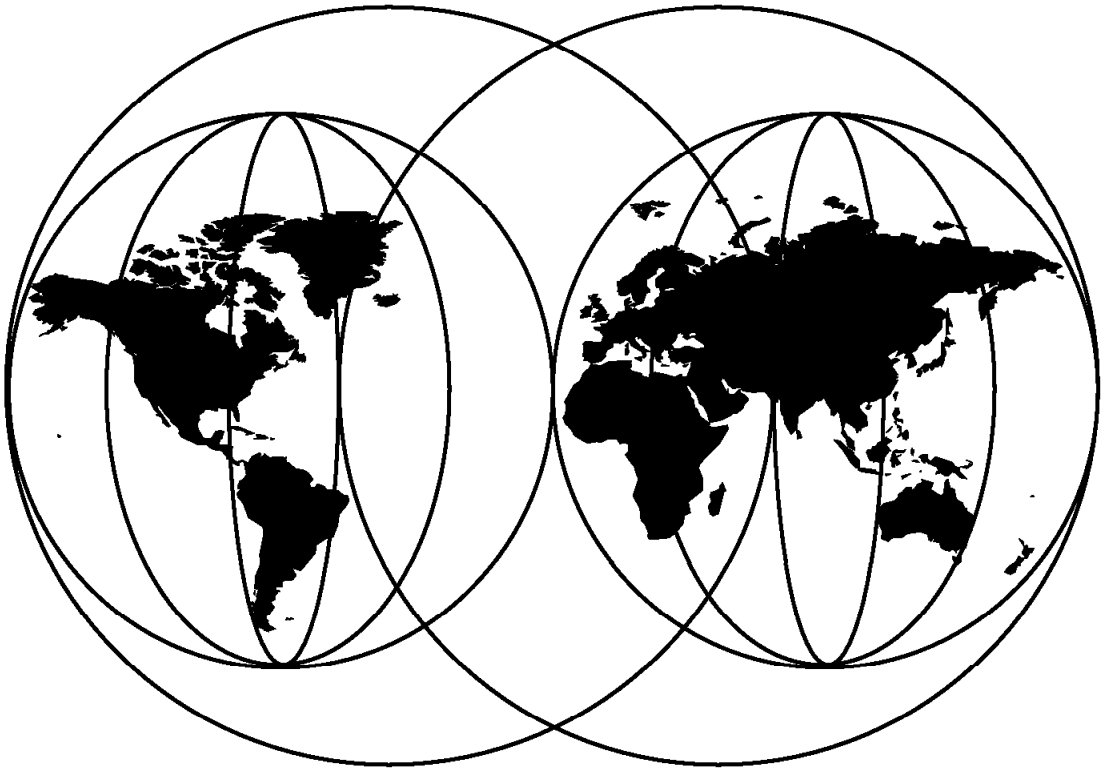
PSSP 2.4 Technical Presentation

Marcelo R. Barrios, Abbas Farazdel
Hajo Kitzhoefer, Yoshimichi Kosuge



International Technical Support Organization

http://www.redbooks.ibm.com

This book was printed at 240 dpi (dots per inch). The final production redbook with the RED cover will be printed at 1200 dpi and will provide superior graphics resolution. Please see "How to Get ITSO Redbooks" at the back of this book for ordering instructions.

SG24-5173-00

International Technical Support Organization

SG24-5173-00

PSSP 2.4 Technical Presentation

June 1998

Take Note!

Before using this information and the product it supports, be sure to read the general information in Appendix A, “Special Notices” on page 255.

First Edition (June 1998)

This edition applies to PSSP Version 2 Release 4 (5765-529) for use with the AIX Version 4 Operating System.

Comments may be addressed to:
IBM Corporation, International Technical Support Organization
Dept. HYJ Mail Station P099
522 South Road
Poughkeepsie, New York 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.

© **Copyright International Business Machines Corporation 1998. All rights reserved.**
Note to U.S. Government Users — Documentation related to restricted rights — Use, duplication or disclosure is subject to restrictions set forth in GSA ADP Schedule Contract with IBM Corp.

Contents

© Copyright IBM Corp. 1998

Tables

© Copyright IBM Corp. 1998

Preface

This redbook provides detailed information about the Parallel System Support Programs (PSSP) Version 2 Release 4, which was announced in April 1998.

The redbook covers all aspects of the announcement, including a revised version of the General Parallel File System for AIX (GPFS) and the Dependent Node (GRF) included in ″Technical Presentation for PSSP 2.3″, SG24-2080.

The contents also include detailed information about the new hardware being announced, such as the new frame (75 inches) and the new SMP nodes (thin and wide) based on the PowerPC 604e chip.

Finally, the book includes a section about Reliability, Availability, and Serviceability (RAS) improvements for the SP Switch. This section contains several examples of all the new facilities for switch problem diagnostics which are part of this new PSSP release.

This redbook is intended to help IBM Customers, Business Partners, IBM System Engineers, and other RS/6000 SP specialists who are involved in PSSP Version 2 Release 4 projects, including the education of professionals responsible for installing, configuring, and administering RS/6000 SP systems.

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 Poughkeepsie Center.

Marcelo R. Barrios is a project leader at the International Technical Support Organization, Poughkeepsie Center. He has been with IBM for five years working in different areas related to RS/6000. Currently he focuses on RS/6000 SP technology by writing redbooks and teaching IBM classes worldwide.

Abbas Farazdel is an Advisory International Technical Support Organization (ITSO) Specialist for the RS/6000 SP at the Poughkeepsie center.

Hajo Kitzhoefer is an Advisory International Technical Support Organization (ITSO) Specialist for RS/6000 SP at the Poughkeepsie center. Before joining the ITSO, he worked as an SP specialist at the RS/6000 and AIX Competence Center, IBM Germany. He has worked at IBM for eight years. His areas of expertise include RS/6000 SP, SMP, and Benchmarks. He now specializes in SP System Management, SP Performance Tuning and SP Hardware. He holds a Ph.D. degree in Electrical Engineering from the Ruhr-University of Bochum (RUB).

Yoshimichi Kosuge is an Advisory RS/6000 SP Project Leader at the International Technical Support Organization (ITSO), Poughkeepsie Center. He is responsible for RS/6000 SP System Management, HACMP ES, and Cluster Technology. Before joining the ITSO, he worked at IBM Japan as an Advisory I/T Specialist for RS/6000 and AIX.

Special thanks to Peter Kes for providing most of the presentation material included in this book.

© Copyright IBM Corp. 1998

Comments Welcome

Your comments are important to us!

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

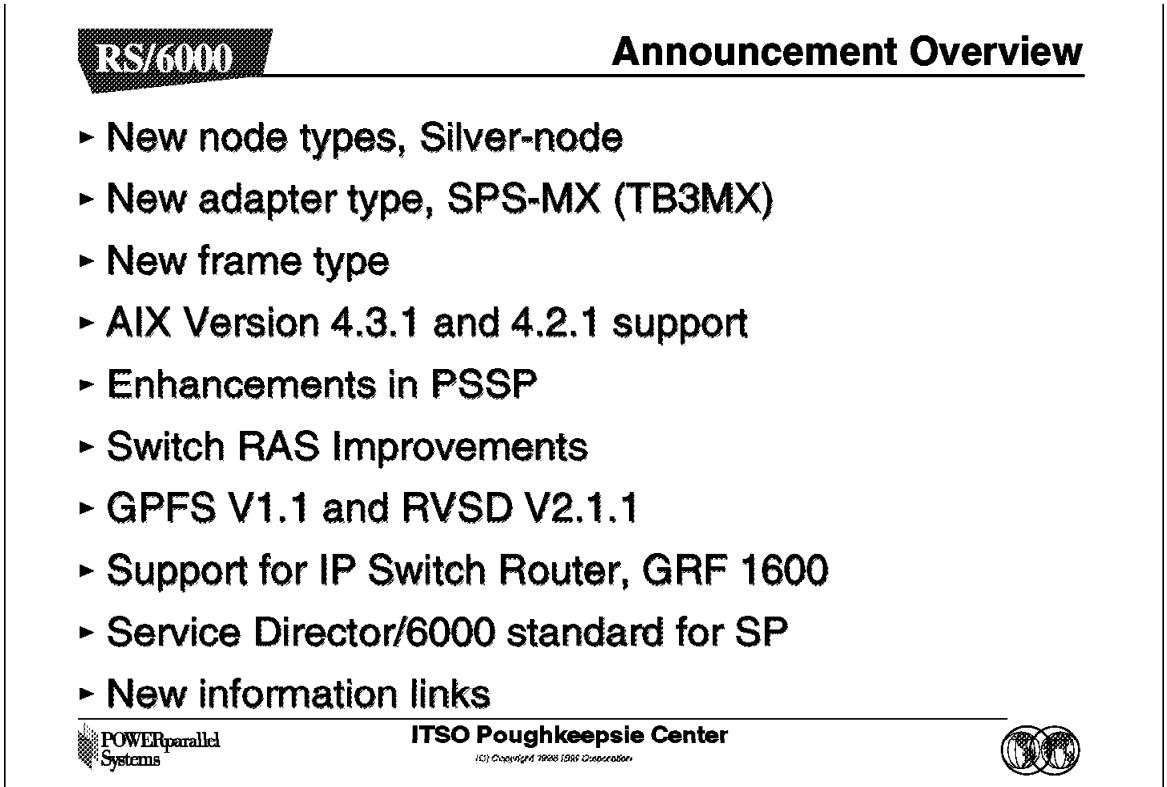
- Fax the evaluation form found in "ITSO Redbook Evaluation" on page 267 to the fax number shown on the form.
- Use the electronic evaluation form found on the Redbooks Web sites:

  For Internet users `http://www.redbooks.ibm.com/`
  For IBM Intranet users `http://w3.itso.ibm.com/`

- Send us a note at the following address:

  `redbook@us.ibm.com`

Chapter 1. Announcement Overview



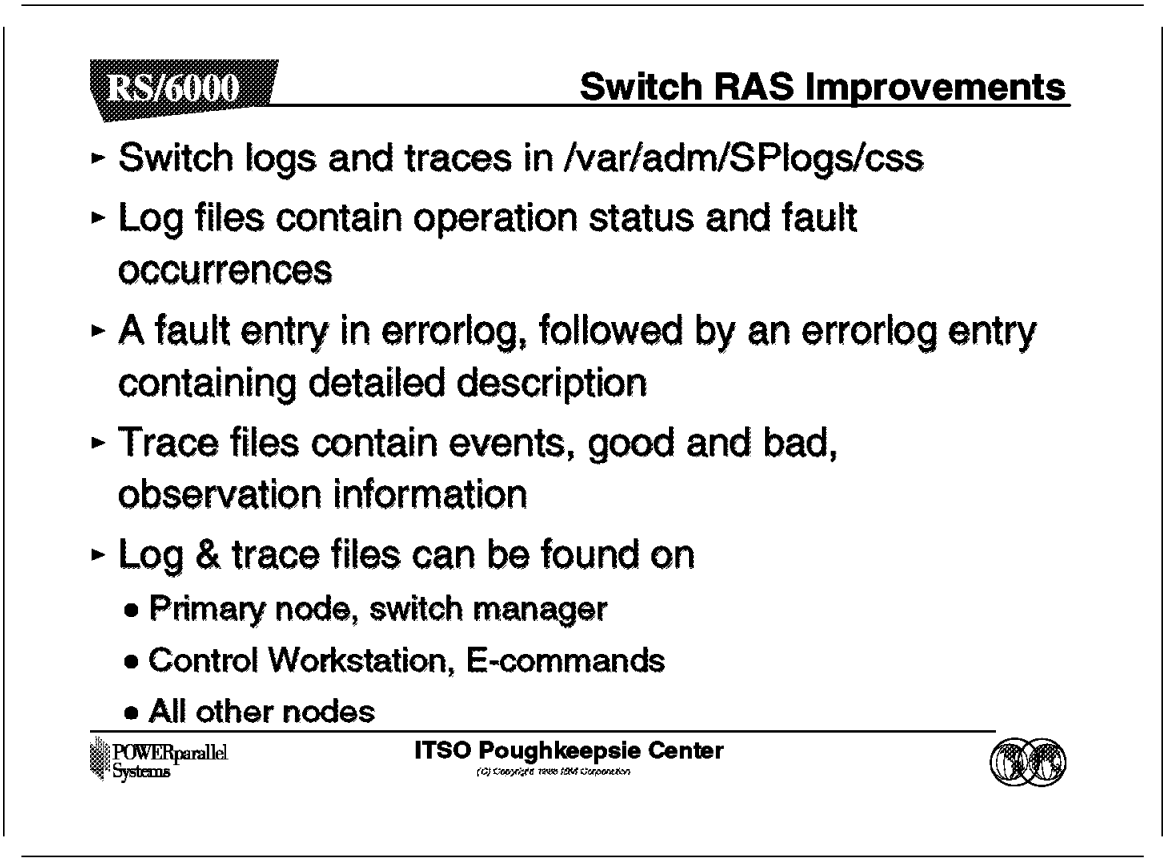
The Parallel System Support Programs (PSSP) for AIX Version 2.4 provides new RS/6000 Scalable POWERparallel (SP) Systems node enablement on AIX Version 4.2.1 or AIX 4.3.1 (in binary compatibility mode), and improved SP Switch diagnostics.

Enhancements include:

- 332MHz SMP wide and thin nodes that you can add to an existing SP system or use as the only node types in a new SP system.
- SP Switch MX Adapter for switch communication with increased SP Switch bandwidth.
- Migration and coexistence. You can upgrade to PSSP V2.4 running on AIX 4.2.1 or later, or on AIX 4.3.1 in binary compatibility mode. Earlier releases of PSSP can coexist with PSSP 2.4.
- Improved SP Switch diagnosis includes diagnostic documentation, messages, and more detailed entries in the AIX error log and switch logfiles for better problem isolation.
- Service Director(R) for RS/6000 is now standard with an SP.

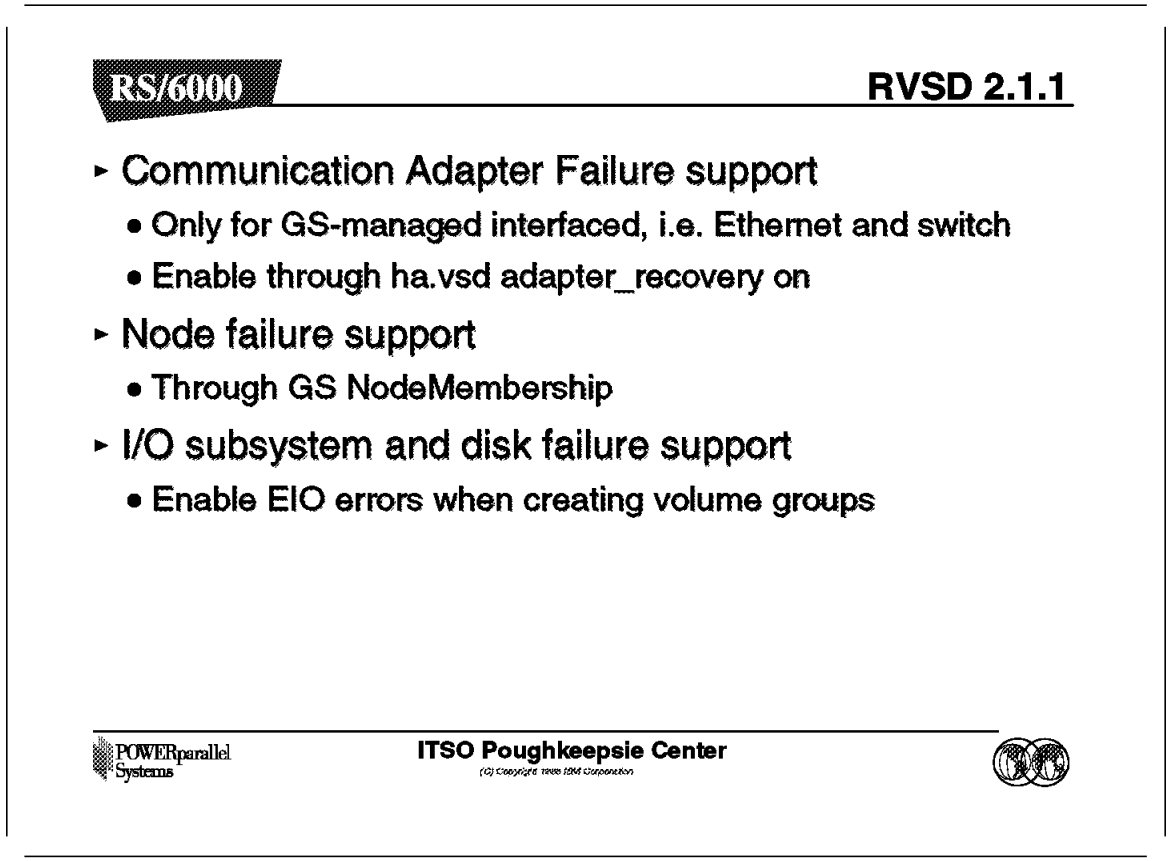
© Copyright IBM Corp. 1998

1.1 Switch RAS Improvements



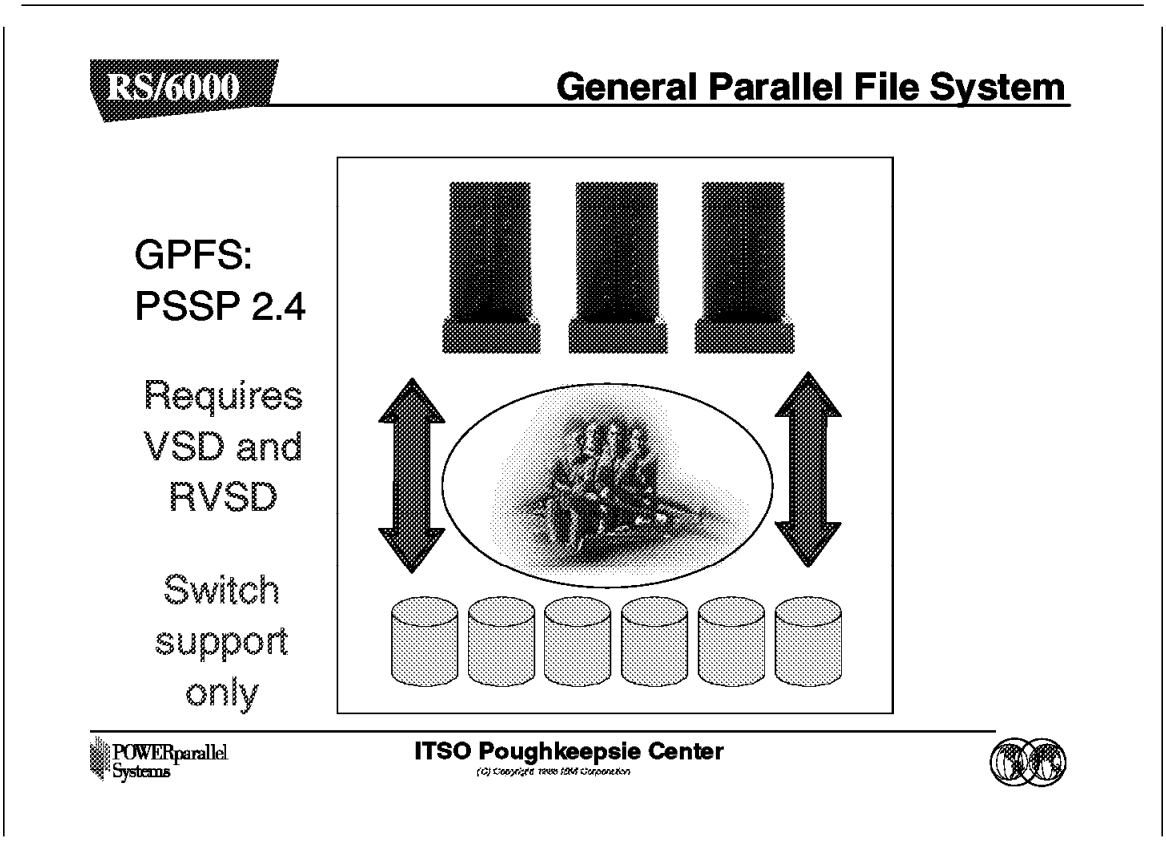
Better error logging and tracing facilities for the SP Switch are available with this new release of PSSP. Log and trace files are available on every node including the Control Workstation.

1.2 RVSD V2.1.1



This new modification level of the IBM Recoverable Virtual Shared Disk provides a higher level of availability for the IBM Virtual Shared Disk subsystem by providing automatic recovery in the case of communication adapter failure. Should there be failure in the communication adapter, the IBM Recoverable Virtual Shared Disk will cause failover to the backup node.

1.3 General Parallel File System (GPFS)



The availability of the General Parallel File System for AIX (GPFS) product is May 29, 1998.

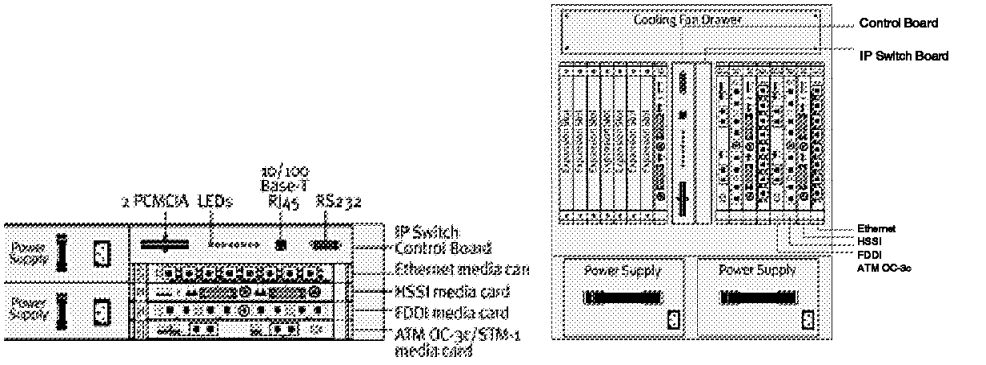
GPFS is a parallel file system that supports most of the interfaces in today's popular UNIX standards. GPFS works on a shared disk model. It allows access to files within an RS/6000 SP system from any node in the system. This provides high-performance file I/O to parallel jobs running on multiple RS/6000 SP nodes or to serial applications that are scheduled to, or executing on, RS/6000 SP nodes.

For details on this product, refer to Software Announcement 297-318 (RFA28501) dated August 26, 1997, and Software Announcement 297-457 (RFA29432) dated November 4, 1997.

1.4 GRF1600

RS/6000 **GRF1600**

- PSSP 2.4 supports the GRF1600 (16 ports model)
- Product number 9077



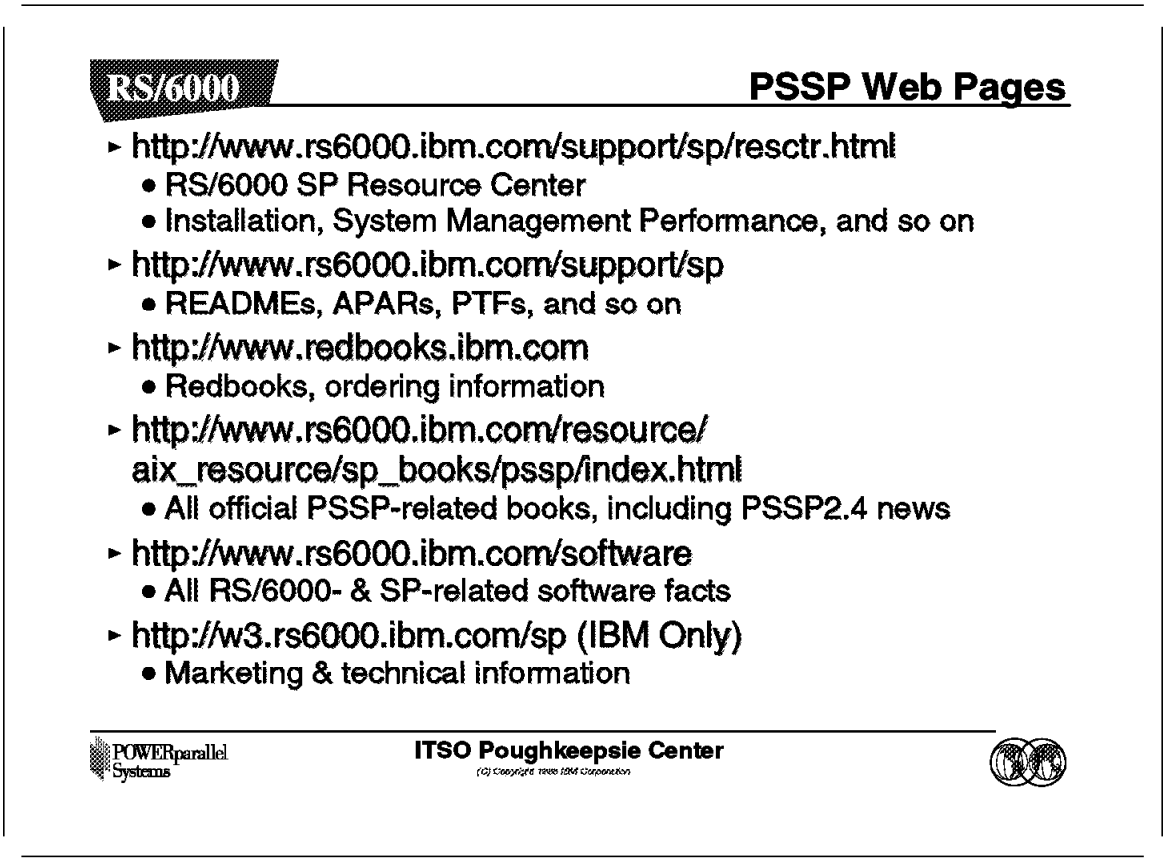
POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

The first router node (GRF400) was announced in 1997. GRF1600 is a new model of the Switch Router which has among other characteristics, 16 slots for media cards. Refer to Chapter 6, “Dependent Node” on page 149 for more details about GRF1600.

1.5 PSSP Web Pages



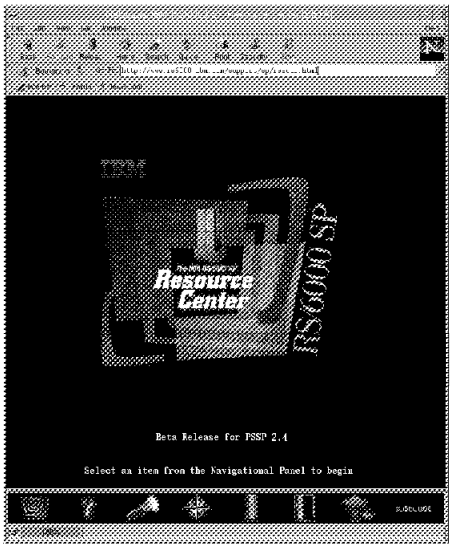
This is a list of URLs that contain useful information about the RS/6000 SP. All of them are external pages, with the exception of the last one, which is a page accessible only through the internal IBM network.

1.6 Resource Center

RS/6000

PSSP2.4: resctr.html

http://www.rs6000.ibm.com/support/sp/resctr.html



- Available at GA date
- Official RS/6000 SP Resource Center
- Provides links to existing Web documents
- Provides new SP-specific information and news
- Part of PSSP2.4 announcement
- Frames can be downloaded for better performance and as basis for own extensions

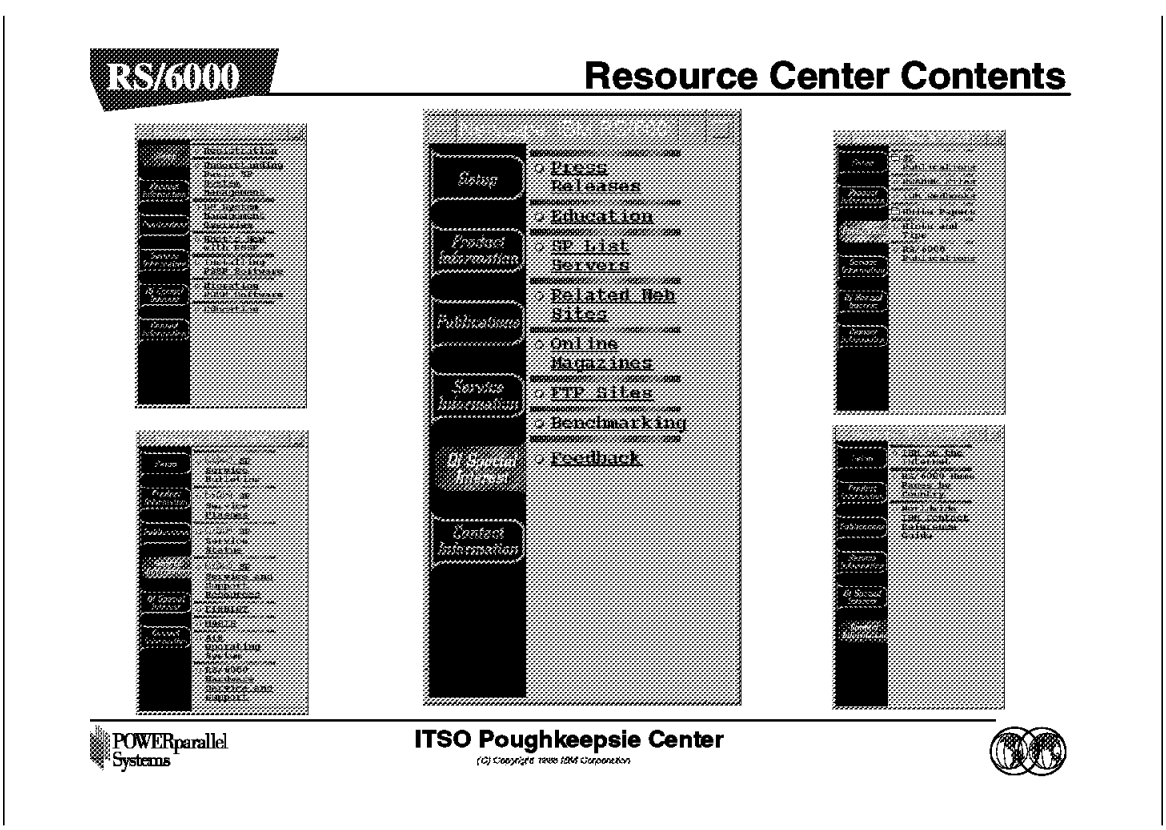
POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

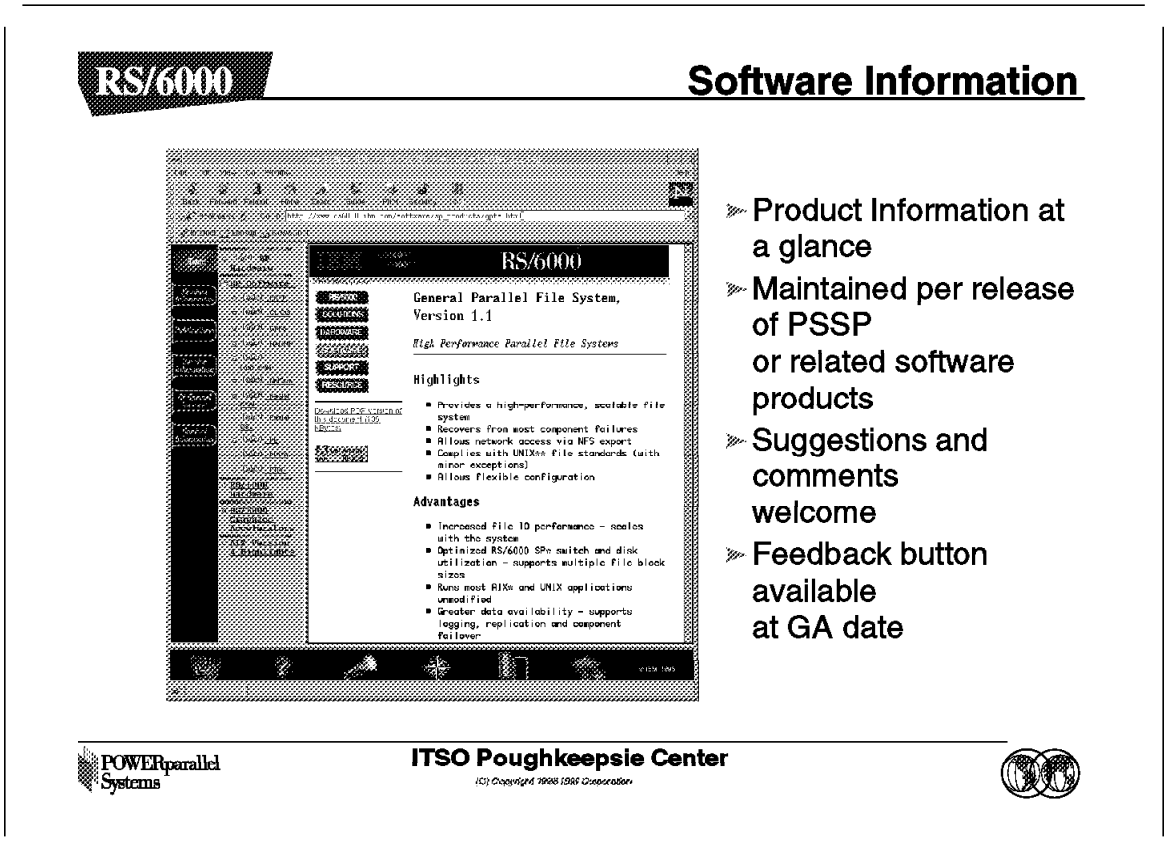
The Resource Center is part of the PSSP 2.4 announcements. It was developed to provide a one-stop shop for RS/6000 SP information. This site contains PSSP manuals, FAQs, and marketing information.

1.7 Resource Center Contents



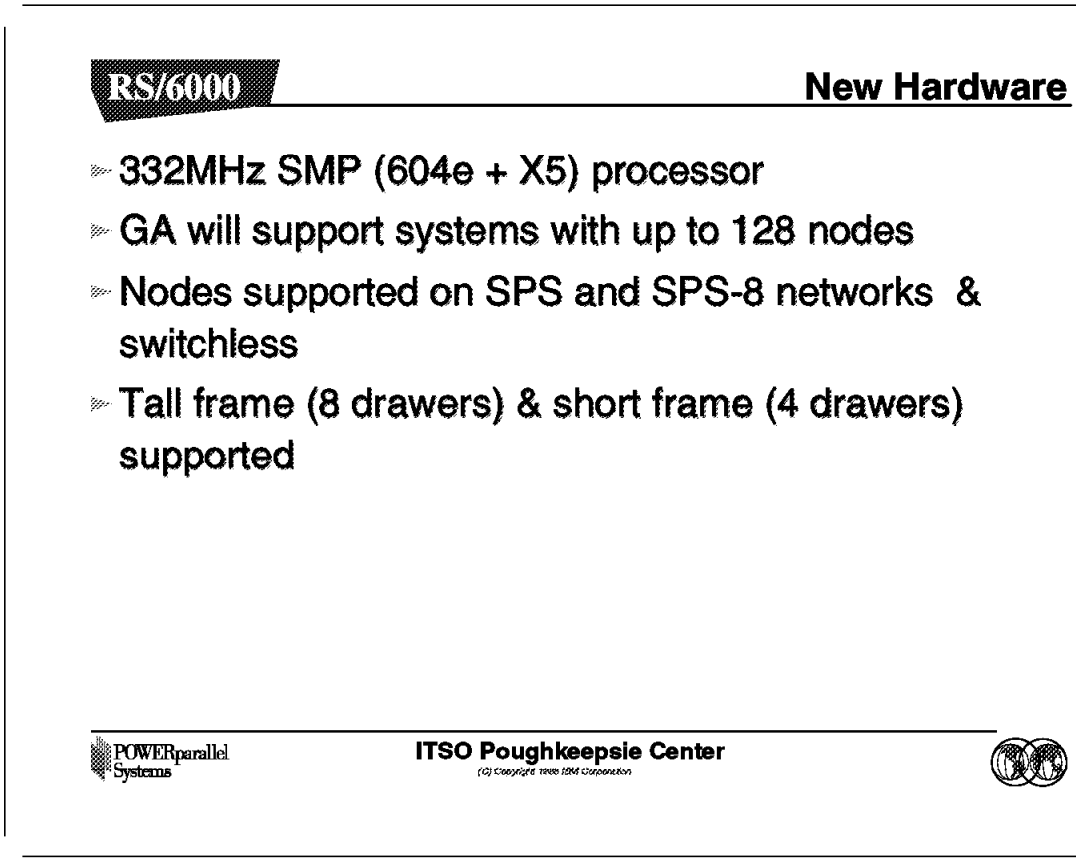
This graphic illustrates what can be found at the Resource Center.

1.8 Software Information



As this is a new tool, feedback about its contents and scope is quite welcome.

Chapter 2. New Hardware



The new hardware announcement includes the following devices:

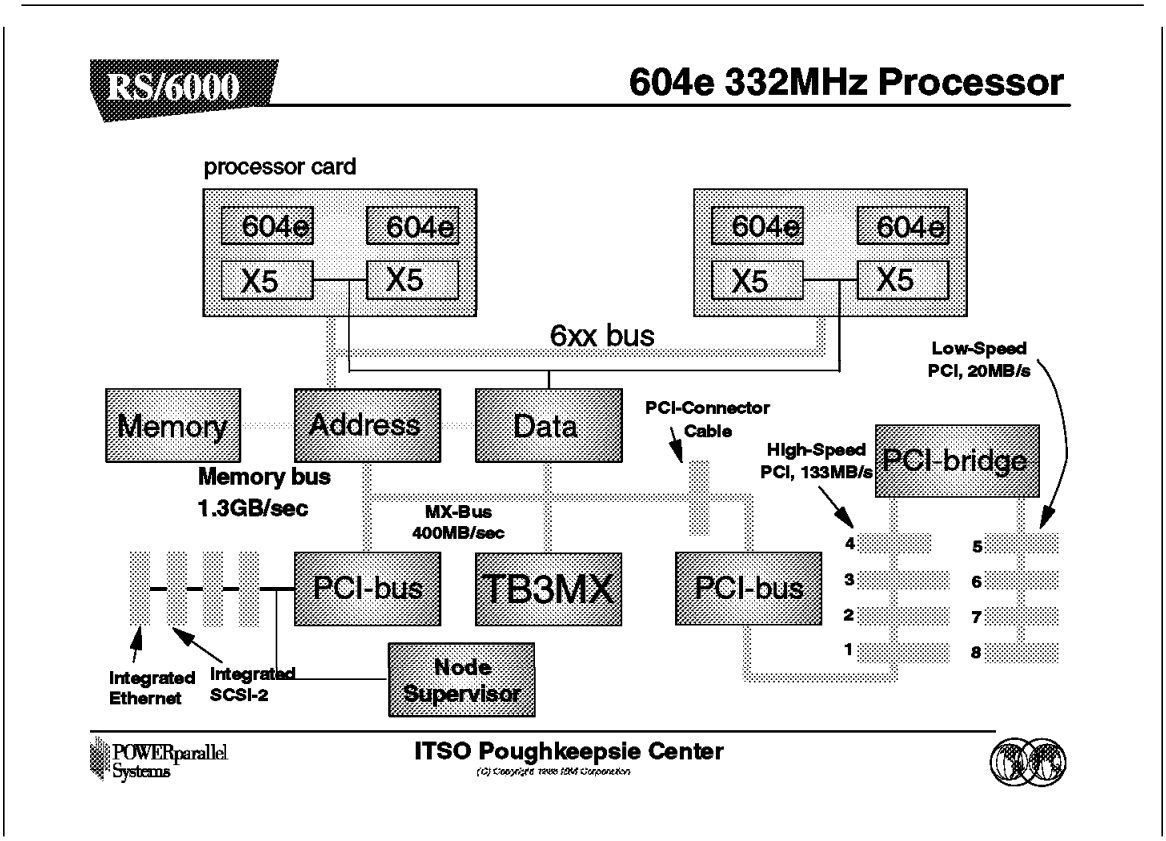
- 332MHz SMP thin node
- 332MHz SMP wide node
- Tall frame
- SPS-MX adapter

© Copyright IBM Corp. 1998

2.1 Nodes

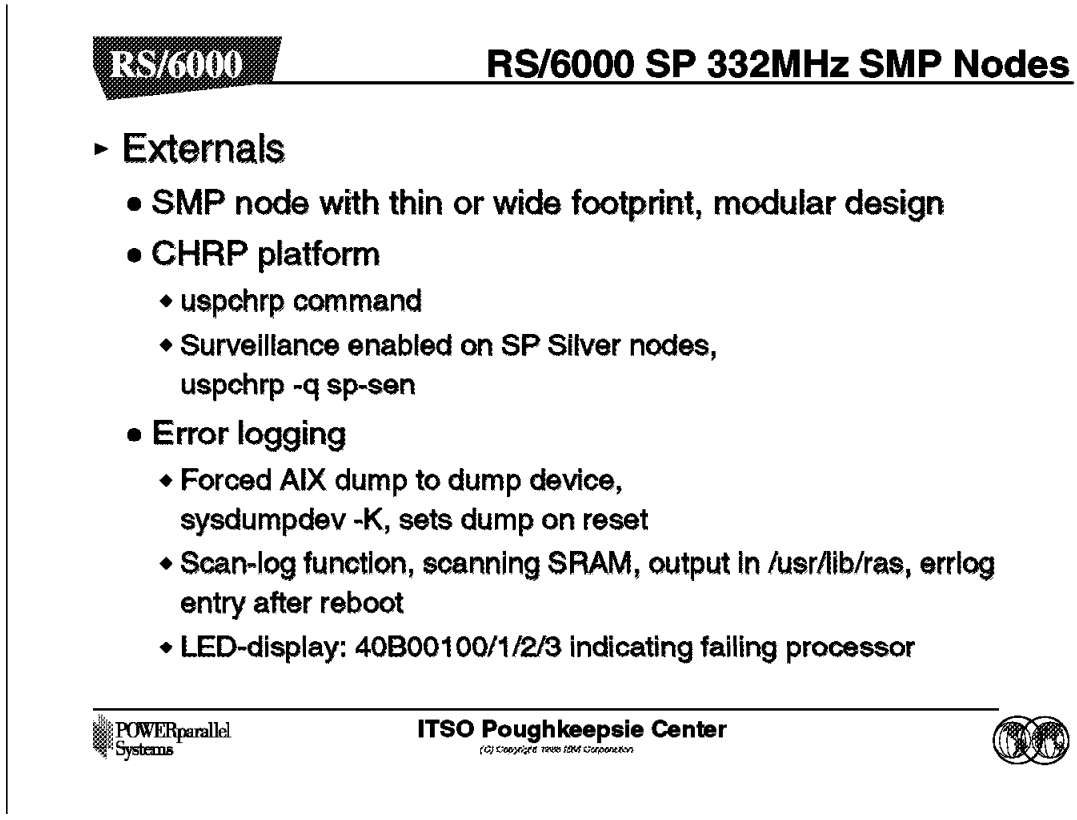
The new SMP thin and wide nodes are based on the PowerPC 604e processor. Some minor modifications have been made to the processor card to adjust the bus and processor speeds. The following sections describe in more detail the new SMP nodes.

2.1.1 604e 332MHz Processor



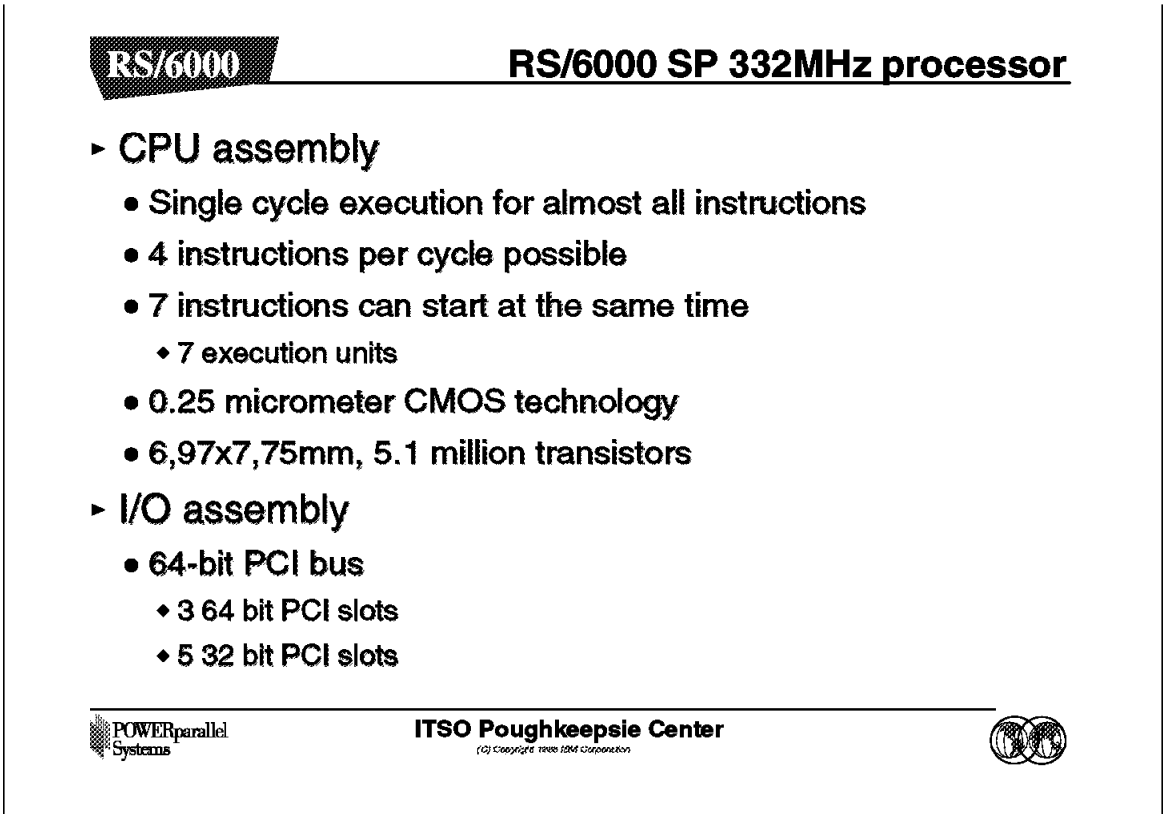
The RS/6000 SP 332MHz SMP nodes are functionally equivalent to an IBM RS/6000 7025-H50 workstation which uses a 6XX system bus, and a Peripheral Component Interconnect (PCI) bus architecture. TB3MX adapter is a variation on the TB3 adapter that plugs into the new Mezzanine (MX) bus.

2.1.2 RS/6000 SP 332MHz SMP Nodes



The RS/6000 SP 332MHz SMP nodes conform to the Common Hardware Reference Platform (CHRP), with some extensions due to the TB3MX adapter.

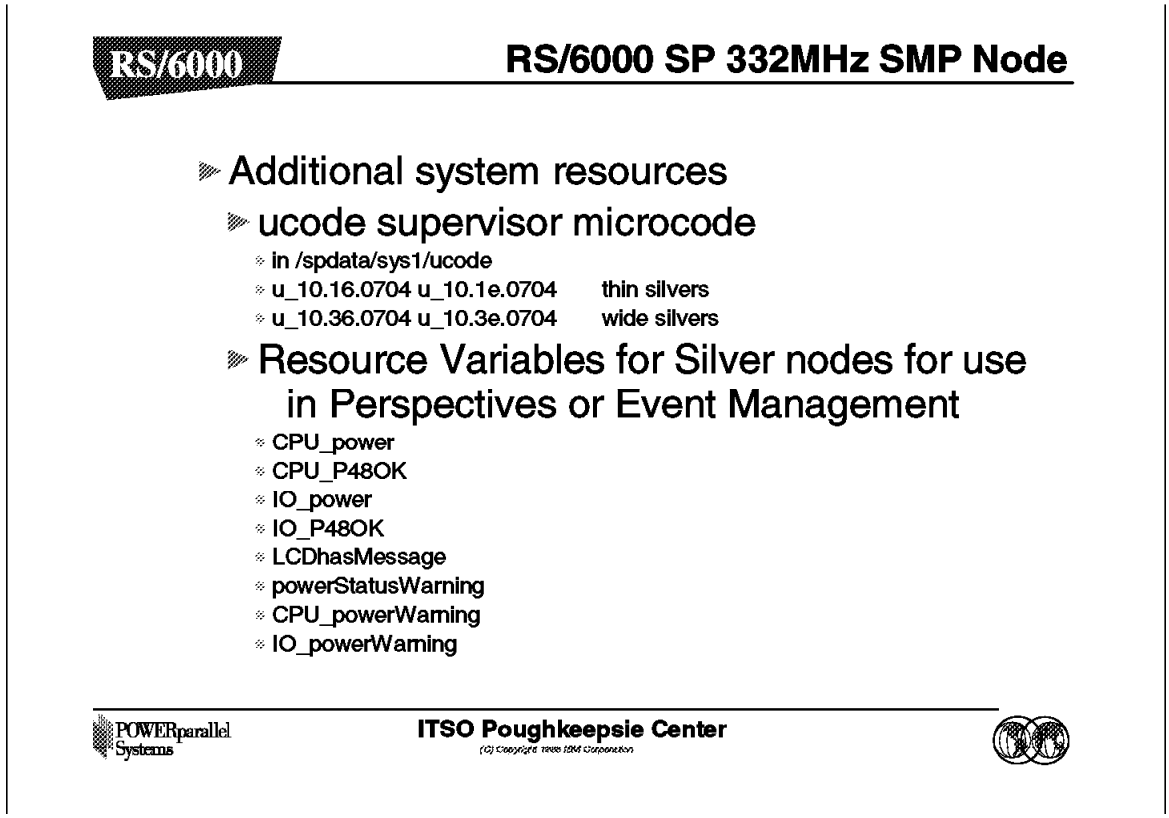
2.1.3 RS/6000 SP 332MHz Processor



The RS/6000 SP 332MHz SMP nodes are comprised of modular components. Both SMP thin nodes and SMP wide nodes have a CPU assembly which contains a node supervisor, two to four CPUs, one TB3MX slot and two PCI slots.

A wide node also has an I/O assembly, which adds eight PCI slots.

2.1.4 RS/6000 SP 332MHz SMP Node



Supervisor microcode for the RS/6000 SP 332MHz SMP nodes is available in the /spdata/sys1/ucode directory. The files u_10.16.0704 and u_10.1e.0704 are for thin nodes, while u_10.36.0704 and u_10.3e.0704 are for wide nodes.

Resource variables for the RS/6000 SP 332MHz nodes are available for Perspectives and Event Management Subsystem, as follows:

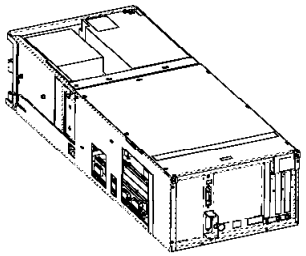
CPU_power	Indicates that the CPU power supply is turned on.
CPU_P48OK	+48 volt input to the CPU is okay.
IO_power	Indicates whether the Sidecar power supply is turned on.
IO_P48OK	+48 volt input to the Sidecar is okay.
LCDhasMessage	The nodes LED and/or LCD contains a message.
powerStatusWarning	The CPU or Sidecar power supply status lights are not consistent.
CPU_powerWarning	Indicates that the node supervisor attempted to power on the node but the CPU power did not turn on.
IO_powerWarning	Indicates that the node supervisor attempted to power on the node but the Sidecar power did not turn on.

2.1.5 RS/6000 SP 332MHz Thin Node

RS/6000

RS/6000 SP 332MHz Thin Node

- Thin Node
  - 2 to 4 CPUs (within 2 CPU slots)
  - 256MB to 3GB memory (within 2 memory card slots)
  - Integrated Ethernet (10Base2 and 10BaseT)
  - Integrated SCSI-2 and serial port
  - One optional SPS-MX switch adapter per node
  - 2 PCI slots
  - 2 DASD bays (4.5 to 18.2GB)
  - Two thin nodes per drawer



POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright IBM Corporation

Thin nodes are now available in both a symmetric multiprocessing (SMP) configuration and in the previous uniprocessor configuration. All RS/6000 SP 332MHz SMP thin nodes have Peripheral Component Interconnect (PCI) architecture.

SMP thin nodes contain either two or four 332MHz PowerPC processors per node. SMP thin nodes are functionally equivalent to an IBM RS/6000 7025-H50 workstation which uses PCI bus architecture. Your IBM RS/6000 SP system must be operating at the PSSP 2.4 level to use SMP thin nodes. Although a power system upgrade is necessary before you can incorporate SMP thin nodes into your SP system, these nodes are fully compatible with all existing SP hardware except for the HiPS. SMP thin nodes are not compatible with any of HiPS.

Note: The SMP thin node is half the width of a wide node, therefore sixteen SMP thin nodes can be housed in a tall frame. SMP thin nodes are ordered singly but must be installed in pairs. For electromagnetic compliance, SMP thin nodes are housed in an SMP Enclosure. The SMP thin node has two PCI adapter slots.

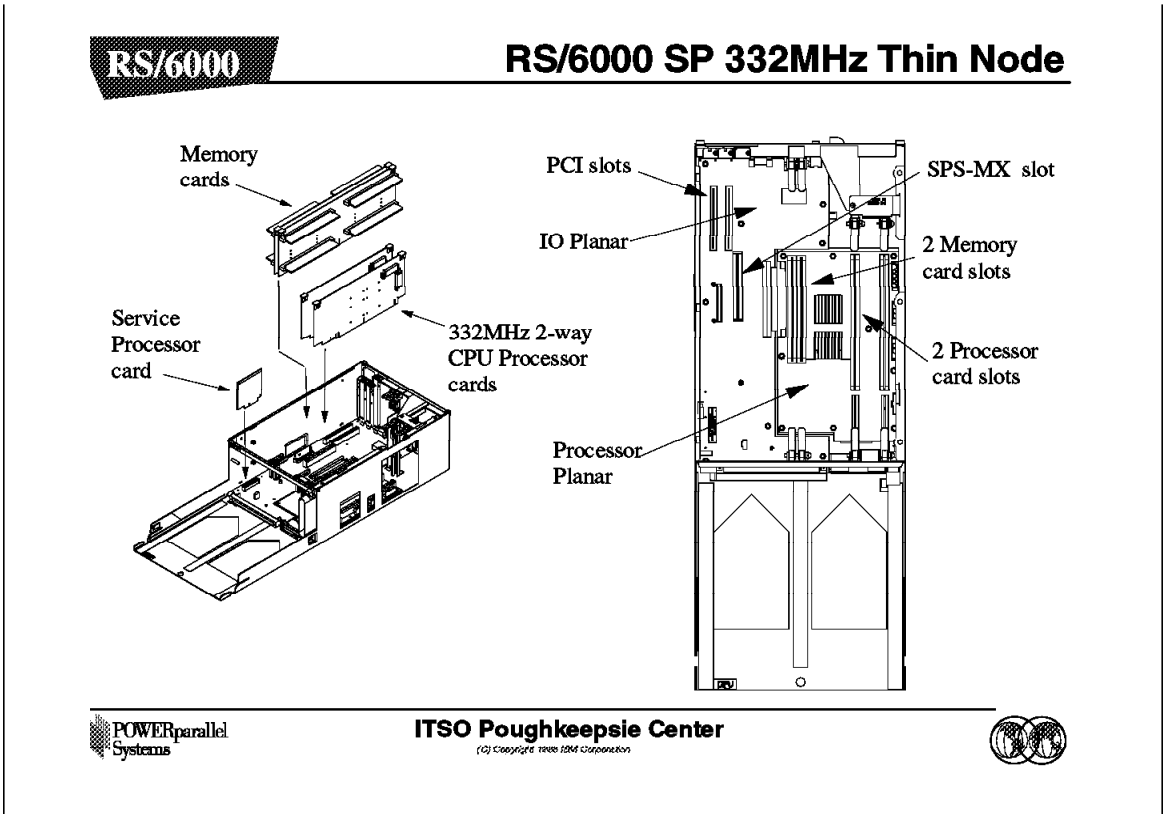
The switch adapter for the SMP thin nodes does not occupy a PCI slot. Instead, the switch adapter for these nodes is installed into the Mezzanine (MX) bus. The MX bus connects the I/O planar with the system planar; locating the switch adapter in the MX bus enables the switch traffic to proceed at higher bandwidths and lower latencies.

SMP thin nodes have two memory cards and require a minimum of 256MB of memory. These nodes will support a maximum of 3GB of memory. Memory is supplied by 128MB DIMMs that must be mounted in pairs (256 MB increments). The memory cards are not required to be configured symmetrically. Each card has the capacity to mount 2GB of DIMMs; however, only 3GB are addressable per node. Memory cards and DIMMs are not interchangeable between SMP and non-SMP thin nodes.

SMP thin nodes can have up to two internal DASD. The SMP thin node requires a minimum of 4.5GB of DASD and has a maximum of 18.2GB of internal disk storage. The SMP thin node also contains an integrated SCSI-2 network for internal DASD installation.

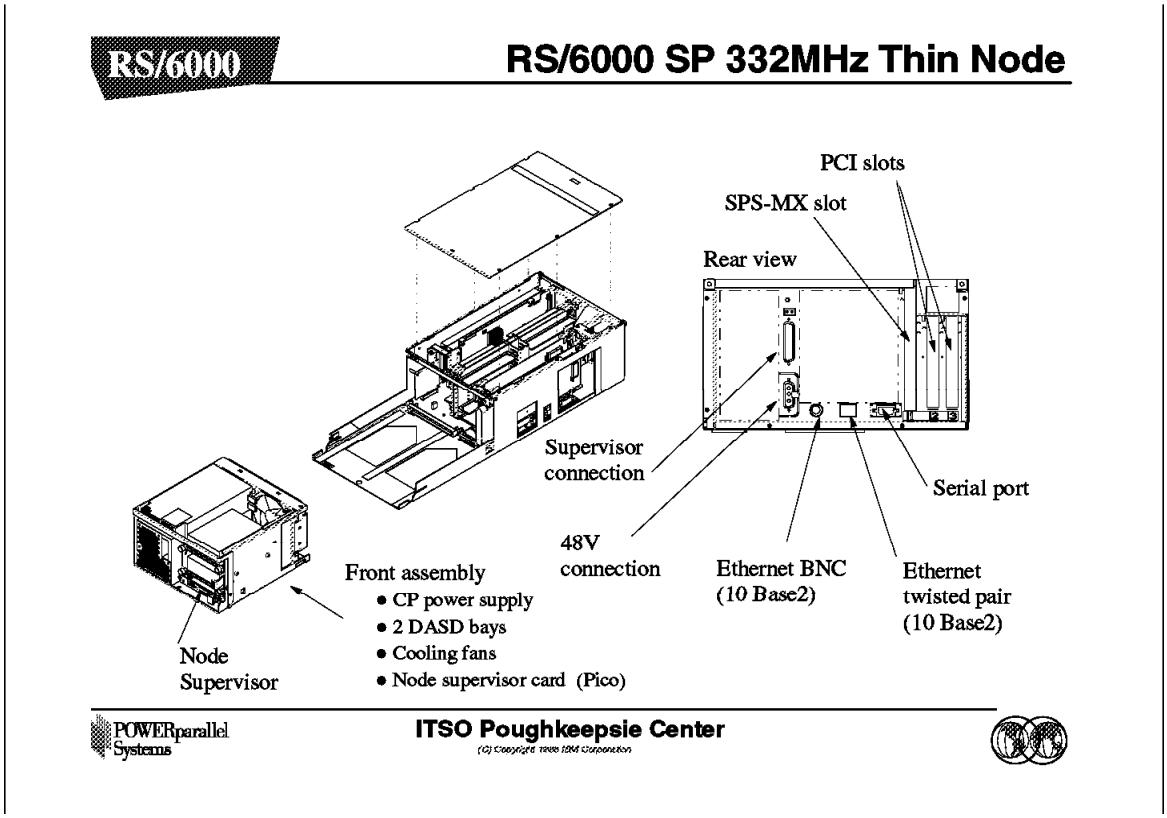
A 10BaseT/10Base2 Ethernet adapter for the SP Ethernet is integrated into the SMP thin node and does not use a PCI slot.

2.1.6 RS/6000 SP 332MHz Thin Node Top View



The SMP thin node is composed with CPU assembly only. It houses both the processor planar and I/O planar. The processor planar mounts the processor and memory, while the I/O planar mounts the SPS-MX slot and PCI slots.

2.1.7 RS/6000 SP 332MHz Thin Node Rear View



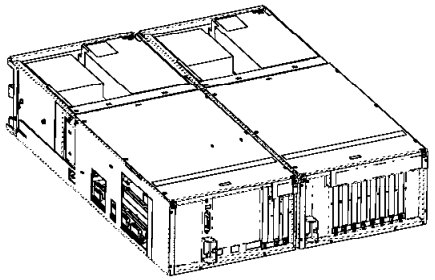
The CPU assembly includes the front assembly, which contains the power supply for the CPU assembly, two DASD bays, and the node supervisor card.

2.1.8 RS/6000 SP 332MHz Wide Node

RS/6000

RS/6000 SP 332MHz Wide Node

- Wide Node
  - 2 to 4 CPUs (within 2 CPU slots)
  - 256MB to 3GB memory (within 2 memory card slots)
  - Integrated Ethernet (10Base2 and 10BaseT)
  - Integrated SCSI-2 and serial port
  - One optional SPS-MX switch adapter per node
  - 10 PCI slots (2+8)
  - 4 DASD bays (4.5 to 36.4GB)
  - One wide node per drawer



POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright IBM Corporation

Wide nodes are now available in both a Symmetric Multiprocessing (SMP) configuration and in the previous uniprocessor configuration. All the RS/6000 SP 332MHz SMP wide nodes have Peripheral Component Interconnect (PCI) bus architecture.

SMP wide nodes contain either two or four 332MHz PowerPC processors per node. SMP wide nodes are functionally equivalent to an IBM RS/6000 7025-H50 workstation which uses PCI bus architecture. Your IBM RS/6000 SP system must be operating at the PSSP 2.4 level to use the SMP wide nodes. Although a power system upgrade is necessary before you can incorporate SMP wide nodes into your SP system, once installed, these nodes are fully compatible with all existing SP hardware except for the HiPS. SMP wide nodes are not compatible with any of the HiPS.

The SMP wide node occupies one full drawer in a frame, therefore up to eight SMP wide nodes can be housed in a tall frame.

Note: If the SMP wide node is the first node in a frame, you *must* install another node drawer in that frame. The second node drawer can be any type of high or wide node, or any type of thin node pair. For electromagnetic compliance, SMP wide nodes are housed in an SMP enclosure.

The SMP wide node has ten PCI slots. The SMP wide node PCI bus is divided into three logical groups of PCI slots, as follows:

- The first slot group (slots I2 and I3) is composed of the two 32-bit slots residing on the *CPU side* of the SMP wide node.

  Note: The *CPU-side* has five slots (I1 through I5). But only two are available for external adapters. The other three are used internally as follows:

  - I1 is used by the Node Supervisor card
  - I4 is used by the integrated SCSI adapter
  - I5 is used by the integrated Ethernet adapter
- The second group resides on the *I/O side* of the node. The second group has four PCI slots (slots I1 through I4), with three 64-bit slots and a single 32-bit slot.
- The third group also resides on the *I/O side* of the node, and it also has four PCI slots (slot I5 through I8), which make up the last four 32-bit slots.

The I1 slot on the CPU side of the node is reserved for the optional SP switch MX adapter. The ten PCI slots in the SMP wide node can be used for any valid RS/6000 SP PCI system adapter. While most PCI adapters will function in any SMP wide node slot, the SSA RAID5 adapter cannot be placed in any of the third group of PCI slots because of performance limitations.

The switch adapter for the SMP wide nodes does not occupy a PCI slot. Instead, the switch adapter for these nodes is installed into the Mezzanine (MX) bus. The MX bus connects the I/O planar with the system planar; locating the switch adapter in the MX bus enables the switch traffic to proceed at higher bandwidths and lower latencies.

SMP wide nodes have two memory cards and require a minimum of 256MB of memory. These nodes will support a maximum of 3GB of memory. Memory is supplied by 128MB DIMMs that must be mounted in pairs (256MB increments). The memory cards are not required to be configured symmetrically. Each card has the capacity to mount 2GB of DIMMs; however, only 3GB are addressable per node. Memory cards and DIMMs are not interchangeable between SMP and non-SMP wide nodes.

SMP wide nodes can have up to four internal DASD. The SMP wide node requires a minimum of 4.5GB of DASD and has a maximum of 36.4GB of internal disk storage. The SMP wide node also contains an integrated SCSI-2 network for internal DASD installation.

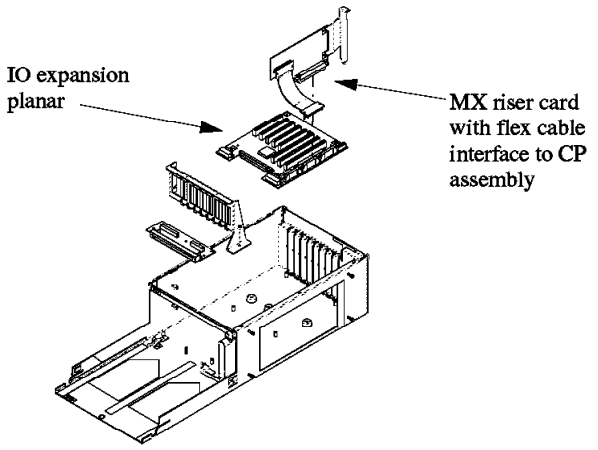
A 10BaseT/10Base2 Ethernet adapter for the SP Ethernet is integrated into the SMP wide node and does not use an external PCI slot.

2.1.9 RS/6000 SP 332MHz Wide Node I/O Assembly

RS/6000

RS/6000 SP 332MHz Wide Node

- I/O Assembly



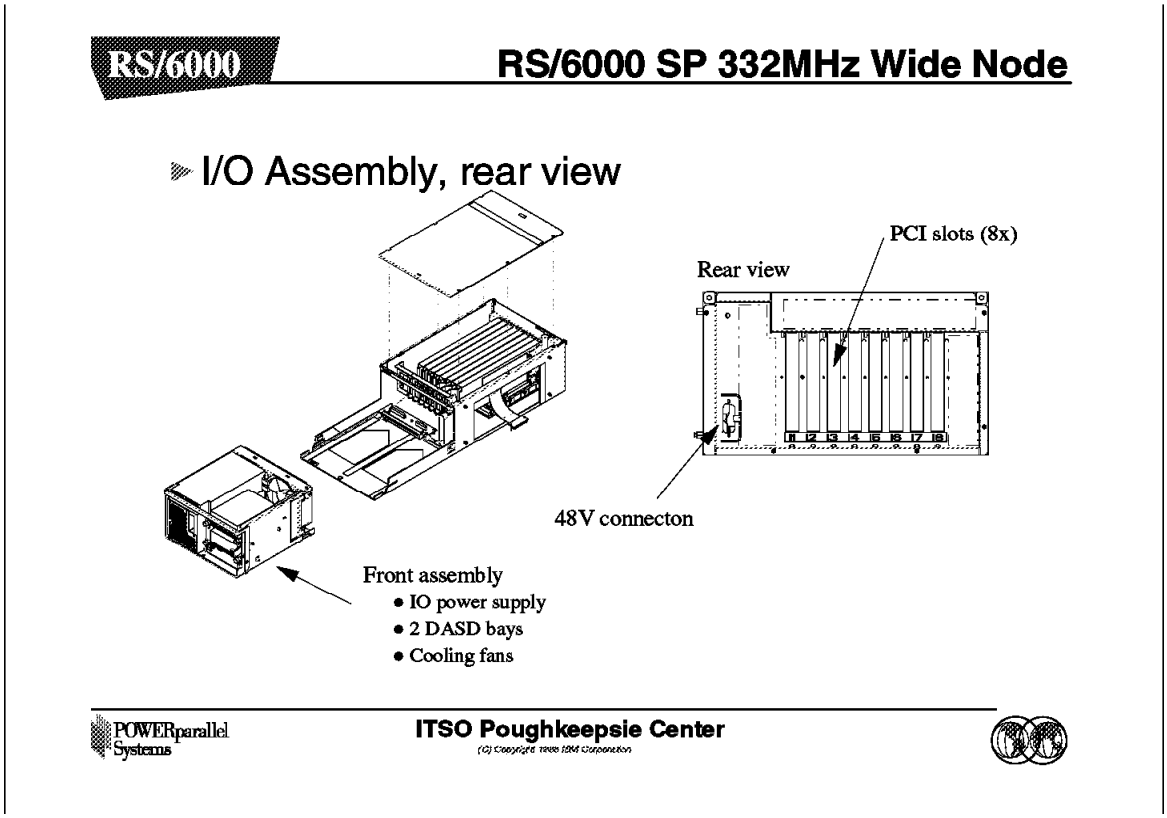
POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright IBM Corporation

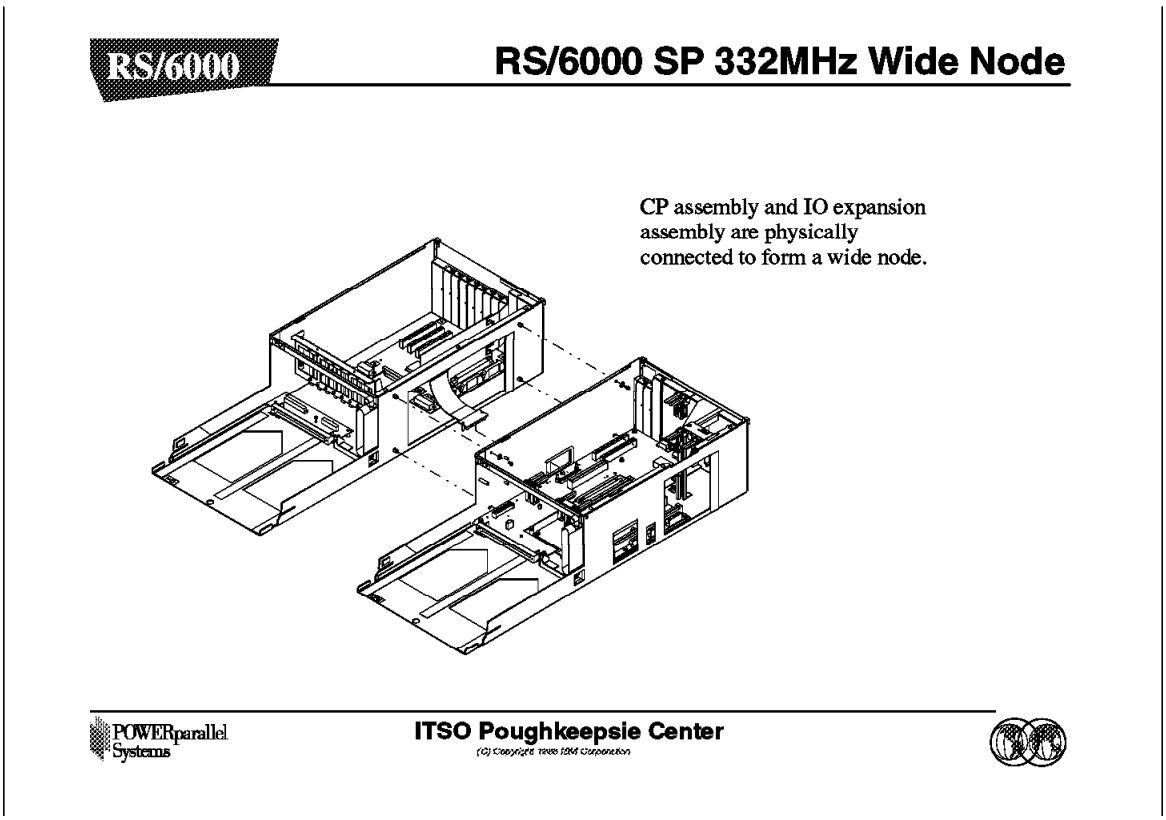
The SMP wide node is composed of a CPU assembly and an I/O assembly. The I/O assembly has an MX riser card, with a flex cable interface to be connected to the CPU assembly.

2.1.10 RS/6000 SP 332MHz Wide Node Rear View



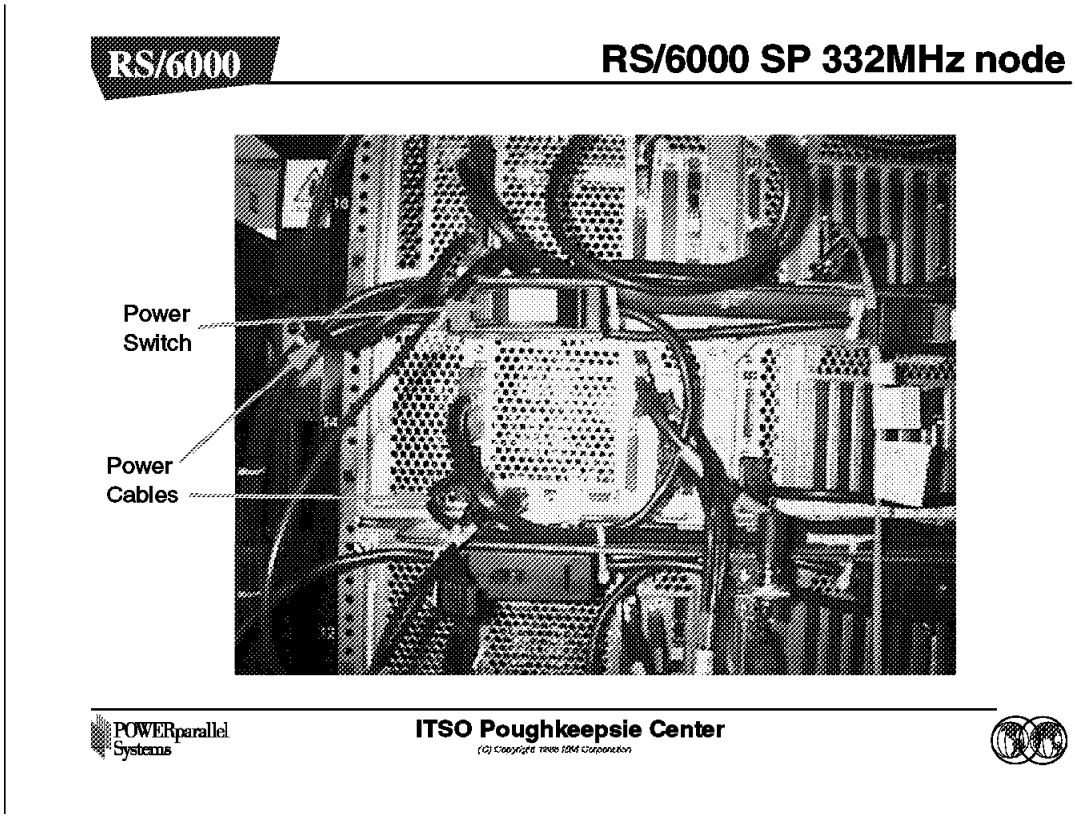
The I/O assembly includes the front assembly, which contains the power supply for the I/O assembly and two DASD bays.

2.1.11 332MHz Wide Node CPU and I/O Assembly



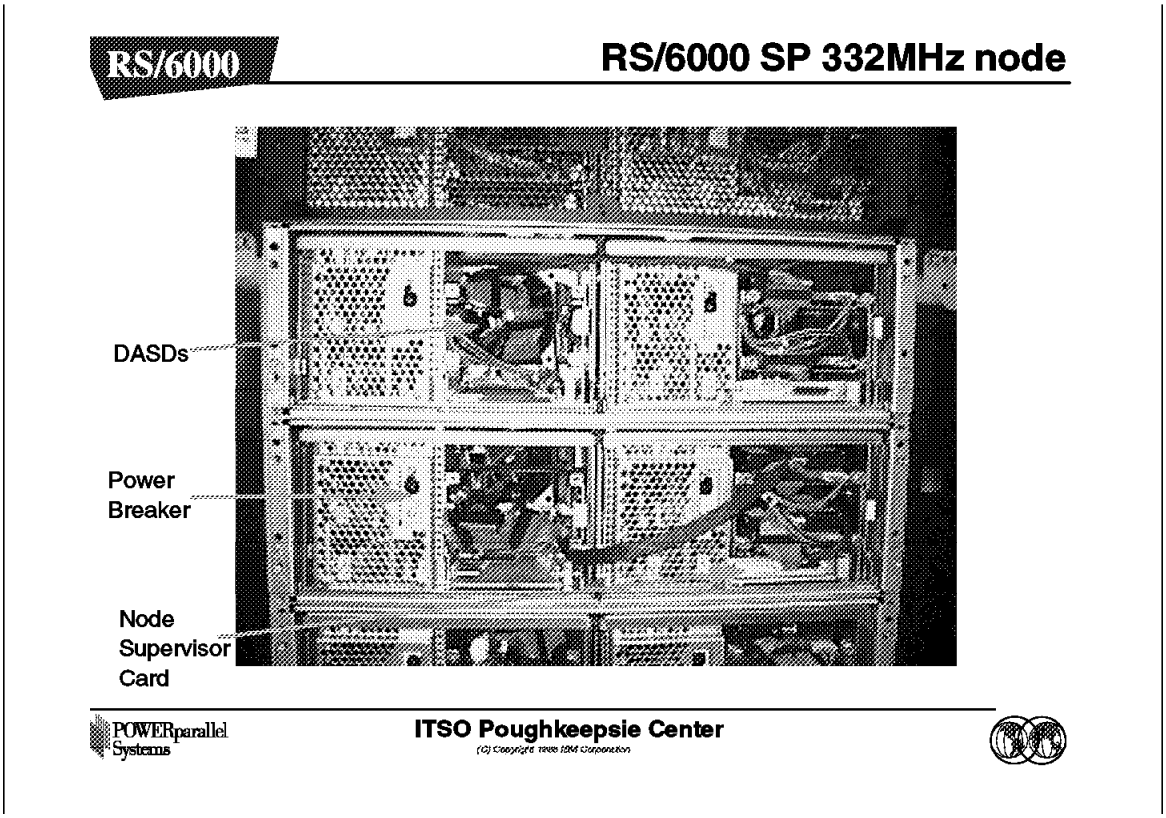
The CPU assembly and the I/O assembly are physically connected to form a wide node.

2.1.12 RS/6000 SP 332MHz Node Rear View



This rear view photograph of an RS/6000 SP 332MHz wide node shows the power cable with switch, the supervisor connection, the 48V power connection, the SP Ethernet connection, and the SPS-MX connection.

2.1.13 RS/6000 SP 332MHz Node Front View

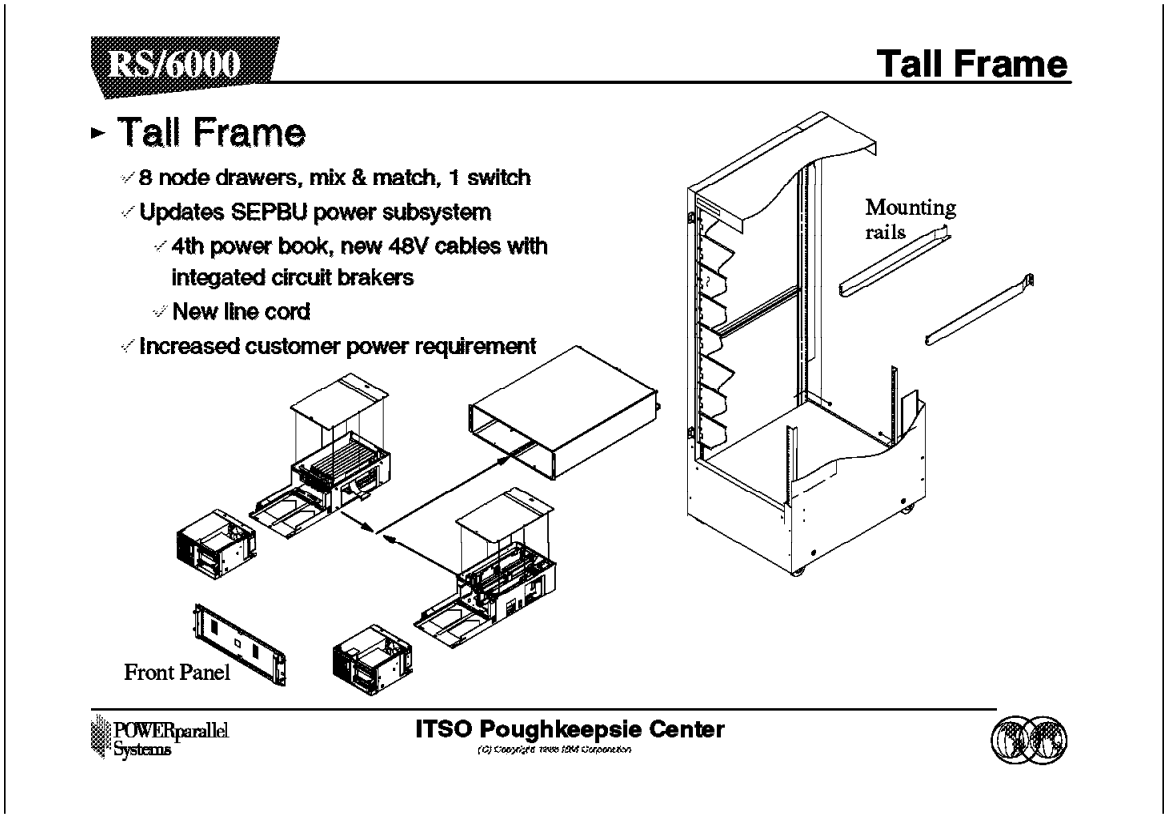


In this front view photograph of RS/6000 SP 332MHz wide node, you can see the DASDs, the node supervisor card, and the power breakers.

2.2 Frames

A new type of frame has been introduced in this announcement. The short frame remains unchanged. The following sections describe this in more detail.

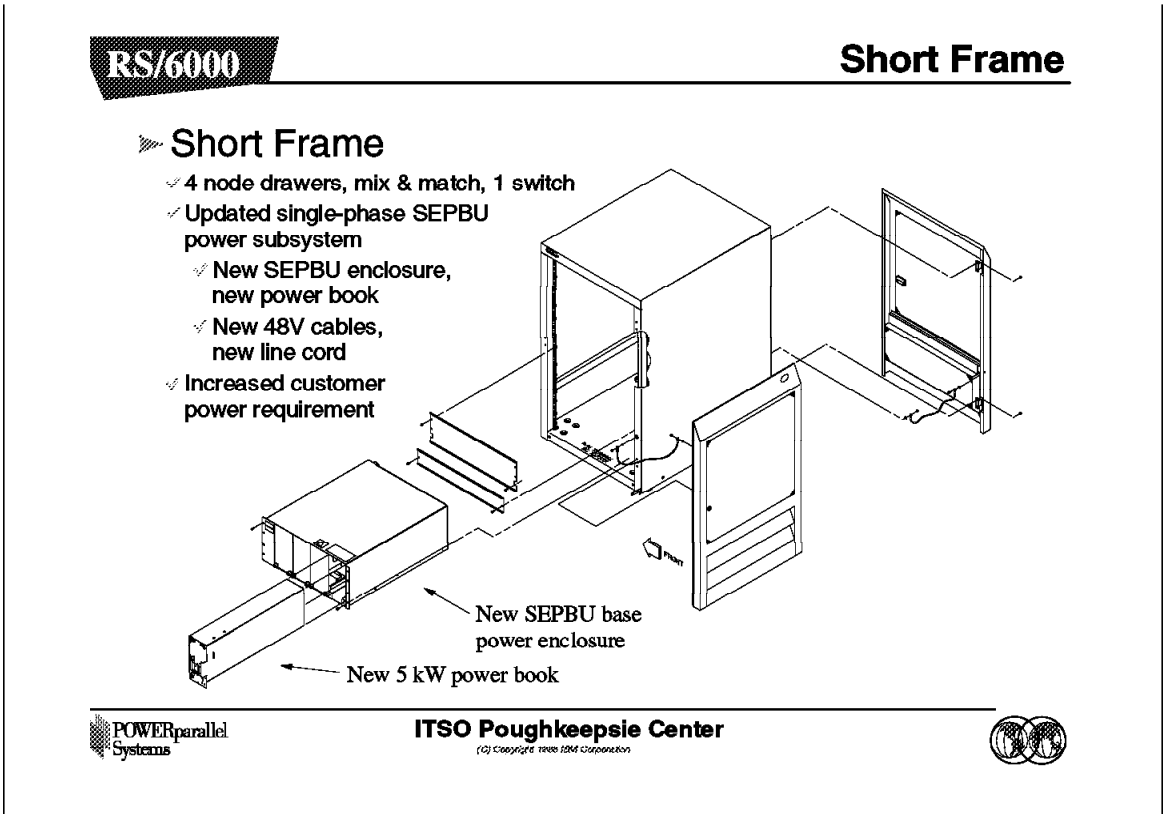
2.2.1 Tall Frame



The most noticeable difference between the new and old tall frames is a reduction in height. The new tall frames are now 1.93 meters high, while the older frames are 2.01 meters high.

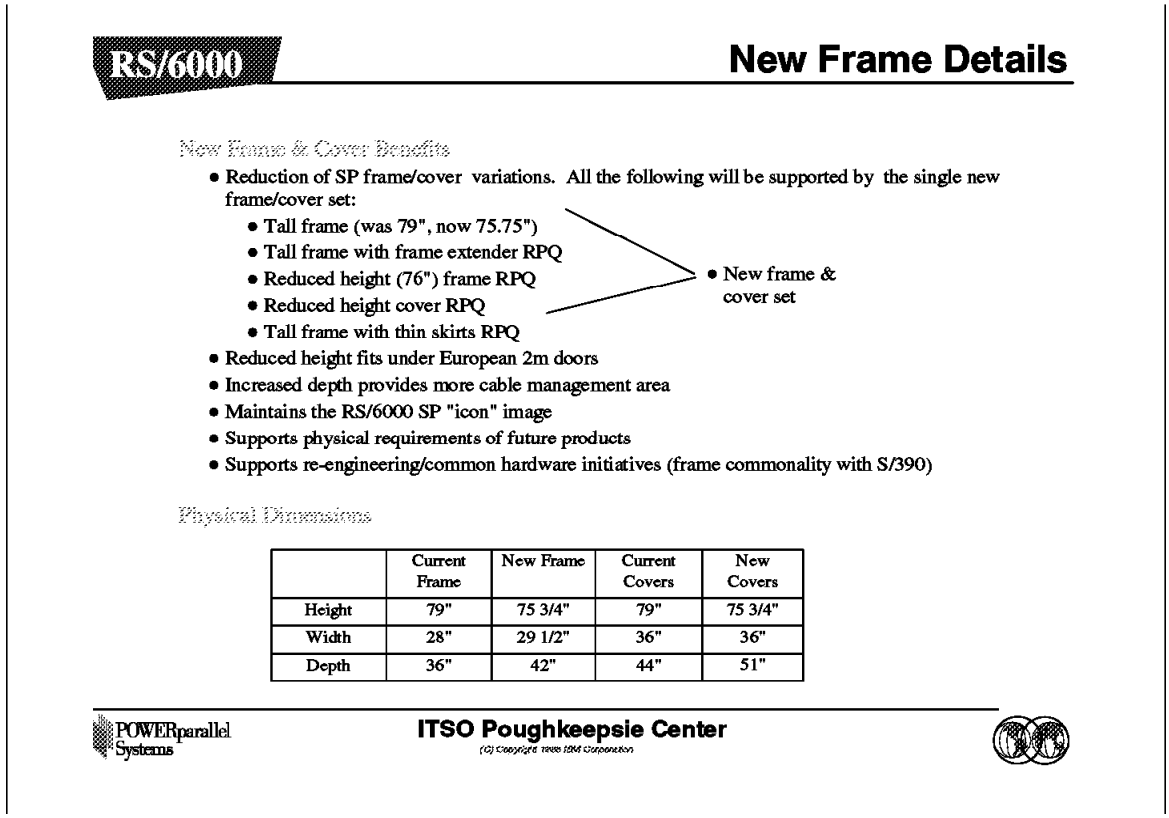
All new frame designs are completely compatible with all valid SP configurations using older equipment. Also, all new nodes can be installed in any existing SP frame, provided that the required power supply upgrades have been implemented in that frame.

2.2.2 Short Frame



Short frames remain 1.25 meters high. They contain an updated single-phase SEPBU power subsystem.

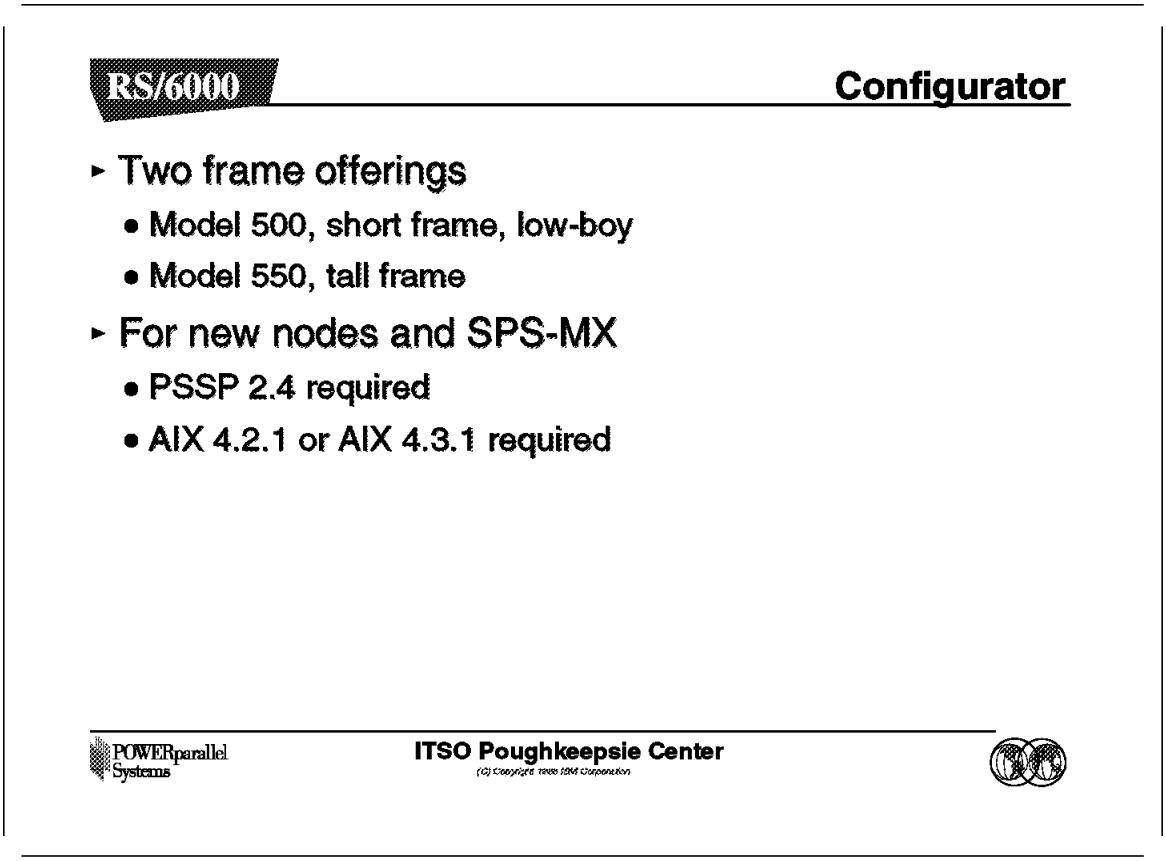
2.2.3 New Frame Details



The previous frame options listed six RS/6000 SP Model Class systems and twelve selections for the various types of expansion frames. In contrast, the simplified frame options listing has a total of five main RS/6000 SP frame options. These options are:

1. A tall model frame
2. A short model frame
3. A tall expansion frame
4. A short expansion frame
5. An SP switch frame

2.2.4 Configurator



The simplified RS/6000 SP model frame classifications has two offerings. These are:

Model 500 Class Short frame with four empty node drawers and a 5.0 KW single-phase SEPBU power supply. All processor nodes and optional switch must be purchased separately for these 1.25 meter frames.

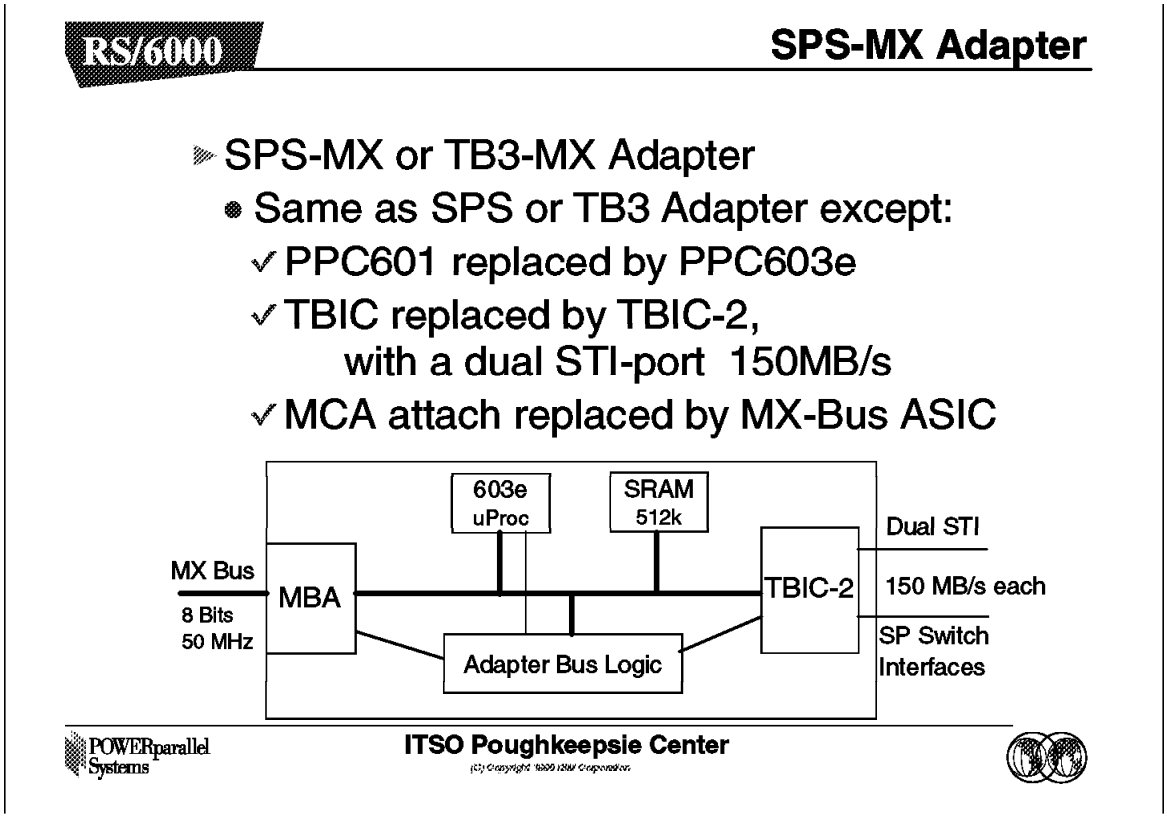
Model 550 Class Tall frame with eight empty node drawers and a 10.5 KW three-phase SEPBU power supply. All processor nodes and optional switches must be purchased separately for these 1.93 meter frames.

RS/6000 SP 332MHz SMP nodes require PSSP 2.4 level and AIX 4.2.1 or AIX 4.3.1 level.

2.3 Switch Adapter

Since the new SMP nodes only have PCI slots, the existing MicroChannel switch adapter cannot be used. A new switch adapter has been developed to allow these new SMP PCI-only nodes to connect to the switch board.

2.3.1 SPS-MX Adapter



The TB3MX adapter is modeled after the TB3 adapter, but with the following changes:

1. The PowerPC 601 is replaced by a PowerPC 603e.
2. The TBIC chip is replaced by the TBIC-2 chip, and so the adapter has two STI switch ports rather than one.
3. The micro-channel interface is replaced by an MX bus ASIC (MBA), which provides the adapter a high bandwidth connection to the I/O bus on the RS/6000 SP 332MHz SMP node system planar.

Note: Only one STI switch port is available with this PSSP release.

Chapter 3. PSSP Enhancements

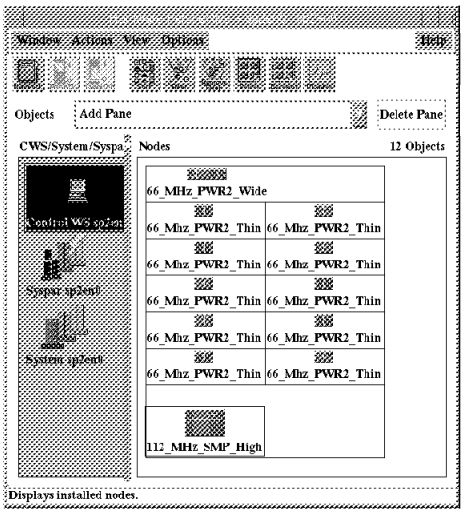
This chapter gives an overview of PSSP 2.4 Enhancements.

© Copyright IBM Corp. 1998

3.1 Description Attributes

RS/6000

PSSP 2.4 Description Attribute



SDR Node Class Description Attribute Values

- 62 Mhz PWR1 Thin
- 66 Mhz PWR2 Thin
- 66 Mhz PWR2 Thin-2
- 120 Mhz P2SC Thin
- 160 Mhz P2SC Thin
- 332 Mhz SMP Thin
- 62 Mhz PWR1 Wide
- 66 Mhz PWR2 Wide
- 66 Mhz PWR2 Wide-2
- 77 Mhz PWR2 Wide
- 77 Mhz P2SC Wide
- 135 Mhz P2SC Wide
- 332 Mhz SMP Wide
- 112 Mhz SMP High
- 200 Mhz SMP High

POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

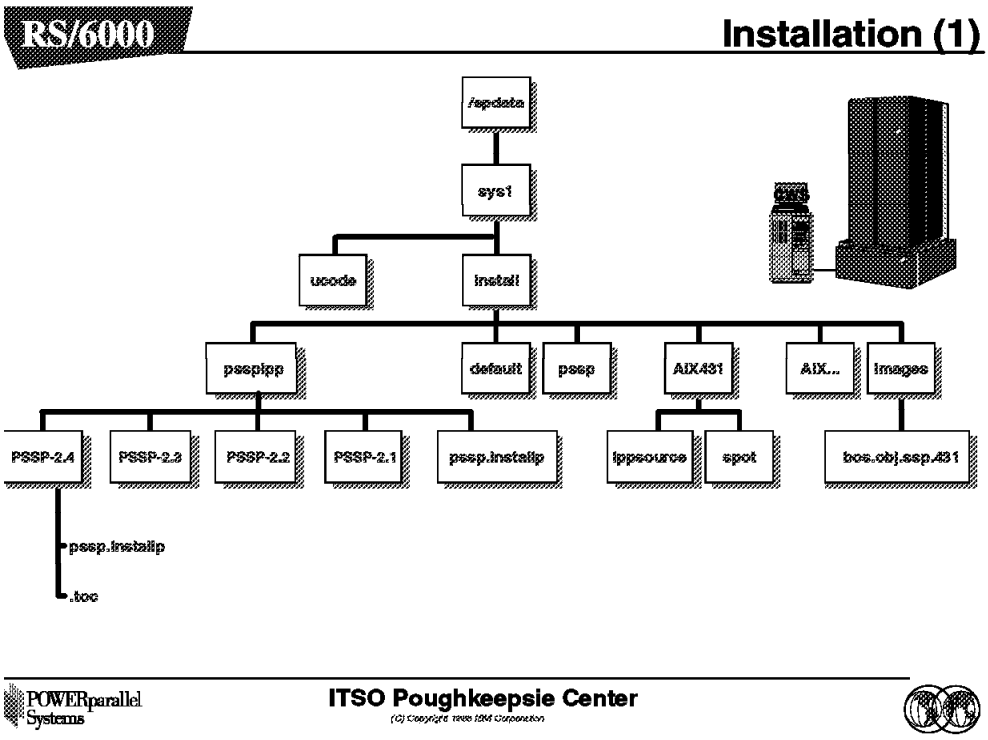
Two new attributes are added to the SDR Class Node:

- Description
- Platform

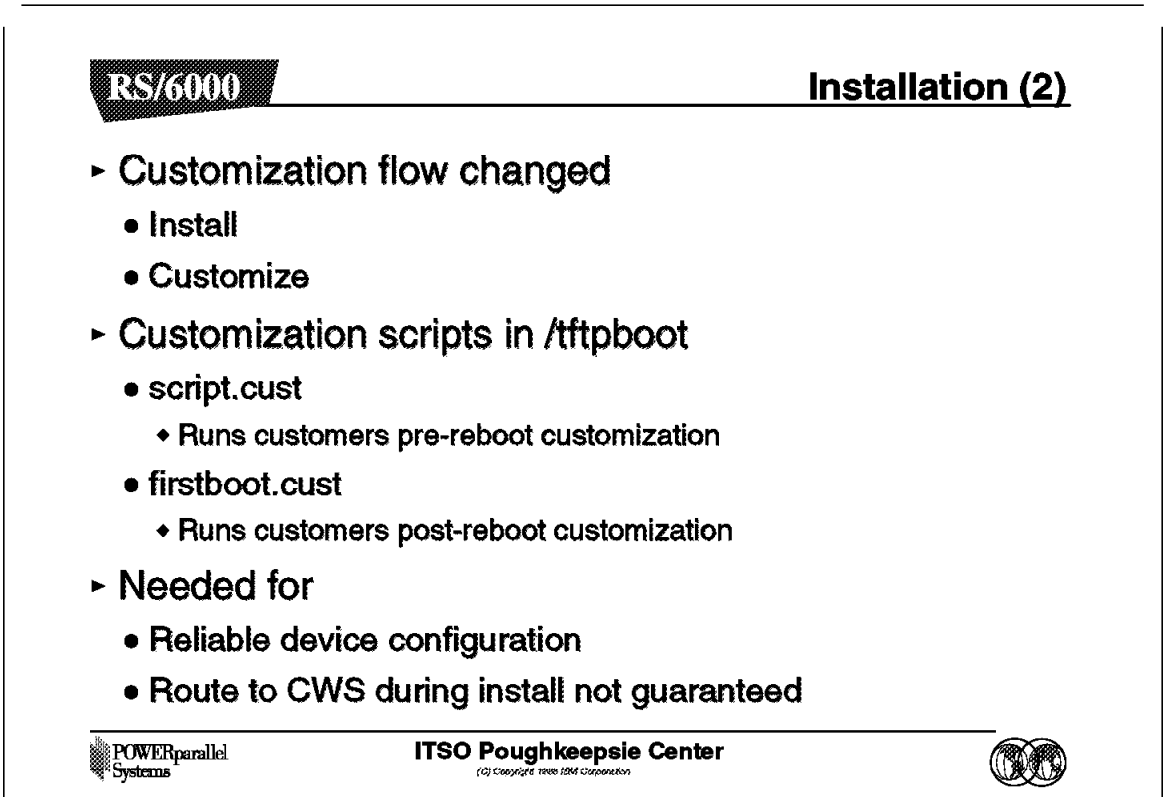
As shown on the foil, the *description* attribute gives more detailed information about the node type.

With PSSP 2.4,PCI-based nodes are supported. For this reason, the SP code needs a method to distinguish the new nodes from the microchannel nodes. The new SDR attribute *platform* addresses this requirement.

3.2 Installation



As shown in the foil, the directory structure has not changed. The PSSP 2.4 software is expected to reside in a separate subdirectory.



In PSSP 2.4, the pssp_script is completely restructured. The device configuration part was moved to a new script, named psspfb_script (pssp firstboot script). The psspfb_script is run during the first reboot of a node after the network installation. Its major task is to complete device configuration on the node.

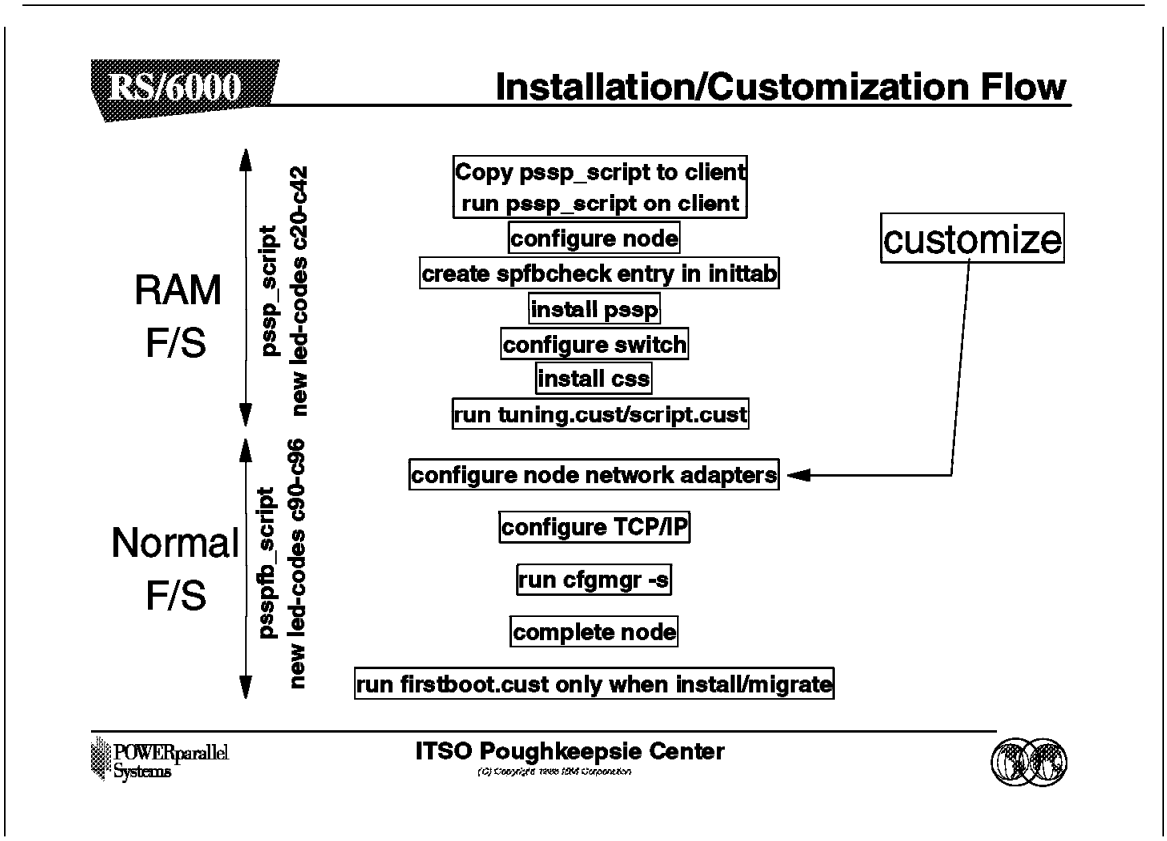
This restructuring was required to support the new PCI-bases nodes and to improve the installation method.

Note

Because pssp_script is now split, script.cust is run before the reboot and firstboot.cust is run after the reboot.

This requires that all device-dependent code from the customer-supplied script.cust needs to move to the firstboot.cust script.

3.3 Installation/Customization Flow



As previously noted, psspfb_script is run during the first reboot of the node. It is run out of /etc/inittab after the /etc/firstboot script. It is invoked by a new module called spfbcheck, which functions in a manner similar to the base AIX module fbcheck, and which causes psspfb_check to be run only after current reboot and then not again until reset by pssp_script.

Due to the new nature of pssp_script and psspfb_script, a restructuring of already existing customer scripts of script.cust and firstboot.cust is required. All device-dependent and routing-dependent code in script.cust needs to be moved to firstboot.cust. Sample scripts are documented in the *Installation and Migration Guide*, GC23-3898 and are also part of the SP code.

Note

See the guidelines for script.cust and firstboot.cust in Appendix C of the *Installation and Migration Guide*, GC23-3898.

In addition, psspfb_script is run out of rc.sp when a node is put in “customize” state and rebooted.

There are a large number of new messages from pssp_script and psspfb_script. Here is an overview of these new messages:

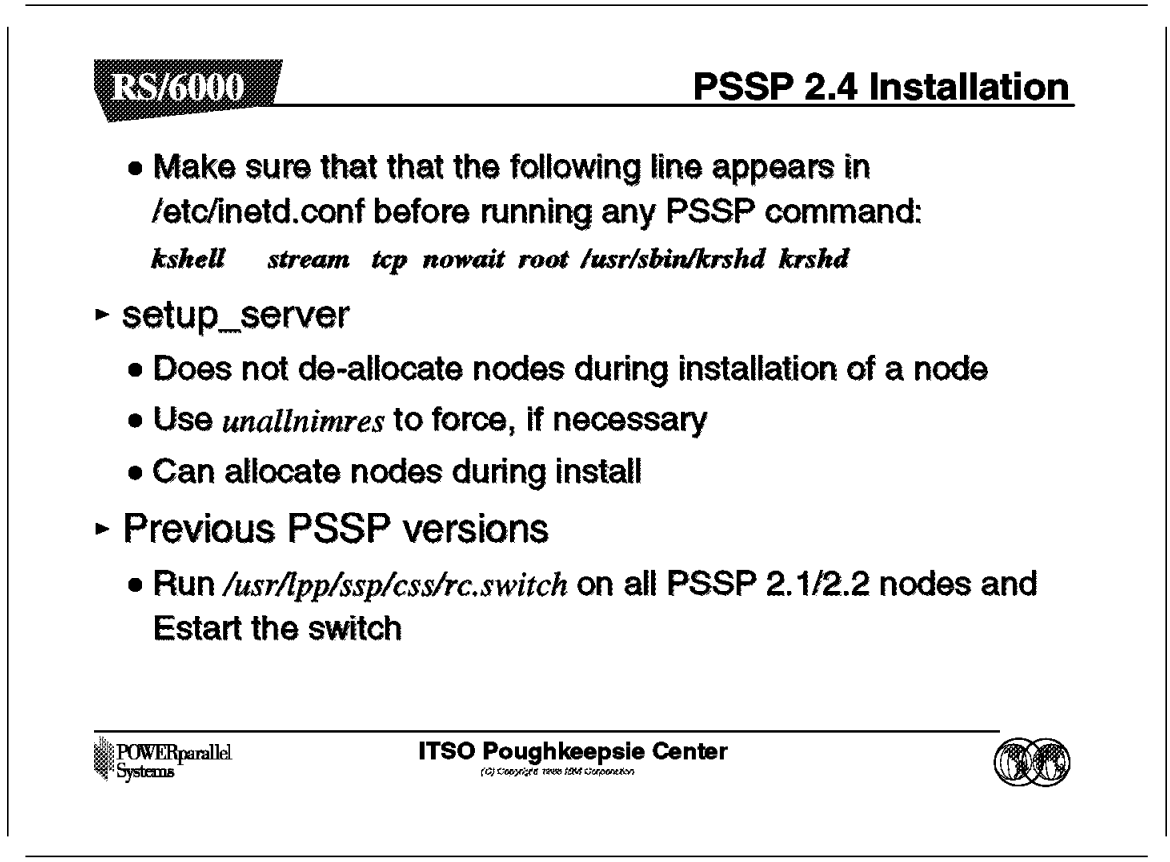
pssp_script

c20 = create_directories	Create PSSP log directories
c21 = setup_environment	Establish environment definitions
c22 = configure_node	Configure node (not adapters)
c23 = create_files	Create /etc/ssp files
c24 = update_etchosts	Update /etc/hosts file
c25 = get_files	Get files from server
c26 = authent_stuff	Do authentication (Kerberos) stuff
c27 = update_etcinittab	Update /etc/inittab file
c28 = upmp_work	Do up/mp handling
c29 = install_prereqs	Install prereq filesets
c30 = install_ssp_clients	Install ssp.clients
c31 = install_ssp_basic	Install ssp.basic
c32 = install_ssp_ha	Install ssp.ha
c33 = install_ssp_sysctl	Install ssp.sysctl
c34 = install_ssp_pman	Install ssp.pman
c35 = install_ssp_css	Install ssp.css
c36 = install_ssp_jm	Install ssp.jm
c37 = delete_master_rhosts	Delete /.rhosts for NIM master
c38 = create_dump_lv	Create dump logical volume
c39 = run_tuning_cust	Run customer's tuning.cust
c40 = run_script_cust	Run customer's script.cust
c41 = config_switch	Add switch odm entries
c42 = Run_psspfb_script	run psspfb_script if "customize"

psspfb_script

c90 = setup_environment	Establish environment definitions
c92 = config_adapters	Configure adapters
c93 = config_inet0	
c94 = run_cfgmgr	
c95 = complete_node	
c96= run_firstboot_cust	

3.4 PSSP 2.4 Installation



Checking the */etc/inetd.conf* entry is only required when running an AIX release of 4.3.1 or higher. Starting with AIX 4.3.1, SP software uses the AIX included security mechanism.

With AIX 4.3.0 (or an earlier AIX release), you will still see an entry like the following:

```
kshell stream  tcp     nowait  root     /usr/lpp/ssp/rcmd/etc/kshd kshd
```

Note that setup_server was originally designed so that it always deallocated NIM resources and then reallocated them. This does not create a problem as long as no one else runs setup_server while an install process is in progress.

However, if this does happen, NIM resources could be deallocated for this currently installed node and the install process would be unsuccessful. To fix that problem, a change was made to PSSP 2.3 (this change was also applied to PSSP 2.2 with a higher PTF level). Now the resources are not deallocated every time setup_server is running.

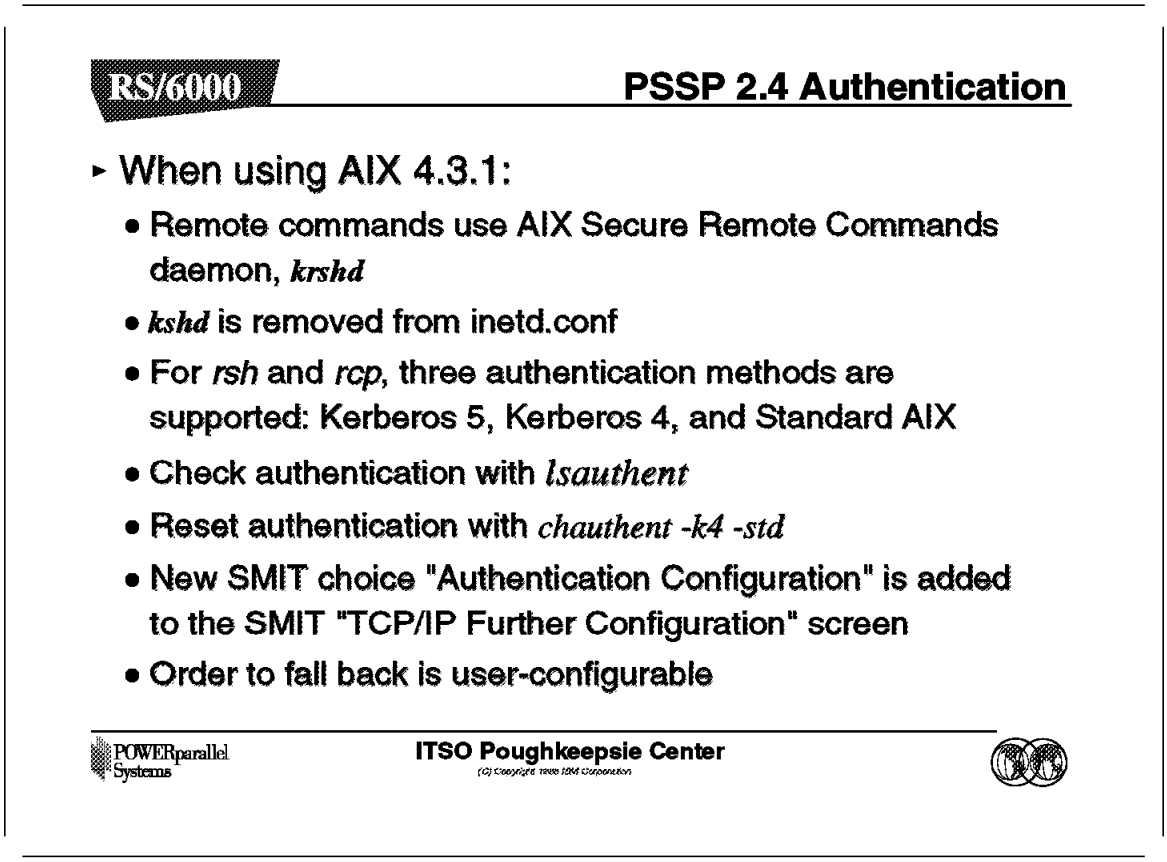
For example, if the node is in maintenance mode and should be changed back to disk, the following commands must be executed:

1. unallnimres -l <node_number>

2. Set the bootp response to disk.
3. Run setup_server.

Note: In nodes running PSSP 2.1 or PSSP 2.2, run the /usr/lpp/ssp/css/rc.switch script to retart the Worm daemon.

3.5 Authentication



The remote commands (rcmds) include `rsh`, `rcp`, `rlogin`, `telnet` and *ftp*.

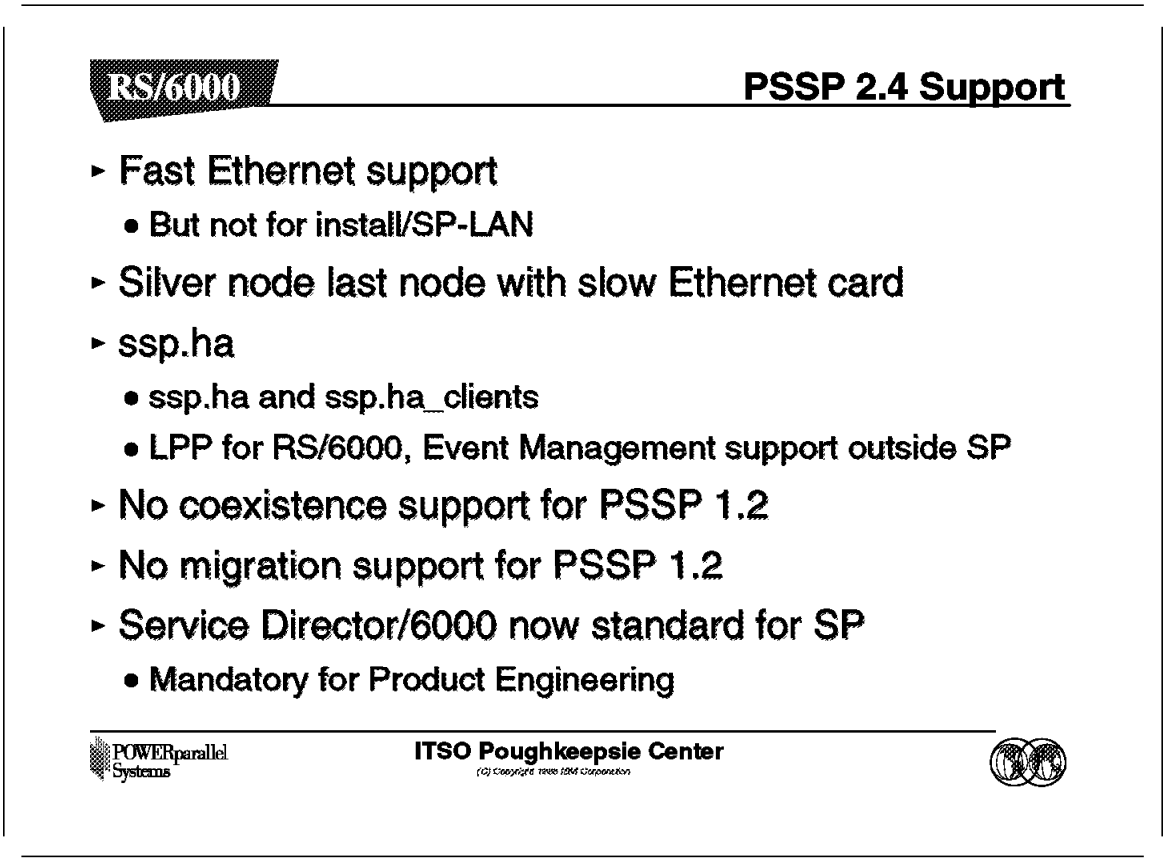
AIX 4.3.1 includes versions of the `rsh` and `rcp` commands that support multiple methods of client-server authentication including:

- Standard AIX (that is using *.rhosts* files)
- Kerberos 4
- Kerberos 5

In AIX 4.3.1, an incoming Kerberos 5 ticket is upgraded to full DCE credentials for all rcmds. Kerberos 4 is included for only `rsh` and `rcp`, in order to provide backward compatibility for the SP systems. Note that Kerberos 4 tickets are not upgraded to DCE credentials.

If an authentication method fails to connect, the commands will fall back to the next one that is configured. If none of the methods configured succeeds, then the command will fail.

3.6 PSSP 2.4 Support



With PSSP 2.4, fast Ethernet adapters (100 BASE-T) are supported but not for the installation of the SP nodes.

Note

Fast Ethernet adapters are only supported for external LANs.

The Wildcat nodes are shipped with integrated 10 Mb/s (slow) Ethernet chips. These will be probably the last nodes with slow Ethernet interfaces.

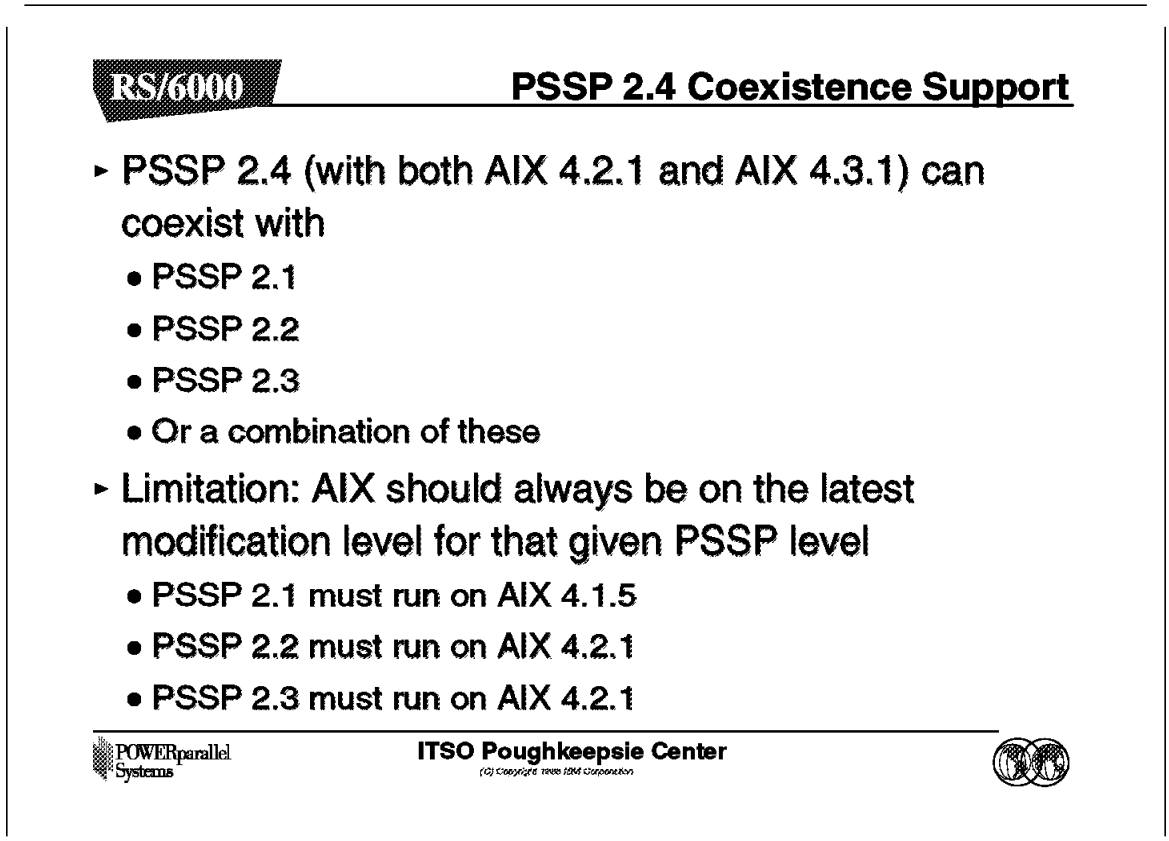
Note that ssp.ha is now a separate lpp (still included in pssp.installp). This is a first step to future support of event management outside the SP.

Note

PSSP 1.2 and AIX 3.2.5 is no longer supported in a mixed system partition when PSSP 2.4 is running on the Control Workstation.

For more information about Service Director/6000, see Chapter 7, “Service Director/6000” on page 247.

3.7 Coexistence Support



The following levels of AIX and PSSP releases coexist with PSSP 2.4:

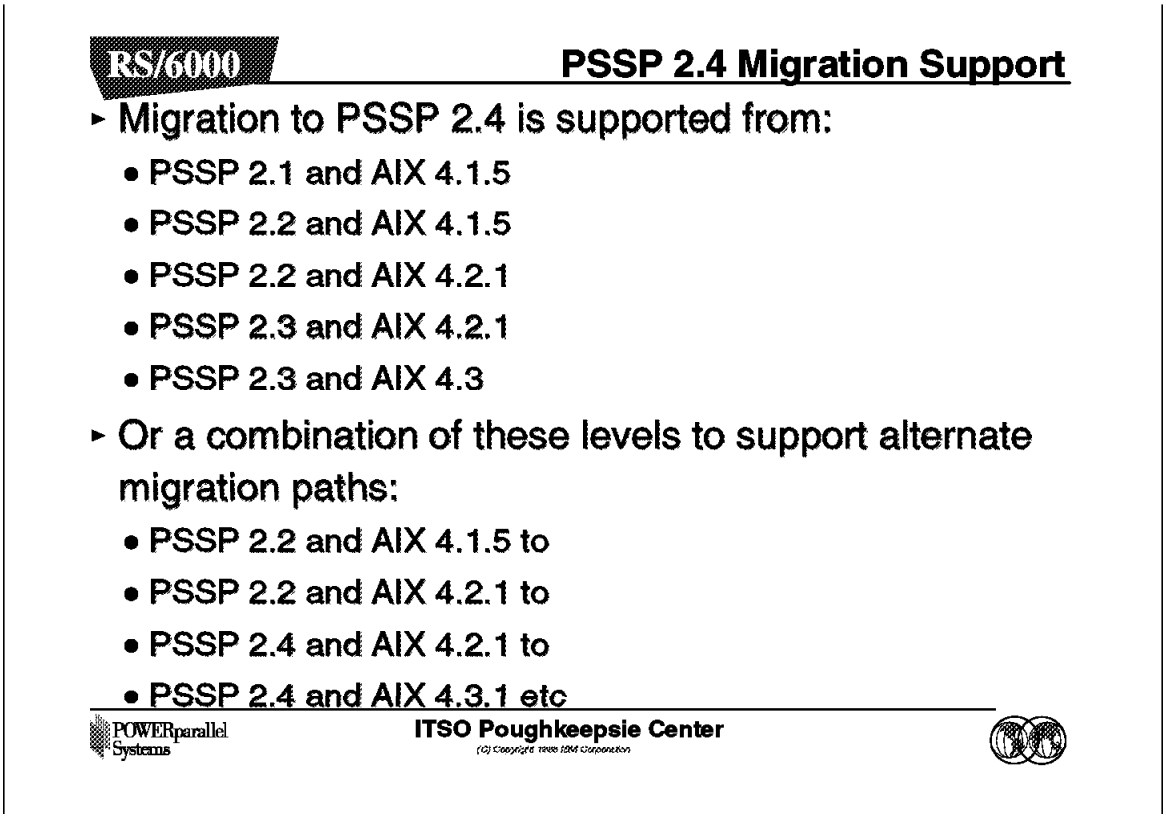
- PSSP 2.1 and AIX 4.1.5
- PSSP 2.2 and AIX 4.1.5, 4.2.1
- PSSP 2.3 and AIX 4.2.1, 4.3
- PSSP 2.4 and AIX 4.2.1, 4.3

AIX should always be on the latest modification level.

Note

PSSP 1.2 and AIX 3.2.5 are no longer supported in a mixed system partition when PSSP 2.4 is running on the Control Workstation.

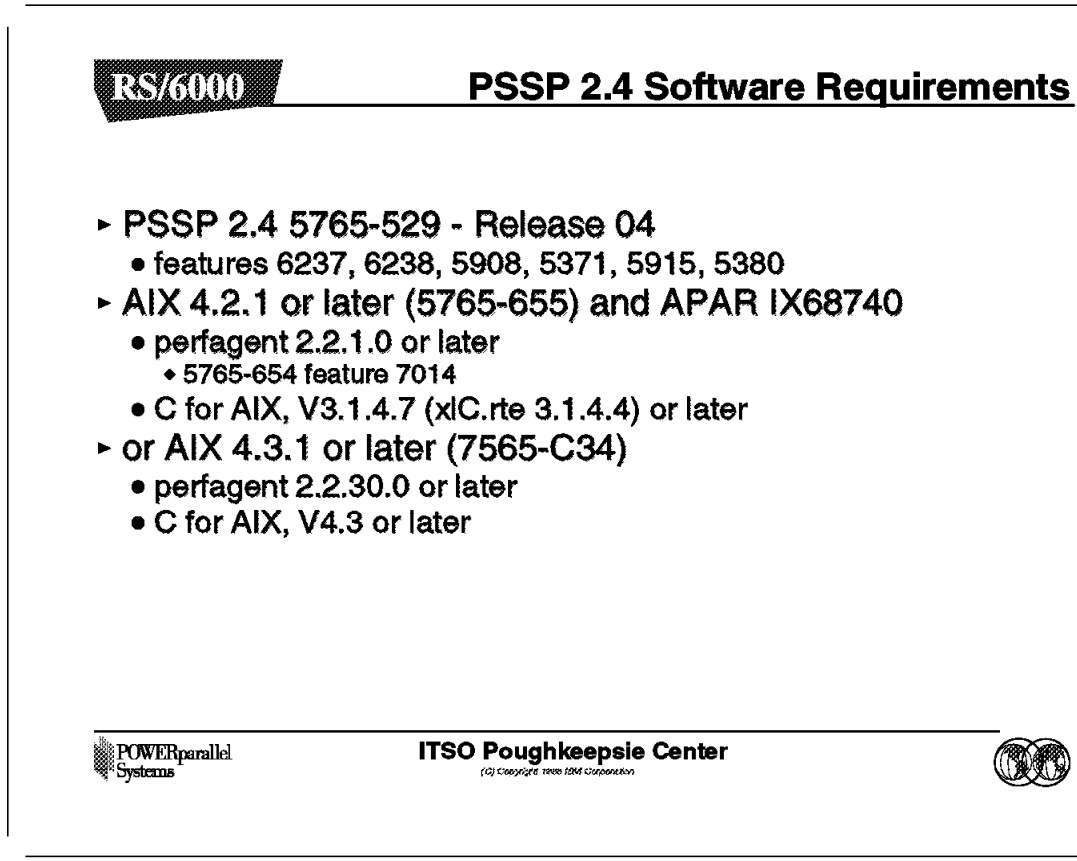
3.7.1 Migration Support



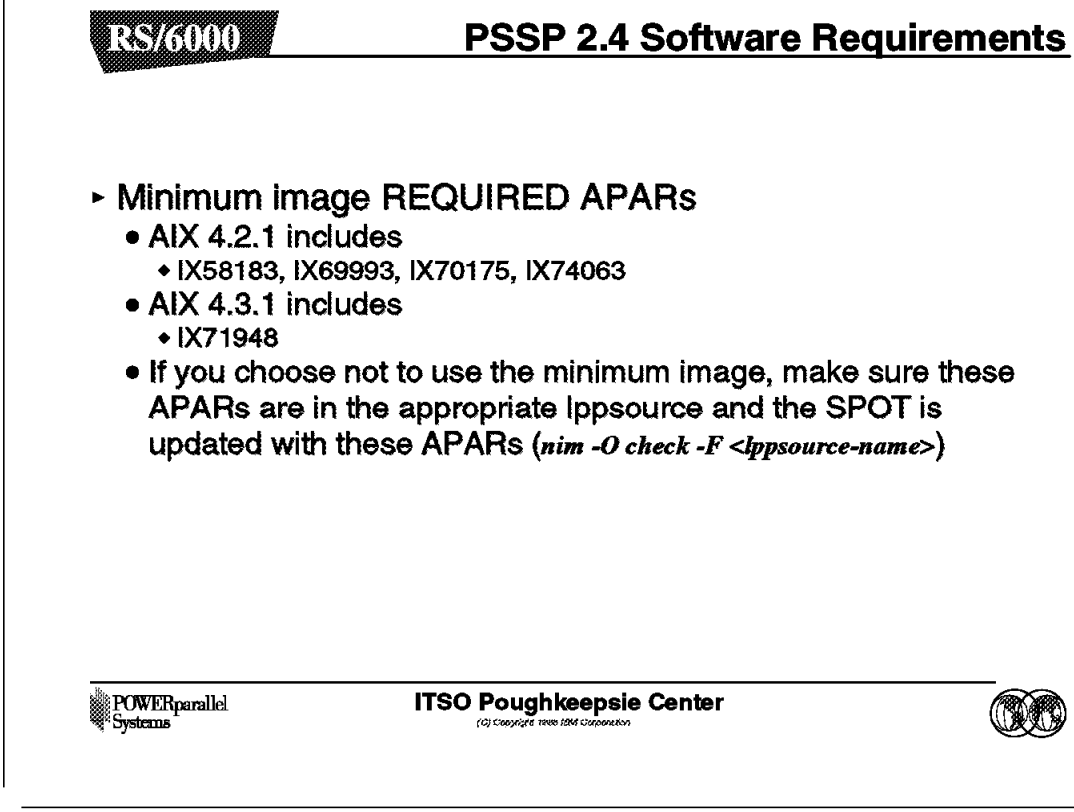
A direct migration path is supported to PSSP 2.4 on AIX 4.2.1 or AIX 4.3 from:

PSSP 2.1 and AIX 4.1.5	to PSSP 2.4 and AIX 4.2.1, 4.3
PSSP 2.2 and AIX 4.1.5, 4.2.1	to PSSP 2.4 and AIX 4.2.1, 4.3
PSSP 2.3 and AIX 4.2.1	to PSSP 2.4 and AIX 4.2.1, 4.3
PSSP 2.3 and AIX 4.3	to PSSP 2.4 and AIX 4.3

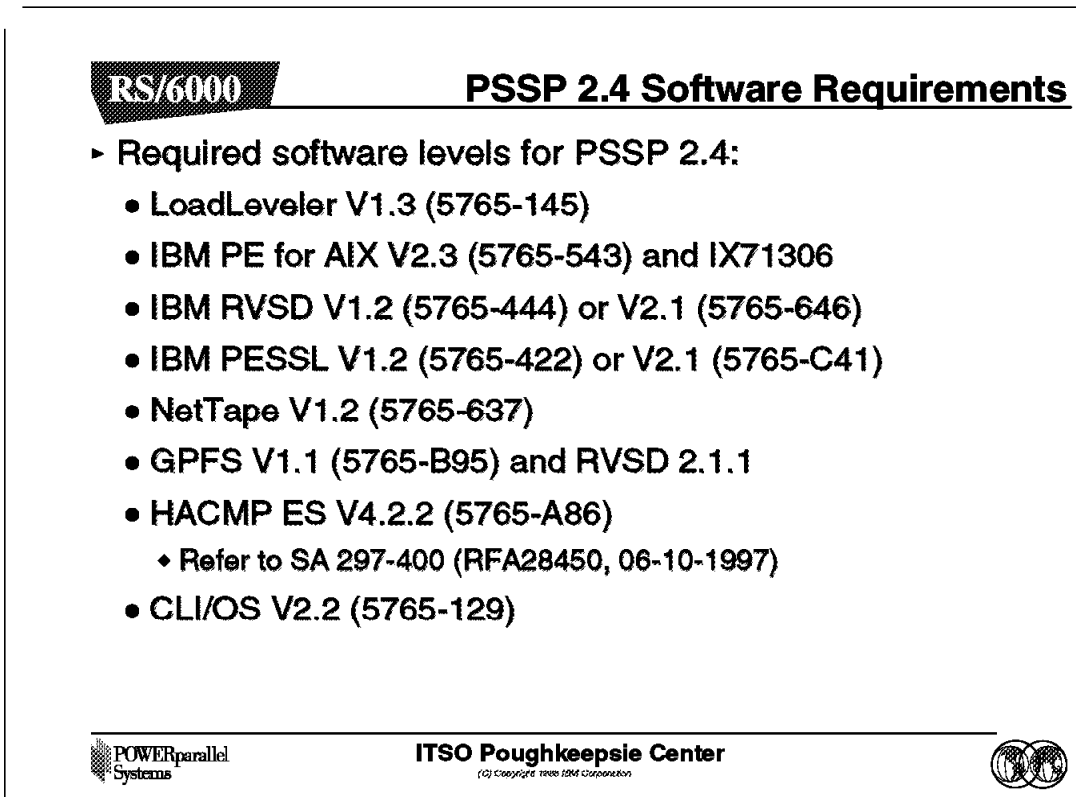
3.8 Software Requirements



This list gives an overview of the required software levels when running PSSP 2.4.

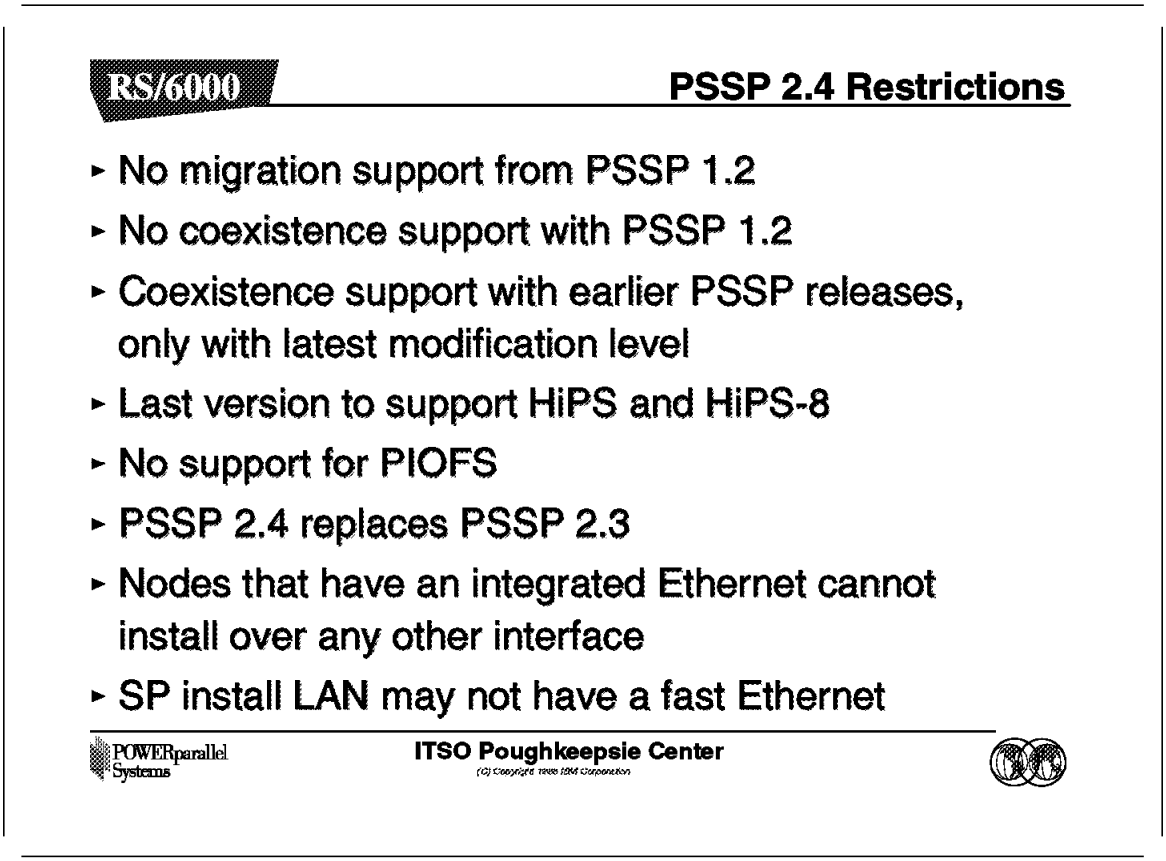


This list gives an overview of the required software levels when running PSSP 2.4.



This list gives an overview of the required software levels when running PSSP 2.4.

3.9 PSSP 2.4 Restrictions



This foil lists limitations that apply when running PSSP 2.4.

Note

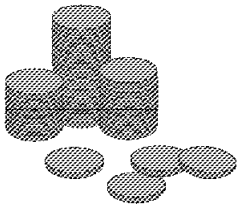
PSSP 1.2 and AIX 3.2.5 are no longer supported in a mixed system partition when PSSP 2.4 is running on the Control Workstation.

3.10 Pricing Facts

RS/6000

PSSP 2.4 Pricing Facts

- Silver node is approximately half the price of old 604e node
- PSSP 2.4 nocharge for customers upgrading from PSSP 2.2 or PSSP 2.3
- PSSP 2.4 charge fee per node for customers upgrading from PSSP 2.1 or PSSP 1.2



POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

The Silver nodes are well-suited for commercial environments. For further information, refer to 3.11, “Silver Node Performance Facts” on page 51.

There is no charge for upgrading from PSSP 2.2 and PSSP 2.3 to PSSP 2.4.

3.10.1 Pricing Details

RS/6000

Pricing Details

- 9077-04s US$53,000
  9077-16s US$72,500 (incl Switch Adapter)
  4021 US$30,000
- PSSP Upgrade Prices to PSSP 2.4 in US$

	4 nodes	16 nodes	32 nodes	64 nodes	128 nodes
PSSP 1.2	750	11,250	22,500	33,750	45,000
PSSP 2.1	500	7,500	15,000	22,500	30,000

POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1996 IBM Corporation

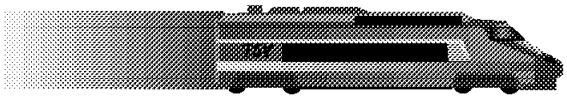
Note that these are only estimated prices for the U.S market. Pricing could be different for your home country.

3.11 Silver Node Performance Facts

RS/6000

Silver Node Performance Facts

- tpmc: App. 10,000 tpm
- Specfp: Half as good as P2SC 160 MHz
- SpecInt: Twice as good as P2SC 160 MHz
- Excellent commercial performance
  - For half the price of the 604e High Node
- SPS-MX: 130 MB/s IP !!!



POWERparallel Systems

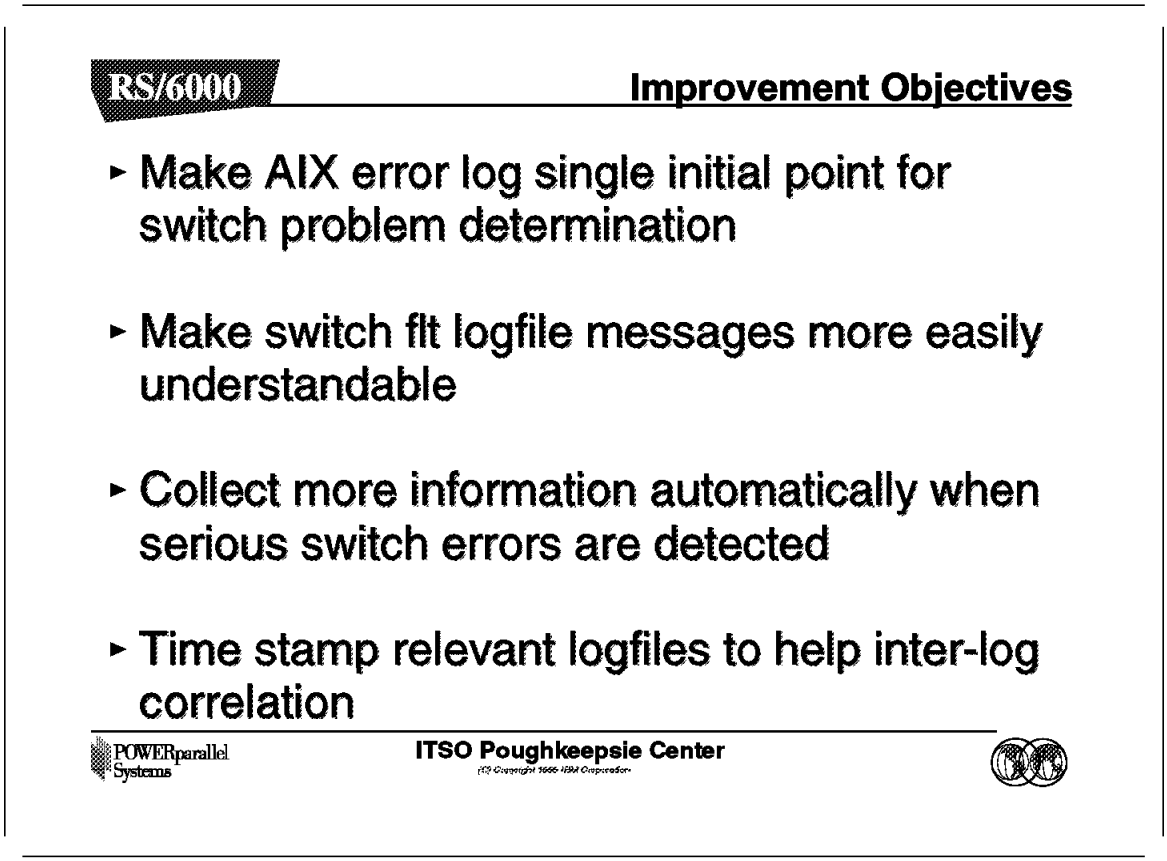
ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

The New Silver node is the first PCI-based node for the RS/6000 SP. It offers an excellent price/performance ratio and is specially suitable for commercial environments.

In addition, the Silver node is the first node exploiting the SPS-MX adapter which provides a maximum transfer rate of 130 MB/s.

Chapter 4. Switch RAS Improvements



This chapter describes some of the reliability, availability, serviceability (RAS) and usability improvements for the switch under PSSP 2.4.

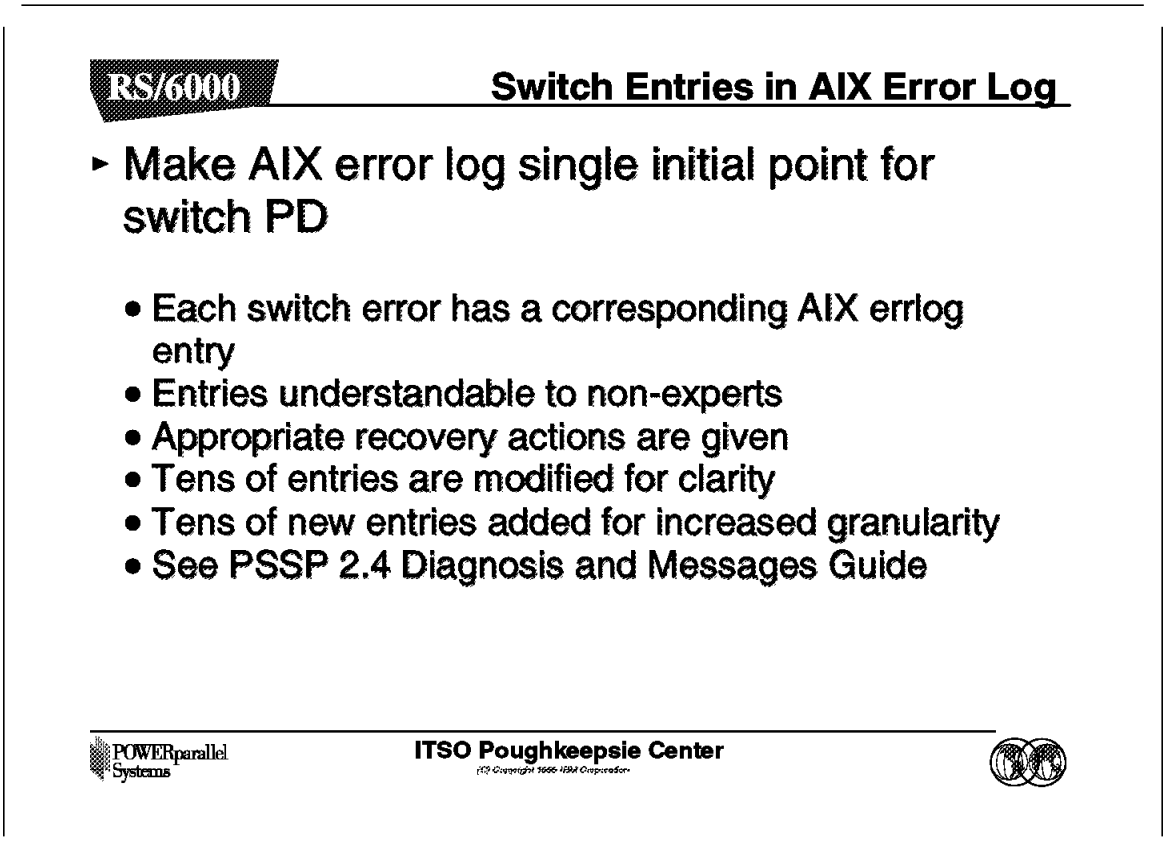
The main goals of these improvements can be summarized as follows:

- Ensure that the AIX error log is made suitable to be used as the starting point for any problem determination activity related to the switch. At a minimum, this requires that any and every switch error has a corresponding entry in the error log.
- Ensure that the switch flt logfile information is more easily intelligible to customers and field personnel.
- Ensure that enough information is collected automatically when severe switch errors occur. This will enable the IBM Software Support to help the customer more quickly and efficiently.
- Ensure that important logfiles are time stamped so that it would be possible to see whether a given event occurred before or after another event.

The last objective is implemented by time stamping the files rc.switch.log, cable_miswire, router.log, and router_failed.log. The next few foils discuss the way the rest of the objectives are implemented.

© Copyright IBM Corp. 1998

4.1 Switch Entries in AIX Error Log



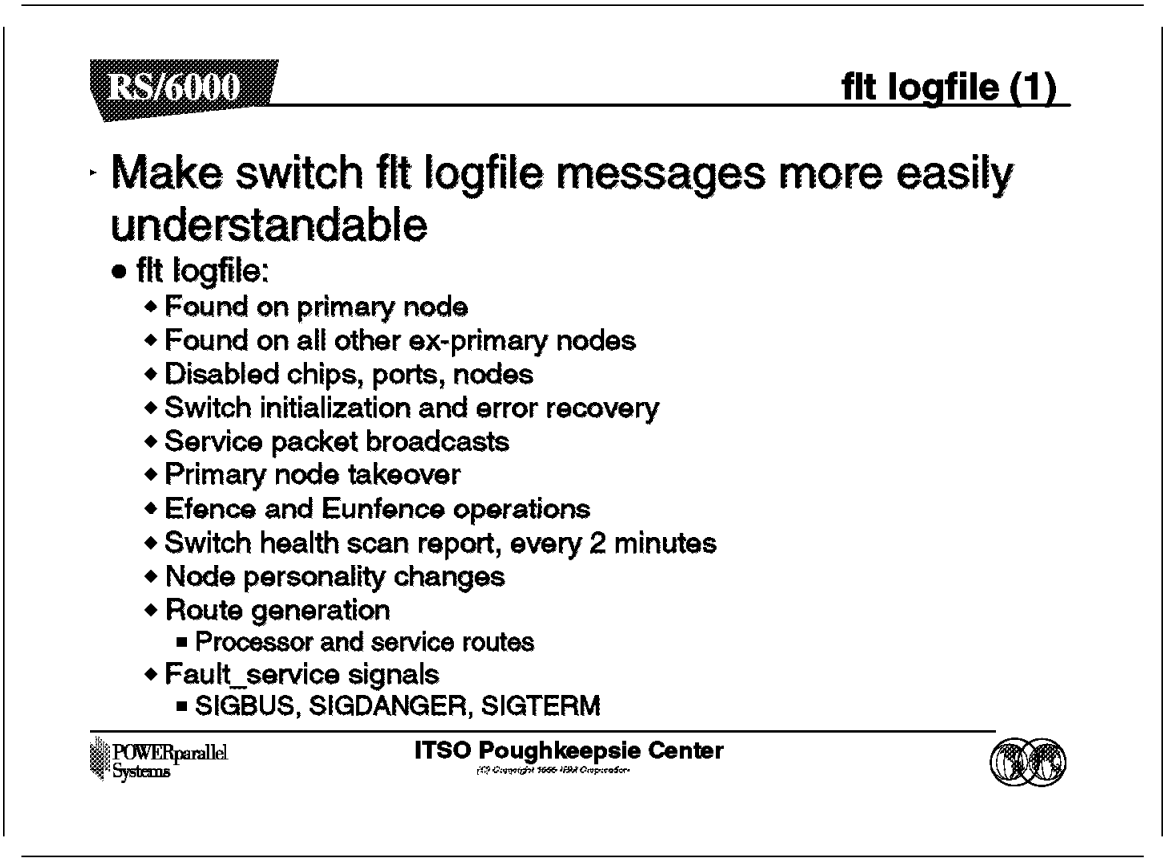
This objective is to allow the AIX error log to be used as the starting point for any problem determination activity related to the switch. To support this, all error conditions that the switch fault service daemon encounters, which it usually reports externally in some way, will have a corresponding AIX error log entry.

These entries describe the problem in terms understandable to non-experts and give the appropriate recovery actions for the problem. In addition, these entries are very granular; that is, each problem is uniquely described by the errlog entry generated for it. No generic entries for whole classes of problems are included unless differentiated clearly by detailed data strings.

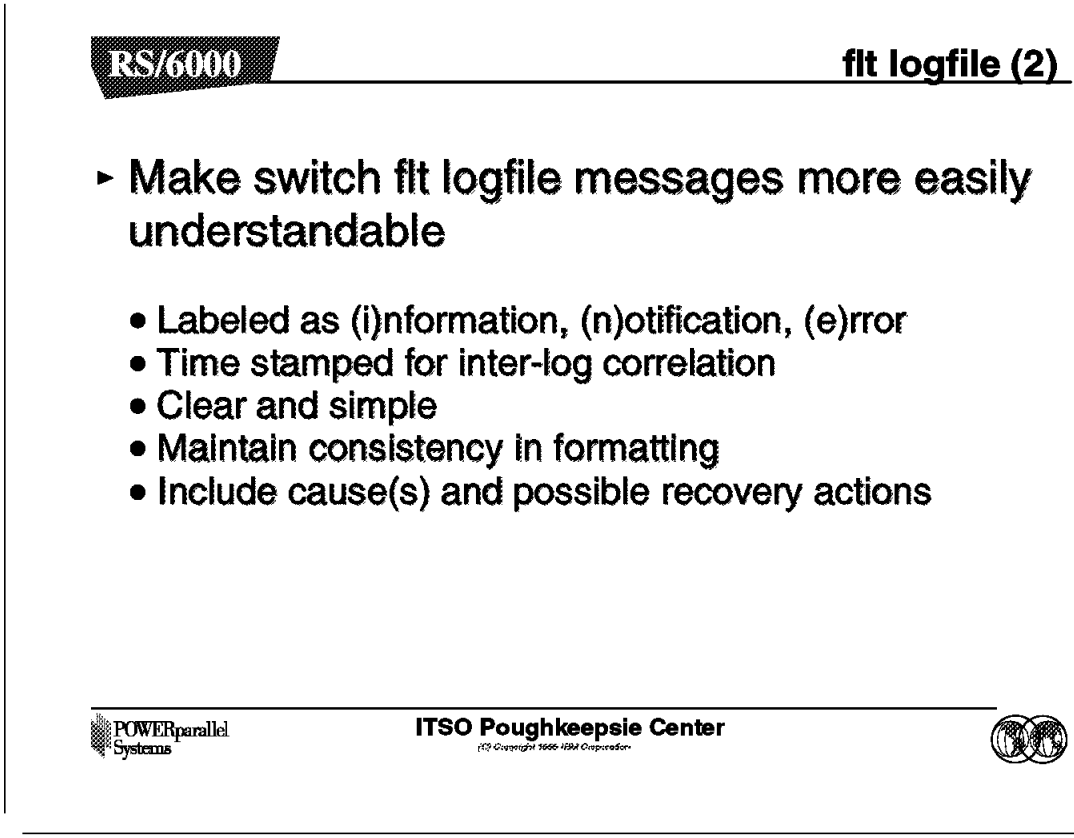
To isolate a switch or adapter error, first view the error log. For a switch-related error, login to the primary node. For adapter problems, login to the suspect node. Then execute the command `errpt` or `errpt -a` in order to view the errors involved. An example of switch errlog entries is given in the following foil. You can find a complete list of these entries in Chapter 8 “Diagnosing Switch Problems” of *PSSP 2.4 Diagnosis and Messages Guide*.

```
                                                                    --
LABEL:  SP_SW_SDR_FAIL_RE
IDENTIFIER: EC771C6B

Date/Time:       Mon Apr 13 12:26:18
Sequence Number: 311
Machine Id:      000168355700
Node Id:         sp21n08
Class:           S
Type:            TEMP
Resource Name:   css

Description
Switch daemon SDR communications failed

Probable Causes
Ethernet overloaded
Excessive SDR traffic
SDR daemon or control workstation down
Software Error

Failure Causes
Excessive ethernet traffic
SDR daemon not running

        Recommended Actions
        Check if SDR daemon is up
        Call software service if problem persists

Detail Data
Software ID String
LPP=PSSP,Fn=set_node_info.c,SID=1.9,L#=469,
Failure cause
switchResponds saw an API failure
---------------------------------------------------------------------------
LABEL:  SP_SW_UNINI_LINK_RE
IDENTIFIER: A388E1C3

Date/Time:       Mon Apr 13 12:26:18
Sequence Number: 310
Machine Id:      000168355700
Node Id:         sp21n08
Class:           S
Type:            INFO
Resource Name:   Worm

Description
Links not initialized during Estart

Probable Causes
Switch cable mis-wired
Switch cable or switch failure

User Causes
Switch cable loose or disconnected
Switch cable mis-wired

        Recommended Actions
        Check / reconnect / replace cable if problem persists
        See /var/adm/SPlogs/css/cable_miswire for possible miswired
           cables.

Detail Data
Software ID String
LPP=PSSP,Fn=fsd_fsm.c,SID=1.56.4.2,L#=4090,
Chip and Port Number(s)
      100015          3
```

4.2 flt logfile (1)



The flt log file was originally designed to log only “faults” in the switch fabric. However, it now contains a good deal of information for problem determination.



The messages generated for the flt file are modified to include a severity field. Three levels of severity are used:

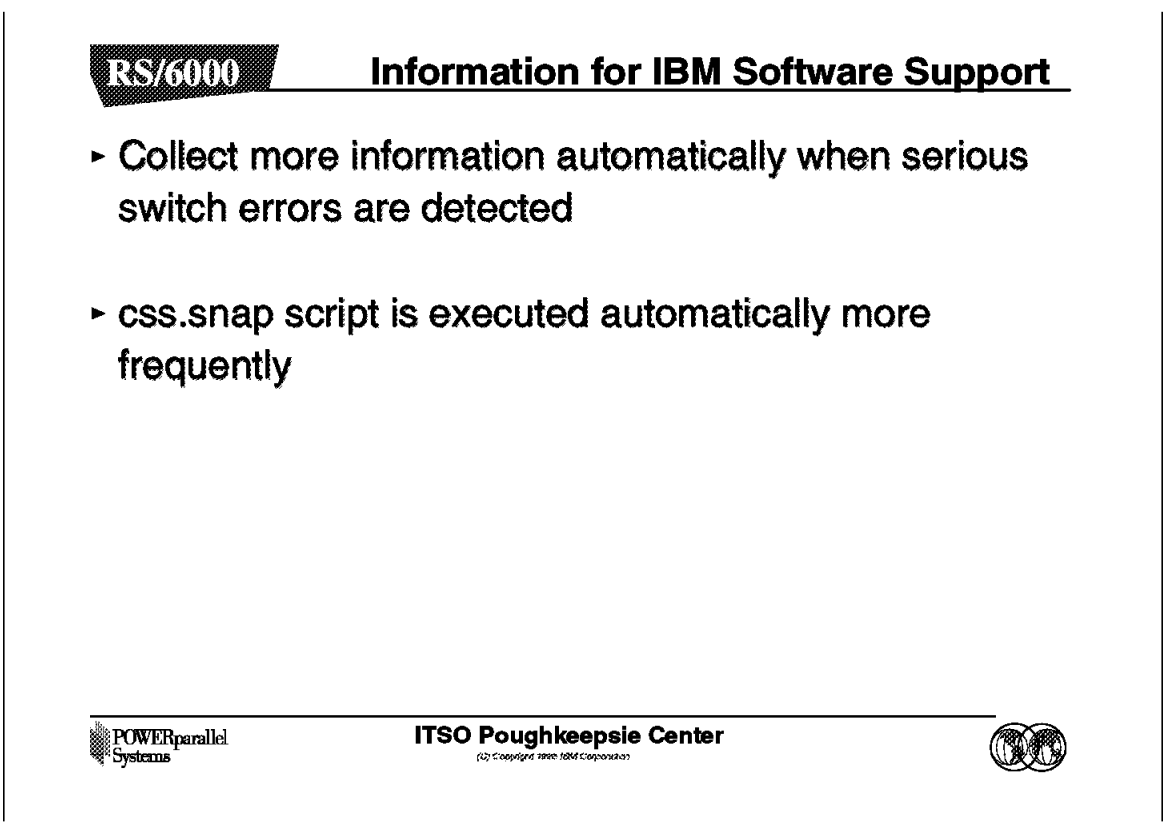
1. Those with least severity, that is those that are meant just to convey the occurrence of an event, are labeled (i) for information.
2. Those messages that deserve notice, but are not strong enough to be classified as error, are labeled (n) for notification.
3. The most severe messages that indicate an error requiring a recovery action, are labeled (e) for error.

Note that two other items also changed in the flt file. First is the use of the *reliable_hostname* in the messages instead of *switch_node_number*. Second is the value displayed for the First Error Capture Register (FECR) and the Second Error Capture Register (SECR). In the past, these were displayed as a hexadecimal string that required a secret decoder to decipher. In PSSP 2.4, the meanings of these bits are displayed in the log.

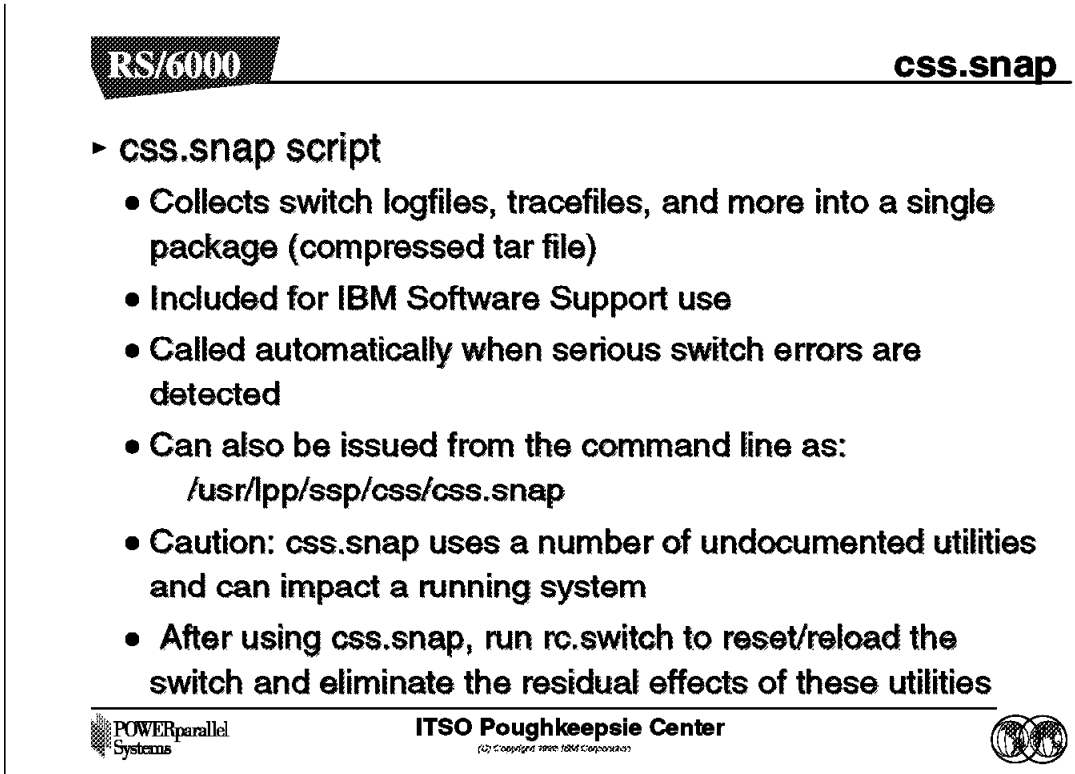
The following is an example of the output that is produced in the flt file. The **boldfaced** text is added by the author in order to describe the type of flt messages that it precedes.

```
Information Messages:
(i) 03/30/98 18:57:43 : 2510-811 Fault service daemon's personality has
    been changed to Primary.
(i) 04/02/98 16:43:25 : 2510-821 The second phase of the switch
    initialization will be retried.
(i) 04/08/98 09:46:43 : 2510-744 Estart initiated.
(i) 04/08/98 15:04:38 : Backup didn't respond to scan.
(i) 04/14/98 16:45:12 : Processor routes down loaded successfully.
(i) 04/20/98 10:35:10 : 2510-744 SP Switch error recovery initiated.

Notification Messages:
(n) 04/02/98 16:43:25 : 2510-828 Error register bits found on device ID
    100022. Disabling the device.
(n) 04/02/98 16:43:25 : 2510-826 Device ID 15 un-initialized during
    switch initialization. Disabling the device.
(n) 04/08/98 09:51:36 : 2510-823 The fault service daemon process has
    exited.
(n) 04/08/98 14:22:31 : 2510-606 A switch Error/Status was service
    packet received during a broadcast operation.
(n) 04/20/98 11:40:32 : 2510-743 Disabling port 0 (jack 10) of chip 5
    on the switch in slot 17 of frame 2
(n) 04/20/98 11:40:32 : 2510-749 Turning off switchResponds bits for
    node 21 in the SDR

Error Messages:
(e) 04/20/98 13:23:37 : 2510-831 Nodes attached to Device 100025 not
    reachable for auto-join.
(e) 04/13/98 10:25:36 : 2510-820 Primary's link to the switch network
    is not in the initialized state. Estart could not be executed.
(e) 04/13/98 18:24:07 : 2510-906 Scan detected a problem with device 100015.
(e) 04/13/98 18:24:07 : 2510-818 Switch Scan failed with a return code
    of 6. Estart will be executed.
(e) 04/14/98 16:46:13 : 2510-894 Error found in handleUnfence()

Use of reliable_hostname instead of switch_node_number:
(i) 03/30/98 18:58:25 : The Primary backup is node k47n16.ppd.pok.ibm.com

Related Notification and Information Message:
(n) 03/30/98 18:58:17 : Switch and Adapter Error bits found during
    switch initialization.
(i) 03/30/98 18:58:17 : Device ID = 100010
(i) 03/30/98 18:58:17 : 2510-793 First  Error Capture Register  = 000001.
(i) 03/30/98 18:58:17 : 2510-741 Second Error Capture Registers =
    00000000 00000000 00000000 00000000 00000000 000001

First Error and Second Error Capture Register Decoder:
(i) __Date__ __Time__ _Msgid__ Reg ___Location___ Ch Po __Type_of_SECR_Error__
(n) 03/30/98 19:01:13 : 2510-767 SEC E01-S00-BH-J20  6  0 Recv Link Sync Failure
(n) 03/30/98 19:01:13 : 2510-778 SEC E01-S00-BH-J20  6  0 Send Link Sync Failure

(i) __Date__ __Time__ _Msgid__ Reg ___Location___ Ch Po __Type_of_SECR_Error__
(n) 04/08/98 15:17:01 : 2510-760 FEC E02-S17-BH-J10  5  0 Incorrect EDC

Sequence of Error Messages:
(e) 04/07/98 14:45:29 : 2510-730 Node k46n05.ppd.pok.ibm.com NOT
    UnFenced, rc = -32.
(e) 04/07/98 14:47:29 : 2510-730 Node k46n05.ppd.pok.ibm.com NOT
    UnFenced, rc = -32.
(e) 04/07/98 14:49:29 : 2510-730 Node k46n05.ppd.pok.ibm.com NOT
    UnFenced, rc = -32.
(e) 04/07/98 14:49:32 : 2510-832 Node k46n05.ppd.pok.ibm.com failed 3
    consecutive autojoin attempts. The node put into the fenced without
    autojoin state.
```

4.3 Information for IBM Software Support



The css.snap script gathers relevant information such as error logs, configuration files, and trace files. The next foil presents more details about this script.



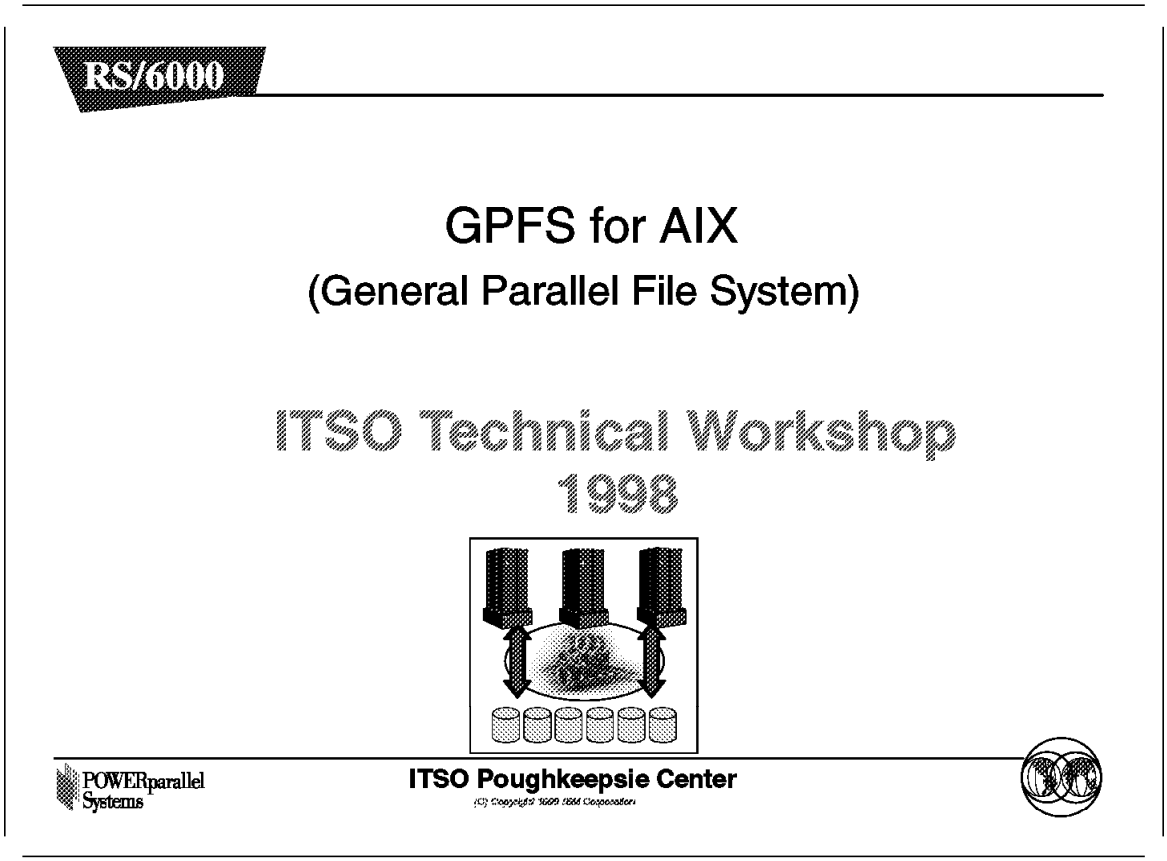
The css.snap script typically collects the following files: cable_miswire, cable_miswire.old, core (fault service daemon dump file), css.snap.log, css_dump.out, daemon.stderr, daemon.stdout, dtbx.trace, dtbx.failed.trace, errpt.out (most recent errpt -a and errpt entries), flt, fs_daemon_print.file, netstat.out (current netstat -l css0 and netstat -m output), out.top, rc.switch.log, regs.out, router.log, scan_out.log, scan_save.log, tb_dump.out, vdidl.out, and worm.trace.

The files ending in .out are produced by running the appropriate command to dump internal (in memory) trace information or dump data to a file. The completed output file will be found in the /var/adm/SPlogs/css directory as a compressed tar file.

The css.snap script avoids flooding /var using the following algorithm:

1. If less than 10% of /var is free, css.snap exits.
2. If the css portion of /var is more than 30% of the total space in /var, css.snap erases old snap files until css′s share sinks below 30%. If it is successful, the snap proceeds. If not, snap exits.

Chapter 5. General Parallel File System (GPFS)



This chapter discusses the General Parallel File System (GPFS), a licensed program product supported with PSSP 2.4.

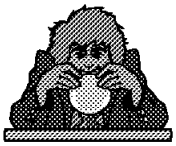
© Copyright IBM Corp. 1998

5.1 The Need for a Parallel File System on the SP

RS/6000

The Need for a Parallel File System

- A serial application can often run out of I/O performance on a single SP Node
- A serial application may often need to access data that is located on a disk physically located on a different SP node
- Capacity requirements may exceed the capabilities of one SP node
- High Availability for a critical file or file system may be required
- Parallel applications need access to disks that are spread across a number of nodes
- Servers outside the SP system need access to data on the SP system



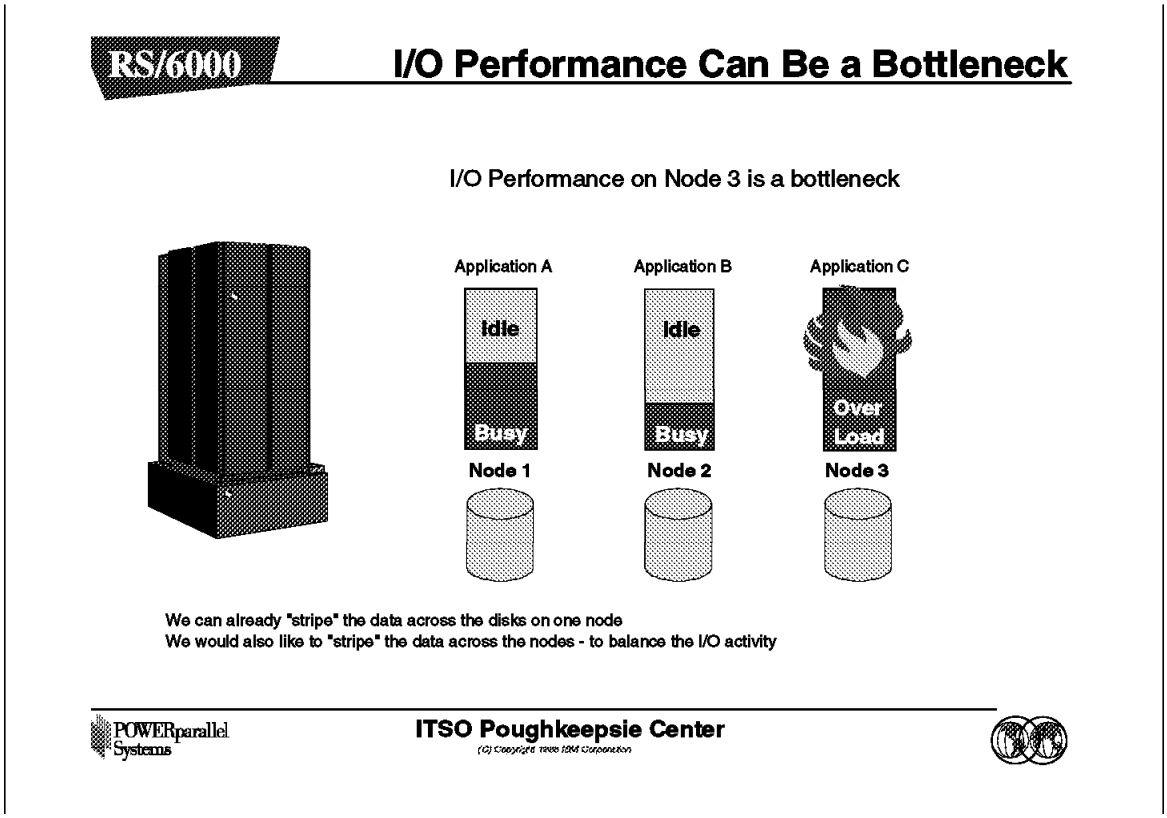
POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

This section deals with the requirements for a parallel file system. These are only partially satisfied today with products such as PIOFS, Virtual Shared Disk, NFS, JFS, and DFS.

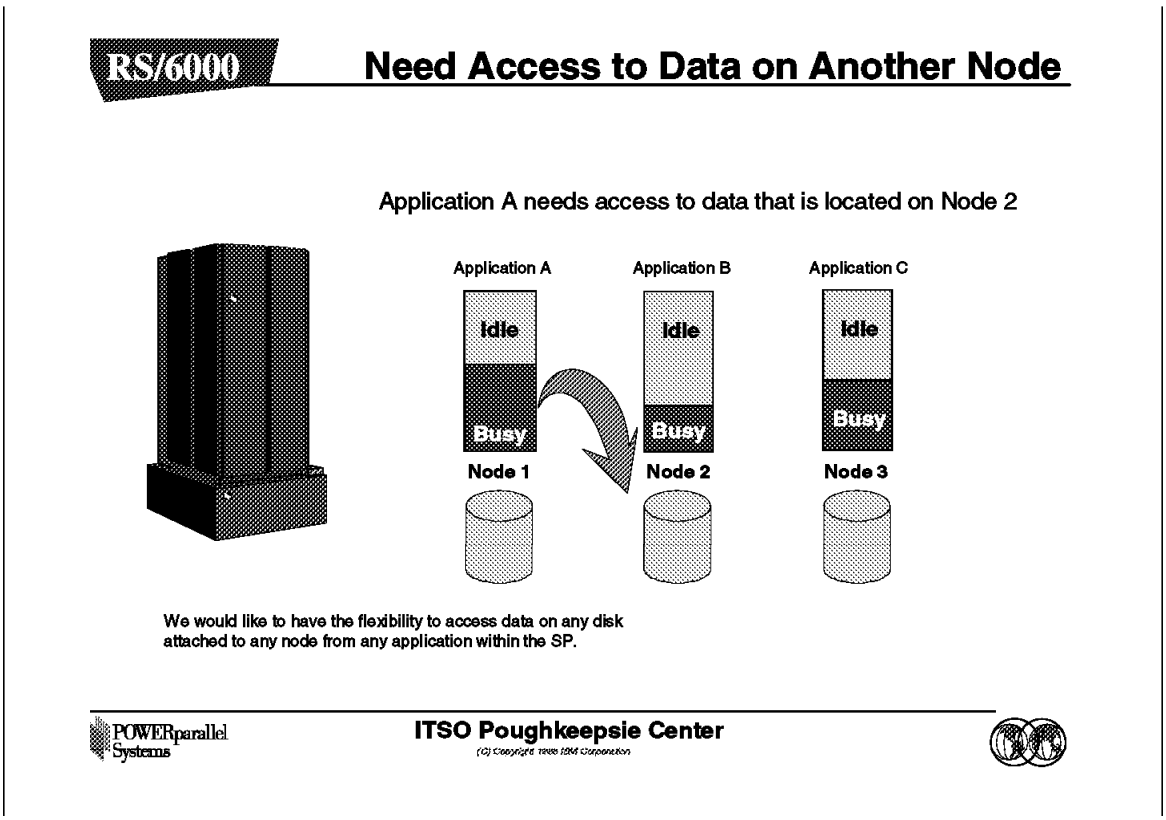
5.1.1 I/O Performance Can Be a Bottleneck



In this example, our application results in a lot of heavy I/O to Node 3. As a result, Node 3 becomes overloaded.

We would like to be able to spread this I/O activity over other nodes. We can already spread the I/O over multiple disks on one node, but the requirement is to also spread the I/O activity over other nodes. This leads to the requirement to have access to the disks on other nodes within the SP. This is only possible *directly* today with Virtual Shared Disk (VSD), but VSD does not support general I/O activity. In particular, VSD does not support file systems. A solution based on Network File System (NFS) will always point to a single node as the source for that data and in addition, in many environments, NFS is not a high-performance solution.

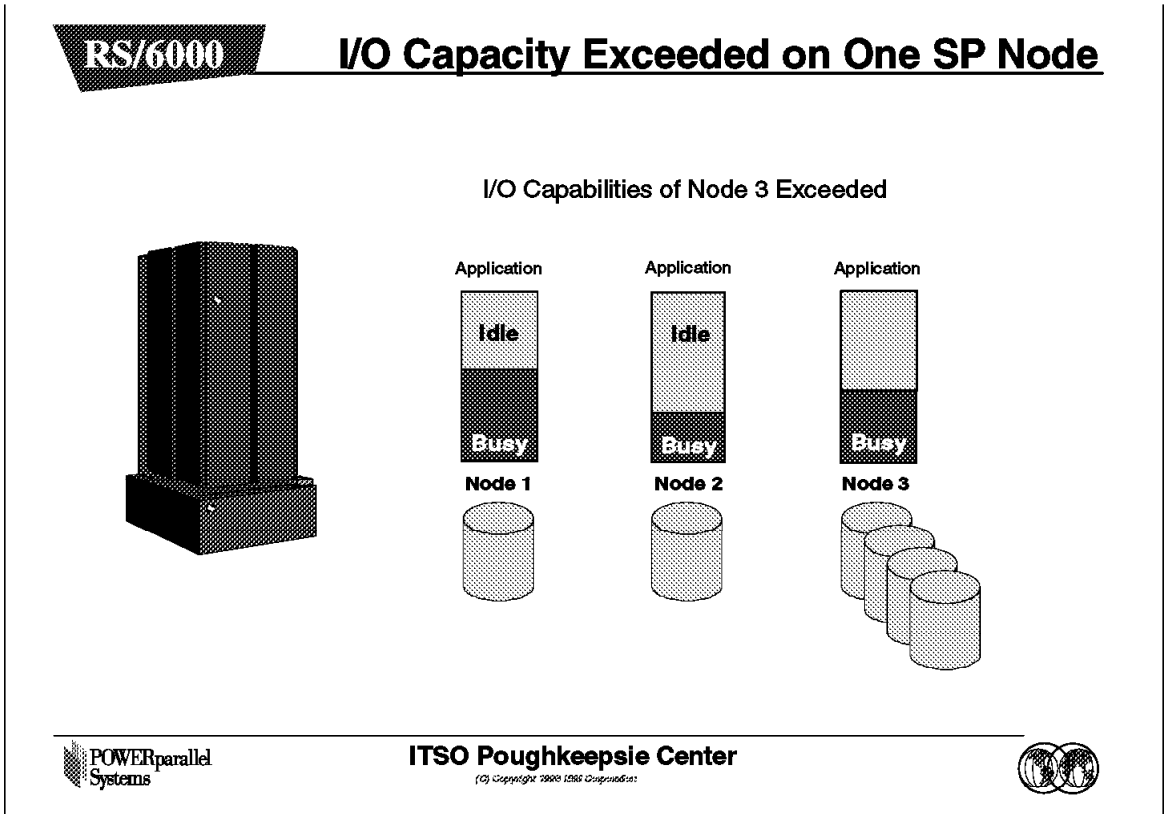
5.1.2 Need Access to Data on Other Nodes



From a pure flexibility point of view, it is often very helpful to have access to data that resides on other nodes within the SP so that an application can run on any node within the SP, yet have access to data that it needs.

Once again, NFS could provide such a solution, but its performance is not adequate in many cases. In particular, the write function of NFS can be slow.

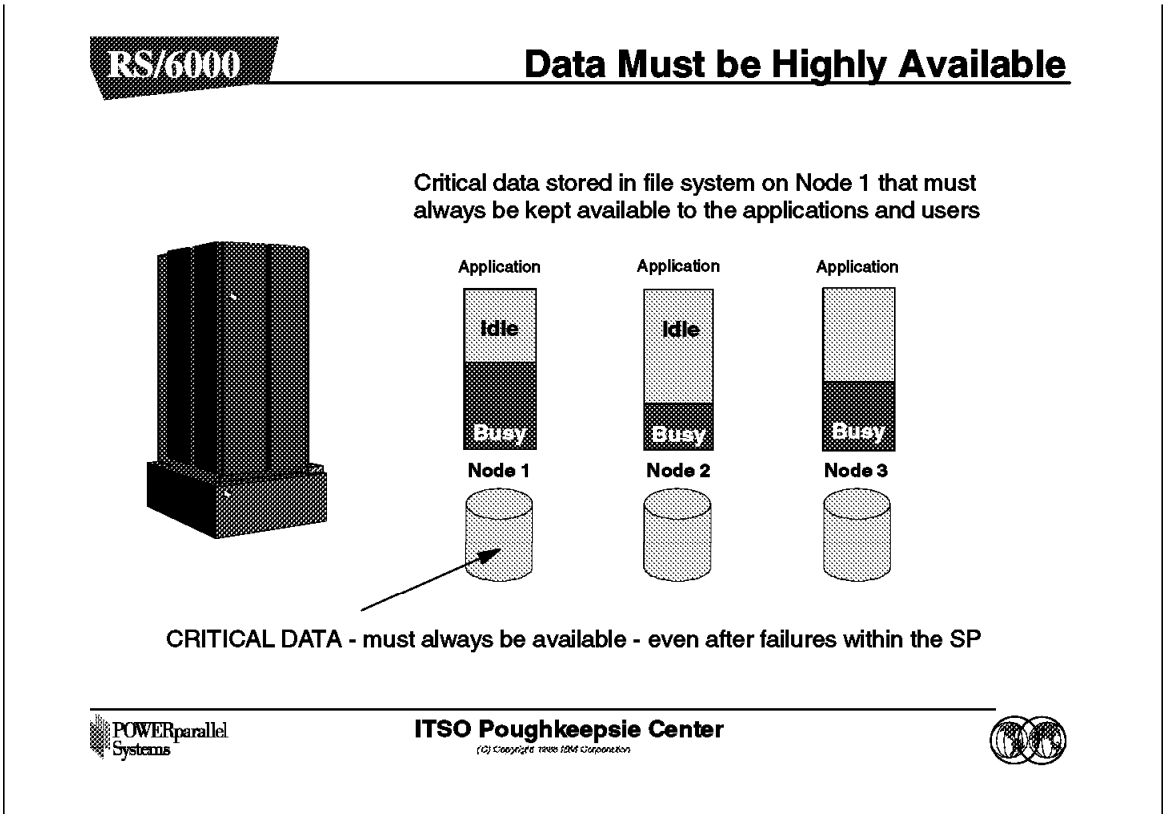
5.1.3 I/O Capacity Exceeded on One SP Node



In this example, the capacity of node 3 has been exceeded from an I/O point of view. For example, adapter slots may have been filled and there is no room for more adapters to connect additional disks.

We would like the ability to access disks that are attached to another node in the SP to take advantage of spare capacity elsewhere.

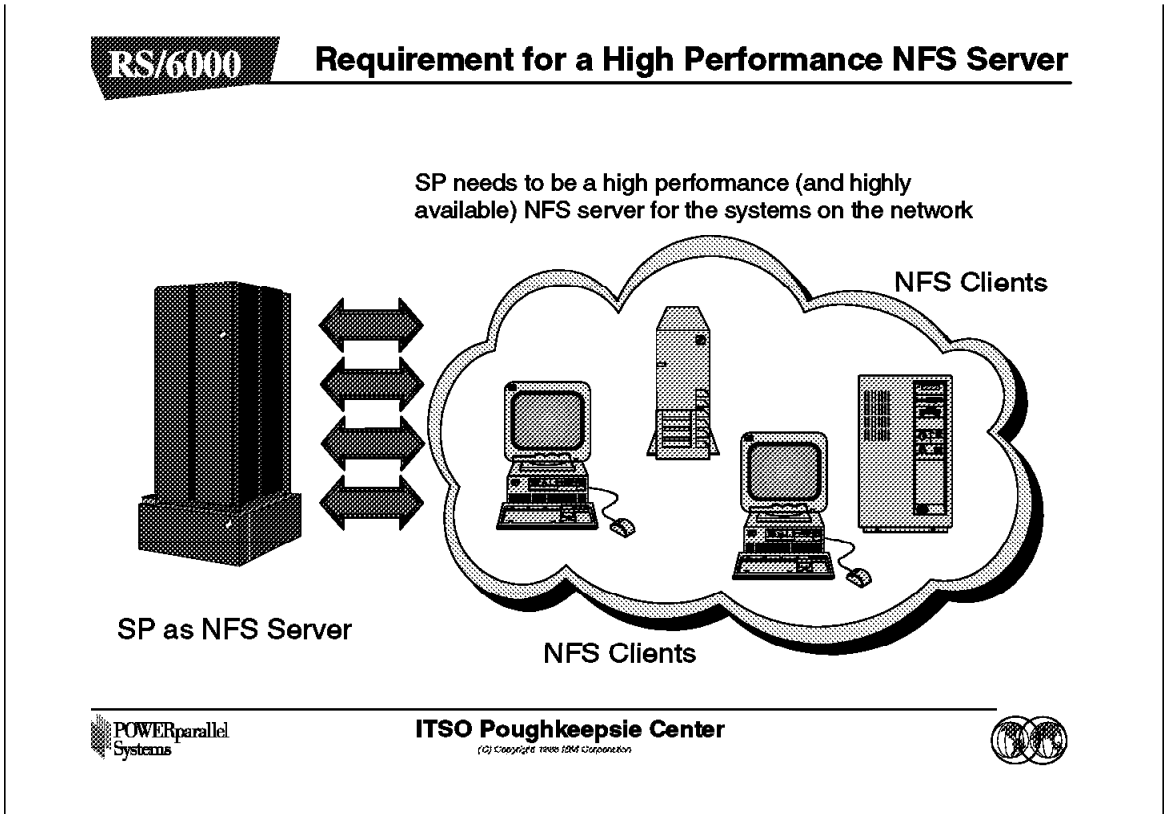
5.1.4 Data Must Be Highly Available



A common requirement is to have critical data on a system that is kept highly available. Solutions do exist to provide this today with NFS, but they have limitations.

As we will see, the GPFS solution is an ideal option in such circumstances.

5.1.5 Requirement for a High Performance NFS Server

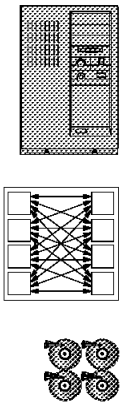


Some customers already have a network of systems that require access to an NFS server or servers. The requirements for such servers, if they are providing data access to a large number of servers, are usually that they must provide high performance, and often high availability in addition.

5.1.6 Trends

RS/6000

Trends



	write	read	Transport	Aggregate
NFS	800KB/s 5MB (V3)	10MB/s	UDP	N/A
PIOFS	8MB/s	12MB/s	IP	Scalable
GPFS	>25MB/s	>25MB/s	IP (VSD)	Scalable

POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright IBM Corporation

The trends in terms of file systems are shown here. Each of these solutions has advantages and disadvantages. The performance estimates are indicative of performance that might be obtained under optimal circumstances and with an optimal configuration.

Depending on the application, performance may well be very different from that shown here.

5.2 What is GPFS - An Overview

RS/6000

What Is GPFS - An Overview

- Provides file system services to parallel and serial applications across the SP
- Allows shared access to file systems and files that may span multiple disks across multiple nodes of the SP
- Similar to the functionality provided by Network File System (NFS) but not a distributed file system and only supported within the SP System - also typically much faster than NFS
- Exploits the IBM Virtual Shared Disk (VSD) subsystem as an underlying technology

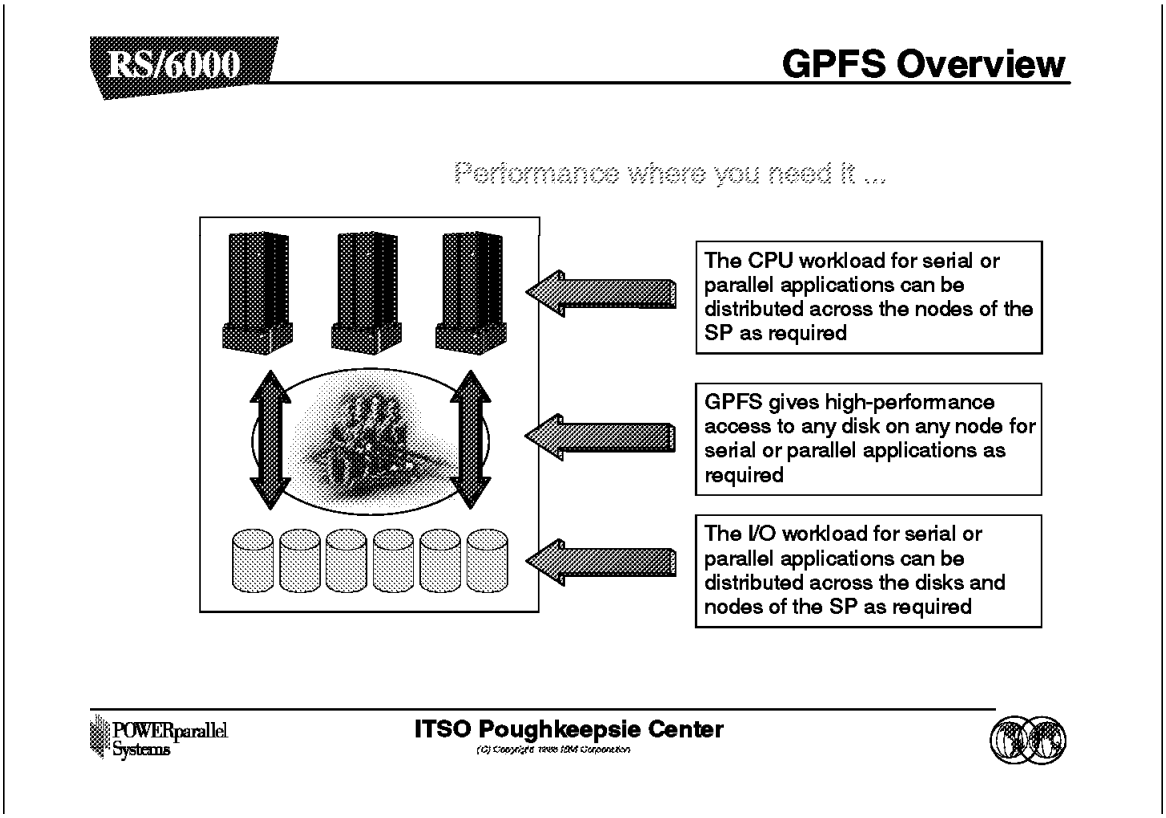
POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

GPFS is a software product that is available on the SP. The software provides the functionality as shown, but it should be remembered that GPFS is only supported on the SP system. Systems external to the SP cannot run the GPFS software.

5.2.1 GPFS Overview



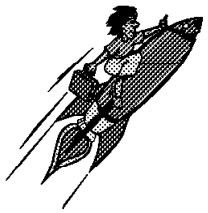
GPFS provides a truly parallel I/O solution for use within the SP. It can be exploited by both serial and parallel applications. It will allow for better balanced performance in many application environments, and will allow an application to exploit a number of nodes for I/O performance, as well as for CPU performance.

5.2.2 GPFS Improves Performance

RS/6000

GPFS Improves Performance

- GPFS allows multiple processes or applications on many nodes of the SP to simultaneously access the same file using standard file system calls
- Increases aggregate bandwidth by spreading reads and writes across multiple disks
- Balances the load evenly across all disks to maximize their total throughput
- Utilizes the SP Switch for fast performance



POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

The GPFS software provides fast access across the SP Switch to disks attached to remote nodes within the SP. This results in extended flexibility and better performance.

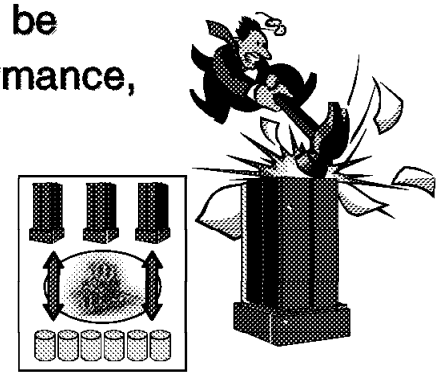
Concurrent access can be supported with the application providing locking in much the same way that NFS provides locking.

5.2.3 GPFS Improves Data Availability

RS/6000

GPFS Improves Data Availability

- GPFS can be used to provide highly available file systems on the SP that can remain accessible even in the event of failures within the SP
- The data can be accessed from systems outside the SP using NFS, so GPFS can be used to provide a high-performance, highly available NFS server



POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1996 IBM Corporation

As we will see, GPFS uses the VSD technology to access remote disks. An extension for VSD is Recoverable Virtual Shared Disk (RVSD). This can provide disk takeover in the event of failure.

GPFS works closely with RVSD and a highly available solution can be provided.

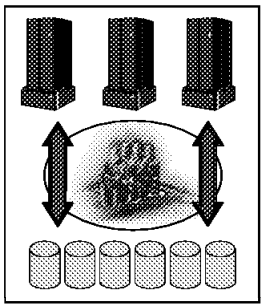
There are a number of design options in this area that are discussed in this chapter. In addition to the usual options of disk mirroring or RAID disk subsystems, GPFS allows for RVSD and replication of data within GPFS.

5.2.4 GPFS Supports Standards

RS/6000

GPFS Supports Standards

- GPFS supports the relevant X/Open standards with minor exceptions:
  - mmap, atime/mtime/ctime
- GPFS allows for coexistence with NFS, to allow NFS access to systems outside the SP



POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1996 IBM Corporation

In most cases, GPFS supports the relevant standards for file systems as defined by X/Open. There are some minor exceptions that are described later.

In many respects, a GPFS file system running within the SP will be seen as a normal file system; it will not normally be obvious that this is a GPFS file system.

Additional commands are delivered with GPFS for supporting GPFS file systems, but many standard AIX file system commands, such as `mount` or `df`, will work as expected.

A GPFS file system, once created, can be exported like any AIX file system, and can therefore be mounted on client systems either within the SP, or outside the SP, using normal NFS commands.

5.2.5 When Can GPFS Be Used?

RS/6000

When Can GPFS be Used?

- GPFS can be used for almost all applications, serial or parallel
- Applications that use NFS will probably be good contenders for using GPFS
- For exceptions, see later
- GPFS is only supported on the IBM SP
- GPFS exploits the High Performance Switch and the SP Switch and cannot be used with other networks
- It can coexist with NFS

POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright IBM Corporation

GPFS can be used in most cases. There are a few cases where it may not provide additional advantages over other solutions such as NFS, but in most cases, SP customers are likely to want to implement and exploit GPFS.

It is applicable for customers running either commercial applications, or scientific and technical applications.

GPFS is only supported on the IBM SP. In particular, it requires a fast network, namely the SP Switch, and uses security facilities within the SP for node-to-node communications.

RS/6000

When Can GPFS Be Used (cont.)

- It will work best with sequential access to large files that are stored with a large block size
- It will also provide better performance than NFS for most types of file access - except when the application continuously opens a file, reads or writes a few bytes and closes it again
- It can be used instead of PIOFS - but does not provide "views"
- It can be used whenever high availability is needed

POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

GPFS can be used in many cases, but the best performance will be seen under certain circumstances.

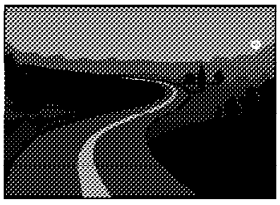
GPFS may be viewed as similar to PIOFS, but there are differences. A comparison between these two solutions is shown later in this chapter. In most aspects, however, GPFS is superior to PIOFS.

5.2.6 Where Does GPFS Come From?

RS/6000

Where Does GPFS Come From?

- GPFS originates from the Almaden research project called "Tiger Shark" and is aimed at providing a high-performance file system for multimedia data
- The technology has so far emerged in three IBM products:
  - Multimedia LAN Server - a server for video and audio data on a LAN with real-time processing
  - Video Charger - videos across a LAN - comes in a Web Server product
  - GPFS
- You will see "mm" appear in GPFS commands as a result of its multimedia ancestry



POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

GPFS originates from the Almaden Research Center. Because of its history as a multimedia file system (mmfs), many of the commands start with the letters mm. This technology has been used in three products to date. GPFS is one of them. On the SP today, only GPFS is supported.

Even if the other products were supported, it is not technically possible to run more than one of these products on the same system.

5.2.7 How Does GPFS Work?

RS/6000

How Does GPFS Work?

- GPFS uses Virtual Shared Disk (VSD) as its underlying structure
- VSD has been part of PSSP for some time, but has only really been exploited by Oracle Parallel Server
- GPFS also requires Recoverable VSD, which is a separate LPP (not part of PSSP)
- GPFS depends on the RS/6000 CT technology and in particular on Group Services, which is part of PSSP

POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

For access to remote disks, GPFS uses the tried and tested Virtual Shared Disk (VSD) that has been part of PSSP for a long time.

VSD is explained in more detail in 5.3, “VSD/RVSD” on page 80.

GPFS is only supported with PSSP 2.4 and VSD Version 2.4.

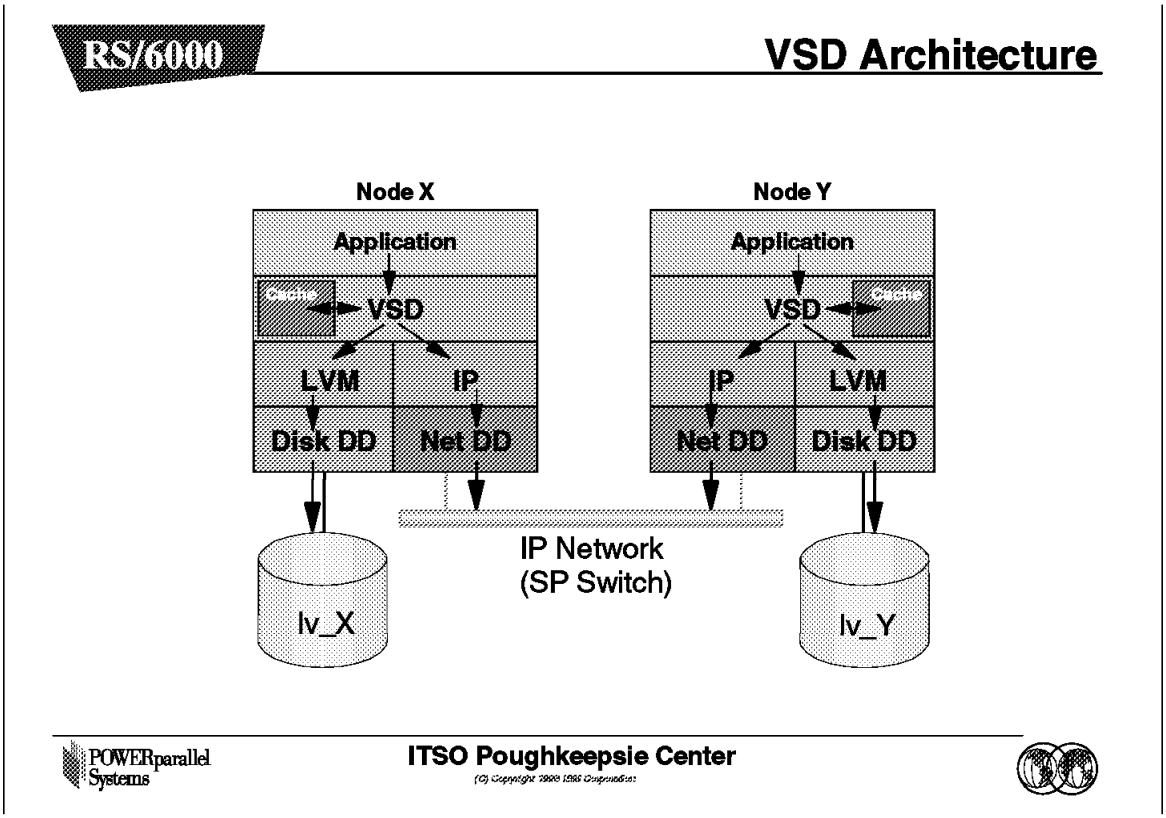
Similarly, GPFS requires RVSD Version 2.1.1. It is required even if twin-tailing of disks is not required because GPFS uses the new node fencing capability that requires RVSD.

GPFS uses the RS/6000 Cluster Technology (RSCT) to synchronize all the actions between the multiple daemons running in different nodes. It especially depends on Group Services (HAGS) for this synchronization.

5.3 VSD/RVSD

The Virtual Shared Disk (VSD) and Recoverable Virtual Shared Disk (RVSD) provide the basic plumbing for remote data access.

5.3.1 VSD Architecture



This diagram shows the structure that allows VSD to gain access to remote disks within the SP system.

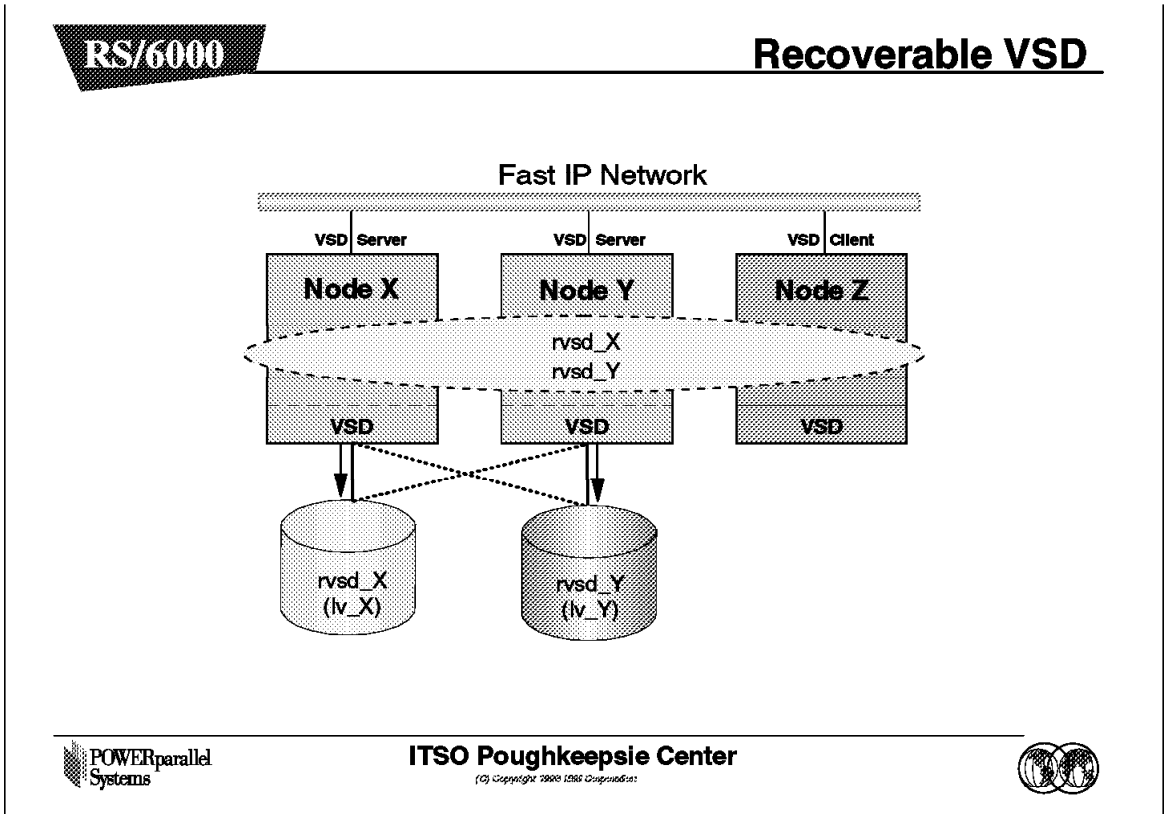
Each disk in an SP is physically cabled to only one node and can only be accessed directly by that node.

VSD allows access across the SP switch to this disk from a remote node. This is achieved by the application communicating with the VSD device driver rather than a disk device driver. The VSD device driver can reroute the I/O request to a remote node if required. Local disk activity occurs in the usual way after going through the VSD device driver.

The VSD software only gives access. It does not provide a locking mechanism to ensure integrity of the data. In addition, a VSD defines only a logical volume, or raw device, and not a file system.

An application such as Oracle Parallel Server is required to provide a global locking mechanism.

5.3.2 Recoverable VSD (RVSD)

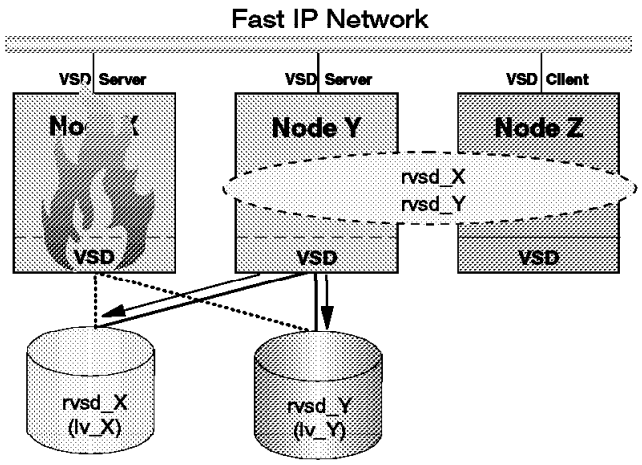


An optional addition that can be used with VSD is Recoverable VSD (RVSD). RVSD is a separate Licensed Program Product (LPP). It is not part of PSSP. It is normally used in conjunction with twin-tailed disks to give high availability for a volume group and the associated VSDs. It is required for GPFS.

This diagram shows the normal operation of a group of nodes with VSD. RVSD is being used to protect the VSD server Node X. In the event of failure of Node X, Node Y will act as the secondary server and take over the volume group.

RS/6000

Recoverable VSD (cont.)



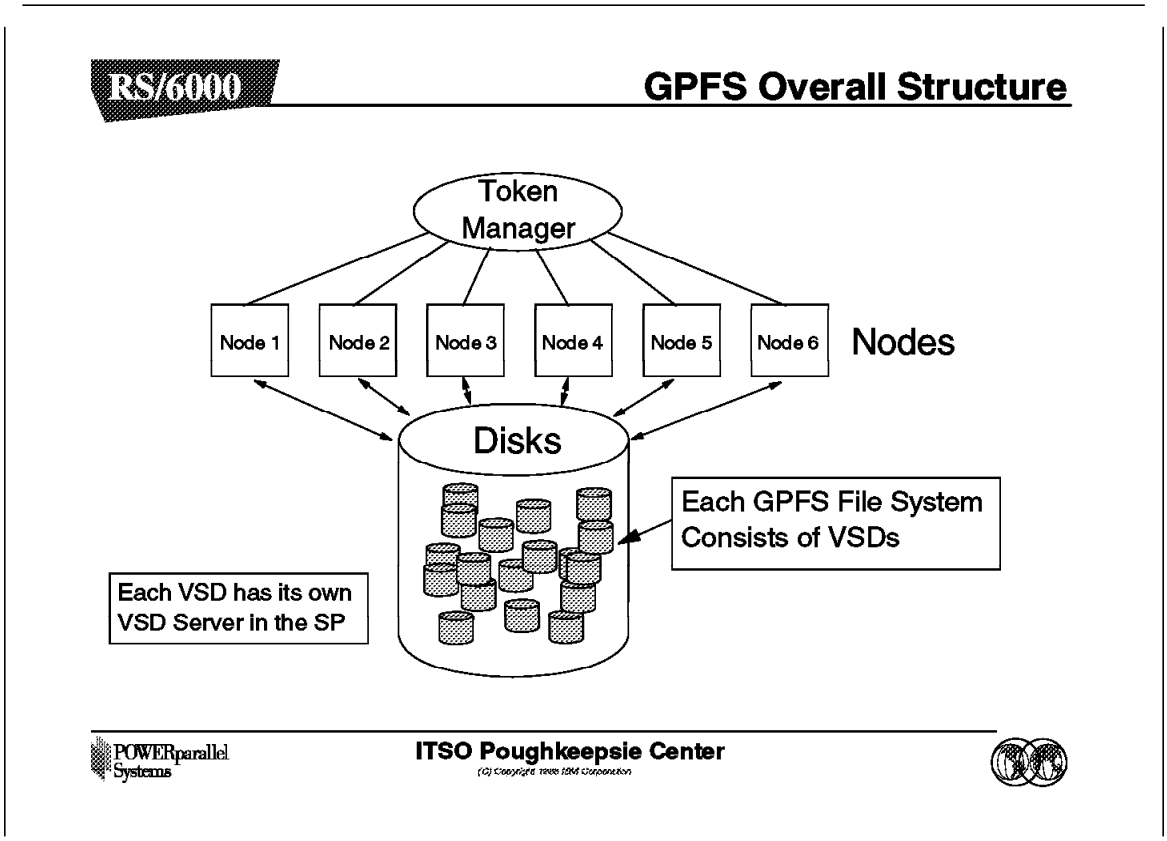
POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

In this diagram, we see that Node X has failed. Node Y has taken ownership of the volume group that is twin-tailed and will vary on the volume group and make the VSDs available. In addition, RVSD will communicate with Group Services to inform other applications, such as Oracle, of the failure and recovery.

5.4 GPFS Overall Structure



This diagram shows the overall structure of GPFS. Each GPFS file system is made up of a number of disks defined as VSDs. Each VSD can reside on any server node within the SP system. The file system writes data by striping across all of these VSDs.

In addition, any GPFS node can mount this file system and have access to the same data.

A Token Manager controls the locking within the GPFS file system.

5.4.1 GPFS Functioning

RS/6000

GPFS Functioning

- GPFS is a "physical" file system rather than a distributed file system (such as NFS, DFS, AFS , and others)
- GPFS is the "equivalent" of a Journaled File System (JFS) in many ways
- GPFS provides a Token Manager and a Lock Manager to provide applications with the mechanism for data integrity

POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright IBM Corporation

GPFS may look similar to NFS in some ways, but is actually very different. It does not have a single server and lots of clients, but instead is a parallel file system that is normally striped across a number of nodes.

As we shall see, it has a number of facilities to provide for high availability. While it is not a journaled file system (jfs) in AIX terminology, it provides the same functions for recovery through a different method.

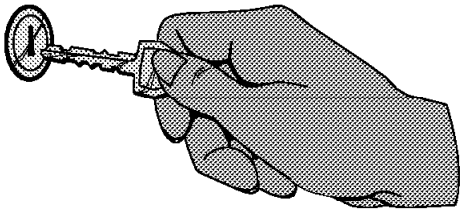
It provides a locking mechanism for applications to prevent data corruption.

5.4.2 GPFS Locking

RS/6000

GPFS Locking

- GPFS uses locking within a node to grant permission to applications to read/write to a file
- GPFS uses tokens across nodes to give permission to grant a lock



POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

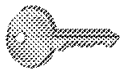
GPFS provides the locking mechanism that VSD does not have.

This allows GPFS file systems to be seen in almost every way as normal file systems.

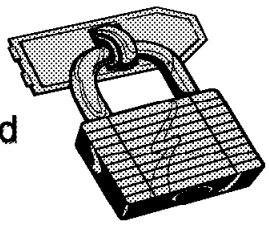
RS/6000

GPFS Locking (cont.)

- Locking controls access to files on a node
- Two copies of tokens are maintained in memory for recovery purposes
- Token Manager controls file access between nodes



- The Lock Manager asks the Token Manager for a lock...
- Locks can be a block or a file
- Lock is either granted immediately or a copy set is sent
- Nodes can negotiate for lock if it is denied
- Read locks can coexist
- Write locks are serial
- Locks are advisory locks (refer to external-to-GPFS locking)



POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

Locking controls access to files from a process point of view on any particular node within the GPFS pool. Two copies of the locks are maintained within the GPFS system to allow for recovery in the event of failure. A Lock Manager runs on each node for the purpose of controlling locking on that node.

Each of the two copies of the token are kept in memory but in separate *Failure Groups* within the SP. Failure Groups are discussed in 5.9.1, “Failure Group” on page 112_, but for the moment we can assume that these copies of locks are kept on separate GPFS nodes.

If the Lock Manager on a particular node is asked for access to a file that exists elsewhere within the SP, then a token needs to be requested from the other node that has the file. This is achieved through the use of tokens. There is one Token Manager for each GPFS file system and this Token Manager (or Stripe Group Manager) runs on one of the nodes within the GPFS pool of nodes.

If other nodes already have the token when it is requested, then the list of nodes in question is passed back to the requesting Lock Manager. This list is called a *copy set*.

It is the responsibility of the requesting Lock Manager to negotiate with nodes in the list to obtain the token.

Under different circumstances, the locking mechanism within GPFS locks different ranges of data.

The request asks for a required range, and also the desired range of data. Depending on who else has locks on sections of the file, a lock may be granted for a larger section of the file. This would improve performance, for example, in the case of a sequential read of a whole file.

There are eight stages to granting a lock, and these cater for contention as well as recovery in the event of failure.

The locks can be read locks or write locks. Read locks are required, for example, so that an application can know whether it can delete a file. It may well be that a file that is being read cannot be deleted.

Multiple concurrent read locks can be granted, whereas write locks are sequential.

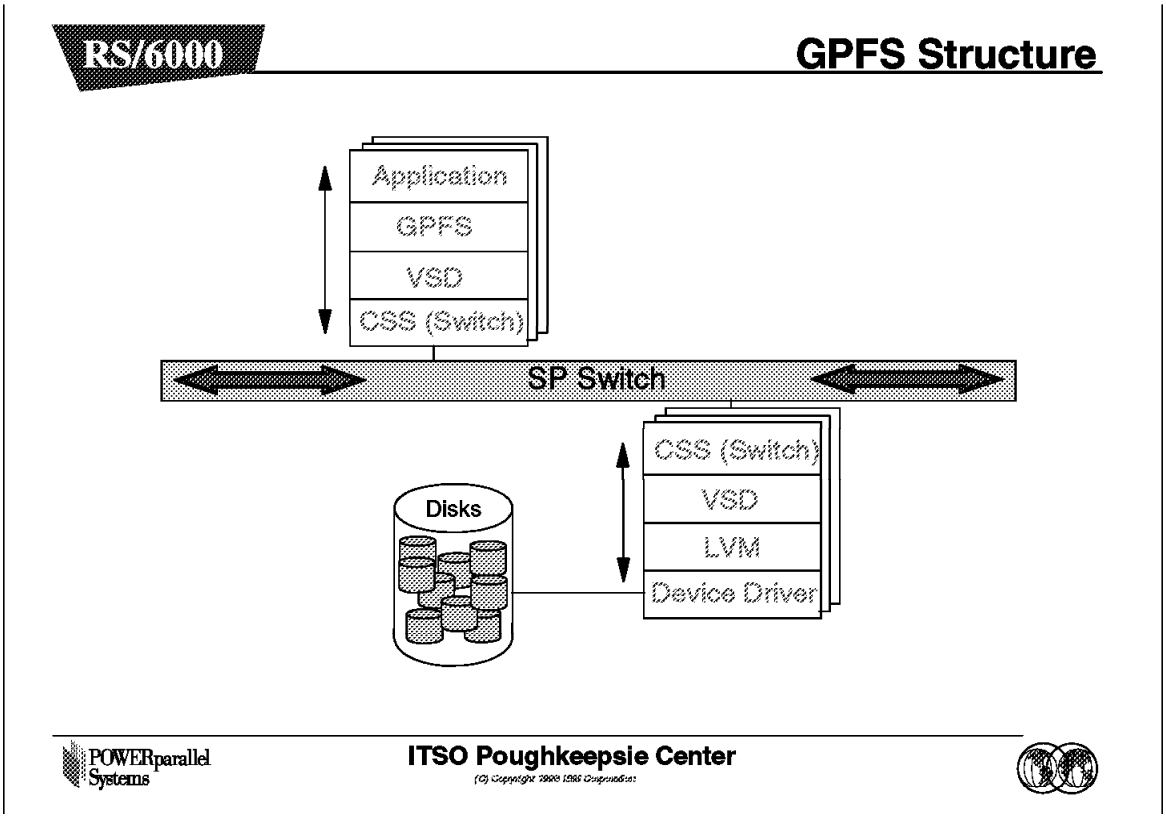
In the event of failure of a node, its logs are replayed before any locks are released to ensure integrity of the data.

In the event of a Stripe Group Manager failure, all the tokens that exist on other nodes can be retrieved to enable the new Stripe Group Manager to have up-to-date information.

Note: This discussion refers to the internal locking mechanisms within GPFS and not to the application locking, such as lockf calls, which are external to GPFS.

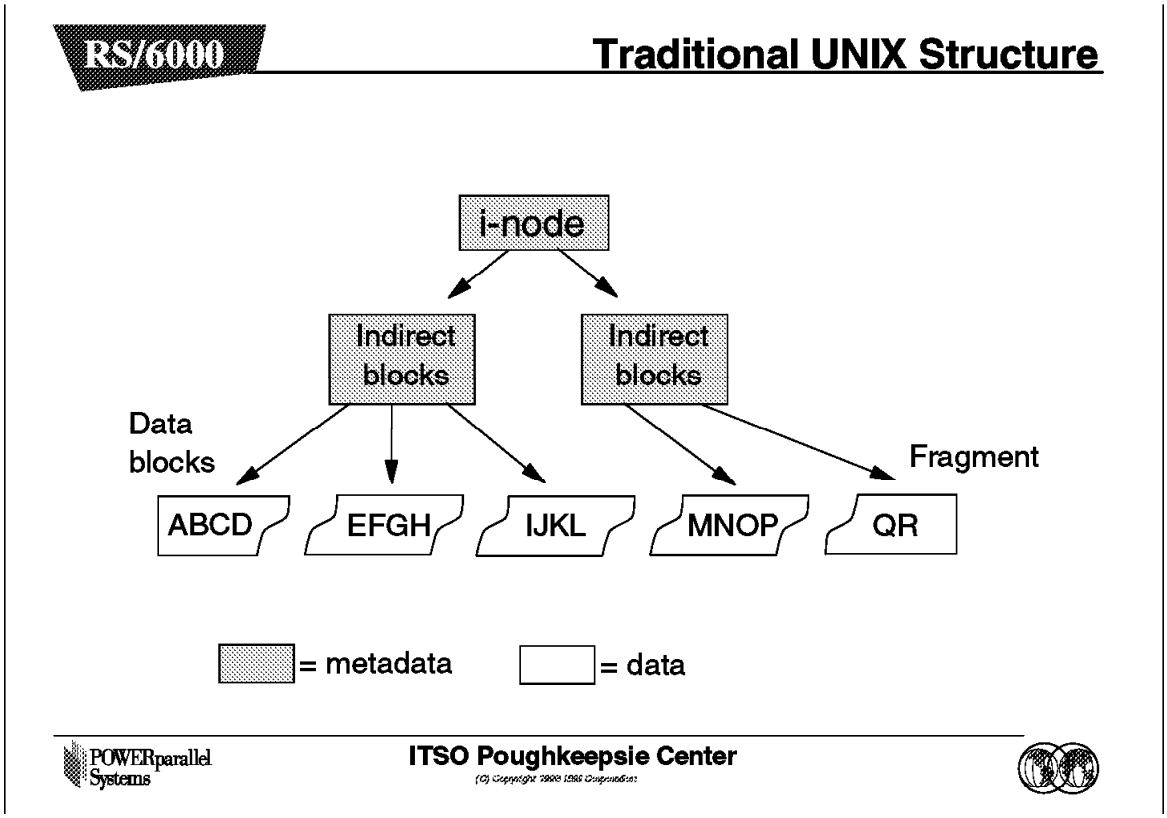
The statement that locks are advisory refers to the *external* locking mechanism.

5.4.3 GPFS Structure



GPFS uses VSD to provide local or remote access to disks, as shown.

5.4.4 Traditional UNIX Structure



GPFS uses the traditional way of structuring file systems, as adopted in the UNIX world.

An *i-node* is a pointer to the actual data on disk. When the amount of data is larger, the i-node will in fact point to another data structure, called an *indirect block*, that in turn will point to the actual addresses of the data on disk.

5.4.5 Quorum

RS/6000

Quorum

- GPFS uses a Quorum concept in the same way as RVSD
- More than 50% of the GPFS nodes must be up for Quorum to be satisfied (50% + 1 node)
- If Quorum fails, GPFS file systems will be unmounted and started again when Quorum is reached
- This protects the GPFS file systems in the event of a failure and subsequent recovery

POWERparallel Systems

ITSO Poughkeepsie Center

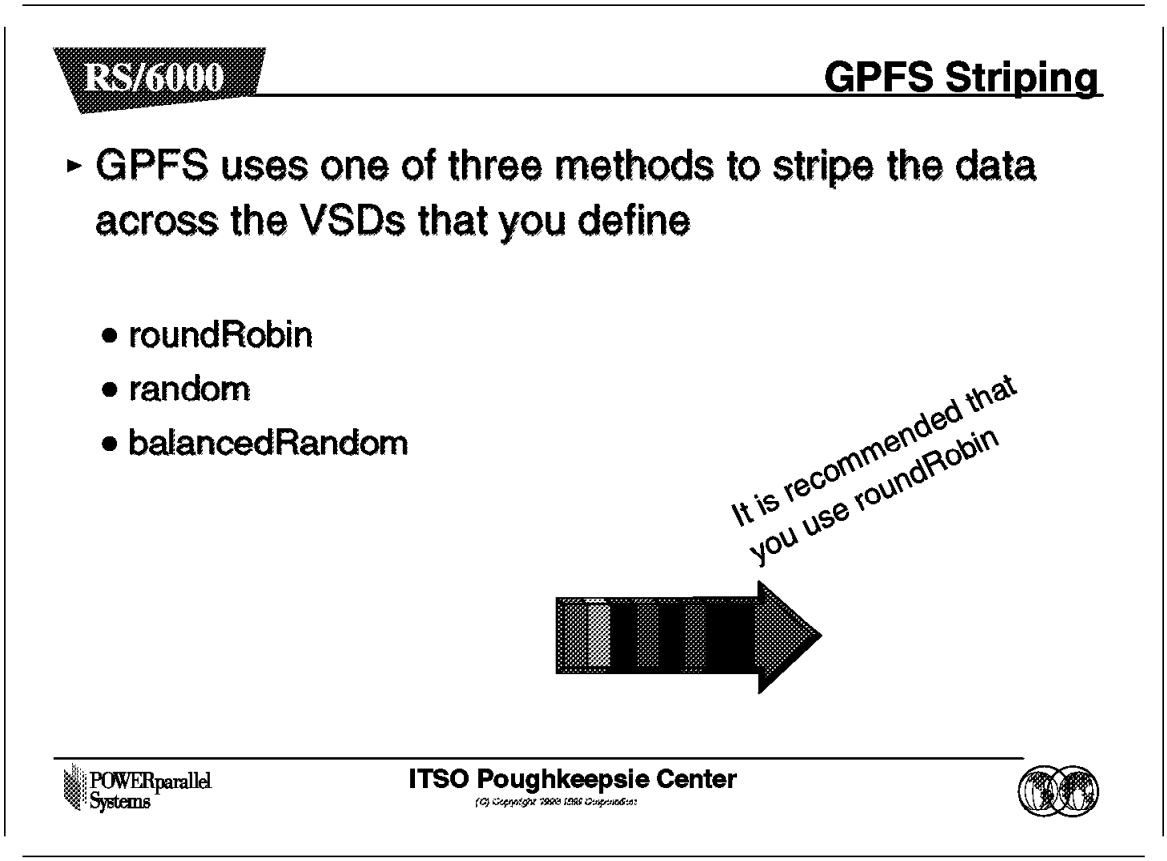
(C) Copyright 1996 IBM Corporation

Quorum is used by GPFS to make sure that no unexpected actions occur during a failure within the SP system.

Quorum ensures that the GPFS group of nodes does not split into two separate groups following a failure of some kind. If there were two separate groups of nodes within the GPFS pool of nodes, there would be two Token Managers controlling file locking. This could have disastrous consequences, with data getting corrupted.

To ensure that this cannot happen and that recovery is carefully controlled and guaranteed, Quorum will not allow GPFS to use a file system if less than half of the nodes are not operational.

5.5 GPFS Striping

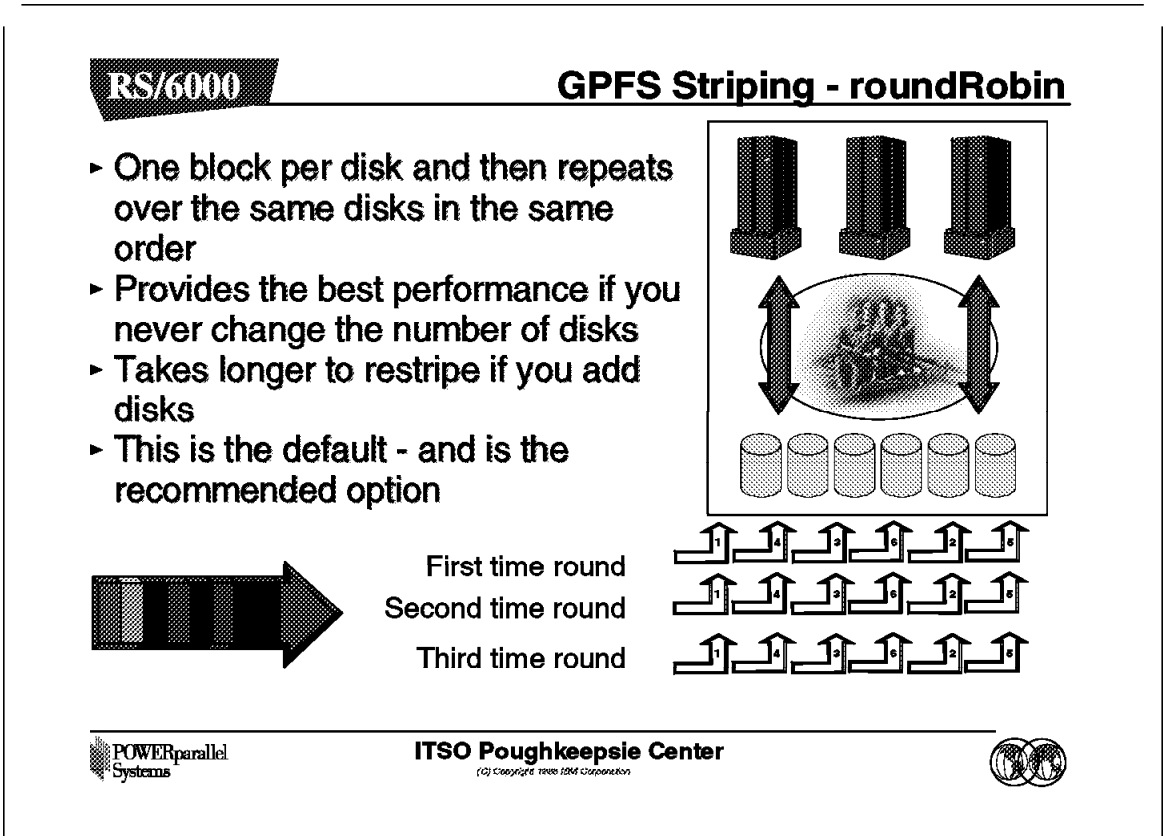


A number of options for GPFS striping exist but the default method, roundRobin, should be the method that you normally use. This is the default selected by GPFS.

It will also be an advantage in most cases, as we will see later, that we use the same number of disks on each node in the GPFS file system and that each of those disks has the same performance and capacity characteristics.

Striping will not use a disk that is offline while the file system is created.

5.5.1 roundRobin Striping

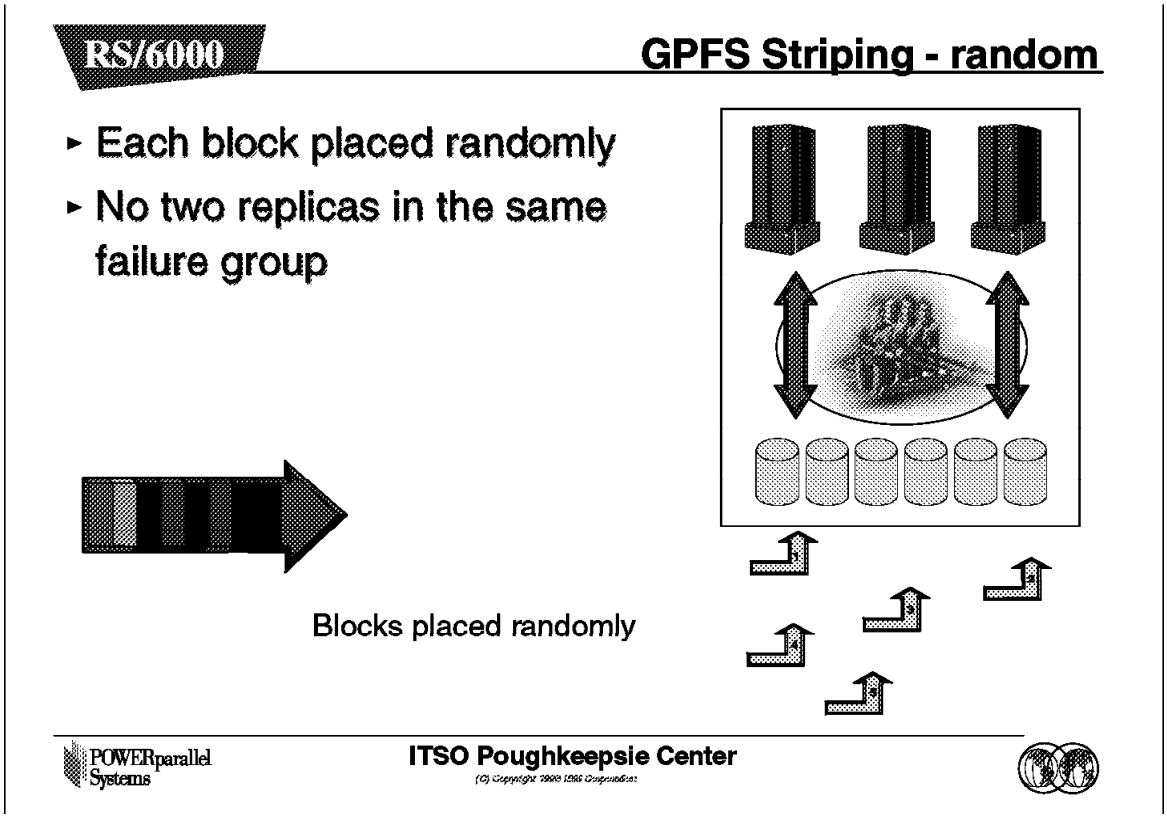


In the case of roundRobin striping, data is written in turn to each disk until all disks have received a block. Then another round of the disks begins, again writing a block each time, and accessing the disks in the same order.

This is the preferred method of striping.

It is clear that equal capacity disks are best in this case.

5.5.2 random Striping



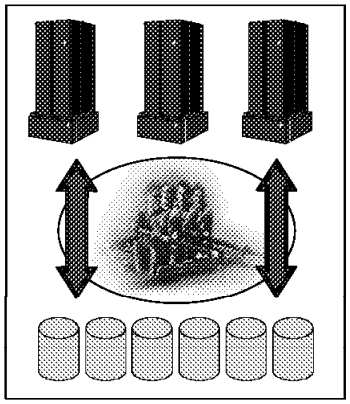
In the case of random striping, data blocks are placed randomly across the disks in the GPFS nodes. If failure groups are defined, the replicated data is stored in separate failure groups to cater for failures.

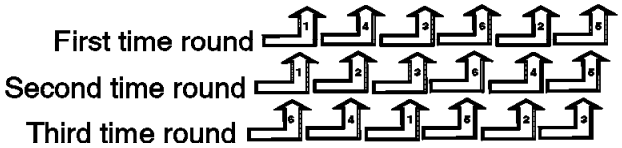
5.5.3 balancedRandom Striping

RS/6000

GPFS Striping - balancedRandom

- One block per disk and then repeats over the same disks, but in any order
- Does not move on to reusing the same disks until one block on each disk
- No two replicas in the same failure group





POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

The balancedRandom striping method writes blocks randomly, but does not return to the same disk until all available disks have been used.

RS/6000

Working with RS/6000 CT

- GPFS requires Group Services
- GPFS can handle multiple failures through Group Services
- One main subsystem: mmfs
- mmfs is the normal recovery daemon
- mmfsrec restarts the recovery of mmfs in the event of a second failure during recovery
- Group Services guarantees order of messages

POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1999 IBM Corporation

Both the mmfs and mmfsrec groups are registered with Group Services. The mmfs subsystem is the normal recovery mechanism. In the event of a second failure during the recovery, mmfsrec is used to ensure correct recovery.

To monitor the attributes of the hags group, you can use the command `lssrc -l -s hags`.

During recovery by mmfs, there is a four-phase process where votes take place to ensure that all nodes are recovering correctly and are in step.

In the event of a second failure, the voting cannot proceed, and in that case, mmfsrec steps in to take action based on the second failure. Recovery can continue with the new information about the second failure now available.

5.6.1 Recovery

RS/6000

Recovery

- If a node fails, mmfs attempts to recover through the use of logs
- mmfs recovers anything that is locked to a consistent state before releasing the lock to another node
- Three phases - cannot move to the next phase until successful completion on each node
- SDR knows the configuration of GPFS: file stored on the CWS

POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1996 IBM Corporation

Whenever a node boots, it checks the two GPFS files that are stored in the SDR to see if configuration changes were applied to the GPFS configuration while the node was down. If this is the case, the changes will be applied to the node at that time.

This allows configuration changes to be made while not all nodes are available. For example, new nodes or disks could be added to the GPFS configuration while some nodes are powered off.

The files in the SDR that are created by GPFS are as follows:

/spdata/sys1/sdr/partitions/9.180.40.16/files/mmsdrcfg1

/spdata/sys1/sdr/partitions/9.180.40.16/files/mmsdrfs

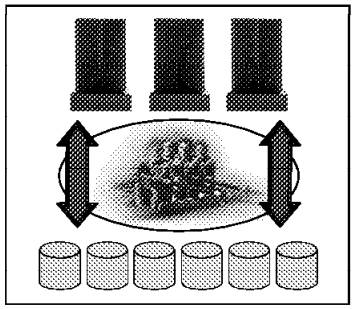
Note: These files should not be edited. They should be backed up when you back up your SDR, and can be used in the event of failure to recover.

5.7 Configuration

RS/6000

Configuration Considerations - Nodes

- Planning Nodes
  - Estimating Node Count
  - Creating a Node List
  - Quorums



- Planning Nodes - recommendations
  - Best to overestimate the number of nodes
  - Prepare a file - do not include the CWS
  - GPFS requires a quorum of nodes to be available (1/2 of the nodes plus 1)

POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

As you create your GPFS system, you will be asked to supply information about which nodes are in the GPFS pool.

As you create an individual GPFS file system, you need to supply other information about nodes, such as how many nodes will participate in this GPFS file system. The default is 32, and you should not underestimate this number. Make sure that you include at least as many nodes as the maximum that you expect in the GPFS pool.

This information is used to create *allocation regions* that will affect the performance of the file system.

5.7.1 Node Count

RS/6000

Node Count

- The number of nodes that you specify initially for your GPFS file system will affect the number of "regions" that are created in the file system data structure
- If you subsequently add more nodes than this number, you will not obtain optimum performance
- Therefore it is best to overestimate this number to some extent
- Do not go overboard on this - or you will waste resources such as memory
- The default is 32 nodes
- You cannot change this value later

POWERparallel Systems

ITSO Poughkeepsie Center

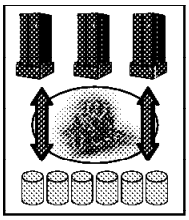
(C) Copyright 1998 IBM Corporation

When determining how many nodes you will specify for your GPFS file system, it is important to remember that this value cannot be changed later.

5.7.2 File System Considerations

RS/6000 **Consideration Configurations - File Systems**

- File System Creation Considerations
  - Choices you make here will have an impact on the maximum file size that you can store in this GPFS file system



  - Block Size
    - 16KB, 64KB or 256KB (default is 256KB)
    - Fragments and Sub-Blocks
  - I-nodes
    - Maximum is 4KB (default is 512 bytes)
  - Indirect Blocks
    - Maximum is 32KB (default is blocksize/16)
  - Replication also affects maximum file size

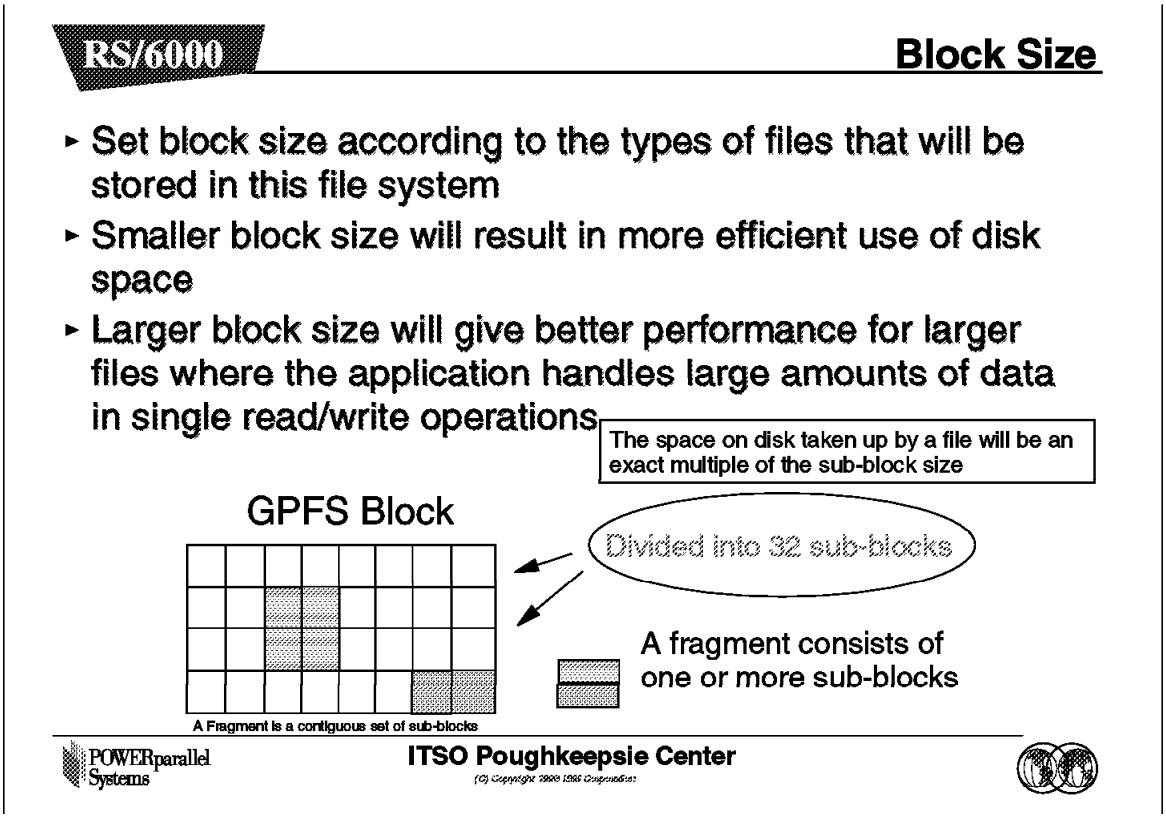
POWERparallel Systems **ITSO Poughkeepsie Center**

(C) Copyright 1996 IBM Corporation

This is a critical area where you will need to make decisions that cannot be changed. Decisions you make will affect the size of the largest file that you can store in this particular GPFS file system.

This will be discussed in more detail in the following foils.

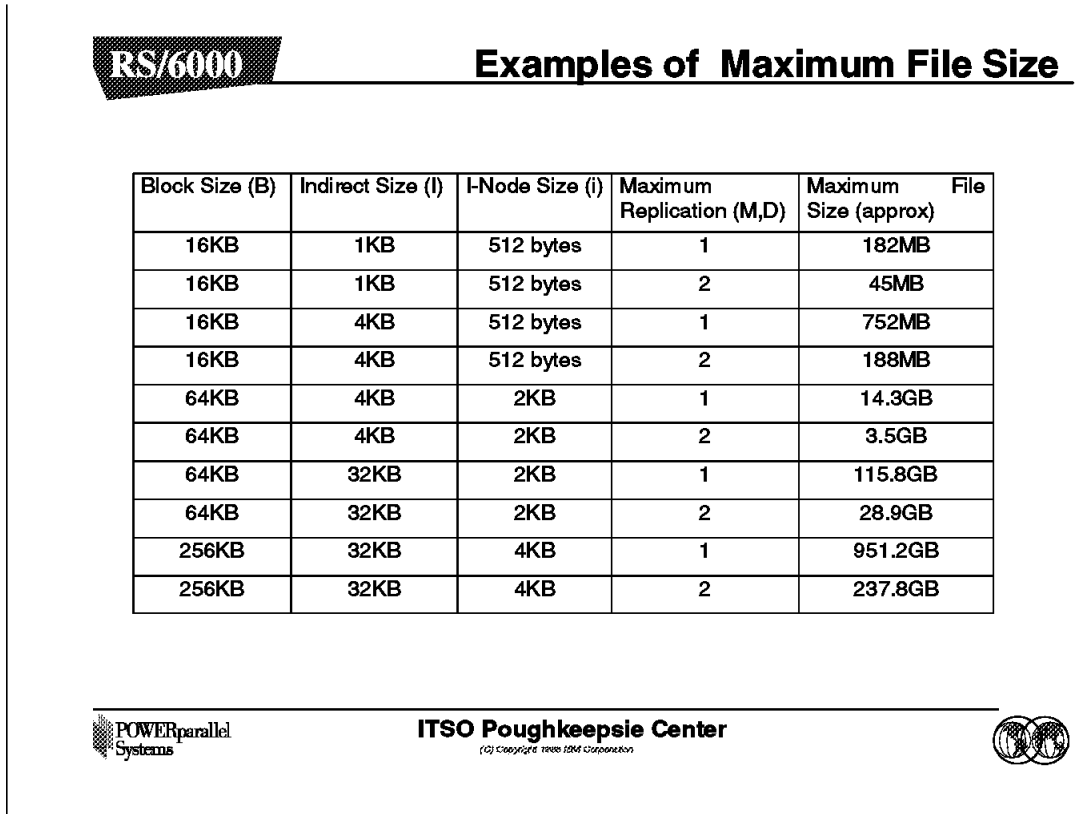
5.7.3 Block Size



You will select the block size for your GPFS file system. This defines the size of the blocks that are written as you stripe over each of the disks in the Stripe Group.

This also defines some other parameters as shown here. The sub-block size results in the smallest area of disk that will allocated to a file. For small files, this has an impact on the space wasted by unused areas of disk.

5.7.4 Examples of Maximum File Size

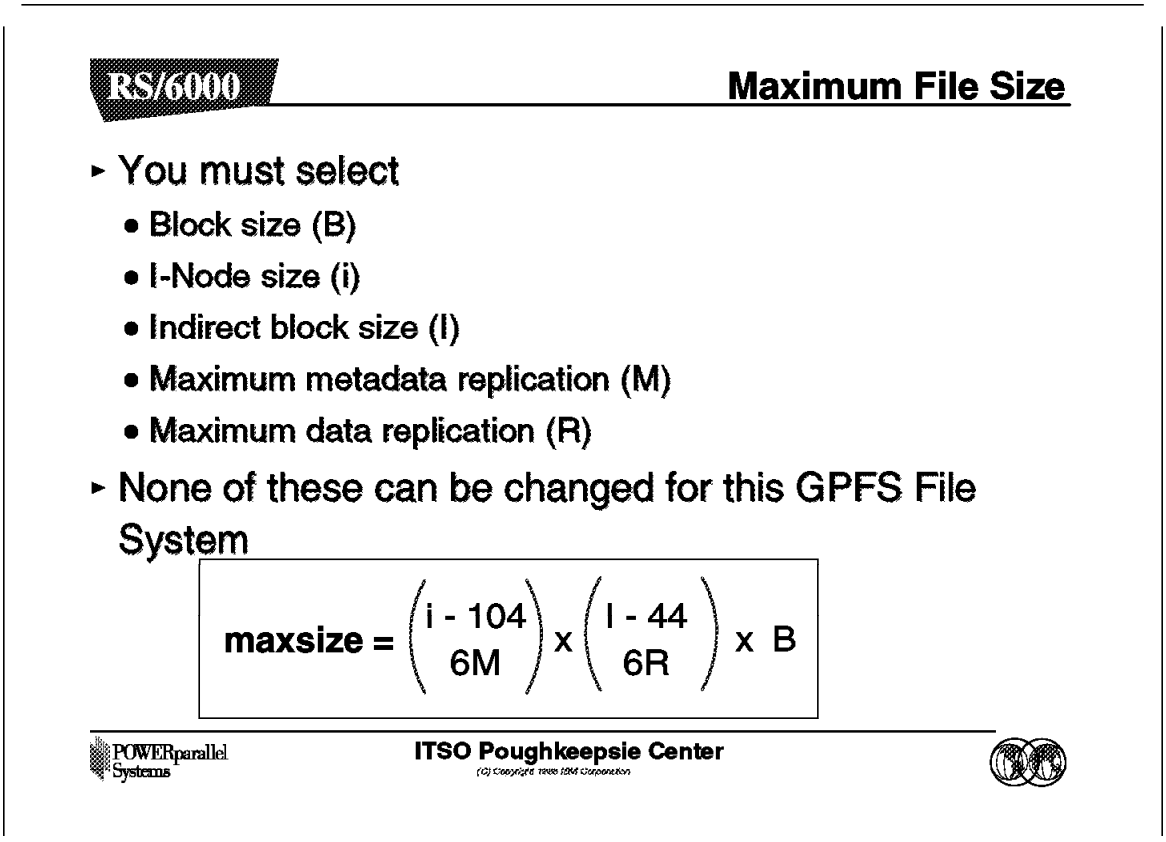


Block Size (B)	Indirect Size (I)	I-Node Size (i)	Maximum Replication (M,D)	Maximum File Size (approx)
16KB	1KB	512 bytes	1	182MB
16KB	1KB	512 bytes	2	45MB
16KB	4KB	512 bytes	1	752MB
16KB	4KB	512 bytes	2	188MB
64KB	4KB	2KB	1	14.3GB
64KB	4KB	2KB	2	3.5GB
64KB	32KB	2KB	1	115.8GB
64KB	32KB	2KB	2	28.9GB
256KB	32KB	4KB	1	951.2GB
256KB	32KB	4KB	2	237.8GB

This table shows some examples of the impact of the decisions you make on the largest file size that you can use in GPFS. You can see the impact of selecting larger block sizes, indirect size, i-node size and replication.

The formula on the next page allows you to calculate these values.

5.7.5 Maximum File Size



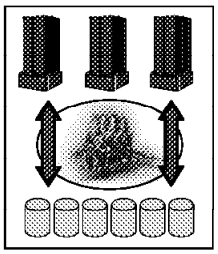
This formula helps you determine maximum file size based on the other decisions you make about this filesystem.

5.7.6 Recovery Considerations

RS/6000

Recovery Considerations

- You can build a highly available file system using GPFS
- You need to decide on how you wish to protect against failures within the SP
- There are a number of options that you can implement depending on your requirements



POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

GPFS is very strong in the area of availability. You will need to plan your GPFS system carefully to handle failures that might occur.

5.7.7 Disk Failure

RS/6000

Disk Failure

- As is usual with RS/6000 or SP, you can protect against disk failure by using AIX mirroring or a RAID disk subsystem
- For truly high availability systems, you may consider having three mirrors and you should consider how you will recover (hot plug disks) in the event of failure

This protects your Logical Volume (and therefore the VSD). This process of protecting disks is no different for GPFS. It is "business as usual".

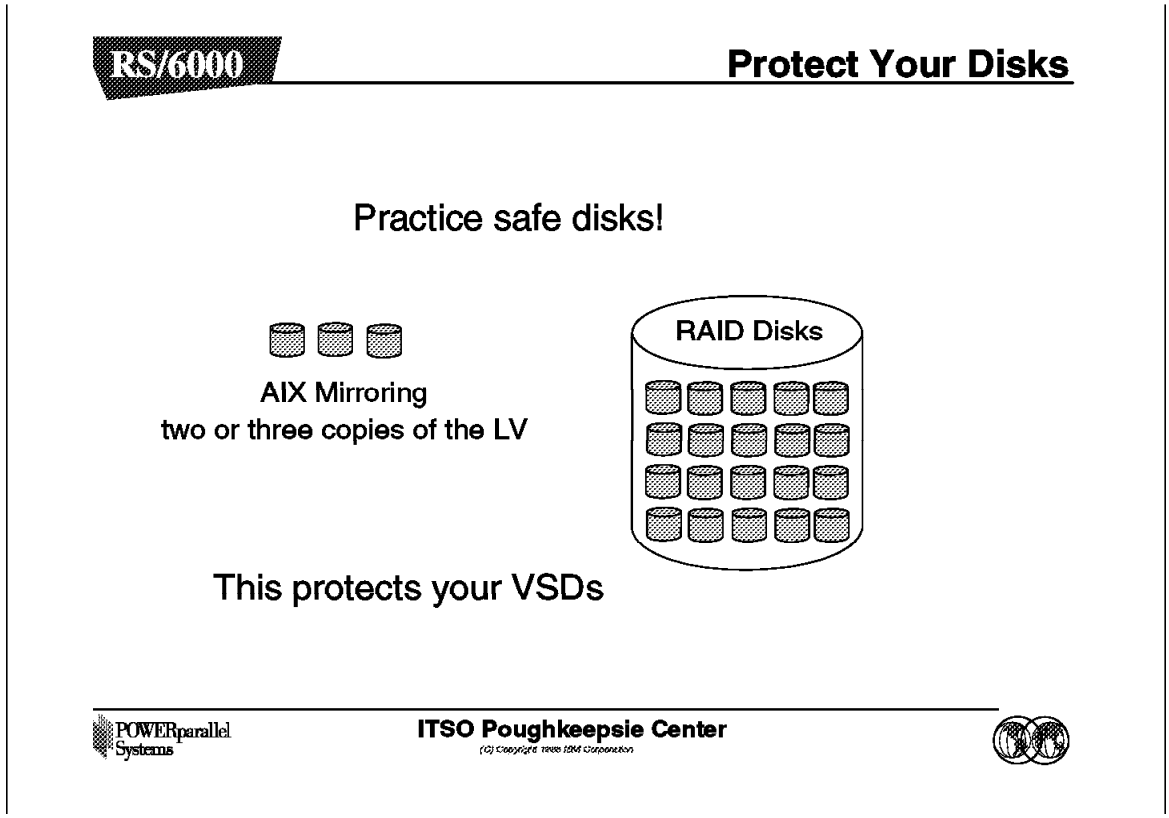
POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright IBM Corporation

Disk mirroring or RAID disks should always be considered. The default solution would normally be SSA disks, with a loop attached to two nodes in the GPFS pool, and using AIX mirroring of the Logical Volume.

5.7.8 Protect Your Disks



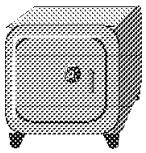
A RAID disk will provide for high availability, but will not provide the high performance of mirroring. However, RAID solutions will often be cheaper.

5.7.9 Practice Safe Nodes

RS/6000

Practice Safe Nodes

- Use twin-tailing of disks with RVSD to protect against a VSD Server failure
- RVSD will manage recovery
- There will be a short delay during takeover - only applications accessing this data will see a delay



- Data or transactions will not be lost in the event of a failure of a VSD server node
- The file system will maintain its integrity

POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright IBM Corporation

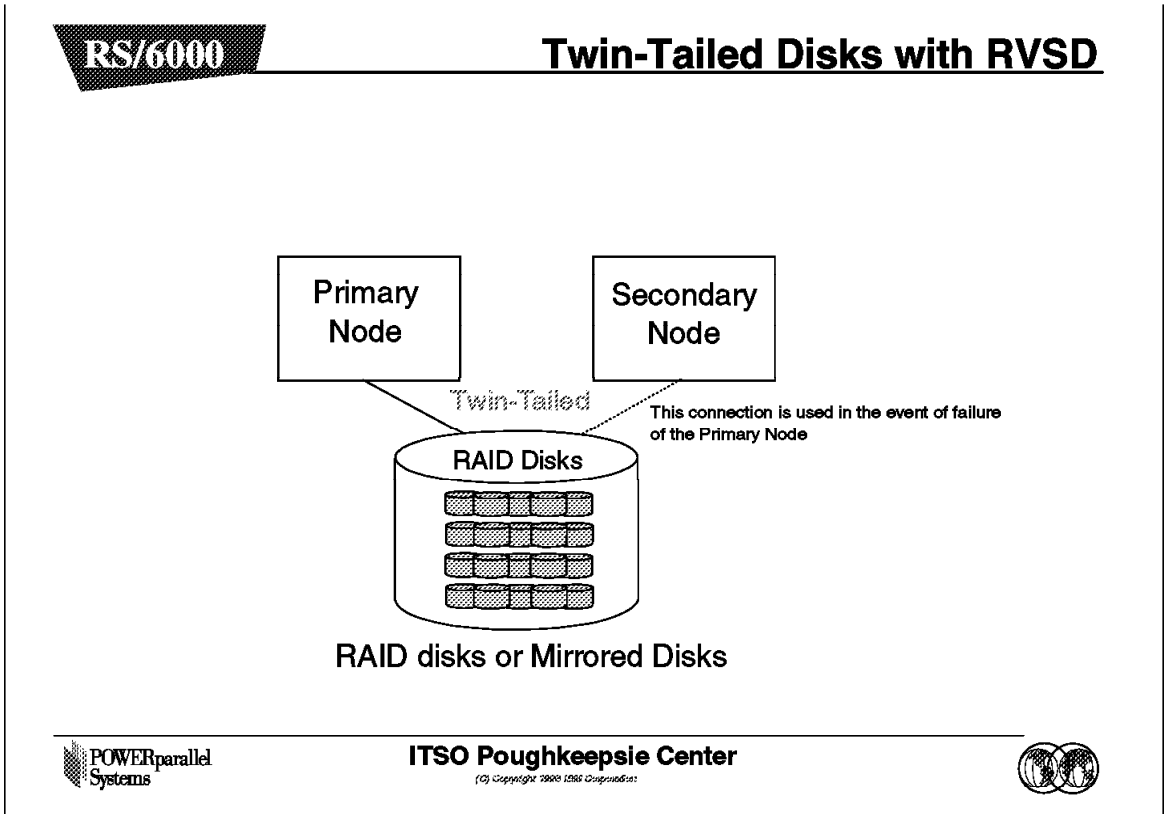
GPFS maintains two copies of logs in separate failure groups within the SP system. These are used for recovery in the event of a power failure, for example, of one node.

GPFS provides equivalent function to journaling in JFS.

To protect against a VSD server failure, you should use twin-tailing and define RVSD to automatically recover the Volume Group.

Twin-tailing your disks should normally be considered if availability is an issue, as it often is. As your file system is spread across a number of nodes, the chances of a failure should lead you to twin-tailing your disks in most circumstances.

5.7.10 Twin-Tailed Disks With RVSD



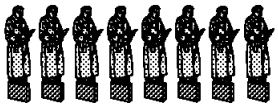
The recovery process shown here is the normal RVSD process working in conjunction with the high availability infrastructure.

5.8 GPFS Replication

RS/6000

GPFS Replication (1)

- GPFS also provides a new replication function
- Useful in a few specific cases
  - No takeover time in the event of a VSD Server failure
  - Could be used for replicating data that cannot be twin-tailed or mirrored (for example, on internal disks)



POWERparallel Systems

ITSO Poughkeepsie Center

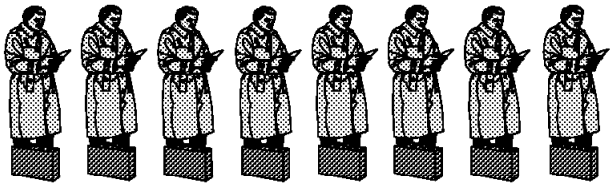
(C) Copyright 1998 IBM Corporation

The Replication function in GPFS provides some overlap with the protection that we have already discussed. Replication will not normally be your first choice. The only obvious cases where it might prove useful are listed here.

RS/6000

GPFS Replication (2)

- GPFS Replication allows you to keep additional copies of data and/or metadata
- It is recommended to always configure two copies for metadata
- The default is no replication
- The maximum number of copies is 2 in this release of GPFS
- Two copies of logs are kept by default



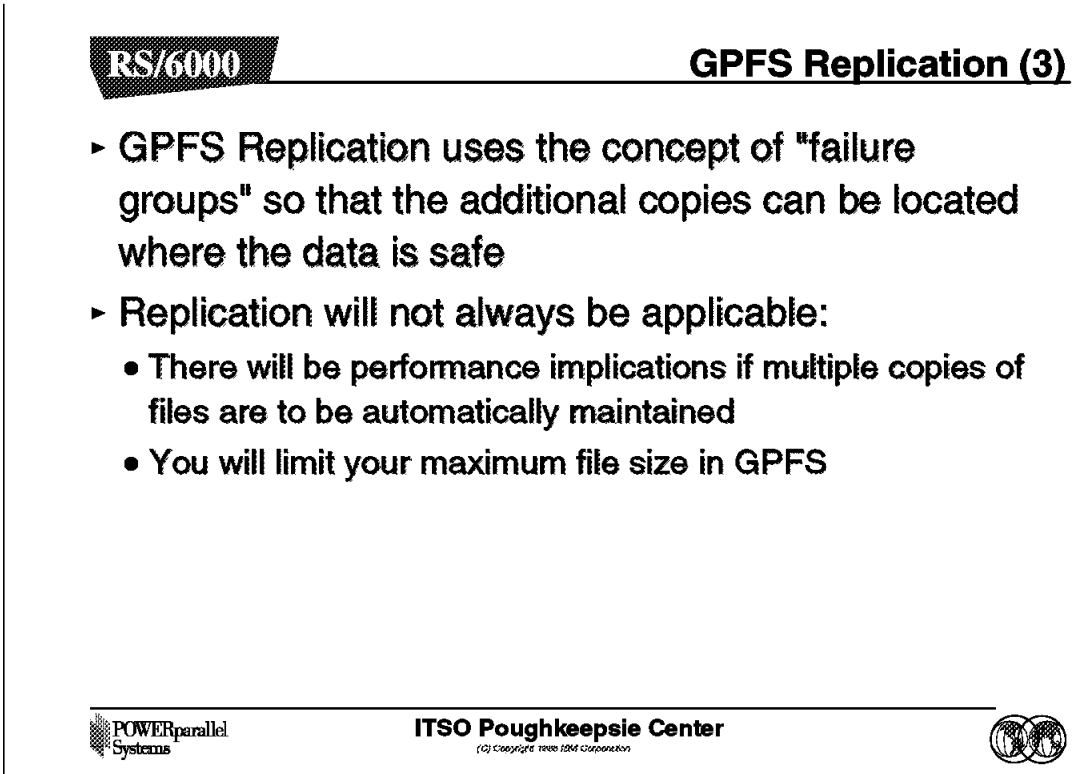
POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

The default for replication is no replication. However, two copies of logs are kept anyway, as already discussed.

You can choose at the file level whether you want to replicate a file and/or its metadata (i-node information).



It would not make sense to have another copy of the data and put it on the same disk, or even on a disk attached to the same node, as there would be occasions when we could not get to the extra copy.

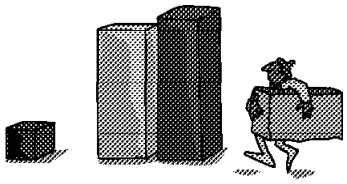
GPFS uses the concept of failure groups to make sure that data is protected from this kind of thing. Failure groups are discussed in more detail in 5.9.1, “Failure Group” on page 112. The default failure group is at the node level, and GPFS will only put replica data in a different failure group.

5.9 GPFS Recovery Parameters

RS/6000

GPFS Recovery Parameters

- At the file system level
  - DefaultMetadataReplicas
  - MaxMetadataReplicas
  - DefaultDataReplicas
  - MaxDataReplicas
- At the disk level
  - Failure Group
  - Metadata
  - Data



POWERparallel Systems

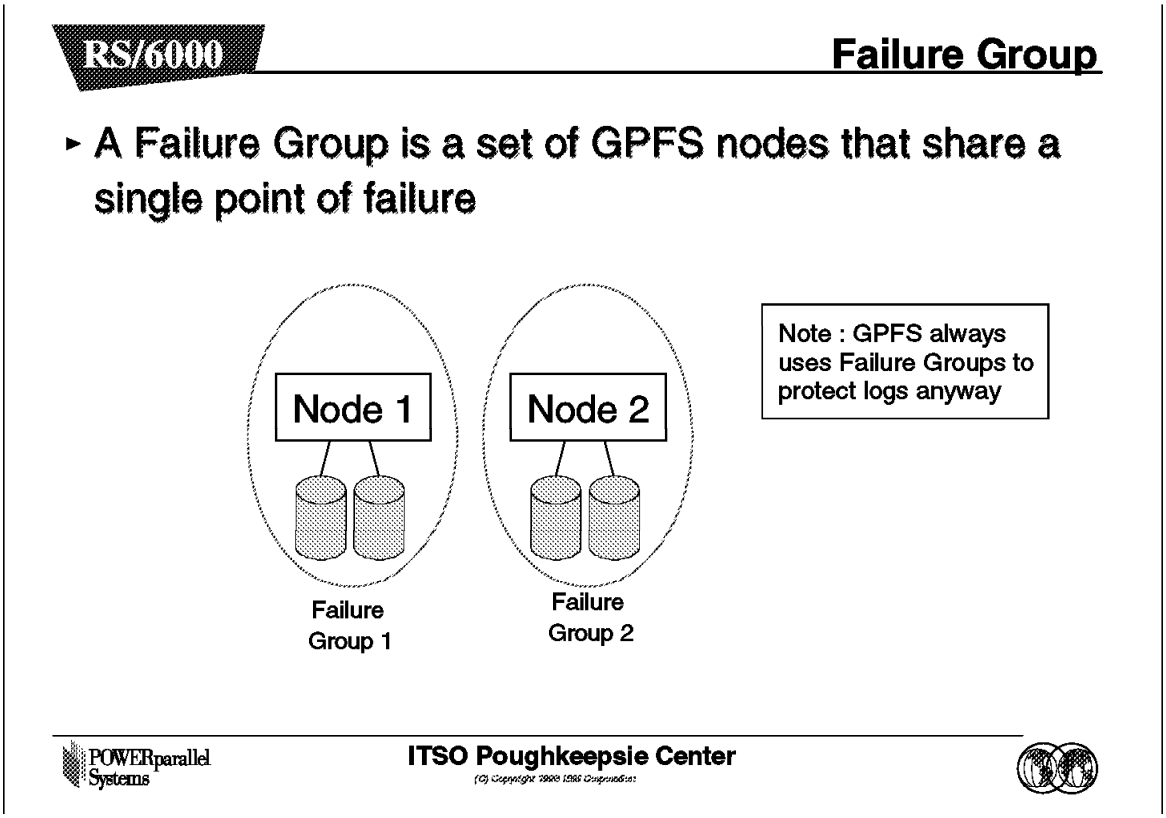
ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

With GPFS, you have a great deal of flexibility. While you cannot change the default or maximum replica settings for a file system once selected, you can change the replication settings at a file level.

A newly created file will always adopt the default settings.

5.9.1 Failure Group



GPFS by default will assign a failure group at the node level for each disk in a GPFS file system.

Typically, each node can be seen as a single point of failure, from the GPFS point of view, and, therefore, constitutes a failure group.

As a result, GPFS will only put replicas of data and metadata into a different failure group when these are configured.

A failure group is defined by a number which can be assigned by the user. The default number is the node number plus 1000. For example, a disk on node 7 will be placed in the failure group 1007.

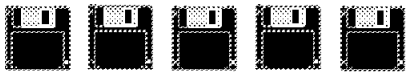
Setting a value of -1 for the failure group says that any considerations with regard to failure groups will be ignored.

5.10 Installation GPFS

RS/6000

Installing GPFS

- Installing GPFS is straightforward; you follow the steps just as you would for any LPP
- Although the CWS will not be one of the GPFS "pools" of nodes, you will need to install part of GPFS on the CWS too
- You do not have to install GPFS on every node - just those that will be in the GPFS pool; select nodes that have enough disk space, if they are VSD servers
- GPFS is only supported on the IBM SP and requires the SP Switch for connectivity between the nodes
- A maximum of 128 nodes is currently supported



POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

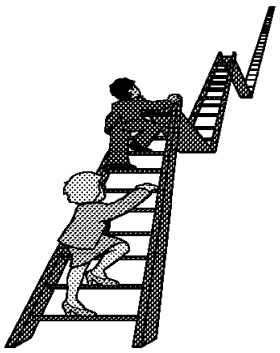
The installation of GPFS is similar to the installation of any LPP.

5.10.1 Installing GPFS - Other Steps

RS/6000

Installing GPFS - Other Steps

- Install the VSD product
  - On all nodes requiring GPFS access
  - Create a dummy VSD on all nodes
- Install RVSD
  - On all nodes requiring GPFS access
  - Start rvsdd
- Install GPFS
  - On all nodes requiring GPFS access
- Configure sysctl
- Tune the SP Switch



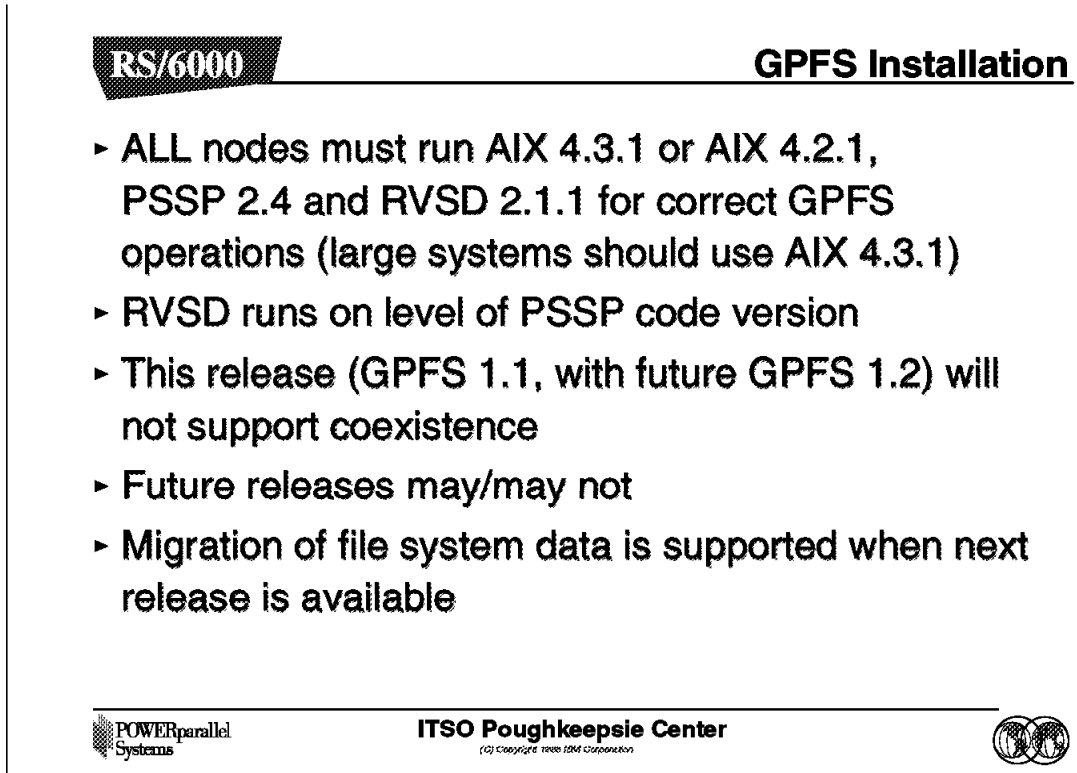
POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright IBM Corporation

There are a few simple but very important steps that you will need to take before you can use GPFS.

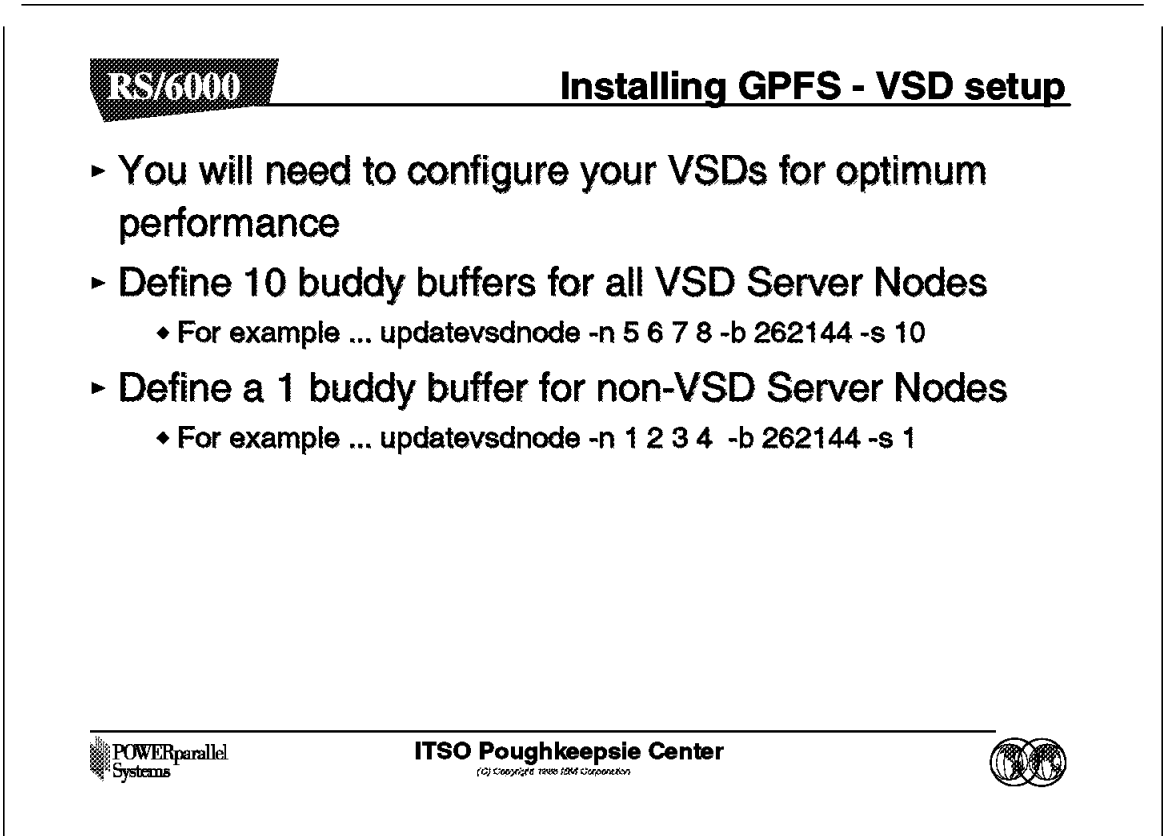
5.10.2 GPFS Installation - Required Software



GPFS requires the following software levels:

- AIX 4.2.1 or AIX 4.3.1 (PSSP 2.4 prerequisite)
- PSSP 2.4
- VSD 2.4
- RVSD 2.1.1 (even if twin-tailed disks are not used)

5.10.3 VSD Setup



Tune your VSDs in the normal way. Refer to the VSD documentation for full details.

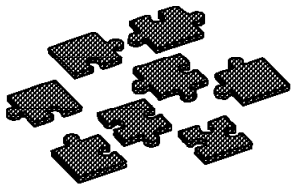
Suggested starting values, in the absence of other guidance, are given here.

5.10.4 Tune the Switch

RS/6000

Installing GPFS - Tune the Switch

- To achieve good GPFS performance, you will need to tune the SP Switch
- Use the dsh command to run the following command on the SP nodes (set it to the maximum value):
  - dsh chgcss -l css0
    -a rpoolsize=16777216
    -a spoolsize=16777216
- NB: rpoolsize and spoolsize use pinned memory



POWERparallel Systems

ITSO Poughkeepsie Center

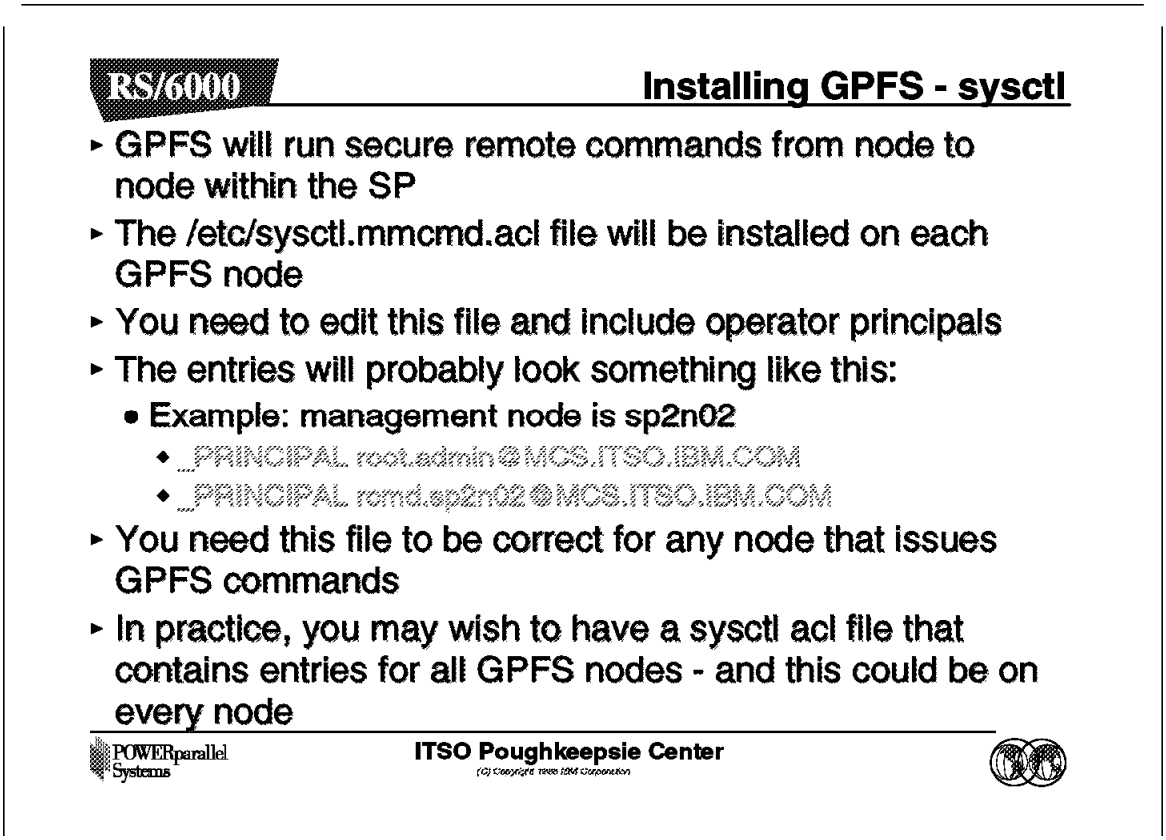
(C) Copyright 1998 IBM Corporation

Tune your switch and TCP/IP in the normal way. Refer to the PSSP documentation for full details.

Suggested starting values, in the absence of other guidance, are given here.

Note: It is recommended that you attend the SP Performance and Tuning class if you are not familiar with these procedures.

5.10.5 Sysctl



GPFS will not work without sysctl. Check carefully that sysctl is configured and working properly before going any further.

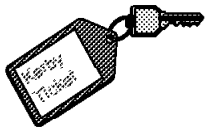
GPFS will exhibit strange errors if you do not have sysctl configured.

5.10.6 Kerberos

RS/6000

Installing GPFS - Kerberos

- GPFS will run secure remote commands from node to node within the SP using sysctl
- The mmremote and mmmkvsd sysctl procedures or commands will be added to the sysctl configuration and will be used internally for many GPFS commands
- To use sysctl, you need a Kerberos ticket on any node that you use to run GPFS commands
- This Kerberos ticket is required to run sysctl remote commands
- Getting a ticket with /usr/lpp/ssp/rcmd/bin/rcmdtgt will suffice
- Use secure rsh/dsh from the CWS to your GPFS "controller" node, open an aixterm, and run all commands at the CWS to be totally secure



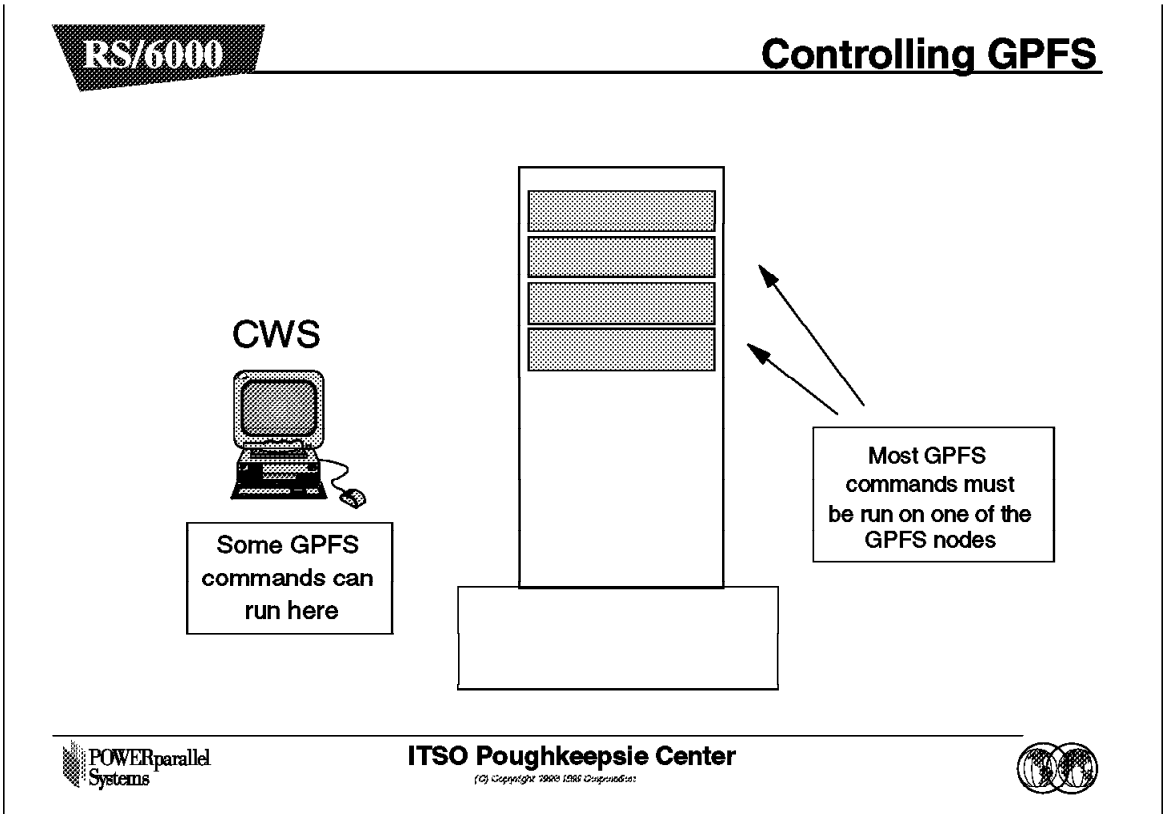
POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

To use sysctl, you need a Kerberos ticket. Rather than type in a `kinit` command on the command line when working on a GPFS node, a remote ticket-granting ticket will be a better option.

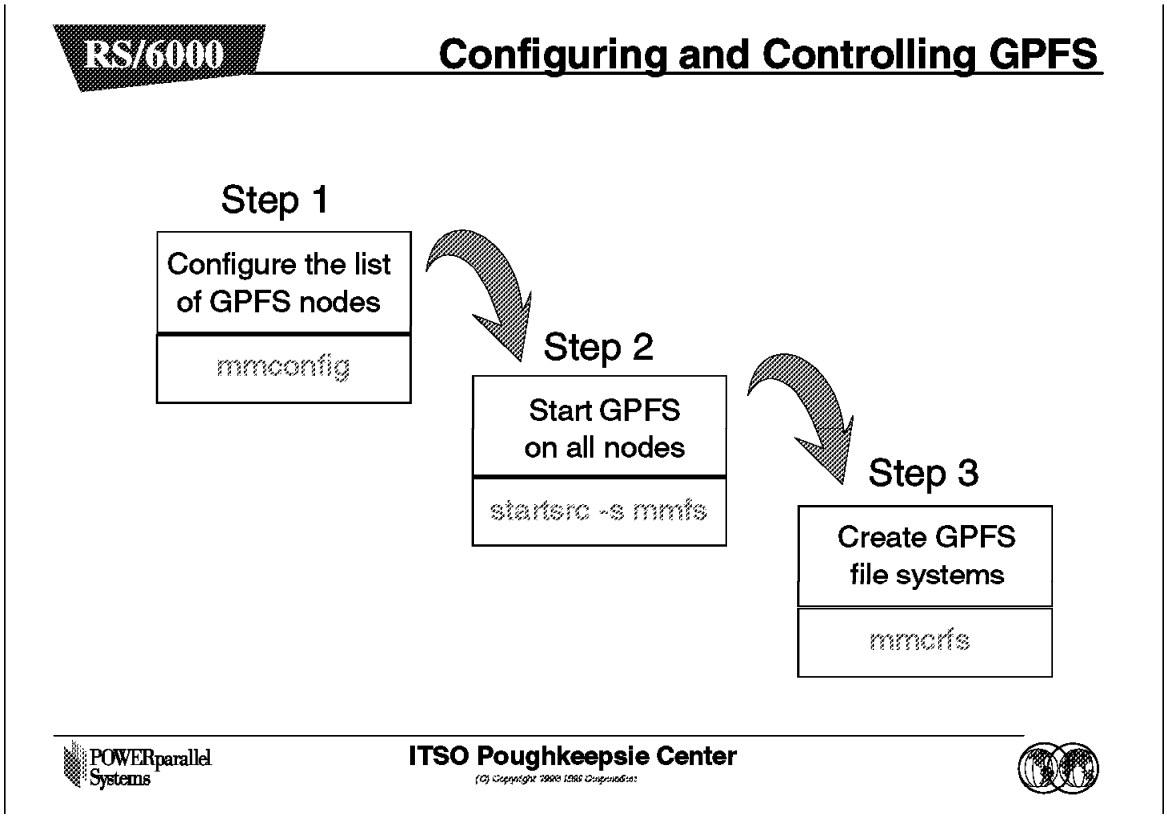
5.10.7 Controlling GPFS



One of the things to note about GPFS is that much of the system management cannot be performed directly at the CWS. You need to be on a GPFS node to execute most GPFS commands.

You can use rsh or dsh from the CWS to achieve this.

5.10.8 Configuring GPFS



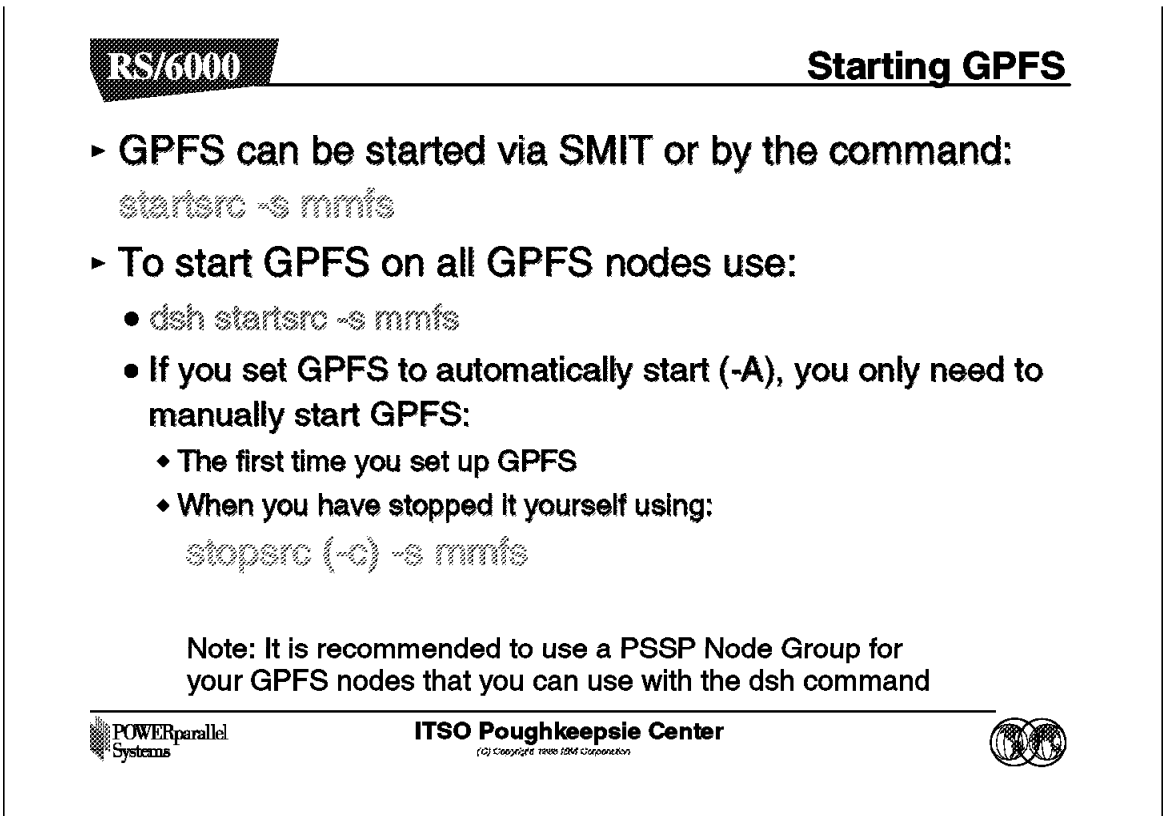
To get going with GPFS there are three distinct things you need to do, from a high-level point of view, as shown in the figure.

There is a “one off” configuration setup command that you will have to run to tell GPFS which nodes are in the GPFS pool. You also provide some other configuration information.

Once this is complete, you can start GPFS (mmfs) on all GPFS nodes.

You are now ready to create and mount GPFS file systems across your GPFS nodes within the SP.

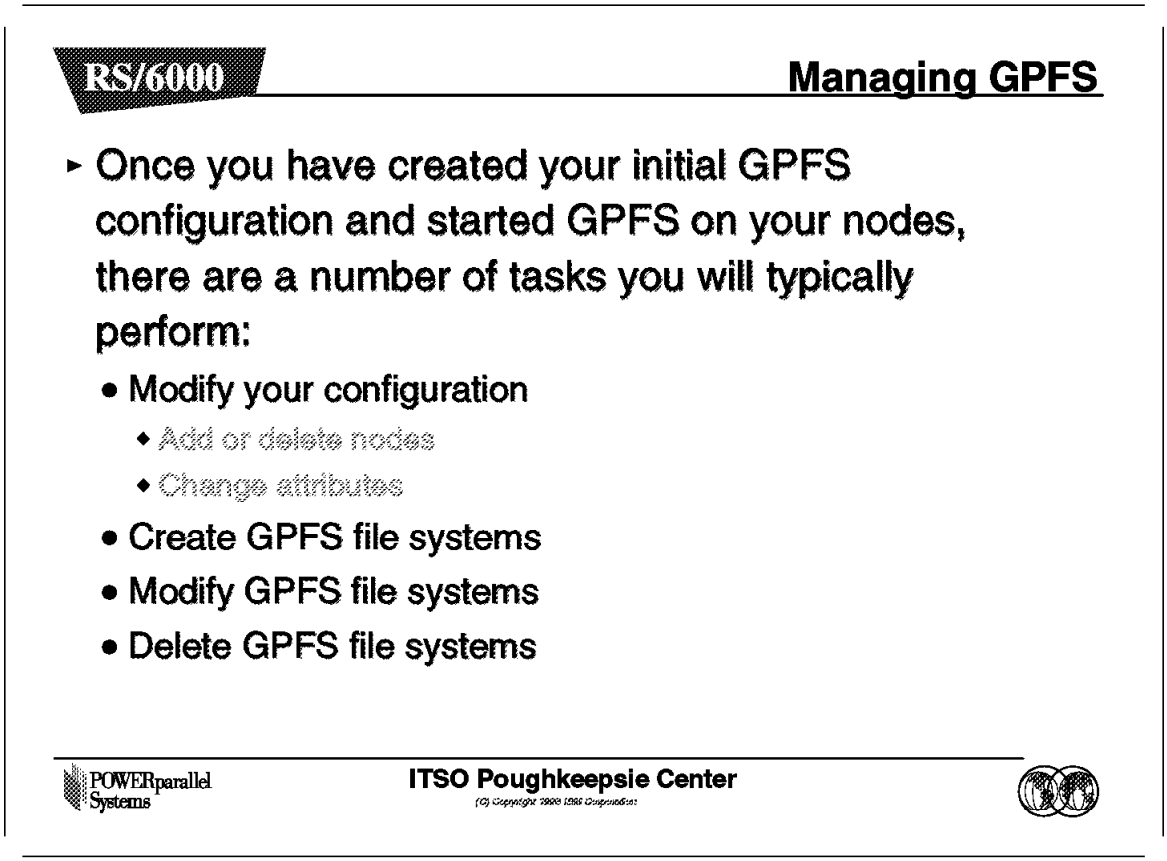
5.10.9 Starting GPFS



Having configured GPFS, you can now start GPFS on your GPFS nodes.

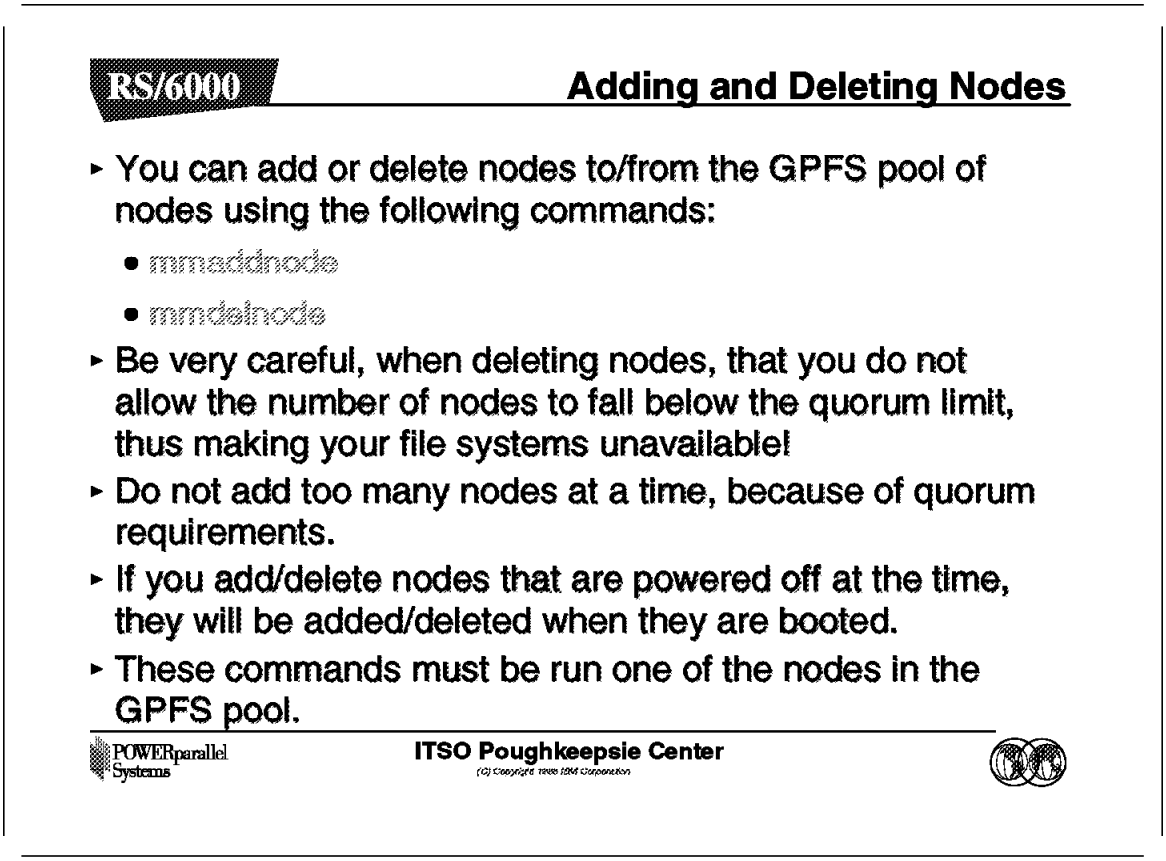
Note: Your CWS will not be one of your GPFS pools and you will not run GPFS on the CWS.

5.11 Managing GPFS



We will consider the options in this area later. However, these are typically the kinds of tasks that you will want to perform.

5.11.1 Adding and Deleting Nodes



If you add nodes into the GPFS pool and these nodes have disks so that they can act as VSD servers, be aware that new files will be striped across these additional disks. Old files will not be striped across these new disks unless you restripe the file system, which is not recommended. It is better to start with the correct number of nodes.

5.11.2 Creating GPFS File Systems

RS/6000

Creating GPFS File Systems - Decisions

- Will you create your own VSDs first, or will you allow GPFS to create VSDs for you (normally preferred)?
- Which disks on which nodes will you use?
- How will you structure the file system?
  - i-node size
  - Indirect size
  - Block size
- Will you use GPFS replication?
- How will you structure your Failure Groups?

POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

You are now ready to create a GPFS file system to store data on the GPFS nodes. Plan what you will do carefully. Many decisions cannot be reversed.

Unless you have good reasons to do otherwise, let GPFS create the VSDs for you.

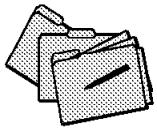
Do not use replication as your first choice for providing availability.

5.11.3 Disk Descriptor Files

RS/6000

Disk Descriptor Files

- It is best to create a disk descriptor file for each GPFS file system
- This will have a line for each VSD that will be part of the GPFS file system



- Each line in this file specifies:

• hdisk names/VSD name	Device names - eg hdisk1, hdisk3
• Primary Server Name	Hostname of primary server (or IP address)
• Secondary Server Name	Hostname of backup server (for use with RVSD)
• Failure Group	A number specifying the Failure Group
• Metadata/Data	Specify: data/metadata/data & metadata

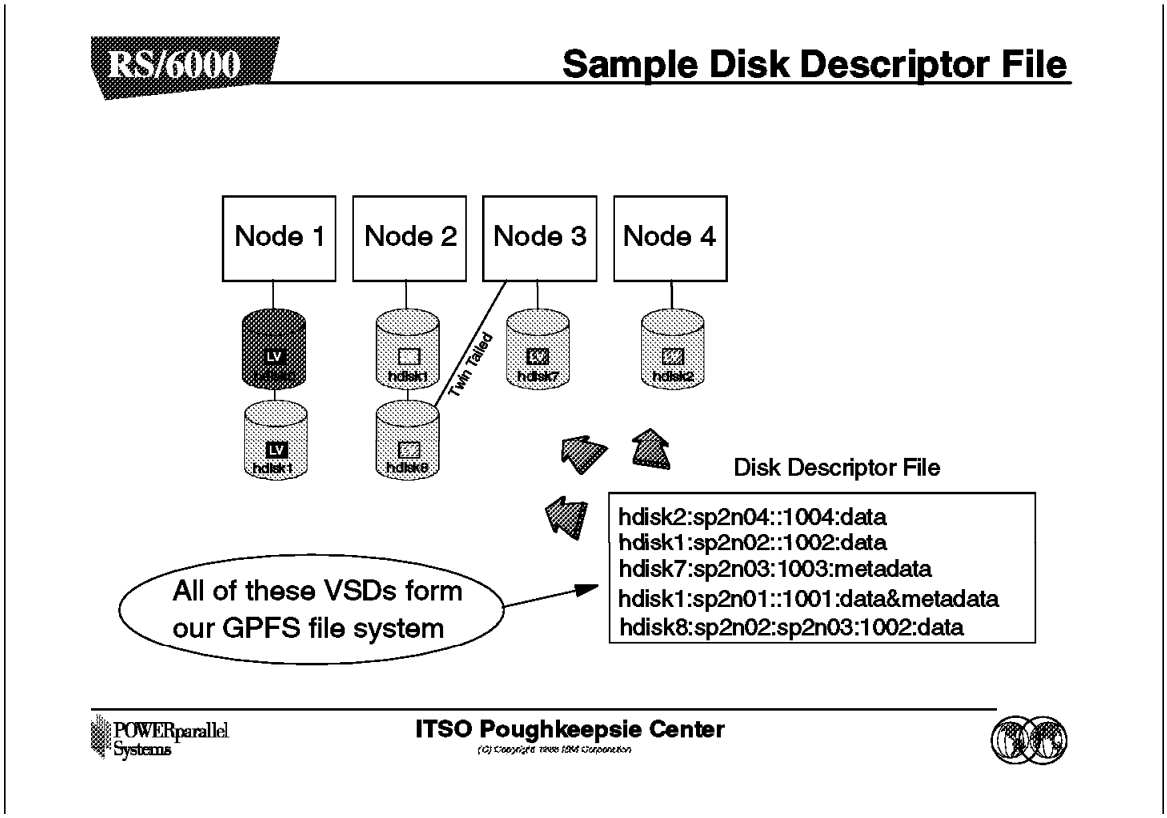
POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright IBM Corporation

The information that you provide in a disk descriptor file is at the heart of a GPFS file system. It defines in detail exactly how the VSDs will be created. This information will be used by the create VSD commands.

5.11.4 Sample Disk Descriptor File



In this example, our disk descriptor file can hold one line for each VSD that goes to make up our GPFS file system.

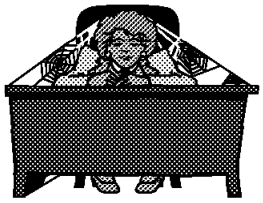
We have a lot of flexibility. Note that this example is not one that we would normally choose to implement because it has an imbalance of disks.

5.11.5 Create GPFS File System Command

RS/6000

Create GPFS File System Command

- You can create a GPFS file system using the mmcrfs command.
- Example:
  - mmcrfs /clivefs2 fs2 -F /gpfsfiles/fs2desc -A yes
- You can also key your VSD parameters into the SMIT screen as you create a file system.
- If you have already created your VSDs, you only need to provide the names of these VSDs in your Disk Descriptor File.



POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright IBM Corporation

The command for creating a GPFS file system is shown here.

5.11.6 Mounting File Systems

RS/6000

Mounting File Systems

- You can now mount your GPFS file system on any GPFS node using the usual AIX mount command:
  - mount /clivefs2
- An entry of the GPFS file system can be found in /etc/filesystems
- Use exportfs to NFS export the GPFS filesystem

POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1995 IBM Corporation

The GPFS file system, once created, can be mounted on any GPFS node in the normal way with the mount command.

5.11.7 Repairing a File System

RS/6000

Repairing a File System

- Repair a GPFS file system using the mmfsck command
- This should not normally be necessary
- The file system must be unmounted

POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

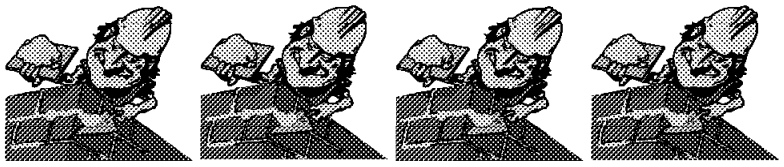
You should not normally need to recover by running mmfsck, but the facility is there if required. Under normal circumstances, including failures, GPFS should repair any file systems itself when required.

5.11.8 Restriping a GPFS File System

RS/6000

Restriping a GPFS File System (1)

- There are a few reasons for restriping a GPFS file system, but it should be avoided if possible
- Potential reasons for restriping include:
  - A new disk has been added to a GPFS file system, and a low level of updates to this file system means that restriping to utilize the new disk would be helpful in balancing performance



POWERparallel Systems

ITSO Poughkeepsie Center

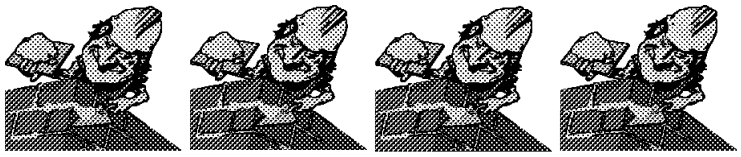
(C) Copyright 1998 IBM Corporation

Restriping is an intensive process and should be avoided if possible.

RS/6000

Restriping a GPFS File System (2)

- Plan to carry out any restriping when activity is low
- To restripe a large GPFS file system (a terabyte or more) can take a very long time
- Before you restripe a GPFS file system, suspend any disks that you want to exclude
- Check that the disks that you want included are up and running OK

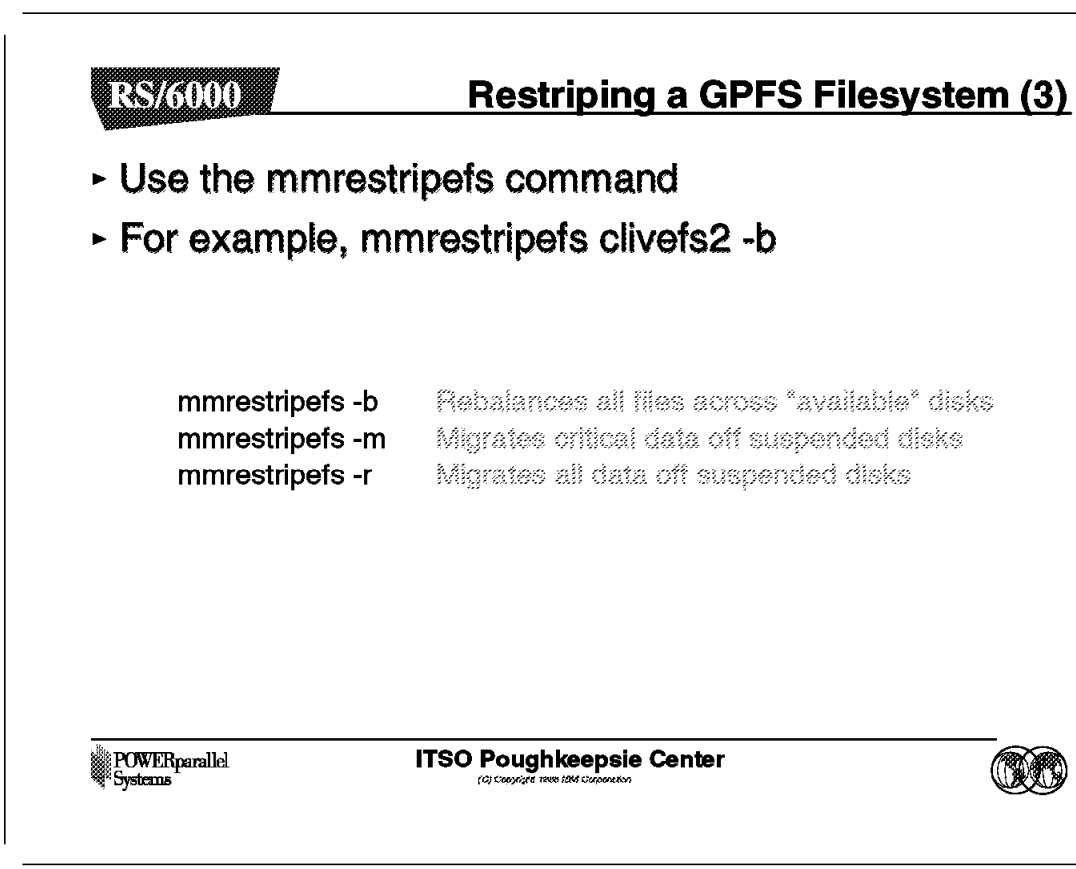


POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

If you must restripe a file system, do it at a time when system activity is low.



Here are some examples of commands to restripe a file system.

5.11.9 Changing Disk States

RS/6000

Changing Disk States

- Changing disk states can be performed using the mmchdisk command, using one of four keywords:
  - suspend
  - resume
  - stop
  - start

POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

You can change the state of disks when required.

5.11.10 Adding or Deleting Disks

RS/6000

Adding or Deleting Disks

- mmadddisk allows you to add a disk to a GPFS file system
- mmdeldisk allows you to delete a disk from a GPFS file system
- mmrpldisk allows you to replace a disk

POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1995 IBM Corporation

You can add or delete disks to the GPFS system. You will be defining VSDs in much the same way as you did when you created your file system.

5.11.11 Deleting a File System

RS/6000

Deleting a File System

- You can keep the VSDs or remove them as you remove a GPFS file system
- The file system must be unmounted on all nodes before it can be removed

POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1996 IBM Corporation

File systems can be removed when no longer required.

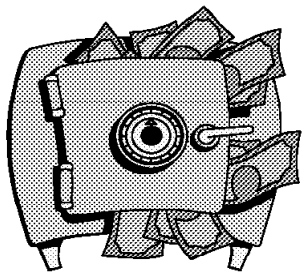
Use the `mmdelfs` command to remove a GPFS file system.

5.11.12 Access Control Lists (ACLs)

RS/6000

Access Control Lists (ACLs)

- Access Control Lists give you additional control over file access to GPFS file systems
- There are new GPFS commands specifically for managing ACLs in GPFS:
  - mmputacl
  - mmeditacl
  - mmgetacl
  - mmdelacl



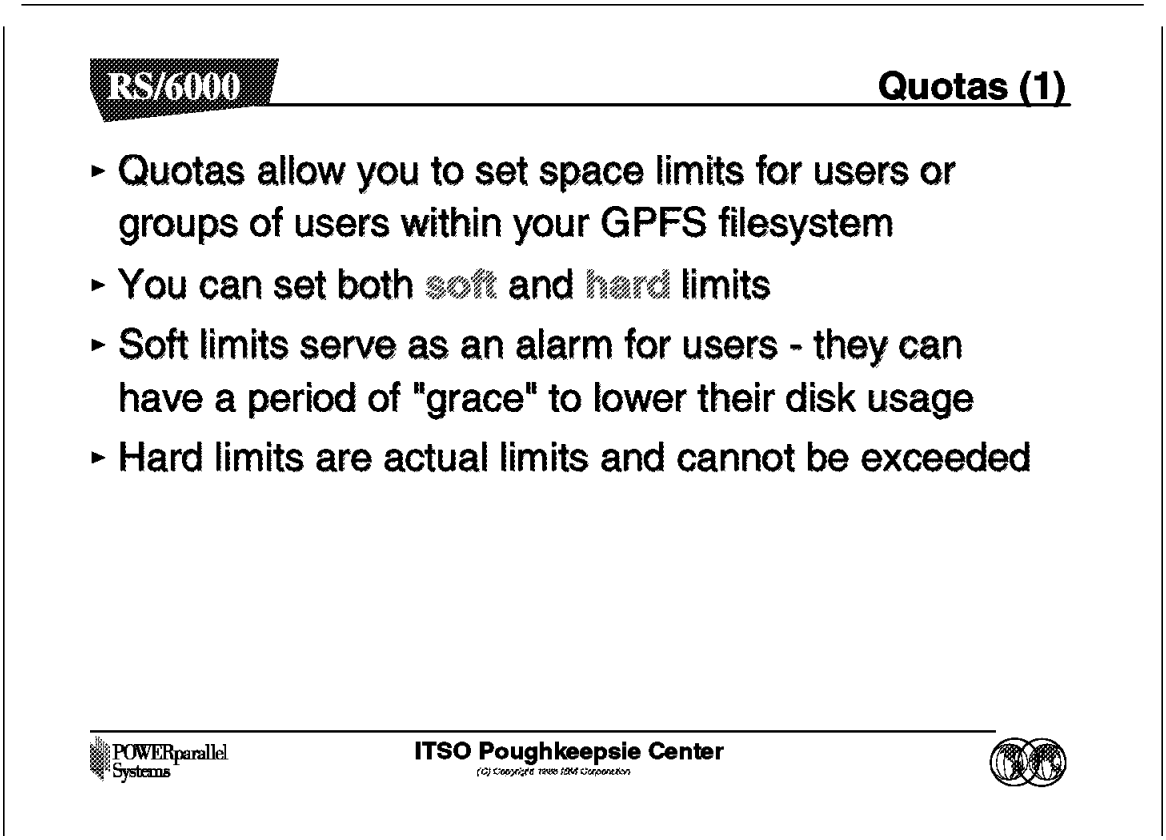
POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

As with standard AIX, ACLs give you additional control over standard file permissions to allow you to give more secure access to files and file systems to users or groups of users.

5.11.13 Quotas



You can set limits or quotas for the space that users or groups of users can use within the GPFS file system.

RS/6000

Quotas (2)

- You can use the following quota commands within GPFS to manage GPFS quotas:
  - mmedquota
  - mmcheckquota
  - mmquotaon
  - mmquotaoff
  - mmrepquota

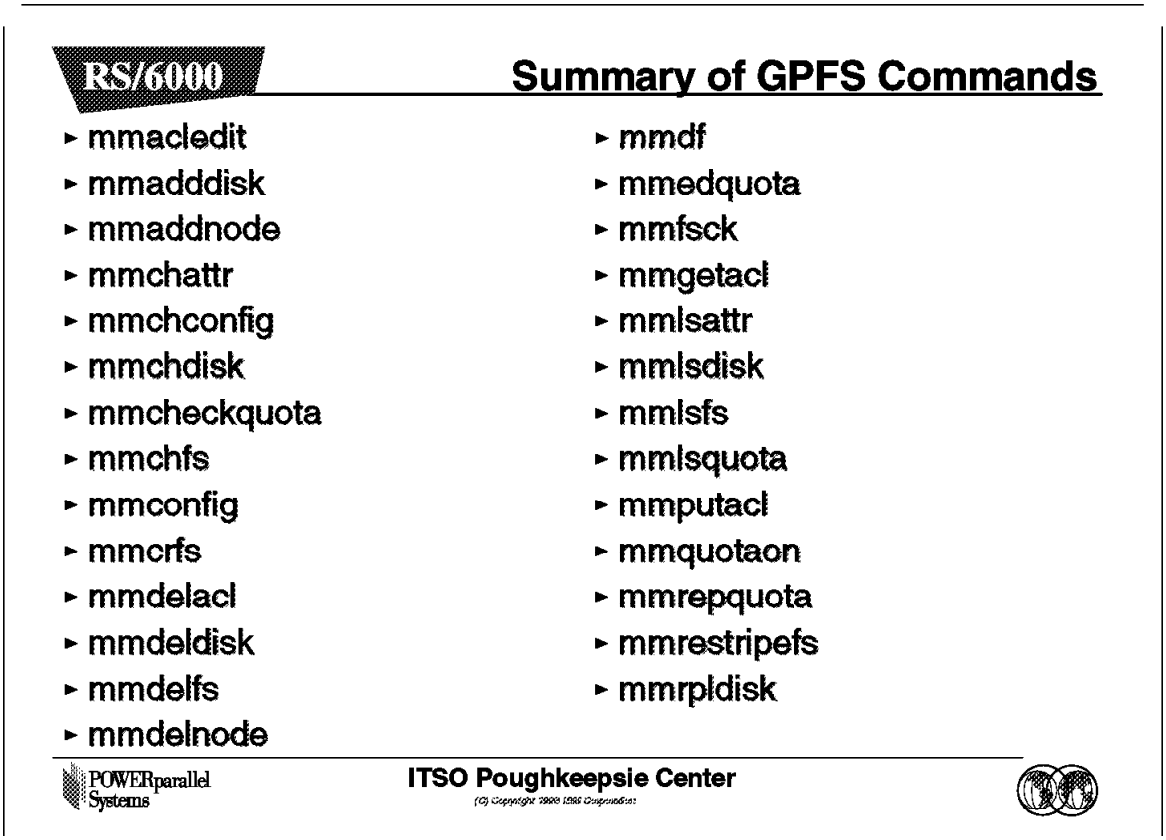
POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

As shown, there are mm commands to allow you to manage quotas.

5.11.14 Summary of GPFS Commands



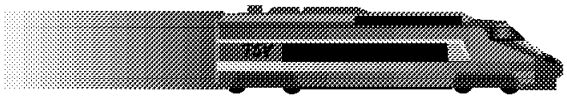
Here is a summary of all of the GPFS (mm) commands.

5.12 GPFS Performance

RS/6000

GPFS Performance

- Good performance is dependent on having a well-tuned SP system
- Areas to consider include:
  - SP Switch tuning
  - TCP/IP tuning
  - LVM tuning
  - VSD tuning
  - GPFS tuning



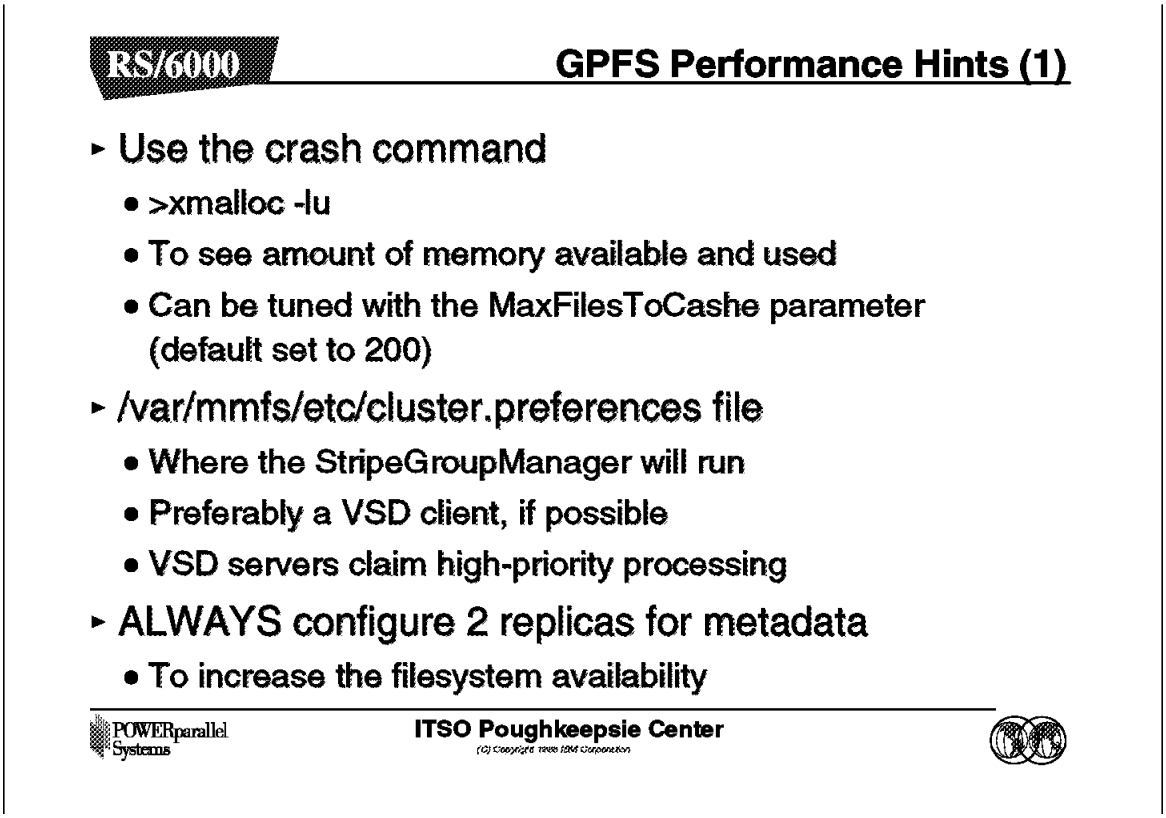
POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

You must tune your SP system and SP Switch to gain optimum performance.

5.12.1 GPFS Performance Hints



A few hints for better GPFS performance are presented here.

RS/6000

GPFS Performance Hints (2)

- Use the /usr/lpp/css/bin/statvsd command:
  - To monitor the VSD behavior
  - Look for the following entries
    - Total retries
    - Total timeouts
  - Indicates too much load/too many disks on that VSD server
- Use netstat -d:
  - To monitor TCP/IP behavior
  - Look for dropped packets

	Ipkts	Opkts	Idrops	Odrops
css_if0	X	Y	<1%X	<1%Y

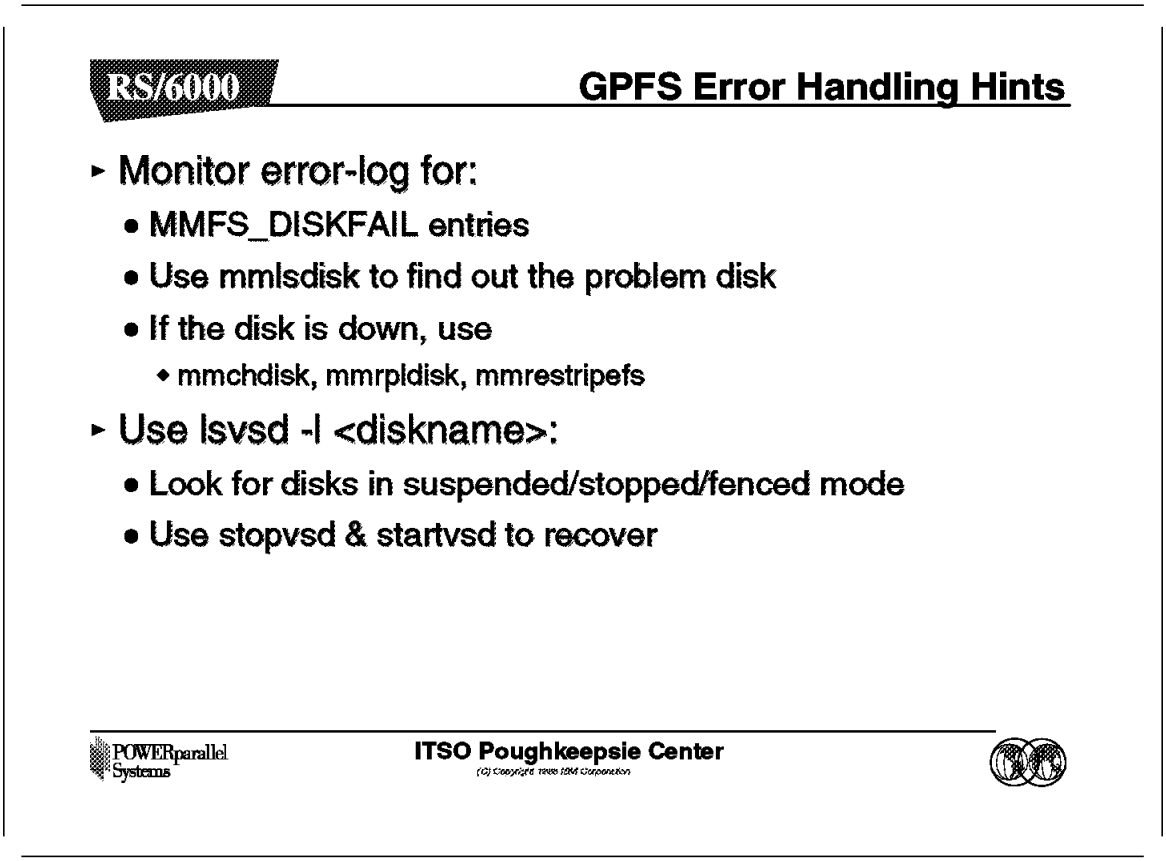
POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

The SP Switch is essential for GPFS performance. Always tune the Switch to get the best performance. VSDs are also a sensitive part of GPFS performance. Good VSD tuning will result in better GPFS performance.

5.13 GPFS Error Handling Hints



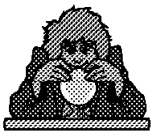
GPFS provides several command and error log entries to make the diagnosis of a problem easier.

5.14 Future GPFS Enhancements

RS/6000

Future GPFS Enhancements

- Multiple terabyte files can be theoretically supported
- Testing will allow larger files to be supported
- Switch-attached disk subsystems support
- S7x support
- DFS support
- mmap support
- MPI/IO support
- No atime/mtime support - too much of a performance penalty
- Adding extra I-nodes capability
- No buffer space allocation - automatically defined
- PRPQ for >128 support



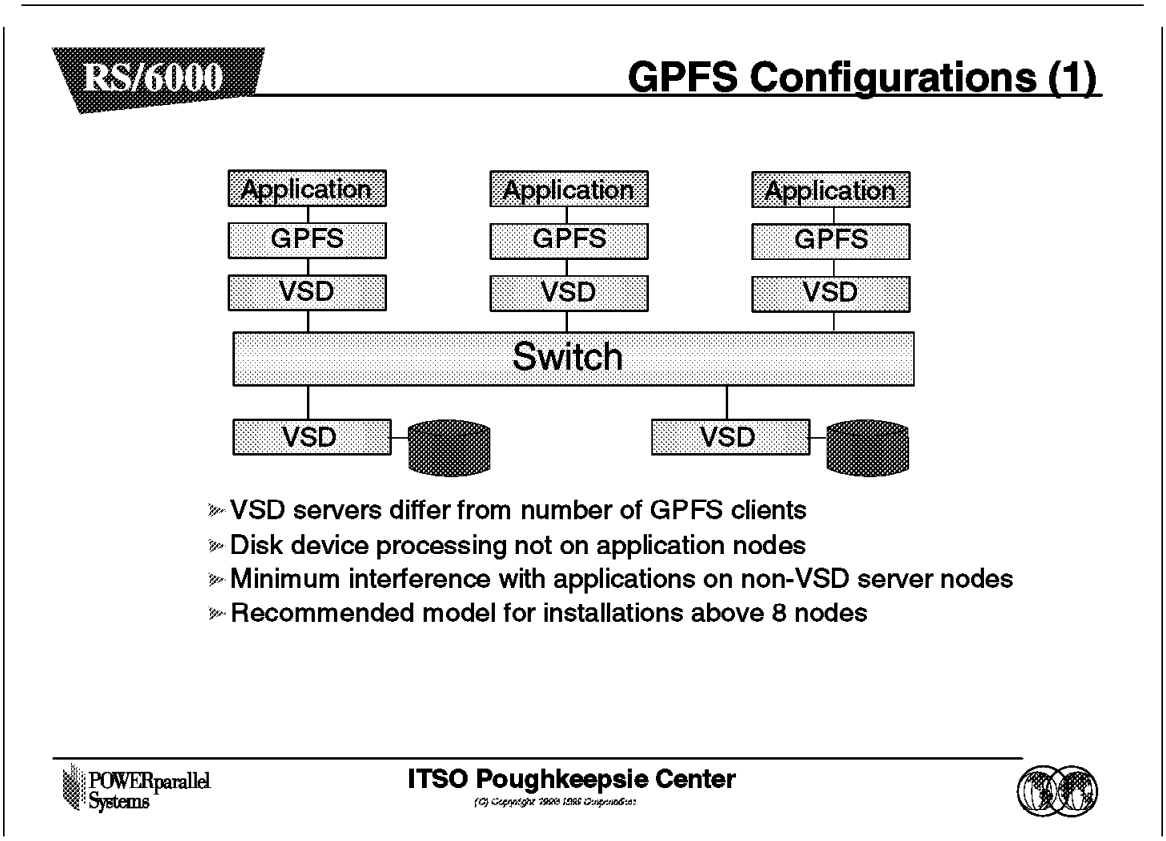
POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

As shown in the foil, these are some of the improvements that will come with future GPFS releases.

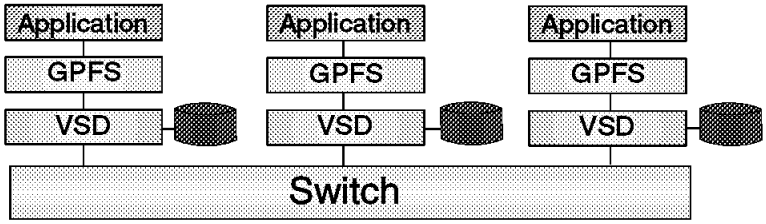
5.15 GPFS Configuration Examples



This example is a preferred solution if such a solution can be cost-justified. The VSD servers are on separate, dedicated nodes.

RS/6000

GPFS Configurations (2)



- Flat model; each server is both client and server
- Highest degree of scalability
- Typical/Recommended for installations upto 8 nodes

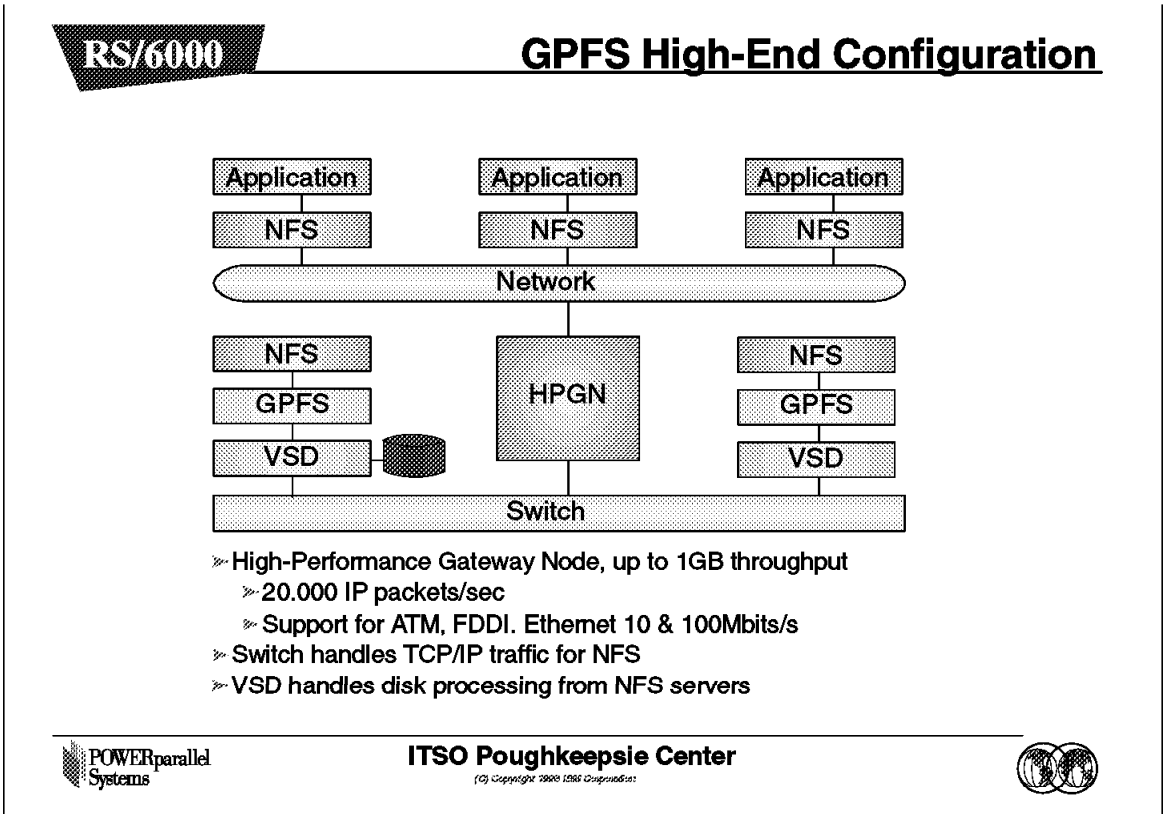
POWERparallel Systems

ITSO Poughkeepsie Center

(C) Copyright 1998 IBM Corporation

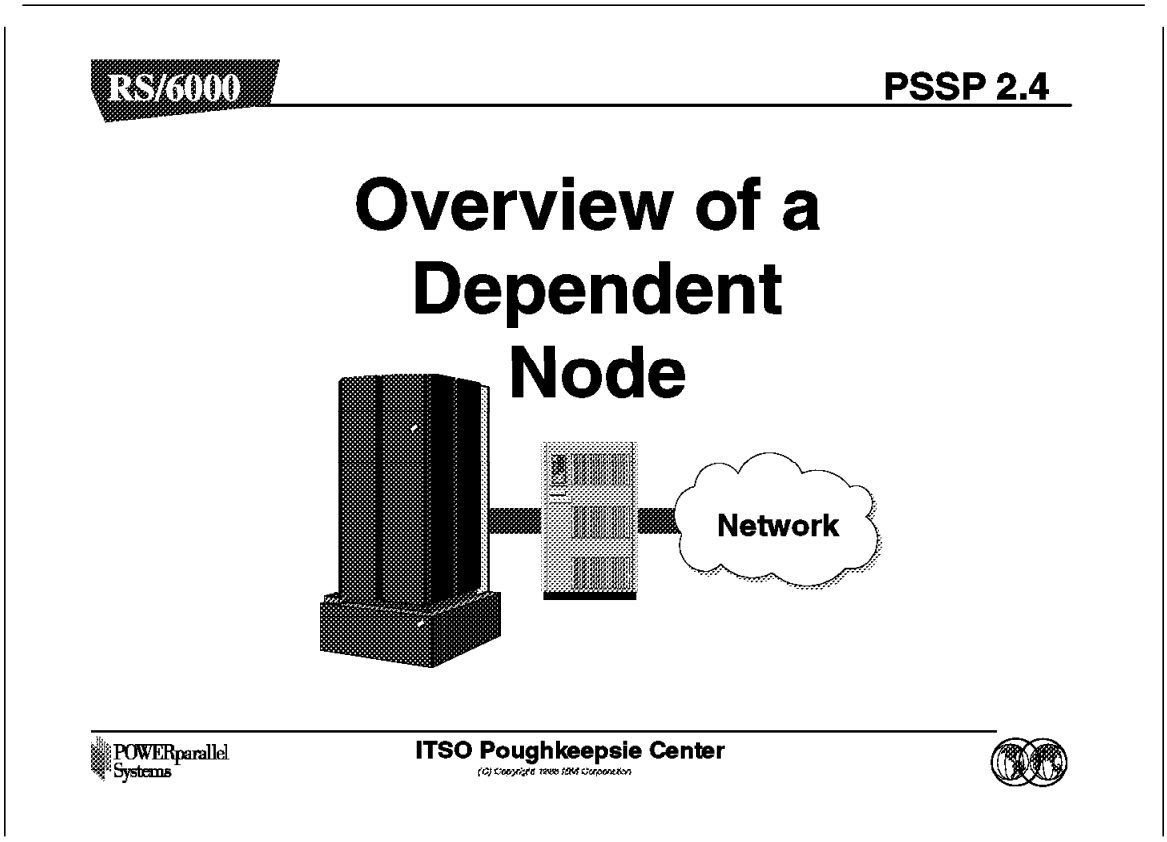
In this example, all nodes are VSD servers and also run the applications. Such a solution would be acceptable for a parallel application that has the same performance requirements across the nodes and exhibits a balanced workload.

5.15.1 GPFS High-End Configuration



This configuration shows the use of the GRF node to allow NFS clients to access the GPFS file system with a high bandwidth.

Chapter 6. Dependent Node



This chapter provides an overview of a dependent node in RS/6000 SP. We start by defining the dependent node and discussing the reasons for its design.

Next we define a router, and introduce the switch node router or GRF.[1] The GRF has a media card that attaches to the SP Switch. Together they form the dependent node. Then we compare the routing process with and without the GRF.

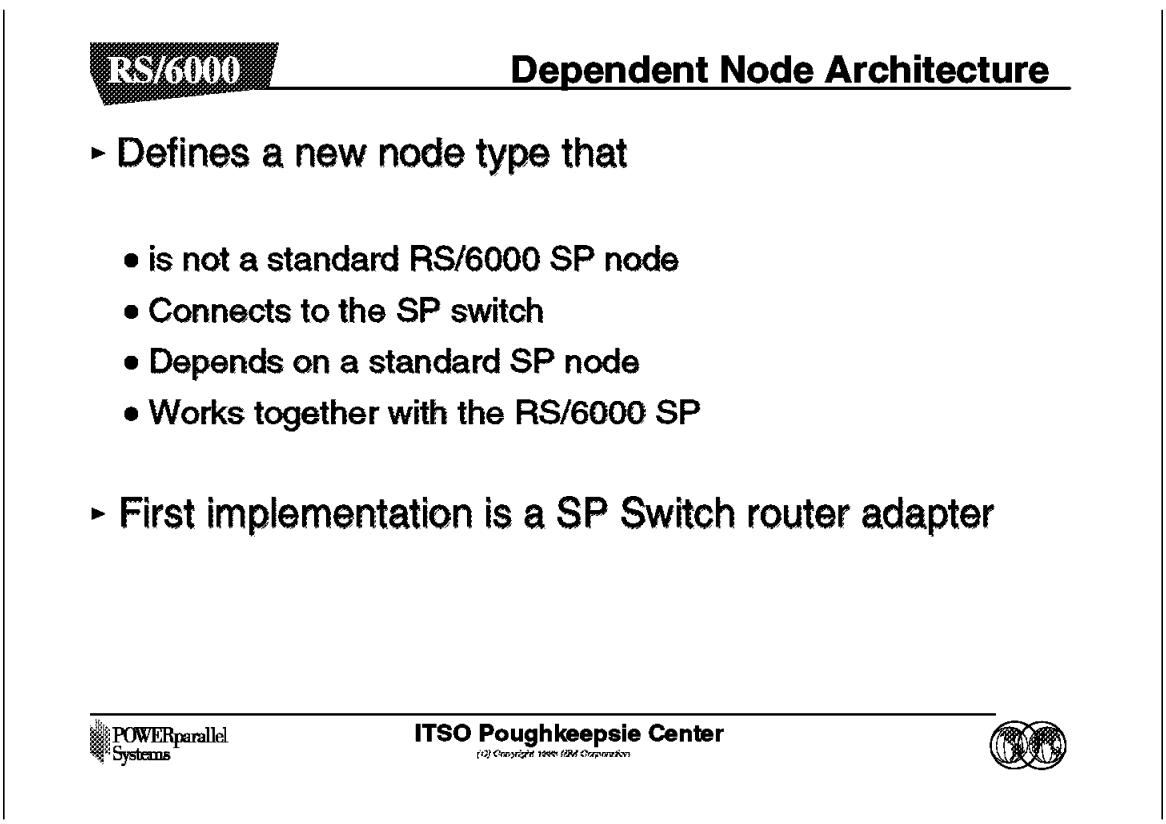
We briefly describe the enhancements to the RS/6000 SP due to the introduction of the dependent node, and discuss tasks such as planning and installation using coexistence and partitioning with the dependent node. We also introduce several sample GRF configurations.

Finally, we end by discussing limitations of the dependent node, and give you hints and tips that are based on both our experience and on common problems typically encountered when dealing with extension nodes.

[1] GRF stands for Goes Really Fast

© Copyright IBM Corp. 1998

6.1 Dependent Node Architecture



The Dependent Node Architecture refers to a processor or node, possibly not provided by IBM, for use with the RS/6000 SP.

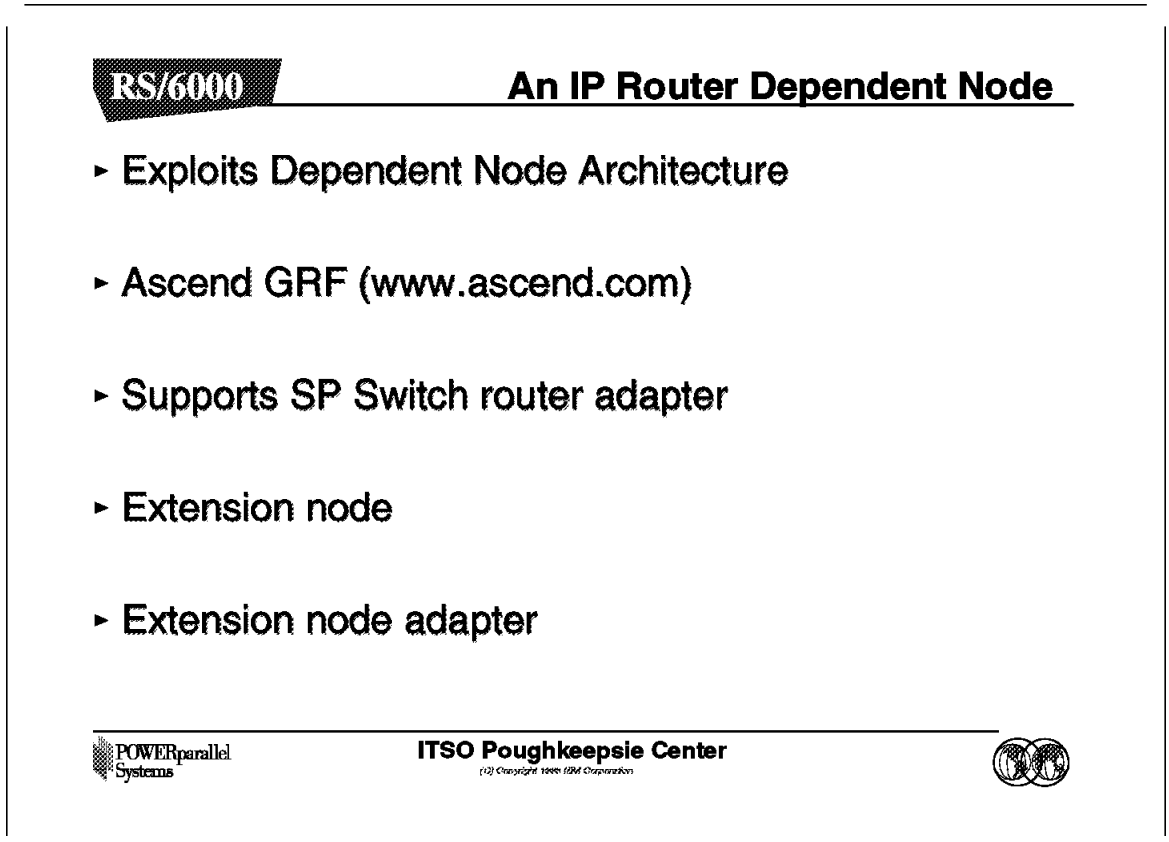
Because this is not a regular RS/6000 SP node, it cannot perform some regular node functions. Instead, it relies on normal RS/6000 SP nodes to do some of its work, which is why it is called “dependent.” For example, it does not include all the functions of the complete fault service (Worm) daemon, as other RS/6000 SP nodes with access to the SP Switch do.

The objective of this architecture is to allow the other processors or hardware to easily work together with the RS/6000 SP, extending the scope and capabilities of the system.

The dependent node connects to the RS/6000 SP Switch.

The SP Switch router adapter in the Ascend GRF is the first product to exploit the Dependent Node Architecture.

6.1.1 IP Routing Dependent Node



The first dependent node is actually a new SP Switch router adapter in a router. The purpose of this adapter is to allow the GRF, manufactured by Ascend, to forward SP Switch IP traffic to other networks. The GRF was known as the High Performance Gateway Node (HPGN) during the development of the adapter. IBM remarkets models of the GRF that connect to the SP Switch as the SP Switch router model 04S (9077-04S) and model 16S (9077-16S). These models are not available directly from Ascent.

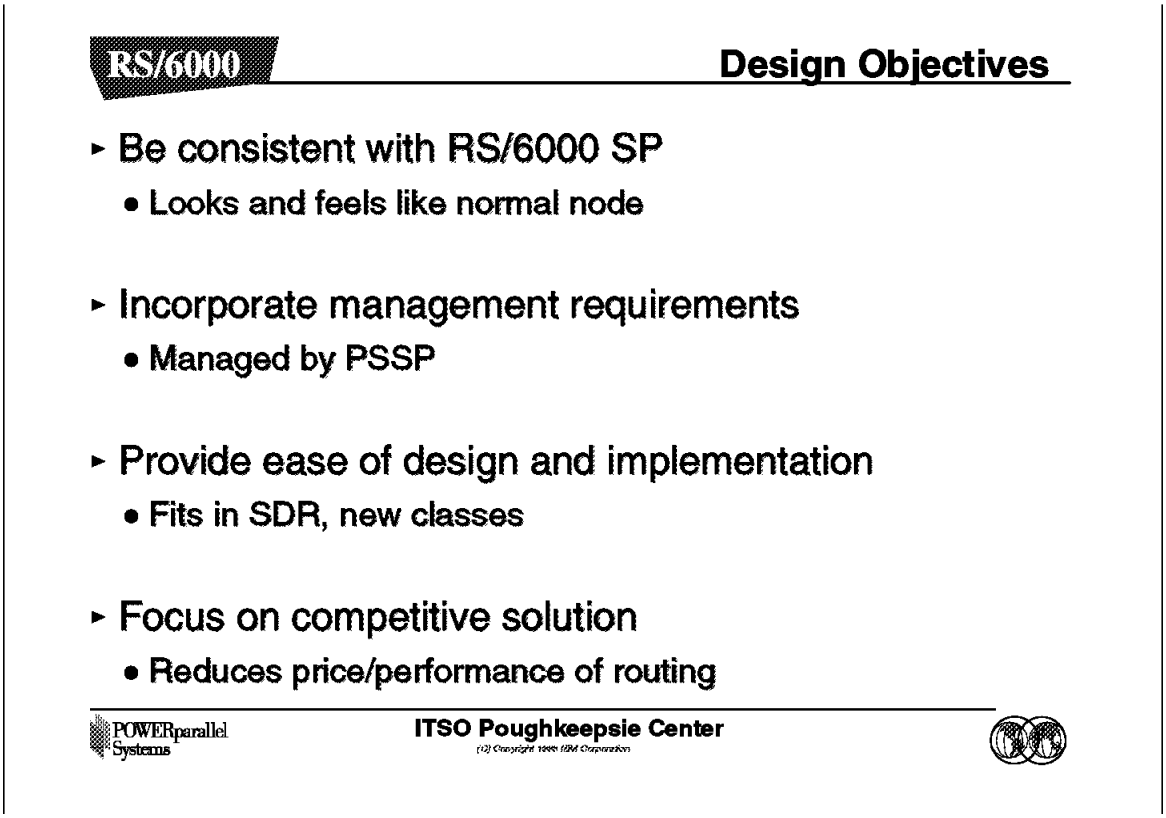
Note: In the remainder of this book, we refer to the SP Switch router as the GRF.

The distinguishing feature of the GRF, when compared with other routers, is that it has an SP Switch router adapter and therefore can connect directly into the SP Switch.

The RS/6000 SP software treats this adapter as an extension node. It is a node, because it takes up one port in the SP Switch and is assigned a node number. It is described as an extension, because it is not a standard RS/6000 SP node, but an adapter card that extends the scope of the RS/6000 SP.

Although the term *extension node* represents the node appearance of the adapter, it does not define the connection. An *extension node adapter* is used for that purpose. Each extension node has an extension node adapter to represent its connection to the SP Switch.

6.1.2 Design Objectives



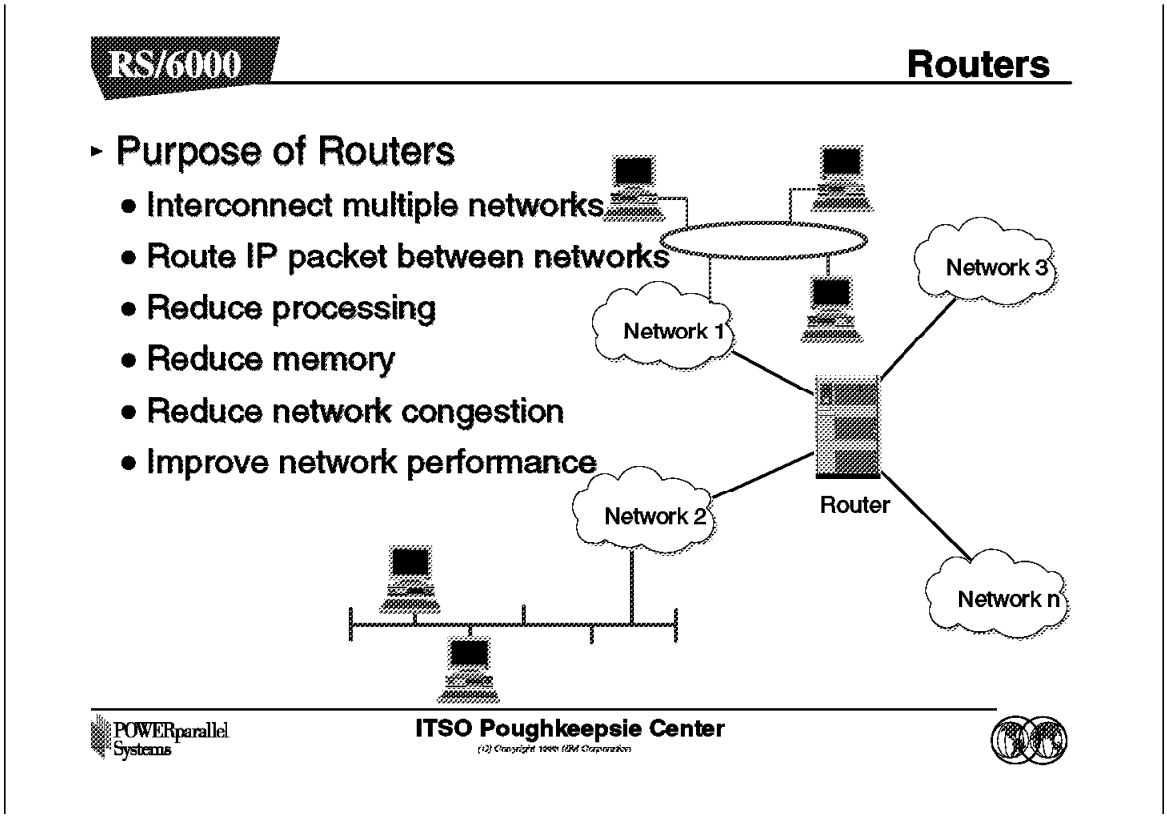
Because the dependent node is part of the RS/6000 SP, it had to be packaged and assign some roles consistent with other RS/6000 SP nodes.

Changes were be made to the RS/6000 SP to incorporate management requirements for the dependent node.

Ease of design and implementation were important factors in the design of its support. These were accomplished by limiting the amount of switch-control protocol for the dependent node.

New SDR classes were created to manage dependent nodes. This was done to minimize the scope of the change and the exposure to side effects that dependent nodes may cause if they were represented as standard nodes in the SDR.

6.1.3 What is a Router?



Routers serve a unique purpose in the world of networks. They interconnect networks so that Internet Protocol (IP) traffic can be routed between the systems in the networks.

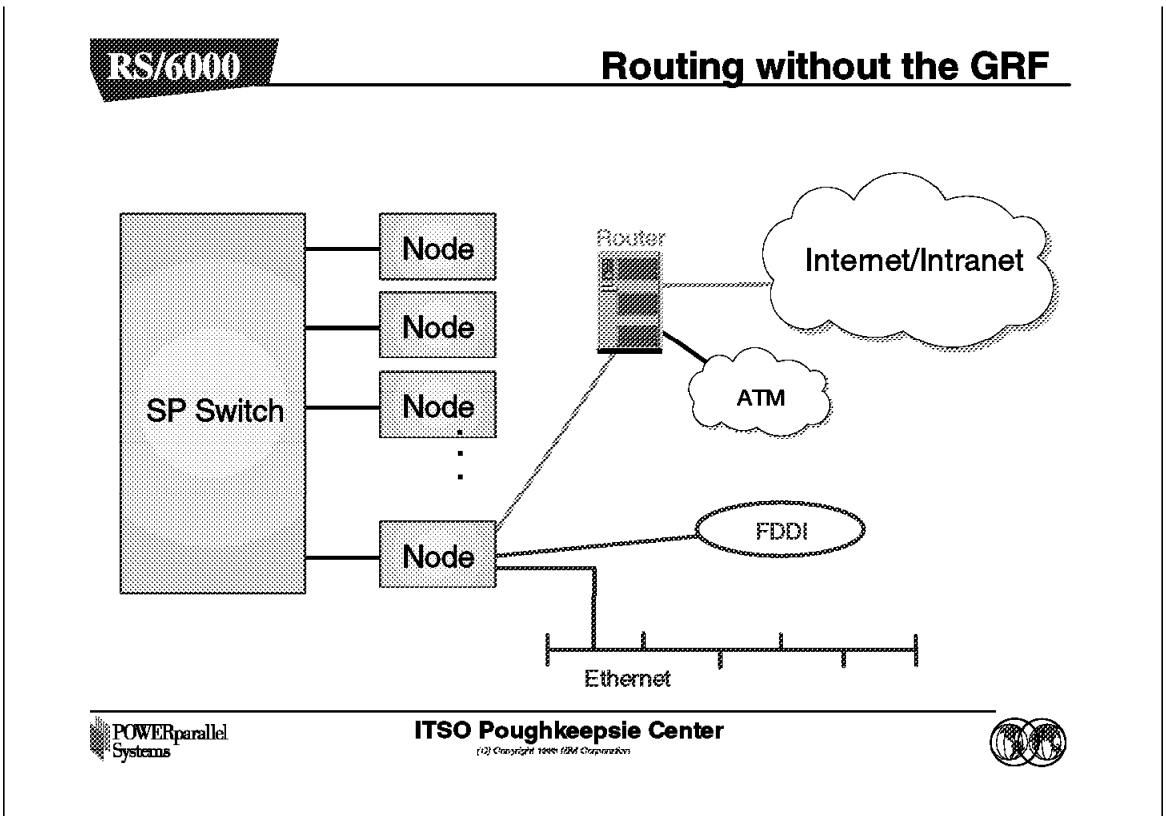
Routers help to reduce the amount of processing required on local systems, since they perform the computation of routes to remote systems. For example, a system can communicate with a remote system not in the local network by passing the message (or packets) to the router. The router works out how to get to the remote system and forwards the message appropriately.

Storing routes on the system takes up memory. But because it does not have to store routes to systems not in its own subnet, the route table uses less storage space, and thereby frees up memory for other work.

The use of routing reduces network traffic, because routers encourage subnetting, and subnetting creates a smaller network of systems. By having smaller networks, network traffic congestion is reduced and overall network performance is improved.

Benefits of reduced network congestion are better network traffic control and improved network performance.

6.1.4 Routing without the GRF



Before the GRF was available, there were only two ways for IP traffic from remote systems to reach the RS/6000 SP nodes:

1. You could put an additional IP adapter into every RS/6000 SP node.
2. You could designate one or two nodes to act as a router (as shown in this figure).

The first case was usually not chosen because of the cost involved. The following points explain why this option is expensive:

- Purchasing multiple IP adapters for each RS/6000 SP node can be expensive.
- The number of I/O slots in the RS/6000 SP node is limited. In addition, these slots are required to perform other tasks for the system, such as connecting to disk or tape. Using these I/O slots to connect IP adapters restricts the functions of the RS/6000 SP node.

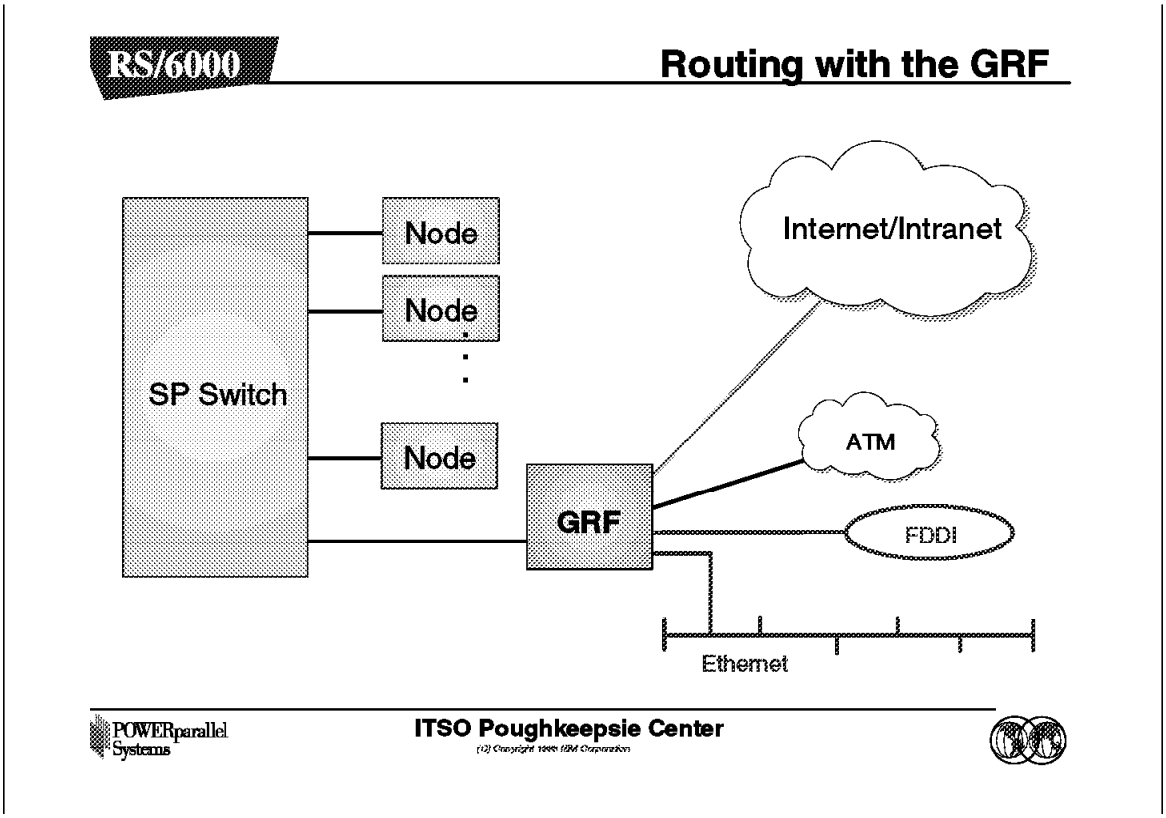
The second case has proven to be very expensive as well. The RS/6000 SP node was not designed for routing. It is not a cost-effective way to route traffic for the following reasons:

- It takes many CPU cycles to process routing. The CPU is not a dedicated router and is very inefficient when used to route IP traffic (this processing can result in usage of up to 90%).

- It takes a lot of memory to store route tables. The memory on the RS/6000 SP node is typically more expensive than router memory.
- The system I/O bus in the RS/6000 SP node is limited. The CPU on a node can only drive it at less than 80MB per second, which is less than what a high-end router can do.

For these reasons, the performance of routers in handling IP traffic from remote systems to the RS/6000 SP nodes was limited.

6.1.5 Routing with the GRF



The GRF is a dedicated, high-performance router. Each SP Switch router adapter can route up to 30,000 packets per second and up to 100MB per second into the SP Switch network.

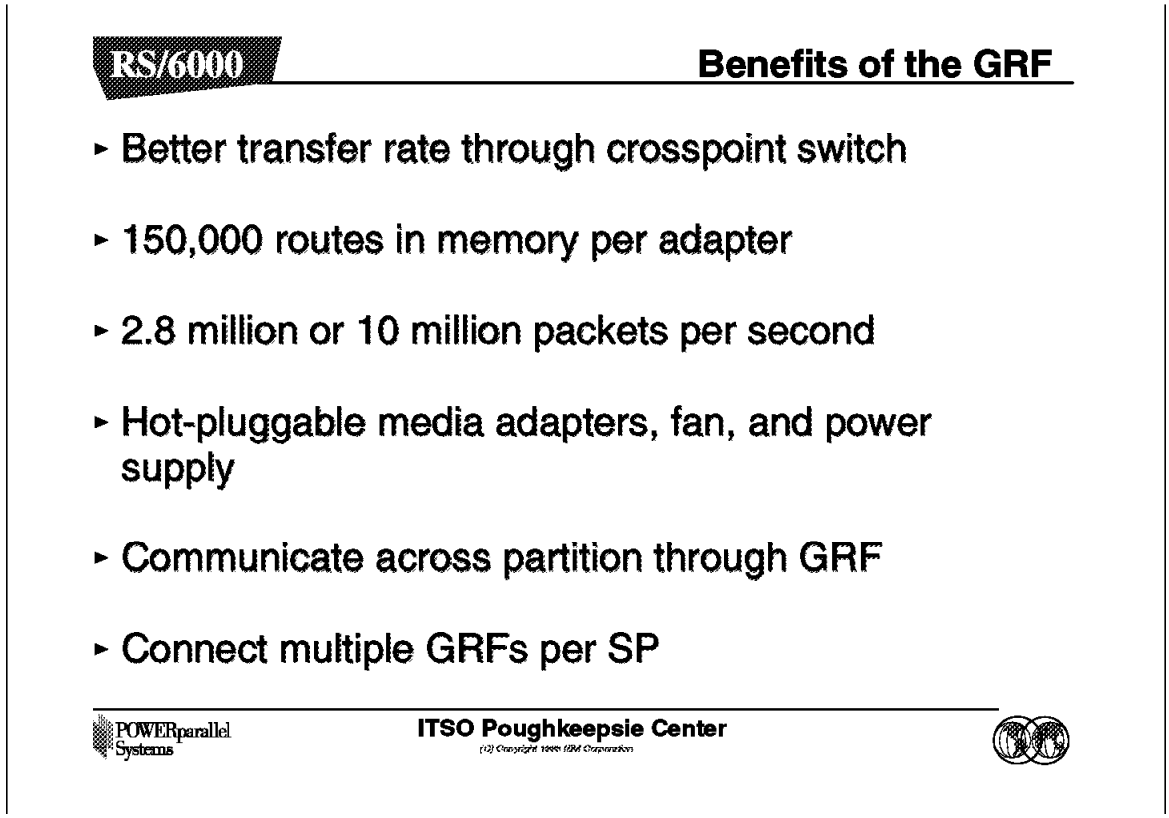
The GRF uses a crosspoint switch instead of an I/O bus to interconnect its adapters. This switch is capable of 4 to 16Gb per second and gives better performance than the MCA bus. Due to the high bandwidth that is available, communication between media adapters is improved.

Other advantages of using GRF are as follows:

- Availability of a redundant power supply
- Availability of a redundant fan
- Availability of a hot-swappable power supply
- Availability of a hot-swappable fan
- Availability of hot-swappable media adapters (to connect to networks)
- Scalability of up to 4 or 16 media adapters, depending on the GRF model

Perhaps the greatest advantage of using the GRF is improved price/performance. As previously mentioned, the GRF is a dedicated router, and as such it is much more cost effective to route IP traffic to the RS/6000 SP nodes than another RS/6000 SP node in many high network throughput configurations.

6.1.6 Benefits of the GRF



The crosspoint switch is a *non-blocking crossbar*. This architecture is faster than an RS/6000 SP node, in which media adapters communicate through a microchannel bus.

To take advantage of the fast I/O provided by the crosspoint switch, fast route table access time is required. The GRF can store up to 150,000 routes in memory, while an RS/6000 SP node can store only hundreds. This means that the GRF is able to retrieve a route faster than an RS/6000 SP node.

The GRF is able to route up to 2.8 million packets per second for the 4-slot model and 10 million packets per second for the 16-slot model.

All the media adapters on the GRF are hot-pluggable. This differs from using an RS/6000 SP node as your router. Should any network adapter on the RS/6000 SP node fail, the node has to be brought down to replace the faulty adapter. As a result, other unaffected network adapters will be brought down as well. The effect of bringing down the router will impact all the networks in the location.

Each RS/6000 SP is allowed to connect to multiple SP Switch Router Adapters, and it does not matter if these adapters are on different GRFs. Connecting multiple SP Switch router adapters to either different partitions in an RS/6000 SP or to different RS/6000 SPs allows them to communicate both with each other and with the other GRF media adapters via the SP Switch. A more detailed

discussion of this is found in the Coexistence figure in 6.4, “PSSP Enhancements” on page 177.

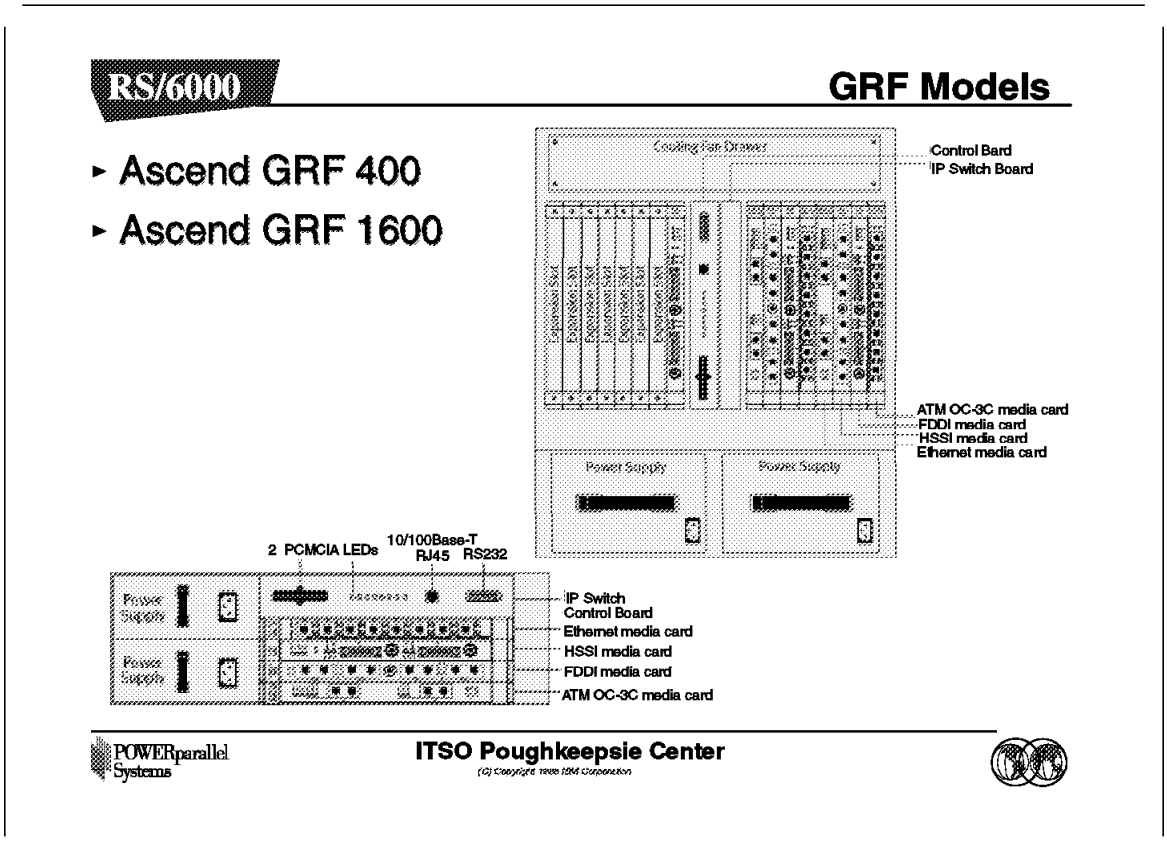
Attention

The SP Switch Router model 04S can support four media cards such as FDDI or ATM. The SP Switch Router model 16S can support 16. In either case, multiple SP Switch router adapters may be installed in the SP Switch Router. Check the final version of the SP product documentation to determine the maximum number of SP Switch router adapters supported in each model.

Note: The number of packets that the GRF can route per second depends on the following:

- The type of media adapter
- The size of the packet

6.2 GRF Modules



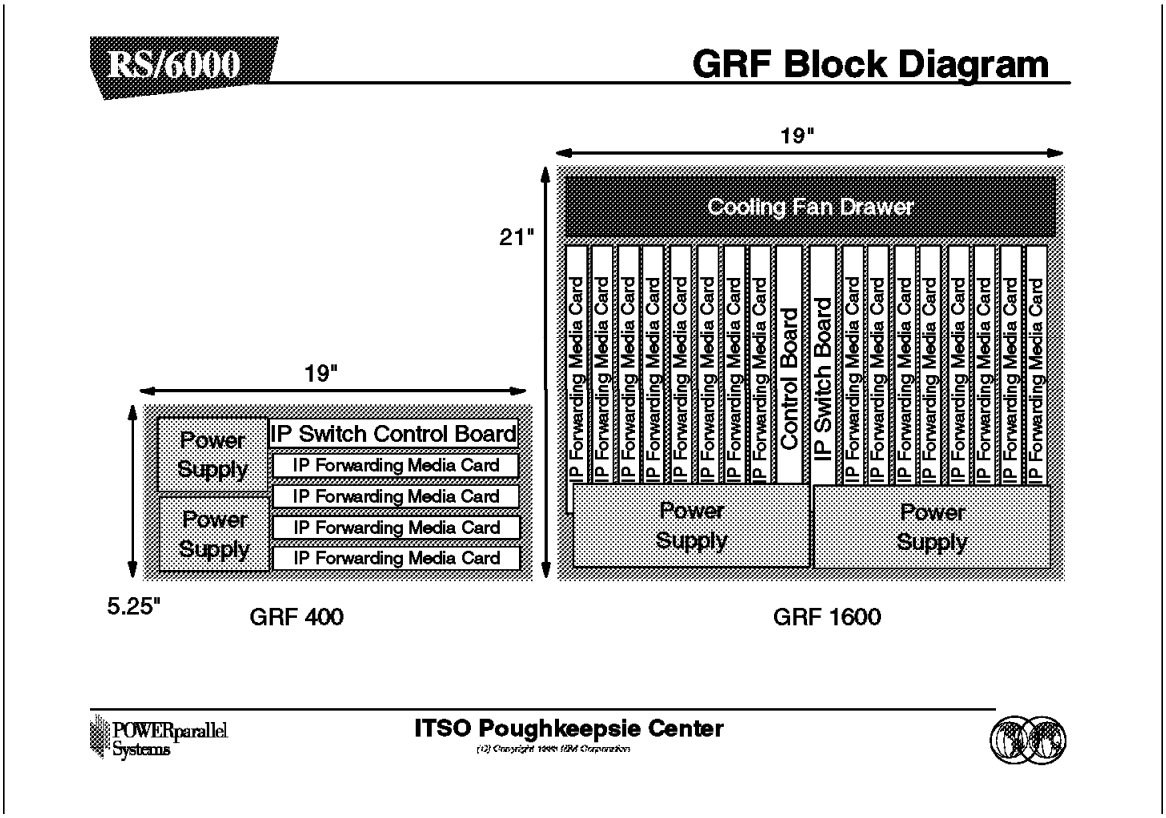
The GRF 400 can accommodate up to four media adapters.

The GRF 1600 can accommodate up to 16 media adapters.

Each adapter allows the GRF to connect to one or more networks.

Each of the models has an additional slot for the IP Switch Control Board, which is used to control the router.

6.2.1 GRF Block Diagram



This figure shows the two GRF models: the 4-slot and the 16-slot model. Detailed descriptions of each follow.

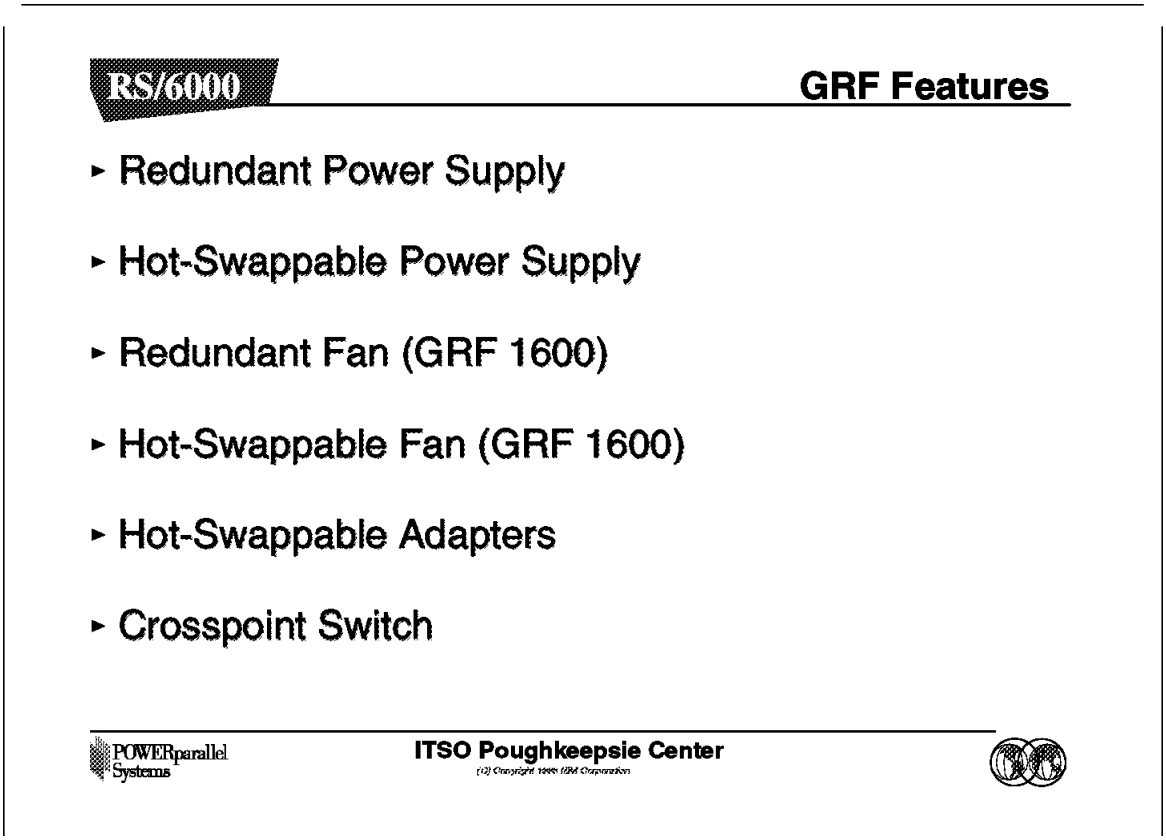
GRF 400

Part	Description
Cooling Fans	These are located at the right side of the chassis and cannot be accessed without bringing down the GRF. There is no redundant fan built into this model, and since the fans can only be accessed by bringing down the GRF, this model is *not* hot-swappable.
Media Cards	There are four media card slots on this chassis. They are slotted horizontally and are located at the bottom of the chassis.
IP Switch Control Board	The IP Switch Control Board is located at the top of the four media slots and is also slotted horizontally.
Power Supply	The left side of the chassis is reserved for the two power supplies that are required for redundancy. The failed power supply can be hot-swapped out of the GRF chassis. The second power supply is optional for this model.

GRF 1600

Part	Description
Cooling Fans	These are located at the top of the chassis, and can be accessed separately from the other parts of the GRF. The redundant fans built into the system are therefore hot-swappable.
Media Cards	There are 16 media card slots on this chassis. They are slotted vertically. Eight of the cards are on the left side of the chassis, and eight are on the right side.
IP Switch Control Board	The IP Switch Control Board is located in the middle of the 16 media slots and is also slotted vertically.
Power Supply	The base of the chassis is reserved for the two power supplies that are required for redundancy. The failed power supply can be hot-swapped out of the GRF chassis.

6.2.2 GRF Features



GRF has the following features:

- Redundant power supply

  Should any power supply fail, a message is sent to the control board. The power supply will automatically reduce its output voltage if the temperature exceeds 90°C (194°F). If the voltage falls below 180V, the GRF will automatically shut down.

- Hot-swappable power supply

  The faulty power supply can be replaced while the GRF is in operation.

- Redundant fan

  For the GRF 1600 model, if one fan breaks down, a message is sent to the control board.

  For both models, when the temperature reaches 53°C (128°F), an audible alarm sounds continuously, and a message is sent to the console and logged into the message log.

  If the temperature exceeds 57.5°C (137°F), the GRF will do an automatic system shutdown.

- Hot-swappable fan

  For the GRF 1600 model, the cooling fan can be replaced while the GRF is in operation.

- Hot-swappable adapters

  There are two types of adapters on the GRF: the media adapters and the IP Switch Control Board.

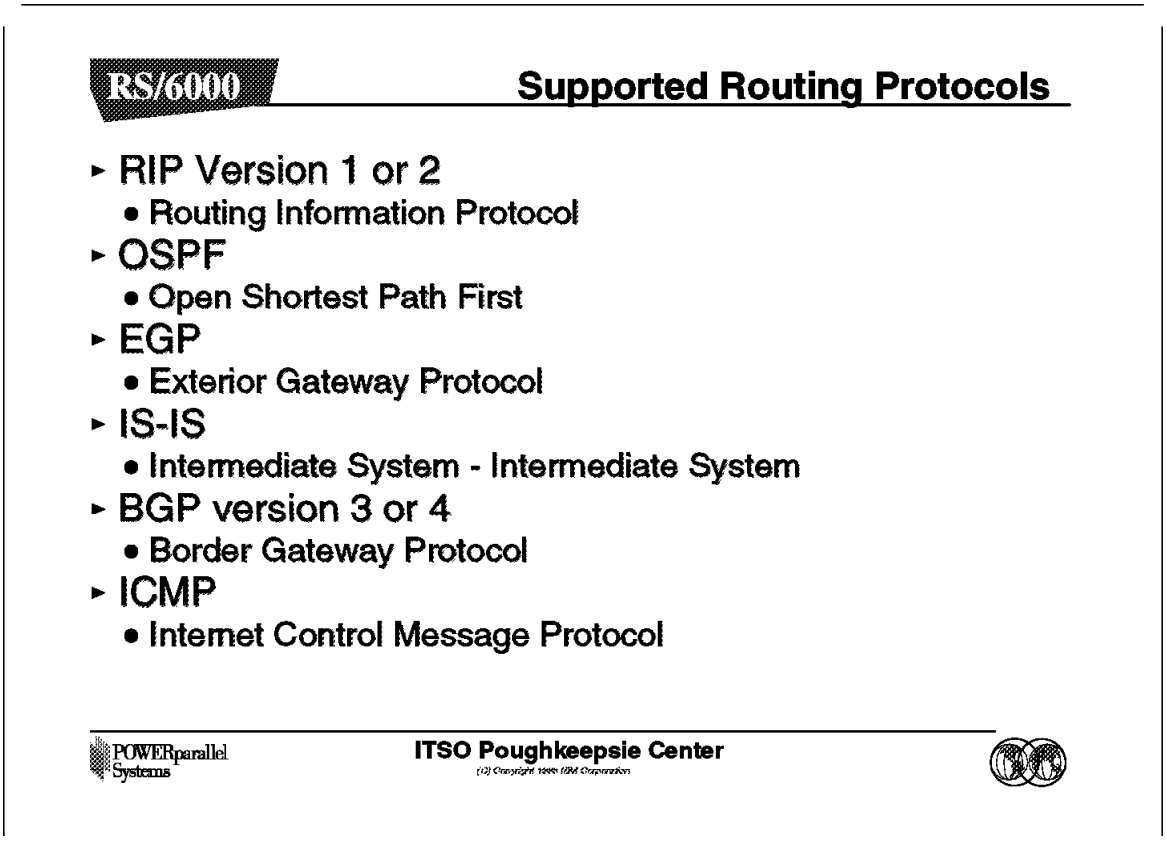
  The media adapters are independent of each other, and can be replaced or removed without affecting any other adapter or the operation of the GRF.

  However, the IP Switch Control Board is critical to the GRF. Should this board be unavailable, the router will fail.

- Crosspoint switch

  The crosspoint switch is a 16x16 (16Gb per second) or 4x4 (4Gb per second) crossbar switch for the GRF 1600 and GRF 400, respectively. It is the I/O path used when the media adapters need to communicate with each other.

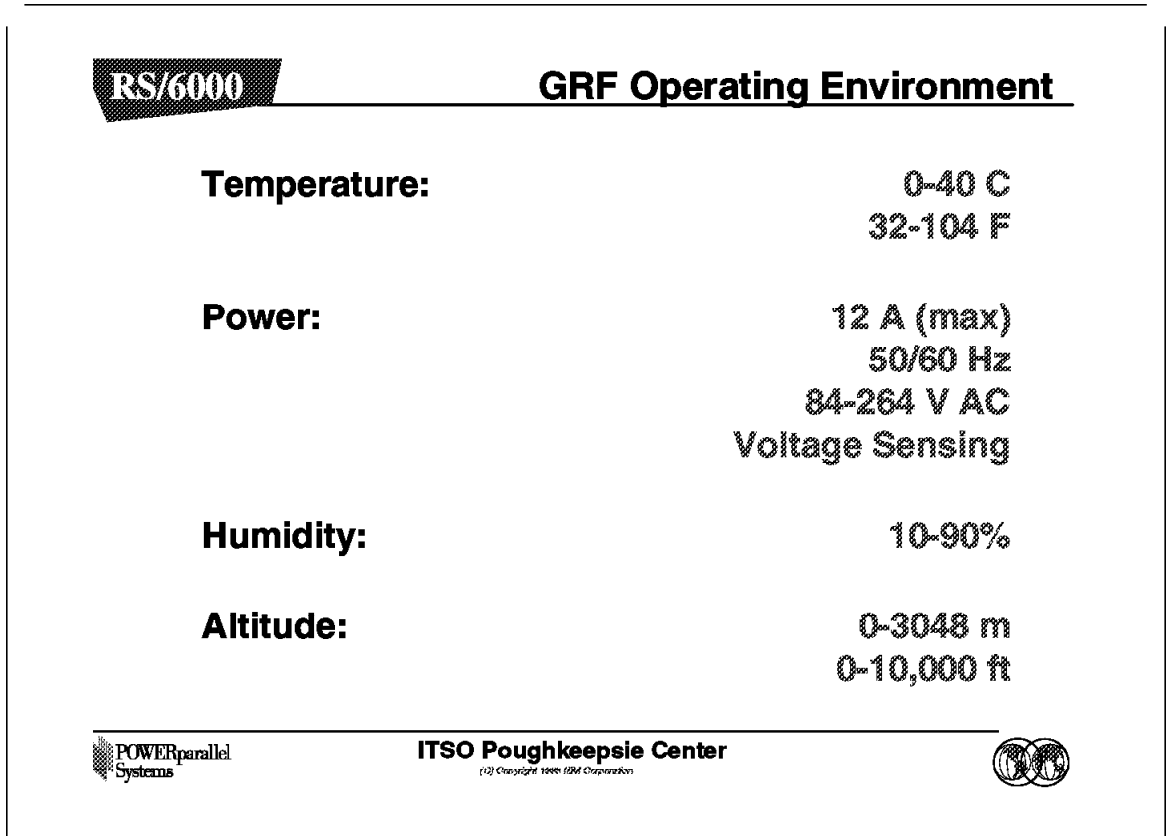
6.2.3 Routing Protocols



In addition to static routes, various routing protocols are available on the GRF, as follows:

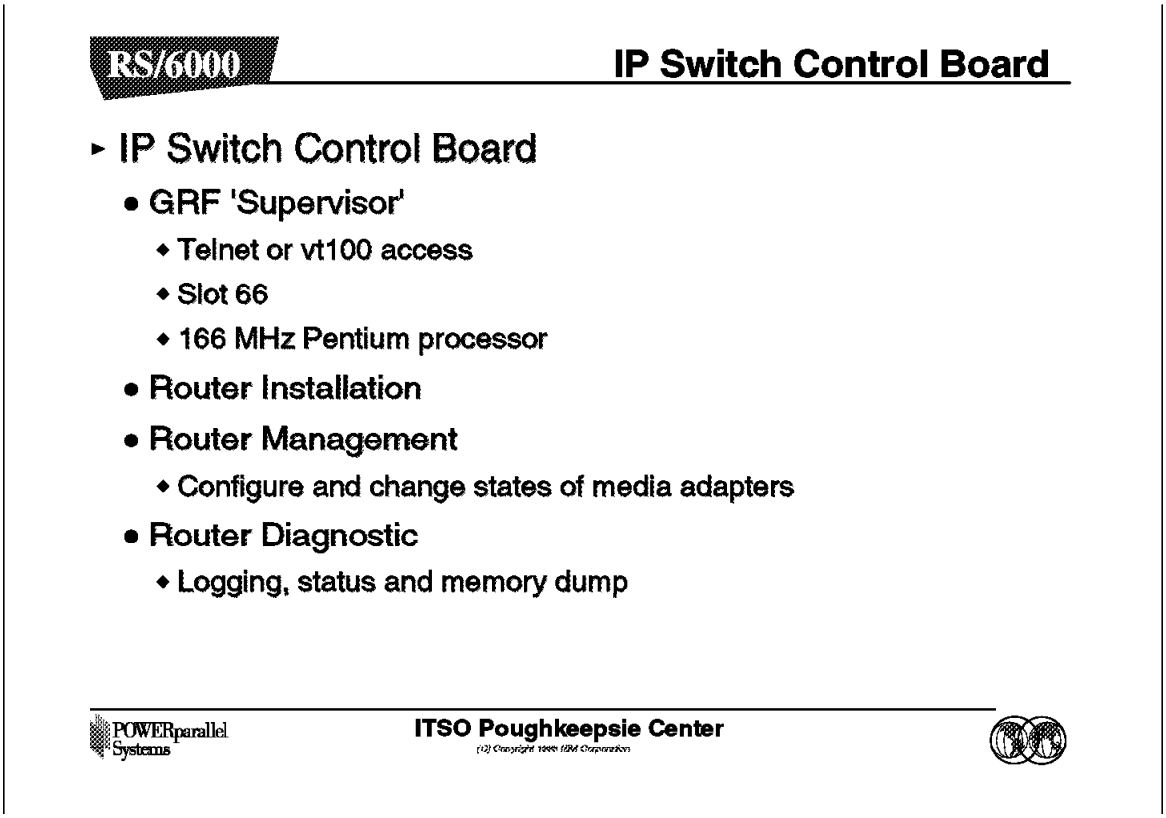
RIP	Routing Information Protocol Version 1 or 2 (RIP 1 or 2)
OSPF	Open Shortest Path First
EGP	Exterior Gateway Protocol
IS-IS	Intermediate System to Intermediate System (an OSI gateway protocol)
BGP	Border Gateway Protocol Version 3 or 4 (BGP 3 or 4)
ICMP	Internet Control Message Protocol

6.2.4 GRF Operating Environment



As previously mentioned, the operating temperature should not exceed 53°C (128°F). Even though there is a buffer between the operating temperature and the warning temperature, it is best to keep the temperature within the operating level in order to minimize the possibility of damage to GRF components.

6.2.5 IP Switch Control Board



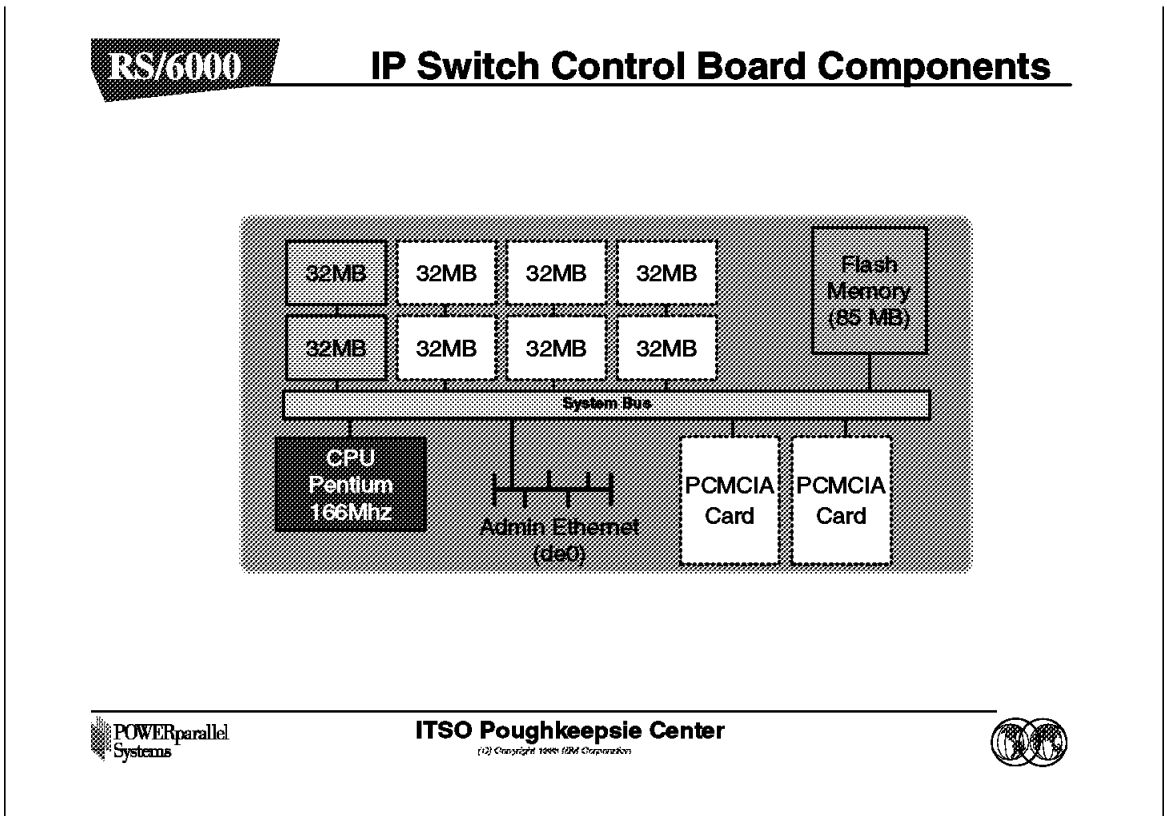
The control board, also known as the IP Switch Control Board, is accessed through Telnet or a locally attached VT100 terminal. The IP Switch Control Board is supplied with the GRF and is necessary for its operation. The VT100 terminal is not supplied with the GRF. It is only required for the installation of the GRF. After installation, all future access to the GRF is through Telnet to the IP Switch Control Board's administrative Ethernet.

The IP Switch Control Board is identified as slot 66 in the GRF. The CPU in the IP Switch Control Board is a 166MHz Pentium processor and runs a variant of BSD UNIX as its operating system. Thus, the GRF administrator is assumed to be proficient in UNIX.

The IP Switch Control Board is used to install, boot, and configure the router and its media adapters.

It is also used for the logging of messages, the dumping of memory and status, and to perform diagnostic checking of both the GRF and the media adapters.

6.2.6 IP Switch Control Board Components



Let us examine the IP Switch Control Board in more detail.

Following are descriptions of its components as shown in the figure:

Item	Description
Memory	The IP Switch Control Board comes standard with 64MB of memory (the two shaded blocks of 32MB of memory in the upper left corner). The IP Switch Control Board memory can be upgraded to 256MB, in increments s of 64MB (the six white blocks of memory). Each column of 64MB of memory is split into two parts. The system uses the bottom half of the memory (32MB) for file system storage. The top half is used for applications such as the SNMP agent, the gated daemon, and for the operating system.
Flash memory	This memory (the 85MB ATA flash memory on the system) is used to store the operating system information and the configuration information for the GRF.

System bus Used by the IP Switch Control Board components to communicate with each other.

Pentium processor This 166MHz processor drives the IP Switch Control Board and the GRF. As previously mentioned, this processor runs a variant of BSD UNIX, and so it is useful for the GRF administrator to have UNIX management skills.

Administrative Ethernet This Ethernet is known to the GRF as de0. This port supports the 10BaseT or the 100BaseT Ethernets, and switches between them automatically, depending on the type of network used.

To use 10Base2 or 10Base5, the user must add a transceiver (supplied by the user).

PCMCIA cards The two white blocks at the bottom right corner of the figure are PCMCIA slots.

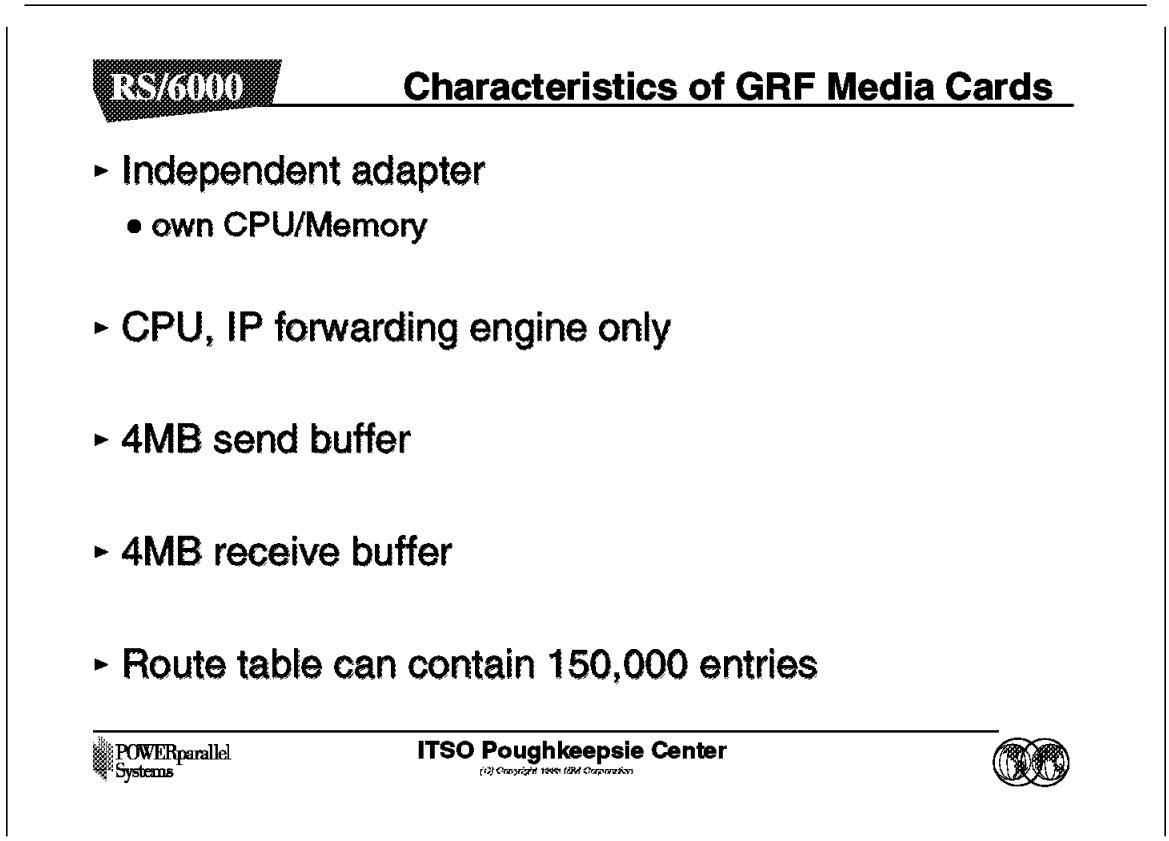
There are two types of PCMCIA cards:

- The PCMCIA 85MB flash memory card, available as an optional device, is used to back up the system. It is similar to a tape drive on a normal system.
- The PCMCIA modem card, also available as an optional device, allows the user to dial into the GRF through a modem to administer it remotely.

  Note: For the initial setup, the console must be available locally, not through the modem.

Additionally, the RS232 port (which is not shown in the figure) allows you to connect the VT100 console by using an RS232 null modem cable. The console and cable must be supplied by the user.

6.3 Characteristics of GRF Media Cards



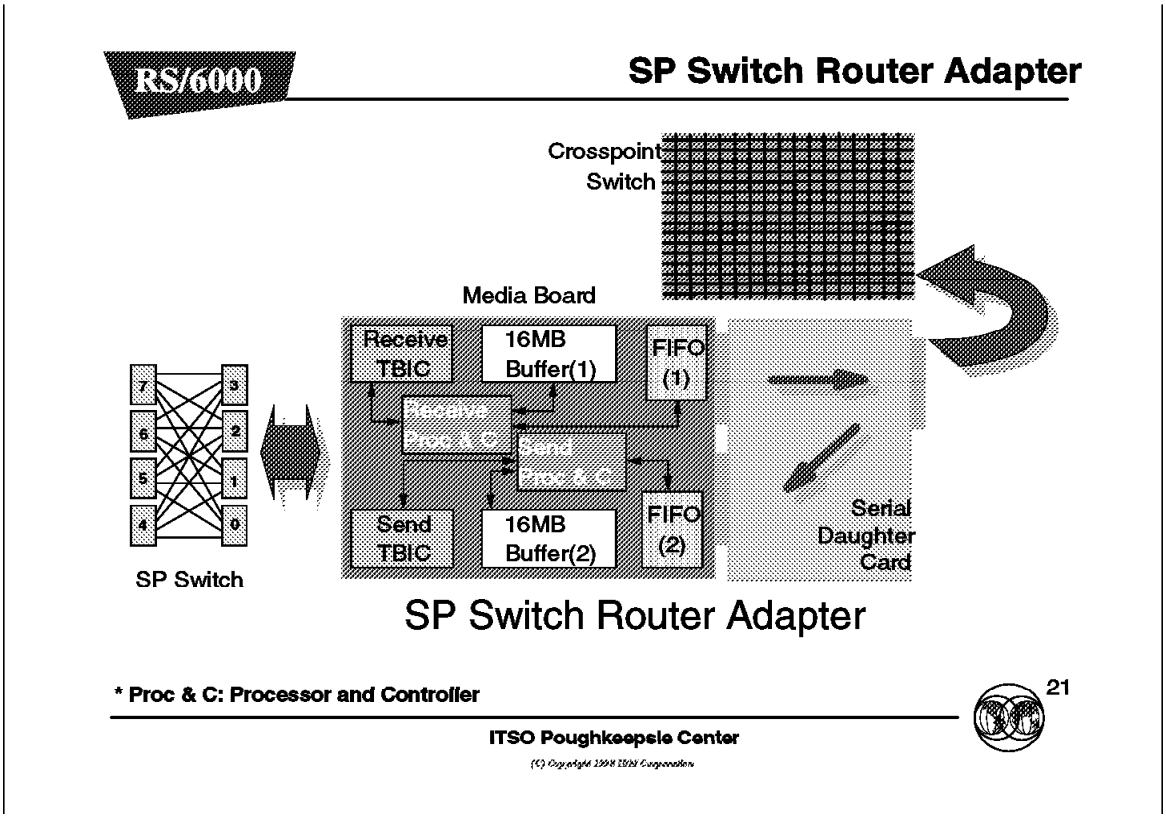
All GRF media cards (media adapters) are self-contained and independent of other media adapters.

Each media card has an onboard processor that is responsible for IP forwarding on the media adapter.

Each media card has two independent memory buffers, a 4MB send buffer and a 4MB receive buffer. These buffers are necessary to balance the speed differences between the media adapters, because they have different transfer rates.

Each onboard processor has local memory that can contain a local route table with up to 150,000 entries, to be used for routing on the media adapter. Because these route entries are in local memory, access to them is very fast. When the media adapter is started up, it gets its initial route entries from the IP Switch Control Board.

6.3.1 SP Switch Router Adapter



The GRF supports a number of media adapters. This figure describes the SP Switch router adapter in detail. This adapter allows the GRF to connect directly into the SP Switch.

The SP Switch router adapter is made up of two parts:

- The media board
- A serial daughter card

The serial daughter card is an interface for the media board into the crosspoint switch. This switch is the medium by which the GRF (media) adapters talk to each other.

The purpose of the media board is to route IP packets to their intended destination through the GRF. The SP Switch router adapter described here is used for routing IP packets to and from the SP Switch to other systems connected directly or indirectly to the GRF. A brief description of the components on the media board follows.

Receive TBIC This component receives data segments from the SP Switch and notifies the Receive Controller and Processor that there is data to be transferred to the buffer.

Receive Controller and Processor

This component recognizes the SP Switch segments and assembles them into IP packets in the 16MB buffer. Up to 256 simultaneous IP datagrams can be handled simultaneously. When a complete IP packet has been received, the Receive Controller sends the packet to the FIFO (1) queue for transfer to the serial daugther card.

Buffer (1)

This component is segmented into 256 64KB IP packet buffers. It is used to reassemble IP packets before being sent to the FIFO queue, as switch data segments may arrive out of order and interleaved with segments belonging to different IP packets.

FIFO (1)

This component is used to transfer complete IP packets to the serial daughter card and even the flow of data between the SP and the GRF backplane.

FIFO (2)

This component receives IP packets from the serial daughter card and transfers them to the Buffer (2).

Buffer (2)

This buffer is used to temporarily store the IP packet while its IP address is examined and a proper SP Switch route is set up to transfer the packet through the SP Switch.

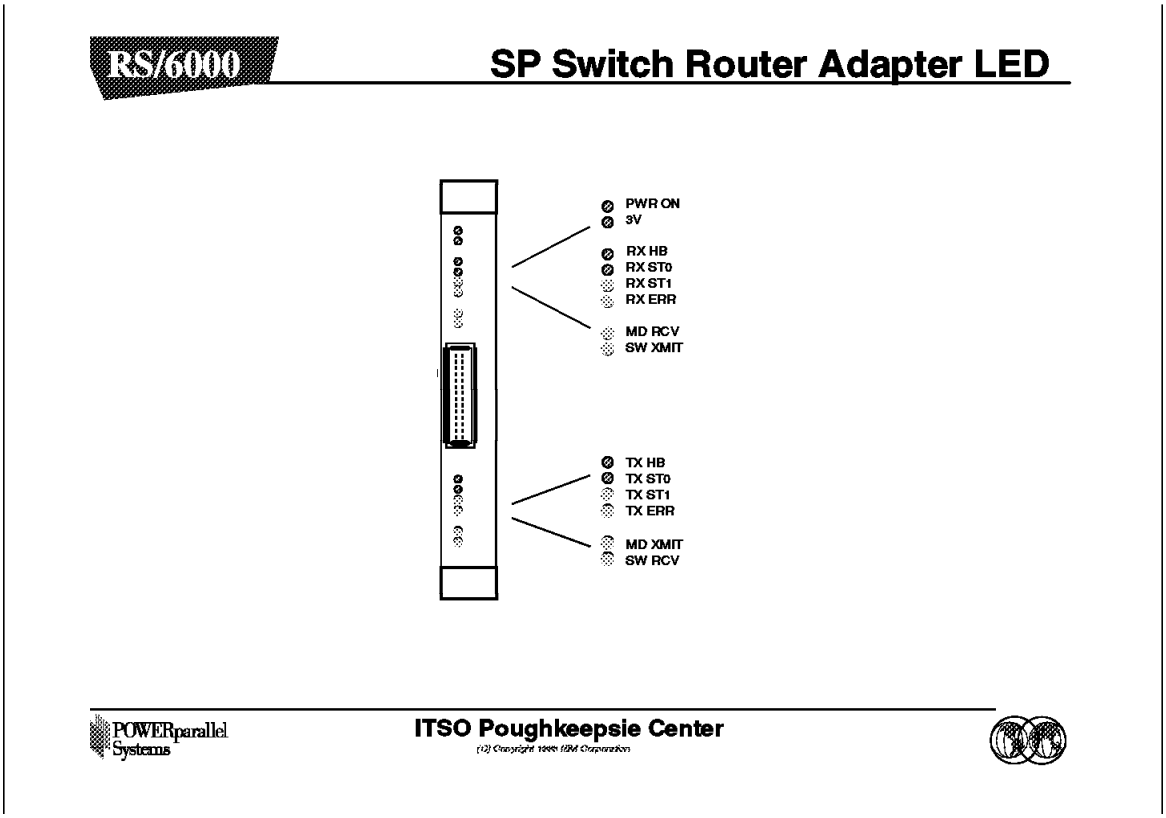
Send Processor and Controller

This component is notified when an IP packet is received in the FIFO (2) queue and sets up a DMA transfer to send the packet to Buffer (2). The Send Processor looks up the IP address in the packet header and determines the SP Switch route for the packet, before notifying the Send Controller to send the packet to the Send TBIC from Buffer (2).

Send TBIC

This component receives data from Buffer (2) and sends it in SP Switch data segments to the SP Switch.

6.3.2 SP Switch Router Adapter LED



LED activities during operations are listed in Table 1, Table 2 on page 173, and Table 3 on page 173.

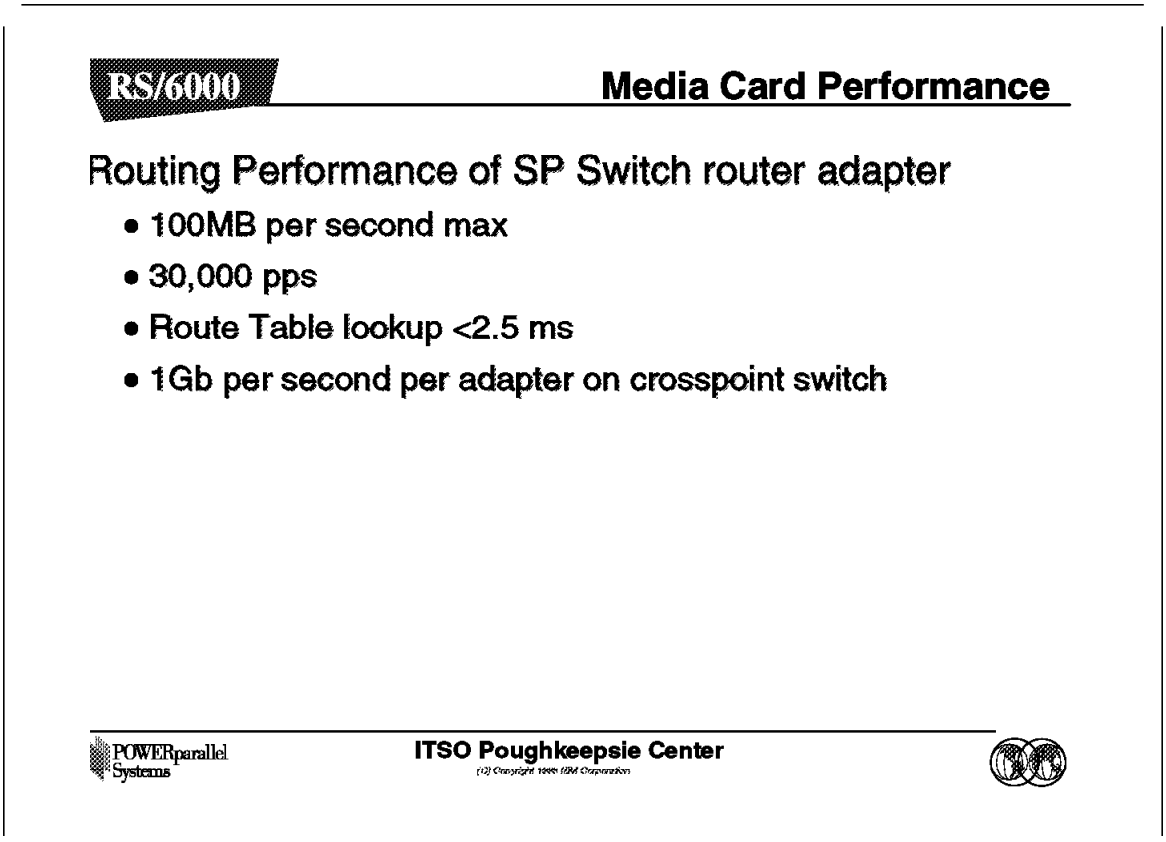
Table 1. SP Switch Router Adapter Media Card LEDs

LED	Description
PWR ON	This green LED is on when 5 volts are present.
3V	This green LED is on when 3 volts are present.
RX HB	This green LED blinks to show the heartbeat pattern for the receive side CPU.
MD RCV	This amber LED turns on when data is received from its media port (RS/6000 SP Switch).
SW XMIT	This amber LED turns on when data is sent to the crosspoint switch (through the serial daughter card).
TX HB	This green LED blinks to show the heartbeat pattern for the transmit side CPU.
MD XMIT	This amber LED turns on when data is transmitted from its media port (RS/6000 SP Switch).
SW RCV	This amber LED turns on when data is received from the crosspoint switch (through the serial daughter card).

Table 2. SP Switch Router Adapter Media Card LEDs - RX/TX			
RX/TX ST0 (green)	**RX/TX ST1 (amber)**	**RX/TX ERR (amber)**	**Description**
on	on	on	STATE_0 for hardware initialization.
off	on	on	STATE_1 for software initialization. Port waiting for configuration parameters.
on	off	on	STATE_2 for configuration parameters in place. Port waiting to be connected.
off	off	on	STATE_3 for port is connected and link is good. The media adapter is ready to be online.
on	off	on	STATE_4 for port is online and running/routing.

Table 3. SP Switch Router Adapter Media Card LEDs During Bootup				
RX/TX HB (green)	**RX/TX ST0 (green)**	**RX/TX ST1 (amber)**	**RX/TX ERR (amber)**	**Description**
on	on	on	on	All LEDs are lit for 0.5 seconds during reset as part of onboard diagnostics.
off	off	off	on	Error condition: checksum error is detected in flash memory.
on	off	on	off	Error condition: SRAM fails memory test.
on	off	off	on	During loading, HB & ST1 flash as each section of the code loads.

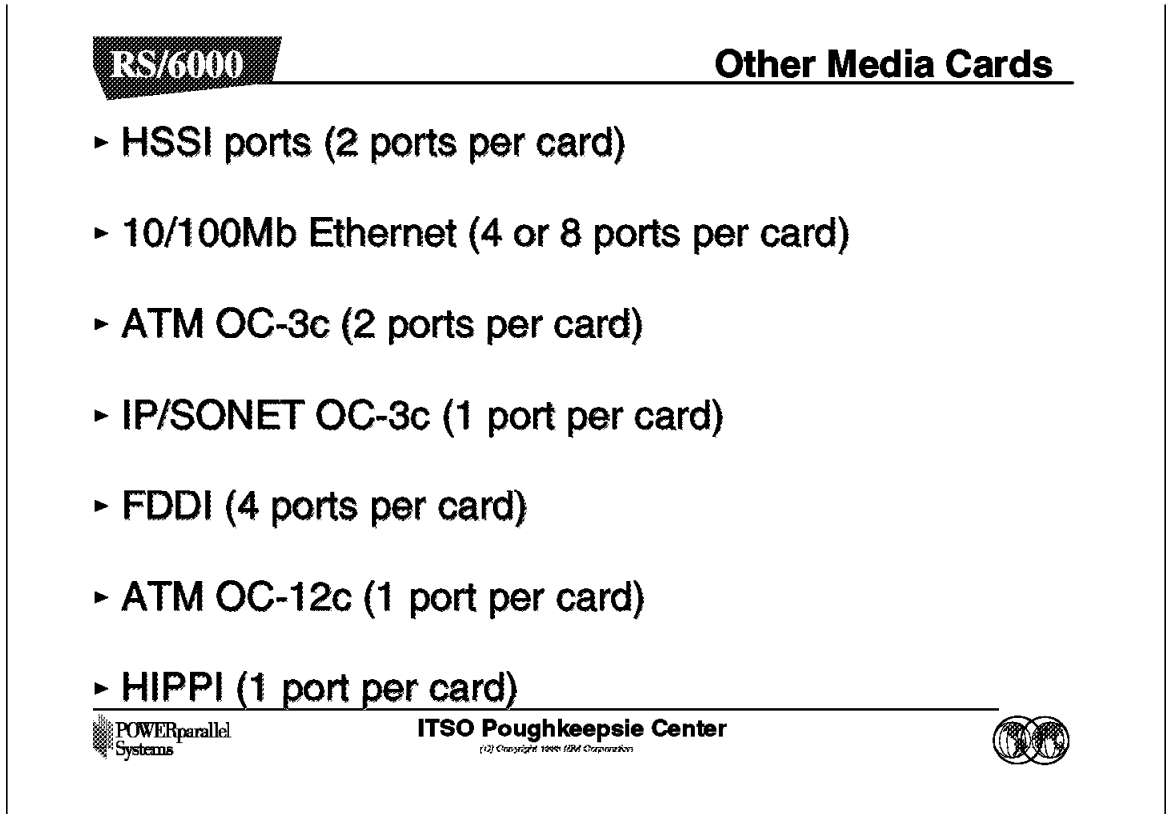
6.3.3 Media Card Performance



The SP Switch router adapter has the following performance characteristics:

- It is able to transfer up to 100MB per second. The limiting factor is the crosspoint switch connection bandwidth.
- It is able to transfer up to 30,000 packets per second. At 20,000 packets per second, each packet needs to be at 5KB in order to achieve the 100MB per second transfer rate mentioned.
- As previously mentioned, each adapter stores its own route tables in memory. Therefore, route table lookup is very fast, that is, less than 2.5 ms.
- Finally, each media adapter has a 1Gb per second dedicated link into the crosspoint switch. That is why the 4-port and 16-port models have an aggregate bandwidth of 4Gb and 16Gb per second, respectively, for the crosspoint switch.

6.3.4 Other Media Cards

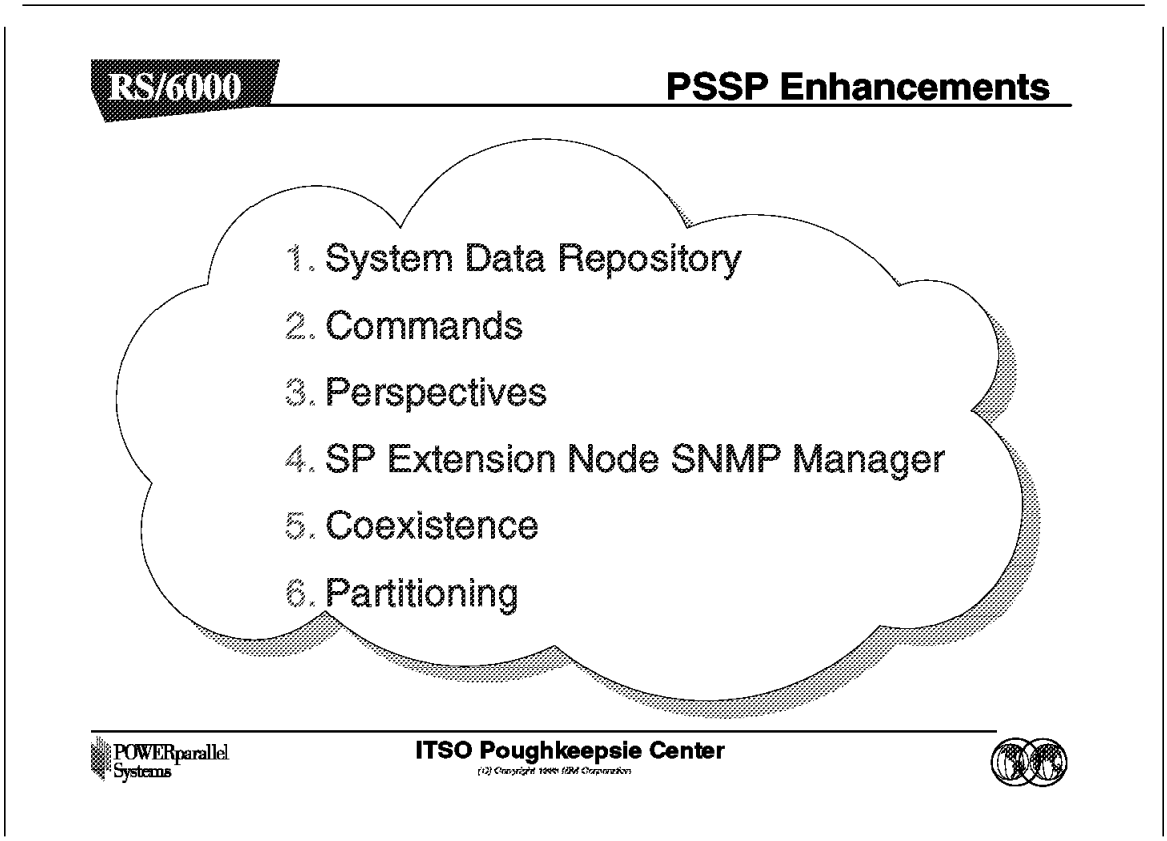


The following are other media cards and adapters currently supported on the GRF:

- The High Speed Serial Interface (HSSI) is a dual-ported media adapter that can connect to two serial networks simultaneously. Each port is capable of up to 45Mb per second.
- The 10/100Mb Ethernet media adapter consists of eight 10/100BaseT Ethernet ports. All ports support only utp cables. Other types of cables require the user to supply the appropriate transceivers.
- The ATM OC-3c media adapter allows the user to connect up to two connections into the ATM network at 155Mb per second.
- The IP/SONET OC-3c is a single-ported card that allows the user to connect to a digital network using a transmission format known as Synchronous Optical Network protocol (SONET). This standard is increasingly popular in the telecommunications industry.
- The FDDI media card provides four ports in the card. These ports allow the media card to be connected into the Fiber Distributed Data Interchange (FDDI). The four ports can be configured such that they support the following:
  - Two dual-ring FDDI networks
  - One dual-ring and two single-ring FDDI networks
  - Four single-ring FDDI networks

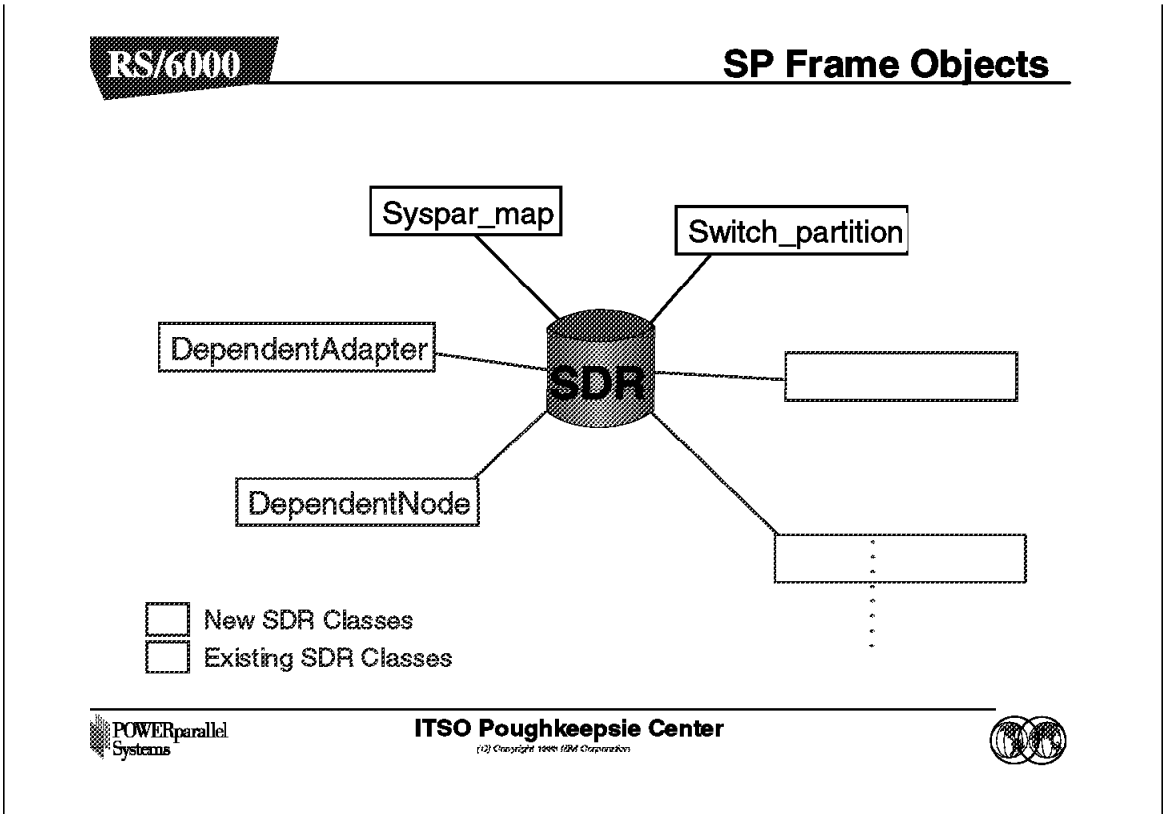
- The HIPPI media adapter is a single-port card that allows the GRF to connect to a High Performance Parallel Interface (HIPPI) network at speeds of up to 800 or 1600Mbits/sec. After deducting the overhead, this medium can support connections of up to 100MB/sec.

6.4 PSSP Enhancements



This section discusses the enhancements made to Parallel Systems Support Programs (PSSP) to accommodate the Dependent Node Architecture.

6.4.1 SP Frame Objects



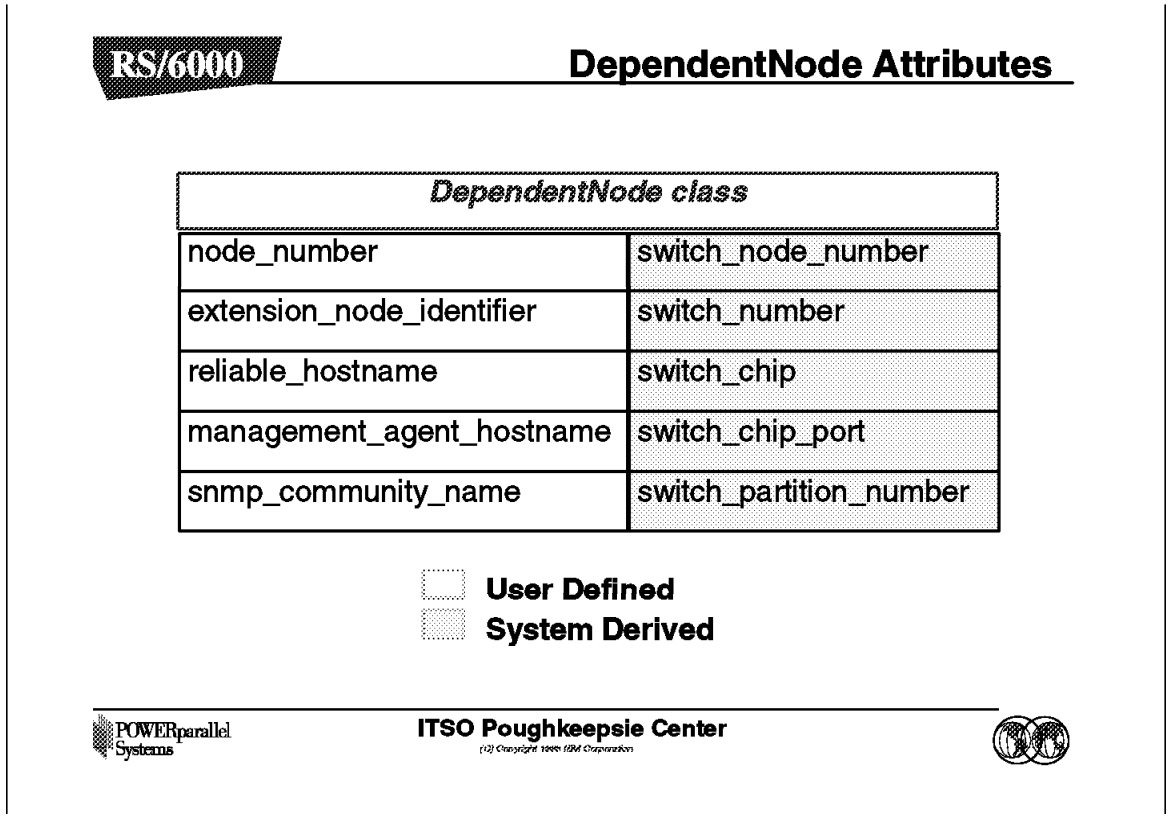
The following two classes have been *added* to the System Data Repository (SDR):

- DependentNode
- DependentAdapter

These classes are described in detail in the next two figures.

Also note that changes were made to the Syspar_map and Switch_partition classes, as described in the Additional Attributes figure.

6.4.2 DependentNode Attributes



This figure shows the attributes of the DependentNode class, which is described in detail as follows:

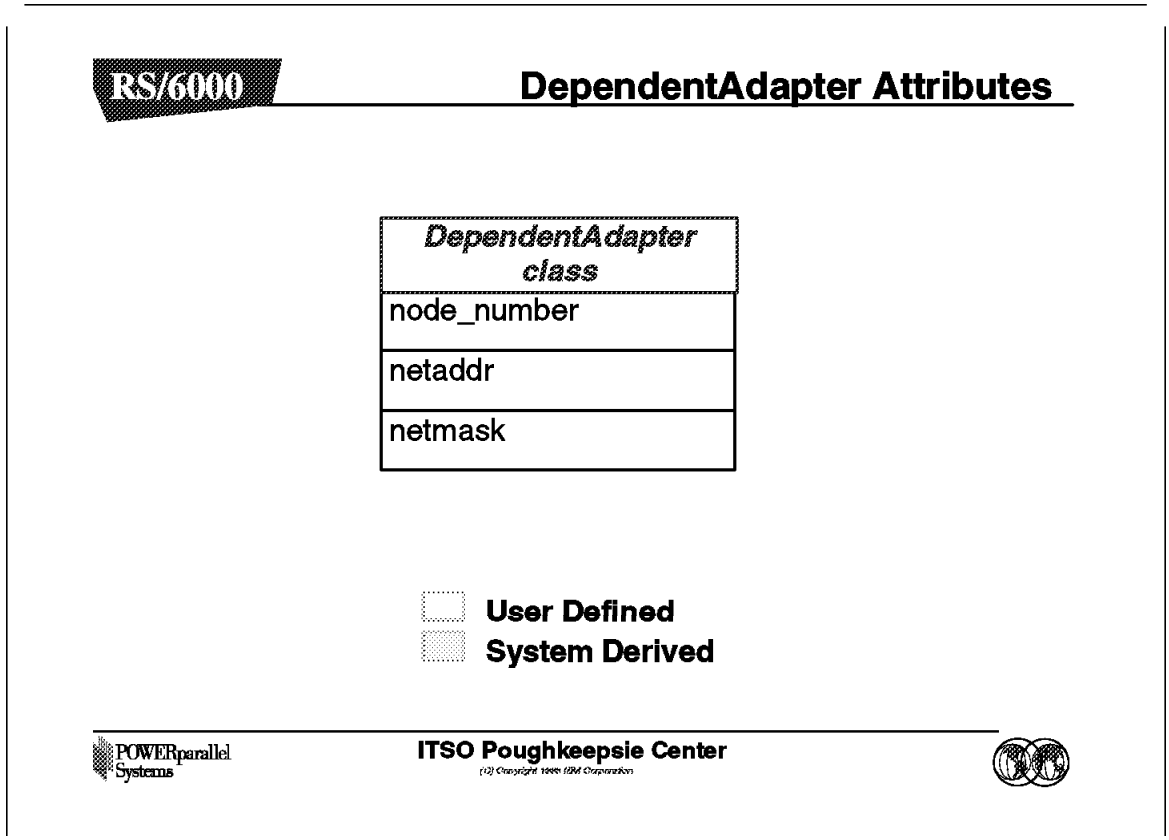
Attribute	Description
node_number	User-supplied node number representing the node position of an unused SP Switch port to be used for the SP Switch router adapter.
extension_node_identifier	This is a 2-digit slot number that the SP Switch router adapter occupies on the GRF. Its range is from 00 to 15.
reliable_hostname	The hostname of the administrative Ethernet, de0, is the GRF's hostname. Use the long version of the hostname when DNS is used.
management_agent_hostname	This attribute is the hostname of the SNMP agent for the GRF. For the GRF dependent node, this is the same as the reliable_hostname.
snmp_community_name	This field contains the SNMP community name that the SP Extension Node SNMP Manager and the GRF's SNMP Agent will send in the corresponding field of the SNMP messages. This value must match the value specified in

	the /etc/snmpd.conf file. If left blank, a default name found in the SP Switch router adapter documentation is used.

The following attributes are derived by the RS/6000 SP system when the SDR_config routine of endefnode is invoked.

Attribute	Description
switch_node_number	The switch port that the dependent node is attached to.
switch_number	The switch board that the dependent node is attached to.
switch_chip	The switch chip that the dependent node is attached to.
switch_chip_port	The switch chip port that the dependent node is attached to.
switch_partition_number	The partition number to which the dependent node belongs.

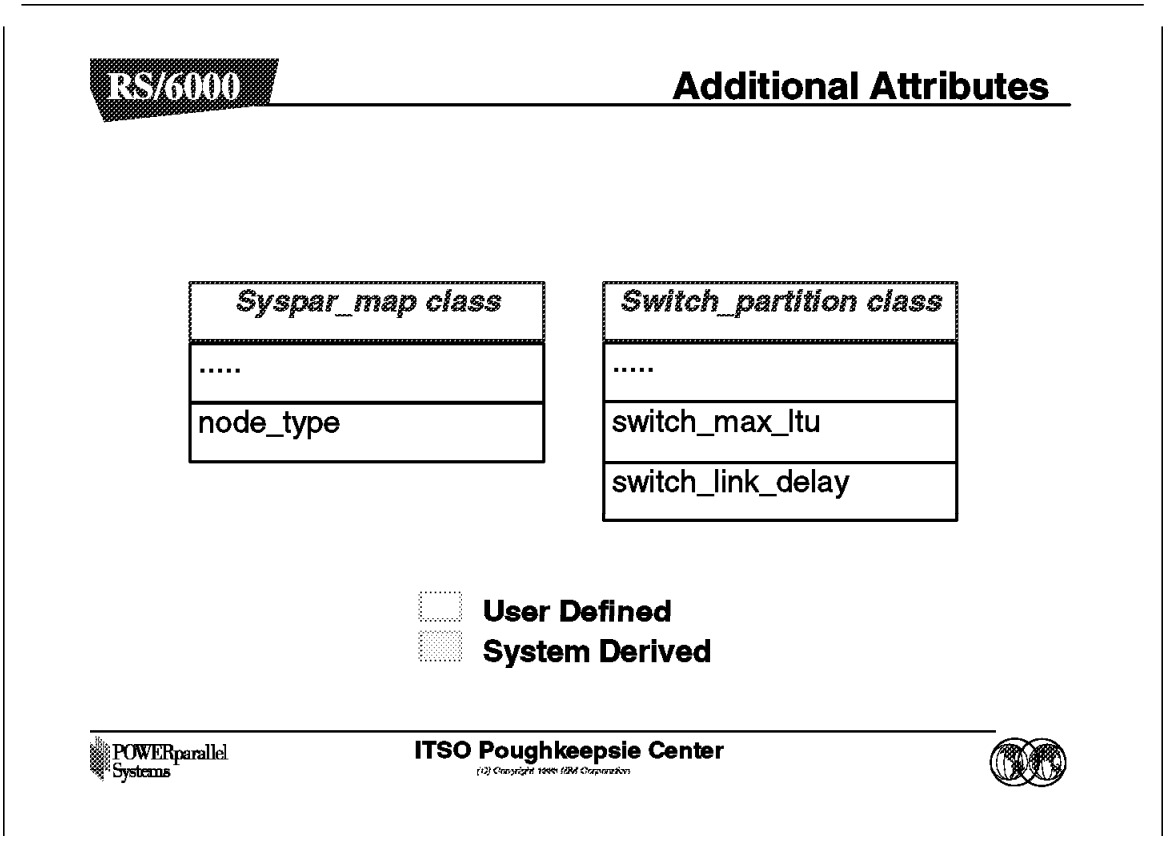
6.4.3 DependentAdapter Attributes



This figure shows the attributes of the DependentAdapter class, which is described in detail as follows:

Attribute	Description
node_number	User-supplied node number representing the node position of an unused SP Switch port to be used by the SP Switch router adapter.
netaddr	This is the IP address of the SP Switch Router Adapter.
netmask	This is the netmask of the SP Switch Router Adapter.

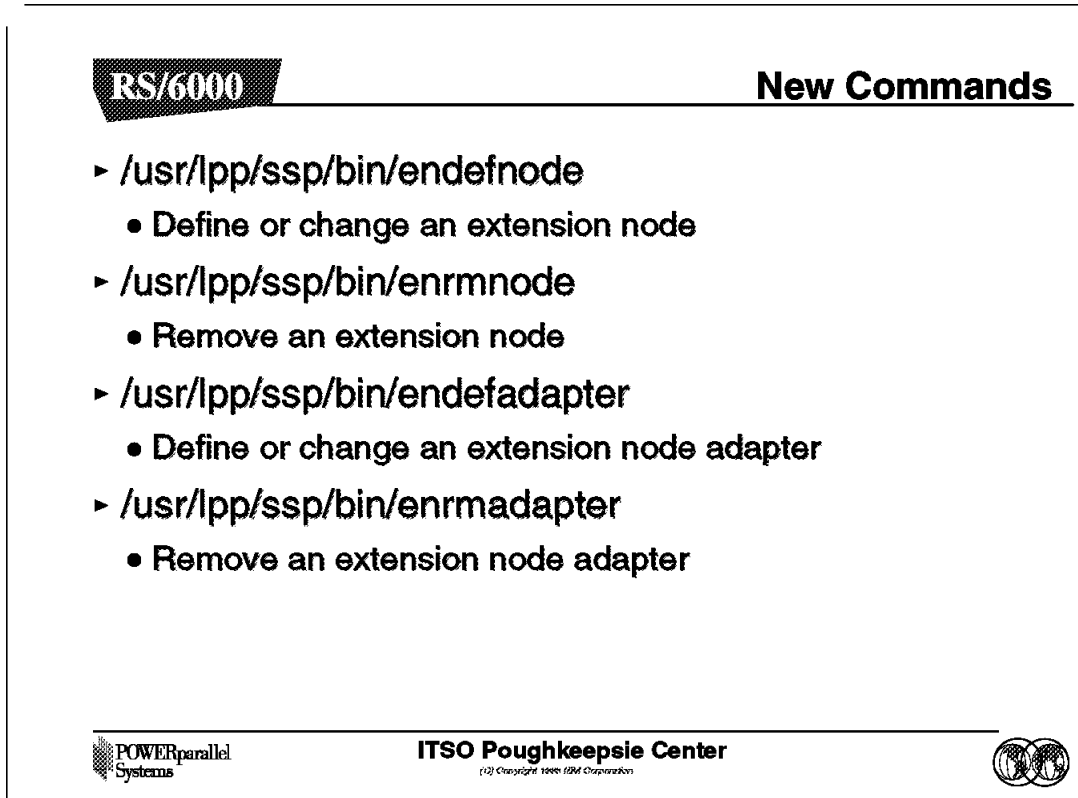
6.4.4 Additional Attributes



This figure shows the additional attributes of the Syspar_map and Switch_partition classes, which are described in detail as follows:

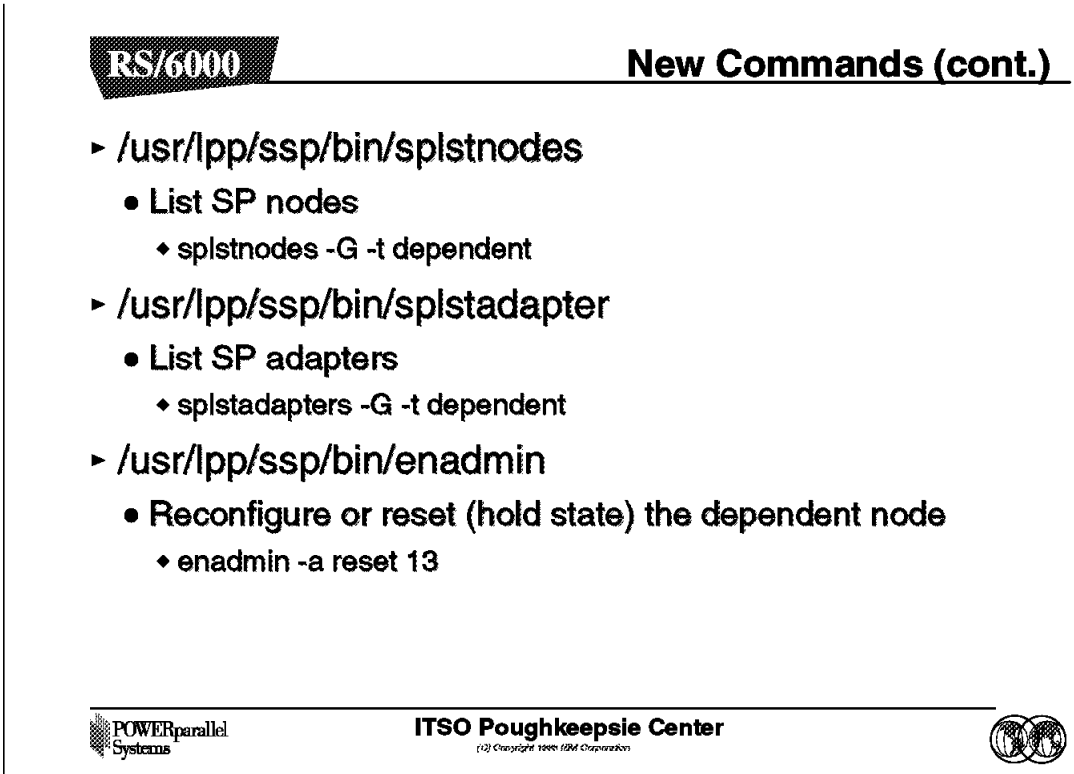
Attribute	Description
node_type	This attribute is set to dependent for GRF and standard for all other RS/6000 SP nodes.
switch_max_ltu	Specifies the maximum packet length of data on the SP Switch; the default is 1024. *Do not change this value for any reason*.
switch_link_delay	Specifies the delay for a message to be sent between the two furthest points on the switch; the default is 31. *Do not change this value for any reason*.

6.4.5 New Commands



This figure shows four new commands that were added to manage the extension node. They have the same characteristics, which are as follows:

- Part of the ssp.basic fileset
- Must only be executed on the Control Workstation
- Can only be executed by the root user
- Only affect the current active partition
- Only affect the SDR, unless the `-r` option is specified (this option is not applicable to `enrmadapter`)
- Return code of 0 if successful, 1 if failed



This figure shows three more commands that were added to manage the extension node.

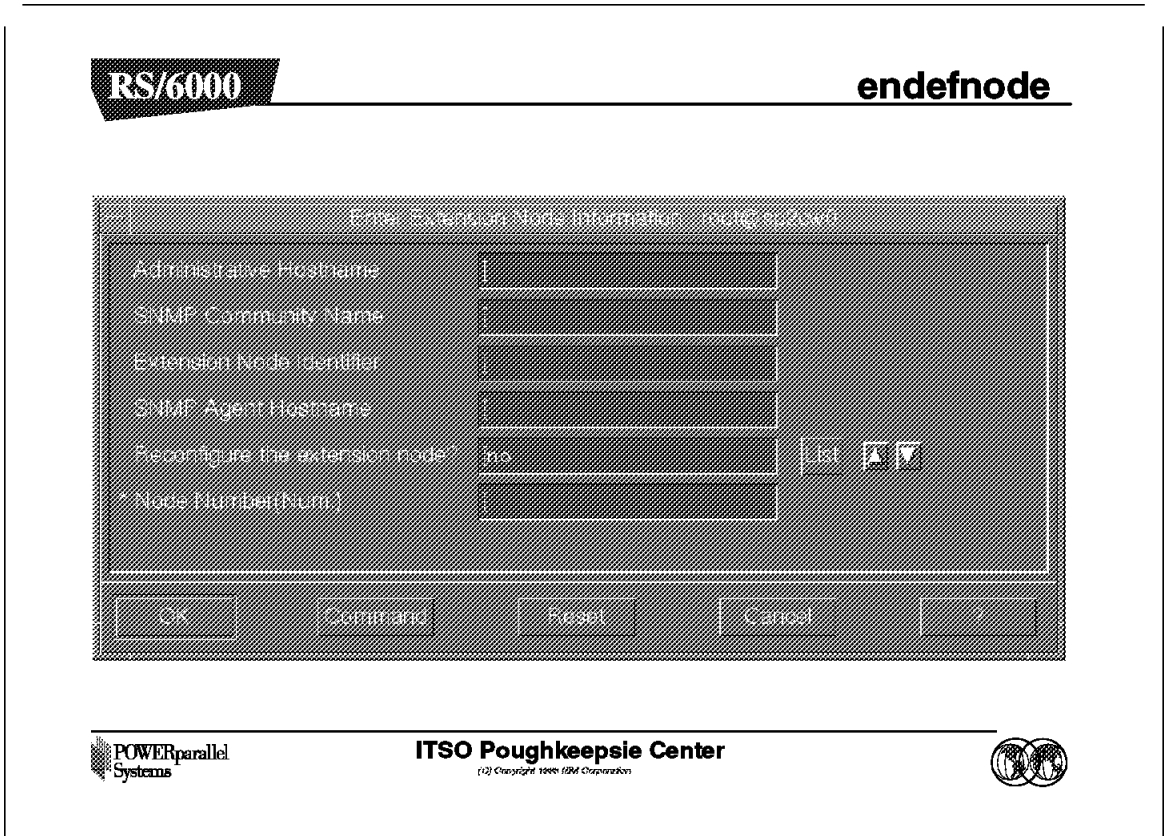
The first two commands, `splstnodes` and `splstadapter`, have the following characteristics:

- Part of the ssp.basic fileset
- Can be executed on any standard RS/6000 SP node
- Can be executed by any user
- Will only affect the current active partition unless the `-G` option is used

The `enadmin` command is used to change the administrative state of a dependent node in the GRF; it has the following characteristics:

- Part of the ssp.spmgr fileset
- Must only be executed on the Control Workstation
- Can only be executed by the root user
- The `-r` option from `endefnode` and `endefadapter` triggers `enadmin -a reconfigure`, while the `-r` option from `enrmnode` triggers `enadmin -a reset`.
- Return code of 0 if successful, 1 if failed

6.4.6 endefnode



The endefnode command can be executed using smit. The fast path for smit is enter_extnode. This command is used to add or change an extension node in the SDR DependentNode class. Its options are shown in Table 4.

Table 4 (Page 1 of 2). endefnode Options

Flag	SMIT Option	Description
-a	Administrative Hostname	This is the hostname of GRF, and the IP name of the GRF's administrative Ethernet, de0. Use long names if DNS is used in the network.
-c	SNMP Community Name	This field contains the SNMP community name that the SP Extension Node SNMP Manager and the GRF's SNMP Agent will send in the corresponding field of the SNMP messages. This value must match the value specified in the /etc/snmpd.conf file on the GRF. If left blank, a default name found in the SP Switch router adapter documentation is used.
-i	Extension Node Identifier	This field contains the two-digit slot number of the SP Switch Router Adapter on the GRF. The value for this field is from 00-15 and is shown on the slots of the GRF.
-s	SNMP Agent Hostname	This field refers to the hostname of the processor running the SNMP Agent for the GRF. In the current version of the GRF, this value is equivalent to that of the Administrative Hostname.

Table 4 (Page 2 of 2). endefnode Options		
Flag	**SMIT Option**	**Description**
-r	Reconfigure the extension node	This field specifies whether the enadmin command is to be activated after the endefnode command completes. It is placed here so that the user does not have to explicitly issue the enadmin command. If the specification is yes, the -r option is part of the command. If the specification is no, the -r option is not part of the command.
	Node Number	This is the node number the extension node logically occupies in the RS/6000 SP.

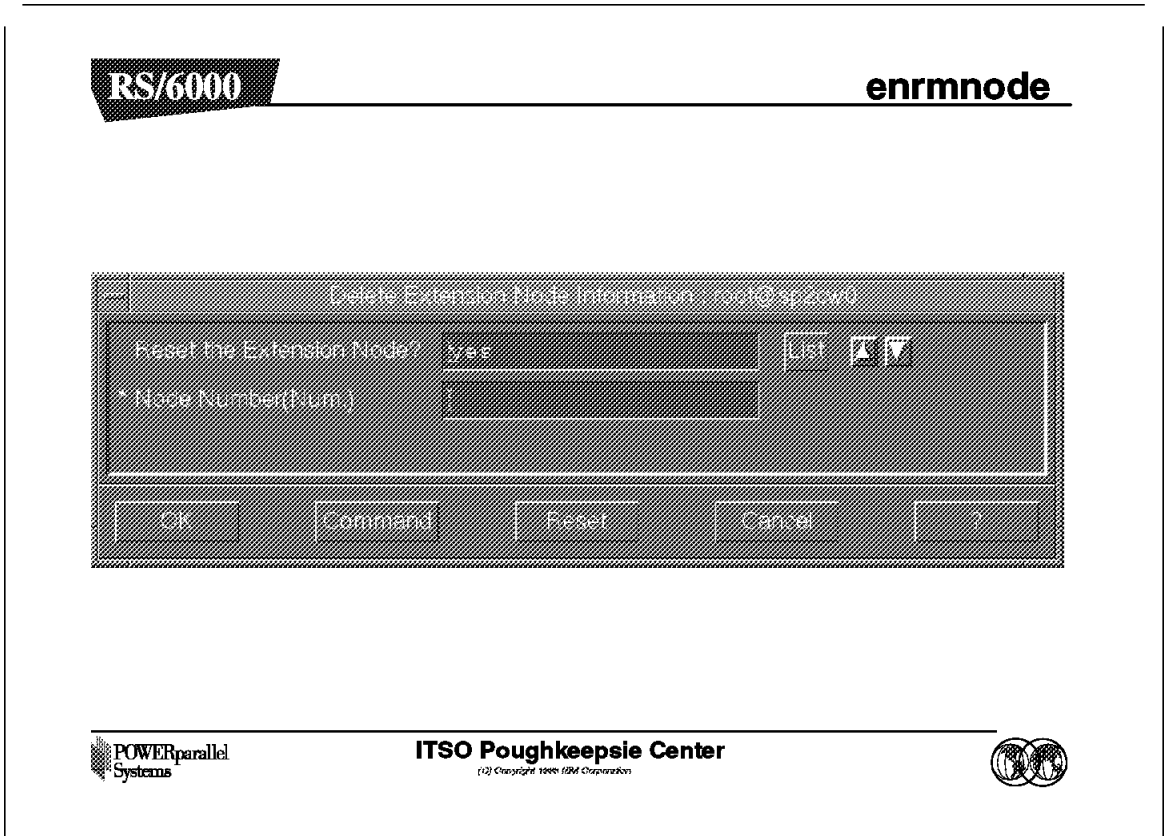
This command adds attribute information for the extension node. The endefadapter command adds IP information, such as IP address and netmask for the extension node. Together, these two commands define the extension node.

Attention

Note that this command only affects the SDR, unless the -r option is used. The -r option should be issued only if endefadapter has been executed for the extension node.

When the GRF is properly configured and powered on, with the SP Switch router adapter inside, it periodically polls the Control Workstation for configuration data. The -r option or enadmin command is not required to activate the polling here.

6.4.7 enrmnode



The enrmnode command is used to remove an extension node from the SDR DependentNode class and can be executed using smit. The fast path for smit is delete_extnode.

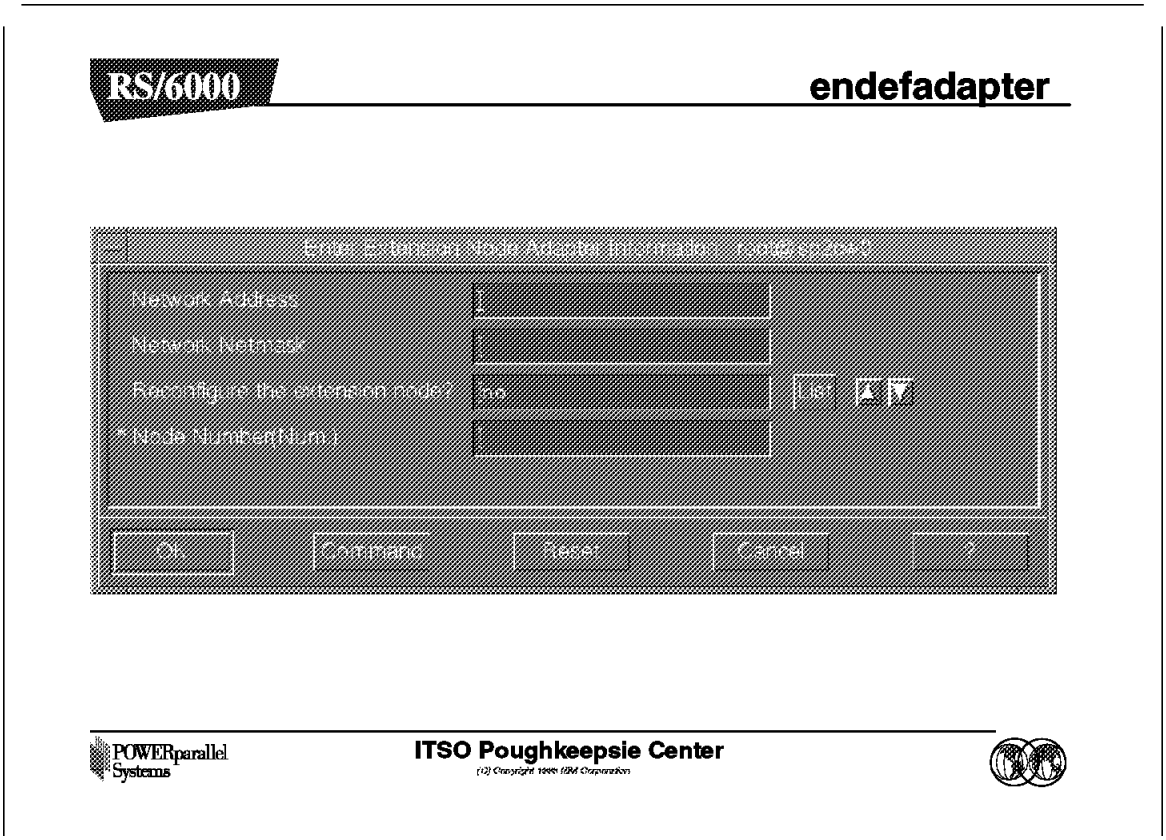
Table 5. enrmnode Options

Flag	SMIT Option	Description
-r	Reset the extension node	Specifies whether the enadmin command is to be activated after the enrmnode command completes. With this option the user does not have to explicitly issue the enadmin command. If the specification is yes, the -r option is part of the command. If the specification is no, the -r option is not part of the command.
	Node Number	This is the node number the extension node logically occupies in the RS/6000 SP.

Attention

Note that this command only affects the SDR unless the -r option is used. This command should be issued with a -r flag, because the enadmin command is not available for the extension node after enrmnode is executed, since the extension node has been removed from the SDR.

6.4.8 endefadapter



The endefadapter command is used to add or change the extension node adapter IP information in the SDR DependentAdapter object, and can be executed using smit. The fast path for smit is enter_extadapter.

Table 6. endefadapter Options

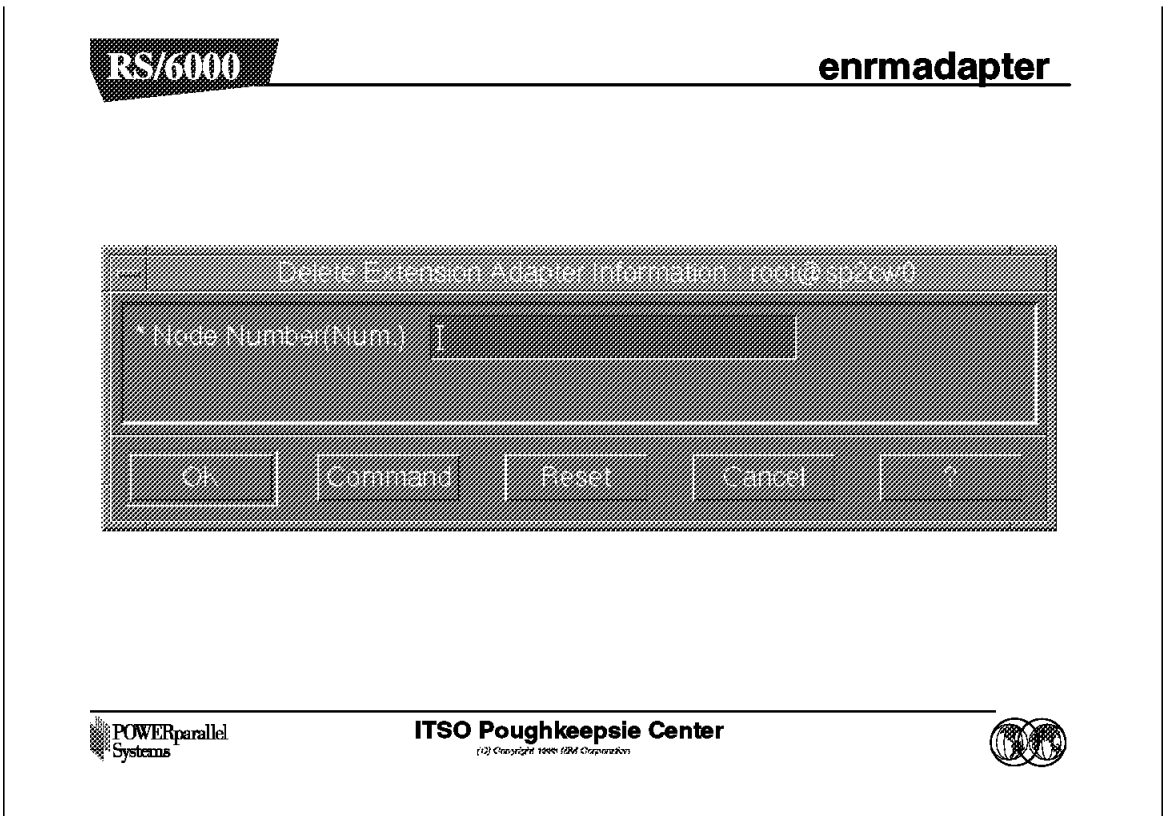
Flags	SMIT Option	Description
-a	Network Address	Specifies the IP address of the extension node.
-m	Network Netmask	Specifies the netmask for the extension node.
-r	Reconfigure the extension node	Specifies if the enadmin command is to be activated after the endefadapter command completes. With this option, the user does not have to explicitly issue the enadmin command. If the specification is yes, the -r option is part of the command. If the specification is no, the -r option is not part of the command.
	Node Number	This is the node number the extension node logically occupies in the RS/6000 SP.

Attention

Note that this command only affects the SDR unless the `-r` option is issued. The `-r` option should be issued only if the `endefnode` has been executed for the extension node.

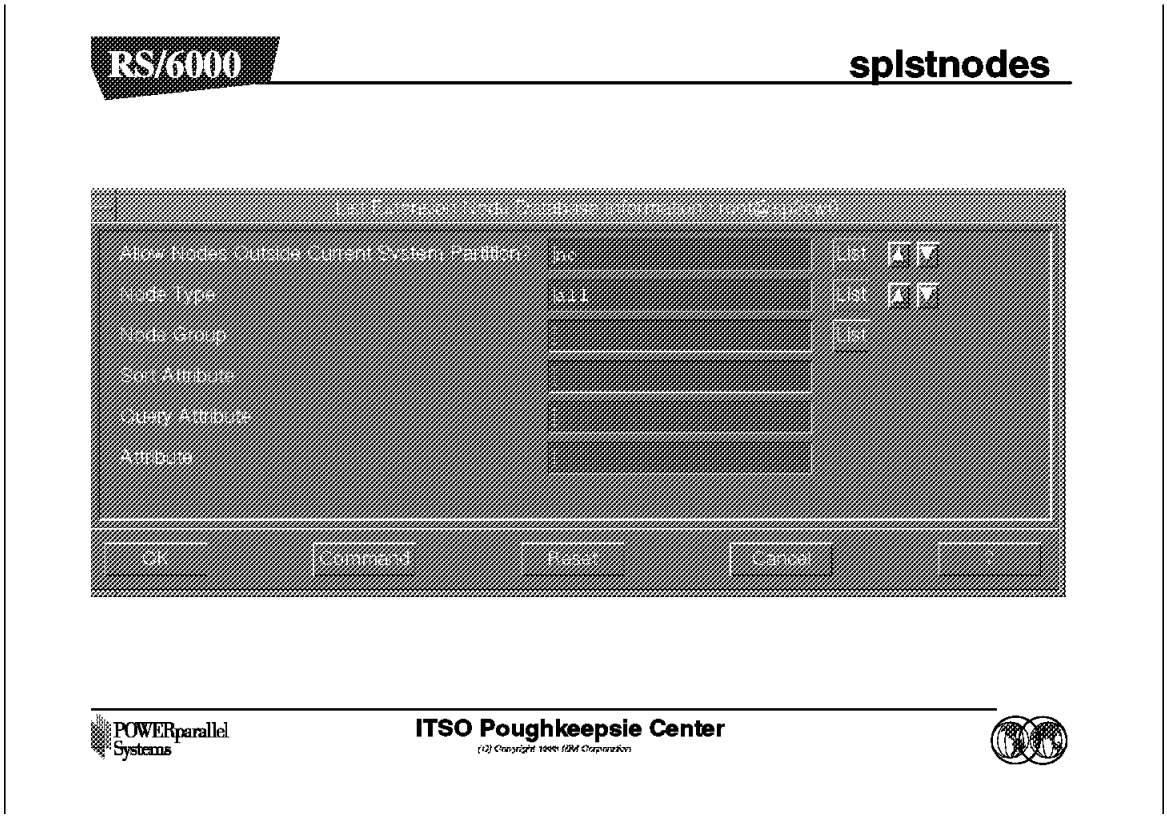
When the GRF is properly configured and powered on, with the SP Switch router adapter inside, it periodically polls the Control Workstation for configuration data. The `-r` option or `enadmin` command is not required to activate the polling here.

6.4.9 enrmadapter



The enrmadapter command is used to remove the SDR DependentAdapter object, and can be executed using smit. The fast path for smit is delete_extadapter.

6.4.10 splstnodes



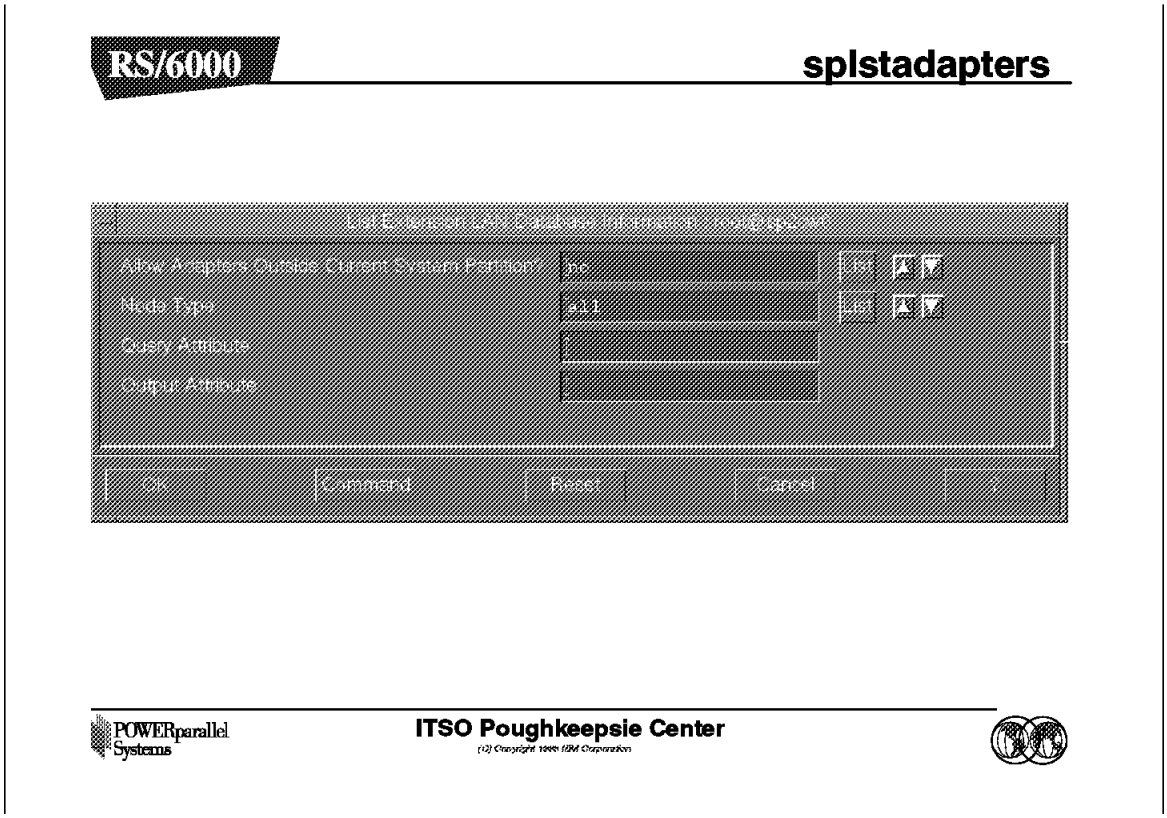
The splstnodes command is used to list the node attributes of all nodes in the SDR, and can be executed using smit. The fast path for smit is list_extnode.

Table 7 (Page 1 of 2). splstnodes Options

Flags	Description
-h	Outputs usage information.
-G	Ignores partition boundaries for its output.
-x.	Inhibits header record in output.
-d <delimiter>	Uses the <delimiter> between its attributes in the output.
-p <string>	Uses the <string> value in the output in place of an attribute that has no value.
-s <attr>	Sorts the output using the <attr> value. In SMIT, this field is known as Sort Attribute.
-t <node-type>	Uses standard to list RS/6000 SP nodes, or dependent. If none is specified, it displays both. In SMIT, this field is known as Node Type.
-N <node_grp>	Restricts the query to the nodes belonging to the node group specified in <node_grp>. If the <node_grp> specified is a system node group, the -G flag is implied.
<attr==value>	This operand is used to filter the output, such that only nodes with attributes that are equivalent to the value specified are displayed. In SMIT, this field is known as Query Attribute.

Table 7 (Page 2 of 2). splstnodes Options	
Flags	**Description**
<attr>	This is a list containing attributes that are displayed by the command. If none is specified, it defaults to node number. This list of attributes can be found in the DependentNode class. In SMIT, this field is known as Attribute.

6.4.11 splstadapters

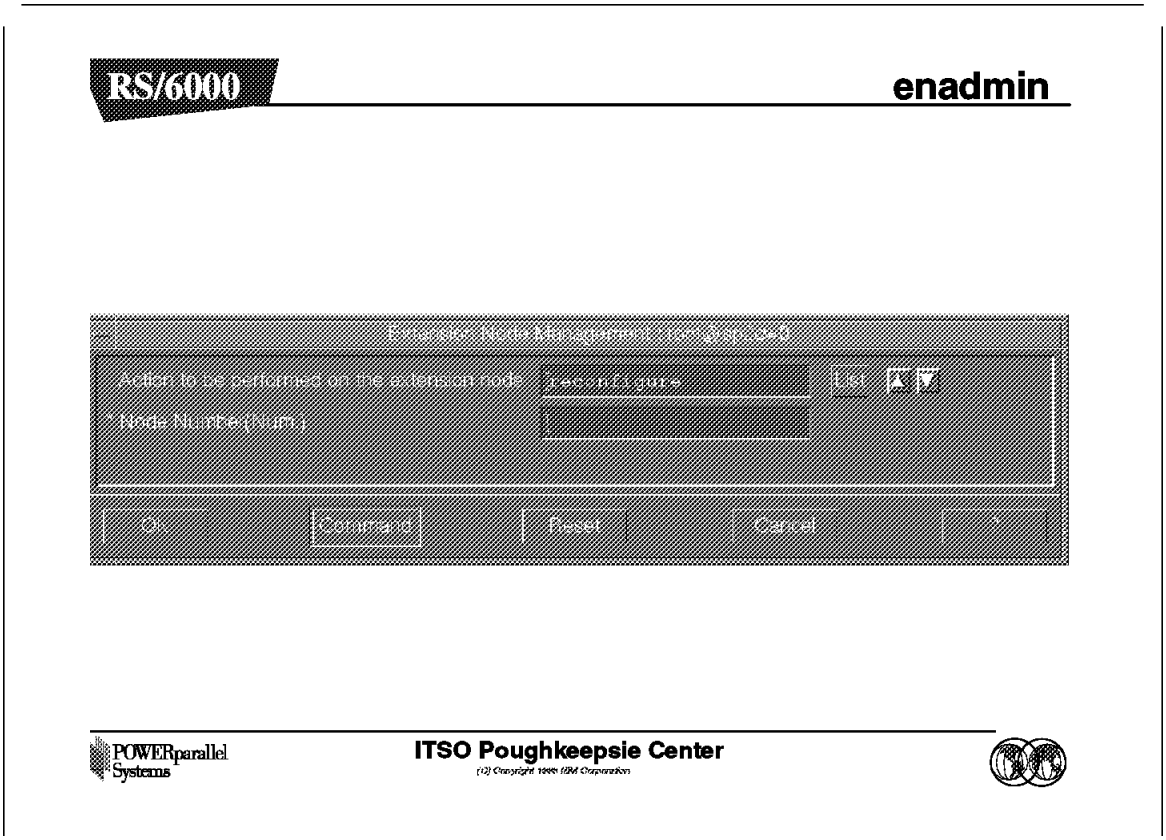


The splstadapers command is used to list the adapter attributes of all nodes in the SDR, and can be executed using smit. The fast path for smit is list_extadapter.

Table 8. splstadapters Options

Flags	Description
-h	Outputs usage information.
-G	Ignores partition boundaries for its output.
-x.	Inhibits header record in output.
-d <delimiter>	Uses the <delimiter> between its attributes in the output.
-p <string>	Uses the <string> value in the output in place of an attribute that has no value.
-t <node-type>	Uses standard to list RS/6000 SP nodes, or dependent. If none is specified, it displays both. In SMIT, this field is known as Node Type.
<attr==value>	This operand is used to filter the output, such that only nodes with attributes that are equivalent to the value specified are displayed. In SMIT, this field is known as Query Attribute.
<attr>	This is a list containing attributes that are displayed by the command. If none is specified, it defaults to node number. This list of attributes can be found in the Adapter and DependentAdapter class. In SMIT, this field is known as Output Attribute.

6.4.12 enadmin

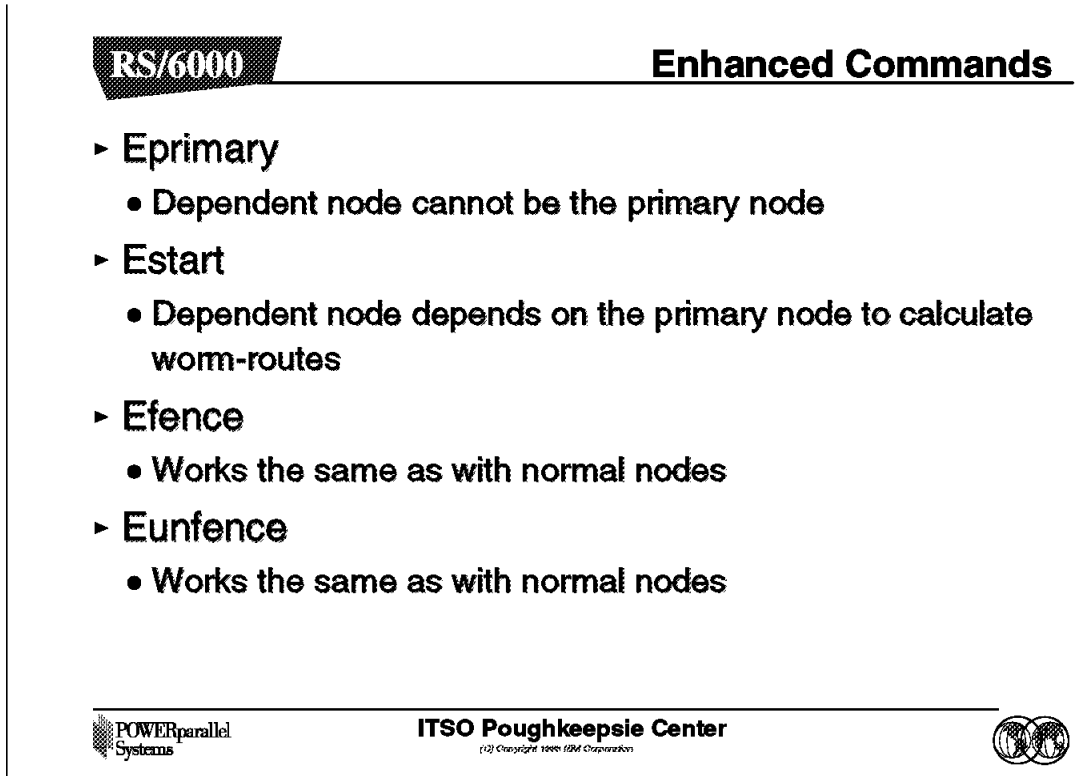


The enadmin command is used to change the status of the SP Switch router adapter in the GRF, and can be executed using smit. The fast path for smit is manage_extnode.

Table 9. enadmin Options

Flags	SMIT Option	Description
-a	Actions to be performed on the extension node.	Either reset or reconfigure. A reset is sent to the extension node SNMP Agent to change the target node to a down state (not active on the SP Switch). A reconfigure is sent to the extension node SNMP Agent to trigger reconfiguration of the target node, which causes the SNMP Agent to request new configuration parameters from the SP Extension Node SNMP Manager, and to reconfigure the target node when the new parameters are received. A more detailed explanation of this is found in the SNMP Flow figure in the Installation section.
	Node Number	This is the node number the extension node logically occupies in the RS/6000 SP.

6.4.13 Enhanced Commands



The following commands have been modified due to the introduction of the dependent node:

- Eprimary

  This command has been modified so that dependent nodes will not be able to act as a Primary or Primary Backup node for the SP Switch in the partition. The dependent node does not run the RS/6000 SP Switch codes like standard RS/6000 SP nodes and therefore does not have the ability to act as the Primary or Primary Backup node.

- Estart

  This command functions as usual with the dependent node in the RS/6000 SP.

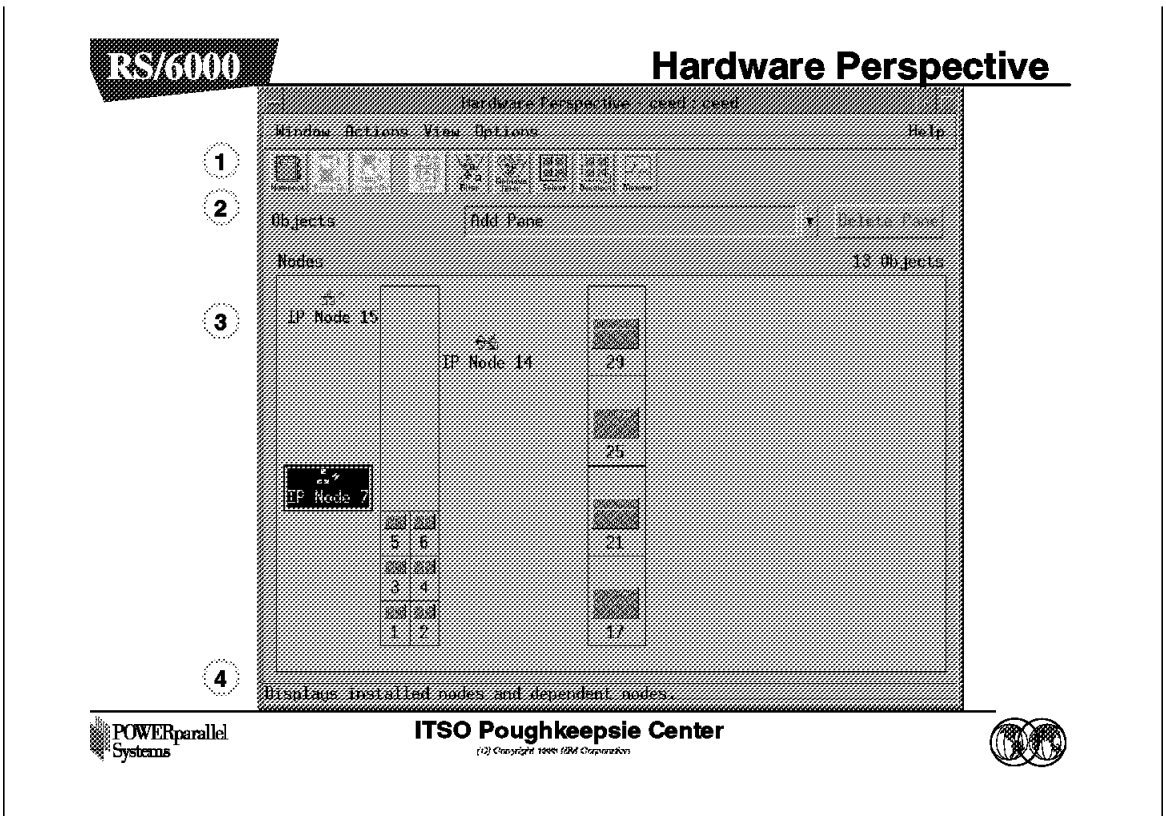
- Efence

  This command functions as usual with the dependent node in the RS/6000 SP. In addition, the dependent node can be fenced from the SP Switch with autojoin like any other standard RS/6000 SP node.

- Eunfence

  This command functions as usual with the dependent node in the RS/6000 SP. In addition, the dependent node can rejoin the SP Switch network with this command, if that node was previously removed from the switch network due to failures or Efence.

6.4.14 Hardware Perspective



IP Node is used in Perspectives as a convenient and short descriptive term easily displayed in the GUI that conveys the role and functions of the dependent node. Currently, this is the only dependent node.

In the following figures, we show the changes made to Perspectives because of the introduction of the IP Node. The changes are restricted to the Hardware and System Partition Aid Perspectives.

This figure shows the Hardware Perspectives, which can be started using the command `perspectives` and selecting the **Hardware** icon. Alternatively, it can be started directly via the command `sphardware`.

The Hardware Perspective consists of the following four parts:

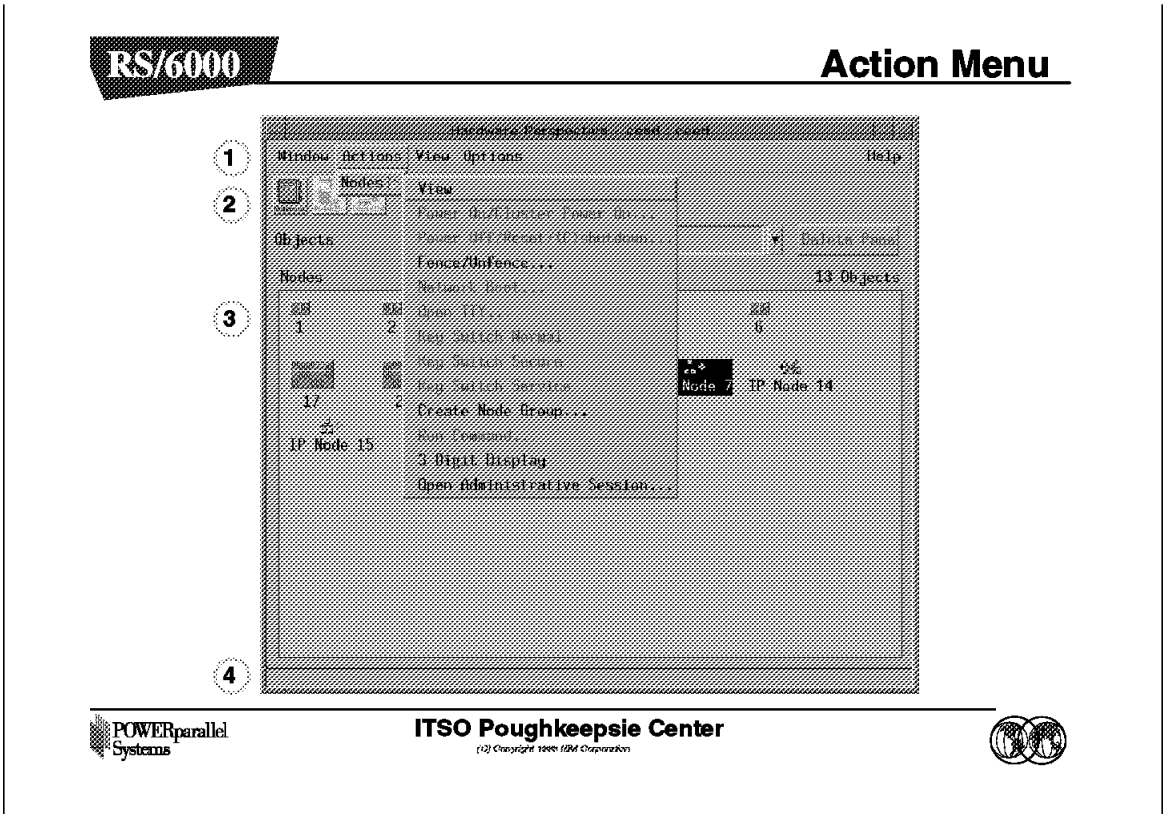
1. Menu bar
2. Toolbar
3. Nodes pane (Frame or Icon View)
4. Information area

The most obvious change is the addition of the IP Node icon as seen in the Nodes pane. (The figure above shows the Frame View.) The default label for this icon is IP Node <node number>.

The IP Node icon is also located on the side of the frame, where a standard node with that node number would be. In this figure the IP Nodes are 7, 14 and 15.

When switch_responds is monitored, it shows the IP Node in two states: green when working with the SP Switch or marked with a red cross when fenced or not operating due to hardware or configuration problems. In the figure, IP Node 7 and 15 are working, while IP Node 14 is down.

6.4.15 Action Menu



In this figure, we see that **IP Node 7** is selected in the Nodes pane, and **Actions→Nodes** is selected in the menu bar (1). We see that only the following five actions are available:

- View

  This will bring up the IP Node's hardware notebook, shown in the next figure.

- Fence/Unfence...

  This will bring up another window to allow us to either fence or unfence an IP Node. If we are fencing the IP Node, we can use the option of autojoin.

- Create Node Group...

  This will bring up another window to allow us to add the RS/6000 SP nodes to a Node Group. This action does not affect the IP Node, even though it is selectable.

- Three-Digit Display

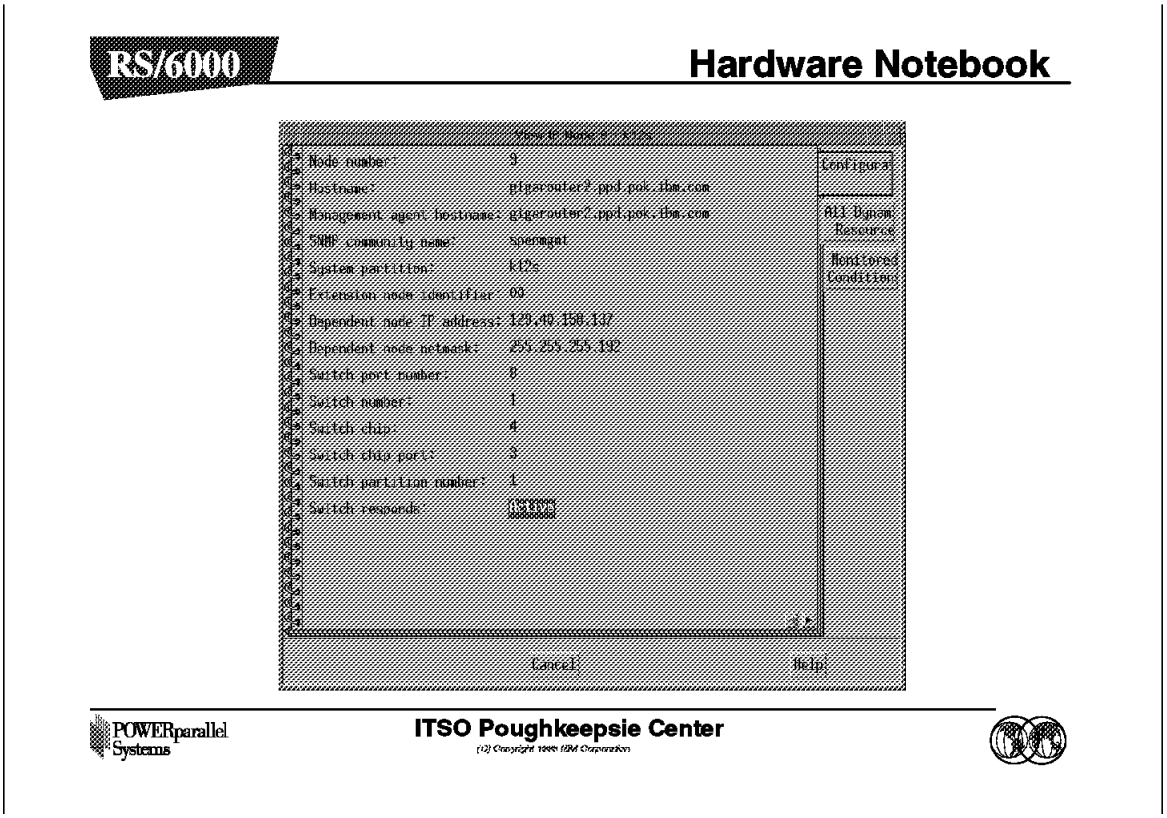
  This will bring up a window to show the three-digit display of all RS/6000 SP standard nodes in the current partition. This action does not apply to the IP Node, even though it is selectable.

- Open Administrative Session...

This action will open a window that is a Telnet session to the GRF, using the reliable_hostname attribute specified in the DependentNode class.

In addition, the Nodes pane in this figure shows the Icon View. In this view, the IP Node icons are always located after all the standard RS/6000 SP node icons. The effects of monitoring the IP Nodes and the icon labels are the same as those of Frame View, mentioned in the previous figure.

6.4.16 Hardware Notebook



This figure shows the IP Node hardware notebook. This notebook can be triggered by selecting the **Notebook** icon on the Hardware Perspective toolbar (2), or selecting **Action**→**Nodes**→**View** in the menu bar (1).

The notebook has three tabs: Configuration, All Dynamic Resource Variables, and Monitored Conditions. This figure shows the Configuration tab.

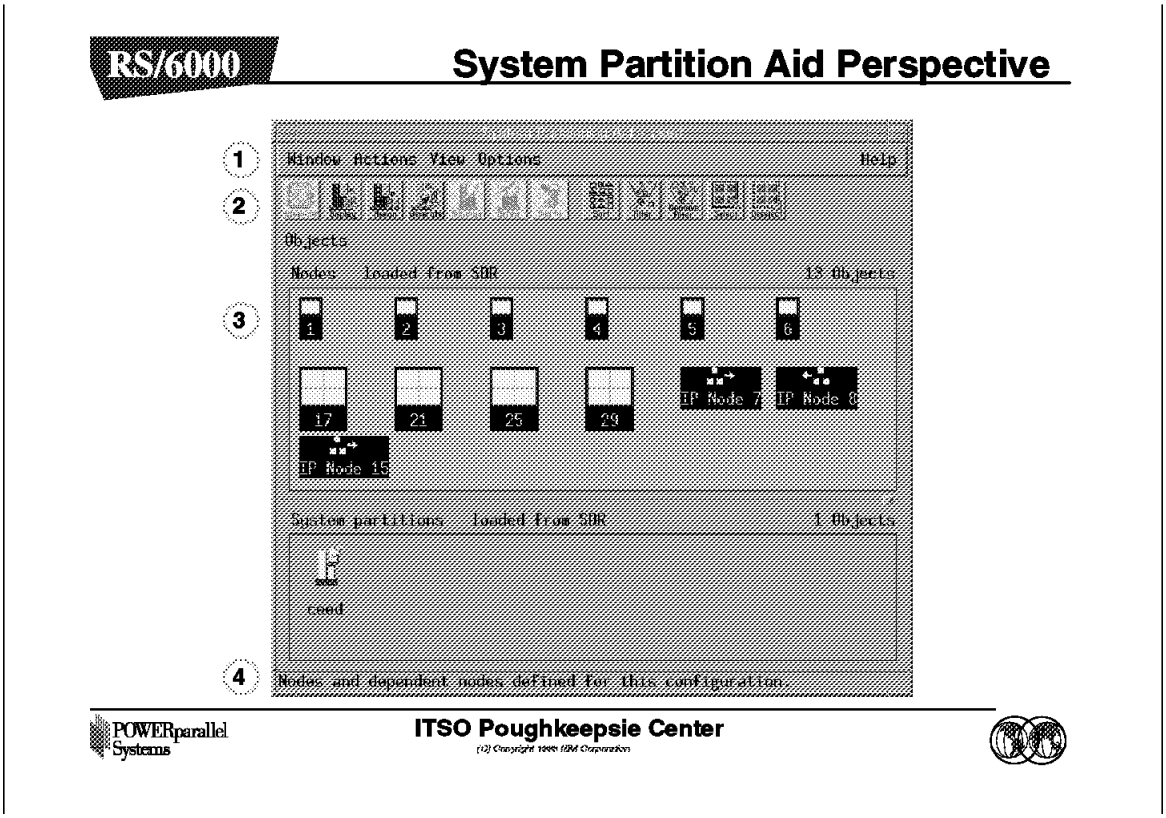
These are the attributes listed in the Configuration tab:

- Node number
- Hostname
- Management agent hostname
- SNMP community name
- System partition
- Extension node identifier
- Dependent node IP address
- Dependent node netmask
- Switch port number
- Switch number
- Switch chip

- Switch chip port
- Switch partition number
- Switch responds

The All Dynamic Resource Variables tab only shows the state of the *Switch Responds*, and the Monitored Conditions tab only shows the value of the *Switch Responds* if it is being monitored.

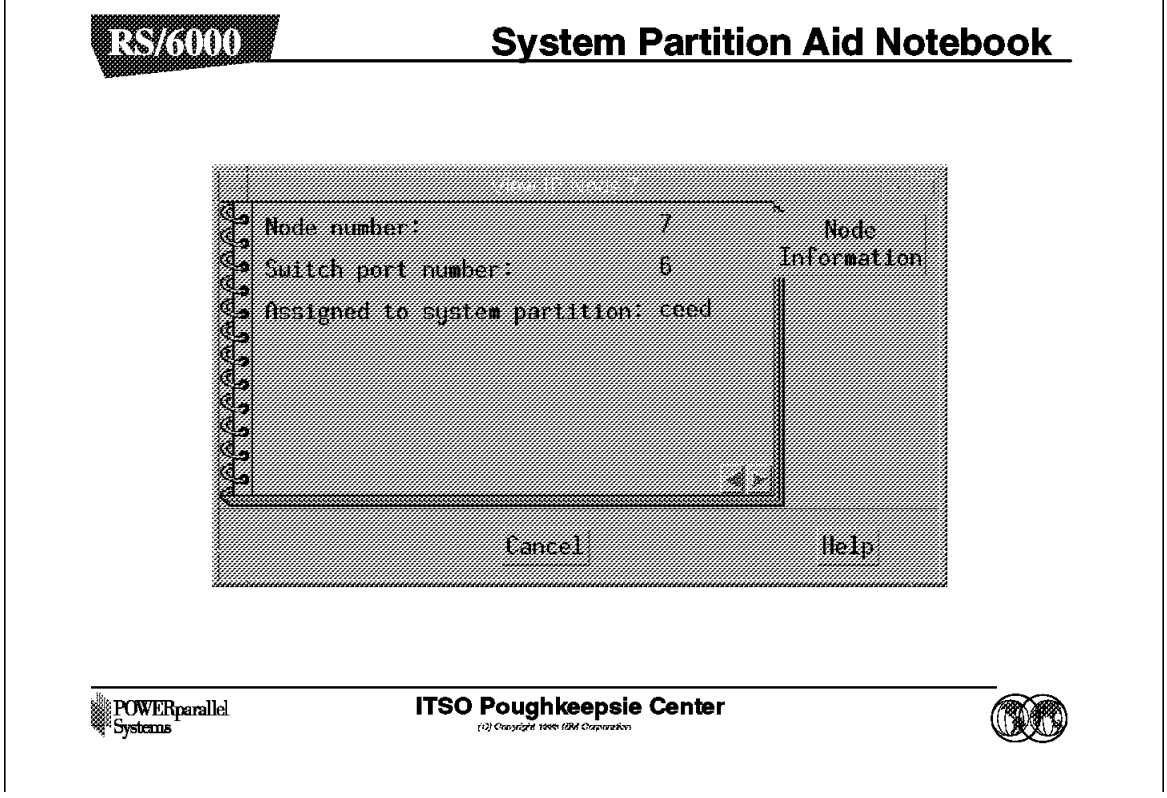
6.4.17 System Partition Aid Perspective



The System Partition Aid Perspective window has two panes, the Nodes pane and the System partitions pane. The Nodes pane (3) in this figure shows the Icon view. Notice that the IP Nodes are displayed after all the standard RS/6000 SP nodes. Also, the node numbers of the IP Nodes are listed below their icons.

The IP Nodes can only be assigned to a partition here. This is done either by using the **Assign** icon in the toolbar (2), or by selecting **Action→Nodes→Assign Nodes to System Partition** on the menu bar (1). Except for the System Partition Notebook, discussed in the next figure, all other actions, though selectable, do not apply to the IP Node.

6.4.18 System Partition Aid Notebook



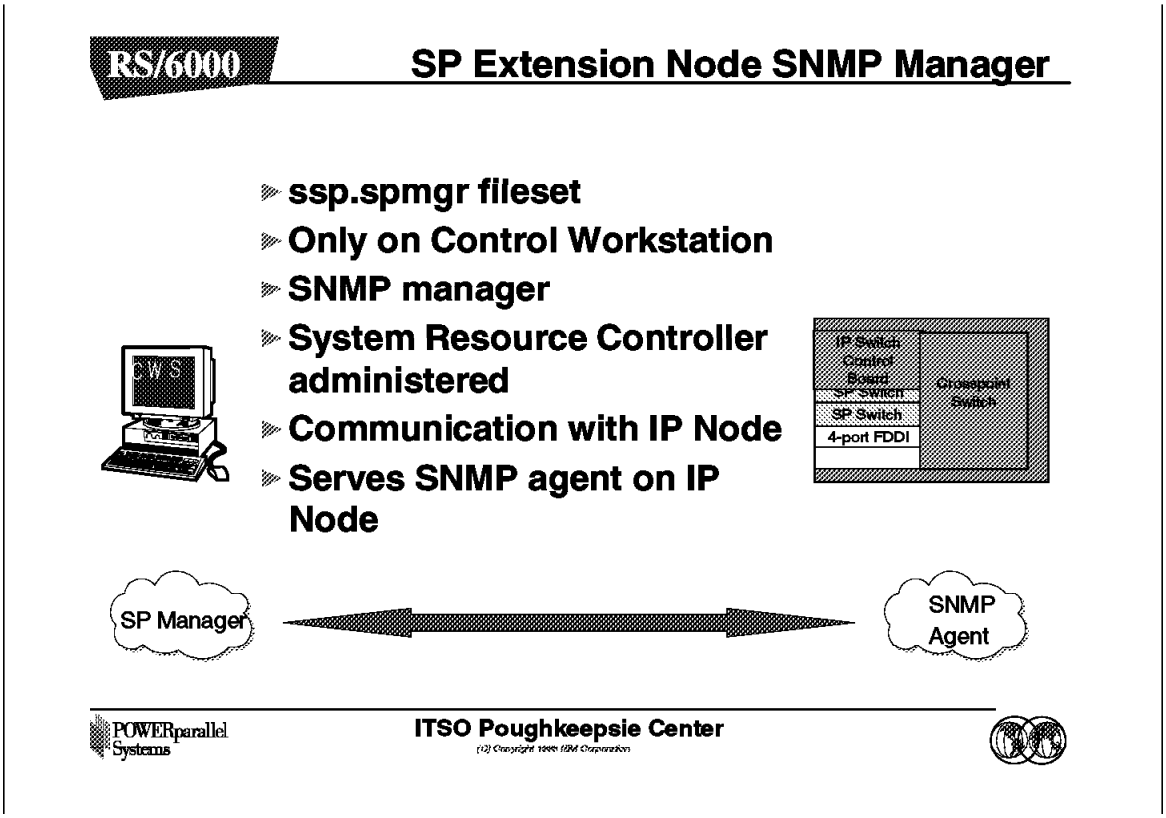
This figure shows the IP Node System Partition Notebook. This notebook can be triggered by selecting the **Notebook** icon on the Hardware Perspective toolbar (2), or selecting **Action**→**Nodes**→**View** on the menu bar (1).

The notebook only has the Node Information tab shown in this figure.

These attributes are listed in the Node Information tab.

- Node number
- Switch port number
- Assigned to system partition

6.4.19 SP Extension Node SNMP Manager

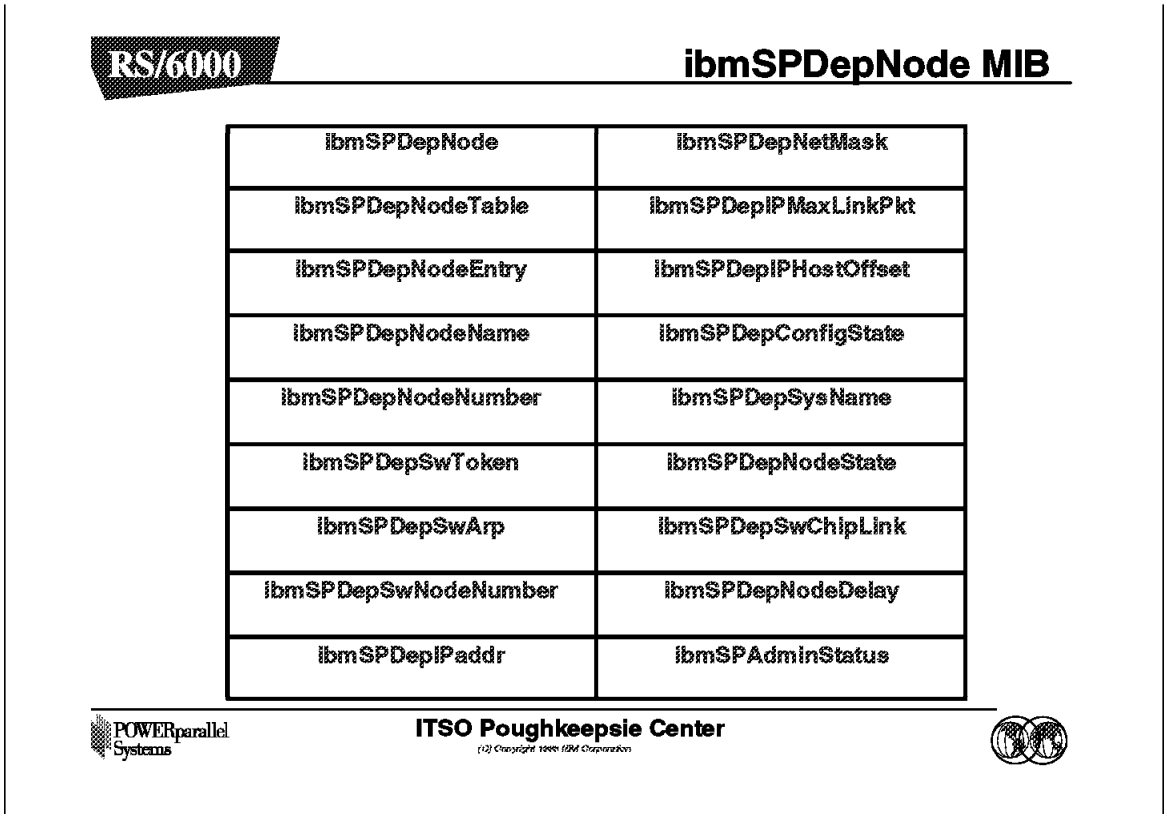


The SP Extension Node SNMP Manager is contained in the ssp.spmgr fileset of PSSP. This fileset must be installed on the Control Workstation in order for the GRF to function as an extension node.

The SP Extension Node SNMP Manager is an SNMP manager administered by the System Resource Controller. The purpose of the SNMP manager is to communicate with the SNMP agent on the GRF. The SNMP Manager and the Agent adhere to Version 1 of the SNMP protocol. The SNMP Manager sends configuration data for an extension node to the SNMP agent on the GRF. The SNMP agent applies the configuration data to the SP Switch router adapter represented by the extension node. The SNMP agent also sends asynchronous notifications in the form of SNMP traps to the SNMP Manager when the extension node changes state. The following commands are available to control the SP Extension Node SNMP Manager:

- `startsrc`
- `stopsrc`
- `lssrc`
- `traceson`
- `tracesoff`

6.4.20 ibmSPDepNode MIB



IBM has defined a dependent node SNMP Management Information Base (MIB) called ibmSPDepNode. This MIB contains definitions of objects representing configuration attributes of each dependent node and its state. The GRF Agent maintains the state and configuration data for each dependent node using the MIB as a conceptual database.

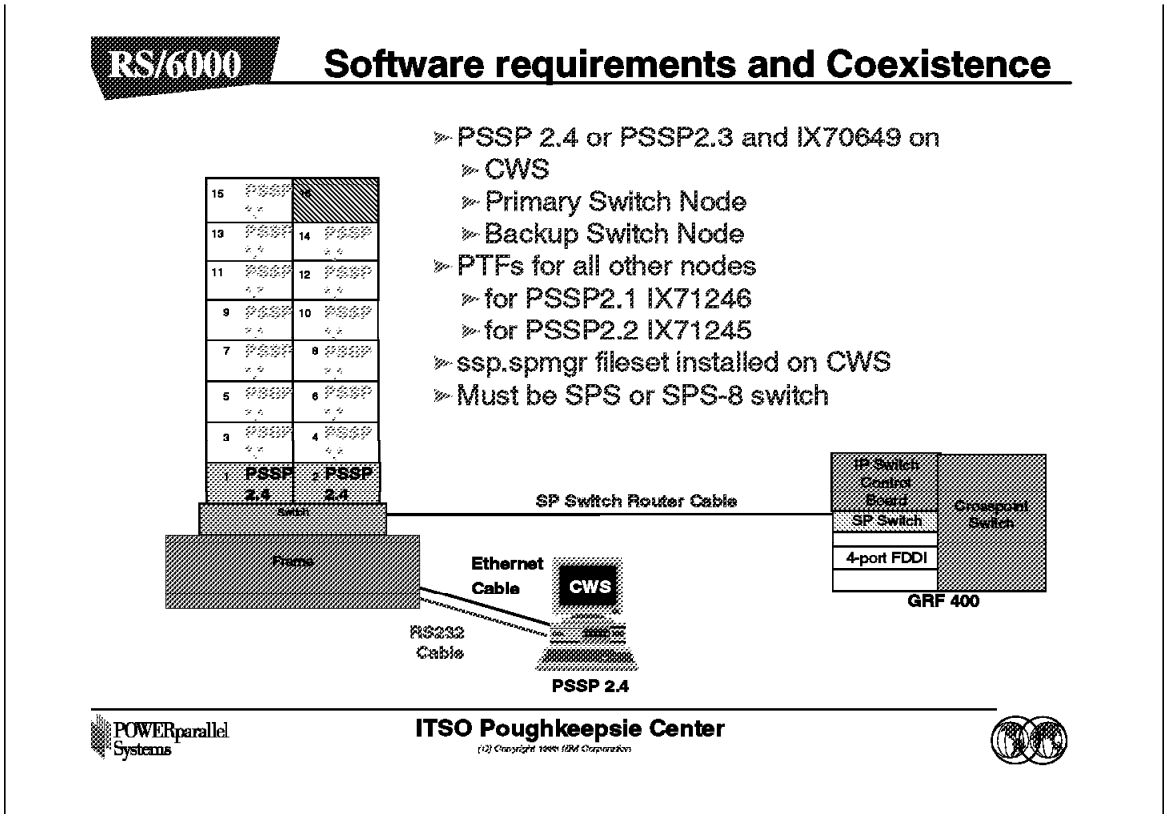
The MIB defines a single table of up to 16 entries representing the adapter slots in the GRF. When a slot is populated by an SP Switch router adapter, the entry in the table, accessed using the extension node identifier, contains the configuration attribute and state values for the adapter in the slot. Also included in the MIB are the definitions of trap messages sent by the GRF Agent to the SP Extension Node SNMP Manager. A copy of the MIB is contained in the file /usr/lpp/ssp/config/spmgrd/ibmSPDepNode.my on the Control Workstation.

Other SNMP managers in the network can query this MIB table to validate the configuration and status of the dependent node and GRF. However, only an SNMP manager using the correct SNMP community name can change the values in the MIB table.

Following is a listing of its entries.

Entry	Definition
ibmSPDepNode	Object identifier for the dependent node in the MIB database.
ibmSPDepNodeTable	Table of entries for dependent nodes.
ibmSPDepNodeEntry	A list of objects comprising a row and a clause in the ibmSPDepNodeTable. The clause indicates which object is used as an index into the table to obtain a table entry.
ibmSPDepNodeName	The extension_node_identifier attribute in the DependentNode class.
ibmSPDepNodeNumber	The node_number attribute in the DependentNode class.
ibmSPDepSwToken	A combination of switch_number, switch_chip and switch_chip_port attributes from the DependentNode class.
ibmSPDepSwArp	The arp_enabled attribute in the Switch_partition class.
ibmSPDepSwNodeNumber	The switch_node_number attribute in the DependentNode class.
ibmSPDepIPaddr	The netaddr attribute in the DependentAdapter class.
ibmSPDepNetMask	The netmask attribute in the DependentAdapter class.
ibmSPDepIPMaxLinkPkt	The switch_max_ltu attribute in the Switch_partition class.
ibmSPDepIPHostOffset	This attribute stores the difference between the host portion of a node′s IP address and its corresponding switch node number. When ARP is disabled on the SP Switch network, this offset is subtracted from the host portion of IP address to calculate the switch node number.
ibmSPDepConfigState	The six config states of the dependent node are: notConfigured, firmwareLoadFailed, driverLoadFailed, diagnosticFailed, microcodeLoadFailed, and fullyConfigured, for use in configuring the adapter.
ibmSPDepSysName	The syspar_name attribute in the Syspar class.
ibmSPDepNodeState	The value of nodeUp or nodeDown, to show the status of the dependent node.
ibmSPDepSwChipLink	The switch_chip_port attribute in the DependentNode class.
ibmSPDepNodeDelay	The switch_link_delay attribute in the Switch_partition class.
ibmSPDepAdminState	The value of up, down, or reconfigure, indicating the desired state of the dependent node. If the dependent node is not in its desired state, the SNMP agent on the GRF will trigger the appropriate action to change its state.

6.4.21 Coexistence



This figure shows a single-frame RS/6000 SP in a single partition with a connection to the GRF. Nodes 1 and 2 are installed with PSSP 2.4. The other nodes are installed with any other version of PSSP that can coexist with PSSP 2.4 to represent coexistence. Also, note that Node 16 is empty, because the SP Switch port for this node is used by the SP Switch router adapter in the GRF.

The dependent node is only supported in PSSP 2.3 and higher PSSP versions. To use it with nodes with PSSP versions less than 2.3 requires the use of *coexistence*. The following conditions are required for the dependent node to communicate with nodes with a lower version than 2.3 using coexistence:

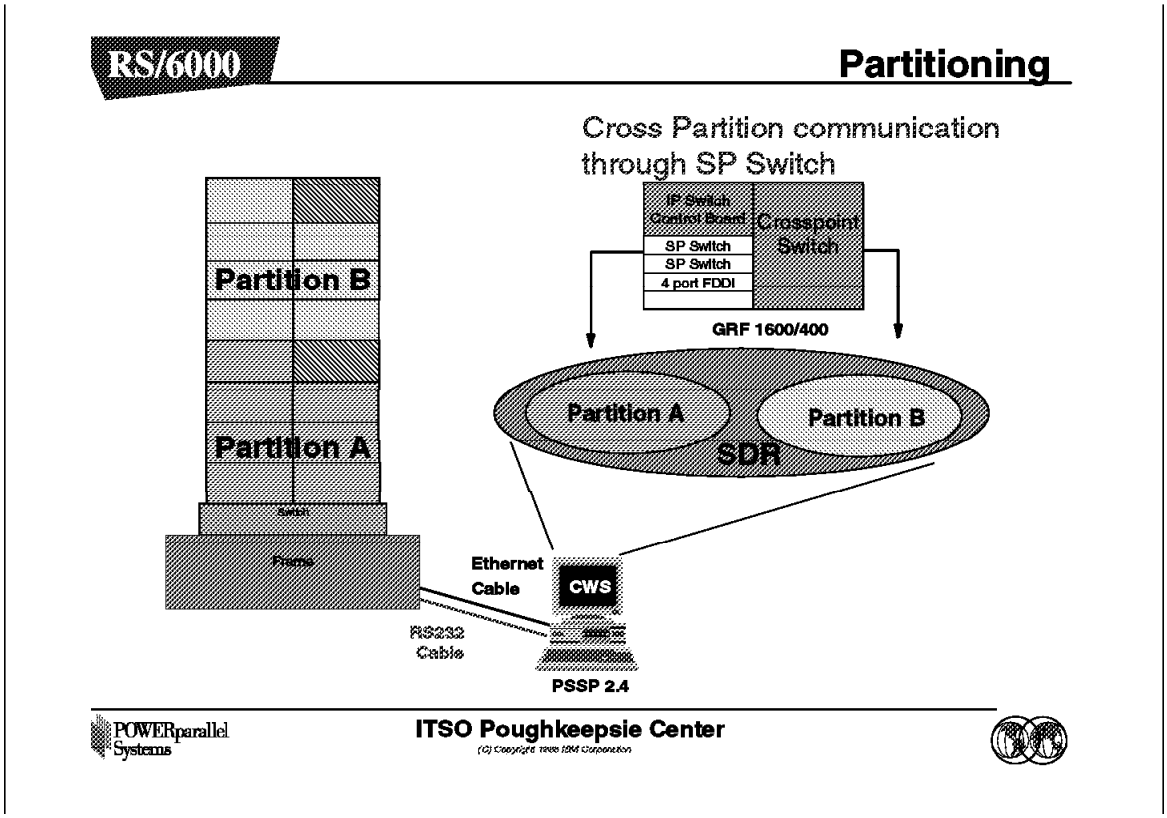
- The Control Workstation must be at PSSP 2.3 or higher to manage dependent nodes.
- The Primary node of the SP Switch must be at PSSP 2.3 or higher, as the Primary node needs to perform some tasks for the dependent node and these functions are only available in PSSP 2.3 and higher PSSP versions.
- The Primary Backup node of the SP Switch should be PSSP 2.3 or higher so that if the Primary node fails, the dependent node can continue to function in the RS/6000 SP when the Backup node takes over.
- All RS/6000 SP nodes with a version less than PSSP 2.3 in the partition need to maintain the right level of fixes (PTFs) in order for coexistence with PSSP 2.4 to take place.

- The ssp.spmgr fileset must be installed on the Control Workstation.
- Because the SP Switch router adapter will only work with the 8-port or 16-port SP Switch, make sure that the switch used in the RS/6000 SP is not a High Performance Switch (HiPS).
- There must be at least one free SP Switch port to install the SP Switch router adapter.

Important

When the Primary Switch node fails, the Primary Backup Switch node will take over as the new Primary switch node. The new Primary Backup Switch node, selected from the current partition, can be a node with a PSSP level below 2.3, even though another node with a PSSP level of 2.3 or higher may exist in that partition. The only way to ensure that the new Backup Switch node is at PSSP 2.3 or higher is to manually check the RS/6000 SP system, and reset it to a node with PSSP 2.3 or higher if one exists.

6.4.22 Partitioning

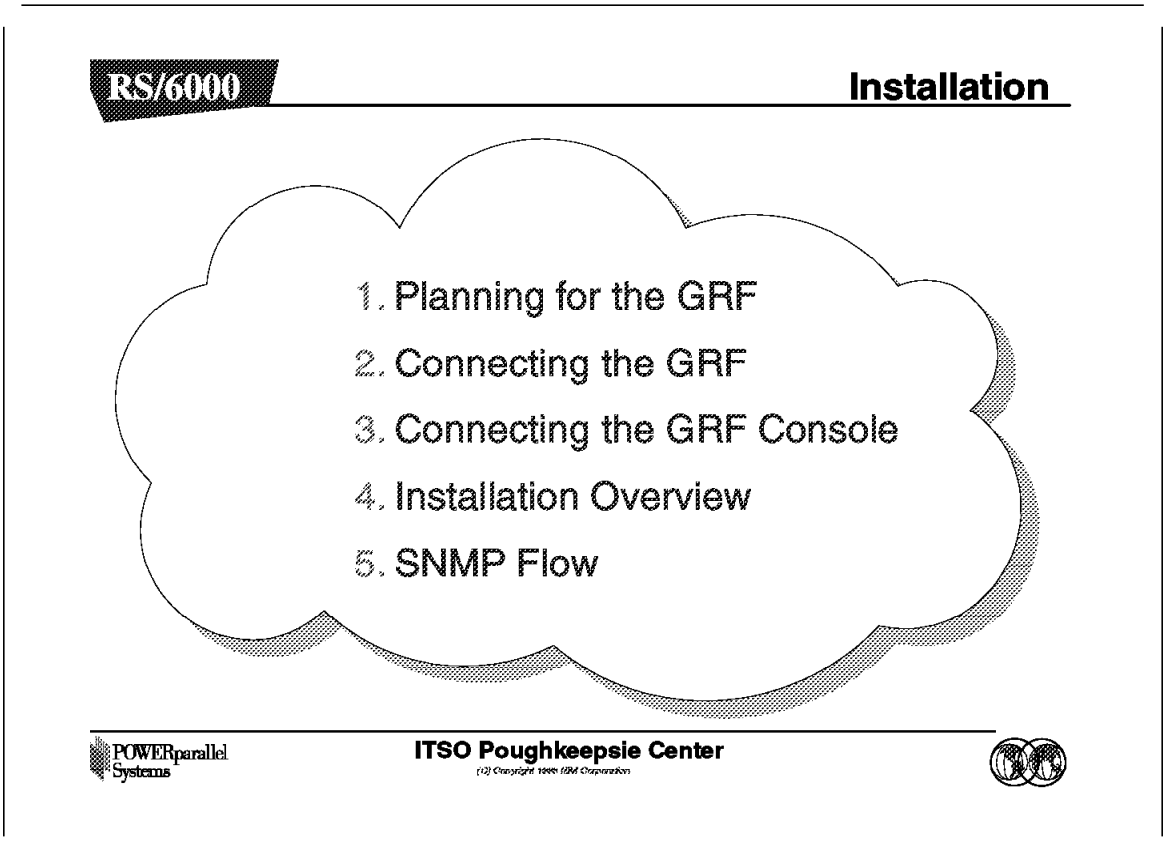


This figure shows a single-frame RS/6000 SP broken into two partitions, Partition A and Partition B. Each partition has seven standard RS/6000 SP nodes and one dependent node. Only seven nodes are allowed in each partition, as a single-frame RS/6000 SP has only 16 SP Switch ports, and two of them are used for the SP Switch router adapter, one for each partition.

Normally, RS/6000 SP nodes in different partitions cannot communicate with each other through the SP Switch. The GRF plays a unique role here by allowing RS/6000 SP nodes to communicate across partitions, when each partition contains at least one SP Switch router adapter, and these adapters are interconnected by TCP/IP.

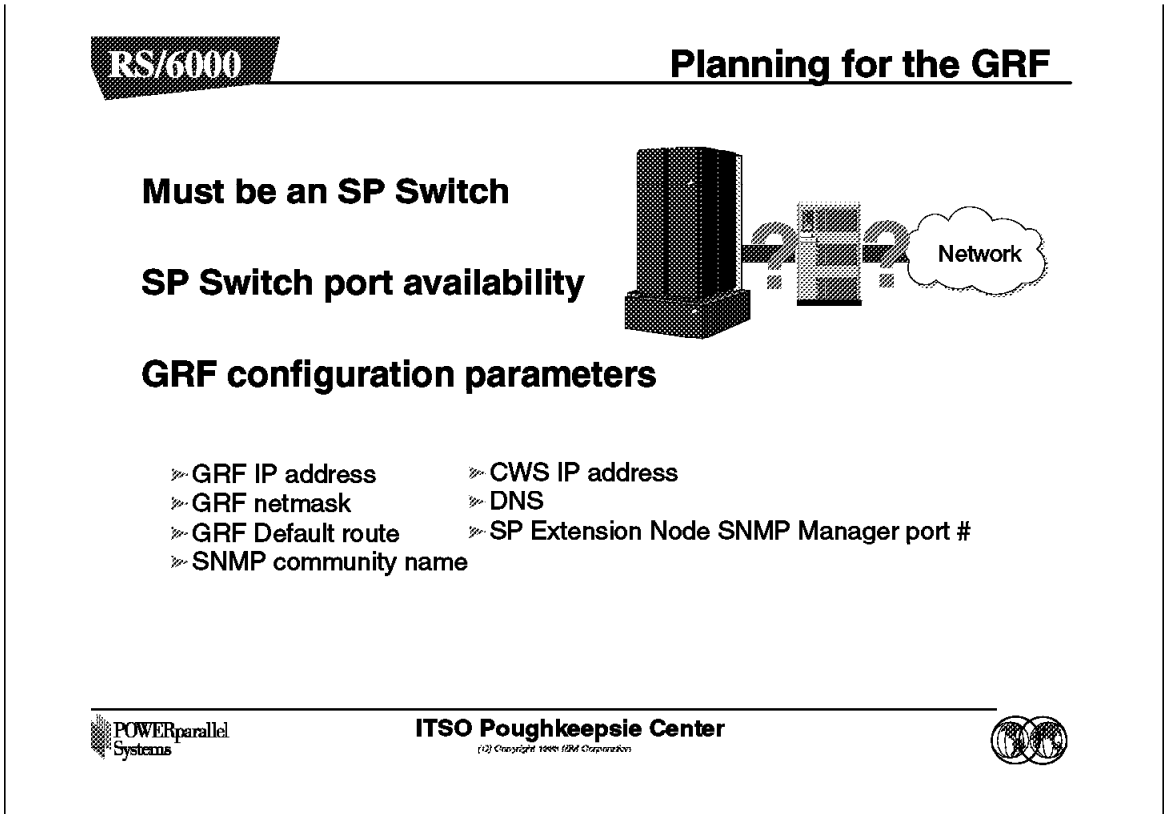
The requirements for partitioning are the same as those for coexistence, with the addition of having at least one free SP Switch port per partition, to connect to the SP Switch router adapter. A more detailed discussion of this situation is given in 6.6, “Sample Configurations” on page 227.

6.5 Installation



This section offers an overview of the installation and planning process.

6.5.1 Planning for the GRF



Before acquiring any model of the SP Switch Router, ensure that there are SP Switch ports available in the designated partition, and that the switch used in the RS/6000 SP is the 8-port or 16-port SP Switch.

Next, ensure that the following parameters are defined:

Parameters	Descriptions
GRF IP address	IP address for GRF administrative Ethernet.
GRF netmask	Netmask for GRF administrative Ethernet.
GRF Default route	The default route of the GRF.
SNMP community name	This attribute describes the SNMP community name that the SP Extension Node SNMP Manager and the GRF's SNMP Agent will send in the corresponding field of the SNMP messages. This value must match the value specified for the same attribute of the corresponding dependent node definition on the SP system. If left blank, a default name found in the SP Switch Router Adapter documentation is used.

CWS IP address

The Control Workstation's IP address. When a GRF contains multiple SP Switch router adapters which are managed by different SNMP Mangers on different RS/6000 SP CWS, each of the Control Workstation IP address should be defined along with a different community name for each Control Workstation.

DNS

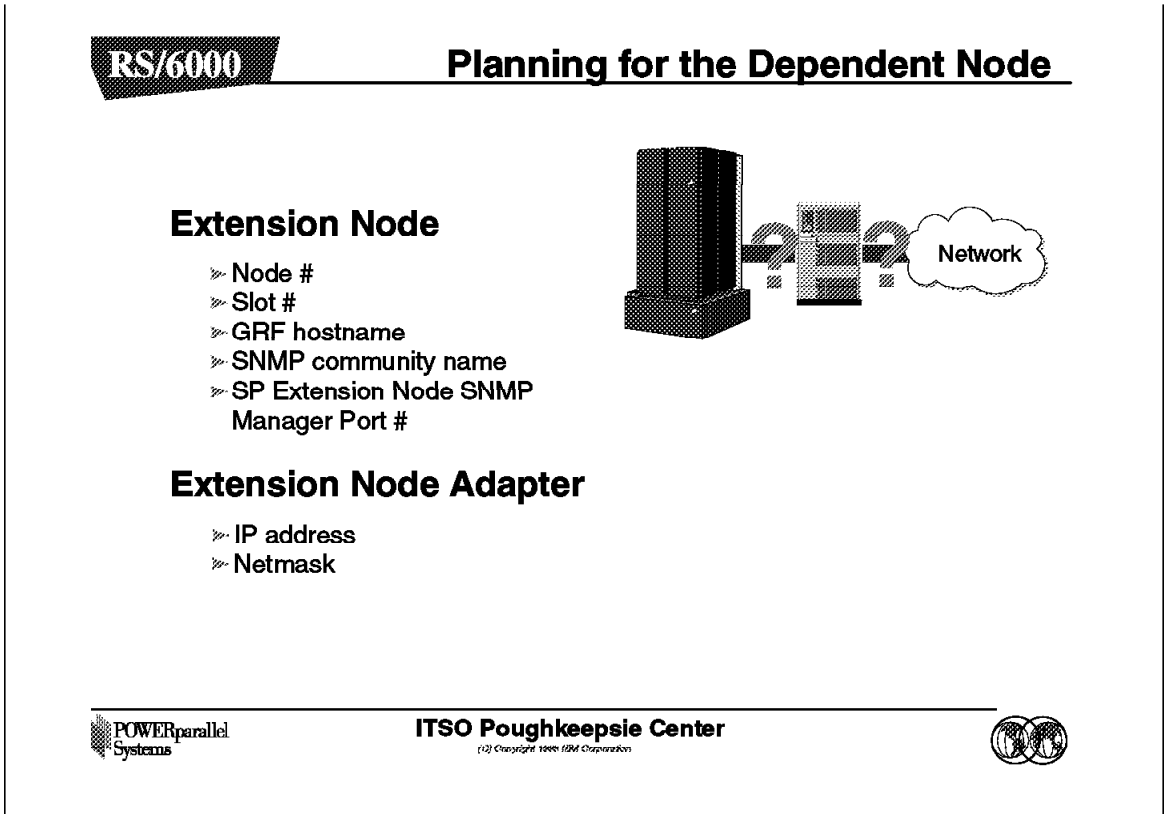
The DNS server and domain name, if used.

SP Extension Node SNMP Manager port #

The SNMP port number used by the SP Extension Node SNMP Manager to communicate with the SNMP agent on the GRF.

This port number is 162 when the SP Extension Node SNMP Manager is the only SNMP manager on the Control Workstation. Otherwise, another port number not used in the /etc/services of the Control Workstation is chosen.

6.5.2 Planning for the Dependent Node



Next, for each dependent node on the RS/6000 SP, define the following:

Parameters	Descriptions
Node #	User-supplied dependent node number representing the node position of an unused SP Switch port to be used by the SP Switch Router Adapter.
Slot #	Slot number which the SP Switch router adapter is located in the GRF.
GRF hostname	Hostname for GRF administrative Ethernet. A long hostname is recommended if the domain name service (DNS) is used in the network. This represents both the Administrative and SNMP Agent Hostname of the dependent node.
SNMP community name	This attribute describes the SNMP community name that the SP Extension Node SNMP Manager and the GRF's SNMP Agent will send in the corresponding field of the SNMP messages. This value must match the value specified in the /etc/snmpd.conf file on the GRF. If left blank, a default name found in the SP Switch Router Adapter documentation is used.

SP Extension Node SNMP Manager port #

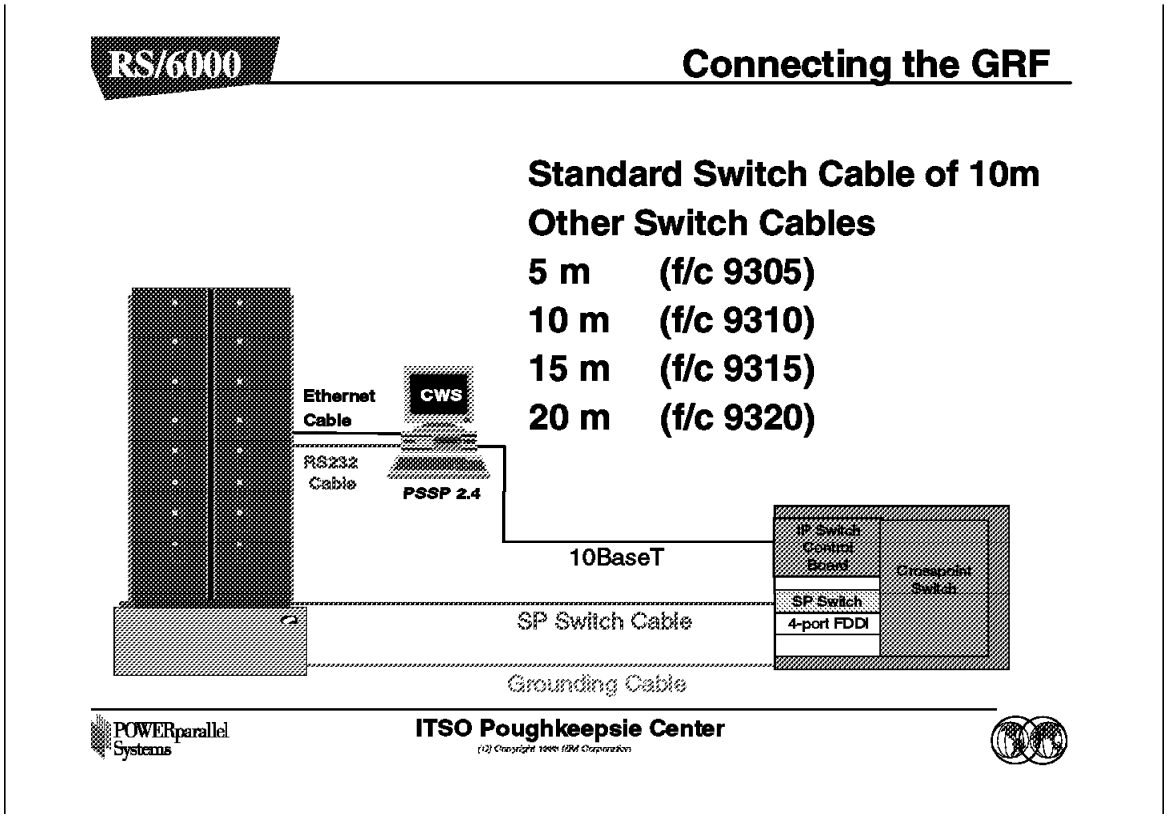
The SNMP port number used by the SP Extension Node SNMP Manager to communicate with the SNMP agent on the GRF.

This port number is 162 when the SP Extension Node SNMP Manager is the only SNMP manager on the Control Workstation. Otherwise, another port number not used in the /etc/services of the Control Workstation is chosen.

Then, for the dependent node adapter, define these parameters:

Parameter	Descriptions
IP address	IP address of this adapter.
Netmask	Netmask of this adapter. Use the same format as that for standard RS/6000 SP nodes.

6.5.3 Connecting the GRF



The GRF, when ordered with the SP Switch router adapter, comes with two cables: a 10 meter SP Switch cable, and a 10 meter grounding cable.

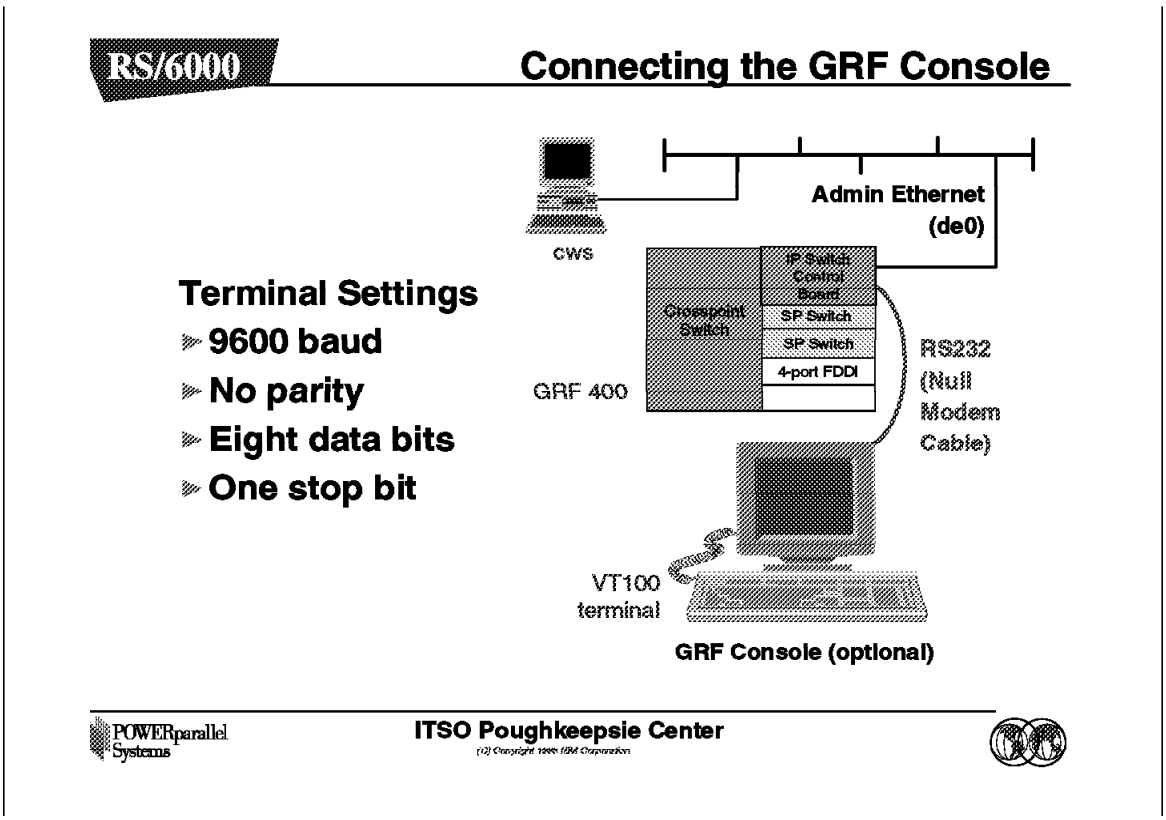
- The SP Switch cable connects the SP Switch port to the SP Switch router adapter on the GRF.
- The grounding cable connects the GRF chassis to the RS/6000 SP chassis for grounding the GRF.

The 10BaseT Ethernet cable is used to connect the GRF's administrative Ethernet to the Control Workstation. The customer must supply the 10BaseT connection to the CWS. Alternatively, this Ethernet can be connected to the SP Ethernet by providing the appropriate bridge.

Note that on RPQ, other switch cable lengths are available.

An alternative to using the GRF-provided SP Switch cable is to use the standard RS/6000 SP Switch cable, which is identical.

6.5.4 Connecting the GRF Console



This figure shows how to connect the console to the GRF.

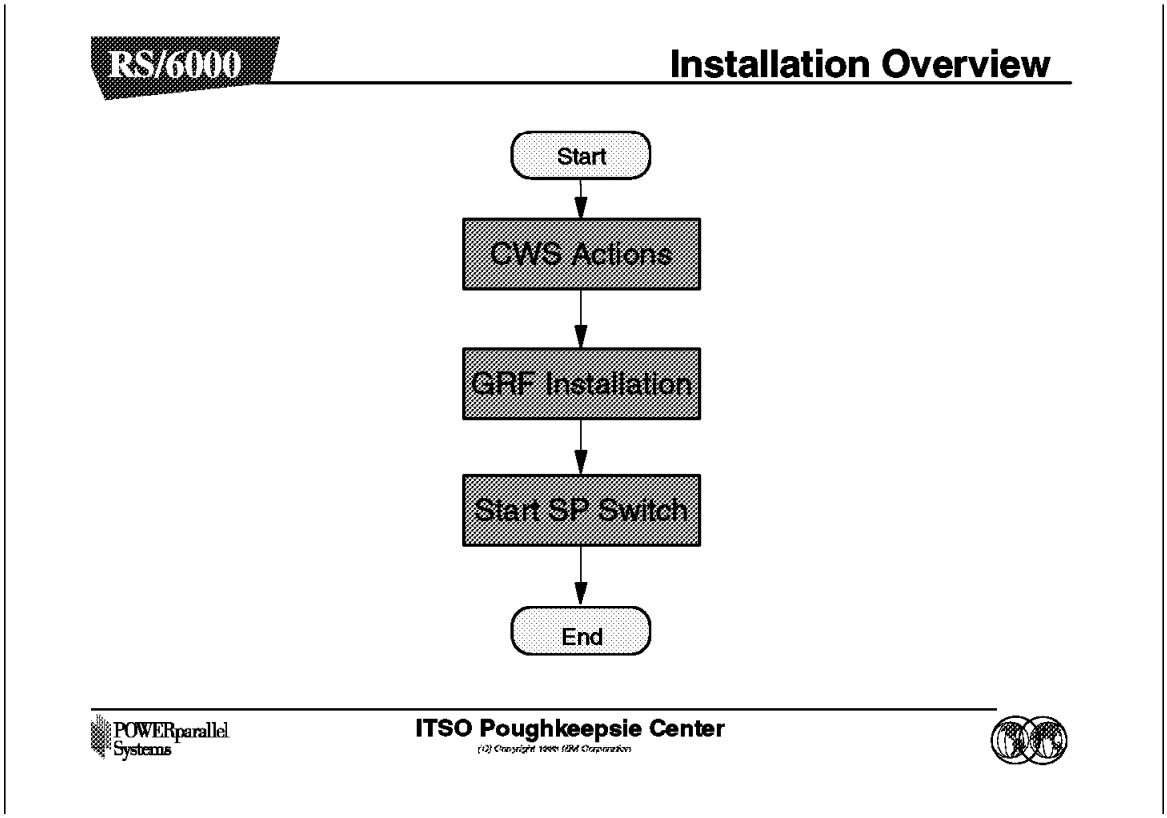
First, you need to supply an RS232 null modem cable and a VT100 terminal. The RS232 null modem cable is used to connect the IP Control Board (9-pin) to the VT100 terminal. The VT100 terminal must have the following settings:

- 9600 baud rate
- No parity
- Eight data bits
- One stop bit

For initial login, the user ID is root and the password is documented in the GRF publications.

Note: Since the VT100 terminal is only required for the initial configuration of the GRF and not for its operation, the user can use a PC to simulate the VT100 terminal.

6.5.5 Installation Overview

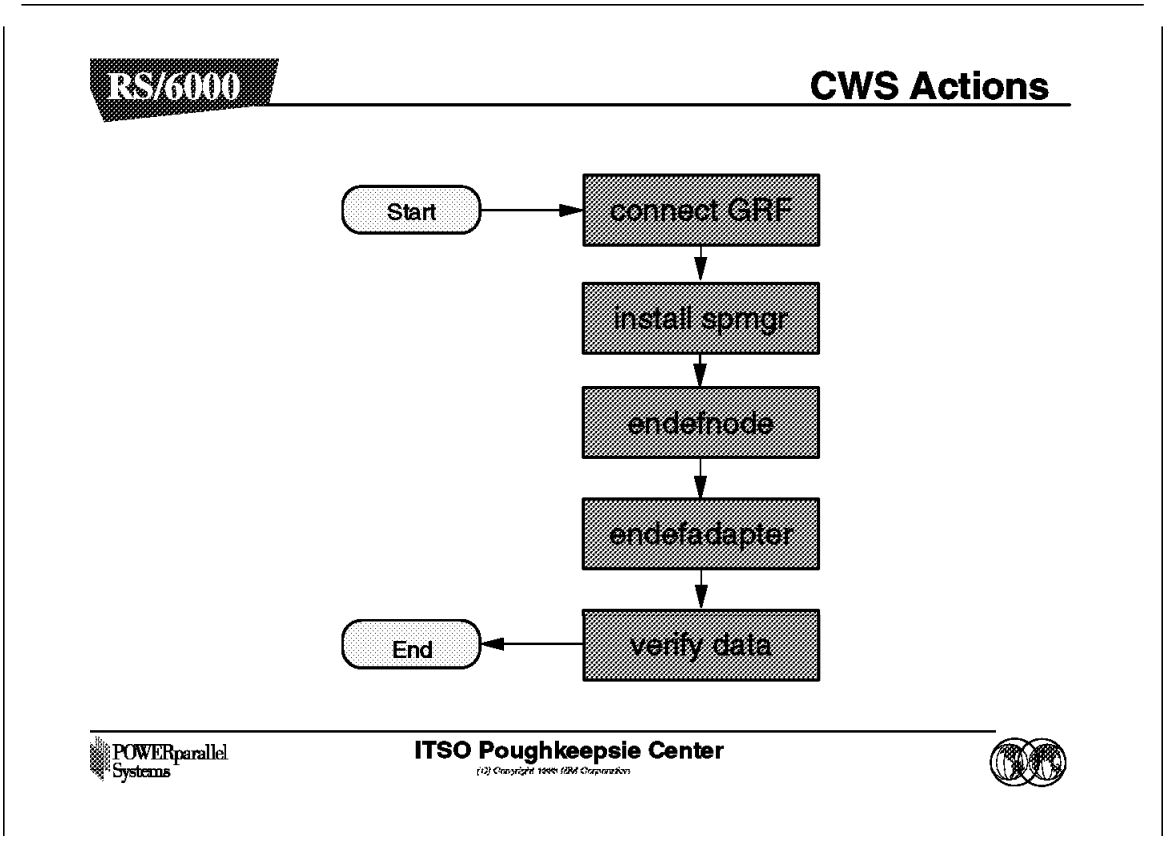


The installation of the dependent node in the RS/6000 SP involves these three steps:

1. Control Workstation actions
2. SP Switch router adapter in the GRF
3. Starting the SP Switch

These steps are discussed in more detail in the next three figures.

6.5.6 CWS Action



The first CWS action is to connect the RS/6000 SP to the GRF. This includes connecting the SP Switch cable, the GRF administrative Ethernet, and the GRF grounding cable.

Next, install the ssp.spmgr fileset on the Control Workstation, and ensure that the spmgr daemon is started.

Next, use the commands endefnode and endefadapter to define the dependent node. These commands are described in 6.4, “PSSP Enhancements” on page 177. Execute the two commands for *all dependent nodes* in the RS/6000 SP.

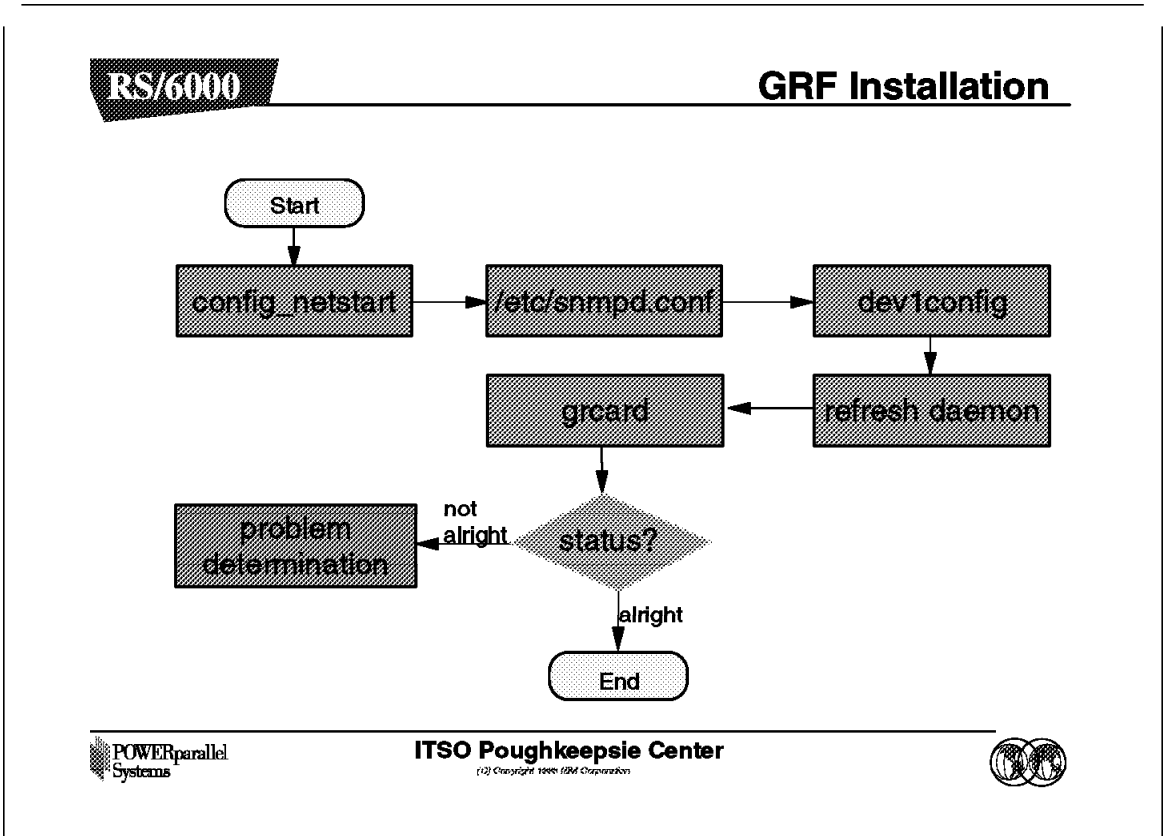
Finally, verify that the data used to define the dependent nodes were correct using the splstnodes and splstadapters commands, as follows:

```
# splstnodes -t dependent node_number switch_node_number \
> switch_chip_port switch_chip switch_number switch_partition_number \
> reliable_hostname management_agent_hostname extension_node_identifier \
> snmp_community_name
node_number  switch_node_number switch_chip_port switch_chip  switch_number
switch_partition_number reliable_hostname management_agent_hostname
extension_node_identifier snmp_community_name
          13           12            1            4            1            1
grf1.ppd.pok.ibm.com grf1.ppd.pok.ibm.com 03          ""
#
```

```
# splstadapters -t dependent node_number netaddr netmask
node_number  netaddr      netmask
         13 129.40.47.77 255.255.255.192
```

Note: To verify the node number used in the endefnode command and the actual switch connection, refer to the Scalable POWERparallel Switch (SPS) Bulkheads figure in the “Installation of RS/6000 SP Optional Features” chapter of *RS/6000 SP Maintenance Information Volume 1: Installation and CE Operations*, GC23-3903.

6.5.7 GRF Installation



1. For the GRF installation, perform the following:

   When the GRF is first powered on, it starts by asking a series of questions in the console to configure itself. These questions can be generated again, to change the GRF configuration, with the command `/etc/sbin/config_netstart` on the GRF. Following is a list of the questions.

   - `Host name for this machine? [ ]`

     Hostname for the GRF. Use the long name if DNS is used in the Control Workstation.

   - `Do you wish to configure the maintenance Ethernet interface? [yes]`

     Press Enter to take the default, yes. This is necessary to set up the GRF to work with the RS/6000 SP.

   - `IP address of this machine? [ ]`

     IP address of the GRF.

   - `Netmask for this network? [ ]`

     Netmask for the GRF.

   - `IP address for router ('none' for no default route)? [ ]`

     Default route for the GRF. Type `none` if none is available. This attribute creates a static route to an external router for routing packets in the administrative Ethernet network.

- Do you wish to go through the questions again? [yes]

  Here, the GRF will list all the parameters that you have typed in. Enter no if they are correct. To make corrections, just press Enter.

- Save a copy of this file as /etc/netstart.bak? [no]

  Specify yes to get a backup copy of the configuration.

2. Edit /etc/snmpd.conf and add the following lines to the end of the file:

```
MANAGER         <Control Workstation IP address>
                SEND ALL TRAPS
                TO PORT <SP manager port #>
                WITH COMMUNITY <SNMP community name>

COMMUNITY       <SNMP community name>
                ALLOW ALL OPERATIONS
                USE NO ENCRYPTION
```

   Replace the values in the <brackets> with site-defined parameters. This value is the same as the SNMP Community Name option defined in the endefnode command in the Control Workstation. To prohibit unauthorized SNMP Managers from configuring an extension node, change the existing public community name access to:

```
COMMUNITY      public
               ALLOW GET, TRAP OPERATIONS
               USE NO ENCRYPTION
```

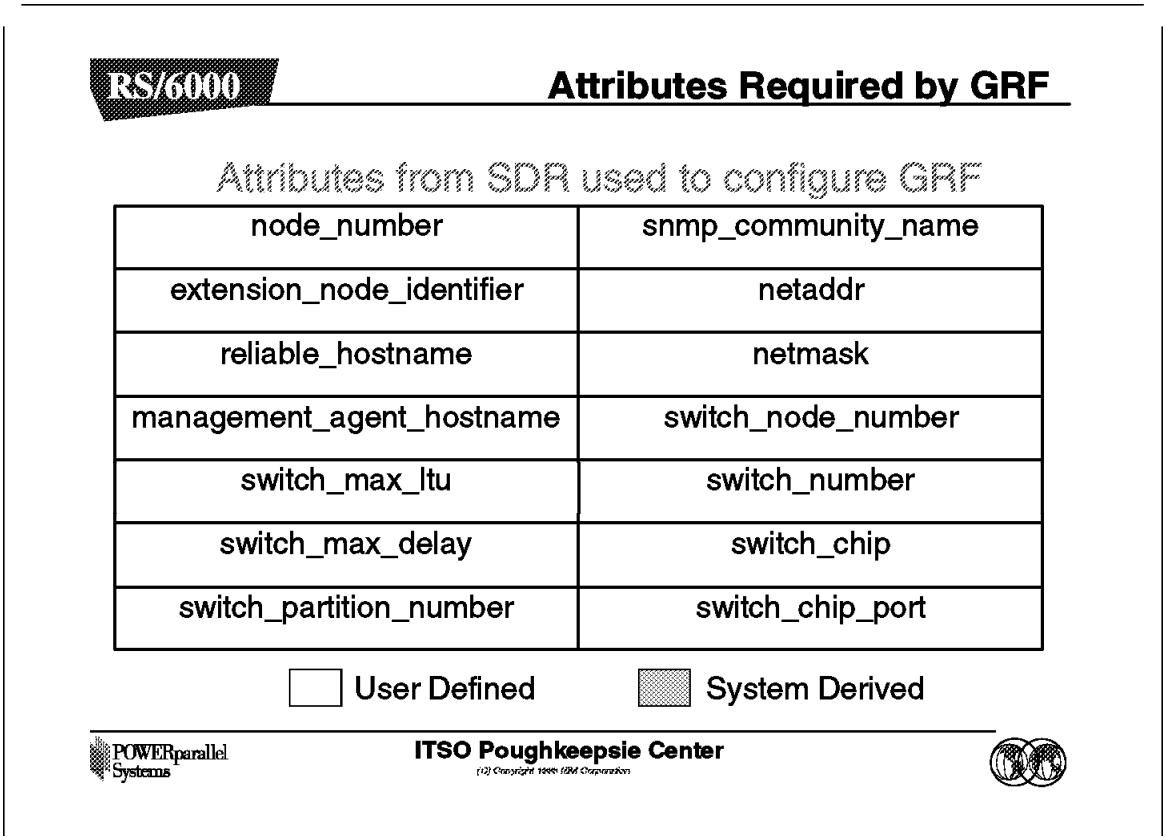
3. Execute dev1config to configure the dependent node on the GRF. Among other things, this command creates the /etc/grdev1.conf and /etc/grdev1.conf.template files, and also updates the /etc/grinchd.conf and /etc/grifconfig.conf files.

4. Next, on the GRF console, refresh the grinch daemon and the SNMP daemon. Use the ps ax and grep commands to list the process ID of the daemons. Execute the kill command on the two process for the two process to respawn themselves. Following is an example of this process:

```
# ps ax]grep grinch
15592  ??  S      0:00.51 grinchd -DNAGER         129.40.47.62
15811  p0  S+     0:00.02 grep grinch
# kill 15592
May  3 04:51:00 grf1 root: grstart:  grinchd exited status 143; restarting.
#
# ps ax]grep snmp
15600  ??  S      0:00.14 snmpd /etc/snmpd.conf /var/run/sn mpd.NOV
# kill 15600
May  3 04:54:43 grf1 mib2d[15605]: mib2d: terminated by master agent
May  3 04:54:43 grf1 root: grstart: snmpd exited status 143; restarting.
May  3 04:54:43 grf1 root: grstart: mib2d exited status 0; restarting.
```

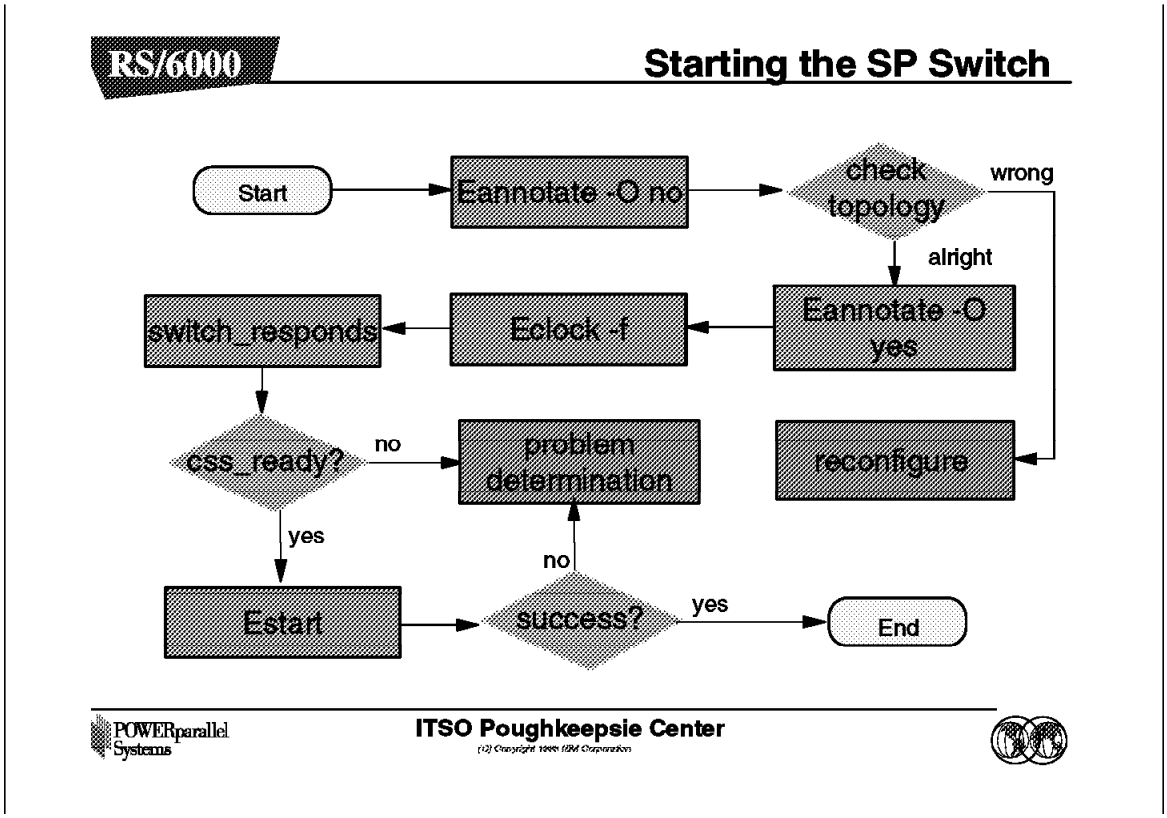
5. Finally, type grcard on the console. Check the status of the SP Switch router adapter. It will show the slot number, the adapter name, and the status of the card. The SP Switch router adapter is known as DEV1_V1 in this listing. If the status is loading, it means that it is polling the Control Workstation using the SNMP InfoNeeded trap to request configuration data, and you are done. If the status is not loading, there is a configuration problem with the SP Switch router adapter.

6.5.8 Attributes Required by GRF



This table shows the attributes required by the GRF to set up the SP Switch router adapter. They are all defined by the `endefnode` and `endefadapter` commands. Explanations of the commands and attributes are found in 6.4, “PSSP Enhancements” on page 177.

6.5.9 Starting the SP Switch



To start the SP Switch, first check the annotated switch file produced by the Eannotator command, without storing the topology file in the SDR.

If the annotated switch topology file shows the dependent node in the RS/6000 SP to be different from that stated in the endefnode command, it means that either the SP Switch router adapter is connected to the wrong SP Switch port, or the node number was entered incorrectly in the endefnode command. You need to either reconnect the SP Switch router adapter, or rerun the endefnode command to correct the problem before continuing. When updating the configuration with the endefnode or endefadapter command, specify the -r flag so that the GRF will be notified of the change and poll the Control Workstation for the update.

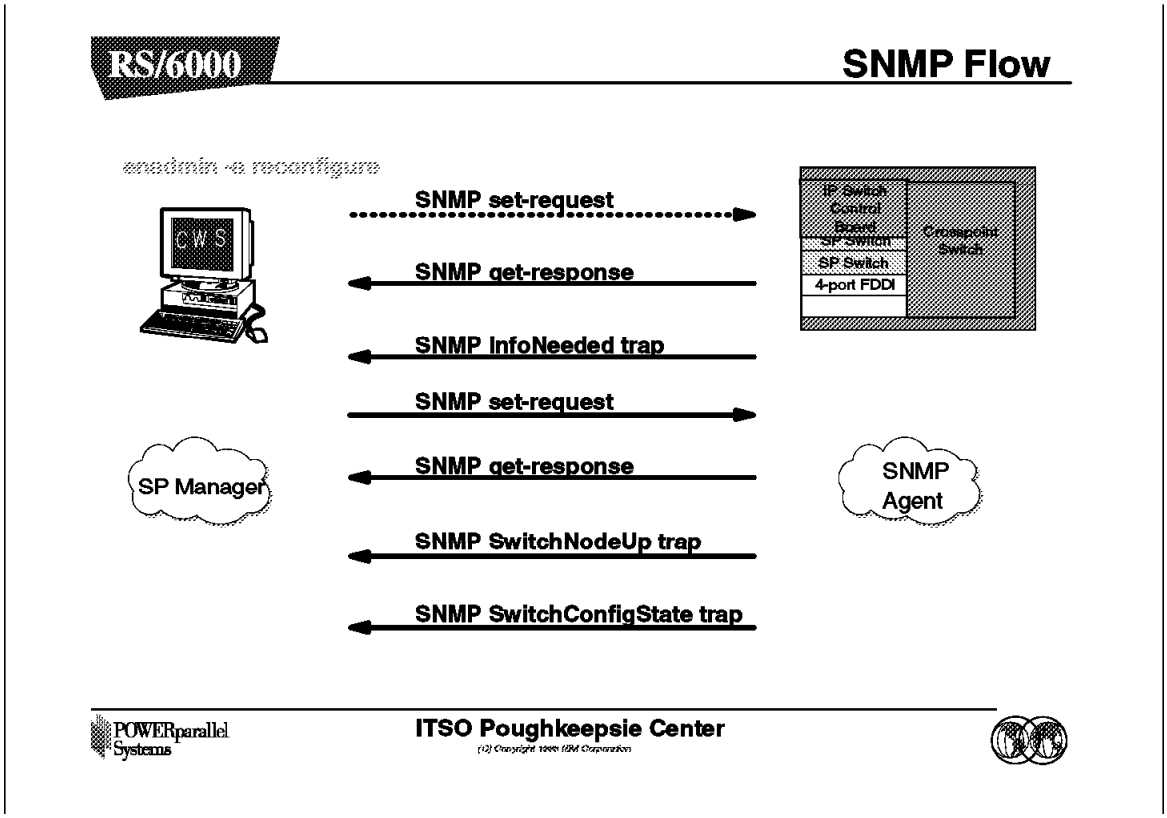
When the annotated file is correct, execute the Eannotator command again to store the topology file in the SDR. Run the Eclock command to reset the SP Switch clock, and reset the worm daemon on all standard RS/6000 SP nodes.

Use the SDRGetObjects switch_responds command to check the adapter_config_state attributes for all the dependent nodes.

If all adapter_config_state attributes are css_ready, run the Estart command. If any of the nodes' adapter_config_state attributes are not css_ready, the Estart will fail for the corresponding node. If any of the dependent nodes'

adapter_config_status is not css_ready, or the Estart fails, perform problem determination using the steps in 6.8, “Hints and Tips” on page 243.

6.5.10 SNMP Flow



The addition of the SP Switch router adapter adds four specific traps to the SNMP:

SNMP InfoNeeded
SwitchNodeUP
SwitchNodeDown
SwitchConfigState.

Except for these traps, most other SNMP traps generated by the GRF are ignored by the SP Extension Node SNMP Manager. However, if the user has another SNMP manager in the network, it can query adapter configuration and state information and monitor the flow of SNMP traps between the GRF Agent and the SP Extension Node SNMP Manager on the Control Workstation.

When the GRF is first powered on, it periodically sends the InfoNeeded SNMP trap to the SP Extension Node SNMP Manager for configuration information. Alternatively, the `enadmin -a reconfig` command will send an SNMP set-request containing an extension node identifier and an administrative status of reconfigure to the GRF to trigger the InfoNeeded trap.

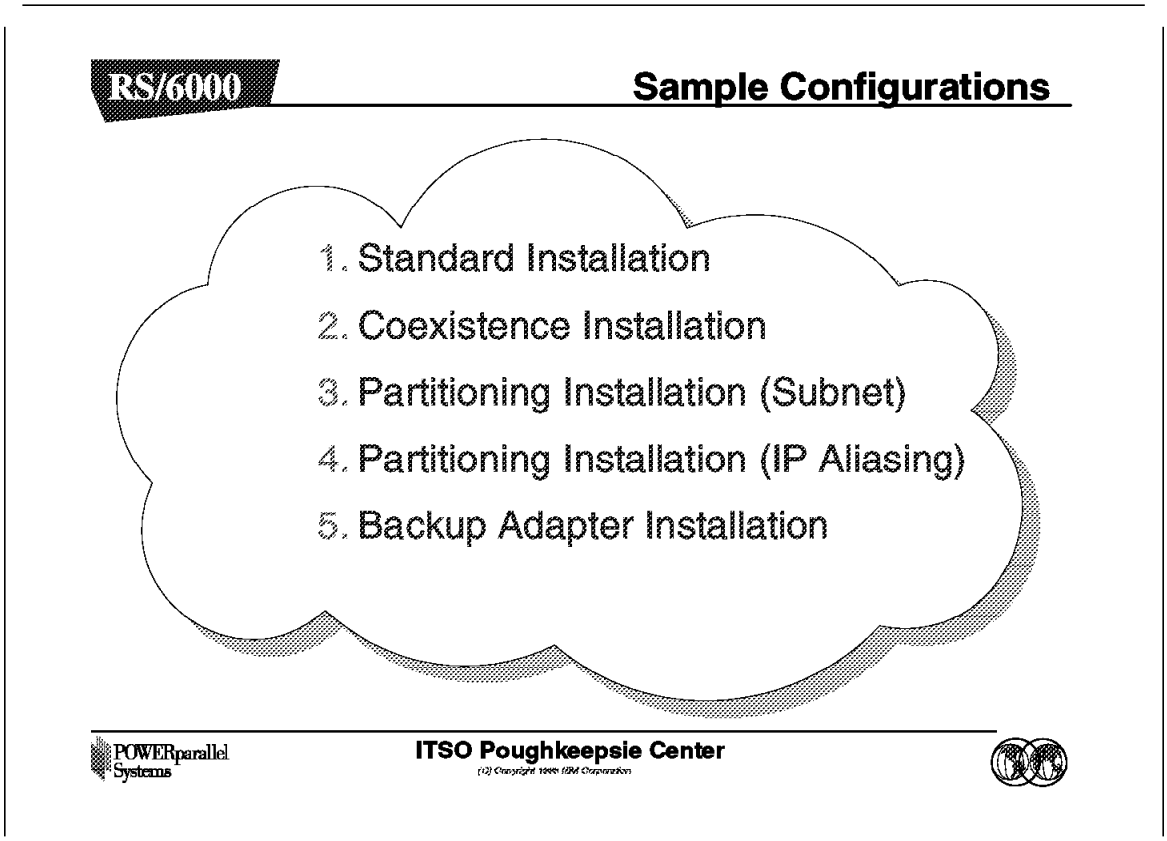
When the Control Workstation receives the InfoNeeded trap, it sends a SNMP set-request containing the extension node identifier and the configuration attributes for that dependent node to the GRF SNMP Agent at UDP port 161. When the GRF Agent has received all the configuration information, it sends an SNMP get-response to the SP Extension Node SNMP Manager on the Control

Workstation. The information is then applied to the SP Switch router adapter, and the GRF sends two SNMP traps, SwitchNodeUp and SwitchConfigState, to indicate that it is ready.

Notes:

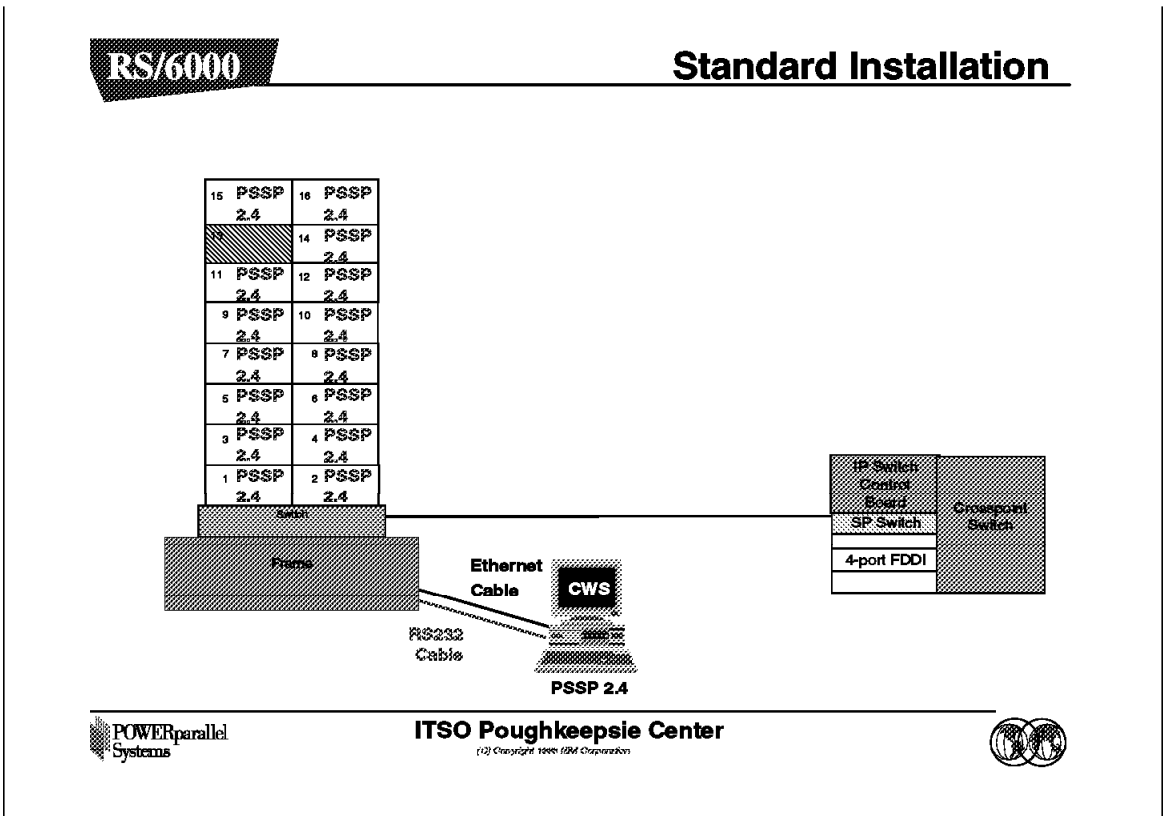
1. When an Estart or Eunfence is issued, processing is done via link-level service packet exchanges between the dependent node and the Primary node using the SP Switch. The Primary node next sets switch_responds for the dependent node.
2. When Efence is issued, a SwitchNodeDown SNMP trap is sent by the GRF SNMP Agent. Via link-level service packet exchanges between the dependent node and Primary node, the Primary node sets the switch_responds for the dependent node.
3. When the dependent node enables or reenables its SP Switch interface, the GRF SNMP Agent sends a SwitchNodeUp SNMP trap to the SP SNMP Manager. If the Efence command was previously issued with the autojoin option to remove the dependent node from the SP Switch network, the SNMP Manager will issue the Eunfence command to allow the dependent node to join the SP Swtich network.

6.6 Sample Configurations



This section offers some sample configurations for using the GRF with the RS/6000 SP.

6.6.1 Standard Installation



This example describes the steps for an installation of GRF into an RS/6000 SP system.

1. On the Control Workstation, we install the SP Extension Node SNMP Manager:

   ```
   # installp -a -d /spdata/sys1/install/pssplpp/PSSP-2.4 -X ssp.spmgr
        .
        .
        .
        .
   Installation Summary
   --------------------
   Name                        Level           Part        Event       Result
   -------------------------------------------------------------------------------
   ssp.spmgr                   2.4.0.0         USR         APPLY       SUCCESS
   ssp.spmgr                   2.4.0.0         ROOT        APPLY       SUCCESS
   ```

   Here we assume that the RS/6000 SP software is in the /spdata/sys1/install/pssplpp/PSSP-2.4 directory. We could have similarly installed it from tape (assuming tape drive 0) using `installp -a -d /dev/rmt0 -X ssp.spmgr`.

2. We check for its status using `lssrc -s spmgr`. Start the SP Extension Node SNMP Manager on the Control Workstation if it is not started. We turn on tracing on the SP Extension Node SNMP Manager to provide us with more information should the dependent node installation fail. The following

example checks for the activity of the spmgr daemon, starts it, verifies that it is active, and turns tracing on for the daemon:

```
# lssrc -s spmgr
Subsystem         Group            PID     Status
 spmgr                                     inoperative
#
# startsrc -s spmgr
0513-059 The spmgr Subsystem has been started. Subsystem PID is 17574.
#
# lssrc -s spmgr
Subsystem         Group            PID     Status
 spmgr                             17574   active
#
# traceson -ls spmgr
Start trace.
0513-091 The request to turn on tracing was completed successfully.
```

If you intend to run with tracing enabled during production, you can limit the size of the trace table by specifying the maximum size as a -s switch value to be passed to the spmgr daemon when it is started (for example: startsrc -s spmgr -a -s <size>).

3. We define the dependent node on the Control Workstation with endefnode and endefadapter, using the following parameters:
   - The GRF hostname is grf1.ppd.pok.ibm.com.
   - SP Switch router adapter is in Slot 2 of the GRF.
   - The dependent node number is 13.
   - The IP address for the dependent node is 129.40.47.77 (the IP address of the SP Switch for node 13 in the RS/6000 SP).
   - The netmask for the dependent node is 255.255.255.192 (the netmask for the SP Switch on the RS/6000 SP).

   ```
   # endefnode -a grf1.ppd.pok.ibm.com -i 02 -s grf1.ppd.pok.ibm.com 13
   The endefnode command has completed successfully.
   #
   # endefadapter -a 129.40.47.77 -m 255.255.255.192 13
   The endefadapter command has completed successfully.
   ```

4. After setting up the dependent node on the RS/6000 SP, we set up the GRF. The following questions are asked when the GRF is powered up for the first time, or these questions can be activated using the config_netstart command on the GRF:
   - The hostname is grf1.ppd.pok.ibm.com (the hostname for the GRF's administrative Ethernet).
   - Answer yes to configure the maintenance Ethernet.
   - Use 129.40.41.47 (the IP address defined for the GRF's administrative Ethernet).
   - Use 255.255.255.0 (netmask for the GRF's administrative Ethernet).
   - Use 129.40.47.62 for the default route.
   - Specify no to avoid going through the questions again.
   - Specify yes to save a copy of the configuration in /etc/netstart.bak.

5. Next, we configure the GRF to communicate with the Control Workstation. Append the following lines to /etc/snmpd.conf in the GRF:

```
MANAGER         129.40.47.62
                SEND ALL TRAPS
                TO PORT 162
                WITH COMMUNITY spenmgmt

COMMUNITY       spenmgmt
                ALLOW ALL OPERATIONS
                USE NO ENCRYPTION
```

6. Next, execute dev1config to configure the SP Switch Router Adapter and refresh the SNMP and grinch daemons:

```
# dev1config
#
# ps ax]grep grinch
15592  ??  S      0:00.51 grinchd                 129.40.47.62
15811  p0  S+     0:00.02 grep grinch
#
# kill 15592
May  3 04:51:00 grf1 root: grstart:  grinchd exited status 143; restarting.
#
# ps ax]grep snmp
15600  ??  S      0:00.14 snmpd /etc/snmpd.conf /var/run/sn mpd.NOV
# kill 15600
May  3 04:54:43 grf1 mib2d[15605]: mib2d: terminated by master agent
May  3 04:54:43 grf1 root: grstart: snmpd exited status 143; restarting.
May  3 04:54:43 grf1 root: grstart: mib2d exited status 0; restarting.
```

7. Execute the grcard command on the GRF, and check to make sure that the SP Switch router adapter, known as DEV1_V1, is running:

```
# grcard
0       ETHER_V1        running
2       DEV1_V1 running
3       HIPPI_V1        running
4       HSSI_V1 running
```

8. We return to the Control Workstation to start the SP Switch. We run the Eannotator and Eclock commands before starting the SP Switch with an Estart.

```
# Eannotator -F /etc/SP/expected.top.2nsb.0isb.0 \
> -f /etc/SP/ann.2nsb.0isb -O no
```

   The annotated file is checked for correct dependent node positioning before storing it into the SDR and setting the SP Switch clock, as follows:

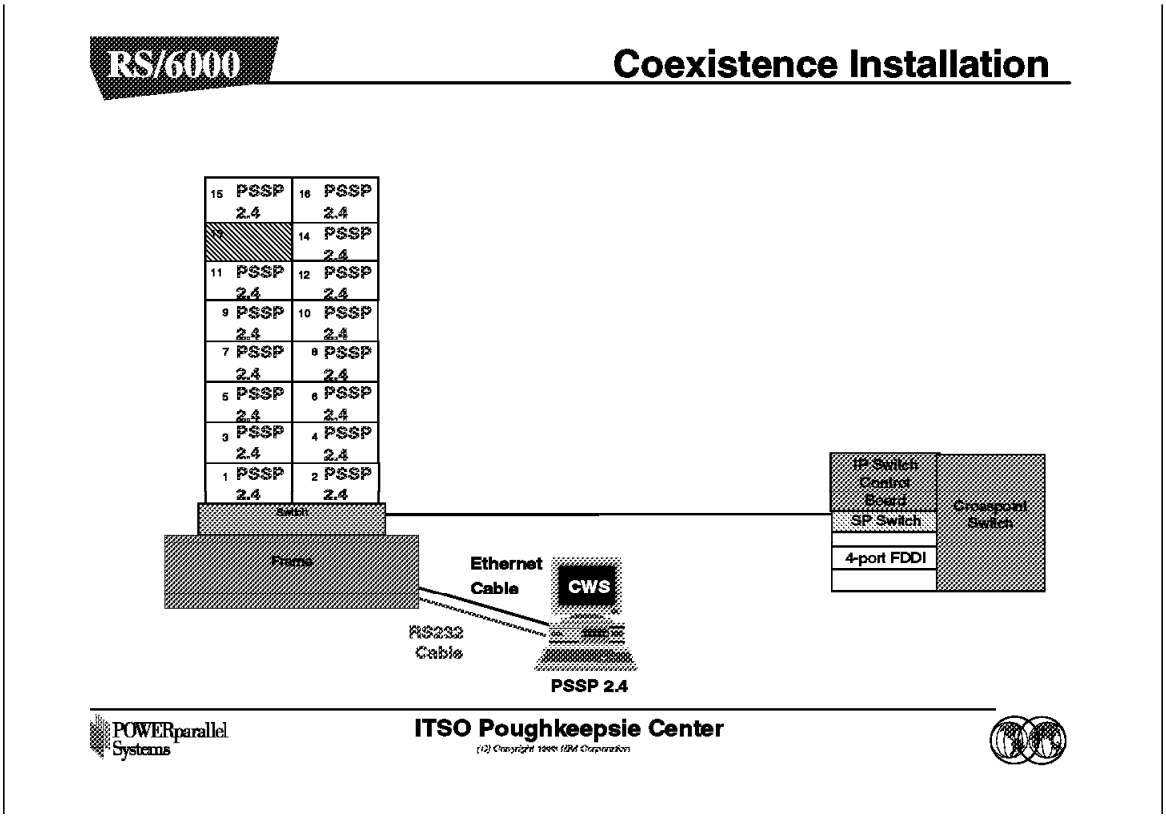
```
# more /etc/SP/ann.2nsb.0isb
     .
     .
     .
s 14 1  tb3 12 0           E01-S17-BH-J33 to E01-N13 # Dependent Node
#
# Eannotator -F /etc/SP/expected.top.2nsb.0isb.0 \
> -f /etc/SP/ann.2nsb.0isb -O yes
# Eclock -f /etc/SP/Eclock.top.2nsb.0isb.0
```

9. Finally, we check the SDR class switch_responds to ensure that the adapter_config_status of the dependent nodes is css_ready before starting the SP Switch:

```
# SDRGetObjects -G switch_responds
node_number  switch_responds autojoin     isolated     adapter_config_status
     .
     .
     .
          13            1            0            0 css_ready
#
# Estart
Switch initialization started on ceed1n05.ppd.pok.ibm.com.
Initialized 5 node(s).
Switch initialization completed.
```

The number of nodes initialized includes both standard and dependent nodes in the RS/6000 SP partition.

6.6.2 Coexistence Installation

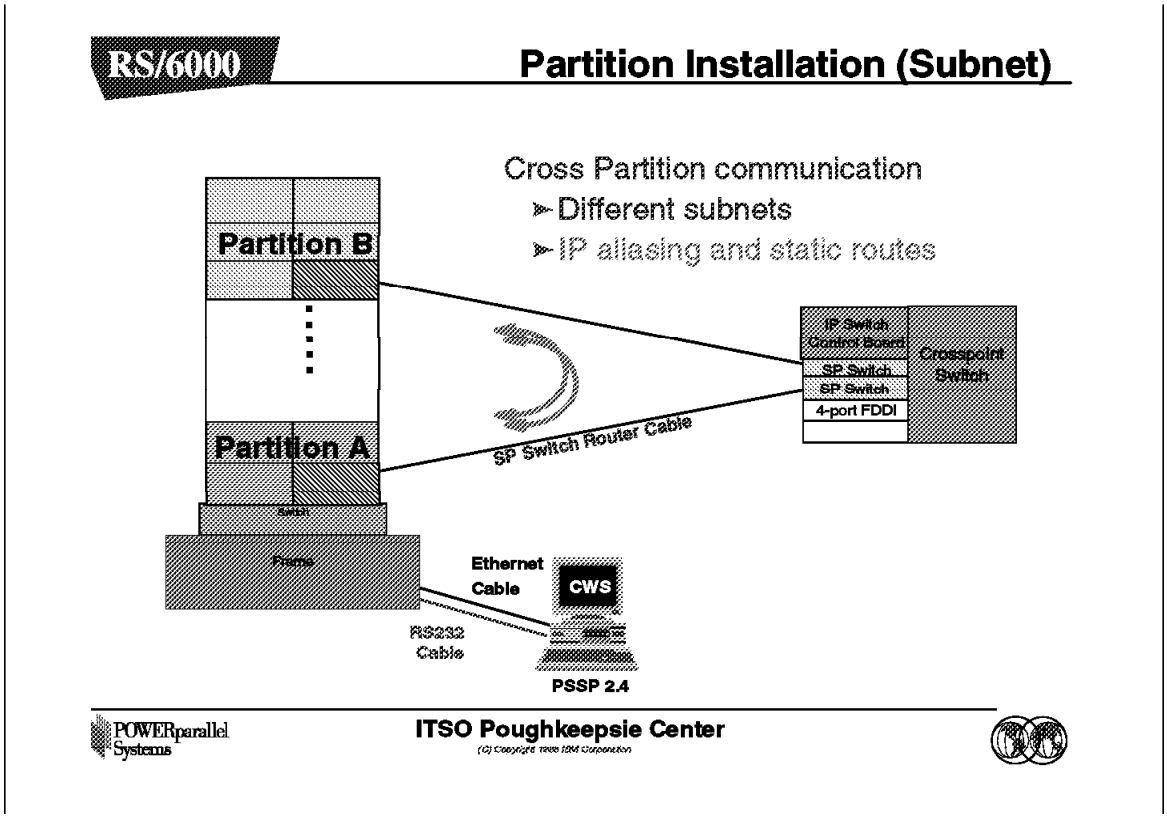


The requirements for a dependent node to be installed in a partition with multiple PSSP levels are outlined in the Coexistence figure of 6.4, “PSSP Enhancements” on page 177.

Here, in our Coexistence Installation example, we assume an RS/6000 SP with multiple PSSP levels in a partition, complying with the coexistence requirements. We also assume that the nodes in the partition are able to communicate with each other through the SP Switch. If these conditions are met, the installation of the dependent node in this scenario is the same as the standard installation shown in the previous figure.

This installation allows the non-PSSP 2.3 RS/6000 SP nodes to work with the dependent node in the partition.

6.6.3 Partition Installation (Subnet)



In this example, we install the RS/6000 SP such that SP Switch adapters in different partitions have different subnets. We assume the availability of a single frame RS/6000 SP system, with two partitions of eight nodes each. The IP address for the SP Switch network with a netmask of 255.255.255.0 is as follows:

- Partition A
  - Node 1: 129.40.47.1 a1
  - Node 2: 129.40.47.2 a2
  - Node 5: 129.40.47.5 a5
  - Node 6: 129.40.47.6 a6
  - Node 9: 129.40.47.9 a9 (dependent node)
- Partition B
  - Node 3: 129.40.48.3 b3
  - Node 4: 129.40.48.4 b4
  - Node 7: 129.40.48.7 b7
  - Node 8: 129.40.48.8 b8
  - Node 11: 129.40.48.11 b11 (dependent node)

Use the instructions of the Standard Installation figure in this section to install the dependent node in each partition.

First perform Steps 1 and 2.

Perform Step 3 twice, once for dependent node 9 and once for dependent node 11. However, in this case use the IP address and netmask of a9 and b11 instead. Use Slots 02 and 03 of the GRF for each of the dependent nodes.

The endefnode and endefadapter commands should be executed in the each partition. Before executing these commands, however, set the appropriate partition by executing export SP_NAME=<partition name>.

Next, perform Steps 4, 5, 6 and 7. For Step 7, grcard should show DEV_V1 running on Slot 2 and 3 instead.

For Step 8, the topology files for a partitioned RS/6000 SP are found in the /spdata/sys1/syspar_configs/topologies directory. The correct topology file to use with Eannotator can be listed by the SDRGetObjects Switch_partition topology_filename command.

Note that the topology file listed in this manner ends with a dot and a number. This is the version number of the topology file stored in the SDR. When using the Eannotator command, ignore this version number. If you list the topology files in the /spdata/sys1/syspar_configs/topologies directories, you will notice that the partitioned RS/6000 SP topology files end with “isb.”

Again, perform Step 8 twice, once for each partition.

Finally, perform Step 9 to complete the definition of the dependent nodes. When SDRGetObjects -G switch_responds is performed, check adapter_config_status for both dependent nodes to ensure that both are in the css_ready state. Do Estart twice, once for each partition.

Next, we set up the RS/6000 SP nodes so that they can communicate with each other across partitions. In partitioning, nodes in one partition do not communicate with nodes in other partitions. They can communicate with the dependent nodes in their own partition. In order for them to communicate across partitions using the GRF, we need to set up routes.

Finally, we set up static routes from each node to enable it to communicate via GRF with the nodes in the other partition. For *every node* in Partition A, execute the following statement to add a static route to Partition B:

```
# route add -net b3 -netmask 255.255.255.0 a9
a9 net b3: gateway a9
#
```

And for *every node* in Partition B, execute the following statement to add a static route to Partition A:

```
# route add -net a1 -netmask 255.255.255.0 b11
b11 net a1: gateway b11
#
```

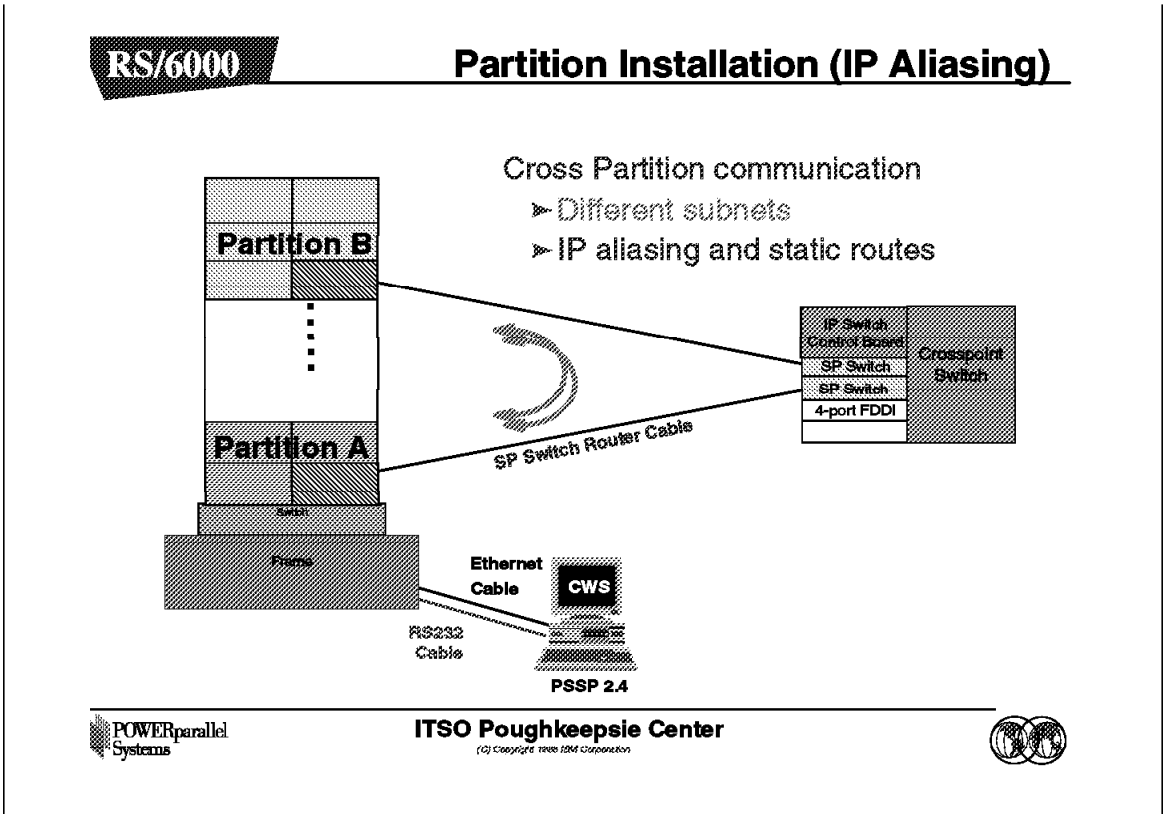
Now the RS/6000 SP nodes in Partition A are able to communicate with nodes in Partition B. In order for the routes to be available after a reboot, add them to the /etc/rc.net file. Alternatively, these routes could be set up using dynamic

routing protocols, such as RIP or OSPF, that are supported both on the RS/6000 SP and the GRF.

Attention

In order for communication to be possible across partitions through the SP Switch, switch_responds must be green for the source node, the destination node and the dependent nodes in each partition.

6.6.4 Partition Installation (IP Aliasing)



Partitions in the RS/6000 SP are separate networks to the system. Therefore, they should use different subnet masks; this is a requirement when we want them to talk to each other through a single GRF. When both partitions are in the same subnet, the routing table on the GRF will only register one of the routes to the RS/6000 SP. Thus, one of the partitions is not reachable through the GRF.

In this example, we show how to make the partitions talk to each other when they are in a single subnet. The IP addresses of the RS/6000 SP Switch and their aliases with a netmask of 255.255.255.0 are as follows:

- RS/6000 SP
  - Node 1: 129.40.49.1 c1
  - Node 2: 129.40.49.2 c2
  - Node 3: 129.40.49.3 c3
  - Node 4: 129.40.49.4 c4
  - Node 5: 129.40.49.5 c5
  - Node 6: 129.40.49.6 c6
  - Node 7: 129.40.49.7 c7
  - Node 8: 129.40.49.8 c8
  - Node 9: 129.40.49.9 c9 (dependent node)

  - Node 11: 129.40.49.11 c11 (dependent node)

- Partition A (aliases)
  - Node 1: 129.40.47.1 a1
  - Node 2: 129.40.47.2 a2
  - Node 5: 129.40.47.5 a5
  - Node 6: 129.40.47.6 a6
  - Node 9: 129.40.47.9 a9 (dependent node)
- Partition B (aliases)
  - Node 3: 129.40.48.3 b3
  - Node 4: 129.40.48.4 b4
  - Node 7: 129.40.48.7 b7
  - Node 8: 129.40.48.8 b8
  - Node 11: 129.40.48.11 b11 (dependent node)

To set up these addresses, follow the instructions described in 6.6.3, “Partition Installation (Subnet)” on page 233. Use the address listed in the RS/6000 SP bullet for the SP Switch here for both partitions instead.

Next, set up the alias on the SP Switch router adapters on the GRF by editing the /etc/grifconfig.conf file in the GRF. Here, the adapters are in Slots 02 and 03.

```
      .
      .
      .
gt020      129.40.49.9     255.255.255.0
gt020      129.40.47.9     255.255.255.0
gt030      129.40.49.11    255.255.255.0
gt030      129.40.48.11    255.255.255.0
```

After inserting the two statements on gt020 and gt030, save the file and reset the two adapters. This activates the aliases.

```
# grreset 2
Ports reset: 2
# grreset 3
Ports reset: 3
May 13 00:40:43 classgig kernel: gt020:  GigaRouter DEV1, GRIT address 0:2:0
May 13 00:40:43 classgig kernel: gt020:  GigaRouter DEV1, GRIT address 0:2:0
May 13 00:40:46 classgig kernel: gt030:  GigaRouter DEV1, GRIT address 0:3:0
May 13 00:40:46 classgig kernel: gt030:  GigaRouter DEV1, GRIT address 0:3:0
#
```

Next, set up the alias for the SP Switch adapter on the RS/6000 SP nodes via the `ifconfig` command. To set up the alias on node 1, use the following command:

```
# ifconfig css0 a1 netmask 255.255.255.0 alias
```

To check whether it was successful, use the `netstat -i` command.

Execute these commands on all RS/6000 SP nodes, using the appropriate IP alias.

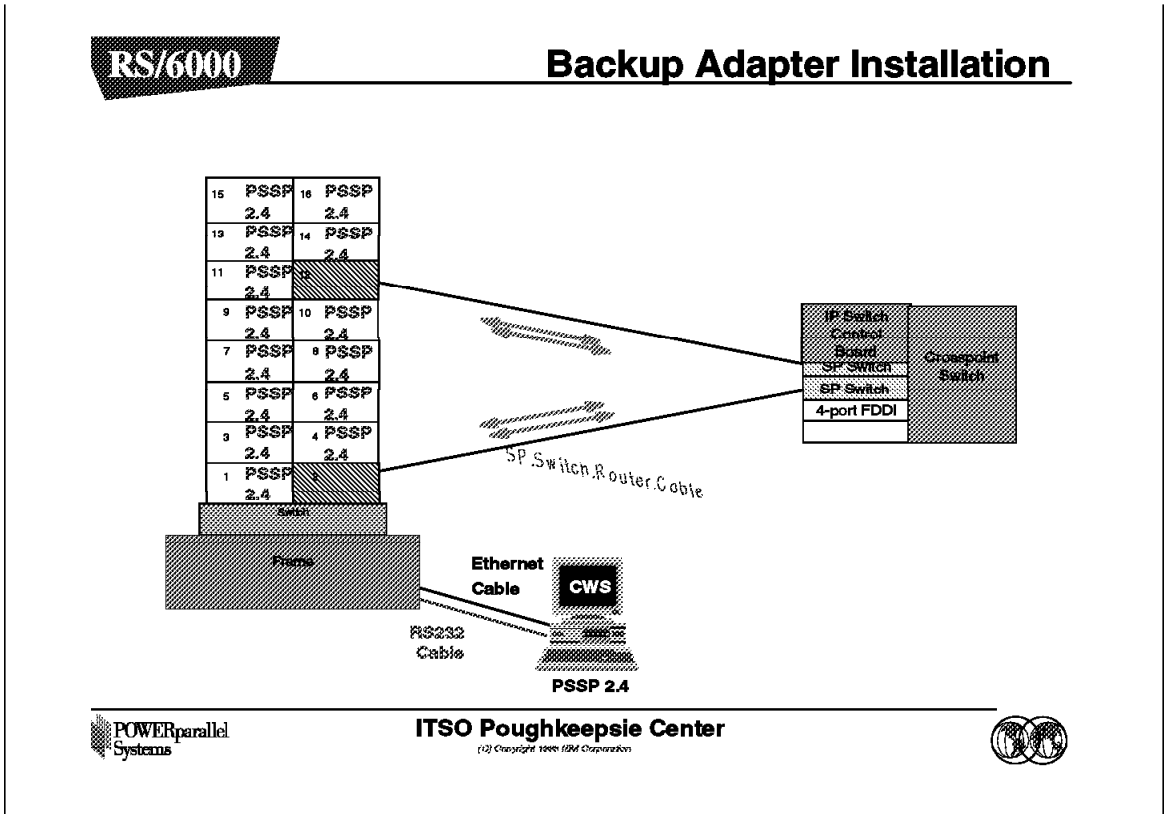
Finally, on each RS/6000 SP node, add a static route to reach the nodes on the other partition:

```
# route add -net b3 -netmask 255.255.255.0 a9
a9 net b3: gateway a9
#
```

This is similar to adding the static routes in the previous example, 6.6.3, “Partition Installation (Subnet)” on page 233.

Now the RS/6000 SP nodes in Partition A can communicate with nodes in Partition B, and vice versa, using the IP aliases. In order for these routes to be available after a reboot, insert the route add commands in the /etc/rc.net file. Alternatively, these routes can be set up using dynamic routing protocols, such as RIP or OSPF, that are supported both on the RS/6000 SP and the GRF.

6.6.5 Backup Adapter Installation



In this example, we show how to install two dependent nodes in one partition. As mentioned in the previous examples, when we have more than one media card with the same subnet on the GRF, only one of them is recorded in the GRF's routing table. Connecting the RS/6000 SP to the GRF in this manner gives us additional availability. Should one of the media cards be unavailable, the other media card will take over. For the RS/6000 SP, the same happens when we connect two SP Switch router adapters to the same partition.

For this example to work, the whole network has to run OSPF. On the RS/6000 SP, at least one node on each partition must run OSPF. OSPF must be running on the GRF as well.

OSPF will configure the routes to the GRF using different weights. Normally, communication between the GRF and the RS/6000 SP uses the SP Switch router adapter with a lower-weight route. When that SP Switch router adapter is unavailable, the corresponding route is also unavailable and all IP traffic (except that using static routes) dynamically reroutes to the other route with the active SP Switch router adapter. In this manner, it enhances the availability of the RS/6000 SP connection to the GRF in that partition.

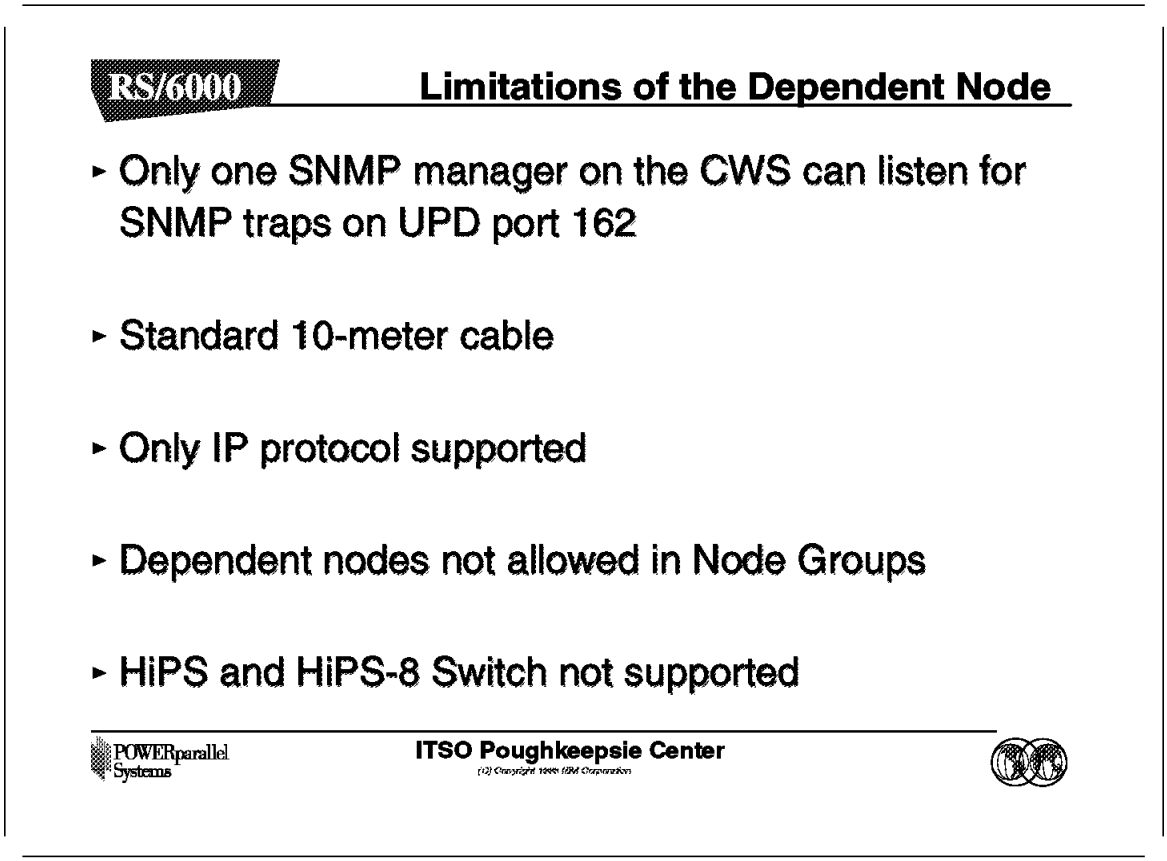
Note: In this example, availability can be enhanced by connecting two GRFs, each with an SP Switch router adapter, to a single partition. Should the GRF in this example fail, all routes going through the SP Switch will be

unavailable. With two GRFs, when one fails, the other will still be available.

The requirements for this example are exactly the same as those for the Backup Adapter Installation example. OSPF must be running on at least one node of the partition and on both GRFs. In addition, both GRFs must be interconnected by a TCP/IP media like HIPPI or FDDI, and this link must be active.

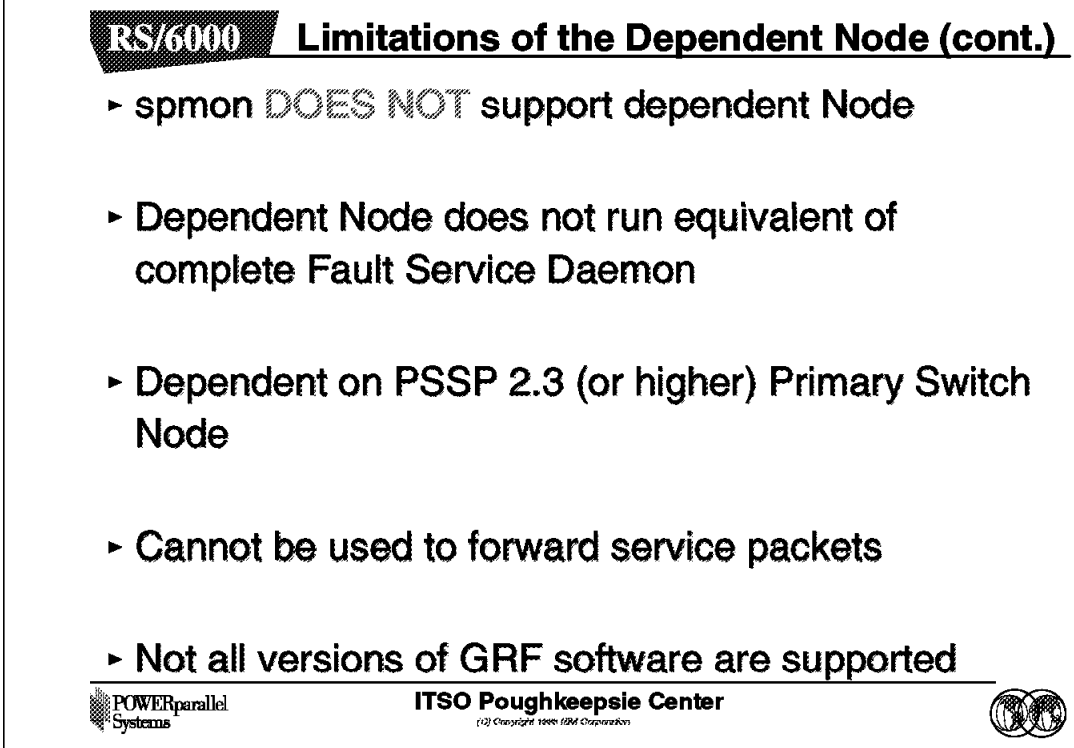
Lastly, in this example, since there are two GRFs, there are also two routing tables available, one on each GRF. Each GRF records the route created by the SP Switch Router Adapter, even though they are in the same subnet. This offers greater flexibility in assigning IP packets between the two routes, and in balancing the IP load.

6.7 Limitations of the Dependent Node



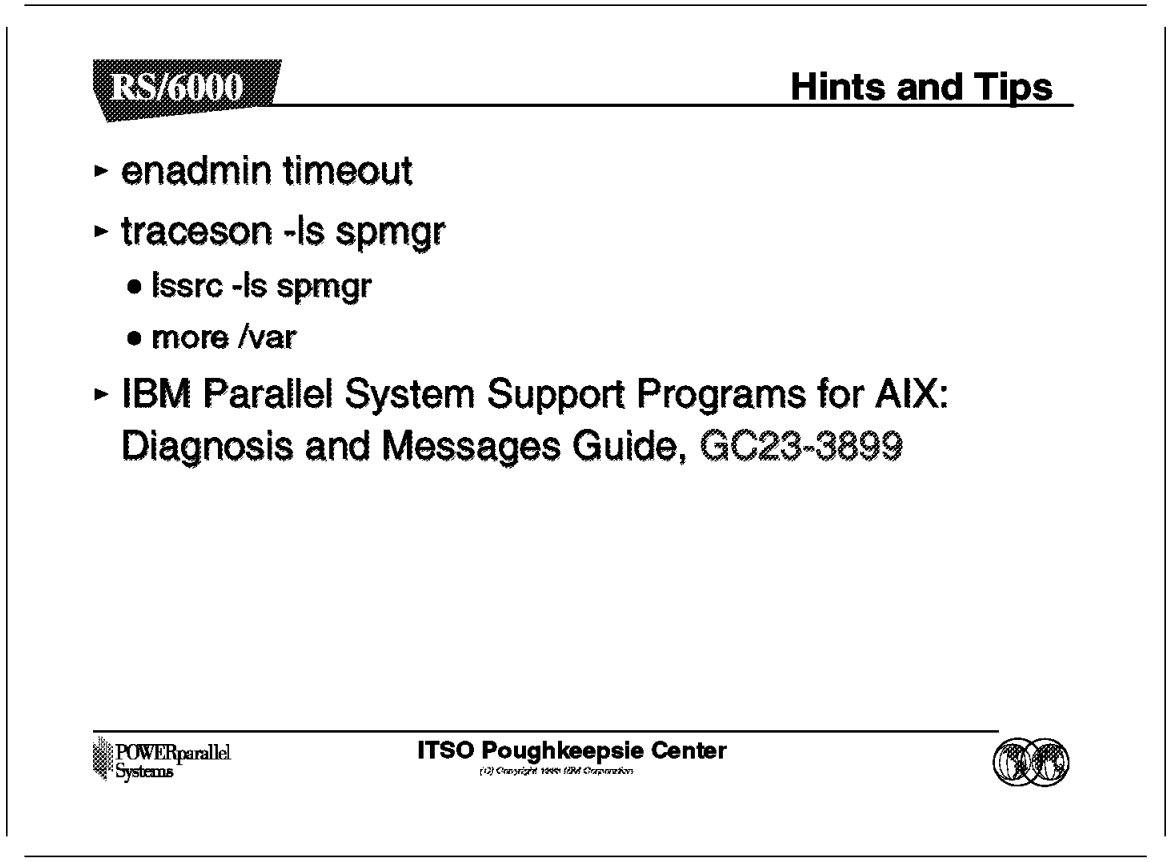
Following are limitations associated with use of the dependent node:

- To use the dependent node in an RS/6000 SP requires the SP Extension Node SNMP Manager to be installed in the Control Workstation. The SP Extension Node SNMP Manager requires UDP port 162 in the Control Workstation. Other SNMP managers, such as Netview, also require this port. To allow the two SNMP managers to coexist, the SP Extension Node SNMP Manager must use an alternative udp port. This process is documented in 6.5, “Installation” on page 210.
- The standard cable provided to attach the GRF to the RS/6000 SP is only 10 meters long. This means that the GRF is within 10 meters of the RS/6000 SP, which may not be far enough for some customers. A longer cable is available through RPQ. The drawback of using a longer cable is that it cannot be wrap tested to see if it is faulty.
- The GRF only supports TCP/IP routing. Thus the dependent node does not support any other protocols such as SNA or user space, that are commonly associated with the RS/6000 SP.
- Dependent nodes are not allowed in Node Groups.
- Only the 8-port and 16-port SP Switch is supported. The 8-port and 16-port High Performance Switch (the old SP switch) cannot be connected to the SP Switch Router Adapter on the GRF.



- The spmon command on the RS/6000 SP is not enhanced to support dependent nodes. Dependent nodes can only be viewed from the perspectives command.
- The fault service daemon runs on all switch nodes in the RS/6000 SP, but not on the dependent node. As such, the dependent node does not have the full functionality of a normal RS/6000 SP Switch node.
- The dependent node requires the SP Switch's Primary node to compute its switch routes. If the Primary node is not at PSSP 2.3, the dependent node cannot work with the RS/6000 SP.
- In the RS/6000 SP, SP Switch nodes occasionally send service packets from one node to the next to keep track of status and links. Sometimes these packets are sent indirectly through another switch node. As the dependent node is not a standard RS/6000 SP Switch node, it cannot be used to forward service packets to other nodes.
- The SP Switch Router comes preloaded with its operating system. To do an upgrade, users will have to download the latest level from the IBM FTP server used for 9077 support. At the time of publication, that server is expected to be service2.boulder.ibm.com. IBM will provide service updates and new levels of the SP Switch Router software on that server. The only GRF software supported on the SP Switch Router will be those versions that are provided by the IBM FTP server.

6.8 Hints and Tips



When installing the dependent node, it is recommended that you turn on tracing for the SP Extension Node SNMP Manager so that valuable information will be available in the snmp log file should the installation fail. To turn on tracing, either specify the `-l` or `-s` flag on the `startsrc` command when the spmgr subsystem is started. Alternatively, use the `traceson -ls spmgr` command if tracing was not specified when the spmgr subsystem was started. To turn it off, use `tracesoff -s spmgr`.

If the output of `enadmin`, `endefnode -r`, or `endefadapter -r` shows a timeout, check the trace file /var/adm/SPlogs/spmgr/spmgrd.log for messages shown in Table 10 in order to perform the corresponding recovery action.

Table 10 (Page 1 of 2). SNMP Trace File Messages

Symptom	Recovery
`init_io failed: udp port in use.`	If you find this message, then port 162 on the Control Workstation is already in use. Change the `spmgr-trap` port number in `/etc/services` in the Control Workstation, and `/etc/snmpd.conf` in the GRF.

Table 10 (Page 2 of 2). SNMP Trace File Messages	
Symptom	**Recovery**
2536-007 An authentication failure notification was received from an SNMP Agent running on the host <hostname> which supports Dependent Nodes.	The SNMP community name in the DependentNode and the GRF do not match. Correct it in the DependentNode using endefnode, or on the GRF by editing the /etc/snmpd.con file.
No authentication error message in the trace file.	Correct the dependent node's management_agent_hostname in the DependentNode class by using endefnode.

Using the command lssrc -ls spmgr, check for the message switchInfoNeeded trap message is not being received. If that is the case, check the IP address of the Control Workstation in the /etc/snmpd.conf file in the GRF. Correct the address and restart the snmp daemon in the GRF.

If lssrc -ls spmgr produces the switchInfoNeeded trap message is being received but not being processed message, check the snmp trace file on the Control Workstation. Table 11 shows the messages found in the trace file and the corresponding recovery action.

Table 11. Additional SNMP Trace File Messages	
Symptom	**Recovery**
Dependent node <ext_id> managed by the SNMP agent on host <CWS hostname> is not configured in the SDR - switchInfoNeeded trap ignored.	Either the wrong dependent node <ext_id> (slot number for the SP Switch router adapter in the GRF) or the wrong management_agent_hostname is placed in the DependentNode class. Correct the attributes and check using lssrc -ls spmgr.
SDR attribute <attr> in class <class> for dependent node <id> has a null value for SNMP Agent on host <hostname>.	Required attribute value is missing either in the DependentNode or in the DependentAdapter class. Add the missing attributes and check using lssrc -ls spmgr.
SDRGetAllObjects() DependentAdapter failed with return code 4.	Same as the previous recovery.
Dependent node <ext_id> managed by the SNMP agent on host <hostname> is configured with a bad community name-switchInfoNeeded trap ignored.	The SNMP community names in the DependentAdapter and the GRF do not match. Correct it in the DependentAdapter using endefnode, or on the GRF by editing the /etc/snmpd.conf file.

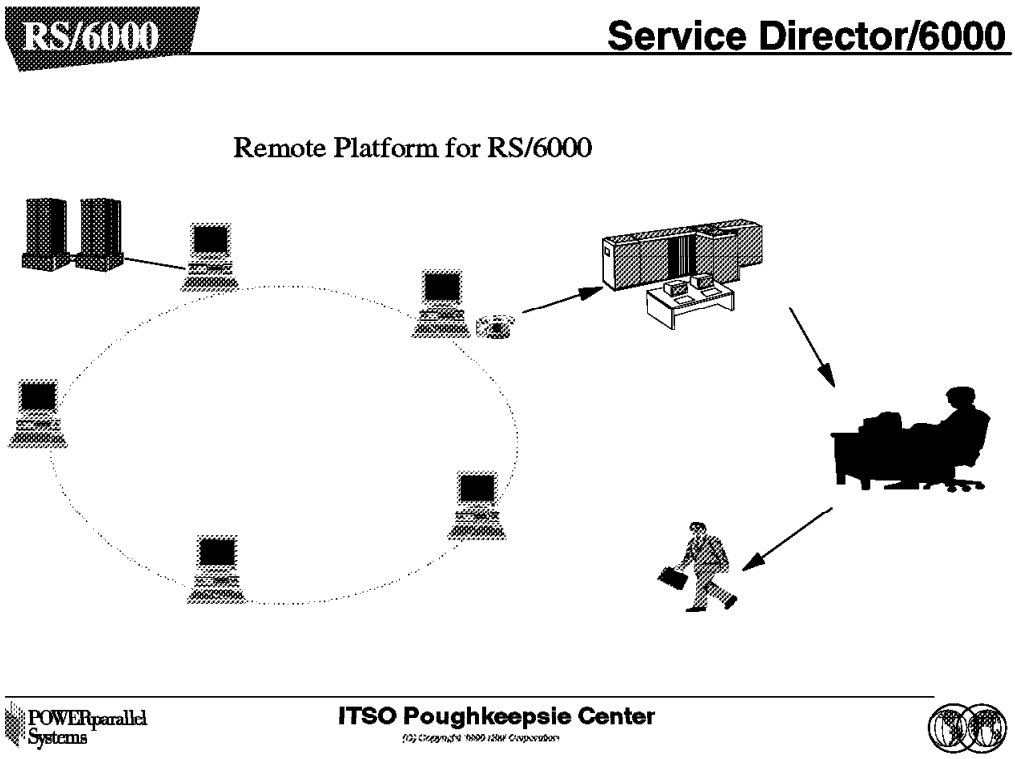
If none of these recovery methods solve the problem, refer to *IBM Parallel System Support Programs for AIX: Diagnosis and Messages Guide*, GC23-3899. Check the *Symptom and Recovery* table in the Diagnosing Switch Problems chapter of GC23-3899 for the proper action. Following is a list of suggestions for performing the recovery:

- If the recovery action is Verify Secondary Nodes, and the failing node is a dependent node, then enter SDRGetObjects switch_responds and check the adapter_config_status of the dependent node. If it is not css_ready, then continue with the following steps.
- Enter SDRGetObjects DependentNode to verify the attributes of the dependent node.
- Login to the GRF to verify the SP Switch router adapter attributes by issuing the following command (assume that the SP Switch router adapter is in slot 2 of the GRF):

```
# grrmb
GR 66> port 2
GR 02> maint 3
GR 00> [RX]
[RX] Configuration Parameters:
[RX]    Slot Number.........: 2
[RX]    Node Number.........: 7
[RX]    Node Name...........: 02
[RX]    SW Token............: 0001000602
[RX]    Arp Enabled.........: 2
[RX]    SW Node Number......: 6
[RX]    IP..................: 0x81282f47
[RX]    IP Mask.............: 0xffffffc0
[RX]    Alias IP............: 0x81283047
[RX]    Max Link Size.......: 1024
[RX]    Host Offset.........: 1
[RX]    Config State........: 1
[RX]    System Name.........: ceedgate
[RX]    Node State..........: 2
[RX]    Switch Link Chip....: 2
[RX]    Transmit Delay......: 31
```

- Verify that the SNMP community name in the /etc/snmpd.conf file on the GRF is the same as that in the CWS.
- When all the preceding items are verified, issue an Eunfence to add the dependent node to the RS/6000 SP.

Chapter 7. Service Director/6000



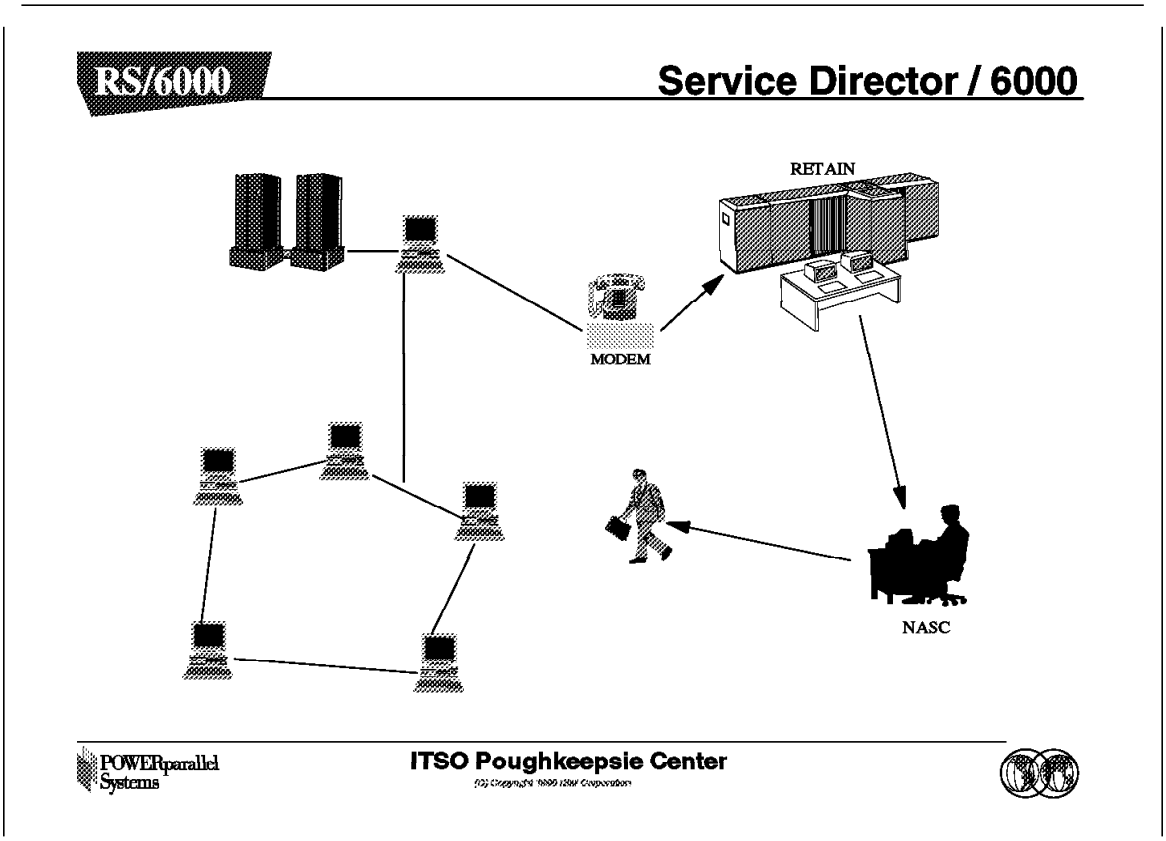
Service Director/6000 for RS/6000 is designed to increase system availability by reducing problem determination and repair time on the customer's RS/6000.

Service Director/6000 enhances IBM's ability to provide maintenance service by monitoring the system and automatically reporting problems. Service Director reports errors immediately for service and notifies the customer of the problem.

Service Director consists of an IBM software application specially designed to enhance IBM service and support for the customer hardware and automate problem reporting responsibilities.

© Copyright IBM Corp. 1998

7.1 Service Director/6000



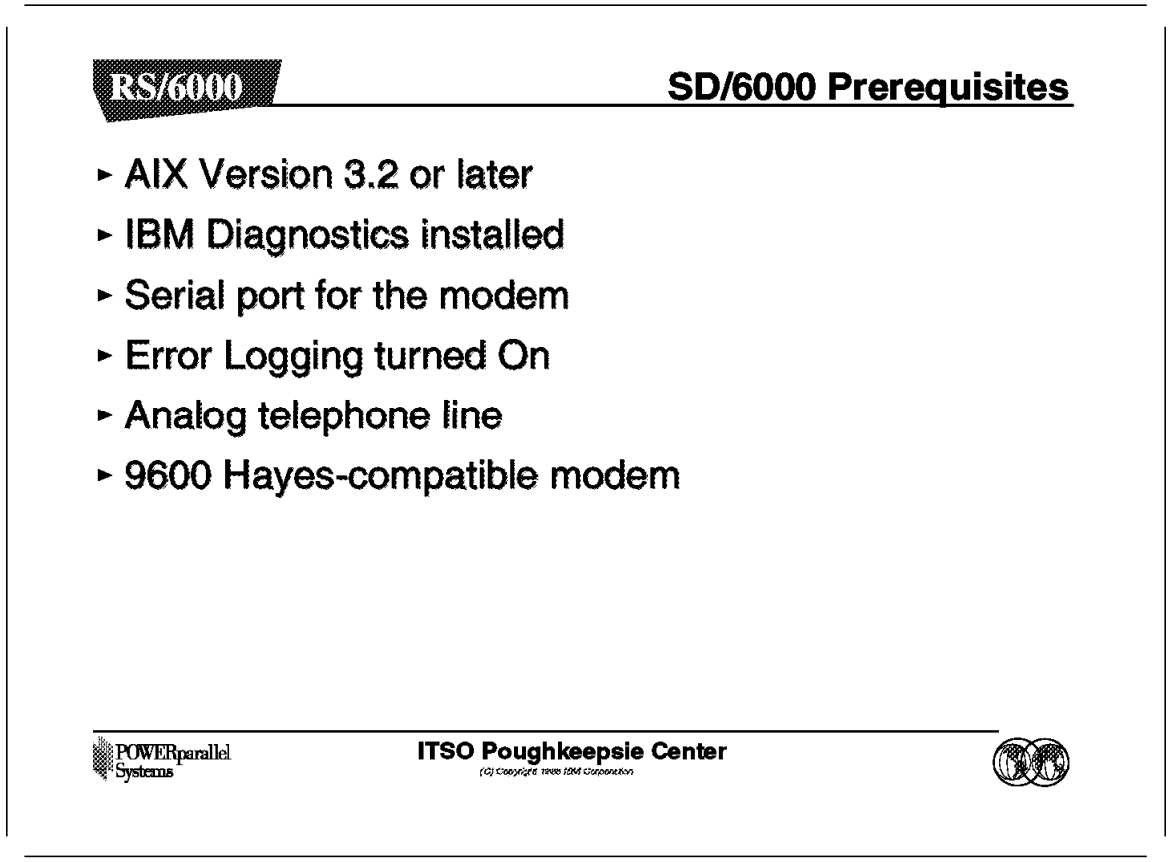
This new IBM service enhancement is part of your IBM warranty or IBM maintenance services agreement and is provided at no additional charge.

The RS/6000 SP requires Service Director/6000 Version 2.1 or greater in order to function properly. This version of Service Director/6000 supports Error Log Analysis of the SP supervisor cards and SP Switch events. Version 2.1 supports all RS/6000models, including the RS/6000 SP.

The application runs dynamically under AIX and can be installed concurrently. It allows several options on installation for customers to customize it to their specific needs. Additional product support application (PSA) can be written to provide additional support. Remote client's logs may be viewed from the Control Workstation.

Prior to warranty expiration, Service Director/6000 will notify the customer at 90, 60, and 30 days, allowing ample time to purchase an IBM maintenance agreement to ensure uninterrupted service.

7.2 SD/6000 Prerequisites



Service Director/6000 can only be installed on an RS/6000 running AIX Version 3.2 or later, with concurrent diagnostics installed.

Service Director/6000 should be applied at the customer's convenience. Installation, registration, and client code distribution may be done concurrently with daily operations. Once the registration key has been installed, the Service Director User Interface may be invoked concurrently.

Service Director requires the installer to have root authority on the local server machine where the installation is done. Note the following installation requirements:

Disk Space Requirements

Check the system the disk space for /usr and make sure that the file system is less than 98% full.

Service Director/6000 requires 4.5 MB for full install (Control Workstation) and 1.3 MB for client code (SP nodes).

TTY Requirements

A TTY device is needed to connect the modem. The characteristics of this TTY are: 9600 baud, no parity, 8 bits, with 1 stop bit.

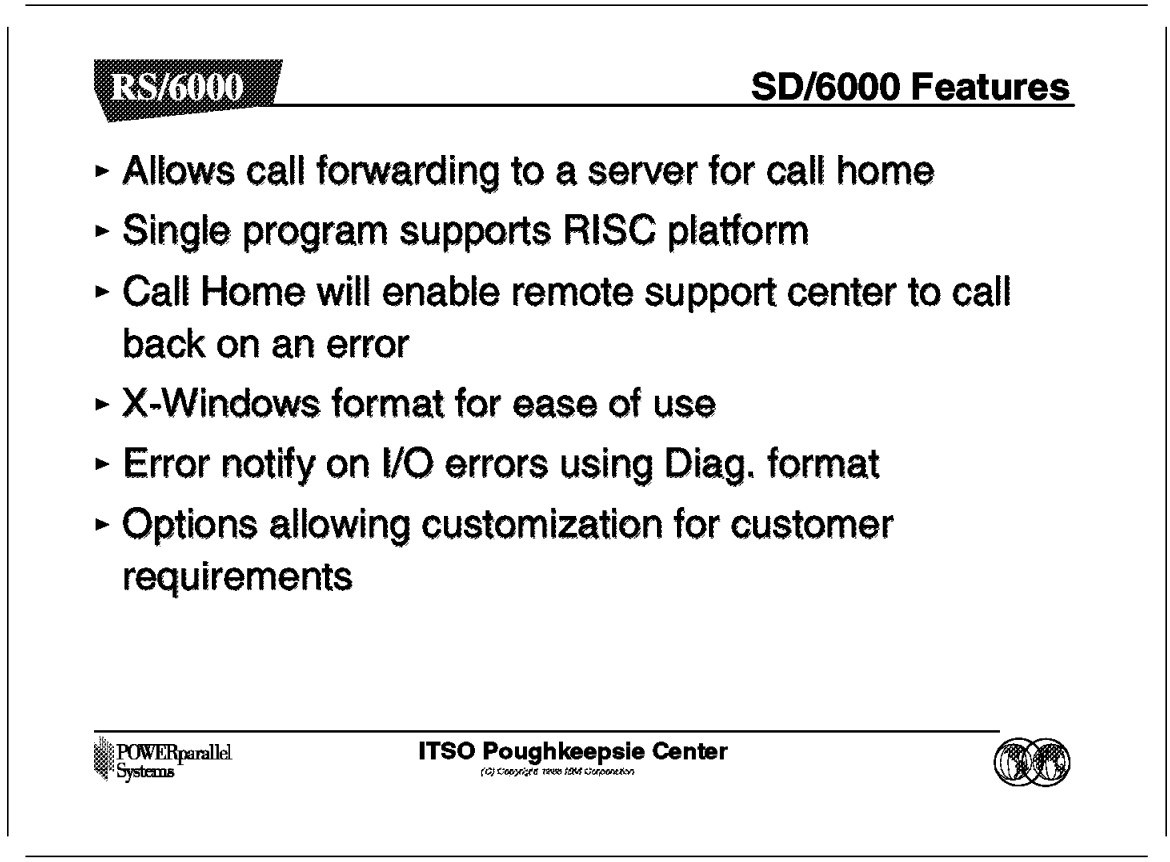
Modem Requirements

Service Director/6000 requires an asynchronous 9600 baud modem or greater with error correction in the United States. Refer to the local procedures in your country to see what the modem requirements are for Service Director/6000.

Diagnostics

IBM Diagnostics must be installed. This is the default configuration on the RS/6000 SP.

7.3 SD/6000 Features



Service Director/6000 is comprised of several software modules that work together to provide the functionality. The product support applications (PSA) monitor error conditions, analyze severity, determine appropriate error disposition, capture and pass any information required to resolve the problem.

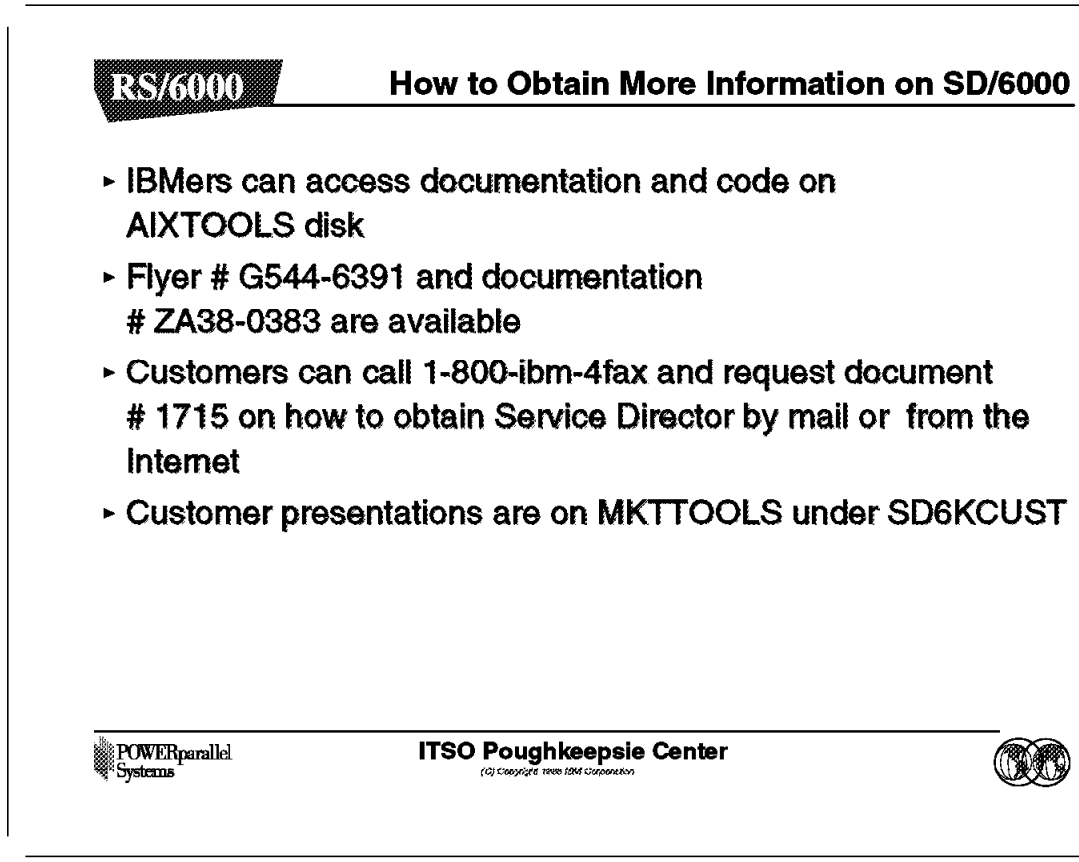
Analysis routine schedules the execution of the PSAs. This routine is configured to run automatically, or on a specific time schedule. Errors or events are logged, and depending on the customer-configured option, it may notify a person(s) within the customer's e-mail structure and automatically transmit hardware errors and associated problem information to an IBM Support Center for remote analysis and action. If needed, IBM Service personnel will be dispatched to the customer site with parts needed to correct the problem reported.

The Display routine of SD provides a structured view of problem management information, such as status of recent hardware events logged by Service Director, history of hardware events, and statistics of the problems managed by the Service Director. Client nodes may be viewed remotely from the Control Workstation.

Logs are maintained at each node and are not consolidated at the Control Workstation.

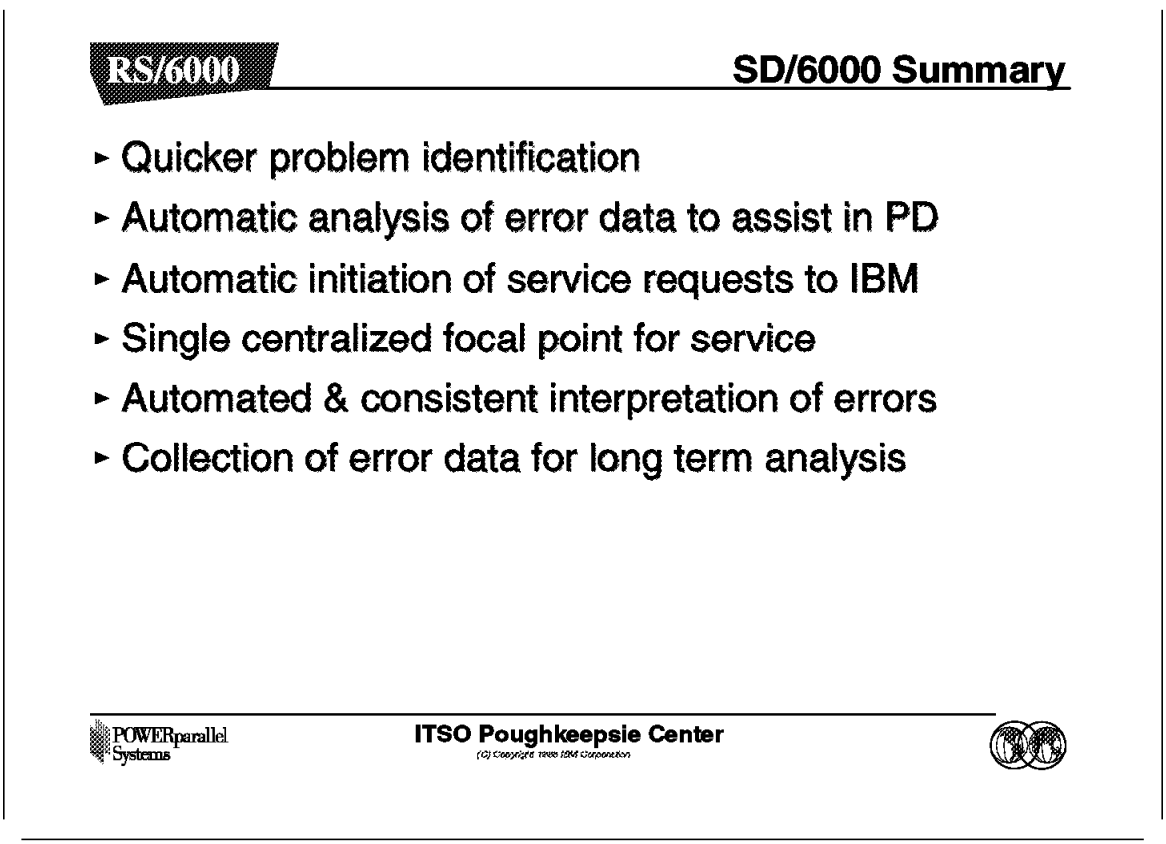
The Call Home function resides only on the local server (CWS) where the modem is attached. The Call Home daemon runs constantly on the server and forwards client calls via the Remote Procedure Call (RPC) process through the modem (TTY) to IBM.

7.4 How to Obtain More Information on SD/6000



Use these information sources if you want learn more, or want to use Service Director/6000.

7.5 SD/6000 Summary



Service Director/6000 can automatically report hardware-related problems to IBM for service via modem on the local server. The Service Director application can automatically do problem analysis on those problems before calling for service. Service Director supports all classic RS/6000 machines, including the RS/6000 SP.

Service Director aids IBM Service personnel in problem source identification, and it can be used to automatically place service calls to IBM for most of the hardware errors. System errors are dynamically monitored and analyzed; no customer intervention is required. Service Director/6000 further simplifies error analysis for some errors once the CE is on site by analysis of the Service Director event log. Therefore, the customer hardware error log can now be reduced, because error records are maintained within Service Director.

Appendix A. Special Notices

This publication is intended to help IBM Customers, Business Partners, IBM System Engineers, and other RS/6000 SP specialists who are involved in Parallel System Support Programs (PSSP) Version 2 Release 4 projects, including the education of RS/6000 SP professionals responsible for installing, configuring, and administering PSSP Version 2 Release 4. The information in this publication is not intended as the specification of any programming interfaces that are provided by Parallel System Support Programs. See the PUBLICATIONS section of the IBM Programming Announcement for PSSP Version 2 Release 4 for more information about what publications are considered to be product documentation.

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

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

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to the IBM Director of Licensing, IBM Corporation, 500 Columbus Avenue, Thornwood, NY 10594 USA.

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 information about non-IBM (″vendor″) products in this manual has been supplied by the vendor and IBM assumes no responsibility for its accuracy or completeness. 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.

© Copyright IBM Corp. 1998

Any performance data contained in this document was determined in a controlled environment, and therefore, the results that may be obtained in other operating environments may vary significantly. Users of this document should verify the applicable data for their specific environment.

Reference to PTF numbers that have not been released through the normal distribution process does not imply general availability. The purpose of including these reference numbers is to alert IBM customers to specific information relative to the implementation of the PTF when it becomes available to each customer according to the normal IBM PTF distribution process.

You can reproduce a page in this document as a transparency, if that page has the copyright notice on it. The copyright notice must appear on each page being reproduced.

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

AIX	AS/400
BookManager	Current
IBM	POWERparallel
PROFS	RS/6000
Service Director	SP
System/390	400

The following terms are trademarks of other companies:

C-bus is a trademark of Corollary, Inc.

Java and HotJava are trademarks of Sun Microsystems, Incorporated.

Microsoft, Windows, Windows NT, and the Windows 95 logo are trademarks or registered trademarks of Microsoft Corporation.

PC Direct is a trademark of Ziff Communications Company and is used by IBM Corporation under license.

Pentium, MMX, ProShare, LANDesk, and ActionMedia are trademarks or registered trademarks of Intel Corporation in the U.S. and other countries.

UNIX is a registered trademark in the United States and other countries licensed exclusively through X/Open Company Limited.

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

Appendix B. Related Publications

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

B.1 International Technical Support Organization Publications

For information on ordering these ITSO publications see “How to Get ITSO Redbooks” on page 259.

- *Technical Presentation for PSSP 2.3*, SG24-2080
- *PSSP 2.2 Technical Presentation*, SG24-4868
- *RS/6000 Scalable POWERparallel System: PSSP Version 2*, SG24-4542
- *GPFS: A Parallel File System*, SG24-5165

B.2 Redbooks on CD-ROMs

Redbooks are also available on CD-ROMs. **Order a subscription** and receive updates 2-4 times a year at significant savings.

CD-ROM Title	Subscription Number	Collection Kit Number
System/390 Redbooks Collection	SBOF-7201	SK2T-2177
Networking and Systems Management Redbooks Collection	SBOF-7370	SK2T-6022
Transaction Processing and Data Management Redbook	SBOF-7240	SK2T-8038
Lotus Redbooks Collection	SBOF-6899	SK2T-8039
Tivoli Redbooks Collection	SBOF-6898	SK2T-8044
AS/400 Redbooks Collection	SBOF-7270	SK2T-2849
RS/6000 Redbooks Collection (HTML, BkMgr)	SBOF-7230	SK2T-8040
RS/6000 Redbooks Collection (PostScript)	SBOF-7205	SK2T-8041
RS/6000 Redbooks Collection (PDF Format)	SBOF-8700	SK2T-8043
Application Development Redbooks Collection	SBOF-7290	SK2T-8037

B.3 Other Publications

These publications are also relevant as further information sources:

- *PSSP Installation and Migration Guide*, GC23-3898
- *PSSP Administration Guide*, GC23-3897
- *PSSP Diagnosis and Messages*, GC23-3899
- *PSSP Command and Technical Reference*, GC23-3900
- *Planning Vol 1, Hardware and Physical Environment*, GA22-7280
- *Planning Vol 2, Control Workstation and Software Environment*, GA22-7281
- *Event Management Programming Guide and Reference*, SC23-3996
- *Group Services Programming Guide and Reference*, SC28-1675
- *Managing Shared Disks*, SA22-7279
- *IBM General Parallel File System for AIX: Installation and Administration Guide*, SA22-7278
- *Service Director for RS/6000: Information Guide*, ZA38-0383

© Copyright IBM Corp. 1998

How to Get ITSO Redbooks

This section explains how both customers and IBM employees can find out about ITSO redbooks, CD-ROMs, workshops, and residencies. A form for ordering books and CD-ROMs is also provided.

This information was current at the time of publication, but is continually subject to change. The latest information may be found at http://www.redbooks.ibm.com/.

How IBM Employees Can Get ITSO Redbooks

Employees may request ITSO deliverables (redbooks, BookManager BOOKs, and CD-ROMs) and information about redbooks, workshops, and residencies in the following ways:

- **Redbooks Web Site on the World Wide Web**

  http://w3.itso.ibm.com/

- **PUBORDER —** to order hardcopies in the United States
- **Tools Disks**

  To get LIST3820s of redbooks, type one of the following commands:

  ```
  TOOLCAT REDPRINT
  TOOLS SENDTO EHONE4 TOOLS2 REDPRINT GET SG24xxxx PACKAGE
  TOOLS SENDTO CANVM2 TOOLS REDPRINT GET SG24xxxx PACKAGE (Canadian users only)
  ```

  To get BookManager BOOKs of redbooks, type the following command:

  ```
  TOOLCAT REDBOOKS
  ```

  To get lists of redbooks, type the following command:

  ```
  TOOLS SENDTO USDIST MKTTOOLS MKTTOOLS GET ITSOCAT TXT
  ```

  To register for information on workshops, residencies, and redbooks, type the following command:

  ```
  TOOLS SENDTO WTSCPOK TOOLS ZDISK GET ITSOREGI 1998
  ```

- **REDBOOKS Category on INEWS**
- **Online** — send orders to: USIB6FPL at IBMMAIL or DKIBMBSH at IBMMAIL

Redpieces

For information so current it is still in the process of being written, look at "Redpieces" on the Redbooks Web Site (http://www.redbooks.ibm.com/redpieces.html). Redpieces are redbooks in progress; not all redbooks become redpieces, and sometimes just a few chapters will be published this way. The intent is to get the information out much quicker than the formal publishing process allows.

© Copyright IBM Corp. 1998

How Customers Can Get ITSO Redbooks

Customers may request ITSO deliverables (redbooks, BookManager BOOKs, and CD-ROMs) and information about redbooks, workshops, and residencies in the following ways:

- **Online Orders** — send orders to:

	IBMMAIL	**Internet**
In United States:	usib6fpl at ibmmail	usib6fpl@ibmmail.com
In Canada:	caibmbkz at ibmmail	lmannix@vnet.ibm.com
Outside North America:	dkibmbsh at ibmmail	bookshop@dk.ibm.com

- **Telephone Orders**

United States (toll free)	1-800-879-2755
Canada (toll free)	1-800-IBM-4YOU
Outside North America	(long distance charges apply)
(+45) 4810-1320 - Danish	(+45) 4810-1020 - German
(+45) 4810-1420 - Dutch	(+45) 4810-1620 - Italian
(+45) 4810-1540 - English	(+45) 4810-1270 - Norwegian
(+45) 4810-1670 - Finnish	(+45) 4810-1120 - Spanish
(+45) 4810-1220 - French	(+45) 4810-1170 - Swedish

- **Mail Orders** — send orders to:

IBM Publications Publications Customer Support P.O. Box 29570 Raleigh, NC 27626-0570 USA	IBM Publications 144-4th Avenue, S.W. Calgary, Alberta T2P 3N5 Canada	IBM Direct Services Sortemosevej 21 DK-3450 Allerød Denmark

- **Fax** — send orders to:

United States (toll free)	1-800-445-9269
Canada	1-403-267-4455
Outside North America	(+45) 48 14 2207 (long distance charge)

- **1-800-IBM-4FAX (United States)** or **(+1)001-408-256-5422 (Outside USA)** — ask for:

Index # 4421 Abstracts of new redbooks
Index # 4422 IBM redbooks
Index # 4420 Redbooks for last six months

- **On the World Wide Web**

Redbooks Web Site	http://www.redbooks.ibm.com/
IBM Direct Publications Catalog	http://www.elink.ibmlink.ibm.com/pbl/pbl

Redpieces

For information so current it is still in the process of being written, look at "Redpieces" on the Redbooks Web Site (`http://www.redbooks.ibm.com/redpieces.html`). Redpieces are redbooks in progress; not all redbooks become redpieces, and sometimes just a few chapters will be published this way. The intent is to get the information out much quicker than the formal publishing process allows.

IBM Redbook Order Form

Please send me the following:

Title	Order Number	Quantity

First name Last name

Company

Address

City Postal code Country

Telephone number Telefax number VAT number

- Invoice to customer number
- Credit card number

Credit card expiration date Card issued to Signature

We accept American Express, Diners, Eurocard, Master Card, and Visa. Payment by credit card not available in all countries. Signature mandatory for credit card payment.

List of Abbreviations

ACL	access control list
AIX	Advanced Interactive Executive
AMG	adapter membership group
ANS	abstract notation syntax
APA	all points addressable
API	application programming interface
BIS	boot-install server
BSD	Berkeley Software Distribution
BUMP	bring-up microprocessor
CP	crown prince
CPU	central processing unit
CSS	communication subsystem
CW	control workstation
DB	database
EM	event management
EMAPI	Event Management Application Programming Interface
EMCDB	Event Management Configuration Database
EMD	Event Manager Daemon
EPROM	erasable programmable read-only memory
FIFO	first in - first out
GB	gigabytes
GL	group leader
GPFS	General Parallel File System for AIX
GS	group services
GSAPI	Group Services Application Programming Interface
GVG	global volume group
HACMP	High Availability Cluster Multiprocessing
hb	heart beat
HPS	High Performance Switch
hrd	host respond daemon
HSD	hashed shared disk
IBM	International Business Machines Corporation
IP	Internet Protocol
ISB	intermediate switch board
ISC	intermediate switch chip
ITSO	International Technical Support Organization
JFS	journal file system
LAN	local area network
LCD	liquid crystal display
LED	light emitter diode
LRU	least recently used
LSC	link switch chip
LVM	logical volume manager
MB	megabytes
MIB	management information base
MPI	message passing interface
MPL	message passing library
MPP	massively parallel processors
NIM	network installation manager
NSB	node switch board
NSC	node switch chip
OID	object ID
ODM	object data manager
PE	parallel environment
PID	process ID
PROFS	Professional Office System
PSSP	Parallel System Support Program
PTC	prepare to commit
PTPE	Performance Toolbox Parallel Extensions
PTX/6000	Performance Toolbox/6000
RAM	random access memory
RCP	remote copy protocol
RM	resource monitor
RMAPI	Resource Monitor Application Programming Interface
RPQ	request for product quotation
RSI	remote statistics interface
RVSD	recoverable virtual shared disk
SBS	structured byte string
SDR	System Data Repository

© Copyright IBM Corp. 1998

SMP	symmetric multiprocessors
SNMP	system network management protocol
SPDM	SP Data Manager
SPMI	System Performance Measurement Interface
SRC	system resource controller
SSI	single system image
TS	topology services
TCP/IP	Transmission Control Protocol/Internet Protocol
UDP	User Datagram Protocol
VSD	virtual shared disk
VSM	visual system management

Index

© Copyright IBM Corp. 1998

T

V

W

X

ITSO Redbook Evaluation

PSSP 2.4 Technical Presentation
SG24-5173-00

Your feedback is very important to help us maintain the quality of ITSO redbooks. **Please complete this questionnaire and return it using one of the following methods:**

- Use the online evaluation form found at http://www.redbooks.ibm.com
- Fax this form to: USA International Access Code + 1 914 432 8264
- Send your comments in an Internet note to redbook@us.ibm.com

Which of the following best describes you?
__Customer __Business Partner __Solution Developer __IBM employee
__None of the above

Please rate your overall satisfaction with this book using the scale:
(1 = very good, 2 = good, 3 = average, 4 = poor, 5 = very poor)

Overall Satisfaction _____

Please answer the following questions:

Was this redbook published in time for your needs? Yes____ No____

If no, please explain:

_____

_____

_____

_____

What other redbooks would you like to see published?

_____

_____

_____

Comments/Suggestions: (THANK YOU FOR YOUR FEEDBACK!)

_____

_____

_____

_____

_____

© Copyright IBM Corp. 1998

SG24-5173-00
Printed in the U.S.A.