Escala Power6 Overview

ESCALA POWER6

REFERENCE
86 A1 32EV 01

ESCALA POWER6

Escala Power6 Overview

Hardware

November 2007

BULL CEDOC
357 AVENUE PATTON
B.P.20845
49008 ANGERS CEDEX 01
FRANCE

REFERENCE
86 A1 32EV 01

Suggestions and criticisms concerning the form, content, and presentation of this book are invited. A form is provided at the end of this book for this purpose.

To order additional copies of this book or other Bull Technical Publications, you are invited to use the Ordering Form also provided at the end of this book.

Trademarks and Acknowledgements

Contents

About this publication

This publication describes the design, components, functions, features, and capabilities of the systems. It is intended for executives, data processing managers, data processing technical staff, consultants, and vendors who want to learn the advantages of the systems.

For information about the accessibility features of this product, for users who have a physical disability, see "Accessibility features," on page 33.

Related publications

To access other supporting information, contact service and support.

How to send your comments

Your feedback is important in helping to provide the most accurate and highest quality information. If you have any comments about this publication, send your comments to us. Be sure to include the name of the book, the form number of the book, and the specific location of the text you are commenting on (for example, a page number or table number).

Chapter 1. Introduction to the POWER6 processor-based systems

The POWER6™ processor-based systems provide a combination of scalability, availability, reliability, and virtualization.

POWER6 technology
: The 5/70 (17M/MA) system contains a system unit based on POWER6 technology and a new PCI Express-based internal I/O subsystem. The system unit consists of up to two 2-core processor cards, a system unit backplane, and up to three dc-dc regulators. The 2-core processor cards contain the POWER6 dual core processor with an integrated memory controller and a 32 MB level 3 (L3) cache chip along with 12 DDR2 dual inline memory module (DIMM) slots.

  POWER6 processors can run 64-bit applications while concurrently supporting 32-bit applications to enhance flexibility. These processors feature simultaneous multithreading, allowing two application threads to be run at the same time, which can significantly reduce the time to complete tasks. The PCI Express internal I/O subsystem contains four (8x) PCI Express slots and two PCI-X DDR slots.

SAS technology
: For POWER6 processor-based systems, new Serial Attached SCSI (SAS) technology is being used which includes the SAS DASD controller chip that provides an SAS interface to the DASD subsystem and a Serialized AT Attachment (SATA) interface to the media devices. A conversion from SATA to IDE occurs on the media backplane, which allows standard IDE media optical devices to be installed.

GX+ bus slots
: Two GX+ bus slots provide the interface to either a RIO-G adapter or the new GX Dual-Port 12X Channel Attach adapter. The GX Dual-Port 12X Channel Attach adapter provides support for the new 14G/30 I/O expansion unit.

Advanced POWER virtualization
: Advanced POWER™ Virtualization allows you to enhance the virtualization capabilities of your system. The Advanced POWER Virtualization feature includes the Virtual I/O Server and enablement for Micro-Partitioning™.

Host Ethernet adapter
: The 17M/MA supports the choice of 1 Gb or 10 Gb integrated host Ethernet adapters (HEA). These ports can be selected at the time of initial order. The 17M/MA supports virtualization of these integrated Ethernet adapters.

Improved RAS
: The Reliability, Availability, and Serviceability (RAS) features help to ensure that the system operates when required, performs reliably, and handles any failures that might occur in an efficient manner. The POWER6 processor-based system offers many features that are designed to increase Reliability, Availability, and Serviceability.

Enhancements

There are many enhancements with the POWER6 processor-based systems including an updated user interface on the HMC, advanced virtualization functions, and enhanced RAS features.

Documentation
: Contact service and support for information on accessing the documentation.

Hardware Management Console
An updated user interface that requires fewer clicks to access key tasks is now available. Added accessibility features allow technology that assists the user.

Advanced POWER virtualization
Advanced POWER virtualization functions facilitate highly efficient system utilization.

Enhanced RAS
Enhanced Reliability, Availability, and Serviceability (RAS) features are designed to improve application availability.

Electronic problem reporting
An updated Electronic Service Agent™ allows server information and problems to be electronically reported to the service and support organization.

Improvements to documentation

Contact service and support for the documentation.

Highlights of the documentation

Highlights of the new information delivery for POWER6 processor-based systems include:

Synchronization of delivery
Information is published and maintained on the same schedule as the product.

Improved online help information on the HMC

- The online help now includes a more powerful search engine, a new interface, and print feature.
- You can operate all features using the keyboard in addition to the mouse.

Adding and exchanging field replaceable units (FRUs) using the HMC
The Service Management menus on the HMC provide an interactive step-by-step process using illustrations and video presentations to help customers and service representatives add and exchange parts if needed.

POWER6 product specifications

The POWER6 processor-based systems include the 5/70 (17M/MA), five I/O expansion units, and two new models of the Hardware Management Console (HMC).

5/70 (17M/MA)
The 17M/MA is a 2-16 core midrange server mounted in an industry-standard 19-inch rack. Multiple systems, each of which is 4-EIA units high, can be joined to build larger *n*-core systems.

Each 17M/MA building block can have up to two 2-core processor cards.

Up to four building blocks can be joined together, so a 16-core system is the maximum configuration.

All configurations use modular symmetric multiprocessor (SMP) architecture. This design allows customers to start with what they need and add additional systems as necessary, without disrupting the base system. Capacity on Demand (CoD) features enable you to activate dormant processors for times as short as one minute.

I/O expansion units
The I/O expansion units that the 17M/MA supports include:

- The 11D/11, which contains six PCI-X slots.
- The 11D/20, which contains seven PCI-X slots and 12 hot-swappable SCSI disk drive bays arranged in two 6-packs.
- The 14G/30, which contains six PCI-X DDR slots.

- The 31D/24 and 31T/24 both include Ultra 320 SCSI interface connections for up to 24 LVD Ultra 320 SCSI disk drives. The models are available in the following configurations:
  – The 31D/24 - rack-mountable configuration
  – The 31T/24 - stand-alone deskside configuration

Note: For additional details on the I/O expansion units, refer to "Expansion unit features" on page 6.

Hardware Management Console (HMC) 42C/06
The 42C/06 is a desktop model that includes one 10/100/1000 Ethernet port, but two additional dual-port 10/100/1000 Gb Ethernet adapters can be added.

Hardware Management Console (HMC) 42C/R4
The 42C/R4 is a 1-EIA unit high, 19-inch rack-mountable model that has two Ethernet ports, but one additional dual-port 10/100/1000 Gb Ethernet adapter can be added.

System specifications

Learn about the system specifications for a single 17M/MA system.

Table 1. 17M/MA specifications

Description	**Range (operating)**
Operating temperature	5 to 35 degrees C (41 to 95 F)
Relative humidity	8% to 80%
Wet bulb	23 degrees C (73 F) (maximum configuration)
Noise level	6.2 to 7.1 bels (operating 4-core configurations)
Operating voltage	200 to 240 V AC 50/60 Hz
Power consumption	1,400 watts (maximum)
Power source loading	1.428 kVA (maximum configuration)
Thermal output	4,778 British thermal unit (BTu)/hour (maximum configuration)

Physical package

The 5/70 (17M/MA) system is available only in the rack-mountable configuration.

The dimensions of the system are listed in Table 2.

Table 2. Physical packaging of the 17M/MA

Dimension	**One 17M/MA system**
Height	174 mm (6.85 inches)
Width	483 mm (19.0 inches)
Depth	793 mm (31.2 inches) from the front rack rail mounting surface to the rear of the power supply
Weight	63.6 kg (140 pounds)

A 17M/MA system can have one to four system units. Review the supplier's installation and planning information for any product-specific installation requirements before installing the system or systems into a rack.

Features of the 17M/MA system

A fully configured 17M/MA has four building blocks.

A fully configured 17M/MA building block includes the following capacities:

- Up to eight processor cards using the POWER6 chip, for a total of 16 processors
- Up to 768 GB of DDR2 memory
- 24 SAS disk drives for an internal storage capacity of 7.2 TB using 300 GB drives
- 24 PCI slots: 8 PCI-X DDR and 16 (8x) PCI Express
- Four SlimLine media bays for optional optical storage devices

For a multiple-drawer server configuration, a processor fabric cable or cables and a service interface cable are required.

The service interface card in the 17M/MA system has the following ports:

- Two 10/100 Ethernet ports
- Two system ports
- Two HMC ports
- Two SPCN ports

Note: For a multiple-drawer server configuration with more than one service interface card, the service interface card in system unit one and system unit two must both be connected to the HMC.

In addition, the 17M/MA building block has one internal SAS controller, redundant hot-swap cooling fans, redundant power supplies, and redundant processor voltage regulators.

Processor card features

Each 17M/MA system unit can contain up to 2-core processor cards with 64-bit, copper-based POWER6 microprocessors running at 3.5, 4.2, or 4.7 GHz.

All processor card features are available only as Capacity on Demand (CoD) features. The initial order of the 17M/MA system must contain the feature code of the processor card, as well as the processor activation feature code.

Table 3 contains the processor card feature codes and processor activation feature codes.

Table 3. Processor card feature codes and processor activation feature codes

Processor card feature code	Description
5620	3.5 GHz POWER6 2-core processor card, 0-core active, 12 DDR2 memory slots. CoD options include: • 5670: 1-way processor activation (permanent) • 5640: Utility CoD (100 processor minutes) • 5642: Utility CoD (one year prepaid) • 5650: CoD (one day billing)
5622	4.2 GHz POWER6 2-core processor card, 0-core active, 12 DDR2 memory slots. CoD options include: • 5672: 1-way processor activation (permanent) • 5641: Utility CoD (100 processor minutes) • 5643: Utility CoD (one year prepaid) • 5653: CoD (one day billing)

Table 3. Processor card feature codes and processor activation feature codes (continued)

Processor card feature code	Description
7380	4.7 GHz POWER6 2-core processor card, 0-core active, 12 DDR2 memory slots. CoD options include: • 5403: 1-way processor activation (permanent) • 5404: Utility CoD (100 processor minutes) • 5408: Utility CoD (one year prepaid) • 5656: CoD (one day billing)

Each processor card features one POWER6 chip with two processor cores and 8MB of L2 cache (each core has a private 4MB L2 cache), 32 MB of L3 cache, and 12 slots of DDR2 memory DIMM technology.

Note: Utility CoD billing for feature codes 5404, 5640, and 5641 provides payment for temporary use of the processor card features for 100 minutes of usage. The purchase of this feature occurs after the customer has 100 minutes of use on processors in the shared processor pool that are not permanently active.

Memory features

The processor cards that are used in the 17M/MA have 12 slots for memory DIMMs.

Table 4 lists the memory feature codes that are available. The 17M/MA system supports CoD options for memory.

Table 4. Memory feature codes

Feature code	Description
5692	0/2GB (4X0.5GB) DIMMS, 667 MHz, DDR2, POWER6™ CoD memory
5693	0/4GB (4X1GB) DIMMS, 667 MHz, DDR2, POWER6 CoD memory
5694	0/8GB (4X2GB) DIMMS, 667 MHz, DDR2, POWER6 CoD memory
5695	0/16GB (4X4GB) DIMMS, 533 MHz, DDR2, POWER6 CoD memory
5696	0/32GB (4X8GB) DIMMS, 400 MHz, DDR2, CoD memory
7954	Activation of on/off memory
5680	Activation of 1GB DDR2 memory for feature codes 5692 through 5696
5681	Activation of 256 GB DDR2 memory
5691	Activation of 1 GB-day on/off memory

Each processor card should have an equal amount of memory to provide balanced memory across the processor cards. This enables memory accesses to be distributed evenly over system components to provide optimal performance.

Disk and media features

Each 17M/MA system has six disk drive bays and one SlimLine media device bay. In a fully configured 17M/MA with four systems, 24 disk drive bays are available, which provide a maximum internal storage capacity of 7.2 TBs. (The minimum configuration includes one 73 GB disk drive.)

Table 5 on page 6 shows the disk drive features that are available.

Table 5. Disk drive feature codes

Feature code	Description
3646	73 GB 15K RPM SAS disk drive
3647	146 GB 15K RPM SAS disk drive
3648	300 GB 15K RPM SAS disk drive

Up to four SlimLine media device bays are available in a fully configured system. Feature code 5629, the optional media enclosure and backplane, is required to support one SlimLine device in each system.

Any combination of the following DVD-ROM and DVD-RAM drives can be installed:

- IDE SlimLine DVD-ROM drive, feature code 5756
- 4.7 GB IDE SlimLine DVD-RAM drive, feature code 5757

Feature code 5629 (the optional media enclosure and backplane), and a DVD-ROM or DVD-RAM device, are required in a system running the Linux® operating system.

Expansion unit features

A single 17M/MA system has four (8x) PCI Express slots and two PCI-X DDR slots. If more PCI-X slots are needed, such as to extend the number of logical partitions, you can attach expansion units.

Up to 20 11D/11 or 11D/20, and up to 32 14G/30 expansion units can be attached to a fully configured 17M/MA system with four system units.

11D/11 expansion unit

Two 11D/11 expansion units fit side-by-side in the 4-EIA units high enclosure (feature code 7311) mounted in a 19-inch rack, such as the 14T/00 or 31T/24

The 11D/11 expansion unit features six PCI-X slots. Only the blind-swap cassettes are supported.

The 11D/11 expansion unit offers a modular growth path for the 17M/MA systems with increasing I/O requirements. A fully configured 17M/MA supports the attachment of up to 20 11D/11 expansion units, and the combined system supports up to 120 PCI-X, eight PCI-X DDR2, and 15 PCI Express adapters.

Note: To attach the 20 11D/11 expansion units to the four 17M/MA systems requires five RIO-2 remote I/O loop adapters (feature code 1800) which blocks one of the PCI Express slots.

The 11D/11 expansion unit has the following attributes:

- 4-EIA units high rack-mountable enclosure (feature code 7311) that can hold one or two 11D/11 expansion units
- Six PCI-X slots: 3.3 V, keyed, 133 MHz blind-swap hot-plug
- Standard redundant hot-plug power and cooling devices
- Two RIO-2 and two SPCN ports

11D/11 expansion unit physical package: Listed below are the physical characteristics of one 11D/11 expansion unit. If you place two expansion units side-by-side, the weight for the two expansion units is also listed:

- Width: 221 mm (8.7 inches)
- Depth: 711 mm (28.0 inches)
- Height: 168 mm (6.6 inches)
- Weight:
  - One expansion unit: 16.8 kg (37 pounds)

– Two expansion units plus the mounting enclosure: 39.1 kg (86 pounds)

11D/20 expansion unit

The 11D/20 expansion unit is a 4-EIA units high full-sized expansion unit that must be mounted in a rack.

The 11D/20 expansion unit offers a modular growth path for the 17M/MA systems with increasing I/O requirements.

A fully configured 17M/MA can have 20 11D/20 expansion units attached. The combined system supports up to 140 PCI-X, eight PCI-X DDR2, and 15 PCI Express adapters.

Note: To attach the 20 11D/20 expansion units to the four 17M/MA systems requires five RIO-2 remote I/O loop adapters (feature code 1800) which blocks one of the PCI Express slots.

PCI-X cards are inserted into the slot from the top of the expansion unit. The adapters are protected by plastic separators, which are designed to prevent grounding and damage when adding or removing adapters.

The 11D/20 expansion unit has the following attributes:

- 4-EIA units high rack-mountable enclosure assembly
- Seven PCI-X slots: 3.3 V, keyed, 133 MHz hot-plug
- 12 hot-swappable SCSI disk drive bays arranged in two 6-packs
- Optional redundant hot-plug power and cooling (feature code 6268)
- Two RIO-2 and two SPCN ports

  Note: The 11D/20 expansion unit initial order, or an existing 11D/20 expansion unit that is migrated from another model, must have the RIO-2 ports available (feature code 6417).

11D/20 expansion unit physical package: The expansion unit has the following physical characteristics:

- Width: 445 mm (17.5 inches)
- Depth: 610 mm (24.0 inches)
- Height: 178 mm (7.0 inches)
- Weight: 45.9 kg (101 pounds)

14G/30 expansion unit

Two 14G/30 expansion units fit side-by-side in the 4-EIA units high enclosure (feature code 7314) mounted in a 19-inch rack, such as the 14T/00 or 14T/42. The 14G/30 expansion unit is designed to be attached to the system unit using the InfiniBand™™ bus and InfiniBand cables.

The 14G/30 expansion unit features six PCI-X DDR slots. Only the blind-swap cassettes are supported.

The 14G/30 expansion unit offers a modular growth path for the 17M/MA systems with increasing I/O requirements. Up to four 14G/30 expansion units can be attached in a loop using the GX Dual-Port 12X Channel Attach adapter (feature code 1802). Two loops for each 17M/MA are supported, allowing up to 32 14G/30 expansion units for a fully configured 17M/MA. The combined system supports up to 200 PCI-X DDR and 15 PCI Express adapters.

Note: To attach the 32 14G/30 expansion units to the four 17M/MA systems requires five GX Dual-Port 12X Channel Attach adapters which blocks one of the PCI Express slots.

The 14G/30 expansion unit has the following attributes:

- 4-EIA units high rack-mountable enclosure (feature code 7314) can hold one or two 14G/30 expansion units

- Six PCI-X DDR 266 MHz adapter slots
- Cassettes can be installed and removed without removing the expansion unit from the rack
- Because the PCI slots support hot-pluggable adapters, adapters can be installed or replaced without turning off the power or removing the covers
- Standard redundant hot-plug power and cooling devices

14G/30 expansion unit physical package: The expansion unit has the following physical characteristics:
- Width: 224 mm (8.8 inches)
- Depth: 800 mm (31.5 inches)
- Height: 172 mm (6.8 inches)
- Weight:
  - One expansion unit: 20 kg (44 pounds)
  - Two expansion units plus the mounting enclosure: 45.9 kg (101 pounds)

31D/24 or 31T/24 expansion unit

The 31D/24 and 31T/24 expansion units provide power, cooling, and Ultra 320 SCSI interface connections for up to 24 LVD Ultra 320 SCSI disk drives.

The 31D/24 and 31T/24 expansion units are available in the following configurations:
- Model 31D/24 - rack-mountable configuration
- Model 31T/24 - stand-alone deskside configuration

The 24 disk drive bays are organized into four independent SCSI groups of six drive bays each. With the use of up to four SCSI repeater cards, you can use either of the following host SCSI bus connection options:
- A single initiator to each SCSI group
- A high-availability dual initiator feature that supports the connection of two adapters to a SCSI group.

The high-availability SCSI connection feature can be used on any or all of the drive groups in the enclosure and together with other drive groups in the enclosure, using the standard connection option. Either model can be set up to use 100-127 V ac or 200-240 V ac.

31D/24 and 31T/24 expansion unit physical packages: The following are the physical characteristics of the31D/24 and 31T/24 expansion units:
- The 31D/24 rack-mountable expansion unit has the following physical characteristics:
  - Width: 447 mm (17.5 inches)
  - Depth: 660 mm (26 inches)
  - Height: 171 mm (6.75 inches)
  - Weight: 54 kg (120 lb.)
- The 31T/24 desk-side model has the following physical characteristics:
  - Width: 305 mm (12.0 inches)
  - Depth: 665 mm (26 inches)
  - Height: 508 mm (20.0 inches)
  - Weight: 66 kg (145 lb.)

PCI adapter slots

Various configurations of I/O expansion units add support for PCI adapter slots for your 17M/MA system.

Table 6 summarizes the maximum number of I/O expansion units that is supported for a 17M/MA system, and the number of PCI adapter slots that are available, when all of the I/O expansion units are the same model type.

Table 6. Maximum number of I/O expansion units supported and total number of PCI adapter slots

17M/MA building block/processor	Maximum number of I/O expansion units	Total number of PCI adapter slots		
		11D/11	11D/20	14G/30
One building block/2-core	4	3 PCI Express 2 PCI-X DDR 24 PCI-X	3 PCI Express 2 PCI-X DDR 28 PCI-X	3 PCI Express 26 PCI-X DDR
One building block/4-core	8	3 PCI Express 2 PCI-X DDR 48 PCI-X	3 PCI Express 2 PCI-X DDR 56 PCI-X	3 PCI Express 50 PCI-X DDR

Hardware Management Console models

The Hardware Management Console (HMC) is a dedicated workstation that provides a graphical user interface for configuring, operating, and performing basic system tasks for your POWER6 processor-based servers.

Table 7 lists the desktop and rack-mountable HMC models available for POWER6 processor-based systems.

Table 7. HMC models available for POWER6 processor-based systems

Type-model	Description
42C/06	42C/06 Desktop Hardware Management Console
42C/R4	42C/R4 Rack-Mounted Hardware Management Console

The 42C/06 is a desktop model that includes one 10/100/1000 Ethernet port; two additional dual-port 10/100/1000 Ethernet adapters can be installed.

The 42C/R4 HMC is a 1-EIA unit high, 19-inch rack-mountable model that has two Ethernet ports and can be extended with one additional two-port 10/100/1000 Gb Ethernet adapter.

One HMC can manage multiple POWER6 processor-based systems. An Ethernet connection is required between the HMC and one of the Ethernet ports on the service processor. Ensure that sufficient Ethernet adapters are available on the HMC to create public and private networks, if you need both.

Two HMCs are recommended in configurations that have high availability requirements. The service processor in the 17M/MA system supports the connection of two HMCs, so there are no additional features needed for an 17M/MA to support a dual HMC environment. The HMCs provide a locking mechanism so that only one HMC at a time has write access to the service processor. In a configuration with multiple systems, the customer is required to provide a switch or hub to connect one HMC to both of the service processors in systems one and two.

Note: When two HMCs are being used for high availability, an Ethernet HUB is required, provided by the customer.

When an HMC is connected to the 17M/MA, the integrated system ports are disabled. If you need serial connections (for example, for HACMP™ heartbeat signals), you must order an additional asynchronous adapter (feature code 5723).

It is a good practice to connect the HMC to the first HMC port on the system, labeled as HMC Port 1, although other network configurations are possible. A second HMC can be attached to HMC Port 2 of the server for redundancy (or vice versa).

The default mechanism for allocation of the IP addresses for the service processor HMC ports is dynamic. The HMC can be configured as a Dynamic Host Configuration Protocol (DHCP) server, providing the IP address at the time the managed server is powered on. If the service processor of the managed server does not receive a DHCP reply before timeout, predefined IP addresses will be set up on both ports. Static IP address allocation is also an option. You can configure the IP address of the service processor ports with a static IP address by using the Advanced System Management Interface (ASMI) menus. See "Service processor" on page 28 for predefined IP addresses and additional information.

Note: If you need to access the ASMI (for example, to set up an IP address of a new POWER6 processor-based server when the HMC is not available or not providing DHCP services), you can connect any PC client to one of the service processor HMC ports with any kind of Ethernet cable, and use a Web browser to access the predefined IP address, such as: https://169.254.2.147

Functions that can be performed using an HMC include:
- Creating and maintaining a multiple logical partition environment
- Displaying a virtual operating system session terminal for each partition
- Displaying a virtual operator panel of contents for each partition
- Detecting, reporting, and storing changes in hardware conditions
- Powering managed systems on and off
- Acting as a service focal point

The HMC provides both a graphical and command-line interface for all management tasks. The command-line interface is also available by using the SSH secure shell connection to the HMC.

Operating system environment

Several operating system environments are available for the POWER6 processor-based systems.

Table 8 displays a summary of the minimum supported operating system levels for the POWER6 processor-based systems.

Table 8. Supported operating systems for POWER6 processor-based systems

Operating system	Version
AIX®	AIX 5L™ Version 5.3 with the 5300-06 Technology Level
Linux (SUSE)	SLES 10 SP1
Virtual I/O Server	1.4

Chapter 2. SAS controllers

In POWER6 processor-based systems, the six drives in the internal hard disk drive enclosure in the system unit use new SAS technology.

Serial Attached SCSI (SAS) architecture defines a serial device interconnection and transportation protocol that defines the rules for information exchange between devices. SAS is an evolution of the parallel SCSI device interface into a serial point-to-point interface.

Benefits of SAS controllers

SAS controllers have a robust SAS expandable architecture that incorporates fibre channel-like functionality (that is, dual path).

Additionally, SAS controllers offer the following benefits:

- An improved signal quality because of a point-to-point connection between device and adapter, or expander
- Improved availability and redundancy, with dual paths to each drive
- Reduced potential customer problems with point-to-point:
  - There is no contention when accessing a drive
  - SAS controllers minimize command time-outs
  - SAS controllers prevent situations where one drive on a bus stops the entire bus
- Performance growth capability
- An improved disk to adapter ratio, providing more addressability: parallel SCSI up to 36, and SAS up to 60
- SAS controllers offer utilization of SCSI commands, providing:
  - Minimal impacts to operating systems
  - Compatibility for High Speed Software (applications)
- SAS controllers ease the ability to identify failing devices

Features of SAS controllers

The SAS controllers are optimized for SAS disk configurations that use dual paths through dual expanders for redundancy and reliability.

Additionally, SAS controllers offer the following features:

- A PCI-X266 system interface or PCI Express system interface
- A physical link speed of 3 Gb per second supporting transfer rates of 300 MB per second
- Support of SAS devices and non-disk Serial Advanced Technology Attachment (SATA) devices
- Manage path redundancy and path switching for multiported SAS devices
- Support for RAID (Redundant Array of Independent Disks)
- Support attachment of other devices, such as non-RAID disks, tape, and optical devices
- RAID disk arrays and non-RAID devices supported as bootable devices

Advanced features of SAS controllers

There are many advanced features included in the SAS controllers.

- Background parity checking

- Background data scrubbing
- Disks formatted to 528 bytes per sector, providing cyclical redundancy checking (CRC) and logically bad block checking
- Optimized skip read/write disk support for transaction workloads
- Supports a maximum of 64 advanced function disks with a total device support maximum of 255 (for example, the number of all physical SAS and SATA devices plus number of logical RAID disk arrays must be fewer than 255 per controller)

Table 9. SAS RAID controller cards

Features	**Card 1**
Custom Card Identification Number (CCIN)	572C
Description	PCI-X266 planar 3Gb SAS
Form factor	Planar integrated
Adapter LED/feature code	2502
Physical links	8
RAID levels supported	0
Write cache size	0
Removable cache card	No
Multi initiator and high availability support	No
Auxiliary write cache (AWC) support	No

Note: AIX supports all of the functions listed in the Table 9. If you are using another operating system, consult the documentation for that operating system regarding support.

SAS physical links are a set of four wires used as two differential signal pairs. One differential signal transmits in one direction while the other differential signal transmits in the opposite direction. Data can be transmitted in both directions simultaneously.

Physical links are contained in ports with each port containing one or more physical link. Each port is a wide port if there is more than one physical link in the port, or a narrow port if there is only one physical link in the port. A port is identified by a unique SAS worldwide name (also called an SAS address).

An SAS controller contains one or more SAS ports. A path is a logical point-to-point link between an SAS initiator port in the controller and an SAS target port in the I/O device (that is, disk). A connection is a temporary association between a controller and I/O device through a path and enables communication to a device. The controller can communicate to the I/O device over this connection using either the SCSI command set or the ATA/ATAPI command set, depending on the device type.

An expander facilitates connections between a controller port and multiple I/O device ports. An expander routes connections between the expander ports. There can exist only a single connection through an expander at any given time. Using expanders creates more nodes in the path from the controller to the I/O device. If an I/O device supports multiple ports, then it is possible for more than one path to the device when there are expander devices on the path. An SAS fabric is the summation of all paths between all controller ports and all I/O device ports in the SAS subsystem.

Note: The external hard disk drive on the POWER6 processor-based systems use the existing SCSI technology.

Disk arrays

RAID technology is used to store data across a group of disks known as a *disk array*. The disk arrays are groups of disks that work together with a specialized array controller to potentially achieve higher data transfer and input and output (I/O) rates than those provided by single large disks. The array controller keeps track of how the data is distributed across the disks.

Depending on the RAID level selected, this storage technique provides the data redundancy required to keep data secure and the system operational. If a disk failure occurs, the disk can usually be replaced without interrupting usual system operation. Disk arrays also have the potential to provide higher data transfer and input and output (I/O) rates than those provided by single large disks.

Each disk array can be used by AIX in the same way that a single non-RAID disk would be used. For example, after creating a disk array, you can create a file system on the disk array or use AIX commands to make the disk array available to the system by adding the disk array to a volume group.

The SAS RAID controller is managed by the SAS Disk Array Manager. The SAS Disk Array Manager serves as the interface to the controller and I/O devices. It also handles the monitoring and recovery features of the controller.

If a disk array will be used as the boot device, you might have to prepare the disks by booting from the diagnostic CD and creating the disk array before installing AIX. You might want to perform this procedure when the original boot drive is part of a disk array. Non-RAID disks formatted to 528 bytes per sector will be automatically put into a single-drive RAID 0 array on system boot. If a single drive RAID 0 array is your boot device, then the single drive RAID 0 array can be installed onto without using the diagnostics CD. Otherwise, use the CD to create the boot drive configuration.

RAID level

The RAID level of a disk array determines how data is stored on the disk array and the level of protection that is provided.

The RAID level supported by the SAS RAID controller has it own attributes and uses a different method of writing data. Currently, RAID level 0 stripes data across the disks in the array, for optimal performance.

RAID level 0 offers a high potential I/O rate, but it is a nonredundant configuration. As a result, there is no redundant data available for the purpose of reconstructing data in the event of a disk failure. There is no error recovery beyond what is usually provided on a single disk. If a physical disk fails in a RAID level 0 disk array, the disk array is marked as *failed*. All data in the array must be backed up regularly to protect against data loss.

Stripe-unit size

With RAID technology, data is striped across an array of physical disks. This data distribution scheme complements the way the operating system requests data.

The granularity at which data is stored on one disk of the array before subsequent data is stored on the next disk of the array is called the stripe-unit size. The collection of stripe units from the first disk of the array to the last disk of the array is called a *stripe*.

You can set the stripe-unit size of an SAS disk array to 16 KB, 64 KB, or 256 KB. You might be able to maximize the performance of your SAS disk array by setting the stripe-unit size to a value that is slightly larger than the size of the average system I/O request. For large system I/O requests, use a stripe-unit size of 256 KB. The optimum strip size is displayed on the screen when you create the disk array.

hdisk and pdisk names

SAS disk arrays are assigned names using the *hdisk* form, in the same way as other disk storage units in AIX.

These names are automatically assigned whenever you create a disk array. The names are deleted when you delete the disk array. The individual physical disks that comprise disk arrays or serve as candidates to be used in disk arrays are represented by *pdisk* names. A *pdisk* is a disk that is formatted to 528 bytes per sector. Disks that are formatted to 512 bytes per sector are assigned names using the *hdisk* form. These disks must be formatted to 528 bytes per sector before they can be used in disk arrays.

The **List SAS Disk Array Configuration** option in the SAS Disk Array Manager can be used to display these *pdisk* and *hdisk* names, along with their associated location codes.

States for disk arrays (hdisks)

The valid states for SAS disk arrays are: *Optimal*, *Degraded*, *Rebuilding*, *Failed*, *Missing*, and *Unknown*.

Degraded
The array's protection against disk failures is degraded or its performance is degraded. When one or more Array Member *pdisks* is in the Failed state, the array is still functional but might no longer be fully protected against disk failures.

Rebuilding
Redundancy data for the array is being reconstructed. After the rebuilding process is complete, the array will return to the Optimal state. Until then, the array is not fully protected against disk failures.

Failed The array is no longer accessible because of disk failures or configuration problems.

Missing
A previously configured disk array no longer exists.

Unknown
The state of the disk array could not be determined.

States for physical disks (pdisks)

The valid states for *pdisks* are: *Active*, *RWProtected*, *Failed*, *Missing*, and *Unknown*.

Active The disk is functioning correctly.

RWProtected
The disk is unavailable because of a hardware or a configuration problem.

Failed
The controller cannot communicate with the disk, or the *pdisk* is the cause of the disk array being in a Degraded state.

Missing
The disk was previously connected to the controller but is no longer detected.

Unknown
The state of the disk could not be determined.

pdisk desriptions

The fourth column in the preceding output is a description of the device. For an array, the description indicates the RAID level of the array. The description of a *pdisk* indicates whether the disk is configured as an *Array Member*, *Hot Spare*, or an *Array Candidate*.

Array Member
A 528 bytes per sector pdisk that is configured as a member of an array.

Array Candidate
A 528 bytes per sector pdisk that is a candidate for becoming an Array Member or a Hot Spare.

Chapter 3. Advanced POWER Virtualization

Learn about the components of the Advanced POWER Virtualization hardware feature, including Micro-Partitioning and the Virtual I/O Server.

Advanced POWER Virtualization is a hardware feature that you can purchase to enhance the vitalization capabilities of your system. In general, the Advanced POWER Virtualization feature includes the Virtual I/O Server and the enablement for Micro-Partitioning.

When you specify the Advanced POWER Virtualization hardware feature with the initial system order, the firmware is activated to support Micro-Partitioning and the Virtual I/O Server. For upgrade orders, a key similar to the Capacity on Demand key is included to enable the firmware. The Virtual I/O Server is a licensed software component of the Advanced POWER Virtualization feature. It contains one charge unit per activated processor, including software maintenance.

Virtual I/O Server

Learn the concepts of the Virtual I/O Server and its primary components.

The Virtual I/O Server is software that is located in a logical partition. This software facilitates the sharing of physical I/O resources between AIX and Linux client logical partitions within the server. The Virtual I/O Server provides virtual SCSI target and Shared Ethernet Adapter capability to client logical partitions within the system, allowing the client logical partitions to share SCSI devices and Ethernet adapters. The Virtual I/O Server software requires that the logical partition be dedicated solely for its use.

The Virtual I/O Server is available as part of the Advanced POWER Virtualization hardware feature.

Using the Virtual I/O Server facilitates the following functions:
- Sharing of physical resources between logical partitions on the system
- Creating logical partitions without requiring additional physical I/O resources
- Creating more logical partitions than there are I/O slots or physical devices available with the ability for partitions to have dedicated I/O, virtual I/O, or both
- Maximizing use of physical resources on the system
- Helping to reduce the Storage Area Network (SAN) infrastructure

The Virtual I/O Server supports client logical partitions running the following operating systems:
- AIX 5.3 or later
- SUSE Linux Enterprise Server 9 (or later)

The Virtual I/O Server comprises the following primary components:
- Virtual SCSI
- Virtual Networking

The following sections provide a brief overview of each of these components.

Virtual SCSI

Physical adapters with attached disks or optical devices on the Virtual I/O Server logical partition can be shared by one or more client logical partitions. The Virtual I/O Server offers a local storage subsystem that provides standard SCSI-compliant logical unit numbers (LUNs). The Virtual I/O Server can export a pool of heterogeneous physical storage as an homogeneous pool of block storage in the form of SCSI disks.

Unlike typical storage subsystems that are physically located in the SAN, the SCSI devices exported by the Virtual I/O Server are limited to the domain within the server. Although the SCSI LUNs are SCSI compliant, they might not meet the needs of all applications, particularly those that exist in a distributed environment.

The following SCSI peripheral-device types are supported:
- Disks backed by a logical volume
- Disks backed by a physical volume
- Optical devices (DVD-RAM and DVD-ROM)

Virtual networking

Shared Ethernet Adapter allows logical partitions on the virtual local area network (VLAN) to share access to a physical Ethernet adapter and to communicate with systems and partitions outside the server. This function enables logical partitions on the internal VLAN to share the VLAN with standalone servers.

Micro-Partitioning

Micro-Partitioning allows multiple logical partitions to share the system's processing power. Use this topic to learn more about Micro-Partitioning and how it functions in a virtual computing environment.

Micro-Partitioning enables you to allocate processors to logical partitions in increments of .1. For example, one partition might have .6 of a processor, while another partition might have 1.4 processors. Such partitions are referred to as *shared processor partitions*. You can choose between dedicated processor partitions and shared processor partitions using Micro-Partitioning.

Micro-Partitioning allows for increased overall use of system resources by automatically applying only the required amount of processor resource needed by each partition. You can configure the POWER hypervisor to continually adjust the amount of processor capacity that is allocated to each shared processor partition based on workload. Tuning parameters provide the system administrator with extensive control over the amount of processor resources that each partition can use.

Chapter 4. RAS and manageability

Many design features help lower the total cost of ownership (TCO) of the POWER6 processor-based systems. The Reliability, Availability, and Serviceability (RAS) technology allows the possibility to improve your TCO architecture by reducing unplanned downtime.

Reliability, Availability, and Serviceability

Reliability, Availability, and Serviceability (RAS) features help to ensure that the system operates when required, performs reliably, and handles failures in an efficient manner.

The POWER6 processor-based systems feature mainframe-inspired RAS features.

Reliability

Highly reliable systems are built with highly reliable components. On POWER6 processor-based systems, this basic premise is expanded upon with a clear *design for reliability* architecture and methodology.

A concentrated, systematic, architecture-based approach is designed to improve overall system reliability with each successive generation of system offerings.

Designed for Reliability

Systems designed with fewer components and interconnects have fewer opportunities to fail. Simple design choices such as integrating two processor cores on a single POWER chip can dramatically reduce the opportunity for system failures. In this case, a 16-core server will include half as many processor chips (and chip socket interfaces) as with a single CPU per processor design. Not only does this reduce the total number of system components, it reduces the total amount of heat generated in the design, resulting in an additional reduction in required power and cooling components.

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, that is, large decreases in component reliability are directly correlated with relatively small increases in temperature, POWER6 processor-based systems are carefully packaged to ensure adequate cooling. Critical system components such as the POWER6 processor chips are positioned on printed circuit cards so they receive fresh air during operation. In addition, POWER6 processor-based systems are built with redundant, variable-speed fans that can automatically increase output to compensate for increased heat in the central electronic complex.

Redundant components and concurrent repair

High-opportunity components, or those that most affect system availability, are protected with redundancy and the ability to be repaired concurrently.

Continuous monitoring

Aided by the First Failure Data Capture (FFDC) methodology and the associated error reporting strategy, commodity managers build an accurate profile of the types of failures that might occur, and initiate programs to enable corrective actions.

The support team also continually analyzes critical system faults, testing to determine if system firmware and maintenance procedures and tools are effectively handling and recording faults as designed.

Availability

The POWER6 processor-based systems include many availability strategy features.

Detecting and deallocating failing components

Runtime correctable or recoverable errors are monitored to determine if there is a pattern of errors. If these components reach a predefined error limit, the service processor initiates an action to deconfigure the ″faulty″ hardware, helping to avoid a potential system outage and to enhance system availability.

To detect and deallocate failing components, the following features are used:

Persistent deallocation

To enhance system availability, a component that is identified for deallocation or deconfiguration on a POWER6 processor-based system is flagged for persistent deallocation. Component removal can occur either dynamically (while the system is running) or at boot-time (IPL), depending both on the type of fault and when the fault is detected.

In addition, runtime unrecoverable hardware faults can be deconfigured from the system after the first occurrence. The system can be rebooted immediately after failure and resume operation on the remaining stable hardware. This prevents the same *faulty* hardware from affecting system operation again, while the repair action is deferred to a more convenient, less critical time.

Dynamic processor deallocation

Dynamic processor deallocation enables automatic deconfiguration of processor cores when patterns of recoverable errors, for example correctable errors on processor caches, are detected. Dynamic processor deallocation can prevent a recoverable error from escalating to an unrecoverable system error (which might otherwise result in an unscheduled server outage). Dynamic processor deallocation relies upon the service processor's ability to use First Failure Data Capture (FFDC)-generated recoverable error information to notify the POWER hypervisor when a processor core reaches its predefined error limit.

- In a shared processor logical partitioning environment, the POWER hypervisor in conjunction with the operating system will drain the run-queue for the failing core, redistribute the work to the remaining CPUs, deallocate the offending CPU, and continue normal operation, although potentially at a lower level of system performance.
- In dedicated processor logical partitioning environment, the platform can request deallocation of the processor from the operating system.

The logical partitioning strategy also enables additional system availability improvements, allowing any processor to be shared with any logical partition on the system.

POWER6 Processor Instruction Retry

POWER6 processor-based systems include a new suite of mainframe-inspired Processor Instruction Retry features that significantly reduces situations that could result in checkstop. The POWER6 processor recovery occurs in the following order:

1. Automatically retry a failed instruction and continue with the task.
2. Interrupt a repeatedly-failing instruction and move it to a new processor and continue with the task.
3. In the event that spare capacity is not found, use the predefined logical partition availability priority list that was created so capacity is obtained from lower demand logical partitions. For example, capacity could be first obtained from a test environment instead of a financial accounting system.
4. When other recovery methods fail, try to contain the termination to the logical partition that is using the faulty core at that instruction.

Memory protection

Memory and cache arrays are comprised of data ″bit lines″ that feed into a memory word. A memory word is addressed by the system as a single element. Depending on the size and addressability of the memory element, each data bit line might include thousands of individual bits (memory cells). For example:

- A single memory module on a memory dual inline memory module (DIMM) can have a capacity of 1 Gb, and supply 8 *bit lines* of data for an ECC word. In this case, each bit line in the ECC word holds 128 Mb behind it, corresponding to more than 128 million memory cell addresses.
- A 32KB L1 cache with a 16-byte memory word, on the other hand, would only have 2 Kb behind each memory bit line.

A memory protection architecture that provides good error resilience for a relatively small L1 cache might be very inadequate for protecting the much larger system main store. Therefore, a variety of different protection methods are used in POWER6 processor-based systems to avoid uncorrectable errors in memory. Memory protection plans must take into account many factors, including size, desired performance, and memory array manufacturing characteristics.

POWER6 processor-based systems have a number of protection schemes designed to prevent, protect, or limit the effect of errors in main memory. These capabilities include:

Hardware scrubbing
: Hardware scrubbing is a method used to deal with transient or soft errors. POWER6 processor-based systems periodically address all memory locations and any memory locations with an ECC error are rewritten with the correct data.

Error correcting code (ECC)
: Error correcting code (ECC) allows a system to detect up to two errors in a memory word and correct one of them. However, without additional correction techniques if more than one bit is corrupted, a system will fail. For example, a burst error (sequential bad bits) or DRAM failure is not tolerated by a system that exclusively uses ECC. For this reason, Chipkill™ memory is used.

Chipkill
: Chipkill is an enhancement to ECC that enables a system to sustain the failure of an entire DRAM. Chipkill spreads the bit lines from a DRAM over multiple ECC words, so that a catastrophic DRAM failure would affect at most one bit in each word. Barring a future single bit error, the system can continue indefinitely in this state with no performance degradation until the failed DIMM can be replaced. To avoid this scenario, POWER6 processor-based systems use a technology called *redundant bit steering*.

Redundant bit steering
: systems use redundant bit steering as a means of avoiding situations where multiple single-bit errors align to create a multi-bit error. In the event that an POWER6 processor-based system detects an abnormal number of errors on a bit line, it can dynamically ″steer″ the data stored at this bit line into one of a number of spare lines. This both reduces exposure to multi-bit errors, and helps to defer maintenance until all redundant bits have been used.

Handling uncorrectable errors

Occasionally an uncorrectable data error can occur in memory or cache. When this happens, the POWER6 processor-based system attempts to limit the impact to the least possible disruption, using a strategy that first considers the data source.

Sometimes an uncorrectable error is transient in nature and occurs in data that can be recovered from another repository. In cases where the data cannot be recovered from an other source, a technique called Special Uncorrectable Error handling is used to determine whether the corruption is a threat to the system. Many times the data is not needed and can be written over, the error condition is voided, and the system will continue to operate normally.

When an uncorrectable error is detected, the system modifies the associated ECC word, thereby signaling to the rest of the system that the ″standard″ ECC is no longer valid. The service processor is then notified and takes appropriate actions. If you are using AIX 5L Version 5.2 or later or Linux, and a process attempts to use the data, the operating system is notified of the error and the operating system will terminate only the specific user program. It is only in the case where the corrupt data is used by the POWER hypervisor that the entire system must be rebooted, thereby preserving overall system integrity.

Cache protection mechanisms

POWER6 processor-based systems are designed with cache protection mechanisms, including cache line delete in both L2 and L3 arrays, processor instruction retry and alternate processor recovery protection on L1-I and L1-D, and redundant *repair* bits in L1-I, L1-D, and L2 caches, as well as L2 and L3 directories.

PCI error recovery

PCI adapters are generally complex designs involving extensive *on-board* instruction processing, often on embedded micro-controllers.

PCI adapters can account for a large portion of the hardware-based error opportunity on a large server. While servers that rely only on *boot time* diagnostics can identify failing components to be replaced by *hot-swap* and reconfiguration, runtime errors pose a more significant problem.

The traditional means of handling problems is through adapter internal error reporting and recovery techniques in combination with operating system device drive management and diagnostics. On older designs, an error in the adapter might cause transmission of bad data on the PCI bus itself, resulting in a hardware-detected parity error that causes a global machine check interrupt, and eventually requiring a system reboot.

A methodology that uses a combination of system firmware enablement and Extended Error Handling (EEH) device drivers allows recovery from intermittent PCI bus errors. This approach works by recovering and resetting the adapter, thereby initiating system recovery for a permanent PCI bus error. Rather than failing immediately, the faulty device is *frozen* and restarted, preventing a machine check. For the POWER6 processor-based systems, this capability has been extended to PCI Express bus errors, and includes expanded Linux support for EEH as well.

Serviceability

The POWER6 serviceability strategy evolves from the service architecture deployed on POWER5 processor-based systems.

The serviceability system package incorporates:
- Easy access to service components
- On-demand service education
- An automated guided repair strategy that uses common service interfaces for a converged service approach across multiple server platforms

Customer control of the service environment extends to firmware maintenance on all of the POWER6 processor-based systems. When taken together, these factors can deliver increased value to the end user. This strategy contributes to higher systems availability with reduced maintenance costs.

Detecting errors

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

While not all errors are a threat to system availability, those that go undetected can be problematic because the system does not have the opportunity to evaluate and act if necessary.

There are many features included in the POWER6 processor-based systems that aid in detecting errors, including:

Error checkers

POWER6 processor-based systems contain specialized hardware detection circuitry, utilized to detect erroneous hardware operations. Error checking hardware ranges from parity error detection coupled with processor instruction retry and bus retry, to ECC correction on caches and system buses. All hardware error checkers have distinct attributes:

- They continually monitor system operations to detect potential calculation errors
- They attempt to isolate physical faults based on runtime detection of each unique failure
- They can initiate a wide variety of recovery mechanisms designed to correct the problem. POWER6 processor-based systems include extensive hardware and firmware recovery logic.

Fault isolation

Error checker signals are captured and stored in hardware Fault Isolation Registers (FIRs). Associated circuitry is used to limit the domain of an error to the first checker that encounters the error. In this way, run time error diagnostics can be deterministic such that for every check station, the unique error domain for that checker is defined and documented. Ultimately, the error domain becomes the field replaceable unit (FRU) callout, and manual interpretation of the data is not usually required.

First Failure Data Capture

First Failure Data Capture (FFDC) is an isolation technique that ensures that when a fault is detected in a system through error checkers or other types of detection methods, the root cause of the fault will be captured without the need to re-create the problem or run extended tracing or a diagnostics program.

For the vast majority of faults, a good FFDC design means that the root cause will be detected automatically without the intervention of a service representative. Pertinent error data related to the fault is captured and saved for analysis. In hardware, FFDC data is collected from the fault isolation registers. In firmware, this data consists of return codes, function calls, and so on.

FFDC ″check stations″ are carefully positioned within the server logic and data paths to ensure that potential errors can be quickly identified and accurately tracked to a field replaceable unit (FRU).

Fault isolation

The service processor interprets error data captured by the FFDC checkers in order to determine the root cause of an error event.

Root cause analysis might indicate that the event is recoverable, meaning that a service action point or need for repair has not been reached. Alternatively, it might indicate that a service action point has been reached, where the event exceeded a predetermined threshold or was unrecoverable. Based upon the isolation analysis, recoverable error threshold counts might be incremented. If the event is recoverable, then a service action might not be necessary.

If the event is deemed to require a service action, additional information will be collected to service the fault. For unrecoverable errors or for recoverable events that meet or exceed their service threshold, meaning a service action points has been reached, a request for service will be initiated though an error logging component.

Diagnosing problems

Using the extensive network of advanced and complementary error detection logic built into the hardware, firmware, and operating systems, POWER6 processor-based systems can perform considerable self diagnosis.

Boot-time

When you start a POWER6 processor-based system, the service processor initializes system hardware. Boot-time diagnostic testing uses a multi-tier approach for system validation, starting with managed low-level diagnostics supplemented with system firmware initialization and configuration of I/O hardware, followed by operating system-initiated software test routines. Boot-time diagnostic routines include:

Built-in self tests (BISTs)

Built-in self tests (BISTs) for both logic components and arrays ensure the internal integrity of components. Because the service processor assists in performing these tests, the system is enabled to perform fault determination isolation whether the system processor is operational or not. The boot-time BISTS might also find faults not detectable by processor-based power-on self test (POST) or diagnostics.

Wire tests
Wire tests discover and precisely identify connection faults between components such as processors, memory, or I/O hub chips.

Initialization of components
During the initialization of components, processes can take place that help aid in isolation of future errors. For example, when ECC memory starts, it writes patterns of data and allows the server to store valid ECC data for each location, which can help isolate errors.

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

Run time
All POWER6 processor-based systems can monitor critical system components during run time, and take corrective actions when recoverable faults occur. The hardware error check architecture provides the ability to report non-critical errors through the service processor to the HMC.

Device drivers
In certain cases, diagnostics are best performed by operating system-specific drivers, most notably I/O devices that are owned directly by a logical partition. In these cases, the operating system device driver will often work in conjunction with I/O device microcode to isolate and or recover from problems. Potential problems are reported to an operating system device driver, which logs the error. I/O devices also might include specific exercisers that can be invoked by the diagnostic facilities for problem recreation if required by service procedures.

Reporting problems

After diagnosing an error, POWER6 processor-based systems report the error through a number of mechanisms. This situation ensures that appropriate entities are aware that the system might be operating in an error state.

However, a crucial piece of solid reporting strategy is ensuring that a single error communicated through multiple error paths is correctly aggregated, so that subsequent notifications are not inadvertently duplicated.

Error logging and analysis
After the root cause of an error has been identified by a fault isolation component, an error log entry is created with some basic data, such as:

- An error code uniquely describing 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
- First Failure Data Capture (FFDC) data

Data that contains information on the effect that the repair will have on the system is also included. Error log routines in the operating system can then use this information and decide whether to contact service and support, send a notification only, or continue without alert.

Remote support
The Remote Management and Control (RMC) application is delivered as part of the base operating system, including the operating system on the Hardware Management Console (HMC). RMC provides a secure transport mechanism across the LAN interface between the operating

system and the HMC and is used by the operating system diagnostic application for transmitting error information. It performs a number of other functions as well, but these are not used for the service infrastructure.

Manage serviceable events using the HMC

A critical requirement in a logically partitioned environment is to ensure that errors are not lost before being reported for service, and that errors are only reported once, regardless of how many logical partitions experience the potential effect of the error. The *Manage Serviceable Events* task on the HMC is responsible for aggregating duplicate error reports, and ensures that all errors are recorded for review and management.

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

The first occurrence of each failure type will be recorded in the *Manage Serviceable Events* task on the HMC. The *Manage Serviceable Events* task will filter and maintain a history of duplicate reports from other logical partitions or the service processor. It then looks across all active service event requests, analyzes the failure to ascertain the root cause and, if enabled, contacts the support organization for service. This methodology ensures that all platform errors will be reported through at least one functional path, ultimately resulting in a single notification for a single problem.

Extended Error Data

Extended error data (EED) is additional data that is collected either automatically at the time of a failure or manually at a later time. The data collected is dependent on the invocation method, but includes information such as firmware levels, operating system levels, additional fault isolation register values, recoverable error threshold register values, system status, and any other pertinent data.

The data is formatted and prepared for transmission back to to assist the service and support organization with preparing a service action plan for the service provider or for additional analysis.

Handling system dumps

In some circumstances, an error might require a dump to be automatically or manually created. In this event, it will be unloaded to the HMC upon reboot. Specific HMC information is included as part of the information that can optionally be sent to support for analysis. If additional information relating to the dump is required, or if it becomes necessary to view the dump remotely, the dump record will contain information that will allow the support center to identify which HMC the dump is located on.

Notifying the appropriate contacts

After an POWER6 processor-based system has detected, diagnosed, and reported an error to an appropriate aggregation point, it then takes steps to notify the customer, and if necessary, the support organization.

Depending upon the assessed severity of the error and the support agreement, this notification could range from a simple notification to a dispatch of a service representative automatically to the customer site with the correct replacement part.

Customer-notify

When an event is important enough to report, but does not indicate a need for a repair action or the need to contact service and support, it is classified as *customer notify*. Customers 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. Examples include:

- Network events, for example, a loss of contact over a local area network (LAN)
- Environmental events, for example, a temperature warning
- Events that need further examination by the customer, but not necessarily require a part or repair action

Customer notify events are serviceable events by definition because they indicate that something has happened that requires customer awareness, in the event they want to take further action. These events can always be reported back to at the customer's discretion.

Contacting the service and support organization
A correctly configured system can contact the service and support organization to initiate an automatic or manual call from a customer location. It can include error data, server status, or other service-related information.

This action invokes the service organization for the appropriate service action to begin, automatically opening a problem report, and in some cases also dispatching field support.

Automated reporting provides faster and potentially more accurate transmittal of error information. While configuring call home is optional, customers are strongly encouraged to configure this feature in order to obtain the full value of service enhancements.

Vital Product Data (VPD) and inventory management
POWER6 processor-based systems store vital product data (VPD) internally, which keeps a record of how much memory is installed, how many processors are installed, manufacturing level of the parts, and so on. These records provide valuable information that can be used by remote support and service representatives, enabling them to provide assistance in keeping the firmware and software on the server up-to-date.

problem management database
At the support center, historical problem data is entered into the service and support problem management database. All of the information related to the error, along with any service actions taken by the service provider are recorded for problem management by the support and development organizations. The problem is then tracked and monitored until the system fault is repaired.

Locating and repairing the problem

The final component of a comprehensive design for serviceability is the ability to effectively locate and replace parts that require service. POWER6 processor-based systems use a combination of visual cues and guided maintenance procedures to ensure that the identified part is replaced correctly.

Guiding Light
Guiding Light is a system that uses a constellation of LEDs, allowing a service provider to quickly and easily identify the location of system components. Because some customer configurations are very complex, Guiding Light is capable of handling multiple error conditions simultaneously.

In the Guiding Light LED implementation, when a fault condition is detected on a POWER6 processor-based system an amber system attention LED will be illuminated. The service provider can engage the *identify* mode by selecting a specific problem. Guiding Light identifies the part that needs to be replaced by flashing the amber *identify* LED.

Operator panel
The operator panel on POWER6 processor-based system is a 4 X 16 element LCD display used to present boot progress codes, indicating advancement through the system power-on and initialization processes. The operator panel is also used to display error and location codes when an error occurs that prevents the system from booting. It includes several buttons allowing a service representative or customer to change various boot-time options, and perform a subset of the service functions that are available on the Advanced System Management (ASM) interface.

Concurrent maintenance
POWER6 processor-based systems are designed with the understanding that certain components

have higher intrinsic failure rates than others. The movement of fans, power supplies, and physical storage devices naturally makes them more susceptible to wear or stress, while other devices such as I/O adapters might begin to wear from repeated plugging or unplugging. For this reason, when correctly configured, these devices are specifically designed to be concurrently maintainable.

In other cases, a customer might be in the process of moving or redesigning a datacenter, or planning a major upgrade. At times like these, flexibility is crucial.POWER6 processor-based systems are designed for redundant and, or concurrently maintainable power, fans, physical storage, and I/O towers.

Blind-swap PCI adapters

Blind-swap PCI adapters represent significant service and ease-of-use enhancements in I/O subsystem design while maintaining high PCI adapter density.

Standard PCI designs supporting *hot-add* and *hot-replace* require top access so that adapters can be slid into the PCI I/O slots vertically. *Blind-swap* allows PCI adapters to be concurrently replaced without having to put the I/O expansion unit into a service position.

Firmware updates

Firmware on the POWER6 processor-based servers is released in a cumulative sequential fix format, packaged as an RPM file for concurrent application and activation. Administrators can install and activate many firmware patches without cycling power or rebooting the server.

The new firmware image is loaded on the HMC using any of the following methods:

- Media such as a CD-ROM
- A problem Fix distribution from the Service and Support repository
- Download from the Web site
- FTP from another server

supports multiple firmware releases in the field, so under expected circumstances a server can operate on an existing firmware release, using concurrent firmware fixes to stay up-to-date with the current patch level. Because changes to some server functions (for example, changing initialization values for chip controls) cannot occur during system operation, a patch in this area will require a system reboot for activation.

Activation of new firmware functions (as opposed to patches) will require the installation of a new firmware release level. This process is disruptive to server operations in that it requires a scheduled outage and full server reboot.

In addition to concurrent and disruptive firmware updates, concurrent patches include functions that are not activated until a subsequent server reboot. A server with these patches will operate normally. Additional fixes will be installed and activated when the system reboots after the next scheduled outage.

Additional capability has been added to the POWER6 firmware to allow viewing of the status of a system power control network background firmware update. This subsystem will update as necessary as migrated nodes or I/O expansion units are added to the configuration. The new firmware will not only provide an interface to view the progress of the update, but also allow starting and stopping the background update if a more convenient time becomes available.

Repair and verify

Repair and verify is a system used to walk a service provider step-by-step through the process of repairing a system and verifying that the problem has been repaired. The steps are customized in the appropriate sequence for the particular repair for the specific system being repaired. Repair scenarios covered by repair and verify include:

- Replacing a defective field replaceable unit (FRU)

- 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 applications after replacing a part
- Adding a new part

Repair and verify procedures are designed to be used both by service providers who are familiar with the task at hand and those who are not. On-demand education content is placed in the procedure at the appropriate location. Throughout the repair and verify 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.

Manageability

Several functions and tools help manageability, and allow you to efficiently and effectively manage your system.

Service processor

The service processor is an embedded controller running the service processor's internal operating system.

The service processor operating system has specific programs and device drivers for the service processor hardware. The host interface is a processor support interface connected to the POWER6 processor. The service processor is always working regardless of main system unit's state. The system unit can be in the following states:

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

The service processor is used to monitor and manage the system hardware resources and devices. The service processor checks the system for errors, ensuring the connection to the HMC for manageability purposes and accepting Advanced System Management Interface (ASMI) SSL network connections. The service processor provides the ability to view and manage the machine-wide settings using the ASMI, and allows complete system and partition management from the HMC.

Note: The service processor enables a system that will not boot to be analyzed. The error log analysis can be performed from either the ASMI or the HMC.

The service processor uses two Ethernet 10/100 Mbps ports:

- Both Ethernet ports are only visible to the service processor and can be used to attach the 17M/MA to an HMC or to access the Advanced System Management (ASM) interface. The ASM interface options can be accessed through an HTTP server that is integrated into the service processor operating environment.
- Both Ethernet ports have a default IP address:
  - Service processor Eth0 or HMC1 port is configured as 169.254.2.147 (This applies to the service processor in drawer 1 or the top drawer.)
  - Service processor Eth1 or HMC2 port is configured as 169.254.3.147 (This applies to the service processor in drawer 1 or the top drawer.)

System diagnostics

The system diagnostics consist of stand-alone diagnostics, which are loaded from the DVD-ROM drive, and online diagnostics (available in AIX).

- Online diagnostics, when installed, are a part of the AIX 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 have access to the AIX error log and the AIX configuration data.
  - **Service mode**, which requires a service mode boot of the system, enables the checking of system devices and features. Service mode provides the most complete checkout of the system resources. All system resources, except the SCSI adapter and the disk drives used for paging, can be tested.
  - **Concurrent mode** enables the normal system functions to continue while selected resources are being checked. Because the system is running in normal operation, some devices might require additional actions by the user or diagnostic application before testing can be done.
  - **Maintenance mode** enables the checking of most system resources. Maintenance mode provides the same test coverage as service mode. The difference between the two modes is the way they are invoked. Maintenance mode requires that all activity on the operating system be stopped. The **shutdown -m** command is used 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.

Electronic Service Agent

Electronic Service Agent, along with the Electronic Services Web site, make up Electronic Services.

Electronic Service Agent automatically monitors and collects hardware problem information and sends this information to the service and support organization. It also can collect information about hardware, software, system configuration, and performance management, which might help the service and support organization assist in diagnosing problems.

Electronic Service Agent is a no-charge software tool that resides on your system to continuously monitor events and periodically send service information to support on a user-definable timetable. This tool tracks and captures service information, hardware error logs, and performance information. It automatically reports hardware error information to support as long as the system is under a maintenance agreement or within the warranty period. Service information and performance information reporting do not require an maintenance agreement.

To access Electronic Service Agent user guides, contact service and support.

To receive maximum coverage, activate Electronic Service Agent on every platform, partition, and Hardware Management Console (HMC) in your network. If your server is managed by an HMC, the HMC will report all hardware problems and the AIX operating system will report only software problems and system information. You must configure the Electronic Service Agent on the HMC in order to have hardware error reporting. The AIX operating system will not report hardware problems for a system managed by an HMC.

Accessing the Electronic Services Web site

The Electronic Services Web site provides the ability to view service information reported by Electronic Service Agent, use the Premium Search function, open and manage service requests, receive support messages by platform or individual, and customize the site to your preferences.

Ensure that you have the appropriate operating system level before installing Electronic Service Agent. You will need AIX 5L Version 5.3 with the 5300-06 Technology Level.

Manage serviceable events with the HMC

Service strategies become more complicated in a partitioned environment. The *Manage Serviceable Events* task in the HMC can help streamline this process.

Each logical partition reports errors it detects, 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 HMC, you can avoid long lists of repetitive call-home information by recognizing that these are repeated errors and correlating them into one error.

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

Hardware user interfaces

In addition to the HMC, other hardware management user interfaces can be used to manage your systems.

Advanced System Management interface

The Advanced System Management (ASM) interface is the interface to the service processor that enables you to manage the operation of the server, such as auto power restart, and to view information about the server, such as the error log and vital product data. Some repair procedures require connection to the ASM interface.

The ASM interface is accessible through the HMC. For details, see "Accessing the ASM interface using an HMC." The ASM interface is also accessible using a Web browser on a system that is connected directly to the service processor (in this case, either a standard Ethernet cable or a crossed cable) or through an Ethernet network. The ASM interface enables the ability to change the service processor IP addresses or to apply some security policies and avoid the access from undesired IP addresses or range.

You might be able to use the service processor's default settings. In that case, accessing the ASM interface is not necessary.

Accessing the ASM interface using an HMC: If configured to do so, the HMC connects directly to the Advanced System Management (ASM) interface for a selected system from this task.

To connect to the Advanced System Management interface from an HMC:
1. Open **Systems Management** from the navigation pane.
2. From the work pane, select one or more managed systems to work with.
3. From the System Management tasks list, select **Operations**.
4. From the Operations task list, select **Advanced System Management (ASM)**.

Accessing the ASMI using a Web browser: The Web interface to the ASMI is accessible through Microsoft® Internet Explorer® 6.0, Netscape 7.1, Mozilla Firefox, or Opera 7.23 running on a PC or mobile computer connected to the service processor. The Web interface is available during all phases of system operation, including the initial program load (IPL) and run time. However, some 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.

Accessing the ASM interface using an ASCII terminal: The ASM interface on an ASCII terminal supports a subset of the functions provided by the Web interface and is available only when the system is in the platform standby state. The ASM interface on an ASCII console is not available during some phases of system operation, such as the IPL and run time.

Graphics terminal

The graphics terminal is available to users who want a graphical user interface (GUI) to their AIX or Linux systems. To use the graphics terminal, plug the graphics adapter into a PCI slot in the back of the server. You can connect a standard monitor, keyboard, and mouse to the adapter to use the terminal. This connection allows you to access the SMS menus, as well as an operating system console.

Appendix. Accessibility features

Accessibility features help users who have a physical disability, such as restricted mobility or limited vision, to use information technology products successfully.

The following list includes the major accessibility features:

- Keyboard-only operation
- Interfaces that are commonly used by screen readers
- Keys that are tactilely discernible and do not activate just by touching them
- Industry-standard devices for ports and connectors
- The attachment of alternative input and output devices

Notices

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

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

The manufacturer may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to the manufacturer.

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

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. The manufacturer 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 Web sites not owned by the manufacturer are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this product and use of those Web sites is at your own risk.

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

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

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

All statements regarding the manufacturer's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

The manufacturer's prices shown are the manufacturer's suggested retail prices, are current and are subject to change without notice. Dealer prices may vary.

This information is for planning purposes only. The information herein is subject to change before the products described become available.

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

If you are viewing this information in softcopy, the photographs and color illustrations may not appear.

The drawings and specifications contained herein shall not be reproduced in whole or in part without the written permission of the manufacturer.

The manufacturer has prepared this information for use with the specific machines indicated. The manufacturer makes no representations that it is suitable for any other purpose.

The manufacturer's computer systems contain mechanisms designed to reduce the possibility of undetected data corruption or loss. This risk, however, cannot be eliminated. Users who experience unplanned outages, system failures, power fluctuations or outages, or component failures must verify the accuracy of operations performed and data saved or transmitted by the system at or near the time of the outage or failure. In addition, users must establish procedures to ensure that there is independent data verification before relying on such data in sensitive or critical operations. Users should periodically check the manufacturer's support websites for updated information and fixes applicable to the system and related software.

Trademarks

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

AIX
Chipkill
Electronic Service Agent
HACMP
Micro-Partitioning
POWER6
ViaVoice

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

InfiniBand and the InfiniBand design marks are trademarks and/or service marks of the InfiniBand Trade Association.

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

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

Terms and conditions

Permissions for the use of these publications is granted subject to the following terms and conditions.

You may not download, export or re-export this information except in full compliance with all applicable laws and regulations, including all United States export laws and regulations.

THE MANUFACTURER MAKES NO GUARANTEE ABOUT THE CONTENT OF THESE PUBLICATIONS. THESE PUBLICATIONS ARE PROVIDED ″AS-IS″ AND WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT, AND FITNESS FOR A PARTICULAR PURPOSE.

Technical publication remarks form

Title :	ESCALA POWER6 Escala Power6 Overview

Reference:	86 A1 32EV 01	**Date:**	November 2007

ERRORS IN PUBLICATION

SUGGESTIONS FOR IMPROVEMENT TO PUBLICATION

Your comments will be promptly investigated by qualified technical personnel and action will be taken as required.
If you require a written reply, please include your complete mailing address below.

NAME : ______________________________ Date : __________

COMPANY : ______________________________

ADDRESS : ______________________________

____

Please give this technical publication remarks form to your BULL representative or mail to:

Bull - Documentation D[ept.]
1 Rue de Provence
BP 208
38432 ECHIROLLES CEDEX
FRANCE
info@frec.bull.fr

Technical publications ordering form

To order additional publications, please fill in a copy of this form and send it via mail to:

BULL CEDOC
357 AVENUE PATTON
B.P.20845
49008 ANGERS CEDEX 01
FRANCE

Phone: +33 (0) 2 41 73 72 66
FAX: +33 (0) 2 41 73 70 66
E-Mail: srv.Duplicopy@bull.net

Reference	Designation	Qty
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
_ _ _ _ _ _ _ _ _ [_ _]		
[_ _] : The latest revision will be provided if no revision number is given.		

NAME: ______________________________ Date: ____________

COMPANY: ______________________________

ADDRESS: ______________________________

____

PHONE: ______________________ FAX: ______________________

E-MAIL: ______________________________

For Bull Subsidiaries:

Identification: ______________________________

For Bull Affiliated Customers:

Customer Code: ______________________________

For Bull Internal Customers:

Budgetary Section: ______________________________

For Others: Please ask your Bull representative.

BULL CEDOC
357 AVENUE PATTON
B.P.20845
49008 ANGERS CEDEX 01
FRANCE

REFERENCE
86 A1 32EV 01