IBM @server pSeries™ 690 Availability Best Practices

Document Number: p690 Availability Best Practices
Last Revised: 5/1/2002
Version: GA2

1.0 About this White Paper

1.1 Purpose

The purpose of this paper is to describe in a level of detail useful to a system administrator or operator how to configure the pSeries 690 system to achieve higher levels of availability utilizing a *Best Practices* approach.

There are many factors which influence the overall availability of a system. Some factors take into account the base design of the system including which components act as single points of failure and which components are replicated through N+1 design or some other form of redundancy. These areas of the pSeries 690 are controlled by development and are minimized through various design techniques (refer to the white paper entitled "*pSeries 690 RAS White Paper*").

Other factors are directly influenced by how the administrator of the system chooses to configure their various operating environments. This is the primary focus of this white paper. Other factors outside the scope of this white paper, but for which consulting services are available from IBM, deal with how the system is administered and the various availability management policies and practices established by the IT department.

1.2 Approach

The first part of this white paper will address the pSeries 690 system running in symmetric multiprocessing (SMP) mode and will form the basis on which the rest of the paper references.

The second part of the white paper will address the availability factors concerned with the Logical PARtition (LPAR) mode of operation. It's recommended that the user interested in configuring an LPAR system for high availability review the SMP sections prior to the LPAR section since most of the LPAR material references back to the SMP section.

The next section will address running HACMP in SMP and LPAR partitions which is followed by the conclusions regarding configuring the for availability.

This paper is best utilized by referencing the following documents or guides to actually perform the suggested actions proposed in this paper:

- pSeries 690 Installation Guide - SA38-0587-00
- Hardware Management Console for pSeries Operations Guide
- IBM @server pSeries 690 User's Guide - SA38-0588-00
- PCI Adapter Placement Reference - SA38-0538
- HACMP for AIX Installation Guide
- Electronic Service Agent for pSeries User's Guide

1.3 Modifications From Last Release

Last Revised Date	Release	Modifications
10/10/2001	Initial p690 release	
04/26/2002	p690 GA2 release	Changes to default options to enable auto restart after failure.

2.0 Configuring the p690 SMP for High Availability

There are many options left to the customer which ultimately can affect the overall system availability. This section of the white paper will address those factors that the customer has control over associated with configuring the pSeries 690 in an SMP mode of operation.

This section will begin by addressing base boot configuration options through use of Service Processor (SP) menus or through alternate methods such as AIX® command line instructions, AIX Diagnostic Service AIDs, Service Management System menus, or Web System Manager (Web SM) interfaces. Once configured, these options remain in effect until specifically changed by the user.

Following this section, Boot and User disk configuration options will be reviewed. Disk configuration and I/O adapter configuration reflect two of the most significant areas that the customer has direct control over to affect overall system availability. Finally, service enablement and remote support topics will be addressed which will help to automate the error report, service call, and dispatch of the service representative with the correct Field Replaceable Units (FRUs) to fix the system quickly and efficiently in the event of a failure.

2.1 Boot Options Utilizing SP Menus and Alternatives

Listed below are a set of Boot options found in the SP menus along with the default and preferred settings to enable for high availability. The setting of these options to the recommended values will have a direct affect on the time to recover from a failure by allowing the system to automatically attempt recovery or will provide a simplified method to recover from the failure.

Each of the individual settings will be explained in additional sections following the table. The last column in the table represents whether there is an alternative method from System Management Services (SMS) menus, the SMIT/SMITTY panels in the AIX operating system or through the AIX Diagnostic Service Aids interfaces to change the setting. The user can choose whichever method they are most familiar with to effect the proposed changes. Each of the options are explained in the sections following the table.

SP Menu	Option	Default Value	Recommended Setting	Alternate Invocation Method
Service Processor Setup Menu				
OS Surveillance Setup Menu	OS Surveillance	Off	On	**Service AID** Configure Service Processor - Configure Surveillance Policy
	Surveillance Time Interval	2 minutes	Customer Determined	
	Surveillance Delay	2 minutes	Customer Determined	
Serial Port Snoop Setup	System Reset String	Unassigned	Choose reset string	
	Snoop Serial Port	Unassigned	1	
Scan Log Dump Policy Menu	Scan Dump Log Policy	1 (As needed)	1 (As needed)	
System Power Control Menu				

SP Menu	Option	Default Value	Recommended Setting	Alternate Invocation Method
Enable/Disable Unattended Start Mode	Enable/Disable Unattended Start Mode	No - GA1 Yes - In systems shipped with GA2 (April 26th,2002) & beyond	Yes	**SMS Menu**- Unattended Start Menu **Service AID** - Reboot/Restart Policy Setup
Reboot/Restart Policy Setup Menu				
Reboot/Restart Policy Setup Menu	Number of reboot attempts	1	1	**Service AID** - Configure Reboot Policy **SMIT** - Automatically Reboot After Crash
	Use OS-Defined restart policy	Yes - GA1 No -In systems shipped with GA2 (April 26th, 2002) & beyond	No	
	Enable supplemental restart policy	No - GA1 Yes - In systems shipped with GA2 (April 26th, 2002) & beyond	Yes	
System Information Menu				
Processor Config/Deconfig Menu	Enable CPU Repeat Gard	Enable	Enable	
Memory Config/Deconfig Menu	Enable Memory Repeat Gard	Enable	Enable	

Refer to the "IBM @server pSeries Installation Guide" for specific instructions on configuring the preferred settings through use of the SP menus.

2.1.1 Boot Time Options

Once these options are set utilizing the Service Processor Setup menus, they are maintained in NVRAM and utilized on every system boot until they are specifically altered by the user.

2.1.2 OS Surveillance Setup

Surveillance is a function in which the service processor monitors the system, and the system monitors the service processor. This monitoring is accomplished by periodic samplings called *heartbeat*s. Surveillance is available during two phases:

1. System firmware bringup (automatic)
2. Operating system runtime (optional)

System Firmware Surveillance

System firmware surveillance is automatically enabled during system power-on. It cannot be disabled by the user, and the surveillance interval and surveillance delay cannot be changed by the user. If the service processor detects no heartbeats during system IPL (for a set period of time), it cycles the system power to attempt a reboot. The maximum number of retries is set from the service processor menus. If the fail condition persists, the service processor leaves the machine powered on, logs an error, and displays menus to the user. The service processor reports the failure to the IBM Hardware Management Console for pSeries (HMC) and displays the operating-system surveillance failure code on the operator panel. The Service Focal Point application running on the HMC will log the error in the Service Action Event (SAE) Log and will call for Service.

Operating System Surveillance

Operating system surveillance provides the service processor with a means to detect hang conditions, as well as hardware or software failures, while the operating system is running. It also provides the operating system with a means to detect a service processor failure caused by the lack of a return heartbeat.

Operating system surveillance is not enabled by default, allowing you to run operating systems that do not support this service processor option. You can also use service processor menus and AIX service aids to enable or disable operating system surveillance.

For operating system surveillance to work correctly, you must set these parameters:

- Surveillance enable/disable
- Surveillance interval
  - The maximum time the service processor should wait for a heartbeat from the operating system before time out.
- Surveillance delay
  - The length of time to wait from the time the operating system is started to when the first heartbeat is expected. Surveillance does not take effect until the next time the operating system is started after the parameters have been set.

If desired, you can initiate surveillance mode immediately from service aids. In addition to the three options above, a fourth option allows you to select immediate surveillance, and rebooting of the system is not necessarily required. If operating system surveillance is enabled (and system firmware has passed control to the operating system), and the service processor does not detect any heartbeats from the operating system, the service processor assumes the system is hung and takes action according to the reboot/restart policy settings. If surveillance is selected from the service processor menus which are only available at bootup, then surveillance is enabled by default as soon as the system boots. From service aids, the selection is optional.

2.1.2.1 Service Processor Menu

OS Surveillance Setup Menu
1. Surveillance:
Currently Enabled
2. Surveillance Time Interval:
2 minutes
3. Surveillance Delay:
2 minutes
98. Return to Previous Menu
0>

– **Surveillance**

Can be set to Enabled or Disabled.

– **Surveillance Time Interval**

Can be set to any number from 2 through 255 (value is in minutes).

– **Surveillance Delay**

Can be set to any number from 0 through 255 (value is in minutes).

2.1.2.2 Service Aid - Configure Service Processor - Configure Surveillance Policy

Note: This service aid runs on CHRP system units only.

This service aid monitors the system for hang conditions; that is, hardware or software failures that cause operating system inactivity. When enabled, and surveillance detects operating system inactivity, a call is placed to report the failure. Use this service aid to display and change the following settings for the Surveillance Policy:

Note: Because of system capability, some of the following settings might not be displayed by this service aid:

- Surveillance (on/off)
- Surveillance Time Interval
  - This is the maximum time between heartbeats from the operating system.
- Surveillance Time Delay
  - This is the time to delay between when the operating system is in control and when to begin operating system surveillance.
- Changes are to Take Effect Immediately
  - Set this to Yes if the changes made to the settings in this menu are to take place immediately. Otherwise the changes take effect beginning with the next system boot.

You can access this service aid directly from the AIX command line, by typing:
/usr/lpp/diagnostics/bin/uspchrp -s

2.1.3 Serial Port Snoop Setup

This menu can be used to set up serial port snooping, in which the user can configure serial port 1 as a “catch-all”...reset device.

2.1.3.1 Service Processor Menu - Serial Port Snoop Setup

From the service processor main menu,
select option 1, service processor setup menu,
then select option 8 (Serial Port Snoop Setup menu).

SERIAL PORT SNOOP SETUP MENU
1.System reset string:
Currently Unassigned
2.Snoop Serial Port:
Currently Unassigned
98.Return to Previous Menu
1>

Use the system reset string option to enter the system reset string, which resets the machine when it is detected on the main console on Serial Port 1.

Use the **Snoop Serial Port** option to select the serial port to snoop.
Note: Only serial port 1 is supported.

After serial port snooping is correctly configured, at any point after the system is booted to AIX, whenever the reset string is typed on the main console, the system uses the service processor reboot policy to restart. Pressing Enter after the reset string is not required, so make sure that the string is not common or trivial. A mixed-case string is recommended.

2.1.4 Scan Dump Log Policy

A scan dump is the collection of chip data that the service processor gathers after a system malfunction, such as a checkstop or hang. The scan dump data may contain chip scan rings, chip trace arrays, and SCOM contents.
The scan dump data are stored in the system control store. The size of the scan dump area is approximately 4MB.

2.1.4.1 Service Processor Menu - Scan Dump Log Policy

The scan dump log policy can be set to:
1 = As needed
This is the default value. In this case, the processor run-time diagnostics record the dump data based on the error type.
2 = Never
3 = Always
4 = Immediately
This option can only be used when the system is in the standby state with power on. It is used to dump the system data after a checkstop or machine check occurs when the system firmware is running, or when the operating system is booting or running.

2.1.5 Unattended Start Mode

Use this option to instruct the service processor to restore the power state of the server after a temporary power failure. Unattended start mode can also be set through the System Management Services (SMS) menus. It is intended to be used on servers that require automatic power-on after a power failure.

When ac power is restored, the system returns to the power state at the time ac loss occurred. For example, if the system was powered-on when ac loss occurred, it reboots/restarts when power is restored. If the system was powered-off when ac loss occurred, it remains off when power is restored.

2.1.5.1 SP Menu

SYSTEM POWER CONTROL MENU
1. Enable/Disable Unattended Start Mode:
Currently Enabled

2.1.5.2 SMS Menu

Unattended Start Mode <ON>: This selection is used to enable or disable unattended start mode. Use this option to instruct the service processor to restore the power-state of the server after a temporary power failure, which is necessary on servers that require automatic power-on after a power failure. The default setting for GA1 was off. The default setting for systems shipped with GA2 (April 26th, 2002) and beyond is on.

2.1.5.3 Service Aid - Reboot/Restart Policy Setup

Enable Unattended Start Mode (1=Yes, 0=No)

When enabled, "Unattended Start Mode"...allows the system to recover from the loss of ac power. If the system was powered-on when the ac loss occurred, the system reboots when power is restored. If the system was powered-off when the ac loss occurred, the system remains off when power is restored. You can access this service aid directly from the AIX command line, by typing: /usr/lpp/diagnostics/bin/uspchrp -b

2.1.6 Reboot/Restart on Failure

The operating system's automatic restart policy (see operating system documentation) indicates the operating system response to a system crash. The service processor can be instructed to refer to that policy by the "Use OS-Defined Restart Policy" setup menu. If the operating system has no automatic restart policy, or if it is disabled, then the service processor-restart policy can be controlled from the service processor menus. Use the "Enable Supplemental Restart Policy" selection.

The purpose of this option is to allow the system to attempt to restart after experiencing a fatal error. This option is defined by a combination of 2 parameters described in the table below. The objective is to allow the system to attempt to restart after failure. Choose the option that you feel most comfortable with to cause that affect.

Use OS-Defined restart policy - The default setting for GA1 was Yes. This causes the service processor to refer to the OS Automatic Restart Policy setting and take action (the same action the operating system would take if it could have responded to the problem causing the restart).

The default for systems shipped with GA2 (April 26th, 2002) & beyond is No. When this setting is No, or if the operating system did not set a policy, the service processor refers to enable supplemental restart policy for its action.

Enable supplemental restart policy - The default setting for GA1 was No. The default setting is Yes for systems shipped with GA2 (April 26th, 2002) & beyond. If set to Yes, the service processor restarts the server when the operating system loses control and either:

The **Use OS-Defined restart policy** is set to No.
OR
The **Use OS-Defined restart policy** is set to Yes and the operating system has no automatic restart policy.

The following table describes the relationship among the operating system and service processor restart controls in SMP mode

Parameter	**Use**	**Recommended Setting**
OS automatic restart policy	Controls reboot on all crashes when "Use OS defined restart policy" is yes	true
"Use OS defined restart policy"	Selects whether OS automatic restart or supplemental restart policy is used	no
"Enable Supplemental Restart Policy"	Controls reboot on all crashes when "Use OS automatic Restart policy" is no	yes
"Enable Unattended Start Mode"	Controls reboot when AC power is restored after a power failure. (true means restart will occur.)	yes

With the recommended settings, the system will automatically attempt restart, regardless of the setting of the OS automatic restart parameter.

The following table describes the interaction of the parameters in LPAR mode:

Parameter	**Use**	**Recommended Setting**
OS automatic restart policy	Controls reboot on partition crashes	true for each LPAR partition
"Use OS defined restart policy"	Undefined	Set to no, if field is available on the service authority partition.
"Enable Supplemental Restart Policy"	Controls reboot on platform crashes (requires "Use O/S defined Reboot policy" set to "no" on firmware releases prior to April 26th, 2002 L3)	true
"Enable Unattended Start Mode"	Controls reboot on when AC power is restored after a power failure. (true means reboot will occur.)	true

With the recommended changes, the system will always automatically attempt reboot if the OS automatic restart parameter is set to true in each LPAR partition.

2.1.6.1 SP Menu - Reboot/Restart Policy Setup Menu

Reboot/Restart Policy Setup Menu
1. Number of reboot attempts:
Currently 1
2. Use OS-Defined restart policy?
Currently No
3. Enable supplemental restart policy?
Currently Yes
4. Call-Out before restart:
Currently Disabled
98. Return to Previous Menu
0>

Reboot is the process of bringing up the system hardware; for example, from a system reset or power on. Restart is activating the operating system after the system hardware is reinitialized. Restart must follow a successful reboot.

– **Number of reboot attempts** - If the server fails to successfully complete the boot process, it attempts to reboot the number of times specified. Entry values equal to or greater than 0 are valid. Only successive failed reboot/restart attempts are counted.

– **Use OS-Defined restart policy** - In SMP mode allows the service processor to react or not react in the same way that the operating system does to major system faults by reading the setting of the operating system parameter **Automatically Restart/Reboot After a System Crash**. This parameter may or may not be defined, depending on the operating system or its version/level.. See your operating system documentation for details on setting up operating system automatic restarts.

For proper operation in LPAR mode, this parameter should always be set to No. With the parameter set to No in LPAR mode, the OS defined restart policy is used for partition crashes and the Supplemental Restart policy is used for system crashes.

The default value for Use OS-Defined restart policy in GA1 was Yes. The default for systems shipped with GA2 (April 26rh, 2002) & beyond is No.

– **Enable supplemental restart policy** - The default setting for GA1 was No. The default for systems shipped with GA2 (April 26th, 2002) & beyond is Yes.
In SMP mode, if set to Yes, the service processor restarts the system when the system loses control as detected by service processor surveillance, and either:
The **Use OS-Defined restart policy** is set to No
OR
The **Use OS-Defined restart policy** is set to Yes, and the operating system has no automatic restart policy.

In LPAR mode, if **Use OS-Defined restart policy** is No, then **Enable supplemental restart policy** controls whether a the service processor restartsthe system when the system loses control as detected by service processor.

– **Call-Out before restart (Enabled/Disabled)** - This should be left in the default disabled state for this system. Call out is handled through the Service Focal Point application running on the IBM hardware management console for pseries. Refer to the “Remote Support” and “Service Enablement” sections later in this white paper for more detail.

2.1.6.2 Service Aid - Configure Reboot Policy

This service aid controls how the system tries to recover from a system crash.
Use this service aid to display and change the following settings for the Reboot Policy.
Note: Because of system capability, some of the following settings might not be displayed by this service aid.

- Maximum Number of Reboot Attempts
  - Enter a number that is 0 or greater.

Note: A value of 0 indicates ’do not attempt to reboot’ to a crashed system. This number is the maximum number of consecutive attempts to reboot the system. The term *reboo*t, in the context of this service aid, is used to describe bringing system hardware back up from scratch; for example, from a system reset or Power-on. When the reboot process completes successfully, the reboot attempts count is reset to 0, and a restart begins. The term *restar*t, in the context of this service aid, is used to describe the operating system activation process. Restart always follows a successful reboot. When a restart fails, and a restart policy is enabled, the system attempts to reboot for the maximum number of attempts.

- Use the O/S Defined Restart Policy (1=Yes, 0=No)

When ’Use the O/S Defined Restart Policy’ is set to Yes, the system attempts to reboot from a crash if the operating system has an enabled Defined Restart or Reboot Policy. When ’Use the O/S Defined Restart Policy’ is set to No, or the operating system restart policy is undefined, then the restart policy is determined by the ’Supplemental Restart Policy’.

- Enable Supplemental Restart Policy (1=Yes, 0=No)

The ’Supplemental Restart Policy’, if enabled, is used when the O/S Defined Restart Policy is undefined, or is set to False. When surveillance detects operating system inactivity during restart, an enabled ’Supplemental Restart Policy’ causes a system reset and the reboot process begins.

- Call-Out Before Restart (on/off)

This option should remain disabled as stated above in the SP menu option.

2.1.7 Processor/Memory Configuration/Deconfiguration Menu

All failures that crash the system with a machine check or check stop, even if intermittent, are reported as a diagnostic call out for service repair. To prevent the recurrence of intermittent problems and improve the availability of the system until a scheduled maintenance window, processors and memory books with a failure history are marked “bad” to prevent their being configured on subsequent boots. A processor or memory book is marked “bad” under the following circumstances:

- A processor or memory book fails built-in self test (BIST) or power-on self test (POST) testing during boot (as determined by the service processor).
- A processor or memory book causes a machine check or check stop during runtime, and the failure can be isolated specifically to that processor or memory book (as determined by the processor runtime diagnostics in the service processor).
- A processor or memory book reaches a threshold of recovered failures that results in a predictive call out (as determined by the processor runtime diagnostics in the service processor).

During boot time, the service processor does not configure processors or memory books that are marked "bad". If a processor or memory book is deconfigured, the processor or memory book remains offline for subsequent reboots until it is replaced or repeat gard is disabled.

The Repeat Gard function also provides the user with the option of manually deconfiguring a processor or memory book, or re-enabling a previously deconfigured processor or memory book. Both of these menus are submenus under the System Information Menu. You can enable or disable CPU Repeat Gard or Memory Repeat Gard using the Processor Configuration/Deconfiguration Menu, which is a submenu under the System Information Menu.

2.1.7.1 SP Menu - System Information Menu - Processor Configuration/Deconfiguration Menu

PROCESSOR CONFIGURATION/DECONFIGURATION MENU
77. Enable/Disable CPU Repeat Gard: Currently Enabled

1. 0 3.0 (00) Configured by system	2. 1 3.1 (00) Deconfigured by system
3. 2 3.2 (00) Configured by system	4. 3 3.3 (00) Configured by system
5. 4 3.4 (00) Configured by system	6. 5 3.5 (00) Deconfigured by system
7. 6 3.6 (00) Configured by system	8. 7 3.7 (00) Configured by system

98. Return to Previous Menu0>

To enable or disable CPU repeat gard, use menu option 77.
CPU repeat gard is enabled by default. If CPU repeat gard is disabled, processors that are in the "deconfigured by system" state will be reconfigured. These reconfigured processors are then tested during the boot process, and if they pass, they remain online. If they fail the boot testing, they are deconfigured even though CPU repeat gard is disabled. The failure history of each CPU is retained. If a processor with a history of failures is brought back online by disabling repeat gard, it remains online if it passes testing during the boot process. However, if repeat gard is enabled, the processor is taken offline again because of its history of failures.

2.1.7.2 SP Menu - System Information Menu - Memory Configuration/Deconfiguration Menu

MEMORY CONFIGURATION/DECONFIGURATION MENU
77. Enable/Disable Memory Repeat Gard: Currently Enabled
1. Memory card
98. Return to Previous Menu

To enable or disable Memory Repeat Gard, use menu option 77 of the Memory Configuration/Deconfiguration Menu.
The failure history of each book is retained. If a book with a history of failures is brought back online by disabling Repeat Gard, it remains online if it passes testing during the boot process. However, if Repeat Gard is enabled, the book is taken offline again because of its history of failures.

2.1.8 SP Menus Which Should Not be Changed from Default Values

The following SP menus should be left in their default state for proper operation of the p690 system with the HMC:

- *Ring Indicate Power on*
- *Enable/Disable Modem*
- *Setup Modem Configuration*
- *Setup Dial-out Phone Numbers*

- *Select Modem Line Speed*

These functions are provided on the HMC and should be configured there since the modem will be supported from the HMC instead of attaching to the serial port on the CEC as on prior products.

2.2 AIX Options

Option		Default Value	Recommended Setting	Invocation Method
Enable CPU Dynamic Deallocation	Enable/Disable CPU Dynamic Deallocation	Disable	Enable	**AIX cmd line** **Web SM** **SMIT -** CPU Gard
Hot-plug Task	PCI Hot-plug Manager	Used for Concurrent Maintenance		**Service AID** - Hot-plug Task
	SCSI Hot-swap Manager			
	RAID Hot-plug Devices			
Periodic Diagnostics	Disable or Enable Automaic Error Log Analysis	Enable	Enable	**Service AID** - Periodic Diagnostics **AIX Command Line**
Save or Restore Hardware Management Policies	Surveillance Policy and Reboot Policy	Used to save hardware and service processor settings		**Service AID** - Save or Restore Hardware Management Policies
Save/Restore Service Processor settings	Service Processor settings			**Service AID** - Save or Restore Service Processor settings
Update System or SP Flash	Command line entry	Used to update System or Service Processor Flash		**Service AID** - Update System or Service Processor Flash

2.2.1 Enable CPU Dynamic Deallocation

L1 instruction cache recoverable errors, L1 data cache correctable errors, and L2 cache correctable errors are monitored by the processor runtime diagnostics (PRD) code running in the service processor. When a predefined error threshold is met, an error log with warning severity and threshold exceeded status is returned to AIX. At the same time, PRD marks the CPU for deconfiguration at the next boot. AIX will attempt to migrate all resources associated with that processor to another processor and then stop the defective processor.

These errors are not fatal and, as long as they remain rare occurrences, can be safely ignored. However, when a pattern of failures seems to be developing on a specific processor, this pattern may indicate that

this component is likely to exhibit a fatal failure in the near future. This prediction is made by the firmware based-on-failure rates and threshold analysis.

AIX, on these systems, implements continuous hardware surveillance and regularly polls the firmware for hardware errors. When the number of processor errors hits a threshold and the firmware recognizes that there is a distinct probability that this system component will fail, the firmware returns an error report to AIX. In all cases, AIX logs the error in the system error log. In addition, on multiprocessor systems, depending on the type of failure, AIX attempts to stop using the untrustworthy processor and deallocate it. This feature is called Dynamic Processor Deallocation.

At this point, the processor is also flagged by the firmware for persistent deallocation for subsequent reboots, until maintenance personnel replaces the processor.

2.2.1.1 Potential Impact to Applications

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

The interface for binding processes or threads to processors uses logical CPU numbers. The logical CPU numbers are in the range [0..N-1] where N is the total number of CPUs. To avoid breaking applications or kernel extensions that assume no "holes" in the CPU numbering, AIX always makes it appear for applications as if it is the "last" (highest numbered) logical CPU to be deallocated. For instance, on an 8-way SMP, the logical CPU numbers are [0..7]. If one processor is deallocated, the total number of available CPUs becomes 7, and they are numbered [0..6]. Externally, it looks like CPU 7 has disappeared, regardless of which physical processor failed. In the rest of this description, the term CPU is used for the logical entity and the term processor for the physical entity.

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

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

Even in the rare cases where the deallocation cannot go through, Dynamic Processor Deallocation still gives advanced warning to system administrators. By recording the error in the error log, it gives them a chance to schedule a maintenance operation on the system to replace the ailing component before a global system failure occurs.

2.2.1.2 Processor Deallocation:

The typical flow of events for processor deallocation is as follows:

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

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

2.2.1.3 AIX Interface for Turning Processor Deallocation On and Off

Dynamic Processor Deallocation can be enabled or disabled by changing the value of the cpugard attribute of the ODM object sys0. The possible values for the attribute are enable and disable.

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

> Note: If processor deallocation is turned off, AIX still reports the errors in the error log and you will see the error indicating that AIX was notified of the problem with a CPU (CPU_FAILURE_PREDICTED, see the following format).

2.2.1.4 Restarting an Aborted Processor Deallocation

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

The syntax for this command is:

```
ha_star -C
```

where -C is for a CPU predictive failure event.

Processor State Considerations

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

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

enable
The processor is used by AIX.
disable
The processor has been dynamically deallocated by AIX.
faulty
The processor was declared defective by the firmware at boot time.

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

2.2.2 Hot-plug Task

The Hot-plug Task provides software function for those devices that support hot-plug or hot-swap capability. This includes PCI adapters, SCSI devices, and some RAID devices. This task was previously known as "SCSI Device Identification and Removal" or "Identify and Remove Resource". The Hot-plug Task has a restriction when running in Standalone or Online Service mode; new devices may not be added to the system unless there is already a device with the same FRU part number installed in the system. This restriction is in place because the device software package for the new device cannot be installed in Standalone or Online Service mode.
Depending on the environment and the software packages installed, selecting this task displays the following three subtasks:

- PCI Hot-plug Manager
- SCSI Hot-swap Manager
- RAID Hot-plug Devices

To run the Hot-plug Task directly from the command line, type the following: Diag -T"identifyRemove"
If you are running the diagnostics in Online Concurrent mode, run the Missing Options Resolution Procedure immediately after adding, removing or replacing any device. Start the Missing Options Resolution Procedure is by running the **diag -a** command. If the Missing Options Resolution Procedure runs with no menus or prompts, then device configuration is complete. Otherwise, work through each menu to complete device configuration.

2.2.2.1 PCI Hot-plug Manager

The PCI Hot-plug Manager task is a SMIT menu that allows you to identify, add, remove, or replace PCI adapters that are hot-pluggable. The following functions are available under this task:

- List PCI hot-plug slots
- Add a PCI hot-plug adapter
- Replace/Remove a PCI hot-plug adapter
- Identify a PCI hot-plug slot
- Unconfigure Devices
- Configure Devices
- Install/Configure Devices Added After IPL

The **List PCI Hot-plug Slots** function lists all PCI hot-plug slots. Empty slots and populated slots are listed. Populated slot information includes the connected logical device. The slot name consists of the physical location code and the description of the physical characteristics for the slot. The **Add a PCI Hot-plug Adapter** function is used to prepare a slot for the addition of a new adapter. The function lists all the empty slots that support hot-plug. When a slot is selected, the visual indicator for the slot blinks at the Identify rate. After the slot location is confirmed, the visual indicator for the specified PCI slot is set to the Action state. This means the power for the PCI slot is off and the new adapter can be plugged in.

The **Replace/Remove a PCI Hot-plug Adapter** function is used to prepare a slot for adapter exchange. The function lists all the PCI slots that support hot-plug and are occupied. The list includes the slot's physical location code and the device name of the resource installed in the slot. The adapter must be in the Defined state before it can be prepared for hot-plug removal. When a slot is selected, the visual indicator for the slot is set to the Identify state. After the slot location is confirmed, the visual indicator for the specified PCI slot is set to the Action state. This means the power for the PCI slot, is off and the adapter can be removed or replaced.

The **Identify a PCI Hot-plug Slot** function is used to help identify the location of a PCI hot-plug adapter. The function lists all the PCI slots that are occupied or empty and support hot-plug. When a slot is selected for identification, the visual indicator for the slot is set to the Identify state.

The **Unconfigure Devices** function attempts to put the selected device, in the PCI hot-plug slot, into the Defined state. This action must be done before any attempted hot-plug function. If the unconfigure function fails, it is possible that the device is still in use by another application. In this case, the customer or system administrator must be notified to quiesce the device.

The **Configure Devices** function allows a newly added adapter to be configured into the system for use. This function should also be done when a new adapter is added to the system.

The **Install/Configure Devices Added After IPL** function attempts to install the necessary software packages for any newly added devices. The software installation media or packages are required for this function.

Standalone Diagnostics has restrictions on using the PCI Hot-plug Manager. For example:

- Adapters that are replaced must be exactly the same FRU part number as the adapter being

replaced.

- New adapters cannot be added unless a device of the same FRU part number already exists in the system, because the configuration information for the new adapter is not known after the Standalone Diagnostics are booted.
- The following functions are not available from the Standalone Diagnostics and will not display in the list:
  - Add a PCI Hot-plug Adapter
  - Configure Device
  - Install/Configure Devices Added After IPL

You can run this task directly from the command line by typing the following command:
diag -d *device* -T"identifyRemove"
However, note that some devices support both the PCI Hot-plug Task and the RAID Hot-plug Devices task. If this is the case for the *device* specified, then the Hot-plug Task displays instead of the PCI Hot-plug Manager menu.

2.2.3 SCSI Hot-swap Manager

This task was known as “SCSI Device Identification and Removal” or “Identify and Remove Resources” in previous releases. This task allows the user to identify, add, remove, and replace a SCSI device in a system unit that uses a SCSI Enclosure Services (SES) device. The following functions are available:

- List the SES Devices
- Identify a Device Attached to an SES Device
- Attach a Device to an SES Device
- Replace/Remove a Device Attached to an SES Device
- Configure Added/Replaced Devices

The **List the SES Devices** function lists all the SCSI hot-swap slots and their contents. Status information about each slot is also available. The status information available includes the slot number, device name, whether the slot is populated and configured, and location.

The **Identify a Device Attached to an SES Device** function is used to help identify the location of a device attached to a SES device. This function lists all the slots that support hot-swap that are occupied or empty. When a slot is selected for identification, the visual indicator for the slot is set to the Identify state.

The **Attach a Device to an SES Device** function lists all empty hot-swap slots that are available for the insertion of a new device. After a slot is selected, the power is removed. If available, the visual indicator for the selected slot is set to the Remove state. After the device is added, the visual indicator for the selected slot is set to the Normal state, and power is restored.

The **Replace/Remove a Device Attached to an SES Device** function lists all populated hot-swap slots that are available for removal or replacement of the devices. After a slot is selected, the device populating that slot is Unconfigured; then the power is removed from that slot. If the Unconfigure operation fails, it is possible that the device is in use by another application. In this case, the customer or system administrator must be notified to quiesce the device. If the Unconfigure operation is successful, the visual

indicator for the selected slot is set to the Remove state. After the device is removed or replaced, the visual indicator, if available for the selected slot, is set to the Normal state, and power is restored.
Note: Be sure that no other host is using the device before you remove it.

The **Configure Added/Replaced Devices** function runs the configuration manager on the parent adapters that had child devices added or removed. This function ensures that the devices in the configuration database are configured correctly. Standalone Diagnostics has restrictions on using the SCSI Hot-plug Manager. For example:

- Devices being used as replacement devices must be exactly the same type of device as the device being replaced.
- New devices may not be added unless a device of the same FRU part number already exists in the system, because the configuration information for the new device is not known after the Standalone Diagnostics are booted. You can run this task directly from the command line.

See the following command syntax: diag -d *device* -T"identifyRemove" OR
diag [-c]-d *device* -T"identifyRemove -a [identify|remove]"
Flag Description
-a Specifies the option under the task.
-c Run the task without displaying menus. Only command line prompts are used. This flag is only applicable when running an option such as identify or remove.
-d Indicates the SCSI device.
-T Specifies the task to run.

2.2.4 RAID Hot-plug Devices

PCI RAID Physical Disk Identify
This selection identifies physical disks connected to a PCI SCSI-2 F/W RAID adapter.
You can run this task directly from the AIX command line. See the following command syntax: diag -c -d *pci RAID adapter* -T identify

2.2.5 Periodic Diagnostics

This selection provides a tool for configuring periodic diagnostics and automatic error log analysis. You can select a hardware resource to be tested once a day, at a user-specified time. By default, the floating point processor diagnostic is run at 4:00 a.m. each day Other devices can be added to the Periodic Diagnostic Device list by setting PDiagAtt->attribute=test_mode. Error log analysis can be directed to run at different times.

Problems are reported by a message to the system console, and a mail message is sent to all members of the system group. The message contains the SRN.

2.2.5.1 Automatic Error Log Analysis (diagela)

Automatic Error Log Analysis (diagela) provides the capability to do error log analysis whenever a permanent hardware error is logged. If the diagela program is enabled and a permanent hardware resource error is logged, the diagela program is started. Automatic Error Log Analysis is enabled by default on all platforms.

The diagela program determines whether the error should be analyzed by the diagnostics. If the error should be analyzed, a diagnostic application will be invoked and the error will be analyzed. No testing is

done. If the diagnostics determines that the error requires a service action, it sends a message to your console and to all system groups. The message contains the SRN, or a corrective action.

If the resource cannot be tested because it is busy, error log analysis is performed. Hardware errors logged against a resource can also be monitored by enabling Automatic Error Log Analysis. This allows error log analysis to be performed every time a hardware error is put into the error log. If a problem is detected, a message is posted to the system console and a mail message sent to the users belonging to the system group containing information about the failure, such as the service request number.
The service aid provides the following functions:

- Add or delete a resource to the periodic test list
- Modify the time to test a resource
- Display the periodic test list
- Modify the error notification mailing list
- Disable or Enable Automatic Error Log Analysis

2.2.5.2 AIX Command Line Invocation

To activate the Automatic Error Log Analysis feature, log in as root and type the following command:

/usr/lpp/diagnostics/bin/diagela ENABLE

To disable the Automatic Error Log Analysis feature, log in as root and type the following command:

/usr/lpp/diagnostics/bin/diagela DISABLE

2.2.6 Save or Restore Hardware Management Policies

Use this service aid to save or restore the settings from Ring Indicate Power-On Policy, Surveillance Policy, Remote Maintenance Policy and Reboot Policy. The following options are available:

- Save Hardware Management Policies

This selection writes all of the settings for the hardware-management policies to the following file: **/etc/lpp/diagnostics/data/hmpolicies**

- Restore Hardware Management Policies

This selection restores all of the settings for the hardware-management policies from the contents of the following file: **/etc/lpp/diagnostics/data/hmpolicies**
You can access this service aid directly from the AIX command line, by typing:
/usr/lpp/diagnostics/bin/uspchrp -a

2.2.7 Save/Restore Service Processor Configuration

Use this service aid to save or restore the Service Processor Configuration to or from a file. The Service Processor Configuration includes the Ring Indicator Power-On Configuration. The following options are available:

- Save Service Processor Configuration

This selection writes all of the settings for the Ring Indicate Power On and the Service Processor to the following file: **/etc/lpp/diagnostics/data/spconfig**

- Restore Service Processor Configuration

This selection restores all of the settings for the Ring Indicate Power On and the Service Processor from the following file: **/etc/lpp/diagnostics/data/spconfig**

2.2.8 Update System or SP Flash

2.2.8.1 AIX Command

You can use the **update_flash** command to perform this function. The command is located in the **/usr/lpp/diagnostics/bin** directory. The command syntax is as follows:
update_flash [-q]-f *file_name* update_flash [-q]-D *device_name* -f *file_name*
update_flash [-q]-D *device_name* -l
Flag Description
-q Forces the **update_flash** command to update the flash EPROM and reboot the system without asking for confirmation.
-D Specifies that the flash update image file is on diskette. The *device_name* variable specifies the diskette drive. The default *device_name* is /dev/fd0 .
-f Flash update image file source. The *file_name* variable specifies the fully qualified path of the flash update image file.
-l Lists the files on a diskette for the user to choose a flash update image file.
Attention: The **update_flash** command reboots the entire system. Do not use this command if more than one user is logged on to the system.

2.2.8.2 Service Aid

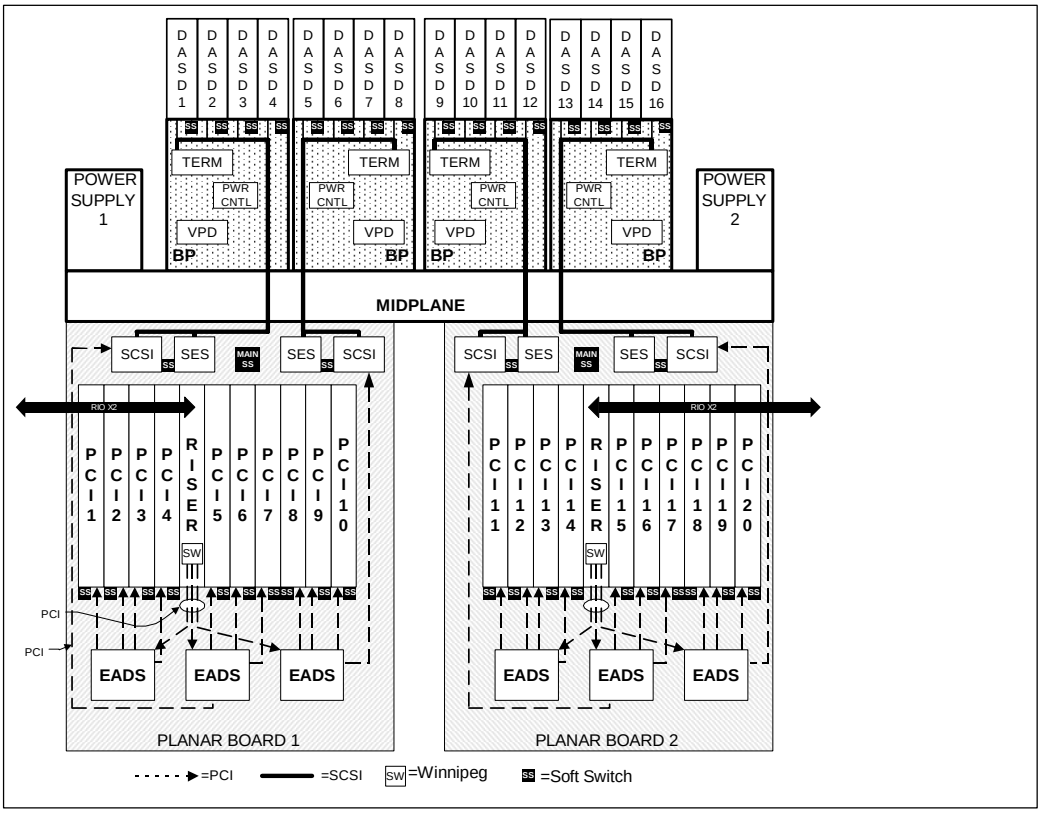
You must know the fully qualified path and file name of the flash update image file that was provided. If the flash update image file is on a diskette, the service aid can list the files on the diskette for selection. The diskette must be a valid backup diskette. Refer to the update instructions, or the service guide for the system unit to determine the level of the system unit or service processor flash. When this service aid is run from online diagnostics, the flash update image file is copied to the **/var** file system. If there is not enough space in the **/var** file system for the flash update image file, an error is reported. If this error occurs, exit the service aid, increase the size of the **/var** file system, and retry the service aid. After the file is copied, a screen requests confirmation before continuing with the update flash. Continuing the update flash reboots the system using the **shutdown -u** command. The system does not return to diagnostics, and the current flash image is not saved. After the reboot, remove the **/var/update_flash_image** file.

When this service aid is run from standalone diagnostics, the flash update image file is copied to the file system from diskette. The user must provide the image on a backup diskette because the user does not have access to remote file systems or any other files that are on the system. If not enough space is available, an error is reported, stating additional system memory is needed. After the file is copied, a screen requests confirmation before continuing with the update flash. Continuing the update flash reboots the system using the **reboot -u** command. You may receive a Caution: some process(es) wouldn't die message during the reboot process. You can ignore this message. The current flash image is not saved.

2.3 Disk Configuration

2.3.1 Boot Disk - Stand-Alone vs. Mirrored

In order to minimize single points of fail the system should be configured with mirrored Boot Disks. In addition to the redundant Boot Disk, the highest availability configuration will consist of two separate I/O drawers, with two separate SCSI adapters in each drawer, each of the mirrored Boot Disks connected to one of these two SCSI adapters. Refer to section "I/O Drawer Additions" for specific placement suggestions for the redundant adapters and disks.



The user will establish the configuration using the following procedure, described here at a high level. The user will need to use SMIT.

1. The user would have to install AIX.
2. Then the user would have to add the appropriate other disk to the root volume group.
3. Then the user would have to mirror the AIX boot logical volume to the other disk.
4. The user would have to use the bootlist command to specify both disks as boot devices.

Below are the commands you will need.

mirrorvg Command

Purpose

Mirrors all the logical volumes that exist on a given volume group. This command only applies to AIX Version 4.2.1 or later.

Syntax

mirrorvg [**-S** | **-s**] [**-Q**] [**-c** *Copies*] [**-m**] *VolumeGroup* [*PhysicalVolume ...*]

Description

The **mirrorvg** command takes all the logical volumes on a given volume group and mirrors those logical volumes. This same functionality may also be accomplished manually if you execute the **mklvcopy** command for each individual logical volume in a volume group. As with **mklvcopy**, the target physical drives to be mirrored with data must already be members of the volume group. To add disks to a volume group, run the **extendvg** command.

By default, **mirrorvg** attempts to mirror the logical volumes onto any of the disks in a volume group. If you wish to control which drives are used for mirroring, you must include the list of disks in the input parameters, *PhysicalVolume.* Mirror strictness is enforced. Additionally, **mirrorvg** mirrors the logical volumes, using the default settings of the logical volume being mirrored. If you wish to violate mirror strictness or affect the policy by which the mirror is created, you must execute the mirroring of all logical volumes manually with the **mklvcopy** command.

When **mirrorvg** is executed, the default behavior of the command requires that the synchronization of the mirrors must complete before the command returns to the user. If you wish to avoid the delay, use the **-S** or **-s** option. Additionally, the default value of 2 copies is always used. To specify a value other than 2, use the **-c** option.

Notes:

1. This command ignores striped logical volumes. Mirroring striped logical volumes is not possible.
2. To use this command, you must either have root user authority or be a member of the **system** group.

Notes:

1. This command ignores striped logical volumes.
2. To use this command, you must either have root user authority or be a member of the **system** group.

Note: To use this command, you must either have root user authority or be a member of the **system** group.

Attention: The **mirrorvg** command may take a significant amount of time before completing because of complex error checking, the amount of logical volumes to mirror in a volume group, and the time is takes to synchronize the new mirrored logical volumes.

You can use the Web-based System Manager Volumes application (**wsm lvm** fast path) to run this command. You could also use the System Management Interface Tool (SMIT) **smit mirrorvg** fast path to run this command.

Flags

-c *Copies*	Specifies the minimum number of copies that each logical volume must have after the **mirrorvg** command has finished executing. It may be possible, through the independent use of **mklvcopy**, that some logical volumes may have more than the minimum number specified after the **mirrorvg** command has executed. Minimum value is 2 and 3 is the maximum value. A value of 1 is ignored.
-m *exact map*	Allows mirroring of logical volumes in the exact physical partition order that the original copy is ordered. This option requires you to specify a `PhysicalVolume(s)` where the exact map copy should be placed. If the space is insufficient for an exact mapping, then the command will fail. You should add new drives or pick a different set of drives that will satisfy an exact logical volume mapping of the entire volume group. The designated disks must be equal to or exceed the size of the drives which are to be exactly mirrored, regardless of if the entire disk is used. Also, if any logical volume to be mirrored is already mirrored, this command will fail.
-Q *Quorum Keep*	By default in **mirrorvg**, when a volume group's contents becomes mirrored, volume group quorum is disabled. If the user wishes to keep the volume group quorum requirement after mirroring is complete, this option should be used in the command. For later quorum changes, refer to the **chvg** command.
-S *Background Sync*	Returns the **mirrorvg** command immediately and starts a background **syncvg** of the volume group. With this option, it is not obvious when the mirrors have completely finished their synchronization. However, as portions of the mirrors become synchronized, they are immediately used by the operating system in mirror usage.
-s *Disable Sync*	Returns the **mirrorvg** command immediately without performing any type of mirror synchronization. If this option is used, the mirror may exist for a logical volume but is not used by the operating system until it has been synchronized with the **syncvg** command.

The following is a description of **rootvg**:

rootvg mirroring	When the **rootvg** mirroring has completed, you must perform three additional tasks: **bosboot, bootlist**, and **reboot**. The **bosboot** command is required to customize the bootrec of the newly mirrored drive. The **bootlist** command needs to be performed to instruct the system which disk and order you prefer the mirrored boot process to start. Finally, the default of this command is for Quorum to be turned off. For this to take effect on a **rootvg** volume group, the system must be rebooted.
non-rootvg mirroring	When this volume group has been mirrored, the default command causes Quorum to deactivated. The user must close all open logical volumes, execute **varyoffvg** and then **varyonvg** on the volume group for the system to understand that quorum is or is not needed for the volume group. If you do not **revaryon** the volume group,

mirror will still work correctly. However, any quorum changes will not have taken effect.

rootvg and non-rootvg mirroring The system dump devices, primary and secondary, should not be mirrored. In some systems, the paging device and the dump device are the same device. However, most users want the paging device mirrored. When **mirrorvg** detects that a dump device and the paging device are the same, the logical volume will be mirrored automatically.

If **mirrorvg** detects that the dump and paging device are different logical volumes, the paging device is automatically mirrored, but the dump logical volume is not. The dump device can be queried and modified with the **sysdumpdev** command.

Examples

1. To triply mirror a volume group, enter:

```
mirrorvg -c 3 workvg
```

The logical partitions in the logical volumes held on `workvg` now have three copies.

2. To get default mirroring of rootvg, enter:

```
mirrorvg rootvg
```

`rootvg` now has two copies.

3. To replace a bad disk drive in a mirrored volume group, enter

```
unmirrorvg workvg hdisk7
reducevg workvg hdisk7
rmdev -l hdisk7 -d
replace the disk drive, let the drive be renamed hdisk7
extendvg workvg hdisk7
mirrorvg workvg
```

Note: By default in this example, **mirrorvg** will try to create 2 copies for logical volumes in `workvg`. It will try to create the new mirrors onto the replaced disk drive. However, if the original system had been triply mirrored, there may be no new mirrors created onto `hdisk7`, as other copies may already exist for the logical volumes.

4. To sync the newly created mirrors in the background, enter:

```
mirrorvg -S -c 3 workvg
```

5. To create an exact mapped volume group, enter:

```
mirrorvg -m datavg hdisk2 hdisk3
```

Implementation Specifics

Software Product/Option:	Base Operating System/ AIX 3.2 to 4.1 Compatibility Links
Standards Compliance:	NONE

Files

/usr/sbin	Directory where the **mirrorvg** command resides.

2.3.2 User Disks - Stand-Alone vs. RAID Disks

Availability advantages can be leveraged for user disks by implementing RAID technology. There are many types of RAID configurations which have varying degrees of availability, performance and cost factors. This section will describe the various RAID modes along with their benefits and detriments.

Redundant Array of Independent Disks (RAID) is a term used to describe the technique of improving data availability through the use of arrays of disks and various data-striping methodologies. Disk arrays are groups of disk drives that work together to achieve higher data-transfer and I/O rates than those provided by single large drives. An array is a set of multiple disk drives plus a specialized controller (an array controller) that keeps track of how data is distributed across the drives. Data for a particular file is written in segments to the different drives in the array rather than being written to a single drive.

Arrays can also provide data redundancy so that no data is lost if a single drive (physical disk) in the array should fail. Depending on the RAID level, data is either mirrored or striped.

Subarrays are contained within an array subsystem. Depending on how you configure it, an array subsystem can contain one or more sub-arrays, also referred to as Logical Units (LUN). Each LUN has its own characteristics (RAID level, logical block size and logical unit size, for example). From the operating system, each subarray is seen as a single hdisk with its own unique name.
RAID algorithms can be implemented as part of the operating system's file system software, or as part of a disk device driver (common for RAID 0 and RAID 1). These algorithms can be performed by a locally embedded processor on a hardware RAID adapter. Hardware RAID adapters generally provide better performance than software RAID because embedded processors offload the main system processor by performing the complex algorithms, sometimes employing specialized circuitry for data transfer and Manipulation.

2.3.2.1 RAID Levels and Their Performance Implications

Each of the RAID levels supported by disk arrays uses a different method of writing data and hence provides different benefits.

RAID 0 is also known as data striping. RAID 0 is only designed to increase performance; there is no redundancy, so any disk failures require reloading from backups. It is not recommended to use this level for critical applications that require high availability.

RAID 1 is also known as disk mirroring. It is most suited to applications that require high data availability, good read response times, and where cost is a secondary issue. The response time for writes can be somewhat slower than for a single disk, depending on the write policy; the writes can either be executed in parallel for speed or serially for safety. Select RAID Level 1 for applications with a high percentage of read operations and where the cost is not the major concern.

RAID 2 is rarely used. It implements the same process as RAID 3, but can utilize multiple disk drives for parity, while RAID 3 can use only one.

RAID 3 and RAID 2 are parallel process array mechanisms, where all drives in the array operate in unison. Similar to data striping, information to be written to disk is split into chunks (a fixed amount of data), and each chunk is written out to the same physical position on separate disks (in parallel). More

advanced versions of RAID 2 and 3 synchronize the disk spindles so that the reads and writes can truly occur simultaneously (minimizing rotational latency buildups between disks). This architecture requires parity information to be written for each stripe of data; the difference between RAID 2 and RAID 3 is that RAID 2 can utilize multiple disk drives for parity, while RAID 3 can use only one. The LVM does not support Raid 3; therefore, a RAID 3 array must be used as a raw device from the host system. RAID 3 provides redundancy without the high overhead incurred by mirroring in RAID 1.

RAID 4 addresses some of the disadvantages of RAID 3 by using larger chunks of data and striping the data across all of the drives except the one reserved for parity. Write requests require a read/modify/update cycle that creates a bottleneck at the single parity drive. Therefore, RAID 4 is not used as often as RAID 5, which implements the same process, but without the parity volume bottleneck.

RAID 5, as has been mentioned, is very similar to RAID 4. The difference is that the parity information is distributed across the same disks used for the data, thereby eliminating the bottleneck. Parity data is never stored on the same drive as the chunks that it protects. This means that concurrent read and write operations can now be performed, and there are performance increases due to the availability of an extra disk (the disk previously used for parity). There are other enhancements possible to further increase data transfer rates, such as caching simultaneous reads from the disks and transferring that information while reading the next blocks. This can generate data transfer rates at up to the adapter speed.

RAID 6 is similar to RAID 5, but with additional parity information written that permits data recovery if two disk drives fail. Extra parity disk drives are required, and write performance is slower than a similar implementation of RAID 5.

The RAID 7 architecture gives data and parity the same privileges. The level 7 implementation allows each individual drive to access data as fast as possible. This is achieved by three features:

1. Independent control and data paths for each I/O device/interface.
2. Each device/interface is connected to a high-speed data bus that has a central cache capable of supporting multiple host I/O paths.
3. A real time, process-oriented operating system is embedded into the disk drive array architecture. The embedded operating system "frees" the drives by allowing each drive head to move independently of the other disk drives. Also, the RAID 7 embedded operating system is enabled to handle a heterogeneous mix of disk drive types and sizes.

RAID 10 - RAID-0+1

RAID-0+1, also known in the industry as RAID 10, implements block interleave data striping and mirroring. RAID 10 is not formally recognized by the RAID Advisory Board (RAB), but, it is an industry standard term. In RAID 10, data is striped across multiple disk drives, and then those drives are mirrored to another set of drives.

RAID 10 provides an enhanced feature for disk mirroring that stripes data and copies the data across all the drives of the array. The first stripe is the data stripe; the second stripe is the mirror (copy) of the first data stripe, but it is shifted over one drive. Because the data is mirrored, the capacity of the logical drive is 50 percent of the physical capacity of the hard disk drives in the array.

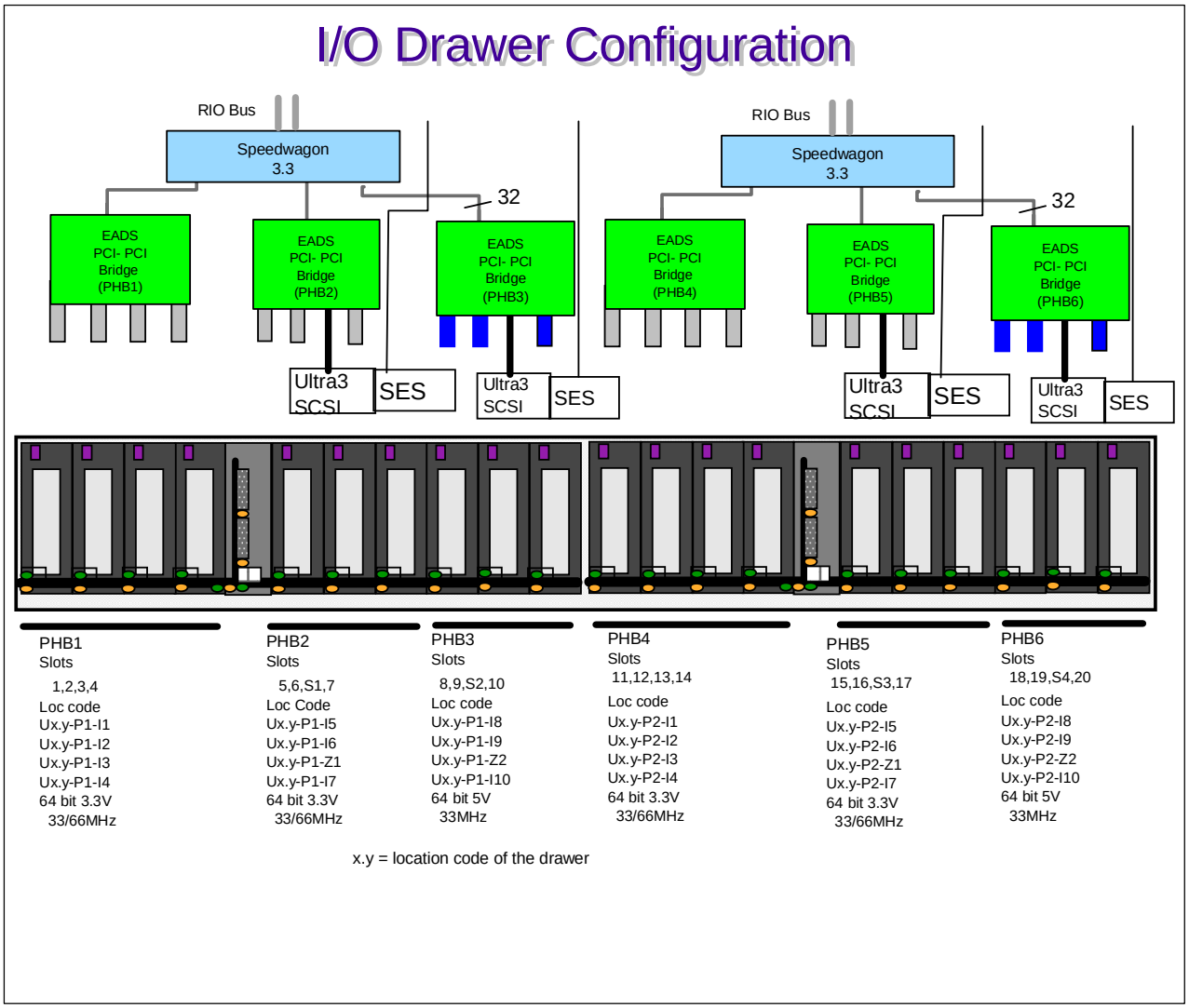
2.4 I/O Drawer and Adapter Configuration

2.4.1 Adapter Placement

The following diagram (with reference to the diagram in the I/O drawer section which follows) will be used to help explain the placement of I/O adapters within and across I/O drawers in order to obtain increasing levels of availability.

The PCI adapter slots controlled by the various PCI Host Bridges are shown in the diagram below.

For placement of redundant adapters or disks, refer to the following section on “I/O Drawer Additions”.



For more information about your device and its capabilities, see the documentation shipped with that device. For a list of supported adapters and a detailed discussion about adapter placement, refer to the *PCI Adapter Placement Referenc*e, order number SA38-0538.

2.4.1.1 Non Enhanced Error Handling (EEH) Adapters

Because some devices do not have enhanced error handling (EEH) capabilities built into their device drivers, non-EEH I/O Adapters (IOA) on a PHB should be solely in one partition - do not split the PHB between partitions as a failure on a non-EEH adapter will affect all other adapters on that PHB.

This description uses the term "EADS", which is IBM unique hardware in the I/O drawer controlling each PCI slot. EADS, firmware, and device drivers act in concert to support EEH.
The exploitation of EADS error handling occurs in three scenarios:

- During firmware probing of the PCI space during boot time configuration
- During PCI hot-plug when RTAS configures a PCI adapter after insertion/replacement
- During normal run-time operation or AIX diagnostic run.

At boot time, firmware can now deconfigure adapters which cause errors on the PCI bus, and continue with the boot process, whereas in the past these types of failures would have caused machine checks and prevented the system from booting. The basic firmware process is as follows:

1. Detect and configure Speedwagon (PHB) bridges.
2. Detect and configure EADS bridges. In each parent PHB, an "ibm,eeh-implemented" property is added to indicate the number and addresses of all freeze mode capable slots under that PHB.
3. Set all EADS bridges to freeze on error (Bridge Arbitration, Config space 0x0040, bit 16 =1)
4. Probe for devices under each EADS bridge.
5. If a return value of 0xFFFFFFFF is returned on any of the PCI config cycles or direct load requests, check the freeze state of the bridge (Bridge Interrupt Status, BAR 0 + 0x1234, bit 25:26)
6. If frozen, bypass the configuration of that adapter, leave in freeze mode, and continue with PCI probing.
7. Before turning control over to AIX, firmware must return all slots to a freeze mode disabled state.
8. Device drivers that support freeze mode detection will re-enable freeze mode for their respective adapters on a slot-by-slot basis, using the *ibm,set-eeh-option* RTAS call.

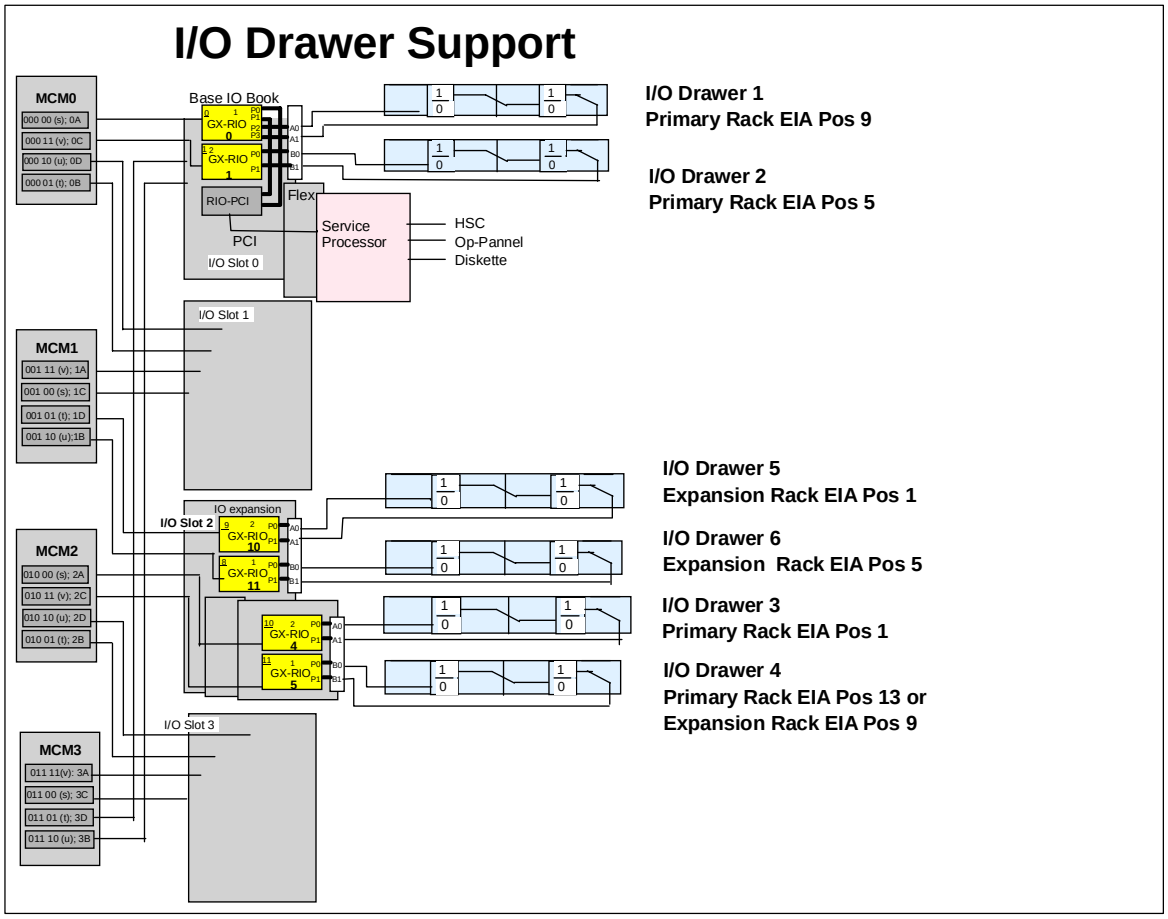
During PCI hot-plug, the firmware scenario would be similar to steps 3-8 above, but targeted to a single EADS bridge and slot that is being configured by the *ibm,configure-connector* RTAS call.

To support EADS-based recovery during run-time operation, AIX device drivers and diagnostic must detect the return values of 0xFFFFFFFF on MMIO Loads to the device address space, and then:

- Use the *ibm,read-slot-reset-state* RTAS function to detect if the bridge is in freeze state
- Reset the device using the *ibm,set-slot-reset* RTAS function
- Attempt any appropriate recovery for device operations

2.4.2 I/O Drawer Additions

The following diagram shows the advantages of configuring redundant adapters and disks across multiple I/O drawers to eliminate potential single points of failure.



The following is the list of least to most preferred configurations for populating redundant adapters and disks in the I/O drawers to provide increasing levels of redundancy and therefore increasing levels of availability.

1. Redundant adapters/disks within same half of one I/O drawer (i.e. same I/O planar)
2. Redundant adapters/disks across I/O planars (half drawers) within same I/O drawer
3. Redundant adapters/disks across I/O drawer pairs on same I/O card (i.e. I/O drawer pairs 1 & 2 or pairs 3 & 4 or pairs 5 & 6)
4. Redundant adapters/disks across I/O Driver Cards (i.e. I/O Drawer 1 or 2 and 3 or 4, etc.)
5. Redundant adapters/disks across Racks (i.e. I/O Drawer 1, 2, 3 or 4 and 5 or 6)

While each increasing level in the list above removes additional single points of failure from the configuration, because of the focus on high reliability components and hardening of single point of failure components, any configuration at or above configuration 2 utilizing redundancy across I/O planars is suggested. The customer will have to decide how much to invest in redundancy to obtain the required level of availability for their environment.

2.5 Service Enablement Through HMC Operations

2.5.1 IBM Hardware Management Console for pSeries (HMC)

The IBM Hardware Management Console for pSeries (HMC) provides features never before available on the IBM @server pSeries line of servers. The HMC uses its serial connection (private Service Link) to one or more Central Electronics Complexes (CECs) to perform various functions.

The HMC's main functions include the following:

- Creating and maintaining a multiple partition environment
- Detecting, reporting, and storing changes in hardware conditions
- Powering managed systems on and off
- Acting as a service focal point for service representatives to determine an appropriate service strategy
- Displaying a virtual operating system session terminal for each partition
- Displaying a virtual 2x16 operator panel of contents for each partition

Service Enablement	HMC Operation	Default Value/Setting	Purpose/Reason
Creating Service Rep ID	User Management	Service Representative	Create Service Rep. User ID for system service
Creating Software Support ID	User Management	hmcpe	Create userid for Software Support Personnel to analyze HMC problems
Establish HMC to OS partition network link for error reporting and management	System Configuration "**Customizing Network Settings"**	No network links established as default. Require network link for OS reported errors	Establish link between HMC and Operating System for service management
Enable systems for automatic call home feature	**Service Focal Point "Enable / Disable Call Home**"	Unknown	Set system enablement for call for service when serviceable action is reached
Configure Service Agent to automatically call for service on error	**Service Agent** "Service Agent UI - Registration/Customization"	Not configured	Configure S.A. Parameters to allow automatic call for service on error
Enable for gathering of extended error data	**Service Focal Point** "**Enable / Disable Extended Error Data Collection**"	Unknown	Enable systems to gather extended error data for service events
Connecting Two HMCs to single system for redundancy	System Management Environment - Navigation Area	None	Configure redundant HMC in navigation area for viewing events
Schedule Critical Console Data Backups	**System Configuration** "**Scheduled Operations**"	None	Backup critical console data (including service data) on scheduled basis

Note: These actions need to be performed on both HMCs if redundant HMCs are installed.

Refer to the "Hardware Management Console for pSeries Operations Guide" for additional information on how to perform the service enablements described.

2.5.2 Creating Service Representative and hmcpe User IDs

Accessing the User Application

To access the User Management application, do the following:
1. In the Navigation area, click the **User** icon.
2. Use the **User** or **Selected** menu to create, modify, delete, or view user information.

Creating a User

This process allows you to create a user.
To create users you, must be a member of the System Administrator role.
To create a user, do the following:
1. In the Navigation area, select the **User** icon.
2. Select **User** from the menu.
3. Select **Ne**w.
4. Select **User** from the cascade menu.
5. In the **Login Name** field, type the login name.
6. In the **Full Name** field, type the full name.
7. Click an item in the role list to select a role for your new user.
8. Click **O**K.
9. In the first field of the Change User Password window, type the user's password.
10. Type the same password again in the **Retype new password** field.
11. Click **O**K.

The new user displays in the Contents area.
Note: It is strongly recommended that you create a user named ***hmcpe*** for software fixes and updates from your software support personnel. Support may need to log on to your HMC using this username when analyzing a problem.

2.5.3 Establish HMC to OS Partition Network Link for Error Reporting and Management

In order for recovered errors which require service to be sent to the Service Focal Point application on the hardware maintenance console, there must be a network link from the operating system image running on the system to the hardware maintenance console. This link is configured on the hardware maintenance console utilizing the System Configuration set of tools.

2.5.3.1 Customizing Network Settings

Use this section to attach the HMC to a network.
Customize your network settings to edit IP (internet protocol) addresses, name services, and routing information.
Note: Changes made to your HMC's network settings do not take effect until you reboot the HMC.
To customize network settings, you must be a member of one of the following roles:

- Advanced Operator
- System Administrator
- Service Representative

2.5.3.2 Setting the IP Address

To customize your HMC's IP address, do the following:
1. In the Navigation area, click the **System Configuration** icon.
2. In the Contents area, click **Customize Network Setting**s. The three-tabbed Network Configuration window displays.
3. Click the **IP Address** tab.
4. Type TCP/IP and gateway information as appropriate. For questions about your network and how it is configured, refer to your network administrator.
5. Click **OK** if you are finished customizing the network.

Note: Changes made to your HMC's network settings do not take effect until you reboot the HMC.

2.5.3.3 Setting Domain Names

Use this section to change the default domain names and enter your own.
1. In the Navigation area, click the **System Configuration** icon.
2. In the Contents area, click **Network Configuratio**n. The three-tabbed Network Configuration window displays.
3. Click the **Name Services** tab.
4. The system displays the default hostname as localhost and the default domain name as localdomain . Replace these names with your Domain Name Service (DNS) information as appropriate. For questions about your network and how it is configured, refer to your network administrator.
5. Click **O**K

.

2.5.3.4 Setting Routing Information

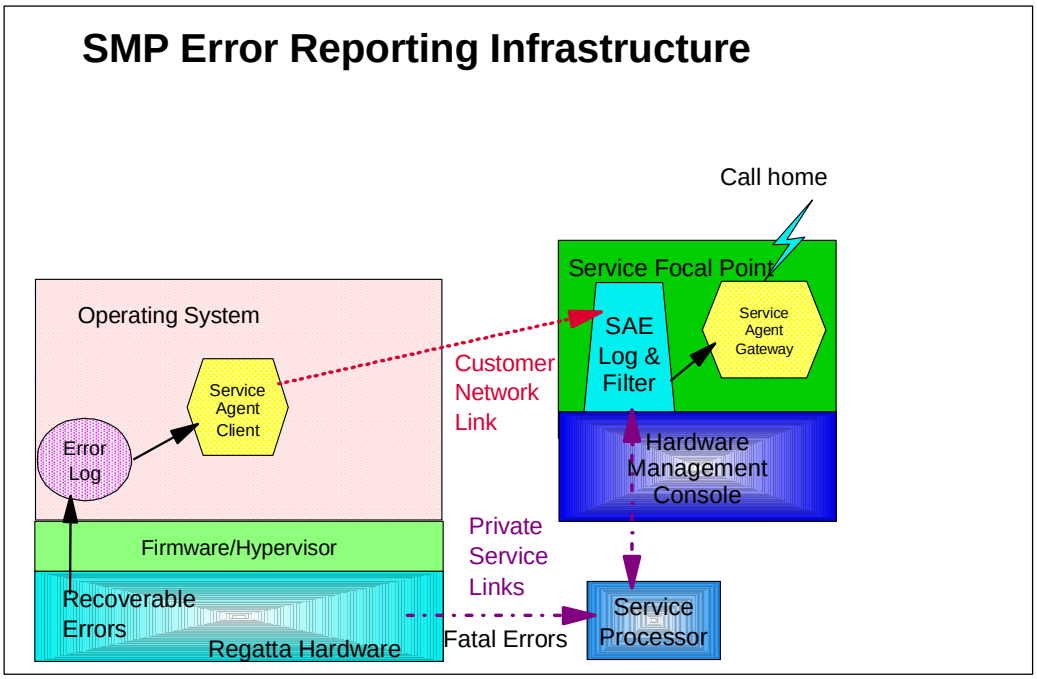
This option allows you to add new routing information, change existing routing information, or delete routing information.
1. In the Navigation area, click the **System Configuration** icon.
2. In the Contents area, click **Network Configuratio**n. The three-tabbed Network Configuration window displays.
3. Click the **Routing** tab.
4. Select **Ne**w, **Chang**e, o**r Delet**e.
5. Type the gateway information in the fields as appropriate. For questions about your network and how it is configured, refer to your network administrator.
6. Click **OK.**

2.5.4 Service Focal Point

Service Representatives use Service Focal Point to start and end their service calls and provides service representatives with event and diagnostic information, as well as Vital Product Data (VPD). The HMC can also notify service representatives of hardware failures automatically by using a feature called *Service Agen*t. You can configure the HMC to use Service Agent's call-home feature to send event information. This information is stored, analyzed, and then acted upon by your service representative.
In SMP and LPAR modes, the HMC must have a LAN connection to each operating system image in order to report errors. You need to configure some parts of Service Focal Point so that the proper information is sent.

2.5.4.1 Enabling Error Reporting to SFP Application

The following diagram will be used to explain how to configure the error reporting paths from the system to the Service Focal Point application on the IBM hardware management console for pseries in SMP mode.



2.5.4.2 Enabling The Automatic Call-Home Feature

The Hardware management console can be configured to automatically call an appropriate service center when it identifies a serviceable event. Use the following processes to configure the call home feature.
To enable and disable the Call Home feature for one or more systems:
Note: It is strongly recommended that you not disable the call-home feature.
1. In the Navigation area, select the Service Focal Point icon.
2. In the Contents area, select **Enable / Disable Call Hom**e.
3. The **Enable / Disable call home** window displays a list of managed systems. Click on the managed system you want to enable or disable from the list.
4. Click **Enable Selected** to enable call-home for the selected system, or click **Disable Selected** to disable call home for the selected system.
5. Click **OK.**

2.5.4.3 Enabling Extended Error Data Collection

This feature allows you to change the state of the collection of extended error data (EED) in one or more managed systems. There are two types of EED that can be activated and deactivated independently: Central Service Processor (CSP) EED and Operating System (OS) EED. CSP EED describes the current partition status on the managed system, and OS EED is the data collected from the erring partition.
To enable or disable extended error data, do the following:

1. In the Navigation area, select the **Service Focal Point** icon.
2. In the Contents area, select **Enable / Disable Error Data Collectio**n.
3. The Enable / Disable Error Data Collection window displays a list of machines. The window also lists each machine's error class, state, and type. Click the machine for which you want to enable or disable extended error data collection.
4. Click **Enable Selected** to enable extended for the selected managed system(s), or click **Disable Selected** to disable call home for the selected managed system(s).
5. Click **OK**.

2.5.5 Service Agent

Service Agent application operates on the HMC and monitors your pSeries server for hardware errors. It reports detected errors, assuming they meet certain criteria for criticality, to IBM for service with no customer intervention.

Note: Service Agent is *not* a replacement for the pSeries Maintenance Package. Use Service Agent as an additional service tool for the server.

IBM uses Service Agent to do the following:

- Analyze problems automatically
- Send service calls to IBM without customer intervention
- View hardware event logs

The Service Agent user interface defines machines and installs Service Agent plug-in. After machines are defined, they are registered with the IBM Service Agent Server (SAS). During the registration process, an electronic key is created, which becomes part of your resident Service Agent program. This key is used each time Service Agent places a call for service. The IBM Service Agent Server checks the current customer service status from the IBM entitlement database; if you are entitled for customer service, then the service call is accepted. If the current call shows that the Maintenance Agreement Expiration date is greater than the local key indicates, the key is extended to the new expiration date and you are sent a message indicating extension of service coverage. Service Agent provides early warning notification of upcoming Warranty or Maintenance Agreement expiration by sending renewal reminders at 90, 60, and 30 days prior to expiration. This feature is activated after you register with IBM. Service Agent reports some information to IBM to help with problem resolution. In some cases, this information may very likely be used by IBM for other purposes. This information consists of the problem or error information itself and Vital Product Data (VPD) or Inventory data. In the event the user is concerned about whether this information is sensitive, you can review the actual data that is being sent to IBM using either the Service Agent User Interface or from the command line using file display programs. If, after reviewing the data and determining you do not want Service Agent to send data, you can use either of the following methods to prevent data from going to IBM.

- Within Service Agent, turn off the VPD gathering feature. VPD is not therefore sent to IBM.

OR

- After registering, turn off the modem itself and configure the Service Agent Notification process to use e-mail to notify a help desk or have the help desk monitor Service Agent (in real time) using the Service Agent Alerts function.

When Service Agent detects an error, you can then call into IBM manually (instead of Service Agent Calling). The only data, besides error information, being sent to IBM is Vital Product Data (VPD), which is generated by either the **lscfg** command or Inventory Scout.

Refer to the "Electronic Service Agent for pSeries User's Guide" for specific information on setting up the modem and Service Agent specific registration data.

2.5.6 Redundant HMCs

While the HMC is considered a requirement for LPAR/NUMA systems, the system is designed not to put any more operational dependency on the HMC than is necessary. The HMC is primarily required to set up or change the LPAR/NUMA system configurations, and most other operations can be performed without it.

- The system will continue normal operation if the HMC fails or is removed, although this does not imply that this is a supported mode of operation for anything other than temporary conditions.
- Without an HMC, it is still possible to bring up a system in its last configured LPAR/NUMA state, including boot of defined partitions, by pressing the power buttons on the operator panel. However, the HMC is required to manually boot, reset, or "power off" individual partitions.
- For most hardware/firmware problems, an error code is still displayed on the physical operator panel, but the HMC would be required to display any error information written to a partition's virtual op panel.
- While HMC does offer a virtual console for each partition, an AIX/Linux console can also be assigned to a physical port on any async adapter in a partition, and telnet support through the network is also available.

If a customer so chooses, a redundant HMC can be configured for system and service management purposes. If a failure causes one of the HMCs to become unavailable, the customer can access the system through the alternate HMC. Note, that either HMC can be used if both are available. They are configured in a peer (as opposed to Master/Slave) relationship in that there is no limiting primary/backup relationship between them, and either can be used at any time.

Options for each HMC must be configured.

2.5.7 Scheduling Critical Console Data Backups

This option enables you to schedule the time and dates for backing up critical console Data.

Using your HMC, you can back up important data. Critical console data includes:

- User preference files
- HMC platform configuration files
- User and group information
- HMC log files

When you back up console data, you cannot choose which piece of the software you want to save. All pertinent HMC data is saved to the DVD-RAM when you back up critical console data. This function saves the data stored on your HMC hard disk and is critical to support HMC operations. Back up the HMC after you have made changes to the HMC or to the information associated with logical partitions. Use this task after customizing your logical partition definitions in any way. You can store a backup copy of hard disk information to your HMC following the repair or replacement of the disk.

To backup critical console data you must be a member of the following roles:

- Operator
- Advanced Operator
- System Administrator
- Service Representative

To backup critical console data, do the following:

1. In the navigation area, click the Software Maintenance icon.
2. In the Contents area, select **Backup Critical Console Dat**a.
3. After reading the dialogue, select **HMC Data** to save critical console information and **HMC Logs** to save log files. You can select either or both.
4. Select **Backup** to store your critical console data on the DVD-RAM disc
5. Click **OK**

2.6 Internal Battery Feature

The p690 system offers the capability to install “Internal Battery Features” (IBF) which act to hold up power to the system for a specified time duration during a loss of power. The IBF can be configured for redundancy. In addition, these batteries can be used in conjunction with an external UPS system to provide the power necessary to meet the customer’s requirements for maintaining system availability during sustained power outages.

3.0 Configuring the LPAR Environment for High Availability

This section of the white paper addresses configuration options when operating the p690 system in **L**ogical **Par**titioning (LPAR) mode. In many cases, this section will build upon or refer back to the SMP sections of the paper rather than repeat information.

3.1 Boot Options Utilizing SP Menus and Alternatives

The following SP menus available in SMP mode on the p690 system are modified or not enabled for the logical partitioning environment.

- *OS Surveillance Setup Menu*
  - *This function is disabled in LPAR mode. Status of each of the partition operating systems is performed on the HMC.*
- *Serial Port Snoop Menu*
  - *This menu is disabled in LPAR mode. Systems can be either "soft reset" or "hard reset" from the **System Management** function on the HSC by choosing the **Operating System Reset** menu.*
- *Use OS-Defined Restart Policy*
  - This policy applies to errors at the system (or *global*) level, not at the partition level.

Based on the above changes for LPAR mode, the following table represents the desired settings for configuring for High Availability utilizing the SP menus or the alternate methods indicated.

SP Menu	Option	Default Value	Recommended Setting	Alternate Invocation Method
Service Processor Setup Menu				
OS Surveillance Setup Menu	OS Surveillance Surveillance Time Interval Surveillance Delay	N/A	N/A	Performed via HMC operation in ***System Management*** *function*
Serial Port Snoop Setup	System Reset String Snoop Serial Port	N/A	N/A	Performed via HMC operation in ***System Management*** *function by choosing the* ***Operating System Reset*** *menu*
Scan Log Dump Policy Menu	Scan Dump Log Policy	1 (As needed)	1 (As needed)	
System Power Control Menu				
Enable/Disable Unattended Start Mode	Enable/Disable Unattended Start Mode	No - GA1 Yes - GA2 (April 26th, 2002) & beyond	Enabled	**SMS Menu**- Unattended Start Menu **Service AID** - Reboot/Restart Policy Setup
Reboot/Restart Policy Setup Menu				

SP Menu	Option	Default Value	Recommended Setting	Alternate Invocation Method
Reboot/Restart Policy Setup Menu ***In LPAR, this menu option works in conjunction with the OS defined restart policy as set in the partition with the "Service Authority" only***	Number of reboot attempts	1	1	**Service AID -** Configure Reboot Policy **SMIT -** Automatically Reboot After Crash ***(as determined by Service Authority partition settings)***
	Use OS-Defined restart policy ***(as determined by Service Authority partition settings)***	Yes - GA1 No - GA2 (April 26th, 2002) & beyond	No if field is available on Service Authority partition.	
	Enable supplemental restart policy	No - GA1 Yes - GA2 (April 26th, 2002) & beyond	Yes	
System Information Menu				
Processor Config/Deconfig Menu	Enable CPU Repeat Gard	Enable	Enable	
Memory Config/Deconfig Menu	Enable Memory Repeat Gard	Enable	Enable	

Refer to the "Boot Options Utilizing SP Menus and Alternatives" section on page 6 of the paper for descriptions on how invoke the recommended settings. Some of the options are only valid when configured in the Service Authority partition as stated in the above table. Otherwise, the options should be enabled in each operating system image in order to be effective for that LPAR image.

3.2 AIX Options

There is only one difference from the SMP defaults chart provided earlier. This difference is in the area of updating the System or Service Processor Flash. Besides this change represented in the table below, each OS partition must perform the recommended settings in the SMP table to have them effective for that LPAR image.

Option		Default Value	Recommended Setting	Invocation Method
Enable CPU Dynamic Deallocation	Enable/Disable CPU Dynamic Deallocation	Disable	Enable	**AIX cmd line Web SM SMIT -** CPU Gard
Hot-plug Task	PCI Hot-plug Manager	Used for Concurrent Maintenance		**Service AID -** Hot-plug Task
	SCSI Hot-swap Manager			
	RAID Hot-plug Devices			
Periodic	Disable or Enable	Enable	Enable	**Service AID -**

Option		Default Value	Recommended Setting	Invocation Method
Diagnostics	Automatic Error Log Analysis			Periodic Diagnostics **AIX Command Line**
Save or Restore Hardware Management Policies	Surveillance Policy and Reboot Policy	Used to save hardware and service processor settings		**Service AID** - Save or Restore Hardware Management Policies
Save/Restore Service Processor settings	Service Processor settings			**Service AID** - Save or Restore Service Processor settings
Update System or SP Flash	Command line entry	Used to update System or Service Processor Flash		**Service AID** - Update System or Service Processor Flash

3.2.1 AIX Defaults

Refer to the "AIX Options" section on page 16 for details on how to set the specified parameters and repeat for each desired OS partition.

3.2.2 Update System or SP Flash

If the system is running in LPAR mode, ask the customer or system administrator if a service partition has been designated. If the service partition is designated, the customer or system administrator shut down all of the other partitions. Perform the firmware update using the service aid or the AIX command line in that partition. If a service partition has not been designated, shut down the system. If the firmware update image is on backup diskettes, perform the firmware update from the service processor menus as a privileged user. If the firmware update image is in a file on the system, reboot the system in SMP mode and follow the normal firmware update procedures. If the system is already in SMP mode, follow the normal firmware update procedures.

Refer to section "**Update System or SP Flash**" on page 23 for AIX command or Service Aid to perform this function.

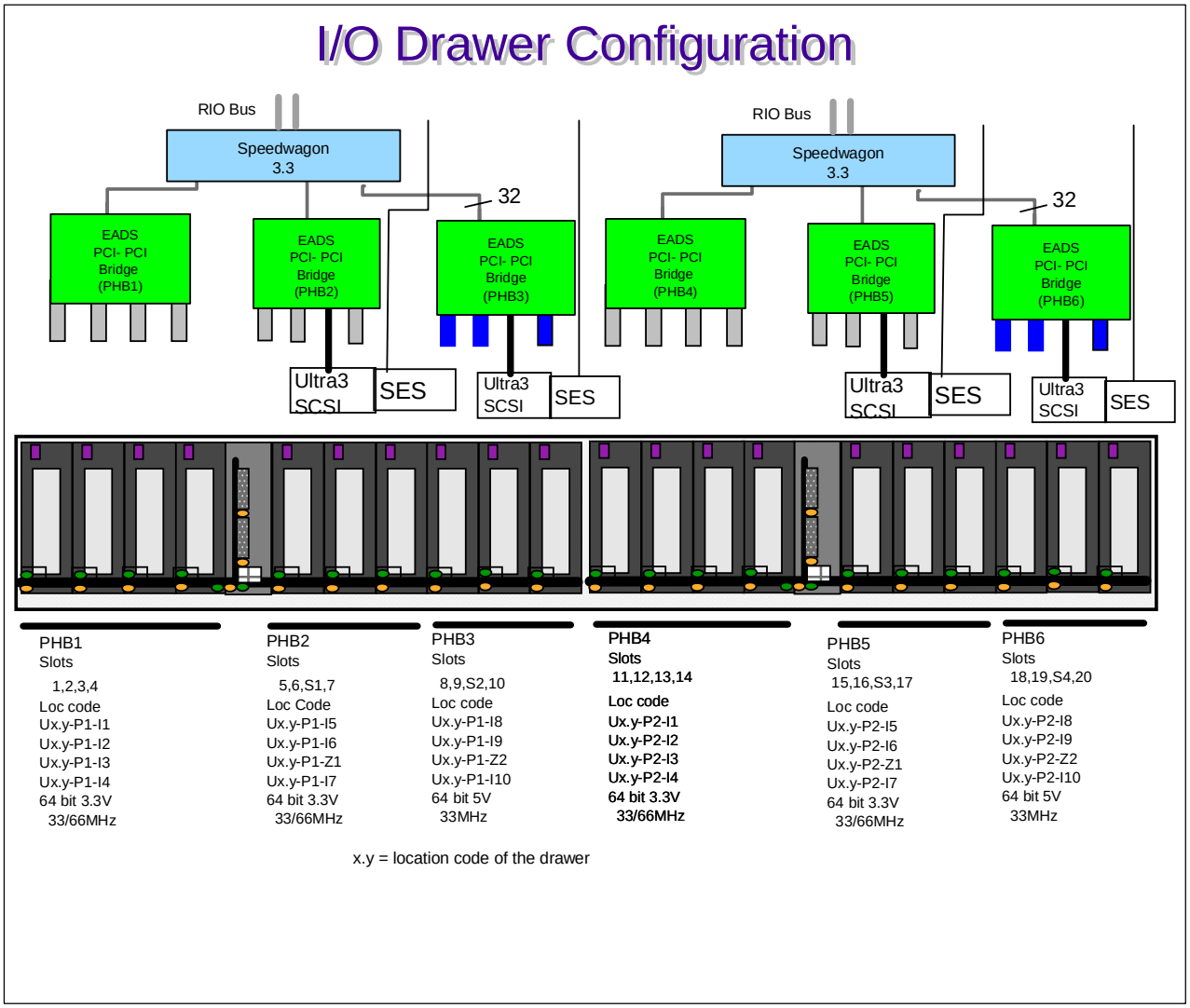
3.3 Disk Configuration

In LPAR mode, each partition must have a boot disk. Options exist to assign various levels of redundancy for the boot disk in each partition such as mirroring as explained in section "Boot Disk - Stand-Alone vs. Mirrored on page 25". For maximum availability, each partition should enable mirrored boots disks utilizing redundant SCSI adapters housed in separate I/O drawers as explained in the referenced section.

3.4 I/O Drawer and Adapter Configuration

3.4.1 Adapter Placement

The following diagram of the IO drawer will be used to help explain the suggested I/O drawer and adapter placement when configuring for high availability.



3.4.2 General Recommendations

- For larger LPARs, have LPARs use entire I/O planars which is half a B & C (10 slots plus 2 integrated SCSI)
- For smaller LPARs, have LPARs use entire PHBs (4 slots or 3 slots plus integrated SCSI)
- For AIX only,
  - If all adapters support EEH, then allow a mix of LPARs on one PHB
  - Linux has to have entire PHBs since DD are non-EEH

- Non-EEH I/O Adapters (IOA) on a PHB should be solely in one partition - do not split the PHB between partitions as a failure on a non-EEH adapter will affect all other adapters on that PHB
- For LPARs that have Non-EEH adapters, place all IOAs together (non-EEH and EEH) since a failure of a Non-EEH adapter will affect the whole partition

3.4.3 I/O Drawer Additions

The only additional concerns related to LPAR when adding I/O drawers over what has been stated in the SMP section is the proper allocation of these resources among the various partitions. The addition of more I/O drawers allows increasing capability for configuring for High Availability because of the increased capacity added to the configuration. Please refer to section "I/O Drawer Additions" on page 33 in conjunction with the above guidelines for assigning I/O adapters to partitions and the "PCI Adapter Placement Reference" for determining preferred configurations.

3.5 Service Enablement Through HMC Operations

3.5.1 Hardware Management Console

The HMC in LPAR mode works much the same way as it did in SMP mode except that now, from a service perspective, it acts as a focal point for consolidating and reporting errors from each of the LPAR operating system images. Therefore, it is extremely important that each Operating System image provide a network link to the HMC to insure that errors are reported and handled appropriately. This concept will be expanded on in section "Enabling Error Reporting to SFP Application" on page 46.

Service Enablement	HMC Operation	Default Value/Setting	Objective
Creating Service Rep ID	User Management	Service Representative	Create Service Rep. User ID for system service
Creating Software Support ID	User Management	hmcpe	Create userid for Software Support Personnel to analyze HMC problems
Establish HMC to OS partition network link for error reporting and management	System Configuration **"Customizing Network Settings"**	No network links established as default. Require network link for OS reported errors	Establish link between HMC and Operating System for service management Note: Should be performed for each OS image
Create "Service Authority" partition to allow install of fw updates & setting of system policies	**Partition Management Tasks**	No partitions defined as service partition	Establish one partition as having Service Authority
Enable systems for	**Service Focal Point**	Unknown	Set system enablement for call for service

automatic call home feature	**"Enable / Disable Call Home"**		when serviceable action is reached
Configure Service Agent to automatically call for service on error	**Service Agent** "Service Agent UI - Registration/Customization"	Not configured	Configure S.A. Parameters to allow automatic call for service on error Note: Should be performed for each OS image
Enable for gathering of extended error data	**Service Focal Point** "**Enable / Disable Extended Error Data Collection**"	Unknown	Enable systems to gather extended error data for service events Note: Should be performed for each OS image
Connecting Two HMCs to single system for redundancy	System Management Environment - Navigation Area	None	Configure redundant HMC in navigation area for viewing events
Schedule Critical Console Data Backups	**System Configuration** "**Scheduled Operations**"	None	Backup critical console data (including service data) on scheduled basis
Note: These actions need to be performed on both HMCs if redundant HMCs are installed.			

Most of the options in the table are described in the section "IBM Hardware Management Console for pseries (HMC)" on page 35 with the exception of Creating a Service Authority Partition which is described below. Note that certain options need to be configured for each operating system image in each of the LPARs as specified in the table.

3.5.1.1 Creating a Service Authority Partition

Certain maintenance functions (such as performing firmware updates and setting system level policies such as reboot/restart) can only be performed in a "Service Authority partition". To create partitions you must be a member of the System Administrator role. To create a partition, do the following:

1. Log in to the HMC.
2. Click on your managed system
3. Click on the partition management icon underneath the HMC hostname to select your preferred partition environment. The Contents area now lists the available managed systems. These systems are listed by default as "CEC001,"..."CEC002,"... and so on. If you have only one managed system, the Contents area lists the CEC as "CEC001."
4. Select the Managed System for which you want to configure partitions
5. With the Managed System selected in the Contents area, choose **Selected** from the menu.
6. Select **Power O**n.
7. Select **boot in partition standby** as a boot mode so that the managed system boots to partition standby mode.

8. Click **OK** to power on the managed system. In the Contents area, the managed system's state changes from *No Power* to *Initializing* . . . and then to *Read*y. When the state reads *Ready* and the Operator Panel Value reads *LPAR* . . . , continue with the next step.
9. Select the managed system in the Contents area.
10. Select **Create.**
11. Select **Partition**.

Note: You cannot create a partition without first creating a default profile.

The system automatically prompts you to begin creating a default partition profile for that partition. Do not click **OK** until you have assigned all the resources you want to your new partition.

12. Name the default partition profile that you are creating in the **General** tab in th**e Profile name** field. Use a unique name for each partition that you create, up to 31 characters long.
13. Click the **Memory** tab on the **Lpar Profile** panel. The HMC shows you the total amount of memory configured for use by the system, and prompts you to enter the number of *desired* and *required* memory. Enter the amount of desired and required memory in 1 gigabyte (GB) increments and 256 megabyte (MB) increments. You must have a minimum of one GB for each partition.

Note: The HMC shows the total number of installed, useable memory. It does not show how much memory the system is using at the time.

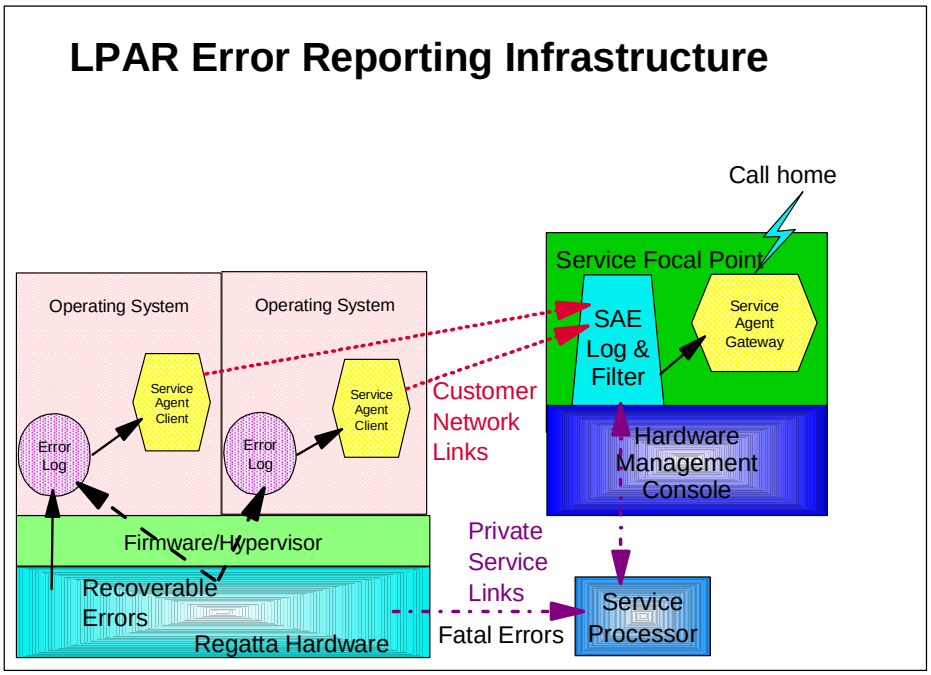
14. Click the **Processor** tab on th**e Lpar Profile** window. The HMC shows you the total number of processors available for use on the system, and prompts you to enter the numbers of processors you *desire* and *requir*e.
15. Click the **I/O** tab on the **Lpar Profile** window. The left side of the dialog displays the I/O drawers available and configured for use. The **I/O Drawers** field also displays the media drawer located on the managed system itself. This grouped I/O is called *Native* I/O and is identified with the prefix *CEC* as shown in the following example. Expand the I/O drawer tree by clicking the icon next to the drawer.
16. Click on the slot to learn more about the adapter installed in that slot. When you select a slot, the field underneath the I/O drawer tree lists the slot's class code (in the following example, *Token Ring controlle*r) and physical location code (*P2-I7*).

Note: Take note of the slot number when selecting a slot. The slots are not listed in sequential order.

17. Select the slot you want to assign to this partition profile and click **Ad**d. If you want to add another slot, repeat this process. Slots are added individually to the profile; you cannot add more than one slot at a time. Minimally, you should add a boot device to the **required** field.
18. Click the **Other** tab on the **Lpar Profile** window. This window allows you to set service authoritry and boot mode policies for this partition profile. Click the box next to **Set Service Authority** if you want this partition to be used by service technicians to perform system firmware updates.
19. Click the button next to the boot mode you want for this partition profile. The following example shows that this user selected service authority for this partition and wants to boot the partition to *Normal* mode.
20. **Note:** Be sure to review each tab of the Partition Properties dialog before clicking **OK** to ensure you have assigned all of the resources you need for this partition profile. Click **OK** when you are finished assigning resources in the Partition Properties dialog. The default profile appears underneath the Managed System tree in the Contents area.
21. Now that you have created a partition, you must install an operating system on it for it to function. To install an operating system on the partition, be sure you have the appropriate resources allocated to the partition you want to activate and refer to the installation information shipped with your operating system.

3.5.1.2 Enabling Error Reporting to SFP Application

The following diagram will be used to explain the error reporting paths from the system to the Service Focal Point application on the IBM hardware management console for pseries in LPAR mode.



The HMC requires a separate ethernet adapter for connection to a customer network to enable direct communication between the HMCs and the operating systems running in the partitions.
Refer to section “Establish HMC to OS partition network link for error reporting and management” on page 36 for a description of how to customize the network link between the HMC and the operating system images. This procedure should be performed for each partition.

3.5.1.3 Modem Configuration

Refer to the “Electronic Service Agent for pSeries User’s Guide” for specific information on setting up the modem and Service Agent specific registration data.

3.5.1.4 Service Agent Setup

Refer to the SMP section “**Service Agent**” on page 39 and to the “Electronic Service Agent for pSeries User’s Guide” for specific information on configuring Service Agent.

3.6 Standalone Diagnostic Considerations

3.6.1 LPAR Considerations for Placement of Regatta System Unit CD-ROM Drive

AIX standalone diagnostics is intended to be used on RS/6000® systems if no other option exists; online diagnostics is preferred due to its greater capabilities, but sometimes running standalone AIX diagnostics is the only viable option, such as when the system is unable to boot AIX (either because of a problem that prevents the system from booting or else because another operating system other than AIX is installed on the system).

Standalone diagnostics may be run either from diagnostics CD-ROM media (of the correct level for the machine you wish to run it on) or booting standalone diagnostics from a NIM server. The first part of this section deals with issues related to booting standalone diagnostics from CD-ROM media, specifically within a static LPAR system, optimizing the logical assignment of the CD-ROM drive (and associated devices attached to the same SCSI bus) so as to minimize the impact to users of the system should it become necessary to run standalone diagnostics on a given partition. The next part discusses NIM boot of standalone diagnostics.

NOTES:

1: Terminology: This information pertains to placement of CD-ROM or DVD-RAM drives (i.e. devices capable of reading an AIX or diagnostics CD-ROM) whose intended use includes booting standalone diagnostics from diagnostics CD-ROM media. This specifically excludes the use of the DVD-RAM drive in the HMC, which is not intended to be used to boot AIX diagnostic media.

2: Definition: a CD-ROM drive is said to be "parked" in a given partition if it is not immanently needed by any partition in the system, but rather assigned to a location where it might be moved to another partition if needed with minimum impact to the users on the system. Another definition would be to leave the CD-ROM drive on a partition where it is most likely to be used. If the CD-ROM drive is temporarily needed by another partition, it (and attached associated resources) may be logically assigned to the requesting partition, but because both the source partition containing the drive and the destination partition must be rebooted whenever the drive is reassigned then it is desirable to minimize the movement of the drive between partitions. In particular, one should avoid "parking" a drive in a partition where the need to reboot that partition in an untimely fashion would negatively impact users of that partition; once the need for the drive in that partition is through (for instance, after running standalone AIX diagnostics on that partition) the drive should be moved back to another partition where it is intended to be "parked", this way the reboot on the partition may be done at a time when it has the least impact to users on both source and destination partitions, rather than having to move the drive at an inopportune time when the destination partition desperately needs to have the CD-ROM drive assigned to it.

General recommendations to maximize availability (minimize the need to reboot partitions) when running standalone diagnostics from CD-ROM:

If the managed system contains more than one CD-ROM drive, if possible, each drive should be physically connected to a SCSI controller associated with a different PCI "slot" (some SCSI adapters consist of 2 different SCSI controllers, both associated with the same PCI "slot" (even if the adapter is

integrated onto the system I/O planar, it can be thought of as occupying a PCI slot). If possible, the controllers in these "slots" should be assigned to different LPAR partitions.

Maximizing the number of CD-ROM drives in the system to the highest number supported, or such that one CD-ROM drive is available in as many partitions as possible (to a maximum of one CD-ROM drive per partition) is desirable to minimize the need for rebooting partitions in order to move a drive from one partition to another.

If the system contains a partition with service authority, ideally, the CD-ROM drive should be "parked" at that partition (or if the system has more than one drive, one drive should be "parked" at the service partition (assuming that the service partition may be more readily rebooted than other partitions, there would be less impact than if the CD-ROM was "parked" on a partition where rebooting the partition would be less desirable.

If devices that require processing supplemental media are present on a system, if possible, they should be in the same partition. The diskette drive should be located in the same partition as the CD-ROM drive that one wants to run standalone diagnostics from.

Standalone diagnostics boot from CD-ROM requires a functioning CD-ROM drive, associated SCSI controller and PCI bus, plus functioning path between PCI bus and the CEC and memory.

3.6.2 Standalone Diagnostics: NIM vs CD-ROM

If possible, to avoid having to reboot partitions when moving the CD-ROM drive between partitions, it is desirable to have one partition (or another system external to the LPAR system) set up as a NIM server. This can allow the partitions to boot standalone diagnostics from the NIM server instead of from the CD-ROM drive (and as an additional benefit, can help administer installation of AIX on partitions from the NIM server). This requires a network adapter tied to the same network as the NIM server be present in each partition in which you want to be able to do a NIM boot of standalone diagnostics on.

Instead of having to move a CD-ROM drive between source and destination partitions to be able to run standalone diagnostics from CD-ROM, a NIM boot of standalone diagnostics only requires a reboot of the partition that one wants to run standalone diagnostics on.

However, the system administrator must be able to properly configure the NIM server, making sure it is up to date with the proper images necessary to support its client partitions. Also, instead of supplemental diskettes for adapters that otherwise require them in CD-ROM standalone diagnostics, the NIM server image must be loaded with support for all the devices that are installed in the client partitions. Because setup of the NIM server and client is not trivial, it is highly recommended that once it is set up, the administrator should attempt to do a NIM boot of standalone diagnostics from each partition that might later rely on it working so that any setup problems might be debugged first on a working system. Also, NIM boot of standalone diagnostics requires a functioning network adapter in the partition, as well as associated PCI bus and path to the CEC and system memory. However, note that this can give the service person an optional way to load standalone diagnostics (such as if the CD-ROM drive or SCSI adapter or associated PCI logic is not functioning in a given partition).

4.0 HACMP

High Availability Cluster Multiprocessing allows the configuration of a cluster of pSeries systems which provides highly available disk, network, and application resources. These systems may be comprised of SMP systems or LPAR based systems. The clusters may be comprised of the following configurations:

- Multiple SMP systems
- One or more SMP system and one or more Operating System images on an LPAR system/s
- LPAR OS images across multiple LPAR systems
  - One or more LPAR images from one system coupled to one or more images from another LPAR system
- LPAR OS images coupled on the same LPAR system
  - Multiple OS images coupled on same HW platform
    - May be susceptible to HW single points of failure
    - Recommended for protection from OS or application faults only

4.1 Uni/SMP Mode

This mode of operation is the typical HACMP environment where multiple Uni or SMP systems are clustered together to eliminate single points of failure. Refer to the "HACMP for AIX Installation Guide" for configurations and operation.

4.2 LPAR Mode

4.2.1 Clusters of LPAR Images Coupled to Uni/SMP systems

This mode of operation supports the clustering of an LPAR image with standalone Uni or SMP systems. In this mode of operation, the partition acts as just another node in the managed cluster and is designed to protect from single points of failure as in the base case.

4.2.2 Clusters of LPAR Images Across LPAR Systems

In this mode of operation, the cluster is comprised of an image from one LPAR system clustered to another image from one or more other LPAR systems. Again, this mode of operation is designed to protect against single points of failure.

4.2.3 HACMP Clusters within One LPAR System

Many customers require their production systems to be highly available, with minimal unplanned system outages. A pervasive opinion is that increased availability can be achieved by partitioning servers, even to the point of claiming that a single partition failure is not significant because the system keeps running. We agree that partitioning provides the ability to perform software upgrades while continuing to run applications in a separate partition, and provides greater isolation for multiple applications that previously ran in the same operating system instance. However, one should not utilize separate partitions in the same system to provide production level high availability failover capability. While a standby partition (not necessarily idle) would help with a software fault or even software failover testing, it will not provide full hardware isolation or hardware failover capabilities.

If high availability is critical, then providing just a partial solution is usually not sufficient. In such situations, we recommend the implementation of high availability failover capability between partitions in separate servers. Small partitions may only require a relatively small separate partition or system.

5.0 Conclusions

The ideas presented forth in this paper are intended to help the system administrator or I/S manager plan to configure the p690 for higher levels of availability. Each suggestion should be evaluated against the ultimate availability objective to determine the benefit obtained versus the potential additional cost incurred by implementing the recommendation.

As stated in the beginning of this document, there are many factors outside the scope of this white paper which influence the ultimate availability achieved by your specific configuration. Topics such as the training of the operations and support personnel, the physical environment and surroundings where the system resides, the operations policies and procedures, etc. are a few examples of these.

The procedures recommended in this white paper coupled with the inherent reliability of the p690 system and its robust RAS features and functions should help facilitate configuring this system to meet your availability requirements for years to come.

Additional assistance with performing availability analyses for your specific environment may be contracted through IBM's Global Services division or obtained through your marketing representative.

IBM Corporation
Marketing Communications
Server Group
Route 100
Somers, NY 10589

Produced in the United States of America
04-02

More details on IBM UNIX hardware, software and solutions may be found at: ibm.com/servers/eserver/pseries.

You can find notices, including applicable legal information, trademark attribution, and notes on benchmark and performance at www.ibm.com/servers/eserver/pseries/hardware/specnote.html

IBM may not offer the products, programs, services or features discussed herein in other countries, and the information may be subject to change without notice.

General availability may vary by geography.

IBM hardware products are manufactured from new parts, or new and used parts. Regardless, our warranty terms apply.

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

Any performance data contained in this document was determined in a controlled environment. Results obtained in other operating environments may vary significantly.