IBM *@*server Clustered Computing

White Paper

May, 2001

Preface

The new e-business environment creates the need for a computing infrastructure comprised of servers that optimize three classes of workloads:

- data transaction servers to manage core business processes,
- Web application servers to manage end-user experience,
- and appliance servers to manage specific network functions.

Computing resources in this diverse computing environment must be managed to effectively balance complex workloads as well as deliver the required quality of service (QoS), capacity on demand (CoD), security, scalability and manageability. This aggregation, or *clustered computing,* will include servers, operating systems, storage, networks, applications, racks and stacks.

This whitepaper gives you an good overview of cluster ready software for IBM @server pSeries cluster server: traditional RS/6000 SP, Blue Hammer and Blue Hammer Jr., and loose cluster.

Authors

Verena Gschell
Web Server Pre Sales Technical Support, Germany
Volker Haug
Web Server Pre Sales Technical Support, Germany
Stefan Kister
Web Server Pre Sales Technical Support, Germany
Gregor Linzmeier
Web Server Pre Sales Technical Support, Germany
Justus Reich
Web Server Pre Sales Technical Support, Germany
Martin Springer
Web Server Pre Sales Technical Support, Germany
Nurcan Tezulas
Web Server Pre Sales Technical Support, Germany

AIX is Cluster Ready

AIX supports Quality of Service QoS models defined by the industry Internet Engineering Task Force (IETF) as part of its base infrastructure. It supports integrated Services and Differentiated Services in the kernel, enabling network policy management of general IP traffic. AIX has many security attributes important for e-business data privacy and secure transactions. This includes the Trusted Computing Base (TCB) for the enforcement of security policies, IBM Network Authentication Services (which includes Kerberos), and IBM Web-based System Manager Security. As a leading, state-of-the-art UNIX OS, AIX is designed with core attributes essential for clustered computing. This cluster software has been demonstrated in major Web serving events such as the Olympics games at Nagano, Atlanta and Sydney, proving that AIX is a Web-ready UNIX.

AIX 5L provides a more simplified cluster management interface, with the integration of selected PSSP (cluster software honed on the RS/6000 SP supercomputer) capabilities including resource monitoring and automated operations. AIX is "cluster ready" out of the box with the same look and feel for single or multi-system management.

Cluster Ready Software Building Blocks

Beside the cluster ready operating system IBM also delivers cluster ready software building blocks, such as:

PSSP - Simplified multi-system management

- Multi-system installation and management
- Remote management capability
- Hardware and performance monitoring
- Distributed, secure parallel commands
- Available for SP, Blue Hammer and Blue Hammer Jr.

GPFS - Cluster file system

- Fast, global parallel file system for BI, server consolidation and technical computing workloads
- Built-in fall-over features
- DMAPI for use with data management applications (e. G. Tivoli Storage Manager)
- Available for SP, Blue Hammer and Blue Hammer Jr.

HACMP - High availability for business critical workloads

- Support for up to 32 nodes
- Data access and backup plans to help optimize application execution and scalability
- Application recovery/restart
- Packaged HA clusters: HA-p640, HA-p660, HA-680, HA-F80, HA-H80, HA-M80, HA-S80

- Cluster Proven Programm - over 20 applications validated in HA environment, plan 15 more in 2001

HAGEO and GeoRM - Disaster Recovery

- HAGEO - remote site disaster survivability with real-time, automated fall-over and reintegration
- GeoRM - remote location disaster survivability of real-time data, and reintegration of data
- No distance limitation
- Rated No. 1 Disaster Recovery solution by DH Brown

Traditional SP

The SP system is basically a group of RS/6000 or pSeries servers working together under one point of control, the Control Workstation (CWS). It manages the whole group of RS/6000 or pSeries nodes as a unique system. The CWS has two connections to each node, an SP-Ethernet network for the system administration and a RS-232-connection for hardware control. Three different types of SP systems are possible:

- The classic SP with the 375 MHz POWER3-II thin, wide and high SP nodes, which only fit in a special rack, the SP frame These systems can scale up to 128 nodes.
- SP system with SP nodes in the SP frame with SP-attached servers. Up to 16 pSeries 680 and RS/6000 S80 servers, or up to 32 midrange RS/6000 M80 and H80 or pSeries 660 rack-mounted servers can be integrated in these SP-clusters.
- Clustered Enterprise Servers, a cluster of non-SP building blocks (RS/6000 S80, H80, M80, pSeries 660 and 680), which has the same building rules and limitations as the SP attachement of these servers.

Nodes or attached enterprise servers can be interconnected by an optional high bandwidth, low-latency switch for high-performance, internode communications. The following switch options are available:

- The SP Switch offers a one-way bandwidth of up to 150MB per second between nodes (300MB bidirectional3) for interconnecting all node types, including PowerPC 332 MHz thin and wide nodes, POWER3 thin, wide and high nodes, S-family enterprise servers and the SP Switch Router.
- The SP Switch2 provides a one-way bandwidth of up to 500MB per second between nodes (1GB bidirectional3) for interconnecting POWER3 high nodes.

The SP Cluster Software Parallel System Support Programs for AIX (PSSP) is a collection of administrative and operational software applications which run on each node of an SP system and on the CWS. Built upon the system management tools and commands of the AIX® Version 4 operating system, PSSP enables system administrators and operators to better manage SP systems and their environments.

Sets of software tools and related utilities, including application programming interfaces (APIs), have been grouped together to offer easier administration of installation, configuration, device management, security administration, error logging, system recovery and resource accounting in the SP environment.

The PSSP system administration and operation component packages together the tools required for SP administrative functions. These include entering and changing configuration information such as:

- The System Data Repository (SDR) is a central repository that contains the specific SP-configuration information and operational information. It only resides on the CWS, where the configuration information is then distribured to the nodes
- System command execution for executing system commands in parallel
- Parallel system management tools and commands for enabling concurrent parallel performance of system management functions across multiple SP nodes
- File collections for managing duplicated files and directories on multiple nodes
- Login control for blocking unauthorized user or group access to a specific SP node or a set of nodes
- Consolidated accounting for centralizing records at the node level (for tracking use by wall clock time rather than processor time) and gathering statistics on parallel jobs.

A consolidated system graphical user interface, RS/6000 SP Perspectives, provides a common launch pad for PSSP system management applications through direct manipulation of system objects represented as icons. This interface is tightly integrated with the problem management infrastructure. It allows users to easily create and monitor system events and provide notification when events occur. The interface is highly scalable for large systems, and can be easily customized to accommodate varying environments.

To significantly improve system availability, PSSP also contains functions and interfaces to other products that can help reduce unplanned outages and minimize the impact of outages that do occur. These include:

- System partitioning, which makes it possible to create a separate logical system partition to test software changes in a non-disruptive manner.
- Installation and migration coexistence that can reduce scheduled maintenance time. The installation process has been restructured to improve verification between installation steps. Administrators can now more effectively utilize the Network Installation Manager (NIM) functions in AIX. PSSP can coexist with up to three previous releases within an SP partition, allowing easier migration to new software levels
- Node isolation, which removes an SP node from active duty and enables it to be reintegrated without causing an SP Switch fault or disrupting switch traffic. This isolation is useful for correcting an error condition or installing new hardware and software without impacting production.
- High Availability Control Workstation (HACWS) connects two RS/6000 workstations (with HACMP installed) to an SP system to provide a backup control workstation in the event the primary one becomes unavailable (only one is active at any time). A twin- tailed disk configuration, along with the IP address takeover afforded by HACMP, enables rapid switchover to the backup control workstation,

with little or no impact on operational access to the SP system or System Data Repository.

The PSSP's RS/6000 Cluster Technology (RSCT) is a collection of services that define hardware and software resources, node relationships and coordinated actions to manage groups of nodes of an SP:

- Event Management monitors hardware and software resources in the SP and notifies an interested application or subsystem when the state of the resource changes. Resources on any node in the SP can be monitored from any other node. Topology Services define the relationships between nodes in a cluster in order to allow seamless takeover of functions in the event of a node failure.
- Group Services provides a set of interfaces which enable a distributed subsystem, such as General Parallel File System (GPFS), to synchronize recovery actions among the processes making up the subsystem.
- The Communication Subsystems Support component contains SP Switch adapter diagnostics, switch initialization and fault-handling software, device driver and configuration methods (config/unconfig), plus parallel communications APIs.

The Virtual Shared Disk (VSD) component is an API that creates logical disk volumes for parallel application access of a real disk device. These can be attached locally or on another SP node. This feature can enhance the performance of applications that provide concurrency control for data integrity, such as Oracle databases.

The IBM Recoverable Virtual Shared Disk function provides recovery from failures of virtual shared disk server nodes, and takes advantage of the availability services provided by PSSP to determine which nodes are up and operational.

PSSP supports a multi-threaded, standards-compliant Message Passing Interface (MPI) via IBM Parallel Environment for AIX (PE), as well as maintaining its single-threaded MPI support. In addition, PSSP includes a Low-level Application Programming Interface (LAPI) with a flexible, active message style, communications programming model on the SP Switch.

The Performance Toolbox Parallel Extensions (PTPE) function of PSSP collects and provides performance data for SP hardware and software through enhancements to the Performance Toolbox for AIX (PTX) product, the preferred performance monitor for AIX systems. It allows PTX to monitor unique SP subsystems, such as VSD, SP Switch and LoadLeveler®

PTPE organizes the SP into a set of performance reporting groups with coordinating managers, and distributes the burden of monitoring nodes throughout the SP system, thus eliminating the need for a dedicated monitoring node. PTPE also provides average performance statistics for the SP system, rather than monitoring every data point on all SP nodes. This can help reduce the computational effort required for run-time monitoring of SP performance.

Offering convenient services to help ease administrative burdens, PSSP also includes the following publicly available software packages:

- Network Time Protocol (NTP) for clock synchronization across SP nodes
- Perl programming language for developing system- wide shell scripts
- Software Update Protocol (SUP) for installing software from the boot file server
- Kerberos IV security for authentication of the execution of remote commands
- Trusted C-shell Language (Tcl) for controlling and extending applications

Blue Hammer and Blue Hammer Jr.

Since the announcement of PSSP 3.2 the SP cluster concept has been extended to area of enterprise server. With PSSP 3.2 the 7017 enterprise server models can be clustered like an SP environment, however an SP Frame is not mandatory anymore (code name blue hammer).
With the new announcement in April 2001 this concept of clustered enterprise server (CES) has been extended to the midrange enterprise server models RS/6000 H80, M80 and pSeries 660 Model 6H1 (code name blue hammer jr.).

In the following sections we want to point out the advantages and also disadvantages of this new IBM cluster strategy.

The SP was originally designed as a supercomputer, but with its outstanding capabilities the SP is also able to build up clusters for commercial requirements. This was the basis for the great success of the SP in scientific and technical computing as well as in commercial environements. With blue hammer and blue hammer jr. most of the SP capabilities are utilized by the enterprise servers.

One of the most important points is the central point of management by a control workstation (CWS). The CWS is responsible for installation and software maintenance of all cluster nodes. All information on the nodes are stored in a central database called SDR. Further more the CWS is the central point of hardware (supervisor environment) and software control (RSCT functionality), a functionality which can be regarded as the build-in availability of an SP/CES cluster environment. With this appoach a higher quality and more cost efficiency of system service are achieved.

With the introduction of the enterprise server into the SP cluster concept an extremely high scalability is possible, because now the horizontal scalability of the SP is combined with the vertical scalability of the enterprise server.

The blue hammer concept also contents the idea of cluster building blocks, because the approved SP technology is no longer related to specific SP hardware. We distinguish cluster hardware and software building blocks.

First this approach increased the variety of clustering possibilities due to the the higher number of cluster nodes (hardware building blocks). We can build up traditional SP (SP Switch or SP Switch2 optional) or SP with SP-attached enterprise server (SP Switch optional, SP Switch2 in 4Q01) as well as pure clustered enterprise server environments. The high speed interconnect SP Switch and SP Switch2 will be available for CES environments at the end of 4Q01 (statement of direction) which will enhance this variety of clustering further. At the moment only the SP Switch is supported and an SP Frame with the SP Switch is mandatory to build up a switched CES environment (SP-attached server).

Second approved specific SP software can be introduced as software building blocks into the area of enterprise server. For example the general parallel file system (GPFS) and the parallel environment (PE) which are very useful in parallel environments such as parallel databases or parallel computing.

The approach of multiple cluster building blocks achieves a high flexibility to adjust our enterprise server solutions to the customers requirements. A very exact sizing to the customers specific workload is possible and due to the extreme scalability of the SP cluster concept the customer can trust to cope increasing workload in future.

The blue hammer concept is the first step into IBM´s strategy on flexible server. The SP cluster concept will assure that the new POWER4 technology (Regatta H) can be integrated into current cluster production environments. Although the new hardware capabilities of Regatta H (LPAR) are controlled by a new console (hardware system console (HSC)) the cluster management is still based on the SP cluster concept. This is a great assurance on investment protection of skills, hardware and applications on today´s SP and clustered enterprise server environments.

However, with the cluster strategy IBM resigns the development of the popular dedicated SP nodes. This high density of servers (up to 16 nodes in one 19 inch frame) cannot be achieved with the enterprise server that are today supported in an SP/CES cluster environment. Of course, it must be considered that up to three midrange server (H80, M80, 6H1) with minimum configuration can be configured into one T42 rack, and the high end enterprise server 7017 models can be configured in a three rack solution (two CEC´s and one I/O Rack with a primary I/O drawer for each CEC).

With the blue hammer jr. announcement IBM also puts its new AIX version 5.1 on the market which is the first AIX 5L version for production environments. However, this new AIX version is not yet supported with the latest version PSSP 3.2, thus the new features and capabilities of AIX 5L cannot not yet be utilized in an SP/CES cluster environment. This support will not be achieved untill the new PSSP version 3.4 will be available in 4Q01.

Of course, it must be considered that all the advantages of an CES environement manageability, scalability, flexibility and performance are connected with extra costs for the control workstation, PSSP licences and if necessary special adapters.

<u>Technical aspects for configuring an CES environment:</u>

⇨ At a maximum up to 32 enterprise server can be configured in a CES environment, thereby only up to 16 can be a high end 7017 enterprise server.

⇨ Software requirements are AIX 4.3.3 or later and PSSP 3.2 or later, for the new CES nodes H80, M80 and 6H1 you need at least AIX 4.3.3 with maintenance level 08 and PSSP 3.2 with APAR IY16350.

⇨ In a pure CES environment the SP Switch as well as the SP Switch2 is not supported yet. For a switched environment with enterprise server you need at least an SP Frame with an SP Switch. Each enterprise node in a switched environment needs an SP Attachment PCI Adapter additionally. The SP Switch2 is not yet supported in this environment either.

⇨ Each 7017 entersprise server has two serial line connections to the control workstation which are delivered with the configuration as an SP/CES node.

⇨ Each midrange enterprise server needs a SAMI Internal Attachment Card (F/C 3154) and a SAMI cable (F/C 3151) as a serial connection interface to the CWS. For the SAMI Internal Attachment Card PCI slot 7 is recommended.

⇨ Each node must be connected to an adminstrative ethernet via the interface en0 (the lowest-numbered Ethernet bus slot in the primary I/O drawer).

⇨ There is only limited support for HACWS solutions because of the lack of stand-by serial line connections in an CES environment.

⇨ The RS/6000 Model F80 is the recommended control workstation for an CES environment.

⇨ Further reading:

→ Manual: *RS/6000 SP Planning Volume, Hardware and Physical Environment (GA22-7280-09)*

→ Redbook: *RS/6000 SP: The Way to Universal Clustering (SG24-5374-01)*

Service Processor Overview

The following section gives you an overview about the Service Processor used in the different RS/6000 or IBM @server pSeries models and how you can use the Service Processor to control several systems without using the RS/6000 SP concept “of a control workstation together with the PSSP software” or “of a Clustered Enterprise Server”.
Both concepts require that you have to use a control workstation connected to each other system over a serial cable and a Ethernet cable. The concepts also require the use of the Parallel System Support Program (PSSP) software.

This document describes how you can control several systems without using the SP control workstation and PSSP.
However, PSSP provides a lot of tools to manage cluster environments as well as a graphical user interface (GUI) to do that.

Service Processor Tasks
A Service Processor is a small hardware feature which is usually placed directly on the system planar. It monitors and diagnostics online the hardware feature of your RS/6000 or IBM @server pSeries system. For example all the fans are sensed by the Service Processor. In case of a problem the Service Processor controls the remaining fans to run at a higher speed to guarantee the cooling of a system and sends out a message to the system console.
The Service Processor also monitors the AIX operating system for software hangs. This is accomplished by heartbeats to check if the AIX operating system is still alive and running.

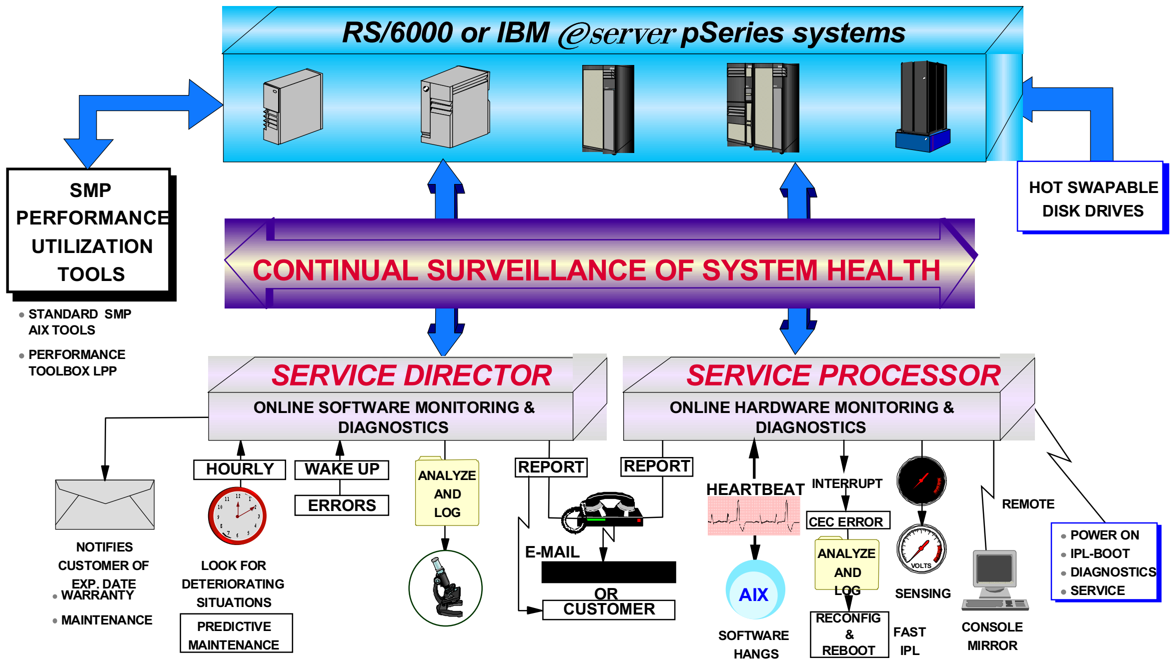
Another aspect of the Service Processor is the “call home” feature which works only with an appropriate service center infrastructure. A server in the service center which has a modem attached waits for incoming calls and forwards them to the support people. They try to solve a problem with remote diagnostics, and if necessary with hardware replacements. Setup and infrastructure for “call home” is not covered in this paper. Please contact IBM Global Services (ITS) for their offerings, your IBM sales representative or your IBM business partner.

There is also a service offering from IBM Global Services (ITS) available which is called Service Director.
Service Director is software which has to be installed on top of the AIX operating system and it monitors and diagnostics online the AIX operating system running on your RS/6000 or IBM @server pSeries system. Service Director cannot maintain your system hardware.

The following picture gives you and overview about the differences between the Service Processor and the Service Director and how an overall high availability can be accomplished with RS/6000 or IBM @server pSeries systems:

System Surveillance Capabilities

... KEEPING BUSINESS-CRITICAL APPLICATIONS OPERATIONAL



RS/6000 or IBM @server pSeries systems

SMP PERFORMANCE UTILIZATION TOOLS

- STANDARD SMP AIX TOOLS
- PERFORMANCE TOOLBOX LPP

HOT SWAPABLE DISK DRIVES

CONTINUAL SURVEILLANCE OF SYSTEM HEALTH

SERVICE DIRECTOR

ONLINE SOFTWARE MONITORING & DIAGNOSTICS

NOTIFIES CUSTOMER OF EXP. DATE

- WARRANTY
- MAINTENANCE

HOURLY

LOOK FOR DETERIORATING SITUATIONS

PREDICTIVE MAINTENANCE

WAKE UP

ERRORS

ANALYZE AND LOG

REPORT

REPORT

E-MAIL

OR

CUSTOMER

SERVICE PROCESSOR

ONLINE HARDWARE MONITORING & DIAGNOSTICS

HEARTBEAT

AIX

SOFTWARE HANGS

INTERRUPT

CEC ERROR

ANALYZE AND LOG

RECONFIG & REBOOT

SENSING

FAST IPL

REMOTE

CONSOLE MIRROR

- POWER ON
- IPL-BOOT
- DIAGNOSTICS
- SERVICE

How to access a Service Processor

Service Processor menus can be accessed locally by connecting an ASCII terminal or a Palm device (PDA) with a vt100 emulation to either serial port. Because the presence of the ASCII terminal cannot be confirmed by the Service Processor, you must press a key on the ASCII terminal to confirm its presence. The Service Processor then prompts you for a password (if
set), and when verified, displays the Service Processor menus.
The Service Processor menus can alternative be accessed remotely by connecting a modem to
serial ports of your server.

Service Processor menus are available locally or remotely when the server is turned off and still connected to a power cable or the Service Processor is operating in standby mode. A subset of the Service Processor menus are also available when server power is on and the AIX operating system is running. Some functionalities can be customized by using the Diagnostics Service Aid under AIX. Invoke the command *diag* from the AIX command line to use these utilities.

For security reasons you can set a password so that a remote user can have only access to a subset of all available options. This is called the General User Menu and can look like the following (depending of the RS/6000 or IBM @server pSeries model):

```
        GENERAL USER MENU

 1. Power-On System
 2. Read VPD Image from Last System Boot
 3. Read Progress Indicators from Last System Boot
 4. Read Service Processor Error Logs
 5. Read System POST Errors
 6. View System Environmental Conditions
99. Exit from Menus

1>
```

Note:
The Service Processor prompt reads either 1> or 2> to indicate which serial port on the system unit is being used to communicate with the Service Processor.
The main menu which can be protected by a privileged password looks like the following (depending of the RS/6000 or IBM @server pSeries model):

```
                     Service Processor Firmware
                     Firmware Level: sh990707
                     Copyright 1997, IBM Corporation
                     SYSTEM NAME

             MAIN MENU

   1. Service Processor Setup Menu
   2. System Power Control Menu
   3. System Information Menu
   4. Language Selection Menu
   5. Call-In/Call-Out Setup Menu
   6. Set System Name
  99. Exit from Menus

1>
```

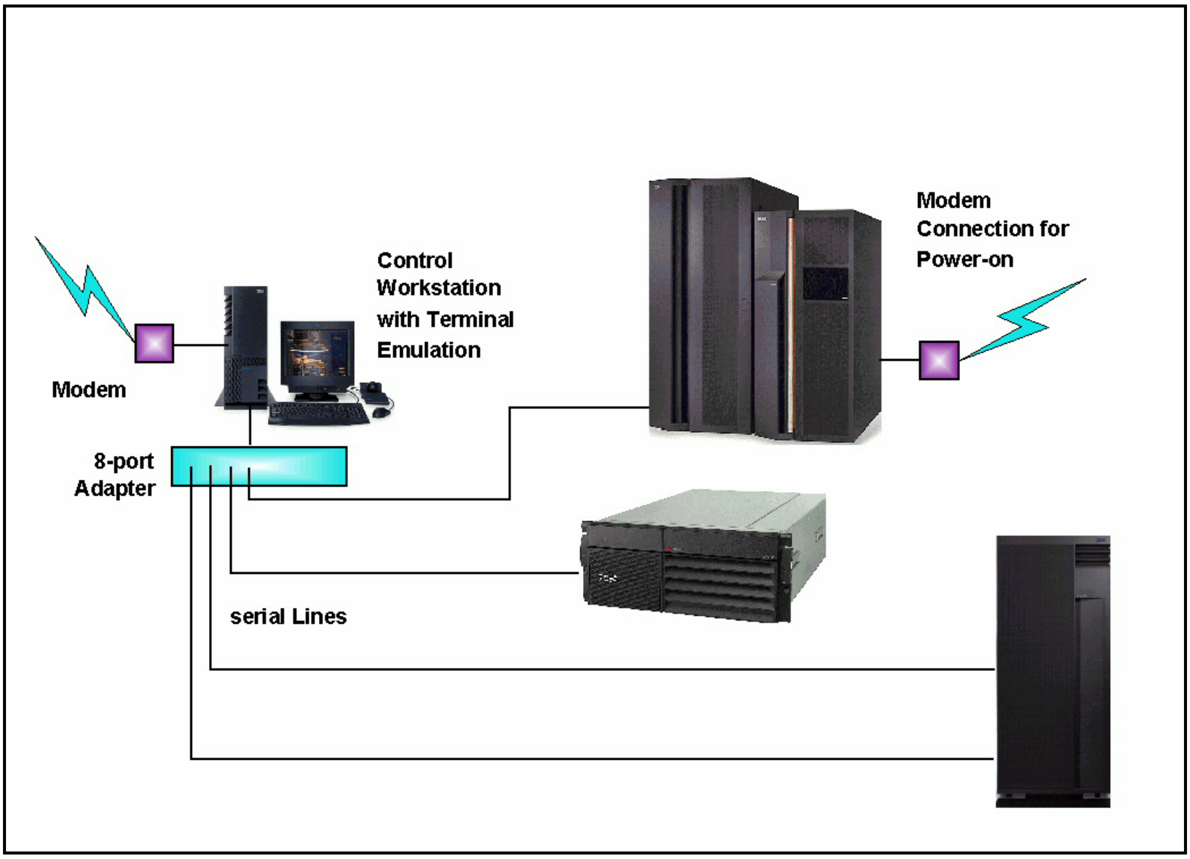
For a detailed description of all different sub menus and functionalities please refer to the Users Guide or Service Guide of your appropriate RS/6000 or IBM @server pSeries system.
You can find the most actual version on the web:
http://www.rs6000.ibm.com/resource/hardware_docs/index.html

Clustering of several systems
The main benefits of the Service Processor to manage several servers without using a Clustered Enterprise Server concept are remote power-on and diagnostics after a

system crash or hardware failure. In the following section you will find detailed hardware configuration examples and a guideline for some typical configurations in AIX.

Configuration Example for remote power-on and remote diagnostics.

The following picture shows a typical remote management environment, which consist of system we use as a control workstation (not to confound with the Control Workstation in a RS/6000 SP environment) and others servers. The control workstation is connected with serial lines (8-wire cabling) to the service processors of each server.


Modem
Control Workstation with Terminal Emulation
Modem Connection for Power-on
8-port Adapter
serial Lines

Control Workstation

A typical control workstation configuration for remote management using the Service Processor is very similar to the RS/6000 SP Control Workstation. You can use any RS/6000 or IBM @server pSeries workstation or entry server with graphical display, mouse and keyboard. You need to run the AIX 4.3.3 or 5.1 operating system, but you don't need to run PSSP.

In the section "Learning from the SP concept" you will find ideas how you can enhance this control workstation with SP like features. The prerequisite is a dedicated control LAN, which should be considered in the hardware planning.

Like the SP Control Workstation you need an 8-port or an 128-port adapter (FC 2943 or FC 2944) in order to increase the number of serial ports.
After connecting the control workstation with the serial port of each server, you can access the Service Processor using the AIX commands *ate* or *cu*.

Here is a configuration example:

Configuration example for *cu*:
Define tty1 using all the defaults, especially “Login disable”
Add the following line in

```
/etc/uucp/Devices:
Direct tty1 – 9600 direct
```

Access the Service Processor connected to tty1 with

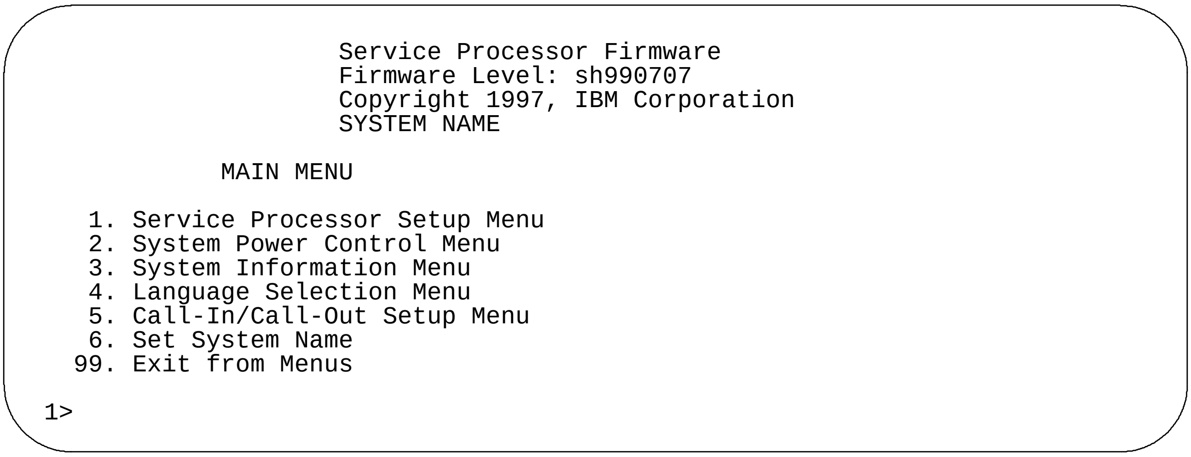
```
cu –ml tty1
```

Terminate with
~. (tilde dot)

Due to cable length limitations of serial cables the control workstation has to be in the same area like the RS/6000 or IBM @server pSeries servers. If needed there a several possibilities to extend serial lines, e.g. serial-to-optical or serial-to-ethernet converters.

If the administrator prefers to manage the cluster from remote, he can login through a TCP/IP connection to the control workstation and issue the *cu* command in his *telnet* session.
A alternative for this scenario is the use of a Terminal Server instead of an RS/6000 or IBM @server pSeries as the control workstation. A Terminal Server provides a TCP/IP connection to serial or graphical (input/output) devices. In our case a Terminal Server is connected to the customers’ LAN and via serial connections to all service processors.

How to power-on a server remotely and use a specific boot mode
First you have to open a serial connection to the Service Processor from your control workstation to your remote system by using the cu -ml tty1 command. You will see the main menu of the Service Processor of your remote system (depending of the RS/6000 or IBM @server pSeries model):

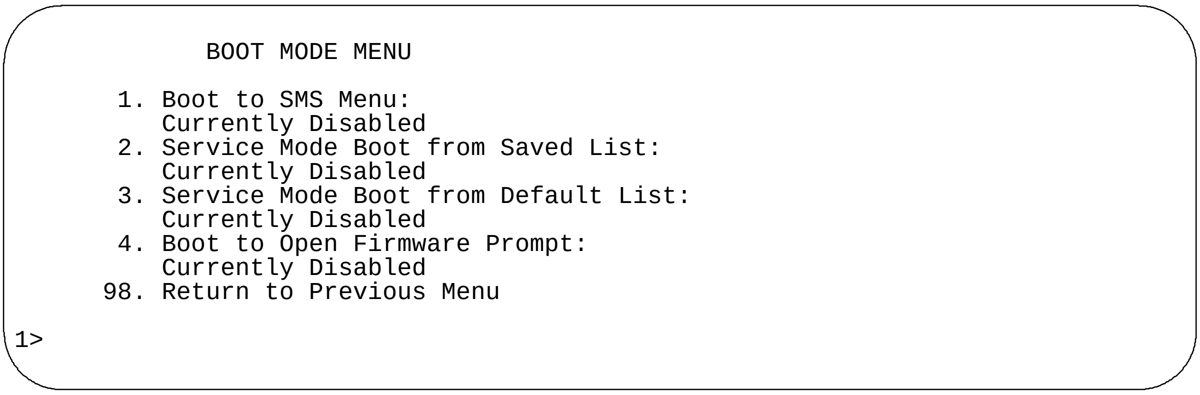


```
                    Service Processor Firmware
                    Firmware Level: sh990707
                    Copyright 1997, IBM Corporation
                    SYSTEM NAME

            MAIN MENU

   1. Service Processor Setup Menu
   2. System Power Control Menu
   3. System Information Menu
   4. Language Selection Menu
   5. Call-In/Call-Out Setup Menu
   6. Set System Name
  99. Exit from Menus

1>
```

Choose “2” for the System Power Control Menu and press ENTER.

You will see the System Power Control Menu:

Choose option 4 to power on the system. The system will ask for a confirmation and then starts to boot the system.



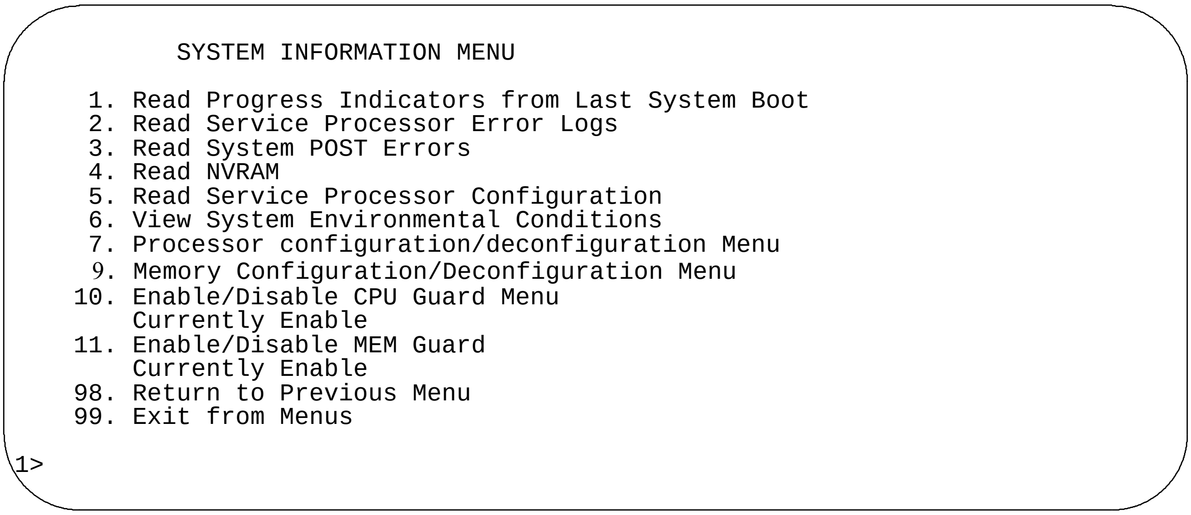
```
             BOOT MODE MENU

       1. Boot to SMS Menu:
          Currently Disabled
       2. Service Mode Boot from Saved List:
          Currently Disabled
       3. Service Mode Boot from Default List:
          Currently Disabled
       4. Boot to Open Firmware Prompt:
          Currently Disabled
      98. Return to Previous Menu

1>
```

You will also see all the LED’s which appear on the operator panel during the BIST (Build In Self Test) an POST (Power On Self Test) on the control workstation. You can monitor the BIST and POST process remotely.
As soon as AIX is booted you’ll no longer see the LED’s coming from AIX.

Choose option 7 if you want to use a specific boot mode. This allows you to set the system to automatically start a specific function on the next system start. This setting applies to the next boot only and is reset to the default state of being disabled following a successful boot attempt. You will see the following screen:

1. **Enabling the Boot to SMS Menu option**
   Causes the system to automatically enter the System Management Services menu during the boot process. Enabling this option is equivalent to pressing the 1 key while the system initialization indicators display on screen.
2. **Enabling the Service Mode Boot from Saved list option**
   Causes the system to automatically enter the stand-alone diagnostics. Enabling this option is equivalent to pressing the 5 key while the system initialization indicators display on screen.
3. **Enabling the Service Mode Boot from Default list option**
   Causes the system to automatically enter the online diagnostics in service mode. Enabling this option is equivalent to pressing the 6 key while the system initialization indicators display on screen.
4. **Enabling the Boot to Open Firmware Prompt option**
   Causes the system to automatically enter open firmware prompt (also called the OK prompt). Enabling this option is equivalent to pressing the 8 key while the system initialization indicators display on screen.

How to change the hardware configuration of your remote system
Before you boot your remote system you can change the system configuration by enabling or disabling processors or memory. From the Service Processor main menu choose option 3 for the System Information menu. It will look like the following:



```
            SYSTEM INFORMATION MENU

      1. Read Progress Indicators from Last System Boot
      2. Read Service Processor Error Logs
      3. Read System POST Errors
      4. Read NVRAM
      5. Read Service Processor Configuration
      6. View System Environmental Conditions
      7. Processor configuration/deconfiguration Menu
      9. Memory Configuration/Deconfiguration Menu
     10. Enable/Disable CPU Guard Menu
         Currently Enable
     11. Enable/Disable MEM Guard
         Currently Enable
     98. Return to Previous Menu
     99. Exit from Menus

1>
```

Choose option 7 or 9 and the you can manually configure or deconfigure any processor or memory DIMM (dual inline memory module), regardless of failure status, through this Service Processor menu. The configuration process takes place during the system power-up. Therefore, the configuration displayed in STANDBY mode reflects the configuration during the last boot. To view the current configuration, access the Service Processor menu after the system starts. When the user selects a processor or memory DIMM, its state toggles between configured and deconfigured. Processors or memory DIMMs that are not present are not listed.

A processor or memory DIMM can be in any of the following four states:

1.**Configured by System**: The processor or memory DIMM is present, and has not exceeded the number of failure threshold. It is configured by the system and is available.

2. **Deconfigured by System**: The processor or memory DIMM is present, but has exceeded the number of failure threshold. It is deconfigured by the system and is currently unavailable.
3. **Manually configured**: The processor or memory DIMM is present and available. It is configured by the user through the Service Processor menus.
4. **Manually deconfigured**: The processor or memory DIMM is present, but unavailable. It has been deconfigured by the user through the Service Processor menus.

Special considerations with RS/6000 S70, S7A, S80 or IBM @server pSeries 680 models

You can access the Service Processor when the hardware is powered-off and connected to a power cable (indicated by a OK in the operator panel) or in the standby state. This is valid for all of the RS/6000 and IBM @server pSeries models, except the RS/6000 Sxx-family or the IBM @server pSeries 680. The Service Processor of the Sxx-family or p680 is only accessible in a certain state during system boot.
This leads to the difficulty that you have to press the power-on switch manually if the system is powered-off (i.e. after a shutdown) even if the system is connected to a power cable.

The only solution to avoid this inconvenience is “ring indicate power-on” using a modem. The ring signal of the modem, which is connected to the Service Processor, powers on the system and after about four minutes (depending on the model) the machine waits about five minutes in a state where you can access the Service Processor as usual with your serial connection. If you don’t access the Service Processor in this time frame, the machine continues the boot process or shuts down again (depending of the configuration).

Important: For security reasons you have to change the Service Processor settings of an RS/6000 Sxx family or IBM @server pSeries 680 model to *call-in enable*, otherwise the machine would start booting, but you wouldn’t be able to access the Service Processor or continue the boot process.

You can invoke the “ring indicate power-on” either with a regular phone or using *cu* again.
Here is an example how to use the cu command:

```
cu -ml tty2
atdt 123456
~.
```

Useful AIX settings for remote management

In a 24x7 environment the systems should restart automatically after an operating system crash. In case of a crash of AIX the system shows a blinking 888 in the display

of the operator panel. The system continues to halt at 888 until you start locally actions to recover. During that halt there is no remote access.
To recover from a 888 message you have to press manually the power-on switch of the machine. The system will then boot again up to the state where you are prompted to choose the device for the dump file.
If you don’t want to do this manual tasks you have to change the following defaults:

```
smitty
-> System Environments
        -> Change/Show Characteristics of Operating System
                -> Automatically REBOOT system after crash true
```

Change the Dump Devices to a dedicated file system with appropriate size.

```
smitty dump
        -> Change/Show Characteristics of Dump Device
                -> Change the Primary Dump Device
                -> Change the Secondary Dump Device
                -> ...
```

Learning from the RS/6000 SP concept
Like in the RS/6000 SP environment you should consider as well a dedicated control LAN for service purposes, like remote installations or back-ups. Using this IP connection our control workstation can fulfill many useful management tasks like the following:

- NFS Server for LPP Sources
- NIM Server for remote installations, software updates and backups (mksysb)
- Synchronizing purposes of system files like /etc/hosts, /etc/services, .rhosts, users and printers (using the AIX commands *rcp* or *rdist*)
- Secure Shell (open or commercial ssh) for secure and convenient remote login.
- Systems management software like Tivoli or Performance Toolbox

Native Serial Port Support for HACMP

When it comes to clustering a group of servers it is not always for management consolidation but also for making some services served by these machines high available. This might be for example file and/or application serving. In this case the tool of choice is HACMP.

HACMP combines up to 32 machines within a cluster which are called cluster nodes and minimizes the risk of loosing one of the provided services by spreading the responsibility of “keeping alive” this service over at least two nodes. So if one node breaks down one of the others takes over and continues providing the services of the lost machine.

The advantage is that it is not necessary to waste resources by running one machine as an idle “hot standby” node which does nothing but waiting for the production machine to crash. It is possible to run the nodes of a HA cluster in a mutual takeover mode which allows them all to provide their own service but having still some resources left to take over the service of a crashing node.

Every sane member of the cluster sends periodically electronic heartbeats across different types of networks so that it can easily be recognized if one node fails. If too many heartbeats are missing or the heartbeat of one node completely stops it will be considered by the other nodes as service disruption and a previously defined takeover process will be initiated. Therefore it is very important to avoid any disturbance on these networks which could be misinterpreted.

Why Is a Serial Network Link Needed in a HACMP Cluster

So as HACMP is an intelligent way of using redundant resources to remove so called “single points of failure” it makes sense that the communication network between the cluster nodes is also layed out redundantly.

The common networks (ethernet, token-ring, FDDI, ...) are using the IP layer for communication which may also be “single point of failure” because if the IP stack fails all networks within one machine are dead. So the redundant layout of the network adapters does not cover all possible network faults. This means there has to be a type of network which is independent from the IP layer. The classical alternative is a TTY connection between the serial ports of two machines.

This TTY link is exclusively to be used by HACMP for heartbeat sending. Besides the fact that this connection is very robust it provides the opportunity for the nodes to communicate even when all other networks are down.

This is very important for the cluster to make the right decision about the type of takeover that should be initiated.

There is also a new technology called “target mode” connection over SSA or SCSI cabling which may be used if there are no serial ports (SCSI) or if longer distances than

15 meters need to be covered (SSA: cable length may be up to 2 * 25 meter). But this document focusses only the serial link connection.

Note:
Remember carefully that there has to be serial network link from every to every node. This means that a machine belonging to a cluster with N nodes needs N-1 serial ports to establish a serial network link to all other nodes of this cluster.

For example: If there is a cluster with 4 nodes. Every machine needs 3 native serial ports. This can normally only be done with an additional adapter card which may extend the number of native serial ports from 8 to 128.

A Serial Port Must Not Be A Serial Port

As it will be shown in the following section the serial ports on the back of a machine do often not seem to be what they appear to be. As an enhancement of serviceability the serial ports have more and more been developed and have been equipped with additional functions so that it is no more possible to use them as simple TTY connections. The ports are transmitting more than just the common serial traffic and this is what disturbs the classical TTY transfer. Therefore not all serial ports can be used for HACMP heartbeat transmission anymore.

An Overview of Serial Port Function

The following table lists the number and function of the serial ports for the current server models. It is distinguished between those ports which may be only used for service processor access and those which are able to transport the serial network heartbeat of HACMP.

Note:
Only the ports dedicated for use with HACMP meet the necessary requirement of a native serial link which guarantees a non disruptive heartbeat communication between two cluster nodes.

Machine Type	Number Of Serial Ports	Ports Used for Service	Ports Used for HACMP
Standalone Models			
7044 - 270	2	S1, S2	(Notes 1)
7025 - F50	3	S1, S2	S3
7025 - F80	4	S1 - S3	S4
Rack Models			

7046 - B50	2	S1, S2	(Notes 2)
7026 - H50	3	S1, S2	S3
7026 - H70	3	S1, S2	S3
pSeries 620 6F1	4	S1, S2	S3, S4
7026 - H80	4	S1 - S3	S4
pSeries 640	3	S1, S2	(S2), S3
7026 - M80	4	S1 - S3	S4
pSeries 660 6H1	4	S1, S2	S3, S4
Enterprise Server			
7017 - S70, S7A, S80	2	S1, S2	(Notes 1)
pSeries 680	2	S1, S2	(Notes 1)
SP Nodes			
All MCA Nodes	1	S1, (Notes 3)	(Notes 4)
POWER3 Thin / Wide 200MHz (Winterhawk 1)	2	S1, (Notes 3)	S2, (Notes 5)
POWER3 Thin / Wide 375MHz (Winterhawk 2)	2	S1, (Notes 3)	S2, (Notes 5)
POWER3 High 222 / 375 MHz	3	S1, S2, (Notes 3)	S3, (Notes 5)

Notes:

1.These models do not support the use of native serial ports in an HACMP/ES RS232 serial network. Configuration of an RS232 serial network requires a PCI multi_port Async card in any of these systems.

2.Due to the minimized configuration of this machine type it is necessary to submit a PRPQ to development for use with HACMP.

3.The serial ports of SP nodes differ from the common serial port architecture. These ports support a special communication protocol (SAMI) which used to monitor several advanced hardware conditions on the SPs control workstation.

4.On the Microchannel SP thin or wide nodes there are no serial ports available. Therefore, any HACMP/ES configurations that require a tty network need to make use of a serial adapter card (8-port async EIA-232 adapter, FC2930), available on the SP as an RPQ.

5.Eventhough it is stated below that the highest serial port can be used for serial heartbeat with HACMP it is not recommended to use it but to use instead a PCI multi_port Async card. There have been performance issues measured which do not garantuee a proper operation of the heartbeat so it is strongly recommended to use the alternative way.

FC 2054/2058, 222MHz/375 MHz POWER3 High (aka Nighthawk 1/2)
Two external, 9-pin, RS-232 connectors on the planar S2 and S3 ports. The S3 port is supported only for HACMP serial heartbeat; the S2 port is not supported for this use. A 9-pin to 25-pin converter cable is included with the node for this connector.

FC 2056/2057, 375 MHz POWER3 Thin/Wide (aka Winterhawk 2)
External nine-pin RS-232 on the planar S2 port (supported only for HACMP serial heartbeat);a 9 to25-pin converter cable is included with the node.

FC 2052/2053, 200MHz POWER3 Thin/Wide (aka Winterhawk 1)
External nine-pin RS-232 on the planar S2 port (supported only for HACMP serial heartbeat);a 9 to25-pin converter cable is included with the node.

The HACMP Level3 Support Intranet Website

An unavoidable focal point for latest information about HACMP and in this special case about the “Native Serial Port Support” is the IBM Intranet Website of the HACMP Level3 Support in Austin

http://hacmp.aix.dfw.ibm.com

NIM in a Clustered Environment

AIX offers with NIM (**N**etwork **I**nstallation **M**anagement) a very powerful tool to cover all possible installation and maintenance scenarios of the operating system. But this does not imply that NIM is difficult to handle.

This chapter will introduce some aspects of the AIX Network Installation Mangement which can be usefully integrated into a cluster management strategy but it will not cover any basic configuration steps.

The major message will be that NIM covers everything about installation, maintenance, backup and recovery of the operating system and meets therefore the best requirements to handle this complex task for all machines within a clustered environment.

An Overview of NIM Advantages

NIM offers a huge variety of ways to install, customize and maintenance the operating system on every RS/6000 and pSeries model. The only requirement is that the network adapter of the installation target machine supports RIPL and is of that kind which is supported by NIM. But this usually fits to all common network adapters.
The easy installation and configuration applies to the clients as well as to the server. Generally all administrative NIM tasks may be managed from the server but it is also possible to initiate some actions like the installation or a software update directly from the client.

Note:

There is a X-based administrative console available called “xnim” which can be operated intuitively very well.

The basic functionality of NIM is to provide all necessary installation resources to a network through a dedicated NIM server machine so that no tape, cdrom or other direct access (e. g. console) to the NIM clients machine is needed and the installation and optional customization or maintenance of the operating system can be performed completely unattended. The installed machines are “ready to go” which might be an interesting point for initial installation as well as for disaster recovery.

During the installation process the client machine runs a copy of the operating system which is provided by the server via the network. This copy is called SPOT (**S**hared **P**roduct **O**bject **T**ree). This remote boot method can not only be used for installation but also for maintenance and diagnostics which offers just one more way of remote management.

The Standard NIM Environment

The standard NIM environment consists of one NIM server and at least one NIM client. The NIM server holds the definition of all resources, clients, the dedicated networks and of course the resources themselves.

There are three types of installation methods: rte, spot, mksysb. But only two of them will be discussed in this context.

rte - is a plain BOSINST installation like it would be performed from a CD or tape device. In this case the installation source is the LPPSource on the NIM server. If there are update filesets in the LPPSource they will be applied and committed during the installation process.

mksysb - The installation from a previously made “mksysb” file offers quite a lot of opportunities. This is a very fast and comfortable way of recovery if the mksysb is installed (restored) to its original machine. But it also can be used for a migration installation when it is allocated together with a LPPSource of a higher operating system level.
When a machine is replaced by a newer, bigger, faster machine it is no big deal to clone the mksysb to a machine with a different architecture or
different adapter types. NIM automatically replaces the old unsuitable filesets with new valid ones from the LPPSource.

By default are there no predefined customization scripts for additional installation or configuration tasks. This all has to be individually adjusted to the given environment. NIM provides the basic installation functionality at first stage and all individual adjustments have to be done manually. This means that NIM is a huge gun but has to be loaded by your own.

The SP NIM Environment

The SP NIM Environment is a good example of a specific customized NIM configuration. In consideration of its main purpose it provides a centralized instance to serve the parallel installation of all SP nodes across the SP internal administration network. The NIM environment is automatically installed and configured by the PSSP software.
Additionally to the basic installation of the operating system NIM installs separatly the PSSP software and takes care of the configuration of all network interfaces of a node.
These two customization tasks have been added to the conventional NIM environment by the PSSP Software to ensure that a node is all set up for joining the SP cluster right after the NIM installation process.

The success of the RS/6000 SP as a preferred platform for server consolidation is not least to attribute to the SP control workstation as a “Single Point of Control” for all major administration and configuration tasks and the centralized event and monitoring management.

NIM in a Clustered Environment

As shown in the description of the SP NIM enviroment before it is possible to centralize the major tasks of the system administration to a dedicated machine.
When a NIM server is now brought onto the same machine that has already been selected to manage the low level administration access through the service processor to all clustered machines then it is obvious that the control workstation concept is no more reserved for the SP enviroment.

The classical NIM methods joined together with the advanced remote machine management features of the new generation of service processors gives the opportunity to implement the concept of a “Single Point of Control” administration for clustered machines of any type.

NIM - Best Practice

“Best Practice” means to use a technology in its most effective way by keeping its configuration and maintenance as simple as possible. The impression of NIM so far is that it is a powerful tool but at the same time this might be deterrent too because of the versatile installation methods that need to be set up manually.
By using of one previously mentioned basic installation functions of NIM - the mksysb installation - it offers a simple but very effective means for disaster recovery as it will be described in the next topic.

Requirements of Fast OS Recovery

What are the expectations of a fast recovery. Of course that it is fast but also that it is easy to initiate and that it is reliable. These three requisites are combined in the concept of NIM.
It is ***fast*** - The restore process of a mksysb runs across the network and takes advantage of the full bandwidth if there is a dedicated network. At this point it should be remembered that NIM supports Gigabit ethernet which provides an enormous bandwidth.
It is ***easy*** - Just set the node to network boot and allocate all necessary installation resources on the NIM server and the machine comes back “read to go” directly after the NIM installation from its own mksysb.
It is ***reliable*** - Restoring an mksysb is the safest way to recover a operating system. And since the mksysb files of all machines are stored in one location on the NIM server it is easy to protect them by mirroring or other means of data backup.
This is another very grave reason why NIM plays an essential role in a centrally managed clustered environment.

Housekeeping your Installation Resources

The configuration data of NIM environment can be considered as static data once everything has been set up. So the expense of maintenance is limited to the housekeeping of the installation resources like there are the baselevel and update filesets of the LPPSource, the SPOT which is created on basis of the LPPSource and finally the mksysb files that shall be used for installation, cloning or recovery.

LPPSource - Normally there is one directory where at first stage all baselevel filesets are stored and in further steps update filesets will be added. This means that a lot of space is wasted and the LPPSource may be destroyed when a broken update is applied. So baselevel and update filesets should first be separated and then linked to a common directory. This gives the opportunity to use the baselevel filesets multiple times to build own LPPSources for different update levels and it protects existing LPPSources and allows different OS levels without wasting space. Having more than one LPPSource minimizes the risk of destroying LPPSource and SPOT during an update and it is very easy to get rid of a superseded version by just deleting the corresponding LPPSource and SPOT directories.

SPOT - is a minimal operating system which is located in a dedicated directory tree and exported via NFS from the server during installation, maintenance or diagnostic processes on the client machine. It is very important that this resource is always available. Each LPPSource creates its own SPOT. If there is only one LPPSource then there is only one SPOT which needs to be updated simultaneously to the LPPSource. If this update fails or some broken filesets are applied to the SPOT then the whole NIM service breaks down while SPOT and LPPSource need to be completely reinstalled which is very time consuming.This emphasizes the strategy of having more than just one LPPSource and not removing an old OS level until the new SPOT and LPPSource have been tested successfully.

mksysb files - There may be existing default install image files which are normally used for initial installation. There may be individually created install image files as starting point for certain types of installation and finally there may be install images from mksysb files which have been created as system backups. The mksysb files should be stored in a mirrored filesystem and additionally be backuped.

In this context only some very general ideas could be provided of what can be done to optimize the management, availability and reliablity of the NIM installation resources. But hopefully it could be shown that NIM is no miracle but a useful and a usable tool.

NIM Documentation

There is a list below which shows several very good Redbooks about the variety of opportunities of NIM. They provide basic and advanced knowledge about implementation, configuration and troubleshooting. This will be a good guidance for setting up your individual NIM strategy.

General NIM

→ Manual**:** *NIM from A to Z in AIX 43 (SG24-5524)*

→ Manual: *AIX LVM from A to Z - Troubleshooting and Commands (SG24-5433)*

SP NIM

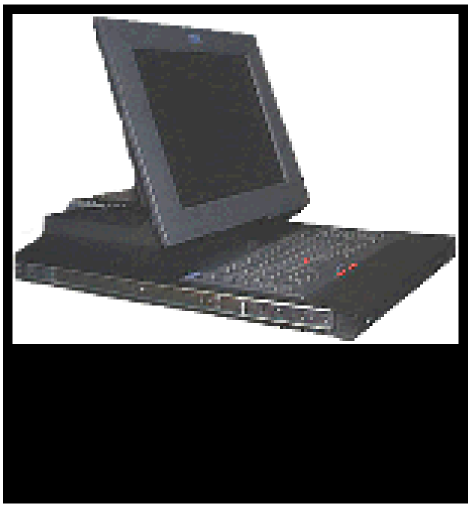
→ Manual: *PSSP Version3 Survival Guide (SG24-5344)*

→ Manual: *RS6000 SP Software Maintenance (SG24-5160)*

→ Manual*: PSSP v3r2 Installation and Migration Guide (GA22-7347-02)*

TF1 Rack Console - A small solution for large environments

The growth in computing areas need more performance and more equipment. Most of the computer require a display, keyboard and mouse for operating and maintenance. In computer centers it becomes more and more importand to save floor space. A comfortable way to collect maschines is a system rack. Stacked in a metal frame a minimum on ground space is used for systems except the peripheral devices.



To solve this problem, IBM offers the 7316 Model TF1 Rack Console. This integrated rack console is an attractive, cost-effective option both for customers who manage multiple systems in large “server farms” and for those who operate individual rack-mounted servers.

The offering features the IBM T54A Flat Panel Color Monitor, a rack-mounted keyboard tray, a flat panel mounting kit, and an IBM Space Saver 2 Keyboard - all in a package that occupies only 3 Units (5.25 inches) in a 19-inch rack enclosure.

The T54A Flat Panel Color Monitor offers advanced display characteristics in a stylish, space-saving package. This display provides a bright, flicker-free 15 inch (304.1 mm x 228.1 mm) viewable image with 1024 x 768 pixel addressability.

The IBM Space Saver 2 Keyboard, a space-efficent version of the popular IBM Trackpoint Keyboard, is only 14.5 inches long and fits easily on the rack keyboard tray. For the convenience of customers worldwide, it s available in a choice of sixteen language configurations.

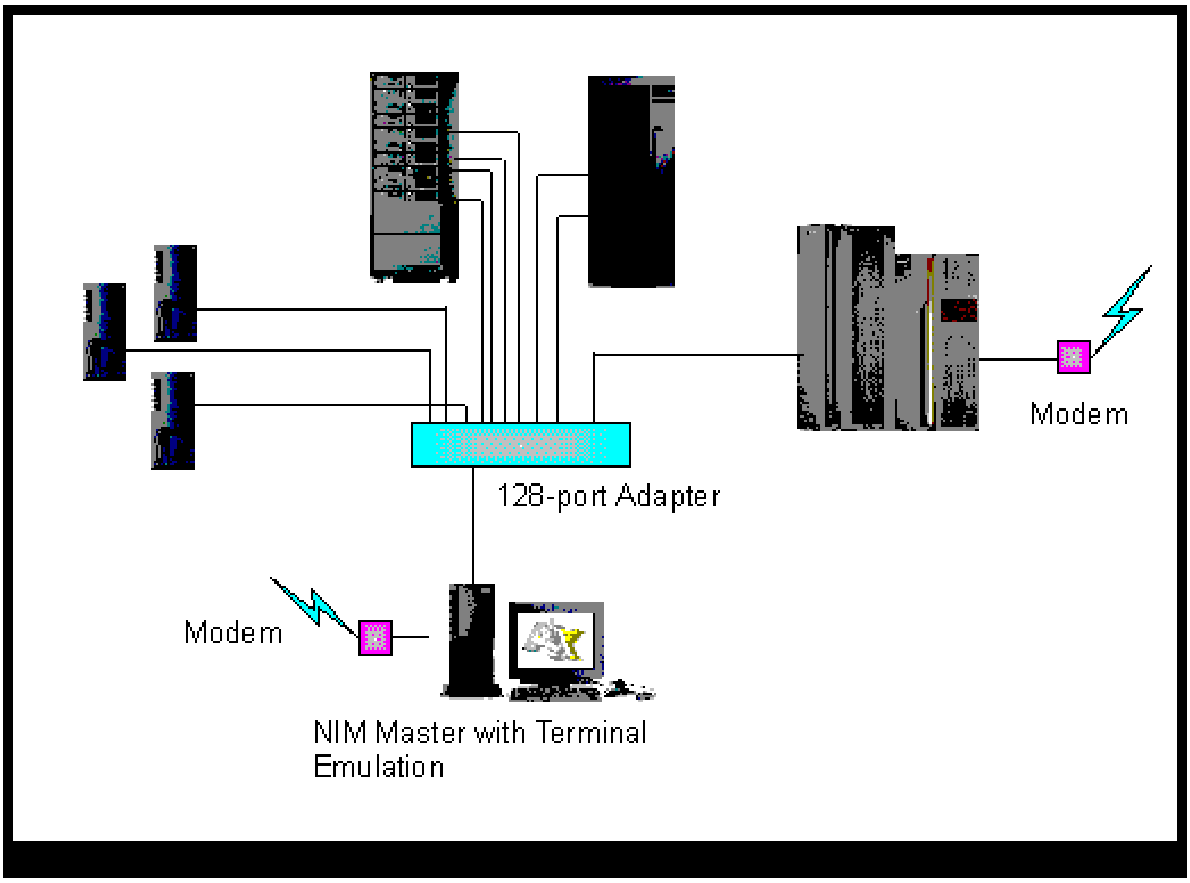
Designed to fit into a variety og IBM and non-IBM rack enclosures, including the IBM Model T00 (1.8 meter, 36 Units) and Model T42 (2.0 meter, 42 Units) racks, the keyboard tray and monitor assembly mounts conveniently on ball bearing slides to enable quick installation, long-term usage and secure cabme management. The console slides out for easy access to system and application management utilities and tucks back into the rack enclosure to maximize use of expensive raised-floor space.

The 7316 Model TF1 Rack Console is supported on IBM @server pSeries 640, 660 and 680 servers, as well as an IBM RS/6000 Models B50, H70, H80, M80 and S80. It can ether be included with a system order of one of the supported server, or be ordered seperately.

Loose Cluster - or what comes for free with AIX

It is amazing: just few customer are using default hardware and software functions which come for free with pSeries and RS/6000 and AIX. Build your own cluster by using the Service Processor and Network Installation Management (NIM).

The following picture shows a typical remote management environment, which consist of a NIM Master and several NIM clients. The NIM Master is connected with serial to the service processors of each server. Like the SP Control Workstation you need an 128-port adapter in order to increase the number of serial ports.



A Terminal Server provides a TCP/IP connection to serial or graphical (input/output) devices. In our case a Terminal Server is connected to the customers' LAN and via serial connections to all service processors.

The system administrators can easily manage a multi-system installation and remote capability.