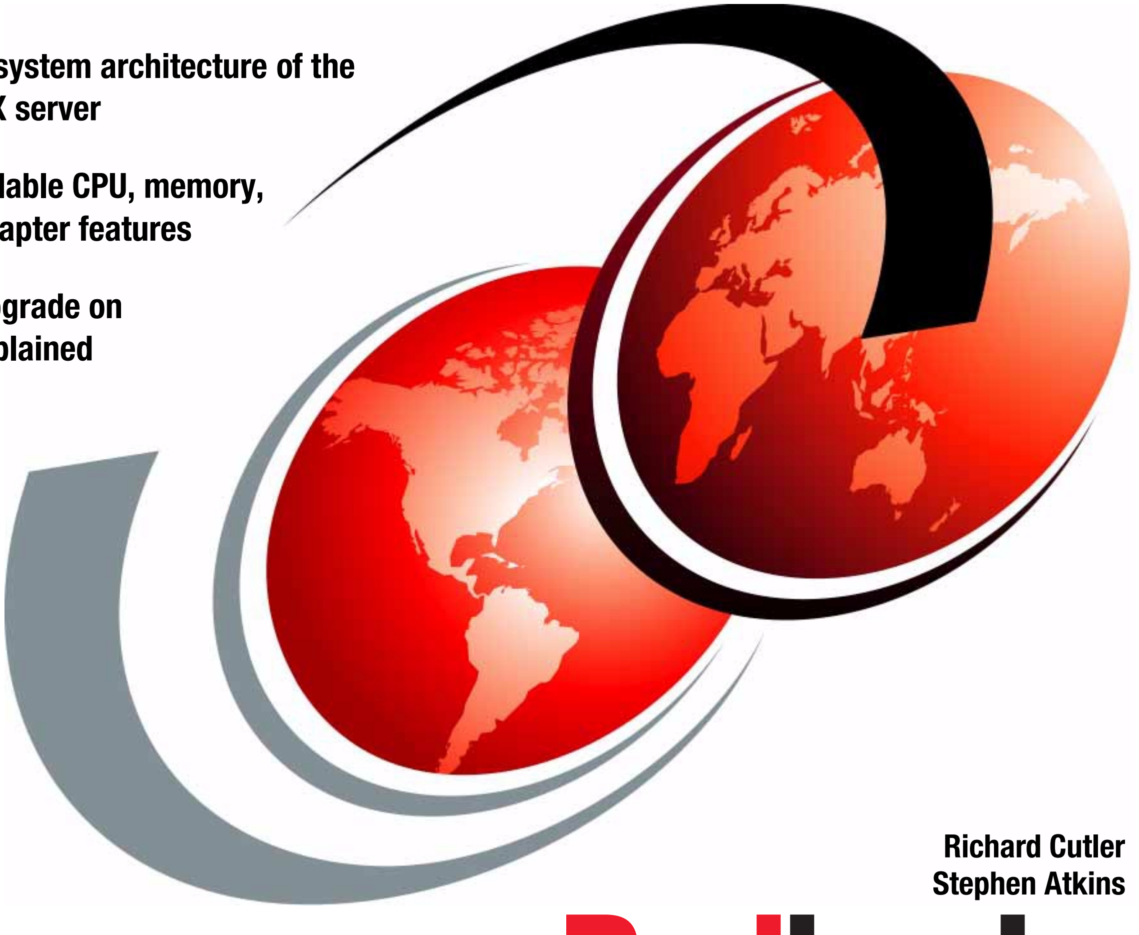
IBM

IBM @server pSeries 680 Handbook

Including RS/6000 Model S80



Details the system architecture of the fastest UNIX server

Covers available CPU, memory, disk and adapter features

Capacity Upgrade on Demand explained

Richard Cutler
Stephen Atkins

Redbooks

ibm.com/redbooks

SG24-6023-00

International Technical Support Organization

IBM @server pSeries 680 Handbook Including RS/6000 Model S80

December 2000

Take Note!

Before using this information and the product it supports, be sure to read the general information in Appendix D, “Special notices” on page 141.

First Edition (December 2000)

This edition applies to the IBM @server pSeries 680 and RS/6000 Model S80 servers for use with the AIX Version 4.3.3 operating system, Program Number 5765-C34.

Comments may be addressed to:
IBM Corporation, International Technical Support Organization
Dept. JN9B Building 003 Internal Zip 2834
11400 Burnet Road
Austin, Texas 78758-3493

When you send information to IBM, you grant IBM a non-exclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Contents

Figures ix

Tables xi

Preface xiii
The team that wrote this redbook xiii
Comments welcome xiv

Chapter 1. Overview of the pSeries 680 and S80 1
1.1 New features 2
1.2 System description 2
1.2.1 Options 3
1.2.2 pSeries 680 specifics 4
1.2.3 Model S80 specifics 6
1.2.4 Common features 7
1.3 Competitive positioning 8
1.3.1 Online transaction processing 9
1.3.2 Oracle Application Standard Benchmark 9
1.3.3 Other Results 9
1.4 AIX Version 4.3 and commercial computing 10
1.4.1 Kernel scalability enhancements 10
1.4.2 AIX Workload Manager 11

Chapter 2. The pSeries 680 13
2.1 Description 13
2.2 Base configuration 14
2.3 Processor 15
2.4 Power 16
2.5 Memory 16
2.6 Disk drives 17
2.7 Boot Devices 18
2.8 SCSI Adapters 18
2.9 I/O Drawers 19
2.10 Cabling 19
2.10.1 System Power Control Network 19
2.10.2 Remote I/O cables 19
2.11 pSeries 680 configurations 20
2.12 Publications 21
2.13 pSeries 680 additional features 22

Chapter 3. The Model S80 25
3.1 Description 25
3.2 Base configuration 26
3.3 Processor 27
3.4 Power 28
3.5 Memory 29
3.6 Disk drives 30
3.7 Boot Devices 30
3.8 SCSI Adapters 31
3.9 I/O Drawers 31
3.10 Cabling 32
3.10.1 System Power Control Network 32
3.10.2 Remote I/O cables 32
3.11 Model S80 configurations 32
3.12 Publications 34
3.13 Model S80 additional features 34

Chapter 4. Remote I/O Subsystem 39
4.1 I/O rack description 39
4.1.1 I/O Drawer description 40
4.2 System Power Control Network (SPCN) 42
4.3 Remote I/O cabling 44

Chapter 5. Using the Configurator 49
5.1 What is the PCRS6000 Configurator? 49
5.2 Configuring pSeries 680 and S80 servers 50
5.2.1 Adapter restrictions 51
5.2.2 Two Servers sharing an I/O rack 52
5.3 High-Availability Solutions 54
5.3.1 Minimum requirements 54
5.3.2 Configuring an HA solution using PCRS6000 55
5.4 Special considerations for SP external node attach 57
5.4.1 Features required on the pSeries 680 and S80 57
5.4.2 Features required on the SP system 58
5.4.3 Features required on the Control Workstation 60
5.5 Clustered Enterprise Servers 60
5.5.1 Features required on the pSeries 680 and S80 61
5.5.2 Configuring CES using PCRS6000 62
5.6 Upgrading an RS/6000 S70 Advanced to a Model S80 63
5.6.1 S70 Advanced to S80 processor conversion 64
5.6.2 S70 Advanced to S80 memory conversion 65

Chapter 6. Hardware architecture 67
6.1 Book packaging 70

6.2 RS64 IV processor and card. 71
6.3 RS64 III processor and card. 73
6.4 Memory controller complex 76
6.5 Memory cards and quads 78
6.6 Hardware Multithreading 80
6.6.1 Limitations 80
6.6.2 Controlling Hardware Multithreading. 82
6.7 Remote I/O connections 82
6.8 I/O Drawer 83
6.9 I/O bridge bus 84
6.10 PCI buses 84

Chapter 7. AIX features for large SMP servers 85
7.1 SMP scalability enhancements. 86
7.1.1 Increased threads and process limits 86
7.1.2 Multiple run queues with load balancing 86
7.1.3 Reduced lock contention 87
7.1.4 Fast device configuration 87
7.2 File system enhancements 87
7.2.1 Online JFS backup 87
7.2.2 Concurrent striping and mirroring 89
7.3 Workload management 91
7.3.1 Overview 91
7.3.2 Managing resources with WLM. 92
7.3.3 Planning for WLM 95
7.3.4 Starting WLM 96
7.3.5 WLM user interfaces. 96
7.4 Capacity Upgrade on Demand 97
7.4.1 Pre-requisites 97
7.4.2 Using the chcod command 98
7.4.3 Additional information. 99

Chapter 8. Reliability, availability, and serviceability 101
8.1 Designed for higher RAS 101
8.2 Features of the pSeries 680 and S80 101
8.2.1 Dynamic error recovery 102
8.3 Hot-swap components 102
8.3.1 Power and cooling 103
8.3.2 Error Recovery for Caches and Memory. 103
8.4 Service processor 104
8.4.1 Introduction 104
8.4.2 Primary functions 105
8.4.3 Firmware updates. 106

8.5 Electronic Service Agent 106
8.6 AIX RAS. 107
8.6.1 AIX Version 4.3.3 enhancements 107
8.6.2 Software service aids 108
8.7 Dynamic CPU de-allocation 108
8.7.1 Potential Impact to Applications 109
8.7.2 Processor De-allocation 110
8.7.3 System Administration 111
8.7.4 Error Log Entries 113

Appendix A. Installation requirements 119
A.1 Central Electronic Complex 119
A.2 I/O rack 120
A.3 10 EIA I/O Drawer. 121
A.4 Physical space requirements 123

Appendix B. A practical guide to the Service Processor 125
B.1 How to access the Service Processor menus. 125
B.1.1 Local access 125
B.1.2 Remote access 126
B.2 Minimum Service Processor configuration 126
B.2.1 Password protection for Service Processor menus 127
B.2.2 Operating system surveillance 128
B.2.3 Configuring call-out. 129
B.2.4 Configuring call-in 132
B.2.5 Power and reboot issues 132
B.2.6 Useful Service Processor utilities 133
B.3 Determining firmware levels 135
B.4 Updating firmware. 136

Appendix C. PCI Dual Channel Ultra2 SCSI Adapter (#6205) 139

Appendix D. Special notices 141

Appendix E. Related publications. 145
E.1 IBM Redbooks 145
E.2 IBM Redbooks collections. 145
E.3 Other resources 145
E.4 Referenced Web sites. 146

How to get IBM Redbooks 147
IBM Redbooks fax order form 148

Abbreviations and acronyms 149

Index 151

IBM Redbooks review 161

Figures

1. IBM @server pSeries 680 5
2. RS/6000 Enterprise Server model S80 6
3. Front view of 10 EIA Drawer 41
4. Rear view of 10 EIA I/O Drawer 42
5. Base SPCN loop with one I/O Drawer 43
6. Fully configured SPCN loop with four I/O Drawers 44
7. RIO loop with one I/O Drawer 45
8. One RIO loop attached to two I/O Drawers 45
9. Two RIO loops attached to two I/O Drawers 46
10. Two RIO loops attached to four I/O Drawers 47
11. The PCRS6000 configurator 50
12. Rack options dialog box 53
13. Starting an HA configuration 55
14. Selecting the HA solution 56
15. HA Solution Validation 57
16. Configuring an SP-attached server 59
17. Identifying a server as a CES machine 62
18. Front view of the pSeries 680 and S80 CEC 67
19. Rear view of the pSeries 680 and S80 CEC 69
20. Book Packaging 70
21. RS64 IV Processor Card 71
22. RS64 III Processor card 74
23. System switch complex 77
24. Memory card placement 79
25. Example share distribution 93
26. Class with independent resource limits 95
27. Main menu for Workload Manager 97
28. pSeries 680 and S80 system service clearances 124

Tables

1. pSeries 680 minimum configuration. 20
2. pSeries 680 system expansion 21
3. Publications shipped with the pSeries 680 21
4. Additional publications available for a fee 21
5. pSeries 680 optional features 22
6. Model S80 minimum configuration. 32
7. Model S80 system expansion 33
8. Publications shipped with the Model S80 34
9. Additional publications available for a fee 34
10. Model S80 optional features 35
11. S70 Advanced to S80 processor conversion. 64
12. S70 Advanced to S80 memory conversion 65
13. Characteristics of different RAID levels 90
14. Central Electronic Complex 119
15. The S00 I/O rack 121
16. The T00 I/O rack 121
17. The T42 I/O rack 121
18. 10 EIA SCSI I/O Drawer 122
19. Privileged and general access passwords. 128

Preface

This redbook covers the IBM @server pSeries 680 and the RS/6000 Enterprise Server Model S80 (hereafter referred to in this redbook as the pSeries 680 and S80 respectively). It will help you understand the architecture of each machine and the similarities and differences between them. An overview of the optional features for each machine is also provided, along with advice on how to use the PCRS6000 Configurator to produce a valid configuration.

This publication is suitable for professionals wishing to acquire a better understanding of the pSeries 680 and S80, including:

- Customers
- Sales and Marketing professionals
- Technical Support professionals
- Business Partners

This publication does not replace the latest marketing materials and tools. It is intended as an additional source of information that, together with existing sources, may be used to enhance your knowledge of IBM Enterprise Server products.

This book is a replacement for the IBM Redbook *RS/6000 S-Series Enterprise Servers Handbook*, SG24-5113.

The team that wrote this redbook

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

Richard Cutler is an AIX and RS/6000 Technical Specialist at the ITSO, Austin Center. Before joining the ITSO, he worked in the RS/6000 Technical Center in the UK, where he assisted customers and independent software vendors with porting their applications to AIX.

Stephen Atkins is a Consulting IT Specialist in EMEA Advanced Technical Support. He has 26 years experience in the IT industry, of which 16 have been spent with IBM. He has seven years experience in AIX and RISC systems. His areas of expertise include symmetric multiprocessors, Scalable Parallel systems, server consolidation and Business Intelligence. He has written extensively on performance, sizing, systems design and systems architecture.

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

IBM Austin
Larry Amy, Margie Blevins, Carolyn Scherrer, Scott Vetter, Wade Wallace

IBM Brazil
Mauro Minomizaki

IBM Dallas
Ron Barker

IBM Rochester
Dave Krolak

Comments welcome

Your comments are important to us!

We want our Redbooks to be as helpful as possible. Please send us your comments about this or other Redbooks in one of the following ways:

- Fax the evaluation form found in “IBM Redbooks review” on page 161 to the fax number shown on the form.
- Use the online evaluation form found at **ibm.com**/redbooks
- Send your comments in an Internet note to redbook@us.ibm.com

Chapter 1. Overview of the pSeries 680 and S80

The pSeries 680 and S80 are 64-bit symmetric multiprocessing (SMP) enterprise servers designed to provide the power, expandability, and reliability needed for the next generation of mission-critical computing.

Both models utilize processors from the RS64 chip family, which incorporates state-of-the art copper technology developed by IBM. The latest version of the RS64 chip, available as an option on the S80 and standard on the pSeries 680, also uses silicon-on-insulator (SOI) technology to deliver increased performance and better reliability.

At the heart of each system is a class-leading data switch architecture, which delivers unrivalled scalability for real-world applications. With 53 PCI slots available for I/O expansion in a typical configuration, and with no compromises on either the number of processors or maximum memory configuration, the pSeries 680 and S80 offer balanced system performance for the most demanding applications.

A new feature of AIX, Capacity Upgrade on Demand, means that inactive processors can be installed on your system to prepare for future growth. They can be enabled quickly and easily as required by your business needs, thus eliminating delays caused by procurement cycles.

Standard availability features include:

- Error checking and correcting (ECC) system memory - including both L2 and L1 data caches.
- A high-availability memory subsystem, which supports *memory scrubbing*, *bit-scattering*, *bit-steering*, and c*hip-kill recovery.*
- An integrated Service Processor for system monitoring.
- Redundant power supplies, power regulators, and blowers.
- I/O link failure recovery.
- Dynamic CPU de-allocation.

For more details on availability features, see Chapter 8, “Reliability, availability, and serviceability” on page 101.

The 64-bit hardware addressing capability of the pSeries 680 and S80 enables current 32-bit applications to run unmodified on systems with up to 96 GB of memory, allowing customers to take advantage of the latest technology without having to recompile applications.

1.1 New features

There are several new features which have been announced for the pSeries 680 and S80 since the publication of the *RS/6000 S-Series Enterprise Server Handbook*, SG24-5113 and which are described in this publication. They are:

- High-availability solutions packages are described in Section 1.2.1.3, “High-availability solutions” on page 3.
- Clustered enterprise servers are described in Section 5.5, “Clustered Enterprise Servers” on page 60.
- The RS64 IV processor is described in Section 6.2, “RS64 IV processor and card” on page 71.
- Hardware multithreading is described in Section 6.6, “Hardware Multithreading” on page 80.
- Support for 96 GB of memory is described in Section 6.5, “Memory cards and quads” on page 78.
- Capacity Upgrade on Demand is described in Section 7.4, “Capacity Upgrade on Demand” on page 97.
- Dynamic CPU de-allocation is described in Section 8.7, “Dynamic CPU de-allocation” on page 108.

1.2 System description

Both the pSeries 680 and S80 have similar physical packaging, although their external appearance is slightly different. In both cases, a single system consists of a system unit enclosure, and up to four I/O drawers housed in up to four 19 inch I/O racks. The system unit, known as the Central Electronic Complex (CEC), houses the data switch, or backplane, together with the processor and memory books. Up to four I/O Drawers can be placed in a single rack, depending on the type of rack selected. The remaining space in the I/O rack can be used for additional peripherals, such as disk and tape subsystems.

The I/O Drawers house the PCI adapters used for connecting the system to peripheral devices. Each I/O Drawer has a total of 14 PCI slots configured across four 33 MHz PCI buses. There are five 64-bit and nine 32-bit slots. Each I/O Drawer also has space for two media devices (tape or CD-ROM) and two SCSI disk bays each holding up to six hot-swap disks, with the first I/O drawer also containing a diskette drive.

The first I/O Drawer in the system is called the primary I/O Drawer and contains the Service Processor. The Service Processor constantly monitors key system hardware components and, if so configured, can detect operating system hangs and force a restart. The Service Processor provides early power-off warnings, has facilities for hardware error analysis and enables local or remote access to the server while the system is off-line. It can be configured to dial IBM Service automatically if the system cannot boot. See Appendix B, “A practical guide to the Service Processor” on page 125 for more details.

The primary I/O Drawer has only 11 PCI slots available for attaching external devices: one 32-bit slot is required for the Service Processor, and two 32-bit slots are used for the Ultra SCSI controllers serving the media bays and the first SCSI disk six-pack (for the boot/paging devices). If a second disk six-pack is installed, for example, for mirroring rootvg, a third Ultra SCSI adapter must be installed, leaving ten for external devices.

The CEC and I/O Drawers are connected by Remote I/O (RIO) and System Power Control Network (SPCN) cables. A system consists of a CEC and between one and four I/O Drawers.

1.2.1 Options

This section describes the configuration options that exist for both the pSeries 680 and S80.

1.2.1.1 SP-attached servers

Up to sixteen pSeries 680 and S80 servers can be attached to an RS/6000 SP, with or without the use of an SP Switch, to provide additional online transaction processing (OLTP) and database capability to the cluster. See Section 5.4, “Special considerations for SP external node attach” on page 57 for more information.

1.2.1.2 Clustered Enterprise Servers

Up to sixteen pSeries 680 and S80 servers may be incorporated into a single cluster managed by the IBM Parallel Systems Support Programs for AIX (PSSP). This implementation requires a 7025-F50 control workstation for cluster control but does not require attachment to a 9076 SP frame. See Section 5.5, “Clustered Enterprise Servers” on page 60 for more information.

1.2.1.3 High-availability solutions

Either system can also be configured for high availability using IBM High Availability Cluster Multi Processing (HACMP) software and redundant hardware components. The HA-S80 Cluster Server and the HA-S85 Cluster Server are two-node, high-availability clusters consisting of:

- Either two Model S80 or two pSeries 680 Servers
- 7133-D40 Serial Disk Systems
- AIX Version 4.3.3 license
- HACMP/ES 4.3.1 high-availability cluster software

The HA-S80 and HA-S85 solutions make a flexible and scalable (vertical and horizontal) framework available to build a highly available cluster system tailored to individual customer needs. See Section 5.3, “High-Availability Solutions” on page 54 for more details.

1.2.1.4 Electronic Service Agent

Additionally, IBM offers a program called Electronic Service Agent (formerly Service Director) to warranty and maintenance customers.

Electronic Service Agent has the following attributes:

- Monitors and analyzes system errors and, if needed, can place a service call to IBM automatically.
- Reduces the effect of business disruptions due to unplanned system outages and failures.
- Performs problem analysis on a subset of hardware-related problems and, with customer authorization, automatically reports the results to IBM Service.

Electronic Service Agent assists customers and IBM Service in achieving higher availability and improved serviceability. See Section 8.5, “Electronic Service Agent” on page 106 for more details.

1.2.1.5 Storage

The pSeries 680 and S80 attach to the latest IBM storage technology products, giving customers a broad choice of tape and disk storage options. Both support advanced communications adapters and provide exceptional performance in online transaction processing, Enterprise Resource Planning (ERP), Business Intelligence (BI), Web serving, and e-commerce.

1.2.2 pSeries 680 specifics

The pSeries 680 base configuration has a 6-way 600 MHz RS64 IV processor card. Each processor has 16 MB of Level 2 (L2) cache. Minimum system memory is 4 GB of SDRAM. Memory must be installed in four-card sets, called quads, which are available in the following sizes: 2 GB (4 x 512 MB), 4 GB (4 x 1024 MB), 8 GB (4 x 2048 MB), 16 GB (4 x 4096 MB), and 32 GB (4 x 8192 MB).

Selectable I/O racks are the 36-EIA T00 and the 42-EIA T42, although existing 32-EIA S00 racks may be used if building high-availability configurations with other 19-inch rack-mounted RS/6000 servers.

The pSeries 680 supports the new AIX performance feature Hardware Multithreading, described in more detail in Section 6.6, “Hardware Multithreading” on page 80.

Also new for the pSeries 680 is a dual line cord feature for the CEC (note that this requires a US/Canada power cord feature).

Figure 1 shows the pSeries 680 with the CEC on the right and a T00 I/O rack on the left.

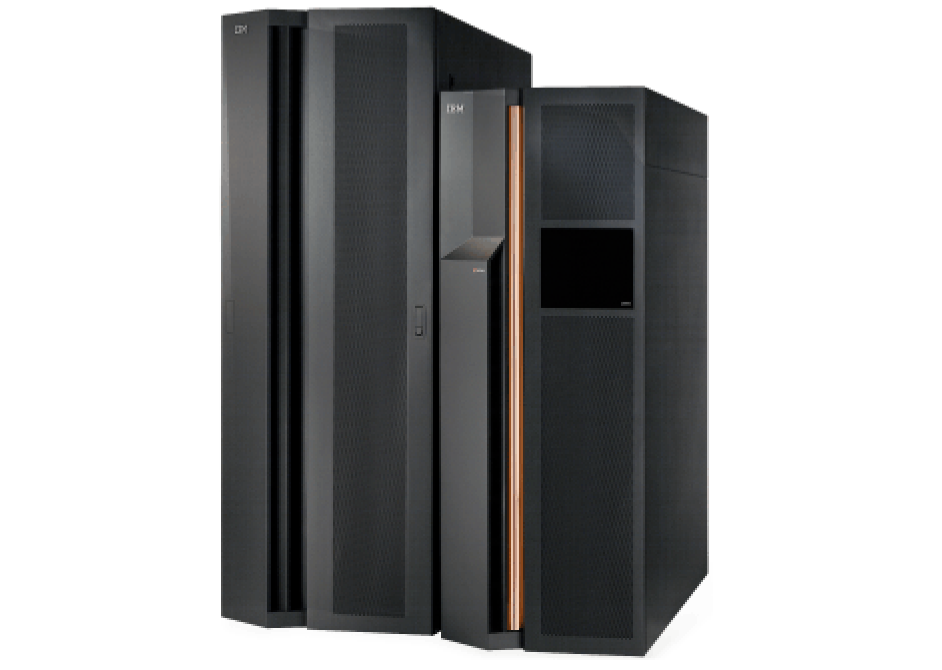


Figure 1. IBM @server pSeries 680

1.2.3 Model S80 specifics

The Model S80 supports both 450 MHz and 600 MHz processors. All processors in the system must be of the same type. New systems can only be ordered with 450 MHz processors (the 600 MHz processors are only available for upgrades to existing systems). Customers requiring 600 MHz processors for new systems should order the pSeries 680.

The base configuration has a 6-way 450 MHz RS64 III processor card. Each processor has 8 MB of Level 2 (L2) cache. Minimum system memory is 2 GB of SDRAM. Memory must be installed in four-card sets, called quads, which are available in the following sizes: 1 GB (4 x 256 MB), 2 GB (4 x 512 MB), 4 GB (4 x 1024 MB), 8 GB (4 x 2048 MB), 16 GB (4 x 4096 MB) and 32 GB (4x 8192 MB).

Selectable I/O racks are the 32-EIA S00, the 36-EIA T00, and the 42-EIA T42.

Figure 2 shows the Model S80 with the CEC on the right and an S00 I/O rack on the left.

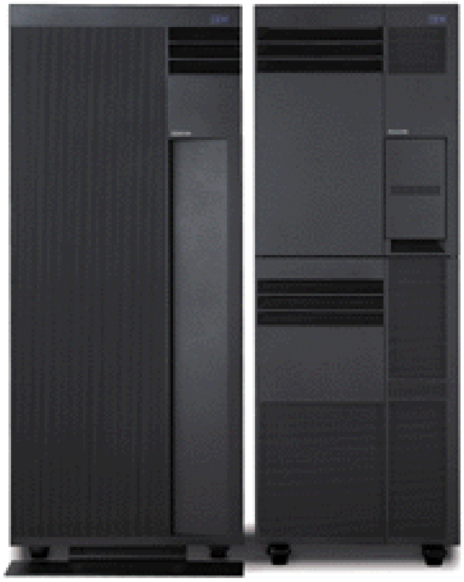


Figure 2. RS/6000 Enterprise Server model S80

1.2.4 Common features

The base configuration for both the pSeries 680 and S80 includes six processors. A 6-way system can be expanded to a 12-, 18-, or 24-way system, with additional processor cards, called books, each containing six processors.

Memory can be increased to 96 GB (this requires maintenance level 433-06 of AIX Version 4.3.3). There are 16 memory book slots, allowing a total of four quads to be installed. Memory cards must be identical within each quad.

The I/O rack holds the primary I/O Drawer, which contains:

- Service processor
- High-performance disk drive
- SCSI-2 CD-ROM
- 1.44 MB 3.5-inch diskette drive
- Two PCI Ultra SCSI controllers

Up to three additional I/O Drawers can be added. Each drawer provides 14 PCI slots across four independent buses. Each drawer may contain two Ultra SCSI hot-pluggable disk 6-packs. Existing RS/6000 racks can also be used for additional rack-mounted storage and communication devices.

A fully-configured system consists of 24 processors, 96 GB of system memory, 53 available PCI adapter slots (the Service Processor and two Ultra SCSI controllers take up three slots in the first I/O Drawer), 48 hot-pluggable disk bays, and seven available media bays. If the PCI Dual Channel Ultra2 SCSI adapter (#6205) is ordered, it can replace two PCI Single-Ended Ultra SCSI adapters, bringing the maximum number of available PCI slots on a fully-configured system to 54. See Appendix C, “PCI Dual Channel Ultra2 SCSI Adapter (#6205)” on page 139 for more details.

PCI adapters may be used to support a wide range of communications and storage subsystems. Supported communications adapters include Gigabit Ethernet, 10/100 Mbps Ethernet, standard Ethernet, token ring, Asynchronous Transfer Mode (ATM), and Fiber Distributed Data Interface (FDDI). Supported storage device protocols include Ultra SCSI, Ultra2 SCSI, Serial Storage Architecture (SSA), and Fibre Channel-Arbitrated Loop (FC-AL).

The pSeries 680 and S80 are both shipped and delivered with internal adapters and devices already installed and configured. Both require AIX

Version 4.3.3 or later, which comes with every system ordered and, if desired, can be pre-installed.

1.3 Competitive positioning

It is important to understand why the performance of the pSeries 680 and S80 are so impressive for real-world applications. The fact that the RS64 processors are running at higher clock speeds, and use more advanced technology, than in the high-end server systems offered by some other vendors explains only a small portion of their advantage. The main reason for the outstanding performance of these systems is SMP scaling efficiency.

For real world applications, the limits of the scalability achieved with an SMP system depend on many factors, some of which are outside the control of the hardware system designer. At a hardware level, though, SMP scalability comes from having a balanced system design: one which matches processor speeds with cache sizes, memory capacity, I/O expansion capability, and bus speeds.

From a technology standpoint, the main challenge facing systems designers today is access to memory, which is largely governed by the speed of the system bus. If the system bus speed is not well-matched to that of the processor, then the processor simply spends much of the time waiting for data to come from memory, and the full advantage of higher processor clock speeds is lost.

In the pSeries 680 and S80 the system bus is architected as a high speed switch, which provides excellent scalability for a very broad class of applications. This technology has been developed over a number of years and began to appear in RS/6000 servers with the J, G and R series. In the pSeries 680 and S80, a 43 GB/sec switch is used. This has dramatically improved the scaling efficiency of the system, even when using much faster processors.

The Multiprocessing (MP) factor is a measure of SMP scalability. It shows the additional performance the server achieves when you double the number of processors. An MP factor of two means that you get double the performance for double the processors. The MP factor for both the pSeries 680 and S80 with 24 processors is greater than 1.75, which is class-leading for large scale UNIX servers and which delivers an outstanding set of commercial benchmark results.

1.3.1 Online transaction processing

Many customers regard the Transaction Processing Council’s TPC-C benchmark as the single most important metric for evaluating the commercial performance of a server system. Using this benchmark, the pSeries 680 is the world's most powerful system for transaction processing. On October 13, 2000, the pSeries 680 performed 220,807 transactions per minute (tpmC) at $43.3 per transaction ($/tpmC). This was over 40 percent better than the Sun StarFire E10000, which has 64 x 400 MHz processors, and nearly 53 percent better than the Compaq AlphaServer GS320, which has 32 x 731 MHz processors.

The pSeries 680 is also the best value among large enterprise servers with the lowest cost per transaction among the ten most powerful systems.

For more information about this benchmark result, refer to the Transaction Processing Performance Council Website at the following URL:

```
http://www.tpc.org
```

1.3.2 Oracle Application Standard Benchmark

The Oracle Applications Standard Benchmark is focused on Enterprise Resource Planning (ERP) applications and represents a mixed workload intended to model the most common transactions operating on the seven most widely used enterprise application modules. As such, it is designed to be representative of typical customer workloads.

In June 2000, using 450 MHz rather than 600 MHz processors, the Model S80 was able to support 14,000 concurrent users with an average response time of only 1.27 seconds. The second highest result was produced by the 32-way 731 MHz Compaq AlphaServer GS320, which supported 11,200 users with a response time of 1.83 seconds, once again demonstrating the benefit of the outstanding SMP scaling efficiency of the pSeries 680 and S80.

For more details of The Oracle Applications Standard Benchmark, refer to the following URL:

```
http://www.oracle.com/apps_benchmark
```

1.3.3 Other Results

At the time of announcement, the pSeries 680 recorded eight *world-record* performances for commercial UNIX systems. Among these were:

- Best SAP Performance (ATO 2-tier), with 8,582 assembly orders per hour
- Best Baan performance, with 11,886 Baan Reference Users (BRU's)

- Best PeopleSoft Performance, with 15,584,416 journal lines per hour and 533,546 transactions per hour.

Just as important for today's e-commerce environment are benchmarks which demonstrate Web-serving ability and Java performance, and the pSeries 680 takes a leadership role here also.

The SPECweb99 benchmark measures the performance of a server that supports multiple Web home pages with rotating advertisements, customized page creation, user registration and other dynamic operations. Using only 12 processors, the pSeries 680 shattered the world record for Internet speed performance by supporting 7,288 simultaneous connections and delivering performance over four times faster than Compaq and Hewlett-Packard UNIX systems.

VolanoMark is a 100 percent pure Java server benchmark that measures the speed of the Java platform and accurately predicts real-world Java performance and scalability. The test has become increasingly important as e-business customers rapidly deploy Java environments as a key component of their server applications. In VolanoMark 2.1.2 local performance testing with 200 connections, a 24-way pSeries 680 transferred an unprecedented 133,251 messages per second.

While hardware improvements were important in setting these new records, AIX enhancements were also very significant. AIX Version 4.3.3 offers several Java environment, Web-performance, and scalability enhancements.

1.4 AIX Version 4.3 and commercial computing

AIX Version 4.3.3 offers a number of performance and system administration enhancements for commercial systems.

1.4.1 Kernel scalability enhancements

Kernel scalability enhancements in AIX Version 4.3.3 greatly increase OLTP throughput. In order to support the very large memory configurations available with the pSeries 680 and S80, increased to 96 GB with the 4330-06 maintenance package, the latest kernel supports multiple lists of free memory frames and multiple page replacement daemons. These constantly-running processes manage real memory by deciding whether the contents of a location in memory should remain where they are or be moved to disk, where it will take more time to retrieve them in the future. Allowing multiple lists and having multiple daemons reduces memory contention and latency (the time needed to retrieve data or instructions needed by a processor).

In another kernel enhancement, vital for effective SMP scaling, runnable threads are assigned to local run queues on a per-processor basis. This simplifies the dispatcher's decision about which thread to run next by reducing lock contention and eliminating time-consuming calculations needed to maintain affinity between a processor and its cached data. Other AIX Version 4.3.3 enhancements include:

- The ability to mirror striped logical volumes.
- A guarantee that mirrored logical volume copies will exist on different disks.
- The ability to do an online backup of a mirrored journaled file system.
- The ability to create a system backup image on a Write Once Read Many (WORM) CD drive.
- Console messages are logged to a file and time stamped.
- Faster reboot times by allowing up to 16 device configuration methods to run in parallel.

Many of these features are discussed in more detail in Chapter 7, "AIX features for large SMP servers" on page 85.

1.4.2 AIX Workload Manager

A new facility called AIX Workload Manager was introduced with AIX Version 4.3.3 and was enhanced in February 2000 with APAR IY06844. It gives a system administrator greater control over how the scheduler and Virtual Memory Manager (VMM) determine priorities and allocate system resources.

Workload management can help provide isolation between user communities that make very different demands on the system. Some users are likely to run interactive or CPU-intensive applications, while others may require batch or memory-intensive capabilities.

Workload management is fundamentally different from operating system partitioning, which is also called logical partitioning or LPAR. LPAR has been implemented on S/390 and AS/400. Partitioning isolates and dedicates hardware resources to multiple copies of the operating system running on a single machine. In contrast, the goal of workload management is to make more efficient and flexible use of CPUs and memory under the control of a single copy of the operating system.

Although the CPU and memory workload management functions described here can be used with any RS/6000 or pSeries running AIX Version 4.3.3, it is expected that they will most commonly be used on large servers running multiple differing workloads that may sometimes compete with one another. A more in-depth discussion of workload management can be found in Section 7.3, “Workload management” on page 91.

Chapter 2. The pSeries 680

This chapter takes an in-depth look at the hardware packaging and features that comprise the pSeries 680. The standard and optional features of each model are described along with the features of the I/O Drawer.

2.1 Description

The pSeries 680 was announced on October 3, 2000 and is a member of a new generation of 64-bit 6-, 12-, 18-, or 24-way symmetric multiprocessing (SMP) enterprise servers. The pSeries 680 can be used as a stand-alone server, but it can also be attached to the RS/6000 SP as an SP-attached server. Up to sixteen pSeries 680 servers can be attached to an RS/6000 SP, with or without the use of an SP Switch, to provide additional online transaction processing (OLTP) and database capability to the cluster.

Alternatively, up to sixteen pSeries 680 servers may be incorporated into a single cluster managed by the IBM Parallel Systems Support Programs for AIX (PSSP). This implementation requires a 7025-F50 control workstation for cluster control, but does not require attachment to a 9076 SP frame.

The pSeries 680 is packaged as a central electronic complex and an I/O rack. The pSeries 680 entry configuration starts with a six-way scalable SMP system that utilizes the 64-bit, 600 MHz RS64 IV processor with 16 MB of Level 2 (L2) cache per processor. The 6-way SMP can be expanded to a 24-way SMP, and the system memory can be expanded to 96 GB.

The I/O rack contains the first I/O Drawer with at least:

- A Service Processor
- One high-performance disk drive
- One SCSI-2 CD-ROM
- A 1.44 MB 3.5-inch diskette drive
- Two PCI SCSI adapters

Up to three additional I/O Drawers can be added. Additional I/O racks can also be ordered with the pSeries 680 (see Section 4.1, “I/O rack description” on page 39 for more information). Existing RS/6000 7015 Model R00 and 7014 Model S00, Model T00 or Model T42 racks can also be used for additional storage and communication drawers. This helps protect your existing investment in SSA or SCSI DASD.

The pSeries 680 is shipped and delivered with all the internal adapters and devices already installed and configured.

AIX Version 4.3.3 software is included with every pSeries 680 and can be pre-installed if desired. Note that maintenance level 433-06 is required for the pSeries 680.

2.2 Base configuration

A new pSeries 680 order must include a minimum of the following items:

- One pSeries 680 system unit, which provides the Central Electronic Complex (CEC) enclosure, system backplane, and CEC power.
- Cabling for connection to the primary I/O Drawer.
  - Two Remote I/O Cable - CEC to I/O Drawer (#3143 or #3144).
  - One System Control and Initialization Cable (#6000).
  - Two Power Control Cable, Processor Complex to I/O rack (SPCN) (#6008).
- One Remote I/O Hub - dual loop (#6503).
- One 6-way processor complex.
  - First RS64 IV Processor, 6-Way SMP, 600 MHz, 16 MB L2 Cache (#5320).
- 4 GB memory minimum. Choose from:
  - Two 2048 MB R1 Memory (4 x 512 MB Cards) (#4191).
  - One 4096 MB R1 Memory (4 x 1024 MB Cards) (#4192).
  - One 8192 MB R1 Memory (4 x 2048 MB Cards) (#4193).
  - One 16384 MB R1 Memory (4 x 4096 MB Cards) (#4194).
  - One 32768 MB R1 Memory (4 x 8192 MB Cards) (#4195).
- One I/O rack (#7036 or #7037), including rack door or trim kit and side panels. Choose from:
  - One Front Trim Kit (#6081 or #6107).
  - One Front Door (#6088 or #6089) and Two Side Panel Kits (#6098).
- One primary I/O Drawer, including:
  - One I/O Drawer, 10 EIA, AC Power, Unconfigured (#6320).
  - One primary I/O Drawer Group (#6321).

  - One Support Processor Group (#6326).
  - One SCSI-2 CD-ROM Drive (#2624).
- 9 GB hard disk (minimum). Choose from:
  - One 9.1 GB Ultra SCSI 16-bit Hot-Swap Disk Drive (#2913).
  - One 9.1 GB 10,000 RPM Ultra SCSI 1-inch (25-mm) Hot-Swap Disk Drive (#3002).
  - One 18.2 GB Ultra SCSI 16-bit Hot-Swap Disk Drive (#3104).
  - One 18.2 GB 10,000 RPM Ultra SCSI Hot-Swap Disk Drive (#3117).
- SCSI backplanes, adapters, and cables. Choose either a single or dual backplane solution.
  - Single backplane solution, consisting of:
    - One SCSI 6-Pack Hot-Swap Backplane/Power Cable (#6547).
    - Two PCI Single-Ended Ultra SCSI Adapter (#6206).
    - One SCSI-2 Backplane-to-DASD 6-Pack Cable (#2447).
  - Dual backplane solution, consisting of:
    - Two SCSI 6-Pack Hot-Swap Backplane/Power Cable (#6547).
    - Three PCI Single-Ended Ultra SCSI Adapter (#6206).
    - Two SCSI-2 Backplane-to-DASD 6-Pack Cable (#2447).

  In either solution, two of the PCI Single-Ended Ultra SCSI Adapters may be replaced by a single PCI Dual Channel Ultra2 SCSI Adapter (#6205). However, the Ultra2 adapter will function at Ultra SCSI speed.

2.3 Processor

Each pSeries 680 processor card has six RS64 IV processors with the associated L2 cache contained on the card. There are 16 MB of L2 per processor. The pSeries 680 can accommodate four CPU cards. The first 600 MHz CPU card (#5320) is included in a base configuration. Up to three additional CPU cards may be added, with differing numbers of active processors. Cards with less than six active processors are enabled with a new feature of AIX described in Section 7.4, “Capacity Upgrade on Demand” on page 97.

The feature codes are:

- (#5320) 600 MHz Card, First
- (#5321) 600 MHz Card, Additional - six processors active

Capacity Upgrade on Demand features codes:

- (#8301) 600 MHz Card, Additional - no processors active
- (#8303) 600 MHz Card, Additional - two processors active
- (#8305) 600 MHz Card, Additional - four processors active

Note that only one of #8303 or #8305 can be installed in a system. A feature conversion to the full 6-way active #5321 processor feature is required before another 2-way or 4-way active feature may be utilized.

2.4 Power

The pSeries 680 comes equipped with sufficient power supplies for a 6- or 12-way system. When ordering a 12- or 24-way system, an additional power supply, processor regulator, and power regulator must be ordered. Choose from:

- (#6913) 1000 Watt AC Power Supply
- (#6914) Programmable Power Regulator
- (#6915) Processor Power Regulator

These additional power components are all installed in the front of the pSeries 680 CEC. The 1000 Watt AC Power Supply is installed in position P05; the Programmable Power Regulator is installed in position R08, and the Processor Power Regulator is installed in position M07. The locations are shown in Figure 18 on page 67.

The pSeries 680 has an optional dual line cord feature (#8622) for the CEC, which may be configured with the initial order or as a field-installable MES upgrade. This feature requires #9800 - Power Cord Specify - United States/Canada.

Optional uninterruptible power supply (UPS) systems are supported and recommended for mission-critical servers.

2.5 Memory

The base configuration for the pSeries 680 includes 4 GB of SDRAM-based memory. The maximum configuration is 96 GB. The pSeries 680 can accommodate up to 16 memory cards. Memory cards are used in sets of four, called quads. Memory cards are available in 512 MB, 1 GB, 2 GB, 4 GB, and 8 GB sizes.

System memory is accessed through four related but distinct ports which can be accessed in a coordinated parallel manner to obtain very high data rates. A system should be configured with a minimum of two memory quads to make best use of the system architecture. A configuration that uses only one port will function properly, but the system can not make use of the full memory bus bandwidth. For example, a system with 4 GB of memory will perform better with two 2 GB features installed than if one 4 GB feature is installed.

For Scientific and Technical applications you should take care to balance the memory across all four ports if possible. For most commercial applications the selection of memory features will usually be decided on cost and on future upgradeability. First, you need to understand and plan the final maximum memory configuration the application is likely to require. You then need to define a growth path to achieve this final maximum without having to discard quads (quads cannot be upgraded). It is acceptable to have different sized quads in a configuration.

Note that it is not possible to have three 32 GB quads in a system. The maximum memory configuration of 96 GB can therefore only be achieved using two 32 GB quads and two 16 GB quads.

The available memory features are as follows:

- (#4191) 2048 MB Memory (4 x 512 MB Cards)
- (#4192) 4096 MB Memory (4 x 1024 MB Cards)
- (#4193) 8192 MB Memory (4 x 2048 MB Cards)
- (#4194) 16384 MB Memory (4 x 4096 MB Cards)
- (#4195) 32768 MB Memory (4 x 8192 MB Cards)

For a detailed description of the pSeries 680 memory subsystem, together with quad installation guidelines, see Chapter 6, “Hardware architecture” on page 67.

2.6 Disk drives

Disk drives are available for installation in the I/O Drawers of the pSeries 680. Disk drives are not supported in the media section of the I/O Drawers. Each I/O Drawer supports up to two SCSI 6-packs. As the name suggests, each 6-pack supports up to six hot-pluggable 16-bit SCSI disk drives. All disk drives must be 16-bit devices. Each 6-pack must be connected to a SCSI adapter using a 6-pack attachment cable.

Disk Drives include:

- (#2913) 9.1 GB Ultra SCSI 16-bit Hot-Swap
- (#3002) 9.1 GB 10,000 RPM Ultra SCSI 1-inch (25-mm) Hot-Swap
- (#3104) 18.2 GB Ultra SCSI 16-bit Hot-Swap
- (#3117) 18.2 GB 10,000 RPM Ultra SCSI Hot-Swap

SCSI 6-pack:

- (#6547) SCSI 6-pack

6-pack Attachment Cables:

- (#2447) Attachment to PCI SCSI/RAID Adapter

2.7 Boot Devices

Boot support is available from local SCSI Adapters (described in the following section), SSA adapters (including the #6230 Advanced SerialRAID adapter, provided a non-RAID SSA disk is included as part of the configuration), or from a network using Ethernet (not #2975 10/100/1000 Base-T Adapter) or token-ring adapters.

The recommended location for a SCSI boot device is within the primary I/O Drawer. This configuration provides service personnel with the maximum amount of diagnostic information if the system encounters errors in the boot sequence. The default boot drive is in the lowest location in the six-pack, in the inner-most bay of the I/O Drawer. Manufacturing installs the boot adapter in slot 13. If a boot source other than internal disk is configured, the supporting adapter must be installed in the primary I/O Drawer.

2.8 SCSI Adapters

SCSI Adapters include:

- (#6205) Dual Channel Ultra2 SCSI Adapter
- (#6206) Ultra SCSI Single-Ended Adapter
- (#6207) Ultra SCSI Differential Adapter

The base configuration of both the pSeries 680 and S80 includes two Ultra SCSI Single-Ended Adapters (#6206). One adapter drives the media bays, and the other drives the first SCSI 6-pack. These two adapter cards can be replaced by a single Dual Channel Ultra2 SCSI Adapter (#6205), which increases the number of available PCI slots in the primary I/O Drawer. However, the Ultra2 adapter will operate at Ultra speed.

2.9 I/O Drawers

See Section 4.1.1, “I/O Drawer description” on page 40, for a full description of the I/O Drawer. The original 7 EIA Model S70 I/O Drawers are not supported on the pSeries 680.

The standard peripherals required in the minimum configuration include the following:

- 1.44 MB Diskette Drive
- (#2624) SCSI-2 CD-ROM

The following feature code is the base I/O Drawer:

- (#6320) Base SCSI I/O Drawer, 10 EIA

The following lists the feature codes of the drawer groups:

- (#6321) Primary I/O Drawer group
- (#6323) Secondary I/O Drawer group

2.10 Cabling

The CEC and the I/O Drawers are connected by various cables. The primary I/O Drawer has additional connections.

2.10.1 System Power Control Network

See Section 4.2, “System Power Control Network (SPCN)” on page 42, for a full description of the purpose and configuration options for SPCN cables.

The available SPCN cables for pSeries 680 systems are as follows:

- (#6006) 2-meter drawer-to-drawer control cable
- (#6007) 15-meter rack-to-rack control cable
- (#6008) 6-meter rack-to-rack control cable

2.10.2 Remote I/O cables

See Section 4.3, “Remote I/O cabling” on page 44, for a full description of the purpose and configuration options for Remote I/O cables.

RIO cables are available in three different lengths. The 3-meter cables can only be used to interconnect two I/O Drawers in the same rack. Manufacturing will determine the placement and cabling of I/O Drawers, based on the quantity of I/O racks and RIO cables ordered.

The following remote I/O cables are available:

- (#3142) 3-meter drawer-to-drawer remote I/O cable
- (#3143) 6-meter rack-to-rack remote I/O cable
- (#3144) 15-meter rack-to-rack remote I/O cable

2.11 pSeries 680 configurations

Table 1 lists the pSeries 680 minimum configuration.

Table 1. pSeries 680 minimum configuration

pSeries 680 Minimum Configuration and Features	
Microprocessor	One 6-way 600 MHz RS64 IV CPU card
Level 1 (L1) cache	128 KB data/128 KB instruction
Level 2 (L2) cache	16 MB per processor
RAM (minimum)	4 GB
Memory bus width	Quad 512-bit
Ports	One parallel, two serial, one keyboard, and one mouse
Internal disk drive	One 9.1 GB Ultra SCSI (hot-swappable)
Media bays	Two (one available)
Expansion slots	Fourteen PCI (eleven available)
PCI bus width	32- and 64-bit
Memory slots	16
CD-ROM drive	Yes
Service processor	Yes
Diskette drive	1.44 MB 3.5-inch diskette drive
SCSI adapters	Two Ultra SCSI PCI adapters
AIX operating system version	Version 4.3 (an unlimited user license is included)

Table 2 lists the pSeries 680 expansion capabilities.

Table 2. pSeries 680 system expansion

pSeries 680 Maximum Configuration	
SMP configurations	Up to 24 600 MHz processors
RAM	Up to 96 GB
Internal PCI slots	Up to 56 per system (53 available)
Internal media bays	Up to eight per system (seven available)
Internal disk bays	Up to 48 (hot-swappable)
Internal disk storage	Up to 873.6 GB

2.12 Publications

Table 3 details the publications shipped with the pSeries 680.

Table 3. Publications shipped with the pSeries 680

Title	Order Number
pSeries 680 Model S85 Installation Guide	SA38-0582
PCI Adapter Placement Reference Guide	SA38-0538
RS/6000 Customer Support Information (U.S. and Canadian customers only)	SA23-2690
System Unit Safety Information	SA23-2652
System Unit Safety Information Supplement	SN32-9075
Warranty Booklet (U.S. customers only)	Z125-4753
Customer Installable Options Library CD-ROM[1]	

[1] The CD-ROM cannot be ordered; it is shipped with the product and no form number is available.

The publications shown in Table 4 are available for a fee:

Table 4. Additional publications available for a fee

Title	Order Number
pSeries 680 Model S85 User's Guide	SA38-0557
pSeries 680 Model S85 Service Guide	SA38-0558

Title	Order Number
Site and Hardware Planning Guide	SA38-0508
Diagnostic Information for Multiple Bus Systems	SA38-0509
Adapters, Devices, and Cable information for Multiple Bus Systems	SA38-0516

2.13 pSeries 680 additional features

This section describes the internal features that can be added to a configuration at additional cost.

The status of a feature is indicative of the following qualifications:

A Indicates features that are available and orderable.

Features not listed in the provided categories indicate that the feature is not supported on this model. Some categories, such as keyboards, cables, and monitors, are not included.

Table 5 lists pSeries 680 optional features and their status.

Table 5. pSeries 680 optional features

Feature Code	Description	Status
Processors		
5320	RS64 IV 6-way 600 MHz first	A
5321	RS64 IV 600 MHz additional, 6 active	A
8301	RS64 IV 600 MHz additional, 0 active	A
8303	RS64 IV 600 MHz additional, 2 active	A
8305	RS64 IV 600 MHz additional, 4 active	A
Memory		
4191	2048 MB (4x512)	A
4192	4096 MB (4x1024)	A
4193	8192 MB (4x2048)	A
4194	16384 MB (4x4096)	A
4195	32768 MB (4x8192) - maximum of two	A

Feature Code	Description	Status
Host Attachment		
2751	ESCON Control Unit	A
8396	SP Attach Adapter	A
Internal Disk Drives		
2913	9.1 GB 1" Ultra SCSI Hot-Swap	A
3002	9.1 GB 10K RPM 1" Ultra SCSI Hot-Swap	A
3104	18.2 GB 1" Ultra SCSI Hot-Swap	A
3117	18.2 GB 10K RPM 1" Ultra SCSI Hot-Swap	A
Internal Tape Drives		
6156	20/40 GB 8 mm (Black)	A
6158	20/40 GB 4 mm	A
6159	12/24 GB 4 mm	A
Internal CD-ROMs		
2624	SCSI-2 CD-ROM	A
Graphics Accelerators		
2830	GXT130P	A
SCSI Adapters		
6204	Ultra SCSI Differential	A
6205	Dual Channel Ultra2 SCSI	A
6206	Ultra SCSI SE	A
SSA Adapters		
6222	SSA Fast-Write Cache Option	A
6230	IBM Advanced Serial RAID Plus	A
6235	IBM Advanced Serial RAID Cache Option	A
Fibre Channel Adapters		
6227	Gigabit Fibre Channel Adapter	A

Feature Code	Description	Status
Async Adapters		
2943	8-Port Async EIA-232/422	A
2944	128-Port Async Controller	A
ARTIC Adapters		
2947	ARTIC960Hx 4-Port Selectable	A
Digital Trunk Adapters		
6310	ARTIC960RxD Quad Digital	A
ATM Adapters		
2946	Turboways 622 PCI MMF ATM	A
2963	Turboways 155 PCI UTP ATM	A
2988	Turboways 155 PCI MMF ATM	A
Token-Ring Adapters		
4959	Token-Ring PCI Adapter	A
Ethernet Adapters		
2968	IBM 10/100 Mbps Ethernet	A
2969	Gigabit Ethernet (fiber)	A
2975	Gigabit Ethernet (UTP)	A
2985	Ethernet BNC / RJ-45	A
2987	Ethernet AUI / RJ-45	A
WAN Adapters		
2962	2-Port Multiprotocol PCI	A
FDDI Adapters		
2741	SysKonnect SK-NET FDDI-LP SAS	A
2742	SysKonnect SK-NET FDDI-LP DAS	A
2743	SysKonnect SK-NET FDDI-UP SAS	A
ISDN Adapter		
2708	Eicon ISDN DIVA PRO 2.0 PCI S/T	A

Chapter 3. The Model S80

This chapter takes an in-depth look at the hardware packaging and features that comprise the Model S80. The standard and optional features of each model are described along with the features of the I/O Drawer.

3.1 Description

The Model S80 was announced on September 13, 1999 and is a member of a new generation of 64-bit 6-, 12-, 18-, or 24-way symmetric multiprocessing (SMP) enterprise servers. The Model S80 can be used as a stand-alone server, but it can also be attached to the RS/6000 SP as an SP-attached server. Up to sixteen Model S80 servers can be attached to an RS/6000 SP, with or without the use of an SP Switch, to provide additional online transaction processing (OLTP) and database capability to the cluster.

Alternatively, up to sixteen Model S80 servers may be incorporated into a single cluster managed by the IBM Parallel Systems Support Programs for AIX (PSSP). This implementation requires a 7025-F50 control workstation for cluster control, but does not require attachment to a 9076 SP frame.

The Model S80 is packaged as a central electronic complex and an I/O rack. The Model S80 entry configuration starts with a six-way scalable SMP system that utilizes the 64-bit, 450 MHz RS64 III processor with 8 MB of Level 2 (L2) cache per processor. The processors may be upgraded to the 64-bit, 600 MHz RS64 IV processors with 16 MB of L2 cache and the 6-way SMP can be expanded to a 24-way SMP. System memory can be expanded to 96 GB.

The I/O rack contains the first I/O Drawer, which includes:

- A Service Processor
- A high-performance disk drive
- One SCSI-2 CD-ROM
- 1.44 MB 3.5-inch diskette drive
- Two PCI SCSI adapters

Up to three additional Model S80 I/O Drawers can be added. Additional I/O racks can also be ordered with the Model S80. Existing RS/6000 7015 Model R00 and 7014 Model S00 racks can also be used for additional storage and communication drawers. This helps to protect your existing investment in SSA or SCSI DASD.

The Model S80 is shipped and delivered with all the internal adapters and devices already installed and configured.

AIX Version 4.3.3 software is included with every Model S80 and can be pre-installed if desired. Note that support for the 600 MHz processors or for more than 64 GB of memory requires maintenance level 433-06 of AIX Version 4.3.3.

3.2 Base configuration

A new Model S80 order must include a minimum of the following items:

- One Model S80 system unit, which provides the CEC (Central Electronic Complex) enclosure, system backplane, and CEC power.
- Cabling for connection to the primary I/O Drawer:
    - Two Remote I/O Cable - CEC to I/O Drawer (#3143 or #3144).
    - One system control and initialization cable (#6000).
    - Two power control cable, Processor Complex to I/O rack (SPCN) (#6008).
- One Remote I/O Hub - dual loop (#6503).
- One 6-way processor complex:
    - First RS64 III Processor, 6-Way SMP 450 MHz, 8 MB L2 Cache (#5318).
- 2 GB memory minimum. Choose from:
    - Two 1024 MB R1 Memory (4 x 256 MB Cards) (#4190).
    - One 2048 MB R1 Memory (4 x 512 MB Cards) (#4191).
    - One 4096 MB R1 Memory (4 x 1024 MB Cards) (#4192).
    - One 8192 MB R1 Memory (4 x 2048 MB Cards) (#4193).
    - One 16384 MB R1 Memory (4 x 4096 MB Cards) (#4194).
    - One 32768 MB R1 Memory (4 x 8192 MB Cards) (#4195).
- One I/O rack. Choose from:
    - (#7000) 32U I/O Rack.
    - (#7036) 36U I/O Rack.
    - (#7037) 42U I/O Rack.
    - (#7036) and (#7037) require either a rack door or trim kit and side panels. Choose from:

    - One Front Trim Kit (#6081 or #6107).
    - One Front Door (#6083 or #6097) and Two Side Panel Kits (#6098).
- One primary I/O Drawer, including:
  - I/O Drawer, 10 EIA, AC Power, Unconfigured (#6320).
  - Primary I/O Drawer group (#6321).
  - Support Processor group (#6326).
  - SCSI-2 CD-ROM Drive (#2624).
- 9 GB hard disk (minimum). Choose from:
  - One 9.1 GB Ultra SCSI 16-bit Hot-Swap Disk Drive (#2913).
  - One 9.1 GB 10,000 RPM Ultra SCSI 1-inch (25-mm) Hot-Swap Disk Drive (#3002).
  - One 18.2 GB Ultra SCSI 16-bit Hot-Swap Disk Drive (#3104).
  - One 18.2 GB 10,000 RPM Ultra SCSI Hot-Swap Disk Drive (#3117).
- SCSI backplanes, adapters, and cables. Choose either single or dual backplane solution.
  - Single backplane solution, consisting of:
    - One - SCSI 6-Pack Hot-Swap Backplane/Power Cable (#6547).
    - Two - PCI Single-Ended Ultra SCSI Adapter (#6206).
    - One - SCSI-2 Backplane-to-DASD 6-Pack Cable (#2447).
  - Dual backplane solution, consisting of:
    - Two - SCSI 6-Pack Hot-Swap Backplane/Power Cable (#6547).
    - Three - PCI Single-Ended Ultra SCSI Adapter (#6206).
    - Two - SCSI-2 Backplane-to-DASD 6-Pack Cable (#2447).

  In either solution, two of the PCI Single-Ended Ultra SCSI Adapters may be replaced by a single PCI Dual Channel Ultra2 SCSI Adapter (#6205). However, the Ultra2 adapter will function at Ultra SCSI speed.

3.3 Processor

Each Model S80 processor card has six RS64 processors with the associated L2 cache contained on the card. Two processor types are available for the Model S80: the 450 MHz RS64 III, with 8 MB of L2 cache per processor, and the 600 MHz RS64 IV, with 16 MB of L2 cache per processor. All the processor cards in a system need to use the same type and speed of processor.

The Model S80 can accommodate 4 CPU cards. The first 450 MHz CPU card (#5318) is included in a base configuration. Up to three additional cards may be added. Feature codes for 600 MHz processors may be installed only as a field upgrade MES (they may not be ordered on a new system).

Feature code for systems using 450 MHz processors include:

- (#5318) 450 MHz Card, First
- (#5319) 450 MHz Card, Additional (six processors active)

Capacity Upgrade on Demand features codes include:

- (#8300) 450 MHz Card, Additional (no processors active)
- (#8302) 450 MHz Card, Additional (two processors active)
- (#8304) 450 MHz Card, Additional (four processors active)

Note that only one of #8302 or #8304 can be installed in a system. A feature conversion to the full 6-way active #5319 processor feature is required before another 2-way or 4-way active feature may be utilized.

Feature code for systems using 600 MHz processors include:

- (#5320) 600 MHz Card, First
- (#5321) 600 MHz Card, Additional (six processors active)

Capacity Upgrade on Demand features codes include:

- (#8301) 600 MHz Card, Additional (no processors active)
- (#8303) 600 MHz Card, Additional (two processors active)
- (#8305) 600 MHz Card, Additional (four processors active)

Note that only one of #8303 or #8305 can be installed in a system. A feature conversion to the full 6-way active #5321 processor feature is required before another 2-way or 4-way active feature may be utilized.

3.4 Power

The model S80 comes equipped with sufficient power supplies for a 6- or 12-way system. When ordering a 12- or 24-way system, an additional power supply, processor regulator, and power regulator must be ordered:

- (#6913) 1000 Watt AC Power Supply

- (#6914) Programmable Power Regulator
- (#6915) Processor Power Regulator

These additional power components are all installed in the front of the Model S80 CEC. The 1000 Watt AC Power Supply is installed in position P05, the Programmable Power Regulator is installed in position R08, and the Processor Power Regulator is installed in position M07. The locations are shown in Figure 18 on page 67.

Note the Model S80 does not have a dual line cord feature for the CEC.

Optional uninterruptible power supply (UPS) systems are supported and recommended for mission-critical servers.

3.5 Memory

The base configuration for the Model S80 includes 2 GB of SDRAM-based memory. The maximum configuration is 96 GB. The Model S80 can accommodate up to 16 memory cards. Memory cards are available in 256 MB, 512 MB, 1 GB, 2 GB, 4 GB, and 8 GB sizes. Memory cards are used in sets of four, called quads.

System memory is accessed through four related but distinct ports which can be accessed in a coordinated parallel manner to obtain very high data rates. A system should be configured with a minimum of two memory quads to make best use of the system architecture. A configuration that uses only one port will function properly, but the system can not make use of the full memory bus bandwidth. For example, a system with 2 GB of memory will perform better with two 1 GB features installed than if one 2 GB feature is installed.

For scientific and technical applications you should take care to balance the memory across all four ports (if possible). For most commercial applications, selection of memory features will usually be decided on cost and on future upgradeability. First, you need to understand and plan the final maximum memory configuration the application is likely to require. You then need to define a growth path to achieve this final maximum without having to discard quads (quads cannot be upgraded). It is acceptable to have different sized quads in a configuration.

Note that it is not possible to have three 32 GB quads in a system. The maximum memory configuration of 96 GB can, therefore, only be achieved using two 32 GB quads and two 16 GB quads.

The available memory features are as follows:

- (#4190) 1024 MB Memory (4 x 256 MB Cards)
- (#4191) 2048 MB Memory (4 x 512 MB Cards)
- (#4192) 4096 MB Memory (4 x 1024 MB Cards)
- (#4193) 8192 MB Memory (4 x 2048 MB Cards)
- (#4194) 16384 MB Memory (4 x 4096 MB Cards)
- (#4195) 32768 MB Memory (4 x 8192 MB Cards)

For a detailed description of the Model S80 memory subsystem, together with quad installation guidelines, see Chapter 6, “Hardware architecture” on page 67.

3.6 Disk drives

Disk drives are available for installation in the I/O Drawers of the Model S80. Disk drives are not supported in the media section of the I/O Drawers. Each I/O Drawer supports up to two SCSI 6-packs. As the name suggests, each 6-pack supports up to six hot-pluggable 16-bit SCSI disk drives. All disk drives must be 16-bit devices. Each 6-pack must be connected to an SCSI adapter using a 6-pack attachment cable.

Disk Drives include:

- (#2913) 9.1 GB Ultra SCSI 16-bit Hot-Swap
- (#3002) 9.1 GB 10,000 RPM Ultra SCSI 1-inch (25-mm) Hot-Swap
- (#3104) 18.2 GB Ultra SCSI 16-bit Hot-Swap
- (#3117) 18.2 GB 10,000 RPM Ultra SCSI Hot-Swap

SCSI 6-pack:

- (#6547) SCSI 6-pack

6-pack Attachment Cables:

- (#2447) Attachment to PCI SCSI/RAID Adapter

3.7 Boot Devices

Boot support is available from local SCSI Adapters (described in the following section), SSA adapters (including the #6230 Advanced SerialRAID adapter, provided a non-RAID SSA disk is included as part of the configuration), or from a network using Ethernet (not #2975 10/100/1000 Base-T Adapter) or token-ring adapters.

The recommended location for the boot device (SCSI) is within the primary I/O Drawer. This configuration provides service personnel with the maximum amount of diagnostic information if the system encounters errors in the boot sequence. The default boot drive is in the lowest location in the six-pack, in the inner-most bay of the I/O Drawer. Manufacturing installs the boot adapter in slot 13. If a boot source other than internal disk is configured, the supporting adapter must be installed in the primary I/O Drawer.

3.8 SCSI Adapters

SCSI Adapters:

- (#6205) Dual Channel Ultra2 SCSI Adapter
- (#6206) Ultra SCSI Single-Ended Adapter
- (#6207) Ultra SCSI Differential Adapter

The base configuration of Model S80 comes with two Ultra SCSI Single-Ended Adapters (#6206). One adapter drives the media bays, and the other drives the first SCSI 6-pack. These two adapter cards can be replaced by a single Dual Channel Ultra2 SCSI Adapter (#6205), which increases the number of available PCI slots in the primary I/O Drawer. However, the Ultra2 adapter will operate at Ultra speed.

3.9 I/O Drawers

See Section 4.1.1, “I/O Drawer description” on page 40, for a full description of the I/O Drawer. The original 7 EIA Model S70 I/O Drawers are not supported on the Model S80.

The standard peripherals required in the minimum configuration include the following:

- 1.44 MB Diskette Drive
- (#2624) SCSI-2 CD-ROM

The following feature code is the base I/O Drawer:

- (#6320) Base SCSI I/O Drawer, 10 EIA

The following lists the feature codes of the drawer groups:

- (#6321) Primary I/O Drawer Group
- (#6323) Secondary I/O Drawer Group

3.10 Cabling

The CEC and the I/O Drawers are connected by various cables. The primary I/O Drawer has additional connections.

3.10.1 System Power Control Network

See Section 4.2, “System Power Control Network (SPCN)” on page 42, for a full description of the purpose and configuration options for SPCN cables.

The available SPCN cables for Model S80 systems are as follows:

- (#6006) 2-meter drawer-to-drawer control cable
- (#6007) 15-meter rack-to-rack control cable
- (#6008) 6-meter rack-to-rack control cable

3.10.2 Remote I/O cables

See Section 4.3, “Remote I/O cabling” on page 44, for a full description of the purpose and configuration options for Remote I/O cables.

RIO cables are available in three different lengths. The 3-meter cables can only be used to interconnect two I/O Drawers in the same rack. Manufacturing will determine the placement and cabling of I/O Drawers, based on the quantity of I/O racks and RIO cables ordered.

The following remote I/O cables are available:

- (#3142) 3-meter drawer-to-drawer remote I/O cable
- (#3143) 6-meter rack-to-rack remote I/O cable
- (#3144) 15-meter rack-to-rack remote I/O cable

3.11 Model S80 configurations

Table 6 lists the S80 minimum configuration.

Table 6. Model S80 minimum configuration

Model S80 Minimum Configuration and Features	
Microprocessor	One 6-way 450 MHz RS64 III CPU card

Model S80 Minimum Configuration and Features	
Level 1 (L1) cache	128 KB data/128 KB instruction
Level 2 (L2) cache	8 MB per processor
RAM (minimum)	2 GB
Memory bus width	Quad 512-bit
Ports	One parallel, two serial, one keyboard, and one mouse
Internal disk drive	One 9.1 GB Ultra SCSI (hot-swappable)
Media bays	Two (one available)
Expansion slots	Fourteen PCI (eleven available)
PCI bus width	32- and 64-bit
Memory slots	16
CD-ROM drive	Yes
Service processor	Yes
Diskette drive	1.44 MB 3.5-inch diskette drive
SCSI adapters	Two Ultra SCSI PCI adapters
AIX operating system version	Version 4.3 (an unlimited user server license is included)

Table 7 lists the Model S80 expansion capabilities.

Table 7. Model S80 system expansion

Model S80 Maximum Configuration	
SMP configurations	Up to 24 600 MHz processors
RAM	Up to 96 GB
Internal PCI slots	Up to 56 per system (53 available)
Internal media bays	Up to eight per system (seven available)
Internal disk bays	Up to 48 (hot-swappable)
Internal disk storage	Up to 873.6 GB

3.12 Publications

Table 8 details the publications shipped with the Model S80.

Table 8. Publications shipped with the Model S80

Title	Order Number
Enterprise Server Model S80 Installation Guide	SA38-0582
PCI Adapter Placement Reference Guide	SA38-0538
RS/6000(R) Customer Support Information (U.S. and Canadian customers only)	SA23-2690
System Unit Safety Information	SA23-2652
System Unit Safety Information Supplement	SN32-9075
Warranty Booklet (U.S. customers only)	Z125-4753
Customer Installable Options Library CD-ROM[1]	

[1] The CD-ROM cannot be ordered; it is shipped with the product and no form number is available.

The publications shown in Table 9 are available for a fee:

Table 9. Additional publications available for a fee

Title	Order Number
Enterprise Server Model S80 User’s Guide	SA38-0557
Enterprise Server Model S80 Service Guide	SA38-0558
Site and Hardware Planning Guide	SA38-0508
Diagnostic Information for Multiple Bus Systems	SA38-0509
Adapters, Devices, and Cable information for Multiple Bus Systems	SA38-0516

3.13 Model S80 additional features

This section describes the internal features that can be added to a configuration at additional cost.

The status of a feature is indicative of the following qualifications:

A Indicates features that are available and orderable.

S Indicates a feature that is supported on a new Model S80 during a model conversion from an S70 Advanced; these features will work on the new model, but additional quantities of these features cannot be ordered (they can only be removed).

W Indicates features that are supported on a Model S80, but which have been withdrawn from marketing and are no longer available.

Features not listed in the provided categories indicate that the feature is not supported on this model. Some categories, such as keyboards, cables, and monitors, are not included.

Table 10 lists Model S80 optional features and their status.

Table 10. Model S80 optional features

Feature Code	Description	Status
Processors		
5318	RS64 III 6-way 450 MHz first	A
5319	RS64 III 6-way 450 MHz additional	A
5320	RS64 IV 6-way 600 MHz first	A
5321	RS64 IV 600 MHz additional, 6 active	A
8300	RS64 III 450 MHz additional, 0 active	A
8301	RS64 IV 600 MHz additional, 0 active	A
8302	RS64 III 450 MHz additional, 2 active	A
8303	RS64 IV 600 MHz additional, 2 active	A
8304	RS64 III 450 MHz additional, 4 active	A
8305	RS64 IV 600 MHz additional, 4 active	A
Memory		
4190	1024 MB (4x256)	A
4191	2048 MB (4x512)	A
4192	4096 MB (4x1024)	A
4193	8192 MB (4x2048)	A

Feature Code	Description	Status
4194	16384 MB (4x4096)	A
4195	32768 MB (4x8192)	A
Host Attachment		
2751	ESCON Control Unit	A
8396	SP Attach Adapter	A
Internal Disk Drives		
2901/9394	4.5 GB Ultra SCSI Hot-Swap	W
2911/3019	9.1 GB Ultra SCSI Hot-Swap	W
2913	9.1 GB 1" Ultra SCSI Hot-Swap	A
3002	9.1 GB 10K RPM 1" Ultra SCSI Hot-Swap	A
3008	9.1 GB 10K RPM Ultra SCSI Hot-Swap	W
3104	18.2 GB 1" Ultra SCSI Hot-Swap	A
3117	18.2 GB 10K RPM 1" Ultra SCSI Hot-Swap	A
Internal Tape Drives		
6142	4/8 GB 4 mm	S
6147	5/10 GB 8 mm	S
6154	20/40GB 8 mm (White)	S
6156	20/40 GB 8 mm (Black)	A
6158	20/40 GB 4 mm	A
6159	12/24 GB 4 mm	A
6160	9-track half-inch tape drawer	S
Internal CD-ROMs		
2619	20X Speed CD-ROM	S
2624	SCSI-2 CD-ROM	A
Graphics Accelerators		
2838	GXT120P	S
2830	GXT130P	A

Feature Code	Description	Status
SCSI Adapters		
6204	Ultra SCSI Differential	A
6205	Dual Channel Ultra2 SCSI	A
6206	Ultra SCSI SE	A
6207	Ultra SCSI Differential	W
6208	SCSI 2 Fast / Wide	S
6209	SCSI Differential Fast/Wide	S
SSA Adapters		
6215	SSA Multi-Initiator / RAID EL	S
6222	SSA Fast-Write Cache Option	A
6225	IBM Advanced Serial RAID	W
6230	IBM Advanced Serial RAID Plus	A
6235	IBM Advanced Serial RAID Cache Option	A
Fibre Channel Adapters		
6227	Gigabit Fibre Channel Adapter	A
Async Adapters		
2943	8-Port Async EIA-232/422	A
2944	128-Port Async Controller	A
ARTIC Adapters		
2947	ARTIC960Hx 4-Port Selectable	A
2948	ARTIC960Hx 4-Port T1/E1 PCI	W
Digital Trunk Adapters		
6310	ARTIC960RxD Quad Digital	A
ATM Adapters		
2946	Turboways 622 PCI MMF ATM	A
2963	Turboways 155 PCI UTP ATM	A
2988	Turboways 155 PCI MMF ATM	A

Feature Code	Description	Status
Token-Ring Adapters		
2920	Token-Ring Adapter	S
2979	Auto LANStreamer Token-Ring	S
4959	Token-Ring PCI Adapter	A
Ethernet Adapters		
2968	IBM 10/100 Mbps Ethernet	A
2969	Gigabit Ethernet (fiber)	A
2975	Gigabit Ethernet (UTP)	A
2985	Ethernet BNC / RJ-45	A
2986	Fast Etherlink XL 3Com	S
2987	Ethernet AUI / RJ-45	A
WAN Adapters		
2962	2-Port Multiprotocol PCI	A
FDDI Adapters		
2741	SysKonnect SK-NET FDDI-LP SAS	A
2742	SysKonnect SK-NET FDDI-LP DAS	A
2743	SysKonnect SK-NET FDDI-UP SAS	A
ISDN Adapter		
2708	Eicon ISDN DIVA PRO 2.0 PCI S/T	A

Chapter 4. Remote I/O Subsystem

The I/O subsystem on both the pSeries 680 and S80 consists of up to four I/O Drawers housed in I/O racks and connected to the system CEC. Each I/O rack can be configured with up to four I/O Drawers (depending on the type of rack chosen).

4.1 I/O rack description

The standard I/O rack supplied with the Model S80 is the 7014-S00 rack. It has 32 EIA units of space and can accommodate two I/O Drawers together with optional disk or tape devices. It is not available as a feature code for a pSeries 680, but can be used as a shared I/O rack.

The standard I/O rack supplied with the pSeries 680 is the 7014-T00 rack. It has 36 EIA units of space and can accommodate three I/O Drawers together with optional disk or tape devices. It is available as an option for the Model S80.

A third rack is also available as an option for both the pSeries 680 and S80. The 7014-T42 has 42 EIA units of space and can accommodate four I/O Drawers together with optional disk or tape devices. You are advised to check for any height restrictions that may impede delivery of this rack prior to ordering. The factory will install drawers only in the lower 32-EIA units of the T42 rack. Installation of drawers in the upper 10-EIA units will be carried out at the customer location.

All 7014 racks are available with three Power Distribution Unit (PDU) options. They are selected with the following base feature codes:

- (#9171) Side-Mounted, 1-Phase
- (#9173) Side-Mounted, 3-Phase
- (#9174) Side-Mounted, 3-Phase, Swiss

Each PDU provides six power outlets. Additional PDUs can be configured in a rack to provide power for additional equipment or to provide a degree of high-availability to devices that have multiple redundant power supplies, such as the 7133-D40 SSA enclosure. The following are the feature codes of available additional Power Distribution Units:

- (#6171) Side-Mounted, 1-Phase
- (#6173) Side-Mounted, 3-Phase

- (#6174) Side-Mounted, 3-Phase, Swiss

Up to three PDUs can be installed in an S00 rack, and up to four in both the T00 and T42 racks. Each additional PDU reduces the available space by one EIA.

The following rules have to be followed for a valid configuration:

- A maximum of four I/O racks can be ordered as features (#7000, #7036, or #7037) per system order. I/O Drawers should ideally be spread between I/O racks.
- Multiple I/O racks can be joined together using rack suite attachment kits. When racks are joined together, side panels are necessary only at the two ends of the suite of racks. A suite of racks, with a quantity of n racks joined together, would typically require a quantity of n-1 rack suite attachment kits and two side panels to cover the two ends of the suite of racks. Joining two racks of unequal height is not recommended.
- I/O racks ordered as feature numbers of the pSeries 680 and S80 (#7000, #7036, or #7037) must contain an I/O Drawer. If additional external communication and storage devices, such as 7133 and 7027, do not fit in the space remaining in the I/O racks, additional empty I/O racks should be ordered. The additional 7014 Racks should be ordered as products rather than features of the pSeries 680 and S80 server. There is no limit to the quantity of 7014 racks that may be ordered.
- Many 3490 and 3590 tape libraries interfere with the front door of an I/O rack. When planning for the installation of this type of equipment, it is recommended that racks be ordered with a trim kit instead of a front door.
- Two IBM 3490 Model F11s can share an 8-EIA space in the rack (order one placement code for each pair of units). Two IBM 3590 Model B11s or E11s may share a 12-EIA space in the rack (order one placement code for each pair of units).

4.1.1 I/O Drawer description

The I/O Drawer offers the advantage of fully-redundant power and fans that can be serviced without taking the system down. In addition to the hot-swappable fans and power supplies, the drawer supports Ultra-SCSI adapters, which are separately cabled to the two disk six packs. The drawer has a local display panel and reports more information for status monitoring.

The primary I/O Drawer contains the I/O planar and the Service Processor card. In addition, the primary I/O Drawer contains six hot-plug disk bays, one available media bay, one floppy disk drive, one SCSI-2 CD-ROM, 14 PCI slots, one keyboard port, one mouse port, two serial ports, and one parallel port. The I/O subsystem is expandable by attaching up to three additional I/O Drawers to a single CEC.

The 14 PCI I/O slots consist of five 64-bit and nine 32-bit PCI slots. Depending on the media and disk configuration chosen, between two and four of the 14 slots in the first I/O Drawer are used for the Service Processor, storage, and media support. The remaining slots are available to support graphics, communications, and storage in the initial I/O Drawer configuration.

There is a maximum of four I/O Drawers (#6320) per I/O rack, depending on rack type. RIO cables and SPCN cables must be ordered for each additional drawer. The manufactured configuration of I/O Drawers in I/O racks is based on cable lengths ordered.

Figure 3 shows a 10 EIA I/O Drawer viewed from the front.

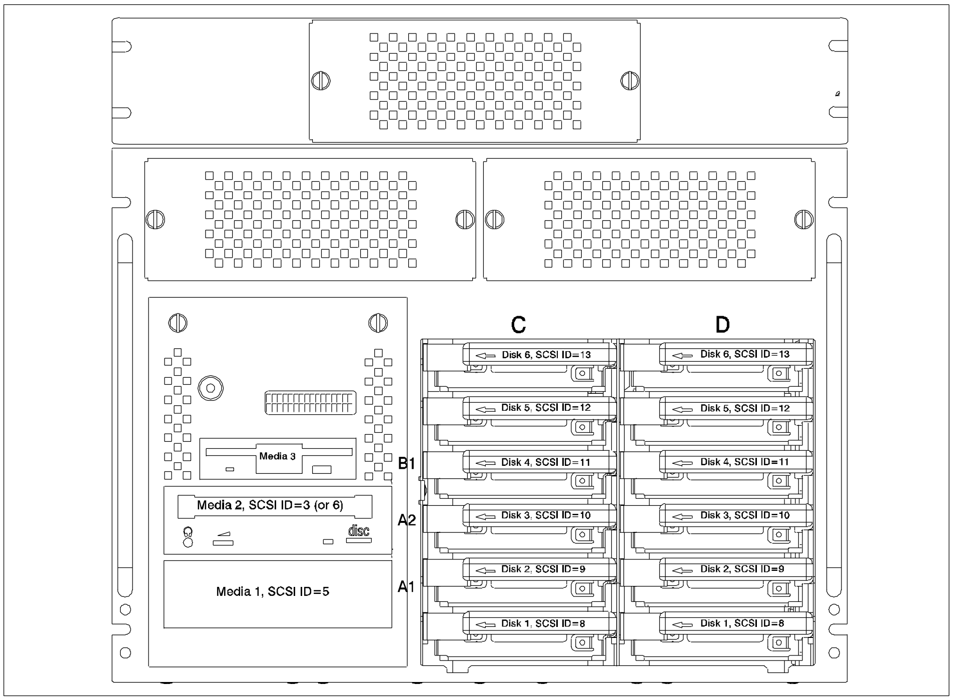


Figure 3. Front view of 10 EIA Drawer

Each drawer has four PCI buses per drawer: Slots 1-4 (PCI bus 1), 5-8 (PCI bus 0), 9-10 (PCI bus 3), and 11-14 (PCI bus 2). Slots 1, 5, 9, 10, and 14 are

64-bit slots. The remaining slots are 32-bit. All supported 32-bit adapters also function in the 64-bit slots. All slots are 33 MHz.

The I/O Drawer has redundant power and cooling. If the I/O racks are ordered with an additional PDU, each drawer can be connected to two PDUs, thus making the I/O system highly available from a power supply point of view.

Figure 4 shows a 10 EIA I/O Drawer viewed from the rear.

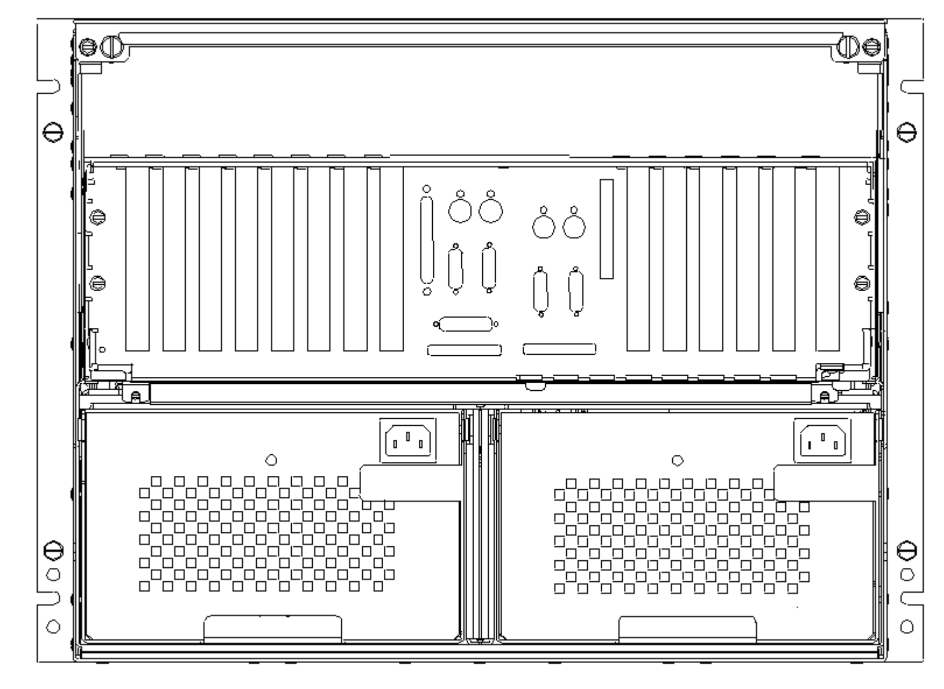


Figure 4. Rear view of 10 EIA I/O Drawer

4.2 System Power Control Network (SPCN)

The function of the SPCN is to allow a single switch on the front of the system CEC to control power to all of the I/O Drawers.

All I/O Drawers and the CEC must be connected in a single SPCN loop. The SPCN can function with any single connection broken, regardless of the location of the open connection.

The rules for cabling the SPCN are the same for pSeries 680 and S80 servers.Two SPCN cables are required for attachment of the first I/O Drawer on each system. Each additional I/O Drawer requires one additional SPCN cable for loop attachment. Figure 5 shows the minimal SPCN loop configuration.

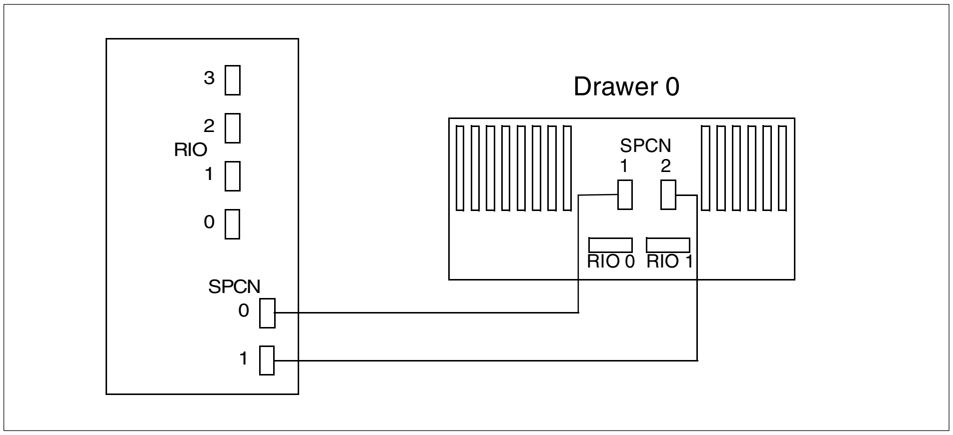


Figure 5. Base SPCN loop with one I/O Drawer

Subsequent I/O Drawers are added to the SPCN loop. The length of SPCN cable chosen will depend on whether the drawer being added is in the same I/O rack as an existing drawer. Figure 6 shows the maximum SPCN configuration of four I/O Drawers.

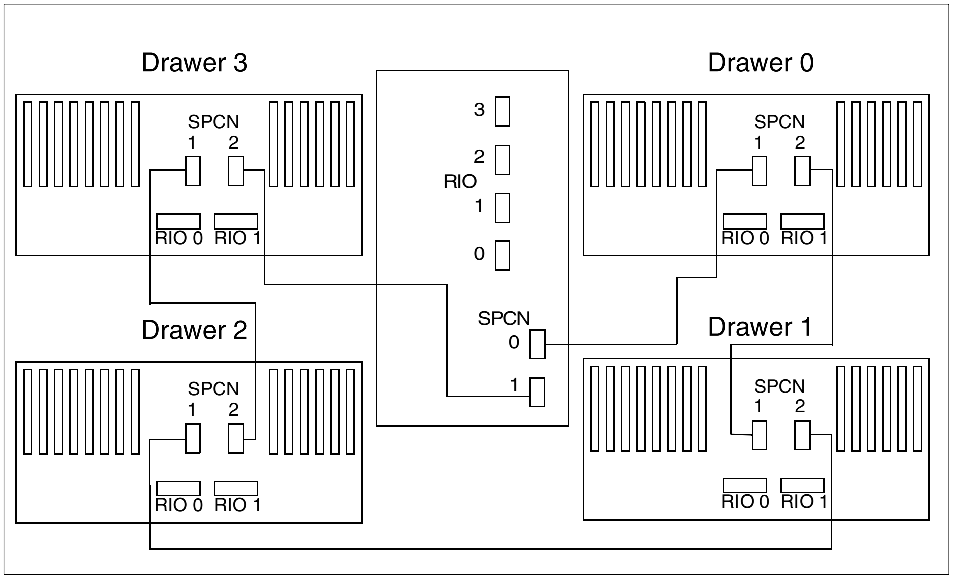


Figure 6. Fully configured SPCN loop with four I/O Drawers

4.3 Remote I/O cabling

The I/O Drawers are connected to the system CEC by a number of different cables. The Remote I/O (RIO) cables provide the means by which data can be transferred between the CPUs and memory in the CEC and the storage and network devices connected using PCI adapters in the I/O Drawers.

RIO connections operate at 250 MB/s in each direction, 500 MB/s duplex. The loop connection provides redundant paths; so, if a failure occurs in part of a cable, a warning message will be displayed, but the system will continue to operate.

The primary I/O Drawer must be installed and connected to RIO port 0 of the CEC. The connection must be made from RI0 port 0 of the CEC to RIO port 0 of the primary I/O Drawer. This connection is required to make the primary drawer the first drawer in the loop that allows the firmware to initialize the system. It also allows the system to find the boot device.

A single RIO loop supports up to two drawers. Each pSeries 680 and S80 server supports up to two RIO loops. The configuration for a single I/O Drawer is very simple. There can only be one RIO loop since there is only one I/O Drawer. An example of this is shown in Figure 7.

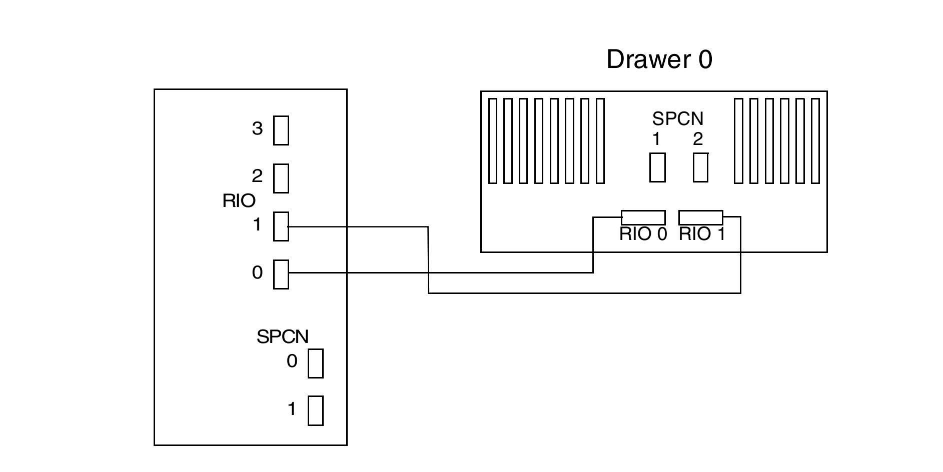


Figure 7. RIO loop with one I/O Drawer

When adding a second drawer, you can cable the additional drawer into the existing RIO loop. This requires one additional RIO cable. An example of this is shown in Figure 8 on page 45.

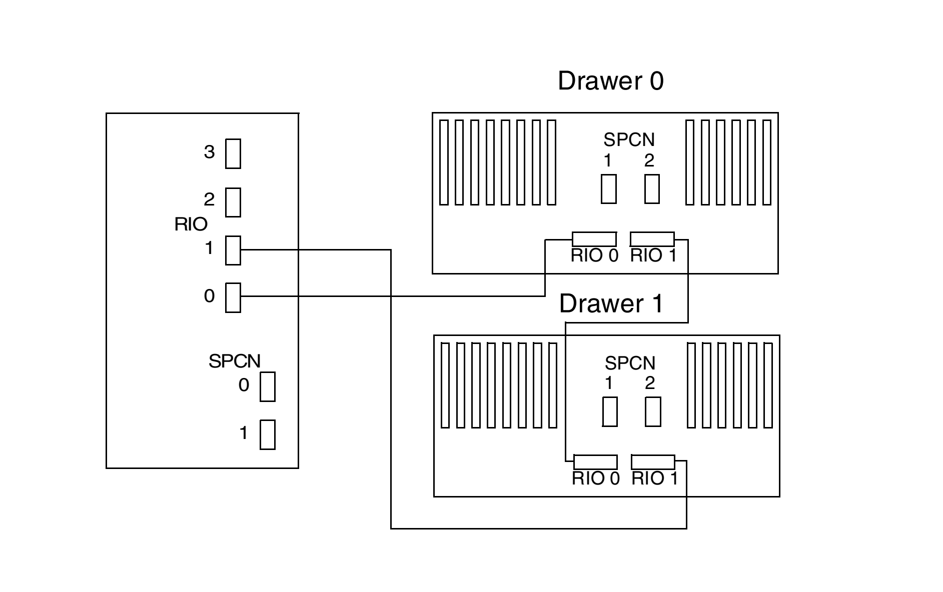


Figure 8. One RIO loop attached to two I/O Drawers

Alternatively, you can cable the second I/O Drawer on a separate loop. This option requires two additional RIO cables and provides each drawer with 500 MB/s of bandwidth. An example of this is shown in Figure 9 on page 46.

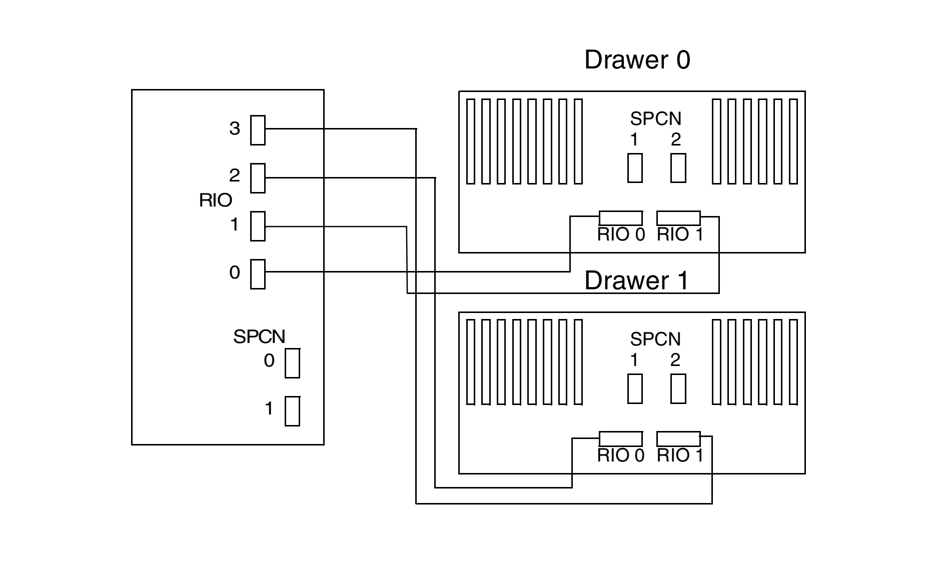


Figure 9. Two RIO loops attached to two I/O Drawers

When adding a third I/O Drawer to a system, the number and length of RIO cables required will depend on the existing RIO loop configuration. If only one RIO loop is configured, adding a third I/O Drawer will require two RIO cables, since the third drawer must be on a new loop. If two RIO loops have been configured, adding a third drawer will require one RIO cable. The length of cable selected will depend on whether or not the third drawer is being installed in the same I/O rack as an existing drawer.

A fully-configured four drawer system must have two RIO loops. An example of this is shown in Figure 10 on page 47.

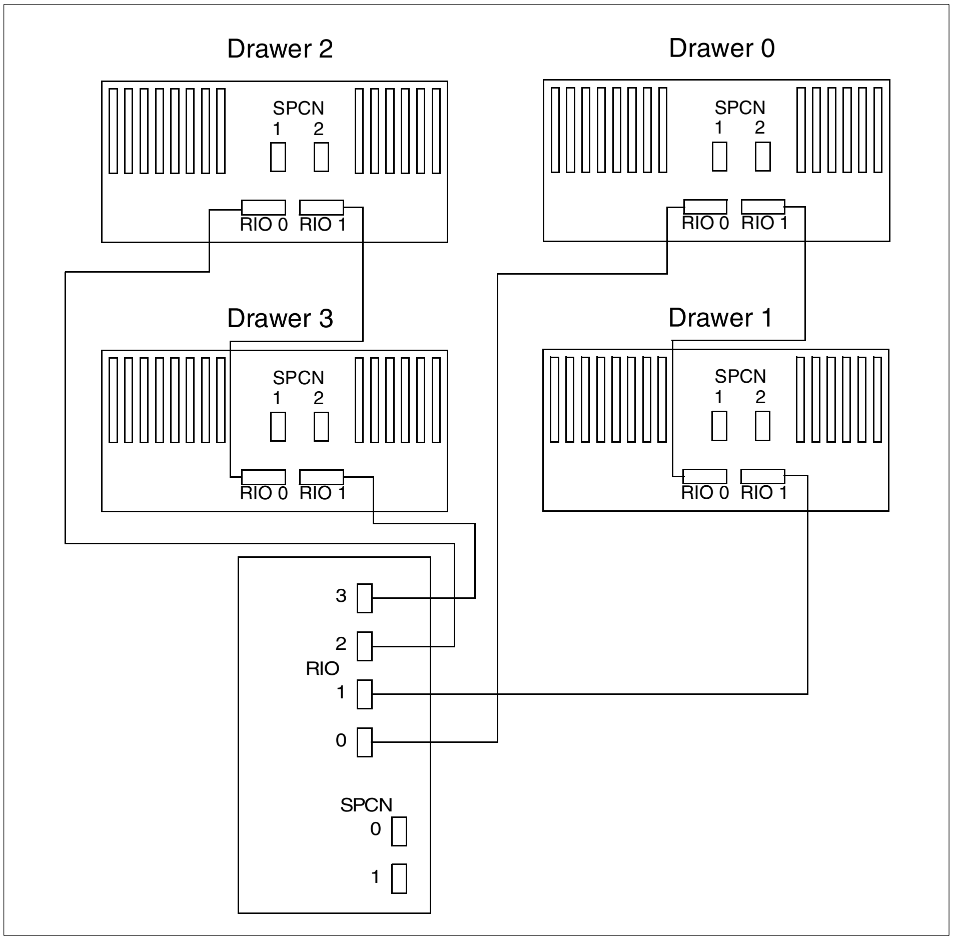


Figure 10. Two RIO loops attached to four I/O Drawers

Chapter 5. Using the Configurator

This chapter includes some guidelines for using the PCRS6000 Configurator to create a valid system order.

Note that the PCRS6000 configurator was used for this section because of the need to configure RS/6000 SP systems for some of the solutions described and to maintain consistency throughout the section. However, the IBM Configurator for e-business should be used for new configurations wherever possible. The IBM Configurator for e-business is available in both a mobile version and a web version from:

```
http://ehone.ibm.com/public/applications/econfig
```

5.1 What is the PCRS6000 Configurator?

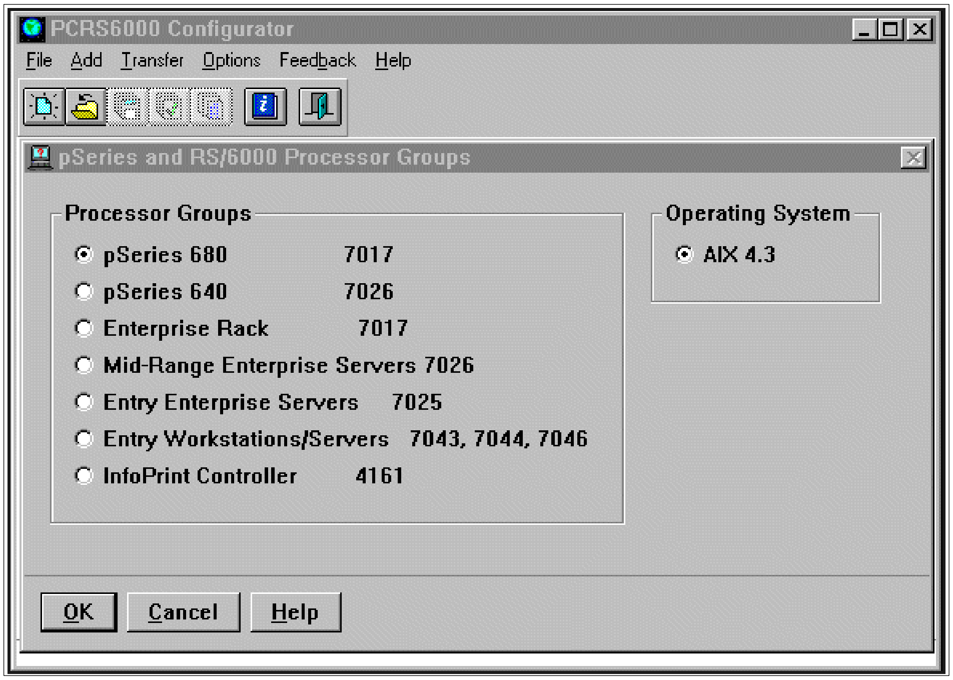
The PCRS6000 Configurator is an application that provides configuration support for hardware, software, and peripherals associated with the IBM @server pSeries and RS/6000 product lines that are available for marketing.

Functions provided by PCRS6000 include:

- The ability to create, save, restore, and print Hardware/Software, hardware only, peripheral only, or software only configurations.
- Hardware/Software interaction for identifying prerequisite and incompatibility conditions.
- Iterative support for reentering product categories and continuous modification and adjustments to the configuration.
- The ability to modify new or existing initial order, MES, or upgrade configurations.
- The ability to modify an installed base prior to beginning MES or upgrade configuration.
- Support for feature exchange and feature conversion.
- The ability to download and upload saved files to the Host/IBMLink.
- Configuration, price, and price total information is displayed as the configuration is built.
- The ability to apply discounts by percentage to individual products or the total configuration.

- Limited native language support is provided for product descriptions and features (for French, German, Italian, Portuguese, Spanish, and Swiss-English).

Figure 11 shows a screenshot of the PCRS6000 configurator.



PCRS6000 Configurator
File Add Transfer Options Feedback Help
pSeries and RS/6000 Processor Groups
OK
Cancel
Help

Processor Groups

● pSeries 680	7017
○ pSeries 640	7026
○ Enterprise Rack	7017
○ Mid-Range Enterprise Servers	7026
○ Entry Enterprise Servers	7025
○ Entry Workstations/Servers	7043, 7044, 7046
○ InfoPrint Controller	4161

Operating System

● AIX 4.3

Figure 11. The PCRS6000 configurator

5.2 Configuring pSeries 680 and S80 servers

When configuring pSeries 680 and S80 Servers, keep the following points in mind:

- In an HACMP configuration, the standard native serial ports on the pSeries 680 and S80 machines are not available for heartbeat cabling; therefore, you must install either the 8-port asynchronous adapter (#2943) with serial-to-serial cable (#3125) or the 128-port asynchronous controller (#2944) with 128-port asynchronous cable (#8136), rack-mountable remote asynchronous node (#8136), RJ-45 to DB-25 converter cable (#8133), and serial-to-serial cable (#3125).
- Either an IBM-supported ASCII terminal with an attachment cable or an IBM-supported graphics display with an attachment cable and graphics

accelerator is required for initial setup and must be available locally for service. OEM asynchronous terminals are not recommended, since they may not transmit the same character sequence and therefore will not be recognized by the Service Processor.

- SP Node Attachment configurations are described in Section 5.4, “Special considerations for SP external node attach” on page 57.

5.2.1 Adapter restrictions

The PCI slot population rules for the pSeries 680 and S80 are very complex and cannot easily be explained in this publication. Refer to the *PCI Adapter Placement Reference*, SA38-0538, for advice and guidance on the supported configurations of adapters in RS/6000 and pSeries systems. The placement reference contains information on both maximum configurations and optimal configurations for best performance.

Extensive configuration rules and checking procedures have been incorporated into the PCRS6000 configurator to ensure valid system configurations. Configurations generated without utilizing either the PCRS6000 configurator or the IBM Configurator for e-business may create orders that cannot be manufactured, resulting in possible order rejection and/or delayed delivery.

There are 14 PCI slots split across 4 PCI buses in each I/O Drawer. Slots assigned to each bus are: 1-4, 5-8, 9-10, and 11-14. Slots 1, 5, 9, 10, and 14 are 64-bit slots. The remaining slots are 32-bit. 32-bit adapters can be used in 64-bit slots.

System maximum limits for adapters and devices may not provide optimal system performance. These limits are given for connectivity and functional assurance.

Configuration limitations have been established to help ensure appropriate PCI bus loading, adapter addressing, and system and adapter functional characteristics when ordering primary and secondary I/O Drawers. These I/O Drawer limitations are in addition to the individual adapter limitations shown in the feature descriptions section.

The Service Processor must always be located in slot 8 of the primary I/O Drawer. To ensure proper functionality of the Service Processor, certain other adapters, such as SSA adapters, can not share a PCI bus with the Service Processor.

The Service Processor provides two serial ports for direct console attachment and support processor use only. These ports are not to be utilized for other functions, such as HACMP heartbeat cables or UPS interface.

Only two SSA adapters are allowed per PCI bus, with a maximum of 26 per system. A maximum of five SSA adapters are allowed in the primary I/O Drawer, while up to seven are allowed in each secondary I/O Drawer.

The S/390 ESCON Channel Adapter (#2751) is limited to four per system and two per I/O Drawer. These adapters must be located in drawers attached using the primary I/O loop for addressing considerations.

The RS/6000 SP System Attachment Adapter (#8396) must always be located in slot 10 of the primary I/O Drawer. No adapters may be installed in slot 9 or 11 of the primary drawer when the SP attachment adapter is installed.

The following adapters are limited to one per system:

- (#2708) Eicon ISDN DIVA PRO 2.0 PCI S/T Adapter
- (#2830) POWER GXT130P Graphics Adapter
- (#2838) POWER GXT120P Graphics Adapter (Model S80 only)
- (#6326) Service Processor
- (#8396) RS/6000 SP System Attachment Adapter

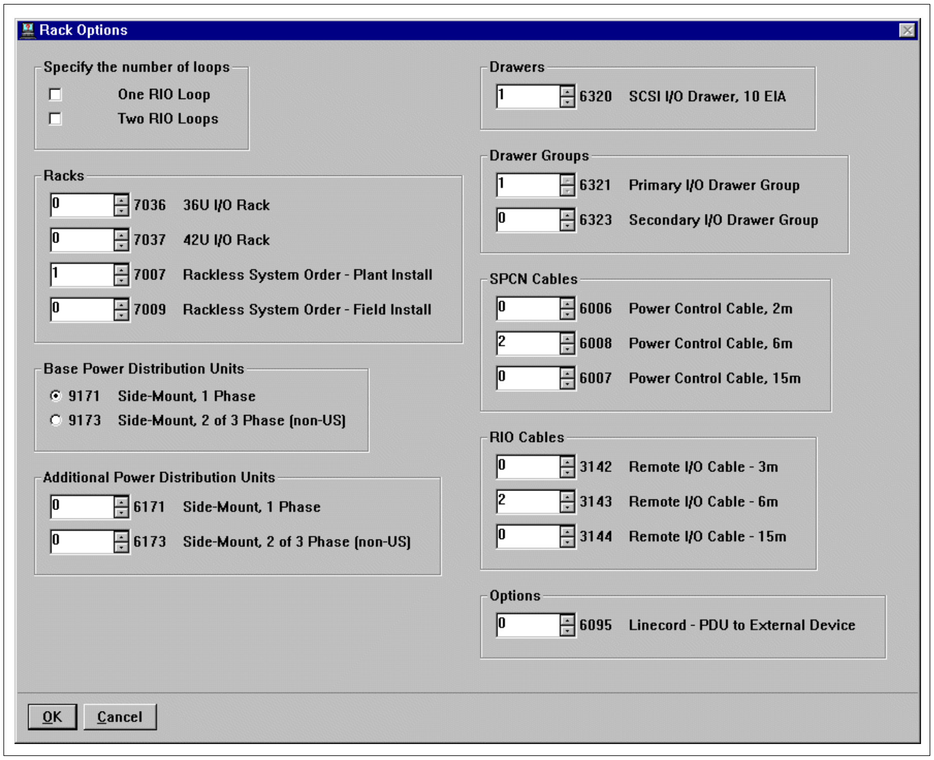
5.2.2 Two Servers sharing an I/O rack

In special circumstances, the pSeries 680 and S80 systems can be ordered without an I/O rack. This configuration can be a good way of having multiple systems without requiring an I/O rack for every system. This option saves machine room floor space by reducing the footprint of the systems. It also reduces the overall cost of the system.

This configuration is not difficult to achieve, although some special considerations must be taken when ordering. If you are ordering a rackless system as part of an HACMP cluster, it may be easier to use the standard `Solution Pack` configuration option described in Section 5.3, “High-Availability Solutions” on page 54.

When placing an order for multiple systems, the first system should be configured as normal. 10 EIA units of space in the I/O rack should be reserved for the I/O Drawer of the second system. This is achieved by selecting feature code #0177 as a rack contents specify feature.

When configuring the second system, there will be a feature change that indicates a rackless system order. The feature is selected in the rack options dialog box, as shown in Figure 12.



Rack Options

Specify the number of loops
□ One RIO Loop
□ Two RIO Loops

Racks
0 7036 36U I/O Rack
0 7037 42U I/O Rack
1 7007 Rackless System Order - Plant Install
0 7009 Rackless System Order - Field Install

Base Power Distribution Units
● 9171 Side-Mount, 1 Phase
○ 9173 Side-Mount, 2 of 3 Phase (non-US)

Additional Power Distribution Units
0 6171 Side-Mount, 1 Phase
0 6173 Side-Mount, 2 of 3 Phase (non-US)

Drawers
1 6320 SCSI I/O Drawer, 10 EIA

Drawer Groups
1 6321 Primary I/O Drawer Group
0 6323 Secondary I/O Drawer Group

SPCN Cables
0 6006 Power Control Cable, 2m
2 6008 Power Control Cable, 6m
0 6007 Power Control Cable, 15m

RIO Cables
0 3142 Remote I/O Cable - 3m
2 3143 Remote I/O Cable - 6m
0 3144 Remote I/O Cable - 15m

Options
0 6095 Linecord - PDU to External Device

Figure 12. Rack options dialog box

There are two ways of configuring a rackless system:

- (#7007) **Rackless System Order - Plant Install**: This feature allows a system containing a single (primary) I/O Drawer to be ordered without an I/O rack. The system primary I/O Drawer will be factory installed in the I/O rack of a companion system that specifies feature code #0177 in order to provide the required 10 EIA units rack space and two power outlets. Both systems must be ordered on the same customer order with the same scheduled factory delivery date.
- (#7009) **Rackless System Order - Field Install**: This feature allows a system containing a single (primary) I/O Drawer to be ordered without an I/O rack. The system primary I/O Drawer will be field installed in the I/O rack of another system that provides the 10 EIA units rack space and two power outlets required for the I/O Drawer.

Note

Model S80 and pSeries 680 I/O Drawers are only supported in Model S80 or pSeries 680 I/O racks. Orders containing feature #7007 are subject to cancellation if a companion system is not ordered with the same factory delivery schedule and feature #0177 specified.

5.3 High-Availability Solutions

Both the pSeries 680 and S80 have high-availability (HA) cluster solutions available. The rules for configuration are identical. This section describes the minimum requirements for an HA cluster solution and how to use the PCRS6000 Configurator to create a valid configuration.

5.3.1 Minimum requirements

Each HA solution order must include a minimum of the following:

- Two Model S80 or pSeries 680 servers, each incorporating the following:
  - One HA Cluster Solution Indicator (#0500).
  - Two I/O rack PDUs (Power Distribution Units) for redundancy.
  - Two 9.1 GB or larger SCSI boot disks mounted in separate 6-pack backplanes within the primary I/O Drawer.
  - One Async adapter for cluster heartbeat and messaging.

    Note: Only one serial cable (#3125 or #8133) is required for the two servers.
  - Two LAN Adapters for LAN attachment and backup heartbeat and messaging.
  - Two PCI SSA Adapters.
  - One HACMP ES Version 4.3.1, or later license.
  - One AIX Version 4.3.3, or later license.

  Note: Additional features may be added, as desired.
- Two - 7133-D40 Serial Disk Subsystems, each incorporating the following:
  - Four Advanced SSA Disk Modules.
  - One Manufacturing Integration Code (Rochester = #0987, Dublin = #0970).
  - Six Advanced SSA Cables.

Note: Additional optional 7133-D40 features may be added, as desired.

5.3.2 Configuring an HA solution using PCRS6000

The PCRS6000 Configurator has options for configuring standard HA solutions. Initiate an HA solutions configuration by selecting **Add** -> **pSeries and RS/6000** -> **Solution Package**, as shown in Figure 13.

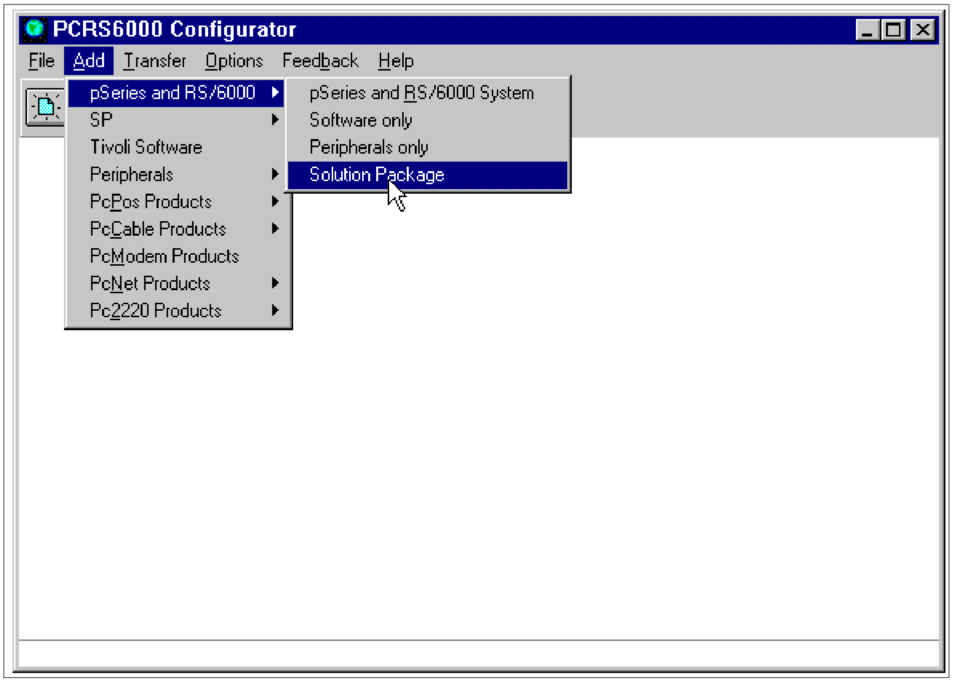
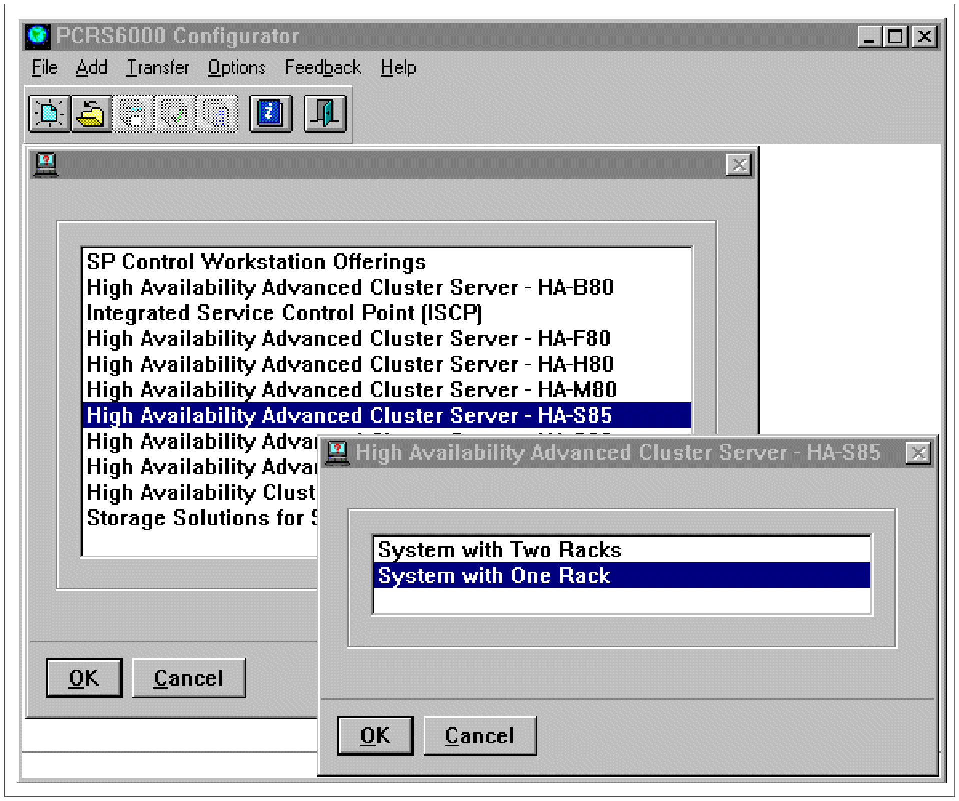


Figure 13. Starting an HA configuration

Now select the appropriate HA solution (either the HA-S80 [for Model S80] or HA-S85 [for pSeries 680]). You will then be presented with a panel asking if you want a single rack or a two rack solution. A single rack solution saves floorspace by using the same rack to hold the primary I/O Drawers from each system. This assumes that you will not be configuring additional racks for external storage devices. Figure 14 on page 56 shows the two panels.



SP Control Workstation Offerings
High Availability Advanced Cluster Server - HA-B80
Integrated Service Control Point (ISCP)
High Availability Advanced Cluster Server - HA-F80
High Availability Advanced Cluster Server - HA-H80
High Availability Advanced Cluster Server - HA-M80
High Availability Advanced Cluster Server - HA-S85
High Availability Advai
High Availability Advai
High Availability Clust
Storage Solutions for S

System with Two Racks
System with One Rack

Figure 14. Selecting the HA solution

The configurator then automatically generates two systems of the required type with the minimum processor, memory and PDU requirements. You must then add appropriate features for AIX, HACMP, LAN adapters, and so forth to meet the minimum requirements, as specified in Section 5.3, “High-Availability Solutions” on page 54, plus any additional features required.

Once you have finished adding features to both systems remember to use the *Validate All* option to make sure that the two systems are validated together. This is shown in Figure 15 on page 57.

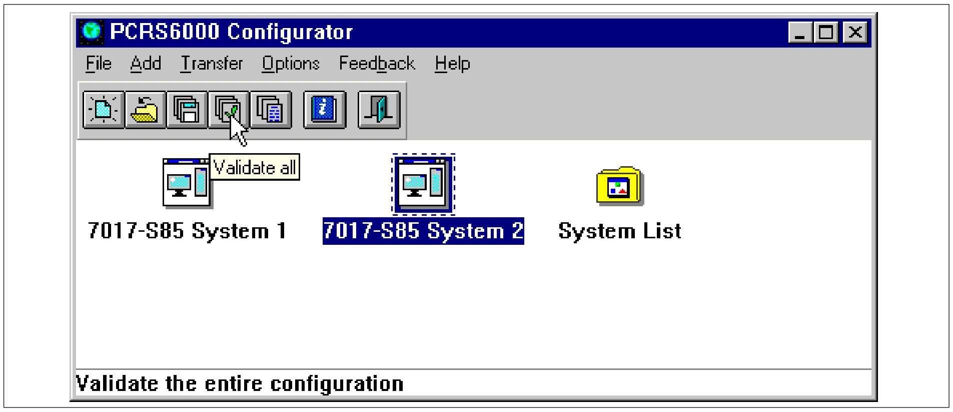


Figure 15. HA Solution Validation

5.4 Special considerations for SP external node attach

The pSeries 680 and S80 can function as an attached SMP server within the IBM RS/6000 SP environment operating under the control of the IBM Parallel Systems Support Programs for AIX. Up to 16 such systems may be attached in a single configuration. This interconnection can be accomplished by utilizing an Ethernet connection and, optionally, using the IBM RS/6000 SP System Attachment Adapter (#8396), which provides connection to the SP Switch. Note that connection to SP systems using the SP Switch2 (Colony) switch are not currently supported.

The SP-attached server requires attachment to an RS/6000 SP system with at least one 79 inch frame containing at least one node and (if switch attached) the SPS (16-port) switch. No other switches or frames are supported.

For full details on configuring pSeries 680 and S80 servers as a node in an SP system, refer to *RS/6000 SP: Planning, Volume 1 Hardware and Physical Environment*, GA22-7280.

5.4.1 Features required on the pSeries 680 and S80

Some I/O adapters available for the pSeries 680 and S80 are not supported in the SP environment and must be removed. Refer to the RS/6000 9076-550 Sales Manual pages for a list of the currently supported adapters. Note that a list is also available in the “Configure RS/6000 models attached as SP nodes” section of the PCRS6000 help file.

A separate chargeable license for IBM Parallel System Support Programs for AIX Version 3.1 (5765-D51) or later must be ordered against each SP-attached server serial number if it is to function as an attached SMP server within the IBM RS/6000 SP environment.

Each pSeries 680 and S80 system that is to function as an attached SMP server within the IBM RS/6000 SP environment must have a minimum of one Ethernet adapter. This adapter must be recognized by the system as *en0* and must reside in slot 5 of the primary I/O Drawer.

The RS/6000 SP-attached server must have the latest system and Service Processor firmware (microcode) installed.

If this is to be a switchless attachment to the pSeries 680 and S80 servers, then the RS/6000 SP system must also be switchless. There must be available (unused) switch node numbers in your SP System Data Repository, even though the switch is not being used. A server cannot be attached to a single frame system where all 16 nodes are occupied.

If the SP system has an SP Switch, then each pSeries 680 and S80 system must have an RS/6000 SP System Attachment Adapter (#8396) installed, and this must always be located in slot 10 of the primary I/O Drawer. No adapters may be installed in slots 9 or 11 of the primary drawer when the SP attachment adapter is installed. Only one RS/6000 SP System Attachment Adapter is permitted in each SP-attached server; this is consistent with the rules for RS/6000 SP system nodes. A 10 meter switch cable is provided with the SP-attach order through the RS/6000 SP system configurator.

5.4.2 Features required on the SP system

If you are configuring a system order for a machine that is to be SP-attached to an existing SP system, you must also perform an MES order against one of the frames in the SP system to configure it for attaching an external node. Figure 16 on page 59 shows the appropriate dialog selection to achieve this process.

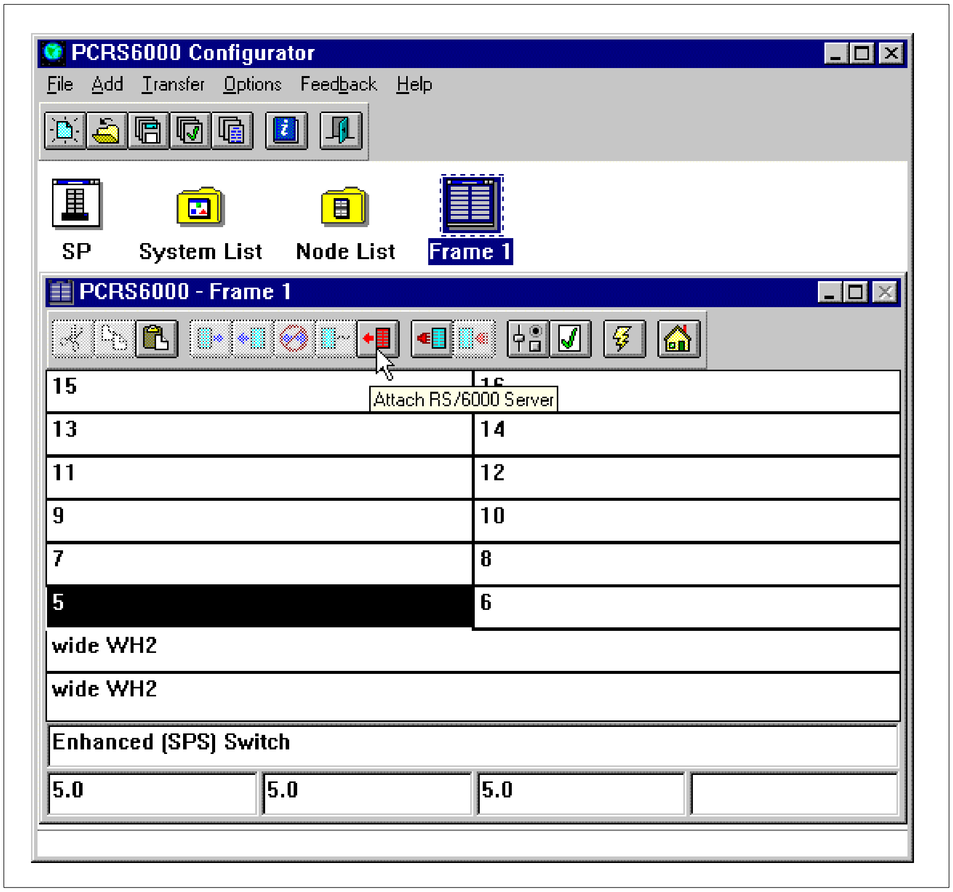


Figure 16. Configuring an SP-attached server

The RS/6000 SP system configurator session for all variations of SP-attach will include the following when adding a SP node attachment:

- (#9122) NODE ATTACHMENT FEATURE: An order for this feature will generate an order for the two 15 Meter RS-232 cables for hardware control and S1TERM connectivity between the CWS and the SP-attached server. It also will generate an order for a 10 meter ground cable. It also traps some data so that the RS/6000 SP system configurator session can keep track of how many nodes (real and logical) are in the system.
- (#9310) SWITCH CONNECTION CABLE: This feature is required only if the SP-attached server is switch attached. It results in the ordering of one 10 meter switch connection cable. The 10 meter cable is the only supported length at this time.

- (#9123) FRAME ATTACHMENT FEATURE: This feature keeps track of how many frames are currently in your RS/6000 SP system. Since the SP-attached server is both a logical node and a logical frame in the PSSP code logic, it is important to track this information to avoid exceeding allowable RS/6000 SP system limits for the number of frames.
- (#9222) NODE ATTACHMENT ETHERNET BNC BOOT FEATURE: This feature will get a BNC cable to allow RS/6000 SP system Ethernet communications and booting with your SP-attached server, whether switch-attached or not.
- (#9223) NODE ATTACHMENT ETHERNET TWISTED PAIR BOOT FEATURE: This feature tracks the choice to incorporate the SP-attached servers as part of an RS/6000 SP system Ethernet Twisted Pair network, but it provides no twisted pair cable. As in the past, the customer is responsible for providing their own twisted pair Ethernet cables.

5.4.3 Features required on the Control Workstation

The Control Workstation (CWS) must have sufficient serial port connections and CPU power to support the RS/6000 SP-attachment. The requirement is two RS-232 attachments for each Model S80 or pSeries 680 server that is to be attached. If the CWS does not have sufficient available RS-232 ports, you may have to add one or more 8-port (#2943) or 128-port (#2944) asynchronous adapters, assuming there are enough free adapter slots (if not, a larger capacity CWS will be required). An RS/6000 Model F50 is the minimum recommended system in this environment.

IBM Parallel System Support Programs for AIX, Version 3.1 (5765-D51) with APAR IY13025, or IBM Parallel System Support Programs for AIX, Version 3.2 (5765-D51) with APAR IY13026 is required for a pSeries 680 system to function as an attached SMP server within the IBM RS/6000 environment.

5.5 Clustered Enterprise Servers

Up to sixteen pSeries 680 and S80 servers may be incorporated into a single cluster managed by the IBM Parallel Systems Support Programs for AIX. This implementation, Clustered Enterprise Servers (CES), requires an 7025-F50 control workstation for cluster control, but does not require attachment to a 9076 SP frame.

5.5.1 Features required on the pSeries 680 and S80

Each server in the cluster must be attached to the control workstation using the following two cable features and customer supplied Ethernet LAN connections:

- Clustered Server Serial Port-to-Control Workstation Cable (#3150)
- Clustered Server Control Panel-to-Control Workstation Cable (#3151)

These cables may be ordered with either the Model S80 or pSeries 680 clustered server, or the control workstation, but it is suggested that you order them with the server.

IBM Parallel Systems Support Programs for AIX, Version 3.2, or later, with APAR IY13026, is required for each Model S80 or pSeries 680 server in the cluster.

AIX Version 4.3.3, or later, is required for each Model S80 or pSeries 680 server in the cluster.

A minimum of one Ethernet adapter is required for each Model S80 or pSeries 680 server in the cluster. This adapter must be recognized by the clustered server as *en0* and must reside in slot 5 of the primary I/O Drawer.

The supported adapters for this system LAN are:

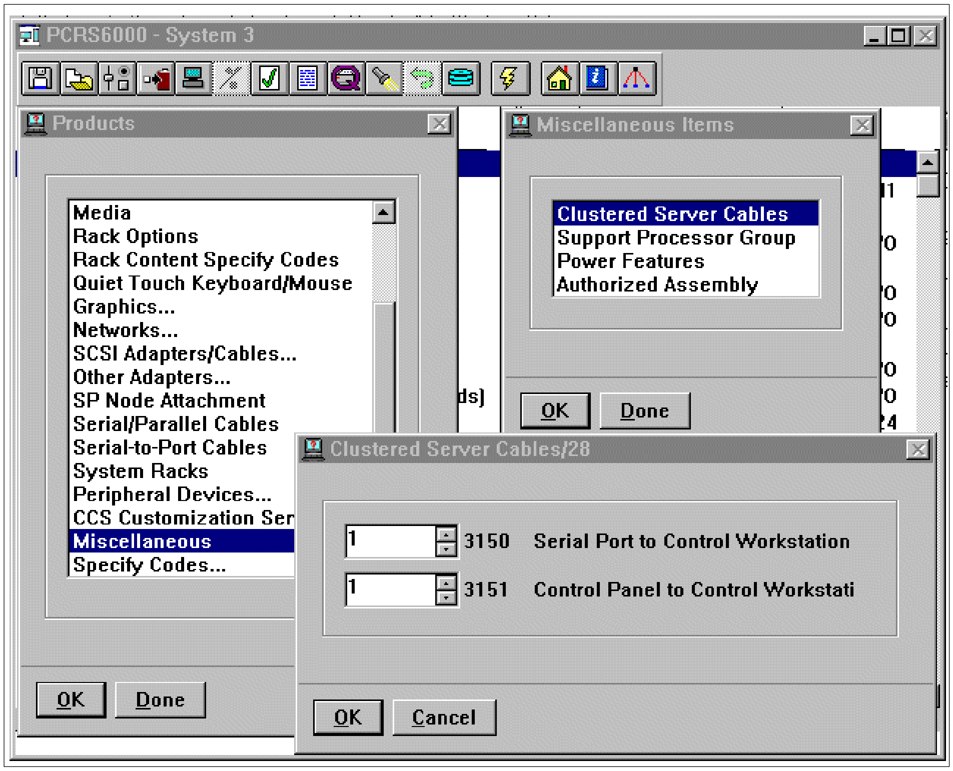
- For twisted-pair cable connection:
  - 10/100 Ethernet 10BaseTX Adapter (#2968)
  - 10 MB AUI/RJ-45 Ethernet Adapter (#2987)
- For BNC cable connection:
  - 10 MB BNC/RJ-45 Ethernet adapter (#2985)

For more information, refer to the document *RS/6000 SP: Planning, Volume 1 Hardware and Physical Environment*, GA22-7280.

Some I/O adapters supported on the Model S80 and pSeries 680 servers, when utilized in a non-clustered environment, are not supported in the CES environment and must be removed. Refer to the RS/6000 9076-550 Sales Manual pages for a list of the currently supported adapters. Note that a list is also available in the *Configure RS/6000 models attached as SP nodes* section of the PCRS6000 help file.

5.5.2 Configuring CES using PCRS6000

You will need to configure each server and a control workstation separately for this solution. When starting the PCRS6000 session, do not select the **This machine will function as an SP node** option. CES is identified to the configurator by selecting the **Clustered Server Cables** option from the **Miscellaneous Items** option of the **Products** panel (see Figure 17). You will need to select one each of #3150 and #3151.



1 3150 Serial Port to Control Workstation

1 3151 Control Panel to Control Workstati

Figure 17. Identifying a server as a CES machine

Configure each server to meet the minimum specifications, as outlined in Section 5.5, “Clustered Enterprise Servers” on page 60, and then configure an RS/6000 F50 as a control workstation by selecting the **Solutions Package** option from the **pSeries and RS/6000** selection list (see Figure 13 on page 55) and then selecting **SP Control Workstation Offerings**, followed by **F50 Control Workstation Small Config**.

5.6 Upgrading an RS/6000 S70 Advanced to a Model S80

Installed Model S70 Advanced systems can be converted to Model S80 systems. This conversion requires replacement of the entire Central Electronic Complex (CEC), including processors and system memory. The support processor group, which includes the Service Processor card, must also be exchanged to provide the proper firmware for interfacing to the Model S80 CEC. The Model S70 Advanced I/O rack and drawers, PCI adapter cards, hard disks, and media devices carry forward to the upgraded Model S80 system. This conversion preserves the existing system serial number.

Note that upgrades are available to a Model S80 with 450 MHz or 600 MHz processors. Upgrades to a pSeries 680 are not permitted.

The model upgrade consists of the following items:

- Central Electronics Complex (CEC): The Model S70 Advanced CEC being replaced is returned to IBM.
- Model S80 labels: Preserving the customers serial number.
- Model S80 publications.

Processor cards and I/O Drawer attachment hardware must be ordered separately. The following items are *required* and must be ordered separately:

- One First RS64 6-way Processor Card. Choose from:
  - 6-Way 600 MHz RS64 IV Processor Card, 16 MB L2 Cache (#5320)
  - 6-Way 450 MHz RS64 III Processor Card, 8 MB L2 Cache (#5318)
- 2 GB Minimum System Memory (See Section 5.6.2, “S70 Advanced to S80 memory conversion” on page 65.)
- Cabling for connection to the primary I/O Drawer includes:
  - One System Control And Initialization Cable - Upgrade Indicator (#8006)
  - Two Power Control Cable - Upgrade Indicator (#8008)
  - Two Remote I/O Cables (#3143 or #3144)
- One Remote I/O Hub - Dual Loop - (#6503)
- One Support Processor Group - Model Upgrade Only (#8326)

Note

The S70 Advanced Support Processor being replaced is returned to IBM.

The following items must be carried forward from the system being upgraded:

- I/O Drawer attachment cables (#8006 and #8008 above)
- I/O rack
- Primary and secondary I/O Drawers (including backplanes, SCSI adapters, and cables)
- PCI adapters
- Hard disk drives
- Media devices

5.6.1 S70 Advanced to S80 processor conversion

Model S70 Advanced systems that are converted to Model S80 systems will require replacement of all system processors. The first Model S70 Advanced 4-way processor card can be replaced with the first Model S80 6-way 450 MHz processor card by feature conversion. This processor conversion is only available at the time of initial model upgrade. A maximum of one S70 Advanced 4-way processor card may be converted to one S80 6-way processor card for each system being converted. The existing Model S70 Advanced 4-way processor card being replaced is returned to IBM.

Table 11 provides the processor feature conversions that are available to Model S70 Advanced customers who convert their systems to Model S80 systems.

Table 11. S70 Advanced to S80 processor conversion

From	To
(#0504) RS64 II, 4-way SMP, 262 MHz	(#5318) RS64 III, 6-way, SMP, 450 MHz
(#5312) RS64 II, 4-way SMP, 262 MHz	(#5318) RS64 III, 6-way, SMP, 450 MHz
(#5314) RS64 II, 4-way SMP, 262 MHz	(#5318) RS64 III, 6-way, SMP, 450 MHz
(#5316) RS64 II, 4-way SMP, 262 MHz	(#5318) RS64 III, 6-way, SMP, 450 MHz
(#9404) RS64, 4-way SMP, 125 MHz	(#5318) RS64 III, 6-way, SMP, 450 MHz
(#0504) RS64 II, 4-way SMP, 262 MHz	(#5320) RS64 IV, 6-way, SMP, 600 MHz
(#5312) RS64 II, 4-way SMP, 262 MHz	(#5320) RS64 IV, 6-way, SMP, 600 MHz
(#5314) RS64 II, 4-way SMP, 262 MHz	(#5320) RS64 IV, 6-way, SMP, 600 MHz
(#5316) RS64 II, 4-way SMP, 262 MHz	(#5320) RS64 IV, 6-way, SMP, 600 MHz
(#9404) RS64, 4-way SMP, 125 MHz	(#5320) RS64 IV, 6-way, SMP, 600 MHz

5.6.2 S70 Advanced to S80 memory conversion

Model S70 Advanced systems that are converted to Model S80 systems will require replacement of all system memory. This memory replacement will be implemented by converting existing Model S70 Advanced memory features to Model S80 memory features. The memory conversions are only available at the time of initial model upgrade and are allowed on a one-for-one feature conversion basis. A maximum of four Model S70 Advanced memory features may be converted to Model S80 memory features for each system being converted from S70 Advanced to S80. The existing Model S70 Advanced memory being replaced is returned to IBM.

The following memory feature conversions are available to Model S70 Advanced customers who convert their systems to Model S80 systems:

Table 12. S70 Advanced to S80 memory conversion

From	To
1024 GB	
(#4173) 1024 MB R1 Memory	(#4190) 1024 MB Memory
(#4174) 1024 MB R1 Memory Select	(#4190) 1024 MB Memory
2048 MB	
(#4171) 512 MB R1 Memory	(#4191) 2048 MB Memory
(#9168) Base 512 MB Memory	(#4191) 2048 MB Memory
(#4173) 1024 MB R1 Memory	(#4191) 2048 MB Memory
(#4174) 1024 MB R1 Memory Select	(#4191) 2048 MB Memory
(#4175) 2048 MB R1 Memory	(#4191) 2048 MB Memory
(#4176) 2048 MB R1 Memory Select	(#4191) 2048 MB Memory
4096 MB	
(#4171) 512 MB R1 Memory	(#4192) 4096 MB Memory
(#9168) Base 512 MB Memory	(#4192) 4096 MB Memory
(#4173) 1024 MB R1 Memory	(#4192) 4096 MB Memory
(#4174) 1024 MB R1 Memory Select	(#4192) 4096 MB Memory
(#4175) 2048 MB R1 Memory	(#4192) 4096 MB Memory
(#4176) 2048 MB R1 Memory Select	(#4192) 4096 MB Memory
(#4177) 4096 MB R1 Memory	(#4192) 4096 MB Memory

From	To
(#4178) 4096 MB R1 Memory Select	(#4192) 4096 MB Memory
8192 MB	
(#4171) 512 MB R1 Memory	(#4193) 8192 MB Memory
(#9168) Base 512 MB Memory	(#4193) 8192 MB Memory
(#4173) 1024 MB R1 Memory	(#4193) 8192 MB Memory
(#4174) 1024 MB R1 Memory Select	(#4193) 8192 MB Memory
(#4175) 2048 MB R1 Memory	(#4193) 8192 MB Memory
(#4176) 2048 MB R1 Memory Select	(#4193) 8192 MB Memory
(#4177) 4096 MB R1 Memory	(#4193) 8192 MB Memory
(#4178) 4096 MB R1 Memory Select	(#4193) 8192 MB Memory
(#4179) 8192 MB R1 Memory	(#4193) 8192 MB Memory
(#4180) 8192 MB R1 Memory Select	(#4193) 8192 MB Memory
16384 MB	
(#4171) 512 MB R1 Memory	(#4194) 16384 MB Memory
(#9168) Base 512 MB Memory	(#4194) 16384 MB Memory
(#4173) 1024 MB R1 Memory	(#4194) 16384 MB Memory
(#4174) 1024 MB R1 Memory Select	(#4194) 16384 MB Memory
(#4175) 2048 MB R1 Memory	(#4194) 16384 MB Memory
(#4176) 2048 MB R1 Memory Select	(#4194) 16384 MB Memory
(#4177) 4096 MB R1 Memory	(#4194) 16384 MB Memory
(#4178) 4096 MB R1 Memory Select	(#4194) 16384 MB Memory
(#4179) 8192 MB R1 Memory	(#4194) 16384 MB Memory
(#4180) 8192 MB R1 Memory Select	(#4194) 16384 MB Memory

Chapter 6. Hardware architecture

The pSeries 680 and S80 have two or more physical enclosures connected by power controller and I/O cables. The first enclosure is the CEC, which contains processor cards, memory, a memory controller, and a Remote I/O (RIO) controller, along with power supplies and cooling fans. The physical layout of the front of the CEC is shown in Figure 18.

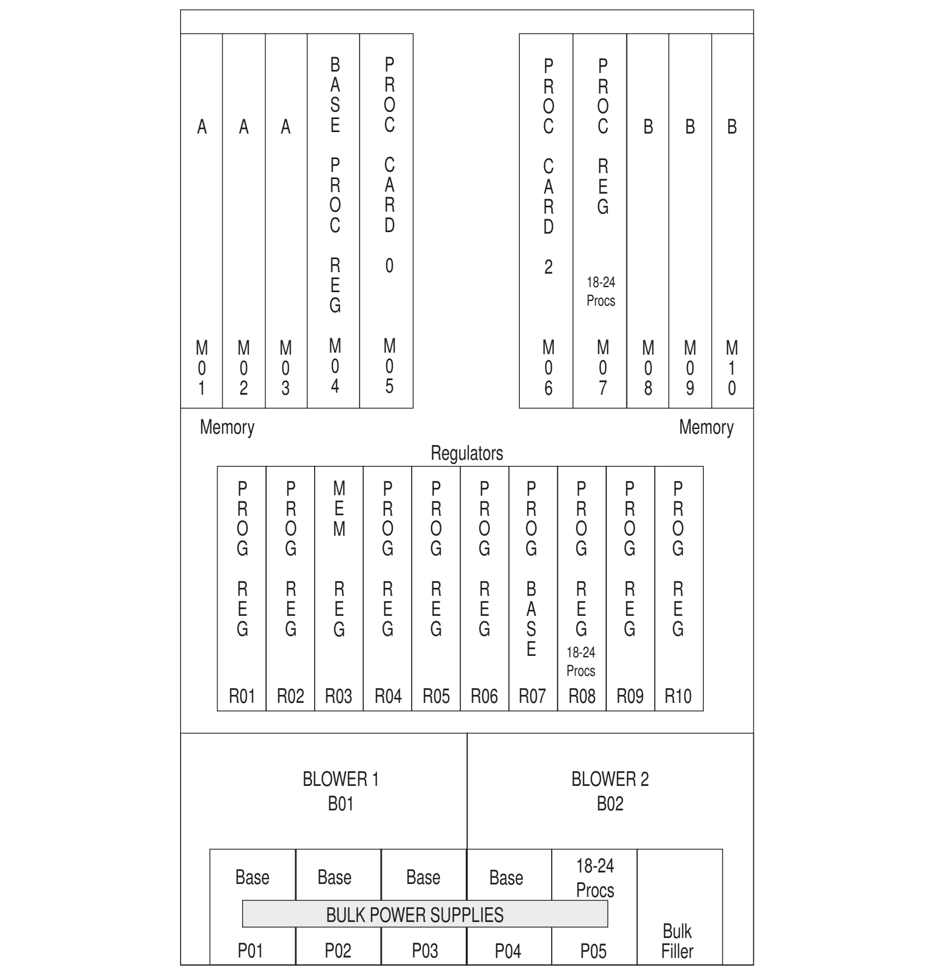


Figure 18. Front view of the pSeries 680 and S80 CEC

The bulk power supplies are installed at the base of the CEC. There will be one more power supply installed than is required to support the chosen configuration (N+1 redundancy). Five power supplies (designated P01 through P05) are required for a fully-configured system. Note that the pSeries 680 and S80 both use grade 1 components for the power subsystem, with separate power regulators. This increases the size of the CEC but adds to overall systems reliability.

Note

Only authorized service personnel should attempt to access components inside the CEC.

Above the bulk power supplies can be seen two of the four blowers used to cool the system. The other two blowers are visible from the rear of the CEC. The blowers use N+1 redundancy and are hot-swappable. The power regulators are above the blowers. Ten are required for a fully configured system.

Above the regulators are ten slots (designated M01 through M10) for holding *books* containing processors, memory or logic cards. Note that each slot has a dedicated function. Slots M05 and M06 are for processors and slots M01 through M03 and M08 through M10 are for memory. The systems architecture allows the maximum number of processors and memory to be installed without compromise.

The books are installed in the slots, which are part of the backplane. The backplane is a rigid construct, encased in sheet metal, running vertically up the middle of the CEC. There are 24 slots on the backplane, ten at the front and fourteen at the rear. Book packaging is described in Section 6.1, “Book packaging” on page 70. Placement rules for memory cards are described in Section 6.5, “Memory cards and quads” on page 78.

The rear of the CEC is shown in Figure 19 on page 69. At the base of the CEC is the AC box, which provides connection to the customer supply. Above this are blowers B03 and B04, and above them the rear of the regulator array. The logic card for the System Power and Control Network (SPCN) is above blower B04. The SPCN allows the I/O drawers to be powered up or down from the CEC (see Section 4.2, “System Power Control Network (SPCN)” on page 42 for more details).

The remaining 14 slots can be seen at the top of the CEC. These hold ten of the 16 memory cards, two processor cards, and two logic cards. The first logic card is the I/O hub card, which controls the Remote I/O (RIO) ports. The RIO subsystem is described in Chapter 4, “Remote I/O Subsystem” on page 39. The other logic card contains the system clock and the JTAG port, which is used to connect the CEC to the Service Processor (see Section 8.4, “Service processor” on page 104 for more details).

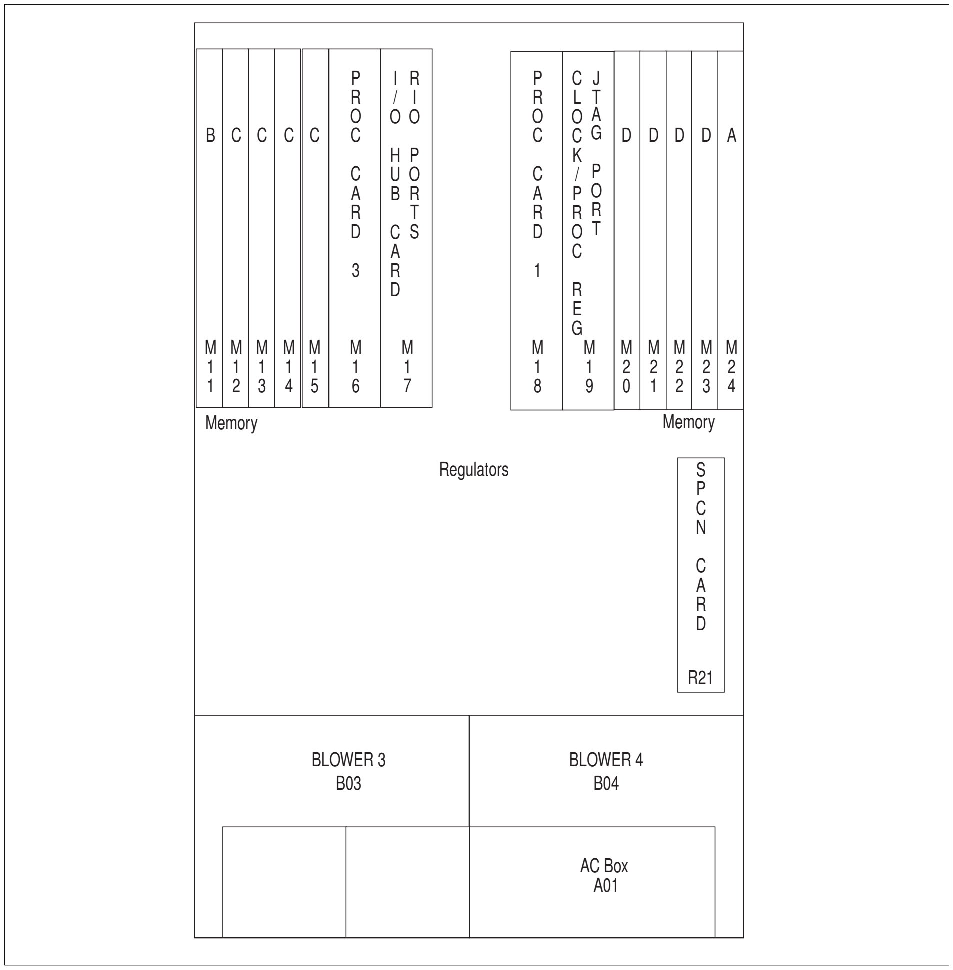


Figure 19. Rear view of the pSeries 680 and S80 CEC

The second enclosure is a standard 19-inch black rack with a door (which, depending on the rack, may be optional). The rack contains the primary I/O Drawer. Up to four drawers are supported per system. The second I/O Drawer can be housed in the same rack as the first or in an additional rack. Up to four I/O Drawers can be installed.

6.1 Book packaging

The components in the CEC are packaged between metal sheets to form what are called books. Book packaging helps protect components from electrostatic discharge and physical damage and also helps to stabilize electronics and distribute air-flow throughout the CEC for proper temperature control. Both memory cards and processors cards used in the pSeries 680 and S80 use book packaging.

Frame guide rails, as shown in Figure 20, help align the books when connecting to the backplane; guidance pins assure final positioning and two book locks secure the book to the backplane and frame cage. This results in a reduction in pin damage when installing processor and memory upgrades, and stabilizes the system for shipment and physical installation.

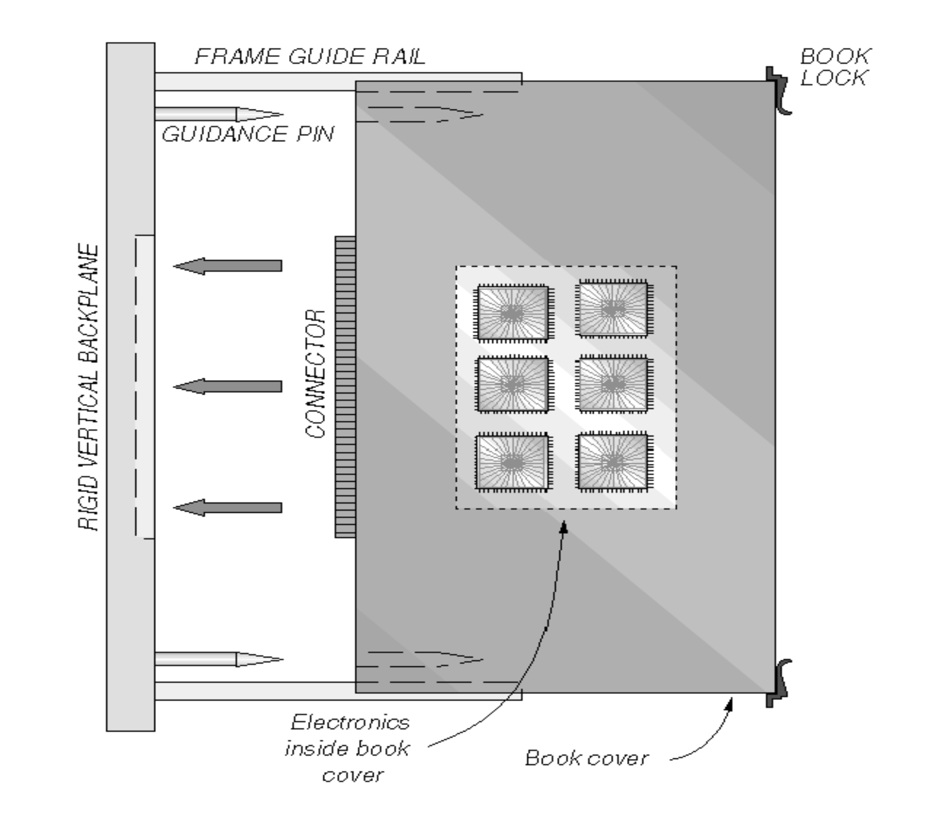
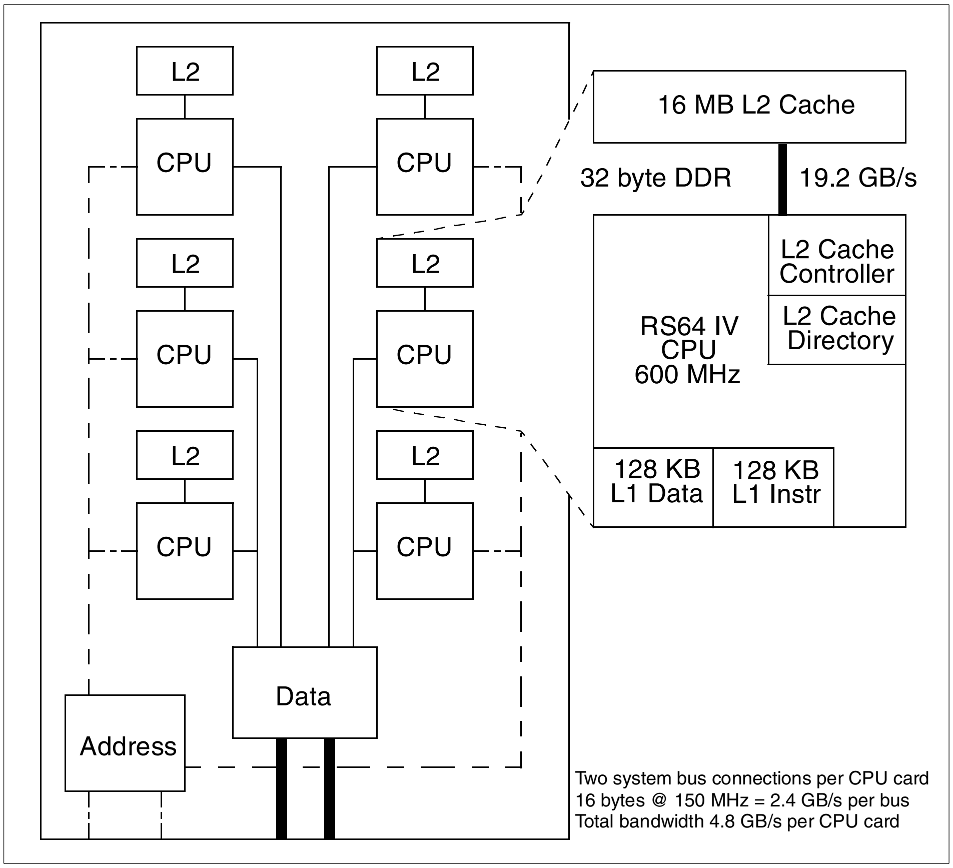


Figure 20. Book Packaging

6.2 RS64 IV processor and card

The 600 MHz RS64 IV processor is based on the 450 MHz RS64 III announced for the Model S80 on September 13, 1999. Both are 64-bit, PowerPC-compatible, four-way superscalar implementations optimized for commercial workloads. The RS64 IV processors in particular are designed for applications that place heavy demands on system memory. The RS64 IV architecture addresses both the need for very large working sets and low latency. Latency is measured by the number of CPU cycles that elapse before requested data or instructions can be utilized by the processor.



Two system bus connections per CPU card
16 bytes @ 150 MHz = 2.4 GB/s per bus
Total bandwidth 4.8 GB/s per CPU card

Figure 21. RS64 IV Processor Card

The new processors combine the latest IBM silicon-on-insulator (SOI) and advanced copper chip technology for improved performance and reliability.

Each processor has an on-chip L2 cache controller and an on-chip directory of L2 cache contents. The cache is four-way set associative, which means that directory information for all four sets is accessed in parallel. Greater associativity results in more cache hits and lower latency, which improves commercial performance. L2 cache performance on the RS64 IV processor has been significantly improved by doubling the L2 size and increasing the speed of the interface technique called Double Data Rate (DDR) that was pioneered for RS/6000 on the RS64 III.

The 16 MB L2 cache is formed of 8 x 16 Mb Static Random Access Memory (SRAM) modules. The new 16 Mb SRAM used for L2 is capable of transferring data twice during each clock cycle. The L2 interface is 32 bytes wide and runs at 300 MHz (half processor speed), but because of the use of DDR, it provides 19.2 GB/s of throughput.

Figure 21 on page 71 contains a diagram of the processor card layout used on the pSeries 680. There are six RS 64 IV processors per card, which share two system buses that connect to the memory controller complex.

The RS64 IV processor has five pipeline execution units: branch, load/store, fixed point, complex fixed point and floating point. The complex fixed point unit provides support for multiplication and division instructions. There is a dispatch buffer that can hold up to 16 current instructions, a technique that helps reduce latency. It also has an eight-deep branch buffer. The processor can sustain a decode and execution rate of up to four instructions per cycle.

Fault detection and correction techniques are used on all processor arrays, where failure would cause more than minor performance degradation. Redundancy, error checking and correction (ECC), parity and retry are all used on the following:

- L1 DCache and directory
- L1 ICache and directory
- L2 Cache and directory
- The linebuffer

Together, these features promote high reliability, availability, and data integrity and enable full fault detection and correction coverage within the CEC. In summary, the RS64 IV features include:

- 128 KB on-chip L1 instruction cache
- 128 KB on-chip L1 data cache with one cycle load-to-use latency

- On-chip L2 cache directory that supports up to 16 MB of off-chip L2 SRAM memory
- 19.2 GBps L2 cache bandwidth
- 32 byte on-chip data buses
- 600 MHz operating frequency
- 4-way superscalar design
- Five stage deep pipeline

6.3 RS64 III processor and card

The 450 MHz RS64 III processor used in the Model S80 is based on the 262 MHz RS64 II announced for the Model S70 Advanced on October 23, 1998.

Both are 64-bit, PowerPC-compatible, four-way superscalar implementations optimized for commercial workloads.

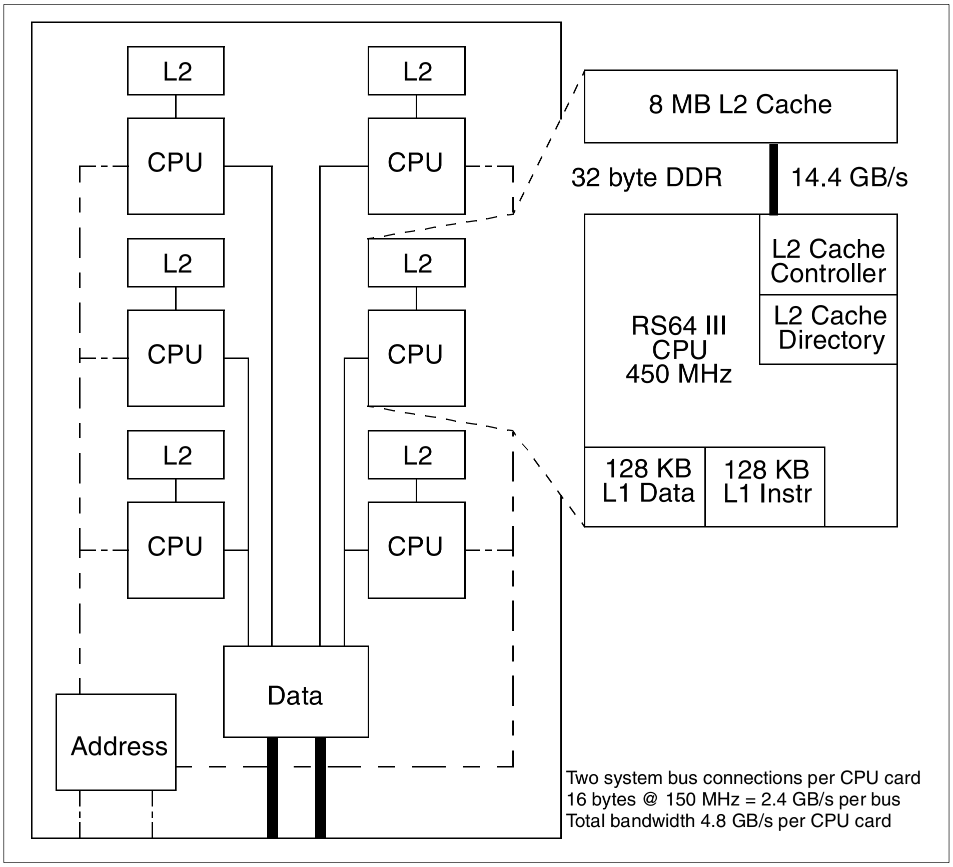


Figure 22. RS64 III Processor card

The RS64 III is designed for applications that place heavy demands on system memory. The architecture addresses both the need for very large working sets and low latency. Latency is measured by the number of CPU cycles that elapse before requested data or instructions can be utilized by the processor.

The RS64 III combines IBM advanced copper chip technology with a redesign of critical timing paths on the chip to achieve greater throughput. The L1 instruction and data caches are 128 KB each - double that of the RS64 II. New circuit design techniques were used to maintain the one cycle load-to-use latency for the L1 data cache.

Each processor has an on-chip L2 cache controller and an on-chip directory of L2 cache contents. The cache is four-way set associative. This means that directory information for all four sets is accessed in parallel. Greater associativity results in more cache hits and lower latency, which improves commercial performance.

Using a technique called Double Data Rate (DDR), the 8 MB Static Random Access Memory (SRAM) used for L2 is capable of transferring data twice during each clock cycle. The L2 interface is 32 bytes wide and runs at 225 MHz (half processor speed), but because of the use of DDR, it provides 14.4 GB/s of throughput.

Figure 22 on page 74 contains a diagram of the processor card layout used on the Model S80. There are six RS64 III processors per CPU card sharing two system buses that connect to the memory controller complex.

The RS64 III processor has five pipeline execution units: branch, load/store, fixed point, complex fixed point, and floating point. The complex fixed point unit provides support for multiplication and division instructions. There is a dispatch buffer that can hold up to 16 current instructions, a technique that helps reduce latency. It also has an eight-deep branch buffer. The processor can sustain a decode and execution rate of up to four instructions per cycle.

Fault detection and correction techniques are used on all processor arrays where failure would cause more than minor performance degradation. Redundancy, error checking and correction (ECC), parity and retry are all used on the following:

- L1 DCache and directory
- L1 ICache and directory
- L2 Cache and directory
- The linebuffer

Together, these features promote high reliability, availability, and data integrity and enable full fault detection and correction coverage within the CEC.

In summary, the RS64 III features include:

- 128 KB on-chip L1 instruction cache
- 128 KB on-chip L1 data cache with one cycle load-to-use latency
- On-chip L2 cache directory that supports up to 8 MB of off-chip L2 SRAM memory
- 14.4 GBps L2 cache bandwidth

- 32 byte on-chip data buses
- 450 MHz operating frequency
- 4-way superscalar design
- Five stage deep pipeline

6.4 Memory controller complex

Ten system data buses link the RS64 processor cards to the memory controller complex and the RIO hub. Each processor card has two buses that connect to two interconnected data flow switches. Each switch connects directly to two quads of Synchronous Dynamic Random Access Memory (SDRAM). Each switch also connects to its own dedicated port on the RIO hub.

Each switch consists of four data flow switch chips and a separate data flow control chip. The data flow switch chips are at the core of the memory controller. An SMP bus arbiter chip in the controller complex prevents switch contention.

Figure 23 on page 77 shows a diagram of the memory controller complex layout. The eight system buses connecting the processor cards and the two system buses connecting the RIO ports are 128 bits wide and run at 150 MHz to provide a total bandwidth of 24 GB/s. Memory ports are 512 bits wide and operate at 75 MHz. The four memory ports have an aggregate bandwidth of 19.2 GB/s. The total memory controller complex switch bandwidth is an impressive 43.2 GB/s. Transfer buffers are used in the switch to queue traffic if the needed connections are not immediately available.

Memory cards are organized into groups of four, called quads. Each quad is accessed through a separate port. Each half of the system switch complex is connected to two quads. Refer to Figure 23 on page 77 for a graphical representation. Memory quads A and D are on one half of the switch, and quads B and C are on the other half. To ensure that both halves of the memory controller complex are used by the system, and to utilize the available switch and system bus bandwidth, it is important that systems be configured with a minimum of two memory quads. See Section 6.5, “Memory cards and quads” on page 78 for more details of memory cards and quads.

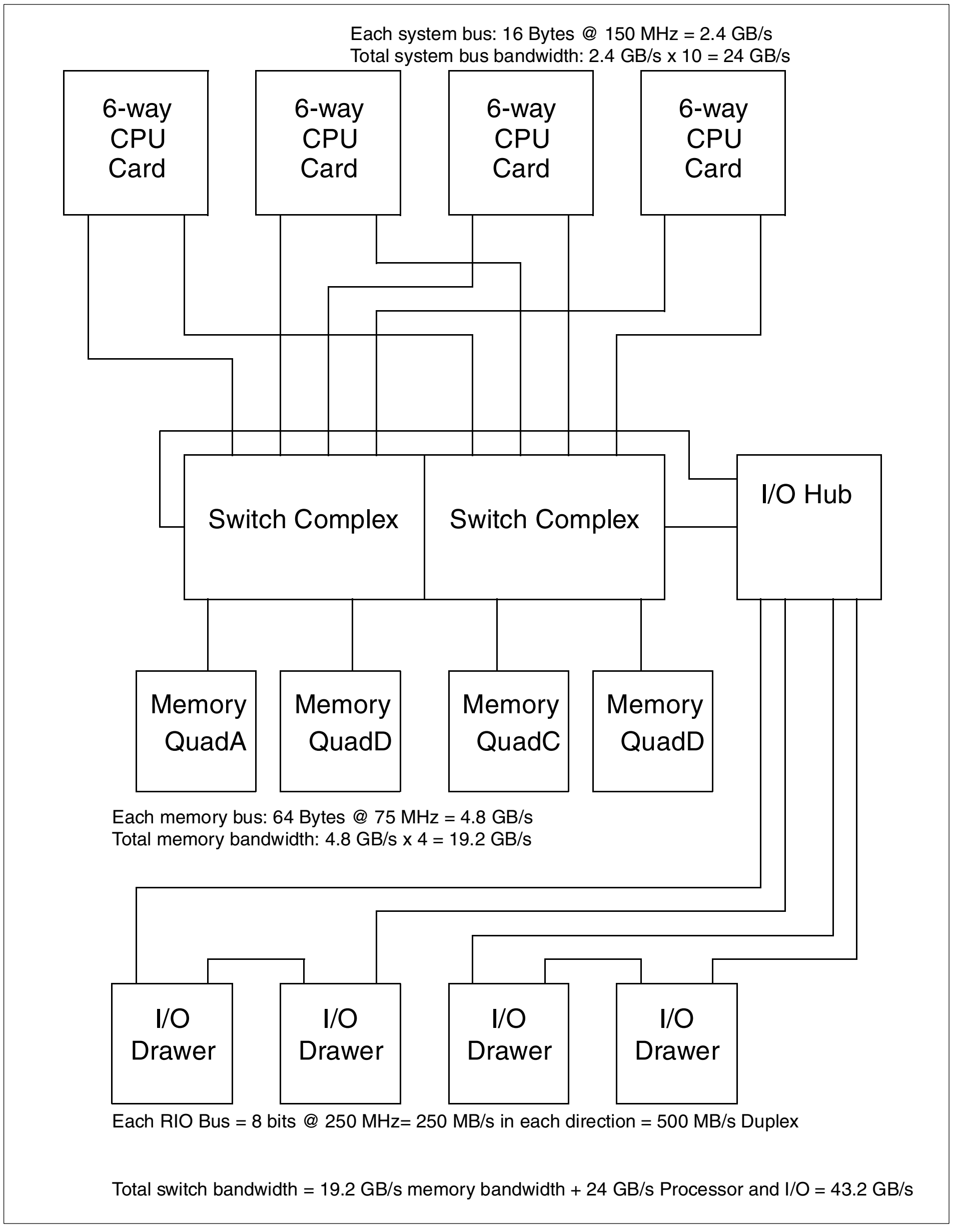

Each system bus: 16 Bytes @ 150 MHz = 2.4 GB/s
Total system bus bandwidth: 2.4 GB/s x 10 = 24 GB/s
6-way CPU Card
6-way CPU Card
6-way CPU Card
6-way CPU Card
Switch Complex
Switch Complex
I/O Hub
Memory QuadA
Memory QuadD
Memory QuadC
Memory QuadD
Each memory bus: 64 Bytes @ 75 MHz = 4.8 GB/s
Total memory bandwidth: 4.8 GB/s x 4 = 19.2 GB/s
I/O Drawer
I/O Drawer
I/O Drawer
I/O Drawer
Each RIO Bus = 8 bits @ 250 MHz= 250 MB/s in each direction = 500 MB/s Duplex
Total switch bandwidth = 19.2 GB/s memory bandwidth + 24 GB/s Processor and I/O = 43.2 GB/s

Figure 23. System switch complex

In addition to the data flow switches, the design incorporates high-function address and data buffers to minimize latency. Addressing is done using a separate 64 bit path. Cross-port traffic between processor cards is queued at the switch as needed.

The high function address and data buffer chips break logical buses into smaller physical units. This allows the clocking frequency and the transfer speed to be increased. The buffers allow each of the system buses to support up to 2.4 GB/s of throughput.

The memory controller complex is mounted on a two-sided active backplane. The processors and memory are inserted as books. The I/O subsystem is connected to the complex using a pair of new RIO hub chips. These chips are on a replaceable I/O interface card that will make upgrades easier in the future. A background task of the memory controller corrects soft, single-bit errors in order to prevent multiple-bit errors from developing; this is called memory scrubbing, and is described in more detail in Section 8.3.2, “Error Recovery for Caches and Memory” on page 103.

One of the features of this memory complex design is its ability to handle different generations of processors. The system is designed to accommodate upgrades without losing its overall excellent balanced performance.

6.5 Memory cards and quads

The base Model S80 comes with 2 GB of SDRAM. The base for the pSeries 680 is 4 GB. The maximum memory on both models is 96 GB.

Both models use a form of memory card designated R1, which uses soldered memory modules in order to minimize failures and faults caused by connectors or sockets. The memory cards for the pSeries 680 and S80 are packaged in books similar to those used for processor packaging, as described in Section 6.1, “Book packaging” on page 70. The memory cards are plugged into reserved slots in the CEC. Some slots are accessible from the front of the CEC and some from the rear.

Memory cards are installed in groups of four called quads. The memory quads are referred to as A, B, C, and D. Memory cards M01 through M03 and M08 through M10 are accessible from the front of the system, and M11 through M15 and M20 through M24 are accessible from the rear. All cards in a quad must be the same size.

Three memory cards from quad A and three cards from quad B are installed on the front of the system. The remaining single card from each quad is installed in the rear of the system. All four cards of each of the C and D quads are installed in the rear of the system.

Figure 24 is the view looking down from the top of the CEC and shows how memory cards are placed for each quad. Refer to *Enterprise Server S80 / pSeries Model S85 Service Guide*, SA38-0558 for more information.

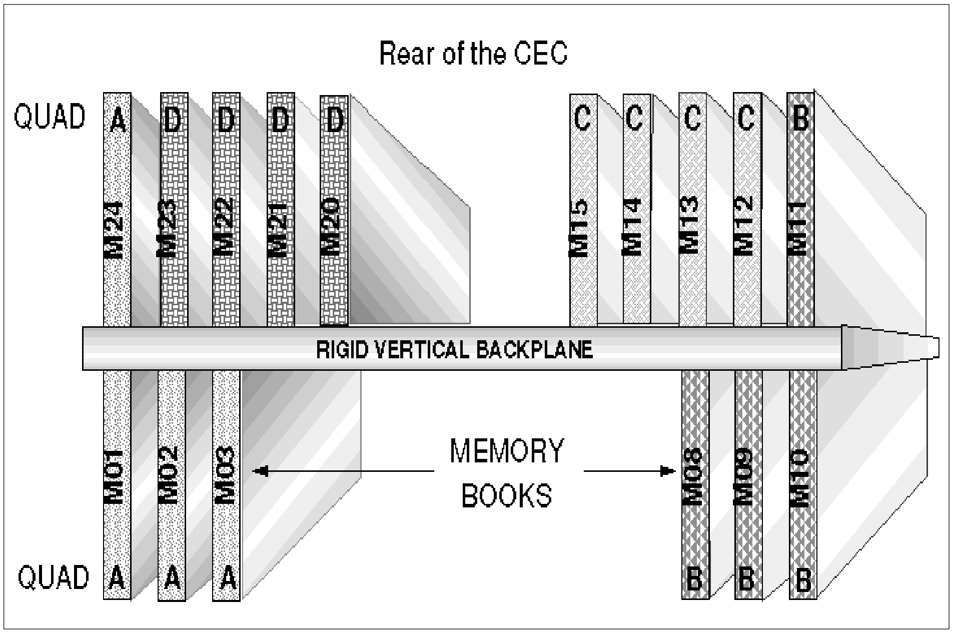


Figure 24. Memory card placement

To get best the performance from either server, sort all memory features by quad size from largest to smallest. Install the largest memory feature in quad A, insert the next largest memory feature in quad B, and so on. Continue inserting memory quads in the available slots alphabetically, with the smaller features following the larger. To ensure that the available switch and system bus bandwidth is fully utilized, it is important that systems be configured with a minimum of two memory quads.

The physical layout of the front of the CEC is shown in Figure 18 on page 67, with the rear of the CEC shown in Figure 19 on page 69.

The memory subsystem has single-bit error correction and double-bit error detection (ECC). The Memory cards feature redundant modules that will support up to one in 14 memory modules not working. The memory is scrubbed by the memory controller when it is idle, a feature that is designed to detect soft errors and reduce hard failures. Memory scrubbing is completely transparent to the operating system and requires no CPU cycles to perform. See Section 8.3.2, “Error Recovery for Caches and Memory” on page 103 for more details.

6.6 Hardware Multithreading

The pSeries 680 supports a new feature of the RS64 processor family - Hardware Multithreading (HMT). The basic technique of HMT is that a processor holds the state of two threads. For example, when a cache miss occurs (L1 or L2), which would normally delay the processor for many cycles while the data is fetched from main memory, the processor switches to another state and executes instructions from that thread. This can help eliminate memory access delays, keeping the CPU more fully utilized, and improving processor throughput.

This new AIX feature provides a performance tuning technique which benefits some workloads, particularly workloads with high multiprogramming levels, such as OLTP. Performance gains of up to 22 percent may be achieved and tend to be highest on systems with eight or less active processors.

However, you need to be very careful before activating HMT. The use of HMT may be very detrimental to system performance when running workloads with low multiprogramming levels (for example, Business Intelligence or batch processing), or when running applications that have scalability issues with a high number of processors. Activating HMT also limits some system functionality (notably dynamic CPU de-allocation) and changes the expected behavior of some AIX commands. Careful planning and testing is therefore recommended prior to using HMT in a production environment.

In order to support this feature you must install new firmware support and APAR IY13075 (available December 19, 2000).

6.6.1 Limitations

When HMT is enabled, the system will appear to have twice as many processors as it actually has. For example, an 8-way SMP will appear to be a 16-way SMP. This gives rise to several limitations:

- When HMT is enabled, unpredictable results (including system crashes) may occur when vendor software has built-in dependencies upon the number of processors in the system.
- Dynamic CPU de-allocation (CPU Gard) is *not* supported when HMT is enabled. If CPU Gard has been configured before HMT is enabled and then HMT is enabled, the CPU Gard subsystem will not be configured on the next system boot and AIX will not de-allocate CPUs. When HMT is

disabled, CPU Gard will be configured on the next system boot (assuming the configuration has not been changed).

- When HMT is enabled, certain performance monitoring tools may provide skewed results. This is due to the hardware threading of a single processor into *two* processors. In particular, some of the sampling based tools are less accurate using HMT with low CPU utilization workloads.
- Measuring scaled throughput when HMT is enabled will not produce useful data.
- The `lsdev -C` and `lscfg` commands still report the actual number of physical processors present. That is, the number of processors reported by these commands will not be doubled when HMT is enabled.
- The `diag` command still tests the actual number of physical processors present. That is, the number of processors tested by diagnostics will not be doubled when HMT is enabled.
- The `bindintcpu` command may now return the error `EXDEV`.
- The `bindintcpu` command may now report the following error message:

  ```
  Unable to assign interrupt level to specified processor.
  ```
- The `i_int2cpu_ppc()` service may now return the error `EXDEV`.
- Capacity Upgrade On Demand (CUoD) constraints will be respected when HMT is enabled.

Kernel developers should take note that there are additional limitations described in the file /usr/lpp/bos/README.HMT, which is supplied as part of APAR IY13075.

IBM has seen mixed results in its performance evaluation of HMT. A report of this evaluation is available at:

`ftp://ftp.software.ibm.com/aix/tools/perftools/perfpmr/misc_documents/HMT_wp.ps`

6.6.2 Controlling Hardware Multithreading

HMT is controlled through the `bosdebug` command, as discussed in the following:

- To enable HMT, issue the following command:

  ```
  bosdebug -H on
  ```

  HMT being enabled will be indicated by the following being displayed:

  ```
  HMT                         on
  ```

  The `bosboot` command must be run and the system rebooted in order for HMT to be enabled.

- To disable HMT, issue the following command:

  ```
  bosdebug -H off
  ```

  The "`HMT     on`" indicator message will no longer be displayed.

  The `bosboot` command must be run and the system rebooted in order for HMT to be disabled.

- To check if HMT is enabled, issue the following command:

  ```
  bosdebug
  ```

  If HMT is enabled the following will be displayed:

  ```
  HMT                         on
  ```

  If HMT is disabled, no HMT message will be displayed.

- To check if HMT is active, issue the following command;

  ```
  bindprocessor -q
  ```

  If HMT is active, double the number physical CPUs will be listed.

When HMT is active, this means that the `bosdebug -H on`, `bosboot`, and a subsequent system reboot have successfully been carried out, and the system is currently running with the HMT processor technique.

6.7 Remote I/O connections

On the pSeries 680 and S80, the Remote I/O (RIO) subsystem connects to the CEC using the I/O hub chips on a new replaceable I/O interface card. Four RIO connections are supported, allowing a maximum of two RIO loops. The RIO connections are scalable, high-speed, point-to-point interfaces designed for low-latency high-bandwidth connections between two boards or boxes. Each RIO bus supports up to 500 MB/s total or 250 MB/s in each direction concurrently. RIO cables connect the CEC to the I/O devices located in the I/O Drawers. The RIO connections are set up as loops. The I/O hub chips direct the traffic around the loop in an optimal way for performance, and they redirect traffic if there are link errors.

The new RIO hub interface chip on the pSeries 680 and S80 offers improved buffering (to enhance the effectiveness of the I/O interface).

These RIO connections are the key to allowing an expandable number of I/O Drawers that are physically separated from the CEC. In turn, this feature also enables the high number of PCI buses and slots.

6.8 I/O Drawer

The pSeries 680 and S80 use an enhanced I/O Drawer. The drawers have fully redundant power supplies and fans that can be serviced without shutting down the system. The drawers improve overall cooling by using more powerful, variable-speed fans that increase air flow when one fails. An LED panel on each drawer displays status information.

The primary I/O Drawer for the pSeries 680 and S80 contains:

- The I/O planar
- The Service Processor
- The native I/O card
- Two independent hot-pluggable disk bays (six-packs)
- One available media bay
- One floppy disk drive
- One SCSI-2 CD-ROM
- 14 PCI slots (11 available)
- One keyboard port
- One mouse port
- Two serial ports and one parallel port

In the primary I/O Drawer, one PCI slot will be used for the Service Processor, and two will be used by the controllers for the media bay and first disk six-pack. If the second six-pack is used, a third controller is required. However, one PCI Dual Channel Ultra2 SCSI adapter (#6205) can be used to replace two PCI Single-Ended Ultra SCSI adapters (#6206), thus freeing one adapter slot. The dual channel controllers will run at Ultra speed. The 14 PCI

I/O slots consist of five 64-bit and nine 32-bit PCI slots. The 64-bit PCI slots are 1, 5, 9, 10, and 14.

The original seven EIA S70 I/O Drawers are not supported on the pSeries 680 and S80.

On the pSeries 680 and S80 I/O Drawers, the majority of the function is implemented on the I/O planar. In each drawer, a single RIO-to-I/O bridge bus chip converts the RIO bus to the local mezzanine bus. The mezzanine bus is also called the I/O bridge bus. The I/O bridge bus drives four PCI bridge chips, each with its own local PCI bus, containing two or four slots each. Each bus works independently. The RIO-to-I/O bridge chips have *IN* and *OUT* RIO ports to enable redundant and chainable loop connection of RIO devices. The I/O bridge bus runs at 66 MHz and has a 528 MB/s bandwidth.

6.9 I/O bridge bus

The local I/O bridge bus is a reduced signal version of the system bus that has been optimized for I/O. The I/O bridge bus uses a multiplexed 64-bit address and data path. The I/O bridge bus is parity checked for address, data, and control errors. Each bus request is range checked and positively acknowledged for improved error detection. The I/O bridge bus operates in pipeline mode. New requests can be issued before previous requests are completed. The bridges and other chips in the I/O path provide significant queuing.

6.10 PCI buses

The PCI bridge chips on both systems convert the I/O bridge bus to PCI. There are 14 PCI-compliant slots running at 33 MHz per I/O Drawer. PCI 2.1 cards are supported. Four PCI bridges are used per I/O planar. One of the PCI bridge chips drives two 64-bit PCI slots. The other three PCI bridges each drive one 64-bit PCI slot and three 32-bit PCI slots. This configuration performance balances the load. Five volts and 3.3 volts are available at the slots. Five volt PCI signaling conventions are used.

The 64-bit PCI slots have a maximum throughput of 266 MB/s. The 32-bit PCI slots have a maximum throughput of 133 MB/s. It is important to note that no PCI-to-PCI bridges are used in this performance-optimized design. PCI-to-PCI bridges significantly limit the useful bandwidth of the related PCI slots.

Chapter 7. AIX features for large SMP servers

AIX is an integrated UNIX 98 branded operating environment for POWER- and PowerPC-based workstations, symmetric multiprocessor, and scalable parallel computing systems. This environment enables both the development and execution of computing applications across the RS/6000 and pSeries product lines. AIX Version 4.3.3 is a third modification release building on the industry-leading strength and stability of AIX Version 4.3.

This chapter describes some of the new features of AIX Version 4.3.3, which are particularly useful in large commercial environments. Although many of these features can be used on any RS/6000 or pSeries server, they are designed to be of most benefit on large SMP systems, such as the pSeries 680 and S80.

AIX Version 4.3.3, with enhanced 64-bit scalability and functionality, provides:

- Significant AIX software scalability enhancements for 24-way SMP systems, described in Section 7.1, “SMP scalability enhancements” on page 86.
- An AIX Workload Management system with a policy-based method for managing system workload and system resources. See Section 7.3, “Workload management” on page 91 for more details.
- Improved system availability with support for online Journaled File System (JFS) backup and concurrent mirroring and striping. See Section 7.2.1, “Online JFS backup” on page 87 and Section 7.2.2, “Concurrent striping and mirroring” on page 89 for more details.
- Capacity Upgrade on Demand, which allows inactive processors to be installed on your system and to be activated quickly and easily as your business needs require. See Section 7.4, “Capacity Upgrade on Demand” on page 97 for more details.

Other new features in AIX 4.3.3 include:

- A port of the Sun Solaris 2.5 NIS+ network information management system (in addition to the current NIS support).
- IBM AIX Developer Kit, Java Technology Edition Version 1.1.8.
- Enhanced Ease-of-Use capabilities, including additional Web-based System Manager Task Guides and SMIT support.
- X11R6.3, the *Broadway release*, OpenGL enhancements, and graPHIGS enhancements.

- AIX exploitation of SecureWay Directory for AIX users and groups.
- Increased scalability, performance, capacity, and capability of e-business with Web serving acceleration, Cisco EtherChannel, and Quality of Service (QoS) administration and support.
- TCP/IP enhancements, such as SOCKS V5, gratuitous ARP, and a Sendmail upgrade to Version 8.9.3 with anti-spamming features.
- Enhanced RAS and improved serviceability features for problem determination

For full details on the enhancements included in AIX Version 4.3.3, refer to *AIX Version 4.3 Differences Guide*, SG24-2014.

7.1 SMP scalability enhancements

Some AIX improvements will be particularly beneficial for large SMP systems. This section gives a brief description of some of these changes.

7.1.1 Increased threads and process limits

In previous AIX versions, the number of system threads and processes were each limited to 131072. In some large SMP systems, the number of threads and processes could reach the limit. In AIX Version 4.3.3, these limits have been extended to 524288 for threads and 174080 for processes.

7.1.2 Multiple run queues with load balancing

On SMP systems, AIX Version 4.3.3 implements multiple run queues. Each processor now has a local run queue in addition to the original system-wide global run queue.

When new threads are created, they are added to the system-wide global run queue. When a processor is looking for a thread to dispatch, AIX examines the global run queue, in addition to its local run queue, and chooses the highest priority thread. Once a thread is on a processors local run queue, it will tend to stay on that run queue and, thus, always be run on the same processor. This increases the probability that the data and instructions required to run the thread will still be in the cache of the processor, reducing the need to fetch instructions and data from main memory.

Periodic load balancing takes place automatically to ensure that all processors have similar numbers of threads in their local run queues and to prevent threads from bottle necking on a single processor.

The multiple run queues improve system performance by dramatically reducing lock contention for the global run queue and increasing processor affinity for individual threads. In turn, this reduces the overhead on the cache and memory subsystem, allowing greater system throughput.

7.1.3 Reduced lock contention

In addition to the changes to the run queue system, many other changes have been made to the process management system. Most of the changes are aimed at minimizing lock contention in the kernel, which is of much more importance now that 24-way SMP systems are supported. Some of the changes include analyzing and improving the code being run while particular kernel locks are held. Other situations called for implementing new more granular locks to reduce contention. One area to benefit from these changes is the process table lock.

7.1.4 Fast device configuration

The AIX configuration manager now allows multiple device configuration methods to run in parallel during system boot. This will produce a faster reboot when multiple devices of the same type, such as multiple SCSI disks, TTYs, and multiport asynchronous adapters, are connected to an AIX system. A serialization mechanism is used when the configuration manager recognizes that new devices have been added to the system. This causes the new devices to be sequentially configured.

Systems with a large number of asynchronous I/O adapters benefit from a reduction in both machine boot time and duration of the `cfgmgr` command. Generally, each asynchronous I/O adapter can take up to five minutes to be configured. With this new feature, up to 16 device configuration methods can run in parallel.

7.2 File system enhancements

This section describes some of the new AIX features that will be of great benefit in the large commercial server environments commonly found on pSeries 680 and S80 servers.

7.2.1 Online JFS backup

The Journaled File System (JFS) has been enhanced to support file system online backup. This capability allows a mirrored copy of a file system to be used for backup purposes. A mirror copy of the file system is split off, mounted read-only, and available for backup. The primary copy of the file system is still mounted and in use. This enables the system administrator to back up a consistent copy of the file system data while another copy is still mounted and in use by users and applications. After the backup is complete, the administrator can reintegrate the backup mirror copy and re-synchronize it with the other mirror copies.

There are some conditions that have to be met in order to perform an online backup:

- The JFS logical volume and JFS log logical volume must be mirrored.
- The number of mirrored copies of both the JFS logical volume and the JFS log logical volume must be the same.

Note

Splitting a mirrored copy of a file system means that one copy is temporarily dedicated to backup activities and is not available to provide data availability. It is recommended that you have a triple mirror of the file system so you can recover from any disk problem that happens before the backup copy has been reintegrated with the file system.

It is recommended to keep file system activity as minimal as possible during online JFS backup events.

After the backup activities are finished, it is possible to remove the read-only file system in order to have maximum data availability.

The process of splitting off and mounting the backup file system is accomplished by use of the `chfs` command. The correct set of options to this command will split off the required copy of the mirrored logical volume and mount the copy in read only mode at a specified mount point. For example, assuming that /data is a file system on a triple mirrored logical volume, the following command splits off the third copy of the mirror and mounts it read only as /backup:

```
# chfs -a splitcopy=/backup -a copy=3 /data
```

When the `chfs` command completes, the snapshot copy of the file system is mounted and ready for the backup procedure to commence.

Once the backup procedure has completed, the snapshot copy of the file system needs to be reintegrated with the live data. This is performed in two steps:

1. Unmount the snapshot copy of the file system. For example:

```
# umount /backup
```

2. Use the `rmfs` command to remove access to the snapshot copy. For example:

```
# rmfs /backup
```

At this point, the snapshot data is not actually deleted. The `rmfs` command removes the temporary logical volume device, which was created to access the snapshot copy. It then starts the syncvg process in the background to re-synchronize the snapshot copy with the live data. Only the data blocks that have changed since the snapshot was taken are actually synchronized. The duration of this operation depends on a number of factors, including the size of the file system, the number of blocks that have changed, and the speed of the disk subsystem on which the file systems are stored.

7.2.2 Concurrent striping and mirroring

For some time, the Logical Volume Manager (LVM) in AIX has provided the ability to create striped or mirrored logical volumes, although a logical volume could not be both striped and mirrored. In this AIX release, the LVM combines RAID 1 (mirror) data availability with RAID 0 (striped) performance by supporting (entirely in software) a striped logical volume with mirrors. This feature further enhances data availability in high-performance striped logical volumes by tolerating disk failures. The remaining disks in the striped mirror copy continue to service striped units contained on these disks. The replacement of a disk where only the partitions on the new disk are synchronized is provided through the `migratepv` and `replacepv` commands.

In addition, all logical volumes can now utilize a new partition allocation policy called *Super Strict*. This Super Strict policy does not allow partitions from one mirror to share a disk with any partitions from a second or third mirror, thus helping to further reduce the probability of data loss resulting from a disk failure.

These new functions are not backward compatible; therefore, new volume groups supporting these features cannot be imported and used with previous versions of AIX.

7.2.2.1 RAID Levels 0 and 1

RAID 0 is also known as data striping. Conventionally, a file is written out sequentially to a single disk. With striping, the information is split into chunks and the chunks written to (or read from) a series of disks in parallel. There are two main performance advantages to this:

- Data transfer rates are higher for sequential operations, due to the overlapping of multiple I/O streams.
- Random access throughput is higher, because access pattern skew is eliminated due to the distribution of the data. This means that with data distributed evenly across a number of disks, random accesses will most likely find the required information spread across multiple disks and benefit from the increased throughput of more than one drive.

Although RAID configurations are normally associated with availability, keep in mind that RAID 0 is only designed to increase performance.

RAID 1 is also known as disk mirroring. In this implementation, duplicate copies of each chunk of data are kept on separate disks. If any disk in the array fails, the mirrored twin can take over. Read performance can be enhanced as the disk with its actuator closest to the required data is always used, thereby, minimizing seek times.

The response time for writes can be somewhat slower than for a single disk depending on the write policy; the writes can either be executed in parallel for speed or serially for safety. Writing in parallel means that the write process will complete in the time taken for the slowest drive to finish in the mirrored pair; this is quick, but it means that a failure in writing to one of the pair of disks is not immediately detectable. In contrast, writing sequentially means that the write to the mirrored copy is not initiated until the first write has successfully completed. This is slower, but any errors are immediately detectable. Mirroring improves response time for read-mostly applications and improves availability at the expense of cost, since twice (or three times) as many disks are required as disk space.

RAID 1 is best-suited to applications that require high data availability and good read response times, and in settings where cost is secondary.

Table 13 compares the relative characteristics of different RAID levels.

Table 13. Characteristics of different RAID levels

RAID Level	Availability Mechanism	Capacity	Performance	Cost
0	none	100%	high	medium
1	mirroring	50%	medium	high
0+1	mirroring	50%	high	high
5	parity	80%	medium	medium

It is important to keep in mind that each different level of RAID configuration has advantages and disadvantages. A combination of striping and mirroring provides high performance and high availability, but there is a high cost.

Note

RAID 0 mainly improves sequential workloads. For heavy random workloads and high availability needs, you should consider using RAID 1 only.

7.3 Workload management

This section describes the AIX Workload Manager (WLM) system, which provides a policy-based method of managing system workload and resources. For more information refer to the IBM Redbook *AIX 5L Workload Manager (WLM)*, SG24-5977.

7.3.1 Overview

Workload management is designed to give the system administrator more control over how the scheduler and virtual memory manager (VMM) allocate resources to processes. This can be used to prevent different classes of jobs from interfering with each other and to allocate resources based on the requirements of different groups of users.

The major use of workload management is expected to be for large systems with many CPUs and large amounts of memory, such as the pSeries 680 and S80. These large systems are often used for *server consolidation*, where workloads from many different server systems, such as print, database, general user, and transaction processing systems, are combined into a single large system to reduce the cost of system maintenance and administration. These workloads often interfere with each other and have different goals and service level agreements. Workload management is designed to address these problems. The same issues can occur in a single environment where the user base has very different system usage characteristics or the system managers have different priority user communities.

Another use of workload management is to provide isolation between user communities with very different system behaviors. This can prevent effective starvation of workloads with certain behaviors, such as interactive or low CPU usage jobs, from workloads with other behaviors, such as batch or high memory usage jobs.

These usages have two different issues that workload management must address. First, targets for the amount of resources available to different workloads are required. These targets are not absolute; rather, they should be achieved over the long term to provide a degree of fairness. The second issue that workload management must address is boundaries on the amount of resources that a workload can receive. These boundaries can be in terms of maximum resources available and minimum resources that must be made available. These boundaries are not intended to be the major means of separating workloads; rather, they are intended to address special situations where targets are not enough to provide sufficient isolation.

7.3.2 Managing resources with WLM

WLM monitors and regulates the CPU utilization and physical memory consumption of the threads and processes that are active on the system. Using a set of class assignment rules provided by the system administrator, new processes are automatically assigned to a class by WLM upon execution. Classes are grouped into tiers, which indicate the relative importance of each class. Each class has a set of minimum and maximum limits for each resource managed by WLM. In addition, each class has a target value for each resource. This target is representative of the amount of the resource that would be optimal for the jobs in the class.

7.3.2.1 Classes

A class defines common resource requirements for a group of processes. It is possible to create up to 29 classes with up to 16 characters for the class name.

The system administrator defines a set of rules that allow the system to determine which class a process belongs to. Class assignment rules are based on the user ID, group ID, and executable name of the process.

When WLM is enabled, the assignment of jobs to classes according to the class rules is automatic. WLM was designed as a *Set & Forget* Administration.

There are two default classes that WLM creates at initial environment setup:

System Class This class defines the resources for the operating system. It includes all kernel processes and all privileged processes that are not automatically assigned to another class.

Default Class This default class is used as a catch-all at the end of the class rules file. All processes that are not assigned to any other class will be assigned to the default class.

7.3.2.2 Tiers

Tiers are the values that indicate the relative importance of classes to WLM. Each class belongs to a tier, with the importance defined by the tier value, which ranges from 0 to 9 (0 being the most important tier and 9 being the least important).

Tier numbers allow the grouping of classes by equivalent importance. Classes in less important tiers get resources that classes in more important tiers do not require.

7.3.2.3 Shares

The usage target for different types of resources is specified with shares. The shares are specified as relative amounts of usage between the different classes.

If a class does not have any processes in it, the available resource is divided among the remaining classes according to their share values.

Possible share values range from 1 to 65,000. The default target share, if unspecified, is 1.

Figure 25 shows what happens when a class with 5 shares has no active jobs. The resources that would have been used by the class are made available to other classes.

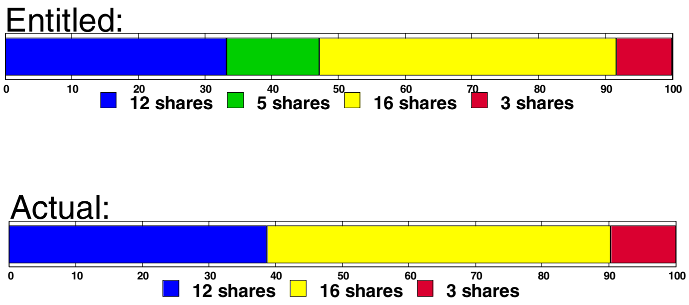


Figure 25. Example share distribution

Each of the active classes still has the same share value, but the real percentage of resource allocated is larger, since there is a smaller total number of shares.

7.3.2.4 Minimum and maximum resource limits

The different resources controlled by WLM can be limited by the following values:

- The minimum percentage of the resource that must be made available when requested. The possible values are integers from 0 to 100. If unspecified, the default value is 0.
- The maximum percentage of a resource that can be made available, even if there is no contention for the resource. The possible values are integers from 1 to 100. If unspecified, the default value is 100.

Keep in mind that the minimum limit reserves resources and the maximum limit restricts resources. The resource target for a class is determined by the number of shares for the class and the number of shares for other active classes. The resource target may also be increased or reduced, because limits take precedence over shares.

Resource limit values are specified by resource type in the resource limit file within stanzas for each class. The limits are specified as a minimum to maximum range separated by a hyphen (-) with whitespace ignored. Each limit value is followed by a percent sign (%).

WLM does not place hard constraints on the values of the resource limits. The following are the only constraints:

- The minimum range must be less than or equal to the maximum range.
- The sum of the minimum of all classes within a tier cannot exceed 100.

WLM enforces the maximum range to ensure that a class or process within a class is not given more resource than the specified value. Note that, in the case of a memory constraint, swapping performance can become very poor for processes within the constrained class. Memory minimums for other classes should be used before imposing a memory maximum for any class.

A minimum value constraint on a class means that processes within the class are always allowed to get resources up to the minimum. WLM cannot guarantee that processes actually reach their minimum limit. This depends on how the processes use their resources and on other limits that may be in effect. For example, a class may not be able to reach its minimum CPU entitlement because it cannot get enough memory.

Figure 26 on page 95 shows a diagram of a class with the limits setup.

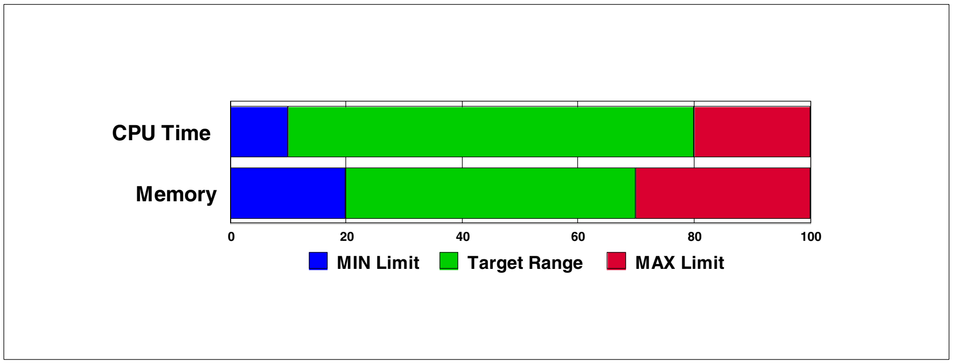


Figure 26. Class with independent resource limits

Each class has independent limits and share values for each resource being controlled by WLM. For example, the class shown in Figure 26 has maximum limits of 80 percent for CPU and 70 percent for memory.

7.3.3 Planning for WLM

WLM offers a fine level of control over resource allocation. However, it is easy to set up conflicting values for the various parameters and obtain undesirable system behaviors. The following tips can help you avoid creating conflicts:

- Know your user base and their basic computing needs when defining classes and class assignment rules.
- Know the resource needs of the main applications.
- Use targets rather than minimum and maximum limits. Targets give the system greater flexibility than hard limits, and targets can help prevent starving applications.
- Try to balance the load using only targets, and monitor the system with the `wlmstat` command. Apply minimum limits for classes that do not receive sufficient share.
- Prioritize some jobs by using tiers.
- Use maximum limits only as a last resort to restrain applications that consume large quantities of resource. Maximum limits can also be used to place hard limits on users' resource consumption (for example, for accounting purposes)

A *passive mode* is provided to help systems administrators understand what the resource requirements of their applications (or application mix) is and thus help them better plan their WLM configuration. In this mode, WLM classifies new and existing processes and gathers statistics about their resource usage, but does not try to regulate this usage. In this mode, the processes compete for resources exactly as they would if WLM was off. The `wlmstat` command can then be used to get snapshots of the resource usage of the classes.

7.3.4 Starting WLM

WLM is an optional service of AIX and can be started manually or automatically. The `wlmcntrl` command allows you to start and stop WLM.

After APAR IY06844, once WLM is started, every new process, when it issues the system call *exec,* will be classified into one of the defined classes based on the class assignment rules pertaining to the user, group and/or application pathname.

In normal system mode, it is best to start WLM early in the system initialization process. For example, WLM could be started by an inittab entry, such as the following:

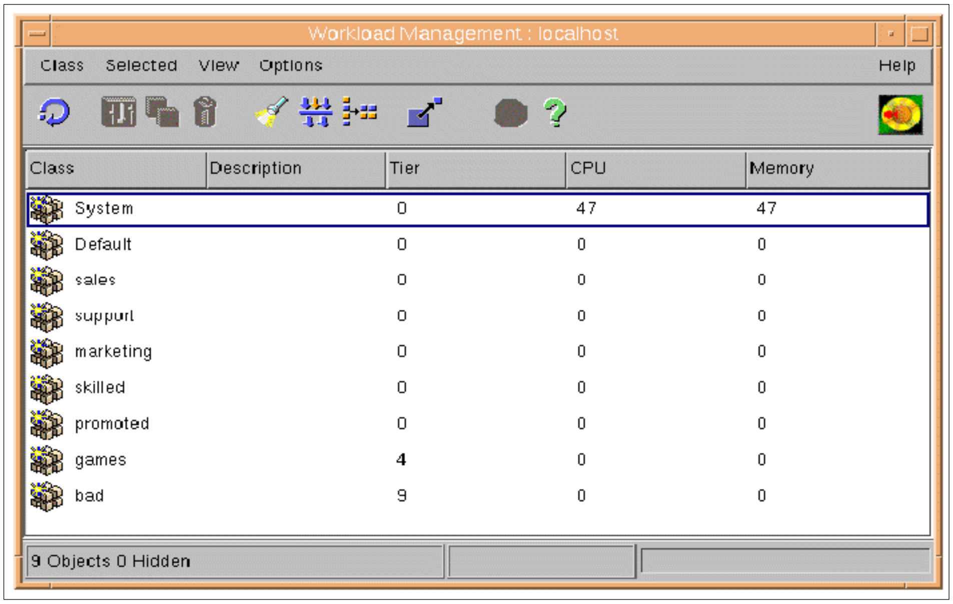
```
wlm:2:once:wlmcntrl -d /etc/wlm/standard > /dev/console 2>&1
```

7.3.5 WLM user interfaces

The administration of WLM can be performed using Web-based System Manager (wsm), SMIT, the command line, or by modifying configuration files directly.

This provides the administrator flexibility to include WLM commands in shell scripts or use a graphical interface to set up or modify the WLM configuration.

Figure 27 on page 97 shows the WLM control application using the wsm user interface.



Class	Description	Tier	CPU	Memory
System		0	47	47
Default		0	0	0
sales		0	0	0
support		0	0	0
marketing		0	0	0
skilled		0	0	0
promoted		0	0	0
games		4	0	0
bad		9	0	0

Figure 27. Main menu for Workload Manager

7.4 Capacity Upgrade on Demand

Capacity Upgrade on Demand (CUoD) is a new feature of AIX that allows you to have inactive processors installed on your system, which can be made active quickly and easily as your business needs require. When more processing capacity is required, you simply issue a new AIX command, `chcod`, to increase the number of active processors, in increments of two, up to the number physically installed in the system. The processors become active at the next system re-boot.

IBM is automatically notified whenever the number of active processors is changed. The customer is responsible for placing an order for the additionally activated processors with their IBM or IBM Business Partner representative.

7.4.1 Pre-requisites

The following pre-requisites are required before activating CUoD processors:

- Electronic Service Agent must be installed and operational.
- The appropriate AIX support is installed on the system; this is available through APAR IY10846 at:

  `http://techsupport.services.ibm.com/rs6k/fixes.html`

- A CUoD processor board is installed and available.

7.4.2 Using the chcod command

The `chcod` command manages CUoD, which allows configuration of more processors on the system than were originally authorized. The additional processors may be enabled if they are available and if the system supports CUoD. The change in the number of processors takes effect after the next system boot. CUoD management also includes displaying the current number of processors which have CUoD support, monitoring the number of processors on the system, and notify the appropriate party.

The syntax of the `chcod` command is as follows:

```
chcod [-r ResourceType -n NbrResources] [-c CustomerInfo] [-h]
```

Flags:

-c *CustomerInfo* Specifies the text string to include in the error log. *CustomerInfo* may not be more than 255 characters. Blank spaces may not be included in the string. Once *CustomerInfo* has been specified, subsequent `chcod` uses do not have to specify the `-c` flag, but you do have the option of changing it. *CustomerInfo* may consist of alphanumerics and any of "." (decimal point), "," (comma), "-" (hyphen), "(" (open parenthesis), or ")" (close parenthesis). This flag is optional.

-h Displays the usage message. This flag is optional.

-n *NbrResources* Specifies the number of *ResourceType* to be authorized on the system. The value of *NbrResources* should be entered in increments of 2. The number that is entered represents the total number of active processors for the system. If you are adding the first additional processor board, the number should be 6, 8, 10, or 12. If it is 0, CUoD will be disabled for the specified *ResourceType*. This flag is optional. If `-n` is specified, then `-r` must be specified as well.

-r *ResourceType* Specifies the *ResourceType* (`proc` for processors) to be enabled and monitored on the system. The system must support Capacity Upgrade on Demand for the specified *ResourceType*. If `-r` is specified, then `-n` must also be specified.

The default invocation (with no flags) displays the current value of *CustomerInfo*, a reserved field named *MailAddr*, the system model name and serial number, and the current value(s) for any *ResourceType*.

When the system performs the command to enable additional processors, information about the new system configuration is added to the error log and sent out through Electronic Service Agent to the service support center. Notification of the number of enabled resources is sent on a monthly basis, in addition to when the number of resources changes.

7.4.2.1 Examples of the chcod command

Here are some examples showing the different formats of the `chcod` command:

1. To activate ten processors, enter:

   ```
   chcod -r proc -n 10 -c"Jane_Doe-Customer_Number_999999-(111)111-1111"
   ```

2. To change the CustomerInfo, enter:

   ```
   chcod -c"Jane_Doe-Customer_Number_999999-(222)222-2222"
   ```

3. To see the current values of the resources with capacity upgrade on demand support, enter:

   ```
   chcod
   ```

   A message similar to the following will be displayed:

   ```
   Current MailAddress = _______Reserved_______
   Current CustInfo = Jane_Doe-Customer_Number_999999-(222)222-2222
   Current Model and System ID = IBM,7017-S80_IBM,000974934C00
   Current number of authorized proc(s) out of 12 installed on system = 10
   ```

7.4.3 Additional information

Unactivated processors cannot be tested by AIX diagnostics, nor do they appear when running the `lscfg` command on the NEW RESOURCE menu, nor on any AIX diagnostic task. However, Systems Management Services and Service Processor menus are not affected by CUoD and all processors are tested at boot time.

Capacity Upgrade on Demand authorizes a number of processors to be activated (not specific processors). If one processor fails at IPL time, the system will still activate the number of authorized processors, marking the remaining good processors as available for future capacity and the failed processors marked as failed.

Inactive processors can also be used to replace system processors which have been de-allocated due to failure; a system re-boot will activate one of the unused Capacity Upgrade on Demand processors to replace the failing processor. This allows the customer to run the system without degradation. However, in such a case, the failing processor card should be replaced at the customer’s convenience to allow the customer to add capacity as needed.

Chapter 8. Reliability, availability, and serviceability

Reliability, availability, and serviceability (described as the collective term RAS) describe how well a system can perform its intended function on demand and how quickly problems and errors can be repaired. Commercial servers running mission-critical applications need strong RAS capabilities. RAS encompasses techniques for reducing faults and minimizing their impacts, shortening repair time, and enabling faster problem resolution.

8.1 Designed for higher RAS

In the pSeries 680 and S80, RAS begins with the development of architectures that give high priority to these characteristics. The emphasis on RAS begins in product development, where designs are tested, evaluated, and optimized. Concern for reliability continues through the manufacturing and distribution processes, where quality is continually evaluated and carefully measured against documented standards. Finally, the focus on RAS extends to service and support, where real-world experience is measured against design criteria. Hardware warranty and maintenance support receive undivided attention within IBM, and significant customer problems are addressed by teams of experts.

The processes IBM follows in designing, testing, manufacturing, and servicing a product are periodically audited and certified for compliance with International Standards Organization (ISO) 9000 guidelines.

During the development of the pSeries 680 and S80, a major effort has been made to analyze single points of failure within the central electronic complex. The processors, L1 and L2 cache, system memory, memory controller complex, and the remote I/O subsystem have been designed to provide mainframe-like levels of reliability. They undergo additional stress testing and screening by suppliers above and beyond what is required of industry-standard components used in many UNIX systems today.

8.2 Features of the pSeries 680 and S80

The pSeries 680 and S80 bring new levels of availability and serviceability to the enterprise server. The pSeries 680 and S80 RAS design enhancements include:

- Automatic error capture and problem isolation
- Dynamic error recovery

- Single-bit error correction, double-bit error detection on internal processor arrays, and L1 and L2 cache and system memory
- Bit steering and scrubbing on memory
- Redundant power supplies and cooling fans, providing fault tolerance and concurrent maintenance for those subsystems
- Predictive failure analysis on processors, memory, RIO, and disk
- Processor boot-time de-allocation of resources based on run-time errors
- Highly-reliable, stress-tested components
- Concurrent diagnostics
- Dynamic CPU de-allocation, described in Section 8.7, “Dynamic CPU de-allocation” on page 108

8.2.1 Dynamic error recovery

The design of the pSeries 680 and S80 aids in the recognition of intermittent errors that are either corrected dynamically or reported for further isolation and repair. Parity checking on the system bus, cyclic redundancy checking (CRC) on the remote I/O bus, and the use of error correcting code on memory and processors contribute to outstanding RAS characteristics. Redundant arrays with cyclic redundancy checking are implemented in the power controller card and the operator panel. The Service Processor arrays are redundant and use CRC checking.

During the boot sequence, built-in self test (BIST) and power-on self test (POST) routines check the processors, cache, and associated hardware required for a successful system start. These tests run every time the system is powered on.

Additional testing can be selected at boot time to fully verify the system memory and check the chip interconnect wiring. When a system reboots after a hard failure, it performs extended mode tests to verify that everything is working properly and that nothing was compromised by the failure. This behavior can be overridden by the systems administrator.

8.3 Hot-swap components

Many server components, such as disk drives, power supplies, and cooling fans, are now hot-swappable, so many common repairs can be made without stopping applications and taking the system offline. For example, with a RAID controller and hot-pluggable disks, normal operation can continue, perhaps in a slightly degraded mode, while a failed drive is repaired.

8.3.1 Power and cooling

Within the CEC, the extra power supply provides redundancy in case of failures in either the bulk or regulated power subsystems. Concurrent repair is supported on the bulk supplies while the regulators will require a controlled system shutdown before repairs can be made. Optional uninterruptible power supply (UPS) systems are supported and recommended for mission-critical servers.

The CEC cooling subsystem has what is called N+1 redundancy. That is, there is one blower more than the number required to keep the system running. If one fails, repairs can be made without shutting down the system. The fans can also adjust their speed to partially compensate for a single failure.

Constant power monitoring hardware assists in the detection of early power loss and notifies the operating system to attempt an orderly shutdown. This same power monitoring hardware detects the loss of redundant bulk power supplies, regulators, fans, and blowers and reports them to the operating system for logging in the system error log and for notification for deferred maintenance.

8.3.2 Error Recovery for Caches and Memory

The RS64 processor L1 cache, the L2 cache, and the memory are protected by error correction code (ECC) logic. The ECC provides single-bit error correction and double-bit error detection for the L2 cache and the memory. All recovered error events are reported by an attention interrupt to the Service Processor, where they are monitored for threshold conditions.

The standard memory card has single error-correct and double-error detect ECC circuitry to correct single-bit memory failures. The double-bit detection helps maintain data integrity by detecting and reporting multiple errors beyond what the ECC circuitry can correct. In many cases, the failure of any specific memory module only affects a single bit within an ECC word (bit scattering), thus allowing for error correction and continued operation in the presence of a complete chip failure (chip kill recovery). A background task of the memory controller corrects soft, single-bit errors in order to prevent multiple-bit errors from developing (memory scrubbing).

The memory data integrity and availability features used in the pSeries 680 and S80 include:

Memory scrubbing A built-in hardware function which performs continuous background reads of data from memory, checking for correctable errors. Correctable errors are corrected and rewritten to memory. A threshold counter is maintained that will signal the Service Processor with a special attention when the threshold is exceeded. Memory scrubbing takes place while the memory controller is idle and is completely transparent to software and uses no CPU cycles.

Bit scattering A technique which scatters memory chip bits across four separate ECC words in order to improve recovery from a memory chip failure (chip-kill recovery). In many cases, the failure of any specific memory module only affects a single bit within an ECC word.

Bit steering Allows memory lines from a spare memory chip to be dynamically reassigned to replace a faulty line in another memory chip.

Chip-kill recovery Continued operation in the presence of a complete chip failure. Redundant modules on the R1 card support up to one in 14 memory modules not working.

If all bits are used up on the spare memory chips, and the threshold is reached, the Service Processor (see Section 8.4, “Service processor” on page 104) will be invoked to request deferred maintenance to replace the memory card at a time convenient to the customer.

8.4 Service processor

The Service Processor and Electronic Service Agent (described in Section 8.5, “Electronic Service Agent” on page 106) work independently and together to achieve high levels of RAS on the pSeries 680 and S80 servers. This section provides a high-level overview of the Service Processor; see Appendix B, “A practical guide to the Service Processor” on page 125 for more details.

8.4.1 Introduction

An independent microprocessor, the Service Processor, occupies a 32-bit PCI slot in the primary I/O Drawer in the pSeries 680 and S80. It comes installed in slot 8. The Service Processor runs its own firmware and has access to non-volatile memory and hardware components within the central electronic complex. It can control system behavior under predetermined conditions. The firmware has an asynchronous menu-driven interface that can be accessed either locally or using a modem from a remote support site.

If desired, the local systems administrator can mirror the remote terminal and keyboard, maintaining oversight and eliminating the need to share passwords with service personnel. This is called console mirroring.

Other capabilities of the Service Processor include the ability to remotely power-off/power-on the system, read its own and system POST error logs, and read vital product data (VPD). The Service Processor can change the bootlist, thus enabling an alternate boot source. It can also view the boot sequence history, which lists the progress indicators that appeared in the operator panel LCD panel during the last boot. The Service Processor also enables the systems administrator to establish password protection of the firmware menus, view the boot sequence history, and change operating system surveillance and reboot policies.

8.4.2 Primary functions

During initialization, the Service Processor monitors diagnostic routines that eventually culminate in control of the system being turned over to the operating system. Even then, the Service Processor continues to monitor hardware, environmental conditions, and, if so configured, the operating system. If a fatal hardware or software problem brings the system down, the Service Processor can automatically call support, the customer's own help desk, or a pager number.

During its run-time monitoring of the hardware, if the Service Processor detects a non-fatal error that could develop into one that would bring the system down, it logs the problem in NVRAM and notifies AIX that it has error information for the error log. On the AIX side, the NVRAM will be accessed by the errdemon, and the problem will be entered in the AIX error log. An error logged in this manner will trigger an error log analysis, and, if Electronic Service Agent has been configured, a service call will be placed to IBM, and the customer's help desk will be notified of the situation.

If a fatal error occurs during run-time, the system will halt. Depending on whether a reboot policy has been set, the system will attempt to restart itself. The Service Processor can be configured to call home and report the problem before attempting the restart. Typically, such calls would go to IBM, the customer's help desk, a pager carried by an on-call technician, or some combination of contacts. If the system is restarted, firmware will cause the Service Processor to de-allocate the failing module.

8.4.3 Firmware updates

Updates to Service Processor and system firmware should be performed by trained personnel. The updates are available from the following support page on the Internet:

```
http://www.rs6000.ibm.com/support/micro
```

Code can be downloaded directly to a server and then installed, or it may be downloaded as a file on a separate workstation. The file is then copied to the machine and installed. The firmware can be updated remotely using diagnostic Service Aids or AIX command line options.

By default, a backup copy of the prior version of firmware is maintained on the system so that it can be booted with the older version (if necessary). At the system administrator's discretion, the new firmware can be copied to the alternate storage location so that the two copies are identical, providing redundant firmware capability in case of corruption.

8.5 Electronic Service Agent

Electronic Service Agent (formerly Service Director) ships automatically on all pSeries 680 and S80 systems. There is no charge for its use as long as the system is covered by warranty or an IBM maintenance agreement. Groups of machines can run Electronic Service Agent with one machine serving as a single point of control and the link to IBM service.

The mission of Electronic Service Agent is to monitor hardware and analyze recoverable faults and report them. Fatal errors that cause the system to crash can be reported by the Service Processor. If desired, Electronic Service Agent can be set up to automatically place a service call to IBM when it detects a problem requiring an on-site visit. Alternatively, it can pass an alert to a help desk so that the customer can decide whether to place a service call, or it could do both. Electronic Service Agent can send alerts as electronic mail to a limited number of addresses specified by the systems administrator.

Electronic Service Agent is menu-driven. The panels can be accessed through the Systems Management Interface Tool (SMIT). The program enables more effective management of hardware support, allowing administrators to quickly and easily view the details of problems that have occurred, the status of open calls, and the machine's service history.

Electronic Service Agent requires a modem. It may share a modem used by the Service Processor.

Electronic Service Agent is designed to automatically report hardware problems based on default settings. However, the customer may modify the settings to prevent Electronic Service Agent from placing a call, such as during hardware upgrades and testing or if the failing component is not covered by an IBM service agreement. The program will also automatically notify an administrator by e-mail before expiration of a warranty or service agreement, giving the customer time to sign or renew a maintenance contract.

8.6 AIX RAS

Reliability, availability, and serviceability is more than just hardware. Multi-user systems, running mission critical applications, require an *industrial strength* operating system in order to meet demanding service level agreements. The standard RAS features of AIX have been well documented: dynamic kernel, journaled file system, and a logical volume manager. AIX Version 4.3.3 has been further enhanced with new RAS features to provide unmatched availability in the UNIX marketplace.

8.6.1 AIX Version 4.3.3 enhancements

With enhanced AIX diagnostics available with AIX Version 4.3.3, system administrators can keep track of diagnostic activity more effectively by using the new diagnostic event log. This log can be viewed by using the Display Previous Diagnostic Results task under concurrent diagnostics. Tasks and service aids within diagnostics have been ordered alphabetically for improved ease of use.

A diagnostic exerciser has been added for processors to enhance problem determination. It provides a means of verifying both memory and processor repairs previously detected by error log analysis.

Auto-restart options for the server, when enabled, are designed to reboot the system automatically following:

- Unrecoverable software error
- Software hang
- Fatal hardware error
- Environmentally-induced (AC power) failure

AIX Version 4.3.3 removes the prohibition against mirroring a dump device. AIX will not complain about the mirrored devices, but it will only write to and read from the primary copy. Additionally, the Logical Volume Manager supports online backup of journaled file systems.

The latest release of AIX also supports mirroring of striped logical volumes, enabling RAID 0 + 1 to improve performance and protect against disk failures.

The Kernel Debugger (kdb) is a tool being added to the AIX system to provide a symbolic debugger for the AIX kernel, kernel extensions, and device drivers. There is also a `kdb` command, which is an alternative to the current `crash` command, to allow examination of system crash dumps.

8.6.2 Software service aids

Serviceability is the ability to diagnose and correct or recover from an error when it occurs. The serviceability capabilities and enablers in AIX are referred to as the software service aids. The primary software service aids are error logging, system dump formatting, and tracing.

The software service aids provide a common set of tools for performing problem determination and problem source identification for AIX software. Problem determination is the name IBM gives to those activities performed in determining whether a malfunction is caused by hardware or software. Those activities that take place after problem determination to determine the specific malfunctioning hardware or software component are referred to as problem source identification.

With proper instrumentation in the software, the software service aids enable the customer to determine whether a problem is caused by a defect in hardware or software. If a problem is due to malfunctioning hardware or software, the service aids lead the customer to place a service call and enable him or her to report an accurate description of the problem. If a problem is not due to malfunctioning hardware or software, the service aids should lead the customer to problem resolution without requiring an unnecessary service call.

8.7 Dynamic CPU de-allocation

Dynamic CPU de-allocation (also known as CPU Gard) is available with AIX Version 4.3.3 on all new pSeries 680 and S80 systems, and on installed Model S80 systems by firmware upgrade. The processors are continuously monitored for errors, such as L2 cache ECC errors. When a predefined error threshold is met, an error log with warning severity and threshold exceeded status is returned to AIX. At the same time, the Service Processor marks the CPU for deconfiguration at the next boot. In the meantime, AIX will attempt to migrate all resources associated with that processor (tasks, interrupts, etc.) to another processor, and then stop the failing processor.

The capability of dynamic CPU de-allocation is only available in systems with more than two processors, because device drivers and kernel extensions which are common to multi-processor and uni-processor systems would change their mode to uni-processor mode with unpredictable results. Note that dynamic CPU de-allocation is disabled by default. See Section 8.7.3.1, “Controlling dynamic CPU de-allocation” on page 111 for details of how to enable it.

8.7.1 Potential Impact to Applications

Dynamic CPU de-allocation is transparent for the vast majority of applications, including drivers and kernel extensions. However, you can use AIX published interfaces to determine whether an application or kernel extension is running on a multiprocessor machine, find out how many processors there are, and bind threads to specific processors.

The interface for binding processes or threads to processors uses logical CPU numbers. The logical CPU numbers are in the range [0..N-1] where N is the total number of CPUs.

To avoid breaking applications or kernel extensions that assume no "holes" in the CPU numbering, AIX always makes it appear for applications as if it is the "last" (highest numbered) logical CPU to be de-allocated. For instance, on an 8-way SMP, the logical CPU numbers are [0..7]. If one processor is de-allocated, the total number of available CPUs becomes 7, and they are numbered [0..6]. Externally, it looks like CPU 7 has disappeared, regardless of which physical processor failed. In the rest of this description, the term CPU is used for the logical entity and the term processor for the physical entity.

Applications or kernel extensions using processes/threads binding could potentially be broken if AIX silently terminated their bound threads or forcefully moved them to another CPU when one of the processors needs to be de-allocated. Dynamic CPU de-allocation provides programming interfaces so that those applications and kernel extensions can be notified that a processor de-allocation is about to happen. When these applications and kernel extensions get this notification, they are responsible for moving their bound threads and associated resources (such as timer request blocks) away from the last logical CPU and adapt themselves to the new CPU configuration.

If, after notification of applications and kernel extensions, some of the threads are still bound to the last logical CPU, the de-allocation is aborted. In this case, AIX logs the fact that the de-allocation has been aborted in the error log and continues using the ailing processor. When the processor ultimately fails, it creates a total system failure. Thus, it is important for applications or kernel extensions binding threads to CPUs to get notice of an impending processor de-allocation, and act on this notice.

Even in the rare cases where the de-allocation can not go through, dynamic CPU de-allocation still gives advanced warning to system administrators. By recording the error in the error log, it gives them a chance to schedule a maintenance operation on the system to replace the ailing component before a global system failure occurs.

8.7.2 Processor De-allocation

The typical flow of events for processor de-allocation is as follows:

1. The firmware detects that a recoverable error threshold has been reached by one of the processors.
2. AIX logs the firmware error report in the system error log, and, when executing on a machine supporting processor de-allocation, start the de-allocation process.
3. AIX notifies non-kernel processes and threads bound to the last logical CPU.
4. AIX waits for all the bound threads to move away from the last logical CPU. If threads remain bound, AIX eventually times out (after ten minutes) and aborts the de-allocation.
5. Otherwise, AIX invokes the previously registered High Availability Event Handlers (HAEHs). An HAEH may return an error that will abort the de-allocation.
6. Otherwise, AIX goes on with the de-allocation process and ultimately stops the failing processor.

In case of failure at any point of the de-allocation, AIX logs the failure with the reason why the de-allocation was aborted. The system administrator can look at the error log, take corrective action (when possible) and restart the de-allocation. For example, if the de-allocation was aborted because at least one application did not unbind its bound threads, the system administrator could stop the application(s), restart the de-allocation (which should go through this time) and restart the application.

8.7.3 System Administration

8.7.3.1 Controlling dynamic CPU de-allocation

Dynamic CPU de-allocation can be enabled or disabled by changing the value of the `cpuguard` attribute of the ODM object `sys0`. The possible values for the attribute are `enable` and `disable`.

The default in AIX Version 4.3.3 is that dynamic CPU de-allocation is disabled (the attribute `cpuguard` has a value of `disable`). System administrators who want to take advantage of this feature must enable it using either the Web-based System Manager system menus, the SMIT System Environments menu, or the `chdev` command.

If dynamic CPU de-allocation is disabled, AIX still reports the errors in the error log and you will see the error indicating that AIX was notified of the problem with a CPU. This type of error log entry is described in Section 8.7.4.2, “CPU_FAIL_PREDICTED” on page 113.

8.7.3.2 Restarting an Aborted Processor De-allocation

Sometimes the processor de-allocation fails because, for example, an application did not move its bound threads away from the last logical CPU. Once this problem has been fixed, by either unbinding (when it is safe to do so) or stopping the application, the system administrator can restart the processor de-allocation process using the `ha_star` command.

The syntax for this command is:

```
ha_star -C
```

where `-C` is for a CPU predictive failure event.

8.7.3.3 Processor State Considerations

Physical processors are represented in the ODM data base by objects named `procn` where *n* is the physical processor number (*n* is a decimal number). Like any other "device" represented in the ODM database, processor objects have a state (Defined/Available) and attributes.

The state of a `proc` object is always *Available* as long as the corresponding processor is present, regardless of whether it is usable by AIX. The state attribute of a `proc` object indicates if the processor is used by AIX and, if not, the reason. This attribute can have three values:

enable The processor is used by AIX.

disable The processor has been dynamically de-allocated by AIX.

faulty The processor was declared defective by the firmware at boot time.

In the case of CPU errors, if a processor for which the firmware reports a predictive failure is successfully de-allocated by AIX, its state goes from `enable` to `disable`. Independently of AIX, this processor is also flagged as defective in the firmware. Upon reboot, it will not be available to AIX and will have its state set to `faulty`. But the ODM `proc` object is still marked available. Only if the defective CPU was physically removed from the system board or CPU board (if it were at all possible) would the `proc` object change to `defined`.

Examples of processor states

Here are some examples of output from the `lsattr` command showing the three possible processor states:

Processor proc4 is working correctly and used by AIX:

```
# lsattr -EH -l proc4
attribute   value               description           user_settable

state       enable              Processor state       False
type        PowerPC_RS64-III    Processor type        False
```

Processor proc4 gets a predictive failure and gets de-allocated by AIX:

```
# lsattr -EH -l proc4
attribute   value               description           user_settable

state       disable             Processor state       False
type        PowerPC_RS64-III    Processor type        False
```

At the next system boot, processor proc4 is reported by firmware as defective and not available to AIX:

```
# lsattr -EH -l proc4
attribute   value               description           user_settable

state       faulty              Processor state       False
type        PowerPC_RS64-III    Processor type        False
```

But in all three cases, the status of processor proc4 is Available:

```
# lsdev -CH -l proc4
name           status                 location               description
proc4          Available              00-04                  Processor
```

8.7.4 Error Log Entries

Following are examples with descriptions of error log entries.

8.7.4.1 errpt short format - summary

Three different error log messages are associated with CPU de-allocation. Following is an example of entries displayed by the `errpt` command (without options):

```
# errpt
IDENTIFIER  TIMESTAMP   T  C  RESOURCE_NAME    DESCRIPTION
804E987A    1008161399  I  O  proc4            CPU DEALLOCATED
8470267F    1008161299  T  S  proc4            CPU DEALLOCATION ABORTED
1B963892    1008160299  P  H  proc4            CPU FAILURE PREDICTED
```

If processor de-allocation is enabled, a `CPU FAILURE PREDICTED` message is always followed by either a `CPU DEALLOCATED` message or a `CPU DEALLOCATION ABORTED` message.

If dynamic CPU de-allocation is not enabled, only the `CPU FAILURE PREDICTED` message is logged. Enabling dynamic CPU de-allocation any time after one or more `CPU FAILURE PREDICTED` messages have been logged initiates the de-allocation process and results in a success or failure error log entry, as described above, for each processor reported failing.

8.7.4.2 CPU_FAIL_PREDICTED

This error indicates that the hardware detected that a processor has a high probability to fail in a near future. It is always logged whether or not processor de-allocation is enabled.

An example error log entry is:

```
LABEL:                  CPU_FAIL_PREDICTED
IDENTIFIER:             1655419A

Date/Time:              Thu Sep 30 13:42:11
Sequence Number:        53
Machine Id:             00002F0E4C00
Node Id:                auntbea
Class:                  H
Type:                   PEND
Resource Name:          proc25
Resource Class:         processor
Resource Type:          proc_rspc
Location:               00-25
```

```
Description
CPU FAILURE PREDICTED

Probable Causes
CPU FAILURE

Failure Causes
CPU FAILURE

Recommended Actions
ENSURE CPU GARD MODE IS ENABLED
RUN SYSTEM DIAGNOSTICS.
```

8.7.4.3 CPU_DEALLOC_SUCCESS

This message is logged by AIX when dynamic CPU de-allocation is enabled, and when the CPU has been successfully de-allocated.

An example error log entry is:

```
LABEL:                  CPU_DEALLOC_SUCCESS
IDENTIFIER:             804E987A

Date/Time:              Thu Sep 30 13:44:13
Sequence Number:        63
Machine Id:             00002F0E4C00
Node Id:                auntbea
Class:                  O
Type:                   INFO
Resource Name:          proc24

Description
CPU DEALLOCATED

Recommended Actions
MAINTENANCE IS REQUIRED BECAUSE OF CPU FAILURE

Detail Data
LOGICAL DEALLOCATED CPU NUMBER
0
```

The preceding example shows that `proc24` was successfully de-allocated and was logical CPU 0 when the failure occurred.

8.7.4.4 CPU_DEALLOC_ABORTED

This message is logged by AIX when dynamic CPU de-allocation is enabled, and when the CPU has not been successfully de-allocated.

The reason code (specified in the `Detail Data` section of the log entry) is a numeric hexadecimal value. The possible reason codes are:

2 One or more processes/threads remain bound to the last logical CPU. In this case, the detailed data lists the PIDs of the offending processes.

3 A registered driver or kernel extension returned an error when notified. In this case, the detailed data field contains the name of the offending driver or kernel extension (ASCII encoded).

4 De-allocating a processor would cause the machine to have less than two available CPUs. AIX does not de-allocate more than N-2 processors on an N-way machine, to avoid confusing applications or kernel extensions using the total number of available processors to determine, whether they are running on a Uni Processor (UP) system, where it is safe to skip the use of multiprocessor locks, or a Symmetric Multi Processor (SMP).

200 (0xC8) Processor de-allocation is disabled (the ODM attribute `cpuguard` has a value of `disable`). You would normally not see this error unless you start `ha_star` manually.

Here are three examples of error log entries for the CPU_DEALLOC_ABORTED event:

Example of kernel extension error:

```
LABEL:                  CPU_DEALLOC_ABORTED
IDENTIFIER:             8470267F
Date/Time:              Thu Sep 30 13:41:10
Sequence Number:        50
Machine Id:             00002F0E4C00
Node Id:                auntbea
Class:                  S
Type:                   TEMP
Resource Name:          proc26

Description
CPU DEALLOCATION ABORTED

Probable Causes
SOFTWARE PROGRAM

Failure Causes
SOFTWARE PROGRAM
```

```
Recommended Actions
MAINTENANCE IS REQUIRED BECAUSE OF CPU FAILURE
SEE USER DOCUMENTATION FOR CPU GARD

Detail Data
DEALLOCATION ABORTED CAUSE
0000 0003
DEALLOCATION ABORTED DATA
6676 6861 6568 3200
```

The preceding example shows that the de-allocation for `proc26` failed. The reason code 3 means that a kernel extension returned an error to the kernel notification routine. The `DEALLOCATION ABORTED DATA` above spells fvhaeh2, which is the name the extension used when registering with the kernel.

Example of a thread still bound to a processor:

```
LABEL:                  CPU_DEALLOC_ABORTED
IDENTIFIER:             8470267F
Date/Time:              Thu Sep 30 14:00:22
Sequence Number:        71
Machine Id:             00002F0E4C00
Node Id:                auntbea
Class:                  S
Type:                   TEMP
Resource Name:          proc19

Description
CPU DEALLOCATION ABORTED

Probable Causes
SOFTWARE PROGRAM

Failure Causes
SOFTWARE PROGRAM

Recommended Actions
MAINTENANCE IS REQUIRED BECAUSE OF CPU FAILURE;
SEE USER DOCUMENTATION FOR CPU GARD

Detail Data
DEALLOCATION ABORTED CAUSE
0000 0002
DEALLOCATION ABORTED DATA
0000 0000 0000 4F4A
```

The preceding example shows that the de-allocation for `proc19` failed. The reason code 2 indicates thread(s) were bound to the last logical processor and did not unbind upon receiving the SIGCPUFAIL signal. The `DEALLOCATION ABORTED DATA` shows that these threads belonged to process 0x4F4A. Options of the `ps` command (-o THREAD, -o BND) allow listings of all threads or processes, with the number of the CPU they are bound to when applicable.

Example of too few active processors:

```
LABEL:                  CPU_DEALLOC_ABORTED
IDENTIFIER:             8470267F

Date/Time:              Thu Sep 30 14:37:34
Sequence Number:        106
Machine Id:             00002F0E4C00
Node Id:                auntbea
Class:                  S
Type:                   TEMP
Resource Name:          proc2

Description
CPU DEALLOCATION ABORTED

Probable Causes
SOFTWARE PROGRAM

Failure Causes
SOFTWARE PROGRAM

Recommended Actions
MAINTENANCE IS REQUIRED BECAUSE OF CPU FAILURE
SEE USER DOCUMENTATION FOR CPU GARD

Detail Data
DEALLOCATION ABORTED CAUSE
0000 0004
DEALLOCATION ABORTED DATA
0000 0000 0000 0000
```

The preceding example shows that the de-allocation of `proc2` failed because there were two or fewer active processors at the time of failure (reason code 4).

Appendix A. Installation requirements

This appendix details the environmental, electrical, and installation requirements of the pSeries 680 and S80 including physical dimensions, clearances, and power requirements.

A fully-configured system consists of a CEC, at least one I/O rack, and at least one I/O Drawer.

A.1 Central Electronic Complex

Table 14 provides the specifications and requirements for the pSeries 680 and S80 Central Electronic Complex (CEC). The CEC component of the pSeries 680 and S80 are only available in 240 Volt single phase configurations. No three-phase options are available. Each CEC requires a single electrical connection.

Table 14. Central Electronic Complex

Dimensions	Metric	Imperial
Height	1577 mm	62.0 in
Width	567 mm	22.3 in
Depth (S80)	1041 mm	40.9 in
Depth (pSeries 680)	1201 mm	47.3 in
Weight	**Metric**	**Imperial**
Minimum	400 kg	880 lbs
Electrical		
Power source loading (Maximum in KVA)	2.129 KVA	
Voltage range (V ac)	200 - 240	
Frequency (Hertz)	50 - 60	
Thermal output (Maximum)	6904 BTU/hr	
Power requirements (Maximum)	2023 watts	
Power factor	0.92 to 0.98	
Inrush current[1]	43 amps	
Maximum altitude	2135 m (7100 ft)	

	Operating	**Non-operating**
Temperature Range	10 to 37.8°C	1 to 60°C
	50 to 100°F	34 to 140°F
Recommended Temperature	24°C (75°F)	
Humidity (Noncondensing)	8 to 80%	8 to 80%
Recommended Humidity	45%	
Wet Bulb Requirements	23°C (73°F)	23°C (73°F)
Noise Emissions	**Operating**	**Idle**
L_{WAd}	7.0 bels	7.0 bels
L_{pAm}	N/A	N/A
$<L_{pA}>_m$	N/A	N/A
Impulsive or prominent discrete tones	No	No

[1] Inrush currents occur only at initial application of power; no inrush occurs during normal power off-on cycle.

A.2 I/O rack

Unless you are planning to install the I/O Drawer in an existing RS/6000 rack, you will require at least one I/O rack. The default I/O rack for the Model S80 is the IBM S00 32-EIA rack; the default I/O rack for the pSeries 680 is the IBM T00 36-EIA rack. You may not configure an S00 rack for a new pSeries 680, although you may use an existing S00 rack if you plan to share with an existing Model S80. The IBM T42 42-EIA rack is available as an option for both systems.

Each I/O rack requires an electrical connection for each Power Distribution Unit (PDU). If you configure an I/O rack with two PDUs, it will require two electrical connections.

The following tables (Table 15, Table 16, and Table 17) provide the dimensions of the three different types of I/O rack:

Table 15. The S00 I/O rack

Dimensions	Metric	Imperial
Height	1577 mm	62.0 in
Width	650 mm	25.5 in
Depth	1019 mm	40.1 in
Weight	**Metric**	**Imperial**
Base rack	159 kg	349 lbs

Table 16. The T00 I/O rack

Dimensions	Metric	Imperial
Height (with AC power)	1804 mm	71.0 in
Width (with side panels)	644 mm	25.4 in
Depth	1147 mm	45.2 in
Weight	**Metric**	**Imperial**
Base rack	244 kg	535 lbs

Table 17. The T42 I/O rack

Dimensions	Metric	Imperial
Height (with AC power)	2015 mm	79.3.0 in
Width (with side panels)	644 mm	25.4 in
Depth	1147 mm	45.2 in
Weight	**Metric**	**Imperial**
Base rack	261 kg	575 lbs

A.3 10 EIA I/O Drawer

The pSeries 680 and S80 use an identical 10 EIA I/O Drawer. The I/O Drawer has dual redundant power supplies and requires two connections on the PDU of the I/O rack in which it is installed. Table 18 details the specifications and requirements of the I/O Drawer.

Table 18. 10 EIA SCSI I/O Drawer

Dimensions	Metric	Imperial
Height	440 mm	17.3 in
Width	443.2 mm	17.5 in
Depth	843.2 mm	33.2 in
Weight	**Metric**	**Imperial**
Minimum configuration	89 kg	195 lbs
Maximum configuration	93 kg	206 lbs
Electrical		
Typical power source loading (in KVA)	0.4 KVA	
Maximum power source loading (in KVA)	1.0 KVA	
Voltage range (V ac)	200 - 240 (autoranging)	
Frequency (Hertz)	50 / 60	
Thermal output (Typical)	1228 BTU/hr	
Thermal output (Maximum)	3071 BTU/hr	
Power requirements (Typical)	360 watts	
Power requirements (Maximum)	900 watts	
Power factor	0.9	
Inrush current	170 amps	
Maximum altitude	2135 m (7000 ft)	
	Operating	**Non-operating**
Temperature Range	10 to 40°C	10 to 52°C
	50 to 104°F	50 to 125.6°F
Humidity (Noncondensing) without tape drive	8 to 80%	8 to 80%
Humidity (Noncondensing) with tape drive	20 to 80%	20 to 80%

Wet Bulb without tape drive	27^{o}C (80^{o}F)	27^{o}C (80^{o}F)
Wet Bulb with tape drive	23^{o}C (73^{o}F)	27^{o}C (80^{o}F)
Noise Emissions	**Operating**	**Idle**
L_{WAd}	5.9 bels	5.3 bels
L_{pAm}	N/A	N/A
$<L_{pA}>_{m}$	N/A	N/A
Impulsive or prominent discrete tones	No	No

A.4 Physical space requirements

Figure 28 on page 124 shows the minimum installation and service clearances required for pSeries 680 and S80 servers.

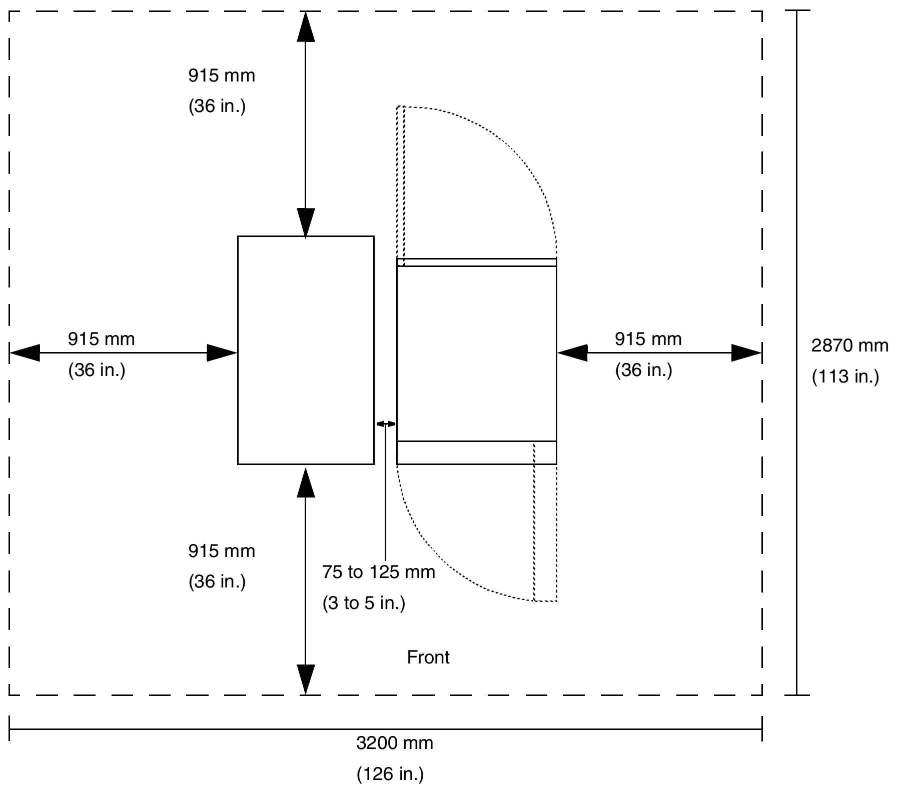


Figure 28. pSeries 680 and S80 system service clearances

Multiple I/O racks can be placed adjacent to one another, with a 75 to 125 mm (3 to 5 in.) gap. Only the racks at each end of the installation require the full 915 mm (36 in.) clearance.

Appendix B. A practical guide to the Service Processor

Both the pSeries 680 and S80 come with a Service Processor to help maintain high reliability, availability, and serviceability (RAS). The Service Processor also makes it possible to perform many maintenance and support tasks from a remote location using a modem.

Even when dial-in support is not contemplated, customers should be aware of the capabilities for calling out to report serious system errors. In addition, there are settings in the Service Processor menus that control the behavior of the system after a power outage and following a hardware or software error. Finally, customers should understand the purpose of system and Service Processor firmware, and they should know how to get updates.

B.1 How to access the Service Processor menus

To access the Service Processor menus, the system administrator needs to access one of the native serial ports: S1 or S2. If remote dial-in support is enabled, one of the ports needs to have an asynchronous modem attached and configured. The other port may have an ASCII terminal attached, or it may have a leased-line or dial-up connection to a remote customer support location. The Service Processor menus give an administrator a great deal of flexibility in deciding which ports to enable and whether to allow dial-out only or to support dial-in as well.

B.1.1 Local access

To access the Service Processor menus from a terminal attached to the S1 or S2 ports, perform the following steps:

1. If the system is running, use the `shutdown` command to bring it to a halt.
2. After the system has powered off, restart it using the white power button on the operator panel.
3. Watch the checkpoints that appear in the operator panel display.
4. Immediately after E04F and just as it enters checkpoint E07A, the system will beep three times. Press any key on the ASCII terminal.
5. Depending on whether a password has been set and which password has been entered, either the Main Menu (privileged access password) or General User Menu (general access password) will appear.

B.1.2 Remote access

To access Service Processor menus from a remote location, an asynchronous modem must be attached to S1 or S2, dial-in must be enabled on the port, and a TTY must be configured in AIX. Perform the following steps:

1. Dial into the system and use the `shutdown` command to power it off.
2. Use the Ring Indicate Power On (RIPO) utility to restart the system. Dial the telephone number of the modem attached to the serial port and let it ring one to three times before hanging up. By default, the RIPO is set to one ring. The number of rings is alterable through Service Aids in diagnostics.
3. Wait about five minutes while the system powers up and pauses for dial-in using the modem. Because it is expecting a dial-in, the system will pause at E07A long enough to allow a remote user to connect to the modem.
4. Depending on whether a password has been set and which password has been entered, either the Main Menu (privileged access password) or General User Menu (general access password) will appear.

B.2 Minimum Service Processor configuration

The default Service Processor values are mostly disabled or unconfigured. In order to use the Service Processor effectively, you need to change the settings to reflect the system behavior you are attempting to control.

The main system behaviors an administrator might want to control include:

- Password protection for the Service Processor menus
- Operating system surveillance to monitor for hangs
- Call-Out policy to report errors
- Call-In policy to allow or disallow remote support
- Determining power and reboot policies
- Determining firmware levels and learning how to update them

In addition, the system administrator should know how processors can be automatically and manually deconfigured and how to cause the system to boot into the Systems Management Services Menu without having to wait for just the right moment to stop the boot process.

B.2.1 Password protection for Service Processor menus

Be sure to set the root password on your server. For your own protection, you should also provide password protection for the Service Processor and Systems Management Services (SMS) menus.

There are two types of passwords for the Service Processor and SMS. One is a general access password that enables a user to look at Vital Product Data (VPD), Service Processor and POST error logs, and boot progress indicators. Additionally, the user can continue the current boot and have the progress indicators displayed on the local or remote terminal being used. The second type of password is for privileged users, such as the system administrator. It controls access to the Main Menu of the Service Processor and SMS utilities.

Passwords can be any combination of up to eight alphanumeric characters. You can enter longer passwords, but the entries are truncated to include only the first eight characters.

The privileged access password can be set from the Service Processor menus or from SMS utilities. The general access password can only be set from the Service Processor menus.

For security purposes, the Service Processor counts the number of unsuccessful attempts to enter a password. After three unsuccessful attempts, the Service Processor responds in one of two ways, depending on whether the attempt was made locally or remotely.

If the attempt was local, the Service Processor resumes the initial program load (IPL). This is based on the assumption that the server is in an adequately secure location to which only authorized users have access. Users must still successfully enter a login password to access AIX.

If the attempt was using a modem, the Service Processor powers the server down in order to prevent potential security attacks by unauthorized remote users.

Table 19 on page 128 illustrates the impact password settings have on which menu you can access on the Service Processor. Pay special attention to the so-called *law of unintended consequences*: If you set password protection on the General User Menu but not the Main Menu, you will have allowed anyone *not* entering a password to access the system's most sensitive settings.

Table 19. Privileged and general access passwords

Privileged Access Password	**General Access Password**	**Resulting Menu**
None	None	Service processor Main Menu.
None	Set	Users with password see General User Menu; those without go to the Main Menu.
Set	None	Users with password see Main Menu, and users without password see General User Menu.
Set	Set	Users with privileged access password see Main Menu; users with general access password see General Users Menu.

To set passwords from the Service Processor menus, perform the following steps:

1. From the Service Processor Main Menu, select **Service Processor Setup Menu**.
2. Select **Change Privileged Access Password**. You should use an eight-character alphanumeric string that can not be easily guessed.
3. Select **Change General Access Password**. You should use a different eight-character alphanumeric string that can not be easily guessed.
4. Return to the Main Menu and exit in order to continue the IPL.

B.2.2 Operating system surveillance

The Service Processor maintains surveillance of the hardware as long as the system is powered on. Using a utility called Repeat-Gard, the firmware automatically marks processors for deconfiguration at the next boot if they experience run-time permanent failures or more than a threshold number of temporary errors. Some errors occur during run-time, but are undetectable by BIST and POST routines. Repeat-Gard will deconfigure the components until they have been replaced. This will prevent getting into a cycle of halting and rebooting, which would interrupt system availability. If a processor is deconfigured, it remains off-line for subsequent reboots until it is replaced.

In addition to monitoring hardware, the Service Processor can be configured to keep track of the operating system by means of a heartbeat. By default, this function is disabled.

To enable it, follow these steps:

1. Access the Service Processor's Main Menu.
2. Select the **Service Processor Setup Menu**.
3. Select **OS Surveillance Setup Menu**.
4. The current setting will be displayed. You may toggle surveillance on or off by pressing the **1** key.
5. Set the surveillance time interval by pressing **2** and entering a value in minutes. The interval is the length of time between heartbeats.
6. Set the surveillance delay in minutes. This is the length of time after the operating system has started before the first heartbeat should be heard.

If the Service Processor does not hear the heartbeat, it assumes an operating system hang has occurred, and it will shut the system down and reboot. For this reason, the system administrator should carefully determine the time interval of the heartbeat and the delay that might be required to avoid a false hang while applications and resources are being brought up.

B.2.3 Configuring call-out

A basic decision needs to be made on whether to allow the system to call out in the event of a failure. This will require a modem attached to either S1 or S2. Enabling call-out does not mean that the system will automatically support dial-in.

When dialing out, the Service Processor can either contact IBM, a customer help desk or a digital pager. IBM and the customer help desk must have programs that know how to decode the information that the Service Processor sends after it connects to a remote modem. IBM Service uses such a program in the United States and many geographies, but you should check locally to see whether this function is available.

If the Service Processor calls IBM, an electronic problem report will be opened in RETAIN. If the customer's help desk receives the call, a so-called *catcher program* must decode the information. Sample catcher code can be found in /usr/samples/syscatch, but customers will need to build their own applications to use this function. If a message is sent to a digital pager, only a phone number will be transmitted. By default, the phone number is the one defined as the system, or modem, telephone line. If dial-in is not permitted, another number could be entered into that field. Whoever is carrying the pager would be responsible for knowing which machine the page came from and what steps to take next.

To enable the call-out feature, you need to configure the modem, configure the serial ports, configure the speed of the serial ports, configure the telephone numbers, determine call-out policy, and enter account information.

B.2.3.1 Configuring the modem

Perform the following steps to configure the modem:

1. Have a modem connected to serial port S1 or S2.
2. Access the Service Processor Main Menu and select the **Call-In/Call-Out Setup Menu**.
3. Select the **Modem Configuration Menu**.
4. Indicate which port will have a modem attached to it and which modem configuration file will be used.
5. After you have linked the correct file to the correct port, save the settings to NVRAM, and configure the modem.
6. Enter `98` to return to the Call-In/Call-Out Setup Menu.

B.2.3.2 Configure the serial ports

Perform the following steps to configure the serial ports:

1. Select the **Serial Port Selection Menu**.
2. Enable Call-Out on the appropriate serial ports.
3. Enter `98` to return to the Call-In/Call-Out Setup Menu.

B.2.3.3 Configure the serial port speed

Perform the following steps to configure the serial port speed:

1. Select the **Serial Port Speed Setup Menu**.
2. Set the speed of the appropriate ports. The default is 9600 baud, which is the slowest recommended speed. Terminal and modem capabilities dictate how much faster you can go. When finished, enter `98` to return to the Call-In/Call-Out Setup Menu.

B.2.3.4 Configure the telephone numbers

Perform the following steps to configure the telephone numbers:

1. Select **Telephone Number Setup Menu**.
2. Assign numbers, appropriate for your geography, for the IBM Service center. If the customer plans to run his own error notification program, configure a number for the help desk. Also, provide entries for Customer Voice Telephone and Customer System Telephone numbers.

> **Note**
>
> If you plan to test call-out, you should initially assign your own telephone number in place of the service center and help desk numbers. That way, you can avoid interrupting ongoing support operations.

3. If you plan to send alerts to a pager, key in the entire string you want to send, including the outside access number (if any), the pager service, the recipient's personal identification number, and any intermediate numeric responses needed before the call-back number. The call-back number is the Customer Voice Telephone number. Be sure to insert pauses to allow for delays caused by prompts from the voice response unit.
4. When finished, enter `98` to return to the Call-In/Call-Out Setup Menu.

B.2.3.5 Set the call-out policy

Perform the following steps to set the call-out policy:

1. Select **Call-Out Policy Setup Menu**.
2. Set the policy to either first or all. The Service Processor will either stop dialing after the first successful connection, or it will dial all three numbers before quitting. You can limit the number of retries.
3. Enter `98` to return to the Call-In/Call-Out Setup Menu.

B.2.3.6 Customer account information

Perform the following steps for customer account information:

1. Select the Customer Account Setup Menu.
2. Enter the customer account number in the first field.
3. If applicable, enter the customer's RETAIN login user ID and password.
4. Enter `98` and return to the Call-In/Call-Out Setup Menu.

B.2.3.7 Test the call-out function

Make sure you have changed the IBM Service and customer help desk numbers to point to your own telephone. If you have a digital pager, make sure the whole string has been entered in the Telephone Numbers Menu.

From the Call-In/Call-Out Setup Menu, select **Call-Out Test**. This will generate a pseudo-error that will trigger a call-out to the first or all of the numbers you have entered, depending on the policy you chose to implement.

B.2.4 Configuring call-in

Configuring call-in is simple once you have entered the parameters for the modems and serial ports.

Dial-in must be enabled on the port to which the modem is attached. To test dial-in, shut down the server and perform the following steps:

1. From any telephone, call the server's telephone number. After you hear three rings, hang up. The server powers on.
2. Give the server five minutes to boot up and prepare to receive another call.
3. From an ASCII terminal or terminal emulator, call the server again. The server answers and presents the Service Processor Menus on your terminal.
4. If required, enter your privileged access password. If no password is required, the Main Menu displays.
5. From the Main Menu, select Continue System Boot to view the IPL progress messages. Depending on your server's configuration, the bootup sequence may take several minutes. Once the bootup completes, the logon prompt displays.

You have successfully called into the Service Processor and brought up the server. Log in and then log out to disconnect from the operating system. Call your server again. The operating system answers and offers the logon prompt.

If these tests are successful, call-in is working correctly.

B.2.5 Power and reboot issues

The pSeries 680 and S80 support unattended start mode and a utility that lets them automatically bring up the operating system after various kinds of failures.

B.2.5.1 Unattended start mode

Unattended start mode means that the Service Processor automatically restores the system power setting after a temporary power failure. It is intended to be used on servers that require automatic power-on after a power failure.

B.2.5.2 Reboot/Restart Setup Menu options

The Service Processor will attempt to reboot the system following a run-time failure. You can control some aspects of that behavior. The options may be set by performing the following steps:

1. From the Main Menu, select the **System Power Control Menu**.
2. From the System Power Control Menu, select **Reboot/Restart Setup Menu**.
3. Select **Number of Reboot Attempts** and enter a value.
4. Select **Use OS-Defined Restart Policy** and toggle the value to give the desired setting. If this is set to Yes, which is the default, the operating system will control whether or not to reboot the system following a system crash. If the value is No, the Service Processor will determine what to do if the system loses control.
5. Select **Enable Supplemental Restart Policy**. The default setting is No. If the operating system has no automatic restart policy, or if it is disabled, this policy will enable reboots/restarts following a hardware or software failure.
6. Select **Call-Out Before Restart**. By selecting the number of this task, you can toggle between enable and disable. If call-out is enabled, a system fault or surveillance failure will cause the Service Processor to dial the configured phone numbers using the current call-out policy.
7. When finished, exit from the menus.

B.2.5.3 Saving customized settings

All of the settings you make (except language) from the Service Processor menus can be saved and used to recover from a fault or to replicate settings to another system that uses the same Service Processor firmware.

Under the diagnostics service aid Save or Restore Hardware Management Policies, you can save your settings to a file. It is strongly recommended that this service aid be used to protect the usefulness of the Service Processor and the availability of the server.

B.2.6 Useful Service Processor utilities

Occasionally, for testing purposes, a system administrator may wish to manually deconfigure one or more processors so that, on the next boot, they will not be initialized. This can be done using a utility accessed through the Main and System Information menus.

B.2.6.1 Configuring/deconfiguring processors

Perform the following steps to configure/deconfigure processors:

1. From the Service Processor Main Menu, select **System Information Menu**.
2. Select **Processor Configuration/Deconfiguration Menu**.
3. Pick the processor or processors that you want to configure or deconfigure, and toggle them on or off. Hexadecimal numbers following each processor denote its status:
   a. 0x00 or 0xFF describe processors configured by the system.
   b. 0x81 is for a processor that has been deconfigured manually.
   c. 0x41 is deconfigured by the system, because the threshold for recoverable run-time errors was exceeded.
   d. 0x21 denotes deconfigured, due to repeated fatal internal errors.
4. Commit the changes and return to the Main Menu.

The system also uses a utility called Repeat-Gard to deconfigure processors that experience either permanent failures or too many temporary errors. The Service Processor deconfigures the processor on the next reboot, which, if autorestart is turned on, would follow the system fault. The processor will remain deconfigured until it is replaced.

B.2.6.2 Boot Mode Menu

There is a firmware option that will control service mode boots. Perform the following steps:

1. From the System Power Control Menu, select **Boot Mode Menu**.
2. The first option is to boot to the SMS Menu. You can toggle this value on and off. If this value has been toggled on, you do not need to manually wait for the right moment to press the terminal enter keys in order to bring up the SMS menu.
3. The next option is Service Mode Boot from Saved List, which boots from the device specified in the service mode boot list saved in NVRAM. If AIX diagnostics has been installed, AIX boots in single-user mode to the diagnostics menu. Using this option to boot the system is the preferred method for running online diagnostics. In this mode, diagnostics has access to the system error log.
4. The next option is Service Mode Boot from Default List, which boots from the devices as defined in system firmware. This is the preferred way of booting from a CD into standalone diagnostics. In this mode, diagnostics

do not have access to disk or the system error log, but can test devices that would be locked if AIX were running.

5. The final option is Boot to Open Firmware Prompt. This option should only be used when you are directed to do so by support personnel or the vendor of a third-party device that uses Open Firmware for installation.

B.3 Determining firmware levels

If AIX is running, the system and Service Processor firmware levels can be determined using the `lscfg` command. If the server is about to be initialized, the firmware levels can be viewed using the SMS utilities for the system and the Service Processor menus for the Service Processor.

Use one of the following procedures to determine your firmware level:

At the AIX command prompt on your system, enter the following command:

```
lscfg -vp | grep -p alterable
```

This command will produce a system configuration report similar to the following:

```
System Firmware:
ROM Level.(alterable).......19990629 (B) 19990621 (A) <= System FW Levels
  Version....................RS6K
System Info Specific.(YL)...P2
SP_CARD_:
  Part Number.................PART_NUM
  EC Level....................EC_LEVEL
  FRU Number..................FRU_NUM_
  Manufacture ID.............IBM
  Serial Number..............SERIAL_#
  Version....................0000RS6K
ROM Level (alterable).......19990630 (B) 19990620 (A) <= SvP FW levels
System Info Specific.(YL)...P2
```

The lines that start *ROM Level (alterable)* list the level numbers of the installed system and Service Processor (SvP) firmware. The levels match the date on which the update was released, in the format YYYYMMDD.

In these examples, the system was booted from type B because it is listed first. The system firmware is, therefore, 19990629. A slightly older copy of the system firmware is in type A, 19990621. You should consider updating your system firmware when new releases become available. When an update is performed, the firmware is written to type B only and can be promoted to type

A at a later time. The system firmware level can also be seen from System Management Services menus.

In the above example, the Service Processor firmware in use is 19990630. Another way of determining this level is to access the Service Processor menus. The firmware level is contained in the heading of the main menu. The following is an example of what the heading looks like:

```
    Service Processor Firmware
        Version: 19990630
Copyright 1998, IBM Corporation
```

The numbers in the second line show the booted firmware level, 19990630. If this level is less than the update level available for your server, you should consider installing the update.

To determine the level of your system firmware using Systems Management Services, perform the following steps:

1. 1. Power the system on after it has been shut down, or reboot the system.
2. Watch for the character-based RS/6000 logo panel and the POST indicators to appear on your terminal.
3. When the word Keyboard appears, immediately press the **1** key. The number 1 key must be pressed before the word Speaker appears.

When the tests have completed and any required passwords have been entered, the System Management Services Utilities menu appears. The System Firmware level is displayed in the top left-hand corner of the display.

When you have read the current firmware level, exit the System Management Services menu as directed on the panel.

If you find the firmware level is less than the update level available for your server, you should consider installing the update.

B.4 Updating firmware

Firmware update packages are available from the RS/6000 Support page on the Internet under RS/6000 Microcode Updates. The URL is:

`http://www.rs6000.ibm.com/support/`

The system firmware and Service Processor firmware are combined into a single download package.

Prior to downloading the firmware, you are asked to read and accept the terms of the Machine Code License Agreement. Once you accept the terms, you are assigned a password. Write down this password, because it is required later to unpack the files you download.

Find the most recent update package for your server. Print the description file, and download one of the format choices, depending on the workstation being used for downloading. The description file provides detailed instructions on downloading and updating the firmware.

Be certain that you do not attempt to load firmware that is inappropriate for your system. Make sure that you follow instructions to the letter, including directions on which module to install first (system or Service Processor). Be aware that installation instructions may vary depending on the level of firmware. Always read the README file before attempting to update firmware.

Appendix C. PCI Dual Channel Ultra2 SCSI Adapter (#6205)

The PCI Dual Channel Ultra2 SCSI Adapter is a 64-bit adapter and is an excellent solution for high-performance SCSI applications. The PCI Dual Channel Ultra2 SCSI Adapter provides two SCSI channels (buses). Each SCSI bus can either be internal or external and supports a data rate of up to 80 MB/s, up to twice the maximum data transfer rate of previous Ultra SCSI adapters, which was 40 MB/s.

In order to achieve an Ultra2 SCSI bus data rate of up to 80 MB/s and also maintain a reasonable drive distance, the adapter utilizes Low Voltage Differential (LVD) drivers and receivers. In order to utilize this Ultra2 80 MB/s performance, all attaching devices or subsystems must also be Ultra2 LVD devices. If any device is not Ultra2 LVD, the adapter will switch its SCSI bus to single-ended (SE) performance and interface at the lower SE SCSI bus data rate of the device.

Two industry-standard VHDCI 68 pin connectors are mounted on the adapter's end bracket, allowing attachment of various LVD and SE external subsystems. The 0.3 meter, VHDCI to P, Mini-68 pin to 68 pin (#2118) converter cable can be used with older external SE subsystems to allow connection to the VHDCI connector on the PCI Dual Channel Ultra2 SCSI Adapter.

Any supported RS/6000 system can be set up to boot from the PCI Dual Channel Ultra2 SCSI Adapter (#6205). If you are running with AIX Version 4.3.3 or later software, this adapter has native boot support as part of that level of AIX software. If you are running with 4.2.1 software, the following procedure applies in order to boot using the PCI Dual Channel Ultra2 SCSI Adapter:

- The designated boot SCSI disk can be located under the covers of a processor unit or in an external SCSI storage unit.
- AIX version 4.2.1 must be loaded to the designated SCSI boot disk using the AIX Network Install Manager (NIM) before booting from the SCSI boot disk.
- The system with a designated SCSI boot disk must have a network connection with another RS/6000 system performing the NIM Master function to perform the install. On RS/6000 SP systems, a similar network install is performed from a control workstation.
- Once AIX Version 4.2.1 with updates is installed on the designated SCSI boot disk and the system is configured for booting, booting takes place

from the boot disk drive without any support from the control processor or NIM Master, and the system does not have to be connected to the network at boot time.

Appendix D. Special notices

This publication is intended to help IBM and Business Partner sales and technical support staff understand the architecture, features, and benefits of the pSeries 680 and S80 servers. The information in this publication is not intended as the specification of any programming interfaces that are provided by the AIX Operating System, Program Number 5765-C34. See the PUBLICATIONS section of the IBM Programming Announcement for the AIX Operating System, Program Number 5765-C34, for more information about what publications are considered to be product documentation.

References in this publication to IBM products, programs or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's product, program, or service may be used. Any functionally equivalent program that does not infringe any of IBM's intellectual property rights may be used instead of the IBM product, program or service.

Information in this book was developed in conjunction with use of the equipment specified, and is limited in application to those specific hardware and software products and levels.

IBM may have patents or pending patent applications covering subject matter 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 IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact IBM Corporation, Dept. 600A, Mail Drop 1329, Somers, NY 10589 USA.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

Any pointers in this publication to external Web sites are provided for convenience only and do not in any manner serve as an endorsement of these Web sites.

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

AIX
e (logo)®
IBM ®
LANStreamer
Netfinity
Redbooks Logo
RS/6000
SecureWay
SP
System/390

AS/400
ESCON
IBMLink
Lotus
Redbooks
RETAIN
S/390
Service Director
Streamer
XT

The following terms are trademarks of other companies:

Tivoli, Manage. Anything. Anywhere.,The Power To Manage., Anything. Anywhere.,TME, NetView, Cross-Site, Tivoli Ready, Tivoli Certified, Planet Tivoli, and Tivoli Enterprise are trademarks or registered trademarks of Tivoli Systems Inc., an IBM company, in the United States, other countries, or both. In Denmark, Tivoli is a trademark licensed from Kjøbenhavns Sommer - Tivoli A/S.

C-bus is a trademark of Corollary, Inc. in the United States and/or other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States and/or other countries.

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

PC Direct is a trademark of Ziff Communications Company in the United States and/or other countries and is used by IBM Corporation under license.

ActionMedia, LANDesk, MMX, Pentium and ProShare are trademarks of Intel Corporation in the United States and/or other countries.

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

SET, SET Secure Electronic Transaction, and the SET Logo are trademarks owned by SET Secure Electronic Transaction LLC.

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

Appendix E. Related publications

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

E.1 IBM Redbooks

For information on ordering these publications see “How to get IBM Redbooks” on page 147.

- *AIX Version 4.3 Differences Guide*, SG24-2014
- *AIX 5L Workload Manager (WLM)*, SG24-5977
- *RS/6000 S-Series Enterprise Servers Handbook*, SG24-5113

E.2 IBM Redbooks collections

Redbooks are also available on the following CD-ROMs. Click the CD-ROMs button at ibm.com/redbooks for information about all the CD-ROMs offered, updates and formats.

CD-ROM Title	**Collection Kit Number**
IBM System/390 Redbooks Collection	SK2T-2177
IBM Networking Redbooks Collection	SK2T-6022
IBM Transaction Processing and Data Management Redbooks Collection	SK2T-8038
IBM Lotus Redbooks Collection	SK2T-8039
Tivoli Redbooks Collection	SK2T-8044
IBM AS/400 Redbooks Collection	SK2T-2849
IBM Netfinity Hardware and Software Redbooks Collection	SK2T-8046
IBM RS/6000 Redbooks Collection	SK2T-8043
IBM Application Development Redbooks Collection	SK2T-8037
IBM Enterprise Storage and Systems Management Solutions	SK3T-3694

E.3 Other resources

These publications are also relevant as further information sources:

- *Enterprise Server Model S80/pSeries 680 Model S85 Installation Guide,* SA38-0582
- *Enterprise Server Model S80/pSeries 680 Model S85 Service Guide*, SA38-0558
- *Enterprise Server Model S80/pSeries 680 Model S85 User’s Guide,* SA38-0557

- *PCI Adapter Placement Reference Guide*, SA38-0538
- *RS/6000 SP: Planning, Volume 1 Hardware and Physical Environment*, GA22-7280

E.4 Referenced Web sites

These Web sites are also relevant as further information sources:

- `http://www.tpc.org/` - Home of the Transaction Processing Council
- `http://www.oracle.com/apps_benchmark/` - Oracle Applications Standard Benchmark description and results
- `http://ehone.ibm.com/public/applications/econfig/` - Download site for the IBM Configurator for e-business
- `http://techsupport.services.ibm.com/rs6k/fixes.html` - Download site for AIX fixes
- `http://www.rs6000.ibm.com/support/micro/` - Download site for pSeries 680 and S80 microcode

How to get IBM Redbooks

This section explains how both customers and IBM employees can find out about IBM Redbooks, redpieces, and CD-ROMs. A form for ordering books and CD-ROMs by fax or e-mail is also provided.

- **Redbooks Web Site** **ibm.com**/redbooks

  Search for, view, download, or order hardcopy/CD-ROM Redbooks from the Redbooks Web site. Also read redpieces and download additional materials (code samples or diskette/CD-ROM images) from this Redbooks site.

  Redpieces are Redbooks in progress; not all Redbooks become redpieces and sometimes just a few chapters will be published this way. The intent is to get the information out much quicker than the formal publishing process allows.

- **E-mail Orders**

  Send orders by e-mail including information from the IBM Redbooks fax order form to:

	e-mail address
In United States or Canada	pubscan@us.ibm.com
Outside North America	Contact information is in the “How to Order” section at this site: http://www.elink.ibmlink.ibm.com/pbl/pbl

- **Telephone Orders**

United States (toll free)	1-800-879-2755
Canada (toll free)	1-800-IBM-4YOU
Outside North America	Country coordinator phone number is in the “How to Order” section at this site: http://www.elink.ibmlink.ibm.com/pbl/pbl

- **Fax Orders**

United States (toll free)	1-800-445-9269
Canada	1-403-267-4455
Outside North America	Fax phone number is in the “How to Order” section at this site: http://www.elink.ibmlink.ibm.com/pbl/pbl

This information was current at the time of publication, but is continually subject to change. The latest information may be found at the Redbooks Web site.

IBM Intranet for Employees

IBM employees may register for information on workshops, residencies, and Redbooks by accessing the IBM Intranet Web site at http://w3.itso.ibm.com/ and clicking the ITSO Mailing List button. Look in the Materials repository for workshops, presentations, papers, and Web pages developed and written by the ITSO technical professionals; click the Additional Materials button. Employees may access MyNews at http://w3.ibm.com/ for redbook, residency, and workshop announcements.

IBM Redbooks fax order form

Please send me the following:

Title	Order Number	Quantity

First name Last name

Company

Address

City Postal code Country

Telephone number Telefax number VAT number

□ Invoice to customer number

□ Credit card number

Credit card expiration date Card issued to Signature

We accept American Express, Diners, Eurocard, Master Card, and Visa. Payment by credit card not available in all countries. Signature mandatory for credit card payment.

Abbreviations and acronyms

ARP	Address Resolution Protocol
ATM	Asynchronous Transfer Mode
BI	Business Intelligence
BIST	Built-In Self Test
CEC	Central Electronic Complex
CPU	Central Processing Unit
CRC	Cyclic Redundancy Checking
CWS	Control Workstation
CUoD	Capacity Upgrade on Demand
DDR	Double Data Rate
ECC	Error Checking and Correction
ERP	Enterprise Resource Planning
ESCON	Enterprise Systems Connection
FC-AL	Fiber Channel-Arbitrated Loop
FDDI	Fiber Distributed Data Interface
Gb	Gigabit
GB	Gigabyte
GB/s	Gigabytes per second
HACMP	High Availability Cluster Multi Processing
HMT	Hardware Multithreading
HTTP	Hypertext Transfer Protocol
IBM	International Business Machines Corporation
I/O	Input/Output
IP	Internet Protocol
IPL	Initial Program Load
ISO	International Standards Organization
ITSO	International Technical Support Organization
JFS	Journal File System
L1	Level 1
L2	Level 2
LCD	Liquid Crystal Display
LPAR	Logical Partitioning
LVD	Low Voltage Differential
LVM	Logical Volume Manager
Mb	Megabit
MB	Megabyte
MB/s	Megabytes per second
MES	Miscellaneous Equipment Specification
MHz	Megahertz
NIM	Network Install Manager
NIS	Network Information System
NVRAM	Non-Volatile Random Access Memory
OLTP	On-line Transaction Processing
PCI	Peripheral Component Interconnect
POST	Power-On Self Test

PSSP	Parallel Systems Support Program
RAID	Redundant Array of Independent Disks
RAS	Reliability Availability Serviceability
RIO	Remote Input/Output
RIPO	Ring Indicator Power On
ROM	Read-only Memory
SCSI	Small Computer System Interface
SDRAM	Synchronous Dynamic Random Access Memory
SRAM	Static Random Access Memory
SE	Single Ended
SMIT	System Management Interface Tool
SMP	Symmetric Multiprocessor
SMS	Service Management System
SOI	Silicon on Insulator
SPEC	System Performance Evaluation Corporation
SSA	Serial Storage Architecture
TCP	Transmission Control Protocol
TPC	Transaction Processing Performance Council
URL	Uniform Resource Locator
VHDCI	Very High Density Cable Interconnect
VMM	Virtual Memory Manager
VPD	Vital Product Data
WLM	Workload Manager

Index

Symbols

$/tpmC result 9

Numerics

10/100 Ethernet 7
2447 Attachment to PCI SCSI Adapter 18, 30
2624 SCSI-2 CD-ROM 19, 31
2708 Eicon ISDN DIVA PRO 2.0 PCI S/T Adapter 52
2751 S/390 ESCON Channel Adapter 52
2830 POWER GXT130P Graphics Adapter 52
2838 POWER GXT120P Graphics Adapter 52
2913 9.1 GB Ultra SCSI 16-bit Hot-Swap 18, 30
2944 128-Port Asynchronous Adapter 50
2975 10/100/1000 Base-T Adapter 18, 31
3002 9.1 GB 10,000 RPM Ultra SCSI 1-inch (25-mm) Hot-Swap 18, 30
3104 18.2 GB Ultra SCSI 16-bit Hot-Swap 18, 30
3117 18.2 GB 10,000 RPM Ultra SCSI Hot-Swap 18, 30
3125 serial-to-serial cable 50
3142 3-meter Remote I/O Cable, Drawer-to-Drawer 20, 32
3143 6-meter Remote I/O Cable, Rack-to-Rack 20, 32
3144 15-meter Remote I/O Cable, Rack-to-Rack 20, 32
32-bit PCI slots 2, 84
4190 1024 MB Memory (4 x 256 MB Cards) 30
4191 2048 MB Memory (4 x 512 MB Cards) 17, 30
4192 4096 MB Memory (4 x 1024 MB Cards) 17, 30
4193 8192 MB Memory (4 x 2048 MB Cards) 17, 30
4194 16384 MB Memory (4 x 4096 MB Cards) 17, 30
4195 32768 MB Memory (4 x 8192 MB Cards) 17
5318 450 MHz Card, First 28
5319 450 MHz Card, Additional 28
5320 600 MHz Card 28
5320 600 MHz Card, First 15
5321 600 MHz Card, Additional 15, 28
5765-D51 PSSP 58
6006 2-meter Drawer-to Drawer Control Cable 19, 32
6007 15-meter Rack-to-Rack Control Cable 19, 32
6008 6-meter Rack-to-Rack Control Cable 19, 32
6171 Side-Mounted, 1-Phase 39
6173 Side-Mounted, 3-Phase 39
6174 Side-Mounted, 3-Phase, Swiss 40
6205 Dual Channel Ultra2 SCSI adapter 7, 18, 31, 83, 139
6206 Ultra SCSI Single-Ended Adapter 18, 31
6207 Ultra SCSI Differential Adapter 18, 31
6230 Advanced SerialRAID adapter 18, 30
6320 Base SCSI I/O Drawer, 10 EIA 19, 31
6321 Primary I/O Drawer Group 19, 32
6323 Secondary I/O Drawer Group 19, 32
64-bit PCI slots 2, 84
6503 Remote I/O Hub - Dual Loop 63
6547 SCSI 6-pack 18, 30
6913 1000 Watt AC Power Supply 16, 28
6914 Programmable Power Regulator 16, 29
6915 Processor Power Regulator 16, 29
7 EIA I/O drawers 19, 31
7007 Rackless System Order - Plant Install 53
7009 Rackless System Order - Field Install 53
8133 RJ-45 to DB-25 converter cable 50
8136 128-port Asynchronous cable 50
8300 450 MHz Card 28
8301 600 MHz Card, Additional 16, 28
8302 450 MHz Card 28
8303 600 MHz Card, Additional 16, 28
8304 450 MHz Card, Additional 28
8305 600 MHz Card, Additional 16, 28
8396 SP System Attachment Adapter 57
8622 dual line cord feature 16
9171 Side-Mounted, 1-Phase 39
9173 Side-Mounted, 3-Phase 39
9174 Side-Mounted, 3-Phase, Swiss 39
9800 power cord specify 16

A

absolute targets (WLM) 92
accessing Service Processor menus 125
active backplane 78
adapter limits 51
adapter placement 51
affinity 11
AIX
- developer kit (JDK) 85
- diagnostics 107

kernel scalability 10
alternate boot source 105
anti-spamming features 86
ARTIC Adapters
supported in Model S80 37
supported in pSeries 680 24
ASCII terminal 50, 132
accessing the Service Processor 125
Async Adapters
supported in Model S80 37
supported in pSeries 680 24
ATM Adapters
supported in Model S80 37
supported in pSeries 680 24
automatic power-on 132
available PCI slots 3

B
Baan 9
backplane 78
backup mirror copy 88
bandwidth
L2 cache 72, 75
memory 76
Remote I/O 44, 82
batch jobs 11, 91
benchmark results 8
Baan 9
Oracle Applications 9
PeopleSoft 10
SAP 9
SPECweb99 10
VolanoMark 10
BI 4
bindintcpu command 81
BIST 102
BNC Ethernet 60
book packaging 70
boot
bootlist 105
device location 18, 31, 44
errors 18, 31
supported devices (Model S80) 30
supported devices (pSeries 680) 18
boot mode menu 134
boot sequence history 105
boot time reduction 87
bosdebug command 82
bound threads 109
branch buffer 72, 75
branch unit 72, 75
built-in self test 102
bus arbiter chip 76
bus frequency 42, 78
bus loading 51
bus transfer speed 78
Business Intelligence 4

C
cached data 11
call home 3, 105
call-in 132
call-out 129
call-out policy 131
Capacity Upgrade on Demand
description 97
APAR IY10846 97
chcod command 98
effect of processor failure 99
feature codes for Model S80 28
feature codes for pSeries 680 16
lscfg command 99
catcher program
decoding electronic problem reports 129
CEC - see Central Electronic Complex
Central Electronic Complex
description 2
cabling 19, 32
physical layout 67
CES - see clustered enterprise servers
cfgmgr command 87
chcod command 98, 99
chdev command 111
chfs command 88
chip interconnect wiring 102
Cisco EtherChannel 86
classes (WLM) 92
clocking frequency 78
clustered enterprise servers
description 3
configuring using PCRS6000 61
support for Model S80 25
support for pSeries 680 13
commands
bindintcpu 81
bosdebug 82

cfgmgr 87
chcod 98, 99
chdev 111
chfs 88
diag 81
errpt 113
ha_star 111
lsattr 112
lscfg 81, 99
lsdev 81, 112
migratepv 89
replacepv 89
rmfs 89
shutdown 125
unmount 89
wlmcntrl 96
wlmstat 95, 96
complex fixed point unit 72, 75
concurrent diagnostics 102, 107
concurrent maintenance 102
configuration manager 87
configuration validation 51
configurator 49
configuring clustered enterprise servers 60
configuring HACMP 50
configuring high-availability solutions 54
configuring rackless sytems 52
HA-S80 54
HA-S85 54
SP-attached servers 57
conflicting resource values (WLM) 95
connectivity assurance 51
console 50, 52
console logging 11
console mirroring 105
constrained class (WLM) 94
contention 10
Control Workstation
features required for SP-attached servers 60
required for clustered enterprise servers 60
cooling fans 67
copper chip 71, 74
copper technology 1
cost per transaction 9
CPU Gard - see dynamic CPU de-allocation
CPU intensive applications 11
CPU utilization - WLM monitoring 92
crash command 108
cross-port traffic 78
CUoD - see Capacity Upgrade on Demand
CWS - see Control Workstation
cyclic redundancy checking 102

D
DASD six-pack 3
data availability 88, 90
data buffers 76
data synchronization 89
DDR 72
deconfiguring processors 134
Default Class (WLM) 92
default target share (WLM) 93
device configuration 11
device configuration methods 87
diag command 81
diagnostic event log 107
diagnostic exerciser 107
diagnostic routines 105
dial-in support 125
dial-out only 125
digital pager
use with Service Processor 129
Digital Trunk Adapters
supported in Model S80 37
supported in pSeries 680 24
disk drives 17, 30
disk six-pack 3
dispatch buffer 72, 75
dispatcher 11
display panel 40
Double Data Rate 72
drawer limitations 51
Dual Channel Ultra2 SCSI adapter 7, 31
dual line cord 16, 29
dynamic CPU de-allocation
description 108
and Hardware Multithreading 80
chdev command 111
error log entries 113
flow of events 110
ha_star command 111
logical CPU numbers 109
lsattr command 112
lsdev command 112
ODM objects 111
reasons for failure 114
SMP support 109

E

ECC 1
e-configurator 49
electrical requirements 119
Electronic Service Agent
  description 4
  details 106
  relationship to Service Processor 104
en0 58
enabling dynamic CPU de-allocation 111
enabling HMT 80, 82
Enterprise Resource Planning 4
environmental requirements 119
ERP 4
errdemon 105
error analysis 3
error correcting code 102
error log 105
error log entries for CPU Gard 113
errpt command 113
Ethernet Adapters
  boot support 18, 30
  supported in Model S80 38
  supported in pSeries 680 24
execution rate 72, 75
extended mode tests 102
external node 58

F

false hang 129
fans 40
fault tolerance 102
FDDI Adapters
  supported in Model S80 38
  supported in pSeries 680 24
feature codes
  supported on Model S80 35
  supported on pSeries 680 22
  see also Numerics section at front of index
Fibre Channel Adapters
  supported in Model S80 37
  supported in pSeries 680 23
file system activity 88
firmware 44
firmware levels 135
fixed point unit 72, 75
floating point unit 72, 75
frame limits 60
functionality assurance 51

G

general access password 125
Gigabit Ethernet 7
global run queue 86
Graphics Accelerators
  supported in Model S80 36
  supported in pSeries 680 23
graPHIGS 85
gratuitous ARP 86
ground cable 59

H

ha_star command 111
HACMP 3, 52
HACMP heartbeat 52
hard constraints (WLM) 94
hardware analysis 3
Hardware Multithreading
  description 80
  and dynamic CPU de-allocation 80
  APAR IY13075 80
  need for planning 80
HA-S80 4, 52
HA-S85 4, 52
heartbeat 52
  detecting system hangs 129
high-availability solutions 3, 52
HMT - see Hardware Multithreading 80
Host Attachment
  feature codes for pSeries 680 23
  supported in Model S80 36
hot-plug six-pack 3
hot-swappable components 102

I

I/O bridge bus 84
I/O Drawer
  description 40
  maximum in 7014-S00 rack 39
  maximum in 7014-T00 rack 39
  maximum in 7014-T42 rack 39
  PCI buses 41
  power and cooling 42
I/O racks
  7014-S00 39

7014-T00 39
7014-T42 39
Model S80 6, 25
pSeries 680 5, 13
IBM Configurator for e-business 49
IBM maintenance agreement 106
independent resource limits (WLM) 95
initial program load 127
inittab entry 96
instruction buffer 72, 75
instructions per cycle 72, 75
interactive applications 11
interactive jobs 91
investment protection 13, 25
IPL 127
ISDN Adapter
supported in Model S80 38
supported in pSeries 680 24

J
JFS log logical volume 88
JFS online backup 11, 87
JTAG port 68

K
kdb command 108
Kernel Debugger 108
kernel extensions 109
kernel processes (WLM) 92
kernel scalability 10

L
L1 cache latency 74
L2 cache 4, 6, 75
L2 cache bandwidth 72, 75
latency 10, 71, 74, 77
Level 2 cache 4, 6
load balancing 86
load/store unit 75
local run queue 86
lock contention 11, 87
logical CPU numbers 109
logical frame 60
logical nodes 59
logical partitioning 11
Logical Volume Manager 89
low voltage differential 139
LPAR 11
lsattr command 112
lscfg 81
lscfg command 99, 135
lsdev command 81, 112
LVD 139
LVM compatibility 89

M
manuals 21, 34
maximum adapter configurations 51
maximum resources (WLM) 92
maximum RIO loop 47
maximum SPCN loop 44
media bays 3
memory
availability features 103
bandwidth 76
books 78
installation rules 79
latency 10
modules 78
quad physical layout 78
quads 4, 6
redundancy features 79
memory contention 10
memory intensive applications 11, 17, 29, 91
mezzanine bus 84
migratepv command 89
minimum resources (WLM) 92
minimum RIO loop 45
minimum SPCN loop 43
mirrored file system copy 87
mirrored logical volumes 89
mirrored striped logical volume 89
mirroring a dump device 107
mission-critical applications 101
mission-critical servers 103
mixing processor types 28
mksysb on CD-R 11
model conversion 63
Model S70 Advanced
upgrading 63
Model S70 I/O drawers 19, 31
Model S80
announce date 25
as an SP-attached server 57
base configuration 6, 26, 32

  CD-ROM 25
  clustered enterprise servers 25
  CUoD feature codes 28
  HA-S80 54
  I/O drawer groups 32
  I/O racks 6, 25
  L2 cache 6
  maximizing memory performance 17, 29
  maximum configuration 7, 33
  media bays 7
  memory feature sizes 29
  memory performance 76
  memory quads 76
  memory slots 7
  minimum system memory 6, 26
  optional features 35
  processor expansion 7
  publications 63
  rackless order 52
  RIO cables 20, 32
  serial ports 50
  SPCN cables 19, 32
  supported boot devices 30
  system bandwidth 76
  system buses 75
modem 106, 125
multiple operating system images 11
multiple page replacement daemons 10
multiple PDUs 42
multiple run queues 11, 86

N
N+1 redundancy 103
native serial ports 50
  accessing the Service Processor 125
network adapters
  supported in Model S80 37
  supported in pSeries 680 24
NIM 139
NIS+ 85
non-volatile memory 104
NVRAM 105

O
OLTP throughput 10
on-chip data buses 76
online JFS backup 11, 87
Open Firmware prompt 135
OpenGL 85
operating system hang 129
operating system partitioning 11
operating system surveillance 105, 128
operator panel 105
optimal adapter performance 51
Oracle Applications 9

P
page replacement daemon 10
parallel device configuration 11
parallel write policy 90
partition allocation policy 89
partitioning 11
passive mode (WLM) 95
password sharing 105
PCI 2.1 84
PCI adapter placement 51
PCI bridge chips 84
PCI bus loading 51
PCI bus speed 2, 42
PCI signaling 84
PCI slot availability 41
PCI throughput 84
PCI-to-PCI bridges 84
PCRS6000 configurator 49
PDU 39, 42, 120
PeopleSoft 10
performance 8
  Hardware Multithreading 80
  transaction processing 9
periodic load balancing 86
persistent CPU de-allocation 112
physical dimensions 119
physical memory management 10
physical memory utilization - WLM monitoring 92
pipeline execution 72, 75
pipeline mode 84
POST 102
power distribution units 39
power failure 132
power outage 125
power requirements 119
power-off warnings 3
power-on self test 102
primary drawer connection 44
privileged access password 125
problem determination 107

problem isolation 101
process limits 86
process table lock 87
Processor
  feature codes for pSeries 680 22
processor affinity 11, 87
processor arrays 72, 75
processor binding 109
processor book 7
processor books 78
processor regulator 16, 28
Processors
  feature codes for Model S80 35
  feature codes for pSeries 680 22
progress indicators 127
pSeries 680
  announce date 13
  as an SP-attached server 57
  base configuration 4, 14, 20
  benchmark results 9, 10
  CD-ROM 13
  clustered enterprise servers 13
  CUoD feature codes 16
  Double Data Rate 72
  Hardware Multithreading 80
  HA-S85 54
  I/O drawer groups 19
  I/O racks 5, 13
  L2 cache 4
  maximum configuration 7, 21
  memory feature sizes 16
  memory performance 76
  memory quads 76
  minimum system memory 4, 14
  optional features 22
  publications 34
  rackless order 52
  serial ports 50
  supported boot devices 18
  transaction processing performance 9
PSSP 57, 60
  support for Model S80 25
  support for pSeries 680 13
PSSP licence 58
publications 21, 34

Q
QoS 86
quads 4, 6
Quality of Service 86

R
rack contents specify code 52
rack ordering rules 40
rack power 39
rackless order 52
RAID 0 89
RAID 1 89
random access 90
RAS 101
reboot attempts 133
reboot policy 105
reducing footprint 52
redundant arrays 102
redundant bulk power supplies 103
redundant cooling 40
redundant data paths 44
redundant hardware components 3
redundant power 39, 40
relative resource amounts 93
remote access 3
remote dial-in support 125
Remote I/O
  bandwidth 44, 82
  cabling
    four I/O Drawers 46
    single I/O Drawer 44
    three I/O Drawers 46
    two I/O Drawers 45
  connections 82
  description 44
  loops 82
Remote I/O hub 76
remote power-on 105
remote support 126
Repeat-Gard 128
replacepv command 89
reserved resources (WLM) 94
resources (WLM)
  allocation 91, 95
  boundaries 92
  rules 94
  targets 92, 94
restricted resources (WLM) 94
RETAIN 129
ring indicate power on 126

RIPO 126
rmfs command 89
RS/6000 SP 13, 25
features required for SP-attached servers 58
RS-232 cables 59
RS-232 ports 60
RS64 III processor 73
RS64 III processor card 6, 27, 75
RS64 IV processor 71
RS64 IV processor card 4, 15
run-time monitoring 105
run-time permanent failures 128

S

S70 Advanced
memory conversion 65
S70 I/O drawer 84
SAP 9
scalability 8
scalability enhancements 86
scaling efficiency 8
scheduler 11, 91
SCSI Adapters
supported in Model S80 37
supported in pSeries 680 23
SCSI adapters 18, 30
SCSI boot support 18, 30
SCSI card 3
SCSI data rate 139
SCSI six-pack 3
SecureWay Directory 86
seek times 90
Sendmail 86
sequential operations 90
serial cables 59
serial port connections 60
serial ports 52
serial write policy 90
serialization mechanism 87
server consolidation
benefit of WLM 91
service agreements 91
Service Aids 106, 126
service clearances 123
Service Director - see Electronic Service Agent
Service Processor
description 3
accessing 125
default settings 126
E04F checkpoint 125
E07A checkpoint 125
general access password 125
location in primary I/O Drawer 51
menus 127, 136
privileged access password 125
progress indicators 127
remote access 126
requirements for SP-attached servers 58
support for OEM asynchronous terminals 51
unsuccessful attempts to enter password 127
service processor 3
shares (WLM) 93
shutdown command 125
silicon-on-insulator 1
silicon-on-insulator (SOI) 71
single frame system 58
single operating system image 11
single phase configurations 119
single point of failure 101
SMIT 96
SMS - see Systems Management Services 127
SOCKS V5 86
SOI 1
SP
features required for SP-attached servers 58
SP node ethernet boot 60
SP-attached server 3, 13, 25
configuring 57
SPCN 3, 42
SPECweb99 10
splitcopy 88
SPS switch 57
SSA Adapters
supported in Model S80 37
supported in pSeries 680 23
SSA boot support 18, 30
Standard Ethernet 7
stress testing 101
striped logical volumes 89
striped mirrored logical volume 89
Super Strict 89
superscalar 71, 74
swapping performance (WLM) 94
switch cable 58
switch contention 76
switch node numbers 58
switch traffic 76

system backplane 1, 68, 78
System Class (WLM) 92
system console 50, 52
system console logging 11
system crash dumps 108
System Data Repository 58
system demands 11
system dispatcher 11
system firmware 44, 106
system hangs 3
system maintenance 91
System Power Control Network 3
system process limits 86
system resource allocation control 11
system restart 3
system scheduler 11
system threads limit 86
system throughput 87
Systems Management Services
    menu 134
    setting passwords 127

T

Tape Drives
    in I/O racks 40
    supported in Model S80 media bay 36
    supported in pSeries 680 media bay 23
terminal 50
terminal emulator 132
testing call-out 131
thread limits 86
thread load balancing 86
threads binding 109
three phase options 119
tiers (WLM) 92
timer request blocks 109
timing paths 74
Token-Ring Adapters
    supported in Model S80 38
    supported in pSeries 680 24
TP Ethernet 60
TPC-C benchmark 9
tpmC result 9
transaction processing 91
transaction processing performance 9
transfer buffers 76
triple mirror 88
twisted pair cables 60
twisted pair Ethernet 60

U

Ultra SCSI card 3
umount command 89
unattended start mode 132
uninterruptible power supply 103
UNIX 98 85
UPS 16, 29, 52

V

VHDCI 139
virtual memory manager 11, 91
vital product data 105, 127
VMM 11
VolanoMark 10
VPD 127

W

WAN Adapters
    supported in Model S80 38
    supported in pSeries 680 24
Web-based System Manager
    use with Workload Manager 96
WLM - see Workload Manager
wlmcntrl command 96
wlmstat command 95, 96
working set 71, 74
workload isolation 11, 91
Workload Manager
    APAR IY06844 96
    classes 92
    comparison with LPAR 11
    conflicting resource values 95
    constrained class 94
    Default Class 92
    default target share 93
    description 11
    hard constraints 94
    independent resource limits 95
    passive mode 95
    planning for 95
    reserved resources 94
    resource allocation 95
    resource rules 94
    resources 92, 94
    restricted resources 94

shares 93
swapping performance 94
System Class 92
tiers 92
Web-based System Manager interface 96
wlmcntrl command 96
wlmstat command 95, 96
write policy 90
wsm - see Web-based System Manager

X
X11R6.3 85

IBM Redbooks review

Your feedback is valued by the Redbook authors. In particular we are interested in situations where a Redbook "made the difference" in a task or problem you encountered. Using one of the following methods, **please review the Redbook, addressing value, subject matter, structure, depth and quality as appropriate.**

- Use the online **Contact us** review redbook form found at **ibm.com**/redbooks
- Fax this form to: USA International Access Code + 1 845 432 8264
- Send your comments in an Internet note to redbook@us.ibm.com

Document Number **Redbook Title**	SG24-6023-00 IBM @server pSeries 680 Handbook Including RS/6000 Model S80
Review	
What other subjects would you like to see IBM Redbooks address?	
Please rate your overall satisfaction:	O Very Good O Good O Average O Poor
Please identify yourself as belonging to one of the following groups:	O Customer O Business Partner O Solution Developer O IBM, Lotus or Tivoli Employee O None of the above
Your e-mail address: The data you provide here may be used to provide you with information from IBM or our business partners about our products, services or activities.	 O Please do not use the information collected here for future marketing or promotional contacts or other communications beyond the scope of this transaction.
Questions about IBM's privacy policy?	The following link explains how we protect your personal information. **ibm.com**/privacy/yourprivacy/

IBM Redbooks

IBM *e*server pSeries 680 Handbook Including RS/6000 Model S80

IBM @server pSeries 680 Handbook

Including RS/6000 Model S80

Details the system architecture of the fastest UNIX server

Covers available CPU, memory, disk and adapter features

Capacity Upgrade on Demand explained

This redbook covers the IBM @server pSeries 680 and RS/6000 Enterprise Server Model S80 (hereafter referred to in this redbook as the pSeries 680 and S80 respectively). It will help you understand the architecture of each machine and the similarities and differences between them. An overview of the optional features for each machine is also provided along with advice on how to use the PCRS6000 Configurator to produce a valid configuration.

This publication is suitable for professionals wishing to acquire a better understanding of the pSeries 680 and Model S80 including:

- Customers
- Sales and Marketing professionals
- Technical Support professionals
- Business Partners

This publication does not replace the latest marketing materials and tools. It is intended as an additional source of information that, together with existing sources, may be used to enhance your knowledge of IBM Enterprise Server products.

This book is a replacement for *RS/6000 S-Series Enterprise Servers Handbook*, SG24-5113.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-6023-00 ISBN 0738419273