Redbooks
ibm.com/redbooks

IBM Power Systems RAID Solutions

Introduction and Technical Overview

Swarna Narendra Babu

Harihara Balakrishnan

Power Systems

IBM

Redpaper

International Technical Support Organization

IBM Power Systems RAID Solutions Introduction and Technical Overview

August 2015

REDP-5234-00

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

First Edition (August 2015)

This edition applies to IBM Power Systems servers based on POWER8 processor technology.

This document was created or updated on December 11, 2018.

Contents

Notices

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

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

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

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

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

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

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

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

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

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

COPYRIGHT LICENSE:

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

Trademarks

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

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

AIX®
DB2®
DS8000®
Easy Tier®
GPFS™
IBM Spectrum™
IBM®
Power Systems™
POWER8™
PowerHA®
PowerVM®
POWER®
Redbooks®
Redpaper™
Redbooks (logo) ®
Storwize®
XIV®

The following terms are trademarks of other companies:

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

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

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

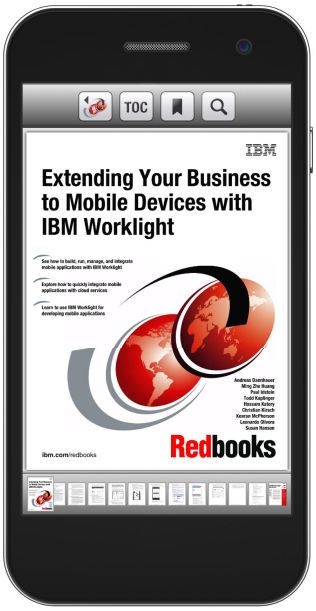
Other company, product, or service names may be trademarks or service marks of others.

Find and read thousands of IBM Redbooks publications

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

Get the latest version of the Redbooks Mobile App

iOS **Download Now** Android



Promote your business in an IBM Redbooks publication

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

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

ibm.com/Redbooks
About Redbooks → Business Partner Programs

THIS PAGE INTENTIONALLY LEFT BLANK

Preface

This IBM® Redpaper™ publication gives an overview and technical introduction to IBM Power Systems™ RAID solutions. The book is organized to start with an introduction to Redundant Array of Independent Disks (RAID), and various RAID levels with their benefits. A brief comparison of Direct Attached Storage (DAS) and networked storage systems such as SAN / NAS is provided with a focus on emerging applications that typically use the DAS model over networked storage models.

The book focuses on IBM Power Systems I/O architecture and various SAS RAID adapters that are supported in IBM POWER8™ processor-based systems. A detailed description of the SAS adapters, along with their feature comparison tables, is included in Chapter 3, “RAID adapters for IBM Power Systems” on page 47.

The book is aimed at readers who have the responsibility of configuring IBM Power Systems for individual solution requirements. This audience includes IT Architects, IBM Technical Sales Teams, IBM Business Partner Solution Architects and Technical Sales teams, and systems administrators who need to understand the SAS RAID hardware and RAID software solutions supported in POWER8 processor-based systems.

Authors

This paper was produced by a team of specialists from around the world working at the International Technical Support Organization, Poughkeepsie Center.

Swarna Narendra Babu is an IT Specialist working for IBM India Software Labs in INDIA. He has 10 years of experience as an UNIX and Windows administrator. He completed his Masters Degree in Information Technology in 2003 from Manipal Academy of Higher Education, Manipal, India. His areas of expertise include IBM AIX®, Linux, Solaris, VMWare, Windows, and HP-UX. He has had documents published in devWorks on IBM POWER® RAID and IBM GPFS™ Cluster:

http://www.ibm.com/developerworks/aix/tutorials/au-aix-raid/index.html
http://www.ibm.com/developerworks/aix/library/au-gpfs-cluster/index.html

Harihara Balakrishnan is a Managing Consultant with the Systems Lab Services team in IBM Singapore. He has more than 10 years of experience in systems administration, technical support, systems consulting, and technical training. He has worked at IBM for three years, and for client organizations earlier in his career. His areas of expertise include IBM Power Systems / AIX Performance, PowerVM® Virtualization, and High Availability using IBM PowerHA®. He is an IBM Certified Consulting IT Specialist, and IBM Certified Cloud Computing Infrastructure Architect. He holds a Bachelor’s degree in Electronics and Communication Engineering from the University of Madras.

The project that produced this publication was managed by:

Scott Vetter, PMP

A special thanks to the following who assisted with this publication:

Hans-Paul Drumm, IBM Germany

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

Ann Lund
International Technical Support Organization, Poughkeepsie Center

Clark Anderson
Sue Baker
Bob Galbraith
James Hermes
John Hock
Brian Horn
Mark Olson
Lakshmi Devi Subramanian
IBM

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

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

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

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

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

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

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

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

- Stay current on recent Redbooks publications with RSS Feeds:

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

1

Introduction to RAID Technologies

This chapter provides an overview and introduction to RAID technology and various levels of RAID. This chapter provides an introduction to different storage models used for various applications in enterprise data centers.

The RAID levels are introduced in detail, followed by a focused description of hardware and software RAID, and RAID levels supported in IBM Power Systems.

This chapter includes the following sections:

- General description
- Introduction to RAID technology

1.1 General description

Enterprises implement two broad types of Storage infrastructure for storing their application data.

- Direct-attached storage (DAS) consists of storage devices or disks that are attached to the server systems that run the application workloads. The devices and the stored data in DAS are private to typically one or two server systems or virtual machines hosted in the physical server hardware. Various high-availability configurations can be configured using server hardware. The data in DAS devices can be replicated to other storage systems, which usually requires replication features provided with certain application software.
- Networked enterprise storage infrastructure typically consists of storage area network (SAN) and network-attached storage (NAS). SAN and NAS provide the infrastructure to share storage devices and data over Fibre Channel or Ethernet networks with a much higher number of server systems and applications within the enterprise data centers. The data in SAN can be replicated across two or more data centers for high availability and disaster recovery requirements by using the required network infrastructure that spans across geographically dispersed data centers.

IT managers have the critical task of protecting application data against loss of data contributed by hardware failures, security invasion, power outages, and natural disasters to name a few. *Redundant Array of Independent / Inexpensive Disks (RAID)* is the technology commonly used in DAS, and also in networked storage *SAN-based* storage subsystems to provide reliable data protection against physical disk failures within an array in the subsystems. This publication provides a technical introduction and overview description of RAID technologies with a focus on RAID solutions that are supported with IBM Power Systems.

IBM Power Systems provide support for multiple options to configure integrated RAID SAS controllers, and many PCIe Gen3 SAS RAID adapters with options to configure them based on specific requirements. All the RAID adapters that are supported in IBM Power Systems are specially designed by IBM for Power Systems and provide industry leading features and performance. These RAID adapters allow users to configure various RAID levels for disk high availability, dual SAS RAID controllers for adapter high availability, and internal cache backed with flash memory. They can also use the support IBM Easy Tier® function for improved application I/O performance. The internal SAS controllers in all of the IBM Power Scale-Out systems also provide all the functions that PCIe Gen3 RAID adapters provide. The internal SAS controllers also allow the system units to connect to expansion SAS Disk drawers for configurations that need extra disk bays.

1.1.1 Emerging workloads, DAS, and RAID

Traditionally, IT architects in enterprises had the task of developing data storage solutions based on application and infrastructure needs. The data that was shared with multiple systems and needed to be replicated for disaster recovery was typically stored in SAN-based devices, and the front end web and application server data was stored in DAS devices. DAS devices provided ease of configuration, simpler device management tools, and typically better performance for certain workloads.

With the increased adoption to Big Data, Analytics, Mobile, Social and other emerging solution implementations, IT enterprises today are faced with the challenge of managing the enormous growth in the size and variety of data. The need to manage a wide variety of data for the new applications has become one of the critical challenges for managing IT infrastructure. The need to treat different data differently in terms of data source, storage

space, high availability, faster access, and data security has created a requirement to enhance storage and data management strategies to adapt to these emerging workloads.

An emerging trend in the design of the new applications is to store and access data from devices that are attached to the servers that run the workloads, typically, DAS, then access data from enterprise storage providers such as SAN and NAS. The reason behind this design is for the high performance characteristics that DAS can provide for these workloads, data isolation from networked storage infrastructure, and relatively lesser costs to use DAS devices than devices in networked storage devices. There is also a security requirement for the IT Infrastructure to provide isolation for the data used by the emerging workloads from the enterprise networked storage systems. This isolation enhances the data security in the enterprises so that front end / user-facing workloads do not get access to the enterprise networked storage subsystems. These new workloads are increasingly becoming another factor in the decision-making process for enterprises to adopt a DAS model.

For example, an enterprise wants to perform operational analytics by using data from specific server, network, and application logs in their IT infrastructure, and isolate these workloads from accessing other storage systems in the enterprise. The solution involves moving, storing, and processing the data that is required for the analytics application to isolated pools of server systems that have DAS devices to store and process data. This configuration provides the required isolation for analytics data and prevents these workloads from accessing the enterprise storage systems. However, there is also a need to provide high availability to the DAS devices that are used for the analytics workloads and other applications that use data in this model. Systems installed with RAID controllers to access the DAS devices provide that required level of High Availability and better application performance when accessing the data.

IBM Data Engine for Analytics (IDEA), one of the IBM Analytics solutions runs on IBM Power Systems, is supported by various PCIe Gen3 RAID adapters.

IBM Power Systems support various hardware- and software-based RAID technologies to provide high availability for both DAS and SAN-based devices. Chapter 4, “Software RAID in Power Systems” on page 81 describes software-based RAID configurations using the various supported operating systems on Power Systems. This book helps readers choose the most effective RAID solution for their application availability and performance requirements.

1.2 Introduction to RAID technology

Redundant Array of Independent/inexpensive Disks (RAID) involves two key design goals: Increased data reliability and increased input/output (I/O) performance. When multiple physical disks are set up to use the RAID technology, they are said to be in a RAID array. Although the array itself is distributed across multiple disks in the disk enclosure, but the array is seen by the computer user and operating system as a single disk. The operating system now accesses the single logical disk, and the RAID adapter handles the data distribution in the multiple disks in the array based on the RAID level with which the array is configured.

Each RAID adapter can support one or more RAID levels, depending on the adapter design. This section provides an introduction to the different RAID levels that exist, software RAID versus hardware RAID, and also detailed information about the RAID levels that are supported in IBM Power Systems.

1.2.1 Benefits and Economics - Various levels of RAID

The different RAID levels can be configured based on two important goals:

- High availability
- Performance

The benefits of RAID technology are providing better performance for data access, high availability for the data, or a combination of the two. RAID levels in general define a trade-off between high availability, performance, and cost. Understanding this equation for each of the RAID levels is required to determine the correct level of RAID to be implemented based on the requirements of the overall application needs. More advanced levels provide a balance of performance and high availability. Some of the basic RAID levels provide one of the two components in the equation. Although advanced levels can provide a certain degree of high availability without impacting performance, these advanced levels require more disks to implement compared to the other basic levels, thereby increasing costs.

The economics of the RAID solution are primarily defined by the ratio of the usable disk space to the total amount of disk space in the array. The RAID levels that provide the best performance and high availability would tend to be less economical because they need more disks to create and store parity data for each RAID set to protect against disk failures. Depending on the application's availability requirements for the level of failure protection, system configurations can be made with two different RAID adapters to protect against adapter failures. However, while this configuration provides more protection against adapter failures, it incurs extra cost for the second RAID controller in the solution.

There are many PCIe based RAID adapters that are supported by IBM Power Systems that provide high performance and availability features. The RAID adapters provide support for all commonly used RAID levels and they also provide unique and advanced RAID levels by using the Easy Tier function in Power Systems. There are supported RAID adapters with or without built-in cache for improved performance. The internal SAS controllers provide all the functions that the PCIe Gen3 RAID adapters provide including the function to connect to external disk drawers. The supported RAID adapters, and their features in each POWER8-based system models are covered in Chapter 3, “RAID adapters for IBM Power Systems” on page 47.

1.2.2 Hardware RAID and software RAID

RAID can be configured using hardware RAID adapters, or using operating system software that provides RAID functionality without the need of a hardware RAID adapter.

Hardware RAID is implemented by using a supported I/O adapter hardware that can be an integrated device in the system board/planar or it can be an external PCI-based RAID adapter. Hardware RAID uses an intelligent and robust disk controller and a redundancy array of disk drives to protect the data if a disk failure happens.

IBM Power Systems support many hardware-based Integrated RAID adapters and PCIe-based RAID adapters. Depending on the selected system model, there are choices for the hardware RAID adapters that can be configured in the system. PCIe-based RAID adapters are also supported in the PCIe Gen3 I/O expansion drawers that are connected to POWER8 processor-based systems. The hardware RAID adapters support different RAID levels that can be configured based on individual system and application needs. The Easy Tier function of the RAID adapters supported in IBM Power Systems provides an efficient way to store data based on the *hotness* of the data. The tier for storing the data is selected and managed by the RAID controller without any manual configuration required. This is an exclusive feature of selected backplane and integrated RAID adapters that are supported in IBM Power Systems.

The RAID adapters supported by IBM Power Systems are specifically designed by IBM for Power Systems, and have superior capabilities and performance characteristics when compared with industry standard RAID adapters available for server systems. The supported PCIe RAID adapters based on the system models are covered in Chapter 3, “RAID adapters for IBM Power Systems” on page 47.

Software RAID is configured by using the RAID functions provided by the operating systems. The RAID levels supported can vary for different operating systems. Most commonly configured software RAID levels are RAID 0 (striping) and RAID 1 (mirroring). For instance, AIX operating system on IBM Power Systems supports three levels of RAID using its Logical Volume Manager function. These three levels are striping, mirroring, and mirroring with striping.

Software RAID provides the functions without using a hardware RAID adapter and can be configured over disks in a just a bunch of disks (JBOD) configuration. However, in general, hardware RAID provides better features than the software RAID, while achieving the same or higher levels of protection against disk and adapter failures.

Table 1-1 provides a high-level comparison in benefits of software and hardware RAID configurations.

The supported RAID levels and configurations for IBM AIX, IBM i, and Linux on Power, Operation Systems are covered in detail in Chapter 4, “Software RAID in Power Systems” on page 81.

Table 1-1 Feature comparison for hardware and software RAID

RAID Type	Description	Advantages
Hardware RAID	► Implemented directly in hardware ► Requires a robust RAID hardware controller	► Data Protection with multiple options ► Higher performance than software-based RAID ► Robust
Software RAID	► Software-based configuration is managed by the operating system / application	► Low cost without need for hardware controller ► Can be configured with a standard SAS controller

1.2.3 Hardware RAID levels and their characteristics

The different types in the RAID Levels can be divided into these categories based on their characteristics:

- Basic RAID Levels
  - RAID 0, RAID 1, and RAID 5 form the basic RAID levels
- Hybrid RAID levels
  - Hybrid Levels are formed using two basic levels in combination.
  - RAID 6, RAID 10, RAID 50, and RAID 60 form the Hybrid RAID Levels.
- Most commonly used RAID levels are 0, 1, 5, 6, and 10

The following sections provide an introduction to each of the hardware RAID levels and a summary table that discusses the benefits and characteristics.

RAID 0

RAID0, also called *striping*, splits data across all the disks without any parity information, redundancy, or fault tolerance. It can be configured with multiple disks with different disk sizes. A minimum of two disks is required for configuring RAID 0. RAID 0 is commonly preferred when performance is most important and data integrity is of less importance. It provides the highest read and write I/O performance. Because no data redundancy is provided by this RAID level, it is not commonly used for mission critical environments with stringent data availability requirements.

Figure 1-1 shows the logical diagram for RAID 0.

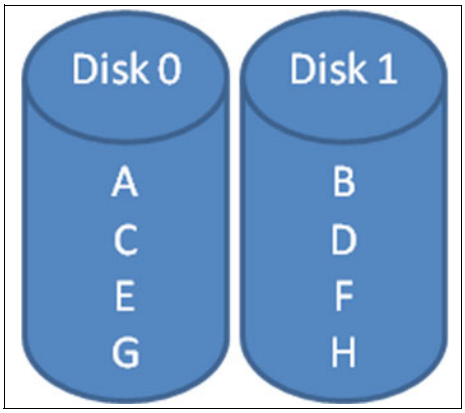


Figure 1-1 RAID 0 logical diagram

RAID 1

RAID 1 is also called *mirroring*. This level writes the same copy of data across all the disks, which are the part of the RAID 1 array. RAID 1 provides the highest data availability because blocks are mirrored across multiple disks and can protect from one or more concurrent disk failures depending on the total number of disks/copies in the array. It provides good performance for write I/O activities and the best performance for read I/O activities. RAID 1 is generally configured when the integrity of the data is of the most importance, and workload does more read IOs than write IOs. Mirroring requires a minimum of two disks in the array to be mirrored.

Figure 1-2 shows the logical diagram for RAID 1.

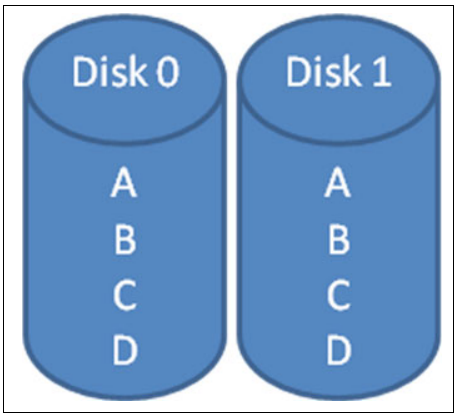


Figure 1-2 RAID 1 logical diagram

RAID 2

RAID 2 stripes the data at the bit level rather than block level. The disks are synchronized by the controller to spin at the same angular orientation. Because all disks are configured with their own error code correction and other configuration complexities, it is rarely implemented. RAID 2 requires at least four raw disks in the array.

Figure 1-3 shows the logical diagram for RAID 2.

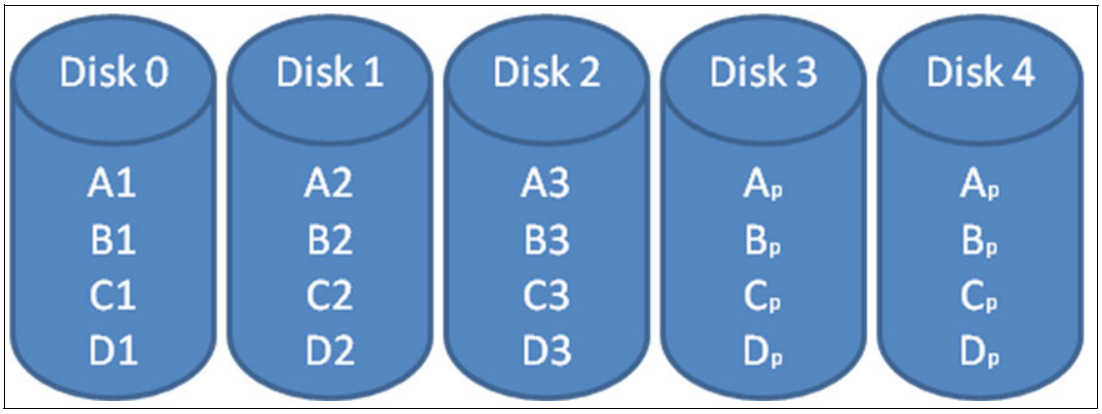


Figure 1-3 RAID 2 logical diagram

RAID 3

RAID 3 consists of byte-level striping with a dedicated parity disk. One of the characteristics of RAID 3 is that it generally cannot service multiple requests simultaneously, which happens because any single block of data will, by definition, be spread across all members of the set and be in the same location. Therefore, any I/O operation adds processor usage to all disks in the arrays, and usually requires synchronized spindles. Both RAID 3 and RAID 4 have been replaced by RAID 5. You need at least four raw disks to configure RAID 3.

Figure 1-4 shows the logical diagram for RAID 3.

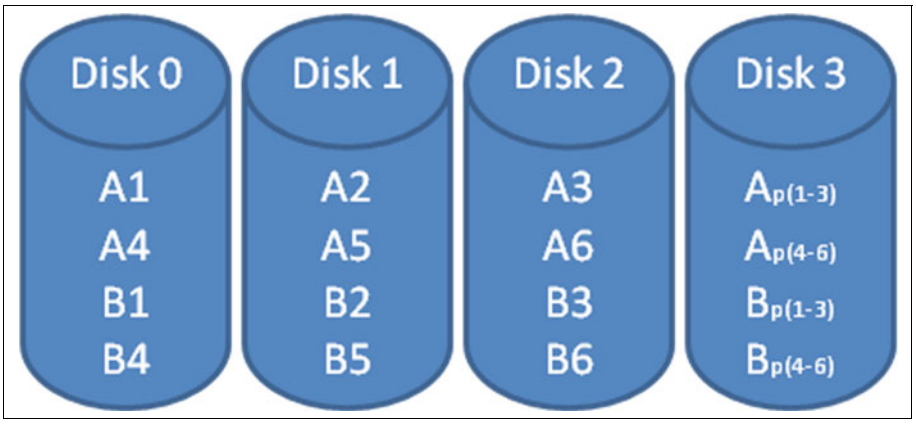


Figure 1-4 RAID 3 logical diagram

RAID 4

RAID 4 is configured using a dedicated parity disk. It provides good performance compared to RAID 2 and RAID 3. RAID 4 requires at least three raw disks to configure the RAID array. It is rarely used.

Figure 1-5 shows the logical diagram for RAID 4.

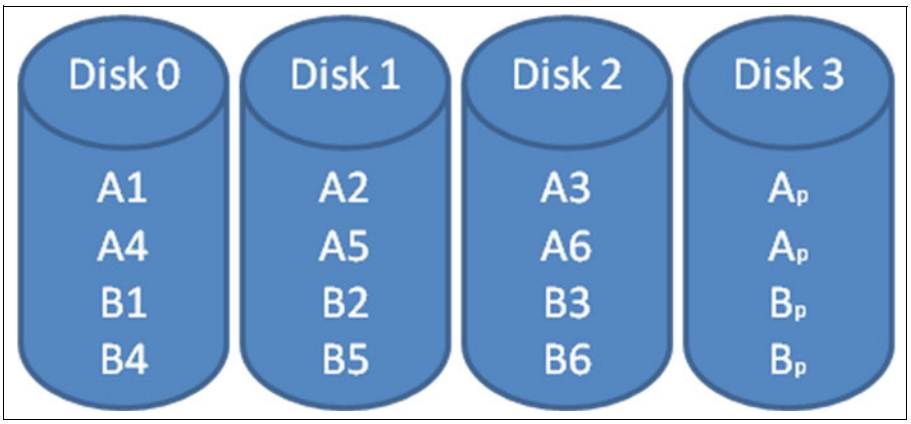


Figure 1-5 RAID 4 logical diagram

RAID 5

RAID 5 consists of block level striping with distributed parity. The parity information is distributed across all the disks in the RAID array. If one disk in the array fails, there is no data loss because all data can be restored to a replacement disk. RAID 5 provides good I/O performance for read and write activities. RAID 5 requires a minimum of three disks to configure.

Figure 1-6 shows the logical diagram for RAID 5.

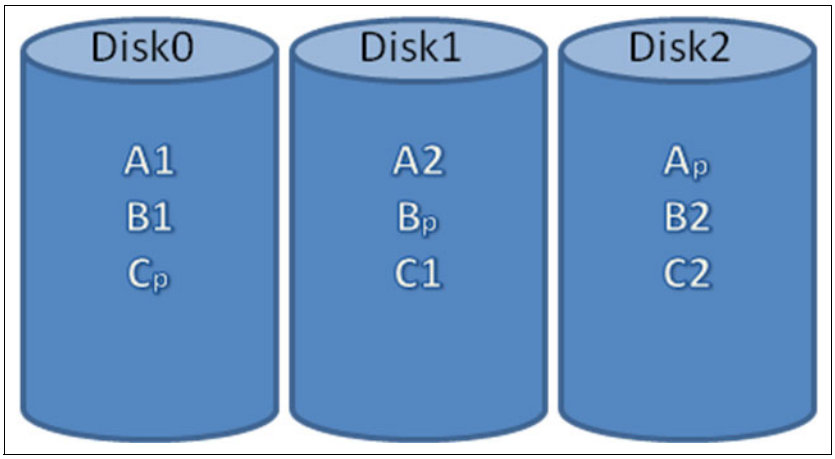


Figure 1-6 RAID 5 logical diagram

RAID 6

RAID 6 consists of block level striping with the same distributed parity disk as RAID 5, and adds an additional or extended parity disk in the array configuration. This level can provide better disk failure protection rate than RAID 5. In comparison to RAID 5 performance, RAID 6 offers better performance for reads than writes because of the extra processing required by the additional parity disk. RAID 6 requires at least four raw disks to configure RAID 6.

Figure 1-7 shows the logical diagram for RAID 6.

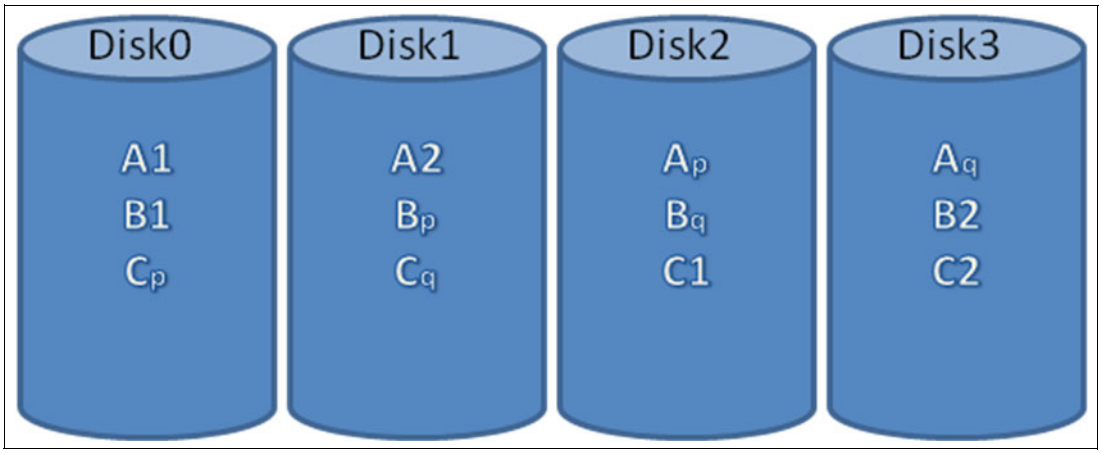


Figure 1-7 RAID 6 logical diagram

RAID 10

RAID 10 is one of the hybrid RAID levels, also known as RAID 1+0. It is a combination of disk mirroring with disk striping. RAID 10 requires a minimum of four disks to configure. It provides the best I/O performance for reads and writes in the array. However, it is not economical because it provides only 50% usable disk space in the total array capacity. This level requires at least four raw disks in the array to be configured.

Figure 1-8 shows the logical diagram for RAID 10.

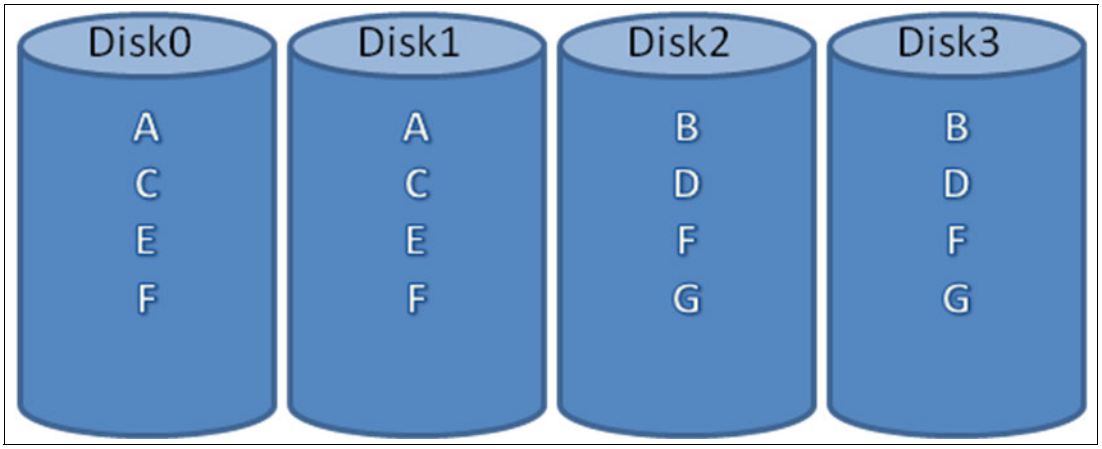


Figure 1-8 RAID 10 logical diagram

RAID 50

RAID 50 is the nonstandard configuration that is also known as RAID 5+0. It is a combination of disk striping with single parity disk. RAID 50 requires a minimum of six disks to configure. It provides the highest data protection even with multiple disk failures.

RAID 60

RAID 60 is the nonstandard configuration that is also known as RAID 6+0. It is a combination of disk striping with multiple parity disks. RAID 60 requires a minimum of eight disks to configure.

Summary of RAID levels

Table 1-2 provides a summary of RAID levels, with an overview of their benefits and economics.

Table 1-2 Summary of RAID levels

RAID Level	Minimum Drives	Protection	Description	Strengths	Weakness
RAID 0	2	None	Data striping without redundancy	Highest performance	No data protection; If one drive fails, all data is lost
RAID 1	2	Single drive failure	Disk mirroring	High performance; Offers data protection; high read performance and good write performance	High redundancy requires extra processing; Because all data is duplicated, twice the storage capacity is required
RAID 2	4	Single drive failure (group)	Bit level striping	High data protection	Expensive
RAID 3	4	Single drive failure	Byte level striping with dedicated parity	Sequential read and write. Provides good I/O performance while reading and writing the data	Not commonly in use due to complexity in implementation
RAID 4	3	Single drive failure	Block level striping with dedicated parity	Good for random read	Not commonly in use due to complexity in implementation

RAID Level	Minimum Drives	Protection	Description	Strengths	Weakness
RAID 5	3	Single drive Failure	Block-level data striping with distributed parity	Best in price/ performance for transaction-orien ted networks; Very high performance; High data protection; Supports multiple simultaneous reads and writes; Can also be optimized for large, sequential requests	Write performance is slower than RAID0 and RAID1
RAID 6	4	Two-drive failure	Same as RAID 5 with double distributed parity across an extra drive	Offers solid performance with the additional fault tolerance of allowing availability to data if two disks in a RAID group fail; Use more drives in RAID group to make up for performance and disk utilization hits compared to RAID5	Must use a minimum of four drives with two of them used for parity, so disk utilization is not as high as RAID3 or RAID5. Performance is slightly lower than RAID5
RAID 10	4	One disk per mirrored stripe (not same mirror)	Combination of RAID0 (disk striping) and RAID1 (mirroring)	Highest performance, highest data protection (can tolerate multiple drive failures)	High redundancy cost extra processing; Because all data is duplicated, twice the storage capacity is required; Requires a minimum of four drives

RAID Level	Minimum Drives	Protection	Description	Strengths	Weakness
RAID 50	6	One disk per mirrored stripe replicated over at least 2 RAID arrays	Combination of RAID0 (disk striping) and RAID5 (single parity drive)	Highest performance, highest data protection (can tolerate multiple drive failures)	High redundancy cost extra processing; Because all data is duplicated, twice the storage capacity is required; Requires a minimum of four drives
RAID 60	8	Two disks per mirrored stripe	Combination of RAID0 (disk striping) and RAID6 (dual-parity drives)	Highest performance, highest data protection (can tolerate multiple drive failures)	High redundancy cost extra processing; Because all data is duplicated, twice the storage capacity is required; Requires minimum of four drives

1.2.4 Supported hardware RAID levels in Power Systems

POWER8 based systems support various levels of hardware RAID configurations. The RAID levels that are supported depend on the SAS RAID adapters used. One of the factors while selecting the RAID adapter in the configuration depends on the RAID level that is required for the configuration. There are other factors in the adapter selection process, such as support for Easy Tier, Internal Cache in the RAID adapters, High Availability configurations, and operating system support. For more information, see Chapter 3, “RAID adapters for IBM Power Systems” on page 47.

RAID level options:

RAID 0 Striping

RAID 1 Mirroring

RAID 5 Striping with distributed parity

RAID 6 Striping with Multiple distributed parity

RAID 10 Striping with Mirroring

Easy Tier options:

RAID 5T2 RAID5 with Easy Tier

RAID 6T2 RAID6 with Easy Tier

RAID 10T2 RAID10 with Easy Tier

1.2.5 Supported software RAID levels in Power Systems

Apart from the RAID Levels supported using hardware RAID controllers, Power Systems also support software RAID configured in the operating systems running in the logical partitions.

Power Systems allow Logical Partitions to be installed with one of the three supported operating systems:

- AIX
- Linux on Power
- IBM i

Each of the operating systems support software RAID levels. AIX and Linux allow configuring RAID levels using their Logical Volume Manager. For more information on software RAID levels, see Chapter 4, “Software RAID in Power Systems” on page 81.

2

IBM Power Systems I/O architecture overview

This chapter provides an overview of the I/O subsystem architecture of the POWER8 processor-based servers. It include a general description for each system model and its I/O system overview with internal and external storage options for the system configurations.

This chapter does not provide detailed information about the architecture and features of POWER8 processor and memory configurations in IBM Power Systems. However, a link to the other relevant IBM Redbooks publications is provided wherever applicable.

This chapter includes the following sections:

- General Description
- PCIe Gen3 Slots
- Coherent Accelerator Processor Interface
- Easy Tier in Power Systems
- IBM Power Systems I/O architecture overview
- Power S814 and S824 I/O Architecture Overview
- Power S822 I/O Architecture Overview
- Power S812L and Power S822L I/O Architecture Overview
- Power E850 I/O Architecture Overview
- Power E870 and Power E880 I/O Architecture Overview
- PCIe Gen3 I/O Expansion Drawer Overview

2.1 General Description

IBM POWER8 systems provide the infrastructure for high performance I/O for application workloads. Power Systems are designed for data with superior memory and I/O bandwidths to cater not only to traditional application and database workloads, but also to emerging workloads such as analytics. All POWER8 systems use the latest industry standard PCIe Gen3 I/O slots in the system units, and also support PCIe Gen3 I/O expansion drawers for system configurations that need more I/O slots.

The superior I/O bandwidth and performance in Power Systems also enables users to consolidate more virtual servers on fewer physical server platforms. The direct benefits of higher consolidation help enterprises to reduce the total cost of ownership (TCO) for the system investments and also reduce the running costs incurred for data center floor space, power, and cooling expenses.

2.2 PCIe Gen3 Slots

POWER8 Systems support PCIe Gen3 slots that provide better bandwidth than the previous generations of PCIe slots. PCIe Gen3 cards/slots provide up to twice the bandwidth of PCIe Gen2 slots/cards, and up to four times the bandwidth of PCIe Gen1 slots/cards.

2.2.1 PCIe Lanes and Bandwidth

A single PCIe serial link is a dual-simplex connection that uses two pairs of wires, one pair for transmit and one pair for receive, and can transmit only one bit per cycle. These pairs of wires are called a lane. PCIe cards/slots are labeled with number of lanes that are supported by the card/slot. For example, x4, x8, and x16 have 4, 8, and 16 lanes as shown in Figure 2-1.

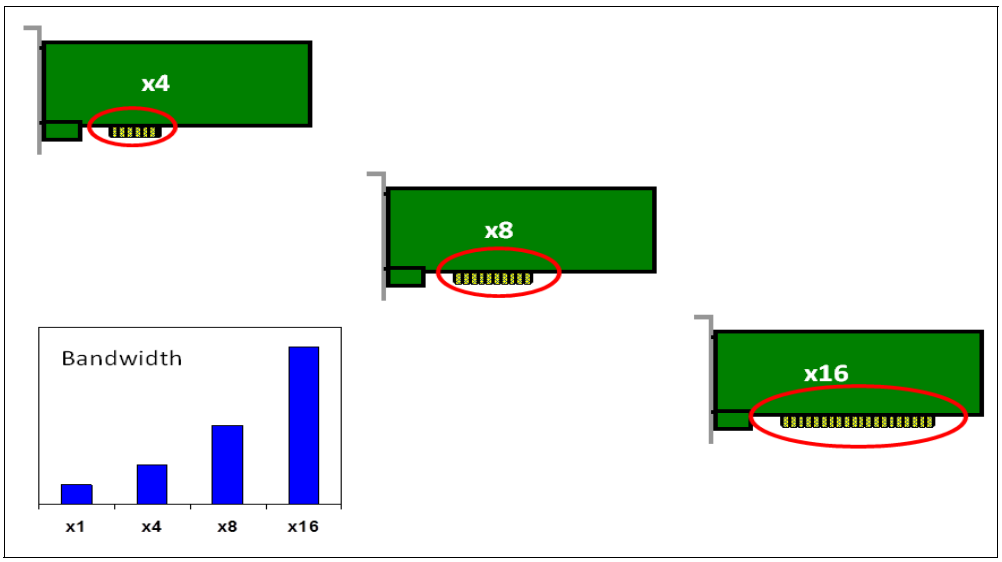


Figure 2-1 PCIe Lanes and Bandwidth

The more lanes the slot supports, the higher the bandwidth provided to the adapter, which is also labeled by lanes. A PCIe x16 adapter installed on a PCIe x16 slot gets the full bandwidth supported by the slot, whereas a PCIe x8 adapter installed on the same slot only uses half of the bandwidth supported by the slot.

Tip: A higher lane PCIe adapter cannot be installed in a lower lane PCIe slot. For example, a PCIe x16 adapter cannot be installed in a PCIe x8 slot. However, the reverse is supported, that is, PCIe x8 card can be installed in a PCIe x16 slot.

The PCIe Lanes and bandwidth are shown in Figure 2-1 on page 16.

With the PCI3 Gen3 (x8 and x16) slots, each slot in a POWER8 system has higher bandwidth than previous generations. This increased bandwidth in each slot and adapter allows for improved I/O performance for the application workloads and provides for higher consolidation of more virtual servers on limited sets of server hardware. They provide best server infrastructure for high-bandwidth adapters such as 10 GbE, 40 GbE adapters, and 16 Gbps FC adapters. A conceptual view of the possible consolidation with PCIe Gen3 slots is shown in Figure 2-2.

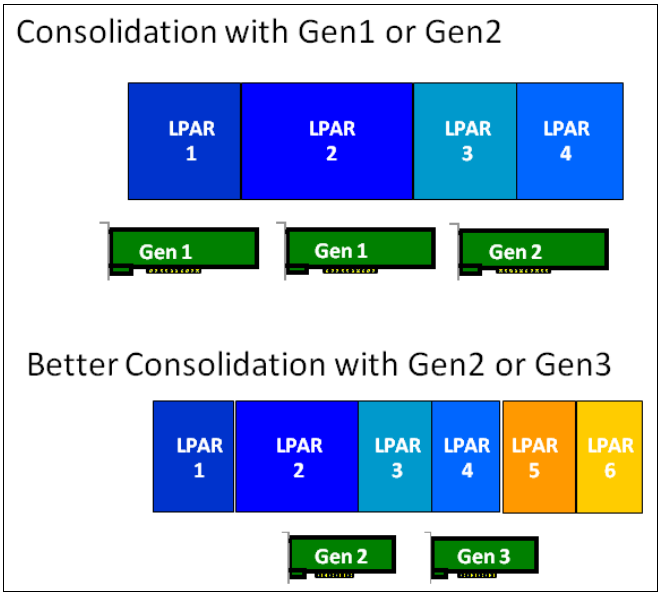


Figure 2-2 Consolidation with PCIe Gen1 and Gen2

The PCIe Gen3 slots with x16 lanes are capable of providing 16 GBps simplex (32 GBps duplex).

PCIe low profile (LP) cards are used with the 2U system unit PCIe slots. These cards are not compatible with P4U system units because of its low height, but have similar cards in another form factors.

PCIe full height and full high cards are not compatible with the 2U system units PCIe Gen3 slots and are designed for the 2U system units only.

All adapters support Enhanced Error Handling (EEH). PCIe adapters use a different type of slot than PCI adapters. If you attempt to force an adapter into the wrong type of slot, you might damage the adapter or the slot.

2.3 Coherent Accelerator Processor Interface

Coherent Accelerator Processor Interface (CAPI) is an innovative I/O interface in the POWER8 based systems that provides a new, efficient interface for applications to use field-programmable gate arrays (FPGAs), graphics processing units (GPUs), and other traditional storage devices efficiently.

An overview of the architecture for CAPI is shown in Figure 2-3.

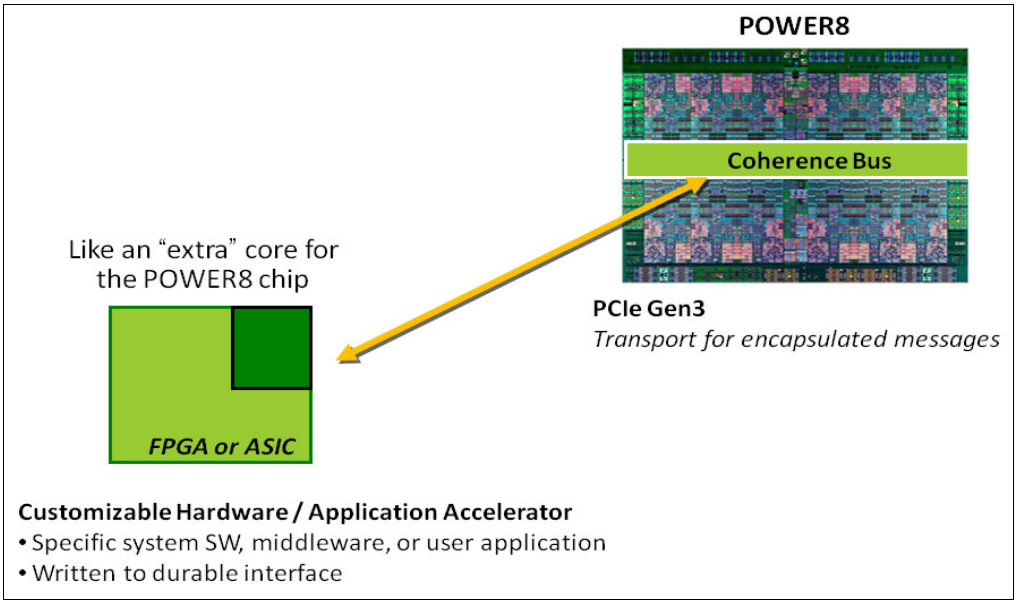


Figure 2-3 CAPI Architecture Overview

CAPI allows accelerators such as FPGAs to have coherent shared memory access with the system processor. This helps to lower the memory access latencies for the accelerators than the traditional implementation in which they depend on the system processor response times for memory access requests. Traditional accelerators can be only as fast as the system processor's response time for the memory requests. CAPI has been designed to address this performance need for the accelerators. With CAPI to provide coherent memory access to the system processors, the accelerators achieve their full and potential I/O performance levels.

Applications that use CAPI-based FPGAs, or Storage devices also negate the extra processing of using device drivers loaded in the Operating System kernel to bypass and perform I/O operations directly to the devices in the CAPI interfaces.

IBM Power Systems using CAPI as the interface to the flash memory can provide an innovative solution to the requirements for higher performance, low I/O latency for applications, transactional processing databases, and emerging workloads such as analytics.

An architecture diagram and notes on key benefits of flash memory connected by using CAPI is shown in Figure 2-4.

Innovative, and industry leading solutions like CAPI, and industry standard technologies like PCIe Gen3 supported in IBM Power Systems enable them to be the ideal platform for application workloads that require high I/O performance.

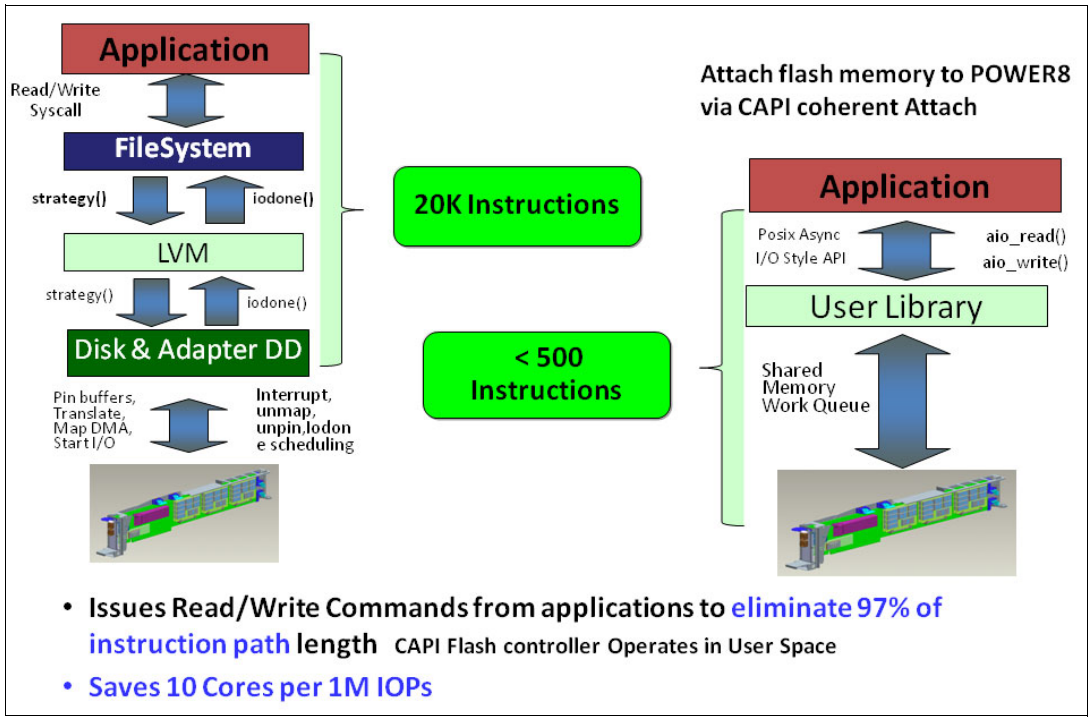


Figure 2-4 Using CAPI for flash memory

2.4 Easy Tier in Power Systems

The internal SAS controllers in the Power Scale Out systems provide the Easy Tier function capability. The Easy Tier is an integrated SAS RAID controller provided feature that helps improve I/O performance for the application workloads running in the system.

The integrated SAS controller can monitor for data access patterns in the RAID array, which consists of hard disk drives (HDDs) and solid-state drives (SSDs). It allows the controller to move hot data to high performance SSD and cold data to HDD. The data movement, also known as tiering, happens dynamically and updates every few seconds or minutes based on the identified access pattern and workload demand.

A conceptual view of the Easy Tier function is shown in Figure 2-5. For more information about this feature, see Chapter 3, “RAID adapters for IBM Power Systems” on page 47.

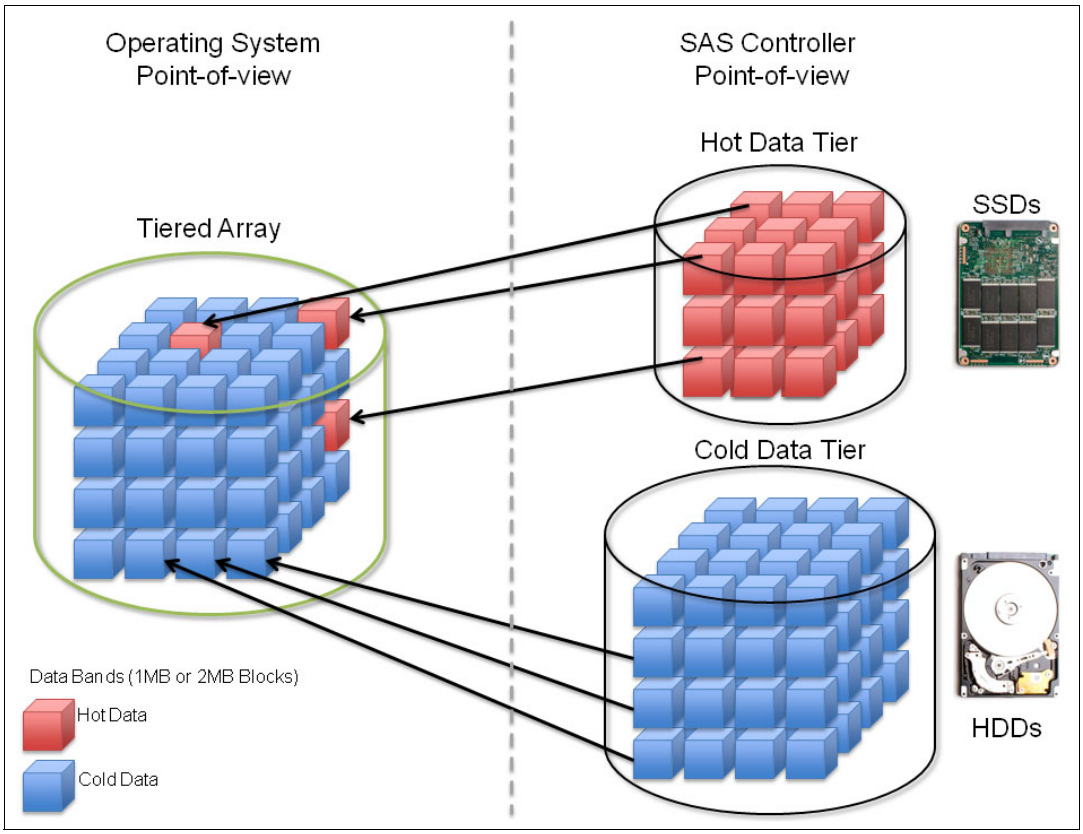


Figure 2-5 Easy Tier functions

2.5 IBM Power Systems I/O architecture overview

This section of this chapter provides information about the I/O architecture, PCIe Gen3 slots availability, and I/O Expansion units for all POWER8 processor-based systems. This section introduces concepts about the PCIe RAID SAS adapters for each of the system models that are covered in detail in Chapter 3, “RAID adapters for IBM Power Systems” on page 47.

More information about each of the system models for topics beyond I/O architecture, such as CPU, Memory, and Virtualization, is available at:

http://www.redbooks.ibm.com/redbooks.nsf/portals/

2.6 Power S814 and S824 I/O Architecture Overview

Power S814 is a single socket server with a maximum of eight POWER8 processor active cores. The system can be configured with 4-core, or 6-core POWER8 processors, in either a rack or tower configuration. The eight activated cores configuration is available only in the rack model. The system offers a maximum memory capacity of 1 TB with a maximum of eight DDR3 CDIMM slots. The system provides a maximum I/O bandwidth of 96 Gbps.

IBM Power S824 is a one or two socket server with a maximum CPU capacity of 24 POWER8 processor cores. The system provides 6-core, 8-core, 12-core, 16-core, and 24-core configurations. The Power S824 system has a maximum memory capacity of 2 TB with a maximum of 16 DDR3 CDIMM slots.

The two systems provide outstanding I/O performance capabilities that include PCIe Gen3 (x8 and x16) I/O internal slots, Easy Tier for Internal Storage disks, CAPI, and an array of supported PCIe Gen3 adapters. These features provide the best performance for data intensive workloads such as Databases, Middleware, and Analytics applications.

The two systems allow configuring PCIe Gen3 I/O Expansion drawer to support extra PCIe Gen3 slots in the system units. Both systems offer options to extend the internal storage using SFF Disk Expansion I/O drawers for configurations that need more internal storage, typically used for analytics workloads. This section provides an overview of I/O configurations of the two system models in detail.

2.6.1 Power S814 PCIe Gen3 Slots

This section describes the characteristics, and availability of the PCIe Gen3 slots in the Power S814 system unit:

- Total Number of PCI slots - Seven hot-pluggable PCIe Gen3 slots.
- There are five x8 PCIe3 slots and two x16 PCIe Gen3 slots available in the system unit.
- The number of slots available for PCIe Gen3 adapters depend on the Storage Backplane option chosen in the configuration.
- The I/O workload is balanced across available processor sockets and I/O switches (PEX) that connect the slots to the processor sockets.
  - The two x16 PCIe Gen3 slots are connected to two different chips in the single POWER8 Dual Chip Module.
  - The five x8 PCIe Gen3 slots are connected to two different PCIe Gen3 switches (PEX) that are connected to two different chips in the single POWER8 Dual Chip Module.
- Use the IBM System Planning Tool to ensure that the adapter placements are valid for the system configuration that you are building.

The IBM System Planning Tool is available at:

http://www-947.ibm.com/systems/support/tools/systemplanningtool/

For a full list of all supported PCIe adapters for S814, see the IBM Knowledge Center or *IBM Power Systems S814 and S824 Technical Introduction and Overview* at:

http://www.redbooks.ibm.com/abstracts/redp5097.html?Open
http://www-01.ibm.com/support/knowledgecenter/POWER8/p8hdx/POWER8welcome.htm

Figure 2-6 shows the logical diagram of Power S814.

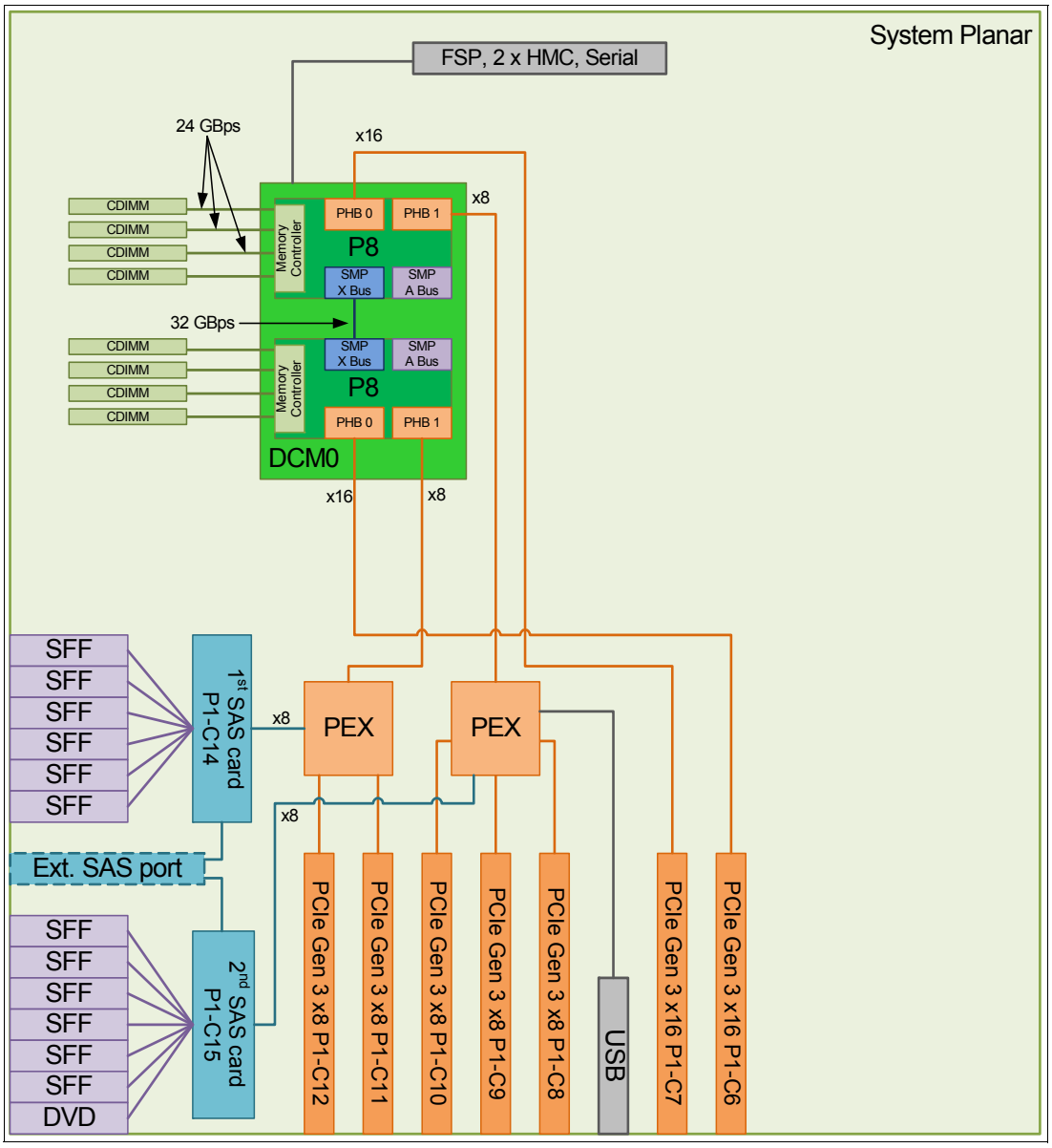


Figure 2-6 Power8 S814 logical system diagram

2.6.2 Power S824 PCIe Gen3 Slots

This section describes the characteristics, and availability of the PCIe Gen3 slots in the Power S824 system unit:

- Total Number of PCI slots - Eleven hot-pluggable PCIe Gen3 when two processor sockets are filled, and seven hot-pluggable PCIe Gen3 slots when one processor socket is filled.
- There are seven x8 PCIe3 slots and four x16 PCIe Gen3 slots available in the system unit.
- The number of slots available for PCIe Gen3 adapters depends on the Storage Backplane option chosen in the configuration, and the number of installed processor sockets.
- The I/O workload is balanced across available processor sockets and I/O switches (PEX) that connect the slots to the processor sockets.
  - The four x16 PCIe Gen3 slots are connected to two different chips with two x16 slots connected to each of the two processor sockets.
  - The five out of the seven x8 PCIe Gen3 slots are connected to two different PCIe Gen3 switches (PEX) that are connected to two different chips in the single POWER8 Dual Chip Module socket.
- The remaining two of the seven x8 PCIe Gen3 slots are connected to the two different chips of one processor socket.
- Use the IBM System Planning Tool to ensure that the adapter placements are valid for the system configuration that you are building.

The IBM System Planning Tool is available at:

http://www-947.ibm.com/systems/support/tools/systemplanningtool/

For a full list of all supported PCIe adapters for S824, see the IBM Knowledge Center or *IBM Power Systems S814 and S824 Technical Introduction and Overview* at:

http://www.redbooks.ibm.com/abstracts/redp5097.html?Open
http://www-01.ibm.com/support/knowledgecenter/POWER8/p8hdx/POWER8welcome.htm

Figure 2-7 shows the logical system diagram for Power S824.

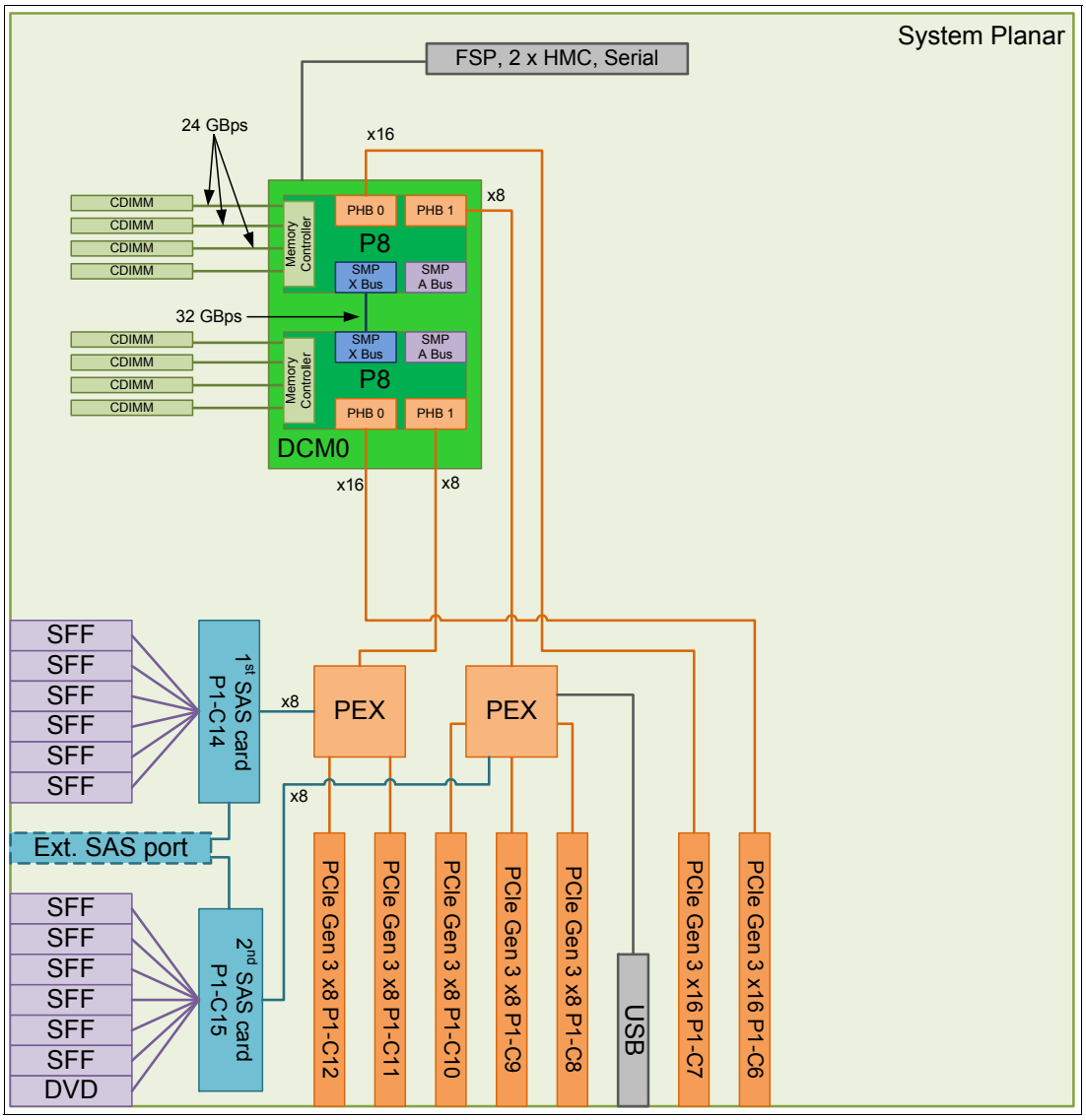


Figure 2-7 Power S824 Logical System Diagram

Note for Power S814 and Power S824

The slot P1-C10 comes populated with a PCIe2 4-port 1Gb Ethernet adapter (#5899). This adapter cannot be used or moved to a different slot if the #EJ0P backplane option is used. Slot P1-C11 gets obstructed when #EJ0P is used for the backplane configuration.

2.6.3 Integrated SAS Controllers

There are three options for the Integrated SAS Controllers for both Power S814 and Power S824 systems:

- #EJ0N - Storage Backplane 12 SFF-3 Bays/DVD Bay
- #EJ0S - Split #EJ0N to 6+6 SFF-3 bays
- #EJ0P - Storage Backplane 18 SFF-3 bays / SSD / Dual IOA write cache

#EJ0N - Storage Backplane 12 SFF-3 Bays/DVD Bay

#EJ0N has these characteristics:

- One SAS Disk controller supports JBOD, RAID 0, RAID 5, RAID 6, and RAID 10.
- The SAS disk controller is placed on slot #P1-C14.
- This single SAS Disk controller is sufficient for all twelve 2.5 inch Gen3 SFF disk HDDs or SSDs.

#EJ0S - Split #EJ0N to 6+6 SFF-3 bays

#EJ0S has these characteristics:

- This option is used to provide the split backplane feature.
- Two SAS Controllers split the disk bays to provide two Disks groups (6+6), each including six 2.5 inch PCIe Gen3 SFF Disks.
- Each SAS Controller can be assigned to different partitions (LPARs).
- The two SAS controllers are placed in P1-C14 and P1-C15 slots.

#EJ0P - Storage Backplane 18 SFF-3 bays / SSD / Dual IOA write cache

#EJ0P has these characteristics:

- This option provides the high performance SSD disks in the SSD Module cage that is included in this configuration.
- System configurations that need Easy Tier feature in the internal storage disks must choose this backplane option.
- Two SAS controllers provide JBOD, RAID 0, RAID 5, RAID 6, RAID10, RAID 5T2, RAID 6T2, and RAID 10T2.
- The two SAS controllers are placed in the dedicated P1-C14 and P1-C15 slots, and are in Active-Active mode configuration.
- This feature by default adds the SSD module cage (#EL0H) with eight 1.8-inch SSD module bays.
- This option also provides the Easy Tier function that provides the best performance for direct-attached storage (DAS) with a mix of SSDs and HDDs:
  - The Easy Tier function in Power Systems can place and move hot data to high performance SSDs and move cold data to attached HDDs.
  - This feature is also required to add the EXP24S SFF Gen2-bay Drawer. (#5887)

2.6.4 PCIe Gen3 I/O Expansion Drawers

This section provides supported PCIe3 I/O Expansion drawer configurations for the Power S814 and Power S824 systems.

Power S814 Guidelines

Power S814 supports one PCIe Gen3 Expansion I/O drawer (EMX0).

The system requires 6-core or 8-core configurations because the 4-core configuration does not allow an Expansion I/O drawer to be added.

One Expansion I/O drawer provides twelve PCIe Gen3 slots, and requires two PCIe Gen3 slots in the system unit to connect to the Expansion I/O drawer.

Power S824 Guideline

Power S824 system supports one PCIe Gen3 Expansion I/O drawer when only one of the two processor sockets are installed.

The system supports two PCIe Gen3 Expansion I/O drawers when two processor sockets are installed.

One Expansion I/O drawer provides twelve PCIe Gen3 slots, and requires two PCIe Gen3 slots in the system unit to connect to the one or two Expansion I/O drawers. For more information about the PCIe Gen3 I/O Expansion drawer, see 2.11, “PCIe Gen3 I/O Expansion Drawer Overview” on page 42.

2.6.5 EXP24S SFF Gen-2 Drawer and internal disk bays

This section provides information about internal disk bays, and supported EXP24S SFF Gen-2 drawers in the Power S814 and S824 system units:

- The internal disk slot bays configuration is the same for both Power S814 and Power S824 systems.
- The system unit in the default configuration consists of two separate SFF-3 disk bays with six disk slots each.
- An optional SFF bay with eight slots SFF-3 for SSDs can be configured in the systems. This configuration requires the #EJ0P Feature code to be enabled.
- The system unit also allows you to add EXP24S SFF Gen2-bay Drawer for configurations that need more disk bays. This configuration requires the #EJ0P Feature code to be enabled. For more information about the #EJ0P feature, see “#EJ0P - Storage Backplane 18 SFF-3 bays / SSD / Dual IOA write cache” on page 25.
- If not using #EJ0P feature code, the EXP24S SFF Gen2-bay Drawer can be connected to the system unit by using PCIe SAS adapters.

2.7 Power S822 I/O Architecture Overview

Power S822 offers a flexible configuration with options to install one or two socket POWER8 processors. It is available in 6-core, 8-core, 12-core, 16-core, and 20-core configurations. The system allows a maximum memory capacity of 1 TB with a maximum of 16 IBM CDIMM slots. The system in the 2U rack dimensions, with all POWER8 enhancements, and performance characteristics, offers an ideal choice for private and public cloud infrastructures. IBM POWER8 processor performance characteristics along with IBM PowerVM virtualization helps with virtualization and consolidation of larger workloads.

All IBM Power Scale-Out system models, including Power S822, also have a guaranteed 65% of sustained system utilization without affecting application performance.

2.7.1 Power S822 PCIe Gen3 Slots

This section describes the characteristics, and availability of the PCIe Gen3 slots in the Power S822 system unit:

- Total Number of PCIe Gen3 slots depends on the number of processor sockets filled:
  - Nine PCIe Gen3 slot when two processor sockets are filled.
  - Six PCIe Gen3 slots when one processor socket is filled.

Single Processor socket configuration

A single-socket configuration has these characteristics:

- In the Single Processor socket configuration, there are four x8 PCIe3 slots and two x16 PCIe Gen3 slots available in the system unit.
- The I/O workload is balanced across available processor sockets and I/O switches (PEX) that connect the slots to the processor sockets:
  - The two x16 PCIe Gen3 slots are connected to two different chips in the single POWER8 Dual Chip Module.
  - The five x8 PCIe Gen3 slots are connected to two different PCIe Gen3 switches (PEX) that are connected to two different chips in the single POWER8 Dual Chip Module.

Figure 2-8 shows the logical system diagram for a single-socket Power S822.

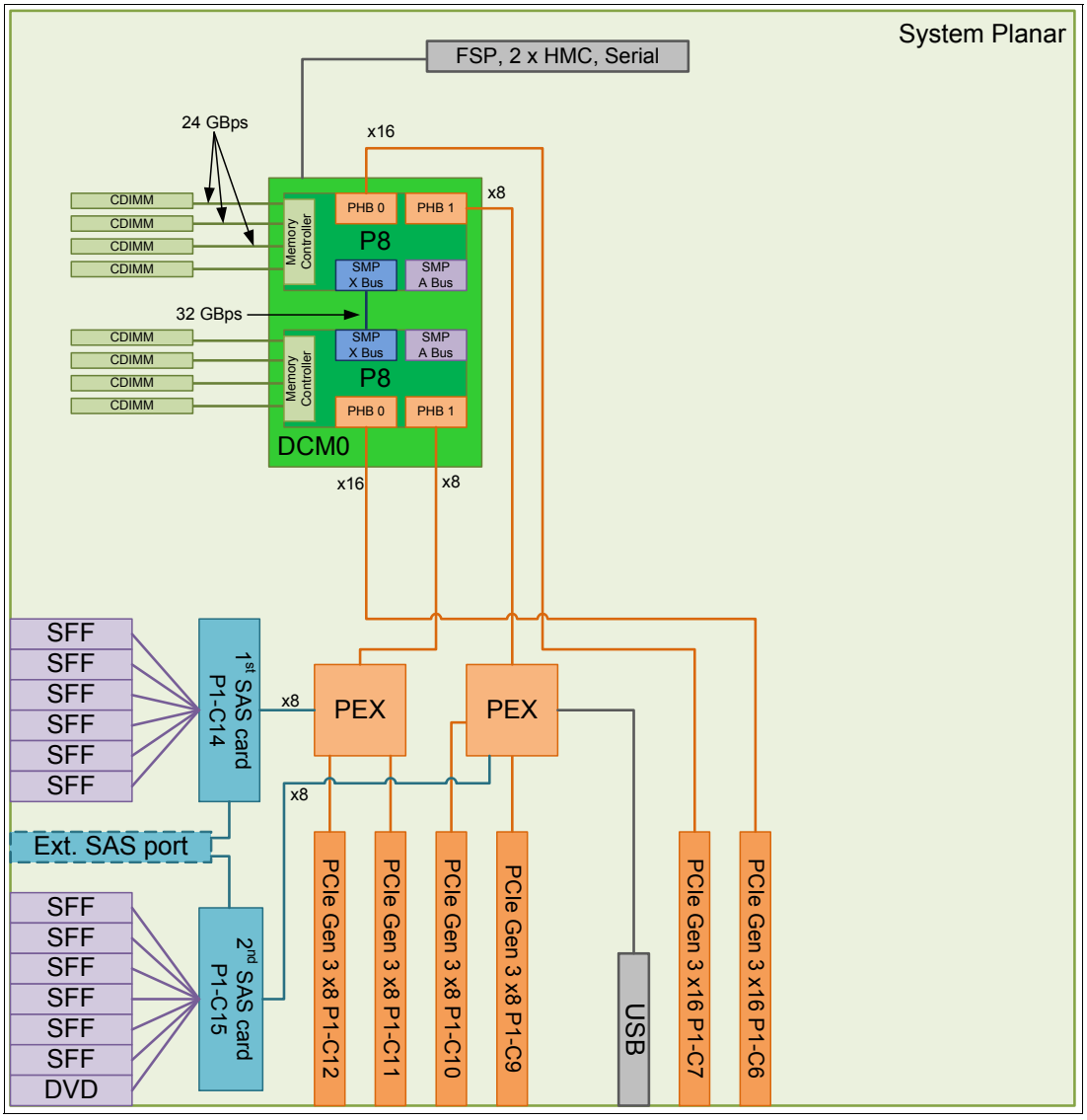


Figure 2-8 Logical system diagram for a one socket Power S822

Two Processor socket configuration

A two-socket configuration has these characteristics:

- In the two processor socket configuration, there are four x8 PCIe3 slots and four x16 PCIe Gen3 slots available in the system unit.
- The I/O workload is balanced across available processor sockets and I/O switches (PEX) that connect the slots to the processor sockets:
  - The four x16 PCIe Gen3 slots - A pair of x16 slots are connected to two different chips in each POWER8 Dual Chip Module.
  - The four x8 PCIe Gen3 slots are connected to two different PCIe Gen3 switches (PEX) that are connected to two different chips in the single POWER8 Dual Chip Module.
  - The fifth x8 PCIe Gen3 slot (P1-C2) is connected to one of the two sockets without a PEX switch.

Figure 2-9 shows the logical system diagram for a two-socket Power S822.

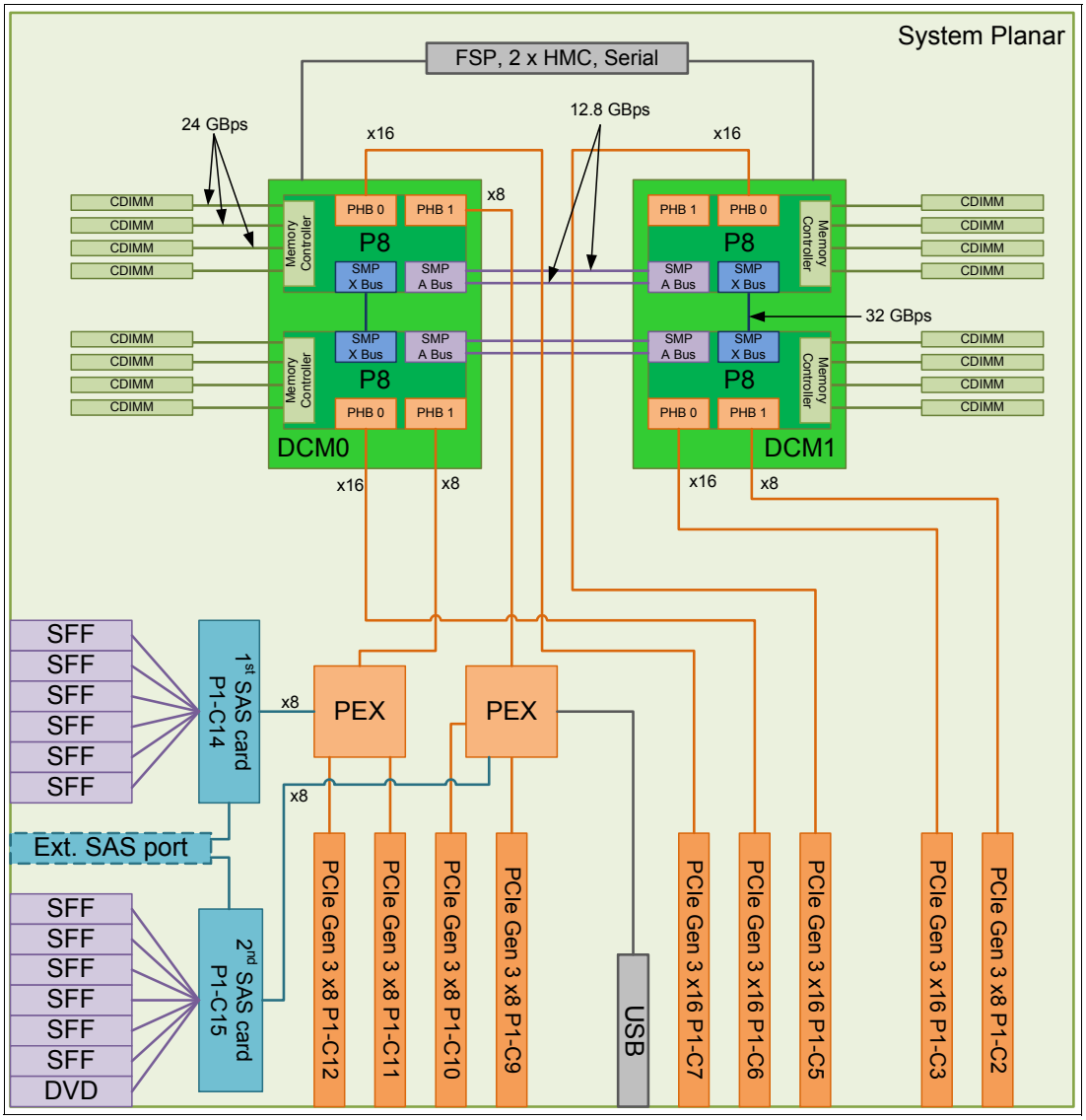


Figure 2-9 Logical system diagram for a two socket Power S822

Other considerations

The slot P1-C10 comes populated with a PCIe2 4-port 1 Gb Ethernet adapter (#5260). Slot P1-C9 gets obstructed when #EJ0U is used for the backplane configuration. Any adapter that is planned for use in P1-C9 must be removed for #EJ0U or a different slot must be chosen.

Use the System Planning Tool to ensure that the adapter placements are valid for the system configuration that you are building.

The IBM System Planning Tool is available at:

http://www-947.ibm.com/systems/support/tools/systemplanningtool/

For a full list of all supported PCIe adapters for S822, see the IBM Knowledge Center or *IBM Power System S822 Technical Overview and Introduction* at:

http://www.redbooks.ibm.com/abstracts/redp5102.html?Open
http://www-01.ibm.com/support/knowledgecenter/POWER8/p8hdx/POWER8welcome.htm

2.7.2 Integrated SAS Controllers

This section provides the configuration details for the Integrated SAS Controllers in Power S822 system:

- #EJ0T - Storage backplane 12 SFF-3 bays / DVD bay
- #EJ0V - Split #EJ0T to 6+6 SFF-3 bays
- #EJ0U - Storage backplane 8 SFF-3 bays/ SSD / Dual IOA write cache

#EJ0T - Storage backplane 12 SFF-3 bays / DVD bay

#EJ0T has these characteristics:

- One SAS Disk controller supports JBOD, RAID 0, RAID 5, RAID 6, and RAID 10.
- The SAS disk controller is placed in slot #P1-C14.
- A Single SAS Disk controller controls all twelve 2.5 inch Gen3 SFF disk HDDs or SSDs.

#EJ0V - Split #EJ0T to 6+6 SFF-3 bays

#EJ0V has these characteristics:

- This option is used to provide the split backplane feature.
- Two SAS Controllers split the disk bays to provide two Disks groups (6+6), each of which has six 2.5 inch PCIe Gen3 SFF Disks.
- Each SAS Controller can be assigned to different LPARs.
- The two SAS controllers are placed in P1-C14 and P1-C15 slots.

#EJ0U - Storage backplane 8 SFF-3 bays/ SSD / Dual IOA write cache

#EJ0U has these characteristics:

- This option provides the storage backplane with two SAS Controllers with eight 2.5 inch PCIe Gen3 SFF Disks.
- This option provides the high performance SSD disks in the #EJTL SSD Module cage that is included in this configuration.
- The #EJTL SSD Module cage has six 1.8 inch SSD disk bays.
- Two SAS controllers provide JBOD, RAID 0, RAID 5, RAID 6, RAID10, RAID 5T2, RAID 6T2, and RAID 10T2.

- The two SAS controllers are placed in the dedicated P1-C14 and P1-C15 slots, and are in the Active-Active mode configuration.
- This feature by default adds the SSD module cage (#EL0H) with six 1.8-inch SSD module bays.
- This option also provides the Easy Tier function that provides the best performance for DAS with mix of SSDs and HDDs.
- Easy Tier function in the Power Systems can place and move hot data to high performance SSDs and move cold data to the HDDs attached. This feature also requires that you add EXP24S SFF Gen2-bay Drawer (#5887).

2.7.3 PCIe Gen3 I/O Expansion Drawers

This section provides supported PCIe3 I/O Expansion drawer configurations for the Power S822 system:

- Power S822 supports one PCIe Gen3 I/O Expansion drawer (EMX0).
- The system configuration filled with one processor socket allows an Expansion drawer with only one fan-out module (half-drawer).
  - One fan-out module provides six PCIe Gen3 slots.
  - Requires two PCIe Gen3 slots in the system unit to connect the Expansion drawer.
- The system configuration filled with one processor socket allows an Expansion drawer with two fan-out modules (full-drawer):
  - Two fan-out modules provide twelve PCIe Gen3 slots.
  - Requires four PCIe Gen3 slots in the system unit to connect the Expansion drawer.

For more details about PCIe Gen3 I/O Expansion drawer, see 2.11, “PCIe Gen3 I/O Expansion Drawer Overview” on page 42

2.7.4 EXP24S SSF Gen-2 Drawer and Internal Disk bays

This section provides information about Internal disk bays, and supported EXP24S SSF Gen-2 Drawers.

- The system unit in the default configuration consists of two separate SFF disk bays with each six disk slots.
- #EJ0P Feature code is required for both of these configurations:
  - An optional SFF bay with six slots for SSDs can be configured in the system.
  - The system unit also allows you to add one EXP24S SFF Gen2-bay Drawer for configurations that need more disk bays.

2.8 Power S812L and Power S822L I/O Architecture Overview

Power S812L is a high performance single socket server with a maximum of twelve POWER8 processor active cores. The system provides two options with 10 or 12 activated cores in the POWER8 processor Dual Chip module. The system offers a maximum memory capacity of 512 GB with a maximum of eight DDR3 IBM CDIMM slots. The system provides a maximum I/O bandwidth of 96 Gbps.

IBM Power S824 is a two socket server with a maximum CPU capacity of 24 POWER8 processor cores. The system offers 16-core, 20-core, and 24-core activations for the processor capacity. The system has a maximum memory capacity of 2 TB with a maximum of 16 CDIMM slots. This high-density server is designed to provide high performance for workloads that are compute and data intensive.

IBM Power S812L and S822L are Linux-only models that can be used for running Linux workloads in a 2U form factor. The high-performance characteristics and virtualization features make these systems of high value for enterprises that need an industry-leading system for their Linux workloads.

The systems are designed with a focus on traditional Linux workloads, emerging workloads such as analytics, and mobile applications designed with Linux as the supported operating system.

PowerKVM as an optional hypervisor to PowerVM hypervisor can be configured to host virtual machines (LPARs) running on one or more supported Linux operating systems on Power platform. PowerKVM enables easy adoption to Power platform to run Linux workloads in environments where KVM is used as a primary hypervisor.

2.8.1 Power S812L PCIe Gen3 Slots

This section describes the characteristics, and availability of the PCIe Gen3 slots in the Power S812L system unit:

- Total Number of PCI slots - Six hot-pluggable PCIe Gen3.
- There are four x8 PCIe3 slots and two x16 PCIe Gen3 slots available in the system unit.
- The I/O workload is balanced across available processor sockets and I/O switches (PEX) that connect the slots to the processor sockets:
  - The two x16 PCIe Gen3 slots are connected to two different chips in the single POWER8 Dual Chip Module.
  - The five x8 PCIe Gen3 slots are connected to two different PCIe Gen3 switches (PEX) that are connected to two different chips in the single POWER8 Dual Chip Module.

Use the System Planning tool to ensure that the adapter placements are valid for the system configuration that you are building.

The IBM System Planning Tool is available at:

http://www-947.ibm.com/systems/support/tools/systemplanningtool/

For a full list of all supported PCIe adapters for S812L, see the IBM Knowledge Center or *IBM Power System S812L and S822L Technical Overview and Introduction* at:

http://www.redbooks.ibm.com/abstracts/redp5098.html?Open
http://www-01.ibm.com/support/knowledgecenter/POWER8/p8hdx/POWER8welcome.htm

Figure 2-10 shows the logical system diagram for a Power S812L.

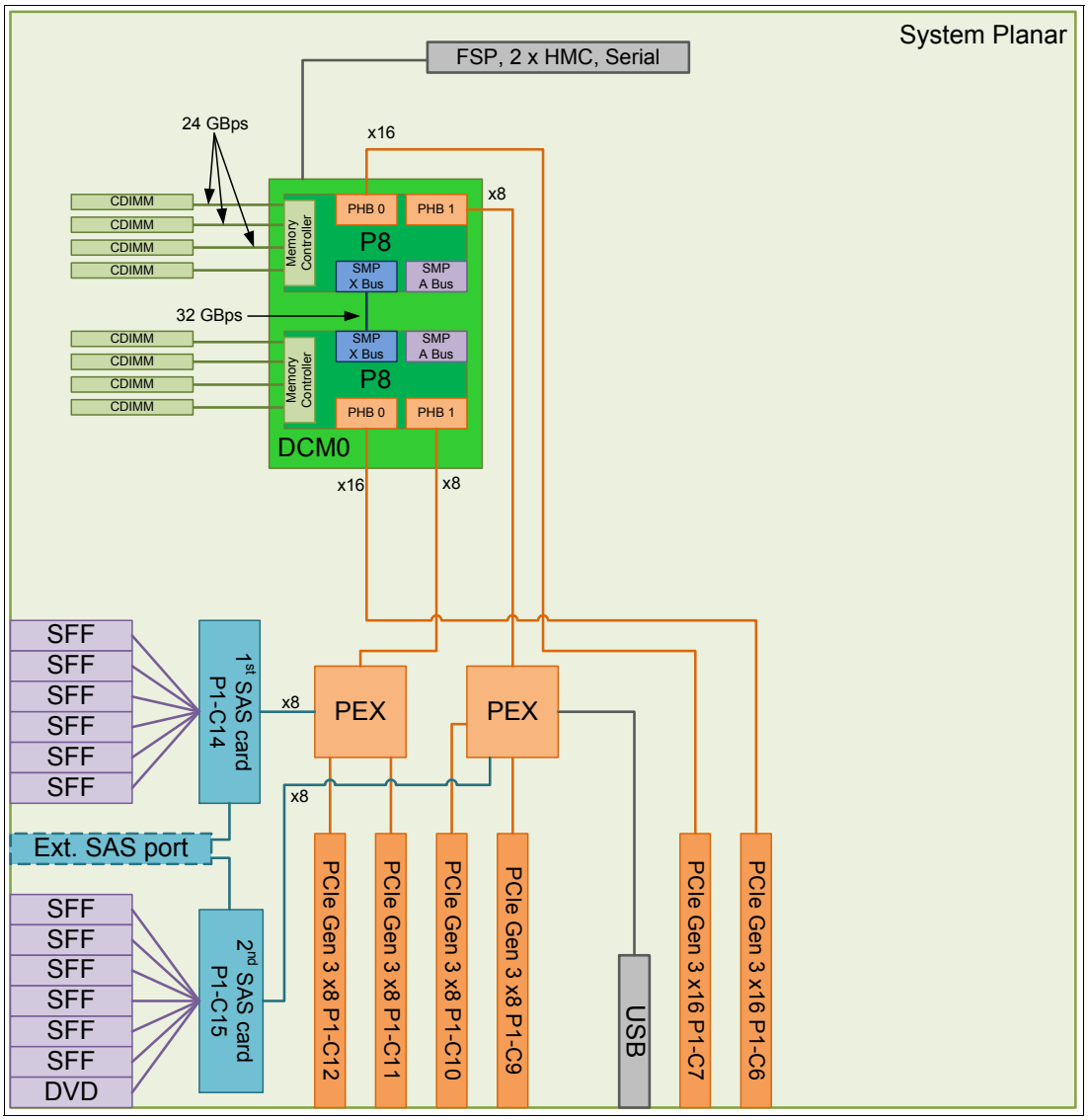


Figure 2-10 Power S812L logical system diagram

2.8.2 Power S822L PCIe Gen3 Slots

This section describes the characteristics, and availability of the PCIe Gen3 slots in the Power S822L system unit:

- Total Number of PCI slots - Nine hot-pluggable PCIe Gen3.
- There are five x8 PCIe3 slots and four x16 PCIe Gen3 slots available in the system unit.
- The I/O workload is balanced across available processor sockets and I/O switches (PEX) that connect the slots to the processor sockets:
  - The four x16 PCIe Gen3 slots - A pair of x16 slots are connected to two different chips in each POWER8 Dual Chip Module.
  - The four x8 PCIe Gen3 slots are connected to two different PCIe Gen3 switches (PEX) that are connected to two different chips in a single POWER8 Dual Chip Module.

– The fifth x8 PCIe Gen3 slot (P1-C2) is connected to one of the two sockets without a PEX switch.

- Use the System Planning tool to ensure that the adapter placements are valid for the system configuration that you are building.

The IBM System Planning Tool is available at:

http://www-947.ibm.com/systems/support/tools/systemplanningtool/

For a full list of all supported PCIe adapters for S822L, see the IBM Knowledge Center or *IBM Power System S812L and S822L Technical Overview and Introduction* at:

http://www.redbooks.ibm.com/abstracts/redp5098.html?Open
http://www-01.ibm.com/support/knowledgecenter/POWER8/p8hdx/POWER8welcome.htm

Figure 2-11 shows the logical system diagram for a Power S822L.

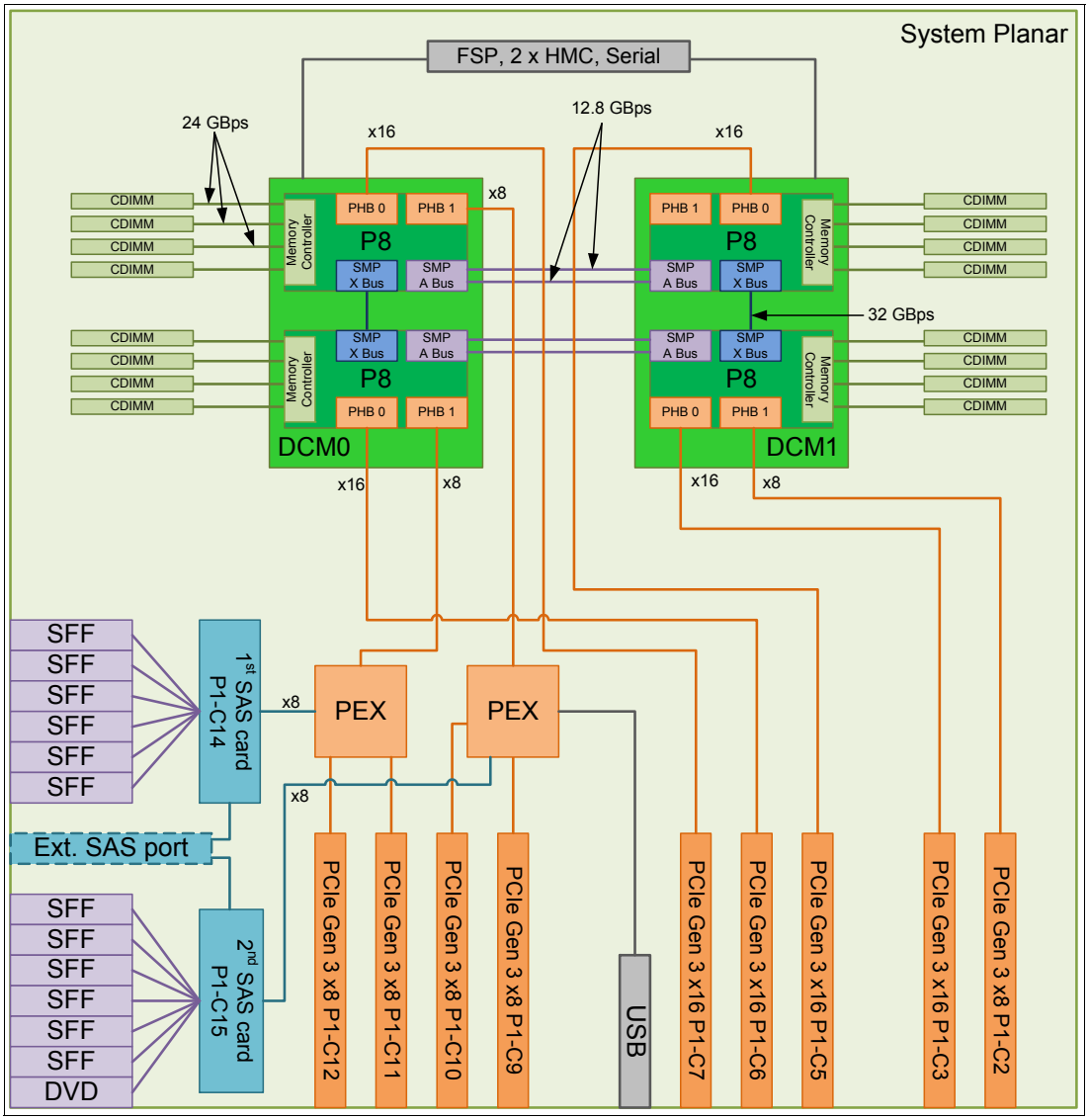


Figure 2-11 Power S822L logical system diagram

Note for Power S812L and Power S822L

The slot P1-C10 comes populated with a PCIe2 4-port 1 Gb Ethernet adapter (#5260). This adapter cannot be used or moved to a different slot if #EL3U backplane option is used. Slot P1-C9 gets obstructed when #EL3U is used for the backplane configuration.

2.8.3 Integrated SAS Controllers

There are three options for the Integrated SAS Controllers for both Power S812L and Power S822L systems:

- #EL3T - Storage backplane 12 SFF-3 bays / DVD bay
- #EL3V - Split #EL3T to 6+6 SFF-3 bays
- #EL3U - Storage backplane 8 SFF-3 bays / SSD / Dual IOA write cache

#EL3T - Storage backplane 12 SFF-3 bays / DVD bay

#EJ3T has these characteristics:

- One SAS Disk controller supports JBOD, RAID 0, RAID 5, RAID 6, and RAID 10.
- The SAS disk controller is placed on slot #P1-C14.
- A single SAS Disk controller controls all twelve 2.5 inch SFF Gen3 HDDs or SSDs.

#EL3V - Split #EL3T to 6+6 SFF-3 bays

#EJ3V has these characteristics:

- This option is used to provide the split backplane feature.
- Two SAS Controllers split the disk bays to provide two disk groups (6+6), each with six 2.5 inch PCIe Gen3 SFF Disks.
- Each SAS Controller can be assigned to one or two different LPARs.
- The two SAS controllers are placed in P1-C14 and P1-C15 slots.

#EL3U - Storage backplane 8 SFF-3 bays / SSD / Dual IOA write cache

#EJ3U has these characteristics:

- This option provides the high performance SSD disks in the SSD Module cage that is included in this configuration.
- Two SAS controllers provide JBOD, RAID 0, RAID 5, RAID 6, RAID10, RAID 5T2, RAID 6T2, and RAID 10T2.
- The two SAS controllers are placed in the dedicated P1-C14 and P1-C15 slots, and are in the Active-Active mode configuration.
  - This feature by default adds the SSD module cage (#EL0H) with eight 1.8-inch SSD module bays only in Power S822L.
  - #EL0H it is not available for Power S812L system.
- This option also provides Easy Tier function that provides the best performance for DAS with mix of SSDs and HDDs.
- Easy Tier function in the Power Systems can place and move hot data to high performance SSDs and move cold data to the attached HDDs.
- This feature also allows the user to add one EXP24S SFF Gen2-bay Drawer.

2.8.4 PCIe Gen3 I/O Expansion Drawers

This section provides supported PCIe3 I/O Expansion drawer configurations for the Power S812L and Power S822L systems.

Power S812L guidelines

Power S812L has these characteristics:

- Power S812L supports one PCIe Gen3 Expansion I/O drawer (EMX0) with one fan-out module (half-drawer):
  - One fan out module provides six PCIe Gen3 slots.
  - Requires two PCIe Gen3 slots in the system unit to connect the Expansion drawer.
- If the configuration needs one full drawer, use a Power822L instead.

Power S822L guidelines

Power S822L has these characteristics:

- Power S822L system supports one PCIe Gen3 Expansion I/O drawer (EMX0).
- One Expansion I/O drawer provides 12 PCIe Gen3 slots, and requires 2 PCIe Gen3 slots in the system unit to connect to one or two Expansion I/O drawers.

For more details about PCIe Gen3 I/O Expansion drawer, see 2.11, “PCIe Gen3 I/O Expansion Drawer Overview” on page 42.

2.8.5 EXP24S SSF Gen-2 Drawer and Internal Disk bays

This section provides information about Internal Disk bays, and supported EXP24S SSF Gen-2 Drawers:

- The Disk slot bays configuration is the same for both Power S812L and Power S822L systems.
- The system unit in the default configuration consists of two separate SFF disk bays, each with six disk slots.
- An optional SFF bay with eight slots SFF-3 for SSDs can be configured only in Power822L:
  - This configuration requires #EL3U Feature code to be enabled.
  - This is not a supported feature in the Power 812L system. Configurations that need this feature should only consider Power 822L system model.
- The system unit also allows you to add EXP24S SFF Gen2-bay Drawer for configurations that need more disk bays. This configuration requires #EL3U Feature code to be enabled. For more information, see “#EL3U - Storage backplane 8 SFF-3 bays / SSD / Dual IOA write cache” on page 34.
- If not using #EL3U feature code, the EXP24S SFF Gen2-bay Drawer can be connected to the system unit by using PCIe SAS adapters.
- Each Power S812L and S822L can support up to 14 EXP24S drawers.

2.9 Power E850 I/O Architecture Overview

Power E850 is a single enclosure, four socket system with enterprise class performance and RAS features. The system offers a maximum CPU capacity of 48 POWER8 processor cores, and a minimum activation of 16 cores. The system allows a maximum memory capacity of 2 TB with a maximum of 32 IBM CDIMM DDR3 slots. The number of available PCIe Gen3 slots depend on the number of processor sockets that are filled in the system.

The 4U system with a maximum of 48 POWER8 cores offers high-density package and capacity to support consolidation for more virtual machines on fewer physical systems, fewer rack space requirements, and a better price/performance ratio. The system also provides enterprise class reliability, availability, and serviceability features to support high availability demands of mission critical workloads.

2.9.1 Power E850 PCIe Gen3 Slots

This section describes the characteristics, and availability of the PCIe Gen3 slots in the Power E850 system unit:

- The total number of PCIe Gen3 slots depends on the number of the processor sockets filled. More number of processor sockets in the system provides more number of PCIe Gen3 slots in the system unit:
  - 7 PCIe Gen3 slot when two processor sockets are filled.
  - 9 PCIe Gen3 slots when three processor sockets are filled.
  - 11 PCIe Gen3 slots when four processor sockets are filled.

Two-processor socket configuration

A two-processor module configuration of Power E850 has these characteristics:

- In the two Processor socket configuration, there are four x16 PCIe3 slots and three x8 PCIe Gen3 slots available in the system unit.
- The I/O workload is balanced across available processor sockets and I/O switches (PEX) that connect the slots to the processor sockets:
  - The first pair of two x16 PCIe Gen3 slots are connected to two different chips in one POWER8 Dual Chip Module (proc mod0).
  - One x8 PCIe Gen3 slot is connected to a different chip in the same POWER8 Dual Chip Module (proc mod0).
  - The second pair of two x16 PCIe Gen3 slots are connected to two different chips in the second POWER8 Dual Chip Module (proc mod1).
  - One pair of two x8 PCIe Gen3 slots is connected to a different switch (PEX) in the second POWER8 Dual Chip Module (proc mod1).

Figure 2-12 shows a two-processor socket configuration of Power E850.

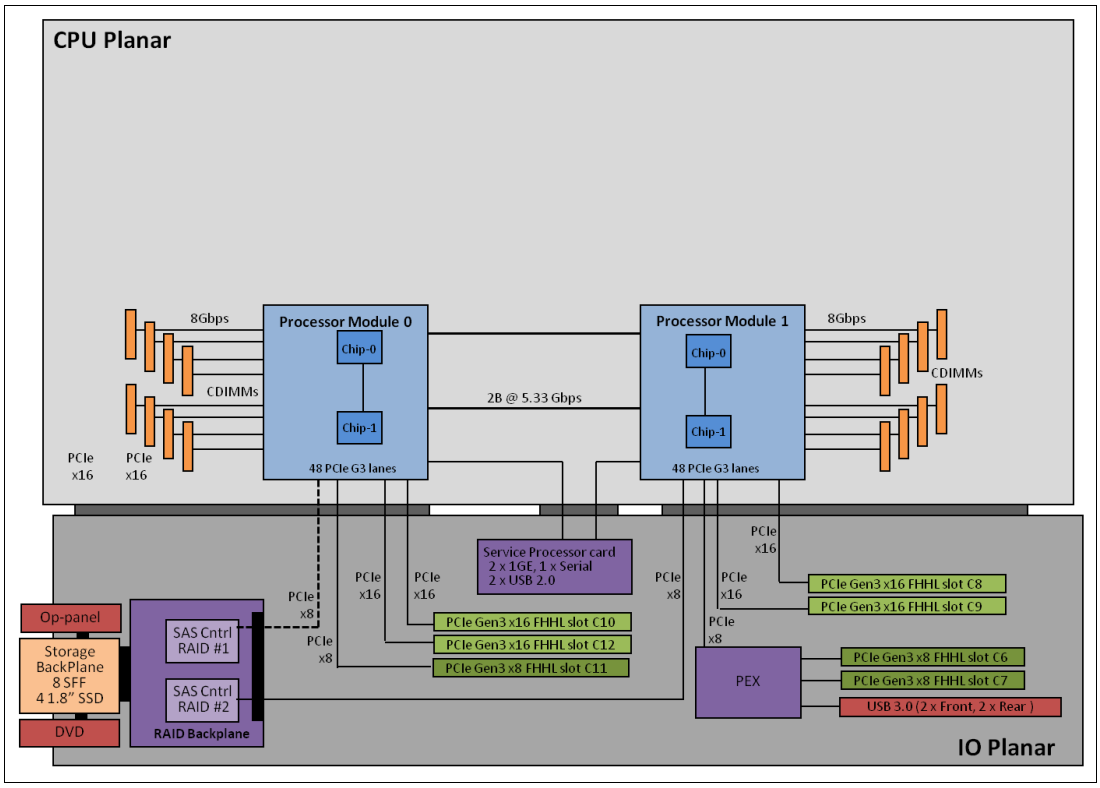


Figure 2-12 Power E850 with two processor modules installed

Three-processor socket configuration

A three-processor module configuration of Power E850 has these characteristics:

- In the three Processor socket configuration, there are six x16 PCIe3 slots and three x8 PCIe Gen3 slots available in the system unit.
- The I/O workload is balanced across available processor sockets and I/O switches (PEX) that connect the slots to the processor sockets:
  - The first pair of two x16 PCIe Gen3 slots is connected to two different chips in one POWER8 Dual Chip Module (proc mod0).
  - One x8 PCIe Gen3 slot is connected to a different chip in the same POWER8 Dual Chip Module (proc mod0).
  - The second pair of two x16 PCIe Gen3 slots is connected to two different chips in the second POWER8 Dual Chip Module (proc mod1).
  - One pair of two x8 PCIe Gen3 slot is connected to a different switch (PEX) in the second POWER8 Dual Chip Module (proc mod1).
  - The third pair of x16 PCIe Gen3 slots is connected to two different chips in the third POWER8 Dual Chip Module (proc mod2).

Figure 2-13 shows a three-processor socket configuration of Power E850.

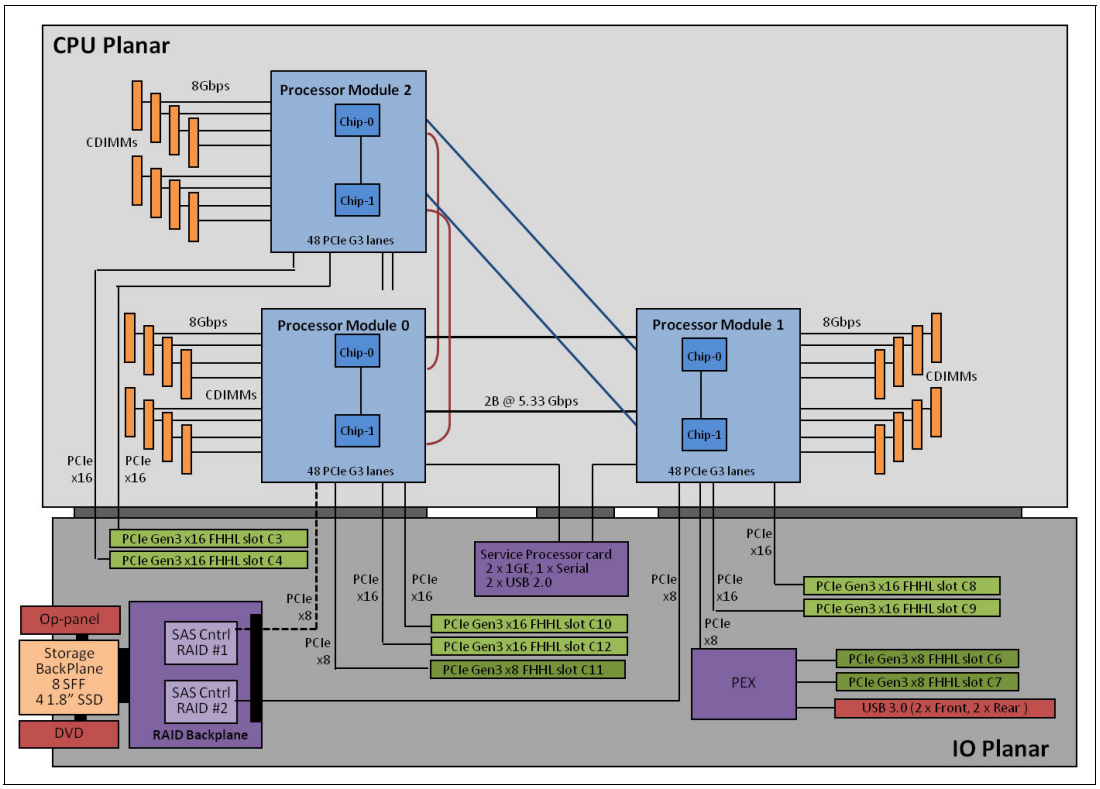


Figure 2-13 Power E850 with three processor modules installed

Four-processor socket configuration

A four-processor module configuration of Power E850 has these characteristics:

- In the four Processor socket configuration, there are eight x16 PCIe3 slots and three x8 PCIe Gen3 slots available in the system unit.
- The I/O workload is balanced across available processor sockets and I/O switches (PEX) that connect the slots to the processor sockets.
  - The first pair of two x16 PCIe Gen3 slots is connected to two different chips in the one POWER8 Dual Chip Module (proc mod0).
  - One x8 PCIe Gen3 slot is connected to a different chip in the same POWER8 Dual Chip Module (proc mod0).
  - The second pair of two x16 PCIe Gen3 slots is connected to two different chips in the second POWER8 Dual Chip Module (proc mod1).
  - One pair of two x8 PCIe Gen3 slot is connected to a different switch (PEX) in the second POWER8 Dual Chip Module (proc mod1).
  - The third pair of x16 PCIe Gen3 slots is connected to two different chips in the third POWER8 Dual Chip Module (proc mod2).
  - The fourth pair of x16 PCIe Gen3 slots is connected to two different chips in the fourth POWER8 Dual Chip Module (proc mod3).

Figure 2-14 shows a four-processor socket configuration of Power E850.

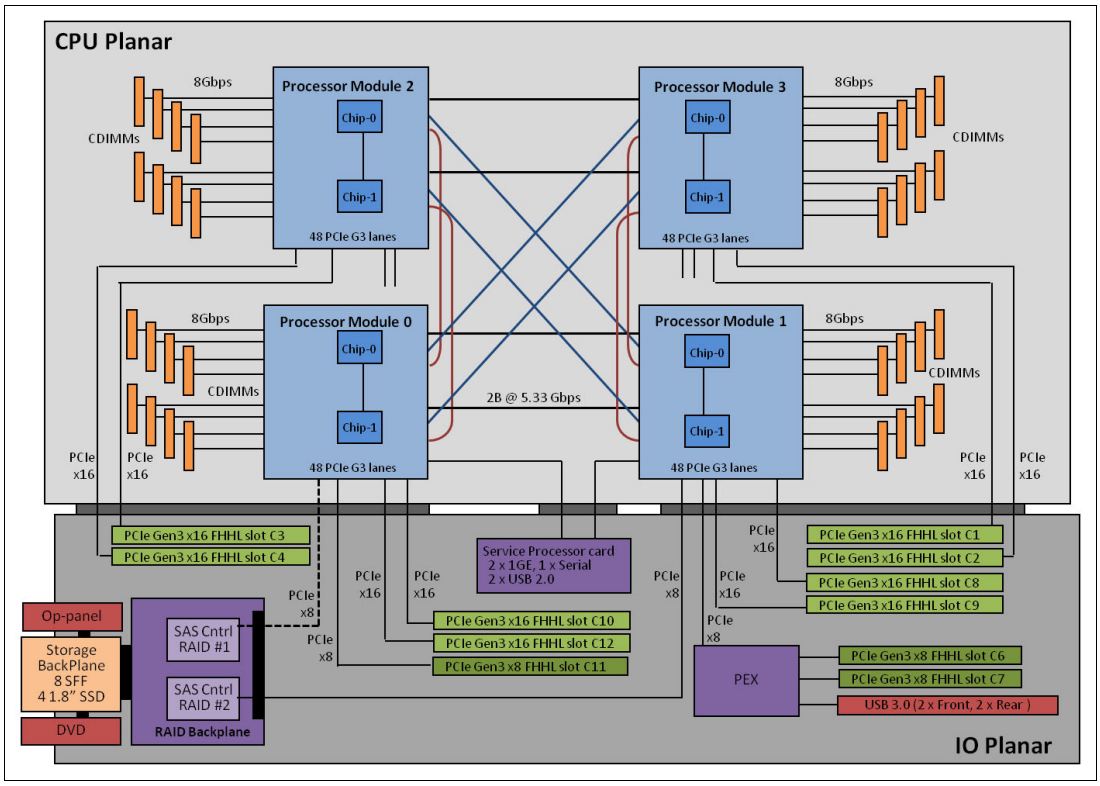


Figure 2-14 Power E850 with four processor modules installed

Note

The slot P1-C11 comes populated with a PCIe2 4-port 1 Gb Ethernet adapter that is required for IBM manufacturing and testing. This slot cannot be used or the adapter must be moved to a different slot if the slot (P1-C11) is required for a different adapter.

Use the IBM System Planning Tool to ensure that the adapter placements are valid for the system configuration that you are building.

The IBM System Planning Tool is available at:

http://www-947.ibm.com/systems/support/tools/systemplanningtool/

For a full list of all supported PCIe adapters for Power E850, see the IBM Knowledge Center or IBM Power System E850 Technical Overview and Introduction:

http://www.redbooks.ibm.com/abstracts/redp5222.html?Open
http://www-01.ibm.com/support/knowledgecenter/POWER8/p8hdx/POWER8welcome.htm

2.9.2 Integrated SAS Controllers

Power E850 system provides three options for the Integrated SAS Controllers configuration:

- #EPVN - Storage backplane dual RAID controllers with write cache
- #EPVP - Storage backplane dual RAID controllers without write cache
- #EPVQ - Split backplane with two RAID controllers without write cache

#EPVN - Storage backplane dual RAID controllers with write cache

#EPVN has these characteristics:

- Dual controller disk backplane with write cache.
- Write cache is 1.8 GB and effectively provides up to 7.2 GB with compression.
- Dual SAS controllers provide redundancy and performance with Active-Active configuration that involves 12 SAS bays with 8 SFF-3 and 4 1.8-inch slots.

#EPVP - Storage backplane dual RAID controllers without write cache

#EPVP has these characteristics:

- Dual controller disk backplane without write cache.
- Zero write cache.
- Dual SAS controllers provide redundancy and performance with Active-Active configuration that involves 12 SAS bays with 8 SFF-3 and 4 1.8-inch slots.

#EPVQ - Split backplane with two RAID controllers without write cache

#EPVQ has these characteristics:

- Split disk backplane (two single controllers) without write cache.
- Each controller gets six SAS slots (6+6) of four SFF-3 and two 1.8 inch slots each.
- The first two options (#EPVN and #EPVP) do not support disks in JBOD configuration.
- All three options provide support for the Easy Tier function. However, the split backplane option (#EPVQ) only supports the RAIDT2 in an Easy Tier configuration. RAID5T2 and RAID6T2 are not supported.

2.9.3 PCIe Gen3 I/O Expansion Drawers

This section provides supported PCIe3 I/O Expansion drawer configurations for the Power E850 system:

- The system supports two PCIe Gen3 Expansion I/O drawers when two processor modules are filled.
- The system supports three PCIe Gen3 Expansion I/O drawers when three processor modules are filled.
- The system supports four PCIe Gen3 Expansion I/O drawers when four processor modules are filled.

For more details about PCIe Gen3 I/O Expansion drawer, see 2.11, “PCIe Gen3 I/O Expansion Drawer Overview” on page 42.

2.9.4 EXP24S SSF Gen-2 Drawer and Internal Disk bays

This section provides information about Internal Disk bays, and supported EXP24S SSF Gen-2 Drawers:

- The system unit internal disk slots vary depending on which of the three backplane options is chosen.
- The integrated SAS controllers do not offer an option to attach an external disk expansion drawer.

- An SAS adapter in the system unit PCIe slots is required if the configuration needs an EXP24S expansion drawer for more disk slots.

2.10 Power E870 and Power E880 I/O Architecture Overview

Power E870 and Power E880 system are highly scalable, which allows greater workload consolidation with large processor and memory capacities. The systems use the latest POWER8 processor technology that is designed to deliver unprecedented performance, scalability, reliability, and manageability for demanding commercial workloads.

Both systems provide PCIe Gen3 x16 slots that support high bandwidths for I/O intensive workloads, and allow for highly dense virtual machine consolidation. The two systems support PCIe Gen3 I/O Expansion drawers for configurations that need extra PCIe Gen3 slots for a higher number of I/O adapters.

Figure 2-15 shows a logical system diagram for Power E870 and E880.

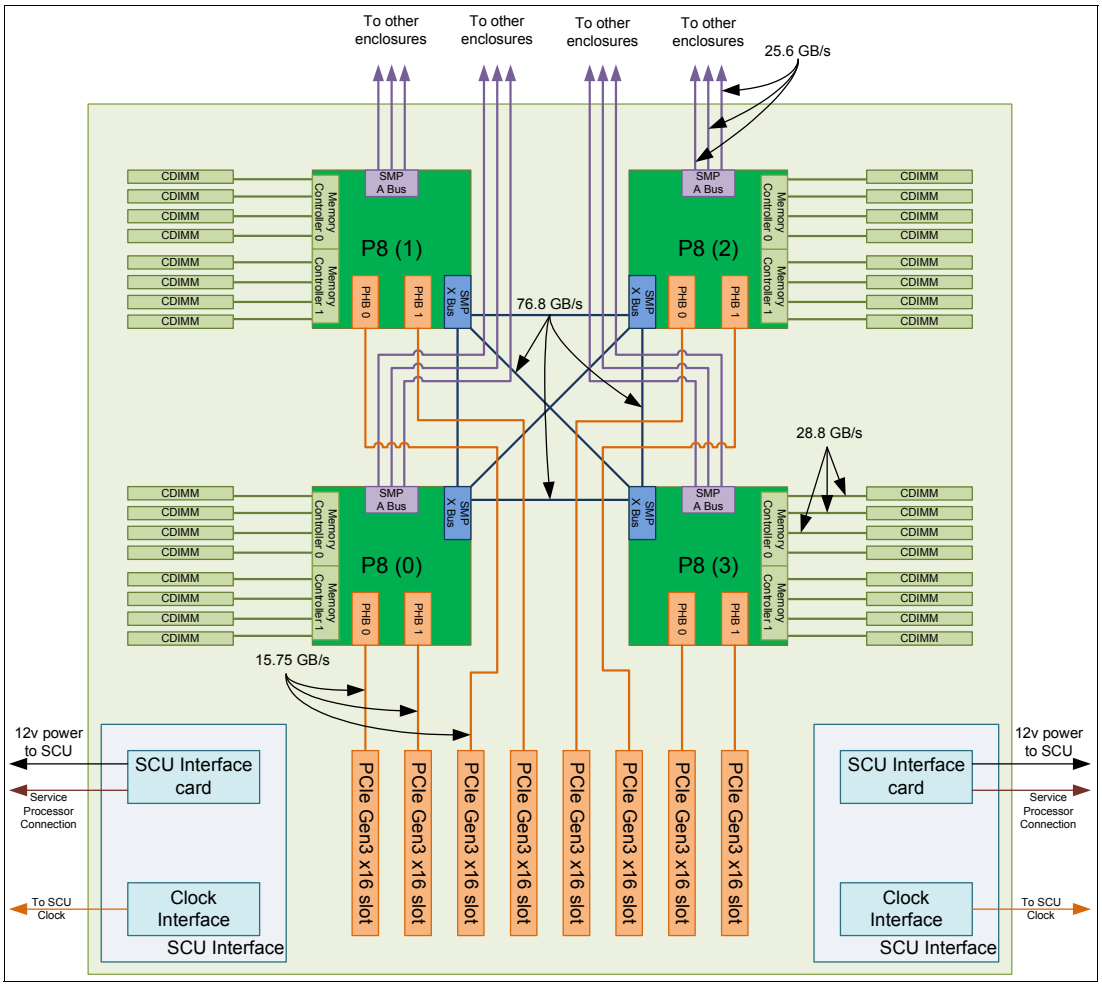


Figure 2-15 Power E870 and E880 Logical System Diagram

Power E870 is a modular-built system that allows minimum of one system node and a maximum of four system nodes:

- Each system node is a four socket enclosure and the two node configuration allows a maximum capacity of 80 POWER8 processor cores.
- The system supports a maximum memory capacity of 8 TB with a total of 64 CDIMM slots.
- The system provides eight PCIe Gen3 x16 slots in each system unit, allows a maximum of eight PCIe Gen3 I/O Expansion drawers, and requires EXP24S SFF Gen2-bay Drawers for internal storage.
- The system units do not support disk bays and require the EXP24S SFF Gen2-bay Drawers for internal storage disks.

Power E880 is modular-built system that allows minimum of one system node and a maximum of four system nodes:

- Each system node is a four socket enclosure and the four node configuration allows a maximum capacity of 192 POWER8 processor cores.
- The System supports a maximum memory capacity of 16 TB with a total of 128 CDIMM slots.
- The system provides eight PCIe Gen3 x16 slots in each system unit, allows a maximum of sixteen PCIe Gen3 I/O Expansion drawers, and requires EXP24S SFF Gen2-bay Drawers for internal storage.
- The system units do not support disk bays and require the EXP24S SFF Gen2-bay Drawers for internal storage disks.

Use the System Planning tool to ensure that the adapter placements are valid for the system configuration that you are building.

The IBM System Planning Tool is available at:

http://www-947.ibm.com/systems/support/tools/systemplanningtool/

For a full list of all supported PCIe adapters for Power E870 and E880, see the IBM Knowledge Center or *IBM Power Systems E870 and E880 Technical Overview and Introduction* at:

http://www.redbooks.ibm.com/abstracts/redp5137.html?Open
http://www-01.ibm.com/support/knowledgecenter/POWER8/p8hdx/POWER8welcome.htm

2.11 PCIe Gen3 I/O Expansion Drawer Overview

The PCIe Gen3 I/O Expansion drawer (#ELMX) is supported on all POWER8 based systems. The drawer is a 4U high, rack-mountable unit. It offers two PCIe Fan Out modules (#ELMF) and each Fan Out module provides six PCIe Gen3 slots.

Figure 2-16 shows the rear view of the PCIe Gen3 drawer.

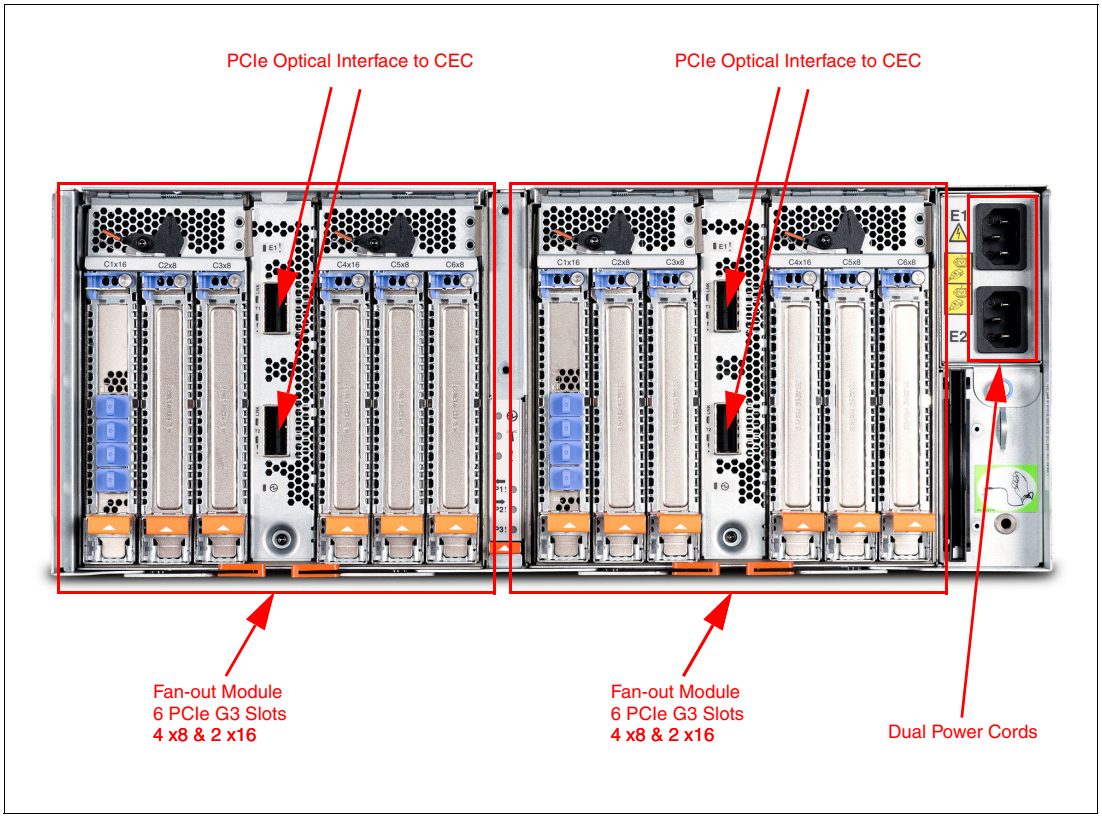


Figure 2-16 Rear View of the PCIe Gen3 I/O expansion drawer

The number of PCIe Gen3 I/O expansion drawers supported in each Scale-Out system varies based on the system model and the number of processor sockets that are installed in the system. Extra processor sockets provide more I/O chips, and so provide more PCIe slots in the system units and more PCIe Gen3 I/O expansion drawers.

Table 2-1 summarizes the maximum number of PCIe Gen3 I/O drawers and the total number of PCIe Gen3 slots for Power Scale-Out systems that are available at the time of writing.

Table 2-1 PCIe Drawer and slot math for Scale-Out Systems

	2U 1-socket	**2U 2-socket One Filled**	**2U 2-socket Two Filled**	**4U 1-socket [a]**	**4U 2-socket One Filled**	**4U 2-socket Two Filled**
	S812L	S822	S822, S822L	S814	S824	S824, S824L*
PCIe slots in System Unit	6	6	9	7	7	11
x16 slots in System unit	2	2	4	2	2	4
Max PCIe Gen3 Expansion Drawer	1/2	1/2	1	1	1	2

	2U 1-socket	2U 2-socket One Filled	2U 2-socket Two Filled	4U 1-socket [a]	4U 2-socket One Filled	4U 2-socket Two Filled
PCIe Gen3 Drawer slots	6	6	12	12	12	24
PCIe slots used for Optical cable adapter	2[b]	2[b]	4[b]	2	2	2
Total Max PCIe Slots	6-2+6= 10	6-2+6= 10	9-4+12= 17	7-2+12= 17	7-2+12= 17	11-4+24= 31

a. Requires a 6-core or 8-core server. The 4-core server does not support an I/O drawer.

b. 2U uses a double-wide Optical Cable Adapter that uses two PCIe slots per fan-out module. Because of the x16 slot location, the system can only use half of the x16 slots for attaching an I/O drawer.

Note: If the S824L is using a GPU, the system doesn't support I/O drawers

Power E870 and E880 support a maximum number of PCIe Gen3 I/O Expansion drawers based on the number of nodes or system units that are used in the server configuration. The number of PCIe Gen3 slots in the system units that are usable for I/O cards depends on the number of I/O expansion drawers because the system unit slots are used by the Optical Adapter for connecting the I/O expansion drawers.

Table 2-2 summarizes the maximum number of PCIe Gen3 I/O expansion drawers based on the number of system units, and the maximum number of configurable PCIe Gen3 slots after reserving the required number of slots in the system units to connect the I/O expansion drawers.

Table 2-2 PCIe Gen3 slots for Power E870 and E880

Possible Number of PCIe drawers on the node	PCIe slots per 1-node server	PCIe slots per 2-node server	PCIe slots per 3-node server	PCIe slots per 4-node server
0	8	16	24	32
1	18 12 in drawer + 6 in system unit	26 12 in drawer + 14 in system unit	34 12 in drawer + 22 in system unit	42 12 in drawer + 30 in system unit
2	28 24 in drawer + 4 in system unit	36 24 in drawer + 12 in system unit	44 24 in drawer + 20 in system unit	52 24 in drawer + 28 in system unit
3	38 36 in drawer + 2 in system unit	46 36 in drawer + 10 in system unit	54 36 in drawer + 18 in system unit	62 36 in drawer + 26 in system unit

Possible Number of PCIe drawers on the node	PCIe slots per 1-node server	PCIe slots per 2-node server	PCIe slots per 3-node server	PCIe slots per 4-node server
4	48 (maximum) 48 in drawer + 0 in system unit	56 48 in drawer + 8 in system unit	64 48 in drawer + 16 in system unit	72 48 in drawer + 24 in system unit
6	n/a	76 72 in drawer + 4 in system unit	84 72 in drawer + 12 in system unit	92 72 in drawer + 20 in system unit
8	n/a	96 (maximum) 96 in drawer + 0 in system unit	104 96 in drawer + 8 in system unit	112 96 in drawer + 16 in system unit
12	n/a	n/a	144 (maximum) 144 in drawer + 0 in system unit	152 144 in drawer + 8 in system unit
16	n/a	n/a	n/a	192 (maximum) 192 in drawer + 0 in system unit

3

RAID adapters for IBM Power Systems

This chapter describes the supported PCIe and Integrated SAS RAID adapters that are supported in POWER8 processor-based systems.

This chapter also describes the following topics:

- High Availability features for SAS RAID adapters
- Performance characteristics
- IBM Disk formatting practices

This chapter includes the following sections:

- IBM Power Systems and options for RAID adapters
- POWER8 processor based systems and supported PCIe RAID adapters
- POWER8 based processor systems and internal RAID adapters
- PCIe SAS RAID adapters with write cache
- General guidelines for selecting SAS RAID adapters
- High Availability feature considerations
- PCIe SAS RAID adapters
- IBM disk formatting practices
- VIOS vSCSI disks and IBM i client partitions
- RAID adapters performance characteristics
- SSDs and Easy Tier array
- SAS RAID adapters performance comparison

3.1 IBM Power Systems and options for RAID adapters

POWER8 processor-based systems provide various options to install DAS devices in RAID configurations. The integrated RAID SAS controllers in supported IBM Power Systems can be used to configure the integrated SFF-3 disk bays in the system units with RAID configurations. Some feature codes of the integrated controllers provide support to connect to expansion drawer EXP24S, which contains 24 SFF-2 disk bays. Some of the POWER8 processor-based system models also support an 1.8 inch cage for SFF-3 SSDs. Certain feature codes in the integrated controllers also provide support for the Easy Tier feature with hard disk drives (HDDs) and solid-state drives (SSDs) for optimal application I/O performance.

IBM Power Systems also support various PCIe based SAS RAID adapters that can be used to connect EXP24S drawer SFF-2 disk bays. This chapter provides detailed information about each of supported adapters and their features. For guidelines about selecting adapters based on specific requirements while performing system configurations, see 3.5, “General guidelines for selecting SAS RAID adapters” on page 51. The following section provides the supported adapters, and the different options to configure the integrated SAS RAID controllers based on specific configuration requirements.

3.2 POWER8 processor based systems and supported PCIe RAID adapters

This section provides the list of supported PCIe based SAS RAID adapters supported in all IBM Power Systems system units and the PCIe3 I/O Expansion Drawers that are supported with Power Systems. The full profile adapters are not supported in the 2U system units, but they are supported in the PCIe Gen3 I/O expansion drawers connected to the 2U system units. The configurations 2U system units require attaching the PCIe Gen3 I/O expansion drawers to support the full height cards.

Power S812L and Power S822L

The following are PCIe SAS RAID adapters for Power S812L and S822L systems:

- #5805 - PCIe 380MB Cache Dual -x4 3Gb SAS RAID Adapter
- #ESA3 - PCIe2 1.8GB Cache RAID SAS Tri-port 6Gb Adapter
- #5913 - PCIe2 1.8GB cache RAID SAS Tri-port 6Gb adapter
- #EJ0L - PCIe3 12GB Cache RAID SAS adapter Quad-port, 6Gb x8
- #EL3B - PCIe3 LP RAID SAS adapter Quad-port 6Gb x8
- #EL59 - PCIe3 RAID SAS adapter Quad-port 6Gb x8

Power S814 and Power S824

The following are supported PCIe SAS RAID adapters for Power S814 and S824 systems:

- #5805 - PCIe 380MB Cache Dual -x4 3Gb SAS RAID Adapter
- #ESA3 - PCIe2 1.8GB Cache RAID SAS Tri-port 6Gb Adapter
- #5913 - PCIe2 1.8GB cache RAID SAS Tri-port 6Gb adapter
- #EJ0J - PCIe3 RAID SAS adapter Quad-port, Low Profile Capable 6Gb x8
  - #EJ0J is for DAS attach only, requires EJ10 for Tape / DVD attach

- #EJ0L - PCIe3 12GB Cache RAID SAS adapter Quad-port, 6Gb x8
- #5901 - PCIe dual-x4 SAS RAID adapter

Power S822

The following are supported PCIe SAS RAID adapters for Power S822 systems:

- #5805 - PCIe 380MB Cache Dual -x4 3Gb SAS RAID Adapter
- #ESA3 - PCIe2 1.8GB Cache RAID SAS Tri-port 6Gb Adapter
- #5913 - PCIe2 1.8GB cache RAID SAS Tri-port 6Gb adapter
- #EJ0J - PCIe3 RAID SAS adapter Quad-port, Low Profile Capable 6Gb x8
  - #EJ0J is for DAS attach only, requires EJ10 for Tape / DVD attach
- #EJ0L - PCIe3 12GB Cache RAID SAS adapter Quad-port, 6Gb x8
- #EJ0M - PCIe3 RAID SAS adapter Quad-port, Low Profile Short, 6Gb x8
  - #EJ0M is for DAS attach only, requires EJ11 for Tape / DVD attach
- #5901 - PCIe dual-x4 SAS RAID adapter

Power S824L

The following are supported PCIe SAS RAID adapters for Power S824L systems:

- #EJ0L - PCIe3 12GB Cache RAID SAS adapter Quad-port, Not Low Profile Capable, 6Gb x8
- #EL59 - PCIe3 RAID SAS adapter Quad-port 6Gb x8

Power E850

The following are supported PCIe SAS RAID adapters for Power E850 systems:

- #EJ0J - PCIe3 RAID SAS adapter Quad-port, Low Profile Capable 6Gb x8
  - #EJ0J is for DAS attach only, and requires EJ10 for Tape / DVD attach
- #EJ0L - PCIe3 12GB Cache RAID SAS adapter Quad-port, 6Gb x8

Power E870 and Power 880

The following are supported PCIe SAS RAID adapters for Power E870 and E880 systems:

- #5805 - PCIe 380MB Cache Dual -x4 3Gb SAS RAID Adapter
- #ESA3 - PCIe2 1.8GB Cache RAID SAS Tri-port 6Gb Adapter
- #5913 - PCIe2 1.8GB cache RAID SAS Tri-port 6Gb adapter
- #EJ0J - PCIe3 RAID SAS adapter Quad-port, Low Profile Capable 6Gb x8
  - EJ0J is for DAS attach only, and requires EJ10 for Tape / DVD attach
- #EJ0L - PCIe3 12GB Cache RAID SAS adapter Quad-port, 6Gb x8
- #EJ0M - PCIe3 RAID SAS adapter Quad-port, Low Profile Short, 6Gb x8
  - #EJ0M is for DAS attach only, and requires EJ11 for Tape / DVD attach
- #5901 - PCIe dual-x4 SAS RAID adapter

Note: #5805, #ESA3, #5913, #5901 RAID adapters (non-PCIe Gen3) are supported in POWER8 processor based systems, but are not available for new ordering.

The Power E850 does not support the #5805 and #5913 adapter feature codes.

PCIe Gen3 I/O Expansion Drawer

The following adapters are supported in PCIe Gen3 I/O Drawer. PCIe Gen3 I/O expansion drawers are required in configurations that need these adapters for 2U system unit models:

- #5805 - PCIe 380MB Cache Dual -x4 3Gb SAS RAID Adapter
- #ESA3 - PCIe2 1.8GB Cache RAID SAS Tri-port 6Gb Adapter
- #5913 - PCIe2 1.8GB cache RAID SAS Tri-port 6Gb adapter
- #EJ0L - PCIe3 12GB Cache RAID SAS adapter Quad-port 6Gb x8
- #EL59 - PCIe3 RAID SAS adapter Quad-port 6Gb x8

3.3 POWER8 based processor systems and internal RAID adapters

This section provides a brief description of the different options that are available to configure the integrated SAS RAID controllers based on specific configuration requirements.

Power S814, S824, S24L

The following list shows the options for internal storage controllers configuration in Power S814, S824, and S24L:

- #EJ0N - Base function, storage backplane 12 SFF-3 bays / DVD bay
- #EJ0P - Expanded function, storage backplane 18 SFF-3 bays / 1.8 inch SSD Attach / DVD bay / Dual IOA write cache
- #EJ0S - Base Function with Split Feature, Split #EJ0N to 6+6 SFF-3 bays

Power S822

The following list shows the options for internal storage controllers configuration in Power S822:

- #EJ0T - Base function, Storage backplane 12 SFF-3 bays / DVD bay
- #EJ0U - Expanded function, Storage backplane quantity 18 SFF-3 bays / quantity 6 1.8 inch SSD Attach / DVD bay / Dual IOA write cache
- #EJ0V - Base Function with Split Feature, Split #EJ0U to 6+6 SFF-3 bays

Power S812L, S822L

The following list shows the options for internal storage controllers configuration in Power S812L and S822L:

- #EL3T - Base Function, Storage backplane 12 SFF-3 bays / DVD bay
- #EL3U - Expanded function, Storage backplane 8 SFF-3 bays / quantity 6 1.8 inch SSD Attach / DVD bay / Dual IOA write cache
- #EL3V - Base Function with Split Feature, Split #EL3T to 6+6 SFF-3 bays

Power E850

The following list shows the options for internal storage controllers configuration in Power S850:

- #EPVN - Storage backplane with dual RAID controllers with write cache, quantity 8 SFF-3 and quantity 4 1.8inch bays
- #EPVP - Storage backplane with dual RAID controllers without write cache, quantity 8 SFF-3 and quantity 4 1.8inch bays
- #EPVQ - Split backplane with two RAID controllers without write cache, quantity 4 SFF-3 and quantity 2 1.8inch + quantity 4SFF-3 and quantity 2 1.8inch bays

3.4 PCIe SAS RAID adapters with write cache

The following is a list of PCIe SAS RAID adapters that provide write cache feature and are supported in POWER8 processor-based systems:

- #EJ0L - PCIe3 12GB Cache RAID SAS adapter Quad-port, 6Gb x8
- #5805 - PCIe 380MB Cache Dual -x4 3Gb SAS RAID Adapter
- #5913 - PCIe2 1.8GB cache RAID SAS Tri-port 6Gb adapter
- #ESA3 - PCIe2 1.8GB Cache RAID SAS Tri-port 6Gb Adapter

3.5 General guidelines for selecting SAS RAID adapters

Building new system configurations requires an understanding of the requirements for RAID levels in the configuration, the features available in all integrated SAS RAID controllers, and the supported PCIe SAS RAID adapters. This section provides general guidelines for selecting from the available RAID adapters based on individual needs:

- First criteria in selecting SAS RAID adapters is based on the required RAID level in the configuration, and the levels supported in the adapter.
- Some of the PCIe3 SAS RAID adapters offer Easy Tier feature. Select the specific RAID adapter if Easy Tier feature is wanted for the configuration.
- Some of the RAID adapters do not support just a bunch of disks (JBOD) configuration for the disk drives for AIX / Linux.
  - SSDs are not supported in JBOD configurations.
- There are different feature codes for the disks to be used in AIX/ Linux or IBM i systems.
- Systems that have only Low Profile PCIe Gen3 slots in the system units (2U) will need PCI3 Gen3 I/O Expansion Drawer to use RAID adapters that are not Low Profile Capable.
- RAID Adapters that have write cache require mandatory pairing, and for adapters with no cache, pairing is optional.
  - Adapters in pairing configuration provide protection against one adapter failure.
- Some Low profile adapters can be paired with equivalent Full profile adapters.
  - The Low Profile adapter in the pair can be in the system unit, and the Full Height adapter in the pair can be in the PCIe Gen3 I/O expansion drawer.
  - This configuration enhances protection against single adapter failure and loss to access to the I/O expansion drawer that is hosting the second adapter in the pair.

- Consider using RAID 0 in place of JBOD configuration:
  - RAID 0 with write cache typically provides better performance than JBOD disks, which do not use the write cache in the adapters.
  - RAID 0 also provides the T10DIF fields in every sector in the RAID formatted disks. For more information about the T10DIF field, see 3.8, “IBM disk formatting practices” on page 70.
- Select Adapters with write cache for workloads that demand high performance and are performance sensitive to disk write operations.
- #5805, #ESA3, #5913 RAID adapters (non-PCIe Gen3) are supported in POWER8 processor based systems, but are not available for new ordering.
- High Availability features of the adapters depend on the operating systems that are used in the LPARs. See “High Availability feature considerations” on page 54 for a detailed comparison of the High Availability features supported by the adapters based on the operating system.

Table 3-1 provides a summary of the features supported by each SAS RAID adapter.

Table 3-1 PCIe RAID Adapters Feature Comparison

Card Feature Code	CCIN	Supported RAID Levels	Easy Tier Support	Write Cache	Pairing
#EJ0L Full Height, Not Low Profile Capable PCIe3 x8	57CE	► RAID 0, RAID 1, RAID 5, RAID 6, and RAID 10 for AIX/Linux ► RAID 5, RAID 6, and RAID 10* for IBM i	None	up to 12 GB	Required
#5913 PCIe2 x8	57B5	► RAID 0, RAID 1, RAID 5, RAID 6, and RAID 10 for AIX/Linux ► RAID 5 and RAID 6 for IBM i	None	1.8 GB	Required
#ESA3 PCIe2 x8	57BB	► RAID 0, RAID 1, RAID 5, RAID 6, and RAID 10 for AIX/Linux ► RAID 5 and RAID 6 for IBM i	None	1.8 GB	Required
#EJ0J Full Height, Low Profile Capable PCIe3 x8	57B4	► RAID 0, RAID 1, RAID 5, RAID 6, and RAID 10 for AIX/Linux ► RAID 5, RAID 6, and RAID 10* for IBM i	None	None	Optional
#5805	574E	► RAID 0, RAID 1, RAID 5, RAID 6, and RAID 10 for AIX/Linux ► RAID 5 and RAID 6 for IBM i	None	380 MB	Required

Card Feature Code	CCIN	Supported RAID Levels	Easy Tier Support	Write Cache	Pairing
#EL3B Low Profile PCIe3 x8	57B4	► RAID 0, RAID 1, RAID 5, RAID 6, and RAID 10 for Linux/VIOS	None	None	Optional (#EL3B or #EJ0J)
#EJ0M Low Profile PCIe3 x8	57B4	► RAID 0, RAID 1, RAID 5, RAID 6, and RAID 10 for AIX/Linux ► RAID 5, RAID 6, and RAID 10* for IBM i	None	None	Optional (#EJ0M or #EJ0J)
#EL59 Low Profile PCIe3 x8	57B4	► RAID 0, RAID 1, RAID 5, RAID 6, and RAID 10 for Linux/VIOS	None	None	Optional (#EL59 or #EJ0J)
#5901	57B3	► RAID 0 and RAID 10 for AIX and Linux	None	None	Optional (#5901)

Note: RAID 10 support for IBM i requires IBM i Version 7.2 or later.

Table 3-2 provides the summary of features supported in the integrated SAS adapters in the system units.

Table 3-2 Integrated Adapters Feature Comparison

Card Feature Code	CCIN	Supported RAID Levels	Easy Tier Support	Write Cache	Pairing
#EJ0N, #EJ0S #EL3T, #EL3V #EJ0T, #EJ0V	57D7	► RAID 0, RAID 1, RAID 5, RAID 6, RAID 10 ► Supports JBOD HDDs only	► None	► None	► No
#EJ0P #EL3U #EJ0U	57D8	► RAID 0, RAID 1, RAID 5, RAID 6, and RAID 10 for AIX/Linux ► RAID 5, RAID 6, and RAID 10* for IBM i	► RAID 5T2, RAID 6T2, RAID 10T2 for AIX/Linux/VIOS	► Upto 7.2 GB	► Yes
#EPVN	2CCA	► RAID 0, RAID 1, RAID 5, RAID 6, RAID 10 for AIX/Linux	► RAID 5T2, RAID 6T2, RAID 10T2 for AIX/Linux/VIOS	► Upto 7.2 GB	► Yes
#EPVP	2CD2	► RAID 0, RAID 1, RAID 5, RAID 6, RAID 10 for AIX/Linux	► RAID 5T2, RAID 6T2, RAID 10T2 for AIX/Linux/VIOS	► None	► Yes

Card Feature Code	CCIN	Supported RAID Levels	Easy Tier Support	Write Cache	Pairing
#EPVQ	2CCD	► RAID 0, RAID 1, RAID 5, RAID 6, RAID 10 for AIX/Linux ► Supports JBOD	► RAID 10T2 for AIX/Linux/VIOS	► None	► No

3.6 High Availability feature considerations

The integrated RAID SAS controllers, and PCIe RAID SAS controllers support High Availability features for providing adapter redundancy. The High Availability features differ based on the operating system that is used in the configuration. There are different sets of features offered in AIX, Linux, and VIOS environments. These are different from the features offered for IBM i environments. This section provides detailed information for all available High Availability features offered by the RAID SAS controllers.

Table 3-3 on page 58 and Table 3-4 on page 59 provide the comparison of High Availability features to help build the system configurations based on the requirements.

3.6.1 High Availability features for AIX and Linux

There are two types of High Availability features exist for AIX and Linux operating system environments. It is important to understand which adapters must be used for the RAID configuration that you will use so you can choose the RAID adapter that is best suited for the system requirements.

3.6.2 High Availability two system RAID

This HA feature allows two RAID controllers to connect the same RAID Disk array, with disks that are formatted with one of the supported RAID levels. The two RAID controllers can be in two different systems, allowing the operating system in the two systems or LPARs to access the same set of disks.

This feature is typically used with High Availability software such as IBM PowerHA for AIX that uses shared disks in two clustered systems or LPARs in two different physical servers. System configurations that need IBM PowerHA clustering and DAS model to access disks, can use this feature to provide shared storage to the cluster without depending on networked storage solutions such as SAN.

See Table 3-3 on page 58 for adapters that support HA Single System RAID configuration for AIX and Linux.

Figure 3-1 shows a typical architecture.

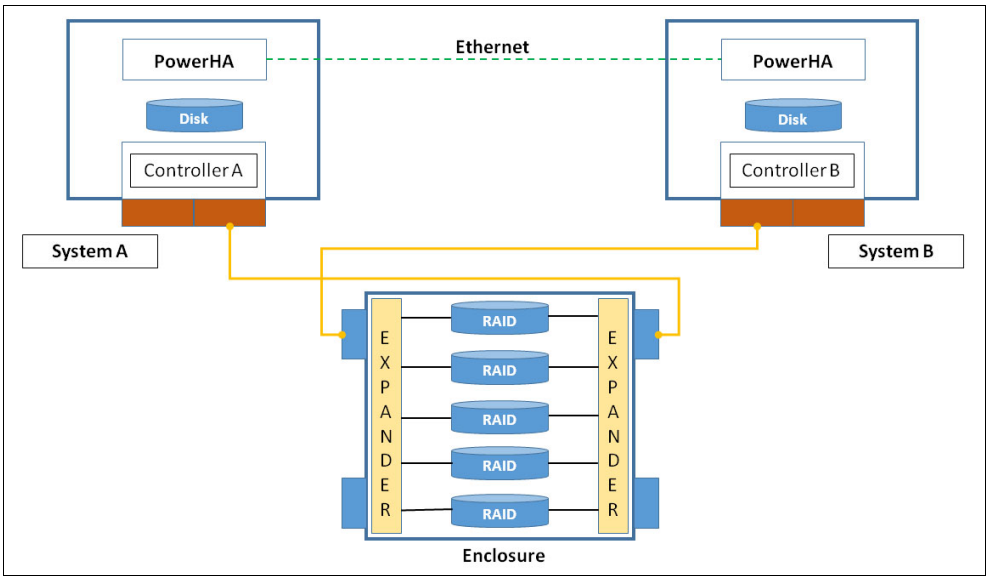


Figure 3-1 Logical diagram for HA two system RAID for AIX/Linux

3.6.3 High Availability single system RAID

This HA feature allows two RAID controllers to connect the same RAID Disk array, with disks that are formatted with one of the supported RAID levels. The two RAID controllers are in the same physical system and can be mapped to one LPAR running the AIX or Linux operating system.

This feature is typically used with the AIX multipath I/O software (MPIO). MPIO provides path management functions to provide SAS RAID adapter redundancy for each RAID formatted disk in the array.

This feature is also supported by the Linux multipath I/O provided by Device-Mapper Multipath in Linux for SAS RAID adapter redundancy for each RAID formatted disk in the array.

For a list of adapters that support HA Single System RAID configuration for AIX and Linux, see Table 3-3 on page 58.

Figure 3-2 provides an overview of a typical architecture for this configuration.

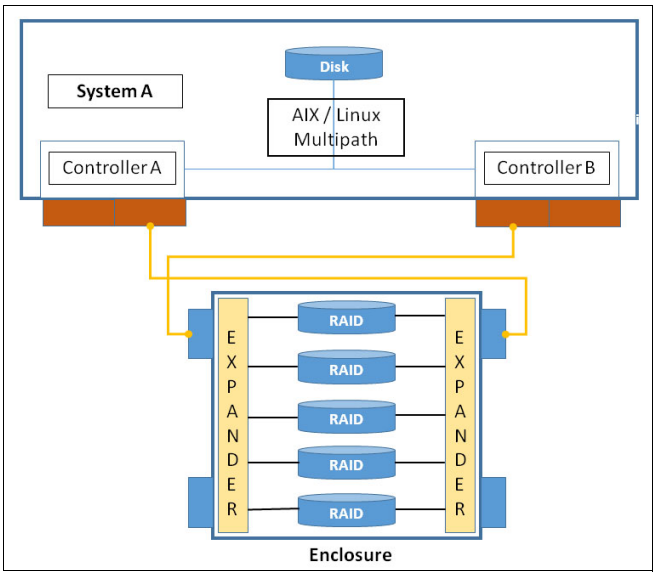


Figure 3-2 Logical diagram for HA Single System for AIX/Linux

3.6.4 High Availability access optimization

When two SAS RAID adapters are used in pairs for the High Availability features, one of the adapters in the pair is chosen to become the optimized adapter for I/O operations to the RAID array. The optimized controller for the RAID array runs the I/O operations for the array, whereas the second adapter is configured in passive mode. The second adapter takes over I/O operations when the primary adapter is lost or failed.

To achieve higher optimization and load balancing across the two available adapters, the preferred practice is configuring multiple RAID arrays in each pair of controllers. Each controller can be defined to be the optimized controller for different RAID arrays managed by the pair.

3.6.5 Just a bunch of disks (JBOD)

JBOD is the configuration to connect to disks that are not RAID formatted and when using the SAS RAID adapter without any RAID levels configured. The controller primarily acts as a pass through device for the I/O operations and does not perform any of the hardware optimizations that RAID enabled adapter performs.

The write cache is not enabled when the adapter is used to connect to non-RAID or JBOD configurations. There are also internal hardware optimizations to access the DRAMs on the adapter hardware when the adapter is configured in one of the supported RAID levels. However, these optimizations are not used when the adapter is connecting to a set of JBOD disks.

Configurations that use JBOD with SAS RAID controllers should configure the adapters with RAID 0 instead of the JBOD configuration. This configuration enables the adapters with write cache and other internal hardware optimizations that provide additional protection than JBOD with no performance penalty for using RAID 0.

Not all SAS RAID adapters in IBM Power Systems support JBOD configurations. See Table 3-3 on page 58 for adapters that support JBOD configuration.

3.6.6 High Availability features for IBM i

The SAS RAID adapter high availability feature for IBM i is provided by the Dual Storage I/O Adapter (IOA) configuration to use two SAS RAID controllers in the same set of RAID arrays.

Figure 3-3 shows the logical diagram of the Dual Storage IOA adapter configuration using two SAS RAID adapters in IBM i system.

To have an optimized Dual Storage IOA configuration and achieve load balance across the two adapters in the pair, use an even number of RAID arrays in each pair. Each of the adapter in the pair is optimized for different RAID arrays, and hence the I/O operations are performed across both adapters to provide enhanced performance and load distributed optimizations. You can choose performance optimization before creating parity sets. This action causes the system to automatically create two or more arrays per adapter pair.

Figure 3-3 shows a single array.

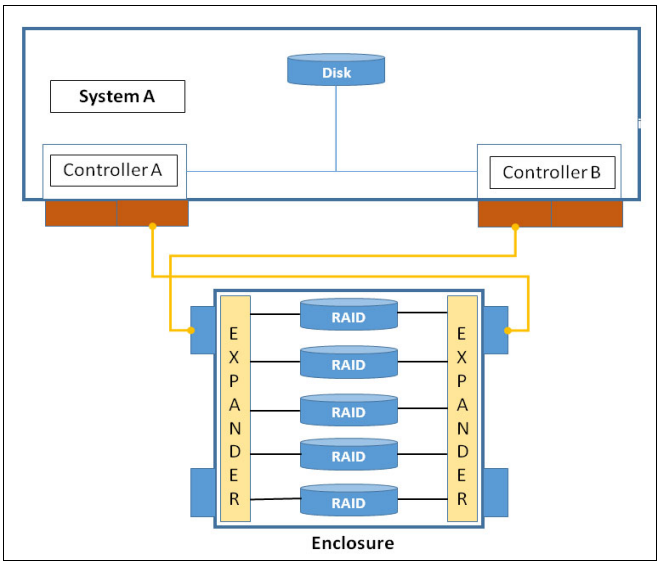


Figure 3-3 Logical diagram for Dual Storage IOA Configuration

Figure 3-4 shows an IBM i system that uses two storage controllers in a Dual IOA storage configuration. The two SAS RAID adapters in the pair are connected to two RAID arrays. Controller A is optimized for RAID array A and handles all read and write operations for RAID array A, while controller B is optimized for RAID array B and handles all read and write operations for RAID array B. The two adapters are also passive adapters to each other in the pair. Therefore, if one of the adapters in the pair fails, the second adapter performs all read and write operations for both arrays.

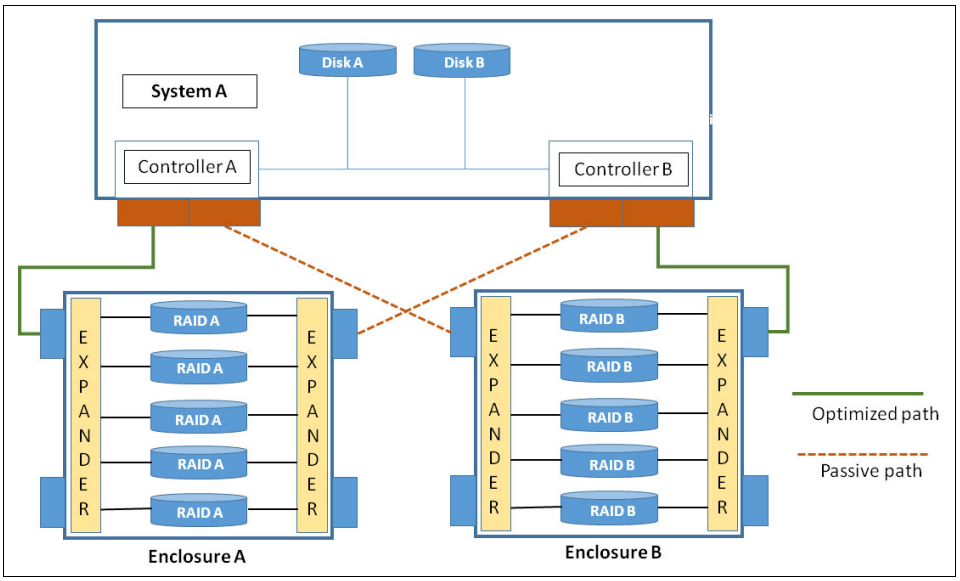


Figure 3-4 Logical diagram for Dual Storage IOA with two RAID arrays

3.6.7 High Availability feature comparison

Table 3-3 provides the High Availability (HA) features provided with the PCIe SAS RAID adapters supported by IBM Power Systems.

Table 3-3 PCIe RAID Adapter high availability features considerations

Card Feature Code	CCIN	HA Features for AIX	HA Features for IBM i	HA Features for Linux
#EJ0L	57CE	► HA two system RAID ► HA single system RAID ► HA RAID required ► No JBOD support	► Requires dual-storage IOA configuration	► HA two system RAID ► HA single system RAID ► HA RAID required ► No JBOD support
#5913	57B5	► HA two system RAID ► HA single system RAID ► HA RAID required ► No JBOD support	► Requires dual-storage IOA configuration	► HA two system RAID ► HA single system RAID ► HA RAID required ► No JBOD support
#ESA3	57BB	► HA two system RAID ► HA single system RAID ► HA RAID required ► No JBOD support	► Requires dual-storage IOA configuration	► HA two system RAID ► HA single system RAID ► HA RAID required ► No JBOD support

Card Feature Code	CCIN	HA Features for AIX	HA Features for IBM i	HA Features for Linux
#EJ0J	57B4	► HA two system RAID ► HA single system RAID ► Supports JBOD for HDDs only	► Dual-storage IOA not supported	► HA two system RAID ► HA single system RAID ► Supports JBOD for HDDs only
#5805	574E	► HA two system RAID ► HA single system RAID ► HA RAID required ► No JBOD support	► Requires dual-storage IOA configuration	► HA two system RAID ► HA single system RAID ► HA RAID required ► No JBOD support
#EL3B	57B4	► HA two system RAID ► HA single system RAID ► Supports JBOD for HDDs only	► Dual-storage IOA not supported	► HA two system RAID ► HA single system RAID ► Supports JBOD for HDDs only
#EJ0M	57B4	► HA two system RAID ► HA single system RAID ► Supports JBOD for HDDs only	► Dual-storage IOA not supported	► HA two system RAID ► HA single system RAID ► Supports JBOD for HDDs only
#EL59	57B4	► HA two system RAID ► HA single system RAID ► Supports JBOD for HDDs only	► Dual-storage IOA not supported	► HA two system RAID ► HA single system RAID ► Supports JBOD for HDDs only
#5901	57B3	► HA Two System RAID ► HA Two System JBOD ► HA single system JBOD ► Supports JBOD for HDDs only	► Dual-storage IOA not supported	► HA Two System RAID ► HA Two System JBOD ► HA single system JBOD ► Supports JBOD for HDDs only

Table 3-4 provides the High Availability features provided with the integrated SAS RAID controllers supported in IBM Power Systems.

Table 3-4 Internal SAS Controllers High Availability Features

Card Feature Code	CCIN	HA Features for AIX	HA Features for IBM i	HA Features for Linux
#EJ0N, #EJ0S #EL3T, #EL3V #EJ0T, #EJ0V	57D7	► Supports JBOD for HDDs only	► Dual-storage IOA not supported	► Supports JBOD for HDDs only
#EJ0P #EL3U #EJ0U	57D8	► HA single system RAID ► HA RAID required ► No JBOD support	► Requires dual-storage IOA configuration	► HA single system RAID ► HA RAID required ► No JBOD support

Card Feature Code	CCIN	HA Features for AIX	HA Features for IBM i	HA Features for Linux
#EPVN	2CCA	► HA single system RAID ► HA RAID required ► No JBOD support	► N/A	► HA single system RAID ► HA RAID required ► No JBOD support
#EPVP	2CD2	► HA Single System RAID ► Requires HA RAID ► No JBOD	► N/A	► HA single system RAID ► HA RAID required ► No JBOD support
#EPVQ	2CCD	► Supports JBOD for HDDs only	► N/A	► Supports JBOD for HDDs only

3.7 PCIe SAS RAID adapters

This section provides detailed information about the PCIe SAS RAID adapters that are supported in POWER8 processor-based systems. The information provided for each adapter includes RAID levels, supported disk drives, adapter pairing configurations, write cache, cabling requirements, and the supported system models.

3.7.1 #5805 - PCIe 380MB cache Dual -x4 3Gb SAS RAID Adapter

A short, full high form factor adapter that supports SAS and SSD devices. The adapter provides 3 Gbs of SAS speed and 380 MB of non-volatile fast write cache that is mirrored with a dual adapter configuration. The adapter supports Concurrent Firmware Updates.

RAID levels and supported disk drives

The following are RAID levels and supported disk drives for this adapter:

- Supports RAID 0, RAID 5, RAID 6, and RAID 10 for AIX/Linux formatted HDD and SSD.
- Supports RAID 5 and RAID 6 for IBM i formatted HDD and SSD.
- SAS bays in EXP24S or PCIe 12X I/O Drawer or EXP 12 Disk Drawer are supported.
- Two adapters in one pair of #5805 can be in two different systems in High Availability configurations for AIX, Linux.
  - IBM i does not support adapter pairing that is in two different systems.
- Supports up to 48 SAS disk drives when configured with two EXP24S (#5887) drawers.
- This adapter supports multiple wide port connections to dual port SAS or SSD for path redundancy.
  - Performs automatic path switching if one of the paths to the disks fail.

Adapter pairing and write cache

This adapter supports the following configurations for adapter pairing in High Availability mode. The adapter's write cache details are also provided below.

- #5805 is always installed in pairs and the pairing provides high availability configuration to protect against failure of one of the adapters in the pair.
- The adapter supports pairing with two options:
  - Two #5805.
  - One #5805 and one #5903.
- The pairs provide mirroring for the write cache.
- Cache mirroring is disabled if the adapter pair is broken or one of the adapters is faulty.
- Write cache provides performance improvement even when configured without RAID 5/6/10.

Cabling requirements

The following are the cabling guidelines for the adapter:

- When attaching to disks in the #5887 EXP24S drawer, cable types SAS (X) #3661, #3662, and #3663 are supported.

For a detailed description of cabling in the system units, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling.htm

For a detailed description of cabling in the PCIe Gen3 Expansion I/O drawer, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling5887.htm

POWER8 processor-based models supported

The following POWER8 processor-based system models are supported by this adapter:

- Power S812L, Power 822L, Power S814, Power S822, Power824
- Power E870, Power E880

3.7.2 #ESA3 - PCIe2 1.8GB Cache RAID SAS Tri-port 6Gb Adapter

A PCIe Gen2 SAS RAID adapter provides 1.8 GB of write cache, and tri-port with 6 Gb of SAS speed

RAID levels and supported disk drives

The following are the RAID levels and supported disk drives for this adapter:

- Supports RAID 0, RAID 5, RAID 6, and RAID 10 for AIX/Linux/VIOS formatted HDD and SSD.
- Supports RAID 5 and RAID 6 for IBM i formatted HDD and SSD.
- SAS bays in EXP24S or PCIe 12X I/O Drawer or EXP12S Disk Drawers are supported.
- The Tri-port adapter supports up to connection for up to three EXP24S (#5887) drawers.

Adapter pairing and write cache

This adapter supports the following mentioned configurations for adapter pairing in High Availability mode. The adapter's write cache details are also provided below.

- #ESA3 is installed in pairs, and the pairing provides high availability configuration to protect against failure of one of the adapters in the pair.
- The adapter supports pairing of two #ESA3 adapters. Pairing with #5913 is not supported.
- The pairs provide mirroring for the write cache.
- Integrated flash memory provides protection of the write cache without batteries in case of power failure.
- An SAS (AA) Cable with HD is required for every #ESA3 pair to communicate status and cache content information.

Cabling requirements

The following are the cabling guidelines for the adapter:

- When attaching to disks in #5887 EXP24S drawer, SAS X, YO, or AT cable types are supported.
- An SAS (AA) with HD is required for adapter High Availability pairing.

For a detailed description of cabling in the system units, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling.htm

For a detailed description of cabling in the PCIe Gen3 Expansion I/O drawer, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling5887.htm

POWER8 processor-based models supported

The following POWER8 processor-based system models are supported by this adapter:

- Power S812L, Power 822L, Power S814, Power S822, Power824
- Power E870, Power E880

3.7.3 #EJ0J - PCIe3 RAID SAS adapter Quad-port 6Gb x8

This is a high performance Full PCIe Gen3 Quad-port x8 SAS RAID adapter. The Low profile capable adapter provides 6 Gb of SAS speed, and does not provide a write cache. Adapter pairing configuration is optional.

RAID levels and supported disk drives

The following are the RAID levels and supported disk drives for this adapter:

- Supports RAID 0, RAID 5, RAID 6, and RAID 10 for AIX/Linux/VIOS formatted HDD and SSD.
- Supports RAID 5 and RAID 6 for IBM i formatted HDD and SSD.
- SAS bays in EXP24S or PCIe 12X I/O Drawers are supported.

- The Quad-port adapter supports connection for up to four EXP24S (#5887) drawers:
  - Maximum of 48 SSDs can be attached per adapter or pair.
  - Maximum of 96 HDDs can be attached per adapter or pair.
- HDDs and SSDs cannot be mixed on the same port, but they can be mixed on the same adapter.

Adapter pairing and write cache

This adapter supports the following configurations for adapter pairing in High Availability mode. The adapter's write cache details are also provided below:

- #EJ0J can be optionally installed in pairs, and the pairing provides performance and high availability configuration to protect against failure of one of the adapters in the pair.
- The adapter supports pairing of two #EJ0J adapters only.
- The #EJ0J does not provide a write cache.
- Configurations for Applications with disk write performance sensitive workloads should consider #EJ0L, which provides write cache.

Cabling requirements

The following are the cabling guidelines for the adapter:

- When attaching to disks in #5887 EXP24S drawer, SAS X, YO, or AT cable types with Mini-SAS HD Narrow connectors such as #ECBJ-ECBL, #ECBT-ECBV, #ECCO-ECC4 are required.

For a detailed description of cabling in the system units, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling.htm

For a detailed description of cabling in the PCIe Gen3 Expansion I/O drawer, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling5887.htm

POWER8 processor-based models supported

The following POWER8 processor-based system models are supported by this adapter:

- Power S814, Power S822, Power824
- Power E850, Power E870, Power E880

3.7.4 #EJ0L - PCIe3 12GB Cache RAID SAS adapter Quad-port 6Gb x8

High performance Full PCIe Gen3 Quad-port x8 SAS RAID adapter, with 12 GB of cache. The adapter provides 6 Gb SAS speed.

RAID levels and supported disk drives

The following are the RAID levels and supported disk drives for this adapter:

- Supports RAID 0, RAID 5, RAID 6, and RAID 10 for AIX/Linux/VIOS formatted HDD and SSD.
- Supports RAID 5 and RAID 6 for IBM i formatted HDD and SSD.
- SAS bays in EXP24S or PCIe 12X I/O Drawer are supported.

- The Quad-port adapter supports connection for up to four EXP24S (#5887) drawers:
  - Maximum of 48 SSDs can be attached per pair.
  - Maximum of 96 HDDs can be attached per pair.

Adapter pairing and write cache

This adapter supports the following configurations for adapter pairing in High Availability mode. The adapter's write cache details are also provided below.

- #EJ0L is installed in pairs and the pairing provides performance and high availability configuration to protect against failure of one of the adapters in the pair.
- The adapter supports pairing of two #EJ0L adapters only.
- The #EJ0L provides a write cache of effectively up to 12 GB.
  - The adapter uses compression in the physical cache of 3 GB.
- Integrated flash memory provides protection of the write cache without batteries in case of power failure.
- Two SAS (AA) Cables with HD are required for every #EJ0L pair to communicate status and cache content information.

Cabling requirements

The following are the cabling guidelines for the adapter:

- When attaching to disks in #5887 EXP24S drawer, SAS X, YO, or AT cable types with Mini-SAS HD Narrow connectors such as #ECBJ-ECBL, #ECBT-ECBV, and #ECCO-ECC4 are required.

For a detailed description of cabling in the system units, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling.htm

For a detailed description of cabling in the PCIe Gen3 Expansion I/O drawer, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling5887.htm

POWER8 processor-based models supported

The following POWER8 processor-based system models are supported by this adapter:

- All POWER8 based systems are supported, but this adapter is not Low Profile capable.

3.7.5 #EJ0M - PCIe3 LP RAID SAS adapter

High performance LP PCIe Gen3 Quad-port x8 SAS RAID adapter. The adapter provides 6 Gb SAS speed. #EJ0M is the low profile version of #EJ0J.

RAID levels and supported disk drives

The following are the RAID levels and supported disk drives for this adapter:

- Supports RAID 0, RAID 5, RAID 6, and RAID 10 for AIX/Linux/VIOS formatted HDD and SSD.
- Supports RAID 5 and RAID 6 for IBM i formatted HDD and SSD.
- SAS bays in EXP24S or PCIe 12X I/O Drawer are supported.

- The Quad-port adapter supports connection for up to four EXP24S (#5887) drawers:
  - Maximum of 48 SSDs can be attached per adapter or pair.
  - Maximum of 96 HDDs can be attached per adapter or pair.
- HDDs and SSDs cannot be mixed on the same port, but can be mixed on the same adapter.
  - 177 GB SSDs are not supported.

Adapter pairing and write cache

This adapter supports the following configurations for adapter pairing in High Availability mode. The adapter's write cache details are also provided below.

- #EJ0M can be installed in pairs, and the pairing provides performance and high availability configuration to protect against failure of one of the adapters in the pair.
- The adapter supports pairing of two #EJ0M adapters, or one #EJ0M and one #EJ0J.
- The #EJ0M does not provide a write cache and hence pairing of adapters is optional.
- One SAS (AA) Cables with HD are required for every #EJ0M pair to communicate status and cache content information.

Cabling requirements

The following are the cabling guidelines for the adapter:

- When attaching to disks in #5887 EXP24S drawer, SAS X, YO, or AT cable types with Mini-SAS HD Narrow connectors such as #ECBJ-ECBL, #ECBT-ECBV, or #ECCO-ECC4 are required.

For a detailed description of cabling in the system units, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling.htm

For a detailed description of cabling in the PCIe Gen3 Expansion I/O drawer, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling5887.htm

POWER8 processor-based models supported

The following POWER8 processor-based system models are supported by this adapter:

- Power S822
- Power E870, Power E880

3.7.6 #EL3B - PCIe3 LP RAID SAS adapter

High performance LP PCIe Gen3 Quad-port x8 SAS RAID adapter. The adapter provides 6 Gb of SAS speed. #EL3B is the low profile version of #EJ0J.

RAID levels and supported disk drives

The following are the RAID levels and supported disk drives for this adapter:

- Supports RAID 0, RAID 5, RAID 6, and RAID 10 for Linux/VIOS formatted HDD and SSD
- #EL3B is supported only in the Linux only models and hence no support for AIX and IBM i operating systems.
- SAS bays in EXP24S or PCIe 12X I/O Drawer are supported.

- The Quad-port adapter supports connection for up to four EXP24S (#5887) drawers:
  - Maximum of 48 SSDs can be attached per adapter or pair.
  - Maximum of 96 HDDs can be attached per adapter or pair.
- HDDs and SSDs cannot be mixed on the same port, but can be mixed on the same adapter.
  - 177 GB SSDs are not supported.

Adapter pairing and write cache

This adapter supports the following configurations for adapter pairing in High Availability mode. The adapter's write cache details are also provided below.

- #EL3B can be installed in pairs and the pairing provides performance and high availability configuration to protect against failure of one of the adapters in the pair.
- The adapter supports pairing of two #EL3B adapters, or one #EL3B and one #EJ0J.
- The #EL3B does not provide a write cache and hence pairing of adapters is optional.
- One SAS (AA) Cable with HD is required for every #EJ0M pair to communicate status and cache content information.

Cabling requirements

The following are the cabling guidelines for the adapter:

- When attaching to disks in #5887 EXP24S drawer, SAS X, YO, or AT cable types with Mini-SAS HD Narrow connectors such as #ECBJ-ECBL, #ECBT-ECBV, or #ECCO-ECC4 are required.

For a detailed description of cabling in the system units, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling.htm

For a detailed description of cabling in the PCIe Gen3 Expansion I/O drawer, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling5887.htm

POWER8 processor-based models supported

The following POWER8 processor-based system models are supported by this adapter:

- Power S812L, Power S822L

3.7.7 #EL59 - PCIe3 LP RAID SAS adapter Quad-port 6Gb x8

High performance PCIe Gen3 Quad-port x8 SAS RAID adapter, and Low Profile Capable. The adapter provides 6 Gb of SAS speed.

RAID levels and supported disk drives

The following are the RAID levels and supported disk drives for this adapter:

- Supports RAID 0, RAID 5, RAID 6, and RAID 10 for Linux/VIOS formatted HDD and SSD.
- #EL59 is supported only in the Linux only models, and provides no support for the AIX and IBM i operating systems.
- SAS bays in EXP24S or PCIe 12X I/O Drawer are supported.

- The Quad-port adapter supports connection for up to four EXP24S (#5887) drawers:
  - Maximum of 48 SSDs can be attached per adapter or pair.
  - Maximum of 96 HDDs can be attached per adapter or pair.
- HDDs and SSDs cannot be mixed on the same port, but can be mixed on the same adapter.

Adapter pairing and write cache

This adapter supports the following configurations for adapter pairing in High Availability mode. The adapter's write cache details are also provided below.

- #EL59 can be optionally installed in pairs, and the pairing provides performance and high availability configuration to protect against failure of one of the adapters in the pair.
- The adapter supports pairing of two #EL59 adapters, or one #EL59 and one #EJ0J, depending on the drawers used.
- The #EL59 does not provide a write cache, and hence pairing is not mandatory.
- Configurations for Applications with disk write performance sensitive workloads should consider #EJ0L, which provides write cache.

Cabling requirements

The following are the cabling guidelines for the adapter:

- When attaching to disks in #5887 EXP24S drawer, SAS X, YO, or AT cable types with Mini-SAS HD Narrow connectors such as #ECBJ-ECBL, #ECBT-ECBV, and #ECCO-ECC4 are required.

For a detailed description of cabling in the system units, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling.htm

For a detailed description of cabling in the PCIe Gen3 Expansion I/O drawer, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling5887.htm

POWER8 processor-based models supported

The following POWER8 processor-based system models are supported by this adapter:

- All POWER8 based systems are supported, but this adapter is not Low Profile capable

3.7.8 #5913 - PCIe2 1.8GB Cache RAID SAS adapter Tri-port 6Gb x8

High performance PCIe2 x8 SAS RAID adapter with 1.8 GB write adapter, provides 6 Gb of SAS speed.

RAID levels and supported disk drives

The following are the RAID levels and supported disk drives for this adapter:

- Supports RAID 0, RAID 5, RAID 6, and RAID 10 for AIX/Linux formatted HDD and SSD.
- Supports RAID 5 and RAID 6 for IBM i formatted HDD and SSD.
- SAS bays in EXP24S drawers, PCIe 12X I/O drawers, and EXP 12 disk drawers are supported.

- The #5913 adapter is always installed in pairs to provide mirrored write cache data and adapter redundancy.
- Two adapters in one pair of #5805 can be in two different systems in High Availability configurations for AIX and Linux.
  - IBM i does not support adapter pairs that are in two different systems.
- Supports connection to disk bays in up to three EXP24S (#5887) drawers or six EXP12S drawers.
- This adapter supports connections to dual port SAS or SSD for path redundancy. It performs automatic path switching if one of the paths to the disks fail.

Adapter pairing and write cache

This adapter supports the following configurations for adapter pairing in High Availability mode. The adapter's write cache details are also provided below.

- #5913 is always installed in pairs, and the pairing provides high availability configuration to protect against failure of one of the adapters in the pair.
- The adapter supports pairing of two #5913 adapters.
- The pairs provide mirroring for the write cache.
- Cache mirroring is disabled if the adapter pair is broken or one of the adapters is faulty.
- Write cache provides performance improvement even when configured without RAID 5/6/10.

Cabling requirements

The following are the cabling guidelines for the adapter:

- X, YO, or AT SAS cables with HD connectors are used to attach the disk bays in the supported expansion drawers.
- An AA SAS cable with HD connectors is attached to the pair to communicate status and cache content information.

For a detailed description of cabling in the system units, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling.htm

For a detailed description of cabling in the PCIe Gen3 Expansion I/O drawer, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling5887.htm

POWER8 processor-based models supported

The following POWER8 processor-based system models are supported by this adapter:

- Power S812L, Power 822L, Power S814, Power S822, Power824
- Power E870, Power E880

3.7.9 #5901 - PCIe Dual-x4 SAS adapter

The PCIe Dual-x4 SAS adapter with low-profile short form factor provides high performance connection to SAS devices with SAS speed of 3 Gbs.

RAID levels and supported disk drives

The following are the RAID levels and supported disk drives for this adapter:

- The adapter supports RAID 0 and RAID 10 for AIX and Linux formatted SAS drives.
- RAID 5 or RAID 6 are not supported for IBM i formatted disks. The adapter can be used for data spreading and mirroring functions that are supported by IBM i.
- This adapter supports connection for disks in the EXP24S expansion drawer.
- Supports up to 48 SAS disks when connected to four EXP24S drawers.
- Provides eight physical links through two mini SAS 4x connectors.

Adapter pairing and write cache

This adapter supports the following configurations for adapter pairing in High Availability mode. The adapter's write cache details are also provided below:

- The adapter supports pairing of two #5901 adapters for high availability configurations to provide adapter redundancy

The adapter does not have the write cache feature.

Cabling requirements

The following are the cabling guidelines for the adapter:

- SAS Y cables attach SAS disk drives in the EXP24S drawers.

For a detailed description of cabling in the system units, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling.htm

For a detailed description of cabling in the PCIe Gen3 Expansion I/O drawer, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling5887.htm

POWER8 processor-based models supported

The following POWER8 processor-based system models are supported by this adapter:

- Power S814, S824, S822, E850, E870, E880

3.7.10 #5278 - PCIe LP Dual-x4 SAS adapter 3Gb

PCIe Low-profile dual-x4 SAS adapter with low-profile short form factor, provides high performance connection to SAS devices with an SAS speed of 3 Gbs.

RAID levels and supported disk drives

The following are the RAID levels and supported disk drives for this adapter:

- The adapter supports RAID 0 and RAID 10 for AIX and Linux formatted SAS drives.
- RAID 5 or RAID 6 are not supported for IBM i formatted disks. The adapter can be used for data spreading and mirroring functions that are supported by IBM i.
- This adapter supports connection for disks in the EXP24S expansion drawer.
- Supports up to 48 SAS disks when connected to four EXP24S drawers.
- Provides eight physical links through two mini SAS 4x connectors.

Adapter pairing and write cache

This adapter supports the following configurations for adapter pairing in High Availability mode. The adapter's write cache details are also provided below.

- The adapter supports pairing of two #5901 adapters for high availability configurations to provide adapter redundancy.

The adapter does not have the write cache feature.

Cabling requirements

The following are the cabling guidelines for the adapter:

- SAS Y cables attach SAS disk drives in the EXP24S drawers.

For a detailed description of cabling in the system units, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling.htm

For a detailed description of cabling in the PCIe Gen3 Expansion I/O drawer, see the IBM Knowledge Center at:

http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling5887.htm

POWER8 processor-based models supported

The following POWER8 processor-based system models are supported by this adapter:

- Power S822

3.8 IBM disk formatting practices

HDDs and SSDs are available in either 5xx byte/block or 4K byte/block formats.

On IBM Power Systems, a 5xx byte/block device can be formatted as 512 bytes/block for operation as a JBOD device, or as 528 bytes/block for use in a RAID array. Likewise, a 4K byte/block device can be formatted as 4096 bytes/block for operation as a JBOD device, or as 4224 (8*528) bytes/block for use in a RAID array.

Traditionally, the storage industry used 5xx byte/block disk formatting. As the typical disk sizes and applications larger block sizes increased greatly, 5xx byte/block disk formatting became less efficient, affecting I/O performance. It is also a challenge to perform error handling in large disks that are formatted with the 5xx byte/block size. To address the problems with modern application requirements and efficient error handling in the disks, the industry has collectively agreed to moving towards adoption of 4K blocks disk formatting. All PCIe3 RAID adapters in POWER8 processor-based systems support disks with 4K block formatting. Most of the disks provide an option to order disks shipped formatted with a 4K blocksize. To help in the transition from 5xx byte/block to 4K block disk formatting, the option to order 5xx disk formatting is also available.

SSDs are only supported in RAID format, that is, 528 or 4224 bytes/block. Additionally, all HDDs and SSDs used under the IBM i are only supported in RAID format because the 8-header on each block of data is needed for virtual memory management by the IBM i storage management.

HDDs shipped from the factory can be formatted in RAID format, and thus AIX/Linux systems that want to use the JBOD configuration need to reformat the HDDs to JBOD format.

3.8.1 Guidelines for choosing disks

Each disk feature code specifies the disk is intended for SFF Gen2 bay or SFF Gen3 bays. The system units in all POWER8 based systems contain SFF-3 bays, and the EXP24S Disk expansion drawers contain SFF-2 bays. Choose between the SFF-2 or SFF-3 disk feature codes based on the planned location for the disks (system unit bays or expansion drawer bays).

3.8.2 ANSI T10 standardized data integrity fields

When HDDs/SSDs are formatted in RAID format, that is, 528 or 4224 bytes/block, each block of data is protected by industry standardized data integrity fields (T10 DIF).

The architecture in Figure 3-5 shows the disk formatting and how different operating systems in Power Systems use each block. The top section of the picture displays what AIX/Linux and IBM i Operating systems use in the disk blocks, and the bottom part of the picture shows the actual formatting used by the SAS adapter. This unique formatting strategy provides an efficient way for the SAS RAID adapters to provide extra protection, and the operating systems use the required space to store data, and header in case of IBM i.

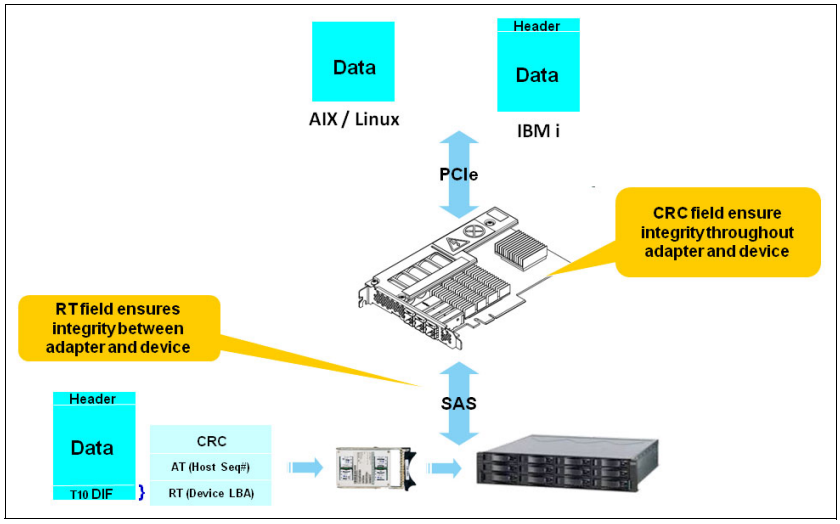


Figure 3-5 Disk formatting practices

3.9 VIOS vSCSI disks and IBM i client partitions

This section in this chapter provides a description on the VIOS vSCSI disks for IBM i client partitions.

Figure 3-6 shows a logical diagram of VIOS vSCSI device mappings to client partitions that run the IBM i operating system. VIOS vSCSI adapters can be used to map to backing devices that can be whole disks (hdisks) in VIOS or logical volumes created in the volume groups can be shared as backing devices to the client partitions.

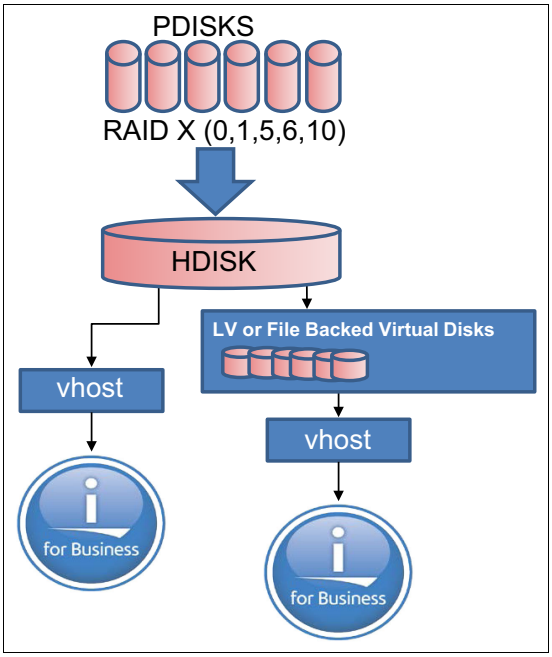


Figure 3-6 VIOS vSCSI for IBM i client

The whole disk (hdisk) backing devices can be in JBOD or RAID configurations from the physical hardware mapped to the VIOS. There are special notes related to performance and disk formatting for IBM i client partitions to be configured correctly to avoid potential performance problems. These special cares are to address the way IBM i requires the disk formatting to be done for better disk performance.

vSCSI backing devices with 5xx sector formatting

When VIOS backing devices to IBM i client partitions have 512-byte sector formatting, this can cause potential I/O performance degradation in the client partition. This is because IBM i uses 520-byte sector formatting, which is an 8-byte header to store virtual address mapping for the data, and 512 byte for the data in the sector. When IBM i is configured with backing devices formatted with 512-byte sector, there is a 9th sector created for the 8-byte headers for eight sectors. This might create potential performance issues at the client partitions. The PCIe2/PCIe3 adapters have internal hardware optimizations that work on 8 sector I/O boundaries.

To overcome this potential performance problem, VIOS can provide virtual disks in 520-byte sector format when backed by SAS adapters that are capable of supporting that format. The disks that are the backing devices for the client IBM i partitions must be configured using a RAID array. RAID 0 can be used for this purpose, which allows single disk (pdisk) in each RAID array (hdisk). Configuring the devices to RAID 0 provides more protection to the RAID array using enhanced error handling methods available in the SAS RAID controllers with no performance penalty.

This optimization is made to IBM i 7.1 TR 7 and later, and is supported when used with physical devices as backing devices for vSCSI and not logical volume backing devices in the volume groups in the VIOS LPAR. Figure 3-7 shows the 5xx sector format.

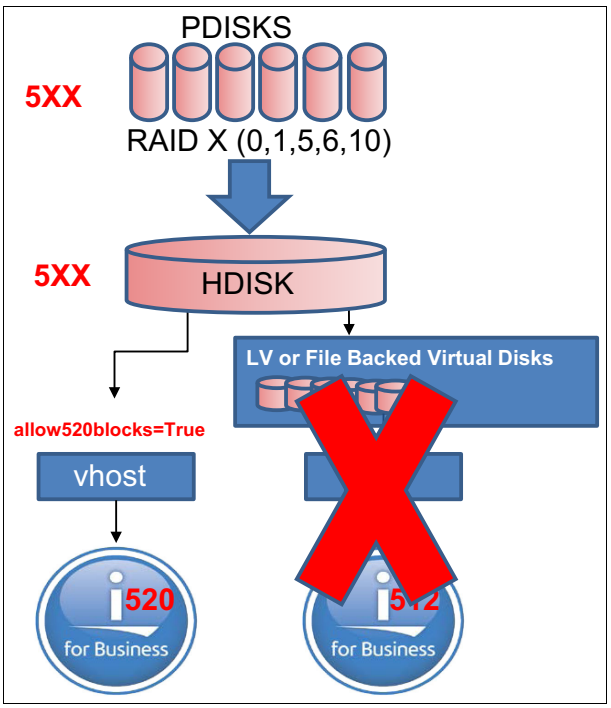


Figure 3-7 5xx sector format and IBM i clients

vSCSI backing devices with 4K sector formatting

Figure 3-8 shows one of the preferred configurations for providing best performance to IBM i client partitions. With the enhancements to the latest version of IBM i, it still prefers to use multiple LUNs/ disks for optimal performance.

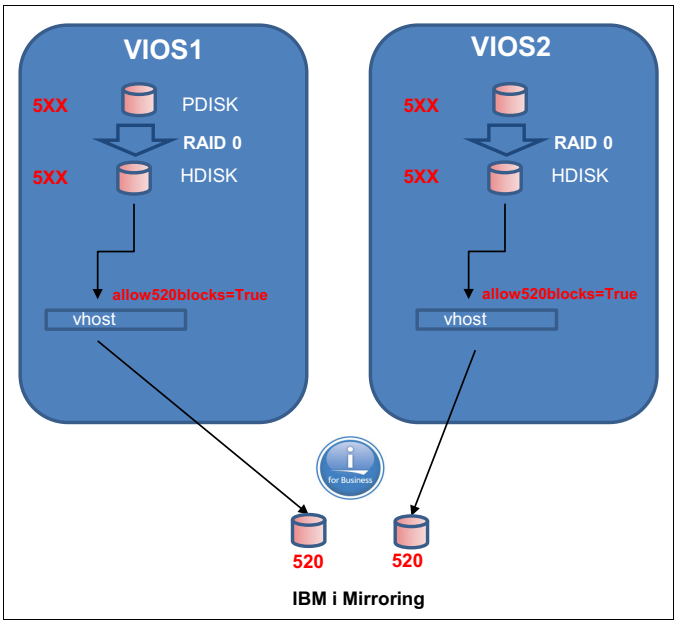


Figure 3-8 IBM i clients, VIOs, and RAID 0 with software mirroring

The best way to achieve this configuration in the VIOS environments is using the SAS RAID adapters with RAID 0 configured backing devices that are mapped from each VIOS. IBM i OS in the client partition now sees two disks, one from each VIOS that will be configured with the software mirroring. This configuration provides the best performance and better protection using RAID 0 in the SAS RAID adapter, and will also use any write cache that is available in the adapter, based on the adapter model used in the configuration.

With POWER8 processor-based systems, all SAS RAID adapters and disk support the use of 4K sector formatting. The 4K sector formatted RAID 0 disks can be used as described in the “vSCSI backing devices with 5xx sector formatting” on page 72 to achieve good performance for IBM i client partitions.

Starting with IBM i version 7.2, enhancements have been added to IBM i operating system to use logical volumes in the VIOS as the backing devices that are formatted with 4K sectors and are connected with one of the supported RAID levels in the SAS RAID adapters (Figure 3-9).

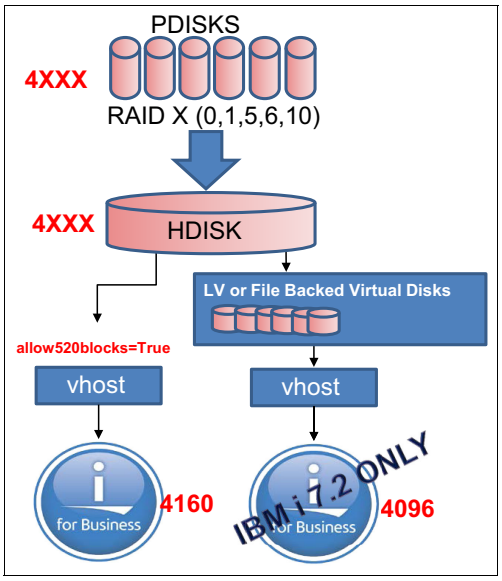


Figure 3-9 IBM i clients, VIOS, and 4K sector format

3.10 RAID adapters performance characteristics

This section in the chapter provides descriptions of the performance configurations on the hardware level and operating systems that are supported by IBM Power systems.

This chapter also discusses various other related topics, including these topics:

- JBOD and RAID 0
- SSDs and Easy Tier Array
- Performance test results

3.10.1 JBOD, RAID 0 and write cache

JBOD is an architecture that involves multiple disk drives, and makes them accessible either as independent hard disk drives or as a spanned single volume without actual RAID functionality. When using JBOD configuration with SAS RAID adapters with write cache, the

write cache in the adapter is disabled. The write cache can be used only when the adapter is configured with one of the supported RAID levels.

SAS RAID adapters provide extra hardware enhancements for error handling and performance with write cache, if write cache is available in the adapters. The additional enhancements in the hardware are also available only when the adapter is configured with one of the supported RAID levels.

RAID 0 with SAS RAID adapters is a better solution that provides more hardware adapter assisted protection for configurations and write cache than using disks in JBOD configurations.

Table 3-5 shows a high-level comparison of the JBOD and RAID 0 configurations.

Table 3-5 Comparison of JBOD versus RAID 0

Function	JBOD	RAID 0
Fault Tolerance	No	No. Additional Error Handling provided by RAID adapters
Write cache	No	Yes (if adapter supports write cache)
Parity Disk	No	No
T10DIF	No	Yes

3.11 SSDs and Easy Tier array

The Easy Tier function automatically places data into optimal tiers of storage, while maintaining or improving performance has been added to some IBM PCIe3 SAS RAID adapters for Power Systems.

Introduced in June 2014, Easy Tier function is combined with the performance advantages of direct-attached storage (DAS) for IBM Power Systems. Easy Tier, at a high level, allows different performance tiers of storage devices to be combined to improve cost and performance of the storage subsystem. Specifically, HDDs and SSDs are combined to provide SSD-like performance while effectively using the higher capacity and lower cost of HDDs. While similar in concept to Easy Tier in IBM storage products such as the DS8000®, V7000, and SAN Volume Controller, this Easy Tier function for Power Systems is managed within the SAS RAID adapter.

The SAS RAID adapter automatically detects hot data and moves it to the SSDs whereas cold data is moved to the HDDs. The hottest data is moved first. The RAID storage subsystem tunes itself based on the workload characteristics of the system to optimize performance. The Easy Tier function continually and dynamically reacts to workload changes in real time, often moving hot data to the SSDs in seconds.

After a tiered RAID array is created, the adapter collects statistics as to how many reads or writes are occurring to each band of data in the array. A *band* is typically 1 MB to 2 MB (automatically selected by the RAID adapter) and represents the amount of data that is swapped between an SSD and HDD when appropriate hot and cold data is identified. As the workload changes, swaps of data between the SSDs and HDDs occur to improve performance.

For more information, see the following Performance Study of IBM Power Systems 'SAS RAID Adapter Easy Tier Function at:

http://w3-01.ibm.com/sales/ssi/cgi-bin/ssialias?subtype=WH&infotype=SA&appname=STGI_PO_PO_USEN&htmlfid=POW03129USEN&attachment=POW03129USEN.PDF

3.11.1 Performance testing with Easy Tier

IBM conducted performance testing in the labs on the SAS RAID adapter that is enabled with Easy Tier feature. The workload used was a Stock Trading application using IBM DB2® that is transaction-based, and configured on an IBM Power System.

The two test result graphs are included here are to display performance benefits to the workload when using Easy Tier adapter with two RAID arrays consisting of SSDs and HDDs, and a comparison to the same workload run using the same SAS RAID adapter with HDDs only, without using Easy Tier configurations.

The graph in Figure 3-10 shows the application transactions per second improvements with Easy Tier configuration over same application being run on HDDs only. The average transactions throughput went from an average of 30 per second on the HDDs only arrays, to over 170 transactions per second when using the Easy Tier arrays. The peak workload differences show the Easy Tier configuration is approximately 8-10 times as efficient as HDD only configurations.

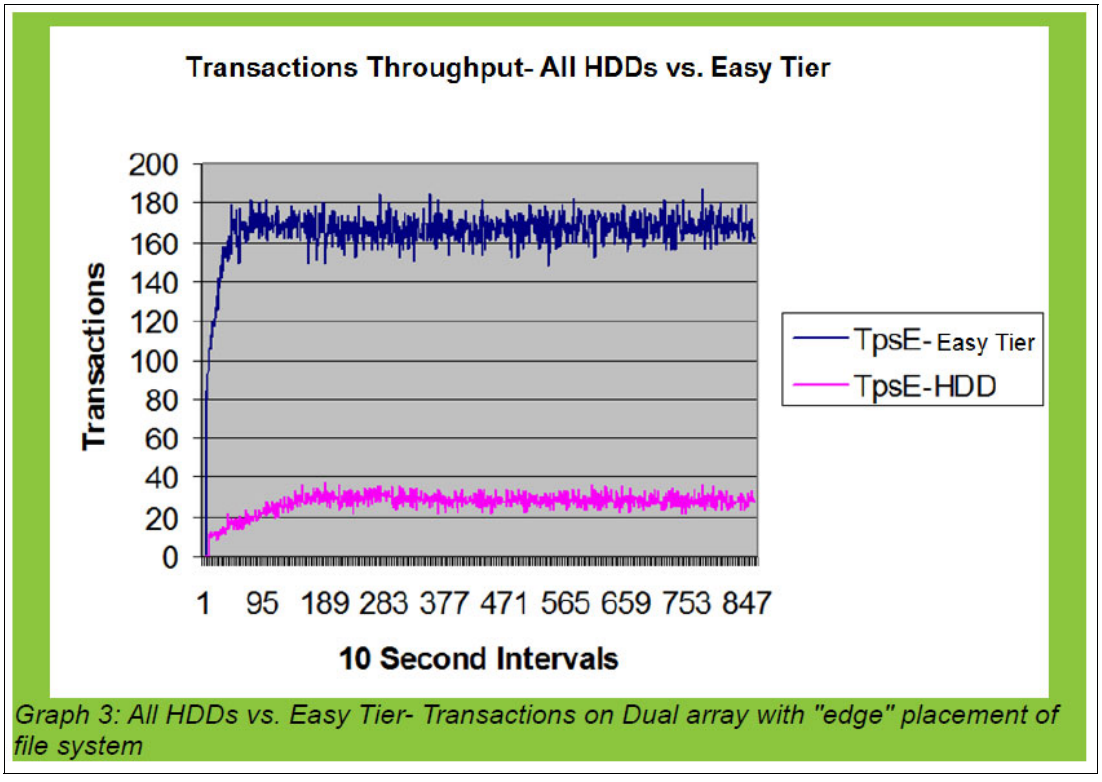


Figure 3-10 Transactions throughput with Easy Tier

The graph in Figure 3-11 shows the response times improvements with Easy Tier configuration over same application being run on HDDs only. The average response times for a complete transaction went down from about 124 milliseconds on the HDDs to about 18 milliseconds on the Easy Tier arrays.

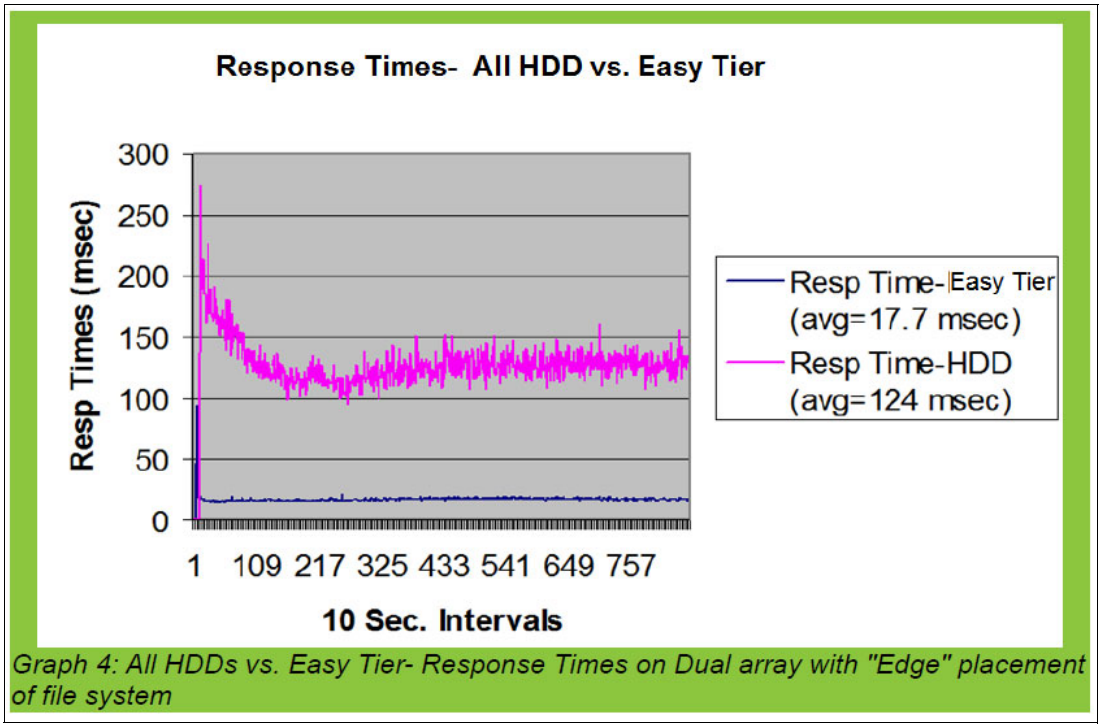


Figure 3-11 I/O Response time with Easy Tier

The tests included many more scenarios to show the performance and price benefits of using Easy Tier configurations compared to HDDs only RAID arrays performance. For all test scenarios with results, and details about the systems configuration and tuning information, download the full test paper at:

http://w3-01.ibm.com/sales/ssi/cgi-bin/ssialias?subtype=WH&infotype=SA&appname=STGI_PO_PO_USEN&htmlfid=POW03129USEN&attachment=POW03129USEN.PDF

3.12 SAS RAID adapters performance comparison

Another round of performance testing conducted by IBM labs provided in-depth analysis of comparisons between various RAID levels performance characteristics, and a comparison of PCIe Gen2 adapter against PCI Gen3 RAID adapter performance. These include the PCIe3 RAID SAS Adapter Quad-port 6 Gb x8, Custom Card Identification Number (CCIN) 57B4, and PCIe3 12GB Cache RAID SAS Adapter Quad-port 6 Gb x8, CCIN #57CE.

Three different workloads were run that mixed reads and writes, various transfer lengths, spatial localities, and request rates. These workloads simulate the I/O done by various OLTP applications:

- OLTP1: Read/write ratio is 60/40 with 4 KB I/O size. I/Os are random over the full capacity of devices, so there is little opportunity for IOA cache hits.
- OLTP2: Read/write ratio is 90/10 with 8 KB I/O size. I/Os are random over the full capacity of each device, so there are almost no IOA cache hits.

- OLTP3: Read/write ratio is 70/30 with 4 KB I/O size. Half of the reads are random and half are intended to elicit cache hits. 33 percent of writes are random, 33 percent are to logical block addresses recently read, and 34 percent are intended to elicit write cache hits.

OLTP 1, 2, 3 test results

The graphs in Figure 3-12, Figure 3-13 on page 79, and Figure 3-14 on page 79 show the test results of three OLTP 1, 2, 3 workloads being run on each adapter configured with RAID 5. The maximum IOPS that we can get with the newer generation adapters is almost twice that of the older generation adapters. The IOPS with the new cached adapters 57CE is also twice as that of the older generation cached adapters. Notice that the 57CE's response time stays much lower than the 57B5's through much higher throughputs. This is due to the larger effective cache size. Also, note that both caching adapters have lower response times than both non-caching adapters.

To download the full paper to view all performance test results, with a full description of the system and adapter configurations, and tuning performed, click the following link:

http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=WH&infotype=SA&appname=STGE_PO_PO_USEN&htmlfid=POW03122USEN&attachment=POW03122USEN.PDF

Figure 3-12 shows the OLTP 1 comparison.

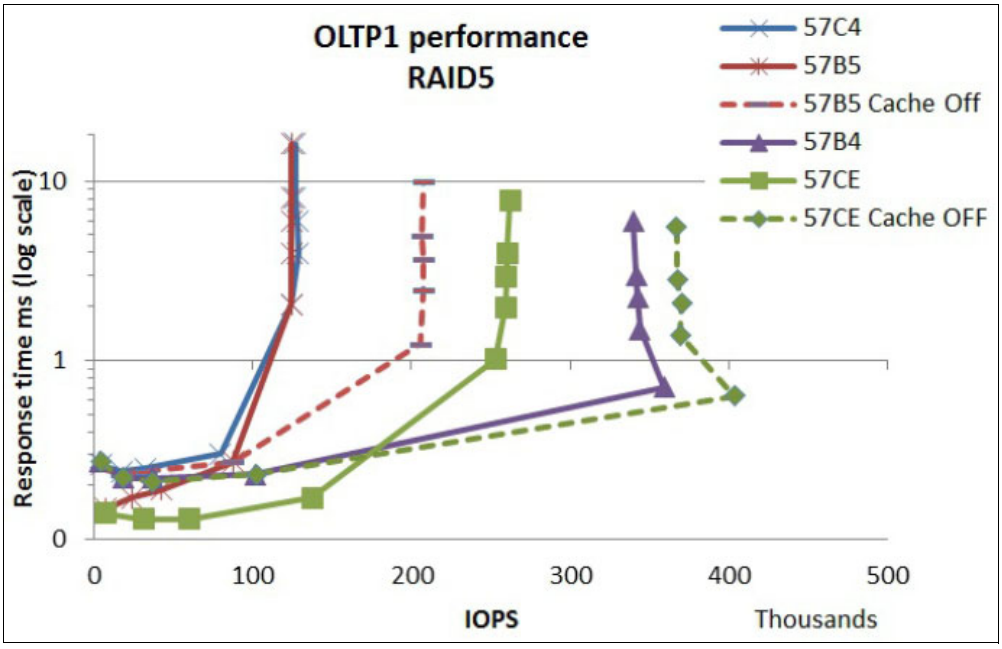


Figure 3-12 OLTP 1 generation comparison

Figure 3-13 shows the OLTP 2 comparison.

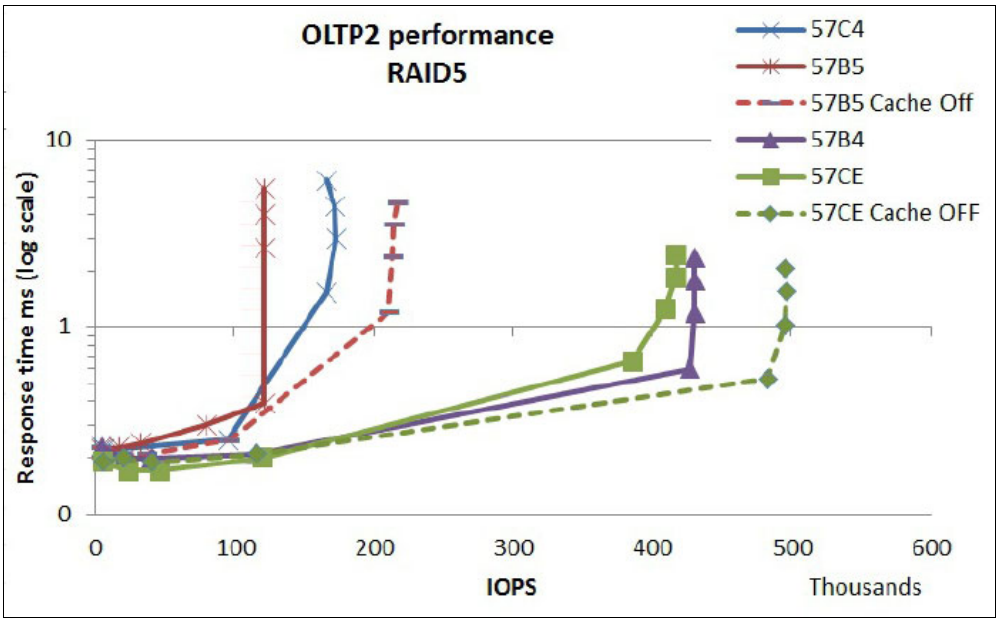


Figure 3-13 OLTP 2 generation comparison

Figure 3-14 shows the OLTP 3 comparison.

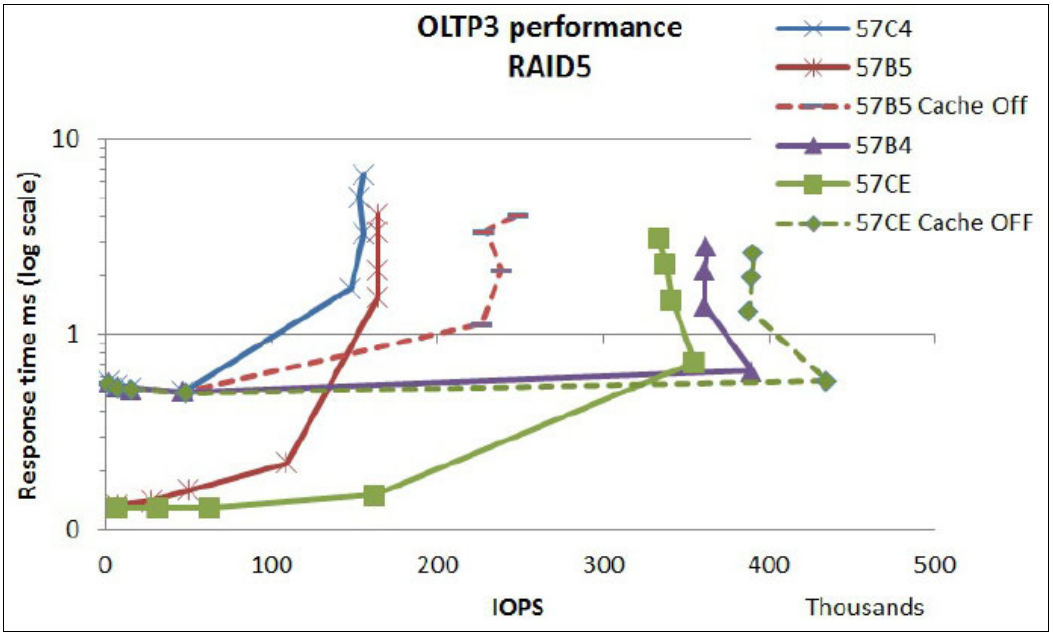


Figure 3-14 OLTP 3 generation comparison

4

Software RAID in Power Systems

This chapter provides descriptions of the Software RAID configurations in the operating systems that supported by IBM Power systems.

- This chapter includes the following sections:
- Software RAID in AIX
- Software RAID in Linux
- Software RAID in IBM i

4.1 Software RAID in AIX

Software RAID functions in AIX are provided by the AIX logical volume manager (LVM). AIX LVM supports striping (RAID 0), mirroring (RAID 1), and the mirror and stripe function (RAID 0+1) for software-based RAID functions without having to use hardware SAS RAID adapters. For configuring software-based RAID in AIX, there are no specific application or drivers to be installed, as the functions are provided with the native LVM in AIX.

Configuring RAID in AIX

AIX provides command-line utilities to configure the software RAID and assist with configuring Hardware RAID through the AIX operating system commands.

AIX LVM supports striping (RAID 0), mirroring (RAID 1), and the mirror and stripe function (RAID 0+1) for software RAID. These can be configured using LVM commands or the System Management Interface Tool (SMIT).

Striping (RAID 0)

AIX uses Volume Group (VG) to extend the space from one disk to another. This configuration requires a minimum of two disks. RAID 0 doesn't provide any data redundancy, so if one disk goes bad, all data access is lost. RAID 0 (Striping) provides excellent performance when reading and writing the data.

Striping (RAID 0) is typically used to provide performance enhancements to certain type of workloads that have many random I/O operations with very high number of I/O operations per second (IOPS). A commonly used practice for such workloads is to have multiple disks in the logical volume that has the application file system.

The data is striped across multiple disks in the logical volume, which allows multiple disks to be accessed in parallel. The queue_depth attribute in the device driver defines the number of IO operations that can be performed by each disk (hdisk) in the logical volume. Striping the data across multiple disks primarily increases the number of disks used to store and access data, thus providing a higher total queue depth for the logical volume.

Figure 4-1 shows the logical structure of AIX LVM striping function.

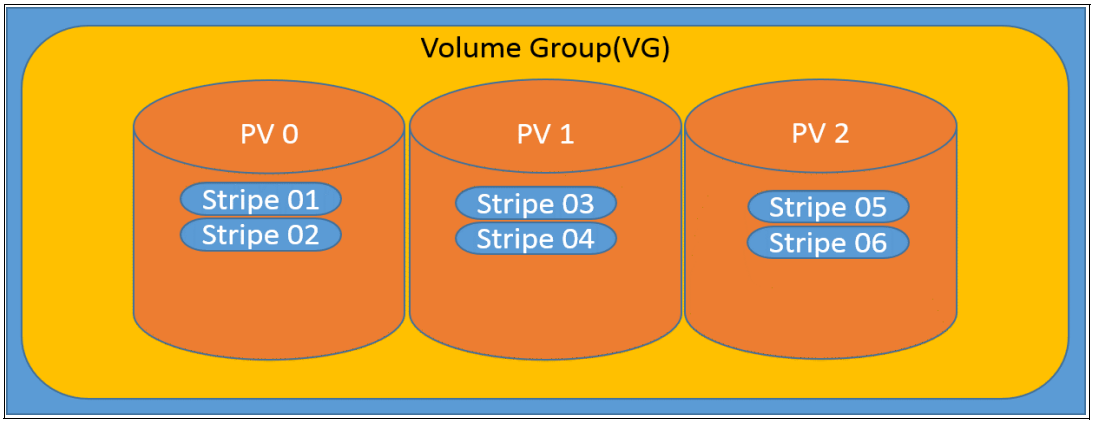


Figure 4-1 AIX LVM Striping

Striping provides no redundancy against disk failures because there is no parity protection or mirroring protection offered by this RAID level. It is common in AIX environments to configure AIX LVM striping using SAN-based LUNs in the logical volume. The storage subsystems in the SAN are often configured with one of the RAID levels that provide protection for each SAN-based LUN. Using SAN-based LUNs that are protected at the storage subsystem level and adding AIX LVM striping over the LUNs allows for better performance and required protection at the storage subsystem level.

Mirroring (RAID 1)

AIX LVM provides the mirroring feature to configure software-based RAID 1 functions using two or three disks for each logical volume. Mirroring for the logical volumes can be configured with a minimum of two or maximum of three copies of the logical volume in separate disks. Mirroring offers protection against one or two disk failures, depending on the maximum number of disks that are used in the mirror.

With the LVM mirroring, the mirroring is configured for each logical volume in a volume group or all logical volumes in the volume group. This mirroring is not based on the disks (hdisk), but rather on the logical volumes, so hdisks are not mirrored. Figure 4-2 shows the logical diagram of AIX LVM mirroring function.

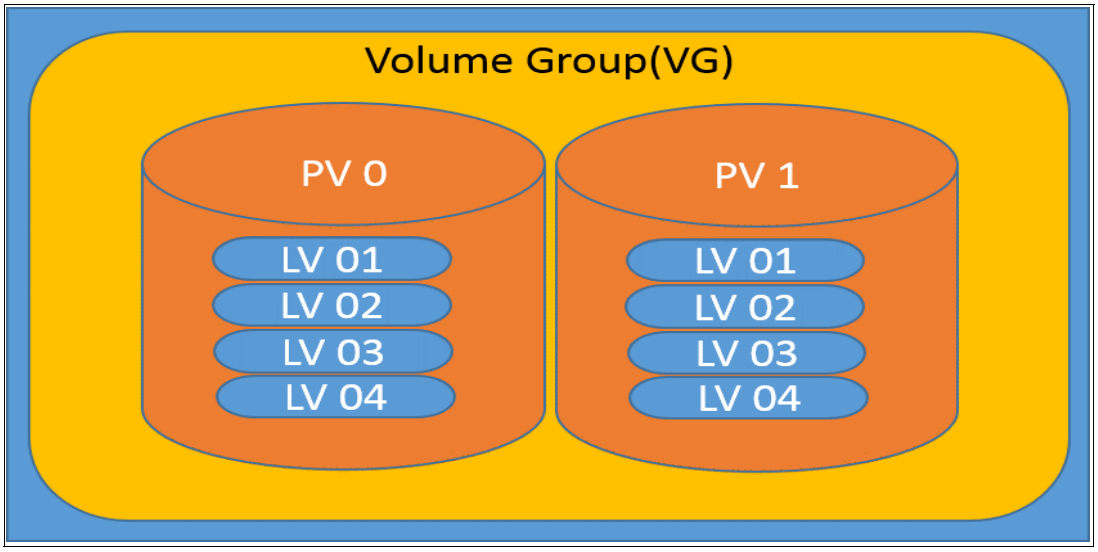


Figure 4-2 AIX LVM mirroring

Strict allocation policy

AIX LVM always attempts to place each copy of the mirrored logical partitions in separate disks. This is achieved by enabling the strict allocation policy for each copy in the mirror to be placed in a separate disk. This is the default policy for all mirrored logical volumes. This policy ensures that no two copies of the same data are located in the same disk, which does not provide the redundancy. If the disk containing the mirrored copies fail, then there is no protection to the stored data. However, users can opt to turn off this policy to have LVM place two copies of the same logical partition in the same disk if there is no free space in one of the disks in the mirror.

Apart from the allocation policy, there are more attributes the LVM provides that can be tuned for individual system or application requirements, which include:

- *Scheduling policy* that allows the user to choose between a sequential scheduling policy that writes data in sequence among the two or three copies in the mirror, or parallel scheduling policy that write data to all copies in parallel.

- *Mirror write consistency* provides a mechanism to keep all copies in a consistent state after a system recovers from a crash or the volume group was closed while IO operations to the logical volumes were in progress.

Figure 4-3 shows mirrored logical volumes and how many copies are in each logical volume.

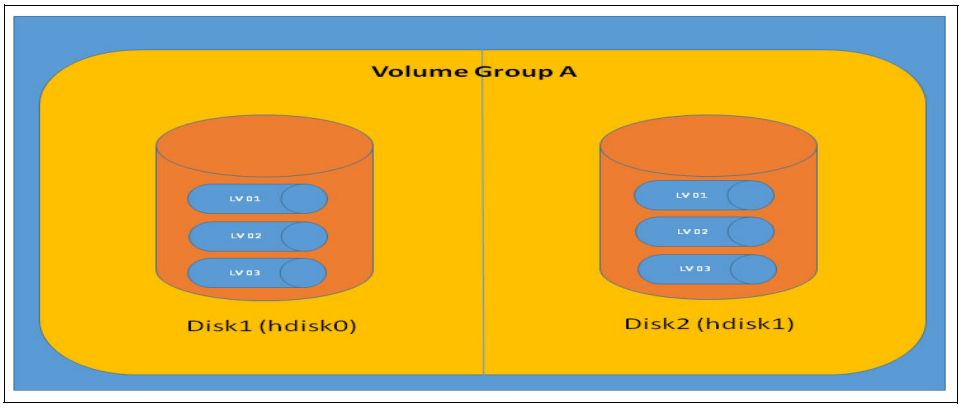


Figure 4-3 Mirrored Logical Volumes

The number of copies in the logical volumes can be identified by observing the number of LPs (logical partitions) and number of PPs (physical partitions).

Note on mirroring AIX rootvg: It is always advised to maintain the boot records on both the physical volumes in the mirror. Boot records can be created in all disks in the mirror by using the **bosboot** command.

The mirror and stripe function

The mirror and stripe function (also known as RAID 0+1) of AIX LVM provides better data availability, but at a higher cost.

In the case of non-mirrored and striped logical volumes, failure of one physical volume causes the loss of the whole striped logical volume. There is no redundancy in the striped logical volume configuration. The mirroring and striping function prevents this situation by creating up to three mirrored copies for each striped copy in the logical volume. If one of the disks holding the copy of the striped logical volume fails, it does not cause loss of access to the logical volume. The remaining copies in the mirrored configuration are still available and provide continuous access to the data in the logical volume.

This function is achieved by setting the partition allocation policy to super strict policy. This policy prohibits partitions from one mirror from sharing a disk with a second or third mirror.

4.2 Software RAID in Linux

This section covers software RAID considerations in Linux.

Linux software RAID

Software RAID is implemented in Linux Operating System using the Multi-Disk (MD) driver that is added to the Linux kernel. The MD driver can be compiled into the kernel, and many Linux distributions do that. If not compiled in the kernel, the MD drivel module can be loaded into the kernel as a module.

Configuring RAID

Linux supports different levels of RAID that can be configured by using the installation wizard at the initial operating system installation time, the **mdadm** command-line utility after installation, and kickstart configuration files for automated installations. The installation wizards in most of the commercial Linux distributions provide an option to create a custom disk layout, specifying which the RAID arrays can be created and the root file system installed in the RAID array. Additional RAID arrays can be configured by using the **mdadm** command-line utility after the installation. The Linux kickstart method for unattended installations can also be used to create RAID arrays by defining the required parameters in the kickstart configuration files to define the disk layouts.

RAID Levels supported

Linux supports the following RAID levels:

- RAID 0
- RAID 1
- RAID 4
- RAID 5
- RAID 6
- RAID 10
- Linear RAID

RAID 0, 1, 4, 5, 6, 10 levels in Linux have the same characteristics as they do with other operating systems or hardware RAID adapters, including a capability to assign spare disks in the RAID arrays.

Linear RAID needs a special mention here. Linear RAID is a way to group smaller disks into a bigger virtual disk. This level provides no protection against disk failures because one disk failure in the group makes the group unusable for I/O operations, and results in loss of data. This weaker protection does not provide better performance. The disks in the group are filled with data in a linear fashion. Therefore, it does not provide any parallel access to the disks to provide any performance benefits to use all disks in parallel.

The most commonly used RAID levels in Linux environments are RAID 5 and 6, in that order. The devices that are used in the RAID array can be physical partitions in the disk that are created using any disk formatting software tool supported in Linux.

RAID devices for root file system

As with many other operating systems, Linux also supports boot devices and root (/) file system to be configured using a RAID array of disks in one of the supported levels. For the system to boot from a RAID disk array, the MD driver module should be available in the kernel to be loaded during the boot process. If the RAID module is not available in the kernel to be loaded, then the system cannot be loaded. Use the root file system before mounting it to load the kernel module.

4.3 Software RAID in IBM i

Storage management in IBM i uses a different approach in comparison to other operating systems discussed in this book. IBM i treats storage as one layer to store and process data. Users do not configure file systems or directories to define the location as it is typically done in other operating systems.

Note: The terms used in this book to describe IBM i concepts are to help readers with different technical backgrounds to understand concepts of IBM i systems. The terms might not follow how they are used in the other IBM i system documentation, such as IBM Knowledge Center for IBM i systems.

IBM i views the storage disk space and memory as a single-level storage space to store and process data. The system decided the location of the storage for any file that the user saves in the system.

A disk unit in IBM i refers to a single or independent unit of storage disk. The disk unit can be locally attached to the system (DASD) or can be mapped to the system from a networked storage system such as SAN-based LUN. IBM i always does striping across multiple disk units, which is similar to the striping feature in other operating systems. This helps to optimize I/O performance and improves even more when multiple disk units are used.

Mirroring protection

The IBM i system supports Mirrored protection that is a software availability function to protect data against disk failures. The data to be written to the storage is mirrored between two different disk units. If one of the disk units in the mirror fails, the system keeps running the application without loss of access to the data by accessing the second disk unit in the mirror. This feature can be used on any model of IBM i systems and is a part of the Licensed Internal Code.

The two disk units can be in the same site, or one of them in the mirror can be at a remote site, thus provide remote mirroring capability. Depending on the specific system configurations, users can configure local mirroring or remote mirroring.

Mirroring with hot spare protection

If one of the disks in the mirrored pair is faulty, the system suspends mirrored protection to the failed disk unit. Applications continue to operate and the system uses the second disk in the mirror for all I/O operations to the data. IBM i mirror protection also allows a hot spare disk to be added to the mirrored pair. If a hot spare disk is available to the mirrored pair, the system replaces the failed disk with the hot spare disk after 5 minutes from the disk is detected to be failed. The 5 minutes wait time allows enough time to ensure that the disk failure is permanent to avoid replacements with hot spare disks for temporary glitches.

The hot spare disk is an optional feature for the mirrored protection functionality. If no hot spare disk is available, user can perform a manual replacement for the failed disk. The system starts to synchronize the data to the replaced disk in the mirrored pair by copying current data from the working disk unit to the replaced disk unit in the mirror. The synchronization process does not require a downtime to the system while it is being run in the system, but there can be a slight performance impact while the synchronization is in progress.

The mirror protection can also be configured using disk units from individual hardware-based RAID arrays. As shown in Figure 4-4, there are two sets of SAS RAID adapters in Dual Storage IOA configuration, and the mirror protection configured in IBM i to uses different RAID disk units to be mirrored.

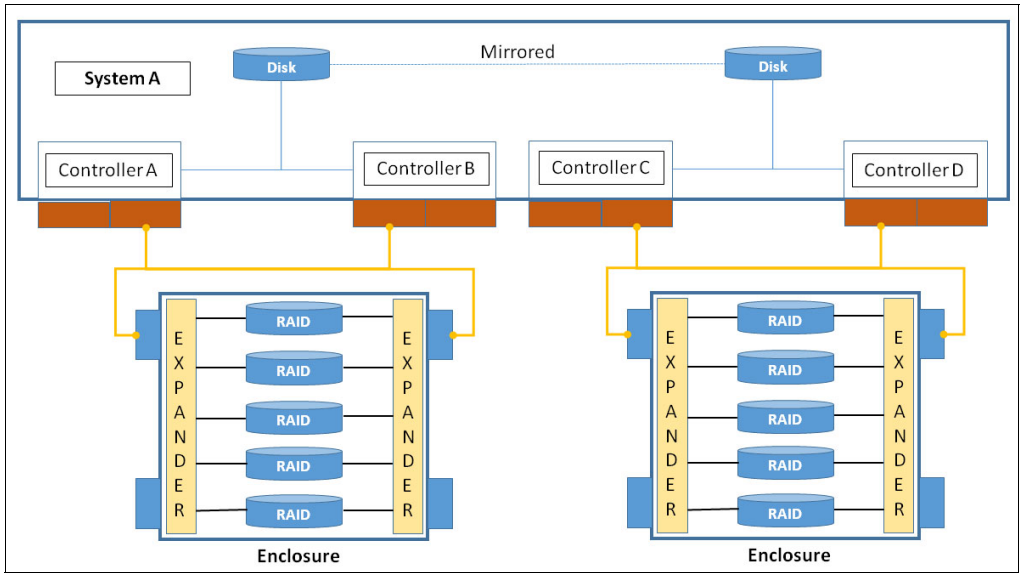


Figure 4-4 IBM i Dual Storage IOA with mirroring

A

RAID in storage subsystems

This appendix provides an introduction to RAID technology used in Storage subsystems that exist in SAN.

In addition, this appendix introduces two innovative IBM solutions that address potential challenges exist with traditional RAID levels:

- IBM XIV® Storage System
- IBM Spectrum™ Scale RAID

This appendix includes the following sections:

- RAID in storage subsystems
- Two innovative solutions

A.1 RAID in storage subsystems

Storage area network (SAN) is a Fibre Channel protocol based network to connect host server systems to the storage subsystems in the enterprise data centers. Typically, SAN consists of hundreds or thousands of server systems hardware and virtual machines, and connect them to hundreds of storage subsystems using SAN or Fibre Channel switches in the network. The SAN is an extension of the traditional direct-attached storage (DAS) that is locally attached in each server system.

The storage subsystems in the SAN provide the disk space for the host server systems or virtual machines to store and process data. It is not uncommon for one storage subsystem in the SAN to provide disk space to multiple server systems that running various operating systems and application workloads. SAN also provides an efficient infrastructure to share disks with multiple servers. They can be shared either in individual storage disk allocations or shared storage allocations for a set of server systems that want to access the data in the shared storage either concurrently or in a non-concurrent share access mode.

The storage subsystems use the SAS protocol for their internal disks. The storage subsystems have controllers that provide the required hardware adapters for host connectivity to the subsystem. They also use SAS RAID adapters to create a virtual disk or logical unit number (LUN) that is configured in one of the supported RAID levels with multiple SAS hard disks based on the level of RAID used.

Most of the storage subsystems, including IBM DS8000 series and IBM Storwize® storage systems, provide support for various levels of RAID to configure their internal SAS HDDs or SDDs, including Easy Tier configurations.

The following RAID levels are supported:

RAID 0 Data striped across one or more drives, no redundancy

RAID 1 Data is mirrored between two drives

RAID 5 Data is striped across a minimum of three drives with one rotating parity

RAID 6 Data is striped across a minimum of five drives with two rotating parities

RAID 10 Data is striped across pairs of mirrored drives, with a minimum of two drives

A.2 Two innovative solutions

This appendix provides a focused description of two innovative approaches that are used in two IBM solutions that address some of the challenges with the traditional RAID levels.

A.2.1 IBM XIV Storage System

IBM XIV Storage System has an innovative approach to providing data protection against disk failures. This approach is different from conventional RAID technology. The primary difference in IBM XIV Storage System is that it is designed to distribute data across all internal resources. This data distribution architecture helps enhance performance, and addresses one of the potential challenges that exist in using traditional RAID models in storage subsystems with large capacity disk drives.

Traditional RAID technologies can potentially consume hours or even days when rebuilding large failed disks, for example, 4 TB or 6 TB. IBM XIV storage subsystem addresses this challenge with its innovative data distribution model that can achieve rebuilding data for large

drives within an hour, and within 30 minutes for smaller drives. This is achieved by a number of functions built into XIV system, two of which are parallelism and data distribution.

A.2.1.1 Parallelism and Data distribution

Parallelism is in the heart of the IBM XIV innovative architecture. It is in the way in which work internal and redundant switch networking connects the individual modules together to create the grid architecture. This grid architecture provides performance enhancements, compatibility, and scalability functions to IBM XIV. The distribution and parallelism in XIV Storage System is depicted in Figure A-1.

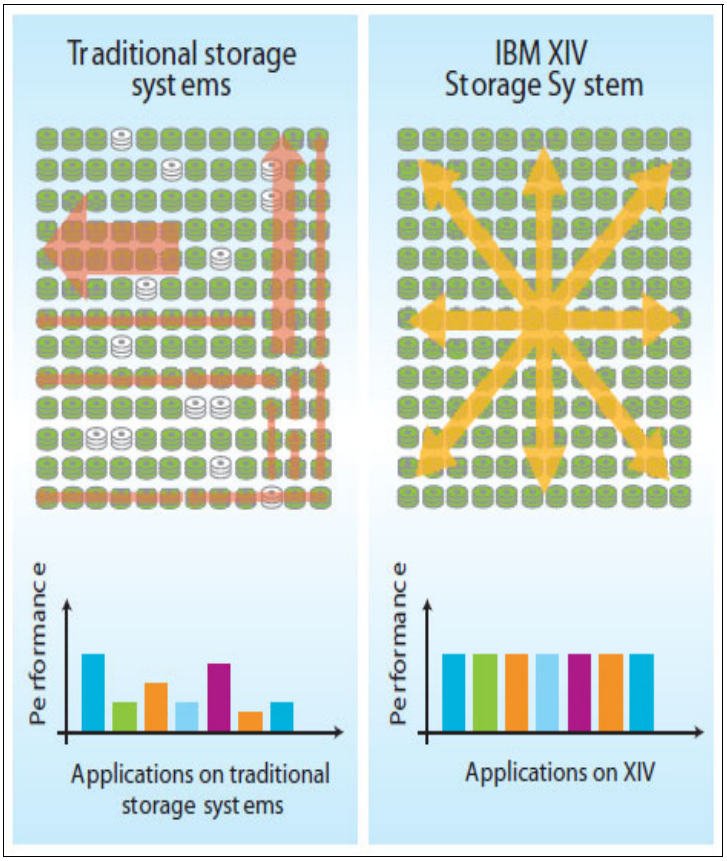


Figure A-1 Distribution with XIV

Along with the hardware parallelism, the software parallelism provides great advantages to the grid architecture. Data distribution algorithms are one of the software parallelism functions, which distributes data across all drives in a pseudo-random fashion. Patented algorithms provide a uniform yet random distribution of data, which is divided into 1 MB partitions across all available disks, to maintain data resilience and redundancy.

Equal use of every component maximizes performance in three ways:

- XIV Storage System engages the performance capabilities of all drives all the time.
- XIV Storage System engages the performance capabilities of every module in the grid.
- XIV Storage System eliminates disk hotspots.

In the traditional RAID technology based subsystems, high performance is achieved by one of the three common practices:

- Break up the application into multiple LUNs and distribute them among several RAID arrays
- Use operating system logical volume striping techniques to distribute the operating system across the performance capabilities of multiple RAID arrays
- Storage subsystem striping techniques that create LUNs that are striped across multiple RAID arrays

With XIV Storage System, there is no need to resort to these techniques. The massive parallelism approach and the way data is effectively distributed within the XIV Storage System grid allows the system to provide the best performance with no manual planning or configurations. The logical diagram of this grid architecture is illustrated in Figure A-2. It is unlikely that an XIV Storage System LUN will ever need to be reconfigured for performance reasons.

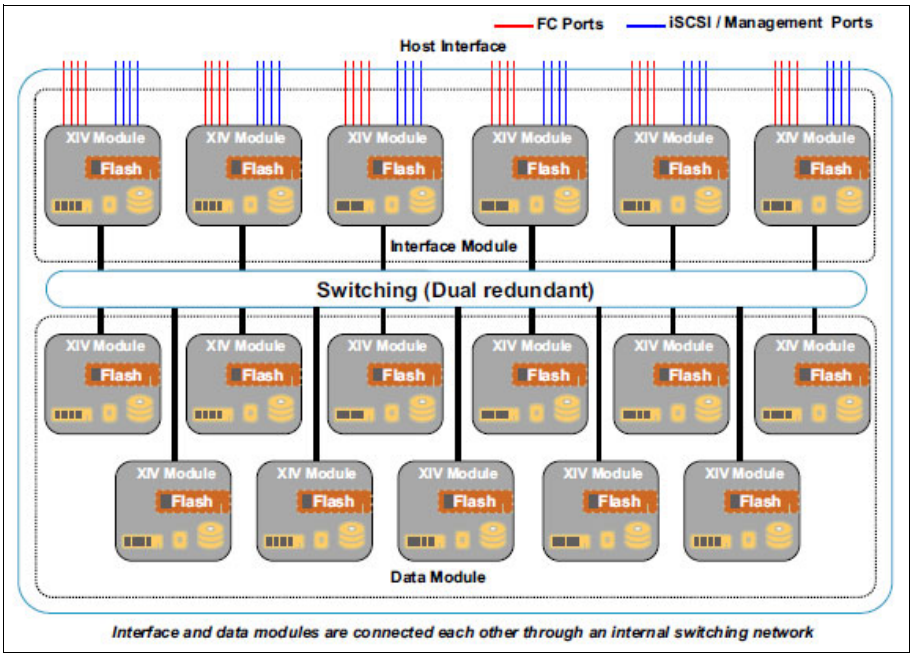


Figure A-2 XIV Grid Architecture

A.2.2 IBM Spectrum Scale RAID (formerly GPFS Native RAID)

The second IBM technology that addresses the challenges in the traditional RAID configurations is the IBM Spectrum Scale RAID. IBM Spectrum Scale RAID is one of the features of the Spectrum Scale software (formerly known as GPFS).

The IBM Spectrum Scale RAID integrates the functions of an advanced storage controller into the Spectrum Scale NSD server. Unlike an external storage controller, where configuration, LUN definition, and maintenance are beyond the control of Spectrum Scale, IBM Spectrum Scale RAID itself takes on the role of controlling, managing, and maintaining physical disks, both hard disk drives (HDDs) and solid-state drives (SSDs).

Compared to conventional RAID, IBM Spectrum Scale RAID implements a sophisticated data and spare space disk layout scheme that allows for arbitrarily sized disk arrays while also reducing the performance effect that clients experience when recovering from disk failures. To accomplish this, Spectrum Scale RAID uniformly spreads or declusters user data, redundancy information, and spare space across all the disks of a declustered array.

Figure A-3 compares a conventional RAID layout versus an equivalent declustered array.

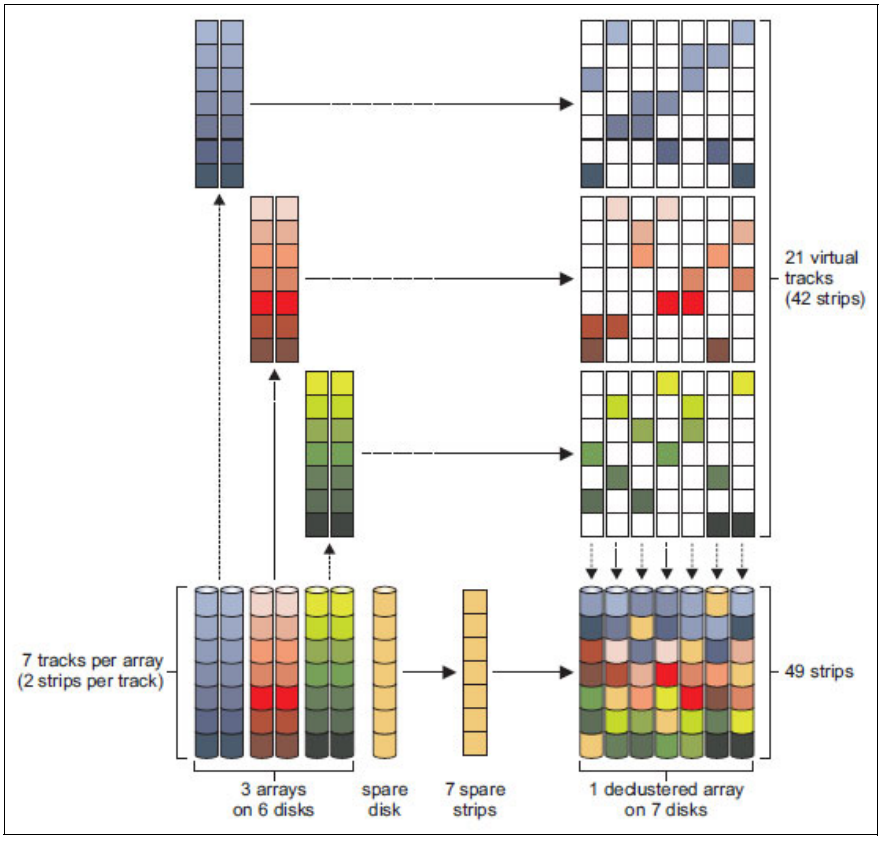


Figure A-3 Comparison of RAID versus Spectrum Scale RAID

As illustrated in Figure A-4, a declustered array can significantly shorten the time that is required to recover from a disk failure, which lowers the rebuild performance effect for client applications. When a disk fails, erased data is rebuilt using all the operational disks in the declustered array, the bandwidth of which is greater than that of the fewer disks of a conventional RAID group. Furthermore, if an additional disk fault occurs during a rebuild, the number of impacted tracks that require repair is markedly less than the previous failure and less than the constant rebuild performance effect of a conventional array.

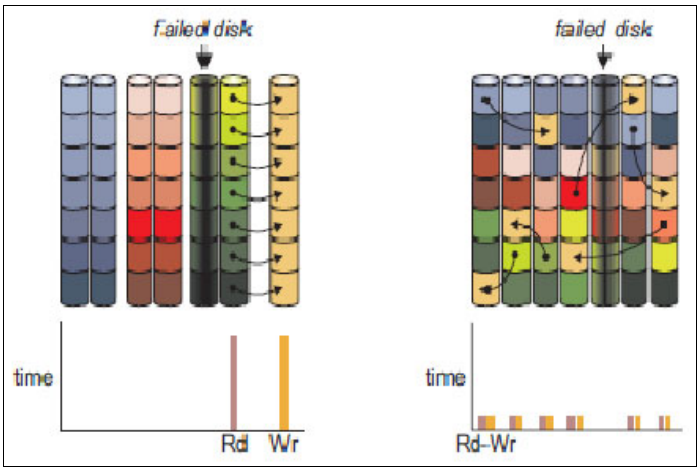


Figure A-4 Rebuilding with Spectrum Scale RAID

The decrease in declustered rebuild impact and client performance effect can be a factor of three to four times less than a conventional RAID. Because GPFS stripes client data across all the storage nodes of a cluster, file system performance becomes less dependent upon the speed of any single rebuilding storage array.

At the time of writing, IBM Spectrum Scale RAID is available for AIX and Linux operating systems, and uses the following configurations:

- IBM Spectrum Scale using IBM POWER8 processor-based systems

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this paper.

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.

- *Deploying Flex System in a BladeCenter Environment*, REDP-5122
- *IBM Power Systems E870 and E880 Technical Overview and Introduction*, REDP-5137
- *IBM Power Systems HMC Implementation and Usage Guide*, SG24-7491
- *IBM Power Systems S812L and S822L Technical Overview and Introduction*, REDP-5098
- *IBM Power Systems S814 and S824 Technical Overview and Introduction*, REDP-5097
- *IBM Power System S822 Technical Overview and Introduction*, REDP-5102
- *IBM Power System S824L Technical Overview and Introduction*, REDP-5139
- *Performance Optimization and Tuning Techniques for IBM Processors, including IBM POWER8*, SG24-8171

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

Online resources

These websites are also relevant as further information sources:

- The IBM System Planning Tool is available at:

  http://www-947.ibm.com/systems/support/tools/systemplanningtool/
- For a full list of all supported PCIe adapters for S814, see the IBM Knowledge Center at:

  http://www-01.ibm.com/support/knowledgecenter/POWER8/p8hdx/POWER8welcome.htm
- For a detailed description of cabling in the system units, see the IBM Knowledge Center at:

  http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling.htm
- For a detailed description of cabling in the PCIe Gen3 Expansion I/O drawer, see the IBM Knowledge Center at:

  http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8had/p8had_sascabling5887.htm

- *Performance Study of IBM Power Systems - SAS RAID Adapter Easy Tier Function*:

  http://w3-01.ibm.com/sales/ssi/cgi-bin/ssialias?subtype=WH&infotype=SA&appname=STGI_PO_PO_USEN&htmlfid=POW03129USEN&attachment=POW03129USEN.PDF

- *Performance Study of 2nd Generation IBM Power Systems SAS RAID Adapters Designed for Solid State Storage*:

  http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=WH&infotype=SA&appname=STGE_PO_PO_USEN&htmlfid=POW03122USEN&attachment=POW03122USEN.PDF

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

REDP-5234-00

ISBN 0738454346

Printed in U.S.A.