IBM Power System E980
Technical Overview and Introduction

James Cruickshank
Yongsheng Li (Victor)
Armin Röll
Note: Before using this information and the product it supports, read the information in “Notices” on page ix.

First Edition (August 2018)
This edition applies to the IBM Power System E980 (9080-M95).

© Copyright International Business Machines Corporation 2018. All rights reserved.
Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.
This document was created or updated on August 15, 2018.

**Important:** At time of publication, this book is based on a pre-GA version of a product. For the most up-to-date information regarding this product, consult the product documentation or subsequent updates of this book.
Contents

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

Chapter 1. General description ................................................................................. 1
1.1 System overview .................................................................................................. 3
  1.1.1 System control unit ......................................................................................... 4
  1.1.2 System nodes .................................................................................................. 5
  1.1.3 Hardware components highlight ..................................................................... 8
1.2 Operating environment ......................................................................................... 9
1.3 Physical package .................................................................................................. 10
  1.3.1 Lift tools ......................................................................................................... 10
1.4 System features ................................................................................................... 10
  1.4.1 Minimum configuration .................................................................................. 11
  1.4.2 Power supply features ................................................................................... 12
  1.4.3 Processor module features ............................................................................ 12
  1.4.4 POWER9 processor highlights ....................................................................... 13
  1.4.5 Memory features ............................................................................................ 14
  1.4.6 System node PCIe slots .................................................................................. 16
  1.4.7 USB ................................................................................................................ 17
  1.4.8 Disk and media features .................................................................................. 18
1.5 I/O drawers ......................................................................................................... 23
  1.5.1 PCIe Gen3 I/O expansion drawer ..................................................................... 23
  1.5.2 I/O drawers and usable PCI slot ..................................................................... 24
  1.5.3 EXP24S SFF Gen2-bay Drawer ..................................................................... 25
  1.5.4 EXP24SX and EXP12SX SAS Storage Enclosures ......................................... 26
1.6 System racks ....................................................................................................... 28
  1.6.1 What’s new? ................................................................................................... 29
  1.6.2 IBM 7014 Model T42 rack ............................................................................. 29
  1.6.3 IBM Enterprise 42U Slim Rack 7965-S42 ....................................................... 31
  1.6.4 Feature Code #0551 1.8 Meter Rack .............................................................. 31
  1.6.5 Feature Code #0553 2.0 Meter Rack .............................................................. 32
  1.6.6 Feature Code #ER05 rack ............................................................................. 32
  1.6.7 The AC power distribution unit and rack content ........................................... 32
  1.6.8 PDU connection limits .................................................................................. 35
  1.6.9 Rack-mounting rules ...................................................................................... 36
  1.6.10 Useful rack additions .................................................................................... 36
  1.6.11 Original equipment manufacturer racks ....................................................... 39
1.7 Hardware Management Console ......................................................................... 40
  1.7.1 What’s new? ................................................................................................... 40
  1.7.2 Hardware Management Console overview ................................................... 40
  1.7.3 Hardware Management Console code level ................................................... 42
  1.7.4 Two architectures of Hardware Management Console .................................... 43
  1.7.5 Connectivity to POWER9 processor-based systems ...................................... 44
Chapter 2. Architecture and technical overview

2.1 The IBM POWER9 processor

2.1.1 POWER9 processor overview

2.1.2 POWER9 processor core

2.1.3 Simultaneous multithreading

2.1.4 POWER9 compatibility modes

2.1.5 Processor feature codes

2.1.6 Memory access

2.1.7 On-chip L3 cache innovation and intelligent caching

2.1.8 Hardware transactional memory

2.1.9 POWER9 accelerator processor interfaces

2.1.10 Power and performance management

2.1.11 Comparison of the POWER9, POWER8, and POWER7+ processors

2.2 Memory subsystem

2.2.1 Custom DIMM

2.2.2 Memory placement rules

2.2.3 Memory activation

2.2.4 Memory throughput

2.2.5 Active Memory Mirroring

2.2.6 Memory Error Correction and Recovery

2.2.7 Special Uncorrectable Error handling

2.3 Capacity on Demand

2.3.1 What’s new?

2.3.2 Capacity Upgrade on Demand (CUoD)

2.3.3 Static activations

2.3.4 Elastic Capacity on Demand (Elastic CoD)

2.3.5 Power enterprise pools and mobile capacity on demand (Mobile CoD)

2.3.6 Utility Capacity on Demand (Utility CoD)

2.3.7 Trial Capacity on Demand (Trial CoD)

2.3.8 Software licensing and CoD

2.4 System bus

2.4.1 PCI Express Gen4

2.4.2 Service Processor Bus

2.5 PCIe adapters

2.5.1 What’s new?

2.5.2 PCI Express (PCIe)

2.5.3 LAN adapters

2.5.4 Graphics adapters

2.5.5 SAS adapters

2.5.6 Fibre Channel adapters

2.5.7 USB adapters

2.5.8 InfiniBand host channel adapters

2.5.9 Cryptographic Coprocessor

2.5.10 CAPI adapters

2.6 Internal storage

2.7 External I/O subsystems

2.7.1 PCIe Gen3 I/O expansion drawer

2.7.2 PCIe Gen3 I/O expansion drawer optical cabling

2.7.3 PCIe Gen3 I/O expansion drawer SPCN cabling

2.8 External disk subsystems

2.8.1 EXP24S SFF Gen2-bay Drawer
4.8.1 Detecting errors .................................................. 147
4.8.2 Error checkers, fault isolation registers, and First-Failure Data Capture ..... 147
4.8.3 Service processor .................................................. 148
4.8.4 Diagnosing ......................................................... 149
4.8.5 Reporting .......................................................... 151
4.8.6 Notifying ........................................................... 153
4.8.7 Locating and servicing ........................................... 154
4.9 Manageability ......................................................... 157
  4.9.1 Service user interfaces ........................................... 157
  4.9.2 IBM Power Systems Firmware maintenance ..................... 161
  4.9.3 Concurrent firmware maintenance improvements ............... 165
  4.9.4 Electronic Services and Electronic Service Agent ............... 165
4.10 Selected POWER9 RAS capabilities by operating system .................. 168

Related publications .................................................. 171
IBM Redbooks .......................................................... 171
Online resources ....................................................... 171
Help from IBM ......................................................... 172
Notices

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

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

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

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

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

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

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

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

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

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

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

COPYRIGHT LICENSE:

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

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

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

<table>
<thead>
<tr>
<th>Active Memory™</th>
<th>Interconnect®</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIX®</td>
<td>Micro-Partitioning®</td>
</tr>
<tr>
<td>AIX 5L™</td>
<td>OpenCAPI™</td>
</tr>
<tr>
<td>C3®</td>
<td>POWER®</td>
</tr>
<tr>
<td>Db2®</td>
<td>POWER Hypervisor™</td>
</tr>
<tr>
<td>DS8000®</td>
<td>Power Systems™</td>
</tr>
<tr>
<td>Easy Tier®</td>
<td>Power Systems Software™</td>
</tr>
<tr>
<td>Electronic Service Agent™</td>
<td>POWER6®</td>
</tr>
<tr>
<td>EnergyScale™</td>
<td>POWER6+™</td>
</tr>
<tr>
<td>eServer™</td>
<td>POWER7®</td>
</tr>
<tr>
<td>IBM®</td>
<td>POWER7+™</td>
</tr>
<tr>
<td>IBM FlashSystem®</td>
<td>POWER8®</td>
</tr>
<tr>
<td>IBM Spectrum™</td>
<td>POWER9™</td>
</tr>
<tr>
<td></td>
<td>PowerHA®</td>
</tr>
<tr>
<td></td>
<td>PowerVM®</td>
</tr>
<tr>
<td></td>
<td>PowerVP™</td>
</tr>
<tr>
<td></td>
<td>Redbooks®</td>
</tr>
<tr>
<td></td>
<td>Redpaper™</td>
</tr>
<tr>
<td></td>
<td>Redbooks (logo)</td>
</tr>
<tr>
<td></td>
<td>RS/6000®</td>
</tr>
<tr>
<td></td>
<td>Storwize®</td>
</tr>
<tr>
<td></td>
<td>System Storage®</td>
</tr>
<tr>
<td></td>
<td>SystemMirror®</td>
</tr>
<tr>
<td></td>
<td>XIV®</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.

LTO, Ultrium, the LTO Logo and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

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

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

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

This IBM® Redpaper™ publication gives a broad understanding of a new architecture of the IBM Power System E980 server that support IBM AIX®, IBM i, and Linux operating systems. The objective of this paper is to introduce the major innovative Power E980 offerings and relevant functions:

- The IBM POWER9™ processor, which is available at frequencies of 3.55 to 4.0 GHz.
- Significantly strengthened cores and larger caches.
- Supports up to 64 TB memory.
- Integrated I/O subsystem and hot-pluggable Peripheral Component Interconnect® Express (PCIe) Gen4 slots, double the bandwidth of Gen3 I/O slots.
- Supports EXP12SX, ESP24SX external disk drawers, which has 12Gb SAS interfaces and double the existing EXP24S drawer bandwidth.
- New IBM EnergyScale™ technology offers new variable processor frequency modes that provide a significant performance boost beyond the static nominal frequency.

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 (ISVs)

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 E980 server.

This paper does not replace the current marketing materials and configuration tools. It is intended as an extra 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.

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 17 years experience working with RS/6000®, pSeries, System p and Power Systems products. James supports customers in the financial services sector in the UK.

Yongsheng Li (Victor) works in the Power Systems Level 2 Support team for IBM in China. He holds a master degree in Computer Application Technology from Graduate University of Chinese Academy of Sciences. Victor has 15 years of experience working on RS/6000, IBM System Storage®, AIX, HMC, System p, and Power Systems products.

Armin Röll works as a Power Systems IT specialist in Germany. He holds a degree in Experimental Physics from the University of Hamburg, Germany. Armin has 22 years of
experience in Power Systems and AIX pre-sales technical support. He co-authored the AIX Version 4.3.3, the AIX 5L™ Version 5.0, the AIX 5L Version 5.3, the AIX Version 6.1, and the AIX 7.1 Differences Guide IBM Redbooks®.

The project that produced this publication was managed by:

Scott Vetter
PMP

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

Brian Allison, Ron Arroyo, Rich Bireta, Gareth M Coates, Arnold Flores, Nigel Griffiths, Daniel Henderson, Dan Hurlimann, Jeff Jajowka, Roxette Johnson, Vic Mahaney, Charles Marino, Michael Mueller, Kaveh Naderi, Todd Rosedahl, David Sheffield, Alan Standridge, Bill Starke, Jeff Stuecheli, IBM

Now you can become a published author, too!

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

Find out more about the residency program, browse the residency index, and apply online at:
ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

▶ Use the online Contact us review Redbooks form found at:
  ibm.com/redbooks

▶ Send your comments in an email to:
  redbooks@us.ibm.com

▶ Mail your comments to:
  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400
Stay connected to IBM Redbooks

- Find us on Facebook:
  http://www.facebook.com/IBMRedbooks
- Follow us on Twitter:
  http://twitter.com/ibmredbooks
- Look for us on LinkedIn:
  http://www.linkedin.com/groups?home=&gid=2130806
- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:
- Stay current on recent Redbooks publications with RSS Feeds:
  http://www.redbooks.ibm.com/rss.html
Chapter 1. General description

This chapter provides general description of the new IBM Power Systems E980 server, which is a logical follow-on to the IBM Power Systems E880 and E880C servers. Due to the new POWER9 architecture, the Power E980 provides improvements in economics of application delivery and IT services via increased price/performance based on increased throughput, reduced response times, and increased memory and I/O bandwidth.

Figure 1-1 shows an exploded view of a Power E980 system node.
Figure 1-1  An exploded view of an E980 node
1.1 System overview

The Power E980 server offers the next generation of IBM Power Systems servers with IBM POWER9 technology. It is built with innovations that can help deliver security and reliability for the data-intensive workloads of today's enterprises. POWER9 technology is designed from the ground up for data-intensive workloads like databases and analytics.

The Power E980 server is the most powerful and scalable server in the Power Systems portfolio.

- Includes massive throughput, performance, and scalability
- Enables large-scale consolidation of older, under-utilized servers
- Improves infrastructure resilience
- Enables rapid service delivery

The Power E980 provides the following hardware and components and software features:

- Up to 192 POWER9 technology-based processor cores
- Up to 64 TB memory
- Up to 32 PCIe Gen4 x16 slots in system nodes
- Initially up to 48 PCIe Gen3 slots with 4 expansion drawers. This will increase to 192 slots with the support of 16 I/O drawers.
- Up to over 4,000 directly attached SAS disks or solid-state drives (SSDs)
- Up to 1,000 VMs (LPARs) per system
- System control unit, providing redundant system master Flexible Service Processor (FSP)
- Support for IBM AIX, IBM i, and Linux environments
- Capacity on demand (CoD) processor and memory options
- Model upgrades from IBM POWER8®-based E870, E870C, E880, and E880C systems
- Power Enterprise Pools, supporting unsurpassed enterprise flexibility for workload balancing and system

The machine type model of the Power E980 is 9080-M9S. At initial availability in September 2018, up to two system nodes are supported. In November up to four system nodes will be supported.

Table 1-1 shows a comparison of the Power E880C and the Power E980.

<table>
<thead>
<tr>
<th>Features</th>
<th>E880C</th>
<th>E980</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor</td>
<td>POWER8</td>
<td>POWER9</td>
</tr>
<tr>
<td>Sockets</td>
<td>4, 8, 12, 16</td>
<td>4, 8, 12, 16</td>
</tr>
<tr>
<td>Cores</td>
<td>Up to 192 cores</td>
<td>Up to 192 cores</td>
</tr>
<tr>
<td>Maximum Memory</td>
<td>32TB</td>
<td>64TB</td>
</tr>
<tr>
<td>Memory Bandwidth</td>
<td>920 GB/sec / drawer</td>
<td>920 GB/sec / drawer</td>
</tr>
<tr>
<td>Predecessor Memory Migration</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>PCIe slots</td>
<td>8 PCIe GEN3 slots / drawer</td>
<td>8 PCIe GEN4 slots / drawer</td>
</tr>
</tbody>
</table>
1.1.1 System control unit

The 2U system control unit (SCU) provides redundant system master service processors (MSPs). Additionally, it contains the Operator Panel and the System vital product data (VPD). One SCU is required for each server. Two service processors (FSPs) in the system control unit are ordered using two #EFFP features. All system nodes connect to the system control unit using the cable features #EFCA, #EFCB, #EFCC, and #EFCD. These cable features also include SMP cables for the server.
The SCU is powered from the system nodes. Two UPIC cables provide redundant power to the SCU. In a single drawer system both UPIC cables will originate from drawer 1. For system with a two or more drawers one UPIC cable will originate from drawer 1 and the other UPIC cable will originate from drawer 2. Just one UPIC cord is enough to power the SCU and the other is in place for redundancy.

SCU highlights:

- Eliminated Clock Cabling
- Provides Front Accessible USB port
- Reduced UPIC Power cabling
- Optional external DVD
- Concurrently maintainable Time of Day clock
- Concurrently maintainable FSPs

Figure 1-3 on page 5 shows the front and rear view of an SCU. The locations of the connectors and features are indicated.

1.1.2 System nodes

Each 5 EIA or 5U system node of the system has four air-cooled Single Chip Modules (SCMs) optimized for performance and scalability. The Power E980 SCMs can have eight, ten, eleven, or twelve POWER9 cores running at up to 4.0 GHz and simultaneous multithreading executing up to eight threads per core. Each SCM has dual memory controllers to support up to 128 GB off-chip eDRAM L4 cache to deliver up to 230 GBps of sustained memory bandwidth or 920 GBps per node. Up to 410 GBps of peak memory bandwidth from the L4 cache to memory DIMMs is provided per SCM or up to 1640 GBps per node.
node. Using tri PCIe Gen4 I/O controllers, which are also integrated on to each SCM to further reduce latency, up to 545.25 GBps I/O bandwidth is available per node. Thus a Power E980 system bandwidth can help deliver over twice the performance per core of competitors, enabling applications to run faster and be more responsive.

Each system node has 32 Custom DIMM (CDIMM) slots and can support up to 16 TB of DDR4 memory. Thus a four-node server can have up to 64 TB of memory. Each system node has eight PCIe slots, which are all PCIe Gen4 x16, low profile. Therefore a four-node server can have up to 32 PCIe slots. PCIe expansion units can optionally expand the number of PCIe slots on the server.

A system node is ordered using a processor feature. Each processor feature will deliver a set of four identical SCMs in one system node. All processor features in the system must be identical. Cable features are required to connect system node drawers to the system control unit and to other system nodes.

Figure 1-4 on page 6 shows the front view of a system node. Fans and Power Supply Units (PSUs) are redundant and concurrently maintainable. Fans are n+1 redundant meaning the system will continue to function when any one fan fails. Power supplies are n+2 redundant meaning the system will continue to function even if any two power supplies fail.

![Figure 1-4 Front view of a system node](image)

Figure 1-5 shows the rear view of a system node. The locations of the connectors and features are indicated.
Figure 1-5   Rear view of a system node

Figure 1-6 on page 8 shows the top view of a system node with the top lid removed.
1.1.3 Hardware components highlight

- Eight, ten, eleven, or twelve POWER9 processors per socket
- Up to 192 POWER9 processor cores; up to 64 TB of 1600 MHz, DDR4 DRAM memory, and eight PCIe Gen4 x16 I/O expansion slots per system node enclosure, with a maximum of 32 per system
- Redundant Clocking in each system node.
- Four NVMe drive bays per system node for boot purposes.
- Integrated USB ports
- System control unit, providing redundant system master Flexible Service Processor (FSP) and support for the operator panel, the system VPD, and external attached DVD
- 19-inch PCIe Gen 3 4U I/O expansion drawer and PCIe FanOut modules, initially supporting a maximum of 48 PCIe slots (2 I/O expansion drawers per node, has max of 2 nodes). This will increase to 192 slots with the support of 16 I/O drawers.
- PCIe Gen1, Gen2, Gen3, and Gen4 adapter cards, supported in the system node and PCIe Gen1, Gen2, and Gen3 adapter cards, supported in I/O expansion drawer
- EXP24SX SFF Drawer with twenty-four 2.5-inch form-factor SAS bays
- Dynamic LPAR support for adjusting workload placement of processor and memory resources
- Active Memory™ Expansion that is assisted by the processor chip
- Power Enterprise Pools that support unsurpassed enterprise flexibility for workload balancing and system maintenance

Figure 1-6  Top view of a system node with top lid removed.
1.2 Operating environment

Table 1-2 details the operating environment for the Power E980 server.

<table>
<thead>
<tr>
<th>Power E980 operating environment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>System</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Temperature</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Relative humidity</td>
</tr>
<tr>
<td>Maximum dew point</td>
</tr>
<tr>
<td>Operating voltage</td>
</tr>
<tr>
<td>Operating frequency</td>
</tr>
<tr>
<td>Maximum power consumption</td>
</tr>
<tr>
<td>Maximum power source loading</td>
</tr>
<tr>
<td>Maximum thermal output</td>
</tr>
<tr>
<td>Maximum altitude</td>
</tr>
<tr>
<td>Maximum noise level</td>
</tr>
</tbody>
</table>

**Environmental assessment:** The IBM Systems Energy Estimator tool can provide more accurate information about power consumption and thermal output of systems based on a specific configuration, including adapter cards and I/O expansion drawers. The IBM Systems Energy Estimator tool can be accessed online:

http://www-912.ibm.com/see/EnergyEstimator

The Power E980 must be installed in a rack with a rear door and side panels for EMC compliance. The native HMC Ethernet ports must use shielded Ethernet cables.

Government regulations, such as those prescribed by OSHA or European Community Directives, may govern noise level exposure in the workplace and may apply to you and your server installation. This IBM system is available with an optional acoustical door feature that can help reduce the noise emitted from this system. The actual sound pressure levels in your installation depend upon a variety of factors, including the number of racks in the installation; the size, materials, and configuration of the room where you designate the racks to be installed; the noise levels from other equipment; the room ambient temperature, and employees' location in relation to the equipment. Further, compliance with such government
regulations also depends upon a variety of additional factors, including the duration of employees’ exposure and whether employees wear hearing protection. IBM recommends that you consult with qualified experts in this field to determine whether you are in compliance with the applicable regulations.

1.3 Physical package

The Power E980 is a modular system that can be constructed from a single system control unit and one, two, three, or four system nodes.

The system control unit requires 2U and each system node requires 5U. Thus, a single-enclosure system requires 7U, a two-enclosure system requires 12U, a three-enclosure system requires 17U, and a four-enclosure system requires 22U.

Table 1-3 lists the physical dimensions of the system node, SCU, and the PCIe Gen3 I/O expansion drawer.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Power E980 system node</th>
<th>Power E980 SCU</th>
<th>PCIe I/O expansion drawer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Width</td>
<td>445.5 mm (17.54 in.)</td>
<td>445.6 mm (17.54 in.)</td>
<td>447.3 mm (17.61 in.)</td>
</tr>
<tr>
<td>Depth</td>
<td>867 mm (34.13 in.)</td>
<td>779.7 mm (30.7 in.)</td>
<td>737 mm (29.0 in.)</td>
</tr>
<tr>
<td>Height</td>
<td>218 mm (8.58 in.) 5 EIA units</td>
<td>86 mm (3.39 in.) 2 EIA units</td>
<td>173 mm (6.8 in.) 4 EIA units</td>
</tr>
<tr>
<td>Weight</td>
<td>86.2 kg (190 lb)</td>
<td>22.7 kg (50 lb)</td>
<td>54.4 kg (120 lb)</td>
</tr>
</tbody>
</table>

1.3.1 Lift tools

It is recommended to have a lift tool available at each site where one or more Power E980s are located to avoid any delays when servicing systems. An optional lift tool #EB2Z is available for order with a Power E980 system. One #EB2Z lift tool can be shared among many servers and I/O drawers. The #EB2Z lift tool provides a hand crank to lift and position up to 159 kg (350 lb). The #EB2Z lift tool is 1.12 meters x 0.62 meters (44 in. x 24.5 in.). Note that a single system node can weigh up to 86.2 kg (190 lb). Also available are a lighter, lower cost lift tool #EB3Z and wedge shelf tool kit for #EB3Z: #EB4Z.

1.4 System features

The following convention is used in the Order type column in all tables in this section.

<table>
<thead>
<tr>
<th>Initial</th>
<th>Only available when ordered as part of a new system</th>
</tr>
</thead>
<tbody>
<tr>
<td>MES</td>
<td>Only available as an MES upgrade</td>
</tr>
<tr>
<td>Both</td>
<td>Available with a new system or as part of an upgrade</td>
</tr>
<tr>
<td>Supported</td>
<td>Unavailable as a new purchase but supported when migrated from another system or as part of a model conversion</td>
</tr>
</tbody>
</table>
You can order system features during the initial system order. You can also add or replace features later on as a miscellaneous equipment specifications (MES). An MES is a hardware change which involves adding, removing, or changing features.

For more information on the available features, see Chapter 2, “Architecture and technical overview” on page 47.

The following features are available on the Power E980 server:

- One to four 5U system nodes
- One 2U system control unit
- One to four processor features per system with four single-chip modules (SCMs) per feature. Each processor feature requires a system node
- 32 CDIMM slots per system node. There are 4 CDIMMs per memory feature
- Active Memory Expansion, assisted by the processor chip (#EM89)
- Eight PCIe Gen4 x16 I/O low-profile expansion slots per system node (maximum 32 in a 4-node system)
- Redundant 2+2 hot-swap AC power supplies in each system node drawer
- Four Hardware Management Console (HMC) ports in system control unit
- Optional PCIe I/O Expansion Drawer with PCIe slots:
  - Zero to four drawers per system node drawer (#EMX0).
  - Each I/O drawer holds one or two 6-slot PCIe Fan-out Modules (#EMXF).
  - Each Fan-out Module attaches to the system node through a PCIe Optical Cable Adapter (#EJ07).

**Note:** At initial availability in September 2018 zero, one, or two PCIe I/O Expansion drawers are supported per system node. In November 2018 zero, one, two, three, or four expansion drawers are supported per system node.

### 1.4.1 Minimum configuration

The minimum configurations is a single system node with 8-core processors, 512 GB of memory, four power supplies and four power cords, an SCU, an operating system indicator, and Language Group Specify.

*Table 1-4 shows the minimum configuration of Power E980*

<table>
<thead>
<tr>
<th>Components</th>
<th>FC</th>
<th>Single Drawer System</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Drawer System</td>
<td>EFN1</td>
<td>1</td>
<td>Base Unit</td>
</tr>
<tr>
<td>Processors 8-cores</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1FC=4processors)</td>
<td>EFP1</td>
<td>4</td>
<td>Four processors per Feature Code EFB1</td>
</tr>
<tr>
<td>CDIMM - Min size is 32GB</td>
<td>EF20</td>
<td>4xEF20</td>
<td></td>
</tr>
</tbody>
</table>
### 1.4.2 Power supply features

The following are the key power supply features:

- **AC power supply - 1950 Watt for Server (200-240V AC):**
  
  Each system drawer has four 1950W bulk power supplies, which provides N+2 redundancy for system bulk power, and the system can continue to operate at full function in nominal mode with any two of the power supplies functioning. Power supplies are hot-swappable allowing replacement of a failed unit without system interruption.

- **AC power supply Conduit for optional PCIe3 Expansion Drawer (#EMXA):**
  
  Provides two 320-C14 inlet electrical connections for two separately ordered AC power cords with C13 connector plugs. Conduit provides electrical power connection between two power supplies located in the front of a PCIe Gen3 I/O Expansion Drawer (#EMX0) and two power cords which connect on the rear of the PCIe Gen3 I/O Expansion Drawer.

- **Specify AC power supply for optional EXP12SX/EXP24SX Storage Enclosure (#ESLA):**
  
  The power supply has a 320-C14 inlet electrical connection for a separately ordered power cord.

### 1.4.3 Processor module features

One to four processor features per system with four single-chip modules (SCMs) per feature:

- **Typical 3.9 to 4.0 GHz 32-core POWER9 processor (#EFP1)**
- **Typical 3.7 to 3.9 GHz, 40-core POWER9 processor (#EFP2)**
- **Typical 3.58 to 3.9 GHz, 44-core POWER9 processor (#EFP4)**
- **Typical 3.55 to 3.9 GHz, 48-core POWER9 processor (#EFP3)**

CoD processor core activation features available on a per-core basis. Each Power E980 system requires a minimum of permanent processor core activations using either static
activations or Power Linux activations that matches the number of processor cores per processor feature. This minimum is per system, not per node. The rest of the cores can be permanently or temporary activated or remain inactive (dark) until needed.

For example, if the core count per socket is 12, then the minimum active cores is 12 for the entire system.

If the count of cores per processor is 8 per socket then a minimum of 8 core activations per system is required.

The activations are not specific to hardware cores or SCMs or nodes. They are known to the system as a total number of activations of different types and used or assigned by the Power hypervisor appropriately.

A variety of activations fit different usage and pricing options. Static activations are permanent and support any type of application environment on this server. Mobile activations are ordered against a specific server, but can be moved to any server within the Power Enterprise Pool and can support any type of application. Mobile-enabled activations are technically static, but can be converted to mobile at no charge when logistically or administratively eligible. Power Linux activations can only run Linux workloads. Temporary activations are used for CoD - Elastic, Utility, and Trial options.

1.4.4 POWER9 processor highlights

- Four processor offerings available (SMT8 cores) - offered both GAs:
  - 12-core, 11-core processor (maximum throughput)
  - 10-core processor
  - 8-core processor (maximum core performance)
- Processor frequencies dynamic for maximum performance
- Capacity on Demand support
- Increased processor to processor fabric interconnect
  - 16 Gb/s X-Bus Fully connected fabric within CEC Drawer
  - 4x increase in O-Bus fabric for Drawer to Drawer interconnect or Accelerator

Figure 1-7 shows P9 processor to processor fabric interconnect

Figure 1-7 O-Bus and X-Bus fabric for Drawer to Drawer interconnect or Accelerator.
1.4.5 Memory features

IBM-custom DIMMs (CDIMMs) are extremely high-performance, high-reliability, high-function memory cards that contain L4 cache, intelligence, and 1600 MHz DRAM memory. Both DDR3 and DDR4 technology is employed and both provide the same 1600 MHz performance. CDIMMs are placed in CDIMM slots in the system node.

Each system node has 32 memory CDIMM slots and at least half of the memory slots must be physically filled.

Eight CDIMM slots are local to each of the four SCMs in the server, but SCMs and their cores have access to all the other memory in the server.

At least half of the eight memory slots for each SCM must physically be filled.

When filling the other four memory slots in the SCM, a quantity of four CDIMMs must be used. Thus the CDIMM slots of the SCMs are either 50% or 100% filled.

The system node (each with four SCMs) CDIMM slots can have 16, 20, 24, 28, or 32 CDIMMs physically installed (quad plugging rules). Memory cards features are as below:

- DDR4 1600 MHz CDIMM memory cards:
  - 128 GB (4 x 32 GB), (#EF20)
  - 256 GB (4 x 64 GB), (#EF21)
  - 512 GB (4 x 128 GB), (#EF22)
  - 1024 GB (4 x 256 GB), (#EF23)
  - 2048 GB (4 x 512 GB), (#EF24)

To assist with the quad plugging rules above, four CDIMMs are ordered using one memory feature.

All CDIMMs must be identical on the same SCM. If using eight CDIMMs, both memory features on an SCM must be identical. A different SCM in the same system node can use a different memory feature. For example, one system node could technically use 128 GB, 256 GB, 512 GB, 1024 GB and 2048 GB memory features. Note that DDR3 and DDR4 memory cannot be mixed on the same system node. DDR3 memory is only available when transferred from an E870, E870C, E880, or E880C system as part of an upgrade.

To provide more flexible pricing, memory activations are ordered separately from the physical memory and can be permanent or temporary. Activation features can be used on DDR4 memory features and used on any size memory feature. Activations are not specific to a CDIMM, but are known as a total quantity to the server. The Power hypervisor determines what physical memory to use.

- CoD memory activation features include:
  - 1 GB (static) Memory Activations (#EMAT)
  - 100 GB (static) Memory Activations (#EMAU)
  - 100 GB Mobile Memory Activations (#EMAV)
  - 100 GB Mobile Enabled Memory Activations (#EMAD)
  - 512 GB Memory Activations for Linux (#ELMD)
  - 8 TB activations (#EMBA) ordered with 8 TB memory using feature EM8Y CDIMMs
  - 4 TB activations (#EMB7) ordered with 4 TB memory using the feature EMB6 package
A minimum of 50% of the total physical memory capacity of a server must have permanent memory activations ordered for that server. For example, a server with a total of 8 TB of physical memory must have at least 4 TB of permanent memory activations ordered for that server.

These activations can be static, mobile-enabled, mobile, or Power Linux. At least 25% must be static activations or Power Linux activations. For example, a server with a total of 8 TB physical memory must have at least 2 TB of static activations or Power Linux activations. The 50% minimum cannot be fulfilled using mobile activations ordered on a different server.

The minimum activations ordered with MES orders of additional physical memory features will depend on the existing total installed physical memory capacity and the existing total installed memory activation features. If you already have installed more than 50% activations for your existing system, then you can order less than 50% activations for the MES ordered memory. The resulting configuration after the MES order of physical memory and any MES activations must meet the same 50% and 25% minimum rules above.

For the best possible performance, it is generally recommended that memory be installed evenly across all system node drawers and all SCMs in the system. Balancing memory across the installed system planar cards enables memory access in a consistent manner and typically results in better performance for your configuration.

Though maximum memory bandwidth is achieved by filling all the memory slots, plans for future memory additions should be taken into account when deciding which memory feature size to use at the time of initial system order.

IBM Active Memory Expansion (AME) is an option that can increase the effective memory capacity of the system. See the AME information later in this section.

Memory Subsystem Highlights

- 230 GB/s Memory Bandwidth and up to 4TB per socket
- 8 DIMM slots per socket, 32 DIMMs per drawer, same CDIMM technology as POWER8
- Quad DIMM granularity with same plug rules as POWER8
- 32, 64, 128, 256, 512 GB DDR4 CDIMMs (16GB DDR4 CDIMM migrate support)
- Memory buffer on CDIMM to improve performance
- Support for migrating POWER8 DDR3 (16GB, 32GB, 64GB, 128GB) and DDR4 CDIMMs
- Mixing DDR3 and DDR4 supported but drawers must be homogeneous
- Capacity on Demand support

Figure 1-8 shows an example of DRAM access via Memory Buffer
1.4.6 System node PCIe slots

Each system node has eight PCIe Gen4 x16 hot-plug slots, a total of 32 slots for a 4-node system.

- All adapter slots in the system node are half-height, half length. The slots are labeled C1 through C8.
- These PCIe slots can be used for either low-profile PCIe adapters or for attaching a PCIe I/O drawer.
- A blind swap cassette (BSC) is used to house the low-profile adapters that go into these slots. The server is shipped with a full set of BSCs, even if the BSCs are empty. A feature
number to order additional low-profile BSCs is not required or announced. Blind swap cassettes allows hot (system running) replace, remove and add for an adapter without having to place a server in a service position or open the drawer. The server BSCs are not the same as the ones used in the I/O drawers.

- If additional PCIe slots beyond the system node slots are required, a system node x16 slot is used to attach a six-slot expansion module in the I/O drawer. An I/O drawer holds two expansion modules that are attached to any two x16 PCIe slots in the same system node or in different system nodes.
- PCIe Gen1, Gen2, and Gen3 adapter cards are supported in these Gen4 slots. The set of PCIe adapters that are supported is found in 2.5, “PCIe adapters” on page 85.
- Concurrent repair and add/removal of PCIe adapter cards is done by HMC-guided menus or by operating system support utilities.
- The system nodes sense which IBM PCIe adapters are installed in the PCIe slots. If an adapter requires higher levels of cooling, fans automatically speed up to increase airflow across these PCIe adapters.
- Each system node supports CAPI 2.0 adapters in all slots. At initial availability in September 2018 no IBM developed or IBM Part Number CAPI adapters supported. The ports have been tested and adapters are available from our Partners such as:
  - Nallatech 250SP [http://www.nallatech.com/250sp](http://www.nallatech.com/250sp)
  - Flyslice FX609
  - ReflexCES XpressVUP LP9P

System node PCIe slots highlights:

- PCIe Gen4:
  - Double the I/O rates of Gen3 slots
  - Gen4 for next generation adapters requiring link speeds in excess of 40Gbps
- Initially most adapters will be Gen3:
  - Fast enough for current adapter speed requirements
  - Can use in Gen3 or Gen4 adapters slots

### 1.4.7 USB

Each system node in the Power E980 provides three USB 3.0 ports for general use by the LPAR or VIOS that they are assigned to. On the first system node, one of the USB ports is rerouted to a port on the front of the SCU. This port is primarily intented for use when attaching a USB DVD drive.

Table 1-8 lists options for USB-attached media available for the Power E980 server.

<table>
<thead>
<tr>
<th>Feature Code</th>
<th>CCIN</th>
<th>Description</th>
<th>Max</th>
<th>OS support</th>
<th>Order typea</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUA5</td>
<td>63BD</td>
<td>Standalone USB DVD drive w/cable</td>
<td>1</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EUA4</td>
<td></td>
<td>RDX USB External Docking Stationb</td>
<td>6</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
</tbody>
</table>
1.4.8 Disk and media features

The Power E980 server can support up to four 800 GB hot-plug mainstream NVMe SSDs (#EC5J) in each system node. The NMCe drives are intended for boot purposes and not as general purpose storage devices. Each NVMe device contains its own controller and connects to the system using PCIe ports. Figure 1-10 on page 18 shows a picture of two NVMe devices.

Figure 1-10  Two NVMe devices

Additional direct attached storage is supported in external disk and media drawers. Table 1-6 lists available disk and SSD features for the Power E980.

Table 1-6  Disk features available on the Power E980

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>Max</th>
<th>OS support</th>
<th>Order type</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES78</td>
<td>5B16</td>
<td>387 GB SFF-2 SSD 5xx eMLC4 for AIX/Linux</td>
<td>2016</td>
<td>AIX and Linux</td>
<td>Supported</td>
</tr>
<tr>
<td>ES79</td>
<td>5B16</td>
<td>387 GB SFF-2 SSD 5xx eMLC4 for IBM i</td>
<td>2016</td>
<td>IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>ES7E</td>
<td>5B17</td>
<td>775 GB SFF-2 SSD 5xx eMLC4 for AIX/Linux</td>
<td>2016</td>
<td>AIX and Linux</td>
<td>Supported</td>
</tr>
<tr>
<td>ES7F</td>
<td>5B17</td>
<td>775 GB SFF-2 SSD 5xx eMLC4 for IBM i</td>
<td>2016</td>
<td>IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>ES85</td>
<td>5B10</td>
<td>387 GB SFF-2 SSD 4k eMLC4 for AIX/Linux</td>
<td>2016</td>
<td>AIX and Linux</td>
<td>Supported</td>
</tr>
<tr>
<td>ES86</td>
<td>5B10</td>
<td>387 GB SFF-2 SSD 4k eMLC4 for IBM i</td>
<td>2016</td>
<td>IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>ES8C</td>
<td>5B11</td>
<td>775 GB SFF-2 SSD 4k eMLC4 for AIX/Linux</td>
<td>2016</td>
<td>AIX and Linux</td>
<td>Supported</td>
</tr>
<tr>
<td>ES8D</td>
<td>5B11</td>
<td>775 GB SFF-2 SSD 4k eMLC4 for IBM i</td>
<td>2016</td>
<td>IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>ES8F</td>
<td>5B12</td>
<td>1.55 TB SFF-2 SSD 4k eMLC4 for AIX/Linux</td>
<td>2016</td>
<td>AIX and Linux</td>
<td>Supported</td>
</tr>
<tr>
<td>Feature code</td>
<td>CCIN</td>
<td>Description</td>
<td>Max</td>
<td>OS support</td>
<td>Order type</td>
</tr>
<tr>
<td>--------------</td>
<td>-------</td>
<td>-------------------------------------------------------------------------------</td>
<td>------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>ES8G</td>
<td>5B12</td>
<td>1.55 TB SFF-2 SSD 4k eMLC4 for IBM i</td>
<td>2016</td>
<td>IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>ES8Y</td>
<td>5B29</td>
<td>931 GB Mainstream SAS 4k SFF-2 SSD for AIX/Linux</td>
<td>2016</td>
<td>Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ES8Z</td>
<td>5B29</td>
<td>931 GB Mainstream SAS 4k SFF-2 SSD for IBM i</td>
<td>2016</td>
<td>IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>ES96</td>
<td>5B21</td>
<td>1.86 TB Mainstream SAS 4k SFF-2 SSD for AIX/Linux</td>
<td>2016</td>
<td>Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ES97</td>
<td>5B21</td>
<td>1.86 TB Mainstream SAS 4k SFF-2 SSD for IBM i</td>
<td>2016</td>
<td>IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>ESE7</td>
<td>5B2D</td>
<td>3.72 TB Mainstream SAS 4k SFF-2 SSD for AIX/Linux</td>
<td>2016</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ESE8</td>
<td>5B2D</td>
<td>3.72 TB Mainstream SAS 4k SFF-2 SSD for IBM i</td>
<td>2016</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ESG5</td>
<td>5B16</td>
<td>387 GB Enterprise SAS 5xx SFF-2 SSD for AIX/Linux</td>
<td>2016</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ESG6</td>
<td>5B16</td>
<td>387 GB Enterprise SAS 5xx SFF-2 SSD for IBM i</td>
<td>2016</td>
<td>AIX and IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>ESGB</td>
<td>5B10</td>
<td>387 GB Enterprise SAS 4k SFF-2 SSD for AIX/Linux</td>
<td>2016</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ESGC</td>
<td>5B10</td>
<td>387 GB Enterprise SAS 4k SFF-2 SSD for IBM i</td>
<td>2016</td>
<td>AIX and IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>ESGF</td>
<td>5B17</td>
<td>775 GB Enterprise SAS 5xx SFF-2 SSD for AIX/Linux</td>
<td>2016</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ESGG</td>
<td>5B17</td>
<td>775 GB Enterprise SAS 5xx SFF-2 SSD for IBM i</td>
<td>2016</td>
<td>AIX and IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>ESGK</td>
<td>5B11</td>
<td>775 GB Enterprise SAS 4k SFF-2 SSD for AIX/Linux</td>
<td>2016</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ESGL</td>
<td>5B11</td>
<td>775 GB Enterprise SAS 4k SFF-2 SSD for IBM i</td>
<td>2016</td>
<td>AIX and IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>ESGP</td>
<td>5B12</td>
<td>1.55 TB Enterprise SAS 4k SFF-2 SSD for AIX/Linux</td>
<td>2016</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ESGQ</td>
<td>5B12</td>
<td>1.55 TB Enterprise SAS 4k SFF-2 SSD for IBM i</td>
<td>2016</td>
<td>AIX and IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>ESHJ</td>
<td>5B29</td>
<td>931 GB Mainstream SAS 4k SFF-2 SSD for AIX/Linux</td>
<td>2016</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ESHK</td>
<td>5B29</td>
<td>931 GB Mainstream SAS 4k SFF-2 SSD for IBM i</td>
<td>2016</td>
<td>AIX and IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>ESHL</td>
<td>5B21</td>
<td>1.86 TB Mainstream SAS 4k SFF-2 SSD for AIX/Linux</td>
<td>2016</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ESHM</td>
<td>5B21</td>
<td>1.86 TB Mainstream SAS 4k SFF-2 SSD for IBM i</td>
<td>2016</td>
<td>AIX and IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>ESHN</td>
<td>5B2F</td>
<td>7.45 TB Mainstream SAS 4k SFF-2 SSD for AIX/Linux</td>
<td>2016</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ESM8</td>
<td>5B2D</td>
<td>3.72 TB Mainstream SAS 4k SFF-2 SSD for AIX/Linux</td>
<td>2016</td>
<td>Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ESM9</td>
<td>5B2D</td>
<td>3.72 TB Mainstream SAS 4k SFF-2 SSD for IBM i</td>
<td>2016</td>
<td>IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>1948</td>
<td>19B1</td>
<td>283 GB 15k RPM SAS SFF-2 Disk Drive (IBM i)</td>
<td>4032</td>
<td>IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>1953</td>
<td>19B1</td>
<td>300 GB 15k RPM SAS SFF-2 Disk Drive (AIX/Linux)</td>
<td>4032</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>1962</td>
<td>19B3</td>
<td>571 GB 10k RPM SAS SFF-2 Disk Drive (IBM i)</td>
<td>4032</td>
<td>IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>1964</td>
<td>19B3</td>
<td>600 GB 10k RPM SAS SFF-2 Disk Drive (AIX/Linux)</td>
<td>4032</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ES62</td>
<td>5B1D</td>
<td>3.86-4.0 TB 7200 RPM 4K SAS LFF-1 Nearline Disk Drive (AIX/Linux)</td>
<td>2016</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>Feature code</td>
<td>CCIN</td>
<td>Description</td>
<td>Max</td>
<td>OS support</td>
<td>Order type(^a)</td>
</tr>
<tr>
<td>--------------</td>
<td>------</td>
<td>-------------</td>
<td>-----</td>
<td>------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>ES64</td>
<td>5B1F</td>
<td>7.72-8.0 TB 7200 RPM 4K SAS LFF-1 Nearline Disk Drive (AIX/Linux)</td>
<td>2016</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ESD2</td>
<td>59CD</td>
<td>1.1 TB 10K RPM SAS SFF-2 Disk Drive (IBMi)</td>
<td>4032</td>
<td>IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>ESD3</td>
<td>59CD</td>
<td>1.2 TB 10K RPM SAS SFF-2 Disk Drive (AIX/Linux)</td>
<td>4032</td>
<td>AIX and Linux</td>
<td>Supported</td>
</tr>
<tr>
<td>ESDN</td>
<td>59CF</td>
<td>571 GB 15K RPM SAS SFF-2 Disk Drive - 528 Block (IBM i)</td>
<td>4032</td>
<td>IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>ESDP</td>
<td>59CF</td>
<td>600 GB 15K RPM SAS SFF-2 Disk Drive - 5xx Block (AIX/Linux)</td>
<td>4032</td>
<td>AIX and Linux</td>
<td>Supported</td>
</tr>
<tr>
<td>ESEU</td>
<td>59D2</td>
<td>571 GB 10K RPM SAS SFF-2 Disk Drive 4K Block - 4224</td>
<td>4032</td>
<td>IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>ESEV</td>
<td>59D2</td>
<td>600 GB 10K RPM SAS SFF-2 Disk Drive 4K Block - 4096</td>
<td>4032</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ESEY</td>
<td>59C9</td>
<td>283 GB 15K RPM SAS SFF-2 4K Block - 4224 Disk Drive</td>
<td>4032</td>
<td>IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>ESEZ</td>
<td>59C9</td>
<td>300 GB 15K RPM SAS SFF-2 4K Block - 4096 Disk Drive</td>
<td>4032</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ESF2</td>
<td>59DA</td>
<td>1.1 TB 10K RPM SAS SFF-2 Disk Drive 4K Block - 4224</td>
<td>4032</td>
<td>IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>ESF3</td>
<td>59DA</td>
<td>1.2 TB 10K RPM SAS SFF-2 Disk Drive 4K Block - 4096</td>
<td>4032</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ESFN</td>
<td>59CC</td>
<td>571 GB 15K RPM SAS SFF-2 4K Block - 4224 Disk Drive</td>
<td>4032</td>
<td>IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>ESFP</td>
<td>59CC</td>
<td>600 GB 15K RPM SAS SFF-2 4K Block - 4096 Disk Drive</td>
<td>4032</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ESFS</td>
<td>59DD</td>
<td>1.7 TB 10K RPM SAS SFF-2 Disk Drive 4K Block - 4224</td>
<td>4032</td>
<td>IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>ESFT</td>
<td>59DD</td>
<td>1.8 TB 10K RPM SAS SFF-2 Disk Drive 4K Block - 4096</td>
<td>4032</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ESNL</td>
<td>5B43</td>
<td>283 GB 15K RPM SAS SFF-2 4k Block Cached Disk Drive (IBM i)</td>
<td>4032</td>
<td>IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>ESNM</td>
<td>5B43</td>
<td>300 GB 15K RPM SAS SFF-2 4k Block Cached Disk Drive (AIX/Linux)</td>
<td>4032</td>
<td>Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ESNQ</td>
<td>5B47</td>
<td>571 GB 15K RPM SAS SFF-2 4k Block Cached Disk Drive (IBM i)</td>
<td>4032</td>
<td>IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>ESNR</td>
<td>5B47</td>
<td>600 GB 15K RPM SAS SFF-2 4k Block Cached Disk Drive (AIX/Linux)</td>
<td>4032</td>
<td>Linux</td>
<td>Both</td>
</tr>
</tbody>
</table>

a. See 1.4.8, “Disk and media features” on page 18 for information on order types.

Table 1-7 shows disk and SSD features available for bulk ordering.
### Table 1-7  Bulk disk and SSD features

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>Max</th>
<th>OS support</th>
<th>Order typea</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQ78</td>
<td>5B16</td>
<td>Quantity 150 of #ES78 387 GB SFF-2 SSD 5xx</td>
<td>13</td>
<td>AIX and Linux</td>
<td>Supported</td>
</tr>
<tr>
<td>EQ79</td>
<td>5B16</td>
<td>Quantity 150 of #ES79 387 GB SFF-2 SSD 5xx</td>
<td>13</td>
<td>IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>EQ7E</td>
<td>5B17</td>
<td>Quantity 150 of #ES7E 775 GB SFF-2 SSD 5xx</td>
<td>13</td>
<td>AIX and Linux</td>
<td>Supported</td>
</tr>
<tr>
<td>EQ7F</td>
<td>5B17</td>
<td>Quantity 150 of #ES7F 775 GB SFF-2 SSD 5xx</td>
<td>13</td>
<td>IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>EQ85</td>
<td>5B10</td>
<td>Quantity 150 of #ES85 387 GB SFF-2 SSD 4k</td>
<td>13</td>
<td>AIX and Linux</td>
<td>Supported</td>
</tr>
<tr>
<td>EQ86</td>
<td>5B10</td>
<td>Quantity 150 of #ES86 387 GB SFF-2 SSD 4k</td>
<td>13</td>
<td>IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>EQ8C</td>
<td>5B11</td>
<td>Quantity 150 of #ES8C 775 GB SFF-2 SSD 4k</td>
<td>13</td>
<td>AIX and Linux</td>
<td>Supported</td>
</tr>
<tr>
<td>EQ8D</td>
<td>5B11</td>
<td>Quantity 150 of #ES8D 775 GB SFF-2 SSD 4k</td>
<td>13</td>
<td>IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>EQ8F</td>
<td>5B12</td>
<td>Quantity 150 of #ES8F 1.55 TB SFF-2 SSD 4k</td>
<td>13</td>
<td>AIX and Linux</td>
<td>Supported</td>
</tr>
<tr>
<td>EQ8G</td>
<td>5B12</td>
<td>Quantity 150 of #ES8 G 1.55 TB SFF-2 SSD 4k</td>
<td>13</td>
<td>IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>EQ8Y</td>
<td>5B29</td>
<td>Quantity 150 of #ES8Y 931 GB SFF-2 SSD 4k</td>
<td>13</td>
<td>Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EQ8Z</td>
<td>5B29</td>
<td>Quantity 150 of #ES8Z 931 GB SFF-2 SSD 4k</td>
<td>13</td>
<td>IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>EQ96</td>
<td>5B21</td>
<td>Quantity 150 of #ES96 1.86 TB SFF-2 SSD 4k</td>
<td>13</td>
<td>Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EQ97</td>
<td>5B21</td>
<td>Quantity 150 of #ES97 1.86 TB SFF-2 SSD 4k</td>
<td>13</td>
<td>IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>EQE7</td>
<td>5B2D</td>
<td>Quantity 150 of #ESE7 3.72 TB SFF-2 SSD 4k</td>
<td>13</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EQE8</td>
<td>5B2D</td>
<td>Quantity 150 of #ESE8 3.72 TB SFF-2 SSD 4k</td>
<td>13</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EQG5</td>
<td>5B16</td>
<td>Quantity 150 of #ESG5 (387 GB SAS 5xx)</td>
<td>13</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EQG6</td>
<td>5B16</td>
<td>Quantity 150 of #ESG6 (387 GB SAS 5xx)</td>
<td>13</td>
<td>AIX and IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>EQGB</td>
<td>5B10</td>
<td>Quantity 150 of #ESGB (387 GB SAS 4k)</td>
<td>13</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EQGC</td>
<td>5B10</td>
<td>Quantity 150 of #ESGC (387 GB SAS 4k)</td>
<td>13</td>
<td>AIX and IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>EQGF</td>
<td>5B17</td>
<td>Quantity 150 of #ESGF (775 GB SAS 5xx)</td>
<td>13</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EQGG</td>
<td>5B17</td>
<td>Quantity 150 of #ESGG (775 GB SAS 5xx)</td>
<td>13</td>
<td>AIX and IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>EQGK</td>
<td>5B11</td>
<td>Quantity 150 of #ESGK (775 GB SAS 4k)</td>
<td>13</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EQGL</td>
<td>5B11</td>
<td>Quantity 150 of #ESGL (775 GB SAS 4k)</td>
<td>13</td>
<td>AIX and IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>EQGP</td>
<td>5B12</td>
<td>Quantity 150 of #ESGP (1.55 TB SAS 4k)</td>
<td>13</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EQGQ</td>
<td>5B12</td>
<td>Quantity 150 of #ESGQ (1.55 TB SAS 4k)</td>
<td>13</td>
<td>AIX and IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>ERHJ</td>
<td>5B29</td>
<td>Quantity 150 of #ESHJ 931 GB SSD 4k SFF-2</td>
<td>13</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ERHK</td>
<td>5B29</td>
<td>Quantity 150 of #ESHK 931 GB SSD 4k SFF-2</td>
<td>13</td>
<td>AIX and IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>ERHL</td>
<td>5B21</td>
<td>Quantity 150 of #ESHL 1.86 TB SSD 4k SFF-2</td>
<td>13</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ERHM</td>
<td>5B21</td>
<td>Quantity 150 of #ESHM 1.86 TB SSD 4k SFF-2</td>
<td>13</td>
<td>AIX and IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>ERHN</td>
<td>5B2F</td>
<td>Quantity 150 of #ESHN 7.45 TB SSD 4k SFF-2</td>
<td>13</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>ERM8</td>
<td>5B2D</td>
<td>Quantity 150 of #ESM8 3.72 TB SSD 4k SFF-2</td>
<td>13</td>
<td>Linux</td>
<td>Both</td>
</tr>
</tbody>
</table>
Table 1-8 lists options for USB-attached media available for the Power E980 server.

### Table 1-8  USB-attached external media

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>Max</th>
<th>OS support</th>
<th>Order type&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERM9</td>
<td>5B2D</td>
<td>Quantity 150 of #ESM9 3.72 TB SSD 4k SFF-2</td>
<td>13</td>
<td>IBM i</td>
<td>Both</td>
</tr>
<tr>
<td>1817</td>
<td>19B3</td>
<td>Quantity 150 of #1962</td>
<td>26</td>
<td>IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>1818</td>
<td>19B3</td>
<td>Quantity 150 of #1964</td>
<td>26</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>1927</td>
<td>19B1</td>
<td>Quantity 150 of #1964</td>
<td>26</td>
<td>IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>1929</td>
<td>19B1</td>
<td>Quantity 150 of #1953</td>
<td>26</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EQ82</td>
<td>5B1D</td>
<td>Quantity 150 of #ES62 3.86-4.0 TB 7200 rpm 4k LFF-1 Disk</td>
<td>13</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EQ64</td>
<td>5B1F</td>
<td>Quantity 150 of #ES64 7.72-8.0 TB 7200 rpm 4k LFF-1 Disk</td>
<td>13</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EQU2</td>
<td>59CD</td>
<td>Quantity 150 of #ESD2 (1.1 TB 10k SFF-2)</td>
<td>26</td>
<td>IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>EQU3</td>
<td>59CD</td>
<td>Quantity 150 of #ESD3 (1.2 TB 10k SFF-2)</td>
<td>26</td>
<td>AIX and Linux</td>
<td>Supported</td>
</tr>
<tr>
<td>EQDN</td>
<td>59CF</td>
<td>Quantity 150 of #ESDN (571 GB 15K RPM SAS SFF-2 for IBM i)</td>
<td>26</td>
<td>IBM i</td>
<td>Supported</td>
</tr>
<tr>
<td>EQD2</td>
<td>59CF</td>
<td>Quantity 150 of #ESDP (600 GB 15K RPM SAS SFF-2 for AIX/LINUX)</td>
<td>26</td>
<td>AIX and Linux</td>
<td>Supported</td>
</tr>
<tr>
<td>EUA5</td>
<td>63BD</td>
<td>Standalone USB DVD drive w/cable</td>
<td>1</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
</tbody>
</table>

<sup>a</sup> See 1.4.8, “Disk and media features” on page 18 for information on order types.
1.5 I/O drawers

If additional PCIe slots beyond the system node slots are required, the Power E980 supports the addition of I/O expansion drawers.

At initial availability in September 2018, the Power E980 supports the attachment of zero, one, or two PCIe Gen3 I/O expansion drawers to each system node.

In November 2018, zero, one, two, three, or four PCIe Gen3 I/O expansion drawers per system node will be supported. To connect an I/O expansion drawer, a PCIe slot is used to attach a 6-slot expansion module in the I/O Drawer. A PCIe Gen3 I/O expansion drawer (#EMX0) holds two expansion modules which are attached to any two PCIe slots in the same system node or in different system nodes.

For connection of SAS disks, disk-only I/O drawers are available. The EXP24S, EXP24SX, and EXP12SX are supported.

1.5.1 PCIe Gen3 I/O expansion drawer

The 19-inch 4 EIA (4U) PCIe Gen3 I/O expansion drawer (#EMX0) and two PCIe FanOut Modules (#EMXG) provide twelve PCIe I/O full-length, full-height slots. One FanOut 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 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 through HMC-guided menus or by operating system support utilities.

A PCIe CXP converter adapter and Active Optical cables (AOC) connect the system node to a PCIe FanOut module in the I/O expansion drawer. Each PCIe Gen3 I/O expansion drawer has two power supplies.

Drawers can be added to the server at a later time, but system downtime must be scheduled for adding a PCIe3 Optical Cable Adapter or a PCIe Gen3 I/O drawer (EMX0) or fan-out module.
Figure 1-11 shows a PCIe Gen3 I/O expansion drawer.

![Figure 1-11 PCIe Gen3 I/O expansion drawer](image)

1.5.2 I/O drawers and usable PCI slot

Figure 1-12 shows the rear view of the PCIe Gen3 I/O expansion drawer with the location codes for the PCIe adapter slots in the PCIe3 6-slot fanout module.

![Figure 1-12 Rear view of a PCIe Gen3 I/O expansion drawer with PCIe slots location codes](image)

Table 1-9 on page 25 provides details of the PCI slots in the PCIe Gen3 I/O expansion drawer.
Table 1-9  PCIe slot locations and descriptions for the PCIe Gen3 I/O expansion drawer

<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, generation 3, blind-swap cassettes.
- Slots C1 and C4 in each PCIe3 6-slot fanout 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 when expansion consists of a single drawer type.

Table 1-10  Maximum number of I/O drawers supported and total number of PCI slots

<table>
<thead>
<tr>
<th>System nodes</th>
<th>Maximum #EMX0 drawers</th>
<th>Total number of slots</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PCIe3, x16</td>
</tr>
<tr>
<td>1 system node</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>2 system nodes</td>
<td>8</td>
<td>32</td>
</tr>
<tr>
<td>3 system nodes</td>
<td>12</td>
<td>48</td>
</tr>
<tr>
<td>4 system nodes</td>
<td>16</td>
<td>64</td>
</tr>
</tbody>
</table>

**Availability:** At initial availability in September 2018 a maximum of two I/O drawers are supported per system node and a maximum of two system nodes is supported. In November 2018 four system nodes and four I/O drawers per node will be supported.

### 1.5.3 EXP24S SFF Gen2-bay Drawer

The EXP24S SFF Gen2-bay Drawer (#5887) is an expansion drawer with 24 2.5-inch form-factor (SFF) SAS bays. The EXP24S supports up to 24 hot-swap SFF-2 SAS hard disk
drives (HDDs) or solid state drives (SSDs). It uses 2 EIA of space in a 19-inch rack. The EXP24S includes redundant AC power supplies and uses two power cords.

With AIX, Linux, and VIOS, you can order the EXP24S with four sets of six bays, two sets of 12 bays, or one set of 24 bays (mode 4, 2, or 1). With IBM i, you can order the EXP24S as one set of 24 bays (mode 1). Mode setting is done by IBM Manufacturing. If you need to change the mode after installation, ask your IBM support representative to refer to:

http://w3-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PR55121

The EXP24S SAS ports are attached to a SAS PCIe adapter or pair of adapters using SAS YO or X cables.

Figure 1-13 shows the EXP24S SFF Gen2-bay drawer.

1.5.4 EXP24SX and EXP12SX SAS Storage Enclosures

If you need more disks than are available with the internal disk bays, you can attach more external disk subsystems, such as an EXP24SX SAS Storage Enclosure (FC #ESLS) or EXP12SX SAS Storage Enclosure (FC #ESLL).

The EXP24SX is a storage expansion enclosure with twenty-four 2.5-inch SFF SAS bays. It supports up to 24 hot-plug HDDs or SSDs in only 2 EIA of space in a 19-inch rack. The EXP24SX SFF bays use SFF Gen2 (SFF-2) carriers or trays.

The EXP12SX is a storage expansion enclosure with twelve 3.5-inch large form factor (LFF) SAS bays. It supports up to 12 hot-plug HDDs in only 2 EIA of space in a 19-inch rack. The EXP12SX SFF bays use LFF Gen1 (LFF-1) carriers/trays. The 4 Kb sector drives (FC #4096 or FC #4224) are supported. SSDs are not supported.

With AIX/Linux/VIOS, the EXP24SX or the EXP12SX can be ordered with four sets of six bays (mode 4), two sets of 12 bays (mode 2), or one set of 24-four bays (mode 1). IBM i supports the EXP24SX with one set of 24 bays (mode 1). It is possible to change the mode setting in the field by using software commands along with a specifically documented procedure.
**Important:** When changing modes, a skilled, technically qualified person should follow the special documented procedures. Improperly changing modes can potentially destroy existing RAID sets, prevent access to existing data, or allow other partitions to access another partition's existing data.

The attachment between the EXP24SX or EXP12SX and the PCIe3 SAS adapters or integrated SAS controllers is through SAS YO12 or X12 cables. All ends of the YO12 and X12 cables have mini-SAS HD narrow connectors. The PCIe Gen3 SAS adapters support 6Gb throughput. The EXP24SX and EXP12SX has been designed to support up to 12Gb throughput if future SAS adapters support that capability.

The EXP24SX or EXP12SX includes redundant AC power supplies and two power cords.

Figure 1-14 shows the EXP24SX drawer.

Figure 1-15 shows the EXP12SX drawer.
1.6 System racks

The Power E980 server is designed to fit a standard 19 inch rack. The server has been certified and tested in the IBM Enterprise racks (7014-T42, 7965-S42, #0551, or #0553). Clients can choose to place the server in other racks if they are confident those racks have the strength, rigidity, depth, and hole pattern characteristics that are needed. Clients should work with IBM Service to determine the appropriateness of other racks. The Power E980 rails can adjust their depth to fit a rack that is 59.06 - 77.3 cm (23.25 - 30.4375 inches) in depth.

**Order information:** It is highly recommended that the Power E980 server be ordered with an IBM 42U enterprise rack (7014-T42, 7965-S42, or #0553). This provides a more complete and higher quality environment for IBM Manufacturing system assembly and testing, and provide a more complete shipping package.

If a system will be installed in a rack or cabinet that is not from IBM, ensure that the rack meets the requirements that are described in 1.6.11, “Original equipment manufacturer racks” on page 39.

**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.
1.6.1 What’s new?

- The new IBM Enterprise 42U Slim Rack 7965-S42 (or feature code #ER05) is available offering 42 EIA units (U) of space in a slim footprint.
- The T00 rack is no longer available to purchase with a Power E980 server. Installing a Power E980 server in an existing T00 rack is still supported.

1.6.2 IBM 7014 Model T42 rack

The 2.0-meter (79.3-inch) Model T42 is compatible with past and present IBM Power Systems servers. The features of the T42 rack are as follows:

- Has 42U (EIA units) of usable space.
- Has optional removable side panels.
- Has optional side-to-side mounting hardware for joining multiple racks.
- Has increased power distribution and weight capacity.
- Supports AC power only.
- Up to four PDUs can be mounted in the PDU bays (see Figure 1-17 on page 33), but others can fit inside the rack. For more information, see 1.6.7, “The AC power distribution unit and rack content” on page 32.

Ruggedized Rack Feature

For enhanced rigidity and stability of the rack, the optional Ruggedized Rack Feature (FC #6080) provides more 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 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.

Weights are as follows:
- T42 base empty rack: 261 kg (575 lb)
- T42 full rack: 930 kg (2045 lb)

Vertical PDUs: All PDUs installed in a rack containing an E980 server must be installed horizontally to allow for cable routing in the sides of the rack.

Some of the available door options for the T42 rack are shown in Figure 1-16.
The door options are explained in the following list:

- The 2.0 m Rack Trim Kit (FC #6272) is used if no front door is used in the rack.
- The Front Door for a 2.0 m Rack (FC #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 some 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 (FC #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 FC #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 some visibility into the rack. The non-acoustic door has a depth of about 134 mm (5.3 in.).

### Rear Door Heat Exchanger

A Rear Door Heat Exchanger (FC #6858) is available to provide more efficient heat removal. This door replaces the standard rear door on the rack. Copper tubes that are attached to the rear door circulate chilled water, provided by the customer. 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 customer-supplied secondary water loop.
For more information about planning for the installation of the IBM Rear Door Heat Exchanger, see IBM Knowledge Center.

1.6.3 IBM Enterprise 42U Slim Rack 7965-S42

The 2.0-meter (79-inch) Model 7965-S42 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 that is 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.

Compared to the 7965-94Y Slim Rack, the Enterprise Slim Rack provides additional strength and additional shipping/installation flexibility.

The 7965-S42 rack has space for up to four PDUs in side pockets. Additional PDUs beyond four will be mounted horizontally and each consume 1U of rack space.

**Vertical PDUs:** All PDUs installed in a rack containing an E980 server must be installed horizontally to allow for cable routing in the sides of the rack.

The Enterprise Slim Rack front door, which can be Basic Black/Flat (#ECRM), High-End appearance (#ECRF), or OEM black (#ECRE), has perforated steel, providing ventilation, physical security, and visibility of indicator lights in the installed equipment within. It comes standard with a lock that is identical to the locks in the rear doors. The door (#ECRG and #ECRE only) can be hinged on either the left or right side.

**Orientation:** Feature ECRF should not be flipped because the IBM logo will be upside down.

1.6.4 Feature Code #0551 1.8 Meter Rack

The 1.8 Meter Rack (FC #0551) is a 36 EIA unit rack. The rack that is delivered as FC #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 FC #0551.

**Order availability:** The #0551 rack is only available when ordered as an MES order. It is not available as an initial order.
1.6.5 Feature Code #0553 2.0 Meter Rack

The 2.0 Meter Rack (FC #0553) is a 42 EIA unit rack. The rack that is delivered as FC #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 FC #0553.

Order availability: The #0551 rack is only available when ordered as an MES order. It is not available as an initial order.

1.6.6 Feature Code #ER05 rack

This feature provides a 19-inch, 2.0-meter high rack with 42 EIA units of total space for installing a rack-mounted central electronics complex 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 FC #ER05:

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

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

If ordering Power Systems equipment in an MES order, use the equivalent rack FC #ER05 instead of 7965-94Y so IBM Manufacturing can ship the hardware in the rack.

1.6.7 The AC power distribution unit and rack content

Power E980 systems that are integrated into a rack at the factory will have power distribution units (PDUs) mounted horizontally in the rack. Each PDU takes 1U of space in the rack. Mounting the PDUs vertically in the side of the rack could cause cable routing issues and interfere with optimal service access.

Two possible PDU ratings are supported: 60A/63A (orderable in most countries) and 30A/32A.

- The 60A/63A PDU supports four system node power supplies and one I/O expansion drawer or eight I/O expansion drawers.
- The 30A/32A PDU supports two system node power supplies and one I/O expansion drawer or four I/O expansion drawer

Rack-integrated system orders require at least two of either feature 7109, 7188, or 7196.

- Intelligent PDU with Universal UTG0247 Connector (#7109) is for an intelligent AC power distribution unit (PDU+) that enables users to monitor the amount of power being used by the devices that are plugged in to this PDU+. This AC power distribution unit provides 12 C13 power outlets. It receives power through a UTG0247 connector. It can be used for many different countries and applications by varying the PDU to Wall Power Cord, which must be ordered separately. Each PDU requires one PDU to Wall Power Cord. Supported power cords include the following features: 6489, 6491, 6492, 6653, 6654, 6655, 6656, 6657, and 6658.
- Power Distribution Unit (#7188) mounts in a 19-inch rack and provides twelve C13 power outlets. The PDU has six 16A circuit breakers, with two power outlets per circuit breaker. System units and expansion units must use a power cord with a C14 plug to connect to the feature 7188. One of the following line cords must be used to distribute power from a wall
outlet to the feature 7188: feature 6489, 6491, 6492, 6653, 6654, 6655, 6656, 6657, or 6658.

- The Three-phase Power Distribution Unit (#7196) provides six C19 power outlets and is rated up to 48A. It has a 4.3 m (14 ft) fixed power cord to attach to the power source (IEC309 60A plug (3P+G)). A separate "to-the-wall" power cord is not required or orderable. Use the Power Cord 2.8 m (9.2 ft), Drawer to Wall/IBM PDU, (250V/10A) (#6665) to connect devices to this PDU. These power cords are different than the ones used on the feature 7188 and 7109 PDUs. Supported countries for the feature 7196 PDU are Antigua and Barbuda, Aruba, Bahamas, Barbados, Belize, Bermuda, Bolivia, Brazil, Canada, Cayman Islands, Colombia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Guam, Guatemala, Haiti, Honduras, Indonesia, Jamaica, Japan, Mexico, Netherlands Antilles, Nicaragua, Panama, Peru, Puerto Rico, Surinam, Taiwan, Trinidad and Tobago, United States, and Venezuela.

Four PDUs can be mounted vertically in the back of the T00 and T42 racks. Figure 1-17 shows the placement of the four vertically mounted PDUs. In the rear of the rack, two more 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 uses 1U per PDU and reduces the space that is available for other racked components. When mounting PDUs horizontally, the preferred approach is to use fillers in the EIA units that are occupied by these PDUs to facilitate proper air-flow and ventilation in the rack.

![Figure 1-17  Power distribution unit placement and view](image-url)

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 1-11 shows the available wall power cord options for the PDU and intelligent power distribution unit (iPDU) features, which must be ordered separately.

<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>IEC 309, 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, 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, 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, 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, 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 appropriate 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 the drawers 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 four 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 IBM Knowledge Center.

For detailed power requirements and power cord details about the 7965-94Y rack, see IBM Knowledge Center.

### 1.6.8 PDU connection limits

Two possible PDU ratings are supported: 60/63 Amps and 30/32 Amps. The PDU rating is determined by the power cord used to connect the PDU to the electricity supply. The number of system nodes and I/O expansion drawers supported by each line cord is shown in Table 1-12 on page 35.

<table>
<thead>
<tr>
<th>Feature Code</th>
<th>Wall plug</th>
<th>PDU Rating</th>
<th>Maximum supported system nodes per PDU pair</th>
<th>Maximum supported I/O drawers with no system nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>6653</td>
<td>IEC 309, 3P+N+G, 16A</td>
<td>60 Amps</td>
<td>2 System nodes and 1 I/O expansion drawer</td>
<td>8</td>
</tr>
<tr>
<td>6489</td>
<td>IEC309 3P+N+G, 32A</td>
<td>60 Amps</td>
<td>2 System nodes and 1 I/O expansion drawer</td>
<td>8</td>
</tr>
<tr>
<td>6654</td>
<td>NEMA L6-30</td>
<td>30 Amps</td>
<td>1 System node and 1 I/O expansion drawer</td>
<td>4</td>
</tr>
<tr>
<td>6655</td>
<td>RS 3750DP (watertight)</td>
<td>30 Amps</td>
<td>1 System node and 1 I/O expansion drawer</td>
<td>4</td>
</tr>
<tr>
<td>6656</td>
<td>IEC 309, P+N+G, 32A</td>
<td>30 Amps</td>
<td>1 System node and 1 I/O expansion drawer</td>
<td>4</td>
</tr>
<tr>
<td>6657</td>
<td>PDL</td>
<td>30 Amps</td>
<td>1 System node and 1 I/O expansion drawer</td>
<td>4</td>
</tr>
<tr>
<td>6658</td>
<td>Korean plug</td>
<td>30 Amps</td>
<td>1 System node and 1 I/O expansion drawer</td>
<td>4</td>
</tr>
<tr>
<td>6492</td>
<td>IEC 309, 2P+G, 60A</td>
<td>60 Amps</td>
<td>2 System node and 1 I/O expansion drawer</td>
<td>4</td>
</tr>
<tr>
<td>6491</td>
<td>IEC 309, P+N+G, 63A</td>
<td>60 Amps</td>
<td>2 System nodes and 1 I/O expansion drawer</td>
<td>8</td>
</tr>
</tbody>
</table>
1.6.9 Rack-mounting rules

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

- The system can be placed at any location in the rack. For rack stability, start filling the rack from the bottom. See Ordering information below for details.
- Any remaining space in the rack can be used to install other systems or peripheral devices. Ensure 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.

Order information: The racking approach for the initial order must be either a 7014-T42 or 7965-S42. If an extra rack is required for I/O expansion drawers as an MES to an existing system, either an FC #0551, FC #0553, or FC #ER05 rack must be ordered.

If installing the E980 server in a T42 rack, 2U of space must be left for cable routing. For bottom cable exit, the 2U must be left at the bottom of the rack. For top cable exit, the 2U must be left at the top of the rack.

If installing the E980 server in an S42 rack, no space is required for cable routing.

1.6.10 Useful rack additions

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

IBM System Storage 7226 Model 1U3 Multi-Media Enclosure
The IBM System Storage 7226 Model 1U3 Multi-Media Enclosure can accommodate up to two tape drives, two RDX removable disk drive docking stations, or up to four DVD-RAM drives.

The IBM System Storage 7226 Multi-Media Enclosure supports LTO Ultrium and DAT160 Tape technology, DVD-RAM, and RDX removable storage requirements on the following IBM systems:

- IBM POWER6® processor-based systems
- IBM POWER7® processor-based systems
- IBM POWER8 processor-based systems
- IBM POWER9 processor-based systems

The IBM System Storage 7226 Multi-Media Enclosure offers an expansive list of drive feature options, as shown in Table 1-13.

<table>
<thead>
<tr>
<th>Feature Code</th>
<th>Description</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>5619</td>
<td>DAT160 SAS Tape Drive</td>
<td>Available</td>
</tr>
<tr>
<td>EU16</td>
<td>DAT160 USB Tape Drive</td>
<td>Available</td>
</tr>
<tr>
<td>1420</td>
<td>DVD-RAM SAS Optical Drive</td>
<td>Available</td>
</tr>
<tr>
<td>1422</td>
<td>DVD-RAM Slim SAS Optical Drive</td>
<td>Available</td>
</tr>
<tr>
<td>5762</td>
<td>DVD-RAM USB Optical Drive</td>
<td>Available</td>
</tr>
</tbody>
</table>
Here are the option descriptions:

- **DAT160 160 GB Tape Drives**: With SAS or USB interface options and a data transfer rate up to 12 MBps (assumes 2:1 compression), the DAT160 drive is read/write compatible with DAT160 and DDS4 data cartridges.

- **LTO Ultrium 5 Half-High 1.5 TB SAS and FC Tape Drive**: With a data transfer rate up to 280 MBps (assuming a 2:1 compression), the LTO Ultrium 5 drive is read/write compatible with LTO Ultrium 5 and 4 data cartridges, and read-only compatible with Ultrium 3 data cartridges. By using data compression, an LTO-5 cartridge can store up to 3 TB of data.

- **LTO Ultrium 6 Half-High 2.5 TB SAS and FC Tape Drive**: With a data transfer rate up to 320 MBps (assuming a 2.5:1 compression), the LTO Ultrium 6 drive is read/write compatible with LTO Ultrium 6 and 5 media, and read-only compatibility with LTO Ultrium 4. By using data compression, an LTO-6 cartridge can store up to 6.25 TB of data.

- **DVD-RAM**: The 9.4 GB SAS Slim Optical Drive with an SAS and USB interface option is compatible with most standard DVD disks.

- **RDX removable disk drives**: The RDX USB docking station is compatible with most RDX removable disk drive cartridges when it is used in the same operating system. The 7226 offers the following RDX removable drive capacity options:
  - 500 GB (FC #1107)
  - 1.0 TB (FC #EU01)
  - 2.0 TB (FC #EU2T)

Removable RDX drives are in a rugged cartridge that inserts into an RDX removable (USB) disk docking station (FC #1103 or FC #EU03). RDX drives are compatible with docking stations, which are installed internally in IBM POWER6, IBM POWER6+, POWER7, IBM POWER7+, POWER8, and POWER9 processor-based servers, where applicable.

Media that is used in the 7226 DAT160 SAS and USB tape drive features are compatible with DAT160 tape drives that are installed internally in IBM POWER6, POWER6+, POWER7, POWER7+, POWER8, and POWER9 processor-based servers.
Media that is used in LTO Ultrium 5 Half-High 1.5 TB tape drives are compatible with Half High LTO5 tape drives that are installed in the IBM TS2250 and TS2350 external tape drives, IBM LTO5 tape libraries, and half-high LTO5 tape drives that are installed internally in IBM POWER6, POWER6+, POWER7, POWER7+, POWER8, and POWER9 processor-based servers.

Figure 1-18 shows the IBM System Storage 7226 Multi-Media Enclosure.

Figure 1-18  IBM System Storage 7226 Multi-Media Enclosure

The IBM System Storage 7226 Multi-Media Enclosure offers customer-replaceable unit (CRU) maintenance service to help make the installation or replacement of new drives efficient. Other 7226 components are also designed for CRU maintenance.

The IBM System Storage 7226 Multi-Media Enclosure is compatible with most IBM POWER6, POWER6+, POWER7, POWER7+, POWER8, and POWER9 processor-based systems that offer current level AIX, IBM i, and Linux operating systems.

Unsupported: IBM i does not support 7226 USB devices.

For a complete list of host software versions and release levels that support the IBM System Storage 7226 Multi-Media Enclosure, see System Storage Interoperation Center (SSIC).

Flat panel display options
The IBM 7316 Model TF4 is a rack-mountable flat panel console kit that can also be configured with the tray pulled forward and the monitor folded up, providing full viewing and keying capability for the HMC operator.

The Model TF4 is a follow-on product to the Model TF3 and offers the following features:

- A slim, sleek, and lightweight monitor design that occupies only 1U (1.75 in.) in a 19-inch standard rack.
- A 18.5-inch (409.8 mm x 230.4 mm) flat panel TFT monitor with truly accurate images and virtually no distortion.
- The ability to mount the IBM Travel Keyboard in the 7316-TF4 rack keyboard tray.
Support for the IBM 1x8 Rack Console Switch (FC #4283) IBM Keyboard/Video/Mouse (KVM) switches.

FC #4283 is a 1x8 Console Switch that fits in the 1U space behind the TF4. It is a CAT5 based switch containing eight rack interface (ARI) ports for connecting either PS/2 or USB console switch cables. It supports chaining of servers that use IBM Conversion Options switch cable FC #4269. This feature provides four cables that connect a KVM switch to a system, or can be used in a daisy-chain scenario to connect up to 128 systems to a single KVM switch. It also supports server-side USB attachments.

1.6.11 Original equipment manufacturer racks

The system can be installed in a suitable original equipment manufacturer (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 IBM Knowledge Center.

The web site 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 (that is, the horizontal width between the vertical columns of holes on the two front-mounting flanges and on the two rear-mounting flanges). Figure 1-19 is a top view showing the specification dimensions.

![Figure 1-19](image)

* Notes:
  - Rails to support tapped hole racks (M5, M6, 10-32, 12-24) are currently not offered for this product.
  - In addition to the dimensions to the left, the following EIA hole shapes are currently supported:
    1. 7.1 mm +/- 0.1 mm round holes (.280" +/- .004")
    2. 9.5 mm +/- 0.1 mm square holes (.375" +/- .004")

*Dimension comes from industry standard EIA-310D specification.
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, making 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 1-20 shows the top front specification dimensions.

1.7 Hardware Management Console

This section discusses Hardware Management Consoles (HMCs) available for use with Power servers.

1.7.1 What’s new?

- New HMCs are now based on systems with Power processors
- Intel x86-based HMCs are supported but no longer available
- Virtual HMCs (vHMC) are available for x86 and Power virtual environments

1.7.2 Hardware Management Console overview

Administrators can use the Hardware Management Console (HMC), which is a dedicated appliance, to configure and manage system resources on IBM Power Systems servers. GUI, command-line interface (CLI), or REST API interfaces are available. The HMC provides basic virtualization management support for configuring logical partitions (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™ (ESA) for faster support.

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

The HMC is available as a hardware appliance or as a virtual appliance (vHMC). The Power E980 servers support several service environments:

- Attachment to one or more HMCs or vHMCs. This is the default configuration for servers supporting multiple logical partitions with dedicated resource or virtual I/O.
  - X86 based HMCs - 7042-CR7, CR8, or CR9
  - POWER® based HMC - 7063-CR1
  - vHMC on x86 or POWER systems LPAR

Hardware support for customer replaceable units (CRUs) comes as standard with the HMC. In addition, users can upgrade this support level to IBM on-site support to be consistent with other Power Systems servers.

**Note:**

- An HMC or vHMC is required for the Power E980.
- Integrated Virtual Management (IVM) is no longer supported.

For more information about vHMC, see Virtual HMC Appliance (vHMC) Overview.

Figure 1-21 shows HMC model selections and tier updates.

![HMC selections](image)

Multiple Power Systems servers can be managed by a single HMC. Each server can be connected to multiple HMC consoles to build extra resiliency into the management platform.

Figure 1-22 shows several examples of HMC configurations.
1.7.3 Hardware Management Console code level

The HMC code must be running at Version 9 Release 2 (V9R2) or later when you use the HMC with the E980 server.

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 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 might verify entitlement through customer number, serial number electronic restrictions, or any other means or methods that are employed by IBM at its discretion.

► HMC V9 supports only the Enhanced+ version of the GUI. The Classic version is no longer available.

► HMC V9R1.911.0 release added support for managing IBM OpenPOWER systems. The same HMC used to manage FSP-based enterprise systems can now also manage the BMC-based AC/LC servers. This provides a consistent and consolidated hardware management solution.

► HMC V9 supports connections to servers that are based on IBM servers that are based on POWER9, POWER8, and POWER7 processors. There is no support in this release for servers that are based on POWER6 processors or earlier.
1.7.4 Two architectures of Hardware Management Console

There are now two options for the HMC hardware: The earlier Intel-based HMCs, and the newer HMCs that are based on an IBM POWER8 processor. x86-based HMCs are no longer available to order but are supported as an option for managing the Power E980.

You may use either architecture to manage the servers. You also may use one Intel-based HMC and one POWER8 based HMC if the software is at the same level.

IBM’s recommendation is to use the new POWER8 processor-based consoles for server management.

Intel-based HMCs

HMCs based on Intel processors that support V9 code are:

- 7042-CR9
- 7042-CR8
- 7042-CR7

7042-CR6 and earlier HMCs are not supported for use with the Power E980.

The 7042-CR9 has the following specifications:

- 2.4 GHz Intel Xeon Processor E5-2620 V3
- 16 GB (1 x 16 GB) of 2.133 GHz DDR4 system memory
- 500 GB SATA SFF HDD
- SATA CD/RW - DVD-RAM
- Four Ethernet ports
- Six USB ports (two front and four rear)
- One PCIe slot

POWER8 processor-based HMC

The POWER processor-based HMC has machine type and model 7063-CR1. It has the following specifications:

- 1U base configuration
- IBM POWER8 120 W 6-core CPU
- 32 GB (4 x 8 GB) of DDR4 system memory
- 2 x 2 TB SATA LFF 3.5-inch HDD RAID 1
- Rail bracket option for round hole rack mounts
- Two USB 3.0 hub ports in the front of the server
- Two USB 3.0 hub ports in the rear of the server
- Redundant 1 kW power supplies
- 4 x 10 Gb Ethernet Ports (RJ-45) (10 Gb/1 Gb/100 Mb)
- 1 x 1 Gb Ethernet Port for Management (baseboard management controller)

All future HMC development will be done for the POWER8 processor-based 7063-CR1 and its successors.

Note: System administrators can remotely start or stop a 7063-CR1 HMC using the ipmitool command or Web UI.
1.7.5 Connectivity to POWER9 processor-based systems

POWER9 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. Additionally, to perform an operation on an LPAR, initiate Live Partition Mobility (LPM), or do PowerVM Active Memory Sharing operations, an Ethernet link to the required on the managed partitions is required. A minimum of two Ethernet ports are needed on the HMC to provide such connectivity.
For the HMC to communicate properly 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 may 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.

Figure 1-23 shows a simple network configuration to enable the connection from the HMC to the server and to allow for dynamic LPAR operations. For more information about HMC and the possible network connections, see IBM Power Systems HMC Implementation and Usage Guide, SG24-7491.

By default, the service processor HMC ports are configured for dynamic IP address allocation. The HMC can be configured as a 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 is 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 by using the Advanced System Management Interface (ASMI) menus.

Notes: The two service processor HMC ports have the following features:
- 1 Gbps connection speed.
- 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
1.7.6 High availability Hardware Management Console 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, such as dynamic LPAR, partition migration that uses PowerVM LPM, or the creation of a partition, cannot be performed without the HMC. Power Systems servers support the capability to have two HMCs attached to a system. This provides redundancy in case one of the HMCs is unavailable.

To achieve HMC redundancy for a POWER9 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 1-24 shows one possible highly available HMC configuration that manages two servers. Each HMC is connected to one FSP port of each managed server.

For simplicity, only the hardware management networks (LAN1 and LAN2) are highly available. 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 IBM Power Systems HMC Implementation and Usage Guide, SG24-7491.
Architecture and technical overview

This chapter describes the overall system architecture for the IBM Power System E980 (9080-M9S). The bandwidths that are provided throughout the section 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 do the performance sizing at the application workload environment level and evaluate performance by using real-world performance measurements and production workloads.

Figure 2-1 on page 48 shows the logical system architecture of the Power E980.
Figure 2-1  Power E980 logical system diagram

Figure 2-2 on page 49 shows the SMP connections between nodes for 2-, 3-, and 4-drawer configurations.
Figure 2-2  SMP cabling
2.1 The IBM POWER9 processor

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

2.1.1 POWER9 processor overview

The POWER9 processors are single-chip modules (SCMs) manufactured on the IBM 14-nm FinFET Silicon-On-Insulator (SOI) architecture. Each module is 68.5 mm x 68.5 mm and contains 8 billion transistors.

As shown in Figure 2-3, the chip contains 12 cores, two memory controllers, Peripheral Component Interconnect Express (PCIe) Gen4 I/O controllers, and an interconnection system that connects all components within the chip at 7 TBps. Each core has 512 KB of level 2 cache, and 10 MB of level 3 embedded DRAM (eDRAM) cache. The interconnect also extends through module and system board technology to other POWER9 processors in addition to memory and various I/O devices.

![Figure 2-3 The POWER9 processor chip](image)

The Power E980 uses memory buffer chips to interface between the POWER9 processor and the DDR3 or DDR4 memory. Each buffer chip also includes an L4 cache to reduce the latency of local memory accesses.
The POWER9 chip provides an embedded algorithm for the following features:

- External Interrupt Virtualization Engine. Reduces the code impact/path length and improves performance compared to the previous architecture.
- Gzip compression and decompression
- PCIe Gen4 support
- Two memory controllers that support buffered connection to DDR3 or DDR4 memory
- Cryptography: Advanced encryption standard (AES) engine
- Random number generator (RNG)
- Secure Hash Algorithm (SHA) engine: SHA-1, SHA-256, and SHA-512, and Message Digest 5 (MD5)
- IBM Data Mover Tool

Table 2-1 provides a summary of the POWER9 processor technology.

**Note:** The total values represent the maximum of 12 cores for the POWER9 architecture. The Power E980 has options for 32, 40, 44, and 48 cores per node.

### Table 2-1  Summary of the POWER9 processor technology

<table>
<thead>
<tr>
<th>Technology</th>
<th>POWER9 processor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die size</td>
<td>68.5 mm × 68.5 mm</td>
</tr>
<tr>
<td>Fabrication technology</td>
<td></td>
</tr>
<tr>
<td></td>
<td>▶ 14-nm lithography</td>
</tr>
<tr>
<td></td>
<td>▶ Copper interconnect</td>
</tr>
<tr>
<td></td>
<td>▶ SOI</td>
</tr>
<tr>
<td></td>
<td>▶ eDRAM</td>
</tr>
<tr>
<td>Maximum processor cores</td>
<td>12</td>
</tr>
<tr>
<td>Maximum execution threads core/module</td>
<td>8/96</td>
</tr>
<tr>
<td>Maximum L2 cache core/module</td>
<td>512 KB/6 MB</td>
</tr>
<tr>
<td>Maximum On-chip L3 cache core/module</td>
<td>10 MB/120 MB</td>
</tr>
<tr>
<td>Number of transistors</td>
<td>8 billion</td>
</tr>
<tr>
<td>Compatibility</td>
<td>With prior generation of POWER processor</td>
</tr>
</tbody>
</table>

### 2.1.2 POWER9 processor core

The POWER9 processor core is a 64-bit implementation of the IBM Power Instruction Set Architecture (ISA) Version 3.0, and has the following features:

- Multi-threaded design, which is capable of up to eight-way simultaneous multithreading (SMT)
- 64 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 that uses 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 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 16 floating point operations per cycle (eight double precision or 16 single precision)
- In-core AES encryption capability
- Hardware data prefetching with 16 independent data streams and software control
- Hardware decimal floating point (DFP) capability

More information about Power ISA Version 3.0, see OpenPOWER: IBM Power ISA Version 3.0B.

Figure 2-4 on page 52 shows a picture of the POWER9 core, with some of the functional units highlighted.

2.1.3 Simultaneous multithreading

POWER9 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 enables a single physical processor core to simultaneously dispatch instructions from more than one hardware thread context. With SMT, each POWER9 core can present eight hardware threads. Because there are multiple hardware threads per physical processor core, more 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.
2.1.4 POWER9 compatibility modes

The POWER9 processor has the ability to run in compatibility modes for previous POWER processor generations. This enables older operating systems to run on POWER9 systems. Compatibility modes also allows for live partition migration from systems based on previous generations of POWER processors. The POWER9 processor can run in the following compatibility modes:

- POWER7
- POWER8
- POWER9 Base

2.1.5 Processor feature codes

Each system enclosure in a Power E980 server has four sockets for processor modules. All sockets must be populated with a matching processor module. In servers with multiple system enclosures, all sockets on all enclosures must be populated with matching processor modules.

Table 2-3 shows the processor features available for the Power E980 server.

<table>
<thead>
<tr>
<th>Feature Code</th>
<th>CCIN</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EFB1</td>
<td>5C35</td>
<td>CBU for Power Enterprise Systems 3.20 GHz, 32-core POWER9 processor</td>
</tr>
<tr>
<td>EFB2</td>
<td>5C36</td>
<td>CBU for Power Enterprise Systems 2.80 GHz, 40-core POWER9 processor</td>
</tr>
<tr>
<td>EFB3</td>
<td>5C39</td>
<td>CBU for Power Enterprise Systems 2.90 GHz, 48-core POWER9 processor</td>
</tr>
<tr>
<td>EFB4</td>
<td>5C46</td>
<td>CBU for Power Enterprise Systems 3.0 GHz, 44-core POWER9 processor</td>
</tr>
<tr>
<td>EFP1</td>
<td>5C35</td>
<td>3.40 GHz 32-core (4x8) POWER9 processor</td>
</tr>
<tr>
<td>EFP2</td>
<td>5C36</td>
<td>3.15 GHz 40-core (4x10) POWER9 processor</td>
</tr>
<tr>
<td>EFP3</td>
<td>5C39</td>
<td>2.90 GHz 48-core (4x12) POWER9 processor</td>
</tr>
<tr>
<td>EFP4</td>
<td>5C46</td>
<td>3.0 GHz 44-core (4x11) POWER9 processor</td>
</tr>
</tbody>
</table>
Processors in the Power E980 system support Capacity on Demand. Further details on Capacity on Demand can be found in Chapter 2.3, “Capacity on Demand” on page 76.

Processor-specific Capacity on Demand features can be found in Table 2-4.

<table>
<thead>
<tr>
<th>Processor feature</th>
<th>Static activation feature</th>
<th>Mobile enabled activation feature</th>
<th>Static activation for Linux feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>#EFP1 3.40 GHz 32-core</td>
<td>#EFPA</td>
<td>#EFPE</td>
<td>#ELBK</td>
</tr>
<tr>
<td>#EFP2 3.15 GHz 40-core</td>
<td>#EFPB</td>
<td>#EFPF</td>
<td>#ELBL</td>
</tr>
<tr>
<td>#EFP3 2.90 GHz 48-core</td>
<td>#EFPC</td>
<td>#EFPG</td>
<td>#ELBM</td>
</tr>
<tr>
<td>#EFP4 3.0 GHz 44-core</td>
<td>#EFP9</td>
<td>#EFPN</td>
<td>#ELBQ</td>
</tr>
</tbody>
</table>

Capacity on Demand features that are independent of the processor feature are found Table 2-5.

<table>
<thead>
<tr>
<th>Feature code</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF2R</td>
<td>Single 5250 Enterprise Enablement</td>
</tr>
<tr>
<td>EF30</td>
<td>Full 5250 Enterprise Enablement</td>
</tr>
<tr>
<td>EFPD</td>
<td>Mobile processor activation for M9S/80H</td>
</tr>
<tr>
<td>EFPH</td>
<td>Mobile processor activation for M9S/80H (Upgrade from P7)</td>
</tr>
<tr>
<td>EP2W</td>
<td>Mobile Processor activation M9S/80H (Upgrade from P8)</td>
</tr>
<tr>
<td>EP9T</td>
<td>90 Days Elastic CoD Processor Core Enablement</td>
</tr>
<tr>
<td>MMC1</td>
<td>Elastic CoD Processor Day for IBM i</td>
</tr>
<tr>
<td>MMCX</td>
<td>Elastic CoD Processor Day for AIX and Linux</td>
</tr>
</tbody>
</table>

2.1.6 Memory access

One POWER9 processor module of a Power E980 high-end system provides two integrated memory controllers to facilitate access to the main memory of the system. One memory controller drives four differential memory interface (DMI) channels with a maximum signaling rate of 9.6 GHz. This yields a peak memory bandwidth of up to 28.8 GBps per memory channel or 230.4 GBps per processor module. Every DMI channel connects to one dedicated memory buffer chip. Each memory buffer chip in turn provides four DDR4 memory ports running at 1,600 MHz signal rate and one 16 MB L4 cache. A single memory buffer chip is mounted with the associated DRAM chips on one circuit board which is referred to as custom DIMM (CDIMM) module. On a new system order the DDR4 technology based CDIMMs are available with 32 GB, 64 GB, 128 GB, 256 GB, and 512 GB capacity. Also, the POWER9 memory channels support the same electrical signaling, transport layer characteristics and high level, agnostic read/write protocol as the POWER8 counterparts on Power E870, Power E870C, Power E880, and Power E880C servers. This enables the option to reuse...
DDR3 and DDR4 CDIMMs when transferred as part of a model upgrade from the named POWER8 high-end servers to a Power E980 system.

The maximum supported memory capacity per processor module is 4 TB which requires the use of 512 GB CDIMMs in all 8 available CDIMM slots of a module. A maximum of 16 TB of main memory can be provided by the four processor modules available in one system node. A 4-node Power E980 system makes up to 64 TB of system memory accessible to configured logical partitions. Figure 2-5 depicts schematically the POWER9 hierarchical memory subsystem of a Power E980 system.

For more detailed information about the CDIMM technology, memory placement rules, memory bandwidth and other topics related to the memory subsystem of the Power E980 server please refer to Chapter 2.2, “Memory subsystem” on page 66.

### 2.1.7 On-chip L3 cache innovation and intelligent caching

Similar to POWER8, the POWER9 processor uses a breakthrough in material engineering and microprocessor fabrication to implement the L3 cache in embedded Dynamic Random-Access Memory (eDRAM) technology and place it on the processor die. The L3 cache is critical to a balanced design, as is the ability to provide good signaling between the L3 cache and other elements of the hierarchy, such as the L2 cache or SMP interconnect.

Like POWER8 the POWER9 processor supports the same L3 non-uniform cache access (NUCA) architecture which provides mechanisms to distribute and share cache footprints across the chip. The on-chip L3 cache is organized into separate areas with differing latency characteristics. Each processor core is associated with a fast 10 MB local region of L3 cache (FLR-L3), but also has access to other L3 cache regions as a shared L3 cache. Additionally, each core can negotiate to use the FLR-L3 cache that is associated with another core, depending on the reference patterns. Data can also be cloned and stored in more than one core’s FLR-L3 cache, again depending on the reference patterns. This intelligent cache management enables the POWER9 processor to optimize the access to L3 cache lines and minimize overall cache latencies. In reference to the POWER8 L3 implementation the POWER9 L3 introduces an enhanced replacement algorithm with data-type and re-use awareness which utilizes information from the core and L2 cache to manage cache.
replacement states. The L3 cache supports an array of prefetch requests from the core including both instruction and data and for different levels of urgency. Prefetch requests for POWER9 include additional information exchange between the core, cache and the memory controller to manage memory bandwidth and to mitigate the prefetch based cache pollution.

The listing below provides an overview of the features offered by the POWER9 L3 cache:

- Private 10-MB L3.0 cache/shared L3.1.
  - Victim cache for local L2 cache (L3.0).
  - Victim cache for other on-chip L3 caches (L3.1).
- 20-way set associative.
- 128-byte cache lines with 64-byte sector support.
- 10 EDRAM banks (interleaved for access overlapping).
- 64-byte wide data bus to L2 for reads.
- 64-byte wide data bus from L2 for L2 castouts.
- Eighty 1 Mb EDRAM macros that are configured in 10 banks, with each bank having a 64-byte wide data bus.
- All cache accesses have the same latency.
- 20-way directory that is organized as four banks, with up to four reads or two reads and two writes every two processor clock cycles to differing banks.

The L3 cache architecture of the 12-core POWER9 processor is identical to the 24-core POWER9 implementation and you may refer to the POWER9 Processor User’s Manual as published by the OpenPOWER Foundation for more details about the L3 cache technology:

https://openpowerfoundation.org/?resource_lib=power9-processor-users-manual


https://www.research.ibm.com/journal/

### 2.1.8 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 like a single operation. Transactional memory is like database transactions where all shared memory accesses and their effects are either committed 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 retrying execution if isolation is violated. Generally, transactional memory enables 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 POWER9 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 enabling the system to operate in maximum performance.

2.1.9 POWER9 accelerator processor interfaces

The POWER9 processor supports three dedicated interfaces and protocols designed to attach advanced accelerator and future memory technologies:

- Coherent Accelerator Processor Interface (CAPI)
- Open Coherent Accelerator Processor Interface (OpenCAPI)
- NVIDIA NVLink

The CAPI protocol utilizes the PCIe Gen4 bus which is natively supported on the POWER9 processor die. CAPI capable accelerators are implemented as adapter cards which are placed in a CAPI enabled PCIe Gen3 or Gen4 slot. To that extend the maximum bandwidth of CAPI accelerator is limited by the PCIe bandwidth which is 64 GBps for a x16 PCIe Gen4 adapter slot in a POWER9 processor-based system. All eight x16 PCIe Gen4 slot in a Power E980 system node are enabled for CAPI support which yields a maximum of 32 CAPI attached accelerators per 4-node system. The CAPI protocol is developed and standardised since 2013 by the OpenPOWER foundation. More information about the CAPI protocol can be found at the following URL:

https://openpowerfoundation.org/

On Power E980 systems OpenCAPI attached accelerators and devices and NVLink graphics processing unit (GPU) connections are supported by two buses per POWER9 processor. They are using the same interconnect technology which facilitates the SMP communication between Power E980 nodes and provide a combined transfer capacity of 48 lanes running with a signaling rate of 25.78 Gbps. Each system node has four interconnect busses referred to as O-buses O0, O1, O2, and O3 which are designed to support the SMP, OpenCAPI, or NVLink protocol. On Power E980 system nodes the buses are configured to support the named protocols as follows:

- O0 – SMP or NVLink or OpenCAPI
- O1 – SMP only
- O2 – SMP only
- O3 – SMP or NVLink or OpenCAPI

The OpenCAPI technology is developed and standardized by the OpenCAPI consortium and in depth information about the consortium's mission and the OpenCAPI protocol specification can be found at the following URL:

https://opencapi.org/

The NVLink 2.0 protocol is natively supported by dedicated logic on the POWER9 processor die and allows to coherently attach NVIDIA graphics processing units (GPUs) through a maximum of two O-buses per processor. Each NVLink O-bus is comprised of two bricks and each brick provides 8 data lanes in NVLink mode running at 25 Gbps signaling rate. The maximum bandwidth of one O-bus used to attach a NVLink GPU is 103.12 GBps:

- 2 bricks · 8 lanes · 25.78 Gbps · 2 full duplex = 103.12 GBps

The NVLink technology is developed by the NVIDIA corporation and more information about the NVLink protocol is available at the following URL:

Coherent Accelerator Processor Interface 2.0

The IBM coherent accelerator processor interface 1.0 (CAPI 1.0) along with the related coherent accelerator interface architecture (CAIA) was introduced with POWER8 in 2014. CAPI 1.0 allows to attach special processing devices or accelerators to the POWER8 processor bus through the native PCIe Gen3 interface. The attached devices exhibit the form factor of a standard PCIe adapter card and typically use field programmable gate arrays (FPGA) or application specific integrated circuits (ASICs) to provide a specific function with significantly enhanced performance. In the CAPI paradigm, the specific algorithm for acceleration is thus contained in a unit on the FPGA or ASIC and is called the accelerator function unit (AFU). One of the key benefits of CAPI is that the devices gain coherent shared memory access with the processors in the server and share full virtual address translation with these processors by using the standard PCIe interconnect logic. In CAPI 1.0 the address translation logic of the attached devices or accelerators is implemented as POWER service layer (PSL) on the given FPGA or ASIC. To ensure cache coherency the PSL exchanges the relevant address information with the coherent accelerator processor proxy (CAPP) unit which resides on the POWER8 processor chip. Applications can have customized functions in FPGAs or ASICs and queue work requests directly into shared memory queues to the accelerator logic. Applications can also have customized functions by using the same effective addresses they use for any threads running on a host processor. From a practical perspective, CAPI enables a specialized hardware accelerator to be seen as an dedicated processor (hollow core) in the system with access to the main system memory and coherent communication with other processors in the system.

CAPI 2.0 was introduced with the POWER9 processor technology and represents the next step in the evolutionary development to enhance the architecture and the protocol for the attachment of accelerators. CAPI 2.0 utilizes the standard PCIe Gen4 interface of the POWER9 processor which provides twice the bandwidth compared to the previous PCIe Gen3 interconnect generation.

A key difference between CAPI 1.0 and CAPI 2.0 relies on the introduction of the Nest Memory Management Unit (NMMU) with POWER9. The NMMU replaces the address translation and page fault logic residing inside the PSL. CAPI 2.0 on POWER9 still requires a PSL in order to control the PCIe bus, to provide the memory mapped IO (MMIO) support, and to generate AFU specific interrupts. However, by taking the address translation and page fault logic out of the POWER9 PSL it has become a “lighter” version of the POWER8 PSL which potentially reduces the complexity of accelerator development. Figure 2-6 shows a block diagram of the CAPI 2.0 POWER9 hardware architecture.

*Note:* The Power E980 system is designed to support the OpenCAPI and the NVLink protocol but at the time of writing neither OpenCAPI attached accelerator or devices nor NVLink attached GPUs are supported. The information pertaining to this technologies are included for future reference only.
The POWER9 processor modules of the Power E980 server provide fault handling for the POWER service layer.

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 reduction in the code path length might occur because the adapter is not operating as a traditional I/O device, and there is no device driver layer to perform processing. CAPI also presents a simpler programming model.
Figure 2-7 shows a comparison of the traditional model, where the accelerator must go through the processor to access memory with CAPI.

As mentioned before with CAPI 1.0 on POWER8 the PSL on the accelerator adapter provides address translation and system memory cache for the accelerator functions. The custom processors on the adapter 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. It also eliminates the need for direct memory access that requires system calls to the OS kernel. By removing these layers, the data transfer operation requires fewer clock cycles in the processor, improving the I/O performance.

With CAPI 2.0 on POWER9 the address translation and page fault logic has been moved to the native NMMU on the POWER9 processor module but since the accelerator has direct access to this functional unit the benefit of reduced path length in the programming model stays the same while the cache coherency control of the unified address space is significantly enhanced.

The implementation of CAPI on the POWER9 processor enables hardware companies to develop solutions for specific application demands. Companies use the performance of the POWER9 processor for general applications and the custom acceleration of specific functions by using a hardware accelerator with a simplified programming model and efficient communication with the processor and memory resources.

Open Coherent Accelerator Processor Interface
In October 2016 the companies AMD, Google, IBM, Mellanox Technologies and Micron formed the OpenCAPI not-for-profit organization to create an open coherent high
performance bus interface based on a new bus standard called Open Coherent Accelerator Processor Interface (OpenCAPI) and to grow the ecosystem that utilizes this interface. This initiative is being driven by the emerging accelerated computing and advanced memory/storage solutions that have introduced significant system bottlenecks in today’s current open bus protocols and requires a technical solution that is openly available.

Two major technology trends heavily impact the industry currently:

- Hardware acceleration will become commonplace as microprocessor technology and design continues to deliver far less than the historical rate of cost/performance improvement per generation.
- New advanced memory technologies will change the economics of computing.

Existing system interfaces are insufficient to address these disruptive forces. Traditional I/O architecture results in very high processor overhead when applications communicate with I/O or accelerator devices at the necessary performance levels. Also they are not able to integrate multiple memory technologies with different access methods and performance attributes.

These challenges are addressed by the OpenCAPI architecture in a way that allows full industry participation. Embracing an open architecture is fundamental to establish sufficient volume base to drive cost down and to ensure the support of a broad ecosystem of software products and attached devices.

OpenCAPI is an open interface architecture that allows any microprocessor to attach to:

- Coherent user-level accelerators and I/O devices
- Advanced memories accessible via read/write or user-level direct memory access (DMA) semantics

OpenCAPI is agnostic to processor architecture and exhibits the following key attributes:

- High-bandwidth, low latency interface optimized to enable streamlined implementations of attached devices
- 25Gbit/sec signaling and protocol built to enable very low latency interface on processors and attached devices
- Complexities of coherence and virtual addressing implemented on host microprocessor to simplify attached devices and facilitate interoperability across multiple CPU architectures
- Attached devices operate natively within an application’s user space and coherently with processors allowing attached devices to fully participate in applications without kernel involvement/overhead
- Support of a wide range of use cases and access semantics:
  - Hardware accelerators
  - High-performance I/O devices
  - Advanced memories

### 2.1.10 Power and performance management

POWER9 processor-based scale-out and scale-up servers implement Workload Optimized Frequency (WOF) as a new feature of the power management EnergyScale technology. With POWER9 EnergyScale the POWER8 dynamic power saver modes which either favor lower power consumption (DPS) or favor performance (DPS-FP) are replaced by two new power saver modes:

- Dynamic performance mode (DPM)
- Maximum performance mode (MPM)
Every POWER9 processor-based scale-out or scale-up system will have either the DPM or the MPM enabled by default. Both modes dynamically adjust processor frequency to maximize performance and enable a much higher processor frequency range in comparison to POWER8 servers. Each of the new power saver modes will deliver consistent system performance without any variation as long as the nominal operating environment limits are met. It also should be noted, that for POWER9 processor-based systems which are under control of the PowerVM hypervisor the DPM and MPM are a system wide configuration setting but each processor module frequency is optimized separately.

Several factors determine the maximum frequency a given processor module can run at:

- Processor utilization:
  Lighter workloads will run at higher frequencies.
- Number of active cores:
  Less number of active cores will run at higher frequencies.
- Environmental conditions:
  At lower ambient temperatures cores will be enabled to run at higher frequencies.

The new power saver modes are defined as follows:

**Dynamic performance mode (DPM)**
In DPM the workload is executed at the highest frequency possible provided the nominal power consumption limit of the processor modules are honored. The frequency of the processor modules is always at the nominal frequency of the POWER9 processor-based system or above the nominal frequency up to the upper limit of the DPM frequency range. This DPM typical frequency range (DTFR) is published as part of the system specifications of a particular POWER9 system if it is running by default in the dynamic performance mode. The system performance is deterministic within the allowed operating environmental limits and as such does not depend on the ambient temperature as long as the temperature stays within the supported range. The idle power saver (IPS) function can be enabled or disabled. If IPS is enabled and all cores in a processor module are idle for 100s of milliseconds the frequency of the cores in the respective module will drop to the pre-defined power save frequency.

**Maximum performance mode (MPM)**
In MPM the workload is executed at the highest frequency possible but unlike in the DPM the processor module is allowed to operate at a higher power consumption level. The higher power draw enables the processor modules to run in a MPM typical frequency range (MTFR) where the lower limit is well above the nominal frequency and the upper limit is given by the system’s maximum frequency. The MTFR is published as part of the system specifications of a particular POWER9 system if it is running by default in MPM. The higher power draw potentially increase the fan speed of the respective system node to meet the higher cooling requirements which in turn will cause a higher noise emission level of up to 15 decibels. The processor frequency will typically stay within the limits given by the MTFR but may get lowered to frequencies between the MTFR lower limit and the nominal frequency at high ambient temperatures above 25 °C (77 °F). If the datacenter ambient environment is less than 25 °C, the frequency in MPM will consistently be in the upper range of the MTFR (roughly 10% to 20% better than nominal). At lower ambient temperatures (below 25 °C), MPM mode also provides deterministic performance. As the ambient temperature increases above 25 °C, determinism can no longer be guaranteed. The idle power saver (IPS) function can be enabled or disabled. If IPS is enabled the frequency is dropped to the static power saver level if the entire system meets the configured idle conditions.
Figure 2-8 shows the frequency ranges for the POWER9 static nominal mode (all modes disabled), the dynamic performance mode, and the maximum performance mode. The frequency adjustments for different workload characteristic, ambient conditions and idle states are also indicated.

![Frequency Diagram](image)

*Figure 2-8  POWER9 power management modes and related frequency ranges*

The Table 2-6 shows the static nominal and the static power saver mode frequencies as well as the frequency ranges of the dynamic performance mode and the maximum performance mode for all four processor module types available for the Power E980 system.

**Table 2-6  Characteristic frequencies and frequency ranges for Power E980**

<table>
<thead>
<tr>
<th>Feature code</th>
<th>Cores per single-chip module</th>
<th>Static nominal frequency [GHz]</th>
<th>Static power saver mode frequency [GHz]</th>
<th>Dynamic performance mode frequency range [GHz]</th>
<th>Maximum performance mode frequency range [GHz]</th>
</tr>
</thead>
<tbody>
<tr>
<td>#EFP1</td>
<td>8</td>
<td>3.4</td>
<td>2.75</td>
<td>3.4 - 4.0</td>
<td>3.9 - 4.0</td>
</tr>
<tr>
<td>#EFP2</td>
<td>10</td>
<td>3.15</td>
<td>2.75</td>
<td>3.15 - 3.9</td>
<td>3.7 - 3.9</td>
</tr>
<tr>
<td>#EFP3</td>
<td>12</td>
<td>2.90</td>
<td>2.75</td>
<td>2.9 - 3.9</td>
<td>3.55 - 3.9</td>
</tr>
<tr>
<td>#EFP4</td>
<td>11</td>
<td>3.00</td>
<td>2.75</td>
<td>3.0 - 3.9</td>
<td>3.58 - 3.9</td>
</tr>
</tbody>
</table>

Figure 2-9 shows schematically the POWER9 processor frequency as a function of power management mode and system utilization.
The default performance mode depends on the POWER9 processor-based server model. For Power E980 systems the dynamic performance mode is enabled by default.

The controls for all power saver modes are available on the advanced system management interface (ASMI) and can be dynamically modified. This includes to enable or to disable the idle power saver function and the change of EnergyScale tunable parameters. A system administrator may also use the hardware management console (HMC) to disable all power saver modes or to enable one of the three available power and performance modes: static power saver mode, dynamic performance mode, or maximum performance mode.

Figure 2-10 shows the ASMI menu for Power and Performance Mode Setup on a Power E980 server.
Figure 2-10 Power E980 ASMI menu for power and performance mode setup

For a more detailed discussion of the POWER9 EnergyScale technology please refer to the IBM EnergyScale for POWER9 Processor-Based Systems white paper available at the following URL -- needs update... points to P8 wp at the moment --:


2.1.11 Comparison of the POWER9, POWER8, and POWER7+ processors

The Power E980 and the Power E950 systems are utilizing exclusively processor modules with twelve cores capable to support eight simultaneous multi-threading (SMT) execution contexts. This 12-core modules are referred to as POWER9 scale-up processor modules. All other POWER9 processor based systems namely Power S914, Power LC921, Power S922, Power H922, Power L922, Power LC922, Power AC922, Power S924, and Power H924 run processor modules with twenty-four cores capable to support four SMT execution contexts. This 24-core modules are referred to as POWER9 scale-out processor modules.

Table 2-7 shows key features and characteristics in comparison between the POWER9 scale-up, POWER9 scale-out, POWER8, and POWER7+ processor implementations.
Table 2-7  Comparison of technology for the POWER9 processor and prior processor generations

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>POWER9 scale-up</th>
<th>POWER9 scale-out</th>
<th>POWER8</th>
<th>POWER7+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>14 nm</td>
<td>14 nm</td>
<td>22 nm</td>
<td>32 nm</td>
</tr>
<tr>
<td>Die size</td>
<td>68.5 mm x 68.5 mm</td>
<td>68.5 mm x 68.5 mm</td>
<td>649 mm²</td>
<td>567 mm²</td>
</tr>
<tr>
<td>Number of transistors</td>
<td>8 billion</td>
<td>8 billion</td>
<td>4.2 billion</td>
<td>2.1 billion</td>
</tr>
<tr>
<td>Maximum cores</td>
<td>12</td>
<td>24</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>Maximum SMT threads per core</td>
<td>8 threads</td>
<td>4 threads</td>
<td>8 threads</td>
<td>4 threads</td>
</tr>
<tr>
<td>Maximum frequency</td>
<td>3.9 - 4.0 GHz</td>
<td>3.8 - 4.0 GHz</td>
<td>4.15 GHz</td>
<td>4.4 GHz</td>
</tr>
<tr>
<td>L2 Cache</td>
<td>512 KB per core</td>
<td>256 KB per core</td>
<td>512 KB per core</td>
<td>256 KB per core</td>
</tr>
<tr>
<td>L3 Cache</td>
<td>10 MB of L3 cache per core with each core having access to the full 120 MB of L3 cache, on-chip eDRAM</td>
<td>10 MB of L3 cache shared by two cores with each core having access to the full 120 MB of L3 cache, on-chip eDRAM</td>
<td>8 MB of 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 L3 cache per core with each core having access to the full 80 MB of L3 cache, on-chip eDRAM</td>
</tr>
<tr>
<td>Memory support</td>
<td>DDR4 and DDR3a</td>
<td>DDR4</td>
<td>DDR3 and DDR4</td>
<td>DDR3</td>
</tr>
<tr>
<td>I/O bus</td>
<td>PCIe Gen4</td>
<td>PCIe Gen4</td>
<td>PCIe Gen3</td>
<td>GX++</td>
</tr>
</tbody>
</table>

a. Only DDR3 memory CDIMMs which are transferred in the context of a model upgrade from Power E870, Power E870C, Power E880, or Power E880C systems to a Power E980 server are supported.

2.2 Memory subsystem

The Power E980 uses the same custom DIMM (CDIMM) technology as found in POWER8-based systems. Each system node has 32 CDIMM slots capable of supporting 16 GB, 32 GB, 64 GB, 128 GB, 256 GB, and 512 GB CDIMMs running at a speed of 1600 MHz.

The Power E980 system supports both DDR3 and DDR4 versions of the CDIMM. Mixing of DDR3 and DDR4 CDIMMs are not supported in the same system node however DDR3 CDIMMs in one system node and DDR4 CDIMMs in another system node is supported in the same Power E980 server.

New orders of Power E980 servers can only be configured with DDR4 CDIMMs of 32 GB, 64 GB, 128 GB, 256 GB, and 512 GB capacities.

To provide significant investment protection, the 16 GB, 32 GB, 64 GB, 128 GB, 256 GB DDR4 CDIMMs and the 16 GB, 32 GB, 64 GB, 128 GB DDR3 CDIMMs of the Power E870, Power E870C, Power E880, and Power E880C servers are supported in the context of model upgrades to Power E980 systems.

The memory subsystem of the Power E980 server allows for a maximum system memory of 16 TB per system node. A 4-node system can support up to 64 TB of system memory.

The memory of Power E980 systems is Capacity on Demand capable, allowing for the purchase of additional physical memory capacity which can then be dynamically activated when needed. It is required that 50% of installed memory capacity is active.
The Power E980 server supports an optional feature called Active Memory Expansion (#EM89). This allows the effective maximum memory capacity to be much larger than the true physical memory. This feature uses a dedicated coprocessor on the POWER9 processor to compress memory pages as they are written to and decompress them as they are read from memory. This can deliver memory expansion of up to 125%, depending on the workload type and its memory usage.

### 2.2.1 Custom DIMM

Custom DIMMs (CDIMMs) are innovative memory DIMMs that house industry-standard DRAM memory chips and a set of components that allow for higher bandwidth, lower latency communications and increased availability. These components include:

- Memory Scheduler
- Memory Management (RAS Decisions & Energy Management)
- Memory Buffer

By adopting this architecture for the memory DIMMs, several decisions and processes regarding memory optimizations are run internally into the CDIMM. This saves bandwidth and allows for faster processor-to-memory communications. This also allows for a more robust RAS. For more information, see Chapter 4, “Reliability, availability, serviceability, and manageability” on page 129.

Depending on the memory capacity the CDIMMs are manufactured in a Tall CDIMM or a Short CDIMM form factor. The 16 GB, 32 GB, and 64 GB CDIMMs are Short CDIMMs and the 128 GB, 256 GB and 512 GB CDIMMs are of the Tall CDIMM form factor. Each design will be comprised of a varying number of 4 Gb or 8 Gb SDRAM chips depending on the total capacity of the CDIMM. The large capacity 256 GB and 512 GB CDIMMs are based on two-high (2H) and four-high (4H) 3D-stacked (3DS) DRAM technology respectively. The CDIMM slots for the Power E980 are Tall CDIMM slots. A filler is added to a Short CDIMM allowing it to properly latch into the same physical location of a Tall CDIMM and ensure proper airflow and ease of handling. Tall CDIMMs slots allow for larger DIMM sizes and potentially a more seamless adoption of future technologies.
A detailed diagram of the CDIMMs available for the Power E980 server can be seen in Figure 2-11.

The Memory Buffer is a L4 cache and is built on eDRAM technology (same as the L3 cache), which has a lower latency than regular SRAM. Each CDIMM has 16 MB of L4 cache and a fully populated Power E980 server has 2 GB of L4 Cache. The L4 Cache performs several functions that have direct impact on performance and bring a series of benefits for the Power E980:

- 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.
- Partial write operations that target the same cache block are gathered within the L4 cache before being written to memory, 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

For a Power E980 system each memory feature code provides four CDIMMs. Therefore a maximum of 8 memory feature codes per system node are allowed in order to fill all the 32 CDIMM slots. To populate all 128 CDIMM slots of a 4-node Power E980 requires 32 memory features.
All the memory CDIMMs are capable of capacity upgrade on demand and must have a minimum of 50% of their physical capacity activated. For example, the minimum installed memory for a Power E980 server is 512 GB, which requires a minimum of 256 GB memory activations.

For the Power E980, the following 1600 MHz DDR4 DRAM memory options are available when placing an initial order:

- 128 GB (4 x 32 GB), (#EF20)
- 256 GB (4 x 64 GB), (#EF21)
- 512 GB (4 x 128 GB), (#EF22)
- 1024 GB (4 x 256 GB), (#EF23)
- 2048 GB (4 x 512 GB), (#EF24)

Each processor module has two memory controllers. These memory controllers must have at least a pair of CDIMMs attached to it. This set of mandatory four CDIMMs is called a memory quad. A logical diagram of a POWER9 processor with its two memory quads attached to the memory controllers MC0 and MC1 can be seen in Figure 2-12.

```
Figure 2-12   Logical diagram of a POWER9 processor and its two memory quads
```

The basic rules for memory placement follow:

- Each feature code designates a set of four physically identical CDIMMs which is also referred to as memory quad.
- One memory quad must be allocated to each installed processor module, which equals to at least one memory feature code per installed processor module.
- Of the 32 CDIMM slots in any given system node a minimum of 16 must be populated.
- A given processor can only have four or eight CDIMMs attached to it.
- Not all CDIMMs connected to the same POWER9 processor module must be identical and it is valid to configure two memory features of different CDIMM capacity per POWER9 processor module. However, for best performance results the CDIMM size is ideally selected to be the same.
- At least 50% of the installed memory must be activated via memory activation features.

The suggested approach is to install memory evenly across all processors and across all system nodes in the server and the chosen CDIMM size is consistently equal for all memory slots. Balancing memory across the installed processors allows memory access in a consistent manner and typically results in the best possible performance for your configuration. You should 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.

A physical diagram with the location codes of the memory CDIMMs of a system node, as well as their grouping as memory quads, can be seen in Figure 2-13. Each system node has eight
memory quads attached to the memory controllers MC0 and MC1 of the respective processor modules. The quads are identified by individually assigned color codes in Figure 2-13.

Each Power E980 system node requires four memory quads to populate the required minimum of 16 CDIMM slots. The location codes of the slots for the memory quads 1 to 4 are shown in the listing below. There is no specific ordering sequence implied since all four quads have to be present in a valid configuration.

- Quad 1: P1-C46, P1-C47, P1-C52, P1-C53 (slots connected to Processor P0)
- Quad 4: P1-C22, P1-C23, P1-C28, P1-C29 (slots connected to Processor P1)
- Quad 2: P1-C38, P1-C39, P1-C44, P1-C45 (slots connected to Processor P2)
- Quad 3: P1-C30, P1-C31, P1-C36, P1-C37 (slots connected to Processor P3)

No mandatory plugging sequence needs to be adhered to for the population of any of the remaining open CDIMM slots. The locations codes for memory quad 5 to 8 are shown in the following listing:

- Quad 5: P1-C48, P1-C49, P1-C50, P1-C51 (slots connected to Processor P0)
- Quad 8: P1-C24, P1-C25, P1-C26, P1-C27 (slots connected to Processor P1)
- Quad 6: P1-C40, P1-C41, P1-C42, P1-C43 (slots connected to Processor P2)
- Quad 7: P1-C32, P1-C33, P1-C34, P1-C35 (slots connected to Processor P3)
As mentioned before the numbering of quad 5 to 8 does not indicate any ordinal sequence and any quad can be assigned to any processor module. The solution designer has the flexibility to assign the additional memory quads to any processor module as long as the minimum memory configuration has been established. Furthermore, in multi system node Power E980 servers, the solution designer can either fully populate one drawer and have the other drawers partially populated or have all the drawers symmetrically populated. For example, consider a 2-node Power E980 server with six additional quads of memory beyond the eight quads needed to fulfill the minimum memory configuration requirement. One option would be to install four quads in one node and two quads in the other node. In an alternative configuration both system nodes could be expanded by three quads each.

### 2.2.3 Memory activation

All the memory CDIMMs are capable of capacity upgrade on demand and must have a minimum of 50% of their physical capacity activated. For example, the minimum physical installed memory for the Power E980 system is 512 GB, which requires a minimum of 256 GB activated.

There are two activation types that can be used to accomplish this:

- **Static memory activations**: Memory activations that are exclusive for a single server
- **Mobile memory activations**: Memory activations that can be moved from server to server in a power enterprise pool.

Both types of memory activations can co-reside in the same system, as long as at least 25% of the memory activations are static. This leads to a maximum of 75% of the memory activations as mobile.

In Figure 2-14, there is an example of the minimum required activations for a system with 1 TB of installed memory.

![Figure 2-14 Example of the minimum required activations for a system with 1 TB of installed memory](image)

The granularity for static memory activation is 1 GB, while for mobile memory activation, the granularity is 100 GB.
Specific feature codes support memory activations for DDR3 or DDR4 memory CDIMMs which were transferred from Power E880 or Power E880C systems in the context of a model upgrade to a Power E980 server.

In Table 2-8, there is a list of the feature codes that can be used to achieve the desired number of activations:

<table>
<thead>
<tr>
<th>Feature code</th>
<th>Feature description</th>
<th>Amount of memory</th>
<th>Type of activation</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMAT</td>
<td>1 GB Memory activation for M9S</td>
<td>1 GB</td>
<td>Static</td>
</tr>
<tr>
<td>EMAU</td>
<td>100 GB Memory activation for M9S</td>
<td>100 GB</td>
<td>Static</td>
</tr>
<tr>
<td>EMAV</td>
<td>100 GB Mobile memory activation for M9S</td>
<td>100 GB</td>
<td>Mobile</td>
</tr>
<tr>
<td>EFA1</td>
<td>1 GB Memory Activation (Upgrade from P8)</td>
<td>1 GB</td>
<td>Static</td>
</tr>
<tr>
<td>EFA2</td>
<td>100 GB Memory Activation (Upgrade from P8)</td>
<td>100 GB</td>
<td>Static</td>
</tr>
</tbody>
</table>

Static memory activations can be converted to mobile memory activations after system installation. In order to enable mobile memory activations, the systems must be part of a power enterprise pool and have feature code #EB35 configured. For more information about power enterprise pools, see 2.3.5, “Power enterprise pools and mobile capacity on demand (Mobile CoD)” on page 80.

### 2.2.4 Memory throughput

The Power E980 system can be configured with up to 16 processor modules. Each processor module drives 8 memory channels at 9.6 GTps speed. Any given transaction can provide 2 byte data read and 1 byte data write simultaneously. Memory bandwidth certainly varies with the workload but the maximum theoretical memory bandwidth when using CDIMM at 1600 Mb/s frequency in all 8 slots of a given processor module is approximately 230 GBps:

\[
\text{9.6 GTps} \times 3 \text{ bytes/channel} \times 8 \text{ channels} = 230.4 \text{ GBps}
\]

The total maximum theoretical memory bandwidth per Power E980 system node is 921.6 GB/s and the total maximum theoretical memory bandwidth per 4-node Power E980 system accumulates to 3686.4 GBps.

As data flows from main memory towards the execution units of the POWER9 processor they pass through the 512 KB L2 and the 64 KB L1 cache. In many cases the 10 MB L3 victim cache may also provide the data needed for the instruction execution.

Table 2-9 below shows the maximum cache bandwidth for a single core as defined by the width of the relevant channels and the related transaction rates on the Power E980 system.

<table>
<thead>
<tr>
<th>Cache level of POWER9 core</th>
<th>Power E980 cache bandwidtha</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.9 to 4.0 GHz core (#EFP1) [GBps]</td>
</tr>
<tr>
<td>L1 64 KB data cache</td>
<td>374 to 384</td>
</tr>
<tr>
<td>L2 512 KB cache</td>
<td>374 to 384</td>
</tr>
</tbody>
</table>
The bandwidth figures for the caches are calculated as follows:

- **L1 data cache:** In one clock cycle, four 16-byte load operations and two 16-byte store operation can be accomplished. The value varies depending on the core frequency and are computed as follows:
  - Core running at 3.55 GHz: \((4 \cdot 16 \text{ B} + 2 \cdot 16 \text{ B}) \cdot 3.55 \text{ GHz} = 340.80 \text{ GBps}\)
  - Core running at 3.58 GHz: \((4 \cdot 16 \text{ B} + 2 \cdot 16 \text{ B}) \cdot 3.58 \text{ GHz} = 343.68 \text{ GBps}\)
  - Core running at 3.70 GHz: \((4 \cdot 16 \text{ B} + 2 \cdot 16 \text{ B}) \cdot 3.70 \text{ GHz} = 355.20 \text{ GBps}\)
  - Core running at 3.90 GHz: \((4 \cdot 16 \text{ B} + 2 \cdot 16 \text{ B}) \cdot 3.90 \text{ GHz} = 374.40 \text{ GBps}\)
  - Core running at 4.00 GHz: \((4 \cdot 16 \text{ B} + 2 \cdot 16 \text{ B}) \cdot 4.00 \text{ GHz} = 384.00 \text{ GBps}\)

- **L2 cache:** In one clock cycle, one 64-byte read operation to the core and two 16-byte store operation from the core can be accomplished. The value varies depending on the core frequency and are computed as follows:
  - Core running at 3.55 GHz: \((1 \cdot 64 \text{ B} + 2 \cdot 16 \text{ B}) \cdot 3.55 \text{ GHz} = 340.80 \text{ GBps}\)
  - Core running at 3.58 GHz: \((1 \cdot 64 \text{ B} + 2 \cdot 16 \text{ B}) \cdot 3.58 \text{ GHz} = 343.68 \text{ GBps}\)
  - Core running at 3.70 GHz: \((1 \cdot 64 \text{ B} + 2 \cdot 16 \text{ B}) \cdot 3.70 \text{ GHz} = 355.20 \text{ GBps}\)
  - Core running at 3.90 GHz: \((1 \cdot 64 \text{ B} + 2 \cdot 16 \text{ B}) \cdot 3.90 \text{ GHz} = 374.40 \text{ GBps}\)
  - Core running at 4.00 GHz: \((1 \cdot 64 \text{ B} + 2 \cdot 16 \text{ B}) \cdot 4.00 \text{ GHz} = 384.00 \text{ GBps}\)

- **L3 cache:** With two clock cycles, one 64-byte read operation to the L2 cache and one 64-byte store operation from the L2 cache can be accomplished. The value varies depending on the core frequency and are computed as follows:
  - Core running at 3.55 GHz: \((1 \cdot 64 \text{ B} + 1 \cdot 64 \text{ B}) \cdot 3.55 \text{ GHz} / 2 = 227.10 \text{ GBps}\)
  - Core running at 3.58 GHz: \((1 \cdot 64 \text{ B} + 1 \cdot 64 \text{ B}) \cdot 3.58 \text{ GHz} / 2 = 229.12 \text{ GBps}\)
  - Core running at 3.70 GHz: \((1 \cdot 64 \text{ B} + 1 \cdot 64 \text{ B}) \cdot 3.70 \text{ GHz} / 2 = 236.80 \text{ GBps}\)
  - Core running at 3.90 GHz: \((1 \cdot 64 \text{ B} + 1 \cdot 64 \text{ B}) \cdot 3.90 \text{ GHz} / 2 = 249.60 \text{ GBps}\)
  - Core running at 4.00 GHz: \((1 \cdot 64 \text{ B} + 1 \cdot 64 \text{ B}) \cdot 4.00 \text{ GHz} / 2 = 256.00 \text{ GBps}\)

For each system node of a Power E980 populated with four processor modules and all its memory CDIMMs filled, the overall bandwidths as defined by the width of the relevant channels and the related transaction rates are shown in Table 2-10.

<table>
<thead>
<tr>
<th>Cache level of POWER9 core</th>
<th>Power E980 cache bandwidth (^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.9 to 4.0 GHz core ((#EFP1)) [GBps]</td>
<td>3.7 to 3.9 GHz core ((#EFP2)) [GBps]</td>
</tr>
<tr>
<td>L1 10 MB cache</td>
<td>250 to 256</td>
</tr>
</tbody>
</table>

\(^a\) Values are rounded to the nearest integer.

<table>
<thead>
<tr>
<th>Memory architecture entity</th>
<th>Power E980 cache and system memory bandwidth per node (^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>32 cores ((#EFP1)) @ 3.9 to 4.0 GHz [GBps]</td>
<td>40 cores ((#EFP2)) @ 3.7 to 3.9 GHz [GBps]</td>
</tr>
<tr>
<td>L1 64KB data cache</td>
<td>11,981 to 12,288</td>
</tr>
<tr>
<td>L2 512 KB cache</td>
<td>11,981 to 12,288</td>
</tr>
<tr>
<td>L3 10 MB cache</td>
<td>7,987 to 8,192</td>
</tr>
<tr>
<td>System memory</td>
<td>922</td>
</tr>
</tbody>
</table>

\(^a\) Values are rounded to the nearest integer.
For the entire Power E980 system configured with four system nodes, the accumulated bandwidth values are shown in Table 2-11 on page 74.

### Table 2-11  Power E980 4-node server total architectural maximum cache and memory bandwidth

<table>
<thead>
<tr>
<th>Memory architecture entity</th>
<th>Power E980 cache and system memory bandwidth for 4-node system(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 64 KB data cache</td>
<td>47,923 to 49,152 GBps</td>
</tr>
<tr>
<td></td>
<td>56,832 to 59,904 GBps</td>
</tr>
<tr>
<td></td>
<td>60,488 to 65,894 GBps</td>
</tr>
<tr>
<td></td>
<td>65,434 to 71,885 GBps</td>
</tr>
<tr>
<td>L2 512 KB cache</td>
<td>47,923 to 49,152 GBps</td>
</tr>
<tr>
<td></td>
<td>56,832 to 59,904 GBps</td>
</tr>
<tr>
<td></td>
<td>60,488 to 65,894 GBps</td>
</tr>
<tr>
<td></td>
<td>65,434 to 71,885 GBps</td>
</tr>
<tr>
<td>L3 10 MB cache</td>
<td>31,949 to 32,768 GBps</td>
</tr>
<tr>
<td></td>
<td>37,888 to 39,936 GBps</td>
</tr>
<tr>
<td></td>
<td>40,325 to 43,603 GBps</td>
</tr>
<tr>
<td>System memory</td>
<td>3,686</td>
</tr>
<tr>
<td></td>
<td>3,686</td>
</tr>
<tr>
<td></td>
<td>3,686</td>
</tr>
<tr>
<td></td>
<td>3,686</td>
</tr>
</tbody>
</table>

\(^a\) Values are rounded to the nearest integer.

#### 2.2.5 Active Memory Mirroring

The Power E980 systems have the ability to provide mirroring of the Power Hypervisor code across multiple memory CDIMMs. If a CDIMM that contains the hypervisor code develops an uncorrectable error, its mirrored partner will enable the system to continue to operate uninterrupted.

Active Memory Mirroring (AMM) is included with all Power E980 systems at no additional charge. It can be enabled, disabled, or re-enabled depending on the user's requirements.

The hypervisor code logical memory blocks will be mirrored on distinct CDIMMs to allow for more usable memory. There is no specific CDIMM that hosts the Hypervisor memory blocks so the mirroring is done at the logical memory block level, not at the CDIMM level. To enable the AMM feature it is mandatory that the server has enough free memory to accommodate the mirrored memory blocks.

Besides the hypervisor code itself, other components that are vital to the server operation are also mirrored:

- Hardware page tables (HPTs), responsible for tracking the state of the memory pages assigned to partitions
- Translation control entities (TCEs), responsible for providing I/O buffers for the partition's communications
- Memory used by the hypervisor to maintain partition configuration, I/O states, virtual I/O information, and partition state

It is possible to check whether the Active Memory Mirroring option is enabled and change its current status through the classical GUI of the HMC, under the Advanced Tab on the CEC Properties panel. If you are using the enhanced GUI of the HMC you find the relevant information and controls in the Memory Mirroring section of the General Settings panel of the selected Power E980 system (Figure 2-15).
After a failure on one of the CDIMMs containing hypervisor data occurs, all the server operations remain active and the flexible service processor (FSP) will isolate the failing CDIMMs. Systems stay in the partially mirrored state until the failing CDIMM is replaced.

There are components that are not mirrored because they are not vital to the regular server operations and require a larger amount of memory to accommodate its data:

- Advanced Memory Sharing Pool
- Memory used to hold the contents of platform dumps

**Partition data:** Active Memory Mirroring will *not* mirror partition data. It was designed to mirror only the hypervisor code and its components, allowing this data to be protected against a DIMM failure.

With AMM, uncorrectable errors in data that are owned by a partition or application are handled by the existing Special Uncorrectable Error handling methods in the hardware, firmware, and operating system.

### 2.2.6 Memory Error Correction and Recovery

The memory has error detection and correction circuitry is designed such that the failure of any one specific memory module within an ECC word can be corrected without any other fault.

In addition, a spare DRAM per rank on each memory port provides for dynamic DRAM device replacement during runtime operation. Also, dynamic lane sparing on the memory channel’s differential memory interface (DMI) link allows for repair of a faulty data lane.

Other memory protection features include retry capabilities for certain faults detected at both the memory controller and the memory buffer.
Memory is also periodically scrubbed to allow for soft errors to be corrected and for solid single-cell errors reported to the Hypervisor, which supports operating system deallocation of a page associated with a hard single-cell fault.

For more details on Memory RAS, see 4.4, “Memory RAS details” on page 139.

2.2.7 Special Uncorrectable Error handling

Special Uncorrectable Error (SUE) handling prevents an uncorrectable error in memory or cache from immediately causing the system to terminate. Rather, the system tags the data and determines whether it will ever be used again. If the error is irrelevant, it does not force a checkstop. If the data is used, termination can be limited to the program/kernel or Hypervisor owning the data, or freeze of the I/O adapters controlled by an I/O hub controller if data is to be transferred to an I/O device.

2.3 Capacity on Demand

Several types of Capacity on Demand (CoD) offerings are optionally available on the Power E980 server to help meet changing resource requirements in an on-demand environment, by using resources that are installed on the system but that are not activated. Activation codes are published at the following location:

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

The following convention is used in the Order type column in all tables in this section.

Initial Only available when ordered as part of a new system
MES Only available as an MES upgrade
Both Available with a new system or as part of an upgrade
Supported Unavailable as a new purchase but supported when migrated from another system or as part of a model conversion

2.3.1 What’s new?

Capacity on Demand for the Power E980 is very similar to capability in the Power E880.

- Minimum of static processor activations or Power Linux activations per system equal to the number of cores per processor feature. If a system uses 12-core processors, there is a system wide minimum of 12 core activations required.
- Minimum of 50% of installed memory capacity must have permanent activations. These activations can be static, mobile-enabled, mobile, or Power Linux.
- At least 25% of memory capacity must have static activations or Power Linux activations.
- The Power E980 can participate in the same Power Enterprise Pool as other Power E980s, as well as with previous generation E870s, E870Cs, E880s, and E880Cs.
- CUoD, Elastic CoD, Utility CoD, and Trial CoD are all available with the Power E980.

2.3.2 Capacity Upgrade on Demand (CUoD)

The Power E980 system includes a number of active processor cores and memory units. They can also include inactive processor cores and memory units. Active processor cores or memory units are processor cores or memory units that are already available for use on your server when it comes from the manufacturer. Inactive processor cores or memory units are
processor cores or memory units that are included with your server, but not available for use until you activate them. Inactive processor cores and memory units can be permanently activated by purchasing an activation feature called Capacity Upgrade on Demand (CUoD) and entering the provided activation code on the hardware management console for the server.

With the CUoD offering, you can purchase additional static processor or memory capacity and dynamically activate them when needed, without requiring you to restart your server or interrupt your business. All the static processor or memory activations are restricted to a single server.

Capacity Upgrade on Demand can have several applications to allow for a more flexible environment. One of its benefits is to allow for a given company to reduce the initial investment on a system. Traditional projects using other technologies require that a system is acquired with all the resources available to support the whole lifecycle of the project. This might incur in costs that would only be necessary on later stages of the project, usually with impacts on software licensing costs and software maintenance.

By using Capacity Upgrade on Demand a company could start with a system with enough installed resources to support the whole project lifecycle but only with enough active resources necessary for the initial project phases. More resources could be added along with the project, adjusting the hardware platform with the project needs. This would allow for a company to reduce the initial investment in hardware and only acquire software licenses that are needed on each project phase, reducing the Total Cost of Ownership and Total Cost of Acquisition of the solution. Figure 2-16 shows a comparison between two scenarios: a fully activated system versus a system with CUoD resources being activated along with the project timeline.

![Figure 2-16 Active cores scenarios comparison during a project lifecycle](image)
### 2.3.3 Static activations

Static processor and memory activations are restricted to a single system. They are enabled by entering a code into the HMC managing the system. The additional cores or memory is immediately available to be used for workloads on the system.

Static activations are available in three varieties:

- **Static activations** This is a standard processor/memory activation that can run AIX, IBM i, and Linux workloads.
- **Static activations for Linux** This is a processor/memory activation that can only run Linux workloads.
- **Mobile-enabled activations** This is a standard processor/memory activation that can run AIX, IBM i, and Linux workloads. Mobile-enabled activations can be converted to mobile activations at no charge. Mobile-enabled activations can be purchased with an initial order or an MES upgrade.

Table 2-12 lists the static processor activation features that are available for initial order on the Power E980.

<table>
<thead>
<tr>
<th>Processor feature</th>
<th>Static activation feature</th>
<th>Mobile enabled activation feature</th>
<th>Static activation for Linux feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>#EFP1 3.40 GHz 32-core</td>
<td>#EFPA</td>
<td>#EFPE</td>
<td>#ELBK</td>
</tr>
<tr>
<td>#EFP2 3.15 GHz 40-core</td>
<td>#EFPB</td>
<td>#EFPF</td>
<td>#ELBL</td>
</tr>
<tr>
<td>#EFP3 2.90 GHz 48-core</td>
<td>#EFPC</td>
<td>#EFPG</td>
<td>#ELBM</td>
</tr>
<tr>
<td>#EFP4 3.0 GHz 44-core</td>
<td>#EFP9</td>
<td>#EFPN</td>
<td>#ELBQ</td>
</tr>
</tbody>
</table>

Table 2-13 lists the static memory activation features that are available for initial order on the Power E980.

<table>
<thead>
<tr>
<th>Feature Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELMD</td>
<td>512 GB Power Linux Memory Activations for M9S/80H</td>
</tr>
<tr>
<td>EMAD</td>
<td>100 GB Mobile Enabled Memory Activations</td>
</tr>
<tr>
<td>EMAS</td>
<td>Base Memory activation (512) for #EHC6®</td>
</tr>
<tr>
<td>EMAT</td>
<td>1 GB Memory activation for M9S</td>
</tr>
<tr>
<td>EMAU</td>
<td>100 GB Memory activation for M9S</td>
</tr>
</tbody>
</table>

Change of name: Some web sites or documents still refer to Elastic Capacity on Demand as On/Off Capacity on Demand.

### 2.3.4 Elastic Capacity on Demand (Elastic CoD)

- a. This feature is only available with the Power E980 Solution Edition for Health care system. This feature is only available in the US and Canada.
With the Elastic CoD offering, you can temporarily activate and deactivate processor cores and memory units to help meet the demands of business peaks such as seasonal activity, period-end, or special promotions. Elastic CoD was previously called On/Off CoD. When you order an Elastic CoD feature, you receive an enablement code that allows a system operator to make requests for additional processor and memory capacity in increments of one processor day or 1 GB memory day. The system monitors the amount and duration of the activations. Both prepaid and post-pay options are available.

Charges are based on usage reporting that is collected monthly. Processors and memory may be activated and turned off an unlimited number of times, when additional processing resources are needed.

This offering provides a system administrator an interface at the HMC to manage the activation and deactivation of resources. A monitor that resides on the server records the usage activity. This usage data must be sent to IBM on a monthly basis. A bill is then generated based on the total amount of processor and memory resources utilized, in increments of processor and memory (1 GB) Days.

The Power E980 supports the 90-day temporary Elastic CoD processor and memory enablement features. These features allow the system to activate processor days and GB days equal to the number of inactive resources multiplied by 90 days. Thus if all resources are activated using Elastic CoD, a new enablement code will need to be ordered every 90 days. If only half of the inactive resources are activated using Elastic CoD a new enablement code will need to be ordered every 180 days.

Before using temporary capacity on your server, you must enable your server. To enable, an enablement feature (MES only) must be ordered and the required contracts must be in place. The 90 day enablement feature for the Power E980 processors is #EP9T. For memory the enablement feature is #EM9V.

If a Power E980 server uses the IBM i operating system in addition to any other supported operating system on the same server, the client must inform IBM which operating system caused the temporary Elastic CoD processor usage so that the correct feature can be used for billing.

The Elastic CoD process consists of three steps, enablement, activation, and billing.

- **Enablement**
  Before requesting temporary capacity on a server, you must enable it for Elastic CoD. To do this, order an enablement feature and sign the required contracts. IBM will generate an enablement code, mail it to you, and post it on the web for you to retrieve and enter on the target server.

- **Activation requests**
  When Elastic CoD temporary capacity is needed, use the HMC menu for On/Off CoD. Specify how many inactive processors or gigabytes of memory are required to be temporarily activated for some number of days. You are billed for the days requested, whether the capacity is assigned to partitions or remains in the shared processor pool.

  At the end of the temporary period (days that were requested), you must ensure that the temporarily activated capacity is available to be reclaimed by the server (not assigned to partitions), or you are billed for any unreturned processor days.

- **Billing**
  The contract, signed by the client before receiving the enablement code, requires the Elastic CoD user to report billing data at least once a month (whether or not activity occurs). This data is used to determine the proper amount to bill at the end of each billing period.
period (calendar quarter). Failure to report billing data for use of temporary processor or memory capacity during a billing quarter can result in default billing equivalent to 90 processor days of temporary capacity.

For more information about registration, enablement, and usage of Elastic CoD, visit the following location:

http://www.ibm.com/systems/power/hardware/cod

Table 2-14 lists Elastic CoD features available for the Power E980.

Table 2-14 Elastic CoD features

<table>
<thead>
<tr>
<th>Feature code</th>
<th>Description</th>
<th>Max</th>
<th>Order type</th>
</tr>
</thead>
<tbody>
<tr>
<td>EP9T</td>
<td>90 Days Elastic CoD Processor Core Enablement</td>
<td>1</td>
<td>MES</td>
</tr>
<tr>
<td>MMC1</td>
<td>ECOD Processor day - IBM i</td>
<td>9999</td>
<td>MES</td>
</tr>
<tr>
<td>MMCB</td>
<td>ECOD GB Memory Day - AIX/Linux</td>
<td>9999</td>
<td>MES</td>
</tr>
<tr>
<td>MMCX</td>
<td>ECOD Processor day - AIX/Linux</td>
<td>9999</td>
<td>MES</td>
</tr>
</tbody>
</table>

a. See 2.3, “Capacity on Demand” on page 76 for information on order types.

2.3.5 Power enterprise pools and mobile capacity on demand (Mobile CoD)

While static activations are valid for a single system, some customers could benefit from moving processor and memory activations among different servers due to workload rebalance or disaster recovery.

IBM power enterprise pools, is a technology for dynamically sharing processor and memory activations among a group (or pool) of IBM Power Systems servers. Using Mobile Capacity on Demand (CoD) activation codes, the systems administrator can perform tasks without contacting IBM.

This capability allows the movement of resources not only between Power E980 systems, but also between Power E980 and previous generation E870, E870C, E880, and E880C systems, resulting in unsurpassed flexibility for workload balancing and system maintenance.

A pool can support systems with different clock speeds or processor generations.

The basic rules for the Mobile Capacity on Demand follows:

- The Power E980 requires a minimum of eight static processor activations.
- The Power 870, Power 870C, Power E880, and Power 880C require a minimum of eight static processor activations.
- For all systems, 25% of the installed memory capacity must have static activations.

All the systems in a pool must be managed by the same HMC or by the same pair of redundant HMCs. If redundant HMCs are used, the HMCs must be connected to a network so that they can communicate with each other. The HMCs must have at least 2 GB of memory.

An HMC can manage multiple power enterprise pools and can also manage systems that are not part of a power enterprise pool. Systems can belong to only one power enterprise pool at a time. Powering down an HMC does not limit the assigned resources of participating systems in a pool but does limit the ability to perform pool change operations.
After a power enterprise pool is created, the HMC can be used to perform the following functions:

- Mobile CoD processor and memory resources can be assigned to systems with inactive resources. Mobile CoD resources remain on the system to which they are assigned until they are removed from the system.
- New systems can be added to the pool and existing systems can be removed from the pool.
- New resources can be added to the pool or existing resources can be removed from the pool.
- Pool information can be viewed, including pool resource assignments, compliance, and history logs.

In order for the Mobile activation features to be configured, it is necessary that a power enterprise pool is registered with IBM as well as the systems that are going to be included as members of the pool. Also, it is necessary that the systems have the feature code #EB35 for mobile enablement configured, and the required contracts must be in place.

Table 2-15 on page 81 lists the mobile processor and memory activation features that are available for the Power E980.

<table>
<thead>
<tr>
<th>Feature Code</th>
<th>Description</th>
<th>Max</th>
<th>Order type</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMAV</td>
<td>100 GB Mobile Memory activation for M9S/80H</td>
<td>124</td>
<td>MES</td>
</tr>
<tr>
<td>EFPD</td>
<td>Mobile processor activation for M9S/80H</td>
<td>184</td>
<td>MES</td>
</tr>
</tbody>
</table>

*a. See 2.3, “Capacity on Demand” on page 76 for information on order types.*

For more information about power enterprise pools, see the publication, *Power Enterprise Pools on IBM Power Systems*, REDP5101:


### 2.3.6 Utility Capacity on Demand (Utility CoD)

Utility CoD automatically provides additional processor performance on a temporary basis within the shared processor pool.

With Utility CoD, you can place a quantity of inactive processors into the server's shared processor pool, which then becomes available to the pool's resource manager. When the server recognizes that the combined processor utilization within the shared processor pool exceeds 100% of the level of base (purchased and active) processors that are assigned across uncapped partitions, then a Utility CoD processor minute is charged and this level of performance is available for the next minute of use.

If additional workload requires a higher level of performance, the system automatically allows the additional Utility CoD processors to be used, and the system automatically and continuously monitors and charges for the performance needed above the base (permanent) level.

Registration and usage reporting for utility CoD is made using a public web site and payment is based on reported usage. Utility CoD requires PowerVM Standard Edition or PowerVM Enterprise Edition to be active.
If a Power E980 server uses the IBM i operating system in addition to any other supported operating system on the same server, the client must inform IBM which operating system caused the temporary Utility CoD processor usage so that the correct feature can be used for billing.

For more information regarding registration, enablement, and use of Utility CoD, visit the following location:

### 2.3.7 Trial Capacity on Demand (Trial CoD)

A **standard request** for Trial CoD requires you to complete a form including contact information and vital product data (VPD) from your Power E980 system with inactive CoD resources.

A standard request activates two processors or 64 GB of memory (or 8 processor cores and 64 GB of memory) for 30 days. Subsequent standard requests can be made after each purchase of a permanent processor activation. An HMC is required to manage Trial CoD activations.

An **exception request** for Trial CoD requires you to complete a form including contact information and VPD from your Power E980 system with inactive CoD resources. An exception request will activate all inactive processors or all inactive memory (or all inactive processor and memory) for 30 days. An exception request can be made only one time over the life of the machine. An HMC is required to manage Trial CoD activations.

To request either a Standard or an Exception Trial, visit the following location:

### 2.3.8 Software licensing and CoD

For software licensing considerations with the various CoD offerings, see the most recent revision of the *Power Systems Capacity on Demand User’s Guide*:
http://www.ibm.com/systems/power/hardware/cod

### 2.4 System bus

This section provides additional information related to the internal buses.

#### 2.4.1 PCI Express Gen4

The internal I/O subsystem on the Power E980 is connected to the PCIe Controllers on a POWER9 processor in the system. Each POWER9 processor module has three PCIe host buses (PHBs). Two of the PHBs (PHB0 and PHB2) have 16 PCIe lanes each. In the Power E980 system they connect directly to two PCIe Gen4 x16 slots providing eight PCIe Gen4 x16 slots per system node. The third PHB on each POWER9 processor is used for other I/O connections:

- **4 x Internal NVMe SSDs**, each using a PCIe Gen4 x4 connection
- **3 x USB ports** that share a PCIe Gen4 x1 connection

Bandwidths for the connections are shown in Table 2-16 on page 83.
The system nodes allow for eight PCIe Gen4 x16 slots. Additional slots can be added by attaching PCIe Expansion Drawers and SAS disks can be attached to EXP24S SFF Gen2 Drawers. The PCIe Expansion Drawer is connected using an #EJ07 adapter. The EXP24S can be either attached to SAS adapters on the system nodes or on the PCIe Expansion Drawer.

The theoretical maximum bandwidth is as follows:

- 8 PCIe Gen4 Slots at 64 GBps = 512 GBps
4 NVMe slots at 8 GBPs = 32 GBps
1 USB controller at 1.25 GBps
545.25 = 512 + 32 + 1.25

For a list of adapters and their supported slots, see 2.5, “PCIe adapters” on page 85.

Disk support: There is no support for SAS disks directly installed on the system nodes and PCIe Expansion Drawers. If directly attached SAS disk are required, they need to be installed in a SAS disk drawer and connected to a supported SAS controller in one of the PCIe slots.

For more information about PCIe Expansion Drawers, see 2.7.1, “PCIe Gen3 I/O expansion drawer” on page 92.

### 2.4.2 Service Processor Bus

The redundant service processor bus connectors are located on the rear of the control unit and the system nodes. All of the service processor (SP) communication between the control unit and the system nodes flows through these cables.

Similar to the previous generation Power E870 and Power E880 systems, the redundant service processors are housed in the System Control Unit. However, the System Control Unit no longer hosts the system clock. Each system node hosts its own redundant clocks. The cables used to provide communications between the control units and system nodes depend on the amount of system nodes installed. When a system node is added, a new set of cables is also added.

The cables necessary for each system node have been grouped under a single feature code, allowing for an easier configuration. Each cable set includes a pair of FSP cables and when applicable SMP cables and UPIC cables. Table 2-17 shows a list of the feature codes available.

<table>
<thead>
<tr>
<th>Feature code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EFCA</td>
<td>System node to system control unit cable set for drawer 1</td>
</tr>
<tr>
<td>EFCB</td>
<td>System node to system control unit cable set for drawer 2</td>
</tr>
<tr>
<td>EFCC</td>
<td>System node to system control unit cable set for drawer 3</td>
</tr>
<tr>
<td>EFCD</td>
<td>System node to system control unit cable set for drawer 4</td>
</tr>
</tbody>
</table>

Cable sets feature codes are incremental and depend on the number of installed drawers as follows:

- 1 system node: #EFCA
- 2 system nodes: #EFCA and #EFCB
- 3 system nodes: #EFCA, #EFCB, and #EFCC
- 4 system nodes: #EFCA, #EFCB, #EFCC, and #EFCD

Disk support: There is no support for SAS disks directly installed on the system nodes and PCIe Expansion Drawers. If directly attached SAS disk are required, they need to be installed in a SAS disk drawer and connected to a supported SAS controller in one of the PCIe slots.

For more information about PCIe Expansion Drawers, see 2.7.1, “PCIe Gen3 I/O expansion drawer” on page 92.

### 2.4.2 Service Processor Bus

The redundant service processor bus connectors are located on the rear of the control unit and the system nodes. All of the service processor (SP) communication between the control unit and the system nodes flows through these cables.

Similar to the previous generation Power E870 and Power E880 systems, the redundant service processors are housed in the System Control Unit. However, the System Control Unit no longer hosts the system clock. Each system node hosts its own redundant clocks. The cables used to provide communications between the control units and system nodes depend on the amount of system nodes installed. When a system node is added, a new set of cables is also added.

The cables necessary for each system node have been grouped under a single feature code, allowing for an easier configuration. Each cable set includes a pair of FSP cables and when applicable SMP cables and UPIC cables. Table 2-17 shows a list of the feature codes available.

<table>
<thead>
<tr>
<th>Feature code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EFCA</td>
<td>System node to system control unit cable set for drawer 1</td>
</tr>
<tr>
<td>EFCB</td>
<td>System node to system control unit cable set for drawer 2</td>
</tr>
<tr>
<td>EFCC</td>
<td>System node to system control unit cable set for drawer 3</td>
</tr>
<tr>
<td>EFCD</td>
<td>System node to system control unit cable set for drawer 4</td>
</tr>
</tbody>
</table>

Cable sets feature codes are incremental and depend on the number of installed drawers as follows:

- 1 system node: #EFCA
- 2 system nodes: #EFCA and #EFCB
- 3 system nodes: #EFCA, #EFCB, and #EFCC
- 4 system nodes: #EFCA, #EFCB, #EFCC, and #EFCD
2.5 PCIe adapters

This section covers the type and functions of the PCIe adapter cards supported by the Power E980 system.

The following convention is used in the Order type column in all tables in this section.

Initial: Only available when ordered as part of a new system

MES: Only available as an MES upgrade

Both: Available with a new system or as part of an upgrade

Supported: Unavailable as a new purchase but supported when migrated from another system or as part of a model conversion

2.5.1 What’s new?

- The Power E980 supports PCIe Gen4 adapters in internal slots.
- PCIe Gen3 adapter are supported both internally and in the PCIe Gen3 I/O Expansion drawer.
- USB ports are integrated on the first two system nodes - no PCIe adapter is required.

2.5.2 PCI Express (PCIe)

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 such configurations, the connection is labelled as x1, x2, x8, x12, x16, or x32, where the number is effectively the number of lanes.

The PCIe interfaces supported on this system nodes are PCIe Gen4, capable of 32 GBps simplex (64 GBps duplex) on a single x16 interface. PCIe Gen4 slots also support previous generations (Gen3, Gen2, and Gen1) adapters, which operate at lower speeds, according to the following rules:

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

The Power E980 also supports expansion beyond the slots available in the system nodes by attaching one or more PCIe Gen3 I/O expansion drawers (#EMX0).

IBM POWER9 processor-based servers can support two different form factors of PCIe adapters:

- PCIe low profile (LP) cards, which are used with system node PCIe slots.
- PCIe full height and full high cards are used in the PCIe Gen3 I/O expansion drawer (#EMX0)

Low-profile PCIe adapter cards are supported only in low-profile PCIe slots, and full-height and full-high cards are supported only in full-high slots. Adapters that are low profile have “LP” in the adapter description.
Before adding or rearranging adapters, use the System Planning Tool to validate the new adapter configuration. For more information, see the System Planning Tool web site:

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 do this, use the IBM Prerequisite web site:

https://www-912.ibm.com/e_dir/eServerPreReq.nsf

The following sections describe the supported adapters and provide tables of orderable feature numbers. The tables indicate operating system support (AIX, IBM i, and Linux) for each of the adapters.

### 2.5.3 LAN adapters

Table 2-18 lists available LAN adapters that are supported in the Power E980 server.

**Table 2-18  Available LAN adapters**

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>OS support</th>
<th>Order type</th>
</tr>
</thead>
<tbody>
<tr>
<td>5260</td>
<td>576F</td>
<td>PCIe2 LP 4-port 1 GbE Adapter</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>5899</td>
<td>576F</td>
<td>PCIe2 4-port 1 GbE Adapter</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EC2R</td>
<td>58FA</td>
<td>PCIe3 LP 2-Port 10 Gb NIC&amp;ROCE SR/Cu Adapter</td>
<td>IBM i and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EC2S</td>
<td>58FA</td>
<td>PCIe3 2-Port 10 Gb NIC&amp;ROCE SR/Cu Adapter</td>
<td>IBM i and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EC2T</td>
<td>58FB</td>
<td>PCIe3 LP 2-Port 25/10 Gb NIC&amp;ROCE SR/Cu Adapter</td>
<td>IBM i and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EC2U</td>
<td>58FB</td>
<td>PCIe3 2-Port 25/10 Gb NIC&amp;ROCE SR/Cu Adapter</td>
<td>IBM i and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EC67</td>
<td>2CF3</td>
<td>PCIe4 LP 2-port 100 Gb ROCE EN LP adapter</td>
<td>IBM i and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EN0H</td>
<td>2B93</td>
<td>PCIe3 4-port (10 Gb FCoE &amp; 1 GbE) SR&amp;RJ45</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EN0J</td>
<td>2B93</td>
<td>PCIe3 LP 4-port (10 Gb FCoE &amp; 1 GbE) SR&amp;RJ45</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EN0K</td>
<td>2CC1</td>
<td>PCIe3 4-port (10 Gb FCoE &amp; 1 GbE) SFP+Copper&amp;RJ45</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EN0L</td>
<td>2CC1</td>
<td>PCIe3 LP 4-port (10 Gb FCoE &amp; 1 GbE) SFP+Copper&amp;RJ45</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EN0S</td>
<td>2CC3</td>
<td>PCIe2 4-Port (10 Gb+1 GbE) SR+RJ45 Adapter</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EN0T</td>
<td>2CC3</td>
<td>PCIe2 LP 4-Port (10 Gb+1 GbE) SR+RJ45 Adapter</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
</tbody>
</table>
2.5.4 Graphics adapters

There are currently no graphics adapters supported in the Power E980. Stay posted of IBM Power System announcements for changes in adapter support.

2.5.5 SAS adapters

Table 2-19 lists the SAS adapters that are available for the Power E980.

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>OS support</th>
<th>Order type</th>
</tr>
</thead>
<tbody>
<tr>
<td>EJ0J 57B4</td>
<td></td>
<td>PCIe3 RAID SAS Adapter Quad-port 6 Gb x8</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EJ0M 57B4</td>
<td></td>
<td>PCIe3 LP RAID SAS Adapter Quad-Port 6 Gb x8</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EJ10 57B4</td>
<td></td>
<td>PCIe3 SAS Tape/DVD Adapter Quad-port 6 Gb x8</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EJ11 57B4</td>
<td></td>
<td>PCIe3 LP SAS Tape/DVD Adapter Quad-port 6 Gb x8</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EJ14 57B1</td>
<td></td>
<td>PCIe3 12 GB Cache RAID PLUS SAS Adapter Quad-port 6 Gb x8</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
</tbody>
</table>

2.5.6 Fibre Channel adapters

Table 2-20 lists the Fibre Channel adapters available for the Power E980.
### Available Fibre Channel adapters

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>OS support</th>
<th>Order type&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>5273</td>
<td>577D</td>
<td>PCIe LP 8 Gb 2-Port Fibre Channel Adapter</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>5729</td>
<td>5729</td>
<td>PCIe2 8 Gb 4-port Fibre Channel Adapter</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>5735</td>
<td>577D</td>
<td>8 Gb PCI Express Dual Port Fibre Channel Adapter</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EN0A</td>
<td>577F</td>
<td>PCIe3 16 Gb 2-port Fibre Channel Adapter</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EN0B</td>
<td>577F</td>
<td>PCIe3 LP 16 Gb 2-port Fibre Channel Adapter</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EN1A</td>
<td>578F</td>
<td>PCIe3 32 Gb 2-port Fibre Channel Adapter</td>
<td>IBM i and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EN1B</td>
<td>578F</td>
<td>PCIe3 LP 32 Gb 2-port Fibre Channel Adapter</td>
<td>IBM i and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EN1C</td>
<td>578E</td>
<td>PCIe3 16 Gb 4-port Fibre Channel Adapter</td>
<td>IBM i and Linux</td>
<td>Both</td>
</tr>
<tr>
<td>EN1D</td>
<td>578E</td>
<td>PCIe3 LP 16 Gb 4-port Fibre Channel Adapter</td>
<td>IBM i and Linux</td>
<td>Both</td>
</tr>
</tbody>
</table>

<sup>a</sup> See 2.5, “PCIe adapters” on page 85 for information on order types.

### 2.5.7 USB adapters

The first and second nodes in any Power E980 system have three built-in USB 3.0 type A ports. If a third or forth node is installed, no USB ports are included on these nodes.

Stay posted of IBM Power System announcements for changes in adapter support.

The ports are found on the rear of the system enclosures as shown in Figure 2-18.
Chapter 2. Architecture and technical overview

2.5.8 InfiniBand host channel adapters

There are currently no InfiniBand PCIe adapters supported in the Power E980.

Stay posted of IBM Power System announcements for changes in adapter support.

2.5.9 Cryptographic Coprocessor

The Cryptographic Coprocessor cards provide both cryptographic coprocessor and cryptographic accelerator functions in a single card.

The IBM PCIe Cryptographic Coprocessor adapter highlights the following features:

- Integrated Dual processors that operate in parallel for higher reliability
- Supports IBM Common Cryptographic Architecture or PKCS#11 standard
- Ability to configure adapter as coprocessor or accelerator
- Support for smart card applications using Europay, MasterCard, and Visa
- Cryptographic key generation and random number generation
- PIN processing: generation, verification, translation
- Encrypt and Decrypt using AES and DES keys

See the following site for the most recent firmware and software updates:

http://www.ibm.com/security/cryptocards/
Table 2-21 lists the cryptographic adapter that is available for the server.

### Table 2-21  Available cryptographic adapters

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>OS support</th>
<th>Order type&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>EJ33</td>
<td>4767</td>
<td>PCIe3 Crypto Coprocessor BSC-Gen3 4767</td>
<td>AIX, IBM i, and Linux</td>
<td>Both</td>
</tr>
</tbody>
</table>

<sup>a</sup> See 2.5, “PCIe adapters” on page 85 for information on order types.

**Adapter height:** Note that the #EJ33 adapter is a full-height adapter so is only supported in the PCI Gen3 I/O Expansion drawer.

### 2.5.10 CAPI adapters

There are currently no supported IBM CAPI adapters or part numbers in the Power E980. The CAPI slots have been tested and cards are available from the following vendors:

- Nallatech 250SP [http://www.nallatech.com/250sp](http://www.nallatech.com/250sp)
- Flyslice FX609
- ReflexCES XpressVUP LP9P [https://www.reflexces.com/products](https://www.reflexces.com/products)

### 2.6 Internal storage

Each Power E980 system node supports up to four internal NVMe U.2 (2.5 inch 7mm form factor) solid state drives. The SSDs are accessible from the rear of the system node and reside in the middle of the PCIe adapter slots as shown in Figure 2-19 on page 90.
The internal SSD drives are intended for boot purposes only and not as a general purpose drive.

Table 2-22 shows available internal SSD drives.

**Table 2-22   Available internal NVMe SSD features**

<table>
<thead>
<tr>
<th>Feature code</th>
<th>CCIN</th>
<th>Description</th>
<th>OS support</th>
<th>Order type(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC5J</td>
<td>59B4</td>
<td>Mainstream 800 GB SSD NVMe U.2 module</td>
<td>AIX and Linux</td>
<td>Both</td>
</tr>
</tbody>
</table>

\(^a\) See 2.5, “PCIe adapters” on page 85 for information on order types.

**Internal SSD plug order**

For redundancy purposes it is ideal to distribute the NVMe drives across the system nodes if present. The recommended approach is as follows:

1. Populate slot C1 in each system node starting with node 1 and then other C1 slots if other nodes are present.
2. Populate slot C3 in each system node starting with node 1 and then other C3 slots if other nodes are present.
3. Populate slot C2 in each system node starting with node 1 and then other C2 slots if other nodes are present.
4. Populate slot C4 in each system node starting with node 1 and then other C4 slots if other nodes are present.

**2.7 External I/O subsystems**

This section describes the PCIe Gen3 I/O expansion drawer (#EMX0) that can be attached to the Power E980 system.

At the initial availability date September 21, 2018 the Power E980 system supports 2 PCIe Gen3 I/O drawer per system node yielding a maximum four I/O drawers per 2-node Power E980 system configuration. One I/O drawer supports 2 fan-out modules which offer 6 PCIe Gen3 adapter slots each. This delivers an additional 24 PCIe Gen3 slot capacity per system node and a maximum of 48 PCIe Gen3 slots per 2-node server.

8 slots in a system nodes are used to cable the four I/O drawers for a total of 56 available slots for a 2-node system.

With the availability of 3-node and 4-node Power E980 configurations in November 16, 2018 the number of supported PCIe Gen3 I/O drawers is raised to 4 per system node with a maximum of 16 per 4-node Power E980 system. A maximum of 48 PCIe Gen3 slots per system node and a maximum of 192 PCIe Gen3 slots per 4-node Power E980 server are available at that date.

Each fan-out module is attached by one optical cable adapter which occupies one x16 PCIe Gen4 slot of a system node. Consequently at the initial availability date a 1-node Power E980 system configuration with 2 I/O drawers attached provides a maximum of 28. With the enhanced configurations options available in November this numbers will increase to a maximum of 48 for a 1-node configuration since all node slots are required to cable the drawers and to a maximum of 192 available PCIe slots for a 4-node Power E980 server with 16 I/O drawers.
2.7.1 PCIe Gen3 I/O expansion drawer

The PCIe Gen3 I/O expansion drawer (#EMX0) is a 4U high, PCI Gen3-based and rack mountable I/O drawer. It offers two PCIe Fanout Modules (#EMXG) each of them providing six PCIe Gen3 full high, full length slots (two x16 and four x8). The older Fanout Modules (#EMXF) as they are used by many Power E870, Power E870C, Power E880, and Power E880C systems are also supported but are now longer available in the context of a new Power E980 system order.

See 1.3, “Physical package” for dimensions.

PCIe3 x16 to optical cable adapter (#EJ07) and 2.0 m (#ECC6), 10.0 m (#ECC8), or 20.0 m (#ECC9) CXP 16X active optical cables (AOC) connect the system node to a PCIe Fanout Module in the I/O expansion drawer. One feature #ECC6, one #ECC8, or one #ECC9 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 2-20 shows the back 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:

- 2.0 m Optical Cable Pair for PCIe3 Expansion Drawer (#ECC6)
- 10.0 m Optical Cable Pair for PCIe3 Expansion Drawer (#ECC8)
- 20.0 m Optical Cable Pair for PCIe3 Expansion Drawer (#ECC9)

**Cable lengths**: Use the 2.0 m cables for intra-rack installations. Use the 10.0 m or 20.0 m cables for inter-rack installations.

A minimum of one PCIe3 Optical Cable Adapter for PCIe3 Expansion Drawer (#EJ07) is required to connect to the PCIe3 6-slot fan-out module\(^1\) in the I/O expansion drawer. The top port of the fan-out module must be cabled to the top port of the #EJ07 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 (#EJ07). However the two fan-out modules in a single I/O expansion drawer can be connected to different system nodes in the same server.

---

\(^1\) Cabling rules and considerations apply to both supported fan-out modules, #EMXG and #EMXF.
Figure 2-21 shows connector locations for the PCIe Gen3 I/O expansion drawer.

Figure 2-22 shows typical optical cable connections.
General rules for the PCI Gen3 I/O expansion drawer configuration

The PCIe3 optical cable adapter (#EJ07) can be in any of the PCIe adapter slots in a Power E980 system node. However, we advise that you first populate the PCIe adapter slots with odd locations codes P1-C1, P1-C3, P1-C5, and P1-C7 and then start to fill the adapter slots with even location codes P1-C2, P1-C4, P1-C6, and P1-C8. Each processor module drives two PCIe Gen4 slots and all slots are equal in respect to their bandwidth characteristics. If you first use the slots with odd location codes you ensure that one PCIe Gen4 slot per processor module is populated before you start to use the second PCIe Gen4 slot of the processor modules. There is no preference in which order you fill the odd sequence or the even sequence locations.

Table 2-23 shows PCIe adapter slot priorities in the Power E980 system. If the sequence within the odd location codes and the sequence within the even locations codes is chosen as shown in the slot priority column the adapters are assigned to the single chip processor modules (SCM) in alignment with the internal enumeration order: SCM0, SCM1, SCM2, SCM3.

Table 2-23  PCIe adapter slot priorities

<table>
<thead>
<tr>
<th>Feature code</th>
<th>Description</th>
<th>Slot priorities</th>
</tr>
</thead>
<tbody>
<tr>
<td>EJ07</td>
<td>PCIe3 Optical Cable Adapter for PCIe3 Expansion Drawer</td>
<td>1, 7, 3, 5, 2, 8, 4, 6</td>
</tr>
</tbody>
</table>

The following figures show several examples of supported configurations. For simplification we have not shown every possible combination of the I/O expansion drawer to server attachments.

Figure 2-23 shows an example of a single system node and two PCI Gen3 I/O expansion drawers.
Figure 2-24 shows an example of two system nodes and two PCI Gen3 I/O expansion drawers.
Figure 2-25 shows an example of two system nodes and four PCI Gen3 I/O expansion drawers.

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 E950:

- EXP24S SFF Gen2-bay Drawer for high-density storage (#5887)
- EXP24SX and EXP12SX SAS Storage Enclosures (#ESLS)
- EXP12SX SAS Storage Enclosures (#ESLL)
- IBM System Storage

2.8.1 EXP24S SFF Gen2-bay Drawer

The EXP24S SFF Gen2-bay Drawer (#5887) is an expansion drawer with twenty-four 2.5-inch form-factor SAS bays. The EXP24S supports up to 24 hot-swap SFF-2 SAS hard disk drives (HDDs) or solid-state drives (SSDs). It uses only 2 EIA of space in a 19-inch rack. The EXP24S includes redundant ac power supplies and uses two power cords.

To maximize configuration flexibility and space utilization, the system node of Power E980 system does not have integrated SAS bays or integrated SAS controllers. PCIe SAS adapters and the EXP24S can be used to provide direct access storage.

To further reduce possible single points of failure, EXP24S configuration rules consistent with previous Power Systems are used. IBM i configurations require the drives to be protected (RAID or mirroring). Protecting the drives is highly advised, but not required for other operating systems. All Power operating system environments that are using SAS adapters with write cache require the cache to be protected by using pairs of adapters.
With AIX, Linux, and VIOS, you can order the EXP24S with four sets of six bays, two sets of 12 bays, or one set of 24 bays (mode 4, 2, or 1). With IBM i, you can order the EXP24S as one set of 24 bays (mode 1). Figure 2-26 shows the front of the unit and the groups of disks on each mode.

Mode setting is done by IBM manufacturing. The stickers indicate whether the enclosure is set to mode 1, mode 2, or mode 4. They are attached to the lower-left shelf of the chassis (A) and the center support between the enclosure services manager modules (B).
Figure 2-27 shows the mode stickers.

The EXP24S SAS ports are attached to a SAS PCIe adapter or pair of adapters using SAS YO or X cables. Cable length varies depending on the feature code, and proper length should be calculated considering routing for proper airflow and ease of handling. A diagram of both types of SAS cables can be seen in Figure 2-28.
The following SAS adapters support the EXP24S:
- PCIe 380MB Cache Dual-x4 3Gb SAS RAID Adapter (#5805)
- PCIe Dual-x4 SAS Adapter (#EJ0J)
- PCIe2 1.8GB Cache RAID SAS Adapter Tri-port 6Gb (#5913)
- PCIe2 1.8GB Cache RAID SAS Adapter Tri-port 6Gb CR (#ESA3)
- PCIe3 RAID SAS Adapter Quad-port 6Gb x8 (#EJ0J)
- PCIe3 12GB Cache RAID SAS Adapter Quad-port 6Gb x8 (#EJ0L)
- PCIe3 12GB Cache RAID Plus SAS Adapter Quad-port 6Gb x8 (#EJ14)
- PCIe3 LP RAID SAS ADAPTER (#EJ0M)

The EXP24S drawer can support up to 24 SAS SFF Gen-2 disks. Table 2-24 lists the available disk options.

### Table 2-24  Available disks for the EXP24S

<table>
<thead>
<tr>
<th>Feature Code</th>
<th>CCIN</th>
<th>Description</th>
<th>OS support</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES0G</td>
<td>775 GB SFF-2 SSD for AIX/Linux</td>
<td>AIX, Linux</td>
<td></td>
</tr>
<tr>
<td>ES0H</td>
<td>775 GB SFF-2 SSD for IBM i</td>
<td>IBM i</td>
<td></td>
</tr>
<tr>
<td>ES0Q</td>
<td>387 GB SFF-2 4 K SSD for AIX/Linux</td>
<td>AIX, Linux</td>
<td></td>
</tr>
<tr>
<td>ES0R</td>
<td>387 GB SFF-2 4 K SSD for IBM i</td>
<td>IBM i</td>
<td></td>
</tr>
<tr>
<td>ES0S</td>
<td>775 GB SFF-2 4 K SSD for AIX/Linux</td>
<td>AIX, Linux</td>
<td></td>
</tr>
<tr>
<td>ES0T</td>
<td>775 GB SFF-2 4 K SSD for IBM i</td>
<td>IBM i</td>
<td></td>
</tr>
<tr>
<td>ES19</td>
<td>387 GB SFF-2 SSD for AIX/Linux</td>
<td>AIX, Linux</td>
<td></td>
</tr>
<tr>
<td>ES1A</td>
<td>387 GB SFF-2 SSD for IBM i</td>
<td>IBM i</td>
<td></td>
</tr>
<tr>
<td>ES2C</td>
<td>387 GB SFF-2 SSD for AIX/Linux</td>
<td>AIX, Linux</td>
<td></td>
</tr>
<tr>
<td>ES2D</td>
<td>387 GB SFF-2 SSD for IBM i</td>
<td>IBM i</td>
<td></td>
</tr>
<tr>
<td>ES62</td>
<td>3.86-4.0 TB 7200 RPM 4K SAS LFF-1 Nearline Disk Drive (AIX/Linux)</td>
<td>AIX, Linux</td>
<td></td>
</tr>
<tr>
<td>ES64</td>
<td>7.72-8.0 TB 7200 RPM 4K SAS LFF-1 Nearline Disk Drive (AIX/Linux)</td>
<td>AIX, Linux</td>
<td></td>
</tr>
<tr>
<td>ES78</td>
<td>387 GB SFF-2 SSD 5xx eMLC4 for AIX/Linux</td>
<td>AIX, Linux</td>
<td></td>
</tr>
<tr>
<td>ES79</td>
<td>387 GB SFF-2 SSD 5xx eMLC4 for IBM i</td>
<td>IBM i</td>
<td></td>
</tr>
<tr>
<td>ES7E</td>
<td>775 GB SFF-2 SSD 5xx eMLC4 for AIX/Linux</td>
<td>AIX, Linux</td>
<td></td>
</tr>
<tr>
<td>ES7F</td>
<td>775 GB SFF-2 SSD 5xx eMLC4 for IBM i</td>
<td>IBM i</td>
<td></td>
</tr>
<tr>
<td>ES80</td>
<td>1.9 TB Read Intensive SAS 4k SFF-2 SSD for AIX/Linux</td>
<td>AIX, Linux</td>
<td></td>
</tr>
<tr>
<td>ES81</td>
<td>1.9 TB Read Intensive SAS 4k SFF-2 SSD for IBM i</td>
<td>IBM i</td>
<td></td>
</tr>
<tr>
<td>ES85</td>
<td>387 GB SFF-2 SSD 4k eMLC4 for AIX/Linux</td>
<td>AIX, Linux</td>
<td></td>
</tr>
<tr>
<td>ES86</td>
<td>387 GB SFF-2 SSD 4k eMLC4 for IBM i</td>
<td>IBM i</td>
<td></td>
</tr>
<tr>
<td>ES8C</td>
<td>775 GB SFF-2 SSD 4k eMLC4 for AIX/Linux</td>
<td>AIX, Linux</td>
<td></td>
</tr>
<tr>
<td>Feature Code</td>
<td>CCIN</td>
<td>Description</td>
<td>OS support</td>
</tr>
<tr>
<td>--------------</td>
<td>------</td>
<td>-------------</td>
<td>------------</td>
</tr>
<tr>
<td>ES8D</td>
<td>77G</td>
<td>775 GB SFF-2 SSD 4k eMLC4 for IBM i</td>
<td>IBM i</td>
</tr>
<tr>
<td>ES8F</td>
<td>1.55</td>
<td>1.55 TB SFF-2 SSD 4k eMLC4 for AIX/Linux</td>
<td>AIX, Linux</td>
</tr>
<tr>
<td>ES8G</td>
<td>1.55</td>
<td>1.55 TB SFF-2 SSD 4k eMLC4 for IBM i</td>
<td>IBM i</td>
</tr>
<tr>
<td>1738</td>
<td>19B4</td>
<td>856 GB 10 K RPM SAS SFF-2 Disk Drive (IBM i)</td>
<td>IBM i</td>
</tr>
<tr>
<td>1752</td>
<td>900</td>
<td>900 GB 10 K RPM SAS SFF-2 Disk Drive (AIX/Linux)</td>
<td>AIX, Linux</td>
</tr>
<tr>
<td>1948</td>
<td>19B1</td>
<td>283 GB 15 K RPM SAS SFF-2 Disk Drive (IBM i)</td>
<td>IBM i</td>
</tr>
<tr>
<td>1953</td>
<td>300</td>
<td>300 GB 15 K RPM SAS SFF-2 Disk Drive (AIX/Linux)</td>
<td>AIX, Linux</td>
</tr>
<tr>
<td>1962</td>
<td>19B3</td>
<td>571 GB 10 K RPM SAS SFF-2 Disk Drive (IBM i)</td>
<td>IBM i</td>
</tr>
<tr>
<td>1964</td>
<td>600</td>
<td>600 GB 10 K RPM SAS SFF-2 Disk Drive (AIX/Linux)</td>
<td>AIX, Linux</td>
</tr>
<tr>
<td>ESD2</td>
<td>59CD</td>
<td>1.1 TB 10 K RPM SAS SFF-2 Disk Drive (IBM i)</td>
<td>IBM i</td>
</tr>
<tr>
<td>ESD3</td>
<td>0</td>
<td>1.2 TB 10 K RPM SAS SFF-2 Disk Drive (AIX/Linux)</td>
<td>AIX, Linux</td>
</tr>
<tr>
<td>ESDN</td>
<td>571</td>
<td>571 GB 15 K RPM SAS SFF-2 Disk Drive - 528 Block (IBM i)</td>
<td>IBM i</td>
</tr>
<tr>
<td>ESDP</td>
<td>0</td>
<td>600 GB 15 K RPM SAS SFF-2 Disk Drive - 5xx Block (AIX/Linux)</td>
<td>AIX, Linux</td>
</tr>
<tr>
<td>ESEU</td>
<td>59D2</td>
<td>571 GB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4224</td>
<td>IBM i</td>
</tr>
<tr>
<td>ESEV</td>
<td>0</td>
<td>600 GB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4096</td>
<td>AIX, Linux</td>
</tr>
<tr>
<td>ESEY</td>
<td>0</td>
<td>283 GB 15 K RPM SAS SFF-2 4 K Block - 4224 Disk Drive</td>
<td>IBM i</td>
</tr>
<tr>
<td>ESEZ</td>
<td>0</td>
<td>300 GB 15 K RPM SAS SFF-2 4 K Block - 4096 Disk Drive</td>
<td>AIX, Linux</td>
</tr>
<tr>
<td>ESF2</td>
<td>58DA</td>
<td>1.1 TB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4224</td>
<td>IBM i</td>
</tr>
<tr>
<td>ESF3</td>
<td>0</td>
<td>1.2 TB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4096</td>
<td>AIX, Linux</td>
</tr>
<tr>
<td>ESFN</td>
<td>0</td>
<td>571 GB 15 K RPM SAS SFF-2 4 K Block - 4224 Disk Drive</td>
<td>IBM i</td>
</tr>
<tr>
<td>ESFP</td>
<td>0</td>
<td>600 GB 15 K RPM SAS SFF-2 4 K Block - 4096 Disk Drive</td>
<td>AIX, Linux</td>
</tr>
<tr>
<td>ESFS</td>
<td>0</td>
<td>1.7 TB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4224</td>
<td>IBM i</td>
</tr>
<tr>
<td>ESFT</td>
<td>0</td>
<td>1.8 TB 10 K RPM SAS SFF-2 Disk Drive 4 K Block - 4096</td>
<td>AIX, Linux</td>
</tr>
</tbody>
</table>

There are six SAS connectors on the rear of the EXP24S drawer to which two SAS adapters or controllers are attached. They are labeled T1, T2, and T3; there are two T1, two T2, and two T3 connectors. While configuring the drawer, special configuration feature codes will indicate for the plant the mode of operation in which the disks and ports will be split:

- 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 2-29 shows the rear connectors of the EXP24S drawer, how they relate with the modes of operation, and disk grouping.

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 with any of the supported SAS adapters or controllers.

**EXP24S common usage scenarios**

The EXP24S drawer is very versatile in the ways that it can be attached to Power Systems. This section describes the most common usage scenarios for EXP24S and Virtual I/O Servers, using standard PCIe SAS adapters #EJ0J.

**Note:** Not all possible scenarios are included. Refer to the “Planning for serial-attached SCSI cables” guide in the IBM Knowledge Center to see more supported scenarios.
Scenario 1: Basic non-redundant connection
This scenario assumes a single Virtual I/O Server with a single PCIe SAS adapter #EJ0J and an EXP24S set on mode 1, allowing for up to 24 disks to be attached to the server. Figure 2-30 shows the connection diagram and components of the solution.

For this scenario, these are the required feature codes:
- One EXP24S drawer #5887 with mode 1
- One PCIe SAS adapter #EJ0J
- One SAS YO cable 3 Gbps with proper length
Scenario 2: Basic redundant connection

This scenario assumes a single Virtual I/O Server with two PCIe SAS adapters #EJ0J and an EXP24S set on mode 1, allowing for up to 24 disks to be attached to the server. Figure 2-31 shows the connection diagram and components of the solution.

![Scenario 2: Basic redundant connection diagram](image)

For this scenario, these are the required feature codes:

- One EXP24S drawer #5887 with mode 1
- Two PCIe SAS adapter #EJ0J
- Two SAS YO cables 3 Gbps with proper length

The ports used on the SAS adapters must be the same for both adapters of the pair. There is no SSD support on this scenario.
Scenario 3: Dual Virtual I/O Servers sharing a single EXP24S

This scenario assumes a dual Virtual I/O Server with two PCIe SAS adapters #EJ0J each and an EXP24S set on mode 2, allowing for up to 12 disks to be attached to the each Virtual I/O Server. Figure 2-32 shows the connection diagram and components of the solution.

For this scenario, these are the required feature codes:

- One EXP24S drawer #5887 with mode 2
- Four PCIe SAS adapter #EJ0J
- Two SAS X cables 3 Gbps with proper length

The ports used on the SAS adapters must be the same for both adapters of the pair. There is no SSD support on this scenario.
Scenario 4: Dual Virtual I/O Servers sharing two EXP24S

This scenario assumes a dual Virtual I/O Server with two PCIe SAS adapters #EJ0J each and two EXP24S set on mode 2, allowing for up to 24 disks to be attached to each Virtual I/O Server (2 per drawer). If compared to scenario 3, this scenario has the benefit to allow disks from different EXP24S drawers to be mirrored, allowing for hot maintenance of the whole EXP24S drawers if all data is properly mirrored. Figure 2-33 shows the connection diagram and components of the solution.

For this scenario, these are the required feature codes:

- Two EXP24S drawers #5887 with mode 2
- Four PCIe SAS adapter #EJ0J
- Four SAS YO cables 3 Gbps with proper length.

There is no SSD support on this scenario.
Scenario 5: Four Virtual I/O Servers sharing two EXP24S

This scenario assumes four Virtual I/O Servers with two PCIe SAS adapters #EJ0J each and two EXP24S set on mode 4, allowing for up to 12 disks to be attached to each Virtual I/O Server (6 per drawer). This scenario has the benefit to allow disks from different EXP24S drawers to be mirrored, allowing for hot maintenance of the whole EXP24S drawers if all data is properly mirrored. Figure 2-34 shows the connection diagram and components of the solution.

For this scenario, these are the required feature codes:
- Two EXP24S drawers #5887 with mode 4
- Eight PCIe SAS adapter #EJ0J
- Four SAS X cables 3 Gbps with proper length.

There is no SSD support on this scenario.

Other scenarios
For direct connection to logical partitions, different adapters, and cables, see “5887 disk drive enclosure” in the IBM Knowledge Center:
2.8.2 EXP24SX and EXP12SX SAS Storage Enclosure

The EXP24SX is a storage expansion enclosure with 24 2.5-inch small form factor (SFF) SAS bays. It supports Hard Disk Drives (HDD) or Solid State Drives (SSD). The EXP12SX is a storage expansion enclosure with twelve 3.5-inch large form factor (LFF) SAS bays. The EXP12SX supports HDD only.

The following PCIe3 SAS adapters support the EXP24SX and EXP 12SX:
- PCIe3 RAID SAS Adapter Quad-port 6 Gb x8 (FC #EJ0J)
- PCIe3 LP RAID SAS Adapter Quad-Port 6Gb x8 (FC #EJ0M)
- PCIe3 RAID SAS quad-port 6 Gb LP Adapter (FC #EL3B)
- PCIe3 12 GB Cache RAID Plus SAS Adapter Quad-port 6 Gb x8 (FC #EJ14)

IBM i configurations require the drives to be protected (RAID or mirroring). Protecting the drives is highly advised, but not required for other operating systems. All Power operating system environments that are using SAS adapters with write cache require the cache to be protected by using pairs of adapters.

The EXP24SX and EXP12SX drawers have many high-reliability design points:
- SAS bays that support hot-swap.
- Redundant and hot-plug power and fan assemblies.
- Dual power cords.
- Redundant and hot-plug ESMs.
- Redundant data paths to all drives.
- LED indicators on drives, bays, ESMs, and power supplies that support problem identification.
- Through the SAS adapters/controllers, drives that can be protected with RAID and mirroring and hot-spare capability.

Note:
- For the EXP24SX, A maximum of 24 2.5-inch SSDs or 2.5-inch HDDs are supported in the ESLS 24 SAS bays. There can be no mixing of HDDs and SSDs in the same mode-1 drawer. HDDs and SSDs can be mixed in a mode-2 or mode-4 drawer, but they cannot be mixed within a logical split of the drawer. For example, in a mode-2 drawer with two sets of 12 bays, one set could hold SSDs and one set could hold HDDs, but you cannot mix SSDs and HDDs in the same set of 12-bays.
- The EXP24S, EXP24SX, and EXP12SX can be mixed on the same server and on the same PCIe3 adapter/s.
- The EXP12SX does not support SSD.

The cables that are used to connect an ESLL or ESLS storage enclosure to a server are different from the cables that are used with the 5887 disk drive enclosure. Attachment between the SAS controller and the storage enclosure SAS ports is through the appropriate SAS YO12 or X12 cables. The PCIe Gen3 SAS adapters support 6Gb throughput. The EXP12SX has been designed to support up to 12Gb throughput if future SAS adapters support that capability.

The cable options are:
There are six SAS connectors at the rear of the EXP24SX and EXP12SX 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.

- 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 a dual adapters configuration.
- In mode 2 or mode 4, four ports are used, two T2s and two T3s, to access all SAS bays.

Figure 2-35 shows connector locations for the EXP24SX and EXP12SX storage enclosures.

Mode setting is done by IBM manufacturing. If you need to change the mode after installation, ask your IBM support representative to refer to:

http://w3-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD106418

For more information about SAS cabling and cabling configurations, see Connecting an ESSL or ELSLS storage enclosure to your system in the IBM Knowledge Center:
2.8.3 IBM System Storage

The IBM System Storage Disk Systems products and offerings provide compelling storage solutions with superior value for all levels of business, from entry-level to high-end storage systems. For more information about the various offerings, see the following web site:

http://www.ibm.com/systems/storage/disk

The following section highlights a few of the offerings.

IBM Storwize family
The IBM Storwize® is part of the IBM Spectrum™ Virtualize family and is the ideal solution to optimize the data architecture for business flexibility and data storage efficiency. Different models, such as the IBM Storwize V3700, IBM Storwize V5000, and IBM Storwize V7000, offer storage virtualization, IBM Real-time Compression, Easy Tier®, and many more functions. For more information, see the following web site:

http://www.ibm.com/systems/storage/storwize

IBM FlashSystem family
The IBM FlashSystem® family delivers extreme performance to derive measurable economic value across the data architecture (servers, software, applications, and storage). IBM offers a comprehensive flash portfolio with the IBM FlashSystem family. For more information, see the following web site:

http://www.ibm.com/systems/storage/flash

IBM XIV Storage System
The IBM XIV® Storage System hardware is part of the Spectrum Accelerate family and is a high-end disk storage system, helping thousands of enterprises meet the challenge of data growth with hotspot-free performance and ease of use. Simple scaling, high service levels for dynamic, heterogeneous workloads, and tight integration with hypervisors and the OpenStack platform enable optimal storage agility for cloud environments.

XIV Storage Systems extend ease of use with integrated management for large and multi-site XIV deployments, reducing operational complexity and enhancing capacity planning. For more information, see the following web site:


IBM System Storage DS8000
The IBM System Storage DS8000® storage subsystem is a high-performance, high-capacity, and secure storage system that is designed to deliver the highest levels of performance, flexibility, scalability, resiliency, and total overall value for the most demanding, heterogeneous storage environments. The system is designed to manage a broad scope of storage workloads that exist in today’s complex data center, doing it effectively and efficiently.

Additionally, the IBM System Storage DS8000 includes a range of features that automate performance optimization and application quality of service, and also provide the highest levels of reliability and system uptime. For more information, see the following web site:

2.9 Operating system support

The Power E980 server supports the following operating systems:

- AIX
- IBM i
- Linux

In addition, the VIOS can be installed in special partitions that provide support to other partitions running AIX or Linux operating systems for using features such as virtualized I/O devices, PowerVM Live Partition Mobility, or PowerVM Active Memory Sharing.

For more information about the software that is available on IBM Power Systems, visit the IBM Power Systems Software™ web site:

https://www.ibm.com/it-infrastructure/power/os

2.9.1 AIX operating system

The following sections describe the various levels of AIX operating system support.

IBM periodically releases maintenance packages (service packs or technology levels) for the AIX operating system. Information about these packages, downloading, and obtaining the CD-ROM is at the Fix Central website:

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

The Fix Central web site also provides information about how to obtain the fixes that are included on CD-ROM.

The Service Update Management Assistant (SUMA), which can help you automate the task of checking and downloading operating system downloads, is part of the base operating system. For more information about the `suma` command, go to the following web site:


Table 2-25 shows minimum supported AIX levels when using any I/O configuration.

Table 2-25 Supported minimum AIX levels for any I/O

<table>
<thead>
<tr>
<th>Version</th>
<th>Technology level</th>
<th>Service pack</th>
<th>Planned availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.2</td>
<td>3</td>
<td>7</td>
<td>7 August 2018</td>
</tr>
<tr>
<td>7.2</td>
<td>2</td>
<td>3</td>
<td>January 2019</td>
</tr>
<tr>
<td>7.2</td>
<td>1</td>
<td>5</td>
<td>14 December 2018</td>
</tr>
<tr>
<td>7.1</td>
<td>5</td>
<td>3</td>
<td>7 August 2018</td>
</tr>
<tr>
<td>7.1</td>
<td>4</td>
<td>7</td>
<td>January 2019</td>
</tr>
<tr>
<td>6.1a</td>
<td>9</td>
<td>12</td>
<td>7 August 2018</td>
</tr>
</tbody>
</table>

a. AIX 6.1 service extension required

Table 2-26 shows minimum supported AIX levels when using virtual I/O only.
Table 2-26  Supported minimum AIX levels for virtual I/O only

<table>
<thead>
<tr>
<th>Version</th>
<th>Technology level</th>
<th>Service pack</th>
<th>Planned availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.2</td>
<td>2</td>
<td>1</td>
<td>7 August 2018</td>
</tr>
<tr>
<td>7.2</td>
<td>1</td>
<td>1</td>
<td>7 August 2018</td>
</tr>
<tr>
<td>7.2</td>
<td>0</td>
<td>2</td>
<td>7 August 2018</td>
</tr>
<tr>
<td>7.1</td>
<td>5</td>
<td>1</td>
<td>7 August 2018</td>
</tr>
<tr>
<td>7.1</td>
<td>4</td>
<td>2</td>
<td>7 August 2018</td>
</tr>
<tr>
<td>6.1a</td>
<td>9</td>
<td>7</td>
<td>7 August 2018</td>
</tr>
</tbody>
</table>

a. AIX 6.1 service extension required

See the IBM Prerequisite web site for compatibility information for hardware features and the corresponding AIX Technology Levels:


2.9.2 IBM i

IBM i is supported on the Power E980 server with the following minimum required levels:

- IBM i 7.2 TR9, or later
- IBM i 7.3 TR5, or later

See the IBM Prerequisite web site for compatibility information for hardware features and the corresponding IBM i Technology Levels:


2.9.3 Linux operating system

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

The supported versions of Linux on the Power E980 server is as follows:

- If installing the Linux operating system (one of these):
  - Red Hat Enterprise Linux 7.5 for Power LE (p8compat), or later
  - SUSE Linux Enterprise Server 12 Service Pack 3, or later
  - SUSE Linux Enterprise Server for SAP with SUSE Linux Enterprise Server 12 Service Pack 3, or later
  - SUSE Linux Enterprise Server for SAP with SUSE Linux Enterprise Server 11 Service Pack 4
  - SUSE Linux Enterprise Server 15
  - Ubuntu 16.04.4

Hardware service diagnostic aids and productivity tools, as well as developer toolkits for Linux operating systems running on IBM servers based on POWER technology. Service and productivity tools are available in a YUM repository that you can use to download, and then
install, all recommended packages for your Red Hat, SUSE Linux or Fedora distribution. It is available from IBM at the following web site:

Learn about developing on the Power architecture, find packages, get access to cloud resources, discover the tools and technologies, see the Linux on IBM Power Systems Developer Portal:
https://developer.ibm.com/linuxonpower/

The IBM Advance Toolchain for Linux on Power is a set of open source compilers, runtime libraries, and development tools which allows users to take leading edge advantage of IBM's latest POWER hardware features on Linux. See the following web site:

For more information about SUSE Linux Enterprise Server, see the following web site:
https://www.suse.com/products/server/

For more information about Red Hat Enterprise Linux, see the following web site:

2.9.4 Virtual I/O Server

The minimum required level of VIOS for the Power E980 is VIOS 2.2.6.31 or later.

IBM regularly updates the VIOS code. To find information about the latest updates, visit the Fix Central web site:
https://www-945.ibm.com/support/fixcentral/
Chapter 3. Virtualization

Virtualization is a key factor for productive and efficient use of IBM Power Systems servers. In this chapter, you find a brief description of virtualization technologies that are available for POWER9. The following IBM Redbooks publications provide more information about the virtualization features.

- *IBM PowerVM Best Practices*, SG24-8062
- *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940
- *IBM PowerVM Virtualization Active Memory Sharing*, REDP-4470
- *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590
3.1 IBM POWER Hypervisor

Power Systems servers that are combined with PowerVM technology offer key capabilities that can help you consolidate and simplify your IT environment:

- Improve server usage and share I/O resources to reduce the total cost of ownership (TCO) and better use IT assets.
- Improve business responsiveness and operational speed by dynamically reallocating resources to applications as needed to better match changing business needs or handle unexpected changes in demand.
- Simplify IT infrastructure management by making workloads independent of hardware resources so that you can make business-driven policies to deliver resources that are based on time, cost, and service-level requirements.

Combined with features in the POWER9 processors, the IBM POWER Hypervisor™ delivers functions that enable other system technologies, including logical partitioning (LPAR) technology, virtualized processors, IEEE VLAN-compatible virtual switch, virtual SCSI adapters, virtual Fibre Channel adapters, and virtual consoles. The POWER Hypervisor is a basic component of the system’s firmware and offers the following functions:

- Provides an abstraction between the physical hardware resources and the LPARs that use them.
- Enforces partition integrity by providing a security layer between LPARs.
- Controls the dispatch of virtual processors to physical processors.
- Saves and restores all processor state information during a logical processor context switch.
- Controls hardware I/O interrupt management facilities for LPARs.
- Provides virtual LAN channels between LPARs that help reduce the need for physical Ethernet adapters for inter-partition communication.
- Monitors the service processor and performs a reset or reload if it detects the loss of the service processor, notifying the operating system if the problem is not corrected.

The POWER Hypervisor is always active, regardless of the system configuration and also when not connected to the managed console. It requires memory to support the resource assignment to the LPARs on the server. The amount of memory that is required by the POWER Hypervisor firmware varies according to several factors:

- Memory that is required for hardware page tables (HPT).
- Memory that is required to support I/O devices.
- Memory that is required for virtualization.

**Memory that required for hardware page tables**

Each partition on the system has its own HPT that contributes to hypervisor memory usage. The HPT is used by the operating system to translate from effective addresses to physical real addresses in the hardware. This translation from effective to real addresses allows multiple operating systems to run simultaneously in their own logical address space. Whenever a virtual processor for a partition is dispatched on a physical processor, the hypervisor indicates to the hardware the location of the partition HPT that should be used when translating addresses.
The amount of memory for the HPT is based on the maximum memory size of the partition and the HPT ratio. The default HPT ratio is either 1/64th of the maximum (for IBM i partitions) or 1/128th (for AIX, Virtual I/O Server (VIOS), and Linux partitions) of the maximum memory size of the partition. AIX, VIOS, and Linux use larger page sizes (16 and 64 KB) instead of using 4 KB pages. Using larger page sizes reduces the overall number of pages that must be tracked, so the overall size of the HPT can be reduced. As an example, for an AIX partition with a maximum memory size of 256 GB, the HPT would be 2 GB.

When defining a partition, the maximum memory size that is specified should be based on the amount of memory that can be dynamically added to the dynamic partition (DLPAR) without having to change the configuration and restart the partition.

In addition to setting the maximum memory size, the HPT ratio can also be configured. The hpt_ratio parameter for the chsyscfg Hardware Management Console (HMC) command can be issued to define the HPT ratio that is used for a partition profile. The valid values are 1:32, 1:64, 1:128, 1:256, or 1:512. Specifying a smaller absolute ratio (1/512 is the smallest value) decreases the overall memory that is assigned to the HPT. Testing is required when changing the HPT ratio because a smaller HPT might incur more CPU consumption because the operating system might need to reload the entries in the HPT more frequently. Most customers have chosen to use the IBM provided default values for the HPT ratios.

Memory usage for I/O devices
In support of I/O operations, the hypervisor maintains structures that are called the Translation Control Entries (TCEs), which provides an information path between I/O devices and partitions. The TCEs provide the address of the I/O buffer, indication of read versus write requests, and other I/O-related attributes. There are many TCEs in use per I/O device, so multiple requests can be active simultaneous to the same physical device. To provide better affinity, the TCEs are spread across multiple processor chips or drawers to improve performance while accessing the TCEs. For physical I/O devices, the base amount of space for the TCEs is defined by the hypervisor based on the number of I/O devices that are supported. A system that supports high-speed adapters can also be configured to allocate more memory to improve I/O performance. Linux is the only operating system that uses these additional TCEs so that the memory can be freed for use by partitions if the system is using only AIX or IBM i.

Memory usage for virtualization features
Virtualization requires more memory to be allocated by the POWER hypervisor for hardware statesave areas and various virtualization technologies. For example, on POWER9 systems, each processor core supports up to eight simultaneous multithreading (SMT) threads of execution, and each thread contains over 80 different registers. The POWER hypervisor must set aside save areas for the register contents for the maximum number of virtual processors that is configured. The greater the number of physical hardware devices, the greater the number of virtual devices, the greater the amount of virtualization, and the more hypervisor memory is required. For efficient memory consumption, wanted and maximum values for various attributes (processors, memory, and virtual adapters) should be based on business needs and not set to values that are significantly higher than actual requirements.

Predicting memory that is used by the POWER hypervisor
The IBM System Planning Tool (SPT) is a resource that can be used to estimate the amount of hypervisor memory that is required for a specific server configuration. After the SPT executable file is downloaded and installed, you can define a configuration by selecting the correct hardware platform, selecting the installed processors and memory, and defining partitions and partition attributes. SPT can estimate the amount of memory that will be
assigned to the hypervisor, which assists you when you change an existing configuration or deploy new servers.

The POWER Hypervisor provides the following types of virtual I/O adapters:

- Virtual SCSI
- Virtual Ethernet
- Virtual Fibre Channel
- Virtual (TTY) console

### 3.1.1 Virtual SCSI

The POWER Hypervisor provides a virtual SCSI mechanism for the virtualization of storage devices. The storage virtualization is accomplished by using two paired adapters: a virtual SCSI server adapter and a virtual SCSI client adapter.

### 3.1.2 Virtual Ethernet

The POWER Hypervisor provides a virtual Ethernet switch function that allows partitions either fast and secure communication on the same server without any need for physical interconnection or connectivity outside of the server if a Layer 2 bridge to a physical Ethernet adapter is set in one VIOS partition, also known as Shared Ethernet Adapter (SEA).

### 3.1.3 Virtual Fibre Channel

A virtual Fibre Channel adapter is a virtual adapter that provides client LPARs with a Fibre Channel connection to a storage area network through the VIOS partition. The VIOS partition provides the connection between the virtual Fibre Channel adapters on the VIOS partition and the physical Fibre Channel adapters on the managed system.

### 3.1.4 Virtual (TTY) console

Each partition must have access to a system console. Tasks such as operating system installation, network setup, and various problem analysis activities require a dedicated system console. The POWER Hypervisor provides the virtual console by using a virtual TTY or serial adapter and a set of Hypervisor calls to operate on them. Virtual TTY does not require the purchase of any additional features or software, such as the PowerVM Edition features.
3.2 POWER processor modes

Although they are not virtualization features, the POWER processor modes are described here because they affect various virtualization features.

On Power System servers, partitions can be configured to run in several modes, including the following modes:

- **POWER7 compatibility mode**
  
  This is the mode for POWER7+ and POWER7 processors, implementing Version 2.06 of the IBM Power Instruction Set Architecture (ISA). For more information, see the following website:
  

- **POWER8 compatibility mode**
  
  This is the native mode for POWER8 processors implementing Version 2.07 of the IBM Power ISA. For more information, see the following address:
  

- **POWER9 compatibility mode**
  
  This is the native mode for POWER9 processors implementing Version 3.0 of the IBM Power ISA. For more information, see the following address:
  
  https://bit.ly/2s1Lbfr
Figure 3-1 shows the available processor modes on a POWER9 system.

Processor compatibility mode is important when Live Partition Mobility (LPM) migration is planned between different generation of servers. An LPAR that potentially might be migrated to a machine that is managed by a processor from another generation must be activated in a specific compatibility mode.
Table 3-1 shows an example where the processor mode must be selected when a migration from POWER9 to POWER8 is planned.

<table>
<thead>
<tr>
<th>Source environment (POWER9 server)</th>
<th>Destination environment (POWER8 server)</th>
<th>Processor compatibility modes for a POWER9 to POWER8 migration</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Active migration</strong></td>
<td><strong>Inactive migration</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Wanted processor compatibility mode</strong></td>
<td><strong>Current processor compatibility mode</strong></td>
<td><strong>Wanted processor compatibility mode</strong></td>
</tr>
<tr>
<td>POWER9</td>
<td>POWER9</td>
<td>Fails because the wanted processor mode is not supported on destination.</td>
</tr>
<tr>
<td>POWER9</td>
<td>POWER8</td>
<td>Fails because the wanted processor mode is not supported on destination.</td>
</tr>
<tr>
<td>Default</td>
<td>POWER9</td>
<td>Fails because the wanted processor mode is not supported on destination.</td>
</tr>
<tr>
<td>POWER8</td>
<td>POWER8</td>
<td>POWER8</td>
</tr>
<tr>
<td>Default</td>
<td>POWER8</td>
<td>Default</td>
</tr>
<tr>
<td>POWER7</td>
<td>POWER7</td>
<td>POWER7</td>
</tr>
</tbody>
</table>

**3.3 Active Memory Expansion**

Active Memory Expansion is an optional feature for the E950 servers.

This feature enables memory expansion on the system. By using compression and decompression of memory, content can effectively expand the maximum memory capacity, providing more server workload capacity and performance.

Active Memory Expansion is a technology that allows the effective maximum memory capacity to be much larger than the true physical memory maximum. Compression and decompression of memory content can allow memory expansion up to 1000% for AIX partitions, which in turn enables a partition to perform more work or support more users with the same physical amount of memory. Similarly, it can allow a server to run more partitions and do more work for the same physical amount of memory.

**Note:** The Active Memory Expansion feature is not supported by IBM i and the Linux operating system.
3.4 Single Root I/O Virtualization

Single Root I/O Virtualization (SR-IOV) is an extension to the Peripheral Component Interconnect Express (PCIe) specification that allows multiple operating systems to simultaneously share a PCIe adapter with little or no runtime involvement from a hypervisor or other virtualization intermediary.

SR-IOV is PCI standard architecture that enables PCIe adapters to become self-virtualizing. It enables adapter consolidation, through sharing, much like logical partitioning enables server consolidation. With an adapter capable of SR-IOV, you can assign virtual slices of a single physical adapter to multiple partitions through logical ports; all of this is done without a VIOS.

POWER9 provides the following SR-IOV enhancements:
- Faster ports: 10 Gb, 25 Gb, 40 Gb, and 100 Gb
- More VFs per port: Sixty VFs per port (120 VFs per adapter) for 100-Gb adapters
- vNIC and vNIC failover support for Linux

Here are the hardware requirements to enable SR-IOV:
- PCIe2 4-port (10-Gb Fibre Channel over Ethernet (FCoE) and 1 GbE) SR & RJ45 Adapter (FC #EN0H)
- PCIe2 4-port (10-Gb FCoE and 1 GbE) SFP+Copper and RJ4 Adapter (FC #EN0K)
- PCIe3 4-port 10GbE SR Adapter (FC #EN15)

For more information, see IBM Power Systems SR-IOV: Technical Overview and Introduction, REDP-5065.

3.5 PowerVM

The PowerVM platform is the family of technologies, capabilities, and offerings that delivers industry-leading virtualization on IBM Power Systems servers. It is the umbrella branding term for Power Systems virtualization (logical partitioning, IBM Micro-Partitioning®, POWER Hypervisor, VIOS, LPM, and more). As with Advanced Power Virtualization in the past, PowerVM is a combination of hardware enablement and software.

Note: PowerVM Enterprise Edition License Entitlement is now included with each Power E980. PowerVM Enterprise Edition is available as a hardware feature (#EPVV) and supports up to 20 partitions per core, VIOS, and multiple shared processor pools. It also offers Live Partition Mobility, Active Memory Sharing, and PowerVP™ performance monitoring.

Logical partitions

LPARs and virtualization increase the usage of system resources and add a level of configuration possibilities.

Logical partitioning is the ability to make a server that is run as though it were two or more independent servers. When you logically partition a server, you divide the resources on the server into subsets called LPARs. You can install software on an LPAR, and the LPAR runs as an independent logical server with the resources that you have allocated to the LPAR. LPAR is the equivalent of a virtual machine (VM).
You can assign processors, memory, and input/output devices to LPARs. You can run AIX, IBM i, Linux, and the VIOS in LPARs. The VIOS provides virtual I/O resources to other LPARs with general-purpose operating systems.

LPARs share a few system attributes, such as the system serial number, system model, and processor FC. All other system attributes can vary from one LPAR to another.

**Micro-Partitioning**

When you use the Micro-Partitioning technology, you can allocate fractions of processors to an LPAR. An LPAR that uses fractions of processors is also known as a *shared processor partition* or *micropartition*. Micropartitions run over a set of processors that is called a *shared processor pool*, and virtual processors are used to let the operating system manage the fractions of processing power that are assigned to the LPAR. From an operating system perspective, a virtual processor cannot be distinguished from a physical processor, unless the operating system is enhanced to determine the difference. Physical processors are abstracted into virtual processors that are available to partitions.

On the POWER9 processors, a partition can be defined with a processor capacity as small as 0.05 processing units. This number represents 0.05 of a physical core. Each physical core can be shared by up to 20 shared processor partitions, and the partition's entitlement can be incremented fractionally by as little as 0.01 of the processor. The shared processor partitions are dispatched and time-sliced on the physical processors under the control of the POWER Hypervisor. The shared processor partitions are created and managed by the HMC.

The Power S914 server supports up to eight cores in a single system. Here are the maximum numbers:
- Eight dedicated partitions
- 160 micropartitions (maximum of 20 micropartitions per physical active core)

The Power S922 server supports up to 20 cores in a single system. Here are the maximum numbers:
- 20 dedicated partitions
- 400 micropartitions (maximum of 20 micropartitions per physical active core)

The Power S924 server supports up to 24 cores in a single system. Here are the maximum numbers:
- 24 dedicated partitions
- 480 micropartitions (maximum of 20 micropartitions per physical active core)

The maximum amounts are supported by the hardware, but the practical limits depend on application workload demands.

**Processing mode**

When you create an LPAR, you can assign entire processors for dedicated use, or you can assign partial processing units from a shared processor pool. This setting defines the processing mode of the LPAR.

**Dedicated mode**

In dedicated mode, physical processors are assigned as a whole to partitions. The SMT feature in the POWER9 processor core allows the core to run instructions from two, four, or eight independent software threads simultaneously.
Shared dedicated mode
On POWER9 processor technology-based servers, you can configure dedicated partitions to become processor donors for idle processors that they own, allowing for the donation of spare CPU cycles from dedicated processor partitions to a shared processor pool. The dedicated partition maintains absolute priority for dedicated CPU cycles. Enabling this feature can help increase system usage without compromising the computing power for critical workloads in a dedicated processor.

Shared mode
In shared mode, LPARs use virtual processors to access fractions of physical processors. Shared partitions can define any number of virtual processors (the maximum number is 20 times the number of processing units that are assigned to the partition). The POWER Hypervisor dispatches virtual processors to physical processors according to the partition's processing units entitlement. One processing unit represents one physical processor's processing capacity. All partitions receive a total CPU time equal to their processing unit's entitlement. The logical processors are defined on top of virtual processors. So, even with a virtual processor, the concept of a logical processor exists, and the number of logical processors depends on whether SMT is turned on or off.

3.5.1 Multiple shared processor pools

Multiple shared processor pools (MSPPs) are supported on POWER9 processor-based servers. This capability allows a system administrator to create a set of micropartitions with the purpose of controlling the processor capacity that can be used from the physical shared processor pool.

Micropartitions are created and then identified as members of either the default processor pool or a user-defined shared processor pool. The virtual processors that exist within the set of micropartitions are monitored by the POWER Hypervisor, and processor capacity is managed according to user-defined attributes.

If the Power Systems server is under heavy load, each micropartition within a shared processor pool is assured of its processor entitlement, plus any capacity that it might be allocated from the reserved pool capacity if the micropartition is uncapped.

If certain micropartitions in a shared processor pool do not use their capacity entitlement, the unused capacity is ceded and other uncapped micropartitions within the same shared processor pool are allocated the additional capacity according to their uncapped weighting. In this way, the entitled pool capacity of a shared processor pool is distributed to the set of micropartitions within that shared processor pool.

All Power Systems servers that support the MSPPs capability have a minimum of one (the default) shared processor pool and up to a maximum of 64 shared processor pools.
### 3.5.2 Virtual I/O Server

The VIOS is part of PowerVM. It is a specific appliance that allows the sharing of physical resources between LPARs to allow more efficient usage (for example, consolidation). In this case, the VIOS owns the physical resources (SCSI, Fibre Channel, network adapters, or optical devices) and allows client partitions to share access to them, thus minimizing the number of physical adapters in the system. The VIOS eliminates the requirement that every partition owns a dedicated network adapter, disk adapter, and disk drive. The VIOS supports OpenSSH for secure remote logins. It also provides a firewall for limiting access by ports, network services, and IP addresses.

Figure 3-2 shows an overview of a VIOS configuration.

![Figure 3-2 Architectural view of the VIOS](image)

It is a preferred practice to run two VIOSes per physical server.

**Shared Ethernet Adapter**

A SEA can be used to connect a physical Ethernet network to a virtual Ethernet network. The SEA provides this access by connecting the POWER Hypervisor VLANs with the VLANs on the external switches. Because the SEA processes packets at Layer 2, the original MAC address and VLAN tags of the packet are visible to other systems on the physical network. IEEE 802.1 VLAN tagging is supported.

By using the SEA, several client partitions can share one physical adapter, and you can connect internal and external VLANs by using a physical adapter. The SEA service can be hosted only in the VIOS, not in a general-purpose AIX or Linux partition, and acts as a Layer 2 network bridge to securely transport network traffic between virtual Ethernet networks (internal) and one or more (Etherchannel) physical network adapters (external). These virtual Ethernet network adapters are defined by the POWER Hypervisor on the VIOS.

**Virtual SCSI**

Virtual SCSI is used to see a virtualized implementation of the SCSI protocol. Virtual SCSI is based on a client/server relationship. The VIOS LPAR owns the physical resources and acts as a server or, in SCSI terms, a target device. The client LPARs access the virtual SCSI backing storage devices that are provided by the VIOS as clients.
The virtual I/O adapters (virtual SCSI server adapter and a virtual SCSI client adapter) are configured by using a managed console or through the Integrated Virtualization Manager on smaller systems. The virtual SCSI server (target) adapter is responsible for running any SCSI commands that it receives. It is owned by the VIOS partition. The virtual SCSI client adapter allows a client partition to access physical SCSI and SAN-attached devices and LUNs that are assigned to the client partition. The provisioning of virtual disk resources is provided by the VIOS.

**N_Port ID Virtualization**

N_Port ID Virtualization (NPIV) is a technology that allows multiple LPARs to access independent physical storage through the same physical Fibre Channel adapter. This adapter is attached to a VIOS partition that acts only as a pass-through, managing the data transfer through the POWER Hypervisor.

Each partition has one or more virtual Fibre Channel adapters, each with their own pair of unique worldwide port names, enabling you to connect each partition to independent physical storage on a SAN. Unlike virtual SCSI, only the client partitions see the disk.

For more information and requirements for NPIV, see *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590.

**3.5.3 Live Partition Mobility**

LPM allows you to move a running LPAR from one system to another without disruption. Inactive partition mobility allows you to move a powered-off LPAR from one system to another one.

LPM provides systems management flexibility and improves system availability by:

- Avoiding planned outages for hardware upgrade or firmware maintenance.
- Avoiding unplanned downtime. With preventive failure management, if a server indicates a potential failure, you can move its LPARs to another server before the failure occurs.

For more information and requirements for NPIV, see *IBM PowerVM Live Partition Mobility*, SG24-7460.

**3.5.4 Active Memory Sharing**

Active Memory Sharing provides system memory virtualization capabilities, allowing multiple partitions to share a common pool of physical memory.

The physical memory of an IBM Power System can be assigned to multiple partitions in either dedicated or shared mode. A system administrator can assign some physical memory to a partition and some physical memory to a pool that is shared by other partitions. A single partition can have either dedicated or shared memory:

- With a pure dedicated memory model, the system administrator’s task is to optimize available memory distribution among partitions. When a partition suffers degradation because of memory constraints and other partitions have unused memory, the administrator can manually issue a dynamic memory reconfiguration.
- With a shared memory model, the system automatically decides the optimal distribution of the physical memory to partitions and adjusts the memory assignment based on partition load. The administrator reserves physical memory for the shared memory pool, assigns partitions to the pool, and provides access limits to the pool.
3.5.5 Active Memory Deduplication

In a virtualized environment, the systems might have a considerable amount of duplicated information that is stored on RAM after each partition has its own operating system, and some of them might even share kinds of applications. On heavily loaded systems, this behavior might lead to a shortage of the available memory resources, forcing paging by the Active Memory Sharing partition operating systems, the Active Memory Deduplication pool, or both, which might decrease overall system performance.

Active Memory Deduplication allows the POWER Hypervisor to map dynamically identical partition memory pages to a single physical memory page within a shared memory pool. This way enables a better usage of the Active Memory Sharing shared memory pool, increasing the system’s overall performance by avoiding paging. Deduplication can cause the hardware to incur fewer cache misses, which also leads to improved performance.

Active Memory Deduplication depends on the Active Memory Sharing feature being available, and it uses CPU cycles that are donated by the Active Memory Sharing pool's VIOS partitions to identify deduplicated pages. The operating systems that are running on the Active Memory Sharing partitions can suggest to the POWER Hypervisor that some pages (such as frequently referenced read-only code pages) are good for deduplication.

3.5.6 Remote Restart

Remote Restart is a high availability option for partitions. If there is an error that causes a server outage, a partition that is configured for Remote Restart can be restarted on a different physical server. At times, it might take longer to start the server, in which case the Remote Restart function can be used for faster reprovisioning of the partition. Typically, this can be done faster than restarting the server that failed and then restarting the partitions. The Remote Restart function relies on technology similar to LPM where a partition is configured with storage on a SAN that is shared (accessible) by the server that will host the partition.

The HMC version V9R2 brings following enhancements to the Remote Restart feature.

- Remote restart a partition with reduced or minimum CPU/memory on the target system.
- Remote restart by choosing a different virtual switch on the target system.
- Remote restart the partition without turning on the partition on the target system.
- Remote restart the partition for test purposes when the source-managed system is in the Operating or Standby state.
- Remote restart through the REST API.
Reliability, availability, serviceability, and manageability

This chapter provides information about reliability, availability, and serviceability (RAS) design and features of the IBM Power E980 system.

The elements of RAS can be described as follows:

**Reliability**
Indicates how infrequently a defect or fault in a server occurs

**Availability**
Indicates how infrequently the functioning of a system or application is impacted 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

The Reliability, availability, serviceability, and manageability chapter includes the following topics:

- 4.1, “Power E980 specific RAS enhancements” on page 130
- 4.2, “Reliability” on page 132
- 4.3, “Processor RAS details” on page 133
- 4.4, "Memory RAS details" on page 139
- 4.5, “PCIe I/O subsystem RAS details” on page 140
- 4.6, “Enterprise systems availability” on page 143
- 4.7, “Availability effects of a solution architecture” on page 144
- 4.8, “Serviceability” on page 147
- 4.9, “Manageability” on page 157
- 4.10, “Selected POWER9 RAS capabilities by operating system” on page 168
4.1 Power E980 specific RAS enhancements

The Power E980 system defines the most current implementation of a modular multi-node design introduced with the POWER5 processor-based IBM eServer™ p5 model 570 (9117-570) in 2004. The p5-570 system could already build on a long tradition of refined reliability, availability and serviceability (RAS) features. With each successor POWER processor generation additional RAS enhancements were released to eventually cumulate in the comprehensive RAS implementation of the Power E980 system. The Power E980 inherits the basic RAS design elements of the predecessor POWER8 processor-based Power E870, Power E870C, Power E880, and Power E880C servers as they are described by Daniel Henderson’s POWER8 Processor-Based Systems RAS white paper.

The remainder of this introductory section provides reference lists of strategic RAS features implemented in Power E980 servers. More detailed explanations for individual features are given in subsequent sections of this chapter and in the latest POWER9 RAS white paper by Daniel Henderson available at the following URL:

-- URL goes here --

The POWER9 processor-based Power E980 RAS enhancements which are unique to this enterprise class server are the following:

- Inter node SMP cable redundancy
- Eight bit wide (x8) DIMM ranks with Chipkill correction
- Custom DIMM (CDIMM) support with additional spare DRAM chips
- Asynchronous clocking across system nodes
- Redundant system clock with concurrent failover per system node
- Redundant flexible service processor (FSP) with concurrent failover
- Dedicated receptacles not on the FSP card itself but on the system control unit rear panel for FSP to system node interconnects improving the serviceability of the FSP card
- System control unit drawer with concurrent maintenance support for fans and with redundant electrical power supplied by up to two system nodes
- Voltage regulators with N+1 phase redundancy and integrated spare

RAS features which are specific to the enterprise class servers and are shared by the Power E980 and Power E950 systems are listed below:

- Core contained checkstops
- Extended L2/L3/L4 cache line delete
- IBM memory buffer support and spare DRAM module capacity with four bit wide (x4) DIMMs
- Memory row repair
- Active Memory Mirroring for Hypervisor
- Internal NVMe drive boot support
- Voltage regulators with N+1 phase redundancy
- Redundant/spare voltage phases on voltage converters for levels feeding processors, Power E980 memory CDIMMs, or Power E950 memory riser cards
- New concurrently maintainable operator panel design comprised of separate base and LED unit and a USB connection
- PCIe3 optical cable adapter with new routing for clock logic within the card as well as additional recovery procedures for faults during IPL
- Time of day (TOD) battery concurrent maintenance

The following section points out POWER9 processor base RAS features which are shared among all POWER9 processor-based systems and also covers important infrastructure
related RAS features which pertain to the Power E980 system but are shared with POWER9 scale-out servers Power S914, Power S922, Power S924, Power H922, and Power H924:

- Traditional POWER9 processor RAS features include first failure data capture, processor instruction retry, L2/L3 cache ECC protection with cache line-delete, power/cooling monitor function integrated into on chip controller (OCC)
- OCC error handling with power safe mode
- New POWER9 cyclic redundancy check (CRC) including retry capability and spare data lane support for the processor fabric bus
- Memory ECC with Chipkill handling
- Memory scrubbing
- Memory preserving IPL
- Dynamic memory relocation
- Enhanced Error Handling (EHH) for all adapters
- I/O adapter concurrent maintenance with PowerVM virtualization or operating system software based redundancy support
- Hot swap direct access storage devices (DASD)
- At least n+1 redundancy and concurrent maintenance support for power supplies and fans of each system node
- Power line cord redundancy
- Redundant vital product data (VPD)
- Emergency power-off (EPOW) reporting
- Concurrent firmware update

Table 4-1 provides a comparison between IBM POWER9 scale-out and POWER9 enterprise class systems in respect to significant RAS features:

<table>
<thead>
<tr>
<th>Base POWER9 Processor RAS features including</th>
<th>POWER9 1- and 2-socket systems(^a)</th>
<th>POWER9 IBM Power System E950</th>
<th>POWER9 IBM Power System E980</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Failure Data Capture</td>
<td>Yes(^b)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Processor Instruction Retry</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2/L3 Cache ECC protection with cache line-delete</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Power/cooling monitor function integrated into processors’ on chip controller</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRC checked processor fabric bus retry with spare data lane</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>POWER9 Enterprise RAS Features</th>
<th>No</th>
<th>Yes</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extended L2/L3 cache line delete</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Core Contained Checkstops</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>POWER9 Multi-node Enterprise RAS</th>
<th>N/A</th>
<th>N/A</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Across node (\frac{1}{2}) bandwidth capability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asynchronous clocking across nodes</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PCIe hot-plug with processor integrated PCIe controller</th>
<th>Yes</th>
<th>Yes</th>
<th>Yes</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Memory DIMM support with ECC checking supporting x4 Chipkill</th>
<th>Yes</th>
<th>Yes</th>
<th>Yes</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>IBM memory buffer support and Spare DRAM module capability with x4 DIMMS</th>
<th>No</th>
<th>Yes</th>
<th>Yes</th>
</tr>
</thead>
</table>
4.2 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 POWER9 systems, elements of all three are used to improve system reliability.

4.2.1 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 POWER9 chip supports a large number of cores per processor module and the PCIe Gen4 I/O controller function is integrated in the processor module, which generates a PCIe bus directly from the processor module. Additionally the POWER9 chip also provides compression and encryption functional units and the integrated circuit logic to attach external accelerators and devices through the coherent accelerator processor interface (CAPI), OpenCAPI and NVLink protocols.

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 4-1).

4.2.2 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 POWER9 processor chips, are positioned on the system board so that they receive clear air flow during operation. POWER9 systems use a premium fan with an extended life to further reduce overall system failure rate and provide adequate cooling for the critical system components.

4.3 Processor RAS details

The more reliable a system or subsystem is, the more available it should be. Nevertheless, considerable effort is made to design systems that can detect faults that do 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 error detection and fault isolation (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 built into each system. The dedicated service processor then works with the hardware to determine the course of action to be taken for each fault.

### 4.3.1 Correctable error introduction

Intermittent or soft errors are typically tolerated within the hardware design by using error correction code 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. There is also the risk 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 such situations. However, unconfiguring components, such as processor cores or entire caches in memory, can reduce the performance or capacity of a system. This in turn typically requires that the failing hardware is replaced in the system. The resulting service action can also temporarily impact system availability.

To avoid such situations in solid faults in POWER9, processors or memory might be candidates for correction by using the “self-healing” features 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 such 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.

### 4.3.2 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. The more advanced RAS design of POWER processor-based systems in some cases might be able to terminate just the application by using the hardware that failed. This RAS design requires that uncorrectable errors are detected by the hardware and reported to software layers, and the software layers is responsible for determining how to minimize the impact of faults.

One additional advantage of the special error correction code (ECC) used in data error detection is that the hardware is able to distinguish between an initial ECC error related to a specific component of the data path and one that has been passed along from earlier data.
transfer stages. This allows the correct component, the one originating the fault, to be reported as the component to be replaced.

The advanced RAS features that are built in to POWER9 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 such a failure.

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

### 4.3.3 Processor core/cache error handling

Layer 2 (L2) and Layer 3 (L3) caches and directories can correct single bit errors and detect double bit errors by error correction code (SEC/DED ECC). When a persistent correctable error occurs in these caches, the system has the capability of purging the data in the cache (writing to another level of the hierarchy) and deleting it. Beyond soft error correction, the intent of the POWER9 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 by using the Power On Reset Engine).

Soft errors that are detected in the level 1 caches are also correctable by a try again operation that is handled by the hardware. But instead of using an error correcting code, intermittent L1 cache errors can be corrected using data from elsewhere in the cache hierarchy. A portion of an L1 cache can be disabled (set delete) to avoid outages due to persistent hard errors. If too many errors are observed across multiple sets the core using the L1 can be predictively deallocated.

Separate from the system caches and the description above are cache directories which provide indexing to the caches. These also have single bit error correction but uncorrectable directory errors will typically result in system checkstops.

Beyond soft error correction, the intent of the POWER9 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.

Beyond the L1, L2, and L3 functional units, single bit correcting ECC is used in multiple areas of the processor as the standard means of protecting data against single bit errors. This includes a number of the internal busses where data is passed between units.

### 4.3.4 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, this is done by purging and deleting a single cache line. Such purge and delete operations are contained within the hardware itself, and 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 code that indicates that the cache line itself has bad data.
Nothing within the system terminates just because such an event is encountered. Rather, the hardware monitors the usage of pages with marks. If such data is never used, hardware replacement is requested, but nothing terminates as a result of the operation. Software layers are not required to handle such 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 such cases, if data is used as owned by a partition, then 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, then the hypervisor might choose to terminate, resulting in a system-wide outage.

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

### 4.3.5 Cyclic redundancy check and lane repair for processor fabric buses

Error correction code (ECC) is used internally in various data paths as data is transmitted between processor units. Externally to the processor, however, high speed data busses can be susceptible to occasional multiple bit errors due to electrical noise, timing drift, and a variety of other factors.

Previous POWER processor implementation like POWER8 use cyclic redundancy check code (CRC) to detect multiple bit errors on the memory bus. The data can be corrected by the memory controller’s ability to retry a faulty operation. If the memory bus experiences multiple CRC errors that must be corrected by retry, the memory controller can be dynamically retrained to re-establish optimal bus performance. If retraining a bus does not correct a persistent soft error, the fault could be due to a faulty bit line on the bus itself. The memory bus contains a dynamic spare bit line (dynamic memory channel repair) function which allows the memory controller to identify a persistently faulty bit line and to substitute a spare.

With the introduction of the POWER9 processor the CRC code for error detection, the retry capability on error conditions, and the ability to substitute a faulty date lane are extended to the processor fabric bus interfaces for both the on planar processor interconnect (X-bus) and the inter node processor interconnect (O-bus).

### 4.3.6 Split inter node connection bus with SMP cable redundancy

A Power E980 system node is using up to three inter node connection buses (O-buses) per processor module to facilitate the SMP communication with the other system nodes which are part of a multi-node configuration. One inter node connection bus is subdivided in two sets of lanes with one SMP cable connected through a dedicated port to each of the two sets. If there is a persistent error that is confined to the data on a single cable the related POWER9 processor module can reduce the bandwidth of the bus and send data across just the remaining SMP cable.

### 4.3.7 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 such 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 an outage 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 POWER9 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 are detected. This complements the try again abilities of the memory buffer module used in POWER9 enterprise class systems Power E950 and Power E980.

4.3.8 Predictive processor deallocation

Because of the amount of self-healing incorporated in POWER9 processor-based systems as well as the extensive error recovery features implemented it is rare that an entire processor core needs to be predictively deallocated due to a persistent recoverable error.

If such cases do occur, PowerVM can invoke a process for deallocating the failing processor dynamically at run-time. This process interacts with the operating system which holds access to the processor in question and requires that the control over the processor can be ceded by the operating system.

4.3.9 Core contained checkstops and PowerVM handled errors

Core hardware faults that processor instruction retry and the other previously described features defined in the hierarchy cannot contain might be handled through PowerVM by a technique called core contained checkstops.

The core contained checkstop technology allows PowerVM to be signaled when such faults occur and terminate code in use by the failing processor core. This allows the outage associated with the fault to be contained to just the logical partition using the core at the time when the uncorrectable fault occurred. In this way the core contained checkstop feature is particularly beneficial for scale-up Power Systems servers like the Power E980 which typically host a large number of logical partitions. However, a core contained checkstop signaling that a fault occurred on a core running a hypervisor instruction will typically result in hypervisor termination and a full system outage.

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

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

4.3.10 PCIe controller and enhanced error handling

Each processor has three elements called PCI express controller (PEC) which generate the various PCIe Gen4 busses used in the Power E980 system. The PEC has the ability to “freeze” operation when certain faults occur, and in certain cases can retry and recover from the fault condition. This PEC freeze behavior prevents faulty data from being written out through the I/O hub system and prevents reliance on faulty data within the processor complex when certain errors are detected.

The PEC freeze behavior supplements the traditional enhanced error handling (EEH) for I/O concept.

This capability signals device drivers when various PCIe bus related faults occur and permits the device driver to reset the slot and retry the failed operation. If retries were not attempted
or did not succeed, the EEH enables the device driver to restart the adapter (EEH recovery) or to terminate the use of the PCI adapter experiencing the faults without directly terminating a partition or system.

A clock error in the PCIe clocking can be signaled and recovered using EEH in any system that incorporates redundant PCIe clocks with dynamic failover enabled such as a Power E980 server.

4.3.11 Memory channel checkstops and hypervisor memory mirroring

The memory controller communicating between the processor and the memory buffer has its own set of methods for containing errors or retry operations.

Some severe faults require that memory under a portion of the controller become inaccessible to prevent reliance on incorrect data. There are cases where the fault can be limited to just one memory channel. In these cases the memory controller asserts what is known as a channel checkstop. In systems without hypervisor memory mirroring, a channel checkstop will usually result in a system outage. However, with hypervisor memory mirroring the hypervisor may continue to operate in the face of a memory channel checkstop.

4.3.12 Persistent guarding of failed elements

Not all processor core or processor module faults can be corrected by using the techniques that are described in this chapter. Therefore, provision is made for faults that require a system-wide outage. In such a “platform” checkstop event, the error detection and fault isolation (ED/FI) capabilities that are built in to the hardware and dedicated service processor work to isolate the root cause of the checkstop and 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 reboot the system automatically following an unrecoverable firmware error, firmware hang, hardware failure, or environmentally induced failure.

The auto-restart (reboot) option must be enabled from the Advanced System Management Interface (ASMI).
4.4 Memory RAS details

One POWER9 processor module of a Power E980 high-end server provides two integrated memory controllers to facilitate access to the main memory of the system. One memory controller drives four differential memory interface (DMI) channels with a maximum signaling rate of 9.6 GHz. Every DMI channel connects to one dedicated memory buffer chip. Each memory buffer chip in turn provides four DDR4 memory ports running at 1,600 MHz signal rate and one 16 MB L4 cache. A single memory buffer chip is mounted with the associated DRAM chips on one circuit board which is referred to as custom DIMM (CDIMM) module.

Figure 4-2 shows the memory subsystem design of a POWER9 processor module based on two memory controllers and eight DMI channels connecting to eight 32 GB CDIMMs each.

The memory buffer chip is manufactured in 22 nm lithography and incorporates similar technologies as used by POWER9 processor functional units to avoid soft errors. The integrated L4 cache is based on eDRAM technology with soft error hardening and persistent error handling features. The memory buffer implements a retry function 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 CDIMM uses cyclic redundancy check (CRC) error detection that is coupled with the ability to retry a memory access operation in case a soft error occurs. The bus features dynamic recalibration capabilities and a spare data lane that can be substituted for a failing bus lane through the recalibration process.

The memory buffer on each CDIMM has four ports for communicating with DRAM modules. For example, the 32 GB DDR4 DIMM features one rank that is composed of four ports of eight columns wide (x8) DRAM modules and each port contains 10 DRAM modules.

For each such port, there are eight DRAM modules worth of data (64 bits) and another DRAM module’s worth of error correction and other such data. There also is a spare DRAM module for each port that can be substituted for a failing DRAM chip.
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 process is also known as *Chipkill correction*. Then, it can correct at least one other bit within the ECC word.

The extra spare DRAM modules are used so that when a CDIMM experiences a Chipkill event within the DRAM modules under a port, the spare DRAM module can be substituted for a failing module. This substitution avoids the need to replace the CDIMM 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 CDIMM without needing to replace the CDIMM. Such a deprecated CDIMM still supports the full memory access bandwidth and is also still capable to correct soft errors within the functional DRAM chips through ECC.

Other CDIMMs are offered for Power E980 systems. A 64 B DDR4 CDIMM has two ranks, where each rank is similar to the 32 GB DDR4 CDIMM with DRAM modules on four ports and each port has 10 x8 DRAM modules. The 128 GB DDR4 CDIMM modules are using two ranks with a total of 152 x 4 Gb capacity DRAM chips. The 256 GB DDR4 CDIMM modules use the same number DRAM chips but the DRAMs are two-high (2H) 3D-stacked (3DS) and the number of ranks doubled to four. The 512 GB CDIMM modules use 152 four-high (4H) 3D-stacked (3DS) DRAM chips on eight ranks to support the specified capacity.

In addition to the protection that is provided by the ECC and sparing capabilities, the memory subsystem implements memory scrubbing to identify and correct single bit soft-errors. The PowerVM hypervisor is 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, such memory page deallocation is not relied upon for repair of faulty hardware.

Finally, should an uncorrectable error in data be encountered, the memory that is affected is marked with a special uncorrectable error code and handled as described in 4.3.4, “Cache uncorrectable error handling” on page 135.

### 4.5 PCIe I/O subsystem RAS details

In POWER8 processor-based systems, the external I/O hub and bridge adapters were eliminated in favor of a topology that integrates the PCI Express (PCIe) controller and the PCI Express host bridge (PHB) logic into the processor module. PCIe busses that are generated directly from a PHB can drive individual I/O slots or a PCIe switch. The integrated PCIe controller supports try again (end-point error recovery) and freezing. With POWER9 this design was carried forward to support PCI Express 4.0 technology.

#### 4.5.1 I/O subsystem availability and Enhanced Error Handling

Multi-path I/O and VIOS for I/O adapters and RAID for storage devices must be used to prevent application outages when I/O adapter faults occur.

To permit soft or intermittent faults to be recovered without failover to an alternative device or I/O path, IBM 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 often 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 or 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 started.

IBM device drivers under AIX are fully EEH-capable. For Linux under PowerVM, EEH support extends to many frequently used devices. There might be various third-party PCI devices that do not provide native EEH support.

4.5.2 PCIe Gen3 I/O Expansion drawer RAS

PCIe Gen3 I/O Expansion Drawers (#EMX0) can be used with Power E980 systems to increase I/O capacity. Figure 4-3 shows the functional components of the #EMX0 drawer.

These I/O drawers are attached using a connecting card called a PCIe3 Optical Cable Adapter (#EJ07) that plugs in to a PCIe slot of a Power E980 system node. The cable cards for POWER9 processor-based servers have been redesigned in certain areas to improve error handling. These improvements include new routing for clock logic within the cable card as well as additional recovery for faults during IPL.

Each I/O drawer contains up to two PCIe Fanout Modules (#EMXG). An I/O module uses 16 PCIe lanes controlled from a processor in a system node. Currently supported is an I/O module that uses a PCIe switch to supply six PCIe slots.

Two active optical cable are used to connect a PCIe3 cable adapter to the equivalent card in the I/O drawer module. While these cables are not redundant, as of FW830 firmware or later; the loss of one cable will simply reduce the I/O bandwidth (number of available lanes available to the I/O module) by 50%.

Figure 4-3  PCI3 Gen3 Expansion Drawer structural diagram
Infrastructure RAS features for the I/O drawer include redundant power supplies, fans, and DC outputs of voltage regulators (phases).

The impact of the failure of an I/O component can be summarized for the most significant cases by Table 4-2 below.

Table 4-2  PCIe Gen3 I/O Expansion Drawer RAS feature matrix

<table>
<thead>
<tr>
<th>Faulty Component</th>
<th>Impact of Failure</th>
<th>Impact of Repair</th>
<th>Prerequisites</th>
</tr>
</thead>
<tbody>
<tr>
<td>I/O adapter in an I/O slot</td>
<td>Loss of function of the I/O adapter</td>
<td>I/O adapter can be repaired while rest of the system continues to operate.</td>
<td>Multipathing I/O adapter redundancy, where implemented, can be used to prevent application outages</td>
</tr>
<tr>
<td>First Fault on a Data lane (in the optics between the PCIe3 cable adapter card in the system and the I/O module)</td>
<td>None: Spare Used</td>
<td>No Repair Needed: Integrated Sparing Feature</td>
<td></td>
</tr>
<tr>
<td>A Second Data Lane fault or other failure of one active optics cable</td>
<td>System continues to run but the number of active lanes available to I/O module will be reduced</td>
<td>Associated I/O module must be taken down for repair; rest of the system can remain active.</td>
<td></td>
</tr>
<tr>
<td>Other Failure of PCIe3 cable adapter card In System or I/O module</td>
<td>Loss of access to all the I/O of the connected I/O Module</td>
<td>Associated I/O module must be taken down for repair; rest of the system can remain active.</td>
<td>Systems with an HMC</td>
</tr>
<tr>
<td>One Fan</td>
<td>System continues to run with remaining fan</td>
<td>Concurrently repairable</td>
<td></td>
</tr>
<tr>
<td>One Power supply</td>
<td>System continues to run with remaining fan</td>
<td>Concurrently repairable</td>
<td></td>
</tr>
<tr>
<td>Voltage Regulator Module associated with an I/O module</td>
<td>System continues to run for a phase failure transition to n mode. Other faults will impact all the I/O in the Module</td>
<td>Associated I/O module cannot be active during repair; rest of the system can remain active.</td>
<td>Systems with an HMC</td>
</tr>
<tr>
<td>Chassis Management Card (CMC)</td>
<td>No Impact to running system, but once powered off, the I/O drawer cannot be re-integrated until CMC is repaired.</td>
<td>I/O drawer must be powered off to repair (loss of use of all I/O in the drawer)</td>
<td>Systems with an HMC</td>
</tr>
<tr>
<td>Midplane</td>
<td>Depends on source of failure, may take down entire I/O drawer</td>
<td>I/O drawer must be powered off to repair (loss of use of all I/O in the drawer)</td>
<td>Systems with an HMC</td>
</tr>
</tbody>
</table>
4.6 Enterprise systems availability

In addition to all of the standard RAS features that were described in this chapter, enterprise class systems allow for increased RAS and availability by including several unique features and redundant components.

The following advanced RAS features pertain to the Power E980 enterprise class system:

- **Redundant Service Processor**
  The service processor is an essential component of a system. It is responsible for the initial power load, setup, monitoring, control, and management. The control units that are present on enterprise class systems house two redundant service processors. If there is a failure in either of the service processors, the second processor ensures continued operation of the system until a replacement is scheduled. Even a system with a single system node included dual service processors in the system control unit.

- **Redundant System Clock Cards**
  Another component that is crucial to system operations is the system reference clock source which is responsible for providing a synchronized clock signal to all functional units. Each system node of a Power E980 server utilizes its own private set of two redundant system clock/control cards. If there is a failure in any of the clock/control cards, the second card ensures continued operation of the system until a replacement is scheduled. Unlike the POWER8 processor-based enterprise class servers Power E870, Power E870C, Power E880, and Power E880C the Power E980 system no longer requires a global reference clock source in the system control.

- **Dynamic Processor Sparing**
  Enterprise class systems are Capacity Upgrade on Demand capable. Processor sparing helps minimize the effect on server performance that is caused by a failed processor. An inactive processor is activated if a failing processor reaches a predetermined error threshold, which helps to maintain performance and improve system availability.

  Dynamic processor sparing happens dynamically and automatically when dynamic logical partitioning (DLPAR) is used and the failing processor is detected before failure. Dynamic processor sparing does not require purchasing an activation code. Instead, it requires only that the system have inactive CUoD processor cores available.

- **Dynamic Memory Sparing**
  Enterprise class systems are Capacity Upgrade on Demand capable. Dynamic memory sparing helps minimize the effect on 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 be in memory DIMMs. If there is a failure of DIMM, the hypervisor can become inoperative. The Active Memory Mirroring for Hypervisor allows for the memory blocks that are used in critical areas of the PowerVM hypervisor to be written in two distinct DIMMs. If an uncorrectable error is encountered during a read, the data is retrieved from the mirrored pair and operations continue normally. Not all of the PowerVM Hypervisor memory is mirrored but only the segments which are critical to keep the Power E980 system in an operational state in case a DIMM failure occurs. This selective mirroring strategy is implemented with fine granularity so that the hypervisor is not required to be permanently confined to a particular reserved memory location. To provide for optimized performance, PowerVM hypervisor code continues to be placed in accordance with affinity rules near the processor which needs the PowerVM service.

4.7 Availability effects of a solution architecture

Any solution should not rely on only the hardware platform. Despite IBM Power Systems being far superior RAS than other comparable systems, it is advisable to design a redundant architecture that surrounds the application to allow for easier maintenance tasks and greater flexibility.

By working in a redundant architecture, some tasks that require that a specific application is brought offline can now be executed with the application running, which allows for greater availability.

When determining a highly available architecture that fits your needs, consider the following questions:

- Will I need to move my workloads off an entire server during service or planned outages?
- If I use a clustering solution to move the workloads, how will the failover time affect my services?
- If I use a server evacuation solution to move the workloads, how long will it take to migrate all the partitions with my current server configuration?

4.7.1 Clustering

IBM Power Systems that is running under PowerVM and IBM i, AIX, and Linux support a spectrum of clustering solutions. These solutions meet requirements for application availability regarding server outages and data center disaster management, reliable data backups, and so forth. These offerings include distributed applications with IBM Db2® PureScale, HA solutions that use clustering technology with IBM PowerHA® SystemMirror®, and disaster management across geographies with PowerHA SystemMirror Enterprise Edition.

For more information, see the following resources:

- IBM PowerHA SystemMirror for i: Using Geographic Mirroring (Volume 4 of 4)
  http://www.redbooks.ibm.com/abstracts/sg248401.html
- IBM PowerHA SystemMirror for i: Using IBM Storwize (Volume 3 of 4)
  http://www.redbooks.ibm.com/abstracts/sg248402.html
4.7.2 Virtual I/O redundancy configurations

Within each server, the partitions can be supported by a single VIOS. However, if a single VIOS is used and that VIOS stops for any reason (hardware or software caused), all of the partitions that use that VIOS stop.

The use of Redundant VIOS servers mitigates this risk. Maintaining the redundancy of adapters within each VIOS (in addition to having redundant VIOS) avoids most faults that keep a VIOS from running. Therefore, multiple paths to networks and SANs are advised. A partition that is accessing data from two distinct Virtual I/O Servers, each one with multiple network and SAN adapters to provide connectivity, is shown in Figure 4-4.

![Partition with dual redundant Virtual I/O Servers](image)
Because each VIOS can be considered as an AIX based partition, each VIOS also must access a boot image, have paging space, and so on, under a root volume group or rootvg. The rootvg can be accessed through a SAN, the same as the data that partitions use.

Alternatively, a VIOS can use the U.2 NVMe internal storage option available for Power E980 systems. One system node of a Power E980 server supports four U.2 NVMe devices which can be individually assigned to independent partitions. To use storage that is locally attached (DASD devices or SSDs) through SAS adapters offers an other option to provide boot devices to VIOS partitions. However accessed, the rootvgs should use mirrored or RAID drives with redundant access to the devices for best availability.

4.7.3 PowerVM Live Partition Mobility

PowerVM Live Partition Mobility allows you to move a running logical partition (including its operating system and running applications) from one system to another without any shutdown and without disrupting the operation of that logical partition. Inactive partition mobility allows you to move a powered-off logical partition from one system to another.

Live Partition Mobility provides systems management flexibility and improves system availability through the following functions:

- Avoid planned outages for hardware or firmware maintenance by moving logical partitions to another server and then performing the maintenance. Live Partition Mobility can help lead to zero downtime for maintenance because you can use it to work around scheduled maintenance activities.
- Avoid downtime for a server upgrade by moving logical partitions to another server and then performing the upgrade. This approach allows your users to continue their work without disruption.
- Avoid unplanned downtime. With preventive failure management, you can move a server’s logical partitions to another server before the failure occurs if a server indicates a potential failure. Partition mobility can help avoid unplanned downtime.
- Take advantage of server optimization:
  - Consolidation: You can consolidate workloads that run on several small, under-used servers onto a single large server.
  - Deconsolidation: You can move workloads from server-to-server to optimize resource use and workload performance within your computing environment. With live partition mobility, you can manage workloads with minimal downtime.

Server Evacuation: This PowerVM function allows you to perform a server evacuation operation. Server Evacuation is used to move all migration-capable LPARs from one system to another if there are no active migrations in progress on the source or the target servers.

With the Server Evacuation feature, multiple migrations can occur based on the concurrency setting of the HMC. Migrations are performed as sets, with the next set of migrations starting when the previous set completes. Any upgrade or maintenance operations can be performed after all the partitions are migrated and the source system is powered off.

You can migrate all the migration-capable AIX, IBM i, and Linux partitions from the source server to the destination server by running the following command from the HMC command line:

```
migrate -o m -m source_server -t target_server --all
```
Hardware and operating system requirements for Live Partition Mobility

Live Partition Mobility is supported by default with enterprise systems. It is also supported in compliance with all operating systems that are compatible with POWER9 technology.

The VIOS partition cannot be migrated.

For more information about Live Partition Mobility and how to implement it, refer to the IBM PowerVM Virtualization Introduction and Configuration Redbook:

http://www.redbooks.ibm.com/abstracts/sg247940.html

4.8 Serviceability

The purpose of serviceability is to repair the system while attempting to minimize or eliminate service cost (within budget objectives) and maintain application availability and high customer satisfaction. Serviceability includes system installation, miscellaneous equipment specification (MES) (system upgrades or downgrades), and system maintenance or repair. Depending on the system and warranty contract, service might be performed by the customer, an IBM System Services 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 SSR Set Up and Customer Installed Features (CIF).
- Error Detection and Fault Isolation (ED/FI).
- First Failure Data Capture (FFDC).
- Guiding Light service indicator architecture is used to control a system of integrated LEDs that lead the individual servicing the machine to the correct part as quickly as possible.
- Service labels, service cards, and service diagrams are available on the system and delivered through the HMC.
- Step-by-step service procedures are available through the HMC.

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 found in all POWER processor-based systems.

4.8.1 Detecting errors

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

Although not all errors are a threat to system availability, those 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 z Systems server-inspired error detection mechanisms, which extend from processor cores and memory to power supplies and hard disk drives (HDDs).

4.8.2 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 fault isolation registers (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 IBM Power Systems design is the concept of first-failure data capture (FFDC). FFDC is a technique that involves sufficient error checking stations and co-ordination 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 many 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. 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.

4.8.3 Service processor

In POWER9 processor-based systems, the flexible service processor (FSP) is a microprocessor that is powered separately from the main instruction processing complex.

The service processor performs the following serviceability functions:

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

The service processor interfaces with the on chip controller (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 appropriate corrective actions before a critical failure threshold is reached. The service processor can also post a warning and start 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.
PowerVM Hypervisor (system firmware) and HMC connection surveillance

The service processor monitors the operation of the firmware during the boot process and 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 the operating system can then take appropriate action, including calling for service. The 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

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

The auto-restart (reboot) option must be enabled from the Advanced System Management Interface (ASMI).

Concurrent access to the service processors menus of the ASMI

This access allows nondisruptive abilities to change system default parameters, interrogate service processor progress and error logs, set and reset service indicators (Light Path for low-end servers), and access all service processor functions without powering 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. Some options, such as changing the hypervisor type, do not take effect until the next boot.

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

4.8.4 Diagnosing

General diagnostic objectives are created 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 IBM back-end system.

By 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-diagnosis.

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 correctly detected and isolated at the time that the failure occurs. Runtime and boot time diagnostic tests fall into this category.

Boot time

When an IBM Power Systems server powers up, the service processor initializes the system hardware. Boot time diagnostic testing uses a multitier 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. This determination based on the way that the system was powered off or on the boot-time selection menu.

**Host Boot IPL**

On POWER9 processor-base systems, the boot process is initializes by the FSP, however, one part of the running firmware is performing the Central Electronics Complex (CEC) chip initialization. A component external to the POWER9 processor that is called the Power NOR (PNOR) chip stores the Host Boot (HB) firmware and the self-boot engine (SBE) code. During system Initial Program Load (IPL) the PNOR starts the SBE which eventually loads the HB base code onto the POWER9 chip.

With this HB initialization, new progress codes are available. An example of an FSP progress code is C1009003. During the HB 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.

**Processor run time diagnostics**

All IBM Power Systems servers can monitor critical system components during run time. They also can take corrective actions when recoverable faults occur. In POWER9 processor-based systems virtualized through PowerVM the IBM hardware error-check architecture in conjunction with the First Failure Data Capture (FFDC) implementation reports errors in the Central Electronics Complex (CEC) to a special service partition which is under the control of the PowerVM Hypervisor. The special service partition runs the processor runtime diagnostics (PRD) code which ingests and processes the error information and directs the error management. The hypervisor can restart the special service partition and reload the PRD code in case the run time diagnostics service became unavailable. On previous POWER7 and POWER8 processor-based systems the PRD code was executed on the Flexible Service Processor (FSP).

Extensive diagnostic and fault analysis routines were developed and improved over many generations of POWER processor-based servers. These routines enable quick and accurate predefined responses to actual and potential system problems. The PRD code running in the special service partition correlates and processes run time 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. 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 operating system 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 start the Call Home request to IBM. I/O devices can also include specific exercisers that can be started by the diagnostic facilities for problem recreation (if required by service procedures).

4.8.5 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 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 that is on the system node to alert the user of an internal hardware problem.

On POWER9 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, forwards the event to the Service Focal Point (SFP) application that is running on the management console, and 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, a Call Home service request is started and the pertinent failure data with service parts information and part locations is sent to the IBM service organization if the system is correctly configured. 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, see “Service Event Manager” on page 166. -- xref need update --

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 the following types of basic data:

- 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. 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, which includes the operating system that runs on the HMC. RMC provides a secure transport mechanism across the LAN interface between the operating system and the optional HMC and is used by the operating system diagnostic application for transmitting error information. It performs several other functions, but those functions 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 being reported for service. Also, an error should be reported only once, regardless of how many logical partitions 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 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, which provides a redundant error-reporting path in case errors are 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 logical partitions 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, starts a Call Home for service. This methodology ensures that all platform errors are reported through at least one functional path, which results 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 data that is collected automatically at the time of a failure or manually later. Although the data that is collected depends on the invocation method, it includes information, such as firmware levels, operating system levels, other fault isolation register values, recoverable error threshold register values, and system status.

The data is formatted and prepared for transmission back to IBM to assist the service support organization with preparing a service action plan for the IBM SSR or for more analysis.

System dump handling
In certain circumstances, an error might require a memory dump to be 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 can be sent to IBM Support for analysis. If more information that relates to the memory dump is required, or if viewing the memory dump remotely becomes necessary, the management console memory dump record notifies the IBM Support center regarding on which managements console the memory dump is located. If no management console is present, the memory dump might be on the FSP or in the operating system, depending on the type of memory dump that was started and whether the operating system is operational.
4.8.6 Notifying

After an IBM Power Systems server detects, diagnoses, and reports an error to an appropriate aggregation point, it notifies 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 having field service personnel automatically dispatched to the client site with the replacement part.

**Client Notify**

When an event is important enough to report but does not indicate the need for a repair action or 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, including the following examples:

- Network events, such as the loss of contact over a local area network (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 customer 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 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. The use of the Call Home feature can result in increased system availability. The ESA application can be configured for automated Call Home. For more information, see 4.9.4, “Electronic Services and Electronic Service Agent” on page 165. -- xref needs update --

**Vital product data and inventory management**

IBM 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, which enables 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, along with 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.
4.8.7 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 cannot 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 become loose during shipping or installation, which prevents a good electrical connection. Positive retention mechanisms, such as latches, levers, thumb-screws, pop Nylatches (U-clips), and cables are included to help prevent loose connections and 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. When a fault condition is detected on the POWER9 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 IBM SSR also can be guided 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 visible only while the system is in standby power. There are two gold caps implemented. The gold cap is used to illuminate the amber LEDs after power is removed from the system. One cap is inside the drawer to identify DIMMs, processors, and VRMs. The other cap is in the RAID assembly.

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

The IBM Knowledge Center is available this website:
http://www.ibm.com/support/knowledgecenter/

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

The following service labels are available:

- Location diagrams
  These 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
  These labels contain procedures that often are found on a cover of the system or in other locations that are accessible to the service representative. These labels provide systematic procedures (including diagrams) that describe how to remove and replace certain serviceable hardware components.

- Numbered arrows
  These arrows are used to 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 mechanisms can engage or disengage. Arrows often improve the ease of serviceability.

Operator panel
The operator panel of the Power E980 resides in the system control unit and is comprised of a base unit and a separate LCD unit which are individually concurrent maintainable. The operator panel is used to present boot progress codes, which indicate advancement through the system power-on and initialization processes. The operator panel also is used to display error and location codes when an error occurs that prevents the system from booting. It includes several buttons, which enable an IBM SSR or client to change various boot-time options and for other limited service functions.

The base operator panel provides LEDs and sensors:

- Power LED:
  - Color: Green
  - Off: Enclosure is off (AC cord is not connected)
  - On Solid: Enclosure is powered on
  - On Blink: Enclosure is in the standby-power state

- Enclosure Identify LED:
  - Color: Blue
  - Off: Normal
– On Solid: Identify State

► System Fault LED:
  – Color: Amber
  – Off: Normal
  – On Solid: Check Error Log

► System Roll-up LED:
  – Color: Amber
  – Off: Normal
  – On Solid: Fault

► Power Button
► System Reset Switch
► Two Thermal Sensors
► One Pressure/Altitude Sensor

The LCD operator panel features two rows of 16 characters and increment, decrement and Enter buttons.

The following functions are available through the operator panel:
  ► Error information
  ► Generate dump
  ► View machine type, model, and serial number
  ► View or change IP addresses of the service processor
  ► Limited set of repair functions

Concurrent maintenance
The IBM POWER9 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 begin to wear from repeated plugging and unplugging. For these reasons, these devices are concurrently maintainable when properly configured. Concurrent maintenance is facilitated by the redundant design for the power supplies, fans, and physical storage.

In addition to these 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 appropriate sequence for the particular repair for the specific system being 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 user engineers and IBM SSR providers who are familiar with the task and those engineers and 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.

4.9 Manageability

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

4.9.1 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.

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

- Light Path (see “Light Path” on page 154 and “Service labels” on page 155)
- 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 includes 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 used to monitor and manage the system hardware resources and devices. The service processor checks the system for errors, which ensures the connection to the management console for manageability purposes and for accepting ASMI Secure Sockets Layer (SSL) network connections. The service processor can view and manage the machine-wide settings by using the ASMI. It also 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 is available in the ASMI and are transferred to the HMC.
The service processor uses two Ethernet ports that run at 1 Gbps speed. Consider the following points:

- 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. Customer-selectable media speed and duplex settings are not available.
- The Ethernet ports have the following default IP address:
  - 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 with which you manage the operation of the server, such as auto-power restart. You also can 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, a standard Ethernet cable or a crossed cable) or through an Ethernet network. ASMI can also be accessed from an ASCII terminal, but this option 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 use the service processor’s default settings to operate your server. If the default settings are used, accessing the ASMI is not necessary. To access ASMI, use one of the following methods:

- **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)**.

- **Web browser**
  
  At the time of this writing, supported web browsers are Netscape 9.0.0.4, Microsoft Internet Explorer 7.0, Opera 9.24, and Mozilla Firefox 2.0.0.11. Later versions of these
browsers might work, but are not officially supported. The JavaScript language and cookies must be enabled and 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 Microsoft Internet Explorer, click **Tools Internet Options.** Clear the **Use TLS 1.0** option, 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 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.

- Command-line start of the ASMI
  On the HMC or when properly configured on a remote system, the ASMI web interface can be started from the HMC command line. Open a 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 properly, a browser window opens on a remote system when issued from there.

**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 the following ways:

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

The operator panel includes the following features:

- 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
Operating system service menu

The system diagnostic tests consist of IBM i service tools, stand-alone diagnostic tests that are loaded from the DVD drive, and online diagnostic tests (available in AIX).

When installed, online diagnostic tests are a part of the AIX or IBM i operating system on the disk or server. They can be booted in single-user mode (service mode), run in maintenance mode, or run concurrently (concurrent mode) with other applications. They can access the AIX error log and the AIX configuration data. IBM i has a service tools problem log, IBM i history log (QHST), and IBM i problem log.

The following modes are available:

- **Service mode**
  This mode requires a service mode boot of the system and enables the checking of system devices and features. Service mode provides the most complete self-check of the system resources. All system resources (except the SCSI adapter and the disk drives that are used for paging) can be tested.

- **Concurrent mode**
  This mode enables the normal system functions to continue while selected resources are being checked. Because the system is running in normal operation, certain devices might require more actions by the user or a diagnostic application before testing can be done.

- **Maintenance mode**
  This mode enables checking most system resources. Maintenance mode provides the same test coverage as service mode. The difference between the two modes is the way that they are started. Maintenance mode requires that all activity on the operating system is stopped. Run `shutdown -m` to stop all activity on the operating system and put the operating system into maintenance mode.

The System Management Services (SMS) error log is accessible on the SMS menus. This error log contains errors that are found by partition firmware when the system or partition is booting.

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 an IBM 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.

IBM i and its associated machine code provide dedicated service tools (DSTs) as part of the IBM i licensed machine code (Licensed Internal Code) and System Service Tools (SSTs) as part of IBM i. DSTs can be run in dedicated mode (no operating system is loaded). DSTs and diagnostic tests are a superset of those available under SSTs.

The IBM i End Subsystem (ENDSBS *ALL) command can shut down all IBM and customer applications subsystems except for the controlling subsystem QTCL. The Power Down System (PWRDWN SYS) command can be set to power down the IBM i partition and restart the partition in DST mode.

You can start SST during normal operations, which keeps all applications running, by using the IBM i Start Service Tools (STRSST) command (when signed onto IBM i with the appropriately secured user ID).
With DSTs and SSTs, you can review various logs, run various diagnostic tests, or take several kinds of system memory dumps or other options.

Depending on the operating system, the following service-level functions are what you often see when you use the operating system service menus:

- Product activity log
- Trace Licensed Internal Code
- Work with communications trace
- Display/Alter/Dump
- Licensed Internal Code log
- Main storage memory dump manager
- Hardware service manager
- Call Home/Customer Notification
- Error information menu
- LED management menu
- Concurrent/Non-concurrent maintenance (within scope of the OS)
- Managing firmware levels:
  - Server
  - Adapter
- Remote support (access varies by OS)

**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 logical partition reports errors that it detects and forwards the event to the SFP application that is running on the management console without determining whether other logical partitions also detect and report the errors. For example, if one logical partition reports an error for a shared resource, such as a managed system power supply, other active logical partitions 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 start service functions on systems and logical partitions, including the exchanging of parts, configuring connectivity, and managing memory dumps.

### 4.9.2 IBM Power Systems Firmware maintenance

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

**Firmware entitlement**

With the HMC Version V8R8.1.0.0, the firmware installations are restricted to entitled servers. The customer 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 some extra time.

The IBM 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. The following 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 on 31 December 2014, and a new firmware release is release 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 can still 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 following IBM CoD Home website:

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

For more information about IBM entitled Software Support, see this 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 customer operations. IBM intends to introduce no more than two new release levels per year. These release levels are supported by service packs. Service packs are intended to contain only firmware fixes and not introduce new functions. A service pack is an update to a release level.

The management console is used for system firmware updates. By using the management console, you can use the CFM option when concurrent service packs are available. CFM is the IBM Power Systems Firmware updates that can be partially or wholly concurrent or nondisruptive. With the introduction of CFM, IBM is increasing its clients’ opportunity to stay on a specific release level for longer periods. Clients that want maximum stability can defer until there is a compelling reason to upgrade, such as the following reasons:

- A release level is approaching its end of service date (that is, it was available for approximately one year and soon service will not be supported).
- They want to move a system to a more standardized release level when there are multiple systems in an environment with similar hardware.
- A new release features a new function that is needed in the environment.
- A scheduled maintenance action causes a platform reboot, which also provides an opportunity to upgrade to a new firmware release.
Updating and upgrading system firmware depends on several factors, including the current firmware that is installed, and what operating systems are running on the system. These scenarios and the associated installation instructions are described in the firmware section of Fix Central, which is available at the following website:

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

You also might want to review the preferred practice white papers that are found 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 from the HMC or a running partition.

IBM Power Systems includes 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 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 following POWER9 section on the IBM Support Portal: -- URL update needed. not available jet --


For POWER9 based IBM Power Systems, select the **POWER9 systems** link.

Within this section, search for “Firmware and HMC updates” to find the resources for keeping your system's firmware current.

If there is an HMC to manage the server, the HMC interface can be used to view the levels of server firmware and power 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 power subsystem firmware:

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

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

- **Accepted level**
  
  This level is the backup level of server or power subsystem firmware. You can return to this level of server or power subsystem firmware if you decide to remove the installed level. It is installed on the permanent side of system firmware.
Use the HMC enhanced GUI to get information about the different firmware levels in effect by selecting Resources → All Systems → chose the system or the systems of interest → Actions → Updates → View system information → check None - Display current values. Figure 4-5 shows the information which has subsequently been collected by the HMC.

![Figure 4-5 HMC Enhanced GUI system firmware information](image)

IBM provides the CFM function on Power E980 servers. This function supports applying nondisruptive system firmware service packs to the system concurrently (without requiring a reboot operation to activate changes).

The concurrent levels of system firmware can (on occasion) contain fixes that are known as deferred. These deferred fixes can be installed concurrently but are not activated until the next IPL. Any deferred fixes 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.

The file-naming convention for system firmware is listed in Table 4-3.

### Table 4-3 Firmware naming convention

<table>
<thead>
<tr>
<th>PP</th>
<th>NN</th>
<th>SSS</th>
<th>FFF</th>
<th>DDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Package identifier</td>
<td>Platform and class</td>
<td>Release indicator</td>
<td>Current fix pack</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 -- P9 update needed --

An installation is disruptive if the following statements are true:

- The release levels (SSS) of the currently installed 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.
4.9.3 Concurrent firmware maintenance 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. However, a patch in this area required a system reboot for activation because changes to some server functions (for example, changing initialization values for chip controls) cannot occur during operation.

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 discernible affect to any applications that are running on a processor. This potentially allows concurrent firmware changes in POWER9, which in earlier designs required a reboot to take effect.

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

4.9.4 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 extra charge enhancement to the service and support that is 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 the following 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
The ESA is software that is on your server. It monitors events and transmits system inventory information to IBM on a periodic, client-defined timetable. The 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 the Service Agent is made available to IBM SSRs when they help answer your questions or diagnose problems. The installation and use of ESA for problem reporting enables IBM to provide better support and service for your IBM server.

For more information about how Electronic Services can work for you, see this website (an IBM ID is required):
http://www.ibm.com/support/electronicsupport
Electronic Services features the following benefits:

- Increased uptime
  The ESA tool enhances the warranty or maintenance agreement by providing faster hardware error reporting and uploading system information to IBM Support. This benefit 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 dependence on human intervention or off-hours customer 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 through the Internet (HTTPS or VPN) or modem. It can be configured to communicate securely through gateways to provide customers a single point of exit from their site.

  Communication is one way. Activating ESA does not enable IBM to call into a customer’s system. System inventory information is stored in a secure database, which is protected behind IBM firewalls. It is viewable only by the customer and IBM. The customer’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, customers are not required to find and send system information, which decreases the risk of misreported or misdiagnosed errors.

  When inside IBM, problem error data 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 following Electronic Support website:

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

  My Systems provides valuable reports of installed hardware and software by using information that is collected from the systems by ESA. Reports are available for any system that is associated with the customers IBM ID. Premium Search combines the function of search and the value of ESA information, which provides advanced search of the technical support knowledge base. By using Premium Search and the ESA information that was collected from your system, your 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/electronicsupport

**Service Event Manager**

The Service Event Manager (SME) allows the user to decide which of the Serviceable Events are called home with the ESA. Certain events can be locked. Some customers might not allow data to be transferred outside their company. After the SEM is enabled, the analysis of the possible problems might take longer.

Consider the following points:

- The SEM can be enabled by running the following command:

  chhmc -c sem -s enable
You can disable SEM mode and specify what 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
```

The basic configuration of the SEM can be accomplished by using the HMC Enhanced GUI. Select **Serviceability → Event Manager for Call Home** (see Figure 4-6) to get access to the Events Manager for Call Home menu.

![Figure 4-6 Service Event Manager configuration through the HMC Enhanced GUI](image)

In the Events Manager for Call Home menu, you can add a HMC that is used to manage the serviceable events to the list of registered management consoles and proceed with further configuration steps, as shown in Figure 4-7.

![Figure 4-7 Event Manager for Call Home menu of the HMC Enhanced GUI](image)
The following configurable options are available:

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

- **Event Criteria**
  Select the filters for filtering the list of serviceable events that are shown. 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 the **All consoles** option to filter the list.

- **Serviceable Events**
  The Serviceable Events table shows the list of events that are based on the filters that are selected. 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 yet approved.

The Help / Learn more function can be used to get more information about the other available windows for the Serviceable Event Manager.

### 4.10 Selected POWER9 RAS capabilities by operating system

The IBM Power Systems RAS capabilities are listed by operating system in Table 4-4. The HMC is an optional feature on scale-out IBM Power Systems servers.

#### Table 4-4  Selected RAS features by operating system

<table>
<thead>
<tr>
<th>RAS feature</th>
<th>AIX</th>
<th>IBM i</th>
<th>Linux</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>V7.2 TL0</td>
<td>V7R1 TR10</td>
<td>RHEL6.5</td>
</tr>
<tr>
<td></td>
<td>V7.1 TL4</td>
<td>V7R2 TR4</td>
<td>RHEL7.1</td>
</tr>
<tr>
<td></td>
<td>V7.1 TL3 SP4</td>
<td>V7R3</td>
<td>SLES11SP3</td>
</tr>
<tr>
<td></td>
<td>V6.1 TL9 SP3</td>
<td></td>
<td>SLES12</td>
</tr>
<tr>
<td>Processor</td>
<td></td>
<td></td>
<td>Ubuntu 16.04</td>
</tr>
<tr>
<td>First failure data capture (FFDC) for fault detection/error isolation</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>RAS feature</td>
<td>AIX</td>
<td>IBM i</td>
<td>Linux</td>
</tr>
<tr>
<td>-------------------------------------------------</td>
<td>-----</td>
<td>-------</td>
<td>---------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynamic Processor Deallocation</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td><strong>I/O subsystem</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCI Express bus enhanced error detection</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>PCI Express bus enhanced error recovery</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>PCI Express card hot-swap</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td><strong>Memory availability</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Memory Page Deallocation</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Special Uncorrectable Error Handling</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td><strong>Fault detection and isolation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Storage Protection Keys</td>
<td>X</td>
<td>Not used by OS</td>
<td>Not used by OS</td>
</tr>
<tr>
<td>Error log analysis</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td><strong>Serviceability</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boot-time progress indicators</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Firmware error codes</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Operating system error codes</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Inventory collection</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Environmental and power warnings</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Hot-swap DASD / media</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Dual disk controllers / Split backplane</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>EED collection</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SP/OS “Call Home” on non-HMC configurations</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>IO adapter/device stand-alone diagnostic tests with PowerVM</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SP mutual surveillance with IBM POWER Hypervisor</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Dynamic firmware update with HMC</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Service Agent Call Home Application</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Service Indicator LED support</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>System dump for memory, POWER Hypervisor, and SP</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>IBM Knowledge Center / IBM Systems Support Site service publications</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>System Service/Support Education</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>RAS feature</td>
<td>AIX</td>
<td>IBM i</td>
<td>Linux</td>
</tr>
<tr>
<td>----------------------------------------------------------------------------</td>
<td>--------</td>
<td>---------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Operating system error reporting to HMC Service Focal Point (SFP) application</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>RMC secure error transmission subsystem</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Healthcheck scheduled operations with HMC</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Operator panel (real or virtual [HMC])</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Concurrent Op Panel Display Maintenance</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Redundant HMCs</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>High availability clustering support</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Repair and Verify Guided Maintenance with HMC</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>PowerVM Live Partition / Live Application Mobility With PowerVM Enterprise Edition</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

**EPOW**

| EPOW errors handling | X | X | X |
Related publications

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

IBM Redbooks

The following IBM Redbooks publications provide more information about the topics in this document. Some publications that are referenced in this list might be available in softcopy only.

- *IBM Power System AC922 Introduction and Technical Overview*, REDP-5472
- *IBM Power System AC922 Introduction and Technical Overview*, REDP-5494
- *IBM Power System L922 Introduction and Technical Overview*, REDP-5496
- *IBM Power Systems LC921 and LC922 Introduction and Technical Overview*, REDP-5495
- *IBM Power System S822LC for High Performance Computing Introduction and Technical Overview*, REDP-5405
- *IBM Power Systems S922, S914 and S924 Introduction and Technical Overview*, REDP-5497
- *IBM Power Systems E950 Technical Overview and Introduction*, REDP-5509
- *IBM PowerVM Best Practices*, SG24-8062
- *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940
- *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590

You can search for, view, download, or order these documents and other Redbooks publications, Redpapers, web docs, drafts, and additional materials, at the following website:

ibm.com/redbooks

Online resources

These websites are also relevant as further information sources:

- IBM Fix Central website
- IBM Knowledge Center
- IBM Knowledge Center: IBM Power Systems Hardware
IBM Knowledge Center: Migration combinations of processor compatibility modes for active Partition Mobility
http://www-01.ibm.com/support/knowledgecenter/api/redirect/powersys/v3r1m5/topic/p7hc3/iphc3pcmcombosact.htm

IBM Portal for OpenPOWER - POWER9 Monza Module

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

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

IBM Systems Energy Estimator
http://www-912.ibm.com/see/EnergyEstimator/

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

NVIDIA Tesla V100

NVIDIA Tesla V100 Performance Guide

OpenCAPI
http://opencapi.org/technical/use-cases/

OpenPOWER Foundation
https://openpowerfoundation.org/

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

Support for IBM Systems website

Help from IBM

IBM Support and downloads
ibm.com/support

IBM Global Services
ibm.com/services