IBM Power System S824L
Technical Overview and Introduction

Linux server built on OpenPOWER technologies

Ready for scientific, Java, and Big Data & Analytics workloads

Dual NVIDIA GPU accelerators supported

Alexandre Bicas Caldeira
YoungHoon Cho
James Cruickshank
Bartłomiej Grabowski
Volker Haug
Andrew Laidlaw
Seulgi Yoppy Sung

ibm.com/redbooks
Contents

Notices ................................................................. vii
Trademarks ........................................................... viii

Preface ................................................................. ix
Authors ................................................................ ix
Now you can become a published author, too! .......... x
Comments welcome ................................................ xi
Stay connected to IBM Redbooks ............................... xi

Chapter 1. General description .................................. 1
1.1 Power S824L server .............................................. 2
1.2 NVIDIA Tesla GPU Accelerators ......................... 3
  1.2.1 NVIDIA CUDA ........................................... 5
1.3 Operating environment ......................................... 5
1.4 Physical package ................................................. 6
1.5 Server features ................................................... 7
  1.5.1 Server features for a Power S824L with a NVIDIA GPU. 7
  1.5.2 Server features for a Power S824L without a GPU. 8
  1.5.3 Minimum features ....................................... 8
  1.5.4 Power supply features ................................ 8
  1.5.5 Processor module features .......................... 9
  1.5.6 Memory features ....................................... 9
1.6 PCIe slots .......................................................... 10
1.7 Disk and media features ...................................... 11
1.8 I/O drawers for Power S824L server .................... 13
  1.8.1 PCIe Gen3 I/O expansion drawer .................... 14
  1.8.2 I/O drawers and usable PCI slot .................... 15
  1.8.3 EXP24S SFF Gen2-bay drawer .................... 16
1.9 Server and virtualization management ................. 17
1.10 System racks .................................................... 18
  1.10.1 IBM 7014 Model T00 rack .......................... 18
  1.10.2 IBM 7014 Model T42 rack .......................... 19
  1.10.3 IBM 42U Slim Rack 7965-94Y ....................... 21
  1.10.4 Feature code 0551 rack ............................ 21
  1.10.5 Feature code 0553 rack ............................ 21
  1.10.6 Feature code ER05 rack ............................ 21
  1.10.7 AC power distribution unit and rack content ...... 22
  1.10.8 Rack-mounting rules ................................ 24
  1.10.9 Useful rack additions ............................... 24
  1.10.10 OEM rack ............................................ 24

Chapter 2. Architecture and technical overview .............. 27
2.1 The IBM POWER8 processor ................................. 28
  2.1.1 POWER8 processor overview ....................... 28
  2.1.2 POWER8 processor core ............................ 31
  2.1.3 Simultaneous multithreading ....................... 32
  2.1.4 Memory access ....................................... 32
  2.1.5 On-chip L3 cache innovation and Intelligent Cache 33
  2.1.6 L4 cache and memory buffer ....................... 34
Chapter 3. Reliability, availability, and serviceability

3.1 Introduction ........................................................................... 74
  3.1.1 RAS enhancements of POWER8 processor-based servers ....... 74
3.2 Reliability ............................................................................. 75
  3.2.1 Designed for reliability .................................................... 76
  3.2.2 Placement of components ................................................ 76
3.3 Processor and memory availability details ......................... 77
3.3.1 Correctable error introduction ........................................ 77
3.3.2 Uncorrectable error introduction .................................... 78
3.3.3 Processor core/cache correctable error handling .................. 78
3.3.4 Processor Instruction Retry and other try again techniques .... 78
3.3.5 Alternative processor recovery and Partition Availability Priority . 79
3.3.6 Core Contained Checkstops and other PowerVM error recovery .... 79
3.3.7 Cache uncorrectable error handling ................................. 79
3.3.8 Other processor chip functions .................................... 80
3.3.9 Other fault error handling ....................................... 80
3.3.10 Memory protection ............................................... 81
3.3.11 I/O subsystem availability and Enhanced Error Handling ........ 82
3.4 Enterprise systems availability details ................................ 83
3.5 Availability effects of a solution architecture ....................... 83
3.6 Serviceability ....................................................... 84
  3.6.1 Detecting introduction .......................................... 84
  3.6.2 Error checkers, fault isolation registers, and First-Failure Data Capture .... 84
  3.6.3 Service processor ............................................. 85
  3.6.4 Diagnosing ................................................... 86
  3.6.5 Reporting .................................................... 88
  3.6.6 Notifying .................................................... 90
  3.6.7 Locating and servicing ....................................... 91
3.7 Manageability ...................................................... 94
  3.7.1 Service user interfaces ....................................... 94
  3.7.2 IBM Power Systems Firmware maintenance ................. 97
  3.7.3 Concurrent firmware update improvements .................. 101
  3.7.4 Electronic Services and Electronic Service Agent .......... 101
3.8 Selected POWER8 RAS capabilities by operating system ........ 105

Related publications .................................................. 107
IBM Redbooks .................................................................. 107
Other publications ..................................................... 107
Online resources ........................................................ 108
Help from IBM .......................................................... 109
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.

© Copyright IBM Corp. 2014. All rights reserved.
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:

<table>
<thead>
<tr>
<th>Active Memory™</th>
<th>Power Systems™</th>
<th>PowerVM®</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIX®</td>
<td>Power Systems Software™</td>
<td>Redbooks®</td>
</tr>
<tr>
<td>DS8000®</td>
<td>POWER6®</td>
<td>Redpaper™</td>
</tr>
<tr>
<td>Easy Tier®</td>
<td>POWER6+™</td>
<td>Redbooks (logo) ®</td>
</tr>
<tr>
<td>Electronic Service Agent™</td>
<td>POWER7®</td>
<td>RS/6000®</td>
</tr>
<tr>
<td>EnergyScale™</td>
<td>POWER7+™</td>
<td>System p®</td>
</tr>
<tr>
<td>Focal Point™</td>
<td>POWER8™</td>
<td>System Storage®</td>
</tr>
<tr>
<td>IBM®</td>
<td>PowerHA®</td>
<td>System z®</td>
</tr>
<tr>
<td>POWER®</td>
<td>PowerLinux™</td>
<td></td>
</tr>
<tr>
<td>POWER Hypervisor™</td>
<td>PowerPC®</td>
<td></td>
</tr>
</tbody>
</table>

The following terms are trademarks of other companies:

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

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

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

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

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.
Preface

This IBM® Redpaper™ publication is a comprehensive guide that covers the IBM Power System S824L (8247-42L) servers that support Linux operating systems. The objective of this paper is to introduce the major innovative Power S824L offerings and their relevant functions:

- The new IBM POWER8™ processor, which is available at frequencies of 3.02 GHz and 3.42 GHz, 3.52 GHz, and 4.15 GHz
- A processor that is designed to accommodate high-wattage adapters, such as NVIDIA graphics processing units (GPUs), that provide acceleration for scientific, engineering, Java, big data analytics, and other technical computing workloads
- Based on OpenPOWER technologies
- Two integrated memory controllers with improved latency and bandwidth
- I/O drawer expansion options offers greater flexibility
- Improved reliability, serviceability, and availability (RAS) functions
- IBM EnergyScale™ technology that provides features, such as power trending, power-saving, power capping, and thermal measurement

This publication is for professionals who want to acquire a better understanding of IBM Power Systems™ products. The intended audience includes the following roles:

- Clients
- Sales and marketing professionals
- Technical support professionals
- IBM Business Partners
- Independent software vendors

This paper expands the current set of IBM Power Systems documentation by providing a desktop reference that offers a detailed technical description of the Power S824L server.

This paper does not replace the latest marketing materials and configuration tools. It is intended as an additional source of information that, together with existing sources, can be used to enhance your knowledge of IBM server solutions.

Authors

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

**Alexandre Bicas Caldeira** is a Certified IT Specialist and a member of the Power Systems Advanced Technical Sales Support team for IBM Brazil. He holds a degree in Computer Science from the Universidade Estadual Paulista (UNESP) and an MBA in Marketing. His major areas of focus are competition, sales, and technical sales support. Alexandre has more than 14 years of experience working with IBM Systems & Technology Group Solutions. He has also worked as an IBM Business Partner on Power Systems hardware, AIX®, and IBM PowerVM® virtualization products.

**YoungHoon Cho** is a Power Systems Top Gun with the post-sales Technical Support Team for IBM in Korea. He has over 10 years of experience working on RS/6000®, System p®, and...
James Cruickshank works in the Power System Client Technical Specialist team for IBM in the UK. He holds an honors degree in Mathematics from the University of Leeds. James has over 13 years experience working with RS/6000, pSeries, System p, and Power Systems products. James supports clients in the financial services sector in the UK.

Bartłomiej Grabowski is a Principal System Support Specialist in DHL IT Services in the Czech Republic. He holds a Bachelor's degree in Computer Science from the Academy of Computer Science and Management in Bielsko-Biala. His areas of expertise include IBM i, PowerHA® solutions that are based on hardware and software replication, Power Systems hardware, and PowerVM. He is an IBM Certified Systems Expert and coauthor of several IBM Redbooks® publications.

Volker Haug is an Open Group Certified IT Specialist within IBM Systems in Germany supporting Power Systems clients and Business Partners. He holds a Diploma degree in Business Management from the University of Applied Studies in Stuttgart. His career includes more than 28 years of experience with Power Systems, AIX, and PowerVM virtualization. He has written several IBM Redbooks publications about Power Systems and PowerVM. Volker is an IBM POWER8 Champion and a member of the German Technical Expert Council, which is an affiliate of the IBM Academy of Technology.

Andrew Laidlaw is a Client Technical Specialist for IBM working in the UK. He supports Service Provider clients within the UK and Ireland, focusing primarily on Power Systems running AIX and Linux workloads. His expertise extends to open source software package including the KVM hypervisor and various management tools. Andrew holds an Honors degree in Mathematics from the University of Leeds, which includes credits from the University of California in Berkeley.

Seulgi Yoppy Sung is a very passionate Engineer, supporting multi-platform systems as a System Services Representative almost three year, include Power System hardware, AIX, high-end and low-end storage DS8000® and V7000. She is very positive and enthusiastic about Power Systems.

The project that created this document was managed by: Scott Vetter, PMP

Thanks to the following people for their contributions to this project:
Tamiikia Barrow, Bruno Blanchard, Ella Buslovich, Mark Clark, Daniel Henderson, Tenley Jackson, Stephanie Jensen, Bob Kovacks, Jai Lei Ma, Cesar D Maciel, Chris Mann, Dwayne Moore, Mark Olson, Monica Sanchez, Bill Starke, Jeff Stuecheli, Doug Szerdi, Jacobo Vargas, Steve Will

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 client 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:
▶ Stay current on recent Redbooks publications with RSS Feeds:
  http://www.redbooks.ibm.com/rss.html
General description

The IBM Power System S824L (8247-42L) server is orderable in two different configurations:

1. Power S824L with a NVIDIA graphics processing unit (GPU)
   This is the first Power Systems server that is designed to accommodate high-performance adapters, such as NVIDIA GPUs. This offering delivers a new class of technology that maximizes performance and efficiency for scientific, engineering, Java, big data analytics, and other technical computing workloads.

2. Power S824L without a NVIDIA graphics processing unit (GPU)
   This Power Systems server is designed to deliver unprecedented performance, scalability, reliability, and manageability for demanding commercial workloads.

Previously before April 2015, the Power S824L server always included one or two GPUs. Now a GPU is optional and the S824L configuration options are tremendously expanded, making it a more general purpose, Linux-only server when a GPU is not included.

**Note:** If the Power S824L server was initially ordered without a GPU, then adding a GPU to that server is not supported.

The IBM Power System S824L servers use the latest POWER8 processor technology that is designed to deliver unprecedented performance, scalability, reliability, and manageability for demanding commercial workloads. Designed to empower the ecosystem of open source development, these new servers support the Linux operating systems (OS) from several distributors.

**Note:** For a Power S824L system that contains a GPU Ubuntu is the supported Linux operating system. The SAP HANA edition is supported by SLES 11. The OPAL Baremetal is supported by Ubuntu and Red Hat. The KVM and PowerVM for OpenPower editions support Red Hat, SLES, and Ubuntu.

The following sections provide detailed information about the Power S824L server.
1.1 Power S824L server

The Power S824L (8247-42L) server is a powerful one to two socket server. The following configurations are available:

- The Power S824L with the NVIDIA GPU is available as a two socket server.
- The Power S824L without the NVIDIA GPU is available as a one to two socket server.

It has one to two POWER8 dual chip module (DCM) processor modules that offer 4.15 GHz, 3.52 GHz, 3.42 GHz and 3.02 GHz performance with 8, 16 20 or 24 fully activated cores. In addition, the Power S824L server can be equipped with at least one NVIDIA GPU, with a maximum of two allowed.

The Power S824L server supports a maximum of 16 DDR3 Custom dual inline memory modules (CDIMM) slots. The memory features (two memory DIMMs per feature) that are supported are 16 GB, 32 GB, 64 GB and 128 GB. They run at speeds of 1600 MHz for a maximum system memory of 2 TB.

The server supports a storage backplane that consists of twelve SFF-3 bays and one DVD bay, which allows for up to 14.4 TB of raw data to be physically installed on the servers. These 2.5-inch or small form factor (SFF) SAS bays can contain SAS drives (hard disk drives (HDDs) mounted on a Gen3 tray or carrier. All SFF-3 bays support concurrent maintenance or hot-plug capability. The storage backplane uses leading-edge, integrated serial-attached SCSI (SAS) RAID controller technology that is designed and patented by IBM.

At the time of writing, Linux operating systems from RedHat, SUSE, and Ubuntu are supported on the system.

The Power S824L server supports PowerVM, KVM, virtualization in a non-GPU configuration. Opal Baremetal, and SAP Hana editions are also available.

The Power S824L comes with one to two POWER8 processor DCMs that provide excellent configuration flexibility and expandability.

Seven hot swap PCI Express (PCIe) Gen3 Full High adapter slots are available with the one POWER8 processor DCM configuration: two x16 slots and five x8 slots. The x16 slots can provide up to twice the bandwidth of an x8 slot because they offer twice as many PCIe lanes. PCIe Gen3 slots can support up to twice the bandwidth of a PCIe Gen2 slot and up to four times the bandwidth of a PCIe Gen1 slot, assuming an equivalent number of PCIe lanes.

Eleven hot swap PCI Express (PCIe) Gen3 Full High adapter slots are available with two POWER8 processor DCMs: four x16 slots and seven x8 slots. The x16 slots can provide up to twice the bandwidth of an x8 slot because they offer twice as many PCIe lanes. PCIe Gen3 slots can support up to twice the bandwidth of a PCIe Gen2 slot and up to four times the bandwidth of a PCIe Gen1 slot, assuming an equivalent number of PCIe lanes.
1.2 NVIDIA Tesla GPU Accelerators

NVIDIA Tesla GPUs are massively parallel accelerators that are based on the NVIDIA Compute Unified Device Architecture (CUDA) parallel computing platform and programming model. Tesla GPUs are designed from the ground up for power-efficient, high performance computing, computational science, supercomputing, big data analytics, and machine learning applications, delivering dramatically higher acceleration than a CPU-only approach.

Two NVIDIA accelerators are available. #EC47 and a 300W #EC4B

Figure 1-2 on page 4 shows a NVIDIA Tesla GPU.
These NVIDIA Tesla GPU Accelerators are based on the NVIDIA Kepler Architecture and designed to run the most demanding scientific models faster and more efficiently. With the introduction of Tesla K40 GPU Accelerators, you can run large scientific models on its 12 GB of GPU accelerator memory, which can process 2x larger data sets and is ideal for big data analytics. It also outperforms CPUs by up to 10x with its GPU Boost feature, which converts power headroom into user-controlled performance boost. Table 1-1 shows a summary of its characteristics.

**Table 1-1  NVIDIA Tesla K40 and K80 specifications**

<table>
<thead>
<tr>
<th>Features</th>
<th>Tesla K40 #EC47</th>
<th>Tesla K80 #EC4B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number and type of GPU</td>
<td>1 Kepler GK110B</td>
<td>2 Kepler GK210</td>
</tr>
<tr>
<td>Peak double precision floating point performance</td>
<td>1.43 Tflops</td>
<td>2.91 Tflops</td>
</tr>
<tr>
<td>Peak single precision floating point performance</td>
<td>4.29 Tflops</td>
<td>8.73 Tflops</td>
</tr>
<tr>
<td>Memory bandwidth (error correction code (ECC) off)</td>
<td>288 GB/sec</td>
<td>480 GB/sec</td>
</tr>
<tr>
<td>Memory size (GDDR5)</td>
<td>12 GB</td>
<td>24 GB</td>
</tr>
<tr>
<td>CUDA cores</td>
<td>2,880</td>
<td>4,992</td>
</tr>
</tbody>
</table>

For more information about the NVIDIA Tesla GPU, see the NVIDIA data sheets: [http://www.nvidia.com/object/tesla-servers.html](http://www.nvidia.com/object/tesla-servers.html)
1.2.1 NVIDIA CUDA

NVIDIA CUDA is a parallel computing platform and programming model that enables dramatic increases in computing performance by harnessing the power of the graphics processing unit (GPU).

Today, the CUDA ecosystem is growing rapidly as more companies provide world-class tools, services, and solutions. If you want to start harnessing the performance of GPUs, the CUDA Toolkit provides a comprehensive development environment for C and C++ developers.

The easiest way to start is to use the plug-in scientific and math libraries that are available in the CUDA Toolkit to quickly accelerate common linear algebra, signal and image processing, and other common operations, such as random number generation and sorting. If you want to write your own code, the Toolkit includes a compiler, and debugging and profiling tools. You’ll also find code samples, programming guides, user manuals, API references, and other documentation to help you get started.

The CUDA Toolkit is available at no charge. Learning to use CUDA is convenient, with comprehensive online training available, and other resources, such as webinars and books. Over 400 universities and colleges teach CUDA programming, including dozens of CUDA Centers of Excellence and CUDA Research and Training Centers. Solutions for Fortran, C#, Python, and other languages are available.

Explore the GPU Computing Ecosystem on CUDA Zone to learn more:

The production release of CUDA 5.5 for POWER8 (and any subsequent release) is available for download on the following link:

1.3 Operating environment

Table 1-2 lists the operating environment specifications for the Power S824L servers.

<table>
<thead>
<tr>
<th>Power S824L operating environment</th>
<th>Operating</th>
<th>Non-operating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temperature</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allowable: 5 - 35 degrees C(^\circ) (41 - 95 degrees F)</td>
<td>5 - 45 degrees C (41 - 113 degrees F)</td>
<td></td>
</tr>
<tr>
<td>Recommended: 18 - 27 degrees C (64 - 80 degrees F)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relative humidity</td>
<td>8 - 80%</td>
<td>8 - 80%</td>
</tr>
<tr>
<td>Maximum dew point</td>
<td>28 degrees C (84 degrees F)</td>
<td>N/A</td>
</tr>
<tr>
<td>Operating voltage</td>
<td>200 - 240 V AC 180 - 400 V DC</td>
<td>N/A</td>
</tr>
<tr>
<td>Operating frequency</td>
<td>47 or 63 Hz AC</td>
<td>N/A</td>
</tr>
<tr>
<td>Power consumption</td>
<td>2,300 watts maximum</td>
<td>N/A</td>
</tr>
</tbody>
</table>
1.4 Physical package

Table 1-3 shows the physical dimensions of the Power S824L chassis. The servers are available only in a rack-mounted form factor and take 4U (4 EIA units) of rack space.

<table>
<thead>
<tr>
<th>Description</th>
<th>Operating</th>
<th>Non-operating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power source loading</td>
<td>2.38 kVA maximum</td>
<td>N/A</td>
</tr>
<tr>
<td>Thermal output</td>
<td>7,848 BTU/hour maximum</td>
<td>N/A</td>
</tr>
<tr>
<td>Maximum altitude</td>
<td>3,050 m (10,000 ft)</td>
<td>N/A</td>
</tr>
<tr>
<td>Noise level and sound power</td>
<td>6.8 bels operating; 6.8 bels idling</td>
<td>N/A</td>
</tr>
</tbody>
</table>

a. Heavy workloads might see some performance degradation above 35 degrees C if internal temperatures trigger a CPU clock reduction.

Tip: The maximum measured value is expected from a fully populated server under an intensive workload. The maximum measured value also accounts for component tolerance and operating conditions that are not ideal. Power consumption and heat load vary greatly by server configuration and usage. Use the IBM Systems Energy Estimator to obtain a heat output estimate that is based on a specific configuration, which is available at the following website:

http://www-912.ibm.com/see/EnergyEstimator
Figure 1-3 shows the rear view of a Power S824L server.

![Rear view of a Power S824L server](image)

**Figure 1-3  Rear view of a Power S824L server**

### 1.5 Server features

The server chassis of the Power S824L contains a maximum of two processor modules. Each POWER8 processor module is either 8-core, 10-core or 12-core and has a 64-bit architecture, up to 512 KB of L2 cache per core, and up to 8 MB of L3 cache per core.

#### 1.5.1 Server features for a Power S824L with a NVIDIA GPU

This summary describes the standard features of the Power S824L with a NVIDIA GPU:

- Rack-mount (4U) chassis
- Dual processor module:
  - 10-core 3.42 GHz processor module
  - 12-core 3.02 GHz processor module
- Up to 2 TB of 1600 MHz DDR3 ECC memory
- One or two NVIDIA graphics processing units (GPUs)
- Storage backplane with 12 SFF-3 bays and one DVD bay
- Hot-swap PCIe Gen 3 slots
- Integrated features:
  - Service Processor
  - EnergyScale technology
  - Hot-swap and redundant cooling
  - Four USB 3.0 ports for general usage (two in back and two in front)
  - Two 1 Gbps Hardware Management Console (HMC) RJ45 ports
  - One system port with RJ45 connector
- Four hot-plug, redundant power supplies (AC only)
- Ubuntu Linux and Red Hat are the supported operating systems
1.5.2 Server features for a Power S824L without a GPU

This summary describes the standard features of the Power S824L without a NVIDIA GPU:

- Rack-mount (4U) chassis
- Dual processor module:
  - 8-core 4.15 GHz processor module
  - 12-core 3.52 GHz processor module
- Up to 2 GB of 1600 MHz DDR3 ECC memory
- Storage backplane with 12 SFF-3 bays and one DVD bay
- Hot-swap PCIe Gen 3 slots
- Optional external PCIe Gen3 I/O drawer
- Integrated features:
  - Service Processor
  - EnergyScale technology
  - Hot-swap and redundant cooling
  - Four USB 3.0 ports for general usage (two in back and two in front)
  - Two 1 Gbps Hardware Management Console (HMC) RJ45 ports
  - One system port with RJ45 connector
- Four hot-plug, redundant power supplies
- RedHat and SUSE (Big Endian) and RedHat, SUSE, Ubuntu (Little Endian) are the supported Linux operating systems

1.5.3 Minimum features

The minimum Power S824L initial order features are listed:

- One processor modules (two processor module are required for a Power S824L with a NVIDIA GPU)
- Processor activations
- 32 GB of memory
- A storage backplane
- One HDD DASD device
- A PCIe2 4-port 1 Gb Ethernet adapter
- A Compute-Intensive Accelerator required for a Power S824L with a NVIDIA GPU
- Four power supplies and power cords (AC or DC)
- An OS indicator
- A cover set indicator
- A Language Group Specify

1.5.4 Power supply features

Four redundant 1400 watt 200 - 240 V AC power supplies (#EL1B) are supported on the Power S824L server.

The server continues to function with two working power supplies. A failed power supply can be hot-swapped but it must remain in the server until the replacement power supply is available for exchange.
1.5.5 Processor module features

The following shows the available processor module feature for the various Power S824L configurations:

**Power S824L with a NVIDIA GPU**

The Power S824L with a NVIDIA GPU supports two processor modules with either 10 processor cores (#ELP2) or 12 processor cores (#ELP5). All processor cores must be activated.

The following list defines the allowed quantities of processor activation entitlements:

- Two 10-core, 3.42 GHz processors (#ELP2) require that 20 processor activation codes are ordered. A maximum of 20 processor activations (#ELA2) are allowed.
- Two 12-core, 3.02 GHz processors (#ELP5) require that 24 processor activation codes are ordered. A maximum of 24 processor activations (#ELA5) are allowed.

Table 1-4 summarizes the processor features that are available for the Power S824L with a NVIDIA GPU.

**Table 1-4  Processor features for the Power S824L with a NVIDIA GPU**

<table>
<thead>
<tr>
<th>Feature code</th>
<th>Processor module description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELP2</td>
<td>10-core 3.42 GHz POWER8 processor card</td>
</tr>
<tr>
<td>ELP5</td>
<td>12-core 3.02 GHz POWER8 processor card</td>
</tr>
</tbody>
</table>

**Power S824L without a NVIDIA GPU**

The Power S824L without a NVIDIA GPU supports one to two processor modules with either 8 processor cores (#ELPH) or 12 processor cores (#ELP5). All processor cores must be activated.

The following list defines the allowed quantities of processor activation entitlements:

- One or two 8-core, 3.42 GHz processors (#ELP2) require that eight or 16 processor activation codes are ordered. A maximum of 16 processor activations (ELAH) are allowed.
- One or two 12-core, 3.52 GHz processors (#ELP5) require that 12 or 24 processor activation codes are ordered. A maximum of 24 processor activations (#ELAJ) are allowed.

Table 1-5 summarizes the processor features that are available for the Power S824L without a NVIDIA GPU.

**Table 1-5  Processor features for the Power S824L without a NVIDIA GPU**

<table>
<thead>
<tr>
<th>Feature code</th>
<th>Processor module description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELPH</td>
<td>8-core 4.15 GHz POWER8 processor card</td>
</tr>
<tr>
<td>ELPJ</td>
<td>12-core 3.52 GHz POWER8 processor card</td>
</tr>
</tbody>
</table>

1.5.6 Memory features

A minimum of 32 GB of memory is required on the Power S824L server. You are required to order memory in pairs. Base memory is two 16 GB, 1600 MHz DDR3 memory features (#EM8B). DDR4 DIMMs (VM91, EM92, EM93, EM94) operate at DDR3 speeds.
Table 1-6 lists the memory features that are available on the Power S824L server.

<table>
<thead>
<tr>
<th>Feature code</th>
<th>DIMM capacity</th>
<th>Access rate</th>
<th>Maximum quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM8B</td>
<td>16 GB</td>
<td>1600 MHz</td>
<td>16</td>
</tr>
<tr>
<td>EM8C</td>
<td>32 GB</td>
<td>1600 MHz</td>
<td>16</td>
</tr>
<tr>
<td>EM8D</td>
<td>64 GB</td>
<td>1600 MHz</td>
<td>16</td>
</tr>
<tr>
<td>EM8E</td>
<td>128 GB</td>
<td>1600 MHz</td>
<td>16</td>
</tr>
<tr>
<td>EM91</td>
<td>16 GB</td>
<td>1600 MHz</td>
<td>16</td>
</tr>
<tr>
<td>EM92</td>
<td>32 GB</td>
<td>1600 MHz</td>
<td>16</td>
</tr>
<tr>
<td>EM93</td>
<td>64 GB</td>
<td>1600 MHz</td>
<td>16</td>
</tr>
<tr>
<td>EM94</td>
<td>128 GB</td>
<td>1600 MHz</td>
<td>16</td>
</tr>
</tbody>
</table>

Consider your plans for future memory upgrades when you decide which memory feature size to use at the time of the initial system order. All memory upgrades are performed in pairs with DIMMs of the same capacity.

1.6 PCIe slots

The following section describes the available PCIe slots.

The one socket Power S824L (no GPU) has up to seven PCIe hot plug Gen 3 slots. Two are x16 Full High slots and five are x8 Gen 3 Full High slots.

The two socket Power S824L (either with or without a GPU) has up to eleven PCIe hot plug Gen 3 slots. Four are x16 Full High slots and seven are x8 Gen 3 Full High slots.

The x16 slots can provide up to twice the bandwidth of x8 slots because they offer twice as many PCIe lanes. PCIe Gen 3 slots can support up to twice the bandwidth of a PCIe Gen 2 slot and up to four times the bandwidth of a PCIe Gen 1 slot, assuming an equivalent number of PCIe lanes.

**Note:** One of the x8 PCIe slots on the Power S824L servers is used for a LAN adapter which can be one of the following:

- EN0W: PCIe2 2-port 10/1GbE BaseT RJ45 Adapter
- EN0S: PCIe2 4-Port (10Gb+1GbE) SR+RJ45 Adapter

The new servers are smarter about energy efficiency for cooling a PCIe adapter environment. They sense which IBM PCIe adapters are installed in their PCIe slots. If an adapter requires higher levels of cooling, the new servers automatically speed up fans to increase the airflow across the PCIe adapters.

If it is required to have more PCIe I/O slots in a systems, you can attach external I/O drawers to the Power S824L. For more information, see 1.8, “I/O drawers for Power S824L server” on page 13.
1.7 Disk and media features

Three backplane options are available for the Power S824L with no GPU configuration. One of the three options must be configured:

1. Storage backplane with 12 SFF-3 bays and one DVD bay (#EJ0N)
2. Storage backplane with 12 SFF-3 bays and one DVD bay (#EJ0N) and feature EJ0S provides split backplane functionality
3. Storage backplane with 18 SFF-3 bays, 1.8-inch SSD cage bays (42A only), one DVD bay, dual integrated SAS controllers with write cache and IBM Easy Tier® functionality (#EJ0P)

Note: If the Power S824L comes with a NVIDIA GPU only the Storage backplane with 12 SFF-3 bays and one DVD bay (#EJ0N) is supported.

Each of the three backplane options provides SFF-3 SAS bays in the system unit. These 2.5-inch or small form factor (SFF) SAS bays can contain SAS drives (HDD or SSD) mounted on a Gen3 tray or carrier. Thus the drives are designated SFF-3. SFF-1 or SFF-2 drives do not fit in an SFF-3 bay. All SFF-3 bays support concurrent maintenance or hot-plug capability.

In addition to supporting HDDs and SSDs in the SFF-3 SAS bays of the Power S824, the storage backplane feature EJ0P supports a mandatory 8-bay, 1.8-inch SSD Cage (#EJTM). All eight bays are accessed by both of the integrated SAS controllers. The bays support concurrent maintenance (hot-plug). The SSD 1.8-inch drive, such as the 387 GB capacity feature EL4Q is supported.

The high performance SAS controllers provide RAID 0, RAID 5, RAID 6, and RAID 10 support. The dual SAS controllers can automatically move hot data to an attached SSD and cold data to an attached HDD by using the Easy Tier function.

Table 1-7 shows the available disk drive feature codes that can be installed in the Power S824L.

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>Max</th>
<th>OS support</th>
</tr>
</thead>
<tbody>
<tr>
<td>EL13</td>
<td>59EA</td>
<td>775 GB SFF-3 SSD for Linux</td>
<td>18</td>
<td>Linux</td>
</tr>
<tr>
<td>EL14</td>
<td>59E6</td>
<td>387 GB SFF-3 SSD for Linux</td>
<td>18</td>
<td>Linux</td>
</tr>
<tr>
<td>EL19</td>
<td></td>
<td>387 GB SFF-2 SSD for Linux</td>
<td>336</td>
<td>Linux</td>
</tr>
<tr>
<td>EL3G</td>
<td></td>
<td>775 GB SFF-2 SSD for Linux</td>
<td>336</td>
<td>Linux</td>
</tr>
<tr>
<td>EL4Q</td>
<td></td>
<td>387 GB SFF-2 4 K SSD for Linux</td>
<td>336</td>
<td>Linux</td>
</tr>
<tr>
<td>EL4S</td>
<td></td>
<td>775 GB SFF-2 4 K SSD for Linux</td>
<td>336</td>
<td>Linux</td>
</tr>
<tr>
<td>EL4U</td>
<td></td>
<td>387 GB SFF-3 4 K SSD for Linux</td>
<td>18</td>
<td>Linux</td>
</tr>
<tr>
<td>EL4W</td>
<td></td>
<td>775 GB SFF-3 4 K SSD for Linux</td>
<td>18</td>
<td>Linux</td>
</tr>
<tr>
<td>Feature code</td>
<td>CCIN</td>
<td>Description</td>
<td>Max</td>
<td>OS support</td>
</tr>
<tr>
<td>-------------</td>
<td>----------</td>
<td>------------------------------------------------------------------------------</td>
<td>-----</td>
<td>------------</td>
</tr>
<tr>
<td>EL1C</td>
<td>387 GB 1.8&quot; SAS 5xx SSD eMLC4 for Linux</td>
<td>8</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL2W</td>
<td>387 GB 1.8&quot; SAS 4k SSD eMLC4 for Linux</td>
<td>8</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL2X</td>
<td>775 GB 1.8&quot; SAS 5xx SSD eMLC4 for Linux</td>
<td>8</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL4K</td>
<td>775 GB 1.8&quot; SAS 4k SSD eMLC4 for Linux</td>
<td>8</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL62</td>
<td>3.82-4.0 TB 7200 RPM 4K SAS LFF-1 Nearline Disk Drive (Linux)</td>
<td>336</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL64</td>
<td>7.72-8.0 TB 7200 RPM 4K SAS LFF-1 Nearline Disk Drive (Linux)</td>
<td>336</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL78</td>
<td>387 GB SFF-2 SSD 5xx eMLC4 for Linux</td>
<td>336</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL7E</td>
<td>775 GB SFF-2 SSD 5xx eMLC4 for Linux</td>
<td>336</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL7K</td>
<td>387 GB SFF-3 SSD 5xx eMLC4 for Linux</td>
<td>18</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL7P</td>
<td>775 GB SFF-3 SSD 5xx eMLC4 for Linux</td>
<td>18</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL80</td>
<td>1.9 TB Read Intensive SAS 4k SFF-2 SSD for Linux</td>
<td>336</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL85</td>
<td>387 GB SFF-2 SSD 4k eMLC4 for Linux</td>
<td>336</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL8C</td>
<td>775 GB SFF-2 SSD 4k eMLC4 for Linux</td>
<td>336</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL8F</td>
<td>1.55 TB SFF-2 SSD 4k eMLC4 for Linux</td>
<td>336</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL8J</td>
<td>1.9 TB Read Intensive SAS 4k SFF-3 SSD for Linux</td>
<td>18</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL8N</td>
<td>387 GB SFF-3 SSD 4k eMLC4 for Linux</td>
<td>18</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL8O</td>
<td>775 GB SFF-3 SSD 4k eMLC4 for Linux</td>
<td>18</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL8V</td>
<td>1.55 TB SFF-3 SSD 4k eMLC4 for Linux</td>
<td>18</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL1P</td>
<td>300 GB 15 K RPM SAS SFF-2 Disk Drive (AIX/Linux)</td>
<td>672</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL1Q</td>
<td>600 GB 10 K RPM SAS SFF-2 Disk Drive (AIX/Linux)</td>
<td>672</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>ELD3</td>
<td>1.2 TB 10 K RPM SAS SFF-2 Disk Drive (Linux)</td>
<td>672</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>ELD5</td>
<td>59D0</td>
<td>600 GB 10 K RPM SAS SFF-3 Disk Drive (Linux)</td>
<td>18</td>
<td>Linux</td>
</tr>
<tr>
<td>ELD9</td>
<td>59D8</td>
<td>1.2 TB 10 K RPM SAS SFF-3 Disk Drive (Linux)</td>
<td>18</td>
<td>Linux</td>
</tr>
<tr>
<td>ELDB</td>
<td>59E0</td>
<td>300 GB 15 K RPM SAS SFF-3 Disk Drive (Linux)</td>
<td>18</td>
<td>Linux</td>
</tr>
<tr>
<td>ELDF</td>
<td>600 GB 15 K RPM SAS SFF-3 Disk Drive - 5xx Block</td>
<td>18</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>ELDP</td>
<td>600 GB 15 K RPM SAS SFF-3 Disk Drive - 5xx Block</td>
<td>672</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>ELEV</td>
<td>600 GB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4096</td>
<td>672</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>ELEZ</td>
<td>59C9</td>
<td>300 GB 15 K RPM SAS SFF-2 4 K Block - 4096 Disk Drive</td>
<td>672</td>
<td>Linux</td>
</tr>
<tr>
<td>ELF3</td>
<td>1.2 TB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4096</td>
<td>672</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>ELF5</td>
<td>600 GB 10 K RPM SAS SFF-3 Disk Drive 4 K Block - 4096</td>
<td>18</td>
<td>Linux</td>
<td></td>
</tr>
</tbody>
</table>
A slimline media bay is included in the #EJ0N backplanes that can optionally house a Serial Advanced Technology Attachment (SATA) DVD-RAM (#5771). The DVD drive is run by the integrated SAS controllers, and a separate PCIe adapter is not required.

The Power S824L supports the RDX USB External Docking Station for Removable Disk Cartridge (#EUA4). The USB External Docking Station accommodates RDX removable disk cartridge of any capacity. The disks are in a protective rugged cartridge enclosure that plug into the docking station. The docking station holds one removable rugged disk drive/cartridge at a time. The rugged removable disk cartridge and docking station backs up similar to tape drive. This can be an excellent alternative to DAT72, DAT160, 8 mm, and VXA-2 and VXA-320 tapes.

Table 1-8 shows the available media device feature codes for the Power S824L server.

Table 1-8 Media device feature code descriptions for Power S8124L

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>Max</th>
<th>OS support</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELF9</td>
<td></td>
<td>1.2 TB 10 K RPM SAS SFF-3 Disk Drive 4 K Block - 4096</td>
<td>18</td>
<td>Linux</td>
</tr>
<tr>
<td>ELFB</td>
<td>59E1</td>
<td>300 GB 15 K RPM SAS SFF-3 4 K Block - 4096 Disk Drive</td>
<td>18</td>
<td>Linux</td>
</tr>
<tr>
<td>ELFF</td>
<td>59E5</td>
<td>600 GB 15 K RPM SAS SFF-3 4 K Block - 4096 Disk Drive</td>
<td>18</td>
<td>Linux</td>
</tr>
<tr>
<td>ELFP</td>
<td>59CC</td>
<td>600 GB 15 K RPM SAS SFF-2 4 K Block - 4096 Disk Drive</td>
<td>672</td>
<td>Linux</td>
</tr>
<tr>
<td>ELFT</td>
<td>1.8 TB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4096</td>
<td>672</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>ELFV</td>
<td>1.8 TB 10 K RPM SAS SFF-3 Disk Drive 4 K Block - 4096</td>
<td>18</td>
<td>Linux</td>
<td></td>
</tr>
</tbody>
</table>

SCSI disks are not supported in the Power S824L disk bays. Also, because there is no PCI SCSI adapter available, you cannot attach existing SCSI disk subsystems.

For more information about the internal disk features, see 2.6, “Internal storage” on page 53.

1.8 I/O drawers for Power S824L server

If additional Gen3 PCIe slots beyond the system node slots are required, a maximum of two PCIe Gen3 I/O drawers can be attached to the Power S824L server if no NVIDIA GPU is installed.

Note: The Power S824L with a NVIDIA GPU configuration does not support the attachment of external I/O drawers.
The disk-only I/O drawer (#5887) is not supported on the Power S824L server. Similarly, the GX++ attached EXP30 Ultra SSD Drawer (#EDR1 or #5888) is not supported. Also, the 3.5-inch-based feature 5886 EXP12S SAS Disk Drawer and feature 5786 EXP24 SCSI Disk Drawer are not supported.

1.8.1 PCIe Gen3 I/O expansion drawer

The 19-inch 4 EIA (4U) PCIe Gen3 I/O expansion drawer (#ELMX) and two PCIe Fan Out Modules (#ELMF) provide twelve PCIe I/O full-length, full-height slots. One Fan Out Module provides six PCIe slots labeled C1 through C6. C1 and C4 are x16 slots and C2, C3, C5, and C6 are x8 slots. PCIe Gen1, Gen2, and Gen3 full-high adapter cards are supported.

A blind swap cassette (BSC) is used to house the full-high adapters that go into these slots. The BSC is the same BSC as used with the previous generation server's 12X attached I/O drawers (#5802, #5803, #5877, #5873). The drawer is shipped with a full set of BSC, even if the BSC is empty.

Concurrent repair and add/removal of PCIe adapter cards is done by HMC guided menus or by operating system support utilities.

A PCIe3 Optical Cable Adapter for PCIe3 Expansion Drawer (#EJ08) and 3.0 m (#ECC7) or 10.0 m (#ECC8) CXP 16X Active Optical cables (AOC) connect the system node to a PCIe Fan Out module in the I/O expansion drawer. One feature #ECC7 or one #ECC8 ships two AOC cables. Each PCIe Gen3 I/O expansion drawer has two power supplies.

A maximum of one PCIe Gen3 I/O expansion drawer is supported on the one socket Power S824L with no GPU configuration.

A maximum of two PCIe Gen3 I/O expansion drawers is supported on the two socket Power S824L with no GPU configuration.

Figure 1-4 shows a PCIe Gen3 I/O expansion drawer.
1.8.2 I/O drawers and usable PCI slot

Figure 1-5 shows the rear view of the PCIe Gen3 I/O expansion drawer equipped with two PCIe3 6-slot Fan Out modules with the location codes for the PCIe adapter slots.

![Rear view of a PCIe Gen3 I/O expansion drawer with PCIe slots location codes](image)

Table 1-9 provides details of the PCI slots in the PCIe Gen3 I/O expansion drawer equipped with two PCIe3 6-slot Fan Out modules.

Table 1-9  PCIe slot locations for the PCIe Gen3 I/O expansion drawer with two Fan Out modules

<table>
<thead>
<tr>
<th>Slot</th>
<th>Location code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slot 1</td>
<td>P1-C1</td>
<td>PCIe3, x16</td>
</tr>
<tr>
<td>Slot 2</td>
<td>P1-C2</td>
<td>PCIe3, x8</td>
</tr>
<tr>
<td>Slot 3</td>
<td>P1-C3</td>
<td>PCIe3, x8</td>
</tr>
<tr>
<td>Slot 4</td>
<td>P1-C4</td>
<td>PCIe3, x16</td>
</tr>
<tr>
<td>Slot 5</td>
<td>P1-C5</td>
<td>PCIe3, x8</td>
</tr>
<tr>
<td>Slot 6</td>
<td>P1-C6</td>
<td>PCIe3, x8</td>
</tr>
<tr>
<td>Slot 7</td>
<td>P2-C1</td>
<td>PCIe3, x16</td>
</tr>
<tr>
<td>Slot 8</td>
<td>P2-C2</td>
<td>PCIe3, x8</td>
</tr>
<tr>
<td>Slot 9</td>
<td>P2-C3</td>
<td>PCIe3, x8</td>
</tr>
<tr>
<td>Slot 10</td>
<td>P2-C4</td>
<td>PCIe3, x16</td>
</tr>
<tr>
<td>Slot 11</td>
<td>P2-C5</td>
<td>PCIe3, x8</td>
</tr>
<tr>
<td>Slot 12</td>
<td>P2-C6</td>
<td>PCIe3, x8</td>
</tr>
</tbody>
</table>
- All slots support full-length, regular-height adapter or short (low-profile) with a regular-height tailstock in single-wide, Gen3, blind-swap cassettes.
- Slots C1 and C4 in each PCIe3 6-slot Fan Out module are x16 PCIe3 buses and slots C2, C3, C5, and C6 are x8 PCIe buses.
- All slots support enhanced error handling (EEH).
- All PCIe slots are hot swappable and support concurrent maintenance.

Table 1-10 summarizes the maximum number of I/O drawers supported and the total number of PCI slots that are available.

<table>
<thead>
<tr>
<th>System</th>
<th>Maximum #EMX0 drawer</th>
<th>Total number of slots</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCIe3, x16</td>
<td>PCIe3, x8</td>
</tr>
<tr>
<td>Power S824L (1-socket)</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Power S824L (2-sockets)</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

### 1.8.3 EXP24S SFF Gen2-bay drawer

If you need more disks than are available with the internal disk bays, you can attach additional external disk subsystems, such as the EXP24S SAS HDD/SSD Expansion Drawer (#EL1S). The EXP24S SFF Gen2-bay drawer is an expansion drawer supporting up to twenty-four 2.5-inch hot-swap SFF SAS HDDs on IBM POWER6®, IBM POWER6+™, POWER7®, POWER7+™, or POWER8 servers in 2U of 19-inch rack space. The EXP24S bays are controlled by SAS adapters or controllers that are attached to the I/O drawer by SAS X or Y cables.

The EXP24S drawer is attached to SAS ports on either a PCIe SAS adapter in the server or to the SAS ports at the rear of the server. Two SAS ports at the rear of the server are enabled with the expanded-function storage backplane with dual IOA support (#EL3U).

A maximum of 28 EXP24S drawers are supported on the Power S824L server.

The SFF bays of the EXP24S differ from the SFF bays of the POWER8 system units. The EXP24S uses Gen2 or SFF-2 SAS drives that physically do not fit in the SFF-3 bays of the POWER8 system unit.

The EXP24S includes redundant AC power supplies and two power cords.

**Note:** Any existing #5887 EXP24S SFF Gen2-bay drawer can be attached to the Power S824L server.
1.9 Server and virtualization management

The Power S824L without a NVIDIA GPU supports PowerVM virtualization.

*Note:* PowerVM, and KVM are not supported on the Power S824L if ordered with a NVIDIA GPU.

If you want to implement partitions, a Hardware Management Console (HMC) or the Integrated Virtualization Manager (IVM) is required to manage the Power S824L server. In general, multiple IBM POWER6, POWER6+, POWER7, POWER7+, and POWER8 processor-based servers can be supported by a single HMC.

If an HMC is used to manage the Power S824L, the HMC must be a rack-mount CR5 or later, or deskside C08 or later.

*Remember:* If you do not use an HMC or IVM, the Power S824L runs in bare-metal mode, which means that a single partition owns all the server resources, and only Red Hat Linux 7 (le) or Ubuntu Linux operating systems can be installed.

In April 2015, IBM announced a new HMC model, machine type 7042-CR9 Hardware features on the CR9 model include a second disk drive (#1998) for RAID 1 data mirroring, and the option of a redundant power supply. If you prefer not to have RAID 1 enabled on the HMC, you can override it in the ordering system and remove the additional HDD from the order. RAID 1 is also offered on the 7042-CR6, 7042-CR7, 7042-CR8, and 7042-CR9 models as a miscellaneous equipment specification (MES) upgrade option.

Starting with HMC V8R8.1.0 code the HMC can manage more LPARs per processor core. A core can be partitioned in up to 20 LPARs (0.05 of a core).

Several HMC models are supported to manage POWER8 processor-based systems. The 7042-CR9 is the only HMCs that are available for ordering at the time of writing, but you can also use one of the withdrawn models that are listed in Table 1-11.

<table>
<thead>
<tr>
<th>Type-model</th>
<th>Availability</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7042-C08</td>
<td>Withdrawn</td>
<td>IBM 7042 Model C08 Deskside Hardware Management Console</td>
</tr>
<tr>
<td>7042-CR5</td>
<td>Withdrawn</td>
<td>IBM 7042 Model CR5 Rack-Mounted Hardware Management Console</td>
</tr>
</tbody>
</table>
At the time of writing the IBM POWER8 processor-based Power S822 server requires HMC V8R8.3.0.

Tip: You can download or order the latest HMC code from the Fix Central website:

http://www.ibm.com/support/fixcentral

If you are attaching an HMC to a new server or adding a function to an existing server that requires a firmware update, the HMC machine code might need to be updated because HMC code must always be equal to or higher than the managed server’s firmware. Access to firmware and machine code updates is conditional on entitlement and license validation in accordance with IBM policy and practice. IBM might verify entitlement through customer number, serial number, electronic restrictions, or any other means or methods that are employed by IBM at its discretion.

1.10 System racks

The Power S824L systems are designed to mount in the 36U 7014-T00 (#0551), the 42U 7014-T42 (#0553), or the IBM 42U Slim Rack (7965-94Y) rack. These racks are built to the 19-inch EIA 310D standard.

**Order information:** The racking approach for the initial order must be either a 7014-T00, 7014-T42, or 7965-94Y. If an additional rack is required for I/O expansion drawers as a miscellaneous equipment specification (MES) to an existing system, either a feature 0551, 0553, or ER05 rack must be ordered.

If a system is to be installed in a rack or cabinet that is not IBM, ensure that the rack meets the requirements that are described in 1.10.10, “OEM rack” on page 24.

**Responsibility:** The client is responsible for ensuring that the installation of the drawer in the preferred rack or cabinet results in a configuration that is stable, serviceable, safe, and compatible with the drawer requirements for power, cooling, cable management, weight, and rail security.

1.10.1 IBM 7014 Model T00 rack

The 1.8-meter (71-inch) model T00 is compatible with past and present IBM Power Systems servers. The T00 rack offers these features:

- 36U (EIA units) of usable space.
- Optional removable side panels.
- Optional side-to-side mounting hardware for joining multiple racks.
Increased power distribution and weight capacity.

Support for both AC and DC configurations.

Up to four power distribution units (PDUs) can be mounted in the PDU bays (see Figure 1-8 on page 22), but others can fit inside the rack. For more information, see 1.10.7, “AC power distribution unit and rack content” on page 22.

For the T00 rack, three door options are available:

- Front Door for 1.8 m Rack (#6068)
  
  This feature provides an attractive black full height rack door. The door is steel with a perforated flat front surface. The perforation pattern extends from the bottom to the top of the door to enhance ventilation and provide visibility into the rack.

- A 1.8 m Rack Acoustic Door (#6248)
  
  This feature provides a front and rear rack door that are designed to reduce acoustic sound levels in a general business environment.

- A 1.8 m Rack Trim Kit (#6263)
  
  If no front door is used in the rack, this feature provides a decorative trim kit for the front.

Ruggedized Rack Feature

For enhanced rigidity and stability of the rack, the optional Ruggedized Rack Feature (#6080) provides additional hardware that reinforces the rack and anchors it to the floor. This hardware is designed primarily for use in locations where earthquakes are a concern. The feature includes a large steel brace or truss that bolts into the rear of the rack.

It is hinged on the left side so that it can swing out of the way for easy access to the rack drawers when necessary. The Ruggedized Rack Feature also includes hardware for bolting the rack to a concrete floor or a similar surface, and bolt-in steel filler panels for any unoccupied spaces in the rack.

The following weights apply to the T00 rack:

- T00 base empty rack: 244 kg (535 lb.).
- T00 full rack: 816 kg (1795 lb.).
- Maximum weight of drawers is 572 kg (1260 lb.).
- Maximum weight of drawers in a zone 4 earthquake environment is 490 kg (1080 lb.).

This number equates to 13.6 kg (30 lb.) per EIA.

**Important:** If additional weight is added to the top of the rack, for example, by adding #6117, the 490 kg (1080 lb.) weight must be reduced by the weight of the addition. As an example, #6117 weighs approximately 45 kg (100 lb.) so the new maximum weight of the drawers that the rack can support in a zone 4 earthquake environment is 445 kg (980 lb.). In the zone 4 earthquake environment, the rack must be configured starting with the heavier drawers at the bottom of the rack.

### 1.10.2 IBM 7014 Model T42 rack

The 2.0-meter (79.3-inch) Model T42 addresses the client requirement for a tall enclosure to house the maximum amount of equipment in the smallest possible floor space. The following features are for the model T42 rack (which differ from the model T00):

- The T42 rack has 42U (EIA units) of usable space (6U of additional space).
- The model T42 supports AC power only.
The following weights apply to the T42 rack:
- T42 base empty rack: 261 kg (575 lb.)
- T42 full rack: 930 kg (2045 lb.)

The available door options for the Model T42 rack are shown in Figure 1-7.

<table>
<thead>
<tr>
<th>Trim kit (no front door)</th>
<th>Plain front door</th>
<th>Acoustic doors (front and rear)</th>
<th>780 logo front door</th>
<th>Optional front door</th>
</tr>
</thead>
<tbody>
<tr>
<td>FC 6272</td>
<td>FC 6069</td>
<td>FC 6249</td>
<td>FC 6250</td>
<td>FC ERG7</td>
</tr>
</tbody>
</table>

The 2.0 m Rack Trim Kit (#6272) is used if no front door is used in the rack.

The Front Door for a 2.0 m Rack (#6069) is made of steel with a perforated flat front surface. The perforation pattern extends from the bottom to the top of the door to enhance ventilation and provide visibility into the rack. This door is non-acoustic and has a depth of about 25 mm (1 in.).

The 2.0 m Rack Acoustic Door (#6249) consists of a front and rear door to reduce noise by approximately 6 dB(A). It has a depth of approximately 191 mm (7.5 in.).

The High-End Appearance Front Door (#6250) provides a front rack door with a field-installed Power 780 logo indicating that the rack contains a Power 780 system. The door is not acoustic and has a depth of about 90 mm (3.5 in.).

**High end**: For the High-End Appearance Front Door (#6250), use the High-End Appearance Side Covers (#6238) to make the rack appear as though it is a high-end server (but in a 19-inch rack format instead of a 24-inch rack).

The feature #ERG7 provides an attractive black full height rack door. The door is steel with a perforated flat front surface. The perforation pattern extends from the bottom to the top of the door to enhance ventilation and provide visibility into the rack. The non-acoustic door has a depth of about 134 mm (5.3 in.).

**Rear Door Heat Exchanger**
To lead away more heat, a special door that is named the Rear Door Heat Exchanger (#EC15) is available. This door replaces the standard rear door on the rack. Copper tubes that are attached to the rear door circulate chilled water, which is provided by the client. The chilled water removes heat from the exhaust air being blown through the servers and attachments that are mounted in the rack. With industry standard quick couplings, the water lines in the door attach to the client-supplied secondary water loop.
For details about planning for the installation of the IBM Rear Door Heat Exchanger, see the following website:


1.10.3 IBM 42U Slim Rack 7965-94Y

The 2.0-meter (79-inch) model 7965-94Y is compatible with past and present IBM Power Systems servers and provides an excellent 19-inch rack enclosure for your data center. Its 600 mm (23.6 in.) width combined with its 1100 mm (43.3 in.) depth plus its 42 EIA enclosure capacity provides great footprint efficiency for your systems and allows it to be easily located on standard 24-inch floor tiles.

The IBM 42U Slim Rack has a lockable perforated front steel door that provides ventilation, physical security, and visibility of indicator lights in the installed equipment within. In the rear, either a lockable perforated rear steel door (#EC02) or a lockable Rear Door Heat Exchanger (RDHX)(1164-95X) is used. Lockable optional side panels (#EC03) increase the rack's aesthetics, help control airflow through the rack, and provide physical security. Multiple 42U Slim Racks can be bolted together to create a rack suite (indicate feature code #EC04).

Up to six optional 1U PDUs can be placed vertically in the sides of the rack. Additional PDUs can be located horizontally, but they each use 1U of space in this position.

1.10.4 Feature code 0551 rack

The 1.8 Meter Rack (#0551) is a 36 EIA unit rack. The rack that is delivered as #0551 is the same rack that is delivered when you order the 7014-T00 rack. The included features might vary. Certain features that are delivered as part of the 7014-T00 must be ordered separately with the #0551.

1.10.5 Feature code 0553 rack

The 2.0 Meter Rack (#0553) is a 42 EIA unit rack. The rack that is delivered as #0553 is the same rack that is delivered when you order the 7014-T42 rack. The included features might vary. Certain features that are delivered as part of the 7014-T42 must be ordered separately with the #0553.

1.10.6 Feature code ER05 rack

This feature provides a 19-inch, 2.0-meter high rack with 42 EIA units of total space for installing rack-mounted Central Electronics Complexes or expansion units. The 600 mm wide rack fits within a data center's 24-inch floor tiles and provides better thermal and cable management capabilities. The following features are required on #ER05:

- #EC01 front door
- #EC02 rear door or #EC05 Rear Door Heat Exchanger (RDHX) indicator

PDUs on the rack are optional. Each #7196 and #7189 PDU consumes one of six vertical mounting bays. Each PDU beyond four consumes 1U of rack space.

If you order Power Systems equipment in an MES order, use the equivalent rack feature ER05 instead of 7965-94Y so that IBM Manufacturing can ship the hardware in the rack.
1.10.7 AC power distribution unit and rack content

For rack models T00 and T42, 12-outlet PDUs are available. These PDUs include the AC power distribution unit #7188 and the AC Intelligent PDU+ #7109. The Intelligent PDU+ is identical to #7188 PDUs, but it is equipped with one Ethernet port, one console serial port, and one RS232 serial port for power monitoring.

The PDUs have 12 client-usable IEC 320-C13 outlets. Six groups of two outlets are fed by six circuit breakers. Each outlet is rated up to 10 amps, but each group of two outlets is fed from one 15 amp circuit breaker.

Four PDUs can be mounted vertically in the back of the T00 and T42 racks. Figure 1-8 shows the placement of the four vertically mounted PDUs. In the rear of the rack, two additional PDUs can be installed horizontally in the T00 rack and three in the T42 rack. The four vertical mounting locations are filled first in the T00 and T42 racks. Mounting PDUs horizontally consumes 1U per PDU and reduces the space available for other racked components. When mounting PDUs horizontally, the best approach is to use fillers in the EIA units that are occupied by these PDUs to facilitate the correct airflow and ventilation in the rack.

Figure 1-8  PDU placement and PDU view
The PDU receives power through a UTG0247 power-line connector. Each PDU requires one PDU-to-wall power cord. Various power cord features are available for various countries and applications by varying the PDU-to-wall power cord, which must be ordered separately. Each power cord provides the unique design characteristics for the specific power requirements. To match new power requirements and save previous investments, these power cords can be requested with an initial order of the rack or with a later upgrade of the rack features.

Table 1-12 shows the available wall power cord options for the PDU and iPDU features, which must be ordered separately.

Table 1-12  Wall power cord options for the PDU and iPDU features

<table>
<thead>
<tr>
<th>Feature code</th>
<th>Wall plug</th>
<th>Rated voltage (Vac)</th>
<th>Phase</th>
<th>Rated amperage</th>
<th>Geography</th>
</tr>
</thead>
<tbody>
<tr>
<td>6653</td>
<td>IEC 309, 3P+N+G, 16A</td>
<td>230</td>
<td>3</td>
<td>16 amps/phase</td>
<td>Internationally available</td>
</tr>
<tr>
<td>6489</td>
<td>IEC309 3P+N+G, 32A</td>
<td>230</td>
<td>3</td>
<td>32 amps/phase</td>
<td>EMEA</td>
</tr>
<tr>
<td>6654</td>
<td>NEMA L6-30</td>
<td>200 - 208, 240</td>
<td>1</td>
<td>24 amps</td>
<td>US, Canada, LA, and Japan</td>
</tr>
<tr>
<td>6655</td>
<td>RS 3750DP (watertight)</td>
<td>200 - 208, 240</td>
<td>1</td>
<td>24 amps</td>
<td>US, Canada, LA, and Japan</td>
</tr>
<tr>
<td>6656</td>
<td>IEC 309, P+N+G, 32A</td>
<td>230</td>
<td>1</td>
<td>24 amps</td>
<td>EMEA</td>
</tr>
<tr>
<td>6657</td>
<td>PDL</td>
<td>230 - 240</td>
<td>1</td>
<td>32 amps</td>
<td>Australia and New Zealand</td>
</tr>
<tr>
<td>6658</td>
<td>Korean plug</td>
<td>220</td>
<td>1</td>
<td>30 amps</td>
<td>North and South Korea</td>
</tr>
<tr>
<td>6492</td>
<td>IEC 309, 2P+G, 60A</td>
<td>200 - 208, 240</td>
<td>1</td>
<td>48 amps</td>
<td>US, Canada, LA, and Japan</td>
</tr>
<tr>
<td>6491</td>
<td>IEC 309, P+N+G, 63A</td>
<td>230</td>
<td>1</td>
<td>63 amps</td>
<td>EMEA</td>
</tr>
</tbody>
</table>

**Notes:** Ensure that the correct power cord feature is configured to support the power that is being supplied. Based on the power cord that is used, the PDU can supply 4.8 - 19.2 kVA. The power of all of the drawers that are plugged into the PDU must not exceed the power cord limitation.

The Universal PDUs are compatible with previous models.

To better enable electrical redundancy, each server has two power supplies that must be connected to separate PDUs, which are not included in the base order.

For maximum availability, a preferred approach is to connect power cords from the same system to two separate PDUs in the rack, and to connect each PDU to independent power sources.

For detailed power requirements and power cord details about the 7014 racks, see the “Planning for power” section in the IBM Power Systems Hardware Knowledge Center website: [http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/topic/p7had/p7hadrpower.ht](http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/topic/p7had/p7hadrpower.ht)
For detailed power requirements and power cord details about the 7965-94Y rack, see the “Planning for power” section in the IBM Power Systems Hardware Knowledge Center website:

http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/topic/p7had/p7hadkickoff795394x.htm

1.10.8 Rack-mounting rules

Consider the following primary rules when you mount the system into a rack:

- The system is designed to be placed at any location in the rack. For rack stability, start filling a rack from the bottom.
- Any remaining space in the rack can be used to install other systems or peripheral devices, if the maximum permissible weight of the rack is not exceeded and the installation rules for these devices are followed.
- Before placing the system into the service position, be sure to follow the rack manufacturer’s safety instructions regarding rack stability.

1.10.9 Useful rack additions

This section highlights several rack addition solutions for IBM Power Systems rack-based systems.

1.10.10 OEM rack

The system can be installed in a suitable OEM rack if that the rack conforms to the EIA-310-D standard for 19-inch racks. This standard is published by the Electrical Industries Alliance. For more information, see the IBM Power Systems Hardware Knowledge Center at the following website:

http://publib.boulder.ibm.com/infocenter/systems/scope/hw/index.jsp

The website mentions the following key points:

- The front rack opening must be 451 mm wide ± 0.75 mm (17.75 in. ± 0.03 in.), and the rail-mounting holes must be 465 mm ± 0.8 mm (18.3 in. ± 0.03 in.) apart on-center (horizontal width between the vertical columns of holes on the two front-mounting flanges and on the two rear-mounting flanges). Figure 1-9 on page 25 is a top view that shows the specification dimensions.
Figure 1-9  Top view of rack specification dimensions (not specific to IBM)
The vertical distance between the mounting holes must consist of sets of three holes spaced (from bottom to top) 15.9 mm (0.625 in.), 15.9 mm (0.625 in.), and 12.67 mm (0.5 in.) on-center, which makes each three-hole set of vertical hole spacing 44.45 mm (1.75 in.) apart on center. Rail-mounting holes must be 7.1 mm ± 0.1 mm (0.28 in. ± 0.004 in.) in diameter. Figure 1-10 shows the top front specification dimensions.

Figure 1-10  Rack specification dimensions top front view
Architecture and technical overview

This chapter describes the overall system architecture for the IBM Power System S824L (8247-42L) servers. The bandwidths that are provided throughout the chapter are theoretical maximums that are used for reference.

The speeds that are shown are at an individual component level. Multiple components and application implementation are key to achieving the best performance.

Always size the performance at the application workload environment level and evaluate performance by using real-world performance measurements and production workloads.
2.1 The IBM POWER8 processor

This section introduces the latest processor in the IBM Power Systems product family and describes its major characteristics and features in general.

2.1.1 POWER8 processor overview

The POWER8 processor is manufactured by using the IBM 22 nm Silicon-On-Insulator (SOI) technology. Each chip is 649 mm² and contains 4.2 billion transistors. As shown in Figure 2-2 on page 29, the chip contains 12 cores, two memory controllers, Peripheral Component Interconnect Express (PCIe) Gen3 I/O controllers, and an interconnection system that connects all components within the chip. Each core has 512 KB of L2 cache, and all cores share 96 MB of L3 embedded dynamic random access memory (eDRAM). The interconnect also extends through module and board technology to other POWER8 processors in addition to DDR3 memory and various I/O devices.
POWER8 systems use memory buffer chips to interface between the POWER8 processor and DDR3 or DDR4 memory. Each buffer chip also includes an L4 cache to reduce the latency of local memory accesses.

Figure 2-2  The POWER8 processor chip

The POWER8 processor is for system offerings from single-socket servers to multi-socket enterprise servers. It incorporates a triple-scope broadcast coherence protocol over local and global SMP links to provide superior scaling attributes. Multiple-scope coherence protocols reduce the amount of symmetric multiprocessor (SMP) link bandwidth that is required by attempting operations on a limited scope (single chip or multi-chip group) when possible. If the operation cannot complete coherently, the operation is reissued by using a larger scope to complete the operation.

The following features can augment the performance of the POWER8 processor:

- Support for DDR3 and DDR4 memory through memory buffer chips that offload the memory support from the POWER8 memory controller.
- L4 cache within the memory buffer chip that reduces the memory latency for local access to memory behind the buffer chip; the operation of the L4 cache is not apparent to applications running on the POWER8 processor. Up to 128 MB of L4 cache can be available for each POWER8 processor.
- Hardware transactional memory.
- On-chip accelerators, including on-chip encryption, compression, and random number generation accelerators.
- Coherent Accelerator Processor Interface (CAPI), which allows accelerators plugged into a PCIe slot to access the processor bus using a low latency, high-speed protocol interface.

1 At the time of writing, the available POWER8 processor-based systems use DDR3 memory.
Adaptive power management.

Two versions of the POWER8 processor chip are available. Both chips use the same building blocks. The scale-out systems use a 6-core version of POWER8. The 6-core chip is installed in pairs in a dual-chip module (DCM) that plugs into a socket in the system board of the servers. Functionally, it works as a single chip.

Figure 2-3 shows a graphical representation of the 6-core processor.

![Figure 2-3 Six-core POWER8 processor chip](image)

Table 2-1 summarizes the technology characteristics of the POWER8 processor.

<table>
<thead>
<tr>
<th>Technology</th>
<th>POWER8 processor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die size</td>
<td>649 mm²</td>
</tr>
<tr>
<td>Fabrication technology</td>
<td>&gt; 22 nm lithography</td>
</tr>
<tr>
<td></td>
<td>&gt; Copper interconnect</td>
</tr>
<tr>
<td></td>
<td>&gt; Silicon-On-Insulator (SOI)</td>
</tr>
<tr>
<td></td>
<td>&gt; eDRAM</td>
</tr>
<tr>
<td>Maximum processor cores</td>
<td>Six or 12</td>
</tr>
<tr>
<td>Maximum execution threads core/chip</td>
<td>8/96</td>
</tr>
<tr>
<td>Maximum L2 cache core/chip</td>
<td>512 KB/6 MB</td>
</tr>
<tr>
<td>Maximum On-chip L3 cache core/chip</td>
<td>8 MB/96 MB</td>
</tr>
<tr>
<td>Maximum L4 cache per chip</td>
<td>128 MB</td>
</tr>
<tr>
<td>Maximum memory controllers</td>
<td>Two</td>
</tr>
<tr>
<td>SMP design-point</td>
<td>Sixteen sockets with IBM POWER8 processors</td>
</tr>
<tr>
<td>Compatibility</td>
<td>With prior generation of POWER® processor</td>
</tr>
</tbody>
</table>
2.1.2 POWER8 processor core

The POWER8 processor core is a 64-bit implementation of the IBM Power Instruction Set Architecture (ISA) Version 2.07 and offers the following features:

- Multi-threaded design that is capable of up to eight-way simultaneous multithreading (SMT)
- 32 KB, eight-way set-associative L1 instruction cache
- 64 KB, eight-way set-associative L1 data cache
- Enhanced prefetch, with instruction speculation awareness and data prefetch depth awareness
- Enhanced branch prediction, using both local and global prediction tables with a selector table to choose the best predictor
- Improved out-of-order execution
- Two symmetric fixed-point execution units
- Two symmetric load/store units and two load units, all four of which can also run simple fixed-point instructions
- An integrated, multi-pipeline vector-scalar floating point unit for running both scalar and single-instruction, multiple-data (SIMD) type instructions, including the Vector Multimedia eXtension (VMX) instruction set and the improved Vector Scalar eXtension (VSX) instruction set, which is capable of up to sixteen floating point operations per cycle (eight double precision or sixteen single precision)
- In-core Advanced Encryption Standard (AES) encryption capability
- Hardware data prefetching with 16 independent data streams and software control
- Hardware decimal floating point (DFP) capability

For more information about Power ISA Version 2.07, see the following website:
https://www.power.org/documentation/power-isa-version-2-07/

Figure 2-4 shows a picture of the POWER8 core with several functional units highlighted.
2.1.3 Simultaneous multithreading

POWER8 processor advancements in multi-core and multi-thread scaling are remarkable. A significant performance opportunity comes from parallelizing workloads to enable the full potential of the microprocessor and the large memory bandwidth. Application scaling is influenced by both multi-core and multi-thread technology.

SMT allows a single physical processor core to simultaneously dispatch instructions from more than one hardware thread context. With SMT, each POWER8 core can present eight hardware threads. Because there are multiple hardware threads per physical processor core, additional instructions can run at the same time. SMT is primarily beneficial in commercial environments where the speed of an individual transaction is not as critical as the total number of transactions that are performed. SMT typically increases the throughput of workloads with large or frequently changing working sets, such as database servers and web servers.

Table 2-2 shows a comparison between the different POWER processors in terms of SMT capabilities that are supported by each processor architecture.

Table 2-2  SMT levels that are supported by POWER processors

<table>
<thead>
<tr>
<th>Technology</th>
<th>Cores/system</th>
<th>Maximum SMT mode</th>
<th>Maximum hardware threads per system</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM POWER4</td>
<td>32</td>
<td>Single Thread (ST)</td>
<td>32</td>
</tr>
<tr>
<td>IBM POWER5</td>
<td>64</td>
<td>SMT2</td>
<td>128</td>
</tr>
<tr>
<td>IBM POWER6</td>
<td>64</td>
<td>SMT2</td>
<td>128</td>
</tr>
<tr>
<td>IBM POWER7</td>
<td>256</td>
<td>SMT4</td>
<td>1024</td>
</tr>
<tr>
<td>IBM POWER8</td>
<td>192</td>
<td>SMT8</td>
<td>1536</td>
</tr>
</tbody>
</table>

The architecture of the POWER8 processor, with its larger caches, larger cache bandwidth, and faster memory, allows threads to have faster access to memory resources, which translates in a more efficient usage of threads. Because of that, POWER8 allows more threads per core to run concurrently, increasing the total throughput of the processor and of the system.

2.1.4 Memory access

On the Power S824L, each POWER8 module has two memory controllers, each connected to four memory channels. Each memory channel operates at 1600 MHz and connects to a dual inline memory module (DIMM). Each DIMM on a POWER8 system has a memory buffer that is responsible for many functions that were previously on the memory controller, such as scheduling logic and energy management. The memory buffer also has 16 MB of L4 cache.

At the time of writing, each memory channel can address up to 128 GB. Therefore, the Power S824L can address up to 2 TB of total memory.
Figure 2-5 gives a simple overview of the POWER8 processor memory access structure in the Power S824L.

2.1.5 On-chip L3 cache innovation and Intelligent Cache

Similar to POWER7 and POWER7+, the POWER8 processor uses a breakthrough in material engineering and microprocessor fabrication to implement the L3 cache in eDRAM and place it on the processor die. L3 cache is critical to a balanced design. The ability to provide good signaling between the L3 cache and other elements of the hierarchy, such as the L2 cache or SMP interconnect, is also critical to a balanced design.

The on-chip L3 cache is organized into separate areas with differing latency characteristics. Each processor core is associated with a fast 8 MB local region of L3 cache (FLR-L3) but also has access to other L3 cache regions as shared L3 cache. Additionally, each core can negotiate to use the FLR-L3 cache that is associated with another core, depending on reference patterns. Data can also be cloned to be stored in more than one core's FLR-L3 cache, again depending on reference patterns. This Intelligent Cache management enables the POWER8 processor to optimize the access to L3 cache lines and minimize overall cache latencies.

Figure 2-2 on page 29 and Figure 2-3 on page 30 show the on-chip L3 cache, and highlight the fast 8 MB L3 region that is closest to a processor core.

The innovation of using eDRAM on the POWER8 processor die is significant for several reasons:

- **Latency improvement**
  A six-to-one latency improvement occurs by moving the L3 cache on-chip compared to L3 accesses on an external (on-ceramic) Application Specific Integrated Circuit (ASIC).

- **Bandwidth improvement**
  A 2x bandwidth improvement occurs with on-chip interconnect. Frequency and bus sizes are increased to and from each core.

- **No off-chip driver or receivers**
  Removing drivers or receivers from the L3 access path lowers interface requirements, conserves energy, and lowers latency.
- Small physical footprint
  The performance of eDRAM when implemented on-chip is similar to conventional static random-access memory (SRAM) but requires far less physical space. IBM on-chip eDRAM uses only a third of the components that conventional SRAM uses (a minimum of six transistors to implement a 1-bit memory cell).

- Low energy consumption
  The on-chip eDRAM uses only 20% of the standby power of SRAM.

2.1.6 L4 cache and memory buffer

POWER8 processor-based systems introduce an additional level in memory hierarchy. The L4 cache is implemented together with the memory buffer in the Custom DIMM (CDIMM). Each memory buffer contains 16 MB of L4 cache. On a Power S824L, you can have up to 256 MB of L4 cache.

Figure 2-6 shows a picture of the memory buffer, where you can see the 16 MB L4 cache, processor links, and memory interfaces.

Table 2-3 on page 35 shows a comparison of the different levels of cache in the POWER7, POWER7+, and POWER8 processors.
2.1.7 Hardware transactional memory

Transactional memory is an alternative to lock-based synchronization. It attempts to simplify parallel programming by grouping read and write operations and running them as a single operation. Transactional memory is similar to database transactions where all shared memory accesses and their effects are either committed all together or discarded as a group. All threads can enter the critical region simultaneously. If there are conflicts in accessing the shared memory data, threads try accessing the shared memory data again or are stopped without updating the shared memory data. Therefore, transactional memory is also called a lock-free synchronization. Transactional memory can be a competitive alternative to lock-based synchronization.

Transactional memory provides a programming model that makes parallel programming easier. A programmer delimits regions of code that access shared data and the hardware runs these regions atomically and in isolation, buffering the results of individual instructions, and trying execution again if isolation is violated. Generally, transactional memory allows programs to use a programming style that is close to coarse-grained locking to achieve performance that is close to fine-grained locking.

Most implementations of transactional memory are based on software. The POWER8 processor-based systems provide a hardware-based implementation of transactional memory that is more efficient than the software implementations and requires no interaction with the processor core, therefore allowing the system to operate at maximum performance.

### Table 2-3  POWER8 cache hierarchy

<table>
<thead>
<tr>
<th>Cache</th>
<th>POWER7</th>
<th>POWER7+</th>
<th>POWER8</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 instruction cache:</td>
<td>32 KB, 4-way</td>
<td>32 KB, 4-way</td>
<td>32 KB, 8-way</td>
</tr>
<tr>
<td>Capacity/associativity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bandwidth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1 data cache:</td>
<td>32 KB, 8-way</td>
<td>32 KB, 8-way</td>
<td>64 KB, 8-way</td>
</tr>
<tr>
<td>Capacity/associativity</td>
<td>Two 16 B reads or</td>
<td>Two 16 B reads or</td>
<td>Four 16 B reads or</td>
</tr>
<tr>
<td>bandwidth</td>
<td>one 16 B write per cycle</td>
<td>one 16 B write per cycle</td>
<td>one 16 B write per cycle</td>
</tr>
<tr>
<td>L2 cache:</td>
<td>256 KB, 8-way</td>
<td>256 KB, 8-way</td>
<td>512 KB, 8-way</td>
</tr>
<tr>
<td>Capacity/associativity</td>
<td>Private</td>
<td>Private</td>
<td>Private</td>
</tr>
<tr>
<td>bandwidth</td>
<td>32 B reads and 16 B writes</td>
<td>32 B reads and 16 B writes</td>
<td>64 B reads and 16 B writes</td>
</tr>
<tr>
<td></td>
<td>per cycle</td>
<td>per cycle</td>
<td>per cycle</td>
</tr>
<tr>
<td>L3 cache:</td>
<td>On-Chip</td>
<td>On-Chip</td>
<td>On-Chip</td>
</tr>
<tr>
<td>Capacity/associativity</td>
<td>4 MB/core, 8-way</td>
<td>10 MB/core, 8-way</td>
<td>8 MB/core, 8-way</td>
</tr>
<tr>
<td>bandwidth</td>
<td>16 B reads and 16 B writes</td>
<td>16 B reads and 16 B writes</td>
<td>32 B reads and 32 B writes</td>
</tr>
<tr>
<td></td>
<td>per cycle</td>
<td>per cycle</td>
<td>per cycle</td>
</tr>
<tr>
<td>L4 cache:</td>
<td>N/A</td>
<td>N/A</td>
<td>Off-Chip</td>
</tr>
<tr>
<td>Capacity/associativity</td>
<td></td>
<td></td>
<td>16 MB/buffer chip, 16-way</td>
</tr>
<tr>
<td>bandwidth</td>
<td></td>
<td></td>
<td>Up to 8 buffer chips per</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>socket</td>
</tr>
</tbody>
</table>

For more information about the POWER8 memory subsystem, see Table 2-4 on page 38.
2.1.8 Coherent Accelerator Processor Interface

The Coherent Accelerator Interface Architecture (CAIA) defines a coherent accelerator interface structure for attaching special processing devices to the POWER8 processor bus.

The CAPI can attach accelerators that have coherent shared memory access with the processors in the server and share full virtual address translation with these processors, using a standard PCIe Gen3 bus.

Applications can have customized functions in Field Programmable Gate Arrays (FPGA) and enqueue work requests directly in shared memory queues to the FPGA, and by using the same effective addresses (pointers) that it uses for any of its threads running on a host processor. From a practical perspective, CAPI allows a specialized hardware accelerator to be seen as an additional processor in the system, with access to the main system memory, and coherent communication with other processors in the system.

The benefits of using CAPI include the ability to access shared memory blocks directly from the accelerator, perform memory transfers directly between the accelerator and processor cache, and reduce the code path length between the adapter and the processors. This is possibly because the adapter is not operating as a traditional I/O device, and there is no device driver layer to perform processing. It also presents a simpler programming model.

Figure 2-7 shows a high-level view of how an accelerator communicates with the POWER8 processor through CAPI. The POWER8 processor provides a Coherent Attached Processor Proxy (CAPP), which is responsible for extending the coherence in the processor communications to an external device. The coherency protocol is tunneled over standard PCIe Gen3, effectively making the accelerator part of the coherency domain.

The accelerator adapter implements the Power Service Layer (PSL), which provides address translation and system memory cache for the accelerator functions. The custom processors on the board, consisting of an FPGA or an ASIC, use this layer to access shared memory regions, and cache areas as though they were a processor in the system. This ability enhances the performance of the data access for the device and simplifies the programming effort to use the device. Instead of treating the hardware accelerator as an I/O device, it is treated as a processor, which eliminates the requirement of a device driver to perform communication, and the need for Direct Memory Access that requires system calls to the operating system (OS) kernel. By removing these layers, the data transfer operation requires fewer clock cycles in the processor, improving the I/O performance.
The implementation of CAPI on the POWER8 processor allows hardware companies to develop solutions for specific application demands and use the performance of the POWER8 processor for general applications and the custom acceleration of specific functions using a hardware accelerator, with a simplified programming model and efficient communication with the processor and memory resources.

For a list of available CAPI adapters, see 2.5.12, “CAPI adapters” on page 53.

2.1.9 Power management and system performance

The POWER8 processor has power saving and performance enhancing features that can be used to lower overall energy usage while yielding higher performance when needed. The following modes can be enabled and modified to use these features.

**Dynamic Power Saver: Favor Performance**
This mode is intended to provide the best performance. If the processor is used even moderately, the frequency is raised to the maximum frequency possible to provide the best performance. If the processors are lightly used, the frequency is lowered to the minimum frequency, which is potentially far below the nominal shipped frequency, to save energy. The top frequency that is achieved is based on system type and it is affected by environmental conditions. Also, when running at the maximum frequency, more energy is consumed, which means that this mode can potentially cause an increase in overall energy consumption.

**Dynamic Power Saver: Favor Power**
This mode is intended to provide the best performance per watt consumed. The processor frequency is adjusted based on the processor usage to maintain the workload throughput without using more energy than required. At high processor usage levels, the frequency is raised above the nominal frequency, as in the Favor Performance mode. Likewise, at low processor usage levels, the frequency is lowered to the minimum frequency. The frequency ranges are the same for the two Dynamic Power Saver modes, but the algorithm that determines the frequency to set is different.

**Dynamic Power Saver: Tunable Parameters**
Dynamic Power Saver: Favor Performance and Dynamic Power Saver: Favor Power are tuned to provide both energy savings and performance increases. However, there might be situations where only top performance is of concern, or, conversely, where peak power consumption is an issue. Tunable Parameters can be used to modify the setting of the processor frequency in these modes to meet these various objectives. Modifying these parameters must be done only by advanced users. If issues must be addressed by the Tunable Parameters, directly involve IBM in the parameter value selection.

**Idle Power Saver**
This mode is intended to save the maximum amount of energy when the system is nearly idle. When the processors are nearly idle, the frequency of all processors is lowered to the minimum. Additionally, workloads are dispatched onto a smaller number of processor cores so that the other processor cores can be put into a low energy usage state. When processor usage increases, the process is reversed. The processor frequency is raised back up to the nominal frequency, and the workloads are spread out again over all of the processor cores. This mode uses no performance boosting, but entering or exiting this mode might affect overall performance. The delay times and usage levels for entering and exiting this mode can be adjusted to allow for more or less aggressive energy savings.
The controls for all these modes are available on the Advanced System Management Interface (ASMI) and are described in more detail in a white paper that is available at the following website:


### 2.1.10 Comparison of the POWER8, POWER7+, and POWER7 processors

Table 2-4 shows comparable characteristics among the generations of POWER8, POWER7+, and POWER7 processors.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>POWER8</th>
<th>POWER7+</th>
<th>POWER7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>22 nm</td>
<td>32 nm</td>
<td>45 nm</td>
</tr>
<tr>
<td>Die size</td>
<td>649 mm$^2$</td>
<td>567 mm$^2$</td>
<td>567 mm$^2$</td>
</tr>
<tr>
<td>Number of transistors</td>
<td>4.2 billion</td>
<td>2.1 billion</td>
<td>1.2 billion</td>
</tr>
<tr>
<td>Maximum cores</td>
<td>12</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Maximum SMT threads per core</td>
<td>Eight threads</td>
<td>Four threads</td>
<td>Four threads</td>
</tr>
<tr>
<td>Maximum frequency</td>
<td>4.15 GHz</td>
<td>4.4 GHz</td>
<td>4.25 GHz</td>
</tr>
<tr>
<td>L2 Cache</td>
<td>512 KB per core</td>
<td>256 KB per core</td>
<td>256 KB per core</td>
</tr>
<tr>
<td>L3 Cache</td>
<td>8 MB of FLR-L3 cache per core with each core having access to the full 96 MB of L3 cache, on-chip eDRAM</td>
<td>10 MB of FLR-L3 cache per core with each core having access to the full 80 MB of L3 cache, on-chip eDRAM</td>
<td>4 MB or 8 MB of FLR-L3 cache per core with each core having access to the full 32 MB of L3 cache, on-chip eDRAM</td>
</tr>
<tr>
<td>Memory support</td>
<td>DDR3 and DDR4</td>
<td>DDR3</td>
<td>DDR3</td>
</tr>
<tr>
<td>I/O bus</td>
<td>PCIe Gen3</td>
<td>GX++</td>
<td>GX++</td>
</tr>
</tbody>
</table>

### 2.2 Memory subsystem

The Power S824L is a two-socket system that supports up to two POWER8 processor modules. The server supports a maximum of 16 DDR3 CDIMM slots, with eight DIMM slots per installed processor. Memory features that are supported are 16 GB, 32 GB, and 64 GB, and they run at speeds of 1600 MHz, allowing for a maximum system memory of 1024 GB.
2.2.1 Custom DIMM

CDIMMs are innovative memory DIMMs that house industry-standard dynamic random access memory (DRAM) memory chips and include a set of components that allow for higher bandwidth and lower latency communications:

- Memory Scheduler
- Memory Management (Reliability, availability, and serviceability (RAS) Decisions & Energy Management)
- Buffer Cache

By adopting this architecture for the memory DIMMs, several decisions and processes regarding memory optimizations are run internally on the CDIMM, saving bandwidth and allowing for faster processor to memory communications. It also allows for more robust RAS. For more information about RAS, see Chapter 3, “Reliability, availability, and serviceability” on page 73.

Figure 2-8 shows a detailed diagram of the CDIMM that is available for the Power S824L.

![CDIMM Diagram](image)

**Figure 2-8  Short CDIMM diagram**

The Buffer Cache is a L4 cache and is built on eDRAM technology (same as the L3 cache), which has lower latency than regular SRAM. Each CDIMM has 16 MB of L4 cache, and a fully populated Power S824L server (two processors and 16 CDIMMs) has 256 MB of L4 cache. The L4 cache performs several functions that directly affect performance and brings a series of benefits for the Power S824L:

- Reduces energy consumption by reducing the number of memory requests.
- Increases memory write performance by acting as a cache and by grouping several random writes into larger transactions.
- “Gathers” partial write operations that target the same cache block within the L4 cache before they are written to memory, therefore becoming a single write operation.
- Reduces latency on memory access. Memory access for cached blocks has up to 55% lower latency than non-cached blocks.
2.2.2 Memory placement rules

The following memory options are orderable:

- 16 GB CDIMM, 1600 MHz DDR3 DRAM (#EM8B)
- 32 GB CDIMM, 1600 MHz DDR3 DRAM (#EM8C)
- 64 GB CDIMM, 1600 MHz DDR3 DRAM (#EM8D)
- 128 GB CDIMM, 1600 MHz DDR3 DRAM (#EM8E)

All memory must be ordered in pairs with a minimum of 32 GB memory for the Power S824L.

The supported maximum memory is 2 TB (sixteen 128 GB CDIMMs), considering two installed processor modules.

The basic rules for memory placement follow:

- Each feature code equates to a single physical CDIMM.
- All memory features must be ordered in pairs.
- All memory CDIMMs must be installed in pairs.
- Each CDIMM within a pair must be of the same capacity.

Figure 2-9 shows the physical memory DIMM topology.

Figure 2-9   Memory DIMM topology for the Power S824L

In general, the preferred approach is to install memory evenly across all processors in the system. Balancing memory across the installed processors allows memory access in a consistent manner and typically results in the best possible performance for your configuration. Account for any plans for future memory upgrades when you decide which memory feature size to use at the time of the initial system order.
For systems with a single processor module installed, use the following plugging order for memory DIMMS:

- The first CDIMM pair is identical and installed at C16 and C18.
- The next CDIMM pair is identical and installed at C21 and C23.
- The next CDIMM pair is identical and installed at C17 and C19.
- The next CDIMM pair is identical and installed at C20 and C22.

For systems with two processor modules that are installed, use the following plugging order for memory DIMMS:

- The first CDIMM pair is identical and installed at C16 and C18.
- The next CDIMM pair is identical and installed at C24 and C26.
- The next CDIMM pair is identical and installed at C21 and C23.
- The next CDIMM pair is identical and installed at C29 and C31.
- The next CDIMM pair is identical and installed at C17 and C19.
- The next CDIMM pair is identical and installed at C25 and C27.
- The next CDIMM pair is identical and installed at C20 and C22.
- The next CDIMM pair is identical and installed at C28 and C30.

### 2.2.3 Memory bandwidth

In this chapter we show the memory bandwidth for the Power S824L with a GPU and the Power S824L in a non-GPU configuration.

**Power S824L with a NVIDIA GPU**

The POWER8 processor has exceptional cache, memory, and interconnect bandwidths. Table 2-5 shows the maximum bandwidth estimates for a single core on the Power S824L with GPU.

<table>
<thead>
<tr>
<th>Single core</th>
<th>Power 824L</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 core @ 3.02 GHz</td>
</tr>
<tr>
<td>L1 (data) cache</td>
<td>144.96 GBps</td>
</tr>
<tr>
<td>L2 cache</td>
<td>144.96 GBps</td>
</tr>
<tr>
<td>L3 cache</td>
<td>193.28 GBps</td>
</tr>
</tbody>
</table>

The bandwidth figures for the caches are calculated in the following manner:

- **L1 cache:** In one clock cycle, two 16-byte load operations and one 16-byte store operation can be accomplished. The value varies depending on the clock of the core and the formula is shown:

  - 3.02 GHz Core: \((2 \times 16 \text{ B} + 1 \times 16 \text{ B}) \times 3.02 \text{ GHz} = 144.96 \text{ GBps}\)
  - 3.42 GHz Core: \((2 \times 16 \text{ B} + 1 \times 16 \text{ B}) \times 3.42 \text{ GHz} = 164.16 \text{ GBps}\)

- **L2 cache:** In one clock cycle, one 32-byte load operation and one 16-byte store operation can be accomplished. The value varies depending on the clock of the core and the formula is shown:

  - 3.02 GHz Core: \((1 \times 32 \text{ B} + 1 \times 16 \text{ B}) \times 3.02 \text{ GHz} = 144.96 \text{ GBps}\)
  - 3.42 GHz Core: \((1 \times 32 \text{ B} + 1 \times 16 \text{ B}) \times 3.42 \text{ GHz} = 164.16 \text{ GBps}\)
- L3 cache: One 32-byte load operation and one 32-byte store operation can be accomplished at half-clock speed and the formula is shown:
  
  3.02 GHz Core: \((1 \times 32 \text{ B} + 1 \times 32 \text{ B}) \times 3.02 \text{ GHz} = 193.28 \text{ GBps}\)
  
  3.42 GHz Core: \((1 \times 32 \text{ B} + 1 \times 32 \text{ B}) \times 3.42 \text{ GHz} = 218.88 \text{ GBps}\)

For the entire Power S824L system populated with two processor modules, the overall bandwidths are shown in Table 2-6.

**Table 2-6  Power S824L with a GPU - total bandwidth estimates**

<table>
<thead>
<tr>
<th>Total bandwidths</th>
<th>Power S824L</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>24 cores @ 3.02 GHz</td>
</tr>
<tr>
<td>L1 (data) cache</td>
<td>3,479 GBps</td>
</tr>
<tr>
<td>L2 cache</td>
<td>3,479 GBps</td>
</tr>
<tr>
<td>L3 cache</td>
<td>4,638 GBps</td>
</tr>
<tr>
<td>Total memory</td>
<td>384 GBps</td>
</tr>
<tr>
<td>SMP Interconnect</td>
<td>25.6 GBps</td>
</tr>
<tr>
<td>PCIe Interconnect</td>
<td>189.048 GBps</td>
</tr>
</tbody>
</table>

For total memory bandwidth, each POWER8 processor has eight memory channels running at 8 GBps capable of writing 2 bytes and reading 1 byte at a time. The bandwidth formula is calculated in the following manner:

\[8 \text{ channels} \times 8 \text{ GBps} \times 3 \text{ Bytes} = 192 \text{ GBps per processor module}\]

For SMP Interconnect, the POWER8 processor has two 2-byte 3-lane A buses working at 6.4 GHz. Each A bus has two active lanes and one spare lane. The bandwidth formula is calculated in the following manner:

\[2 \text{ A buses} \times 2 \text{ Bytes} \times 6.4 \text{ GHz} = 25.6 \text{ GBps}\]

For PCIe Interconnect, each POWER8 processor has 48 PCIe lanes running at 7.877 Gbps full-duplex. The bandwidth formula is calculated in the following manner:

\[48 \text{ lanes} \times 2 \text{ processors} \times 8 \text{ Gbps} \times 2 = 189.048 \text{ GBps}\]

**Rounding:** The bandwidths listed here might appear slightly differently in other materials due to the rounding of some of the numbers.

**Power S824L without a NVIDIA GPU**

The POWER8 processor has exceptional cache, memory, and interconnect bandwidths. Table 2-7 shows the maximum bandwidth estimates for a single core on the Power S824L without a GPU.

**Table 2-7  Power S824L with no GPU - single core bandwidth estimates**

<table>
<thead>
<tr>
<th>Single core</th>
<th>Power 824L</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 core @ 3.52 GHz</td>
</tr>
<tr>
<td>L1 (data) cache</td>
<td>168.96 GBps</td>
</tr>
<tr>
<td>L2 cache</td>
<td>168.96 GBps</td>
</tr>
<tr>
<td>L3 cache</td>
<td>225.28 GBps</td>
</tr>
</tbody>
</table>
The bandwidth figures for the caches are calculated in the following manner:

- **L1 cache**: In one clock cycle, two 16-byte load operations and one 16-byte store operation can be accomplished. The value varies depending on the clock of the core and the formula is shown:
  - 3.52 GHz Core: \( (2 \times 16 \text{ B} + 1 \times 16 \text{ B}) \times 3.52 \text{ GHz} = 168.96 \text{ GBps} \)
  - 4.15 GHz Core: \( (2 \times 16 \text{ B} + 1 \times 16 \text{ B}) \times 4.15 \text{ GHz} = 199.2 \text{ GBps} \)

- **L2 cache**: In one clock cycle, one 32-byte load operation and one 16-byte store operation can be accomplished. The value varies depending on the clock of the core and the formula is shown:
  - 3.52 GHz Core: \( (1 \times 32 \text{ B} + 1 \times 16 \text{ B}) \times 3.52 \text{ GHz} = 168.96 \text{ GBps} \)
  - 4.15 GHz Core: \( (1 \times 32 \text{ B} + 1 \times 16 \text{ B}) \times 4.15 \text{ GHz} = 199.2 \text{ GBps} \)

- **L3 cache**: One 32-byte load operation and one 32-byte store operation can be accomplished at half-clock speed and the formula is shown:
  - 3.52 GHz Core: \( (1 \times 32 \text{ B} + 1 \times 32 \text{ B}) \times 3.52 \text{ GHz} = 225.28 \text{ GBps} \)
  - 4.15 GHz Core: \( (1 \times 32 \text{ B} + 1 \times 32 \text{ B}) \times 4.15 \text{ GHz} = 265.6 \text{ GBps} \)

For the entire Power S824L with no GPU system populated with two processor modules, the overall bandwidths are shown in Table 2-8.

### Table 2-8  Power S824L with no GPU - total bandwidth estimates

<table>
<thead>
<tr>
<th>Total bandwidths</th>
<th>Power S824L</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>24 cores @ 3.52 GHz</td>
</tr>
<tr>
<td>L1 (data) cache</td>
<td>4,055 GBps</td>
</tr>
<tr>
<td>L2 cache</td>
<td>4,055 GBps</td>
</tr>
<tr>
<td>L3 cache</td>
<td>5,407 GBps</td>
</tr>
<tr>
<td>Total memory</td>
<td>384 GBps</td>
</tr>
<tr>
<td>SMP Interconnect</td>
<td>25.6 GBps</td>
</tr>
<tr>
<td>PCIe Interconnect</td>
<td>189.048 GBps</td>
</tr>
</tbody>
</table>

For total memory bandwidth, each POWER8 processor has eight memory channels running at 8 GBps capable of writing 2 bytes and reading 1 byte at a time. The bandwidth formula is calculated in the following manner:

\[
8 \text{ channels} \times 8 \text{ GBps} \times 3 \text{ Bytes} = 192 \text{ GBps per processor module}
\]

For SMP Interconnect, the POWER8 processor has two 2-byte 3-lane A buses working at 6.4 GHz. Each A bus has two active lanes and one spare lane. The bandwidth formula is calculated in the following manner:

\[
2 \text{ A buses} \times 2 \text{ Bytes} \times 6.4 \text{ GHz} = 25.6 \text{ GBps}
\]

For PCIe Interconnect, each POWER8 processor has 48 PCIe lanes running at 7.877 GBps full-duplex. The bandwidth formula is calculated in the following manner:

\[
48 \text{ lanes} \times 2 \text{ processors} \times 8 \text{ Gbps} \times 2 = 189.048 \text{ GBps}
\]

**Rounding**: The bandwidths listed here might appear slightly differently in other materials due to the rounding of some of the numbers.
2.3 System bus

This section provides more information about the internal buses.

The Power S824L systems have internal I/O connectivity through Peripheral Component Interconnect Express Gen3 (PCI Express Gen3 or PCIe Gen3) slots and also external connectivity through serial-attached SCSI (SAS) adapters.

The internal I/O subsystem on the systems is connected to the PCIe Controllers on a POWER8 processor in the system. Each POWER8 processor has a bus that has 48 PCIe lanes running at 7.877 Gbps full-duplex and provides 96 GBps of I/O connectivity to the PCIe slots, serial-attached SCSI (SAS) internal adapters, and USB ports.

Some PCIe devices are connected directly to the PCIe Gen3 buses on the processors, and other devices are connected to these buses through PCIe Gen3 Switches. The PCIe Gen3 Switches are high-speed devices (512 GBps - 768 GBps each) that allow for the optimal usage of the processor’s PCIe Gen3 x16 buses by grouping slower x8 or x4 devices that may be plugged into a x16 slot, therefore not using the slot’s full bandwidth. For more information about which slots are connected directly to the processor and which slots are attached to PCIe Gen3 Switches (referred as PEX), see 2.1, “The IBM POWER8 processor” on page 28.

A diagram that compares the architecture of the POWER7 and POWER8 I/O buses is shown in Figure 2-10.

![Comparison of the architecture of the POWER7 and POWER8 I/O buses](image)

Table 2-9 lists the I/O bandwidth of Power S824L processor configurations.

<table>
<thead>
<tr>
<th>I/O</th>
<th>I/O bandwidth (maximum theoretical)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total I/O bandwidth</td>
<td>94.524 GBps simplex</td>
</tr>
<tr>
<td></td>
<td>189.048 GBps duplex</td>
</tr>
</tbody>
</table>
For PCIe Interconnect, each POWER8 processor has 48 PCIe lanes running at 7.877 Gbps full-duplex. The bandwidth formula is calculated in the following manner:

\[ 48 \text{ lanes} \times 2 \text{ processors} \times 7.877 \text{ Gbps} \times 2 = 189.048 \text{ Gbps} \]

## 2.4 Internal I/O subsystem

The internal I/O subsystem is on the system board, which supports PCIe slots. PCIe adapters on the Power S824L are hot-pluggable.

All PCIe slots support Enhanced Error Handling (EEH). PCI EEH-enabled adapters respond to a special data packet that is generated from the affected PCIe slot hardware by calling system firmware, which examines the affected bus and allows the device driver to reset it, and continues without a system restart. For Linux, EEH support extends to many frequently used devices, although certain third-party PCI devices might not provide native EEH support.

### 2.4.1 Slot configuration

The number of the PCIe slots available on IBM Power Scale-out servers depends on the processors installed. It is mandatory for the Power S824L to have both processor modules installed so that all its PCIe slots are active. The available slots are listed:

- Qty 4 - x16 Gen3 low profile slots plus
- Qty 7 - x8 Gen3 low profile slots

Table 2-10 shows the PCIe Gen3 slot configuration for the Power S824L.

<table>
<thead>
<tr>
<th>Slot</th>
<th>Description</th>
<th>Location code</th>
<th>Card size</th>
<th>CAPI capable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slot 1</td>
<td>PCIe Gen3 x8</td>
<td>P1-C2</td>
<td>Full height, half length</td>
<td>No</td>
</tr>
<tr>
<td>Slot 2</td>
<td>PCIe Gen3 x16</td>
<td>P1-C3</td>
<td>Full height, full length</td>
<td>Yes</td>
</tr>
<tr>
<td>Slot 3</td>
<td>PCIe Gen3 x8</td>
<td>P1-C4</td>
<td>Full height, half length</td>
<td>No</td>
</tr>
<tr>
<td>Slot 4</td>
<td>PCIe Gen3 x16</td>
<td>P1-C5</td>
<td>Full height, full length</td>
<td>Yes</td>
</tr>
<tr>
<td>Slot 5</td>
<td>PCIe Gen3 x16</td>
<td>P1-C6</td>
<td>Full height, full length</td>
<td>Yes</td>
</tr>
<tr>
<td>Slot 6</td>
<td>PCIe Gen3 x16</td>
<td>P1-C7</td>
<td>Full height, full length</td>
<td>Yes</td>
</tr>
<tr>
<td>Slot 7</td>
<td>PCIe Gen3 x8</td>
<td>P1-C8</td>
<td>Full height, half length</td>
<td>No</td>
</tr>
<tr>
<td>Slot 8</td>
<td>PCIe Gen3 x8</td>
<td>P1-C9</td>
<td>Full height, half length</td>
<td>No</td>
</tr>
<tr>
<td>Slot 9a</td>
<td>PCIe Gen3 x8</td>
<td>P1-C10</td>
<td>Full height, half length</td>
<td>No</td>
</tr>
<tr>
<td>Slot 10</td>
<td>PCIe Gen3 x8</td>
<td>P1-C11</td>
<td>Full height, half length</td>
<td>No</td>
</tr>
</tbody>
</table>
Figure 2-11 is a diagram of the top view of the Power S824L server. PCIe slots on the server with the default backplane can be seen.

At least one PCIe Ethernet adapter is required on the server and therefore one of the x8 PCIe slots is used for this required adapter, which is identified as the P1-C10 slot. Included on all base configurations, as of the time of writing, this adapter is the PCIe2 4-port 1 Gb Ethernet Adapter (#EN0W). It can be replaced or moved in field or as part of an upgrade if the server still contains at least one Ethernet adapter.

**Remember:** Slot 9 (P1-C10) comes with the PCIe2 4-port 1 Gb Ethernet Adapter (#EN0W) installed. If this slot is needed, the adapter must be moved in the field to another suitable slot or replaced by another Ethernet adapter.

### 2.4.2 System ports

The system board has one serial port that is called a system port. When a Hardware Management Console (HMC) is connected to the server, the integrated system port of the server is rendered non-functional:

- The integrated system port is not supported under Linux when the HMC ports are connected to an HMC. Either the HMC ports or the integrated system port can be used, but not both.
- The integrated system port is supported for modem and asynchronous terminal connections. Any other application that uses serial ports requires that a serial port adapter
is installed in a PCI slot. The integrated system ports do not support IBM PowerHA configurations.

- The configuration of the integrated system port, including basic port settings (baud rate, and so on), modem selection, and Call Home and call-in policy, can be performed with the Advanced Systems Management Interface (ASMI).

**Remember:** The integrated console/modem port usage that is described is for systems that are configured as a single, system-wide partition. When the system is configured with multiple partitions, the integrated console/modem ports are disabled because the TTY console and Call Home functions are performed with the HMC.

### 2.5 PCI adapters

This section covers the various types and functions of the PCI adapters that are supported by the Power S824L servers.

#### 2.5.1 PCI Express

Peripheral Component Interconnect Express (PCIe) uses a serial interface and allows for point-to-point interconnections between devices (using a directly wired interface between these connection points). 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 two pairs of wires are called a *lane*. A PCIe link can consist of multiple lanes. In these configurations, the connection is labeled as x1, x2, x8, x12, x16, or x32, where the number is effectively the number of lanes.

The PCIe interfaces that are supported on this server are PCIe Gen3, which are capable of 16 GBps simplex (32 GBps duplex) on a single x16 interface. PCIe Gen3 slots also support previous generation (Gen2 and Gen1) adapters, which operate at lower speeds, according to the following rules:

- Place x1, x4, x8, and x16 speed adapters in the same size connector slots first, before mixing adapter speed with connector slot size.
- Adapters with lower speeds are allowed in larger sized PCIe connectors but larger speed adapters are not compatible in smaller connector sizes (that is, a x16 adapter cannot go in an x8 PCIe slot connector).

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.

IBM POWER8 processor-based servers can support two different form factors of PCIe adapters:

- PCIe low profile (LP) cards, which are used with the Power S822L PCIe slots. These cards are not compatible with Power S824L servers because the tail stock of the adapter is physically too small to fit. The same adapters are often available with the full height tail stock but they will have a different feature code.
- PCIe full height and full high cards are designed for the following scale-out servers:
  - Power S814
  - Power S824
  - Power S824L
Before adding or rearranging adapters, use the System Planning Tool to validate the new adapter configuration. For more information, see the System Planning Tool website: http://www.ibm.com/systems/support/tools/systemplanningtool/

If you are installing a new feature, ensure that you have the software that is required to support the new feature and determine whether there are any existing update prerequisites to install. To obtain this information, use the IBM prerequisite website: https://www-912.ibm.com/e_dir/eServerPreReq.nsf

The following sections describe the supported adapters and provide tables of orderable feature code numbers.

### 2.5.2 LAN adapters

To connect the Power S824L servers to a local area network (LAN), you can use the LAN adapters that are supported in the PCIe slots of the system unit. Table 2-11 lists the available LAN adapters.

**Table 2-11** Available LAN adapters

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>Max</th>
<th>OS support</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC3B</td>
<td>57B6</td>
<td>PCIe3 2-Port 40 GbE NIC RoCE QSFP+ Adapter</td>
<td>30</td>
<td>Linux</td>
</tr>
<tr>
<td>EC3M</td>
<td></td>
<td>PCIe3 2-port 100GbE (NIC&amp;RoCE) QSFP28 Adapter x16</td>
<td>4</td>
<td>Linux</td>
</tr>
<tr>
<td>EL4L</td>
<td></td>
<td>PCIe2 4-port 1GbE Adapter</td>
<td>31</td>
<td>Linux</td>
</tr>
<tr>
<td>EL50</td>
<td></td>
<td>PCIe3 2-port 56Gb FDR IB Adapter x16</td>
<td>4</td>
<td>Linux</td>
</tr>
<tr>
<td>EL53</td>
<td></td>
<td>PCIe3 2-port 10 GbE NIC&amp;RoCE SFP+ Copper Adapter</td>
<td>30</td>
<td>Linux</td>
</tr>
<tr>
<td>EL54</td>
<td></td>
<td>PCIe3 2-port 10 GbE NIC&amp;RoCE SR Adapter</td>
<td>30</td>
<td>Linux</td>
</tr>
<tr>
<td>EL55</td>
<td></td>
<td>PCIe2 2-port 10/1 GbE BaseT RJ45 Adapter</td>
<td>31</td>
<td>Linux</td>
</tr>
<tr>
<td>EN0S</td>
<td>2CC3</td>
<td>PCIe2 4-Port (10Gb+1 GbE) SR+RJ45 Adapter</td>
<td>2</td>
<td>Linux</td>
</tr>
<tr>
<td>EN15</td>
<td>2CE3</td>
<td>PCIe3 4-port 10 GbE SR Adapter</td>
<td>30</td>
<td>Linux</td>
</tr>
<tr>
<td>EN17</td>
<td>2CE4</td>
<td>PCIe3 4-port 10 GbE SFP+ Copper Adapter</td>
<td>30</td>
<td>Linux</td>
</tr>
</tbody>
</table>

### 2.5.3 SAS adapters

Table 2-12 lists the SAS adapters that are available for Power S824L system.

**Table 2-12** Available SAS adapters

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>Max</th>
<th>OS support</th>
</tr>
</thead>
<tbody>
<tr>
<td>EJ0L</td>
<td>57CE</td>
<td>PCIe3 12 GB Cache RAID SAS Adapter Quad-port 6Gb x8</td>
<td>22</td>
<td>Linux</td>
</tr>
<tr>
<td>EL59</td>
<td></td>
<td>PCIe3 RAID SAS Adapter Quad-port 6Gb x8</td>
<td>22</td>
<td>Linux</td>
</tr>
</tbody>
</table>
2.5.4 Graphics accelerator adapters

Table 2-13 lists the available graphics accelerator adapter. The adapter can be configured to operate in either 8-bit or 24-bit color modes. The adapter supports both analog and digital monitors.

Table 2-13  Available graphics accelerator adapter

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>Max</th>
<th>OS support</th>
</tr>
</thead>
<tbody>
<tr>
<td>EJ14</td>
<td>57B1</td>
<td>PCIe3 12 GB Cache RAID PLUS SAS Adapter</td>
<td>22</td>
<td>Linux</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Quad-port 6 Gb x8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EJ1P</td>
<td>57B3</td>
<td>PCIe1 SAS Tape/DVD Dual-port 3Gb x8 Adapter</td>
<td>8</td>
<td>Linux</td>
</tr>
</tbody>
</table>

2.5.5 Fibre Channel adapters

The servers support direct or SAN connection to devices that use Fibre Channel adapters. Table 2-14 summarizes the available Fibre Channel adapters, which all have LC connectors.

If you are attaching a device or switch with an SC type fiber connector, then an LC-SC 50 Micron Fibre Converter Cable (#2456) or an LC-SC 62.5 Micron Fibre Converter Cable (#2459) is required.

Table 2-14  Available Fibre Channel adapters

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>Max</th>
<th>OS support</th>
</tr>
</thead>
<tbody>
<tr>
<td>EL58</td>
<td>5748</td>
<td>PCIe 8 Gb 2-port Fibre Channel Adapter</td>
<td>30</td>
<td>Linux</td>
</tr>
<tr>
<td>EL5B</td>
<td>57B3</td>
<td>PCIe3 16 Gb 2-port Fibre Channel Adapter</td>
<td>30</td>
<td>Linux</td>
</tr>
<tr>
<td>EL5Z</td>
<td>57B4</td>
<td>PCIe2 8 Gb 2-Port Fibre Channel Adapter</td>
<td>30</td>
<td>Linux</td>
</tr>
</tbody>
</table>
2.5.6 Fibre Channel over Ethernet

Fibre Channel over Ethernet (FCoE) allows for the convergence of Fibre Channel and Ethernet traffic onto a single adapter and a converged fabric.

Figure 2-12 compares existing Fibre Channel and network connections and FCoE connections.

Table 2-15 lists the available Fibre Channel over Ethernet Adapters. They are high-performance Converged Network Adapters (CNAs) with several port combination options (SR, LR, SFP+, and copper). Each port can simultaneously provide network interface card (NIC) traffic and Fibre Channel functions. The ports can also handle network only traffic, providing 10 Gbps Ethernet network ports.

Table 2-15 Available FCoE adapters

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>Max</th>
<th>OS support</th>
</tr>
</thead>
<tbody>
<tr>
<td>EL56</td>
<td>PCIe2 4-port (10Gb FCoE &amp; 1 GbE) SR&amp;RJ45</td>
<td>30</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>EL57</td>
<td>PCIe2 4-port (10Gb FCoE &amp; 1 GbE) SFP+Copper&amp;RJ45</td>
<td>30</td>
<td>Linux</td>
<td></td>
</tr>
</tbody>
</table>

For more information about FCoE, see An Introduction to Fibre Channel over Ethernet, and Fibre Channel over Convergence Enhanced Ethernet, REDP-4493.
2.5.7 Asynchronous and USB adapters

Asynchronous PCIe adapters connect asynchronous EIA-232 or RS-422 devices. If you have a cluster configuration or high-availability configuration and plan to connect the IBM Power Systems using a serial connection, you can use the asynchronous adapters.

Table 2-16 lists the available Asynchronous and USB adapters.

Table 2-16  Available Asynchronous and USB adapters

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>Max</th>
<th>OS support</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC46</td>
<td>PCIe2 4-Port USB 3.0 Adapter</td>
<td>30</td>
<td>Linux</td>
<td></td>
</tr>
</tbody>
</table>

2.5.8 Cryptographic coprocessor

Table 2-17 lists the available cryptographic adapters.

Table 2-17  Available Asynchronous and USB adapters

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>Max</th>
<th>OS support</th>
</tr>
</thead>
<tbody>
<tr>
<td>EJ32</td>
<td>PCIe3 Crypto Coprocessor no BSC 4767</td>
<td>10</td>
<td>Linux</td>
<td></td>
</tr>
<tr>
<td>ELJ33</td>
<td>PCIe3 Crypto Coprocessor BSC-Gen3 4767</td>
<td>10</td>
<td>Linux</td>
<td></td>
</tr>
</tbody>
</table>

2.5.9 Field-programmable gate array adapters

The field-programmable gate array (FPGA) adapters are PCIe adapters that are based on a semiconductor device that can be programmed.

Unlike an ASIC, which is designed and programmed to perform a single function, an FPGA can be programmed to run different product functions, adapt to new standards, and reconfigure its hardware for specific applications even after the product is installed in the field. An FPGA can implement any logical function that an ASIC can perform, but can have its code updated to include more functions or perform a different role.

Today, FPGAs have logic elements that can be reprogrammed, and include SRAM memory, high-speed interconnects that can range up to 400 Gbps, logic blocks, and routing. FPGAs allow for versatile solutions because a single adapter can perform several distinct functions depending on the code that is deployed on it.

By having a highly optimized software stack, the FPGAs can act as coprocessor for the server CPUs, running repetitive and complex functions at a fraction of the time and power, while allowing for the server CPUs to perform other functions at the same time.

The FPGA adapter that is supported for the Power S824L is shown in Table 2-18.

Table 2-18  Available FPGA adapter

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>Max</th>
<th>OS support</th>
</tr>
</thead>
<tbody>
<tr>
<td>EJ12</td>
<td>59AB</td>
<td>PCIe3 FPGA Accelerator Adapter</td>
<td>2</td>
<td>RHEL</td>
</tr>
</tbody>
</table>
The #EJ12 is an FPGA adapter that is based on Altera Stratix V 28 nm hardware. In addition to its logic components, it has 8 GB DDR3 RAM.

The initial implementation for the #EJ12 is an adapter that can perform gzip compression. The zlib API, a software library that is responsible for data compression and is used in several software packages, can move the tasks directly to the FPGA, which allows increased compression performance, increased compression rates, and decreased CPU usage. Java 7.1 is already enabled to take advantage of this acceleration.

Other applications, such as big data, can take advantage of this approach to compression after the compression rates are higher than the ones that are achieved through software. The compression processes complete in a shorter time by allowing for more data density, disk savings, and faster data analysis.

2.5.10 Compute Intensive Accelerator

Compute Intensive Accelerators are graphics processing units (GPUs) that are developed by NVIDIA. With NVIDIA GPUs, the Power S824L is designed to offload processor-intensive operations to a GPU accelerator and boost performance. The Power S824L aims to deliver a new class of technology that maximizes performance and efficiency for all types of scientific, engineering, Java, big data analytics, and other technical computing workloads.

The available NVIDIA GPUs are based on the K40 Tesla adapters, which are double-wide cards and therefore occupy two PCIe slots. The mandatory first card occupies slots P1-C6 and P1-C7. The optional secondary card occupies slots P1-C3 and P1-C4.

Table 2-19 lists the available Compute Intensive Accelerators.

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>Maximum per system</th>
<th>OS support</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC47</td>
<td>2CE8</td>
<td>Compute Intensive Accelerator (NVIDIA K40Tesla GPU)</td>
<td>2</td>
<td>Linux</td>
</tr>
<tr>
<td>EC4B</td>
<td></td>
<td>Compute Intensive Accelerator (NVIDIA K80 Tesla GPU)</td>
<td>2</td>
<td>Linux</td>
</tr>
</tbody>
</table>

For more information about the NVIDIA K40 Tesla GPU, see this website:

2.5.11 Flash storage adapters

The available flash storage adapters are shown in Table 2-20.

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>Max</th>
<th>OS support</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC55</td>
<td>58CB</td>
<td>PCIe3 1.6TB NVMe Flash Adapter</td>
<td>8</td>
<td>Linux</td>
</tr>
<tr>
<td>EC57</td>
<td>58CC</td>
<td>PCIe3 3.2TB NVMe Flash Adapter</td>
<td>8</td>
<td>Linux</td>
</tr>
</tbody>
</table>
2.5.12 CAPI adapters

The available CAPI adapters are shown in Table 2-21.

*Table 2-21  Available CAPI adapters*

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>Max</th>
<th>OS support</th>
</tr>
</thead>
<tbody>
<tr>
<td>EJ1A</td>
<td>2CF0</td>
<td>CAPI Compression Accelerator Adapter</td>
<td>2</td>
<td>Ubuntu, RHEL</td>
</tr>
</tbody>
</table>

This CAPI FPGA (Field Programmable Gate Array) adapter acts as a co-processor for the POWER8 processor chip handling specialized, repetitive function extremely efficiently. The adapter is preloaded with a GZIP application and is intended to run as a gzip accelerator. The GZIP application maximum bandwidth is a PCIe Gen3 interface bandwidth.

2.6 Internal storage

The internal storage on the Power S824L server depends on the DASD/Media backplane that is used. The server supports two DASD/Media backplanes if it has no GPU installed: #EJ0N and #EJ0P.

The #EJ0N storage backplane contains the following items:

- A storage backplane for twelve 2.5-inch SFF Gen3 HDDs or SSDs.
- One SAS disk controller that is capable of RAID 0, RAID 5, RAID 6, and RAID 10, placed in a dedicated SAS controller slot (P1-C14).
- The optional split backplane feature #EJ0S adds a secondary SAS disk controller and allows DASDs to be split into two groups of six (6+6). This secondary controller is placed in the second dedicated SAS controller slot (P1-C15).

The #EJ0P storage backplane contains the following items:

- A storage backplane for eighteen 2.5-inch SFF Gen3 HDDs or SSDs.
- Two active-active SAS disk controllers that are capable of RAID 0, RAID 5, RAID 6, RAID 10, RAID 5T2, RAID 6T2, and RAID 10T2, which can be placed in dedicated SAS controller slots P1-C14 and P1-C15.
- The #EJTM SSD Module Cage, which is automatically added by e-config when #EJ0P is ordered, adds a secondary disk cage with eight 1.8-inch SSD module bays.
- Two external SAS ports for DASD drawers connectivity (through slot P1-C11) support one EXP24S SFF Gen2-bay Drawer (#5887).
- The storage split backplane function is not supported.

*Note:* If the Power S824L comes with a NVIDIA GPU only the Storage backplane with 12 SFF-3 bays and one DVD bay (#EJ0N) is supported.

Table 2-22 presents a summarized view of these features.
The 2.5-inch or small form factor (SFF) SAS bays can contain SAS drives (HDD or SSD) that are mounted on a Gen3 tray or carrier (also known as SFF-3). SFF-1 or SFF-2 drives do not fit in an SFF-3 bay. All SFF-3 bays support concurrent maintenance or hot-plug capability.

Additionally, as an option for the #EJ0P backplane, the feature #EJTM adds an 8-bay 1.8-inch SSD Cage behind the server bezel. All eight bays are accessed by both of the SAS controllers and the bays support concurrent maintenance (hot-plug).

The server front view is shown in Figure 2-13.

The internal connections to the physical disks are shown in Figure 2-14.
2.6.1 RAID support

There are multiple protection options for HDDs in the Power S824L systems in the SAS SFF bays in the system unit. Although protecting drives is always recommended, Linux users can choose to leave a few or all drives unprotected at their own risk, and IBM supports these configurations.

Drive protection
HDD drive protection can be provided by Linux, or by the HDD hardware controllers.

The default storage backplane #EJ0N contains one SAS HDD controller and provides support for just a bunch of disks (JBOD) and RAID 0, 5, 6, and 10 for Linux.

Table 2-23 lists the RAID support configurations by the storage backplane.

<table>
<thead>
<tr>
<th>Storage backplane</th>
<th>JBOD</th>
<th>RAID 0, 5, 6, and 10</th>
<th>RAID 0, 5, 6, and 10 and Easy Tier (RAID 5T2, 6T2, and 10T2)</th>
<th>Split backplane</th>
<th>External SAS port</th>
</tr>
</thead>
<tbody>
<tr>
<td>EJ0N</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Linux can use disk drives that are formatted with 512-byte blocks when they are mirrored by the operating system. These disk drives must be reformatted to 528-byte sectors when used.
in RAID arrays. Although a small percentage of the drive’s capacity is lost, additional data protection, such as error-correcting code (ECC) and bad block detection, is gained in this reformatting. For example, a 300 GB disk drive, when reformatted, provides approximately 283 GB. SSDs are always formatted with 528-byte sectors.

**Supported RAID functions**

The base hardware supports RAID 0, 5, 6, and 10:

- RAID 0 provides striping for performance, but it does not offer any fault tolerance.
  
  The failure of a single drive results in the loss of all data on the array. This version of RAID increases I/O bandwidth by simultaneously accessing multiple data paths.

- RAID 5 uses block-level data striping with distributed parity.
  
  RAID 5 stripes both data and parity information across three or more drives. Fault tolerance is maintained by ensuring that the parity information for any specific block of data is placed on a drive that is separate from the drives that are used to store the data. This version of RAID provides data resiliency if a single drive fails in a RAID 5 array.

- RAID 6 uses block-level data striping with dual distributed parity.
  
  RAID 6 is the same as RAID 5 except that it uses a second level of independently calculated and distributed parity information for additional fault tolerance. A RAID 6 configuration requires N+2 drives to accommodate the additional parity data, making it less cost-effective than RAID 5 for equivalent storage capacity. This version of RAID provides data resiliency if one or two drives fail in a RAID 6 array. When you work with large capacity disks, RAID 6 allows you to sustain data parity during the rebuild process.

- RAID 10 is a striped set of mirrored arrays.
  
  It is a combination of RAID 0 and RAID 1. A RAID 0 stripe set of the data is created across a two-disk array for performance benefits. A duplicate of the first stripe set is then mirrored on another two-disk array for fault tolerance. This version of RAID provides data resiliency if a single drive fails, and it can provide resiliency for multiple drive failures.

### 2.6.2 Media bays

A slimline media bay that can optionally house a Serial Advanced Technology Attachment (SATA) DVD-RAM (#5771) is included in the EJ0N backplane. The direct dock and hot-plug of the DVD media device are supported.

The DVD drive and media device do not have an independent SAS adapter and cannot be assigned to a logical partition (LPAR) independently of the HDDs in the system.

### 2.7 External I/O subsystems

This section describes the PCIe Gen3 I/O expansion drawer that can be attached to the Power S824L without a NVIDIA GPU configuration.

#### 2.7.1 PCIe Gen3 I/O expansion drawer

The PCIe Gen3 I/O expansion drawer (#ELMX) is a 4U high, PCI Gen3-based and rack mountable I/O drawer. It offers two PCIe Fan Out Modules (#ELMF) each of them providing six PCIe slots.
The physical dimensions of the drawer are 444.5 mm (17.5 in.) wide by 177.8 mm (7.0 in.) high by 736.6 mm (29.0 in.) deep for use in a 19-inch rack.

A PCIe x16 to Optical CXP converter adapter (#EJ08) and 3.0 m (#ECC7), 10.0 m (#ECC8) CXP 16X Active Optical cables (AOC) connect the system node to a PCIe Fan Out module in the I/O expansion drawer. One feature #ECC7, one #ECC8 ships two AOC cables.

Concurrent repair and add/removal of PCIe adapter cards is done by HMC guided menus or by operating system support utilities.

A blind swap cassette (BSC) is used to house the full high adapters which go into these slots. The BSC is the same BSC as used with the previous generation server's #5802/5803/5877/5873 12X attached I/O drawers.

Figure 2-15 shows the back view of the PCIe Gen3 I/O expansion drawer.

Figure 2-15  Rear view of the PCIe Gen3 I/O expansion drawer

2.7.2 PCIe Gen3 I/O expansion drawer optical cabling

I/O drawers are connected to the adapters in the system node with data transfer cables:

- 3.0 m Optical Cable Pair for PCIe3 Expansion Drawer (#ECC7)
- 10.0 m Optical Cable Pair for PCIe3 Expansion Drawer (#ECC8)

Cable lengths: Use the 3.0 m cables for intra-rack installations. Use the 10.0 m cables for inter-rack installations.
A minimum of one PCIe3 Optical Cable Adapter for PCIe3 Expansion Drawer (#EJ08) is required to connect to the PCIe3 6-slot Fan Out module in the I/O expansion drawer. The top port of the fan out module must be cabled to the top port of the #EJ08 port. Likewise, the bottom two ports must be cabled together.

1. Connect an active optical cable to connector T1 on the PCIe3 optical cable adapter in your server.
2. Connect the other end of the optical cable to connector T1 on one of the PCIe3 6-slot Fan Out modules in your expansion drawer.
3. Connect another cable to connector T2 on the PCIe3 optical cable adapter in your server.
4. Connect the other end of the cable to connector T2 on the PCIe3 6-slot Fan Out module in your expansion drawer.
5. Repeat the four steps above for the other PCIe3 6-slot Fan Out module in the expansion drawer, if required.

**Drawer connections:** Each Fan Out module in a PCIe3 Expansion Drawer can only be connected to a single PCIe3 Optical Cable Adapter for PCIe3 Expansion Drawer (#EJ08). However the two Fan Out modules in a single I/O expansion drawer can be connected to different system nodes in the same server.

Figure 2-16 shows the connector locations for the PCIe Gen3 I/O expansion drawer.
Figure 2-17 shows typical optical cable connections.

**General rules for the PCI Gen3 I/O expansion drawer configuration**

The PCIe3 optical cable adapter can be in any of the PCIe adapter slots in the Power S824L server. However, we advise that you use the PCIe adapter slot priority information while selecting slots for installing PCIe3 Optical Cable Adapter for PCIe3 Expansion Drawer (#EJ08).

Table 2-24 shows PCIe adapter slot priorities in the Power S824L system.

**Table 2-24  PCIe adapter slot priorities**

<table>
<thead>
<tr>
<th>Feature code</th>
<th>Description</th>
<th>Slot priorities</th>
</tr>
</thead>
<tbody>
<tr>
<td>EJ08</td>
<td>PCIe3 Optical Cable Adapter for PCIe3 Drawer</td>
<td>2, 8, 4, 6, 1, 7, 3, 5</td>
</tr>
</tbody>
</table>

The following figures show the supported configurations. Figure 2-18 shows an example of the Power S824L with no GPU and one PCI Gen3 I/O expansion drawers with two Fan Out modules.
Figure 2-19 shows an example of a two socket Power S824L with no GPU and two PCI Gen3 I/O expansion drawers with four Fan Out modules.

![Diagram of two socket Power S824L with no GPU and two I/O drawers](image)

**2.7.3 PCIe Gen3 I/O expansion drawer SPCN cabling**

There is no system power control network (SPCN) used to control and monitor the status of power and cooling within the I/O drawer. SPCN capabilities are integrated in the optical cables.

**2.8 External disk subsystems**

This section describes the following external disk subsystems that can be attached to the Power S824L server if it is not installed with a NVIDIA GPU:

- EXP24S SFF Gen2-bay drawer for high-density storage (#EL1S)
- IBM System Storage®

**Note:** The EXP30 Ultra SSD Drawer (#EDR1 or #5888), the EXP12S SAS Disk Drawer (#5886), and the EXP24 SCSI Disk Drawer (#5786) are not supported on the Power S824L server.

**2.8.1 EXP24S SFF Gen2-bay drawer**

The EXP24S SFF Gen2-bay Drawer (#EL1S) is an expansion drawer that supports up to 24 hot-swap 2.5-inch SFF SAS HDDs on POWER6, POWER6+, POWER7, POWER7+, or POWER8 servers in 2U of 19-inch rack space. The EXP24S drawer includes redundant AC power supplies and two power cords.

**Note:** Any existing EXP24S SFF Gen2-bay Drawer (#5887) is also supported. For a new order of a EXP24S drawer, the #EL1S must be configured.
The EXP24S uses Gen2 or SFF-2 SAS drives that physically do not fit in the SFF-3 bays of the Power S822L system unit.

The EXP24S drawer is attached to SAS ports on either a PCIe SAS adapter in the server or to the SAS ports at the rear of the server. Two SAS ports on the rear of the server are enabled with the expanded-function storage backplane with dual IOA support (#EJ0P).

The SAS controller and the EXP24S SAS ports are attached by using the appropriate SAS Y or X cables.

The following internal SAS adapters support the EXP24S:

- PCIe3 12 GB Cache RAID SAS Adapter Quad-port 6Gb x8 (#EJ0L, CCIN 57CE)
- PCIe3 RAID SAS Adapter Quad-port 6Gb x8 (#EL59)

The SAS disk drives that are contained in the EXP24S SFF Gen2-bay Drawer are controlled by one or two PCIe SAS adapters that are connected to the EXP24S through SAS cables. The SAS cable varies, depending on the adapter being used, the OS being used, and the protection you want.

The EXP24S SFF Gen2-bay drawer can be ordered in one of three possible mode settings, which are configured by manufacturing (not customer set up), of 1, 2, or 4 sets of disk bays.

With Linux, the EXP24S drawer can be ordered with four sets of six bays (mode 4), two sets of 12 bays (mode 2), or one set of 24 bays (mode 1).

There are six SAS connectors at the rear of the EXP24S drawer to which SAS adapters or controllers are attached. They are labeled T1, T2, and T3; there are two T1, two T2, and two T3 connectors. Figure 2-20 shows the rear connectors of the EXP24S drawer:

- In mode 1, two or four of the six ports are used. Two T2 ports are used for a single SAS adapter, and two T2 and two T3 ports are used with a paired set of two adapters or dual adapters configuration.
- In mode 2 or mode 4, four ports are used, two T2 and two T3, to access all SAS bays.

![Figure 2-20  EXP24S SFF Gen2-bay Drawer rear connectors](image)

An EXP24S drawer in mode 4 can be attached to two or four SAS controllers and provide high configuration flexibility. An EXP24S in mode 2 has similar flexibility. Up to 24 HDDs can be supported by any of the supported SAS adapters or controllers.

Any EXP24S order includes the EXP24S drawer no-charge specify codes to indicate to IBM manufacturing the mode to which the drawer should be set. The drawer is delivered with this configuration. If multiple EXP24S drawers are ordered, mixing modes should not be within that order. There is no externally visible indicator regarding the drawer's mode.
2.9 Hardware Management Console (optional)

The Hardware Management Console (HMC) is a dedicated appliance that allows administrators to configure and manage system resources on IBM Power Systems servers that use IBM POWER6, POWER6+, POWER7, POWER7+, and POWER8 processors and the PowerVM hypervisor. The HMC provides basic virtualization management support for configuring LPARs and dynamic resource allocation, including processor and memory settings for selected Power Systems servers.

The HMC also supports advanced service functions, including guided repair and verification, concurrent firmware updates for managed systems, and around-the-clock error reporting through IBM Electronic Service Agent™ for faster support.

The HMC management features help improve server usage, simplify systems management, and accelerate provisioning of server resources by using the PowerVM virtualization technology.

**Requirements:** When you use the HMC with the non-GPU Power S824L servers, the HMC code must be running at V8R8.3.0 level, or later.
The Power S824L platforms support two main service environments:

- Attachment to one or more HMCs. This environment is the common configuration for managing the server.
- No HMC attachment. Systems that are not attached to an HMC have a single partition that owns all the server resources and only one operating system can be installed. Hardware support for client-replaceable units is standard with the HMC. In addition, users can upgrade this support level to IBM onsite support to be consistent with other Power Systems servers.

### 2.9.1 HMC code level

If you are attaching an HMC to a new server or adding a function to an existing server that requires a firmware update, the HMC machine code might need to be updated to support the firmware level of the server. In a dual HMC configuration, both HMCs must be at the same version and release of the HMC code.

To determine the HMC machine code level that is required for the firmware level on any server, go to the following website to access the Fix Level Recommendation Tool (FLRT) on or after the planned availability date for this product:


FLRT identifies the correct HMC machine code for the selected system firmware level.

**Note:** Access to firmware and machine code updates is conditional on entitlement and license validation in accordance with IBM policy and practice. IBM can verify entitlement through client number, serial number electronic restrictions, or any other means or methods that are employed by IBM at its discretion.

### 2.9.2 HMC RAID 1 support

HMCs now offer a high-availability feature. The 7042-CR8, and CR9 by default, includes two HDDs with RAID 1 configured. RAID 1 is also offered on the 7042-CR6, 7042-CR7, and 7042-CR8 (if the feature was removed from the initial order) as a miscellaneous equipment specification (MES) upgrade option.

RAID 1 uses data mirroring. Two physical drives are combined into an array, and the same data is written to both drives. This makes the drives a **mirror image** of each other. If one of the drives experiences a failure, it is taken offline and the HMC continues operating with the other drive.

To use an existing HMC to manage any POWER8 processor-based server, the HMC must be a model CR5, or later, rack-mounted HMC, or model C08, or later, deskside HMC. The latest HMC model is the 7042-CR9. For your reference, Table 2-25 on page 64 lists a comparison between the 7042-CR8 and the 7042-CR9 HMC models.

**Note:** The 7042-CR9 ships with 16 GB of memory, and is expandable to 192 GB with an upgrade feature. 16 GB is advised for large environments or where external utilities, such as PowerVC and other third party monitors, are to be implemented.
Table 2-25 Comparison between 7042-CR8 and 7042-CR9 models

<table>
<thead>
<tr>
<th>Feature</th>
<th>CR8</th>
<th>CR9</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM System x model</td>
<td>x3550 M4 7914 PCH</td>
<td>x3550 M5 5463 AC1</td>
</tr>
<tr>
<td>HMC model</td>
<td>7042-CR8</td>
<td>7042-CR9</td>
</tr>
<tr>
<td>Processor</td>
<td>Intel 8-Core Xeon v2 2.00 GHz</td>
<td>Intel 18-core Xeon v3 2.4 GHz</td>
</tr>
<tr>
<td>Memory max:</td>
<td>16 GB (when featured)</td>
<td>16 GB DDR4 expandable to 192 GB</td>
</tr>
<tr>
<td>DASD</td>
<td>500 GB</td>
<td>500 GB</td>
</tr>
<tr>
<td>RAID 1</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>USB ports</td>
<td>Two front, four back</td>
<td>Two front, four rear</td>
</tr>
<tr>
<td>Integrated network</td>
<td>Four 1 Gb Ethernet</td>
<td>Four 1 Gb Ethernet</td>
</tr>
<tr>
<td>I/O slots</td>
<td>One PCI Express 3.0 slot</td>
<td>One PCI Express 3.0 slot</td>
</tr>
</tbody>
</table>

Notes: The two service processor HMC ports have the following features:
- Run at a speed of 1 Gbps
- Are visible only to the service processor and can be used to attach the server to an HMC or to access the ASMI options from a client directly from a client web browser
- Use the following network configuration if no IP addresses are set:
  - Service processor eth0 (HMC1 port): 169.254.2.147 with netmask 255.255.255.0
  - Service processor eth1 (HMC2 port): 169.254.3.147 with netmask 255.255.255.0

For more information about the service processor, see “Service processor” on page 94.

2.9.3 HMC connectivity to the POWER8 processor-based systems

POWER8 processor-based servers, and their predecessor systems, that are managed by an HMC require Ethernet connectivity between the HMC and the server's service processor.

For the HMC to communicate correctly with the managed server, eth0 of the HMC must be connected to either the HMC1 or HMC2 ports of the managed server, although other network configurations are possible. You can attach a second HMC to the remaining HMC port of the server for redundancy. The two HMC ports must be addressed by two separate subnets.

By default, the service processor HMC ports are configured for dynamic IP address allocation. The HMC can be configured as a Dynamic Host Configuration Protocol (DHCP) server, providing an IP address at the time that the managed server is powered on. In this case, the flexible service processor (FSP) is allocated an IP address from a set of address ranges that are predefined in the HMC software.

If the service processor of the managed server does not receive a DHCP reply before timeout, predefined IP addresses are set up on both ports. Static IP address allocation is also an option and can be configured using the ASMI menus.

2.9.4 High availability HMC configuration

The HMC is an important hardware component. Although Power Systems servers and their hosted partitions can continue to operate when the managing HMC becomes unavailable, certain operations cannot be performed without the HMC. To avoid these situations, consider...
installing a second HMC, in a redundant configuration, to be available when the other HMC is not (during maintenance, for example).

To achieve HMC redundancy for a POWER8 processor-based server, the server must be connected to two HMCs:

- The HMCs must be running the same level of HMC code.
- The HMCs must use different subnets to connect to the service processor.
- The HMCs must be able to communicate with the server's partitions over a public network to allow for full synchronization and functionality.

Figure 2-21 shows one possible highly available HMC configuration that is managing two servers. Each HMC is connected to one flexible service processor (FSP) port of each managed server.

For simplicity, only the hardware management networks (LAN1 and LAN2) are highly available (Figure 2-21). However, the open network (LAN3) can be made highly available by using a similar concept and adding a second network between the partitions and HMCs.

For more information about redundant HMCs, see the *IBM Power Systems HMC Implementation and Usage Guide*, SG24-7491.

### 2.10 Operating system support

Linux is an open source, cross-platform operating system that runs on numerous platforms from embedded systems to mainframe computers. It provides a UNIX like implementation across many computer architectures.

PowerVM supports Big Endian (BE) and Little Endian (LE) mode.
Table 2-26 lists the supported versions of Linux on the Power S824L server.

Table 2-26 Linux support matrix

<table>
<thead>
<tr>
<th>System</th>
<th>Type of virtualization</th>
<th>PowerVM</th>
<th>KVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power S824L with NVIDIA GPU</td>
<td>Bare-metal(^a)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>▶ Ubuntu 14.04.02+</td>
<td>▶ n/a</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td>▶ Ubuntu 14.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>▶ Ubuntu 15.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>▶ Ubuntu 15.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>▶ Red Hat Enterprise</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Linux 7.2+</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Power S824L with no NVIDIA GPU</td>
<td>Bare-metal(^a)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>▶ Ubuntu 14.04.02+</td>
<td>▶ Big Endian (BE):</td>
<td></td>
</tr>
<tr>
<td></td>
<td>▶ Ubuntu 14.10</td>
<td>▶ Red Hat Enterprise</td>
<td></td>
</tr>
<tr>
<td></td>
<td>▶ Ubuntu 15.04</td>
<td>Linux 6.5, 7.0, or</td>
<td></td>
</tr>
<tr>
<td></td>
<td>▶ Ubuntu 15.10</td>
<td>later</td>
<td></td>
</tr>
<tr>
<td></td>
<td>▶ Red Hat Enterprise</td>
<td>▶ SUSE Linux</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Linux 7.2+</td>
<td>Enterprise Server</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>11 Service Pack 3,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>or later</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>▶ Ubuntu 15.04</td>
<td>▶ Red Hat Enterprise</td>
<td></td>
</tr>
<tr>
<td></td>
<td>▶ Ubuntu 15.10</td>
<td>Linux 7.1, or later</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>▶ SUSE Linux</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Enterprise Server</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>12, or later</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>▶ Ubuntu 15.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>▶ Ubuntu 15.10</td>
<td></td>
</tr>
</tbody>
</table>

\(^{a}\) Not supported with a EXP24S external disk drawer and not with a external PCIe Gen3 I/O drawer

Linux supports almost all of the Power System I/O and the configurator verifies support on order.

For more information about the software that is available on IBM Power Systems, see the IBM Power Systems Software™ website:


Be sure to connect your systems to the IBM Service and Productivity Tools for PowerLinux™ repository and keep up-to-date with the latest Linux service and productivity tools, which are available from IBM at the following website:


For information about the PowerLinux Community, see the following website:

https://www.ibm.com/developerworks/group/tpl

For information about the features and external devices that are supported by Linux, see the following website:


For more information about SUSE Linux Enterprise Server, see the following website:

http://www.novell.com/products/server
For more information about Red Hat Enterprise Linux Advanced Server, see the following website:
http://www.redhat.com/rhel/features

For more information about Ubuntu Server, see the following website:
http://www.ubuntu.com/server

2.10.1 Java

When running Java applications on the POWER8 processor, the pre-packaged Java that is part of a Linux distribution is designed to meet the most common requirements.

If you require a different level of Java, there are several resources available.

Current information about IBM Java and tested Linux distributions are available here:

Additional information about the OpenJDK port for Linux on PPC64 LE, as well as some pre-generated builds can be found here:
http://cr.openjdk.java.net/~simonis/ppc-aix-port/

Launchpad.net has resources for Ubuntu builds. You can find out about them here:
https://launchpad.net/ubuntu/+source/openjdk-9
https://launchpad.net/ubuntu/+source/openjdk-8
https://launchpad.net/ubuntu/+source/openjdk-7

2.11 Energy management

The Power S824L systems have features to help clients become more energy efficient. EnergyScale technology enables advanced energy management features to conserve power dramatically and dynamically and further improve energy efficiency. Intelligent Energy optimization capabilities enable the POWER8 processor to operate at a higher frequency for increased performance and performance per watt, or to dramatically reduce frequency to save energy.

2.11.1 IBM EnergyScale technology

IBM EnergyScale technology provides functions to help the user understand and dynamically optimize processor performance versus processor energy consumption, and system workload, to control IBM Power Systems power and cooling usage.

EnergyScale uses power and thermal information that is collected from the system to implement policies that can lead to better performance or better energy usage. IBM EnergyScale offers the following features:

► Power trending

EnergyScale provides continuous collection of real-time server energy consumption. It enables administrators to predict power consumption across their infrastructure and to react to business and processing needs. For example, administrators can use this
information to predict data center energy consumption at various times of the day, week, or month.

- **Power saver mode**
  Power saver mode lowers the processor frequency and voltage on a fixed amount, reducing the energy consumption of the system while still delivering predictable performance. This percentage is predetermined to be within a safe operating limit and is not user configurable. The server is designed for a fixed frequency drop of almost 50% down from nominal frequency (the actual value depends on the server type and configuration).

  Power saver mode is not supported during system start, although it is a persistent condition that is sustained after the start when the system starts running instructions.

- **Dynamic power saver mode**
  Dynamic power saver mode varies processor frequency and voltage based on the usage of the POWER8 processors. Processor frequency and usage are inversely proportional for most workloads, implying that as the frequency of a processor increases, its usage decreases, given a constant workload. Dynamic power saver mode takes advantage of this relationship to detect opportunities to save power, based on measured real-time system usage.

  When a system is idle, the system firmware lowers the frequency and voltage to power energy saver mode values. When fully used, the maximum frequency varies, depending on whether the user favors power savings or system performance. If an administrator prefers energy savings and a system is fully used, the system is designed to reduce the maximum frequency to about 95% of nominal values. If performance is favored over energy consumption, the maximum frequency can be increased up to 111.3% of nominal frequency for extra performance.

  Dynamic power saver mode is mutually exclusive with power saver mode. Only one of these modes can be enabled at a time.

- **Power capping**
  Power capping enforces a user-specified limit on power usage. Power capping is not a power-saving mechanism. It enforces power caps by throttling the processors in the system, degrading performance significantly. The idea of a power cap is to set a limit that must never be reached but that frees extra power that was never used in the data center. The *margined* power is this amount of extra power that is allocated to a server during its installation in a data center. It is based on the server environmental specifications that usually are never reached because server specifications are always based on maximum configurations and worst-case scenarios.

- **Soft power capping**
  There are two power ranges into which the power cap can be set: power capping, as described previously, and soft power capping. Soft power capping extends the allowed energy capping range further, beyond a region that can be ensured in all configurations and conditions. If the energy management goal is to meet a particular consumption limit, soft power capping is the mechanism to use.
Processor core nap mode

IBM POWER8 processor uses a low-power mode that is called nap that stops processor execution when there is no work to do on that processor core. The latency of exiting nap mode is small, typically not generating any impact on applications that are running. Therefore, the IBM POWER Hypervisor™ can use nap mode as a general-purpose idle state. When the operating system detects that a processor thread is idle, it yields control of a hardware thread to the POWER Hypervisor. The POWER Hypervisor immediately puts the thread into nap mode. Nap mode allows the hardware to turn off the clock on most of the circuits in the processor core. Reducing active energy consumption by turning off the clocks allows the temperature to fall, which further reduces leakage (static) power of the circuits and causes a cumulative effect. Nap mode saves 10 - 15% of power consumption in the processor core.

Processor core sleep mode

To save even more energy, the POWER8 processor has an even lower power mode referred to as sleep. Before a core and its associated private L2 cache enter sleep mode, the cache is flushed, transition lookaside buffers (TLB) are invalidated, and the hardware clock is turned off in the core and in the cache. Voltage is reduced to minimize leakage current. Processor cores that are inactive in the system (such as capacity on demand (CoD) processor cores) are kept in sleep mode. Sleep mode saves about 80% of the power consumption in the processor core and its associated private L2 cache.

Processor chip winkle mode

The most energy can be saved when a whole POWER8 chiplet enters the winkle mode. In this mode, the entire chiplet is turned off, including the L3 cache. This mode can save more than 95% power consumption.

Fan control and altitude input

System firmware dynamically adjusts fan speed based on energy consumption, altitude, ambient temperature, and energy savings modes. Power Systems are designed to operate in worst-case environments, in hot ambient temperatures, at high altitudes, and with high-power components. In a typical case, one or more of these constraints are not valid. When no power savings setting is enabled, fan speed is based on ambient temperature and assumes a high-altitude environment. When a power savings setting is enforced (either Power Energy Saver Mode or Dynamic Power Saver Mode), the fan speed varies based on power consumption and ambient temperature.

Processor folding

Processor folding is a consolidation technique that dynamically adjusts, over the short term, the number of processors that are available for dispatch to match the number of processors that are demanded by the workload. As the workload increases, the number of processors made available increases. As the workload decreases, the number of processors that are made available decreases. Processor folding increases energy savings during periods of low to moderate workload because unavailable processors remain in low-power idle states (nap or sleep) longer.

EnergyScale for I/O

IBM POWER8 processor-based systems automatically power off hot-pluggable PCI adapter slots that are empty or not being used. System firmware automatically scans all pluggable PCI slots at regular intervals, looking for those slots that meet the criteria for being not in use and powering them off. This support is available for all POWER8 processor-based servers and the expansion units that they support.
Server power down

If overall data center processor usage is low, workloads can be consolidated on fewer servers so that certain servers can be turned off completely. Consolidation makes sense when there are long periods of low usage, such as weekends. Live Partition Mobility can be used to move workloads to consolidate partitions onto fewer systems, reducing the number of servers that are powered on and therefore reducing the power usage.

On POWER8 processor-based systems, several EnergyScale technologies are embedded in the hardware and do not require an operating system or external management component. Fan control, environmental monitoring, and system energy management are controlled by the On Chip Controller (OCC) and associated components. The power mode can also be set up without external tools, by using the ASMI interface, as shown in Figure 2-22.

![Figure 2-22 Setting the power mode in ASMI](image)

2.11.2 On Chip Controller

To maintain the power dissipation of POWER7+ with its large increase in performance and bandwidth, POWER8 invested significantly in power management innovations. A new On Chip Controller (OCC) that uses an embedded IBM PowerPC® core with 512 KB of SRAM runs real-time control firmware to respond to workload variations by adjusting the per-core frequency and voltage based on activity, thermal, voltage, and current sensors.
The on-die nature of the OCC allows for approximately 100× speedup in response to workload changes over POWER7+, enabling reaction under the timescale of a typical OS time slice and allowing for multi-socket, scalable systems to be supported. It also enables more granularity in controlling the energy parameters in the processor, and increases reliability in energy management by having one controller in each processor that can perform certain functions independently of the others.

POWER8 also includes an internal voltage regulation capability that enables each core to run at a different voltage. Optimizing both voltage and frequency for workload variation enables a better increase in power savings versus optimizing frequency only.

2.11.3 Energy consumption estimation

Often, for Power Systems, various energy-related values are important:

- Maximum power consumption and power source loading values
  
  These values are important for site planning and are described in the POWER8 systems information IBM Knowledge Center at the following website:
  
  http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/index.jsp

  Search for type and model number and “server specifications”. For example, for the Power S14 and Power 824 system, search for “8286-41A” or “8286-42A server specifications”.

- An estimation of the energy consumption for a certain configuration

  Calculate the energy consumption for a certain configuration in the IBM Systems Energy Estimator at the following website:
  
  http://www-912.ibm.com/see/EnergyEstimator

  In that tool, select the type and model for the system, and enter details about the configuration and CPU usage that you want. As a result, the tool shows the estimated energy consumption and the waste heat at the usage that you want and also at full usage.
Reliability, availability, and serviceability

This chapter provides information about IBM Power Systems reliability, availability, and serviceability (RAS) design and features.

The elements of RAS are described:

- **Reliability**: Indicates how infrequently a defect or fault in a server occurs.
- **Availability**: Indicates how infrequently the functioning of a system or application is affected by a fault or defect.
- **Serviceability**: Indicates how well faults and their effects are communicated to system managers and how efficiently and nondisruptively the faults are repaired.
3.1 Introduction

The POWER8 processor-based servers are available in two classes:

- **Scale-out systems**: For environments consisting of multiple systems working in concert. In these environments, application availability is enhanced by the superior availability characteristics of each system.

- **Enterprise systems**: For environments requiring systems with increased availability. In these environments, mission-critical applications can take full advantage of the scale-up characteristics, increased performance, flexibility, and enterprise availability characteristics of each system.

One key differentiator of the IBM POWER8 processor-based servers is that they use all the advanced RAS characteristics of the POWER8 processor through the whole portfolio, offering reliability and availability features that often are not seen in other scale-out servers. Several of these features are improvements for POWER8 or features that were previously only in higher-end Power Systems.

The POWER8 processor modules support an enterprise level of reliability and availability. The processor design has extensive error detection and fault isolation (ED/FI) capabilities to allow for a precise analysis of faults, whether they are hard or soft. They use advanced technology, including stacked latches and Silicon-on-Insulator (SOI) technology, to reduce susceptibility to soft errors, and advanced design features within the processor for correction or try again after soft error events. In addition, the design incorporates spare capacity that is integrated into many elements to tolerate certain faults without requiring an outage or parts replacement. Advanced availability techniques are used to mitigate the impact of other faults that are not directly correctable in the hardware.

Features within the processor and throughout systems are incorporated to support design verification. During the design and development process, subsystems go through rigorous verification and integration testing processes by using these features. During system manufacturing, systems go through a thorough testing process to help ensure high product quality levels, again taking advantage of the designed ED/FI capabilities.

Fault isolation and recovery of the POWER8 processor and memory subsystems are designed to use a dedicated service processor and are meant to be largely independent of any operating system or application that is deployed.

The Power S824L POWER8 processor-based servers are designed to support a “scale-out” system environment that consists of multiple systems working in concert. In these environments, application availability is enhanced by the superior availability characteristics of each system.

3.1.1 RAS enhancements of POWER8 processor-based servers

Several features are included in the entire portfolio of the POWER8 processor-based servers. Many of these features are improvements for POWER8 or features that were previously only in higher-end Power Systems, using a higher RAS even for scale-out equipment.
These features are summarized:

- **Processor Enhancements Integration**
  
  POWER8 processor chips are implemented using 22 nm technology and integrated onto SOI modules.

  The processor design now supports a spare data lane on each fabric bus, which is used to communicate between processor modules. A spare data lane can be substituted for a failing data lane dynamically during system operation.

  A POWER8 processor module offers improved performance compared to POWER7+, including support of a maximum of 12 cores compared to a maximum of eight cores in POWER7+. Performing more work with less hardware in a system provides greater reliability, by concentrating the processing power and reducing the need for additional communication fabrics and components.

  The processor module integrates a new On Chip Controller (OCC). This OCC is used to handle power management and thermal monitoring without the need for a separate controller, which was required in POWER7+. In addition, the OCC can also be programmed to run other RAS-related functions independently of any host processor.

  The memory controller within the processor is redesigned. From a RAS standpoint, the ability to use a replay buffer to recover from soft errors is added.

- **I/O Subsystem**
  
  The POWER8 processor now integrates PCIe controllers. PCIe slots that are directly driven by PCIe controllers can be used to support I/O adapters directly in the systems or used to attach external I/O drawers. For greater I/O capacity, the POWER8 processor-based scale-out servers also support a PCIe switch to provide additional integrated I/O capacity.

  These integrated I/O adapters can be repaired in these servers concurrently, which is an improvement over comparable POWER7/7+ scale-out systems that did not allow an adapter hot-plug. Enterprise systems have offered this technology since POWER6 was introduced.

- **Memory Subsystem**
  
  Custom dual inline memory modules (DIMMs) (CDIMMS) are used to correct a single dynamic random access memory (DRAM) fault within an error-correcting code (ECC) word (and then an additional bit fault) to avoid unplanned outages. They also contain a spare DRAM module per port (per nine DRAMs for x8 DIMMs), which can be used to avoid replacing memory.

### 3.2 Reliability

Highly reliable systems are built with highly reliable components. On IBM POWER processor-based systems, this basic principle is expanded upon with a clear design for reliability architecture and methodology. A concentrated, systematic, and architecture-based approach is designed to improve overall system reliability with each successive generation of system offerings. Reliability can be improved in primarily three ways:

- Reducing the number of components
- Using higher reliability grade parts
- Reducing the stress on the components

In the POWER8 systems, elements of all three methods are used to improve system reliability.
3.2.1 Designed for reliability

Systems that are designed with fewer components and interconnects have fewer opportunities to fail. Simple design choices, such as integrating processor cores on a single POWER chip, can reduce the opportunity for system failures. The POWER8 chip has more cores per processor module, and the I/O Hub Controller function is integrated in the processor module, which generates a PCIe BUS directly from the Processor module. Parts selection also plays a critical role in overall system reliability.

IBM uses stringent design criteria to select server grade components that are extensively tested and qualified to meet and exceed a minimum design life of seven years. By selecting higher reliability grade components, the frequency of all failures is lowered, and wear-out is not expected within the operating system life. Component failure rates can be further improved by burning in select components or running the system before shipping it to the client. This period of high stress removes the weaker components with higher failure rates, that is, it cuts off the front end of the traditional failure rate bathtub curve (see Figure 3-1).

![Failure rate bathtub curve](image)

3.2.2 Placement of components

Packaging is designed to deliver both high performance and high reliability. For example, the reliability of electronic components is directly related to their thermal environment. Large decreases in component reliability are directly correlated to relatively small increases in temperature. All POWER processor-based systems are packaged to ensure adequate cooling. Critical system components, such as the POWER8 processor chips, are positioned on the system board so that they receive clear air flow during operation. POWER8 systems use a premium fan with an extended life to further reduce the overall system failure rate and provide adequate cooling for the critical system components.
3.3 Processor and memory availability details

The more reliable a system or subsystem is, the more available it can be. Nevertheless, considerable effort is made to design systems that can detect faults that occur and take steps to minimize or eliminate the outages that are associated with them. These design capabilities extend availability beyond what can be obtained through the underlying reliability of the hardware.

This design for availability begins with implementing an architecture for ED/FI.

First-failure data capture (FFDC) is the capability of IBM hardware and microcode to continuously monitor hardware functions. Within the processor and memory subsystem, detailed monitoring is done by circuits within the hardware components themselves. Fault information is gathered into fault isolation registers (FIRs) and reported to the appropriate components for handling.

Processor and memory errors that are recoverable in nature are typically reported to the dedicated service processor that is built into each system. The dedicated service processor then works with the hardware to determine the course of action to take for each fault.

3.3.1 Correctable error introduction

Intermittent or soft errors are typically tolerated within the hardware design by using error correction code (ECC) or advanced techniques to try operations again after a fault.

Tolerating a correctable solid fault runs the risk that the fault aligns with a soft error and causes an uncorrectable error situation. Also, the risk exists that a correctable error is predictive of a fault that continues to worsen over time, resulting in an uncorrectable error condition.

You can predictively deallocate a component (to prevent correctable errors from aligning with soft errors or other hardware faults and causing uncorrectable errors) to avoid these situations. However, unconfiguring components, such as processor cores or entire caches in memory, can reduce the performance or capacity of a system. Unconfiguring components in turn typically requires that the failing hardware is replaced in the system. The resulting service action can also temporarily affect system availability.

To avoid these situations in solid faults in POWER8, processors or memory might be candidates for correction by using the “self-healing” features that are built into the hardware, such as taking advantage of a spare DRAM module within a memory DIMM, a spare data lane on a processor or memory bus, or spare capacity within a cache module.

When this self-healing is successful, the need to replace any hardware for a solid correctable fault is avoided. The ability to predictively unconfigure a processor core is still available for faults that cannot be repaired by self-healing techniques or because the sparing or self-healing capacity is exhausted.
3.3.2 Uncorrectable error introduction

An uncorrectable error can be defined as a fault that can cause incorrect instruction execution within logic functions, or an uncorrectable error in data that is stored in caches, registers, or other data structures. In less sophisticated designs, a detected uncorrectable error nearly always results in the termination of an entire system. More advanced system designs in certain cases might be able to terminate just the application by using the hardware that failed. These designs might require that uncorrectable errors are detected by the hardware and reported to software layers, and the software layers must then be responsible for determining how to minimize the impact of faults.

The advanced RAS features that are built into POWER8 processor-based systems handle certain “uncorrectable” errors in ways that minimize the impact of the faults, even keeping an entire system up and running after experiencing a failure.

Depending on the fault, this recovery can use the virtualization capabilities of PowerVM in a way that the operating system or any applications that are running in the system are not affected or must participate in the recovery.

3.3.3 Processor core/cache correctable error handling

Layer 2 (L2) and Layer 3 (L3) caches and directories can correct single bit errors (single-error correcting (SEC)) and detect double bit errors (double-error detecting (DED)), which is known as SEC/DED error correction code (ECC) protection. Soft errors that are detected in the level 1 caches are also correctable by a try again operation that is handled by the hardware. Internal and external processor “fabric” buses have SEC/DED ECC protection, as well.

SEC/DED capabilities are also included in other data arrays that are not directly visible to clients.

Beyond soft error correction, the intent of the POWER8 design is to manage a solid correctable error in an L2 or L3 cache by using techniques to delete a cache line with a persistent issue, or to repair a column of an L3 cache dynamically by using spare capability.

Information about column and row repair operations is stored persistently for processors so that more permanent repairs can be made during processor reinitialization (during system reboot, or individual Core Power on Reset using the Power On Reset Engine).

3.3.4 Processor Instruction Retry and other try again techniques

Within the processor core, soft error events might occur that interfere with the various computation units. When an event can be detected before a failing instruction is completed, the processor hardware might be able to try the operation again by using the advanced RAS feature that is known as Processor Instruction Retry.

Processor Instruction Retry allows the system to recover from soft faults that otherwise result in outages of applications or the entire server.

Try again techniques are used in other parts of the system, as well. Faults that are detected on the memory bus that connects processor memory controllers to DIMMs can be tried again. In POWER8 systems, the memory controller is designed with a replay buffer that allows memory transactions to be tried again after certain faults internal to the memory controller faults are detected. This replay buffer complements the try again capabilities of the memory buffer module.
3.3.5 Alternative processor recovery and Partition Availability Priority

If Processor Instruction Retry for a fault within a core occurs multiple times without success, the fault is considered to be a solid failure. In certain instances, PowerVM can work with the processor hardware to migrate a workload running on the failing processor to a spare or alternative processor. This migration is accomplished by migrating the pertinent processor core state from one core to another core with the new core taking over at the instruction that failed on the faulty core. Successful migration keeps the application running during the migration without needing to terminate the failing application.

Successful migration requires sufficient available spare capacity to reduce the overall processing capacity within the system by one processor core. Typically, in highly virtualized environments, the requirements of partitions can be reduced to accomplish this task without any further impact to running applications.

In systems without sufficient reserve capacity, it might be necessary to terminate at least one partition to free resources for the migration. In advance, PowerVM users can identify the partitions with the highest priority. When you use this Partition Priority feature of PowerVM, if a partition must be terminated for alternative processor recovery to complete, the system can terminate lower priority partitions to keep the higher priority partitions up and running, even when an unrecoverable error occurred on a core running the highest priority workload.

Partition Availability Priority is assigned to partitions by using a weight value or integer rating. The lowest priority partition is rated at 0 (zero) and the highest priority partition is rated at 255. The default value is set to 127 for standard partitions and 192 for Virtual I/O Server (VIOS) partitions. Priorities can be modified through the Hardware Management Console (HMC).

3.3.6 Core Contained Checkstops and other PowerVM error recovery

PowerVM can handle certain other hardware faults without terminating applications, such as an error in certain data structures (faults in translation tables or lookaside buffers).

Other core hardware faults that alternative processor recovery or Processor Instruction Retry cannot contain might be handled in PowerVM by a technique called Core Contained Checkstops. This technique allows PowerVM to be signaled when these faults occur and terminate code that is in use by the failing processor core (typically just a partition, although potentially PowerVM if the failing instruction is in a critical area of PowerVM code).

Processor designs without Processor Instruction Retry typically must resort to techniques for all faults that can be contained to an instruction in a processor core.

3.3.7 Cache uncorrectable error handling

If a fault within a cache occurs that cannot be corrected with SEC/DED ECC, the faulty cache element is unconfigured from the system, typically, by purging and deleting a single cache line. These purge and delete operations are contained within the hardware, and they prevent a faulty cache line from being reused and causing multiple errors.

During the cache purge operation, the data that is stored in the cache line is corrected where possible. If correction is not possible, the associated cache line is marked with a special ECC that indicates that the cache line has bad data.
Nothing within the system terminates just because this type of an event is encountered. Rather, the hardware monitors the usage of pages with marks. If this data is never used, hardware replacement is requested, but nothing terminates as a result of the operation. Software layers are not required to handle these faults.

Only when data is loaded to be processed by a processor core, or sent out to an I/O adapter, is any further action needed. In these cases, if data is used as owned by a partition, the partition operating system might be responsible for terminating itself or just the program using the marked page. If data is owned by the hypervisor, the hypervisor might choose to terminate, resulting in a system-wide outage.

However, the exposure to these events is minimized because cache lines can be deleted. The deletion eliminates the repetition of an uncorrectable fault that is in a particular cache line.

### 3.3.8 Other processor chip functions

Functions other than processor cores are available within a processor chip.

POWER8 processors have built-in accelerators that can be used as application resources to handle functions, such as random number generation. POWER8 also introduces a controller for attaching cache-coherent adapters that are external to the processor module. The POWER8 design contains a function to freeze the function that is associated with several of these elements, without taking a system-wide checkstop. Depending on the code that uses these features, a freeze event might be handled without an application or partition outage.

Single bit errors, even solid faults, within internal or external processor fabric buses, are corrected by the ECC that is used. POWER8 processor-to-processor module fabric buses also use a spare data lane so that a single failure can be repaired without calling for the replacement of hardware.

### 3.3.9 Other fault error handling

Not all processor module faults can be corrected by these techniques. Therefore, a provision is still made for certain faults that cause a system-wide outage. In a platform checkstop event, the ED/FI capabilities that are built into the hardware and dedicated service processor work to isolate the root cause of the checkstop. The ED/FI capabilities then unconfigure the faulty element where possible so that the system can reboot with the failed component unconfigured from the system.

The auto-restart (reboot) option, when enabled, can restart the system automatically following an unrecoverable firmware error, firmware stop, hardware failure, or environmentally induced (AC power) failure.

The auto-restart (reboot) option must be enabled from the Advanced System Management Interface (ASMI) or from the Control (Operator) Panel.
3.3.10 Memory protection

POWER8 processor-based systems have a three-part memory subsystem design. This design consists of two memory controllers in each processor module. The two memory controllers communicate to buffer modules on memory DIMMS through memory channels and access the DRAM memory modules on DIMMs, as shown in Figure 3-2.

<table>
<thead>
<tr>
<th>Memory Controller</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supports 128 Byte Cache Line</td>
</tr>
<tr>
<td>hardened “Stacked” Latches for Soft Error Protection</td>
</tr>
<tr>
<td>And retry buffer to retry after soft internal faults</td>
</tr>
<tr>
<td>Special Uncorrectable error handling for soft faults</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Memory Bus</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRC protection with recalibration and retry on error</td>
</tr>
<tr>
<td>Spare Data lane can be dynamically substituted for failed one</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Memory Buffer</th>
</tr>
</thead>
<tbody>
<tr>
<td>same technology as POWER8 Processor Chips</td>
</tr>
<tr>
<td>Hardened “Stacked” Latches for Soft Error Protection</td>
</tr>
<tr>
<td>Can retry after internal soft Errors</td>
</tr>
<tr>
<td>L4 Cache implemented in eDRAM</td>
</tr>
<tr>
<td>DED/SEC ECC Code</td>
</tr>
<tr>
<td>Persistent correctable error handling</td>
</tr>
</tbody>
</table>

16 GB DIMM

<table>
<thead>
<tr>
<th>4 Ports of Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>- 10 DRAMs x8 DRAM modules attached to each port</td>
</tr>
<tr>
<td>- 8 Modules Needed For Data</td>
</tr>
<tr>
<td>- 1 Needed For Error Correction Coding</td>
</tr>
<tr>
<td>- 1 Additional Spare</td>
</tr>
<tr>
<td>2 Ports are combined to form a 128 bit ECC word</td>
</tr>
<tr>
<td>- 8 Reads fill a processor cache</td>
</tr>
<tr>
<td>Second port can be used to fill a second cache line</td>
</tr>
<tr>
<td>(Much like having 2 DIMMs under one Memory Buffer but housed in the same physical DIMM)</td>
</tr>
</tbody>
</table>

Figure 3-2 Memory protection features

The memory buffer chip is made by the same 22 nm technology that is used to make the POWER8 processor chip, and the memory buffer chip incorporates the same features in the technology to avoid soft errors. It implements a try again for many internally detected faults. This function complements a replay buffer in the memory controller in the processor, which also handles internally detected soft errors.

The bus between a processor memory controller and a DIMM uses cyclic redundancy check (CRC) error detection that is coupled with the ability to try soft errors again. The bus features dynamic recalibration capabilities plus a spare data lane that can be substituted for a failing bus lane through the recalibration process.

The buffer module implements an integrated L4 cache using eDRAM technology (with soft error hardening) and persistent error handling features.

The memory buffer on each DIMM has four ports for communicating with DRAM modules. The 16 GB DIMM, for example, has one rank that is composed of four ports of x8 DRAM modules, with each port containing 10 DRAM modules.

For each of these ports, there are eight DRAM modules worth of data (64 bits) plus another DRAM module’s worth of error correction and other data. There is also a spare DRAM module for each port that can be substituted for a failing port.
Two ports are combined into an ECC word and supply 128 bits of data. The ECC that is deployed can correct the result of an entire DRAM module that is faulty. (This correction is also known as Chipkill correction). Then, it can correct at least an additional bit within the ECC word.

The additional spare DRAM modules are used when a DIMM experiences a Chipkill event within the DRAM modules under a port. The spare DRAM module can be substituted for a failing module, avoiding the need to replace the DIMM for a single Chipkill event.

Depending on how DRAM modules fail, it might be possible to tolerate up to four DRAM modules failing on a single DIMM without needing to replace the DIMM. Then, you can still correct an additional DRAM module that is failing within the DIMM.

There are other DIMMs offered with these systems. A 32 GB DIMM has two ranks, where each rank is similar to the 16 GB DIMM with DRAM modules on four ports. Each port has 10 x8 DRAM modules.

In addition, a 64 GB DIMM is offered through x4 DRAM modules that are organized in four ranks.

In addition to the protection that is provided by the ECC and sparing capabilities, the memory subsystem also implements scrubbing of memory to identify and correct single bit soft errors. Hypervisors are informed of incidents of single-cell persistent (hard) faults for deallocation of associated pages. However, because of the ECC and sparing capabilities that are used, this memory page deallocation is not relied on for the repair of faulty hardware.

If an uncorrectable error in data is encountered, the memory that is affected is marked with a special uncorrectable error code and handled as described for cache uncorrectable errors.

3.3.11 I/O subsystem availability and Enhanced Error Handling

Use multi-path I/O and VIOS for I/O adapters and RAID for storage devices to prevent application outages when I/O adapter faults occur.

To recover soft or intermittent faults without failover to an alternative device or I/O path, Power Systems hardware supports Enhanced Error Handling (EEH) for I/O adapters and PCIe bus faults.

EEH allows EEH-aware device drivers to try again after certain non-fatal I/O events to avoid failover, especially in cases where a soft error is encountered. EEH also allows device drivers to terminate if there is an intermittent hard error or other unrecoverable errors, while protecting against reliance on data that cannot be corrected. This action typically is done by freezing access to the I/O subsystem with the fault. Freezing prevents data from flowing to and from an I/O adapter and causes the hardware/firmware to respond with a defined error signature whenever an attempt is made to access the device. If necessary, a special uncorrectable error code can be used to mark a section of data as bad when the freeze is first initiated.

In POWER8 processor-based systems, the external I/O hub and bridge adapters were eliminated in favor of a topology that integrates PCIe Host Bridges into the processor module. PCIe buses that are generated directly from a host bridge can drive individual I/O slots or a PCIe switch. The integrated PCIe controller supports try again (endpoint error recovery) and freezing.

For Linux under PowerVM, EEH support extends to many frequently used devices. Various third-party PCI devices might not provide native EEH support.
3.4 Enterprise systems availability details

In addition to the standard RAS features that are described, enterprise class systems offer increased RAS by including several features and redundant components.

The following major features are exclusive to the enterprise class systems:

- **Redundant Service Processor**

  The service processor is the main component of a system and responsible for the IPL, setup, monitoring, control, and management. The control units, which are on enterprise class systems house two redundant service processors. In a failure in any of the service processors, the second service processor allows for continued operation of the system until a replacement is scheduled. Even a system with a single system node has dual service processors on the control unit.

- **Redundant System Clock Cards**

  The system clock cards are critical components that are crucial to the system operations. They are responsible for providing synchronized clock signals for the whole system. The control units on enterprise class systems house two redundant system clock cards. In a failure in any of the clock cards, the second clock card allows for continued operation of the system until a replacement is scheduled. Even a system with a single system node has dual clock cards on the control unit.

- **Dynamic Processor Sparing**

  Enterprise class systems are Capacity Upgrade on Demand (CUoD) capable. Processor sparing helps minimize the impact to server performance that is caused by a failed processor. An inactive processor is activated if a failing processor reaches a predetermined error threshold, therefore helping to maintain performance and improve system availability. Dynamic processor sparing happens automatically when you use dynamic logical partitioning (DLPAR) and the failing processor is detected before it fails. Dynamic Processor Sparing does not require the purchase of an activation code; it requires available inactive CUoD processor cores on the system.

- **Dynamic Memory Sparing**

  Enterprise class systems are CUoD capable. Dynamic memory sparing helps minimize the impact to server performance that is caused by a failed memory feature. Memory sparing occurs when on-demand inactive memory is automatically activated by the system to temporarily replace failed memory until a service action can be performed.

- **Active Memory™ Mirroring for Hypervisor**

  The hypervisor is the core part of the virtualization layer. Although minimal, its operational data must reside in memory CDIMMs. In a failure of a CDIMM, the hypervisor can become inoperative. The Active Memory Mirroring for Hypervisor allows for the memory blocks that are used by the hypervisor to be written in two distinct CDIMMs. If an uncorrectable error is encountered during a read, the data is retrieved from the mirrored pair and operations continue normally.

3.5 Availability effects of a solution architecture

Any solution must not rely only on the hardware platform. Despite IBM Power Systems RAS superiority, it is advisable to design a redundant architecture surrounding the application to allow easier maintenance and greater flexibility.
By working in a redundant architecture, certain tasks that require that a specific application is offline can now execute with the application running, which offers even greater availability.

3.6 Serviceability

The purpose of serviceability is to repair the system while attempting to minimize or eliminate service cost (within budget objectives) and maintaining application availability and high client satisfaction. Serviceability includes system installation, miscellaneous equipment specification (MES) (system upgrades/downgrades), and system maintenance/repair. Depending on the system and warranty contract, service might be performed by the client, an IBM Service Support Representative (SSR), or an authorized warranty service provider.

The serviceability features that are delivered in this system provide a highly efficient service environment by incorporating the following attributes:
- Design for client setup (CSU), client-installed features (CIF), and client-replaceable units (CRU)
- ED/FI incorporating first-failure data capture (FFDC)
- Converged service approach across multiple IBM server platforms
- Concurrent Firmware Maintenance (CFM)

This section provides an overview of how these attributes contribute to efficient service in the progressive steps of error detection, analysis, reporting, notification, and repair in all POWER processor-based systems.

3.6.1 Detecting introduction

The first and most crucial component of a solid serviceability strategy is the ability to detect errors accurately and effectively when they occur.

Although not all errors are a guaranteed threat to system availability, errors that go undetected can cause problems because the system has no opportunity to evaluate and act if necessary. POWER processor-based systems employ IBM System z® server-inspired error detection mechanisms, extending from processor cores and memory to power supplies and hard disk drives (HDDs).

3.6.2 Error checkers, fault isolation registers, and First-Failure Data Capture

IBM POWER processor-based systems contain specialized hardware detection circuitry that is used to detect erroneous hardware operations. Error checking hardware ranges from parity error detection that is coupled with Processor Instruction Retry and bus try again, to ECC correction on caches and system buses.

Within the processor/memory subsystem error checker, error checker signals are captured and stored in hardware FIRs. The associated logic circuitry is used to limit the domain of an error to the first checker that encounters the error. In this way, runtime error diagnostic tests can be deterministic so that for every check station, the unique error domain for that checker is defined and mapped to field-replaceable units (FRUs) that can be repaired when necessary.
Integral to the Power Systems design is the concept of FFDC. FFDC is a technique that involves sufficient error checking stations and coordination of faults so that faults are detected and the root cause of the fault is isolated. FFDC also expects that necessary fault information can be collected at the time of failure without needing to re-create the problem or run an extended tracing or diagnostics program.

For most faults, a good FFDC design means that the root cause is isolated at the time of the failure without intervention by an IBM SSR. For all faults, good FFDC design still makes failure information available to the IBM SSR, and this information can be used to confirm the automatic diagnosis. More detailed information can be collected by an IBM SSR for rare cases where the automatic diagnosis is not adequate for fault isolation.

### 3.6.3 Service processor

In POWER8 processor-based systems with a dedicated service processor, the dedicated service processor is primarily responsible for the fault analysis of processor and memory errors.

The service processor is a microprocessor that is powered separately from the main instruction processing complex.

In addition to FFDC functions, the service processor performs many serviceability functions:

- Several remote power control options
- Reset and boot features
- Environmental monitoring

The service processor interfaces with the OCC function, which monitors the server’s built-in temperature sensors and sends instructions to the system fans to increase rotational speed when the ambient temperature is above the normal operating range. By using a designed operating system interface, the service processor notifies the operating system of potential environmentally related problems so that the system administrator can take corrective actions before a critical failure threshold is reached. The service processor can also post a warning and initiate an orderly system shutdown in the following circumstances:

- The operating temperature exceeds the critical level (for example, failure of air conditioning or air circulation around the system).
- The system fan speed is out of operational specification (for example, because of multiple fan failures).
- The server input voltages are out of operational specification. The service processor can shut down a system in the following circumstances:
  - The temperature exceeds the critical level or remains above the warning level for too long.
  - Internal component temperatures reach critical levels.
  - Non-redundant fan failures occur.
POWER Hypervisor (system firmware) and HMC connection surveillance

The service processor monitors the operation of the firmware during the boot process, and also monitors the hypervisor for termination. The hypervisor monitors the service processor and can perform a reset and reload if it detects the loss of the service processor. If the reset/reload operation does not correct the problem with the service processor, the hypervisor notifies the operating system, and then the operating system can act, including calling for service. The flexible service processor (FSP) also monitors the connection to the HMC and can report loss of connectivity to the operating system partitions for system administrator notification.

Uncorrectable error recovery

The auto-restart (reboot) option, when enabled, can reboot the system automatically following an unrecoverable firmware error, firmware stop, hardware failure, or environmentally induced (AC power) failure.

The auto-restart (reboot) option must be enabled from the ASMI or from the Control (Operator) Panel.

Concurrent access to the service processor menus of the ASMI

This access allows nondisruptive capabilities to change system default parameters, interrogate service processor progress and error logs, and set and reset service indicators (Light Path for low-end servers). All service processor functions can be accessed without needing to power down the system to the standby state. The administrator or IBM SSR dynamically can access the menus from any web browser-enabled console that is attached to the Ethernet service network, concurrently with normal system operation. Certain options, such as changing the hypervisor type, do not take effect until the next boot.

Management of the interfaces for connecting uninterruptible power source systems to the POWER processor-based systems and performing timed power-on (TPO) sequences

3.6.4 Diagnosing

General diagnostic objectives are to detect and identify problems so that they can be resolved quickly. The IBM diagnostic strategy includes the following elements:

- Provide a common error code format that is equivalent to a system reference code, system reference number, checkpoint, or firmware error code.
- Provide fault detection and problem isolation procedures. Support a remote connection ability that is used by the IBM Remote Support Center or IBM Designated Service.
- Provide interactive intelligence within the diagnostic tests with detailed online failure information while connected to the IBM back-end system.

Using the extensive network of advanced and complementary error detection logic that is built directly into hardware, firmware, and operating systems, the IBM Power Systems servers can perform considerable self-diagnoses.

Because of the FFDC technology that is designed into IBM servers, re-creating diagnostic tests for failures or requiring user intervention is unnecessary. Solid and intermittent errors are designed to be correctly detected and isolated at the time that the failure occurs. Runtime and boot time diagnostic tests are in this category.
Boot time
When an IBM Power Systems server starts, the service processor initializes the system hardware. Boot-time diagnostic testing uses a multi-tier approach for system validation, starting with managed low-level diagnostic tests that are supplemented with system firmware initialization and configuration of I/O hardware, followed by OS-initiated software test routines.

To minimize boot time, the system determines which of the diagnostic tests are required to be started to ensure correct operation, which is based on the way that the system was powered off, or on the boot-time selection menu.

Host Boot IPL
In POWER8, the initialization process during IPL changed. The Flexible Service Processor (FSP) is no longer the only instance that initializes and runs the boot process. With POWER8, the FSP initializes the boot processes, but on the POWER8 processor, one part of the firmware is running and performing the Central Electronics Complex chip initialization. A new chip stores the Host Boot firmware. The Self Boot Engine (SBE) is an internal part of the POWER8 chip. The SBE is used to boot the chip.

With this Host Boot initialization, new progress codes are available. An example of an FSP progress code is C1009003. During the Host Boot IPL, progress codes, such as CC009344, appear.

If there is a failure during the Host Boot process, a new Host Boot System Dump is collected and stored. This type of memory dump includes Host Boot memory and is offloaded to the HMC when it is available.

Run time
All Power Systems servers can monitor critical system components during run time, and they can take corrective actions when recoverable faults occur. The IBM hardware error-check architecture can report non-critical errors in the Central Electronics Complex in an out-of-band communications path to the service processor without affecting system performance.

A significant part of the IBM Runtime Diagnostics capabilities originates with the service processor. Extensive diagnostic and fault analysis routines were developed and improved over many generations of POWER processor-based servers, and they enable quick and accurate predefined responses to both actual and potential system problems.

The service processor correlates and processes runtime error information by using logic that is derived from IBM engineering expertise to count recoverable errors (called thresholding) and predict when corrective actions must be automatically initiated by the system. These actions can include the following items:

- Requests for a part to be replaced
- Dynamic invocation of built-in redundancy for automatic replacement of a failing part
- Dynamic deallocation of failing components so that system availability is maintained
Device drivers
In certain cases, diagnostic tests are best performed by operating system-specific drivers, most notably adapters or I/O devices that are owned directly by a logical partition (LPAR). In these cases, the operating system device driver often works with I/O device microcode to isolate and recover from problems. Potential problems are reported to an operating system device driver, which logs the error. In non-HMC managed servers, the OS can start the Call Home application to report the service event to IBM. For optional HMC managed servers, the event is reported to the HMC, which can initiate the Call Home request to IBM. I/O devices can also include specific exercisers that can be started by the diagnostic facilities to re-create problems (if required by service procedures).

3.6.5 Reporting
In the unlikely event that a system hardware or environmentally induced failure is diagnosed, IBM Power Systems servers report the error through various mechanisms. The analysis result is stored in system nonvolatile random access memory (NVRAM). Error log analysis (ELA) can be used to display the failure cause and the physical location of the failing hardware.

Using the Call Home infrastructure, the system automatically can send an alert through a phone line to a pager, or call for service if there is a critical system failure. A hardware fault also illuminates the amber system fault LED, which is on the system unit, to alert the user of an internal hardware problem.

On POWER8 processor-based servers, hardware and software failures are recorded in the system log. When a management console is attached, an ELA routine analyzes the error and forwards the event to the Service Focal Point™ (SFP) application running on the management console. The ELA routine can notify the system administrator that it isolated a likely cause of the system problem. The service processor event log also records unrecoverable checkstop conditions, forwards them to the SFP application, and notifies the system administrator. After the information is logged in the SFP application, if the system is correctly configured, a Call Home service request is initiated and the pertinent failure data with service parts information and part locations is sent to the IBM service organization. This information also contains the client contact information as defined in the IBM Electronic Service Agent (ESA) guided setup wizard. With the new HMC V8R8.1.0, a Serviceable Event Manager is available to block problems from being automatically transferred to IBM. For more information about this topic, see “Serviceable Event Manager” on page 102.

Error logging and analysis
When the root cause of an error is identified by a fault isolation component, an error log entry is created with basic data, for example:
- An error code that uniquely describes the error event
- The location of the failing component
- The part number of the component to be replaced, including pertinent data, such as engineering and manufacturing levels
- Return codes
- Resource identifiers
- FFDC data
Data that contains information about the effect that the repair has on the system is also included. Error log routines in the operating system and FSP can then use this information and decide whether the fault is a Call Home candidate. If the fault requires support intervention, a call is placed with service and support, and a notification is sent to the contact that is defined in the ESA-guided setup wizard.

Remote support
The Remote Management and Control (RMC) subsystem is delivered as part of the base operating system, including the operating system that runs on the HMC. RMC provides a secure transport mechanism across the local area network (LAN) interface between the operating system and the optional HMC. RMC is used by the operating system diagnostic application for transmitting error information. It performs several other functions, but they are not used for the service infrastructure.

Service Focal Point application for partitioned systems
A critical requirement in a logically partitioned environment is to ensure that errors are not lost before they are reported for service, and that an error is reported only once, regardless of how many LPARs experience the potential effect of the error. The SFP application on the management console or in the Integrated Virtualization Manager (IVM) is responsible for aggregating duplicate error reports, and it ensures that all errors are recorded for review and management. The SFP application provides other service-related functions, such as controlling service indicators, setting up Call Home, and providing guided maintenance.

When a local or globally reported service request is made to the operating system, the operating system diagnostic subsystem uses the RMC subsystem to relay error information to the optional HMC. For global events (platform unrecoverable errors, for example), the service processor also forwards error notification of these events to the HMC, providing a redundant error-reporting path if errors exist in the RMC subsystem network.

The first occurrence of each failure type is recorded in the Manage Serviceable Events task on the management console. This task then filters and maintains a history of duplicate reports from other LPARs or from the service processor. It then looks at all active service event requests within a predefined timespan, analyzes the failure to ascertain the root cause and, if enabled, initiates a Call Home for service. This methodology ensures that all platform errors are reported through at least one functional path, ultimately resulting in a single notification for a single problem. Similar service functionality is provided through the SFP application on the IVM for providing service functions and interfaces on non-HMC partitioned servers.

Extended error data
Extended error data (EED) is additional data that is collected either automatically at the time of a failure or manually later. The data that is collected depends on the invocation method, but it includes the following information:

- Firmware levels
- Operating system levels
- Additional fault isolation register values
- Recoverable error threshold register values
- System status
- Any other pertinent data

The data is formatted and prepared for transmission back to IBM either to assist the service support organization with preparing a service action plan for the IBM SSR or for additional analysis.
System dump handling
In certain circumstances, an error might require that a memory dump is automatically or manually created. In this event, the memory dump can be offloaded to the optional HMC. Specific management console information is included as part of the information that optionally can be sent to IBM Support for analysis. If additional information that relates to the memory dump is required, or if viewing the memory dump remotely becomes necessary, the management console memory dump record reports where the memory dump is located (on which management console) to the IBM Support center. If no management console is present, the memory dump might be either on the FSP or in the operating system, depending on the type of memory dump that was initiated and whether the operating system is operational.

3.6.6 Notifying
After a Power Systems server detects, diagnoses, and reports an error to an aggregation point, it then takes steps to notify the client and, if necessary, the IBM Support organization. Depending on the assessed severity of the error and support agreement, this client notification might range from a simple notification to automatically dispatching field service personnel to the client site with the correct replacement part.

Client Notify
When an event is important enough to report, but it does not indicate the need for a repair action or the need to call home to IBM Support, it is classified as Client Notify. Clients are notified because these events might be of interest to an administrator. The event might be a symptom of an expected systemic change, such as a network reconfiguration or failover testing of redundant power or cooling systems. These events include the following examples:

- Network events, such as the loss of contact over a LAN
- Environmental events, such as ambient temperature warnings
- Events that need further examination by the client (although these events do not necessarily require a part replacement or repair action)

Client Notify events are serviceable events because they indicate that something happened that requires client awareness if the client wants to take further action. These events can be reported to IBM at the discretion of the client.

Call Home
Call Home refers to an automatic or manual call from a client location to an IBM Support structure with error log data, server status, or other service-related information. The Call Home feature starts the service organization application so that the appropriate service action can begin. Call Home can be done through HMC or most non-HMC managed systems. Although configuring a Call Home function is optional, clients are encouraged to implement this feature to obtain service enhancements, such as reduced problem determination and faster and potentially more accurate transmission of error information. In general, using the Call Home feature can result in increased system availability. The ESA application can be configured for automated Call Home. For more information, see 3.7.4, “Electronic Services and Electronic Service Agent” on page 101.
Vital product data and inventory management
Power Systems store vital product data (VPD) internally, which keeps a record of how much memory is installed, how many processors are installed, the manufacturing level of the parts, and so on. These records provide valuable information that can be used by remote support and IBM SSRs, enabling the IBM SSRs to help keep the firmware and software current on the server.

IBM Service and Support Problem Management database
At the IBM Support center, historical problem data is entered into the IBM Service and Support Problem Management database. All of the information that is related to the error and any service actions that are taken by the IBM SSR, is recorded for problem management by the support and development organizations. The problem is then tracked and monitored until the system fault is repaired.

3.6.7 Locating and servicing
The final component of a comprehensive design for serviceability is the ability to effectively locate and replace parts that require service. POWER processor-based systems use a combination of visual cues and guided maintenance procedures to ensure that the identified part is replaced correctly, every time.

Packaging for service
The following service enhancements are included in the physical packaging of the systems to facilitate service:

- Color coding (touch points)
  Terracotta-colored touch points indicate that a component (FRU or CRU) can be concurrently maintained.
  Blue-colored touch points delineate components that might not be concurrently maintained; they might require that the system is turned off for removal or repair.

- Tool-less design
  Selected IBM systems support tool-less or simple tool designs. These designs require no tools, or require basic tools, such as flathead screw drivers, to service the hardware components.

- Positive retention
  Positive retention mechanisms help ensure proper connections between hardware components, such as from cables to connectors, and between two cards that attach to each other. Without positive retention, hardware components risk becoming loose during shipping or installation, which prevents a good electrical connection. Positive retention mechanisms, such as latches, levers, thumbscrews, pop Nylatches (U-clips), and cables are included to help prevent loose connections and to aid in installing (seating) parts correctly. These positive retention items do not require tools.

Light Path
The Light Path LED function is for scale-out systems that can be repaired by clients. In the Light Path LED implementation, when a fault condition is detected on the POWER8 processor-based system, an amber FRU fault LED is illuminated (turned on solid), which is then rolled up to the system fault LED. The Light Path system pinpoints the exact part by lighting the amber FRU fault LED that is associated with the part that must be replaced.
The service person can clearly identify components for replacement by using specific component level identify LEDs. The service person can also guide the IBM SSR directly to the component by signaling (flashing) the FRU component identify LED, and rolling up to the blue enclosure Locate LED.

After the repair, the LEDs shut off automatically when the problem is fixed. The Light Path LEDs are only visible while the system is in standby power. There is no gold cap or battery implemented.

**Service labels**

Service providers use these labels to assist with maintenance actions. Service labels are in various formats and positions. They are intended to transmit readily available information to the IBM SSR during the repair process.

Several of these service labels and their purposes are described:

- *Location diagrams* are strategically positioned on the system hardware and relate information about the placement of hardware components. Location diagrams can include location codes, drawings of physical locations, concurrent maintenance status, or other data that is pertinent to a repair. Location diagrams are especially useful when multiple components are installed, such as DIMMs, sockets, processor cards, fans, adapter, LEDs, and power supplies.

- *Remove or replace procedure labels* contain procedures that are often on a cover of the system or in other locations that are accessible to the IBM SSR. These labels provide systematic procedures, including diagrams that describe how to remove and replace certain serviceable hardware components.

- *Numbered arrows* indicate the order of operation and serviceability direction of components. Various serviceable parts, such as latches, levers, and touch points, must be pulled or pushed in a certain direction and order so that the mechanical mechanisms can engage or disengage. Numbered arrows generally improve the ease of serviceability.

**The operator panel**

The operator panel on a POWER processor-based system is an LCD display (two rows by 16 elements) that is used to present boot progress codes, indicating advancement through the system power-on and initialization processes. The operator panel is also used to display error and location codes when an error occurs that prevents the system from booting. It includes several buttons, enabling an IBM SSR or client to change various boot-time options and for other limited service functions.

**Concurrent maintenance**

The IBM POWER8 processor-based systems are designed with the understanding that certain components have higher intrinsic failure rates than others. These components can include fans, power supplies, and physical storage devices. Other devices, such as I/O adapters, can wear from repeated plugging and unplugging. For these reasons, these devices are concurrently maintainable when correctly configured. Concurrent maintenance is facilitated because of the redundant design for the power supplies, fans, and physical storage.

In addition to the previously mentioned components, the operator panel can be replaced concurrently by using service functions of the ASMI menu.
Repair and verify services

Repair and verify (R&V) services are automated service procedures that are used to guide a service provider, step-by-step, through the process of repairing a system and verifying that the problem was repaired. The steps are customized in the correct sequence for the particular repair for the specific system that is serviced. The following scenarios are covered by R&V services:

- Replacing a defective FRU or a CRU
- Reattaching a loose or disconnected component
- Correcting a configuration error
- Removing or replacing an incompatible FRU
- Updating firmware, device drivers, operating systems, middleware components, and IBM applications after replacing a part

R&V procedures can be used by IBM SSR providers who are familiar with the task and service providers who are not. Education-on-demand content is placed in the procedure at the appropriate locations. Throughout the R&V procedure, repair history is collected and provided to the Service and Support Problem Management Database for storage with the serviceable event to ensure that the guided maintenance procedures are operating correctly.

Clients can subscribe through the subscription services on the IBM Support Portal to obtain notifications about the latest updates that are available for service-related documentation.

IBM Knowledge Center

IBM Knowledge Center provides you with a single information center where you can access product documentation for IBM systems hardware, operating systems, and server software.

The latest version of the documentation is accessible through the Internet; however, a CD-ROM based version is also available.

The purpose of Knowledge Center, in addition to providing client related product information, is to provide softcopy information to diagnose and fix any problems that might occur with the system. Because the information is electronically maintained, changes due to updates or addition of new capabilities can be used by service representatives immediately.

The Knowledge Center contains sections specific to each server model, and include detailed service procedures for a number of potential repair situations. The service procedure repository for a particular server model can be found in the “Troubleshooting, service and support” section.

The IBM Knowledge Center can be found online at:
http://www.ibm.com/support/knowledgecenter/

QR code labels for servicing information

A label containing a QR code can be found on the top service cover of the Power S824L server. This can be scanned with an appropriate app on a mobile device to link to a number of sources of information that simplify the servicing of the system.

From this quick access link you can find information on topics including:

- Installing and configuring the system
- Troubleshooting and problem analysis
- Reference code lookup tables
- Part location guides
- Removing and replacing field replaceable units
- Video guides for removal and installation of customer replaceable units
3.7 Manageability

Several functions and tools help manageability so that you can efficiently and effectively manage your system.

3.7.1 Service user interfaces

The service interface allows support personnel or the client to communicate with the service support applications in a server by using a console, interface, or terminal. Delivering a clear, concise view of available service applications, the service interface allows the support team to manage system resources and service information in an efficient and effective way. Applications that are available through the service interface are carefully configured and placed to give service providers access to important service functions.

Various service interfaces are used, depending on the state of the system and its operating environment. The primary service interfaces are listed:

- Light Path (See “Light Path” on page 91 and “Service labels” on page 92.)
- Service processor and ASMI
- Operator panel
- Operating system service menu
- SFP on the HMC
- SFP Lite on IVM

Service processor

The service processor is a controller that is running its own operating system. It is a component of the service interface card.

The service processor operating system has specific programs and device drivers for the service processor hardware. The host interface is a processor support interface that is connected to the POWER processor. The service processor is always working, regardless of the main system unit’s state. The system unit can be in the following states:

- Standby (power off)
- Operating, ready to start partitions
- Operating with running LPARs

The service processor is used to monitor and manage the system hardware resources and devices. The service processor checks the system for errors, ensuring the connection to the management console for manageability purposes and accepting ASMI Secure Sockets Layer (SSL) network connections. The service processor can view and manage the machine-wide settings by using the ASMI, and it enables complete system and partition management from the HMC.

Analyzing a system that does not boot: The FSP can analyze a system that does not boot. Reference codes and detailed data are available in the ASMI and transferred to the HMC.
The service processor uses two Ethernet ports that run at 100 Mbps speed. Consider the following information:

- Both Ethernet ports are visible only to the service processor and can be used to attach the server to an HMC or to access the ASMI. The ASMI options can be accessed through an HTTP server that is integrated into the service processor operating environment.
- Both Ethernet ports support only auto-negotiation. Client-selectable media speed and duplex settings are not available.
- Both Ethernet ports have a default IP address, for example:
  - Service processor eth0 (HMC1 port) is configured as 169.254.2.147.
  - Service processor eth1 (HMC2 port) is configured as 169.254.3.147.

The following functions are available through the service processor:

- Call Home
- ASMI
- Error information (error code, part number, and location codes) menu
- View of guarded components
- Limited repair procedures
- Generate dump
- LED Management menu
- Remote view of ASMI menus
- Firmware update through a USB key

**Advanced System Management Interface**

ASMI is the interface to the service processor to manage the operation of the server, such as auto-power restart, and to view information about the server, such as the error log and VPD. Various repair procedures require connection to the ASMI.

The ASMI is accessible through the management console. It is also accessible by using a web browser on a system that is connected directly to the service processor (in this case, either a standard Ethernet cable or a crossed cable) or through an Ethernet network. ASMI can also be accessed from an ASCII terminal, but this function is available only while the system is in the platform powered-off mode.

Use the ASMI to change the service processor IP addresses or to apply certain security policies and prevent access from unwanted IP addresses or ranges.

You might be able to use the service processor’s default settings. In that case, accessing the ASMI is not necessary. To access the ASMI, use one of the following methods:

- Use a management console.
  
  If configured to do so, the management console connects directly to the ASMI for a selected system from this task.
  
  To connect to the ASMI from a management console, complete the following steps:
  
  a. Open **Systems Management** from the navigation pane.
  b. From the work window, select one of the managed systems.
  c. From the System Management tasks list, click **Operations → Launch Advanced System Management (ASM)**.

- Use a web browser.
  
  At the time of writing, supported web browsers are Microsoft Internet Explorer (Version 10.0.9200.16439), Mozilla Firefox ESR (Version 24), and Chrome (Version 30). Later versions of these browsers might work, but they are not officially supported. The...
JavaScript language and cookies must be enabled and Transport Layer Security (TLS) 1.2 might need to be enabled.

The web interface is available during all phases of system operation, including the initial program load (IPL) and run time. However, several of the menu options in the web interface are unavailable during IPL or run time to prevent usage or ownership conflicts if the system resources are in use during that phase. The ASMI provides an SSL web connection to the service processor. To establish an SSL connection, open your browser by using the following address:

https://<ip_address_of_service_processor>

Note: To make the connection through Internet Explorer, click Tools Internet Options. Clear the Use TLS 1.0 check box, and click OK.

Use an ASCII terminal.

The ASMI on an ASCII terminal supports a subset of the functions that are provided by the web interface and it is available only when the system is in the platform powered-off mode. The ASMI on an ASCII console is not available during several phases of system operation, such as the IPL and run time.

Use a command-line start of the ASMI.

Either on the HMC itself or when correctly configured on a remote system, it is possible to start the ASMI web interface from the HMC command line. Open a terminal window on the HMC or access the HMC with a terminal emulation and run the following command:

`asmmenu --ip <ip_address>`

On the HMC, a browser window opens automatically with the ASMI window and, when configured correctly, a browser window opens on a remote system when the command is issued from there.

The operator panel

The service processor provides an interface to the operator panel, which is used to display system status and diagnostic information.

The operator panel can be accessed in two ways:

- By using the normal operational front view
- By pulling it out to access the switches and viewing the LCD display

Several operator panel features are listed:

- A 2 x 16 character LCD display
- Reset, enter, power On/Off, increment, and decrement buttons
- Amber System Information/Attention, and a green Power LED
- Blue Enclosure Identify LED
- Altitude sensor
- USB Port
- Speaker/Beeper

The following functions are available through the operator panel:

- Error information
- Generate dump
- View machine type, model, and serial number
- Limited set of repair functions
Service menu
The System Management Services (SMS) error log is accessible on the SMS menus. This error log contains errors that are identified by partition firmware when the system or partition is starting.

The service processor’s error log can be accessed on the ASMI menus.

You can also access the system diagnostics from a Network Installation Management (NIM) server.

Alternative method: When you order a Power System, a DVD-ROM or DVD-RAM might be an option. An alternative method for maintaining and servicing the system must be available if you do not order the DVD-ROM or DVD-RAM.

Service Focal Point on the Hardware Management Console
Service strategies become more complicated in a partitioned environment. The Manage Serviceable Events task in the management console can help streamline this process.

Each LPAR reports errors that it detects and forwards the event to the SFP application that is running on the management console, without determining whether other LPARs also detect and report the errors. For example, if one LPAR reports an error for a shared resource, such as a managed system power supply, other active LPARs might report the same error.

By using the Manage Serviceable Events task in the management console, you can avoid long lists of repetitive Call Home information by recognizing that these errors are repeated errors and consolidating them into one error.

In addition, you can use the Manage Serviceable Events task to initiate service functions on systems and LPARs, including exchanging parts, configuring connectivity, and managing memory dumps.

3.7.2 IBM Power Systems Firmware maintenance

The IBM Power Systems Client-Managed Microcode is a methodology to manage and install microcode updates on Power Systems and their associated I/O adapters.

Firmware entitlement
With the new HMC Version V8R8.1.0.0 and Power Systems servers, the firmware installations are restricted to entitled servers. The client must be registered with IBM and entitled with a service contract. During the initial machine warranty period, the access key is installed in the machine by manufacturing. The key is valid for the regular warranty period, plus additional time. The Power Systems Firmware is relocated from the public repository to the access control repository. The I/O firmware remains on the public repository, but the server must be entitled for installation. When the lslic command is run to display the firmware levels, a new value, `update_access_key_exp_date`, is added. The HMC GUI and the ASMI menu show the Update access key expiration date.

When the system is no longer entitled, the firmware updates fail. New System Reference Code (SRC) packages are available:

- E302FA06: Acquisition entitlement check failed
- E302FA08: Installation entitlement check failed

Any firmware release that was made available during the entitled time frame can still be installed. For example, if the entitlement period ends 31 December 2014 and a new firmware
release is released before the end of that entitlement period, it can still be installed. If that firmware is downloaded after 31 December 2014, but it was made available before the end of the entitlement period, it still can be installed. Any newer release requires a new update access key.

**Note:** The update access key expiration date requires a valid entitlement of the system to perform firmware updates.

You can find an update access key at the IBM CoD Home website:

http://www-912.ibm.com/pod/pod

To access the IBM entitled Software Support page for details, go to the following website:

http://www.ibm.com/servers/eserver/ess

**Firmware updates**

System firmware is delivered as a release level or a service pack. Release levels support the general availability (GA) of new functions or features and new machine types or models. Upgrading to a higher release level is disruptive to client operations. These release levels are not supported by service packs. Service packs are intended to contain only firmware fixes and do not introduce new functions. A *service pack* is an update to an existing release level.

If the system is managed by a management console, you use the management console for firmware updates. By using the management console, you can take advantage of the concurrent firmware maintenance (CFM) option when concurrent service packs are available. CFM is the IBM Power Systems Firmware update that can be partially or wholly concurrent or nondisruptive.

With the introduction of CFM, clients that want maximum stability can defer until a compelling reason to upgrade arises, for example:

- A release level is approaching its end-of-service date (available for approximately a year and soon service will not be supported).
- The client wants to move a system to a more standardized release level when there are multiple systems in an environment with similar hardware.
- A new release has a new function that is needed in the environment.
- A scheduled maintenance action causes a platform restart, which provides an opportunity to also upgrade to a new firmware release.

Updating and upgrading system firmware depends on several factors, such as whether the system is stand-alone or managed by a management console, the current firmware that is installed, and the operating systems that are running on the system. These scenarios and the associated installation instructions are comprehensively outlined in the firmware section of Fix Central at the following website:

http://www.ibm.com/support/fixcentral/

You might also want to review the preferred practice white papers at the following website:


**Firmware update steps**

The system firmware consists of service processor microcode, Open Firmware microcode, and Systems Power Control Network (SPCN) microcode.
The firmware and microcode can be downloaded and installed either from an HMC, from a running partition, or from USB port number 1 on the rear of the system, if that system is not managed by an HMC.

Power Systems has a permanent firmware boot side (A side) and a temporary firmware boot side (B side). New levels of firmware must be installed first on the temporary side to test the update’s compatibility with existing applications. When the new level of firmware is approved, it can be copied to the permanent side.

For access to the initial websites that address this capability, see the Support for IBM Systems website:

http://www.ibm.com/systems/support

Follow these steps:

1. For Power Systems, select the **Power** link.
2. Although the content under the Popular links section can change, click the **Firmware and HMC updates** link to go to the resources to keep your system’s firmware current.
3. If there is an HMC to manage the server, the HMC interface can be used to view the levels of server firmware and subsystem firmware that are installed and that are available to download and install.

Each IBM Power Systems server has the following levels of server firmware and subsystem firmware:

- **Installed level**
  This level of server firmware or subsystem firmware is installed and will be installed into memory after the managed system is powered off and then powered on. It is installed on the temporary side of the system firmware.

- **Activated level**
  This level of server firmware or subsystem firmware is active and running in memory.
Accepted level

This level is the backup level of server or subsystem firmware. You can return to this level of server or subsystem firmware if you decide to remove the installed level. It is installed on the permanent side of system firmware.

Figure 3-3 shows the levels in the HMC.

![Figure 3-3 HMC system firmware Updates window](image)

IBM provides the CFM function on selected Power Systems. This function supports applying nondisruptive system firmware service packs to the system concurrently (without requiring a reboot operation to activate the changes). For systems that are not managed by an HMC, the installation of system firmware is always disruptive.

The concurrent levels of system firmware can contain fixes that are known as deferred. These deferred fixes can be installed concurrently but they are not activated until the next IPL. Deferred fixes, if any, are identified in the Firmware Update Descriptions table of the firmware document. For deferred fixes within a service pack, only the fixes in the service pack that cannot be concurrently activated are deferred. Table 3-1 shows the file-naming convention for system firmware.

<table>
<thead>
<tr>
<th>PPNNSSS_FFF_DDD</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP</td>
<td>Package identifier</td>
</tr>
<tr>
<td>NN</td>
<td>Platform and class</td>
</tr>
<tr>
<td>SSS</td>
<td>Release indicator</td>
</tr>
<tr>
<td>FFF</td>
<td>Current fix pack</td>
</tr>
<tr>
<td>DDD</td>
<td>Last disruptive fix pack</td>
</tr>
</tbody>
</table>

The following example uses the convention:

01SV810_030_030 = POWER8 Entry Systems Firmware for 8286-41A and 8286-42A
An installation is disruptive if the following statements are true:

- The release levels (SSS) of the currently installed firmware and the new firmware differ.
- The service pack level (FFF) and the last disruptive service pack level (DDD) are equal in the new firmware.

Otherwise, an installation is concurrent if the service pack level (FFF) of the new firmware is higher than the service pack level that is installed on the system and the conditions for disruptive installation are not met.

### 3.7.3 Concurrent firmware update improvements

Since POWER6, firmware service packs are concurrently applied and take effect immediately. Occasionally, a service pack is shipped where most of the features can be concurrently applied, but because changes to certain server functions (for example, changing initialization values for chip controls) cannot occur during operation, a patch in this area required a system reboot for activation.

With the Power On Reset Engine (PORE), the firmware can now dynamically power off processor components, change the registers, and reinitialize while the system is running, without a discernible effect on any applications that are running on a processor. This capability potentially allows concurrent firmware changes in POWER8, which in earlier designs required a reboot to take effect.

Activating certain new firmware functions requires the installation of a firmware release level. This process is disruptive to server operations and requires a scheduled outage and full server reboot.

### 3.7.4 Electronic Services and Electronic Service Agent

IBM transformed its delivery of hardware and software support services to help you achieve higher system availability. Electronic Services is a web-enabled solution that offers an exclusive, no additional charge enhancement to the service and support that are available for IBM servers. These services provide the opportunity for greater system availability with faster problem resolution and preemptive monitoring. The Electronic Services solution consists of two separate, but complementary, elements:

- Electronic Services news page
- Electronic Service Agent

**Electronic Services news page**

The Electronic Services news page is a single Internet entry point that replaces the multiple entry points that traditionally are used to access IBM Internet services and support. With the news page, you can gain easier access to IBM resources for assistance in resolving technical problems.

**Electronic Service Agent**

ESA is software on your server. It monitors events and transmits system inventory information to IBM on a periodic, client-defined timetable. ESA automatically reports hardware problems to IBM.

Early knowledge about potential problems enables IBM to deliver proactive service that can result in higher system availability and performance. In addition, information that is collected through ESA is available to IBM SSRs when they help answer your questions or diagnose
problems. Installation and use of ESA for problem reporting enables IBM to provide better support and service for your IBM server.

To learn how Electronic Services can work for you, see the following website (an IBM ID is required):

http://www.ibm.com/support/electronic

Electronic Services offers these benefits:

- Increased uptime
  The ESA tool enhances the warranty or maintenance agreement by providing faster hardware error reporting and system information uploading to IBM Support. This function can translate to less time that is wasted monitoring the symptoms, diagnosing the error, and manually calling IBM Support to open a problem record.
  Its 24x7 monitoring and reporting mean no more dependency on human intervention or off-hours client personnel when errors are encountered in the middle of the night.

- Security
  The ESA tool is secure in monitoring, reporting, and storing the data at IBM. The ESA tool securely transmits either through the Internet (HTTPS or VPN) or modem. The ESA tool can be configured to communicate securely through gateways to provide clients a single point of exit from their site.
  Communication is one way. Activating ESA does not enable IBM to call into a client's system. System inventory information is stored in a secure database, which is protected behind IBM firewalls. System inventory information is viewable only by the client and IBM. The client's business applications or business data is never transmitted to IBM.

- More accurate reporting
  Because system information and error logs are automatically uploaded to the IBM Support Center with the service request, clients are not required to find and send system information, which decreases the risk of misreported or misdiagnosed errors.
  When the problem error data is inside IBM, it is run through a data knowledge management system and knowledge articles are appended to the problem record.

- Customized support
  By using the IBM ID that you enter during activation, you can view system and support information by selecting My Systems at the Electronic Support website:

  http://www.ibm.com/support/electronic

  My Systems provides valuable reports about installed hardware and software, using information that is collected from the systems by ESA. Reports are available for any system that is associated with the client's IBM ID. Premium Search combines the function of search and the value of ESA information, providing an advanced search of the technical support knowledge base. Using Premium Search and the ESA information that was collected from your system, clients can see search results that apply specifically to their systems.
  For more information about how to use the power of IBM Electronic Services, contact your IBM SSR, or see the following website:

  http://www.ibm.com/support/electronic

**Serviceable Event Manager**

Serviceable Event Manager (SEM) allows the user to decide which of the serviceable events are called home with ESA. It is possible to lock certain events. Certain clients might not allow
data to be transferred outside their company. After SEM is enabled, the analysis of the possible problems might take longer.

Enable SEM by running the `chhmc -c sem -s enable` command. You can disable SEM mode and specify the state in which to leave the Call Home feature by running the following commands:

- `chhmc -c sem -s disable --callhome disable`
- `chhmc -c sem -s disable --callhome enable`

Configure SEM from the HMC GUI. After you select the Serviceable Event Manager, as shown in Figure 3-4, you must add the HMC console.
In the next window, you can configure the HMC that is used to manage the Serviceable Events and proceed with further configuration steps, as shown in Figure 3-5.

![Figure 3-5  Initial SEM window](image)

The configurable options in Figure 3-5 are described:

- **Registered Management Consoles**
  The Total consoles value lists the number of consoles that are registered. Select **Manage Consoles** to manage the list of Registered Management Consoles.

- **Event Criteria**
  Select the filters for the list of serviceable events. After the selections are made, click **Refresh** to refresh the list based on the filter values.

  - **Approval state**
    Select the value for approval state to filter the list.

  - **Status**
    Select the value for the status to filter the list.

  - **Originating HMC**
    Select a single registered console or All consoles to filter the list.

- **Serviceable Events**
  The Serviceable Events table shows the list of events based on the selected filters. To refresh the list, click **Refresh**.

The following menu options are available when you select an event in the table:

- **View Details**
  Shows the details of this event.

- **View Files**
  Shows the files that are associated with this event.
Approve Call Home
Approves the Call Home of this event. This option is available only if the event is not approved already.
The Help/Learn more function can be used to get more information about the other available windows for the Serviceable Event Manager.

3.8 Selected POWER8 RAS capabilities by operating system

Table 3-2 provides a list of the Power Systems RAS capabilities by operating system. The HMC is an optional feature on scale-out Power Systems servers.

<table>
<thead>
<tr>
<th>RAS feature</th>
<th>Linux</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RHEL6.5</td>
</tr>
<tr>
<td></td>
<td>RHEL7</td>
</tr>
<tr>
<td></td>
<td>SLES11SP3</td>
</tr>
<tr>
<td></td>
<td>Ubuntu 14.04</td>
</tr>
<tr>
<td></td>
<td>Ubuntu 15.04</td>
</tr>
<tr>
<td>Processor</td>
<td></td>
</tr>
<tr>
<td>FFDC for fault detection/error isolation</td>
<td>X</td>
</tr>
<tr>
<td>Dynamic Processor Deallocation</td>
<td>X&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Core Error Recovery</td>
<td></td>
</tr>
<tr>
<td>Alternative processor recovery</td>
<td>X&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Partition Core Contained Checkstop</td>
<td>X&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>I/O subsystem</td>
<td></td>
</tr>
<tr>
<td>PCI Express bus enhanced error detection</td>
<td>X</td>
</tr>
<tr>
<td>PCI Express bus enhanced error recovery</td>
<td>X&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>PCI Express card hot-swap</td>
<td>X&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Memory availability</td>
<td></td>
</tr>
<tr>
<td>Memory Page Deallocation</td>
<td>X</td>
</tr>
<tr>
<td>Special Uncorrectable Error Handling</td>
<td>X</td>
</tr>
<tr>
<td>Fault detection and isolation</td>
<td></td>
</tr>
<tr>
<td>Storage Protection Keys</td>
<td>Not used by OS</td>
</tr>
<tr>
<td>Error log analysis</td>
<td>X&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Serviceability</td>
<td></td>
</tr>
<tr>
<td>Boot-time progress indicators</td>
<td>X</td>
</tr>
<tr>
<td>Firmware error codes</td>
<td>X</td>
</tr>
<tr>
<td>Operating system error codes</td>
<td>X&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Inventory collection</td>
<td>X</td>
</tr>
<tr>
<td>RAS feature</td>
<td>Linux</td>
</tr>
<tr>
<td>---------------------------------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td></td>
<td>RHEL6.5</td>
</tr>
<tr>
<td></td>
<td>RHEL7</td>
</tr>
<tr>
<td></td>
<td>SLES11SP3</td>
</tr>
<tr>
<td></td>
<td>Ubuntu 14.04</td>
</tr>
<tr>
<td></td>
<td>Ubuntu 15.04</td>
</tr>
<tr>
<td>Environmental and power warnings</td>
<td>X</td>
</tr>
<tr>
<td>Hot-swap DASD/media</td>
<td>X</td>
</tr>
<tr>
<td>Dual disk controllers/Split backplane</td>
<td>X</td>
</tr>
<tr>
<td>EED collection</td>
<td>X</td>
</tr>
<tr>
<td>SP “Call Home” on non-HMC configurations</td>
<td>$^a$X</td>
</tr>
<tr>
<td>IO adapter/device stand-alone diagnostic tests with PowerVM</td>
<td>X</td>
</tr>
<tr>
<td>SP mutual surveillance with POWER Hypervisor</td>
<td>X</td>
</tr>
<tr>
<td>Dynamic firmware update with HMC</td>
<td>X</td>
</tr>
<tr>
<td>Service Agent Call Home Application</td>
<td>$^a$X</td>
</tr>
<tr>
<td>Service Indicator LED support</td>
<td>X</td>
</tr>
<tr>
<td>System dump for memory, POWER Hypervisor, and SP</td>
<td>X</td>
</tr>
<tr>
<td>Information center/IBM Systems Support Site service publications</td>
<td>X</td>
</tr>
<tr>
<td>System Support Site education</td>
<td>X</td>
</tr>
<tr>
<td>Operating system error reporting to HMC SFP application</td>
<td>X</td>
</tr>
<tr>
<td>RMC secure error transmission subsystem</td>
<td>X</td>
</tr>
<tr>
<td>Healthcheck scheduled operations with HMC</td>
<td>X</td>
</tr>
<tr>
<td>Operator panel (real or virtual)</td>
<td>X</td>
</tr>
<tr>
<td>Concurrent Op Panel Maintenance</td>
<td>X</td>
</tr>
<tr>
<td>Redundant HMCs</td>
<td>X</td>
</tr>
<tr>
<td>Automated server recovery/restart</td>
<td>X</td>
</tr>
<tr>
<td>High availability clustering support</td>
<td>X</td>
</tr>
<tr>
<td>Repair and Verify Guided Maintenance with HMC</td>
<td>X</td>
</tr>
<tr>
<td>PowerVM Live Partition/Live Application Mobility With PowerVM Enterprise Edition</td>
<td>X</td>
</tr>
<tr>
<td><strong>EPOW</strong></td>
<td>$^a$X</td>
</tr>
<tr>
<td>EPOW error handling</td>
<td>$^a$X</td>
</tr>
</tbody>
</table>

a. Supported in POWER Hypervisor, but not supported in a KVM environment
b. Supported in POWER Hypervisor, with limited support in a KVM environment
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.

- *IBM Power Systems S812L and S822L Technical Overview and Introduction*, REDP-5098
- *IBM Power System S822 Technical Overview and Introduction*, REDP-5102
- *IBM Power Systems S814 and S824 Technical Overview and Introduction*, REDP-5097
- *IBM Power Systems E850 Technical Overview and Introduction*, REDP-5222
- *IBM Power Systems E870 and E880 Technical Overview and Introduction*, REDP-5137
- *IBM PowerVM Best Practices*, SG24-8062
- *IBM PowerVM Enhancements What is New in 2013*, SG24-8198
- *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940
- *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590
- *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](http://ibm.com/redbooks)

Other publications

These websites are also relevant as further information sources:

- *Active Memory Expansion: Overview and Usage Guide*
  

- *IBM EnergyScale for POWER8 Processor-Based Systems* white paper
  

- *IBM Power Facts and Features - IBM Power Systems, IBM PureFlex System, and Power Blades*
  

- *IBM Power System S812L server specifications*
  
IBM Power System S814 server specifications
http://www.ibm.com/systems/power/hardware/s814/specs.html

IBM Power System S822 server specifications
http://www.ibm.com/systems/power/hardware/s822/specs.html

IBM Power System S822L server specifications
http://www.ibm.com/systems/power/hardware/s822l/specs.html

IBM Power System S824 server specifications
http://www.ibm.com/systems/power/hardware/s824/specs.html

IBM Power System S824L server specifications:
http://www.ibm.com/systems/power/hardware/s824l/specs.html

IBM Power System E850 server specifications:
http://www.ibm.com/systems/power/hardware/e850/specs.html

IBM Power System E870 server specifications:
http://www.ibm.com/systems/power/hardware/e870/specs.html

IBM Power System E870 server specifications:
http://www.ibm.com/systems/power/hardware/e870/specs.html

Specific storage devices that are supported for Virtual I/O Server

System RAS - Introduction to Power Systems Reliability, Availability, and Serviceability

Online resources

These websites are also relevant as further information sources:

IBM Fix Central website
http://www.ibm.com/support/fixcentral/

IBM Knowledge Center
http://www.ibm.com/support/knowledgecenter/

IBM Power Systems website
http://www.ibm.com/systems/power/

IBM Power Systems Hardware information center
http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/index.jsp

IBM Storage website
http://www.ibm.com/systems/storage/

IBM System Planning Tool website
http://www.ibm.com/systems/support/tools/systemplanningtool/

IBM Systems Energy Estimator
http://www-912.ibm.com/see/EnergyEstimator/
- Migration combinations of processor compatibility modes for active Partition Mobility
  http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/topic/p7hc3/iphc3pcmco
  mbosact.htm

- Power Systems Capacity on Demand website
  http://www.ibm.com/systems/power/hardware/cod/

- Support for IBM Systems website

- Current information about IBM Java and tested Linux distributions are available here:

- Additional information about the OpenJDK port for Linux on PPC64 LE, as well as some
  pre-generated builds can be found here:
  http://cr.openjdk.java.net/~simonis/ppc-aix-port/

- Launchpad.net has resources for Ubuntu builds. You can find out about them here:
  https://launchpad.net/ubuntu/+source/openjdk-9
  https://launchpad.net/ubuntu/+source/openjdk-8
  https://launchpad.net/ubuntu/+source/openjdk-7

**Help from IBM**

IBM Support and downloads
ibm.com/support

IBM Global Services
ibm.com/services
IBM Power System S824L
Technical Overview and Introduction

Linux server built on OpenPOWER technologies

Ready for scientific, Java, and Big Data & Analytics workloads

Dual NVIDIA GPU accelerators supported

This IBM Redpaper publication is a comprehensive guide that covers the IBM Power System S824L (8247-42L) servers that support Linux operating systems. The objective of this paper is to introduce the major innovative Power S824L offerings and their relevant functions:

- The new IBM POWER8™ processor, which is available at frequencies of 3.02 GHz and 3.42 GHz, 3.52 GHz, and 4.15 GHz
- A processor that is designed to accommodate high-wattage adapters, such as NVIDIA graphics processing units (GPUs), that provide acceleration for scientific, engineering, Java, big data analytics, and other technical computing workloads
- Based on OpenPOWER technologies
- I/O drawer expansion options offers greater flexibility
- Two integrated memory controllers with improved latency and bandwidth
- Improved reliability, serviceability, and availability (RAS) functions
- IBM EnergyScale technology that provides features, such as power trending, power-saving, power capping, and thermal measurement

This publication is for professionals who want to acquire a better understanding of IBM Power Systems products.

This paper expands the current set of IBM Power Systems documentation by providing a desktop reference that offers a detailed technical description of the Power S824L server.

This paper does not replace the latest marketing materials and configuration tools. It is intended as an additional source of information that, together with existing sources, can be used to enhance your...