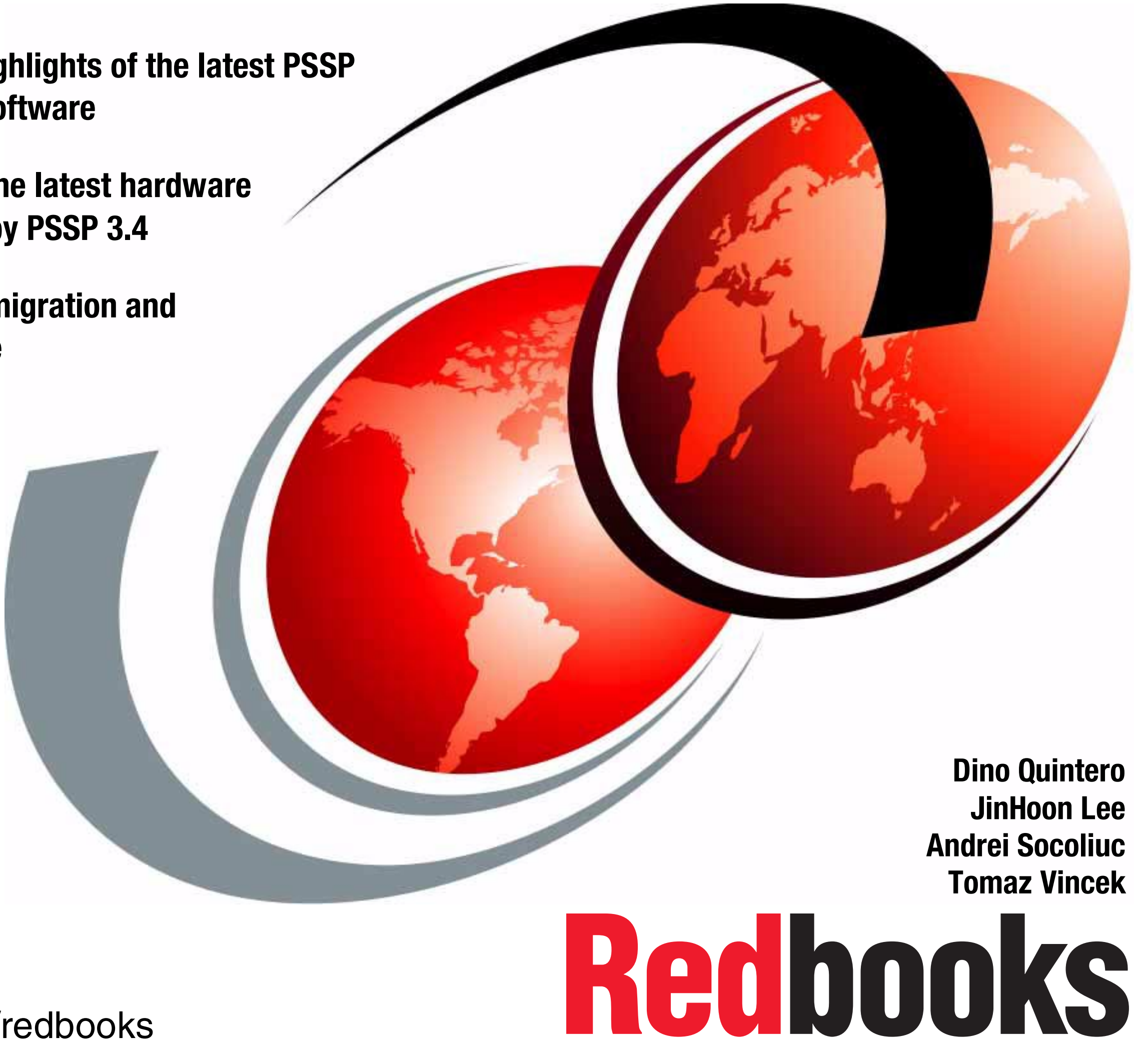
IBM @server Cluster 1600 and PSSP 3.4 Cluster Enhancements



Provides highlights of the latest PSSP clustered software

Describes the latest hardware supported by PSSP 3.4

Discusses migration and coexistence

Dino Quintero
JinHoon Lee
Andrei Socoliuc
Tomaz Vincek

Redbooks

ibm.com/redbooks

International Technical Support Organization

IBM @server Cluster 1600 and PSSP 3.4 Cluster Enhancements

December 2001

SG24-6604-00

Take Note! Before using this information and the product it supports, be sure to read the general information in “Special notices” on page 189.

First Edition (December 2001)

This edition applies to Parallel System Support Program Verison 3, Release 4 and LoadLeveler Version 3 Release 1 for use with the AIX Operating System Version 5 Release 1.

Comments may be addressed to:
IBM Corporation, International Technical Support Organization
Dept. JN9B Mail Station P099
2455 South Road
Poughkeepsie, NY 12601-5400

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

Tables

Preface

This redbook applies to IBM Parallel System Support Programs for AIX (PSSP) Version 3, Release 4 for use with the AIX operating system Version 5, Release 1 and Version 4, Release 3, modification 3.

This redbook details the new features and functions of PSSP 3.4, including supported hardware. We describe changes in the product to give cluster professionals a convenient and detailed look at the latest PSSP enhancements.

IBM has also announced the release of additional RS/6000 Scalable POWERparallel Systems (RS/6000 SP) related products. This redbook discusses the following:

- General Parallel File System (GPFS) 1.5
- LoadLeveler 3.1
- Parallel Environment (PE) 3.2
- Engineering and Scientific Subroutine Library (ESSL) 3.3 and Parallel ESSL 2.3

This redbook is for IBM customers, IBM Business Partners, IBM technical and marketing professionals, and anyone seeking an understanding of the new hardware and software components and improvements included in this IBM @server announcement.

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 (ITSO), Austin Center.

Dino Quintero is a project leader at the ITSO, Poughkeepsie Center. He has 10 years experience in the information technology field. He holds a BS in computer science and an MS degree in computer science/information systems from Marist College. Before joining the ITSO, he worked as a performance analyst for the Enterprise Systems Group, and as a disaster recovery architect for IBM Global Services. He has been with IBM since 1996. His areas of expertise include enterprise backup and recovery, disaster recovery planning and implementation, RS/6000 SP, and pSeries servers. Currently, he focuses on RS/6000 Cluster Technology by writing redbooks and teaching IBM classes worldwide.

JinHoon Lee is an AIX system support specialist for IBM Korea. He has worked with the RS/6000 and IBM @server pSeries post-sales support team since joining IBM 5 years ago. His areas of expertise include AIX, system performance tuning, High Availability Cluster Multi-Processor (HACMP), PSSP, and Tivoli Storage Manage (TSM).

Andrei Socoliuc is an IT specialist at IBM Global Services in Romania. He provides technical support for the RS/6000 and IBM @server pSeries. He is an IBM Certified Specialist in AIX and HACMP. His areas of expertise include AIX, HACMP, PSSP, and TSM. He holds an MS degree in computer science from the Politehnica University in Bucharest.

Tomaz Vincek is a senior IT specialist working at the IBM International Technical Center in Ljubljana, Slovenia. He is a team leader of the RS/6000 and IBM @server pSeries technical support group for Central and Eastern Europe/Middle East/Africa (CEMA). He is an RS/6000 SP expert and a member of the high-end business intelligence technical focus group for CEMA.

Team member photo (left to right):

(rear) Peter Zutenis (workshop team), Dino Quintero (project leader), Andrei Socoliuc (redbook team).

(front) JinHoon Lee (redbook team), Loretta Balerini (workshop team), Tomaz Vincek (redbook team).

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

International Technical Support Organization, Austin Center
Susan Schmerling, Scott Vetter, Wade Wallace

IBM Poughkeepsie
David Ayd, Mike Cavanaugh, Deborah Lawrence, Bernard King-Smith, Joan McComb, Mary Nisley, Shujun Zhou, Lissa Valletta, Brian Herr, Dave Delia, Duane Witherspoon, Bruno Bonetti, Stephen Hughes, Janet Elsworth, Gordon McPheeters, Bob Curran, Allison White, Dick Treumann, Bill Tuel, Robert Blackmore, Chris DeRobertis, Mike Coffey, Jim Dykman, Corky Norton, Paul

Swaitocha, Bruce Potter, Brian Croswell, William LaPera, Dave Wong

IBM Austin
Yuri L. VoloBuev

Special notice

This publication is intended to help IBM customers, IBM Business Partners, IBM sales professionals, IBM IT specialists, and the IBM Technical Support community in the proposing of RS/6000 cluster-based solutions. The information in this publication is not intended as the specification of any programming interfaces that are provided by RS/6000 hardware, AIX software, or PSSP software. See the PUBLICATIONS section of the IBM Programming Announcement for RS/6000 for more information about what publications are considered to be product documentation.

IBM trademarks

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

AFS®
AIX 5L™
DFS™
eLiza™
IBM ®
LoadLeveler®
Notes®
Perform™
PowerPC®
pSeries™
Redbooks Logo
SecureWay®
VisualAge®

AIX®
Chipkill™
e (logo)®
Enterprise Storage Server™
IBM.COM™
Netfinity®
Open Class®
POWERparallel®
PowerPC 604™
Redbooks
RS/6000®
SP™

Comments welcome

Your comments are important to us!

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

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

  ibm.com/redbooks

- Send your comments in an Internet note to:

  redbook@us.ibm.com

- Mail your comments to the address on page ii.

1

Overview

This chapter begins with a brief summary of the Parallel System Support Programs for AIX 5L (PSSP) announcement in Section 1.1, “Announcement summary” on page 2.

Section 1.2, “Cluster definition” on page 4 introduces the concept of clustering, the new IBM @server Cluster 1600, and the new 9076-556 frame.

Section 1.3, “Feature codes and naming conventions” on page 7 presents feature codes and naming conventions used throughout this redbook.

Section 1.4, “Announcement highlights” on page 8 presents highlights of the improvements announced for PSSP 3.4.

Section 1.5, “New in PSSP 3.4” on page 10 looks at new software and hardware enhancements.

Section 1.6, “New product releases” on page 16 discusses new PSSP-related product releases.

Section 1.7, “Software modifications in PSSP 3.4” on page 21 deals with software enhancements in PSSP 3.4.

Section 1.8, “Features removed” on page 23 covers PSSP software features removed from Version 3, Release 4.

Additional useful information is given in Section 1.9, “Software support notes” on page 23.

1.1 Announcement summary

The Parallel System Support Programs for AIX 5L (PSSP) software provides a comprehensive suite of applications for the installation, operation, management, and administration of the RS/6000 SP, attached servers, and Clustered Enterprise Servers (CES) from a single point of control. PSSP 3.4 concentrates on the connection to the SP Switch2 PC Interface (PCI) of RS/6000 servers S80, H80, M80 and pSeries 6H1, 6H0, 6M1 and p690 as SP-attached servers using the SP Switch2 PCI Attachment Adapter (F/C 8397). PSSP 3.4 supports Winterhawk1, Winterhawk2, and Silver SP nodes within the SP Switch2 PCI Attachment Adapter connection to the SP Switch2 PCI.

PSSP 3.4 support for the SP Switch2 PCI Attachment Adapter connected to the SP nodes and SP-attached servers required changes in:

- PSSP systems management/configuration support for a single-port adapter for the SP Switch2-PCI.
- PSSP systems management/configuration support for an SP-attached server connected to the SP Switch2-PCI.
- Connectivity subsystem (CSS) and switch communication subsystems changes to recognize the new node and switch combinations to invoke the correct data copy routines that are dependent upon specific hardware characteristics.

PSSP 3.4 allows the SP to attach logical-partition (LPAR) servers such as p690 LPAR. The basic approach is to extend the SP-attached server work to support an attached system as a “frame” with each LPAR AIX image represented as a “node.” Integrated hardware control is provided by an application programming interface (API) through the Hardware Management Console (HMC) exploited by a *hardmon* extension. Each LPAR AIX image runs a complete copy of PSSP, and all PSSP services are available. Most SP software does not need to be aware it is running in an LPAR. PSSP 3.4 supports a maximum of 16 LPARs per server.

The IBM @server Cluster 1600 extends and enhances IBM’s innovative and proven AIX and RS/6000 SP clustering technologies. The Cluster 1600, machine type 9078 Model 160, includes both the legacy SP system and new clusters made up of IBM @server pSeries servers.

The new PSSP release extends cluster management to include a broader range of cluster-ready servers. New attachment adapters for the SP Switch2 allow more pSeries servers to take advantage of PSSP high-performance, high-bandwidth server-to-server communications.

The benefits of PSSP 3.4 include the following:

- Cost-effective management of your IT infrastructure
- Anytime, anywhere access to business-critical data and applications
- IT infrastructure adaptability to meet your ever-changing business demands
- Extreme scalability (terabytes of data, billions of Web transactions, trillions of floating-point operations)
- Investment protection through the coexistence of old and new technologies

Existing RS/6000 SP customers can expand their systems with improved software and new SP Switch2 adapters. New customers can buy a cluster of mid-range servers, a high-end server, or a combination of the two. All Cluster 1600 servers (see Section 1.2.1, “The Cluster 1600” on page 4) are managed by the new PSSP 3.4, the same single-point-of-control software used by the RS/6000 SP system.

PSSP cluster configurations improve upon the current SP capabilities of high-performance server-to-server interconnect and processor performance. The improvements include the following:

- SP Switch2 two-plane configurations are designed to improve performance and reliability, availability, and serviceability (RAS) characteristics across the switch fabric.
- The SP Switch2 PCI Attachment Adapter allows the connection of legacy and new PCI-based servers to the SP Switch2. You can add more pSeries servers to take advantage of high-performance, high-bandwidth, server-to-server communications.
- PSSP supports up to two SP Switch2 PCI Attachment Adapters in each server for two-plane configurations.
- PSSP cluster configurations are supported with the SP Switch, the SP Switch2, or no switch.
- Rack-mount, multiple SP Switch2s within the new 9076 Model 556 frame improve machine-room efficiencies when more than one SP Switch2 is required for a cluster (refer to Section 1.2.2, “The new 9076-556 frame” on page 5).
- The SP Switch2 MX2 Adapter allows connection of 332 MHz symmetrical multiprocessor system (SMP) Thin and Wide Nodes, POWER3 SMP Thin and Wide Nodes, and 375 MHz POWER3 SMP Thin and Wide Nodes (single-plane) to the SP Switch2 fabric.

1.2 Cluster definition

A cluster is a collection of interconnected computers used as a unified computing resource. Usually two or more servers are managed from a single point of control.

There are three major cluster types:

- High performance
- High availability
- Horizontal scaling

Clustering, in a computing context, has become popular during the last five years for a number of reasons:

- Clusters offer scalability, which is essential to resolving computing problems.
- Consolidation, with a single point of control, has become essential to reduce management costs.
- Improved availability of resources is accomplished through sharing, replication, and redundancy within a cluster.

We discuss the new clustering definitions, the new hardware support announced, and more in the remainder of this redbook.

1.2.1 The Cluster 1600

The IBM @server Cluster 1600 extends and enhances IBM's innovative, proven AIX and RS/6000 SP clustering technologies. The Cluster 1600, machine type 9078-160, includes both the legacy SP system and new clusters made up of IBM @server pSeries servers. The Cluster 1600 introduces a single cluster serial number. Under the 9078-160, there are feature codes for each of the currently available pSeries or RS/6000 cluster servers or cluster switches:

- 7017-S85
- 7017-S80
- 7026-6M1
- 7026-6H1
- 7026-6H0
- 9076-550
- 9076-555
- 9076-556

New feature codes are enabled under the newly announced Cluster 1600 (9078-160), as shown in Table 1-1.

Table 1-1 9078-160 feature codes

MT	Model	Feature	Description
9078	160	0001	7017 type server
9078	160	0002	7026 type server
9078	160	0003	9076-555 type frame (SP Switch)
9078	160	0004	9076-556 type frame (SP Switch2)
9078	160	0005	9076-SP part of the cluster
9078	160	0006	9076 SP expansion frame part of the cluster
9078	160	0007	Control workstation
		8397	SP Switch2 PCI Attachment Adapter used in SP-attached server and switched CES (M/T 7017 and 7026)
		9941	Frame extender for 556 and 1556
		1556	Multiple SP Switch2 expansion frames

1.2.2 The new 9076-556 frame

The IBM RS/6000 SP 9076-556 allows you to connect clustered RS/6000 or pSeries servers to the SP Switch2 without the need for an SP internal node. Up to 32 RS/6000 or pSeries servers may be clustered in the switched environment. These environments add server interaction through the higher switched bandwidth. PSSP Version 3.4 supports switch-to-switch connectivity between the 9076-556 and legacy 9076-550 (across the SP Switch2).

The 9076 Model 556 is a 75-inch tall frame with 10.5 kW of power and a three-phase power system that includes N+1 power capabilities. SP Frame redundant power input features are available. A separate order of the SP Switch2 is required for the frame. You may place additional SP Switch2 orders for a total of up to four switches, to support up to 32 servers in a two-plane SP Switch2 environment.

The 9076-556 supports multiple SP Switch2 node switch boards within a single no-node SP frame for the application of switched Clustered Enterprise Servers, in both single-plane and two-plane configurations. The server attachment adapter is the SP Switch2 PCI Attachment Adapter. The initial support requirement within Model 556 is one to four SP Switch2 node switch boards, which supports up to 32 servers in a two-plane application (identified by specifying code 9977). You can connect the 9076-556 to the 9076-550's SP Switch2 fabric through switch-to-switch connections.

Important:

- Only Network Switch Boards (NSBs) can be placed in these frames. The frames cannot include SP nodes or I/O drawers.
- 4.3.3 support is included to aid migration. We do not recommend it as a long-term solution.

The switch plane cabling for two planes in the Model 556 is interleaved. The first switch plane (switch plane 0) is located in the first drawer. The second switch plane (switch plane 1) is located in the second drawer.

The SP Switch2 interposers, blank interposers, and cables are driven by the configurator. You use an F/C 9302 (Switch-to-Node SP Switch cable) as a Switch-to-Switch cable connection within the multiple SP Switch2 node switch board frame.

Attention: The entire system (nodes and Control Workstation) must be at PSSP 3.4.

IBM introduced the 9076 Model 555 frame in April 2001 as a no-node SP system/frame for use in switched Clustered Enterprise Servers, with the SP Switch attached to the TB3PCI cards in the server. A second SP Switch to support more than 16 Clustered Enterprise Servers requires an expansion frame.

The new 9076-556 frame allows you to connect clustered pSeries and RS/6000 servers to the SP Switch2. When required, multiple SP Switch2s can be rack-mounted within the 9076-556 frame to improve machine-room efficiencies.

Attention: The 9076-556 requires a minimum of two clustered RS/6000 or pSeries servers and a supported Control Workstation.

New feature codes are enabled for the IBM RS/6000 SP 9076, as shown in Table 1-2.

Table 1-2 New RS/6000 SP 9076 feature codes

MT	Model	Feature	Description
9076	556		SP Switch2 frame

MT	Model	Feature	Description
9076	556	9125	7026 attachment cable
9076	556	9977	Two-plane support for SP Switch2

1.3 Feature codes and naming conventions

Table 1-3 shows the new hardware and software feature codes and naming conventions used throughout this redbook.

Table 1-3 Software and hardware feature codes and descriptions

Feature code	Description
9435	AIX Version 4.3.3 with PSSP 3.4
9534	AIX 5L with PSSP 3.4
2033	Two-plane support for F/C 2032
4012	SP Switch2
4025	SP Switch2 adapter
4026	SP Switch2 MX2 adapter
8397	SP Switch2 PCI Attachment Adapter
6203	Genie Ultra3 SCSI adapter
4957	Lumbee-F Refresh asynchronous transfer mode (ATM) adapter
4953	Lumbee-U Refresh ATM adapter
4962	Scurry 10/100 Ethernet adapter
4963	Stonewall Cryptographic Coprocessor (FIPCS-4) adapter

For more information on the latest software and hardware features and descriptions, refer to *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment,* GA22-7280, and *IBM RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment,* SG22-7281.

1.4 Announcement highlights

The RS/6000 SP system delivers solutions to some of the most large and complex technical and commercial applications by simultaneously bringing dozens of RISC processing nodes to a computing problem. This parallel processing capability enhances computing performance and throughput many times over serial computing.

PSSP cluster configurations continue to expand and extend present SP interconnect and processor performance capabilities to clusters, with the following improvements:

- Software support for AIX 5L Version 5.1.
  - PSSP 3.4 and AIX 5L Version 5.1.
  - PSSP 3.4 and AIX Version 4.3.3.
- Secure Shell (SSH) authentication and encryption.
  - Allows you to choose any secure remote command software conforming to the Internet Engineering Task Force (IETF) SSH protocol.
  - Rapidly replacing older, less secure tools like Telnet.
- Firewall support within the SP.
- Perl 5 support (conversion to AIX Perl).
- Cluster support for all p690 configuration options.
  - Standard (up to 32-way).
  - Turbo (up to 32-way).
  - High Performance Computing (HPC) up to 16-way.
  - Central management point extended to include logical partition (LPAR) support.
  - Ability to add p690 to existing clusters of legacy servers.
  - PSSP 3.4 support of 96 GB of memory per LPAR.
  - Industry-standard communication support (Fast Ethernet, Gigabit Ethernet, and so on.)
- p690 cluster interconnect support for the SP Switch2.
  - Initial scaling limit of 16 p690 servers.
  - Up to eight SP Switch2-connected LPARs per server with optional dual plane support.
  - Ability to mix SP Switch2- and LAN-connected LPARs.
  - Up to eight Network Switch Boards (NSBs) per 9076-556 frame.

- p690 cluster interconnect support for the SP Switch.
  - Up to eight SP Switch-connected LPARs per server (single plane only).
  - Mandatory SP Switch connection for all LPARs.
  - One NSB per 9076-555 frame.
- The SP Switch2 PCI Attachment Adapter (F/C 8397) allows you to connect both legacy and new PCI-based servers to the SP Switch2. You can add more pSeries servers to take advantage of high-performance, high-bandwidth, server-to-server communications.
  - p680, p660 (6M1, 6H1, and 6H0).
  - Legacy RS/6000 S70, S7A, S80, M80, and H80.
  - pSeries 690 (1H/2002).
- SP Switch2 two-plane configurations are designed to improve performance and RAS characteristics across the switch fabric.
  - Up to two adapters per node, server, or LPAR (each connected to a separate SP Switch2 plane).
  - Increases bandwidth per node, server, or LPAR.
  - Improves reliability through node and switch plane redundancy.
- Support for PSSP cluster configurations with the SP Switch, the SP Switch2, or no switch.
- Support for additional configuration and expansion options as a result of the new software's ability to also run on the AIX 5L Version 5.1 for POWER Version 5.1 operating system.
- The ability to rack-mount multiple SP Switch2s within the new 9076-556 frame, to improve machine-room efficiencies when a cluster requires more than one SP Switch2.
- The ability provided by the SP Switch2 MX2 Adapter to connect 332 MHz SMP Thin and Wide Nodes, POWER3 SMP Thin and Wide Nodes, and 375 MHz POWER3 SMP Thin and Wide Nodes (single-plane) to the SP Switch2 fabric.
- The ability to mix switched and non-switched nodes in an SP Switch2 PCI-switched SP system (nodes on and off the Switch).
- Increased flexibility due to the SP Switch2 easing the:
  - Placement of attached servers.
  - Placement of SP nodes.
  - Physical connections to the switch.
- Multiport Ethernet card adapter support.

- Enhanced Fibre Channel SAN support.
  - AIX install or boot from Fibre Channel SAN disks.
  - Virtual Shared Disk (VSD) applications (such as the General Parallel File System (GPFS) or Oracle Parallel Server) now work with IBM Enterprise Storage Server.
- Enhancements to parallel message passing support for HPC.
  - Low-level Application Programming Interface (LAPI) performance enhancements.
  - 64-bit Message Passing Interface (MPI), LAPI, and tools.
  - LAPI-based message passing over shared memory.
- New product releases.
  - GPFS 1.5.
  - LoadLeveler 3.1.
  - Parallel Environment 3.2.
  - Engineering and Scientific Subroutine Library (ESSL) 3.3 and Parallel ESSL 2.3.

PSSP cluster configurations continue to expand and evolve with the following additional software releases planned:

- PSSP cluster support for pSeries 690, first half of 2002
- High Availability Cluster Multi-Processor (HACMP) 4.4.1 support on pSeries 690, first quarter of 2002
- HACMP 4.4.1 support for 64-bit Kernel, first quarter of 2002
- LoadLeveler Checkpoint/Restart, first half of 2002

1.5 New in PSSP 3.4

PSSP 3.4 is a new version of the Parallel System Support Programs that run on RS/6000 SP and pSeries clusters. The pSeries clusters can be attached, clustered together without an SP frame, and run in either a switched or switchless environment.

1.5.1 New hardware support

The RS/6000 SP includes several improvements in hardware components, such as support for pSeries 690-attached servers and support for additional PCI adapters.

Multiple SP Switch2 node switch board within a single frame

Refer to Section 1.2.2, “The new 9076-556 frame” on page 5 for more information on multiple SP Switch2 switch boards within a single frame.

The SP Switch2 MX2 Adapter

SP Switch2 attachment applications are now enabled for the following SP nodes: Silver (thin and wide), Winterhawk-1 (thin and wide), and Winterhawk-2 (thin and wide). The SP Switch2 MX2 Adapter, also known as the Corsair-MX Adapter, plugs into the MX slot within the PCI-based thin and wide nodes to enable attachment to the SP Switch2 fabric.

> **Attention:** Support for the SP Switch2 MX2 Adapter is for single-plane SP Switch2 applications only.

The SP Switch2 PCI Attachment Adapter

SP Switch2 attachment applications are now enabled for SP-attached servers and switched Clustered Enterprise Servers supported through the 9076-556 frame. This is accomplished through the use of the SP Switch2 PCI Attachment Adapter F/C 8397.

> **Attention:** PSSP 3.4 is required software support for the SP Switch2 PCI Attachment Adapter.

Regatta clusters

PSSP 3.4 allows the SP to attach logical-partition (LPAR) servers such as the Regatta-H LPAR. The attached cluster has the following characteristics:

- It extends the SP-attached server to support an attached system as “frame,” with each LPAR AIX image represented as a “node.”
- An API provides integrated hardware control through the Hardware Management Console (HMC) exploited by the hardmon extension.
- Each LPAR can have a PCI switch adapter (two for SP Switch2) for switch connectivity.
- Each LPAR AIX image runs a complete copy of PSSP, and all PSSP services are available. Most SP software does not need to be aware it is running in an LPAR.
- PSSP 3.4 supports a maximum of 16 LPARs per server.

The SP-attached server 7026-6M1

The 7026-6M1, the successor to the 7026-M80, is supported as an SP-attached server as well as a Clustered Enterprise Server. Both the TB3PCI SP Attachment Adapter and the SP Switch2 Attachment Adapter are supported.

PCI I/O adapters

The SP system supports the following PCI adapter feature codes:

- Genie Ultra3 SCSI adapter, F/C 6203
- Lumbee-F Refresh asynchronous transfer mode (ATM) adapter, F/C 4957
- Lumbee-U Refresh ATM adapter, F/C 4962
- Scurry 10/100 Ethernet adapter, F/C 4962
- Stonewall Cryptographic Coprocessor (FIPS-4) adapter, F/C 4963

For more information about the latest hardware support enhancements with PSSP 3.4, refer to Chapter 2, “Hardware support” on page 25.

1.5.2 Hardware migration scenarios

PSSP 3.4 supports the following hardware migration scenarios:

- SP Switch-to-SP Switch2 upgrades are supported in the new 9076-556 frame in contrast to the 9076-555 frame.
- Additional SP Switch2 switches may be added for a total of four switches within the 9076-556 frame.
- Additions of the SP System Attachment Adapter2 (Corsair, F/C 8397) within the servers are enabled.
- Upgrades from the SP System Attachment Adapter (TB3PCI, F/C 8396) to the SP System Attachment Adapter2 (Corsair, F/C 8397) are enabled.
- To support the SP Switch2 two-plane application, a second set of four SP Switch2 switches is assigned an F/C of 2033 to support up to 128 nodes within a single switch frame (F/C 2032) without the need for the request for price quotation (RPQ) 8P2109.
- Upgrades from F/C 4022 (TB3 MX) and F/C 4023 (TB3 MX2) to F/C 4026 (Corsair MX) are enabled.
- Additions of the Corsair-MX card within Silver/Winterhawk-1/Winterhawk-2 thin and wide nodes are enabled.

1.5.3 SP Switch2 improvements

The new PSSP release includes the following enhancements to the switch system management software:

- The switch_plane class has attributes similar to switch_partition (one object per plane).
- The switch_adapter_port class has connection information on an adapter-per-port basis.
- The switch class has switch_plane and switch_plane_seq attributes.
- The adapter class now contains an adapter_status attribute.
- The switch_responds class has vectorized attributes for switch_responds isolated.
- A new **spswplane** command has been added to specify the number of switch planes in use.
- Emaster displays the Master Switch Sequencing node (MSS) for an SP Switch2.
- Ecommands can be extended with a –p when you need plane actions. For example, the following command starts plane 1:

```
Estart –p 1
```

Attention: If the –p parameter is not specified, the default is to perform the operation for all valid switch planes.

SP Switch2 RAS concepts

A switch adapter or a port failure results in the loss of node connectivity to the switch. The SP Switch2 boasts the following improvements in reliability, availability, and serviceability (RAS):

- A two switch-plane SP Switch2 system provides redundant node-to-node connectivity.
  - Each node contains two separate and independent switch adapters.
  - Each adapter contains one switch port.
  - Each switch port connects to a separate and independent switch-plane.
  - Each switch-plane provides multiple paths connecting any two nodes.

Notes:

- Local adapter error recovery for the SP Switch2 adapter is performed by the fault service daemon.
- Local adapter error recovery for the SP Switch2 PCI adapter is performed by a new daemon, the local adapter error recovery daemon la_event_d.

Aggregate IP addressing

In addition to the RAS improvements described in “SP Switch2 RAS concepts” on page 13, a new aggregate IP addressing concept has been introduced. Aggregate IP addressing has the following characteristics:

- The addition of a third IP address, which aggregates the addressing of the two SP Switch2 adapters on a single node.
- The aggregate IP address is optional.
- The aggregate IP address must be explicitly assigned by the system administrator.
- The System Data Repository (SDR) class, Common Messaging Interface (CMI) commands, and System Management Interface Tool (SMIT) panels are supported.
- A third pseudo-IP driver multilinks and aggregates the two real IP devices.

Functional characteristics of the SP Switch2

The SP Switch2 provides the system administrator new logging characteristics as well as new administrative features. The following list highlights the functional changes:

- More unified logging. For example, the fs_daemon_print.file captures all information including what was in the work_print.file.
- Time stamp on all logs.
- Node-level files.
  - /var/adm/SPlogs/css.
- Adapter-level files.
  - /var/adm/SPlogs/css0.
  - /var/adm/SPlogs/css1.

- Port-level files.
  - /var/adm/SPlogs/css0/p0.
  - /var/adm/SPlogs/css1/p0.

  > **Attention:** Most common debug files (for example, flt, cable_miswire, and so on) are in the port-level directories.

- Easy-to-use log files.
  - adapter.log or la_event_d log located in /var/adm/SPlogs/cssX.
  - flt and fs_daemon_print.file located in /var/adm/SPlogs/cssX/p0.
- Less convenient log files for the switch.
  - The switch number in the out.top and log files is the switch_plane_seq number.
  - s 15 has various meanings:
    - In out.top, it means switch1, chip 5.
    - On plane 0, it means switch 1 in frame 1.
    - On plane 1, it means switch 2 in frame 2 (switch_plane_seq in the Switch object is 1).

    > **Important:** Make sure you are looking at the plane you want to work with.

1.5.4 Security

In PSSP 3.4 you remove root-level PSSP dependencies on root using **rsh**/**rcp** from the Control Workstation (CWS) to the node. The PSSP code allows you to use a secure remote command method, such as **ssh**/**scp** (Secure Shell) as an alternative to **rsh**/**rcp** (AIX remote shell), for copying files with executing commands on remote machines.

Attention: For `ssh` to be enabled, all nodes must be running PSSP 3.2 or later, the CWS must be at PSSP 3.4, and Remote Root Access (RRA) must be enabled.

In addition, you have the ability to select "none" for AIX remote command authorization methods on a partition. This prevents PSSP from automatically generating entries for root in the authorization files for the nodes in that partition, while preserving the ability of root on the CWS to issue remote shell commands to the nodes.

Attention: To enable the AIX Authorization Remote Commands Methods = none selection, all nodes in the partition must be installed with PSSP 3.4, and `ssh` must be enabled.

For more information on the latest security features of PSSP 3.4, refer to Chapter 5, "Security enhancements" on page 101.

1.5.5 Migration and coexistence

PSSP 3.4 has the following migration and coexistence requirements:

- PSSP 3.4 requires AIX Version 4.3.3.75 or AIX 5L Version 5.1.0.10 or later.
- The CWS must be running at the highest level of PSSP and AIX.
- PSSP 3.4 can coexist with previous levels of PSSP in the same partition (PSSP Versions 3.2 and 3.1.1).
- PSSP 2.4 is only supported for migration purposes.
- Migration is supported in the following cases:
  - From PSSP 2.4 on AIX 4.2.1 or AIX Version 4.3.3
  - From PSSP 3.1.1 on AIX Version 4.3.3
  - From PSSP 3.2 on AIX Version 4.3.3

For more information on the latest PSSP 3.4 migration and coexistence requirements, refer to Chapter 4, "Installation, migration, and coexistence" on page 77.

1.6 New product releases

Along with PSSP 3.4, new releases for several licensed products have been announced.

1.6.1 The General Parallel File System (GPFS) 1.5

The General Parallel File System (GPFS) provides file system services to parallel and serial applications running in the AIX operating environment. For scientific and technical computing, GPFS provides a single global file system for the SP/Cluster system and is ideal for high-performance parallel file transfer and parallel I/O to single or multiple files into and out of the SP. For business intelligence (BI) usage, GPFS provides parallel enablement for applications such as Statistical Analysis System (SAS), improves performance of database-loading utilities, and supports parallel online analysis processing using tools that are not themselves enabled for parallel use. The key features of this release for users of the applications we have mentioned are the following:

- GPFS supports Fibre Channel (FC)-attached disks; storage area networks (SANs) extend GPFS cluster support to FC attachment of disks with switches.
- GPFS supports 100 TB file systems.
- GPFS supports default quotas for file systems.
- GPFS provides FC disk support in HACMP cluster environments.
- GPFS supports AIX 5L Version 5.1AIX 5L Version 5.1.
- GPFS supports SP Switch2 (double switch planes).
- GPFS provides a new **`mmcrlv`** command for cluster environments.

For more information, refer to the GPFS Web site:

`http://www.ibm.com/servers/eserver/clusters/software/`

This Web site includes links for both AIX and Linux versions of GPFS.

Online GPFS documentation can be found at:

`http://www.rs6000.ibm.com/resource/aix_resource/sp_books/gpfs/index.html`

For more information on the latest GPFS enhancements, refer to Chapter 8, “GPFS 1.5” on page 155.

1.6.2 LoadLeveler 3.1

LoadLeveler is a key tool for scheduling scientific and technical applications to a pool of SP nodes and/or clusters of RS/6000 workstations and servers. LoadLeveler is ideal for users who do not need to submit jobs according to some preset schedule. Other job schedulers providing rule-based batch workload management for inter-job dependencies and predetermined job scheduling can be integrated with the LoadLeveler functionality. LoadLeveler supports load balancing for interactive Parallel Operating Environment sessions, and focuses on parallel job scheduling and scalability.

The key enhancements of the new LoadLeveler 3.1 for AIX 5L Version 5.1 for users of these applications are in capability and scaling. This enhanced functionality includes:

- System-initiated parallel Checkpoint/Restart

  Enhances the checkpoint and restart facilities in LoadLeveler and the Parallel Environment (PE). Both LoadLeveler and PE exploit the new kernel-level checkpoint functionality of the base AIX operating system.

- Support for Gang Scheduling

  Provides a parallel job-scheduling algorithm allowing tightly synchronized parallel applications to run and share resources on the same set of nodes.

- 64-bit application enablement

  Provides full 64-bit support for interactive and batch jobs. LoadLeveler 3.1 has been enhanced so that users and administrators can specify and request enforcement of:

  - The large 64-bit system resource limits available with AIX 5L Version 5.1 or higher
  - History files with 64-bit statistics on completed jobs

  Checkpoint batch applications running under LoadLeveler can be 64-bit applications, as can user applications linked to the LoadLeveler API library.

- Integration of LoadLeveler and AIX WorkLoad Manager (WLM)

  LoadLeveler manages resource limits such as CPU, physical memory, and paging I/O on an individual basis using AIX job class. The enhanced LoadLeveler provides both consistent resource management control across all nodes (for example, an equivalent bind-processor() function with more than one processor), and a fair-share function across clusters with AIX WLM capability.

For more information on the latest LoadLeveler enhancements, refer to Chapter 7, “LoadLeveler 3.1” on page 137.

1.6.3 The Parallel Environment (PE) 3.2

The Parallel Environment for AIX (PE) Version 5, Release 1 is a highly functional development and execution environment for parallel applications using either the RS/6000 SP system or one or more RS/6000 processors. PE is widely used for Fortran and C/C++ parallel application development, as well as execution of scientific and technical applications. The key enhancements of this release for users of these applications are the following:

- Message Passing Interface (MPI) I/O performance and GPFS optimization.

- MPI-2 external interfaces and enhancements. The semantics of all MPI-2 functions is discussed in detail in the MPI-2 standard.
- 64-bit support for MPI, LAPI, and tools.

  Most numerically intensive computing applications are memory-intensive. The cost of communication is high relative to the cost of local memory access.
- LAPI performance enhancements and LAPI shared memory.
  - Specific customer concerns have been addressed in the adapting of LAPI to a changing SP architectural paradigm.
  - Future concerns have been addressed in Dynamic Probe Class Library (DPCL) enhancements including the following:
    - Shared library support
    - Elimination of prelink steps
    - New interfaces required by the PE performance tools developed on top of DPCL
- The PE benchmark tools.

  The application performance tools are a set of DPCL-based tools consisting of the following components:
  - The performance data collection tool

    Allows users to collect MPI, system, and user event traces, or operating system and hardware performance counters, through a Java-based GUI or text-based command line. All data collected is from an application perspective and is dynamically controlled by the user.
  - Trace analysis utilities

    A set of utilities to gather and analyze trace statistics and to convert and/or merge trace files into a scalable file format that can be viewed by third-party tools.
  - The profile visualization tool (for operating-system and hardware-counter files only)

    A Java-based GUI and text-based command line application that allows users to display and analyze operating system and hardware counter statistics on a per-function and per-thread basis, through visual histograms and statistical reports.

For more information on the latest PE enhancements, refer to Chapter 6, “Parallel programming” on page 115.

1.6.4 ESSL 3.3 and Parallel ESSL 2.3

The Engineering and Scientific Subroutine Library (ESSL) is a collection of mathematical subroutines useful for many different scientific and technical computing applications. ESSL provides a thread-safe serial library and a symmetrical multiprocessor system (SMP) library that support 32-bit and 64-bit applications. ESSL for AIX 5L Version 5.1, Version 3.3 includes POWER4 tuning, new LAPACK subroutines, and support for the AIX 5L Version 5.1 32-bit and 64-bit kernels.

Parallel ESSL is a collection of mathematical subroutines useful for many different scientific and technical computing message-passing applications. Parallel ESSL provides a serial library for use with the PE MPI Signals Library and an SMP library for use with the PE MPI threaded library. Parallel ESSL for AIX 5L Version 5.1, Version 2.3 provides 64-bit application support in the SMP library and new ScaLAPACK subroutines.

> **Note:** Parallel Environment, LoadLeveler, ESSL, and Parallel ESSL support AIX 5L Version 5.1 only. AIX 4.3 support is not provided in the new levels of these licensed products.

For more information, refer to the ESSL and Parallel ESSL home page at:

http://www.ibm.com/servers/eserver/pseries/software/sp/essl.html

You can find online documentation at:

http://www.ibm.com/servers/eserver/pseries/library

For more information on ESSL and Parallel ESSL enhancements, refer to Section 6.5, “Parallel ESSL 2.3 and ESSL 3.3” on page 132.

1.6.5 The High Availability Cluster Multi-Processor for AIX (HACMP)

HACMP is designed to give your application availability by detecting system or network failures and managing failover to a recovery processor with a minimal loss of end-user time. HACMP makes use of redundant hardware in the cluster to keep your application running, restarting it on a backup processor if necessary. HACMP clusters can be configured to meet complicated application availability and recovery needs, maximizing application throughput and investments in hardware and software. HACMP 4.4.1 adds support for pSeries 690 and support of the AIX 5L Version 5.1 64-bit kernel.

The HACMP 4.4.1 enhancements are as follows:

- Enhanced failover support, including the ability to handle combinations of multiple failures in certain circumstances

- Improved usability, including an enhanced SMIT interface for configuring volume groups and networks, customizable pager notifications, and RS/6000 High Availability Geographic Cluster system (HAGEO) site configuration data
- Use of single point of control (C-SPOC) to replace a failed disk
- Expanded device support, including multiple logical interfaces on the same ATM network adapter, original equipment manufacturer (OEM) disk API, and 32 nodes in concurrent access mode volume groups
- Replacement of hot-plug-capable PCI network adapters
- Demonstration code (sample scripts) for new highly available network service

For more information on HACMP enhancements, refer to Section 3.7, “HACMP” on page 73.

1.7 Software modifications in PSSP 3.4

The following sections describe the software modifications announced with PSSP 3.4.

1.7.1 Virtual Shared Disk 3.4

Virtual Share Disk is the software that enables nodes in the RS/6000 SP to share disks with another nodes in the same system partition. The following enhancements are supported with Virtual Shared Disk (VSD) 3.4.

Subsystem Device Driver (SDD) support

The SDD is an Enterprise Storage Server (ESS) device driver that provides:

- Enhanced data availability
- Automatic path failover and recovery to an alternate path
- Dynamic load balance of multiple paths
- Concurrent microcode upgrade

Attention: The *IBM Subsystem Device Driver: Installation and User’s Guide,* SC26-7425, is required reading before any attempt to make use of VSD SDD support.

When redundant paths are configured to ESS logical units and SDD is installed and configured, the AIX **lspv** command displays the multiple hard disks (hdisks) as well as a new construct called a *vpath*. To view further information, use the **lsvpcfg** command.

The IBM Virtual Shared Disk subsystem supports virtual shared disks defined in SDD volume groups, which can be referred to as *vpath volume groups*. To exploit the SDD functionality (including automatic path failover), the shared disk volume groups must be created as or converted to vpath volume groups. There are three ways to configure VSDs using SDD:

- To convert an existing hdisk volume group to a vpath volume group you can use the **dpovgfix volgrp** command or the **hd2vp volgrp** command. This removes the physical volume identifier (PVID) from the existing hdisk paths and creates the vpath volume group.
- The **createvsd** and **createhsd** commands (this includes creating shared disks via perspectives) have been modified to accept hdisks or vpaths. hdisks and vpaths cannot be specified on the same invocation. If you create the shared disks from hdisks as opposed to vpaths, you need to convert the volume groups to vpath volume groups as specified in the previous step. Refer to *PSSP for AIX: Managing Shared Disks*, SA22-7349 for more information on the creation and management of shared disks.
- Perform the following steps:
  a. Use the SDD command **mkvg4vp** to build volume groups.
  b. Use the AIX Logical Volume Manager (LVM) commands to build logical volumes.
  c. Use VSD commands to associate the LVM constructs with VSD constructs.

Note: If SDD volume groups (vpaths) are used, all nodes that access those volume groups must access them as vpath volume groups. For example, you must not have one node accessing a normal volume group and another node accessing the same volume group as a vpath volume group.

Aggregate Adapter support

The vsdnode (**updatevsdnode**) command specifies which communication adapter to use.

- All nodes must use the same adapter in a fully connected network (all nodes must be able to send messages directly to each other). The only adapter names supported are css0 and ml0.
- ml0 provides data striping and recovery over multiple adapters. ml0 can be specified when the nodes support two switch adapters and a multilink IP address has been configured.

For more information on the latest VSD enhancements with PSSP 3.4, refer to Section 3.6, “IBM Virtual Shared Disk 3.4” on page 60.

1.8 Features removed

The following functionalities have been removed from PSSP 3.4:

- SP Taskguides (ssp.tguides) have been removed.
- ssp.jm was already removed in PSSP 3.1.1. Its major executable is removed during node migration from Version 2.4, but manually removed during CWS migration from Version 2.4.
- The Performance Toolbox Parallel Extension (PTPE) was not supported in PSSP 3.2 with Distributed Computing Environment (DCE) security, and the PTPE perspectives GUI (ssp.ptpegui) has now been removed in 3.4.
- In PSSP 3.2, the trace visualization portion of the visualization tools (VT) was withdrawn from Parallel Environment (PE) in order to maintain a migration path from VT to follow-on tools that were in development, but the trace collection macros embedded within the MPI library were left intact. This remaining portion of VT has been removed from PE in PSSP 3.4.
- The GUI-based parallel debugger Parallel Environment Debugger (PEDB) is no longer shipped as part of PE.
- Netfinity Processor Extension nodes have been removed.

1.9 Software support notes

According to the software support timetable, the following support ends as indicated below:

- Support for PSSP 2.4 on AIX 4.2.1 or AIX Version 4.3.3 ends December 2001.
- Support for PSSP 3.1.1 on AIX Version 4.3.3 ends December 2002.
- Support for PSSP 3.2 on AIX Version 4.3.3 ends December 2002.

2

Hardware support

This chapter describes the following:

- RS/6000 SP Switch2 enhancements
- IBM @server p690 model 681 clusters
- SP-attached server 7026-6M1
- IBM @server Cluster 1600
- PC Interface (PCI) I/O adapters
- Reliability, availability, and serviceability (RAS) enhancements

2.1 Enhancements to the RS/6000 SP Switch2

The SP Switch2 was introduced with the PSSP 3.2 announcement (for details see the IBM Redbook *PSSP 3.2: RS/6000 SP Software Enhancements,* SG24-5673). This section describes the new features of SP Switch2 hardware, which is supported only in PSSP 3.2 or higher.

The SP Switch2 has the following enhancements:

- Two-plane switch fabric support (software support begins with PSSP 3.4).
- New connectivity features (see Section 2.1.4, “The SP Switch2 MX2 and SP Switch2 PCI adapters” on page 29, Section 2.2, “p690 clusters” on page 35 and Section 2.3, “The SP-attached server 7026-6M1” on page 40). The software support begins with PSSP 3.4.
- Switch flexibility. Nodes or servers in the same cluster can be either on or off SP Switch2 (see “Switched and non-switch nodes” on page 48).
- Node placement flexibility. The SP Switch2 has no restrictions on node placement—a node can be placed anywhere the physical constraints allow (see Chapter 3, “Node placement with the SP Switch2”, in *IBM RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment,* SG22-7281).
- Automatic switch port number assignment. The switch port numbers for an SP Switch2 system are automatically assigned sequentially by the PSSP 3.4 communication subsystems support component (CSS). As a node is assigned a CSS adapter it is given the lowest available switch node number from 0 through 511. There is no correlation between the switch port number and any hardware connections (see Chapter 3, "Node placement with the SP Switch2", in *IBM RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment,* SG22-7281).
- Packaging (see Section 2.1.2, “SP Switch2 configurations” on page 27).

2.1.1 Two-plane option

With the release of PSSP 3.4, SP Switch2 (SPS2) usage is extended to include two-plane switch fabric support.

You create a two-plane application by adding a second switch adapter to the nodes and a second switch (including interposers, blank interposers, and cables). Each plane has the same functionality and characteristics as a single SP Switch2 plane (see Figure 2-1 on page 27). The addition of a second switch plane is specified by the feature code 9977, with which you order two switch planes:

- Switch plane 0
- Switch plane 1

The SP Switch2 two-plane option increases reliability through node and plane redundancy. This redundancy is accomplished by having two adapters per node, server, or logical partition (LPAR), each connected to a separate SP Switch2 plane.

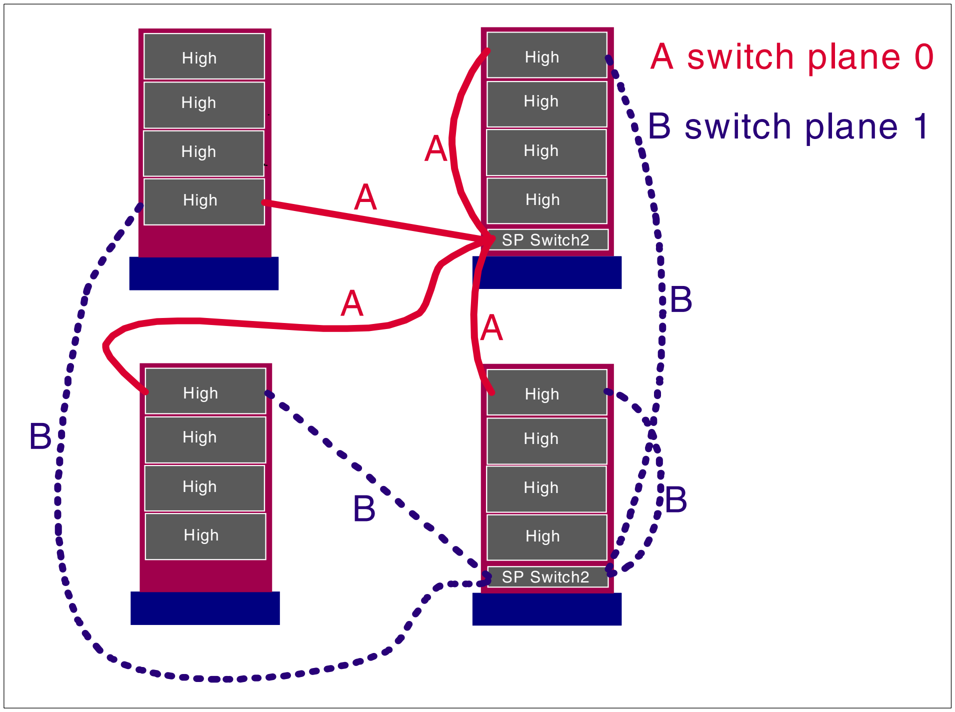


Figure 2-1 Two-plane application

2.1.2 SP Switch2 configurations

Two configurations are supported for an SP system with the SP Switch2 (SPS2):

- SP Switch2 Single-Single configuration, which is supported for PSSP 3.2 or later.
  - Each node has a single SPS2 adapter installed.
- SP Switch2 Double-Single configuration, which is supported for PSSP 3.4 or later.
  - A node has two SPS2 adapters installed and each is installed to a different SPS2 switch fabric.

  - We recommend that you connect ports on the same node to the port with the same number on each switch board, to avoid port-number confusion if you use the Double-Single configuration. For example, the logical adapter name css0 should be connected to plane 0.

> **Note:** In a Double-Single configuration a node should be connected to both SPS2 switch planes or to neither one.

For more information about SP Switch2 configurations, refer to the following publications:

- IBM Redbook *PSSP 3.2: RS/6000 SP Software Enhancements,* SG24-5673
- *IBM RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment,* SG22-7281
- *PSSP for AIX: Administration Guide,* SA22-7348

> **Note:** A Double-Double configuration, as described in the IBM Redbook *PSSP 3.2: RS/6000 SP Software Enhancements,* SG24-5673, is not supported.

2.1.3 SP Switch2 deployment considerations

The following considerations must be taken into account before deployment of a cluster with an SP Switch2:

- PSSP 3.4 is required with AIX 5L Version 5.1 or AIX Version 4.3.3 aids migration but is not recommended as a long-term solution.
- The two-plane option is not supported for either thin or wide nodes.
- Switch plane 0 and switch plane 1 cannot be connected.
- You cannot deploy an SP Switch2-SP Switch mix or an SP Switch2-High Performance Switch (HPS) mix.
- You cannot connect the SPS2 to the SP Switch Router.
- The greatest switch-to-switch cable length is 20 m.
- The maximum number of SP-attached servers (or nodes in a Clustered Enterprise Server (CES)) is 32.
  - A maximum of 16 servers using the SAMI protocol (7017-S80, IBM @server p680) is allowed.
  - A maximum of 32 servers using the Common Service Processor (CSP) protocol (IBM @server p660-6M1, IBM @server- 6H1 and -6H0, 7026-H80, 7026-M80) is allowed.

  - A maximum of 32 combined Service and Manufacturing Interface (SAMI) and CSP servers is allowed.

> **Note:** The 7017-S70 server is not supported on an SP Switch2.

- In a two-plane configuration a node, an SP-attached server, or a CES must be connected to both planes or to neither one.

2.1.4 The SP Switch2 MX2 and SP Switch2 PCI adapters

Until now, only Nighthawk nodes (222 MHz and 375 MHz POWER3 High Nodes) are supported on the SP Switch2. Two new adapters have been released to allow you to take advantage of the high-performance, high-bandwidth, server-to-server communication of the SP Switch2 with both legacy servers and PCI-based servers (refer to Figure 2-4 on page 32 for more details):

- The SP Switch2 MX2 Adapter (F/C 4026).
- The SP Switch2 PCI Adapter (F/C 8397).

The SP Switch2 MX2 adapter (F/C 4026)

The SP Switch2 MX2 adapter supports Single-Single applications only, and enables connectivity of the following:

- Silver nodes (332 Mhz SMP Thin and Wide Nodes)
- Winterhawk-1 nodes (200 MHz POWER3 SMP Thin and Wide Nodes)
- Winterhawk-2 nodes (375 MHz POWER3 SMP Thin and Wide Nodes)

The adapter plugs into the MX slot within the PCI-based thin/wide nodes (see Figure 2-2 on page 30 and Figure 2-3 on page 31 to locate the MX bus).

Note: You may not be able to locate the bus by viewing the following schematic diagrams. However, the customer engineer (CE) performs the addition of the adapter.

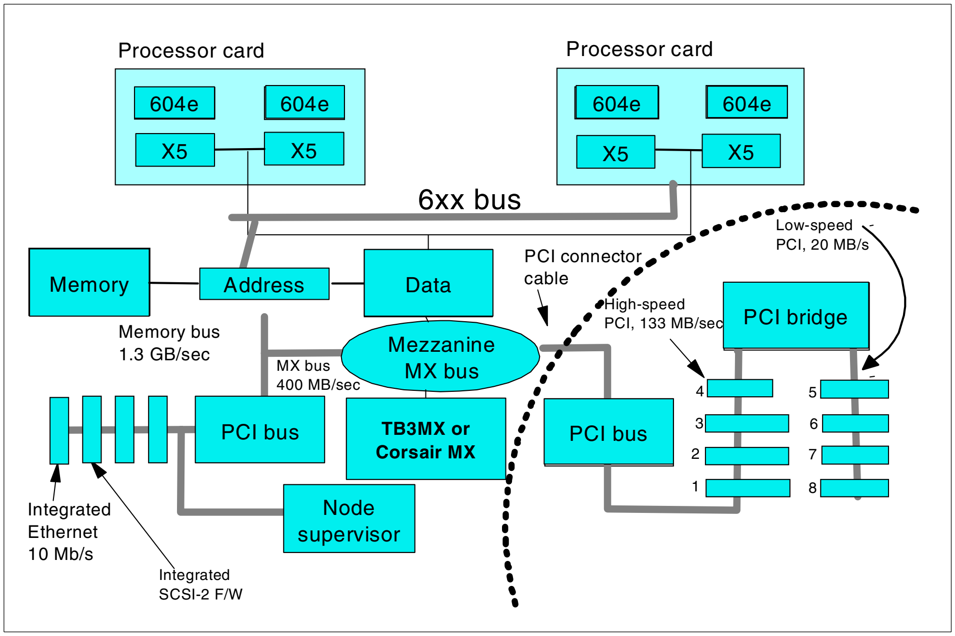


Figure 2-2   PowerPC 604e 332 MHz processor

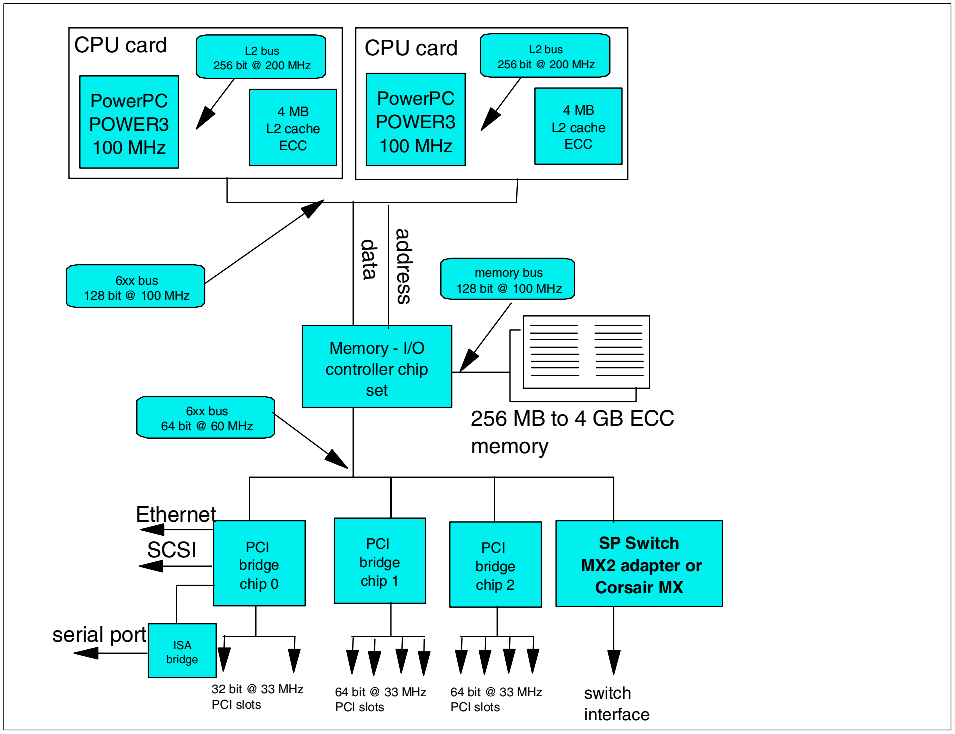


Figure 2-3 POWER3 SMP system architecture

The SP Switch2 PCI adapter (F/C 8397)

RS/6000 servers and pSeries servers might be attached to an SP system or might be configured in a cluster configuration (Clustered Enterprise Servers). The SP Switch2 PCI Adapter provides a solution to connect these nodes to an SPS2 in a single-plane or a two-plane configuration. The nodes benefiting are the SP-attached servers or switched CESs. The switched CESs are supported through the 9076-556 (see Section 2.1.5, “Multiple SP Switch2 node switch boards within a single frame” on page 32).

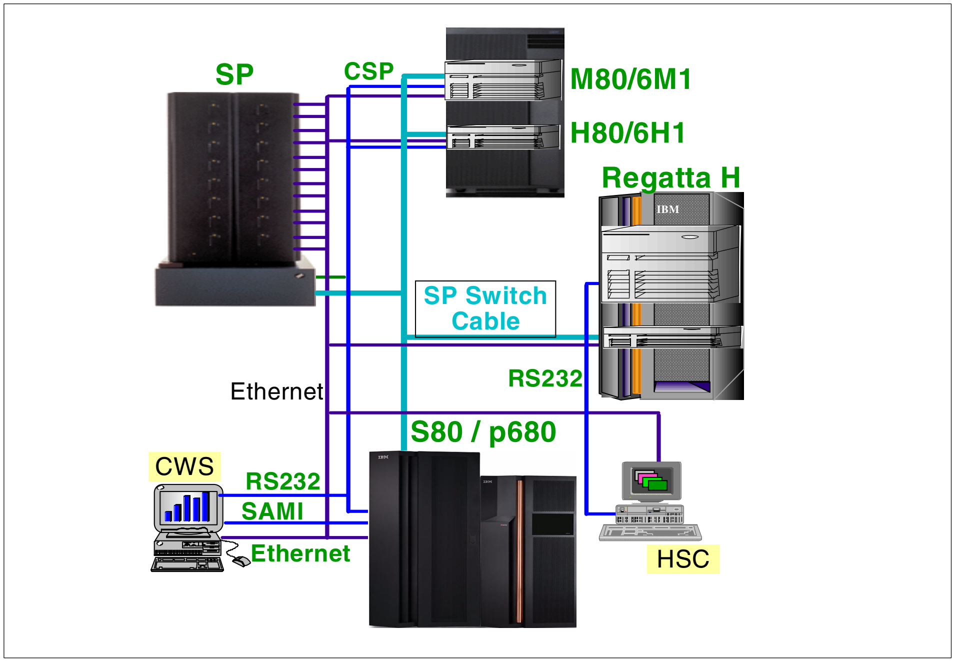


Figure 2-4 A switched cluster

2.1.5 Multiple SP Switch2 node switch boards within a single frame

The 9076-555 frame announced in April 2001 connects clustered pSeries servers to the SP Switch without the need for an SP internal node. In addition to the 9076-555 frame, IBM now offers a new 9076-556 frame.

> **Note:** The 9076 frames supports NSBs only. They cannot accommodate any SP node or I/O drawer.

We can illustrate multiple SP Switch2 node switch boards within a single 9076-556 frame with examples of output obtained using the **`splstdata`** command with various options. Example 2-1 lists frame database information, obtained by using the **`splstdata -f`** command.

Example 2-1 splstdata -f command

```
 List Frame Database Information

frame# tty              s1_tty           frame_type      hardware_protocol
control_ipaddrs domain_name
------ ---------------- ---------------- --------------- ------------------
--------------- -----------
```

```
    1 /dev/tty0      ""            switch        SP             ""
""
    2 /dev/tty1      ""            switch        SP             ""
""
    5 /dev/tty4      ""            multinsb      SP             ""
""
     6 ""               ""               ""              HMC
9.114.213.120   SeaBisquit
```

Example 2-2 lists node configuration information, obtained by using the **splstdata -n** command.

Example 2-2   splstdata -n command

```
 List Node Configuration Information

node# frame# slot# slots initial_hostname  reliable_hostname dce_hostname
default_route   processor_type processors_installed description     on_switch
LPAR_name
----- ------ ----- ----- ----------------- ----------------- -----------------
--------------- -------------- -------------------- --------------- ---------
---------
    1      1     1     4 c179n01.ppd.pok.i c179n01.ppd.pok.i ""
9.114.213.125   MP                                2 POWER3_SMP_High         1
""
    5      1     5     4 c179n05.ppd.pok.i c179n05.ppd.pok.i ""
9.114.213.125   MP                                2 POWER3_SMP_High         1
""
    9      1     9     4 c179n09.ppd.pok.i c179n09.ppd.pok.i ""
9.114.213.125   MP                                8 375_MHz_POWER3_         1
""
   13      1    13     4 c179n13.ppd.pok.i c179n13.ppd.pok.i ""
9.114.213.125   MP                                2 POWER3_SMP_High         1
""
   17      2     1     4 e179n01.ppd.pok.i e179n01.ppd.pok.i ""
9.114.213.125   MP                                8 POWER3_SMP_High         1
""
   21      2     5     4 e179n05.ppd.pok.i e179n05.ppd.pok.i ""
9.114.213.125   MP                                2 POWER3_SMP_High         1
""
   25      2     9     4 e179n09.ppd.pok.i e179n09.ppd.pok.i ""
9.114.213.125   MP                                2 POWER3_SMP_High         1
""
   29      2    13     4 e179n13.ppd.pok.i e179n13.ppd.pok.i ""
9.114.213.125   MP                                2 POWER3_SMP_High         1
""
```

```
   81      6     1     1 e159rp01.ppd.pok. e159rp01.ppd.pok. ""
9.114.213.125   MP                                4 7040-681                0
SeaBisquit
```

Example 2-3 lists node switch information, obtained by using the **splstdata -s** command.

Example 2-3 splstdata -s command

```
 List Node Switch Information
node# initial_hostname  switch_node# switch_protocol
----- ---------------- ------------ ---------------
    1 c179n01.ppd.pok.i           0             IP
    5 c179n05.ppd.pok.i           1             IP
    9 c179n09.ppd.pok.i           2             IP
   13 c179n13.ppd.pok.i           3             IP
   17 e179n01.ppd.pok.i           4             IP
   21 e179n05.ppd.pok.i           5             IP
   25 e179n09.ppd.pok.i           6             IP
   29 e179n13.ppd.pok.i           7             IP
      Plane   0         Plane   1         Plane   2         Plane   3
node# switch# chip port switch# chip port switch# chip port switch# chip port
----- ------- ---- ---- ------- ---- ---- ------- ---- ---- ------- ---- ----
    1       1    5    3       1    5    3       -    -    -       -    -    -
    5       1    5    1       1    5    1       -    -    -       -    -    -
    9       1    4    3       1    4    3       -    -    -       -    -    -
   13       1    4    1       1    4    1       -    -    -       -    -    -
   17       1    5    2       1    5    2       -    -    -       -    -    -
   21       1    5    0       1    5    0       -    -    -       -    -    -
   25       -    -    -       -    -    -       -    -    -       -    -    -
   29       2    5    2       2    5    2       -    -    -       -    -    -
switch# frame# slot# switch_partition# switch_type clock_input switch_name
switch_plane#
------- ------ ----- ----------------- ----------- ----------- -----------
-------------
      1     1    17               1        132          0 SP_Switch2
0
      2     2    17               1        132          0 SP_Switch2
1
      3     5     2               1        132          0 SP_Switch2
0
      4     5     4               1        132          0 SP_Switch2
1

master_switch_sequence number_switch_planes
---------------------- --------------------
                    83                    2
switch_plane# topology_filename           number_nodes_success max_ltu
link_delay
```

```
------------- -------------------------- ------------------- ---------
----------
            0 default.top.annotated.p0.89                  1      1024
31
            1 default.top.annotated.p1.89                  1      1024
31
switch_plane# primary_name     primary_backup   oncoming_primary
oncoming_primary_backup
------------- ---------------- ---------------- ----------------
-----------------------
            0 e179n13.ppd.pok. c179n13.ppd.pok. e159rp03.ppd.pok
c179n05.ppd.pok.ibm.com
            1 e179n13.ppd.pok. c179n13.ppd.pok. e159rp03.ppd.pok
c179n05.ppd.pok.ibm.com
```

The new 9076-556 enables support of multiple SP Switch2 node switch boards within a single no-node SP frame. It is used for single-plane and two-plane configurations of switched CESs and SP-attached servers. The Corsair adapters (see Section 2.1.4, “The SP Switch2 MX2 and SP Switch2 PCI adapters” on page 29) are used for the SP Switch2 server attachments.

Model 556 currently supports one to four SP Switch2 switches, which allows up to 32 servers in a two-plane configuration. IBM intends to support up to eight SP Switch2 switches in the future. You are still able to connect your existing 9076-550 SPS2 environment through switch-to-switch connections.

Note: If you want to include the 9076-556 frame in your environment, the entire system (nodes and CWS) must be at PSSP 3.4 or later level.

The switch plane cabling for two-plane configuration is interleaved. The first switch plane (switch plane 0) is located in the first drawer, while the second switch plane (switch plane 1) is located in the second drawer. Additional switches for each of the planes may be added later in alternate drawer positions.

2.2 p690 clusters

The IBM @server p690 model 681 (7040-681) is a high-end 32-way 1.1 GHz or 1.3 GHz processor POWER4 system. The physical resources can be configured in up to 16 logical partitions (LPARs).

PSSP presents each p690 server as a single frame, as shown in Figure 2-5 on page 36 and Figure 2-6 on page 36. Figure 2-5 illustrates a frame with one node in symmetrical multiprocessor system (SMP) mode. In Figure 2-6, showing four nodes in LPAR mode, each LPAR represents a node inside the frame.

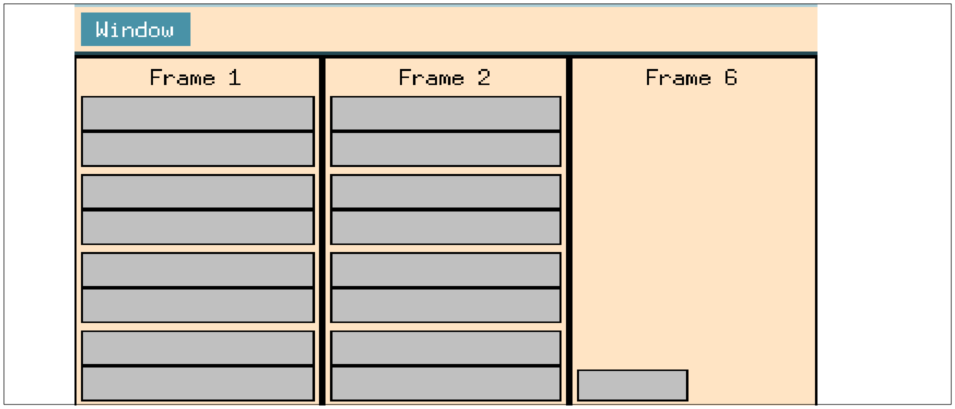


Figure 2-5 p690 in frame 6 (1 frame, 1 node in SMP mode)

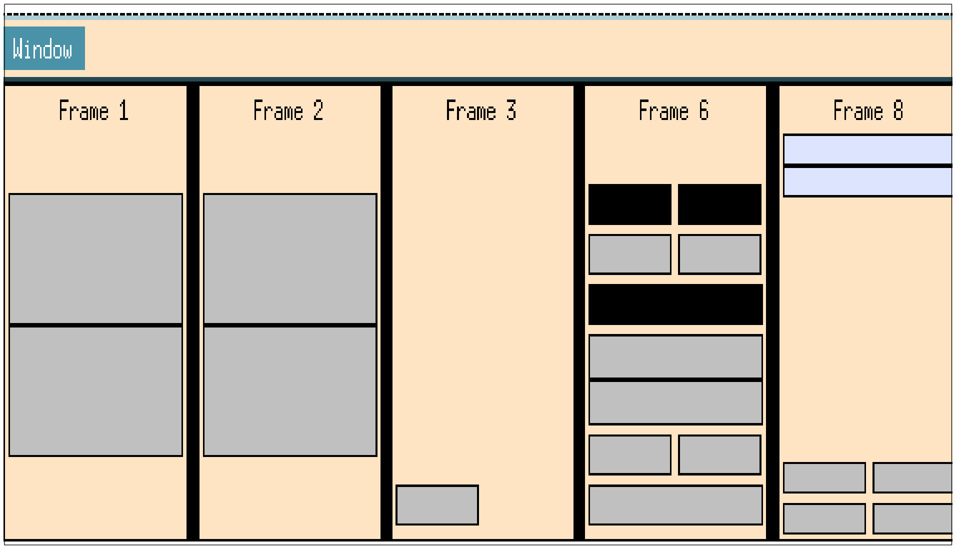


Figure 2-6 p690 in Frame 8 (1 frame, 4 nodes in LPAR mode)

You may use up to 16 p690 servers, with a maximum of 48 LPARs in a cluster environment (SP-attached servers or CESs). The IBM @serverp690 server enables very flexible cluster configurations (refer to Section 2.2.4, “p690 configurations” on page 38, for more details).

2.2.1 Cluster connectivity

The SP Ethernet LAN (SPLAN) connects all the nodes in a system running PSSP to the CWS. Since each LPAR functions as an SP node, it must meet the SPLAN requirements. For all the nodes in a cluster other than p690 nodes, you must use the adapter name en0.

> **Note:** For p690 nodes, we recommend using the physical location of the supported Ethernet adapter (for example, U1.9-P1-I2/E1).

There is no serial connection to any LPAR or p690 server. Connectivity from CWS to p690 is achieved through the Hardware Management Console (HMC) via a network connection to the SPLAN.

The p690 is supported as an SP-attached server or in a CES configuration with the SP Switch2, the SP Switch, or no switch.

2.2.2 Cluster operating system

Each p690 LPAR must run on AIX 5L Version 5.1 + PTF and PSSP 3.4 or later.

2.2.3 The Hardware Management Console (HMC)

The IBM Hardware Management Console (HMC) for pSeries is an installation and service support processor that runs only HMC software. The HMC provides the following functions for the p690 server:

- Creates and maintains a multiple partition environment.
- Detects, reports, and stores changes in hardware conditions.
- Acts as a focal point for service representatives to determine an appropriate service strategy.

One HMC can be used for up to four p690 servers.

For an IBM @server pSeries 690 server to run the PSSP software, an HMC is required with a network connection to the CWS. PSSP does not provide an interface to logically partition the p690 server; you need to use the WebSM (WebSMIT) facility provided on the HMC console. You can, however, display the HMC GUI on the CWS monitor. An interface to launch a remote WebSM session to the HMC console is provided by the SP Perspectives GUI.

Connectivity from the SP CWS to the p690 server is done through the HMC by a network connection to the SPLAN.

For stand-alone servers, the HMC is optional if the LPAR function is not needed. It is mandatory whenever the p690 server is part of a clustered environment and managed by PSSP.

2.2.4 p690 configurations

A p690 server in a clustered environment that is not partitioned is called an *SMP server* (or SMP configuration). A partitioned p690 server in a clustered environment is called an *LPAR server* (or LPAR configuration). In both configurations, several rules and restrictions apply when a p690 server is part of the cluster. These rules and restrictions are discussed in the following sections.

As a general rule there is an upper limit of 16 p690 servers, with a maximum of 48 LPARs supported in a clustered environment.

SP Switch cluster environments

If the p690 server is configured within an SP Switch environment, then all LPARs must be connected to the SP Switch with a TB3PCI adapter (only one is supported per LPAR).

Restriction: A maximum of two p690 servers and a maximum of sixteen LPARs for the system are supported (8 LPARS per server).

The TB3PCI adapter must be connected to slot 8 (see Figure 2-7 on page 39). The adapter takes two slots due to its large heatsink and cannot be mixed with the Corsair PCI adapter (see Section 2.1.4, “The SP Switch2 MX2 and SP Switch2 PCI adapters” on page 29). The SPLAN adapter must be in the same LPAR but need not be in the same drawer.

Note: We recommend that no other adapters be plugged into slots supported by an EADS chip (refer to Figure 2-7 on page 39) that handles any switch (TB3PCI or Corsair PCI) adapter.

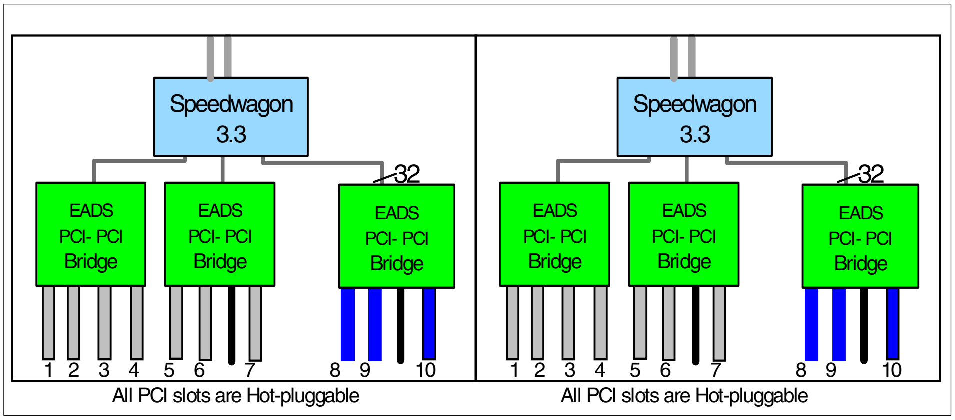


Figure 2-7 PCI slots in a p690 Bonnie & Clyde drawer

SP Switch2 cluster environments

The LPAR is connected to the SP Switch2 using the Corsair PCI adapter (refer to Section 2.1.4, “The SP Switch2 MX2 and SP Switch2 PCI adapters” on page 29). Due to its large heatsink, each adapter takes two slots.

The Corsair PCI adapter plugs into slot 3 or 5 (see Figure 2-7) if a single-plane switch is used or into slot 3 for css0 (switch plane 0) and to slot 5 for css1 (switch plane 1) in a two-plane SP Switch2 environment.

For a switchless system or for an SP Switch2 system

You can have up to sixteen p690 servers in one system in these environments, and any p690 can have up to sixteen LPARs. The total limit of LPARs is still 48.

An SP Switch2 system with all nodes running PSSP 3.4

The LPARs can be optionally connected to the switch. Though you can have up to 16 LPARs in one p690, only eight of them can be connected to the SP Switch2 regardless of the number of switch planes used. Any LPARs exceeding the count of eight in one p690 server can be in the system but not connected to the switch. Keep in mind that the total count of p690 servers cannot exceed sixteen and the total count of LPARs cannot exceed 48.

2.3 The SP-attached server 7026-6M1

The IBM @server p660 model 6M1 announced in September 2001 is now supported as an SP-attached server as well as a CWS (see Figure 2-4 on page 32). The 6M1 may be connected to an SP Switch with a TB3PCI SP Attachment Adapter or to an SP Switch2 with the Corsair PCI Adapter (refer to Section 2.1.4, “The SP Switch2 MX2 and SP Switch2 PCI adapters” on page 29). The 6M1 can participate in a single-plane or a two-plane configuration.

The IBM @server p660-6M1 replaces the IBM RS/6000 M80 server. It incorporates a higher-performance processor than its predecessor and double the main system memory.

The capacity upgrade on demand (CUoD) is also enabled with the p660-6M1. The CUoD delivers more processor capacity than is enabled and needed initially.

2.3.1 SP-attached server cable for 7026 servers

if the 7026 rack-mounted servers are used as SP-attached servers or as a CES, they only require a serial cable—the same as the SAMI cable with 7017 models—but no additional S1 cable or ground-to-ground cable as with 7017 models. In order to eliminate shipping of additional cables and any possible confusion regarding the cabling, a new F/C 9125 is used for 7026 servers.

2.4 The Cluster 1600

The RS/6000 SP has been successful for several reasons:

- Reliability, scalability, and serviceability
- High-speed computing power
- High-bandwidth communication
- Single system-wide serial number

IBM introduced the IBM @server Cluster 1600 (M/T 9078-160) to incorporate the same benefits when SP-attached servers or CESs are involved.

Under the 9078-160, there are content F/Cs for each of the cluster-enabled systems and RS/6000 SP systems—whether with clusters or legacy nodes/servers. This includes the 9076-555 and 9076-556 frame models and any previously installed SP-attached servers or CESs. See Table 1-1 on page 5 and Table 1-2 on page 6 for more information on the new hardware and software F/Cs.

For more information on the IBM @server Cluster 1600, refer to the IBM *@server Cluster 1600: Planning, Installation and Service*, GA22-7863.

2.5 PCI I/O adapters

Four new PCI I/O adapters are available for the SP system and one in the SP-attached servers and CESs:

- Dual-channel Ultra3 SCSI Adapter (F/C 6203)

  This adapter provides ultra3 support for internal ultra3 SCSI disks and attaches ultra3-capable 2104-DU3/TU3 Expandable Storage Plus rack/tower models. It is also compatible with the devices supported by the existing ultra2 adapter.

- Unshielded twisted pair (UTP) 155 Mbps ATM adapter (F/C 4953) and multimode fibre 155 Mbps ATM adapter (F/C 4957)

  These 64-bit/66 MHz adapters are both compatible with existing ATM infrastructure.

- Enhanced 10/100 Mbps Ethernet adapter (F/C 4962)
- Four-port Ethernet capability extended to SP-attached servers and CESs

For more information on the new PCI I/O adapters, refer to the *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment,* GA22-7280.

2.6 RAS enhancements

The complexity of today's servers and the complexity of the businesses they participate in are leading to the idea of self-managing systems. The result is the IBM eLiza project of Figure 2-8 and the following Web site:

http://www.ibm.com/servers/eserver/introducing/eliza/index.html

Figure 2-8 Project eLiza

An IBM @server initiative, eLiza strives to make your e-business infrastructure an autonomous, self-managing system. eLiza extends the IBM @server platform to help to get a system that is:

- Self-configuring
- Self-healing
- Self-optimizing
- Self-protecting

eLiza consists of several hardware and software implementations. For more information, refer to Figure 2-9.

2.6.1 Self-configuration

Self-configuration is achieved through replacement of defective parts while the system is running. Online hot-plug of electro-mechanical components keeps the system operational:

- Hot-swappable disk drives
- Hot-plug fans
- Hot-plug power subsystems
- Hot-swap PCI adapters

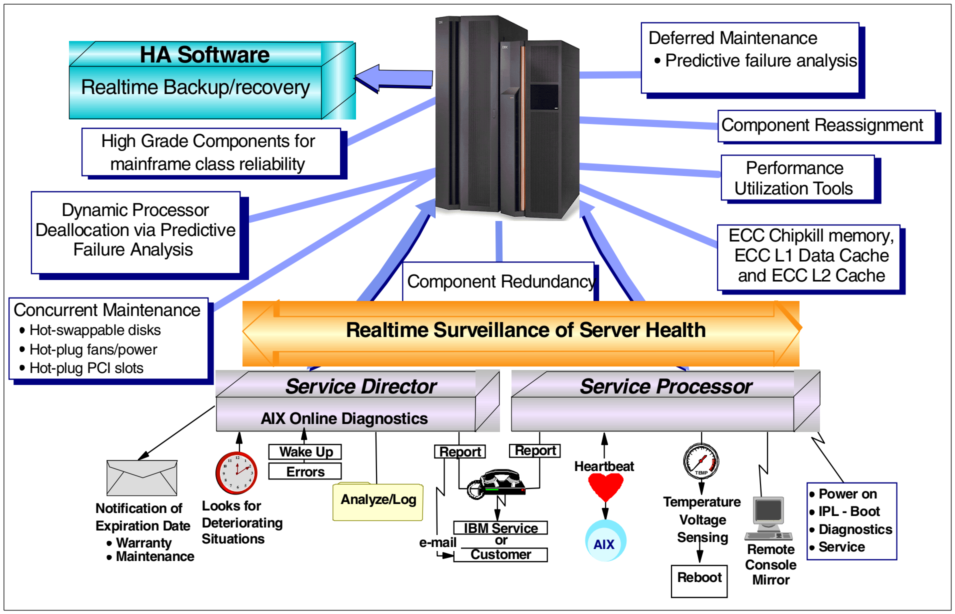


Figure 2-9 pSeries RAS features

2.6.2 Self-healing

The key to self-diagnosis is the First Failure Data Capture (FFDC) technology. It is exclusive IBM technology with specialized hardware designed to capture hardware problems.

FFDC constantly monitors every critical part of the system and provides real-time data on machine status. The system identifies the precise component that is failing and takes actions to prevent a failure, correct, or isolate the failing component. If needed, the system contacts IBM service.

There are several error prevention methods:

- Error prevention through retries
  - Error checking and correction (ECC) cache and memory (single bit correction)
  - Chipkill memory (double bit correction)
- Error prevention through isolation
  - Dynamic processor deallocation
  - PCI bus deallocation
- Error prevention through reassignment
  - Bit steering memory

3

Software enhancements

In this chapter, we introduce the following software enhancements:

- PSSP-related software enhancements
  - Corsair support
  - LPAR
  - secure shell remote program
  - Perl
  - Switched and non-switched nodes
- AIX 5L Version 5.1
- Cluster system management:
  - PSSP
  - AIX CSM
- Switch-management software
- Fibre Channel Boot support
- Reliability, availability, and serviceability
- High Availability Cluster Multi-Processor (HACMP)
- Virtual Shared Disk (VSD)

3.1 PSSP 3.4

PSSP 3.4 provides enhanced software quality, hardware support, and enablement of new nodes on which it can operate via the AIX 5L Version 5.1 and AIX Version 4.3.3 operating systems. Functional enhancements include new or expanded support in the following areas:

- Hardware
- The SP Switch2
- Boot-install from Fibre Channel SAN Direct Access Storage Devices (DASDs)
- Secure remote command processing
- The Low-Level Application Programming Interface
- IBM Virtual Shared Disk support for Subsystem Device Driver
- Support for 64-bit applications running with AIX 5L Version 5.1 on performance optimization with enhanced RISC (POWER) nodes
- Migration and coexistence support
- PSSP-related licensed programs

3.1.1 Management support

PSSP 3.4 provides new SP Switch2 connectivity for several IBM @server pSeries and RS/6000 servers (p660, p680, p690) as SP-attached servers and existing SP PCI (Silver, WinterHawk-1, and WinterHawk-2) nodes using the latest SP Switch2 Attachment Adapters (see Section 2.1.4, “The SP Switch2 MX2 and SP Switch2 PCI adapters” on page 29) for the SP Switch2 environment.

SP Switch2 attachment adapter support

PSSP 3.4 support for the SP Switch2 MX2 and SP Switch2 PCI attachment adapters implies enhancements in:

- PSSP systems management/configuration supports a single-port SP Switch2 PCI attachment adapter for the SP Switch2 on SP PCI nodes.
- PSSP systems management/configuration supports an SP-attached server connected to the SP Switch2 using the SP Switch2 PCI attachment adapter.
- The connectivity subsystem (CSS) and switch communication subsystem recognizes the new node and switch combinations to invoke the correct data copy routines that are dependent upon specific hardware characteristics.

p690 support

PSSP 3.4 allows the SP to attach p690 servers in SMP or LPAR configurations:

- The SP recognizes each attached system as a frame, with each LPAR AIX image represented as a node.
- Integrated hardware control is provided by an API through the Hardware Management Console (HMC) exploited by a hardmon extension.
- Each LPAR can have a PCI switch adapter (up to two for the SP Switch2) for switch connectivity.
- Each LPAR AIX image runs a complete copy of PSSP, and all PSSP services are available. Most of the SP software does not need to be aware of the p690 server running in LPAR mode.
- PSSP 3.4 supports a maximum of 16 LPARs per server (depending on the system configuration).

Secure shell remote program

The secure shell remote program is a follow-on item for a PSSP 3.2 functionality that removed a PSSP authorization requirement for the root user on all SP nodes to issue the **rsh** and **rcp** commands to all other SP nodes and the SP control workstation (CWS).

The work started in PSSP 3.2 is extended in PSSP 3.4 by:

- Removing the current PSSP requirement for the use of the **rsh** and **rcp** services.
- Restricting the ability of the root user on an SP node to modify critical SP system information.

Previously, SP system management components rely on the ability to issue **rsh** and **rcp** commands from SP nodes to the SP control workstation (CWS) and from the CWS to SP nodes. As a result, the SP system is configured to allow processes running under the root user ID on the SP CWS or any SP node to issue **rsh** and **rcp** commands to any other node or the CWS. This ability ensures that a security compromise of any SP node results in the compromise of the entire SP system (all nodes and the CWS).

Now, PSSP allows the customer to chose a secure shell remote program (see Section 5.3, “Secure remote command processes” on page 107) as an alternative to **rsh**. This program can be used by the PSSP software on the CWS when root is issuing remote commands from the CWS to the nodes, in order to provide a more secure, encrypted communication environment between untrusted hosts over an insecure network. This secure shell remote program, when enabled, removes all requirements in the PSSP software for **rsh** and **rcp** commands.

Perl

There is no longer a dependency on Perl 4 for a subset of the System Management files. This dependency has been removed.

Perl conversion to AIX Perl and error code enhancements are:

- Existing SP Perl 5 version scripts work with the version of Perl supported in AIX.
- Scripts using SP Perl 4 work with the version of Perl 5 supported in AIX.
- Error code handling for Perl scripts has been improved.

Switched and non-switch nodes

Mixes of switched and non-switched nodes in an SP Switch2 system are allowed (refer to Section 3.4.1, “Switched and non-switched nodes” on page 54). There is no similar support in an SP Switch system.

3.2 AIX 5L Version 5.1

AIX 5L Version 5.1 represents the next generation of AIX. Fortified with open technologies from some of the world's top providers, AIX 5L Version 5.1 builds on a solid heritage of supplying integrated, enterprise-class support for RS/6000 and IBM @server pSeries systems.

With AIX Version 5 Release 1, IBM provides an industrial-strength UNIX operating system with increased levels of integration, flexibility, and performance for meeting the high demands of today's mission-critical workloads. The features provided by AIX 5L Version 5.1 include:

- 64-bit kernels, device drivers, and application environment
- New hardware support: pSeries family, p690 servers (POWER4) with logical partitioning (LPAR) enabled
- WorkLoad Manager GUI and functional updates
- Web-based System Manager distributed framework
- (Logical Volume Manager (LVM) scalability and new JFS2 file system support (for file systems up to four petabytes and files up to one terabyte in size)
- Improved RAS functions through error logging, trace, and monitoring
- Performance and debug tool enhancements
- TCP/IPv6 and Web Serving enhancements
- SecureWay Directory, Version 3.2

- Java2 Version 1.3 updates; JDK, and runtime included in the base operating system
- Linux affinity and AIX Toolbox for Linux Applications

The next paragraphs detail some of these features related to RS/6000 SP systems and Clustered Enterprise Servers (CESs).

3.2.1 AIX 5L Version 5.1 hardware support

As new hardware platforms are developed, AIX 5L Version 5.1 brings the support necessary for exploiting their architecture. AIX 5L Version 5.1 supports the new SP-attached servers:

- IBM @server pSeries 660 (6M1). This server was supported in AIX Version 4.3.3.
- IBM @server pSeries 690—supported only by AIX 5L Version 5.1

IBM @server pSeries 690 machines can be used as SMP systems, taking advantage of all their resources, or as partitioned systems. The logical partitioning requires a Hardware Management Console (HMC) attached to the pSeries 690 machine. The attachment between the server and the HMC is a serial connection. Logical partitioning means dividing a single multiprocessor computer into separate systems, called logical partitions (LPARs), each running its own operating-system image. The following characteristics apply to pSeries 690 LPARs:

- There can be up to 16 partitions per SMP machine (PSSP puts limitations on this).
- The minimum partition size is one processor.
- Partitions cannot extend beyond the base SMP machine.
- There is no sharing of resources between running LPARs, but more profiles are allowed for an LPAR.

In an LPAR environment, the platform architecture is significantly extended, to create interfaces to what is called the platform *hypervisor*, which provides several low-level functions traditionally performed in the AIX kernel.

AIX software support is also provided for new I/O adapters with special relevance to RS/6000 SP and attached servers:

- Enhanced 10/100 Ethernet (F/C 4962)
- Dual-channel Ultra3 SCSI Adapter (FC 6203)
- UTP 155 Mbps ATM adapter (F/C 4953) and multimode fibre 155 Mbps ATM adapter (F/C 4957)

- Two new SP Switch2 adapters (supported by version AIX 5L Version 5.1):
  - SP Switch2 PCI attachment adapter (FC 8397)
  - SP Switch2 MX2 attachment adapter (FC 4026)

See also Chapter 2, “Hardware support” on page 25 for more information on the latest hardware supported.

3.2.2 64-bit kernel and application binary interfaces

AIX 5L Version 5.1 provides new kernel code capabilities by enabling 64-bit support. The new, scalable 64-bit kernel has the following features:

- Provides simplified data and I/O device sharing for multiple applications on the same system.
- Provides more scalable kernel extensions and device drivers that make full use of the kernel’s system resources and capabilities.
- Allows for future hardware development that will provide even larger single-image systems ideal for server consolidation or workload scalability.

AIX 5L Version 5.1 provides kernel enhancements and 64-bit API changes to support industry standards. You can see the differences between AIX Version 4.3.3 an AIX 5L Version 5.1 in Figure 3-1 on page 51.

AIX 5L Version 5.1 can use two types of kernels: 32-bit and 64-bit. The 64-bit kernel can be enabled at installation time or later, by recreating the symbolic-link files to a 64-bit kernel file. However, you should check that the 64-bit kernel is compatible with your platform by issuing **`bootinfo -y`** command. See the *AIX 5L Differences Guide Version 5.1 Edition,* SG24-5765, for further details.

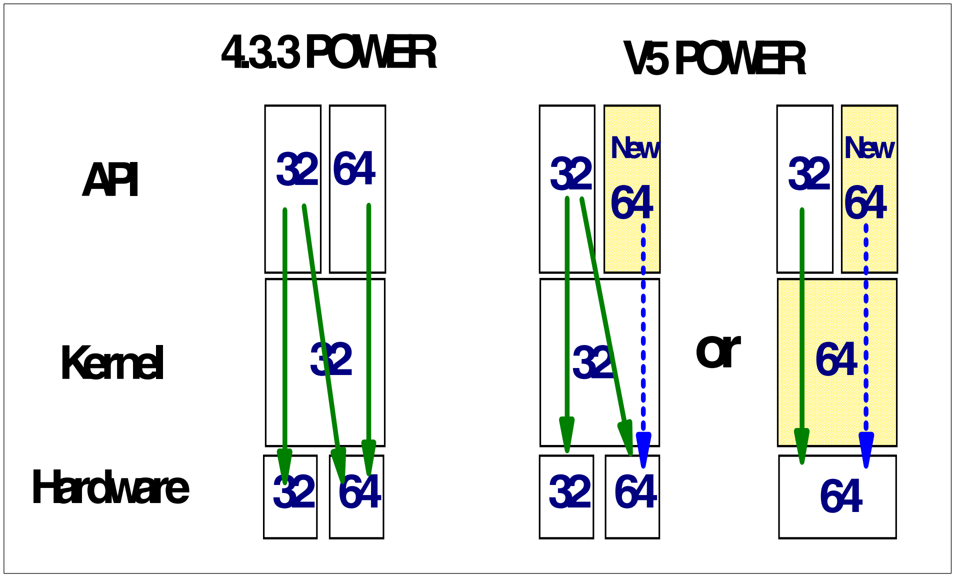


Figure 3-1 AIX 5L Version 5.1- New 64 bit Kernel and API

32-bit applications and products have upward compatibility, but applications developed for AIX 5L Version 5.1 (32/64 bit) do not have backward binary compatibility. AIX 64-bit applications running on AIX 4.3 are not binary compatible with AIX 5L Version 5.1. They need to be recompiled using an upgraded 64-bit compiler for AIX 5L Version 5.1.

3.2.3 PSSP changes with AIX 5L Version 5.1

PSSP 3.4 uses the 32-bit kernel of AIX 5L Version 5.1 but still can run 64-bit applications. The new 64-bit API allows applications to better exploit the 64-bit hardware architecture.

For MPI libraries, 64-bit support is supplied by Parallel Environment (PE) Version 3.2, which runs only on AIX 5L Version 5.1.

AIX 5L Version 5.1 does not support General Parallel File System (GPFS) 1.4 or earlier releases, so the new release 5 is required for AIX 5L Version 5.1 environments.

A new version of LoadLeveler, 3.1 is now provided for AIX 5L Version 5.1. The previous versions of LoadLeveler do not run in AIX 5L Version 5.1 environments. New improvements in LoadLeveler 3.1 allows it to set up AIX Workload Manager to control CPU and memory usage. See Chapter 7, “LoadLeveler 3.1” on page 137 for more details on LoadLeveler 3.1 enhancements.

Reliable Scalable Cluster Technology (RSCT) 2.2 is now supplied by AIX. The older RSCT 1.2.1 running on AIX Version 4.3.3 is automatically updated in migration to AIX 5L Version 5.1.

3.2.4 Other features of AIX 5L Version 5.1

There are some other features in AIX 5L Version 5.1 that are of interest for RS/6000 SP and CES environments:

- Capability to perform boot/install from SAN DASD

  Boot from SAN attached disks, as well as NIM install from SAN attached disks is provided on both 6227 and Flipper 64 adapters. To minimize boot time, booting can be limited to the first four adapters configured. Furthermore, it can be assumed that the boot adapter does not have more than 256 volumes attached to it. Even though it is possible to create an extremely large environment with Fibre Channel; these environments do not need to be configured before finding the boot volume.

- New and enhanced performance analysis tools

  New tools include truss, wlmmon, and PMAPI. Improvements were also made to iostat, tprof, and other performance tools. AIX 5L Version 5.1 also provides Performance Toolbox and Performance AIDE Version 3.0.

- Electronic licensing in AIX 5L Version 5.1

  AIX 5L Version 5.1 has been enhanced to handle electronic software license agreements. These are new features to administer license agreements and associated documents. Information about all available license agreements on the system is kept in the /usr/lib/objrepos/lag agreement database file.

  The agreement database only includes license-agreement information, and does not contain information about usage licenses, such as those administered by the Licensed User Management (LUM). The agreement text itself is stored in the /usr/swlag/<locale> directory. The license agreement database is designed so that license information from non-IBM installation programs can be integrated.

  Electronic licensing is now part of the AIX 5L Version 5.1 installation. During the install or migration to AIX 5L Version 5.1, you are required to accept all software license agreements. PSSP 3.4 was enhanced to bypass this interactive step for all nodes. See Chapter 4, “Installation, migration, and coexistence” on page 77 for more details.

3.3 Cluster system management software

Cluster management features allow a single point of control for all clustered nodes and resources. The cluster management software simplifies system management and reduces the administrative and operational cost of managing distributed servers.

> **Attention:** There are two cluster system management software available: PSSP and CSM. Do not confuse Parallel System Support Programs (PSSP) and Cluster System Management (CSM).

PSSP

PSSP has always been the main system-management software used in SP implementations. Its functionality and usability have evolved over the years, and now PSSP is enhanced to support a cluster of stand-alone IBM @server pSeries servers with or without the standard SP frames and nodes.

The single point of control within PSSP governs:

- Installation and configuration
- Hardware control
- Base cluster technology
- System monitoring and problem management
- System administration
- Cluster security

CSM

IBM's Cluster Systems Management for AIX 5L Version 5.1 lets you manage and monitor multiple AIX 5L Version 5.1 machines from a single point of control. This solution for distributed system management lets a system administrator form a cluster of up to 32 IBM @server pSeries machines or RS/6000 servers (nodes) that run the AIX 5L Version 5.1 operating system. By using functions such as monitoring and configuration file management, an administrator can easily set up a cluster and maintain it. CSM includes the following features:

- You can add, remove, change, or list nodes (with persistent configuration information displayed about each node in the list).
- You can run commands across nodes or node groups in the cluster, and gather responses.
- You can monitor nodes and applications to see whether they are up or down.

- You can monitor CPU, memory, and system utilization, and run automated responses when events occur in the cluster.
- A configuration file manager provides synchronization of files across multiple nodes.

A single management server is the control point for the CSM cluster. Note that CSM manages a loose cluster of machines: It does not provide high-availability services or failover technology, although high-availability clusters can be part of the set of machines that CSM manages. In subsequent releases, IBM plans to make available greater scalability built into the software.

CSM information is available at the following address:

http://www.alphaworks.ibm.com/tech/aixcsm

3.4 Switch management software

PSSP 3.4 and the SP Switch2 technologies give you the ability to use nodes that are not able to participate, or unwanted, in a switched environment. However, if they are connected to the SP Switch2, you may use only one IP address and one host name to reference a node.

3.4.1 Switched and non-switched nodes

With the control workstation and all the nodes running PSSP 3.4, you have optional switch connectivity. This means a system using the SP Switch2 can have some nodes that are not connected to the switch. This way you can use the SP Switch2 and still keep older nodes in your system that are not connected to the switch, if they are not supported on the SP Switch2.

Note:

- The Switch node numbers for nodes not on the switch is -1.
- If the css adapter has been defined, the on_switch value is 1.
- If no css adapter has been defined, the on_switch value is 0.

If an SP node is installed in a two-plane environment (see Section 2.1.1, “Two-plane option” on page 26), then the SP Switch2 must be connected to both SP Switch2 planes or to neither of them.

Note: The only switch environment allowing non-switched nodes is the SP Switch2 environment.

3.4.2 Aggregate IP addressing

The current PSSP 3.2 design provides on two plane systems, two independent IP interfaces for each SP Switch2 adapter (css0 and css1) each having its own unique IP address. Although each node has access to both switch networks, there is no physical or logical connection between them.

The Aggregate IP address function in the new PSSP 3.4 provides aggregate/multilink IP device abstraction for IP SP Switch2 networks. This pseudo-device driver enables SP administrators to use one IP address over two SP Switch2 fabrics.

Two benefits accrue from configuring an aggregate IP address in your system:

- Higher data throughput, allowed by data striping across both switch adapters.
- High availability. If one path is broken, the other one is used without interrupting the switch communication.

The third IP address, associated with the ml0 device, allows IP messages to be transmitted in a more economical manner called *striping*. The striping technique provides the capability to transmit consecutive IP data across two fully operational adapters, taking advantage of their combined bandwidth. For example, when an IP message is sent between nodes and both nodes have access to both available switch networks, consecutive datagrams are sent in the pattern: ...css0, css1, css0, css1...

In addition, ml0 can ensure that a single failure in the SP Switch2 subsystem does not cause a complete outage to a node or subsystem dependent on an SP Switch2. If a fault occurs between a node and one of the two configured switches, a transparent failover occurs using the ml0 interface to access the remaining functional switch. Using the example above, if adapter –0 (css0) malfunctions, the resulting data flow is ...css1, css1, css1, css1...

Important: Aggregate IP addressing can be used only with SP Switch2 systems (see Section 2.1, “Enhancements to the RS/6000 SP Switch2” on page 26), only one aggregate IP adapter (ml0) can exist per node, and css0, ccs1 and ml0 requires unique IP addresses.

SDR changes

Table 3-1 lists the new Aggregate_IP SDR class and associated attributes used to support aggregate IP addressing.

Table 3-1 Attributes of the Aggregate_IP SDR class

Attribute	Type	Use
node_number	integer	Numeric node number containing the adapters to be aggregated
device_name	string	Alphanumeric device name that is associated with the aggregate IP address, for example, ml0
ip_address	string	IP address associated with the device_name
netmask	string	Network mask associated with the IP address
agg_list	string	List of adapter names that are aggregate, for example, css0, css1
update_interval	integer	Numeric time interval between aggregate network route table refreshes, for example, 3
update_threshold	integer	Numeric count of missed refresh updates before the network connection is terminated, for example, 10

Example 3-1 provides the SDR content for a two-switch plane configuration.

Example 3-1 An SDRGetObjects output of the Aggregate_IP SDR class

```
SP4:/usr>SDRGetObjects Aggregate_IP
node_number  device_name  ip_address   netmask      agg_list
update_interval update_threshold
```

```
1            ml0          192.168.20.1 255.255.255.0 css0,css1
3               10
5            ml0          192.168.20.5 255.255.255.0 css0,css1
3               10
```

New Object Data Manager (ODM) changes

When the ml0 adapter is configured using the **psspfb_script**, new Object Data Manager (ODM) entries are created (see Table 3-2).

Table 3-2 New ODM attributes for ml0 device

Attribute	Sample Value	Type	Generic	Rep	NLS	Note
agnetaddr	192.168.20.1	‘R’	‘DU’	String	Yes	IP address
agnetmask	255.255.255.0	‘R’	‘DU’	String	Yes	Network mask
agglist	css0, css1	‘R’	‘DU’	String	Yes	List of adapter name to be aggregated
aginterval	3	‘R’	‘DU’	Number	Yes	Time interval between aggregate network route table refreshes
agthreshold	10	‘R’	‘DU’	Number	Yes	Count of missed refresh updates before network is dropped

New commands

In order to manage the information relating to the aggregate IP address configuration, new commands are provided:

- **spaggip**

  This Perl script allows you to represent up to two SP Switch2 adapters on a node by providing a single IP address, netmask, and device name for the adapters.

Important: Both adapters still need their own IP addresses.

When you add an aggregate IP address object to the SDR, the command assumes a device_name attribute IP of ml0. The following rules apply:

- The CSS switch adapter objects (for example, css0,css1) must exist in the SDR Adapter class.
- IP addressing for the ml0 adapter must be distinct from any other adapter in the system.
- The IP address of the ml0 adapter must be resolvable.
- The node must be reinstalled or customized for the changes to take effect.

The syntax and externals of the **`spaggip`** command are detailed in the man page. You can also use the SMIT panel to configure the aggregate IP address issuing the fast path **`smitty sp_agg_dialog`**.

- **`spdelagg`**

  This Perl script allows you to remove an aggregate object from the SDR. The command assumes the deletion of the aggregate IP object ml0. You must re-install or customize the nodes for the changes to take effect.

  The syntax and externals of the **`spdelagg`** command are detailed in the man page. Alternatively, you can use the SMIT fast path **`smitty delete_aggip_dialog`**.

Command changes

The following commands have been modified in PSSP Version 3.4:

- **`spdeladap`**

  This command has been modified to check for the presence of a virtual device name associated with the adapter being deleted. If the SP Switch2 adapter is represented by an Aggregate_IP object and can be removed, the associated virtual device name is removed only if the virtual device is representing no other switch adapters. For example, if ml0 represents the *css0* and *css1* adapters, *ml0* is removed from the SDR only if both *css0* and *css1* are deleted. If only css0 is removed, then the *ml0* interface remains as a representation for css1.

- **`spdelnode`**

  This command has been modified to check for the presence of a virtual device associated with a node. If the node can be deleted, all Aggregate_IP objects associated with it are also deleted.

- **splstdata -g**

  Use this command option to display aggregate IP information, specifically aggregate IP information associated with a particular adapter. See Example 3-2.

Example 3-2 IP aggregate information gathered with the splstdata -g command

```
SP4:/>splstdata -g
                  List Aggregate IP Database Information

node# adapt  netaddr         netmask         hostname          devicename
update_interval   update_threshold
----- ------ --------------- --------------- ----------------- -------------
----------------- ------------------
    1 ml0    192.168.20.1    255.255.255.0   sp4ml01.msc.itso. css0,css1
3                 10
    5 ml0    192.168.20.5    255.255.255.0   sp4ml05.msc.itso. css0,css1
3                 10
```

Other aggregate IP changes

The **mkconfig** command adds information to the <node>.config_info file to configure the ml0 adapter correctly in the **psspfb_script**.

psspfb_script configures the adapters defined in a node's SDR Adapter and Aggregate_IP objects. It creates new ODM entries for the aggregate IP address, and it obtains the value for these entries from the <node>.config_info file generated using the **mkconfig** command. The additional information appended to the end of the config_info entry, ci_agglist, ci_intvl, ci_count, is supported only in PSSP 3.4.

Changes have also been made to some security scripts, such as **setup_CWS** and **setupdce**, to add the security credentials for the ml0 adapter.

3.5 Fibre Channel boot

PSSP 3.4 supports boot and Network Installation Manager (NIM) installs from Direct Access Storage Devices (DASDs) attached by Fibre Channel in a storage area network (SAN). This support incorporates new functions developed in system firmware, AIX 5L Version 5.1, and Fibre Channel adapter firmware to natively support Fibre Channel SAN DASDs as AIX 5L Version 5.1 installation and boot devices.

The SDR Volume_Group class (see Table 3-3) is changed accordingly to support this new feature.

Table 3-3 The SDR Volume_Group class pv_list attribute

Attribute name	Type	Description	Comments
pv_list	S	A list of physical volumes (pv) Valid format for hdisk specification is: hdiskn, hdiskn+1,...,hdiskm Valid format for conn where, location, PVID, or SAN_DISKID specification is the corresponding AIX attribute value for any combination of those physical volumes separated by colon: pv:pv:...:pv	Initially set to hdisk0.

3.6 IBM Virtual Shared Disk 3.4

This section covers IBM Virtual Shared Disk (VSD) topics. In Section 3.7.1 we introduce VSD concepts. Section 3.7.2 points out enhancements in VSD version 3.4, and Section 3.7.3 presents some issues related to migration and coexistence.

3.6.1 IBM Virtual Shared Disk overview

This section is a brief introduction to IBM Virtual Shared Disk (VSD) concepts and components. For further details, refer to *PSSP for AIX: Managing Shared Disks,* SA22-7349.

Virtual Shared Disk

The IBM VSD software is a component of PSSP that allows you to share data on disks attached to some SP nodes, with other nodes using a high-speed interconnect network like SP Switch or SP Switch2.

A VSD is a logical volume that can be accessed both by the system to which it belongs and by other nodes in the system. The VSD nodes that share data may reside in a single system partition.

VSD software allows applications running on different nodes to have access to raw logical volumes as if they were local. I/O requests to VSDs are routed by the VSD device driver, which is loaded as a kernel extension on each node, thus making raw logical volumes accessible to other nodes in the system.

Depending on the VSD function, a node can be:

- A VSD server, which has local attached disks. It is able to complete I/O requests from VSD clients by using a communication network inside the system.
- A VSD client, which is a node requesting access to VSDs.

Note: It is possible for a node inside the system to be both a server and a client for VSD services.

The communication protocols used by VSD nodes can be:

- TCP/IP
- Kernel Low-Level Application Programming Interface (KLAPI)

KLAPI is a protocol designed to improve communications and to extend the capabilities of VSD and applications like General Parallel File System (GPFS) that can exploit it. KLAPI provides transport services to kernel subsystems that need to communicate via SP Switch or SP Switch2.

Recoverable Virtual Shared Disk

Using Recoverable Virtual Shared Disk (RVSD) along with twin-tailed disks or disk arrays allows a secondary node to take over the server function of the primary VSD server, in case that server fails.

The RVSD uses the Group Services subsystem of the Reliable Scalable Cluster Technology (RSCT) component to monitor failures and automatically manage the VSDs for several types of failures:

- Node failures
- Adapter errors
- Disk failures

RVSD can switch disk access from the primary node to the secondary node so that applications can continue to operate normally. It allows you to dynamically change the server node by using the `vsdchgserver` command.

Concurrent Virtual Shared Disk

The Concurrent Virtual Shared Disk feature of PSSP allows multiple VSD servers to simultaneously access logical volumes inside a volume group, using the Concurrent Logical Volume Manager (CLVM) component supplied by AIX. I/O requests from nodes that do not have locally attached disks are spread across VSD servers, thus improving raw logical volume access.

When you use Concurrent Virtual Shared Disk, recovery from node failure is much faster because the failed node is marked as unavailable to all other nodes, and its access to the physical disk is *fenced* while the other nodes can continue to access the disks.

Hashed Shared DIsk

Hashed Shared Disk (HSD) support extends the VSD concept by striping data across multiple nodes and multiple VSDs, thus reducing I/O bottlenecks.

The principal contribution of the HSD component is to provide data distribution across physical disks and nodes while being transparent to the application program, which uses VSDs.

3.6.2 Enhancements to VSD 3.4

VSD Version 3.4 comes with the following enhancements:

- Support for the Subsystem Device Driver (SDD). See “SSD support in VSD” on page 62 for more details on this feature.
- Support for two switch adapters in SP Switch2 configurations. See “Multilink adapter support” on page 71 for more information.

SSD support in VSD

This section describes the Subsystem Device Driver (SDD) and VSD support for SDD.

SDD overview

The IBM SDD software resides in the host server with the disk device driver for the Enterprise Storage Server (ESS). It uses redundant connections between the host server and ESS disk storage to enhance performance and data availability.

The SDD provides the following functions:

- Enhanced data availability
- Automatic path failover and recovery to an alternate path
- Dynamic load balancing of multiple paths
- Concurrent microcode upgrade

Host servers are usually configured with multiple host adapters with SCSI or Fibre Channel connections to an ESS that, in turn, provides internal component redundancy. With dual clusters and multiple host interface adapters, the ESS provides more flexibility in the number of available I/O paths, as shown in Figure 3-2.

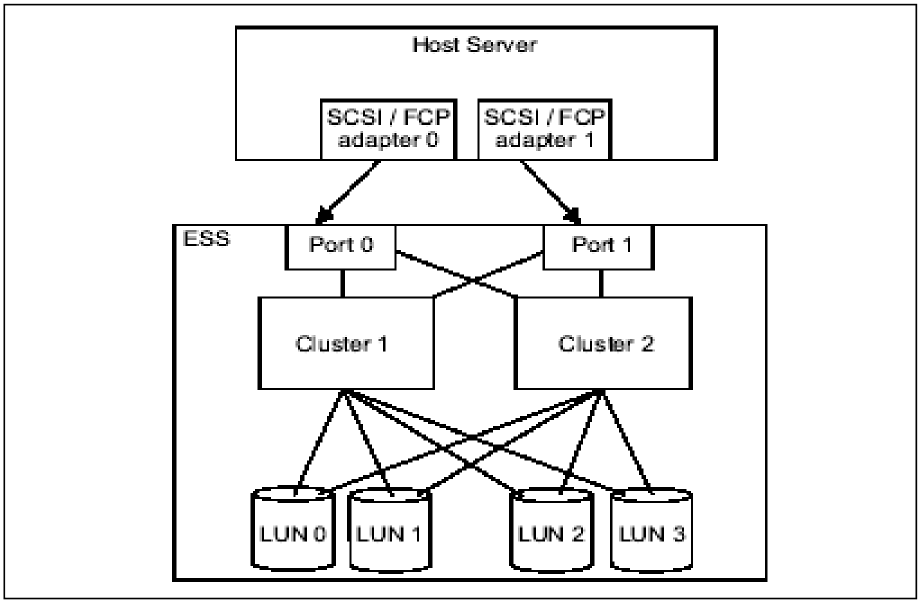


Figure 3-2 ESS disk configuration with multipath connections to a host system

The SDD failover system is designed to provide recovery upon failure of a data path. If a data path fails, the failover system selects an alternate data path and minimizes any disruptions in operation. The failover process consists of the following actions:

1. Detecting a failure
2. Signaling to the AIX host that a failure has occurred
3. Compensating for the failure by selecting an alternate data path

When a failure occurs, the SDD reroutes I/O operations from the failed path to the remaining available paths. In this way the SDD provides fault tolerance to failures in the following:

- Bus adapter on the host server
- External SCSI or Fibre Channel cable
- Host interface adapter on the ESS

SDD also exploits the multiple paths provided by ESS configurations by load balancing of data flow to prevent a single path from becoming overloaded with I/O operations.

When configuring SDD for an ESS attached to your AIX system, you must consider two drivers:

- The ESS driver (for example, the ibm2105.rte ESS package)
- The SDD (for example, the ibmSdd_433.rte package)

In host systems using multiple paths to access the logical unit number (LUN) inside the ESS, an hdisk is created for every path accessing the LUN. Figure 3-3 on page 65 shows two systems connecting to an ESS using a Fibre Channel (FC) switch. Node1 has 4 Fibre-Channel adapters, and the switch is connected to two ports of the ESS. As a result, a total of eight hdisks are created for the eight paths to the same LUN. Node2, having two adapters, has four hdisks created. Note that the hdisks are created as the result of installing the ESS drivers and running the **cfgmgr** command.

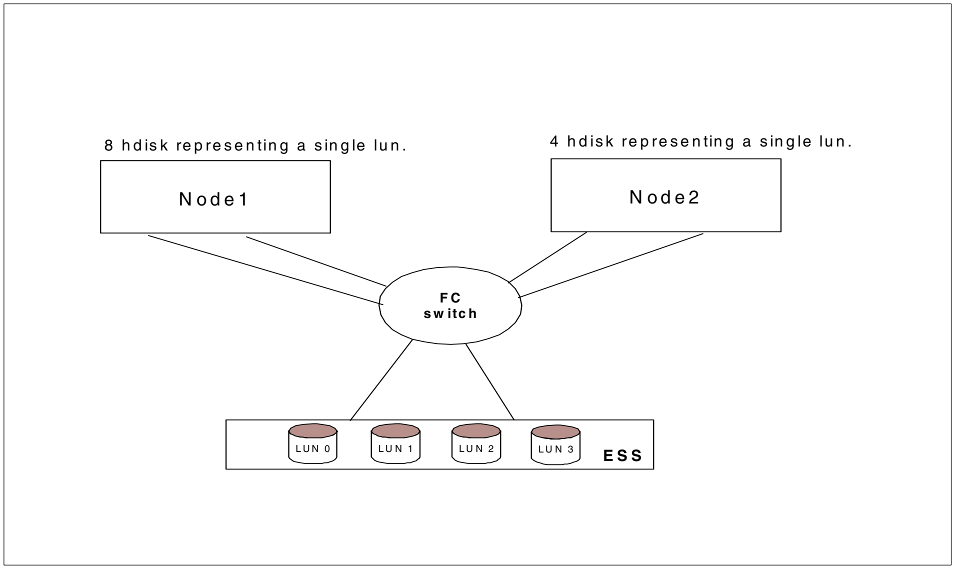

8 hdisk representing a single lun.
4 hdisk representing a single lun.
Node1
Node2
FC switch
LUN 0
LUN 1
LUN 2
LUN 3
ESS

Figure 3-3 Fibre Channel switch configuration: two hosts connecting to an ESS

When an SDD is installed and configured in a system, a special device called a *vpath* is created for accessing an external LUN in the ESS. Depending on the configuration, a vpath can be made up of one or multiple hdisks.

The AIX **lspv** command displays multiple hdisks in addition to vpath devices. In order to get further information on vpath devices you must issue the **lsvpcfg** command. Example 3-3 gives sample output for these commands.

Example 3-3 Sample lspv and lsvpcfg output

```
c185n01:/# lspv
      hdisk0         000047690001d59d    rootvg
      hdisk1         000047694d8ce8b6    None
      hdisk18        000047694caaba22    None
      hdisk19        000047694caadf9a    None
      hdisk20        none                None
      hdisk21        none                None
      hdisk22        000047694cab2963    None
      hdisk23        none                None
      hdisk24        none                None
      vpath0         none                None
      vpath1         none                None
      vpath2         000047694cab0b35    gpfs1scsivg
      vpath3         000047694cab1d27    gpfs2scsivg

c185n01:/# lsvpcfg
      vpath0 (Avail ) 502FCA01 = hdisk18 (Avail pv)
```

```
vpath1 (Avail ) 503FCA01 = hdisk19 (Avail pv)
vpath2 (Avail pv gpfs1scsivg) 407FCA01 = hdisk20 (Avail) hdisk24 (Avail)
vpath3 (Avail pv gpfs2scsivg) 408FCA01 = hdisk21 (Avail) hdisk23 (Avail)
```

There are a few key points to understand in Example 3-3:

- vpath0 is made up of a single path (hdisk18) and therefore does not provide failover protection. Also, hdisk18 is defined to AIX as a physical volume (note the pv flag) and has a PVID, as seen in the **lspv** output. vpath0 has no PVID since hdisk18 and vpath0 cannot be configured at the same time as independent disk volumes with their own PVIDs.
- vpath2 has two paths (hdisk20 and hdisk24) and has a volume group defined on it. Notice that in the **lspv** output hdisk20 and hdisk24 look like newly installed disks with no PVIDs.

For gathering more information related to the SDD configuration a special set of tools is provided. The **datapath** commands can help you in querying as well as setting the adapter or device status. Table 3-4 provides a brief description of the **datapath** commands.

Table 3-4 Datapath commands

Command	Description
datapath query adapter	Displays adapter information
datapath query adaptstats	Displays performance information for all SCSI and frame check sequence (FCS) adapters that are attached to SDD devices
datapath query device	Displays device information
datapath query devstats	Displays performance information for a single SDD device or all SDD devices
datapath set adapter	Sets all device paths that are attached to an adapter to online or offline
datapath set device	Sets the path of a device to online or offline

Example 3-4 on page 67 provides the output of the **datapath query device** command for an SDD configuration using two SCSI adapters connected to the two ports of an ESS as shown in Figure 3-2 on page 63. The argument of the command is the vpath device number.

Example 3-4 Querying an SDD device

```
c185n01:/# datapath query device 0

Total Devices : 14


DEV#:   0  DEVICE NAME: vpath0  TYPE: 2105F20   SERIAL: 51CFCA16
================================================================
Path#              Adapter/Hard Disk    State     Mode     Select     Errors
    0                  fscsi0/hdisk2    OPEN    NORMAL        114          0
    1                 fscsi0/hdisk17    OPEN    NORMAL        144          0
    2                 fscsi1/hdisk32    OPEN    NORMAL        136          0
    3                 fscsi1/hdisk47    OPEN    NORMAL        155          0
```

In order to use vpath devices in volume groups you must use the SDD utilities. These are capable of creating a vpath volume group from scratch or converting an hdisk volume group to a vpath volume group and vice versa. Table 3-5 provides a brief overview of the most important utilities used in volume group operations.

Table 3-5 SDD utilities

Command	Description
hd2vp	Converts a volume group from original ESS hdisks into SDD vpaths.
vp2hd	Converts a vpath volume group to original ESS hdisks
dpovgfix	Recovers from mixed volume groups. A mixed volume group is a volume group which contains both vpath and hdisk volumes.
mkvg4vp	Use this instead of **mkvg** for creating a vpath volume group.
extendvg4vp	Use this instead of **extendvg** for adding vpath volumes to a vpath volume group

Note: Mixed volume groups using both vpaths and hdisks are allowed to function in SDD configurations, but neither fault tolerance nor load balancing is provided for hdisks. You should convert mixed volume groups to vpath volume groups, using the **`dpovgfix`** command.

The operations performed by SDD inside a vpath are transparent to the applications using vpath volume groups. Creating logical volumes and file systems inside a vpath volume group is just like using hdisk volume groups.

VSD support for SDD

The IBM VSD subsystem supports virtual shared disks defined in SDD volume groups (vpath volume groups). In order to exploit the features of SDD (including automatic path failover), you must use SDD utilities to create or convert to vpath volume groups before using the VSD commands for creating virtual shared disks.

Attention: The *IBM Subsystem Device Driver: Installation and User's Guide,* SC26-7425 is a prerequisite reading for configuring VSD to use the SDD.

There are several ways you can configure VSDs using SDD:

- Convert an existing hdisk volume group to a vpath volume group.

  You can use the **`hd2vp`** or **`dpovgfix`** command. These commands remove the PVID from the existing hdisk paths and convert the hdisk volume group to a vpath volume group.

  If VSDs have already been created on an hdisk volume group, you must follow the next steps in order to exploit the SDD vpath devices instead of hdisks:

  a. Stop RVSD by issuing the **`ha.vsd stop`** command.

  b. Run **`hd2vp`** *`VGname`*. Alternatively, you can use **`dpovgfix`** *`VGname`*.

  c. When conversion is completed, start the VSDs using the **`ha.vsd start`** command.

- Use the **`createvsd`** or **`createhsd`** command

  These commands can be issued via the command line, SMIT panels, or perspectives. They are now enabled to receive hdisks as well as vpath devices as parameters.

Important: hdisk and vpaths cannot be specified in the same invocation of the **createvsd** or **createhsd** command. If you create the shared disks on hdisks rather than vpaths, you must convert the volume groups to vpath volume groups as specified in “Convert an existing hdisk volume group to a vpath volume group.” on page 68.

For example, we assume two high nodes in an SP system have the following SDD disk configuration:

```
>lspv
      hdisk20         none                 None
      hdisk21         none                 None
      hdisk22         none                 None
      hdisk23         none                 None
      vpath0          none                 None
      vpath1          none                 None

>lsvpcfg
      vpath0 (Avail ) 407FCA01 = hdisk20 (Avail ) hdisk22 (Avail )
      vpath1 (Avail ) 408FCA01 = hdisk21 (Avail ) hdisk23 (Avail )
```

Issue the following command on the CWS to create a virtual shared disk definitions on the primary node 1 and the secondary node 5, where the logical volume created on nodes 1 and 5 spans two vpaths devices and the global volume group spans two vpaths:

```
# createvsd -n 1/5:vpath0,vpath1/ -s 128 -g datavg -v sdd -T 16
```

This command creates the virtual shared disk sdd1n1, with logical volume lvsdd1n1 defined on a volume group with the global volume group name datavgn1b5 on node 1, imported to node 5. The volume group datavg spans vpath0 and vpath1. The logical volume lvsdd1n1 is contained in vpath0.

The volume group datavg is created with a 16 MB physical partition. Logical volume lvsdd1n1 thus contains eight physical partitions (PPs).

Note: If you want -g datavg1, then the globalvg would be datavg1n1b5 and the volume group would be datavg1.

- Create vpath volume groups and logical volumes.

  You must build the vpath volume groups, then the logical volumes inside them using AIX LVM commands. Finally, you use the VSD commands to associate the LVM constructs with IBM VSD constructs. There are a few steps that you now must perform manually. For example, we will use the same SDD configuration as described in “Use the `createvsd` or `createhsd` command” on page 68.

a. Use the **mkvg4vp** command to create volume group datavgn1 on node 1:

```
> /usr/sbin/mkvg4vp -f -y'datavgn1' -s'16' -V'80' vpath0 vpath1
> lspv
      hdisk20        none                None
      hdisk21        none                None
      hdisk22        none                None
      hdisk23        none                None
      vpath0         000047694cab0b35    datavgn1
      vpath1         000047694cab1d27    datavgn1
```

b. Build an AIX lvsdd1n1 logical volume of 128 MB on node 1:

```
# mklv -y lvsdd1n1 datavgn1 8 vpath0
# lsvg -l datavgn1
datavgn1:
LV NAME             TYPE       LPs   PPs   PVs  LV STATE      MOUNT
POINT
lvsdd1n1            jfs        8     8     1    closed/syncd  N/A
```

c. Import the volume group definition on node 5 (volume group datavgn1 is varied off on node 1):

```
# importvg -y datavgn1 -V80 vpath0
datavgn1
# lsvg -l datavgn1
LV NAME             TYPE       LPs   PPs   PVs  LV STATE      MOUNT
POINT
lvsdd1n1            jfs        8     8     1    closed/syncd  N/A
```

d. Use the **vsdvg** and **defvsd** commands to associate the lvsdd1n1 with the sdd1n1 VSD.The volume group datavgn1 is activated at node 1. On the CWS, create the global volume group shared by nodes 1 and 5. Then, define the sdd1n1 VSD using the **defvsd** command:

```
#vsdvg -g datavgn1 datavgn1 1 5
#vsdatalst -g
VSD Global Volume Group Information
                                                Server Node Numbers
Global Volume Group name        Local VG name   primary      backup
eio_recovery     recovery     server_list      vsd_type
------------------------------ --------------- -------      ------
------------     --------    ----------------------
datavgn1                     datavgn1           1        5
1                0            0                VSD

#defvsd lvsdd1n1 datavgn1 sdd1n1
#vsdatalst -v
VSD Table
VSD name                    logical volume  Global Volume Group
minor# option    size_in_MB
```

```
------------------------------- ---------------
------------------------------- ------ --------- ----------
sdd1n1                      lvsdd1n1      datavgn1
4      nocache    128
```

Important: If SDD volume groups (vpath volume groups) are used, all nodes that access those volume groups must access them as vpath volume groups. It is not possible for one node to use the vpath volume group while another node accesses the same volume group as an hdisk volume group.

Multilink adapter support

Virtual Shared Disk 3.4 supports configurations of two switch adapters in a node connected to two SP Switch2 planes and using the IP aggregate option. For more information about IP aggregates see Section 3.4.2, “Aggregate IP addressing” on page 55.

Configuring VSD with an IP aggregate

VSD 3.4 supports multilink IP addresses for the IP network configuration. In order to use this option you must first configure the ml0 adapter in the SDR using the **spaggip** command. See Section 3.4.2, “Aggregate IP addressing” on page 55 for more details.

When assigning the VSD nodes, specify ml0 as the adapter_name in the **vsdnode** command. You can use SMIT panels as in Example 3-5.

Example 3-5 Designating VSD nodes to use the ml0 adapter

```
VSD Node Database Information

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                  [Entry Fields]
* Nodes                                        [1 5]                +
* Adapter Name for VSD Communications          ml0                  +
* Initial Cache Buffer Count                   [64]                 #
* Maximum Cache Buffer Count                   [256]                #
* VSD Request Count                            [256]                #
* Read/Write Request Count                     [48]                 #
* VSD Minimum Buddy Buffer Size                [4096]               #
* VSD Maximum Buddy Buffer Size                [131072]             #
* VSD Number of Max-sized Buddy Buffers        [4]                  #
* VSD Maximum IP Message Size                  [61440]              #
  VSD Cluster Name                             []
F1=Help          F2=Refresh                F3=Cancel            F4=List
F5=Reset         F6=Command                F7=Edit              F8=Image
F9=Shell         F10=Exit                  Enter=Do
```

3.6.3 Migration and coexistence considerations

You can have mixed PSSP levels and any supported level of the IBM Recoverable Virtual Shared Disk licensed program (RVSD) in the same system partition with PSSP 3.4, but the CWS must have RVSD Version 3.4.

Nodes migrated to RVSD Version 3.4 interoperate with different versions of RVSD components in the same system partition, but their RVSD functional level must be at the lowest level used in the partition.

To set the level at which the RVSD subsystem runs, use the **rvsdrestrict** command on these nodes. Table 3-6 provides the argument that should be passed to the command, depending on the RVSD versions existing in the system partition.

Table 3-6 Levels for the rvsdrestrict command

IBM Recoverable Virtual Shared Disk	Level Value for rvsdrestrict Command
3.4	RVSD3.4
3.2	RVSD3.2
3.1.1	RVSD3.1
2.1.1	RVSD2.1

For example, if you have some nodes running RVSD 2.1.1 and you just installed some Version 3.4 nodes that you want to coexist with them; you need to set the RVSD 3.4 functioning level to RVSD 2.1. The command to do this is:

```
/usr/lpp/csd/bin/rvsdrestrict -s RVSD2.1
```

The **rvsdrestrict** command does not dynamically change the RVSD subsystem run level across the SP—an instance of the RVSD subsystem only reacts to the setting after it is restarted. To override the level of an active RVSD subsystem, do the following on each node:

1. Stop the RSVD subsystem on all nodes including the CWS.
2. Run the **rvsdrestrict** command.
3. Restart the RVSD subsystem.

Further information related to installation and migration of VSD components can be obtained from the *PSSP for AIX: Managing Shared Disks,* SA22-7349.

3.7 HACMP

High Availability Cluster Multi-Processing (HACMP) for AIX provides a highly available environment that ensures that critical applications at your site are available to end users. As a system administrator, your job is to make sure that HACMP is stable and operational. Before undertaking any management task, you should carefully read the following documents to familiarize yourself with the facilities and capabilities of HACMP for AIX:

- *HACMP V4.3 Concepts and Facilities,* SC23-4276
- *HACMP V4.3 AIX Planning Guide,* SC23-4277
- *HACMP V4.3 AIX: Installation Guide,* SC23-4278

In Section 3.7.1, we summarize the new features and enhancements in HACMP 4.4.1. Section 3.7.2 on page 75 discusses some considerations for HACMP and the High Availability Control Workstation (HACWS). Coexistence of HACMP and High Availability Cluster Multi-Processing for AIX/Enhanced Scalability (HACMP/ES) in SP clusters is the topic of Section 3.7.3 on page 75.

3.7.1 New features and enhancements

- Recovery from Resource Group Acquisition Failure (ES only)

  With this feature, HACMP/ES keeps trying to recover a resource group until it has exhausted all the nodes. HACMP/ES can recover from multiple failures.

- Dynamic Node Priority (ES only)

  With Dynamic Node Priority in HACMP/ES, node preference is determined when fallover happens rather than at configuration time.

- Selective Fallover (ES only)

  When a local network down occurs on a node, Selective Failover recovers by moving only the affected resource groups to another node.

- Forced Down (ES only)

  This feature is included with HACMP/ES 4.4.1. It is also available in HACMP/ES 4.4.0 with APAR IY15968.

- Extended Inactive Takeover (ES only)

  Inactive takeover is now relevant to every node, while in previous versions inactive takeover was only relevant to the first node.

- Configuration Discovery for Network

  This feature provides the possible network configurations, using the addresses defined in the Object Data Manager (ODM) to identify and propose possible HACMP networks.

- Configuration Discovery for Disk

  This feature lists the volume groups that can be added to a given resource group and ensures that shared volume groups are correctly configured on every node in the owning resource group.

- Mount All Filesystems

  This feature makes a reasonable guess about unconfigured file systems or volume groups.

- Pager Notification

  Users have configured pre-event and post-event scripts to send messages to a pager. Pager Notification has been designed for easier setup with enhanced reliability.

- ATM enhancements

  HACMP can support classic IP networks and emulated LAN networks on the same ATM adapter.

- Support for Hot-Pluggable PCI Adapter

  You can replace faulty communications adapters with minimal interruption of service.

- Concurrent Mode enhancement (ES only)

  HACMP/ES concurrent mode has been validated on as many as 16 nodes, using Fibre-Channel connections to ESS.

- Cluster Single Point of Control (CSPOC) disk replacement

  Failed disks in SCSI and Serial Storage Architecture (SSA) disk arrays can be swapped without stopping the cluster.

- HACMP on AIX 5L Version 5.1

  HACMP 4.4.1 and HACMP/ES 4.4.1 can be run on AIX 5L Version 5.1 with some limitations. HACMP 4.4.0 requires IY17684.

- User Interface enhancements

  - CSPOC user password—the force change flag has been made optional.
  - The Configuring Network Module has been reorganized and corrected.
  - User-defined events (ES only)—The rules.hacmprd file has been replaced by two ODM classes, HACMPrules and HACMPude.
  - The New Event Summary feature (ES only) makes it easier to extract principle information from hacmp.out.

3.7.2 HACMP and HACWS considerations

The following requirements must be considered if you are planning to deploy HACMP and HACWS:

- Each cluster node must have AIX Version 4.3.3.25 or later for HACMP 4.4.1.
- For use of HACMP 4.4.1 in an SP cluster environment, you need PSSP 3.4 or later. An HACMP cluster does not span system partition boundaries.
- HACWS does not support HACMP Concurrent Resource Manager.
- You cannot use IPv6 aliasing with DCE, HACMP, and HACWS.
- An HACWS configuration at the PSSP 3.4 level requires the version of HACMP that can run with any level of AIX supported by PSSP 3.4.
- HACMP 4.4.1 is not compatible with any of the earlier versions, but there is a version compatibility function for migration times.

3.7.3 Coexistence in SP clusters

HACMP and HACMP/ES can coexist in a mixed system partition containing nodes running supported combinations of PSSP if the following conditions are met:

- HACMP nodes and HACMP/ES nodes cannot coexist in the same cluster.
- HACMP nodes and HACMP/ES nodes can coexist in the same mixed-system partition provided that the nodes running HACMP are in an HACMP cluster and those running HACMP/ES are in an HACMP/ES cluster.
- Any given node can only be in one cluster.

HACMP and HACMP/ES clusters do not interoperate.

Migration in SP clusters

HACMP 4.4.1 on an SP requires PSSP 3.4 and AIX Version 4.3.3 or later. Once the migration is completed, each node in an HACMP cluster must be at the same AIX and HACMP release levels, including all PTFs. You have the following migration options:

- Since HACMP and HACMP/ES cannot run in the same cluster, migration from HACMP to HACMP/ES cannot be done one node at a time without first dividing the cluster.
- Migrating from HACMP 4.2.1, 4.2.2, 4.3.0, 4.3.1, or 4.4.0 to 4.4.1 involves reinstalling HACMP on all nodes in the cluster. The version compatibility function allows you to upgrade the cluster one node at a time without taking the entire cluster offline. The new function is available after the whole cluster has been migrated.

For more information on HACMP and HACWS in SP clusters, refer to *IBM RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment,* SG22-7281, and *PSSP for AIX: Installation and Migration Guide,* GA22-7347.

4

Installation, migration, and coexistence

This chapter highlights the major changes which occur in the installation and migration process for PSSP 3.4. We also present some coexistence considerations for PSSP 3.4 programs.

This chapter is not intended as a substitute for the installation manual for PSSP 3.4. The detailed steps involved in installing or migrating to PSSP 3.4 are presented in the *PSSP for AIX: Installation and Migration Guide,* GA22-7347.

Before starting the installation process, you should carefully plan your system environment. For hardware planning, refer to *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment,* GA22-7280. For software planning, refer to *IBM RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment,* SG22-7281.

4.1 Changes and enhancements

This section highlights the changes and enhancements in the installation and migration process.

The Parallel System Support Programs for AIX Version 3, Release 4 is supported on:

- AIX Version 4.3.3
- AIX 5L Version 5.1.0.10

In order to install the base PSSP 3.4 code, you need to bring your operating system up to the maintenance level indicated in Table 4-1.

Table 4-1 AIX maintenance level required for PSSP 3.4

Operating system	Maintenance level
AIX Version 4.3.3	4.3.3.75
AIX 5L Version 5.1	AIX 5L Version 5.1.0.10

> **Note:** The IBM @server pSeries 690 requires that AIX 5L Version 5.1.C be installed on the control workstation.

Download the required maintenance-level package for your operating system from the following Web site:

`http://techsupport.services.ibm.com/server/fixes`

In migration to PSSP 3.4, the following platforms are supported (see also Section 4.4, “Migration considerations” on page 96):

- PSSP 2.4 on AIX 4.2.1 or AIX Version 4.3.3
- PSSP 3.1.1 on AIX Version 4.3.3
- PSSP 3.2 on AIX Version 4.3.3

4.1.1 Installation changes

This section describes the installation changes and the new fileset requirements for PSSP 3.4.

New prerequisites

The following run-time C++ filesets are required for installing PSSP 3.4:

- vacpp.ioc.aix43.rte 5.0.2.0 and xlC.aix43.rte.5.0.2.0 (AIX Version 4.3.3)

- vacpp.ioc.aix50.rte 5.0.2.0 and xlC.aix50.rte.5.0.2.0 (AIX 5L Version 5.1)

These filesets have their own prerequisites, and all are shipped on the PSSP 3.4 CD-ROM. Table 4-2 gives a brief description of the VisualAge C++ Licensed Program Products (LPPs).

Table 4-2 VisualAge C++ LPPs

LPP name	Description
vacpp.cmp	Contains the VisualAge C++ compiler
vacpp.ioc	Contains the VisualAge C++ IBM Open Class library run time fileset
xlC.aix43	Contains the VisualAge C++ run time filesets specific to AIX 4.3
xlC.aix50	Contains the VisualAge C++ run time filesets specific to AIX 5L 5.1
xlC.rte	Contains the VisualAge C++ run time filesets

The control workstation needs two additional filesets to support IBM @server pSeries 690 systems:

- Java130.xml4j.usr.1.3.0.0
- openCIMOM-0.61-1.aix5.1.noarch.rpm (can be found on the AIX toolbox for Linux applications CD or it can be downloaded at:

  http://www-1.ibm.com/servers/aix/products/aixos/linux/download.html

Important: To install LoadLeveler 3.1 or Parallel Environment 3.2, the bos.cpr fileset is required.

Additional filesets are required to complete a minimal fileset list for AIX 5L Version 5.1:

- rpm.rte—Red Hat Package Manager (RPM)
- bos.mp64—contains the 64-bit kernel

New path for lppsource

To permit new installation media formats, changes have been made in AIX 5L Version 5.1 to allow installation with installers other than **installp**. With AIX 5L Version 5.1, new **geninstall** and **gencopy** commands have been introduced which call **installp**, **bffcreate**, or other commands as appropriate.

Additional subdirectories have also been added in the AIX 5L Version 5.1 Network Installation Manager (NIM) LPP_SOURCE. For NIM, rather than simply put everything in the LPP_SOURCE directory, you create new directories using the **gencopy** and **bffcreate** commands. You copy the images to their respective directories based on the format of the install package.

You must copy the AIX filesets under the following directory on your control workstation (CWS):

/spdata/sys1/installp/<cw_lppsource_name>/lppsource

The **bffcreate** command places the files in the appropriate directories, which are created automatically. See Table 4-3.

Table 4-3 LPP source directories in PSSP 3.4

AIX Version	LPP source directories
AIX Version 4.3.3	/spdata/sys1/install/<cw_lppsource_name>/lppsource
AIX 5L Version 5.1	/spdata/sys1/install/<cw_lppsource_name>/lppsource/installp/ppc /spdata/sys1/install/<cw_lppsource_name>/lppsource/rpm/ppc

No changes were made to the location of PSSP LPP. The /spdata directory structure is shown in Figure 4-1 on page 81.

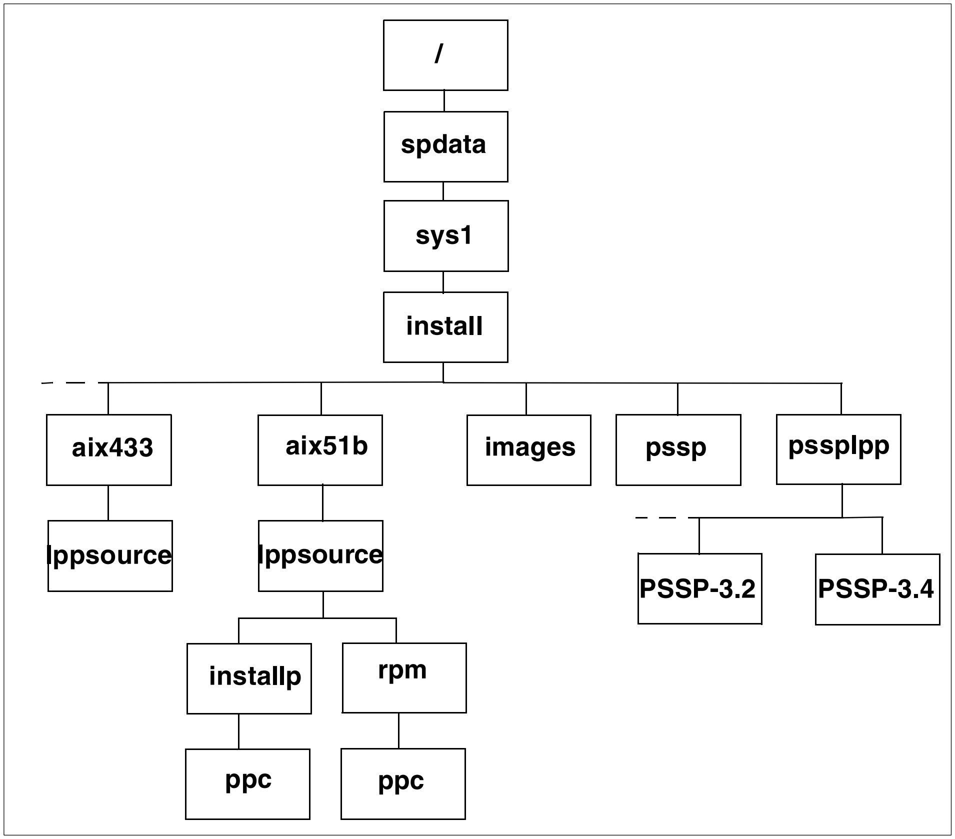


Figure 4-1 /spdata directory structure

AIX electronic licensing

Electronic licensing is now part of AIX 5L Version 5.1. An additional step in the install/migration process requires you to accept the terms and conditions of the software license agreements.

The electronic licensing enabled in AIX 5L Version 5.1 affects the SP and CES systems in two ways:

- While installing or migrating to AIX 5L Version 5.1 on a CWS from the CD, you need to accept the AIX license electronically through the GUI interface.
- PSSP 3.4 supports automated installs of AIX and PSSP on nodes in the SP or Clustered Enterprise Server (CES) systems.

To bypass user interaction with the automated installation program, PSSP performs a combination of the following steps to prevent licensing from stopping the automatic installation processes:

a. Set the environment variable ACCEPT_LICENSES to “yes.”

b. Set the environment variable NIM_LICENSE_ACCEPT to “yes.”

c. Set the accept_licenses NIM attribute to yes in a bosinst.data file and allocate that resource during the NIM bos_inst operation.

Reliable Scalable Cluster Technology packaging

Reliable Scalable Cluster Technology (RSCT) provides the high availability infrastructure daemons used in cluster configurations such as the SP and the CES. RSCT is a required PSSP component.

Changes in RSCT packaging have been made in PSSP 3.4. For AIX Version 4.3.3 environments, you must install RSCT Version 1.2.1, which is packaged with PSSP 3.4.

Because the use of RSCT in AIX 5L Version 5.1 is growing, it is now packaged with AIX 5L Version 5.1. AIX 5L Version 5.1 provides a new Version 2.2 of RSCT. As a prerequisite for installing PSSP 3.4 on your system with AIX 5L Version 5.1, RSCT 2.2.0.10 or higher is required. RSCT Version 2.2 is automatically installed during migration from AIX Version 4.3.3 to AIX 5L Version 5.1.

> **Important:** Additional PTFs for RSCT are required. Refer to the Read This First document. For the most up-to-date copy of this document, please look under READ THIS FIRST the following site:
>
> `http://www.rs6000.ibm.com/resource/aix_resource/sp_books/pssp/index.html`

4.2 Planning for installation and migration

You must plan the software installation and migration of your cluster system carefully. This section provides a brief overview of the hardware requirements.

4.2.1 Hardware planning

Since previous versions of PSSP supported SP-attached servers, this support is included in PSSP 3.4 along with improved hardware support for SP-attached servers. The following is a list of supported servers grouped according to communication protocol:

- SAMI protocol servers

- RS/6000 S70
- RS/6000 S7A
- RS/6000 S80 and S80+
- IBM @server pSeries 680

- CSP protocol servers:
  - RS/6000 H80
  - RS/6000 M80
  - IBM @server pSeries 660 (6H0, 6H1, 6M1)
- HMC protocol servers
  - IBM @server pSeries 690

These systems are supported as SP-attached servers as well as CESs and are also referred to as *nodes* in PSSP configurations.

Additional support for the SP Switch2 has been added in PSSP 3.4 in case you are planning to use a high-speed switch in your cluster system:

- You can have nodes in a system with optional connectivity to the SP Switch2, as long as there is no CSS adapter definition in the System Data Repository (SDR) for that node.
- SP Switch2 attachment adapters (PCI and MX) are introduced to cover both SP-node- and CES connectivity to the SP Switch2.
- Double-plane SP Switch2 configurations are now supported.

For further details on the new SP Switch2 support features refer to Section 2.1, “Enhancements to the RS/6000 SP Switch2” on page 26.

4.2.2 Control workstation space (CWS) requirements

As a general orientation in space allocation for your file systems, we provide in this section information related to LPP requirements for installing PSSP 3.4 on the CWS. This information actually refers to space requirements for the /spdata file system, which hosts all the LPP packages needed for installing the nodes.

You must take into consideration the sum of the following requirements:

- AIX lppsource

  Size depends on the AIX filesets included. Downloading the minimum number of filesets requires approximately 1.5 GB. We recommend downloading all the filesets, which requires about 2 GB of space in the lppsource directory.

- spimg mksysb image

  This is the image used by NIM to restore the base operating system on the nodes. The size for a typical restoration image ranges from about 204 MB to about 1.5 GB.

- The PSSP LPPs. See Appendix A, “Additional information” on page 171.
- Other optional software components: Distributed Computing Environment (DCE), LoadLeveler, and so on.

4.2.3 Security issues

PSSP 3.4 allows you to use the secure remote commands **ssh** and **scp** instead of **rsh** and **rcp**. This security enhancement can be enabled only if all the nodes in the system are at least at PSSP 3.2 and the control workstation is at PSSP 3.4. Restricted Root Access (RRA) must also be enabled.

An additional AIX authorization method for remote commands is provided under the name of *none*. Selecting none as your authorization method for remote commands requires installing and enabling the secure remote commands. Also, all the nodes in the systems must be at PSSP 3.4. When “none” is selected, all PSSP generated entries are removed from the authorization files.

4.2.4 Reserving port numbers

Components of the high-availability infrastructure such as topology services, group services, and event management need port numbers from the 10000 to 10100 inclusive range. You must therefore reserve these port numbers.

4.3 Installation

This section is not intended as an installation guide. It details the changes made in the installation process. As a reference guide, you should use the *PSSP for AIX: Installation and Migration Guide,* GA22-7347*.*

Installation involves the following steps:

1. Preparing the CWS with PSSP and AIX
2. Entering configuration information for nodes that are new to the system
3. Installing and customizing the nodes

4.3.1 Preparing the control workstation

The first step in installing the SP system is to prepare the CWS. This involves connecting your SP frames and attached servers to the serial ports of your CWS, configuring the Ethernet connections on your CWS, and verifying name resolution. You must then install the PSSP code on the CWS and apply the necessary PTFs.

The control workstation uses one serial port per SP frame for hardware monitoring and control. For attached servers, the number of serial ports you must allocate depends on the servers:

- SAMI protocol servers use two tty ports for communicating with the CWS.
- CSP protocol servers use one tty port.
- For IBM @server pSeries 690 servers you do not need to reserve a tty port, but the Hardware Management Console (HMC) must be connected to the SP Ethernet administrative LAN.

Because the CWS is the default boot/install server, it must have enough disk space to store all the data required by the Network Installation Manager (NIM). See Section 4.2.2, “Control workstation space (CWS) requirements” on page 83. You must also define the /spdata file system on the CWS and the following directory structures inside it:

- /spdata/sys1/install/<name>/lppsource

  The location for the AIX filesets. <name> is the lpp_source name for the nodes (for example, AIX_5.1). Different directories can be created for a system installing more than one AIX version on the nodes.

- /spdata/sys1/install/images

  Used for storing all AIX mksysb images.

- /spdata/sys1/install/pssplpp/<code_version>

  The location of PSSP installp filesets. Other versions of PSSP code can be installed in your system (for example, PSSP 3.4 and PSSP 3.2). <code_version> is the PSSP level.

Copying AIX and PSSP LPPs to /spdata

You must copy the AIX filesets into the lppsource directory on your CWS. For AIX Version 4.3.3, the files go into /spdata/sys1/install/<name>/lppsource. AIX 5L Version 5.1 permits installation with installers other than installp, to allow new installation media formats. See Section 4.1.1, “Installation changes” on page 78.

You can download all the AIX filesets (a very large number) or only the minimum number of required AIX filesets. Download the AIX filesets and the required AIX LPPs to /spdata/sys1/install/<name>/lppsource; these filesets and required LPPs must reside in this directory or the AIX 5L Version 5.1 directory structure. Links to filesets in other directories are not allowed.

You may want to add additional files to your lppsource directory:

- bos.acct.* are required if you plan to use PSSP accounting.
- bos.cpr.* are required to install LoadLeveler 3.1 or Parallel Environment 3.2.
- dce.* are required only if DCE is configured by PSSP anywhere on the system. You need the client portion of the DCE filesets because the installation code installs the DCE client code.
- Java130.xml4j is required for pSeries 690 servers.
- CIMOM must be copied from the AIX toolbox for Linux applications CD. It is required for pSeries 690 servers.

The Performance Toolbox for AIX, Agent Component (PAIDE) is required. You need to install the correct AIX PAIDE level (perfagent.tools) on the CWS and copy it to all the lppsource directories. The perfagent.tools fileset is part of AIX Version 4.3.3 and AIX 5L Version 5.1, and the perfagent level depends upon the AIX and PSSP levels (see Table 4-4).

Table 4-4 perfagent filesets

AIX	PSSP	perfagent fileset
AIX 4.2.1	PSSP 2.4	perfagent server 2.2.1.x, where x≥2
AIX Version 4.3.3	PSSP 2.4	perfagent.server 2.2.33
AIX Version 4.3.3	PSSP 3.1.1	perfagent.tools 2.2.33
AIX Version 4.3.3	PSSP 3.2	perfagent.tools 2.2.33
AIX Version 4.3.3	PSSP 3.4	perfagent.tools 2.2.33
AIX 5L Version 5.1	PSSP 3.4	perfagent.tools 5.1.0

After you install PAIDE, you must copy the PSSP installp images in the PSSP lppsource directory:

/spdata/sys1/install/pssplpp/PSSP-3.4

Installing PSSP 3.4 on the CWS

Install the PSSP prerequisite files:

- bos.net files
- perfagent.tools
- Run time files
- RSCT files

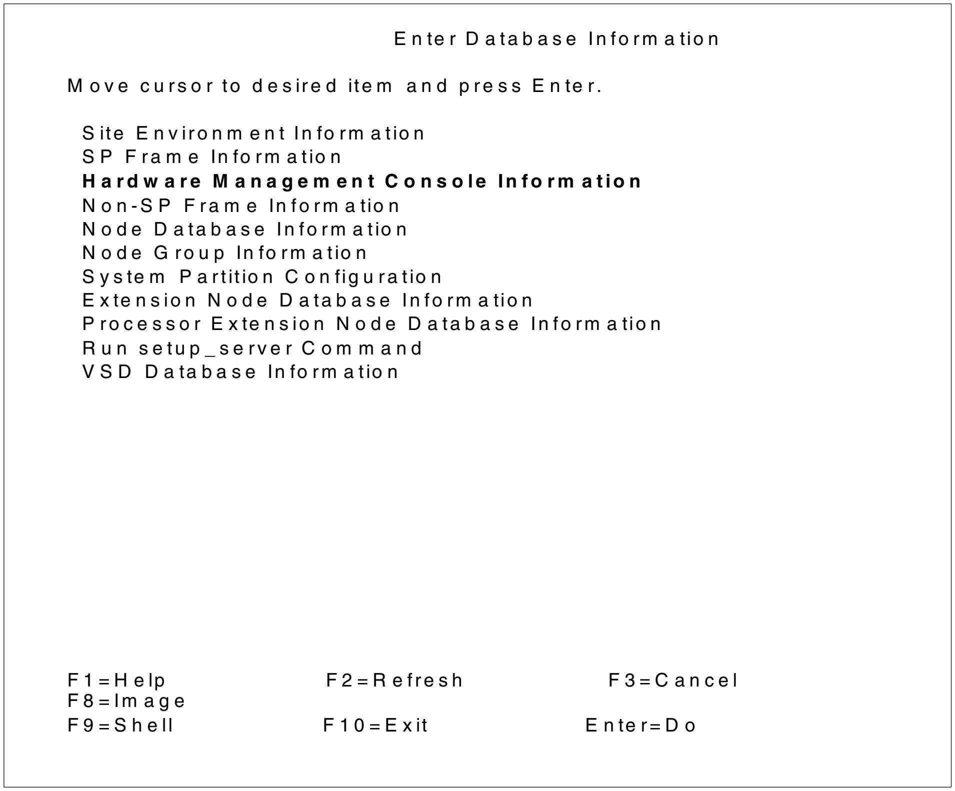
If you have pSeries 690 systems in your environment you must install the additional filesets required by the pSeries 690 server. See “Copying AIX and PSSP LPPs to /spdata” on page 85.

Continue with the procedure for installing the PSSP 3.4 package on the CWS, set the authentication methods for remote commands on the CWS, and complete the CWS preparation by running the install_cw script.

4.3.2 Entering configuration information

After preparing the control workstation, you can begin configuring the nodes. Based on the information you have collected (see the worksheets provided in *IBM RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment,* SG22-7281), you must enter the configuration data in the SDR. You can do this by using SMIT panels or the corresponding PSSP commands.

To start entering information in the SDR, issue the SMIT fast path command **`smitty enter_data`**. The SMIT panels are enhanced to allow you to introduce the configuration data for the IBM @server pSeries 690. See Figure 4-2 on page 88.



```
                              Enter Database Information

  Move cursor to desired item and press Enter.

    Site Environment Information
    SP Frame Information
    Hardware Management Console Information
    Non-SP Frame Information
    Node Database Information
    Node Group Information
    System Partition Configuration
    Extension Node Database Information
    Processor Extension Node Database Information
    Run setup_server Command
    VSD Database Information









  F1=Help                  F2=Refresh                F3=Cancel
  F8=Image
  F9=Shell                 F10=Exit                Enter=Do
```

Figure 4-2 SMIT panel for entering data in the SDR

Entering site environment information

At this point you must enter information related to your environment. You can use the **spsitenv** command or the SMIT panels. PSSP 3.4 provides additional security settings.

If you plan to use the secure remote commands (**ssh** and **scp**) instead of standard AIX remote commands (**rsh** and **rcp**) and have already installed them on your CWS, you must set the remote command method to secrshell in the site environment configuration. You may also customize the location of the secure remote commands. As a requirement for using the secure remote commands, you must also enable the restricted root access (RRA) mode. For further details, see Section 5.2, “Restricted root access (RRA)” on page 103.

Note: To list the environment settings, issue the `splstdata -e` command.

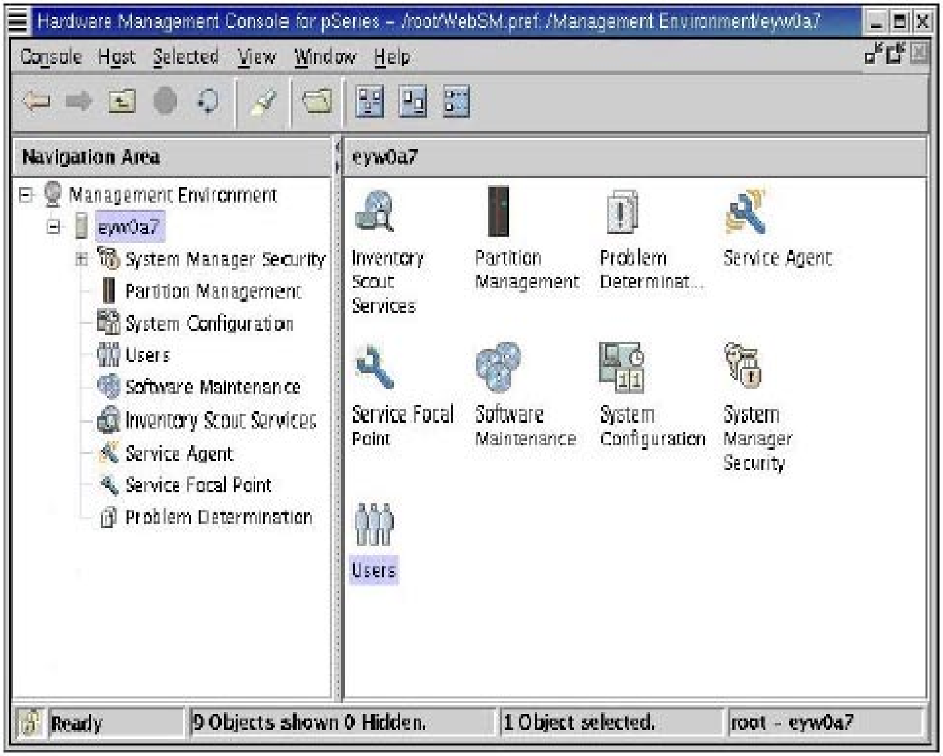
Entering Hardware Management Console (HMC) information

If your system contains IBM @server pSeries 690 servers, you must enter HMC information in the SDR. The HMC performs all the management functions for the server. You must properly set up your pSeries690 using the HMC before you can attach it to the SP system or to a CES system.

Preparing the HMC

In order to perform the operations needed to configure the pSeries 690 system, refer to the *Hardware Management Console for pSeries: Operation Guide,* SA38-0590.

When you log in to the HMC, the HMC management panel opens and the management environment is already selected (see Figure 4-3). This is the starting point you need for configuring your pSeries 690 system.



Ready | 9 Objects shown 0 Hidden. | 1 Object selected. | root - eyw0a7

Figure 4-3 HMC management panel

The following are the steps you must perform on the HMC to integrate the IBM @server pSeries 690 with the PSSP managed system:

1. Configure the HMC network. The HMC must be installed and configured to operate on the SP Ethernet administrative LAN network.
2. Define a user ID with the role of system administrator. The CWS uses this user ID in SDR data entry. On the HMC, the default user ID is hmcroot. It is recommended that this be changed.
3. Verify that the HMC recognizes the pSeries 690. Use the Partition Management interface from the Hardware Management Console for pSeries graphical application.
4. View the system name associated with the global server, and change the default it you wish. This information is needed in SDR frame information entry.

> **Important:** Instead of the system name, it is recommended to use the central electronic complex (CEC) name. But the system name is fine to use. CEC name is not system name technically; CEC name is a domain name for this pSeries 690 complex.

5. Select the power-on mode you want for the system:
   a. Full-system partition mode (when the system is used as an SMP)
   b. Logical partition standby mode
   c. Physical partition mode

   Select b or c when partitioning your system. Use the operator guide to create the partitions (LPARs) at this time—they do not need to be installed or activated before you enter the SDR data.

   LPARs may use different profiles when they are activated (powered on), but one of them is the default profile used by the CWS when powering on the associated node.

> **Note:** As each partition is created, note the partition ID used by PSSP to assign an SP slot number and node number to the node associated with that partition.

Entering HMC data

Before entering data into the SDR, check the network connectivity to the HMC using `ping` *hmc_host*, where hmc_host is the name or IP address of the HMC.

You must define on the CWS the user created in the step “Preparing the HMC” on page 89, in order for hardmon to be able to establish a remote client session with the HMC. Issue the following command:

```
/usr/lpp/ssp/bin/sphmcid hmc_hostname user ID
```

where:

- hmc_hostname is the host name or IP address of the HMC.
- user ID is the system-administrator user ID you created.

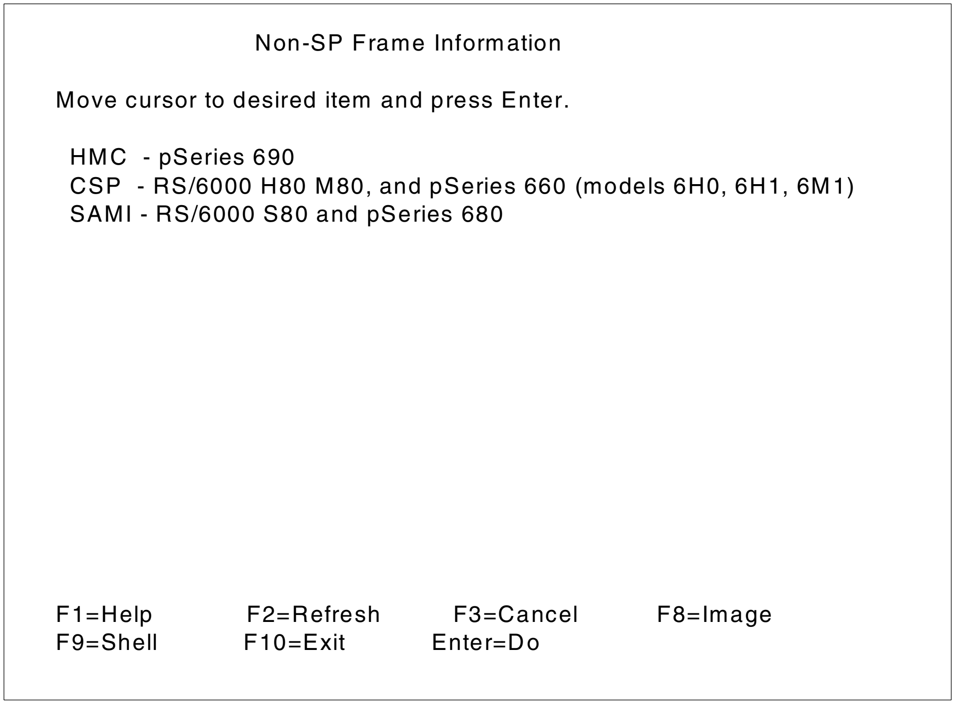
Entering information about frames with multiple NSBs

PSSP 3.4 supports frames containing multiple Network Switch Boards (NSBs). This configuration is supported only for SP Swich2 systems.

To enter information about such frames in the SDR, use the **`spframe -m`** command. For example, for a multiple-NSB system contained in an SP frame and attached to the CWS /dev/tty1 port, enter **`spframe -m 1 1 /dev/tty1`**. The frame will be frame 1.

Entering non-SP frame information

SP-attached servers and CESs need frame information in the SDR. The enhanced PSSP 3.4 allows you to enter pSeries 690 non-SP frames in the SDR and also provides new SMIT panels. You can display the appropriate SMIT panel by issuing the SMIT fast path command **`smitty non_sp_frame_menu`** (see Figure 4-4 on page 92).



```
                    Non-SP Frame Information

 Move cursor to desired item and press Enter.

  HMC  - pSeries 690
  CSP  - RS/6000 H80 M80, and pSeries 660 (models 6H0, 6H1, 6M1)
  SAMI - RS/6000 S80 and pSeries 680











 F1=Help          F2=Refresh       F3=Cancel        F8=Image
 F9=Shell         F10=Exit         Enter=Do
```

Figure 4-4 SMIT menu for entering non-SP frame information

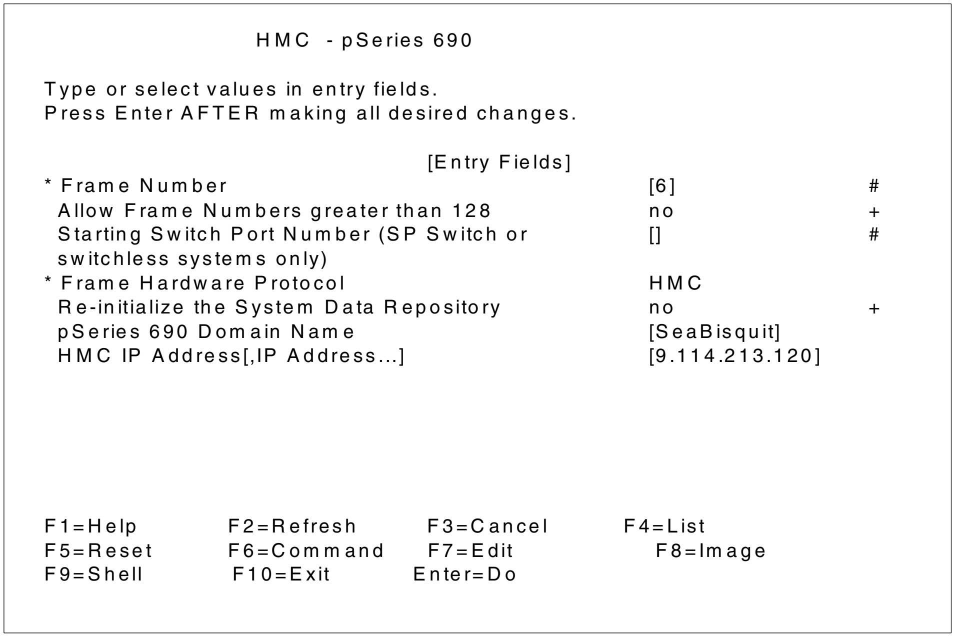
There are three protocols available:

HMC Used for communicating with pSeries 690 systems through the HMC.

CSP Serial communication protocol between the CWS and the following systems: RS/6000 H80, M80, and IBM @server pSeries 660 servers (6H0, 6H1,6M1).

SAMI Serial communication protocol for RS/6000 S70, S7A, and S80 or IBM @server pSeries 680 servers.

Select **HMC—pSeries 690** from the menu, and enter the information for your pSeries 690 server as illustrated in Figure 4-5 on page 93.



```
                    HMC  - pSeries 690

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                   [Entry Fields]
* Frame Number                                      [6]              #
  Allow Frame Numbers greater than 128              no               +
  Starting Switch Port Number (SP Switch or         []               #
  switchless systems only)
* Frame Hardware Protocol                           HMC
  Re-initialize the System Data Repository          no               +
  pSeries 690 Domain Name                           [SeaBisquit]
  HMC IP Address[,IP Address...]                    [9.114.213.120]






F1=Help        F2=Refresh      F3=Cancel       F4=List
F5=Reset       F6=Command      F7=Edit            F8=Image
F9=Shell        F10=Exit      Enter=Do
```

Figure 4-5   New SMIT panel for entering pSeries 690 frame information

You can also use the **spframe** command instead of the SMIT menus. To list the frame information for your system, use the **splstdata** command, as shown in Example 4-1.

Example 4-1   Listing the frames installed in your system

```
# splstdata -f
                  List Frame Database Information

frame# tty              s1_tty           frame_type      hardware_protocol
control_ipaddrs domain_name
------ ---------------- ---------------- --------------- ------------------
--------------- -----------
    1 /dev/tty0      ""           switch        SP             ""
""
    2 /dev/tty1      ""           switch        SP             ""
""
     6 ""               ""               ""              HMC
9.114.213.120   SeaBisquit
```

Entering required node information

In this step, you proceed to acquire physical adapter information for the nodes.

> **Attention:** Do not issue the **`spadapter_loc`** command on a node that is running. It stops the node and acquires the physical adapter location information and hardware Ethernet addresses.

As an example, to acquire physical adapter information for a system with one frame and two nodes, enter **`spadaptr_loc 1 1 2`**. The output of this command is shown in Example 4-2.

Example 4-2 Physical adapter information for a node

```
/usr/lpp/ssp/bin/spadaptr_loc 1 1 2
Acquiring adapter physical location codes for node 1
Acquiring adapter physical location codes for node 5
node# adapter_type physical_location_code MAC_address
----- ------------ ---------------------- ------------
    1 Ethernet     U1.1-P2-I1/E1 0004ACEC12CB
    1 Ethernet     U1.1-P2-I5/E1 0004AC7CE5F3
    5 Ethernet     U1.5-P2-I1/E1 0004ACEC12C7
    5 Ethernet     U1.5-P2-I5/E1 0004AC7CE5F4
```

> **Note:** pSeries 690 systems do not require an en0 adapter for an SP Ethernet administrative LAN. We therefore recommend using the physical location codes obtained with the **`spadaptr_loc`** command to define the adapter used for an SP Ethernet LAN.

Example 4-3 uses the physical location code for node 1 acquired in Example 4-2 to define that node's SP Ethernet adapter for the administrative LAN of the SP system.

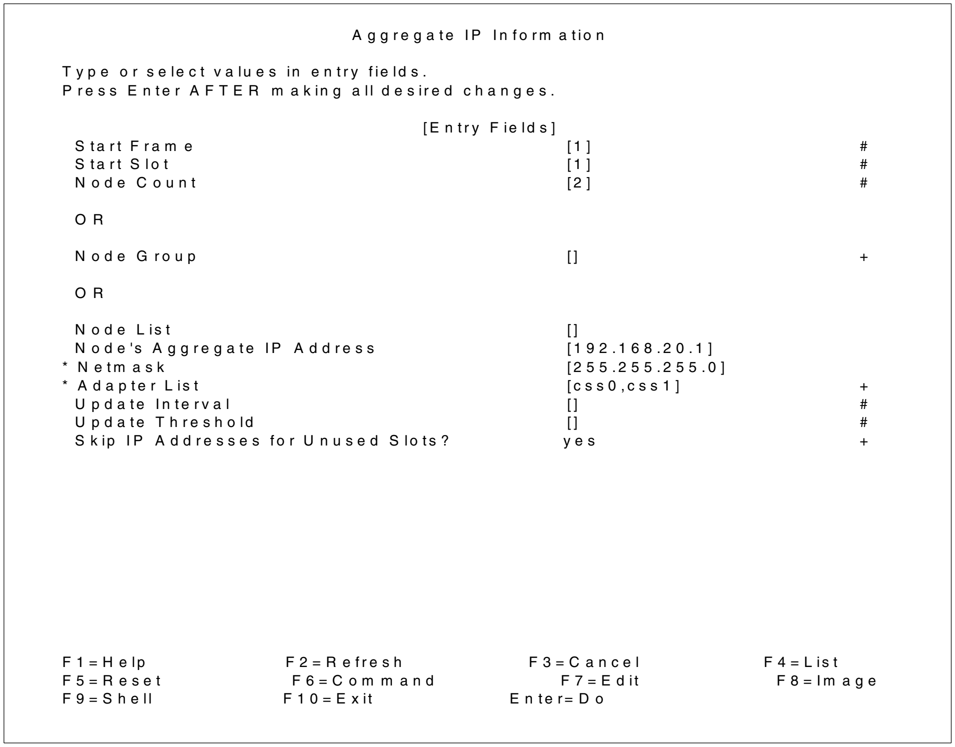
Example 4-3 Defining an SP Ethernet adapter using physical location codes

```
/usr/lpp/ssp/bin/spadaptrs -P 'U1.1-P2-I5/E1' -t 'tp' -d 'auto' -f 'auto' 6 1\
1 en 192.168.14.1 255.255.255.0
```

Configuring the Aggregate IP interface

The IP Aggregate interface can be configured only for systems using the SP Switch2. In order to configure the IP Aggregate interface you must have defined the switch adapters in the SDR. For more information about the IP Aggregate, refer to Section 3.4.2, "Aggregate IP addressing" on page 55.

To configure the IP aggregate interface, issue the **spaggip** command or use SMIT as in Figure 4-6. This will create the pseudodevice ml0 on the specified nodes after installation. If you plan to configure the IP Aggregate interface after the nodes are installed, follow the same procedure for entering data in the SDR, and then customize the nodes.



```
                         Aggregate IP Information

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                  [Entry Fields]
   Start Frame                                    [1]                  #
   Start Slot                                     [1]                  #
   Node Count                                     [2]                  #

   OR

   Node Group                                     []                   +

   OR

   Node List                                      []
   Node's Aggregate IP Address                    [192.168.20.1]
 * Netmask                                        [255.255.255.0]
 * Adapter List                                   [css0,css1]          +
   Update Interval                                []                   #
   Update Threshold                               []                   #
   Skip IP Addresses for Unused Slots?            yes                  +







  F1=Help             F2=Refresh           F3=Cancel          F4=List
  F5=Reset            F6=Command           F7=Edit            F8=Image
  F9=Shell            F10=Exit             Enter=Do
```

Figure 4-6   smitty sp_agg_dialog

Configuring the security services

At this step you can configure and customize the SP authentication and authorization methods you have selected.

You can use the **spsetauth** command to set the security capabilities to be installed on the nodes in the partition you specify. You can choose one or more of the following security options:

- Kerberos V4
- Distributed Computing Environment (DCE)
- AIX standard

As indicated in Section 4.2.3, “Security issues” on page 84, PSSP 3.4 provides a new option for the remote commands authorization methods in a system partition called none. If the none option is selected, no other authorization methods can be selected for the system partition. This setting does not generate entries in the .k5login, .klogin and .rhosts files. For further details, refer to Chapter 5, “Security enhancements” on page 101.

4.3.3 Installing and customizing the nodes

There are no changes in the node installation and customization steps in PSSP 3.4.

4.4 Migration considerations

PSSP 3.4 is supported in both AIX Version 4.3.3 and AIX 5L Version 5.1 environments. Since migrating to PSSP 3.4 on AIX 5L Version 5.1 is the final target, we refer to the migration of PSSP 3.4 on AIX Version 4.3.3 as an intermediate stage. Note also that some PSSP 3.4-related products such as LoadLeveler 3.1 and Parallel Environment 3.2 require AIX 5L Version 5.1.

When migrating to PSSP 3.4 and AIX 5L Version 5.1 on the CWS, PSSP must first be migrated to PSSP 3.4 on AIX Version 4.3.3, and then AIX must be migrated to AIX 5L Version 5.1.

For the nodes, direct migration to PSSP 3.4 and AIX 5L Version 5.1 is allowed. If you plan to use AIX Version 4.3.3, you can migrate to PSSP 3.4 on AIX Version 4.3.3 and migrate to AIX Version 5.1 later. See Table 4-6 on page 97 for the PSSP-related software available for PSSP 3.4 and AIX Version 4.3.3.

4.4.1 Migration paths

You can follow two paths to complete your migration to PSSP 3.4 on AIX 5L Version 5.1:

- The intermediate migration path uses two stages to migrate to PSSP 3.4 and AIX 5L Version 5.1. See “Intermediate migration path” on page 97.
- The direct migration path uses one stage to migrate to PSSP 3.4 and AIX 5L Version 5.1. See “Direct migration path” on page 97.

Before you migrate any node in your system, you must migrate the CWS to the latest level of AIX and PSSP represented on the nodes you want to serve. You can use a single node or a group of your nodes for test purposes to migrate to PSSP 3.4 and the desired AIX level. The following PSSP and AIX releases are supported as starting levels:

- PSSP 2.4 on AIX 4.2.1 or AIX Version 4.3.3
- PSSP 3.1.1 on AIX Version 4.3.3
- PSSP 3.2 on AIX Version 4.3.3

Intermediate migration path

The control workstation and the nodes can take this path to complete the migration to PSSP 3.4 on AIX 5L Version 5.1. Table 4-5, Table 4-6, and Table 4-7 detail the migration path for PSSP, AIX, and PSSP-related software.

Table 4-5 Starting levels

PSSP	AIX	GPFS	LL	PE
2.4	4.2.1	none	1.3	2.3
2.4	4.3.3	none	1.3	2.3
3.1.1	4.3.3	1.2	2.1	2.4
3.2	4.3.3	1.3/1.4	2.2	3.1

Table 4-6 Intermediate levels

PSSP	AIX	GPFS	LL	PE
3.4	4.3.3	1.3/1.4/1.5	2.2	3.1

Attention: To complete the migration process to PSSP 3.4 and AIX 5L Version 5.1, you should first migrate GPFS to Version 1.5. Then migrate AIX, PE, and LoadLeveler simultaneously to the levels indicated in Table 4-7.

Table 4-7 Final migration levels

PSSP	AIX	RSCT	RVSD	GPFS	LL	PE	Security
3.4	5.1	2.2 (with AIX)	3.4	1.5	3.1	3.2	No changes

Direct migration path

Direct migration is a procedure for bringing the nodes to PSSP 3.4 and AIX 5L Version 5.1 without utilizing the intermediate migration stage. You can use direct migration to PSSP 3.4 and AIX 5L Version 5.1 only for the nodes.

Note: You cannot apply direct migration from PSSP 2.4 on AIX 4.2.1 to PSSP 3.4 on AIX 5L Version 5.1. You must migrate your system to PSSP 3.4 on AIX Version 4.3.3 and then migrate from AIX Version 4.3.3 to AIX 5L Version 5.1. The two steps cannot be combined.

Table 4-8 Starting levels

PSSP	AIX	GPFS	LL	PE
2.4	4.3.3	none	1.3	2.3
3.1.1	4.3.3	1.2	2.1	2.4
3.2	4.3.3	1.3/1.4	2.2	3.1

Table 4-9 Final migration levels

PSSP	AIX	GPFS	LL	PE
3.4	5.1	1.5	3.1	3.2

LoadLeveler 3.1 includes the warmstart function after migration from Version 2.2 to Version 3.1. LoadLeveler 3.1 and PE 3.2 run on PSSP 3.4 with AIX 5L Version 5.1; they do not run on AIX Version 4.3.3. Also, take into consideration that no previous version of GPFS supports AIX 5L Version 5.1. Refer to Table 4-8 and Table 4-9 for information on the starting and ending software levels for PSSP, AIX, and PSSP-related software.

4.4.2 CWS migration

Perform he following steps to migrate the CWS to PSSP 3.4 and AIX 5L Version 5.1:

1. Migrate to AIX Version 4.3.3 and upgrade to the necessary maintenance level. See Section 4.1, “Changes and enhancements” on page 78.
2. Migrate the CWS to PSSP 3.4 on AIX Version 4.3.3.
3. Migrate to AIX 5L Version 5.1. During the migration, the PSSP 64-bit libraries are replaced with the new 64-bit libraries, which are incompatible with AIX Version 4.3.3.
4. Update vacpp.ioc.aix50.rte 5.0.2.0; for xlC the update takes place automatically with AIX migration.

4.4.3 Node migration

You can use the following options:

- Customize the node for migration from PSSP 2.4, 3.1.1, or 3.2 to PSSP 3.4 on AIX Version 4.3.3.
- Migrate the node from any PSSP version on AIX Version 4.3.3 to PSSP 3.4 on AIX 5L Version 5.1.0.10.

From PSSP 2.4 on AIX 4.2.1, you should perform the following steps:

1. Migrate to PSSP 3.4 on AIX Version 4.3.3
2. Migrate to PSSP 3.4 on AIX 5L Version 5.1.

4.5 Coexistence

PSSP 3.4 can coexist with PSSP 2.4 and later. Table 4-10 shows the levels of AIX and PSSP software supported by PSSP in the same system partition.

Table 4-10 Coexistence levels for PSSP components

PSSP	AIX	RSCT	RVSD	GPFS	LL	PE
2.4	4.2.1	2.4	2.1.1	none	1.3	2.3
2.4	4.3.3	2.4	2.1.1	none	1.3	2.3
3.1.1	4.3.3	1.1	3.1	1.2	2.1	2.4
3.2	4.3.3	1.2	3.2	1.3/1.4	2.2	3.1
3.4	4.3.3	1.2.1	3.4	1.3/1.4/1.5	2.2	3.1
3.4	5.1	2.2	3.4	1.5	3.1	3.2

Note that different RSCT packages are used for AIX Version 4.3.3 and AIX 5L: RSCT 2.2 is now included with the AIX package. See Section 3.2, "AIX 5L Version 5.1" on page 48.

LoadLeveler 2.2 running on both PSSP 3.2 and PSSP 3.4 in a single LoadLeveler cluster may interoperate with LoadLeveler 3.1 running on PSSP 3.4. In this environment, the new functions of LoadLeveler 3.1 such as gang scheduling and checkpoint/restart are not available. A PTF is required for LoadLeveler 2.2 to properly interoperate with LoadLeveler 3.1 in a DCE environment. To Find more information about LoadLeveler 3.1, see Chapter 7, "LoadLeveler 3.1" on page 137.

PE 3.1 runs on PSSP 3.2 or PSSP 3.4, but a single parallel job cannot use nodes with different PSSP versions. See more information about PEt 3.2 in Chapter 6, “Parallel programming” on page 115.

PSSP 3.4 supports GPFS Versions 1.3 and 1.4 with AIX Version 4.3.3. GPFS 1.5 is the native version for PSSP 3.4 and runs in both AIX Version 4.3.3 and AIX 5L Version 5.1 environments. GPFS releases do not interoperate, but they may coexist inside a GPFS cluster. See Chapter 8, “GPFS 1.5” on page 155.

5

Security enhancements

This chapter contains information about the new security features implemented in PSSP 3.4.

5.1 Prerequisites

As a common prerequisite for all new (and old) security features, we recommend you read the PSSP planning guide *IBM RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment,* SG22-7281, especially the chapter “Planning for security."

Since this chapter discusses several security features introduced in PSSP 3.4, we state the requirements for each feature separately:

- Restricted root access (RRA)

  PSSP 3.2 or higher is needed on the nodes and on the control workstation (CWS), and the restrict_root_rcmd attribute in the SP_restricted SDR class should be set to true.

- Secure remote command process (secrshell)

  PSSP 3.4 or higher must be installed on the CWS, the nodes must be installed with at least PSSP 3.2, and RRA must be enabled. You also need a distribution of secrshell and working installations of Zlib and OpenSSL, which are prerequisites for secure remote shell. Finally, you need a C compiler (which is still a PSSP prerequisite) to compile the secrshell, Zlib, and Open SSL sources.

- Using none as an AIX authorization remote command method

  To enable this feature, you need PSSP 3.4 or higher on all nodes in the partition and also on the CWS. Set the auth_methods attribute in the Syspar SDR class to none*,* and enable secrshell.

- Firewalled RS/6000 SP system

  - Software requirements

    - All SP system nodes and the CWS must be at PSSP 3.4 or higher.
    - For AIX Version 4.3.3 systems, fix U475875.bff via IY20117 is required.
    - For AIX 5L systems, fix U476341.bff via IY18836 is required.

  - Firewall software requirements

    - Make sure any firewall application you select runs on the level of AIX you are using.

  - Optional software

    - IBM Distributed Computing Environment Version 3.1 or higher for AIX and Solaris, plus the latest cumulative PTFs.
    - Kerberos V4 supplied with PSSP, or equivalent.

- PSSP security requirements
  - *One* of the following PSSP Trusted Services security methods must be enabled: compat, dce, or dce and compat (dce:compat).
  - Restricted Root Access (RRA) *must* be enabled.
  - Secure remote command process must be enabled.

In the following sections, we describe these features in more detail.

5.2 Restricted root access (RRA)

Restricted root access (RRA) was introduced with PSSP 3.2 (see the IBM Redbook *PSSP 3.2: RS/6000 SP Software Enhancements,* SG24-5673*)*, but it needs to be described here as other new features rely on it.

5.2.1 Overview

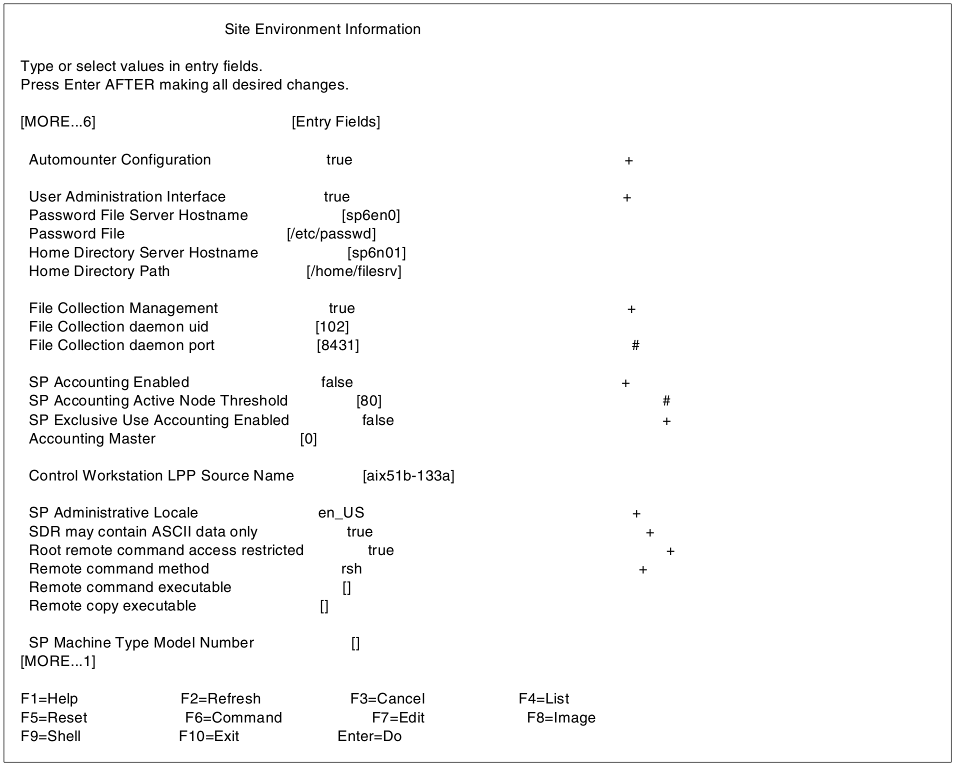
When RRA is activated, the root user is not automatically authorized to issue a **rsh** or **rcp** command from any SP node—including boot-install servers (BISs)—back to the CWS or to another SP node.

The system administrator may still manually set up the SP system to allow the root user on a node to issue **rsh** or **rcp** to CWS or to another SP node. When RRA is active, any such actions can only be run from the CWS or from SP nodes explicitly configured to authorize them. RRA restricts root **rsh** and **rcp** authorizations from the nodes to the CWS, but permits CWS-to-node **rsh** and **rcp** access.

To activate RRA, change the restrict_root_cmd attribute to true in the SP_restricted class of the System Data Repository (SDR). This can only be done by user root on the CWS. Use the Site Environment SMIT panel shown in Figure 5-1 on page 104, or use the **spsitenv** command.

In some situations (such as when you use a BIS), the PSSP uses commands that still need **rsh** or **rcp**. All the commands that need **rsh** or **rcp** check the restrict_root_cmd attribute and use the sysctl method when RRA is enabled

Note: The sysctl daemon becomes a critical part of an SP environment when you are operating with RRA enabled. Make sure that sysctl uses the same default TCP/IP port number on all nodes and on the CWS.



```
                        Site Environment Information

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

[MORE...6]                                  [Entry Fields]

  Automounter Configuration                  true                          +

  User Administration Interface              true                          +
  Password File Server Hostname              [sp6en0]
  Password File                              [/etc/passwd]
  Home Directory Server Hostname             [sp6n01]
  Home Directory Path                        [/home/filesrv]

  File Collection Management                 true                          +
  File Collection daemon uid                 [102]
  File Collection daemon port                [8431]                        #

  SP Accounting Enabled                      false                         +
  SP Accounting Active Node Threshold        [80]                          #
  SP Exclusive Use Accounting Enabled        false                         +
  Accounting Master                          [0]

  Control Workstation LPP Source Name        [aix51b-133a]

  SP Administrative Locale                   en_US                         +
  SDR may contain ASCII data only            true                          +
  Root remote command access restricted      true                          +
  Remote command method                      rsh                           +
  Remote command executable                  []
  Remote copy executable                     []

  SP Machine Type Model Number               []
[MORE...1]

F1=Help             F2=Refresh          F3=Cancel           F4=List
F5=Reset            F6=Command          F7=Edit             F8=Image
F9=Shell            F10=Exit            Enter=Do
```

Figure 5-1 Restricted root access

RRA fundamentally changes authorization methods, and this has consequences for the following PSSP components as well as some AIX software programs that are used in an SP environment:

- Coexistence
- Virtual Shared Disk (VSD) and General Parallel File System (GPFS)
- High Availability Cluster Multi-Processor (HACMP)
- High Availability Control Workstation (HACWS)
- Boot-install servers (BISs)
- Ecommands
- System management commands

For a complete list of limitations see Chapter 2 “Security Features of the SP” in the *PSSP for AIX: Administration Guide,* SA22-7348, and Chapter 6 “Planning for Security” in *IBM RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment,* SG22-7281.

The RRA concept also impacts SP user management. For example, RRA does not prevent root from logging in to another node using the **telnet** command, from using **su** to another user and exploiting that user's remote-command permissions, or from using other AIX and PSSP remote-command capabilities to access other nodes or the CWS.

Therefore it is obvious that using RRA with the minimal security combination none/std is of little value. Instead, we suggest you use the PSSP security methods compat, dce or dce:compat.

In the following section we describe the authorization routines and files involved in the RRA concept.

5.2.2 Authorization routine and files

Whenever you change the restrict_root_cmd attribute the updauthfiles script is activated. The script updates and, if necessary, creates the AIX authorization files for root on the CWS and on the nodes.

When RRA is enabled, the script removes all SP-generated entries in the remote authorization files (/.klogin, /.rhosts, and /.k5login) on the nodes.

For more information, refer to the chapter "Planning for security" in *IBM RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment,* SG22-7281.

Several authorization files are generated in the SP environment based on the authentication method activated for the SP trusted services. The following sections each give examples of these files, for a case where the RRA mode is not activated and for one where it is.

Kerberos 4 as an authorization method

If you are in compat mode and not in RRA mode, you get a /.klogin file on the CWS or the nodes like Example 5-1 on page 105.

Example 5-1 /.klogin file with RRA disabled

```
rcmd.sp6en0@MSC.ITSO.IBM.COM
root.admin@MSC.ITSO.IBM.COM
root.SPbgAdm@MSC.ITSO.IBM.COM
rcmd.sp6n01@MSC.ITSO.IBM.COM
rcmd.sp6n03@MSC.ITSO.IBM.COM
rcmd.sp6n05@MSC.ITSO.IBM.COM
rcmd.sp6n07@MSC.ITSO.IBM.COM
rcmd.sp6n09@MSC.ITSO.IBM.COM
rcmd.sp6n10@MSC.ITSO.IBM.COM
rcmd.sp6n11@MSC.ITSO.IBM.COM
rcmd.sp6n12@MSC.ITSO.IBM.COM
rcmd.sp6n13@MSC.ITSO.IBM.COM
rcmd.sp6n14@MSC.ITSO.IBM.COM
```

Once RRA is enabled, and still in compat mode, the /.klogin file is pruned on the nodes to look like Example 5-2, so that only the CWS can send remote commands to the nodes.

Example 5-2 /.klogin file with RRA enabled

```
rcmd.sp6en0@MSC.ITSO.IBM.COM
```

DCE as an authorization method

Before RRA is enabled the /.k5login files look like Example 5-3.

Example 5-3 /.k5login file with RRA disabled

```
hosts/sp6en0.msc.pok.ibm.com/self@sp6dcecell
ssp/sp6en0.msc.pok.ibm.com/spbgroot@sp6dcecel
hosts/sp6n01.msc.pok.ibm.com/self@sp6dcecell
ssp/sp6n01.msc.pok.ibm.com/spbgroot@sp6dcecel
hosts/sp6n03.msc.pok.ibm.com/self@sp6dcecell
ssp/sp6n03.msc.pok.ibm.com/spbgroot@sp6dcecel
hosts/sp6n05.msc.pok.ibm.com/self@sp6dcecell
ssp/sp6n05.msc.pok.ibm.com/spbgroot@sp6dcecel
hosts/sp6n07.msc.pok.ibm.com/self@sp6dcecell
ssp/sp6n07.msc.pok.ibm.com/spbgroot@sp6dcecel
hosts/sp6n09.msc.pok.ibm.com/self@sp6dcecell
ssp/sp6n09.msc.pok.ibm.com/spbgroot@sp6dcecel
hosts/sp6n10.msc.pok.ibm.com/self@sp6dcecell
ssp/sp6n10.msc.pok.ibm.com/spbgroot@sp6dcecel
hosts/sp6n11.msc.pok.ibm.com/self@sp6dcecell
ssp/sp6n11.msc.pok.ibm.com/spbgroot@sp6dcecel
hosts/sp6n12.msc.pok.ibm.com/self@sp6dcecell
ssp/sp6n12.msc.pok.ibm.com/spbgroot@sp6dcecel
hosts/sp6n13.msc.pok.ibm.com/self@sp6dcecell
ssp/sp6n13.msc.pok.ibm.com/spbgroot@sp6dcecel
hosts/sp6n14.msc.pok.ibm.com/self@sp6dcecell
ssp/sp6n14.msc.pok.ibm.com/spbgroot@sp6dcecel
```

When the RRA is enabled the /.klogin files looks like Example 5-4.

Example 5-4 /.k5login file with RRA enabled

```
hosts/sp6en0.msc.pok.ibm.com/self@sp6dcecell
ssp/sp6en0.msc.pok.ibm.com/spbgroot@sp6dcecel
```

std as the standard authorization method

Before RRA is enabled the /.rhosts file looks like Example 5-5.

Example 5-5 /.rhosts file with RRA disabled

```
sp6en0
sp6en0.msc.pok.ibm.com
sp6n01.msc.pok.ibm.com
sp6n03.msc.pok.ibm.com
sp6n05.msc.pok.ibm.com
sp6n07.msc.pok.ibm.com
sp6n09.msc.pok.ibm.com
sp6n10.msc.pok.ibm.com
sp6n11.msc.pok.ibm.com
sp6n12.msc.pok.ibm.com
sp6n13.msc.pok.ibm.com
sp6n14.msc.pok.ibm.com
```

When RRA is enabled the /.rhosts file looks like Example 5-6.

Example 5-6 /.rhosts file with RRA enabled

```
sp6en0
sp6en0.msc.pok.ibm.com
```

5.3 Secure remote command processes

The PSSP code allows you to select a secure remote command method as an alternative to traditional remote shell commands.

5.3.1 Overview

In addition to RRA (see Section 5.2, “Restricted root access (RRA)” on page 103), PSSP 3.4 also permits removal of root-level PSSP dependencies on root using **rsh** or **rcp** from the control workstation (CWS) to the node.

The design is independent of the version or configuration of secrshell. It has no dependency on the AIX **rsh** or **rcp** commands and does not prevent the root user or any other user from using these AIX commands.

PSSP 3.4 does not provide a secure remote program. You are responsible for obtaining a copy of existing versions of secure remote programs.

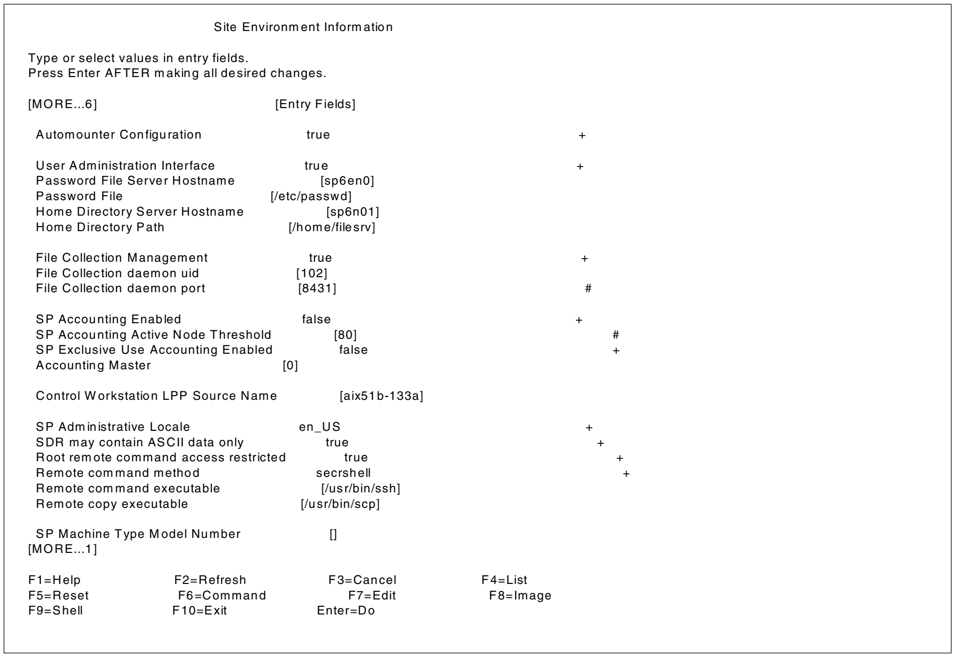
5.3.2 Setting up secrshell

The prerequisites for secrshell are described in Section 5.1, “Prerequisites” on page 102.

If you want to use secrshell you need to define it in the SDR with the **spsitenv** command, as shown in Example 5-7, or by using the SMIT site_env_dialog panel, as shown in Figure 5-2.

Example 5-7 setting up secrshell with spsitenv

```
spsitenv restrict_root_rcmd=true rcmd_pgm=secrshell dsh_remote_cmd=<remote
        command executable> remote_copy_cmd=<remote copy executable>
```



```
                              Site Environment Information

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

[MORE...6]                                  [Entry Fields]

  Automounter Configuration                   true                          +

  User Administration Interface               true                          +
  Password File Server Hostname               [sp6en0]
  Password File                               [/etc/passwd]
  Home Directory Server Hostname              [sp6n01]
  Home Directory Path                         [/home/filesrv]

  File Collection Management                  true                          +
  File Collection daemon uid                  [102]
  File Collection daemon port                 [8431]                        #

  SP Accounting Enabled                       false                         +
  SP Accounting Active Node Threshold         [80]                          #
  SP Exclusive Use Accounting Enabled         false                         +
  Accounting Master                           [0]

  Control Workstation LPP Source Name         [aix51b-133a]

  SP Administrative Locale                    en_US                         +
  SDR may contain ASCII data only             true                          +
  Root remote command access restricted       true                          +
  Remote command method                       secrshell                     +
  Remote command executable                   [/usr/bin/ssh]
  Remote copy executable                      [/usr/bin/scp]

  SP Machine Type Model Number                []
[MORE...1]

F1=Help            F2=Refresh          F3=Cancel           F4=List
F5=Reset           F6=Command          F7=Edit             F8=Image
F9=Shell           F10=Exit            Enter=Do
```

Figure 5-2 smit site_env_dialog panel for setting up secrshell

The SP software does not install, configure, or set up your chosen secure remote command program; this is your responsibility.

The root user must be able to issue the **ssh** and **scp** commands without being prompted for a password or passphrase. There are several ways to achieve this. Two possibilities are:

- A secure way (which we recommend) using generation of public and private keys

- A non-secure way involving changing the /usr/local/etc/ssh_config file (a default location in distribution that we use) as follows:
  - StrictHostKeyChecking=no
  - BatchMode=yes

In addition, PSSP 3.4 provides an example of how to create a script.cust (Example A-1 on page 177) to install and set up secrshell on the nodes during node installation.

5.3.3 Changes to PSSP code and to SDR

PSSP 3.4 supports three environment variables to let you pick whether you want the PSSP system management software to use the AIX **rsh** and **rcp** remote commands or a secure remote command process for parallel remote commands like **dsh**, **pcp**, and others. The following are the environment variables and how to use them:

- RCMD_PGM

  Enable use of the executables named by the DSH_REMOTE_CMD and REMOTE_COPY_CMD environment variables. The default is **rsh**. Set the value to secrshell to enable a secure remote command process.

- DSH_REMOTE_CMD

  Specify the path and name of the remote command executable. The default with rsh is **/bin/rsh**. The default with secrshell is **/bin/ssh**.

- REMOTE_COPY_CMD

  Specify the path and name of the remote copy command executable. The default with rsh is **/bin/rcp**. The default with secrshell is **/bin/scp**.

The commands described depend on entries in the SDR. Three new attributes have been added to the SDR SP_Restricted class (refer to “dsh” on page 172): rcmd_pgm, dsh_remote_cmd and remote_copy_cmd. They have their corresponding environment variables with the same name, as described at the beginning of this section. The SMIT site environment menu and the **spsitenv** command have been changed accordingly to manipulate these new attributes in the SDR (see Figure 5-2 on page 108 and Example 5-7 on page 108).

5.4 Using none as an AIX remote command authorization method

When very high security is required in your SP system, you might consider using this new feature as your remote command authorization method.

5.4.1 Overview

Using none as an authorization method for AIX remote commands eliminates automatic generation of the entries for root in the authorization files (Kerberos V4: /.klogin, Kerberos V5: /.k5login, standard AIX: /.rhosts).

The change requires you to enable a secure remote command method other than AIX **rsh** in the PSSP code. It does not remove the need to remote shell to the nodes.

You may want to set none for AIX remote command authorization if you want to add a level of security at which even the root user is no longer automatically authorized to issue the **rsh** or **rcp** commands from the CWS to the nodes.

5.4.2 Changes to PSSP code

The SMIT Select AIX Authorization methods for remote commands menu has been enhanced to support a none option.

When the none option is selected, the **updauthfiles** script (see Section 5.2.2, “Authorization routine and files” on page 105) runs and removes all the PSSP entries in the /.rhosts, /.klogin and /.k5login files.

5.5 Migration

For detailed description of migration processes refer to Chapter 4, “Installation, migration, and coexistence” on page 77.

5.5.1 Authorization = none

If the AIX remote command authorization method is set to none, the rcmd authorization files (Kerberos V4 .klogin, Kerberos V5 .k5login, standard AIX .rhosts) on the nodes and the CWS have no entries. If a new node is added, entries are not added. Because it is a requirement that all nodes be at PSSP 3.4 for none to be set, there is no migration of a node from PSSP 3.2 to PSSP 3.4 in this environment.

5.5.2 When secrshell is used and RRA enabled

When secrshell is enabled, the migration is the same as for RRA. Enabling secrshell does not remove any additional authorization methods and supports the migration of nodes from PSSP 3.2 to PSSP 3.4 with no changes. Because it is a requirement that all nodes to be at least at PSSP 3.2 for secrshell to be set, there is no migration of a node from a PSSP version lower than PSSP 3.2 to PSSP 3.4 in this environment.

Important: If you have enabled RRA and secrshell, you must ensure that the required customization in script.cust, firstboot.cust and firstboot.cmds is available and up-to-date.

Attention: Public keys must be generated and a known_hosts file set up, or strict_hostname checking must be disabled so that the PSSP code can issue secure remote commands as root without being prompted.

Refer to the *PSSP for AIX: Installation and Migration Guide,* GA22-7347.

5.6 Firewalled RS/6000 SP system

If you need to logically and physically separate your SP system into two “sides,” then in order to control communication between these sides you should consider implementing the firewalled SP system.

5.6.1 Overview

A firewalled SP system is the same SP system divided into two sets of nodes:

- A trusted or private side (nodes behind the firewall)
- An untrusted or public side (nodes in front of the firewall)

The trusted side consists of the CWS and *trusted nodes*, and the untrusted side consists of *untrusted nodes*. An SP node serves as a firewall between the two sets (see Figure 5-3 on page 112). Obviously the firewall node belongs to both sides, trusted and untrusted, but the CWS must be on the trusted side.

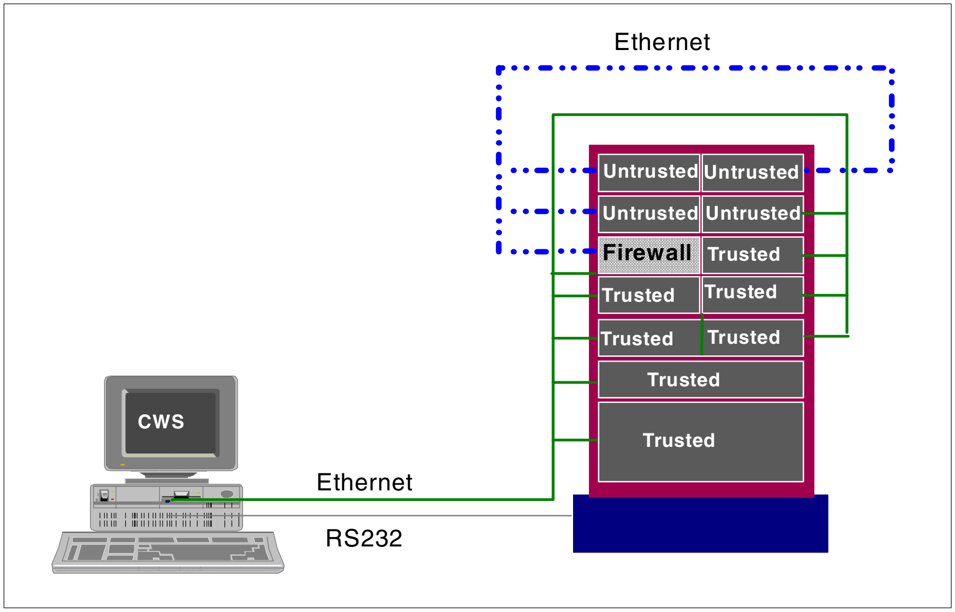


Figure 5-3 A firewalled RS/6000 SP system

The firewall node separates the two sides physically and logically. Because it regulates communication between the two sides, it must be connected to both of them. The firewall node has two separate LANs:

- en0, the original SP Ethernet LAN, for the trusted side
- en1 for the untrusted side

The firewall node permits or blocks communication between trusted and untrusted sides. This includes traffic to and from the firewall node and to and from the CWS. Communication is regulated by a set of conditions, or *firewall rules*, which are combinations of several values (host names and IP addresses, protocols, services, and port numbers).

Note: It is up to you to select the brand or type of firewall product you use, but it must run on the level of AIX required within a firewalled SP system.

Please read *Implementing a Firewalled RS/6000 SP System, Version 3, Release 4,* GA22-7874*l*.

Before you start to implement a firewalled SP system you should:

- Have a basic understanding of firewall technology
- Be familiar with firewall words and phrases such as “rule”, “rule set”, “service/service object”, “host object”, “network object”, and “security policy”
- Understand what is meant by adding and enabling rules to your firewall application
- Understand what is meant by disabling rules in your firewall application

5.6.2 Restrictions

A firewalled SP system must adhere to the guidelines in *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment,* GA22-7280, and *IBM RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment,* SG22-7281.

Also, there are restrictions unique to a firewalled SP system. The following items are *not* supported in a firewalled SP system:

- A PSSP Trusted Services security method of none
- An SP Switch
- An SP Switch2 on the untrusted nodes and the firewall node. Trusted nodes can be on the SP Switch2
- Switch commands through the firewall, if the SP Switch2 is used
- Multiple SP system partitions
- HACMP on the firewall node
- HACMP on the untrusted nodes when the takeover node is a trusted node or the firewall node
- HACMP on the trusted nodes when the takeover node is an untrusted node or the firewall node
- HACWS
- High availability firewall node

- Multiple firewall nodes
- The CWS connected directly to the untrusted side
- The CWS as the firewall node
- The CWS as a BIS server for the untrusted nodes
- A trusted node as a BIS for the untrusted nodes or the firewall node
- An IBM @server pSeries 690 and corresponding Hardware Management Console cannot be on the untrusted side
- GPFS
- An IBM Virtual Shared Disk

The IBM @server pSeries 690 server is supported within a firewalled SP system under the following conditions:

- All the IBM @server p690's LPARs and its associated HMC controller must be network-connected to the trusted side of a firewalled SP system.
- An IBM @server p690 cannot be the firewall node.

The following applications are not supported on the untrusted nodes and the firewall node, though they are allowed on the trusted nodes:

- Dynamic Probe Class Library (DPCL)
- LoadLeveler
- Parallel Operating Environment (POE)

6

Parallel programming

This chapter discusses the following:

- A brief introduction to IBM Parallel Environment for AIX (PE) 3.2 enhancements in Section 6.1, “The IBM Parallel Environment for AIX (PE) 3.2” on page 116
- Message Passing Interface (MPI) enhancements in Section 6.2, “The Message Passing Interface (MPI)” on page 118
- 65-way User Space (US) communication in Section 6.3, “65-way multiple user space processes per adapter (MUSPPA)” on page 125
- Low-level Application Programming Interface (LAPI) enhancements in Section 6.4, “The Low-Level Application Programming Interface” on page 126
- Engineering and Scientific Subroutine Library (ESSL) and Parallel ESSL enhancements in Section 6.5, “Parallel ESSL 2.3 and ESSL 3.3” on page 132

6.1 The IBM Parallel Environment for AIX (PE) 3.2

The IBM Parallel Environment for AIX (PE) 3.2 is a set of programs designed to help you develop parallel Fortran, C, or C++ programs and execute them. You can run them on the IBM RS/6000 SP system or a cluster of IBM @server pSeries or RS/6000 systems.

6.1.1 New in PE Version 3, Release 2

PE 3.2 contains a number of functional enhancements:

- 64-bit support for the PE, including Parallel Operating Environment (POE), MPI, LAPI, Xprofiler, Dynamic Probe Class Library (DPCL), and the PE Benchmarker toolkit
- MPI enhancements

  These enhancements are discussed in Section 6.2, “The Message Passing Interface (MPI)” on page 118.
- Parallel checkpoint and restart capabilities

  In previous releases a checkpoint sequence, for example, could be initiated only by tasks in the parallel MPI program. In Version 3.2, a user, a single task in a parallel MPI or LAPI job, a system administrator, or LoadLeveler can be allowed to initiate a checkpoint sequence.
- New PE Benchmarker toolkit

  This toolkit, called *PE benchmarker,* enables you to analyze program performance. PE benchmarker is built on the DPCL and consists of the following:

  – Performance Collection Tool (PCT)

    This tool enables you to collect either MPI and user event data or hardware and operating system profiles for one or more application processes.

  – Unified Trace Environment (UTE) utilities

    When you collect MPI and user event traces using the PCT, the collected information is saved as a standard AIX event trace file. The UTE utilities enable you to convert one or more of these AIX trace files into UTE interval files that are easier to visualize than the traditional AIX files.

  – Profile Visualization Tool (PVT)

    This tool enables you to view hardware and operating-system profiles collected by the PCT.

For more detailed information related to PE benchmarker, refer to *Parallel Environment for AIX: Operation and Use, Volume 2,* SA22-7426.

Software considerations

PE 3.2 includes these changes:

- PE 3.2 requires AIX 5L Version 5.1 or later and PSSP 3.4.
- PE 2.2 is no longer supported.
- The DPCL is no longer a part of PE, though it is still shipped with PE. Instead, DPCL is now an open-source offering that supports PE.
- The pedb debugger has been removed.
- The visualization tool (VT) parallel tracing facility has been removed.

Coexistence

All SP nodes involved in a parallel job must be running the same level of PE.

When LoadLeveler and PE coexist on a node, they must be conform to one of these combinations:

- LoadLeveler 3.1 with PE 3.2
- LoadLeveler 2.2 with PE 3.1
- LoadLeveler 2.2 with PE 2.4, with PSSP 3.1 only
- LoadLeveler 2.1 with PE 2.4
- LoadLeveler 1.3 with PE 2.3

The PE libraries used to run a job on a given node must be compatible with the PSSP libraries on that node. The following combinations are compatible:

- PE 3.2 with PSSP 3.4
- PE 3.1 with PSSP 3.4, 32-bit applications on AIX Version 4.3.3 only
- PE 3.1 with PSSP 3.2
- PE 2.4 with PSSP 3.1
- PE 2.3 with PSSP 2.4
- PE 2.3 with PSSP 2.3

Migration

PE does not support node-by-node migration. You must migrate all the nodes in a system partition to a new level of PE at the same time. To migrate to PE 3.2, we suggest that you first migrate to PSSP 3.4, then migrate to AIX 5L Version 5.1 and PE 3.2 at the same time.

6.2 The Message Passing Interface (MPI)

The MPI is the result of a standardization effort for an expressive and flexible message passing API for parallel computing that can be implemented on almost any platform. An MPI program can be compiled to run with a compliant MPI implementation.

This section discusses MPI enhancements such as 64-bit support, MPI-IO, collective communication, and others. For more detailed information about PE MPI, refer to *IBM Parallel Environment for AIX: Installation,* GA22-7418, *IBM Parallel Environment for AIX: MPI Programming Guide,* SA22-7422, and *IBM Parallel Environment for AIX: MPI Subroutine Reference,* SA22-7423.

For information about the Message Passing Interface Forum and documents concerning MPI standards refer to the following Web site:

http://www.mpi-forum.org

6.2.1 Porting to 64-bit MPI

In a 32-bit addressing scheme, a data type can map 2 GB. In 64-bit mode, users can create MPI data types for more than 2 GB.

In C programs, some of the parameters in the MPI bindings use MPI_Aint instead of int for 64-bit support. The MPI Standard defines MPI_Aint as an *address-size integer*. In the MPI library, MPI_Aint is an *unsigned long integer*. Other parameters impacted by the address range still use int, such as count parameters in MPI data type constructors, count parameters in MPI_SEND and MPI_RECV, and outputs of MPI_TYPE_SIZE and MPI_PACK_SIZE.

The FORTRAN parameters matching C MPI_Aint parameters are INTEGER in MPI-1. This works well for a 32-bit environment, but the definition is incorrect for a 64-bit environment. MPI-2 provides replacements for routines with this problem and uses INTEGER (KIND=MPI_ADDRESS_KIND). For example, the following functions have new MPI-2 replacement versions:

- MPI_TYPE_HVECTOR → MPI_TYPE_CREATE_HVECTOR
- MPI_TYPE_STRUCT → MPI_TYPE_CREATE_STRUCT
- MPI_ADDRESS → MPI_GET_ADDRESS
- MPI_TYPE_EXTENT → MPI_TYPE_GET_EXTENT

Important: For FORTRAN INTEGER arguments cases in which the argument value fits in 32 bits, the old functions will work. But if the value does not fit, truncation is silent. You are urged to modify the code when you port to 64 bits to avoid surprises later.

6.2.2 MPI-IO performance enhancements

MPI-IO is the I/O component of MPI-2 and provides a set of interfaces aimed at performing portable and efficient parallel I/O. MPI-IO in PE MPI is targeted to the IBM General Parallel File System (GPFS) for production use. File access through MPI-IO normally requires that a single GPFS file system image be available across all tasks of an MPI job. PE MPI with MPI-IO can be used for program development on any other file system that supports a Portable Operating System Interface For Computer Environments (POSIX) interface (AFS, DFS™, JFS, or NFS), as long as all tasks run on a single node or workstation. This restriction is not expected to yield a useful model for production use of MPI-IO. PE MPI can be used without all nodes on a single GPFS file system image by using the MP_IONODEFILE environment variable.

To improve MPI-IO performance, some optimizations have been made in PE 3.2:

- Round-robin scheme for shipping commands and data

  In PE 3.1, all driver threads issuing I/O requests communicate with I/O agents 0, 1, 2, 3, ... in the same order for shipping commands and data. PE 3.2 provides a round-robin scheme: each driver thread starts with a distinct I/O agent and goes around until it has completed a full cycle. This provides a more even load and task serving as I/O agents.

- Use of GPFS multiple-access range hint

  The GPFS multiple-access range hint allows you to explicitly control selective prefetching of GPFS blocks and free buffer space by releasing GPFS access ranges after they are accessed. This hint should only be provided to GPFS when the access pattern is neither sequential nor strided-sequential, because GPFS handles these patterns well without any hints.

- Use of double buffering at the responder

  In PE 3.1, only one data buffer at each responder is used for a given collective data access operation. PE 3.2 allows overlapping of data transfer between driver threads and responders, and GPFS I/O calls at the responders. It accomplishes this by providing two data buffers, one for data transfer and the other for GPFS I/O calls.

The MPI-IO library functions remain functionally equivalent in PE 3.2. The only change in functionality is the introduction of one new hint and two new environment variables.

The IBM_sparse_access hint

MPI-2 provides a hints facility. Hints provide the implementation with information about things, such as the structure of the application and the type of expected file accesses for running a set of tasks.

The new file hint, IBM_sparse_access, enables you to specify whether the file access requests from participating tasks are sparse or dense. This lets you specify the future file access pattern of the application for the associated file. You can use MPI_FILE_OPEN, MPI_FILE_SET_INFO and MPI_FILE_SET_VIEW subroutines to set this hint. For more information about file hints, refer to *IBM Parallel Environment for AIX: MPI Subroutine Reference,* SA22-7423.

POE environment variables

Two new environment variables have been added:

- MP_IO_BUFFER_SIZE

  This variable enables you to set the default size of buffers used by I/O agents, which is the global value of the IBM_io_buffer_size hint. Example 6-1 sets the size of the MPI-IO data buffer to 16 MB.

  > **Note:** The MPI library rounds the number up to match an integer number of file blocks, if necessary.

Example 6-1 MP_IO_BUFFER_SIZE

```
export MP_IO_BUFFER_SIZE=16M
       or
poe -io_buffer_size 16M
```

- MP_IO_ERRLOG

  This variable enables you to log errors occurring at the file system level throughout the MPI-IO application run and to identify the lower-level problems. If you turn on error logging as shown in Example 6-2, a line of error information is logged into file /tmp/mpi_io_errdump.<job_name>.<userid>.<taskid>.

Example 6-2 MP_IO_ERRLOG

```
export MP_IO_ERRLOG=yes
       or
poe -io_errlog yes
```

6.2.3 Collective communication enhancements

PE 3.2 provides the MPI_IN_PLACE and intercommunicator semantics that MPI-2 has added to a number of MPI 1.1 collective communication subroutines:

- MPI_BCAST
- MPI_GATHER, MPI_GATHERV
- MPI_SCATTER, MPI_SCATTERV
- MPI_ALLGATHER, MPI_ALLGATHERV
- MPI_ALLTOALL, MPI_ALLTOALLV
- MPI_REDUCE, MPI_ALLREDUCE
- MPI_REDUCE_SCATTER
- MPI_BARRIER

The MPI_IN_PLACE parameter

MPI_IN_PLACE is supported as the value of the send buffer at the root in the following collective functions:

- MPI_GATHER
- MPI_GATHERV
- MPI_REDUCE
- MPI_ALLREDUCE

MPI_IN_PLACE is supported as the value of the receive buffer at the root in the following collective functions:

- MPI_SCATTER
- MPI_SCATTERV

MPI_IN_PLACE is supported as the value of the send buffer on all tasks in the following collective functions:

- MPI_ALLGATHER
- MPI_ALLGATHERV
- MPI_REDUCE_SCATTER
- MPI_SCAN

New collective communication functions

Two new subroutines, MPI_ALLTOALLW and MPI_EXSCAN, are also provided:

- MPI_ALLTOALLW

This subroutine is an extension of MPI_ALLTOALLV allowing separate specification of count, displacement, and data type. In addition, to allow maximum flexibility, the displacement of blocks within the send and receive buffers is specified in bytes.

- MPI_EXSCAN

  This subroutine performs a prefix reduction on data distributed across the group.

6.2.4 Miscellaneous

Some of the functions included in the “miscellaneous” chapter of MPI-2 were provided in prior releases of PE. PE 3.2 provides the rest of these functions.

Additional command for starting MPI jobs

While you can continue to use the **poe** command to start MPI jobs, release 3.2 of PE provides support for the **mpiexec** command described in the MPI-2 standard. This command is not meant to replace the **poe** command; instead it is provided as a portable way to start MPI programs and should prove helpful for applications that target multiple implementations of MPI.

Thread utility functions

The MPI thread library supports either a single-threaded or full multiple-threaded environment. This is controlled by the MPI_SINGLE_THREAD environment variable. The default is full multithreaded support. These threads functions provide a portable way to specify desired thread support and to check what is available. The level of threads support in PE/MPI has been equivalent to MPI_THREAD_MULTIPLE for several releases and add portability but not new capability. The following are new functions for portable MPI programming with threads:

- MPI_INIT_THREAD

  This subroutine initializes MPI in the same way that a call to MPI_INIT does. In PE MPI, MPI_INIT_THREAD is equivalent to MPI_INIT.

- MPI_QUERY_THREAD

  This subroutine returns the current level of thread support into the argument provided.

- MPI_IS_THREAD_MAIN

  A thread can call this subroutine to find out whether it is the main thread.

Canonical MPI_PACK and MPI_UNPACK functions

These functions provide a way for an MPI program to convert between the native format of data in memory and a portable format called external32, matching the external32 MPI_FILE format. The following do not replace MPI_PACK and MPI_UNPACK which are used to pack/unpack data used within an MPI job:

- MPI_PACK_EXTERNAL

  This subroutine packs the message in the specified send buffer into the specified buffer space.

- MPI_UNPACK_EXTERNAL

  This subroutine unpacks the message into the specified receive buffer from the specified packed buffer.

- MPI_PACK_EXTERNAL_SIZE

  This subroutine returns the number of bytes required to hold the data.

MPI_COMM_SELF deleted at MPI_FINALIZE

This change allows programs to run user provided cleanup code when MPI_FINALIZE is called. The side effect of deleting MPI_COMM_self is that any attributes a user has attached will be deleted. Deleting an attribute includes running any attribute delete function linked to the attribute.

Generalized requests

For a generalized request, the MPI non-blocking operations associated with the request are designed by the application programmer and performed by the application. A generalized request can be used with MPI_WAIT, MPI_TEST, or MPI_CANCEL. To use generalized requests, the application will normally spawn a thread on which the application defined operation will asynchronously proceed and eventually finish.

The following are new functions for generalized requests:

- MPI_GREQUEST_START

  This subroutine initializes a generalized request and returns a handle to it in the request argument. The application must notify MPI when the operation has finished. It does this by making a call to MPI_GREQUEST_COMPLETE.

- MPI_GREQUEST_COMPLETE

  This subroutine informs MPI that the operations represented by the generalized request are complete.

- MPI_STATUS_SET_ELEMENTS

  This subroutine defines element information for a generalized request and places it in the status argument.

- MPI_STATUS_SET_CANCELED

  This subroutine defines cancellation information for a generalized request and places it in the status argument.

Others

The followings are miscellaneous new MPI functions:

- MPI_TYPE_CREATE_INDEXED_BLOCK
- MPI_ADD_ERROR_CLASS
- MPI_ADD_ERROR_CODE
- MPI_ADD_ERROR_STRING
- MPI_COMM_SET_NAME
- MPI_COMM_GET_NAME
- MPI_TYPE_SET_NAME
- MPI_TYPE_GET_NAME
- MPI_WIN_SET_NAME
- MPI_WIN_GET_NAME
- MPI_COMM_CALL_ERRHANDLER
- MPI_WIN_CALL_ERRHANDLER
- MPI_FILE_CALL_ERRHANDLER
- MPI_REQUEST_GET_STATUS
- MPI_FINALIZED

For more information about functions, refer to *IBM Parallel Environment for AIX: MPI Subroutine Reference,* SA22-7423.

6.2.5 Language bindings

PE 3.2 provides the C++ bindings described by MPI-2. C++ programmers can now use PE MPI more naturally, as they no longer need to use the C bindings. FORTRAN 90 programmers can now use the mpi module in place of mpif.h.

C++ bindings

MPI has supported C++ with C bindings since the beginning. PE 3.2 provides true C++ bindings based on the concept of classes that match the original object oriented structure of MPI. That is, MPI is based on communicators, datatypes, files, groups, and so on. These pre-existing MPI “classes” map to C + + “classes”. The C++ bindings correspond to the existing MPI routines, because for the most part MPI routines are members of one of these classes.

FORTRAN 90 improvements

The addition of FORTRAN 90 bindings to MPI allows users to more fully access the functionality of MPI with FORTRAN, while still allowing use of FORTRAN 77.

This extended FORTRAN support includes an MPI module that can used in FORTRAN 90 programs. The module defines all named MPI constants and declares MPI functions that return a value. Extended support also provides some additional routines to support FORTRAN intrinsic numeric types. The following are new functions for creating FORTRAN 90 types:

- MPI_TYPE_CREATE_F90_REAL
- MPI_TYPE_CREATE_F90_COMPLEX
- MPI_TYPE_CREATE_F90_INTEGER
- MPI_TYPE_MATCH_SIZE
- MPI_SIZEOF

MPI_SIZEOF is only supported by the FORTRAN 90 bindings and is the only routine that requires USE MPI in place of INCLUDE MPI F.4.

6.3 65-way multiple user space processes per adapter (MUSPPA)

User Space (US) communication requested by a task running on LoadLeveler, Parallel Environment (PE), and PSSP communication requires at least one adapter window per processor. 32 US tasks per node are supported by using the SP Switch2 PCI attachment adapter attached to legacy RS/6000 servers and IBM @server pSeries 690, 680, and 660. Each US task can use both LAPI and MPI; thus, 64 US windows are required. A sixty-fifth window is reserved for the GPFS token manager via the US LAPI.

SP Switch2 PCI attachment adapter

The SP Switch2 PCI attachment adapter has the following characteristics:

- Only this adapter can support 65 US windows. If a pair of SP Switch2 PCI attachment adapters are used in a dual switch-plane environment, striped communication can be specified through PE.
- There are also three non-US windows for IP, service packet, and VSD/KHAL on this adapter.
- PSSP 3.4 is required for enabling 65 US windows.

An SP Switch2 MX2 adapter on POWER3 SMP Thin and Wide Nodes, 375 MHz POWER3 SMP Thin and Wide Nodes, and 332 MHz SMP Thin and Wide Nodes supports up to nine US windows.

6.4 The Low-Level Application Programming Interface

The Low-Level Application Programming Interface (LAPI) is a non-standard application programming interface (API) designed to provide optimal communication performance on an SP switch. It is based on an active message programming mechanism that provides a one-sided communications model, that is, one process initiates an operation and the completion of that operation does not require any other process to take a complementary action. The LAPI is designed for use by libraries and power programmers for whom performance is more important than code portability.

The following are some considerations for using PSSP 3.4 LAPI:

- LAPI can run in a mixed environment of nodes with PSSP 3.2 or later.
- LAPI supports both SP Switch2 and SP Switch.
- You can migrate PSSP from any supported release of PSSP to PSSP 3.4 one node at a time. Applications that use LAPI cannot migrate one node at a time. LAPI users do not need to recompile the programs. LAPI maintains binary compatibility between releases.
- You must use PE 3.1 or later.
- 64-bit support for LAPI is available on AIX 5L Version 5.1 or later.
  - All nodes in a job must run the same level of PE, PSSP, and the application.
  - Applications under 64-bit addressing must use AIX 5L Version 5.1 and PSSP 3.4.

This section describes LAPI enhancements such as shared memory support and a new LAPI vector function. For more detailed information, refer to *PSSP for AIX: Administration Guide,* SA22-7348.

6.4.1 Shared memory support

In PSSP 3.2, the LAPI supports only User Space communication using the SP Switch. In PSSP 3.4, it is possible for communication protocol clients to use an efficient shared memory protocol in intranode communication.

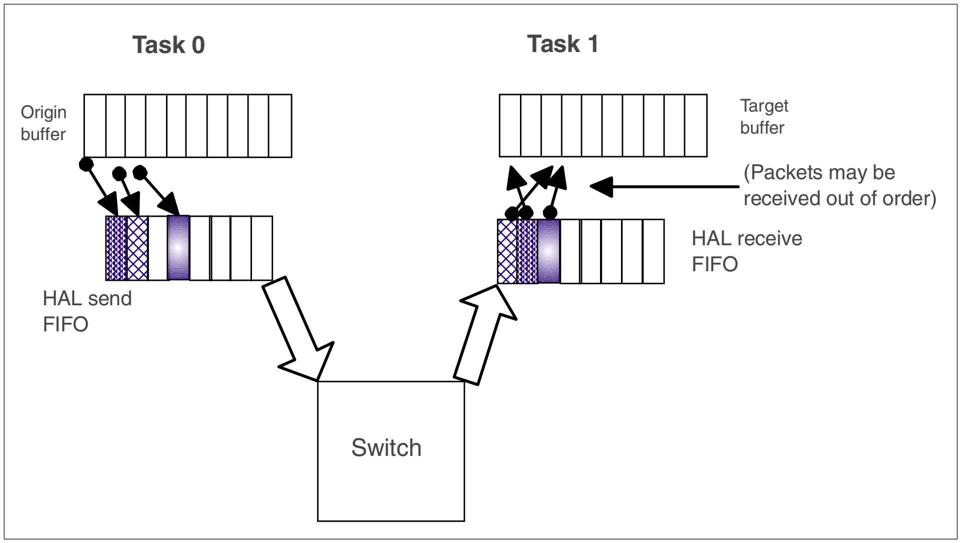


Figure 6-1 Intranode communication with switch

Figure 6-1 shows that LAPI interacts with SP Switch through the hardware abstraction layer (HAL). Figure 6-2 on page 128 shows that LAPI uses shared memory segments on the same node. Tasks using the shared memory segments are controlled by semaphore. This causes two copies of the LAPI communication and should be used for small messages.

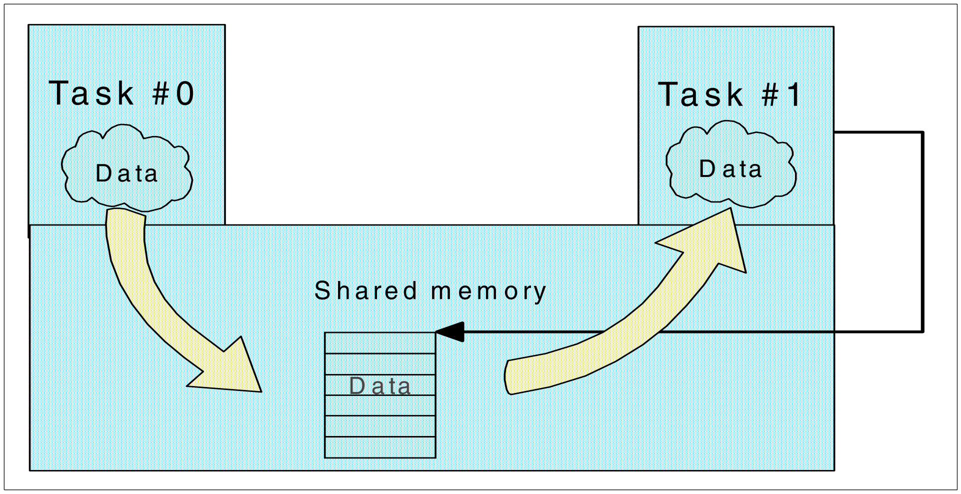


Figure 6-2 Intranode communication with shared memory segment

The shared memory kernel extension

For large messages, a kernel extension is provided to support the LAPI shared memory model loaded as part of AIX initialization. This kernel extension allows one task to export a portion of its address space to another task of the LAPI parallel application. The exported portion of the address space is attached to the address space of another task and then the data transfer is done by a simple copy, so only one copy is needed.

Figure 6-3 on page 129 reflects an intranode LAPI communication using a kernel extension:

- Task 1 registers a shared memory segment containing data and gets the segment identifier for that segment.
- Task 1 copies the segment identifier to the shared memory segment.
- Task 1 invokes Task 2.
- Task 2 retrieves the segment identifier.
- Task 2 attaches the segment containing the data to its address space.
- Task 2 copies the data.

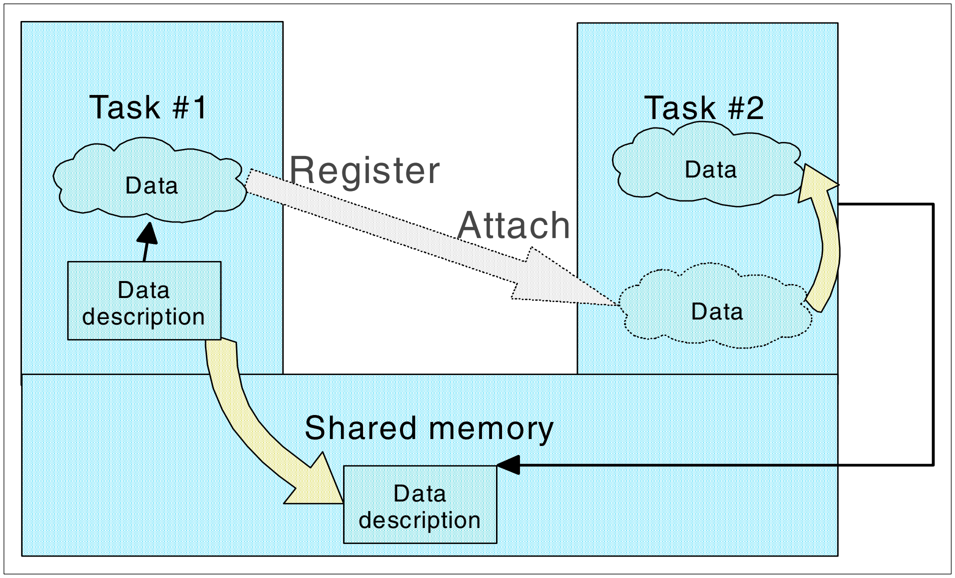


Figure 6-3 Shared memory kernel extension

LAPI shared memory requirements

The requirements for using LAPI shared memory are as follows:

- LAPI shared memory must be used on AIX 4.3 or later and used only in a POE environment.
- The maximum number of tasks supported on a node is 32.
- The kernel extension must be loaded.
- The environment variable LAPI_USE_SHM must be set to either YES or ONLY. If it is set to ONLY, LAPI uses the shared memory mechanism only, and an error message is returned if not all tasks are on the same node or shared memory setup is not successful. If it is set to YES, LAPI uses both the shared memory mechanism and the switch path. The switch path is used when shared memory initialization fails or tasks on different nodes are involved.

6.4.2 New LAPI vector function

PSSP 3.2 introduced two LAPI vector functions to improve performance and ease of use in transporting non-contiguous data between tasks using LAPI. These functions were *general I/O vector transfer* and *general strided transfer.* PSSP 3.4 supports another LAPI vector function, *generic I/O vector transfer,* which is a generalized form of general I/O vector transfer.

All the LAPI vector functions use the following I/O vector structure:

```
typedef struct {
   lapi_vectype_t vec_type; /* operation code */
   uint num_vecs; /* number of vectors */
   void **info; /* vector of information */
   uint *len; /* vector of lengths */
} lapi_vec_t;
```

General I/O vector transfer

With vec_type set to LAPI_GEN_IOVECTOR we have a general I/O vector transfer.

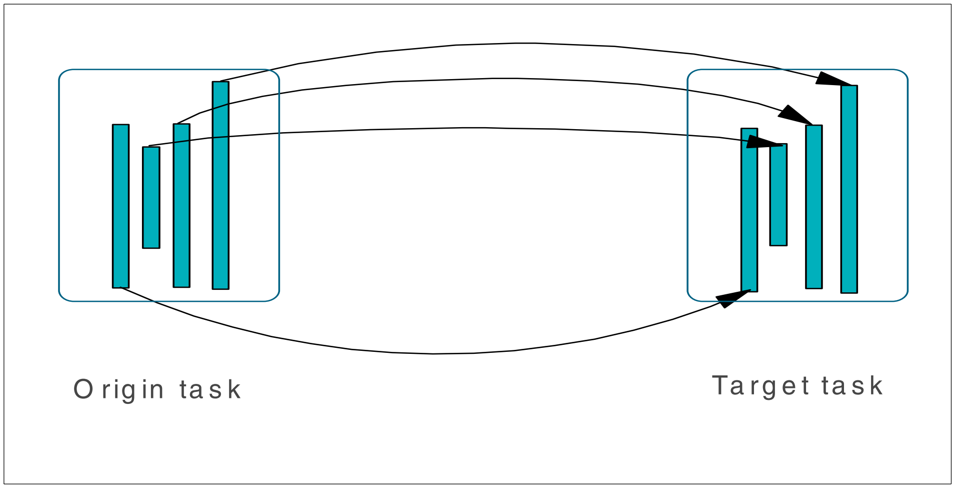


Figure 6-4 General I/O vector transfer

Figure 6-4 shows an example of general I/O vector transfer. Four vectors with different lengths are transferred from the origin task to the target task by one vector transfer operation.

Strided vector transfer

With vec_type set to LAPI_GEN_STRIDED_XFER we have a strided vector transfer. Figure 6-5 on page 131 shows an example of strided vector transfer. Three strided vectors are transferred from the origin task to the target task by one vector transfer operation.

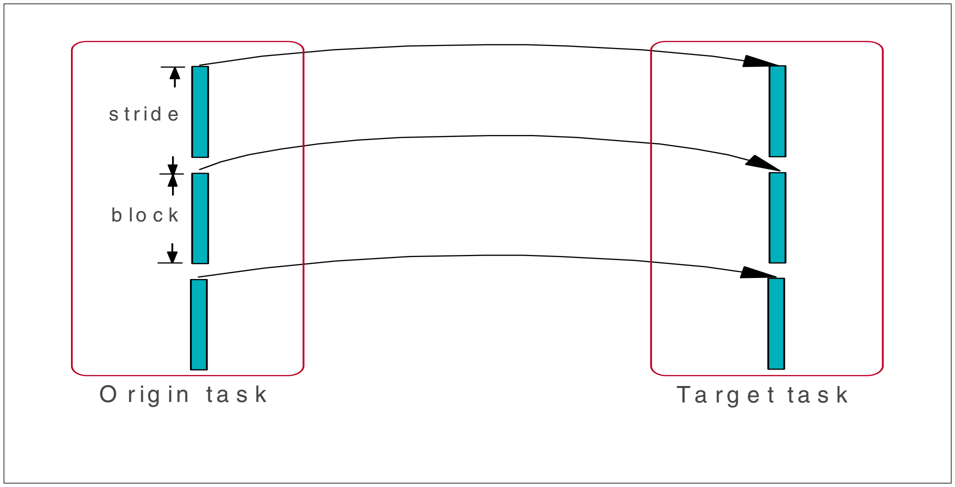


Figure 6-5 Strided vector transfer

Generic I/O vector transfer

With vec_type set to LAPI_GEN_GENERIC we have a generic I/O vector transfer. Figure 6-6 shows an example of generic I/O vector transfer. The active message function (LAPI_Amsendv) is used for this. A given number of bytes in non-contiguous buffers are transferred from the origin task to another number of bytes in non-contiguous buffers specified by the target vector structure returned by the header handler of the LAPI_Amsendv function.

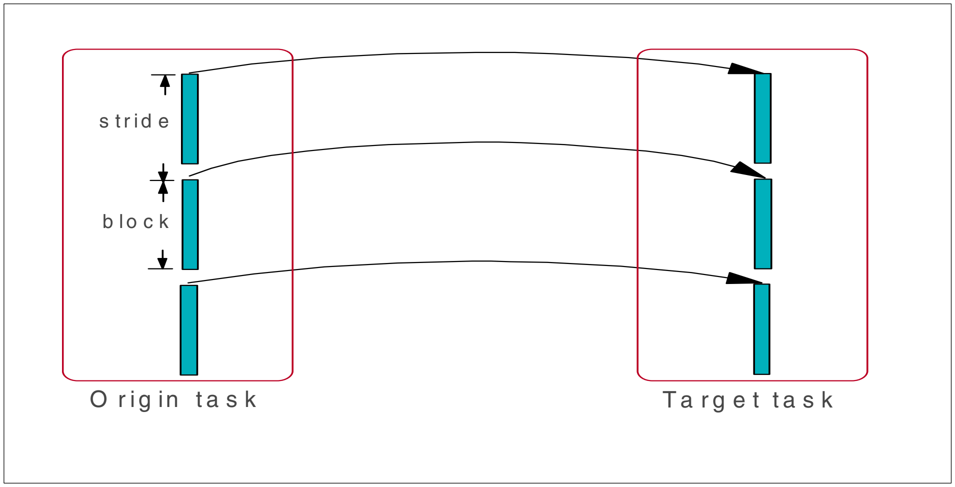


Figure 6-6 Generic I/O vector transfer

6.4.3 Other performance improvements

PSSP 3.4 LAPI provides greater bandwidth and increased payload size, because the LAPI message packet header has been reduced. There are multiple completion handler functions, such as LAPI_Waitcntr, LAPI_Getcntr, and LAPI_Setcntr, for better exploitation of the symmetrical multiprocessor system (SMP) system and for improving PUT latency.

Striped communication is provided for Kernel Lightweight Directory Access Protocol (KLAPI) in a dual switch-plane environment.

6.5 Parallel ESSL 2.3 and ESSL 3.3

The Engineering and Scientific Subroutine Library (ESSL) family of products is a state-of-the-art collection of mathematical subroutines that provides a wide range of high-performance mathematical functions for many different scientific and engineering applications.

ESSL and Parallel ESSL libraries can be used to develop and enable many different types of scientific and engineering applications. New applications can be designed and developed to take full advantage of all the capabilities of the ESSL family. Existing applications can be enabled by replacing comparable subroutines and in line code with calls to ESSL subroutines.

6.5.1 ESSL

ESSL contains over 450 high-performance mathematical subroutines tuned to IBM @server pSeries and RS/6000 servers, workstations, and SP systems. The ESSL subroutines can be called from application programs written in Fortran, C, or C++.

The mathematical subroutines fall into nine computational areas:

- Linear algebra subprograms
- Matrix operations
- Linear algebraic equations
- Eigensystems analysis
- Fourier transforms, convolutions and correlations, and related computations
- Sorting and searching
- Interpolation
- Numerical quadrature

- Random-number generation

New in ESSL Version 3, Release 3

The following are new features in ESSL 3.3:

- The ESSL libraries are tuned for the performance optimization with enhanced RISC architecture (POWER4).
- ESSL supports AIX 5L Version 5.1 32-bit and 64-bit kernels.
- The ESSL header file supports the C++ Standard Numerics Library facilities for complex arithmetic.
- The dense linear algebraic equation subroutines include new linear algebra package (LAPACK) subroutines:
  - General matrix inverse
  - Positive definite real symmetric or complex Hermitian matrix factorization
  - Positive definite real symmetric or complex Hermitian matrix multiple right-hand sides
  - Positive definite real symmetric matrix inverse
  - Triangular matrix inverse
- The linear least squares subroutines include a new LAPACK subroutine:
  - linear least squares solution for a general matrix

ESSL libraries

ESSL libraries support both 32-bit and 64-bit application environments and are tuned for Power4, Power3, Power3-II, PowerPC, and POWER processors.

ESSL 3.3 provides two run-time libraries:

- ESSL SMP (OpenMP) Library

  This library provides thread-safe versions of ESSL subroutines for use on SMP processors. In addition, a subset of these subroutines are multithreaded versions; that is, they support the shared memory parallel processing programming model. You do not have to change your existing application programs that call ESSL to take advantage of the increased SMP processor performance; you can simply relink your existing application programs. For a list of the multithreaded subroutines in the ESSL SMP Library, see *Parallel ESSL for AIX V2R3 Guide and Reference,* SA22-7273.

- ESSL Serial (thread-safe) Library

  This library provides thread-safe versions of the ESSL subroutines for use on all IBM @server pSeries and RS/6000 processors. You may use this library to develop your own multithreaded applications.

All libraries are designed to provide high performance levels for numerically intensive computing jobs. All versions provide mathematically equivalent results.

Requirements

The software products you need when using ESSL 3.3 are as follows:

- AIX 5L Version 5.1 32-bit or 64-bit kernel
- XL FORTRAN for AIX, Version 7.1.1 (5765-E02), VisualAge C++ Professional for AIX, Version 5.0.2, or C for AIX, Version 5.0.2, for compiling
- XL FORTRAN Run-Time Environment for AIX, Version 7.1.1 (5765-E03), and C libraries, for linking, loading, or running. AIX includes the C and math libraries in the Application Developer's Kit.

Migration

The following considerations apply to migration to ESSL 3.3:

- No changes to FORTRAN or C application programs are required if you are migrating from ESSL 3.2 to ESSL 3.3.
- Changes may be required in C++ application programs. For detailed information, refer to *Parallel ESSL for AIX V2R3 Guide and Reference,* SA22-7273.
- All 64-bit applications and libraries must be recompiled on AIX 5L Version 5.1.

6.5.2 Parallel ESSL

Parallel ESSL contains more than 75 single program-multiple data (SPMD) mathematical subroutines tuned to exploit the full power of the SP hardware with scalability across the range of system configuration. In addition to SP systems, Parallel ESSL runs on clusters of IBM @server pSeries and RS/6000 servers and/or workstations.The Parallel ESSL subroutines can be called from application programs written in FORTRAN, C and C++.

Parallel processing subroutines are provided in these areas:

- Basic linear algebra communication sub-programs (BLACS)
- Level 2 and Level 3 Parallel basic linear algebra sub-programs (BLAS)
- Linear algebraic equations (dense, banded, and sparse subroutines)
- Eigensystems analysis

- Fourier transforms
- Random-number generation

New in Parallel ESSL Version 2, Release 3

The following are new features in Parallel ESSL 2.3:

- The Parallel ESSL libraries are tuned for the performance optimization with enhanced RISC architecture (POWER4). IBM intends to announce a PTF in the first half of 2002.
- Parallel ESSL supports the AIX 5L Version 5.1 32-bit kernel.
- The Parallel ESSL SMP Libraries support 32-bit and 64-bit environment applications.
- The Parallel ESSL header file supports the C++ Standard Numerics Library facilities for complex arithmetic.
- The dense linear algebraic equations subroutines include new scalable linear algebra package (ScaLAPACK) subroutines:
  - Complex general matrix QR factorization
  - Least squares solution to linear systems of equations for complex general matrix
- The eigensystem analysis subroutines include new ScaLAPACK subroutines:
  - Selected eigenvalues and optionally the eigenvectors of a complex Hermitian positive definite generalized eigenproblem
  - Reduction of a complex Hermitian positive definite generalized eigenproblem to standard form

Parallel ESSL libraries

Parallel ESSL 2.3 provides two run-time libraries:

- Parallel ESSL SMP Libraries

  These libraries are provided for use with the MPI threaded library. You may run single or multithreaded applications on all types of nodes. However, you cannot simultaneously call Parallel ESSL from multiple threads. Use these Parallel ESSL libraries if you are using both MPI and LAPI. The SMP library is for use on SMP processors.

  The Parallel ESSL SMP Libraries support both 32-bit and 64-bit environment applications.

- Parallel ESSL Serial Libraries

  These libraries are provided for use with the MPI signal-handling library on all types of nodes.

The Parallel ESSL Serial Libraries support only 32-bit environment applications.

Requirements

The software products you need when using Parallel ESSL 2.3 are as follows:

- AIX 5L Version 5.1 32-bit kernel
- Parallel Environment (PE) 3.2
- ESSL 3.3
- XL FORTRAN for AIX, Version 7.1.1 (5765-E02), VisualAge C++ Professional for AIX, Version 5.0.2, or C for AIX, Version 5.0.2, for compiling
- XL FORTRAN Run-Time Environment for AIX, Version 7.1.1 (5765-E03), C libraries, and PE 3.2, and ESSL 3.3, for linking, loading, or running
  - AIX includes the C and math libraries in the Application Developer’s Kit.

Migration

No changes to your application programs are required if you are migrating from Parallel ESSL 2.2 to Parallel ESSL 2.3.

7

LoadLeveler 3.1

This chapter discusses the following LoadLeveler 3.1 features:

- Gang scheduler
- Checkpoint and restart
- LoadLeveler and Workload Manager integration
- 64-bit support
- File system monitoring
- Striped communication
- Parallel Operating Environment (POE) considerations

7.1 LoadLeveler overview

LoadLeveler is a job management system that allows users to run more jobs in less time by matching the jobs' processing needs with the available resources. LoadLeveler schedules jobs and provides functions for building, submitting, and processing jobs quickly and efficiently in a dynamic environment.

7.1.1 Migration

When migrating from LoadLeveler Version 2.2 to Version 3.1, you must migrate the *central manager* first. All other nodes can be migrated on a node-by-node basis without bringing the entire cluster down. After migration, all LoadLeveler commands are available to users and administrator. In a mixed cluster, we recommend that you execute the **llacctmrg** and the **llsummary** commands from a version 3.1 node.

7.1.2 Compatibility

The following compatibility issues must be taken into account if you are planning to use LoadLeveler:

- LoadLeveler 3.1 can coexist with LoadLeveler 2.2. In a mixed cluster, the central manager and all LoadL_schedd daemons must be on a machine operating with LoadLeveler 3.1.
- AIX Workload Manager (WLM), checkpointing, gang scheduling, and 64-bit values are only supported if all nodes are at LoadLeveler 3.1.
  - Resource usage with WLM is not enforced on 2.2 nodes.
  - Incorrect or incomplete checkpoint files are generated.
  - LoadLeveler does start if gang scheduler is requested.
  - Values greater than 32-bit representations are truncated when sent from 3.1 nodes to 2.2 nodes.

7.1.3 Prerequisites

If you are planning to use LoadLeveler 3.1, the following are needed:

- PSSP 3.4
- AIX 5L Version 5.1.0.10
- POE 3.2

7.2 LoadLeveler features

This section describes the new enhancements in gang scheduler, checkpoint and restart, LoadLeveler WLM, 64-bit support, file system monitoring, and striped communication.

7.3 Gang scheduler

Gang scheduler allocates the resources of a node to a set (gang) of jobs, each of which takes turns using the resources exclusively or semi-exclusively. Gang scheduling provides good overall system utilization and responsiveness to interactive workloads.

Gang scheduling is designed for the parallel environment and features:

- Support for coordinated context switching for both time-sharing and space-sharing applications
- Resource sharing that matches applications with available resources
- Support for preempting jobs

Gang scheduler runs all tasks for a job step on all nodes for a brief time called a *time slice*. Gang scheduler runs and stops all tasks for the job step at the same time.

Supported hardware

Gang scheduling supports the following adapters:

- SP Switch MX adapter
- SP Switch MX2 adapter
- SP Switch2 adapter
- SP Switch2 PCI adapter

7.3.1 Gang scheduling matrix

The negotiator daemon on the managing machine in a LoadLeveler cluster, known as the *central manager*, monitors the status of each job and machine in the cluster with a *gang scheduling matrix* containing the following:

- A synchronization mechanism
- The time-slice size
- The lists of steps running on all processors on a set of nodes

The gang scheduling matrix is sent to the startd daemon for each node it covers. The startd daemon uses the synchronization mechanism to determine which job steps should be running. It can start, suspend, resume, or ignore a job step. For example, it starts any steps that should be running but have not been started.

Hierarchical communication

The negotiator daemon sends the gang scheduling matrix to the master daemon at the root of the hierarchy that is running on the first destination or node in the execution machines. The master daemon removes local matrix columns and sends them to the local startd daemon. The master daemon splits the remaining matrix into *n* smaller matrices and sends them to the next *n* destinations (*n=fanout)*. The startd daemon receives the local column and makes it a current matrix. Figure 7-1 shows an example of hierarchical communication.

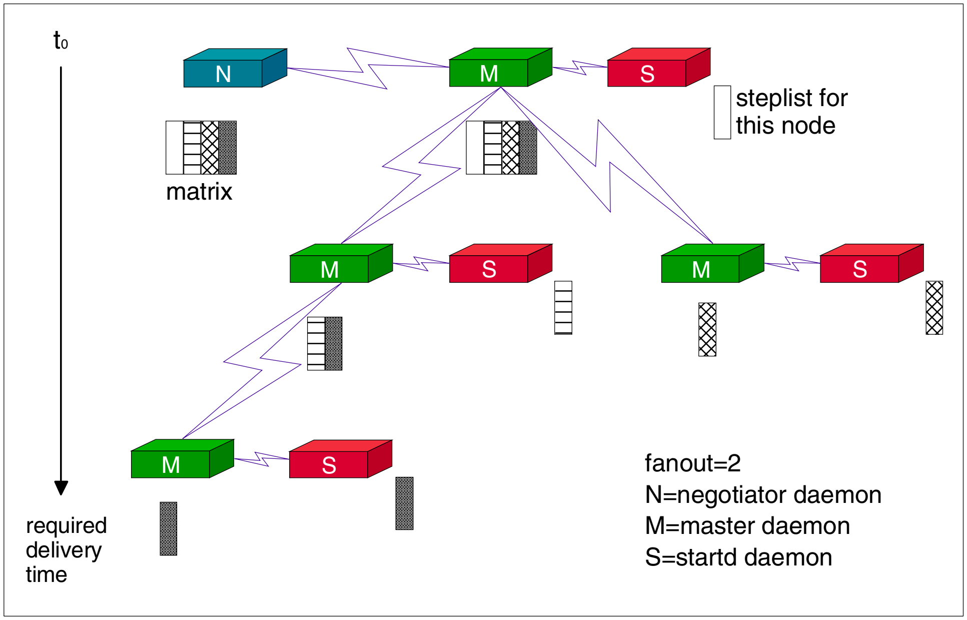


Figure 7-1 Hierarchical communication path

7.3.2 Preemption

Gang scheduling enables a new feature called *preemption*. Using preemption, resources are switched from executing jobs to new jobs that have been authorized to take control of cluster resources. When a job has been preempted, the job is suspended and remains in the swap space of the nodes that were running it; so, gang scheduling can release the preempted job's resources and reassign them to the incoming job.

There are two types of preemption:

- System-initiated preemption, which has the following characteristics:
  - Automatically enforced by LoadLeveler.
  - Controlled by the PREEMPT_CLASS keyword in the global configuration file.
  - Preempted job steps are resumed, not redispatched, when resources become available according to START_CLASS rules.
  - System-preempted job steps are only resumed by the system, and not by using the **`llpreempt`** command or ll_preempt subroutine.

  > **Notes:**
  >
  > - The **`llmodify -x 99`** command makes a job step non-preemptable. All other jobs running on the same nodes are preempted until this job finishes running.
  > - The **`llmodify -x 1`**, **`2`**, or **`3`** command undoes the **`llmodify -x 99`** command, thus allowing you to preempt the job step.

- User-initiated preemption, which has the following characteristics:
  - Manually enforced by the user.
  - Controlled by the **`llpreempt`** command or the ll_preempt subroutine.
  - Both the command and the subroutine can resume a user-preempted job step.
  - Cannot automatically resume user-preempted steps.

Figure 7-2 on page 142 shows an example of the preemption rules. ENOUGH indicates that the incoming class listed on the left (class B) will only preempt the number of lower priority tasks (class C) needed to free up the required resources. *ALL* prevents the preempted class jobs (class B, C) from being resumed until the preempting class job (class A) completes. Two START_CLASS lines indicate that class B and class C jobs can only start on nodes that do not have any class A jobs running.

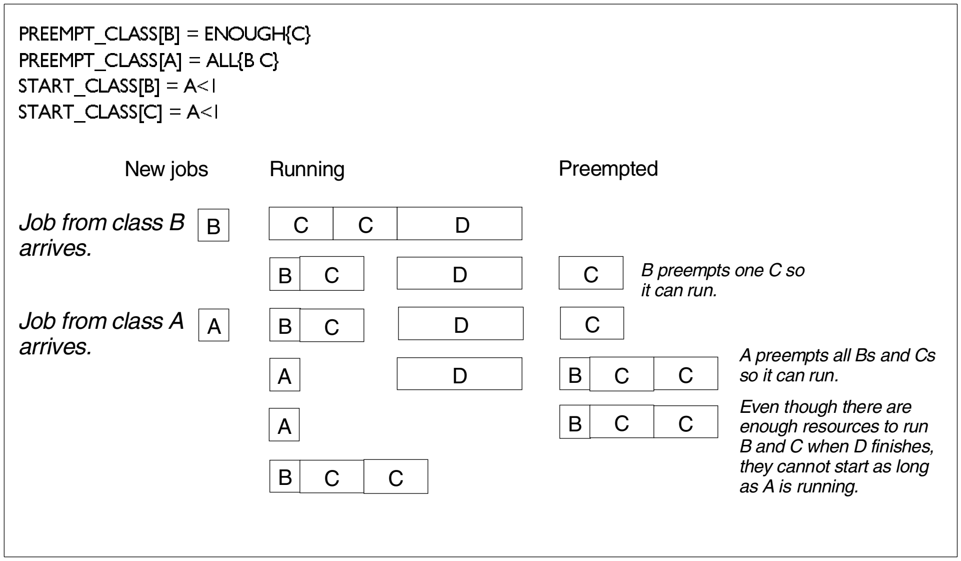

PREEMPT_CLASS[B] = ENOUGH{C}
PREEMPT_CLASS[A] = ALL{B C}
START_CLASS[B] = A<1
START_CLASS[C] = A<1
New jobs
Running
Preempted
Job from class B arrives.
B
C
C
D
B
C
D
C
B preempts one C so it can run.
Job from class A arrives.
A
B
C
D
C
A
D
B
C
C
A preempts all Bs and Cs so it can run.
A
B
C
C
Even though there are enough resources to run B and C when D finishes, they cannot start as long as A is running.
B
C
C

Figure 7-2 PREEMPT_CLASS and START_CLASS

7.3.3 New keywords for gang scheduling

There are several new keywords in the configuration file, administration file, and job command file.

New keywords in the configuration file

The following keywords can be found in the configuration file:

- HIERARCHICAL_FANOUT

  Specifies how many child nodes a hierarchical message is distributed to at each level of the hierarchy. The default value for this keyword is 2. The minimum value is 1, while the maximum is the number of nodes in the cluster. If the value is greater than the number of nodes, the first node sends the message to all other nodes.

- SCHEDULER_TYPE

  Specifies the type of scheduling algorithm. The default value for this keyword is LL_DEFAULT. In order to use the gang scheduler, users must assign GANG to this value:

  - SCHEDULER_TYPE = GANG

- GANG_MATRIX_TIME_SLICE

  Specifies a time slice. The minimum (and default) value for this keyword is 60 seconds, while the maximum is 3600 seconds.

- GANG_MATRIX_NODE_SUBSET_SIZE

  Specifies the ideal number of nodes for processing user jobs. The default value for this keyword is 512, and the range is any positive integer value.

- GANG_MATRIX_REORG_CYCLE

  Specifies the number of negotiation loops during which the scheduler waits to reorganize the gang scheduling matrix into subsets whose sizes are as close as possible to the ideal GANG_MATRIX_NODE_SUBSET_SIZE. The default value for this keyword is 16, and the range is all positive integers.

- GANG_MATRIX_BROADCAST_CYCLE

  Specifies the time that the scheduler will wait before sending the complete matrix information to the startd daemon. The default value for this keyword is 300 seconds and the range is all positive integers.

- PREEMPT_CLASS

  Specifies the preemption rules used by an incoming job class. When a job step is preempted by the rules specified in this keyword, the preemption is automatic or system-initiated. System-initiated preemtion cannot be undone using the **llpreempt** command or through the ll_preempt subroutine. If this keyword is not defined, the gang scheduler default does not allow preemption.

- START_CLASS

  Specifies the rules for starting a job of the designated class. By default, the system does not place a limit on the number of jobs that can have a START_CLASS definition. The START_CLASS rules are applied whenever the scheduler decides which job step should begin.

New keywords in the administration file

The following keywords can be found in the administration file:

- execution_factor

  Specifies the number of time slices for job steps in this class. This keyword appears in the class stanza. The default value for this keyword is 1. The range is 1, 2, and 3.

- max_smp_tasks

  Specifies the number of tasks that run concurrently on a machine. This keyword appears in the machine stanza. The default value of this keyword is the number of processors in the node. The value must be less than or equal to the number of processors in the node.

- max_total_tasks

  Specifies the number of tasks that the gang scheduler allows a user, group, or class to run at any given time. This keyword appears in the user, group, and class stanzas. The default value for this keyword is –1, which allows an unlimited number of tasks. The range is –1, 0, and all positive integers.

Modification in the job command file

The node_usage keyword has only been modified for use with gang scheduling. For gang scheduling, node_usage has a new value of slice_not_shared, which specifies that nodes are not shared during the time slice when the job is running.

7.3.4 New commands for gang scheduling

The following commands can be used with gang scheduling:

`llclass` Returns information about classes.

`llmatrix` Returns the gang matrix information in the LoadLeveler cluster when gang scheduling is used.

`llmodify` Changes the attributes or characteristics of a submitted job.

`llpreempt` Places a specified job step in the (user-initiated) preempted state. The job step stays in that state until the action is undone with the -u flag. When **`llpreempt`** is undone the job resumes its normal state, controlled by the LoadLeveler scheduling and preemption rules.

`llq` Returns information about job steps in the LoadLeveler queues. The status field in the **`llq -l`** output indicates whether the preemption was system initiated or user initiated.

`llsubmit` Submits a job to LoadLeveler to be dispatched based upon job requirements in the job command file.

For more detailed information, refer to *Loadleveler V2R2 for AIX: Using and Administering,* SA22-7311.

7.4 Checkpoint and restart

Checkpointing is a method of periodically saving the state of a running job so that job execution can be restarted from the saved state. The job can only be restarted on machines that satisfy the environment in which the checkpoint was made.

LoadLeveler supports two mechanisms for initiating a checkpoint:

- Application- (or user-) initiated

  The user's application program determines when the checkpoint is taken through the ll_init_ckpt application programming interface (API) call for serial jobs, and through the mpc_init_ckpt API call for parallel jobs.

- Externally (or system-) initiated

  The checkpoint is initiated from the **`llckpt`** command or from another program such as a monitoring program or external scheduler that invokes the checkpoint API, or at administrator-defined intervals.

7.4.1 How to checkpoint and restart jobs

When you enable checkpointing for a job, if that job does not complete it can be restarted from the saved state.

How to checkpoint a job

We list three ways to checkpoint a job. You can also select these options on the Build a Job window of the GUI.

- checkpoint = yes in the job command file

  This value allows application or external initiation through an API or command.

- checkpoint = interval

  LoadLeveler checkpoints a job automatically. System administrators must set two keywords, MIN_CKPT_INTERVAL and MAX_CKPT_INTERVAL, in the configuration file to specify when the checkpoints are taken. A minimum value of 900 seconds and a maximum value of 7200 seconds are the defaults. The first checkpoint is taken after a period of time has passed that is equal to the MIN_CKPT_INTERVAL. The interval value is doubled when the checkpoint is taken. LoadLeveler repeats this step. Finally, if the new value is greater than MAX_CKPT_INTERVAL, the interval value is set to MAX_CKPT_INTERVAL.

- checkpoint = no

  This value disallows checkpointing and is the default.

How to restart a job

The restart_from_ckpt = yes keyword allows you to restart the job from an existing checkpoint file. You can also select this option on the Build a Job window of the GUI.

7.4.2 Checkpoint files

At checkpoint time a checkpoint file and potentially a control file and/or error file is created. The directory containing the checkpoint file must preexist and have sufficient space and permissions for the checkpoint files to be written.

Batch checkpoint file

The name and location of the checkpoint file are controlled through keywords in the job command file or the LoadLeveler configuration:

- If ckpt_file is specified as a fully qualified name in the job command file, both the base file and directory names are set to chpt_file.
- If ckpt_dir is specified in the job command file, the base directory name is set to ckpt_dir.
- If chpt_dir is specified in the class stanza of the LoadLeveler admin file, the base directory name is set to ckpt_dir.
- If ckpt_dir is not specified, the base directory name is the initial working directory.
- If ckpt_file is specified in the job command file but is not a fully qualified name, the base name is constructed from base_dir/ckpt_file.
- If ckpt_file is not specified, the base name is constructed from base_dir/[jobname.]job_step_id.

Attention: Two or more job steps running at the same time cannot write to the same checkpoint file, since the file would be corrupted.

The checkpoint file for a serial job or for the master task of a parallel job has a name with the format *basename.tag,* where tag is a number that is the same for all files generated from the same checkpoint. The file names for the other tasks of a parallel job follow the format *basename.taskid.tag.*

Interactive checkpoint file

The ckpt_dir and ckpt_file are not used for interactive parallel jobs.

The checkpoint file name is obtained from the following in the order listed:

1. If MP_CKPTFILE is specified as a fully qualified name in the job command file, both the base file and directory names are set to MP_CKPTFILE.
2. If MP_CKPTDIR is specified in the job command file, the base directory name is set to MP_CKPTDIR.
3. If MP_CKPTDIR is not specified, the base directory name is the initial working directory.
4. If MP_CKPTFILE is specified in the job command file but is not a fully qualified name, the base name is constructed from base_dir/MP_CKPTFILE.
5. If MP_CKPTFILE is not specified, the base name is constructed from base_dir/poe.ckpt.pid.

7.4.3 Checkpoint and restart considerations

There are several considerations or limitations for using checkpoint and restart; for example, a job can be restarted on machines other than the one where it was checkpointed; the characteristics of the machines can be changed since the job was checkpointed; programs can be run under a debugger or may have a semaphore, message queue, shared memory, sockets, or pipes.

The following is a sample consideration at the time of a restart. The node on which a process is restarted must have:

- The same operating-system level (including PTFs) as the node where the checkpoint occurred
- The same switch type (SP Switch or SP Switch2)
- Any processor IDs to which threads were bound at checkpoint time

For more detailed information, refer to *Loadleveler V2R2 for AIX: Using and Administering,* SA22-7311.

7.5 LoadLeveler and Workload Manager integration

LoadLeveler enforces the usage of consumable resources, ConsumableCpus and ConsumableMemory, using the AIX Workload Manager (WLM). If resources in the LoadLeveler cluster are in contention, WLM adjusts the priority of jobs exceeding their resource usage.

New keywords in the configuration file

There are two new keywords for LoadLeveler and WLM integration:

- ENFORCE_RESOURCE_USAGE

  Specifies that LoadLeveler enforces the resource usage of ConsumableCpus and ConsumableMemory defined on the SCHEDULE_BY_RESOURCES keyword in the configuration file. These two variable must be specified on the SCHEDULE_BY_RESOURCES keyword to be enforced by WLM.

  The deactivate variable can be specified on this keyword. If the value is specified, LoadLeveler deactivates WLM on all the nodes in the LoadLeveler cluster.

- ENFORCE_RESOURCE_SUBMISSION = true | false

  Specifies whether LoadLeveler checks all jobs at submission time.

Restrictions

There are a few restrictions to consider:

- Resource usage is only enforced on LoadLeveler 3.1 nodes.
- Only 27 job steps can be running at a time on a machine with resource usage enforced.
- If gang scheduling is used, 27 simultaneous job steps can be running on a machine.
- The twenty-eighth job remains undone until one of the first 27 job steps finishes.

7.6 64-bit support

LoadLeveler version 3.1 has been enhanced to provide 64-bit support for interactive and batch jobs. Users and administrators can specify and request enforcement of the large 64-bit system resource limits that are available under AIX 5L Version 5.1 or higher.

7.6.1 Keywords supporting 64-bit data

64-bit integer values can be assigned to selected keywords and used in expressions in the job command, configuration, and administration files.

The job command file

The following job command file keywords support 64-bit data:

- image_size

- data_limit
- file_limit
- core_limit
- stack_limit
- rss_limit
- resources
- requirements
- preferences

The following keywords are still 32-bit values:

- cpu_limit
- job_cpu_limit
- wall_clock_limit
- ckpt_time_limit

If a value beyond the range of an int32_t data type is assigned to one of these limits, it is truncated to either INT32_MAX (2147483647) or INT32_MIN (–2147483648).

The administration file

There are new keywords in the administration file:

- resources (machine stanza)
- data_limit (class stanza)
- file_limit (class stanza)
- core_limit (class stanza)
- stack_limit (class stanza)
- rss_limit (class stanza)
- default_resources (class stanza)

The configuration file

There are new keywords in the configuration file:

- floating_resources
- Memory
- VirtualMemory
- FreeRealMemory

- Disk
- ConsumableMemory
- ConsumableVirtualMemory
- ConsumableCpus
- PagesScanned
- PagesFreed

Units for 64-bit keywords

Table 7-1 specifies the units for 32-bit and 64-bit keywords.

Table 7-1 Keyword units

Unit	Abbreviation	Value
Units for 32-bit and 64-bit keywords		
byte	b	1
word	w	4 bytes
kilobyte	kb	2^{10} bytes
kiloword	kw	2^{12} bytes
megabyte	mb	2^{20} bytes
megaword	mw	2^{22} bytes
gigabyte	gb	2^{30} bytes
gigaword	gw	2^{32} bytes
Additional units for 64-bit enhanced keywords		
terabyte	tb	2^{40} bytes
teraword	tw	2^{42} bytes
petabyte	pb	2^{50} bytes
petaword	pw	2^{52} bytes
exabyte	eb	2^{60} bytes
exaword	ew	2^{62} bytes

7.6.2 LoadLeveler APIs

Both 32-bit and 64-bit application development libraries are available for LoadLeveler APIs and Message Passing Interface (MPI) checkpointing. In Version 2.2, the LoadLeveler checkpointing library and the LoadLeveler API library are two distinct entities. But now in Version 3.1 the checkpoint library is part of the LoadLeveler API library and the MPI library. Developers attempting to exploit the 64-bit capabilities of the LoadLeveler API library should take into consideration the following points:

- If the Distributed Computing Environment (DCE) is not enabled,
  - Both 32-bit and 64-bit libraries are available.
  - Only the 32-bit library is available when you use the Interactive Session Support (ISS) 1.3.
- If DCE is enabled (DCE_ENABLEMENT = TRUE), only the 32-bit library is available.
- The 32-bit and 64-bit LoadLeveler checkpointing libraries are available even when DCE is enabled.

Most API subroutines support both 32-bit and 64-bit data:

- Checkpointing API
- Submit API
- Data access API
- Parallel job API
- Workload management API
- Query API

In the accounting API there are some changes in the history file and the methods accessing this file:

- Job statistics are preserved in the history file as rusage64, rlimit64,and data items of type int64_t.
- In previous versions of LoadLeveler, users could access the history file via the GetHistory interface. In LoadLeveler 3.1, users of the GetHistory interface can access 64-bit data items as either int64_t data or int32_t data.
- The **llsummary** command has been modified to display 64-bit information.

Any API we have not listed supports only 32-bit data.

7.7 File system monitoring

The file system keywords monitor all file systems used by LoadLeveler and increase system fault tolerance. This is intended to avoid problems caused by insufficient space for writing logs or saving executables.

New keywords for file system monitoring

There are new keywords for file system monitoring:

- FS_INTERVAL = second

  Specifies the file system checking interval (in seconds). LoadLeveler does not check a file system if the value of FS_INTERVAL is:

  – Not specified

  – Set to zero

  – Set to a negative integer

- FS_NOTIFY = low_threshold,high_threshold

  Specifies the point when the administrator is notified of a problem and when the administrator is notified that the problem is resolved. If file system free space drops below the low_threshold, LoadLeveler sends a message to the administrator indicating that logging problems may occur. If the free space rises above the high_threshold, LoadLeveler sends a message to the administrator indicating that the problems have been resolved.

- FS_SUSPEND = low_threshold,high_threshold

  Specifies the point when LoadLeveler is suspended on a machine and the point at which it is resumed. If file system free space drops below low_threshold, LoadLeveler suspends the schedd and the startd daemons if they are running on a node. If the free space rises above the high_threshold, LoadLeveler signals the schedd and the startd daemons to resume.

- FS_TERMINATE = low_threshold,high_threshold

  Specifies the point when LoadLeveler is terminated on a machine. LoadLeveler is not automatically restarted. If file system space drops below low_threshold, LoadLeveler sends a SIGTERM signal to the Master daemon, and all LoadLeveler daemons are terminated.

7.8 Striped communication

LoadLeveler provides striped communication between the nodes for parallel jobs, which can use all the communication paths to the node. Striped communication is meant for parallel jobs, since submitting a serial job using striping brings no benefits. There are two methods of striping: user space (US) striping and IP striping.

US striping

All switch connections on a node are represented by a single virtual device named *csss*. The csss device is generated automatically and is not specified in the administration file. You can use this virtual adapter when a job requires striped communication. US striping protocol is passed to the communication subsystem through POE. Note that each instance of the US mode requested by a task running on the SP Switch requires an adapter window.

Example 7-1 shows a network statement using US striping.

Example 7-1 US striping

```
network.protocol = network_type [, [usage] [, mode [, comm_level]]]

network.MPI = csss,shared,US
```

IP striping

IP striping involves three sets of adapters, css0, css1, and ml0. It is known as multi-link IP addressing or *aggregate IP addressing* (see Section 3.4.2, "Aggregate IP addressing" on page 55).

Only nodes that have aggregate IP addresses are used in this method, that is, adapters within an aggregate IP network can communicate with each other using the ml0 aggregate IP addresses. If aggregate IP addresses are not available, an IP address of css0 or css1 is used. Example 7-2 shows a network statement using IP striping.

Example 7-2 IP striping

```
network.protocol = network_type [, [usage] [, mode [, comm_level]]]

network.MPI = csss,shared,IP
```

7.9 POE considerations

LoadLeveler imposes some restrictions on POE:

- All machines running a POE job must have the same level of POE.
- POE 3.2 requires LoadLeveler 3.1, and POE 3.1 requires LoadLeveler 2.2.
- To provide a usable environment for parallel jobs, we recommend that you partition a mixed cluster into 2.2 and 3.1 pools, using the pool_list keyword in the administration file. You can then use the requirements keyword in the job command file to direct a job to the desired pool for a batch job. Machines also can be assigned to the desired pool based on the level of POE installed, using the POE -rmpool option for an interactive parallel job.
- To avoid copying incompatable levels of the POE executable between nodes in a mixed-cluster environment, the job command file of a POE batch job submitted to LoadLeveler should not contain the executable keyword. If you specify this keyword, it causes the executable from the submitting node to be copied to the executing node.

8

GPFS 1.5

This chapter presents the IBM General Parallel File System (GPFS) Version1.5.

GPFS is a high-performance, scalable file system designed for cluster environments. Version 1.5 is provided for PSSP 3.4 environments.

This chapter introduces GPFS concepts and configurations, then details the enhancements in version 1.5, and finally discusses some considerations related to migration, coexistence, and compatibility between GPFS versions.

8.1 GPFS overview

GPFS is a high-performance, scalable file system designed for cluster environments. GPFS provides *high performance* by spreading I/O activity across multiple disks, *high scalability* with the SP Switch and the SP Switch2, and *high availability* through logging and replication mechanisms. It can be configured to provide fault tolerance to both disk and server malfunctions.

GPFS allows both parallel and serial applications running on different nodes to share data spanning multiple disk drives attached to multiple nodes. GPFS provides the following features:

- Conventional Portable Operating System Interface For Computer Environments (POSIX)

  Most programs run without any changes or recompilation. Access to data is transparent for applications; the standard UNIX file system semantics are used. GPFS also supports UNIX file system utilities, so users can use the UNIX commands for ordinary file operations. JFS file systems and GPFS file systems can be NFS-exported.

- High-performance data access

  Wide striping of read and write requests across disks and nodes leads to performance improvements and multiple GB/s bandwidth for transferring data to or from one file.

- Scalability

  GPFS is flexible as cluster growth requirements appear. Unlike other distributed file systems, GPFS file performance scales as additional file server nodes and disks are added to the GPFS cluster.

- High storage capabilities

  GPFS allows multi-TB files and file systems. Large numbers of data blocks can be stored in a single flat file, which is accessed concurrently by applications residing on multiple nodes.

- Parallel data and metadata access

  GPFS performs all file system functions, including metadata functions, on all members of the cluster, both within a file and across different files in a file system.

- Reliability and fault tolerance

  GPFS provides recoverability for different failure scenarios, including failure of the following components: node, disk, disk adapter, and communication adapter.

- Online system management

  GPFS provides dynamic configuration and monitoring. Disks can be added or deleted while the file system is mounted. You can also add new nodes without stopping and restarting the GPFS daemon.

- Simplified administration

  A single GPFS multinode command can perform a file system function across the entire GPFS cluster and can be performed from any node in the cluster.

GPFS is able to support large data environments requiring a high degree of parallelism in accessing shared data. GPFS applications include:

- Parallel or serial applications that require fast, scalable access to large amounts of data like:
  - Weather forecasting
  - Seismic data processing
- Applications with very large data that exceeds the capacity limits of other file systems:
  - Digital libraries
  - Access to large computer-graphics aided three-dimensional interactive application (CATIA) filesets
  - Business intelligence
- Applications requiring data transfer rates that exceed the capabilities of other file systems:
  - Large aggregate of scratch space for commercial or scientific applications
  - Internet content distribution
- Applications that require file systems with high-availability features

8.2 GPFS clusters

GPFS is a clustered file system defined over multiple nodes. The overall set of nodes over which GPFS is defined is known as a *GPFS cluster*.

There are several types of GPFS clusters, depending on the operating environment:

SP environment The SP environment is based on the IBM Virtual Shared Disk (VSD), a component of PSSP. GPFS uses the high-speed interconnect, the SP Switch or the SP Switch2 network, inside the SP system and the shared disk component provided by VSD. For more information, refer to Section 8.2.1, “GPFS in an SP environment.”

HACMP environment GPFS in an HACMP environment is based on High Availability Cluster Multi-Processing for AIX/Enhanced Scalability (HACMP/ES). Refer to Section 8.2.2, “GPFS in an HACMP environment” on page 160 for more details on this configuration.

Linux cluster GPFS currently runs on Netfinity Intel 32-bit processors with the Red Hat Linux distribution. It supports both direct attached disks and Network Shared Disk (NSD) configurations. For more information related to GPFS for a Linux environment go to the following Web site:

`http://www-1.ibm.com/servers/eserver/clusters/software/gpfs.html`

Within a GPFS cluster, the nodes are divided into one or more GPFS *nodesets*. Nodes inside a nodeset have the same GPFS version. However, it is possible for different GPFS versions to coexist in the same cluster. Refer to Section 8.5, “Migration, coexistence and compatibility” on page 168.

Each nodeset has his own set of file systems shared among the nodes inside it. The shared file systems are not accessible for the nodes outside the nodeset.

8.2.1 GPFS in an SP environment

GPFS in an SP environment utilizes the Virtual Shared Disk (VSD) and the Recoverable Virtual Shared Disk (RVSD) components. The VSD component of PSSP is used as the underlying data transport method. It actually emulates a storage area network (SAN) using switch connectivity.

Since VSD uses the switch network, the boundaries of the GPFS cluster in an SP environment depend on the switch type. For systems with the SP Switch, the GPFS cluster is equal to the corresponding SP partition. In a system with an SP Switch2, the GPFS cluster is equal to all the nodes in the system.

GPFS is designed to operate with the following software products:

- AIX provides basic operating system services and receives GPFS file system calls, which are redirected to the GPFS daemon.
- PSSP provides:
  - The VSD and RVSD components. RVSD is a prerequisite for GPFS. RVSD fences a node from accessing certain disks, which is needed for

successful recovery of that node. It also provides transparent failover in case of VSD server failure.

- The Group Services (GS) subsystem of the Reliable Scalable Cluster Technology (RSCT). GS is used for failure notification and sequencing of recovery on multiple nodes.
- The SP security services, which are needed by the sysctl routines and remote commands for performing VSD and GPFS functions.
- The SDR on the Control Workstation (CWS), which is used to retrieve configuration data when a nodeset is created or a node is added to the nodeset. Such data might include:
  - Node number
  - Adapter type
  - IP address

> **Note:** The complete configuration data maintained by GPFS is stored in GPFS-specific files and stored in the System Data Repository (SDR).

- Distributed File System (DFS) for AIX, which is used for DFS interoperability.

Figure 8-1 shows the components of a GPFS cluster for an SP environment and the relations between them.

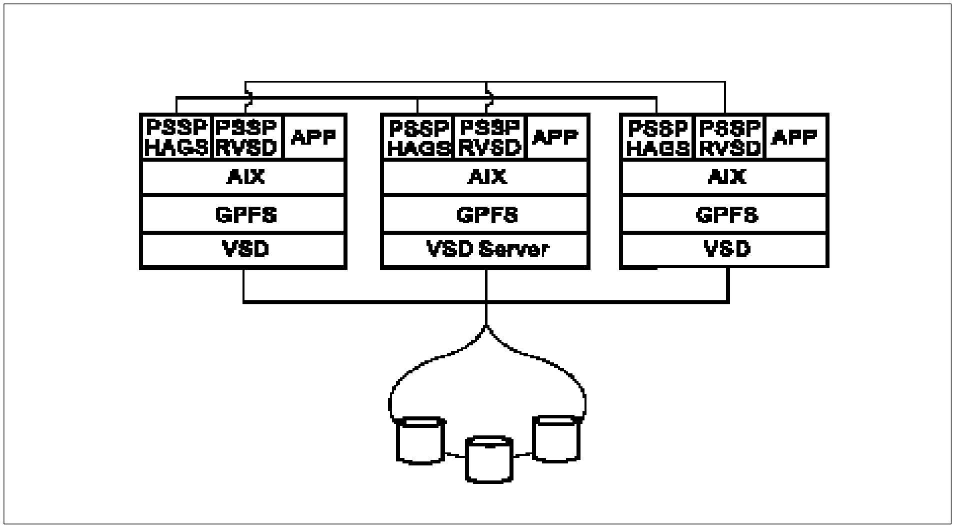


Figure 8-1 A GPFS configuration in an SP environment

8.2.2 GPFS in an HACMP environment

A GPFS cluster in an HACMP environment is formed by a group of RS/6000 and IBM @server pSeries machines taking part in an HACMP/ES cluster.

In this environment, every node in the GPFS cluster must be connected to the disks through their physical attachment, so that each disk is available to all nodes for concurrent access (see Figure 8-2 on page 161). In these conditions, a switchless SP system with nodes belonging to an HACMP/ES cluster is also consider an HACMP environment.

The GPFS cluster in an HACMP environment is a subset of nodes belonging to an HACMP/ES cluster. A single GPFS cluster can be defined within an HACMP/ES cluster, and more nodesets can be defined within the cluster. The boundaries of the GPFS cluster are established with the **mmcrcluster** command. The cluster size depends on the available disk technology. For GPFS 1.5 cluster sizes, see Section 8.4, “Hardware and software requirements” on page 165.

The configuration data for the GPFS cluster is kept on a primary server, but an optional secondary server may be specified at creation time using the **mmcrcluster** command. The two servers are members of the same cluster.

GPFS is designed to operate with the following software products:

- AIX, which provides:
  - The basic operating system services. AIX receives GPFS file system calls which are redirected to the GPFS daemon.
  - The Logical Volume Manager (LVM) subsystem. The volume groups containing the GPFS data are v*aried on* by GPFS administrative scripts to each node in the cluster at proper times for file system operations. This GPFS operation is done without locking the disks that make up the volume group, using the **varyonvg** command with the -u flag.

  > **Note:** The Concurrent Logical Volume Manager (CLVM) component of AIX is not required for a GPFS cluster in an HACMP environment.

- HACMP/ES, which provides:
  - The basic cluster environment for GPFS. The HACMP/ES Object Data Manager (ODM) is used for management operations.
  - The Group Services subsystem needed in failure situations.
- DFS for AIX, which provides DFS interoperability.

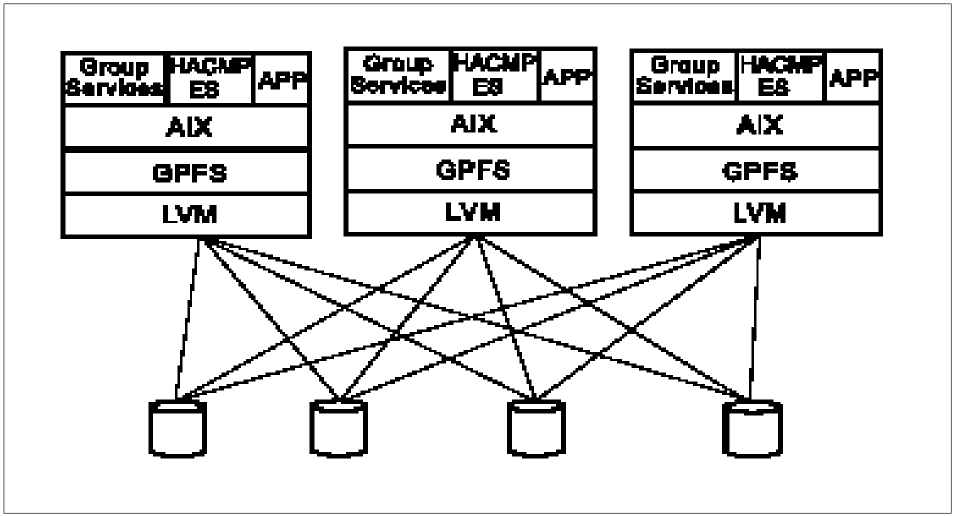


Figure 8-2 A GPFS configuration for an HACMP cluster environment

8.3 GPFS 1.5 enhancements

GPFS 1.5 improves performance and extends its capabilities by allowing new hardware attachments, supports new software releases and enhances command set capabilities.

8.3.1 Review of changes in GPFS 1.4

Though delivered without any new PSSP release, GPFA 1.4 brings enhancements that have been inherited by GPFS 1.5 for PSSP 3.4. GPFS 1.4 extends the capabilities of previous versions to clusters of RS/6000 and IBM @server pSeries systems using HACMP/ES software. Release 4 provides data-sharing capabilities across Serial Storage Architecture (SSA) disks attached to all the systems in the cluster. The SSA disks must be accessed by every node in the cluster via its physical attachment. Due to limitations of the SSA adapter, a maximum of eight nodes may form a nodeset.

Along with cluster support in GPFS 1.4, new commands are provided:

- `mmcrcluster`

  Used for creating a GPFS cluster from a set of nodes belonging to an HACMP/ES cluster. There can be one GPFS cluster per HACMP/ES cluster.

- **mmchcluster**

  Used to change/synchronize the primary or secondary GPFS SDR server for a cluster.

- **mmaddcluster**

  Adds nodes to a GPFS cluster.

- **mmdelcluster**

  Deletes nodes from a GPFS cluster.

- **mmlscluster**

  Displays the current configuration for a GPFS cluster.

For more information about GPFS commands in an HACMP cluster environment, refer to *IBM General Parallel File System for AIX: Administration and Programming Reference,* SA22-7452.

GPFS 1.4 also provides enhancements for both SP and HACMP environments:

- The **mmstartup** command starts GPFS daemons on one or multiple nodes. Various flags allow the GPFS subsystem to start for a nodeset or all nodes from a nodeset or from specified nodes.
- With the support of the dynamic **mmaddnode** command, there is no need to shut down the cluster to integrate new nodes when the communication protocol is TCP/IP and a quorum exists for the nodeset.
- The **mmlsconfig** command lists configuration data for all nodes from a nodeset. The -C flag has been provided for listing configuration information for a nodeset.
- Support for a single-node quorum in a two-node environment is provided. The single-node quorum algorithm allows the GPFS daemon to continue operation even if the peer node fails.

8.3.2 New hardware support

GPFS 1.4 supports only SSA disk attachments in an HACMP environment. The new Version 1.5 with Fibre Channel (FC) disk support now extends GPFS cluster capabilities. FC attachment support for HACMP environments includes:

- 2105 Fibre Channel—Enterprise Storage Server (ESS).
- 9x00 Hashed Shared Disk (HDS) Disk Storage.

Restriction: Clusters using FC disks do not support single-node quorum configurations.

GPFS 1.5 also uses 2105 FC disks for SP environments since they are now directly supported in VSD 3.4. Support is also available with PSSP 3.2 or later.

FC ESS 2105 systems can be configured using the Subsystem Device Driver (SDD) in both HACMP and SP environments. For more information about SDD, see “SDD overview” on page 62.

SP environments using GPFS 1.5 can exploit SP Switch2 configurations including single- and double-plane configurations. In double-plane configurations, two switch adapters are used on a node. Aggregate IP addressing can be used to improve switch communication performance and provide high availability features. Refer to “Multilink adapter support” on page 71 for detailed information.

8.3.3 Support for AIX 5L

GPFS 1.5 supports AIX 5L. GPFS 1.5 can be installed in both AIX 5L Version 5.1 and AIX Version 4.3.3 environments. However, if your system has AIX 5L Version 5.1 installed you must install GPFS 1.5. No previous version is supported in AIX 5L Version 5.1.

Applications running on top of GPFS 1.5 are able to exploit the 64-bit enhanced support of AIX 5L Version 5.1 through GPFS programming interfaces. In order to use the 64-bit versions of the programming interface, you must recompile the code using the 64-bit options for the compiler. However, AIX kernel operations are supported in 32-bit mode only.

8.3.4 Increased maximum file system size

In previous releases of GPFS, file system size was limited to 9 TB. GPFS 1.5 supports file systems of up to 100 TB.

8.3.5 Default quota limits

Default quota values can be applied to users or groups writing to a GPFS file system. The default values complement the explicit quota setting associated with users and groups of a GPFS file system. If default quotas are not used, then all users and groups without an explicit quota setting have no limits.

Three new commands are provided to support default quotas:

- **mmdefquotaon**

  Activates default quota limits for a file system. At this point new file system users or groups have a default quota limit of 0, unless you change it using the **mmdefedquota** command. The default quota remains active across a file system after you restart GPFS services and remount the file system. To disable the default quota setting on a file system, use the **mmdefquotaoff** command.

- **mmdefedquota**

  Sets or changes the default quota limits for file system users or groups. This command can be applied only if the file system was created or updated using the -Q yes option. For more information, refer to the entries for the **mmcrfs** and **mmchfs** commands in the *IBM General Parallel File System for AIX: Administration and Programming Reference,* SA22-7452.

- **mmdefquotaoff**

  Turns off the default quota limits for a file system. If the default quota limits are deactivated, new file system users or groups have no default quota limit.

Some changes have been applied to the **mmlsquota** and **mmrepquota** commands to allow you to add new information concerning default quota settings. For example:

- **mmlsquota -d {-u | -g }**.
- The **mmrepquota** command has been updated to reflect user and group default quotas in a file system. It has also been enhanced to display a numerical user ID (-n flag).

For further information about these commands, refer to the *IBM General Parallel File System for AIX: Administration and Programming Reference,* SA22-7452.

See also Example A-2 on page 181 for an example of enabling a file system default quota.

8.3.6 Software enhancements for cluster environments

HACMP/ES Version 4.4.1 is supported by GPFS 1.5 for a HACMP environment.

The new **mmcrlv** command complements the **mmcrvsd** command used for an SP environment. The **mmcrlv** command creates volume groups and logical volumes for use by GPFS, simplifying the LVM tasks associated with logical volume creation.

The **mmcrlv** command receives a disk descriptor file as a parameter, which contains the target disks along with their disk usage and failure group. Consider an example using the format *DiskName:::DiskUsage:FailureGroup:*

DiskName	The device name of your disk
DIskUsage	May be data and metadata, data only, or metadata only
FailureGroup	A number for grouping disks with a common failure point

As a result of issuing the **mmcrlv**, for each disk, a volume group containing a logical volume is created. The **mmcrlv** command imports all volume groups to all existing nodes in the GPFS cluster, not just those from the nodeset. For details about **mmcrlv**, see the *IBM General Parallel File System for AIX: Administration and Programming Reference,* SA22-7452.

The **mmaddcluster** command has been improved. When you add a new node to a GPFS cluster, existing GPFS disks can be in use while the cluster definition is imported to the newly added node.

8.4 Hardware and software requirements

This section lists the GPFS 1.5 hardware and software requirements for both SP and HACMP environments.

8.4.1 For an SP environment

This section describes the hardware and software requirements for an SP environment.

Hardware requirements

The following list provides a description of the hardware requirements for GPFS 1.5 in an SP environment:

- An RS/6000 SP cluster system with switch connectivity. The following systems can be considered:
  - An RS/6000 SP system
  - An RS/6000 SP system with attached RS/6000 servers or IBM @server pSeries servers
  - A Cluster of Enterprise Servers (CES) with RS/6000 servers or IBM @server pSeries servers
- An SP Switch, SP Switch2, or a two-plane SP Switch2
- Sufficient disk space for the file systems

Attention: The disks drives supported depend on VSD. The disk drives supported are:

- SCSI
- SSA
- Fibre Channel (including 2105, 3552, and EMC Symmetrix)

- There are some optional considerations to discuss. If you are planning to use LAPI as the communication protocol, one of the following switch adapters is required:
  - A TB3MX adapter or TB3MX2 adapter for SP Switch configurations
  - An SP Switch2 adapter for SP Switch2 configurations

  For more information about switch- and switch adapter support refer to *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment,* GA22-7280.

Software requirements

The following list provides a description of the software requirements for running GPFS 1.5 in an SP environment:

- AIX operating system. The following versions are supported by GPFS 1.5:
  - AIX Version 4.3.3 or later
  - AIX 5L Version 5.1 or later
- PSSP 3.4 or later
- The following software options:
  - ssp.basic
  - ssp.css
  - rsct.basic.rte
  - ssp.sysctl
  - ssp.ha_topsvcs.compat
  - vsd.cmi
  - vsd.sysctl
  - vsd.vsdd
  - vsd.rvsd.hc
  - vsd.rvsd.rvsdd
  - vsd.rvsd.scripts

- Optionally, DFS for AIX Version 3.1.0 or later for exploiting DFS interoperability

8.4.2 For an HACMP environment

The following sections describe the hardware and software requirements for GPFS 1.5 in an HACMP environment.

Hardware requirements

The following list provides a description of the hardware requirements for GPFS 1.5 in an HACMP environment:

- An HACMP/ES cluster of RS/6000 servers and IBM @server pSeries servers.
- A pool of external disks with the following characteristics:
  - SSA or Fibre Channel attachment. The cluster size in these environments depends on the latest disk technology available:
    - Two to eight nodes for SSA
    - Three to 32 nodes for Fibre Channel (the single-node quorum configuration is not supported)
  - A maximum 1024 shared disks or disk arrays
  - A disk configuration allowing each disk to be accessible to any node in the cluster
- An IP network with sufficient bandwidth (we recommend a minimum of 100 Mb per second)

> **Notes:**
>
> - A proper disk configuration should take into consideration the type of disks (attachment, capacity, I/O speed), Redundant Array of Independent Disks (RAID) configuration, and the requirements of other applications running on the cluster nodes.
> - Protection against disk failure is possible through GPFS replication or RAID devices, but not through LVM mirroring.

Software requirements

The following list provides a description of the software requirements for running GPFS 1.5 in an HACMP environment:

- AIX operating system. The following releases are supported:
  - AIX Version 4.3.3 or later

  - AIX 5L Version 5.1 or later

- HACMP/ES Version 4.4.1 or later. The nodes must be configured as members of the HACMP cluster.
- Optionally, DFS for AIX 3.1.0 or a later modification, to exploit DFS interoperability.

8.5 Migration, coexistence and compatibility

This section describes migration, coexistence, and compatibility considerations with GPFS 1.5.

8.5.1 Migration considerations

All nodes within a GPFS nodeset must be upgraded to the same level of GPFS, and at the same time.

> **Note:** A system using GPFS nodesets with GPFS 1.5 installed and other nodesets with lower levels installed must have GPFS 1.5 installed on the CWS.

In order to use the new file system functions in GPFS 1.5, you must explicitly authorize these changes by issuing the **`mmchfs -V`** command. Once all the nodesets are running GPFS 1.5, issue the **`mmchconfig release=LATEST`** command to change the GPFS system files to the latest format. This reduces the number of configuration files stored in the GPFS cluster data, improving the performance of the GPFS administration commands.

> **Attention:**
>
> - Once you have issued the **`mmchfs -V`** command, it is not possible for a previous version of GPFS to use the file system.
> - Issue the **`mmchconfig release=LATEST`** command only if all the nodesets are at the GPFS 1.5 level. After these steps are done, you may not revert to the previous level of GPFS.

In order to use the 64-bit versions of the GPFS programming interfaces, you must recompile your code using the appropriate 64-bit options for your compiler.

Staged migration

If you plan to test the new version of GPFS on a multi-nodeset cluster, you can use one of the nodesets to upgrade to the latest release.

Alternatively, you can create an additional GPFS nodeset for test purposes by deleting nodes from the existing nodeset using the **mmdeletenode** command and then grouping them in a new nodeset using the **mmconfig** command. You can delete a node from a nodeset using the **mmdelnode** command and add it to a new nodeset using the **mmaddnode** command. You need not move a file system to the nodes, but you can associate it with a nodeset by issuing the **mmchfs** command with the -C option. If you plan to use the file system at the latest level, convert it using **mmchfs -V** (bearing in mind that you will then be unable to use it with a previous version of GPFS).

You can now migrate the remaining nodes to GPFS 1.5. Once you have decided to permanently accept this release for all the nodes, convert the file systems to the latest version and complete the migration by issuing the **mmchconfig release=LATEST** command. The GPFS administration commands no longer maintain two versions of the SDR configuration data.

Attention: We recommend a minimum of six nodes in the system in order to use multiple GPFS nodesets and perform a staged migration. Otherwise, it is better to perform a full migration.

For details on the steps involved in the staged migration process, refer to the *IBM General Parallel File System for AIX: Concepts, Planning and Installation Guide,* GA22-7453.

Full migration

In this method, all the nodes in the system are upgraded to the new Release 5. You have to stop GPFS on all nodes and perform the migration on all nodes at the same time.

Once the you have decide to accept the new level, issue the **mmchfs -V** command for each file system to convert them to the latest level, and issue the **mmchconfig release=LATEST** command on each nodeset.

For details on the steps involved in the full migration process, refer to *IBM General Parallel File System for AIX: Concepts, Planning and Installation Guide,* GA22-7453.

8.5.2 Coexistence

Due to the changes in the token management function, it is not possible to run different levels of GPFS in the same nodeset. However, it is possible to run multiple nodesets at different levels of GPFS. If a system is going to have some GPFS nodesets with GPFS 1.5 installed and others with lower levels installed, the CWS must have GPFS 1.5 installed.

A GPFS file system may be shared only between members of the same nodeset. A node from a GPFS cluster may be contained only in a single nodeset. In order to use LAPI as the communication protocol, all nodes in the GPFS nodeset must use it. There is no coexistence between LAPI and TCP/IP.

Due to common components shared by GPFS, IBM Multi-Media Server, and IBM Video Charger, the kernel extensions for GPFS cannot coexist with these products on the same system.

8.5.3 Compatibility

All applications that executed on previous releases of GPFS execute on the new level of GPFS. File systems created under previous GPFS releases may continue to be used at the new level.

However, once a GPFS file system has been explicitly changed by issuing the **mmchfs -V** command, the disk images can no longer be read by a back-level file system. You must recreate the file system from a backup medium and restore the content if you choose to go back after this command has been issued.

A

Additional information

This appendix contains the following:

- Command changes
- Software disk space requirements
- Useful scripts

New or changed commands

The following describes changes that allow **dsh** to work in a secure remote-command environment.

dsh

Purpose

dsh issues commands to a group of hosts in parallel.

Syntax

```
dsh [-q]

dsh [-h]

dsh [-i] [-v] [-c] [-a] [-t] [-G] [-d] [-D] [-l login_name]
[-N node_group, node_group,...] [-w { host_name[, host_name...] | .}]
[-f fanout_value] [-o "flags_and_parms"] [command]
```

New or changed flags

-t Specifies the target shell syntax. This flag is used to create the syntax of the internal environment variables passed with the command. Supported shells are ksh and csh. The default is ksh.

-d Forwards the DCE credentials for authentication by using the -f flag on the **rsh** command. If this flag is not set, the **dsh** command does not forward DCE credentials. This option is relevant only for use with **rsh**.

-D Recursively forwards the DCE credentials for authentication by using the -F flag on the **rsh** command. If this flag is not set, the **dsh** command does not forward DCE credentials. This option is only for use with **rsh**.

-l login_name Specifies a remote user name under which to execute the commands. If -l is not used, the remote user name is the same as your local user name. (This is lowercase l, as in list.)

-o flags_and_parms Specifies flags and parameters to be passed to the remote shell command. You must surround the set of flags and parameters with either single or double quotes. The flags and parameters must be passed to your remote command program, which defines their use.

Environment variables

The **dsh** command uses the AIX **rsh** command to function or a secure remote command of your choosing depending on the setting of the DSH_REMOTE_CMD and the RCMD_PGM environment variables. If the environment variables are not set, the default is DSH_REMOTE_CMD=/bin/rsh and RCMD_PGM=rsh. If RCMD_PGM=secrshell and DSH_REMOTE_CMD is not specified, the default is DSH_REMOTE_CMD=/bin/ssh.

DSH_REMOTE_CMD Remote command executable.

RCMD_PGM Remote command method, either rsh or secrshell.

HN_METHOD Reliable or initial.

This environment variable is used with the **dsh -a** or **dsh -a -v** option to determine whether to use the reliable_hostname or initial_hostname when building the working collective. The default is the initial_hostname.

DSHPATH PATH to the command on the remote node.

The path used when resolving the **dsh** command on the target nodes.

SDR changes

Table A-1 describes the new SDR SP_Restricted class attributes, which holds information about which remote command should be used in the SP environment.

Table A-1 SP_Restricted class attributes

Attribute name	Type	Description	Comments
restrict_root_rcmd	S	Whether restricted root access is enabled or not	Value=true or false; default is false.
rcmd_pgm	S	Name of the remote command executable	Value=rsh or secrshell; default is rsh.
dsh_remote_cmd	S	Name of the remote command executable	Default in the SDR is null.
remote_copy_cmd	S	Name of the remote copy command executable	Default in the SDR is null.

The Volume_Group class SDR pv_list attribute now also holds information about SAN-attached boot disks, as shown in Table A-2.

Table A-2 The SDR Volume_Group class pv_list attribute

Attribute name	Type	Description	Comments
pv_list	S	A list of physical volumes (pv). Valid format for hdisk specification is hdiskn,hdiskn,...,hdiskn Valid format for connwhere, location, PVID, or SAN_DISKID specification is the corresponding AIX attribute value for any combination of those physical volumes, separated by colons: pv:pv:...:pv	Initially set to hdisk0.

LPP sizes and disk space requirements

The following sections provide information on LPP sizes and disk space requirements.

Space requirements for AIX install

Table A-3 provides a comparison of disk space allocated per directory according to base operating system level.

Table A-3 Space allocated during base AIX install

Directory	AIX Version 4.3.3	AIX 5L Version 5.1
/	4 MB	8 MB
/usr	294 MB	385 MB
/var	4 MB	4 MB
/tmp	16 MB	32 MB
/opt	N/A	4 MB

PSSP installp image sizes

Table A-4 provides information on the disk space required for storing the different PSSP installp images.

Table A-4 Space required for storing PSSP installp images

Installp image|	Space required	Description
rsct.basic	16 MB, rsct.basic.rte, rsct.basic.sp	This has the RSCT basic components required in all realms on the control workstation.
rsct.clients	500 KB rsct.clients.rte, rsct.clients.sp	RSCT clients required in all realms.
rsct.core	900 KB rsct.core.utils	This has the RSCT core components required in all realms on the control workstation and all nodes.
ssp	150 MB	This has the base PSSP components. It must be on the control workstation.
ssp.resctr	4 MB	The resource center image is optional on the control workstation and nodes.
vsd ssp.vsdgui	8 MB	The VSD, HSD, and RVSD optional image must be on the control workstation and nodes that will have or use virtual shared disks. The ssp.vsdgui image is the IBM Virtual Shared Disk Perspective. It must be on the control workstation and is optional on nodes.

VSD LPP space requirements

Table A-5 provides information on disk space requirements for storing the different VSD and RVSD LPPs.

Table A-5 Space requirements for VSD and RVSD LPPs

Fileset	root during installation	usr during installation	var during execution
vsd.cmi	100 KB	270 KB	0
vsd.vsdd	16 MB for /var/adm/csd file system created in rootvg	490 KB	16 MB only if /var/adm/csd cannot be created in rootvg
vsd.hsd	0	220 KB	0
vsd.sysctl	0	490 KB	0
vsd.rvsd.scripts	0	250 KB	8 MB only if /var/adm/csd could not be created in rootvg
vsd.rvsd.rvsdd	0	320 KB	0
vsd.rvsd.hc	0	300 KB	0

LoadLeveler disk space requirements

Table A-6 provides information on disk space required during LoadLeveler installation.

Table A-6 Disk space required during LoadLeveler install

Directory	Space required
release directory (/usr/lpp/LoadL/full)	27 MB
local directory	15 MB (minimum)
home directory	no limits unless same as release or local directory
release directory for Submit-only (/usr/lpp/LoadL/so)	11 MB
PDF documentation directory (/usr/lpp/LoadL/pdf)	2 MB

Directory	Space required
HTML pages directory (/usr/lpp/LoadL/html)	2 MB
Configuration Tasks Directory (usr/lpp/LoadL/codebase)	1 MB

PE disk space requirements

Table A-7 provides information on the disk space required during PE installation.

Table A-7 Disk space required during PE install

PE fileset	/usr	/tmp	/etc
ppe.html	15 MB	not applicable	not applicable
ppe.man	3.7 MB	not applicable	not applicable
ppe.pdf	10 MB	not applicable	not applicable
ppe.poe	13 MB	256 KB	5 KB
ppe.xprofiler	3.2 MB	not applicable	not applicable
ppe.perf	23.5 MB	not applicable	5 KB
ppe.pvt	2 MB	not applicable	not applicable
ppe.dpcl	42 MB	not applicable	15 KB
Note: ppe.dpcl is required when installing ppe.perf.			

Useful scripts

The script.cust script functions to build up the secure remote-command environment, as shown in Example A-1.

Example: A-1 script.cust (secrshell setup portion)

```
#---------------------------------------------------------------------------#
# Secure Remote Commands                                                    #
#The following code gives an example of how a secure remote command        #
#product can be automatically installed and configured on a node           #
#being installed by PSSP. This example shows a specific product.          #
#You should modify it to install the secure remote command product of      #
#your choice.                                                              #
#                                                                           #
#Before this code is executed the following items must be completed        #
#on the control workstation:                                               #
```

```
#   1) Choose any secure remote command product, install and configure  #
#      it on the CWS. The secure remote command daemons must be active      #
#      before node installation is started. Make sure the root home         #
#      directory is owned by root on the CWS. This is necessary for AIX #
#     releases prior to AIX 5L Version 5.1.
#
#   2) Considerations for strict hostname checking:                         #
#      * This example will work if strict hostname                          #
#        checking is disabled for your secure remote command product.       #
#      * To install a node with strict hostname checking enabled the        #
#        administrator must add additional function to firstboot.cmds to    #
#        build a known_hosts file on the CWS and on each Boot Install Node  #
#        with an entry for each node to be installed by that BIS            #
#         using secure remote commands.                                     #
#   3) Put the product installation images into a directory which can be    #
#       exported from the CWS.  This code will nfs mount that directory on  #
#       each node to be installed.                                          #
#       This example uses directory /spdata/sys1/ssh.                       #
#   4) The DSA/RSA keys for root must be already generated on the CWS and   #
#       on any BIS nodes before a node can be installed from a BIS node.    #
#       *  Place a copy of the CWS root public RSA/DSA key(s) in            #
#          /tftpboot with read-any permissions.                             #
#       *  If there are any BIS nodes, place a copy of the                  #
#          BIS node's root's public RSA/DSA key(s) into its                 #
#          /tftpboot directory with read-any permissions.                  #
#       *  When script.cust runs the public keys (RSA/DSA) must be copied    #
#          from the CWS and the BIS node (if not the CWS) to the node       #
#          being installed.                                                 #
#       *  Some versions of secure remote copy require that the secure      #
#          remote command code be in or linked to the /usr/local/bin        #
#          directory. Create this directory and link the secure remote      #
#          command executable. This must be done before the node install on #
#          the Control Work Station.                                        #
# Below is a sample of a script that installs a secure shell product on a node#
#    * Increase space in the /usr, /var, and /tmp directories to prevent    #
#      failures during the install                                          #
#        (chfs calls)                                                       #
#    * Change owner of root home directory to root. This is only needed    #
#      for AIX 4.3.3                                                       #
#    * Mount the directory containing the installation images and           #
#      perform the install                                                  #
#        (mkdir through unmount call)
#
#                                                                           #
#    * Modify /etc/inittab to start the ssh daemon before any PSSP code    #
#   is started on the nodes. The placement in inittab is critical,      #
#       because the ssh daemon must be running before the PSSP code         #
#   uses it during the first boot process.                              #
#       When migrating, the sshd daemon must be removed from /etc/inittab   #
#       in order to ensure that it is put back in the correct order.        #
#       (code for rmitab, mkitab)                                          #
#                                                                           #
#    *  tftp keyfiles from the CWS and BIS nodes to the node being installed#
#                                                                           #
#    *  Link the secure remote command executable                           #
#---------------------------------------------------------------------------#
#SECRSHELL=/bin/ssh
#if [[ ! -x $SECRSHELL ]];then
# /usr/sbin/chfs -a size=+60000 /usr
# /usr/sbin/chfs -a size=+60000 /tmp
# /usr/sbin/chfs -a size=+60000 /var
# /bin/mkdir /spdata/sys1/ssh
# mount c55s:/spdata/sys1/ssh /spdata/sys1/ssh
# chown root <root_home_dir>
# /bin/mkdir /usr/local
# cd /spdata/sys1/ssh/zlib-1.1.3
# make install
# cd /spdata/sys1/ssh/openssl-0.9.6
# make install
# mkdir /tmp/ssh
# cp -rp /spdata/sys1/ssh/openssh-2.9p1  /tmp/ssh
# cd /tmp/ssh/openssh-2.9p1
# make install
# cd /tmp
# rm -r /tmp/ssh
# /usr/sbin/umount /spdata/sys1/ssh
# /usr/sbin/rmitab sshd
# /usr/sbin/mkitab -i rctcpip “sshd:2:once:/usr/sbin/sshd”
#fi

#---------------------------------------------------------------------------#
# If this is a BIS node, the DSA/RSA keys for root must be generated and the#
# public key must be copied to the /tftpboot directory for the node(s)  the #
# BIS node serves.  In addition the public key must be sent to the CWS, in  #
# order to issue secure remote calls to the CWS during the install.         #
# The public key will be obtained by the nodes when running script.cust.    #
# To copy the public key to the CWS, the firstboot.cust file must be        #
# must be modifed to run firstboot.cmd which will have the CWS request the  #
# tftp of the public key to the CWS and install it in the secure remote     #
# command authorization file.  See firstboot.cust for information on        #
# enabling firstboot.cmds.                                                  #
# Secondly the secure remote command configuration file should be copied    #
# the CWS to the node. This configuration file should have strict host name #
# checking disabled.                                                        #
#---------------------------------------------------------------------------#
# /bin/ssh-keygen -f /.ssh/identity -N ““
# /bin/ssh-keygen -d -f /.ssh/id_dsa -N ““
# /bin/cp /.ssh/identity.pub /tftpboot/identity.pub.$HOSTNAME
# /bin/cp /.ssh/id_dsa.pub /tftpboot/id_dsa.pub.$HOSTNAME

#tftp -o /usr/etc/ssh_config <cws_ip_address> /tftpboot/ssh_config
#---------------------------------------------------------------------------#
# The following will enable PSSP code to use secure remote commands to the  #
# nodes without prompts by copying the root’s public RSA/DSA keys to the    #
# nodes being installed from the CWS and BIS nodes.                         #
# The DSA/RSA   keys  for root must already have been generated on the CWS  #
# and the BIS nodes. The public key(s) must have been copied to the /tftpboot#
# directory on the CWS and the BIS nodes.This must be done for each secure  #
# remote command protocol enabled (RSA/DSA).                                #
# The public key for root will be tftp’d from the /tftpboot directory       #
# on the CWS and the boot install node (BIS), if the BIS is not the CWS,    #
# to the node after installation and before reboot.                        #
# This setup enables PSSP installation and configuration scripts to run     #
# without password prompts using a secure remote method as long as strict  #
# host checking is disabled during the install.                             #
#Note: Fill in the appropriate root home directory below <root_home_dir>     #
#     and control workstation IP address below <cws_ip_addr>                #
#---------------------------------------------------------------------------#
#Gets root public key from the BIS to the nodes for protocol 1
#tftp -o <root_home_dir>/.ssh/identity.pub.tmp $SERVER_IP_ADDR
/tftpboot/identity.pub
#cat <root_home_dir>/.ssh/identity.pub.tmp >>
<root_home_dir>/.ssh/authorized_keys

#Gets root public key from the BIS to the nodes for protocol 2
#tftp -o <root_home_dir>/.ssh/id_dsa.pub.tmp $SERVER_IP_ADDR
/tftpboot/id_dsa.pub
#cat <root_home_dir>/.ssh/id_dsa.pub.tmp >>
<root_home_dir>/.ssh/authorized_keys2

#IF BIS node is not the CWS, need to tftp the keyfiles for the CWS to the node.
#Fill in the <cws_ip_addr> with the correct control workstation ip address

#Gets root public key from the CWS to the nodes for protocol 1
#tftp -o <root_home_dir>/.ssh/identity.pub.tmp <cws_ip_addr>
/tftpboot/identity.pub
#cat <root_home_dir>/.ssh/identity.pub.tmp >>
<root_home_dir>/.ssh/authorized_keys


#Gets root public key from the CWS to the nodes for protocol 2
#tftp -o <root_home_dir>/.ssh/id_dsa.pub.tmp <cws_ip_addr> /tftpboot/id_dsa.pub
#cat <root_home_dir>/.ssh/id_dsa.pub.tmp >>
<root_home_dir>/.ssh/authorized_keys2

#---------------------------------------------------------------------------#
```

```
# The following will create and link the secure remote command            #
# executable in the /usr/local/bin directory. This is necessary for some    #
# implementations of secure remote copy                                     #
# which automatically look for ssh in this directory.                       #
#---------------------------------------------------------------------------#
# /bin/mkdir /usr/local/bin
# ln -s /bin/ssh /usr/local/bin/ssh
```

Example A-2 explains how the default quota for a GPFS file system is assigned.

Example: A-2 Assigning a default quota for a GPFS file system

```
The example scenario was tested an SP environment. A file system is created
over 3 VSDs. We suppose that GPFS services are started and a nodeset gpfs1 was
created. The commands are issued from a node in the nodeset.

Step1. Create a new file system
root@sp4n01:> mmcrfs /gpfs1 /dev/gpfsvsd \
>“gpfs1vsd1:::dataAndMetadata:;\
>gpfs1vsd2:::dataAndMetadata:;\
>gpfs1vsd3:::dataAndMetadata:”\
>-C gpfs1 -Q yes


mmrts: Executing "tsctl showCfgValue maxblocksize" on node
sp4n01.msc.itso.ibm.com
mmrts: Executing "tscrfs /dev/gpfsvsd -F /var/mmfs/tmp/tsddFile.mmcrfs.25738 -c
0 -I 16384 -i 512 -M 1 -n 32 -R 1 -s roundRobin
 -w 0" on node sp4n01.msc.itso.ibm.com

GPFS: 6027-531 The following disks of gpfsvsd will be formatted on node
sp4n01.msc.itso.ibm.com:
    gpfs1vsd1: size 983040 KB
    gpfs1vsd2: size 983040 KB
    gpfs1vsd3: size 983040 KB
GPFS: 6027-540 Formatting file system ...
Creating Inode File
Creating Allocation Maps
Clearing Inode Allocation Map
Clearing Block Allocation Map
Flushing Allocation Maps
GPFS: 6027-572 Completed creation of file system /dev/gpfsvsd.
GPFS: 6027-623 All disks up and ready
mmcrfs: Propagating the changes to all affected nodes.
This is an asynchronous process.
```

```
root@sp4n01:> mmlsfs gpfsvsd -Q
flag value          description
---- -------------- -----------------------------------------------------
 -Q  user;group     Quotas enforced
     none           Default quotas enabled


Step2. Mount de filesystem:
root@sp4n01:> mount /gpfs1

Step3. Activate default quota on /gpfs1 file system:
root@sp4n01:> mmdefquotaon gpfsvsd
root@sp4n01:> mmlsfs gpfsvsd
flag value          description
---- -------------- -----------------------------------------------------
 -Q  user;group     Quotas enforced
     user;group     Default quotas enabled

Step4. Edit default quota ( as example for users only )
root@sp4n01:> mmdefedquota -u /dev/gpfsvsd
```

```
*** Edit quota limits for USR DEFAULT entry
NOTE: block limits will be rounded up to the next multiple of the block size.
gpfsvsd: blocks in use: OK, limits (soft = 4096M, hard = 5120M)
        inodes in use: 0, limits (soft = 0, hard = 0)
~






:wq
```

```
After some user operations with the file system, the output of mmlsquota looks
like:

root@sp4n01:> mmlsquota -d -u
*** Report for USR and GRP quotas on gpfsvsd
                     Block Limits                            |
File Limits
Name       type          KB      quota      limit   in_doubt    grace |
files   quota    limit in_doubt    grace entryType
root       USR            0          0          0          0     none |
0       0        0        0     none default on
andrei     USR         6224    4194304    5242880       4016     none |
1       0        0       19     none d
```

```
marcel     USR            0    4194304    5242880       5120     none |
0       0        0        1    none d
mihai      USR            8    4194304    5242880       2040     none |
1       0        0        0    none d
system     GRP         6232          0          0      14248     none |
2       0        0       38    none default on
staff      GRP            0          0          0      10240     none |
0       0        0       40    none d
```

Related publications

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

IBM Redbooks

For information on ordering these publications, see “How to get IBM Redbooks” on page 187.

- *PSSP 3.2: RS/6000 SP Software Enhancements*, SG24-5673

Other resources

These publications are also relevant as further information sources:

- *ESSL Products General Information*, GC23-0529
- *ESSL V3R3M0 for AIX Guide and Reference*, SA22-7272
- *HACMP V4.3 AIX: Administration Guide*, SC23-4279
- *HACMP V4.3 AIX: Installation Guide*, SC23-4278
- *HACMP V4.3 AIX: Planning Guide*, SC23-4277
- *HACMP V4.3 Concepts and Facilities*, SC23-4276

- *IBM General Parallel File System for AIX: Concepts, Planning and Installation Guide*, GA22-7453
- *IBM LoadLeveler for AIX 5L: Using and Administering*, SA22-7311
- *IBM Parallel Environment for AIX: Installation*, GA22-7418
- *IBM RS/6000 SP: Planning Volume 1, Hardware and Physical Environment*, GA22-7280
- *IBM RS/6000 SP: Planning Volume 2, Control Workstation and Software Environment*, GA22-7281
- *IBM Subsystem Device Driver: Installation and User's Guide*, SC26-7425
- *Implementing a Firewalled RS/6000 SP System Version 3, Release 4*, GA22-7874
- *Installation Memo for IBM LoadLeveler V2R2*, GI10-0642
- *Parallel ESSL for AIX V2R3 Guide and Reference*, SA22-7273
- *Parallel Environment for AIX: Hitchhiker's Guide*, SA22-7424
- *Parallel Environment for AIX: Messages*, GA22-7419
- *Parallel Environment for AIX: MPI Programming Guide*, SA22-7422
- *Parallel Environment for AIX: MPI Subroutine Reference*, SA22-7423
- *Parallel Environment for AIX: Operations and Use, Volume 1*, SA22-7425
- *Parallel Environment for AIX: Operations and Use, Volume 2*, SA22-7426
- *PSSP for AIX: Administration Guide*, SA22-7348
- *PSSP for AIX: Installation and Migration Guide*, GA22-7347
- *PSSP for AIX: Managing Shared Disks*, SA22-7349

Referenced Web sites

These Web sites are also relevant as further information sources:

- GPFS for Linux environment

  `http://www-1.ibm.com/servers/eserver/clusters/software/gpfs.html`

- GPFS home page

  `http://www-1.ibm.com/servers/eserver/clusters/software`

- Online documentation for GPFS on AIX and Linux

  `http://www.rs6000.ibm.com/resource/aix_resource/sp_books/gpfs/index.html`

- ESSL and Parallel ESSL online documentation

  `http://www.ibm.com/servers/eserver/pseries/library`
- ESSL Library

  `http://www.ibm.com/servers/eserver/pseries/software/sp/essl.html`
- Project eLiza

  `http://www.ibm.com/servers/eserver/introducing/eliza/index.html`
- AIX CSM

  `http://www.alphaworks.ibm.com/tech/aixcsm`
- AIX Toolbox for Linux Applications

  `http://www-1.ibm.com/servers/aix/products/aixos/linux/download.html`
- IBM @server Support - Fixes

  `http://techsupport.services.ibm.com/server/fixes`
- Message Passage Interface Forum

  `http://www.mpi-forum.org`

How to get IBM Redbooks

You can search for additional Redbooks or Redpieces, view, download, or order hardcopy from the Redbooks Web site:

ibm.com/redbooks

You can also download additional materials (code samples or diskette/CD-ROM images) from this Redbooks site.

Redpieces are Redbooks in progress; not all Redpieces become Redbooks, and sometimes just a few chapters are published this way. The intent is to get the information out much faster than the formal publishing process allows.

IBM Redbook collections

Redbooks are also available on CD-ROM. Click the CD-ROM button on the Redbooks Web site for information about all the CD-ROMs offered, as well as updates and formats.

Special notices

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 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

Abbreviations and acronyms

AIX	Advanced Interactive Executive
APAR	Authorized Program Analysis Report
API	Application Programming Interface
ATM	Asynchronous Transfer Mode
BI	Business Intelligence
BIS	Boot Install Server
BLACS	Basic Linear Algebra Communication Sub-programs
BLAS	Basic Linear Algebra Sub-programs
CATIA	Computer-graphics Aided Three-dimensional Interactive Application
CES	Clustered Enterprise Servers
CLVM	Concurrent Logical Volume Manager
CMI	Common Messaging Interface
CSM	Cluster System Management
CSP	Common Service Processor
CSPOC	Cluster Single Point of Control
CSS	Communication Subsystem Support
CUoD	Capacity Upgrade on Demand
CVSD	Concurrent Virtual Shared Disk
CWS	Control WorkStation
DASD	Direct Access Storage Device
DCE	Distributed Computing Environment
DFS	Distributed File System
DMAPI	Data Management API
DPCL	Dynamic Probe Class Library
ECC	Error Checking and Correction
ESS	Enterprise Storage Server
ESSL	Engineering and Scientific Subroutine Library
F/C	Feature Code
FCS	Frame Check Sequence
FFDC	First Failure Data Capture
GeoRM	Geographic Remote Mirror
GPFS	General Parallel File System
GS	Group Services
HACMP	High Availability Cluster Multi-Processing
HACMP/ES	High Availability Cluster Multiprocessing/Enhanced Scalability
HACWS	High Availability Control Workstation
HAGEO	RS/6000 High Availability Geographic cluster system
HAL	Hardware Abstraction Layer
HMC	Hardware Management Console
HPC	High Performance Computing
HPS	High Performance Switch
HSD	Hashed Shared Disk
IBM	International Business Machines Corporation
IETF	Internet Engineering Task Force
IKE	IP Key Encryption
ISS	Interactive Session Support

ITSO	International Technical Support Organization
JFS	Journaled File System
KDC	Key Distribution Center
KLAPI	Kernel Low Level Application Programming Interface
LAPACK	Linear Algebra Package
LAPI	Low-level Application Programming Interface
LDAP	Lightweight Directory Access Protocol
LL	LoadLeveler
LPAR	Logical Partition
LPP	Licensed Program Product
LUM	Licensed User Management
LUN	Logical Unit Number
LVM	Logical Volume Manager
M/T	Machine Type
MPI	Message Passing Interface
MSS	Master Switch Sequencing node
MUSPPA	Multiple User Space Processes Per Adapter
NAS	Network Authentication Service
NIM	Network Installation Manager
NSB	Network Switch Board
NSD	Network Shared Disk
ODM	Object Data Manager
OEM	Original Equipment Manufacturer
Parallel ESSL	Parallel Engineering and Scientific Subroutine Library
PCI	Peripheral Component Interconnect
PCT	Performance Collection Tool
PE	Parallel Environment
POSIX	Portable Operating System Interface for Computer Environments
POWER	Performance Optimization With Enhanced RISC (architecture)
PSSP	AIX Parallel System Support Programs
PTF	Program Temporary Fix
PTPE	Performance Toolbox Parallel Extension
PV	Physical Volume
PVID	Physical Volume Identifier
PVT	Profile Visualization Tool
RAID	Redundant Array of Independent Disks
RAS	Reliability, Availability, Serviceability
RISC	Reduced Instruction Set Computer/Cycles
RPM	Red Hat Package Manager
RRA	Restricted Root Access
RSCT	RS/6000 Cluster Technology
RVSD	Recoverable Virtual Shared Disk
SAMI	Service and Manufacturing Interface
SAN	Storage Area Network
SAS	Statistical Analysis System
ScaLAPACK	Scalable Linear Algebra Package
SCSI	Small Computer System Interface
SDD	Subsystem Device Driver
SDR	System Data Repository
SMIT	System Management Interface Tool
SMP	Symmetric MultiProcessors/ MultiProcessing

SP	IBM RS/6000 Scalable POWERparallel Systems (RS/6000 SP)
SPLAN	SP LAN
SPMD	Single Program-Multiple Data
SPS2	SP Switch2
SSA	Storage System Architecture
SSH	Secure Shell/Secret Shell
TB3PCI	SP Switch PCI Adapter
TSM	Tivoli Storage Manager
US	User Space
UTE	Unified Trace Environment
UTP	Unshielded Twisted Pair
VSD	Virtual Shared Disk
VT	Visualization Tool
WebSM	WebSMIT
WLM	WorkLoad Manager

Index

IBM

Redbooks

IBM @server Cluster 1600 and PSSP 3.4 Cluster Enhancements

IBM @server Cluster 1600 and PSSP 3.4 Cluster Enhancements

Provides highlights of the latest PSSP clustered software

Describes the latest hardware supported by PSSP 3.4

Discusses migration and coexistence

This redbook applies to PSSP Version 3, Release 4 for use with the AIX Operating System Version 5, Release 1, and Version 4, Release 3, modification 3.

This redbook details the new features and functions of PSSP 3.4, including the supported hardware. The redbook describes the changes to the product, allowing cluster professionals a convenient and detailed look at the latest PSSP enhancements.

Other SP-related products have also announced new releases. This redbook discusses the following:

- General Parallel File System (GPFS) 1.5
- LoadLeveler 3.1
- Parallel Environment (PE) 3.2
- Engineering and Scientific Subroutine Library (ESSL) 3.3 and Parallel ESSL 2.3

This redbook is for IBM customers, Business Partners, IBM technical and marketing professionals and anyone seeking an understanding of the new hardware and software components and improvements included in this IBM @server announcement.

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

SG24-6604-00 ISBN 0738423440