration flexibility across and within the models to address widely varying customer performance and capacity needs. The best way to determine which DS8900F model and specific configuration is most appropriate for a given workload is to create a performance model of the actual workload. The performance modeling tool we use is StorM. To learn more about performance modeling and StorM see “IBM Storage Modeller” on page 126.

The sections that follow briefly discuss the scalable components and highlight the differences in scalability between the DS8900F models, where applicable.

2.2 Central Processor Complexes & System Memory

The DS8900F Central Processor Complexes are based on POWER9 symmetric multiprocessors (SMP) technology. The CPCs comprise “Layer 2” in the three layer processor architecture shown and described in Figure 2-1 on page 16. Layer 2 is discussed first because it is the main processor layer. The CPCs perform the “heavy lifting” which includes managing I/O to and from the hosts, cache operations, and all the advanced functions which includes copy services (such as Metro Mirror, Global Mirror, Global Copy, FlashCopy and Safeguarded Copy) and Easy Tier.

The pair of CPCs in the two DS8910F models are IBM Power System 9009-22A servers; the DS8950F and DS8980F each utilize a pair of 9009-42A servers. The number of cores in those CPCs and available system memory options are directly related and detailed below. On all DS8900F models, each central processor complex has its own dedicated system memory. Within each processor complex, the system memory is divided into:

- Memory used for the DS8000 control program
- Cache
- Non-Volatile Storage (NVS) or persistent memory

The DS8900F models and their respective number of processor cores and system memory options are shown below. The amount allocated as NVS or persistent memory scales according to the system memory configured:

- DS8910F - Flexibility Class (models 993 & 994)
  - 16 POWER9 cores with system memory options:
    - 192 GB (NVS memory 8 GB) or
    - 512 GB (NVS memory 32 GB)
- DS8950F - Agility Class (model 996)
  - 20 POWER9 cores with 512 GB system memory (NVS memory 32 GB)
  - 40 POWER9 cores with system memory options:
    - 1,024 GB (NVS memory 64 GB) or
    - 2,048 GB (NVS memory 128 GB) or
    - 3.4 TB (NVS memory 128 GB)
- DS8980F - Analytic Class (model 998)
  - 44 POWER9 cores with 4.3 TB system memory (NVS memory 128 GB)

2.2.1 DS8000 caching algorithms

Most, if not all, high-end storage systems have an internal cache that is integrated into the system design. How effectively and efficiently they each utilize their cache differs considerably. Competitive offerings use large and inefficient cache segment sizings (16 KB or 64 KB) and rely on simplistic caching algorithms such as Least Recently Used (LRU) to manage the cache. That, in turn, forces customers to acquire oversized cache memory to accommodate those inefficiencies. In other words, it is an attempt to solve a software problem with hardware. What differentiates the DS8900F (and earlier DS8000 generations), from competitive offerings, are its unique use of a very granular cache segment size (4 KB) and its superior caching algorithms. Working together they enable the DS8000 to deliver extremely effective and efficient cache utilization and extremely low response times (latency).

IBM DS8000 Development has a long-standing partnership with IBM Research which develops the DS8000's caching algorithms which include:

- Sequential Adaptive Replacement Cache (SARC)
- Adaptive Multi-stream Prefetching (AMP)
- Intelligent Write Caching (IWC)
- List Prefetch Optimizer (LPO)

The DS8900F can be equipped with up to 4.3 TB of system memory, most of which is configured as cache. With its POWER9 processors, the server architecture of the DS8900F makes it possible to manage such large caches with small cache segments of 4 KB (and large segment tables). The POWER9 processors have the power to support these sophisticated caching algorithms, which contribute to the outstanding performance that is available in the DS8900F.

These algorithms and the small cache segment size optimize cache hits and cache utilization. Cache hits are also optimized for different workloads, such as sequential workloads and transaction-oriented random workloads, which might be active at the same time. Therefore, the DS8900F provides excellent I/O response times.

Write data is always protected by maintaining a second copy of cached write data in the NVS of the other internal server until the data is destaged to the flash media.

Sequential Adaptive Replacement Cache

Sequential Adaptive Replacement Cache (SARC) is a self-tuning, self-optimizing solution for a wide range of workloads with a varying mix of sequential and random I/O streams. SARC is inspired by the Adaptive Replacement Cache (ARC) algorithm and inherits many of its features.

SARC attempts to determine the following cache characteristics:

- When data is copied into the cache.
- Which data is copied into the cache.
- Which data is evicted when the cache becomes full.
- How the algorithm dynamically adapts to different workloads.

The DS8900F series cache is organized in 4-KB pages that are called cache pages or slots. This unit of allocation (which is smaller than the values that are used in other storage systems) ensures that small I/Os do not waste cache memory. The decision to copy data into the DS8900F cache can be triggered from the following policies:

- Demand paging

  Eight disk blocks (a 4 K cache page) are brought in only on a cache miss. Demand paging is always active for all volumes and ensures that I/O patterns with locality discover at least recently used (LRU) data in the cache.

- Prefetching

  Data is copied into the cache speculatively even before it is requested. To prefetch, a prediction of likely data accesses is needed. Because effective, sophisticated prediction schemes need an extensive history of page accesses (which is not feasible in real systems), SARC uses prefetching for sequential workloads. Sequential access patterns naturally arise in video-on-demand, database scans, copy, backup, and recovery. The goal of sequential prefetching is to detect sequential access and effectively prefetch the likely cache data to minimize cache misses. Today, prefetching is ubiquitously applied in web servers and clients, databases, file servers, on-disk caches, and multimedia servers.

For prefetching, the cache management uses tracks. A track is a set of 128 disk blocks (16 cache pages). To detect a sequential access pattern, counters are maintained with every track to record whether a track was accessed together with its predecessor. Sequential prefetching becomes active only when these counters suggest a sequential access pattern.

In this manner, the DS8900F monitors application read I/O patterns and dynamically determines whether it is optimal to stage into cache the following I/O elements:

- Only the page requested.
- The page that is requested plus the remaining data on the disk track.
- An entire disk track (or a set of disk tracks) that was not requested.

The decision of when and what to prefetch is made in accordance with the Adaptive Multi-stream Prefetching (AMP) algorithm. This algorithm dynamically adapts the number and timing of prefetches optimally on a per-application basis (rather than a system-wide basis). For more information about AMP, see "Adaptive Multi-stream Prefetching" on page 19.

To decide which pages are evicted when the cache is full, sequential and random (non-sequential) data is separated into separate lists. The SARC algorithm for random and sequential data is shown in Figure 2-2.

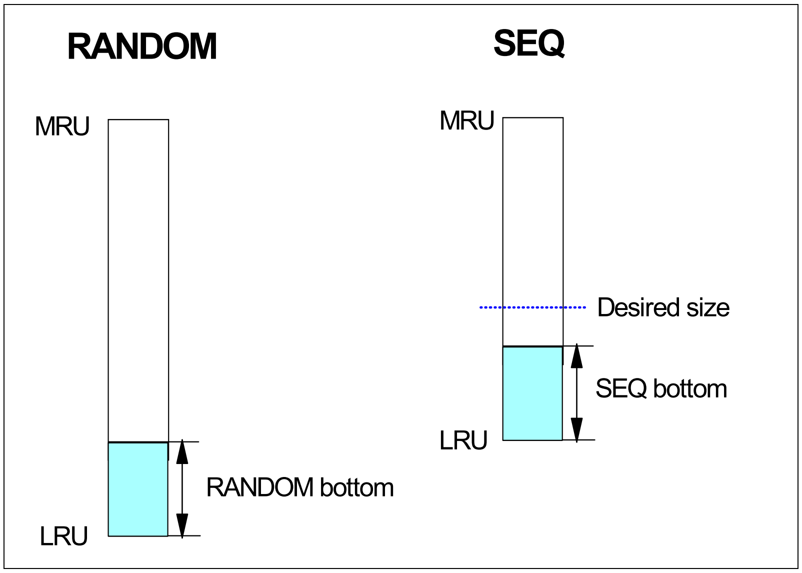


Figure 2-2 Sequential Adaptive Replacement Cache

A page that was brought into the cache by simple demand paging is added to the head of the Most Recently Used (MRU) section of the RANDOM list. Without further I/O access, it goes down to the bottom of LRU section. A page that was brought into the cache by a sequential access or by sequential prefetching is added to the head of MRU of the SEQ list and then goes in that list. Other rules control the migration of pages between the lists so that the system does not keep the same pages in memory twice.

Adaptive Multi-stream Prefetching

SARC dynamically divides the cache between the RANDOM and SEQ lists, where the SEQ list maintains pages that are brought into the cache by sequential access or sequential prefetching.

In the DS8900F, AMP, an algorithm that was developed by IBM Research, manages the SEQ list. AMP is an autonomic, workload-responsive, and self-optimizing prefetching technology that adapts the amount of prefetch and the timing of prefetch on a per-application basis to maximize the performance of the system. The AMP algorithm solves the following problems that plague most other prefetching algorithms:

- Prefetch wastage occurs when prefetched data is evicted from the cache before it can be used.
- Cache pollution occurs when less useful data is prefetched instead of more useful data.

By carefully choosing the prefetching parameters, AMP provides optimal sequential read performance and maximizes the aggregate sequential read throughput of the system. The amount that is prefetched for each stream is dynamically adapted according to the application's needs and the space that is available in the SEQ list. The timing of the prefetches is also continuously adapted for each stream to avoid misses and any cache pollution.

SARC and AMP play complementary roles. SARC carefully divides the cache between the RANDOM and the SEQ lists to maximize the overall hit ratio. AMP manages the contents of the SEQ list to maximize the throughput that is obtained for the sequential workloads. SARC affects cases that involve both random and sequential workloads. However, AMP helps any workload that has a sequential read component, including pure sequential read workloads.

AMP dramatically improves performance for common sequential and batch processing workloads. It also provides excellent performance synergy with Db2 by preventing table scans from being I/O-bound and improves performance of index scans and Db2 utilities, such as Copy and Recover. Furthermore, AMP reduces the potential for array hot spots that result from extreme sequential workload demands

Intelligent Write Caching

Another cache algorithm, which is referred to as Intelligent Write Caching (IWC), was implemented in the DS8900F series. IWC improves performance through better write cache management and a better destaging order of writes. This algorithm is a combination of CLOCK, a predominantly read cache algorithm, and CSCAN, an efficient write cache algorithm. Out of this combination, IBM produced a powerful and widely applicable write cache algorithm.

The CLOCK algorithm uses temporal ordering. It keeps a circular list of pages in memory, with the clock hand that points to the oldest page in the list. When a page must be inserted in the cache, then an R (recency) bit is inspected at the clock hand's location. If R is zero, the new page is put in place of the page the clock hand points to and R is set to 1. Otherwise, the R bit is cleared and set to zero. Then, the clock hand moves one step clockwise forward and the process is repeated until a page is replaced.

The CSCAN algorithm uses spatial ordering. The CSCAN algorithm is the circular variation of the SCAN algorithm. The SCAN algorithm tries to minimize the disk head movement when the disk head services read and write requests. It maintains a sorted list of pending requests with the position on the drive of the request. Requests are processed in the current direction of the disk head until it reaches the edge of the disk. At that point, the direction changes. In the CSCAN algorithm, the requests are always served in the same direction. After the head arrives at the outer edge of the disk, it returns to the beginning of the disk and services the new requests in this one direction only. This process results in more equal performance for all head positions.

The basic idea of IWC is to maintain a sorted list of write groups, as in the CSCAN algorithm. The smallest and the highest write groups are joined, forming a circular queue. The new idea is to maintain a recency bit for each write group, as in the CLOCK algorithm. A write group is always inserted in its correct sorted position and the recency bit is set to zero at the beginning. When a write hit occurs, the recency bit is set to one.

The destage operation proceeds, where a destage pointer is maintained that scans the circular list and looks for destage victims. Then, this algorithm allows destaging of only write groups whose recency bit is zero. The write groups with a recency bit of one are skipped and the recent bit is then turned off and reset to zero. This process gives an extra life to those write groups that were hit since the last time the destage pointer visited them. The concept of this mechanism is shown in Figure 5-3 on page 150.

In the DS8900F implementation, an IWC list is maintained for each rank. The dynamically adapted size of each IWC list is based on workload intensity on each rank. The rate of destage is proportional to the portion of NVS that is occupied by an IWC list. The NVS is shared across all ranks in a cluster. Furthermore, destages are smoothed out so that write bursts are not converted into destage bursts.

Another enhancement to IWC is an update to the cache algorithm that increases the residency time of data in NVS. This improvement focuses on maximizing throughput with excellent average response time (latency).

In summary, IWC has better or comparable peak throughput to the best of CSCAN and CLOCK across a wide variety of write cache sizes and workload configurations. In addition, even at lower throughputs, IWC has lower average response times than CSCAN and CLOCK.

List Prefetch Optimizer

The Adaptive List Prefetch (ALP) algorithm, which is also known as the LPO, enables prefetch of a list of non-sequential tracks, providing improved performance for Db2 workloads. ALP, when working together with AMP, dramatically improves the performance of record identifier (RID) list scans. These features are an essential part of Data Warehouse and Business Intelligence workloads, and help analytics applications benefit from being run on a DS8900F.

2.3 Host adapters

Host Adapters (HAs) comprise Layer 1 of the DS8000 Three Layer processor architecture which is illustrated in Figure 2-1 on page 16. They operate on the “front-end” of the DS8000s. Host Adapters, which are based on IBM PowerPC® processor technology, serve two functions. One, they provide I/O connectivity between the host server and the DS8000 storage system for reads and writes. Two, they carry synchronous (Metro Mirror) and asynchronous (Global Mirror and Global Copy) replication between source and target DS8000s. Host I/O can be either FICON or Fibre Channel Protocol (FCP); DS8000 replication between DS8000s is via FCP.

The DS8900F models offer both 16 Gbps and 32 Gbps four-port Fibre Channel/IBM FICON Host Adapters. Each adapter can auto-negotiate up to two generations back. That is the fiber channel standard for backwards compatibility and is known as N - 2 compatibility. The 16 Gbps adapter can auto-negotiate down to 8 Gbps and 4 Gbps. Similarly, the 32 Gbps adapter can auto negotiate down to 16 Gbps and 8 Gbps. Both HAs are available with either all short wave (SW) or all long wave (LW) SFPs. An individual HA does not offer an intermix of SW and LW nor an intermix of 16 Gbps and 32 Gbps ports. But, a DS8900F can have an intermix of SW and LW HAs running at either 16 Gbps or 32 Gbps.

With flexibility like that, you can benefit from the higher performance while maintaining compatibility with existing host processor and SAN infrastructures. Additionally, you can configure ports individually (and dynamically) which enables FICON and FCP intermix on the same adapter. It is also possible to scale-up host adapter connectivity within the first frame (all DS8900F models) or into a second frame (models DS8950F and DS8980F) without adding more flash capacity, unless the workload requires it.The following lists the minimum and maximum host adapters by model:

- DS8910F - Flexibility Class
  - 2 to 8 - model 993
  - 2 to 16 - model 994
- DS8950F - Agility Class (model 996)
  - 2 to 16 - for a single frame system
  - 2 to 32 - for a two-frame system
- DS8980F - Analytic Class (model 998)
  - 2 to 16 - for a single frame system
  - 2 to 32 - for a two-frame system

For more information on DS8000 Host Adapter best practices, see “DS8000 Host Adapter Configuration Guidelines”, created and maintained by IBM’s Advanced Technology Group. It is available at:

https://www.ibm.com/support/pages/node/6354509

2.4 zHyperLinks

zHyperLink is a point-to-point optical connection designed to provide ultra-low response times (latency) between IBM Z servers (z14 and later) and DS8900F (and DS8880) storage systems. It works alongside the FICON infrastructure whether its direct point-to-point connectivity or via a storage area network (SAN); zHyperLink does not replace FICON connectivity. It reduces I/O latency, for selected eligible I/Os, to the point that z/OS no longer needs to undispatch the running task. zHyperLink enables response times (latency) as low a ~20 μ seconds (micro seconds) and can be enabled for both read and write I/Os.

zHyperLinks are typically added in pairs for redundancy and have a point-to-point distance limitation of 150 m (492 feet) between the System Z and DS8000 storage system. The shorter the route the lower the latency added by the physical infrastructure. It requires specific zHyperlink adapters on System Z, zHyperLink cabling, and zHyperLink adapters on the DS8900F (and DS8880). For two-frame DS8900F (and DS8880) storage systems we recommend spreading the zHyperLinks evenly, or as evenly as possible, across the frames.

zHyperLink adapters by DS8900F model:

- DS8910F - Flexibility Class (models 993 & 994)
  - 0 to 4 zHyperLinks
- DS8950F - Agility Class (model 996)
  - 0 to 6 zHyperLinks for a single frame system with 20 cores
  - 0 to 8 zHyperLinks for a single frame system with 40 cores
  - 0 to 12 zHyperLinks for a two-frame system
- DS8980F - Analytic Class (model 998)
  - 0 to 8 zHyperLinks for a single frame system
  - 0 to 12 zHyperLinks for a two-frame system

zHyperLink is one of the many Z Synergy features on the DS8900Fs. For more information about zHyperLink, see *IBM DS8000 and IBM Z Synergy*, REDP-5186 at:

https://www.redbooks.ibm.com/redpapers/pdfs/redp5186.pdf

2.5 High-Performance Flash Enclosure (HPFE) Gen-2

A DS8000 High-Performance Flash Enclosure - Gen2 (HPFE) is a 2U storage enclosure that houses a redundant pair of Device Adapters and one, two, or three flash drivesets. HPFEs are part of the physical “back-end” infrastructure of a DS8000. The number of HPFEs installed directly impacts the DS8000’s physical capacity (scalability) of flash. The following lists the minimum and maximum HPFEs by DS8900F model:

- DS8910F - Flexibility Class
  - 1 - model 993 (when installed in an IBM Z or LinuxONE rack)
  - 1 to 2 - model 993 (when installed in a customer provided rack)
  - 1 to 4 - model 994
- DS8950F - Agility Class (model 996)
  - 1 to 4 - for a single frame system
  - 1 to 8 - for a two-frame system
- DS8980F - Analytic Class (model 998)
  - 1 to 4 - for a single frame system
  - 1 to 8 - for a two-frame system

The DS8900F models, like the predecessor DS8880 all flash array models, can be configured with custom flash drive placement (custom placement) to scale the “back-end” or internal bandwidth. There are some workloads that would benefit from more Device Adapters (which provide the internal bandwidth), but do not require additional storage capacity. In other words, it is not necessary to add additional flash drivesets in order to add more device adapters. Two examples of custom placement are shown below to introduce the concept. More elaborate configurations are possible.

The first example shows a DS8910F, on the left, configured with 2 x 3.84 TB HCF Tier 1 flash drivesets installed in 1 x HPFE. The DS8910F, on the right in the same example, is configured with 4 x 1.92 TB HCF Tier 2 flash drivesets, spread evenly across 2 x HPFEs via custom placement. For reference see Figure 2-3 on page 24.

Both configurations have the same usable capacity (~73.35 TiBs). The differences are noted on the diagram and restated here. The DS8910F on the left has flash drives with a higher performance profile (HCF Tier 1). The DS8910F on the right has flash drives with a lower performance profile (HCF Tier 2), but twice the internal bandwidth due to the 2 x HPFEs which add a second set of device adapters. Both may be suitable for the same workload. In some cases the higher performing drives may be required; in other cases the additional internal bandwidth may be the deciding factor. Or, physical space limitations may ultimately be the determining factor which is to say, in this example, the configuration on the left may be required.

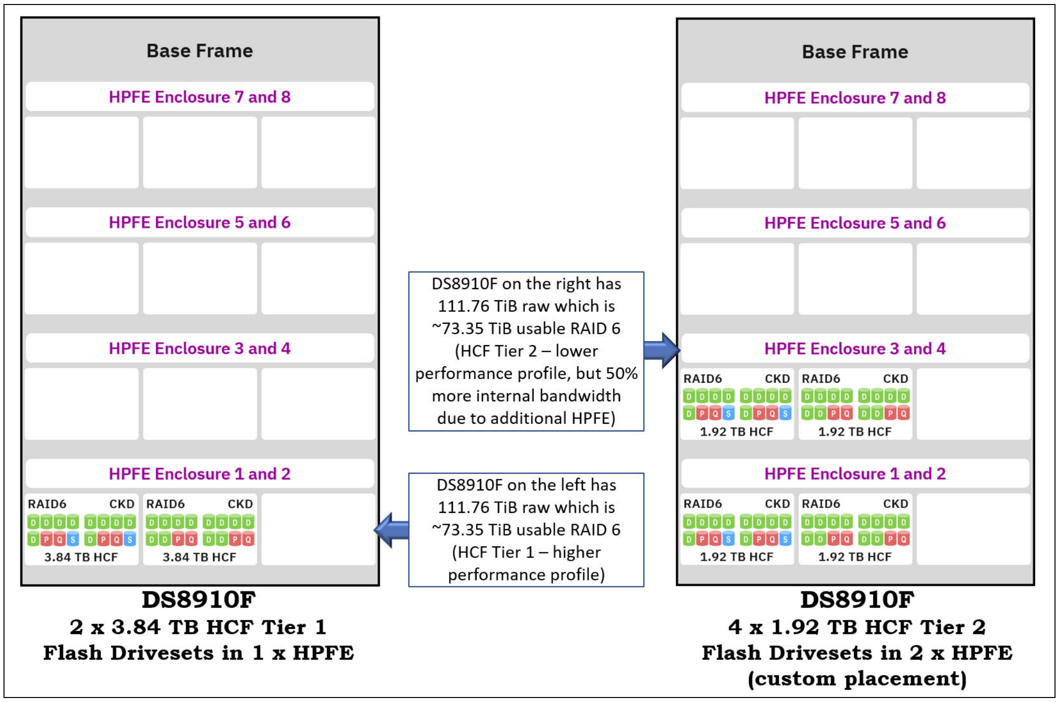

Base Frame
HPFE Enclosure 7 and 8
HPFE Enclosure 5 and 6
HPFE Enclosure 3 and 4
HPFE Enclosure 1 and 2
RAID6
CKD
3.84 TB HCF
3.84 TB HCF
DS8910F
2 x 3.84 TB HCF Tier 1
Flash Drivesets in 1 x HPFE
DS8910F on the right has 111.76 TiB raw which is ~73.35 TiB usable RAID 6 (HCF Tier 2 – lower performance profile, but 50% more internal bandwidth due to additional HPFE)
DS8910F on the left has 111.76 TiB raw which is ~73.35 TiB usable RAID 6 (HCF Tier 1 – higher performance profile)
Base Frame
HPFE Enclosure 7 and 8
HPFE Enclosure 5 and 6
HPFE Enclosure 3 and 4
HPFE Enclosure 1 and 2
RAID6
CKD
1.92 TB HCF
1.92 TB HCF
DS8910F
4 x 1.92 TB HCF Tier 2
Flash Drivesets in 2 x HPFE
(custom placement)

Figure 2-3 DS8910Fs with custom placement - Example 1 - 3.84 TBs vs 1.92 TBs (1 vs 2 HPFEs)

The second example shows a DS8950F, on the left, configured with 6 x 3.84 TB HCF Tier 1 flash drivesets installed in 2 x HPFEs. On the right in this example, the DS8950F is also configured with 6 x 3.84 TB HCF Tier 1 flash drivesets, but they are spread evenly across 3 x HPFEs via custom placement. That provides 50% additional internal bandwidth due to an additional pair of device adapters. For reference see Figure 2-4 on page 25.

The two DS8950Fs are configured with the same number of 3.84 TB HCF Tier 1 flash drivesets, but there is a slight difference in usable capacity. As noted on the diagram, the DS8950F on the left has ~227 Tebibyte (TiB) usable capacity, but the DS8950F on the right has ~220 TiB because it has 2 x additional spare drives (see blue disk icons labeled with an S). Both may be capable of delivering the same response time (latency) for a particular workload. If there is a significant sequential workload, requiring more internal bandwidth, then the configuration on the right may be more appropriate.

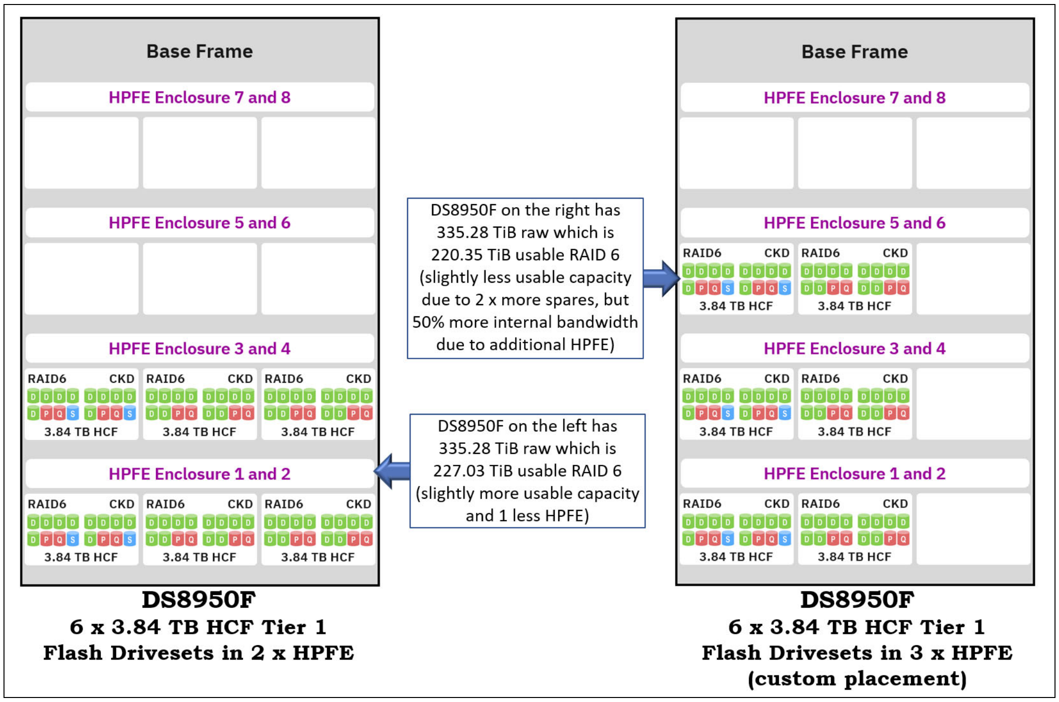

Base Frame
HPFE Enclosure 7 and 8
HPFE Enclosure 5 and 6
HPFE Enclosure 3 and 4
HPFE Enclosure 1 and 2
RAID6
CKD
3.84 TB HCF
DS8950F on the right has 335.28 TiB raw which is 220.35 TiB usable RAID 6 (slightly less usable capacity due to 2 x more spares, but 50% more internal bandwidth due to additional HPFE)
DS8950F on the left has 335.28 TiB raw which is 227.03 TiB usable RAID 6 (slightly more usable capacity and 1 less HPFE)
DS8950F
6 x 3.84 TB HCF Tier 1
Flash Drivesets in 2 x HPFE
DS8950F
6 x 3.84 TB HCF Tier 1
Flash Drivesets in 3 x HPFE
(custom placement)

Figure 2-4 DS8950Fs with custom placement - Example 2 - 6 x 3.84 TB drivesets in 2 vs 3 HPFEs

Device Adapters and Flash Tiers are discussed in the sections that follow immediately below.

For more information about HPFEs, see *IBM DS8000 High-Performance Flash Enclosure Gen2*, REDP-5422 at:

https://www.redbooks.ibm.com/redpapers/pdfs/redp5422.pdf

2.5.1 Device Adapters

Device Adapters (DAs), also known as “Flash RAID adapters”, comprise “Layer 3” of the DS8000 “Three Layer” processor architecture illustrated in Figure 2-1 on page 16. Like host adapters, Flash RAID adapters, or DAs, are based on IBM PowerPC technology. And, they too, serve two functions, but they operate on the “back-end” of the DS8000. One, they manage RAID protection and sparing for the flash arrays or ranks. Two, they manage read and write I/Os to and from those ranks, respectively.

2.5.2 Flash Tiers

The three tiers of flash drives available on the DS8900Fs have their own respective performance profile. All three DS8900F models offer the same set of flash choices. That means that there are a variety of possible configurations that can satisfy both performance and capacity requirements. Typically, DS8900Fs are configured with just one or two flash tiers. Depending on the particular workload, that may mean a single tier of high performance flash or high capacity flash or a combination of high performance and high capacity flash.

The goal is to configure a storage system that meets the performance and capacity requirements *effectively and efficiently*. In other words, achieve a balanced price for performance configuration. The best way to determine the appropriate flash tier or tiers is to create a performance model of the actual workload. For details on performance modeling, see “IBM Storage Modeller” on page 126.

2.5.3 Multiple paths to Open Systems servers

For high availability, each host system must use a multipathing device driver and each host system must have a minimum of two host connections to HA cards in different I/O enclosures on the DS8880 storage system.

After you determine your throughput workload requirements, you must choose the appropriate number of connections to put between your Open Systems hosts and the DS8000 storage system to sustain this throughput. Use an appropriate number of HA cards to satisfy high throughput demands. The number of host connections per host system is primarily determined by the required bandwidth.

Host connections frequently go through various external connections between the server and the DS8000 storage system. Therefore, you need enough host connections for each server so that if half of the connections fail, processing can continue at the level before the failure. This availability-oriented approach requires that each connection carry only half the data traffic that it otherwise might carry. These multiple lightly loaded connections also help to minimize the instances when spikes in activity might cause bottlenecks at the HA or port. A multiple-path environment requires at least two connections. Four connections are typical, and eight connections are not unusual. Typically, these connections are spread across as many I/O enclosures in the DS8880 storage system that you equipped with HAs.

Usually, SAN directors or switches are used. Use two separate switches to avoid a single point of failure.

2.5.4 Multiple paths to z Systems servers

In the z Systems environment, the preferred practice is to provide multiple paths from each host to a storage system. Typically, four paths are installed. The channels in each host that can access each Logical Control Unit (LCU) in the DS8000 storage system are defined in the hardware configuration definition (HCD) or I/O configuration data set (IOCDS) for that host. Dynamic Path Selection (DPS) allows the channel subsystem to select any available (non-busy) path to initiate an operation to the storage system. Dynamic Path Reconnect (DPR) allows the DS8000 storage system to select any available path to a host to reconnect and resume a disconnected operation, for example, to transfer data after disconnection because of a cache miss. These functions are part of the z Systems architecture and are managed by the channel subsystem on the host and the DS8000 storage system.

In a z Systems environment, you must select a SAN switch or director that also supports FICON. An availability-oriented approach applies to the z Systems environments similar to the Open Systems approach. Plan enough host connections for each server so that if half of the connections fail, processing can continue at the level before the failure.

2.5.5 Spreading host attachments

A common question is how to distribute the server connections. Take the example of four host connections from each of four servers, all running a similar type workload, and four HAs. Spread the host connections with each host attached to one port on each of four adapters. Now, consider the scenario in which the workloads differ. You probably want to isolate mission-critical workloads (for example, customer order processing) from workloads that are lower priority but I/O intensive (for example, data mining), or prevent the I/O-intensive workload from dominating the HA. If one of the four servers is running an I/O-intensive workload, acquire two additional HAs and attach the four connections of the I/O-intensive server to these adapters. Connect two host connections on each adapter. The other three servers remain attached to the original four adapters.

Here are some general guidelines:

- Spread the paths from all host systems across the available I/O ports, HA cards, and I/O enclosures to optimize workload distribution across the available resources, depending on your workload sharing and isolation considerations.
- Spread the host paths that access the same set of volumes as evenly as possible across the available HA cards and I/O enclosures. This approach balances workload across hardware resources and helps ensure that a hardware failure does not result in a loss of access.
- Ensure that each host system uses a multipathing device driver, such as an SDD, and a minimum of two host connections to different HA cards in different I/O enclosures on the DS8880 storage system. Preferably, evenly distribute them between left-side (even-numbered) I/O enclosures and right-side (odd-numbered) I/O enclosures for the highest availability and a balanced workload across I/O enclosures and HA cards.
- Do not use the same DS8000 I/O port for host attachment and Copy Services remote replication (such as Metro Mirror or Global Mirror). Ideally, even separate off HA pairs (with all their ports) for Copy Services traffic only if enough adapters are available.
- Consider using separate HA cards for the FICON protocol and FCP. Even though I/O ports on the same HA can be configured independently for the FCP and the FICON protocol, it might be preferable to isolate your z/OS environment (FICON) from your Open Systems environment (FCP).
- Do not use the 8-port (8 Gbps) HAs with all their ports when you have high throughputs (GBps) to handle.

For more information, see 4.10.1, “Fibre Channel host adapters: 16 Gbps and 32 Gbps” on page 100.

2.6 Tools to aid in hardware planning

This chapter described the hardware components of the DS8000 storage system. Each component provides high performance through its specific function and meshes well with the other hardware components to provide a well-balanced, high-performance storage system. There are many tools that can assist you in planning your specific hardware configuration.

2.6.1 White papers

IBM publishes white papers that document the performance of specific DS8000 configurations. Typically, workloads are run on multiple configurations, and performance results are compiled so that you can compare the configurations. For example, workloads can be run by using different DS8000 models, different numbers of HAs, or different types of flash. By reviewing these white papers, you can make inferences about the relative performance benefits of different components to help you to choose the type and quantities of components to fit your particular workload requirements. A selection of these papers can be found at:

https://www.ibm.com/support/pages/ibm-techdocs-technical-sales-library

Your IBM representative or IBM Business Partner has access to these and additional white papers and can provide them to you.

2.6.2 Cache and I/O operations

Caching is a fundamental technique for avoiding (or reducing) I/O latency. Cache is used to keep both the data read and written by the host servers. The host does not need to wait for the flash media to either obtain or store the data that is needed because cache can be used as an intermediate repository. Prefetching data from flash media for read operations, and operations of writing to the flash media, are done by the DS8000 storage system asynchronously from the host I/O processing.

Cache processing improves the performance of the I/O operations that are done by the host systems that attach to the DS8000 storage system. Cache size, the efficient internal structure, and algorithms of the DS8000 storage system are factors that improve I/O performance. The significance of this benefit is determined by the type of workload that is run.

Read operations

These operations occur when a host sends a read request to the DS8000 storage system:

- A cache hit occurs if the requested data is in the cache. In this case, the I/O operation does not disconnect from the channel/bus until the read is complete. A read hit provides the highest performance.
- A cache miss occurs if the data is not in the cache. The I/O operation logically disconnects from the host. Other I/O operations occur over the same interface. A stage operation from the flash MEDIA back-end occurs.

Write operations: Fast writes

A fast write hit occurs when the write I/O operation completes when the data that is received from the host is transferred to the cache and a copy is made in persistent memory. Data that is written to a DS8000 storage system is almost 100% fast-write hits. The host is notified that the I/O operation is complete when the data is stored in the two locations.

The data remains in the cache and persistent memory until it is destaged, at which point it is flushed from cache. Destage operations of sequential write operations to RAID 5 arrays are done in parallel mode, writing a stripe to all disks in the RAID set as a single operation. An entire stripe of data is written across all the disks in the RAID array. The parity is generated one time for all the data simultaneously and written to the parity disk. This approach reduces the parity generation penalty that is associated with write operations to RAID 5 arrays. For RAID 6, data is striped on a block level across a set of drives, similar to RAID 5 configurations. A second set of parity is calculated and written across all the drives. This technique does not apply for the RAID 10 arrays because there is no parity generation that is required. Therefore, no penalty is involved other than a double write when writing to RAID 10 arrays.

It is possible that the DS8000 storage system cannot copy write data to the persistent cache because it is full, which can occur if all data in the persistent cache waits for destage to disk. In this case, instead of a fast write hit, the DS8000 storage system sends a command to the host to retry the write operation. Having full persistent cache is not a good situation because it delays all write operations. On the DS8000 storage system, the amount of persistent cache is sized according to the total amount of system memory. The amount of persistent cache is designed so that the probability of full persistent cache occurring in normal processing is low.

2.6.3 Determining the correct amount of cache storage

A common question is “How much cache do I need in my DS8000 storage system?” Unfortunately, there is no quick and easy answer.

There are a number of factors that influence cache requirements:

- Is the workload sequential or random?
- Are the attached host servers z Systems or Open Systems?
- What is the mix of reads to writes?
- What is the probability that data is needed again after its initial access?
- Is the workload cache friendly (a cache-friendly workload performs better with relatively large amounts of cache)?

It is a common approach to base the amount of cache on the amount of disk capacity, as shown in these common general rules:

- For Open Systems, each TB of drive capacity needs 1 GB - 2 GB of cache.
- For z Systems, each TB of drive capacity needs 2 GB - 4 GB of cache.

For SAN Volume Controller attachments, consider the SAN Volume Controller node cache in this calculation, which might lead to a slightly smaller DS8000 cache. However, most installations come with a minimum of 128 GB of DS8000 cache. Using flash in the DS8000 storage system does not typically change the prior values. HPFE with its flash media are beneficial with cache-unfriendly workload profiles because they reduce the cost of cache misses.

Most storage servers support a mix of workloads. These general rules can work well, but many times they do not. Use them like any general rule, but only if you have no other information on which to base your selection.

When coming from an existing disk storage server environment and you intend to consolidate this environment into DS8000 storage systems, follow these recommendations:

- Choose a cache size for the DS8000 series that has a similar ratio between cache size and disk storage to that of the configuration that you use.
- When you consolidate multiple disk storage servers, configure the sum of all cache from the source disk storage servers for the target DS8000 processor memory or cache size.

For example, consider replacing four DS8880 storage systems, each with 200 TB and 512 GB cache, with a single DS8950F storage system. The ratio between cache size and disk storage for each DS8880 storage system is 0.25% (512 GB/200 TB). The new DS8950F storage system is configured with 900 TB to consolidate the four 200 TB DS8880 storage systems, plus provide some capacity for growth. This DS8950F storage system should be fine with 2 TB of cache to keep mainly the original cache-to-disk storage ratio. If the requirements are somewhere in between or in doubt, round up to the next available memory size. When using a SAN Volume Controller in front, round down in some cases for the DS8000 cache.

The cache size is not an isolated factor when estimating the overall DS8000 performance. Consider it with the DS8000 model, the capacity and speed of the disk drives, and the number and type of HAs. Larger cache sizes mean that more reads are satisfied from the cache, which reduces the load on DAs and the disk drive back end that is associated with reading data from disk. To see the effects of different amounts of cache on the performance of the DS8000 storage system, run a StorM or Disk Magic model, which is described in Chapter 6.1, “IBM Storage Modeller” on page 126.

3

Logical configuration concepts and terminology

This chapter summarizes the important concepts that must be understood for the preparation of the DS8000 logical configuration and about performance tuning. You can obtain more information about the DS8000 logical configuration concepts in *IBM DS8900F Architecture and Implementation*, SG24-8456.

This chapter includes the following topics:

- The abstraction layers for logical configuration
  - ArraySites
  - Arrays
  - Ranks
  - Extent Pools
  - Logical Volumes
  - Easy Tier
- Data placement on ranks and extent pools

3.1 Logical Configuration

Logical configuration refers to the creation of Redundant Array of Independent Disks (RAID) arrays and pools, and the assignment of the configured capacity to servers. The initial logical configuration and all subsequent modifications are client responsibilities.The logical configuration can be created, applied, and modified by using the Data Storage Graphical User Interface (DS GUI), Data Storage Command-inline Interface (DS CLI), or DS Open application programming interface (DS Open API).

DS8900F uses the process of virtualization which is the abstraction of a physical drive to one or more logical volumes. This virtual drive is presented to hosts and systems as though it is a physical drive. This process allows the host to think that it is using a storage device that belongs to it, but the device is implemented in the storage system.

From a physical point of view, the DS8900F offers all flash storage. Flash Storage has no moving parts and a lower energy consumption. The performance advantages are the fast seek time and the average access time. Within flash, there is a wide choice between high-performing Flash Tier 0 and the more economical Flash Tier 1 and 2 drives. In a combination, these tiers provide a solution that is optimized for both performance and cost.

Virtualization builds upon the flash physical drives as a series of layers:

- Array Sites
- Arrays
- Ranks
- Extent Pools
- Logical Volumes
- LSSs

3.1.1 Array Sites

An array site is formed from a group of eight identical drives with the same capacity, speed, and drive class. The specific drives that are assigned to an array site are automatically chosen by the system at installation time to balance the array site capacity across both drive enclosures in a pair and across the connections to both storage servers. The array site also determines the drives that are required to be reserved as spares. No predetermined storage server affinity exists for array sites.

Array sites are the building blocks that are used to define arrays. See Figure 3-1 on page 33.

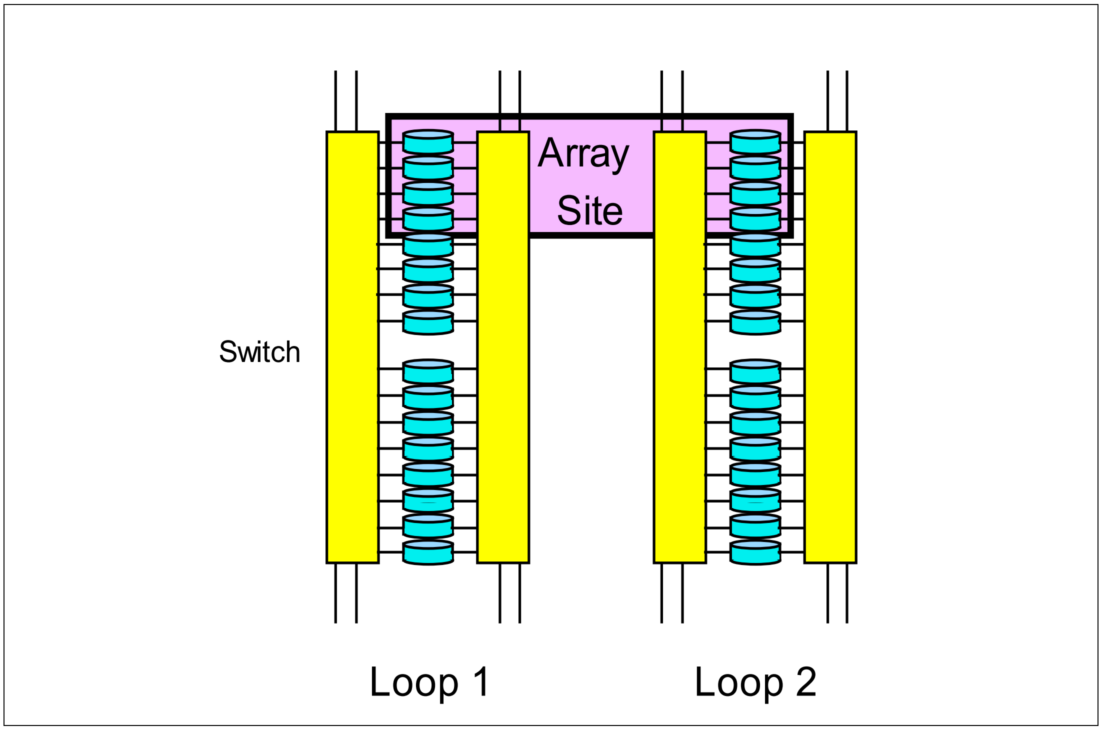


Figure 3-1 Array Site

3.1.2 Arrays

An array is created from one array site. When an array is created, its RAID level, array type, and array configuration are defined. This process is also called defining an array. In all IBM DS8000 series implementations, one array is always defined as using one array site.

The following RAID levels are supported in DS8900F:

- RAID 6 (default)
- RAID 10
- RAID 5 (possible for flash drive sizes below 1 TB)

A request for price quotation (RPQ) is required for flash drives greater than 1 TB (RPQ is not available for drive sizes of 4 TB and over).

Each HPFE Gen2 pair can contain up to six array sites. The first set of 16 flash drives creates two 8-flash-drive array sites. RAID 6 arrays are created by default on each array site. RAID 5 is optional for flash drives smaller than 1 TB, but is not recommended. RAID 10 is optional for all flash drive sizes.

During logical configuration, RAID 6 arrays and the required number of spares are created. Each HPFE Gen2 pair has two global spares that are created from the first increment of 16 flash drives. The first two arrays to be created from these array sites are 5+P+Q. Subsequent RAID 6 arrays in the same HPFE Gen2 Pair are 6+P+Q.

> **Important:** Using RAID 6 is recommended, and it is the default in the DS Graphical User Interface (DS GUI). As with large drives in particular, the RAID rebuild times (after one drive failure) get ever larger. Using RAID 6 reduces the danger of data loss due to a double-drive failure.

Tip: Larger drives have a longer rebuild time. Only RAID 6 can recover from a second error during a rebuild due to the extra parity. RAID 6 is the best choice for systems that require high availability (HA), so it is the default.

Smart Rebuild

Smart Rebuild is a function that is designed to help reduce the possibility of secondary failures and data loss of RAID arrays. It can be used to rebuild a RAID 6 array when certain drive errors occur and a normal determination is made that it is time to use a spare to proactively replace a failing flash drive. If the suspect drive is still available for I/O, it is kept in the array rather than being rejected as under a standard RAID rebuild. A spare is brought into the array, as an extra member, concurrently.

Note: Smart Rebuild for flash drives is available for DS8900F. It is available for RAID 6 arrays only.

Smart Rebuild is not applicable in all situations, so it is not always used. Smart Rebuild runs only for healthy RAID 6 arrays. If two drives with errors are in a RAID 6 configuration, or if the drive mechanism failed to the point that it cannot accept any I/O, the standard RAID rebuild procedure is used for the RAID array. If communications across a drive fabric are compromised, such as an SAS path link error that causes the drive to be bypassed, standard RAID rebuild procedures are used because the suspect drive is not available for a one-to-one copy with a spare. If Smart Rebuild is not possible or does not provide the designed benefits, a standard RAID rebuild occurs.

Example 3-1 DS8900F Arrays created from ArraySites

```
dscli> lsarraysite -l
arsite DA Pair dkcap (10^9B) diskrpm State    Array diskclass  encrypt
========================================================================
S1     10             7680.0   65000 Assigned A3    FlashTier2 supported
S2     10             7680.0   65000 Assigned A1    FlashTier2 supported
S3     11             1600.0   65000 Assigned A4    FlashTier0 supported
S4     11             1600.0   65000 Assigned A5    FlashTier0 supported
S5     11             1600.0   65000 Assigned A2    FlashTier0 supported
S6     11             1600.0   65000 Assigned A0    FlashTier0 supported

dscli> lsarray -l
Array State    Data   RAIDtype    arsite Rank DA Pair DDMcap (10^9B) diskclass  encrypt
=========================================================================================
A0    Assigned Normal 6 (6+P+Q)   S6     R0   11              1600.0 FlashTier0 supported
A1    Assigned Normal 6 (5+P+Q+S) S2     R1   10              7680.0 FlashTier2 supported
A2    Assigned Normal 6 (6+P+Q)   S5     R2   11              1600.0 FlashTier0 supported
A3    Assigned Normal 6 (5+P+Q+S) S1     R3   10              7680.0 FlashTier2 supported
A4    Assigned Normal 6 (5+P+Q+S) S3     R4   11              1600.0 FlashTier0 supported
A5    Assigned Normal 6 (5+P+Q+S) S4     R5   11              1600.0 FlashTier0 supported
```

In Example 3-1 we have a DS8980F with 6 Array Sites:

- 2 × High-Capacity Flash of 7.68 TB (FlashTier2)
- 4 × High-Performance Flash of 1.6 TB (FlashTier0)

Using the DS CLI command mkarray or through the DS GUI interface, 6 Arrays (1 Arraysite <-> 1 Array) were created:

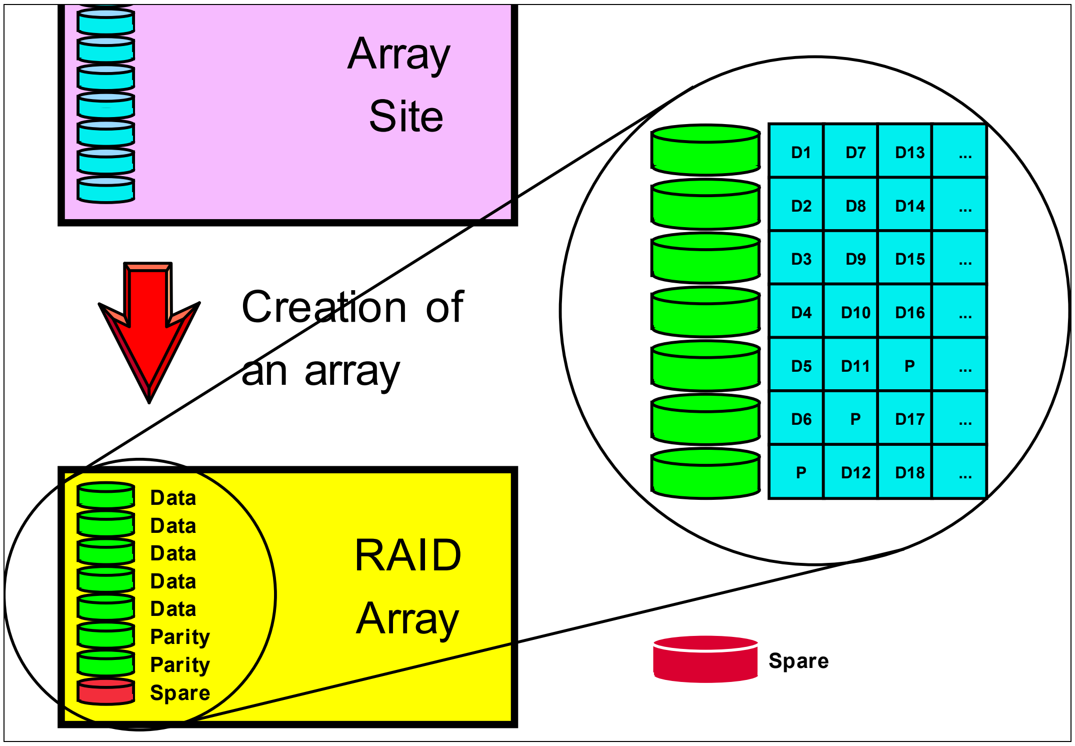
- A3 <- S1 (5+P+Q+S)

- A1 <- S2 (5+P+Q+S)
- A4 <- S3 (5+P+Q+S)
- A5 <- S4 (5+P+Q+S)
- A2 <- S5 (6+P+Q)
- A0 <- S6 (6+P+Q)

RAID 6 (default) was used to create the Arrays, and as result we have the following:

- 5+P+Q+S RAID 6 configuration: The array consists of five data drives and two parity drives (P and Q). The remaining drive on the array site is used as a spare (S).
- 6+P+Q RAID 6 configuration: The array consists of six data drives and two parity drives

See Figure 3-2 on page 35 to see how a RAID 6 is created from an Array Site and Figure 3-3 on page 36 for RAID 5, RAID 6 and RAID 10 Array types and resulting digit configurations.



D1	D7	D13	...
D2	D8	D14	...
D3	D9	D15	...
D4	D10	D16	...
D5	D11	P	...
D6	P	D17	...
P	D12	D18	...

Figure 3-2 Creation of a RAID 6 Array

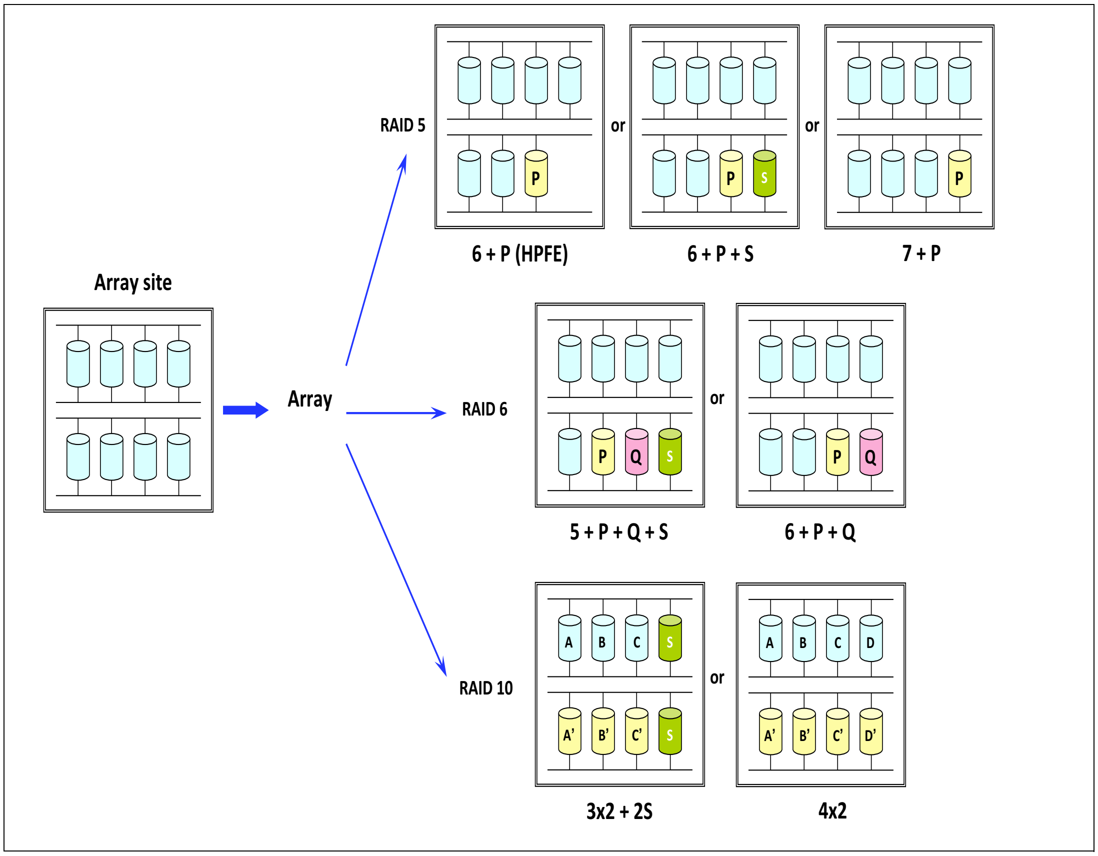


Figure 3-3 RAID Array types

3.1.3 Ranks

After the arrays are created, the next task is to define a rank. A rank is a logical representation of the physical array that is formatted for use as FB or CKD storage types. In the DS8900F, ranks are defined in a one-to-one relationship to arrays.

The process of defining a rank accomplishes the following objectives:

- The array is formatted for FB data for open systems or CKD for IBM Z data.
- The capacity of the array is subdivided into partitions, which are called extents. The extent size depends on the extent type, which is FB or CKD
- You can choose between large extents and small extents

An FB rank features an extent size of either 1 GB (more precisely a gibibyte (GiB), which is a binary gigabyte that is equal to 2^{30} bytes), called large extents, or an extent size of 16 MiB, called small extents.

IBM Z storage CKD is defined in terms of the original 3390 volume sizes. A 3390 Model 3 is three times the size of a Model 1. A Model 1 features 1113 cylinders, which are about 0.946 GB. The extent size of a CKD rank is one 3390 Model 1, or 1113 cylinders. A 3390 Model 1 (1113 cylinders) is the large extent size for CKD ranks. The smallest CKD extent size is 21 cylinders, which corresponds to the z/OS allocation unit for Extended Address Volume (EAV) volumes larger than 65520 cylinders. z/OS changes the addressing modes and allocates storage in 21 cylinder units.

Whether you use small or large extents depends on the goals that you want to achieve. Small extents provide better capacity utilization, particularly for thin-provisioned storage. With thin-provisioned volumes, using small extents is preferred. However, managing many small extents causes some small performance degradation during initial allocation. For example, a format write of 1 GB requires one storage allocation with large extents, but 64 storage allocations with small extents. Otherwise, host performance should not be adversely affected.

DS8900F ranks are created using the DS CLI command **mkrank** or through the DS GUI.

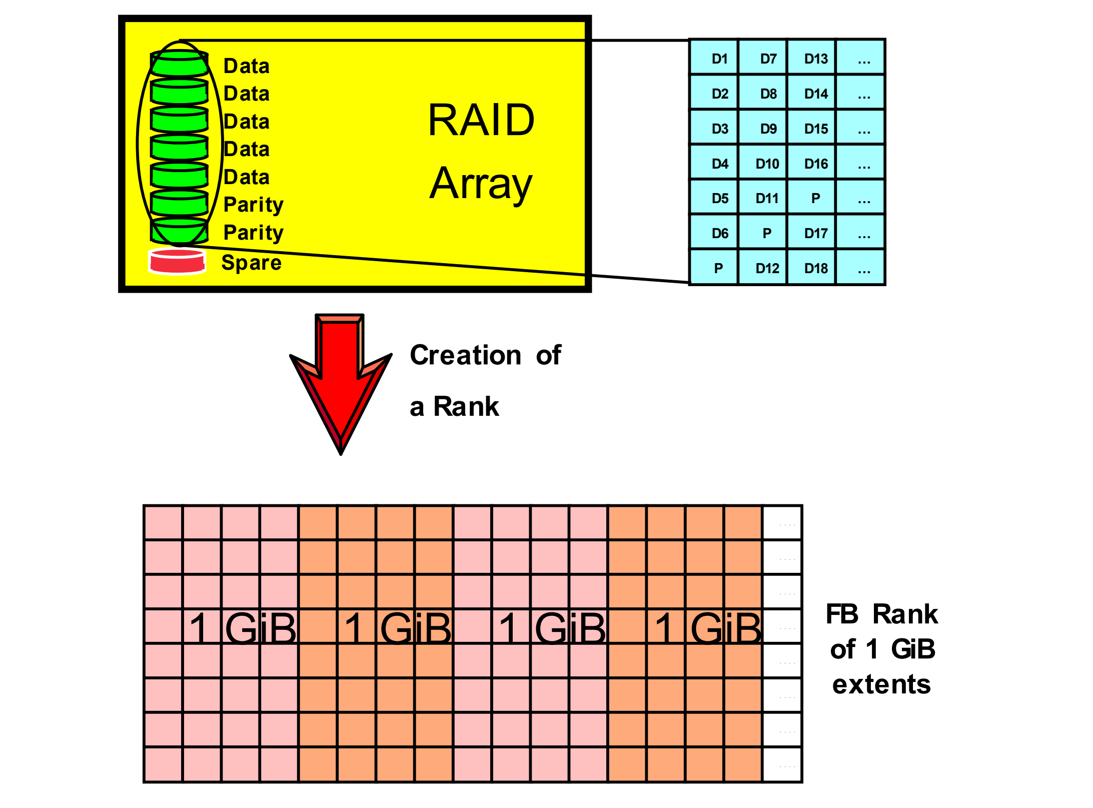
Example 3-2 lsrank command

```
dscli> lsrank -l
ID Group State  datastate Array RAIDtype extpoolID extpoolnam stgtype exts    usedexts keygrp marray extsize (cap)
==================================================================================================================
R0     0 Normal Normal    A0           6 P0        Open_ET    fb      554647    546499      1 MA6    16MiB
R1     0 Normal Normal    A1           6 P0        Open_ET    fb      2219583   279004      1 MA2    16MiB
R2     1 Normal Normal    A2           6 P1        Open_ET    fb      554647    553660      1 MA5    16MiB
R3     1 Normal Normal    A3           6 P1        Open_ET    fb      2219583  1844141      1 MA1    16MiB
R4     0 Normal Normal    A4           6 P2        CKD_z/OS   ckd     429258     36695      1 MA3    21cyl
R5     1 Normal Normal    A5           6 P3        CKD_z/OS   ckd     429258     54314      1 MA4    21cyl
```

6 Ranks were created (see Example 3-2) and can verify the following:

- 4 Fixed Block (fb) Ranks - R0 to R3
  - Small Extents of 16 MiB
- 2 CKD (ckd) - R4 and R5
  - Small Extents of 21 cyl

Figure 3-4 on page 38 shows an example of an array that is formatted for FB data with 1 GiB extents (the squares in the rank indicate that the extent is composed of several blocks from DDMs).



D1	D7	D13	...
D2	D8	D14	...
D3	D9	D15	...
D4	D10	D16	...
D5	D11	P	...
D6	P	D17	...
P	D12	D18	...

Figure 3-4 Creation of a FB Rank

3.1.4 Extent pools

An extent pool is a logical construct to aggregate the extents from a set of ranks, and it forms a domain for extent allocation to a logical volume. Originally, extent pools were used to separate drives with different revolutions per minute (RPM) and capacity in different pools that have homogeneous characteristics. You still might want to use extent pools to separate Tier 0, Tier 1, and Tier 2 flash drives.

If you want Easy Tier to automatically optimize rank utilization, configure more than one rank in an extent pool. A rank can be assigned to only one extent pool. As many extent pools as ranks can exist.

Heterogeneous extent pools, with a mixture of Tier 0, Tier 1, and Tier 2 flash drives can take advantage of the capabilities of Easy Tier to optimize I/O throughput. Easy Tier moves data across different storage tiering levels to optimize the placement of the data within the extent pool.

With storage pool striping, you can create logical volumes that are striped across multiple ranks to enhance performance. To benefit from storage pool striping, more than one rank in an extent pool is required.

Storage pool striping can enhance performance significantly. However, in the unlikely event that a whole RAID array fails, the loss of the associated rank affects the entire extent pool because data is striped across all ranks in the pool. For data protection, consider mirroring your data to another DS8000 storage system.

Tip: With Easy Tier, storage pool striping is irrelevant. It is a much better approach to let Easy Tier manage extents across ranks (arrays).

When an extent pool is defined, it must be assigned the following attributes:

- Internal storage server affinity
- Extent type (FB or CKD)
- Encryption group

A minimum of two extent pools must be configured to balance the capacity and workload between the two servers. One extent pool is assigned to internal server 0. The other extent pool is assigned to internal server 1. In a system with both FB and CKD volumes, four extent pools provide one FB pool for each server and one CKD pool for each server.

Example 3-3 Four Extent Pools - 2 FB; 2 CKD

```
dscli> lsextpool -l
Name     ID stgtype rankgrp status   availstor (2^30B) %allocated available reserved numvols numranks keygrp numtiers etmanaged extsize
=======================================================================================================================================
Open_ET  P0 fb            0    below             30470         29   1950091      128      22        2      1        2 yes       16MiB
Open_ET  P1 fb            1 exceeded              5939         86    380094      128      22        2      1        2 yes       16MiB
CKD_z/OS P2 ckd           0    below              6525          8    392494       53    1536        1      1        1 yes       21cyl
CKD_z/OS P3 ckd           1    below              6232         12    374891       53    1536        1      1        1 yes       21cyl
```

Example 3-3 shows that 4 Extent Pools were created:

- 2 fb Extent Pools (P0 and P1): each with 2 ranks and multitier with 2 Tiers.
- 2 ckd Extent Pools (P2 and P3): each with 1 Rank and Single Tier.

Example 3-4 on page 39 shows the tier details of Extent Pool 1.

Example 3-4 Tier P1 with 2 Tiers

```
===========Tier Distribution============
Tier       Cap (GiB/mod1) %allocated %assigned
==============================================
FlashTier0 8666           100        0
FlashTier2 34680          83         0
```

This is an example of a mixed environment with CKD and FB extent pools. Additional extent pools might be wanted to segregate workloads. See Figure 3-5 on page 40.

Extent pools are expanded by adding more ranks to the pool. Ranks are organized into two rank groups: Rank group 0 is controlled by storage server 0 (processor complex 0), and rank Group 1 is controlled by storage server 1 (processor complex 1).

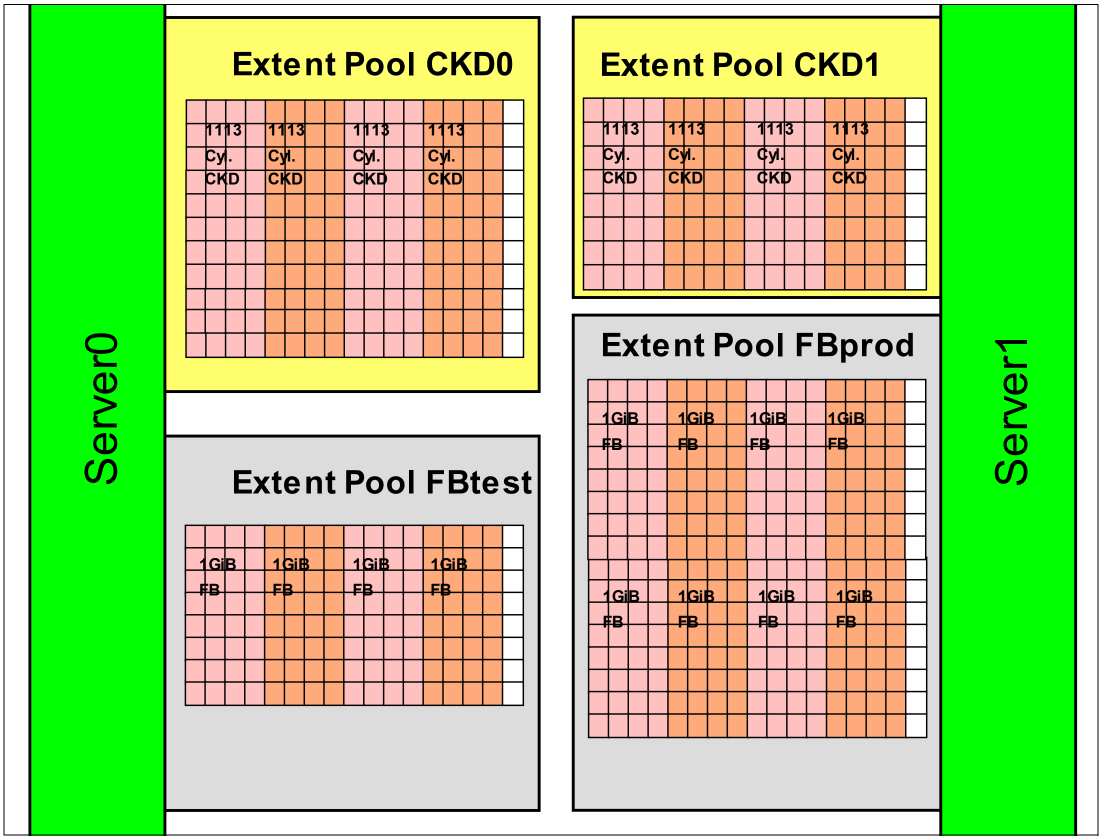


Figure 3-5 Extent Pools allocated to Server 0 and 1

3.1.5 Logical volumes

A logical volume consists of a set of extents from one extent pool. The DS8900F supports up to 65,280 logical volumes (64 K CKD or 64K FB volumes, or a mixture of both, up to a maximum of 64 K total volumes). The abbreviation 64 K is used in this section, even though it is 65,536 minus 256, which is not quite 64 K in binary.

Fixed-block LUNs

A logical volume that is composed of FB extents is called a LUN. An FB LUN is composed of one or more 1 GiB large extents or one or more 16 MiB small extents from one FB extent pool. A LUN cannot span multiple extent pools, but a LUN can have extents from multiple ranks within the same extent pool. You can construct LUNs up to 16 TiB when using large extents.

Count key data volumes

An IBM Z CKD volume is composed of one or more extents from one CKD extent pool. CKD extents are of the size of 3390 Model 1, which features 1113 cylinders for large extents or 21 cylinders for small extents. When you define an IBM Z CKD volume, specify the size of the volume as a multiple of 3390 Model 1 volumes or the number of cylinders that you want for the volume.

Before a CKD volume can be created, an LCU must be defined that provides up to 256 possible addresses that can be used for CKD volumes. Up to 255 LCUs can be defined.

CKD Alias Volumes

Another type of CKD volume is the PAV alias volume. PAVs are used by z/OS to send parallel I/Os to the same CKD volume. Alias Volumes are defined within the same LCU as their corresponding base volumes. Although they have no size, each alias volume needs an address, which is linked to a base volume. The total of base and alias addresses cannot exceed the maximum of 256 for an LCU.

HyperPAV and SuperPAV

With HyperPAV, the alias can be assigned to access a base device within the same LCU on an I/O basis. However, the system can exhaust the alias resources that are needed within an LCU and cause an increase in I/O queue time (the queue time of software that is waiting for a device to issue the I/O). In this case you need more aliases to access host volumes.

Classically, to start an I/O to a base volume, z/OS can select any alias address only from the same LCU to perform the I/O. With SuperPAV, the z/OS can use alias addresses from other LCUs to perform an I/O for a base address.

The restriction is that the LCU of the alias address belongs to the same DS8000 server (processor complex). In other words, if the base address is from an even (odd) LCU, the alias address that z/OS can select must also be from an even (odd) LCU.

Extended Address Volume

On a DS8900F, you can define CKD volumes with up to 1,182,006 cylinders, or about 1 TB. This volume capacity is an EAV supported by the 3390 Model A. A CKD volume cannot span multiple extent pools, but a volume can have extents from different ranks in the same extent pool. You also can stripe a volume across the ranks, as described below.

Extent Allocation Method

Two extent allocation methods (EAMs) are available for the DS8900F: Storage pool striping (rotate extents) and rotate volumes.

Storage pool striping: Extent rotation

The storage allocation method is chosen when a LUN or volume is created. The extents of a volume can be striped across several ranks. An extent pool with more than one rank is needed to use this storage allocation method.

The DS8900F maintains a sequence of ranks. The first rank in the list is randomly picked at each power-on of the storage system. The DS8900F tracks the rank in which the last allocation started. The allocation of the first extent for the next volume starts from the next rank in that sequence. See Figure 3-6 on page 42.

> **Note:** Although the preferred storage allocation method was storage pool striping in the past, it is now a better choice to let Easy Tier manage the storage pool extents. This item describes rotate extents for the sake of completeness, but it is now mostly irrelevant.

Rotate volumes allocation method

Extents can be allocated sequentially. In this case, all extents are taken from the same rank until enough extents are available for the requested volume size or the rank is full. In this case, the allocation continues with the next rank in the extent pool.

If more than one volume is created in one operation, the allocation for each volume starts in another rank.

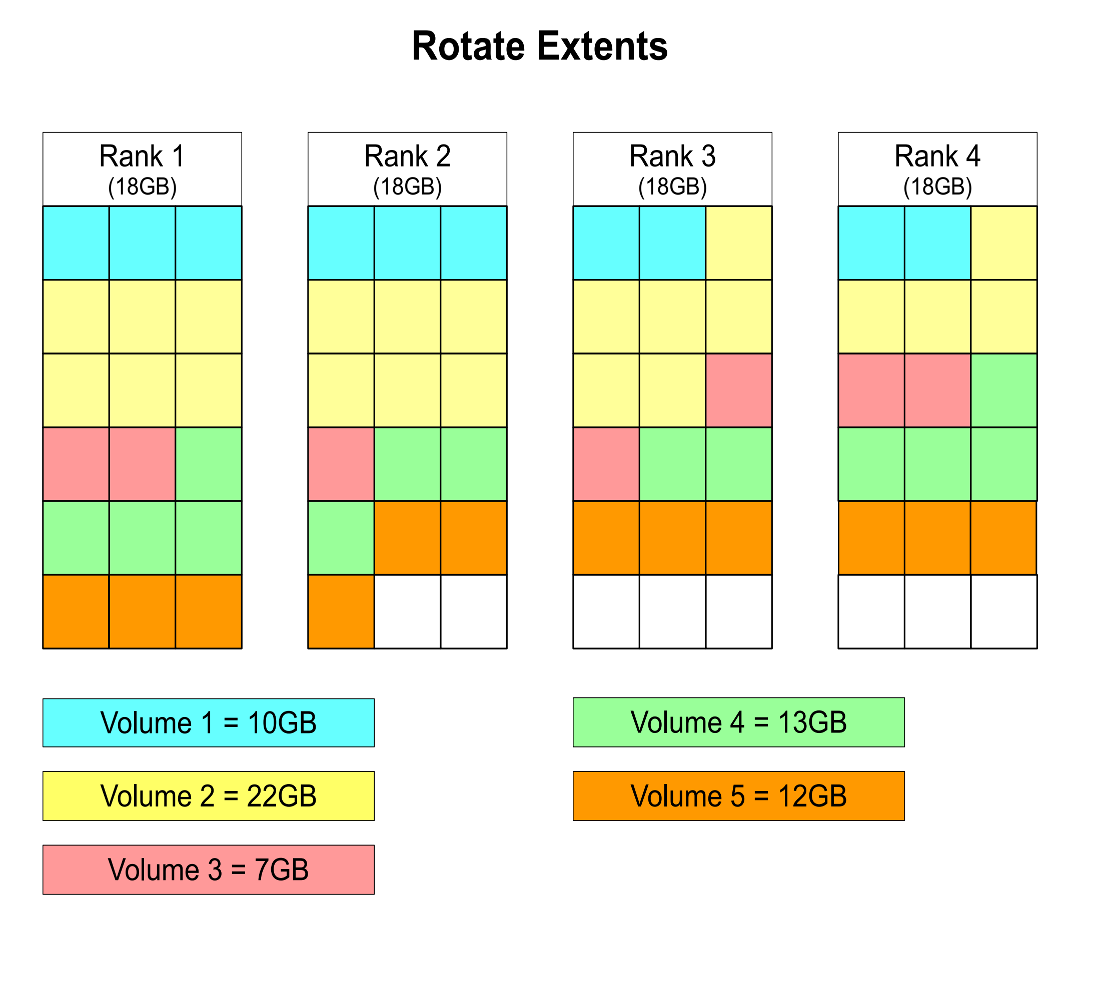


Figure 3-6 Rotate Extents

3.1.6 Easy Tier

Easy Tier enables the DS8000 to automatically balance I/O access to disk drives to avoid hot spots on disk or flash arrays. Easy Tier can place data in the storage tier that best suits the access frequency of the data. Highly accessed data can be moved non-disruptively to a higher tier, for example, to a High-Performance 1.6 TB Flash Tier 0 drive. Cold data or data that is primarily accessed sequentially is moved to a lower tier (Tier 1 or Tier 2) of High-Capacity drives.

Easy Tier includes more components:

- Easy Tier Application is an application-aware storage utility to help deploy storage more efficiently by enabling applications and middleware to direct more optimal placement of the data by communicating important information about current workload activity and application performance requirements. It is possible for Db2 applications in z/OS environments to give hints of data placement to Easy Tier at the data set level.
- Easy Tier Heat Map Transfer (HMT) can take the data placement algorithm on the Metro Mirror (MM) Global Copy (GC) and Global Mirror (GM) primary site and reapply it to the MM / GC / GM secondary site when failover occurs by using the Easy Tier Heat Map Transfer Utility (HMTU) (or respective CSM/GDPS functions). With this capability, the DS8000 systems can maintain application-level performance. The Easy Tier HMT

functions support Metro/Global Mirror (MGM) to transfer a heat map automatically to a tertiary site.

Dynamic extent pool merge

Dynamic extent pool merge is a capability that is provided by the Easy Tier manual mode feature.

Dynamic extent pool merge allows one extent pool to be merged into another extent pool while the logical volumes in both extent pools remain accessible to the host servers. Dynamic extent pool merge can be used for the following scenarios:

- Use dynamic extent pool merge for the consolidation of smaller extent pools of the same storage type into a larger homogeneous extent pool that uses storage pool striping. Creating a larger extent pool allows logical volumes to be distributed evenly over a larger number of ranks, which improves overall performance by minimizing skew and reducing the risk of a single rank that becomes a hot spot. In this case, a manual volume rebalance must be initiated to restripe all existing volumes evenly across all available ranks in the new pool. Newly created volumes in the merged extent pool allocate capacity automatically across all available ranks by using the rotate extents EAM (storage pool striping), which is the default.
- Use dynamic extent pool merge for consolidating extent pools with different storage tiers to create a merged multitier extent pool with a mix of storage classes (High-Performance flash and High-Capacity flash) for automated management by Easy Tier automatic mode.
- Under certain circumstances, use dynamic extent pool merge for consolidating extent pools of the same storage tier but different drive types or RAID levels that can eventually benefit from storage pool striping and Easy Tier automatic mode intra-tier management (auto-rebalance) by using the Easy Tier micro-tiering capabilities.

Important: You can apply dynamic extent pool merge only among extent pools that are associated with the same DS8000 storage system affinity (storage server 0 or storage server 1) or rank group. All even-numbered extent pools (P0, P2, P4, and so on) belong to rank group 0 and are serviced by storage server 0. All odd-numbered extent pools (P1, P3, P5, and so on) belong to rank group 1 and are serviced by storage server 1 (unless one DS8000 storage system failed or is quiesced with a failover to the alternative storage system).

Additionally, the dynamic extent pool merge is not supported in these situations:

- If source and target pools have different storage types (FB and CKD).
- If you select an extent pool that contains volumes that are being migrated.

For more information about dynamic extent pool merge, see *IBM DS8000 Easy Tier*, REDP-4667.

General considerations when adding capacity to an extent pool

If you must add capacity to an extent pool that is almost fully allocated on a DS8000 storage system without the Easy Tier feature, add several ranks to this pool at the same time, not just one. With this approach, the new volumes can be striped across the newly added ranks and reduce skew and hot spots.

With the Easy Tier feature, you can easily add capacity and even single ranks to existing extent pools without concern about performance.

Manual volume rebalance

Manual volume rebalance is a feature of the Easy Tier manual mode and is available only in non-managed, single-tier (homogeneous) extent pools. It allows a balanced redistribution of the extents of a volume across all ranks in the pool. This feature is not available in managed or hybrid pools where Easy Tier is supposed to manage the placement of the extents automatically on the ranks in the pool based on their actual workload pattern, available storage tiers, and rank utilizations.

Manual volume rebalance is designed to redistribute the extents of volumes within a non-managed, single-tier (homogeneous) pool so that workload skew and hot spots are less likely to occur on the ranks. This feature is useful for redistributing extents after adding new ranks to a non-managed extent pool. Also, this feature is useful after merging homogeneous extent pools to balance the capacity and the workload of the volumes across all available ranks in a pool.

Manual volume rebalance provides manual performance optimization by rebalancing the capacity of a volume within a non-managed homogeneous extent pool. It is also called capacity rebalance because it balances the extents of the volume across the ranks in an extent pool without considering any workload statistics or device usage. A balanced distribution of the extents of a volume across all available ranks in a pool is supposed to provide a balanced workload distribution and minimize skew and hot spots. It is also called volume restriping, and it is supported for standard and Extent Space Efficient (ESE) thin-provisioned volumes.

For further information about manual volume rebalance, see *IBM DS8000 Easy Tier*, REDP-4667.

Auto-rebalance

With Easy Tier automatic mode enabled for single-tier or multitier extent pools, you can benefit from Easy Tier automated intratier performance management (auto-rebalance), which relocates extents based on rank utilization, and reduces skew and avoids rank hot spots. Easy Tier relocates subvolume data on extent level based on actual workload pattern and rank utilization (workload rebalance) rather than balance the capacity of a volume across all ranks in the pool (capacity rebalance, as achieved with manual volume rebalance).

When adding capacity to managed pools, Easy Tier automatic mode performance management, auto-rebalance, takes advantage of the new ranks and automatically populates the new ranks that are added to the pool when rebalancing the workload within a storage tier and relocating subvolume data. Auto-rebalance can be enabled for hybrid and homogeneous extent pools.

> **Tip:** For brand new DS8000 storage systems, the Easy Tier automatic mode switch is set to **Tiered**, which means that Easy Tier is working only in hybrid pools. Have Easy Tier automatic mode working in *all* pools, including single-tier pools. To do so, set the Easy Tier automode switch to **All**.

For more information about auto-rebalance, see *IBM DS8000 Easy Tier*, REDP-4667.

Extent allocation in hybrid and managed extent pools

When you create a volume in a managed extent pool, that is, an extent pool that is managed by Easy Tier automatic mode, the EAM of the volume always becomes *managed*. This situation is true no matter which extent allocation method is specified at volume creation. The volume is under control of Easy Tier. Easy Tier moves extents to the most appropriate storage tier and rank in the pool based on performance aspects. Any specified EAM, such as rotate extents or rotate volumes, is ignored. In managed extent pools, an initial EAM similar to rotate extents for new volumes is used.

In managed homogeneous extent pools with only a single storage tier, the initial extent allocation for a new volume is the same as with rotate extents or storage-pool striping. For a volume, the appropriate DSCLI command, `showfbvol` or `showckdvol`, which is used with the `-rank` option, allows the user to list the number of allocated extents of a volume on each associated rank in the extent pool.

The EAM attribute of any volume that is created or already in a managed extent pool is changed to *managed* after Easy Tier automatic mode is enabled for the pool. When enabling Easy Tier automatic mode for all extent pools, that is, hybrid and homogeneous extent pools, all volumes immediately become managed by Easy Tier. Once set to managed, the EAM attribute setting for the volume is permanent. All previous volume EAM attribute information, such as rotate extents or rotate volumes, is lost.

Initial Extent Allocation

Newer code releases have two allocation policy options:

- High Utilization
- High Performance

New volumes are allocated to the home tier. When the allocation policy High Utilization is in effect, the home tier is some middle tier, and when High Performance is in effect, the home tier is the highest available flash tier. The extents of a new volume are distributed in a *rotate extents* or *storage pool striping* fashion across all available ranks in this home tier in the extent pool if sufficient capacity is available. Only when all capacity on the home tier in an extent pool is used does volume creation continue on the ranks of another tier. The allocation order is the following:

- High Utilization: Flash Tier 1 -> Flash Tier 2 -> Flash Tier 0
- High Performance: Flash Tier 0 -> Flash Tier 1 -> Flash Tier 2

Note: The default allocation order is High Performance for all-flash pools

You can change the allocation policy according to your needs with the **chsi** command and the **-ettierorder highutil** or **highperf** option.

Dynamic volume expansion

The size of a LUN or CKD volume can be expanded without losing the data. On the DS8000 storage system, you add extents to the volume. The operating system must support this resizing capability. If a logical volume is dynamically expanded, the extent allocation follows the volume EAM. Dynamically reducing the size of a logical volume is not supported because most operating systems do not support this feature.

Important: Before you can expand a volume, you must remove any Copy Services relationships that involve that volume.

Dynamic volume relocation

Dynamic volume relocation (DVR) or volume migration is a capability that is provided as part of the Easy Tier manual mode feature. It allows a logical volume to be migrated from its current extent pool to another extent pool or even back into the same extent pool (to relocate the extents of a volume) while the logical volume and its data remain accessible to the attached hosts. The user can request DVR by using the migrate volume function that is available through the DS8000 Storage Manager GUI or DSCLI (the **managefbvol** or **manageckdvol** command). DVR allows the user to specify a target extent pool and an EAM.

> **Important:** DVR can be applied among extent pools that are associated with the same DS8000 storage system affinity (storage server 0 or storage server 1) or rank group only. All volumes in even-numbered extent pools (P0, P2, P4, and so on) belong to rank group 0 and are serviced by storage server 0. All volumes in odd-numbered extent pools (P1, P3, P5, and so on) belong to rank group 1 and are serviced by storage server 1. Additionally, the DVR is not supported if source and target pools are different storage types (FB and CKD).

If the same extent pool is specified and rotate extents is used as the EAM, the volume migration is carried out as manual volume rebalance, as described in “Manual volume rebalance” on page 44. Manual volume rebalance is designed to redistribute the extents of volumes within a non-managed, single-tier (homogeneous) pool so that workload skew and hot spots are less likely to occur on the ranks. During extent relocation, only one extent at a time is allocated rather than preallocating the full volume and only a minimum amount of free capacity is required in the extent pool.

> **Important:** A volume migration with DVR back into the same extent pool (for example, manual volume rebalance for restriping purposes) is not supported in managed or hybrid extent pools. Hybrid pools are always supposed to be prepared for Easy Tier automatic management. In pools under control of Easy Tier automatic mode, the volume placement is managed automatically by Easy Tier. It relocates extents across ranks and storage tiers to optimize storage performance and storage efficiency. However, it is always possible to migrate volumes across extent pools, no matter if those pools are managed, non-managed, or hybrid pools.

For more information about this topic, see *IBM DS8000 Easy Tier*, REDP-4667.

3.1.7 Logical subsystem

A logical subsystem (LSS) is another logical construct. It groups logical volumes and LUNs in groups of up to 256 logical volumes.

On the DS8000 series, there is no fixed binding between a rank and an LSS. The capacity of one or more ranks can be aggregated into an extent pool. The logical volumes that are configured in that extent pool are not necessarily bound to a specific rank. Different logical volumes on the same LSS can even be configured in separate extent pools. The available capacity of the storage facility can be flexibly allocated across LSSs and logical volumes. You can define up to 255 LSSs on a DS8000 storage system.

For each LUN or CKD volume, you must select an LSS when creating the volume. The LSS is part of the volume ID ‘`abcd`’ and must to be specified upon volume creation.

> **Hexadecimal notation:** To emphasize the hexadecimal notation of the DS8000 volume IDs, LSS IDs, and address groups, an ‘X’ is used in front of the ID in part of the book.

You can have up to 256 volumes in one LSS. However, there is one restriction. Volumes are created from extents of an extent pool. However, an extent pool is associated with one DS8000 storage system (also called a central processor complex (CPC)): server 0 or server 1. The LSS number also reflects this affinity to one of these DS8000 storage systems. All even-numbered LSSs (X'00', X'02', X'04', up to X'FE') are serviced by storage server 0 (rank group 0). All odd-numbered LSSs (X'01', X'03', X'05', up to X'FD') are serviced by storage server 1 (rank group 1). LSS X’FF’ is reserved.

All logical volumes in an LSS must be either CKD or FB. LSSs are even grouped into address groups of 16 LSSs. All LSSs within one address group must be of the same storage type, either CKD or FB. The first digit of the LSS ID or volume ID specifies the address group. For more information, see 3.1.8, “Address groups” on page 47.

IBM Z users are familiar with a logical control unit (LCU). z Systems operating systems configure LCUs to create device addresses. There is a one-to-one relationship between an LCU and a CKD LSS (LSS X'ab' maps to LCU X'ab'). Logical volumes have a logical volume number X'abcd' in hexadecimal notation where X'ab' identifies the LSS and X'cd' is one of the 256 logical volumes on the LSS. This logical volume number is assigned to a logical volume when a logical volume is created and determines with which LSS the logical volume is associated. The 256 possible logical volumes that are associated with an LSS are mapped to the 256 possible device addresses on an LCU (logical volume X'abcd' maps to device address X'cd' on LCU X'ab'). When creating CKD logical volumes and assigning their logical volume numbers, consider whether PAVs are required on the LCU and reserve addresses on the LCU for alias addresses.

For Open Systems, LSSs do not play an important role other than associating a volume with a specific rank group and server affinity (storage server 0 or storage server 1) or grouping hosts and applications together under selected LSSs for the DS8000 Copy Services relationships and management.

Tip: Certain management actions in Metro Mirror, Global Mirror, or Global Copy operate at the LSS level. For example, the freezing of pairs to preserve data consistency across all pairs is at the LSS level. The option to put all or a set of volumes of a certain application in one LSS can make the management of remote copy operations easier under certain circumstances.

Important: LSSs for FB volumes are created automatically when the first FB logical volume on the LSS is created, and deleted automatically when the last FB logical volume on the LSS is deleted. CKD LSSs require user parameters to be specified and must be created before the first CKD logical volume can be created on the LSS. They must be deleted manually after the last CKD logical volume on the LSS is deleted.

3.1.8 Address groups

Address groups are created automatically when the first LSS that is associated with the address group is created, and deleted automatically when the last LSS in the address group is deleted. The DSCLI **lsaddressgrp** command displays a list of address groups in use for the storage image and the storage type that is associated with it, either FB or CKD.

All devices in an address group must be either CKD or FB. LSSs are grouped into address groups of 16 LSSs. LSSs are numbered X'ab', where a is the address group. So, all LSSs within one address group must be of the same type, CKD or FB. The first LSS defined in an address group sets the type of that address group. For example, LSS X'10' to LSS X'1F' are all in the same address group and therefore can all be used only for the same storage type, either FB or CKD. Figure 3-7 on page 48 shows the concept of volume IDs, LSSs, and address groups.

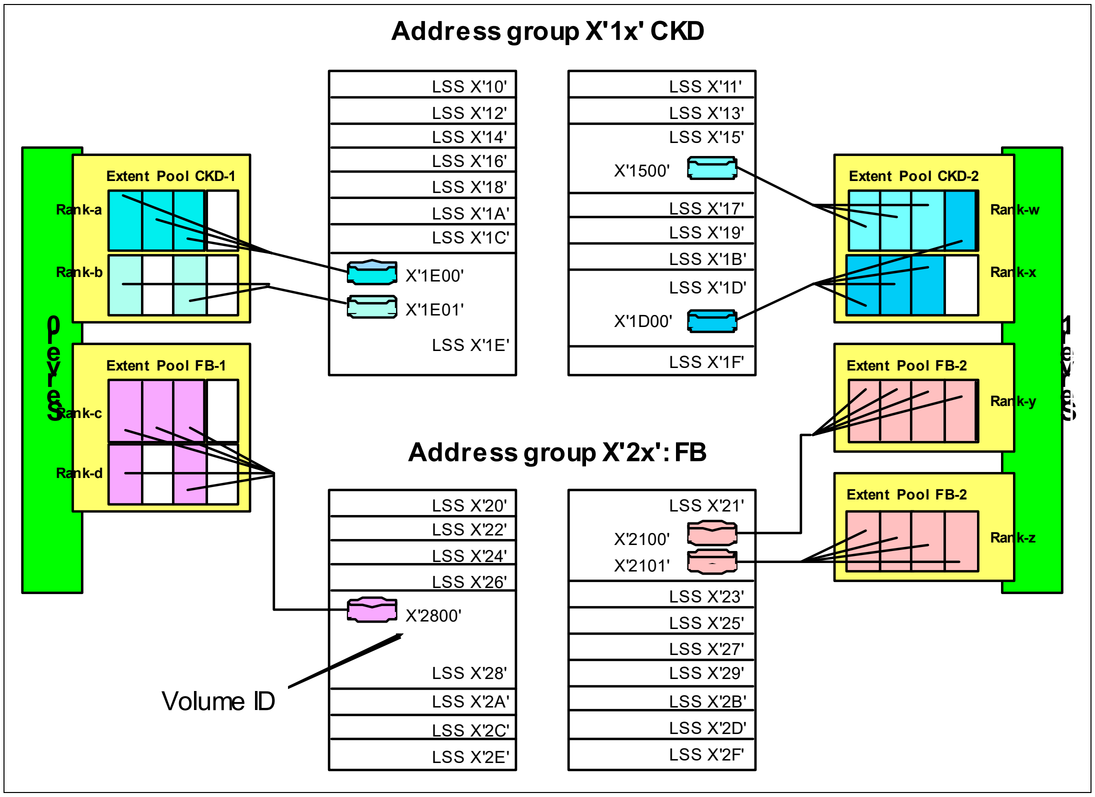


Figure 3-7 Volume IDs, logical subsystems, and address groups on the DS8000 storage systems

The volume ID X'*gabb*' in hexadecimal notation is composed of the address group X'*g*', the LSS ID X'g*a*', and the volume number X'*bb*' within the LSS. For example, LUN X'2101' denotes the second (X'01') LUN in LSS X'21' of address group 2.

3.1.9 Volume access

A DS8000 storage system provides mechanisms to control host access to LUNs. In most cases, a server or host has two or more host bus adapters (HBAs) and needs access to a group of LUNs. For easy management of server access to logical volumes, the DS8000 storage system uses the concept of host attachments and volume groups.

Host attachment

Host bus adapters (HBAs) are identified to the DS8000 storage system in a host attachment or host connection construct that specifies the HBA worldwide port names (WWPNs). A set of host ports (host connections) can be associated through a port group attribute in the DSCLI that allows a set of HBAs to be managed collectively. This group is called a host attachment within the GUI.

Each host attachment can be associated with a volume group to define which LUNs that HBA is allowed to access. Multiple host attachments can share the volume group. The host attachment can also specify a port mask that controls which DS8000 I/O ports the HBA is allowed to log in to. Whichever ports the HBA logs in to, it sees the same volume group that is defined on the host attachment that is associated with this HBA.

The maximum number of host attachments on a DS8000 is 8,192.

Volume group

A volume group is a named construct that defines a set of logical volumes. When used with CKD hosts, there is a default volume group that contains all CKD volumes. Any CKD host that logs in to a FICON I/O port has access to the volumes in this volume group. CKD logical volumes are automatically added to this volume group when they are created and automatically removed from this volume group when they are deleted.

When used with Open Systems hosts, a host attachment object that identifies the HBA is linked to a specific volume group. You must define the volume group by indicating which FB logical volumes are to be placed in the volume group. Logical volumes can be added to or removed from any volume group dynamically.

One host connection can be assigned to one volume group only. However, the same volume group can be assigned to multiple host connections. An FB logical volume can be assigned to one or more volume groups. Assigning a logical volume to different volume groups allows a LUN to be shared by hosts, each configured with its own dedicated volume group and set of volumes (in case a set of volumes that is not identical is shared between the hosts).

The maximum number of volume groups is 8,320 for the DS8000 storage system.

3.1.10 Summary of the DS8000 logical configuration hierarchy

Describing the virtualization hierarchy, this section started with multiple disks that were grouped in array sites. An array site is transformed into an array, eventually with spare disks. The array was further transformed into a rank with extents formatted for FB data or CKD. Next, the extents were added to an extent pool that determined which storage server served the ranks and aggregated the extents of all ranks in the extent pool for later allocation to one or more logical volumes. Within the extent pool, you can optionally reserve storage for space-efficient volumes.

Next, this section described the creation of logical volumes within the extent pools (optionally striping the volumes), assigning them a logical volume number that determined to which LSS they are associated and indicated which server manages them. Space-efficient volumes can be created immediately or within a repository of the extent pool. Then, the LUNs can be assigned to one or more volume groups. Finally, the HBAs were configured into a host attachment that is associated with a volume group.

The above concept is seen when working with the DSCLI. When working with the DS Storage Manager GUI, some complexity is reduced externally: Instead of array sites, arrays and ranks, the GUI just speaks of Arrays or Managed Arrays, and these are included under Pools. Also, the concept of a volume group is not directly visible externally when working with the GUI.

This virtualization concept provides for greater flexibility. Logical volumes can dynamically be created, deleted, migrated, and resized. They can be grouped logically to simplify storage management. Large LUNs and CKD volumes reduce the required total number of volumes, which also contributes to a reduction of management efforts.

Figure 3-8 on page 50 summarizes the virtualization hierarchy.

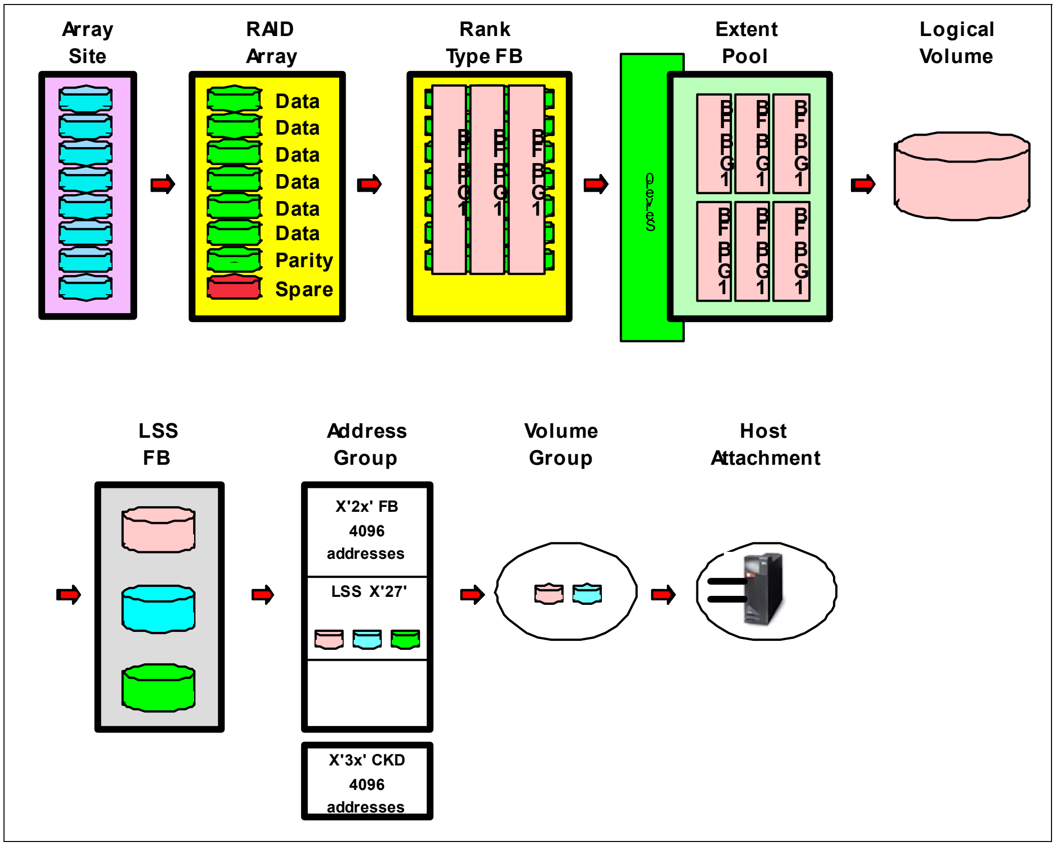


Figure 3-8 DS8000 virtualization hierarchy

3.2 Data placement on ranks and extent pools

To determine the optimal DS8900F layout, define the I/O Performance requirements of the servers and applications first because they are a large factor in the physical and logical configuration of the required disk system.

3.2.1 Workload characteristics

It is very important to understand the workload requirements. The following information must be considered for a detailed modeling:

- Number of IOPS
- I/O density
- Megabytes per second
- Read/Write ratio
- Random or sequential access characteristics
- Cache-hit ratio
- Response time

As a best practice, use IBM Storage Modeler to size the cache and I/O adapter requirements.

3.2.2 DS8900F Resource utilization

After having determined the disk system throughput, it is necessary to optimize the DS8900F resource utilization using the following guidelines:

- Balance the ranks and extent pools between the Servers.
- Spread the logical volume workload across servers 0 and 1.
- Use as many disks as possible and avoid idle disks.
- Distribute capacity and workload across DA pairs.
- Use multi-rank extent pools.
- Stripe your logical volumes across several ranks.
- Consider placing specific database objects(logs) on separate ranks.
- For an application, use volumes from even-numbered and odd-numbered extent pools.

It is important to understand how the volume data is placed on ranks and extent pools. This understanding helps you decide how to create extent pools and choose the required number of ranks within an extent pool. It also helps you understand and detect performance problems or optimally tweak overall system performance.

It can also help you fine-tune system performance from an extent pool perspective, for example, sharing the resources of an extent pool evenly between application workloads or isolating application workloads to dedicated extent pools. Data placement can help you when planning for dedicated extent pools with different performance characteristics and storage tiers without using Easy Tier automatic management. Plan your configuration carefully to meet your performance goals by minimizing potential performance limitations that might be introduced by single resources that become a bottleneck because of workload skew. For example, use rotate extents as the default EAM to help reduce the risk of single ranks that become a hot spot and limit the overall system performance because of workload skew.

If workload isolation is required in your environment, you can isolate workloads and I/O on the rank and DA levels on the DS8000 storage systems, if required.

You can manually manage storage tiers that are related to different homogeneous extent pools of the same storage class and plan for appropriate extent pools for your specific performance needs. Easy Tier provides optimum performance and a balanced resource utilization at a minimum configuration and management effort.

3.2.3 Extent pool considerations

The data placement and thus the workload distribution of your volumes on the ranks in a non-managed homogeneous extent pool is determined by the selected EAM, as described in “Extent Allocation Method” on page 41 and the number of ranks in the pool.

Even in single-tier homogeneous extent pools, you can benefit from Easy Tier automatic mode (by running the DSCLI `chsi ETautomode=all` command). It manages the subvolume data placement within the managed pool based on rank utilization and thus reduces workload skew and hot spots (auto-rebalance).

In multitier hybrid extent pools, you can fully benefit from Easy Tier automatic mode (by running the DSCLI `chsi ETautomode=all|tiered` command). It provides full automatic storage performance and storage economics management by optimizing subvolume data placement in a managed extent pool across different storage tiers and even across ranks within each storage tier (auto-rebalance). Easy Tier automatic mode and hybrid extent pool configurations offer the most efficient way to use different storage tiers. It optimizes storage performance and storage economics across three drive tiers to manage more applications effectively and efficiently with a single DS8000 storage system at an optimum price versus the performance and footprint ratio.

The data placement and extent distribution of a volume across the ranks in an extent pool can be displayed by running the DSCLI **showfbvol -rank** or **showckdvol -rank** command, as shown in Example 3-5 .

Before configuring extent pools and volumes, be aware of the basic configuration principles about workload sharing, isolation, and spreading, as described in 4.2, “Configuration principles for optimal performance” on page 66.

Single-tier extent pools

A single-tier or homogeneous extent pool is formed by using drives of the same storage tier with the same drive characteristics, which means not combining High-Performance flash with High-Capacity flash in the same extent pool.

The first example, which is shown in Figure 3-9 on page 53, illustrates an extent pool with only one rank, which is also referred to a *single-rank extent pool*. This approach is common if you plan to use a configuration that uses the maximum isolation that you can achieve on the rank/extpool level. In this type of a single-rank extent pool configuration, all volumes that are created are bound to a single rank. This type of configuration requires careful logical configuration and performance planning because single ranks are likely to become a hot spot and might limit overall system performance. It also requires the highest administration and management effort because workload skew typically varies over time. You might constantly monitor your system performance and need to react to hot spots. It also considerably limits the benefits that a DS8000 storage system can offer regarding its virtualization and Easy Tier automatic management capabilities.

In these configurations, use host-based striping methods to achieve a balanced distribution of the data and the I/O workload across the ranks and back-end disk resources. For example, you can use IBM AIX® LVM or SAN Volume Controller to stripe the volume data across ranks, or use SVC-based Easy Tier.

Figure 3-9 on page 53 shows the data placement of two volumes created in an extent pool with a single rank. Volumes that are created in this extent pool always use extents from rank R6, and are limited to the capacity and performance capability of this single rank. Without the use of any host-based data and workload striping methods across multiple volumes from different extent pools and ranks, this rank is likely to experience rank hot spots and performance bottlenecks.

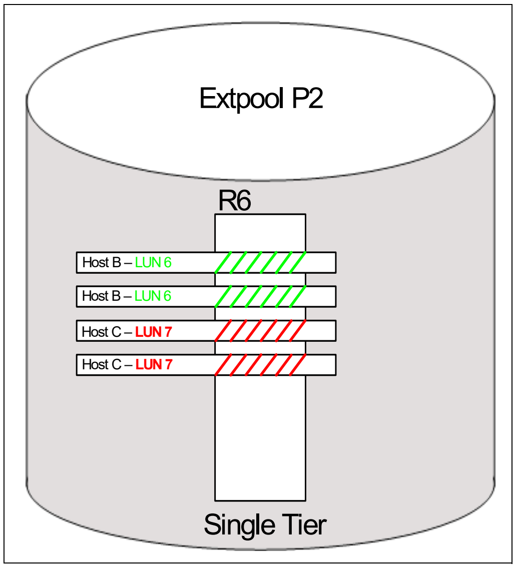


Figure 3-9 Extent pool with a single rank

Also, in this example, one host can easily degrade the whole rank, depending on its I/O workload, and affect multiple hosts that share volumes on the same rank if you have more than one LUN allocated in this extent pool.

The second example, which is shown in Figure 3-10 on page 53, illustrates an extent pool with multiple ranks of the same storage class or storage tier, which is referred to as a homogeneous or single-tier extent pool. In general, an extent pool with multiple ranks also is called a multi-rank extent pool.

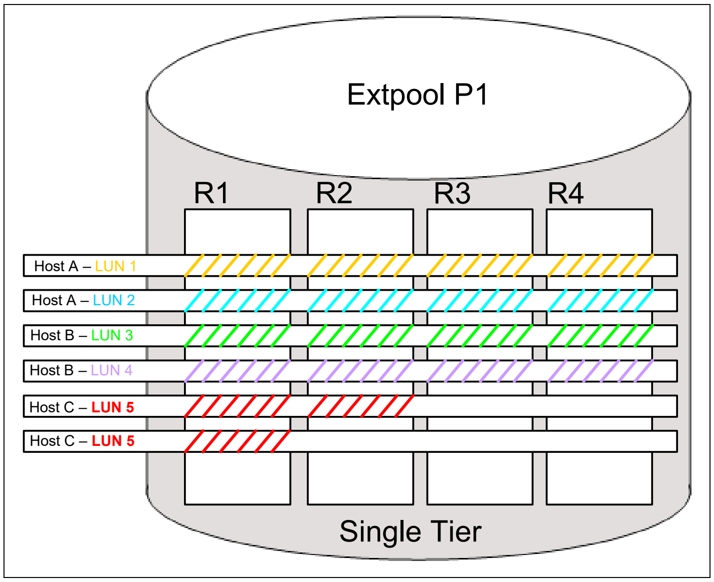


Figure 3-10 Single-tier extent pool with multiple ranks

Although in principle both EAMs (rotate extents and rotate volumes) are available for non-managed homogeneous extent pools, it is preferable to use the default allocation method of rotate extents (storage pool striping). Use this EAM to distribute the data and thus the workload evenly across all ranks in the extent pool and minimize the risk of workload skew and a single rank that becomes a hot spot.

The use of the EAM rotate volumes still can isolate volumes to separate ranks, even in multi-rank extent pools, wherever such configurations are required. This EAM minimizes the configuration effort because a set of volumes that is distributed across different ranks can be created with a single command. Plan your configuration and performance needs to implement host-level-based methods to balance the workload evenly across all volumes and ranks. However, it is not a preferred practice to use both EAMs in the same extent pool without an efficient host-level-based striping method for the non-striped volumes. This approach easily forfeits the benefits of storage pool striping and likely leads to imbalanced workloads across ranks and potential single-rank performance bottlenecks.

Figure 3-10 on page 53 is an example of storage-pool striping for LUNs 1 - 4. It shows more than one host and more than one LUN distributed across the ranks. In contrast to the preferred practice, it also shows an example of LUN 5 being created with the rotate volumes EAM in the same pool. The storage system tries to allocate the continuous space available for this volume on a single rank (R1) until there is insufficient capacity that is left on this rank and then it spills over to the next available rank (R2). All workload on this LUN is limited to these two ranks. This approach considerably increases the workload skew across all ranks in the pool and the likelihood that these two ranks might become a bottleneck for all volumes in the pool, which reduces overall pool performance.

Multiple hosts with multiple LUNs, as shown in Figure 3-10 on page 53, share the resources (resource sharing) in the extent pool, that is, ranks, DAs, and physical spindles. If one host or LUN has a high workload, I/O contention can result and easily affect the other application workloads in the pool, especially if all applications have their workload peaks at the same time. Alternatively, applications can benefit from a much larger amount of disk spindles and thus larger performance capabilities in a shared environment in contrast to workload isolation and only dedicated resources. With resource sharing, expect that not all applications peak at the same time, so that each application typically benefits from the larger amount of disk resources that it can use. The resource sharing and storage-pool striping in non-managed extent pools method is a good approach for most cases if no other requirements, such as workload isolation or a specific quality of service (QoS) requirements, dictate another approach.

Enabling Easy Tier automatic mode for homogeneous, single-tier extent pools always is an additional option, and is preferred, to let the DS8000 storage system manage system performance in the pools based on rank utilization (auto-rebalance). The EAM of all volumes in the pool becomes managed in this case. With Easy Tier and its advanced micro-tiering capabilities that take different RAID levels and drive characteristics into account for determining the rank utilization in managed pools, even a mix of different drive characteristics and RAID levels of the same storage tier might be an option for certain environments.

With Easy Tier the DS8900F family offers advanced features when taking advantage of resource sharing to minimize administration efforts and reduce workload skew and hot spots while benefiting from automatic storage performance, storage economics, and workload priority management. The use of these features in the DS8000 environments is highly encouraged. These features generally help provide excellent overall system performance while ensuring (QoS) levels by prioritizing workloads in shared environments at a minimum of administration effort and at an optimum price-performance ratio.

Multitier extent pools

A multitier or hybrid extent pool is formed by combining ranks of different storage classes or different storage tiers within the same extent pool. Hybrid pools are always supposed to be managed by Easy Tier automatic mode, which provides automated storage performance and storage economics management. It provides full automated cross-tier and intra-tier management by dynamically relocating subvolume data (at the extent level) to the appropriate rank and storage tier within an extent pool based on its access pattern. Easy Tier supports automatic management of up to three storage tiers that use High-Performance and High-Capacity flash drives. Using Easy Tier to let the system manage data placement and performance is highly encouraged with DS8900F storage systems.

When you create a volume in a managed extent pool, that is, an extent pool that is managed by Easy Tier automatic mode, the EAM of the volume always becomes managed. This situation is true no matter which EAM is specified at volume creation. The volume is under control of Easy Tier. Easy Tier moves extents to the most appropriate storage tier and rank in the pool based on performance aspects. Any specified EAM, such as rotate extents or rotate volumes, is ignored.

Mixing different storage tiers combined with Easy Tier automatic performance and economics management on a subvolume level can considerably increase the performance versus price ratio, increase energy savings, and reduce the overall footprint. The use of Easy Tier automated subvolume data relocation and the addition of an flash tier are good for mixed environments with applications that demand both IOPS and bandwidth at the same time. For example, database systems might have different I/O demands according to their architecture. Costs might be too high to allocate a whole database on High-Performance flash.

Mixing different drive technologies, for example, High-Performance flash with High-Capacity flash, and efficiently allocating the data capacity on the subvolume level across the tiers with Easy Tier can highly optimize price, performance, the footprint, and the energy usage. Only the hot part of the data allocates into High-Performance flash instead of provisioning this type of flash capacity for full volumes. Therefore, you can achieve considerable system performance at a reduced cost and footprint with only a few High-Performance flash drives.

The ratio of High-Performance flash to High-Capacity flash in a hybrid pool depends on the workload characteristics and skew.

For a two-tier configuration always configure an equal or greater number of HPFE Enclosure Pairs and drive sets for the upper tier compared to the lower tier. You should not use less drives of High-Performance flash (Tier 0) than of High-Capacity flash (Tier 2). For instance, mixing High-Performance flash of 3.2 TB with High-Capacity flash of 15.36 TB under these conditions, you already come to almost 20% of net capacity in Flash Tier 0.

The DS8000 GUI also provides guidance for High-Performance flash planning based on the existing workloads on a DS8000 storage system with Easy Tier monitoring capabilities. For more information about the DS8900F GUI, see https://www.ibm.com/docs/en/ds8900/ .

Use the hybrid extent pool configurations under automated Easy Tier management. It provides ease of use with minimum administration and performance management efforts while optimizing the system performance, price, footprint, and energy costs.

3.2.4 Easy Tier considerations

When an extent pool is under control of Easy Tier automatic mode, the EAM of all volumes in this pool becomes managed. Easy Tier automatically manages subvolume data placement on the extent level and relocates the extents based on their access pattern to the appropriate storage tier in the pool (cross-tier or inter-tier management). Easy Tier also rebalances the extents across the ranks of the same storage tier based on resource usage (intra-tier management or auto-rebalance) to minimize skew and avoid hot spots in the extent pool. Easy Tier can be enabled for multitier hybrid extent pools with up to three storage tiers or single-tier homogeneous extent pools and automatically manage the data placement of the volumes in these pools.

The initial allocation of extents for a volume in a managed single-tier pool is similar to the rotate extents EAM or storage pool striping. So, the extents are evenly distributed across all ranks in the pool right after the volume creation. The initial allocation of volumes in *hybrid* extent pools differs slightly.

After the initial extent allocation of a volume in the pool, the extents and their placement on the different storage tiers and ranks are managed by Easy Tier. Easy Tier collects workload statistics for each extent in the pool and creates migration plans to relocate the extents to the appropriate storage tiers and ranks. The extents are promoted to higher tiers or demoted to lower tiers based on their actual workload patterns. The data placement of a volume in a managed pool is no longer static or determined by its initial extent allocation. The data placement of the volume across the ranks in a managed extent pool is subject to change over time to constantly optimize storage performance and storage economics in the pool. This process is ongoing and always adapting to changing workload conditions. After Easy Tier data collection and automatic mode are enabled, it might take a few hours before the first migration plan is created and applied. For more information about Easy Tier migration plan creation and timings, see *IBM DS8000 Easy Tier*, REDP-4667.

The DSCLI **`showfbvol -rank`** or **`showckdvol -rank`** commands, and the **`showfbvol -tier`** or **`showckdvol -tier`** commands, can help show the current extent distribution of a volume across the ranks and tiers, as shown in Example 3-5. In this example, volume 8099 is managed by Easy Tier and distributed across ranks R0 (FlashTier0 - High-Performance flash), R1 (FlashTier2 - High-Capacity flash). You can use the **`lsarray -l -rank Rxy`** command to show the storage class and DA pair of a specific rank R*xy*.

Example 3-5 showfbvol -tier and showfbvol -rank commands to show the volume-to-rank relationship in a multitier pool

```
dscli> showfbvol -tier 8099
extpool          P0
===========Tier Distribution============
Tier       %allocated
=====================
FlashTier0         78
FlashTier2         22

dscli> showfbvol -rank 8099
==============Rank extents==============
rank extents capacity (MiB/cyl) metadata
========================================
R0    507786 8124576            yes
R1    147574 2361184            no

dscli> lsarray -l -rank R0
Array State    Data   RAIDtype  arsite Rank DA Pair DDMcap (10^9B) diskclass  encrypt
=======================================================================================
A0    Assigned Normal 6 (6+P+Q) S6     R0   11              1600.0 FlashTier0

dscli> lsarray -l -rank R1
Array State    Data   RAIDtype  arsite Rank DA Pair DDMcap (10^9B) diskclass encrypt
======================================================================================
A1    Assigned Normal 6 (5+P+Q+S)S2     R1   10              7680.0 FlashTier2
```

The volume heat distribution (volume heat map), which is provided by DS GUI helps you identify hot, warm, and cold extents for each volume and its distribution across the storage tiers in the pool. For more information about the DS GUI volume heat distribution, see Chapter 5 of *IBM DS8000 Easy Tier*, REDP-4667.

Easy Tier constantly adapts to changing workload conditions and relocates extents, so the extent locations of a volume are subject to a constant change over time that depends on its workload pattern. However, the number of relocations decreases and eventually becomes marginal if the workload pattern about the decision windows of Easy Tier remains steady.

3.2.5 Configuration Hints

Often, High-Capacity Flash Tier 2 drives in a single tier configuration fulfill the given requirements of a normal load completely.

- For small storage system configurations, may try everything in High-Capacity Flash of 1.92 TB (100%)
- For large storage system configurations, may try everything in High-Capacity Flash of 7.68 TB (100%)

Preferential 2-tier combinations (specially if workload is heavier) are:

- High-Performance Flash of 800 GB combined with High-Capacity Flash of 1.92 TB
- High-Performance Flash of 1.6 TB combined with High-Capacity Flash of 7.68 TB
- High-Performance Flash of 3.2 TB combined with High-Capacity Flash of 15.36 TB

Another good configuration for heavier loads and very heavy access densities is 100% of the capacity in Flash Tier 1 - 3.84 TB.

Only in case of very high loads and very heavy access densities, having a storage system configuration of 100% High-Performance Flash (Flash Tier 0) is usually needed.

Three-tier combinations are uncommon, and mostly not economical.

These above are general configuration hints for a storage system to begin with and do not replace a proper sizing with StorM.

> **Tip:** For a two-tier configuration always configure an equal or greater number of HPFE Enclosure Pairs and drive sets for the upper tier (for instance Tier 0) compared to the lower tier (for instance Tier 2).

Please also note that for certain types of large sequential load, 2 drivesets per HPFE instead of 3 drivesets, can be recommended, and a configuration with 3 HPFEs of 2 drivesets 1.92 TB Flash Tier 2 each may be faster and cheaper for that kind of workload than a configuration with 2 HPFEs filled with 3 drivesets 1.6 TB Flash Tier 0 each.

Again, a StorM sizing will give the precise answer for your workload profile.

4

Logical configuration performance considerations

Important: Before reading this chapter, familiarize yourself with the material that is covered in Chapter 3, "Logical configuration concepts and terminology" on page 31.

This chapter introduces a step-by-step approach to configuring the IBM Storage System DS8900F depending on workload and performance considerations:

- DS8900F Starter drive choices
- Reviewing the tiered storage concepts and Easy Tier
- Understanding the configuration principles for optimal performance:
  - Workload isolation
  - Workload resource-sharing
  - Workload spreading
- Analyzing workload characteristics to determine isolation or resource-sharing
- Planning allocation of the DS8900F drive and host connection capacity to identified workloads
- Planning spreading volumes and host connections for the identified workloads
- Planning array sites
- Planning RAID arrays and ranks with RAID-level performance considerations
- Planning extent pools with single-tier and multitier extent pool considerations
- Planning address groups, Logical SubSystems (LSSs), volume IDs, and Count Key Data (CKD) Parallel Access Volumes (PAVs)
- Planning I/O port IDs, host attachments, and volume groups
- Logical configuration

4.1 DS8900F Models

The DS8900F family is the most advanced of the DS8000 offerings to date. The DS8900F family consists of the following high-performance Models:

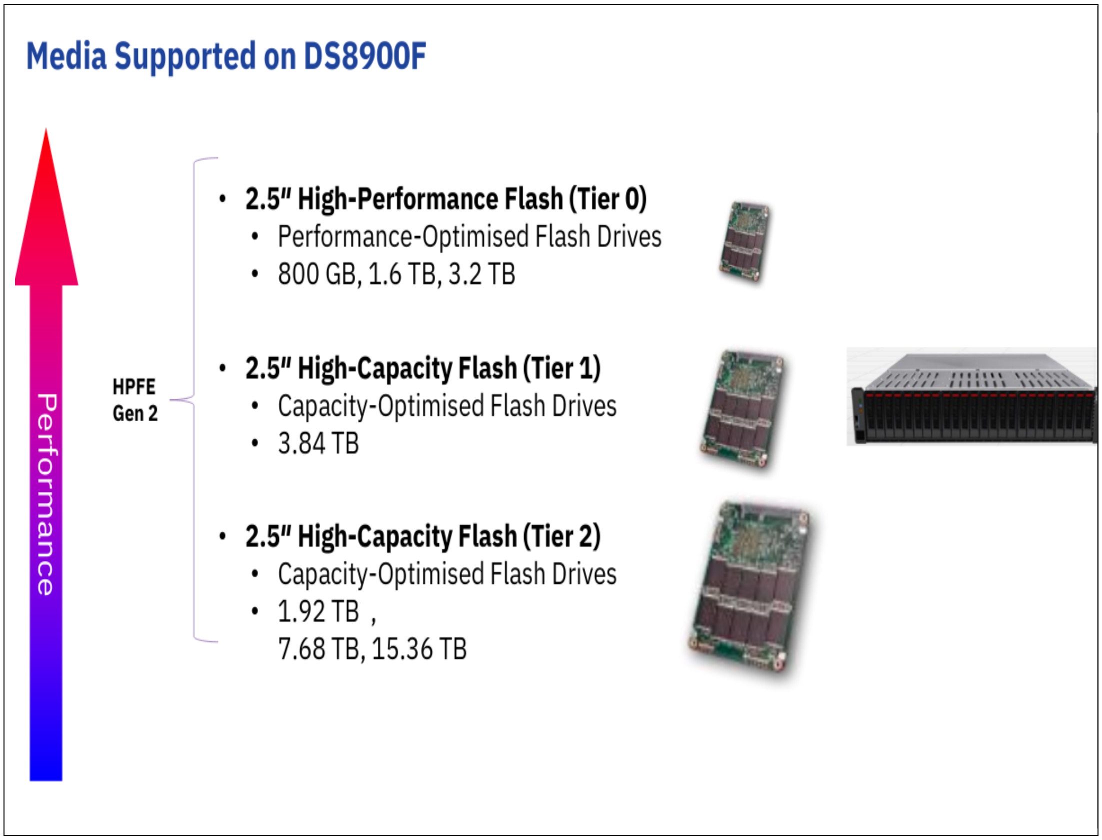
- DS8910F - Flexibility Class (rack-mounted and racked)
- DS8950F - Agility Class
- DS8980F - Analytic Class

4.1.1 DS8910F

DS8910F has the following characteristics:

- Two IBM Power Systems POWER9 MTM 9009-22A
- 8 POWER9 cores per CEC
- From 192 GB to 512 GB of System Memory
- From 8 to 64 Host Adapter Ports
- From 16 to 192 Flash drives
- From 12.8 TB to 2,949 TB of capacity (with 15.36 TB Flash Drives)
- 860K of Maximum IOps (4K 70/30 R/W mix)
- Maximum of 21 GB/s (Read)/17 GB/s (Write) of Sequential Read - Write

Below we can see the flash drive types supported by DS8900F:



Media Supported on DS8900F

- 2.5" High-Performance Flash (Tier 0)
  - Performance-Optimised Flash Drives
  - 800 GB, 1.6 TB, 3.2 TB
- 2.5" High-Capacity Flash (Tier 1)
  - Capacity-Optimised Flash Drives
  - 3.84 TB
- 2.5" High-Capacity Flash (Tier 2)
  - Capacity-Optimised Flash Drives
  - 1.92 TB ,
    7.68 TB, 15.36 TB

Figure 4-1 Flash Drives supported by DS8900F

For the DS8910F we have the following suggested Starter drive choices: See Table 4-1 on page 61.

Use these options and a storage sizing tool such as StorM:

Table 4-1   DS8910F Starter Drive choices

Configuration size	Typical Workload	Heavy Workload
< 60 TB	1.92 TB drives	800 GB drives
60 to 200 TB	1.92 TB drives	1.6 TB drives or 800 GB & 1.92 TB drives
200 TB to 500 TB	1.6 TB & 7.68 TB drives	3.84 TB drives
500 TB to 1 PB	7.68 TB drives or use DS8950F	Use DS8950F

Observe the following configuration rules:

- For a tiered configuration always configure an equal or greater number of HPFE Enclosure Pairs and drive sets for the upper tier compared to the lower tier.
- For optimal performance configure at least two HPFE Enclosure Pairs spreading drive sets over the enclosure pairs.
- Use the “Typical” workload selection unless there is a heavy sequential write activity either due to host workload or extensive use of FlashCopy combined with significant host write workload.
- The capacity limits are not intended to be exact, so capacity requirements close to the boundary between two sizes could use either option.
- Use a sizing tool such as StorM.

4.1.2 DS8950F

DS8950F has the following characteristics:

- 2 × IBM Power Systems POWER9 MTM 9009-42A
- 20/40 POWER9 cores per CEC
- From 512 GB to 3,456 GB of System Memory
- From 8 to 128 Host Adapter Ports
- From 16 to 384 Flash drives
- From 12.8 TB to 5,898 TB of capacity (with 15.36 TB Flash Drives)
- 2,300K of Maximum IOps (4K 70/30 R/W mix)
- Maximum of 63 GB/s (Read)/ 32 GB/s (Write) of Sequential Read - Write

Table 4-2 shows suggested Starter drive choices for the DS8950F, which can be used along with some storage sizing tool such as StorM.

Table 4-2   DS8950F Starter drive choices

Configuration size	Typical Workload	Heavy Workload
< 60 TB	Use DS8910F	800 GB drives
60 to 200 TB	3.84 TB drives or use DS8910F	1.6 TB drives
200 TB to 800 TB	1.6 TB & 7.68 TB drives	3.84 TB drives
800 TB to 1.2 PB	7.68 TB drives	3.2 TB & 15.36 TB drives
1.2 PB to 4 PB	15.36 TB drives	Use multiple DS8950F

Observe these configuration rules:

- For a tiered configuration always configure an equal or greater number of HPFE Enclosure Pairs and drive sets for the upper tier compared to the lower tier.
- For optimal performance configure two frames and four HPFE Enclosure Pairs and spread HPFE enclosures, Drive sets and Host Adapters across both frames.
- With a single frame, configure at least two HPFE Enclosure Pairs.
- Use the Typical Workload selection unless there is a heavy sequential write activity either due to host workload or extensive use of FlashCopy combined with significant host write workload.
- The capacity limits are not intended to be exact, so capacity requirements close to the boundary between two sizes could use either option.
- Use a sizing tool such as StorM.

4.1.3 DS8980F

DS8980F has the following characteristics:

- 2 × IBM Power Systems POWER9 MTM 9009-42A
- 44 POWER9 cores per CEC
- 4,352 GB of System Memory
- From 8 to 128 Host Adapter Ports
- From 16 to 384 Flash drives
- From 12.8 TB to 5,898 TB of capacity (with 15.36 TB Flash Drives)
- 2,300K of Maximum IOps (4K 70/30 R/W mix)
- Maximum of 63 GB/s (Read)/ 32 GB/s (Write) of Sequential Read - Write

The DS8980F will usually be intended for very heavy workloads, and in occasions of bigger consolidation of several older DS8000 into one new DS. As suggested, Starter drive choice, the one for Heavy Workload as shown for the DS8950F in Table 4-2 on page 61, could apply – but again, use a tool like StorM for the final sizing.

4.1.4 Examples using Starter drive choices

Example 1

In the first example shown in Figure 4-2 , we have around 100 TiB of High-Performance Flash drives of 1.6 TB using two HPFEs in the above left picture.

Using the DS8910F Starter drive choices with 100 TB for a typical workload and then the additional custom-placement option (Feature 06060), it is often better to have 1.92 TB drive distributed in 3 HPFEs as it is shown in the above right-hand picture. The slightly higher costs for the additional HPFE are more than offset by the lower price of Flash Tier 2 drives. And especially if the workload profile is heavily sequential, the right-hand alternative would give you even bigger MB/sec throughput, at a lower price.

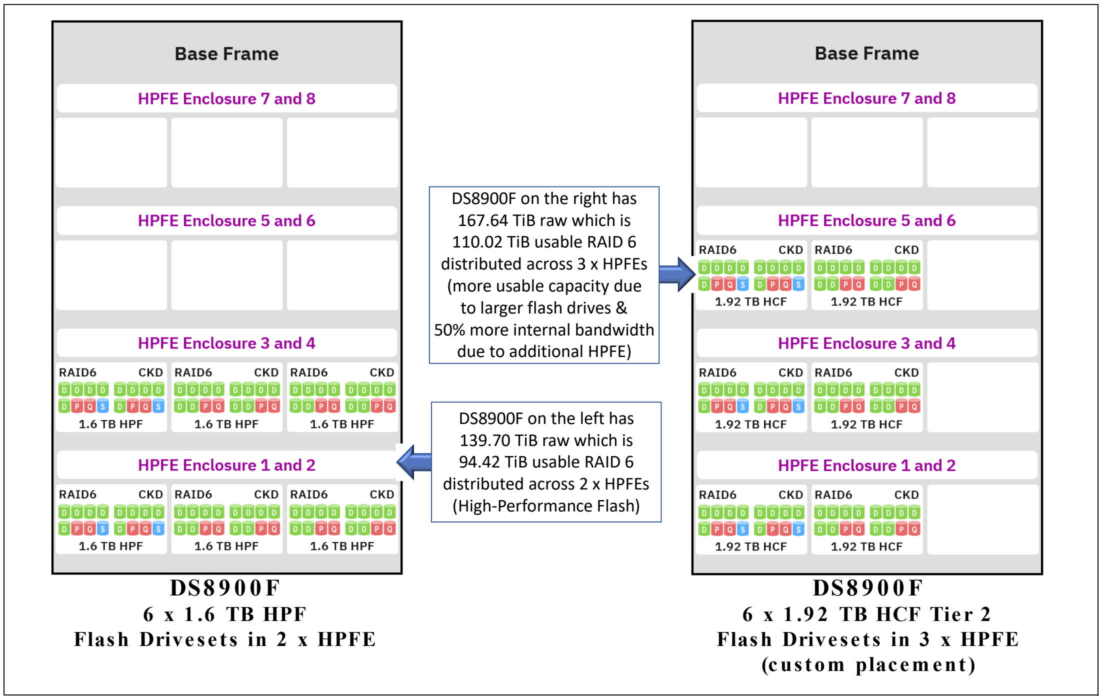

Base Frame
HPFE Enclosure 7 and 8
HPFE Enclosure 5 and 6
HPFE Enclosure 3 and 4
RAID6 CKD
1.6 TB HPF
HPFE Enclosure 1 and 2
1.92 TB HCF
DS8900F on the right has 167.64 TiB raw which is 110.02 TiB usable RAID 6 distributed across 3 x HPFEs (more usable capacity due to larger flash drives & 50% more internal bandwidth due to additional HPFE)
DS8900F on the left has 139.70 TiB raw which is 94.42 TiB usable RAID 6 distributed across 2 x HPFEs (High-Performance Flash)
DS8900F
6 x 1.6 TB HPF
Flash Drivesets in 2 x HPFE
DS8900F
6 x 1.92 TB HCF Tier 2
Flash Drivesets in 3 x HPFE
(custom placement)

Figure 4-2 Example 1: Custom placement with around 100 TiB

Example 2

In the second example, Figure 4-3 , we have around 300 TB being distributed between High-Performance Flash drives of 1.6 TB with High-Capacity Flash of 7.68 TB. The left-hand alternative uses only 2 HPFEs together.

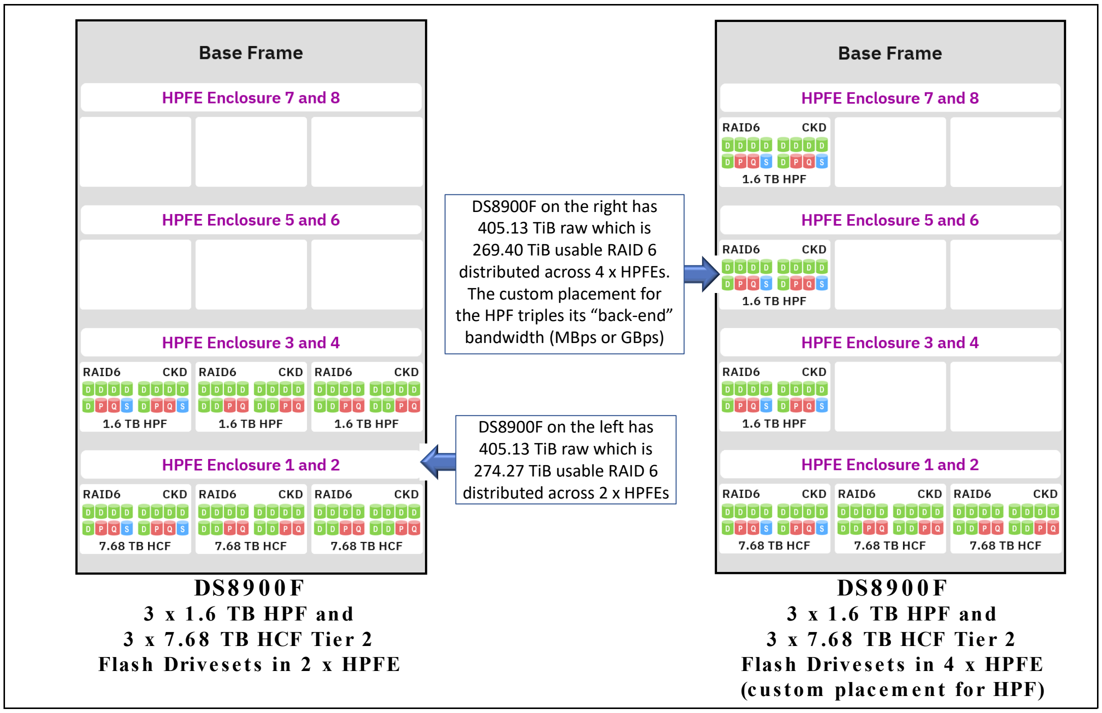

Base Frame
HPFE Enclosure 7 and 8
HPFE Enclosure 5 and 6
HPFE Enclosure 3 and 4
RAID6 CKD
1.6 TB HPF
HPFE Enclosure 1 and 2
7.68 TB HCF
DS8900F on the right has 405.13 TiB raw which is 269.40 TiB usable RAID 6 distributed across 4 x HPFEs. The custom placement for the HPF triples its "back-end" bandwidth (MBps or GBps)
DS8900F on the left has 405.13 TiB raw which is 274.27 TiB usable RAID 6 distributed across 2 x HPFEs
DS8900F
3 x 1.6 TB HPF and
3 x 7.68 TB HCF Tier 2
Flash Drivesets in 2 x HPFE
DS8900F
3 x 1.6 TB HPF and
3 x 7.68 TB HCF Tier 2
Flash Drivesets in 4 x HPFE
(custom placement for HPF)

Figure 4-3 Example 2: Custom placement with around 300 TB

For workloads which are heavily skewed, and in case of very specific workload profiles, too much load may end up on the one HPFE pair serving the 1.6 TB drives. In that specific example and given workload profile, again using the custom-placement option (FC0606) to order additional HPFEs for the upper tier might give a lot more performance, and could justify the small additional price for the HPFEs.

Using the sizing tool with your specific workload profile will tell if such an alternative option is beneficial for your case.

4.1.5 Easy Tier Overview

Many storage environments support a diversity of needs and use disparate technologies that cause storage sprawl. In a large-scale storage infrastructure, this environment yields a suboptimal storage design that can be improved only with a focus on data access characteristics analysis and management to provide optimum performance.

For Performance Analysis, I/O has always been the most tricky resource to analyze and where there are some rules to follow such as:

- Avoid many I/O operations in the same volume or LUN.
- No I/O is the best I/O. Or in other words: make use of all kinds of cache:
  - Memory cache – use of the memory to avoid I/O operations.
  - DS8900F cache – this is the reason for bigger DS8000 caches and better microcode caching techniques. A very high cache-hit ratio is always desirable.
- Placement of high important I/Os or heavy workloads in the Flash High-Performance technology and less important I/Os or near to zero workloads into Flash High-Capacity technology.

Before the Easy Tier technology on DS8000, performance analysts tried to make all kinds of I/O studies (called Seek Analysis) in order to understand on how the I/O operations were distributed, and using this information they allocate the datasets, files, databases into the different disk technologies using the performance objectives.

In a certain way this was very frustrating because some time later (weeks, months) the above behavior changed and they would have to change again the placement of the data.

Easy Tier is a DS8900F microcode technique and when enabled allows a permanent I/O analysis where the hot I/O goes automatically to the best technology and the cold I/O goes to less expensive disks.

4.1.6 IBM System Storage Easy Tier

DS8900F is an All-Flash Array (AFA) and it comes with HPFE Gen2 which is an enclosure that can support both High-Performance Flash Drives and High-Capacity Flash Drives.

The enclosures must be installed in pairs. The HPFEs connect to the I/O enclosures over a PCIe fabric, which increases bandwidth and transaction-processing capability.

Flash Drives types and Tiers:

- High-Performance Flash Drives (Tier 0)
  - 800 GB
  - 1.6 TB
  - 3.2 TB
- High-Capacity Flash Drives

  - 1.92 TB (Tier 2)
  - 3.84 TB (Tier 1)
  - 7.68 TB (Tier 2)
  - 15.36 TB (Tier 2

With dramatically high I/O rates, low response times, and IOPS-energy-efficient characteristics, flash addresses the highest performance needs and also potentially can achieve significant savings in operational costs. It is critical to choose the correct mix of storage tiers and the correct data placement to achieve optimal storage performance and economics across all tiers at a low cost.

With the DS8900F storage system, you can easily implement tiered storage environments that use high-performance flash and high-capacity flash storage tiers. Still, different storage tiers can be isolated to separate extent pools and volume placement can be managed manually across extent pools where required. Or, better and highly encouraged, volume placement can be managed automatically on a subvolume level (extent level) in hybrid extent pools by Easy Tier automatic mode with minimum management effort for the storage administrator. Easy Tier is a no-cost feature on DS8900F storage systems. For more information about Easy Tier, see 1.2.4, “Easy Tier” on page 11.

Consider Easy Tier automatic mode and hybrid or multi-tier extent pools for managing tiered storage on the DS8900F storage system. The overall management and performance monitoring effort increases considerably when manually managing storage capacity and storage performance needs across multiple storage classes and does not achieve the efficiency provided with Easy Tier automatic mode data relocation on the subvolume level (extent level). With Easy Tier, client configurations show less potential to waste flash capacity than with volume-based tiering methods.

With Easy Tier, you can configure hybrid or multi-tier extent pools (mixed high-performance flash/high-capacity flash storage pools) and turn on Easy Tier on. It then provides automated data relocation across the storage tiers and ranks in the extent pool to optimize storage performance and storage economics. It also rebalances the workload across the ranks within each storage tier (auto-rebalance) based on rank utilization to minimize skew and hot spots. Furthermore, it constantly adapts to changing workload conditions. There is no need anymore to bother with tiering policies that must be manually applied to accommodate changing workload dynamics.

In environments with homogeneous system configurations or isolated storage tiers that are bound to different homogeneous extent pools, you can benefit from Easy Tier automatic mode. Easy Tier provides automatic intra-tier performance management by rebalancing the workload across ranks (auto-rebalance) in homogeneous single-tier pools based on rank utilization. Easy Tier automatically minimizes skew and rank hot spots and helps to reduce the overall management effort for the storage administrator.

Depending on the particular storage requirements in your environment, with the DS8900F architecture, you can address a vast range of storage needs combined with ease of management. On a single DS8900F storage system, you can perform these tasks:

- Isolate workloads to selected extent pools (or down to selected ranks and DAs).
- Share resources of other extent pools with different workloads.
- Use Easy Tier to manage automatically multitier extent pools with different storage tiers (or homogeneous extent pools).
- Adapt your logical configuration easily and dynamically at any time to changing performance or capacity needs by migrating volumes across extent pools, merging extent pools, or removing ranks from one extent pool (rank depopulation) and moving them to another pool.

Easy Tier helps you consolidate more workloads onto a single DS8900F storage system by automating storage performance and storage economics management across up to three drive tiers.

For many initial installations, an approach with two extent pools (with or without different storage tiers) and enabled Easy Tier automatic management might be the simplest way to start if you have FB or CKD storage only; otherwise, four extent pools are required. You can plan for more extent pools based on your specific environment and storage needs, for example, workload isolation for some pools, different resource sharing pools for different departments or clients, or specific Copy Services considerations.

Considerations, as described in 4.2, “Configuration principles for optimal performance” on page 66, apply for planning advanced logical configurations in complex environments, depending on your specific requirements.

4.2 Configuration principles for optimal performance

There are three major principles for achieving a logical configuration on a DS8900F storage system for optimal performance when planning extent pools:

- Workload isolation
- Workload resource-sharing
- Workload spreading

Easy Tier provides a significant benefit for mixed workloads, so consider it for resource-sharing workloads and isolated workloads dedicated to a specific set of resources. Furthermore, Easy Tier automatically supports the goal of workload spreading by distributing the workload in an optimum way across all the dedicated resources in an extent pool. It provides automated storage performance and storage economics optimization through dynamic data relocation on extent level across multiple storage tiers and ranks based on their access patterns. With auto-rebalance, it rebalances the workload across the ranks within a storage tier based on utilization to reduce skew and avoid hot spots. Auto-rebalance applies to managed multitier pools and single-tier pools and helps to rebalance the workloads evenly across ranks to provide an overall balanced rank utilization within a storage tier or managed single-tier extent pool. Figure 4-4 shows the effect of auto-rebalance in a single-tier extent pool that starts with a highly imbalanced workload across the ranks at T1. Auto-rebalance rebalances the workload and optimizes the rank utilization over time.

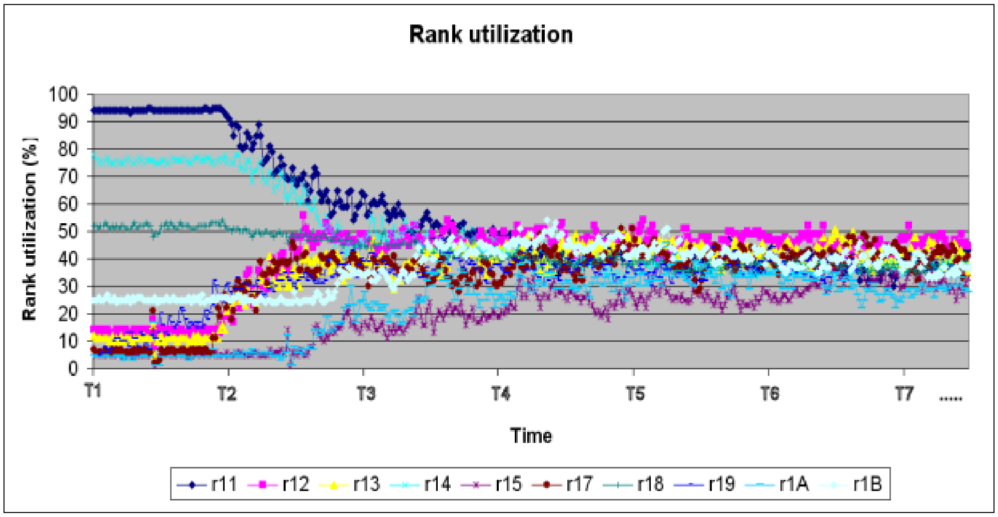


Figure 4-4 Effect of auto-rebalance on individual rank utilization in the system

4.2.1 Workload isolation

With workload isolation, a high-priority workload uses dedicated DS8000 hardware resources to reduce the impact of less important workloads. Workload isolation can also mean limiting a lower-priority workload to a subset of the DS8000 hardware resources so that it does not affect more important workloads by fully using all hardware resources.

Isolation provides ensured availability of the hardware resources that are dedicated to the isolated workload. It removes contention with other applications for those resources.

However, isolation limits the isolated workload to a subset of the total DS8900F hardware so that its maximum potential performance might be reduced. Unless an application has an entire DS8900F storage system that is dedicated to its use, there is potential for contention with other applications for any hardware (such as cache and processor resources) that is not dedicated. Typically, isolation is implemented to improve the performance of certain workloads by separating different workload types.

One traditional practice to isolation is to identify lower-priority workloads with heavy I/O demands and to separate them from all of the more important workloads. You might be able to isolate multiple lower priority workloads with heavy I/O demands to a single set of hardware resources and still meet their lower service-level requirements, particularly if their peak I/O demands are at different times.

Important: For convenience, this chapter sometimes describes isolation as a single isolated workload in contrast to multiple resource-sharing workloads, but the approach also applies to multiple isolated workloads.

DS8900F flash drive capacity isolation

The level of drive capacity isolation that is required for a workload depends on the scale of its I/O demands as compared to the DS8900F array and DA capabilities, and organizational considerations, such as the importance of the workload and application administrator requests for workload isolation.

You can partition the DS8900F flash drive capacity for isolation at several levels:

- Rank level: Certain ranks are dedicated to a workload, that is, volumes for one workload are allocated on these ranks. The ranks can have a different disk type (capacity or speed), a different RAID array type (RAID 5, RAID 6, or RAID 10, arrays with spares or arrays without spares, RAID 6 being the default), or a different storage type (CKD or FB) than the drive types, RAID array types, or storage types that are used by other workloads. Workloads that require different types of the above can dictate rank, extent pool, and address group isolation. You might consider workloads with heavy random activity for rank isolation, for example.
- Extent pool level: Extent pools are logical constructs that represent a group of ranks that are serviced by storage server 0 or storage server 1. You can isolate different workloads to different extent pools, but you always must be aware of the rank and DA pair (HPFE) associations. Although physical isolation on rank and DA/HPFE level involves building appropriate extent pools with a selected set of ranks or ranks from a specific HPFE/DA pair, different extent pools with a subset of ranks from different DA pairs still can share DAs. Isolated workloads to different extent pools might share a DA adapter as a physical resource, which can be a potential limiting physical resource under certain extreme conditions. However, given the capabilities of one DA or HPFE, some mutual interference there is rare, and isolation (if needed) on extent pool level is effective if the workloads are disk-bound.
- DA (HPFE) level: All ranks on one or more DA pairs or a HPFE pair are dedicated to a workload, that is, only volumes for this workload are allocated on the ranks that are associated with one or more DAs. These ranks can be a different disk type (capacity or speed), RAID array type (RAID 5, RAID 6, or RAID 10, arrays with spares or arrays without spares), or storage type (CKD or FB) than the disk types, RAID types, or storage types that are used by other workloads. Consider only huge (multiple GBps) workloads with heavy, large blocksize, and sequential activity for DA-level/HPFE-level isolation because these workloads tend to consume all of the available DA resources.
- Processor complex level: All ranks that are assigned to extent pools managed by processor complex 0 or all ranks that are assigned to extent pools managed by processor complex 1 are dedicated to a workload. This approach is not preferable because it can reduce the processor and cache resources and the back-end bandwidth that is available to the workload by 50%.
- Storage unit level: All ranks in a physical DS8900F storage system are dedicated to a workload, that is, the physical DS8900F storage system runs one workload.
- Pinning volumes to a tier in hybrid pools: This level is not complete isolation, but it can be a preferred way to go in many cases. Even if you have just one pool pair with two large three-tier pools, you may use the High-Performance flash tier only for some volumes and the High-Capacity flash tier only for others. Certain volumes may float only among ranks of the upper tier or between the upper tiers only, but lower tiers are excluded for them, and the other way round for other volumes. Such a setup allows a simple extent pool structure, and at the same time less-prioritized volumes stay on lower tiers only and highly prioritized volumes are treated preferentially. The way you do such pinning could also be changed at any time.

DS8000 host connection isolation

The level of host connection isolation that is required for a workload depends on the scale of its I/O demands as compared to the DS8900F I/O port and host adapter (HA) capabilities. It also depends on organizational considerations, such as the importance of the workload and administrator requests for workload isolation.

The DS8900F host connection subsetting for isolation can also be done at several levels:

- I/O port level: Certain DS8900F I/O ports are dedicated to a workload, which is a common case. Workloads that require Fibre Channel connection (FICON) and Fibre Channel Protocol (FCP) must be isolated at the I/O port level anyway because each I/O port on a FCP/FICON-capable HA card can be configured to support only one of these protocols. Although Open Systems host servers and remote mirroring links use the same protocol (FCP), they are typically isolated to different I/O ports. You must also consider workloads with heavy large-block sequential activity for HA isolation because they tend to consume all of the I/O port resources that are available to them.
- HA level: Certain HAs are dedicated to a workload. FICON and FCP workloads do not necessarily require HA isolation because separate I/O ports on the same FCP/FICON-capable HA card can be configured to support each protocol (FICON or FCP). However, it is a preferred practice to separate FCP and FICON to different HBAs. Furthermore, host connection requirements might dictate a unique type of HA card (longwave (LW) or shortwave (SW)) for a workload. Workloads with heavy large-block sequential activity must be considered for HA isolation because they tend to consume all of the I/O port resources that are available to them.
- I/O enclosure level: Certain I/O enclosures are dedicated to a workload. This approach is not necessary.

4.2.2 Workload resource-sharing

Workload resource-sharing means multiple workloads use a common set of the DS8900F hardware resources:

- Ranks
- DAs and HPFEs
- I/O ports
- HAs

Multiple resource-sharing workloads can have logical volumes on the same ranks and can access the same DS8900F HAs or I/O ports. Resource-sharing allows a workload to access more DS8900F hardware than can be dedicated to the workload, providing greater potential performance, but this hardware sharing can result in resource contention between applications that impacts overall performance at times. It is important to allow resource-sharing only for workloads that do not consume all of the DS8900F hardware resources that are available to them. Pinning volumes to one certain tier can also be considered temporarily, and then you can release these volumes again.

Easy Tier extent pools typically are shared by multiple workloads because Easy Tier with its automatic data relocation and performance optimization across multiple storage tiers provides the most benefit for mixed workloads.

To better understand the resource-sharing principle for workloads on disk arrays, see 3.2.3, “Extent pool considerations” on page 51.

4.2.3 Workload spreading

Workload spreading means balancing and distributing overall workload evenly across all of the DS8900F hardware resources that are available:

- Processor complex 0 and processor complex 1
- HPFEs and DAs

- Ranks
- I/O enclosures
- HAs

Spreading applies to both isolated workloads and resource-sharing workloads.

You must allocate the DS8900F hardware resources to either an isolated workload or multiple resource-sharing workloads in a balanced manner, that is, you must allocate either an isolated workload or resource-sharing workloads to the DS8900F ranks that are assigned to DAs and both processor complexes in a balanced manner. You must allocate either type of workload to I/O ports that are spread across HAs and I/O enclosures in a balanced manner.

You must distribute volumes and host connections for either an isolated workload or a resource-sharing workload in a balanced manner across all DS8900F hardware resources that are allocated to that workload.

You should create volumes as evenly distributed as possible across all ranks and DAs allocated to those workloads.

One exception to the recommendation of spreading volumes might be when specific files or data sets are never accessed simultaneously, such as multiple log files for the same application where only one log file is in use at a time. In that case, you can place the volumes required by these data sets or files on the same resources.

You must also configure host connections as evenly distributed as possible across the I/O ports, HAs, and I/O enclosures that are available to either an isolated or a resource-sharing workload. Then, you can use host server multipathing software to optimize performance over multiple host connections. For more information about multipathing software, see Chapter 8, “Host attachment” on page 187.

4.2.4 Using workload isolation, resource-sharing, and spreading

When you perform DS8900F performance optimization, you must first identify any workload that has the potential to negatively impact the performance of other workloads by fully using all of the DS8900F I/O ports and the DS8900F ranks available to it.

Additionally, you might identify any workload that is so critical that its performance can never be allowed to be negatively impacted by other workloads.

Then, identify the remaining workloads that are considered appropriate for resource-sharing.

Next, define a balanced set of hardware resources that can be dedicated to any isolated workloads, if required. Then, allocate the remaining DS8000 hardware for sharing among the resource-sharing workloads. Carefully consider the appropriate resources and storage tiers for Easy Tier and multitier extent pools in a balanced manner.

The next step is planning extent pools and assigning volumes and host connections to all workloads in a way that is balanced and spread. By default, the standard allocation method when creating volumes is stripes with one-extent granularity across all arrays in a pool, so on the rank level, this distribution is done automatically.

Without the explicit need for workload isolation or any other requirements for multiple extent pools, starting with two extent pools (with or without different storage tiers) and a balanced distribution of the ranks and DAs/HPFEs might be the simplest configuration to start with using resource-sharing throughout the whole DS8900F storage system and Easy Tier automatic management if you have either FB or CKD storage. Otherwise, four extent pools are required for a reasonable minimum configuration, two for FB storage and two for CKD storage, and each pair is distributed across both DS8900F storage servers.

The final step is the implementation of host-level striping (when appropriate) and multipathing software, if needed. If you planned for Easy Tier, do not consider host-level striping because it dilutes the workload skew and is counterproductive to the Easy Tier optimization.

4.3 Analyzing application workload characteristics

The first and most important step in creating a successful logical configuration for the DS8900F storage system is analyzing the workload characteristics for the applications that access the DS8900F storage system. The DS8900F hardware resources, such as RAID arrays and I/O ports, must be correctly allocated to workloads for isolation and resource-sharing considerations. If planning for shared multitier configurations and Easy Tier, it is important to determine the skew of the workload and plan for the amount of required storage capacity on the appropriate storage tiers. You must perform this workload analysis during the DS8900F capacity planning process, and you must complete it before ordering the DS8000 hardware.

4.3.1 Determining skew and storage requirements for Easy Tier

For Easy Tier configurations, it is important to determine the *skew* of the workload and plan for the amount of required storage capacity on each of the storage tiers. Plan the optimum initial hardware configuration for managed multitier environments so that you determine the overall distribution of the I/O workload against the amount of data (data heat distribution) to understand how much of the data is doing how much (or most) of the I/O workload. The workload pattern, small block random or large block sequential read/write operations, also is important. A good understanding of the workload heat distribution and skew helps to evaluate the benefit of an Easy Tier configuration.

For example, the ratio of High-Performance flash capacity to High-Capacity flash capacity in a hybrid pool depends on the workload characteristics and skew.

For a two-tier configuration always configure an equal or greater number of HPFE Enclosure Pairs and drive sets for the upper tier compared to the lower tier. You should not use less drives of High-Performance flash (Tier 0) than of High-Capacity flash (Tier 2). For instance, mixing High-Performance flash of 3.2 TB with High-Capacity flash of 15.36 TB under these conditions, you already come to almost 20% of net capacity in Flash Tier 0.

The DS8900F Storage Management GUI also can provide guidance for capacity planning of the available storage tiers based on the existing workloads on a DS8900F storage system with Easy Tier monitoring enabled.

4.3.2 Determining isolation requirements

The objective of this analysis is to identify workloads that require isolated (dedicated) DS8900F hardware resources because this determination ultimately affects the total amount of disk capacity that is required and the total number of disk drive types that is required, and the number and type of HAs that is required. The result of this first analysis indicates which workloads require isolation and the level of isolation that is required.

You must also consider organizational and business considerations in determining which workloads to isolate. Workload priority (the importance of a workload to the business) is a key consideration. Application administrators typically request dedicated resources for high priority workloads. For example, certain database online transaction processing (OLTP) workloads might require dedicated resources to ensure service levels.

The most important consideration is preventing lower-priority workloads with heavy I/O requirements from impacting higher priority workloads. Lower-priority workloads with heavy random activity must be evaluated for rank isolation. Lower-priority workloads with heavy, large blocksize, and sequential activity must be evaluated for I/O port, and eventually DA (HPFE) isolation.

Workloads that require different disk drive types (capacity and speed), different RAID types (RAID 5, RAID 6, or RAID 10), or different storage types (CKD or FB) dictate isolation to different DS8000 arrays, ranks, and extent pools, unless this situation can be solved by pinning volumes to one certain tier. For more information about the performance implications of various RAID types, see “RAID-level performance considerations” on page 77.

Workloads that use different I/O protocols (FCP or FICON) dictate isolation to different I/O ports. However, workloads that use the same drive types, RAID type, storage type, and I/O protocol can be evaluated for separation or isolation requirements.

Workloads with heavy, continuous I/O access patterns must be considered for isolation to prevent them from consuming all available DS8900F hardware resources and impacting the performance of other types of workloads. Workloads with large blocksize and sequential activity can be considered for separation from those workloads with small blocksize and random activity.

Isolation of only a few workloads that are known to have high I/O demands can allow all the remaining workloads (including the high-priority workloads) to share hardware resources and achieve acceptable levels of performance. More than one workload with high I/O demands might be able to share the isolated DS8900F resources, depending on the service level requirements and the times of peak activity.

The following examples are I/O workloads, files, or data sets that might have heavy and continuous I/O access patterns:

- Sequential workloads (especially those workloads with large-blocksize transfers)
- Log files or data sets
- Sort or work data sets or files
- Business Intelligence and Data Mining
- Disk copies (including Point-in-Time Copy background copies, remote mirroring target volumes, and tape simulation on disk)
- Video and imaging applications
- Engineering and scientific applications
- Certain batch workloads

You must consider workloads for all applications for which DS8900F storage is allocated, including current workloads to be migrated from other installed storage systems and new workloads that are planned for the DS8900F storage system. Also, consider projected growth for both current and new workloads.

For existing applications, consider historical experience first. For example, is there an application where certain data sets or files are known to have heavy, continuous I/O access patterns? Is there a combination of multiple workloads that might result in unacceptable performance if their peak I/O times occur simultaneously? Consider workload importance (workloads of critical importance and workloads of lesser importance).

For existing applications, you can also use performance monitoring tools that are available for the existing storage systems and server platforms to understand current application workload characteristics:

- Read/write ratio
- Random/sequential ratio
- Average transfer size (blocksize)
- Peak workload (IOPS for random access and MB per second for sequential access)
- Peak workload periods (time of day and time of month)
- Copy Services requirements (Point-in-Time Copy and Remote Mirroring)
- Host connection utilization and throughput (FCP host connections and FICON)
- Remote mirroring link utilization and throughput

Estimate the requirements for new application workloads and for current application workload growth. You can obtain information about general workload characteristics in Chapter 5, “Understanding your workload” on page 107.

As new applications are rolled out and current applications grow, you must monitor performance and adjust projections and allocations. You can obtain more information about this topic in Chapter 7, “Practical performance management” on page 139.

You can use the StorM modeling tool to model the current or projected workload and estimate the required DS8900F hardware resources. They are described in Chapter 6, “Performance planning tools” on page 125.

The DS8900F Storage Management GUI can also provide workload information and capacity planning recommendations that are associated with a specific workload to reconsider the need for isolation and evaluate the potential benefit when using a multitier configuration and Easy Tier.

4.3.3 Reviewing remaining workloads for feasibility of resource-sharing

After workloads with the highest priority or the highest I/O demands are identified for isolation, the I/O characteristics of the remaining workloads must be reviewed to determine whether a single group of resource-sharing workloads is appropriate, or whether it makes sense to split the remaining applications into multiple resource-sharing groups. The result of this step is the addition of one or more groups of resource-sharing workloads to the DS8900F configuration plan.

4.4 Planning allocation of disk and host connection capacity

You must plan the allocation of specific DS8900F hardware first for any isolated workload, and then for the resource-sharing workloads, including Easy Tier hybrid pools. Use the workload analysis in 4.3.2, “Determining isolation requirements” on page 71 to define the flash capacity and host connection capacity that is required for the workloads. For any workload, the required flash capacity is determined by both the amount of space that is needed for data and the number of arrays (of a specific speed) that are needed to provide the needed level of performance. The result of this step is a plan that indicates the number of ranks (including disk drive type) and associated DAs and the number of I/O adapters and associated I/O enclosures that are required for any isolated workload and for any group of resource-sharing workloads.

Planning DS8000 hardware resources for isolated workloads

For the DS8900F flash allocation, isolation requirements might dictate the allocation of certain individual ranks or all of the ranks on certain DAs to one workload. For the DS8900F I/O port allocation, isolation requirements might dictate the allocation of certain I/O ports or all of the I/O ports on certain HAs to one workload.

Choose the DS8900F resources to dedicate in a balanced manner. If ranks are planned for workloads in multiples of two, half of the ranks can later be assigned to extent pools managed by processor complex 0, and the other ranks can be assigned to extent pools managed by processor complex 1. You may also note the DAs and HPFEs to be used. If I/O ports are allocated in multiples of four, they can later be spread evenly across all I/O enclosures in a DS8900F frame if four or more HA cards are installed. If I/O ports are allocated in multiples of two, they can later be spread evenly across left and right I/O enclosures.

Planning DS8900F hardware resources for resource-sharing workloads

Review the DS8900F resources to share for balance. If ranks are planned for resource-sharing workloads in multiples of two, half of the ranks can later be assigned to processor complex 0 extent pools, and the other ranks can be assigned to processor complex 1 extent pools. If I/O ports are allocated for resource-sharing workloads in multiples of four, they can later be spread evenly across all I/O enclosures in a DS8900F frame if four or more HA cards are installed. If I/O ports are allocated in multiples of two, they can later be spread evenly across left and right I/O enclosures.

Easy Tier later provides automatic intra-tier management in single-tier and multitier pools (auto-rebalance) and cross-tier management in multitier pools for the resource-sharing workloads.

4.5 Planning volume and host connection spreading

After hardware resources are allocated for both isolated and resource-sharing workloads, plan the volume and host connection spreading for all of the workloads.

> **Host connection:** In this chapter, we use host connection in a general sense to represent a connection between a host server (either z Operating Systems or Open Systems) and the DS8000 storage system.

The result of this step is a plan that includes this information:

- The specific number and size of volumes for each isolated workload or group of resource-sharing workloads and how they are allocated to ranks and DAs
- The specific number of I/O ports for each workload or group of resource-sharing workloads and how they are allocated to HAs and I/O enclosures

After the spreading plan is complete, use the DS8900F hardware resources that are identified in the plan as input to order the DS8900F hardware.

4.5.1 Spreading volumes for isolated and resource-sharing workloads

Now, consider the requirements of each workload for the number and size of logical volumes. For a specific amount of required disk capacity from the perspective of the DS8900F storage system, there are typically no significant DS8900F performance implications of using more small volumes as compared to fewer large volumes. However, using one or a few standard volume sizes can simplify management.

However, there are host server performance considerations related to the number and size of volumes. For example, for z Systems servers, the number of PAVs that are needed can vary with volume size. For more information about PAVs, see 10.2, “DS8000 and z Systems planning and configuration” on page 245.

There also can be Open Systems host server or multipathing software considerations that are related to the number or the size of volumes, so you must consider these factors in addition to workload requirements.

There are significant performance implications with the assignment of logical volumes to ranks and DAs. The goal of the entire logical configuration planning process is to ensure that volumes for each workload are on ranks and DAs that allow all workloads to meet performance objectives.

To spread volumes across allocated hardware for each isolated workload, and then for each workload in a group of resource-sharing workloads, complete the following steps:

1. Review the required number and the size of the logical volumes that are identified during the workload analysis.
2. Review the number of ranks that are allocated to the workload (or group of resource-sharing workloads) and the associated DA pairs.
3. Evaluate the use of multi-rank or multitier extent pools. Evaluate the use of Easy Tier in automatic mode to automatically manage data placement and performance.
4. Assign the volumes, preferably with the default allocation method rotate extents (DSCLI term: rotateexts, GUI: rotate capacity).

4.5.2 Spreading host connections for isolated and resource-sharing workloads

Next, consider the requirements of each workload for the number and type of host connections. In addition to workload requirements, you also might need to consider the host server or multipathing software in relation to the number of host connections. For more information about multipathing software, see Chapter 8, “Host attachment” on page 187.

There are significant performance implications from the assignment of host connections to I/O ports, HAs, and I/O enclosures. The goal of the entire logical configuration planning process is to ensure that host connections for each workload access I/O ports and HAs that allow all workloads to meet the performance objectives.

To spread host connections across allocated hardware for each isolated workload, and then for each workload in a group of resource-sharing workloads, complete the following steps:

1. Review the required number and type (SW, LW, FCP, or FICON) of host connections that are identified in the workload analysis. You must use a minimum of *two* host connections to different DS8900F HA cards to ensure availability. Some Open Systems hosts might impose limits on the number of paths and volumes. In such cases, you might consider not exceeding four paths per volume, which in general is a good approach for performance and availability. The DS8900F front-end host ports are 16 Gbps and 32 Gbps capable and if the expected workload is not explicitly saturating the adapter and port bandwidth with high sequential loads, you might share ports with many hosts.
2. Review the HAs that are allocated to the workload (or group of resource-sharing workloads) and the associated I/O enclosures.
3. Review requirements that need I/O port isolation, for example, remote replication Copy Services, or SAN Volume Controller. If possible, try to split them as you split hosts among

hardware resources. Do not mix them with other Open Systems because they can have different workload characteristics.

4. Assign each required host connection to a different HA in a different I/O enclosure if possible, balancing them across the left and right I/O enclosures:
   - If the required number of host connections is less or equal than the available number of I/O enclosures (which can be typical for certain Open Systems servers), assign an equal number of host connections to left I/O enclosures (0, 2, 4, and 6) as to right I/O enclosures (1, 3, 5, and 7).
   - Within an I/O enclosure, assign each required host connection to the HA of the required type (SW FCP/FICON-capable or LW FCP/FICON-capable) with the greatest number of unused ports. When HAs have an equal number of unused ports, assign the host connection to the adapter that has the fewest connections for this workload.
   - If the number of required host connections is greater than the number of I/O enclosures, assign the additional connections to different HAs with the most unused ports within the I/O enclosures. When HAs have an equal number of unused ports, assign the host connection to the adapter that has the fewest connections for this workload.

4.6 Planning array sites

During the DS8900F installation, array sites are dynamically created and assigned to HPFEs and their DA pairs. Array site IDs (S*x*) do not have any fixed or predetermined relationship to drive physical locations or to the flash enclosure installation order. The relationship between array site IDs and physical drive locations or DA assignments can differ between the DS8900F storage systems, even on the DS8900F storage systems with the same number and type of flash drives.

When using the DS Storage Manager GUI to create managed arrays and pools, the GUI automatically chooses a good distribution of the arrays across all DAs, and initial formatting with the GUI gives optimal results for many cases. Only for specific requirements (for example, isolation by DA pairs) is the command-line interface (DSCLI) advantageous because it gives more options for a certain specific configuration.

After the DS8900F hardware is installed, you can use the output of the DS8000 DSCLI `lsarraysite` command to display and document array site information, including flash drive type and DA pair. Check the disk drive type and DA pair for each array site to ensure that arrays, ranks, and ultimately volumes that are created from the array site are created on the DS8900F hardware resources required for the isolated or resource-sharing workloads.

The result of this step is the addition of specific array site IDs to the plan of workload assignment to ranks.

4.7 Planning RAID arrays and ranks

The next step is planning the RAID arrays and ranks. When using DSCLI, take the specific array sites that are planned for isolated or resource-sharing workloads and define their assignments to RAID arrays and CKD or FB ranks, and thus define array IDs and rank IDs. Because there is a one-to-one correspondence between an array and a rank on the DS8900F storage system, you can plan arrays and ranks in a single step. However, array creation and rank creation require separate steps.

The sequence of steps when creating the arrays and ranks with the DSCLI finally determines the numbering scheme of array IDs and rank IDs because these IDs are chosen automatically by the system during creation. The logical configuration does not depend on a specific ID numbering scheme, but a specific ID numbering scheme might help you plan the configuration and manage performance more easily.

Storage servers: Array sites, arrays, and ranks do not have a fixed or predetermined relationship to any DS8900F processor complex (storage server) before they are finally assigned to an extent pool and a rank group (rank group 0/1 is managed by processor complex 0/1).

RAID-level performance considerations

The available drive options provide industry-class capacity and performance to address a wide range of business requirements. The DS8900F storage arrays can be configured as RAID 6, RAID 10, or RAID 5, depending on the drive type.

RAID 6 is now the default and preferred setting for the DS8900F. RAID 5 can be configured for drives of less than 1 TB, but this configuration is not preferred and requires a risk acceptance. Flash Tier 0 drive sizes larger than 1 TB can be configured by using RAID 5, but require an RPQ and an internal control switch to be enabled. RAID 10 continues to be an option for all drive types.

When configuring arrays from array sites, you must specify the RAID level, either RAID 5, RAID 6, or RAID 10. These RAID levels meet different requirements for performance, usable storage capacity, and data protection. However, you must determine the correct RAID types and the physical flash drives (speed and capacity) that are related to initial workload performance objectives, capacity requirements, and availability considerations before you order the DS8900F hardware.

Each HPFE Gen2 pair can contain up to six array sites. The first set of 16 flash drives creates two 8-flash drive array sites. RAID 6 arrays are created by default on each array site.

During logical configuration, RAID 6 arrays and the required number of spares are created. Each HPFE Gen2 pair has two global spares, created from the first increment of 16 flash drives. The first two arrays to be created from these array sites are 5+P+Q. Subsequent RAID 6 arrays in the same HPFE Gen2 Pair will be 6+P+Q.

RAID 5 had been for long one of the most commonly used levels of RAID protection because it optimizes cost-effective performance while emphasizing usable capacity through data striping. It provides fault tolerance if one disk drive fails by using XOR parity for redundancy. Hot spots within an array are avoided by distributing data and parity information across all of the drives in the array. The capacity of one drive in the RAID array is lost because it holds the parity information. RAID 5 provides a good balance of performance and usable storage capacity.

RAID 6 provides a higher level of fault tolerance than RAID 5 in disk failures, but also provides less usable capacity than RAID 5 because the capacity of two drives in the array is set aside to hold the parity information. As with RAID 5, hot spots within an array are avoided by distributing data and parity information across all of the drives in the array. Still, RAID 6 offers more usable capacity than RAID 10 by providing an efficient method of data protection in double disk errors, such as two drive failures, two coincident medium errors, or a drive failure and a medium error during a rebuild. Because the likelihood of media errors increases with the capacity of the physical disk drives, consider the use of RAID 6 with large capacity disk drives and higher data availability requirements. For example, consider RAID 6 where rebuilding the array in a drive failure takes a long time.

RAID 6 arrays are created by default on each array site.

RAID 10 optimizes high performance while maintaining fault tolerance for disk drive failures. The data is striped across several disks, and the first set of disk drives is mirrored to an identical set. RAID 10 can tolerate at least one, and in most cases, multiple disk failures if the primary copy and the secondary copy of a mirrored disk pair do not fail at the same time.

Regarding random-write I/O operations, the different RAID levels vary considerably in their performance characteristics. With RAID 10, each write operation at the disk back end initiates two disk operations to the rank. With RAID 5, an individual random small-block write operation to the disk back end typically causes a RAID 5 write penalty, which initiates four I/O operations to the rank by reading the old data and the old parity block before finally writing the new data and the new parity block. For RAID 6 with two parity blocks, the write penalty increases to six required I/O operations at the back end for a single random small-block write operation. This assumption is a worst-case scenario that is helpful for understanding the back-end impact of random workloads with a certain read/write ratio for the various RAID levels. It permits a rough estimate of the expected back-end I/O workload and helps you to plan for the correct number of arrays. On a heavily loaded system, it might take fewer I/O operations than expected on average for RAID 5 and RAID 6 arrays. The optimization of the queue of write I/Os waiting in cache for the next destage operation can lead to a high number of partial or full stripe writes to the arrays with fewer required back-end disk operations for the parity calculation.

On modern disk systems, such as the DS8900F storage system, write operations are cached by the storage subsystem and thus handled asynchronously with short write response times for the attached host systems. So, any RAID 5 or RAID 6 write penalties are shielded from the attached host systems in disk response time. Typically, a write request that is sent to the DS8900F subsystem is written into storage server cache and persistent cache, and the I/O operation is then acknowledged immediately to the host system as complete. If there is free space in these cache areas, the response time that is seen by the application is only the time to get data into the cache, and it does not matter whether RAID 5, RAID 6, or RAID 10 is used.

There is also the concept of rewrites. If you update a cache segment that is still in write cache and not yet destaged, update segment in the cache and eliminate the RAID penalty for the previous write step. However, if the host systems send data to the cache areas faster than the storage server can destage the data to the arrays (that is, move it from cache to the physical disks), the cache can occasionally fill up with no space for the next write request. Therefore, the storage server signals the host system to retry the I/O write operation. In the time that it takes the host system to retry the I/O write operation, the storage server likely can destage part of the data, which provides free space in the cache and allows the I/O operation to complete on the retry attempt.

RAID 10 is not as commonly used as RAID 5 or RAID 6 for the following reason: RAID 10 requires more raw disk capacity for every TB of effective capacity.

Table 4-3 shows a short overview of the advantages and disadvantages for the RAID level reliability, space efficiency, and random write performance.

Table 4-3 RAID-level comparison of reliability, space efficiency, and write penalty

RAID level	Reliability (number of erasures)	Space efficiency[a]	Performance write penalty (number of disk operations)
RAID 5 (7+P)	1	87.5%	4

RAID level	Reliability (number of erasures)	Space efficiency[a]	Performance write penalty (number of disk operations)
RAID 6 (6+P+Q)	2	75%	6
RAID 10 (4x2)	At least 1	50%	2

[a] The space efficiency in this table is based on the number of disks that remain available for data storage. The actual usable decimal capacities are up to 5% less.

Because RAID 5, RAID 6, and RAID 10 perform equally well for both random and sequential *read* operations, RAID 5 or RAID 6 might be a good choice for space efficiency and performance for standard workloads with many read requests. RAID 6 offers a higher level of data protection than RAID 5, especially for large capacity drives.

For array rebuilds, RAID 5, RAID 6, and RAID 10 require approximately the same elapsed time, although RAID 5 and RAID 6 require more disk operations and therefore are more likely to affect other disk activity on the same disk array.

Today, High-Capacity flash drives are mostly used as some Tier 1 (High-Capacity flash of 3.84 TB) and lower tiers in a hybrid pool where most of the IOPS are handled by Tier 0 High-Performance flash drives. Yet, even if the High-Performance flash tier might handle, for example, 70% and more of the load, the High-Capacity flash drives still handle a considerable workload amount because of their large bulk capacity.

Finally, using the **lsarray -l**, and **lsrank -l** commands can give you an idea of which DA pair (HPFE) is used by each array and rank respectively, as shown in Example 4-1. You can do further planning from here.

Example 4-1 lsarray -l and lsrank -l commands showing Array ID sequence and DA pair

```
dscli> lsarray -l
Array State    Data   RAIDtype    arsite Rank DA Pair DDMcap (10^9B) diskclass  encrypt
=========================================================================================
A0    Assigned Normal 6 (6+P+Q)   S6     R0   11             1600.0 FlashTier0 supported
A1    Assigned Normal 6 (5+P+Q+S) S2     R1   10             7680.0 FlashTier2 supported
A2    Assigned Normal 6 (6+P+Q)   S5     R2   11             1600.0 FlashTier0 supported
A3    Assigned Normal 6 (5+P+Q+S) S1     R3   10             7680.0 FlashTier2 supported
A4    Assigned Normal 6 (5+P+Q+S) S3     R4   11             1600.0 FlashTier0 supported
A5    Assigned Normal 6 (5+P+Q+S) S4     R5   11             1600.0 FlashTier0 supported

dscli> lsrank -l
ID Group State  datastate Array RAIDtype extpoolID extpoolnam stgtype exts    usedexts keygrp marray extsi
===========================================================================================================
R0     0 Normal Normal    A0           6 P0        Open_ET    fb      554647   547943      1 MA6    16MiB
R1     0 Normal Normal    A1           6 P0        Open_ET    fb      2219583  276917      1 MA2    16MiB
R2     1 Normal Normal    A2           6 P1        Open_ET    fb      554647   553725      1 MA5    16MiB
R3     1 Normal Normal    A3           6 P1        Open_ET    fb      2219583 1840518      1 MA1    16MiB
R4     0 Normal Normal    A4           6 P2        CKD_z/OS   ckd     429258    35783      1 MA3    21cyl
R5     1 Normal Normal    A5           6 P3        CKD_z/OS   ckd     429258    39931      1 MA4    21cyl
```

4.8 Planning extent pools

After planning the arrays and the ranks, the next step is to plan the extent pools, which means taking the planned ranks and defining their assignment to extent pools and rank groups, including planning the extent pool IDs.

Extent pools are automatically numbered with system-generated IDs starting with P0, P1, and P2 in the sequence in which they are created. Extent pools that are created for rank group 0 are managed by processor complex 0 and have even-numbered IDs (P0, P2, and P4, for example). Extent pools that are created for rank group 1 are managed by processor complex 1 and have odd-numbered IDs (P1, P3, and P5, for example). Only in a failure condition or during a concurrent code load is the ownership of a certain rank group temporarily moved to the alternative processor complex.

To achieve a uniform storage system I/O performance and avoid single resources that become bottlenecks (called hot spots), it is preferable to distribute volumes and workloads evenly across all of the ranks and DA pairs that are dedicated to a workload by creating appropriate extent pool configurations.

The assignment of the ranks to extent pools together with an appropriate concept for the logical configuration and volume layout is the most essential step to optimize overall storage system performance. A rank can be assigned to any extent pool or rank group. Each rank provides a particular number of storage extents of a certain storage type (either FB or CKD) to an extent pool. Finally, an extent pool aggregates the extents from the assigned ranks and provides the logical storage capacity for the creation of logical volumes for the attached host systems.

On the DS8900F storage system, you can configure homogeneous single-tier extent pools with ranks of the same storage class, and hybrid multitier extent pools with ranks from different storage classes. The Extent Allocation Methods (EAM), such as rotate extents or storage-pool striping, provide easy-to-use capacity-based methods of spreading the workload data across the ranks in an extent pool. Furthermore, the use of Easy Tier automatic mode to automatically manage and maintain an optimal workload distribution across these resources over time provides excellent workload spreading with the best performance at a minimum administrative effort.

The following sections present concepts for the configuration of single-tier and multitier extent pools to spread the workloads evenly across the available hardware resources. Also, the benefits of Easy Tier with different extent pool configurations are outlined. Unless otherwise noted, assume that enabling Easy Tier automatic mode refers to enabling the automatic management capabilities of Easy Tier and Easy Tier monitoring.

4.8.1 Single-tier extent pools

Single-tier or homogeneous extent pools are pools that contain ranks from only one storage class or storage tier:

- HPFE High-Performance Flash
- HPFE High-Capacity Flash

Single-tier extent pools consist of one or more ranks that can be referred to as single-rank or multi-rank extent pools.

Single-rank extent pools

With single-rank extent pools, there is a direct relationship between volumes and ranks based on the extent pool of the volume. This relationship helps you manage and analyze performance down to rank level more easily, especially with host-based tools, such as IBM Resource Measurement Facility (IBM RMF) on z Systems in combination with a hardware-related assignment of LSS/LCU IDs. However, the administrative effort increases considerably because you must create the volumes for a specific workload in multiple steps from each extent pool separately when distributing the workload across multiple ranks.

With single-rank extent pools, you choose a configuration design that limits the capabilities of a created volume to the capabilities of a single rank for capacity and performance. A single volume cannot exceed the capacity or the I/O performance provided by a single rank. So, for demanding workloads, you must create multiple volumes from enough ranks from different extent pools and use host-level-based striping techniques, such as volume manager striping, to spread the workload evenly across the ranks dedicated to a specific workload. You are also likely to waste storage capacity easily if extents remain left on ranks in different extent pools because a single volume can be created only from extents within a single extent pool, not across extent pools.

Furthermore, you benefit less from the advanced DS8000 virtualization features, such as dynamic volume expansion (DVE), storage pool striping, Easy Tier automatic performance management, and workload spreading, which use the capabilities of multiple ranks within a single extent pool.

Single-rank extent pools are selected for environments where isolation or management of volumes on the rank level is needed, such as in some z/OS environments. Single-rank extent pools are selected for configurations by using storage appliances, such as the SAN Volume Controller, where the selected RAID arrays are provided to the appliance as simple back-end storage capacity and where the advanced virtualization features on the DS8000 storage system are not required or not wanted to avoid multiple layers of data striping. However, the use of homogeneous multi-rank extent pools and storage pool striping to minimize the storage administrative effort by shifting the performance management from the rank to the extent pool level and letting the DS8000 storage system maintain a balanced data distribution across the ranks within a specific pool is popular. It provides excellent performance in relation to the reduced management effort.

Also, you do not need to strictly use *only* single-rank extent pools or *only* multi-rank extent pools on a storage system. You can base your decision on individual considerations for each workload group that is assigned to a set of ranks and thus extent pools. The decision to use single-rank and multi-rank extent pools depends on the logical configuration concept that is chosen for the distribution of the identified workloads or workload groups for isolation and resource-sharing.

In general, single-rank extent pools might not be good in the current complex and mixed environments unless you know that this level of isolation and micro-performance management is required for your specific environment. If not managed correctly, workload skew and rank hot spots that limit overall system performance are likely to occur.

Multi-rank homogeneous extent pools (with only one storage tier)

If the logical configuration concept aims to balance certain workloads or workload groups (especially large resource-sharing workload groups) across multiple ranks with the allocation of volumes or extents on successive ranks, use multi-rank extent pools for these workloads.

With a homogeneous multi-rank extent pool, you take advantage of the advanced DS8900F virtualization features to spread the workload evenly across the ranks in an extent pool to achieve a well-balanced data distribution with considerably less management effort. Performance management is shifted from the rank level to the extent pool level. An extent pool represents a set of merged ranks (a larger set of disk spindles) with a uniform workload distribution. So, the level of complexity for standard performance and configuration management is reduced from managing many individual ranks (micro-performance management) to a few multi-rank extent pools (macro-performance management).

The DS8900F capacity allocation methods take care of spreading the volumes and thus the individual workloads evenly across the ranks within homogeneous multi-rank extent pools. Rotate extents is the default and preferred EAM to distribute the extents of each volume successively across all ranks in a pool to achieve a well-balanced capacity-based distribution of the workload. Rotate volumes is rarely used today, but it can help to implement a strict volume-to-rank relationship. It reduces the configuration effort compared to single-rank extent pools by easily distributing a set of volumes to different ranks in a specific extent pool for workloads where the use of host-based striping methods is still preferred.

The size of the volumes must fit the available capacity on each rank. The number of volumes that are created for this workload in a specific extent pool must match the number of ranks (or be at least a multiple of this number). Otherwise, the result is an imbalanced volume and workload distribution across the ranks and rank bottlenecks might emerge. However, efficient host-based striping must be ensured in this case to spread the workload evenly across all ranks, eventually from two or more extent pools. For more information about the EAMs and how the volume data is spread across the ranks in an extent pool, see , “Extent Allocation Method” on page 41.

Even multi-rank extent pools that are not managed by Easy Tier provide some level of control over the volume placement across the ranks in cases where it is necessary to enforce manually a special volume allocation scheme: You can use the DSCLI command **chrank -reserve** to reserve all of the extents from a rank in an extent pool from being used for the next creation of volumes. Alternatively, you can use the DSCLI command **chrank -release** to release a rank and make the extents available again.

Multi-rank extent pools that use storage pool striping are the general configuration approach today on modern DS8900F storage systems to spread the data evenly across the ranks in a homogeneous multi-rank extent pool and thus reduce skew and the likelihood of single-rank hot spots. Without Easy Tier automatic mode management, such non-managed, homogeneous multitier extent pools consist only of ranks of the same drive type and RAID level. Although not required (and probably not realizable for smaller or heterogeneous configurations), you can take the effective rank capacity into account, grouping ranks with and without spares into different extent pools when using storage pool striping to ensure a strict balanced workload distribution across all ranks up to the last extent. Otherwise, take additional considerations for the volumes that are created from the last extents in a mixed homogeneous extent pool that contains ranks with and without spares because these volumes are probably allocated only on part of the ranks with the larger capacity and without spares.

In combination with Easy Tier, a more efficient and automated way of spreading the workloads evenly across all ranks in homogeneous multi-rank extent pool is available. The automated intra-tier performance management (auto-rebalance) of Easy Tier efficiently spreads the workload evenly across all ranks. It automatically relocates the data across the ranks of the same storage class in an extent pool based on rank utilization to achieve and maintain a balanced distribution of the workload, minimizing skew and avoiding rank hot spots. You can enable auto-rebalance for homogeneous extent pool by setting the Easy Tier management scope to *all* extent pools (`ETautomode=all`).

In addition, Easy Tier automatic mode can also handle storage device variations within a tier that uses a micro-tiering capability.

Note: Easy Tier automated data relocation across tiers to optimize performance and storage economics based on the hotness of the particular extent takes place only between different storage tiers. If these drives are mixed in the same managed extent pool, the Easy Tier auto-rebalance algorithm balances the workload only across all ranks of this tier based on overall rank utilization, taking the performance capabilities of each rank (micro-tiering capability) into account.

With multi-rank extent pools, you can fully use the features of the DS8900F virtualization architecture and Easy Tier that provide ease of use when you manage more applications effectively and efficiently with a single DS8900F storage system. Consider multi-rank extent pools and the use of Easy Tier automatic management especially for mixed workloads that will be spread across multiple ranks. Multi-rank extent pools help simplify management and volume creation. They also allow the creation of single volumes that can span multiple ranks and thus exceed the capacity and performance limits of a single rank.

Easy Tier manual mode features, such as dynamic extent pool merge, dynamic volume relocation (volume migration), and rank depopulation also help to manage easily complex configurations with different extent pools. You can migrate volumes from one highly used extent pool to another less used one, or from an extent pool with a lower storage class to another one associated with a higher storage class, and merge smaller extent pools to larger ones. You can also redistribute the data of a volume within a pool by using the manual volume rebalance feature, for example, after capacity is added to a pool or two pools are merged, to optimize manually the data distribution and workload spreading within a pool. However, manual extent pool optimization and performance management, such as manual volume rebalance, is not required (and not supported) if the pools are managed by Easy Tier automatic mode. Easy Tier automatically places the data in these pools even if the pools are merged or capacity is added to a pool.

For more information about data placement in extent pool configurations, see 3.2.3, “Extent pool considerations” on page 51.

Important: Multi-rank extent pools offer numerous advantages with respect to ease of use, space efficiency, and the DS8900F virtualization features. Multi-rank extent pools, in combination with Easy Tier automatic mode, provide both ease of use and excellent performance for standard environments with workload groups that share a set of homogeneous resources.

4.8.2 Multitier extent pools

Multi-rank or hybrid extent pools consist of ranks from different storage classes or storage tiers (referred to as multitier or hybrid extent pools). Data placement within and across these tiers can automatically be managed by Easy Tier providing automated storage performance and storage economics optimization.

A multitier extent pool can consist of one of the following storage class combinations with up to three storage tiers, for instance:

- HPFE High-Performance flash cards (Tier 0) + HPFE High-Capacity flash cards (Tier 2)
- HPFE High-Performance flash cards (Tier 0) + HPFE High-Capacity of 3.84 TB (Tier1) + HPFE High-Capacity flash cards (Tier 2)

Multitier extent pools are especially suited for mixed, resource-sharing workloads. Tiered storage, as described in 4.1, “DS8900F Models” on page 60, is an approach of using types of storage throughout the storage infrastructure. It is a mix of higher-performing/higher-cost storage with lower-performing/lower-cost storage and placing data based on its specific I/O access characteristics. Correctly balancing all the tiers eventually leads to the lowest cost and best performance solution.

Always create hybrid extent pools for Easy Tier automatic mode management. The extent allocation for volumes in hybrid extent pools differs from the extent allocation in homogeneous pools. Any specified EAM, such as rotate extents or rotate volumes, is ignored when a new volume is created in, or migrated into, a hybrid pool. The EAM is changed to managed when the Easy Tier automatic mode is enabled for the pool, and the volume is under the control of Easy Tier. Easy Tier then automatically moves extents to the most appropriate storage tier and rank in the pool based on performance aspects.

Easy Tier automatically spreads workload across the resources (ranks and DAs) in a managed hybrid pool. Easy Tier automatic mode adapts to changing workload conditions and automatically promotes hot extents from the lower tier to the next upper tier. It demotes colder extents from the higher tier to the next lower tier. The auto-rebalance feature evenly distributes extents across the rank of the same tier based on rank utilization to minimize skew and avoid hot spots. Auto-rebalance takes different device characteristics into account when different devices or RAID levels are mixed within the same storage tier (micro-tiering).

Regarding the requirements of your workloads, you can create one or multiple pairs of extent pools with different two-tier or three-tier combinations that depend on your needs and available hardware resources. You can create three-tier extent pools for mixed, large resource-sharing workload groups and benefit from fully automated storage performance and economics management at a minimum management effort. You can boost the performance of your high-demand workloads with High-Performance flash and reduce the footprint and costs with High-Capacity flash for the lower-demand data.

You can use the DSCLI `showfbvol/showckdvol -rank` or `-tier` commands to display the current extent distribution of a volume across the ranks and tiers, as shown in Example 3-5 on page 56. Additionally, the volume heat distribution (volume heat map), provided by the Easy Tier Reporting in the DS8900F GUI, can help identify the amount of hot, warm, and cold extents for each volume and its distribution across the storage tiers in the pool. For more information about Easy Tier Reporting in the DS8900F GUI, see *IBM DS8000 Easy Tier*, REDP-4667.

The ratio of High-Performance flash and High-Capacity flash drive capacity in a hybrid pool depends on the workload characteristics and skew and must be planned when ordering the drive hardware for the identified workloads.

With the Easy Tier manual mode features, such as dynamic extent pool merge, dynamic volume relocation, and rank depopulation, you can modify existing configurations easily, depending on your needs. You can grow from a manually managed single-tier configuration into a partially or fully automatically managed tiered storage configuration. You add tiers or merge appropriate extent pools and enable Easy Tier at any time. For more information about Easy Tier, see *IBM DS8000 Easy Tier*, REDP-4667.

Important: Multitier extent pools and Easy Tier help you to implement a tiered storage architecture on a single DS8900F storage system with all its benefits at a minimum management effort and ease of use. Easy Tier and its automatic data placement within and across tiers spread the workload efficiently across the available resources in an extent pool. Easy Tier constantly optimizes storage performance and storage economics and adapts to changing workload conditions. Easy Tier can reduce overall performance management efforts and help consolidate more workloads efficiently and effectively on a single DS8000 storage system. It optimizes performance and reduces energy costs and the footprint.

Pinning volumes to a tier

By using the DSCLI commands `manageckdvol -action tierassign` or `managefbvol -action tierassign`, volumes in multitier pools can be assigned to a High-Performance flash tier (`-tier flashtier0`), to a High-Capacity flash of 3.84 TB tier (`-tier flashtier1`), or to a High-Capacity flash tier (`-tier flashtier2`). Such a pinning can be done permanently, or temporarily. For example, shortly before an important end-of-year application run of one database that is of top priority, you might move volumes of this database to a flash tier. After the yearly run is finished, you release these volumes again (`-action tierunassign`) and let them migrate freely between all available tiers. Such setups allow flexibly to make good use of a multitier extent pool structure and still give some short-term boost, or throttle, to some certain volumes if needed for a certain period.

Copy Services replication scenarios

When doing replication between two DS8000 storage systems, each using a hybrid multitier extent pool, all recent DS8000 models offer a heatmap transfer from the primary DS8000 storage system to the secondary DS8000 storage system, so that in case of a sudden failure of the primary system, the performance on the secondary system is immediately the same as before the DR failover. This Easy Tier heatmap transfer can be carried out by Copy Services Manager, by IBM GDPS, or by a small tool that regularly connects to both DS8000 storage systems. For each Copy Services pair, the local learning at each PPRC target volume for Easy Tier in that case is disabled, and instead the heatmap of the primary volume is used. For more information about this topic, see *IBM DS8870 Easy Tier Heat Map Transfer*, REDP-5015.

4.8.3 Additional implementation considerations for multitier extent pools

For multitier extent pools managed by Easy Tier automatic mode, consider how to allocate the initial volume capacity and how to introduce gradually new workloads to managed pools.

Using thin provisioning with Easy Tier in multitier extent pools

In environments that use FlashCopy, which supports the usage of thin-provisioned volumes (ESE), you might start using thin-provisioned volumes in managed hybrid extent pools, especially with three tiers.

In this case, only used capacity is allocated in the pool and Easy Tier does not move unused extents around or move hot extents on a large scale up from the lower tiers to the upper tiers. However, thin-provisioned volumes are not fully supported by all DS8000 Copy Services or advanced functions and platforms yet, so it might not be a valid approach for all environments at this time. For more information about the initial volume allocation in hybrid extent pools, see “Extent allocation in hybrid and managed extent pools” on page 44.

Staged implementation approach for multitier extent pools

Another control is the allocation policy for new volumes. In implementations before microcode 8.3, new storage was allocated from Tier 1 if available and then from Tier 2, but Tier 0 was not preferred.

You can change the allocation policy according to your needs with the chsi command and the `t-ettierorder highutil` or `highperf option`.

The default allocation order is High Performance for all-flash pools (DS8900F) accordingly with the following priorities:

- High Performance:
  1. Flash tier 0 (high performance)
  2. Flash tier 1 (high capacity)
  3. Flash tier 2 (high capacity)
- High Capacity:
  1. Flash tier 1 (high capacity)
  2. Flash tier 2 (high capacity)
  3. Flash tier 0 (high performance)

In a migration scenario, if you are migrating all the servers at once, some server volumes, for reasons of space, might be placed completely on Flash tier 0 first, although these servers also might have higher performance requirements. For more information about the initial volume allocation in hybrid extent pools, see “Extent allocation in hybrid and managed extent pools” on page 44.

When bringing a new DS8900F storage system into production to replace an older one, with the older storage system often not using Easy Tier, consider the timeline of the implementation stages by which you migrate all servers from the older to the new storage system.

One good option is to consider a staged approach when migrating servers to a new multitier DS8900F storage system:

- Assign the resources for the high-performing and response time sensitive workloads first, then add the less performing workloads.
- Split your servers into several subgroups, where you migrate each subgroup one by one, and not all at once. Then, allow Easy Tier several days to learn and optimize. Some extents are moved to High-Performance flash and some extents are moved to High-Capacity flash. After a server subgroup learns and reaches a steady state, the next server subgroup can be migrated. You gradually allocate the capacity in the hybrid extent pool by optimizing the extent distribution of each application one by one.
- Another option that can help in some cases of new deployments is to reset the Easy Tier learning heatmap for a certain subset of volumes, or for some pools. This action cuts off all the previous days of Easy Tier learning, and the next upcoming internal auto-migration plan is based on brand new workload patterns only.

4.8.4 Extent allocation in homogeneous multi-rank extent pools

When creating a new volume, the extent allocation method (EAM) determines *how* a volume is created within a multi-rank extent pool for the allocation of the extents on the available ranks. The rotate extents algorithm spreads the extents of a single volume and the I/O activity of each volume across all the ranks in an extent pool. Furthermore, it ensures that the allocation of the first extent of successive volumes starts on different ranks rotating through all available ranks in the extent pool to optimize workload spreading across all ranks.

With the default rotate extents (`rotateexts` in DSCLI, **Rotate capacity** in the GUI) algorithm, the extents (1 GiB for FB volumes and 1113 cylinders or approximately 0.86 GiB for CKD volumes) of each single volume are spread across all ranks within an extent pool and thus across more drives. This approach reduces the occurrences of I/O hot spots at the rank level within the storage system. Storage pool striping helps to balance the overall workload evenly across the back-end resources. It reduces the risk of single ranks that become performance bottlenecks while providing ease of use with less administrative effort.

When using the optional rotate volumes (`rotatevols` in DSCLI) EAM, each volume is placed on a single separate rank with a successive distribution across all ranks in a round-robin fashion.

The rotate extents and rotate volumes EAMs determine the initial data distribution of a volume and thus the spreading of workloads in non-managed, single-tier extent pools. With Easy Tier automatic mode enabled for single-tier (homogeneous) or multitier (hybrid) extent pools, this selection becomes unimportant. The data placement and thus the workload spreading is managed by Easy Tier. The use of Easy Tier automatic mode for single-tier extent pools is highly encouraged for an optimal spreading of the workloads across the resources. In single-tier extent pools, you can benefit from the Easy Tier automatic mode feature auto-rebalance. Auto-rebalance constantly and automatically balances the workload across ranks of the same storage tier based on rank utilization, minimizing skew and avoiding the occurrence of single-rank hot spots.

Additional considerations for specific applications and environments

This section provides considerations when manually spreading the workload in single-tier extent pools and selecting the correct EAM without using Easy Tier. The use of Easy Tier automatic mode also for single-tier extent pools is highly encouraged. Shared environments with mixed workloads that benefit from storage-pool striping and multitier configurations also benefit from Easy Tier automatic management. However, Easy Tier automatic management might not be an option in highly optimized environments where a fixed volume-to-rank relationship and well-planned volume configuration is required.

Certain, if not most, application environments might benefit from the use of storage pool striping (rotate extents):

- Operating systems that do not directly support host-level striping.
- VMware datastores.
- Microsoft Exchange.
- Windows clustering environments.
- Older Solaris environments.
- Environments that need to suballocate storage from a large pool.
- Applications with multiple volumes and volume access patterns that differ from day to day.
- Resource sharing workload groups that are dedicated to many ranks with host operating systems that do not all use or support host-level striping techniques or application-level striping techniques.

However, there might also be valid reasons for *not* using storage-pool striping. You might use it to avoid unnecessary layers of striping and reorganizing I/O requests, which might increase latency and not help achieve a more evenly balanced workload distribution. Multiple independent striping layers might be counterproductive. For example, creating a number of volumes from a single multi-rank extent pool that uses storage pool striping and then, additionally, use host-level striping or application-based striping on the same set of volumes might compromise performance. In this case, two layers of striping are combined with no overall performance benefit. In contrast, creating four volumes from four different extent pools from both rank groups that use storage pool striping and then use host-based striping or application-based striping on these four volumes to aggregate the performance of the ranks in all four extent pools and both processor complexes is reasonable.

Consider the following products for specific environments:

- SAN Volume Controller: SAN Volume Controller is a storage appliance with its own methods of striping and its own implementation of IBM Easy Tier. So, you can use striping or Easy Tier on the SAN Volume Controller or the DS8000 storage system.

- IBM i: IBM i has also its own methods of spreading workloads and data sets across volumes. For more information about IBM i storage configuration suggestions and Easy Tier benefits, see Chapter 9, “Performance considerations for the IBM i system” on page 205.
- z Systems: z/OS typically implements data striping with storage management subsystem (SMS) facilities. To make the SMS striping effective, storage administrators must plan to ensure a correct data distribution across the physical resources. For this reason, single-rank extent pool configuration was preferred in the past. Using storage-pool striping and multi-rank extent pools can considerably reduce storage configuration and management efforts because a uniform striping within an extent pool can be provided by the DS8900F storage system. Multi-rank extent pools, storage-pool striping, and Easy Tier are reasonable options today for z Systems configurations. For more information, see Chapter 10, “Performance considerations for IBM z Systems servers” on page 225.
- Database volumes: If these volumes are used by databases or applications that explicitly manage the workload distribution by themselves, these applications might achieve maximum performance by using their native techniques for spreading their workload across independent LUNs from different ranks. Especially with IBM Db2 or Oracle, where the vendor suggests specific volume configurations, for example, with Db2 Database Partitioning Feature (DPF) and Db2 Balanced Warehouses (BCUs), or with Oracle Automatic Storage Management (ASM), it is preferable to follow those suggestions. For more information about this topic, see Chapter 11, “Database for IBM z/OS performance” on page 265.
- Applications that evolved particular storage strategies over a long time, with proven benefits, and where it is unclear whether they can additionally benefit from using storage pool striping. When in doubt, follow the vendor recommendations.

DS8000 storage-pool striping is based on spreading extents across different ranks. So, with extents of 1 GiB (FB) or 0.86 GiB (1113 cylinders/CKD), the size of a data chunk is rather large. For distributing random I/O requests, which are evenly spread across the capacity of each volume, this chunk size is appropriate. However, depending on the individual access pattern of a specific application and the distribution of the I/O activity across the volume capacity, certain applications perform better. Use more granular stripe sizes for optimizing the distribution of the application I/O requests across different RAID arrays by using host-level striping techniques or have the application manage the workload distribution across independent volumes from different ranks.

Consider the following points for selected applications or environments to use storage-pool striping in homogeneous configurations:

- Db2: Excellent opportunity to simplify storage management by using storage-pool striping. You might prefer to use Db2 traditional recommendations for Db2 striping for performance-sensitive environments.
- Db2 and similar data warehouse applications, where the database manages storage and parallel access to data. Consider independent volumes on individual ranks with a careful volume layout strategy that does *not* use storage-pool striping. Containers or database partitions are configured according to suggestions from the database vendor.
- Oracle: Excellent opportunity to simplify storage management for Oracle. You might prefer to use Oracle traditional suggestions that involve ASM and Oracle striping capabilities for performance-sensitive environments.
- Small, highly active logs or files: Small highly active files or storage areas smaller than 1 GiB with a high access density might require spreading across multiple ranks for performance reasons. However, storage-pool striping offers a striping granularity on extent levels only around 1 GiB, which is too large in this case. Continue to use host-level striping techniques or application-level striping techniques that support smaller stripe sizes. For

example, assume a 0.8 GiB log file exists with extreme write content, and you want to spread this log file across several RAID arrays. Assume that you intend to spread its activity across four ranks. At least four 1 GiB extents must be allocated, one extent on each rank (which is the smallest possible allocation). Creating four separate volumes, each with a 1 GiB extent from each rank, and then using Logical Volume Manager (LVM) striping with a relatively small stripe size (for example, 16 MiB) effectively distributes the workload across all four ranks. Creating a single LUN of four extents, which is also distributed across the four ranks by using DS8900F storage-pool striping, cannot effectively spread the file workload evenly across all four ranks because of the large stripe size of one extent, which is larger than the actual size of the file.

- IBM Spectrum® Protect/Tivoli Storage Manager storage pools: IBM Spectrum Protect storage pools work well in striped pools. But, Spectrum Protect suggests that the Spectrum Protect databases be allocated in a separate pool or pools.
- AIX volume groups (VGs): LVM and physical partition (PP) striping continue to be tools for managing performance. In combination with storage-pool striping, now considerably fewer stripes are required for common environments. Instead of striping across a large set of volumes from many ranks (for example, 32 volumes from 32 ranks), striping is required only across a few volumes from a small set of different multi-rank extent pools from both DS8000 rank groups that use storage-pool striping. For example, use four volumes from four extent pools, each with eight ranks. For specific workloads that use the advanced AIX LVM striping capabilities with a smaller granularity on the KiB or MiB level, instead of storage-pool striping with 1 GiB extents (FB), might be preferable to achieve the highest performance.
- Microsoft Windows volumes: Typically, only a few large LUNs per host system are preferred, and host-level striping is not commonly used. So, storage-pool striping is an ideal option for Windows environments. You can create single, large-capacity volumes that offer the performance capabilities of multiple ranks. A volume is not limited by the performance limits of a single rank, and the DS8000 storage system spreads the I/O load across multiple ranks.
- Microsoft Exchange: Storage-pool striping makes it easier for the DS8000 storage system to conform to Microsoft sizing suggestions for Microsoft Exchange databases and logs.
- Microsoft SQL Server: Storage-pool striping makes it easier for the DS8000 storage system to conform to Microsoft sizing suggestions for Microsoft SQL Server databases and logs.
- VMware datastore for Virtual Machine Storage Technologies (VMware ESX Server Filesystem (VMFS)) or virtual raw device mapping (RDM) access: Because data stores concatenate LUNs rather than striping them, allocate the LUNs inside a striped storage pool. Estimating the number of disks (or ranks) to support any specific I/O load is straightforward based on the specific requirements.

In general, storage-pool striping helps improve overall performance and reduces the effort of performance management by evenly distributing data and workloads across a larger set of ranks, which reduces skew and hot spots. Certain application workloads can also benefit from the higher number of disk spindles behind one volume. But, there are cases where host-level striping or application-level striping might achieve a higher performance, at the cost of higher overall administrative effort. Storage-pool striping might deliver good performance in these cases with less management effort, but manual striping with careful configuration planning can achieve the ultimate preferred levels of performance. So, for overall performance and ease of use, storage-pool striping might offer an excellent compromise for many environments, especially for larger workload groups where host-level striping techniques or application-level striping techniques are not widely used or available.

Note: Business- and performance-critical applications always require careful configuration planning and individual decisions on a case-by-case basis. Determine whether to use storage-pool striping or LUNs from dedicated ranks together with host-level striping techniques or application-level striping techniques for the preferred performance.

Storage-pool striping is well-suited for new extent pools.

4.8.5 Balancing workload across available resources

To achieve a balanced utilization of all available resources of the DS8900F storage system, you must distribute the I/O workloads evenly across the available back-end resources:

- Ranks (disk drive modules)
- DA pairs (HPFEs)

You must distribute the I/O workloads evenly across the available front-end resources:

- I/O ports
- HA cards
- I/O enclosures

You must distribute the I/O workloads evenly across both DS8900F processor complexes (called storage server `0/CEC#0` and storage server `1/CEC#1`) as well.

Configuring the extent pools determines the balance of the workloads across the available back-end resources, ranks, DA pairs, and both processor complexes.

Tip: If you use the GUI for an initial configuration, most of this balancing is done automatically for you.

Each extent pool is associated with an extent pool ID (P0, P1, and P2, for example). Each rank has a relationship to a specific DA pair and can be assigned to only one extent pool. You can have as many (non-empty) extent pools as you have ranks. Extent pools can be expanded by adding more ranks to the pool. However, when assigning a rank to a specific extent pool, the affinity of this rank to a specific DS8900F processor complex is determined. By hardware, a predefined affinity of ranks to a processor complex does not exist. All ranks that are assigned to even-numbered extent pools (P0, P2, and P4, for example) form rank group 0 and are serviced by DS8900F processor complex 0. All ranks that are assigned to odd-numbered extent pools (P1, P3, and P5, for example) form rank group 1 and are serviced by DS8900F processor complex 1.

To spread the overall workload across both DS8900F processor complexes, a minimum of two extent pools is required: one assigned to processor complex 0 (for example, `P0`) and one assigned to processor complex 1 (for example, `P1`).

For a balanced distribution of the overall workload across both processor complexes and both DA cards of each DA pair, apply the following rules. For each type of rank and its RAID level, storage type (FB or CKD), and drive characteristics (flash type, HCF/HPF, and capacity), apply these rules:

- Assign half of the ranks to even-numbered extent pools (rank group 0) and assign half of them to odd-numbered extent pools (rank group 1).
- Spread ranks with and without spares evenly across both rank groups.
- Distribute ranks from each DA pair evenly across both rank groups.

It is important to understand that you might seriously limit the available back-end bandwidth and thus the system overall throughput if, for example, all ranks of a DA pair are assigned to only one rank group and thus a single processor complex. In this case, only one DA card of the DA pair is used to service all the ranks of this DA pair and thus only half of the available DA pair bandwidth is available.

Use the GUI for creating a new configuration; even if there are fewer controls, as far as the balancing is concerned, the GUI takes care of rank and DA distribution when creating pools.

4.8.6 Assigning workloads to extent pools

Extent pools can contain only ranks of the same storage type, either FB for Open Systems or IBM i, or CKD for FICON-based z Systems. You can have multiple extent pools in various configurations on a single DS8900F storage system, for example, managed or non-managed single-tier (homogeneous) extent pools with ranks of only one storage class. You can have multitier (hybrid) extent pools with any two storage tiers, or up to three storage tiers managed by Easy Tier.

Multiple homogeneous extent pools, each with different storage classes, easily allow tiered storage concepts with dedicated extent pools and manual cross-tier management. For example, you can have extent pools with slow, large-capacity drives for backup purposes and other extent pools with high-speed, small capacity drives or flash for performance-critical transaction applications. Also, you can use hybrid pools with Easy Tier and introduce fully automated cross-tier storage performance and economics management.

Using dedicated extent pools with an appropriate number of ranks and DA pairs for selected workloads is a suitable approach for isolating workloads.

The minimum number of required extent pools depends on the following considerations:

- The number of isolated and resource-sharing workload groups
- The number of different storage types, either FB for Open Systems or IBM i, or CKD for z Systems
- Definition of failure boundaries (for example, separating logs and table spaces to different extent pools)
- in some cases, Copy Services considerations

Although you are not restricted from assigning all ranks to only one extent pool, the minimum number of extent pools, even with only one workload on a homogeneously configured DS8000 storage system, must be two (for example, P0 and P1). You need one extent pool for each rank group (or storage server) so that the overall workload is balanced across both processor complexes

To optimize performance, the ranks for each workload group (either isolated or resource-sharing workload groups) must be split across at least two extent pools with an equal number of ranks from each rank group. So, at the workload level, each workload is balanced across both processor complexes. Typically, you assign an equal number of ranks from each DA pair to extent pools assigned to processor complex 0 (rank group 0: P0, P2, and P4, for example) and to extent pools assigned to processor complex 1 (rank group 1: P1, P3, and P5, for example). In environments with FB and CKD storage (Open Systems and z Systems), you additionally need separate extent pools for CKD and FB volumes. It is often useful to have a minimum of four extent pools to balance the capacity and I/O workload between the two DS8900F processor complexes. Additional extent pools might be needed to meet individual needs, such as ease of use, implementing tiered storage concepts, or separating ranks for different drive types, RAID types, clients, applications, performance, or Copy Services requirements.

However, the maximum number of extent pools is given by the number of available ranks (that is, creating one extent pool for each rank).

In most cases, accepting the configurations that the GUI offers when doing initial setup and formatting already gives excellent results. For specific situations, creating dedicated extent pools on the DS8900F storage system with dedicated back-end resources for separate workloads allows individual performance management for business and performance-critical applications. Compared to share and spread everything storage systems without the possibility to implement workload isolation concepts, creating dedicated extent pools on the DS8900F storage system with dedicated back-end resources for separate workloads is an outstanding feature of the DS8900F storage system as an enterprise-class storage system. With this feature, you can consolidate and manage various application demands with different performance profiles, which are typical in enterprise environments, on a single storage system.

Before configuring the extent pools, collect all the hardware-related information of each rank for the associated DA pair, disk type, available storage capacity, RAID level, and storage type (CKD or FB) in a spreadsheet. Then, plan the distribution of the workloads across the ranks and their assignments to extent pools.

Plan an initial assignment of ranks to your planned workload groups, either isolated or resource-sharing, and extent pools for your capacity requirements. After this initial assignment of ranks to extent pools and appropriate workload groups, you can create additional spreadsheets to hold more details about the logical configuration and finally the volume layout of the array site IDs, array IDs, rank IDs, DA pair association, extent pools IDs, and volume IDs, and their assignments to volume groups and host connections.

4.9 Planning address groups, LSSs, volume IDs, and CKD PAVs

After creating the extent pools and evenly distributing the back-end resources (DA pairs and ranks) across both DS8900F processor complexes, you can create host volumes from these pools. When creating the host volumes, it is important to follow a volume layout scheme that evenly spreads the volumes of each application workload across all ranks and extent pools that are dedicated to this workload to achieve a balanced I/O workload distribution across ranks, DA pairs, and the DS8900F processor complexes.

So, the next step is to plan the volume layout and thus the mapping of address groups and LSSs to volumes created from the various extent pools for the identified workloads and workload groups. For performance management and analysis reasons, it can be useful to relate easily volumes, which are related to a specific I/O workload, to ranks, which finally provide the physical disk spindles for servicing the workload I/O requests and determining the I/O processing capabilities. Therefore, an overall logical configuration concept that easily relates volumes to workloads, extent pools, and ranks is wanted.

Each volume is associated with a hexadecimal four-digit volume ID that must be specified when creating the volume. An example for volume ID 1101 is shown in Table 4-4.

Table 4-4 Understand the volume ID relationship to address groups and LSSs/LCUs

Volume ID	Digits	Description
1101	First digit: **1***xxx*	Address group (0 - F) (16 address groups on a DS8000 storage system)
	First and second digits: **11***xx*	Logical subsystem (LSS) ID for FB Logical control unit (LCU) ID for CKD (x0 - xF: 16 LSSs or LCUs per address group)
	Third and fourth digits: *xx***01**	Volume number within an LSS or LCU (00 - FF: 256 volumes per LSS or LCU)

The first digit of the hexadecimal volume ID specifies the address group, 0 - F, of that volume. Each address group can be used only by a single storage type, either FB or CKD. The first and second digit together specify the logical subsystem ID (LSS ID) for Open Systems volumes (FB) or the logical control unit ID (LCU ID) for z Systems volumes (CKD). There are 16 LSS/LCU IDs per address group. The third and fourth digits specify the volume number within the LSS/LCU, 00 - FF. There are 256 volumes per LSS/LCU. The volume with volume ID 1101 is the volume with volume number 01 of LSS 11, and it belongs to address group 1 (first digit).

The LSS/LCU ID is related to a rank group. Even LSS/LCU IDs are restricted to volumes that are created from rank group 0 and serviced by processor complex 0. Odd LSS/LCU IDs are restricted to volumes that are created from rank group 1 and serviced by processor complex 1. So, the volume ID also reflects the affinity of that volume to a DS8000 processor complex. All volumes, which are created from even-numbered extent pools (P0, P2, and P4, for example) have even LSS IDs and are managed by DS8000 processor complex 0. All volumes that are created from odd-numbered extent pools (P1, P3, and P5, for example) have odd LSS IDs and are managed by DS8000 processor complex 1.

There is no direct DS8000 performance implication as a result of the number of defined LSSs/LCUs. For the z/OS CKD environment, a DS8000 volume ID is also required for each PAV. The maximum of 256 addresses per LCU includes both CKD base volumes and PAVs, so the number of volumes and PAVs determines the number of required LCUs.

In the past, for performance analysis reasons, it was useful to identify easily the association of specific volumes to ranks or extent pools when investigating resource contention. But, since the introduction of storage-pool striping, the use of multi-rank extent pools is the preferred configuration approach for most environments. Multitier extent pools are managed by Easy Tier automatic mode anyway, constantly providing automatic storage intra-tier and cross-tier performance and storage economics optimization. For single-tier pools, turn on Easy Tier management. In managed pools, Easy Tier automatically relocates the data to the appropriate ranks and storage tiers based on the access pattern, so the extent allocation across the ranks for a specific volume is likely to change over time. With storage-pool striping or extent pools that are managed by Easy Tier, you no longer have a fixed relationship between the performance of a specific volume and a single rank. Therefore, planning for a hardware-based LSS/LCU scheme and relating LSS/LCU IDs to hardware resources, such as ranks, is no longer reasonable. Performance management focus is shifted from ranks to extent pools. However, a numbering scheme that relates only to the extent pool might still be viable, but it is less common and less practical.

The common approach that is still valid today with Easy Tier and storage pool striping is to relate an LSS/LCU to a specific application workload with a meaningful numbering scheme for the volume IDs for the distribution across the extent pools. Each LSS can have 256 volumes, with volume numbers 00 - FF. So, relating the LSS/LCU to a certain application workload and additionally reserving a specific range of volume numbers for different extent pools is a reasonable choice, especially in Open Systems environments. Because volume IDs are transparent to the attached host systems, this approach helps the administrator of the host system to determine easily the relationship of volumes to extent pools by the volume ID. Therefore, this approach helps you easily identify physically independent volumes from different extent pools when setting up host-level striping across pools. This approach helps you when separating, for example, database table spaces from database logs on to volumes from physically different drives in different pools.

This approach provides a logical configuration concept that provides ease of use for storage management operations and reduces management efforts when using the DS8900F related Copy Services because basic Copy Services management steps (such as establishing Peer-to-Peer Remote Copy (PPRC) paths and consistency groups) are related to LSSs. If Copy Services are not planned, plan the volume layout because overall management is easier if you must introduce Copy Services in the future (for example, when migrating to a new DS8000 storage system that uses Copy Services).

However, the strategy for the assignment of LSS/LCU IDs to resources and workloads can still vary depending on the particular requirements in an environment.

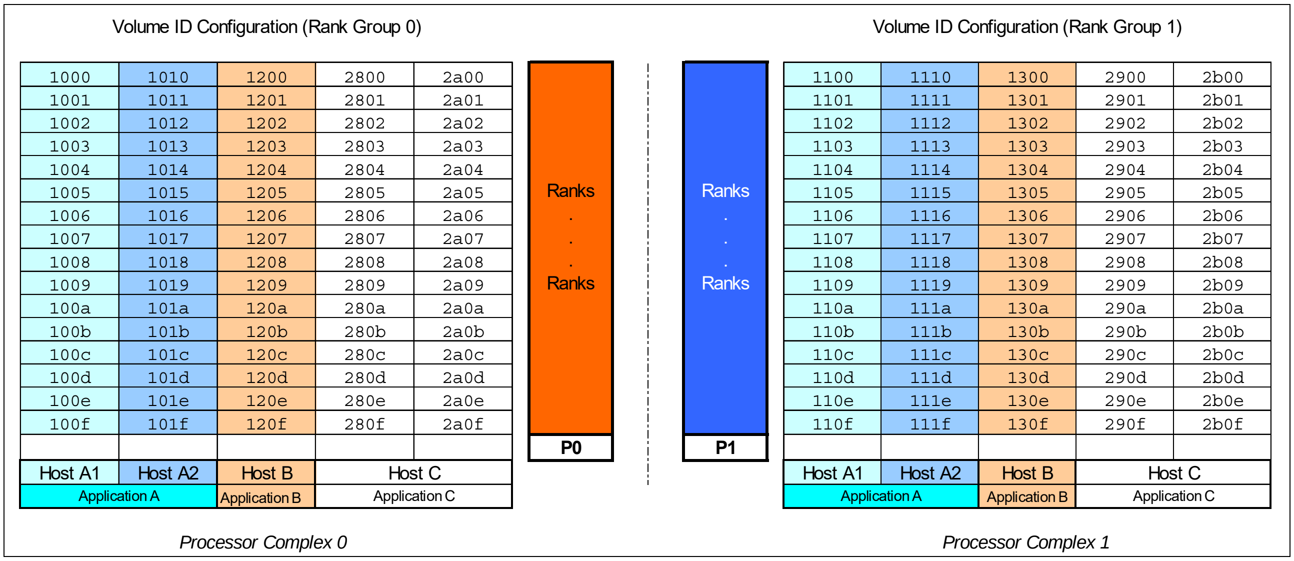
The following section introduces suggestions for LSS/LCU and volume ID numbering schemes to help to relate volume IDs to application workloads and extent pools.

4.9.1 Volume configuration scheme by using application-related LSS/LCU IDs

This section describes several suggestions for a volume ID numbering scheme in a multi-rank extent pool configuration where the volumes are evenly spread across a set of ranks. These examples refer to a workload group where multiple workloads share the set of resources. The following suggestions apply mainly to Open Systems. For CKD environments, the LSS/LCU layout is defined at the operating system level with the Input/Output Configuration Program (IOCP) facility. The LSS/LCU definitions on the storage server must match exactly the IOCP statements. For this reason, any consideration on the LSS/LCU layout must be made first during the IOCP planning phase and afterward mapped to the storage server.

Typically, when using LSS/LCU IDs that relate to application workloads, the simplest approach is to reserve a suitable number of LSS/LCU IDs according to the total number of volumes requested by the application. Then, populate the LSS/LCUs in sequence, creating the volumes from offset 00. Ideally, all volumes that belong to a certain application workload or a group of related host systems are within the same LSS. However, because the volumes must be spread evenly across both DS8000 processor complexes, at least two logical subsystems are typically required per application workload. One even LSS is for the volumes that are managed by processor complex 0, and one odd LSS is for volumes managed by processor complex 1 (for example, LSS 10 and LSS 11). Moreover, consider the future capacity demand of the application when planning the number of LSSs to be reserved for an application. So, for those applications that are likely to increase the number of volumes beyond the range of one LSS pair (256 volumes per LSS), reserve a suitable number of LSS pair IDs for them from the beginning.

Figure 4-5 shows an example of an application-based LSS numbering scheme. This example shows three applications, application A, B, and C, that share two large extent pools. Hosts A1 and A2 both belong to application A and are assigned to LSS 10 and LSS 11, each using a different volume ID range from the same LSS range. LSS 12 and LSS 13 are assigned to application B, which runs on host B. Application C is likely to require more than 512 volumes, so use LSS pairs 28/29 and 2a/2b for this application.



Volume ID Configuration (Rank Group 0)

1000	1010	1200	2800	2a00
1001	1011	1201	2801	2a01
1002	1012	1202	2802	2a02
1003	1013	1203	2803	2a03
1004	1014	1204	2804	2a04
1005	1015	1205	2805	2a05
1006	1016	1206	2806	2a06
1007	1017	1207	2807	2a07
1008	1018	1208	2808	2a08
1009	1019	1209	2809	2a09
100a	101a	120a	280a	2a0a
100b	101b	120b	280b	2a0b
100c	101c	120c	280c	2a0c
100d	101d	120d	280d	2a0d
100e	101e	120e	280e	2a0e
100f	101f	120f	280f	2a0f
Host A1	Host A2	Host B	Host C	
Application A		Application B	Application C	

Volume ID Configuration (Rank Group 1)

1100	1110	1300	2900	2b00
1101	1111	1301	2901	2b01
1102	1112	1302	2902	2b02
1103	1113	1303	2903	2b03
1104	1114	1304	2904	2b04
1105	1115	1305	2905	2b05
1106	1116	1306	2906	2b06
1107	1117	1307	2907	2b07
1108	1118	1308	2908	2b08
1109	1119	1309	2909	2b09
110a	111a	130a	290a	2b0a
110b	111b	130b	290b	2b0b
110c	111c	130c	290c	2b0c
110d	111d	130d	290d	2b0d
110e	111e	130e	290e	2b0e
110f	111f	130f	290f	2b0f
Host A1	Host A2	Host B	Host C	
Application A		Application B	Application C	

Figure 4-5 Application-related volume layout example for two shared extent pools

If an application workload is distributed across multiple extent pools on each processor complex, for example, in a four extent pool configuration, as shown in Figure 4-6 , you can expand this approach. Define a different volume range for each extent pool so that the system administrator can easily identify the extent pool behind a volume from the volume ID, which is transparent to the host systems.

In Figure 4-6 , the workloads are spread across four extent pools. Again, assign two LSS/LCU IDs (one even, one odd) to each workload to spread the I/O activity evenly across both processor complexes (both rank groups). Additionally, reserve a certain volume ID range for each extent pool based on the third digit of the volume ID. With this approach, you can quickly create volumes with successive volume IDs for a specific workload per extent pool with a single DSCLI **`mkfbvol`** or **`mkckdvol`** command.

Hosts A1 and A2 belong to the same application A and are assigned to LSS 10 and LSS 11. For this workload, use volume IDs 1000 - 100f in extent pool P0 and 1010 - 101f in extent pool P2 on processor complex 0. Use volume IDs 1100 - 110f in extent pool P1 and 1110 - 111f in extent pool P3 on processor complex 1. In this case, the administrator of the host system can easily relate volumes to different extent pools and thus different physical resources on the same processor complex by looking at the third digit of the volume ID. This numbering scheme can be helpful when separating, for example, DB table spaces from DB logs on to volumes from physically different pools.

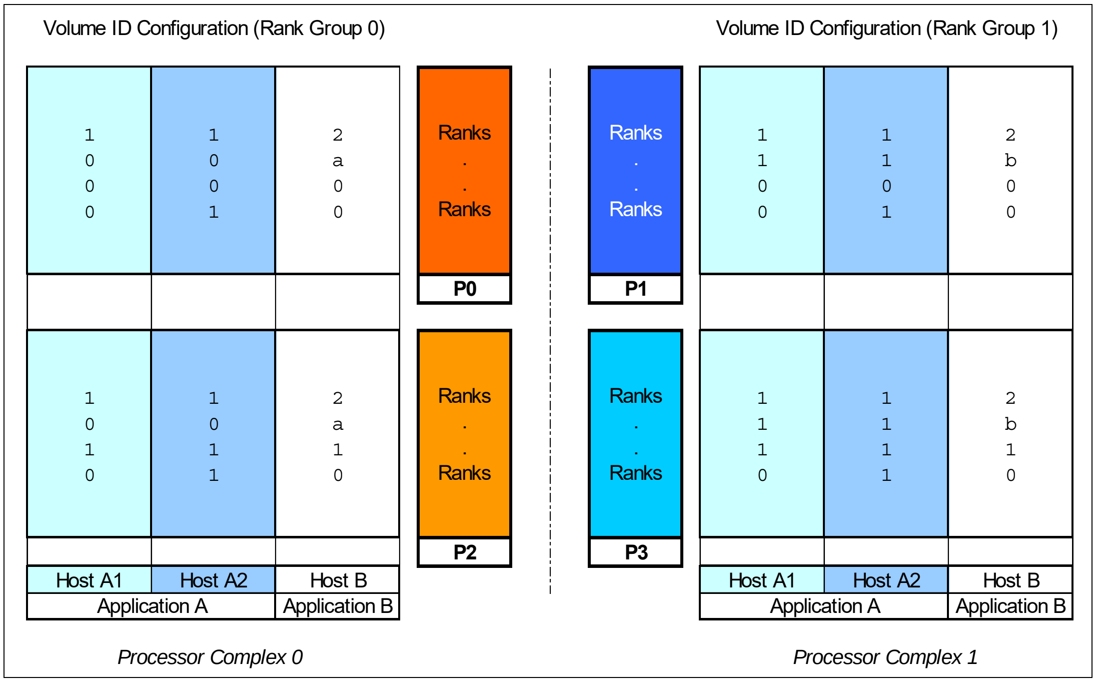
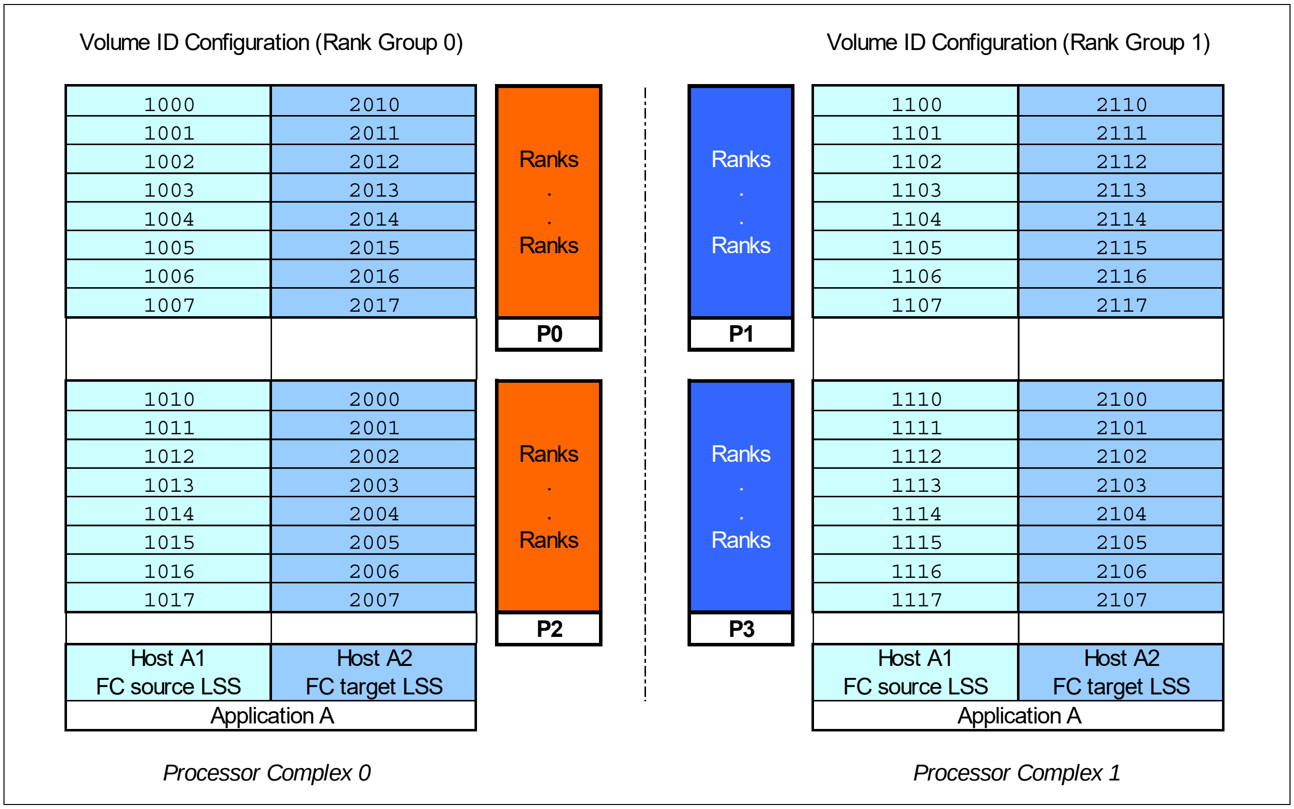


Figure 4-6 Application and extent pool-related volume layout example for four shared extent pools

The example that is depicted in Figure 4-7 on page 97 provides a numbering scheme that can be used in a FlashCopy scenario. Two different pairs of LSS are used for source and target volumes. The address group identifies the role in the FlashCopy relationship: address group 1 is assigned to source volumes, and address group 2 is used for target volumes. This numbering scheme allows a symmetrical distribution of the FlashCopy relationships across source and target LSSs. For example, source volume 1007 in P0 uses the volume 2007 in P2 as the FlashCopy target. In this example, use the third digit of the volume ID within an LSS as a marker to indicate that source volumes 1007 and 1017 are from different extent pools. The same approach applies to the target volumes, for example, volumes 2007 and 2017 are from different pools.

However, for the simplicity of the Copy Services management, you can choose a different extent pool numbering scheme for source and target volumes (so 1007 and 2007 are not from the same pool) to implement the recommended extent pool selection of source and target volumes in accordance with the FlashCopy guidelines. Source and target volumes must stay on the same rank group but different ranks or extent pools. For more information about this topic, see the FlashCopy performance chapters in *IBM DS8000 Copy Services*, SG24-8367.



Volume ID Configuration (Rank Group 0)

Host A1 FC source LSS	Host A2 FC target LSS
1000	2010
1001	2011
1002	2012
1003	2013
1004	2014
1005	2015
1006	2016
1007	2017
1010	2000
1011	2001
1012	2002
1013	2003
1014	2004
1015	2005
1016	2006
1017	2007

Application A

Processor Complex 0

Volume ID Configuration (Rank Group 1)

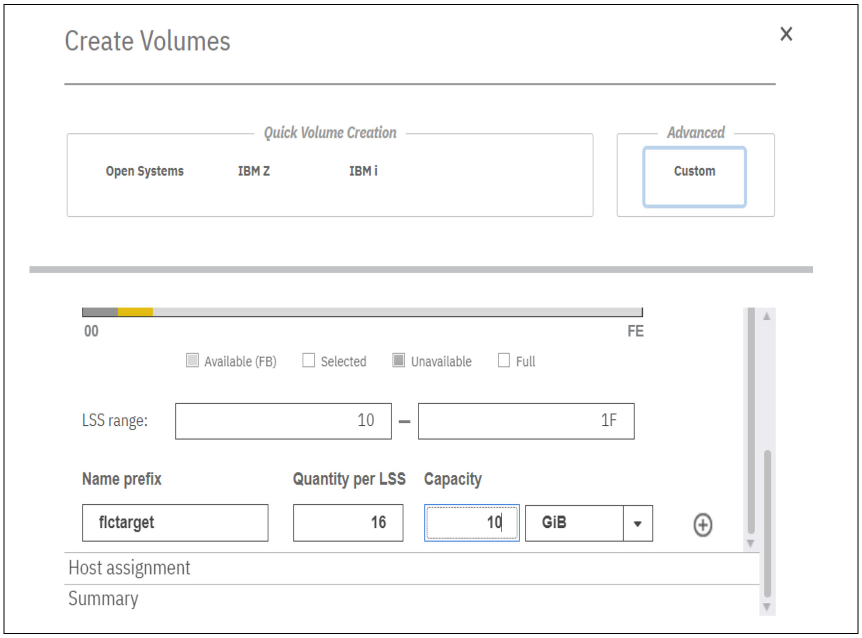
Host A1 FC source LSS	Host A2 FC target LSS
1100	2110
1101	2111
1102	2112
1103	2113
1104	2114
1105	2115
1106	2116
1107	2117
1110	2100
1111	2101
1112	2102
1113	2103
1114	2104
1115	2105
1116	2106
1117	2107

Application A

Processor Complex 1

Figure 4-7 Application and extent pool-related volume layout example in a FlashCopy scenario

> **Tip:** Use the GUI advanced Custom mode to select a specific LSS range when creating volumes. Choose this Custom mode and also the appropriate Volume definition mode, as shown in Figure 4-8 .



Create Volumes

LSS range: 10 – 1F

Name prefix	Quantity per LSS	Capacity	
flctarget	16	10	GiB

Figure 4-8 Create volumes in Custom mode and specify an LSS range

4.10 I/O port IDs, host attachments, and volume groups

Finally, when planning the attachment of the host system to the storage system HA I/O ports, you must achieve a balanced workload distribution across the available front-end resources for each workload with the appropriate isolation and resource-sharing considerations. Therefore, distribute the FC connections from the host systems evenly across the DS8900F HA ports, HA cards, and I/O enclosures.

For high availability, each host system must use a multipathing device driver, such as the native MPIO of the respective operating system. Each host system must have a minimum of two host connections to HA cards in different I/O enclosures on the DS8900F storage system. Preferably, they are evenly distributed between left side (even-numbered) I/O enclosures and right side (odd-numbered) I/O enclosures. The number of host connections per host system is primarily determined by the required bandwidth. Use an appropriate number of HA cards to satisfy high throughput demands.

With typical transaction-driven workloads that show high numbers of random, small-blocksize I/O operations, all ports in a HA card can be used likewise. For the preferred performance of workloads with different I/O characteristics, consider the isolation of large-block sequential and small-block random workloads at the I/O port level or the HA card level.

The preferred practice is to use dedicated I/O ports for Copy Services paths and host connections. For more information about performance aspects that are related to Copy Services, see the performance-related chapters in *IBM DS8000 Copy Services*, SG24-8367.

To assign FB volumes to the attached Open Systems hosts by using LUN masking, when using the DSCLI, these volumes must be grouped in the DS8900F volume groups. A volume group can be assigned to multiple host connections, and each host connection is specified by the worldwide port name (WWPN) of the host FC port. A set of host connections from the same host system is called a host attachment. The same volume group can be assigned to multiple host connections; however, a host connection can be associated only with one volume group. To share volumes between multiple host systems, the most convenient way is to create a separate volume group for each host system and assign the shared volumes to each of the individual volume groups as required. A single volume can be assigned to multiple volume groups. Only if a group of host systems shares a set of volumes, and there is no need to assign additional non-shared volumes independently to particular hosts of this group, can you consider using a single shared volume group for all host systems to simplify management. Typically, there are no significant DS8000 performance implications because of the number of DS8000 volume groups or the assignment of host attachments and volumes to the DS8000 volume groups.

Do not omit additional host attachment and host system considerations, such as SAN zoning, multipathing software, and host-level striping.

After the DS8900F storage system is installed, you can use the DSCLI **lsioport** command to display and document I/O port information, including the I/O ports, HA type, I/O enclosure location, and WWPN. Use this information to add specific I/O port IDs, the required protocol (FICON or FCP), and the DS8000 I/O port WWPNs to the plan of host and remote mirroring connections that are identified in 4.4, “Planning allocation of disk and host connection capacity” on page 73.

Additionally, the I/O port IDs might be required as input to the DS8900F host definitions if host connections must be restricted to specific DS8900F I/O ports by using the **-ioport** option of the **mkhostconnect** DSCLI command. If host connections are configured to allow access to all DS8000 I/O ports, which is the default, typically the paths must be restricted by SAN zoning. The I/O port WWPNs are required as input for SAN zoning. The **lshostconnect -login**

DSCLI command might help verify the final allocation of host attachments to the DS8900F I/O ports because it lists host port WWPNs that are logged in, sorted by the DS8900F I/O port IDs for known connections. The **`lshostconnect -unknown`** DSCLI command might further help identify host port WWPNs, which are not yet configured to host connections, when creating host attachments by using the **`mkhostconnect`** DSCLI command.

The DSCLI **`lsioport`** output identifies this information:

- The number of I/O ports on each installed HA
- The type of installed HAs (SW FCP/FICON-capable or LW FCP/FICON-capable)
- The distribution of HAs across I/O enclosures
- The WWPN of each I/O port

The DS8900F I/O ports use predetermined, fixed DS8900F *logical* port IDs in the form I0*xyz*, where:

- x: I/O enclosure
- y: Slot number within the I/O enclosure
- z: Port within the adapter

For example, `I0101` is the I/O port ID for these devices:

- I/O enclosure `1`
- Slot `0`
- Second port

Slot numbers: The slot numbers for logical I/O port IDs are one less than the physical location numbers for HA cards, as shown on the physical labels and in IBM Spectrum Control/Storage Insights, for example, `I0101` is R1-XI2-C1-T2.

A simplified example of spreading the DS8000 I/O ports evenly to two redundant SAN fabrics is shown in Figure 4-9 . The SAN implementations can vary, depending on individual requirements, workload considerations for isolation and resource-sharing, and available hardware resources.

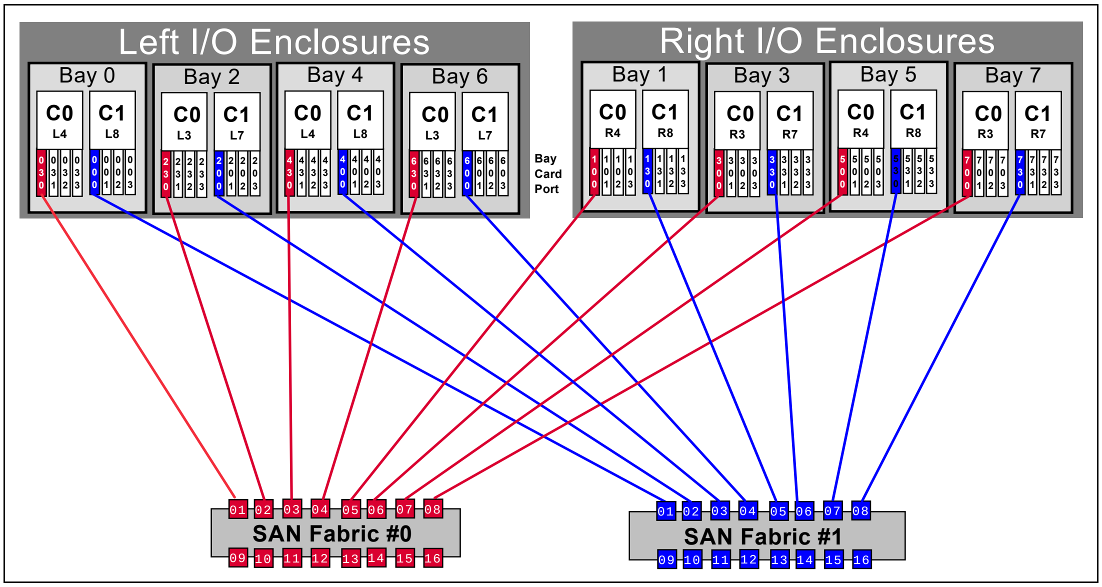


Figure 4-9 Example of spreading DS8000 I/O ports evenly across two redundant SAN fabrics

4.10.1 Fibre Channel host adapters: 16 Gbps and 32 Gbps

The DS8900F supports up to 32 Fibre Channel (FC) host adapters, with four FC ports for each adapter. Each port can be independently configured to support Fibre Channel connection (IBM FICON) or Fibre Channel Protocol (FCP).

Each adapter type is available in both longwave (LW) and shortwave (SW) versions. The DS8900F I/O bays support up to four host adapters for each bay, allowing up to 128 ports maximum for each storage system. This configuration results in a theoretical aggregated host I/O bandwidth around 128 x 32 Gbps. Each port provides industry-leading throughput and I/O rates for FICON and FCP.

The host adapters that are available in the DS8900F have the following characteristics:

- 32 Gbps FC HBAs (32 GFC):
  - Four FC ports
  - FC Gen7 technology
  - New IBM Custom ASIC with Gen3 PCIe interface
  - Quad-core PowerPC processor
  - Negotiation to 32, 16, or 8 Gbps (4 Gbps or less is not possible)
- 16 Gbps FC HBAs (16 GFC):
  - Four FC ports
  - Gen2 PCIe interface
  - Quad-core PowerPC processor
  - Negotiation to 16, 8, or 4 Gbps (2 Gbps or less is not possible)

The DS8900F supports a mixture of 32 Gbps and 16 Gbps FC adapters. Hosts with slower FC speeds like 4 Gbps are still supported if their HBAs are connected through a switch.

The 32 Gbps FC adapter is encryption-capable. Encrypting the host bus adapter (HBA) traffic usually does not cause any measurable performance degradation.

With FC adapters that are configured for FICON, the DS8900F series provides the following configuration capabilities:

- Fabric or point-to-point topologies
- A maximum of 128 host adapter ports, depending on the DS8900F system memory and processor features
- A maximum of 509 logins for each FC port
- A maximum of 8192 logins for each storage unit
- A maximum of 1280 logical paths on each FC port
- Access to all 255 control-unit images (65,280 Count Key Data (CKD) devices) over each FICON port
- A maximum of 512 logical paths for each control unit image

An IBM Z server supports 32,000 devices per FICON host channel. To fully access 65,280 devices, it is necessary to connect multiple FICON host channels to the storage system. You can access the devices through an FC switch or FICON director to a single storage system FICON port.

The 32 GFC host adapter doubles the data throughput of 16 GFC links. When comparing the two adapter types, the 32 GFC adapters provide I/O improvements in full adapter I/Os per second (IOPS) and reduced latency.

Double bandwidth with 32 GFC HA

The latest DS8000 storage system support 32-gigabit fiber channel host adapters (32 GFC HA). This capability effectively doubles the bandwidth performance compared to 16 GFC HA. The 32 GFC HA is able to transfer data at 3.2 GBPS per port. It provides 2x bandwidth capacity at the port and the HA layer when compared to single 16 GFC HA. The 32 GFC HA ports are compatible with 16 GFC and 8 GFC infrastructure and can negotiate the transmission speed to 800 or 1600 MBPS depending on the Fibre Channel port they are connected to. With support for FICON, zHPF, and FCP protocols provided on an individual port basis, each port can support any one of these protocols so that a single 4-port HA card can have all supported protocols running concurrently.

Figure 4-10 on page 101 illustrates read/write performance with zHPF protocol

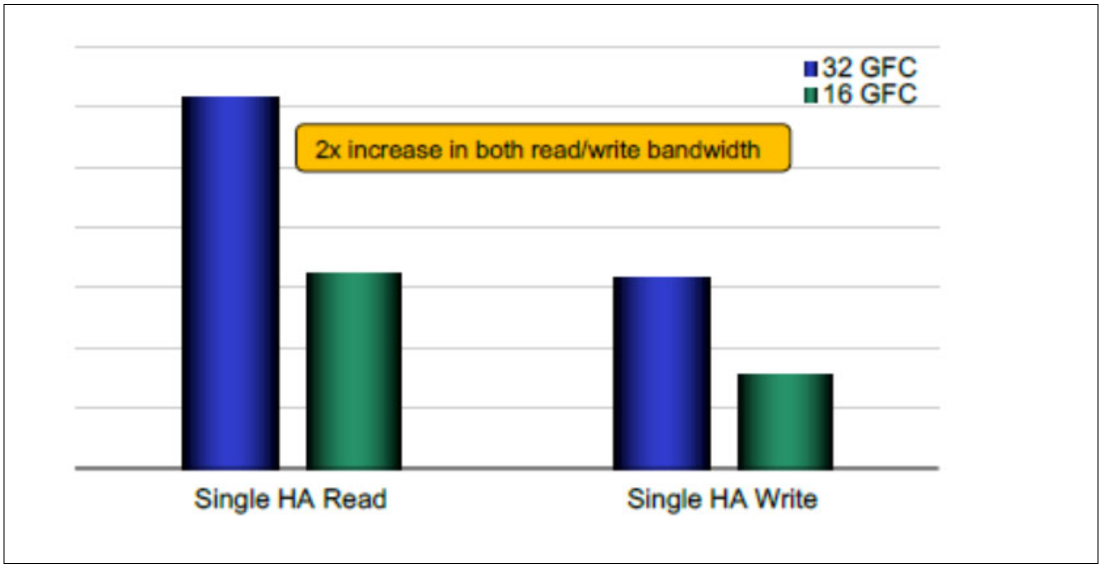


Figure 4-10 CKD/zHPF - 32 GFC vs. 16 GFC HA, Read/Write

Figure 4-11 illustrates Metro Mirror Bandwidth

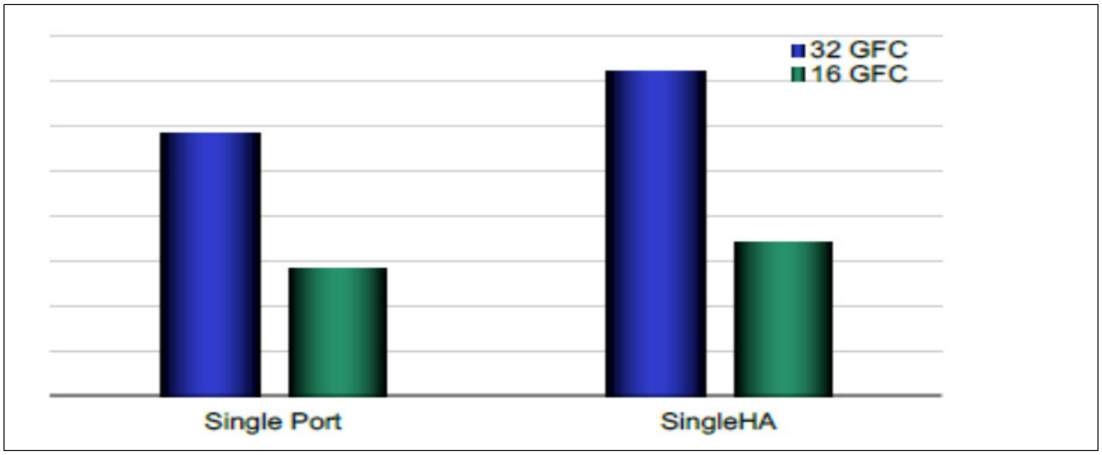


Figure 4-11 Metro Mirror Link Bandwidth 32 GFC vs. 16 GFC HA

Note The above measurements were derived by running simulated workloads in a Lab environment. They are being provided purely to demonstrate comparative/indicative performance improvement between different HA and port types. Your current environment setup will determine the scope for performance improvements

General considerations about 16 and 32 GFC and I/O Ports

The following recommendations applies to 16 and 32 GFC and I/O Ports:

- Spread the paths from all host systems across the available I/O ports, HA cards, and I/O enclosures to optimize workload distribution across the available resources depending on your workload sharing and isolation considerations.
- Spread the host paths that access the same set of volumes as evenly as possible across the available HA cards and I/O enclosures. This approach balances workload across hardware resources, and it ensures that a hardware failure does not result in a loss of access.
- Plan the paths for the attached host systems with a minimum of two host connections to different HA cards in different I/O enclosures on the DS8900F storage system. Preferably, evenly distribute them between left (even-numbered) I/O enclosures and right (odd-numbered) I/O enclosures for the highest availability and a balanced workload distribution across I/O enclosures and HA cards.
- Use separate DS8900F I/O ports for host attachment and Copy Services remote replication connections (such as Metro Mirror, Global Mirror, and zGM data mover). If additional HAs are available, consider using separate HAs for Copy Services remote replication connections and host attachments to avoid any possible interference between remote replication and host workloads.
- Spread Copy Services remote replication connections at least across two HA cards in different I/O enclosures.
- Consider using separate HA cards for FICON protocol and FCP. Although I/O ports on the same HA can be configured independently for the FCP protocol and the FICON protocol, it might be preferable to isolate your z/OS environment (FICON) from your Open Systems environment (FCP).

Look at Example 4-2, which shows a DS8900F storage system with a selection of different HAs:

Example 4-2 DS8000 HBA example - DSCLI lsioport command output (shortened)

```
dscli> lsioport -l
ID    WWPN             State   Type             topo     portgrp  Speed   Frame I/O Enclosure HA
Card
=================================================================================================
I0200 500507630A1013E7  Online Fibre Channel-SW SCSI-FCP 0         16 Gb/s 1     3             1
I0201 500507630A1053E7  Online Fibre Channel-SW FICON    0         16 Gb/s 1     3             1
I0202 500507630A1093E7  Online Fibre Channel-SW FICON    0         16 Gb/s 1     3             1
I0203 500507630A10D3E7  Online Fibre Channel-SW FICON    0         16 Gb/s 1     3             1
I0230 500507630A1313E7 Offline Fibre Channel-LW -        0         32 Gb/s 1     3             4
I0231 500507630A1353E7 Offline Fibre Channel-LW -        0         32 Gb/s 1     3             4
I0232 500507630A1393E7 Offline Fibre Channel-LW -        0         32 Gb/s 1     3             4
I0233 500507630A13D3E7 Offline Fibre Channel-LW -        0         32 Gb/s 1     3             4
I0240 500507630A1413E7  Online Fibre Channel-SW FICON    0         16 Gb/s 1     3             5
I0241 500507630A1453E7 Offline Fibre Channel-SW -        0         16 Gb/s 1     3             5
I0242 500507630A1493E7  Online Fibre Channel-SW SCSI-FCP 0         16 Gb/s 1     3             5
I0243 500507630A14D3E7 Offline Fibre Channel-SW -        0         16 Gb/s 1     3             5
```

```
I0300 500507630A1813E7 Offline Fibre Channel-LW -        0         32 Gb/s 1     4            1
I0301 500507630A1853E7 Offline Fibre Channel-LW -        0         32 Gb/s 1     4            1
I0302 500507630A1893E7 Offline Fibre Channel-LW -        0         32 Gb/s 1     4            1
I0303 500507630A18D3E7 Offline Fibre Channel-LW -        0         32 Gb/s 1     4            1
I0310 500507630A1913E7  Online Fibre Channel-SW SCSI-FCP 0         16 Gb/s 1     4            2
I0311 500507630A1953E7  Online Fibre Channel-SW FICON    0         16 Gb/s 1     4            2
I0312 500507630A1993E7  Online Fibre Channel-SW FICON    0         16 Gb/s 1     4            2
I0313 500507630A19D3E7  Online Fibre Channel-SW FICON    0         16 Gb/s 1     4            2
I0330 500507630A1B13E7  Online Fibre Channel-SW FICON    0         16 Gb/s 1     4            4
I0331 500507630A1B53E7 Offline Fibre Channel-SW -        0         16 Gb/s 1     4            4
I0332 500507630A1B93E7  Online Fibre Channel-SW SCSI-FCP 0         16 Gb/s 1     4            4
I0333 500507630A1BD3E7 Offline Fibre Channel-SW -        0         16 Gb/s 1     4            4
```

When planning the paths for the host systems, ensure that each host system uses a multipathing device driver and a minimum of two host connections to two different HA cards in different I/O enclosures on the DS8900F. Preferably, they are evenly distributed between left side (even-numbered) I/O enclosures and the right side (odd-numbered) I/O enclosures for highest availability. Multipathing additionally optimizes workload spreading across the available I/O ports, HA cards, and I/O enclosures.

You must tune the SAN zoning scheme to balance both the oversubscription and the estimated total throughput for each I/O port to avoid congestion and performance bottlenecks.

4.10.2 Further optimization

Sometimes the IBM support staff is asked if any further tuning possibilities exist with respect to HBAs, and paths. Although the impact is usually minimal, take the following hints and tips into consideration, whenever possible:

PPRC fine-tuning

For optimal use of your PPRC connections, observe the following guidelines:

- Avoid using FCP ports for both host attachment and replication.
- With bi-directional PPRC: Avoid using a specific adapter port for both replication output and input I/Os, rather consider to use each port in one direction only.
- When having a larger number of replication ports, assign even and odd LSS/LCU to different ports (example: With 4 paths, could use two only for all even LSSs to be replicated, and the other two for all odd LSSs to be replicated).

These guidelines help the adapters to reduce internal switching, what slightly increases their throughput for PPRC.

I/O Bay fine-tuning

The DS8000 models use specific algorithms to distribute internal CPU resources across the I/O Bays, depending on the specific submodel and cores available. As a result of this, some configurations are more favorable than others, and the following guidelines apply:

- Balance your host attachments well by spreading across all available I/O Bays.
- Balance your PPRC connections well by spreading them across all available I/O Bays.
- Avoid to order a DS8000 with many cores (DS8000s with 1 TB and larger caches) but then with only 2 I/O Bays in a frame. That also means for instance, if it is a stronger 2-frame storage system and it is decided that just 4 I/O Bays in total are enough, have these all in one frame.

- Spreading a particular workload over host adapters in only either the top bays of both frames, or only the bottom bays of both frames, would also not lead to the most optimal result – hence spread each bigger load (attachment/core switches, PPRC) across all the various rows of bays that you have.
- These guidelines would also apply in case of incoming PPRC load (Secondary DS8000), or when doing an even/odd load split with the PPRC paths as mentioned in the PPRC section. So if you do such a split, spread out the paths across all I/O Bays in each case, like across the 4 bays in one frame.

4.11 Documentation of the logical configuration

You can use the DS Storage Manager GUI to export information in a spreadsheet format (CSV file). Figure 4-12 the shows the export of the GUI System Summary. The resulting spreadsheet file contains a lot of detailed information, such as cache and drives installed, HAs installed, licenses installed, arrays created and extent pools, all volumes created, hosts and users, and code levels. You can also consider this function for regular usage.

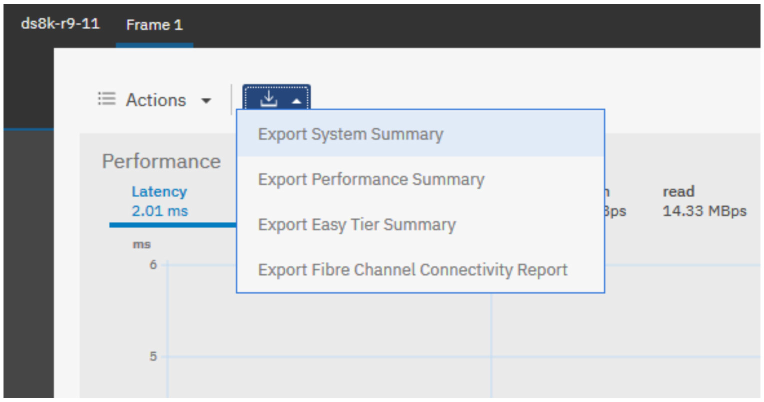


Figure 4-12 Export System Summary

Part 2

IBM System Storage DS8000 performance management

This part starts with a brief introduction of different workload types and introduces various tools for effective performance planning, monitoring, and management on the DS8000 storage system.

This part includes the following topics:

- Understanding your workload
- Performance planning tools
- Practical performance management

5

Understanding your workload

This chapter presents and describes the various workload types that an application can generate. This characterization can be useful for understanding performance documents and reports and categorizing the various workloads in your installation.

Information in this chapter is not dedicated to IBM System Storage DS8900F. You can apply this information generally to other flash or disk storage systems.

This chapter includes the following topics:

- General workload types
- Database Workload
- Application workload
- Profiling workloads in the design phase
- Understanding your workload type

5.1 General workload types

The correct understanding of the existing or planned workload is the key element of the entire planning and sizing process. Understanding the workload means having the description of the workload pattern:

- Expected or existing number of IOPS
- Size of the I/O requests
- Read and write ratio
- Random and sequential access ratio
- General purpose of the application and the workload

You might also collect the following information:

- Expected cache hit ratio

  The number of requests serviced from cache. This number is important for read requests because write requests always get into the cache first. If out of 1000 requests 100 of them are serviced from cache, you have a 10% of cache hit ratio. The higher this parameter, the lower the overall response time.

- Expected seek ratio
- This parameter is not applicable to the flash drives that have no disk arms by design.

  The percentage of the I/O requests for which the disk arm must move from its location. Moving the disk arm requires more time than rotating the disk, which rotates anyway and is fast enough. So, by not moving the disk arm, the whole track or cylinder can be read, which generally means large amount of data. This parameter is mostly indicative of how disk systems worked a long time ago, and it is now turned into a sort of the quality value of the random nature of the workload. A random workload shows this value close to 100%.

- Expected write efficiency

  The write efficiency is a number that represents the number of times a block is written to before being de-staged to the Flash module or disk. Actual applications, especially databases, update the information: write-read again-change-write with changes. So, the data for the single flash module or disk block can be served several times from cache before it is written to the flash module or disk. A value of 0% means that a de-stage is assumed for every write operation and the characteristic of the “pure random small block write workload pattern”, which is unlikely. A value of 50% means that a de-stage occurs after the track is written to twice. A value of 100% is unlikely also because it means that writes come to the one track only and are never destaged to disk.

In general, you describe the workload in these terms. The following sections cover the details and describe the different workload types.

5.1.1 Typical Online Transaction Processing workload

This workload is characterized by mostly the random access of small-sized I/O records (less than or equal to 16 KB) with a mix of 70% reads and 30% writes. This workload is also characterized by low read-hit ratios in the flash or disk system cache (less than 20%). This workload might be representative of various online applications, for example, the SAP NetWeaver application or many database applications. This type of workload is typical because it is the basis for most of the benchmark and performance tests. However the following online transaction processing (OLTP) patterns are spread:

- 90% read, 8 - 16 - 32 - 64 KB blocks, 30% sequential, 30 - 40% cache hit
- 80% read, 16 - 32 - 64 KB blocks, 40 - 50% sequential, 50 - 60% cache hit

A 100% random access workload is rare, which you must remember when you size the disk system.

5.1.2 Microsoft Exchange Server Workload

This type of workload can be similar to an OLTP workload, but it has a different read/write balance. It is characterized by many write operations, up to 60%, with high random numbers. Also, the size of the I/O can be high, with blocks up to 128 KB, which is explained by the nature of the application. Additionally, it acts as a database and the data warehouse.

To help you with the implementation of storage systems with Microsoft Exchange Server 2019 is the Microsoft Jetstess tool. This tool provides the ability to simulate and verify the performance and stability of storage subsystems before putting them into production environments. Microsoft Exchange documentation can be found at:

https://www.microsoft.com/en-us/download/details.aspx?id=36849

See also Microsoft Exchange Solution Reviewed Program (ESRP) Storage which is primarily designed for the testing of third-party storage solutions. Refer to the following Microsoft Exchange article:

https://docs.microsoft.com/en-us/exchange/esrp-storage?view=exchserver-2019

5.1.3 Sequential Workload

Sequential access is one of the original workload types for data storage. Tape is the best example of sequential access. Tape uses several large blocks at one read operation, and it uses buffers first. Sequential access does not change for the flash modules or disks. Sequential workload is good for prefetch and putting data into cache because blocks are accessed one by one, and the storage system can read many blocks to unload the flash modules or disks. Sequential write requests work well because the storage system can optimize the access to the flash modules disks and write several tracks or cylinders at a time. Block sizes are typically 256 KB or more and response time is high, but that time does not matter. Sequential workload is about bandwidth, not response time. The following environments and applications are likely sequential workloads:

- Backup/restore applications
- Database log files
- Batch processing
- File servers
- Web servers
- Media streaming applications
- Graphical software

5.1.4 Batch Jobs Workload

Batch workloads have several common characteristics:

- Mixture of random database access, skip-sequential, pure sequential, and sorting.
- Large block sizes up to 128 - 256 KB
- High volume of write activity, and read activity
- The same volume extents might scramble for write and read at the same time.
- Batch workloads include large data transfers and high path utilizations
- Batch workloads are often constrained to operate within a particular window of time when online operation is restricted or shut down. Poor or improved performance is often not recognized unless it affects this window.

Plan when to run batch jobs: Plan all batch workload activity for the end of the day or at a slower time of day. Normal activity can be negatively affected with the batch activity.

5.1.5 Sort Jobs Workload

Most sorting applications are characterized by large transfers for input, output, and work data sets. DFSORT is IBM's high-performance sort, merge, copy, analysis and reporting product and is an optional feature of z/OS.

For more information about z/OS DFSORT, see the following:

- http://www.ibm.com/support/docview.wss?rs=114&uid=isg3T7000077
- https://www.ibm.com/docs/en/zos

5.1.6 Read-intensive Cache Friendly and Unfriendly Workloads

Use the cache hit ratio to estimate the cache-friendly read workload. If the ratio is more than 50%, the workload is cache friendly. If you have two serviced I/O's to the same data and one I/O is serviced from cache, it is a 50% cache hit ratio.

It is not as easy to divide known workload types into cache friendly and cache unfriendly. An application can change its behavior during the day several times. When users work with data, it is cache friendly. When the batch processing or reporting starts, it is not cache friendly. High random-access numbers mean a not cache-friendly workload type. However, if the amount of data that is accessed randomly is not large, 10% for example, it can be placed totally into the disk system cache and becomes cache friendly.

Sequential workloads are always cache-friendly because of prefetch algorithms that exist in the DS8900 storage system. Sequential workload is easy to prefetch. You know that the next 10 or 100 blocks are accessed, and you can read them in advance. For the random workloads, it is different. There are no purely random workloads in the actual applications, and it is possible to predict some moments. The DS8900 storage systems use the following powerful read-caching algorithms to deal with cache unfriendly workloads:

- Sequential Prefetching in Adaptive Replacement Cache (SARC)
- Adaptive Multi-stream Prefetching (AMP)
- Intelligent Write Caching (IWC)

The write workload is always cache friendly because every write request comes to the cache first and the application gets the reply when the request is placed into cache. Write requests are served at least two times longer by the back end than read requests. You always need to wait for the write acknowledgment, which is why cache is used for every write request.

To learn more about the DS8900 caching algorithms, see Chapter 2.2.1, "DS8000 caching algorithms" on page 17.

Table 5-1 on page 111 provides a summary of the characteristics of the various types of workloads.

Table 5-1 Workload types

Workload type	Characteristics	Representative of this type of process
Sequential read	Sequential 128 - 1024 KB blocks Read hit ratio: 100%	Database backups Batch processing Media streaming
Sequential write	Sequential 128 - 1024 KB blocks Write ratio: 100%	Database restores and loads Batch processing File access
z/OS cache uniform	Random 4 KB record R/W ratio 3.4 Read hit ratio 84%	Average database CICS/VSAM IBM IMS
z/OS cache standard	Random 4 KB record R/W ratio 3.0 Read hit ratio 78%	Representative of typical database conditions
z/OS cache friendly	Random 4 KB record R/W ratio 5.0 Read hit ratio 92%	Interactive Existing software
z/OS cache hostile	Random 4 KB record R/W ratio 2.0 Read hit ratio 40%	Db2 logging
Open read-intensive	Random 4 - 8 - 16 - 32 KB record Read% = 70 - 90% Hit ratio 28%, 1%	Databases (Oracle and Db2) Large DB inquiry Decision support Warehousing
Open standard	Random 4 - 8 - 16 - 32 KB record Read% = 70% Hit ratio 50%	OLTP General Parallel File System (IBM GPFS)

5.2 Database Workload

Database workload does not come with the database initially. It depends on the application that is written for the database and the type of work that this application performs. The workload can be an OLTP workload or a data warehousing workload in the same database. A reference to a database in this section mostly means Db2 and Oracle databases, but this section can apply to other databases. For more information, see Chapter 11, “Database for IBM z/OS performance” on page 265.

The database environment is often difficult to typify because I/O characteristics differ greatly. A database query has a high read content and is of a sequential nature. It also can be random, depending on the query type and data structure. Transaction environments are more random in behavior and are sometimes cache unfriendly. At other times, they have good hit ratios. You can implement several enhancements in databases, such as sequential prefetch and the exploitation of I/O priority queuing, that affect the I/O characteristics. Users must understand the unique characteristics of their database capabilities before generalizing the performance.

5.2.1 Database query workload

Database query is a common type of database workload. This term includes transaction processing that is typically random, and sequential data reading, writing, and updating. The query can have following properties:

- High read content
- Mix of write and read content
- Random access, and sequential access
- Small or large transfer size

A well-tuned database keeps characteristics of the queries closer to a sequential read workload. A database can use all available caching algorithms, both its own and the storage system algorithms. This function, which caches data that has the most probability to be accessed, provides performance improvements for most database queries.

5.2.2 Database logging workload

The logging system is an important part of the database. It is the main component to preserve the data integrity and provide a transaction mechanism. There are several types of log files in the database:

- Online transaction logs: This type of log is used to restore the last condition of the database when the latest transaction failed. A transaction can be complex and require several steps to complete. Each step means changes in the data. Because data in the database must be in a consistent state, an incomplete transaction must be rolled back to the initial state of the data. Online transaction logs have a rotation mechanism that creates and uses several small files for about an hour each.
- Archive transaction logs: This type is used to restore a database state up to the specified date. Typically, it is used with incremental or differential backups. For example, if you identify a data error that occurred a couple of days ago, you can restore the data back to its prior condition with only the archive logs. This type of log uses a rotation mechanism also.

The workload pattern for the logging is sequential writes mostly. Block size is about 64 KB. Reads are rare and might not be considered. The write capability and location of the online transaction logs are most important. The entire performance of the database depends on the writes to the online transaction logs, if the database is very write intensive, consider RAID-10. If the configuration has several extent pool pairs in its logical layout also consider physically separating log files from the flash modules or disks on which the data and index files are.

5.2.3 Database transaction environment workload

Database transaction workloads have these characteristics:

- Low to moderate read hits, depending on the size of the database buffers
- Cache unfriendly for certain applications
- Deferred writes that cause low write-hit ratios, which means that cached write data is rarely required for reading
- Deferred write chains with multiple locate-record commands in chain
- Low read/write ratio because of reads that are satisfied in a large database buffer pool
- High random-read access values, which are cache unfriendly

The enhanced prefetch cache algorithms, together with the high storage back-end bandwidth, provide high system throughput and high transaction rates for database transaction-based workloads.

A database can benefit from using a large amount of server memory for the large buffer pool. For example, the database large buffer pool, when managed correctly, can avoid a large percentage of the accesses to flash or disk. Depending on the application and the size of the buffer pool, this large buffer pool can convert poor cache hit ratios into synchronous reads in Db2. You can spread data across several RAID arrays to increase the throughput even if all accesses are read misses. Db2 administrators often require that table spaces and their indexes are placed on separate volumes. This configuration improves both availability and performance.

5.2.4 Database utilities workload

Database utilities, such as loads, reorganizations, copies, and recovers, generate high read and write sequential and sometimes random operations. This type of workload takes advantage of the sequential bandwidth performance of the back-end storage connection, such as the PCI-Express bus for the device adapter (DA) pairs, and the use of higher RPM (15 K) disk drives or flash drives and Easy Tier automatic mode enabled on both.

5.3 Application workload

This section categorizes various types of common applications according to their I/O behavior. There are four typical categories:

- Need for high throughput. These applications need more bandwidth (the more, the better). Transfers are large, read-only I/O’s that are sequential access. These applications use database management systems (DBMSs); however, random DBMS access might also exist.
- Need for high throughput and a mix of read/write (R/W), similar to the first category (large transfer sizes). In addition to 100% read operations, this category mixes reads and writes in 70/30 and 50/50 ratios. The DBMS is typically sequential, but random and 100% write operations also exist.
- Need for high I/O rate and throughput. This category requires both performance characteristics of IOPS and megabytes per second (MBps). Depending on the application, the profile is typically sequential access, medium to large transfer sizes (16 KB, 32 KB, and 64 KB), and 100/0, 0/100, and 50/50 R/W ratios.
- Need for high I/O rate. With many users and applications that run simultaneously, this category can consist of a combination of small to medium-sized transfers (4 KB, 8 KB, 16 KB, and 32 KB), 50/50 and 70/30 R/W ratios, and a random DBMS.

> **Synchronous activities:** Certain applications have synchronous activities, such as locking database tables during an online backup, or logging activities. These types of applications are highly sensitive to any increase in flash module or disk drive response time and must be handled with care.

Table 5-2 summarizes these workload categories and common applications.

Table 5-2 Application workload types

Category	Application	Read/write ratio	I/O size	Access type
4	General file serving	Expect 50/50	64 - 256 KB	Sequential mostly because of good file system caching.
4	Online transaction processing	50/50, 70/30	4 KB, 8 KB, or 16 KB	Random mostly for writes and reads. Bad cache hits
4	Batch update	Expect 50/50	16 KB, 32 KB, 64 KB, or 128 KB	Almost 50/50 mix of sequential and random. Moderate cache hits.
1	Data mining	90/10	32 KB, 64 KB, or larger	Mainly sequential, some random.
1	Video on demand	100/0	256 KB and larger	Sequential, good caching.
2	Data warehousing	90/10, 70/30, or 50/50	64 KB or larger	Mainly sequential, rarely random, and good caching.
2	Engineering and scientific	100/0, 0/100, 70/30, or 50/50	64 KB or larger	Sequential mostly, good caching.
3	Digital video editing	100/0, 0/100, or 50/50	128 KB, 256 - 1024 KB	Sequential, good caching.
3	Image processing	100/0, 0/100, or 50/50	64 KB, 128 KB	Sequential, good caching.
1	Backup, restore	100/0, 0/100	256 - 1024 KB	Sequential, good caching.

5.3.1 General file serving

This application type consists of many users who run many different applications, all with varying file access sizes and mixtures of read/write ratios, all occurring simultaneously. Applications can include file server, LAN storage and even internet/intranet servers. There is no standard profile. General file serving fits this application type because this profile covers almost all transfer sizes and R/W ratios.

5.3.2 Online Transaction Processing

This application category typically has many users, all accessing the same storage system and a common set of files. The file access typically is under the control of a DBMS, and each user might work on the same or unrelated activities. The I/O requests are typically spread across many files; therefore, the file sizes are typically small and randomly accessed. A typical application consists of a network file server or a storage system that is accessed by a sales department that enters order information.

5.3.3 Data mining

Databases are the repository of most data, and every time that information is needed, a database is accessed. Data mining is the process of extracting valid, previously unknown, and ultimately comprehensive information from large databases to make crucial business decisions. This application category consists of several operations, each of which is supported by various techniques, such as rule induction, neural networks, conceptual clustering, and association discovery. In these applications, the DBMS extracts only large sequential or possibly random files, depending on the DBMS access algorithms.

5.3.4 Video on Demand

Video on demand consists of video playback that can be used to broadcast quality video for either satellite transmission or a commercial application, such as in-room movies. Fortunately for the storage industry, the data rates that are needed for this type of transfer are reduced dramatically because of data compression developments. A broadcast quality video stream, for example, full HD video, may need about 4 - 5 Mbps bandwidth to serve a single user. These advancements reduce the need for higher speed interfaces and can be serviced with the current interface. However, these applications demand numerous concurrent users that interactively access multiple files within the same storage system. This requirement changed the environment of video applications because the storage system is specified by several video streams that they can service simultaneously. In this application, the DBMS extracts only large sequential files.

5.3.5 Data Warehousing

A data warehouse supports information processing by providing a solid platform of integrated, historical data from which to perform analysis. A data warehouse organizes and stores the data that is needed for informational and analytical processing over a long historical period. A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data that is used to support the management decision-making process. A data warehouse is always a physically separate store of data that spans a spectrum of time, and many relationships exist in the data warehouse.

An example of a data warehouse is a design around a financial institution and its functions, such as loans, savings, bank cards, and trusts for a financial institution. In this application, there are three kinds of operations: initial loading of the data, access to the data, and updating of the data. However, because of the fundamental characteristics of a warehouse, these operations can occur simultaneously. At times, this application can perform 100% reads when accessing the warehouse, 70% reads and 30% writes when accessing data while record updating occurs simultaneously, or even 50% reads and 50% writes when the user load is heavy. The data within the warehouse is a series of snapshots and after the snapshot of data is made, the data in the warehouse does not change. Therefore, there is typically a higher read ratio when using the data warehouse.

5.3.6 Engineering and Scientific Applications

The engineering and scientific arena includes hundreds of applications. Typical applications are computer-assisted design (CAD), Finite Element Analysis, simulations and modeling, and large-scale physics applications. Transfers can consist of 1 GB of data for 16 users. Other transfers might require 20 GB of data and hundreds of users. The engineering and scientific areas of business are more concerned with the manipulation of spatial data and series data. This application typically goes beyond standard relational DBMSs, which manipulate only flat (two-dimensional) data. Spatial or multi-dimensional issues and the ability to handle complex data types are commonplace in engineering and scientific applications.

Object-Relational DBMSs (ORDBMSs) are being developed, and they offer traditional relational DBMS features and support complex data types. Objects can be stored and manipulated, and complex queries at the database level can be run. Object data is data about real objects, including information about their location, geometry, and topology. Location describes their position, geometry relates to their shape, and topology includes their relationship to other objects. These applications essentially have an identical profile to that of the data warehouse application.

5.3.7 Digital Video Editing

Digital video editing is popular in the movie industry. The idea that a film editor can load entire feature films onto flash storage systems and interactively edit and immediately replay the edited clips has become a reality. This application combines the ability to store huge volumes of digital audio and video data onto relatively affordable storage devices to process a feature film.

Depending on the host and operating system that are used to perform this application, transfers are typically medium to large and access is always sequential. Image processing consists of moving huge image files for editing. In these applications, the user regularly moves huge high-resolution images between the storage device and the host system. These applications service many desktop publishing and workstation applications. Editing sessions can include loading large files of up to 16 MB into host memory, where users edit, render, modify, and store data onto the storage system. High interface transfer rates are needed for these applications, or the users waste huge amounts of time by waiting to see results. If the interface can move data to and from the storage device at over 32 MBps, an entire 16 MB image can be stored and retrieved in less than 1 second. The need for throughput is all important to these applications and along with the additional load of many users, I/O operations per second (IOPS) are also a major requirement.

5.4 Profiling workloads in the design phase

Assessing the I/O profile before you build and deploy the application requires methods of evaluating the workload profile without measurement data. In these cases, as a preferred practice, use a combination of general rules based on application type and the development of an application I/O profile by the application architect or the performance architect. The following examples are basic examples that are designed to provide an idea of how to approach workload profiling in the design phase.

For general rules for application types, see Table 5-1 on page 111.

The following requirements apply to developing an application I/O profile:

- User population

  Determining the user population requires understanding the total number of potential users, which for an online banking application might represent the total number of customers. From this total population, you must derive the active population that represents the average number of persons that use the application at any specific time, which is derived from experiences with other similar applications.

  In Table 5-3, you use 1% of the total population. From the average population, you can estimate the peak. The peak workload is some multiplier of the average and is typically derived based on experience with similar applications. In this example, we use a multiple of 3.

  Table 5-3 User population

Total potential users	Average active users	Peak active users
50000	500	1500

- Transaction distribution

  Table 5-4 breaks down the number of times that key application transactions are run by the average user and how much I/O is generated per transaction. Detailed application and database knowledge is required to identify the number of I/O’s and the type of I/O’s per transaction. The following information is a sample.

  Table 5-4 Transaction distribution

Transaction	Iterations per user	I/O’s	I/O type
Look up savings account	1	4	Random read
Look up checking account	1	4	Random read
Transfer money to checking	0.5	4 reads/4 writes	Random read/write
Configure new bill payee	0.5	4 reads/4 writes	Random read/write
Submit payment	1	4 writes	Random write
Look up payment history	1	24 reads	Random read

- Logical I/O profile

  An I/O profile is created by combining the user population and the transaction distribution. Table 5-5 provides an example of a logical I/O profile.

  Table 5-5 Logical I/O profile from user population and transaction profiles

Transaction	Iterations per user	I/O’s	I/O type	Average user I/O’s	Peak users
Look up savings account	1	4	Random read (RR) I/O’s	2000	6000
Look up checking account	1	4	RR	2000	6000
Transfer money to checking	0.5	4 reads/4 writes	RR, random write I/O’s (RW)	1000, 1000	3000 R/W

Transaction	Iterations per user	I/O’s	I/O type	Average user I/O’s	Peak users
Configure new bill payee	0.5	4 reads/4 writes	RR, RW	1000, 1000	3000 R/W
Submit payment	1	4 writes	RW	2000	6000 R/W
Look up payment history	1	24 reads	RR	12000	36000

- Physical I/O profile

  The physical I/O profile is based on the logical I/O with the assumption that the database provides cache hits to 90% of the read I/O’s. All write I/O’s are assumed to require a physical I/O. This physical I/O profile results in a read miss ratio of (1 - 0.9) = 0.1 or 10%. Table 5-6 is an example, and every application has different characteristics.

Table 5-6 Physical I/O profile

Transaction	Average user logical I/O’s	Average active users physical I/O’s	Peak active users physical I/O’s
Look up savings account	2000	200 RR	600 RR
Look up checking account	2000	200 RR	600 RR
Transfer money to checking	1000, 1000	100 RR, 1000 RW	300 RR, 3000 RW
Configure new bill payee	1000, 1000	100 RR, 1000 RW	300 RR, 3000 RW
Submit payment	2000	200 RR	600 RR
Look up payment history	12000	1200 SR	3600 RR
Totals	20000 R, and 2000 W	2000 RR 2000 RW	6000 RR 6000 RW

As you can see in Table 5-6, to meet the peak workloads, you must design an I/O subsystem to support 6000 random reads/sec and 6000 random writes/sec:

Physical I/O’s	The number of physical I/O’s per second from the host perspective
RR	Random Read I/O’s
RW	Random Write I/O’s

Determine the appropriate configuration to support your unique workload.

5.5 Understanding your workload type

To understand the workload, you need the performance data from the operating system and from the disk system. Combined analysis of these two sets of performance data can give you the entire picture and help you understand your workload. Separate analysis might be not accurate. This section describes various performance monitoring tools.

5.5.1 Monitoring the DS8000 workload

The following performance monitoring tools are available for the IBM System Storage DS8900:

- IBM Spectrum Control

  IBM Spectrum Control is the tool to monitor the workload on your DS8000 storage for a long period and collect historical data. This tool can also create reports and provide alerts. For more information, see 7.2.1, “IBM Spectrum Control and IBM Storage Insights Pro overview” on page 141.

- IBM Storage Insights

  Is a cloud based service which enables quick deployment to allow optimization of storage and SAN resources using with proprietary analytics from IBM Research. For information on IBM Storage Insights, refer to:

  https://www.ibm.com/products/analytics-driven-data-management

5.5.2 Monitoring the host workload

The following sections list the host-based performance measurement and reporting tools under the UNIX, Linux, Windows, IBM i, and z/OS environments.

Open Systems servers

This section lists the most common tools that are available on Open Systems servers to monitor the workload.

UNIX and Linux Open Systems servers

To get host information about I/O subsystems, processor activities, virtual memory, and the use of the physical memory, use the following common UNIX and Linux commands:

- **`iostat`**
- **`vmstat`**
- **`sar`**
- **`nmon`**
- **`topas`**
- **`filemon`**

These commands are standard tools that are available with most UNIX and UNIX like (Linux) systems. Use **`iostat`** to obtain the data that you need to evaluate your host I/O levels. Specific monitoring tools are also available for AIX, Linux, Hewlett-Packard UNIX (HP-UX), and Oracle Solaris.

Microsoft Windows servers

Common Microsoft Windows Server monitoring tools include the Windows Performance Monitor (**`perfmon`**). Performance Monitor has the flexibility to customize the monitoring to capture various categories of Windows server system resources, including processor and memory. You can also monitor disk I/O by using **`perfmon`**.

To activate:

1. Open Windows Start
2. Search for Performance Monitor and click on perfmon app.

IBM i environment

IBM i provides a vast selection of performance tools that can be used in performance-related cases with external storage. Several of the tools, such as Collection services, are integrated in the IBM i system. Other tools are a part of an IBM i licensed product. The management of many IBM i performance tools is integrated into IBM Systems Director Navigator for IBM i, or into IBM iDoctor.

The IBM i tools, such as Performance Explorer and iDoctor, are used to analyze the hot data in IBM i and to size flash modules for this environment. Other tools, such as Job Watcher, are used mostly in solving performance problems, together with the tools for monitoring the DS8900 storage system such as Storage Insights.

For more information about the IBM i tools and their usage, see 9.4.1, “IBM i performance tools” on page 217.

z Systems environment

The z/OS systems have proven performance monitoring and management tools that are available to use for performance analysis. Resource Measurement Facility (RMF), a z/OS performance tool, collects performance data and reports it for the wanted interval. It also provides cache reports. The cache reports are similar to the disk-to-cache and cache-to-disk reports that are available in IBM Spectrum Control, except that the RMF cache reports are in text format. RMF collects the performance statistics of the DS8900 storage system that are related to the link or port and also to the rank and extent pool. The **REPORTS(ESS)** parameter in the RMF report generator produces the reports that are related to those resources.

The RMF Spreadsheet Reporter is an easy way to create Microsoft Excel Charts based on RMF postprocessor reports. It is used to convert your RMF data to spreadsheet format and generate representative charts for all performance charts for all performance-relevant areas, and is described here:

https://www.ibm.com/docs/en/zos/2.5.0?topic=workstation-rmf-spreadsheet-reporter

For more information, see Chapter 10, “Performance considerations for IBM z Systems servers” on page 225.

5.5.3 Modeling the workload and sizing the system

Workload modeling is used to predict the behavior of the system under the workload to identify the limits and potential bottlenecks, and to model the growth of the system and plan for the future.

IBM and IBM Business Partner specialists use the IBM Storage Modeler (StorM) tool for performance modeling of the workload and capacity planning on the systems.

IBM StorM can be used to help to plan the DS8900 hardware configuration. With IBM StorM, you model the DS8900 performance when migrating from another disk system or when making changes to an existing DS8000 configuration and the I/O workload. IBM StorM is for use with both z Systems and Open Systems server workloads. In addition, IBM StorM also models storage capacity requirements.

You can model the following major DS8000 components by using IBM StorM:

- Supported DS8000 model: DS8884F, DS8884, DS8886F, DS8886, DS8888F, DS8882F, DS8980F, DS8910F and DS8950F models
- Capacity sizing in IBM i
- Importing data for performance assessments

- Configuration capacity using flash modules and RAID type
- Type and number of DS8000 host adapters (HAs)
- Remote Copy options

When working with IBM StorM, always ensure that you input accurate and representative workload information because IBM StorM results depend on the input data that you provide. Also, carefully estimate the future demand growth that you input to IBM StorM for modeling projections. The hardware configuration decisions are based on these estimates.

For more information about using StorM, see Chapter 6.1, “IBM Storage Modeller” on page 126.

Sizing the system for the workload

With the performance data collected or estimated, and the model of the data that is created, you can size the planned system. Systems are sized with application-specific tools that are provided by the application vendors. There are several tools:

- General storage sizing: StorM.
- MS Exchange sizing: MS Exchange sizing tool. You can find more information about this tool at the following websites:
  - http://technet.microsoft.com/en-us/library/bb124558.aspx
  - http://technet.microsoft.com/en-us/library/ff367907.aspx
- Oracle, Db2, SAP NetWeaver: IBM Techline and specific tools.

Workload testing

There are various reasons for conducting I/O load tests. They all start with a hypothesis and have defined performance requirements. The objective of the test is to determine whether the hypothesis is true or false. For example, a hypothesis might be that you think that a DS8900 storage system with 18 flash arrays and 256 GB of cache can support 100,000 IOPS with a 70/30/50 workload and the following response time requirements:

- Read response times: 95th percentile < 10 ms
- Write response times: 95th percentile < 5 ms

To test, complete the following generic steps:

1. Define the hypothesis.
2. Simulate the workload by using an artificial or actual workload.
3. Measure the workload.
4. Compare workload measurements with objectives.
5. If the results support your hypothesis, publish the results and make recommendations. If the results do not support your hypothesis, determine why and make adjustments.

Microsoft Windows environment

The following example tests might be appropriate for a Windows environment:

- Pre-deployment hardware validation. Ensure that the operating system, multipathing, and host bus adapter (HBA) drivers are at the current levels and supported. Before you deploy any solution and especially a complex solution, such as Microsoft cluster servers, ensure that the configuration is supported. Refer to the following interoperability website for more information:

  http://www.ibm.com/systems/support/storage/ssic/interoperability.wss

- Application-specific requirements. Often, you receive inquiries about the DS8000 storage system and Microsoft Exchange. To simulate MS Exchange 2019 storage I/O load on a server, use the MS Exchange Server Jetstress 2016 tool. For more information, refer to:

https://www.microsoft.com/en-us/download/details.aspx?id=36849

The universal workload generator and benchmark tool is the Iometer (`http://www.iometer.org`). Iometer is both a workload generator (it performs I/O operations to stress the system) and a measurement tool (it examines and records the performance of its I/O operations and their effect on the system). It can be configured to emulate the disk or network I/O load of any program or benchmark, or it can be used to generate entirely synthetic I/O loads. It can generate and measure loads on single or multiple (networked) systems.

Iometer can be used for the following measurements and characterizations:

- Performance of disk and network controllers
- Bandwidth and latency capabilities of buses
- Network throughput to attached drives
- Shared bus performance
- System-level hard disk drive performance
- System-level network performance

You can use Iometer to configure these settings:

- Read/write ratios
- Sequential/random
- Arrival rate and queue depth
- Block Size
- Number of concurrent streams

With these configuration settings, you can simulate and test most types of workloads. Specify the workload characteristics to reflect the workload in your environment.

Unix and Linux environment

The Unix and Linux **dd** command is a very useful tool to drive sequential read workloads or sequential write workloads against the DS8900 storage system.

From 2019 Subsystem Device Driver Specific Module (SDDDSM) and Subsystem Device Driver Path Control Module (SDDPCM) are no longer supported for DS8000. Users will need to use native operating system device drivers for multipath support. For information, refer to:

https://www.ibm.com/support/pages/sdddsm-and-sddpcm-end-support-ds8000

This section describes how to perform these tasks using Multipath I/O devices:

Useful Multipath I/O commands for AIX

- lsattr -El hdisk*xx* - lists device attributes
- lsmpio - l hdisk*xx* - lists the number of available paths to a logical volume
  - Determine the sequential read speed that an individual logical unit number (LUN)) can provide in your environment.
  - Measure sequential read and write speeds for file systems.

To test the sequential read speed of a rank, run the following command:

```
time dd if=/dev/rhdiskxx of=/dev/null bs=128k count=781
```

The rhdisk*xx* is the character or raw device file for the logical unit numbers (LUN) that is presented to the operating system by Multipath I/O (MPIO). This command reads 100 MB off rhdisk*xx* and reports how long it takes in seconds. Take 100 MB and divide by the number of seconds that is reported to determine the MBps read speed.

Run the following command and start the **nmon** monitor or **iostat -k 1** command in Linux:

```
dd if=/dev/rhdiskxx of=/dev/null bs=128k
```

Your **nmon** monitor (the **e** option) reports that this previous command imposed a sustained 100 MBps bandwidth with a blocksize=128 K on rhdiskxx. Notice the xfers/sec column; xfers/sec is IOPS. Now, if your **dd** command did not error out because it reached the end of the disk, press Ctrl+c to stop the process. Now, **nmon** reports an idle status. Next, run the following **dd** command with a 4 KB blocksize and put it in the background:

```
dd if=/dev/rhdiskxx of=/dev/null bs=4k &
```

For this command, **nmon** reports a lower MBps but a higher IOPS, which is the nature of I/O as a function of blocksize. Run your **dd** sequential read command with a bs=1024 and you see a high MBps but a reduced IOPS.

The following commands perform sequential writes to your LUNs:

- `dd if=/dev/zero of=/dev/rhdiskxx bs=128k`
- `dd if=/dev/zero of=/dev/rhdiskxx bs=1024k`
- `time dd if=/dev/zero of=/dev/rhdiskxx bs=128k count=781`

Try different block sizes, different raw hdisk devices and combinations of reads and writes. Run the commands against the block device (`/dev/hdiskxx`) and notice that block size does not affect performance.

Because the **dd** command generates a sequential workload, you still must generate the random workload. You can use a no-charge open source tool, such as Vdbench.

Vdbench is a disk and tape I/O workload generator for verifying data integrity and measuring the performance of direct-attached and network-connected storage on Windows, AIX, Linux, Solaris, OS X, and HP-UX. It uses workload profiles as the inputs for the workload modeling and has its own reporting system. All output is presented in HTML files as reports and can be analyzed later. For more information, refer to:

http://www.oracle.com/technetwork/server-storage/vdbench-downloads-1901681.html

AIX Tuning Parameters

There are several disk and disk adapter kernel tunable parameters in the AIX operating system that are useful when used in conjunction with DS8000 storage servers.

- Disk Adapter Outstanding-Requests Limit
- Fibre Channel Adapter Outstanding-Requests Limit
- Disk Drive Queue Depth

For full details see the following website:

https://www.ibm.com/docs/en/aix/7.2?topic=parameters-disk-disk-adapter-tunable

6

Performance planning tools

This chapter gives an overview about the IBM Storage Modeller, which is the storage modeling and sizing tool for IBM storage systems.

We also consider special aspects of the sizing like planning for Multi-Target PPRC, or working with DS8000 storage systems using Easy Tier.

6.1 IBM Storage Modeller

IBM Storage Modeler (StorM) is a sizing and modeling tool for IBM Storage Systems, based on an open client/server architecture, prepared to be connected to other tools, using a web-based graphical user interface. Its ability to collaborate by sharing project files is a huge differentiator, and the interactive canvas enables easy design and visualization of a solution architecture. IBM Storage Modeler is platform agnostic and runs on most current web browsers.

Only IBMers and registered Business Partners can access StorM.

A precise storage modeling, for both capacity and performance, is paramount, and the StorM tool follows the latest mathematical modeling for it. Storage systems which are supported include the DS8000 models starting with the DS8880 (POWER8 based) generation.

Some aspects of the outgoing Disk Magic tool are still mentioned here, given that this tool can also model DS8000 generations before DS8880 or given its still excellent means of importing collected host performance data and doing a quick number crunching on them. Disk Magic however does not pick up anymore the latest support for product features coming new in 2021 and later.

This chapter gives and idea about the basic usage of these tools. Clients who want to have such a study run best contact their IBM or BP representative for it.

6.1.1 The need for performance planning and modeling tools

There are two ways to achieve your planning and modeling:

- Perform a complex workload benchmark based on your workload.
- Do a modeling by using performance data retrieved from your system.

Performing an extensive and elaborate lab benchmark by using the correct hardware and software provides a more accurate result because it is real-life testing. Unfortunately, this approach requires much planning, time, and preparation, plus a significant amount of resources, such as technical expertise and hardware/software in an equipped lab.

Doing a study with the sizing tool requires much less effort and resources. Retrieving the performance data of the workload and getting the configuration data from the servers and the storage systems is all that is required from the client. Especially when replacing a storage system, or when consolidating several, these tools are valuable. In the past we often had small amounts of flash like 5% and less, and then many 10K-rpm HDDs in multi-frame installations. Now we sharply shrink footprint by switching to all-flash installations with some major amount of high-capacity flash, and often we consolidate several older and smaller DS8000 models into one bigger of a newer generation: all that can be simulated. But we can also simulate an upgrade for an existing storage system that we want to keep for longer: For instance: Does it help to upgrade cache, and what would be the positive effect on latency; or am I having a utilization problem with my host bus adapters, and how many should I add to get away with it.

Now for collecting your performance data, especially when on host side: What matters is, that the time frames are really representative, and also for most busy times of a year even. And: Different applications can peak at different times. For example, a processor-intensive online application might drive processor utilization to a peak while users are actively using the system. However, the bus or drive utilization might be at a peak when the files are backed up during off-hours. So, you might need to model multiple intervals to get a complete picture of your processing environment.

6.1.2 Configuration data of your legacy storage systems

Generally, to perform a sizing and modeling study, we should obtain the following information about our existing storage systems, that we want to upgrade or to potentially replace:

- Control unit type and model
- Cache size
- Disk drive sizes (flash, HDD), their numbers and their speeds, and RAID formats
- Number, type, and speed of channels
- Number and speed of host adapter (HA) and (for Z:) FICON ports
- Parallel access volume (PAV) for Z
- Speed in the SAN, and when this is due for replacement

For Remote Copy, you need the following information:

- Remote Copy type(s) (synchronous, asynchronous; any multi-target?)
- Distance
- Number of links and speed

But also the capacity planning must be clear: is today‘s capacity to be retained 1:1? Or would it grow further, and by what factor. And for the performance: shall we also factor in some load growth, and to what extent.

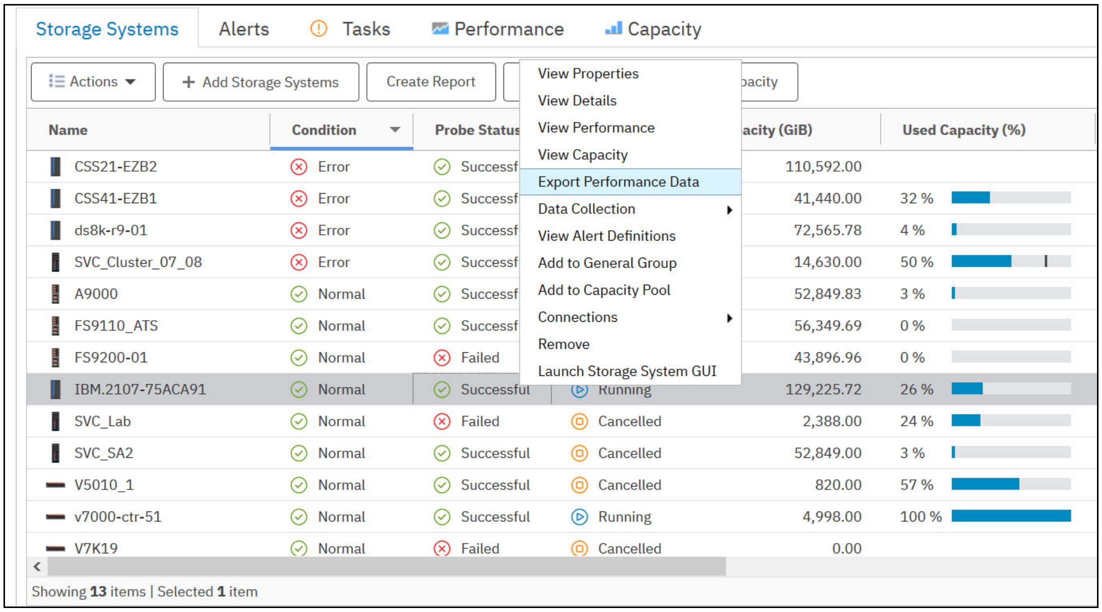
If architectural changes are planned, like introducing further multi-target replication, or introducing Safeguarded Copy, it should definitely be noted and taken into account.

A part of these data can be obtained from either the collected Z host performance data, or when using a tool like IBM Spectrum Control, or IBM Storage Insights.

6.2 Data from IBM Spectrum Control and Storage Insights

IBM Storage Insights, or Spectrum Control, are strategic single-pane-of-glass monitoring tools that can be recommended for every IBM storage customer. Since the recommendation also is to have a measurement from your real storage system and workload data, some monitoring tools like these can be used. Especially for distributed systems, where a some large-scale host based performance monitoring (like z/OS clients using RMF) is typically not available.

In both Storage Insights and IBM Spectrum Control, when going to the Block Storage Systems and then doing a right-mouse click, the menu opens to “Export Performance Data”, as in Figure 6-1 on page 128. This performance data package gives valuable and detailed performance and configuration information, for instance, on drive types and quantity, RAID type, extent pool structure, number of HBAs, ports, and their speeds, if PPRC is used, and more.



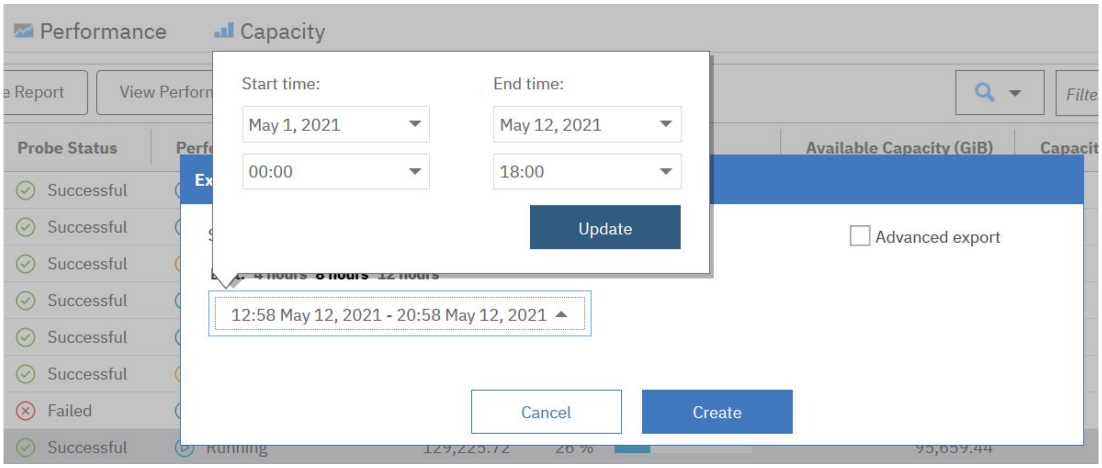
Name	Condition	Probe Status		acity (GiB)	Used Capacity (%)
CSS21-EZB2	Error	Successf		110,592.00	
CSS41-EZB1	Error	Successf		41,440.00	32 %
ds8k-r9-01	Error	Successf		72,565.78	4 %
SVC_Cluster_07_08	Error	Successf		14,630.00	50 %
A9000	Normal	Successf		52,849.83	3 %
FS9110_ATS	Normal	Successf		56,349.69	0 %
FS9200-01	Normal	Failed		43,896.96	0 %
IBM.2107-75ACA91	Normal	Successful	Running	129,225.72	26 %
SVC_Lab	Normal	Failed	Cancelled	2,388.00	24 %
SVC_SA2	Normal	Successful	Cancelled	52,849.00	3 %
V5010_1	Normal	Successful	Cancelled	820.00	57 %
v7000-ctr-51	Normal	Successful	Running	4,998.00	100 %
V7K19	Normal	Failed	Cancelled	0.00	

Showing **13** items | Selected **1** item

Figure 6-1 SC/SI Export Performance Data package

It also shows those ports that are not used or RAID arrays outside pools and, therefore, not used.

Before exporting the package, extend the period from the default of a few hours to the desired longer interval, incorporating many, and representative, busy days, as in Figure 6-2.



Start time: May 1, 2021 00:00

End time: May 12, 2021 18:00

Update

□ Advanced export

12:58 May 12, 2021 - 20:58 May 12, 2021

Cancel Create

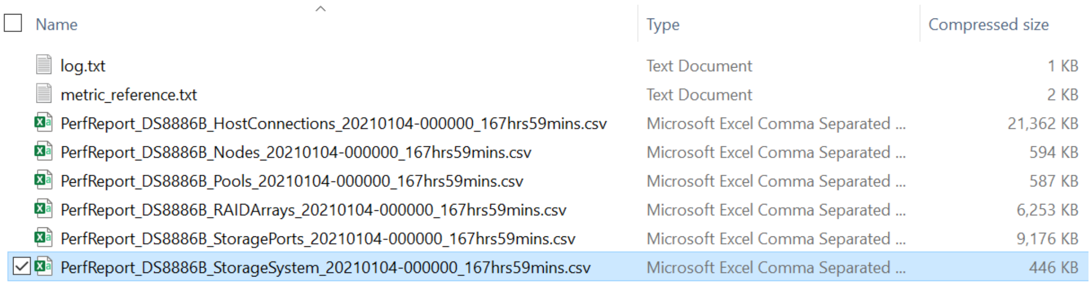
Figure 6-2 SC/SI exporting performance data: Select the right time duration

You can consider may aspects of the configuration, including PPRC (synchronous, asynchronous; what is the distance in km, Metro/Global Mirror, or multi-site).

The resulting performance data come in the following format:

SC_PerfPkg_<DS8000_name>_<date>_<duration>.ZIP

and contain mainly CSV files, as shown in Figure 6-3 on page 129.



☐ Name	Type	Compressed size
log.txt	Text Document	1 KB
metric_reference.txt	Text Document	2 KB
PerfReport_DS8886B_HostConnections_20210104-000000_167hrs59mins.csv	Microsoft Excel Comma Separated ...	21,362 KB
PerfReport_DS8886B_Nodes_20210104-000000_167hrs59mins.csv	Microsoft Excel Comma Separated ...	594 KB
PerfReport_DS8886B_Pools_20210104-000000_167hrs59mins.csv	Microsoft Excel Comma Separated ...	587 KB
PerfReport_DS8886B_RAIDArrays_20210104-000000_167hrs59mins.csv	Microsoft Excel Comma Separated ...	6,253 KB
PerfReport_DS8886B_StoragePorts_20210104-000000_167hrs59mins.csv	Microsoft Excel Comma Separated ...	9,176 KB
☑ PerfReport_DS8886B_StorageSystem_20210104-000000_167hrs59mins.csv	Microsoft Excel Comma Separated ...	446 KB

Figure 6-3 DS8000 SC/SI Performance Data Package content

The StorageSystem CSV file is essential, as it contains the view on the entire storage system in terms of performance data to the SAN and to the hosts. But for setting up your model, the other ones also give you important information beforehand:

- From the Nodes CSV, get some information on the DS8000 cache available and used.
- In the Pools CSV, get an information how many extent pools exist.
- In the RAIDArrays CSV, get the drive types used, and the type of RAID formatting for each rank (e.g., RAID-6). When you see arrays which are consistent with zero load, don't take them into your modeling.
- In the StoragePorts CSV, see how many HBAs are installed, how many ports thereof are in use, and their speeds. Also, you see how many of these HBAs and ports do PPRC traffic.
  Look at different points of time: If some ports (or even HBAs) are consistent with zero load, we take them out of the modeling.
- Looking at the HostConnections CSV, you might want to check if, eventually, not all hosts take part in the PPRC.
- The StorageSystem CSV finally tells you about the exact DS8000 model (e.g., 986) and finally gives the primary peak data for the modeling. Again here, you also can check the amount of PPRC activity.

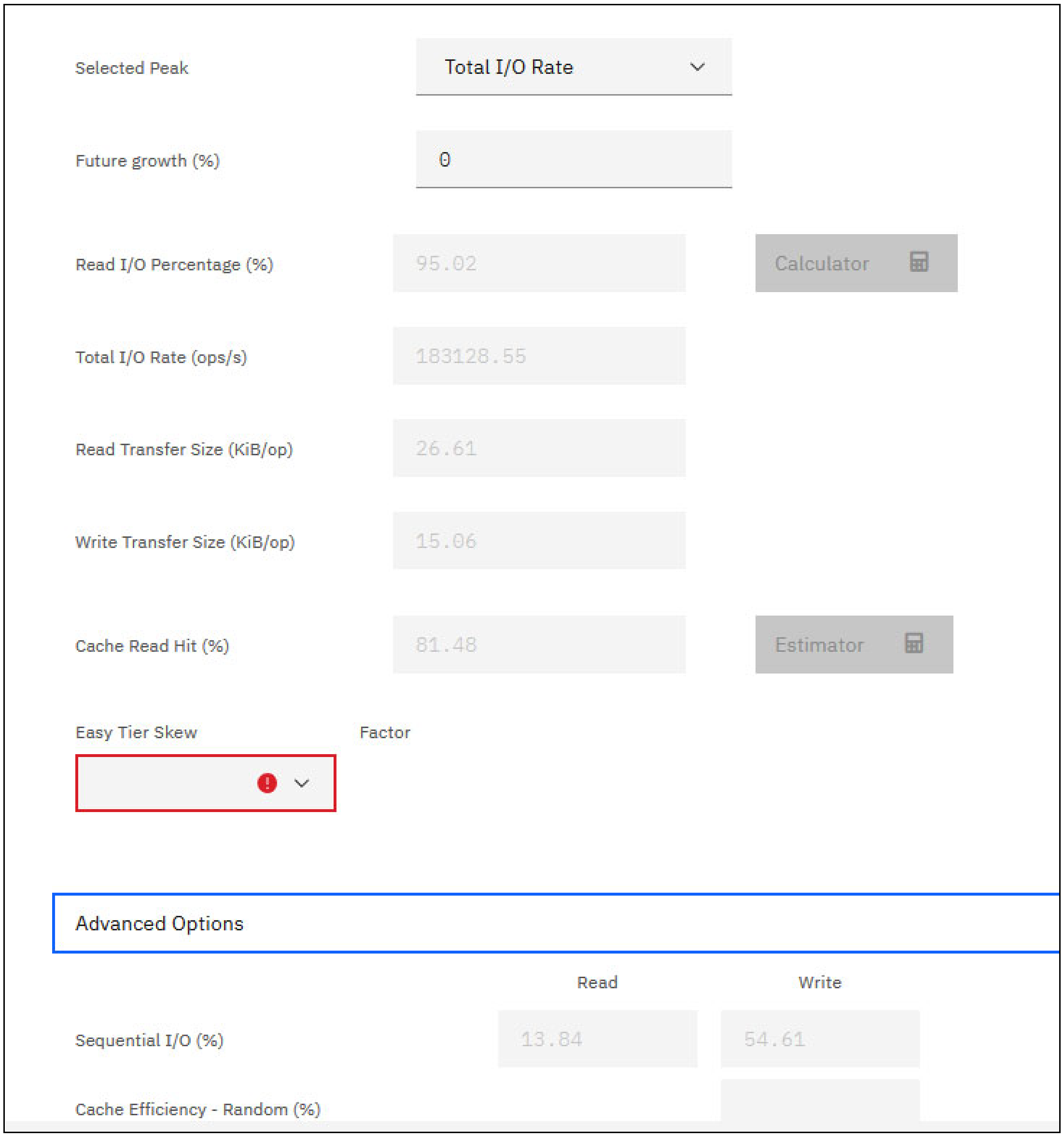
May also check if there are any exceptionally high values somewhere, which gives additional ideas about why a customer might ask for exchanging the storage system.

All this information from the several CSVs can build up your configuration of an existing DS8000. But competition storage systems can be also monitored by SC/SI.

The CSV, or ZIP file, can then be imported into StorM, in an automated way, to let the tool determine the peak load intervals, and use them for your monitoring.

Workload profiles

After importing the workload file into the StorM tool, we see several categories there that typically make up a full workload profile: The Read I/O Percentage (or, Read:Write ratio), the transfer sizes for reads and writes respectively, the read-cache hit ratio, and the sequential ratios for reads and for writes. The latter ones are optional, in case you would enter a workload manually, and hence are under the Advanced Options, as in Figure 6-4 on page 130.



Selected Peak	Total I/O Rate	
Future growth (%)	0	
Read I/O Percentage (%)	95.02	Calculator
Total I/O Rate (ops/s)	183128.55	
Read Transfer Size (KiB/op)	26.61	
Write Transfer Size (KiB/op)	15.06	
Cache Read Hit (%)	81.48	Estimator
Easy Tier Skew	Factor	

Advanced Options

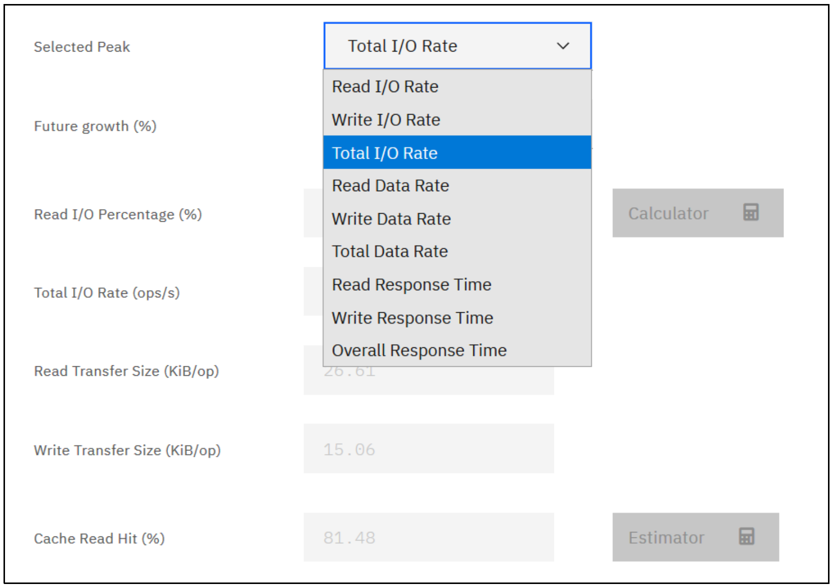
	Read	Write
Sequential I/O (%)	13.84	54.61
Cache Efficiency - Random (%)		

Figure 6-4   Open Systems workload after import into StorM

The workload peak profiles thus obtained can then be scaled further with a factor for “Future growth”. For this value, it is even possible to enter negative values like -50 %, for instance in case you would want to cut your workload in half when splitting it into pools and into the StorM “Data Spaces”.

The Cache Read Hit percentage is shown exactly as measured from the previous legacy storage system. However in case you already are modeling a replacement system with a bigger cache, can switch from automated input to manual and then use the Estimator button to calculate the new cache-hit ratio on the new storage system.

Once the data are read in automatically, you have up to 9 different kinds of peak workload intervals to choose from when doing your modeling, like for instance the Total I/O Rate (= usually the most important overall), or the Total Data Rate (= also important and should always be looked at), or response time peaks (= which are interesting to consider because here we often have low cache-hit ratios, and deviating read:write ratios or suddenly higher block sizes), or the peak of the Write Data Rate (= this one should be additionally considered for instance when doing modeling for PPRC projects, but also flash drives can be limited by a high write throughput). See Figure 6-5 for this selection of the peak intervals.



Selected Peak: Total I/O Rate

Future growth (%)

Read I/O Percentage (%)

Total I/O Rate (ops/s)

Read Transfer Size (KiB/op): 26.61

Write Transfer Size (KiB/op): 15.06

Cache Read Hit (%): 81.48

Figure 6-5 Peak interval selection in StorM

6.3 StorM with IBM i

Modeling for an IBM i host can also be done by gathering storage data via IBM Storage Insights or Spectrum Control. When the customer mixes IBM i with other platforms on their storage system, SC/SI can be the preferred choice for the performance data collection.

However, a significant number of IBM i clients is entirely focused on that platform, and they prefer to collect their storage performance data on the host platform.

On host side, one of the options is then the Performance Tools (PT1) reports. IBM i clients sometimes can have very heavy short-term I/O peaks, and if it’s an i-only platform, then a data collection with the PT1 reports can also give you a precise idea of what is happening on the storage side.

Such PT1 (Performance Tools) reports consist of:

- System Report
  - Disk Utilization
  - Storage Pool Utilization
- Component Report
  - Disk Activity
- Resource Interval Report
  - Disk Utilization Summary

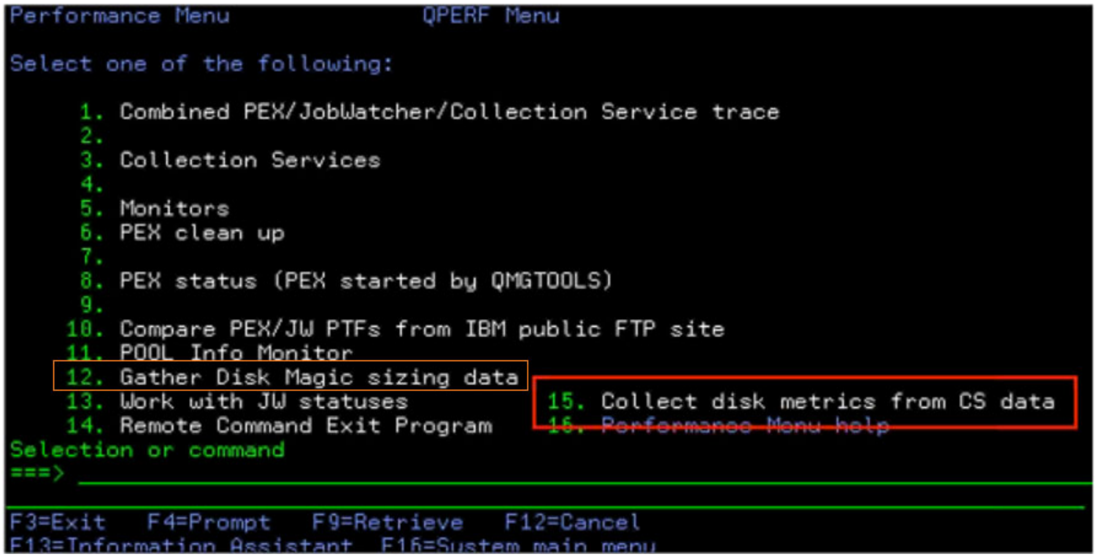
From the PT1 reports, IO peaks can be manually transfered into StorM.

With StorM, a newer reporting option can be used. It is also based on the IBM MustGather Data Capture tool for IBM i. A QMGTOOLS build of 23 June 2021 or later is necessary to exploit that new reporting option.

Use a collection interval of 5 min or less with it, and find more background on this tool and method at:

https://www.ibm.com/support/pages/node/684977

Figure 6-6 shows in the QMGTOOLS QPERF Menu, which options exist to gather data for modeling storage performance.



```
Performance Menu                QPERF Menu

Select one of the following:

     1. Combined PEX/JobWatcher/Collection Service trace
     2.
     3. Collection Services
     4.
     5. Monitors
     6. PEX clean up
     7.
     8. PEX status (PEX started by QMGTOOLS)
     9.
    10. Compare PEX/JW PTFs from IBM public FTP site
    11. POOL Info Monitor
    12. Gather Disk Magic sizing data
    13. Work with JW statuses              15. Collect disk metrics from CS data
    14. Remote Command Exit Program        16. Performance Menu help
Selection or command
===> 

F3=Exit   F4=Prompt   F9=Retrieve   F12=Cancel
F13=Information Assistant  F16=System main menu
```

Figure 6-6 QMGTOOLS QPERF menu: Gathering data for storage performance modeling

The newer option now is 15, to collect the needed metrics that can directly be used with the StorM tool.

IIBM i has a very low Easy Tier skew typically. Unless you have a precisely measured skew factor, enter “Easy Tier Skew = Very Low” (Factor 2.0) in the Performance tab of StorM.

6.4 Modeling and sizing for IBM Z

In a z/OS environment, gathering data for performance modeling requires the System Management Facilities (SMF) record types 70 - 78. Namely, for a representative time period, would extract the following SMF records: **73, 74-1, 74-5, 74-7, 74-8, 78-3**.

The SMF data are packed by using **`RMFPACK`**. The installation package, when expanded, shows a user guide and two XMIT files: RMFPACK.JCL.XMIT and RMFPACK.LOAD.XMIT. To pack the SMF data set into a `ZRF` file, complete the following steps:

1. Install **`RMFPACK`** on your z/OS system.
2. Prepare the collection of SMF data.
3. Run the $1SORT job to sort the SMF records.
4. Run the $2PACK job to compress the SMF records and to create the `ZRF` file.

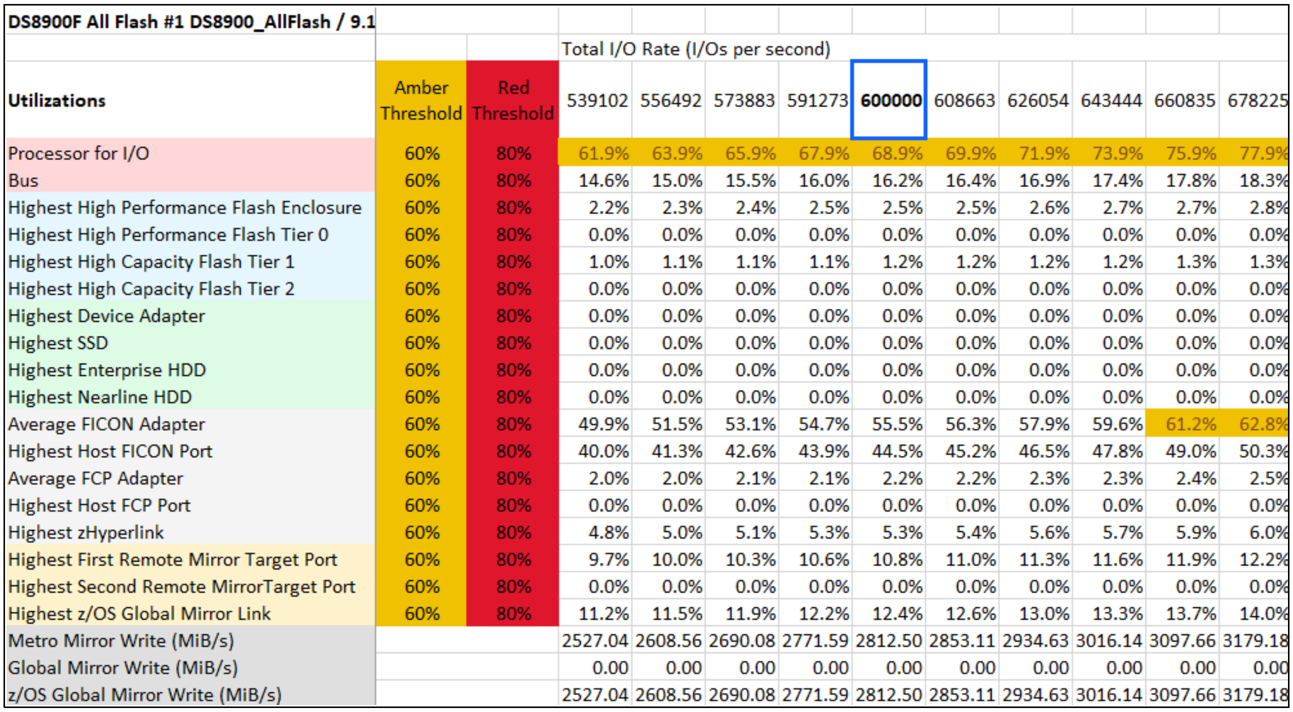
For most mainframe customers, sampling periods of 15 min are fine. The main thing is to cover those days which are really busy, as they can occur only at certain times in a month, or even a year potentially. So 1 or 2 busy days to have performance data of could be sufficient, but have the right days for this, and cover both nights and daytimes, as workload patterns in the night (batch) often differ from more transaction-processing workloads during daytimes.

Storage modeling output: Utilization

The performance modeling checks for several things, not only that it calculates the expected new latency (response time) for a replacement storage system, or upgraded storage system. It can also calculate the expected internal utilization levels of all the various components in such a storage system.

Figure 6-7 shows the modeling of DS8000 component utilization for a given peak workload interval. We can make the following observations:

- For the selected (peak) workload of 600 KIOps, and with the selected number of drives, host bus adapters, amount of cache, number of DS8000 processor cores and so on, how busy each internal component would with this model. We see utilization numbers for the FICON adapters and for the FICON ports, to be around 50%. This utilization rate could easily be lowered, by adding more HBAs even though the utilization is still below the critical “amber” threshold value.
- The utilization numbers for the HPFE and for the Flash Tier 1 drives (consisting of 3.84 TB drives) are extremely low. As a result, you could consider a cheaper configuration using 7.68 TB Flash Tier 2 drives.
- Expected utilization levels for the internal DS8950 bus and for the processing capability shows an expected utilization of about 69%, which is a bit too high. Assuming the storage system is currently a DS8950F with dual 10-core processors, you could upgrade it to a dual 20-core instead.



DS8900F All Flash #1 DS8900_AllFlash / 9.1												
			Total I/O Rate (I/Os per second)									
Utilizations	Amber Threshold	Red Threshold	539102	556492	573883	591273	**600000**	608663	626054	643444	660835	678225
Processor for I/O	60%	80%	61.9%	63.9%	65.9%	67.9%	68.9%	69.9%	71.9%	73.9%	75.9%	77.9%
Bus	60%	80%	14.6%	15.0%	15.5%	16.0%	16.2%	16.4%	16.9%	17.4%	17.8%	18.3%
Highest High Performance Flash Enclosure	60%	80%	2.2%	2.3%	2.4%	2.5%	2.5%	2.5%	2.6%	2.7%	2.7%	2.8%
Highest High Performance Flash Tier 0	60%	80%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Highest High Capacity Flash Tier 1	60%	80%	1.0%	1.1%	1.1%	1.1%	1.2%	1.2%	1.2%	1.2%	1.3%	1.3%
Highest High Capacity Flash Tier 2	60%	80%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Highest Device Adapter	60%	80%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Highest SSD	60%	80%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Highest Enterprise HDD	60%	80%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Highest Nearline HDD	60%	80%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Average FICON Adapter	60%	80%	49.9%	51.5%	53.1%	54.7%	55.5%	56.3%	57.9%	59.6%	61.2%	62.8%
Highest Host FICON Port	60%	80%	40.0%	41.3%	42.6%	43.9%	44.5%	45.2%	46.5%	47.8%	49.0%	50.3%
Average FCP Adapter	60%	80%	2.0%	2.0%	2.1%	2.1%	2.2%	2.2%	2.3%	2.3%	2.4%	2.5%
Highest Host FCP Port	60%	80%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Highest zHyperlink	60%	80%	4.8%	5.0%	5.1%	5.3%	5.3%	5.4%	5.6%	5.7%	5.9%	6.0%
Highest First Remote Mirror Target Port	60%	80%	9.7%	10.0%	10.3%	10.6%	10.8%	11.0%	11.3%	11.6%	11.9%	12.2%
Highest Second Remote MirrorTarget Port	60%	80%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Highest z/OS Global Mirror Link	60%	80%	11.2%	11.5%	11.9%	12.2%	12.4%	12.6%	13.0%	13.3%	13.7%	14.0%
Metro Mirror Write (MiB/s)			2527.04	2608.56	2690.08	2771.59	2812.50	2853.11	2934.63	3016.14	3097.66	3179.18
Global Mirror Write (MiB/s)			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
z/OS Global Mirror Write (MiB/s)			2527.04	2608.56	2690.08	2771.59	2812.50	2853.11	2934.63	3016.14	3097.66	3179.18

Figure 6-7 Storage Systems component Utilization chart for a DS8950F

We also see that this is a storage system actively using zHyperLink, Metro Mirror, and z/OS Global Mirror. Mirroring consumes a significant percentage of the internal storage system processing resources and must be modeled specifically. Metro Mirror also triggers important FICON adapters usage, and a Metro Mirror modeling will allow you to see the throughput at the DS8000 host adapter which could be too high. It could also be limited by the bandwidth of your Wide Area Network. You must also take into consideration that a typical workload increases overtime, and that your Metro Mirroring bandwidth will have to increase by the same factor.

The sizing must be done in a way that optimizes the utilization of all internal components while making certain that all peak workload profiles stay below the critical amber or red threshold. When coming close to these thresholds, elongated response time occurs and in hitting the threshold the system latency will start to exhibit a non-linear behavior and any workload increase at that stage will deteriorate the response time. This is a situation to be definitely avoided.

Modeling output: Expected response time

StorM can predict for the given workload profile, that the response time is expected to be 0.34 ms. But there is more information: IBM Storage Modeler predicts when the response time starts increasing exponentially, and it gives more information about the individual response time components.

What do the Response time Components tell you?

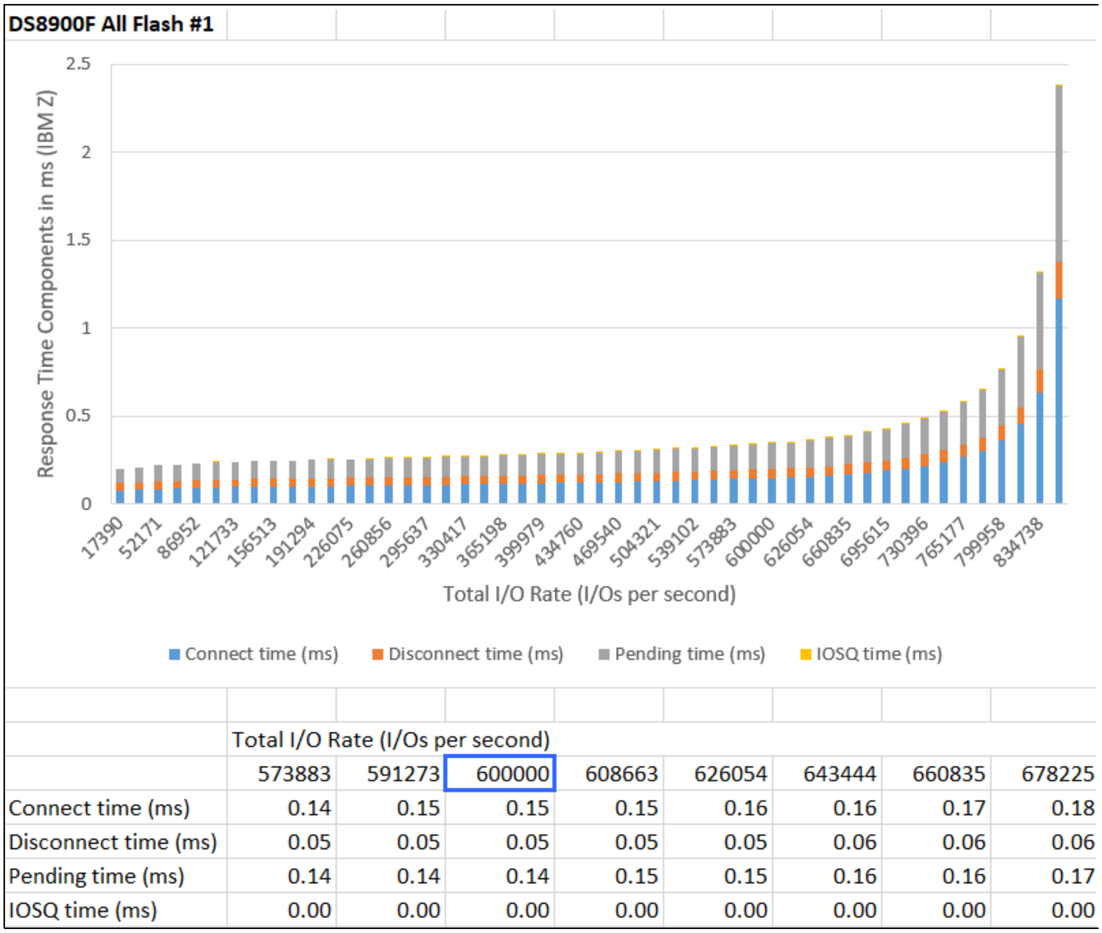
Figure 6-8 on page 135 provides the average response time for a minimum to maximum workload period for each interval for the measured data. The response time components are Connect time, Disconnect time, Pending time, and IOSQ time:

- Connect time: The device was connected to a channel path and transferring data between the device and central storage (processor). High connect time occurs because of contention in the FICON channels.
- Disconnect time: The value reflects the time when the device was in use but not transferring data. Because the device is likely waiting for Metro Mirror copy, Random-read workload (read miss), or Write de-stage activity.
- Pending time: I/O request must wait in the hardware between the processor and storage system. This value also includes the time waiting for an available channel path and control unit as well as the delay due to shared DASD contention.
- IOSQ time: I/O request must wait on an IOS queue before the disk becomes free.

Being at 0.34 ms overall response for the 600 KIOps workload, we can compare these expected 0.34 ms (new/upgraded system) with what the customer has today, and make sure that there is an improvement. But we can also make sure that we are still in some near-linear range of workload increase, so that further I/O increases will not lead to sudden bigger spikes in latency. The total latency is shown with the four components that make up the response time.

Being below 700 KIOps here the response time remains at a lower level. This indicates the configured system has enough room for growth against the customer's peak workload data which was collected using RMF Utility. Based on the modeled latency and response time curve, the proposed storage system is suitable for a customer proposal.

But, again, we can go back to the Utilization figures and make sure that these are all in a lower “green”' range first. If not the case, expand and upgrade the configuration still. So both aspects, expected utilization levels and expected response time values, need to fit - or the configuration upgraded.



	Total I/O Rate (I/Os per second)							
	573883	591273	600000	608663	626054	643444	660835	678225
Connect time (ms)	0.14	0.15	0.15	0.15	0.16	0.16	0.17	0.18
Disconnect time (ms)	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.06
Pending time (ms)	0.14	0.14	0.14	0.15	0.15	0.16	0.16	0.17
IOSQ time (ms)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Figure 6-8 Response time components chart for DS8950F

IBM Z Batch Network Analyzer (zBNA) tool

zBNA is a PC-based productivity tool designed to provide a means of estimating the elapsed time for batch jobs. It provides a powerful, graphic demonstration of the z/OS batch window. The zBNA application accepts SMF Record Types 14, 15, 16, 30, 42, 70, 72, 74, 78 and 113 data to analyze such a single batch window of user-defined length.

In case no zHyperLink is used yet, the zBNA tool can for instance be used to determine how big the zHyperLink ratio is expected to be, once zHyperLink is switched on. Find the tool and more background information at:

https://www.ibm.com/support/pages/node/6354321

6.5 Easy Tier modeling

StorM supports DS8000 Easy Tier modeling for multitier configurations. The performance modeling is based on the workload skew level concept. The skew level describes how the I/O activity is distributed across the capacity for a specific workload. A workload with a low skew level has the I/O activity distributed evenly across the available capacity. A heavily skewed workload has many I/Os to only a small portion of the data. The workload skewing affects the Easy Tier effectiveness, especially if there is a limited number of high performance ranks. The heavily skewed workloads benefit most from the Easy Tier capabilities because even when moving a small amount of data, the overall performance improves. With lightly skewed workloads, Easy Tier is less effective because the I/O activity is distributed in such a large amount of data that it cannot be moved to a higher performance tier.

Lightly skewed workloads might be solved already with a 100% High-Capacity Flash design, and in many cases, this can even mean using Flash Tier 2 drives only. Whereas with a higher skew and higher load, a mix of High-Performance Flash (Flash Tier 0) and High-Capacity Flash (then mostly Flash Tier 2) is usually done, and the question is, which capacity ratio to best choose between the various tiers.

There are three different approaches on how to model Easy Tier on a DS8000. Here are the three options:

- Use one of the predefined skew levels.
- Use an existing skew level based on the current workload on the current DS8000 (measuring on the host side).
- Use heatmap data from a DS8000 (measuring on storage system side).

6.5.1 Predefined skew levels

StorM supports five predefined skew levels for use in Easy Tier predictions:

- Skew level 2: Very low skew
- Skew level 3.5: Low Skew
- Skew level 7: Intermediate skew
- Skew level 14: High skew
- Skew level 24: Very high skew

Storm uses this setting to predict the number of I/Os that are serviced by the higher performance tier.

A skew level value of 1 means that the workload does not have any skew at all, meaning that the I/Os are distributed evenly across all ranks.

The skew level settings affect the modeling tool predictions. A heavy skew level selection results in a more aggressive sizing of the higher performance tier. A low skew level selection provides a conservative prediction. It is important to understand which skew level best matches the actual workload before you start the modeling.

6.5.2 Getting the exact skew

For IBM Z workloads, getting an exact skew level is very easy, when you have RMF measurements. Additionally, most Open Systems clients have skew levels between “high” and “intermediate”. IBM i clients, however, typically have a “very low” skew level. In case you do some modeling for IBM i and could not retrieve an exact value for this level, best go with the assumption of “very low” (2.0), for i.

Your IBM support or Business Partner personnel have access to additional utilities that allows them to retrieve the exact skew directly from the DS8000, from the Easy Tier summary.

At the DS8000 Graphical User Interface, you can, at the top, export the Easy Tier Summary, as in Figure 6-9:

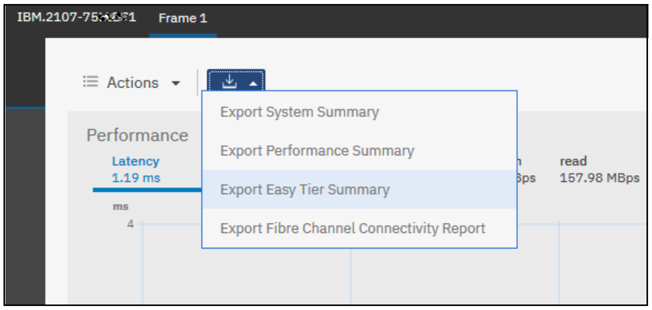


Figure 6-9   DS8900 GUI: Export Easy Tier Summary

The output ZIP that is created contains CSV file reports on the Easy Tier activity that can be used for additional planning and more accurate sizing. The skew curve is available in a format to help you differentiate the skew according to whether the skew is for small or large blocks, reads or writes, and IOPS or MBps, and it is described in more detail in *IBM DS8000 Easy Tier*, REDP-4667.

The comparable DSCLI command, which yields the same data, is shown in Example 6-1.

Example 6-1   DSCLI export of Easy Tier data

```
dscli> offloadfile -etdataCSV \tmp
Date/Time: 13 July 2021 14:51:55 CEST IBM DSCLI Version: 7.9.20.440 DS: IBM.2107-75HAL91
CMUC00428I offloadfile: The etdataCSV file has been offloaded to
\tmp\et_data_20210713145155.zip.
```

7

Practical performance management

This chapter describes the tools, data, and activities that are available for supporting the DS8000 performance management processes by using IBM Storage Insights and IBM Spectrum Control.

This chapter includes the following topics:

- Introduction to practical performance management
- Performance management tools
- IBM Storage Insights Pro data collection considerations
- IBM Storage Insights Pro performance metrics
- IBM Storage Insights Pro reporting options
- Using IBM Storage Insights Pro network functions
- End-to-end analysis of I/O performance problems
- Performance analysis examples
- IBM Storage Insights Pro in mixed environments

7.1 Introduction to practical performance management

Performance management processes go along with inputs, actors, and roles. Performance management processes include operational processes, such as data collection and alerting, tactical processes, such as performance problem determination and analysis, and strategic processes, such as long-term trending. This chapter defines the tools, metrics, and processes that are required to support the operational, tactical, and strategic performance management processes by using IBM Spectrum Control, IBM Storage Insights and IBM Storage Insights Pro. These tools do very similar functions with the following differences. IBM Spectrum Control is installed on premises, IBM Storage Insights is available free of charge to all IBM block storage customers with a subset of the full functionality in the paid version of IBM Storage Insights Pro. Both of these are in the IBM cloud with a data collector installed inside the data center. With a license for IBM Spectrum Control the user is entitled to IBM Storage Insights Pro.

7.2 Performance management tools

Tools for collecting, monitoring, and reporting on DS8000 performance are critical to the performance management processes. The tools provide support for the DS8000 performance management processes with the following features, which are briefly described in Table 7-1. Furthermore, this chapter describes some of the most important Tactical/Strategic items.

Table 7-1 IBM Spectrum Control and Storage Insights supported activities for performance processes

Process	Activities	Feature
Operational	Performance data collection for ports, arrays, volumes, pools, nodes (formerly called controllers), and host connections. Switch performance metrics can also be collected.	Performance monitor jobs.
Operational	Alerting.	Alerts and threshold violations.
Tactical/ Strategic	Performance reporting of ports, pools, array, volumes, nodes, host connections, and switches.	Web-based GUI, IBM Cognos® (including predefined reports), and TPCTOOL.
Tactical	Performance analysis and tuning.	Tool facilitates thorough data collection and reporting.
Tactical	Short-term reporting.	GUI charting with the option to export data to analytical tools.
Tactical	Advanced Analytics in IBM Spectrum Control.	Tiering and Balancing Analysis.
Strategic	Long-term reporting.	GUI charting with the option to export data to analytical tools.

Additional performance management processes that complement IBM Spectrum Control are shown in Table 7-2 on page 141.

Table 7-2 Additional tools

Process	Activity	Alternative
Strategic	Sizing	Storage Modeler and general rules (see Chapter 6, "Performance planning tools" on page 125)
Strategic	Planning	Logical configuration performance considerations (see Chapter 4, "Logical configuration performance considerations" on page 59).
Operational	Host data collection performance and alerting	Native host tools.
Tactical	Host performance analysis and tuning	Native host tools.
Operational	Monitoring performance of SAN switches	Native switch tools.

7.2.1 IBM Spectrum Control and IBM Storage Insights Pro overview

IBM Spectrum Control and IBM Storage Insights Pro reduce the complexity of managing SAN storage devices by allowing administrators to configure, manage, and monitor storage devices and switches from a single console. Most of the examples that will be used in this chapter will focus on IBM Storage Insights Pro. Just remember that you have the ability to see as much or more in IBM Spectrum Control and a subset for a shorter period of time in IBM Storage Insights.

IBM Storage Insights

IBM Storage Insights provides a level of visibility across your storage environment to help you manage complex storage infrastructures including IBM block storage systems as well as SAN fabric devices. Is is a cloud-based service that enables you to deploy quickly and save storage administration time while optimizing your storage. It allows you to monitor the basic health, status and performance of your storage systems. It also helps automate aspects of the support process to enable faster resolution of issues by allowing you to open tickets directly from the web browser interface, upload data to support and monitor your tickets. Storage Insights requires a data collector to be installed in the data center to gather capacity and performance statistics that get uploaded to the IBM Cloud®. The user is able to view all of the information available in Storage Insights for 24 hours.

IBM Storage Insights Pro

IBM Storage Insights Pro is based on a capacity based subscription. It includes all of the features found in the basic version plus a more comprehensive view of performance, capacity and health. There are additional reporting features available and the data collected can be viewed for up to two years.

IBM Spectrum Control

IBM Spectrum Control is installed on a server in the data center and includes all the capabilities of Storage Insights/Pro however it does not send data to IBM. It can maintain the data at the user's specified time frame, it collects data from the monitored systems as a more frequent rate, and has the ability to integrate with system's management applications (such as LDAP). If the user is licensed for IBM Spectrum Control they can also use Storage Insights Pro at no additional cost.

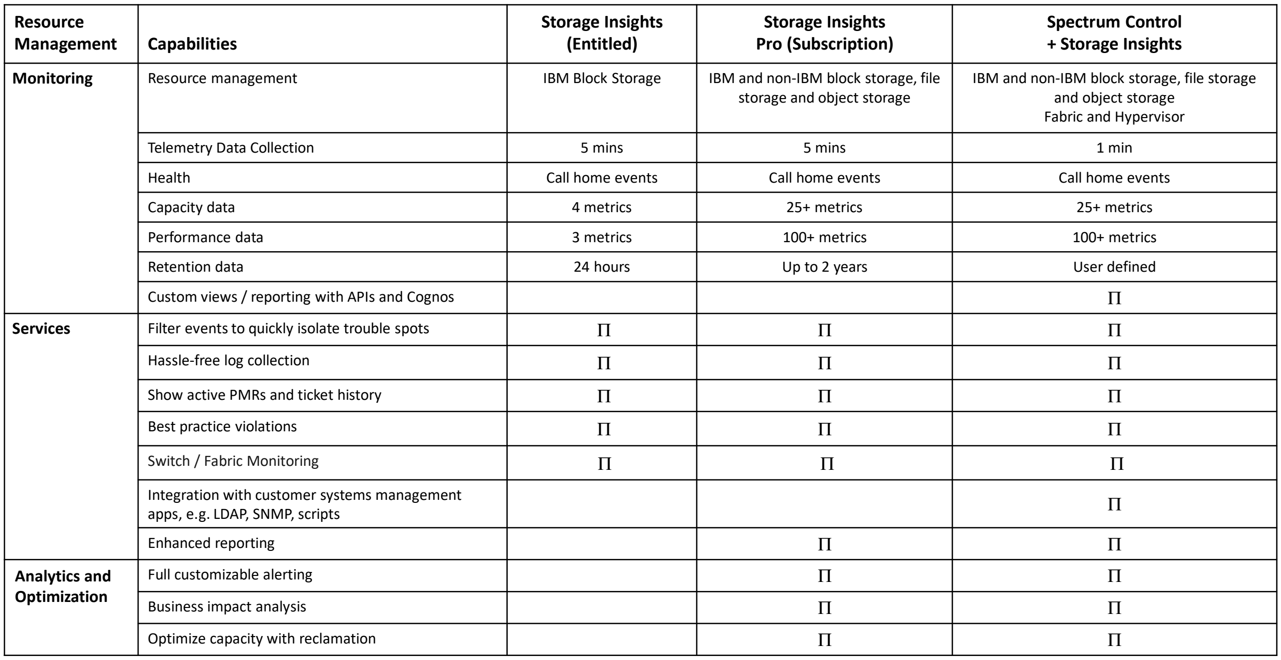
For a full list of the features that are provided in IBM Spectrum Control and Storage Insights, go to the following IBM websites:

- https://www.ibm.com/products/spectrum-control
- https://www.ibm.com/products/analytics-driven-data-management

For more information about the configuration and deployment of storage by using IBM Spectrum Control, see these publications:

- *IBM Spectrum Family: IBM Spectrum Control Standard Edition*, SG24-8321
- *Regain Control of your Environment with IBM Storage Insights,* REDP-5231

The table in Figure 7-1 on page 142 provides a side by side comparison of these very powerful options.



Resource Management	Capabilities	Storage Insights (Entitled)	Storage Insights Pro (Subscription)	Spectrum Control + Storage Insights
Monitoring	Resource management	IBM Block Storage	IBM and non-IBM block storage, file storage and object storage	IBM and non-IBM block storage, file storage and object storage Fabric and Hypervisor
	Telemetry Data Collection	5 mins	5 mins	1 min
	Health	Call home events	Call home events	Call home events
	Capacity data	4 metrics	25+ metrics	25+ metrics
	Performance data	3 metrics	100+ metrics	100+ metrics
	Retention data	24 hours	Up to 2 years	User defined
	Custom views / reporting with APIs and Cognos			Π
Services	Filter events to quickly isolate trouble spots	Π	Π	Π
	Hassle-free log collection	Π	Π	Π
	Show active PMRs and ticket history	Π	Π	Π
	Best practice violations	Π	Π	Π
	Switch / Fabric Monitoring	Π	Π	Π
	Integration with customer systems management apps, e.g. LDAP, SNMP, scripts			Π
	Enhanced reporting		Π	Π
Analytics and Optimization	Full customizable alerting		Π	Π
	Business impact analysis		Π	Π
	Optimize capacity with reclamation		Π	Π

Figure 7-1 Capabilities for IBM Storage Insights, Storage Insights Pro and Spectrum Control

On the IBM Spectrum Control and IBM Storage Insights Pro Overview window of a selected DS8000 storage system, performance statistics for that device are displayed. Figure 7-2 on page 143 is an example of IBM Spectrum Control. While IBM Storage Insights Pro has a slightly different look and feel, the same information can be seen when selecting the storage system, as shown in Figure 7-3.

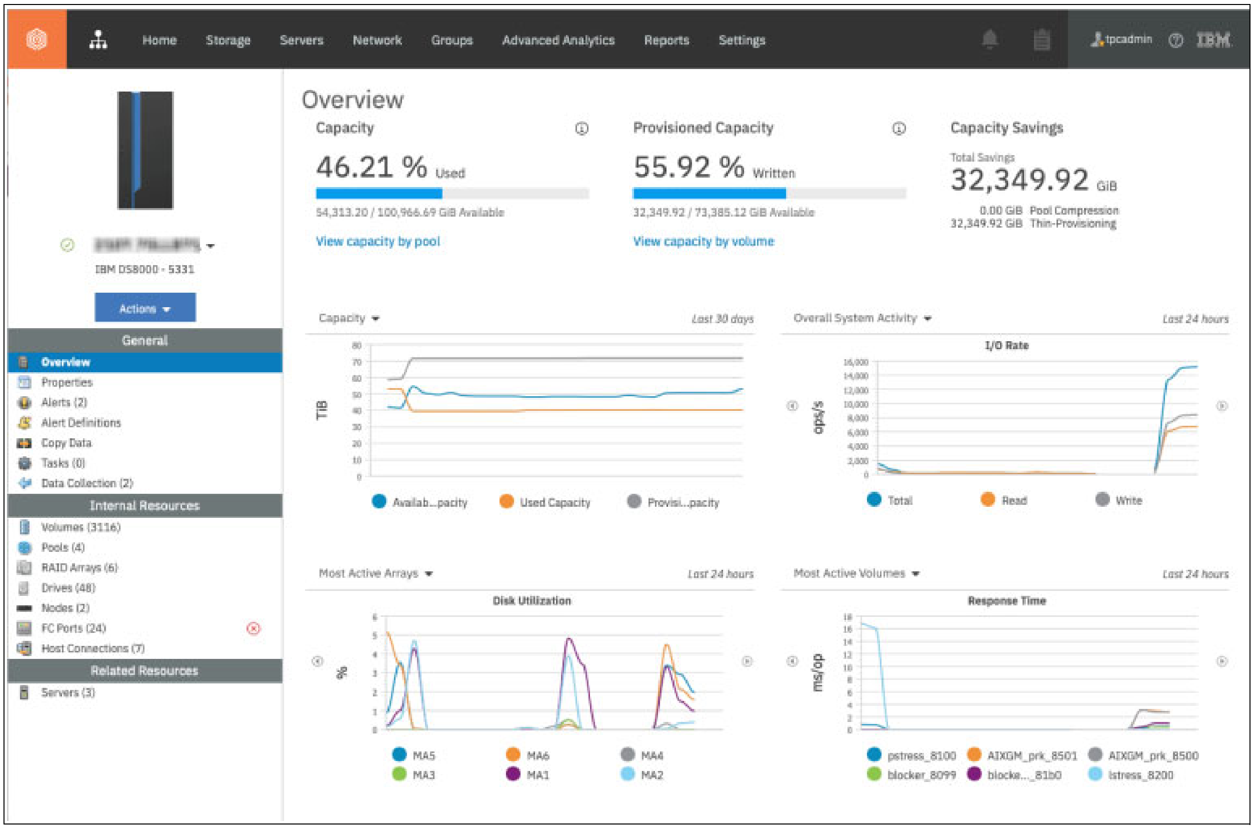


Figure 7-2 Overview window of a DS8000 storage system in IBM Spectrum Control

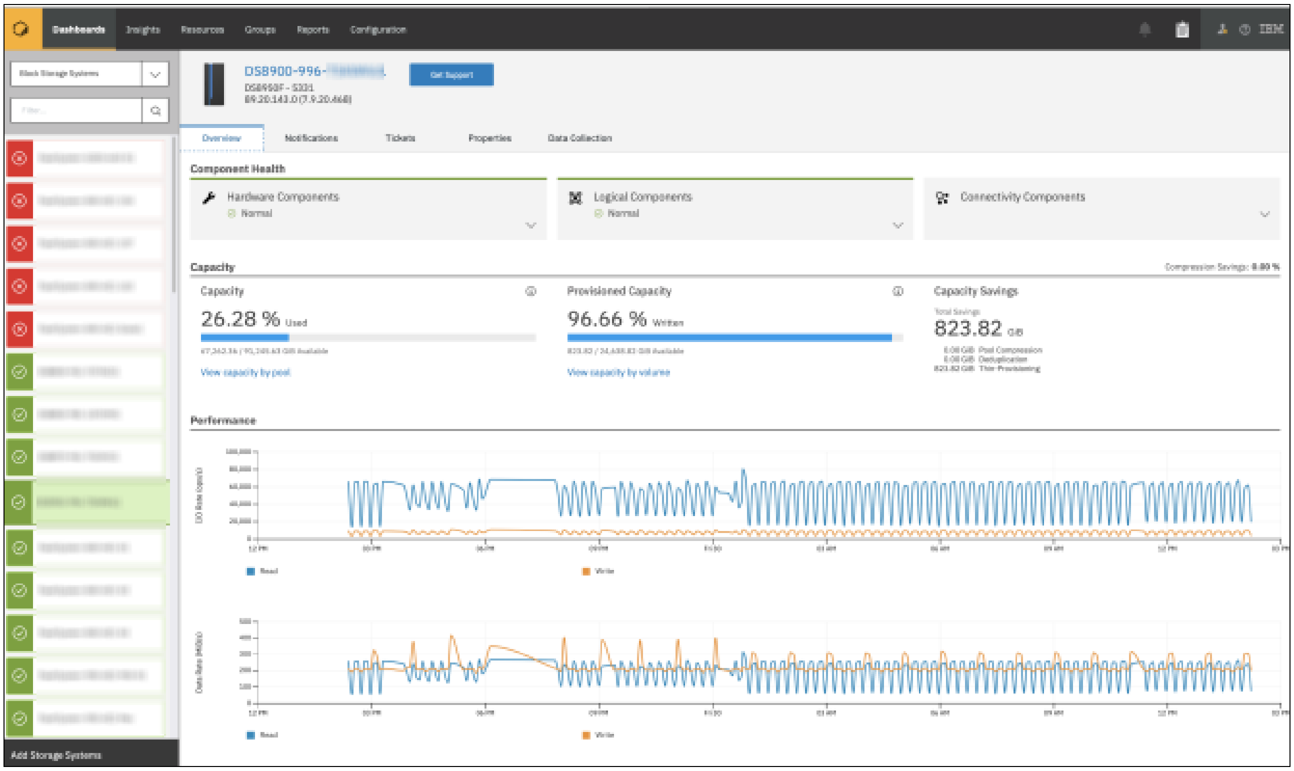


Figure 7-3 IBM Storage Insights Pro Overview

In the left navigation section of IBM Spectrum Control, you can select the aggregated status of internal or related resources. The **Actions** dropdown menu allows you do view performance, capacity as well as additional functions supported in IBM Spectrum Control. The dropdown list next to the system name allows you to select any one of the systems being monitored. This is shown in Figure 7-4 on page 144.

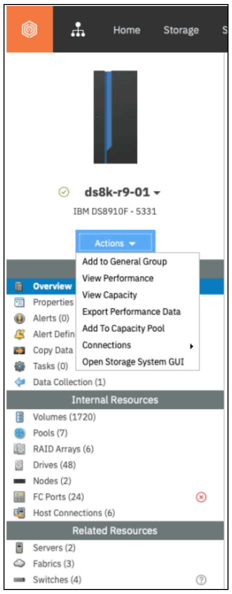


Figure 7-4 IBM Spectrum Control resources and actions

To view similar items in IBM Storage Insights Pro simply click on the down arrow under the category or one of the tabs as shown in Figure 7-5 on page 144.

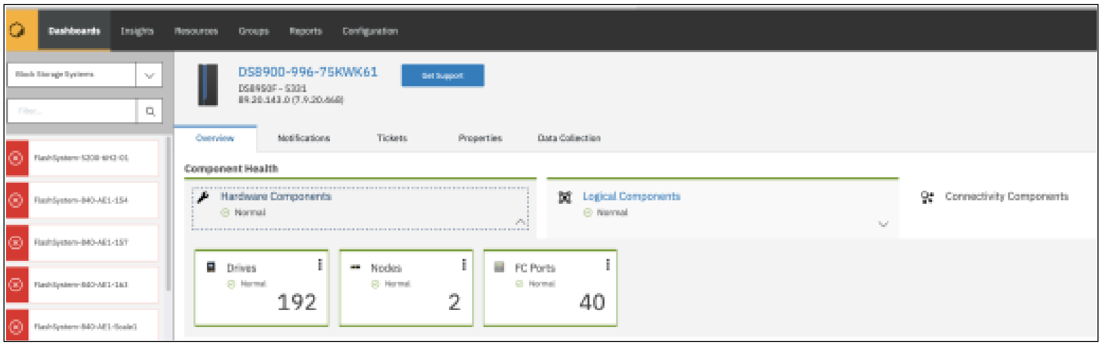


Figure 7-5 Finding resources in IBM Storage Insights.

7.2.2 IBM Spectrum Control measurement of DS8000 components

IBM Spectrum Control and Storage Insights/Pro can gather information about the component levels, as shown in Figure 7-6, for the DS8000 storage system. Displaying a metric depends on the ability of the storage system and its mechanisms to provide the performance data and related information, such as the usage of its components. Figure 7-6 drills down from the top-level subsystem view to give you a better understanding of how and what data is displayed.

Both IBM Spectrum Control and Storage Insights Pro monitor metrics for the following internal resources:

- Volumes
- Pools
- Arrays
- Nodes
- Fibre Channel Ports
- Host Connections

Additional monitoring is available for servers, fabrics and switches.

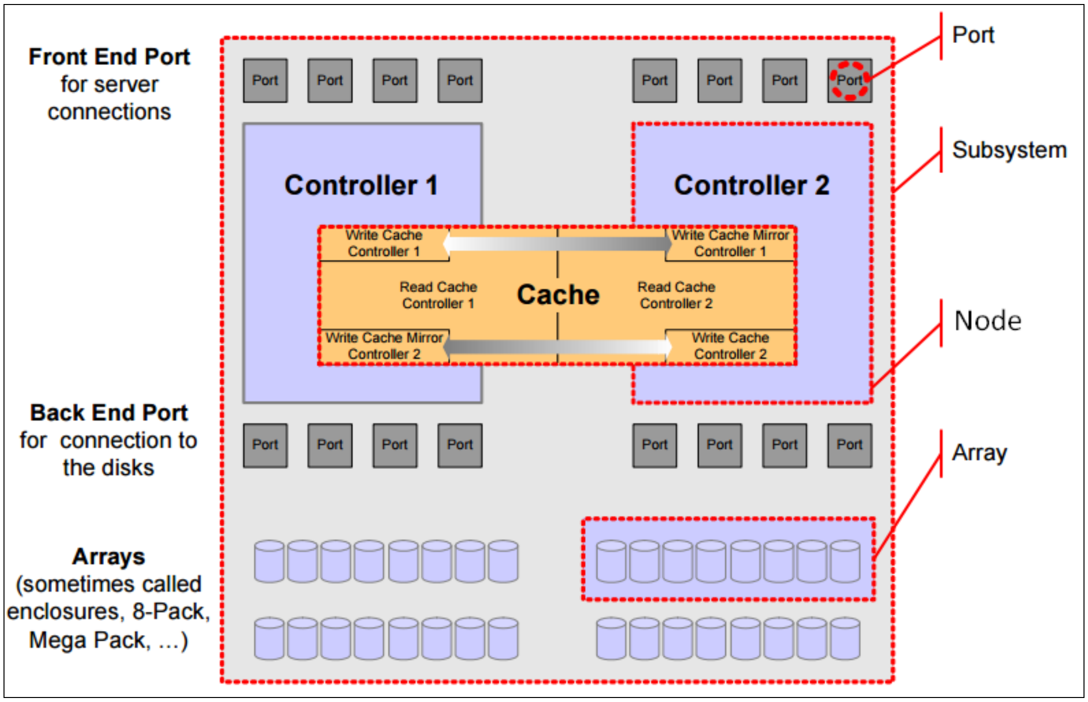


Figure 7-6 Physical view compared to IBM Spectrum Control Performance Report Items

Metrics: A *metric* is a numerical value that is derived from the information that is provided by a device. It is the raw data and a calculated value. For example, the raw data is the transferred bytes, but the metric uses this value and the interval to show the bytes/second.

For the DS8000 storage system, the native application programming interface (NAPI) is used to collect performance data, in contrast to the SMI-S Standard that is used for third-party devices.

The DS8000 storage system interacts with the NAPI in the following ways:

- Access method used: Enterprise Storage Server® Network Interface (ESSNI)
- Failover:
  - For the communication with a DS8000 storage system, IBM Spectrum Control and Storage Insights use the ESSNI client. This library is basically the same library that is included in any DS8000 command-line interface (DSCLI). Because this component has built-in capabilities to fail over from one Hardware Management Console (HMC) to another HMC, a good approach is to specify the secondary HMC IP address of your DS8000 storage system.

  - During failover the query job may fail, but the next command that is sent to the device will use the redundant connection.
- Network: No special network considerations exist. IBM Spectrum Control and Storage Insights need to be able to connect to the HMC.

Subsystem

On the subsystem level, metrics are aggregated from multiple records to a single value per metric to give the performance of a storage subsystem from a high-level view, based on the metrics of other components. This aggregation is done by adding values, or calculating average values, depending on the metric.

Cache

The cache in Figure 7-6 plays a crucial role in the performance of any storage subsystem.

Metrics, such as disk-to-cache operations, show the number of data transfer operations from drives to cache. The number of data transfer operations from drives to cache is called staging for a specific volume. Disk-to-cache operations are directly linked to read activity from hosts. When data is not found in the DS8000 cache, the data is first staged from back-end drives into the cache of the DS8000 storage system and then transferred to the host.

Read hits occur when all the data that is requested for a read data access is in cache. The DS8000 storage system improves the performance of read caching by using Sequential Prefetching in Adaptive Replacement Cache (SARC) staging algorithms. For more information about the SARC algorithm, see 1.2.2, “Advanced caching algorithms” on page 9. The SARC algorithm seeks to store those data tracks that have the greatest probability of being accessed by a read operation in cache.

The cache-to-disk operation shows the number of data transfer operations from cache to drives, which is called destaging for a specific volume. Cache-to-disk operations are directly linked to write activity from hosts to this volume. Data that is written is first stored in the persistent memory (also known as nonvolatile storage (NVS)) at the DS8000 storage system and then destaged to the back-end drive. The DS8000 destaging is enhanced automatically by striping the volume across all the drives in one or several ranks (depending on your configuration). This striping, or volume management that is done by Easy Tier, provides automatic load balancing across DDMs in ranks and an elimination of the hot spots.

The Write-cache Delay I/O Rate or Write-cache Delay Percentage because of persistent memory allocation gives you information about the cache usage for write activities. The DS8000 storage system stores data in the persistent memory before sending an acknowledgment to the host. If the persistent memory is full of data (no space available), the host receives a retry for its write request. In parallel, the subsystem must destage the data that is stored in its persistent memory to the back-end drive before accepting new write operations from any host.

If a volume experiences delays in the write operations due to persistent memory constraint, consider moving, consider moving the volume to a less busy rank or spread this volume on multiple ranks (increase the number of DDMs used). If this solution does not fix the persistent memory constraint problem, consider adding cache capacity to your DS8000 storage system.

As shown in Figure 7-7 on page 147 and Figure 7-8, you can use IBM Spectrum Control and Storage Insights Pro to monitor the cache metrics easily.

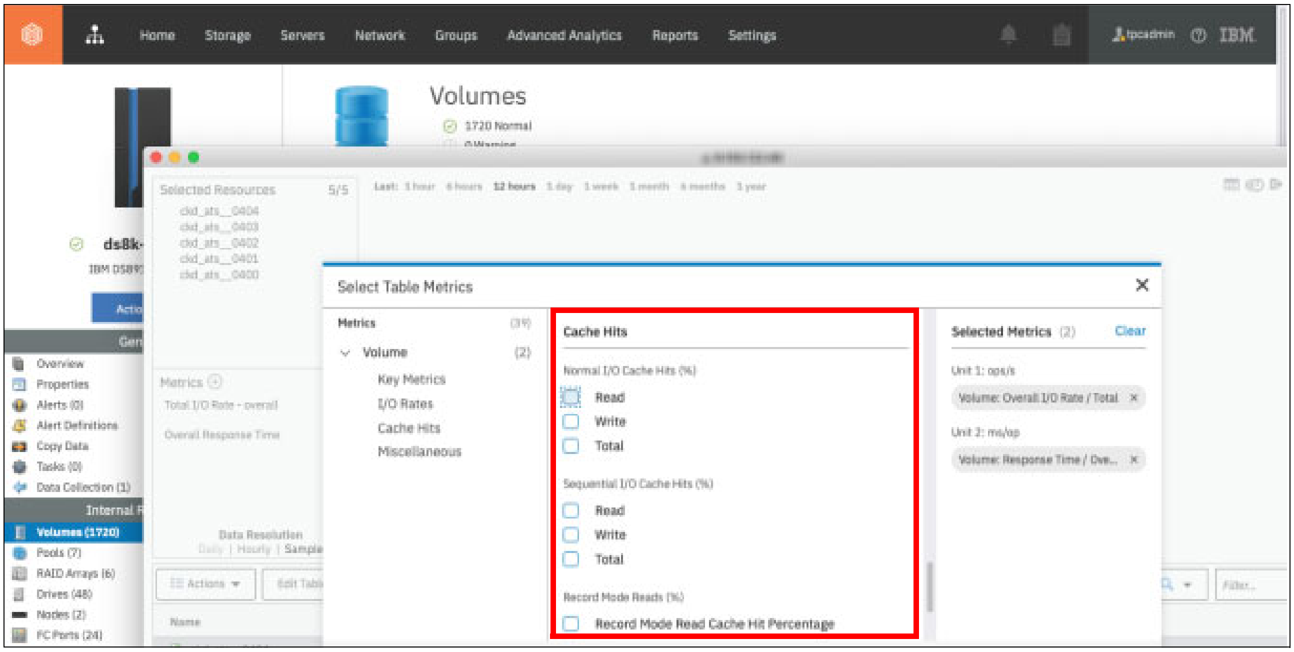


Figure 7-7 Available cache metrics in IBM Spectrum Control

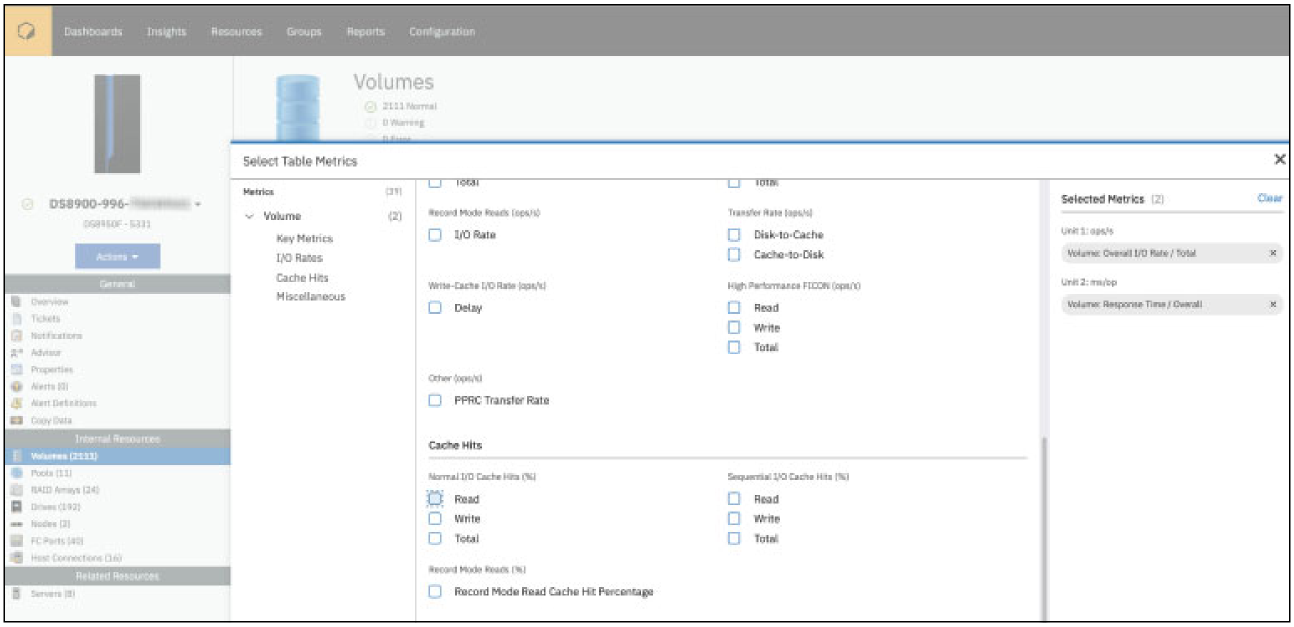
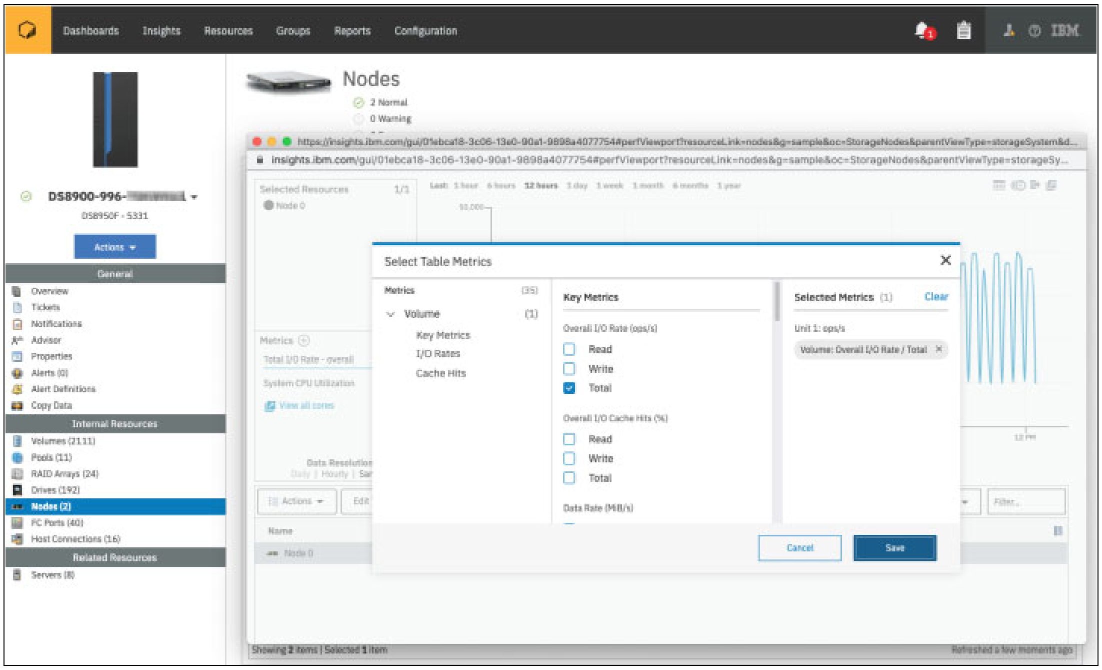


Figure 7-8 Cache Metrics in Storage Insights Pro

Controller/Nodes

The DS8000 processor complexes are referred to as Nodes. A DS8000 storage system has two processor complexes, and each processor complex independently provides major functions for the storage system. Examples include directing host adapters (HAs) for data transfer to and from host processors, managing cache resources, and directing lower device interfaces for data transfer to and from the flash media. To analyze performance data, you must know that most volumes can be assigned/used by only one controller at a time.

Node metrics can be selected in IBM Storage Insights Pro from the Nodes view and Edit Table Metrics, as shown in Figure 7-9 on page 148.



Select Table Metrics

Metrics	(35)
Volume	(1)
Key Metrics	
I/O Rates	
Cache Hits	

Key Metrics

Overall I/O Rate (ops/s)
- □ Read
- □ Write
- ☑ Total

Overall I/O Cache Hits (%)
- □ Read
- □ Write
- □ Total

Data Rate (MiB/s)

Selected Metrics (1) Clear

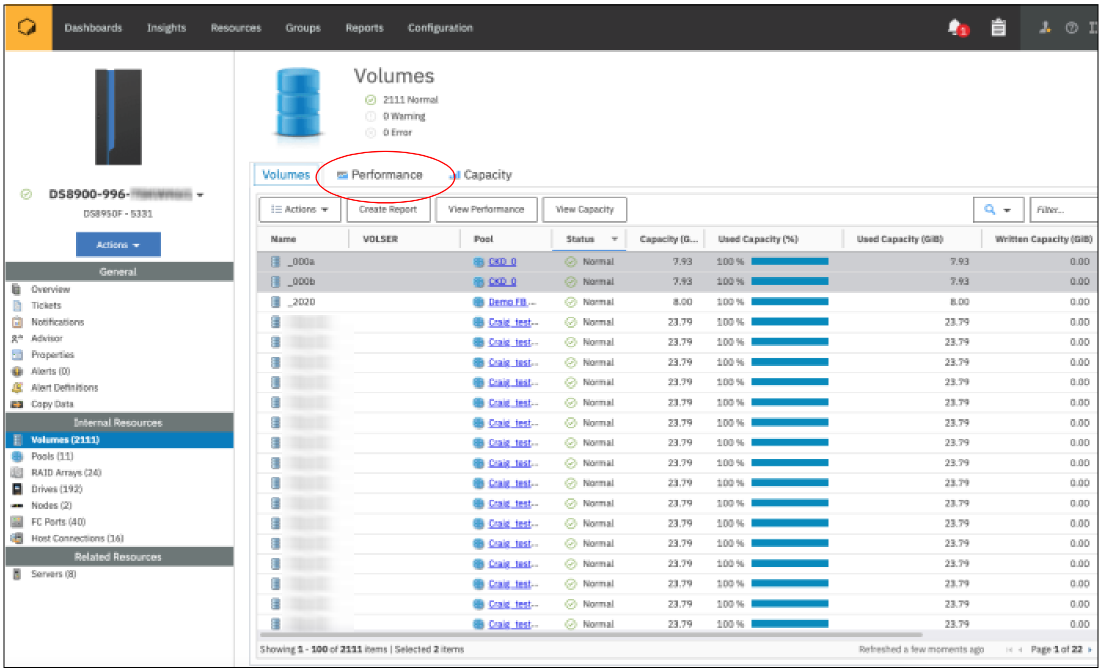
Unit 1: ops/s

Volume: Overall I/O Rate / Total ✕

Cancel Save

Figure 7-9 Select node metrics in IBM Storage Insights Pro

You can view the performance of the volumes (all or a subset) by selecting Volumes in the Internal Resources section and then selecting the Performance tab, as shown in Example 7-10.



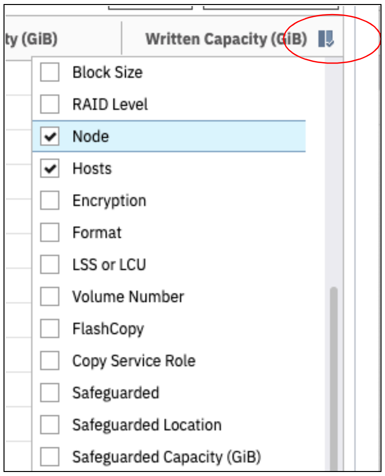
Name	VOLSER	Pool	Status	Capacity (G...	Used Capacity (%)	Used Capacity (GiB)	Written Capacity (GiB)
_000a		CKD_0	Normal	7.93	100 %	7.93	0.00
_000b		CKD_0	Normal	7.93	100 %	7.93	0.00
_2020		Demo FB...	Normal	8.00	100 %	8.00	0.00
[redacted]		Craig_test...	Normal	23.79	100 %	23.79	0.00
[redacted]		Craig_test...	Normal	23.79	100 %	23.79	0.00
[redacted]		Craig_test...	Normal	23.79	100 %	23.79	0.00
[redacted]		Craig_test...	Normal	23.79	100 %	23.79	0.00
[redacted]		Craig_test...	Normal	23.79	100 %	23.79	0.00
[redacted]		Craig_test...	Normal	23.79	100 %	23.79	0.00
[redacted]		Craig_test...	Normal	23.79	100 %	23.79	0.00
[redacted]		Craig_test...	Normal	23.79	100 %	23.79	0.00
[redacted]		Craig_test...	Normal	23.79	100 %	23.79	0.00
[redacted]		Craig_test...	Normal	23.79	100 %	23.79	0.00
[redacted]		Craig_test...	Normal	23.79	100 %	23.79	0.00
[redacted]		Craig_test...	Normal	23.79	100 %	23.79	0.00
[redacted]		Craig_test...	Normal	23.79	100 %	23.79	0.00
[redacted]		Craig_test...	Normal	23.79	100 %	23.79	0.00
[redacted]		Craig_test...	Normal	23.79	100 %	23.79	0.00
[redacted]		Craig_test...	Normal	23.79	100 %	23.79	0.00
[redacted]		Craig_test...	Normal	23.79	100 %	23.79	0.00

Showing 1 - 100 of 2111 items | Selected 2 items

Refreshed a few moments ago Page 1 of 22

Figure 7-10 View volume performance in IBM Storage Insights Pro

You also can see which volumes are assigned to which node in the volumes window by selecting Node from the available columns in the table, as shown in Figure 7-11 on page 149. There are many other points that can be selected to view in the table.



- □ Block Size
- □ RAID Level
- ☑ Node
- ☑ Hosts
- □ Encryption
- □ Format
- □ LSS or LCU
- □ Volume Number
- □ FlashCopy
- □ Copy Service Role
- □ Safeguarded
- □ Safeguarded Location
- □ Safeguarded Capacity (GiB)

Figure 7-11 Column selections

Ports

The fiber channel port information reflects the performance metrics for the front-end DS8000 ports that connect the DS8000 storage system to the SAN switches or hosts. Additionally, port error rate metrics, such as Error Frame Rate, are also available. The DS8000 HA card has four or eight ports. The WebUI does not reflect this aggregation, but if necessary, custom reports can be created with native SQL statements to show port performance data that is grouped by the HA to which they belong. Monitoring and analyzing the ports that belong to the same card are beneficial because the aggregated throughput is less than the sum of the stated bandwidth of the individual ports.

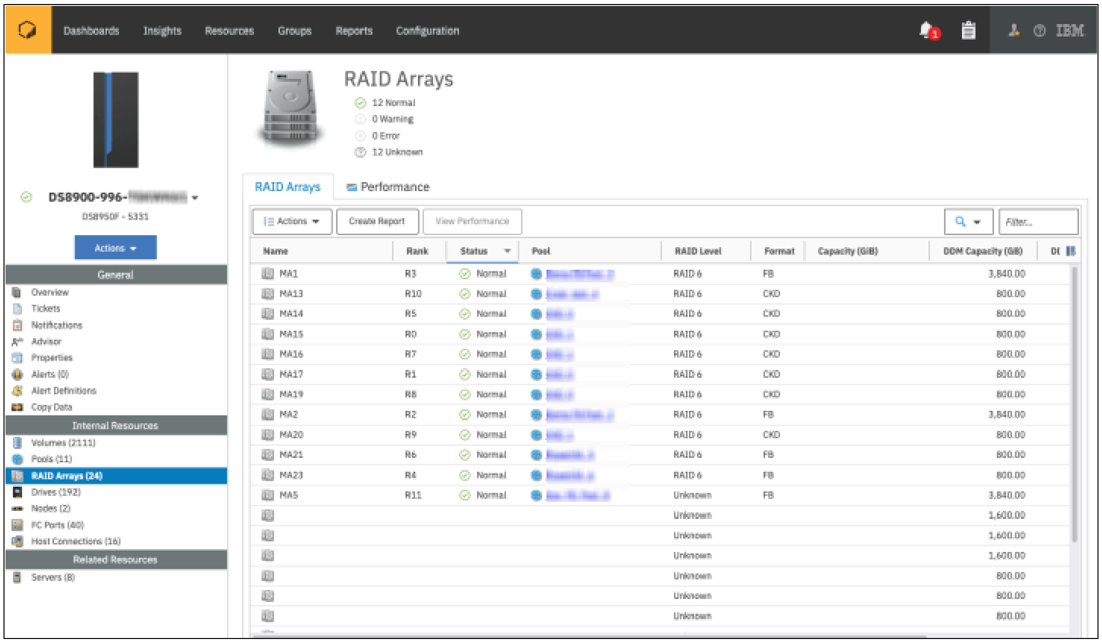
For more information about the DS8000 port cards, see Chapter 2.3, “Host adapters” on page 21.

.

Port metrics: IBM Spectrum Control and Storage Insights Pro report on many port metrics, so the ports on the DS8000 storage system are the front-end part of the storage device.

Array

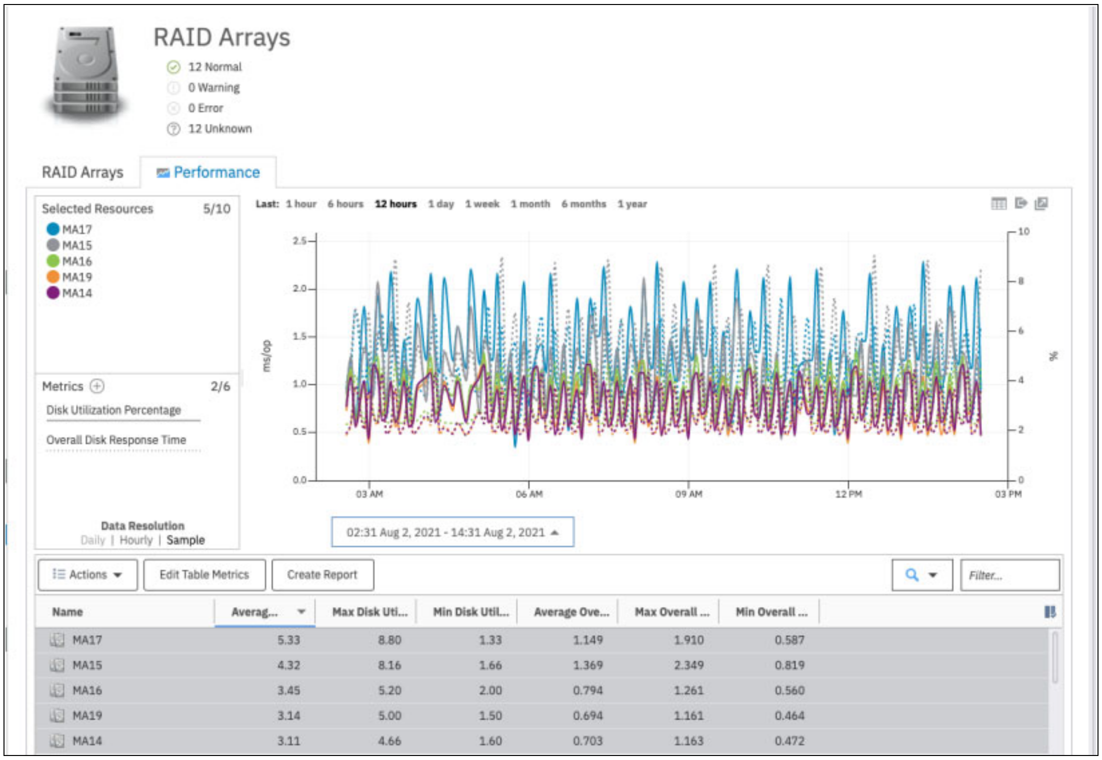
The array name that is shown in the WebUI, as shown in Figure 7-12, directly refers to the array on the DS8000 storage system as listed in the DS GUI or DSCLI.



Name	Rank	Status	Pool	RAID Level	Format	Capacity (GiB)	DDM Capacity (GiB)
MA1	R3	Normal	[redacted]	RAID 6	FB		3,840.00
MA13	R10	Normal	[redacted]	RAID 6	CKD		800.00
MA14	R5	Normal	[redacted]	RAID 6	CKD		800.00
MA15	R0	Normal	[redacted]	RAID 6	CKD		800.00
MA16	R7	Normal	[redacted]	RAID 6	CKD		800.00
MA17	R1	Normal	[redacted]	RAID 6	CKD		800.00
MA19	R8	Normal	[redacted]	RAID 6	CKD		800.00
MA2	R2	Normal	[redacted]	RAID 6	FB		3,840.00
MA20	R9	Normal	[redacted]	RAID 6	CKD		800.00
MA21	R6	Normal	[redacted]	RAID 6	FB		800.00
MA23	R4	Normal	[redacted]	RAID 6	FB		800.00
MA5	R11	Normal	[redacted]	Unknown	FB		3,840.00
				Unknown			1,600.00
				Unknown			1,600.00
				Unknown			1,600.00
				Unknown			800.00
				Unknown			800.00
				Unknown			800.00

Figure 7-12   Viewing RAID Arrays in IBM Storage Insights Pro

When you click the **Performance** tab, the top five performing arrays are displayed with their corresponding graphs, as shown in Figure 7-13 on page 150.



Name	Averag...	Max Disk Uti...	Min Disk Util...	Average Ove...	Max Overall ...	Min Overall ...
MA17	5.33	8.80	1.33	1.149	1.910	0.587
MA15	4.32	8.16	1.66	1.369	2.349	0.819
MA16	3.45	5.20	2.00	0.794	1.261	0.560
MA19	3.14	5.00	1.50	0.694	1.161	0.464
MA14	3.11	4.66	1.60	0.703	1.163	0.472

Figure 7-13   Top 5 RAID arrays' performance

Volumes on the DS8000 storage systems are primarily associated with a pool, and a pool relates to a set of ranks. To associate quickly all arrays with their related ranks and pools, use the output of the DSCLI **lsarray -l** and **lsrank -l** commands, as shown in Example 7-1.

Example 7-1 DS8000 array site, array, and rank association

```
dscli> showrank r23
ID             R23
SN             -
Group          1
State          Normal
datastate      Normal
Array          A15
RAIDtype       6
extpoolID      P5
extpoolnam     rd6_fb_1
volumes        1100,1101,1102,1103,1104,1105
stgtype        fb
exts           12922
usedexts       174
widearrays     0
nararrays      1
trksize        128
strpsize       384
strpesize      0
extsize        16384
encryptgrp     -
migrating(in)  0
migrating(out) 0

dscli> lsrank -l R23
ID  Group State  datastate Array RAIDtype extpoolID extpoolnam stgtype exts  usedexts
=====================================================================================
R23     1 Normal Normal    A15          6 P5        rd6_fb_1   fb      12922      174

dscli> lsarray -l A15
Array State    Data   RAIDtype    arsite Rank DA Pair DDMcap (10^9B) diskclass encrypt
==========================================================================================
A15   Assigned Normal 6 (5+P+Q+S) S16    R23  2               3000.0 NL        unsupported

dscli> lsarraysite S16
arsite DA Pair dkcap (10^9B) State    Array
===========================================
S16    2              3000.0 Assigned A15
```

A DS8000 array is defined on an array site with a specific RAID type. A rank is a logical construct to which an array is assigned. A rank provides a number of extents that are used to create one or several volumes. A volume can use the DS8000 extents from one or several ranks but all within the same pool. For more information, see Chapter 3.1, “Logical Configuration” on page 32, and Chapter 3.2.4, “Easy Tier considerations” on page 55.

Associations: On a configured DS8000 storage system, there is a 1:1 relationship between an array site, an array, and a rank. However, the numbering sequence can differ for arrays, ranks, and array sites, for example, array site S16 = array A15 = rank R23.

In most common logical configurations, users typically sequence them in order, for example, Array Site = S1 = Array A0 = Rank R0. If they are not in order, you must understand on which array the analysis is performed.

Example 7-1 on page 151 shows the relationships among a DS8000 rank, an array, and an array site with a typical divergent numbering scheme by using DSCLI commands. Use the `showrank` command to show which volumes have extents on the specified rank.

In the Array performance chart, you can include both front-end and back-end metrics. The back-end metrics can be selected on the Disk Metrics Tab. They provide metrics from the perspective of the controller to the back-end array sites. The front-end metrics relate to the activity between the server and the controller.

There is a relationship between array operations, cache hit ratio, and percentage of read requests:

- When the cache hit ratio is low, the DS8000 storage system has frequent transfers from drives to cache (staging).
- When the percentage of read requests is high and the cache hit ratio is also high, most of the I/O requests can be satisfied without accessing the drives because of the cache management prefetching algorithm.
- When there is heavy write activity, it leads to frequent transfers from cache to drives (destaging).

Comparing the performance of different arrays shows whether the global workload is equally spread on the drives of your DS8000 storage system. Spreading data across multiple arrays increases the number of drives that is used and optimizes the overall performance.

Volumes

The volumes, called LUNs, are shown in Figure 7-14 on page 153. The host server sees the volumes as physical drives and treats them as physical drives.

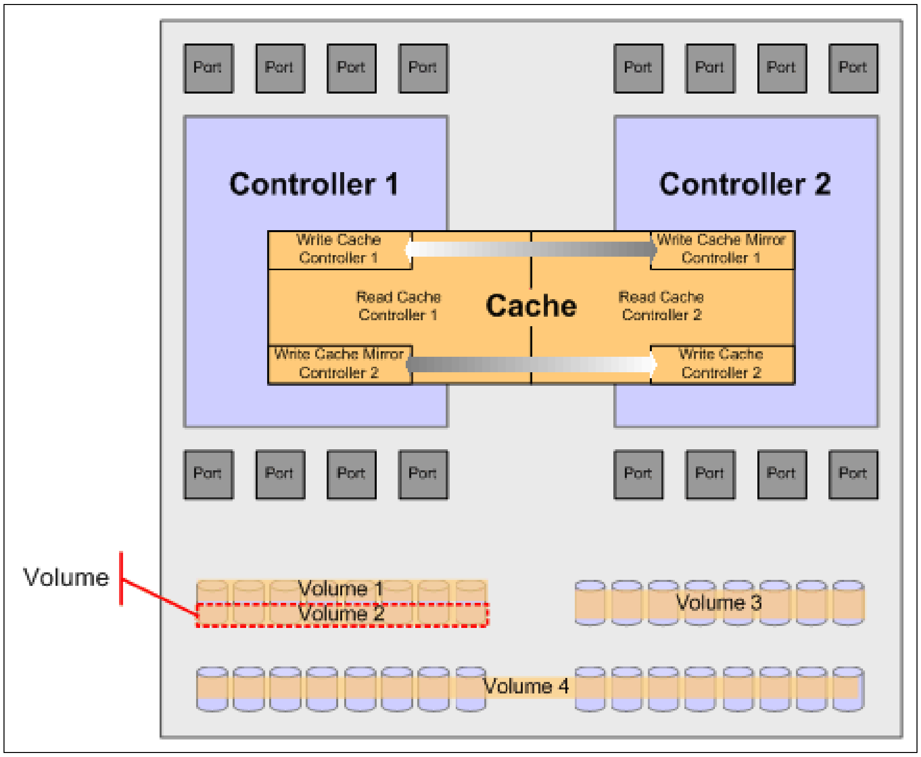
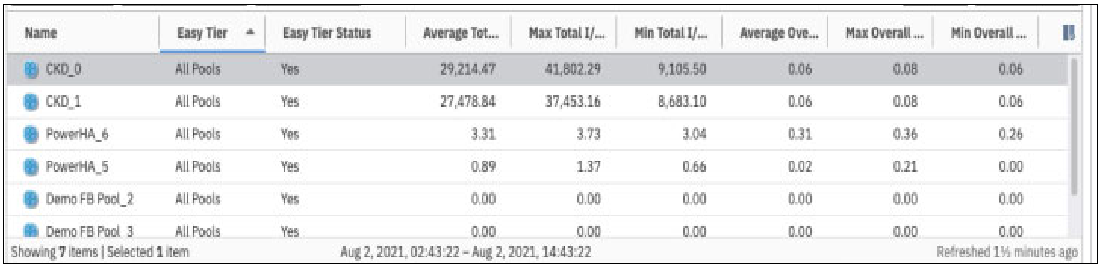


Figure 7-14 DS8000 volume

Analysis of volume data facilitates the understanding of the I/O workload distribution among volumes, and workload characteristics (random or sequential and cache hit ratios). A DS8000 volume can belong to one or several ranks, as shown in Figure 7-14 on page 153 (for more information, see Chapter 3.2, “Data placement on ranks and extent pools” on page 50). Especially in managed multi-rank extent pools with Easy Tier automatic data relocation enabled, the distribution of a certain volume across the ranks in the extent pool can change over time.

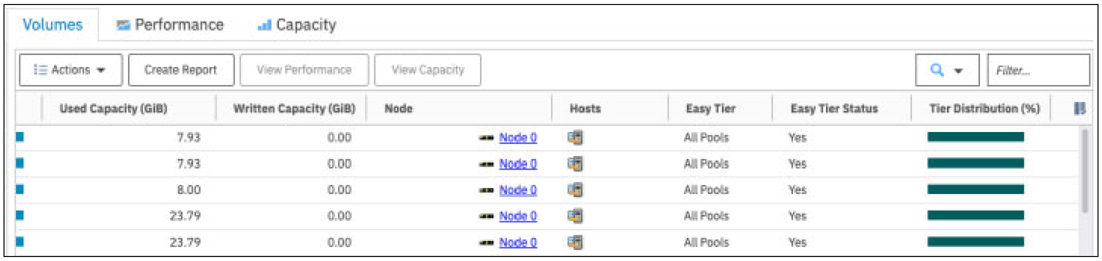
With IBM Spectrum Control and Storage Insights Pro, you can see the Easy Tier, Easy Tier Status, and the capacity values for pools, shown in Figure 7-15 on page 153, and volumes, shown in Figure 7-16.



Name	Easy Tier ▲	Easy Tier Status	Average Tot...	Max Total I/...	Min Total I/...	Average Ove...	Max Overall ...	Min Overall ...
CKD_0	All Pools	Yes	29,214.47	41,802.29	9,105.50	0.06	0.08	0.06
CKD_1	All Pools	Yes	27,478.84	37,453.16	8,683.10	0.06	0.08	0.06
PowerHA_6	All Pools	Yes	3.31	3.73	3.04	0.31	0.36	0.26
PowerHA_5	All Pools	Yes	0.89	1.37	0.66	0.02	0.21	0.00
Demo FB Pool_2	All Pools	Yes	0.00	0.00	0.00	0.00	0.00	0.00
Demo FB Pool_3	All Pools	Yes	0.00	0.00	0.00	0.00	0.00	0.00

Showing 7 items | Selected 1 item   Aug 2, 2021, 02:43:22 – Aug 2, 2021, 14:43:22   Refreshed 1½ minutes ago

Figure 7-15 Easy Tier pool information in IBM Storage Insights Pro



Used Capacity (GiB)	Written Capacity (GiB)	Node	Hosts	Easy Tier	Easy Tier Status	Tier Distribution (%)
7.93	0.00	Node 0		All Pools	Yes	
7.93	0.00	Node 0		All Pools	Yes	
8.00	0.00	Node 0		All Pools	Yes	
23.79	0.00	Node 0		All Pools	Yes	
23.79	0.00	Node 0		All Pools	Yes	

Figure 7-16 Volume Easy Tier information in IBM Storage Insights Pro

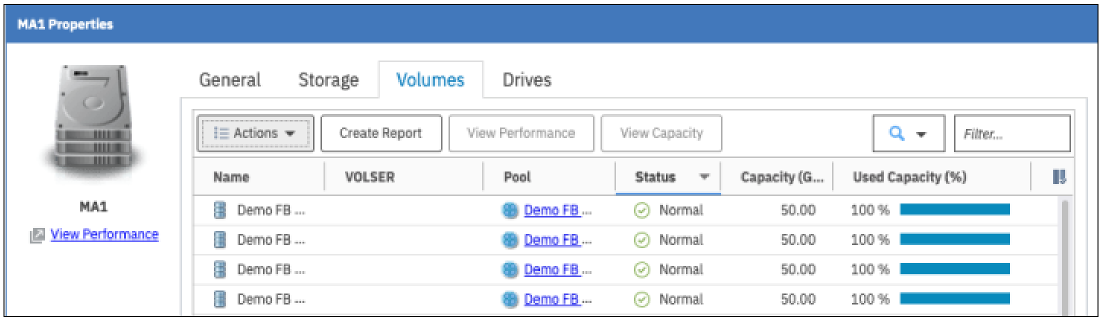
The analysis of volume metrics shows the activity of the volumes on your DS8000 storage system and can help you perform these tasks:

- Determine where the most accessed data is and what performance you get from the volume.
- Understand the type of workload that your application generates (sequential or random and the read or write operation ratio).
- Determine the cache benefits for the read operation (cache management prefetching algorithm SARC).
- Determine cache bottlenecks for write operations.
- Compare the I/O response observed on the DS8000 storage system with the I/O response time observed on the host.

The relationship of certain RAID arrays and ranks to the DS8000 pools can be derived from the RAID array list window, which is shown in Figure 7-12.

From there, you can easily see the volumes that belong to a certain pool by right-clicking a pool, selecting **View Properties**, and clicking the **Volumes** tab.

Figure 7-17 on page 154 shows the relationship of the raid array, pools, and volumes for a raid array.



Name	VOLSER	Pool	Status	Capacity (G...	Used Capacity (%)
Demo FB ...		Demo FB ...	Normal	50.00	100 %
Demo FB ...		Demo FB ...	Normal	50.00	100 %
Demo FB ...		Demo FB ...	Normal	50.00	100 %
Demo FB ...		Demo FB ...	Normal	50.00	100 %

Figure 7-17 Volumes view within Pool Properties

In addition, to associate quickly the DS8000 arrays to array sites and ranks, you might use the output of the DSCLI commands `lsrank -l` and `lsarray -l`, as shown in Example 7-1 on page 151.

7.2.3 General IBM Spectrum Control and IBM Storage Insights Pro measurement considerations

To understand the IBM Spectrum Control and IBM Storage Insights Pro measurements of the DS8000 components, it is helpful to understand the context for the measurement. The measurement facilitates insight into the behavior of the DS8000 storage system and its ability to service I/O requests. The DS8000 storage system handles various types of I/O requests differently. Table 7-3 shows the behavior of the DS8000 storage system for various I/O types.

Table 7-3 DS8000 I/O types and behavior

I/O type	DS8000 high-level behavior
Sequential read	Pre-stage reads in cache to increase cache hit ratio.
Random read	Attempt to find data in cache. If not present in cache, read from back end.
Sequential write	Write data to the NVS of the processor complex owning volume and send a copy of the data to cache in the other processor complex. Upon back-end destaging, perform prefetching of read data and parity into cache to reduce the number of disk operations on the back end.
Random write	Write data to NVS of the processor complex owning volume and send a copy of the data to cache in the other processor complex. Destage modified data from NVS to disk as determined by Licensed Internal Code.

Understanding writes to a DS8000 storage system

When the DS8000 storage system accepts a write request, it processes it without physically writing to the drives. The data is written into both the processor complex to which the volume belongs and the persistent memory of the second processor complex in the DS8000 storage system. Later, the DS8000 storage system asynchronously destages the modified data out to the drives. In cases where back-end resources are constrained, NVS delays might occur. IBM Spectrum Control and Storage Insights report on these conditions with the following front-end metrics: Write Cache Delay I/O Rate and Write Cache Delay I/O Percentage.

Understanding reads on a DS8000 storage system

If the DS8000 storage system cannot satisfy the read I/O requests within the cache, it transfers data from the drives. The DS8000 storage system suspends the I/O request until it reads the data. This situation is called cache-miss. If an I/O request is cache-miss, the response time includes the data transfer time between host and cache, and also the time that it takes to read the data from drives to cache before sending it to the host. The various read hit ratio metrics show how efficiently cache works on the DS8000 storage system.

The read hit ratio depends on the characteristics of data on your DS8000 storage system and applications that use the data. If you have a database and it has a high locality of reference, it shows a high cache hit ratio because most of the data that is referenced can remain in the cache. If your database has a low locality of reference, but it has the appropriate sets of indexes, it might also have a high cache hit ratio because the entire index can remain in the cache.

A database can be cache-unfriendly by nature. An example of a cache-unfriendly workload is a workload that consists of large sequential reads to a highly fragmented file system. If an application reads this file, the cache hit ratio is low because the application never reads the same data because of the nature of sequential access. In this case, defragmentation of the file system improves the performance. You cannot determine whether increasing the size of cache improves the I/O performance without knowing the characteristics of the data on your DS8000 storage system.

Monitor the read hit ratio over an extended period:

- If the cache hit ratio is historically low, it is most likely because of the nature of the data access patterns. Defragmenting the file system and making indexes if none exist might help more than adding cache.
- If you have a high cache hit ratio initially and it decreases as the workload increases, adding cache or moving part of the data to volumes that are associated with the other processor complex might help.

Interpreting the read-to-write ratio

The read-to-write ratio depends on how the application programs issue I/O requests. In general, the overall average read-to-write ratio is in the range of 75% - 80% reads.

For a logical volume that has sequential files, you must understand the application types that access those sequential files. Normally, these sequential files are used for either read only or write only at the time of their use. The DS8000 cache management prefetching algorithm (SARC) determines whether the data access pattern is sequential. If the access is sequential, contiguous data is prefetched into cache in anticipation of the next read request.

IBM Storage Insights Pro report the reads and writes through various metrics. For a description of these metrics in greater detail, see 7.3, “IBM Storage Insights Pro data collection considerations” on page 156.

7.3 IBM Storage Insights Pro data collection considerations

This section describes the performance data collection considerations, such as time stamps, durations, and intervals.

7.3.1 Time stamps

IBM Storage Insights Pro uses the time stamp of the source devices when it inserts data into the database. If the server clock is not synchronized with the rest of your environment, it does not include any additional offset because you might need to compare the performance data of the DS8000 storage system with the data gathered on a server or other connected SAN devices, such as FlashSystem or switch metrics.

The data is collected at the indicated resource time stamp in the server time zone. It receives the time zone information from the devices (or the NAPIs) and uses this information to adjust the time in the reports to the local time. Certain devices might convert the time into Coordinate Universal Time time stamps and not provide any time zone information.

This complexity is necessary to compare the information from two subsystems in different time zones from a single administration point. This administration point is the GUI. If you open the GUI in different time zones, a performance diagram might show a distinct peak at different times, depending on its local time zone.

When using IBM Storage Insights Pro to compare data from a server (for example, iostat data) with the data of the storage subsystem, it is important to know the time stamp of the storage subsystem. The time zone of the device is shown in the DS8000 Properties window.

To ensure that the time stamps on the DS8000 storage system are synchronized with the other infrastructure components, the DS8000 storage system provides features for configuring a Network Time Protocol (NTP) server. To modify the time and configure the HMC to use an NTP server, see the following publications:

- *IBM DS8870 Architecture and Implementation (Release 7.5)*, SG24-8085
- *IBM DS8880 Architecture and Implementation (Release 8)*, SG24-8323
- *IBM DS8900F Architecture and Implementation (Release 9.2)*, SG24-8456

As IBM Storage Insights Pro can synchronize multiple performance charts that are opened in the WebUI to display the metrics at the same time, use an NTP server for all components in the SAN environment.

7.3.2 Duration

IBM Storage Insights Pro collects data continuously. From a performance management perspective, collecting data continuously means that performance data exists to facilitate reactive, proactive, and even predictive processes, as described in Chapter 7, "Practical performance management" on page 139.

7.3.3 Intervals

In IBM Storage Insights Pro, the data collection interval is referred to as the *sample interval*. The sample interval for the DS8000 performance data collection tasks is 5- 60 minutes. A shorter sample interval results in a more granular view of performance data at the expense of requiring additional database space. The appropriate sample interval depends on the objective of the data collection. Table 7-4 on page 157 displays example data collection objectives and reasonable values for a sample interval.

Table 7-4 Sample interval examples

Objective	Sample interval
Probe of the entire system	once a day
Performance	5 minutes

Changes in the logical configuration of a system will result in a non-scheduled probe in both IBM Storage Insights Pro and IBM Spectrum Control.

In support of ongoing performance management, a reasonable sample interval is 5 minutes. An interval of 5 minutes provides enough granularity to facilitate reactive performance management. In certain cases, the level of granularity that is required to identify the performance issue is less than 5 minutes. In these cases, you can reduce the sample interval to a 1-minute interval in IBM Spectrum Control only. IBM Spectrum Control also provides reporting at higher intervals, including hourly and daily. It provides these views automatically.

Attention: Although the 1-minute interval collection is the default for some devices, the 5-minute interval is considered a good average for reviewing data and is the only option in IBM Storage Insights. However, some issues can occur within this time interval resulting in peaks being averaged out, which might not be as apparent as they are with 1-minute interval collection. In such situations, the 1-minute interval is most appropriate for offering a more granular view that results in a more effective analysis. However, a 1-minute interval results in copious amount of data being collected, resulting in a fast growing database. Therefore, it should be used only for troubleshooting purposes and is only available in IBM Spectrum Control.

7.4 IBM Storage Insights Pro performance metrics

IBM Storage Insights Pro has many metrics available for reporting the health and performance of the DS8000 storage system.

For a list of available performance metrics for the DS8000 storage system, see the IBM Storage Insights Pro in IBM Documentation Center:

https://www.ibm.com/docs/en/storage-insights

7.4.1 DS8000 key performance indicator thresholds

This section provides some additional information about a subset of critical metrics. It provides suggested threshold values as general recommendations for alerting for various DS8000 components, as described in 7.4.2, “Alerts and thresholds” on page 159. As with any recommendation, you must adjust them for the performance requirements of your environment.

Note: IBM Storage Insights Pro and IBM Spectrum Control also have other metrics that can be adjusted and configured for your specific environment to suit customer demands.

Colors are used to distinguish the components that are shown in Table 7-5.

Table 7-5 DS8000 key performance indicator thresholds

Component	Tab	Metric	Threshold	Comment
Node	Volume Metrics	Cache Holding Time	< 200	Indicates high cache track turnover and possibly cache constraint.
Node	Volume Metrics	Write Cache Delay Percentage	> 1%	Indicates writes delayed because of insufficient memory resources.
Array	Drive Metrics	Utilization Percentage	> 70%	Indicates drive saturation. For IBM Storage Insights Pro, the default value on this threshold is 50%.
Array	Drive Metrics	Overall Response Time	> 35	Indicates busy drives.
Array	Drive Metrics	Write Response Time	> 35	Indicates busy drives.
Array	Drive Metrics	Read Response Time	> 35	Indicates busy drives.
Port	Port Metrics	Total Port I/O Rate	Depends	Indicates transaction intensive load. The configuration depends on the HBA, switch, and other components.

Component	Tab	Metric	Threshold	Comment
Port	Port Metrics	Total Port Data Rate	Depends	If the port data rate is close to the bandwidth, this rate indicates saturation. The configuration depends on the HBA, switch, and other components.
Port	Port Metrics	Port Send Response Time	> 2	Indicates contention on I/O path from the DS8000 storage system to the host.
Port	Port Metrics	Port Receive Response Time	> 2	Indicates a potential issue on the I/O path or the DS8000 storage system back end.
Port	Port Metrics	Total Port Response Time	> 2	Indicates a potential issue on the I/O path or the DS8000 storage system back end.

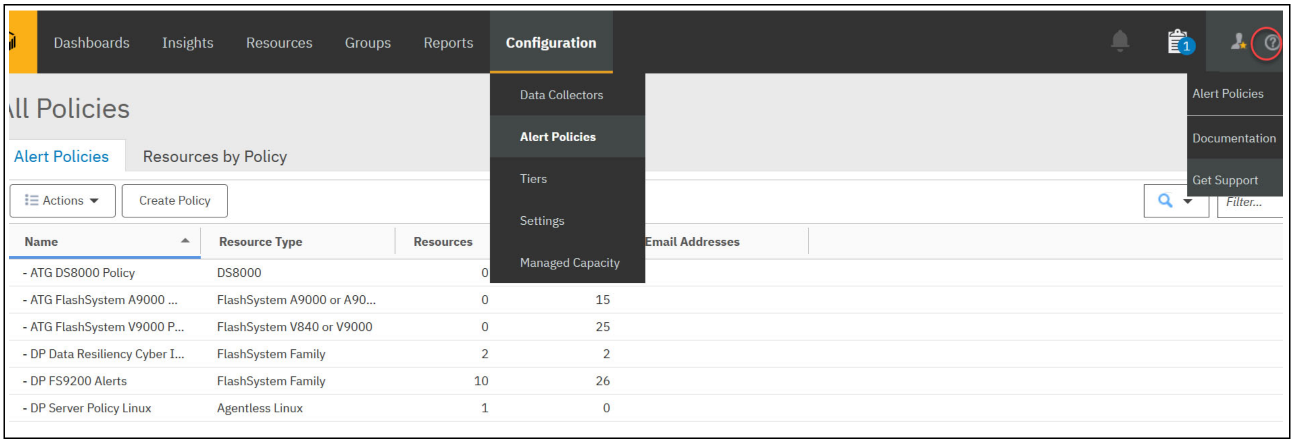
7.4.2 Alerts and thresholds

IBM Storage Insights Pro uses the term data collector for the process that is set up to gather data from a subsystem. This is shown under the Configuration > Data Collection window. The data collector collects information at 5 minute intervals and stores the data in an IBM cloud database. The performance data is retained for 1 year and is available for analysis by using several methods that are described in 7.5, “IBM Storage Insights Pro reporting options” on page 162.

You can use IBM Storage Insights Pro to define performance-related alerts that can trigger an event when the defined thresholds are reached. Even though it works in a similar manner to a monitor without user intervention, the actions are still performed at specified intervals of the data collection job.

With Storage Insights Pro you can create alert policies to manage the alert conditions for multiple resources. You can create a new policy or copy an existing policy to modify the alerts definitions and to add resources to monitor. To create a new policy for DS8000 complete the following steps:

1. From the dashboard select **Configuration** > **Alert Policies** as shown in Figure 7-18 on page 159. You can copy and modify an existing policy or create a new policy.

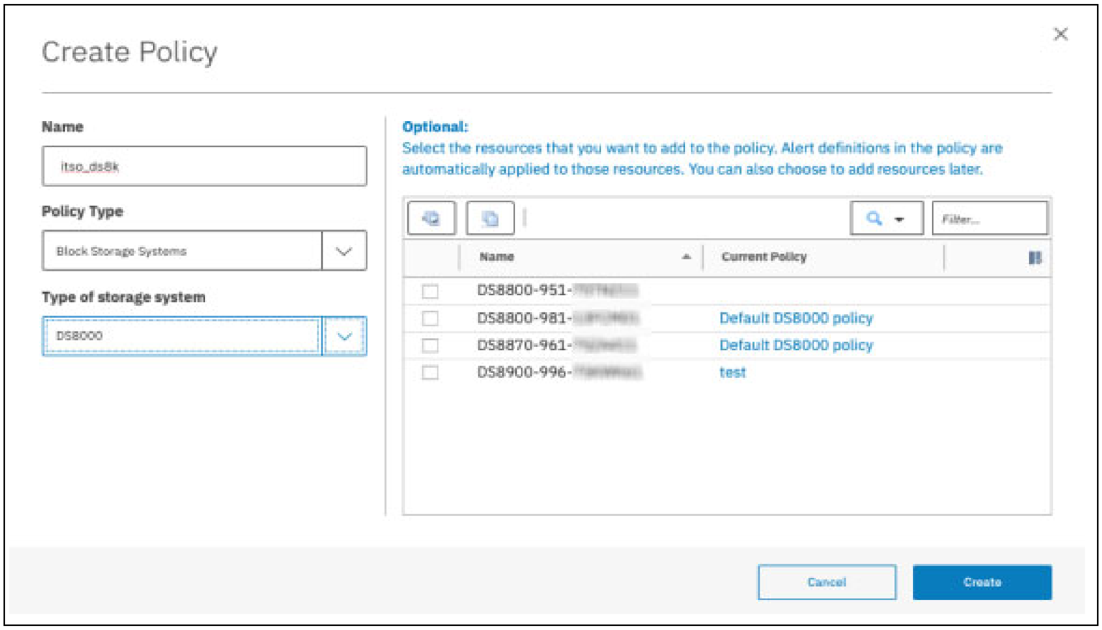


Name	Resource Type	Resources		Email Addresses
- ATG DS8000 Policy	DS8000	0		
- ATG FlashSystem A9000 ...	FlashSystem A9000 or A90...	0	15	
- ATG FlashSystem V9000 P...	FlashSystem V840 or V9000	0	25	
- DP Data Resiliency Cyber I...	FlashSystem Family	2	2	
- DP FS9200 Alerts	FlashSystem Family	10	26	
- DP Server Policy Linux	Agentless Linux	1	0	

Figure 7-18 Configuring Alert policies

2. To create a new policy, select **Create Policy.** Enter a Name and select the type of resource from the list, for example block storage. Select the type of storage system you want to monitor for example DS8000. A list of DS8000 names managed by Storage

Insights Pro will be displayed. You can choose any or all DS8000 to add to this policy click create as shown in Figure 7-19. Notice that if a system already has a policy associated with it, that will be displayed under “Current Policy”.



Create Policy

Name: itso_ds8k

Policy Type: Block Storage Systems

Type of storage system: DS8000

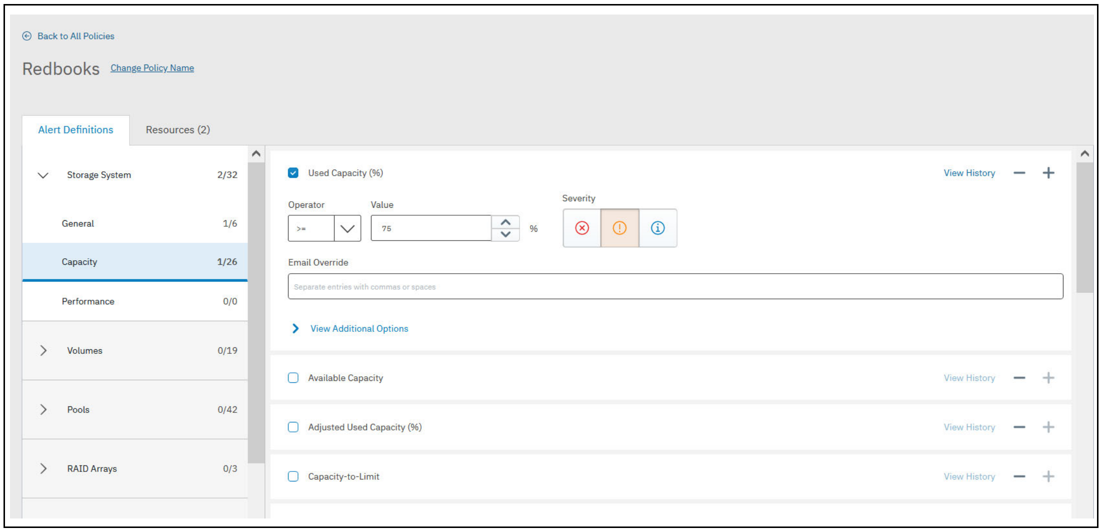
Optional: Select the resources that you want to add to the policy. Alert definitions in the policy are automatically applied to those resources. You can also choose to add resources later.

	Name	Current Policy
□	DS8800-951-[redacted]	
□	DS8800-981-[redacted]	Default DS8000 policy
□	DS8870-961-[redacted]	Default DS8000 policy
□	DS8900-996-[redacted]	test

Figure 7-19 Create a new Policy

Next define alerts for the category of the attributes that you want to alert on:

3. **General**: Attributes for the key properties of a resource, such as status, data collection status and firmware.
4. **Capacity**: Attributes for capacity statistics of a resource, such as available capacity, used capacity, drive capacity, Safeguarded capacity, and more. Total of 26 alerts can be set as shown in Figure 7-20 on page 160.



Back to All Policies

Redbooks Change Policy Name

Alert Definitions | Resources (2)

Storage System	2/32
General	1/6
Capacity	1/26
Performance	0/0
Volumes	0/19
Pools	0/42
RAID Arrays	0/3

☑ Used Capacity (%)

Operator: >= Value: 75 % Severity

Email Override: Separate entries with commas or spaces

View Additional Options

□ Available Capacity

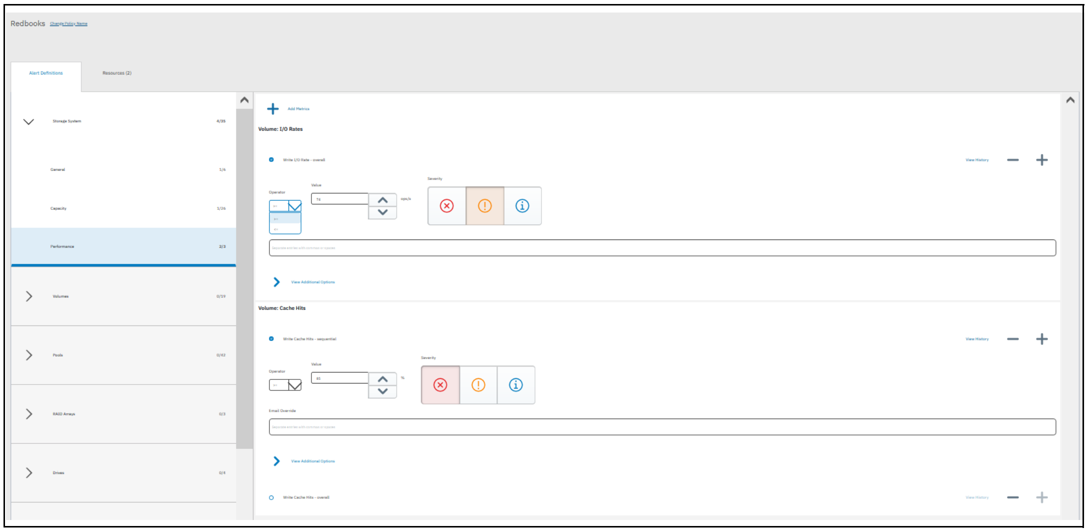
□ Adjusted Used Capacity (%)

□ Capacity-to-Limit

Figure 7-20 Defining Alerts for Capacity categories

5. **Performance**: You can define alerts that are triggered when the performance of a resource falls outside a specified threshold. You:

   a. Click **Add Metric**.
   b. Select the check box for the metric for which you want to set the threshold for Volume, Drive or Port.
   c. Click **save**.
6. Once the performance metrics are selected you can specify the conditions for generating an alert as shown in Figure 7-21 on page 161:
   a. Select >= or <=.
   b. Enter a threshold value.
   c. Select a severity for the alert.
   d. Click save.



Volume: I/O Rates

Write I/O Rate - overall

Operator >= Value 74 ops/s

Volume: Cache Hits

Write Cache Hits - sequential

Operator >= Value 85 %

Write Cache Hits - overall

Figure 7-21 Setting Performance Alerts

> **Reference:** For more information about setting Thresholds and Alert suppressions in IBM Storage Insights Pro, see IBM Documentation Storage Insights at:
>
> https://www.ibm.com/docs/en/storage-insights?topic=product-overview.

You should configure the thresholds that are most important and most relevant to the environmental needs to assist with good planning.

IBM Storage Insights Pro provides recommended values for threshold values that do not vary much between environments. However, for metrics that measure throughput and response times, thresholds can vary because of workload, model of hardware, amount of cache memory, and other factors. In these cases, there are no recommended values. To help determine threshold values for a resource, collect performance data over time to establish a baseline of the normal and expected performance behavior for that resource.

After you determine a set of baseline values, define alerts to trigger if the measured performance behavior falls outside of the normally expected range.

The alerts for a DS8000 storage system can be seen, filtered, removed, acknowledged, or exported in the storage system Alert window, as shown in Figure 7-22.

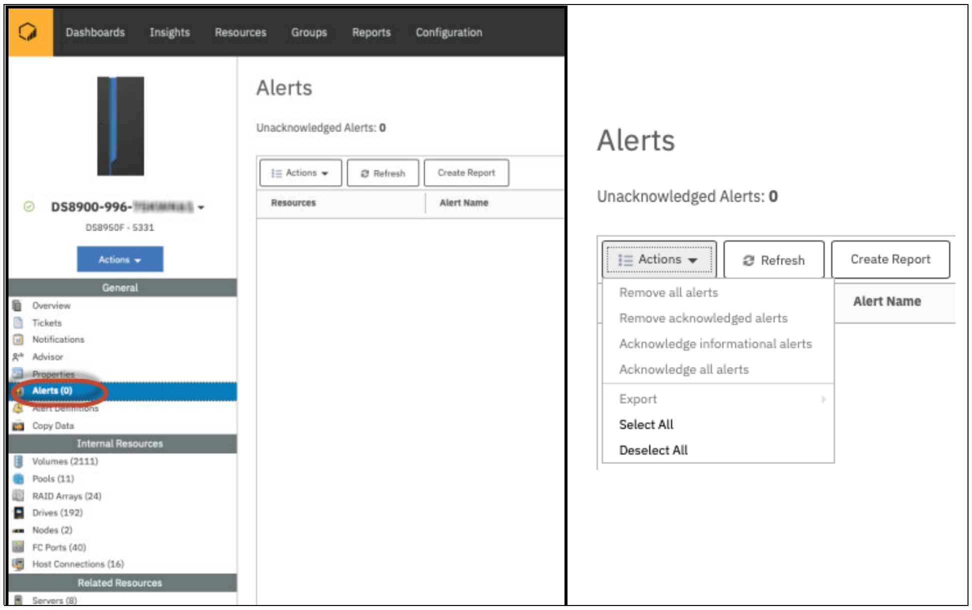


Figure 7-22 Viewing alerts for a system

Limitations to alert definitions

There are a few limitations to alert levels:

- Thresholds are always active. They cannot be set to exclude specific periods. This setting can be mitigated by using alert suppression settings.
- Detailed knowledge of the workload is required to use thresholds effectively.

> **False positive alerts:** Configuring thresholds too conservatively can lead to an excessive number of false positive alerts.

7.5 IBM Storage Insights Pro reporting options

IBM Storage Insights Pro can easily create inventory reports, capacity and performance reports, and reports about storage consumption. This section provides an overview of the various options and their appropriate usage in ongoing performance management of a DS8000 storage system and all the monitored resources.

In Storage Insights Pro you can create a number of reports that you can schedule and send by email:

- Predefined capacity reports that you can configure and refine the information that is included in the report. Create a predefined report about storage systems, pools, or tiered pools.
- Custom reports that you can create to include asset, capacity, configuration, or health status or performance information about your storage resources. You can specify a relative time range for the capacity information in the report, such as the last hour, 6 hours, 12 hours, day, week, month, 6 months, or year. Depending on the time range that you specify, the aggregated values for the performance information are shown in the report.
- Consumer reports that you can create to help plan capacity purchases and make your organization aware of the cost and the amount of the storage that is used by storage consumers, create chargeback and consumer reports.

For the available tutorials related to creating reports see the IBM Storage Insights Pro in IBM Documentation Center at:

https://www.ibm.com/docs/en/storage-insights

Note: You must have an Administrator role to create, edit, or delete custom, predefined capacity and inventory reports, and chargeback and storage consumption reports. Users with a Monitor role can run and view the reports that are shown on the Reports page, but they can't edit or delete reports.

Additional reporting

In IBM Storage Insights Pro, export options are available. To export the data that is used in the performance chart, use the export function, which is found in the upper right of the chart as shown in Figure 7-23 on page 163, or under the Actions menu in other views.

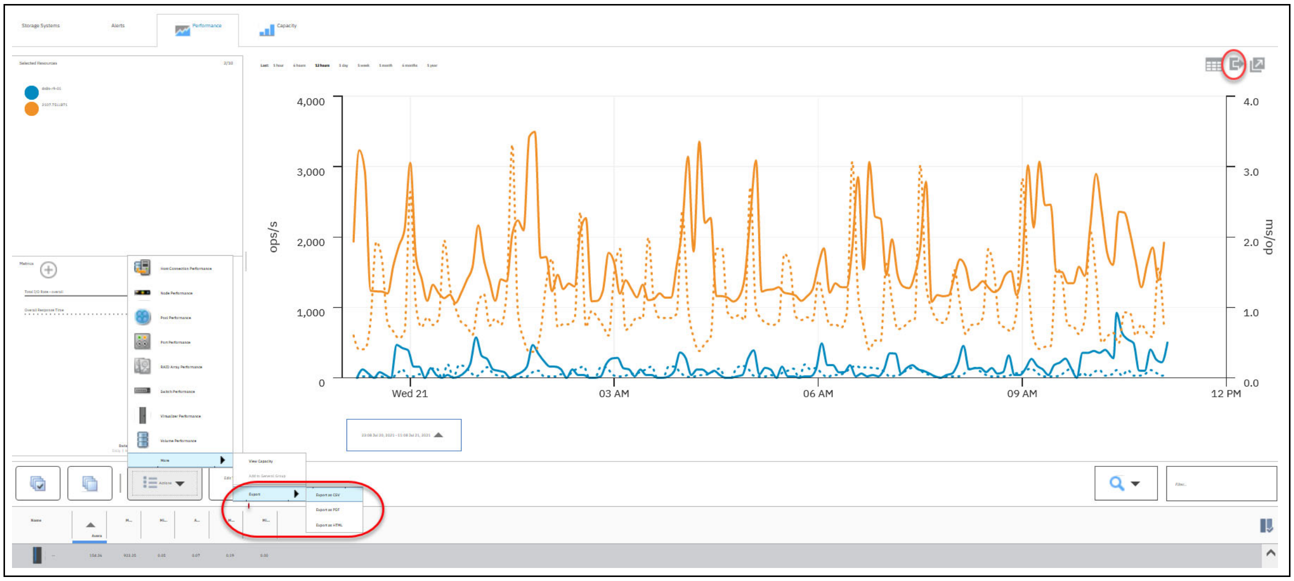


Figure 7-23 Performance export functions

To export the summary table underneath the chart, click **Action** → **More** → **Export**, and select the desired format.

Charts are automatically generated for most of the predefined reports. Depending on the type of resource, the charts show statistics for space usage, workload activity, bandwidth percentage, and other statistics. You can schedule reports and specify the report output as HTML, PDF, and CSV formats. You can also configure reports to save the report output to your local file system, and to send reports as mail attachments.

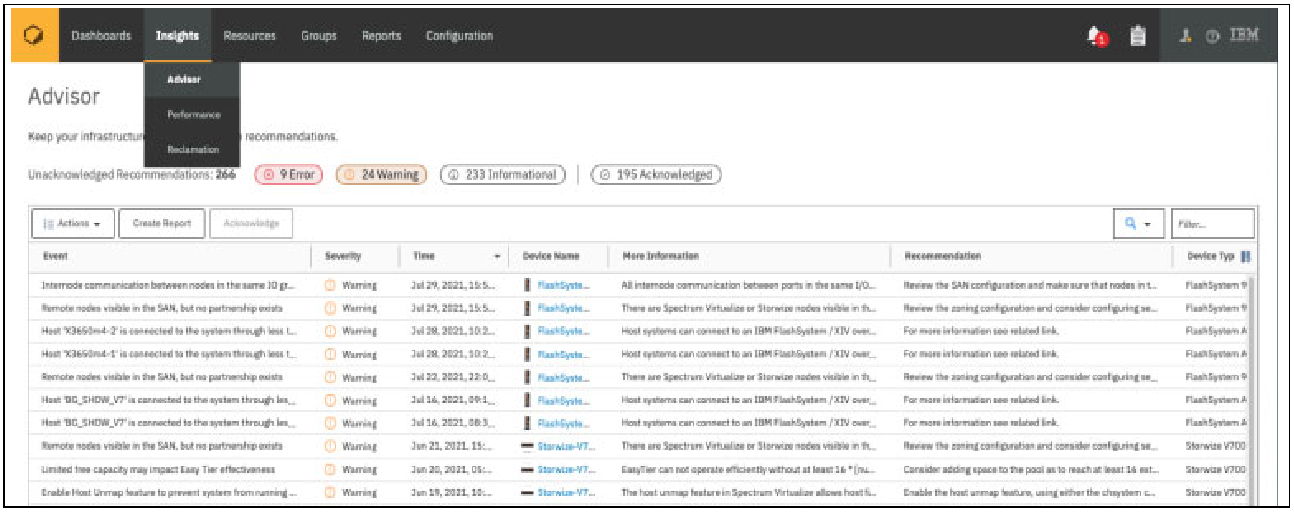
7.6 Insights

IBM Storage Insights Pro provides additional functions to gain insights into the performance of the resources that are monitored in your storage environment. You can view the recommendations that help you address issues in your storage environment and gain insights into storage reclamation and performance.

These functions are not included in the entitled version of IBM Storage Insights.

7.6.1 Advisor

To display the Advisor page Click **Insights** > **Advisor**, shown in Figure 7-24 on page 164. You can view, evaluate, and manage events that include a recommended action. For each event that includes recommended actions, you can view details such as the severity, the time it occurred, the device, and the internal resource on which it occurred. You can acknowledge recommended actions, and create reports about them. You can also use the filter tags on the upper part of the dashboard to quickly filter your recommended actions. You can also customize the view by selecting the columns you want to see, just below the filter box.



Advisor

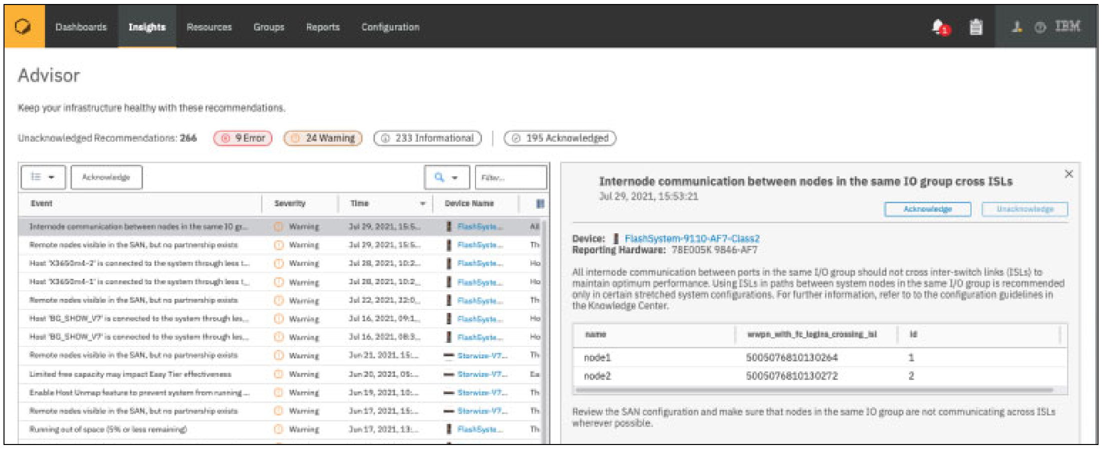
Keep your infrastructur[e] ... recommendations.

Unacknowledged Recommendations: **266** 9 Error 24 Warning 233 Informational 195 Acknowledged

Event	Severity	Time	Device Name	More Information	Recommendation	Device Typ
Internode communication between nodes in the same IO gr...	Warning	Jul 29, 2021, 15:5...	FlashSyste...	All internode communication between ports in the same I/O...	Review the SAN configuration and make sure that nodes in t...	FlashSystem 9
Remote nodes visible in the SAN, but no partnership exists	Warning	Jul 29, 2021, 15:5...	FlashSyste...	There are Spectrum Virtualize or Storwize nodes visible in th...	Review the zoning configuration and consider configuring se...	FlashSystem 9
Host 'X3650m4-2' is connected to the system through less t...	Warning	Jul 28, 2021, 10:2...	FlashSyste...	Host systems can connect to an IBM FlashSystem / XIV over...	For more information see related link.	FlashSystem A
Host 'X3650m4-1' is connected to the system through less t...	Warning	Jul 28, 2021, 10:2...	FlashSyste...	Host systems can connect to an IBM FlashSystem / XIV over...	For more information see related link.	FlashSystem A
Remote nodes visible in the SAN, but no partnership exists	Warning	Jul 22, 2021, 22:0...	FlashSyste...	There are Spectrum Virtualize or Storwize nodes visible in th...	Review the zoning configuration and consider configuring se...	FlashSystem 9
Host 'BG_SHOW_V7' is connected to the system through les...	Warning	Jul 16, 2021, 09:1...	FlashSyste...	Host systems can connect to an IBM FlashSystem / XIV over...	For more information see related link.	FlashSystem A
Host 'BG_SHOW_V7' is connected to the system through les...	Warning	Jul 16, 2021, 08:3...	FlashSyste...	Host systems can connect to an IBM FlashSystem / XIV over...	For more information see related link.	FlashSystem A
Remote nodes visible in the SAN, but no partnership exists	Warning	Jun 21, 2021, 15:...	Storwize-V7...	There are Spectrum Virtualize or Storwize nodes visible in th...	Review the zoning configuration and consider configuring se...	Storwize V700
Limited free capacity may impact Easy Tier effectiveness	Warning	Jun 20, 2021, 05:...	Storwize-V7...	EasyTier can not operate efficiently without at least 16 * [nu...	Consider adding space to the pool as to reach at least 16 ext...	Storwize V700
Enable Host Unmap feature to prevent system from running ...	Warning	Jun 19, 2021, 10:...	Storwize-V7...	The host unmap feature in Spectrum Virtualize allows host fi...	Enable the host unmap feature, using either the chsystem c...	Storwize V700

Figure 7-24 Advisor page

Click the recommendation detail from the Action list to view the details of a recommended action; an example of this is shown in Figure 7-25 on page 164. You can also **export** the table that displays the recommended actions to a file.



Advisor

Keep your infrastructure healthy with these recommendations.

Unacknowledged Recommendations: **266** 9 Error 24 Warning 233 Informational 195 Acknowledged

Event	Severity	Time	Device Name	
Internode communication between nodes in the same IO gr...	Warning	Jul 29, 2021, 15:5...	FlashSyste...	All
Remote nodes visible in the SAN, but no partnership exists	Warning	Jul 29, 2021, 15:5...	FlashSyste...	Th
Host 'X3650m4-2' is connected to the system through less t...	Warning	Jul 28, 2021, 10:2...	FlashSyste...	Ho
Host 'X3650m4-1' is connected to the system through less t...	Warning	Jul 28, 2021, 10:2...	FlashSyste...	Ho
Remote nodes visible in the SAN, but no partnership exists	Warning	Jul 22, 2021, 22:0...	FlashSyste...	Th
Host 'BG_SHOW_V7' is connected to the system through les...	Warning	Jul 16, 2021, 09:1...	FlashSyste...	Ho
Host 'BG_SHOW_V7' is connected to the system through les...	Warning	Jul 16, 2021, 08:3...	FlashSyste...	Ho
Remote nodes visible in the SAN, but no partnership exists	Warning	Jun 21, 2021, 15:...	Storwize-V7...	Th
Limited free capacity may impact Easy Tier effectiveness	Warning	Jun 20, 2021, 05:...	Storwize-V7...	Ea
Enable Host Unmap feature to prevent system from running ...	Warning	Jun 19, 2021, 10:...	Storwize-V7...	Th
Remote nodes visible in the SAN, but no partnership exists	Warning	Jun 17, 2021, 15:...	Storwize-V7...	Th
Running out of space (5% or less remaining)	Warning	Jun 17, 2021, 13:...	FlashSyste...	Th

Internode communication between nodes in the same IO group cross ISLs
Jul 29, 2021, 15:53:21

Device: FlashSystem-9110-AF7-Class2
Reporting Hardware: 78E005K 9846-AF7

All internode communication between ports in the same I/O group should not cross inter-switch links (ISLs) to maintain optimum performance. Using ISLs in paths between system nodes in the same I/O group is recommended only in certain stretched system configurations. For further information, refer to to the configuration guidelines in the Knowledge Center.

name	wwpn_with_fc_logins_crossing_isl	id
node1	5005076810130264	1
node2	5005076810130272	2

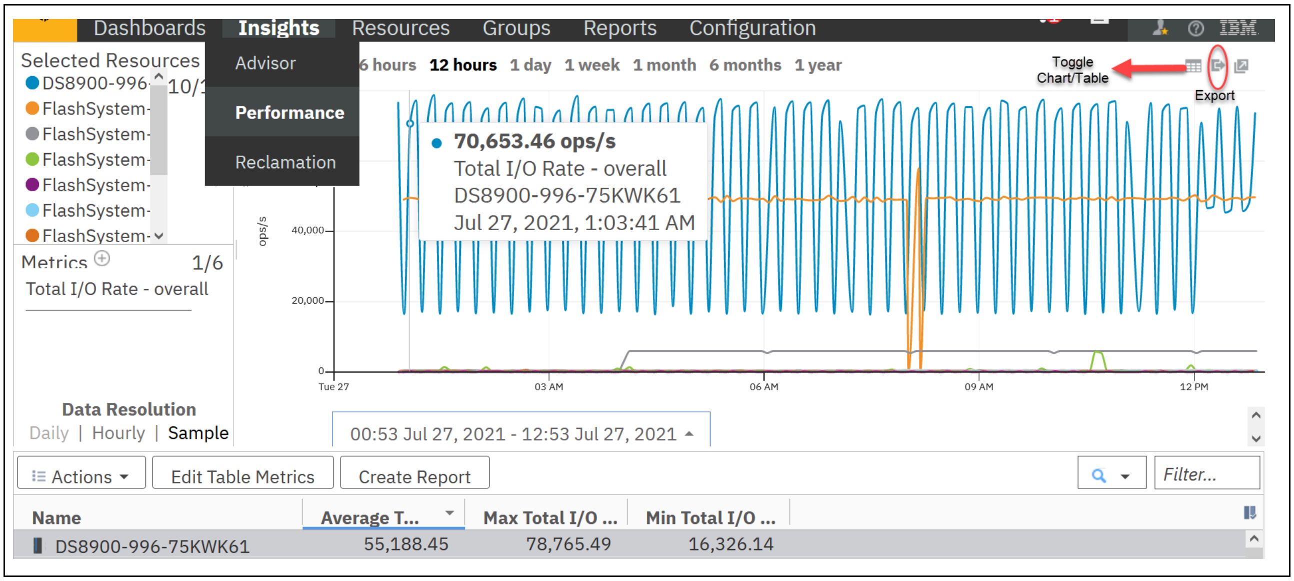
Review the SAN configuration and make sure that nodes in the same IO group are not communicating across ISLs wherever possible.

Figure 7-25 Viewing recommended actions

7.6.2 Performance

You can use performance metrics for volumes, drives, or ports to help you measure, identify, and troubleshoot performance issues and bottlenecks in storage systems.

To display the performance page, click **Insights** > **Performance** as shown in Figure 7-26**.** Up to 10 storage systems with the highest overall total I/O rate over a 12-hour period are displayed in the chart. These storage systems are listed under Resources on the left and are highlighted at the top of the performance chart legend, which is sorted by the Total I/O Rate - overall (ops/s) column. Switch between the chart and table views by selecting the toggle bottom on the to right top section of performance page as shown in Figure 7-26. View the specific details about a system by selecting just that system. Also view over different intervals, from six hours up to a year in IBM Storage Insights Pro (longer periods are available in IBM Spectrum Control). The Performance chart legend at the bottom section of the performance view is a table that shows more information about the storage systems.



70,653.46 ops/s
Total I/O Rate - overall
DS8900-996-75KWK61
Jul 27, 2021, 1:03:41 AM

Metrics 1/6
Total I/O Rate - overall

00:53 Jul 27, 2021 - 12:53 Jul 27, 2021

Name	Average T...	Max Total I/O ...	Min Total I/O ...
DS8900-996-75KWK61	55,188.45	78,765.49	16,326.14

Figure 7-26 Performance page

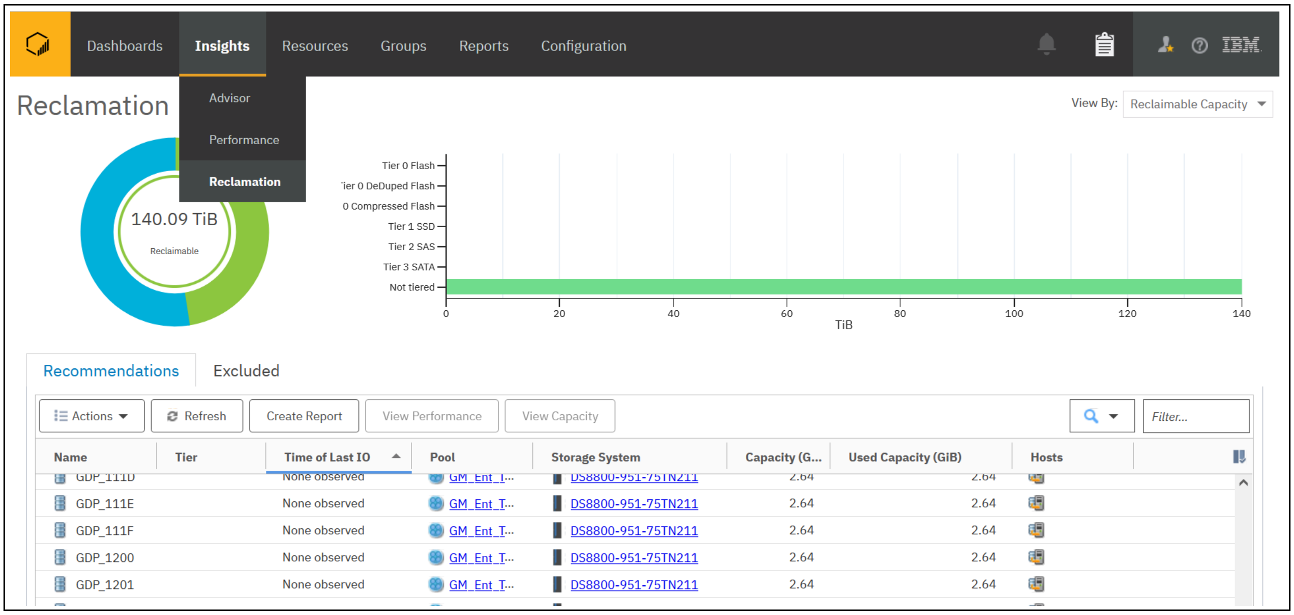
To view performance trends, select metrics for volumes, drives, ports, or node and specify a time range.

For more information about Insights to performance see the IBM Documentation center for Insights Pro at:

https://www.ibm.com/docs/en/storage-insights

7.6.3 Reclamation

Use the recommendations on the reclamation page to determine if you can reclaim capacity before planning new capacity purchases. Click **Insights** > **Reclamation** to see how much capacity can be reclaimed as shown in Figure 7-27. See the savings that can be made by reclaiming capacity for tiered and non-tiered storage and a list of the reclaimable volumes. To identify the volumes that are not being used, the storage resources that you add for monitoring are regularly analyzed. A list of the volumes that are not being used is generated. You can decide in accordance with the internal procedures of your organization which of the volumes in the list can be decommissioned.



Name	Tier	Time of Last IO	Pool	Storage System	Capacity (G...	Used Capacity (GiB)	Hosts
GDP_111D		None observed	GM_Ent_T...	DS8800-951-75TN211	2.64	2.64	
GDP_111E		None observed	GM_Ent_T...	DS8800-951-75TN211	2.64	2.64	
GDP_111F		None observed	GM_Ent_T...	DS8800-951-75TN211	2.64	2.64	
GDP_1200		None observed	GM_Ent_T...	DS8800-951-75TN211	2.64	2.64	
GDP_1201		None observed	GM_Ent_T...	DS8800-951-75TN211	2.64	2.64	

Figure 7-27 Reclamation view

Before you delete or reclaim space that was identified by IBM Storage Insights Pro, keep in mind for the volumes on IBM Spectrum Accelerate and CKD volumes on DS8000, the volumes are identified as reclaimable based on I/O activity, because information about the assignment of volumes to servers is not available. To exclude volumes in the reclamation analysis, right-click the volumes and click **Exclude from Analysis**. Additional actions can be performed on the volumes in the recommended reclamation list such as add to an application or general group.

For more information about reclamation with IBM Insights Pro see IBM Documentation center for Insights Pro at:

https://www.ibm.com/docs/en/storage-insights

7.7 Using IBM Storage Insights Pro network functions

All SAN switch and director vendors provide management software that includes performance monitoring capabilities. The real-time SAN statistics, such as port utilization and throughput information available from SAN management software, can be used to complement the performance information that is provided by host servers or storage subsystems.

For more information about monitoring performance through a SAN switch or director point product, see the following websites:

- https://www.broadcom.com
- http://www.cisco.com

Most SAN management software includes options to create SNMP alerts based on performance criteria, and to create historical reports for trend analysis. Certain SAN vendors offer advanced performance monitoring capabilities, such as measuring I/O traffic between specific pairs of source and destination ports, and measuring I/O traffic for specific LUNs.

In addition to the vendor point products, IBM Storage Insights Pro can be used as a monitoring and reporting tool for switch and fabric environments. It collects and reports on data for a one hour (15 minute increments) up to 1 year at 24 hour intervals for performance analysis.

IBM Storage Insights Pro provides facilities to report on fabric topology, configurations and switches, and port performance and errors. In addition, you can use IBM Storage Insights Pro to configure alerts or thresholds for port congestion. send and receive bandwidth and others. Configuration options allow the creation of events to be triggered if thresholds are exceeded. Seamless ticket management from opening and automatically uploading diagnostic information to updating and tracking tickets:

- Ability to store performance data from multiple switch vendors in a common database
- Advanced reporting and correlation between host data and switch data through custom reports
- Centralized management and reporting
- Aggregation of port performance data for the entire switch

In general, you need to analyze SAN statistics for these reasons:

- Ensure that there are no SAN bottlenecks that limit the DS8000 I/O traffic, for example, analyze any link utilization over 80%.
- Confirm that multipathing/load balancing software operates as expected.
- Isolate the I/O activity contributed by adapters on different host servers that share storage subsystem I/O ports.
- Isolate the I/O activity contributed by different storage subsystems accessed by the same host server.

For more information about IBM Storage Insights Pro functions and how to work with them, refer to:

https://www.ibm.com/docs/en/storage-insights

7.8 End-to-end analysis of I/O performance problems

To support tactical performance management processes, problem determination skills and processes must exist. This section explains the logical steps that are required to perform successful problem determination for I/O performance issues. The process of I/O performance problem determination consists of the following steps:

- Define the problem.
- Classify the problem.
- Identify the I/O bottleneck.
- Implement changes to remove the I/O bottleneck.
- Validate that the changes that were made resolved the issue.

Perceived or actual I/O bottlenecks can result from hardware failures on the I/O path, contention on the server, contention on the SAN Fabric, contention on the DS8000 front-end ports, or contention on the back-end device adapters or arrays. This section provides a process for diagnosing these scenarios by using IBM Storage Insights Pro. This process was developed for identifying specific types of problems and is not a substitute for common sense, knowledge of the environment, and experience. Figure 7-28 on page 168 shows the high-level process flow.

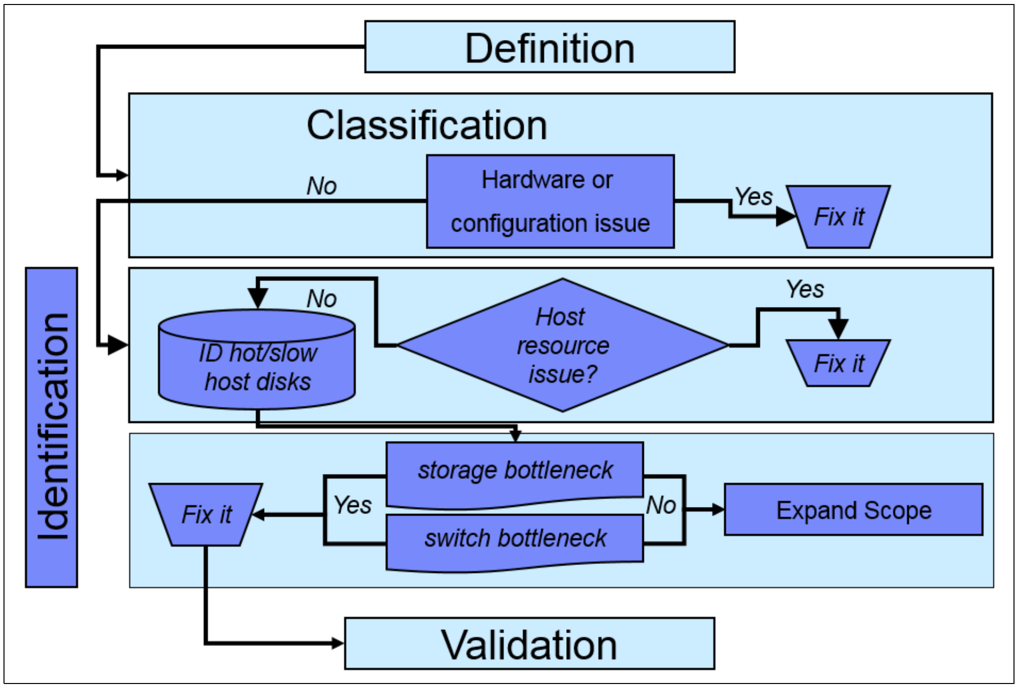


Figure 7-28 I/O performance analysis process

I/O bottlenecks that are referenced in this section relate to one or more components on the I/O path that reached a saturation point and can no longer achieve the I/O performance requirements. I/O performance requirements are typically throughput-oriented or transaction-oriented. Heavy sequential workloads, such as tape backups or data warehouse environments, might require maximum bandwidth and use large sequential transfers. However, they might not have stringent response time requirements. Transaction-oriented workloads, such as online banking systems, might have stringent response time requirements, but have no requirements for throughput.

If a server processor or memory resource shortage is identified, it is important to take the necessary remedial actions. These actions might include but are not limited to adding additional processors, optimizing processes or applications, or adding additional memory. If there are not any resources that are constrained on the server but the end-to-end I/O response time is higher than expected for the DS8000 storage system, a resource constraint likely exists in one or more of the SAN components.

To troubleshoot performance problems, IBM Storage Insights Pro data must be augmented with host performance and configuration data. Figure 7-29 on page 169 shows a logical end-to-end view from a measurement perspective.

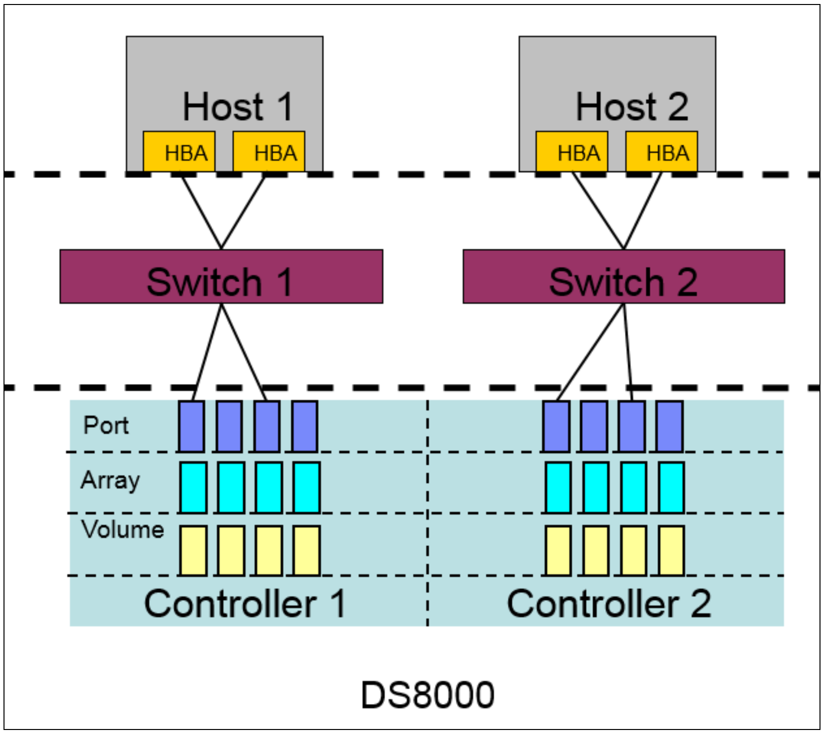


Figure 7-29 End-to-end measurement

Although IBM Insights Pro does not provide host performance, configuration, or error data, IBM Insights Pro provides performance data from host connections, SAN switches, and the DS8000 storage system, and configuration information and error logs from SAN switches and the DS8000 storage system.

You can create and update support tickets and automatically upload logs directly in the IBM Storage Insights interface. You can also give IBM Support permission to collect and upload log packages for storage systems without contacting you every time.

Tip: Performance analysis and troubleshooting must always start top-down, starting with the application (for example, database design and layout), then the operating system, server hardware, SAN, and then storage. The tuning potential is greater at the higher levels. The best I/O tuning is never carried out because server caching or a better database design eliminated the need for it.

Process assumptions

This process assumes that the following conditions exist:

- The server is connected to the DS8000 storage system natively.
- Tools exist to collect the necessary performance and configuration data for each component along the I/O path (server disk, SAN fabric, and the DS8000 arrays, ports, and volumes).
- Skills exist to use the tools, extract data, and analyze data.
- Data is collected in a continuous fashion to facilitate performance management.

Process flow

The order in which you conduct the analysis is important. Use the following process:

1. Define the problem. The goal is to assist you in determining the problem background and understand how the performance requirements are not being met.

   Changes: Before proceeding any further, ensure that an adequate investigation is carried out to identify any changes that were made in the environment. Experience has proven that a correlation exists between changes made to the environment and sudden “unexpected” performance issues.

2. Consider checking the application level first. Has all potential tuning on the database level been performed? Does the layout adhere to the vendor recommendations, and is the server adequately sized (RAM, processor, and buses) and configured?
3. Correctly classify the problem by identifying hardware or configuration issues. Hardware failures often manifest themselves as performance issues because I/O is degraded on one or more paths. If a hardware issue is identified, all problem determination efforts must focus on identifying the root cause of the hardware errors:
   a. Gather any errors on any of the host paths.

      Physical component: If you notice significant errors when querying the path and the errors increase, there is most likely a problem with a physical component on the I/O path.

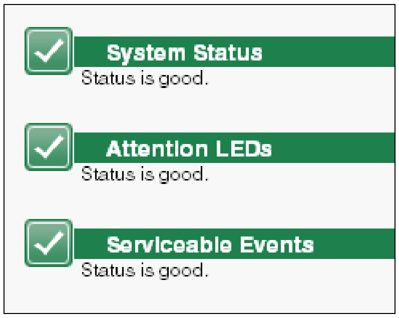
   b. Gather the host error report and look for Small Computer System Interface (SCSI) or Fibre errors.

      Hardware: Often a hardware error that relates to a component on the I/O path shows as a TEMP error. A TEMP error does not exclude a hardware failure. You must perform diagnostic tests on all hardware components in the I/O path, including the host bus adapter (HBA), SAN switch ports, and the DS8000 HBA ports.

   c. Gather the SAN switch configuration and errors. Every switch vendor provides different management software. All of the SAN switch software provides error monitoring and a way to identify whether there is a hardware failure with a port or application-specific integrated circuit (ASIC). For more information about identifying hardware failures, see your vendor-specific manuals or contact vendor support.

      Patterns: As you move from the host to external resources, remember any patterns. A common error pattern that you see involves errors that affect only those paths on the same HBA. If both paths on the same HBA experience errors, the errors are a result of a common component. The common component is likely to be the host HBA, the cable from the host HBA to the SAN switch, or the SAN switch port. Ensure that all of these components are thoroughly reviewed before proceeding.

   d. If errors exist on one or more of the host paths, determine whether there are any DS8000 hardware errors. Log on to the HMC as "customer" and look to ensure that there are no hardware alerts. Figure 7-30 on page 171 provides a sample of a healthy DS8000 storage system. If there are any errors, you might need to open a hardware case with DS8000 hardware support.



☑ System Status
Status is good.

☑ Attention LEDs
Status is good.

☑ Serviceable Events
Status is good.

Figure 7-30 DS8000 storage system - healthy HMC

4. After validating that no hardware failures exist, analyze server performance data and identify any disk bottlenecks. The fundamental premise of this methodology is that I/O performance degradation that relates to SAN component contention can be observed at the server through analysis of the key server-based I/O metrics.

   Degraded end-to-end I/O response time is the strongest indication of I/O path contention. Typically, server physical disk response times measure the time that a physical I/O request takes from the moment that the request was initiated by the device driver until the device driver receives an interrupt from the controller that the I/O completed. The measurements are displayed as either service time or response time. They are averaged over the measurement interval. Typically, server wait or queue metrics refer to time spent waiting at the HBA, which is usually an indication of HBA saturation. In general, you need to interpret the service times as response times because they include potential queuing at various storage subsystem components, for example:

   – Switch
   – Storage HBA
   – Storage cache
   – Storage back-end drive controller
   – Storage back-end paths
   – Drives

   Important: Subsystem-specific load-balancing software usually does not add any performance impact and can be viewed as a pass-through layer.

In addition to the drive response time and queuing data, gather the drive activity rates, including read I/Os, write I/Os, and total I/Os because they show which drives are active:

a. Gather performance data, as shown in Table 7-6.

Table 7-6 Native tools and key metrics

OS	Native tool	Command/Object	Metric/Counter
AIX	iostat, filemon	**`iostat -D`**, **`filemon -o /tmp/fmon.log -O all`**	read time(ms) write time(ms) reads, writes queue length
HP-UX	sar	**`sar -d`**	avserv(ms) avque blks/s
Linux	iostat	**`iostat -d`**	svctm(ms) avgqu-sz tps
Solaris	iostat	**`iostat -xn`**	svc_t(ms) Avque blks/s
Microsoft Windows Server	perfmon	Physical disk	Avg Disk Sec/Read Avg Disk Sec/Write Read Disk Queue Length Write Disk Queue Length Disk Reads/sec Disk Writes/sec
IBM z	Resource Measurement Facility (RMF)/System Management Facilities (SMF)	See Chapter 10, “Performance considerations for IBM z Systems servers” on page 225.	N/A

I/O-intensive disks: The number of total I/Os per second indicates the relative activity of the device. This relative activity provides a metric to prioritize the analysis. Those devices with high response times and high activity are more important to understand than devices with high response time and infrequent access. If analyzing the data in a spreadsheet, consider creating a combined metric of Average I/Os × Average Response Time to provide a method for identifying the most I/O-intensive disks. You can obtain additional detail about OS-specific server analysis in the OS-specific chapters.

b. Gather configuration data from the default multipathing software for distributed systems. In addition to the multipathing configuration data, you must collect configuration information for the host and DS8000 HBAs, which includes the bandwidth of each adapter.

Multipathing: Ensure that multipathing works as designed. For example, if there are two paths that are zoned per HBA to the DS8000 storage system, there must be four active paths per LUN. The default MPIO should be configured to use an active/active configuration of multipathing, which means that traffic flows across all the traffic fairly evenly. For native DS8000 connections, the absence of activity on one or more paths indicates a problem with the multipathing.

c. Format the data and correlate the host LUNs with their associated DS8000 resources. Formatting the data is not required for analysis, but it is easier to analyze formatted data in a spreadsheet.

   The following steps represent the logical steps that are required to format the data and do not represent literal steps. You can codify these steps in scripts:

   i. Read the configuration file.

   ii. Build a hdisk hash with `key = hdisk` and `value = LUN SN`.

   iii. Read I/O response time data.

   iv. Create hashes for each of the following values with hdisk as the key: Date, Start time, Physical Volume, Reads, Avg Read Time, Avg Read Size, Writes, Avg Write Time, and Avg Write Size.

   v. Print the data to a file with headers and commas to separate the fields.

   vi. Iterate through the hdisk hash and use the common hdisk key to index into the other hashes and print those hashes that have values.

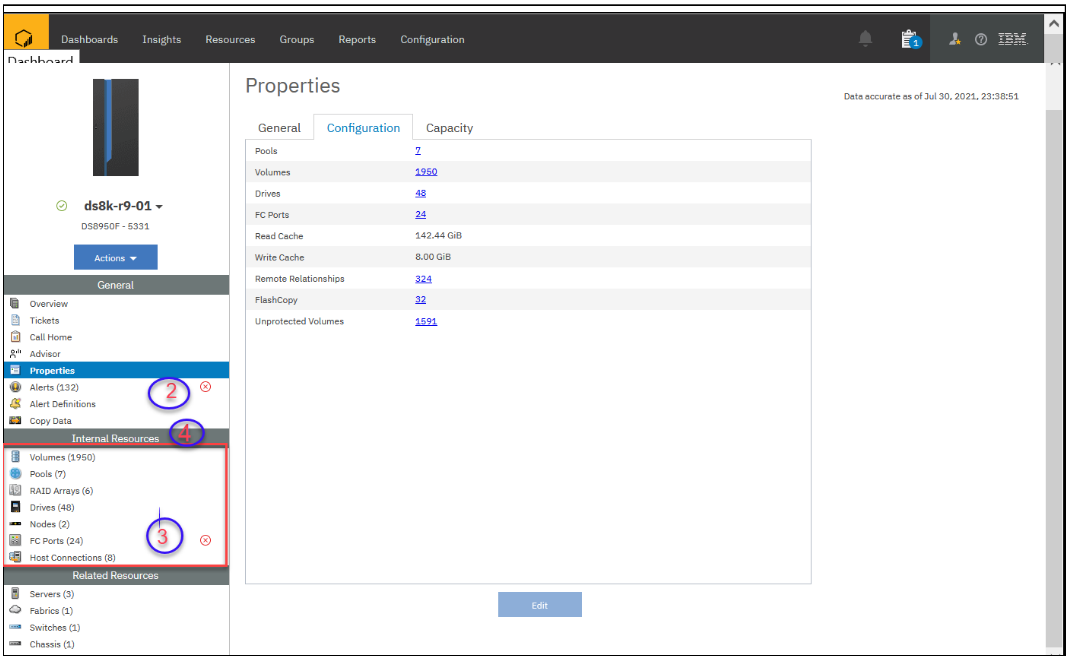
d. Analyze the host performance data:

   i. Determine whether I/O bottlenecks exist by summarizing the data and analyzing key performance metrics for values in excess of the thresholds. Identify those vpaths/LUNs with poor response time. Hardware errors and multipathing configuration issues must already be excluded. The hot LUNs must already be identified. Proceed to step 5 on page 173 to determine the root cause of the performance issue.

   ii. If no degraded disk response times exist, the issue is likely not internal to the server.

5. If there are drive constraints that are identified, continue the identification of the root cause by collecting and analyzing the DS8000 configuration and performance data:

   a. Gather the configuration information. IBM Storage Insights Pro can also be used to gather configuration data through the Properties window, as shown in Figure 7-31 on page 174.



Data accurate as of Jul 30, 2021, 23:38:51

Pools	7
Volumes	1950
Drives	48
FC Ports	24
Read Cache	142.44 GiB
Write Cache	8.00 GiB
Remote Relationships	324
FlashCopy	32
Unprotected Volumes	1591

Figure 7-31 DS8000 Properties window in IBM Storage Insights Pro

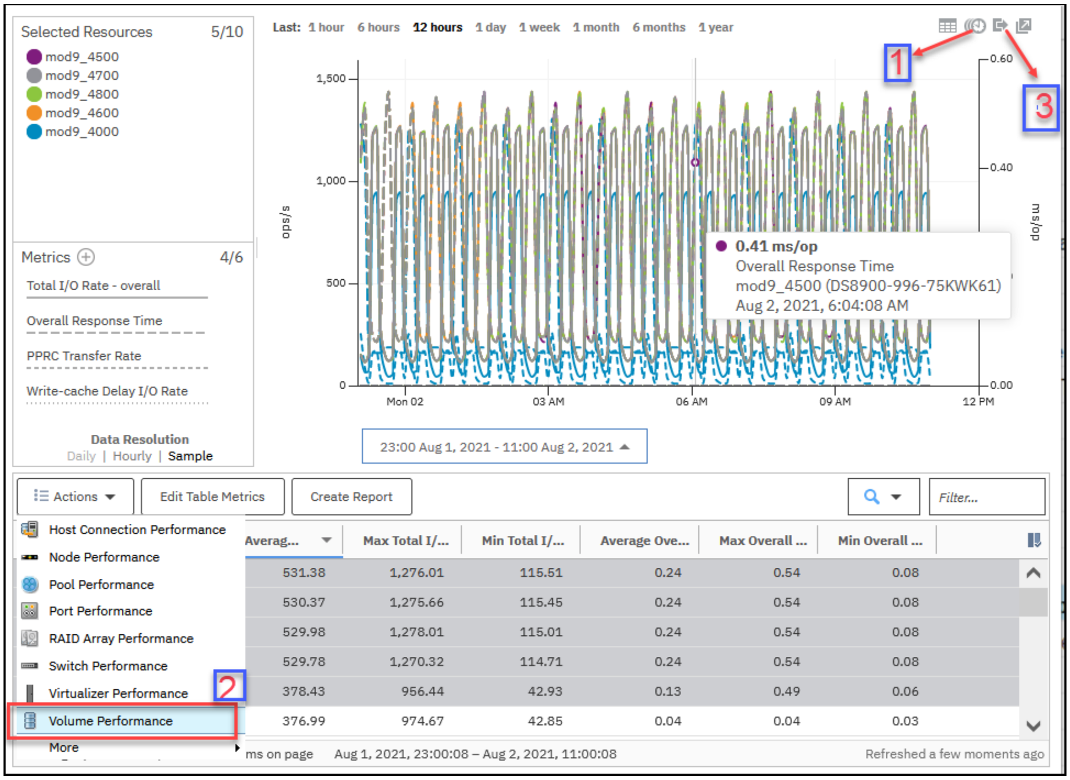
> **Analyze the DS8000 performance data first:** Check for Alerts (2) and errors (3) in the left navigation. Then, look at the performance data of the internal resources (4). Analysis of the SAN fabric and the DS8000 performance data can be completed in either order. However, SAN bottlenecks occur less frequently than drive bottlenecks, so it can be more efficient to analyze the DS8000 performance data first.

b. Use IBM Storage Insights Pro to gather the DS8000 performance data for fiber channel ports, pools, arrays, volumes, nodes, and host connections. Compare the key performance indicators from Table 7-5 on page 158 with the performance data. To analyze the performance, complete the following steps:

i. For those server LUNs that show poor response time, analyze the associated volumes during the same period. If the problem is on the DS8000 storage system, a correlation exists between the high response times observed on the host and the volume response times observed on the DS8000 storage system.

> **Compare the same period:** Meaningful correlation with the host performance measurement and the previously identified hot LUNs requires analysis of the DS8000 performance data for the same period that the host data was collected. The synchronize time function of IBM Storage Insights Pro can help you with this task (see Figure 7-32 on page 175 (1)). For more information, see IBM Documentation: https://www.ibm.com/docs/en/storage-insights. For more information about time stamps, see 7.3.1, “Time stamps” on page 156.

ii. Correlate the hot LUNs with their associated arrays. When using the IBM Storage Insights Pro, the relationships are provided automatically in the drill-down feature, as shown in Figure 7-32 on page 175 (2).



0.41 ms/op
Overall Response Time
mod9_4500 (DS8900-996-75KWK61)
Aug 2, 2021, 6:04:08 AM

Averag...	Max Total I/...	Min Total I/...	Average Ove...	Max Overall ...	Min Overall ...
531.38	1,276.01	115.51	0.24	0.54	0.08
530.37	1,275.66	115.45	0.24	0.54	0.08
529.98	1,278.01	115.01	0.24	0.54	0.08
529.78	1,270.32	114.71	0.24	0.54	0.08
378.43	956.44	42.93	0.13	0.49	0.06
376.99	974.67	42.85	0.04	0.04	0.03

Figure 7-32 Drill-down function of IBM Storage Insights Pro

If you use export reports, shown in Figure 7-32 on page 175 (3), and want to correlate the volume data to the rank data, you can correlate the volume data to the rank data manually or by using the script. If multiple ranks per extent pool and storage pool striping, or Easy Tier managed pools are used, one volume can exist on multiple ranks. Easy Tier can help alleviate any hot spots with automatic rebalancing.

Analyze storage subsystem ports for the ports associated with the server in question.

6. Continue the identification of the root cause by collecting and analyzing SAN fabric configuration and performance data:

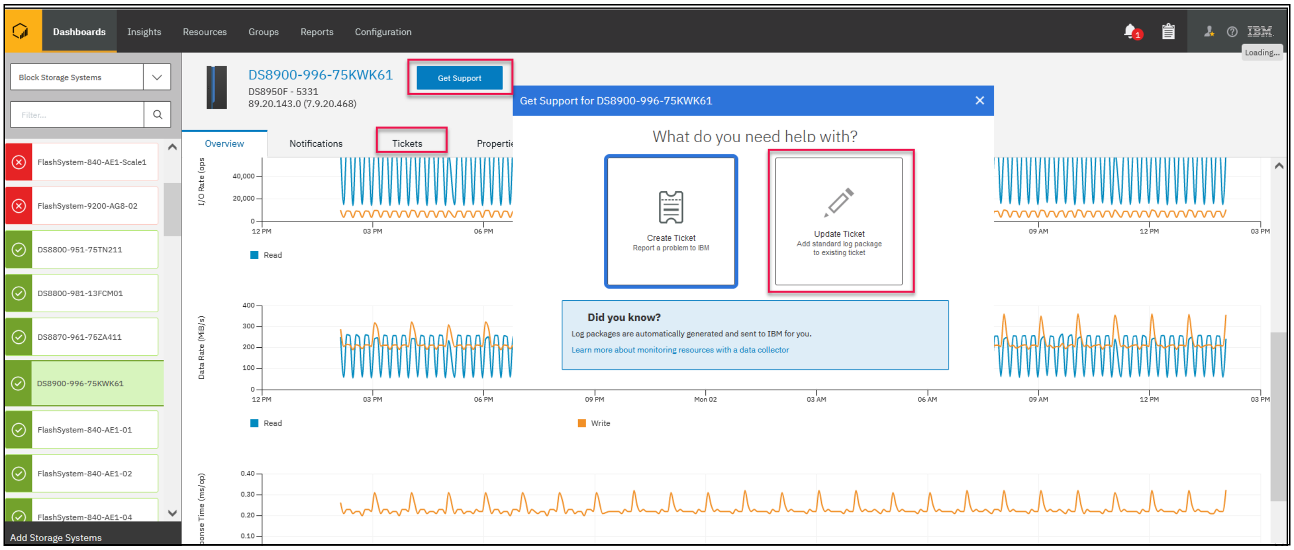
   a. Gather the connectivity information and establish a visual diagram of the environment.

   > **Visualize the environment:** Sophisticated tools are not necessary for creating this type of view; however, the configuration, zoning, and connectivity information must be available to create a logical visual representation of the environment.

   b. Gather the SAN performance data. Each vendor provides SAN management applications that provide the alerting capability and some level of performance management. Often, the performance management software is limited to real-time monitoring, and historical data collection features require additional licenses. In addition to the vendor-provided solutions, IBM Storage Insights Pro can collect further metrics, which are shown in Table 7-5 on page 158.

   c. Consider graphing the Overall Port Response Time, Port Bandwidth Percentage, and Total Port Data Rate metrics to determine whether any of the ports along the I/O path are saturated during the time when the response time is degraded. If the Total Port Data Rate is close to the maximum expected throughput for the link or the bandwidth percentages that exceed their thresholds, this situation is likely a contention point. You can add additional bandwidth to mitigate this type of issue either by adding additional links or by adding faster links. Adding links might require upgrades of the server HBAs and the DS8000 HAs to take advantage of the additional switch link capacity.

d. Create a **Ticket** (**Case**) and automatically upload logs directly in IBM Storage Insights Pro. Click **Dashboards** > **Operations** then select the storage system you want to create a case for and click **Get Support** as shown in Figure 7-33 on page 176. You can view updates and add logs to a ticket: Click the **Tickets** tab, select the ticket, then click **Add Log Package to Ticket**.



What do you need help with?

Did you know?

Log packages are automatically generated and sent to IBM for you.

Learn more about monitoring resources with a data collector

Figure 7-33 Get Support in IBM Storage Insights Pro

> **Note:** If you are eligible for Premium Support, it's recommended that you call IBM Support or open a support case at https://www.ibm.com/mysupport/ to access that service for your issue. Ensure that you have your Direct Access Code (DAC) number ready so IBM can best assist you.

7.9 Performance analysis examples

This section provides sample performance data, analysis, and recommendations for the following performance scenarios by using the process that is described in 7.8, “End-to-end analysis of I/O performance problems” on page 167. The examples highlight the key performance data that is appropriate for each problem type. It provides the host configuration or errors only in the cases where that information is critical to determine the outcome.

7.9.1 Example 1: Array bottleneck

The most common type of performance problem is an array bottleneck. Similar to other types of I/O performance problems, an array bottleneck usually manifests itself in high response time on the host. In many cases, the write response times are excellent because of cache hits, but reads often require immediate disk access.

Defining the problem

The application owner complains of poor response time for transactions during certain times of the day.

Classifying the problem

There are no hardware errors, configuration issues, or host performance constraints.

Identifying the root cause

Figure 7-34 shows the average read response time for a Windows Server that performs a random workload in which the response time increases steadily over time.

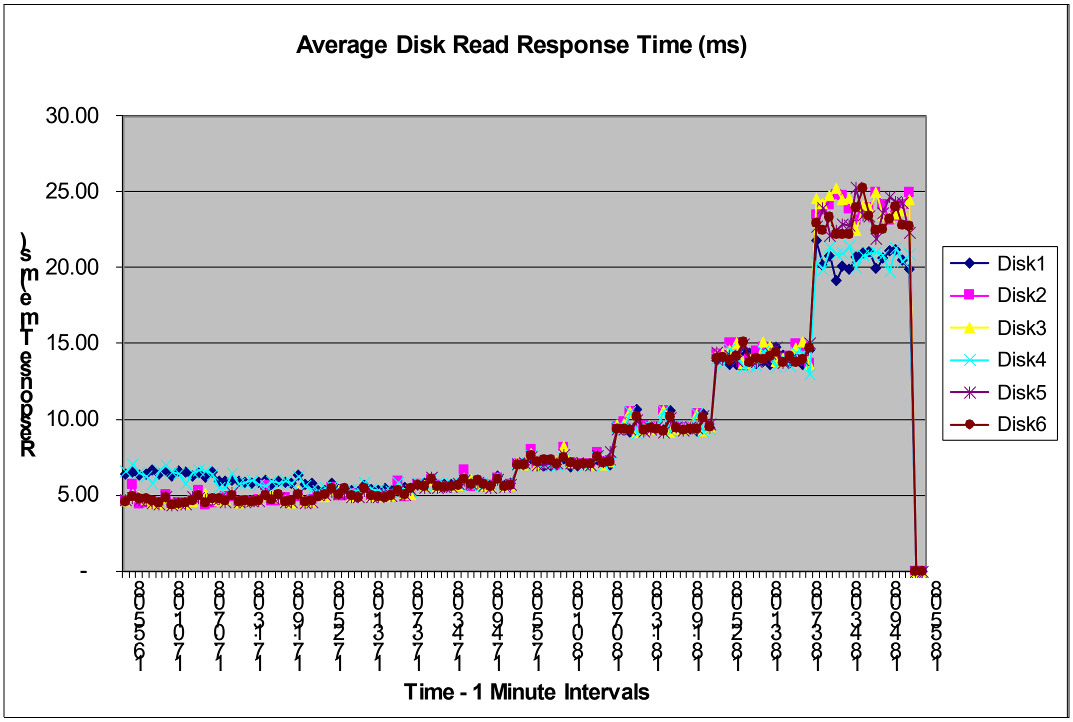


Figure 7-34 Windows Server perfmon - Average Physical Disk Read Response Time

At approximately 18:39 hours, the average read response time jumps from approximately 15 ms to 25 ms. Further investigation of the host reveals that the increase in response time correlates with an increase in load, as shown in Figure 7-35 on page 178.

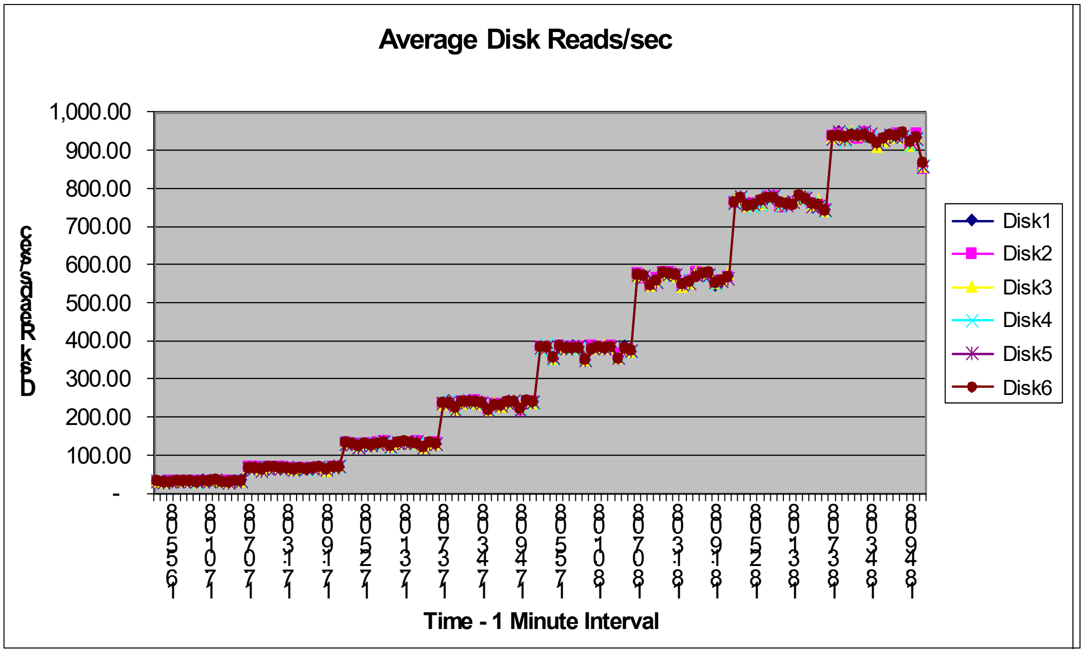


Figure 7-35 Windows Server perfmon - Average Disk Reads/sec

As described in 7.8, “End-to-end analysis of I/O performance problems” on page 167, there are several possibilities for high average disk read response time:

- DS8000 array contention
- DS8000 port contention
- SAN fabric contention
- Host HBA saturation

Because the most probable reason for the elevated response times is the drive utilization on the array, gather and analyze this metric first. Figure 7-36 on page 178 shows the drive utilization on the DS8000 storage system.

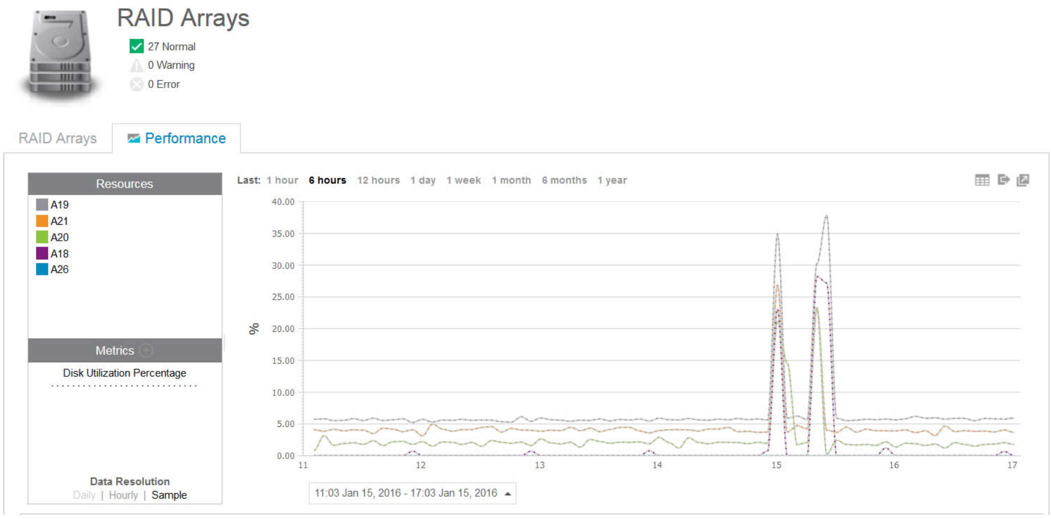


Figure 7-36 IBM Storage Insights Pro array drive utilization

Implementing changes to resolve the problem

Add volumes on additional drives. For environments where host striping is configured, you might need to re-create or migrate the host volumes to spread the I/O from an existing workload across the new volumes.

Validating the problem resolution

Gather performance data to determine whether the issue is resolved.

7.9.2 Example 2: Hardware connectivity part 1

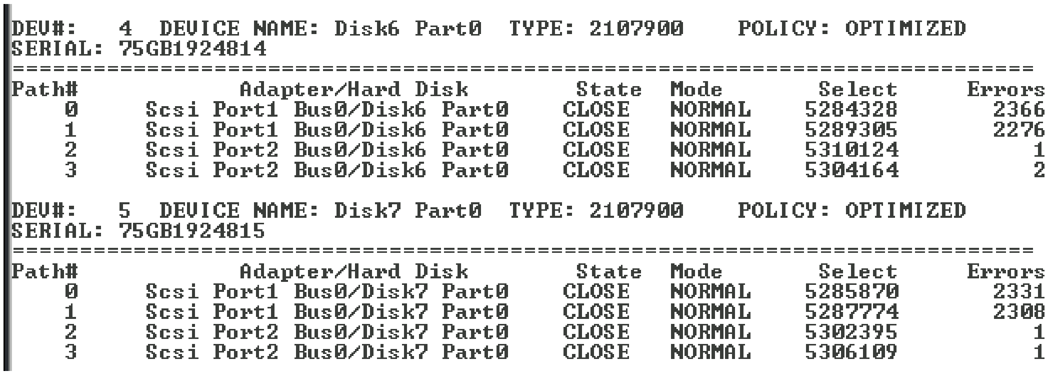
Infrequent connectivity issues occur as a result of broken or damaged components in the I/O path. The following example illustrates the required steps to identify and resolve these types of issues.

Problem definition

The online transactions for a Windows Server SQL server appear to take longer than normal and time out in certain cases.

Problem classification

After reviewing the hardware configuration and the error reports for all hardware components, we determined that there are errors on the paths associated with one of the host HBAs, as shown in Figure 7-37 on page 179. This output shows the errors on path 0 and path 1, which are both on the same HBA (SCSI port 1). For a Windows Server that runs MPIO, additional information about the HAs is available. The command that you use to identify errors depends on the multipathing software installation.



```
DEV#:   4  DEVICE NAME: Disk6 Part0  TYPE: 2107900    POLICY: OPTIMIZED
SERIAL: 75GB1924814
============================================================================
Path#            Adapter/Hard Disk        State   Mode       Select     Errors
    0      Scsi Port1 Bus0/Disk6 Part0    CLOSE   NORMAL    5284328       2366
    1      Scsi Port1 Bus0/Disk6 Part0    CLOSE   NORMAL    5289305       2276
    2      Scsi Port2 Bus0/Disk6 Part0    CLOSE   NORMAL    5310124          1
    3      Scsi Port2 Bus0/Disk6 Part0    CLOSE   NORMAL    5304164          2

DEV#:   5  DEVICE NAME: Disk7 Part0  TYPE: 2107900    POLICY: OPTIMIZED
SERIAL: 75GB1924815
============================================================================
Path#            Adapter/Hard Disk        State   Mode       Select     Errors
    0      Scsi Port1 Bus0/Disk7 Part0    CLOSE   NORMAL    5285870       2331
    1      Scsi Port1 Bus0/Disk7 Part0    CLOSE   NORMAL    5287774       2308
    2      Scsi Port2 Bus0/Disk7 Part0    CLOSE   NORMAL    5302395          1
    3      Scsi Port2 Bus0/Disk7 Part0    CLOSE   NORMAL    5306109          1
```

Figure 7-37 Example of information available

Identifying the root cause

A further review of the switch software revealed significant errors on the switch port associated with the paths in question. A visual inspection of the environment revealed a kink in the cable from the host to the switch.

Implementing changes to resolve the problem

Replace the cable.

Validating the problem resolution

After you implement the change, the error counts do not increase and the nightly backups complete within the backup window.

7.9.3 Example 3: Hardware connectivity part 2

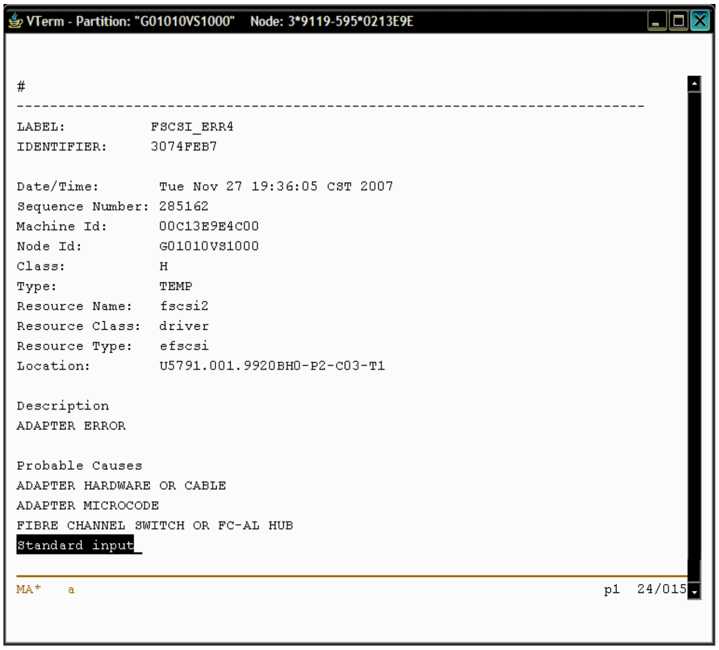
Infrequent connectivity issues occur as a result of broken or damaged components in the I/O path. The following example illustrates the required steps to identify and resolve these types of issues.

Defining the problem

Users report that the data warehouse application on an AIX server does not complete jobs in a reasonable amount of time. Online transactions also time out.

Classifying the problem

A review of the host error log shows a significant number of hardware errors. An example of the errors is shown in Figure 7-38 on page 180.


VTerm - Partition: "G01010VS1000"   Node: 3*9119-595*0213E9E

```
#
---------------------------------------------------------------------------
LABEL:          FSCSI_ERR4
IDENTIFIER:     3074FEB7

Date/Time:       Tue Nov 27 19:36:05 CST 2007
Sequence Number: 285162
Machine Id:      00C13E9E4C00
Node Id:         G01010VS1000
Class:           H
Type:            TEMP
Resource Name:   fscsi2
Resource Class:  driver
Resource Type:   efscsi
Location:        U5791.001.9920BH0-P2-C03-T1

Description
ADAPTER ERROR

Probable Causes
ADAPTER HARDWARE OR CABLE
ADAPTER MICROCODE
FIBRE CHANNEL SWITCH OR FC-AL HUB
Standard input_

MA*   a                                                                 p1  24/015
```

Figure 7-38 AIX error log

Identifying the root cause

The IBM Service Support Representative (SSR) ran IBM diagnostic tests on the host HBA, and the card did not pass the diagnostic tests.

> **Disabling a path:** In cases where there is a path with significant errors, you can disable the path with the multipathing software, which allows the non-working paths to be disabled without causing performance degradation to the working paths.

Implementing changes to resolve the problem

Replace the card.

Validating the problem resolution

The errors did not persist after the card is replaced and the paths are brought online.

7.9.4 Example 4: Port bottleneck

DS8000 port bottlenecks do not occur often, but they are a component that is typically oversubscribed.

Defining the problem

The production server batch runs exceed their batch window.

Classifying the problem

There are no hardware errors, configuration issues, or host performance constraints.

Identifying the root cause

The production server throughput diminishes at approximately 18:30 hours daily. Concurrently, development workloads that run on the same DS8000 ports increase. Figure 7-39 on page 181 and Figure 7-40 on page 182 demonstrate the overall workload from both the production server and the development server.

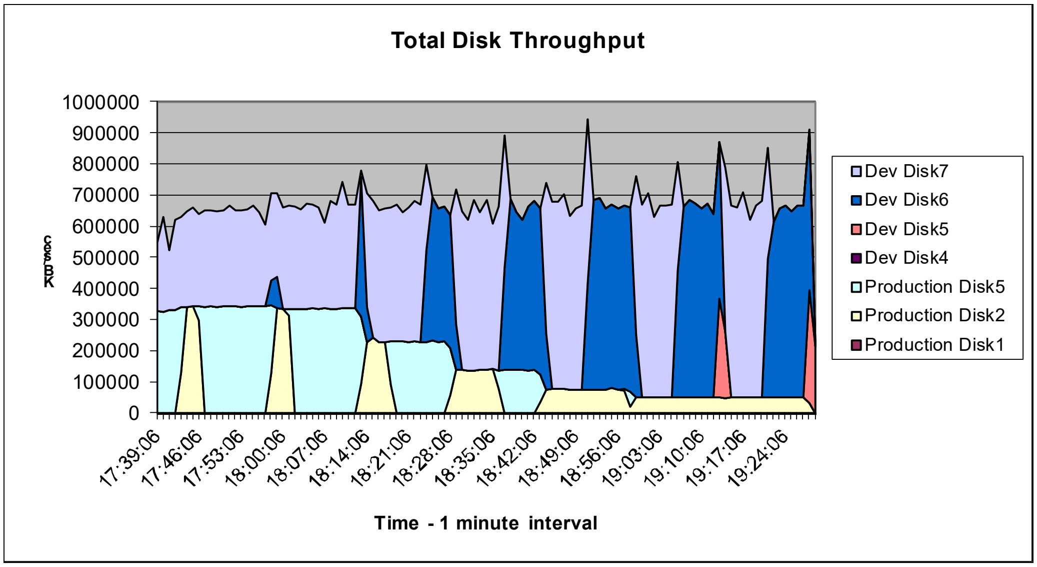


Figure 7-39 Production throughput compared to development throughput

The DS8000 port data reveals a peak throughput of around 300 MBps per 4-Gbps port.

> **Note:** The HA ports speed might differ from DS8000 models, so it is a value that depends on the hardware configuration of the system and the SAN environment.

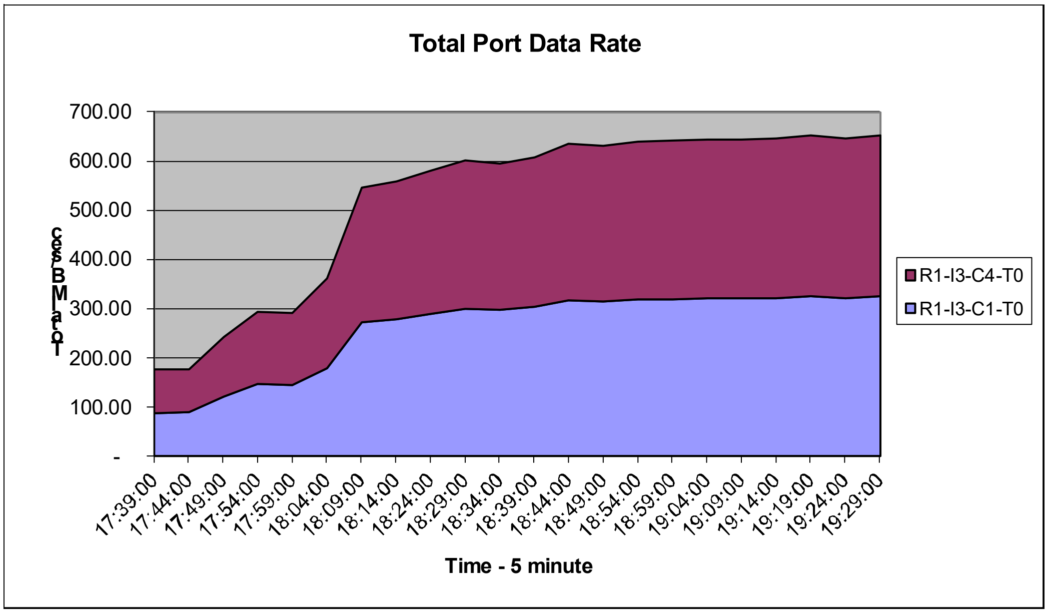


Figure 7-40 Total port data rate << image reference missing in text >>

Implementing changes to resolve the problem

Rezone ports for production servers and development servers so that they do not use the same DS8000 ports. Add additional ports so that each server HBA is zoned to two DS8000 ports.

Validating the problem resolution

After implementing the new zoning that separates the production server and the development server, the storage ports are no longer the bottleneck.

7.10 IBM Storage Insights Pro in mixed environments

A benefit of IBM Storage Insights Pro is its capability to analyze both distributed fixed block (FB) and mainframe Count Key Data (CKD) workloads. When the DS8000 storage system is attached to multiple hosts that run on different platforms, distributed hosts might affect your mainframe workload, and the mainframe workload might affect the distributed workloads. If you use a mixed environment, looking at the RMF reports is insufficient. You also need the information about the Open Systems hosts. IBM Storage Insights Pro informs you about the cache and I/O activity.

Before beginning the diagnostic process, you must understand your workload and your physical configuration. You must know how your system resources are allocated, and understand your path and channel configuration for all attached servers.

Assume that you have an environment with a DS8000 storage system attached to a z/OS host, an AIX on IBM Power Systems host, and several Windows Server hosts. You noticed that your z/OS online users experience a performance degradation 07:30 - 08:00 hours each morning.

You might notice that there are 3390 volumes that indicate high disconnect times, or high device busy delay time for several volumes in the RMF device activity reports. Unlike UNIX or

Windows Server, you might notice the response time and its breakdown to connect, disconnect, pending, and IOS queuing.

Disconnect time is an indication of cache-miss activity or destage wait (because of persistent memory high utilization) for logical drives behind the DS8000 storage systems.

Device busy delay is an indication that another system locks up a volume, and an extent conflict occurs among z/OS hosts or applications in the same host when using Parallel Access Volumes (PAVs). The DS8000 multiple allegiance or PAVs capability allows it to process multiple I/Os against the same volume at the same time. However, if a read or write request against an extent is pending while another I/O is writing to the extent, or if a write request against an extent is pending while another I/O is reading or writing data from the extent, the DS8000 storage system delays the I/O by queuing. This condition is referred as extent conflict. Queuing time because of extent conflict is accumulated to device busy (DB) delay time. An extent is a sphere of access; the unit of increment is a track. Usually, I/O drivers or system routines decide and declare the sphere.

To determine the possible cause of high disconnect times, check the read cache hit ratios, read-to-write ratios, and bypass I/Os for those volumes. If you see that the cache hit ratio is lower than usual and you did not add other workloads to your IBM z environment, I/Os against FB volumes might be the cause of the problem. It is possible that FB volumes which are defined on the same server have a cache-unfriendly workload, thus affecting your IBM z volumes hit ratio.

To get more information about cache usage, you can check the cache statistics of the FB volumes that belong to the same server. You might be able to identify the FB volumes that have a low read hit ratio and short cache holding time. Moving the workload of the FB logical disks, or the CKD volumes, that you are concerned about to the other side of the cluster improves the situation by concentrating cache-friendly I/O workload across both clusters. If you cannot or if the condition does not improve after this move, consider balancing the I/O distribution on more ranks. Balancing the I/O distribution on more ranks optimizes the staging and destaging operation.

The scenarios that use IBM Storage Insights Pro as described in this chapter might not cover all the possible situations that can be encountered. You might need to include more information, such as application and host operating system-based performance statistics or other data collections to analyze and solve a specific performance problem.

Part 3

Performance considerations for host systems and databases

This part provides performance considerations for various host systems or appliances that are attached to the IBM System Storage DS8000 storage system, and for databases.

This part includes the following topics:

- Host attachment
- Performance considerations for the IBM i system
- Performance considerations for IBM z Systems servers
- Database for IBM z/OS performance

8

Host attachment

This chapter describes the following attachment topics and considerations between host systems and the DS8000 series for availability and performance:

- DS8000 host attachment types.
- Attaching Open Systems hosts.
- Attaching IBM Z hosts.

This chapter provides detailed information about performance tuning considerations for specific operating systems in later chapters of this book.

8.1 DS8000 host attachment

The DS8000 enterprise storage solution provides various host attachments that allow exceptional performance and superior data throughput. At a minimum, have two connections to any host, and the connections must be on different host adapters (HAs) in different I/O enclosures. You can consolidate storage capacity and workloads for Open Systems hosts and z Systems hosts by using the following adapter types and protocols:

- Fibre Channel Protocol (FCP)-attached Open Systems hosts.
- FCP/Fibre Connection (FICON)-attached z Systems hosts.
- zHyperLink Connection (zHyperLink)-attached z Systems hosts.

The DS8900F supports up to 32 Fibre Channel (FC) host adapters, with four FC ports for each adapter. Each port can be independently configured to support Fibre Channel connection (IBM FICON) or Fibre Channel Protocol (FCP).

Each adapter type is available in both longwave (LW) and shortwave (SW) versions. The DS8900F I/O bays support up to four host adapters for each bay, allowing up to 128 ports maximum for each storage system. This configuration results in a theoretical aggregated host I/O bandwidth around 128 x 32 Gbps. Each port provides industry-leading throughput and I/O rates for FICON and FCP.

The host adapters that are available in the DS8900F have the following characteristics:

- 32 Gbps FC HBAs (32 Gigabit Fibre Channel - GFC):
  - Four FC ports
  - FC Gen7 technology
  - New IBM Custom ASIC with Gen3 PCIe interface
  - Quad-core PowerPC processor
  - Negotiation to 32, 16, or 8 Gbps; 4 Gbps or less is not possible
- 16 Gbps FC HBAs (16 GFC):
  - Four FC ports
  - Gen2 PCIe interface
  - Quad-core PowerPC processor
  - Negotiation to 16, 8, or 4 Gbps (2 Gbps or less is not possible)

The DS8900F supports a mixture of 32 Gbps and 16 Gbps FC adapters. Hosts with slower FC speeds like 4 Gbps are still supported if their HBAs are connected through a switch.

The 32 Gbps FC adapter is encryption-capable. Encrypting the host bus adapter (HBA) traffic usually does not cause any measurable performance degradation.

With FC adapters that are configured for FICON, the DS8900F series provides the following configuration capabilities:

- Fabric or point-to-point topologies.
- A maximum of 128 host adapter ports, depending on the DS8900F system memory and processor features.
- A maximum of 509 logins for each FC port.
- A maximum of 8192 logins for each storage unit.
- A maximum of 1280 logical paths on each FC port.

- Access to all 255 control-unit images (65,280 CKD devices) over each FICON port.
- A maximum of 512 logical paths for each control unit image.

An IBM Z server supports 32,768 devices per FICON host channel. To fully access 65,280 devices, it is necessary to connect multiple FICON host channels to the storage system. You can access the devices through an FC switch or FICON director to a single storage system FICON port.

The 32 GFC host adapter doubles the data throughput of 16 GFC links. When comparing the two adapter types, the 32 GFC adapters provide I/O improvements in full adapter I/Os per second (IOPS) and reduced latency.

The DS8000 storage system can support host and remote mirroring links by using Peer-to-Peer Remote Copy (PPRC) on the same I/O port. However, it is preferable to use dedicated I/O ports for remote mirroring links.

zHyperLink reduces the I/O latency by providing a point-to-point communication path between IBM z14®, z14 ZR1 or z15™ Central Processor Complex (CPC) and the I/O bay of DS8000 family through optical fiber cables and zHyperLink adapters. zHyperLink is intended to complement FICON and speed up I/O requests that are typically used for transaction processing. This goal is accomplished by installing zHyperLink adapters on the z/OS host and the IBM DS8900 hardware family, connecting them using zHyperLink cables to perform synchronous I/O operations. zHyperLink connections are limited to a distance of 150 meters (492 feet). All DS8900 models have zHyperLink capability, and the maximum number of adapters is 12 on DS8950F and DS8980F models and 4 adapters on the DS8910F.

Planning and sizing the HAs for performance are not easy tasks, so use modeling tools, such as Storage Modeller, see Chapter 6.1, “IBM Storage Modeller” on page 126.

The factors that might affect the performance at the HA level are typically the aggregate throughput and the workload mix that the adapter can handle. All connections on a HA share bandwidth in a balanced manner. Therefore, host attachments that require maximum I/O port performance may be connected to HAs that are not fully populated. You must allocate host connections across I/O ports, HAs, and I/O enclosures in a balanced manner (workload spreading).

Find further fine-tuning guidelines in Chapter 4.10.2, “Further optimization” on page 103.

8.2 Attaching Open Systems hosts

This section describes the host system requirements and attachment considerations for Open Systems hosts running AIX, Linux, Hewlett-Packard UNIX (HP-UX), Oracle Solaris, VMware, and Microsoft Windows to the DS8000 series with Fibre Channel (FC) adapters.

> **No Ethernet/iSCSI:** There is no direct Ethernet/iSCSI, or Small Computer System Interface (SCSI) attachment support for the DS8000 storage system.

8.2.1 Fibre Channel

FC is a full-duplex, serial communications technology to interconnect I/O devices and host systems that might be separated by tens of kilometers. The DS8900 storage system supports 32, 16, 8, and 4 Gbps connections and it negotiates the link speed automatically.

Supported Fibre Channel-attached hosts

For specific considerations that apply to each server platform, and for the *current* information about supported servers (the list is updated periodically), refer to:

http://www.ibm.com/systems/support/storage/config/ssic

Fibre Channel topologies

The DS8900F architecture supports two FC interconnection topologies:

- Direct connect
- Switched fabric

Note: With DS8900F, Fibre Channel Arbitrated Loop (FC-AL) is no longer supported.

For maximum flexibility and performance, use a switched fabric topology.

The next section describes best practices for implementing a switched fabric.

8.2.2 Storage area network implementations

This section describes a basic storage area network (SAN) network and how to implement it for maximum performance and availability. It shows examples of a correctly connected SAN network to maximize the throughput of disk I/O.

Description and characteristics of a storage area network

With a SAN, you can connect heterogeneous Open Systems servers to a high-speed network and share storage devices, such as disk storage and tape libraries. Instead of each server having its own locally attached storage and tape drives, a SAN shares centralized storage components, and you can efficiently allocate storage to hosts.

Storage area network cabling for availability and performance

For availability and performance, you must connect to different adapters in different I/O enclosures whenever possible. You must use multiple FC switches or directors to avoid a potential single point of failure. You can use inter-switch links (ISLs) for connectivity.

Importance of establishing zones

For FC attachments in a SAN, it is important to establish *zones* to prevent interaction from HAs. Every time a HA joins the fabric, it issues a Registered State Change Notification (RSCN). An RSCN does not cross zone boundaries, but it affects every device or HA in the same zone.

If a HA fails and starts logging in and out of the switched fabric, or a server must be restarted several times, you do not want it to disturb the I/O to other hosts. Figure 8-1 shows zones that include only a single HA and multiple DS8900F ports (single initiator zone). This approach is the preferred way to create zones to prevent interaction between server HAs.

Tip: Each zone contains a single host system adapter with the wanted number of ports attached to the DS8000 storage system.

By establishing zones, you reduce the possibility of interactions between system adapters in switched configurations. You can establish the zones by using either of two zoning methods:

- Port number
- Worldwide port name (WWPN)

You can configure switch ports that are attached to the DS8000 storage system in more than one zone, which enables multiple host system adapters to share access to the DS8000 HA ports. Shared access to a DS8000 HA port might be from host platforms that support a combination of bus adapter types and operating systems.

Important: A DS8000 HA port configured to run with the FICON topology cannot be shared in a zone with non-z/OS hosts. Ports with non-FICON topology cannot be shared in a zone with z/OS hosts.

LUN masking

In FC attachment, LUN affinity is based on the WWPN of the adapter on the host, which is independent of the DS8000 HA port to which the host is attached. This LUN masking function on the DS8000 storage system is provided through the definition of DS8000 volume groups. A volume group is defined by using the DS Storage Manager or DS8000 command-line interface (DSCLI), and host WWPNs are connected to the volume group. The LUNs to be accessed by the hosts that are connected to the volume group are defined to be in that volume group.

Although it is possible to limit through which DS8000 HA ports a certain WWPN connects to volume groups, it is preferable to define the WWPNs to have access to all available DS8000 HA ports. Then, by using the preferred process of creating FC zones, as described in “Importance of establishing zones” on page 190, you can limit the wanted HA ports through the FC zones. In a switched fabric with multiple connections to the DS8000 storage system, this concept of LUN affinity enables the host to see the same LUNs on different paths.

Configuring logical disks in a storage area network

In a SAN, carefully plan the configuration to prevent many disk device images from being presented to the attached hosts. Presenting many disk devices to a host can cause longer failover times in cluster environments. Also, boot times can take longer because the device discovery steps take longer.

The number of times that a DS8000 logical disk is presented as a disk device to an open host depends on the number of paths from each HA to the DS8000 storage system. The number of paths from an open server to the DS8000 storage system is determined by these factors:

- The number of HAs installed in the server
- The number of connections between the SAN switches and the DS8000 storage system
- The zone definitions created by the SAN switch software

Physical paths: Each physical path to a logical disk on the DS8000 storage system is presented to the host operating system as a disk device.

Consider a SAN configuration, as shown in Figure 8-1 :

- The host has two connections to the SAN switches, and each SAN switch in turn has four connections to the DS8000 storage system.
- Zone A includes one FC card (FC0) and two paths from SAN switch A to the DS8000 storage system.
- Zone B includes one FC card (FC1) and two paths from SAN switch B to the DS8000 storage system.
- This host uses only four of the eight possible paths to the DS8000 storage system in this zoning configuration.

By cabling the SAN components and creating zones, as shown in Figure 8-1 on page 192, each logical disk on the DS8000 storage system is presented to the host server four times because there are four unique physical paths from the host to the DS8000 storage system. As you can see in Figure 8-1 on page 192, Zone A shows that FC0 has access through DS8000 host ports I0000 and I0130. Zone B shows that FC1 has access through DS8000 host ports I0230 and I0300. So, in combination, this configuration provides four paths to each logical disk presented by the DS8000 storage system. If Zone A and Zone B are modified to include four paths each to the DS8000 storage system, the host has a total of eight paths to the DS8000 storage system. In that case, each logical disk that is assigned to the host is presented as eight physical disks to the host operating system. Additional DS8000 paths are shown as connected to Switch A and Switch B, but they are not in use for this example.

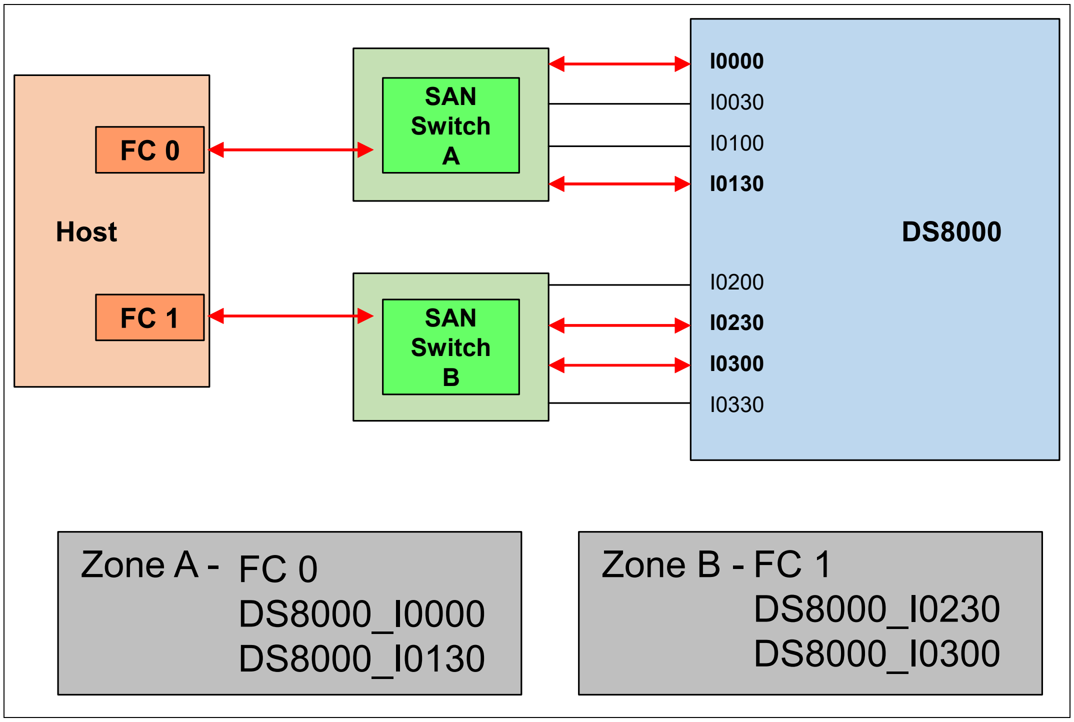


Figure 8-1 Zoning in a SAN environment

You can see how the number of logical devices that are presented to a host can increase rapidly in a SAN environment if you are not careful about selecting the size of logical disks and the number of paths from the host to the DS8000 storage system.

Typically, it is preferable to cable the switches and create zones in the SAN switch software for dual-attached hosts so that each server HA has 2 - 4 paths from the switch to the DS8000 storage system. With hosts configured this way, you can allow the multipathing module to balance the load across the four HAs in the DS8000 storage system.

Between 2 and 4 paths for each specific LUN in total are usually a good compromise for smaller and mid-size servers. Consider eight paths if a high data rate is required. Zoning more paths to a certain LUN, such as more than eight connections from the host to the DS8000 storage system, does not improve SAN performance and can cause too many devices to be presented to the operating system.

8.2.3 Multipathing

Multipathing describes a technique to attach one host to an external storage device through more than one path. Multipathing can improve fault tolerance and the performance of the overall system because the fault of a single component in the environment can be tolerated without an impact to the host. Also, you can increase the overall system bandwidth, which positively influences the performance of the system.

As illustrated in Figure 8-2 on page 193, attaching a host system by using a single-path connection implements a solution that depends on several single points of failure. In this example, as a single link, failure either between the host system and the switch, between the switch and the storage system, or a failure of the HA on the host system, the DS8000 storage system, or even a failure of the switch leads to a loss of access of the host system. Additionally, the path performance of the whole system is reduced by the slowest component in the link.

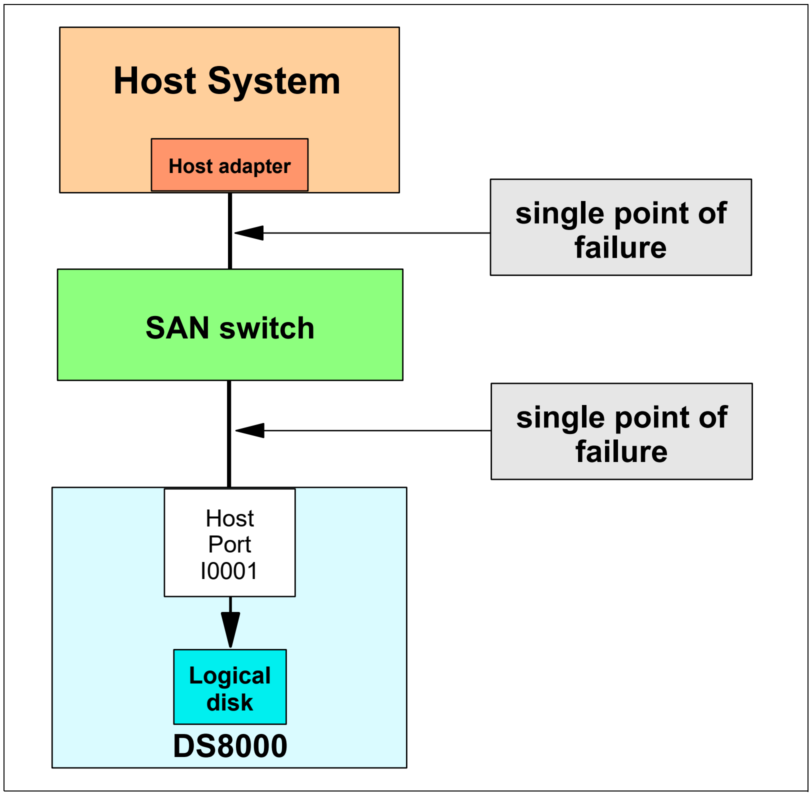


Figure 8-2 SAN single-path connection

Adding additional paths requires you to use multipathing software (Figure 8-3). Otherwise, the same LUN behind each path is handled as a separate disk from the operating system side, which does not allow failover support.

Multipathing provides the DS8000 attached Open Systems hosts that run Windows, AIX, HP-UX, Oracle Solaris, VMware, KVM, or Linux with these capabilities:

- Support for several paths per LUN.
- Load balancing between multiple paths when there is more than one path from a host server to the DS8000 storage system. This approach might eliminate I/O bottlenecks that occur when many I/O operations are directed to common devices through the same I/O path, thus improving the I/O performance.

- Automatic path management, failover protection, and enhanced data availability for users that have more than one path from a host server to the DS8000 storage system. It eliminates a potential single point of failure by automatically rerouting I/O operations to the remaining active paths from a failed data path.
- Dynamic reconfiguration after changing the environment, including zoning, LUN masking, and adding or removing physical paths.

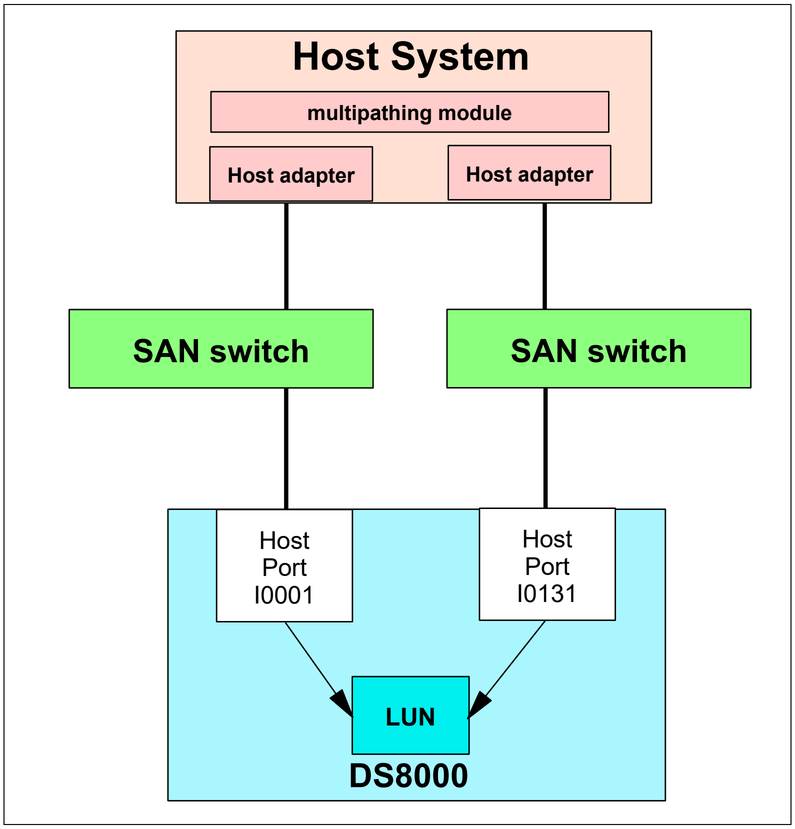


Figure 8-3 DS8000 multipathing implementation that uses two paths

The DS8000 storage system supports several multipathing implementations. Depending on the environment, host type, and operating system, only a subset of the multipathing implementations is available. This section introduces the multipathing concepts and provides general information about implementation, usage, and specific benefits.

> **Important:** Do not intermix several multipathing solutions within one host system. Usually, the multipathing software solutions cannot coexist.

Multipath I/O

Multipath I/O (MPIO) summarizes native multipathing technologies that are available in several operating systems, such as AIX, Linux, and Windows. Although the implementation differs for each of the operating systems, the basic concept is almost the same:

- The multipathing module is delivered with the operating system.
- The multipathing module supports failover and load balancing for standard SCSI devices, such as simple SCSI disks or SCSI arrays.

- To add device-specific support and functions for a specific storage device, each storage vendor might provide a device-specific module that implements advanced functions for managing the specific storage device.

8.2.4 Example: AIX

AIX brings its own MPIO, and some of these commands are very powerful to see what is going on in the system.

Example 8-1 shows a smaller AIX LPAR which has 8 volumes, belonging to two different DS8000 storage systems. The **lsmpio -q** command gives us the sizes and the DS8000 volume serial numbers. **lsmpio -a** gives us the WWPNs of the host being used, that we have already told the DS8000 to make our volumes known to. **lsmpio -ar** additionally gives us the WWPNs of the DS8000 host ports being used, and how many SAN paths are being used between a DS8000 port and the IBM Power server port.

Example 8-1 Working with AIX MPIO

```
[p8-e870-01v20:root:/home/root:] lsmpio -q
Device           Vendor Id  Product Id       Size       Volume Name
--------------------------------------------------------------------------------
hdisk0           IBM        2107900            50.00GiB 8050
hdisk1           IBM        2107900            50.00GiB 8150
hdisk2           IBM        2107900            80.00GiB 8000
hdisk3           IBM        2107900            80.00GiB 8001
hdisk4           IBM        2107900            80.00GiB 8100
hdisk5           IBM        2107900            80.00GiB 8101
hdisk6           IBM        2107900            10.00TiB 8099
hdisk7           IBM        2107900             4.00TiB 81B0


[p8-e870-01v20:root:/home/root:] lsmpio -ar
Adapter Driver: fscsi0 -> AIX PCM
    Adapter WWPN:  c05076082d8201ba
    Link State:    Up
                          Paths      Paths      Paths      Paths
    Remote Ports        Enabled   Disabled     Failed    Missing         ID
    500507630311513e          2          0          0          0    0xc1d00
    500507630a1013e7          4          0          0          0    0xc1000

Adapter Driver: fscsi1 -> AIX PCM
    Adapter WWPN:  c05076082d8201bc
    Link State:    Up
                          Paths      Paths      Paths      Paths
    Remote Ports        Enabled   Disabled     Failed    Missing         ID
    500507630313113e          2          0          0          0    0xd1500
    500507630a1b93e7          4          0          0          0    0xd1000
```

In the DS8000, we can use **lshostconnect -login** to see that our server is logged in with its adapter WWPNs, and through which DS8000 ports.

With **lsmpio -q -l hdisk*X*** we get additional information about one specific volume. As seen in Example 8-2 on page 196, we see the capacity and which DS8000 storage system is hosting that LUN. We see that it is a 2107-998 model (DS8980F), and in the Volume Serial, the WWPN of the DS8000 is also encoded.

Example 8-2 Getting more information on one specific volume

```
[p8-e870-01v20:root:/home/root:] lsmpio -q -l hdisk6
Device:  hdisk6
          Vendor Id:  IBM
         Product Id:  2107900
           Revision:  .431
           Capacity:  10.00TiB
       Machine Type:  2107
       Model Number:  998
         Host Group:  V1
        Volume Name:  8099
      Volume Serial:  600507630AFFD3E70000000000008099  (Page 83 NAA)
```

As we see in Example 8-3 with the **lsmpio -l hdisk*X*** command, initially, only one of the paths is selected by the operating system, for this LUN. **lsattr -El** shows us having the algorithm fail_over, for that LUN (and besides, here it also shows the serial number of the DS8000 we use).

Example 8-3 hdisk before algorithm change

```
[p8-e870-01v20:root:/home/root:] lsmpio -l hdisk6
name    path_id  status   path_status  parent  connection

hdisk6  0        Enabled  Sel          fscsi0  500507630a1013e7,4080409900000000
hdisk6  1        Enabled               fscsi1  500507630a1b93e7,4080409900000000


[p8-e870-01v20:root:/home/root:] lsattr -El hdisk6
DIF_prot_type   none                             T10 protection type              False
DIF_protection  no                               T10 protection support           True
FC3_REC         false                            Use FC Class 3 Error Recovery    True
PCM             PCM/friend/aixmpiods8k           Path Control Module              False
PR_key_value    none                             Persistant Reserve Key Value     True+
algorithm       fail_over                        Algorithm                        True+
...
max_retry_delay 60                               Maximum Quiesce Time             True
max_transfer    0x80000                          Maximum TRANSFER Size            True
node_name       0x500507630affd3e7               FC Node Name                     False
...
queue_depth     20                               Queue DEPTH                      True+
reassign_to     120                              REASSIGN time out value          True
reserve_policy  single_path                      Reserve Policy                   True+
...
timeout_policy  fail_path                        Timeout Policy                   True+
unique_id       200B75HAL91809907210790003IBMfcp Unique device identifier         False
ww_name         0x500507630a1013e7               FC World Wide Name               False
```

But we can change this path algorithm, with the **chdev -l -a** command, as in Example 8-4. For this to work, the hdisk needs to be taken offline.

Example 8-4 Changing the path algorithm to shortest_queue

```
[p8-e870-01v20:root:/home/root:] chdev -l hdisk6 -a algorithm=shortest_queue -a
reserve_policy=no_reserve
hdisk6 changed
```

As seen in Example 8-5, after the hdisk is taken online, both power adapters and paths are selected, and the algorithm is changed to shortest_queue, including balancing for this individual LUN.

Example 8-5 hdisk after algorithm change

```
[p8-e870-01v20:root:/home/root:] lsmpio -l hdisk6
name    path_id  status   path_status  parent  connection

hdisk6  0        Enabled  Sel          fscsi0  500507630a1013e7,4080409900000000
hdisk6  1        Enabled  Sel          fscsi1  500507630a1b93e7,4080409900000000


[p8-e870-01v20:root:/home/root:] lsattr -El hdisk6
...
PCM             PCM/friend/aixmpiods8k          Path Control Module                  False
PR_key_value    none                            Persistant Reserve Key Value         True+
algorithm       shortest_queue                  Algorithm                            True+
...
reserve_policy  no_reserve                      Reserve Policy                       True+
...
```

Find more hints on setting up AIX multipathing at:

https://developer.ibm.com/articles/au-aix-mpio

8.3 Attaching z Systems hosts

This section describes the host system requirements and attachment considerations for attaching the z Systems hosts (z/OS, IBM z/VM, IBM z/VSE®, Linux on z Systems, KVM, and Transaction Processing Facility (TPF)) to the DS8000 series. The following sections describe attachment through FICON adapters only.

FCP: z/VM, z/VSE, and Linux for z Systems can also be attached to the DS8000 storage system with FCP. Then, the same considerations as for Open Systems hosts apply.

8.3.1 FICON

FICON is a Fibre Connection used with z Systems servers. Each storage unit HA has either four or eight ports, and each port has a unique WWPN. You can configure the port to operate with the FICON upper-layer protocol. When configured for FICON, the storage unit provides the following configurations:

- Either fabric or point-to-point topology.
- A maximum of 128 host ports for a DS8950F and DS8980F models and a maximum of 64 host ports for a DS8910F model.
- A maximum of 1280 logical paths per DS8000 HA port.
- Access to all 255 control unit images (65280 count key data (CKD) devices) over each FICON port. FICON HAs support 4, 8, 16or 32 Gbps link speeds in DS8000 storage systems.

Improvement of FICON generations

IBM introduced FICON channels in the IBM 9672 G5 and G6 servers with the capability to run at 1 Gbps. Since that time, IBM has introduced several generations of FICON channels. The FICON Express16S channels make up the current generation. They support 16 Gbps link speeds and can auto-negotiate to 4 or 8 Gbps. The speed also depends on the capability of the director or control unit port at the other end of the link.

Operating at 16 Gbps speeds, FICON Express16S channels achieve up to 2600 MBps for a mix of large sequential read and write I/O operations, as shown in the following charts. Figure 8-4 shows a comparison of the overall throughput capabilities of various generations of channel technology.

Due to the implementation of new Application-Specific Integrated Circuits (ASIC), and new internal design, the FICON Express16S+ and FICON EXPRESS16SA on IBM z15, represent a significant improvement in maximum bandwidth capability compared to FICON Express16S channels and previous FICON offerings. The response time improvements are expected to be noticeable for large data transfers. The FICON Express16SA channel operates at the same 16 Gbps line rate as the previous generation FICON Express16S+ channel. Therefore, the FICON Express16SA channel provides similar throughput performance when compared to the FICON Express16S+.

With the introduction of the FEx16SA channel, improvements can be seen in both response times and maximum throughput for IO/sec and MB/sec for workloads using the zHPF protocol on FEx16SA channels.

With Fibre Channel Endpoint Security encryption enabled, the FEx16SA channel has the added benefit of Encryption of Data In Flight for FCP (EDIF) operations with less than 4% impact on maximum channel throughput.

High Performance FICON for IBM Z (zHPF) is implemented for throughput and latency, which it optimizes by reducing the number of information units (IU) that are processed. Enhancements to the z/Architecture and the FICON protocol provide optimizations for online transaction processing (OLTP) workloads.

The size of most online database transaction processing (OLTP) workload I/O operations is 4K bytes. In laboratory measurements, a FICON Express16SA channel, using the zHPF protocol and small data transfer I/O operations, achieved a maximum of 310 KIO/sec. Also, using FICON Express16SA in a z15 using the zHPF protocol and small data transfer FC-ES Encryption enabled achieved a maximum of 304 KIO/sec.

Figure 8-4 on page 199 shows the maximum single channel throughput capacities of each generation of FICON Express channels supported on z15. For each of the generations of FICON Express channels, it displays the maximum capability using the zHPF protocol exclusively.

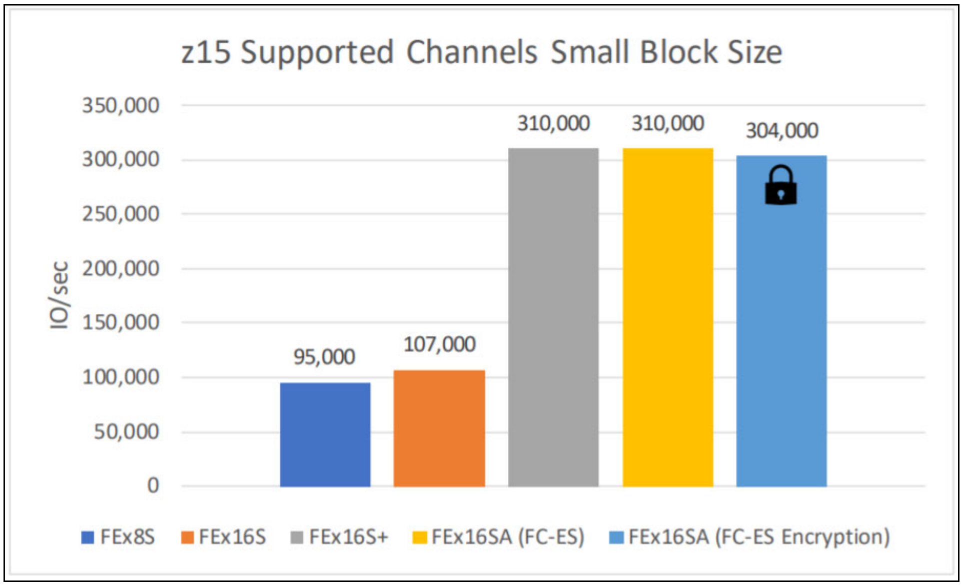


Figure 8-4 FICON Express Channels Maximum IOPS

Figure 8-5 displays the maximum READ/WRITE (mix) MB/sec for each channel. Using FEx16SA in a z15 with the zHPF protocol and a mix of large sequential read and update write data transfer I/O operations, laboratory measurements achieved a maximum throughput of 3,200 READ/WRITE (mix) MB/sec. Also, using FEx16SA in a z15 with the zHPF protocol with FC-ES Encryption and a mix of large sequential read and update write data transfer I/O operations achieved a maximum throughput of 3,200 READ/WRITE (mix) MB/sec.

.

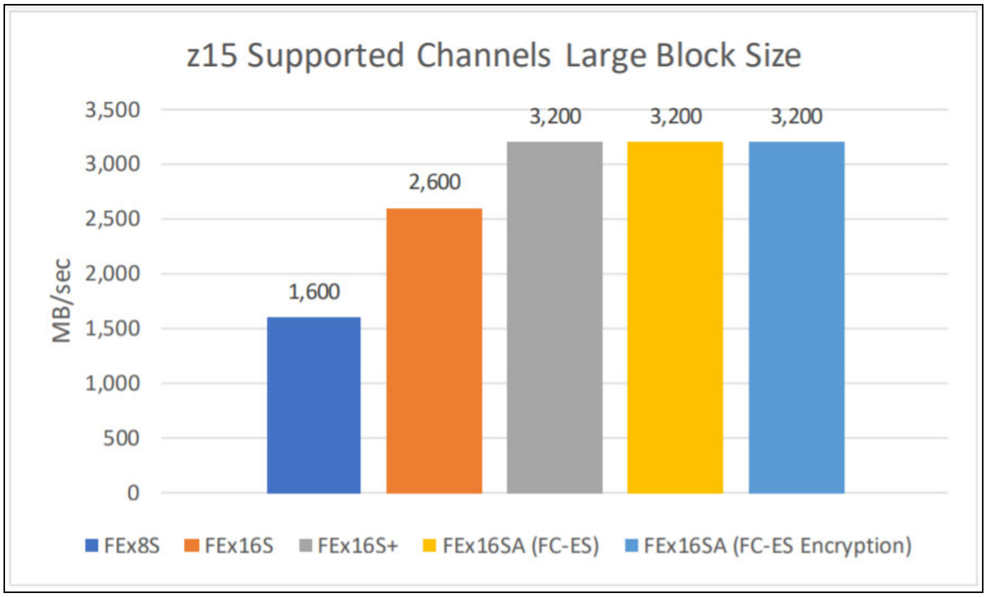


Figure 8-5 FICON Express Channels Maximum MB/sec over latest generations

The z14 and z15 CPCs offer FICON Express16S and FICON Express 16S+ SX and LX features with two independent channels. Each feature occupies a single I/O slot and uses one CHPID per channel. Each channel supports 4 Gbps, 8 Gbps, and 16 Gbps link data rates with auto-negotiation to support existing switches, directors, and storage devices.

FICON Support:

- The FICON Express16S is supported on the z13®, z14 and z15 CPCs.
- The FICON Exprss16S+ is available on the z14 and z15 CPCs.
- The FICON Express16SA is available only on the z15 CPC.

For any generation of FICON channels, you can attach directly to a DS8000 storage system, or you can attach through a FICON capable FC switch.

8.3.2 FICON configuration and sizing considerations

When you use an FC/FICON HA to attach to FICON channels, either directly or through a switch, the port is dedicated to FICON attachment and cannot be simultaneously attached to FCP hosts. When you attach a DS8000 storage system to FICON channels through one or more switches, the maximum number of FICON *logical paths* is 1280 per DS8000 HA port. The FICON directors (switches) provide high availability with redundant components and no single points of failure. A *single* director between servers and a DS8000 storage system is not preferable because it can be a single point of failure. This section describes FICON connectivity between z Systems and a DS8000 storage system and provides recommendations and considerations.

FICON topologies

As shown in Figure 8-6 on page 201, FICON channels in FICON native mode, which means CHPID type FC in Input/Output Configuration Program (IOCP), can access the DS8000 storage system through the following topologies:

- Point-to-Point (direct connection)
- Switched Point-to-Point (through a single FC switch)
- Cascaded FICON Directors (through two FC switches)

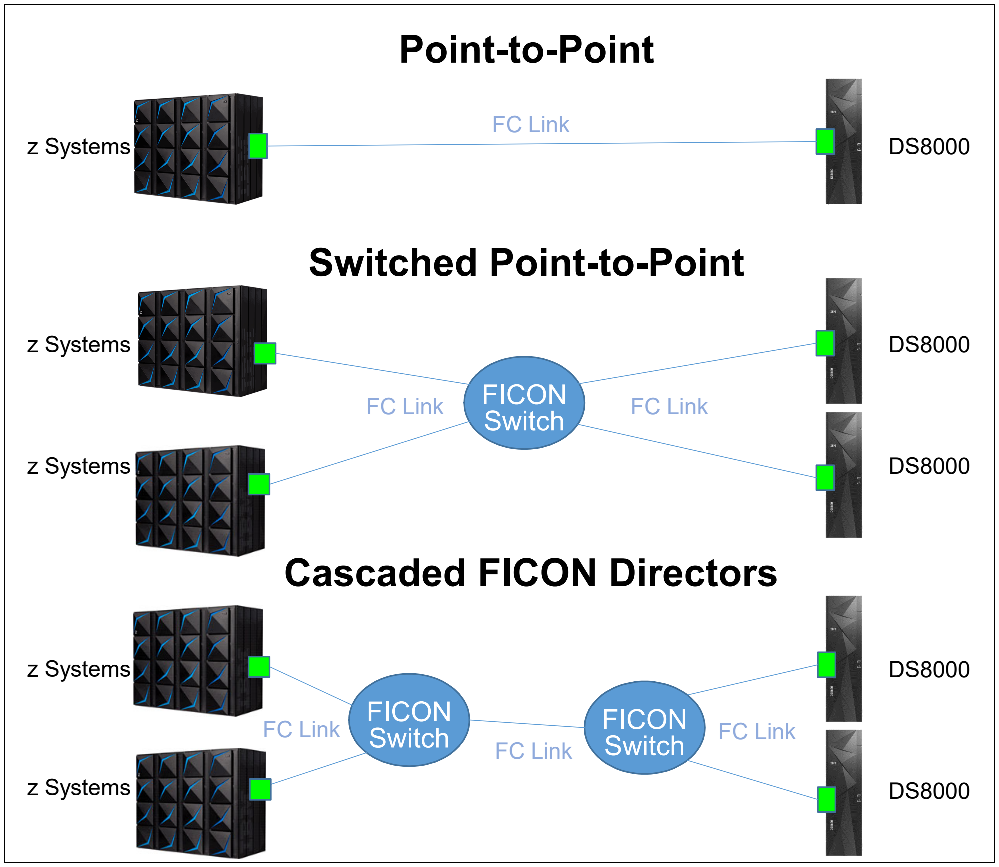


Figure 8-6 FICON topologies between z Systems and a DS8000 storage system

FICON connectivity

Usually, in IBM Z environments, a one-to-one connection between FICON channels and storage HAs is preferred because the FICON channels are shared among multiple logical partitions (LPARs) and heavily used. Carefully plan the oversubscription of HA ports to avoid any bottlenecks.

Figure 8-7 on page 202 shows an example of FICON attachment that connects a z Systems server through FICON switches. This example uses 16 FICON channel paths to eight HA ports on the DS8000 storage system and addresses eight logical control units (LCUs). This channel consolidation might be possible when your aggregate host workload does not exceed the performance capabilities of the DS8000 HA.

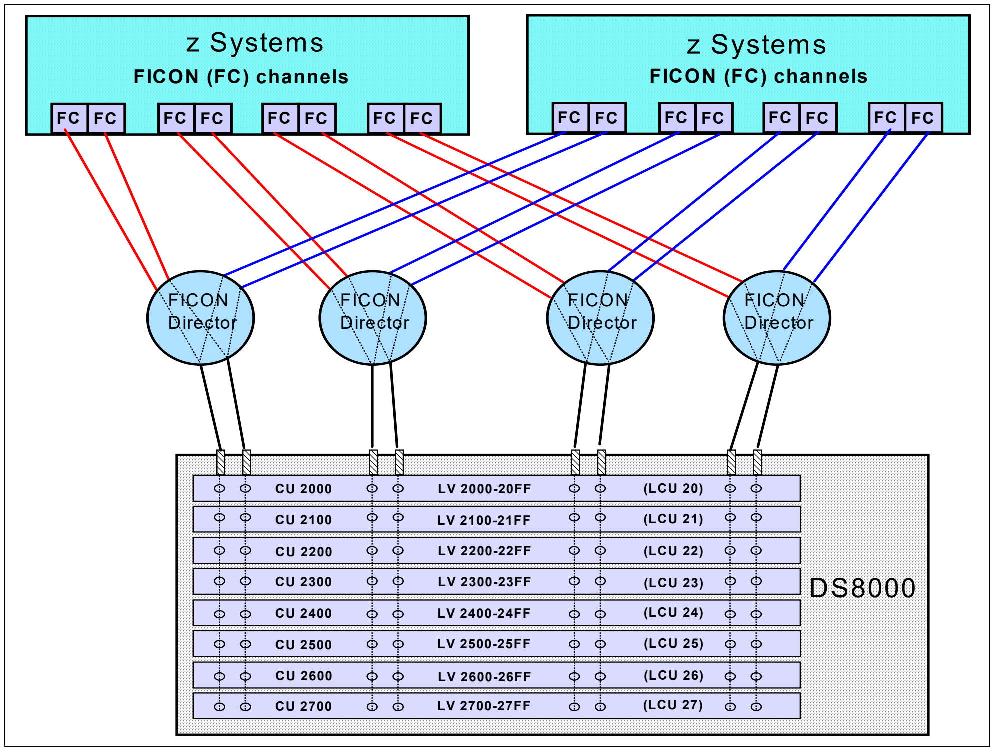

z Systems
FICON (FC) channels
FC FC FC FC FC FC FC FC
z Systems
FICON (FC) channels
FC FC FC FC FC FC FC FC
FICON Director
FICON Director
FICON Director
FICON Director
CU 2000 LV 2000-20FF (LCU 20)
CU 2100 LV 2100-21FF (LCU 21)
CU 2200 LV 2200-22FF (LCU 22)
CU 2300 LV 2300-23FF (LCU 23)
CU 2400 LV 2400-24FF (LCU 24)
CU 2500 LV 2500-25FF (LCU 25)
CU 2600 LV 2600-26FF (LCU 26)
CU 2700 LV 2700-27FF (LCU 27)
DS8000

Figure 8-7 Many-to-one FICON attachment

A one-to-many configuration is also possible, as shown in Figure 8-8 on page 202, but careful planning is needed to avoid performance issues.

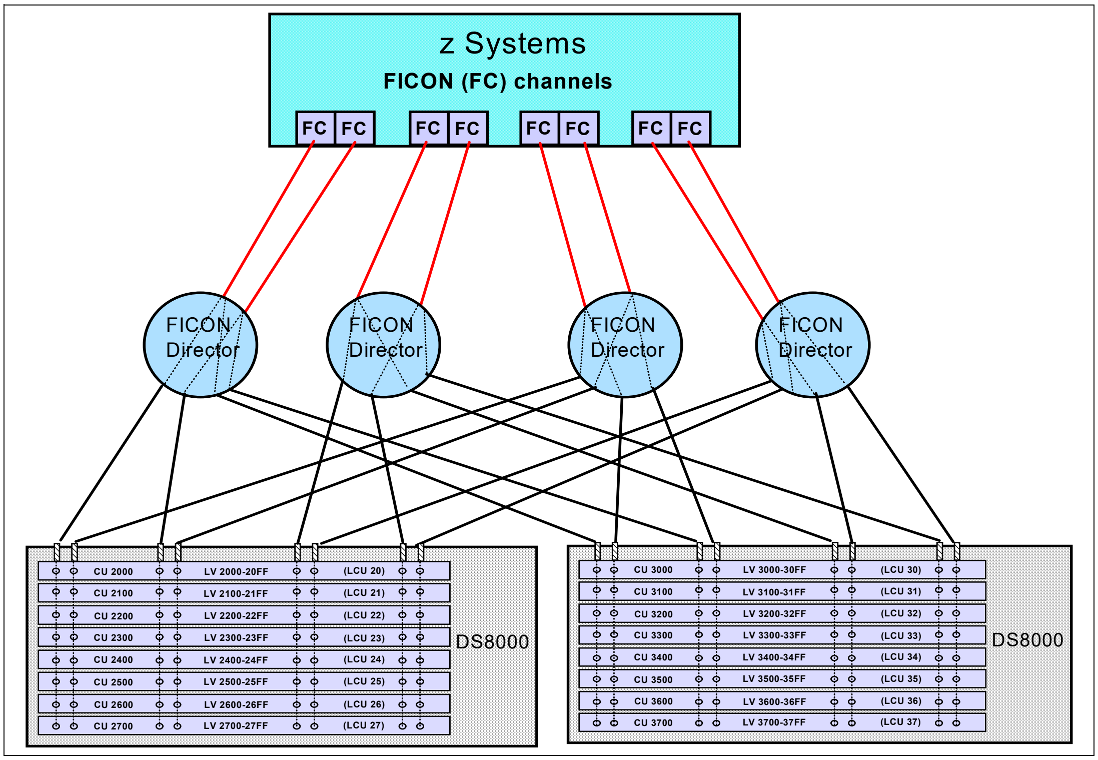

z Systems
FICON (FC) channels
FC FC FC FC FC FC FC FC
FICON Director
FICON Director
FICON Director
FICON Director
CU 2000 LV 2000-20FF (LCU 20)
CU 2100 LV 2100-21FF (LCU 21)
CU 2200 LV 2200-22FF (LCU 22)
CU 2300 LV 2300-23FF (LCU 23)
CU 2400 LV 2400-24FF (LCU 24)
CU 2500 LV 2500-25FF (LCU 25)
CU 2600 LV 2600-26FF (LCU 26)
CU 2700 LV 2700-27FF (LCU 27)
DS8000
CU 3000 LV 3000-30FF (LCU 30)
CU 3100 LV 3100-31FF (LCU 31)
CU 3200 LV 3200-32FF (LCU 32)
CU 3300 LV 3300-33FF (LCU 33)
CU 3400 LV 3400-34FF (LCU 34)
CU 3500 LV 3500-35FF (LCU 35)
CU 3600 LV 3600-36FF (LCU 36)
CU 3700 LV 3700-37FF (LCU 37)
DS8000

Figure 8-8 One-to-many FICON attachment

Sizing FICON connectivity is not an easy task. You must consider many factors. As a preferred practice, create a detailed analysis of the specific environment. Use these guidelines before you begin sizing the attachment environment:

- For FICON Express CHPID utilization, the preferred maximum utilization level is 50%.
- For the FICON Bus busy utilization, the preferred maximum utilization level is 40%.
- For the FICON Express Link utilization with an estimated link throughput of 4 Gbps, 8 Gbps, or 16 Gbps, the preferred maximum utilization threshold level is 70%.

For more information about DS8000 FICON support, see *IBM System Storage DS8000 Host Systems Attachment Guide*, SC26-7917, and *FICON Native Implementation and Reference Guide*, SG24-6266.

You can monitor the FICON channel utilization for each CHPID in the RMF Channel Path Activity report. For more information about the Channel Path Activity report, see Chapter 10, “Performance considerations for IBM z Systems servers” on page 225.

The following statements are some considerations and best practices for paths in z/OS systems to optimize performance and redundancy:

- Do not mix paths to one LCU with different link speeds in a path group on one z/OS. It does not matter in the following cases, even if those paths are on the same CPC:
  - The paths with different speeds, from one z/OS to different multiple LCUs
  - The paths with different speeds, from each z/OS to one LCU
- Place each path in a path group on different I/O bays.
- Do not have two paths from the same path group sharing a card.

8.3.3 z/VM, z/VSE, KVM, and Linux on z Systems attachment

z Systems FICON features in FCP mode provide full fabric and point-to-point attachments of Fixed Block (FB) devices to the operating system images. With this attachment, z/VM, z/VSE, KVM, and Linux on z Systems can access industry-standard FCP storage controllers and devices. This capability can facilitate the consolidation of UNIX server farms onto IBM Z servers and help protect investments in SCSI-based storage.

The FICON features support FC devices to z/VM, z/VSE, KVM, and Linux on z Systems, which means that these features can access industry-standard SCSI devices. These FCP storage devices use FB 512-byte sectors rather than Extended Count Key Data (IBM ECKD) format for disk applications. All available FICON features can be defined in FCP mode.

Linux FCP connectivity

You can use either direct or switched attachment to attach a storage unit to an IBM Z or LinuxONE host system that runs SUSE Linux Enterprise Server (SLES), or Red Hat Enterprise Linux (RHEL), or Ubuntu Linux, with current maintenance updates for FICON. For more information, refer to:

- http://www.ibm.com/systems/z/os/linux/resources/testedplatforms.html
- https://developer.ibm.com/technologies/linux

9

Performance considerations for the IBM i system

This chapter describes the topics that are related to the DS8900 performance with an IBM i host. The performance of IBM i database applications and batch jobs is sensitive to the disk response time. Therefore, it is important to understand how to plan, implement, and analyze the DS8900 performance with the IBM i system.

This chapter includes the following topics:

- IBM i storage architecture
- Fibre Channel adapters and Multipath
- Performance guidelines for drives in a DS8000 storage system with IBM i
- Analyzing performance data
- Easy Tier with the IBM i system

9.1 IBM i storage architecture

To understand the performance of the DS8900F storage system with the IBM i system, you need insight into IBM i storage architecture. This section explains this part of IBM i architecture and how it works with the DS8900F storage system.

The following IBM i specific features are important for the performance of external storage:

- Single-level storage
- Object-based architecture
- Storage management
- Types of storage pools

This section describes these features and explains how they relate to the performance of a connected DS8900F storage system.

9.1.1 Single-level storage

The IBM i system uses the same architectural component that is used by the iSeries and AS/400 platform: single-level storage. It treats the main memory and the flash unit space as one storage area. It uses the same set of 64-bit virtual addresses to cover both main memory and flash space. Paging in this virtual address space is performed in 4 KB memory pages.

9.1.2 Object-based architecture

One of the differences between the IBM i system and other operating systems is the concept of objects. For example, data files, programs, libraries, queues, user profiles, and device descriptions are all types of objects in the IBM i system. Every object on the IBM i system is packaged with the set of rules for how it can be used, enhancing integrity, security, and virus-resistance.

The IBM i system takes responsibility for managing the information in auxiliary disk pools. When you create an object, for example, a file, the system places the file in the best location that ensures the best performance. It normally spreads the data in the file across multiple flash units. Advantages of such design are ease of use, self-management, automation of using the added flash units, and so on. IBM i object-based architecture is shown on Figure 9-1 on page 207.

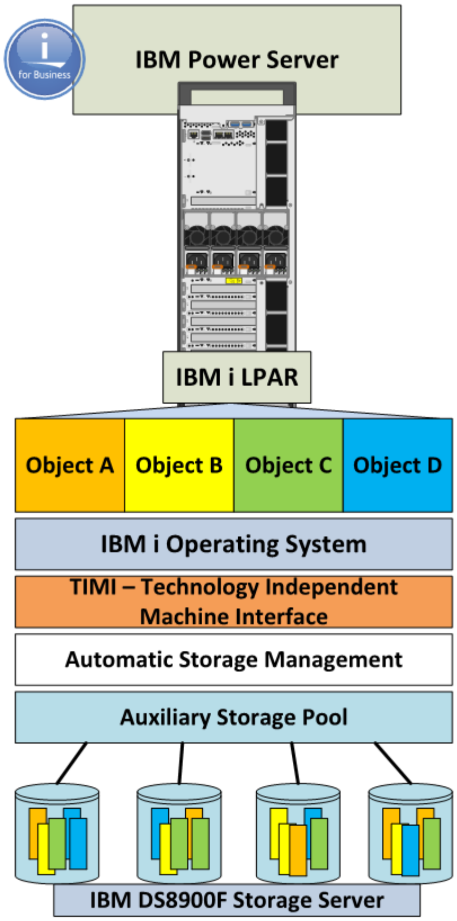


Figure 9-1 IBM i object-based architecture

9.1.3 Storage management

Storage management is a part of the IBM i Licensed Internal Code that manages the I/O operations to store and place data on storage. Storage management handles the I/O operations in the following way.

When the application performs an I/O operation, the portion of the program that contains read or write instructions is first brought into LPAR main memory where the instructions are then run.

With the read request, the virtual addresses of the needed record are resolved, and for each needed page, storage management first looks to see whether it is in LPAR main memory. If the page is there, it is used to resolve the read request. However, if the corresponding page is not in LPAR main memory, a page fault is encountered and it must be retrieved from the Auxiliary Storage Pool (ASP). When a page is retrieved, it replaces another page in LPAR main memory that recently was not used; the replaced page is paged out to the ASP, which resides inside the DS8900F storage server.

Similarly, writing a new record or updating an existing record is done in LPAR main memory, and the affected pages are marked as changed. A changed page normally remains in LPAR main memory until it is written to the ASP as a result of a page fault. Pages are also written to the ASP when a file is closed or when write-to-storage is forced by a user through commands and parameters. The handling of I/O operations is shown in Figure 9-2 on page 208.

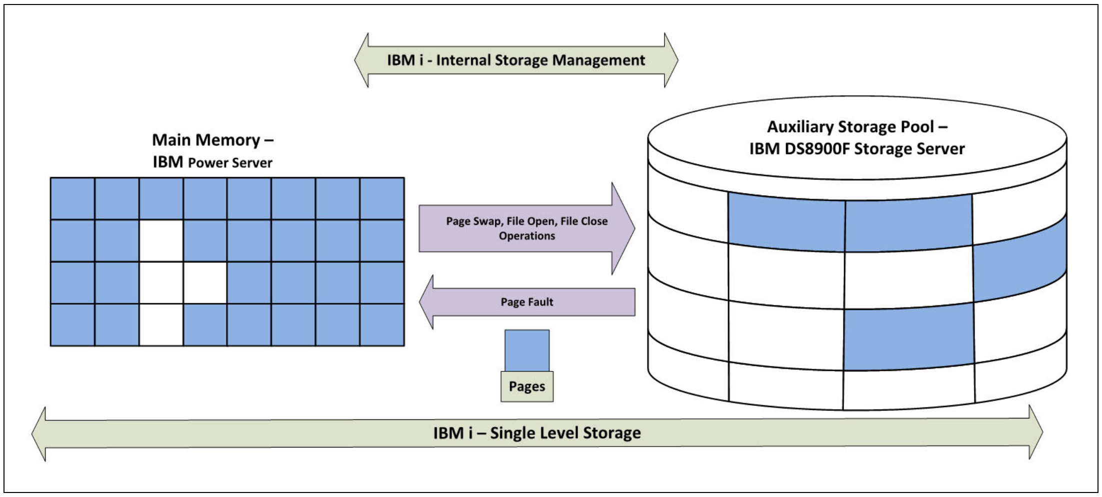

IBM i - Internal Storage Management
Main Memory –
IBM Power Server
Page Swap, File Open, File Close Operations
Page Fault
Pages
Auxiliary Storage Pool –
IBM DS8900F Storage Server
IBM i – Single Level Storage

Figure 9-2 Storage management handling I/O operations

When resolving virtual addresses for I/O operations, storage management directories map the flash and sector to a virtual address. For a read operation, a directory lookup is performed to get the needed information for mapping. For a write operation, the information is retrieved from the page tables.

9.1.4 Disk pools in the IBM i system

The flash pools in the IBM i system are referred to as Auxiliary Storage Pools (ASPs). The following types of disk pools exist in the IBM i system:

- System ASP
- User ASP
- Independent ASP (IASP)

System ASP

The system ASP is the basic flash pool for the IBM i system. This ASP contains the IBM i system boot flash (load source device), system libraries, indexes, user profiles, and other system objects. The system ASP is always present in the IBM i system and is needed for the IBM i system to operate. The IBM i system does not start if the system ASP is inaccessible.

User ASP

A user ASP separates the storage for different objects for easier management. For example, the libraries and database objects that belong to one application are in one user ASP, and the objects of another application are in a different user ASP. If user ASPs are defined in the IBM i system, they are needed for the IBM i system to start.

Independent ASP

The IASP is an independent flash pool that can switch among two or more IBM i systems that are in a cluster. The IBM i system can start without accessing the IASP. Typically, the objects that belong to a particular application are in this flash pool. If the IBM i system with IASP fails, the independent flash pool can be switched to another system in a cluster. If the IASP is on the DS8900F storage system, the copy of IASP (FlashCopy, Metro Mirror, or Global Mirror copy) is made available to another IBM i system and the cluster, and the application continues to work from another IBM i system.

9.2 Fibre Channel adapters and Multipath

This section explains the usage of Fibre Channel (FC) adapters to connect the DS8900F storage system to the IBM i system, with multiple ways to connect. It describes the performance capabilities of the adapters and the performance enhancement of using Multipath.

The DS8900F storage system can connect to the IBM i system in one of the following ways:

- Native: FC adapters in the IBM i system are connected through a Storage Area Network (SAN) to the Host Bus Adapters (HBAs) in the DS8900F storage system.
- With Virtual I/O Server Node Port ID Virtualization (VIOS NPIV): FC adapters in the VIOS are connected through a SAN to the HBAs in the DS8900F storage system. The IBM i system is a client of the VIOS and uses virtual FC adapters; each virtual FC adapter is mapped to a port in an FC adapter in the VIOS.

  For more information about connecting the DS8900F storage system to the IBM i system with VIOS_NPIV, see *DS8000 Copy Services for IBM i with VIOS*, REDP-4584, and *IBM System Storage DS8000: Host Attachment and Interoperability*, SG24-8887.
- With VIOS: FC adapters in the VIOS are connected through a SAN to the HBAs in the DS8900F storage system. The IBM i system is a client of the VIOS, and virtual SCSI adapters in VIOS are connected to the virtual SCSI adapters in the IBM i system.

  For more information about connecting storage systems to the IBM i system with the VIOS, see *IBM i and Midrange External Storage*, SG24-7668.

  Most installations use the native connection of the DS8900F storage system to the IBM i system or the connection with VIOS_NPIV.

> **IBM i I/O processors:** The information that is provided in this section refers to connection with IBM i I/O processor (IOP)-less adapters. For similar information about older IOP-based adapters, see *IBM i and IBM System Storage: A Guide to Implementing External Disks on IBM i*, SG24-7120.

9.2.1 FC adapters for native connection

The following FC adapters are used to connect the DS8900F storage system natively to an IBM i partition in a POWER server:

- 2-Port 16 Gb PCIe3 Generation-3 Fibre Channel Adapter, Feature Code EN0A (EN0B for Low Profile (LP) adapter)
- 4-Port 16Gb PCIe3 Generation-3 Fibre Channel Adapter, Feature Code EN1C (EN1D for LP adapter)
- 2-Port 16Gb PCIe3 Generation-3 Fibre Channel Adapter, Feature Code EN2A (EN2B for LP adapter)

- 2-Port 32 Gb PCIe3 Generation-3 Fibre Channel Adapter, Feature Code EN1A (EN1B for (LP) adapter)

> **Note:** The supported FC adapters listed above require DS8900F R9.1 or higher microcode and the IBM i 7.4TR4 operating system installed on Power 8/9 servers.
>
> For detailed specifications, see the IBM System Storage Interoperation Centre (SSIC) at:
>
> https://www.ibm.com/systems/support/storage/ssic/interoperability.wss

All listed adapters are IOP-less adapters. They do not require an I/O processor card to offload the data management. Instead, the processor manages the I/O and communicates directly with adapter. Thus, the IOP-less FC technology takes full advantage of the performance potential in the IBM i system.

Before the availability of IOP-less adapters, the DS8900F storage system connected to IOP-based FC adapters that required the I/O processor card.

IOP-less FC architecture enables two technology functions that are important for the performance of the DS8900F storage system with the IBM i system: Tag Command Queuing and Header Strip Merge.

Tagged Command Queuing

Tagged Command Queuing allows the IBM i system to issue multiple commands to the DS8900F storage system on the same path to a logical volume (LV). In the past, the IBM i system sent one command only per LUN path. Up to six I/O operations to the same LUN through one path are possible with Tag Command Queuing in the natively connected DS8900F storage system. With the natively connected DS8900F storage system, the queue depth on a LUN is 6.

Header Strip Merge

Header Strip Merge allows the IBM i system to bundle data into 4 KB chunks. By merging the data together, it reduces the amount of storage that is required for the management of smaller data chunks.

9.2.2 FC adapters in VIOS

The following Fibre Channel (FC) adapters are used to connect the DS8900F storage system to VIOS in a POWER server to implement VIOS_NPIV connection for the IBM i system:

- 2-Port 16 Gb PCIe3 Generation-3 FC Adapter, Feature Code EN0A (EN0B for LP adapter)
- 2-Port 16Gb PCIe3 Generation-3 FCI Adapter, Feature Code EN2A (EN2B for LP adapter)
- 4-Port 16 Gb PCIe3 Generation-3 FC Adapter, Feature Code EN1E (EN1F for LP adapter)
- 4-Port 16 Gb PCIe3 Generation-3 FC Adapter, Feature Code EN1C (EN1D for LP adapter
- 2-Port 32 Gb PCIe3 Generation-3 FC Adapter, Feature Code EN1A (EN1B for LP adapter
- 2-Port 32 Gb PCIe4 Generation-4 FC Adapter, Feature Code EN1J (EN1K for LP adapter

Queue depth and the number of command elements in the VIOS

When you connect the DS8900F storage system to the IBM i system through the VIOS, consider the following types of queue depths:

- The queue depth per LUN: SCSI Command Tag Queuing in the IBM i operating system enables up to 32 I/O operations to one LUN at the same time. This queue depth is valid for either connection with VIOS VSCSI or connection with VIOS_NPIV.
- If the DS8900F storage system is connected in VIOS VSCSI, consider the queue depth 32 per physical disk (hdisk) in the VIOS. This queue depth indicates the maximum number of I/O requests that can be outstanding on a physical disk in the VIOS at one time.
- The number of command elements per port in a physical adapter in the VIOS is 500 by default; you can increase it to 2048. This queue depth indicates the maximum number of I/O requests that can be outstanding on a port in a physical adapter in the VIOS at one time, either in a VIOS VSCSI connection or in VIOS_NPIV. The IBM i operating system has a fixed queue depth of 32, which is not changeable. However, the queue depth for a physical disk in the VIOS can be set up by a user. If needed, set the queue depth per physical disk in the VIOS to 32 by using the **`chdev -dev hdisk`*`xx`*` -attr queue_depth=32`** command.

9.2.3 Multipath

The IBM i system allows multiple connections from different ports on a single IBM i partition to the same LVs in the DS8900F storage system. This multiple connections support provides an extra level of availability and error recovery between the IBM i system and the DS8900F storage system. If one IBM i adapter fails, or one connection to the DS8900F storage system is lost, you can continue using the other connections and continue communicating with the disk unit. The IBM i system supports up to eight active connections (paths) to a single LUN in the DS8900F storage system.

IBM i multi-pathing is built into the IBM i System Licensed Internal Code (SLIC) and does not require a separate driver package to be installed.

In addition to high availability, multiple paths to the same LUN provide load balancing. A Round-Robin algorithm is used to select the path for sending the I/O requests. This algorithm enhances the performance of the IBM i system with DS8900F connected LUNs.

When the DS8900F storage system connects to the IBM i system through the VIOS, Multipath in the IBM i system is implemented so that each path to a LUN uses a different VIOS. Therefore, at least two VIOSs are required to implement Multipath for an IBM i client. This way of multipathing provides additional resiliency if one VIOS fails. In addition to IBM i Multipath with two or more VIOS, the FC adapters in each VIOS can multipath to the connected DS8900F storage system to provide additional resiliency and enhance performance.

9.3 Performance guidelines for drives in a DS8000 storage system with IBM i

This section describes the guidelines to use when planning and implementing flash drives in a DS8900F storage system for an IBM i system to achieve the required performance.

9.3.1 RAID level

The DS8900F supports three types of RAID configurations, RAID 6, RAID 10 or by Request for Price Quotation (RPQ) RAID 5.

The default and preferred RAID configuration for the DS8900F is now RAID 6. RAID 6 arrays provide better resiliency than RAID 5 with the use of two parity flash modules inside each array allowing an array to be rebuilt which has two failed flash modules.

Alternatively RAID 10 provides better resiliency and in some cases enables better performance than RAID 6. The difference in performance is because of the lower RAID penalty that is experienced with RAID 10 compared to RAID 6. The workloads with a low read/write ratio and with many random writes benefit the most from RAID 10.

Consider RAID 10 for IBM i systems especially for the following types of workloads:

- Workloads with large I/O rates
- Workloads with many write operations (low read/write ratio)
- Workloads with many random writes
- Workloads with low write-cache efficiency

9.3.2 Number of ranks

To better understand why the number of disk drives or the number of ranks is important for an IBM i workload, here is a short explanation of how an IBM i system spreads the I/O over flash modules.

When an IBM i page or a block of data is written to flash space, storage management spreads it over multiple flash modules. By spreading data over multiple flash modules, multiple flash arms work in parallel for any request to this piece of data, so writes and reads are faster.

When using external storage with the IBM i system, storage management sees an Logical Volume (LUN) in the DS8900F storage system as a “physical” flash module.

IBM i LUNs can be created with one of two Extent Allocation Methods (EAM) using the mkfbvol command:

- Rotate Volumes (rotatevols) EAM. This method occupies multiple stripes of a single rank.
- Rotate Extents (rotateexts) EAM. This method is composed of multiple stripes of different ranks.

Figure 9-3 on page 213 shows the use of the DS8900F disk with IBM i LUNs created with the rotatexts EAM.

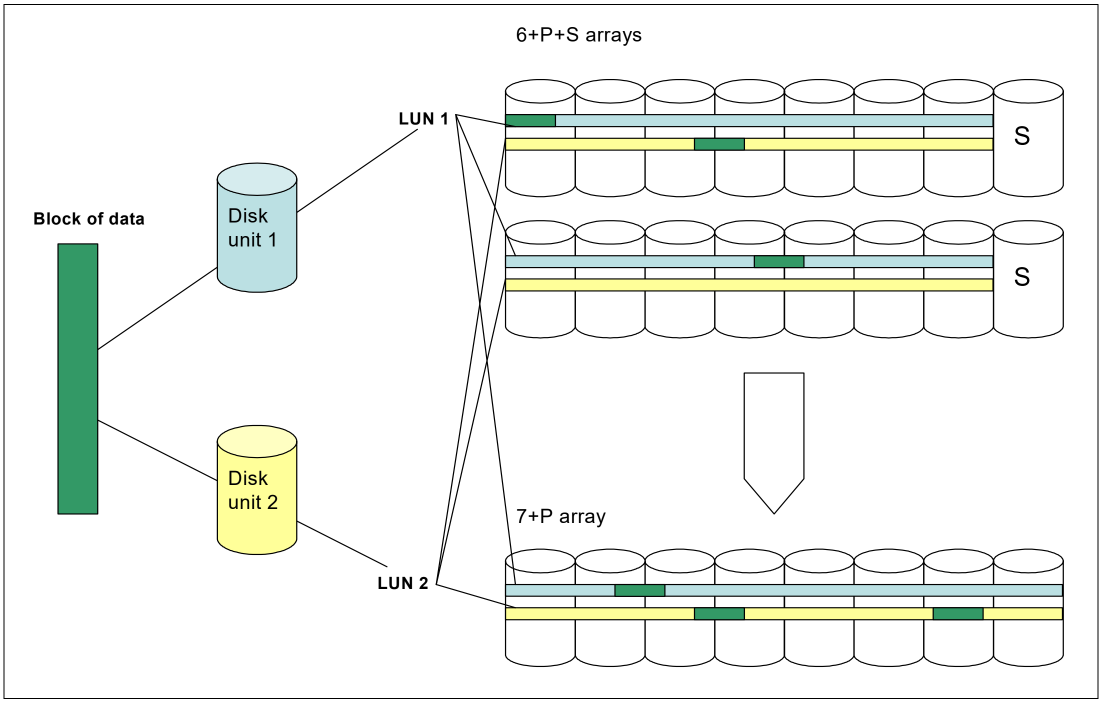


Figure 9-3 Use of disk arms with LUNs created in the rotate extents method

Therefore, a LUN uses multiple DS8900F flash arms in parallel. The same DS8900F flash arms are used by multiple LUNs that belong to the same IBM i workload, or even to different IBM i workloads. To support efficiently this structure of I/O and data spreading across LUNs and flash modules, it is important to provide enough disk arms to an IBM i workload.

Use the new StorM tool when planning the number of ranks in the DS8900F storage system for an IBM i workload.

To provide a good starting point for the StorM modeler, consider the number of ranks that is needed to keep disk utilization under 60% for your IBM i workload.

Table 9-1 shows the maximal number of IBM i I/O/sec for one rank to keep the disk utilization under 60%, for the workloads with read/write ratios 70/30 and 50/50.

Table 9-1 Host IO/sec for an DS8950F flash array[1]

RAID array, disk drive	Host I/O per second at 70% reads	Host I/O per second at 50% reads
RAID 6, 3.84 TB Flash Module	55,509	39,670
RAID 10, 3.84 TB Flash Module	84,629	68,157
RAID 6, 1.92 TB Flash Module	55,509	39,670

Use the following steps to calculate the necessary number of ranks for your workload by using Table 9-1:

1. Decide which read/write ratio (70/30 or 50/50) is appropriate for your workload.
2. Decide which RAID level to use for the workload.

[1] The calculations for the values in Table 9-1 are based on the measurements of how many I/O operations one rank can handle in a certain RAID level, assuming 20% read cache hit and 30% write cache efficiency for the IBM i workload. Assume that half of the used ranks have a spare and half are without a spare.

3. Look for the corresponding number in Table 9-1.
4. Divide the I/O/sec of your workload by the number from the table to get the number of ranks.

9.3.3 Number and size of LUNs

For the performance of an IBM i system, ensure that the IBM i system uses many flash modules. The more flash modules that are available to an IBM i system, the more server tasks are available for the IBM i storage management to use for managing the I/O operations to the ASP space. The result is improved I/O operational performance.

With IBM i internal disks can potentially be physical disk units or flash modules. With a connected DS8900F storage system, usable capacity created from flash modules is called a LUN. Therefore, it is important to provide many LUNs to an IBM i system.

> **Number of disk drives in the DS8900F storage system:** In addition to the suggestion for many LUNs, use a sufficient number of flash modules in the DS8900F storage system to achieve good IBM i performance, as described in 9.3.2, “Number of ranks” on page 212.

Since the introduction of the IBM i 7.2 TR7 and 7.3 TR3 operating systems up to 127 LUN’s are now supported per IBM i 16 Gb (or higher) physical or 8 Gb (or higher) virtual Fibre Channel (FC) adapter.

For IBM i operating systems before IBM i 7.2 TR7 and 7.3 TR3, the limit on the number of LUNs was 64 per FC adapter port.

Another reason why you should define smaller LUNs for an IBM i system is the queue depth in Tagged Command Queuing. With a natively connected DS8900F storage system, an IBM i system manages the queue depth of six concurrent I/O operations to a LUN. With the DS8900F storage system connected through VIOS, the queue depth for a LUN is 32 concurrent I/O operations. Both of these queue depths are modest numbers compared to other operating systems. Therefore, you must define sufficiently small LUNs for an IBM i system to not exceed the queue depth with I/O operations.

Also, by considering the manageability and limitations of external storage and an IBM i system, define LUN sizes of about 70.5 GB - 141 GB.

> **Note:** For optimal performance do not create LUNs of different sizes within the same ASP or IASP, choose a suitable LUN capacity from the outset and stick to it.
>
> Since the introduction of code bundle 87.10.xx.xx, two new IBM i volume data types were introduced to support variable volume sizes:
>
> - A50 an unprotected variable size volume.
> - A99 a protected variable size volume.
>
> Using these volume data types provide greater flexibility in choosing an optimum LUN size for your requirements.

9.3.4 DS8900F Ranks

A rank is a logical representation of the physical array formatted for use as FB or CKD storage types. In the DS8900F, ranks are defined in a one-to-one relationship to arrays.

Once the rank is formatted for IBM i data, the formatting determines the size of the set of data that is contained on one flash module within a stripe on the array. The capacity of the rank is subdivided into partitions, which are called extents.

An extent size can depend on the extent type and which type of operating system will be using them, in this case IBM i which is FB. The extents are the capacity building blocks of the LUNs and you can choose between large extents and small extents when creating the ranks during initial DS8900F configuration.

A Fixed Block (FB) rank can potentially have an extent size of either:

- Large Extent 1 GiB
- Small Extent size of 16 MiB

Small or Large extents

Small extents provide a better capacity utilization. However, managing many small extents causes some small performance degradation during initial allocation. For example, a format write of 1 GB requires one storage allocation with large extents, but 64 storage allocations with small extents. Otherwise, host performance should not be adversely affected.

To utilize more effective capacity utilization within the DS8900F the newly available SCSI Un-Map capability takes advantage of the small extents feature and is now the recommended extent type for IBM i storage implementations.

9.3.5 DS8900F extent pools for IBM i workloads

This section describes how to create extent pools and LUNs for an IBM i system and dedicate or share ranks for IBM i workloads.

Number of extent pools

Create two extent pools for an IBM i workload with each pool in one rank group.

Rotate volumes or rotate extents EAMs for defining IBM i LUNs

IBM i storage management spreads each block of data across multiple LUNs. Therefore, even if the LUNs are created with the rotate volumes EAM, performing an IBM i I/O operation uses the flash modules of multiple ranks.

You might think that the rotate volumes EAM for creating IBM i LUNs provides sufficient flash modules for I/O operations and that the use of the rotate extents EAM is “over-virtualizing”. However, based on the performance measurements and preferred practices, the rotate extents EAM of defining LUNs for an IBM i system still provides the preferred performance, so use it.

Dedicating or sharing the ranks for IBM i workloads

When multiple IBM i logical partitions (LPARs) use disk space on the DS8900F storage system, there is always a question whether to dedicate ranks (extent pools) to each of them or to share ranks among the IBM i systems.

Sharing the ranks among the IBM i systems enables the efficient use of the DS8900F resources. However, the performance of each LPAR is influenced by the workloads in the other LPARs.

For example, two extent pools are shared among IBM i LPARs A, B, and C. LPAR A experiences a long peak with large block sizes that causes a high I/O load on the DS8900F ranks. During that time, the performance of B and the performance of C decrease. But, when the workload in A is low, B and C experience good response times because they can use most of the disk arms in the shared extent pool. In these periods, the response times in B and C are possibly better than if they use dedicated ranks.

You cannot predict when the peaks in each LPAR happen, so you cannot predict how the performance in the other LPARs is influenced.

Many IBM i data centers successfully share the ranks with little unpredictable performance because the flash modules and cache in the DS8900F storage system are used more efficiently this way.

Other IBM i data centers prefer the stable and predictable performance of each system even at the cost of more DS8900F resources. These data centers dedicate extent pools to each of the IBM i LPARs.

Many IBM i installations have one or two LPARs with important workloads and several smaller, less important LPARs. These data centers dedicate ranks to the large systems and share the ranks among the smaller ones.

9.3.6 Number of ports in an IBM i system

An important aspect of DS8900F to IBM i SAN configuration is the number of paths to assign to a volume.

Recommendations as to how many fiber ports to use can be found in the following publication. You should take into account that a fiber port should not exceed 70% utilization when running peak workload. An IBM i client partition supports up to eight multi-path connections to a single flash module based LUN. Recommendations can be found in the following publication:

- *Limitations and Restrictions for IBM i Client Logical Partitions*:

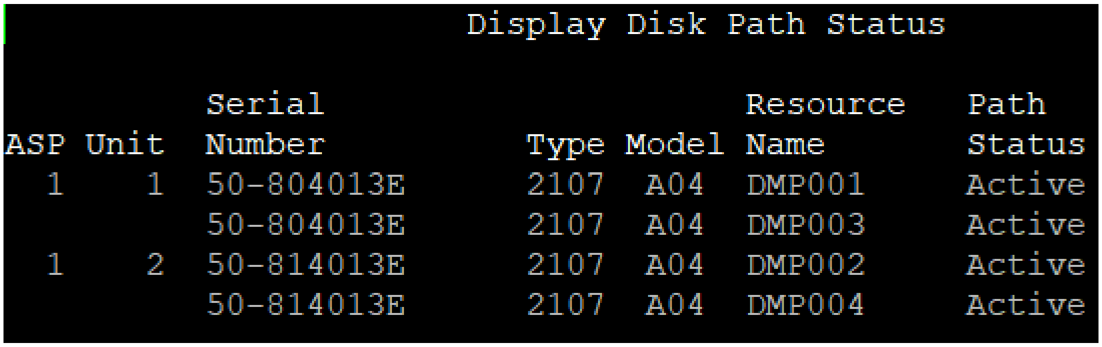
  https://www.ibm.com/docs/en/power9?topic=planning-i-restrictions

When implementing a multi-pathing design for the IBM i and DS8900F, plan to zone one fiber port in an IBM i system with one fiber port in the DS8900F storage system when running the workload on flash modules.

Using the Service Tools Function within the IBM i partition, the number of paths to a LUN can be seen:

1. On IBM i LPAR command Line type STRSST
2. Take Option 3 - Work With Disk Units
3. Take Option 2 - Work With Disk Configuration
4. Take Option 1 - Display Disk Configuration
5. Take Option 9 - Display Disk Path Status

The following is an extract of a typical IBM i LPAR.



```
                       Display Disk Path Status

          Serial                           Resource    Path
ASP Unit  Number         Type Model  Name          Status
  1    1  50-804013E     2107  A04   DMP001        Active
          50-804013E     2107  A04   DMP003        Active
  1    2  50-814013E     2107  A04   DMP002        Active
          50-814013E     2107  A04   DMP004        Active
```

Two disk units are shown (1 & 2) with two paths allocated to each. For a single path disk the Resource Name column would show DDxxx. Instead as each disk has more than one path available the Resource Name begins with (Disk Multi Pathing) DMP001 and DMP003 for Disk Unit 1.

The IBM i multi-pathing driver supports up to 8 active paths (+8 standby paths for HyperSwap) configurations to the DS8000 storage systems.

Since the introduction of operating system level IBM i 7.1 TR2 the multi-pathing driver uses an advanced load balancing algorithm which accounts for path usage by the amount of outstanding I/O per path.

From a performance perspective the use of more than 2 or 4 active paths with DS8000 storage systems is typically not required.

9.4 Analyzing performance data

For performance issues with IBM i workloads that run on the DS8900F storage system, you must determine and use the most appropriate performance tools in an IBM i system and in the DS8900F storage system.

9.4.1 IBM i performance tools

This section presents the performance tools that are used for an IBM i system. It also indicates which of the tools we employ with the planning and implementation of the DS8900F storage system for an IBM i system in this example.

To help you better understand the tool functions, they are divided into two groups: performance data collectors (the tools that collect performance data) and performance data investigators (the tools to analyze the collected data).

The following tools are the IBM i performance data collectors:

- Collection Services
- IBM i Job Watcher
- IBM i Disk Watcher
- Performance Explorer (PEX)

Collectors can be managed by IBM System Director Navigator for i, IBM System i Navigator, or IBM i commands.

The following tools are or contain the IBM i performance data investigators:

- IBM Performance Tools for i
- IBM System Director Navigator for i
- iDoctor

Most of these comprehensive planning tools address the entire spectrum of workload performance on System i, including processor, system memory, disks, and adapters. To plan or analyze performance for the DS8900F storage system with an IBM i system, use the parts of the tools or their reports that show the disk performance.

Collection Services

Collection Services for IBM i can collect system and job level performance data. It can run all the time and will sample system and job level performance data with collection intervals as low as 15 seconds or alternatively up to an hour. It provides data for performance health checks or analysis of a sudden performance problem. For detailed documentation, refer to:

https://www.ibm.com/docs/en/i/7.4?topic=collectors-collection-services

Collection Services can look at jobs, threads, processor, disk, and communications. It also has a set of specific statistics for the DS8900F storage system. For example, it shows which IBM i storage units are located within the DS8900F LUNs, whether they are connected in a single path or multipath, the disk service time, and wait time.

The following tools can be used to manage the data collection and report creation of Collection Services:

- IBM System i Navigator
- IBM System Director navigator
- IBM Performance Tools for i

iDoctor Collection Service Investigator can be used to create graphs and reports based on Collection Services data. For more information about iDoctor, see the IBM i iDoctor online documentation at:

`https://www.ibm.com/it-infrastructure/services/lab-services/idoctorl`

With IBM i level V7R4, the Collection Services tool offers additional data collection categories, including a category for external storage. This category supports the collection of nonstandard data that is associated with certain external storage subsystems that are attached to an IBM i partition. This data can be viewed within iDoctor, which is described in “iDoctor” on page 220.

Job Watcher

Job Watcher is an advanced tool for collecting and analyzing performance information to help you effectively monitor your system or to analyze a performance issue. It is job-centric and thread-centric and can collect data at intervals of seconds. To use IBM i Job Watcher functions and content require the installation of IBM Performance Tools for i (5770-PT1) Option 3 - Job Watcher.

For more information about Job Watcher, refer to:

https://www.ibm.com/docs/en/i/7.4?topic=interface-managing-i-job-watcher

Disk Watcher

Disk Watcher is a function of an IBM i system that provides disk data to help identify the source of disk-related performance problems on the IBM i platform. Its functions require the installation of IBM Performance Tools for i (5770-PT1) Option 1 Manager Feature.

For documented command strings and file layouts, refer to:

https://www.ibm.com/docs/en/i/7.4?topic=interface-managing-i-disk-watcher

Disk Watcher gathers detailed information that is associated with I/O operations to flash modules and provides data beyond the data that is available in other IBM i integrated tools, such as Work with Disk Status (**WRKDSKSTS**), Work with System Status (**WRKSYSSTS**), and Work with System Activity (**WKSYSACT**).

Performance Explorer

Performance Explorer is a data collection tool that helps the user identify the causes of performance problems that cannot be identified by collecting data using Collection Services or by doing general trend analysis.

Two reasons to use Performance Explorer include:

- Isolating performance problems to the system resource, application, program, procedure, or method that is causing the problem
- Analyzing the performance of applications.

Find more details on the Performance Explorer here:

https://www.ibm.com/docs/en/i/7.4?topic=collectors-performance-explorer

IBM Performance Tools for i

IBM Performance Tools for i is a licensed program product that includes features that provide further detailed analysis over the basic performance tools that are available in the V7.4 operating system. This software is described in more detail here:

https://www.ibm.com/docs/en/i/7.4?topic=data-performance-tools-i

Performance Tools helps you gain insight into IBM i performance features, such as dynamic tuning, expert cache, job priorities, activity levels, and pool sizes. You can also identify ways to use these services better. The tool also provides analysis of collected performance data and produces conclusions and recommendations to improve system performance.

The Job Watcher part of Performance Tools analyzes the Job Watcher data through the IBM Systems Director Navigator for i Performance Data Visualizer.

Collection Services reports about disk utilization and activity, which are created with IBM Performance Tools for i, are used for sizing and Disk Magic modeling of the DS8900F storage system for the IBM i system:

- The Disk Utilization section of the System report
- The Disk Utilization section of the Resource report
- The Disk Activity section of the Component report

IBM Navigator for i

The IBM Navigator for i is a web-based console that provides a single, easy-to-use view of the IBM i system. IBM Navigator for i provides a strategic tool for managing a specific IBM i partition.

The Performance section of IBM Systems Director Navigator for i provides tasks to manage the collection of performance data and view the collections to investigate potential performance issues. Figure 9-4 on page 220 shows the menu of performance functions in the IBM Systems Director Navigator for i.

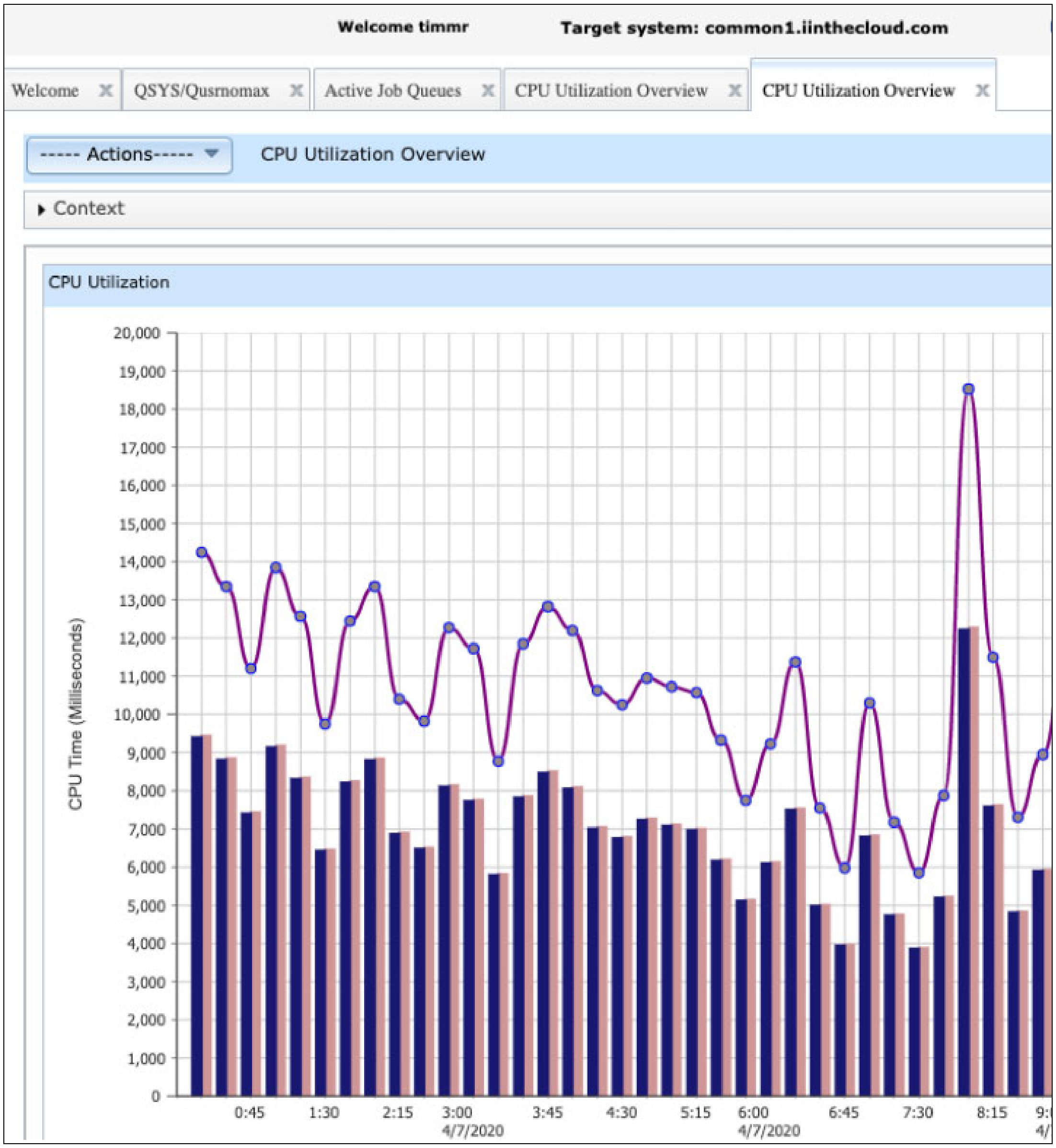


Figure 9-4 Performance tools of Navigator for i

iDoctor

iDoctor is a suite of tools that is used to manage the collection of data, investigate performance data, and analyze performance data on the IBM i system. The goals of iDoctor are to broaden the user base for performance investigation, simplify and automate processes of collecting and investigating the performance data, provide immediate access to collected data, and offer more analysis options.

The iDoctor tools are used to monitor the overall system health at a high level or to drill down to the performance details within jobs, flash modules and programs. Use iDoctor to analyze data that is collected during performance situations. iDoctor is frequently used by IBM, clients, and consultants to help solve complex performance issues quickly. Further information about these tools can be found at:

https://www.ibm.com/it-infrastructure/services/lab-services/idoctor

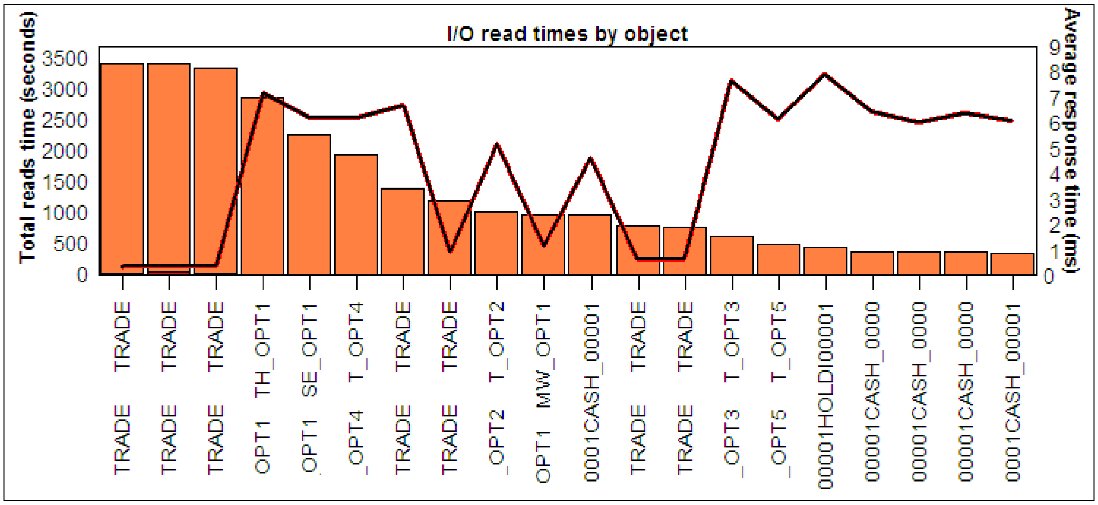


Figure 9-5 iDoctor - query of I/O read times by object

9.4.2 DS8900F performance tools

As a preferred practice, use Storage Insights web portal for analyzing the DS8900F performance data for an IBM i workload. For more information about this product, see Chapter 7, “Practical performance management” on page 139.

9.5 Easy Tier with the IBM i system

This section describes how to use Easy Tier with the IBM i system.

9.5.1 Hot data in an IBM i workload

An important feature of the IBM i system is object-based architecture. Everything on the system that can be worked with is considered an object. An IBM i library is an object. A database table, an index file, a temporary space, a job queue, and a user profile are objects. The intensity of I/Os is split by objects. The I/O rates are high on busy objects, such as application database files and index files. The I/Os rates are lower on user profiles.

IBM i Storage Manager spreads the IBM i data across the available flash modules (LUNs) contained within an extent pool so that each flash module is about equally occupied. The data is spread in extents that are 4 KB - 1 MB or even 16 MB. The extents of each object usually span as many LUNs as possible to provide many volumes to serve the particular object. Therefore, if an object experiences a high I/O rate, this rate is evenly split among the LUNs. The extents that belong to the particular object on each LUN are I/O-intense.

Many of the IBM i performance tools work on the object level; they show different types of read and write rates on each object and disk service times on the objects. For more information about the IBM i performance tools, see 9.4.1, “IBM i performance tools” on page 217. You can relocate hot data by objects by using the Media preference method, which is described in “IBM i Media preference” on page 222.

Also, the Easy Tier tool monitors and relocates data on the 1 GB extent level. IBM i ASP balancing, which is used to relocate data to Flash, works on the 1 MB extent level. Monitoring extents and relocating extents do not depend on the object to which the extents belong; they occur on the subobject level.

9.5.2 IBM i methods for hot-spot management

You can choose your tools for monitoring and moving data to faster drives in the DS8900F storage system. Because the IBM i system recognizes the LUNs on SSDs in a natively connected DS8900F storage system, you can use the IBM i tools for monitoring and relocating hot data. The following IBM i methods are available:

- IBM i Media preference
- ASP balancing
- A process where you create a separate ASP with LUNs on SSDs and restore the application to run in the ASP

IBM i Media preference

This method provides monitoring capability and the relocation of the data on the object level. You are in control. You decide the criteria for which objects are hot and you control which objects to relocate to the Flash Module. Some clients prefer Media preference to Easy Tier or ASP balancing. IBM i Media preference involves the following steps:

1. Use PEX to collect disk events

   Carefully decide which disk events to trace. In certain occasions, you might collect read flash operations only, which might benefit the most from improved performance, or you might collect both flash read and write information for future decisions. Carefully select the peak period in which the PEX data is collected.

2. Examine the collected PEX data by using the PEX-Analyzer tool of iDoctor or by using the user-written queries to the PEX collection.

   It is a good idea to examine the accumulated read service time and the read I/O rate on specific objects. The objects with the highest accumulated read service time and the highest read I/O rate can be selected to relocate to Flash Modules. It is also helpful to analyze the write operations on particular objects. You might decide to relocate the objects with high read I/O and rather modest write I/O to flash benefit from lower service times and wait times. Sometimes, you must distinguish among the types of read and write operations on the objects. iDoctor queries provide the rates of asynchronous and synchronous database reads and page faults.

   In certain cases, queries must be created to run on the PEX collection to provide specific information, for example, the query that provides information about which jobs and threads use the objects with the highest read service time. You might also need to run a query to provide the block sizes of the read operations because you expect that the reads with smaller block sizes profit the most from flash. If these queries are needed, contact IBM Lab Services to create them.

3. Based on the PEX analysis, decide which database objects to relocate to Flash Modules in the DS8900F storage system. Then, use IBM i commands such as Change Physical File (**`CHGPF`**) with the **`UNIT(*SSD)`** parameter, or use the SQL command **`ALTER TABLE UNIT SSD`**, which sets on the file a preferred media attribute that starts dynamic data movement. The preferred media attribute can be set on database tables and indexes, and on User-Defined File Systems (UDFS).

   For more information about the UDFS, refer to:

   http://www.ibm.com/support/knowledgecenter/ssw_ibm_i/welcome

ASP balancing

This IBM i method is similar to DS8900F Easy Tier because it is based on the data movement within an ASP by IBM i ASP balancing. The ASP balancing function is designed to improve

IBM i system performance by balancing disk utilization across all of the disk units (or LUNs) in an ASP. It provides three ways to balance an ASP:

- Hierarchical Storage Management (HSM) balancing
- Media Preference Balancing
- ASP Balancer Migration Priority

The HSM balancer function, which traditionally supports data migration between high-performance and low-performance internal disk drives, is extended for the support of data migration between Flash Modules and HDDs. The flash drives can be internal or on the DS8900F storage system. The data movement is based on the weighted read I/O count statistics for each 1 MB extent of an ASP. Data monitoring and relocation is achieved by the following two steps:

1. Run the ASP balancer tracing function during the important period by using the **`TRCASPBAL`** command. This function collects the relevant data statistics.
2. By using the **`STRASPBAL TYPE(*HSM)`** command, you move the data to Flash and HDD based on the statistics that you collected in the previous step.

The Media preference balancer function is the ASP balancing function that helps to correct any issues with Media preference-flagged database objects or UDFS files not on their preferred media type, which is either Flash or HDD, based on the specified subtype parameter.

The function is started by the **`STRASPBAL TYPE(*MP)`** command with the **`SUBTYPE`** parameter equal to either `*CALC` (for data migration to both Flash (SSD) and HDD), `*SSD`, or `*HDD`.

The ASP balancer migration priority is an option in the ASP balancer so that you can specify the migration priority for certain balancing operations, including *HSM or *MP in levels of either *LOW, *MEDIUM, or *HIGH, thus influencing the speed of data migration.

> **Location:** For data relocation with Media preference or ASP balancing, the LUNs defined on Flash and on HDD must be in the same IBM i ASP. It is not necessary that they are in the same extent pool in the DS8900F storage system.

Additional information

For more information about the IBM i methods for hot-spot management, including the information about IBM i prerequisites, refer to:

https://www.ibm.com/support/pages/ibm-i-technology-refresh-information-ibm-i-74

10

Performance considerations for IBM z Systems servers

This chapter covers some z Systems specific performance topics. First, it explains how to collect and interpret z Systems I/O performance data from the Resource Measurement Facility (RMF). Next, it provides a section with hints, tips, and best practices to avoid potential bottlenecks up front. The last part describes a strategy to isolate bottlenecks and develop strategies to prevent or eliminate them.

> **A note about DS8000 sizing:** z Systems I/O workload is complex. Use RMF data, StorM and Disk Magic models for sizing, where appropriate.
>
> For more information about StorM with IBM Z see "Modeling and sizing for IBM Z" on page 132. For more information about the sizing tools, see Chapter 6.1, "IBM Storage Modeller" on page 126.
>
> **Disk Magic for IBM i**s the legacy as-is performance modeling and hardware configuration planning tool for legacy IBM data storage solutions and is not updated for product currency effective January 1, 2020. Users will continue to have access to the Magic tools to support legacy product sizings.

IBM z Systems® and the IBM System Storage DS8000 storage systems family have a long common history. Numerous features were added to the whole stack of server and storage hardware, firmware, operating systems, and applications to improve I/O performance. This level of synergy is unique to the market and is possible only because IBM is the owner of the complete stack. This chapter does not describe these features because they are explained in other places. For an overview and description of IBM Z Systems synergy features, see *IBM DS8000 and IBM Z Synergy*, REDP-5186.

This chapter includes the following topics:

- DS8000 performance monitoring with RMF
- DS8000 and z Systems planning and configuration
- Problem determination and resolution

10.1 DS8000 performance monitoring with RMF

This section describes how to gather and interpret disk I/O related performance data from Resource Management Facility (RMF), which is part of the z/OS operating system. It provides performance information for the DS8000 storage systemand other storage systems. RMF can help with monitoring the following performance components:

- I/O response time
- IOP/SAP
- FICON host channel
- FICON director
- zHyperLink
- Symmetric multiprocessing (SMP)
- Cache and nonvolatile storage (NVS)
- Enterprise Storage Servers
  - FICON/Fibre port and host adapter (HA)
  - Extent pool and rank/array

Sample performance measurement data: Any sample performance measurement data that is provided in this chapter is for explanatory purposes only. It does not represent the capabilities of a real system. The data was collected in controlled laboratory environment at a specific point by using a specific configuration with hardware and firmware levels available then. Performance in real-world environments will be different.

Contact your IBM representative or IBM Business Partner if you have questions about the expected performance capability of IBM products in your environment.

10.1.1 RMF Overview

Figure 10-1 shows an overview of the RMF infrastructure.

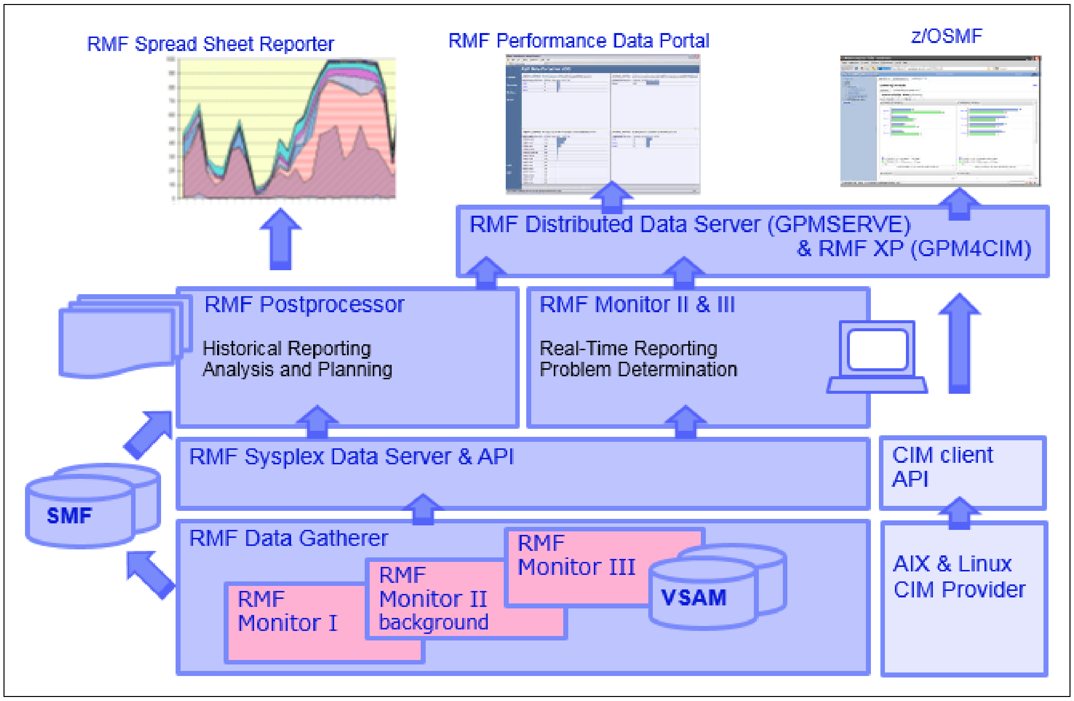


Figure 10-1 RMF overview

RMF gathers data by using three monitors:

- Monitor I: Long-term data collector for all types of resources and workloads. The SMF data collected by Monitor I is mainly used for capacity planning and performance analysis.
- Monitor II: Snapshot data collector for address space states and resource usage. A subset of Monitor II data is also displayed by the IBM z/OS System Display and Search Facility (SDSF) product.
- Monitor III: Short-term data collector for problem determination, workflow delay monitoring, and goal attainment supervision. This data is also used by the RMF PM Java Client and the RMF Monitor III Data Portal.

The data is then stored and processed in several ways. The ones that are described and used in this chapter are:

- Data that the three monitors collect can be stored as SMF records (SMF types 70 - 79) for later reporting.
- RMF Monitor III can write VSAM records to in-storage buffer or into VSAM data sets.
- The RMF Postprocessor is the function to extract historical reports for Monitor I data.

Other methods of working with RMF data, which are not described in this chapter, are:

- The RMF Spread Sheet Reporter provides a graphical presentation of long-term Postprocessor data. It helps you view and analyze performance at a glance or for system health checks.
- The RMF Distributed Data Server (DDS) supports HTTP requests to retrieve RMF Postprocessor data. The data is returned as an XML document so that a web browser can act as Data Portal to RMF data.
- z/OS Management Facility provides the presentation for DDS data.
- RMF XP enables Cross Platform Performance Monitoring for installation running operating systems other than z/OS, namely: AIX on Power, Linux on System x, Linux on z Systems.

The following sections describe how to gather and store RMF Monitor I data and then extract it as reports using the RMF Postprocessor.

RMF Monitor I gatherer session options and write SMF record types

To specify which types of data RMF is collecting, you specify Monitor I session gatherer options in the ERBRMFxx parmlib member.

Table 10-1 shows the Monitor I session options and associated SMF record types that are related to monitoring I/O performance. The defaults are emphasized.

Table 10-1 Monitor I gatherer session options and write SMF record types

Activities	**Session options in ERBRMFxx parmlib member**	**SMF record types (Long-term Monitor I)**
Direct Access Device Activity	*DEVICE* / NODEVICE	74.1
I/O queuing Activity	*IOQ* / NOIOQ	78.3
Channel Path Activity	*CHAN* / NOCHAN	73
FICON Director Activity (H)	FCD / *NOFCD*	74.7

Activities	Session options in ERBRMFxx parmlib member	SMF record types (Long-term Monitor I)
Cache Subsystem Activity (H)	*CACHE* / NOCACHE	74.5
Enterprise Storage Server (link and rank statistics) (H)	ESS / *NOESS*	74.8

Note: The Enterprise Storage Server activity is not collected by default. Change this and turn on the collection if you have DS8000 storage systems installed. It provides valuable information about DS8000 internal resources.

Note: FCD performance data is collected from the FICON directors. You must have the FICON Control Unit Port (CUP) feature licensed and installed on your directors.

Certain measurements are performed by the storage hardware. The associated RMF record types are 74.5, 74.7, and 74.8. They are marked with an (H) in Table 10-1 . They do not have to be collected by each attached z/OS system separately; it is sufficient to get them from one or, for redundancy reasons, two systems.

Note: Many customers, who have several z/OS systems sharing disk systems, typically collect these records from two production systems that are always up and not running at the same time.

In the ERBRMFxx parmlib member, you also find a TIMING section, where you can set the RMF sampling cycle. It defaults to 1 second, which should be good for most cases. The RMF cycle does not determine the amount of data that is stored in SMF records.

To store the collected RMF data, you must make sure that the associated SMF record types (70 - 78) are included in the `SYS` statement in the SMFPRMxx parmlib member.

You also must specify the interval at which RMF data is stored. You can either do this explicitly for RMF in the ERBRMFxx parmlib member or use the system wide SMF interval. Depending on the type of data, RMF samples are added up or averaged for each interval. The number of SMF record types you store and the interval make up the amount of data that is stored.

For more information about setting up RMF and the ERBRMFxx and SMFPRMxx parmlib members, see *z/OS Resource Measurement Facility User's Guide*, SC34-2664-40.

Important: The shorter your interval, the more accurate your data is. However, there is always a trade-off between shorter intervals and the size of the SMF data sets.

Preparing SMF records for post processing

The SMF writer routine stores SMF records to VSAM data sets that are named SYS1.MANx by default. Before you can run the postprocessor to create RMF reports, you must dump the SMF records into sequential data sets using the IFASMFDP program. Example 10-1 on page 229 shows a sample IFASMFDP job.

Example 10-1 Sample IFASMFDP job

```
//SMFDUMP EXEC PGM=IFASMFDP
//DUMPIN1  DD DSN=SYS1.MAN1,DISP=SHR
//DUMPOUT  DD DSN=hlq.SMFDUMP.D151027,DISP=(,CATLG),
//            SPACE=(CYL,(10,100),RLSE),VOL=SER=ABC123,UNIT=SYSALLDA
//SYSPRINT DD SYSOUT=*
//SYSIN DD *
   INDD(DUMPIN1,OPTIONS(DUMP))
   OUTDD(DUMPOUT,TYPE(70:79))
   DATE(2015301,2015301)
   START(1200)
   END(1230)
/*
```

IFASMFDP can also be used to extract and concatenate certain record types or time ranges from existing sequential SMF dump data sets. For more information about the invocation and control of ISASMFDP, see *z/OS MVS System Management Facilities (SMF)*, SA38-0667-40.

To create meaningful RMF reports or analysis, the records in the dump data set must be chronological. This is important if you plan to analyze RMF data from several LPARs. Use a SORT program to combine the individual data sets and sort them by date and time. Example 10-2 shows a sample job snippet with the required sort parameters by using the DFSORT program.

Example 10-2 Sample combine and sort job

```
//RMFSORT  EXEC  PGM=SORT
//SORTIN   DD    DISP=SHR,DSN=<input_smfdata_system1>
//         DD    DISP=SHR,DSN=<input_smfdata_system2>
//SORTOUT  DD    DSN=<output_cobined_sorted_smfdata>
//SYSIN    DD   *
  SORT FIELDS=(11,4,CH,A,7,4,CH,A,15,4,CH,A),EQUALS
  MODS E15=(ERBPPE15,36000,,N),E35=(ERBPPE35,3000,,N)
```

RMF postprocessor

The RMF postprocessor analyses and summarizes RMF data into human readable reports. Example 10-3 shows a sample job to run the ERBRMFPP post processing program.

Example 10-3 Sample ERBRMFPP job

```
//RMFPP    EXEC PGM=ERBRMFPP
//MFPINPUT DD  DISP=SHR,DSN=hlq.SMFDUMP.D151027
//MFPMSGDS DD SYSOUT=*
//SYSIN    DD  *
  NOSUMMARY
  DATE(10272015,10272015)
  REPORTS(CACHE,CHAN,DEVICE,ESS,IOQ,FCD)
  RTOD(1200,1230)
  SYSOUT(H)
/*
```

In the control statements, you specify the reports that you want to get by using the **REPORTS** keyword. Other control statements define the time frame, intervals, and summary points for the reports to create. For more information about the available control statements, see *z/OS Resource Measurement Facility User's Guide*, SC34-2664-40.

The following sections explain the reports that might be useful in analyzing I/O performance issues, and provide some basic rules to identify potential bottlenecks. For more information about all available RMF reports, see *z/OS Resource Measurement Facility Report Analysis*, SC34-2665-40.

> **Note:** You can also generate and start the postprocessor batch job from the ISPF menu of RMF.

10.1.2 Direct Access Device Activity report

The RMF Direct Access Device Activity report (see Example 10-4) is a good starting point in analyzing the storage systems' performance. It lists the activity of individual devices, and their average response time for the reporting period. The response time is also split up into several components.

To get a first impression, you can sort volumes by their I/O intensity, which is the I/O rate multiplied by Service Time (PEND + DISC + CONN component). Also, look for the largest component of the response time. Try to identify the bottleneck that causes this problem. Do not pay too much attention to volumes with low or no Device Activity Rate, even if they show high I/O response time. The following sections provide more detailed explanations.

The device activity report accounts for all activity to a base and all of its associated alias addresses. Activity on alias addresses is not reported separately, but accumulated into the base address.

The Parallel Access Volume (PAV) value is the number of addresses assigned to a unit control block (UCB), including the base address and the number of aliases assigned to that base address.

RMF reports the number of PAV addresses (or in RMF terms, exposures) that are used by a device. In a HyperPAV environment, the number of PAVs is shown in this format: *n.nH*. The *H* indicates that this volume is supported by HyperPAV. The *n.n* is a one decimal number that shows the average number of PAVs assigned to the address during the RMF report period. Example 10-4 shows that address 7010 has an average of 1.5 PAVs assigned to it during this RMF period. When a volume has no I/O activity, the PAV is always 1, which means that no alias is assigned to this base address. In HyperPAV an alias is used or assigned to a base address only during the period that is required to run an I/O. The alias is then released and put back into the alias pool after the I/O is completed.

> **Important:** The number of PAVs includes the base address plus the number of aliases assigned to it. Thus, a PAV=1 means that the base address has no aliases assigned to it.

Example 10-4 Direct Access Device Activity report

```
                                     D I R E C T   A C C E S S   D E V I C E   A C T I V I T Y

                                                DEVICE   AVG  AVG   AVG  AVG  AVG   AVG  AVG  AVG     %     %     %     AVG    %
STORAGE  DEV  DEVICE   NUMBER  VOLUME PAV  LCU  ACTIVITY RESP IOSQ  CMR  DB   INT   PEND DISC CONN    DEV   DEV   DEV   NUMBER ANY
 GROUP   NUM  TYPE     OF CYL  SERIAL           RATE     TIME TIME  DLY  DLY  DLY   TIME TIME TIME   CONN  UTIL  RESV  ALLOC  ALLOC
         7010 33909     10017  ST7010 1.5H 0114 689.000  1.43 .048  .046 .000       .163 .474 .741  33.19  54.41   0.0   2.0   100.0
         7011 33909     10017  ST7011 1.5H 0114 728.400  1.40 .092  .046 .000       .163 .521 .628  29.72  54.37   0.0   2.0   100.0
YGTST00  7100 33909     60102  YG7100 1.0H 003B   1.591  12.6 .000  .077 .000 .067  .163 12.0 .413   0.07   1.96   0.0  26.9   100.0
YGTST00  7101 33909     60102  YG7101 1.0H 003B   2.120  6.64 .000  .042 .000 .051  .135 6.27 .232   0.05   1.37   0.0  21.9   100.0
```

10.1.3 I/O response time components

The device activity report provides various components of the overall response time for an I/O operation:

- IOS Queuing (IOSQ) time: Queuing at the host level.
- Pending (PEND) time: Response time of the storage hardware (fabric and HA).
- Disconnect (DISC) time: Stage or de-stage data to or from cache, and remote replication impact.
- Connect (CONN) time: Data transfer time.
- I/O Interrupt delay (INT): Time between finishing an I/O operation and interrupt.

Figure 10-2 illustrates how these components relate to each other and to the common response and service time definitions.

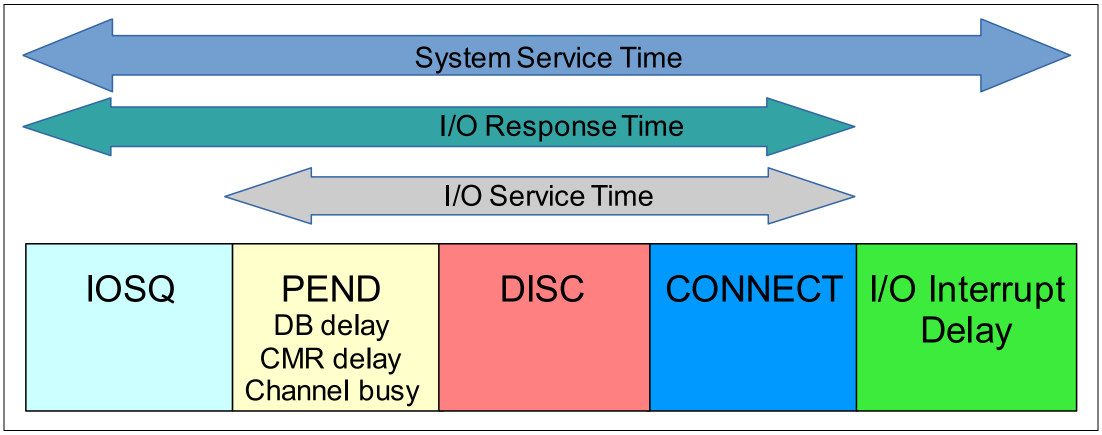


Figure 10-2 I/O interrupt delay time

Before learning about the individual response time components, you should know about the different service time definitions in simple terms:

- I/O service time is the time that is required to fulfill an I/O request after it is dispatched to the storage hardware. It includes locating and transferring the data and the required handshaking.
- I/O response time is the I/O service time plus the time the I/O request spends in the I/O queue of the host.
- System service time is the I/O response time plus the time it takes to notify the requesting application of the completion.

Only the I/O service time is directly related to the capabilities of the storage hardware. The additional components that make up I/O response or system service time are related to host system capabilities or configuration.

The following sections describe all response time components in more detail. They also provide possible causes for unusual values.

IOSQ time

IOSQ time is the time an I/O request spends on the host I/O queue after being issued by the operating system. During normal operation, IOSQ time should be zero, or at least only a fraction of the total response time. The following situations can cause high IOSQ time:

- In many cases, high IOSQ time is because of the unavailability of aliases to initiate an I/O request. Implementing HyperPAV/SuperPAV and adding/increasing alias addresses for the affected LCUs improves the situation.
- RESERVEs can be a cause of a shortage of aliases. When active, they can hold off much of the other I/O. After the reserve is released, there is a large burst of activity that uses up all the available aliases and results in increased IOSQ. Avoid RESERVEs altogether and change the I/O serialization to Global Resource Serialization (GRS).
- There is also a slight possibility that the IOSQ is caused by a long busy condition during device error recovery.
- If you see a high IOSQ time, also look at whether other response time components are higher than expected.

PEND time

PEND time represents the time that an I/O request waits in the hardware. It can become increased by the following conditions:

- High DS8000 HA utilization:
  - An HA can be saturated even if the individual ports have not yet reached their limits. HA utilization is not directly reported by RMF.
  - The Command response (CMR) delay, which is part of PEND, can be an indicator for high HA utilization. It represents the time that a Start- or Resume Subchannel function needs until the device accepts the first command. It should not exceed a few hundred microseconds.
  - The Enterprise Storage Server report can help you further to find the reason for increased PEND caused by a DS8000 HA. For more information, see “Enterprise Storage Server” on page 261.
- High FICON Director port utilization:
  - High FICON Director port or DS8000 HA port utilization can occur due to a number of factors:
    - Multiple FICON channels connecting to the same outbound switch port causing over commission on the port.

      In this case, the FICON channel utilization as seen from the host might be low, but the combination or sum of the utilization of these channels that share the outbound port can be significant. A one-to-one mapping between one FICON channel and one outbound switch port is recommended to avoid any over utilization of the outbound port issues.
    - Faulty Small form-factor plugable (SFP) transceivers at the DS8000 or Switch port end where the DS8000 connects.

      In this case, the SFP will undergo recovery to try and maintain a viable connection increasing the port utilization.
  - The FICON Director report can help you isolate the ports that cause increased PEND. For more information, see 10.1.6, “FICON Director Activity report” on page 237.
- Device Busy (DB) delay: Time that an I/O request waits because the device is busy. Today, mainly because of the Multiple Allegiance feature, DB delay is rare. But it can occur due to:

  - The device is RESERVED by another system. Use GRS to avoid hardware RESERVEs, as already indicated in “IOSQ time” on page 232.
  - If both DASD and Tape workloads are resident on the same switched infrastructure. Tape being a slower device compared to DASD, long running tape backup jobs can cause a tape to become a slow-drain device which, if left unchecked for an extended period of time, can cause port/switch buffer credits to exhaust, leading to contention across the switch and, in severe cases, devices becoming busy. Avoid mixing DASD and Tape workloads on the same switched infrastructure. If not achievable, then implement appropriate quality of service to separate DASD and Tape workloads within the switch.
- SAP impact: Indicates the time the I/O Processor (IOP/SAP) needs to handle the I/O request. For more information, see 10.1.4, “I/O Queuing Activity report” on page 234.

DISC time

DISC is the time that the storage system needs to process data internally. During this time, it disconnects from the channel to free it for other operations:

- The most frequent cause of high DISC time is waiting for data to be staged from the storage back end into cache because of a read cache miss. This time can be elongated because of the following conditions:
  - Low read hit ratio. For more information, see 10.1.7, “Cache and NVS report” on page 238. The lower the read hit ratio, the more read operations must wait for the data to be staged from the DDMs to the cache. Adding/Increasing cache to the DS8000 storage system can increase the read hit ratio.
  - High rank utilization. You can verify this condition with the Enterprise Storage Server Rank Statistics report, as described in “Enterprise Storage Server” on page 261.
- Heavy write workloads sometimes cause an NVS full condition. Persistent memory full condition or NVS full condition can also elongate the DISC time. For more information, see 10.1.7, “Cache and NVS report” on page 238.
- In a Metro Mirror environment, the time needed to send and store the data to the secondary volume also adds to the DISCONNECT component.

CONN time

For each I/O operation, the channel subsystem measures the time that storage system, channel, and CPC are connected for data transmission. CONN depends primarily on the amount of data that is transferred per I/O. Large I/Os naturally have a higher CONN component than small ones.

If there is a high level of utilization of resources, time can be spent in contention rather than transferring data. Several reasons exist for higher than expected CONN time:

- FICON channel saturation. CONN time increases if the channel or BUS utilization is high. FICON data is transmitted in frames. When multiple I/Os share a channel, frames from an I/O are interleaved with those from other I/Os. This elongates the time that it takes to transfer all of the frames of that I/O. The total of this time, including the transmission time of the interleaved frames, is counted as CONN time. For details and thresholds, see “FICON channels” on page 247.
- Contention in the FICON Director. A DS8000 HA port can also affect CONN time, although they primarily affect PEND time.

I/O interrupt delay time

The interrupt delay time represents the time from when the I/O completes until the operating system issues TSCH to bring in the status and process the interrupt. It includes the time that it takes for the Hypervisor (IBM PR/SM) to dispatch the LPAR on an available processor. Possible examples of why the time might be high include the **CPENABLE** setting, processor weights for LPAR dispatching, MSU Capping (Hard/Soft) in effect, and so on. It usually is lower than 0.1 ms.

Tip: A CPENABLE=(5,15) setting for all z/OS LPARs running on z14 and later processors is recommended.

For more information about CPENABLE, refer to:

https://www.ibm.com/support/pages/system/files/inline-files/CPENABLE_v2020.pdf

Note: I/O interrupt delay time is *not* counted as part of the I/O response time.

10.1.4 I/O Queuing Activity report

The I/O Queuing Activity report is grouped into three sections:

- Shows how busy the I/O processors (IOPs) in a z Systems server are. IOPs are special purpose processors (SAP) that handle the I/O operations.
- Shows Alias Management Group (AMG) details. An AMG is a set of logical subsystems (LSS, also known as logical control units, or simply control units), each containing base and alias devices. Within an AMG, an alias device can be used for I/O to any base PAV device in the AMG. SuperPAV functionality then allows the use of PAVs across all the Logical control units (LCUs) or LSSs that have an affinity to the same DS8000 internal server and have the same paths to the DS8000 storage subsystem.
- Shows I/O Queuing Activity broken down by Logical control units (LCU).

If the utilization (% IOP BUSY) is unbalanced and certain IOPs are saturated, it can help to redistribute the channels assigned to the storage systems. An IOP is assigned to handle a certain set of channel paths. Assigning all of the channels from one IOP to access a busy disk system can cause a saturation on that particular IOP. For more information, see the hardware manual of the CPC that you use.

Example 10-5 shows an I/O Queuing Activity report for the IOPs section.

Example 10-5 I/O Queuing Activity report - Input/Output Processors

```
                                                I/O   Q U E U I N G   A C T I V I T Y
-------------------------------------------------------------------------------------------------------------------------------------
                                                      INPUT/OUTPUT PROCESSORS
 ------------------------------------------------------------------------------------------------------------------------------------
     -INITIATIVE QUEUE-  ----------- IOP UTILIZATION ------------   -- % I/O REQUESTS RETRIED --   -------- RETRIES / SSCH ---------
 IOP   ACTIVITY  AVG Q   % IOP  % CMPR  % SCM I/O START INTERRUPT          CP    DP    CU    DV             CP     DP     CU     DV
         RATE    LNGTH    BUSY   BUSY    BUSY    RATE      RATE     ALL  BUSY  BUSY  BUSY  BUSY      ALL  BUSY   BUSY   BUSY   BUSY
  00   8877.770   0.00    3.72    0.00   0.00  8877.766  13297.95    0.0   0.0   0.0   0.0   0.0     0.00  0.00   0.00   0.00   0.00
  01   8878.297   0.00    3.89    0.00   0.00  8878.297  13296.23    0.0   0.0   0.0   0.0   0.0     0.00  0.00   0.00   0.00   0.00
  02   8877.566   0.00    1.61    0.00   0.00  8877.566    62.403    0.0   0.0   0.0   0.0   0.0     0.00  0.00   0.00   0.00   0.00
 SYS   26633.64   0.00    3.07    0.00   0.00  26633.63  26656.59    0.0   0.0   0.0   0.0   0.0     0.00  0.00   0.00   0.00   0.00
```

In a SuperPAV environment, the AMG section of the I/O Queueing Activity report shows the combined LCU performance attributed to each AMG. This report is to be used in conjunction with the LCU section of the I/O Queuing Activity report. The LCU section shows a breakdown of each LCU and the related AMG.

Example 10-6 on page 235 show an I/O Queuing Activity report with SuperPAV enabled, AMGs formed and LCUs related to the AMGs.

Example 10-6   I/O Queuing Activity report - Alias Management Groups and LCUs related to AMGs

```
                                                I/O   Q U E U I N G   A C T I V I T Y
---------------------------------------------------------------------------------------------------------------------------
                                                     ALIAS MANAGEMENT GROUPS
---------------------------------------------------------------------------------------------------------------------------
                                                        AVG   AVG             DELAY   AVG               AVG   DATA
 AMG         DCM GROUP  CHAN     CHPID    % DP   % CU   CUB   CMR CONTENTION      Q   CSS     HPAV      OPEN  XFER
            MIN MAX DEF PATHS    TAKEN    BUSY   BUSY   DLY   DLY    RATE     LNGTH   DLY   WAIT  MAX   EXCH  CONC
 00000004               82       1.453    0.00   0.00   0.0   0.0
                        85       1.340    0.00   0.00   0.0   0.0
                        83       1.340    0.00   0.00   0.0   0.0
                        86       1.333    0.00   0.00   0.0   0.0
                         *       5.467    0.00   0.00   0.0   0.0     0.000    0.00   0.1  0.000    0   0.00  0.00
 00000006               80       0.683    0.00   0.00   0.0   0.0
                        81       0.627    0.00   0.00   0.0   0.0
                        84       0.627    0.00   0.00   0.0   0.0
                        87       0.627    0.00   0.00   0.0   0.0
                         *       2.563    0.00   0.00   0.0   0.0     0.000    0.00   0.1  0.000    0   0.00  0.00

---------------------------------------------------------------------------------------------------------------------------
                                                LOGICAL CONTROL UNITS
---------------------------------------------------------------------------------------------------------------------------
                                                             AVG   AVG             DELAY   AVG               AVG   DATA
 LCU/      CU     DCM GROUP  CHAN     CHPID    % DP   % CU   CUB   CMR CONTENTION      Q   CSS     HPAV      OPEN  XFER
  AMG            MIN MAX DEF PATHS    TAKEN    BUSY   BUSY   DLY   DLY    RATE     LNGTH   DLY   WAIT  MAX   EXCH  CONC
 015A      8201              82       0.030    0.00   0.00   0.0   0.0
  00000007                   85       0.020    0.00   0.00   0.0   0.0
                             83       0.030    0.00   0.00   0.0   0.0
                             86       0.023    0.00   0.00   0.0   0.0
                              *       0.103    0.00   0.00   0.0   0.0     0.000    0.00   0.1  0.000    0   0.00  0.00
 015C      8401              82       0.010    0.00   0.00   0.0   0.0
  00000007                   85       0.010    0.00   0.00   0.0   0.0
                             83       0.013    0.00   0.00   0.0   0.0
                             86       0.013    0.00   0.00   0.0   0.0
                              *       0.047    0.00   0.00   0.0   0.0     0.000    0.00   0.1  0.000    0   0.00  0.00
```

> **Note:** If the system is not running in SuperPAV mode or the conditions to enable SuperPAV are not met, no AMGs can be formed and the AMG section is not shown in the I/O Queuing Activity report.

In a HyperPAV environment, you can also check the usage of HyperPAV alias addresses. Example 10-7 shows the LCU section of the I/O Queueing Activity report. It reports on HyperPAV alias usage in the HPAV MAX column. Here, a maximum of 2 PAV alias addresses were used for that LCU during the reporting interval. You can compare this value to the number of aliases that are defined for that LCU. If all are used, you might experience delays because of a lack of aliases.

The `HPAV WAIT` value also indicates this condition. It is calculated as the ratio of the number of I/O requests that cannot start because no HyperPAV aliases are available to the total number of I/O requests for that LCU. If it is nonzero in a significant number of intervals, you might consider defining more aliases for this LCU.

Example 10-7   I/O Queuing Activity report that uses HyperPAV

```
                                                I/O   Q U E U I N G   A C T I V I T Y
------------------------------------------------------------------------------------------------------------------------------------
                                                     LOGICAL CONTROL UNITS
------------------------------------------------------------------------------------------------------------------------------------
                                                             AVG   AVG             DELAY   AVG               AVG   DATA
 LCU/      CU     DCM GROUP  CHAN     CHPID    % DP   % CU   CUB   CMR CONTENTION      Q   CSS     HPAV      OPEN  XFER
  AMG            MIN MAX DEF PATHS    TAKEN    BUSY   BUSY   DLY   DLY    RATE     LNGTH   DLY   WAIT  MAX   EXCH  CONC
 015A      8201              82      1931.8    0.00   0.00   0.0   0.0
                             85      1931.8    0.00   0.00   0.0   0.0
                             83      1931.9    0.00   0.00   0.0   0.0
                             86      1931.9    0.00   0.00   0.0   0.0
```

```
              *      7727.3    0.00    0.00    0.0    0.0     0.000    0.00   0.1  0.000    2   2.11  2.11
```

Note: If your HPAV MAX value is constantly below the number of defined alias addresses, you can consider unassigning some aliases and use these addresses for additional base devices. Do this only if you are short of device addresses. Monitor HPAV MAX over an extended period to ensure that you do not miss periods of higher demand for PAV.

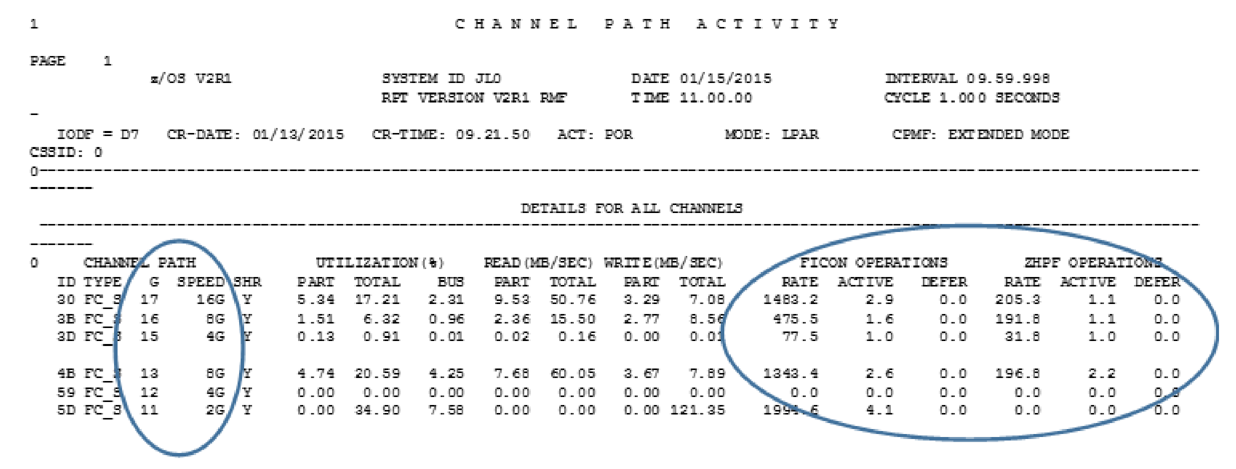
10.1.5 FICON host channel report

The CHANNEL PATH ACTIVITY report, which is shown in Figure 10-3, shows the FICON channel statistics:

- The PART Utilization is the utilization of the channel hardware related to the I/O activity on this LPAR.
- The Total Utilization is the sum from all LPARs that share the channel. It should not exceed 50% during normal operation
- The BUS utilization indicates the usage of the internal bus that connects the channel to the CPC. The suggested maximum value is 40%.
- The FICON link utilization is not explicitly reported. For a rough approach, take the higher of the TOTAL READ or TOTAL WRITE values and divide it by the maximum achievable data rate, as indicated by the SPEED field. Consider that SPEED is in Gbps, and READ and WRITE values are in MBps. The link utilization should not exceed 70%.

Note: We explain this estimation with CHPID 30 in the example in Figure 10-3. The SPEED value is 16 Gbps, which roughly converts to 1600 MBps. TOTAL READ is 50.76 MBps, which is higher than TOTAL WRITE. Therefore, the link utilization is approximately 50.78 divided by 1600, which results in 0.032 or 3.2%

Exceeding the thresholds significantly causes frame pacing, which eventually leads to higher than necessary CONNECT times. If this happens only for a few intervals, it is most likely no problem.



```
1                                        C H A N N E L   P A T H   A C T I V I T Y

PAGE    1
            z/OS V2R1                SYSTEM ID JL0              DATE 01/15/2015            INTERVAL 09.59.998
                                     RPT VERSION V2R1 RMF       TIME 11.00.00              CYCLE 1.000 SECONDS
-
   IODF = D7   CR-DATE: 01/13/2015   CR-TIME: 09.21.50   ACT: POR          MODE: LPAR       CPMF: EXTENDED MODE
CSSID: 0
0-------------------------------------------------------------------------------------------------------------------------
-------
                                                  DETAILS FOR ALL CHANNELS
 -------------------------------------------------------------------------------------------------------------------------
-------
0     CHANNEL PATH          UTILIZATION(%)    READ(MB/SEC) WRITE(MB/SEC)       FICON OPERATIONS          ZHPF OPERATIONS
   ID TYPE   G  SPEED SHR   PART  TOTAL   BUS   PART  TOTAL   PART  TOTAL     RATE  ACTIVE   DEFER    RATE  ACTIVE  DEFER
   30 FC_S  17    16G  Y    5.34  17.21  2.31   9.53  50.76   3.29   7.08   1483.2     2.9     0.0   205.3     1.1    0.0
   3B FC_S  16     8G  Y    1.51   6.32  0.96   2.36  15.50   2.77   8.56    475.5     1.6     0.0   191.8     1.1    0.0
   3D FC_S  15     4G  Y    0.13   0.91  0.01   0.02   0.16   0.00   0.01     77.5     1.0     0.0    31.8     1.0    0.0

   4B FC_S  13     8G  Y    4.74  20.59  4.25   7.68  60.05   3.67   7.89   1343.4     2.6     0.0   196.8     2.2    0.0
   59 FC_S  12     4G  Y    0.00   0.00  0.00   0.00   0.00   0.00   0.00      0.0     0.0     0.0     0.0     0.0    0.0
   5D FC_S  11     2G  Y    0.00  34.90  7.58   0.00   0.00   0.00 121.35   1994.6     4.1     0.0     0.0     0.0    0.0
```

Figure 10-3 Channel Path Activity report

For small block transfers, the BUS utilization is less than the FICON channel utilization. For large block transfers, the BUS utilization is greater than the FICON channel utilization.

The Generation (G) field in the channel report shows the combination of the FICON channel generation that is installed and the speed of the FICON channel link for this CHPID at the time of the storage system start. The G field does not include any information about the link between the director and the DS8000 storage system.

Depending on the type of Fibre Channel host adapter installed in the DS8000 storage system, the link between the director and the DS8000 storage system can run at 32Gbps and 16Gbps and negotiate down to 8Gbps and 4Gbps.

If the channel is point-to-point connected to the DS8000 HA port, the G field indicates the negotiated speed between the FICON channel and the DS8000 port. The SPEED column indicates the actual channel path speed at the end of the interval.

The RATE field in the FICON OPERATIONS or zHPF OPERATIONS columns means the number of FICON or zHPF I/Os per second initiated at the physical channel level. It is not broken down by LPAR.

10.1.6 FICON Director Activity report

The FICON director is the switch that connects the host FICON channel to the DS8000 HA port. FICON director performance statistics are collected in the SMF record type 74 subtype 7. The FICON Director Activity report (Example 10-8) provides information about director and port activities.

> **Note:** To get the data that is related to the FICON Director Activity report, the CUP device must be online on the gathering z/OS system.

The measurements that are provided are on a director port level. It represents the total I/O passing through this port and is not broken down by LPAR or device.

The CONNECTION field indicates how a port is connected:

- CHP: The port is connected to a FICON channel on the host.
- CHP-H: The port is connected to a FICON channel on the host that requested this report.
- CU: This port is connected to a port on a disk subsystem.
- SWITCH: This port is connected to another FICON director.

The important performance metric is AVG FRAME PACING. This metric shows the average time (in microseconds) that a frame waited before it was transmitted. The higher the contention on the director port, the higher the average frame pacing metric. High frame pacing negatively influences the CONNECT time.

Example 10-8 FICON director activity report

```
                              F I C O N   D I R E C T O R   A C T I V I T Y

PORT    ---------CONNECTION--------   AVG FRAME     AVG FRAME SIZE     PORT BANDWIDTH (MB/SEC)     ERROR
ADDR    UNIT      ID  SERIAL NUMBER     PACING        READ   WRITE     -- READ --  -- WRITE --     COUNT
 05     CHP       FA   0000000ABC11          0         808     285          50.04        10.50         0
 07     CHP       4A   0000000ABC11          0         149     964          20.55         5.01         0
 09     CHP       FC   0000000ABC11          0         558    1424          50.07        10.53         0
 0B     CHP-H     F4   0000000ABC12          0         872     896          50.00        10.56         0
 12     CHP       D5   0000000ABC11          0          73     574          20.51         5.07         0
 13     CHP       C8   0000000ABC11          0         868    1134          70.52         2.08         1
 14     SWITCH  ----   0ABCDEFGHIJK          0         962     287          50.03        10.59         0
 15     CU      C800   0000000XYG11          0        1188     731          20.54         5.00         0
```

10.1.7 Cache and NVS report

The CACHE SUBSYSTEM ACTIVITY report provides useful information for analyzing the reason for high DISC time. It is provided on an LCU level. Example 10-9 shows a sample cache overview report.

Note: Cache reports by LCU calculate the total activities of volumes that are online.

Example 10-9 Cache Subsystem Activity summary

```
                                          C A C H E   S U B S Y S T E M   A C T I V I T Y

------------------------------------------------------------------------------------------------------------------------------------
                                                      CACHE SUBSYSTEM OVERVIEW
------------------------------------------------------------------------------------------------------------------------------------
TOTAL I/O    12180   CACHE I/O    12180
TOTAL H/R    1.000   CACHE H/R    1.000
CACHE I/O   -------------READ I/O REQUESTS-------------     ----------------------WRITE I/O REQUESTS----------------------         %
REQUESTS       COUNT     RATE      HITS     RATE     H/R      COUNT     RATE      FAST     RATE      HITS     RATE     H/R      READ
NORMAL          6772     22.6      6770     22.6   1.000       3556     11.9      3556     11.9      3556     11.9   1.000      65.6
SEQUENTIAL       871      2.9       871      2.9   1.000        981      3.3       981      3.3       981      3.3   1.000      47.0
CFW DATA           0      0.0         0      0.0     N/A          0      0.0         0      0.0         0      0.0     N/A       N/A
TOTAL           7643     25.5      7641     25.5   1.000       4537     15.1      4537     15.1      4537     15.1   1.000      62.8
-----------------------CACHE MISSES-----------------------            ----------------MISC---------------
REQUESTS      READ    RATE   WRITE    RATE  TRACKS    RATE                                   COUNT   RATE
                                                                      DELAYED DUE TO NVS         0    0.0
NORMAL           2     0.0       0     0.0       4     0.0            DELAYED DUE TO CACHE       0    0.0
SEQUENTIAL       0     0.0       0     0.0       0     0.0            DFW INHIBIT                0    0.0
CFW DATA         0     0.0       0     0.0                            ASYNC (TRKS)            1093    3.6
TOTAL            2    RATE     0.0
---CKD STATISTICS---     ---RECORD CACHING--                                -HOST ADAPTER ACTIVITY-     --------DISK ACTIVITY-------
                                                                                       BYTES  BYTES              RESP   BYTES  BYTES
WRITE           3333     READ MISSES       0                                            /REQ   /SEC              TIME    /REQ   /SEC
WRITE HITS      3332     WRITE PROM     3449                                READ       16.1K 410.7K     READ    4.000   32.8K    437
                                                                            WRITE      10.2K 154.2K     WRITE  13.261   38.4K 104.0K
```

The report breaks down the LCU level I/O requests by read and by write. It shows the rate, the hit rate, and the hit ratio of the read and the write activities. The read-to-write ratio is also calculated.

Note: The total I/O requests here can be higher than the I/O rate shown in the DASD report. In the DASD report, one channel program is counted as one I/O. However, in the cache report, if there are multiple Locate Record commands in a channel program, each Locate Record command is counted as one I/O request.

In this report, you can check the value of the read hit ratio. Low read hit ratios contribute to higher DISC time. For a cache-friendly workload, you see a read hit ratio of better than 90%. The write hit ratio is usually 100%.

High DELAYED DUE TO NVS is an indication that persistent memory or NVS is overcommitted. It means that write I/Os cannot be completed because persistent memory is full and must be retried. If this Rate is higher than 1%, the write retry operations can affect the DISC time. It is an indication of insufficient back-end resources because write cache destaging operations are not fast enough.

Note: In cases with high DELAYED DUE TO NVS, you usually see high rank utilization in the DS8000 rank statistics, which are described in 10.1.8, “Enterprise Disk Systems report” on page 240.

The DISK ACTIVITY part of the report can give you a rough indication of the back-end performance. This information, along with the ESS report (see 10.1.8, “Enterprise Disk

Systems report” on page 240), will provide a thorough understanding of the amount of read and write requests and the related impact on the ranks and extent pools of the DS8000 storage system.

The report also shows the number of sequential I/O as SEQUENTIAL row and random I/O as NORMAL row for read and write operations. These metrics can also help to analyze and specify I/O bottlenecks.

Example 10-10 is the second part of the CACHE SUBSYSTEM ACTIVITY report, providing measurements for each volume in the LCU. You can also see to which extent pool each volume belongs.

Example 10-10   Cache Subsystem Activity by volume serial number

```
                                          C A C H E   S U B S Y S T E M   A C T I V I T Y

------------------------------------------------------------------------------------------------------------------------------------
                                                      CACHE SUBSYSTEM DEVICE OVERVIEW
------------------------------------------------------------------------------------------------------------------------------------
VOLUME      DEV XTNT      %    I/O   ---CACHE HIT RATE--   ----DASD I/O RATE----    ASYNC   TOTAL    READ   WRITE       %
SERIAL      NUM POOL   I/O   RATE    READ    DFW    CFW   STAGE  DEL NVS  OTHER     RATE     H/R     H/R     H/R    READ
*ALL                 100.0   7723    3869   3854    0.0     0.0      0.0    0.0     9323   1.000   1.000   1.000    50.1
AT8200    08200 0002  50.0   3858    3858    0.0    0.0     0.0      0.0    0.0      0.0   1.000   1.000     N/A   100.0
AT8201    08201 0002   0.0    0.0     0.0    0.0    0.0     0.0      0.0    0.0      0.0     N/A     N/A     N/A     N/A
AT8202    08202 0002   0.0    0.0     0.0    0.0    0.0     0.0      0.0    0.0      0.0     N/A     N/A     N/A     N/A
AT8203    08203 0002   0.0    0.0     0.0    0.0    0.0     0.0      0.0    0.0      0.0     N/A     N/A     N/A     N/A
AT8230    08230 0002   0.7   55.1     0.1   55.1    0.0     0.0      0.0    0.0     55.1   1.000   1.000   1.000     0.1
AT8231    08231 0002   0.0    0.1     0.1    0.0    0.0     0.0      0.0    0.0     55.1   1.000   1.000   1.000    75.9
AT8232    08232 0002   0.7   55.1     0.1   55.1    0.0     0.0      0.0    0.0     55.1   1.000   1.000   1.000     0.1
AT8233    08233 0002   0.0    0.1     0.1    0.0    0.0     0.0      0.0    0.0     55.1   1.000   1.000   1.000    75.9
```

If you specify **REPORTS(CACHE(DEVICE))** when running the cache report, you get the complete report for each volume, as shown in Example 10-11. You get detailed cache statistics of each volume. By specifying **REPORTS(CACHE(SSID(nnnn)))**, you can limit this report to only certain LCUs.

Example 10-11   Cache Device Activity report detail by volume

```
C A C H E   D E V I C E   A C T I V I T Y
VOLSER     AT8200      NUM    08200    EXTENT POOL   0002
------------------------------------------------------------------------------------------------------------------------------------
                                                        CACHE DEVICE STATUS
------------------------------------------------------------------------------------------------------------------------------------
CACHE STATUS
DASD FAST WRITE     - ACTIVE
PINNED DATA         - NONE
------------------------------------------------------------------------------------------------------------------------------------
                                                       CACHE DEVICE ACTIVITY
------------------------------------------------------------------------------------------------------------------------------------
TOTAL I/O    1157K   CACHE I/O    1157K
TOTAL H/R    1.000   CACHE H/R    1.000
CACHE I/O    -------------READ I/O REQUESTS-------------    ----------------------WRITE I/O REQUESTS----------------------         %
REQUESTS       COUNT     RATE      HITS     RATE     H/R      COUNT     RATE      FAST     RATE      HITS     RATE     H/R      READ
NORMAL          2375      7.9      2375      7.9   1.000          0      0.0         0      0.0         0      0.0     N/A     100.0
SEQUENTIAL     1155K     3850     1155K     3850   1.000          0      0.0         0      0.0         0      0.0     N/A     100.0
CFW DATA           0      0.0         0      0.0     N/A          0      0.0         0      0.0         0      0.0     N/A       N/A
TOTAL          1157K     3858     1157K     3858   1.000          0      0.0         0      0.0         0      0.0     N/A     100.0
-----------------------CACHE MISSES-----------------------            ----------------MISC---------------
REQUESTS      READ    RATE   WRITE    RATE  TRACKS      RATE                                   COUNT   RATE
                                                                      DELAYED DUE TO NVS         0    0.0
NORMAL           0     0.0       0     0.0   44630     148.8            DELAYED DUE TO CACHE       0    0.0
SEQUENTIAL       0     0.0       0     0.0   1402K      4672            DFW INHIBIT                0    0.0
CFW DATA         0     0.0       0     0.0                            ASYNC (TRKS)               2    0.0
TOTAL            0    RATE     0.0
---CKD STATISTICS---     ---RECORD CACHING--     --SYNCH I/O ACTIVITY--     -HOST ADAPTER ACTIVITY-     --------DISK ACTIVITY-------
                                                           REQ     HITS                BYTES  BYTES              RESP   BYTES  BYTES
WRITE              0     READ MISSES       0              /SEC     /REQ                 /REQ   /SEC              TIME    /REQ   /SEC
WRITE HITS         0     WRITE PROM        0     READ      0.0    0.000     READ       49.5K 190.9M     READ    0.000       0      0
                                                 WRITE     0.0    0.000     WRITE          0      0     WRITE   0.000       0      0
```

10.1.8 Enterprise Disk Systems report

The Enterprise Disk Systems (ESS) report provides the following measurement:

- DS8000 rank activity as ESS RANK STATISTICS
- DS8000 FICON/Fibre Channel port and HA activity as ESS LINK STATISTICS
- DS8000 extent pool information as ESS EXTENT POOL STATISTICS

> **Important:** Enterprise Storage Server data is not gathered by default. Make sure that Enterprise Storage Server data is collected as processed, as described in 10.1.1, “RMF Overview” on page 226.

DS8000 rank

ESS RANK STATISTICS provides measurements of back-end activity on the extent pool and RAID array (rank) levels, such as OPS/SEC, BYTES/OP, BYTES/SEC, and RTIME/OP, for read and write operations.

Example 10-12 shows rank statistics for a system with multi-rank extent pools, which contain High Performance and High Capacity Flash Enclosures (i.e., an All Flash storage system with multiple tiers of flash).

Example 10-12 Rank statistics example

```
                                             E S S   R A N K   S T A T I S T I C S

                            ----- READ OPERATIONS  ----   ----- WRITE OPERATIONS ----
--EXTENT POOL-      ADAPT      OPS  BYTES  BYTES  RTIME      OPS BYTES  BYTES   RTIME  -----ARRAY----  MIN    RANK  RAID
 ID  TYPE      RRID    ID     /SEC    /OP   /SEC    /OP     /SEC   /OP   /SEC     /OP  SSD  NUM  WDTH  RPM     CAP  TYPE
0000 FIBRE 1Gb 0000  0003   2939.7  41.5K 122.0M  0.439   3405.7  60.1K 204.6M  3.498   Y     1     6  N/A   9600G  RAID 6
               0001  0002     70.5  32.7K   2.3M  0.424     88.3 372.2K  32.9M  1.874   Y     1     5  N/A  38400G  RAID 6
               POOL         3010.2  41.3K 124.3M  0.439   3494.0  68.0K 237.5M  3.457   Y     2    11  N/A  48000G  RAID 6
0001 FIBRE 1Gb 0002  0003     20.3  65.5K   1.3M  0.367     21.3  74.6K   1.6M  1.091   Y     1     6  N/A   9600G  RAID 6
               0003  0002      0.0   0.0    0.0   0.000      1.9 611.7K   1.2M  3.194   Y     1     5  N/A  38400G  RAID 6
               POOL           20.3  65.5K   1.3M  0.367     23.2 119.2K   2.8M  1.266   Y     2    11  N/A  48000G  RAID 6
0002 CKD 1Gb   0004  0003   4843.7  57.3K 277.5M  0.950   1146.8 467.5K 536.2M  8.837   Y     1     5  N/A   8000G  RAID 6
0003 CKD 1Gb   0005  0003      0.1  57.3K   6.1K  0.500      0.7   8.2K   6.1K  1.500   Y     1     5  N/A   8000G  RAID 6
```

> **Note:** This report also shows the relationship of ranks to DA pairs (ADAPT ID). With all Flash DS8000 storage systems, the MIN RPM value will be N/A.

If I/O response elongation and rank saturation is suspected, it is important to check IOPS (OPS/SEC) and throughput (BYTES/SEC) for both read and write rank activities. Also, you must make sure what kind of workloads were run and how was the ratio of those workloads. If your workload is of a random type, the IOPS is the significant figure. If it is more sequential, the throughput is more significant.

The rank response times (RTIME) can be indicators for saturation. In a balanced and sized system with growth potential, the read and write response times are in the order of 1 - 2 ms for HPFE ranks. Ranks with response times reaching the range of 3 - 5 ms for HPFE are approaching saturation.

> **Note:** Both Spectrum Control and Storage Insights can also provide relevant information.
>
> For more information see: 6.2, “Data from IBM Spectrum Control and Storage Insights” on page 127 and 7.2.1, “IBM Spectrum Control and IBM Storage Insights Pro overview” on page 141

DS8000 FICON/Fibre port and host adapter

The Enterprise Storage Server Link Statistics report includes data for all DS8000 HA ports, regardless of their configuration (FICON, FCP, or PPRC). Example 10-13 shows a sample that contains FICON ports (link type ECKD READ and WRITE) and PPRC ports (PPRC SEND and PPRC RECEIVE). It also shows that the ports are running at 8 Gbps.

Example 10-13 DS8000 link statistics

```
                                            E S S   L I N K   S T A T I S T I C S

------ADAPTER------   --LINK TYPE--       BYTES      BYTES         OPERATIONS      RESP TIME         I/O
SAID  TYPE                                /SEC       /OPERATION    /SEC            /OPERATION        INTENSITY
0200  FIBRE 8Gb       SCSI READ           124.2M     34.2K        3627.4           0.114            412.0
0200  FIBRE 8Gb       SCSI WRITE           61.2M     29.3K        2086.3           0.501           1046.0
                                                                                                     ------
                                                                                                     1458.1
0201  FIBRE 8Gb       ECKD READ            46.5M     49.3K         943.1           0.118            111.1
0201  FIBRE 8Gb       ECKD WRITE           49.8M     48.9K        1017.5           0.274            278.6
                                                                                                     ------
                                                                                                      389.7
0240  FIBRE 8Gb       ECKD READ            49.4M     49.3K        1001.8           0.118            118.1
0240  FIBRE 8Gb       ECKD WRITE           46.5M     48.8K         952.4           0.274            261.2
                                                                                                     ------
                                                                                                      379.3
0242  FIBRE 8Gb       PPRC SEND             0.0        0.0           0.0           0.000              0.0
0242  FIBRE 8Gb       PPRC RECEIVE         57.0M     25.0K        2280.9           0.146            333.3
                                                                                                     ------
                                                                                                      333.3
```

For a definition of the HA port ID (SAID), see *IBM DS8900F Architecture and Implementation*, SG24-8456.

The I/O INTENSITY is the result of multiplication of the operations per second and the response time per operation. For FICON ports, it is calculated for both the read and write operations, and for PPRC ports, it is calculated for both the send and receive operations. The total I/O intensity is the sum of those two numbers on each port.

For FICON ports, the I/O intensity should be below 2000 for most of the time. With higher values, the interface might start becoming ineffective. Much higher I/O intensities can affect the response time, especially PEND and CONN components. With a value of 4000, it is saturated. For more information, see the description of PEND and CONN times in “PEND time” on page 232 and “CONN time” on page 233. This rule does not apply for PPRC ports, especially if the distance between the primary and secondary sites is significant.

> **Note:** IBM Spectrum Control performance monitoring also provides a calculated port utilization value. For more information, see 7.2.1, “IBM Spectrum Control and IBM Storage Insights Pro overview” on page 141.

If you must monitor or analyze statistics for specific ports over time, you can generate an overview report and filter for the subject port IDs. Provide post processor control statements like the ones in Example 10-14 and you get an overview report, as shown in Example 10-15 on page 242.

Example 10-14 Control statements for DS8000 link overview report

```
OVW(TSEND003(ESTRPSD(SERN(0000ABC11),SAID(0003))))
OVW(RSEND003(ESRTPSD(SERN(0000ABC11),SAID(0003))))
OVW(TSEND133(ESTRPSD(SERN(0000ABC11),SAID(0133))))
OVW(RSEND133(ESRTPSD(SERN(0000ABC11),SAID(0133))))
```

Example 10-15 RMF overview report

```
                         R M F   O V E R V I E W   R E P O R T

DATE    TIME      INT       TSEND033  RSEND033  TSEND233  RSEND233
MM/DD HH.MM.SS HH.MM.SS
06/20 20.53.00 00.01.00     133.7M     39.0     136.9M     39.6
06/20 20.54.00 00.00.59     123.4M     42.4     123.4M     42.7
06/20 20.55.00 00.00.59     121.5M     41.8     114.6M     40.4
06/20 20.56.00 00.01.00     121.3M     43.4     124.1M     42.1
06/20 20.57.00 00.00.59     118.2M     35.3     117.4M     36.4
06/20 20.58.00 00.01.00     103.8M     34.1     105.3M     35.0
06/20 20.59.00 00.01.00      93.9M     28.5      88.9M     27.1
06/20 21.00.00 00.00.59      86.5M     28.3      88.5M     29.0
```

10.1.9 Alternatives and supplements to RMF

There are several other tools available that can help with monitoring the performance of the DS8000 storage system:

IBM Z Monitoring Suite

The IBM Z Monitoring Suite brings together a range of IBM Z Monitoring tools into a single integrated solution. It is designed to provide both operations and deep domain experts with monitoring for health, availability and performance management in near real-time and historical metrics, across IBM z/OS operating system, networks, storage subsystems, Java runtimes, IBM Db2, IBM CICS®, IBM IMS, IBM MQ for z/OS and IBM WebSphere® Application Server. For more information, refer to:

https://www.ibm.com/products/z-monitoring-suite

RMF Data Portal for z/OS

One of the key components for the sysplex wide access of Monitor III data is the RMF Distributed Data Server (DDS). With RMF, DDS offers an HTTP API that can access short-term data from the Monitor III and long-term data from the Postprocessor. An application program can send an HTTP request for selected performance data to the DDS. Since the requested data are returned as XML documents, a web browser can act as Data Portal to retrieve and display RMF Postprocessor data.

The DDS exploiters of Monitor III performance data are, among others, z/OS Capacity Provisioning, z/OSMF, or RMF PM. If you want to monitor systems in a sysplex, you must set up a Distributed Data Server (DDS) host session on the system in the sysplex with the highest RMF release. If you want to monitor several sysplexes, each one needs to have an active DDS. DDS supports the request of performance data in a z/OS environment via HTTP requests.

With current z/OS versions, RMF Data Portal for z/OS is an additional exploiter of the RMF DDS data. The RMF Data Portal for z/OS panel is shown in Figure 10-4 on page 243. The web browser based user interface allows access to a cross-sysplex RMF Monitor III and RMF Postprocessor Reports from a single point of control.

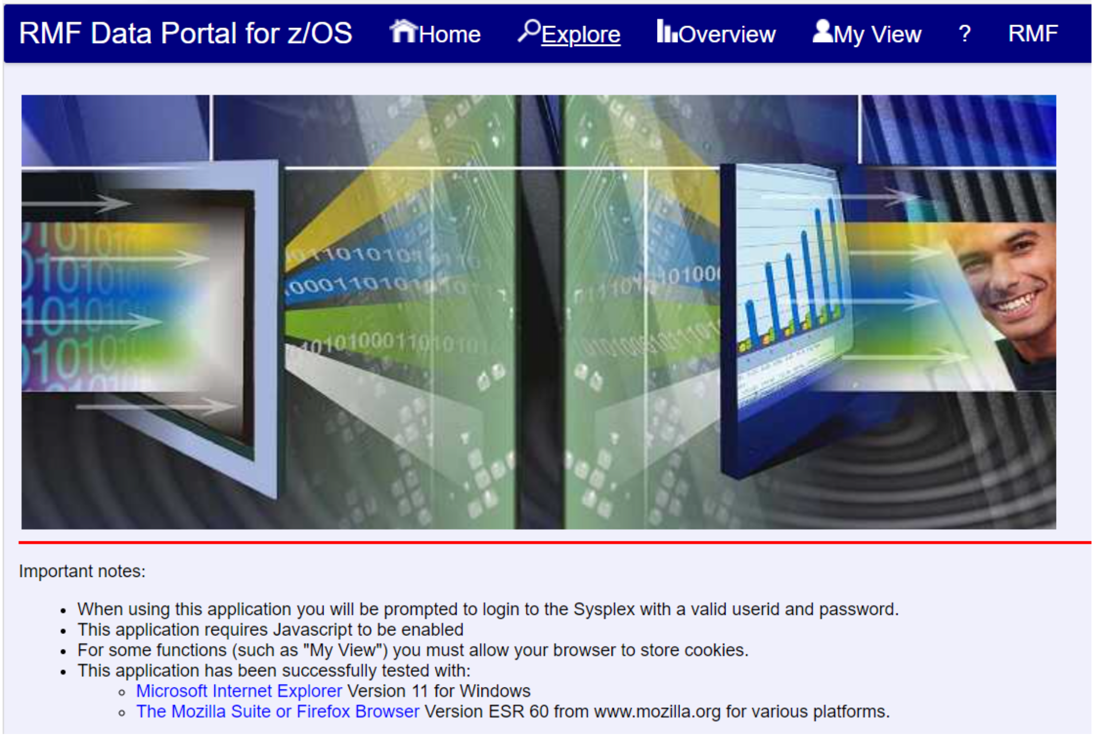
To access the RMF Data Portal for z/OS, after performing the proper customizations, point your browser to:

```
http://<ddshost>:<ddsport>
```

where:

- – ddshost is the IP address or the symbolic name of the DDS server.
- – ddsport (default port is 8803) is the port number of the DDS server (GPMSERVE or GPM4CIM).

Figure 10-4, "Browser access to RMF Data Portal for z/OS"



Important notes:

- When using this application you will be prompted to login to the Sysplex with a valid userid and password.
- This application requires Javascript to be enabled
- For some functions (such as "My View") you must allow your browser to store cookies.
- This application has been successfully tested with:
  - Microsoft Internet Explorer Version 11 for Windows
  - The Mozilla Suite or Firefox Browser Version ESR 60 from www.mozilla.org for various platforms.

Figure 10-4   Browser access to RMF Data Portal for z/OS

At the top of the RMF Data Portal for z/OS page, under the Explore link, you gain access to the RMF Monitor III Data, Resources, Metrics, and to the RMF Postprocessor Reports. Selecting the Metrics link, under the RMF Monitor III Data, you have the option to view the Full RMF Reports and individual metrics with their description and explanation.

Under the RMF Post Processor Reports panel, you can run any of the reports available (CACHE, CHAN, CPU, CRYPTO, DEVICE, EADM, ENQ, ESS, FCD, HFS, IOQ, OMVS PAGESP, PAGING, PCIE, SDELAY, XCF, CF, SDEVICE, WLMGL and OVW) by providing the filter options to tailor the display output. An explanation about the format of data and values to be entered in the option fields is shown when you mouse over them. Figure 10-13 on page 256 shows a PCIE report obtained from the RMF Data Portal for z/OS explore option.

Note: For systems in a z/OS environment, the DDS GPMSERVE gathers data from RMF instances running on the Sysplex members. The HTTP API of the DDS can also serve requests for AIX and Linux performance data, which are directed against an active GPM4CIM instance. RMF XP supports the following operating systems:

- ► AIX on System p.
- ► Linux on System z®.
- ► Linux on System x.

For more information about RMF Distributed Data Server (DDS) refer to:

z/OS Resource Management Facility Reporter Programmers Guide, GC27-4937-40

Performance data in a z/OS environment can be retrieved by sending HTTP requests to the DDS server on the monitored z/OS Sysplex. Here is an example request for a certain performance metric for a specific resource; it requests the response time of volume SYSLIB on system SYSA:

```
http://ddshost:8803/ICBM/perform.xml?resource=SYSA,SYSLIB,VOLUME&id=8D10F0
```

Application programs can exploit the extended DDS HTTP API by sending standard URL requests for historical RMF Postprocessor data. An example request for the Postprocessor CPU and CRYPTO reports looks similar to the following:

```
http://ddshost:8803/ICBM/rmfpp.xml?reports=CPU,CRYPTO&date=20210701,202107
04
```

In most installations, access to historical data is needed for in-depth performance analysis. This allows keeping track of whether a critical situation has been persistent or not. The existing HTTP API of the RMF Distributed Data Server (DDS) already provides sysplex-wide access of the data collected by RMF Monitor III. As mentioned, this API grants instant and easy access to RMF long-term historical data as reported by the RMF Postprocessor.

For a complete list of RMF documentation, refer to:

https://www.ibm.com/docs/en/zos/2.4.0?topic=zos-rmf

DS8000 GUI Performance Reports

The DS8000 GUI provides s a performance monitoring function. It also provides the function to export performance data in the CSV format. This data is presented online in graphical format, and it is easy to access but provides less detail than RMF. The following GUI Performance Reports are available:

- System Summary Report
- Performance Summary Report
- Easy Tier Summary Report
- Fiber Channel Connectivity Report (new)

IBM Spectrum Control

IBM Spectrum Control does not provide detailed host related performance data like RMF. However, it might be useful as follows:

In a Remote Mirror and Copy environment, you can use IBM Spectrum Control to monitor the performance of the remote disk system if there is no z/OS system that accesses it to gather RMF data.

In mixed (mainframe and distributed) environments, IBM Spectrum Control can also help to gain a quick view of the complete loads which run on the DS8000.

For more information about IBM Spectrum Control, see 7.2.1, “IBM Spectrum Control and IBM Storage Insights Pro overview” on page 141.

IBM Copy Services Manager (CSM)

CSM provides management and monitoring capability of data-copy functions. It can provide:

- Overall health and status of sessions, storage systems
- Information about Copy sessions:
  - Type/properties of session

  - Current session progress, pair status/progress and state of the session
  - Global Mirror RPO historical data and LSS Out of sync track information
- Information about the storage system, connection, logical paths, volume details
- Ability to create a system diagnostic for a specific storage system

All of the above can be combined to develop a comprehensive overview of the replication performance for a DS8000 storage system.

There are several products that are provided by other vendors that use RMF data to provide performance analysis capabilities.

> **Note:** In a DS8000 remote replication configuration, it is important to also collect the performance data of the remote storage systems, even if they are not online to production systems. As mentioned, you can use IBM Spectrum Control for this purpose. In a GDPS controlled environment, the GDPS controlling system on the remote site can collect this data by way of RMF as well.

10.2 DS8000 and z Systems planning and configuration

This section describes general aspects and suggestions for planning the DS8000 configuration. For a less generic and more detailed analysis that accounts for your particular environment, the StorM modeling tool is available to IBM representatives and IBM Business Partners who can help you in the planning activities. StorM can be used to help understand the performance effects of various configuration options, such as the number of ports and HAs, disk drive capacity, and number of disks. For more information, see 6.4, “Modeling and sizing for IBM Z” on page 132.

10.2.1 Sizing considerations

When you plan to configure a new DS8000 storage system, you have many decisions to make. Each of these decisions might affect cost, scalability and performance capabilities of the new system. For the current model of the DS8000 family, the flagship DS8900F storage systems, the choices include, but are not limited to, the following items:

- DS8900F models: DS8910F (Racked or Rack mounted), DS8950F or DS8980F storage systems
- Cache size: 192 GB - 4352GB
- Flash resources (number, capacity, and type)
- HAs (connectivity and throughput requirements)
- Licensing
- Logical Corruption capabilities

A z Systems workload is too complex to be estimated and described with only a few numbers and general guidelines. Furthermore, the capabilities of the storage systems are not just the sum of the capabilities of their components. Advanced technologies, such as zHyperLink and zHyperWrite, can positively influence the complete solution's throughput and other functions, such as point-in-time copies or remote replication, add additional workload.

Most of these factors can be accounted for by modeling the new solution with StorM or Disk Magic. For z Systems, the modeling is based on RMF data and real workload characteristics. You can compare the current to potential new configurations and consider growth (capacity and workload) and the influence of Copy Services and other Z synergy features.

For more information about StorM and Disk Magic, see Chapter 6.1, “IBM Storage Modeller” on page 126, which describes the following items:

- How to get access to the tool or someone who can use it
- The capabilities and limitations
- The data that is required for a proper model

> **Note:** It is important, particularly for mainframe workloads, which often are I/O response time sensitive, to observe the thresholds and limitations provided by the StorM and Disk Magic model.

10.2.2 Optimizing performance

This section describes optimizing the performance of a z Systems server with a DS8000 storage system.

Easy Tier

Easy Tier is a performance enhancement to the DS8000 family of storage systems that helps avoid issues that might occur if the back-end workload is not balanced across all the available resources. For a description of Easy Tier, its capabilities, and how it works, see 1.2.4, “Easy Tier” on page 11.

There are some IBM Redbooks publications that you can refer to if you need more information about Easy Tier:

- *IBM DS8000 Easy Tier*, REDP-4667-08
- *DS8870 Easy Tier Application*, REDP-5014
- *IBM DS8870 Easy Tier Heat Map Transfer*, REDP-5015

Easy Tier provides several capabilities that you can use to solve performance issues. Here are some examples:

- Manage skewed workload in multitier (hybrid) extent pools: If part of your back-end storage is overutilized (hot), consider adding some more flash arrays to the affected pools and let Easy Tier move the hot data to the faster resources.
- Distribute workload across all resources within a storage tier: In both uniform or hybrid extent pools, Easy Tier automatically and transparently moves data from heavily used resources to less used ones within the same tier.
- Give new workloads a good start: If you deploy a new application, you can manually assign its storage space (volumes) to a specific tier by using the Easy Tier Application feature. After the application starts and Easy Tier learns about its requirements, you can switch it to automatic mode.
- Add or remove resources: Using Easy Tier manual mode, you can transparently add new resources (arrays) to an extent pool or remove them from it. Easy Tier automatically redistributes the data for best performance so that you can optimize the use of the back-end resources you have available.

The predominant feature of Easy Tier for a Hybrid storage system is to manage data distribution in hybrid pools. For an all-flash system, the primary objective is to enable a lower-cost configuration to perform equally to or better than a higher cost configuration. It is important to know how much of each resource type is required to optimize performance while keeping the cost as low as possible. Since the latency characteristics of different types of Flash drives are very similar, the main purpose of Easy Tier is to avoid overloading the lower tier especially with regards to large block and sequential writes.

Easy Tier is designed to provide a balanced and stable workload distribution. It aims at minimizing the amount of data that must be moved as part of the optimization. To achieve this task, it monitors data access permanently and establishes migration plans based on current and past patterns. It will not immediately start moving data as soon as it is accessed more frequently than before. There is a certain learning phase. You can use Easy Tier Application to overrule learning if necessary.

FICON channels

The StorM model established for your configuration indicates the number of host channels, DS8000 HAs, and HA FICON ports necessary to meet the performance requirements. The modeling results are valid only if the workload is evenly distributed across all available resources. It is your responsibility to make sure this really is the case.

In addition, there are certain conditions and constraints that you must consider:

- IBM mainframe systems support a maximum of eight channel paths to each logical control unit (LCU). A channel path is a connection from a host channel to a DS8000 HA port.
- A FICON channel port can be shared between several z/OS images and can access several DS8000 HA ports.
- A DS8000 HA port can be accessed by several z/OS images or even multiple z Systems servers.
- A FICON host channel or DS8000 HA port has specific throughput capabilities that are not necessarily equivalent to the link speed they support.
- A z Systems FICON feature or DS8000 HA card has specific throughput capabilities that might be less than the sum of all individual ports on this card. This is especially true for the DS8000 8-port HA cards.

The following sections provide some examples that can help to select the best connection topology.

The simplest case is that you must connect one host to one storage system. If the StorM model indicates that you need eight or less host channels and DS8000 host ports, you can use a one-to-one connection scheme, as shown in Figure 10-5.

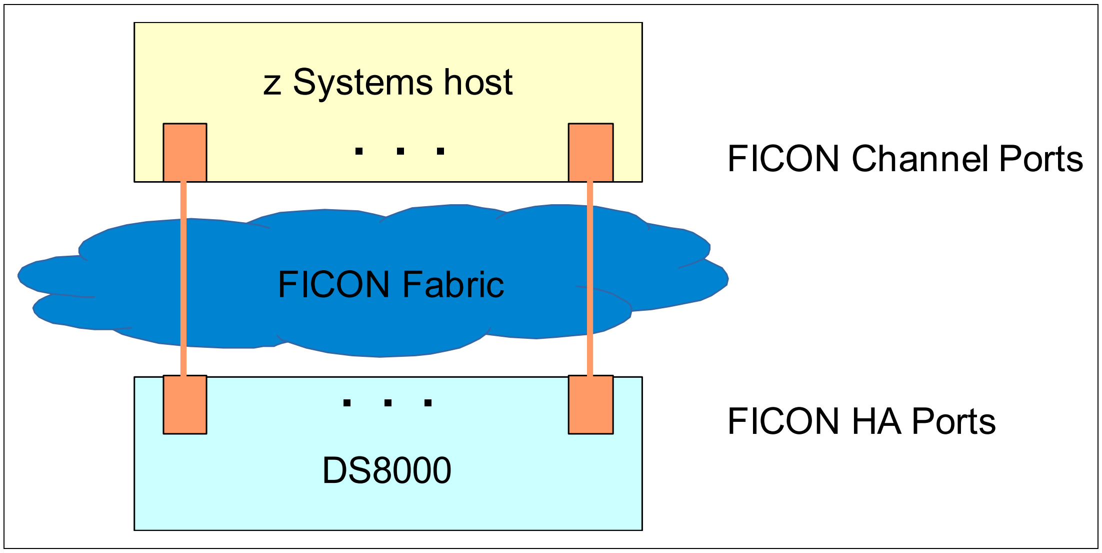


Figure 10-5 One to one host connection

To simplify the figure, it shows only one fabric “cloud”, where you normally have at least two for redundancy reasons. The orange lines that here directly connect one port to another via the fabric. It can range from a direct connection without any switch components to a cascaded FICON configuration.

If the model indicates that more than eight connections are required, you must distribute the connections between resources. One example is to split the DS8000 LCUs into groups and assign a certain number of connections to each of them. This concept is shown in Figure 10-6.

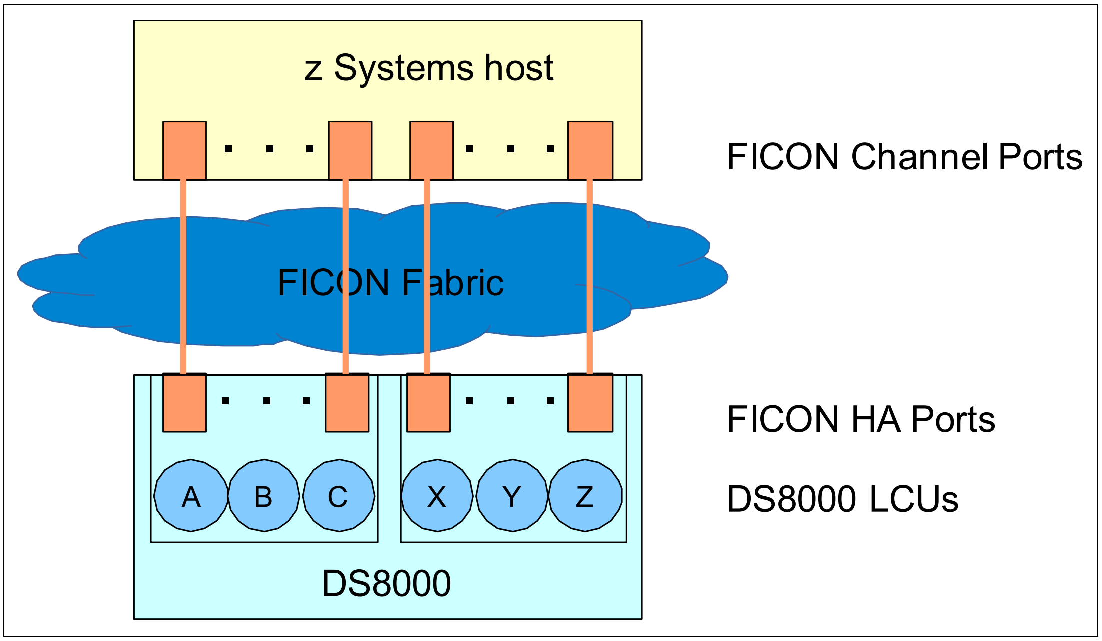


Figure 10-6 Split LCUs into groups and assign them to connections

Therefore, you can have up to eight connections and host and storage ports in each group. It is your responsibility to define the LCU split in a way that each group gets the amount of workload that the assigned number of connections can sustain.

Note: StorM models can be created on several levels. To determine the best LCU split, you might need LCU level modeling.

Therefore, in many environments, more than one host system shares the data and accesses the same storage system. If eight or less storage ports are sufficient according to your StorM model, you can implement a configuration as shown in Figure 10-7, where the storage ports are shared between the host ports.

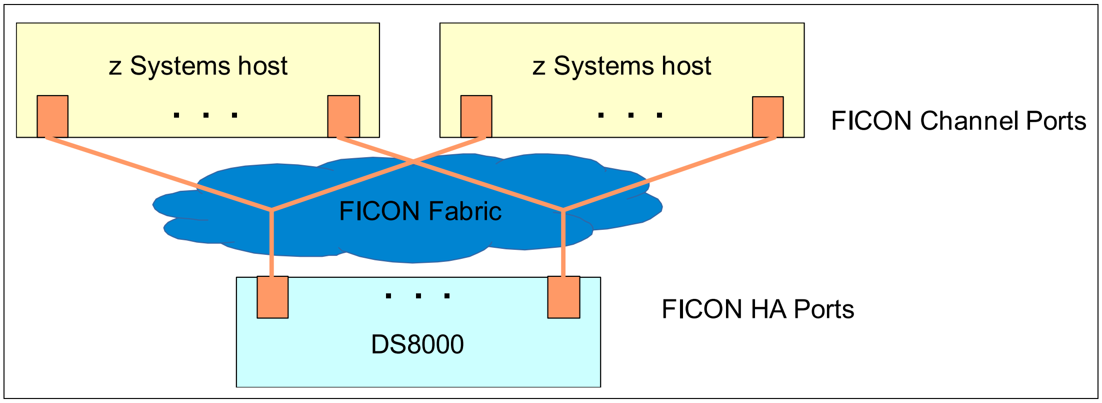


Figure 10-7 Several hosts accessing the same storage system - sharing all ports

If the model indicates that more than eight storage ports are required, the configuration might look like the one that is shown in Figure 10-8.

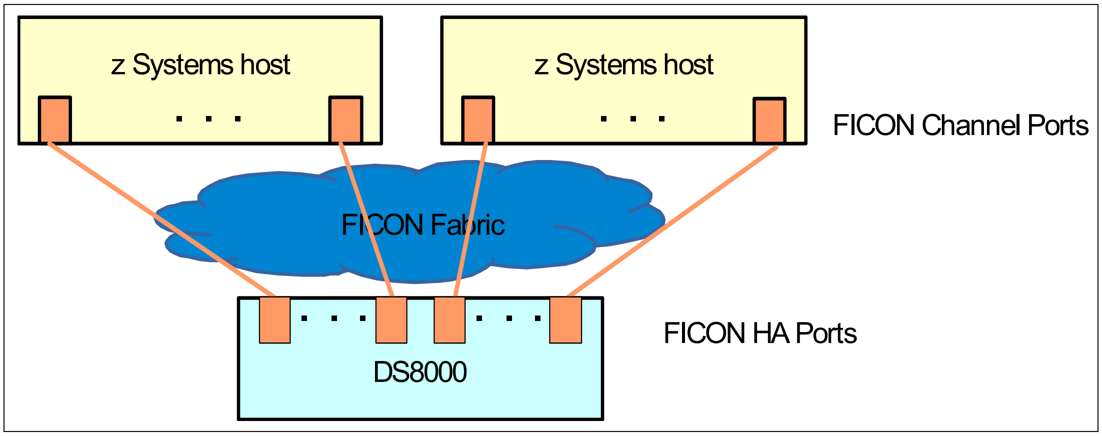


Figure 10-8   Several hosts accessing the same storage system - distributed ports

As you split storage system resources between the host systems, you must ensure that you assign a sufficient number of ports to each host to sustain its workload.

Figure 10-9 shows another variation, where more than one storage system is connected to a host. In this example, all storage system ports share the host ports. This works well if the StorM model does not indicate that more than eight host ports are required for the workload.

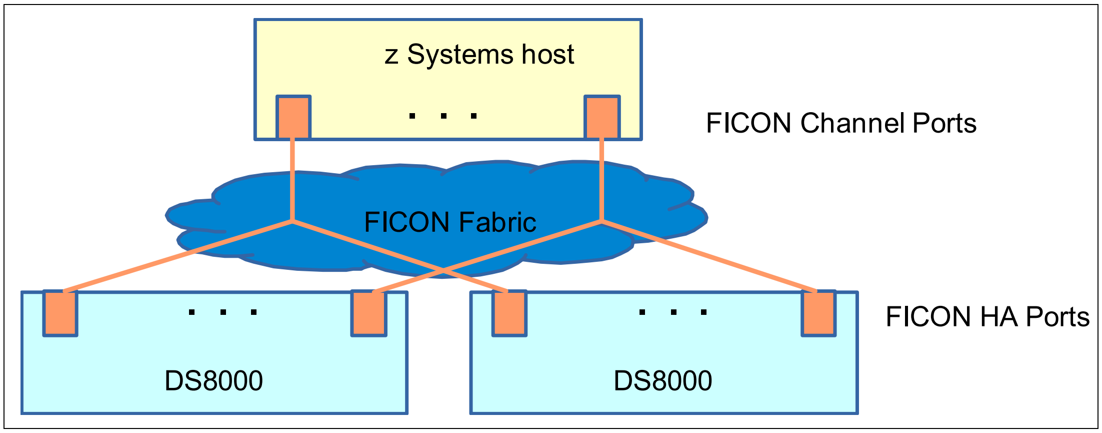


Figure 10-9   More than one storage system connected to a host

If more than eight host ports are required, you must split the host ports between storage resources again. One way of doing this is to provide a separate set of host ports for each storage system, as shown in Figure 10-10.

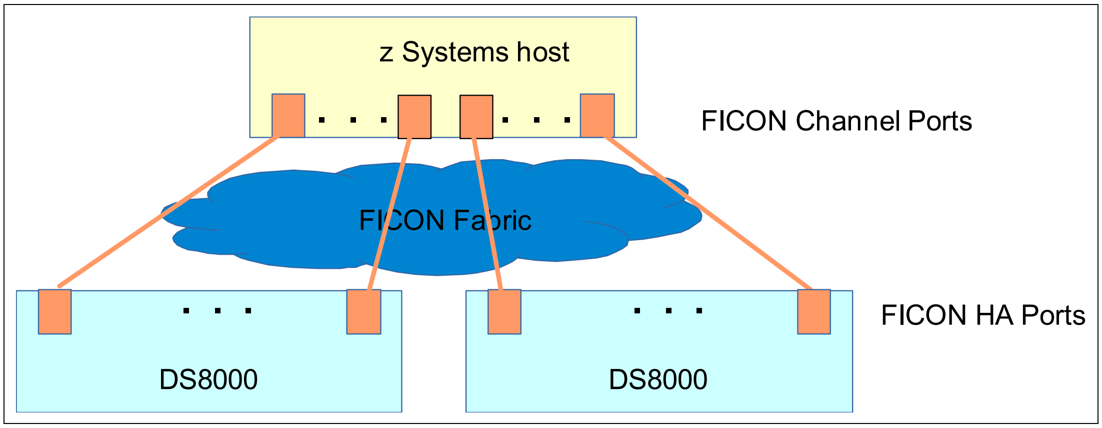


Figure 10-10 Separate host ports for each storage system

Another way of splitting HA ports is by LCU, in a similar fashion as shown in Figure 10-6 on page 248. As in the earlier examples with split resources, you must make sure that you provide enough host ports for each storage system to sustain the workload to which it is subject.

Mixed workload: A DS8000 HA has several ports. You can set each individual port to run FICON or FCP topology. There is nothing wrong with using an HA for FICON, FCP, and remote replication connectivity concurrently. However, if you need the highest possible throughput or lowest possible response time for a given topology, consider isolating this topology on separate HAs.

Remote replication: To optimize the HA throughput if you have remote replication active, consider not sharing HA ports for the following items:

- Synchronous and asynchronous remote replication
- For FCP host I/O and remote replication

High Performance FICON (zHPF): zHPF is an enhancement to FICON, which improves the channel throughput. It was implemented in several steps, both in hardware and software, on the host and the storage side, over a couple of years. At the time of writing, zHPF is used by most access methods and all DB2® workloads use it. For a more detailed description and how to enable or disable the feature, see *IBM DS8000 and IBM Z Synergy*, REDP-5186.

zHPF Extended Distance II: With DS8000 and IBM z Systems current models, zHPF is further improved to deliver better response times for long write operations, as used, for example, by DB2 utilities, at greater distances. It reduces the required round trips on the FICON link. This benefits environments that are IBM HyperSwap enabled and where the auxiliary storage system is further away.

zHyperLink: Introduced with IBM z14, zHyperLink provides a short distance direct connection of up to 150 meters (492 feet) to the DS8000. zHyperLink introduced a new synchronous I/O paradigm that eliminates z/OS dispatcher delays, I/O interrupt processing, and the time needed to reload the processor caches that occurs after regaining control of the processor when I/O completes. See “IBM zHyperLink” on page 254.

For more information about the performance implications of host connectivity, see 8.2.4, “Example: AIX” on page 195 and 4.10.1, “Fibre Channel host adapters: 16 Gbps and 32 Gbps” on page 100.

Logical configuration

This section provides z Systems specific considerations for the logical configuration of a DS8000 storage system. For a detailed and host system independent description of this topic, see Chapter 4, “Logical configuration performance considerations” on page 59.

> **Note:** The DS8000 architecture is symmetrical, based on the two CPCs. Many resources, like cache, device adapters (DAs), and RAID arrays, become associated to a CPC. When defining a logical configuration, you assign each array to one of the CPCs. Make sure to spread them evenly, not only by count, but also by their capabilities. The I/O workload must also be distributed across the CPCs as evenly as possible.
>
> The preferred way to achieve this situation is to create a symmetrical logical configuration.

A fundamental question that you must consider is whether there is any special workload that must be isolated, either because it has high performance needs or because it is of low importance and should never influence other workloads.

Sharing or isolating resources

One of the most important aspects during logical configuration is to spread the I/O workload across all available DS8000 resources. Regardless of whether you assign isolated resources to your workloads, share everything, or configure a mix of both methods, the resources you assigned must be used as evenly as possible. Otherwise, your sizing will not fit and the overall performance will not be adequate.

The major groups of resources in a DS8000 storage system are the storage resources, such as RAID arrays and DAs, and the connectivity resources, such as HAs. The traditional approach in mainframe environments to assign storage resources is to divide them into the smallest possible entity (single rank extent pool) and distribute the workload either manually or managed by the Workload Manager (WLM) and System Managed Storage (SMS) as evenly as possible. This approach still has its advantages:

- RMF data provides granular results, which can be linked directly to a resource.
- If you detect a resource contention, you can use host system tools to fix it, for example, by moving a data set to a different volume in a different pool.
- It is easy to detect applications or workloads that cause contention on a resource.
- Isolation of critical applications is easy.

On the contrary, this approach comes with some significant disadvantages, especially with modern storage systems that support automated tiering and autonomous balancing of resources:

- Statically assigned resources might be over- or underutilized for various reasons:
  - Monitoring is infrequent and only based on events or issues
  - Too many or too few resources are allocated to certain workloads
  - Workloads change without resources being adapted
- All rebalancing actions can be performed only on a volume level
- Modern automatic workload balancing methods cannot be used:
  - Storage pool striping
  - Easy Tier automatic tiering
  - Easy Tier intra-tier rebalancing

It is preferred practice to share as many storage resources as possible and let the storage system take care of the balancing and tiering. There might be situations where resource or application isolation is beneficial. In such cases, you must take great care to assign sufficient resources to the isolated application to fulfill its requirements. Isolation alone does not solve a resource constrain issue.

For the host connectivity resources (HAs and FICON links), similar considerations apply. You can share FICON connections by defining them equally for all accessed LCUs in the z Systems I/O definitions. That way, the z Systems I/O subsystem takes care of balancing the load over all available connections. If there is a need to isolate a certain workload, you can define specific paths for their LCUs and volumes. This is a normal requirement if the storage system is used in a multi-tenancy environment.

Volume sizes

The DS8000 storage system now supports CKD logical volumes of any size 1 - 1182006 cylinders, which is 1062 times the capacity of a 3390-1 (1113 cylinders).

> **Note:** The DS8000 storage system provides options for allocating storage with a granularity of:
>
> - Large extents which is the equivalent of 1113 cylinders.
> - Small extents which are 21 cylinders.
>
> Choosing large or small extents is dependent on requirements. Small extents provide better capacity utilization, particularly for thin-provisioned storage but the management of many small extents can cause small performance degradation during initial allocation. For example, a format write of 1 GB requires one storage allocation with large extents, but 64 storage allocations with small extents. Otherwise, host performance should not be adversely affected have a small performance.
>
> For System Z workloads, it is best to choose small extents.

When planning the CKD volume configuration and sizes, a key factor to consider is the limited number of devices a z/OS system can address within one Subchannel Set (65,535). You must define volumes with enough capacity so that you satisfy you storage requirements within this supported address range, including room for PAV aliases and future growth.

Apart from saving device addresses, using large volumes brings additional benefits:

- Simplified storage administration
- Reduced number of X37 abends and allocation failures because of larger pools of free space
- Reduced number of multivolume data sets to manage

One large volume performs the same as though you allocated the same capacity in several smaller ones, if you use the DS8000 built-in features to distribute and balance the workload across resources. There are two factors to consider to avoid potential I/O bottlenecks when using large volumes:

- Use HyperPAV/SuperPAV to reduce IOSQ.

  With equal I/O density (I/Os per GB), the larger a volume, the more I/Os it gets. To avoid excessive queuing, the use of PAV is of key importance. With HyperPAV, you can reduce the total number of alias addresses because they are assigned automatically as needed. SuperPAV goes one step further by using aliases among "like" (=same DS8000 cluster affinity) control units. For more information about the performance implications of PAV, see “Parallel Access Volumes” on page 253.

- Eliminate unnecessary reserves.

  As the volume sizes grow larger, more data on a single CKD device address will be accessed in parallel. There is a danger of performance bottlenecks when frequent activities reserve an entire volume or its VTOC/VVDS. For the most part, activities/batch jobs requiring volume reserves are run during low utilization/overnight periods. This is a mitigation, but the overall issue remains and needs to be managed as part of general housekeeping.

Parallel Access Volumes

PAVs allow multiple concurrent I/Os to the same volume at the same time from applications that run on the same z/OS system image. This concurrency helps applications better share the logical volumes with reduced contention. The ability to send multiple concurrent I/O requests to the same volume nearly eliminates I/O queuing in the operating system, thus reducing I/O responses times.

PAV is implemented by defining alias addresses to the conventional base address. The alias address provides the mechanism for z/OS to initiate parallel I/O to a volume. An alias is another address/UCB that can be used to access the volume that is defined on the base address. An alias can be associated with a base address that is defined in the same LCU only. The maximum number of addresses that you can define in an LCU is 256. Theoretically, you can define one base address plus 255 aliases in an LCU.

Aliases are initially defined to be associated to a specific base address. In a traditional static PAV environment, the alias is always associated with the same base address, which requires many aliases and manual tuning.

In Dynamic PAV or HyperPAV environments, an alias can be reassigned to any base address as your needs dictate. Therefore, you need less aliases and no manual tuning.

With dynamic PAV, the z/OS WLM takes care of the alias assignment. Therefore, it determines the need for additional aliases at fixed time intervals and adapts to workload changes rather slowly.

The more modern approach of HyperPAV assigns aliases in real-time, based on outstanding I/O requests to a volume. The function is performed by the I/O subsystem with the storage system. HyperPAV reacts immediately to changes. With HyperPAV, you achieve better average response times and higher total throughput. Today, there is no technical reason anymore to use either static or dynamic PAV.

With the introduction of SuperPAV (an extension of HyperPAV), the pool of aliases has increased since aliases of “like” control units can be used to drastically reduce IOS queue time (IOSQ). The word “like” means that the even LCUs go with other even LCUs, and the odd LCUs go with other odd LCUs. Therefore all aliases in the even LCUs will be pooled together and can be used by any even LCU when required. The same paradigm will apply to the odd LCUs.

You can check the usage of alias addresses by using RMF data. Chapter 10.1.4, “I/O Queuing Activity report” on page 234 provides an overview of SuperPAV and also provides sample reports showing SuperPAV reporting. You can use such reports to determine whether you assigned enough alias addresses for an LCU.

Number of aliases: With HyperPAV/SuperPAV, you need fewer aliases than with the older PAV algorithms. Assigning 32 aliases per LCU is a good starting point for most workloads. It is a preferred practice to leave a certain number of device addresses in an LCU initially unassigned in case it turns out that a higher number of aliases is required.

Special IBM zSynergy features

This section describes how certain features can help you avoid or overcome I/O performance issues.

IBM zHyperwrite

IBM zHyperWrite is a technology that is provided by the DS8000 storage system, and used by DFSMS to accelerate Db2 and IMS log writes in a HyperSwap enabled Metro Mirror environment. The environment can be managed by GDPS or CSM. zHyperWrite is also supported in Multi-Target Metro Mirror (MTMM) environments and can be used in each relationship independently, depending on the state of the relationship.

When an application sends a write I/O request to a volume that is in synchronous data replication, the response time is affected by the latency caused by the replication management. The distance between the source and the target disk control units also adds to the latency.

With zHyperWrite, Database Log writes for Db2 and IMS are not replicated by PPRC but with the help of Media Manager (DHFSMS), the writes are written to the primary and secondary volumes simultaneously by the host itself. The application, DFSMS, the I/O subsystem, and the DS8000 storage system are closely coordinating the process to maintain data consistency.

Note: At the time of writing, only Db2 and IMS uses zHyperwrite for log writes.

IBM zHyperLink

zHyperLink is a ultra-low latency, synchronous I/O interface which is point-to-point connected (up to a maximum cable length of 150 meters or 492 feet) to an IBM z System server (z14 onwards). This feature whilst representing a substantial leap over zHPF, it does not replace FICON. Instead, zHyperLink works with FICON and zHPF to bring about significant application latency reductions.

Low latencies are provided for read and write operations for storage systems by using a point-to-point link from IBM Z to the storage system’s I/O bay. Low I/O latencies deliver value through improved workload elapsed times and faster transactional response times. The DS8000 implementation of zHyperLink I/O delivers service times fast enough to enable a synchronous I/O model in high performance IBM Z Servers. zHyperLink speeds Db2 for z/OS transaction processing and improve logs throughput.

Synchronous I/O, in a nutshell, means that the entire path handling an I/O request stays within the process context that initiated the I/O. When synchronous I/O is performed, the CPU waits or “spins” until the I/O is completed, or the time-out value is reached. zHyperLink can significantly reduce the time that is required to complete the I/O because the dispatching, interrupt handling, CPU queue time, and CPU cache reload activities are no longer necessary

zHyperLink is fast enough the CPU can just wait for the data:

- No Un-dispatch of the running task.
- No CPU Queueing Delays to resume it.
- No host CPU cache disruption.
- Very small I/O service time.
- Operating System and Middleware (e.g. Db2) are changed to keep running over an I/O.

> **Note:** Not all I/Os are eligible for zHyperLink. Additionally, if a zHyperlink I/O is not successful (for example, because of a read cache miss), the I/O is rerouted over FICON zHPF.

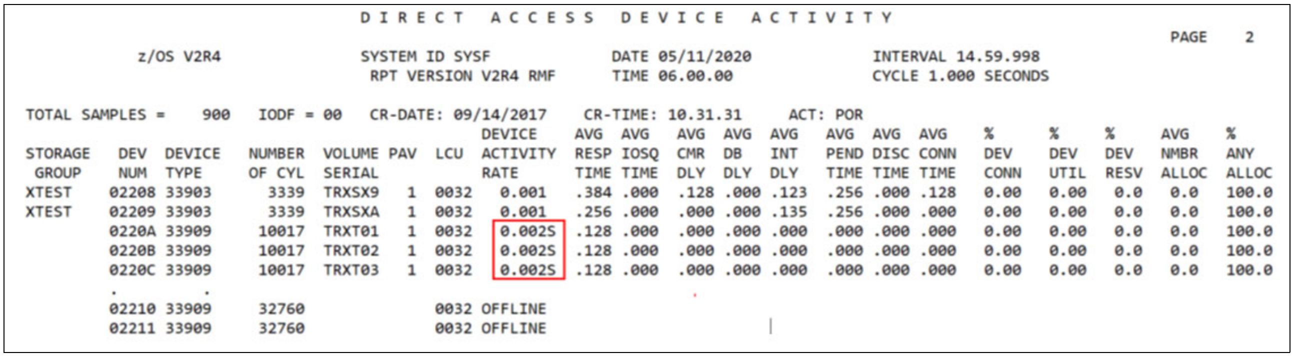
For more information about zHyperLink, see *Getting Started with IBM zHyperLink for z/OS*, REDP-5493.

zHyperLink RMF Reports

DASD devices that actively performed synchronous I/O requests are marked by the character S, following the device activity rate in the Direct Access Device Activity report as shown in Figure 10-11.

RMF is enhanced to monitor and report on performance of:

- PCIE Synchronous I/O links, defined by a zHyperlink port of a server, a fiber optic cable and the target port on a connected storage controller
- DASD devices that actively perform synchronous I/O reads and writes

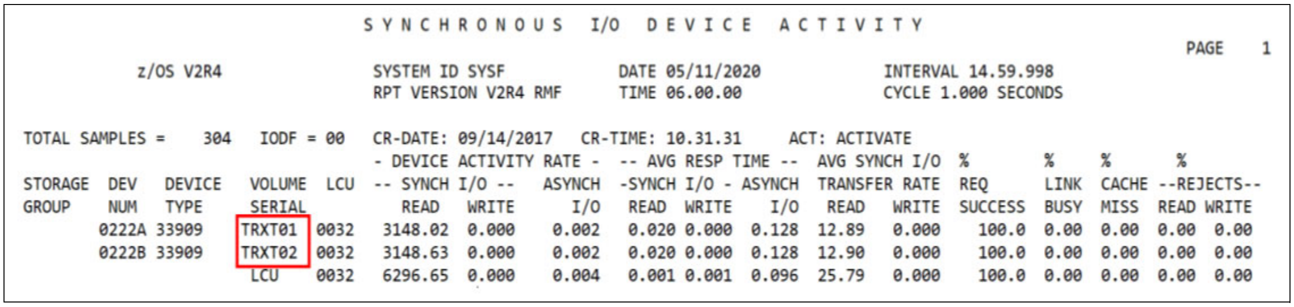


```
                                 D I R E C T   A C C E S S   D E V I C E   A C T I V I T Y
                                                                                                         PAGE    2
           z/OS V2R4              SYSTEM ID SYSF          DATE 05/11/2020         INTERVAL 14.59.998
                                   RPT VERSION V2R4 RMF    TIME 06.00.00           CYCLE 1.000 SECONDS

 TOTAL SAMPLES =     900   IODF = 00   CR-DATE: 09/14/2017   CR-TIME: 10.31.31    ACT: POR
                                                DEVICE   AVG  AVG   AVG  AVG  AVG   AVG  AVG  AVG    %      %     %    AVG    %
 STORAGE   DEV  DEVICE   NUMBER  VOLUME PAV  LCU  ACTIVITY RESP IOSQ  CMR  DB   INT   PEND DISC CONN   DEV    DEV   DEV  NMBR   ANY
  GROUP    NUM  TYPE     OF CYL  SERIAL           RATE     TIME TIME  DLY  DLY  DLY   TIME TIME TIME   CONN   UTIL  RESV ALLOC  ALLOC
 XTEST    02208 33903      3339  TRXSX9   1  0032   0.001  .384 .000  .128 .000 .123  .256 .000 .128   0.00   0.00   0.0  0.0   100.0
 XTEST    02209 33903      3339  TRXSXA   1  0032   0.001  .256 .000  .000 .000 .135  .256 .000 .000   0.00   0.00   0.0  0.0   100.0
          0220A 33909     10017  TRXT01   1  0032   0.002S .128 .000  .000 .000 .000  .000 .000 .000   0.00   0.00   0.0  0.0   100.0
          0220B 33909     10017  TRXT02   1  0032   0.002S .128 .000  .000 .000 .000  .000 .000 .000   0.00   0.00   0.0  0.0   100.0
          0220C 33909     10017  TRXT03   1  0032   0.002S .128 .000  .000 .000 .000  .000 .000 .000   0.00   0.00   0.0  0.0   100.0
            .       .
          02210 33909     32760              0032 OFFLINE
          02211 33909     32760              0032 OFFLINE
```

Figure 10-11 RMF Postprocessor Device Activity Report with I/O devices performing Synchronous I/O

Synchronous I/O Device Activity report is available if at least one DASD device actively performed synchronous I/O requests using IBM zHyperLink. Detailed synchronous I/O performance statistics for the marked DASD devices are displayed in the Synchronous I/O Device Activity section of the report, as shown in Figure 10-12 on page 256. For easy comparison this report lists the devices' activity rate and average response time for asynchronous I/O requests adjacent to the appropriate measurements for synchronous I/O read and write requests and provides additional rates and percentages related to synchronous I/O processing. A device with synchronous I/O activity may be mapped back to the synchronous I/O link by which it is reached by looking up the serial number and node descriptor information of the device's storage controller in the RMF Cache Subsystem Device Overview report.



```
                               S Y N C H R O N O U S   I / O   D E V I C E   A C T I V I T Y
                                                                                                        PAGE    1
           z/OS V2R4               SYSTEM ID SYSF          DATE 05/11/2020             INTERVAL 14.59.998
                                   RPT VERSION V2R4 RMF    TIME 06.00.00               CYCLE 1.000 SECONDS

TOTAL SAMPLES =    304   IODF = 00    CR-DATE: 09/14/2017   CR-TIME: 10.31.31     ACT: ACTIVATE
                                      - DEVICE ACTIVITY RATE -  -- AVG RESP TIME --  AVG SYNCH I/O  %        %    %       %
STORAGE  DEV   DEVICE    VOLUME  LCU  -- SYNCH I/O --   ASYNCH  -SYNCH I/O - ASYNCH  TRANSFER RATE  REQ      LINK  CACHE --REJECTS--
GROUP    NUM   TYPE      SERIAL          READ   WRITE      I/O   READ  WRITE    I/O   READ    WRITE  SUCCESS  BUSY  MISS  READ WRITE
      0222A 33909     TRXT01  0032   3148.02  0.000    0.002  0.020 0.000  0.128  12.89    0.000    100.0  0.00  0.00  0.00  0.00
      0222B 33909     TRXT02  0032   3148.63  0.000    0.002  0.020 0.000  0.128  12.90    0.000    100.0  0.00  0.00  0.00  0.00
                       LCU    0032   6296.65  0.000    0.004  0.001 0.001  0.096  25.79    0.000    100.0  0.00  0.00  0.00  0.00
```

Figure 10-12 RMF Synchronous I/O Device Activity Report

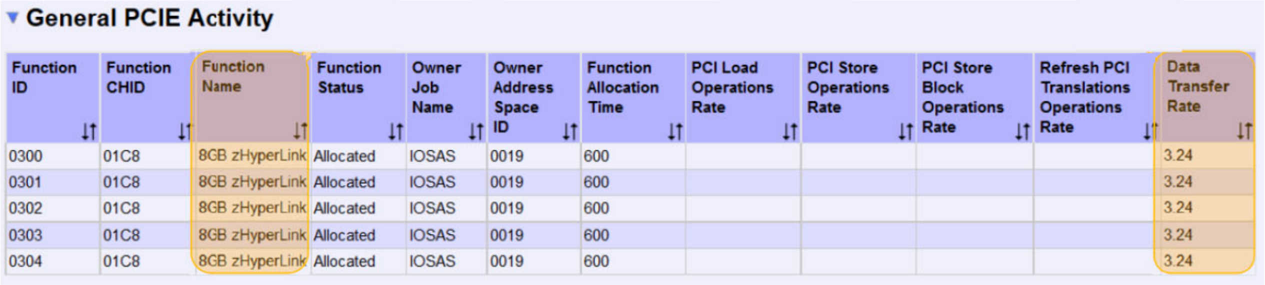
RMF collects and reports new zHyperlink synchronous I/O statistics in following reports:

1. PCIE Activity Report. Figure 10-13
   – The RMF Monitor III PCIE Activity Report is generated using SMF record type 74.9.
2. Synchronous I/O Link Activity. Figure 10-14 on page 257
   – The synchronous I/O statistics are stored in SMF 74-1 and SMF 74-9.
3. Synchronous I/O Response Time Distribution. Figure 10-15 on page 257

General PCIE Activity

In the General PCIE Activity section of the report, zHyperLinks that were active on the system are displayed as PCIe function name 8GB zHyperLink. The Read and write transfer rates are combined in a single Data Transfer Rate value.

Figure 10-13 shows the output of a PCIE report extracted using RMF Data Portal for z/OS. For more information see, “RMF Data Portal for z/OS” on page 242.



▾ General PCIE Activity

Function ID	Function CHID	Function Name	Function Status	Owner Job Name	Owner Address Space ID	Function Allocation Time	PCI Load Operations Rate	PCI Store Operations Rate	PCI Store Block Operations Rate	Refresh PCI Translations Operations Rate	Data Transfer Rate
0300	01C8	8GB zHyperLink	Allocated	IOSAS	0019	600					3.24
0301	01C8	8GB zHyperLink	Allocated	IOSAS	0019	600					3.24
0302	01C8	8GB zHyperLink	Allocated	IOSAS	0019	600					3.24
0303	01C8	8GB zHyperLink	Allocated	IOSAS	0019	600					3.24
0304	01C8	8GB zHyperLink	Allocated	IOSAS	0019	600					3.24

Figure 10-13 General PCIE Activity Report

Synchronous I/O Link Activity

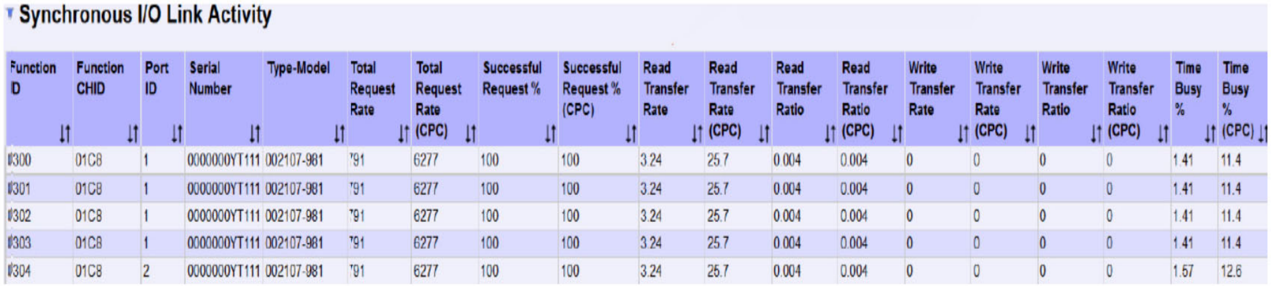
The Synchronous I/O Link Activity section has either system scope (metrics showing values per function) or CPC scope (metrics showing values with a CPC wide view).

For IBM zHyperLinks the section displays:

- Interconnection metrics: These are metrics like port ID, serial number, type and model of the storage controller the synchronous I/O link is connected to.
- Transfer metrics,. For example: data transfer rate.
- Metrics on requests processed, like percentage of successful requests.
- Utilization metrics, like time busy percentage.

> **Note:** Values on CPC level are only reported if Global Performance Reporting is enabled in the LPAR image profile of the Hardware Management Console (HMC).

The storage controller's serial number and type-model can be looked up in the Synchronous I/O Link Activity section of the RMF Postprocessing PCIE Activity report to help identify the appropriate synchronous I/O link, as shown in Figure 10-14



Synchronous I/O Link Activity

Function ID	Function CHID	Port ID	Serial Number	Type-Model	Total Request Rate	Total Request Rate (CPC)	Successful Request %	Successful Request % (CPC)	Read Transfer Rate	Read Transfer Rate (CPC)	Read Transfer Ratio	Read Transfer Ratio (CPC)	Write Transfer Rate	Write Transfer Rate (CPC)	Write Transfer Ratio	Write Transfer Ratio (CPC)	Time Busy %	Time Busy % (CPC)
0300	01C8	1	0000000YT111	002107-981	791	6277	100	100	3.24	25.7	0.004	0.004	0	0	0	0	1.41	11.4
0301	01C8	1	0000000YT111	002107-981	791	6277	100	100	3.24	25.7	0.004	0.004	0	0	0	0	1.41	11.4
0302	01C8	1	0000000YT111	002107-981	791	6277	100	100	3.24	25.7	0.004	0.004	0	0	0	0	1.41	11.4
0303	01C8	1	0000000YT111	002107-981	791	6277	100	100	3.24	25.7	0.004	0.004	0	0	0	0	1.41	11.4
0304	01C8	2	0000000YT111	002107-981	791	6277	100	100	3.24	25.7	0.004	0.004	0	0	0	0	1.57	12.6

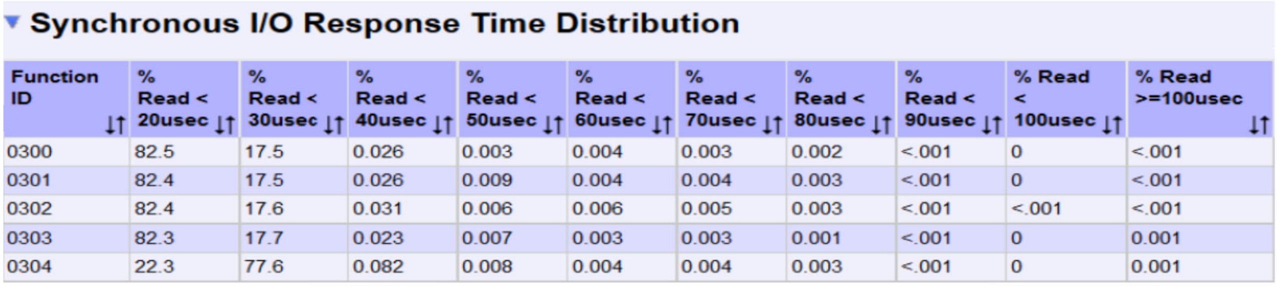
Figure 10-14 Synchronous I/O Link Activity Report

Synchronous I/O Response Time Distribution

The Synchronous I/O Response Time Distribution section provides a response time distribution overview for read and write instructions executed on the allocated synchronous I/O links. The response time of synchronous I/Os are sampled in 10 buckets: Left-most bucket: Percentage of I/Os with a response time less than n microseconds. Right-most bucket: Percentage of I/Os with a response time greater or equal n microseconds.

Other buckets: Percentage of I/Os with a response time less than n microseconds and greater or equal to the prior bucket time limit.

Example: % Read <30 msec = 17.5 means that 17.5 percent of the read I/Os had a response time of more than or equal to 20 microseconds but less than 30 microseconds. Refer to Figure 10-15.



Synchronous I/O Response Time Distribution

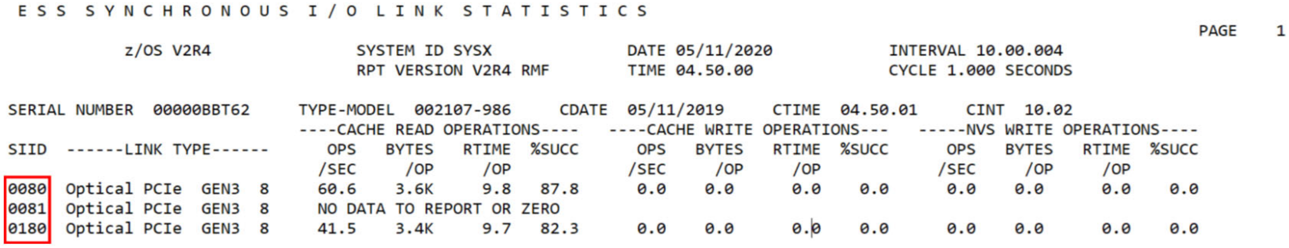
Function ID	% Read < 20usec	% Read < 30usec	% Read < 40usec	% Read < 50usec	% Read < 60usec	% Read < 70usec	% Read < 80usec	% Read < 90usec	% Read < 100usec	% Read >=100usec
0300	82.5	17.5	0.026	0.003	0.004	0.003	0.002	<.001	0	<.001
0301	82.4	17.5	0.026	0.009	0.004	0.004	0.003	<.001	0	<.001
0302	82.4	17.6	0.031	0.006	0.006	0.005	0.003	<.001	<.001	<.001
0303	82.3	17.7	0.023	0.007	0.003	0.003	0.001	<.001	0	0.001
0304	22.3	77.6	0.082	0.008	0.004	0.004	0.003	<.001	0	0.001

Figure 10-15 Synchronous I/O Response Time Distribution

RMF is enhanced to monitor and report on performance of:

- PCIE Synchronous I/O links, defined by a zHyperlink port of a server, a fiber optic cable and the target port on a connected storage controller.
- DASD devices that actively perform synchronous I/O reads and writes.

The RMF Postprocessor ESS Report is collected by RMF Monitor I and reported in the ESS Synchronous I/O Link Statistics report section of RMF Postprocessor Enterprise Disk Systems (ESS) report, provides new performance metrics for the IBM zHyperLink connections to the DS8000. Figure 10-16 on page 258 shows a Hardware (DS8000) view of synchronous I/O activity over the zHyperLink.

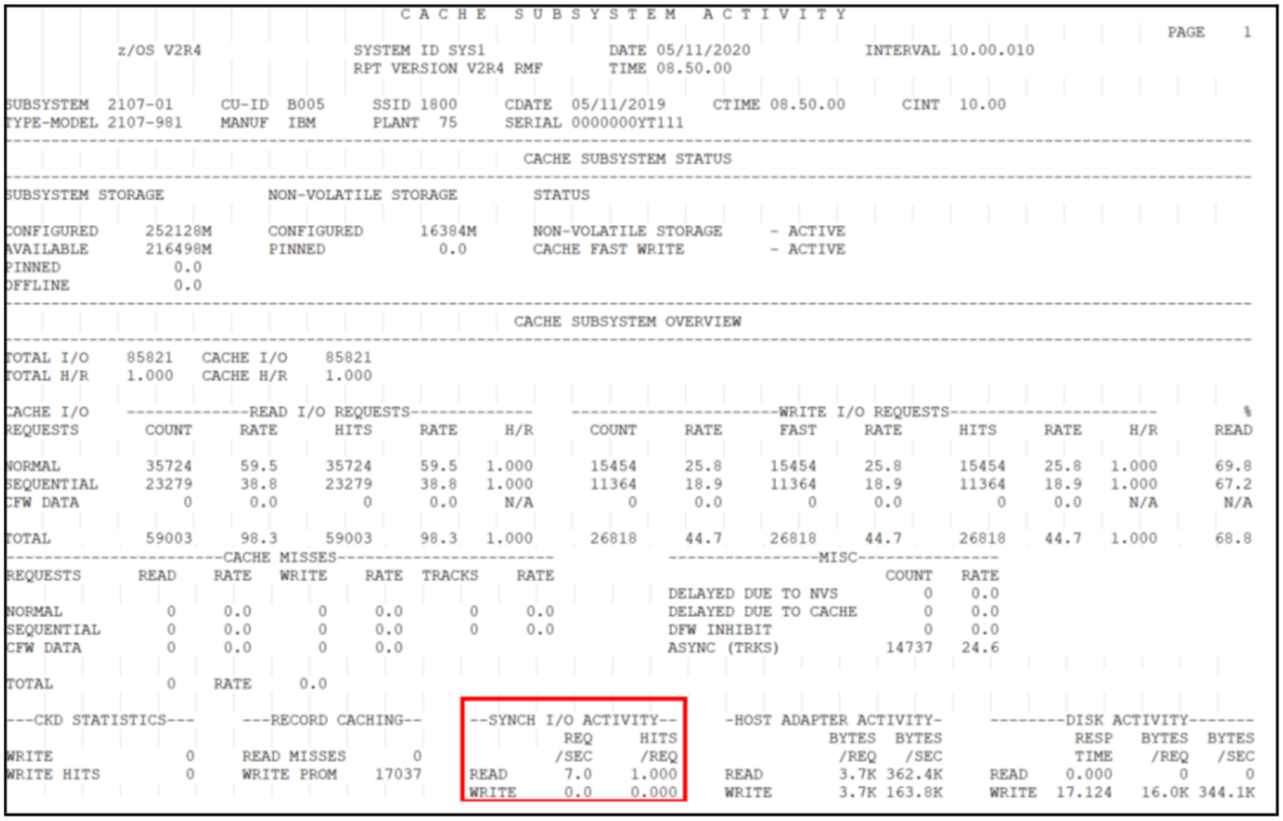


```
 E S S  S Y N C H R O N O U S  I / O  L I N K  S T A T I S T I C S
                                                                                                                 PAGE    1
           z/OS V2R4                SYSTEM ID SYSX             DATE 05/11/2020            INTERVAL 10.00.004
                                    RPT VERSION V2R4 RMF       TIME 04.50.00              CYCLE 1.000 SECONDS

SERIAL NUMBER  00000BBT62     TYPE-MODEL  002107-986     CDATE  05/11/2019     CTIME  04.50.01     CINT  10.02
                              ----CACHE READ OPERATIONS----   ----CACHE WRITE OPERATIONS---   -----NVS WRITE OPERATIONS----
SIID   ------LINK TYPE------    OPS    BYTES   RTIME  %SUCC     OPS    BYTES   RTIME  %SUCC     OPS    BYTES   RTIME  %SUCC
                               /SEC      /OP     /OP           /SEC      /OP     /OP           /SEC      /OP     /OP
0080  Optical PCIe  GEN3  8    60.6     3.6K     9.8   87.8     0.0     0.0      0.0    0.0     0.0     0.0      0.0    0.0
0081  Optical PCIe  GEN3  8    NO DATA TO REPORT OR ZERO
0180  Optical PCIe  GEN3  8    41.5     3.4K     9.7   82.3     0.0     0.0      0.0    0.0     0.0     0.0      0.0    0.0
```

Figure 10-16 ESS Synchronous I/O Link Statistics

The RMF Postprocessor Cache Subsystem Activity (CACHE) report is enhanced to provide performance metrics for synchronous I/O over IBM zHyperLink. Synchronous I/O performance statistics are displayed in the Cache Subsystem Overview and Cache Device Activity sections of the report as shown in Figure 10-17. A set of new OVW conditions was introduced which can be used to report the new synchronous I/O metrics in SMF 74-5 and SMF 74-8 records.



```
                                         C A C H E   S U B S Y S T E M   A C T I V I T Y
                                                                                                            PAGE    1
            z/OS V2R4                 SYSTEM ID SYS1              DATE 05/11/2020            INTERVAL 10.00.010
                                      RPT VERSION V2R4 RMF        TIME 08.50.00

SUBSYSTEM  2107-01      CU-ID  B005      SSID 1800      CDATE  05/11/2019     CTIME 08.50.00      CINT  10.00
TYPE-MODEL 2107-981     MANUF  IBM       PLANT  75      SERIAL 0000000YT111
------------------------------------------------------------------------------------------------------------------------------
                                                   CACHE SUBSYSTEM STATUS
------------------------------------------------------------------------------------------------------------------------------
SUBSYSTEM STORAGE          NON-VOLATILE STORAGE         STATUS

CONFIGURED      252128M    CONFIGURED       16384M      NON-VOLATILE STORAGE     - ACTIVE
AVAILABLE       216498M    PINNED              0.0      CACHE FAST WRITE         - ACTIVE
PINNED              0.0
OFFLINE             0.0
------------------------------------------------------------------------------------------------------------------------------
                                                  CACHE SUBSYSTEM OVERVIEW
------------------------------------------------------------------------------------------------------------------------------
TOTAL I/O     85821    CACHE I/O     85821
TOTAL H/R     1.000    CACHE H/R     1.000

CACHE I/O     ------------READ I/O REQUESTS------------    --------------------WRITE I/O REQUESTS--------------------       %
REQUESTS       COUNT      RATE      HITS     RATE    H/R      COUNT     RATE     FAST     RATE      HITS     RATE    H/R     READ

NORMAL         35724      59.5     35724     59.5  1.000      15454     25.8    15454     25.8     15454     25.8  1.000     69.8
SEQUENTIAL     23279      38.8     23279     38.8  1.000      11364     18.9    11364     18.9     11364     18.9  1.000     67.2
CFW DATA           0       0.0         0      0.0    N/A          0      0.0        0      0.0         0      0.0    N/A      N/A

TOTAL          59003      98.3     59003     98.3  1.000      26818     44.7    26818     44.7     26818     44.7  1.000     68.8
----------------------CACHE MISSES---------------------         ---------------MISC--------------
REQUESTS       READ    RATE    WRITE     RATE  TRACKS     RATE                                      COUNT    RATE
                                                                DELAYED DUE TO NVS                      0     0.0
NORMAL            0     0.0        0      0.0       0      0.0  DELAYED DUE TO CACHE                    0     0.0
SEQUENTIAL        0     0.0        0      0.0       0      0.0  DFW INHIBIT                             0     0.0
CFW DATA          0     0.0        0      0.0                   ASYNC (TRKS)                        14737    24.6

TOTAL             0    RATE      0.0

---CKD STATISTICS---     ---RECORD CACHING--     --SYNCH I/O ACTIVITY--     -HOST ADAPTER ACTIVITY-     --------DISK ACTIVITY-------
                                                              REQ      HITS                BYTES   BYTES              RESP   BYTES   BYTES
WRITE                0   READ MISSES        0                /SEC      /REQ                 /REQ    /SEC              TIME    /REQ    /SEC
WRITE HITS           0   WRITE PROM     17037     READ        7.0     1.000     READ        3.7K  362.4K    READ     0.000       0       0
                                                  WRITE       0.0     0.000     WRITE       3.7K  163.8K    WRITE   17.124   16.0K  344.1K
```

Figure 10-17 Cache Subsystem Activity / Cache Subsystem Overview showing Synch I/O Activity

Copy Services considerations

The performance aspects of all DS8000 Copy Services can vary greatly and are dependent on workload/data center setup. Here are some general aspects to consider from a performance point of view:

- Remote data replication, such as Metro Mirror, workload can be accounted for in StorM. However, not all parameters and values are directly derived from the RMF data. Make sure all values that are used in the modeling are correct.

Remote data replication is used universally and among all the available types of replication, it is synchronous data replication (Metro Mirror) which can have a performance impact due to the distance between data centers and the ever present/never changing speed of light. To help mitigate the performance implications, the DS8000 replication firmware includes enhancements to the standard fiber channel protocol where pre-deposit writes improve the standard FCP and optimize the number of protocol handshakes during an MM I/O operation. Several I/O requests are sent before a handshake signals to the sending storage system that everything arrived safely at the secondary site.

The number of protocol exchanges that is needed to run the I/O chain is minimized compared to standard FCP. The goal here is to combine several aspects, such as how many I/Os to send before the handshake, considering the resources that are needed for buffers, and how smart error detection and error recovery are to run at the maximal speed and performance.

- FlashCopy workload is not reflected in the RMF data that is used as StorM or Disk Magic input. If there is significant FlashCopy activity, a thorough analysis is needed to ascertain current back-end utilization and use that information to provide enough headroom for back-end resources.

10.3 Problem determination and resolution

This section provides some basic tips and methods for problem determination and troubleshooting for when you experience insufficient performance that appears storage system-related.

This is only an introduction. It cannot replace a thorough analysis by IBM Technical Support in more complex situations or if there are product issues.

10.3.1 Sources of information

If your client or the business you provide IT services for approaches you with a performance issue, you most likely get statements like the following ones:

- “We cannot complete our jobs in the batch window.”
- “The backup takes too long.”
- “Users complain about the response time of application XY.”

Being responsible for the I/O part of your infrastructure, you must convert such general statements to I/O terms and discover whether the problem is I/O-related. You need more information from several instances. First, get as much detail as possible about how the issue appears. Ask the client or user who reports the problem:

- How is the issue perceived?
- At which times does it occur?
- Is it reproducible?
- Does it show up under specific circumstances?
- What kind of workload is running at the time?
- Was there already any analysis that links the issue to I/O? If yes, get the details.
- Was anything changed, either before or after the issue started to appear?

Next, gather performance data from the system. For I/O related investigations, use RMF data. For a description about how to collect, process, and interpret RMF data, see 10.1, “DS8000 performance monitoring with RMF” on page 226. There might be many data that you must analyze. The faster you can isolate the issue up front, both from a time and a device point of view, the more selective your RMF analysis can be.

There are other tools or methods that you can use to gather performance data for a DS8000 storage system. They most likely are of limited value in a mainframe environment because they do not take the host system into account. However, they can be useful in situations where RMF data does not cover the complete configuration, for example:

- Mixed environments (mainframe, open, and IBM i)
- Special copy services configurations, such as Global Mirror secondary

For more information about these other tools, see 10.1.9, “Alternatives and supplements to RMF” on page 242.

To match physical resources to logical devices, you also need the exact logical configuration of the affected DS8000 storage systems, and the I/O definition of the host systems you are analyzing.

10.3.2 Identifying critical and restrained resources

With the information you collected, you can try to discover whether there is an I/O-related issue, and if yes, isolate the responsible resources. In the RMF data you collected, look for periods and resources with outlier/out of normal range values:

- High response time
- Exceptionally high or low throughput/utilization
- High elapsed time for jobs

The following sections point you to some key metrics in the RMF reports, which might help you isolate the cause of a performance issue.

RMF Summary Report

Attention: RMF provides many different measurements. It can be complex to associate them with conditions that lead to performance issues. This section can cover only the most common symptoms in this section. If the matter is too complex or you need a more general analysis of your performance situation, consider a performance study. IBM offers this study as a charged service. Contact your IBM representative or IBM Business Partner if you are interested. Similar services might also be available from other companies.

The RMF Summary Report gives an overview over the system load. From an I/O perspective, it provides only LPAR-wide I/O rate and average DASD response time. Use it to get an overall impression and discover periods of peak I/O load or response time. See whether these periods match the times for which the performance problems are reported.

Attention: Because the summary report provides a high-level overview, there might be issues with individual components that are not directly visible here.

Direct Access Device Activity

This report shows the activity of all DASD devices within the LPAR scope. Search this report for devices with high response times. The term high in this context means either of the following items:

- Higher than other devices
- Higher than usual for certain times

After you isolate a certain time and a set of volumes that are conspicuous, you can analyze further. Discover which of the response time components are higher than usual. A description of these components and why they can be increased is provided in 10.1.3, "I/O response time components" on page 231.

If you also need this information on a Sysplex scope, create the Shared Direct Access Device Activity report. It provides a similar set of measurements for each volume by LPAR and also summarized for the complete Sysplex.

> **Important:** Most of the time, devices with no or almost no activity should not be considered. Their response time values are not relevant and might be inaccurate. This rule can also be applied to all other reports and measurements.
>
> The exception is when the issue is outbound to the storage system such as on the SAN/switching layer, then the Device utilization can show up as very low or no activity. In this case, the SAN/switching layer needs to be interrogated to identify the cause for low utilization. Port Statistics on the switch will provide clues if error counters are incrementing at a substantial rate over a short period of time.

Cache Subsystem Summary

This report provides cache statistics at a volume or LCU level. RMF gathers this information directly from the storage system. You can use it to analyze important values like read/write ratio, read cache hit ratio, and the Delayed due to NVS rate.

You specifically might want to check the volumes you identified in the previous section and see whether they have the following characteristics:

- They have a high write ratio.
- They show a DELAYED DUE TO NVS rate greater than 0, which indicates that you are running into an NVS Full condition.

Enterprise Storage Server

The data for the Enterprise Storage Server report is also gathered directly from the storage system. It provides measurements for the DS8000 HAs and the disk back end (RAID arrays / ranks).

Use the Enterprise Storage Server Link Statistics to analyze the throughput of the DS8000 HA ports. Pay particular attention to those that have higher response time than others. Also, use the I/O intensity value to determine whether a link might be close to its limitations. All HA ports are listed here. You can also analyze remote replication and Open Systems workload.

The Enterprise Storage Server Rank Statistics show the back-end I/O load for each of the installed RAID arrays. Again, look for those that stand out, either with a high load, or much higher response time.

Channel Path activity

This report contains measurements of all defined z/OS channel paths. You see their utilization and throughput and the activity, where it distinguishes between classic FICON and zHPF. The activity (I/O rate) is the number of FICON and zHPF operations per second, which is different than the device I/O rate. A device I/O requires more several FICON operations.

Look for channel paths that have the following features:

- Have much higher throughput, utilization, or activity than others, which indicates an unbalanced configuration.
- Have high classic FICON activity, which indicates that systems or application do not use the more efficient zHPF protocol.

- Have significant high defer values, which can be an indicator for frame pacing (buffer credit exhaustion at the switch/fabric layer).

I/O Queueing activity

This report consists of two parts. The first part shows the queuing activity and utilization of the z Systems I/O processors. Check whether there are IOPs with are more busy than others.

The second part of the report provides queuing details about an LCU and channel path level. You can see whether I/Os are delayed on their way to the device. Check for LCUs / paths that have the following features:

- Higher average control unit busy delay (AVG CUB DLY), which can mean that devices are in use or reserved by another system.
- Higher average command response delay (AVG CMR DLY), which can indicate a saturation of certain DS8000 resources, such as HA, internal bus, or processor.
- Nonzero HPAV wait times and HPAV max values in the order of the number of defined alias addresses, which can indicate that the number of alias addresses is not sufficient.

10.3.3 Corrective actions

If you can identify where the performance issue is, you can devise and perform measures to correct it or avoid it in the future. Depending on what the cause is, there are several approaches that you can follow.

In many cases, your analysis shows that one or more resources either on the host system or DS8000 storage system are saturated or overloaded. The first thing to check is whether your storage system is configured to use all available features that improve performance or automate resource balancing.

The most important ones are the following ones:

- Easy Tier automated mode to avoid or eliminate back-end hot spots
- zHPF to improve the efficiency of the FICON links
- HyperPAV to improve parallelism

If these features do not solve the issue, consider the following actions:

- Distribute the workload further over additional existing resources with less utilization.
- Add more resources of the same type, if there is room.
- Exchange the existing, saturated resources for different ones (other or newer technology) with higher capabilities.

If you isolated applications (for example, by using their own set of HPFE ranks) but still experience poor response times, check the following items:

- Are the dedicated resources saturated? If yes, you can add more resources, or consider switching to a shared resource model.
- Is the application doing something that the dedicated resources are not suited for (for example, mostly sequential read and write operations on HPFE ranks)? If yes, consider changing the resource type, or again, switching to a shared model.
- Does the contention come from other resources that are not dedicated, such as HA ports in our example with dedicated HPFE ranks? Here, you can consider increasing the isolation by dedicating host ports to the application as well.

If you are running in a resource sharing model and find that your overall I/O performance is good, but there is one critical application that suffers from poor response times, you can consider moving to an isolation model, and dedicate certain resources to this application. If the issue is limited to the back end, another solution might be to use advanced functions:

- Changes in DHSMS/DFSMSHSM/WLM to adjust/meet performance goals.
- Easy Tier Application to manually assign certain data to a specific storage tier

If you cannot identify a saturated resource, but still have an application that experiences insufficient throughput, it might not use the I/O stack optimally. For example, modern storage systems can process many I/Os in parallel. If an application does not use this capability and serializes all I/Os, it might not get the required throughput, although the response times of individual I/Os are good.

11

Database for IBM z/OS performance

This chapter reviews the major IBM database systems and the performance characteristics and considerations when they are used with the DS8000 storage system. This chapter describes the following databases:

- IBM Db2 in a z/OS environment
- IMS in a z/OS environment

You can obtain additional information about IBM Db2 and IMS at these websites:

- https://www.ibm.com/analytics/db2
- https://www.ibm.com/it-infrastructure/z/ims

This chapter includes the following topics:

- DS8000 considerations for Db2 and IMS

11.1 DS8000 considerations for Db2 and IMS

The DS8000 storage family was designed to have a synergistic relationship with z/OS, Db2, and IMS which enables numerous performance benefits while providing ease of management for business as usual activities while providing ease of management for typical day-to-day business activities.

11.1.1 Balancing workload across DS8000 resources

Configuring only two storage pools (one for CPC or Server 0; the other for CPC or Server 1) and leaving Easy Tier enabled to autonomically and dynamically enforce effective & efficient use of back-end resources; and, spreading Channel Path Groups across multiple DS8000 Host Adapters, will balance workload activity across the DS8000 resources.

11.1.2 Using DFSMS for Db2 and IMS dataset placement

Placing Db2 and IMS datasets using DFSMS (Data Facility Storage Management Subsystem) will ensure that data is placed appropriately using SMS constructs. Combining SMS and EasyTier leads to best of worlds approach where SMS is in control of the datasets and EasyTier looks after I/O activity. This tight integration allows for transparent and best in class service from a DS8000 perspective.

Db2 and IMS data sets can be allocated in the extended addressing space (EAS) of a extended address volume (EAV) to significantly reduce addressing footprint and ongoing management.

> **Note:** The maximum amount of data that you can store in a single Db2 table space or index space is the same for extended and non-extended address volumes. The same Db2 data sets might use more space on extended address volumes than on non-extended address volumes because space allocations in the extended area are multiples of 21 cylinders on extended address volumes. For more information, see:
>
> - https://www.ibm.com/docs/en/db2-for-zos/12?topic=data-general-approach-estimating-storage
> - https://www.ibm.com/docs/en/db2-for-zos/12?topic=requirements-active-log-data-sets-storage

IMS Datasets such as:

- GSAM database data sets
- OSAM database data sets
- VSAM database data sets
- Online log data sets (OLDSs), including large OLDS (for example, greater than 64 KB tracks)
- Write ahead data sets (WADSs)
- Database Image Copy utility (DFSUDMP0) data sets
- Database Image Copy 2 utility (DFSUDMT0) data sets
- MFS map library data sets produced by the MFS Language and Service utilities (IMS.FORMAT)
- IMS Trace facility external trace data sets
- IMS Monitor output data sets

Full Data set list can be reviewed at:

https://www.ibm.com/docs/en/ims/15.2.0?topic=requirements-dasd

For more information on EAV volumes see: *IBM DS8000 and IBM Z Synergy*, REDP-5186-05

11.1.3 Parallel Access Volumes

Db2 and IMS takes advantage of the parallel access volume (HyperPAV/SuperPAV) function that allows multiple I/Os to the same volume concurrently from applications that run on a z/OS system image. This will be helpful if IOSQ time is being experienced in the environment. At a minimum, HyperPAV should be enabled in a System Z environment with periodic checking of RMF reports to ascertain if more HyperPAV devices are required or SuperPAV functionality needs to be enabled.

For more information about IOSQ, see: “IOSQ time” on page 232.

For more information about HyperPAV/SuperPAV and enablement, see: *IBM DS8000 and IBM Z Synergy*, REDP-5186-05

11.1.4 High Performance FICON (zHPF)

High Performance FICON, also known as zHPF, is not new. It was introduced in 2009 and has seen multiple enhancements which has bought about significant performance improvements for Db2 and IMS I/Os.

The HPFEs in the DS8900F storage family deliver performance to such an extent that it leads to more stress on the channel subsystem on z Systems. zHPF in combination with flash provides an effective way to reduce this channel subsystem stress. zHPF should be enabled in the environment.

Latest versions of Db2 (Db2 v12) combined with zHPF and current FICON hardware (FICON Express cards 16S/16S+/16SA) deliver improvements to the following Db2 functions:

- Db2 queries
- Table scans
- Index-to-data access, especially when the index cluster ratio is low
- Index scans, especially when the index is disorganized
- Reads of fragmented large objects (LOBs)
- New extent allocation during inserts
- Db2 REORG
- Sequential reads
- Writes to the shadow objects
- Reads from a non-partitioned index
- Log applies
- Db2 LOAD and REBUILD
- Db2 Incremental COPY
- RECOVER and RESTORE
- Db2 RUNSTATS table sampling

Important: All Db2 I/Os are eligible for zHPF.

11.1.5 Db2 List Prefetch

zHPF is enhanced to support Db2 list prefetch. The enhancements include a new cache optimization algorithm that can greatly improve performance and hardware efficiency. When combined with the latest releases of z/OS and Db2, it can demonstrate up to a 14x - 60x increase in sequential or batch-processing performance.

Db2 typically performs a maximum of two parallel list prefetch I/Os. With List Prefetch, zHPF sends a list of tracks in a Locate Record command followed by several read commands, which allows many parallel retrievals from disk. Also, the use of the Locate Record command now sends a list of non-contiguous records to prefetch, which increases the cache-hit rate for subsequent read commands.

In addition, Db2 can benefit from the new caching algorithm at the DS8900F level, which is called List Prefetch Optimizer (LPO).

For more information about list prefetch, see *DB2 for z/OS and List Prefetch*, REDP-4862.

11.1.6 Db2 Castout Accelerator

In Db2, a castout refers to the process of writing pages from the group buffer pool to disk. Db2 writes long chains that typically contain multiple locate record domains. Traditionally, each I/O in the chain is synchronized individually. This process add to the response time and increases latency.

Now the DS8900F can be notified through the Db2 Media Manager that the multiple I/Os in a castout can be treated as a single logical I/O even though there are multiple embedded I/Os. In other words, the data hardening requirement is for the entire I/O chain. This enhancement brings significant response time reduction.

11.1.7 IBM FlashCopy with Db2

IBM FlashCopy (with all its options) can be used for:

- Db2 image copies using FlashCopy fast replication parameter.
- Data set level FlashCopy for Db2 for z/OS utilities.
- Seamless Point-in-time backup and recovery with Db2 System Level Backup (SLB) and Restore utility.

  With z/OS DFSMS copy pool setup and IBM Fast Replication Backup command FRBACKUP, database and storage administrators can create backup policies for IBM Db2 databases with minimum effect on applications. Db2 uses this function for Db2 system-level backup and to control the serialization effort that ensures a consistent set of the complete Db2 region (including all Db2 table space and logs). Furthermore, Db2 also maintains its repository, for example, by updating the bootstrap data set (BSDS) to reflect all backup activity. Similarly, for recovery DFSMS and IBM Fast Replication Restore command, FRRECOV enables restore of a complete Db2 region but also provides granularity of restoring an individual Db2 object from within the full system backup. The above is made possible due to the close synergy provided between DFSMS components, DFSMSdfp, DFSMShsm, DFSMSdss and Db2, all working together to provide a complete backup and restore package.

  For more information about Copy Pools and setting them up in SMS, see:

  – https://www.ibm.com/docs/en/zos/2.4.0?topic=groups-copy-pools

  – https://www.ibm.com/docs/en/zos/2.4.0?topic=administration-defining-copy-pools

  – *IBM DS8000 and IBM Z Synergy*, REDP-5186-05

11.1.8 IBM FlashCopy with IMS

Using FlashCopy with IMS enables cloning of entire IMS regions for testing purposes with minimal disruption. IMS Database utilities such as Database Image Copy, Online Database Image Copy, Database Recovery all use DFSMSdss to invoke the FlashCopy function. This allows for backup of a database or a subset of data at a point in time with minimum downtime for the database.

For more information see *IMS High Performance Image Copy User's Guide*, SC19-2756-04.

11.1.9 zHyperWrite with Db2

In a synchronous replication (Metro Mirror) environment, all writes (including Db2 log writes) are mirrored synchronously to the secondary device, which increases transaction response times.

With zHyperWrite Db2 log writes are performed to the primary and secondary volumes in parallel, which reduces Db2 log write response times. Implementation of zHyperWrite requires that HyperSwap is enabled through either IBM Geographically Dispersed Parallel Sysplex® (IBM GDPS) or CSM. zHyperWrite combines Metro Mirror (PPRC) synchronous replication and software mirroring through media manager (DFSMS) to provide substantial improvements in Db2 log write latency.

For more information about zHyperWrite (z/OS and Db2 enablement and display), see:

- https://www.ibm.com/docs/en/zos/2.4.0?topic=parameters-hyperwrite
- https://www.ibm.com/support/pages/apar/PI25747
- https://www.ibm.com/docs/en/zos/2.4.0?topic=command-displaying-zhyperwrite-data-replication-status-ioshyperwrite

11.1.10 WADS and OLDS supported by zHyperWrite

The IMS WADS (write-ahead log data set) and OLDS (online log data set) are the most active data sets in an IMS environment as far as write activity is concerned. Because of its nature, the volumes that contain these data sets have the biggest challenge in the storage system. Due to the amount of I/O activity generated to the these volumes, it is strongly recommended to place these data sets on flash drives making the IBM DS8900F family the perfect storage system for this type of workload. It is also strongly recommended to allocate the WADS and OLDS across LCUs to try get the IO evenly balanced at the backend.

Furthermore, IMS v15.1 introduced the use of DFSMS Media Manager to write data to the WADS and OLDS. This change enables IMS to use important I/O features, such as zHPF, which increases I/O throughput, and zHyperWrite, which reduces latency time for synchronous replication.

The use of zHPF and zHyperWrite can be specially useful for data sets with high write rates, such as WADS and OLDS, which increases logging speed. Service times for WADS and OLDS can be reduced by up to 50%, depending on the environment setup.

For more information about zHyperWrite enablement, WADS and OLDS support for zHyperWrite, including migration considerations, see:

- https://www.ibm.com/docs/en/ims/15.1.0?topic=zhyperwrite-wads-support
- https://www.ibm.com/docs/en/ims/15.1.0?topic=zhyperwrite-olds-support
- https://www.ibm.com/docs/en/ims/15.1.0?topic=zhyperwrite-dynamic-enablement-enhancement

11.1.11 zHyperLink

zHyperLink provides Ultra-low latency for database page reads in online transactions using synchronous I/O operations. Compared to a normal asynchronous I/O operation through FICON or zHPF, with zHyperLink there are no I/O interrupts or dispatcher delays. All the overhead associated with FICON and zHPF are avoided entirely leading to less than 20 microseconds of observed latency. This allows for processing of huge volumes of transactions at a much faster rate.

For more information about Db2 v12 enablement, see:

- https://www.ibm.com/docs/en/db2-for-zos/12?topic=panel-db2-zhyperlinks-scope-field-zhyperlink-subsystem-parameter

For more information about zHyperLink (z/OS enablement), see:

- https://www.ibm.com/docs/en/zos/2.4.0?topic=parameters-zhyperlink
- *Getting Started with IBM zHyperLink for z/OS*, REDP-5493.

11.1.12 zHyperLink + zHyperWrite

In a Metro Mirror environment, specifically for writes, both zHyperLink and zHyperWrite can be used together to further provide the benefits of reduced latency for reads (for all VSAM reads) and writes (Db2 v12 log writes and IMS v15 WADS/OLDS support).

For Dynamically changing zHyperLink eligibility of a dataset, see:

- https://www.ibm.com/docs/en/zos/2.4.0?topic=command-changing-zhyperlink-eligibility-data-set
- Extensive and increasing support of zHyperLink and Advanced Copy Services.

> **Note:** At the time of writing, zHyperLink provides writes support for Metro Mirror (MM), Multi-Target Metro Mirror (MTMM), Metro Global Mirror (MGM) and Global Mirror (GM) replication flavors.
>
> For zHyperLink write support in MM, MTMM and MGM, zHyperWrite is also required on MM leg.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- *IBM DS8900F Architecture and Implementation*, SG24-8456
- *IBM DS8000 Copy Services*, SG24-8367
- *IBM DS8000 Easy Tier*, REDP-4667
- *IBM DS8000 and IBM Z Synergy*, REDP-5186
- *IBM DS8000 High-Performance Flash Enclosure Gen2*, REDP-5422
- *Getting Started with IBM zHyperLink for z/OS*, REDP-5439
- *FICON Native Implementation and Reference Guide*, SG24-6266
- *IBM Storage Modeller Guidance*, ZG24-8401

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

These publications are also relevant as further information sources:

- H.A. Fetene, N. Clayton, T. Camacho, D.V. Valverde, S.E. Williams, Y. Xu: *IBM System Storage DS8900F Performance Whitepaper, DS8900F Release R9.0,* WP102814
- C. Gordon: *DS8000 Host Adapter Configuration Guidelines*, https://www.ibm.com/support/pages/ds8000-host-adapter-configuration-guidelines

Online resources

These websites are also relevant as further information sources:

- *IBM Documentation Center*

  https://www.ibm.com/docs/en/ds8900
- *IBM System Storage DS8000 Host Systems Attachment Guide,* SC26-7917

  http://publibfp.dhe.ibm.com/epubs/pdf/c2795631.pdf
- *IBM Storage Modeller Help*

  https://www.ibm.com/tools/storage-modeller/help/

- *IBM Z Batch Network Analyzer (zBNA) Tool*

  https://www.ibm.com/support/pages/ibm-z-batch-network-analyzer-zbna-tool-0
- *IBM i 7.4: Systems management, Disk management*

  https://www.ibm.com/docs/en/ssw_ibm_i_74/rzaly/rzalypdf.pdf
- *MustGather: How to obtain and install QMGTOOLS and keep it current*

  https://www.ibm.com/support/pages/mustgather-how-obtain-and-install-qmgtools-and-keep-it-current

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Redbooks

IBM DS8900F

Performance Best Practices and Monitoring

SG24-8501-00

ISBN 0738460192

(1.5” spine)
1.5”<-> 1.998”
789 <->1051 pages

Redbooks

IBM DS8900F

Performance Best Practices and Monitoring

SG24-8501-00

ISBN 0738460192

(1.0” spine)
0.875”<->1.498”
460 <-> 788 pages

Redbooks

IBM DS8900F Performance Best Practices and Monitoring

SG24-8501-00

ISBN 0738460192

(0.5” spine)
0.475”<->0.873”
250 <-> 459 pages

Redbooks

IBM DS8900F Performance Best Practices and Monitoring

(0.2”spine)
0.17”<->0.473”
90<->249 pages

(0.1”spine)
0.1”<->0.169”
53<->89 pages

Redbooks

IBM DS8900F Performance Best

SG24-8501-00

ISBN 0738460192

IBM

(2.5” spine)
2.5”<->nnn.n”
1315<-> nnnn pages

Redbooks

IBM DS8900F Performance Best Practices and Monitoring

SG24-8501-00

ISBN 0738460192

IBM

(2.0” spine)
2.0” <-> 2.498”
1052 <-> 1314 pages

SG24-8501-00

ISBN 0738460192

Printed in U.S.A.