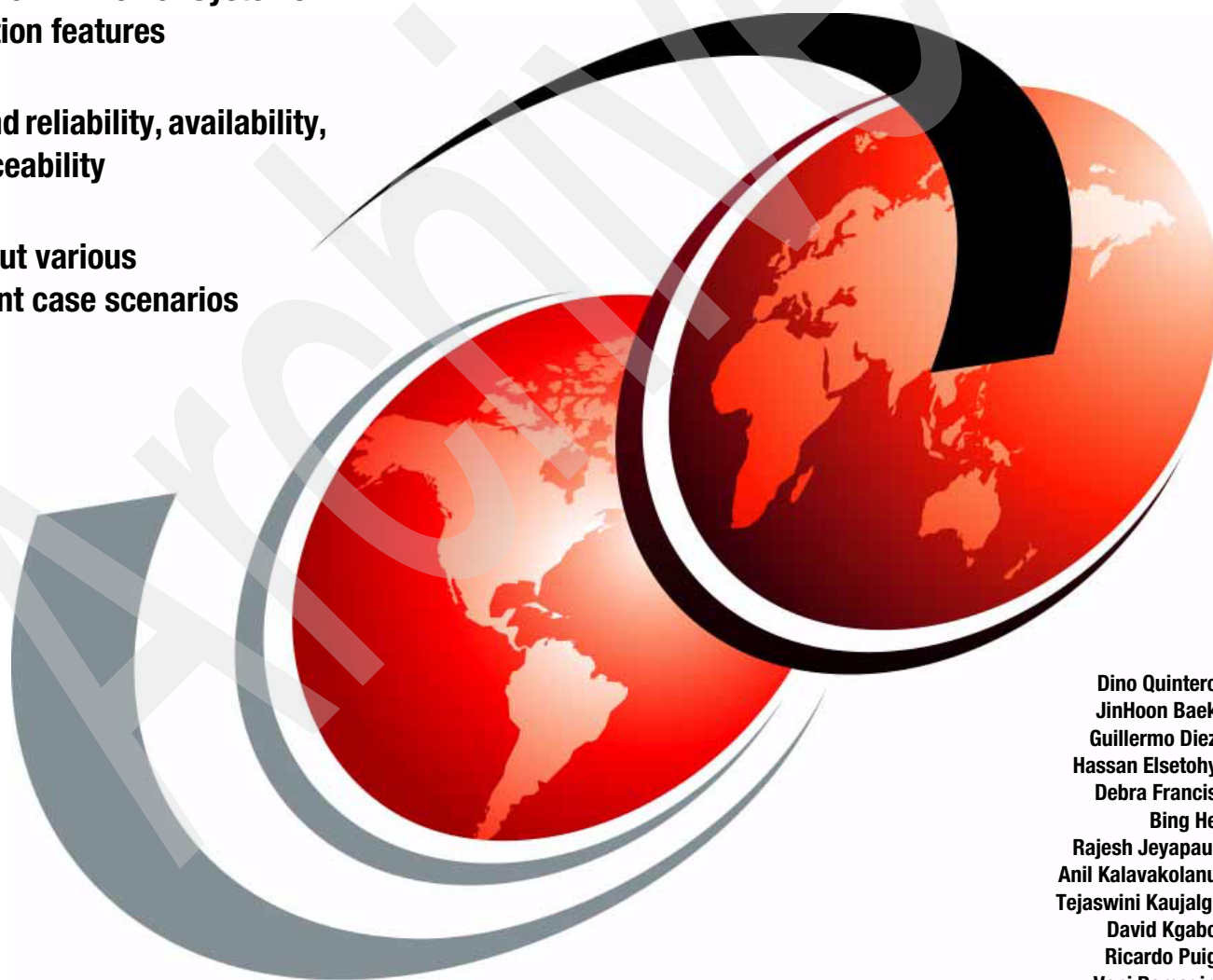


Power Systems Enterprise Servers with PowerVM Virtualization and RAS

Unleash the IBM Power Systems
virtualization features

Understand reliability, availability,
and serviceability

Learn about various
deployment case scenarios



Dino Quintero
JinHoon Baek
Guillermo Diez
Hassan Elsetohy
Debra Francis
Bing He
Rajesh Jeyapaul
Anil Kalavakolanu
Tejaswini Kaujalgi
David Kgabo
Ricardo Puig
Vani Ramagiri

Redbooks



International Technical Support Organization

**Power Systems Enterprise Servers with PowerVM
Virtualization and RAS**

December 2011

Archived

Note: Before using this information and the product it supports, read the information in “Notices” on page ix.

First Edition (December 2011)

This edition applies to AIX 7.1 SP 3, IBM SDD PCM for AIX V61 Version 2.5.2.0, HMC code level 7.3.5, and IBM Systems Director Version 6.2.1.2.

© Copyright International Business Machines Corporation 2011. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Notices	ix
Trademarks	x
Preface	xi
The team who wrote this book	xi
Now you can become a published author, too!	xiii
Comments welcome	xiii
Stay connected to IBM Redbooks	xiii
Chapter 1. Introducing POWER7 Enterprise Server RAS and virtualization features ..	1
1.1 High availability in today's business environments	2
1.2 Introduction to RAS and virtualization	2
1.2.1 Reliability, availability, and serviceability (RAS)	3
1.2.2 Virtualization	5
1.2.3 Latest available feature enhancements	8
Chapter 2. Exploring RAS and virtualization features in more detail	11
2.1 New RAS and virtualization features with POWER7	13
2.1.1 Active Memory Mirroring for the hypervisor on Power 795	13
2.1.2 Hot GX adapter repair	16
2.1.3 Improved memory RAS features	19
2.1.4 Active Memory Expansion	19
2.2 Significant features	22
2.2.1 Active Memory Mirroring for the hypervisor on the Power 795	22
2.2.2 Persistent hardware deallocation	22
2.2.3 First Failure Data Capture (FFDC)	23
2.2.4 Processor RAS features	23
2.2.5 Memory RAS features	25
2.2.6 Dynamic service processor (SP) failover at run time and redundant SP	25
2.2.7 Hot node add and repair	28
2.2.8 Hot node upgrade (memory)	28
2.3 TurboCore and MaxCore technology	28
2.3.1 Enabling and disabling TurboCore mode	32
2.4 Hypervisor and firmware technologies	33
2.4.1 Hypervisor	33
2.4.2 Firmware	35
2.4.3 Dynamic firmware update	35
2.4.4 Firmware update and upgrade strategies	36
2.5 Power management	36
2.5.1 Differences in dynamic power saver from POWER6 to POWER7	37
2.6 Rapid deployment of PowerVM clients	38
2.6.1 Deployment using the VMControl plug-in	38
2.6.2 File-backed virtual optical devices	38
2.6.3 Deployment using the System Planning Tool (SPT)	40
2.7 I/O considerations	41
2.7.1 Virtual SCSI	41
2.7.2 N_Port ID Virtualization (NPIV)	43
2.8 Active Memory Sharing	45
2.8.1 Shared memory pool	46

2.8.2	Paging virtual I/O server	47
2.8.3	Client LPAR requirements	47
2.8.4	Active Memory Sharing and Active Memory Expansion	47
2.8.5	Active Memory Sharing with Live Partition Mobility (LPM)	47
2.9	Integrated Virtual Ethernet	48
2.10	Partitioning	49
2.10.1	Creating a simple LPAR	52
2.10.2	Dynamically changing the LPAR configurations (DLAR)	58
Chapter 3.	Enhancing virtualization and RAS for higher availability	63
3.1	Live Partition Mobility (LPM)	64
3.1.1	Partition migration	65
3.1.2	Migration preparation	66
3.1.3	Inactive migration	68
3.1.4	Active migration	68
3.2	WPAR	69
3.2.1	Types of WPARs	70
3.2.2	Creating a WPAR	72
3.2.3	Live Application Mobility (LPM)	75
3.3	Partition hibernation	77
3.4	IBM SystemMirror PowerHA	80
3.4.1	Comparing PowerHA with other high-availability solutions	80
3.4.2	PowerHA 7.1, AIX, and PowerVM	82
3.5	IBM Power Flex	82
3.5.1	Power Flex Overview: RPQ 8A1830	83
3.5.2	Power Flex usage options	84
3.6	Cluster Aware AIX (CAA)	87
3.6.1	Cluster Aware AIX Services	89
3.6.2	Cluster Aware AIX event infrastructure	90
3.7	Electronic services and electronic service agent	92
3.7.1	Benefits of ESA for your IT organization and your Power systems	93
3.7.2	Secure connection methods	94
Chapter 4.	Planning for virtualization and RAS in POWER7 high-end servers	99
4.1	Physical environment planning	100
4.1.1	Site planning	100
4.1.2	Power and power distribution units (PDUs)	100
4.1.3	Networks and storage area networks (SAN)	102
4.2	Hardware planning	103
4.2.1	Adapters	105
4.2.2	Additional Power 795-specific considerations	106
4.2.3	Planning for additional Power server features	107
4.2.4	System management planning	108
4.2.5	HMC planning and multiple networks	112
4.2.6	Planning for Power virtualization	112
4.2.7	Planning for Live Partition Mobility (LPM)	113
4.3	CEC Hot Add Repair Maintenance (CHARM)	121
4.3.1	Hot add or upgrade	121
4.3.2	Hot repair	123
4.3.3	Planning guidelines and prerequisites	124
4.4	Software planning	131
4.5	HMC server and partition support limits	132
4.6	Migrating from POWER6 to POWER7	132

4.6.1 Migrating hardware from POWER6 and POWER6+ to POWER7	132
4.6.2 Migrating the operating system from previous Power servers to POWER7	133
4.6.3 Disk-based migrations.	138
4.6.4 SAN-based migration with physical adapters	138
4.6.5 After migration to POWER7	145
4.7 Technical and Delivery Assessment (TDA).	150
4.8 System Planning Tool (SPT).	151
4.9 General planning guidelines for highly available systems.	154
Chapter 5. POWER7 system management consoles	157
5.1 SDMC features	158
5.1.1 Installing the SDMC	158
5.1.2 SDMC transition	158
5.1.3 SDMC key functionalities	159
5.1.4 HMC versus SDMC.	160
5.1.5 Statement of direction for support HMC	164
5.2 Virtualization management: Systems Director VMControl	165
5.2.1 VMControl terminology	165
5.2.2 VMControl planning and installation	167
5.2.3 Managing a virtual server	171
5.2.4 Relocating a virtual server.	173
5.2.5 Managing virtual appliances	174
5.2.6 Creating a workload	178
5.2.7 Managing server system pools	179
5.3 IBM Systems Director Active Energy Management (AEM)	183
5.3.1 Active Energy Manager (AEM) overview	183
5.3.2 AEM planning, installation, and uninstallation.	184
5.3.3 AEM and the managed systems.	185
5.3.4 Managing and monitoring the consumed power using AEM.	187
5.4 High availability Systems Director management consoles	193
Chapter 6. Scenarios	195
6.1 Hot node add and repair	196
6.1.1 Hot node add	196
6.1.2 Hot node repair	202
6.2 Hot GX adapter add and repair	206
6.2.1 Hot GX adapter add	206
6.2.2 Hot GX adapter repair.	207
6.3 Live Partition Mobility (LPM) using the HMC and SDMC	210
6.3.1 Inactive migration from POWER6 to POWER7 using HMC and SDMC.	210
6.4 Active migration example	214
6.5 Building a configuration from the beginning	217
6.5.1 Virtual I/O servers	218
6.5.2 HEA port configuration for dedicated SEA use	221
6.5.3 NIB and SEA failover configuration.	221
6.5.4 Active Memory Sharing configuration	223
6.5.5 NPIV planning	226
6.5.6 Client LPAR creation (virtual servers).	227
6.5.7 Server-side NPIV configuration.	229
6.6 LPM and PowerHA	241
6.6.1 The LPM operation	243
6.6.2 The PowerHA operation	245
Chapter 7. POWER7 Enterprise Server performance considerations	247

7.1	Introduction	248
7.2	Performance design for POWER7 Enterprise Servers	248
7.2.1	Balanced architecture of POWER7	248
7.2.2	Processor eDRAM technology	251
7.2.3	Processor compatibility mode	251
7.2.4	MaxCore and TurboCore modes	252
7.2.5	Active Memory Expansion	252
7.2.6	Power management's effect on system performance	252
7.3	POWER7 Servers performance considerations	254
7.3.1	Processor compatibility mode	254
7.3.2	TurboCore and MaxCore modes	259
7.3.3	Active Memory Expansion (AME)	261
7.3.4	Logical memory block size	264
7.3.5	System huge-page memory	266
7.4	Performance considerations with hardware RAS features	271
7.4.1	Active Memory Mirroring for the hypervisor	271
7.5	Performance considerations with Power virtualization features	272
7.5.1	Dynamic logical partitioning (DLPAR)	272
7.5.2	Micro-partitioning	273
7.5.3	PowerVM Lx86	275
7.5.4	Virtual I/O server	277
7.5.5	Active Memory Sharing	280
7.5.6	Live Partition Mobility	282
7.6	Performance considerations with AIX	284
7.6.1	Olson and POSIX time zones	284
7.6.2	Large page size	286
7.6.3	One TB segment aliasing	289
7.6.4	Memory affinity	291
7.6.5	Hardware memory prefetch	292
7.6.6	Simultaneous multithreading (SMT)	294
7.6.7	New features of XL C/C++ V11.1	297
7.6.8	How to deal with unbalanced core and memory placement	298
7.6.9	AIX performance tuning web resources	300
7.7	IBM i performance considerations	301
7.7.1	Overview	301
7.7.2	Optimizing POWER7 performance through tuning system resources	303
7.8	Enhanced performance tools of AIX for POWER7	305
7.8.1	Monitoring POWER7 processor utilization	305
7.8.2	Monitoring power saving modes	307
7.8.3	Monitoring CPU frequency using lparstat	308
7.8.4	Monitoring hypervisor statistics	309
7.8.5	Capabilities for 1024 CPU support	312
7.8.6	Monitoring block IO statistics	315
7.8.7	Monitoring Active Memory Expansion (AME) statistics	317
7.8.8	Monitoring memory affinity statistics	323
7.8.9	Monitoring the available CPU units in a processor pool	326
7.8.10	Monitoring remote node statistics in a clustered AIX environment	328
7.9	Performance Management for Power Systems	328
7.9.1	Levels of support available within PM for Power Systems	329
7.9.2	Benefits of PM for Power Systems	330
7.9.3	Data collection	331
7.9.4	Accessing the PM for Power Systems website	332

Chapter 8. PowerCare Services offerings for Power Enterprise Servers.	335
8.1 PowerCare highlights	336
8.2 PowerCare Services offerings.	336
8.2.1 Availability optimization services.	337
8.2.2 Systems Director and VMControl enablement	339
8.2.3 Systems Director Active Energy Manager enablement.	341
8.2.4 IBM Systems Director Management Console	341
8.2.5 Security assessment.	342
8.2.6 Performance optimization assessment.	346
8.2.7 Power Flex enablement	347
8.2.8 Power 795 upgrade implementation services.	349
8.2.9 PowerCare technical training	350
Appendix A. Administration concepts	353
Making a root volume group (rootvg) easier to manage	354
Example importing non-root volume group	355
A dynamic LPAR operation using the HMC	357
Setting up Secure Shell keys between two management consoles.	360
Simple cluster installation.	360
Installing and configuring PowerHA	363
Appendix B. Performance concepts	375
Performance concepts	376
Throughput versus response time	376
Performance and computing resources	377
Central processing unit	378
Multiple core systems	378
Memory architecture	380
Server I/O storage.	382
Performance metrics.	387
Performance benchmarks.	388
Appendix C. ITSO Power Systems testing environment	393
Austin environment	394
Poughkeepsie benchmark center environment	394
ITSO Poughkeepsie environment	395
Related publications	397
IBM Redbooks publications	397
Other publications	398
Online resources	398
Help from IBM	399
Index	401

Archived

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.


COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at <http://www.ibm.com/legal/copytrade.shtml>

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

Active Memory™	Informix®	pSeries®
AIX 5L™	iSeries®	Redbooks®
AIX®	Orchestrate®	Redpaper™
BladeCenter®	Power Architecture®	Redbooks (logo)  ®
DB2®	Power Systems™	System i®
DS8000®	POWER4™	System p®
Electronic Service Agent™	POWER5™	System Storage®
EnergyScale™	POWER6+™	System x®
eServer™	POWER6®	System z®
GPFS™	POWER7™	Systems Director VMControl™
HACMP™	POWER7 Systems™	WebSphere®
IBM Systems Director Active Energy Manager™	PowerHA™	XIV®
IBM®	PowerVM™	
	POWER®	

The following terms are trademarks of other companies:

Intel, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

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

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

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

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

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

Preface

This IBM® Redbooks® publication illustrates implementation, testing, and helpful scenarios with IBM Power® Systems 780 and 795 using the comprehensive set of the Power virtualization features. We focus on the Power Systems functional improvements, in particular, highlighting the reliability, availability, and serviceability (RAS) features of the enterprise servers.

This document highlights IBM Power Systems Enterprise Server features, such as system scalability, virtualization features, and logical partitioning among others. This book provides a documented deployment model for Power 780 and Power 795 within a virtualized environment, which allows clients to plan a foundation for exploiting and using the latest features of the IBM Power Systems Enterprise Servers.

The target audience for this book includes technical professionals (IT consultants, technical support staff, IT Architects, and IT Specialists) responsible for providing IBM Power Systems solutions and support.

The team who wrote this book

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

Dino Quintero is a Project Leader and IT generalist with the ITSO in Poughkeepsie, NY. His areas of knowledge include enterprise continuous availability planning and implementation, enterprise systems management, virtualization, and clustering solutions. He is currently an Open Group Master Certified IT Specialist. Dino holds a Master of Computing Information Systems degree and a Bachelor of Science degree in Computer Science from Marist College.

JinHoon Baek is a Certified Product Support Professional and Senior System Service Representative (SSR) in IBM Korea, working in Maintenance and Technical Support. He is also an IBM Certified Advanced Technical Expert in IBM System p® and AIX® 5L™ with seven years of experience in AIX and IBM Power Systems. His areas of expertise include high-end storage systems, including SAN, Power Systems, and PowerVM™, as well as AIX, GPFS™ and PowerHA™.

Guillermo Diez is a Certified IT Specialist and IBM Certified Systems Expert in Virtualization Technical Support for AIX and Linux working at the Service Delivery Center in IBM Uruguay. He joined IBM in 2003 and works as the Team Leader for the UNIX and Storage administration teams since 2007. His areas of expertise include AIX, Linux, PowerVM, performance tuning, TCP/IP, and midrange storage systems. Guillermo also holds a Computer Engineer degree from the Universidad Catolica del Uruguay (UCUDAL).

Hassan Elsetohy is dual-certified professional. He is both a Certified IT Architect and a Certified IT Specialist. He performs the lead architect role for full life-cycle engagements, and also undertakes the Method Exponent role in large engagements, in addition to his lead architect role. He also sometimes performs the SME role in his in-depth areas of expertise, such as Storage and AIX/System p. Hassan Joined IBM in 1994 directly from university after attaining his Bachelor of Engineering in Electrical Engineering. Hassan also attained his Masters degree in VLSI Design - Course Work - in 1996.

Debra Francis is a Senior Managing Consultant with IBM STG Lab Services and Training out of Rochester, MN, with over 25 years of experience with IBM midrange and Power Systems. She is part of the Worldwide PowerCare team that works with Power Enterprise Server clients around the globe. This team tackles the clients' IT availability demands to meet the business expectations of today and to provide input and availability consulting as part of a solid IT resiliency strategy.

Bing He is a Senior I/T Specialist of the IBM Advanced Technical Skills (ATS) team in China. He has 11 years of experience with IBM Power Systems. He has worked at IBM for over four years. His areas of expertise include PowerHA, PowerVM, and performance tuning on AIX.

Rajesh Jeyapaul is the technical lead for IBM Systems Director POWER Server management. His focus is on the PowerHA SystemMirror plug-in and PowerRF interface for Virtualization Management Control (VMC) Plug-in on System Director. He is part of the Technical advocate team that works closely work with clients to tackle their POWER Server-related issues. Rajesh holds a Master in Software Systems degree from the University of BITS, India, and a Master of Business Administration (MBA) degree from the University of MKU, India.

Anil Kalavakolanu is a Senior Engineer and also Technical Lead in the AIX Development Support Organization. He has 18 years of experience supporting AIX and POWER. He holds a Masters degree in Electrical Engineering from University of Alabama, in Tuscaloosa, AL. His areas of expertise include AIX, PowerVM, and SAN.

Tejaswini Kaujalgi is currently working as a Systems Software Engineer in the IBM AIX UNIX Product Testing team, Bangalore, India. Her expertise lies in the areas of AIX, PowerHA, Security, and Virtualization. She has also worked on various client configurations using LDAP, Kerberos, RBAC, PowerHA, and AIX. She is an IBM Certified System p Administrator. She has published articles in developerWorks Forum, as well.

David Kgabo is a Specialist Systems Programmer working for ABSA in South Africa. He has 15 years of experience in IT, nine of which he worked on Enterprise POWER systems. His areas of expertise include AIX, Virtualization, disaster recovery, Clustering PowerHA, and GPFS.

Ricardo Puig has been working as an AIX Support Engineer since 1998 and is a leading expert in installation and disaster recovery procedures for AIX.

Vani Ramagiri is a Virtual I/O Server Specialist in the Development Support Organization in Austin, Texas. She has 12 years of experience supporting AIX and has worked as a Lead in PowerVM since its inception in 2004. She holds a Masters degree in Computer Science from Texas State University.

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

David Bennin, Richard Conway, Don Brennan
International Technical Support Organization, Poughkeepsie Center

Bob Maher, Christopher Tulloch, Duane Witherspoon
IBM Poughkeepsie

Basu Vaidyanathan, Jayesh Patel, Daniel Henderson, Liang Jiang, Vasu Vallabhaneni,
Morgan Jeff Rosas
IBM Austin

Cesar Maciel
IBM Atlanta

Gottfried Schimunek
IBM Rochester

Ralf Schmidt-Dannert
IBM US

Priya Kannan, Kiran Grover, Saravanan Devendra, Venkatakrishnan Ganesan, Shubha Joshi,
Jaipaul K Antony
IBM India

Martin Abeleira
IBM Uruguay

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and client satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:
ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

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

ibm.com/redbooks

- Send your comments in an email to:

redbooks@us.ibm.com

- Mail your comments to:

IBM Corporation, International Technical Support Organization
Dept. HYTD Mail Station P099
2455 South Road
Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on Facebook:

<http://www.facebook.com/IBMRedbooks>

- Follow us on Twitter:

<http://twitter.com/ibmredbooks>

- Look for us on LinkedIn:

<http://www.linkedin.com/groups?home=&gid=2130806>

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

<https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm>

- Stay current on recent Redbooks publications with RSS Feeds:

<http://www.redbooks.ibm.com/rss.html>



Introducing POWER7 Enterprise Server RAS and virtualization features

In this chapter, we introduce reliability, availability, and serviceability (RAS), and virtualization concepts for the IBM Enterprise Power Systems servers.

The following topics are discussed in this chapter:

- ▶ High availability in today's business environments
- ▶ Introduction to RAS and virtualization

1.1 High availability in today's business environments

Availability is a well-established concept in today's IT environments. Tremendous growth in system capabilities, along with business demands for around the clock operations, has put increased demands on efforts to provide the highest levels of system availability.

IBM Power Systems™ servers are especially designed to help achieve high availability; however, the need for planned downtime is required for periodic maintenance (for both hardware and software) and cannot be completely eliminated. Even though both types of outages affect the overall availability of a server, we need to understand the distinctions between planned and unplanned downtime in today's business environments:

- ▶ Planned downtime is scheduled and typically is a result of a maintenance action to the hardware, operating system, or an application. Scheduled downtime is used to ensure that the server can operate optimally and reliably in the future. Because this type of event can be planned for in advance, it can be scheduled at a time that least affects system or application availability.
- ▶ Unplanned downtime is downtime that occurs as a result of a type of physical event or failure, along with human error, and cannot be planned in advance.

Understanding the causes of downtime and how the IBM Power Systems Enterprise Servers can help you address both of them is a key aspect for improving IT operations in every business.

"Despite the potential consequences of unplanned downtime, less than 10% of all downtime can be attributed to unplanned events, and only a fraction of that is due to a site disaster. The other 90+ %—the kind that companies face on a regular basis—are those caused by system maintenance tasks."

—Vision Solutions, Inc. white paper, *An introduction to System i® High Availability 2010*

The following typical system maintenance tasks are included in planned downtime:

- ▶ Data backups (nightly, weekly, and monthly)
- ▶ Reorganization of files to reclaim disk space and improve performance
- ▶ Vendor software upgrades and data conversions
- ▶ IBM software release upgrades and patches (program temporary fixes (PTFs))
- ▶ New application software installations
- ▶ Hardware upgrades
- ▶ System migrations

The number of unplanned outages continues to shrink quickly as hardware and software technology becomes more resilient. Although unplanned outages must still be eliminated, in the desire to achieve 24x365 availability, planned downtime has now become a primary focus.

1.2 Introduction to RAS and virtualization

Servers must be designed to help avoid every possible outage, focusing on applications availability. For almost two decades now, the IBM Power Systems development teams have worked to integrate industry-leading IBM System z® mainframe reliability features and capabilities into the IBM Power Systems servers line. These RAS capabilities together with the IBM Power Systems virtualization features help implement fully virtualized and highly available environments.

In the following section, we present a brief introduction to RAS and virtualization concepts for IBM Power Systems servers.

1.2.1 Reliability, availability, and serviceability (RAS)

Hardware RAS is defined this way:¹

- ▶ Reliability: How infrequently a defect or fault is seen in a server.
- ▶ Availability: How infrequently the functionality of a system or application is impacted by a fault or defect.
- ▶ Serviceability: How well faults and their impacts are communicated to users and services, and how efficiently and non-disruptively they are repaired.

Defined this way, reliability in hardware is all about how often a hardware fault requires a system to be serviced (the less frequent the failures, the greater the reliability). Availability is how infrequently such a failure impacts the operation of the system or application. For high levels of availability, correct system operation must not be adversely affected by hardware faults. A highly available system design ensures that most hardware failures do not result in application outages. Serviceability relates to identifying what fails and ensuring an efficient repair (of that component, firmware, or software).

IBM POWER7 is designed for RAS by including technologies among others to detect and isolate component faults the first time that they appear without the need to recreate the situation or perform further tests. This technology helps to minimize the risk of the same error repeating itself and causing similar or even larger problems.

Table 1-1 summarizes the RAS features that are available in the IBM POWER6® and IBM POWER7 Enterprise Servers.

Table 1-1 System support for selected RAS features (✓=capable, X=incapable)

RAS feature	Power 595	Power 780	Power 795
Processor			
Processor fabric bus protection	✓	✓	✓
Dynamic Processor Deallocation	✓	✓	✓
Dynamic Processor Sparing			
Using CoD cores	✓	✓	✓
Using capacity from spare pool	✓	✓	✓
Core Error Recovery			
Processor Instruction Retry	✓	✓	✓
Alternate Processor Recovery	✓	✓	✓
Partition core contained checkstop	✓	✓	✓
Persistent processor deallocation	✓	✓	✓
Midplane connection for inter-nodal communication	✓	X	✓
I/O subsystem			
GX+ bus persistent deallocation	✓	✓	✓
Optional ECC I/O hub with freeze behavior	✓	✓	✓
PCI bus enhanced error detection	✓	✓	✓
PCI bus enhanced error recovery	✓	✓	✓
PCI-PCI bridge enhanced error handling	✓	✓	✓
Redundant 12x Channel link	✓	✓	✓

¹ D. Henderson, J.Mitchell, G. Ahrens. "POWER7™ System RAS Key Aspects of Power Systems Reliability, Availability, and Serviceability" POW03056.doc, November 2010

RAS feature	Power 595	Power 780	Power 795
Clocks and service processor			
Dynamic SP failover at run time / Redundant SP	✓	✓	✓
Clock failover at run time / Redundant Clock	✓	✓	✓
Memory availability			
ECC in L2 and L3 cache	✓	✓	✓
Error detection/correction			
Chipkill memory plus additional 1/2 symbol correct	✓	✓	✓
Memory DRAM sparing	X ^a	✓	✓
Memory sparing with CoD at IPL time	✓	✓	✓
CRC plus retry on memory data bus (CPU to buffer)	X ^b	✓	✓
Data bus (memory buffer to DRAM) ECC plus retry	✓	✓	✓
DRAM sparing on x8+1 memory	X	✓	✓
Dynamic memory channel repair	✓	✓	✓
Processor memory controller memory scrubbing	✓	✓	✓
Memory page deallocation	✓	✓	✓
L1 parity check plus retry/set delete	✓	✓	✓
L2 cache line delete	✓	✓	✓
L3 cache line delete	✓	✓	✓
Special Uncorrectable Error handling	✓	✓	✓
Active Memory™ Mirroring for hypervisor	X	✓	✓
Fault detection and isolation			
FFDC for fault detection and isolation	✓	✓	✓
Storage Protection Keys	✓	✓	✓
Error log analysis	✓	✓	✓
Serviceability			
Boot-time progress indicators	✓	✓	✓
Firmware error codes	✓	✓	✓
Operating system error codes	✓	✓	✓
Inventory collection	✓	✓	✓
Environmental and power warnings	✓	✓	✓
PCI card hot-swap	✓	✓	✓
Hot-swap DASD/media	✓	✓	✓
Dual disk controllers/split backplane	✓	✓	✓
Extended error data collection	✓	✓	✓
I/O drawer redundant connections	✓	✓	✓
I/O drawer hot add and concurrent repair	✓	✓	✓
Hot GX adapter add and repair	X	✓	✓
Concurrent add of powered I/O rack	✓	✓	✓
SP mutual surveillance with the Power hypervisor	✓	✓	✓
Dynamic firmware update with HMC	✓	✓	✓
Service Agent Call Home application	✓	✓	✓
Service indicators – guiding light or light path LEDs	✓	✓	✓
Service processor support for BIST for logic/arrays, wire tests, and component initialization	✓	✓	✓
System dump for memory, Power hypervisor, SP	✓	✓	✓
Operating system error reporting to HMC SFP application	✓	✓	✓
RMC secure error transmission subsystem	✓	✓	✓
Health check scheduled operations with HMC	✓	✓	✓
Operator panel (real or virtual)	✓	✓	✓

RAS feature	Power 595	Power 780	Power 795
Redundant HMCs	✓	✓	✓
Automated server recovery/restart	✓	✓	✓
Hot-node add/cold node repair	✓	✓	✓
Hot-node repair/hot memory upgrade	✓	✓	✓
Hot-node repair/hot memory Add for all nodes	✓	✓	✓
PowerVM Live Partition/Live Application Mobility	✓	✓	✓
Power and cooling			
Redundant, hot swap fans and blowers for CEC	✓	✓	✓
Redundant, hot swap power supplies for CEC	✓	✓	✓
Redundant voltage regulator outputs	✓	✓	✓
TPMD/MDC for system power and thermal management	✓	✓	✓
CEC power/thermal sensors (CPU and memory)	✓	✓	✓
Redundant power for I/O drawers	✓	✓	✓

a. The Power 595 does not have the Memory DRAM sparing feature, but it has redundant bit steering.

b. In the Power 595, there is ECC on the memory bus with spare lanes.

1.2.2 Virtualization

First introduced in the 1960s, computer virtualization was created to logically divide mainframe systems to improve resource utilization. After many years of continuous evolution, IT organizations all over the world use or implement various levels of virtualization.

Built upon the Power systems RAS hardware platform, IBM virtualization features allow for great flexibility, hardware optimization, simple management, and secure and low cost hardware-assisted virtualization solutions.

The following section summarizes the available virtualization technologies for the IBM Power Systems Enterprise Servers.

IBM PowerVM

IBM PowerVM is a combination of hardware and software that enable the virtualization platform for AIX, Linux, and IBM i environments for IBM Power Systems. By implementing PowerVM you can perform these functions:

- ▶ Easily and quickly deploy new partitions.
- ▶ Execute isolated workloads for production, development, and test systems.
- ▶ Reduce costs by consolidating AIX, IBM i, and Linux workloads into one high-end IBM Power System.
- ▶ Optimize resource utilization by effectively allocating resources to those workloads that need them.
- ▶ Optimize the utilization of I/O adapters.
- ▶ Reduce the complexity and management of the environment.
- ▶ Increase your overall availability by making workloads independent of the physical hardware and by adding the capabilities to move those workloads to another server without disruption, thus eliminating planned downtime.

There are three editions of PowerVM that are suitable for these purposes:

► **PowerVM Express Edition**

PowerVM Express Edition is designed for clients looking for an introduction to virtualization features at an affordable price. The Express Edition is not available for the IBM Power Systems Enterprise Servers.

► **PowerVM Standard Edition**

PowerVM Standard Edition provides advanced virtualization functionality for AIX, IBM i, and Linux operating systems. PowerVM Standard Edition is supported on all POWER processor-based servers and includes features designed to help businesses increase system utilization.

► **PowerVM Enterprise Edition**

PowerVM Enterprise Edition includes all the features of PowerVM Standard Edition, plus two new industry-leading capabilities that are called Active Memory Sharing and Live Partition Mobility. This option provides complete virtualization for AIX, IBM i, and Linux operating systems. Active Memory Sharing intelligently flows system memory from one partition to another as workload demands change. Live Partition Mobility allows for the movement of a running partition from one server to another with no application downtime, resulting in better system utilization, improved application availability, and energy savings. With Live Partition Mobility, planned application downtime due to regular server maintenance can be a thing of the past.

Table 1-2 lists the feature codes of the PowerVM editions that are available on the Power 780 and 795 Enterprise Servers.

Table 1-2 Availability of PowerVM editions on the Power 780 and 795 Enterprise Servers

PowerVM editions	Express	Standard	Enterprise
IBM Power 780	N/A	7942	7995
IBM Power 795	N/A	7943	8002

Table 1-3 outlines the functional elements of the available PowerVM editions for both the Power 780 and 795.

Table 1-3 PowerVM capabilities and features on Power 780 and 795

PowerVM editions	Standard	Enterprise
PowerVM Hypervisor	Yes	Yes
Dynamic Logical Partitioning	Yes	Yes
Maximum partitions	1000 per server	1000 per server
Management	VMControl, HMC, and SDMC	VMControl, HMC, and SDMC
Virtual I/O server	Yes (Maximum supported 10)	Yes (Maximum supported 10)
PowerVM Lx86	Yes	Yes
Suspend/Resume	Yes	Yes
N_port ID Virtualization	Yes	Yes
Multiple Shared Processor Pool	Yes	Yes
Shared Storage Pools	Yes	Yes

PowerVM editions	Standard	Enterprise
Thin Provisioning	Yes	Yes
Active Memory Sharing	No	Yes
Live Partition Mobility	No	Yes

Hardware management Console: The IBM Power 780 and 795 must be managed with the Hardware Management Console (HMC) or the Systems Director Management Console (SDMC). The Integrated Virtualization Manager (IVM) is not supported.

The PowerVM Standard Edition can be upgraded to the PowerVM Enterprise Edition by entering a key code in the HMC. This upgrade operation is non-disruptive. If you have an existing Power 595 (machine type 9119-FHA) with PowerVM Standard Edition (feature code 7943) or PowerVM Enterprise Edition (feature code 8002), you can also migrate the licenses for PowerVM, if you migrate from a Power 595 to a Power 795.

Operating system versions supported

The following operating system versions are supported:

- ▶ AIX 5.3, AIX 6.1, and AIX 7
- ▶ IBM i 6.1 and IBM i 7.1
- ▶ Red Hat Enterprise Linux 5 and Red Hat Enterprise Linux 6
- ▶ SUSE Linux Enterprise Server 10 and SUSE Linux Enterprise Server 11

Table 1-4 summarizes the PowerVM features that are supported by the operating systems that are compatible with the POWER7 processor-based servers.

Table 1-4 PowerVM features supported by AIX, IBM i, and Linux on Power 780 and 795

Feature	AIX 5.3	AIX 6.1	AIX 7.1	IBM i 6.1.1	IBM i 7.1	RHEL 5.5	SLES10 SP3	SLES11 SP1
Simultaneous Multi-Threading (SMT)	Yes ^a	Yes ^b	Yes	Yes ^c	Yes	Yes ^a	Yes ^a	Yes
Dynamic LPAR I/O adapter add/remove	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dynamic LPAR processor add/remove	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dynamic memory add	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dynamic memory remove	Yes	Yes	Yes	Yes	Yes	No	No	Yes
Capacity on Demand	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Micro-partitioning	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Shared Dedicated Capacity	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Multiple Shared Processor Pools	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Virtual I/O server	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Feature	AIX 5.3	AIX 6.1	AIX 7.1	IBM i 6.1.1	IBM i 7.1	RHEL 5.5	SLES10 SP3	SLES11 SP1
Virtual SCSI	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Virtual Ethernet	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N_Port ID Virtualization (NPIV)	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Live Partition Mobility	Yes	Yes	Yes	No	No	Yes	Yes	Yes
Workload Partitions	No	Yes	Yes	No	No	No	No	No
Active Memory Sharing	Yes	Yes	Yes	Yes	Yes	No	No	Yes
Active Memory Expansion	No	Yes ^d	Yes	No	No	No	No	No

a. Only supports two threads.

b. AIX 6.1 up to TL4 SP2 only supports two threads, and supports four threads as of TL4 SP3.

c. IBM i 6.1.1 and up support SMT4.

d. On AIX 6.1 with TL4 SP2 and later.

You can obtain additional information about the PowerVM Editions at the IBM PowerVM Editions website:

<http://www.ibm.com/systems/power/software/virtualization/editions/index.html>

You can obtain detailed information about the use of PowerVM technology in the following IBM Redbooks publications:

- ▶ *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940-04
- ▶ *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590

Refer to Chapter 2, “Exploring RAS and virtualization features in more detail” on page 11 for detailed PowerVM information.

Other virtualization features

In addition to the PowerVM features, the IBM POWER7 Systems™ Enterprise Servers introduce Active Memory Expansion (AME) and also the Integrated Virtual Ethernet (IVE) adapter, which were previously available in the low and midrange Power servers.

We cover both AME and IVE in detail in 2.1, “New RAS and virtualization features with POWER7” on page 13.

1.2.3 Latest available feature enhancements

The latest available virtualization feature contains the following enhancements:

- ▶ LPAR maximums increased up to 1000 partitions per server as shown in Table 1-5.

Table 1-5 High-end Power Systems features

POWER7 model	Maximum cores	Original maximum LPARs	May 2011 maximum LPARs
780	64	256	640
795	256	256	1000

Requirement: Using maximum LPARs requires PowerVM Standard or PowerVM Enterprise and the latest system firmware, which is 730_035 or later.

► Trial PowerVM Live Partition Mobility

This feature enables a client to evaluate Live Partition Mobility at no-charge for 60 days. At the conclusion of the trial period, clients can place an upgrade order for a permanent PowerVM Enterprise Edition to maintain continuity. At the end of the trial period (60 days), the client's system automatically returns to the PowerVM Standard Edition. Live Partition Mobility is available only with PowerVM Enterprise Edition. It allows for the movement of a running partition from one Power System server to another with no application downtime, resulting in better system utilization, improved application availability, and energy savings. With Live Partition Mobility, planned application downtime due to regular server maintenance is a challenge of the past.

Requirement: This is a 60-day trial version of PowerVM Enterprise Edition. Using this trial version requires PowerVM Standard Edition and firmware 730_035 or later.

Archived

Exploring RAS and virtualization features in more detail

Each successive generation of IBM servers is designed to be more reliable than the previous server family. The IBM POWER7 processor-based servers have new features to support new levels of virtualization, help ease administrative burden, and increase system utilization.

POWER7 Enterprise Servers use several innovative technologies that offer industry-leading processing speed and virtualization capabilities while using less energy and operating at a lower cost per transaction.

In this chapter, we investigate in more detail the new POWER7 reliability, availability, and serviceability (RAS) features, along with other significant RAS and virtualization features. You will become familiar with their benefits and understand how these capabilities strengthen your overall IBM Power Systems server availability environment. Figure 2-1 shows the additional features that are available only on the POWER7 Enterprise Server 780 and 795.

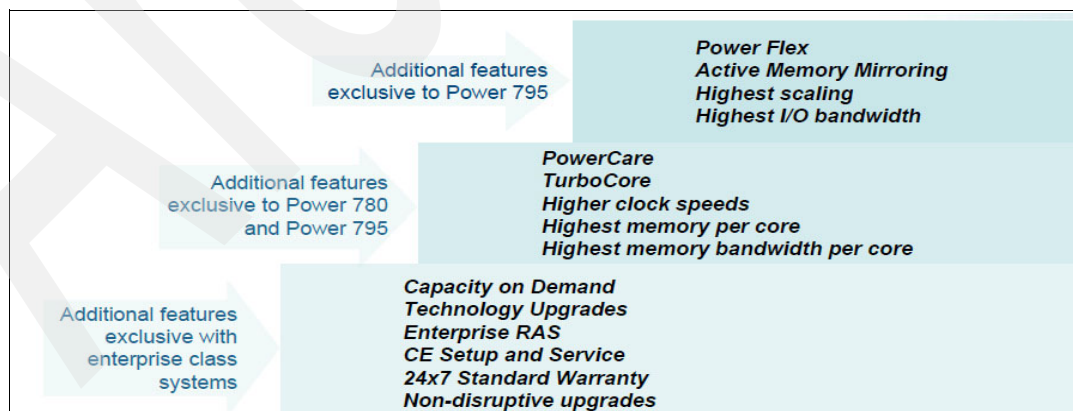


Figure 2-1 Additional exclusive features on POWER7 Enterprise Servers

In the following sections, we discuss the key features that IBM Power Systems provide in detail, providing guidelines for implementing these features to take full advantage of their capabilities.

We discuss the following topics in this chapter:

- ▶ New RAS and virtualization features with POWER7
- ▶ Significant features
- ▶ TurboCore and MaxCore technology
- ▶ Hypervisor and firmware technologies
- ▶ Power management
- ▶ Rapid deployment of PowerVM clients
- ▶ I/O considerations
- ▶ Active Memory Sharing
- ▶ Integrated Virtual Ethernet
- ▶ Partitioning

2.1 New RAS and virtualization features with POWER7

A number of RAS and virtualization features are introduced with POWER7 servers. In this section, we discuss the following features in more detail:

- ▶ Active Memory Mirroring for the hypervisor
- ▶ Hot GX adapter add/repair
- ▶ Improved memory RAS features
- ▶ Active Memory Expansion (AME)
- ▶ Hibernation or suspend/resume (refer to 3.3, “Partition hibernation” on page 77)

For more in-depth information about POWER7 RAS features, see the *POWER7 System RAS Key Aspects of Power Systems Reliability, Availability, and Serviceability* white paper at this website:

<http://www-03.ibm.com/systems/power/hardware/whitepapers/ras7.html>

2.1.1 Active Memory Mirroring for the hypervisor on Power 795

Active Memory Mirroring for the hypervisor is a new RAS function that is provided with POWER7 and is only available on the Power 795 server. This feature is also sometimes referred to as *system firmware mirroring*. Do not confuse it with other memory technologies, such as Active Memory Sharing and Active Memory Expansion, which are discussed in 2.1.4, “Active Memory Expansion” on page 19 and 2.8, “Active Memory Sharing” on page 45.

Active Memory Mirroring for the hypervisor is designed to mirror the main memory that is used by the system firmware to ensure greater memory availability by performing advance error-checking functions. This level of sophistication in memory reliability on Power systems translates into outstanding business value. When enabled, an uncorrectable error that results from a failure of main memory used by the system firmware will not cause a system-wide outage. The system maintains two identical copies of the system hypervisor in memory at all times, as shown in Figure 2-2.

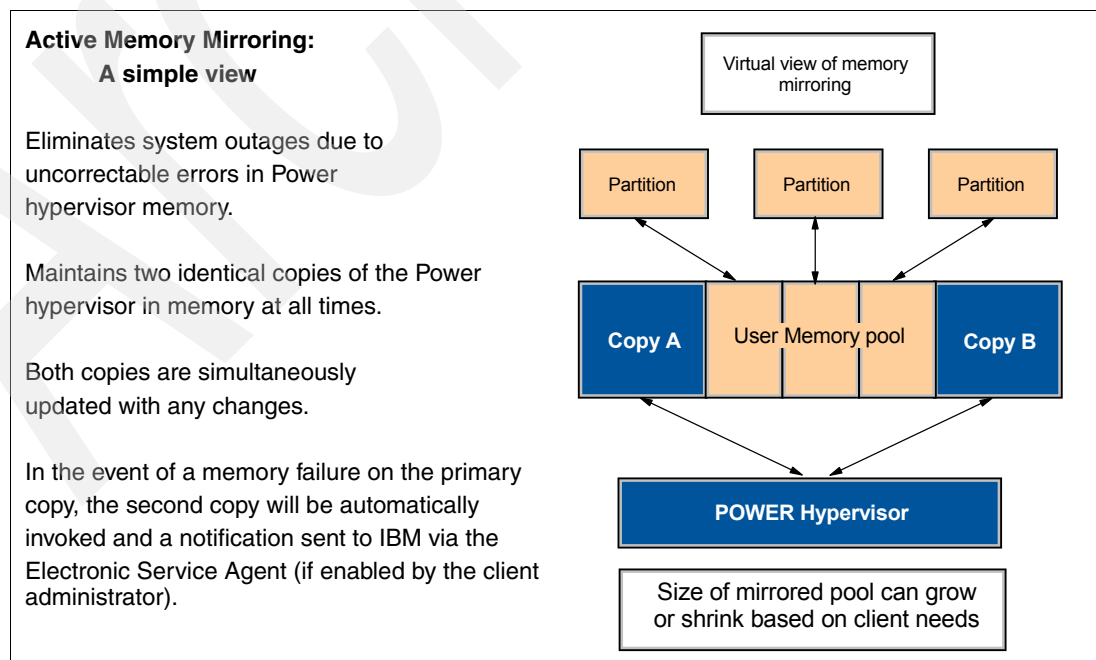


Figure 2-2 A simple view of Active Memory Mirroring

When a failure occurs on the primary copy of memory, the second copy is automatically invoked and a notification is sent to IBM via the Electronic Service Agent™ (ESA). Implementing the Active Memory Mirroring function requires additional memory; therefore, you must consider this requirement when designing your server. Depending on the system I/O and partition configuration, between 5% and 15% of the total system memory is used by hypervisor functions on a system on which Active Memory Mirroring is not being used. Use of Active Memory Mirroring for the hypervisor doubles the amount of memory that is used by the hypervisor, so appropriate memory planning must be performed. The System Planning Tool (SPT) can help estimate the amount of memory that is required. See Chapter 4, “Planning for virtualization and RAS in POWER7 high-end servers” on page 99 for more details.

Active Memory Mirroring for the hypervisor is provided as part of the hypervisor, so there is no feature code that needs to be ordered that provides this function. The feature is enabled by default on a Power 795 server. An optimization tool for memory defragmentation is also included as part of the Active Memory Mirroring feature.

Disabling Active Memory Mirroring: Active Memory Mirroring can be disabled on a system if required, but you must remember that disabling this feature leaves your Power server exposed to possible memory failures that can result in a system-wide outage.

The only requirement of a Power 795 system to support Active Memory Mirroring is that in each node at least one processor module must be fully configured with eight dual inline memory modules (DIMMs). Figure 2-3 shows the layout of a processor book and its components.

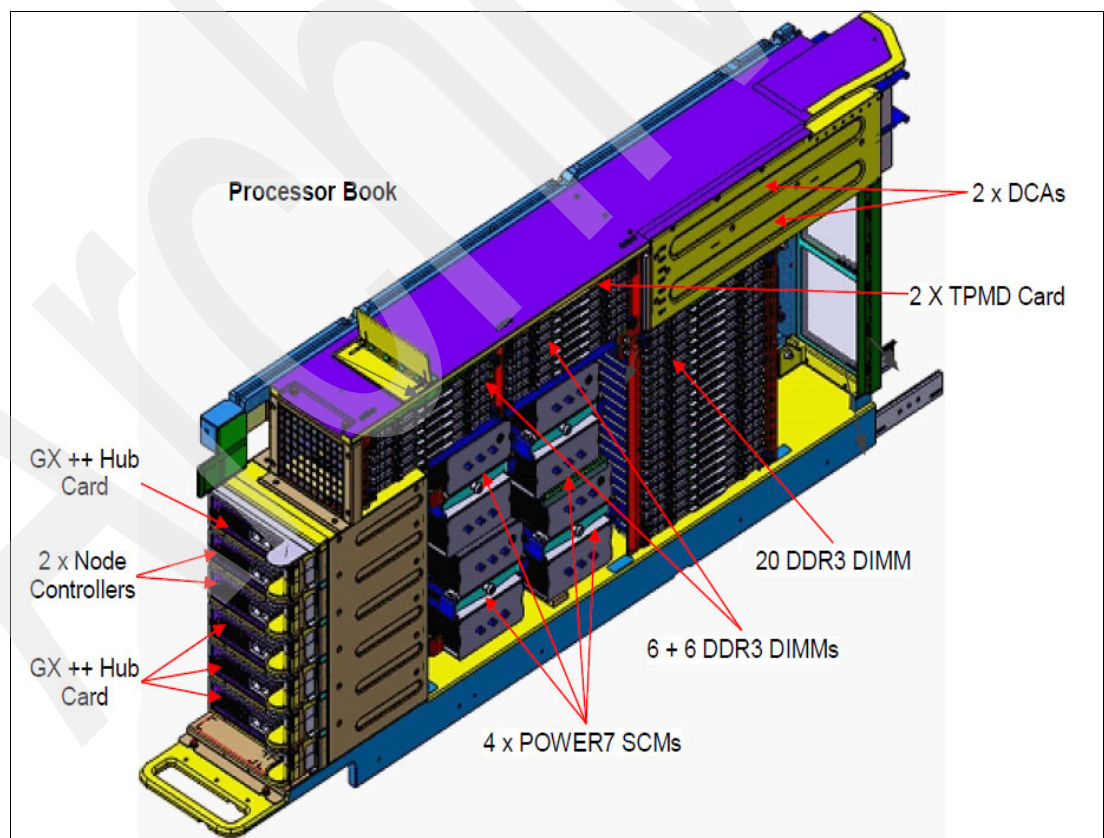


Figure 2-3 A POWER7 processor book and its components

Beginning with Hardware Management Console (HMC) V7 R7.2.0, new commands have been added for the Active Memory Mirroring support:

- ▶ `optmem -x<sysX> -o start -t mirror -q <xxxMB> --minutes<timeout>`
- ▶ `optmem -x<sysX> -o stop`
- ▶ `lsmemopt -m<sysX>`

This command lists the status and progress information of the most recent defragmentation operation.

The `lshwres` command on the HMC, which lists the hardware resources of the managed system, has been enhanced to support Active Memory Mirroring on the IBM POWER7 servers only and specifically the Power 795. Also, the `chhwres` command, which dynamically changes the hardware resource configuration, supports Active Memory Mirroring. The following commands are also valid on the IBM Systems Director Management Console (SDMC) using the command-line interface (CLI). Each command is preceded by `smcli`:

- ▶ `smcli optmem -x<sysX> -o start -t mirror -q <xxxMB> --minutes<timeout>`
- ▶ `smcli optmem -x<sysX> -o stop`
- ▶ `smcli lsmemopt -m<sysX>`

You also have the option of configuring Active Memory Mirroring via the Advanced System Management Interface (ASMI) interface. To perform this operation, you must have one of the following authority levels:

- ▶ Administrator
- ▶ Authorized service provider

To configure Active Memory Mirroring, perform the following steps:

1. On the ASMI Welcome pane, specify your user ID and password, and click **Log In**.
2. In the navigation area, expand **System Configuration** → **Selective Memory Mirroring**.
3. In the right pane, select the Requested mode (Enabled or Disabled) and click **Save settings**, as shown in Figure 2-4.

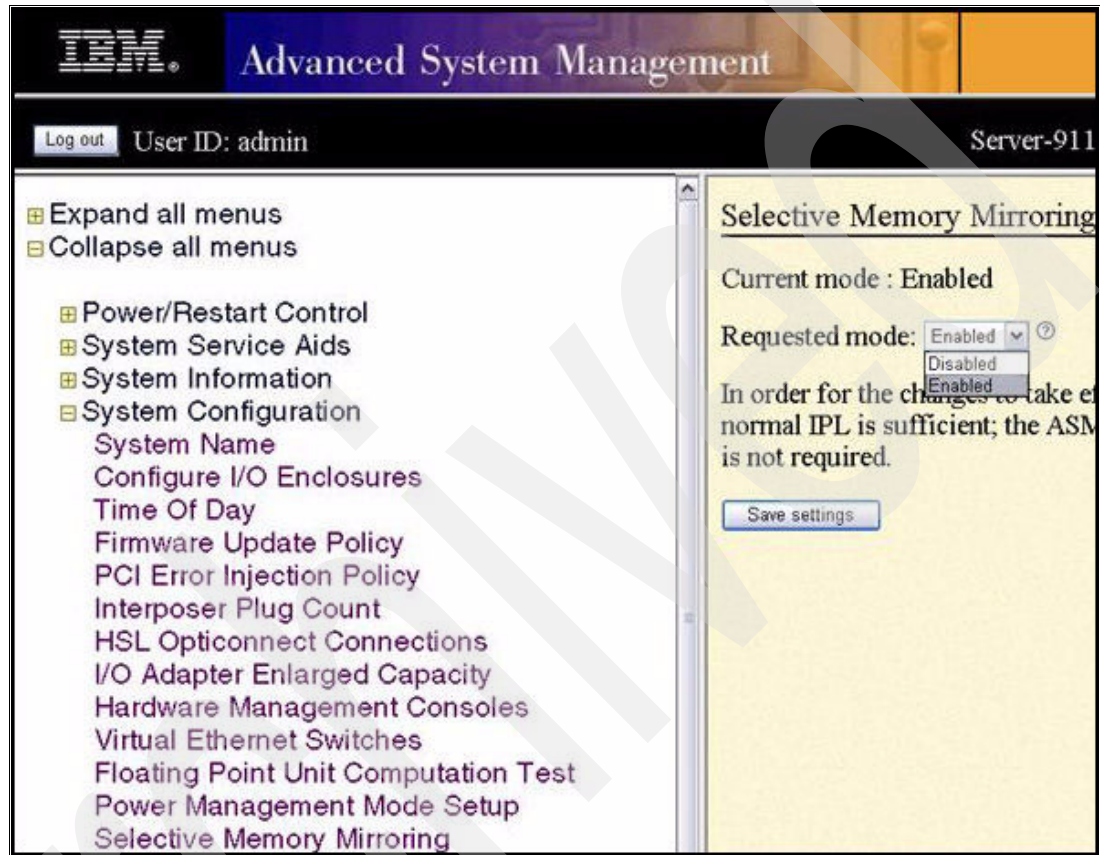


Figure 2-4 Memory Mirroring enablement via the ASMI interface

2.1.2 Hot GX adapter repair

The IBM GX host channel adapter (HCA) provides server connectivity to InfiniBand fabrics and I/O drawers. The POWER7 server provides the following GX adapter capabilities:

- ▶ GX+ adapters run at 5 GB/second
- ▶ GX++ adapters run at 20 GB/second

Concurrent maintenance has been available on Power Systems since 1997. POWER6 (2007-2009) introduced the ability to have Central Electronic Complex (CEC) concurrent maintenance functions. The CEC consists of the processor, memory, systems clocks, I/O hubs, and so on. Hot GX adapter ADD with COLD repair has been a RAS feature since POWER6, but we did not have the capability to perform Hot GX adapter repair until POWER7. Hot GX adapter repair enables the repair and replacement of the component with reduced impact to systems operations:

- ▶ *Cold Repair:* The hardware being repaired is electrically isolated from the system.
- ▶ *Hot Repair:* The hardware being repaired is electrically connected to the system.

With POWER7, we introduced CEC Hot Add Repair Maintenance (CHARM) for Power 780 and Power 795 servers. CHARM offers new capabilities in reliability, availability, and serviceability (RAS). Hot GX adapter repair enables the repair and replacement of the component with reduced impact to systems operations, if all prerequisites have been met.

Important: Accomplishing CHARM requires careful advance planning and meeting all the prerequisites.

CHARM operations are complex and, therefore, require the following additional prerequisites:

- ▶ **Off-peak schedule:** It is highly recommended that repairs are done during non-peak operational hours.
- ▶ **Redundant I/O:** It is a prerequisite that all I/O resources are configured with redundant paths. Redundant I/O paths need to be configured through separate nodes and GX adapters. Redundant I/O adapters must be located in separate I/O expansion units that are attached to separate GX adapters that are located in separate nodes.

Redundant I/O can be either directly attached I/O or virtual I/O that is provided by dual VIO servers housed in separate nodes.

- ▶ **ESA must be enabled:** Electronic Service Agent (ESA) must be enabled on the POWER7 systems. ESA systems show decreased unscheduled repair actions and provides invaluable statistics to gauge field performance.
- ▶ **Quiesce/Live Partition Mobility (LPM) critical applications:** Critical applications must be quiesced or moved to another server using LPM, if available.

Hardware concurrent maintenance entails numerous steps that are performed by both you and your IBM service personnel while the system is powered on. The likelihood of failure increases with the complexity of the maintenance function. Therefore, IBM recommends that all hardware concurrent maintenance operations be performed during off-peak hours.

The “Prepare for Hot Repair/Upgrade” (PHRU) utility on the HMC must be run by the system administrator to determine the processor, memory, and I/O resources that must be freed up prior to the start of the concurrent repair operation.

Important: All serviceable hardware events must be repaired and closed before starting an upgrade.

The Prepare for Hot Repair, Upgrade HMC, or SDMC utility is a tool for the system administrator to identify the effects to system resources in preparation for a hot node repair, hot node upgrade, or hot GX adapter repair operation. The utility provides an overview of platform conditions, partition I/O, and processor and memory resources that must be freed up for a node evacuation.

Figure 2-5 displays the HMC Partitions tab showing error messages for the AIX resources.

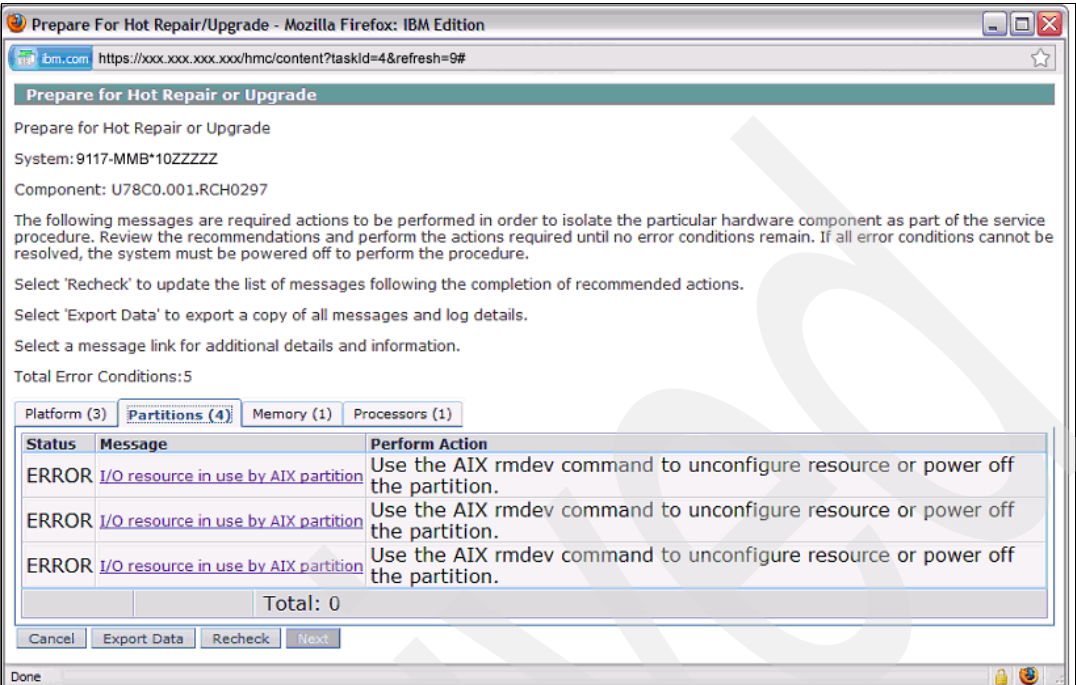


Figure 2-5 Prepare for hot repair/upgrade utility

Figure 2-6 shows a message about the I/O resource in use by the AIX partition.

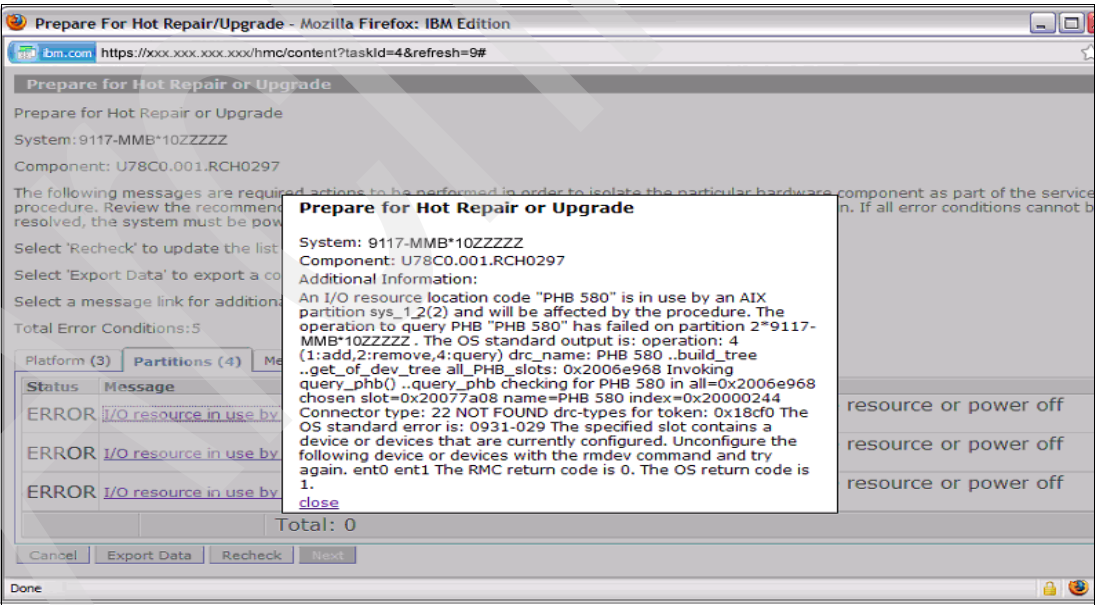


Figure 2-6 Details for the I/O resource in use by the AIX partition

You can obtain detailed information about the CHARM process in 4.3, “CEC Hot Add Repair Maintenance (CHARM)” on page 121.

2.1.3 Improved memory RAS features

The following section provides details about the improved memory RAS features:

- **Chipkill**

Chipkill is an enhancement that enables a system to sustain the failure of an entire dynamic random access memory (DRAM) chip. An error correction code (ECC) word uses 18 DRAM chips from two DIMM pairs, and a failure on any of the DRAM chips can be fully recovered by the ECC algorithm. The system can continue indefinitely in this state with no performance degradation until the failed DIMM can be replaced.

- **72-byte ECC (cyclic redundancy check (CRC) plus retry on memory data bus (CPU to buffer)**

In POWER7, an ECC word consists of 72 bytes of data. Of these, 64 bytes are used to hold application data. The remaining eight bytes are used to hold check bits and additional information about the ECC word.

This innovative ECC algorithm from IBM research works on DIMM pairs on a rank basis (a rank is a group of 10 DRAM chips on the Power 795). With this ECC code, the system can dynamically recover from an entire DRAM failure (chipkill). It can also correct an error even if another symbol (a byte, accessed by a 2-bit line pair) experiences a fault. This capability is an improvement from the Double Error Detection/Single Error Correction ECC implementation that is on the POWER6 processor-based systems.

- **DRAM sparing**

IBM Power 780 and 795 servers have a spare DRAM chip per rank on each DIMM that can be used to replace a failed DIMM in a rank (chipkill event). Effectively, this protection means that a DIMM pair can sustain two, and in certain cases, three DRAM chip failures and correct the errors without any performance degradation.

2.1.4 Active Memory Expansion

First introduced in POWER7, Active Memory Expansion (AME) is an innovative IBM technology that enables the memory assigned to a logical partition (LPAR) to expand beyond its physical limits.

AME relies on the real-time compression of data stored in memory to increase the amount of available memory. When AME is enabled, the operating system compresses a portion of the real memory, generating two pools: compressed and uncompressed memory. The size of each pool varies, according to the application's requirements. This process is completely transparent to users and applications.

Processing capability: Because AME relies on memory compression by the operating system, additional processing capacity is required to use AME.

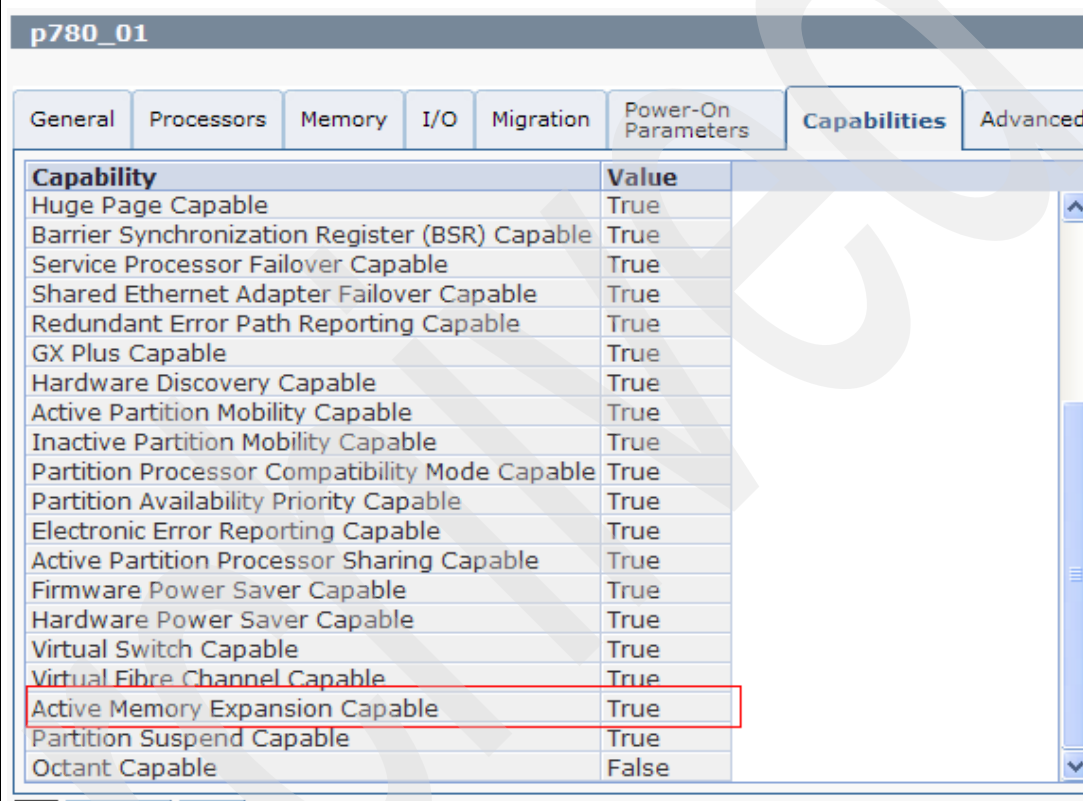
After using AME, the system's effective memory increases. Then, you are able to perform these functions:

- **Optimize memory utilization:** By consolidating larger workloads using less real memory than needed
- **Increase an LPAR's throughput:** By allowing a single LPAR to expand its memory capacity beyond the physical memory assigned

Active Memory Expansion License

AME is a special feature that needs to be licensed before it can be used. Check your system configuration for the AME feature:

1. Log in into the server's HMC.
2. In the navigation pane, expand **Management** → **Servers**, and select the system that you want to check.
3. Open the Properties page for the selected server.
4. Check the Capabilities tab, as shown in Figure 2-7. The Active Memory Expansion Capable value must be set to True.



p780_01		
General Processors Memory I/O Migration Power-On Parameters Capabilities Advanced		
Capability	Value	
Huge Page Capable	True	
Barrier Synchronization Register (BSR) Capable	True	
Service Processor Failover Capable	True	
Shared Ethernet Adapter Failover Capable	True	
Redundant Error Path Reporting Capable	True	
GX Plus Capable	True	
Hardware Discovery Capable	True	
Active Partition Mobility Capable	True	
Inactive Partition Mobility Capable	True	
Partition Processor Compatibility Mode Capable	True	
Partition Availability Priority Capable	True	
Electronic Error Reporting Capable	True	
Active Partition Processor Sharing Capable	True	
Firmware Power Saver Capable	True	
Hardware Power Saver Capable	True	
Virtual Switch Capable	True	
Virtual Fibre Channel Capable	True	
Active Memory Expansion Capable	True	
Partition Suspend Capable	True	
Octant Capable	False	

Figure 2-7 AME-capable server

AME feature: If the value is False, you need to obtain a license for the AME feature, or you can request a free 60-day trial at this website:

https://www-912.ibm.com/tcod_reg.nsf/TrialCod?OpenForm

Expansion factor

When using AME, you only need to perform one configuration, the memory expansion factor. This parameter specifies the new amount of memory that is available for a specific LPAR and thus defines how much memory the system tries to compress.

The new LPAR memory size is calculated this way:

$$\text{LPAR_expanded_memory_size} = \text{LPAR_true_memory_size} * \text{LPAR_expansion_factor}$$

The expansion factor is defined on a per LPAR basis using the HMC, as shown in Figure 2-8.

Logical Partition Profile Properties: default @ lpar1_p780_1 @ p780_01 - lpar1_p780_1

General

Processors

Memory

I/O

Virtual Adapters

Power Controlling

Settings

Logical Host Ethernet Adapters (LHEA)

Detailed below are the current memory settings for this partition profile.

Dedicated Memory

Installed memory (MB):262144

Current memory available for partition usage (MB) : 129024

Minimum memory : 0 GB512 MB

Desired memory : 1 GB0 MB

Maximum memory : 2 GB0 MB

Specify the Barrier Synchronization Register BSR for this profile

Available BSR arrays:256

BSR arrays for this profile:0

Huge Page Memory

Page size (in GB) :16

Configurable pages :0

Minimum pages : 0

Desired pages : 0

Maximum pages : 0

Active Memory Expansion

☒ Active memory expansion factor (1.00 - 10.00)1.2

Figure 2-8 Active Memory Expansion factor setting

AME: When considering AME, use the amepat tool to determine the best expansion factor for your specific workload.

Chapter 2. Exploring RAS and virtualization features in more detail21

Figure 2-9 presents a scenario where memory expansion is used in a 20 GB RAM LPAR using an expansion factor of 1.5.

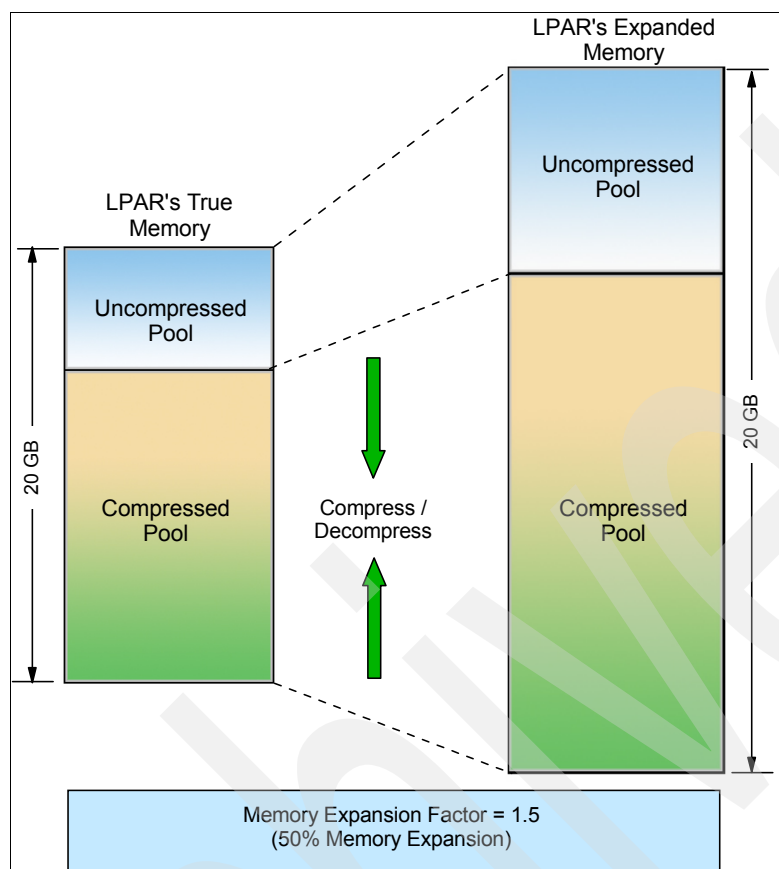


Figure 2-9 Active Memory Expansion example

2.2 Significant features

In this section, we describe previous features that play a key role in the POWER7 server RAS and virtualization strategy.

2.2.1 Active Memory Mirroring for the hypervisor on the Power 795

Active Memory Mirroring for the hypervisor is a new RAS feature being introduced on the Power 795 that is designed to eliminate the potential for a complete system outage as a result of an uncorrectable error in memory.

2.2.2 Persistent hardware deallocation

For overall system availability purposes, a component that is identified as failing on a POWER processor-based system is flagged for persistent deallocation. Component removal can occur dynamically or at boot time, depending in the type of fault and the moment that the fault is detected.

By deallocating failed components, we prevent faulty hardware from affecting the entire system operation. The repair action is deferred to a more convenient, less critical time. The

affected components for this function are processors, L2/L3 cache lines, memory, and I/O adapters.

2.2.3 First Failure Data Capture (FFDC)

IBM Power Systems servers First Failure Data Capture ensures that when a fault is detected in the system, the root cause is isolated in the first appearance of the problem without needing any additional tests or reproducing the problem. FFDC relies on built-in checkers to capture and identify error conditions. FFDC is the technique that is used to check all the components in the system Central Electronic Complex (CEC): processors, memory buffers, and I/O controllers.

First Failure Data Capture (FFDC) is a serviceability technique where a program that detects an error preserves all the data that is required for the subsequent analysis and resolution of the problem. The intent is to eliminate the need to wait for or to force a second occurrence of the error to allow specially applied traps or traces to gather the data that is required to diagnose the problem.

AIX V5.3 TL3 introduced the First Failure Data Capture (FFDC) capabilities. The set of FFDC features is further expanded in AIX V5.3 TL5 and AIX V6.1. The following features are described in the following sections:

- ▶ Lightweight Memory Trace (LMT)
- ▶ Run-Time Error Checking (RTEC)
- ▶ Component Trace (CT)
- ▶ Live Dump

These features are enabled by default at levels that provide valuable FFDC information with minimal performance effects. The advanced FFDC features can be individually manipulated. Additionally, a SMIT dialog has been provided as a convenient way to persistently (across reboots) disable or enable the features through a single command. To enable or disable all four advanced FFDC features, enter the following command:

```
#smitty ffdc
```

This SMIT dialog specifies whether the advanced memory tracing, live dump, and error checking facilities are enabled or disabled. Note that disabling these features reduces system RAS.

2.2.4 Processor RAS features

POWER-based servers are designed to recover from processor failures in many scenarios. When a soft or transient failure is detected in a processor core, the processor instruction retry algorithm retries the failed instruction in the same processor core. If that failure becomes a solid or persistent failure, the alternate processor recovery algorithm tries to execute the failed instruction in another processor core.

If the systems detects the existence of an error-prone processor, it takes the failed processor out of service before it causes an unrecoverable system error. This process is called *dynamic processor deallocation*. This features relies on the service processor's ability to use FFDC algorithms to notify the hypervisor of a failing processor. The Power hypervisor then deconfigures the failing processor.

While dynamic processor deallocation can potentially reduce the overall system performance, it can be coupled with dynamic processor sparing to automatically replace the failed processor. This entire process is transparent to the partitions.

Similar to the alternate processor recovery technique, dynamic processor sparing tries to get a free processor from the capacity on demand (CoD) pool. If not available, it uses an unused available processor (from both shared processor pools or dedicated processor partitions). If it cannot find an available processor, the hypervisor tries to release a processor from one active partition based on the partition availability priority settings.

The POWER7 processor also provides the single processor checkstop functionality (already present in POWER6). This feature is invoked in the case that neither of the techniques described is able to manage the existing processor error. With single processor checkstop, the system reduces the probability of one failed processor to affect the overall system availability by containing the processor checkstop to the partition using the failing processor.

For more information: For a more detailed explanation of this process, see the white paper *POWER7 System RAS Key Aspects of Power Systems Reliability, Availability, and Serviceability*.

<http://www-03.ibm.com/systems/power/hardware/whitepapers/ras7.html>

Partition availability priority

POWER6 and POWER7 systems allow systems administrators to specify the availability priority of their partition. If an alternate processor recovery event requires a spare processor and there is no way to obtain the spare resource, and the system needs to shut down a partition to maintain the server's overall availability, the process selects the partitions with the lowest availability priority. Priorities are assigned from 0 - 255 with 255 being the highest priority (most critical partition). The partition availability priority default for a normal partition is 127 and for a VIO server is 191.

Partition availability priority settings define critical partitions so that the hypervisor can select the best reconfiguration option after a process deallocation without sparing.

To check or modify the partition availability priority (refer to Figure 2-10) for a specific server, use the following steps:

1. Log in into the system HMC.
2. In the navigation pane, select **Systems Management** → **Servers** and select your server.
3. Under the Configuration menu, select **Partition Availability Priority**.

Partition Availability Priority: p780_01

You can change the partition availability priority for the following partitions by first selecting one or more partitions and then choosing an availability priority from the field below the table. Click OK to submit your changes.

Select	Partition Name	Partition Type	Processing units	Processing Mode	Availability priority
<input checked="" type="checkbox"/>	lpar1_p780_1	AIX or Linux	0.2	Shared	127
<input type="checkbox"/>	vios1_p780_1	Virtual I/O Server	2.0	Shared	191
<input type="checkbox"/>	vios2_p780_1	Virtual I/O Server	2.0	Shared	191

Availability priority:

Figure 2-10 Partition availability priority settings

2.2.5 Memory RAS features

As defined in the *IBM Power 770 and 780 (9117-MMB, 9179-MHB) Technical Overview and Introduction*, REDP-4639-00, and the *IBM Power 795 (9119-FHB) Technical Overview and Introduction*, REDP-4640-00, IBM POWER7-based systems include a variety of protection methods, which were already present in POWER6 servers, that are designed to prevent, protect, or limit the consequences of memory errors in the system.

Memory error detection schemes

The following methods are the memory error detection schemes:

- ▶ Hardware scrubbing

Hardware scrubbing is a method that is used to deal with intermittent errors. IBM POWER processor-based systems periodically address all memory locations; any memory locations with a correctable error are rewritten with the correct data.

- ▶ CRC

The bus that transfers data between the processor and the memory uses CRC error detection with a failed operation-retry mechanism and the ability to dynamically return bus parameters when a fault occurs. In addition, the memory bus has spare capacity to substitute a data bit-line, whenever it is determined to be faulty.

Memory page deallocation

IBM POWER processor systems can contain cell errors in memory chips using memory page deallocation. When a memory address experiences repeated correctable errors or an uncorrectable error, the service processor notifies the hypervisor and the memory page is marked for deallocation. The operating system that uses the page is asked to move data to another memory page, then the page is deallocated and no longer can be used by any partition or the hypervisor. This action does not require user intervention. If the page is owned by the hypervisor, it is deallocated as soon as the hypervisor releases that page.

The hypervisor maintains a list of deallocated pages. The list is used to continuously deallocate pages upon system or partition reboots.

Memory persistent deallocation

At boot time, during system self-test, defective memory is deallocated and is not used in subsequent reboots. If the server has available CoD memory, the hypervisor attempts to replace the faulty memory with CoD unlicensed memory, and if properly configured, the system triggers a service call. After memory deallocation, a service repair action needs to be scheduled to replace the failed memory chips.

If, after system reboot, the amount of memory configured does not allow the system to activate one or more partitions, the hypervisor reduces the memory assigned to one or more partitions, based on the partition availability priority setting.

2.2.6 Dynamic service processor (SP) failover at run time and redundant SP

The ability to have redundant service processors (SP) to address demanding availability requirements continues today with the Power 795 and the Power 780 systems with multi-node capability (2 or more CECs). The redundant service processor capability enables you to configure a secondary service processor that is activated when the primary service processor fails. You must have either a Hardware Management Console (HMC) or System Director Management Console (SDMC) to enable and disable the redundant service processor capability.

Figure 2-11 depicts the correct configuration for redundant service processors and redundant HMCs.

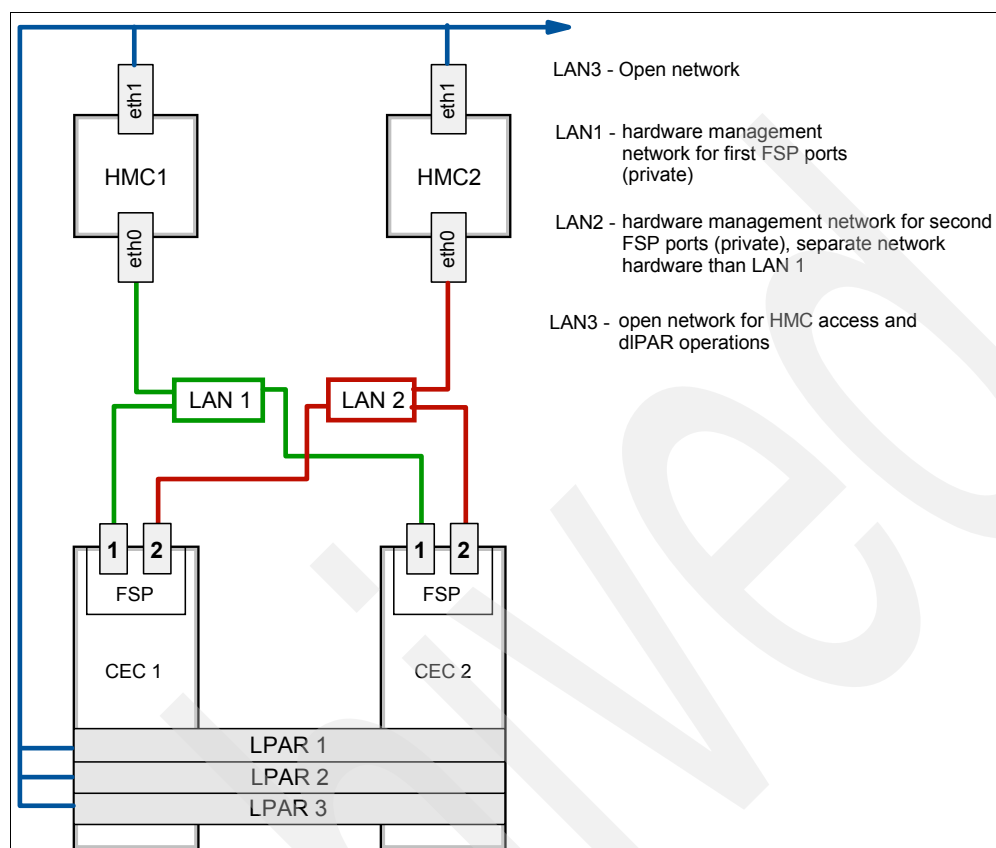


Figure 2-11 A correct service processor and HMC configuration

One HMC must connect to the port labeled as HMC Port 1 on the first two CEC drawers of each system. A second HMC must be attached to HMC Port 2 on the first two CEC drawers of each system. This type of solution provides redundancy for the service processors and the HMCs.

It is important to understand exactly what a service processor is and the functions that it provides to fully appreciate why having redundant SP with dynamic service processor failover capability is an important RAS feature for POWER7 Enterprise Servers.

The *service processor* is a dedicated microprocessor that is independent of the other POWER7 microprocessors and is separately powered. Its main job is to correlate and process error information that is received from other system components and to engineer error recovery mechanisms along with the hardware and the Power hypervisor. The service processor uses error “thresholding” and other techniques to determine when corrective action needs to be taken. *Thresholding*, as defined for Power RAS, is the ability to use historical data and engineering expertise that is built into the feature to count recoverable errors and accurately predict when corrective actions must be initiated by the system. Power systems require a service processor to perform system power-on to initialize the system hardware and to monitor error events during operation.

The service processor (SP) and the Power hypervisor work together to monitor and detect errors. While the service processor is monitoring the operation of the Power hypervisor firmware, the Power hypervisor monitors the service processor activity. The service processor

can take the proper action, which includes calling IBM for service, when it detects that the Power hypervisor has lost control.

Similarly, the Power hypervisor automatically performs a reset and reload of the service processor when it detects an error. A service processor reset/reload is not disruptive and does not effect system operations. SP resets can also be initiated by the service processor itself when necessary. When a SP does not respond to a reset request or the reset/reload threshold is reached, the system dynamically performs a failover from one service processor to the secondary SP during run time.

Service processor failover is enabled through an HMC or SDMC. The default for new systems is to enable automatic failover if the primary service processor fails.

Important: Verify that service processor redundancy is enabled on your server.

To enable/disable service processor redundancy failover on your managed system on the HMC, in the navigation area of your managed server, select **Serviceability** → **FSP Failover** → **Setup**, as seen in Figure 2-12.

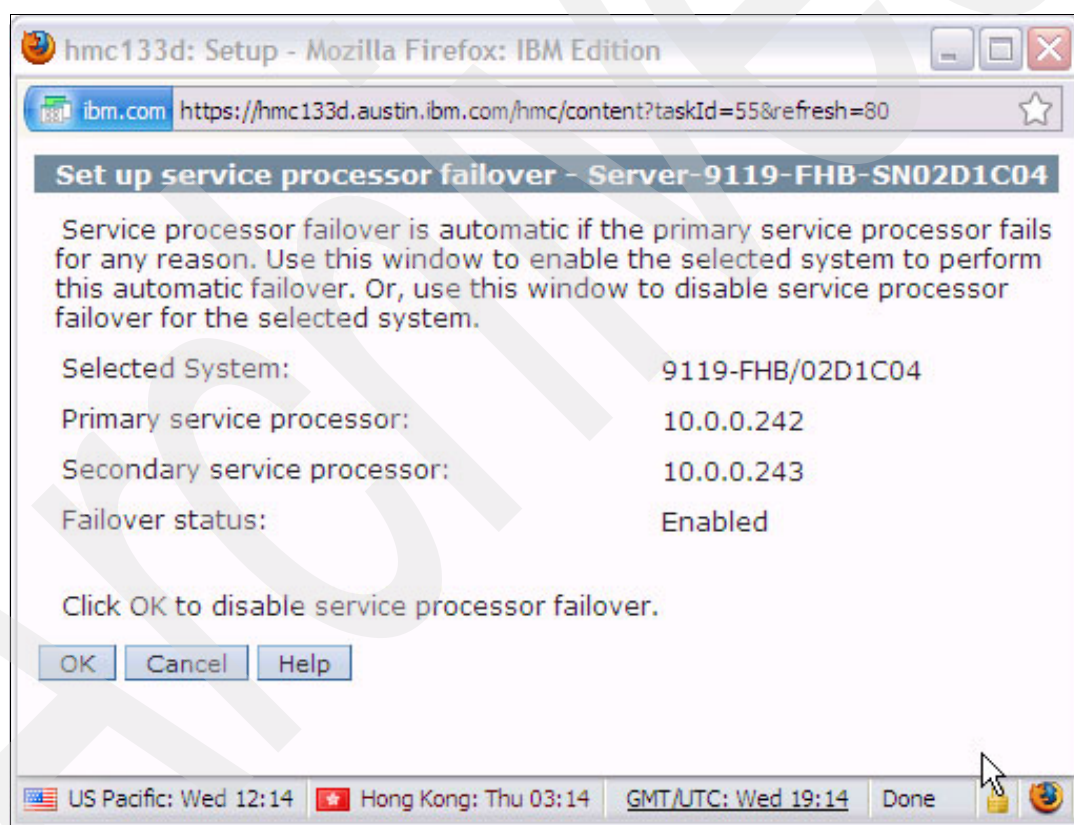


Figure 2-12 HMC enable/disable service processor failover pane

To verify that the service processor redundancy is enabled on an SDMC, select your server by clicking **Navigate Resources** → **Actions** → **System Configuration** → **Edit host** → **Capabilities**.

One of the prerequisites for the ability to perform CEC Hot Add and Repair maintenance is that the service processor failover must be enabled (see Figure 2-13). See Chapter 4, “Planning for virtualization and RAS in POWER7 high-end servers” on page 99 for a detailed overview.

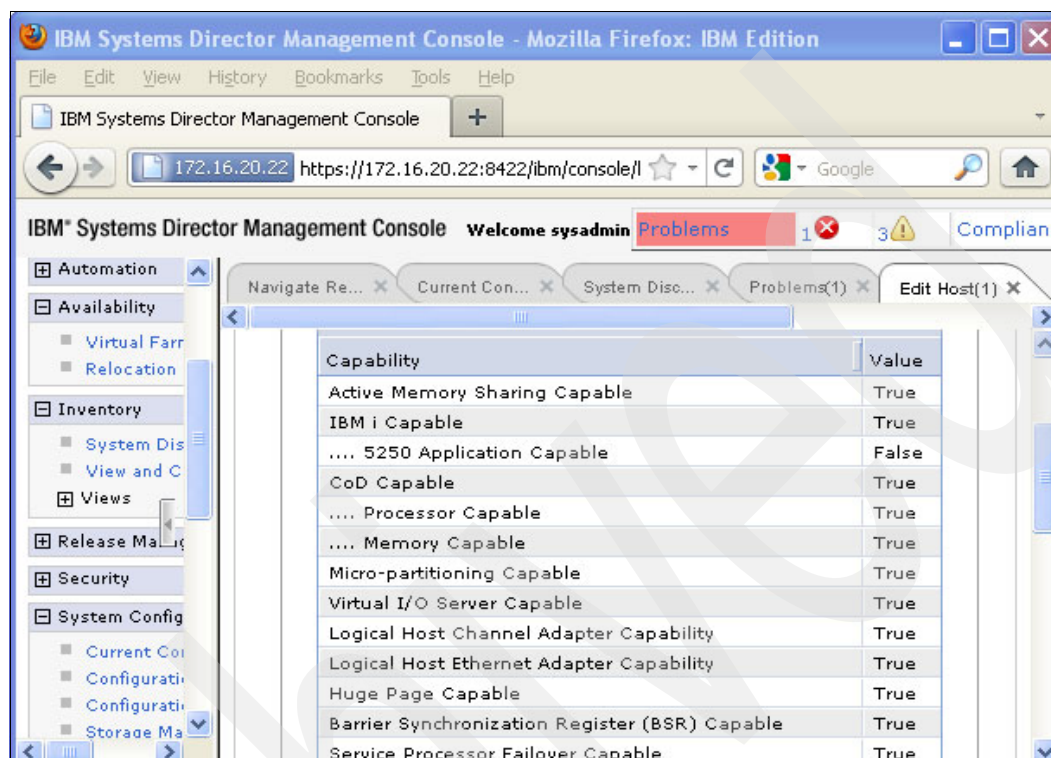


Figure 2-13 SDMC enable/disable service processor failover pane

2.2.7 Hot node add and repair

Hot node add and hot node repair functions allow you to add or repair a system node without causing downtime to the system. After the node installation, the firmware integrates the new hardware and makes it available to existing or new partitions. These features are part of the CHARM process. See 4.3.1, “Hot add or upgrade” on page 121 for more detailed information.

2.2.8 Hot node upgrade (memory)

Hot node upgrade allows you to increase the memory capacity in a system by adding or replacing (exchanging) installed memory DIMMs with higher capacity DIMMs. This feature is part of the CHARM process. See 4.3.1, “Hot add or upgrade” on page 121 for more detailed information.

2.3 TurboCore and MaxCore technology

An innovative new feature that is offered on the POWER7 Enterprise Servers is the ability to switch between the standard MaxCore mode, which is optimized for throughput, and our unique TurboCore mode, where performance per core is boosted with access to both additional cache and additional clock speed. TurboCore mode can run up to four active cores for database and other transaction-oriented processing. Standard MaxCore mode can run up

to eight active cores for Internet-oriented processing. Figure 2-14 shows the POWER7 chip design in detail.

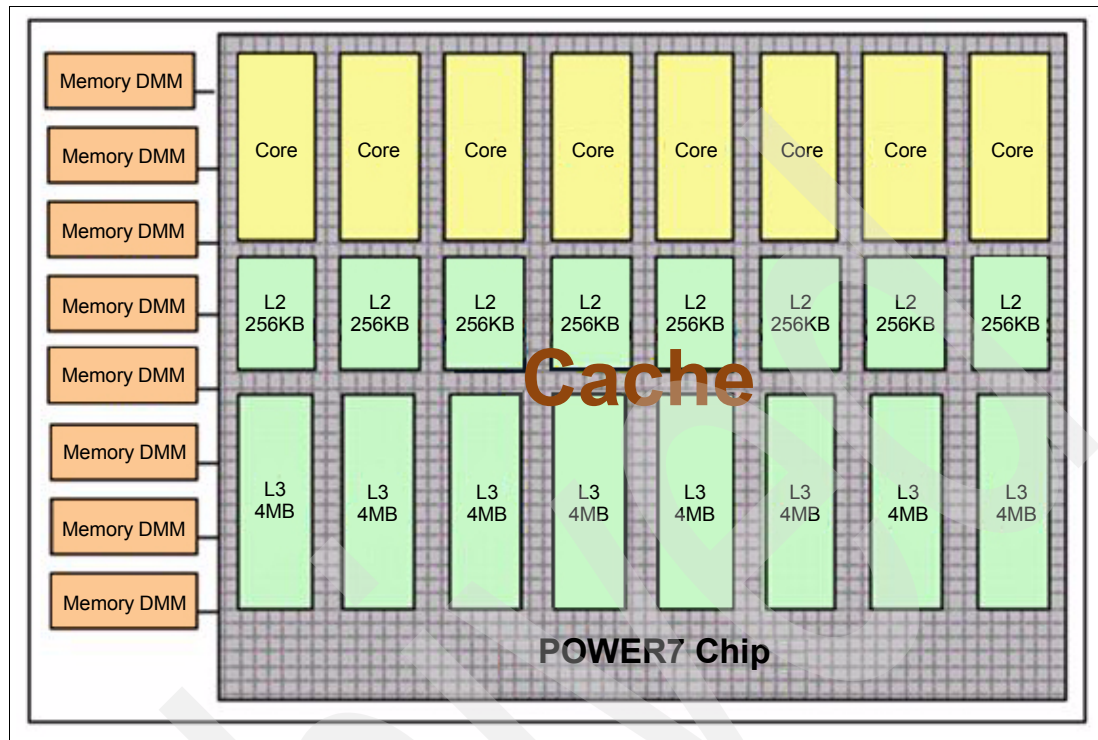


Figure 2-14 POWER7 chip design

TurboCore is a special processing mode where only four cores per chip are activated. A POWER7 chip consists of eight processor cores, each with on-core L1 instruction and data caches, a rapid access L2 cache, and a larger longer-access L3 cache. With only four active cores, ease of cooling allows the active cores to provide a frequency faster (~7.25%) than the nominal rate. This capability also means that there is more processor cache available per core. Both the higher frequency and the greater amount of cache per core are techniques that can provide better performance.

Figure 2-15 shows the differences between the MaxCore and TurboCore frequencies.

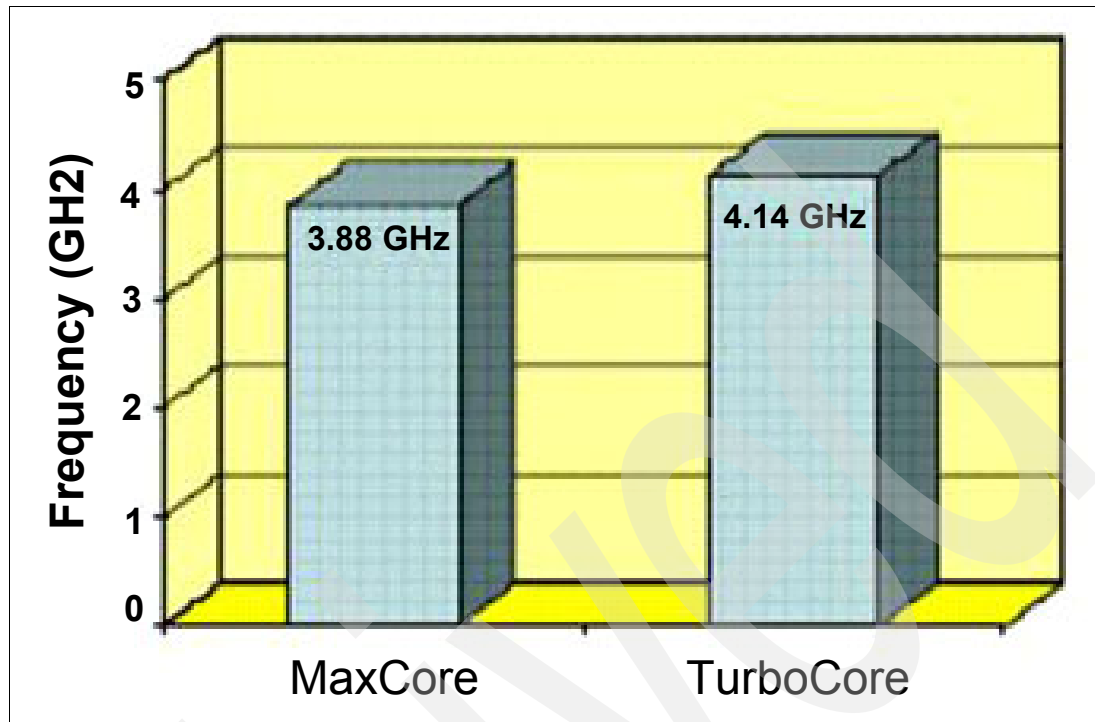


Figure 2-15 MaxCore compared to TurboCore frequency

Using MaxCore mode, which is the standard mode, all eight cores are used at a frequency of 3.86 GHz (at the time of writing this publication). Therefore, the 32 MB of L3 cache is shared evenly across the eight cores, for example, 4 MB of L3 cache per core.

In the TurboCore mode, four cores out of the eight cores are switched off, while the other four cores get a performance boost through running at a higher frequency, 4.1 GHz, that is, approximately 7.25% of the nominal rate. The TurboCore four cores access the full 32 MB of L3 cache. For example, each core gets 8 MB of L3 cache, which is double the amount in the MaxCore mode:

- ▶ MaxCore up to 256 cores @ 4 GHz (versus 64 cores @ 3.86 GHz on 780)
- ▶ TurboCore up to 128 cores at 4.25 GHz (versus 32 cores @ 4.1 GHz on 780)

In TurboCore mode, up to half of the processor cores on each single-chip module (SCM) are disabled, and their L3 cache is made available to the active processor cores on the chip. This design provides a performance boost to the active cores. In general, the number of cores used in TurboCore mode is equal to the number of processors activated, but only up to a maximum of half the number of cores physically installed.

Important: To be able to have a system that is capable of running in TurboCore mode, you need extra processor cores physically installed. Only half of the processor cores are used and you do not need to have extra activations.

Both the Power 780 and Power 795 systems support the MaxCore and TurboCore technology.

The Power 780 can be configured with up to four system enclosures. Each system enclosure contains one processor card, as shown in Figure 2-16 on page 31. Each processor card

contains two POWER7 sockets, and each socket has eight cores with 2 MB of L2 cache and 32 MB of L3 cache.

Figure 2-16 contains a top view of the Power 780 system.

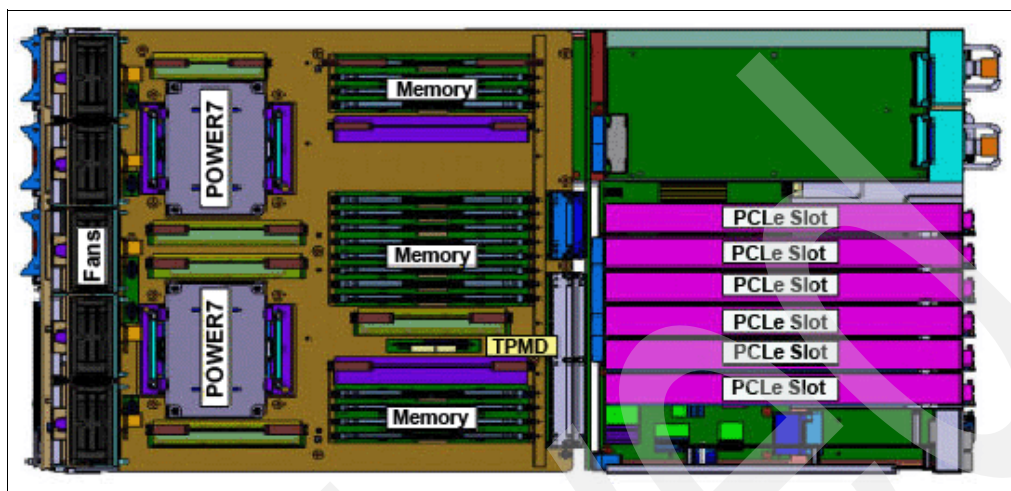


Figure 2-16 Power 780 CEC top view

The Power 795 system has a 20U-tall CEC housing that contains the system backplane cooling fans, system electronic components, and mounting slots for up to eight processor books. One to eight POWER7 processor books can be configured. Each processor book can contain either 24 or 32 cores, for example, the 6-core or the 8-core offering. These cores are packaged on four POWER7 processor chips. Each processor chip contains 2 MB of on-chip L2 cache and 32 MB of eDRAM L3 cache, and each core supports four hardware threads.

There are two types of processor nodes available on the Power 795 offering with the following features:

- ▶ Four 6-core POWER7 single chip glass ceramic modules with 24 MB of L3 cache (24 cores per processor node) at 3.7 GHz (feature code 4702)
- ▶ Four 8-core POWER7 single chip glass ceramic modules with 32 MB of L3 cache (32 cores per processor node) at 4.0 GHz (feature code 4700)

TurboCore mode: Only the 8-core processor card supports the TurboCore mode.

Both the Power 780s and the Power 795s 8-core processor cards can be configured in either of two modes, MaxCore or TurboCore.

In the MaxCore mode, the POWER7 cache design has 4 MB of L3 cache per core. Although it might look as if there is a private L3 cache per core, this cache can be shared between cores. The cache state from an active core's L3 cache can be saved into the L3 cache of less active cores.

In the case of TurboCore, the cache state from the four active cores can be saved into the L3 cache of the TurboCore's inactive cores. The result is more accessible cache per core.

It is important to note that in using TurboCore Mode, the processor chip's core count has decreased from eight cores per chip to four. An 8-core partition formally residing on one processor chip now must reside on two.

A Power 780 system needing sixteen cores and packaged in a single drawer as in Figure 2-16 on page 31 requires two drawers when using TurboCore.

Another effect, though, stems from the fact that chip crossings introduce extra time for storage accesses. For the same number of cores, there are often more chips required with TurboCore. So, more chips often imply a higher probability of longer latency for storage accesses. The performance chapter, in 7.2.4, “MaxCore and TurboCore modes” on page 252, addresses the performance effects of using TurboCore mode versus MaxCore mode.

Example 2-1, Example 2-2, and Example 2-3 explain the relationship between physical processors, activated processors, and TurboCore mode.

Example 2-1 Relationship between physical processors - 32 physical processors

A server has 32 physical processor cores with 14 activated, running in MaxCore mode. If you re-IPL the system and switch to TurboCore mode, you now have 14 processor cores running in TurboCore mode.

Example 2-2 Relationship between physical processors - 48 physical processors

A server has 48 physical processor cores with 21 activated, running in MaxCore mode. If you re-IPL the system and switch to TurboCore mode, you will have 21 processors running in TurboCore mode. There is no requirement to have an even number of processors running in TurboCore mode.

Example 2-3 Relationship between physical processors - 40 physical processors with 29 active

A server has 48 physical processor cores with 29 activated, running in MaxCore mode. If you re-IPL the system and switch to TurboCore mode, you will have 24 processors running in TurboCore mode and 5 extra activations, which are not used, because the maximum number of cores that can be used in TurboCore mode is half the number of cores physically installed (24 out of 48 in this case).

The rules for a minimum number of processor activations still apply when you configure a POWER7 Enterprise server 780 or 795 for TurboCore mode:

- ▶ Model 780: A minimum of four processor cores must be activated.
- ▶ Model 795: A minimum of three feature code 4700 processor books for TurboCore mode support 96 cores. You must activate a minimum of 25% or 24 of the installed processors, whichever is greater.

2.3.1 Enabling and disabling TurboCore mode

TurboCore mode is enabled/disabled through the ASMI interface on either your HMC or SDMC. The POWER7 server must meet the requirements reviewed earlier in this chapter to support the TurboCore settings. All processors in the system must support TurboCore for the processors to be enabled. If processors are installed in the system that do not support TurboCore, a message similar to Example 2-4 is displayed.

Example 2-4 TurboCore mode not supported

Unable to process the request because some processors are not capable of supporting TurboCore settings.

The location codes of the processors that do not support TurboCore are also displayed. To perform this operation, you must have one of the following authority levels:

- ▶ Administrator
- ▶ Authorized service provider

To set the TurboCore settings, perform the following steps:

1. On the ASMI Welcome pane, specify your user ID and password, and click **Log in**.
2. In the navigation area, expand **Performance Setup**.
3. Click **Turbo Core Settings**.
4. In the right pane, select the settings that you want.
5. Click **Save settings**.

Activate settings: To enable or disable TurboCore settings, perform an initial program load (IPL) to power off, and then power on a managed system.

Figure 2-17 shows the TurboCore enablement through the ASMI from the SDMC.

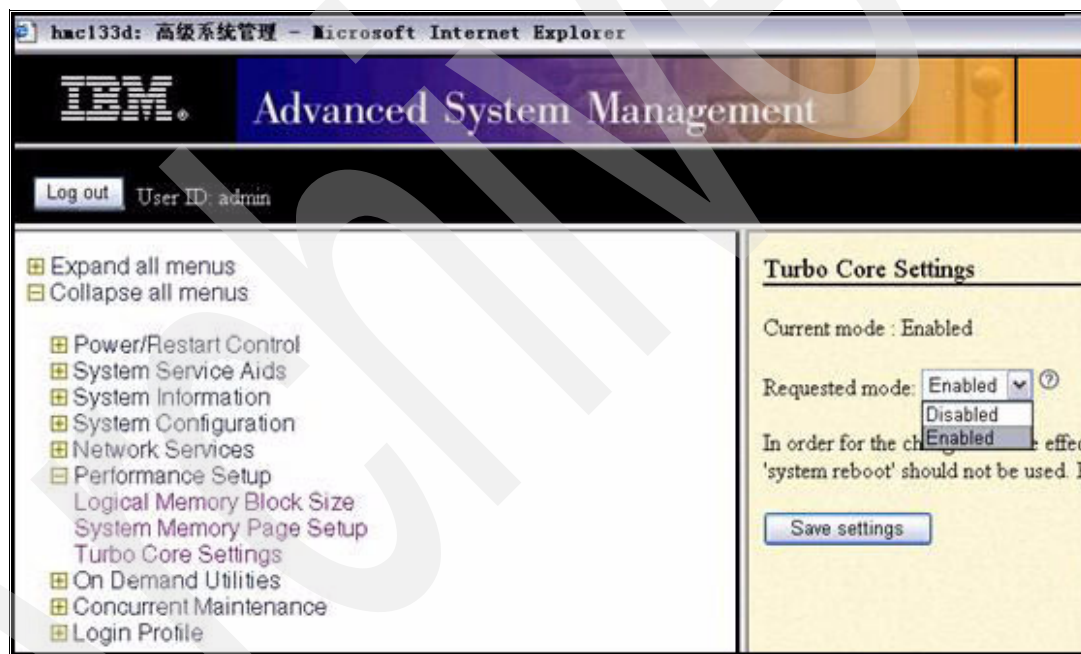


Figure 2-17 TurboCore settings

2.4 Hypervisor and firmware technologies

In this section, we describe the hypervisor and firmware technologies.

2.4.1 Hypervisor

The technology behind the virtualization of the IBM POWER7 systems is from a piece of firmware that is known as the Power hypervisor (PHYP), which resides in flash memory. This firmware performs the initialization and configuration of the POWER7 processors, along with the required virtualization support to now run up to 1,000 partitions concurrently on the IBM

POWER 795 server. The Power hypervisor is an essential element of the IBM virtualization engine. The Power hypervisor is the key component of the functions that are shown in Figure 2-18 and performs the following tasks:

- Provides an abstraction layer between the physical hardware resources and the LPARs
- Enforces partition integrity by providing a security layer between the LPARs
- Controls the dispatch of virtual processors to physical processors
- Saves and restores all processor state information during logical processor context switch
- Controls hardware I/O interrupts to management facilities for LPARs

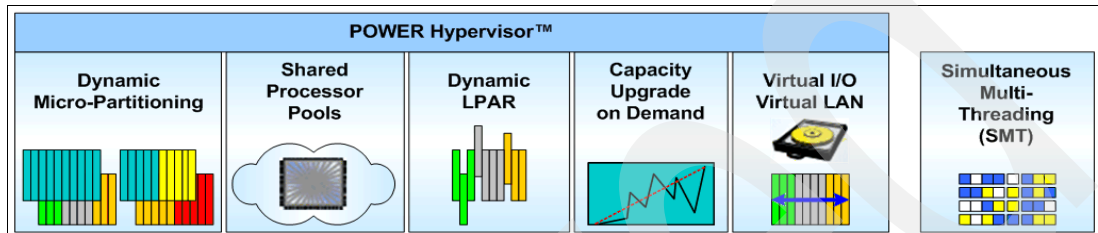


Figure 2-18 Power hypervisor functions

The Power hypervisor, acting as the abstraction layer between the system hardware and the LPARs, allows multiple operating systems to run on POWER7 technology with little or no modifications.

The Power hypervisor is a component of the system firmware that is always installed and activated, regardless of your system configuration. It operates as a hidden partition, with no processor resources assigned to it. The hypervisor provides privileged and protected access to assigned partition hardware resources and enables the use of advanced Power virtualization features by receiving and responding to requests using specialized hypervisor calls.

The Power hypervisor requires both system processor and memory resources to perform its tasks. The performance impact is relatively minor for most workloads, but it can increase with extensive amounts of page-mapping activities. Refer to Chapter 7, “POWER7 Enterprise Server performance considerations” on page 247 for more information about performance considerations.

Implement the virtual I/O server: While not required, we highly advise that you implement the virtual I/O server for use with the Power hypervisor technology to take advantage of virtualization capabilities when sharing physical I/O resources between LPARs.

Micro-partitioning technology, provided by the hypervisor, allows for increased overall use of system resources by automatically applying only the required amount of processor resources that each partition needs. Micro-partitioning technology allows for multiple partitions to share one physical processor. Partitions using micro-partitioning technology are referred to as *shared processor partitions*. You can choose between dedicated processor partitions and shared processor partitions using micro-partitioning technology on POWER7. Therefore, you are able to have both dedicated and shared processor partitions running on the same system at the same time.

The hypervisor schedules shared processor partitions from a set of physical processors that is called the *shared processor pool*. By definition, these processors are not associated with dedicated partitions.

The hypervisor continually adjusts the amount of processor capacity that is allocated to each shared processor partition and any excess capacity that is unallocated based on current partition profiles within a shared pool. Tuning parameters allow the administrator extensive control over the amount of processor resources that each partition can use.

For IBM i, the Technology Independent Machine Interface (TIMI) and the layers above the hypervisor are still in place. System Licensed Internal Code, however, was changed back in POWER5™ and enabled for interfacing with the hypervisor. The hypervisor code is based on the i Partition Licensed Internal Code and is now part of the hypervisor.

2.4.2 Firmware

Server firmware is the licensed machine code that resides in system flash memory. Server firmware includes a number of subcomponents, including Power hypervisor power control, the service processor, and LPAR firmware that is loaded into your AIX, IBM i, and Linux LPARs. There are several types of upgrades:

- ▶ **Concurrent Firmware maintenance (CFM)** is one way to perform maintenance for the hypervisor. *Concurrent Firmware maintenance* on your IBM POWER7 server firmware refers to a system that can execute the firmware upgrade without having to reboot the system.
- ▶ *Deferred updates* (delay in the upgrade) refer to firmware upgrades that can be completed in the concurrent mode, but afterwards, specific firmware upgrade functions can only be activated after the next system IPL.
- ▶ *Disruptive upgrades* require that you must perform a full system reboot before the contents of the firmware upgrade take effect.

System firmware is delivered as a release level or a service pack. Release levels support both new function or features, as well as new machine types or models. Upgrading to a higher release level is disruptive to client operations. IBM intends to introduce no more than two new release levels per year. These release levels are supported by service packs. Service packs are intended to contain only firmware fixes and not to introduce new functionality. A service pack is an update to an existing release level. The Power code matrix website also provides the life cycle for each system firmware release level. Use life-cycle information, along with the supported code combination tables, to assist you with long-term planning for upgrades to your HMC and system firmware.

2.4.3 Dynamic firmware update

To increase systems availability, it is important that defects in the service processor firmware, Power hypervisor, and other system firmware can be fixed without experiencing an outage. IBM Power servers can operate in a given supported firmware release, using concurrent firmware fixes to be updated to the current release level.

Normally, IBM provides patches for a firmware release level for up to two years after its release. Then, clients must plan to upgrade to a new release in order to stay on a supported firmware release level.

In addition to concurrent and disruptive firmware updates, IBM offers concurrent updates, which include functions that are activated in the next server reboot.

Disruptive firmware patches: Certain firmware patches, for example, patches changing the initialization values for chips, and the activation of new firmware functions, which require the installation of a new firmware release level, are disruptive processes that require a scheduled outage and full server reboot.

2.4.4 Firmware update and upgrade strategies

This section describes the firmware update and upgrade strategies.

Upgrade strategy (release level)

New functions are released via firmware release levels. The installation of a new release level is disruptive. At the time of this book, there is no plan for non-destructive versions.

Unless you require the functions or features that are introduced by the latest release level, it is generally prudent to wait a few months until the release level stability has been demonstrated in the field. Release levels are supported with fixes (delivered via service packs) for approximately two years. Supported releases overlap, so fixes usually are made in multiple service packs. Typically, clients are not required to upgrade to the latest level to obtain fixes, except when the client's release level has reached the end of service. Therefore, clients can stay on an older (typically more stable) release level and still obtain fixes (via service packs) for problems. Because the number of changes in a service pack is substantially fewer than a release level, the risk of destabilizing the firmware by installing a service pack update is much lower.

Update strategy (service pack within a release)

The strategy to update to the latest service pack needs to be more aggressive than upgrading to the latest release level. Service packs contain fixes to problems that have been discovered in testing and reported from the field. The IBM fix strategy is to encourage the installation of a fix before the problem is encountered in the field. The firmware download page provides downloads for the *N* and *N-1* service packs, unless a problem was introduced with the *N-1* service pack. In such cases, only the latest level is available. Our goal is to prevent the installation of a broken service pack. When product and development engineering determines that the latest available (*N*) service pack has sufficient field penetration and experience (30 - 60 days), the download for the older (*N-1*) service pack is removed from the firmware download web page.

Website: Refer to the following website to ensure that you have the latest release levels for the supported IBM Power Systems firmware:

<http://www14.software.ibm.com/webapp/set2/sas/f/power5cm/power7.html>

2.5 Power management

In response to rising energy costs, which can be prohibitive to business growth, and also in support of green initiatives, IBM developed the EnergyScale™ technology for IBM Power Systems. This technology allows the system architects to monitor and control energy consumption for power, cooling, planning, and management purposes.

POWER7 processor-based systems support EnergyScale features and support IBM Systems Director Active Energy Manager™, which is a comprehensive energy management tool that monitors and controls IBM Power servers. Support and awareness of EnergyScale extends

throughout the system software stack, and is included in AIX, IBM i, and Linux operating systems. Table 2-1 reviews the Power management features and indicates which features require Active Energy Manager (AEM).

Table 2-1 Power management features

Feature	Requires AEM	Description
Power Trending	Yes	Collects and reports power consumption information for a server.
Thermal Reporting	Yes	Collects and reports inlet and exhaust temperatures (where applicable).
Static Power Server	No	Provides predictable performance with power savings by reducing CPU frequency by a fixed amount.
Dynamic Power Saver	Yes	Allows a system to implement algorithms to adjust the processor core frequency to favor system performance (saving power where applicable) or to balance power and performance. Core frequency can exceed 100% at times.
Power Capping	Yes	Enforces a user-specified power budget on a system.
Energy-Optimized Fans	No	System fans respond dynamically to temperatures of system components.
Processor Core Nap	No	Enables low-power modes in POWER7 when cores are unused.
Processor Folding	No	Dynamically re-allocates which processor cores execute a task to optimize energy efficiency of the entire system.
EnergyScale for I/O	No	Powers on I/O slots only when needed.
Server Power Down	Yes	Provides information that is necessary to dynamically migrate workloads off of a lightly utilized system, allowing the entire system to be powered off.
Partition Power Management	Yes	Provides power savings settings for certain partitions and the system processor pool. Not available on IBM PS70x Blades.

2.5.1 Differences in dynamic power saver from POWER6 to POWER7

Dynamic power saver differs from POWER6 to POWER7 systems.

In POWER6 systems, maximum frequencies varied based on whether Favor Power or Favor Performance was selected in Active Energy Manager. Favor Power guaranteed power savings by limiting the maximum frequency of the system under peak utilization. Favor Performance allowed a higher frequency range. In both cases, the firmware increased the processor frequency only under high utilization.

In POWER7 systems running system firmware EM710, EnergyScale Dynamic Power Saver maintains compatibility with POWER6 implementations. In POWER7 systems running EM711 system firmware or later, Dynamic Power Saver has been enhanced so that the full frequency range is available to a system (including frequencies in excess of 100% where applicable) regardless of whether power or performance is selected in Active Energy Manager.

Instead of controlling frequency ranges, POWER7 EnergyScale with EM711 firmware or newer selects from various power and performance control algorithms, depending on the selected mode. In Dynamic Power Saver, Favor Power mode, system firmware balances performance and power consumption, only increasing processor core frequency when the system is heavily utilized. In Dynamic Power Saver, Favor Performance mode, system firmware defaults to the maximum processor core frequency that is allowed for a given system's environment and configuration, and reduces frequency only when a system is lightly utilized or idle.

2.6 Rapid deployment of PowerVM clients

This section presents three methods for parallel deployment of virtual servers (LPARs) on a high-end POWER7 system to deliver improvements in both efficiency and flexibility.

2.6.1 Deployment using the VMControl plug-in

VMControl allows you to both provision an empty server and deploy an operating system image to it. For instructions to provision a new virtual server via VMControl, see "Creating a virtual server" under 4.1.1, "Managing virtual servers" in *IBM Systems Director VMControl Implementation Guide on IBM Power Systems*, SG24-7829:

<http://www.redbooks.ibm.com/redbooks/pdfs/sg247829.pdf>

For instructions to deploy a complete operating system on a POWER7 system, see 4.2.4, "Deploying a virtual appliance to a host" in *IBM Systems Director VMControl Implementation Guide on IBM Power Systems*, SG24-7829:

<http://www.redbooks.ibm.com/redbooks/pdfs/sg247829.pdf>

Section 5.2.5, "Managing virtual appliances", in *IBM Systems Director VMControl Implementation Guide on IBM Power Systems*, SG24-7829, also describes in detail creating image repositories for AIX and Linux, and capturing a mksysb image or resource to create/deploy a virtual appliance. A *virtual appliance* contains an image of a full operating system, and it can also contain software applications and metadata describing the virtual server that the image requires.

2.6.2 File-backed virtual optical devices

The virtual I/O server support for virtual optical devices allows sharing of a physical CD or DVD drive that is assigned to the virtual I/O server between multiple AIX, IBM i, and Linux client partitions. This feature has been available since Virtual I/O Server Version 1.5. This feature supports provisioning of the DVD drive on the virtual I/O server by mapping it to each of the partitions. This drive can only be utilized by one partition at a time.

A new feature, the file-backed virtual optical device, which was introduced in Virtual I/O Server Version 1.5, allows simultaneous use of the installation image by all the virtual servers in parallel. This technology is the preferred method for deployment of operating system images on high-end systems where a large number of virtual servers can be configured.

Using file-backed virtual optical devices provides the flexibility to use an ISO image as a virtual device and share it among all the virtual servers on your system as a virtualized optical drive. The virtual media repository is used to store virtual optical media for use by the file-backed virtual optical devices. This design is analogous to the file-backed virtual optical device being a juke box and the virtual optical media repository as its CD library.

The following procedure illustrates how to install an AIX client partition using file-backed virtual optical devices. This procedure can also be used in installing Linux and IBM i partitions:

1. The first step is to check for a defined storage pool in which you need to create the virtual media repository, as shown in Example 2-5.

Example 2-5 Checking for the defined storage pools

```
$ lspp
Pool              Size(mb)  Free(mb)  Alloc Size(mb)  BDs  Type
rootvg            69632    41984          512      0  LVP00L
isopool           16304    16304          16      0  LVP00L
```

The **lspp** command that is shown in Example 2-5 lists the storage pools that are defined in the virtual I/O server.

2. Create a virtual media repository on the virtual I/O server using the **mkrep** command, as shown in Example 2-6.

Example 2-6 Creating a virtual media repository

```
$ mkrep -sp isopool -size 10G
Virtual Media Repository Created
```

3. Copy the ISO images to the virtual I/O server and place them in a directory that is created in `/home/padmin`.
4. Create a virtual optical media disk in the virtual media repository using the **mkvopt** command, as shown in Example 2-7.

Example 2-7 Creating a virtual optical media disk

```
$ mkvopt -name AIXcd -file /home/padmin/AIXiso/dvd.710_GOLD.v1.iso
```

The **mkvopt** command that is shown in Example 2-7 creates a virtual optical media disk named **AIXcd** from the `dvd.710_GOLD.v1.iso` image that is located in the `/home/padmin/AIXiso` directory.

5. In order to show that the new virtual optical media disk was added with the **mkvopt** command, use the **lsrep** command as shown in Example 2-8.

Example 2-8 Showing that the virtual optical media was added

```
$ lsrep
Size(mb)  Free(mb)  Parent Pool      Parent Size      Parent Free
10198     7083     isopool          16304            6064

Name      File Size Optical      Access
AIXcd     3115  None                rw
```

6. Remove the `iso` file to save space, because it is already in the repository.
7. Now, map the virtual optical media disk to a virtual server adapter that is mapped to the AIX client with the **mkvdev** command, as shown in Example 2-9.

Example 2-9 Mapping the virtual optical media disk

```
$ mkvdev -fbo -vadapter vhost0
vtopt0 Available
```

The **mkdev** command makes a vtop0 device available, as shown in Example 2-9 on page 39.

8. The next step is to load the virtual media in the virtual optical device using the **loadopt** command, as shown in Example 2-10.

Example 2-10 Loading the virtual media

```
$ loadopt -disk AIXcd -vtd vtop0
```

9. Verify the mapping with the **lsmmap** command. The output is similar to Example 2-11.

Example 2-11 Output of the lsmmap command

```
$ lsmmap -vadapter vhost# (replace # with your adapter number)
VTD          vtop0
Status       Available
LUN          0x8400000000000000
Backing device /var/vio/VMLibrary/AIXcd
Physloc
Mirrored     N/A
```

10. Use the **lsdev** command to show the newly created file-backed optical device. Refer to Example 2-12.

Example 2-12 Showing the file-backed optical device

name	status	description
ent2	Available	Virtual I/O Ethernet Adapter (1-lan)
ent3	Available	Virtual I/O Ethernet Adapter (1-lan)
ent4	Available	Virtual I/O Ethernet Adapter (1-lan)
ent5	Available	Virtual I/O Ethernet Adapter (1-lan)
vasi0	Available	Virtual Asynchronous Services Interface (VASI)
vbsd0	Available	Virtual Block Storage Device (VBSD)
vhost0	Available	Virtual SCSI Server Adapter
vsa0	Available	LPAR Virtual Serial Adapter
name	status	description
max_tranVTD	Available	Virtual Target Device - Disk
skessd6s_hd01	Available	Virtual Target Device - Disk
vtd01	Available	Virtual Target Device - Disk
vtop0	Available	Virtual Target Device - File-backed Optical <- NEW
name	status	description
ent6	Available	Shared Ethernet Adapter

Finally, we can use the vtop0 device to boot the AIX client partition and install from it.

2.6.3 Deployment using the System Planning Tool (SPT)

The IBM System Planning Tool is the third method to deploy a system that is based on existing performance data or based on new workloads. The system plans that are generated by the SPT can be deployed on the system by the Hardware Management Console (HMC).

We describe the SPT in detail in Chapter 4 of this book (4.8, “System Planning Tool (SPT)” on page 151). For the latest detailed information about SPT, refer to the IBM System Planning Tool at this website:

<http://www-947.ibm.com/systems/support/tools/systemplanningtool/>

2.7 I/O considerations

In a Power system, up to 10 partitions can be defined per physical processor. With a maximum of 1,000 LPARs per server, typically each LPAR uses at least one Ethernet adapter and one adapter to access the back-end storage. This design results in at least 2,000 adapters that are needed to exploit the full capacity of the server. This number doubles if we consider redundant adapters. You also need to consider the number of cables and both LAN and SAN switch ports that are required for this type of a configuration.

Available systems cannot fulfill this I/O requirement in terms of physical adapters. However, by using PowerVM, you can implement several storage virtualization technologies that allow you to share the disk I/O resources.

Although many backing devices are supported, not all of them offer redundancy or virtualization capabilities. The following devices are available backing devices:

- ▶ Internal storage (virtualized or physically assigned)
- ▶ Physically attached external storage
- ▶ Virtualized SAN storage (either virtual SCSI (vSCSI) or N_Port ID Virtualization (NPIV))
- ▶ Disk backed by file
- ▶ Disk backed by logical volume (LV)
- ▶ Optical CD/DVD
- ▶ Optical DVD-RAM backed by file
- ▶ Tape devices

Because we focus on virtualized highly available solutions, we only advise the use of SAN-attached devices that are virtualized via vSCSI or NPIV technologies.

Supported storage systems and drivers: For an up-to-date list of supported storage systems and drivers, consult the SSIC website:

<http://www-03.ibm.com/system/support/storage/ssic/interoperability.wss>

2.7.1 Virtual SCSI

Virtual SCSI (vSCSI) is a virtual implementation of the SCSI protocol. Available since POWER5, it provides Virtual SCSI support for AIX 5.3 and later, selected Linux distributions, and IBM i (POWER6 needed).

Virtual SCSI is a client/server technology. The virtual I/O server owns both the physical adapter and the physical disk and acts as a server (target device). The client LPAR accesses the physical disk as a SCSI client. Both client and server virtual SCSI adapters are configured using HMC, SDMC, or Integrated Virtualization Manager (IVM). Physical connection between the adapters is emulated by mapping the corresponding adapter slots at the management console and by configuring the devices in both client and virtual I/O server.

After the adapter's configuration is ready in the client and the server, the back-end disk must be mapped to the virtual client. This process is done at the VIO server or servers. Then, the disk is configured at the client partition as a regular physical disk.

As shown in Figure 2-19 on page 42, for availability purposes, always consider a dual virtual I/O server environment with at least two paths to the physical storage at each virtual I/O server.

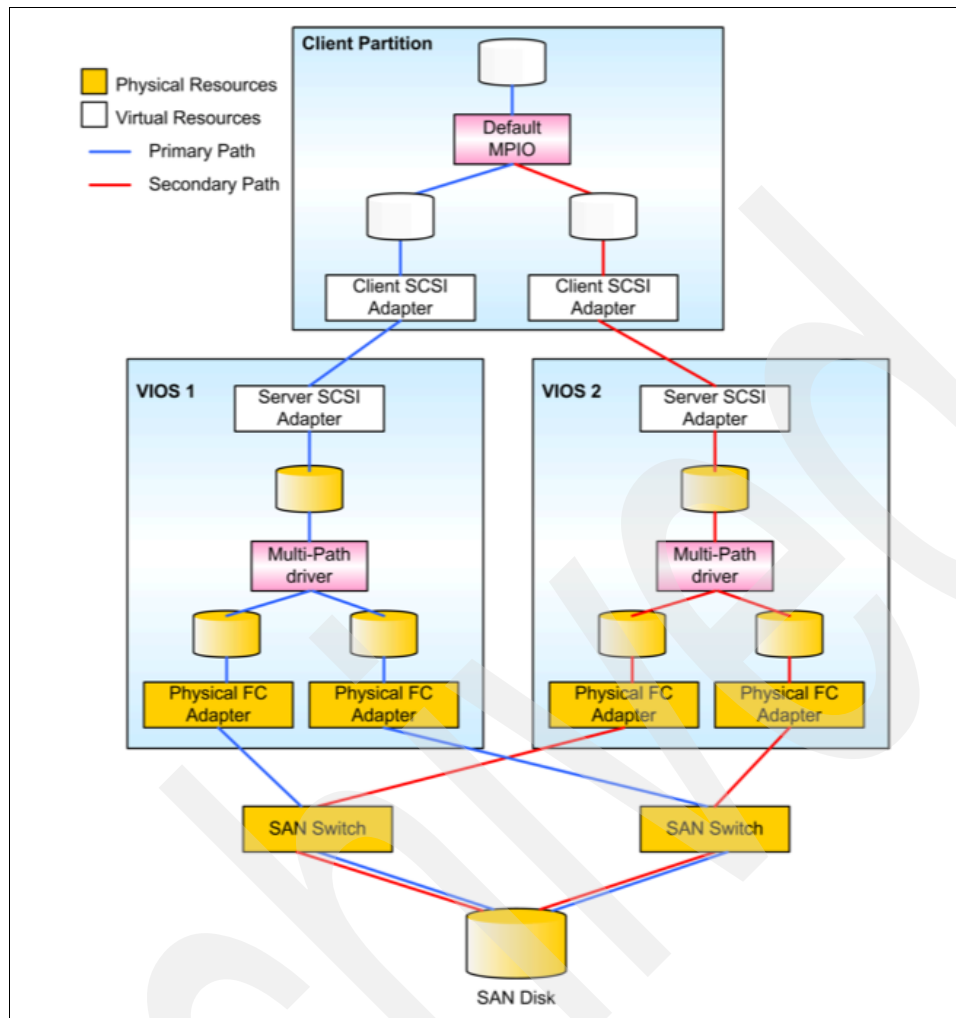


Figure 2-19 Redundant vSCSI configuration

Figure 2-19 shows these components:

- ▶ Two virtual I/O servers.
- ▶ Each virtual I/O server has two paths to the SAN storage.
- ▶ Using the multipath I/O (MPIO) driver, the client partition can access its backing device using the prioritized path.
- ▶ The backing disks are located in a SAN environment.

Important: Shared storage pool redundancy is not supported at the time of developing this book.

SCSI Remote Direct Memory Access (RDMA)

By implementing the RDMA protocol, PowerVM SCSI implementation has the ability to directly transfer information between SCSI initiators and target memory address spaces.

As shown in Figure 2-20, the SCSI request and responses are sent over the vSCSI adapters using the Power hypervisor, but the actual data transfer is done directly between the LPAR and the physical adapter using the RDMA protocol.

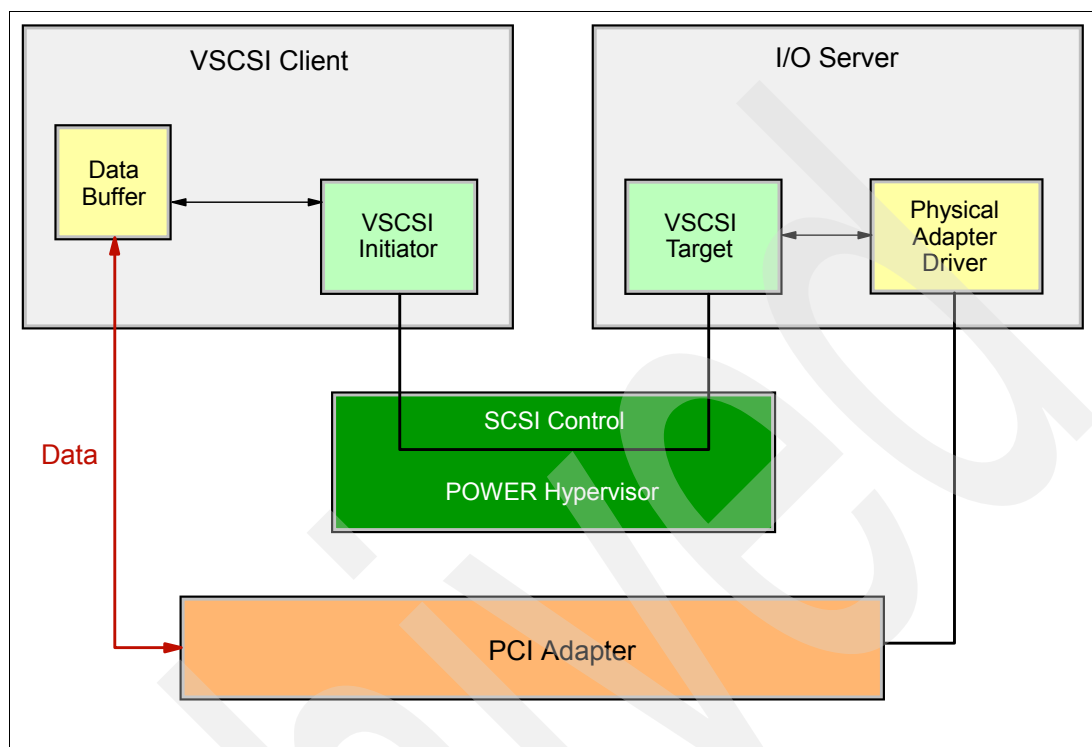


Figure 2-20 Logical Remote Direct Memory Access (RDMA)

2.7.2 N_Port ID Virtualization (NPIV)

In a typical SAN environment, when using FC, the logical unit numbers (LUNs) are created using the physical storage and then mapped to physical host bus adapters (HBAs). Each physical port on each physical FC adapter is identified by its own unique worldwide port name (WWPN).

NPIV is an FC adapter virtualization technology. By implementing NPIV, you can configure your system so that each LPAR has independent access to the storage system that shares a physical adapter.

To enable NPIV in your managed system, you need to install one or more virtual I/O servers. By creating servers' and clients' virtual FC adapters, the virtual I/O server provides a pass through to enable the client virtual server to communicate with the storage subsystem using a shared HBA.

Using NPIV, you can have FC redundancy either by using MPIO or by mirroring at the client partition. Because the virtual I/O server is just a pass through, the redundancy occurs totally in the client.

As shown in Figure 2-21, to achieve high levels of availability, NPIV must be configured in a dual VIO environment with at least two HBAs at each virtual I/O server.

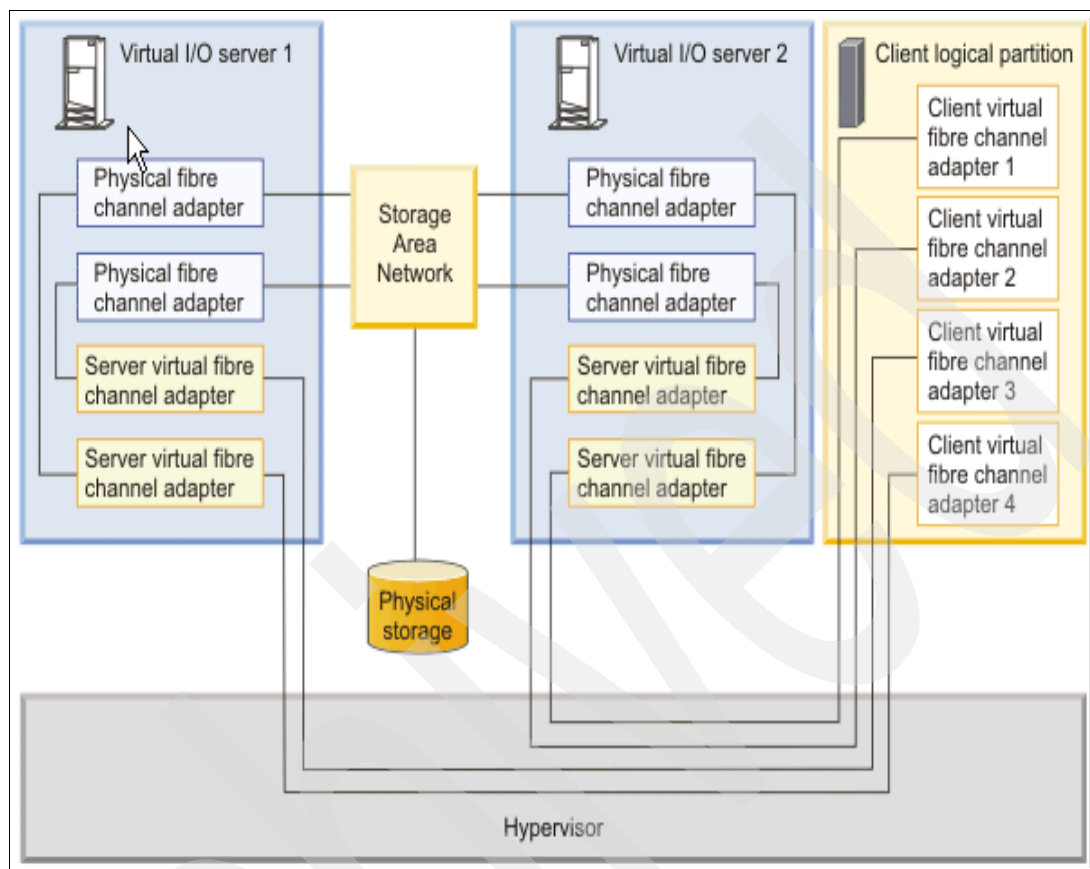


Figure 2-21 NPIV with redundancy

Figure 2-21 shows the following scenario:

- ▶ There are two virtual I/O servers to provide redundancy at the VIO level.
- ▶ There are two independent HBAs at each VIO and they are connected to the storage.
- ▶ The client can access the storage through both virtual I/O servers using any of the configured HBAs.

Using this configuration, you protect your client partitions from both HBA and virtual I/O server problems.

Adding more HBAs: Depending on your environment, consider adding more HBAs for both load balancing and availability purposes.

NPIV memory prerequisites

When planning for an NPIV configuration, remember that, as in any other high-speed adapters, virtual FC adapters require memory to operate. Because the virtual I/O server only copies packets from the physical adapter to the virtual client adapter, it does not require extra memory. The hypervisor reserves the required memory.

Important: Consider reserving 140 MB for each client virtual FC adapter. However, the amount of memory that is needed is much less than the amount that is required to manage physical adapters. For more information, refer to the following website:

https://www.ibm.com/developerworks/mydeveloperworks/blogs/aixpert/entry/virtual_fibre_channel_for_npiv_requires_memory_too59?lang=en

Mixed configurations

As shown in Figure 2-22, an NPIV and vSCSI mixed configuration is also supported. A configuration in which you also use physical adapters in the client partitions is supported, too. However, consider that if a physical adapter is present, you need to move your disks to a virtual adapter before performing a Live Partition Mobility operation.

By using NPIV and vSCSI in the same scenario, you can take advantage of the specific capabilities of both technologies. Refer to *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940-04, for a complete comparison between NPIV and vSCSI.

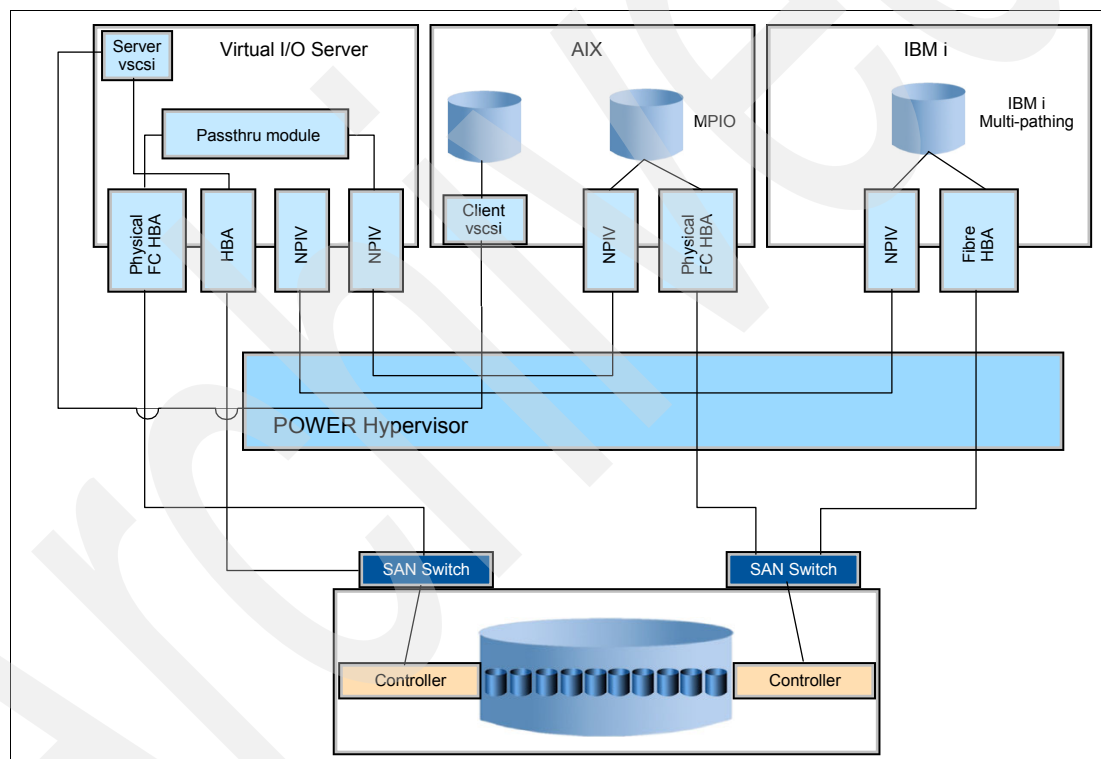


Figure 2-22 vSCSI, NPIV, and physical adapter in the same client partition

2.8 Active Memory Sharing

Since the earlier version of PowerVM, formerly known as Advanced Power Virtualization (APV), IBM Power Systems have had the ability to virtualize processor use. Initially known as *micro-partitioning*, this feature allows one processor core to be shared with up to 10 LPARs.

Introduced in 2009, Active Memory Sharing (AMS) is an advanced PowerVM memory virtualization technology that allows multiple partitions in a Power System to share a common pool of physical memory.

AMS can be used to improve memory utilization in a Power System's server by reducing the total assigned memory or by creating more LPARs using the same amount of physical memory.

AMS allows the administrator to define shared memory pools and to assign LPARs to these pools. As shown in Figure 2-23, the same Power System server can host both shared and dedicated memory LPARs.

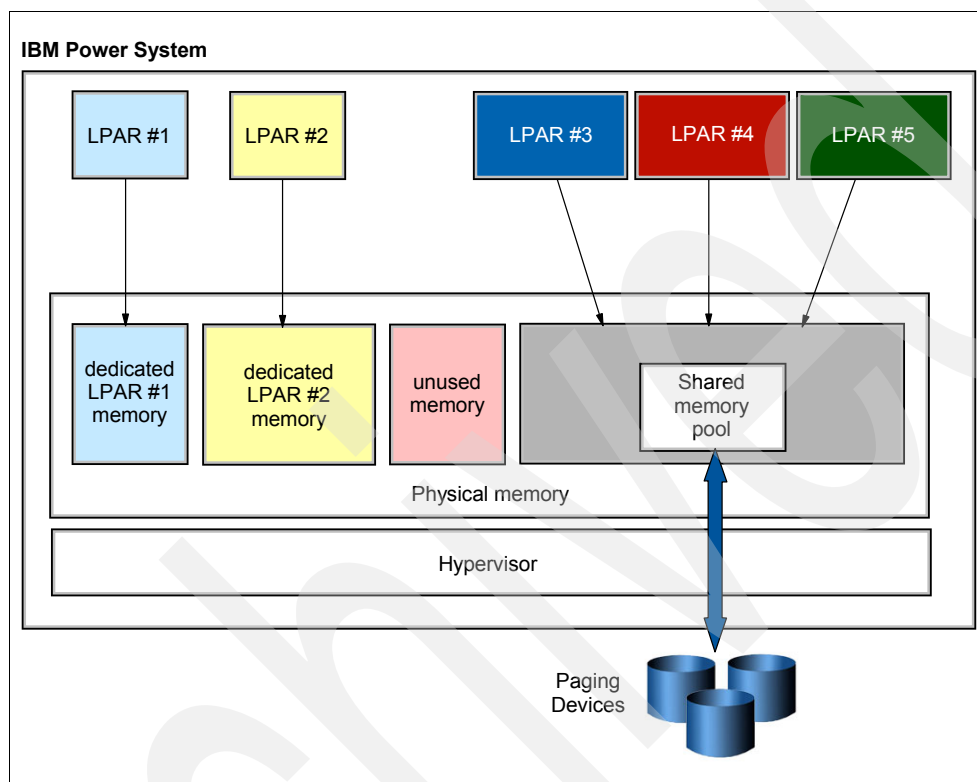


Figure 2-23 Power System with dedicated and shared memory partitions

Using AMS, memory flows from one LPAR to another LPAR according to the real needs of each LPAR, not based on fixed assignments. Using AMS, the sum of logical memory that is assigned to a pool of shared memory partitions can exceed the amount of memory in the pool, allowing the creation of more LPARs than if using dedicated memory. This capability is known as *memory logical overcommitment*.

2.8.1 Shared memory pool

The *shared memory pool* is a collection of physical memory that is reserved for shared memory partitions' exclusive use. In a system that is enabled to use the AMS feature, the administrator must define a shared memory pool before creating shared memory partitions.

Shared memory option: If there is no shared memory pool defined, the shared memory option does not appear in the Create logical partition wizard. However, you can define your partition as a dedicated memory partition and later change this setting. A reboot is required for this operation.

The shared memory pool size can be increased and reduced dynamically. If no shared memory partitions are available, the pool can also be deleted.

Upper limit: Up to 1,000 partitions are supported at each shared memory pool.

In a shared memory environment, each shared memory partition requires a dedicated paging space that is assigned to the shared memory pool. If there is no available paging device in the pool, the activation fails. Paging devices are automatically designed based on the maximum logical configuration of the LPAR.

Paging devices can be dynamically added or removed from the shared memory pool if not in use by any virtual server.

2.8.2 Paging virtual I/O server

In memory overcommitment situations, the hypervisor needs to free memory pages in the shared processor pool. As in regular paging, the data in those memory pages needs to be moved to paging devices to be restored when needed. In active memory sharing environments, this paging activity is performed by the virtual I/O server that is assigned to that shared memory pool following the hypervisor requests.

For availability, we highly suggest to implement a dual VIO to provide redundancy for shared memory pool paging activities.

As described in 2.8.1, “Shared memory pool” on page 46, each shared memory LPAR requires its own paging device. Although many backup devices can be used, for performance and high availability, we advise that you only use these backup devices:

- ▶ Mirrored volumes (distributed over many physical disks)
- ▶ Located on SAN environments
- ▶ Accessible for both paging virtual I/O servers

2.8.3 Client LPAR requirements

In order to use AMS, an LPAR must meet the following prerequisites:

- ▶ Use shared processors rather than dedicated processors
- ▶ Use virtual I/O resources
- ▶ AIX Level 6.1 TL 03 or later
- ▶ Novell SUSE SLES11 kernel 2.6.27.25-0.1-ppc64 or later
- ▶ IBM i Version V6R1M1 PTF SI32798 or later

2.8.4 Active Memory Sharing and Active Memory Expansion

Active Memory Sharing is a PowerVM feature. Active Memory Expansion is a virtualization technology that allows a partition to perform memory compression and expand its memory.

Although these technologies differ, they are both memory virtualization technologies that can work independently or together.

2.8.5 Active Memory Sharing with Live Partition Mobility (LPM)

Shared memory partitions are eligible for LPM operations if there is a shared memory pool in the destination server with available, suitable paging devices to be allocated to the migrated partition.

2.9 Integrated Virtual Ethernet

First introduced in POWER6 servers, Integrated Virtual Ethernet (IVE) was not present in the high-end servers. With the release of POWER7 servers, this high-speed, virtualizable network technology is now available for the IBM POWER7 780 servers.

Also called the Host Ethernet Adapter (HEA), IVE enables the sharing of integrated high-speed Ethernet ports. IVE includes hardware features to provide logical Ethernet ports for inter-partition and external network communication without using any other component, such as the virtual I/O server.

The IVE is a physical Ethernet adapter that is connected to the GX+ bus of the Power processor instead of being connected to the PCI buses, as illustrated in Figure 2-25 on page 49. This configuration provides a high throughput adapter. IVE also includes hardware features to enable the adapter to provide logical adapters. These logical adapters appear as regular Ethernet adapters to the virtual servers.

As shown in Figure 2-24, IVE logical adapters communicate directly between the LPAR and external networks, reducing the interaction with the hypervisor. Previously, this communication was performed using virtual Ethernet and Shared Ethernet Adapter (SEA) adapters by moving packages from one LPAR to another LPAR through the Power hypervisor.

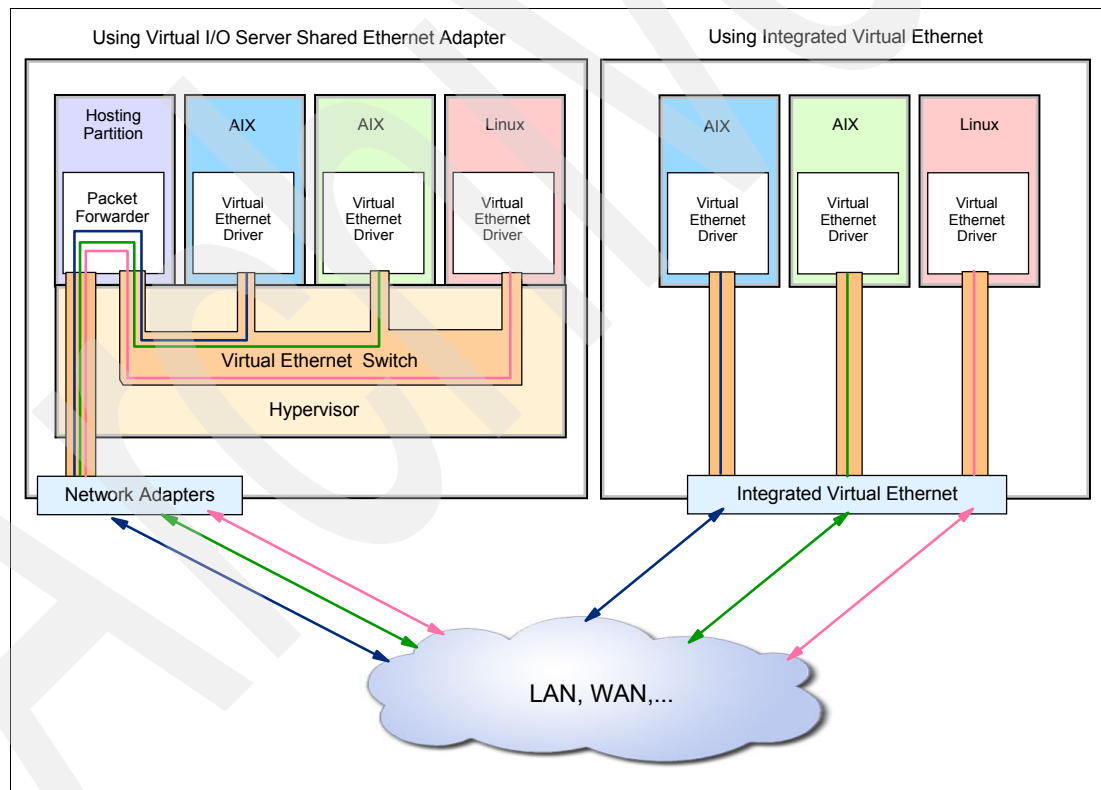


Figure 2-24 SEA and IVE model comparison

Figure 2-25 shows the IVE and processor data connections.

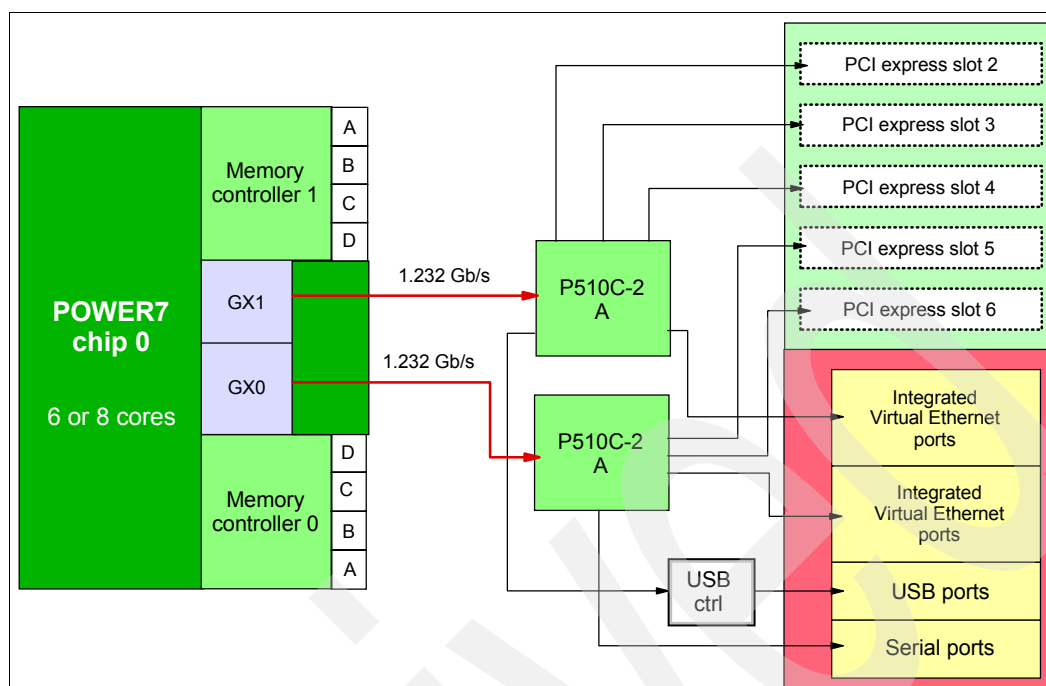


Figure 2-25 IVE and processor data connections

At the time that this publication was written, three IVE adapters were available on the Power 780:

- ▶ Feature code 1803: Four 1 Gbps Ethernet ports
- ▶ Feature code 1804: Two 10 Gbps SFP+ optical (SR only) Ethernet ports and two 1 Gbps copper Ethernet ports
- ▶ Feature code 1813: Two 10 Gbps SFP+ copper twinaxial ports and two 1 Gbps Ethernet ports

IVE supports 64 LPARs maximum: Each IVE feature code can address up to 64 logical Ethernet ports to support up to 64 LPARs. If you plan to have more than 64 LPARs, use the Shared Ethernet Adapter.

For more information about IVE features and configuration options, see the *Integrated Virtual Ethernet Adapter Technical Overview and Introduction*, REDP-4340-00.

2.10 Partitioning

One of the earlier concepts that was introduced in Power Systems was the creation of LPARs. LPARs were introduced in Power Systems beginning with POWER4™ systems. Since this introduction, IBM has continued to improve the virtualization technologies. The goal of this publication is to familiarize you with most of these virtual technologies.

SDMC: Two names have changed with the SDMC:

- ▶ LPARs are called *virtual servers* on the SDMC.
- ▶ Managed servers are called *hosts* on the SDMC.

Prior to the introduction of partitioning, IBM clients experienced the following situations:

- ▶ Clients installed a number of related applications on a single machine, which led to possible incompatibilities in system requirements and competition for resources:
 - One example was an application requiring certain network parameter settings; however, another application required a separate setting. These parameters might not be valid for the third application. The use of settings, such as Asynchronous I/O (aio) in AIX, can be used by one application, but not another application.
 - Another example was an application using more CPU time than other applications, thus slowing down the whole system. This situation is analogous to a 1 Tier Architecture. IBM, with AIX Version 4, attempted to remedy the situation by introducing a feature called *Workload Management*. The challenge was that, even with Workload Management, all the applications still executed in the same operating system space.
- ▶ Application designers needed to separate the machines on which the applications are installed. For each new application, a new physical machine had to be purchased for each application that was deployed. Clients ended up with an unmanageable number of physical machines. Data center floors were full of machines. This situation is analogous to an *n*-Tier system.

To avoid these situations and to make better use of deployed technology and to gain full power of IT investments, server consolidation became a necessity. IBM introduced POWER4 systems that were capable of logical partitioning. Each partition was seen as an individual machine. POWER4 logical partitioning had limitations, including the inability to perform Dynamic Reconfigurations (DLPAR). Most DLPAR operations were hardware-based. Clients were unable to virtualize processors.

IBM POWER improved with the virtualization of system resources. A number of publications exist that explain more about LPARs. Consider reading *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940, which lists in detail, virtualization features, operating system support, and hardware support. In general terms, you can virtualize the processor, memory, I/O, and operating system. A single copy of AIX can be virtualized into multiple workload partitions (WPARs).

This section describes the process of creating an LPAR. Although we concentrate on creating the LPAR using the IBM Systems Director Management Console (SDMC), all these steps can be performed using an HMC.

Naming: Be aware of the name changes between the HMC and the SDMC. We will use both virtual server (VS) and LPAR interchangeably throughout the book. Do not confuse a virtual server with a virtual I/O server.

Assumptions

We create an LPAR with the following assumptions:

- ▶ We assume that the POWER7 server is installed, powered on, and connected to either an HMC or SDMC. Chapter 4, “Planning for virtualization and RAS in POWER7 high-end servers” on page 99 contains a discussion about planning and installation.

- We assume that the sizing of the LPAR has been determined and that the system administrator has enough resources to meet the minimum requirements to start the partition. You can choose to use the entire POWER7 as a single partition system. If this design is done, after the partition is started, no other partitions can be started, because there are no resources available. You can create other partition profiles, but the system partition must be stopped before any other partitions can be started.

Figure 2-26 shows the creation of a full system partition. Remember that you are resource-bounded when you create a partition that uses all the resources of the server.

- We also assume that you have decided which virtualization technology to use for access to the network and storage. We discuss these existing options in this book:
 - NPIV for Virtual Fibre Channel Adapter or vSCSI
 - Internal disks for storage (limits mobility)
 - Storage pools if required (limits mobility)
 - IVE as a Logical Host Ethernet Adapter (HEA) versus a Shared Ethernet Adapter (SEA)

Technologies: You can use any combination of these technologies, depending on your environment. You need to understand the features, as well as the challenges that exist for each feature. For example, using HEA and not SEA hinders LPM.

You can also use internal disks along with physical adapters. This publication does not discuss these options, because our goal is to introduce features that make your environment virtualized, highly available, flexible, and mobile.

- We assume that you are connected to either an SDMC or an HMC as a valid user with proper authorization to create an LPAR.

Figure 2-26 shows the window to create a single partition system.

Figure 2-26 Creating a single partition system

When connected to the SDMC, select the managed server on which to create a virtual server. You see a window similar to Figure 2-27 when you are connected and have expanded the Hosts selection. The HMC window differs slightly. The names of the LPARs on Figure 2-27 are IBM ITSO names. You use names that are specific to your environment.

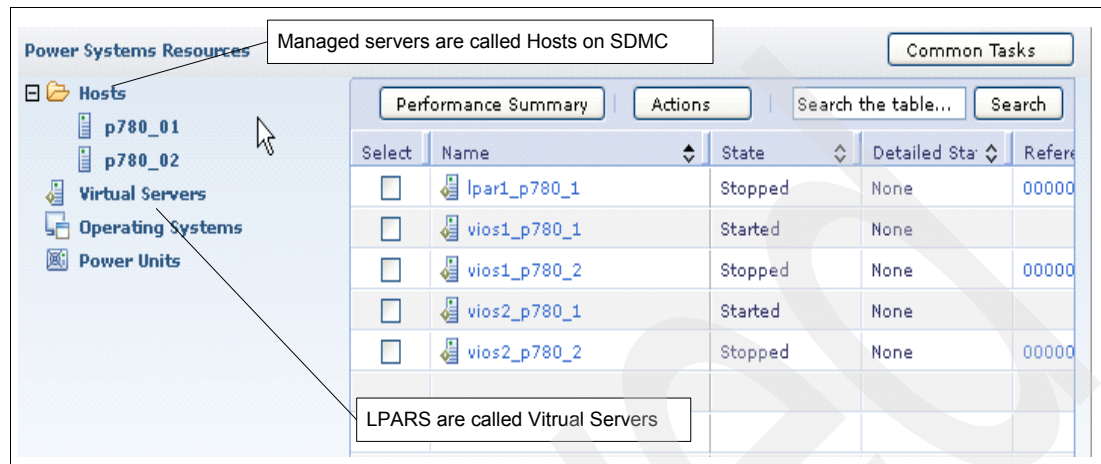


Figure 2-27 SDMC listing of available hosts

An LPAR requires the following elements:

- ▶ Processors
- ▶ Memory
- ▶ Storage
- ▶ Ethernet access to install the operating system through Network Installation Management (NIM)
- ▶ HMC for Resource Monitoring and Control (RMC) connectivity

2.10.1 Creating a simple LPAR

We now guide you through the creation of an LPAR using the SDMC. It is possible to perform the LPAR creation operations via the HMC, HMC command-line interface, or SDMC command-line interface. Refer to the *Hardware Management Console V7 Handbook*, SG24-7491, for command-line syntax.

Creating an LPAR

From the Power system resource menu, select the host on which to create an LPAR by selecting **Select Actions** → **Select the managed server (host)** → **System configurations** → **Create Virtual Server** (Figure 2-28).

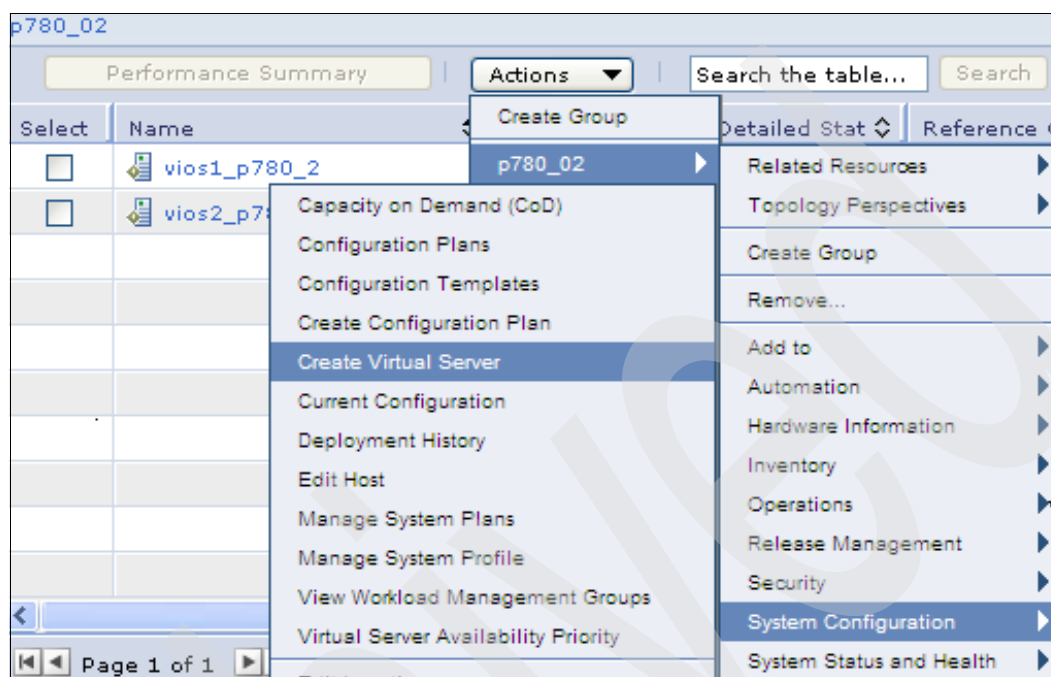


Figure 2-28 Creating a virtual server on P780_02 using the SDMC

Figure 2-28 shows the creation process. This operation opens the LPAR Basic Information tab.

Enter the following information based on your configuration:

- ▶ **Virtual server name (LPAR name):** This is the name with which the HMC or SDMC identifies your virtual server. This name is not the host name that is used by the application. It is not the operating system host name. To simplify the environment, you might make the host name and the LPAR name the same name. Example 2-13 shows an AIX command to check if the hostname and virtual server names differ. If you plan to migrate your virtual server using partition mobility features, you are required to have unique names for all partitions that might be migrated. Both source and destination hosts must have unique names.

Example 2-13 Using AIX command to see LPAR name and hostname

```
# lparstat -i | head
Node Name                : testlpar
Partition Name           : lpar1
Partition Number         : 6
Type                     : Shared-SMT
Mode                     : Uncapped
Entitled Capacity        : 0.30
Partition Group-ID       : 32774
Shared Pool ID           : 0
Online Virtual CPUs      : 2
Maximum Virtual CPUs     : 4
#
```

```
# hostname  
testlpar
```

- ▶ Virtual Server: IDID Number used by the SDMC to identify the virtual server.
- ▶ Select Environment: AIX or LINUX/i/virtual I/O server.
- ▶ Suspend capable: A *suspend capable* virtual server can be put “On Hold”. Refer to 3.1, “Live Partition Mobility (LPM)” on page 64, which discusses LPM in detail.

Setting up memory requirements

To set up the memory requirements, select the Memory Mode. Figure 2-29 shows the available modes.

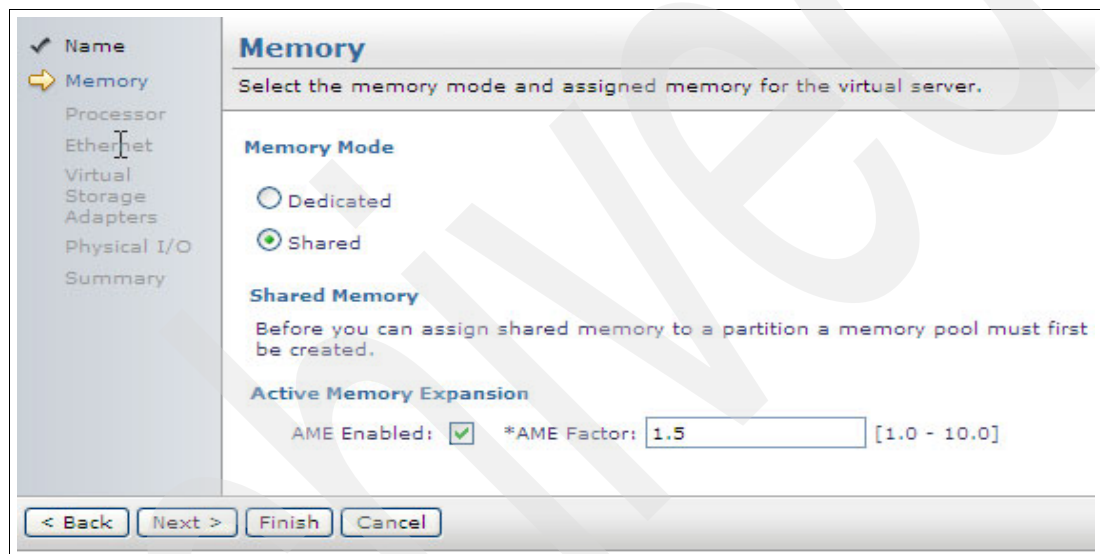


Figure 2-29 Available Memory Modes with AME and AMS

Memory mode: For memory mode, you have the following options:

- ▶ Dedicated memory mode, with or without Active Memory Expansion (AME)
- ▶ Shared Memory (AMS) with or without AME

Setting up processor requirements

To set up the processor requirements, select the processor mode. The processor can be shared or dedicated. There is a slight difference between SDMC and HMC menus. With the HMC, you can select the initial settings for *Minimum*, *Desired*, and *Maximum* processor and memory values. You can also specify virtual processors. In the SDMC, you use a single value to create the partition. You can later modify the values and attributes using Virtual Server Management. With the SDMC LPAR creation wizard, the memory is configured before the processors.

Figure 2-30 shows the initial processor settings and compares the SDMC settings with the HMC settings.

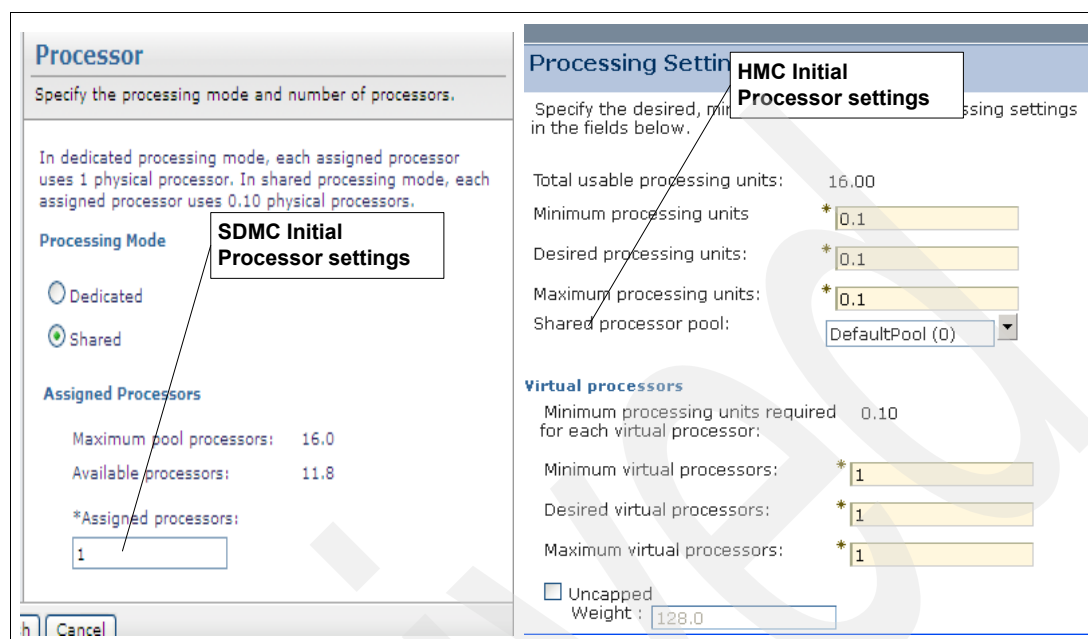


Figure 2-30 Processor settings on SDMC (left) and HMC (right)

Setting up Ethernet adapters

The choices include Shared Ethernet Adapter (SEA) and Integrated Virtual Ethernet (IVE) adapter. We explain the Shared Ethernet Adapter creation in 6.5.1, “Virtual I/O servers” on page 218.

You can use the Integrated Virtual Ethernet adapter for the LPAR, but because we are creating a mobile-capable LPAR, we use SEA and not the Host Ethernet Adapter.

Figure 2-31 on page 56 shows where to select the Virtual Ethernet V/S Host Ethernet adapter.

Mobile-capable LPAR: For a mobile-capable LPAR, select a VLAN that is part of a Shared Ethernet Adapter on one or more virtual I/O servers. Do not use the Host Ethernet Adapter. The virtual I/O server can use PCI or IVE adapters.

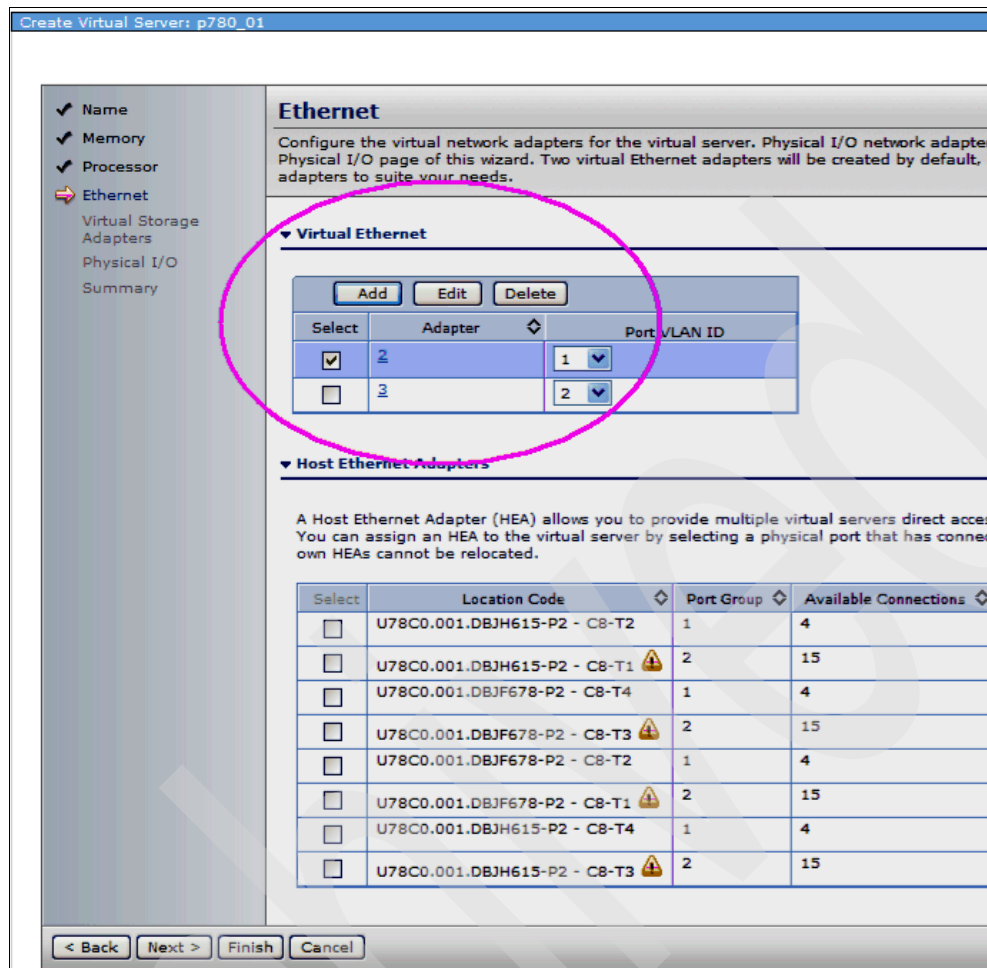


Figure 2-31 Virtual Ethernet

Selecting storage adapters

For the purpose of mobility and virtualization, do not use the physical adapters. An LPAR using AMS cannot have physical adapters allocated to it.

The SDMC provides an option to let the VIO manage your virtual adapter allocations, or you can manage them explicitly. The options that are shown in Figure 2-32 on page 57 do not exist on the HMC. For this example, we manage the resources manually.

We explain the options that are shown in Figure 2-32 on page 57:

Virtual Disks

This option allows a selected virtual I/O server to create a logical volume as a virtual disk to be used by the LPAR. This option is not a recommended method of creating disks for an LPAR. The virtual I/O server becomes a single point of failure (SPOF) and the LPAR cannot be migrated.

Physical Volumes

If you use Virtual SCSI adapters, you are able to select any physical disk that is visible to your selected virtual I/O server. The management of the disk allocation is done via the VIO `mkvdev` command.

Fibre Channel

This option creates an NPIV FC adapter, which is explained in detail in 2.7.2, “N_Port ID Virtualization (NPIV)” on page 43.

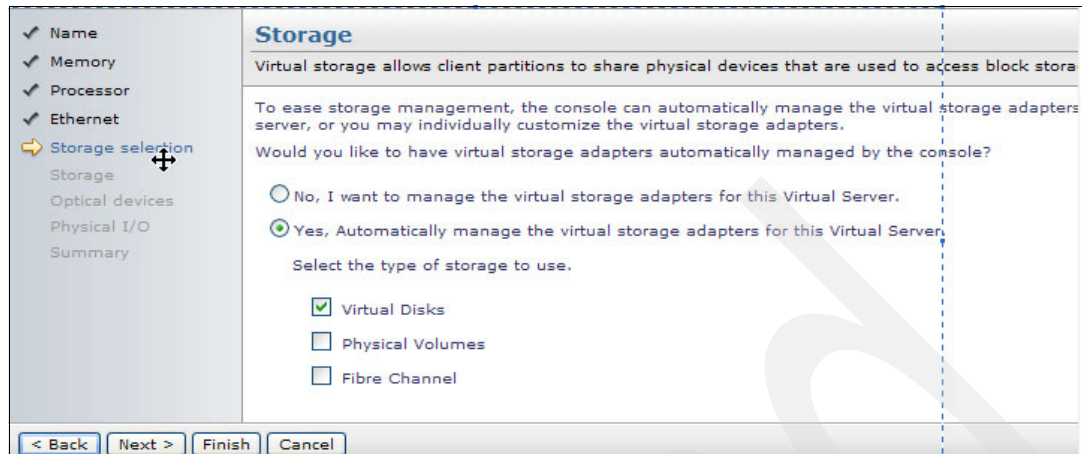


Figure 2-32 Virtual adapter management

Now, select the number of virtual adapter the LPAR can have (refer to Figure 2-33). This affects the adapter ID created in the next section. Specify the number and select create. A virtual adapter creation box is shown in Figure 2-33.

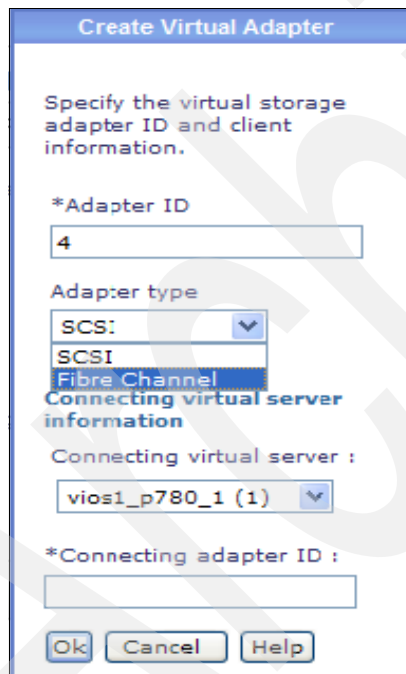


Figure 2-33 Virtual adapter selection

Although the adapter ID does not matter, the connecting adapter ID must match the VIO connecting adapter. This creates the “path” through which the hypervisor transfers the data.

After you complete this step, you see a summary of the setup. Review the setup summary and select **Finish** to complete the creation of the LPAR. See Figure 2-34 on page 58.

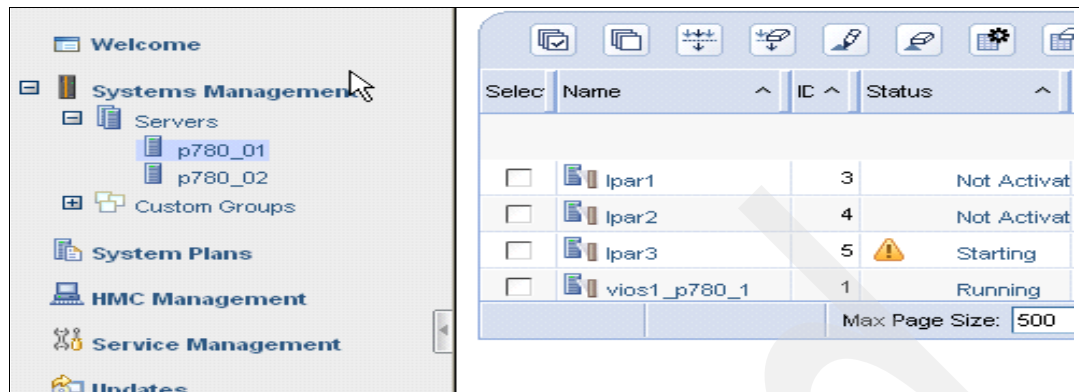


Figure 2-34 LPARS on the HMC

The created LPAR appears in the list of LPARS in the HMC, as shown in Figure 2-34, and it also appears in the list of virtual servers in the SDMC, as shown in Figure 2-35.

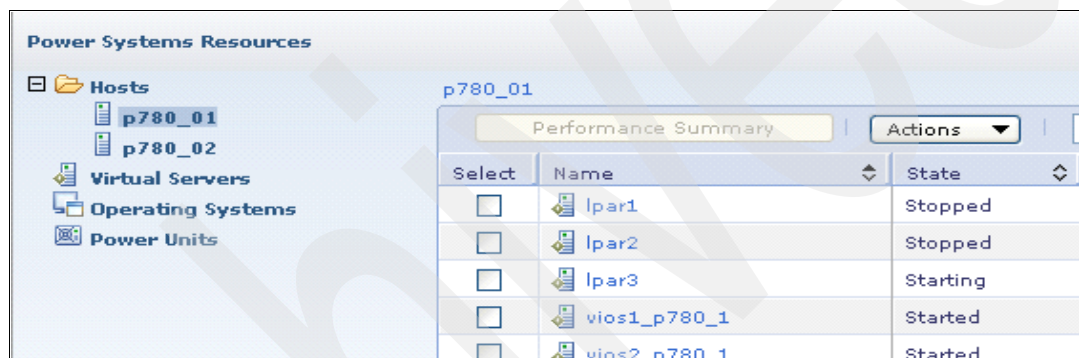


Figure 2-35 Virtual Servers on the SDMC

2.10.2 Dynamically changing the LPAR configurations (DLAR)

It is possible to change resource configuration of a running LPAR, for an example, if you need to add or remove virtual or physical adapters, or increase the allocated size of memory or processors on a running partition. When performing a dynamic LPAR operation, decide whether the resource must be added permanently on the LPAR or if it is temporary. If you need the resource permanently, you must indicate that you need the resource permanently on the LPAR profile.

Reconfiguring adapters

To add the adapter dynamically, the adapter must be available and not in use by any other partition. If there is another partition using the adapter, the adapter can be removed using a dynamic LPAR operation, which is shown in “A dynamic LPAR operation using the HMC” on page 357. Do not try to allocate resources in a partition profile as required, unless they are actually critical to the operation. Adapters that are allocated as required cannot be removed from a running LPAR. Consider this rule when creating an LPAR profile.

Reconfiguring an adapter using SDMC

Refer to the *IBM Systems Director Management Console: Introduction and Overview*, SG24-7860, for more LPAR operations using the SDMC. In this example, we only show the adapter reconfiguration operations using the SDMC. You can also reconfigure an adapter using the HMC.

Before adding an adapter, confirm which adapters are present in the LPAR, as shown on Example 2-14. Then, continue with the steps to add the adapters to the virtual server.

Example 2-14 Listing available adapters

```
# lsdev -Cc adapter
ent0 Available          Virtual I/O Ethernet Adapter (1-lan)
fcs0 Available 20-T1    Virtual Fibre Channel Client Adapter
fcs1 Available 21-T1    Virtual Fibre Channel Client Adapter
fcs2 Available 22-T1    Virtual Fibre Channel Client Adapter
fcs3 Available 23-T1    Virtual Fibre Channel Client Adapter
vsa0 Available          LPAR Virtual Serial Adapter
```

Follow these steps to add the adapters to the virtual server:

- 1. Log on to the **SDMC** and select **Virtual Servers** → **Virtual LPAR** → **Actions**, as summarized and shown in Figure 2-36.

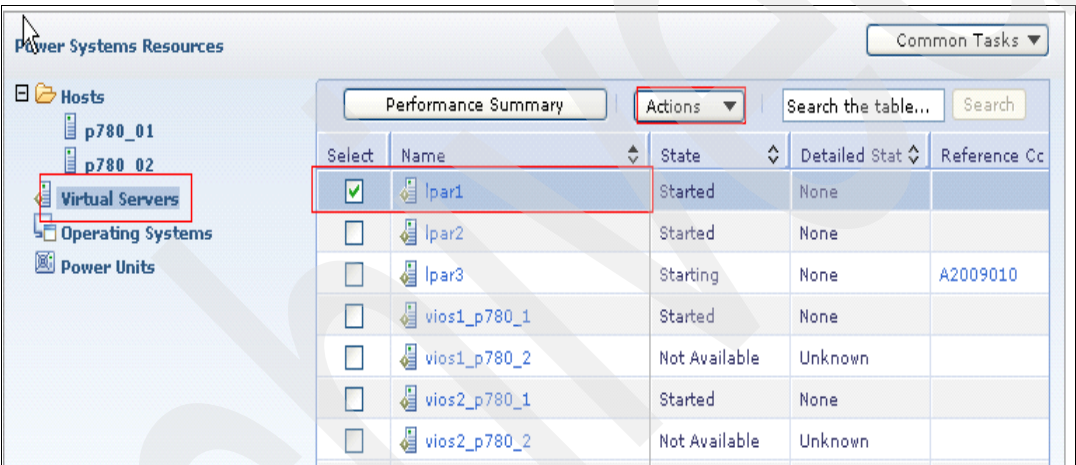


Figure 2-36 Selecting Actions on an LPAR

2. The Actions menu appears. Select **System Configuration** → **Manage Virtual Server**, as shown on Figure 2-37.

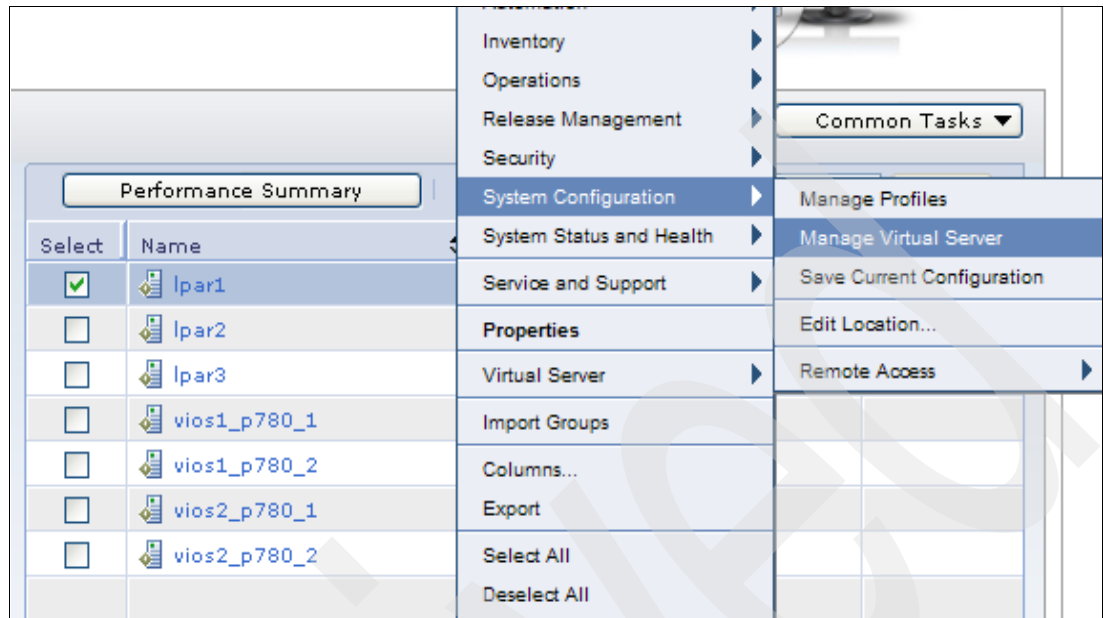


Figure 2-37 Managing an LPAR (Virtual Server) configurations

3. Now, we add a virtual adapter. The SDMC differentiates between the physical I/O adapters and the virtual adapters. The storage adapters refer to PowerVM virtual adapters. If you are adding physical adapters, you select the physical I/O. In this case, we use virtual devices. Select **Storage Adapters** → **Add**. Figure 2-38 shows adding a virtual adapter.

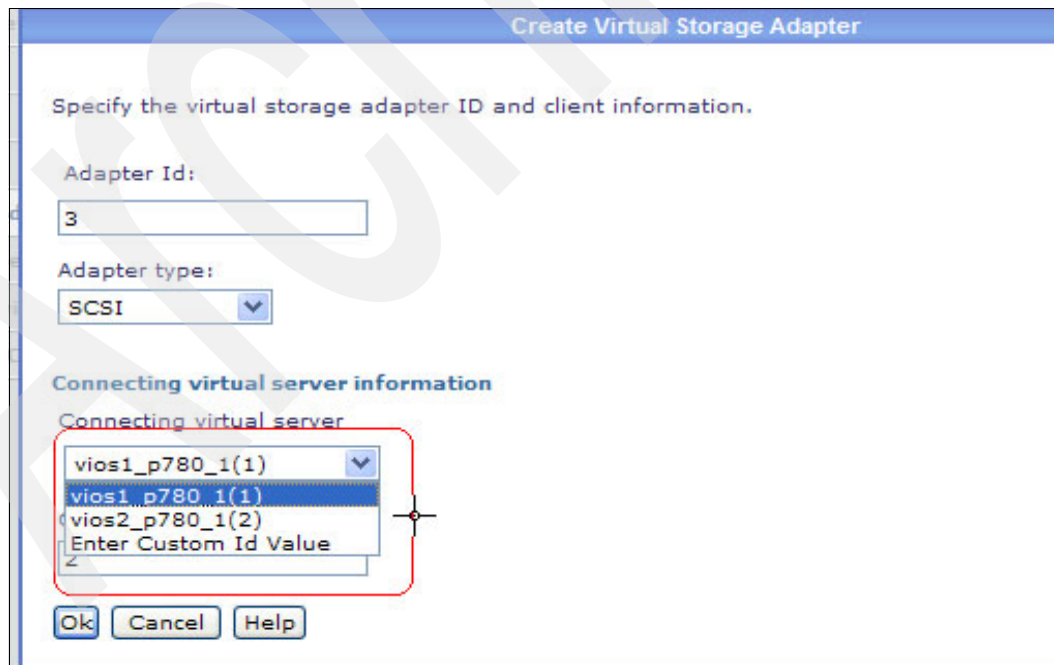


Figure 2-38 Adding a virtual adapter dynamically

Virtual I/O server connection: Figure 2-38 on page 60 shows the connecting device information. For a VIO client, you select the virtual I/O server to which this device connects. If this device was created for a virtual I/O server, the heading of the list box is “Connecting virtual server”, but the drop-down list contains other LPARs and not virtual I/O servers.

4. Click **Apply**. If you do not select **Apply**, the dynamic LPAR operation is rolled back.
5. On the LPAR command line, run the **cfgmgr** command to configure the dynamically added adapter, as shown in Example 2-15.

Example 2-15 Configuring an adapter with the cfgmgr command

```
# lsdev -Cc adapter
ent0 Available          Virtual I/O Ethernet Adapter (1-lan)
fcs0 Available 20-T1 Virtual Fibre Channel Client Adapter
fcs1 Available 21-T1 Virtual Fibre Channel Client Adapter
fcs2 Available 22-T1 Virtual Fibre Channel Client Adapter
fcs3 Available 23-T1 Virtual Fibre Channel Client Adapter
vsa0 Available          LPAR Virtual Serial Adapter
#
# cfgmgr
# lsdev -Cc adapter
ent0 Available          Virtual I/O Ethernet Adapter (1-lan)
fcs0 Available 20-T1 Virtual Fibre Channel Client Adapter
fcs1 Available 21-T1 Virtual Fibre Channel Client Adapter
fcs2 Available 22-T1 Virtual Fibre Channel Client Adapter
fcs3 Available 23-T1 Virtual Fibre Channel Client Adapter
vsa0 Available          LPAR Virtual Serial Adapter
vscsi0 Defined          Virtual SCSI Client Adapter
vscsi1 Available        Virtual SCSI Client Adapter

# cfgmgr
Method error (/usr/lib/methods/cfg_vclient -l vscsi0 ):
0514-040 Error initializing a device into the kernel.
```

Note: The **cfgmgr** command in Example 2-15 shows a method error on vscsi0, “vscsi1 has an associated Virtual I/O adapter and shows available and not defined”. The resource that was created requires a connecting resource on a virtual I/O server, and that resource does not exist. You must have a Virtual Server adapter for each Client adapter. Refer to *IBM Systems Director Management Console: Introduction and Overview*, SG24-7860.

6. If the adapter must be added permanently, you can add it into the profile. One way is to save the current configurations from the system Virtual Server Management tab. Select **Tasks** → **Save current configurations**, as shown in Figure 2-39 on page 62.

Note: The adapter reconfiguration using the HMC is shown in “A dynamic LPAR operation using the HMC” on page 357. We used the steps that are shown in the appendix to dynamically remove the adapter that was added in this example.

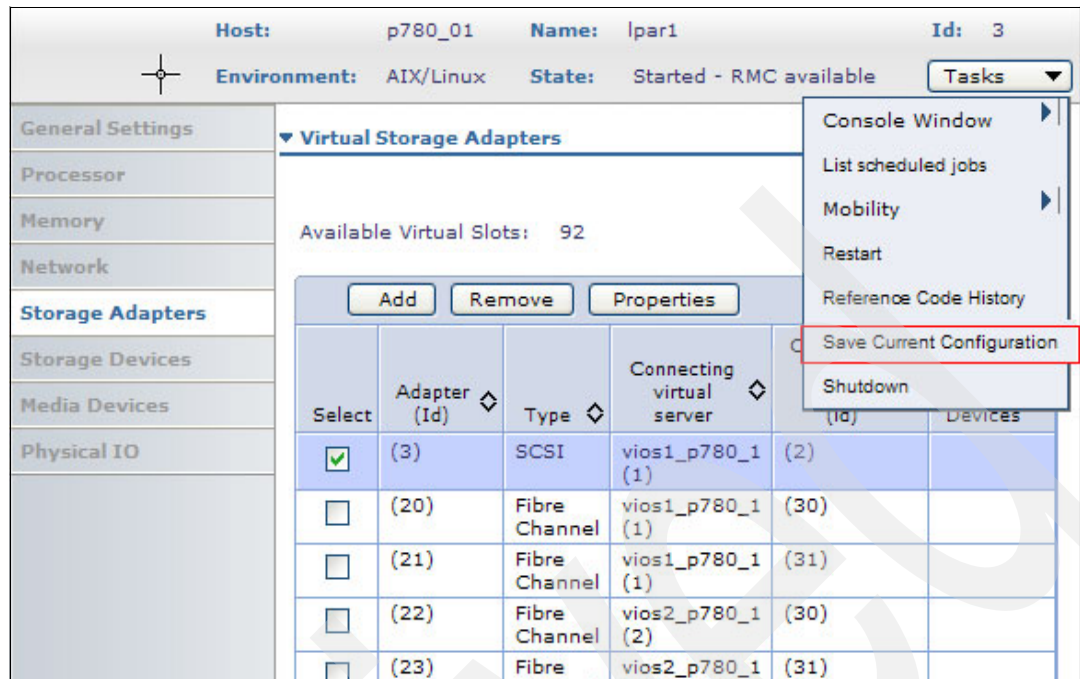


Figure 2-39 Permanently saving a dynamic LPAR operation

Memory and processor DLPAR operations

When performing operations on your managed server, certain resources are used by the hypervisor. The system must have enough available resources to allocate to the LPAR. If you are removing resources from the LPAR, this requirement does not apply. The dynamic LPAR (DLPAR) operation for both memory and processor reallocation is similar to the storage adapter example that was shown previously, except that you select *Processor* or *Memory* instead of the storage adapters.

If there are no resources available to allocate, you can reduce resources from other LPARs using dynamic LPAR, and then move them to the requesting partition.

Enhancing virtualization and RAS for higher availability

This chapter covers details of the RAS and virtualization features of the IBM Power Systems Enterprise Servers to help provide high systems availability to the applications hosted in these servers.

In this chapter, we describe the following topics:

- ▶ Live Partition Mobility (LPM)
- ▶ WPAR
- ▶ Partition hibernation
- ▶ IBM SystemMirror PowerHA
- ▶ IBM Power Flex
- ▶ Cluster Aware AIX (CAA)
- ▶ Electronic services and electronic service agent

3.1 Live Partition Mobility (LPM)

Live Partition Mobility is a PowerVM feature that allows you to move an LPAR from one physical Power system server to another physical Power system server. The following process shows the manual steps that are required to move a partition if you do *not* have LPM functionality. You can skip the manual steps and go directly to 3.1.1, “Partition migration” on page 65.

We list the steps to move a partition from one server to another server. This process is cumbersome and requires manual user intervention. We explain the following steps in more detail in 4.2.7, “Planning for Live Partition Mobility (LPM)” on page 113:

1. Create a new LPAR (client LPAR) on the destination server.
2. Create client virtual adapters for the LPAR on the destination server.
3. Dynamically create server virtual adapters for the virtual I/O server for the client.
4. Update the destination virtual I/O server profile with the same virtual adapters that are added dynamically.
5. Either create client VTD mappings if using virtual SCSI (vSCSI) for storage or create and zone N_Port ID Virtualization (NPIV) Fibre Channels (FCs). The disks being mapped must be the exact disks that are used by the moving partition on the source server.
6. Shut down the moving partition from the source server.
7. Start the partition that you created on the destination.
8. Remove the source LPAR virtual adapter from the virtual I/O server both dynamically and on the profile.
9. Remove the VTD mappings and NPIV mapping on the source virtual I/O server.
10. Remove the client partition from the source server.

LPM automates these steps, making the process faster because the manual intervention is minimized.

Moving a non-PowerVM environment: It is possible to move a non-PowerVM environment, which we explain in the 4.6, “Migrating from POWER6 to POWER7” on page 132. This procedure is still valid for systems prior to POWER6.

Using LPM offers many advantages, including the ability to migrate a partition without having to stop the application. The *IBM PowerVM Live Partition Mobility*, SG24-7460, publication explains LPM in detail. This IBM Redbooks publication also provides the reasons to use LPM and the advantages of LPM. The advantages include server consolidation, saving electricity, the ability to perform scheduled maintenance of the server, such as firmware upgrades, and the ability to handle growth where the current system cannot handle the load. For example, you can move a workload from a POWER6 550 to a POWER 780 by using LPM.

We only briefly introduce the LPM technology in this book, because the *IBM PowerVM Live Partition Mobility*, SG24-7460, publication explains LPM in detail. The Redbooks publication further details all components that might be involved in the mobility process, several of which might be transparent to the administrator. The following components are involved in the mobility process:

- ▶ Systems Director Management Console (SDMC) or Hardware Management Console (HMC)
- ▶ Resource Monitoring and Control (RMC)

- ▶ Dynamic LPAR resource manager
- ▶ Virtual Asynchronous Service interface (VASI)
- ▶ Time reference
- ▶ Mover service partition
- ▶ Power hypervisor
- ▶ Partition profile
- ▶ Virtual I/O server

Migration road map for LPARs

The following methods are used to migrate a partition from one server to another server. We list the methods in order of our preference and recommendation. We discuss these methods in 4.2.7, “Planning for Live Partition Mobility (LPM)” on page 113.

- ▶ Active migration
- ▶ Inactive migration
- ▶ Manual migration using virtual I/O server but not the LPM option
- ▶ Manual migration by pointing SAN volumes to the new server
- ▶ Manual migration using **alt_disk_copy**
- ▶ Manual migration using **mksysb**

High-availability clusters versus LPM

LPM complements cluster technologies, but it does not provide cluster functionality. Table 3-1 explains the differences.

Table 3-1 Clusters versus LPM

Cluster	LPM
Handles unplanned failures of applications, servers, and components.	Handles planned migration of LPARs from one server to another server
Operating system is not the same image.	Same operating system image
Update and changes must be applied on each node on the cluster.	Same operating system image
Basic operating system configurations, such as the IP addresses, might not be the same.	Same operating system image

3.1.1 Partition migration

Partition Mobility uses two methods:

- ▶ Inactive: Where the services are stopped and the LPAR is shut down
- ▶ Active: Services remain running

The SDMC/HMC controls whether the migration is active or inactive based on the partition status. If the partition is shut down, it assumes an inactive migration. If it is running, it assumes an active migration. The SDMC/HMC also performs a validation. If the requirements for LPM are not met, the migration is not attempted.

Two methods exist for performing the migration:

- ▶ Point in time: This option is administrator driven and requires an administrator to be connected to the HMC and to follow a simple wizard.

- Automated: This option can be incorporated and scheduled into system management scripts.

3.1.2 Migration preparation

In both active and inactive migrations, you must ensure that you meet the prerequisites before LPM can be executed. Table 3-2 lists the prerequisites.

Prepare the SDMC/HMC for migration. The SDMC/HMC controls the migration process. It copies LPAR profile information from the source server to the destination server's hypervisor.

- Use *Dynamic LPAR* operations to remove physical or dedicated devices. Depending on the LPAR profile, it might not be possible to dynamically remove specific resources, especially physical I/O adapters that are marked as required. In this case, you have to shut down the partition.
- Run the migration validation. At the validation stage, no resource changes are done.

When the validation is complete, the SDMC/HMC on the destination server creates a shell partition that is similar to the source server. This shell partition ensures that the required resources are reserved.

Only partition profiles that have been activated can be migrated. To make sure that a profile is capable of being migrated, activate that profile. You do not need to start the operating system. You can activate the profile into System Management Services (SMS) mode and then shut it down again.

Selecting a capable virtual I/O server

Because a virtual I/O server is a requirement for the migration process, you can confirm if there is a capable virtual I/O server by running the migration validation. An example of validation is shown on Figure 3-1 on page 69. You can also use the `lslparmigr` command.

Selecting a capable mover service

At least one virtual I/O server in the same VLAN as the moving partition must be selected as a mover service partition. This mover service partition is a requirement for migrating an active partition.

Table 3-2 lists items to consider before attempting the migration. Use the table as a checklist or quick guide. Although these considerations are listed under the LPM discussion, most of them apply to manual migration, as well.

Table 3-2 Migration readiness checklist

Consideration	Requirements	Valid for
POWER6 and later	Confirm that all devices are supported between source and destination	Active and inactive migrations
CPU and memory resources	Make sure that the destination server has available resources	Active migration reserves resources. An inactive migration does not need to reserve resources.
Server on battery	Source server can be running on battery power. The destination server cannot be running on battery power.	Destination server only

Consideration	Requirements	Valid for
Logical memory block size (LMB)	Same size between source and destination server	Only active migration
Barrier synchronization registers (BSR)	Active migration LPAR must not use BSR.	Only active migration
Large pages	Active mobile LPAR cannot use large pages. Inactive LPARs can use large pages.	Only active migration
Storage pools	Not supported	Active and inactive migrations
Logical Volume VTDs	Not supported	Active and inactive migration
SAN	Additional settings must be confirmed, including reserve policy	Active and inactive. Must be vSCSI or NPIV
Internal disks	Not supported	Active and inactive migration
IVE	Supported only if it is part of a shared Ethernet adapter (SEA) on a VIO. All clients to use SEA.	Both active and inactive
Other physical adapter	Remove all physical adapters using DLAR	Active migration. These adapters are removed with inactive migration
VLAN	Both virtual I/O servers must access the VLAN used by the mobile LPAR	Active and inactive migration
LPAR name	Must be unique across the source and destination	Active and inactive migration
LPAR state	Depends on migration	Shut down for inactive and running for active
HMC communication with virtual I/O server	RMC communication is required between HMC and a managed service provider.	Only active migration
Inter HMC/SDMC ssh keys	Allows one managed server to communicate with another without prompting for a password	Remote HMC/SDMC
HMC communication with moving partition	RMC communication is required between HMC and LPAR to get memory state	Only active migration
HMC versions	Specifies required version: HMC Version 7 Release 7.1.0 SP1 or later	Active and inactive migration
Virtual I/O server as an MSP	At least one VIO on the server and destination must be used as an MSP partition	Only valid for active migration
Virtual I/O server version	Minimum 2.1.12 or higher required for on POWER7 Remote migration	Active and inactive migration

Consideration	Requirements	Valid for
Type of LPAR	Only AIX and Linux. Virtual I/O server and IBM i servers cannot be migrated. Ensure that the version supports the hardware.	Active and inactive migration

3.1.3 Inactive migration

Before performing an inactive migration, check that the partition is shut down. Inactive migration does not need a mover service partition. Check that you meet the requirements that are specified in Table 3-2 on page 66. Notice that inactive migration requires fewer prerequisites than active migration. Refer to 6.3.1, “Inactive migration from POWER6 to POWER7 using HMC and SDMC” on page 210 where an example of inactive migration is performed.

3.1.4 Active migration

Active migration is performed with running clients. A *Mover Service Partition*, which can be any virtual I/O server on both the source and destination system, is required to keep the services available. The SDMC/HMC copies the physical memory to the destination server.

LPM keeps the applications running. Regardless of the size of the memory that is used by the partition, the services are not interrupted, the I/O continues accessing the disk, and network connections keep transferring data. The *IBM PowerVM Live Partition Mobility*, SG24-7460, publication lists the following memory states and the migration sequence. LPAR active run time states that are copied to the destination server include the following information:

- ▶ Partition's memory
- ▶ Hardware page table (HPT)
- ▶ Processor state
- ▶ Virtual adapter state
- ▶ Non-volatile RAM (NVRAM)
- ▶ Time of day (ToD)
- ▶ Partition configuration
- ▶ State of each resource

The mover service partitions on the source and destination, under the control of the SDMC/HMC, move these states between the two systems. See the flow indicators in Figure 3-1 on page 69.

For active partition migration, the transfer of the partition state follows this path:

1. From the mobile partition to the source system's hypervisor
2. From the source system's hypervisor to the source mover service partition
3. From the source mover service partition to the destination mover service partition
4. From the destination mover service partition to the destination system's hypervisor
5. From the destination system's hypervisor to the partition shell on the destination system

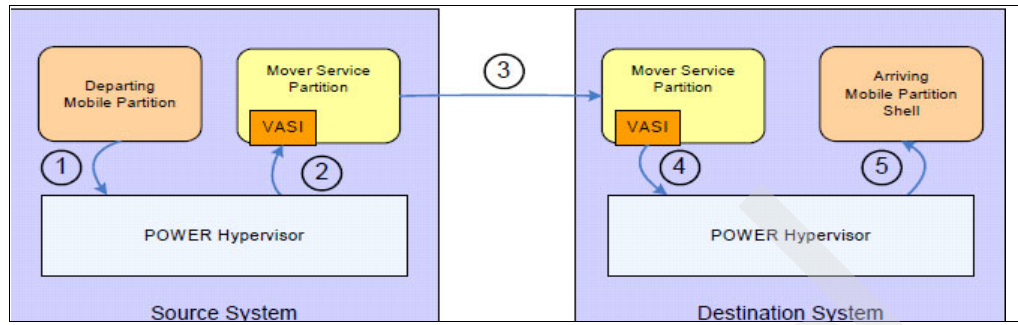


Figure 3-1 Active migration partition state transfer path

We show migration scenarios in 6.3, “Live Partition Mobility (LPM) using the HMC and SDMC” on page 210 and in 6.4, “Active migration example” on page 214.

3.2 WPAR

The IBM PowerVM offering extends the virtualization further to include software virtualization. This function is called *Workload Partition (WPAR)*. A *logical partition (LPAR)* is a hardware-based partitioning feature that allows you to create multiple independent operating system environments, which are called LPARs. Each LPAR can run a version of either AIX, Linux, or IBM i. A WPAR is built within an AIX partition. Therefore, WPARs are software-created, virtualized operating system environments within a single instance of the AIX operating system.

WPAR: The concept of a WPAR was introduced in AIX 6.1.

Each WPAR is seen as a separate operating system that is independent of any other WPAR within the same LPAR. We differentiate between a WPAR and an LPAR by referring to the LPAR in which a WPAR operates as the *Global Environment*, because the LPAR has a global view of all resources, and the WPARs are created within the LPAR. Each WPAR hosts applications that are invisible to other WPARs within the same Global Environment. The Global Environment is an AIX LPAR. We suggest that you do not use the LPAR to host applications while it hosts WPARs.

You define a hypervisor profile in the Global Environment. Devices are attached to the Global Environment with hardware resources. You can see the global/LPAR in the system partition list on the HMC or Virtual Server on the SDMC. You cannot see a WPAR on the HMC/SDMC.

The Global Environment has full control of the WPARs. But the WPARs cannot view the global WPARs on the same LPAR and cannot overlap. One WPAR is not aware of any other WPAR in the same AIX Global Environment. Only by using standard TCP/IP can two or more WPARs communicate; although, the actual communication is a loopback.

We describe the level of granularity by using the following hierarchy:

- **Managed server:** This server is a Power System. In our environment, it is an IBM POWER7.
- **An LPAR:** Within a managed server, you create one or more LPARs. Each partition is capable of running its own operating system space. The operating system version might differ per LPAR, which is called the Global Environment.

- ▶ A WPAR: Inside a single running copy of AIX, you can create partitions (WPARs). These partitions have the same characteristics as an independent LPAR. All WPARs must be on the same operating system level as the global LPAR. Detached/rootvg WPARs have a separate /usr and /opt, and can have software installed in them that is not part of the LPAR (Global Environment).

An LPAR can be created on an IBM Power server starting from POWER4 and later. A WPAR does not depend on the hardware. It is a software virtualization feature. The minimum operating system requirement is AIX Version 6.1 or later. Thus, if you can install AIX Version 6.1 on a Power server, you can set up a WPAR.

An application sees a WPAR as an LPAR, and the WPAR still has the following characteristics displayed in LPARs:

- ▶ Private execution environments
- ▶ Isolation from other processes outside the WPAR
- ▶ Dedicated network addresses and filesystems
- ▶ Interprocess communication that is restricted to processes executing only in the same Workload Partition

Because a system WPAR can be viewed as an independent operating system environment, consider separating the WPAR system administrator from the Global system administrator.

3.2.1 Types of WPARs

This section describes the kinds of WPARs.

System WPAR

A *System WPAR* has the following characteristics:

- ▶ A System WPAR is a flexible, complete copy of an AIX instance.
- ▶ It has a mixture of shared and dedicated file systems.
- ▶ Each system WPAR has separate init processes, daemons, users, resources, file systems, user IDs, process IDs, and network addresses. Applications and interprocess communication (IPC) are restricted to processes running in the same workload partition.
- ▶ Can be attached or detached WPARs:
 - Attached System WPARs have shared /opt and /usr with the Global Environment.
 - Detached System WPARs have dedicated /opt and /usr as the Global Environment.

Application WPAR

An *Application WPAR* is a process-based workload environment. It starts when a process is called and stops when the process terminates. The Application WPAR cannot be detached from the Global Environment. Applications in workload partitions are isolated in terms of process and signal, and they can be isolated in the file system space.

An Application WPAR has the process isolation that a System WPAR provides, except that it shares file system namespace with the Global Environment and any other Application WPAR that is defined in the system. Other than the application itself, a typical Application WPAR runs an additional lightweight init process in the WPAR.

Figure 3-2 on page 71 shows a diagram that was adapted from the *Exploiting IBM AIX Workload Partitions*, SG24-7955, publication. Refer to this publication for more information about WPARs.

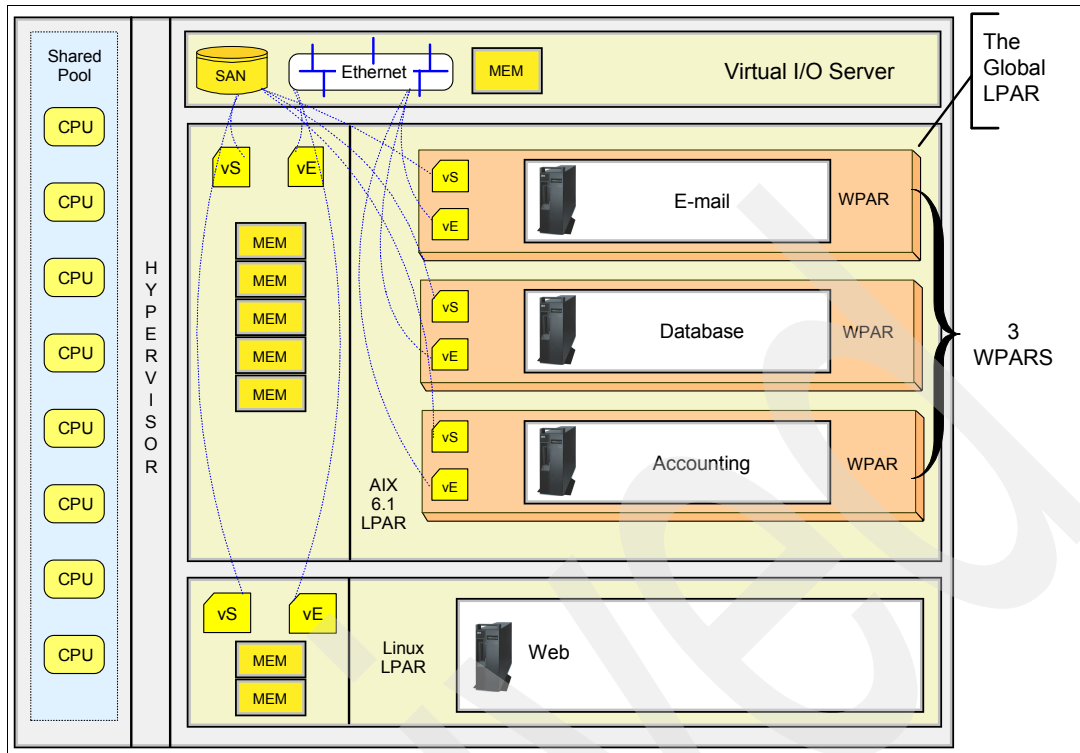


Figure 3-2 Three WPARs within an LPAR

WPARs have improved since AIX 6.1. Most of the improvements, as well as detailed information, are discussed in *Exploiting IBM AIX Workload Partitions*, SG24-7955. Table 3-3 summarizes the improvements since WPARs in AIX 6.1.

Table 3-3 WPAR improvements with the AIX 6.1 and later operating systems

AIX version	WPAR improvement
AIX 6.1 Base Level (GA)	<ul style="list-style-type: none"> ▶ Initial support, including mobility using synchronous checkpoint/restart ▶ First WPAR manager release
AIX 6.1 TL1	<ul style="list-style-type: none"> ▶ Network File System (NFS) support for WPAR
AIX 6.1 TL2	<ul style="list-style-type: none"> ▶ Asynchronous mobility ▶ Per-WPAR routing ▶ Name-mapped network interfaces ▶ Network Installation Management (NIM) support for WPAR
AIX 6.1 TL3	<ul style="list-style-type: none"> ▶ Storage disk devices support
AIX 6.1 TL4	<ul style="list-style-type: none"> ▶ rootvg WPAR ▶ SAN mobility ▶ WPAR manager integration with IBM Systems Director ▶ VxFS support
AIX 6.1 TL5	<ul style="list-style-type: none"> ▶ WPAR Error Logging Framework (RAS)
AIX 6.1 TL6	<ul style="list-style-type: none"> ▶ Virtual SCSI (vSCSI) disk support ▶ WPAR migration to AIX 7.1

AIX version	WPAR improvement
AIX 7.1 Base Level (GA)	<ul style="list-style-type: none"> Everything that is supported in AIX 6.1, plus Fiber Channel (FC) adapter support, Versioned WPARs running AIX 5.2, and Trusted Kernel extension support

Advantages of WPARs over LPARs

WPARs have the following advantages over LPARs:

- ▶ WPARs are much simpler to manage, and they can actually be created from the AIX command line or through SMIT unlike the LPARs.
- ▶ It is a requirement to install patches and technology upgrades to every LPAR. Each LPAR requires its own archiving strategy and disaster recovery strategy. However, this requirement is not the same with a WPAR, because it is part of a single LPAR.
- ▶ As many as 8,000 WPARs can be created in a single LPAR, which means that 8,000 applications can run in an isolated environment.

Rather than a replacement for LPARs, WPARs are a complement to LPARs. WPARs allow you to further virtualize application workloads through operating system virtualization. WPARs allow new applications to be deployed much more quickly, which is an important feature.

3.2.2 Creating a WPAR

In this section, we show an example of creating, starting, and stopping a WPAR, as shown in Example 3-1. Creating this WPAR took one minute and 15 seconds. Example 3-1 only shows the beginning and ending output lines of the WPAR creation process.

For further WPAR administration and management, refer to Chapter 4 of *Exploiting IBM AIX Workload Partitions*, SG24-7955.

Example 3-1 Creating a WPAR on a global called rflpar20

```
# uname -a
AIX rflpar20 1 6 00EE14614C00
# mkwpar -n wpartest1 -N address=172.16.21.61 netmask=255.255.252.0
mkwpar: Creating file systems...
/
/home
/opt
/proc
/tmp
/usr
/var
Mounting all workload partition file systems.
x ./usr
x ./lib
x ./admin
x ./admin/tmp
x ./audit
x ./dev
x ./etc
x ./etc/check_config.files
x ./etc/consdef
x ./etc/cronlog.conf
x ./etc/csh.cshrc
```

```

.
.
.
A few lines were skipped
.
.
rsct.core.hostrm          3.1.0.1      ROOT      COMMIT    SUCCESS
rsct.core.microsensor     3.1.0.1      ROOT      COMMIT    SUCCESS
syncroot: Processing root part installation status.
syncroot: Installp root packages are currently synchronized.
syncroot: RPM root packages are currently synchronized.
syncroot: Root part is currently synchronized.
syncroot: Returns Status = SUCCESS
Workload partition wpartest1 created successfully.
mkwpar: 0960-390 To start the workload partition, execute the following as root:
startwpar [-v] wpartest1

#

```

After creating the WPAR, you can see it and start it, as shown in Example 3-2.

Example 3-2 Listing and starting the WPAR

```

# lswpar
Name          State  Type  Hostname  Directory          RootVG WPAR
-----
wpartest1    D      S      wpartest1 /wpars/wpartest1  no
#
#
#
# startwpar wpartest1
Starting workload partition wpartest1.
Mounting all workload partition file systems.
Loading workload partition.
Exporting workload partition devices.
Starting workload partition subsystem cor_wpartest1.
0513-059 The cor_wpartest1 Subsystem has been started. Subsystem PID is 6553710.
Verifying workload partition startup.
#
#
# lswpar
Name          State  Type  Hostname  Directory          RootVG WPAR
-----
wpartest1    A      S      wpartest1 /wpars/wpartest1  no

```

Because a WPAR behaves like a normal LPAR, we can access it, as shown in Example 3-3.

Example 3-3 Accessing a WPAR using telnet

```

# telnet 172.16.21.61
Trying...
Connected to 172.16.21.61.
Escape character is '^'.

```

```
telnet (wpartest1)
```

AIX Version 6
Copyright IBM Corporation, 1982, 2010.
login:

In Example 3-4, we show a few AIX commands that have executed within the WPAR.

Example 3-4 Executing AIX commands within a WPAR

```
AIX Version 6
Copyright IBM Corporation, 1982, 2010.
login: root
*****
*
*
* Welcome to AIX Version 6.1!
*
*
* Please see the README file in /usr/lpp/bos for information pertinent to
* this release of the AIX Operating System.
*
*
*****

# hostname
wpartest1
# ifconfig -a
en0:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT
,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
    inet 172.16.21.61 netmask 0xfffffc00 broadcast 172.16.23.255
    tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
lo0:
flags=e08084b,c0<UP,BROADCAST,LOOPBACK,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT,LAR
GESEND,CHAIN>
    inet 127.0.0.1 netmask 0xff000000 broadcast 127.255.255.255
    inet6 ::1%1/0
    tcp_sendspace 131072 tcp_recvspace 131072 rfc1323 1

# df
Filesystem      512-blocks      Free %Used    Iused %Iused Mounted on
Global          196608        138576   30%      1845    11% /
Global           65536         63808    3%         5     1% /home
Global          884736        555544   38%      8394   11% /opt
Global           -             -         -         -     - /proc
Global          196608        193376    2%         10    1% /tmp
Global          5505024       1535416   73%      41390   18% /usr
Global          262144        110016   59%       2797   19% /var

# exit
Connection closed.
```

In Example 3-5, we stop and remove the WPAR.

Example 3-5 Stopping and removing a WPAR

```
# stop wpar wpartest1
ksh: wpar: Specify a process identifier or a %job number.
# stopwpar wpartest1
```

```

Stopping workload partition wpartest1.
Stopping workload partition subsystem cor_wpartest1.
0513-044 The cor_wpartest1 Subsystem was requested to stop.
stopwpar: 0960-261 Waiting up to 600 seconds for workload partition to halt.
Shutting down all workload partition processes.
wio0 Defined
Unmounting all workload partition file systems.

# rmwpar wpartest1
rmwpar: Removing file system /wpars/wpartest1/var.
rmlv: Logical volume fslv03 is removed.
rmwpar: Removing file system /wpars/wpartest1/usr.
rmwpar: Removing file system /wpars/wpartest1/tmp.
rmlv: Logical volume fslv02 is removed.
rmwpar: Removing file system /wpars/wpartest1/proc.
rmwpar: Removing file system /wpars/wpartest1/opt.
rmwpar: Removing file system /wpars/wpartest1/home.
rmlv: Logical volume fslv01 is removed.
rmwpar: Removing file system /wpars/wpartest1.
rmlv: Logical volume fslv00 is removed.

```

3.2.3 Live Application Mobility (LPM)

WPARs also have the capability to be actively relocated (or migrated) from one AIX LPAR to another AIX LPAR. This process is called *Live Application Mobility*. Live Application Mobility refers to the ability to relocate a WPAR from one Global AIX LPAR to another Global AIX LPAR. It uses *checkpoint/restart* capabilities that allow the WPAR to hold the application state. The Global LPAR can be on the same server or a separate server, which shows the flexibility and portability of a WPAR. Live Application Mobility is an operating system feature and independent of the hardware. Because a WPAR resides over an LPAR, migrating the LPAR to a separate server using LPM also migrates the WPARs within the LPAR. The WPAR has three ways of migrating:

- ▶ Explicit migration of a WPAR to a separate LPAR (Global Environment) within the same server
- ▶ Explicit migration of a WPAR to a separate LPAR
- ▶ Implicit migration of a WPAR due to the Global Environment migration using LPM. In this case, the WPAR remains part of the same LPAR.

Any hardware that can run AIX 6.1 is supported for Live Application Mobility. That is, you can migrate a WPAR from a Global Environment running on a POWER5 server to a Global Environment running on a POWER7 server and vice versa. LPM is a POWER6 and POWER7 feature.

Mobility can be inactive or active. In an inactive migration, a WPAR has to be shut down. In an active migration, the WPAR is migrated while the applications are active.

Explicit mobility can be performed in one of two ways

You can perform explicit mobility by using either an NFS-mounted WPAR migration or a rootvg WPAR:

- ▶ NFS-mounted WPAR migration

The relocation of a WPAR involves moving its executable code from a source LPAR to a destination LPAR while keeping application data on a common Network File System (NFS)

that is visible and accessible to both the source and destination LPARs. AIX operating system binaries can be stored in file systems that are local to the hosting LPAR.

Figure 3-3 shows an NFS-mounted WPAR migration.

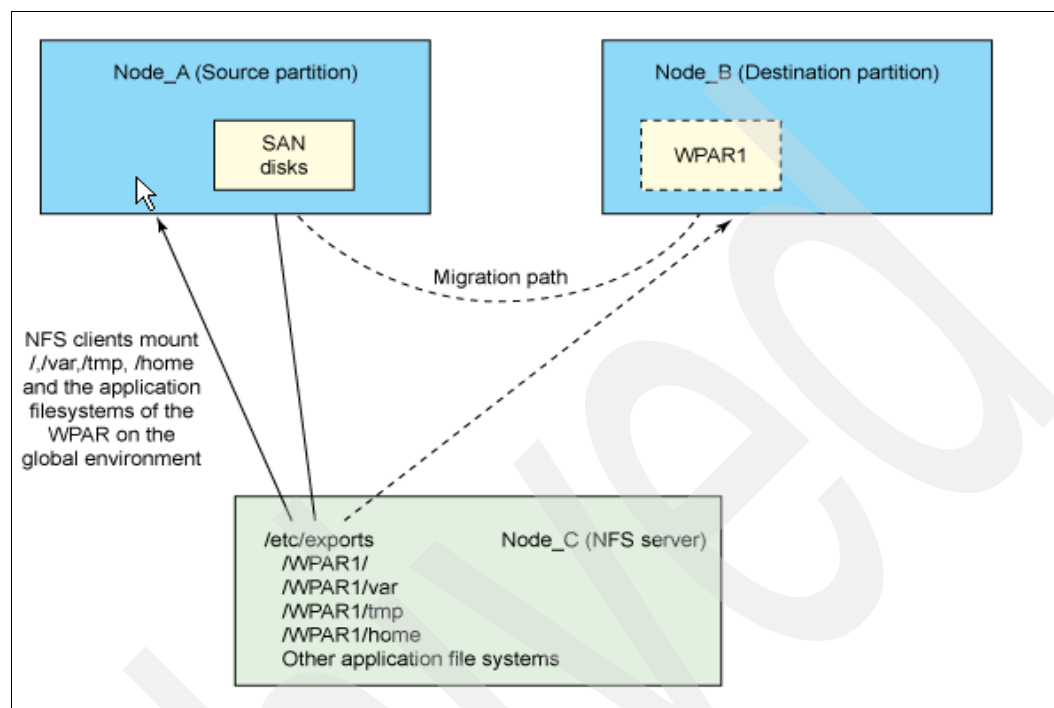


Figure 3-3 NFS-mounted WPAR migration

The setup that is shown in Figure 3-3 consists of two LPARs, NODE_A and NODE_B, and NODE_C. NODE_A is the source LPAR, which hosts WPAR1. Node_B is the destination LPAR to which WPAR1 will be migrated. Node_C is an NFS server that is required to support workload partition mobility. Before creating WPAR1, create its file systems (/ , /var, /tmp, /home, and the application file systems) on the NFS server. These file systems are exported to Node_A, Node_B, and WPAR1. While creating WPAR1, its file systems are mounted on Node_A and WPAR1. When WPAR1 migrates from Node_A to Node_B, its file systems are mounted on Node_B and unmounted from Node_A. In this way, the WPAR migration has to rely on common NFS file systems that are hosted on a separate NFS server.

However, the drawback of this setup is that certain applications might not support data access over NFS. Also, to eliminate the need of NFS services, the concept of “*rootvg WPAR*” was introduced.

► rootvg WPAR

In Figure 3-4, SAN disks are allocated to both Node A and Node B. The disks are shared across both nodes. The SAN storage disks are assigned to the System WPARs while creating the WPAR itself. The root file systems (/ , /usr, /opt, /home, /tmp and /var file systems) of the WPAR are created on the storage disks that are assigned to the WPAR.

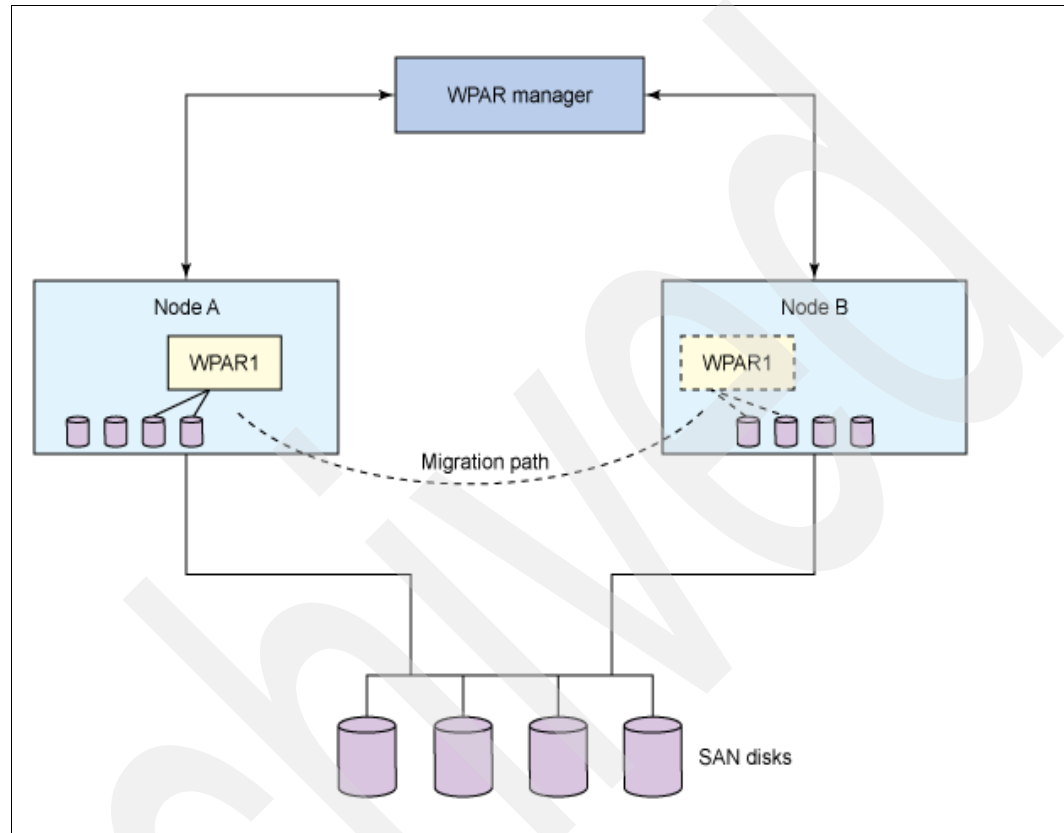


Figure 3-4 SAN-based WPAR mobility

WPAR1 is assigned the disks from the SAN subsystem, and these disks are also seen by Node B. Here, the migration can be done from NODE A to NODE B without needing NFS services. This kind of setup has been supported since AIX 6.1 TL4.

WPAR Manager: WPAR migration requires the use of WPAR Manager, which is a plug-in to the IBM Systems Director and the SDMC.

3.3 Partition hibernation

POWER7 provides another virtualization feature: *hibernation* or *suspend/resume*. With LPM, the memory state is transferred from one server hypervisor to another server hypervisor. Hibernation takes the memory state and stores it on a non-volatile storage device, which provides the ability to suspend a running partition and restart it at a later stage. On resume, the hypervisor reads the memory structures from a virtual I/O server back into the partition so that all applications that were running can continue where they left off. Resuming a partition is not limited to a single server. You can suspend a partition on one server, move it with inactive partition mobility, and resume it on another server.

For example, you suspend a data warehouse partition on a Power 795 to run payroll. The payroll runs longer than expected and you need the data warehouse server up and running. You can migrate the suspended data warehouse partition to a Power 780, and resume it there.

You might suspend a partition for the following reasons:

- ▶ Long running applications. The batch runs can be suspended during online periods and resumed at the end of the day.
- ▶ Disruptive firmware upgrades. Remember, updates are not disruptive.
- ▶ Disruptive hardware maintenance. CEC Hot Add Repair Maintenance (CHARM) allows certain hot hardware maintenance.
- ▶ LPARs use IVE and physical adapters that cannot be migrated due to LPM prerequisites, or if you do not have a server with spare capacity to which to migrate.

Previously, you stopped the process and started over.

Consider these factors before implementing partition hibernation:

- ▶ HMC Version 7 Release 7.3
- ▶ POWER7 server
- ▶ The partition can be running on either POWER6 or POWER7 mode
- ▶ Set up the partition as suspendible
- ▶ AIX 7.1 SP1 and AIX 6.1 TL6 SP1. Linux and IBM System i are not supported.
- ▶ PowerVM Virtual I/O Server 2.2.0
- ▶ Storage pools (used to save the partition state)
- ▶ No huge pages
- ▶ The virtual I/O server cannot be suspended, but it can be restarted while clients are suspended, because the memory states are on non-volatile storage.

Active Memory Sharing: If used with Active Memory Sharing (AMS), plan for a paging device that is larger than the AMS memory storage pool. The suspend resume uses the same paging devices that are used for AMS, which is also used for hibernation.

To configure partition hibernation, use the following steps:

1. Confirm that your server is capable of suspend and resume. On the HMC, select **System Management** → **Servers**. Select the server you want. Click **Properties** → **Capabilities**. See Figure 3-5 on page 79.

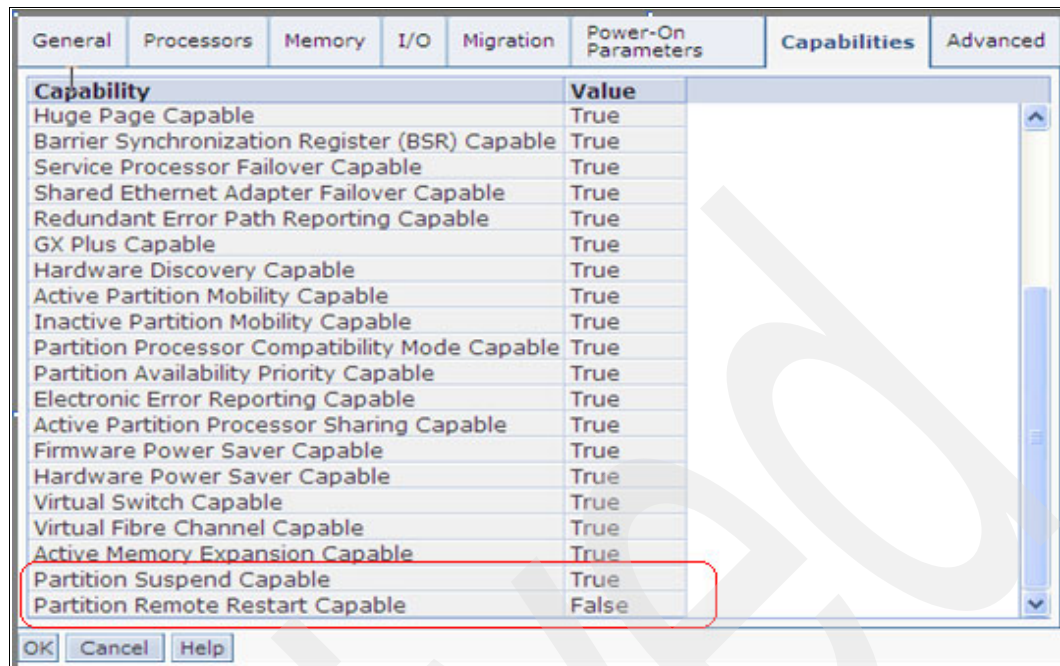


Figure 3-5 System capabilities for suspend resume

2. Make sure that the LPAR is suspend-capable. Select **LPAR** → **Properties**. See Figure 3-6.

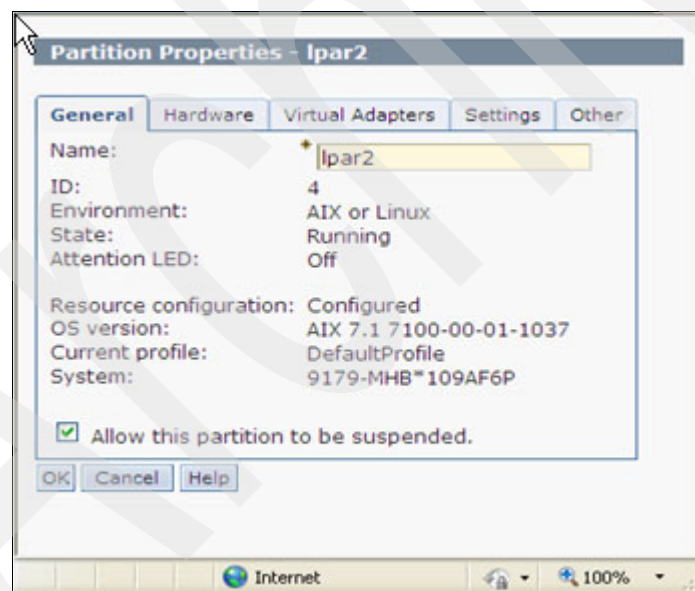


Figure 3-6 Suspend and resume option for the LPAR

Suspend and resume is a dynamic feature, and it can be enabled or disabled on a running partition. After you confirm the system capability and LPAR ability to be suspended, you need to configure the storage pools on one or more virtual I/O servers. *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590, explains this concept in detail with examples.

3.4 IBM SystemMirror PowerHA

PowerHA provides a highly available cluster environment that enhances business continuity. PowerHA provides an *infrastructure* that enables mission-critical resources to remain available in case of failures. It also allows the quick recovery of an application to another server if one or more servers fail.

Several factors make PowerHA a beneficial utility:

- ▶ PowerHA is a mature cluster technology (was the High Availability Cluster Multi-Processing (HACMP™) technology since 1992)
- ▶ It can be deployed on standard hardware. If the hardware supports the required version of AIX, IBM i, or Linux, PowerHA can be installed.
- ▶ PowerHA allows you to change many configurations without having to shut down the cluster, thus eliminating planned downtime. This concept is known as Dynamic Automatic Reconfiguration (DARE). PowerHA is flexible.
- ▶ PowerHA complements virtualization technologies, including Capacity on Demand (CoD), WPAR, and LPM. These technologies are introduced and documented throughout this publication.
- ▶ PowerHA monitors the resources within its control. In the failure of any resource, PowerHA takes an appropriate action to either restart the resource or move the resource to another node in the cluster. Another node is typically on a separate server. Node failure, application failure, and component failures are monitored by PowerHA.

3.4.1 Comparing PowerHA with other high-availability solutions

We provide a technical comparison between PowerHA and other high-availability solutions.

PowerHA with fault-tolerant systems

Fault-tolerant systems are costly and use specialized hardware and software. PowerHA uses standard hardware and software. PowerHA can be deployed on a system that possesses RAS features, such as the Power 795 and 780.

Figure 3-7 shows the cost and benefit of the available technologies.

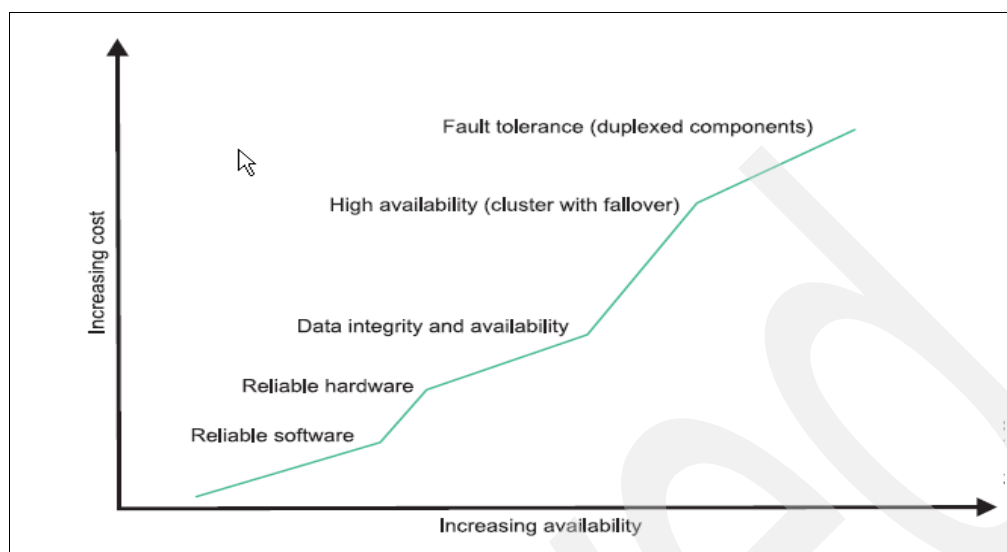


Figure 3-7 Cost/benefit graph of high-availability technologies

PowerHA is also compared with software-based clusters

It is not our intention to compare PowerHA with software clusters. PowerHA provides a platform for both software-capable clustering and applications that do not have clustering capability to be highly available. Most cluster software takes advantage of the PowerHA infrastructure.

For example, if you have an application, which was developed in-house and does not have clustering capabilities, and the application is critical, you can identify the components of the application. Create a start script, stop script, and a script to monitor the state of the application as you normally do on a single system. Integrate the application into PowerHA, and you have an in-house-developed high-availability solution.

Choosing the type of cluster affects how you will set up the resource groups.

Planning for PowerHA

Planning and design form the basis of reliable environments. Chapter 4, “Planning for virtualization and RAS in POWER7 high-end servers” on page 99 discusses how to eliminate single points of failures (SPOFs). The chapter covers most of the planning requirements for PowerHA.

Consider these factors:

Nodes	Names and number of nodes. Refer to the Cluster Aware AIX (CAA) documentation regarding node names. Note that the naming convention affects the names that are used when creating the PowerHA cluster.
Networks	Multicast IP address, number of interfaces, supported/unsupported networks, and IPv6.
SAN	PowerHA recommends multipathing.
CAA	Cluster repository for the CAA. We discuss CAA in 3.6, “Cluster Aware AIX (CAA)” on page 87. Notice that the cluster repository does not support Logical Volume Manager (LVM) Mirroring.

LVM	Mirroring for volume groups.
Resource group	File system, logical volumes, volume groups, and network and application servers under the control of PowerHA.
Cluster type	Mutual failover, standby, one-sided takeover, mutual takeover, and multi-tiered applications (both nodes active).

Resource groups contain the resources that PowerHA keeps highly available.

PowerHA management and configuration

You can use either the System Management Interface Tool (Smitty), command line, or a plug-in to IBM Systems Director. From our experiences, we recommend IBM Systems Director and Smitty. PowerHA 7.1 has a method that disables certain smitty commands, such as **chfs**, **chlv**, and **chgrp**. You can overwrite this feature, but we do not advise that you do.

The PowerHA 7.1 SMIT menu differs slightly from PowerHA 6.1 and prior versions. A list of these changes is included in the *IBM PowerHA SystemMirror 7.1 for AIX*, SG24-7845, publication.

3.4.2 PowerHA 7.1, AIX, and PowerVM

PowerHA 7.1 introduces many improvements from PowerHA 6.1 and prior versions. Most of the improvements come from changes and improvements to the base AIX operating system. AIX Version 7 is enhanced to be able to maintain a few nodes using CAA capabilities. CAA is introduced in 3.6, “Cluster Aware AIX (CAA)” on page 87. Most of the PowerHA monitoring infrastructure is part of the base operating system. For earlier versions of AIX to take advantage of PowerHA 7.1, CAA is included in AIX 6.1 TL6. CAA requires Reliable Scalable Cluster Technology (RSCT) 3.1.

PowerHA classically has subnet requirements and needs a number of interfaces, because PowerHA monitors failures and moves IP addresses to a surviving interface prior to moving to the next node. AIX provides methods of monitoring the network adapter using EtherChannel, which is implemented either as link aggregation or a network interface backup. The use of EtherChannel eliminates the need for configuring multiple interfaces because this requirement is taken care of by implementing EtherChannel.

The virtual I/O server provides a method of creating a Shared Ethernet Adapter Failover (SEA Failover), which allows the virtual I/O server to provide required redundancy. An example of an SEA configuration is shown in 6.5.3, “NIB and SEA failover configuration” on page 221. An SEA configuration also removes the need to create multiple interfaces in the PowerHA configuration.

Refer to the *IBM PowerHA SystemMirror 7.1 for AIX*, SG24-7845, and *PowerHA for AIX Cookbook*, SG24-7739, publications to get started with PowerHA or to migrate to PowerHA 7.1.

3.5 IBM Power Flex

Power Flex was introduced with POWER7 and is a multi-system Power 795 infrastructure offering from IBM. It provides a highly available and flexible IT environment to support large-scale server consolidation and an enterprise’s most demanding business resiliency objectives. Power Flex is designed to enable you to more easily use your purchased

processor and memory activations across a pool of 2 - 4 Power 795 systems. This flexibility leads to the increased utilization of the resources and to enhanced application availability.

3.5.1 Power Flex Overview: RPQ 8A1830

Power Flex has these highlights:

- ▶ Supports multi-system infrastructure support for active-active availability
- ▶ Allocates and rebalances processor and memory resources
- ▶ Uses LPM for flexible workload movement
- ▶ Delivers seamless growth with Capacity on Demand (CoD)
- ▶ Includes On/Off processor days for extra capacity

Power 795 servers in a Power Flex environment are allowed to share large portions of their virtual processor and memory resources to provide capacity where it is most needed and to best support application availability during occasional planned system maintenance activity.

Power Flex consists of two to four Power 795 systems, each with four or more 4.0 GHz or 4.25 GHz processor books, and 50% or more permanent processor and memory activations to support its applications. Capacity higher than 25% on these systems can be used as a Flex Capacity Upgrade on Demand resource and rebalanced to and from another Power 795 system in the same Power Flex pool of systems, up to twelve times per year.

Power Flex enablement has these prerequisites:

- ▶ All servers within a Power Flex capacity pool must have equivalent IBM hardware maintenance status.
- ▶ Each feature code 4700 processor book installed must include enough feature code 4713s (core activation feature) so that a minimum of 50% of the total number of processor cores that are configured on the server are activated. For example, you can have 16 x feature code 4713 (1-core feature) for each feature code 4700 processor book, or 8 x feature code 4713 for each feature code 4700, if TurboCore mode (feature code 9982) is specified.
- ▶ The total number of RPQ 8A1830 that is installed on the server must equal the total number of feature code 4700 processor books installed.
- ▶ The Power Flex attachment and supplement must be signed and returned to enable the 960 On/Off processor credit days per book and to allow the rebalancing of resources on the system to which RPQ 8A1830 applies.

A summary of the Power Flex details are shown in Figure 3-8 on page 84.

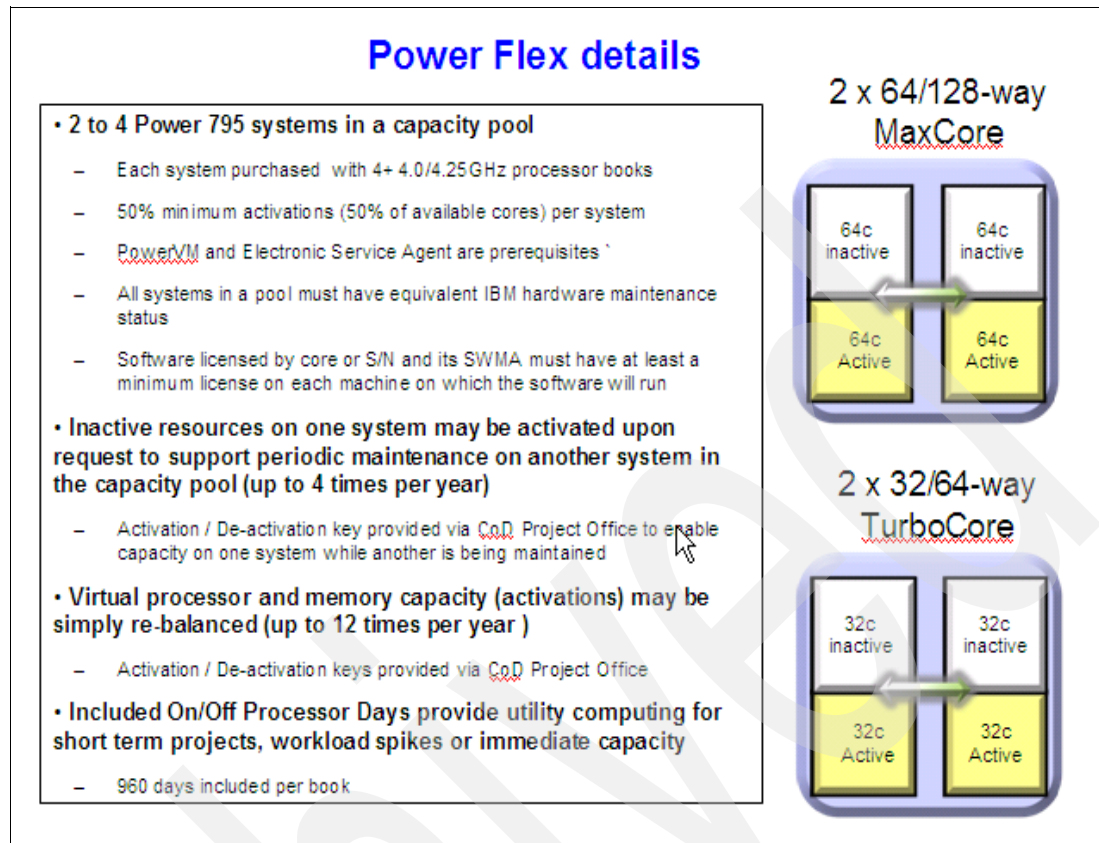


Figure 3-8 Power Flex details

3.5.2 Power Flex usage options

PowerFlex options are designed to allow your multi-system virtualization infrastructure to provide a highly available and flexible IBM Power Systems IT environment.

Maintaining application availability

Flex Capacity Upgrade on Demand, working in conjunction with PowerVM Live Partition Mobility, can help clients maintain application availability more affordably during planned maintenance activities on a Power Flex system. Up to four times per year, a client can request to temporarily activate inactive resources on one Power Flex system to support the planned maintenance activities on another Power Flex system. While virtual processor and memory resources are not being used on a system that is being maintained, they can be used on another system for productive use without having to first be deactivated on the system being maintained.

Any resources activated as part of an advanced planning event are to be deactivated within seven days.

Important: Advanced planning event capacity (key) requests to the Power CoD project office require a minimum of two business days to ensure the receipt of the activation and deactivation codes for a system, prior to commencing planned maintenance activities.

Workload rebalancing with Power Flex Capacity Upgrade on Demand

Flex Capacity Upgrade on Demand (CUoD) processor and memory activations on a Power Flex system can be temporarily rebalanced to be allowed to execute on another installed Power Flex system within the same enterprise and country. Each Power 795's Flex CUoD resources are the processor and memory activations above 25% of its total capacity. These resources can be rebalanced up to 12 times per year to execute on another Power Flex system in the same pool to support changing capacity requirements.

Unique to a Power Flex environment, rebalancing capacity can be activated on a target system prior to the capacity being deactivated on its donor system to better facilitate any transition of applications from one system to another system. While resources on one system are activated, corresponding resources are to be deactivated in the donating system within seven days.

Rebalanced processor activation resources are not permanently moved to another system. They are temporarily allowed to execute on systems within a Power Flex capacity pool, yet they are retained on an initial system for inventory and maintenance purposes. Power Flex merely allows clients to make use of these CUoD resources on more than a single system. Any rebalanced activations are to be reconciled with inventory records in the event that the system is upgraded or sold and must be returned to the original system upon any lease termination. See Figure 3-9 on page 86 for an example of Power Flex in action.

Important: Requests for Flex Capacity Upgrade on Demand activation/deactivation keys are initiated via e-mail to the Power CoD project office (pcod@us.ibm.com) at least two business days in advance of rebalancing activity.

Utility computing via On/Off processor days

Each Power Flex system ships with a quantity of included On/Off Capacity on Demand processor days (approximately 60 days of the inactive resources on each purchased Power Flex 64/128-core system, or 960 days per 32-core processor book). These On/Off processor days are enabled via the normal Capacity on Demand resource enablement process and contracts. The On/Off days are credited to the client's account upon completion of initial CoD and Power Flex contracts. They can be used at a client's discretion to provide utility computing for short-term projects, workload spikes, or in the event of an immediate maintenance activity where an advanced planning event or rebalancing request has not been requested.

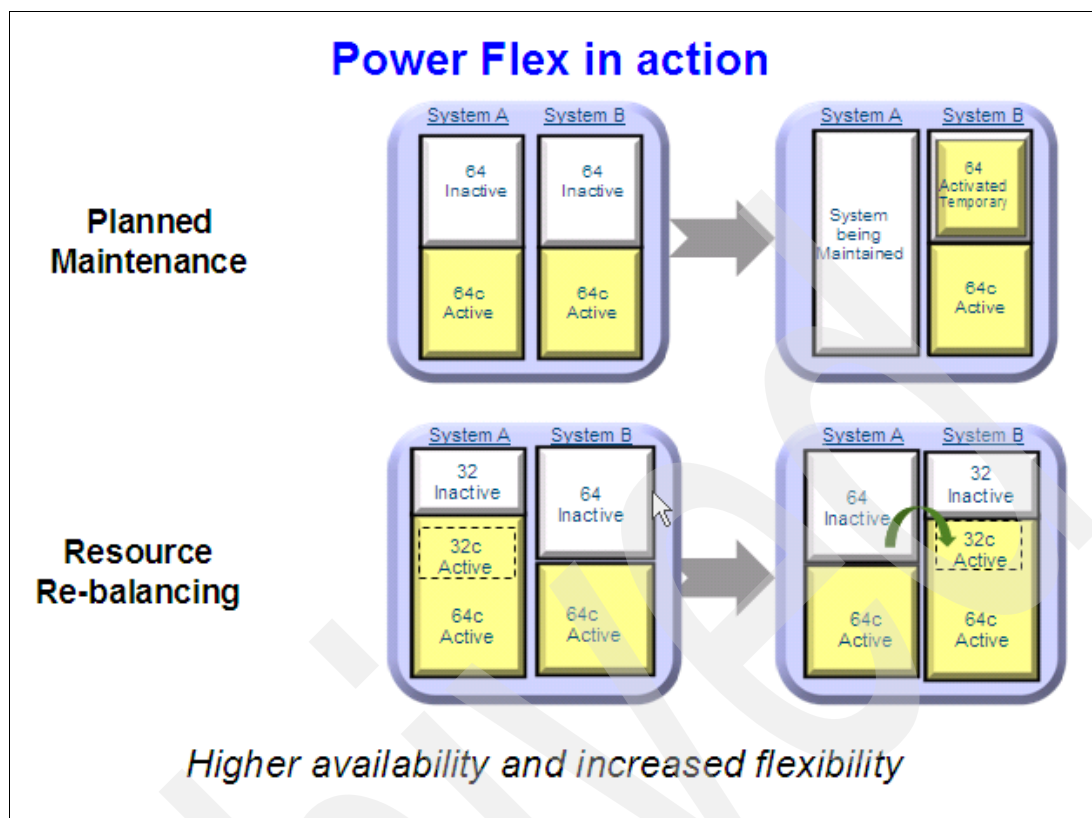


Figure 3-9 Depicts Power Flex in action

IBM Capacity on Demand offerings

The On/Off Capacity on Demand processor days that are included with Power Flex are separate from the IBM Capacity on Demand (CoD) offerings. With the IBM CoD offerings, you can dynamically activate one or more resources on your POWER7 server as your business activity peaks dictate. You can activate inactive processor cores or memory units that are already installed on your server on a temporary and permanent basis. Inactive processor cores and inactive memory units are resources that are installed as part of your server, but they are not available for use until you activate them.

Table 3-4 provides a brief description of each CoD offering. For additional information, review the *Power Systems Capacity on Demand* document:

<https://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/topic/p7ha2/p7ha2.pdf>

Or, consult your IBM sales representative or IBM Business Partner.

Table 3-4 CoD offerings from IBM

CoD offering	Description
Capacity Upgrade on Demand	Permanently activate inactive processor cores and memory units by purchasing an activation feature and entering the provided activation code.
Trial Capacity on Demand	Evaluate the use of inactive processor cores, memory, or both, at no charge using Trial CoD. After it is started, the trial period is available for 30 power-on days.

CoD offering	Description
On/Off Capacity on Demand	Activate processor cores or memory units for a number of days by using the HMC to activate resources on a temporary basis.
Utility Capacity on Demand	Used when you have unpredictable, short workload spikes. Automatically provides additional processor capacity on a temporary basis within the shared processor pool. Use is measured in processor minute increments and is reported at the Utility CoD website.
Capacity BackUp	Used to provide an off-site, disaster recovery server using On/Off CoD capabilities. The offering has a minimum set of active processor cores that can be used for any workload and a large number of inactive processor cores that can be activated using On/Off CoD in the event of a disaster. A specified number of no-charge On/Off CoD processor days are provided with Capacity BackUp.

3.6 Cluster Aware AIX (CAA)

Cluster Aware AIX (CAA) services and tools are among the most important new features in AIX 7.1. AIX 7.1 is the first AIX release to provide for built-in clustering. The latest editions of PowerHA SystemMirror and PowerVM are designed to exploit the CAA cluster infrastructure to facilitate high availability and advanced virtualization capabilities. Administrators are now able to create a cluster of AIX systems using features of the AIX 7 kernel. IBM introduced the “built-in” clustering capabilities to AIX OS to simplify the configuration and management of highly available clusters and high availability. Cluster Aware AIX functionality is primarily intended to provide a reliable, scalable clustering infrastructure for products, such as PowerHA SystemMirror and PowerVM. The new AIX clustering capabilities are designed to offer these benefits:

- ▶ Significantly simplify cluster construction, configuration, and maintenance
- ▶ Improve availability by reducing the required time to discover failures
- ▶ Offer capabilities, such as common device naming for shared devices, to help optimize administration
- ▶ Provide built-in event management and monitoring
- ▶ Offer a foundation for future AIX capabilities and the next generation of PowerVM and PowerHA SystemMirror

AIX 7 runs on our latest generation of Power processor POWER7 systems, as well as systems based on POWER4, POWER5, and POWER6. Most of the new features of AIX 7 are available on earlier Power processor-based platforms, but the most capability is delivered on systems built with the POWER6 and POWER7 processors.

CAA is not designed as a high-availability replacement for PowerHA SystemMirror, but it does change the way in which AIX integrates with cluster solutions, such as PowerHA (HACMP). IBM’s mature RSCT technology is still an important element of AIX and PowerHA configurations. IBM PowerHA now uses components of CAA, instead of RSCT, to handle the cluster topology, including heartbeats, configuration information, and live notification events. PowerHA still communicates with RSCT Group Services (grpsvcs replaced by cthags), but PowerHA has replaced the topsvcs (topology services) function with the new CAA function. CAA reports the status of the topology to cthags by using Autonomic Health Advisory File System API (AHAFS) events, which are fed up to cthagsrhosts. Refer to Figure 3-10.

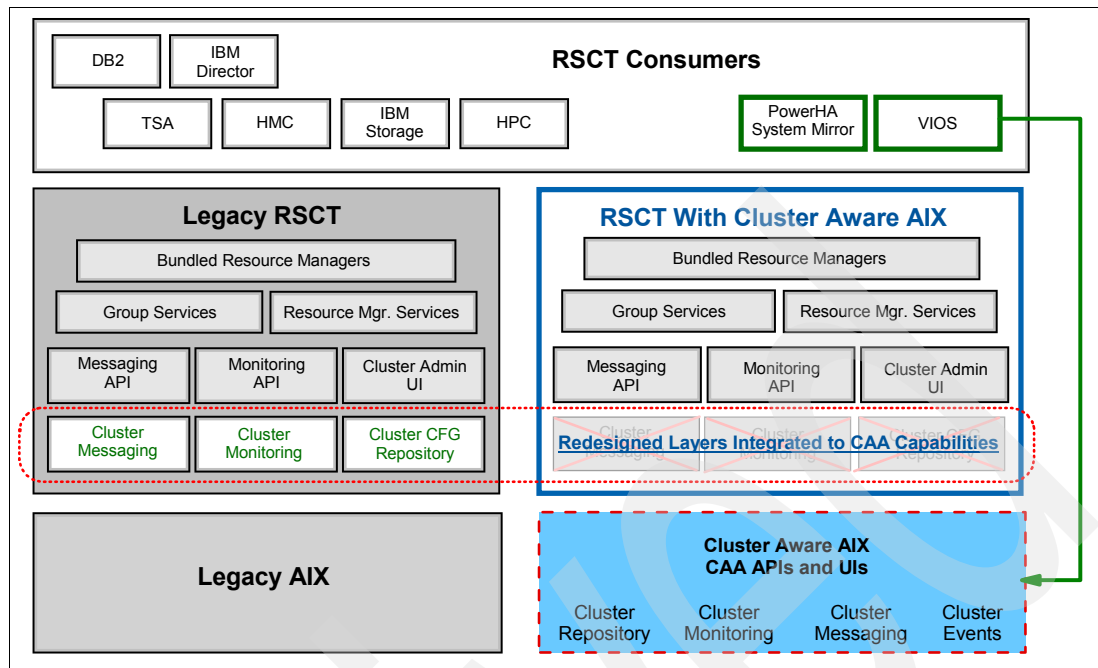


Figure 3-10 Cluster aware AIX exploiters

IBM Reliable Scalable Cluster Technology (RSCT) is a set of software components that together provide a comprehensive clustering environment for AIX and Linux. RSCT is the infrastructure that is used by a variety of IBM products to provide clusters with improved system availability, scalability, and ease of use. RSCT includes daemons, which are responsible for monitoring the state of the cluster (for example, a node network adapter, network interface card (NIC), or network crash) and coordinates the response to these events. PowerHA is an RSCT-aware client. RSCT is distributed with AIX. The following list includes the major RSCT components:

- ▶ **Resource Monitoring and Control (RMC) subsystem**, which is the scalable, reliable backbone of RSCT. It runs on a single machine or on each node (operating system image) of a cluster and provides a common abstraction for the resources of the individual system or the cluster of nodes. You can use RMC for single system monitoring, or for monitoring nodes in a cluster. In a cluster, however, RMC provides global access to subsystems and resources throughout the cluster, thus providing a single monitoring/management infrastructure for clusters. It is also used for dynamic LPAR, sfp, invscout, and so on.
- ▶ **RSCT core resource managers.** A *resource manager* is a software layer between a resource (a hardware or software entity that provides services to another component) and RMC. A resource manager maps programmatic abstractions in RMC into the actual calls and commands of a resource.
- ▶ **RSCT cluster security services** provides the security infrastructure that enables RSCT components to authenticate the identities of other parties.
- ▶ **The Topology Services subsystem** provides node/network failure detection in certain cluster configurations.
- ▶ **The Group Services subsystem** provides cross-node/process coordination in certain cluster configurations.

RSCT Version 3.1 is the first version that supports Cluster Aware AIX (CAA). RSCT 3.1 can operate without CAA in a “non-CAA” mode.

You use the non-CAA mode if you use one of the following products:

- ▶ PowerHA versions before PowerHA 7.1
- ▶ A mixed cluster with PowerHA 7.1 and prior PowerHA versions
- ▶ Existing RSCT Peer Domains (RPD) that were created before RSCT 3.1
- ▶ A new RPD cluster, when you specify during creation that the system must not use or create a CAA cluster

Figure 3-11 shows both modes in which RSCT 3.1 can be used (with or without CAA). On the left diagram, you can see how the non-CAA mode works, which is equal to the older RSCT versions. The diagram on the right side shows the CAA-based mode. The difference between these modes is that Topology Services has been replaced with CAA.

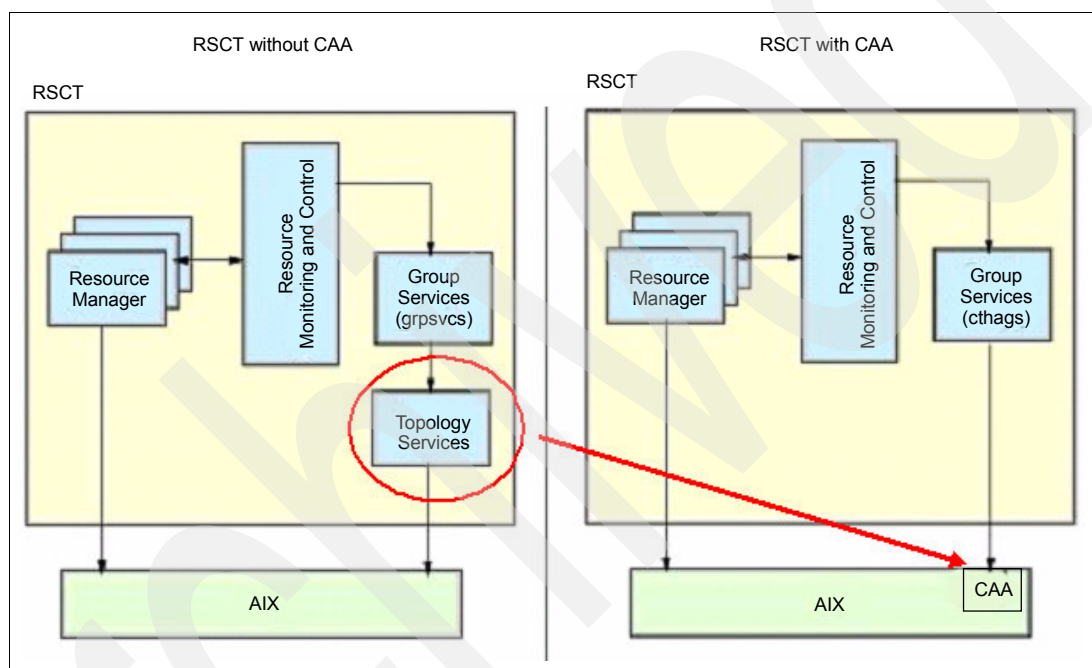


Figure 3-11 RSCT 3.1 modes

Important: RSCT 3.1 is available for both AIX 6.1 and AIX 7.1. To use CAA, for RSCT 3.1 on AIX 6.1, you must have TL 6 or later installed.

The use of CAA on AIX 6.1 TL 6 is enabled only for PowerHA 7.1 and not for earlier versions.

3.6.1 Cluster Aware AIX Services

Cluster Aware AIX (CAA) is set of services and tools embedded in AIX to help you manage a cluster of AIX nodes and help you run cluster software on AIX. CAA services provide these functions:

- ▶ CAA configuration and database
 - Cluster verification that is performed when the cluster is defined or modified.

- ▶ CAA communication

Communication between nodes within the cluster is achieved using multicasting over the IP-based network and also using storage interface communication through FC and serial-attached SCSI (SAS) adapters.
- ▶ CAA monitoring (nodes, networks, and storage):
 - All monitors are implemented at low levels of the AIX kernel and are largely insensitive to system load.
 - All communication interfaces are monitored: network and storage.
- ▶ CAA device-naming services:
 - When a cluster is defined or modified, AIX interfaces automatically create a consistent shared device view.
 - When managed by Cluster Aware AIX, device files that are associated with the disks shared across the nodes in the cluster have a common name across the nodes in the cluster that have access to the disks.

Global device names, such as `cldisk1`, refer to the same physical disk from any cluster node.
- ▶ CAA cluster-wide event management:
 - AIX event infrastructure allows event propagation across the cluster.
 - Applications can monitor events from any cluster node.
- ▶ CAA cluster-wide command distribution:
 - Many of the security and storage-related AIX commands are enhanced to support the operation across the cluster.
 - The `c1cmd` command provides a facility to distribute a command to a set of nodes that are cluster members.

3.6.2 Cluster Aware AIX event infrastructure

AIX event infrastructure for AIX and AIX Clusters, which was introduced in AIX 6.1, provided an event monitoring framework for monitoring predefined and user-defined events. Enhancements in AIX 7.1 include support for cluster-wide event notifications for certain events (for example, network and disk errors) with continuous monitoring and additional producers. An *event* is defined as any change of a state or a value that can be detected by the kernel or a kernel extension at the time that the change occurs. The events that can be monitored are represented as files in a pseudo file system named the Autonomic Health Advisor FileSystem (AHAFS). Cluster Aware AIX generates granular storage and network events that are used by PowerHA to provide for better decision making for high-availability management.

Four components make up the AIX event infrastructure (refer to Figure 3-12 on page 91):

- ▶ The kernel extension implementing the pseudo file system
- ▶ The event consumers that consume the events
- ▶ The event producers that produce events
- ▶ The kernel component that serves as an interface between the kernel extension and the event producers

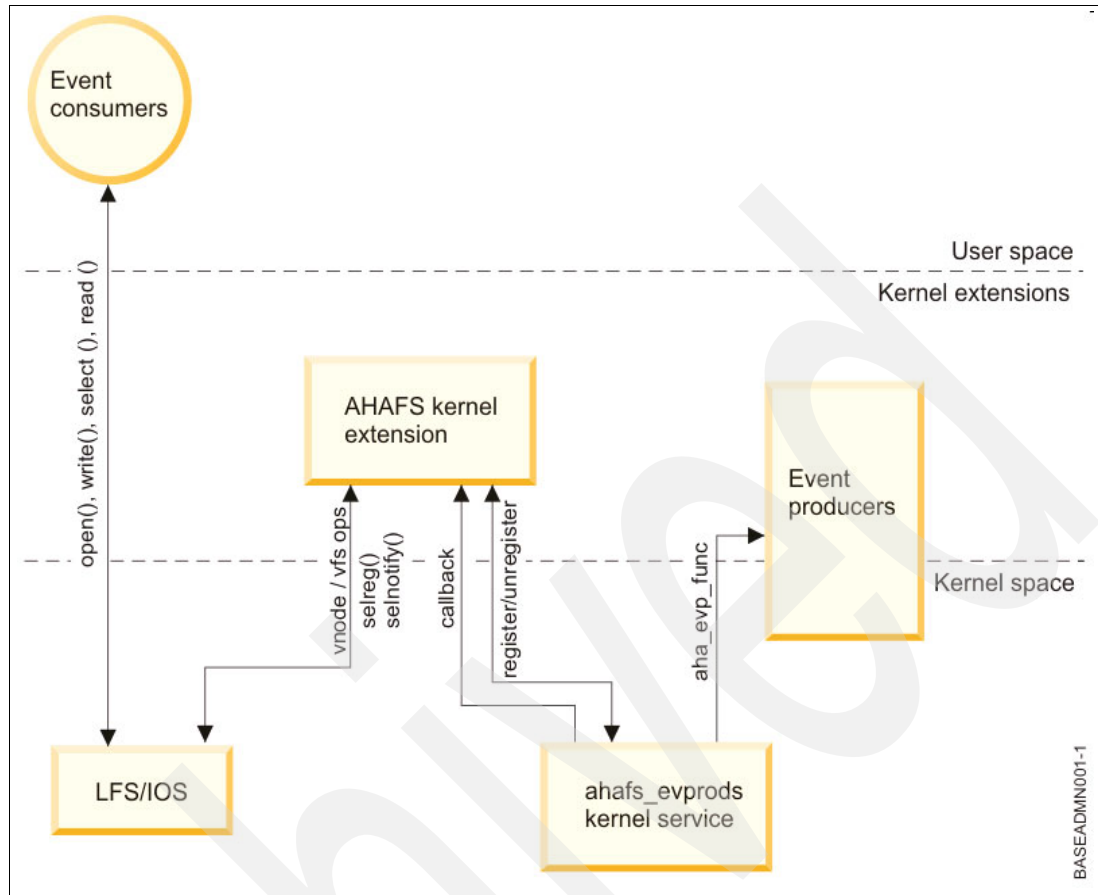


Figure 3-12 AIX event infrastructure components

The Cluster Aware AIX event infrastructure is designed to provide these functions:

- ▶ Event framework for monitoring events efficiently and without the need for polling
 - System events are defined as a change in state or value, which can be detected in the AIX kernel or kernel extensions as it occurs.
- ▶ A pseudo-file system named Autonomic Health Advisor FileSystem (AHAFS):
 - The AHAFS kernel extension was first introduced in AIX 6.1 TL 04 (October 2009) fileset `bos.ahafs`.
 - Further extensions were included in AIX 6.1 TL 06 and AIX 7.
 - Loadable kernel extension and root mount of the AHAFS file system, for example: `mount ahafs /aha /aha`
 - In-memory only file system allocated from a pinned heap.
 - Monitoring applications can use standard file system interfaces (for example, `open`, `write`, `select`, `read`, `close`, and so on) to perform monitoring instead of having to use a special set of APIs.
- ▶ Monitor the health, security, and RAS of AIX:
 - Events triggered by an event producer must originate from either the kernel or in a kernel extension.
 - Event producers can dynamically register and unregister themselves with the AHAFS framework.

- The authorization to monitor specific events is determined by each event producer.
- Detailed information about an event (stack trace, user, and process information) is provided.
- Control is handed to the AIX event infrastructure at the exact time the event occurs.

The *IBM AIX Version 6.1 Differences Guide*, SG24-7559 contains detailed information about Cluster Aware AIX functionality.

3.7 Electronic services and electronic service agent

Electronic services is an IBM support approach, which consists of the Electronic Service Agent™ (ESA) and the electronic services website, as shown in Figure 3-13. The ESA automatically monitors and collects hardware problem information and sends this information to IBM support. It can also collect hardware, software, system configuration, and performance management information, which might help the IBM support team to assist in diagnosing problems. IBM electronic services reaches across all IBM systems in all countries and regions where IBM does business. Electronic services can provide the electronic support relationship for a single machine environment or a multinational complex of many servers.

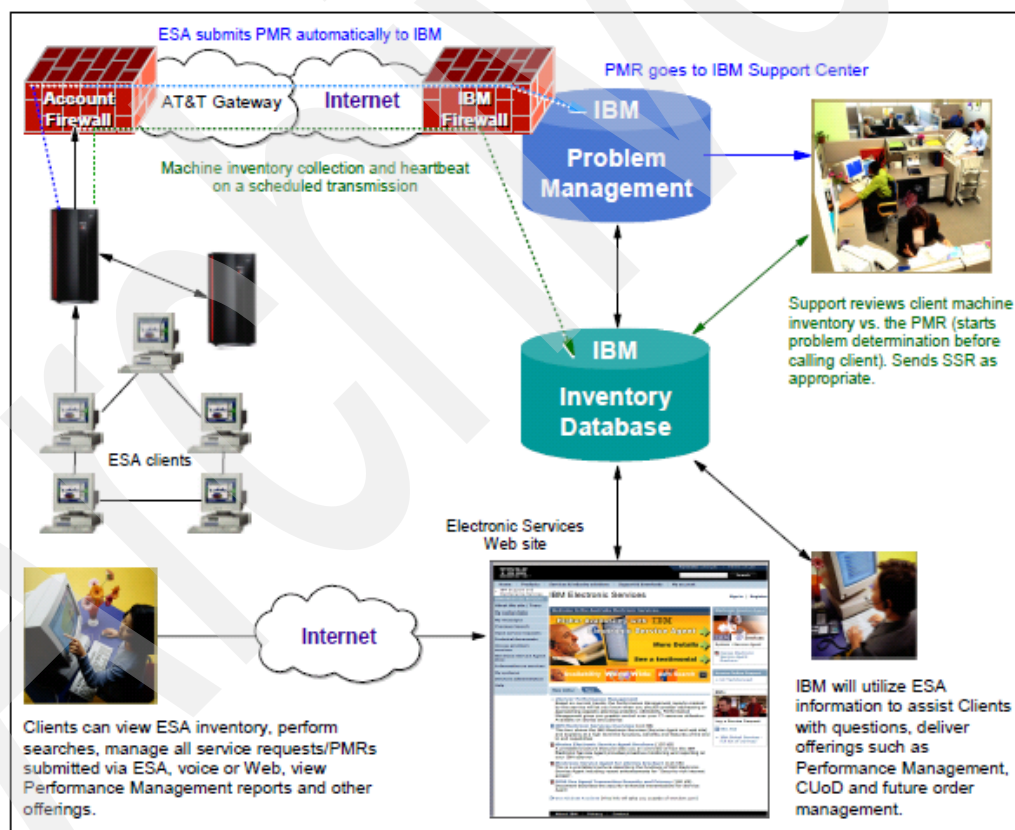


Figure 3-13 Electronic services overview

Electronic service agent (ESA) is a no-charge software tool that resides on your system to continuously monitor events and periodically send service information to IBM support on a user-definable time table. This tool tracks and captures service information, hardware error logs, and performance information. It automatically reports hardware error information to IBM

support as long as the system is under an IBM maintenance agreement or within the IBM warranty period. Service information reporting and performance information reporting do not require an IBM maintenance agreement and do not need to be within the IBM warranty period to be reported. Information that is collected by the ESA application is available to IBM service support representatives to help them diagnose problems.

Previous ESA products were unique to the platform or operating system on which they were designed to run. Because the ESA products were unique, each ESA product offered its own interface to manage and control the ESA and its functions. Because networks can have separate platforms with separate operating systems, administrators had to learn a separate interface for each separate platform and operating system in their network. Multiple interfaces added to the burden of administering the network and reporting problem service information to IBM support.

ESA now installs on platforms that are running separate operating systems. It offers a consistent interface to ESA functions, reducing the burden of administering a network with various platforms and operating systems. ESA is operating system specific. Each operating system needs its own compatible version of ESA. To access ESA user guides, go to the electronic services website and select **Electronic Service Agent** on the left side of the navigation page. In the contents pane, select **Reference Guides** → Select a *platform*. Select **Operating System or Software**.

On your POWER7 platform, you can have one or more operating systems. No matter how many partitions are configured, or which operating systems are running, the IBM ESA must be installed and activated on each partition, operating system, and HMC or SDMC.

ESA: For HMC-controlled or SDMC-controlled environments, ESA must be activated on the HMC or SDMC for hardware error reporting.

For system inventory reporting, Resource Monitoring and Control (RMC) must be configured in the partition. Additional activation of ESA on the partitions sends back OS-specific (AIX or IBM i) and software inventory data.

You configure ESA on AIX 5.3, 6.1, and 7.1 from the command line by entering `smit esa_main` and then selecting **Configure Electronic Service Agent**.

Important: It is important to activate ESA on every platform, partition, and Hardware Management Console (HMC) or Systems Director Management Console (SDMC) in your network to get the maximum coverage and utilization of the ESA capabilities.

3.7.1 Benefits of ESA for your IT organization and your Power systems

ESA offers the following electronic services benefits for both your IT organization and systems:

- ▶ No additional charge for systems under warranty or maintenance agreements
- ▶ Helps achieve higher availability and shorter downtime
- ▶ Automatically contacts IBM support
- ▶ Immediately uploads error logs
- ▶ Faster diagnosis and time to repair
- ▶ Automatically gathers and reports required system information by ESA, thus reducing data entry errors or the risk of misreading system information

- ▶ Less personnel time providing and gathering information and reporting problems
- ▶ Routes calls to the correct resource the first time with the required information to provide an end-to-end, automated, closed loop support process
- ▶ Access to web delivered services, such as viewing ESA information and tracking and managing reported problems
- ▶ Standard install for Power 780 and Power 795 systems

ESA: ESA enablement is a prerequisite for POWER7 systems performing CEC hot node add, hot node upgrade (memory), hot node repair, or hot GX adapter repair. It is also required for Power Flex enablement. ESA-enabled systems show improved concurrent operations results.

3.7.2 Secure connection methods

IBM knows that your security and information privacy are extremely important. *No client business data is ever transmitted to IBM through ESA.* We provide several secure connectivity methods from which you can choose:

- ▶ Internet
- ▶ VPN
- ▶ Dial-up

We provide both proxy and authenticating firewall support when using ESA: security protocols, including https (SSL and Transport Layer Security (TLS)) and 128-bit encryption that uses encryption keys, certificates, and tokens. ESA and Call-Home follow the industry standards for protecting data during network transport by using the TLS protocol. ESA and Call-Home also protect your Call-Home and IBM support accounts by generating unique passwords for these accounts. Call-Home uses protected channels, for example, TLS and VPN, to transfer data from the HMC to IBM support. The channels provide confidentiality and integrity protection for the data that is sent between the two entities. Figure 3-14 on page 95 shows the various connectivity methods that are available for ESA.

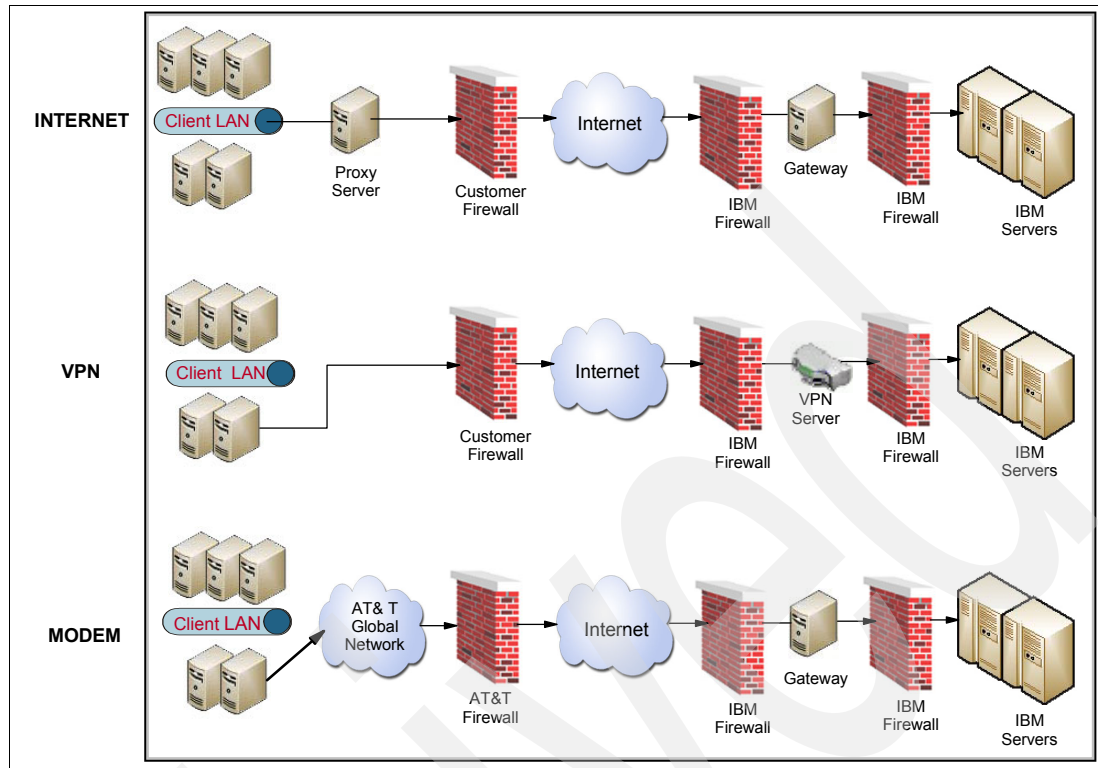


Figure 3-14 Call home information paths

ESA has no inbound capability. It cannot accept incoming connection attempts. ESA initiates a connection with IBM support, and then IBM support replies. IBM support never initiates a connection to the ESA.

IBM provides secure storage for all data that is transmitted using ESA. Your system information is stored in a secure database behind two firewalls and is accessible by you with a protected password. The database is accessible only by authorized IBM support representatives. All access to the database is tracked and logged, and it is certified by the IBM security policy.

The IBM Power 780 or 795 system has an attached HMC or SDMC, so there are additional considerations when using ESA. For HMC/SDMC managed environments, ESA must be activated on the HMC/SDMC for hardware error reporting.

The HMC and the SDMC include their own versions of ESA. ESA on the HMC and SDMC monitors the system and AIX, IBM i, and Linux partitions for errors, and ESA reports these errors to IBM. It also collects and reports hardware service information and performance management information to IBM support. ESA on a partition does not collect hardware information; it collects other service information, such as software information.

To access the ESA user guide for HMC, go to the electronic services website and select **Electronic Service Agent** from the left navigation page. In the contents pane, select **Reference Guides** → a *platform* → **Operating System or Software**.

To activate the ESA for problem hardware reporting from your HMC, as shown in Figure 3-15 on page 96, perform the following steps:

1. Log in to your HMC interface.
2. Select **Guided Setup Wizard**.

3. Select **Yes** to launch the Call-Home Setup Wizard, as shown in Figure 3-15.

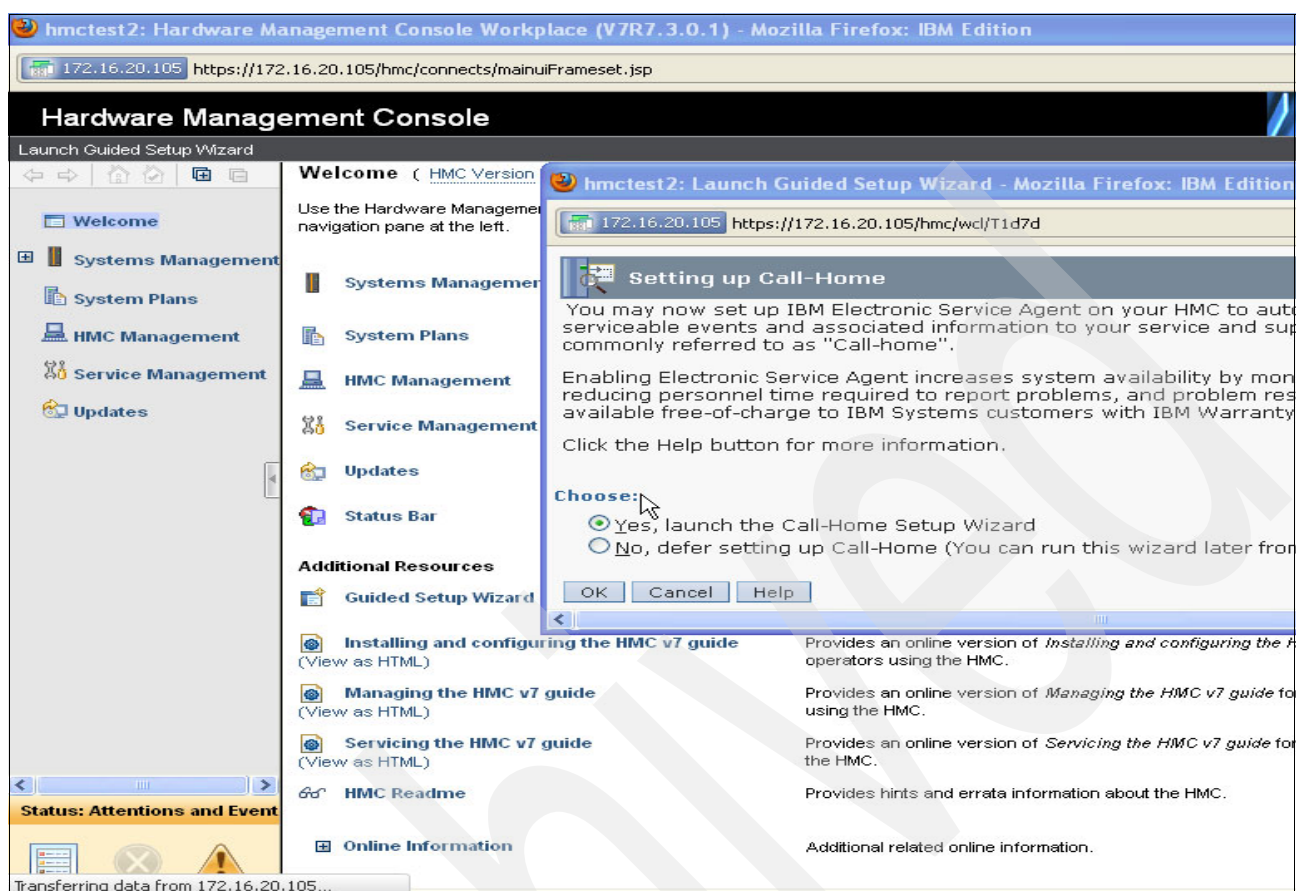


Figure 3-15 Launching the Call-Home Setup Wizard

Worksheet: A preinstallation configuration worksheet is available to assist you with the identification of prerequisite information:

<https://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7ecq/arecqconfigurethehmc.htm>

Figure 3-16 on page 97 shows the welcome page for configuring ESA on your HMC.

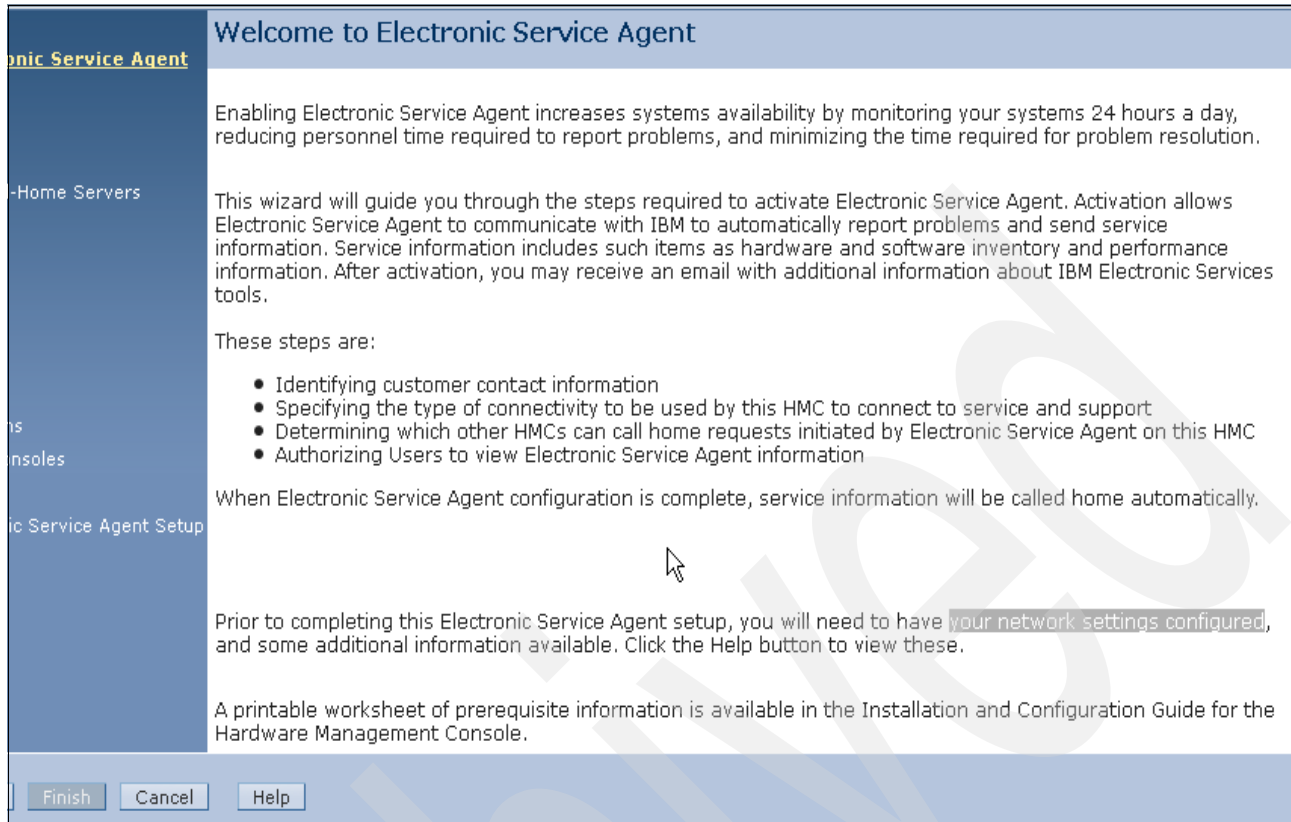


Figure 3-16 Electronic Service Agent welcome pane

You can also review and control whether Call-Home requests can be created for the HMC or a managed system by choosing **Service Management** → **Connectivity** → **Enable Electronic Service Agent**.

You can also configure ESA from your SDMC. *IBM Electronic Services Support using Automation and Web Tools*, SG24-6323, is an excellent source of information about using IBM electronic services.

Archived

Planning for virtualization and RAS in POWER7 high-end servers

This chapter provides information about the suggested planning to help you enable your Power System servers (Power 780 and Power 795) to exploit the RAS and virtualization features.

In this chapter, we describe the following topics:

- ▶ Physical environment planning
- ▶ Hardware planning
- ▶ CEC Hot Add Repair Maintenance (CHARM)
- ▶ Software planning
- ▶ HMC server and partition support limits
- ▶ Migrating from POWER6 to POWER7
- ▶ Technical and Delivery Assessment (TDA)
- ▶ System Planning Tool (SPT)
- ▶ General planning guidelines for highly available systems

4.1 Physical environment planning

In this section, we introduce points to consider before installing the IBM Power Systems hardware. We introduce important considerations when deploying a reliability, availability, and serviceability (RAS)-capable server, such as the Power 795 and Power 780. It is not our intention in this chapter to provide a comprehensive environmental planning guide. We exclude the site planning and assume that it is already completed. Concepts that are not in the scope of this book include power and cooling, raised floors, air distribution, shock, and vibrations.

Insufficient planning can affect the effectiveness of the RAS features of the intended hardware. You must include the following items in the planning of the environment where the server will be installed:

- ▶ Site planning
- ▶ Power distribution units (PDU)
- ▶ Networks and switches
- ▶ SANs and SAN switches

4.1.1 Site planning

A site inspection must be performed before a server is installed. After the site inspection, a floor plan must be updated with a clearly marked location for the server and its expansion drawers. If the expansion is not considered in the beginning, you might need to move the machine at a later stage. This move can cause downtime to users unless IBM Live Partition Mobility (LPM) is used. See 3.1, “Live Partition Mobility (LPM)” on page 64. For example, if you share a rack between a 780 and other systems, you must plan for up to 16Us in case you need to expand, even if you only purchased two system enclosures.

When considering a site, include the following factors:

- ▶ Floor construction.
- ▶ Access routes and entrances.
- ▶ Floor space, which also must include working space for future maintenance of the server.
- ▶ Distances from the network and SAN switching devices.
- ▶ In case continuous operation is required, you must also have another separate site available if you totally lose this site. This second site must be part of business continuity and disaster recovery planning.
- ▶ Other environmental considerations.

For a comprehensive list of all site planning considerations, consult the hardware pages on the information center:

<https://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp>

4.1.2 Power and power distribution units (PDUs)

Power and air conditioning form part of system planning. Refer to the handbook for your system to ensure that your environment meets the energy requirements, as well as the air conditioning requirements.

Power installation

Consider using an uninterruptible power supply that is designed to provide clean power wherever possible. Depending on the power management and distribution panel design and the manufacturer of your power management, clean power ensures surge protection and avoids blackouts and excess current. Using an uninterruptible power supply is not required for AC-powered systems, but it is recommended.

PDU installation

To take advantage of RAS and availability features, including the ability to implement CEC Hot Add Repair Maintenance (CHARM), as introduced in 4.3, “CEC Hot Add Repair Maintenance (CHARM)” on page 121, remember to avoid single points of failure (SPOFs). It is advisable to ensure that you install PDUs from two separate power distribution panels. See Figure 4-1. In the case of a Power 780, you have to make sure that each power supply is connected to a separate PDU.

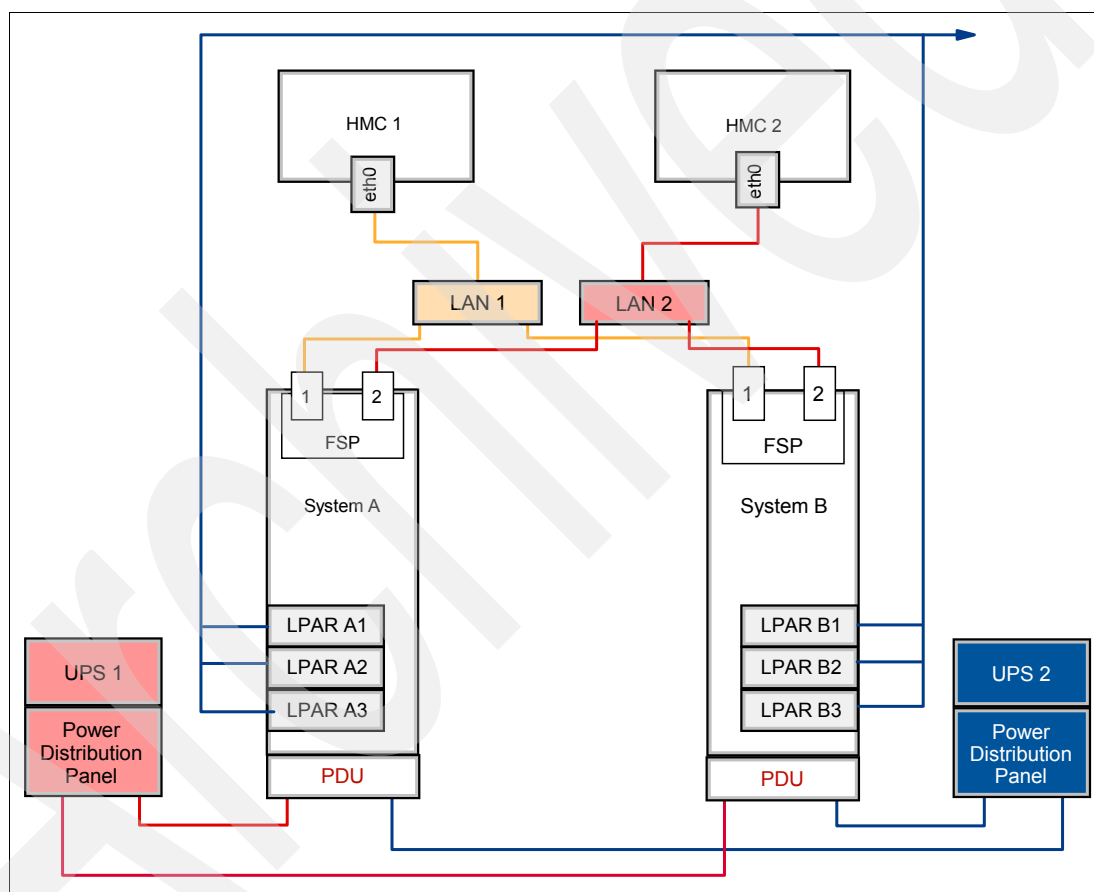


Figure 4-1 Installing from dual power sources

Optional battery backup: The Power 795 has an optional battery backup that can be used to complete transactions and shut the server down gracefully in the case of a total power loss.

4.1.3 Networks and storage area networks (SAN)

Network and SAN design falls outside of the scope of this publication. It is beneficial, however, for the administrator to understand and suggest the network connection between the managed server and the switches. Consider the following factors when connecting the managed server:

- ▶ For system management connectivity between the managed server and IBM Systems Director Management Console (SDMC) or Hardware Management Console (HMC). Make sure that each SDMC/HMC port connects to a separate switch. The port connected to the first SDMC/HMC switch must be on a separate virtual LAN (VLAN) from the port on the second switch. One SDMC or one HMC with the required code level must be connected to each of the two VLANs. A dedicated management server port must be in a VLAN with only one SDMC/HMC connecting to it for Dynamic Host Configuration Protocol (DHCP) requests.
- ▶ For public Ethernet connection, at least four Ethernet cables must be connected to either PCI slots or to the Integrated Virtual Ethernet Adapter (IVE) if you use dual virtual I/O servers, or at least a pair must be connected to each network switching device. For link aggregation, remember to consult the switch manufacturer's documentation for its capabilities. EtherChannel network interface backup (NIB) does not dictate how the connections are made to the switch or switches. If you are using a dual virtual I/O server, remember that each virtual I/O server uses redundant adapters connected to redundant switches to avoid either a switch, a cable, or an adapter single-point-of-failure (SPOF).
- ▶ SAN connection follows the same guidelines: Two cables per virtual I/O server, each on a separate host bus adapter (HBA) and each HBA connected to a separate SAN switch.
- ▶ Be careful when using dual port adapters because the adapter can be a SPOF.

Example 4-1 shows a typical installation list.

Example 4-1 Typical list for initial installation of a POWER7 system

Between SDMC/HMC and Management server

2 N/W Ports and 2 Cables connected to the Servers 2 SDMC/HMC ports : No IPs needed
DHCP

For LPARS : viopok1 and viopok2, 4 Ports, 2 on NWSwitch1 and 2 on NWSwitch2, 4
Cables accordingly 2 per VIO NIB or Link Aggregation

4 SAN ports, cabled to HBAs intended for VIO servers

The list in Example 4-1 helps to connect the managed server in the format that is shown in Figure 4-2 on page 103.

The diagram shows the following design:

- ▶ Two enclosures for CHARM
- ▶ Two SAN switches recommended
- ▶ Two network switches recommended
- ▶ Two management consoles (HMC or SDMC)

Notice how carefully the ports and cables are allocated to the virtual I/O server.

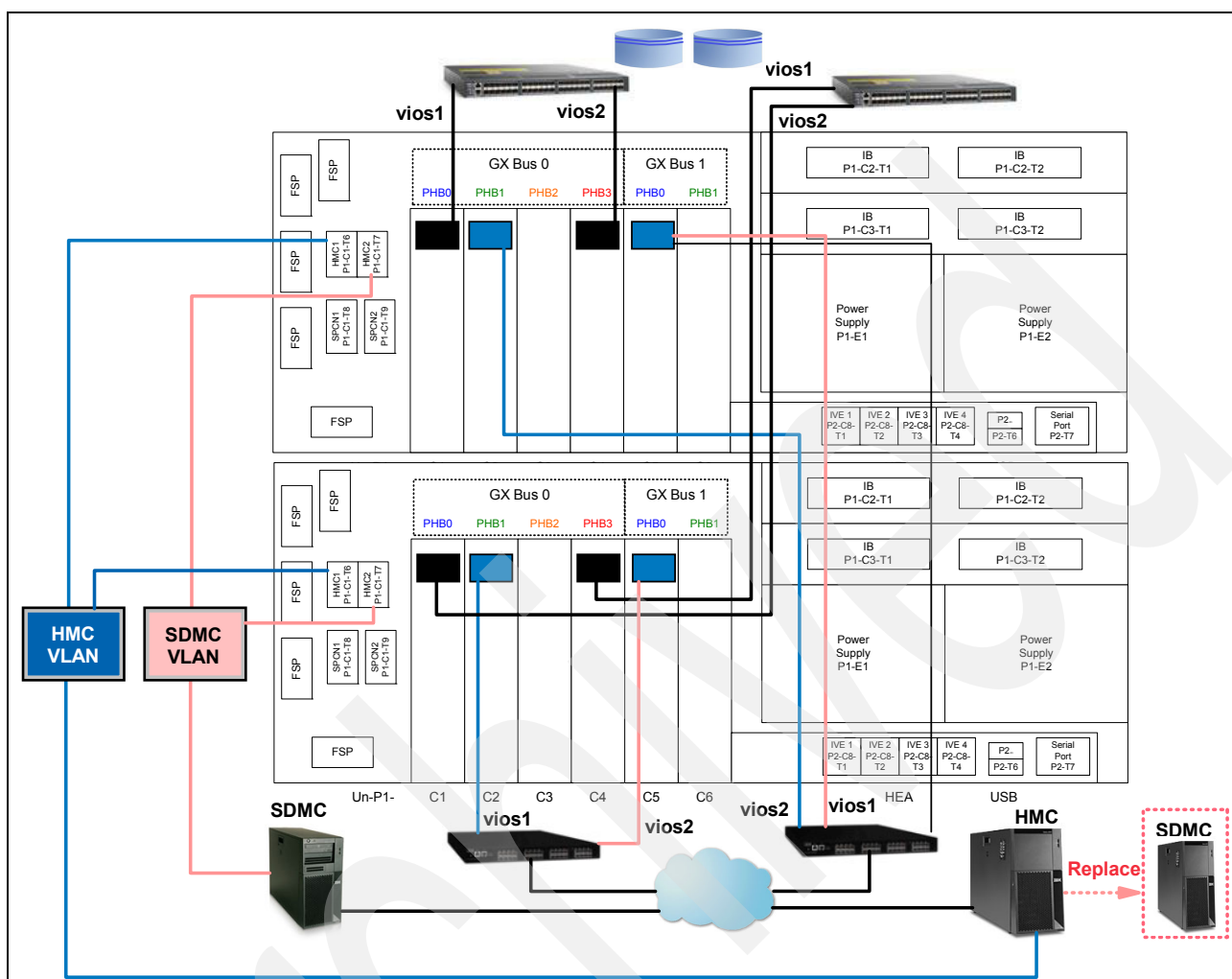


Figure 4-2 Setting up the Power7 to provide redundancy

4.2 Hardware planning

Physical hardware planning involves the placement of the hardware components to ensure that the RAS features, as well as the virtualization features, can be exploited. Certain reliability features are embedded within the server. There is no need to cater for them. These reliability features include processes, reliability retries, caches, memory, such as chipkill, and clock cards. The system takes care of these components transparently.

We do not intend to dictate the way that all environments are designed, but we do however attempt to provide preferred practices. Before any system installation or changes to the current system, use the System Planning Tool (SPT) to validate the intended configuration. We describe the SPT in 2.6.3, “Deployment using the System Planning Tool (SPT)” on page 40. Download the STP from this website:

<http://www.ibm.com/systems/support/tools/systemplanningtool>

The placement of the systems, racks, and cables affects the ability to implement CHARM, as discussed in 4.3, “CEC Hot Add Repair Maintenance (CHARM)” on page 121. When cabling,

remember that the cables must have enough length “slag”. Cable length must be long enough to allow an IBM service support representative (SSR) to pull out the components of the system that need to be serviced or replaced. Sufficient cable length is even more important if more than one Power system share the same rack. Use the provided cable arms to lead the cables so that they do not end up in disarray. Normally, using the provided cable arms to lead the cable ensures that the cable has enough length.

System planning

To use RAS features on an enterprise Power server, you need to have at least two system enclosures. These enclosures are interconnected with flexible symmetric multiprocessor (SMP) cables and service processor cables. These cables are designed to support scalability (hot add), hot repair, and concurrent maintenance. Figure 4-3 and Figure 4-4 on page 105 have been taken from the IBM Redpaper™ publication, *IBM Power 770 and 780 Technical Overview and Introduction*, REDP-4639.

Refer to this Redpaper for more information regarding the cabling of the IBM Power 780. This IBM Redpaper also covers installation of the GX++ adapters. Figure 4-3 shows the SMP cable installation.

We used the following steps in installing a four-node Power 780 (Refer to Figure 4-3):

1. The first system enclosure was installed on the topmost 4U of the rack. For example, if the rack is empty, you install the first system enclosure on U12-16. In this manner, you reserve the first 12Us for growth and avoid having to relocate in the future. This design is only a guideline and a preferred practice from our experiences.
2. Install a system enclosure Flex Cable from system enclosure 1 to system enclosure 2.
3. Add a third system enclosure Flex Cable from system enclosure 1 and system enclosure 2 to system enclosure 3.
4. Add a fourth node Flex Cable from system enclosure 1 and system enclosure 2 to system enclosure 4.

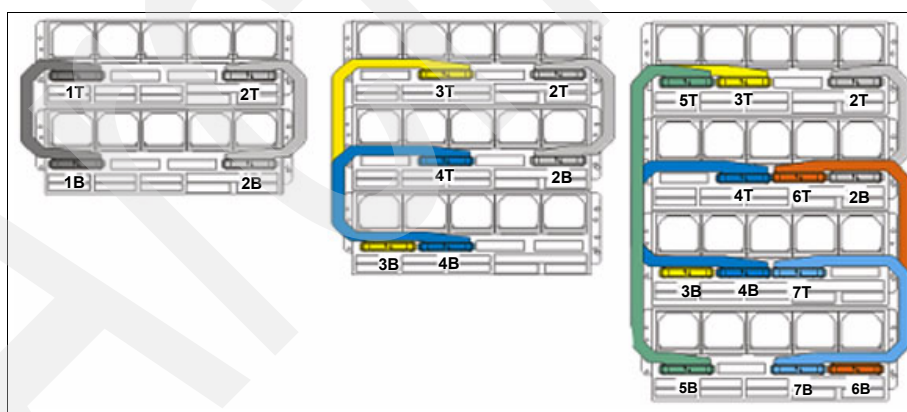


Figure 4-3 SMP cable installation order

Figure 4-4 shows the Flex Cable installation.

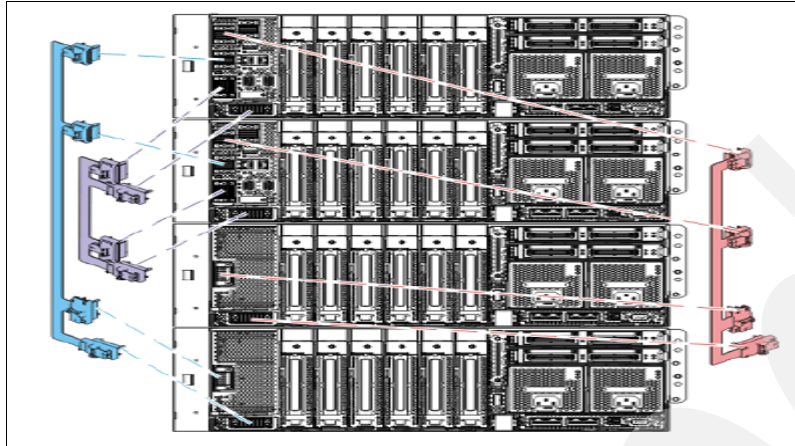


Figure 4-4 Flexible Service Processor (FSP) Flex cables

4.2.1 Adapters

To take advantage of CHARM, RAS, or high-availability features, the number of adapters connected to the server is important. For example, PCIe adapters are hot swappable, but the devices on the slot are not available when the adapter is maintained. You must have redundant adapters to avoid losing service. Many publications, including the IBM PowerHA publications, provide advice about the number of adapters, types, and placement of the adapters to take advantage of RAS features, as well as high-availability offerings. Although suggestions exist, an important concept is introduced to help eliminate single points of failure (SPOF).

SPOF concepts are explained in detail in the PowerHA publications. Both SAN and Ethernet adapters must have multiple ports. The use of multiple ports on a single adapter for the same environment might hinder RAS operation. Although the use of multiple ports on a single adapter might provide better throughput in the case of link aggregation for Ethernet and multi-pathing for the HBA (host bus adapter), the loss or failure of the adapter affects both ports. Loss or failure of the adapter also affects both ports on the Integrated Virtual Ethernet (IVE), which can, depending on the feature code, provide a total of 64 logical Ethernet ports called Local Host Ethernet Adapters (LHEA).

Choice of internal storage

When using internal disks, remember to use Logical Volume Mirroring (LVM) or mirrors. A number of choices are available to consider regarding disk types and placements. Types of disks range from storage area network (SAN), Small Computer System Interface (SCSI), IP-based SCSI (iSCSI), Serial Storage Architecture (SSA), serial-attached SCSI (SAS), and solid-state drives (SSD). Mirroring between SSD and other disk types is not supported. Refer to 2.7, “I/O considerations” on page 41 regarding storage considerations.

External I/O

If you need to implement an external I/O subsystem, use multiple GX++ busses. Using I/O devices that are connected to single I/O drawer is limited to a single system. These devices cannot be used for Logical Partition Mobility (LPM). And, these devices cannot be used in clustering that depends on operating system mirroring LVM. Wherever possible, use SAN-attached disks. IBM offers a range of external storage from which you can choose.

We list several types of external storage offered by IBM, as well as links for more details:

- ▶ IBM System Storage® N series
<http://www.ibm.com/systems/storage/network>
- ▶ IBM System Storage DS3000 family
<http://www.ibm.com/systems/storage/disk/ds3000/index.html>
- ▶ IBM System Storage DS5020 Express
<http://www.ibm.com/systems/storage/disk/ds5020/index.html>
- ▶ IBM System Storage DS5000
<http://www.ibm.com/systems/storage/disk/ds5000>
- ▶ IBM XIV® Storage System
<http://www.ibm.com/systems/storage/disk/xiv/index.html>
- ▶ IBM System Storage DS8700
<http://www.ibm.com/systems/storage/disk/ds8000/index.html>

4.2.2 Additional Power 795-specific considerations

The following list provides details about considerations that are specific to Power 795 servers:

Number of cores	Be careful when choosing specific Power 795 servers. Processor books can have either 24 cores in a book or 32 cores in a book. These processor books cannot be mixed. A 24-core processor book machine scales up to 192 cores. The 32-core processor book machine scales up to 256 cores. You can use the 32-core machine in <i>MaxCore</i> Or <i>Turbo Core</i> modes. See 2.3, “TurboCore and MaxCore technology” on page 28 for additional details.
Processors per LPAR	The maximum number of processors per LPAR.
Space requirements	Depending on the workload, if you expand beyond a single primary Power 795 rack, you might need to plan for another rack, either 24U/19 inches (5.79 meters) or 32U/24 inches (7.3 meters). Up to two expansion racks can be added to the Power 795.
Power source	Each expansion rack has redundant bulk power assemblies (BPA). Power all racks as shown in Figure 4-1 on page 101.
I/O drawers	Power 795 I/O drawers are separated into halves, which are identified by either P1 or P2. You can run the <code>lscfg -v1</code> command to see the slot number. Slots on the I/O drawers are hot swappable. Refer to documentation to confirm what is supported by which drawer (PCI, PCI-X, PCIe, and so on). I/O drawers can be connected to the processor book in either single-loop or dual-loop mode. Dual-loop mode is preferable whenever possible, because it provides the maximum bandwidth between the I/O drawer and the processor book, as well as independent paths to each of the I/O drawer planars.

Important: You must discuss all the previous considerations with the sales representative prior to the configuration. A few types of expansion racks are available (powered and non-powered). We only pointed out a few considerations. See the technical overview documents for a list of the available drawers that are specific to the server that you intend to deploy.

4.2.3 Planning for additional Power server features

This section discusses additional features for you to consider in detail. Certain features are standard with Power 795 and Power 780 servers. Other features might need to be enabled. The eConfig and System Planning Tool (SPT) are used to configure and specify the features. We discuss SPT in 4.8, “System Planning Tool (SPT)” on page 151. Consider the following points when planning to configure your system:

- ▶ Is the feature supported on the intended server? For example, Turbo core is available on Power 780, but not in the Power 770.
- ▶ Is the feature standard with the server or must it be specified on eConfig? For example, Active Memory Sharing (AMS) is standard with PowerVM on enterprise servers, but Active Memory Expansion (AME) must be configured and requires a specific server license.
- ▶ Can the feature be enabled after configuration? Can I upgrade between Editions (Express to Standard to Enterprise)?
- ▶ What are the supported server firmware, HMC, and SDMC code levels?
- ▶ What is the supported operating system level to take advantage of the features? For example, AME requires AIX 6.1 TL 4 with SP2 or later.
- ▶ What is the feature code, and how can I see, after installation, if the feature is enabled?
- ▶ Other considerations:
 - How dynamic is the feature enablement?
 - Do I need to restart the LPAR after enabling the feature?
 - Do I need to shut the LPAR down completely to reread the profile?
 - Do I need to shut the server (host) down before the feature is enabled?

Consider the following Power server features.

Active memory expansion

AME is the ability to compress memory pages up to twice the size of the actual memory. We discuss AME in 2.1.4, “Active Memory Expansion” on page 19.

Power management

We discuss power management in 2.5, “Power management” on page 36.

Active Memory Mirroring

Memory mirroring of the hypervisor is designed to mirror the main memory that is used by the system firmware to ensure greater memory availability by performing advanced error-checking functions. We discuss Active Memory Mirroring in 2.1.1, “Active Memory Mirroring for the hypervisor on Power 795” on page 13.

4.2.4 System management planning

The following sections provide information that is recommended for system management planning.

Management console

For the system management console (HMC or SDMC), we suggest the components:

- ▶ At least one 7310-CR3 or 7310-C05 is recommended for the HMC. Several HMC models are supported to manage POWER7 systems:
 - Desktop Hardware Management Console (HMC): 7310-C05, 7310-C06, 7042-C06, 7042-C07, or 7042-C08
 - Rack-mounted Hardware Management Console: 7310-CR3, 7042-CR4, 7042-CR5, or 7042-CR6
- ▶ At least a 7042-CR6 with feature code 0963 is suggested for an SDMC.
- ▶ The V7R710 code level or later is suggested for the HMC.

For the IBM Power 795, the licensed machine code Version 7 Revision 720 is required. For the IBM Power 780, the licensed machine code Version 7 Revision 710 SP1 is required. Note that an HMC is required, even if you plan to implement a full system partition server.

HMC code: You can download or order the latest HMC code from the Fix Central website:

<http://www.ibm.com/support/fixcentral>

- ▶ Director Version 6.2.1.2 or higher is suggested for the SDMC.

Check the HMC software level for compatibility with the entire configuration using the IBM Fix Level Recommendation Tool (FLRT):

<http://www14.software.ibm.com/webapp/set2/flrt/home>

Planning for the SDMC or HMC

IBM introduces the IBM Systems Director Management Console (SDMC). Due to the amount of detail that is discussed in Chapter 5, “POWER7 system management consoles” on page 157, we do not discuss SDMC planning in this chapter. You need to implement either an HMC or an SDMC to manage the resources of your Power system, particularly the Power 795 and Power 780. It is worthwhile to know that the SDMC provides the functionality of the HMC, as well as the functionality of the Integrated Virtualization Manager (IVM). If you still use the HMC, consider using the dual HMCs per managed server.

HMC: You must turn on the HMC before the managed server, because the managed server requests an IP address from the HMC. If more than one HMC is on that virtual LAN (VLAN), there is no way to know which HMC is managing the FSP. This situation also occurs if both ports of the managed server are on the same VLAN. You might end up with one HMC managing both ports. You have to search through all the HMCs on the same private VLAN to find the HMC that is managing the managed server.

Use Figure 4-5 on page 109 as a guideline for connecting the dual HMCs. Notice how two VLANs are specified in the diagram. Figure 4-5 on page 109 originated in the publication *Hardware Management Console V7 Handbook*, SG24-7491.

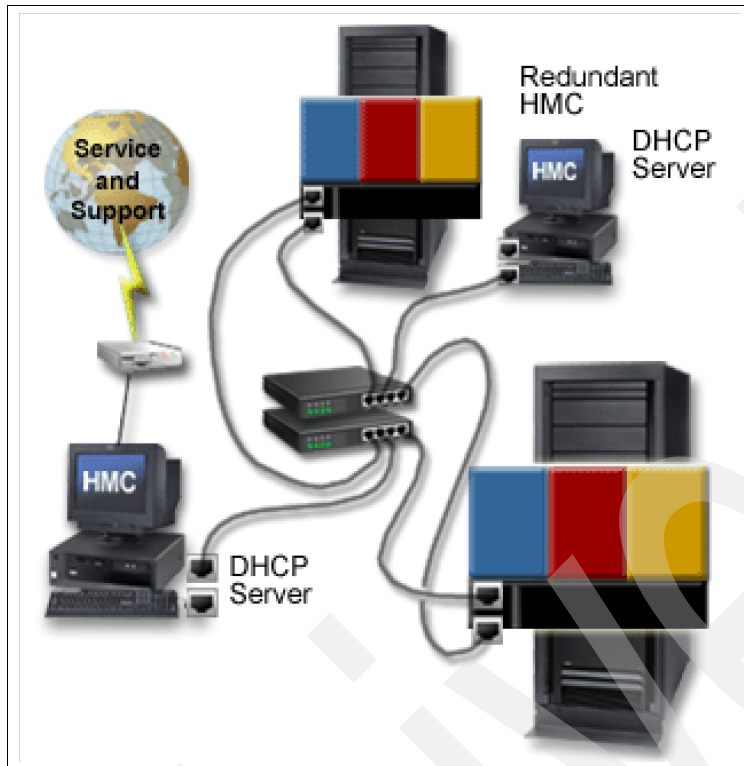


Figure 4-5 A redundant HMC setup

The following figures show a few possible HMC connections.

For a single drawer with dual HMCs, a recommended connection is shown on Figure 4-6 on page 110. This configuration provides dual HMC connection. It is however a single drawer and does not allow many RAS features to be performed. Depending on the redundancy configurations of each drawer, certain CHARM operations cannot be performed (refer to the 4.3, “CEC Hot Add Repair Maintenance (CHARM)” on page 121). Thus, there is no FSP redundancy.

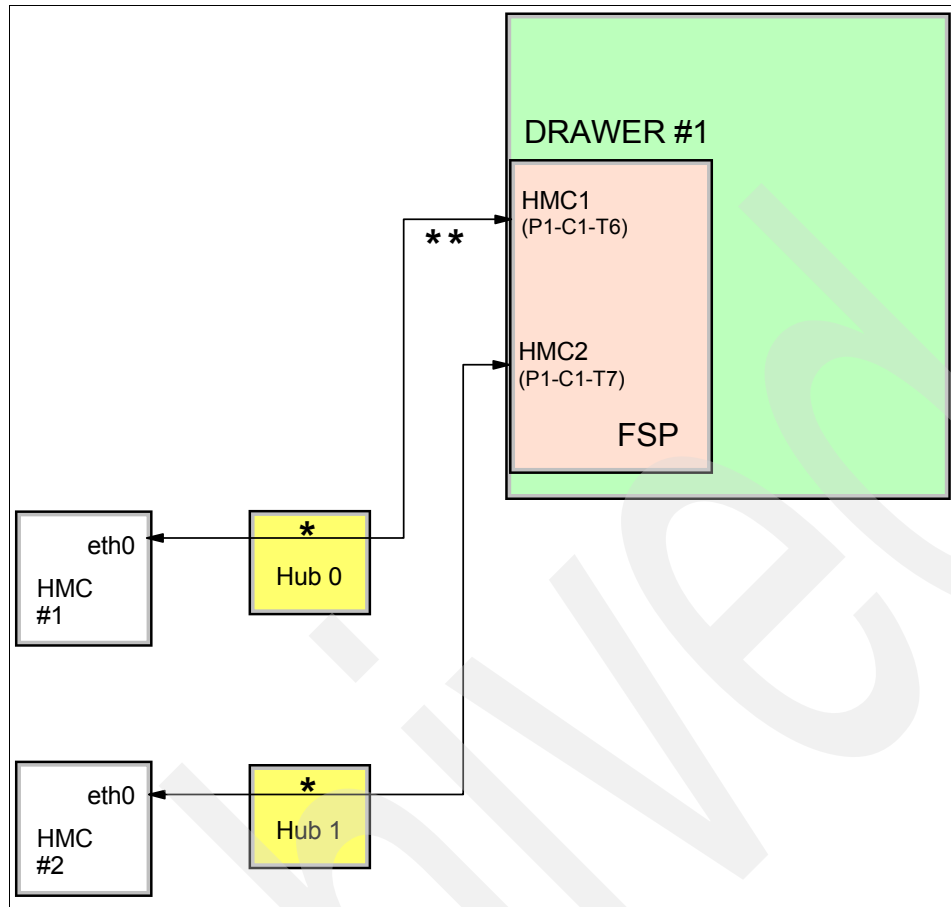


Figure 4-6 A single drawer Power 780 with dual HMC

Figure 4-7 on page 111 shows a preferred dual HMC connection where two drawers are connected. Each drawer has both HMC port 0 and port 1 connected to dual HMCs. This configuration can display redundancy on both HMC connections, as well as loss of FSP, depending on how the VIO servers are set up. With redundant LPAR/Virtual I/O setup, this configuration allows you to perform all CHARM operations without loss of clients. Refer to Figure 4-7 on page 111, which suggests how to cable dual VIO servers to provide the required level of redundancy.

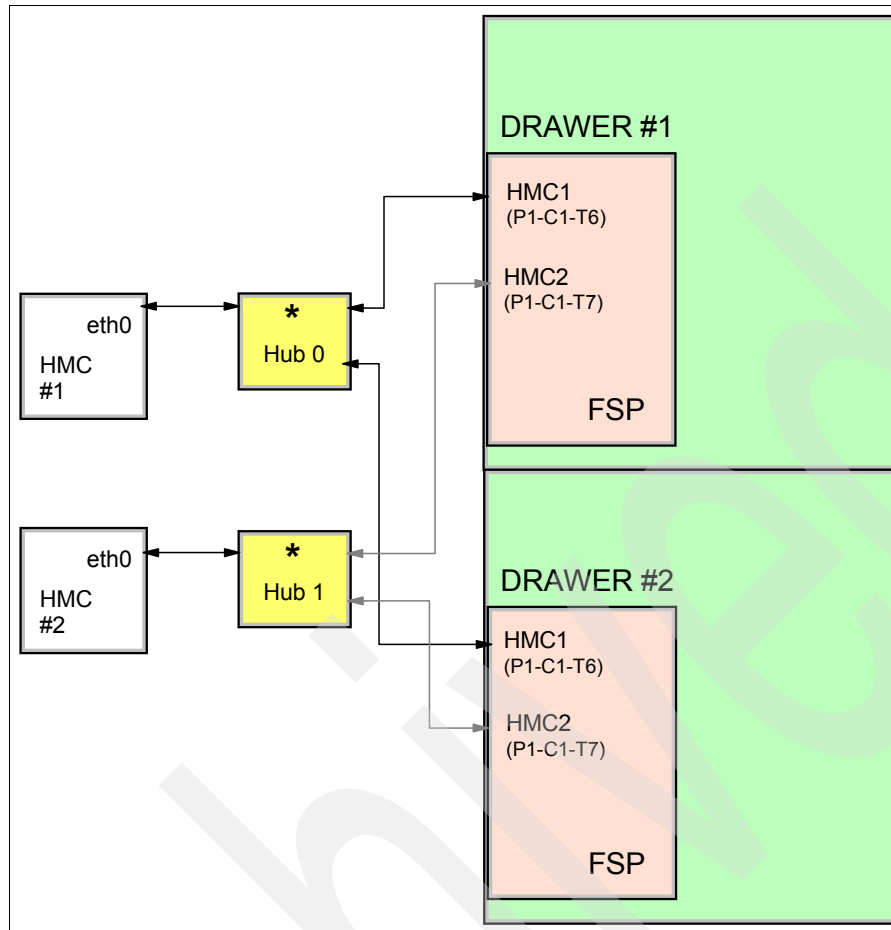


Figure 4-7 Dual HMC with redundant FSP

The version of HMC to support Power 770 and Power 780 servers is V7R710 SP1 or later. HMC functionality has been enhanced to support new features. These features include, but are not limited to, the following features:

- ▶ Support for POWER7 class servers
- ▶ Support for Active Memory Expansion
- ▶ Concurrent add/repair
- ▶ Redundant service processors
- ▶ Removal of the limit of 128 Active Memory Sharing partitions

Before updating the HMC to the latest version, consider the following factors:

- ▶ If the HMC has managed systems connected to it, and it is on the latest version, no actions need to be taken.
- ▶ If the HMC has managed systems, and the HMC is on an earlier version, upgrade the HMC to the appropriate version:
 - If the managed server's firmware is on a supported version for the HMC, you can simply upgrade the HMC to the required level.
 - If the managed server firmware is not on a supported version for the HMC, upgrade the server firmware to a supported level. Depending on the firmware level, you might have to either perform a concurrent firmware update where there is no need to shut down the managed server, or a firmware upgrade where the server needs to be shut down

for the firmware to take effect. Consider 3.1, “Live Partition Mobility (LPM)” on page 64 as an option to minimize the effect on service-level agreements (SLA) with your clients.

Update: Although it is possible to perform a concurrent update, We advise that you update in a negotiated maintenance window with a scheduled change.

- ▶ If a new HMC or the POWER7 server is the first device to connect to the HMC, consider upgrading the HMC to the latest version. Any server joining or connecting to the same HMC at a later stage must have its firmware upgraded to the supported level.

HMC and the network

The following sections describe the HMC with the private and public network characteristics.

HMC and the private network

The HMC is used to manage the server resources. You need to perform several operations through the HMC to manage the managed server resource allocation. The managed server continues to function in the absence of an HMC (for high-end systems, an HMC or SDMC is required), but server management is not possible:

- ▶ You cannot increase or decrease system resources.
- ▶ You cannot start up a partition that is not activated.
- ▶ You cannot perform LPM functions.

In the case of losing a single HMC, reconnect a new HMC as soon as possible. Backing up the HMC data (HMC backup or the data in the managed server FSP) allows the new HMC to populate and rebuild the managed server farm.

HMC and the public network

Certain computer center rules do not allow administrators into the data center without following certain rules, which makes it difficult to manage machines behind the HMC. The HMC has a port that can be connected to the public network. Through a web browser (previously websm), an administrator can manage the system from almost any location.

4.2.5 HMC planning and multiple networks

With the processing power and redundancy that are built into the Power 780 and Power 795, it is common to find a few business units hosted on the same server. This is the idea behind server virtualization. The challenge with this approach is that the HMC communicates with the LPARs on a managed server. The HMC must establish a Resource Monitoring and Control (RMC) connection with the LPARs, as well as the Virtual I/O servers on the machine. If the LPARs are not on the same network as the HMC, certain functions, such as dynamic LPAR, are not possible. Consult with the network administrator to discuss the firewalls and open rules that allow two-way RMC port connections between the networks.

4.2.6 Planning for Power virtualization

The IBM virtualization technology provides other offerings from which you can choose. Knowledge of these virtualization technology offerings helps you make the correct choice of the virtualization technology to use. The following virtual technologies are provided by the Power hypervisor:

- ▶ Logical partitioning
- ▶ Virtualized processors
- ▶ Hypervisor IEEE VLAN-compatible virtual switches for Virtual Ethernet

- ▶ Virtual SCSI adapters
- ▶ Virtual Fibre Channel (FC) adapters
- ▶ Virtual console (TTY)

We discuss these virtual resources in separate chapters throughout this publication.

Most of these virtualization features are managed by the Power hypervisor, which provides a logical mapping between the physical hardware and the virtual resources. This hypervisor functionality requires memory and CPU cycles to function. The hypervisor uses several of the server resources. Remember to check that there are resources available for any virtual request, for example, dynamic LPAR operations. When there are not enough resources for a hypervisor to perform an operation, an error appears, such as the error that is shown in Figure 4-8.

Certain factors might affect the amount of memory that is used by the hypervisor:

- ▶ Number of partitions
- ▶ Number of both physical and virtual devices
- ▶ Number of Virtual I/O servers created
- ▶ Number of N_Port ID Virtualization (NPIV) adapters created
- ▶ Partition profiles, for instance, the maximum memory that is specified on the profile

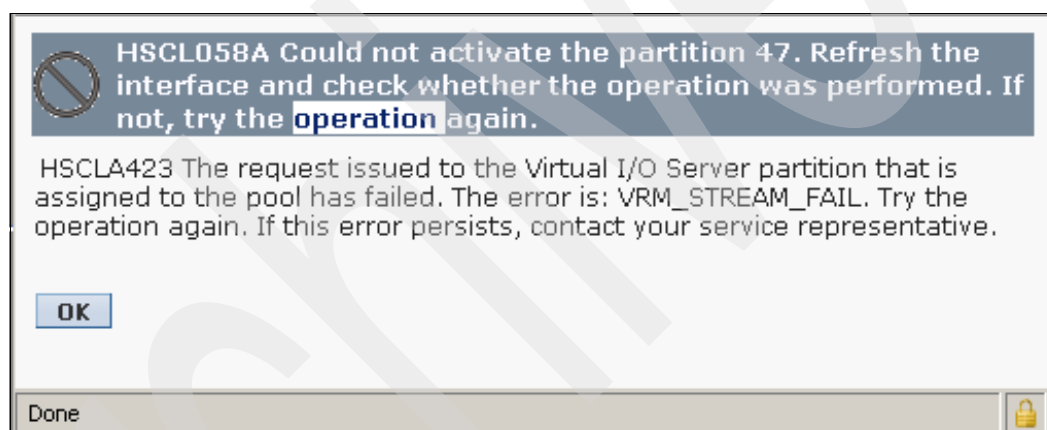


Figure 4-8 Example of hypervisor error due to resource constraints

In the next sections, we discuss planning for the other RAS features.

4.2.7 Planning for Live Partition Mobility (LPM)

Live Partition Mobility is designed to enable the migration of a logical partition (LPAR) from one system to another compatible system. You must ensure that the mobile partition (the partition that is to be moved) is configured correctly so that you can successfully move it from the source server to the destination server. Live Partition Mobility has specific requirements in terms of the operating system level, firmware level, storage layout, and network interfaces. A successful migration, regardless of whether it is active or inactive, requires careful deployment planning in advance. Sometimes, you can qualify a partition for mobility by taking additional steps, such as removing physical adapters (non-virtual adapters) or using a dynamic logical partitioning (DLPAR) operation.

The following requirements, as shown in Table 4-1 on page 114, are for the Power 780 and 795 servers to be eligible for migration. We describe all the requirements in detail later in this chapter. There might be separate requirements for other POWER7 models, so remember to check for the latest requirements.

Table 4-1 Power 780 and 795 server requirements for migration

Component	Prerequisites
Hardware levels	
HMC	CR7310-CR3 or later or the 7310-C05
SDMC	7042-CR6 with feature code 0963
Power Systems	POWER7 or POWER6
Software levels	
HMC	Version 7.7.1.0 or later for POWER7
SDMC	Director Version 6.2.1.2 or higher
System firmware	<ul style="list-style-type: none"> ▶ Ax710_065 or later, where x is an M for Midrange servers, such as 780 (or MHB) and an H for Enterprise Servers, such as 795 (or FHB) ▶ Source and destination systems can have separate levels of firmware, but the level on the source system must be compatible with the destination server's firmware.
PowerVM	PowerVM Enterprise Edition must be licensed and activated on both the source and destination servers.
Virtual I/O server	Minimum of one virtual I/O server on both source and destination systems at Release level 2.12.12 with Fix Pack 22.1 and Service Pack 1 or later for POWER7 servers
AIX	<ul style="list-style-type: none"> ▶ AIX Version 5.3 TL09 and Service Pack 7 or later for POWER7 ▶ AIX Version 6.1 TL02 and Service Pack 8 or later ▶ AIX Version 7.1
Red Hat Linux	RHEL Version 5 Update 5 or later (with the required kernel security update)
SUSE Linux	SUSE Enterprise Server 10 (SLES 10) Service Pack 3 or later (with the required kernel security update)
AMS (Not required for LPM)	<ul style="list-style-type: none"> ▶ PowerVM Enterprise Edition on both systems ▶ Firmware level EH340_075 or later ▶ HMC Version 7.3.4 Service Pack 2 for HMC-managed systems ▶ Virtual I/O Server Version 2.1.0.1 with Fix Pack 21 ▶ AIX 6.1 TL03 or AIX V7.1 ▶ SUSE SLES 11
NPIV (If used)	<ul style="list-style-type: none"> ▶ HMC Version 7.3.4 or later ▶ SDMC Version 6.2.1.2 or later ▶ Virtual I/O Server Version 2.1 with Fix Pack 20.1 ▶ Virtual I/O Server Version 2.3.1 required for NPIV on FCoCEE ▶ AIX V5.3 TL9 or later ▶ AIX V6.1 TL2 SP2 or later ▶ AIX V7.1

Component	Prerequisites
Environmental	
Storage	<ul style="list-style-type: none"> ▶ Must be shared by and accessible by at least one virtual I/O server on both the source and destination servers ▶ Must not have any required dedicated physical adapters ▶ Logical unit numbers (LUNs) must be zoned and masked to the virtual I/O servers on both systems ▶ Storage pools are not supported ▶ SCSI reservation must be disabled ▶ All shared disks have the reserve_policy set to "no_reserve"
Network	<ul style="list-style-type: none"> ▶ One or more physical IP networks or LANs that provide the necessary network connectivity for the mobile partition through a virtual I/O server partition on both the source and destination servers ▶ At least one virtual I/O server on each system must be bridged to the same physical network and the same subnet
Restrictions	
Source and destination server	<ul style="list-style-type: none"> ▶ Must be managed by the same HMC or SDMC (or redundant pair) ▶ Memory and processor resources required for current entitlements must be available on the destination server ▶ Logical memory block size must be the same on the source and destination systems
Destination server	Cannot be running on battery power at the time of the migration
Mobile partition	<ul style="list-style-type: none"> ▶ The partition must have a unique name ▶ Cannot use the Barrier Synchronization Register with an active migration ▶ The partition cannot be a virtual I/O server ▶ All I/O must be virtualized through the virtual I/O server ▶ All storage must reside on shared disks, not LVs ▶ The moving partition cannot be designated as a redundant error path reporting partition

Live Partition Mobility (LPM) is a feature of the PowerVM Enterprise Edition. There are two types of LPM: Active Partition Mobility and Inactive Partition Mobility. The migration process can be performed either with a live partition by Active Partition Mobility or with a powered-off partition by Inactive Partition Mobility.

Active Partition Mobility

Active Partition Mobility has the following characteristics:

- ▶ This type of migration allows you to migrate a running LPAR, including its operating system and applications, from a source system to a destination system.
- ▶ The operating system, the applications, and the services running on the migrated partition are not stopped during the migration.
- ▶ This type of migration allows you to balance workloads and resources among servers without any effect on the users.

Inactive Partition Mobility

Inactive Partition Mobility has the following characteristics:

- ▶ This type of migration allows you to migrate a powered-off LPAR from a source system to a destination system.
- ▶ Inactive Partition Mobility is executed in a controlled way and with minimal administrator interaction so that it can be safely and reliably performed.

To make full use of LPM, you must meet the following requirements and considerations.

Management console requirements and considerations

Beginning with HMC Version 7 Release 3.4, the destination system can be managed by a remote HMC. So, it is possible to migrate an LPAR between two IBM Power System servers, each of which is managed by a separate HMC. The following considerations must be in place:

- ▶ The source HMC and the destination HMC must be connected to the same network so that they can communicate with each other. This rule applies to the SDMC, as well.
- ▶ The source and destination systems, which can be under the control of a single HMC, can also include a redundant HMC.
- ▶ The source and destination systems, which can be under the control of a single SDMC, can also include a redundant SDMC.
- ▶ The source system is managed by an HMC and the destination system is managed by an SDMC.
- ▶ The source system is managed by an SDMC and the destination system is managed by an HMC.

Capacity: The HMC or SDMC can handle multiple migrations simultaneously. However, the maximum number of concurrent partition migrations is limited by the processing capacity of the HMC or SDMC.

Source and destination system requirements and considerations

The source and destination servers have these considerations:

- ▶ The source and destination systems must be an IBM POWER6-based model (and higher).
- ▶ Migration between systems with separate processor types is possible. You can obtain detailed information in the *IBM PowerVM Live Partition Mobility*, SG24-7460.

Firmware

The servers have these firmware considerations:

- ▶ **System firmware:** The firmware must be Ax710_065 or later, where the x is an M for Midrange servers, such as 780 (or MHB), and the x is an H for Enterprise Servers, such as 795 (or FHB).

Note: Source and destination systems can have separate levels of firmware. The level of source system firmware must be compatible with the destination firmware.

- ▶ Ensure that the firmware levels on the source and destination servers are compatible before upgrading and if you plan to use LPM.

Table 4-2 on page 117 shows the values in the left column that represent the firmware level from which you are migrating, and the values in the top row represent the firmware level to which you

are migrating. For each combination, *blocked* entries are blocked by code from migrating; *not supported* entries are not blocked from migrating, but are not supported by IBM; *mobile* entries are eligible for migration.

Table 4-2 Partition Mobility firmware support matrix

From/To	320_xxx	330_034 +	340_039 +	350_xxx +	710_xxx	720_xxx
320_xxx	Not supported	Not supported	Not supported	Not supported	Blocked	Blocked
330_034 +	Not supported	Mobile	Mobile	Mobile	Mobile	Blocked
340_039 +	Not supported	Mobile	Mobile	Mobile	Mobile	Mobile
350_xxx +	Not supported	Mobile	Mobile	Mobile	Mobile	Mobile
710_xxx	Blocked	Mobile	Mobile	Mobile	Mobile	Mobile
720_xxx	Blocked	Blocked	Mobile	Mobile	Mobile	Mobile

- ▶ Both source and destination systems must have PowerVM Enterprise Edition installed and activated.
- ▶ Ensure that the Logical Memory Block (LMB) is the same on both the source and destination systems (Refer to Figure 4-9).

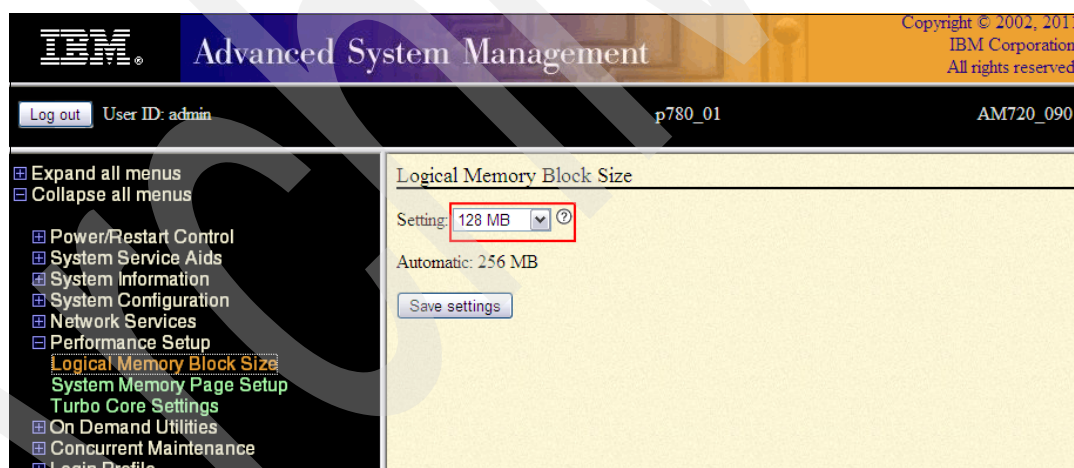


Figure 4-9 Verify LMB size from the HMC

- ▶ Destination systems must have enough available CPU and memory resources to host the mobile partitions.
- ▶ The destination server cannot be running on battery power at the time of migration.

To determine whether the destination server has enough available physical memory to support your mobile partition, complete the following steps from the HMC:

1. Identify the amount of physical memory that the mobile partition requires:
 - a. In the navigation pane, expand **Systems Management** → **Servers**.
 - b. Click the source server on which the mobile partition is located.
 - c. In the work pane, select the mobile partition.

- d. From the Tasks menu, click **Properties**. The Partition Properties window opens.
 - e. Click the **Hardware** tab.
 - f. Click the **Memory** tab.
 - g. Record the dedicated minimum, assigned, and maximum memory settings.
 - h. Click **OK**.
2. Identify the amount of physical memory that is available on the destination server:
 - a. In the navigation pane, expand **Systems Management** → **Servers**.
 - b. In the work pane, select the destination server to which you plan to move the mobile partition.
 - c. From the Tasks menu, click **Properties**.
 - d. Click the **Memory** tab.
 - e. Record the Current memory available for partition usage.
 - f. Click **OK**.
 3. Compare the values from steps 1 and 2.

Make sure that the server has enough processor memory by completing the previous steps from the HMC.

AME: In order to move an LPAR using AME through LPM to another system, the target system must support AME. The target system must have AME activated through the software key. If the target system does not have AME activated, the mobility operation fails during the pre-mobility check phase, and an appropriate error message is displayed.

Source and destination virtual I/O server requirements and considerations

The following considerations are for the virtual I/O server:

- ▶ Power 795: A dual Virtual I/O Server at V2.2 or higher must be installed on both the source and destination systems.
- ▶ Power 780: A Virtual I/O Server at V2.1.2.12 with Fix Pack 22.1 and Service Pack 2 or higher must be installed on both the source and destination systems.

You can obtain more information about the virtual I/O server and the latest downloads at the virtual I/O server website:

<http://www14.software.ibm.com/webapp/set2/sas/f/vios/download/home.html>

Operating system requirements

The operating system that runs in the mobile partition must be AIX or Linux:

- ▶ Power 795:
 - AIX Version 5.3.10.5 or later, 5.3.11.5 or later, or 5.3.12.1 or later.
 - AIX Version 6.1.0.6 or later (CSM 1.7.0.1 or later). Install APAR IZ95265 if the AIX level is 6100-06.
 - AIX Version 7.1.
 - Red Hat Enterprise Linux Version 5 (RHEL5) Update 5 or later (with the required kernel security update).
 - SUSE Linux Enterprise Server 10 (SLES 10) Service Pack 3 or later (with the required kernel security update).

- SUSE Linux Enterprise Server 11 (SLES 11) Service Pack 1 or later (with the required kernel security update).
- ▶ Power 780:
 - AIX Version 5.3.9.7 or later, 5.3.10.4 or later, or 5.3.11.2 or later.
 - AIX Version 6.1.2.8 or later (CSM 1.7.0.1 or later), 6.1.3.5 or later, or 6.1.4.3 or later. Install APAR IZ95265 if AIX level is 6100-06.
 - AIX Version 7.1.
 - SUSE Linux Enterprise Server 10 (SLES 10) Service Pack 3 or later (with the required kernel security update).
 - SUSE Linux Enterprise Server 11 (SLES 11) or later (with the required kernel security update).

To download the Linux kernel security updates, refer to the following website:

<http://www14.software.ibm.com/webapp/set2/sas/f/pm/component.html>

The previous versions of AIX and Linux can participate in inactive partition migration if the operating systems support virtual devices on IBM POWER6-based servers and POWER7-based servers.

Storage requirements

The following storage is required:

- ▶ For vSCSI:
 - Storage must be shared by and accessible by at least one virtual I/O server on both the source and destination systems.
 - SAN LUNs must be zoned and masked to at least one virtual I/O server on both the source and destination systems.
- ▶ For NPIV
 - You must zone both worldwide names (WWNs) of each of the virtual FC adapters.
- ▶ Storage pools and logical volumes are not supported.
- ▶ SCSI reservation must be disabled.
- ▶ All shared disks must have reserve_policy set to “no_reserve”.
- ▶ You must not have any required dedicated physical adapters for active migration.
- ▶ NPIV (if used):
 - HMC Version 7.3.4 or later
 - SDMC Version 6.2.1.2 or later
 - Virtual I/O Server Version 2.1 with Fix Pack 20.1
 - Virtual I/O Server Version 2.3.1 required for NPIV on FCoCEE
 - AIX V5.3 TL9 or later
 - AIX V6.1 TL2 SP2 or later
 - AIX V7.1

For a list of supported disks and optical devices, see the virtual I/O server data sheet for the virtual I/O server:

<http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/datasheet.html>

Network requirements

The migrating partition uses the virtual LAN (VLAN) for network access. Consider the following network requirements. The VLAN must be bridged to a physical network using a Shared Ethernet Adapter in the virtual I/O server partition. If there are multiple VLANs, the additional VLANs also must be bridged. Your LAN must be configured so that migrating partitions can continue to communicate with other necessary clients and servers after a migration is completed. At least one virtual I/O server on each machine must be bridged to the same physical network (same subnet). An RMC connection must be set up between the HMC and the LPARs, and it must be operational at all times.

Mobile partition requirement and considerations

The following requirements exist for the mobile partition:

- ▶ The mobile partition's OS must be installed in a SAN environment (external disk).
- ▶ The mobile partition cannot use the BSR for active migration, as shown in Figure 4-10.
- ▶ All I/O must be virtualized through the virtual I/O server.
- ▶ All storage must reside on shared disks (not LVs).

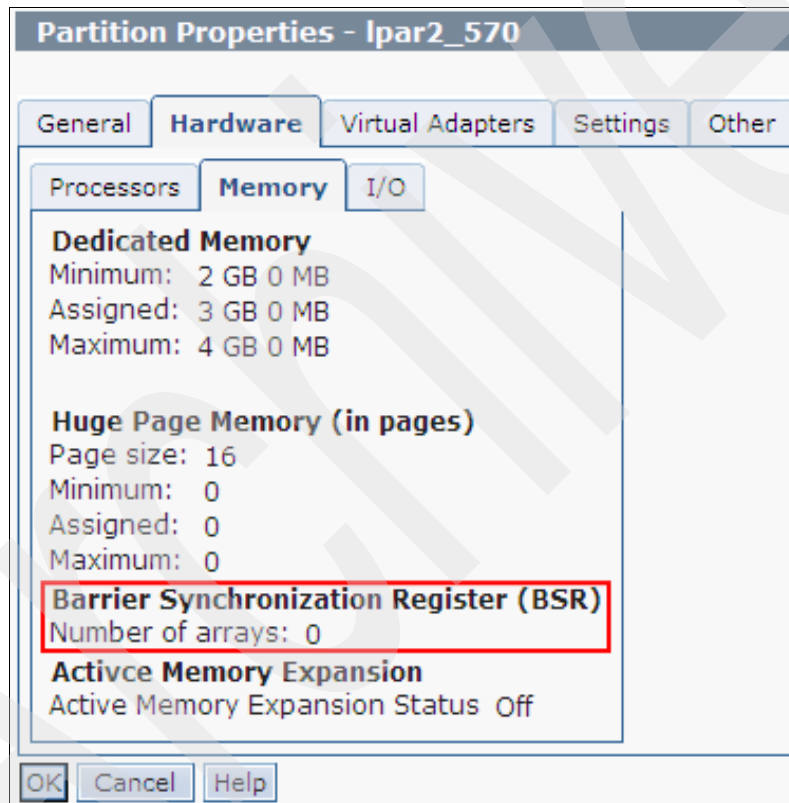


Figure 4-10 Barrier Synchronization Register (BSR)

- ▶ Ensure that the mobile partition's name is unique across both frames.

Mobile partition application requirement

Certain applications that are tied to the hardware identification information, such as license compliance managers, must be aware of the migration. You can obtain detailed information about the use of the IBM PowerVM Live Partition Mobility in *IBM PowerVM Live Partition Mobility*, SG24-7460.

You can obtain the latest information about LPM at the following website:

<http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7hc3/iphc3whatsnew.htm>

4.3 CEC Hot Add Repair Maintenance (CHARM)

Concurrent add and repair capabilities for Power Systems servers have been introduced incrementally since 1997, starting with the power supply, fan, I/O device, PCI adapter, and I/O enclosure/drawer. In 2008, IBM introduced significant enhancements to Enterprise Power Systems 595 and 570 that highlighted the ability to perform node add/upgrade/maintenance concurrently, without powering down the system. CEC hot add and repair maintenance (CHARM) offers new capabilities in reliability, availability, and serviceability (RAS). With the introduction of POWER7 in 2010, these capabilities continue, but the terminology has changed:

- ▶ CEC Concurrent Maintenance (CCM) for Power 570 and Power 595
- ▶ CEC Hot Add Repair Maintenance (CHARM) for Power 770, Power 780, and Power 795

Table 4-3 shows the changed POWER7 terminology compared to the POWER6 terminology.

Table 4-3 New POWER7 terminology

POWER6 CEC CCM terminology	New POWER7 terminology
"Concurrent" when referring to CEC hardware	"Hot" when referring to CEC hardware
CCM: CEC Concurrent Maintenance	CHARM: CEC Hot Add and Repair Maintenance
Concurrent Node Add	Hot Node Add
Concurrent Node Upgrade (memory)	Hot Node Upgrade (memory)
Concurrent Hot Node Repair Concurrent Cold Node Repair	Hot Node Repair
Concurrent GX Adapter Add	Concurrent GX Adapter Add
Concurrent Cold GX Adapter Repair	Hot GX Adapter Repair
Concurrent System Controller Repair	Concurrent System Controller Repair

Hot GX Adapter repair is supported from POWER7.

4.3.1 Hot add or upgrade

The CHARM functions provide the ability to add/upgrade system capacity and repair the Central Electronic Complex (CEC), including processors, memory, GX adapters, system clock, and service processor without powering down the system. The *hot node add* function adds a node to a system to increase the processor, memory, and I/O capacity of the system. The *hot node upgrade* (memory) function adds additional memory dual inline memory modules (DIMMs) to a node, or upgrade (exchange) existing memory with higher-capacity memory DIMMs. The system must have two or more nodes to utilize the hot node upgrade function. *To take full advantage of hot node add or upgrade, partition profiles must reflect higher maximum processor and memory values than the values that existed before the upgrade. Then, the new resources can be added dynamically after the add or upgrade.*

Important: Higher maximum memory values in the partition profiles increase the system memory set aside for partition page tables; changing maximum memory values requires the partition reactivation of a new profile.

You can estimate the increased system memory with SPT. Figure 4-11 shows a modification of the maximum memory setting for a partition changing from 32768 to 47616, and the corresponding change to the hypervisor memory.

IBM System Planning Tool

System plan: ITSO_test System: ITSO_test [IBM Power 780 Server (9179-MHB)]

System **Partitions** Hardware Networking Virtual Storage Consoles Summary

Partition properties Processors **Memory**

Memory

System memory (MB): 65536
Configured memory (MB): 33792
Hypervisor memory (MB): 2304
Unassigned memory (MB): 29440
Logical memory block size (MB): 256

Memory for Partitions

Name	ID	Operating System	Min	Desired	Max
LPAR1	1	Virtual I/O Server	512	512	512
LPAR2	2	Virtual I/O Server	512	512	512
LPAR3	3	AIX 6.1	7168	32768	32768

OK Apply Save... Cancel Report Help

Memory

System memory (MB): 65536
Configured memory (MB): 33792
Hypervisor memory (MB): 3072
Unassigned memory (MB): 28672
Logical memory block size (MB): 256

Memory for Partitions

Name	ID	Operating System	Min	Desired	Max
LPAR1	1	Virtual I/O Server	512	512	512
LPAR2	2	Virtual I/O Server	512	512	512
LPAR3	3	AIX 6.1	7168	32768	47616

OK Apply Save... Cancel Report Help

Figure 4-11 Checking the increased system memory by using IBM System Planning Tool

The *concurrent GX Adapter add function* adds a GX adapter to increase the I/O capacity of the system. You can add one GX adapter concurrently to a Power 770/780 system without planning for a GX memory reservation. To concurrently add additional GX adapters, additional planning is required (refer the following note and to Figure 4-12 for more details).

Memory reservations: The system firmware automatically makes GX adapter memory reservations to support a concurrent GX adapter add. The default Translation Control Entry (TCE) memory reservations are made in the following manner:

- ▶ One GX adapter (128 MB) maximum, if an empty slot is available
- ▶ One adapter slot per node, two slots maximum for 795, and one adapter slot for 780

The system administrator can change the default value for a GX adapter from zero to the total number of empty slots in the system via the service processor Advanced System Management Interface (ASMI) menu. The change takes effect on the next system IPL.

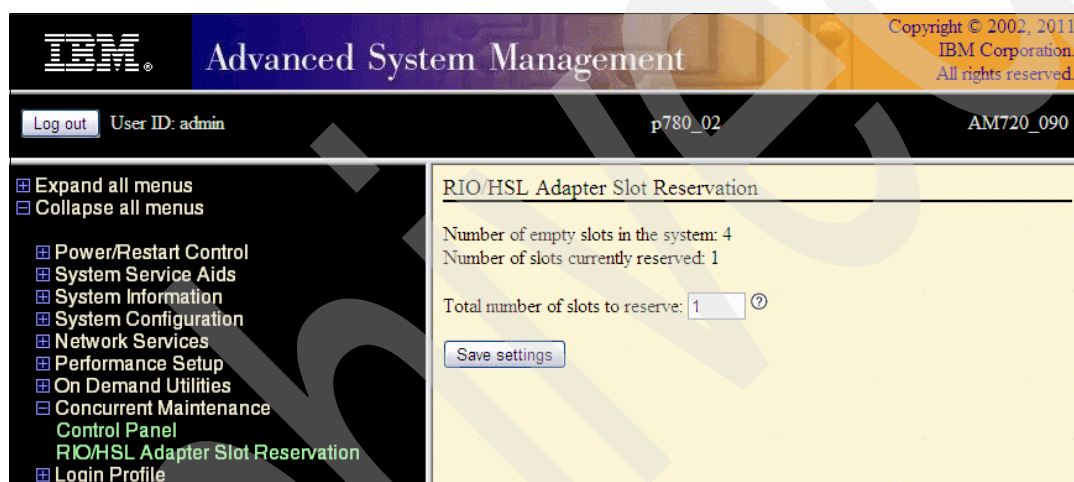


Figure 4-12 Altering the GX adapter memory reservation with ASM

4.3.2 Hot repair

Hot node repair repairs defective hardware in a node of a system. *The system must have two or more nodes to utilize the hot node repair function.* *Hot GX adapter repair* repairs a defective GX adapter in the system. And, *system controller repair* (795) repairs a defective service processor.

Node and I/O evacuation:

- ▶ For hot node upgrade or repair, processors and memory in use within the target node are relocated to alternate nodes with available resources.
- ▶ I/O devices that are attached to the target node or I/O hub must be removed from usage by the system administrator.

CoD resources: Unlicensed Capacity on Demand (CoD) resources are used by system firmware automatically without a CoD usage charge to meet node evacuation needs during the CHARM operation.

4.3.3 Planning guidelines and prerequisites

Implementing CHARM requires careful advanced planning and meeting all prerequisites. You need to request the free pre-sales “I/O Optimization for RAS” services offering. The system must have spare processor and memory capacity to allow a node to be taken offline for hot repair or upgrade.

Capacity: If you do not have spare processor and memory capacity, you can use either the dynamic LPAR operation to reduce processor and memory to minimum size or LPM to move a partition or several partitions to another server. Otherwise, shut down the low priority or unnecessary partitions.

You must configure all critical I/O resources using an operating system multi-path I/O redundancy configuration, for example, multi-path I/O (MPIO), SDDPCM, PowerPath, HDLM, and so on).

You must configure redundant I/O paths through separate nodes and GX adapters, because the I/O expansion units that are attached to the GX adapters in that node are unavailable during a hot node repair or upgrade procedure. These separate nodes and GX adapters can be either directly attached I/O or virtual I/O that is provided by dual virtual I/O servers housed in separate nodes (Figure 4-13).

Figure 4-13 shows the system configuration with redundant virtual I/O servers and redundant I/O adapters to improve the I/O availability and reduce the effect of a hot node repair or upgrade operation.

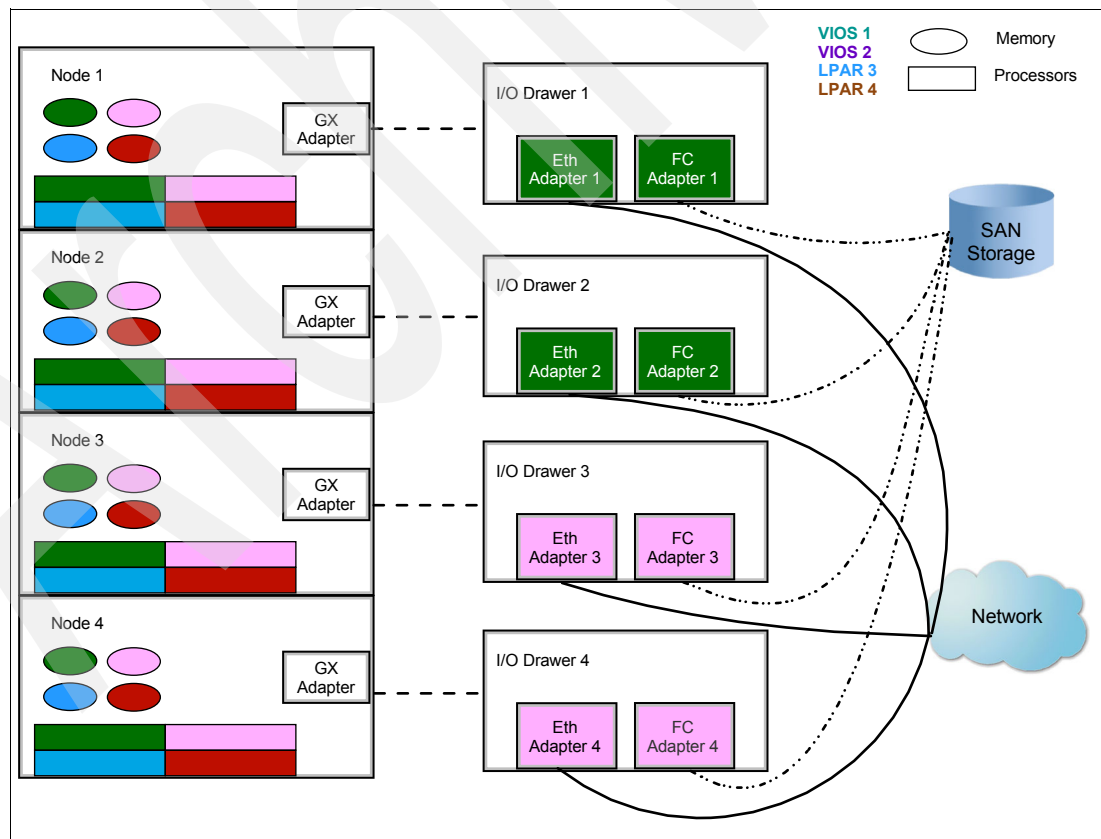


Figure 4-13 System configuration with redundant virtual I/O servers and I/O paths

Figure 4-14 shows the connections that are disrupted during the hot repair of node 1. A partition with paths to the I/O configured through node 1 and at least one other node continue to have access to the Ethernet and storage networks.

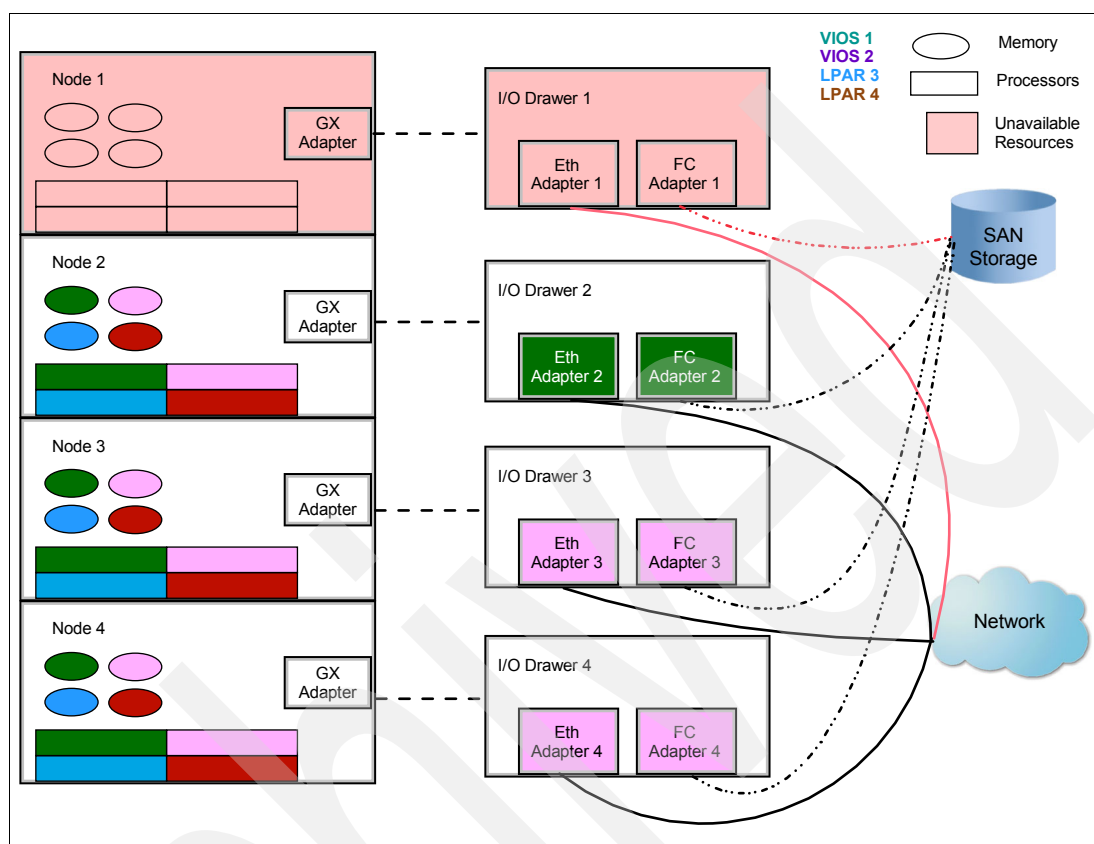


Figure 4-14 Redundant virtual I/O servers and I/O paths during the hot repair of node 1

Consider these points about hot repair:

- ▶ It is strongly recommended that you perform all scheduled hot adds, upgrades, or repairs during off-peak hours.
- ▶ Electronic Service Agent (ESA) or Call-Home must be enabled.
- ▶ All critical business applications are moved to another server using LPM, if available. Or, critical applications are quiesced for hot node add, hot node repair, hot node upgrade, and hot GX adapter repair.
- ▶ All LPARs must have an RMC network connection with the HMC.
- ▶ You must configure the HMC with a redundant service network for both service processors in the CEC for hot repair or upgrade of the 780.
- ▶ Do not select the “Power off the system after all the logical partitions are powered off” property system setting (Figure 4-15 on page 126).

p780_02

General Processors Memory I/O Migration Power-On Parameters Capabilities Advanced

Name: * **p780_02**

Serial number: 109AF7P
 Type/Model: 9179-MHB
 State: Operating
 Attention LED: Off
 Service processor version: 00070000
 Manufacturing default configuration: False
 Maximum number of partitions: 320
 Service partition: Unassigned

☐ Power off the system after all the logical partitions are powered off.

OK Cancel Help

Figure 4-15 Do not select Power off the system after all the logical partitions are powered off

Table 4-4 summarizes the minimum enablement criteria for individual CHARM functions, as well as other concurrent maintenance functions.

Table 4-4 CCM/CHARM minimum enablement criteria

Functions	Criteria			
	Off-peak ^a	Redundant I/O ^b	ESA-enabled ^c	LPM or quiesce ^d
Fan/Blower/Control Add, Repair	Recommend			
Power Supply/Bulk Power Add, Repair	Recommend			
Operator Panel	Recommend			
DASD/Media Drive & Drawer Add	Recommend			
DASD/Media Drive & Drawer Repair	Recommend	Prerequisite		
PCI Adapter Add	Recommend			
PCI Adapter Repair	Recommend	Prerequisite		
I/O Drawer Add	Recommend			
I/O Drawer Repair, Remove	Recommend	Prerequisite		
System Controller Repair	Recommend			
GX Adapter Add	Recommend		Prerequisite	
GX Adapter Repair	Recommend	Prerequisite	Prerequisite	Prerequisite
Node Add	Recommend		Prerequisite	Prerequisite
Node Upgrade (Memory ^e)	Recommend	Prerequisite	Prerequisite	Prerequisite
Hot Node Repair	Recommend	Prerequisite	Prerequisite	Prerequisite

- a. Highly recommend that schedule upgrades or repairs are done during “non-peak” operational hours.
- b. Prerequisite that critical I/O resources are configured with redundant paths.
- c. Electronic Service Agent (ESA) enablement highly recommended for POWER6 systems and prerequisite for POWER7 systems.
- d. Prerequisite that business applications are moved to another server using LPM, if available, or critical applications quiesced.
- e. IBM recommends that you not dynamically change the size of the 16 M large page pool in AIX partitions with the `vmo` command while a CCM/CHARM operation is in progress.

Next, we describe the supported firmware levels. Table 4-5 provides the minimum and recommended system firmware, HMC levels, and IBM Systems Director Management Console (SDMC) levels for CEC hot node add and hot node repair maintenance operations on Power 780. Refer to Table 4-6 on page 128 provides the system firmware, HMC levels, and SDMC levels for Power 795.

Table 4-5 System firmware, HMC levels, and SDMC levels for add/repair on Power 780

Function	Minimum system firmware, HMC levels, and SDMC levels	Recommended system firmware, HMC levels, and SDMC levels
Hot node add/ Hot node repair	AM720_064 or later V7R7.2.0 + MH01235	AM720_084 or later V7R7.2.0 + MH01246
Hot memory add or upgrade	AM720_064 or later V7R7.2.0 + MH01235	AM720_084 or later V7R7.2.0 + MH01246
Hot GX adapter add	All levels V7R7.1.0	AM720_084 or later V7R7.2.0 + MH01246
Hot GX adapter repair	AM720_064 or later V7R7.2.0 + MH01235	AM720_084 or later V7R7.2.0 + MH01246

For more details and the latest update on the minimum and recommended firmware levels for CHARM on Power 780, refer to the IBM Power Systems Hardware Information Center:

http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7ed3/p7ed3cm_matrix_mmb.htm

Important: If there are two HMCs or SDMCs attached to the system, both HMCs or SDMCs must be at the same level. If not, the HMC or SDMC that is not at the required level must be disconnected from the managed system and powered off.

To view the HMC machine code version and release, follow these steps:

1. In the Navigation area, click **Updates**.
2. In the Work area, view and record the information that appears under the HMC Code Level heading, including the HMC version, release, maintenance level, build level, and base versions.

To view the SDMC appliance code version and release, follow these steps:

1. On the SDMC command line, type `lsconfig -v`.
2. View and record the information that is displayed under the SDMC Code Level heading, including the SDMC version, release, service pack, build level, and base versions.

To disconnect an HMC: To disconnect an HMC from a managed system, follow these steps:

1. On the ASMI Welcome pane, specify your user ID and password, and click **Log In**.
2. In the navigation area, expand **System Configuration**.
3. Select **Hardware Management Consoles**.
4. Select the desired HMC.
5. Click **Remove connection**.

To disconnect an SDMC: To disconnect an SDMC from a managed system, refer to this website:

http://publib.boulder.ibm.com/infocenter/director/v6r2x/topic/dpsm/dpsm_trouble_shooting/dpsm_troubleshooting_managedsystemstate_conn_prob.html

Table 4-6 System firmware, HMC levels, and SDMC levels for Power 795

Function	Minimum system firmware, HMC levels, and SDMC levels	Recommended system firmware, HMC levels, and SDMC levels
Hot node add/ Hot node repair	AH730_0xx or later V7R7.3.0 + MHyyyy	AH730_0xx or later V7R7.3.0 + MHyyyy
Hot memory add or upgrade	AH730_0xx or later V7R7.3.0 + MHyyyy	AH730_0xx or later V7R7.3.0 + MHyyyy
Hot GX adapter add	AH730_0xx or later V7R7.3.0 + MHyyyy	AH730_0xx or later V7R7.3.0 + MHyyyy
24-inch I/O drawer add/removal	All levels V7R7.2.0	All levels V7R7.2.0

For more details and the latest update on the minimum and recommended firmware levels for CHARM on Power 795, refer to the IBM Power Systems Hardware Information Center:

http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7ed3/p7ed3cm_matrix_mmb_9119.htm

Next, we discuss IBM i planning considerations. To allow for a hot node repair/memory upgrade to take place with i partitions running, the following PTFs are also required:

- ▶ V5R4: MF45678
- ▶ V6R1: MF45581

If the PTFs are not activated, the IBM i partitions have to be powered off before the CHARM operation can proceed.

Features not supported: The following features and capabilities are not supported in conjunction with CHARM:

- ▶ Systems clustered using RIO-SAN technology (this technology is used only by IBM i users clustering using switchable towers and virtual OptiConnect technologies).
- ▶ I/O Processors (IOPs) used by IBM i partitions do not support CHARM (any IBM i partitions that have IOPs assigned must either have the IOPs powered off or the partition must be powered off).
- ▶ Systems clustered using InfiniBand technology (this capability is typically used by High Performance Computing clients using an InfiniBand switch).
- ▶ Sixteen GB memory pages, which are also known as huge pages, do not support memory relocation (partitions with 16 GB pages must be powered off to allow CHARM).

Table 4-7 provides estimated times in minutes for each activity (by role) for a CHARM operation on a Power 780 Server. The times are shown in minutes, and they are approximations (~). The estimated times are for a single operation. For a large MES upgrade with multiple nodes or GX adapters, careful planning by the system administrator and IBM system service representative (SSR) must be done to optimize the overall upgrade window.

Table 4-7 Estimated time for CHARM operation on a Power 780

Operation	System administrator time (minutes)		SSR time (minutes)		
	Prepare for node/GX evacuation	Resource allocation/restore	Memory relocation (32-512GB)	Firmware deactivate/activate	Physically remove/install
Node Add	N/A	~30	N/A	~30 - 45	~60
Node Upgrade	~30 - 60	~30	~11 - 77	~25 - 40	~15
Node Repair	~30 - 60	~30	~11 - 102	~25 - 40	~15 - 20
GX Add	N/A	~15	N/A	~10	~5
GX Repair	~10 - 30	~15	N/A	~15 - 20	~8

There are rules for CHARM operations:

- ▶ Only a single hot add or repair operation can be performed at one time from one HMC.
- ▶ In a dual management console environment, all CHARM operations must be performed from the primary management console.

Non-primary management console: In V7R7.3.x.x of HMC and SDMC, a new feature was added that if you start a repair from the non-primary management console, it asks if you want to make the console from which you are running the procedure the primary management console. It then tries to renegotiate the role of the primary management console. If the non-primary HMC can become the primary management console, it allows you to continue with the procedure on this console. Refer to Figure 6-6 on page 199.

- ▶ The “Prepare for Hot Repair/Upgrade” task must be run by the system administrator to determine the processor, memory, and I/O resources that must be freed up prior to the start of concurrent operation (this task is for the system administrator).

Note: The Prepare for Hot Repair/Upgrade utility is a tool for the system administrator to identify the effects to system resources in preparation for a hot node repair, hot node upgrade, or hot GX adapter repair operation. Refer to Figure 4-16. This utility provides an overview of platform conditions, partition I/O, and processor and memory resources that must be freed up for a node evacuation. A node is a drawer in a 9117-MMB, 9179-MHB, or 9119-FHB system. For more details about the Prepare for Hot Repair/Upgrade utility, refer to the IBM Power Systems Hardware Information Center: <http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/topic/p7ed3/ared3nodeevac.htm>

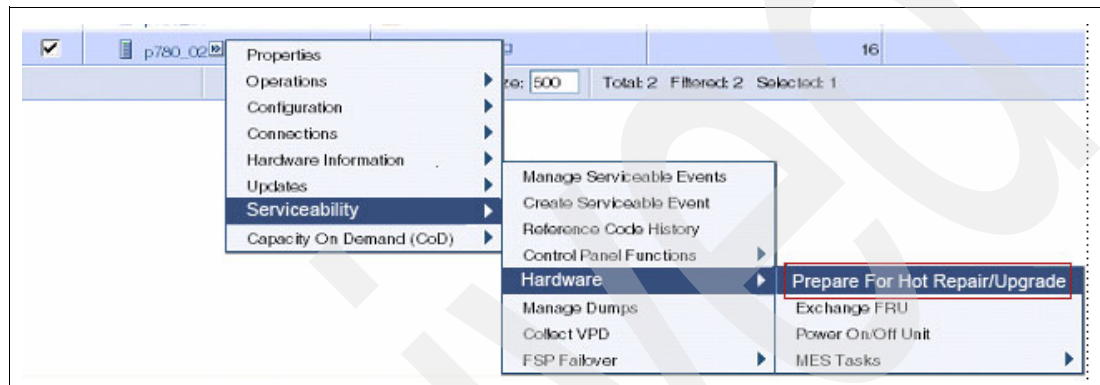


Figure 4-16 Prepare for Hot Repair/Upgrade utility

- ▶ A second hot add or repair operation cannot be started until the first one has completed successfully. If, at first, the hot operation fails, the same operation must be restarted and completed before attempting another operation.
- ▶ Multiple hot add or repair operations must be completed by performing a series of single hot add or repair operations.
- ▶ You must enable the service processor redundancy capability, if it has been disabled, before a CHARM operation, except on a Power 780 with a single node.
- ▶ An IBM service representative (SSR) must perform the execution of CHARM procedures and the physical hardware removal and replacement.

You can find additional information about CHARM on the Power 780 Server at the following website:

- ▶ Planning for CEC hot node add and hot node repair maintenance
http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7ed3/abstract_ared3.htm
- ▶ Planning for concurrent GX adapter or hot node add
<http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7ed3/ared3addhardware.htm>
- ▶ Planning for hot GX adapter or node repair
<http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7ed3/ared3repairhardware.htm>
- ▶ Planning for adding or upgrading memory
<http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7ed3/ared3kickoff.htm>

For the latest version of the planning checklist, refer to the “Planning for CEC hot-node add and hot-node repair maintenance” section of the IBM Power Systems Hardware Information Center:

<http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7ed3/ar ed3kickoff.htm>

4.4 Software planning

Next, we describe the operating system support requirements and key prerequisites for the Power 780 and Power 795.

First, we discuss the following AIX operating system requirements for POWER7 systems:

- ▶ AIX V7.1
- ▶ AIX V6.1, with the 6100-06 Technology Level
- ▶ AIX V5.3, with the 5300-12 Technology Level and Service Pack 1, or later
- ▶ AIX V5.3, with the 5300-11 Technology Level and Service Pack 5, or later (availability 30 September 2010)
- ▶ AIX V5.3, with the 5300-10 Technology Level and Service Pack 5, or later (availability 30 September 2010)

Although AIX 7.1 is supported on older Power servers, when running on the POWER7 servers, AIX 7.1 is the first OS version that allows an application to scale beyond 64 cores/128 threads to reach 256 cores and 1,024 threads in a single instance. Also, with AIX 7.1, you are able to control which processors (cores) are allowed to be interrupted in order to handle typical system interrupt requests. This new capability in AIX and Power enables a much more friendly environment for business critical applications that require true real-time processing.

If installing the IBM i operating system on POWER7 systems, the following versions are supported:

- ▶ IBM i 7.1, or later
- ▶ IBM i 6.1, with 6.1.1 machine code, or later

If installing the Linux operating system on POWER7, the following versions are supported:

- ▶ Red Hat Enterprise Linux AP 5 Update 5 for Power, or later
- ▶ SUSE Linux Enterprise Server 10 Service Pack 3, or later
- ▶ SUSE Linux Enterprise Server 11 Service Pack 1, or later

For systems ordered with the Linux operating system, IBM ships the most current version that is available from the distributor. If you require a separate version than the version that is shipped by IBM, you must obtain it via a download from the Linux distributor's website. Information concerning access to a distributor's website is located on the product registration card that is delivered to you as part of your Linux operating system order.

If you are installing virtual I/O server, Virtual I/O Server 2.2 or later is required.

There are unique considerations when running Java 1.4.2 on POWER7. For the best exploitation of the outstanding performance capabilities and most recent improvements of POWER7 technology, IBM recommends upgrading Java-based applications to Java 6 or Java 5 whenever possible. For more information, refer to the following website:

<http://www.ibm.com/developerworks/java/jdk/aix/service.html>

IBM Systems Director Version 6.2.1.2 or later is required for CHARM on POWER7.

4.5 HMC server and partition support limits

HMC Version 7.7 supports managing a maximum of 48 servers (non-Power 590/595 models) or 32 IBM Power 590/595 servers with a maximum of 1,024 partitions across the managed servers. The number of servers that each HMC can control varies by server size and complexity. Each server partition must have a physical connection to the network, and the HMC must be logically connected to each partition via the network connection. For additional details about the number of servers and LPARs supported, go to this website:

<http://www.software.ibm.com/webapp/set2/sas/f/hmc/>

4.6 Migrating from POWER6 to POWER7

Before attempting cross-platform migration, familiarize yourself with the binary compatibility statement of POWER7 with previous generations. This statement applies to the version of AIX that you might plan to install or use on your POWER7 server.

4.6.1 Migrating hardware from POWER6 and POWER6+ to POWER7

Hardware migration from POWER5 and earlier to POWER7 is currently not possible. However, it is possible to migrate your current POWER6 and POWER6+™ to a POWER7 server and vice versa. You can obtain a complete list of adapters that can be reused from the POWER6 on the POWER7 in the *IBM Power 795 Technical Overview and Introduction*, REDP-4640. You might need to review features, including CoD and PowerVM, with your sales representative before this migration.

Hardware upgrade path to an IBM Power 780

IBM will replace the following components:

- ▶ The CEC
- ▶ Dynamic device reconfiguration (DDR2) to DDR3
- ▶ Trim kits
- ▶ Enterprise enablement

Depending on your system configuration, you might not replace the following components:

- ▶ The rack
- ▶ PCIe adapters
- ▶ Cables, line cords, keyboards, and displays
- ▶ I/O drawers

Hardware upgrade path to an IBM Power 795

POWER6 machine type 9119-FHA can be migrated to POWER7 machine type 9119-FHB. The upgrade includes the replacement of the processor books and memory in the 9119-FHA CEC frame. You must reorder the CoD enablements.

Components that are not replaced

You do not replace the following components:

- ▶ The current POWER6 bulk power distribution and bulk regulator assemblies
- ▶ Bulk power regulators
- ▶ Bulk power distributors
- ▶ Some 12X PCI-X and PCI-e

The *IBM Power 795 Technical Overview and Introduction*, REDP-4640, specifies the complete list and associated feature codes.

4.6.2 Migrating the operating system from previous Power servers to POWER7

The following sections provide information about software migrations from previous Power Systems servers to POWER7 servers.

POWER6 to POWER7 migration

POWER6 TO POWER7 migration offers the following possibilities:

- ▶ Hardware migration, which is discussed in 4.6.1, “Migrating hardware from POWER6 and POWER6+ to POWER7” on page 132.
- ▶ Active and inactive migration using LPM, which is introduced in 3.2.3, “Live Application Mobility (LPM)” on page 75.
- ▶ Offline migration, which is a similar process to Migrating from POWER5 and earlier systems to POWER7.

Migrating from POWER5 and earlier systems to POWER7

In this section, an application refers to any non-operating system software, including vendor off-the-shelf, packaged applications, databases, custom-made applications, and scripts.

The AIX installation, whether Network Installation Management (NIM) or media, provides a few installation options. These options include a new and complete overwrite, preservation install, and migration. *NIM from A to Z in AIX 5L*, SG24-7296, explains these options in detail. The following examples show NIM installations and not media-based installation. The preferred I/O environment takes advantage of the PowerVM virtual I/O setup.

There is no hardware-based path to move from POWER5. Therefore, there is no *active migration*. The migration options in this section apply to POWER6/POWER7 migration. Our preferred option is *LPM*. The options that follow require that you create an LPAR. After you create the LPAR, you can perform one of these migrations:

- ▶ New and complete installation
- ▶ mksysb restore
- ▶ SAN-based migration using physical HBAs
- ▶ SAN-based migration using virtual adapters
- ▶ Alternate disk installation using SAN

New and complete overwrite installation

This option is an installation of AIX (BOS), which is often called a “new and complete overwrite”. Ensure that you have enough disk space to perform the installation. The installation media can be CD or DVD. However, NIM is the preferred and recommended method.

With the “New and complete overwrite” method, consider the following information:

- ▶ All data on the selected disk is lost.
- ▶ All customized system settings are lost. These settings might be required by the applications. Examples are custom network settings, including static routes, and file system settings, including VIO users. You must set these options again.
- ▶ All applications must be reinstalled.
- ▶ If data resided on the same disk, the data must be restored from backup media.
- ▶ If data resided on a separate volume group, recover it by importing the volume group as shown in “Example importing non-root volume group” on page 355.

Follow these steps for a new complete overwrite installation on a new server:

1. Prepare the LPAR.
2. Make disks available to the LPAR using either NPIV or vSCSI.
3. Prepare the NIM environment or boot from installation media.
4. Initiate Base Operating System (BOS) installation.
5. Start up the LPAR and select the new and complete overwrite option. The following steps are explained in *NIM from A to Z in AIX 5L*, SG24-7296:
 - a. Start the LPAR from the HMC into SMS
 - b. Select **Boot Options**.
 - c. Select **Boot install devices**.
 - d. Select **install devices**.
 - e. Choose **Option 6: Network**.
 - f. Select **Bootp**.
 - g. Choose the appropriate Network Adapter.
 - h. Select **Normal Boot mode**.
 - i. Select **Yes** to start.
 - j. Type 1 and press Enter to use this terminal as the system console.

- k. Select option **2** Change/Show Installation Settings and Install on the Base Operating System installation window, as shown in Figure 4-17.

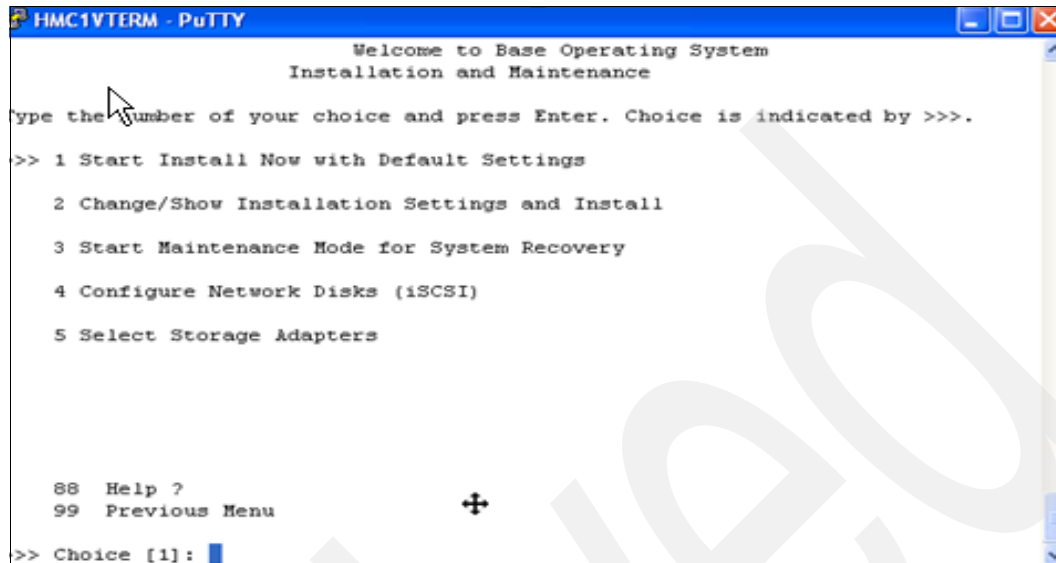


Figure 4-17 Base Operating System Installation and Maintenance window

- l. You are prompted either install with the current installation settings or make changes, as shown in Figure 4-18.

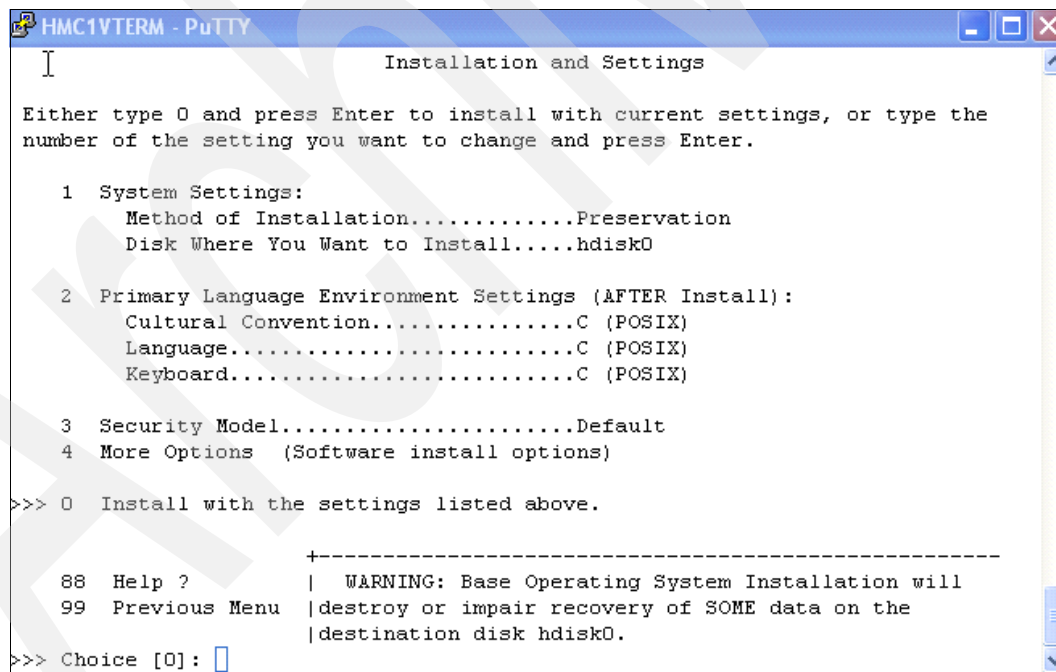


Figure 4-18 BOS Installation and Settings window

- m. Select option 1 New and Complete Overwrite. This option overwrites everything on the selected disk, as shown in Figure 4-19.

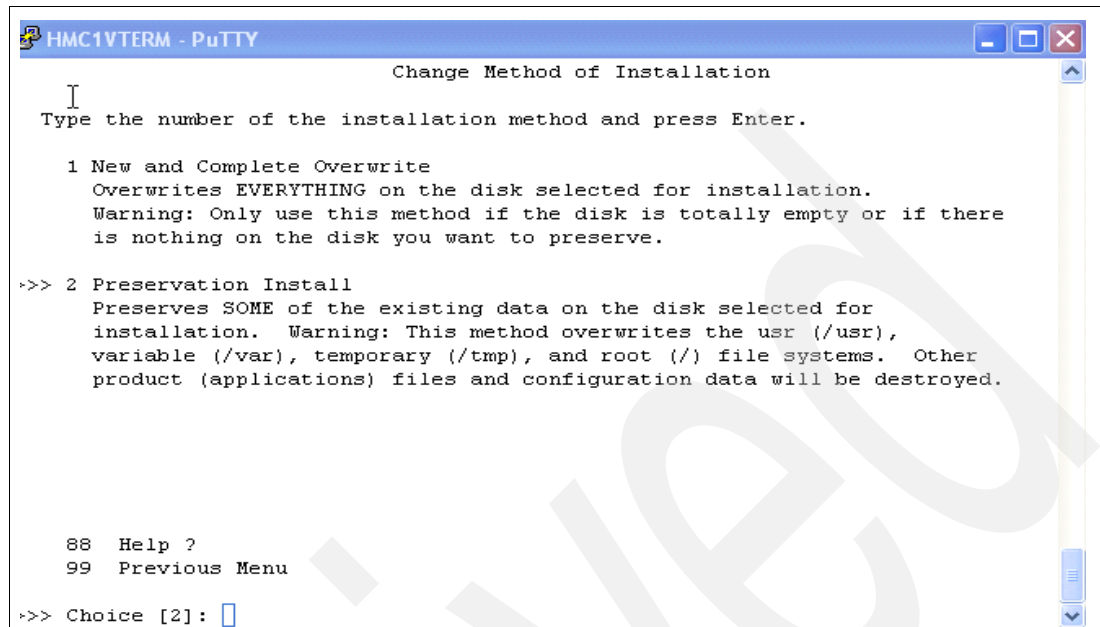


Figure 4-19 Change Method of Installation

- n. Selecting option 1 continues with the installation and overwrites the contents of the selected disk.

The `mksysb restore` command

The **`mksysb restore`** command allows you to restore operating system files and configurations. This method restores all file systems that were mounted when the backup was taken. You must not use the **`mksysb`** command as a backup strategy for non-operating system files (data). Follow these steps for using the **`mksysb restore`** procedure:

1. Prepare the LPAR.
2. Make disks available to the LPAR using either NPIV or vSCSI. Physical devices are not recommended for this volume.
3. Prepare the NIM environment:
 - If initiating a restore from installation media, you must boot either from a tape that was created using the **`mksysb`** command or from a CD/DVD that was created using the **`mkdvd`** command.
 - If a NIM server is used, you need a valid **`mksysb`** image that was created with the **`mksysb`** commands, a spot, and an **`lpp_source`**.

The `lpp_source`: The **`lpp_source`** is required, because it contains device files that might not be included in the **`mksysb`**. These devices are POWER7-specific devices that might not be available on POWER5. The required device files include virtual devices that might not have been installed on the source machine.

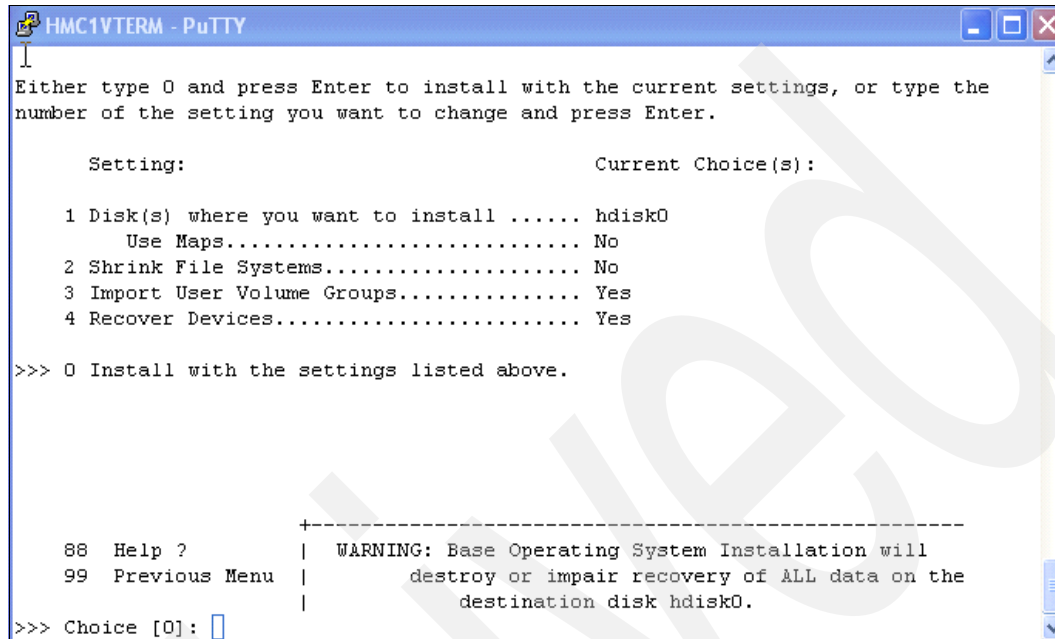
4. Initiate the BOS installation. When using NIM to initiate booting, select **mksysb - Install from a mksysb** and not **rte**, as shown in Figure 4-20.

```
+-----+-----+
|                                     Select the installation TYPE
|
| Move cursor to desired item and press Enter.
|
|   rte - Install from installation images
|   mksysb - Install from a mksysb
|   spot - Install a copy of a SPOT resource
|
| F1=Help           F2=Refresh           F3=Cancel
| F8=Image          F10=Exit              Enter=Do
| /=Find            n=Find Next
+-----+-----+
```

Figure 4-20 NIM BOS installation type selection

5. Boot the LPAR into maintenance mode and follow these steps, which help to start an LPAR using media:
 - a. Start the LPAR from the HMC into SMS.
 - b. Select **Boot Options**.
 - c. Select **Boot install devices**.
 - d. Select **install devices**.
 - e. Choose **Option 6: Network**.
 - f. Select **Bootp**.
 - g. Choose the appropriate Network Adapter.
 - h. Select **Normal Boot mode**.
 - i. Select **Yes** to start.
 - j. Type 1 and press Enter to use this terminal as the system console.

- k. On the Base Operating System installation window that is shown in Figure 4-21, select Option **2 Shrink File Systems**. Figure 4-21 shows that there is no preservation or migration option. There is only an option to select a target disk. After the target disk is selected, the mksysb restore process continues.



```
HMC1VTERM - PuTTY
Either type 0 and press Enter to install with the current settings, or type the
number of the setting you want to change and press Enter.

Setting:                                Current Choice(s):

1 Disk(s) where you want to install ..... hdisk0
  Use Maps..... No
2 Shrink File Systems..... No
3 Import User Volume Groups..... Yes
4 Recover Devices..... Yes

>>> 0 Install with the settings listed above.

88 Help ? | +-----+
89 Previous Menu | | WARNING: Base Operating System Installation will
                | | destroy or impair recovery of ALL data on the
                | | destination disk hdisk0.
                | |
>>> Choice [0]: [ ]
```

Figure 4-21 Selecting a disk on which to restore an mksysb

Consider the following information when using the **mksysb restore** command:

- ▶ All data on the selected target disk is lost.
- ▶ Customized system settings and parameters are restored from the mksysb image.
- ▶ All application binaries and data residing on the mksysb disk are restored, with the exception of any directories and subdirectories that are listed in the `/etc/exclude.rootvg` file.
- ▶ If there is any data residing on a separate volume group, recover it by importing the volume group, as shown in “Example importing non-root volume group” on page 355. This method is a safer option than a new and complete overwrite. The process is still cumbersome.

4.6.3 Disk-based migrations

The following methods remove the need to install or restore the operating system or the application data. The operating system and the data is “Taken” as it is by pointing physical volumes to host-based adapters (HBAs) on the POWER7 servers. Your storage team must be involved in this process.

4.6.4 SAN-based migration with physical adapters

In this method, HBAs can be allocated directly to an LPAR. We are not using the virtual I/O server. Although we mention this method, we recommend that you use virtual I/O server, which is part of the PowerVM virtualization offering. Follow these steps:

1. Identify the disks that are allocated to the LPAR on the POWER5 server:

- Many commands exist to identify which disks are connected to the LPAR. Most of these commands are vendor-based multipath software commands, such as **pcmpath query device**. Other commands are AIX commands and VIO commands. Although Example 4-2 shows two commands that can be used to get the serial numbers of the disks that must be zoned or mapped to the target system, there are a number of available commands, depending on the installed devices and drivers.

Example 4-2 Using pcmpath and lsattr -El to identify a LUN serial

```
# pcmpath query device 1
```

```
DEV#: 1 DEVICE NAME: hdisk1 TYPE: 1814 ALGORITHM: Load Balance
SERIAL: 600A0B800026B28200007ADC4DD13BD8
```

```
=====
Path#    Adapter/Path Name    State    Mode    Select    Errors
  0      fscsi0/path0        CLOSE    NORMAL    0         0
  1      fscsi2/path2        CLOSE    NORMAL    0         0
  2      fscsi0/path1        CLOSE    NORMAL   162       0
  3      fscsi2/path3        CLOSE    NORMAL   144       0
```

Using lsattr

```
# lsattr -El hdisk1 -a unique_id
```

```
unique_id 3E213600A0B800026B28200007ADC4DD13BD80F1814 FASTT03IBMfcpc PCM False
```

- Check that none of the disks are internal disks. If any of the disks are internal, you must either replace them with SAN-attached disks, or you must migrate them to other disks on the same volume group using either the **migratepv** or **replacepv** command. *PowerVM Migration from Physical to Virtual Storage*, SG24-7825, explains other options of migrations from physical to virtual. Use the **lsvg** and **lsdev** commands to confirm if there are any internal disks that are allocated on a volume group. Example 4-3 shows a root volume group with internal disks. These disks must be migrated to SAN storage.

Example 4-3 A rootvg with internal disks

```
lsvg -p rootvg
```

```
root@nimres1 / # lsvg -p rootvg
```

```
rootvg:
```

```
PV_NAME      PV STATE      TOTAL PPs    FREE PPs      FREE DISTRIBUTION
hdisk0       active        546          11            00..00..00..00..11
```

```
root@nimres1 / #
```

Notice that hdisk0 is allocated to rootvg. lsdev -Cc disk shows hdisk0 is physically attached to the server.

```
root@nimres1 / # lsdev -Cc disk
```

```
hdisk0 Available 04-08-00-3,0 16 Bit LVD SCSI Disk Drive
hdisk1 Available 04-08-00-4,0 16 Bit LVD SCSI Disk Drive
hdisk2 Available 04-08-00-5,0 16 Bit LVD SCSI Disk Drive
hdisk3 Available 04-08-00-8,0 16 Bit LVD SCSI Disk Drive
```

- Example 4-4 shows a root volume group using SAN-attached disks. The rootvg on Example 4-3 cannot be “zoned” to the POWER7, because it is internal to the POWER5 hardware.

Example 4-4 A rootvg with SAN-attached disks

```
# lsvg -p rootvg
```

```
rootvg:
```

PV_NAME	PV STATE	TOTAL PPs	FREE PPs	FREE DISTRIBUTION
hdisk0	active	79	2	00..00..00..00..02

#

Notice that hdisk0 is allocated to rootvg. lsdev -Cc disk shows hdisk0 is a Multi Path I/O device (mpio)

```
# lsdev -Cc disk
hdisk0 Available 02-00-02 IBM MPI0 DS4700 Array Disk
hdisk1 Available 02-00-02 IBM MPI0 DS4700 Array Disk
hdisk2 Available 02-00-02 IBM MPI0 DS4700 Array Disk
hdisk3 Available 02-00-02 IBM MPI0 DS4700 Array Disk
hdisk4 Available 02-00-02 IBM MPI0 DS4700 Array Disk
hdisk5 Available 02-00-02 IBM MPI0 DS4700 Array Disk
hdisk6 Available 02-00-02 IBM MPI0 DS4700 Array Disk
hdisk7 Available 02-00-02 IBM MPI0 DS4700 Array Disk
hdisk8 Available 00-00-02 IBM MPI0 DS4700 Array Disk
hdisk9 Available 00-00-02 IBM MPI0 DS4700 Array Disk
```

2. Prepare the LPAR, as shown in 2.10.1, “Creating a simple LPAR” on page 52.
3. Make disks available to the LPAR using either NPIV or vSCSI. 2.7.2, “N_Port ID Virtualization (NPIV)” on page 43 shows the process.
4. Shut down the LPAR on the source server.

Note: If you do not shut down the LPAR on the source server before starting it, the operating system does not start. The file systems, including the rootvg file systems, will be corrupted. This situation creates a code 553 or 557 on the destination server and file system corruptions on the source. You will have to restore from mksysb. This condition does not show immediately, but it shows when the server is rebooted, as shown in Figure 4-22. The IPL stops at this point. The HMC or SDMC displays the error code. The first diagram shows the open console. The IPL stops at the window that is shown in Figure 4-22. Figure 4-23 on page 141 shows the error code.

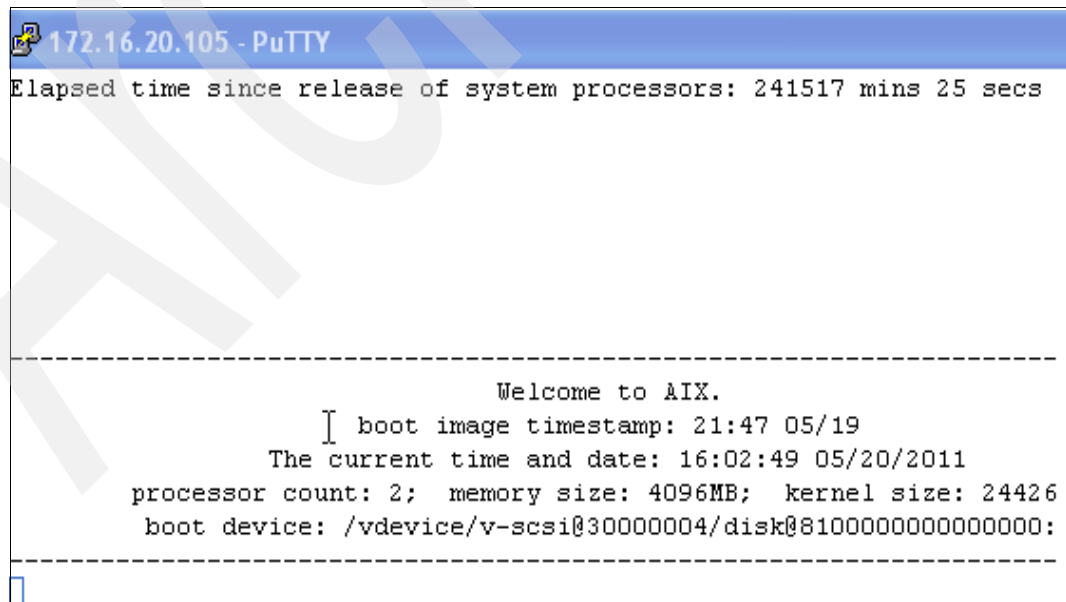


Figure 4-22 Stage where IPL stops due to corrupted root file systems and boot device

https://172.16.20.105/hmc/connects/mainuiFrameSet.jsp

Hardware Management Console

hscroot | Help |

Systems Management > Servers > p570_170

Filter Tasks Views

Select	Name ^	ID ^	Status ^	Processing Units ^	Memory (GB) ^	Active Profile ^	Environment ^	Reference Code ^
<input type="checkbox"/>	570_1_VIO_1	1	Running	0.2	4	default	Virtual I/O Server	
<input type="checkbox"/>	lpar1	6	Running	0.3	4	default	AIX or Linux	0557
<input type="checkbox"/>	lpar2	7	Running	0.3	4	default	AIX or Linux	
<input type="checkbox"/>	miglp1	4	Running	0.4	8	miglp1	AIX or Linux	

Figure 4-23 Code caused by corrupt file system or boot device

5. Start the LPAR on the POWER7 in system management services (SMS) by using either NIM rte or media. Go into system maintenance mode by selecting option **3 Start Maintenance Mode for System Recovery** on the Base Operating System Installation and Maintenance menu, as shown on Figure 4-24. Refer to *NIM from A to Z in AIX 5L*, SG24-7296, which shows how to prepare the NIM server to get to the Base Operating System Installation and Maintenance menu.

```

172.16.20.105 - PuTTY
Welcome to Base Operating System
Installation and Maintenance

Type the number of your choice and press Enter. Choice is indicated by >>>.
>>> 1 Start Install Now with Default Settings
    2 Change/Show Installation Settings and Install
    3 Start Maintenance Mode for System Recovery
    4 Configure Network Disks (iSCSI)
    5 Select Storage Adapters

    88 Help ?
    99 Previous Menu

>>> Choice [1]: 

```

Figure 4-24 Selecting the option to start an LPAR in maintenance mode

6. Select option **1** Access a Root Volume Group, as shown in Figure 4-25.

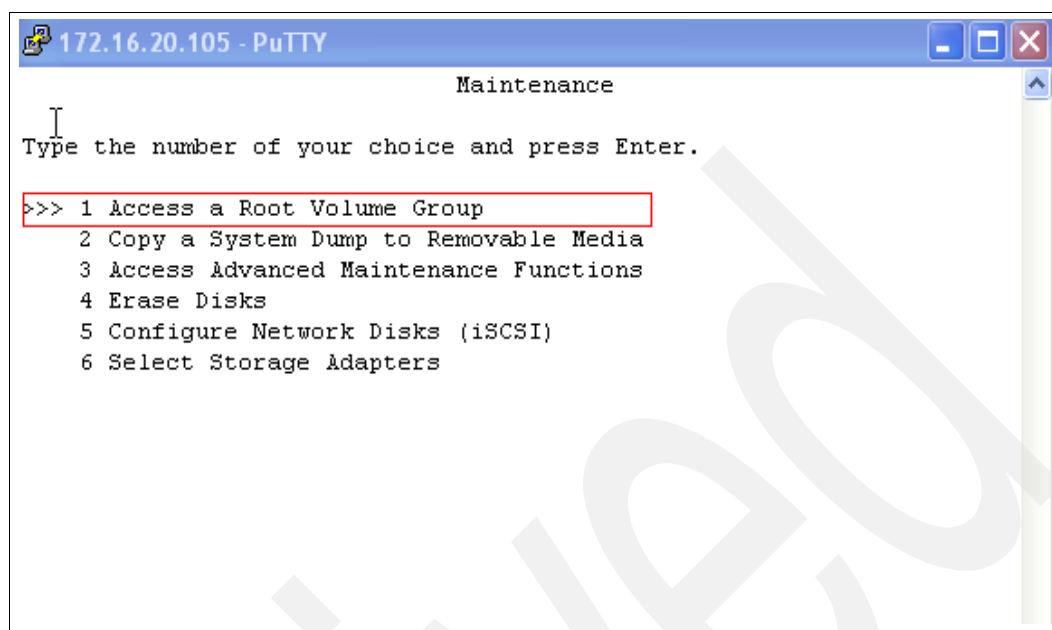


Figure 4-25 Accessing a root volume group for system maintenance

7. Select **0** to continue when prompted.
8. Select the disk that contains the root volume group.
9. Select the **Access the volume group and start the shell** option. You have access to the disk. The controlling image is the spot from NIM, not the operating system on your server. When you enter the `df` command, the command shows which file systems are mounted. The NIM file systems are also displayed. See Figure 4-26.

```
df
```

filesystem	512-blocks	Free	%Used	Iused	%Iused	Mounted on
dev/ram0	393216	71064	82%	11926	56%	/
nimres1:/nimrepo/spot/spot6103/usr				-	-	- /SPOT/usr
nimres1:/nimrepo/lpp_source/aix6103				-	-	- /SPOT/usr/sys/ins
.images						
proc	393216	71064	82%	11926	56%	/proc
dev/hd4	393216	71064	82%	11926	56%	/
dev/hd2	3735552	247072	94%	34250	53%	/usr
dev/hd3	163840	160544	3%	34	1%	/tmp
dev/hd9var	491520	137520	73%	6496	29%	/var
dev/hd10opt	720896	242064	67%	8143	23%	/opt

Figure 4-26 RAMFS file systems in system maintenance mode

10. After you are in system maintenance, run the following commands (the output is shown in Example 4-5 on page 143):
 - a. Run the `cfgmgr` command for the Device configuration manager.
 - b. Run the `bosboot` command to recreate the boot image.
 - c. Run `bootlist` to confirm the bootlist.

Example 4-5 Configuring POWER7 devices that might not be on the boot image on disk

```
# cfgmgr
# bosboot -ad /dev/hdisk0

bosboot: Boot image is 29083 512 byte blocks.
# bootlist -m normal -o
ent0 bserver=172.16.20.40 client=172.16.21.35 gateway=172.16.20.40
ent1 bserver=172.16.20.40 client=172.16.21.35 gateway=172.16.20.40
hdisk0 blv=hd5
# bootlist -m normal hdisk0
```

11. After the **cfgmgr** command completes successfully, and the **bosboot** command completes without a failure, restart the LPAR by running **shutdown -Fr**.

Alternate disk installation (alt_disk_clone)

This method is similar to 4.6.4, “SAN-based migration with physical adapters” on page 138 with the following differences:

- ▶ With **alt_disk_clone**, you clone the operating system to an alternate disk before making the disk available to the destination server.
- ▶ You can allocate either the original disk or the cloned disk to the target server.
- ▶ The added safety of alternate disk is that you can return to the original server in its original state, and the operating system has no added drivers and filesets.

After the **alt_disk_copy** is completed and cleaned, the alternate disk must be removed from the source and allocated to the target server. You can follow the process in 4.6.4, “SAN-based migration with physical adapters” on page 138.

Example 4-6 shows the commands to create a clone.

Example 4-6 Commands showing how to create a clone

```
# hostname
rflpar20
# lspv
hdisk0          00c1f170c2c44e75          rootvg          active
hdisk1          00f69af6dbccc5ed          None
hdisk2          00f69af6dbccc57f          datavg
# alt_disk_copy -d hdisk1
Calling mkoszfile to create new /image.data file.
Checking disk sizes.
Creating cloned rootvg volume group and associated logical volumes.
Creating logical volume alt_hd5
Creating logical volume alt_hd6
Creating logical volume alt_hd8
Creating logical volume alt_hd4
Creating logical volume alt_hd2
Creating logical volume alt_hd9var
Creating logical volume alt_hd3
Creating logical volume alt_hd1
Creating logical volume alt_hd10opt
Creating logical volume alt_hd11admin
Creating logical volume alt_lg_dumplv
Creating logical volume alt_livedump
Creating /alt_inst/ file system.
```

```

/alt_inst filesystem not converted.
    Small inode extents are already enabled.
Creating /alt_inst/admin file system.
/alt_inst/admin filesystem not converted.
    Small inode extents are already enabled.
Creating /alt_inst/home file system.
/alt_inst/home filesystem not converted.
    Small inode extents are already enabled.
Creating /alt_inst/opt file system.
/alt_inst/opt filesystem not converted.
    Small inode extents are already enabled.
Creating /alt_inst/tmp file system.
/alt_inst/tmp filesystem not converted.
    Small inode extents are already enabled.
Creating /alt_inst/usr file system.
/alt_inst/usr filesystem not converted.
    Small inode extents are already enabled.
Creating /alt_inst/var file system.
/alt_inst/var filesystem not converted.
    Small inode extents are already enabled.
Creating /alt_inst/var/adm/ras/livedump file system.
/alt_inst/var/adm/ras/livedump filesystem not converted.
    Small inode extents are already enabled.
Generating a list of files
for backup and restore into the alternate file system...
Backing-up the rootvg files and restoring them to the
alternate file system...
Modifying ODM on cloned disk.
Building boot image on cloned disk.
forced unmount of /alt_inst/var/adm/ras/livedump
forced unmount of /alt_inst/var/adm/ras/livedump
forced unmount of /alt_inst/var
forced unmount of /alt_inst/var
forced unmount of /alt_inst/usr
forced unmount of /alt_inst/usr
forced unmount of /alt_inst/tmp
forced unmount of /alt_inst/tmp
forced unmount of /alt_inst/opt
forced unmount of /alt_inst/opt
forced unmount of /alt_inst/home
forced unmount of /alt_inst/home
forced unmount of /alt_inst/admin
forced unmount of /alt_inst/admin
forced unmount of /alt_inst
forced unmount of /alt_inst
Changing logical volume names in volume group descriptor area.
Fixing LV control blocks...
Fixing file system superblocks...
Bootlist is set to the boot disk: hdisk1 blv=hd5
# bootlist -m normal -o
hdisk1 blv=hd5
# bootlist -m normal hdisk0

```

The size and contents of your disks affect the time that it takes for the alternate disk to complete.

VIO server-based migration using virtual adapters

This method requires the creation of Virtual SCSI or Virtual Fibre. Refer to “Creating Virtual FC adapters” on page 229. After allocating the disks, follow the process that is described in 4.6.4, “SAN-based migration with physical adapters” on page 138.

Supported device drivers: In all the migration procedures that we have discussed, check that you have supported device drivers. One way to resolve this issue if you do not is to install all device support with the Base Operating System installation. This method requires disk space, makes the installation longer, and makes the upgrades longer. If you do not have all the supported devices, LPM might not work.

4.6.5 After migration to POWER7

Review the AIX prerequisites for running POWER7. The version of AIX has an effect on the mode in which your POWER7 server runs. To take full advantage of the POWER7 features, upgrade AIX to Version 7.1. Consult your application vendors to confirm compatibility. Also, refer to *Exploiting IBM AIX Workload Partitions*, SG24-7599, for migrating an AIX Version 5.2 LPAR. This Redbooks publication shows the creation of a Versioned WPAR, which can run AIX Version 5.2 TL 8 and later.

Reasons to consider running the latest AIX version:

- ▶ AIX 5.3: With 5300-09 TL and service pack 7, or later, the LPAR only runs in POWER6 or POWER6+ mode. You are not able to run smt2. Thus, you cannot run four threads per core as designed for POWER7.
- ▶ AIX Version 6.1: Prior to TL 6, the LPAR mode was POWER6 or POWER6+.
- ▶ AIX Version 6.1 TL 6 and later: The LPAR runs in POWER7 mode, but it is limited to 64 cores.
- ▶ AIX Version 7.1: It exploits all the capabilities of the POWER7 architecture.

Table 4-8 on page 146 from the *IBM Power 770 and 780 Technical Overview and Introduction*, REDP-4639, shows the benefits that you can derive from running in POWER7 mode compared to POWER6.

Table 4-8 Benefits of running POWER7 mode

POWER6 (and POWER6+) mode	POWER7 mode	Client value
Two-thread SMT	Four-thread SMT	Throughput performance, processor core utilization
Vector Multimedia Extension (VME)/Altivec	Vector Scalar Extension (VSX)	High-performance computing
Affinity OFF by default	Three-tier memory, Micro-partition Affinity	Improved system performance for system images spanning sockets and nodes
<ul style="list-style-type: none"> ▶ Barrier Synchronization ▶ Fixed 128-byte Array; Kernel Extension Access 	<ul style="list-style-type: none"> ▶ Enhanced Barrier Synchronization ▶ Variable-Sized Array; User Shared Memory Access 	High-performance computing, parallel programming, synchronization facility
64-core and 128-thread scaling	32-core and 128-thread scaling 64-core and 256-thread scaling 256-core and 1,024-thread scaling	Performance and scalability for large scale-up single system image workloads, such as online transaction processing (OLTP), ERP scale-up, and WPAR consolidation
EnergyScale CPU Idle	EnergyScale CPU Idle and Folding with NAP and SLEEP	Improved energy efficiency

You can set up an LPAR running AIX Version 6 TL 6 and AIX Version 7.1 to an earlier processor mode using the LPAR Profile Processors tab. We recommend that you leave the option as the default. The processor mode changes based on the operating system level.

The next sections discuss the LPAR mode and provide examples.

Notice on the Processing Settings tab, when creating an LPAR, there is no option to choose which processor mode to use. See Figure 4-27.

Processing Settings

Specify the desired, minimum, and maximum processing settings in the fields below.

Total usable processing units: 4.00

Minimum processing units * 0.1

Desired processing units * 0.1

Maximum processing units * 0.1

Shared processor pool: DefaultPool (0)

Virtual processors

Minimum processing units required for each virtual processor: 0.10

Minimum virtual processors * 1

Desired virtual processors * 1

Maximum virtual processors * 1

☐ Uncapped

Weight : 128.0

Figure 4-27 Initial creation of an LPAR: No processor mode option

After the LPAR creation, you can change the processor mode on the LPAR when you are on the POWER7 system. Follow these steps:

1. Log on to the **SDMC**.
2. Select **hosts**.
3. Select **Virtual Server** → **Action** → **System Configuration** → **Manage Profile**.

These choices are shown in Figure 4-28.

In the following figures, we show you how to use an HMC to change the system mode. We explain how to use an SDMC in detail in 5.1, “SDMC features” on page 158.

4. Log in to the HMC.
5. Select **System Management**.
6. Select the system. This option lists a few LPARs.
7. Click on the LPAR.
8. Select **Tasks** → **Configuration** → **Manage Profiles**. The window that is shown in Figure 4-28 opens.

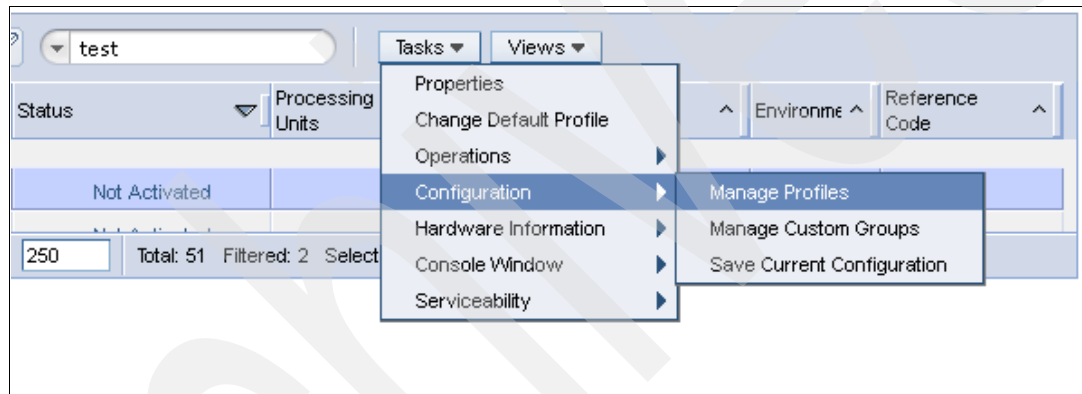


Figure 4-28 Editing a virtual server profile

9. Select the profile that you need to edit. Select the **Processor** tab. On the Processor tab, select the appropriate mode on the Processor compatibility mode list box, as shown in Figure 4-29 on page 149.

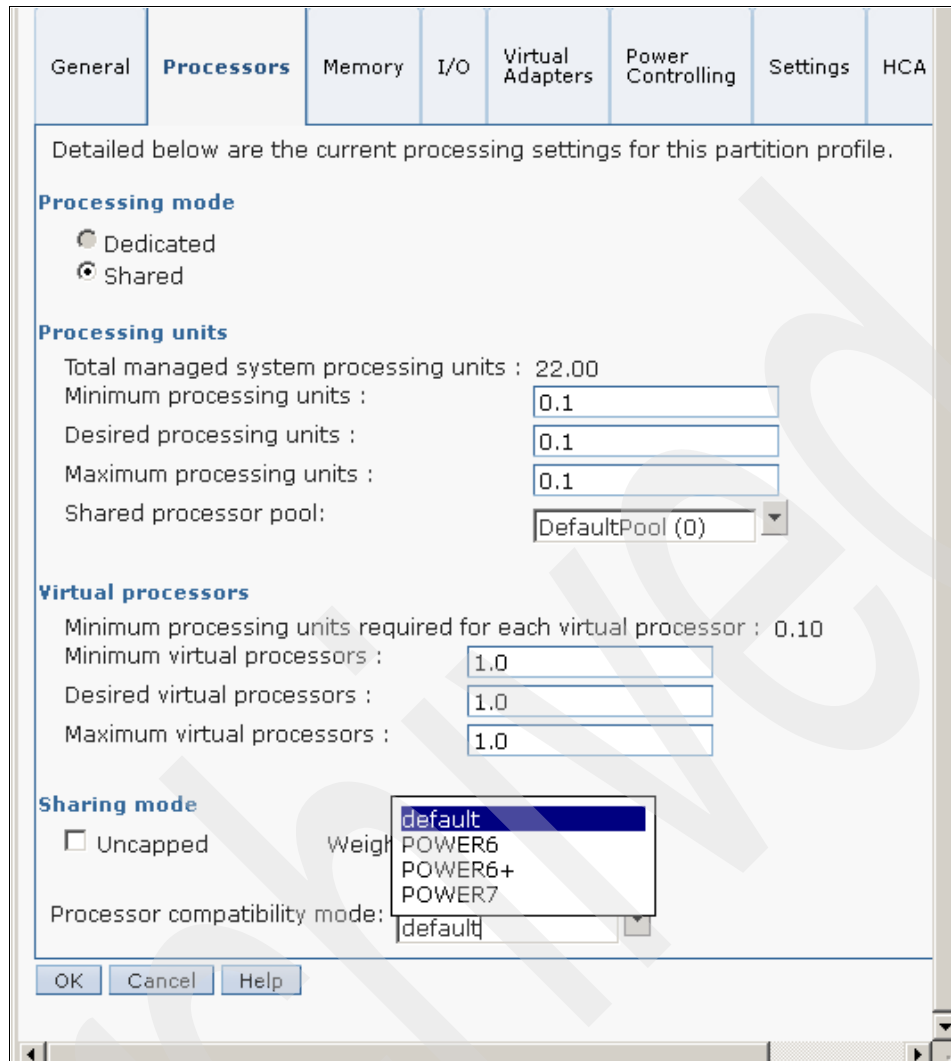


Figure 4-29 POWER7 Processor compatibility Mode

Example 4-7 shows the `lsconf` command, which allows you to see the processor mode of a running AIX LPAR.

Example 4-7 Showing the processor mode of an AIX LPAR

```
# lsconf | head
System Model: IBM,9117-MMA
Machine Serial Number: 101F170
Processor Type: PowerPC_POWER6
Processor Implementation Mode: POWER 6
Processor Version: PV_6_Compat
Number Of Processors: 2
Processor Clock Speed: 4208 MHz
CPU Type: 64-bit
Kernel Type: 64-bit
LPAR Info: 6 lpar2_570
```

4.7 Technical and Delivery Assessment (TDA)

IBM is continually striving to improve worldwide solution quality. The Technical and Delivery Assessment (TDA) is an objective, third-party technical expert inspection of your completed Power System solution design to answer three questions:

- ▶ Will your solution work?
- ▶ Is IBM prepared to implement it successfully?
- ▶ Will your proposed solution meet your requirements and expectations?

The Power 780 and Power 795 servers require a mandatory pre-installation TDA before the order ships. The pre-installation TDA is designed to evaluate your (the client) readiness to install, implement, and support your new Power System solution. In addition, the TDA is designed to minimize installation problems, minimize IBM and IBM Business Partner costs to support, and document the actions that are required for success.

We want to ensure that you are getting the correct solution to meet your business requirements. This process uses IBM technical support to provide expert skills to identify activities that are required for a successful solution implementation.

Our solution assurance works because we have designed the process over time using actual client experiences. The preparation of the TDA document is an exercise that can be extremely revealing. It pulls the entire solution together from one central view and can reveal if there are missing components. During the TDA review, experts review the overall solution and help identify what might have been overlooked in the design. It provides many perspectives with a single consistent approach.

When IBM conducts your TDA review, all appropriate team members are invited (IBM, IBM Business Partner, and client) and required to attend. The pre-installation review needs to be completed one to two weeks before the Power system ships, or earlier if significant porting, moving, or site preparation tasks are required. During this review, the experts discuss the following items:

- ▶ Power requirements
- ▶ Space requirements
- ▶ Cabling requirements
- ▶ Installation plan and responsibilities
- ▶ Upgrade plan and responsibilities
- ▶ Services and support

There are three possible outcomes to a TDA. The solution stage assessments are “passed”, “passed with contingency on action items”, and “not recommended”.

Passed

If the subject matter experts (SMEs) approve the solution as presented, the proposed design and solution proceeds. A result of “Passed” does not mean that there are no outstanding action items, there might be many outstanding action items. However, the outstanding action items associated with a review that receives this rating must have a predictable outcome that does not alter the viability of the overall solution. For instance, an action item might be to review options that you as the client have for maintenance offerings. This item can be performed, and the outcome does not alter the nature of the solution technically.

Passed with contingency on action items

This outcome is a conditional approval that depends on the results of certain specified action items. For example, suppose that a certain version of an application is an absolute prerequisite to support your proposed Power server, but it was not known whether that

version was actually available. The reviewers might elect to approve contingent on it being verified that the required version can be installed or upgraded to the required release.

A contingency on action item differs from an ordinary action item in that its outcome is uncertain, yet critical to the viability of the proposed solution. In the case of a “Passed With Contingency on Action Items” result, your IBM sales team must take steps to execute the contingent action items and ensure that the outcomes are the ones needed to satisfy the TDA conditions.

Not recommended

This result means that the reviewers do not agree that the solution is technically viable. A solution might be “Not recommended” due to the lack of sufficiently detailed information to evaluate the solution. The “Not recommended” result does not occur often. We list several reasons that reviewers might conclude that a solution is “Not recommended”:

- ▶ The solution, as presented, fails to meet the requirements articulated by the client and cannot be rescued with minor adjustments.
- ▶ The solution presenter cannot provide sufficient information to allow the reviewers to judge the technical viability of the solution.
- ▶ The technical risk that is associated with the solution is unreasonably high.

Your TDA review must document two solution risk assessment ratings:

- ▶ **Before Action Items Are Completed:** Risk assessment is for the solution “as is”, at the time of the TDA.
- ▶ **After Action Items Are Completed:** Risk assessment is for the solution with the assumption that all recommended action items that result from the review are completed on schedule.

There are three risk assessment levels:

- ▶ High
- ▶ Medium
- ▶ Low

The members of your IBM sales and technical sales team need to ensure that all action items are completed correctly.

4.8 System Planning Tool (SPT)

The System Planning Tool (SPT) is a browser-based application that helps you design your system configurations and is particularly useful for designing logically partitioned systems. It is available to assist in the design of an LPAR system and to provide an LPAR validation report that reflects your system requirements while not exceeding IBM’s LPAR recommendations. It is also used to provide input for your specified hardware placement requirements. System plans that are generated by the SPT can be deployed on the system by the HMC, SDMC, and the IVM. The SPT is intended to be run on the user’s personal computer, and it is provided as is with no implied or expressed warranty of any kind.

SPT: The SPT is available for download at this website:

<http://www-947.ibm.com/systems/support/tools/systemplanningtool/>

You can use the SPT to design both a logically partitioned system and a non-partitioned system. You can create an entirely new system configuration from nothing or you can create a system configuration based upon any of the following information:

- ▶ Performance data from an existing system that the new system will replace
- ▶ A performance estimate that anticipates future workload requirements
- ▶ Sample systems that you can customize to fit your needs

After designing a system with SPT, you can generate the following information:

- ▶ Reports that detail the system configuration that you have architected
- ▶ System-plan files that can be moved to the HMC, SDMC, or IVM that are used to actually deploy your system plan

SPT uses a file format called `.sysplan`, which is used on your management console to systematically distribute your system plan. The `.sysplan` file can be renamed to `.zip` and an XML file can be extracted for possible manipulation outside of the SPT tool and the HMC or SDMC.

For managed systems with virtual I/O server installed, the HMC code must be at 7.3.3 (or greater) and the virtual I/O server must be at Fix Pack 10.1 (or greater) to generate SPT files from your HMC.

You can review the following reports from the SPT viewer:

- ▶ Partition-wise processor summary report
- ▶ Partition-wise memory summary report
- ▶ Virtual SCSI server-client slot mappings
- ▶ Virtual FC server-client slot mappings
- ▶ Verify dual virtual I/O server configurations for preferred practices

We highly suggest that you create a system plan using the SPT before and after any hardware changes are made. Additionally, any major changes or new systems need to be built in SPT before an order is placed to ensure their validity.

To use the HMC or SDMC to create a system plan successfully, you need to ensure that your system meets a number of prerequisite conditions.

A system plan that you create by using HMC V7.3.3 or later, or the SDMC V6.2.1.2 or later, contains hardware information that the management console was able to obtain from your selected managed system. However, the amount of hardware information that can be captured for the system plan varies based on the method that was used to gather the hardware information.

The management console can potentially use two methods: inventory gathering and hardware discovery. When using hardware discovery, the HMC/SDMC can detect information about hardware that is unassigned to a partition or that is assigned to an inactive partition. Additionally, the HMC/SDMC can use one or both of these methods to detect disk information for IBM i LPARs. You will collect better quality data and a more accurate quantity of data for the system plans if you use the hardware discovery process.

The IBM POWER7 information center gives the detailed requirements for both inventory gathering and hardware discovery at this website:

<http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp>

To create a system plan by using the HMC, complete the following steps:

1. In the navigation area, select **System Plans**. The System Plans page opens.

2. In the Tasks area, select **Create System Plan**. The Create System Plan window opens.
3. Select the managed system that you want to use as the basis for the new system plan.
4. Enter a name and description for the new system plan.
5. Optional: Select whether you want to retrieve inactive and unallocated hardware resources. This option appears only if the managed system is capable of hardware discovery, and the option is selected by default.

Important: If you do not select the “Retrieve inactive and unallocated hardware resources” option, the HMC does not perform a new hardware discovery, but instead uses the data in the inventory cache on the system. The HMC still performs inventory gathering and retrieves hardware information for any active LPARs on the managed server. The resulting new system plan contains hardware information from the inventory-gathering process and hardware information from the hardware inventory cache on the system.

6. Optional: Select whether you want to view the system plan immediately after the HMC creates it.
7. Click **Create**.

Figure 4-30 show the HMC panels to create your Sysplan.

Figure 4-30 Creating the system plan pane

Now that you have created a new system plan, you can export the system plan, import it onto another managed system, and deploy the system plan to that managed system.

Creating a system plan: As an alternative to the HMC web-user interface, you can use the following methods to create a system plan that is based on the configuration of an existing managed system.

There are several methods to create a system plan:

- ▶ Run the `mksysplan` command from the HMC command-line interface (CLI).
- ▶ Run the `mksysplan` command from the SDMC CLI.
- ▶ Use the SDMC web user interface.

The POWER7 Enterprise Servers support the Customer Specified Placement (CSP) of I/O adapters and I/O devices within the CEC and I/O drawers. Through the use of the CSP feature, IBM Manufacturing can provide customization of your Power server order to match your hardware placement request according to the slot in the drawer hardware placement, before the server arrives at your site. We strongly advise that you use CSP for all Power 780 and 795 orders.

Without CSP, IBM Manufacturing makes an effort to distribute adapters evenly across busses, planars, and drawers. However, this default placement might not be optimum for your specific performance, availability, or LPAR connectivity requirements.

CSP specifications are collected using the SPT and processed through eConfig, or placement requirements can be specified directly in eConfig using the Placement view. An advantage of using SPT is that it allows the CSP information to be copied and preserved.

CSP requires your IBM account team to submit the `cfsreport` output of eConfig to IBM Manufacturing in a timely manner (within 24 hours) via the CSP website. It also requires your account team to assure that the eConfig output submitted reflects the actual order placed.

We strongly advise that you create a system plan using the SPT before and after any changes are made to existing hardware configuration. Additionally, any major changes or new systems need to be built in SPT before an order is placed to ensure that the changes or new systems are valid.

Disaster recovery planning: The SPT is also an excellent tool for documentation and needs to be included as input into your disaster recovery plan.

4.9 General planning guidelines for highly available systems

For a highly available operating environment that takes advantage of reduced planned and unplanned outages, planning is important. Plan toward eliminating single points of failure (SPOFs) within a single system or cluster of interconnected systems that support an application or applications. The *IBM PowerHA SystemMirror Planning Guide*, SC23-6758, suggests the following considerations when eliminating SPOFs.

Considerations within a single managed system

The following considerations help eliminate SPOFs in a single server system:

- ▶ Power source: Use multiple circuits or uninterruptible power supplies.
- ▶ Networks: Use multiple networks to connect nodes. Use redundant network adapters.
- ▶ TCP/IP subsystems: Use as many TCP/IP subsystems as required to connect to users.
- ▶ Disk adapters: Use redundant disk adapters.
- ▶ Controllers: Use redundant disk controllers.
- ▶ Disks: Use redundant hardware and disk mirroring.
- ▶ Cluster repository: Use RAID protection.
- ▶ Virtual I/O server: Use redundant VIO servers.
- ▶ System management: Use redundant HMCs, SDMCs, or a combination of HMCs and SDMCs.

Considerations for enhancing availability

The following considerations help enhance availability:

- ▶ **Nodes:** We suggest that you use multiple physical nodes.
- ▶ **Applications:** Use clustering, such as PowerHA, Cluster Aware AIX (CAA), or high availability disaster recovery (HADR). You need to assign additional nodes for the takeover.
- ▶ **Mobility:** Use either Live Application Mobility (LAM) or Live Partition Mobility (LPM).
- ▶ **Sites:** Use more than one site. Also, disaster recovery uses multiple sites.

You must complement all planning and implementations with testing. Skipping the planning stage can result in infrequent, high-impact errors occurring. The more scenarios that you can test assist in building resilience around the solutions that are provided.

Together with planning and testing comes training. Users need to be trained both on the job and formally to be able to take advantage of the features that are provided. Clients need to plan for the components on a single system that might have to be repaired.

Archived

POWER7 system management consoles

This section explores the POWER Server management console solutions through the Hardware Management Console (HMC), IBM Systems Director Management Console (SDMC), and the IBM Systems Director console.

We discuss the following topics in this section:

- ▶ SDMC features
- ▶ Virtualization management: Systems Director VMControl
- ▶ IBM Systems Director Active Energy Management (AEM)
- ▶ High availability Systems Director management consoles

5.1 SDMC features

The SDMC is designed to be a successor to both the Hardware Management Console (HMC) and the Integrated Virtualization Manager (IVM) for Power Systems administration. The Power Systems management is integrated into the Systems Director framework, which allows for the management of many systems of various types.

5.1.1 Installing the SDMC

The SDMC installation involves the following tasks:

- ▶ The installation of the hardware appliance that is required for all midrange and high-end systems.
- ▶ The installation of the software appliance that replaces IVM.
- ▶ The use of the setup wizard at the end of the installation process to set up and perform the initial configuration of the SDMC.

The SDMC virtual machine contains Linux as the base operating system. The virtualization layer for the hardware appliance is fixed and cannot be changed. The hardware is provided by IBM and it uses the Red Hat Enterprise Virtualization hypervisor (RHEV-H hypervisor). The software appliance can be installed on either VMware or a kernel-based virtual machine (KVM), and the client supplies the hardware.

For then detailed step-by-step installation procedure for the SDMC, see Chapter 2, “Installation” of the *IBM Systems Director Management Console Introduction and Overview*, SG24-7860, which is located at the following website:

<http://www.redbooks.ibm.com/redbooks/pdfs/sg247860.pdf>

5.1.2 SDMC transition

Although the move from the HMC to the SDMC might at first seem daunting, the SDMC has been designed to allow for as smooth a transition as possible. First, you can run the HMC and the SDMC in parallel, co-managing the same hardware, during the transition period. To operate in parallel, both consoles must be at the same level of code.

Section 4.3, “HMC to SDMC transition”, in the *IBM Systems Director Management Console Introduction and Overview*, SG24-7860, describes the procedure to launch the transition wizard to move a system that is managed by an HMC to the SDMC. The publication is located at the following website:

<http://www.redbooks.ibm.com/redbooks/pdfs/sg247860.pdf>

To use any of the advanced managers, such as the VMControl or Advanced Energy Manager (AEM) plug-ins, at the SDMC launch, you must have the configuration that is shown in the Figure 5-1 on page 159.

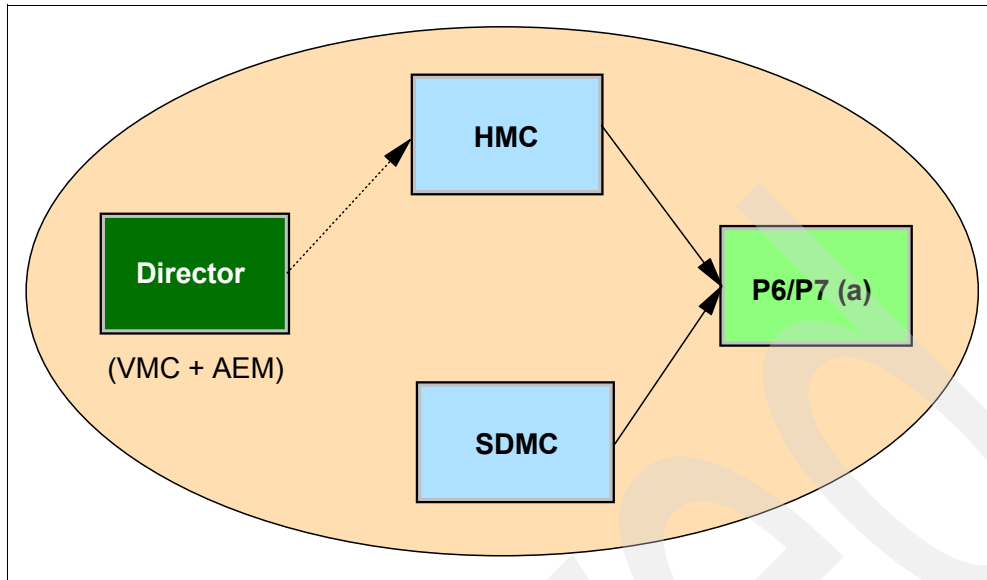


Figure 5-1 Recommended configuration

Figure 5-1 describes the parallel management of an HMC or SDMC on a single POWER6 or POWER7 frame. It also shows a hierarchical management configuration in which the Systems Director, with the advanced management plug-ins installed, uses the HMC's management interface to the server to facilitate the use of the plug-ins.

As the HMC transitions out of use, the Systems Director will be able to manage the POWER6/POWER7 either directly or hierarchically through the SDMC.

Requirements: The configuration that is shown in Figure 5-1 requires these supported levels:

- ▶ The level of the HMC is 7.3.5 or higher
- ▶ The level of the IBM Systems Director is 6.2.1.2 or higher

5.1.3 SDMC key functionalities

SDMC allows a single management point for many systems in your enterprise. The SDMC is extremely similar to the HMC. The goal in the redesign of the single point of control is to provide the Systems Director with a combined hardware and software control user experience. With SDMC, you can perform these functions:

- ▶ Manage and provision multiple systems of heterogeneous infrastructure
- ▶ Reconfigure systems by using logical partition (LPAR) and dynamic LPAR (DLPAR) capabilities
- ▶ Enable certain hardware enhancements, such as POWER6 compatibility mode, on the POWER7
- ▶ Orchestrate Live Partition Mobility (LPM) operations
- ▶ Coordinate the suspend and resume of virtual servers
- ▶ Modify the resource assignment of your virtual servers even when they are in a stopped state
- ▶ Manage virtual slots automatically, leading to enhanced virtual I/O server management

- Create users that use Lightweight Directory Access Protocol (LDAP) or Kerberos for authentication
- Back up the whole virtual machine onto removable media or to a remote FTP server
- Schedule operations for managed systems and virtual servers
- Preserve the HMC's active-active redundancy model in addition to the active-passive availability model that is provided by the Systems Director

The *IBM Systems Director Management Console Introduction and Overview*, SG24-7860, describe all these functionalities in detail. This book is located at the following website:

<http://www.redbooks.ibm.com/redbooks/pdfs/sg247860.pdf>

5.1.4 HMC versus SDMC

The SDMC represents the consolidation of several system management offerings that are currently offered by IBM. It brings together the features in the Systems Director, IVM, and HMC. Tasks that you can perform independently on any of these platforms can also conveniently be performed on the SDMC. The SDMC, which includes all the traditional server and virtualization management functions that are provided by the latest HMC, also provides the simplicity of IVM in its functions.

The HMC administers entry-level systems to high-end systems. The SDMC manages many systems of multiple types. It manages both POWER processor-based blades, systems that were previously managed by the IVM, and high-end systems. The SDMC is available as a hardware appliance similar to the HMC. The SDMC, however, unlike the HMC, is available in a virtual appliance form, as well, for installation into existing virtual machine environments. The software appliance is intended for the management of low-end and midrange servers. The hardware appliance is targeted for use with midrange to high-end servers.

Due to the integration with the Systems Director, an inherently cross-platform management server, you might notice changes in terminology. Logical partitions (LPARs) are, for example, in SDMC referred to as “*virtual servers*”, and managed systems are referred to as “*hosts*” or “*servers*”. The SDMC also demonstrates tighter integration with virtual I/O server through a more automatic management of virtual slots than the HMC. For users who prefer to use the HMC command-line interface (CLI), the CLI transitioned fairly intact, although with a few minor syntax changes. The commands of the HMC are run with a prefix of “**smcli**”.

For example, to list the virtual Small Computer System Interface (SCSI) resources of a host, prefix the HMC **lshwres** command with **smcli**, as shown:

```
sysadmin@dd172:~>smcli lshwres -r virtualio --rsubtype scsi -m
Server-8233-E8B-SN100042P --level lpar
```

The preceding command lists all virtual SCSI adapters on the managed system, Server-8233-E8B-SN100042P.

Experienced users also appreciate the enhancements to dynamic LPAR (DLPAR) management to make it more intuitive, such as the ability to modify resource allocations regardless of whether the partition is On or Off. You can modify the processor, memory, and adapter assignments for a virtual server even when it is in a stopped state.

Perform the following steps to add a virtual Ethernet adapter using a DLPAR operation on a virtual server that is in the stopped state:

1. Use the Manage Virtual Server task to change the general properties and perform dynamic logical partitioning.

For dynamic LPAR operations on the virtual server, click **Manage Virtual Server** to locate the virtual server on which the DLPAR operation will be performed. Select the virtual server, click **Add** and click **OK**, as shown in Figure 5-2.

The screenshot shows the 'Targets' page in a web interface. At the top, there's a tab labeled 'Targets'. Below it, a message says 'Select the targets on which the job will run.' The main section is titled 'Manage Virtual Server' and contains the instruction 'Select a valid target then add it to the selected list.' There's a 'Show:' dropdown menu set to 'All Targets'. Below this, there are two panels: 'Available:' and 'Selected:'. The 'Available:' panel contains a table with columns 'Select', 'Name', and 'Type'. The table lists several virtual servers, all of type 'Virtual Server'. The 'Selected:' panel shows 'danO' as the selected target. Between the two panels are buttons for 'Add >' and '< Remove'.

Targets

Select the targets on which the job will run.

Manage Virtual Server
Select a valid target then add it to the selected list.

Show: All Targets

Available:

All Targets

Select	Name	Type
<input type="radio"/>	danO	Virtual Server
<input type="radio"/>	redbook_new	Virtual Server
<input type="radio"/>	vioc1-vmobso	Virtual Server
<input type="radio"/>	vioc2-lrddbso	Virtual Server
<input type="radio"/>	vioc3-adamsbso	Virtual Server
<input type="radio"/>	VIOS-1	Virtual Server

Selected:

danO

Add >

< Remove

Figure 5-2 Page showing the Manage Virtual Server task

2. Figure 5-3 shows the next page with tabs on the left side that can be used to modify the processor, memory, and adapter assignments of the selected virtual server.

Host: SP-9117-MMA-SN1059020

Name: danO

Id: 8

Environment: AIX/Linux

State: Stopped

Tasks

General Settings

Processor

Memory

Network

Storage Adapters

Storage Devices

Media Devices

Physical IO

Virtual Ethernet

Available Virtual Slots: 4

AddRemoveProperties

Select	Adapter(Id)	PVID	Additional VLAN
<input type="checkbox"/>	(3)	1 - ent0(VIOS-1)	20
<input type="checkbox"/>	(4)	2	49
<input type="checkbox"/>	(5)	1 - ent0(VIOS-1)	20

Logical Host Ethernet Adapters

Host Ethernet Adapters can't be configured for virtual server with shared memory mode.

Figure 5-3 Shows the properties of the adapter to be created

3. Figure 5-4 shows the attributes that have been selected to create the virtual Ethernet adapter. We selected the Adapter ID 6, Port Virtual Ethernet 1, checked the IEEE 802.1q compatible adapter box, added 20 as the additional VLAN ID, and selected the default VSwitch **ETHERNET0**.

Create Virtual Ethernet Adapter

Specify an adapter ID and virtual Ethernet for this adapter.

Adapter Id
6

Port Virtual Ethernet
1

IEEE Settings
Select this option to allow additional virtual LAN IDs for the adapter.

☒ IEEE 802.1q compatible adapter*

Maximum number of VLANs: 20

Additional VLAN IDs: *
20 1,20,48,...

Shared Ethernet Settings
Select Ethernet bridging to link(bridge) the virtual Ethernet to a physical network.

☐ Use this adapter for Ethernet bridging

Priority:
(1 or 2)

Virtual Switch
VSwitch: ETHERNET0

Ok Cancel Help

Figure 5-4 Shows the attributes to add to create the adapter

Figure 5-5 shows that the virtual Ethernet adapter in slot number 6 has been added.

Host: SP-9117-MMA-SN1059020Name: dan0Id: 8Environment: AIX/LinuxState: StoppedTasks

General SettingsProcessorMemoryNetworkStorage AdaptersStorage DevicesMedia DevicesPhysical IO

Virtual Ethernet

Available Virtual Slots: 3

AddRemoveProperties

Select	Adapter(Id)	PVID	Additional VLAN
<input type="checkbox"/>	(3)	1 - ent0(VIOS-1)	20
<input type="checkbox"/>	(4)	2	49
<input type="checkbox"/>	(5)	1 - ent0(VIOS-1)	20
<input type="checkbox"/>	(6)	1	20

Logical Host Ethernet Adapters

Host Ethernet Adapters can't be configured for virtual server with shared memory mode.

Figure 5-5 Shows that the virtual Ethernet adapter has been added

There are a few features in the HMC that are not in the SDMC: the system plans feature, the management of POWER5 systems, and the capability to disconnect and reconnect to old sessions.

5.1.5 Statement of direction for support HMC

It is expected that most new users of IBM Power Systems will use the new, enhanced SDMC offering as their systems management of choice; therefore, the SDMC has been designed to support only POWER6 and higher servers.

POWER4 and POWER5: Users of POWER4 and POWER5 platforms have to use the HMC to manage their servers.

The HMC then takes on the role of an older management server. The amount of new functionality that is added to the HMC ends over the next two years, and the POWER7 server series will be the last systems to be able to be managed by the HMC.

IBM advises clients to consider adopting the SDMC in their environment in the near future.

5.2 Virtualization management: Systems Director VMControl

This section provides an overview of IBM Systems Director VMControl™ and its functionality to manage the virtualization of Power servers. Systems Director VMControl is available in three editions: Express Edition, Standard Edition, and Enterprise Edition. Figure 5-6 describes the features of each edition. The Express Edition is a free download. The Standard Edition and the Enterprise Edition require a valid license after a 60-day evaluation period.

VMControl Features			
Features	Express	Standard	Enterprise
Create and manage virtual servers	✓	✓	✓
Virtual server relocation	✓	✓	✓
Import, create, edit, and delete virtual appliances		✓	✓
Deploy virtual appliances		✓	✓
Maintain virtual images in repository		✓	✓
Manage virtual workloads in system pools			✓

Figure 5-6 VMC features supported by the various editions

5.2.1 VMControl terminology

This section explains the VMControl terminology.

Virtual server

A *virtual server* is associated with a host system. It is called an LPAR, a partition, or a virtual machine.

Virtual appliance

A *virtual appliance* contains an image of a full operating system, and it can contain software applications and middleware. The virtual appliance contains metadata describing the virtual server.

Workload

A *workload* represents a deployed virtual appliance that allows you to monitor and manage one or more virtual servers as a single entity. For example, a workload that might contain both a web server and a database server can be monitored and managed as a single entity. A workload is automatically created when a virtual appliance is deployed.

System pools

A *system pool* groups similar resources and manages the resources within the system pool as a single unit. Storage system pools and server system pools are examples of system pools.

A *server system pool* consists of multiple hosts and their associated virtual servers, along with the attached shared storage.

Important: Hosts that are not connected to the same shared storage as the server system pool cannot be added to the system pool.

Image repositories

The created virtual appliances are stored in storage that is considered to be an *image repository*. The image repositories can be a Network Installation Management (NIM)-based storage system or virtual I/O server-based storage system.

VMControl subagents

The following characteristics apply to the VMControl subagents:

- ▶ For VMControl to see the images in the repositories, agents need to be installed in the repository system.
- ▶ If NIM is used to manage the virtual appliance, the subagent needs to be installed in the NIM master.
- ▶ If virtual I/O server is used to manage the image repository, the subagent needs to be installed in the virtual I/O server partition.
- ▶ In both cases, the subagents are installed on top of the common agent.

Import

The *import* task enables you to import a virtual appliance package, storing the virtual appliance that it contains within VMControl. Then, the virtual appliance can be deployed.

Virtual farms

A *virtual farm* logically groups like hosts and facilitates the relocation task: moving a virtual server from one host to another host within the virtual farm. A virtual farm can contain multiple hosts and their associated virtual servers.

Relocation

The following policies are *relocation* policies:

- ▶ **Manual relocation**
This policy relocates one or more virtual servers from an existing host at any time. To relocate within virtual farms, choose the relocation target. If relocating within server system pools, the relocation target is automatically identified.
- ▶ **Policy-based relocation**
This policy activates a resiliency policy on a workload so that VMControl can detect a predicted hardware failure problem that relates to processors, memory subsystems, power source, or storage and relocate the virtual servers to another host in the server system pool. Policy-based relocation can be done with approval or without approval.
- ▶ **Automatic relocation**
The VMControl (refer to Figure 5-7 on page 167) server system pools can predict hardware failure problems and relocate the virtual servers to maintain resilience.
For example, you can activate a threshold to monitor high and low values for CPU utilization in workloads. You can create an automation plan to automatically relocate the virtual server's system pool when the thresholds are crossed.

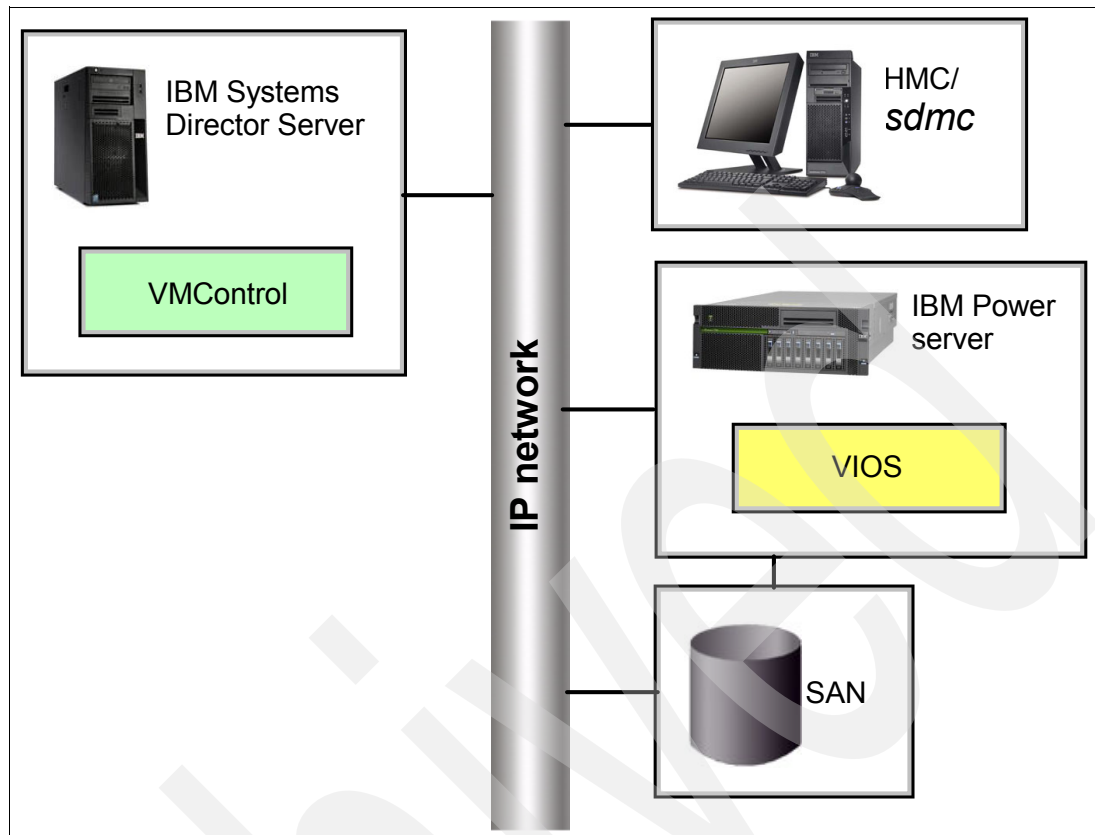


Figure 5-7 VMControl environment with Systems Director, HMC/SDMC, virtual I/O server, and storage

5.2.2 VMControl planning and installation

When planning for VMControl, the following components are required:

- ▶ Systems Director installation
- ▶ VMControl plug-in installation in the Director server
- ▶ VMControl subagents installation in the NIM and virtual I/O server partition

Installation steps

Follow these Systems Director and the VMControl plug-in installation steps:

1. Use the detailed installation information at these links to download and install the Systems Director:
 - Director download link:
<http://www.ibm.com/systems/software/director/resources.html>
 - Installation link:
http://publib.boulder.ibm.com/infocenter/director/v6r2x/index.jsp?topic=/com.ibm.director.main.helps.doc/fqm0_main.html

Use the information at this link to log in to the Systems Director server for the first time:

http://publib.boulder.ibm.com/infocenter/director/v6r2x/index.jsp?topic=/com.ibm.director.main.helps.doc/fqm0_main.html

2. Follow these VMControl plug-in installation steps:

Systems Director VMControl is installed on systems running Systems Director server Version 6.2.1 or higher.

a. Download the VMControl plug-in from the following link:

<http://www-03.ibm.com/systems/software/director/downloads/plugins.html>

b. Select the download package for the operating system that is running on your Systems Director server:

- For AIX/Linux: SysDir_VMControl_<ver>_Linux/AIX.tar.gz
- For AIX: SysDir_VMControl_2_2_AIX.tar.gz
- For Microsoft Windows: SysDir_VMControl_Windows.zip

c. Copy the download package to a directory or folder in the Systems Director server and extract the contents of the package:

```
gzip -cd SysDir_VMControl_<ver>_Linux/AIX.tar.gz | tar -xvf -
```

d. Change to the extracted folder and install the VMControl plug-in:

- For AIX/Linux: IBMSystems-Director-VMControl-Setup.sh
- For Microsoft Windows: IBMSystems-Director-VMControl-Setup.exe

e. Edit the following lines in the `installer.properties` file for silent mode installation:

```
INSTALLER_UI=silent  
LICENSE_ACCEPTED=true  
START_SERVER=true (this entry starts the director server on reboot)
```

f. Follow the instructions in the installation wizard to install Systems Director VMControl. Ensure that you restart the Systems Director server.

g. Check the log file to see if the installation completed successfully:

- For AIX/Linux: `/opt/ibm/Director/VMControlManager/installLog.txt`
- For Microsoft Windows: `\Director\VMControlManager\installLog.txt` (path where Director is installed)

h. Go to this link to obtain the hardware requirement for VMControl installation:

http://publib.boulder.ibm.com/infocenter/director/v6r2x/topic/com.ibm.director.plan.helps.doc/fqm0_r_supported_hardware_and_software_requirements.html

i. Download the VMControl installation from this link:

<http://www-03.ibm.com/systems/software/director/downloads/plugins.html>

Combined installation: At the time of writing this book, the current VMControl version was Version 2.3.1. In later releases, IBM intends to make the VMControl plug-in installation a part of the Systems Director installation. Then, no separate installation will be required.

3. Installing VMControl agents and subagents:

– NIM subagent

VMControl uses the NIM Master to manage the virtual appliance. For VMControl to connect with the NIM Master, a subagent needs to be installed in the NIM Master, as shown in Figure 5-8 on page 169.

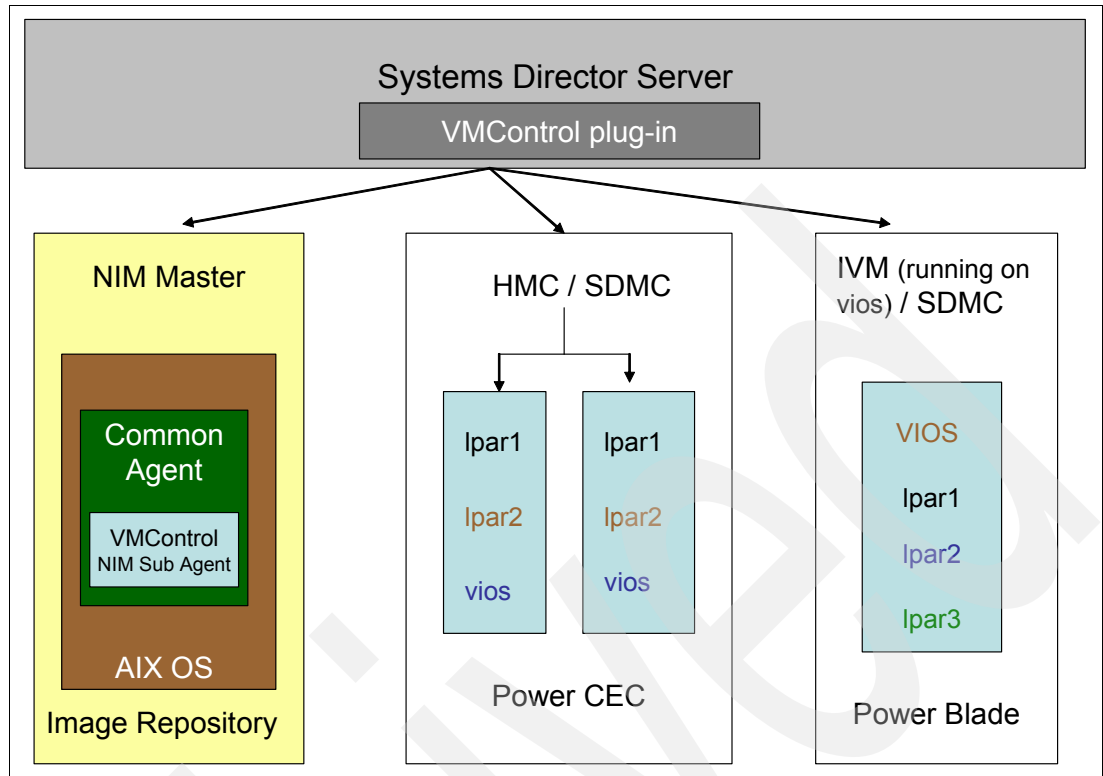


Figure 5-8 NIM subagent installation in the NIM Master server

– Common repository subagent:

- VMControl makes use of the virtual I/O server partition to store the raw disk images that are associated with AIX or Linux virtual appliances. The storage is allocated from the SAN and provided through the virtual I/O server.
- For VMControl to connect with the image repository, both the Systems Director Common Agent and VMControl Common repository subagents need to be installed in the virtual I/O server partition, as shown in Figure 5-9 on page 170.

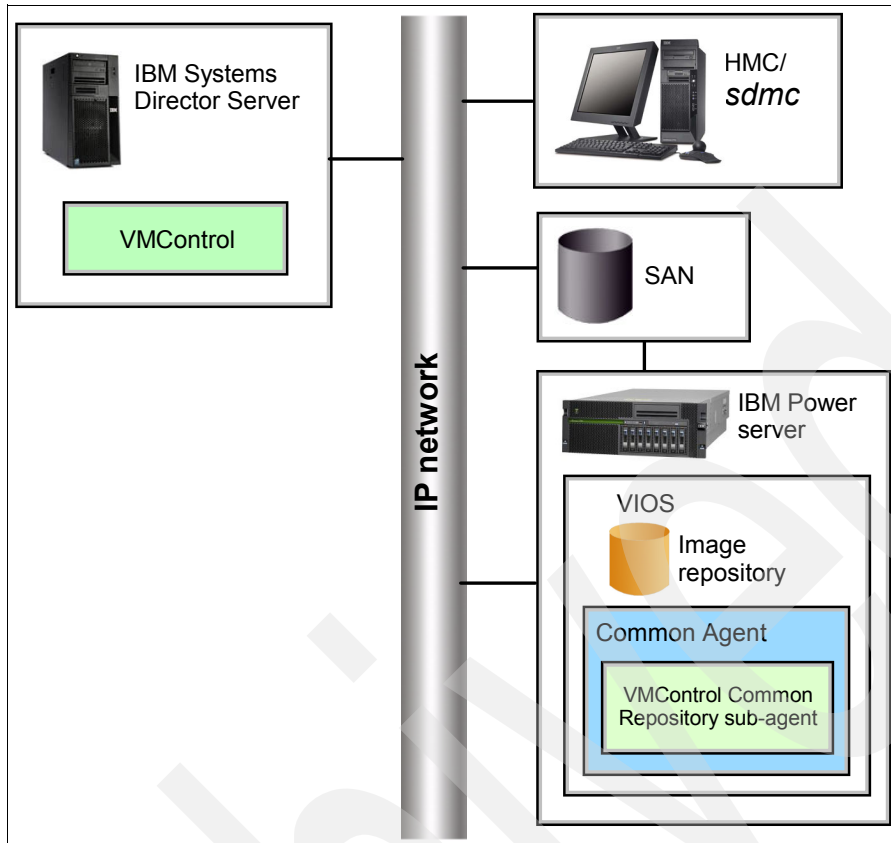


Figure 5-9 Repository subagent

Important: Ensure that the Systems Director's Version 6.2.1 or higher common agent is installed on the target NIM Master and virtual I/O server partition.

- Subagent installation can be done through the following methods:
 Subagent installation through the Systems Director's release management task:
 - i. In the IBM Systems Director navigation pane, expand **Release management**.
 - ii. Click **Agents**.
 - iii. On the Agents page, click **Common Agent Subagent Packages**.
 - iv. From the Common Agent Subagent Packages view, select the subagent that needs to be installed.
 - v. Click **Actions** on the menu bar.
 - vi. Select **Release Management** → **Install Agent**.
 - vii. Follow the instructions in the installation wizard to install the subagent.
 Manual subagent installation steps:
 - i. Locate the subagent in the following folder:
 For AIX/Linux: /opt/ibm/director/packaging/agent
 For Microsoft Windows: C:\Program Files\IBM\Director\packaging\agent
 - ii. Copy the subagent to a temporary folder, for example, /tmp.
 For NIM, the agent is CommonAgentSubagent_VMControl_NIM_2.3.1.

For the common repository, the agent is
CommonAgentSubagent_VMControl_CommonRepository-2.3.1.

iii. Change the directory to the Systems Director system bin directory:
/opt/ibm/director/agent/bin

iv. Use the `./lwiupdatemgr.sh` command to install the subagent.

For the NIM subagent:

```
./lwiupdatemgr.sh -installFeatures -featureId  
com.ibm.director.im.rf.nim.subagent -fromSite  
jar:file:/tmp/com.ibm.director.im.rf.nim.subagent.zip\!/site.xml -toSite  
"file:/var/opt/tivoli/ep/runtime/agent/subagents/eclipse/"
```

For the common repository subagent:

```
./lwiupdatemgr.sh -installFeatures -featureId  
com.ibm.director.im.cr.agent.installer-fromSite  
jar:file:/tmp/scom.ibm.director.im.cr.agent.installer.zip\!/site.xml -toSite  
"file:/opt/ibm/director/agent/runtime/agent/subagents/eclipse/"
```

For the Linux environment involving VMware, you must install the following subagents:

- VMware vCenter 4.x subagent: CommonAgentSubagent_VSM_VC4x-6.2.1
- VMware VirtualCenter 2.x subagent: CommonAgentSubagent_VSM_VC2x-6.2.1
- VMware ESX 4.x subagent: CommonAgentSubagent_VSM_ESX4x-6.2.1
- VMware ESX 3.x subagent: CommonAgentSubagent_VSM_ESX3x-6.2.1

5.2.3 Managing a virtual server

With Systems Director VMControl, virtual appliances can be deployed to virtual servers. Virtual servers can be created, edited, and deleted through the Systems Director (without the need of the VMControl plug-in).

Creating a virtual server

The following steps provide instructions to create a virtual server. The Systems Director provides wizards to create a virtual server. We outline the steps for creating a virtual server next. After the virtual server is created, the virtual appliance can be deployed through VMControl.

Important:

- ▶ Ensure that the managed system has been discovered through the Systems Director.
- ▶ The virtual server can also be created through the HMC or the SDMC.

Follow these steps to create a virtual server:

1. In the Systems Director, click **Navigate Resources** to locate the host.
2. Select the host.
3. Click **Actions** from the menu bar.
4. Click **System Configuration**.
5. Select **Create Virtual Server**, as shown in Figure 5-10 on page 172.
6. Follow the Create Virtual Server wizard to set up the virtual server.

Note: The Create Virtual Server task gives you an option to run it immediately or schedule it to run at a later time.

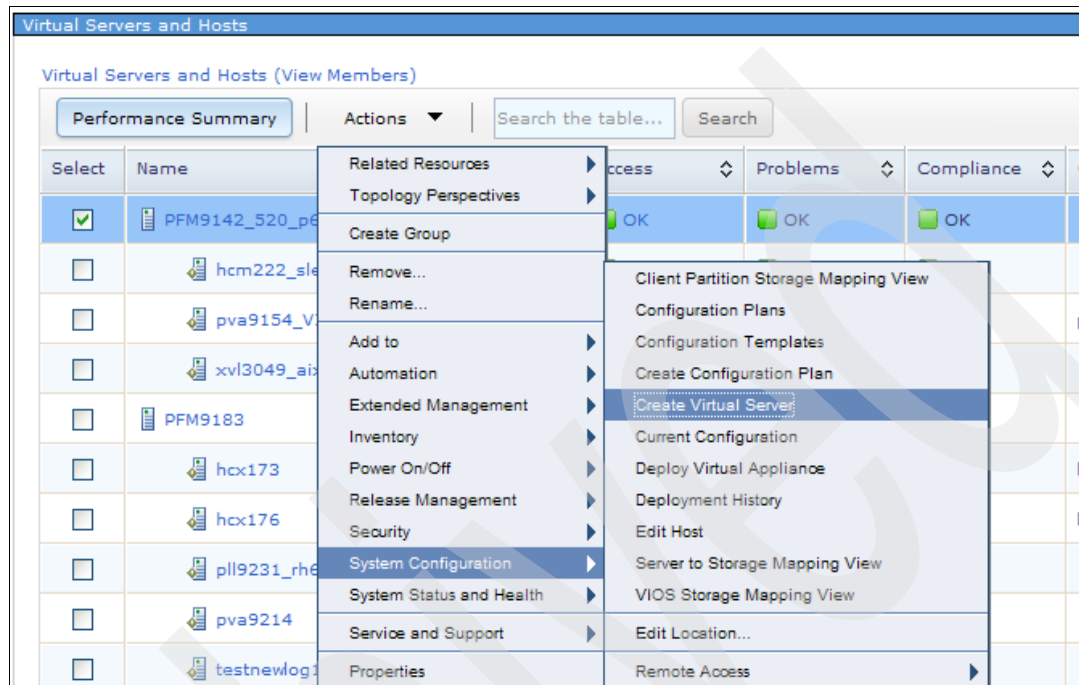


Figure 5-10 Create Virtual Server

Editing a virtual server

This section illustrates how to edit a virtual server. After creating virtual servers through the Systems Director, the virtual servers can be edited, as well, through the Systems Director. Follow these steps to edit a virtual server:

1. In the Systems Director, click **Navigate Resources** to locate the virtual server.
2. Select the virtual server.
3. Click **Actions** from the menu bar.
4. Click **System Configuration**.
5. Select **Edit Virtual Server**.

Editing: The Systems Director allows you to edit processor and memory details. For any other modification, the Systems Director provides options to launch other platform management utilities, such as the HMC and the SDMC.

Deleting a virtual server

You can delete a virtual server permanently from the host, as shown in Figure 5-11 on page 173.

Important: Power off the virtual server first to delete it permanently.

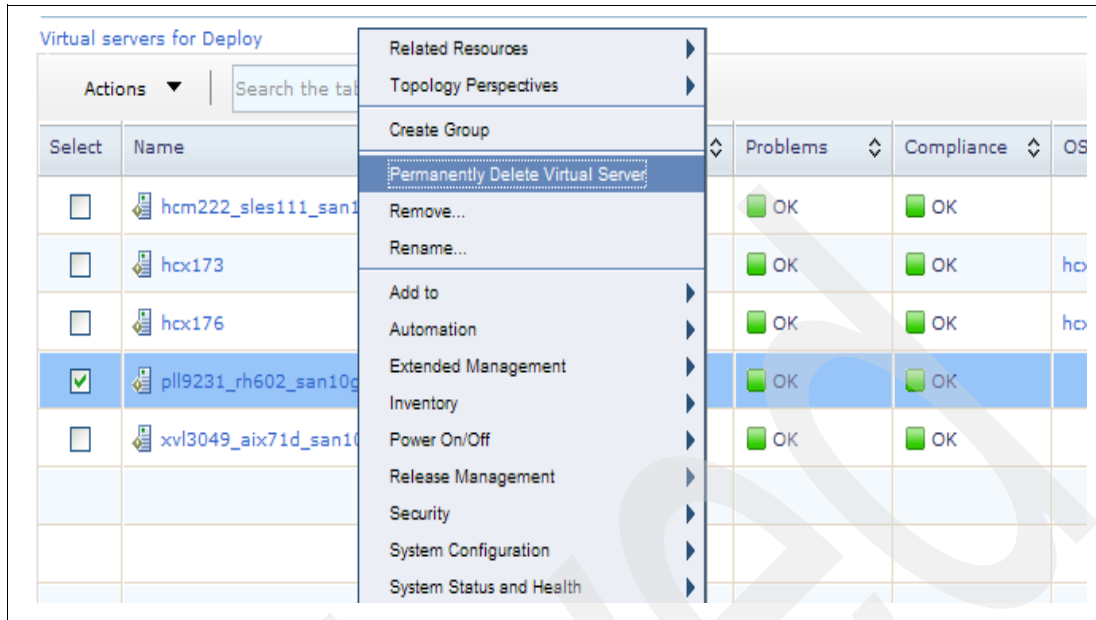


Figure 5-11 Permanently Delete Virtual Server

5.2.4 Relocating a virtual server

VMControl can relocate virtual servers in response to predicted hardware failures related to processors, memory subsystems, a power source, or storage. Also, a virtual server can be relocated for planned maintenance or downtime or to adjust resources to improve performance.

Live Partition Mobility: The relocation feature makes use of the Live Partition Mobility (LPM) functionality from the IBM Power Systems servers.

You can perform relocation in the following ways:

- ▶ Static relocation

With static relocation, if the virtual server is powered on, the relocation operation powers off the virtual server at the beginning of the relocation process and powers the virtual server on when the relocation is complete.

- ▶ Live relocation

With live relocation, if the virtual server is powered on, the relocation occurs without powering the server off. There are three options from which to choose for the relocation:

- Manually relocate virtual servers at any time.
- Activate a resilience policy on a workload to relocate virtual servers automatically to prevent predicted hardware failures from affecting the availability of the workload.
- Create an automation plan to relocate the virtual servers when certain events occur.

Relocating virtual servers manually

Systems Director VMControl chooses the target host from similarly configured hosts in the server system pool and displays the proposed relocation actions. You either accept or cancel the relocation operation based on your requirements.

Relocating virtual servers using the resilience policy

The resilience policy enables Systems Director VMControl to relocate virtual servers automatically to maintain the resilience (high availability) of workloads.

When the resilience policy is activated, Systems Director VMControl can automatically relocate virtual servers when a predicted hardware failure is detected. VMControl moves virtual servers away from a failing host system, and it relocates them to a host that the server system pool determines has adequate resources.

Approval required: By default, a prompt to approve any policy-based action, such as relocation, appears before the move is performed.

Relocating virtual servers automatically

You create an automation plan to relocate the virtual servers from the host with a critical event (for example, a hardware problem, high CPU utilization, and so on) to a host that the server system pool determines has adequate resources.

5.2.5 Managing virtual appliances

As described under 5.2.1, “VMControl terminology” on page 165, a virtual appliance is the bundle of the operating system image and the application that is installed on top of it. In addition, the virtual appliance also has information about the virtual server configuration details that are bundled.

The first step in managing virtual appliances is creating image repositories. The following entities must exist before managing virtual appliances:

- ▶ **Image repositories:** Virtual appliances are stored in image repositories:
 - **AIX:** It has two options: NIM image repositories and virtual I/O server image repositories
 - **Linux:** Virtual I/O server image repositories
- ▶ **Agents:** They are specific to the environment. NIM-based subagents are installed in the NIM Master, and storage-based subagents are installed in the virtual I/O server partition.
- ▶ **Discovery:** VMControl must have discovered the image repositories and virtual appliances.

Creating image repositories

We discuss the creation of the image repositories.

Discovering NIM image repositories for AIX

The following actions are required to have NIM image repositories for AIX:

1. Discover and request access to the NIM server.
2. At this stage, image repositories, such as mksysb, are already created in the NIM server.
3. Ensure that the VMControl NIM subagent is installed in the NIM server, as explained in 5.2.2, “VMControl planning and installation” on page 167.
4. Run inventory collection on the NIM server.
5. From the VMControl summary page, go to the **Basics** tab and click **Discover virtual appliances** to discover your repositories (mksysb) and virtual appliances, as shown in Figure 5-12 on page 175. The virtual appliances that are already present in your

repositories and that have been imported or captured using VMControl are detected by VMControl.

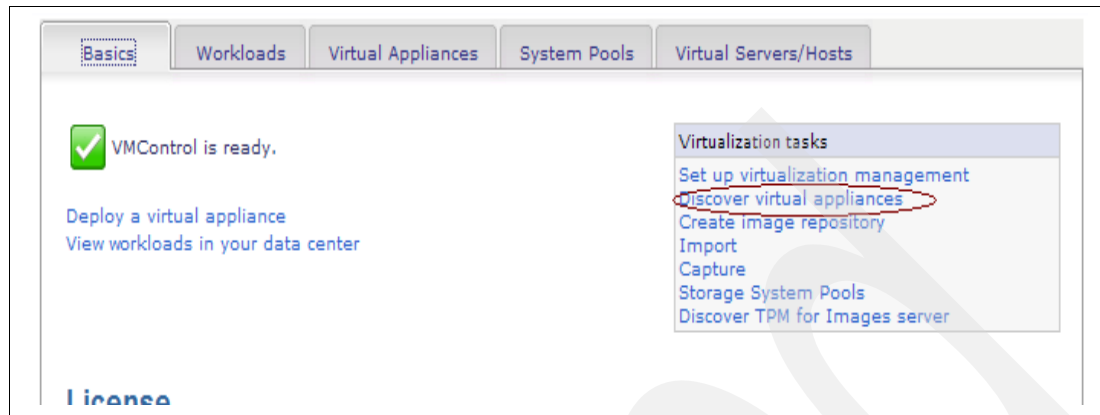


Figure 5-12 Discovering virtual appliances in the NIM image repository

Creating and discovering VIOS image repositories for AIX and Linux

You must follow these steps for VMControl to create image repositories:

1. Set up SAN storage pools, and set up a virtual I/O server that has access to the storage pool.
2. Discover and request access to the storage and the operating system of the virtual I/O server.
3. Ensure that the VMControl common repository subagent software is installed on the virtual I/O server that hosts the image repository, as explained in 5.2.2, “VMControl planning and installation” on page 167.
4. Run the full inventory collection on the operating system of the virtual I/O server to gather information about the image repository subagent.
5. Create an image repository. From the VMControl summary page, go to the **Virtual Appliances** tab and click **Create image repository**, as shown in Figure 5-13.

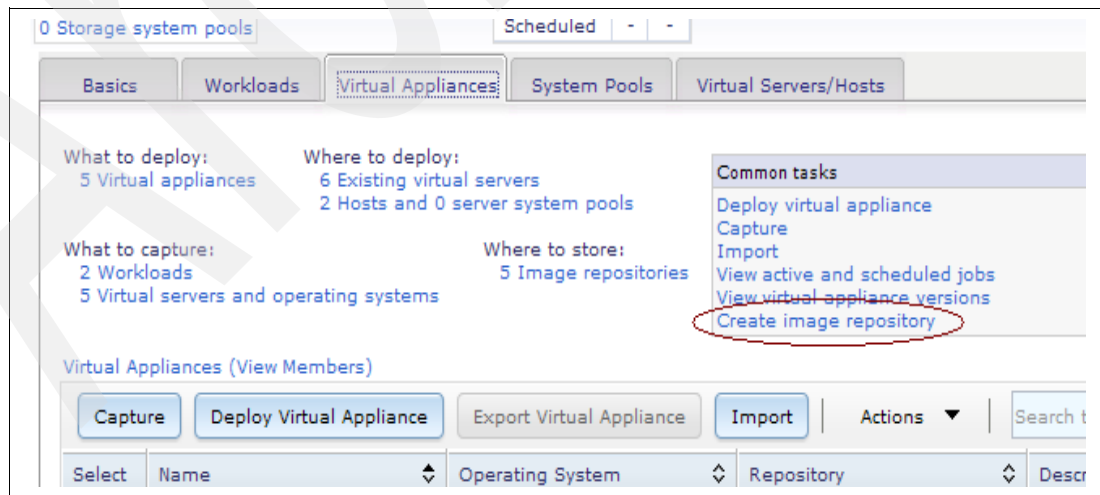


Figure 5-13 Create image repository: Virtual I/O server image repository

6. Follow the instructions in the Create Image Repository wizard, as shown in Figure 5-14 on page 176.

Figure 5-14 Create Image Repository wizard

Importing a virtual appliance package

Follow these steps to import a virtual appliance package:

1. Go to the VMControl **Basics** tab.
2. Under Common tasks, select the **Import** option (as seen in Figure 5-13 on page 175) to import the virtual appliance package in Open Virtualization Format (ovf) format into the image repository.

Figure 5-15 shows the Import virtual appliance wizard.

Figure 5-15 Import virtual appliance: Importing the .ovf file

Capturing virtual appliances

To capture virtual appliances for a virtual server or workload to create a virtual appliance, or for a mkysb image or resource to create a virtual appliance, or for a NIM lpp_source resource or directory to create a virtual appliance, refer to the following steps:

1. To capture the virtual appliance image, click **System Configuration** → **VMControl** → **Virtual Appliances** tab, as shown in Figure 5-16 on page 177.

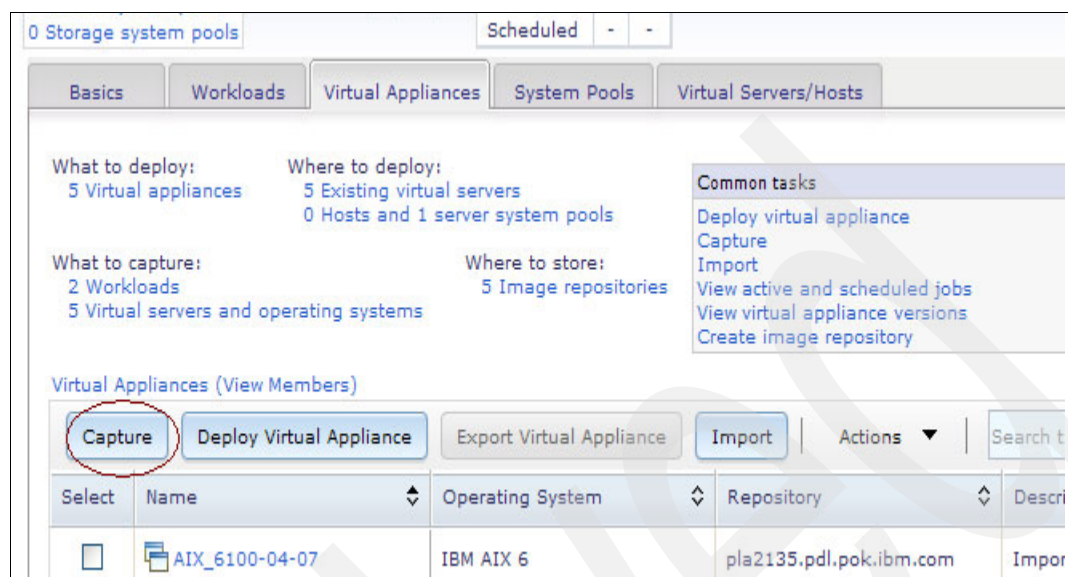


Figure 5-16 Capture the virtual appliance under the VMControl window

2. The Capture virtual appliance wizard takes you through the steps involved to provide the source virtual server repository where you want to store the image that is associated with the new virtual appliance, as shown in Figure 5-17.

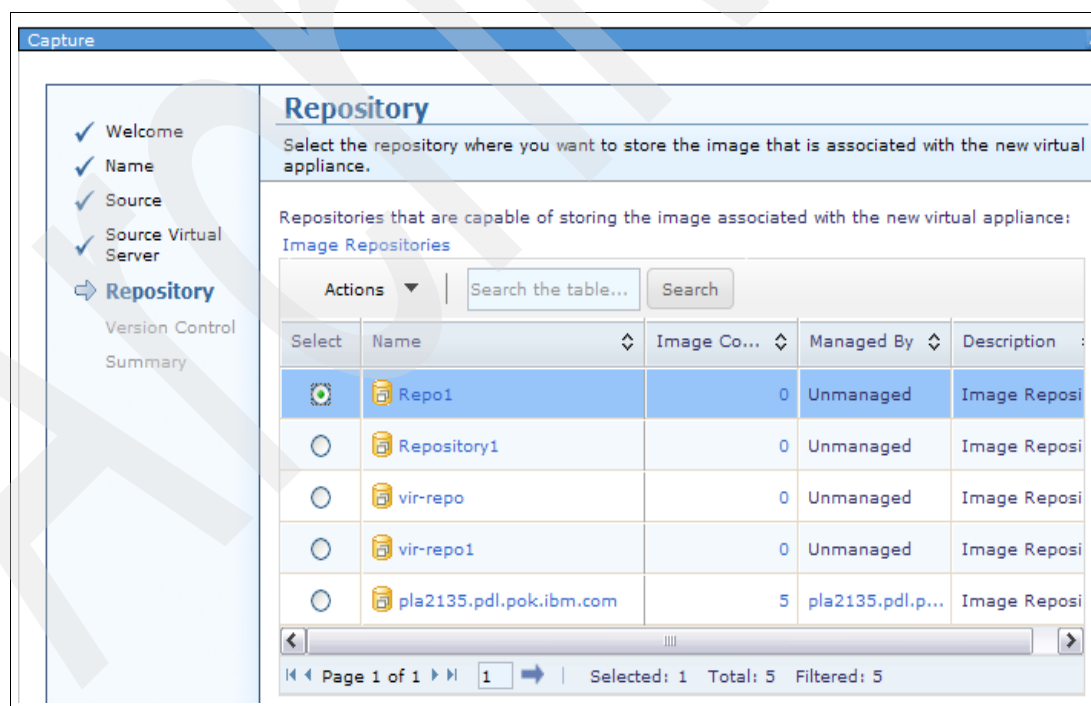


Figure 5-17 Capture the virtual appliance: Selecting the repository into which to store the image

Deploying the virtual appliance

Similar to capturing virtual appliances, the Deploy Virtual Appliance wizard helps in deploying the virtual appliance, as shown in Figure 5-18 on page 178.

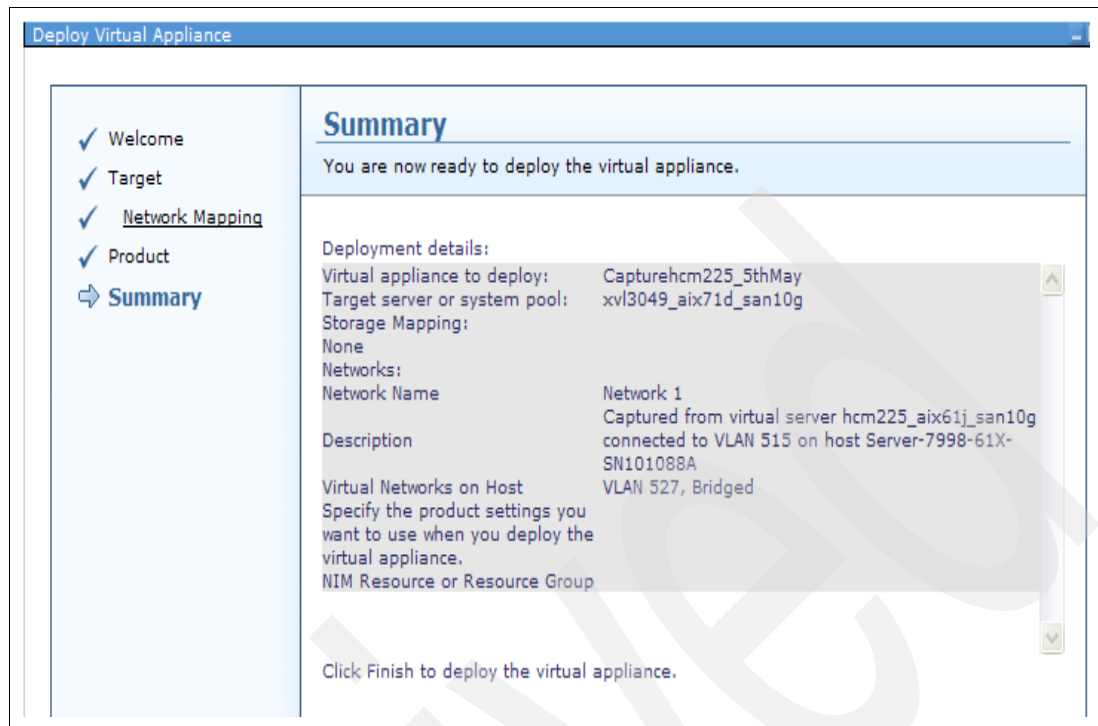


Figure 5-18 Deploying virtual appliances

5.2.6 Creating a workload

As discussed in the VMControl terminology in 5.2.1, “VMControl terminology” on page 165, each virtual appliance is considered a *workload*. At the same time, using the following step, you can group virtual appliances to create a workload for better monitoring and management.

To create a workload, click **System Configuration** → **VMControl** → **Workloads** tab → **Create workload**, as shown in Figure 5-19 on page 179.

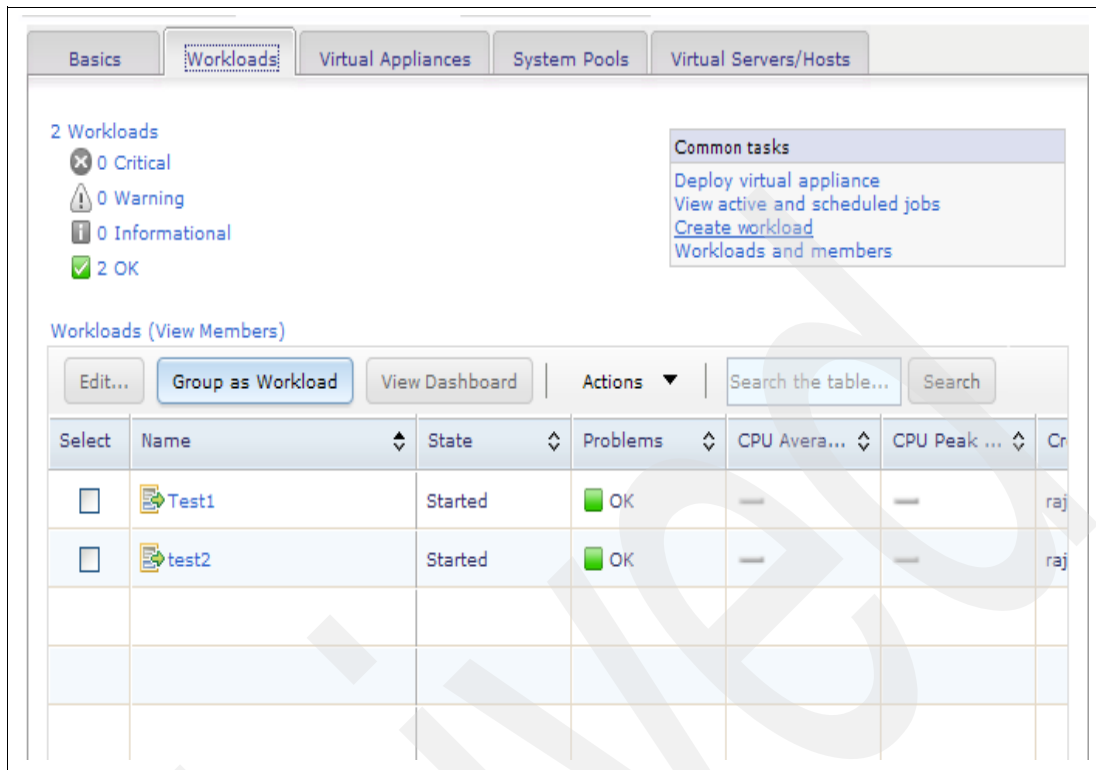


Figure 5-19 Create a workload

5.2.7 Managing server system pools

Server system pools enable grouping similar hosts. It is advantageous to manage and monitor hosts through a pool. This method provides more resiliency through relocation within a server pool.

Requirements: Ensure that the required VMControl plug-in and the subagent are already installed. 5.2.2, “VMControl planning and installation” on page 167 offers more details about the required plug-in, agents, and installation steps.

Creating a server system pool

Use the following steps to create a server system pool:

1. Click **System Configuration** → **VMControl** → **System Pools** tab, as shown in Figure 5-20 on page 180.

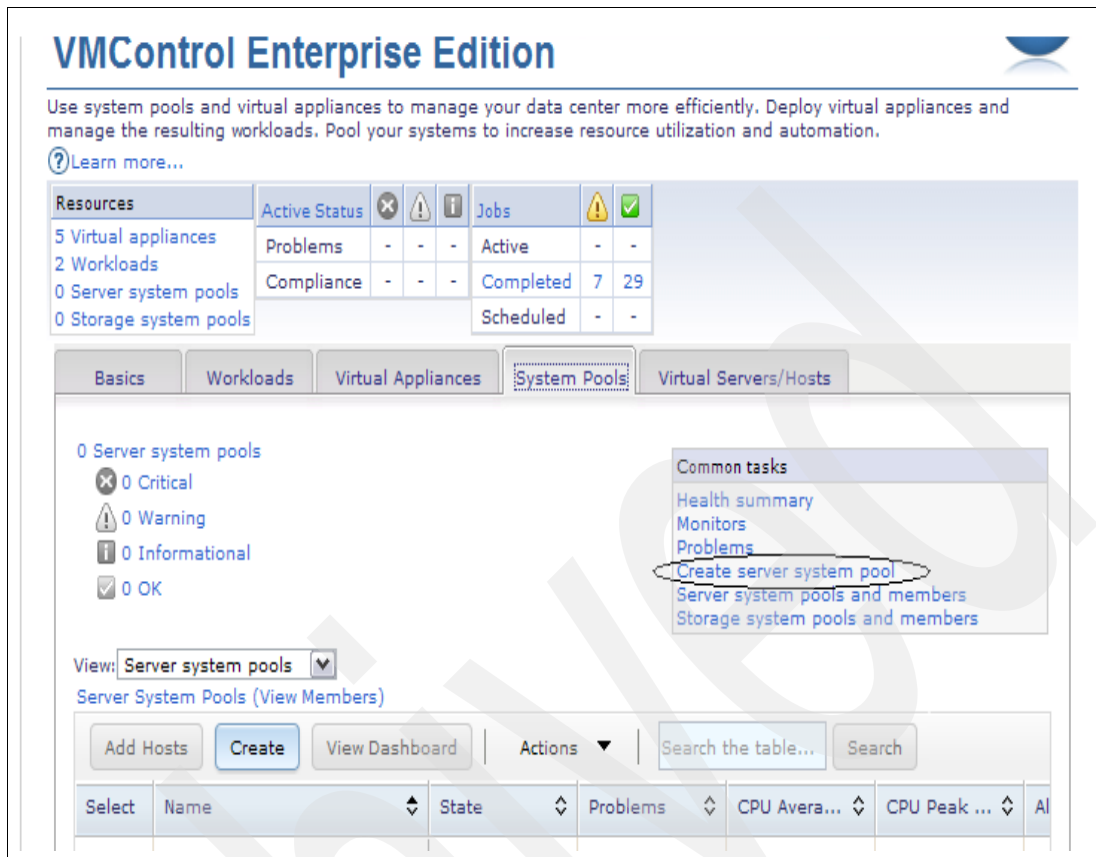


Figure 5-20 Create server system pool

2. Use the wizard to create a server pool.
3. While creating the server pool, you are presented with options to select the pool resilience criteria, as shown in Figure 5-21 on page 181.

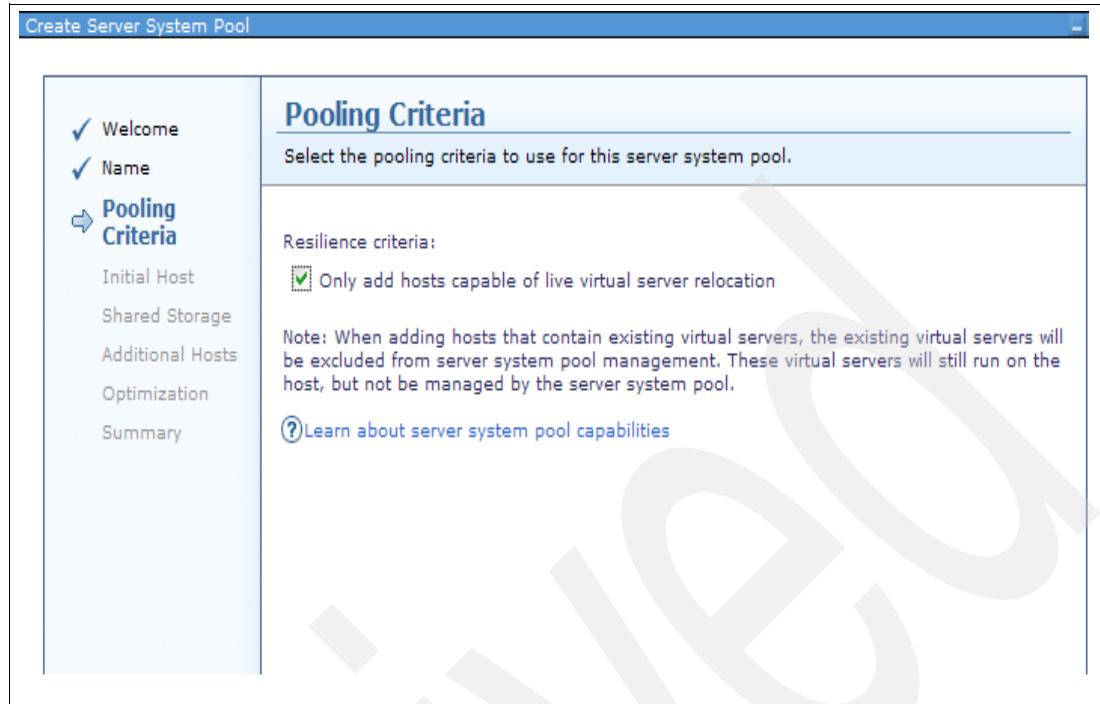


Figure 5-21 Server pool resilience criteria

4. All hosts in the server system pool must use the same shared storage. Available shared storage is listed for selection, as shown in Figure 5-22.

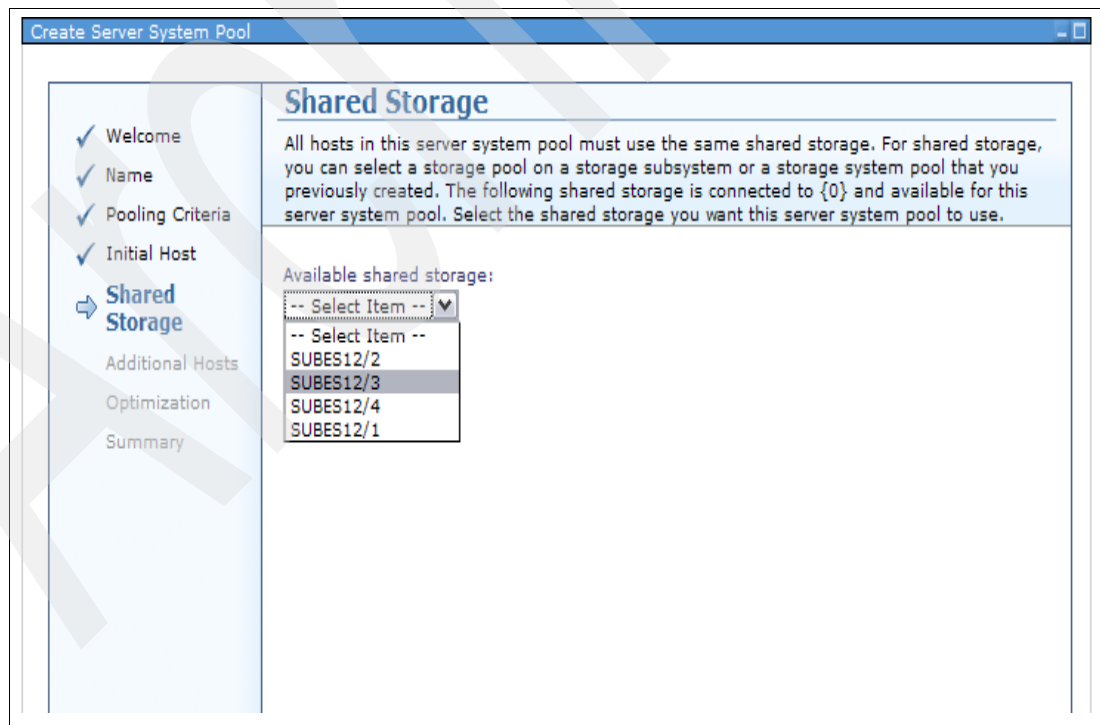


Figure 5-22 System pool shared storage

5. The optimization function analyzes the server system pool and periodically optimizes for performance. The process can relocate the virtual server to the most efficient host. We provide more details about optimization next. You must select the optimization option while creating the server system pool, as shown in Figure 5-23 on page 183.

Configuring SAN storage for server system pools

This section illustrates how to configure SAN storage for server system pools.

Add/remove host from server system pool

In addition to server pool creation, VMControl provides options to add/remove the hosts from the pool.

Server system pool optimization

Optimization enables the analysis and periodic performance improvement of all the virtual servers within a server system pool based on specified needs, such as the relocation of the virtual servers within a workload to the most efficient hosts.

When optimization is run, the system pool is examined for performance hot spots, that is, systems in the pool that are heavily used. If a hot spot is identified, workloads are relocated to better balance resource usage in the environment.

Important: The hosts in the server system pool must support the relocation of their virtual servers for workload resiliency, as shown in Figure 5-23 on page 183. The panel for configuring optimization is displayed only if all the hosts in the server system pool support workload resiliency.

There are two types of optimization:

- ▶ *Manual optimization:* Optimize manually whenever you want any task, such as relocation, to be started at convenient off-peak times.
- ▶ *Automated optimization:* In automatic optimization, system analysis determines whether a workload must be distributed.

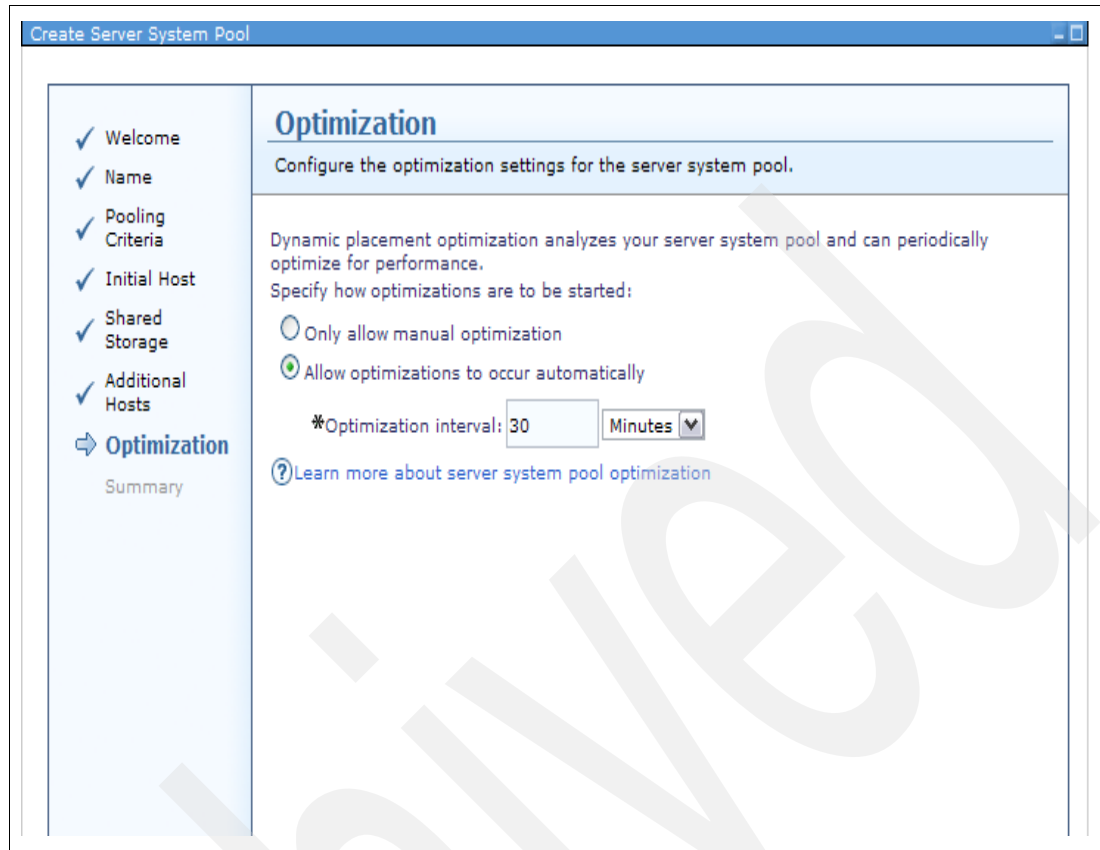


Figure 5-23 Server system pool optimization

We have described a few of the key features of VMControl. VMControl saves time and reduces the configuration complexity that is involved with virtualization. In the next section, we explain how another advanced software function from Systems Director called *Active Energy Management (AEM)* can be used to enable power saving options for IBM Power servers.

5.3 IBM Systems Director Active Energy Management (AEM)

In this section, we provide an overview of the Active Energy Manager (AEM) plug-in. We describe briefly the installation and uninstallation of AEM. We concentrate on how effectively the power can be managed and monitored on the IBM POWER7 Systems. We cover functionalities, such as power saving and power capping.

5.3.1 Active Energy Manager (AEM) overview

Active Energy Manager (AEM) is a Systems Director plug-in that offers power and thermal monitoring. AEM also comes with many management capabilities, which can be used to gain a better understanding of the data center's power usage. Overall, this feature helps you to better utilize the available power. You can also use AEM to plan for your future energy needs.

AEM directly monitors IBM Power Systems, System z, System x®, and IBM BladeCenter® servers. AEM can also indirectly monitor non-IBM equipment by using external metering devices, such as power distribution units (PDUs) and sensor devices.

5.3.2 AEM planning, installation, and uninstallation

In the following section, we describe how to implement AEM.

Prerequisite: Systems Director must be installed and configured before the installation of the AEM plug-in. You can obtain more information about how to configure Systems Director in this document:

<http://www.redbooks.ibm.com/redbooks/pdfs/sg247694.pdf>

Use this link to download the IBM Systems Director software:

<http://www-03.ibm.com/systems/software/director/downloads/plugins.html>

Installation steps

You must install AEM on systems running IBM Systems Director server Version 6.2.1 or later. The link in the previous shaded box provides the details about Systems Director. Follow these AEM plug-in installation steps:

1. Download the AEM plug-in from the following link:
<http://www-03.ibm.com/systems/software/director/downloads/plugins.html>
2. Select the download package for the operating system that is running on the Systems Director server.
3. Copy the downloaded package to a directory or folder on the Systems Director server and extract the contents of the package.
4. Set a few attributes in the `installer.properties` file. Change the value of the three following attributes for unattended installation:

```
INSTALLER_UI=silent:  
LICENSE_ACCEPTED=true  
START_SERVER=true
```

If the `START_SERVER` option is set to false, you need to manually restart AEM. Perform these steps to manually restart AEM:

- a. Run `#/opt/ibm/director/bin/smstop`.
- b. To see the status of the Systems Director, issue the `smstatus` command.
- c. To start the Systems Director, issue the `/opt/ibm/director/bin/smstart` command.

Removal steps

Before removing AEM from the Systems Director database, ensure that both AEM and the Systems Director server are running. The AEM group only gets removed if the Systems Director server is active. Follow these AEM plug-in removal steps:

1. To uninstall AEM on a system running AIX/LINUX, edit the `installer.properties` to change the value of these attributes:

```
INSTALLER_UI=silent  
DATABASE_CLEANUP=true
```

2. Launch the uninstaller.

5.3.3 AEM and the managed systems

After installing AEM, we need to connect the managed systems to the Systems Director. Figure 5-24 illustrates how a server can be connected to Systems Director.

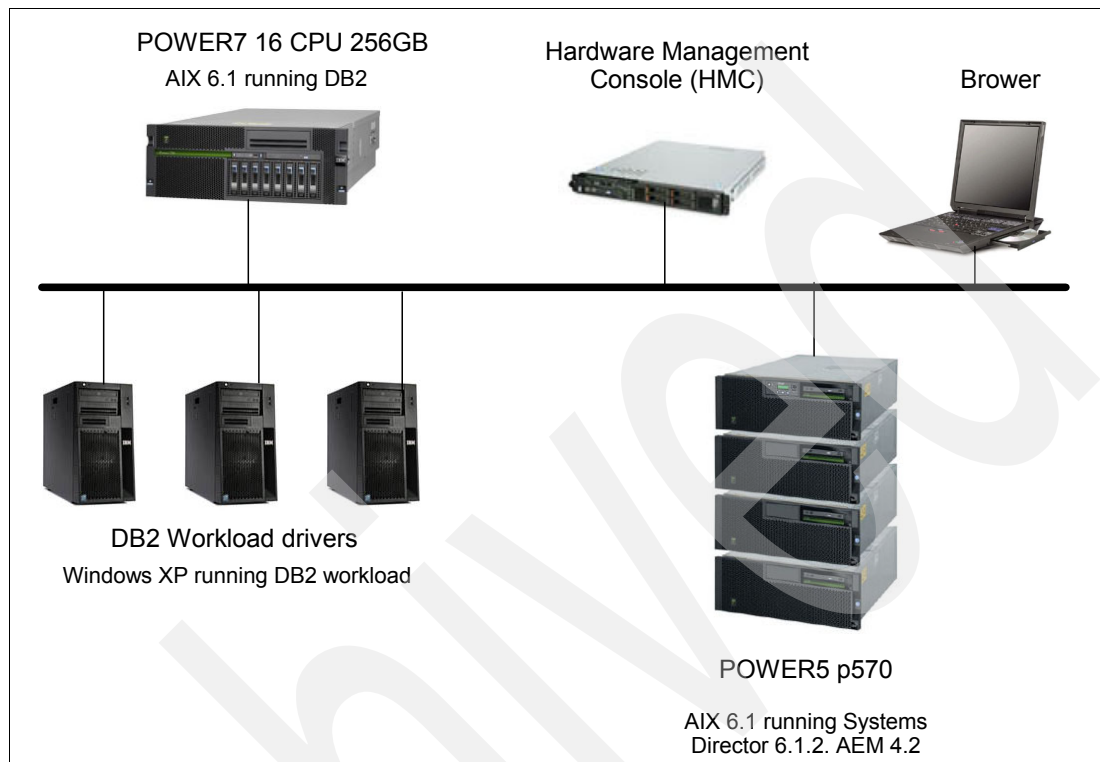


Figure 5-24 Connection between Systems Director and the managed server

AEM, running on a Systems Director, communicates to the HMC, which communicates to the server. You can access the Systems Director GUI via a browser. The DB2® workload is running on the server (refer to Figure 5-24) managed by the HMC. The DB2 server simulates the workload that is generated on the production servers. AEM manages and monitors the energy that is consumed by the servers.

We need to connect the managed systems to the Systems Director. Follow these steps:

1. The first step is to discover the HMC, which manages the server. Log in to IBM Systems Director server as a root user.
2. Select the **Inventory** tab.
3. Select the **System Discovery** tab.
4. Enter the IP address of the HMC.

- Click **Discover Now**, as shown in Figure 5-25.

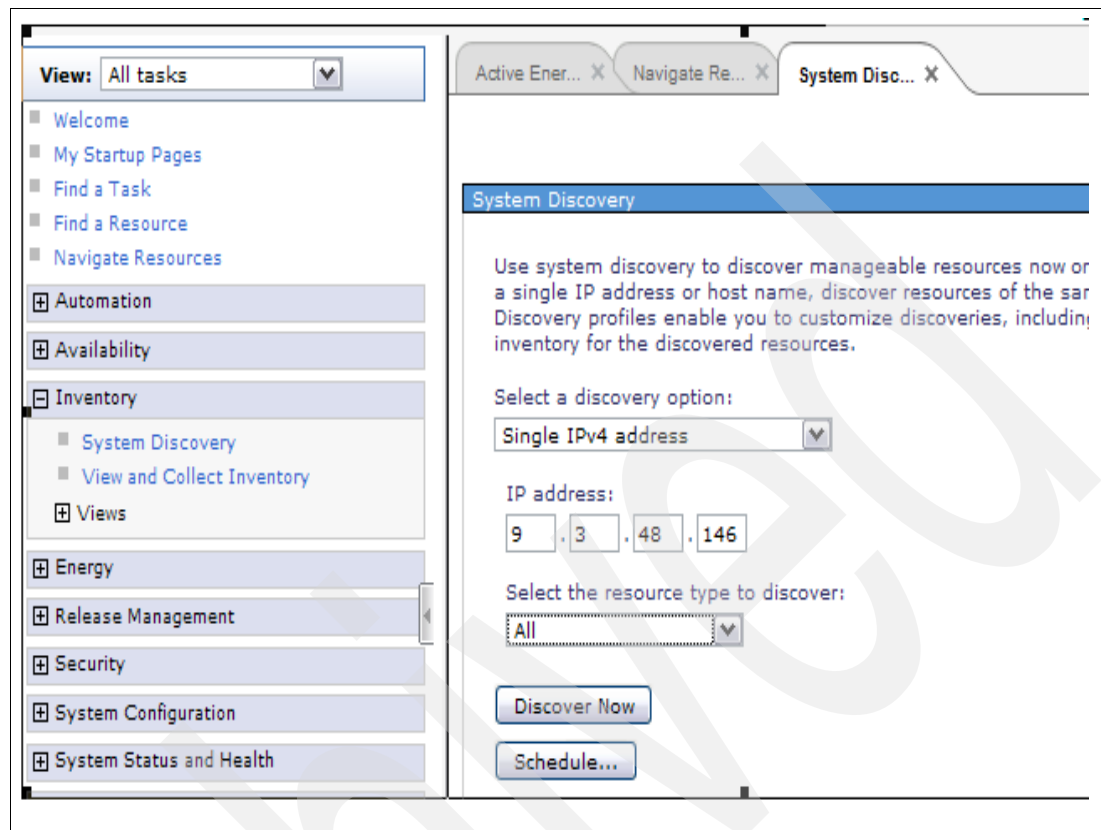


Figure 5-25 HMC discovery

- After the HMC is discovered, you need to give access to the system. Under the column heading **Access**, click **No Access**, as shown in Figure 5-26.

Actions ▼				
Name ▲	Discovered ▲	Type ▼	Access ▼	Problems ▼
guandu3	New	Hardware Ma...	No access	OK
guandu3.upt.austin.ibm.c...	New	Operating Sy...	No access	OK

Figure 5-26 Accessing the discovered system

- Provide the user ID and password for the HMC

- Click **Request Access**, as shown in Figure 5-27.

Figure 5-27 Accessing the discovered system

- After the access is granted, click **Inventory** → **Views** → **Platform Managers and Members**, as shown in Figure 5-28. In this pane, you see the HMC, as well as the servers that are managed by the HMC.

	ostName		Information	
<input type="checkbox"/>	ostName	OK	OK	OK
<input type="checkbox"/>	ostName	OK	OK	OK
<input type="checkbox"/>	ostName	OK	OK	OK
<input type="checkbox"/>	destiny3	Offline	OK	OK
<input type="checkbox"/>	ostName	Unknown	OK	OK
<input type="checkbox"/>	guandu1.upt.austin.ibm.c...	OK	OK	OK
<input type="checkbox"/>	%HostName%	OK	OK	OK

Figure 5-28 Discovered HMCs and the servers that are managed by the HMCs

5.3.4 Managing and monitoring the consumed power using AEM

There are two ways to manage power: the *power capping* method and the *power savings* method.

Power capping

You can use power capping to limit the power that is used by the server. Setting a power capping value ensures that system power consumption stays at or beneath the value that is defined by the setting. You can specify this value in terms of an absolute value or in terms of a percentage of the maximum power cap:

- ▶ **Absolute value:** This option is useful for a single object or for a group of similar objects for which the same power cap value is appropriate.
- ▶ **Percentage value:** This option is particularly useful in the case of a group of heterogeneous or unlike systems, where a specific power cap value is inappropriate, but percentage capping makes sense.

Licensing: The use of the power capping and power savings functions requires an Active Energy Manager license. You are granted a 60-day evaluation license when you begin using AEM. When the evaluation license expires, the power savings and power capping functions, on systems where these functions were activated, are turned off. The policies that were previously saved in the AEM still exist when the license expires, but they cannot be applied to any resource.

Power savings

Power savings is helpful in saving the amount of power that is used by the servers. There are two types of power savings:

- ▶ **Static power saving:** This mode lowers the processor frequency and voltage on a system by a fixed amount, therefore reducing the power consumption of the system while still delivering predictable performance. This percentage is predetermined to be within a safe operating limit and is not configurable by the user.

Static power saving can be enabled based on regular variations in workloads, such as predictable dips in utilization overnight, or over weekends. It can be used to reduce peak energy consumption, which can lower the cost of all power used. Note that when static power saving is enabled for certain workloads with low CPU utilization, workload performance is not affected, although CPU utilization might increase due to the reduced processor frequency.

- ▶ **Dynamic power saving:** Dynamic power saving allows for two kinds of modes: the favor performance mode and the favor power mode:
 - **Favor performance mode:** This mode allows the CPU cycles to go above the nominal frequency. The system increases the CPU frequency above the nominal frequency if the workload demands. With a lesser workload, the system runs at a lower frequency to save energy.
 - **Favor power mode:** In this mode, the CPU frequency is capped at a certain level, which means that the CPU frequency cannot go higher than that level, even if the workload demands it. The system increases the CPU frequency up to the nominal frequency if demanded by the workload. With no workload, the system runs at a slower frequency to save energy.

Power capping versus power savings

It is important to understand the differences between power capping and power savings. Power capping is used to allow the user to allocate less power to a system, which in turn helps to cool the system. This mode can help you save on the data center infrastructure costs and then potentially allow more servers to be put into an existing infrastructure. However, power savings is used to put the server into a mode that consumes less energy.

Configuring dynamic power savings

Follow these steps to configure dynamic power savings:

1. Log on to the IBM Systems Director server.

2. Expand the **Energy** tab to view the Active Energy Manager pane, as shown in Figure 5-29.

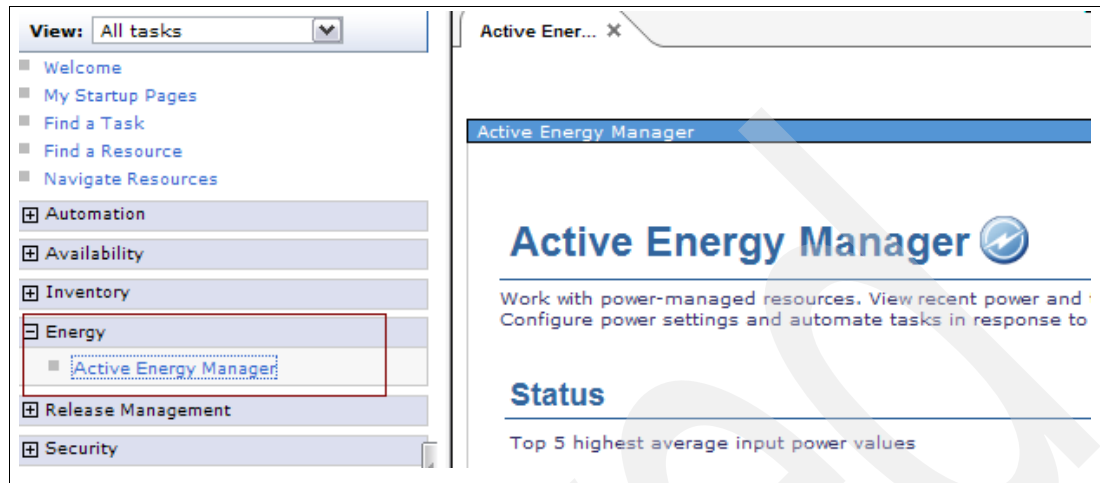


Figure 5-29 Active Energy Manager

3. From the AEM GUI, select **Active Energy Managed Resources**, which shows the list of the systems that the Systems Director can manage, as shown in Figure 5-30.

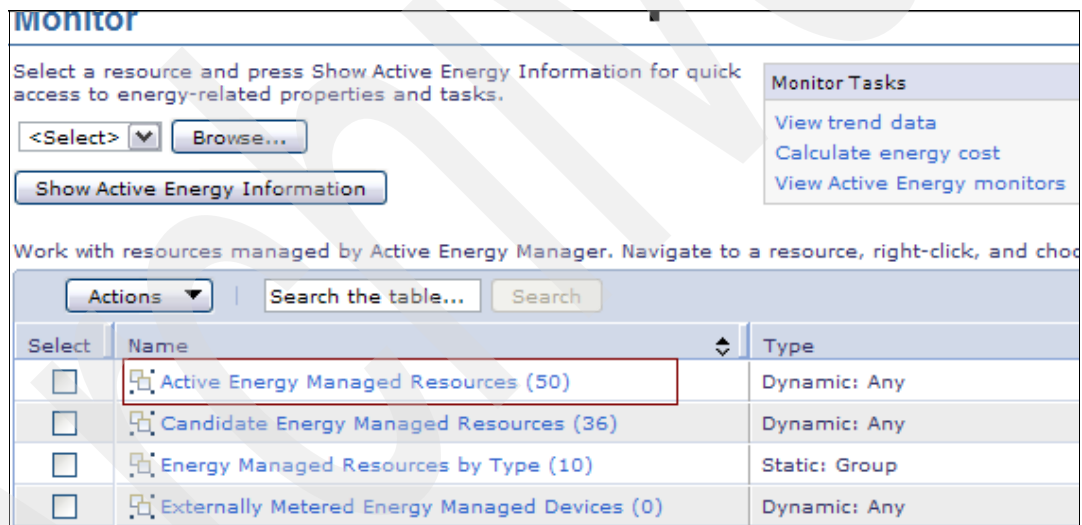


Figure 5-30 Active Energy Managed Resources

4. Select one of the systems. Then, click the Actions drop-down list box to navigate through the power management panel. Select **Energy** → **Manage Power** → **Power Savings**, as shown in Figure 5-31.

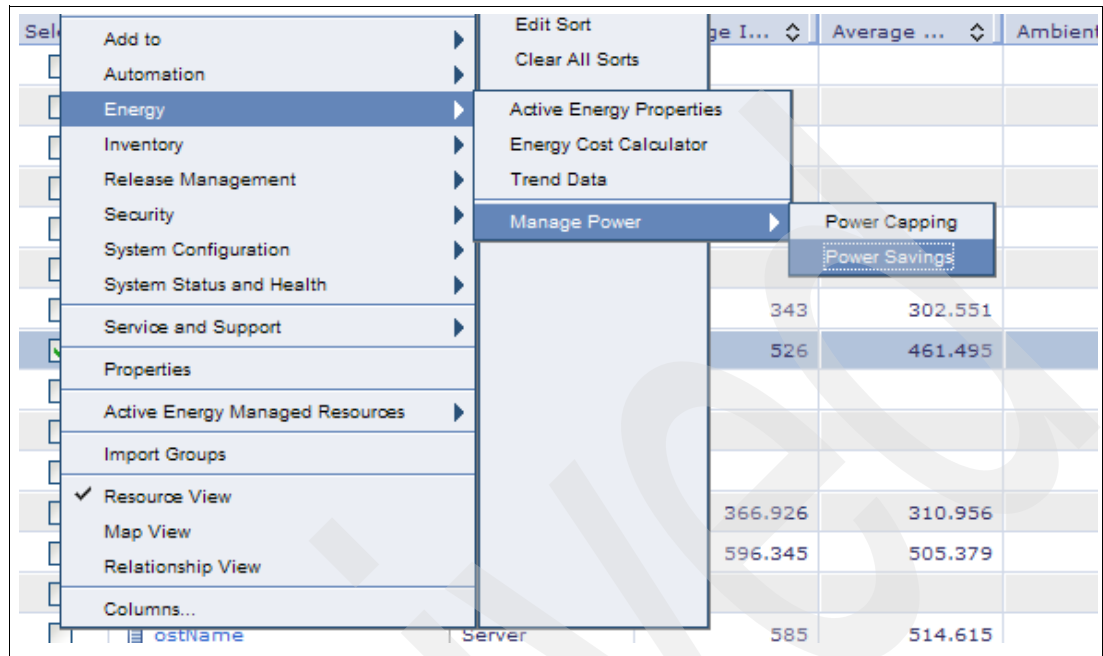


Figure 5-31 Managing power

5. Select **Dynamic power savings**. Choose from two options: Favor Power mode and Favor Performance mode, as shown in Figure 5-32.

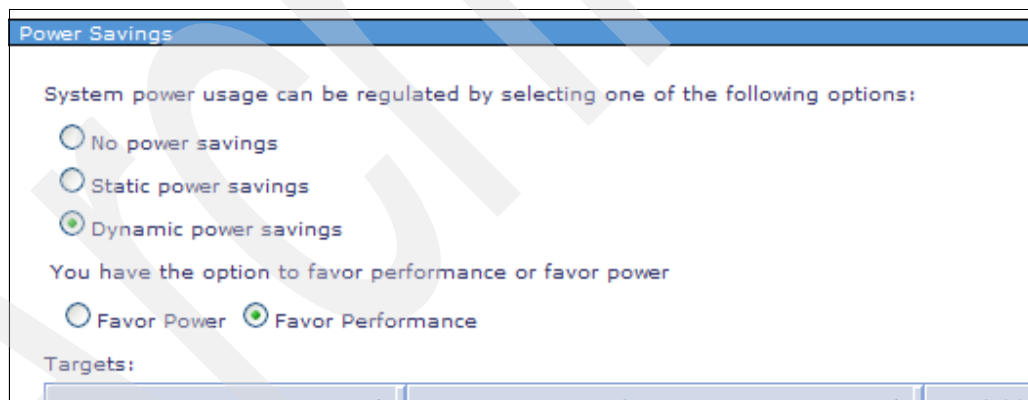


Figure 5-32 Power saving mechanisms

You have completed the setup for the AEM. To view the data and the graphical representation of the power consumed, as well the CPU frequency, use the Trend Data option. Select **Actions** → **Energy** → **Trend Data**, as shown in Figure 5-33.

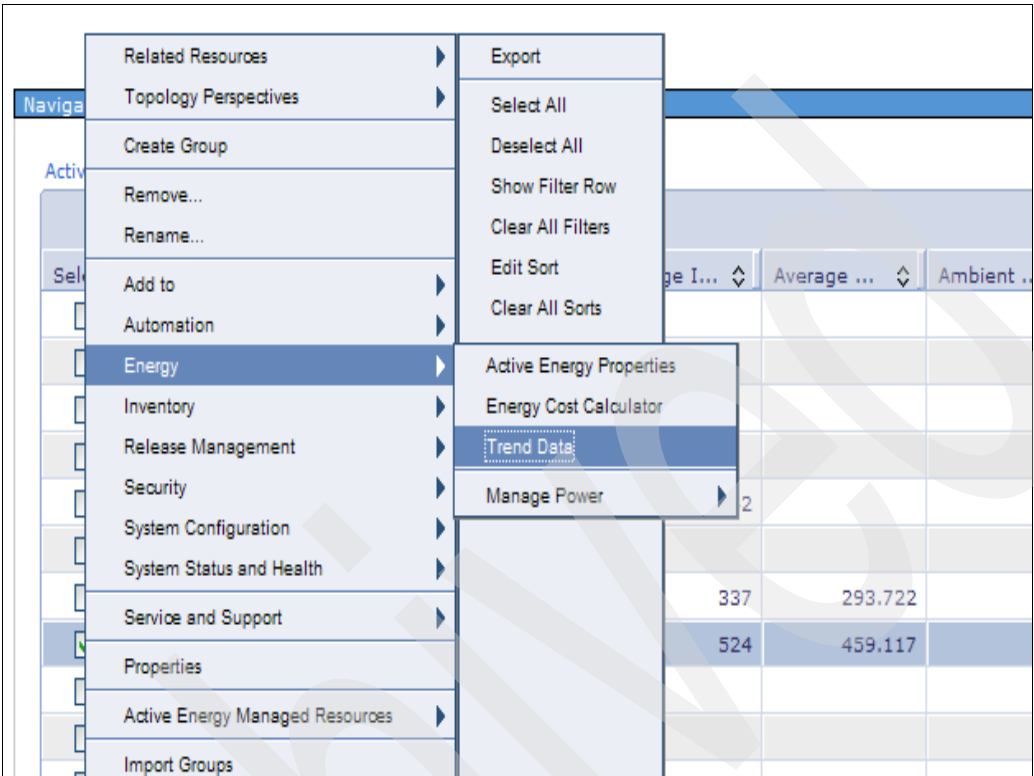


Figure 5-33 Selecting to view the Trend Data of the server

Figure 5-34 shows the variation in the Maximum Input Power when in Favor Performance mode compared to Favor Power mode.

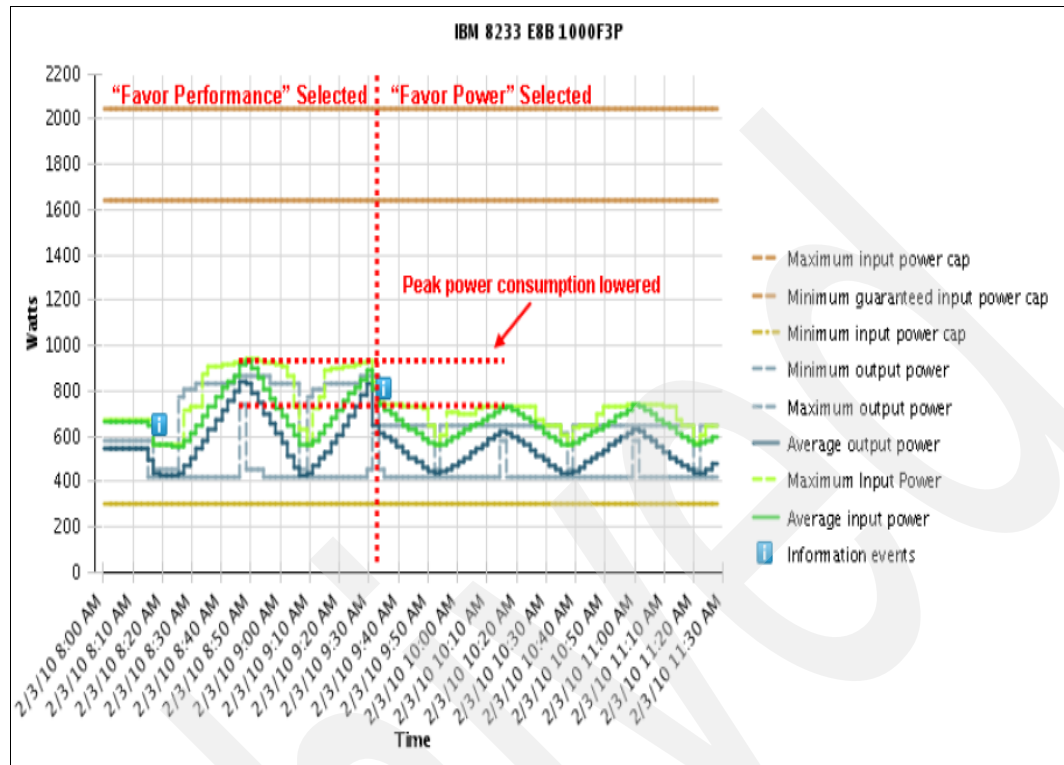


Figure 5-34 Shows the peak power consumption lowered

Figure 5-35 shows the variation in the Processor Frequency in the Favor Power mode and the Favor Performance mode.

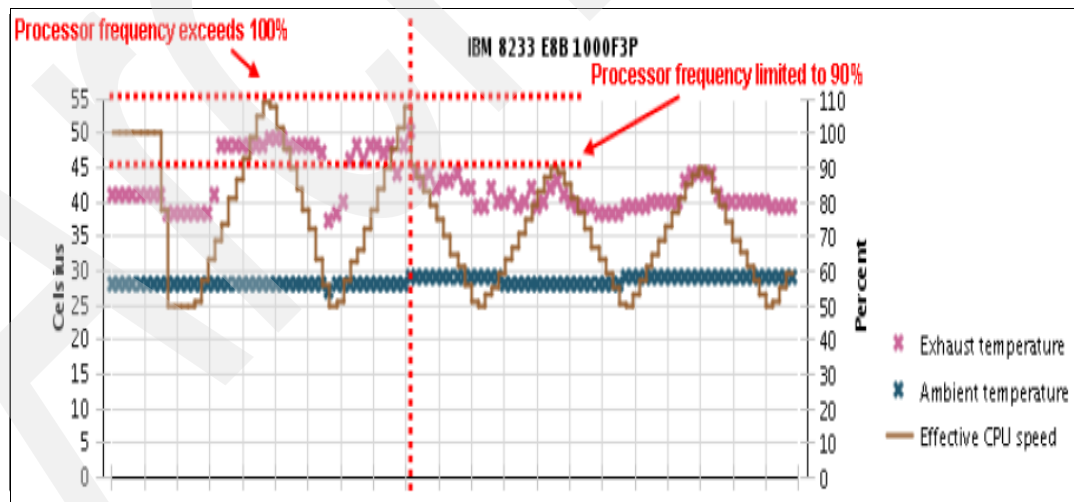


Figure 5-35 Variation of processor frequency

These features of AEM help to save energy.

For more information: For further details, refer to the *IBM Systems Director Active Energy Manager Installation and User's Guide*, which is available at the following website:
http://publib.boulder.ibm.com/infocenter/director/v6r2x/topic/com.ibm.director.aem.helps.doc/frb0_aem4.3_docs_user.pdf

5.4 High availability Systems Director management consoles

High availability of the server and its applications is maintained through the following solutions:

- ▶ PowerHA
- ▶ Workload partition (WPAR)

The Systems Director provides management consoles to configure and manage both PowerHA and WPAR. Both solutions are available as an additional plug-ins to the Systems Director.

The following references provide details about how to install and configure these plug-ins to manage high availability.

For more information:

- ▶ PowerHA Plug-in: See Chapter 12, "Creating and managing a cluster using IBM Systems Director" in *IBM PowerHA SystemMirror 7.1 for AIX*:
<http://www.redbooks.ibm.com/redbooks/pdfs/sg247845.pdf>
- ▶ WPAR Plug-in:
http://publib.boulder.ibm.com/infocenter/director/v6r1x/index.jsp?topic=wparlpp_210/wparlpp-overview.html

Requirements:

- ▶ The PowerHA plug-in needs AIX 7.1 and PowerHA7.1 or later
- ▶ WPAR requires AIX Version 6.1 with the 6100-02 TL or later

Archived

Scenarios

This chapter provides sample scenarios that show the various configurations in which IBM Power Systems high-end servers can participate. The scenarios show the Power Systems flexibility, high availability, reliability, availability, and serviceability (RAS) capabilities, and the ease of administration with the IBM Systems Director.

In this chapter, we discuss the following topics:

- ▶ Hot node add and repair
- ▶ Hot GX adapter add and repair
- ▶ Live Partition Mobility (LPM) using the HMC and SDMC
- ▶ Active migration example
- ▶ Building a configuration from the beginning
- ▶ LPM and PowerHA

6.1 Hot node add and repair

The CEC Hot Add Repair Maintenance (CHARM) functions provide the ability to add/upgrade system capacity and repair the Central Electronic Complex (CEC), including processors, memory, GX adapters, systems clock, and service processor without powering down the system. You can obtain more detailed information about CHARM in 4.3, “CEC Hot Add Repair Maintenance (CHARM)” on page 121.

In this section, we show the configuration and repair steps using CHARM to help you understand the operation of hot node add and repair, and hot GX adapter add and repair through a simple scenario.

Physical environment of the test system

Table 6-1 shows the hardware specifications and the location code for each drawer in the test system.

Table 6-1 Hardware specification for the test system on a Power780

	Processor Number	Memory size	Drawer serial number
First CEC drawer	16	128 GB	DBJH613
Second CEC drawer	16 (CoD processor)	128 GB (CoD memory)	DBJG781

6.1.1 Hot node add

In this scenario, the second CEC drawer is added to a single CEC drawer using the hot node add feature, as shown in Figure 6-1. Our hot node add and repair is tested under the control of a single Hardware Management Console (HMC). We do not provide you with all the steps, but we show specific, useful steps.

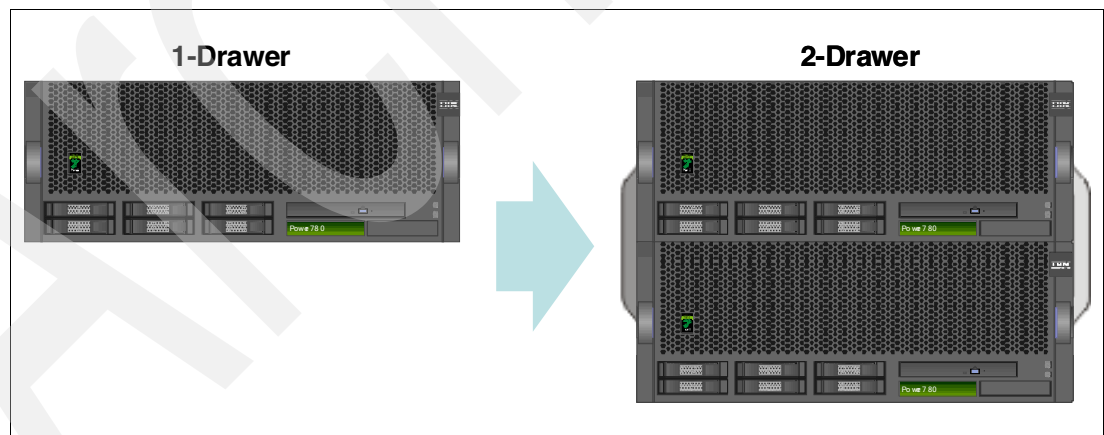


Figure 6-1 Hot node add to a single CEC drawer

Figure 6-2 shows the processor number before adding the second CEC drawer. The total number of processors is 16.

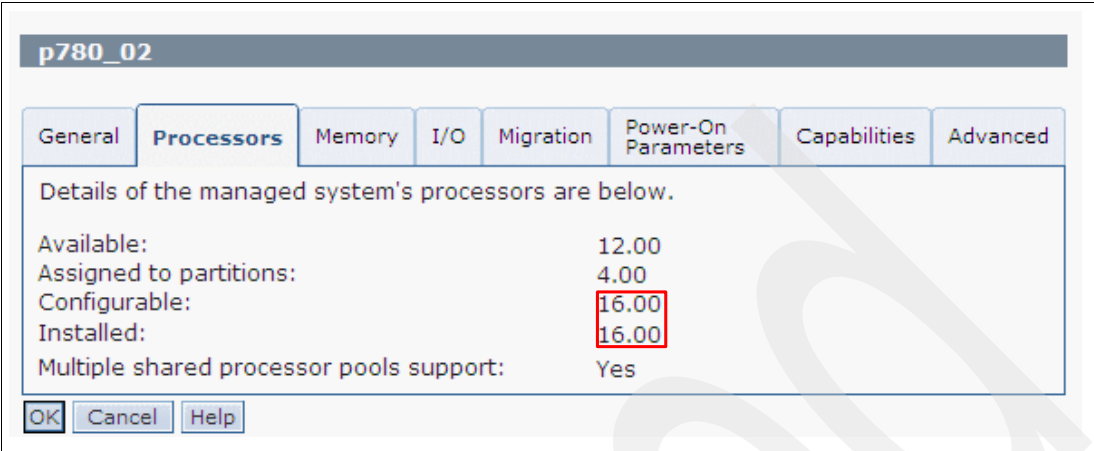


Figure 6-2 Processor number before the hot node add procedure

Figure 6-3 shows the memory size before adding the second CEC drawer. The total memory size is 128 GB.

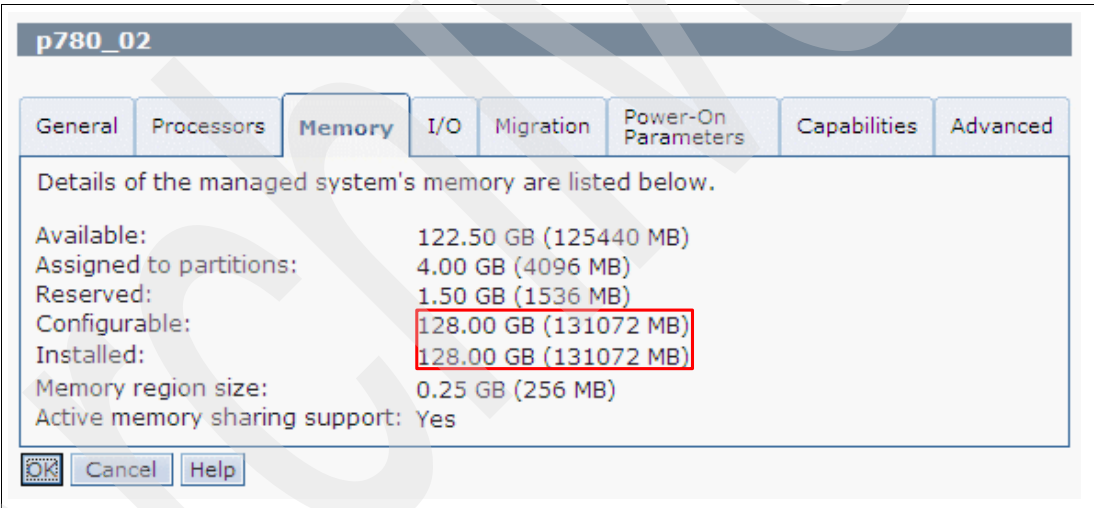


Figure 6-3 Memory size before the hot node add procedure

Figure 6-4 shows the physical I/O resources before adding the second CEC drawer.

Slot	Description	Bus	I/O Pool Id	Owner	Type
U78C0.001.DBJH613	P2-C8-T5 Universal Serial Bus UHC Spec	512	Unassigned	Unassigned	
U78C0.001.DBJH613	P2-C1 Fibre Channel Serial Bus	516	Unassigned	vios1_p780_2	
U78C0.001.DBJH613	P2-C2 Ethernet controller	517	Unassigned	Unassigned	
U78C0.001.DBJH613	P2-C3 Empty slot	518	Unassigned	Unassigned	
U78C0.001.DBJH613	P2-C4 Empty slot	519	Unassigned	Unassigned	
U78C0.001.DBJH613	P2-T3 RAID Controller	520	Unassigned	Unassigned	
U78C0.001.DBJH613	P2-C8-T7 Generic XT-Comptable Serial Controller	521	Unassigned	Unassigned	
U78C0.001.DBJH613	P2-C5 Fibre Channel Serial Bus	524	Unassigned	vios2_p780_2	
U78C0.001.DBJH613	P2-C6 Ethernet controller	525	Unassigned	Unassigned	
U78C0.001.DBJH613	P2-C9-T2 PCI-E SAS Controller	526	Unassigned	Unassigned	
U78C0.001.DBJH613	P2-C9-T1 PCI-E SAS Controller	527	Unassigned	Unassigned	

Total: 11 Filtered: 11

Figure 6-4 Physical I/O resources before the hot node add procedure

Prerequisites for the hot node add procedure

Prior to adding the second CEC drawer, check that the prerequisites are met. Table 6-2 shows the system firmware and the HMC levels for the hot node add procedure for the Power 780.

Table 6-2 System firmware and HMC levels for the hot node add procedure for the Power 780

	Minimum recommended level	Test system level
System firmware	AM720_064 or later	AM720_090
HMC	V7R7.2.0 + MH01235	V7R7.2.0.1

If the prerequisites are met, continue with the preparation steps.

Figure 6-5 shows that the logical partition (LPAR) is running on the node during the CHARM operation. We added a workload using over 90% of the processor during the hot node repair and succeeded without any problem. However, we strongly advise that you quiesce a critical application prior to starting the hot node repair.

Select	Name	ID	Status	Processing Units	Memory (GB)	Active Profile
<input type="checkbox"/>	vios1_p780_2	1	Running	2	2	node_hot_add
<input type="checkbox"/>	vios2_p780_2	2	Not Activated	2	2	default

Max Page Size: 500 Total: 2 Filtered: 2 Selected: 0

Figure 6-5 Notice that the LPAR is running during hot node add procedure

Important: We strongly advise that all scheduled hot adds, upgrades, or repairs are performed during off-peak hours.

You must move all critical business applications to another server using Live Partition Mobility (LPM), if available, or quiesce critical applications for hot node add, hot node repair, hot node upgrade, and hot GX adapter repair.

Follow these steps to add the hot node:

1. In the navigation pane, select **Systems Management**.
2. Select Add Enclosure by selecting the server → **Serviceability** → **Hardware** → **MES Tasks** → **Add Enclosure** (refer to Figure 6-6).

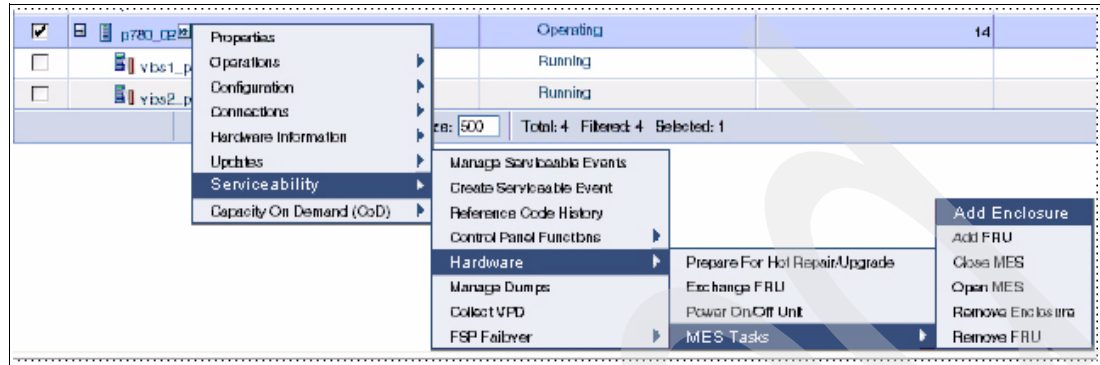


Figure 6-6 Add enclosure

3. Select a machine type - model to add. Click **add** (refer to Figure 6-7).

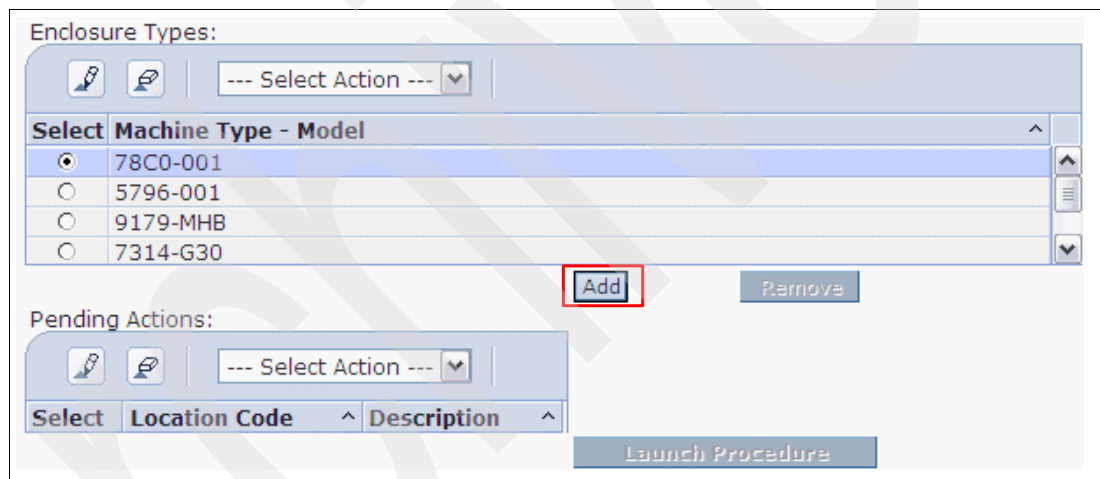


Figure 6-7 Step to select an enclosure type to add

4. Click **Launch Procedure**. Read every step carefully. Click **Next**.

5. At this step (refer to Figure 6-8), connect the SMP and the FSP cable between the first CEC drawer and the second CEC drawer. Click **Next**.

PBMesUseWciiInstallProcedure - p780_02

This MES upgrade procedure must be performed using the hardware mechanical installation instructions provided for this MES.

Note: Record the following information for reference while performing the hardware mechanical installation procedure. Record the entire location code and serial number. The serial number should match the label on the drawer being serviced.

Fru Location Code: U78C0.001.__TMP__

Machine Type: 78C0

Model: 001

Serial Number: __TMP__

Click **Next** once you have performed the hardware mechanical installation procedure, and you are ready to continue with these instructions.

Figure 6-8 Step to connect SMP and FSP cable between two nodes

6. After a few minutes, you see the message that is shown in Figure 6-9.

PBCommonVerifySuccessful - p780_02

The verification procedure completed successfully.

No problems were detected.

Next

Figure 6-9 Successful completion of hot node add

Figure 6-10 shows the processor number after the hot node add. You can see the increased installed number of processors. The configurable number of processors is not changed, because the processors of the second CEC drawer are Capacity on Demand (CoD) processors. The total installed processor number is 32.

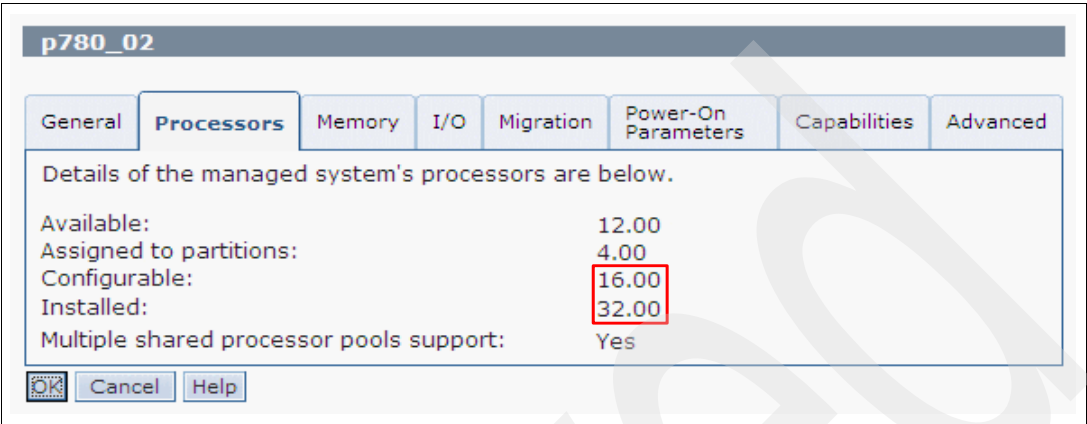


Figure 6-10 Processor number increase after the hot node add

Figure 6-11 shows the memory size after the hot node add. The configurable size of the memory is not changed, because the memory of the second CEC drawer is CoD memory. The total memory size is 256 GB.

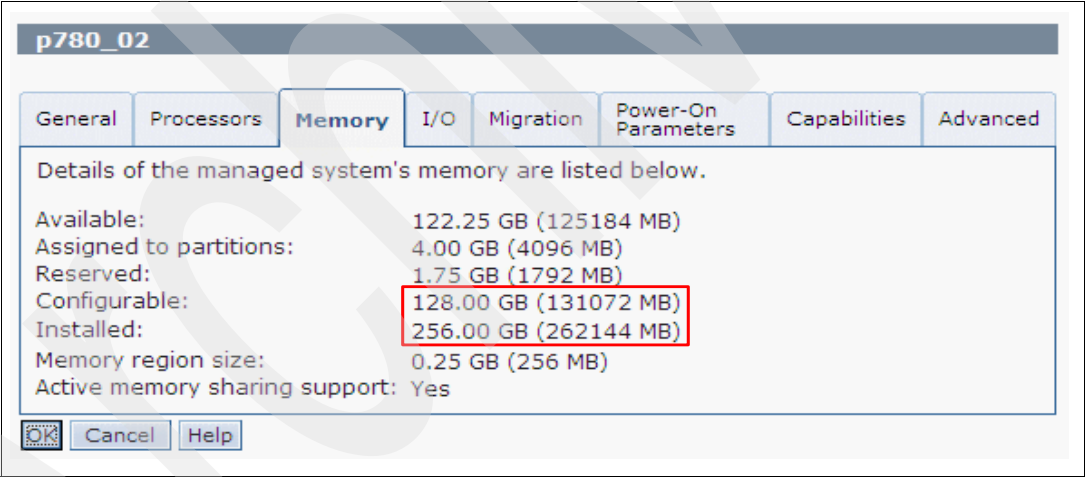


Figure 6-11 Memory size increase after the hot node add

Figure 6-12 shows the physical I/O resources after the hot node add. You can see the additional I/O location codes.

Slot	Description	Bus	I/O Pool Id	Owner	Type
U78C0.001.DBJH613-P2-C5	Fibre Channel Serial Bus	524	Unassigned	vios2_p780_2	
U78C0.001.DBJH613-P2-C6	Ethernet controller	525	Unassigned	Unassigned	
U78C0.001.DBJH613-P2-C9-T2	PCI-E SAS Controller	526	Unassigned	Unassigned	
U78C0.001.DBJH613-P2-C9-T1	PCI-E SAS Controller	527	Unassigned	Unassigned	
U78C0.001.DBJG781-P2-C8-T5	Universal Serial Bus UHC Spec	576	Unassigned	Unassigned	
U78C0.001.DBJG781-P2-C1	Fibre Channel Serial Bus	580	Unassigned	Unassigned	
U78C0.001.DBJG781-P2-T3	RAID Controller	584	Unassigned	Unassigned	
U78C0.001.DBJG781-P2-C8-T7	Unknown	585	Unassigned	Unassigned	
U78C0.001.DBJG781-P2-C2	Ethernet controller	581	Unassigned	Unassigned	
U78C0.001.DBJG781-P2-C5	Fibre Channel Serial Bus	588	Unassigned	vios2_p780_2	
U78C0.001.DBJG781-P2-C3	Empty slot	582	Unassigned	Unassigned	
U78C0.001.DBJG781-P2-C4	Empty slot	583	Unassigned	Unassigned	
U78C0.001.DBJG781-P2-C6	Ethernet controller	589	Unassigned	Unassigned	
U78C0.001.DBJG781-P2-C9-T2	PCI-E SAS Controller	590	Unassigned	Unassigned	
U78C0.001.DBJG781-P2-C9-T1	PCI-E SAS Controller	591	Unassigned	Unassigned	
Total: 22 Filtered: 22					

Figure 6-12 I/O resources increase after the hot node add

6.1.2 Hot node repair

In this scenario, we replace a memory dual inline memory module (DIMM) in the first CEC drawer.

Prerequisites for hot node repair

As already explained in CHARM operations on 4.3.2, “Hot repair” on page 123, you must have the following prerequisites installed before the hot node repair:

- ▶ The system must have two or more nodes to use the hot node repair function.
- ▶ Verify that the service processor redundancy capability is enabled.
 - In the navigation pane, select **Systems Management** → **Servers**.
 - In the work pane, select the **Server** → **Serviceability** → **FSP Failover** → **Setup**. (Refer to Figure 6-13 on page 203).
- ▶ You must configure the HMC with a redundant service network with both service processors. Refer to Figure 6-14 on page 203.

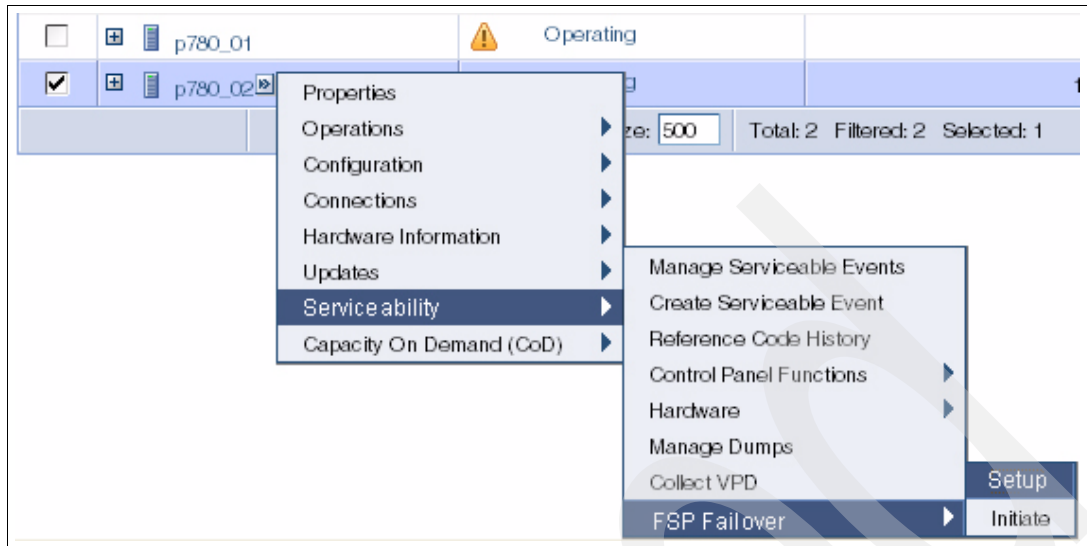


Figure 6-13 FSP Failover

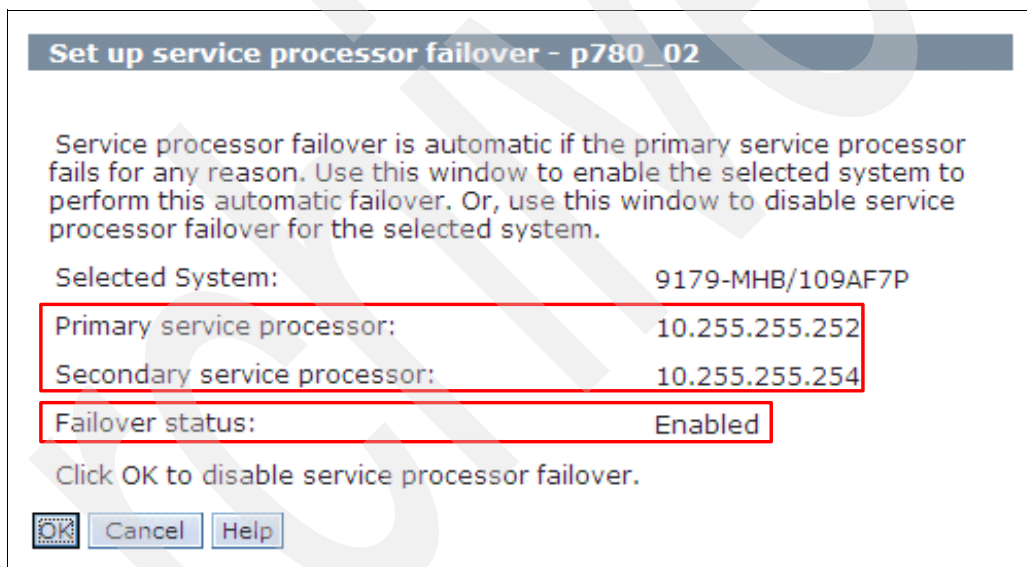


Figure 6-14 Service processor redundancy enabled

Preparing for the hot repair or upgrade (PHRU) utility

Prior to the start of the hot node repair, you need to run the Prepare For Hot Repair or Upgrade tool first. All resources that are identified by the Prepare for Hot Repair or Upgrade utility must be freed up by the system administrator prior to the start of the hot upgrade or repair procedure.

The Prepare for Hot Repair or Upgrade utility is automatically run during every service procedure requiring the evacuation of a node. This utility ensures that all requirements are addressed prior to the execution of the repair or upgrade procedure.

The test system is managed by an HMC. Follow this procedure:

1. In the navigation pane, select **Systems Management** → **Servers**.
2. In the work pane, select the server name on which the procedure will be performed. Select **Serviceability** → **Hardware** → **Prepare for Hot Repair/Upgrade**.

3. Select the base location code that contains the field-replaceable unit (FRU) to be serviced, as shown in Figure 6-15.

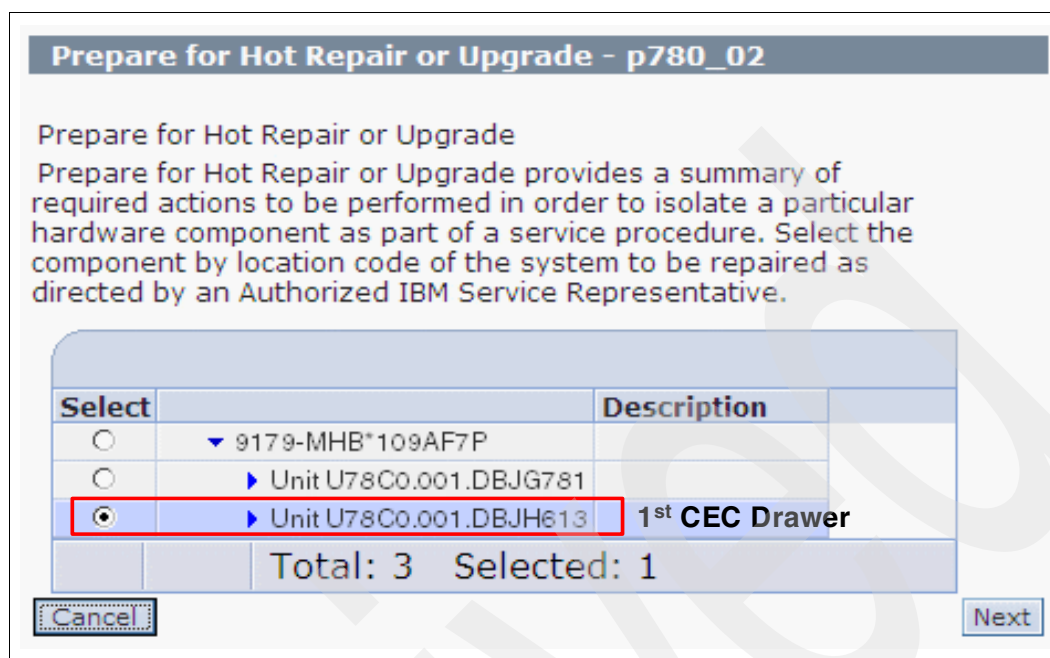


Figure 6-15 FRU selection on the Prepare for Hot Repair or Upgrade utility

4. Click **Next**.
5. Click **OK** when prompted to continue.

The Prepare for Hot Repair or Upgrade utility (PHRU) displays a window listing the set of actions that must be performed for the node evacuation to be successful, as shown in Figure 6-16.

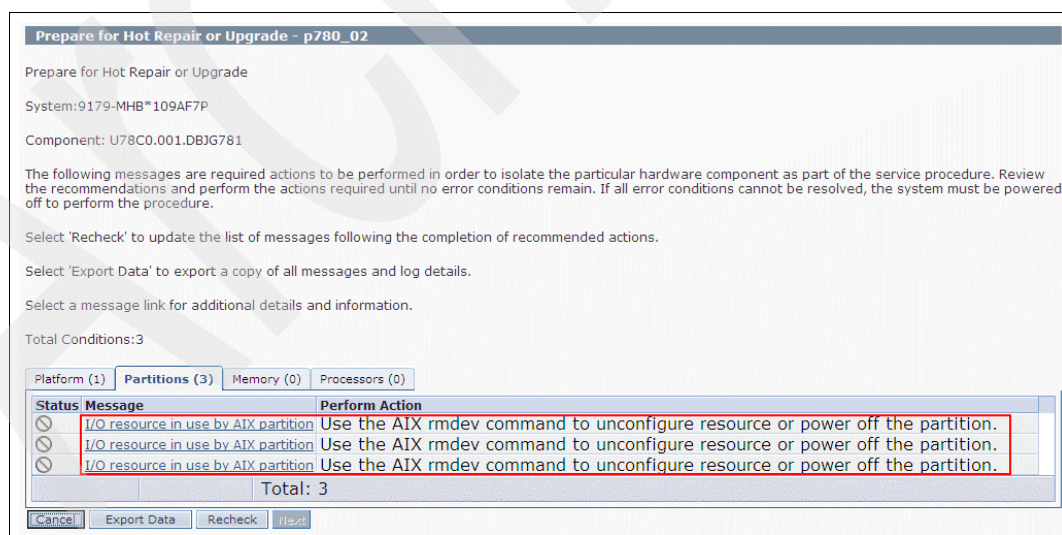


Figure 6-16 Prepare for Hot Repair or Upgrade window

Click the message text to display information about the resources that are being used, as shown in Figure 6-17.

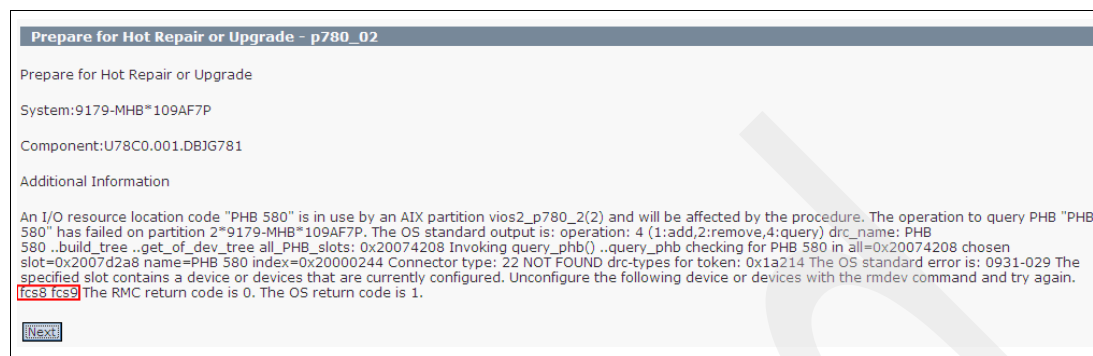


Figure 6-17 Information about the adapter to be removed prior to start the hot repair

The I/O resource is removed by using the **rmdev** command. All errors are corrected and a node can be evacuated for the host node repair.

Hot repair

During the hot node repair of replacing a memory DIMM in the first CEC drawer, the control panel must be moved to the alternate CEC drawer, as shown in the Figure 6-18.

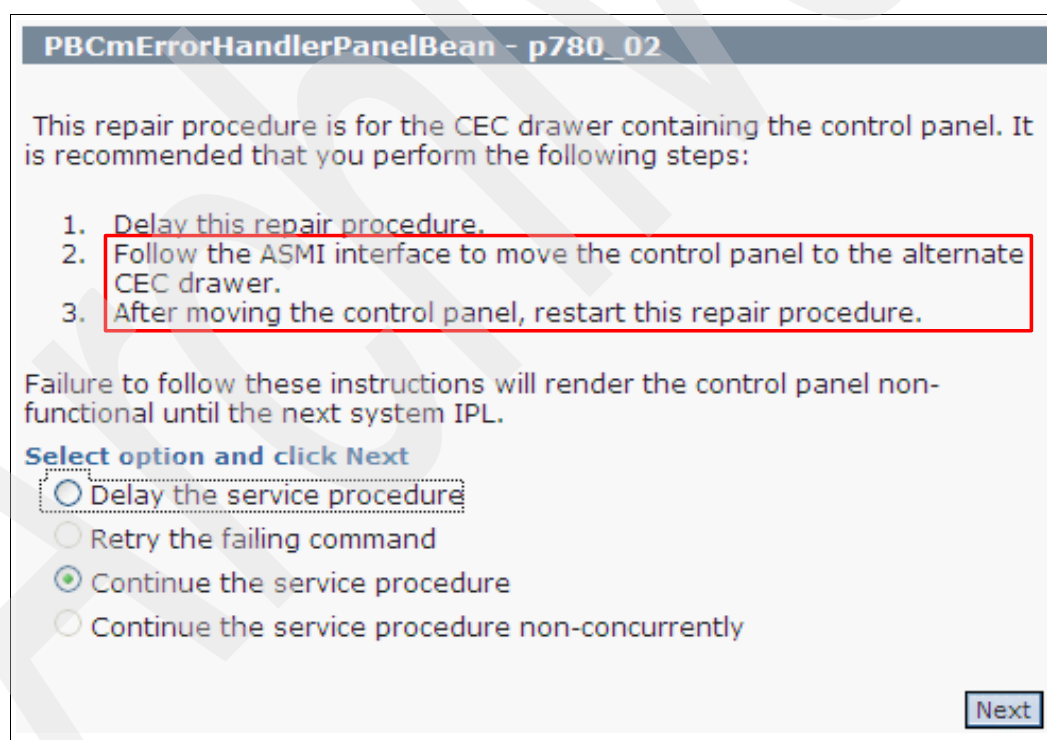


Figure 6-18 Move the control panel to the alternate CEC drawer

At this step, the control panel is removed from the first CEC drawer and installed on the second CEC drawer using the Advanced System Management Interface (ASM) function, as shown in the Figure 6-19 and Figure 6-20.

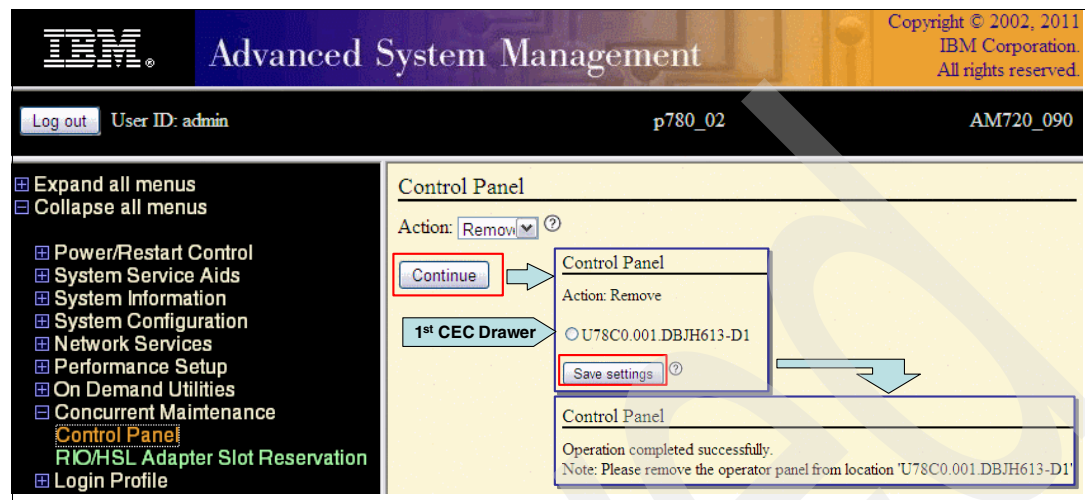


Figure 6-19 Remove control panel from the first CEC drawer using ASM

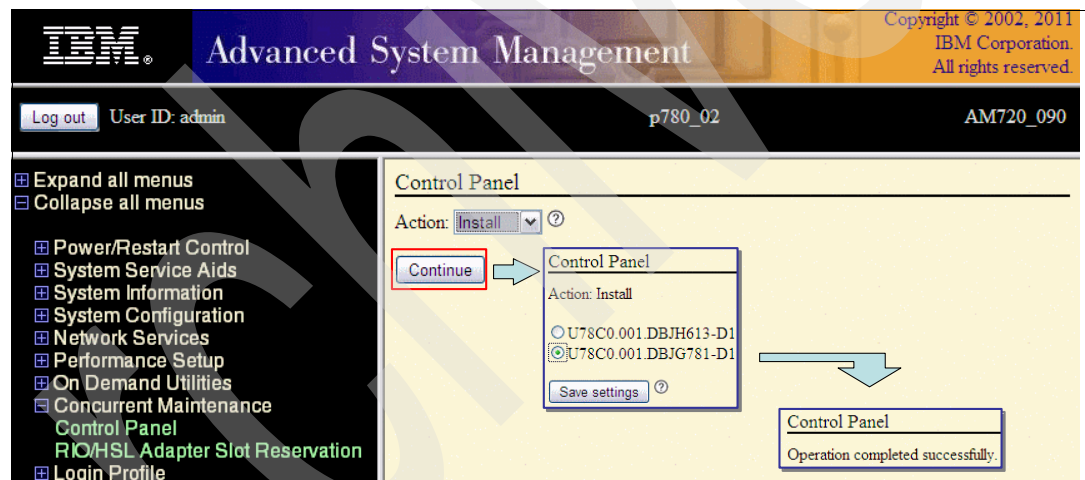


Figure 6-20 Install the control panel at the second CEC drawer using the ASM

6.2 Hot GX adapter add and repair

In this test scenario, the hot GX adapter add and repair is done under a dual management console with the HMC and the Systems Director Management Console (SDMC). We do not show all the steps, but we show specific, useful steps.

6.2.1 Hot GX adapter add

The following prerequisites are necessary for the hot GX adapter add.

Prerequisites for hot GX adapter add

Prior to adding a GX adapter, check that the prerequisites are met. Table 6-3 shows the system firmware and HMC levels prior to the hot GX adapter add in the Power 780.

Table 6-3 System firmware and HMC levels for the hot GX adapter add for the Power 780

	Minimum recommended level	Test system level
System firmware	All levels	AM720_101
HMC	V7R7.1.0	V7R7.3.0.1

If the prerequisites are met, continue with the next steps.

In the dual management console environment, you must perform all the CHARM operations from the primary management console. But, with V7R7.3.x.x of the HMC and SDMC, a new feature was added that if you start an add or repair process from the non-primary management console, you are prompted if you want to make the console at which you are performing the procedure the primary management console. The operation then tries to renegotiate the role of the primary console. If the non-primary HMC can become the primary HMC, the process allows you to continue with the procedure on this console.

Primary console: In a dual management console environment, generally the primary console is the first console that is connected. So, the primary console is not fixed, and it can be the HMC or SDMC if you use a mixed console of HMC and SDMC. This concept is the same as two HMCs or two SDMCs.

We omitted most of the hot GX adapter add steps because they are similar to the steps of the hot node add. During our hot GX adapter test, we see a message, as shown in Figure 6-21, and we select **Yes, force the local management console to become the primary**. Then, the management console becomes the primary, and we continue with the procedure on this console.

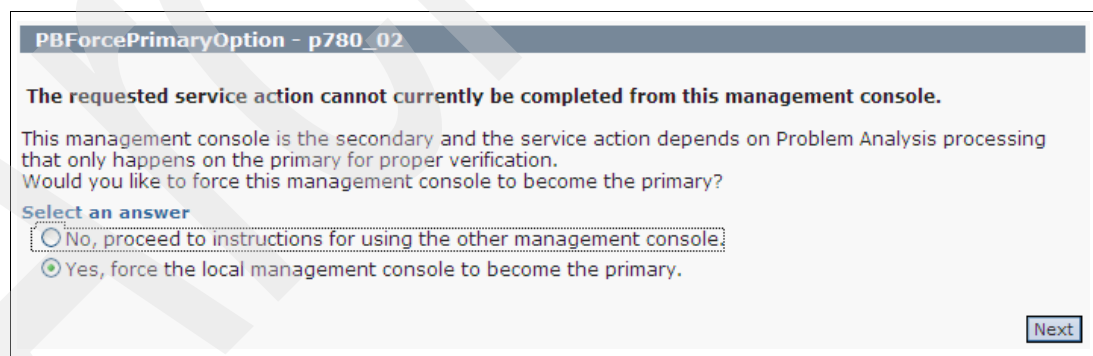


Figure 6-21 Force the local management console to become the primary

After finishing the hot GX adapter add, it is necessary to add an I/O drawer separately.

6.2.2 Hot GX adapter repair

In this scenario, we replace a GX adapter in the first CEC drawer. The prerequisites for the hot GX adapter repairs are presented in the next section.

Prerequisites for the hot GX adapter repair

Prior to repairing a GX adapter, check that the prerequisites are met. Table 6-4 shows the system firmware and HMC levels prior to the hot GX adapter repair in the Power 780.

Table 6-4 System firmware, HMC levels for hot GX adapter repair for Power 780

	Minimum recommended level	Test system level
System firmware	AM720_064 or later	AM720_101
HMC	V7R7.2.0 + MH01235	V7R7.3.0.1

If the prerequisites are met, continue with the following steps.

Follow these steps to perform the hot GX adapter repair:

1. In the navigation pane, select **Systems Management** → **Servers**.
2. In the work pane, select the server name on which the procedure will be performed. Select **Serviceability** → **Hardware** → **Exchange FRU** (see Figure 6-22).



Figure 6-22 Window for exchanging the FRU

3. Select the proper enclosure types and FRU to be replaced, as shown in Figure 6-23. After this window, read every step carefully and click **Next**.

Select an installed enclosure type to get the list of FRU types which may be installed in the selected enclosure type. Choose the FRU type to be replaced, then click Next to locate and replace the FRU.

Selected System: 9179-MHB*109AF7P

Installed Enclosure Types: System Unit, Model MHB

FRU Types: Expansion Unit, Feature Code 581
System Unit, Model MHB

--- Select Action ---

Select	Description
<input type="radio"/>	SAS Cable
<input type="radio"/>	Memory DIMM
<input type="radio"/>	SPCN Cable
<input type="radio"/>	VPD Card
<input type="radio"/>	Operator Panel
<input type="radio"/>	Cache Battery Pack
<input checked="" type="radio"/>	GX Adapter Card
<input type="radio"/>	Serial Cable
<input type="radio"/>	Thermal Power Management Device (TPMD) Card
<input type="radio"/>	RAID Enablement Card
<input type="radio"/>	HMC Cable
<input type="radio"/>	Processor Card
<input type="radio"/>	Flexible Service Processor Card
<input type="radio"/>	USB Cable
<input type="radio"/>	Power Cable
<input type="radio"/>	DASD Backplane
<input type="radio"/>	Power Supply
<input type="radio"/>	PCI Adapter Card
<input type="radio"/>	HEA Card
<input type="radio"/>	Disk Drive (DASD)
<input type="radio"/>	Battery
<input type="radio"/>	GX Cable
<input type="radio"/>	Air Moving Device
<input type="radio"/>	SMP Cable
<input type="radio"/>	FSP Flex Cable

< Back Next > Finish Cancel Help

Figure 6-23 Select a FRU for a hot repair

Even though you skip the step to identify which I/O resources are to be removed with the PHRU. Figure 6-24 shows you the step that notifies you to remove an adapter during the hot GX adapter repair.

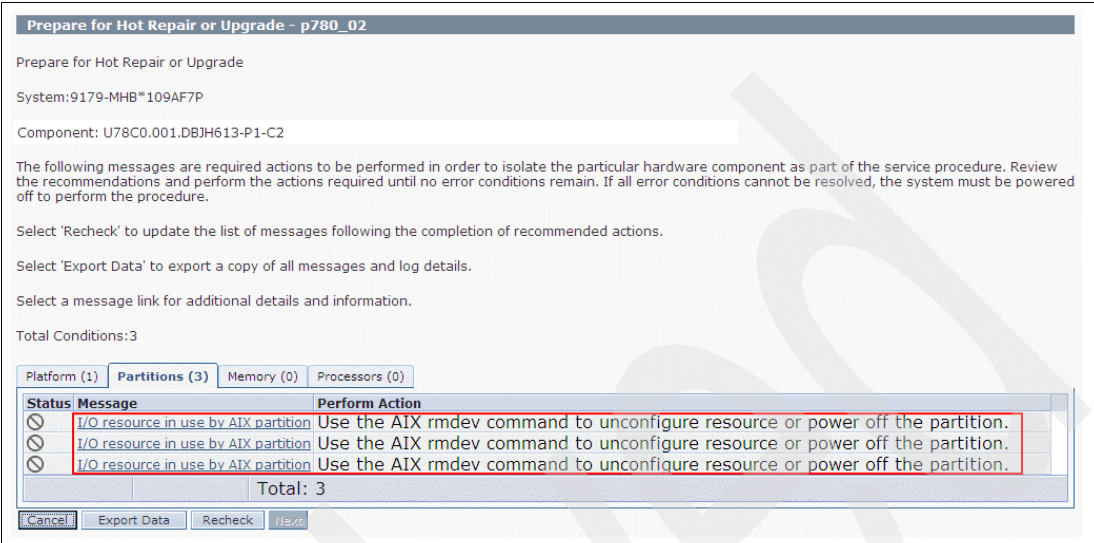


Figure 6-24 Prepare for Hot Repair or Upgrade main window

Prior to continuing to the next step, all the adapters must be removed first. To continue, read y every step carefully and follow each step.

6.3 Live Partition Mobility (LPM) using the HMC and SDMC

We use this HMC and SDMC for this scenario:

- Source: hmctest2 (172.16.20.109)
- Destination: sdmc1 (172.16.20.22)

One node is attached to the HMC, and the other node is attached to the SDMC.

6.3.1 Inactive migration from POWER6 to POWER7 using HMC and SDMC

The POWER6 server and the POWER7 server are shown in “ITSO Poughkeepsie environment” on page 395.

In this example, we perform an inactive migration from a POWER6 server to a POWER7 server. The POWER6 server is managed by an HMC, and the POWER7 Server is managed by the SDMC. We deliberately ignored the LPM prerequisites to demonstrate the importance of planning prior to the process. You notice that most of the errors relate to items that are listed in Table 3-2 on page 66. We attempt to show the resolution to several of the issues in this example. We demonstrate a migration that has met all the prerequisites in 6.4, “Active migration example” on page 214.

Follow these steps for the migration:

1. Log on to the HMC command-line interface (CLI) and confirm if the HMC can perform the remote mobility:

```
hscroot@hmctest4:~> lsiparmigr -r manager
remote_lpar_mobility_capable=1
```

Notice that a message equal to 1 means that the HMC is capable of performing the mobility.

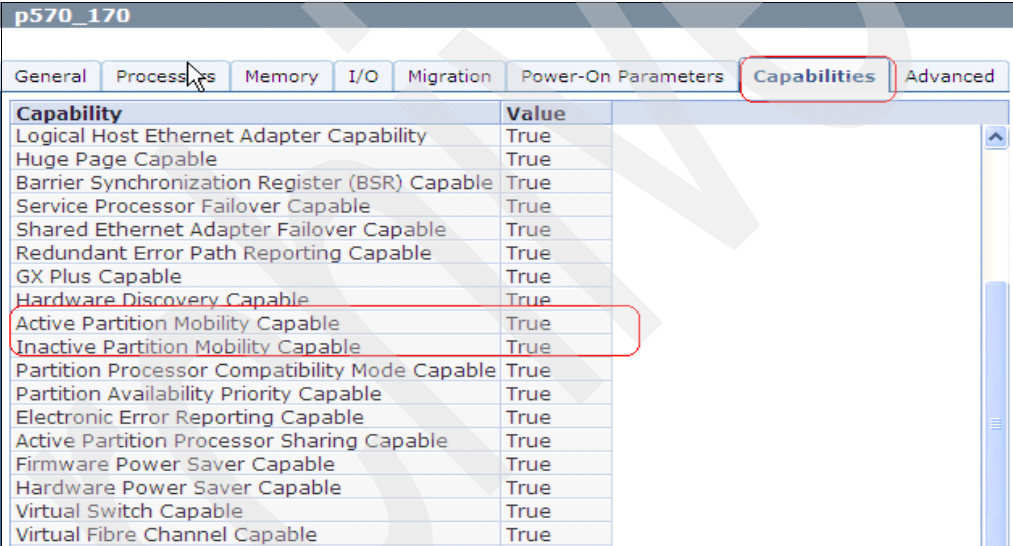
2. Log on to the SDMC and confirm that it is also capable of remote mobility:

```
sysadmin@sdmc1:~> lsiparmigr -r manager
remote_lpar_mobility_capable=1
```

3. Confirm that the managed server is mobile capable. On the HMC, follow these steps:

- a. Log on to the HMC.
- b. Click **System Management** → **Servers** → **Select properties**.
- c. Select the **Capabilities** tab.

4. Look for Mobility Capable, as shown on Figure 6-25.



Capability	Value
Logical Host Ethernet Adapter Capability	True
Huge Page Capable	True
Barrier Synchronization Register (BSR) Capable	True
Service Processor Failover Capable	True
Shared Ethernet Adapter Failover Capable	True
Redundant Error Path Reporting Capable	True
GX Plus Capable	True
Hardware Discovery Capable	True
Active Partition Mobility Capable	True
Inactive Partition Mobility Capable	True
Partition Processor Compatibility Mode Capable	True
Partition Availability Priority Capable	True
Electronic Error Reporting Capable	True
Active Partition Processor Sharing Capable	True
Firmware Power Saver Capable	True
Hardware Power Saver Capable	True
Virtual Switch Capable	True
Virtual Fibre Channel Capable	True

Figure 6-25 Capabilities of the POWER6 570

Assuming that all prerequisites are met, proceed with the rest of the steps.

- Perform a validation by selecting **Managed Server**, which lists the valid partitions. Select a partition. Click the pop-up arrow icon. Click **Operations** → **Mobility** → **Validate**, as shown on Figure 6-26.

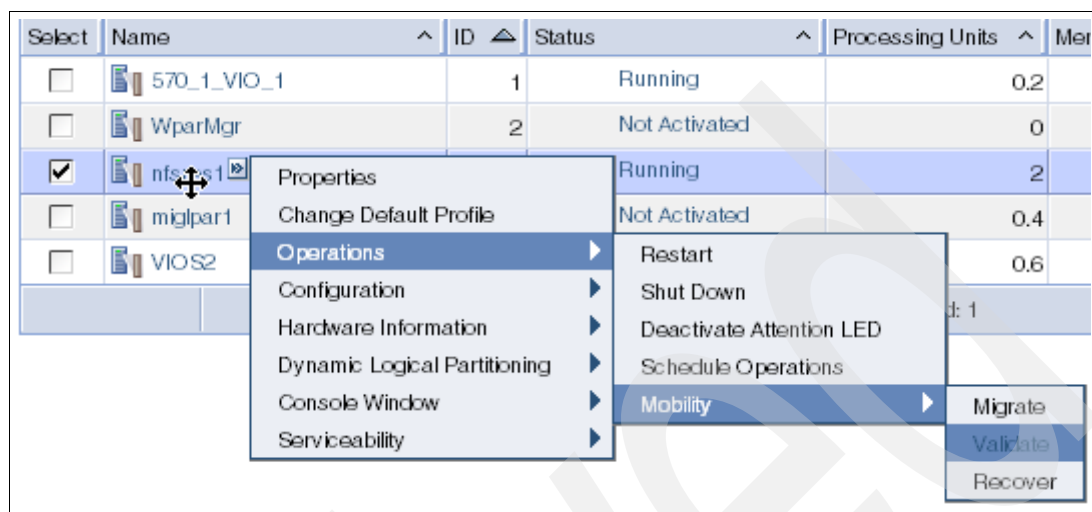


Figure 6-26 Validating the partition mobility

- Enter the remote SDMC, as requested. We are performing a remote LPM, as shown in Figure 6-27. Click **Refresh Destination System** to propagate the list of systems that are managed by the SDMC.

Partition Migration Validation - p570_170 - lpar2

Fill in the following information to set up a migration of the partition to a different managed system. Click **Validate** to ensure that all requirements are met for this migration. You cannot migrate until the migration set up has been verified.

Source system : p570_170
 Migrating partition: lpar2
 Remote HMC: 172.16.20.22
 Remote User: sysadmin
 Destination system:
 Destination profile name:
 Destination shared processor pool:
 Refresh Destination System
 Override virtual network errors when possible: ☐
 Override virtual storage errors when possible: ☐
 Virtual Storage assignments :

Select	Source Slot ID	Slot Type	Destination VIOS
<input type="checkbox"/>			

 View VLAN Settings... Validate Migrate Cancel Hel

Figure 6-27 Destination and validation window

This request fails with an error, because the Secure Shell (SSH) authorization keys are not set up between the HMC and the SDMC. To resolve this error, refer to “Setting up Secure Shell keys between two management consoles” on page 360. Figure 6-28 on page 213 shows the error.

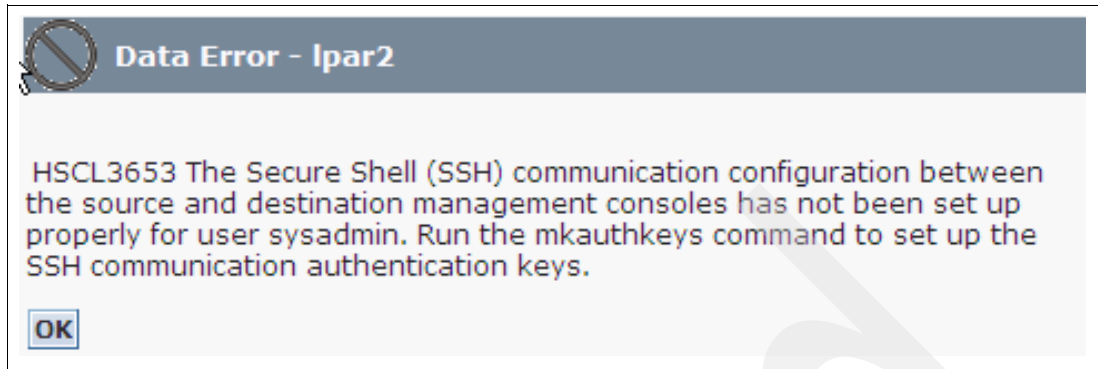


Figure 6-28 SSH authentication error

After you resolve the error, retry the validate operation. Again, you see an error code HSCLA27C, which is generic and depends on your environment. The error can relate to N_Port ID Virtualization (NPIV) or virtual Small Computer System Interface (vSCSI). Or, the error might be because the addition of a virtual adapter is not required. Another possibility is that there are Resource Monitoring and Control (RMC) communication issues between the HMC and the virtual I/O servers. Review the settings and retry the operation.

Avoid using clones: Using cloned images, such as the `alt_disk_install`, to clone servers creates RMC challenges. Avoid using cloned images. The previous error was partly caused because RMC uses a node ID to communicate with the LPARs. The node ID is stored in `/etc/ct_node_id`. If one or more LPARs, including the VIO servers on the same network, have the same `node_id`, RMC cannot confirm with which LPAR it is communicating. Either avoid `alt_disk_clone` to install another LPAR or clean the `node_id` immediately after the cloning. Other `node_id` symptoms include the inability to perform dynamic LPAR (DLPAR) operations.

Figure 6-29 shows an example of an error. The error is caused by RMC communication issues between the HMC/SDMC and the virtual I/O server.

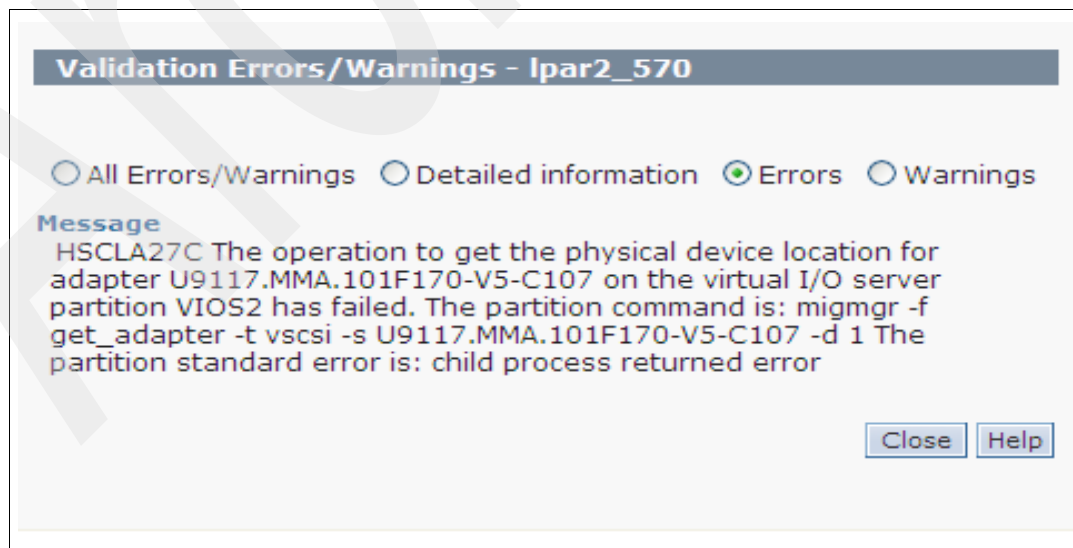


Figure 6-29 LPM validation error

When migrating to POWER7, check that the operating system level is on a TL that supports the destination server processor mode. Failure to do so might result in the migration request failing, which is shown with the error code HSCL366B in Figure 6-30.

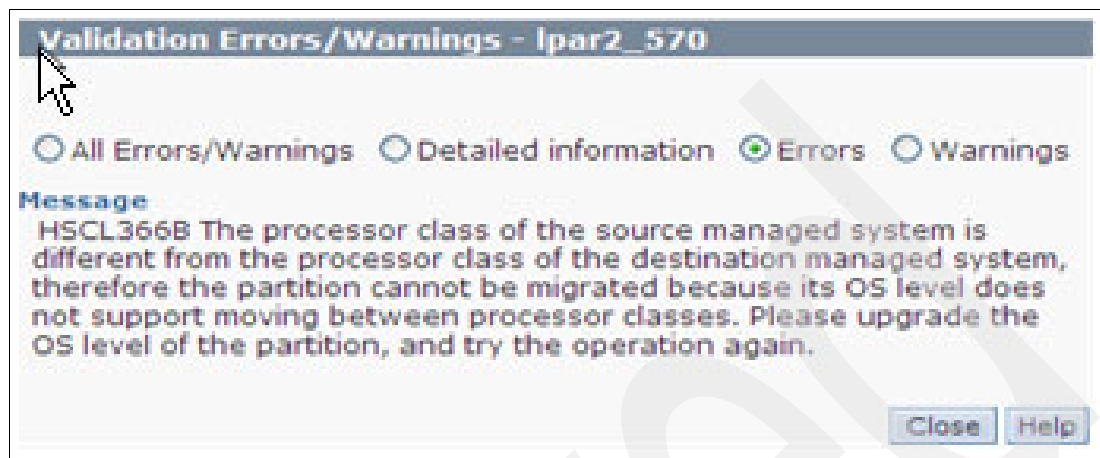


Figure 6-30 LPM processor mode failure

In this example, we tried to migrate an LPAR with AIX 6.11 TL3 to a POWER7 machine. This attempt failed due to the processor mode. To resolve the issue, we upgraded AIX to AIX 6.1 TL 6 and retried the operation. The migration checked for the virtual I/O servers. After a suitable virtual I/O server was selected, the migration completed successfully, as shown in Figure 6-31.

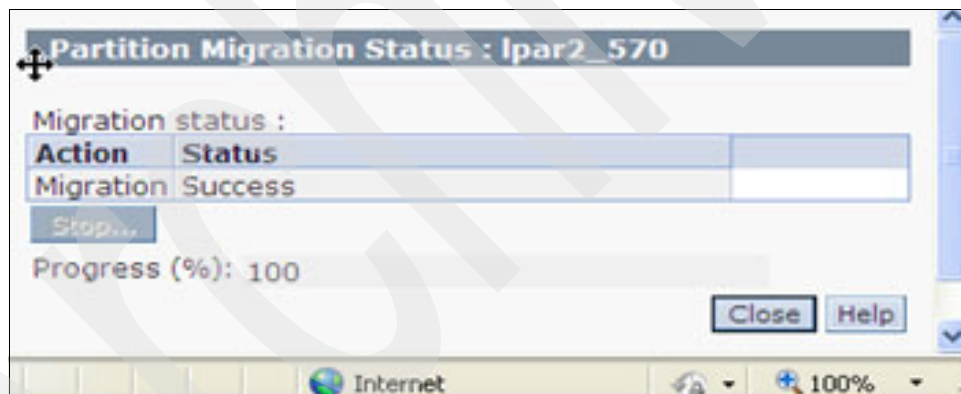


Figure 6-31 LPM migration status

6.4 Active migration example

In this example, we migrate a POWER7 LPAR to a POWER6 server that is on a separate management console. The POWER6 server is managed by the HMC. The POWER7 server is managed by the SDMC. Prior to this migration example, we performed a similar migration of the same LPAR from POWER6 to POWER7 without meeting the requirements outlined in the planning section. That migration kept failing due to requirements that were not met (see 6.3.1, “Inactive migration from POWER6 to POWER7 using HMC and SDMC” on page 210 for that example). Refer to the scenario to see which errors might be experienced. In this section, we learned from our mistakes and experiences, and we completed the prerequisites in the following example. Follow these steps:

1. Log on to the LPAR and run **lsconf**, as shown in Example 6-1.

Example 6-1 Running lsconf on the server to confirm the system model

```
# uname -a
AIX rflpar20 1 6 00F69AF64C00

# lsconf | head -15
System Model: IBM,9179-MHB
Machine Serial Number: 109AF6P
Processor Type: PowerPC_POWER7
Processor Implementation Mode: POWER 6
Processor Version: PV_7_Compat
Number Of Processors: 2
Processor Clock Speed: 3864 MHz
CPU Type: 64-bit
Kernel Type: 64-bit
LPAR Info: 6 lpar2_570
Memory Size: 3072 MB
Good Memory Size: 3072 MB
Platform Firmware level: AM720_090
Firmware Version: IBM,AM720_090
```

Example 6-1 shows that rflpar20 is on a POWER7 Server running in POWER6 mode. Before attempting a migration, ensure that the destination server is capable of handling the running mode. Refer to 7.2.3, “Processor compatibility mode” on page 251.

2. Follow these steps to initiate an LPM operation:
 - a. Log on to the SDMC.
 - b. Select **Hosts**.
 - c. Select the Virtual Server.
 - d. Click **Action** → **Operations** → **Mobility** → **Validate**.

Figure 6-26 on page 212 shows these steps using an HMC instead of the SDMC.

- At the validation window, enter the destination management system (SDMC/HMC) with the appropriate user, and select **Refresh Destination System**. This action propagates the list of managed servers that are managed by the remote SDMC/HMC. Select **Validate**, as shown in Figure 6-32.

The screenshot shows a configuration window for validating LPM between separate management consoles. It includes fields for source and destination systems, remote management console, remote user, destination system, destination profile name, destination shared processor pool, source and destination mover service partitions, wait time, and checkboxes for overriding virtual network and storage errors. A table for virtual storage assignments is at the bottom, currently empty. Numbered callouts indicate the steps: 1) Enter Remote HMC and user, 2) Refresh List from remote, 3) Select Destination Managed Server, and 4) Validate.

Source system: p780_01
 Migrating virtual server: lpar2_570
 Remote management console: 172.16.20.109
 Remote user: hscroot
 *Destination system: p570_170
 Destination profile name: p570_170
 Destination shared processor pool: Server-8233-E8B-SN106076P
 Source mover service partition:
 Destination mover service partition:
 Wait time (in min): 5
 Override virtual network errors when possible: ☐
 Override virtual storage errors when possible: ☐
 Virtual Storage assignments :
 There is no data to display.
 Total: 0, Displayed: 0
 Actions: View VLAN Settings... Validate Migrate

Figure 6-32 Validating the LPM between separate management consoles

- After validation, the capable virtual I/O servers are listed. Select the appropriate virtual I/O server and click **Migrate**.
- After the migration completes, rerun the `lsconf` command to confirm that you are on a separate server, as shown in Example 6-2.

Example 6-2 The lsconf command confirming that the LPAR has moved

```
# uname -a;lsconf | head -15
AIX rflpar20 1 6 00C1F1704C00
```

System Model: IBM,9117-MMA
Machine Serial Number: 101F170
Processor Type: PowerPC_POWER6
Processor Implementation Mode: POWER 6
Processor Version: PV_6_Compat
Number Of Processors: 2
Processor Clock Speed: 4208 MHz
CPU Type: 64-bit
Kernel Type: 64-bit
LPAR Info: 6 lpar2_570
Memory Size: 3072 MB
Good Memory Size: 3072 MB
Platform Firmware level: EM350_085
Firmware Version: IBM,EM350_085
Console Login: enable

6.5 Building a configuration from the beginning

The following scenario shows a complete HA virtual solution built from nothing. For this scenario, we use two IBM Power 780 servers and implemented Active Memory Sharing (AMS), Active Memory Expansion (AME), LPM, and PowerHA features using NPIV.

Figure 6-33 on page 218 illustrates the scenario to be configured in this section. Figure 6-33 on page 218 represents the configuration use for each server in this scenario. We create the same configuration for both of the 780 servers in our ITSO environment. We described our environment in “ITSO Poughkeepsie environment” on page 395.

Notice that this scenario is a high availability solution. However, you can increase the levels of redundancy by adding, for example, more Ethernet adapters or host bus adapters (HBAs) and additional paths to the storage area network (SAN).

When implementing your solution, remember to check that each physical Ethernet adapter is located in a separate CEC (in this scenario, we use Integrated Virtual Ethernet (IVE) adapters). Also, check that each physical Fibre Channel (FC) adapter is located in a separate CEC.

In the following sections, we describe these configurations:

- ▶ Virtual I/O server definition and installation
- ▶ HEA port configuration for dedicated SEA use
- ▶ NIB and SEA failover configuration
- ▶ Active Memory Sharing configuration
- ▶ Server-side NPIV configuration
- ▶ Active Memory Expansion (AME) configuration
- ▶ The LPM operation
- ▶ The PowerHA operation

RMC connection: The RMC connection between the virtual servers and the HMC or the SDMC are key to the implementation. You can check if it works correctly by executing the command `lspartition -dlpar` at your HMC or SDMC.

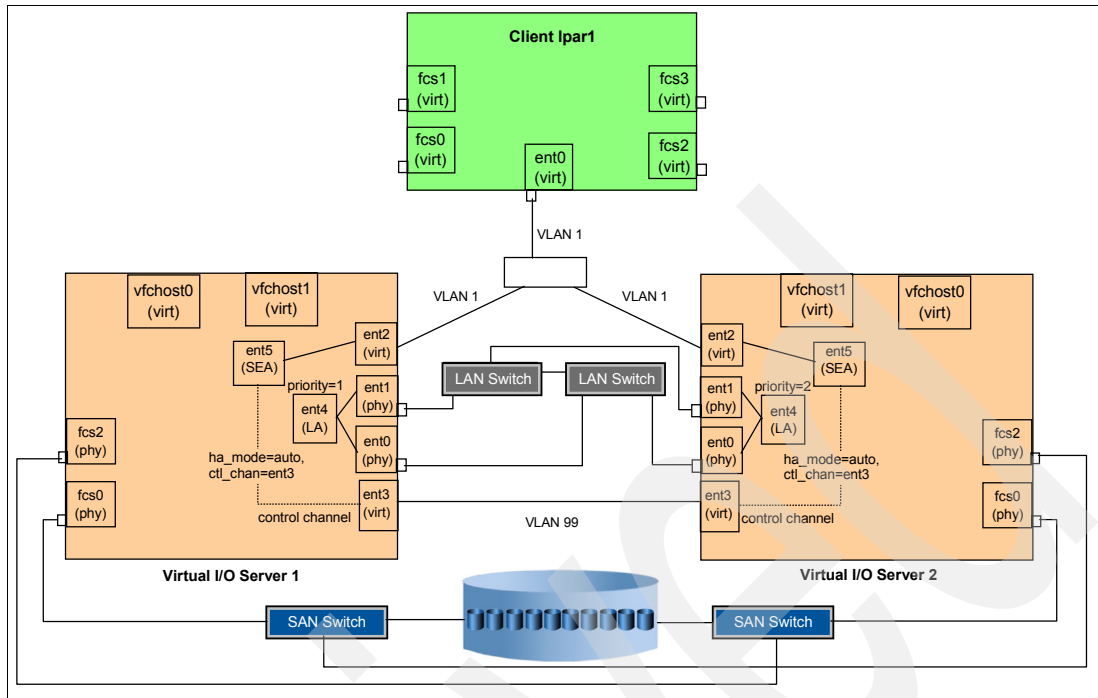


Figure 6-33 Example of a minimum configuration for high availability in one server

Differences: We reference our environment's adapter numbers in this scenario. Be aware that adapter numbers can change according to your environment.

The same network configuration needs to be done in the second server. The NPIV configuration is only performed in the first server for the client LPAR installation.

Planning: Perform exhaustive planning of your SAN zoning before starting the Power Systems configuration.

Refer to the following SAN considerations:

- ▶ Create a group for both VIO servers in each Power 780 server. AMS paging disks are assigned to the group.
- ▶ The VIO operating system's disks must be assigned only to the corresponding VIO server.
- ▶ AMS paging disks will be duplicated for each Power 780 server.
- ▶ Because we use NPIV, additional configurations are performed after we create the client virtual servers.

The following sections guide you to configure this environment. These scenarios are not intended to show you a step-by-step configuration, but they provide a consistent guide for the installation. For more details and specific tasks, refer to each specific product installation guide. We advise you to read the entire example before performing your implementation.

6.5.1 Virtual I/O servers

This section guides you to perform the installation and initial configuration of your Virtual I/O servers.

Virtual I/O server definition and installation

We define the first virtual I/O server LPAR (virtual server in the SDMC) and direct you through the installation process.

Creating the virtual I/O server LPAR profile

Important: We do not include each step for the LPAR creation, but we provide details about the key parts.

Follow these steps to create the virtual I/O server LPAR profile:

1. Log on to the SDMC.
2. In the welcome window, select the server with which you want to work. Click **Action** → **System Configuration** → **Create Virtual Server**. Then, the Create Virtual Server window opens.
3. Type the partition name and select **VIOS** in the Environment box. Click **Next**. Refer to Figure 6-34.

Create Virtual Server: p780_02

Name

This wizard helps you create and assign resources to a virtual server.

Host name: p780_02

*Virtual server name: vios1_p780_2

Virtual server ID: 1

Environment: VIOS

< Back Next > Finish Cancel

Figure 6-34 VIO server creation using SDMC 1

4. Configure the memory and processor values for your VIO (we review these values later).

5. In the Virtual Ethernet profile, as shown in Figure 6-35, configure two virtual Ethernet adapters.

Select	Adapter	Port VLAN ID	Bridge	Priority
<input checked="" type="checkbox"/>	2	1	Yes	1
<input type="checkbox"/>	3	99	No	

Figure 6-35 Virtual Ethernet adapter configuration

Both adapters are used for the Shared Ethernet Adapter (SEA) configuration.

Trunk priority: Notice that we configure the trunk priority with a value of 1 for the first VIO and with a value of 2 in the second VIO for each server. This priority helps you to configure the SEA failover feature in later steps.

In the Host Ethernet Adapter section, select the T1 adapters for the first virtual I/O server, and select the T3 adapters for the second virtual I/O server. Click **Next**.

6. In the Virtual storage adapters profile, set the Maximum number of virtual adapters to **1000**. Click **Next**. Do not configure the virtual storage adapters at this point.
7. In the Physical I/O adapters window, select the physical adapters (HBAs) that you want to configure in the VIO server.

Use separate CECs: As you did with the Host Ethernet Adapter (HEA) ports, the adapters that you assign to each virtual I/O server need to be located in separate CECs to help maximize availability.

8. In the summary window, review the settings and click **Finish**.
9. After the virtual server is created, edit the virtual server profile to adjust the following values:
 - Check the processor values for minimum, desired, and maximum processing units and virtual processors.
 - Configure the partition as uncapped with a weight of 255.
 - Configure the memory values that you want.
 - In the Optional settings window, select **Enable connection monitoring**.
10. Activate your new virtual I/O server LPAR, and install the latest available virtual I/O server image.
11. Perform the usual virtual I/O server configuration, for example, with date and time settings.
12. Install the multi-path I/O (MPIO) driver according to your environment. Consult the System Storage Information Center (SSIC) website for more information about the available multi-pathing drivers.

Configurations: Repeat this configuration for the second virtual I/O server in this server and for the two virtual I/O servers in your second POWER7 server.

6.5.2 HEA port configuration for dedicated SEA use

After creating your virtual I/O servers, you need to configure the corresponding IVE ports in promiscuous mode. As explained in 2.9, “Integrated Virtual Ethernet” on page 48, this action helps us to configure an SEA using the IVE ports.

To perform this configuration, refer to the following steps:

1. Log on to the SDMC.
2. In the Welcome window, select the server with which you want to work.
3. Click **Action** → **Hardware Information** → **Adapter** → **Host Ethernet**, and the Host Ethernet Adapter window opens.
4. Select the HEA port that you want to configure and click **Configure** (in this scenario, we use port T1 for the first VIO and port T3 for the second VIO on both CECs).
5. In the Promiscuous virtual server field, select the virtual server that will use the HEA port. In this scenario, we put the first virtual I/O server as the virtual server for each T1 adapter and the second virtual I/O server as the virtual server for each T3 adapter.
6. Select **Enable flow control**.
7. Click **OK**.

Important: Review the IVE documentation to determine the best values for the remaining configuration parameters according to your network infrastructure.

8. Repeat this process for the HEA ports that you configure in the VIO servers and plan to use as part of the SEA adapters.

6.5.3 NIB and SEA failover configuration

In this section, we explain the Network Interface Backup (NIB) and Shared Ethernet Adapter (SEA) configurations.

In this configuration example, we use NIB as the aggregation technology for network redundancy. Follow these steps:

1. Check the adapter numbers for your physical Ethernet adapters. You create the NIB adapter using the two physical adapters, which in this case are ent0 and ent1.
2. Create your NIB adapter by executing the commands that are shown in Example 6-3. The command output is shown in Figure 6-12 on page 202.

Example 6-3 NIB adapter configuration

```
$ mkvdev -lnagg ent0
ent4 Available
en4
et4
$ cfglnagg -add -backup ent4 ent1
```

3. Using the virtual slot placement, identify the adapter in VLAN1 and the adapter in VLAN99. The adapter in VLAN99 is the control channel adapter. The adapter in VLAN1 is the default adapter of the SEA.
4. Define the SEA adapter in the virtual I/O server using the **mkvdev** command, as shown in Example 6-4 on page 222.

Example 6-4 Shared Ethernet Adapter with failover creation

```
$ mkvdev -sea ent4 -vadapter ent2 -default ent2 -defaultid 1 -attr ha_mode=auto
ctl_chan=ent3 netaddr=172.16.20.1 largesend=1
ent5 Available
en5
et5
```

5. Configure the IP address for the SEA adapter using the **cfgassist** command, as shown in Figure 6-36. Remember to use a separate IP address for the VIO servers.

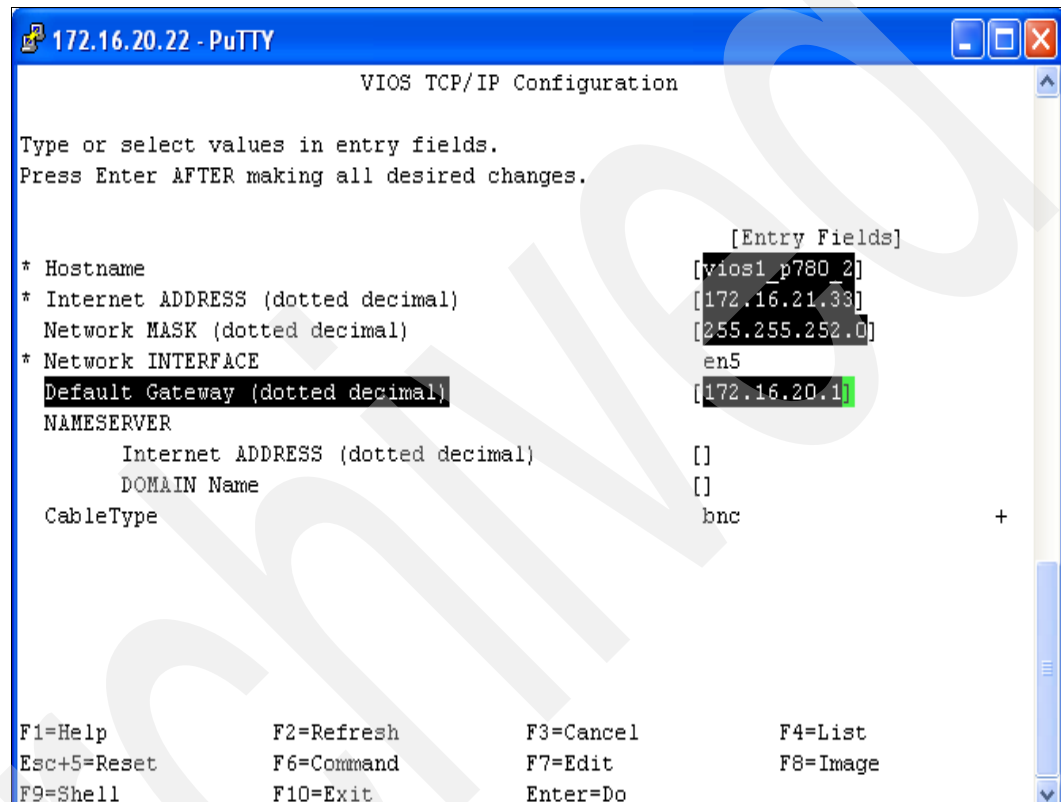


Figure 6-36 SEA IP address configuration using the **cfgassist** menu

6. Dynamically configure the virtual server for each virtual I/O server to be a *mover service partition*. Overwrite the virtual I/O server LPAR profile to make the change permanent.

After you perform this configuration on both VIO servers in each server, you have addressed the needs for the virtual I/O server and Ethernet configurations, and you can continue with the rest of the configuration steps.

Note: After you configure the SEA adapters in both VIO servers and install a client LPAR, test the adapter failover before going into production. Refer to section 5.1.2 of *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940-04, for more information about this test.

6.5.4 Active Memory Sharing configuration

In this section, we describe the process for the AMS configuration:

- ▶ Creating the paging devices on the Virtual I/O servers
- ▶ Creating the shared memory pool using the SDMC

Console: For this part of the example, we use the SDMC. Or, you can also use the HMC.

Creating the paging devices on the Virtual I/O servers

Because we are deploying a redundant configuration, we need our paging devices located in the SAN environment and accessible to both VIO servers in the IBM Power server. Also, because we plan to use LPM, the paging spaces must be duplicated in each server. For example, if we have two 8 GB and two 4 GB paging devices on server 1 (available to both virtual I/O servers in the server), we need to have another two 8 GB and two 4 GB paging devices at server 2, and assign them to both VIO servers. Follow these steps to create the paging devices:

1. As shown in Figure 6-37, create your paging spaces in the SAN and assign them to both VIO servers in the first server and to both VIO servers in the second server.

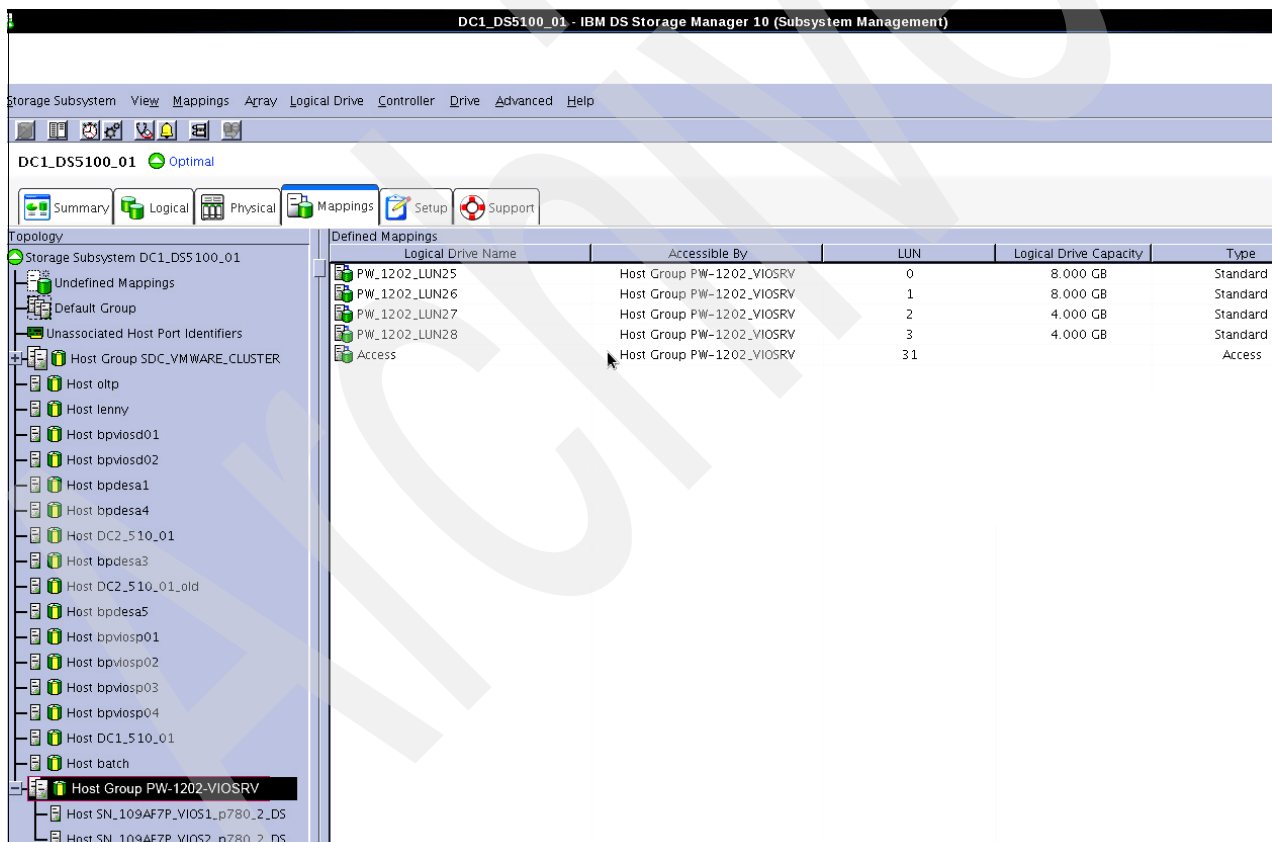


Figure 6-37 Paging space creation

2. Configure the device on each VIO server by executing the **cfgdev** command.
3. On your SDMC Welcome page, select the server where you will create the shared memory pool.

4. As shown in Figure 6-38, select **Action** → **System Configuration** → **Virtual Resources** → **Shared Memory Pool Management**.



Figure 6-38 Shared memory pool creation with SDMC (page 1 of 4)

5. The Create Shared Memory Pool window opens, as shown in Figure 6-39.
 - Specify the Maximum Pool Size and the Pool Size.
 - Check that both Virtual I/O servers appear as paging devices.
 - Click **Add Paging Devices**.

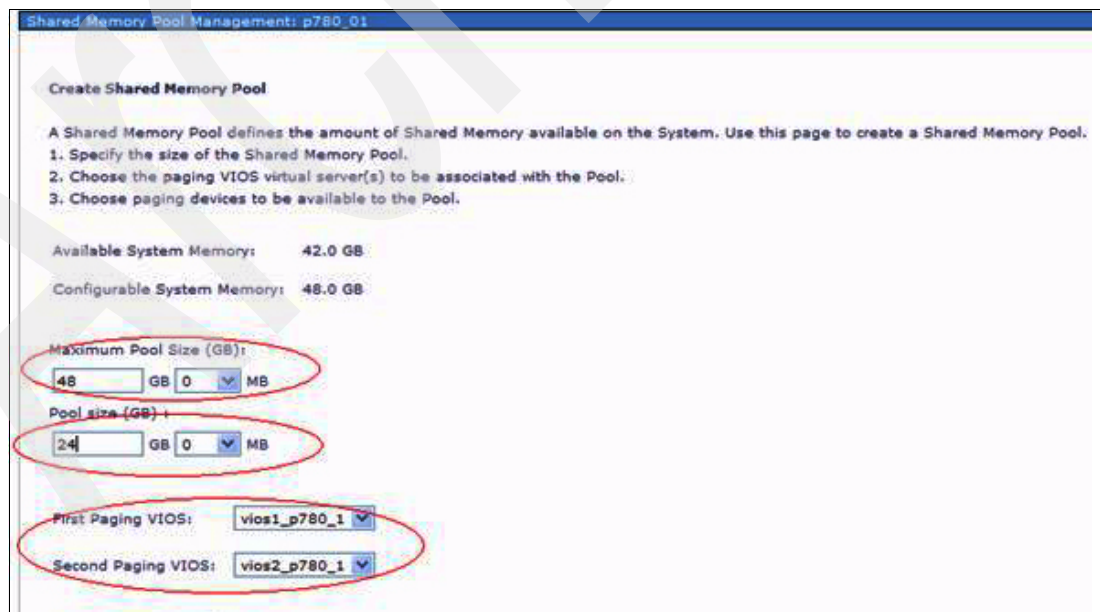


Figure 6-39 Shared memory pool creation with SDMC (page 2 of 4)

6. The Add Paging Devices window opens.

- Select **Physical** as the Device Type.
- Click **View Devices**.

As shown in Figure 6-40, the SDMC detects the available disks and presents them in a grid. After the available disks appear, select the devices that you plan to add into the shared memory pool and click **OK**.

Add Paging Devices

To display the available paging space devices in the device lists, you must first populate the filter parameters and then click View Devices. You can list all available paging space devices by selecting All as the device type, or you can narrow your search by selecting a device type and populating the maximum size, or minimum size.

Device Type: ▼

Maximum Size (GB):

Minimum Size (GB):

Choose from the following list of devices. You can choose more than one paging device to be added to the pool. Paging space devices should be assigned to only one shared memory pool at a time. You should not assign a paging space device to this shared memory pool if it is already assigned to another shared memory pool on another system. After you have made your selections, select the OK button to assign the selected devices to the memory pool.

Select	VIOS Name	Device Name	Device Size (MB)	Redundancy Capable
<input checked="" type="checkbox"/>	vios1_p780_1,vios2_p780_1	hdisk1,hdisk1	8192	true
<input checked="" type="checkbox"/>	vios1_p780_1,vios2_p780_1	hdisk2,hdisk2	8192	true
<input checked="" type="checkbox"/>	vios1_p780_1,vios2_p780_1	hdisk3,hdisk3	4096	true
<input checked="" type="checkbox"/>	vios1_p780_1,vios2_p780_1	hdisk4,hdisk4	4096	true
<input type="checkbox"/>	vios1_p780_1,vios2_p780_1	hdisk7,hdisk7	10240	true

Figure 6-40 Shared memory pool creation with SDMC (page 3 of 4)

- The selected paging space devices are added to the shared memory pool, as shown in Figure 6-41.

Create Shared Memory Pool

A Shared Memory Pool defines the amount of Shared Memory available on the System. Use this page to create a Shared Memory Pool.

- Specify the size of the Shared Memory Pool.
- Choose the paging VIOS virtual server(s) to be associated with the Pool.
- Choose paging devices to be available to the Pool.

Available System Memory: 42.0 GB

Configurable System Memory: 48.0 GB

Maximum Pool Size (GB):

48 GB 0 MB

Pool size (GB):

24 GB 0 MB

First Paging VIOS: vios1_p780_1

Second Paging VIOS: vios2_p780_1

Paging space device(s):

Select	Virtual Server ID	VIOS Name	Device Name	Device Size	Device Status	Redundancy	Phys
<input type="checkbox"/>		vios1_p780_1.vic	hdisk1,hdisk1	8192		true	U78
<input type="checkbox"/>		vios1_p780_1.vic	hdisk2,hdisk2	8192		true	U78
<input type="checkbox"/>		vios1_p780_1.vic	hdisk3,hdisk3	4096		true	U78
<input type="checkbox"/>		vios1_p780_1.vic	hdisk4,hdisk4	4096		true	U78

Page 1 of 1 | Selected: 0 Total: 4 Filtered: 4

Figure 6-41 Shared memory pool creation with SDMC (page 4 of 4)

Verification: Notice that the Redundancy attribute is set to true. Both VIO servers can access the device. Also, notice that there are two 4 GB and two 8 GB paging devices. We can activate two LPARs with up to 4 GB and two LPARs with up to 8 GB memory with this pool.

- Click **OK** to create the pool.

6.5.5 NPIV planning

Because client partitions must exist before the creation of the virtual FC server adapters, you need to plan the slot assignments and start by creating the client partition. Then, you can add the server virtual adapters.

Figure 6-42 on page 227 shows the configuration that we will follow for each LPAR that we define. The slot numbers vary on the virtual I/O server side.

Table 6-5 on page 227 shows the virtual adapter placement for this configuration. We use this information to create both the client and server adapters.

Table 6-5 Virtual FC initial slot assignment

Virtual I/O server	Virtual I/O server slot	Client partition	Client partition slot
vios1_p780_1	30	lpar1(3)	20
vios1_p780_1	31	lpar1(3)	21
vios2_p780_1	30	lpar1(3)	22
vios2_p780_1	31	lpar1(3)	23
vios1_p780_1	40	lpar2(4)	20
vios1_p780_1	41	lpar2(5)	21
vios2_p780_1	40	lpar2(5)	22
vios2_p780_1	41	lpar2(5)	23

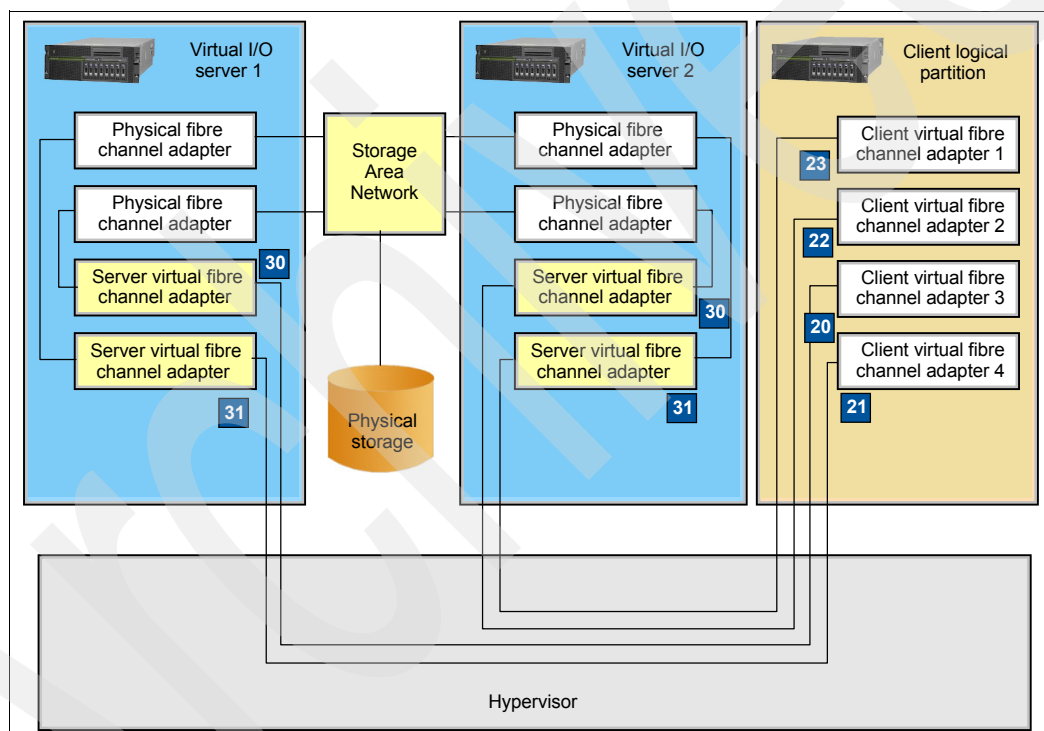


Figure 6-42 NPIV redundant configuration

6.5.6 Client LPAR creation (virtual servers)

In this section, we provide the steps to create the client LPARs or virtual servers:

1. In your SDMC Welcome window, select the server, and then, click **Action** → **System Configuration** → **Create virtual Server**.
2. In the Name window:
 - Enter the LPAR name.
 - For the environment, select **AIX/Linux**.

Important: If you click the suspend capable check box, be aware that you add 10% extra space in your paging device to activate the LPAR.

3. In the Memory window:
 - Select **Shared** for the memory mode.
 - Select the Assigned Memory (in this scenario, we create two 6 GB partitions).
4. In the Processor window, assign one processor to the LPAR.
5. In the Ethernet window, be sure that you only have one virtual Ethernet adapter that is located in VLAN 1, as shown in Figure 6-43.

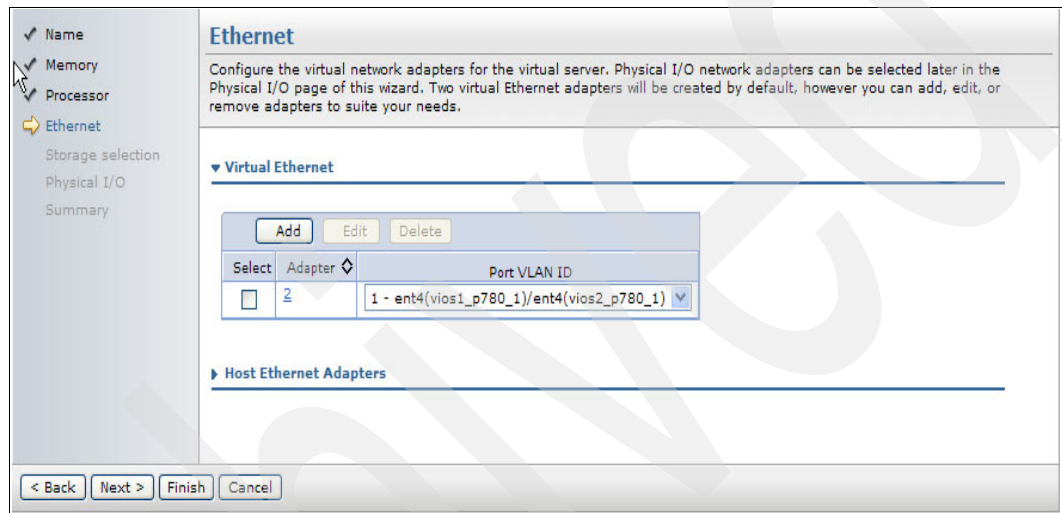


Figure 6-43 Ethernet configuration for the virtual server

6. In the Storage Selection window, select **No** (you want to manage the virtual storage adapters for this Virtual Server).

This option allows you to manually configure the slot assignment for the virtual FC adapters.

7. Follow these steps in the Virtual Storage Adapters window:
 - Enter 100 for the Maximum number of virtual adapters.
 - Configure the virtual FC adapters according to the values in Table 6-5 on page 227.
 Figure 6-44 shows the lpar1 virtual adapters configuration.

Select	Adapter ID	Type	Connecting Virtual Server	Connecting Adapter ID
<input type="checkbox"/>	20	Fibre Channel	vios1_p780_1 (1)	30
<input type="checkbox"/>	21	Fibre Channel	vios1_p780_1 (1)	31
<input type="checkbox"/>	22	Fibre Channel	vios2_p780_1 (2)	30
<input checked="" type="checkbox"/>	23	Fibre Channel	vios2_p780_1 (2)	31

Figure 6-44 Client virtual FC adapters creation using SDMC

8. In the Physical I/O adapters window, do not configure any adapter.
9. In the Summary window, review the settings and click **Finish**. You have created the virtual server at this point.

Slot assignments: You select slot assignments merely to have references and a proper initial assignment. After your LPAR becomes operational, these numbers change.

Because the SDMC has an Integrated Virtualization Manager (IVM) approach in the virtual server creation, you need to review the settings in the profile and configure them:

- ▶ **Paging virtual I/O server:** Ensure that you have a primary and secondary paging virtual I/O server.
- ▶ **Processor:** Adjust the values to your desired values.
- ▶ **Memory:** Consider maximum memory size and paging devices for AMS.

Now that the LPARs are created, proceed to create the server virtual FC adapters.

6.5.7 Server-side NPIV configuration

In this section, we show you how to create the server virtual FC adapters and to perform the rest of the configuration to enable your virtual server clients to access the back-end storage devices.

Creating Virtual FC adapters

After you have defined the client virtual servers, we need to create the server virtual FC adapters. You can perform this operation dynamically using dynamic LPAR (DLPAR), or you can shut down your virtual I/O server and modify the partition profile.

Follow these steps to perform the operation dynamically:

1. Select the virtual I/O server.

2. Click **Action** → **System Configuration** → **Manage Virtual Server**. The Manage Virtual Server pane appears.
3. Navigate to the storage adapters section.
4. Click **Add** and configure the virtual FC adapters according to Table 6-5 on page 227.
5. Configure the devices in the virtual I/O server by using the **cfgdev** command.

Save the configuration: After you perform the DLPAR operation, you must save the current configuration to avoid losing the configuration at the next partition shutdown.

Follow these steps to modify the partition profile:

1. Select the virtual I/O server.
2. Click **Action** → **System Configuration** → **Manage Profiles**.
3. Select the profile that you want to modify, and click **Action** → **Edit**.
4. Navigate to the **Virtual Adapters** tab.
5. Click **Action** → **Create Virtual Adapter** → **Fibre Channel Adapter**.
6. Configure the virtual FC adapters according to Table 6-5 on page 227.
7. Shut down the partition and start it again with the modified profile.

SAN zoning for client virtual servers

Because the NPIV technology presents newly independent worldwide port names (WWPNs) to the SAN devices, you need to perform specific zoning and storage configuration for each server.

To discover the WWPN information for each virtual FC client adapter, follow these steps:

1. Select the LPAR that you want to configure.
2. Click **Action** → **System Configuration** → **Manage Profiles**.
3. Select the profile with which you want to work.
4. Click **Action** → **Edit**.
5. Navigate to the **Virtual Adapters** tab.
6. Select the Client FC adapter with which you want to work. Click **Action** → **Properties**. The window that is shown in Figure 6-45 opens.

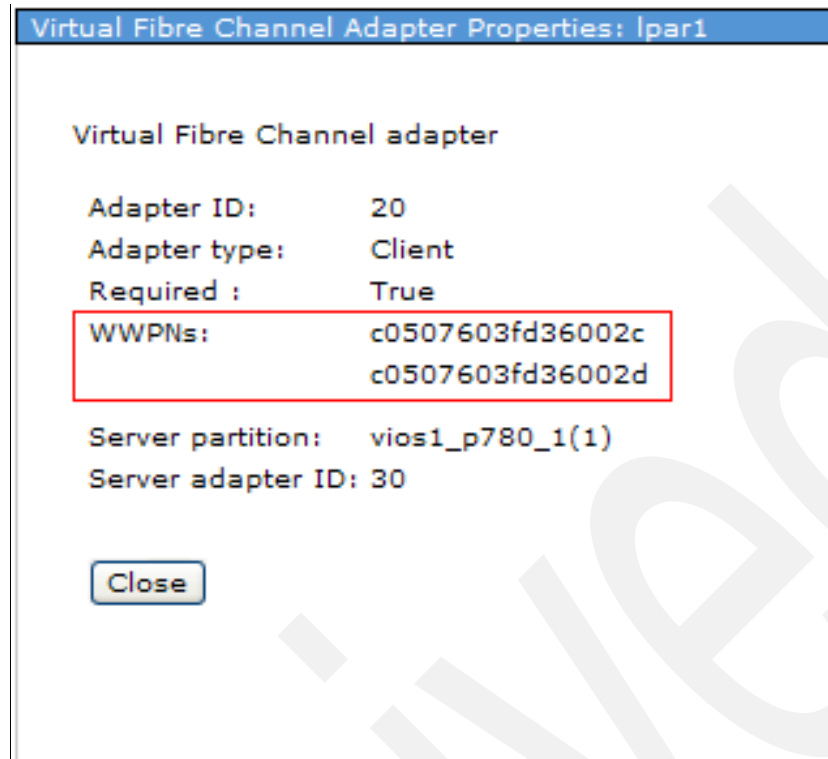


Figure 6-45 WWN configuration for a client virtual FC adapter

7. Document the WWPNs and repeat the process for each client virtual FC adapter.
8. Perform the zoning and storage configuration task to enable these WWPNs to access the storage logical unit numbers (LUNs).
9. Repeat this process for all the adapters in your virtual server client.

Important: You need to zone both WWPNs to be able to perform LPM operations. You must perform the zoning and storage configuration manually. The virtual WWPNs do not appear in the switch fabric as the physical adapters.

Mapping virtual FC adapters to physical HBAs

We create the client and VIO partitions, the client virtual FC adapters, the server virtual FC adapters, and the zoning tasks. We now need to map the server virtual FC adapters to the physical HBAs.

The first step is to check that the SAN ports to which we are connected are NPIV-capable ports:

1. In the virtual I/O server, execute the **lsnports** command, as shown in Figure 6-46.

```

$ lsnports
name          physloc          fabric tports aports swwpns awwpns
fcs2          U78C0.001.DBJF678-P2-C1-T1  1      64      64    2048    2046
fcs0          U78C0.001.DBJH615-P2-C1-T1  1      64      64    2048    2046
$

```

Figure 6-46 The lsnports command execution

Important: If the command output shows a value of 1 in the fabric attribute, the port switch is NPIV capable. If it shows 0, you need to change your SAN port configuration.

2. Use the **vfcmmap** and the **lsmmap** commands to perform the physical to virtual adapter mapping, as shown in Example 6-5.

Example 6-5 vfcmmap and lsmmap commands example

```

$ vfcmmap -vadapter vfchost0 -fcp fcs0
$ vfcmmap -vadapter vfchost1 -fcp fcs2
$ vfcmmap -vadapter vfchost2 -fcp fcs0
$ vfcmmap -vadapter vfchost3 -fcp fcs2
$ lsmmap -all -npiv

```

Name	Physloc	ClntID	ClntName	ClntOS
vfchost0	U9179.MHB.109AF6P-V1-C30	3		

```

Status:NOT_LOGGED_IN
FC name:fcs0                      FC loc code:U78C0.001.DBJH615-P2-C1-T1
Ports logged in:0
Flags:4<NOT_LOGGED>
VFC client name:                  VFC client DRC:

```

Name	Physloc	ClntID	ClntName	ClntOS
vfchost1	U9179.MHB.109AF6P-V1-C31	3		

```

Status:NOT_LOGGED_IN
FC name:fcs2                      FC loc code:U78C0.001.DBJF678-P2-C1-T1
Ports logged in:0
Flags:4<NOT_LOGGED>
VFC client name:                  VFC client DRC:

```

Name	Physloc	CIntID	CIntName	CIntOS
vfchost2	U9179.MHB.109AF6P-V1-C40	4		

Status:NOT_LOGGED_IN
 FC name:fcs0 FC loc code:U78C0.001.DBJH615-P2-C1-T1
 Ports logged in:0
 Flags:4<NOT_LOGGED>
 VFC client name: VFC client DRC:

Name	Physloc	CIntID	CIntName	CIntOS
vfchost3	U9179.MHB.109AF6P-V1-C41	4		

Status:NOT_LOGGED_IN
 FC name:fcs2 FC loc code:U78C0.001.DBJF678-P2-C1-T1
 Ports logged in:0
 Flags:4<NOT_LOGGED>
 VFC client name: VFC client DRC:

Status attribute: Check the Status attribute. In Example 6-5, the Status attribute shows as NOT_LOGGED_IN, because the client virtual server did not yet connect using this adapter. After you finish your client partition configuration, this attribute changes to LOGGED_IN.

Client installation

Continue with the client installation:

1. At this point, boot your client virtual server and install AIX (the same way that you install it for any LPAR).
2. Install the corresponding MPIO driver in your client LPAR.
3. Test the available path to the storage.
4. Perform your usual AIX configurations.

Active Memory Expansion (AME) configuration

The AME configuration is a specific configuration that depends on multiple factors: the environment and applications. In this section, we present a configuration example that is based in a newly installed shared memory partition in which we use a memory stress tool to generate workloads.

Note: To get the memory stress tool, download it from the following website:
<http://www.ibm.com/developerworks/wikis/display/WikiPtype/nstress>

In this scenario, we use the **nmem64**, **dbstart**, and **webstart** scripts.

We use an LPAR that is one of the partitions that we installed in this scenario. It has the following configuration:

- Six GB RAM

- ▶ Four GB paging space
- ▶ Two CPUs
- ▶ Four virtual CPUs
- ▶ Four SMT enabled
- ▶ AMS is enabled
- ▶ AME is disabled

As shown in Example 6-6, initially there is no workload in the LPAR.

Example 6-6 The Topas Monitor output in an idle partition

Topas Monitor for host:lp1							EVENTS/QUEUES		FILE/TTY	
Thu May 26 09:45:12 2011 Interval:2							Cswitch	200	Readch	1906
							Syscall	205	Writech	177
CPU	User%	Kern%	Wait%	Idle%	Physc	Entc%	Reads	20	Rawin	0
Total	0.0	0.3	0.0	99.7	0.01	0.56	Writes	0	Ttyout	177
							Forks	0	Igets	0
Network	BPS	I-Pkts	O-Pkts	B-In	B-Out	Execs	0	Namei	23	
Total	505.0	6.50	0.50	299.0	206.0	Runqueue	1.00	Dirblk	0	
							Waitqueue	0.0		
Disk	Busy%	BPS	TPS	B-Read	B-Writ	MEMORY				
Total	0.0	0	0	0	0	PAGING	Real,MB		6144	
							Faults	OK	% Comp	17
FileSystem	BPS	TPS	B-Read	B-Writ	Steals		OK	% Noncomp	0	
Total	1.86K	20.00	1.86K	0	PgspIn		0	% Client	0	
							PgspOut	OK		
Name	PID	CPU%	PgSp	Owner	PageIn		0	PAGING SPACE		
xmhc	786456	0.0	60.0K	root	PageOut		OK	Size,MB	4096	
topas	3473578	0.0	2.40M	root	Sios		OK	% Used	6	
getty	8585228	0.0	588K	root				% Free	94	
gil	1900602	0.0	124K	root	NFS (calls/sec)					
clstrmgr	5570762	0.0	1.31M	root	SerV2		0	WPAR Activ	0	
clcomd	4587562	0.0	1.71M	root	CliV2		0	WPAR Total	0	
rpc.lock	5439664	0.0	208K	root	SerV3		0	Press: "h"-help		
netm	1835064	0.0	60.0K	root	CliV3		0	"q"-quit		

In order to generate a workload, we execute the nstress tool with four memory stress processes, each one consuming 2,000 MB of RAM memory during a five-minute period. Refer to Example 6-7.

Example 6-7 nmem64 command execution

```
# nohup ./nmem64 -m 2000 -s 300 &
[1] 7602190
# Sending output to nohup.out
nohup ./nmem64 -m 2000 -s 300 &
[2] 7405586
# Sending output to nohup.out
nohup ./nmem64 -m 2000 -s 300 &
[3] 8388704
# Sending output to nohup.out
nohup ./nmem64 -m 2000 -s 300 &
[4] 8126552
# Sending output to nohup.out
```

You can see the Topas Monitor output in Example 6-8. Notice that there is significant paging activity in the server.

Example 6-8 Topas Monitor output in a memory stressed partition without memory compression

Topas Monitor for host:lp1r1							EVENTS/QUEUES		FILE/TTY	
Thu May 26 09:48:14 2011 Interval:2							Cswitch	2533	Readch	1906
							Syscall	348	Writech	196
CPU	User%	Kern%	Wait%	Idle%	PhySc	Entc%	Reads	20	Rawin	0
Total	1.0	32.7	5.8	60.5	1.08	53.87	Writes	0	Ttyout	196
							Forks	0	Igets	0
Network	BPS	I-Pkts	O-Pkts		B-In	B-Out	Execs	0	Namei	23
Total	407.0	4.00	0.50		184.0	223.0	Runqueue	2.50	Dirblk	0
							Waitqueue	4.5		
Disk	Busy%	BPS	TPS	B-Read	B-Writ	MEMORY				
Total	20.0	8.85M	1.16K	4.31M	4.54M	PAGING		Real,MB	6144	
							Faults	1128K	% Comp	99
FileSystem		BPS	TPS	B-Read	B-Writ	Steals	1161K	% Noncomp	0	
Total		1.86K	20.00	1.86K	0	PgspIn	1103	% Client	0	
							PgspOut	1161K		
Name	PID	CPU%	PgSp	Owner		PageIn	1103	PAGING SPACE		
amepat	5832710	25.0	220K	root		PageOut	1161K	Size,MB	4096	
lrud	262152	0.0	92.0K	root		Sios	1925K	% Used	80	
java	9044006	0.0	77.6M	root				% Free	20	
nmem64	8388704	0.0	1.95G	root		NFS (calls/sec)				
nmem64	8126552	0.0	1.95G	root		SerV2	0	WPAR Activ	0	
nmem64	7405586	0.0	1.95G	root		Cliv2	0	WPAR Total	0	
nmem64	7602190	0.0	1.95G	root		SerV3	0	Press: "h"-help		
topas	3473578	0.0	2.88M	root		Cliv3	0	"q"-quit		

While the memory stress processes run, we execute the **amepat** command to analyze the memory behavior and get the AME recommendation. Example 6-9 shows the command output.

Example 6-9 The amepat command output during high paging activity

# amepat 3	
Command Invoked	: amepat 3
Date/Time of invocation	: Thu May 26 09:47:37 EDT 2011
Total Monitored time	: 4 mins 48 secs
Total Samples Collected	: 3
System Configuration:	

Partition Name	: lp1r1_p780
Processor Implementation Mode	: POWER7
Number Of Logical CPUs	: 16
Processor Entitled Capacity	: 2.00
Processor Max. Capacity	: 4.00
True Memory	: 6.00 GB
SMT Threads	: 4
Shared Processor Mode	: Enabled-Uncapped
Active Memory Sharing	: Enabled
Active Memory Expansion	: Disabled

System Resource Statistics:	Average	Min
Max		
-----	-----	-----
CPU Util (Phys. Processors)	0.32 [8%]	0.16 [4%]
0.65 [16%]		
Virtual Memory Size (MB)	8477 [138%]	6697 [109%]
9368 [152%]		
True Memory In-Use (MB)	6136 [100%]	6136 [100%]
6136 [100%]		
Pinned Memory (MB)	1050 [17%]	1050 [17%]
1050 [17%]		
File Cache Size (MB)	36 [1%]	31 [1%]
48 [1%]		
Available Memory (MB)	0 [0%]	0 [0%]
0 [0%]		

Active Memory Expansion Modeled Statistics:

Modeled Expanded Memory Size : 6.00 GB
Average Compression Ratio : 2.00

Expansion Factor	Modeled True Memory Size	Modeled Memory Gain	CPU Usage Estimate
-----	-----	-----	-----
1.00	6.00 GB	0.00 KB [0%]	0.00 [0%]
1.12	5.38 GB	640.00 MB [12%]	0.00 [0%]
1.20	5.00 GB	1.00 GB [20%]	0.00 [0%]
1.30	4.62 GB	1.38 GB [30%]	0.00 [0%]
1.42	4.25 GB	1.75 GB [41%]	0.00 [0%]
1.50	4.00 GB	2.00 GB [50%]	0.00 [0%]
1.60	3.75 GB	2.25 GB [60%]	0.03 [1%]

Active Memory Expansion Recommendation:

The recommended AME configuration for this workload is to configure the LPAR with a memory size of 3.75 GB and to configure a memory expansion factor of 1.60. This will result in a memory gain of 60%. With this configuration, the estimated CPU usage due to AME is approximately 0.03 physical processors, and the estimated overall peak CPU resource required for the LPAR is 0.68 physical processors.

NOTE: amepat's recommendations are based on the workload's utilization level during the monitored period. If there is a change in the workload's utilization level or a change in workload itself, amepat should be run again.

The modeled Active Memory Expansion CPU usage reported by amepat is just an estimate. The actual CPU usage used for Active Memory Expansion may be lower or higher depending on the workload.

In Example 6-9 on page 235, you can observe the **amepat** command execution output during the period in which we stress the server memory. In the recommendation section, The **amepat** command specifies that we need to configure 3.75 GB and a 1.60 expansion factor. In this case, we have two options:

- Follow the recommendation as is: In this case, the amount of physical memory that is consumed by the virtual server is reduced, but the paging activity remains.
- Configure the 1.60 expansion factor and continue using the 6 GB of logical RAM memory (remember we are using AMS, also). In this case, the paging activity disappears.

We present both scenarios. We must start by enabling the AME feature. Follow these steps:

1. Shut down the virtual server.
2. Modify the partition profile. In the Memory tab, select AME and enter a 1.6 active memory expansion factor.
3. Reduce the assigned memory to the desired memory of 3.75 GB, as shown in Figure 6-47.

Figure 6-47 AME configuration options

4. Start the partition.
5. We now execute the memory stress test with 3.75 GB RAM and a 1.6 AME expansion factor in the virtual server. Example 6-10 shows the results.

Example 6-10 The Topas Monitor output in a memory-stressed partition

Topas Monitor for host:lp1							EVENTS/QUEUES		FILE/TTY	
Thu May 26 10:15:31 2011 Interval:2							Cswitch	2225	Readch	1906
							Syscall	206	Writech	151
CPU	User%	Kern%	Wait%	Idle%	PhySc	Entc%	Reads	20	Rawin	0
Total	2.5	2.5	13.1	82.0	0.17	8.37	Writes	0	Ttyout	151
							Forks	0	Igets	0
Network	BPS	I-Pkts	O-Pkts	B-In	B-Out	Execs	0	Namei	23	

Total	270.0	2.00	0.50	92.00	178.0	Runqueue	2.50	Dirblk	0
						Waitqueue	3.5		
Disk	Busy%	BPS	TPS	B-Read	B-Writ			MEMORY	
Total	19.9	7.97M	1.06K	3.96M	4.01M	PAGING		Real,MB	6144
						Faults	2687K	% Comp	99
FileSystem		BPS	TPS	B-Read	B-Writ	Steals	2709K	% Noncomp	0
Total		1.86K	20.00	1.86K	0	PgspIn	1014	% Client	0
						PgspOut	1027K		
Name	PID	CPU%	PgSp	Owner		PageIn	1014	PAGING SPACE	
cmemd	655380	0.0	180K	root		PageOut	1027K	Size,MB	4096
nmem64	7733342	0.0	1.95G	root		Sios	1914K	% Used	75
nmem64	7209204	0.0	1.95G	root				% Free	25
nmem64	6619252	0.0	1.95G	root		AME			
nmem64	8585308	0.0	1.95G	root		TMEM	3.75G	WPAR Activ	0
lrud	262152	0.0	92.0K	root		CMEM	2.25G	WPAR Total	0
topas	7602196	0.0	2.79M	root		EF[T/A]	1.6/1.6	Press: "h"-help	
java	5243082	0.0	70.2M	root		CI:1.66K	CO:1.64K	"q"-quit	

In Example 6-10, observe that you still have significant paging activity. However, the virtual server thinks it has 6 GB of RAM memory. In reality, it only has 3.75 GB, and there is 2.25 GB of compressed memory.

6. Dynamically, we add 2.25 GB of memory to the virtual server. The total amount is now 6 GB of RAM memory.
7. We execute the tests again. Example 6-11 presents the results.

Example 6-11 Topas output in a memory-stressed partition with memory compression and 6 GB RAM

Topas Monitor for host:lp1							EVENTS/QUEUES		FILE/TTY	
Thu May 26 10:22:25 2011 Interval:2							Cswitch	2145	Readch	1909
							Syscall	245	Writech	668
CPU	User%	Kern%	Wait%	Idle%	Physc	Entc%	Reads	20	Rawin	0
Total	38.4	48.6	0.8	12.2	3.74	187.07	Writes	6	Ttyout	215
							Forks	0	Igets	0
Network	BPS	I-Pkts	O-Pkts	B-In	B-Out		Execs	0	Namei	30
Total	850.0	10.52	1.00	540.0	310.1		Runqueue	9.52	Dirblk	0
							Waitqueue	0.0		
Disk	Busy%	BPS	TPS	B-Read	B-Writ				MEMORY	
Total	0.0	246K	50.59	246K	0		PAGING		Real,MB	9728
							Faults	80140K	% Comp	95
FileSystem		BPS	TPS	B-Read	B-Writ		Steals	80190K	% Noncomp	0
Total		1.88K	20.54	1.87K	17.03		PgspIn	0	% Client	0
							PgspOut	0K		
Name	PID	CPU%	PgSp	Owner			PageIn	49	PAGING SPACE	
cmemd	655380	40.3	180K	root			PageOut	0K	Size,MB	4096
lrud	262152	13.4	92.0K	root			Sios	49K	% Used	1
nmem64	8585310	13.4	1.95G	root				% Free	99	
nmem64	7209206	13.4	1.95G	root			AME			
nmem64	7733344	13.4	1.95G	root			TMEM	6.00G	WPAR Activ	0
nmem64	6619254	0.0	1.95G	root			CMEM	3.09G	WPAR Total	0
slp_svr	4718600	0.0	484K	root			EF[T/A]	1.6/1.6	Press: "h"-help	
topas	7602196	0.0	2.79M	root			CI:80.0K	CO:77.8K	"q"-quit	

As you can see in Example 6-11 on page 238, there is no paging activity in the server with the new configuration.

Live Partition Mobility (LPM) operations

We now move lpar1 from the original server to the secondary server. Before performing the LPM operation, we generate memory and CPU activity with the nstress tool.

Environment: In this scenario, we generate workloads in the test partitions. These workloads exist in a real production environment.

Follow these steps to move the partition:

1. Execute the **dbstart.sh** script, which creates a fake database.
2. Execute the **webstart.sh** script, which creates a fake web server.
3. Execute the memory stress test: **nohup ./nmem64 -m 2000 -s 3000**. In this example, we execute it four times.
4. Observe the Topas Monitor output, as shown in Example 6-12.

Example 6-12 Topas Monitor output for migration candidate partition

Topas Monitor for host:lp1r1							EVENTS/QUEUES		FILE/TTY		
Fri May 27 13:06:50 2011 Interval:2							Cswitch	2489	Readch	0	
							Syscall	224	Writech	231	
CPU	User%	Kern%	Wait%	Idle%	Physc	Entc%	Reads	0	Rawin	0	
Total	46.4	35.9	0.7	17.0	3.67	183.75	Writes	0	Ttyout	231	
							Forks	0	Igets	0	
Network	BPS	I-Pkts	O-Pkts		B-In	B-Out	Execs	0	Namei	4	
Total	1.08K	18.48	0.50		850.1	258.7	Runqueue	6.49	Dirblk	0	
							Waitqueue	0.0			
Disk	Busy%	BPS	TPS	B-Read	B-Writ	MEMORY					
Total	0.1	0	0	0	0	PAGING		Real,MB	9728		
						Faults	64742	% Comp	87		
FileSystem	BPS	TPS	B-Read	B-Writ	Steals	64716	% Noncomp	0			
Total	1.86K	19.98	1.86K	0	PgspIn	0	% Client	0			
						PgspOut	0				
Name	PID	CPU%	PgSp	Owner	PageIn	0	PAGING SPACE				
cmemd	655380	40.8	180K	root	PageOut	0	Size,MB	4096			
nmem64	8388742	13.6	1.95G	root	Sios	0	% Used	1			
webserve	14418146	13.6	4.18M	root	% Free						99
lrud	262152	13.6	92.0K	root	AME						
db	12451978	13.6	64.2M	root	TMEM	6.00G	WPAP Activ	0			
nmem64	12386446	0.0	1.64G	root	CMEM	2.38G	WPAP Total	0			
nmem64	14221532	0.0	1.95G	root	EF[T/A]	1.6/1.6	Press: "h"-help				
nmem64	7340044	0.0	1.35G	root	CI:40.9K	CO:63.2K	"q"-quit				
webserve	10420390	0.0	108K	root							

5. Check the machine serial number, as shown in Example 6-13.

Example 6-13 Checking the machine serial number

```
# hostname
lp1r1_p780
# uname -u
IBM,02109AF6P
```

6. In the SDMC, select the virtual server to move (in this case, lp1r1).
7. Click **Action** → **Operations** → **Mobility** → **Migrate**.

8. Complete the migration wizard with the default options.
9. On the Summary window, you see a window similar to Figure 6-48.

Figure 6-48 LPM Summary window

10. Click **Finish** to perform the LPM operation. The Virtual server migration status appears, as shown in Figure 6-49 on page 240.

Figure 6-49 LPM operation in progress

11. After the migration completes, a window similar to Figure 6-50 opens.

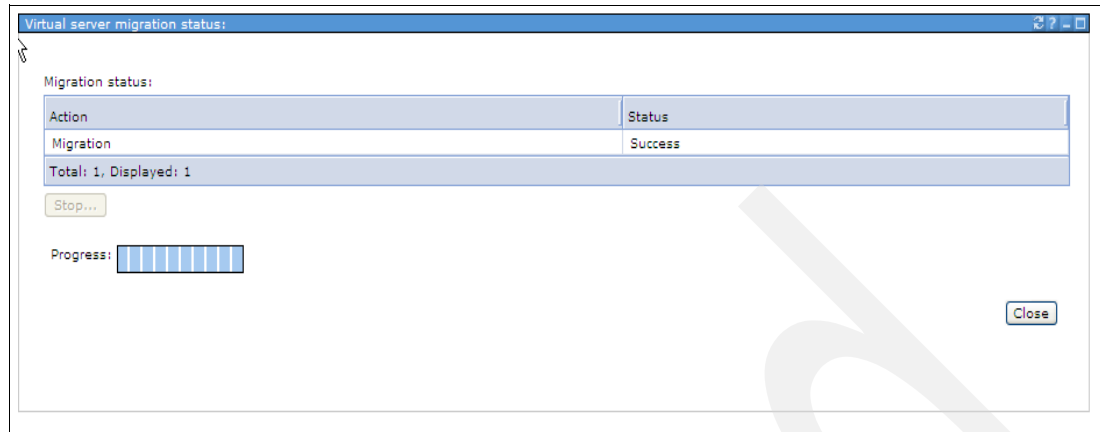


Figure 6-50 Successful LPM operation

12. Check the machine serial number, as shown in Example 6-14.

Example 6-14 New serial number for lpar1 LPAR

```
# hostname
lpar1_p780
# uname -u
IBM,02109AF7P
```

At this point, the partition has successfully migrated to the second Power server without disruption in the services.

6.6 LPM and PowerHA

In this section, we perform an LPM with a simple script running, followed by a PowerHA failover test. “Simple cluster installation” on page 360 shows the cluster configuration.

These items are in the setup:

- ▶ Nodes: rflpar10 and rflpar20
- ▶ Resource groups: lpar1svcr and lpar2svcr
- ▶ Application controllers: lpar2appserver and lpar2appserver
- ▶ Service IP labels: rflpar10_svc and rflpar20_svc

Example 6-15 shows the application server scripts for the PowerHA failover. There are simple DB2 start and stop scripts, and they must not be used in production environments.

Example 6-15 A simple application server is used in this example to show the failover test

```
cat /hascripts/startlpar1.sh
echo "0 `hostname` 0" > /home/db2inst1/sqllib/db2nodes.cfg
su - db2inst1 -c db2start
#
/hascripts/stoplpar1.sh
echo "0 `hostname` 0" > /home/db2inst1/sqllib/db2nodes.cfg
su - db2inst1 -c db2stop force
```

We also used a small script to confirm the following components:

- ▶ The physical server
- ▶ The LPAR that we are currently using
- ▶ The IP addresses
- ▶ The state of the database, which is shown after connecting to the database by using **db2 connect to test1 user db2inst1 using password**

Example 6-16 shows the script.

Example 6-16 Script to track the state of the LPAR

```
while true
do
echo ----- | tee -a /home/db2inst1/status.log
lsconf | head -2 | tee -a /home/db2inst1/status.log
hostname | tee -a /home/db2inst1/status.log
ifconfig en0 | tee -a /home/db2inst1/status.log
echo ----- | tee -a /home/db2inst1/status.log
db2 select tabname from syscat.tables fetch first 2 rows only |grep -v "\-\" | \
tee -a /home/db2inst1/status.log
date | tee -a /home/db2inst1/status.log
who -r
sleep 10
echo "===== \n"
done
```

Example 6-17 on page 242 shows the results of the script that is shown in Example 6-16. Look at the following components in Example 6-17 on page 242:

- ▶ System model
- ▶ Serial number
- ▶ lparname
- ▶ IP addresses
- ▶ Date and time
- ▶ In all instances, the db2 select statement result is the same

Example 6-17 Initial results before the LPM test

```
System Model: IBM,9117-MMA
Machine Serial Number: 101F170
rflpar10
en0:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT
,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
    inet 172.16.21.36 netmask 0xfffffc00 broadcast 172.16.23.255
    inet 172.16.21.40 netmask 0xfffffc00 broadcast 172.16.23.255
    inet 172.16.21.60 netmask 0xfffffc00 broadcast 172.16.23.255
    tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
-----

TABNAME
ATTRIBUTES
AUDITPOLICIES

    2 record(s) selected.

Fri Jun  3 12:45:50 EDT 2011
```

Example 6-18 shows the cluster status.

Example 6-18 Cluster status before LPM

```
/usr/es/sbin/cluster/clstat
clstat - HACMP Cluster Status Monitor
-----

Cluster: pough (1111817142)
Fri Jun 3 12:48:29 2011
      State: UP                      Nodes: 2
      SubState: STABLE

Node: rflpar10      State: UP
  Interface: rflpar10 (0)      Address: 172.16.21.36
                                State: UP
  Interface: rflpar10_svc (0)  Address: 172.16.21.60
                                State: UP
  Resource Group: lpar1svcrs  State: On line

Node: rflpar20      State: UP
  Interface: rflpar20 (0)      Address: 172.16.21.35
                                State: UP
  Interface: rflpar20_svc (0)  Address: 172.16.21.61
                                State: UP

***** f/forward, b/back, r/refresh, q/quit *****
```

6.6.1 The LPM operation

The LPM operation is performed from an IBM POWER6 570 to an IBM Power 780, as explained in 6.3, “Live Partition Mobility (LPM) using the HMC and SDMC” on page 210.

The migration is performed based on the information that is shown in Figure 6-51.

Partition Migration Validation - p570_170 - rflpar10

Fill in the following information to set up a migration of the partition to a different managed system. Click Validate to ensure that all requirements are met for this migration. You cannot migrate until migration set up has been verified.

Source system : p570_170
 Migrating partition: rflpar10
 Remote HMC: 172.16.20.22
 Remote User: sysadmin
 Destination system: p780_01 Refresh Destination System
 Destination profile name: default
 Destination shared processor pool: DefaultPool (0)
 Source mover service partition: 570_1_VIO_1 MSP Pairing...
 Destination mover service partition: vios1_p780_1
 Wait time (in min): 5
 Override virtual network errors when possible: ☐
 Override virtual storage errors when possible: ☐

Virtual Storage assignments :

Select	Source Slot ID	Slot Type	Destination VIOS
<input type="checkbox"/>	107	SCSI	vios1_p780_1
<input checked="" type="checkbox"/>	107	SCSI	vios2_p780_1

View VLAN Settings... Validate Migrate Cancel

Figure 6-51 Migration setup

After the LPM operation, we observed the following results, as shown in Example 6-19:

- ▶ The cluster status that is shown with the `clstat` command does not change.
- ▶ We did not lose the session on which we ran the `while` statement.
- ▶ The model and serial number are the only items that changed.
- ▶ The IP addresses do not change.
- ▶ The script continued running.

Example 6-19 Results from the LPM operation

```
System Model: IBM,9117-MMA
Machine Serial Number: 109AF6P
rflpar10
en0:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT
,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
    inet 172.16.21.36 netmask 0xfffffc00 broadcast 172.16.23.255
    inet 172.16.21.40 netmask 0xfffffc00 broadcast 172.16.23.255
    inet 172.16.21.60 netmask 0xfffffc00 broadcast 172.16.23.255
    tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
-----

TABNAME
ATTRIBUTES
AUDITPOLICIES
```

2 record(s) selected.


```

Fri Jun  3 12:55:01 EDT 2011
.          run-level 2 Jun 03 12:54          2    0    S
=====

System Model: IBM,9179-MHB
Machine Serial Number: 109AF6P
rflpar10
en0:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT
,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
    inet 172.16.21.36 netmask 0xfffffc00 broadcast 172.16.23.255
    inet 172.16.21.40 netmask 0xfffffc00 broadcast 172.16.23.255
    inet 172.16.21.60 netmask 0xfffffc00 broadcast 172.16.23.255
    tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
-----

TABNAME
ATTRIBUTES
AUDITPOLICIES

2 record(s) selected.

Fri Jun  3 12:55:28 EDT 2011
.          run-level 2 Jun 03 12:54          2    0    S

```

6.6.2 The PowerHA operation

We continue. We run the same scripts as we ran in 6.6, “LPM and PowerHA” on page 241 and failed over using PowerHA. To force a failover, we forced one LPAR down.

We observed the following conditions, as shown in Example 6-20:

- ▶ We had to restart the session and rerun the status script.
- ▶ All IP addresses that were active on the failing node moved to rflpar20.

Example 6-20 Results of a failover

```

System Model: IBM,9179-MHB
Machine Serial Number: 109AF6P
rflpar20
en0:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT
,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
    inet 172.16.21.35 netmask 0xfffffc00 broadcast 172.16.23.255
    inet 172.16.21.41 netmask 0xfffffc00 broadcast 172.16.23.255
    inet 172.16.21.61 netmask 0xfffffc00 broadcast 172.16.23.255
    inet 172.16.21.60 netmask 0xfffffc00 broadcast 172.16.23.255
    tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
-----

TABNAME
ATTRIBUTES
AUDITPOLICIES

```

2 record(s) selected.

```
Fri Jun  3 13:10:29 EDT 2011
.          run-level 2 Jun 03 11:25      2    0    S
```

The cluster status changed. One node showed a *down* status, as shown in Example 6-21.

Example 6-21 Cluster status after the failover test

clstat - PowerHA SystemMirror Cluster Status Monitor

```
-----
Cluster: pough (1111817142)
Fri Jun  3 13:05:28 EDT 2011
      State: UP                      Nodes: 2
      SubState: STABLE

Node: rflpar10      State: DOWN
  Interface: rflpar10 (0)      Address: 172.16.21.36
                                State:   DOWN

Node: rflpar20      State: UP
  Interface: rflpar20 (0)      Address: 172.16.21.35
                                State:   UP
  Interface: rflpar10_svc (0)  Address: 172.16.21.60
                                State:   UP
  Interface: rflpar20_svc (0)  Address: 172.16.21.61
                                State:   UP
  Resource Group: lpar1svcrs   State: On line State: UP
  Resource Group: lpar2svcrs   State: On line
```

```
***** f/forward, b/back, r/refresh, q/quit *****
```

Example 6-20 on page 245 and Example 6-21 show the difference between PowerHA and LPM.

POWER7 Enterprise Server performance considerations

This chapter discusses the performance aspects of the POWER7 Enterprise Servers. We start by introducing the performance design of our POWER7 servers. We also introduce key considerations with POWER7 Enterprise Servers, such as the reliability, availability, and serviceability (RAS) features and virtualization features. We also discuss specific AIX and IBM i operating system considerations. In addition, we explain enhanced monitoring methods for POWER7 servers. In the last few sections, we discuss IBM performance management tools.

In this chapter, we discuss the following topics:

- ▶ Performance design for POWER7 Enterprise Servers
- ▶ POWER7 Servers performance considerations
- ▶ Performance considerations with hardware RAS features
- ▶ Performance considerations with Power virtualization features
- ▶ Performance considerations with AIX
- ▶ IBM i performance considerations
- ▶ Enhanced performance tools of AIX for POWER7
- ▶ Performance Management for Power Systems

7.1 Introduction

The IBM Power development team in Austin, TX, has aggressively pursued the integration of industry-leading mainframe reliability technologies into Power Systems servers. With the introduction of POWER7, there are successive generations of new RAS features included in the server line. One core principle that guides the IBM RAS architecture engineering design team is that systems must be configurable to achieve the required levels of availability without compromising performance, utilization, or virtualization. Hardware and firmware RAS features are independent of the operating system and, therefore, do not affect operating system or application performance. However, the hardware RAS features can provide the key enablement of availability features built into the AIX, IBM i, and Linux operating systems and benefits that contribute to the overall system availability.

7.2 Performance design for POWER7 Enterprise Servers

This section contains descriptions of the Power 780 and Power 795 performance features:

- ▶ Balanced architecture
- ▶ Processor embedded dynamic random access memory (eDRAM) technology
- ▶ Processor compatibility mode
- ▶ MaxCore and TurboCore modes
- ▶ Active Memory Expansion (AME)
- ▶ Power management's effect on system performance

7.2.1 Balanced architecture of POWER7

Multi-core processor technologies face major challenges to continue delivering growing throughput and performance. These challenges include the constraints of physics, power consumption, and socket pin count limitations.

To overcome these limitations, a balanced architecture is required. Many processor design elements need to be balanced on a server in order to deliver maximum throughput.

In many cases, IBM has been innovative in order to achieve the required levels of throughput and bandwidth. Areas of innovation for the POWER7 processor and POWER7 processor-based systems include (but are not limited to) these areas:

- ▶ On-chip L3 cache implemented in eDRAM
- ▶ Cache hierarchy and component innovation
- ▶ Advances in the memory subsystem
- ▶ Advances in off-chip signaling
- ▶ Exploitation of the long-term investment in coherence innovation

For example, POWER6, POWER5, and POWER4 systems derive large benefits from high bandwidth access to large, off-chip cache. However, socket pin count constraints prevent scaling the off-chip cache interface to support eight cores, which is a feature of the POWER7 processor.

Figure 7-1 on page 249 illustrates the POWER5 and POWER6 large L2 cache technology.

Cache Hierarchy Requirement for POWER Servers

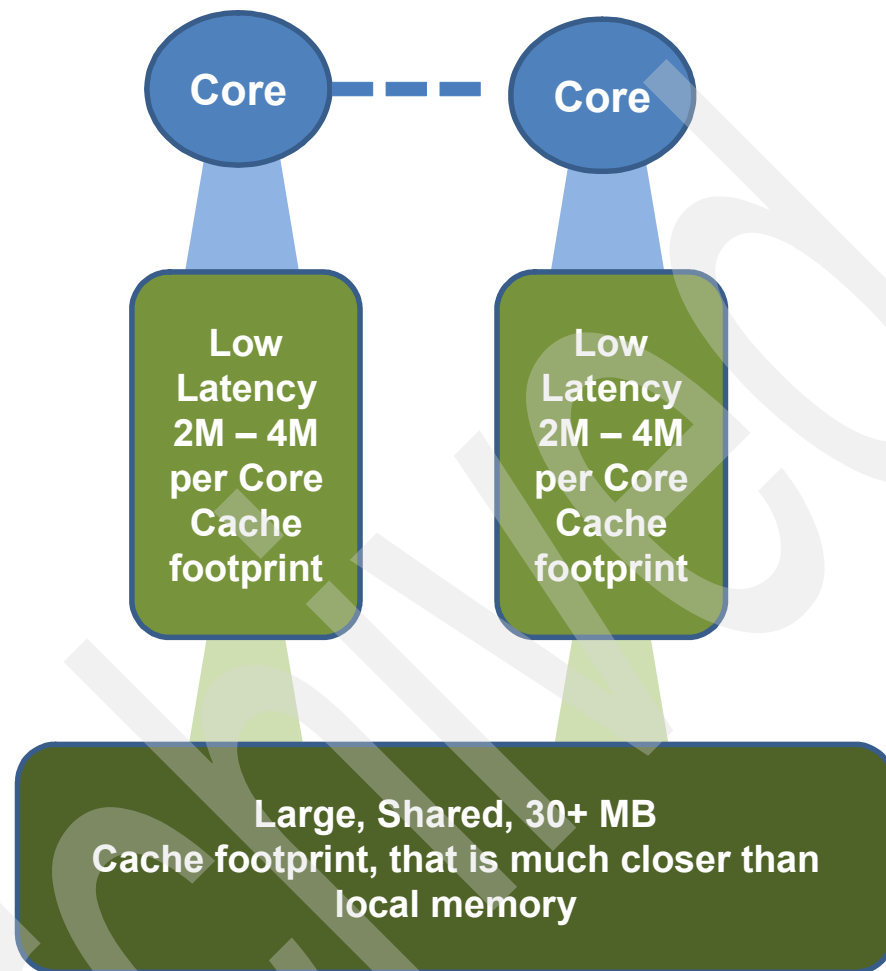


Figure 7-1 The POWER5 and POWER6 large L2 cache technology

IBM was able to overcome this challenge by introducing an innovative solution: high speed eDRAM on the processor chip. With POWER7, IBM introduces on-processor, high-speed, custom eDRAM, combining the dense, low power attributes of eDRAM with the speed and bandwidth of static random access memory (SRAM).

Figure 7-2 on page 250 illustrates the various memory technologies.

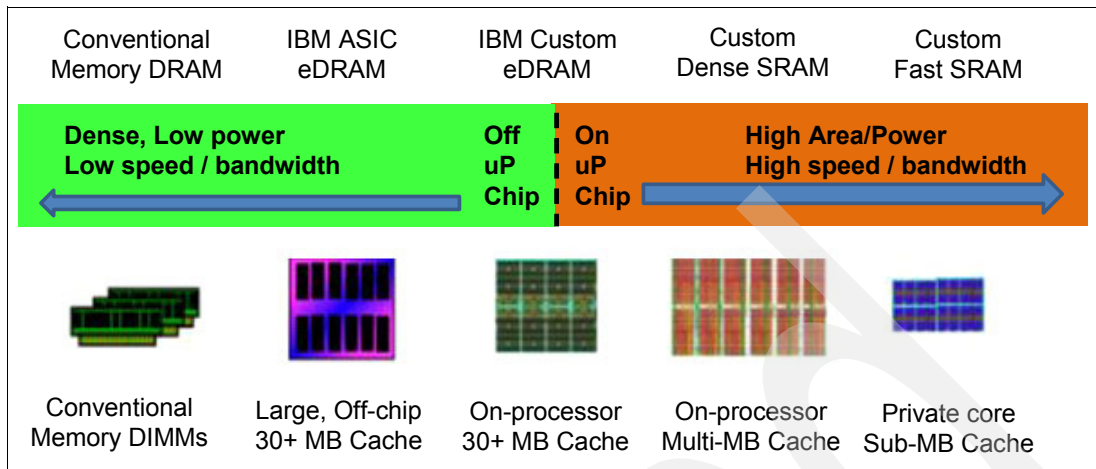


Figure 7-2 The various memory technologies

Another challenge is the need to satisfy both caching requirements: the low latency per core cache and the large cache with one cache.

IBM introduced an innovative solution called the hybrid L3 “Fluid” cache structure, which has these characteristics:

- Keeps multiple footprints at ~3X lower latency than local memory
- Automatically migrates private footprints (up to 4M) to the fast local region (per core) at ~5X lower latency than full L3 cache
- Automatically clones shared data to multiple private regions

Figure 7-3 illustrates the Hybrid L3 Fluid Cache Structure.

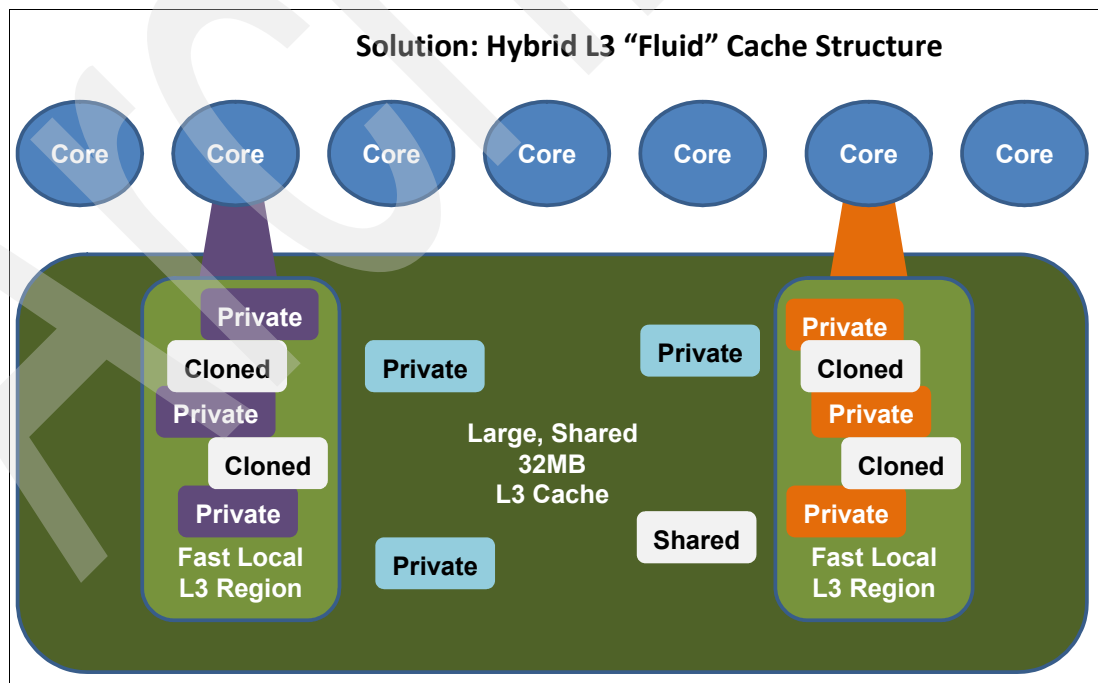


Figure 7-3 Hybrid L3 Fluid Cache

7.2.2 Processor eDRAM technology

In many cases, IBM has been innovative in order to achieve the required levels of throughput and bandwidth. Areas of innovation for the POWER7 processor and POWER7 processor-based systems include (but are not limited to) these areas:

- ▶ On-chip L3 cache implemented in embedded dynamic random access memory (eDRAM)
- ▶ Cache hierarchy and component innovation
- ▶ Advances in the memory subsystem
- ▶ Advances in off-chip signalling
- ▶ Exploitation of long-term investment in coherence innovation

The innovation of using eDRAM on the POWER7 processor chip is significant for several reasons:

- ▶ Latency improvement: A six-to-one latency improvement occurs by moving the L3 cache on-chip compared to L3 accesses on an external (on-ceramic) ASIC.
- ▶ Bandwidth improvement: A 2x bandwidth improvement occurs with on-chip interconnect. Frequency and bus sizes are increased to and from each core.
- ▶ No off-chip driver or receivers.
- ▶ Removing drivers or receivers from the L3 access path lowers interface requirements, conserves energy, and lowers latency.
- ▶ Small physical footprint: The eDRAM L3 cache requires far less physical space than an equivalent L3 cache that is implemented with conventional SRAM. IBM on-chip eDRAM uses only a third of the components that are used in conventional SRAM, which has a minimum of six transistors to implement a 1-bit memory cell.
- ▶ Low energy consumption: The on-chip eDRAM uses only 20% of the standby power of SRAM.

7.2.3 Processor compatibility mode

POWER7 supports partition mobility with POWER6 and POWER6+ systems by providing compatibility modes. Partitions running in POWER6 or POWER6+ compatibility mode can run in single thread (ST) or simultaneous multi-thread (SMT2). SMT4 and single-instruction, multiple-data (SIMD) double-precision floating-point (VSX) are not available in compatibility mode.

Applications that are single process and single threaded might benefit from running in ST mode. Multithreaded and multi-process applications typically benefit more running in SMT2 or SMT4 mode. ST mode can be beneficial in the case of a multi-process application where the number of application processes is smaller than the number of cores assigned to the partition.

Applications that do not scale with a larger number of CPUs might also benefit from running in ST or SMT2 mode instead of SMT4, because the lower number of SMT threads means a lower number of logical CPUs.

You can set ST, SMT2, or SMT4 mode through the `smtctl` command. The default mode is SMT4. For more information about performance considerations about processor compatibility mode, refer to 7.3.1, “Processor compatibility mode” on page 254.

7.2.4 MaxCore and TurboCore modes

TurboCore provides a higher frequency core and more cache per core, which are normally extremely good things for performance. Also, to make use of these positive attributes, the system's active core placement needs to change from having eight cores per chip to having four cores per chip. So, using TurboCore can also mean a change in the number of chips and perhaps in the number of drawers. For a given partition, this in turn can mean an increase in the probability of longer access latencies to memory and cache. But partition placement and workload type can influence these probabilities. As a result, the positive benefits of TurboCore's higher frequency and increased cache also have the potential of being offset to various extents by these longer latency storage accesses.

In cases in which the cross-chip accesses are relatively limited, all of TurboCore's benefits can remain. For information about MaxCore performance considerations, refer to 7.3.2, "TurboCore and MaxCore modes" on page 259.

7.2.5 Active Memory Expansion

Active Memory Expansion (AME) is an innovative POWER7 technology that uses compression and decompression to effectively expand the true physical memory that is available for client workloads. Often a small amount of processor resource provides a significant increase in the effective memory maximum.

Actual expansion results depend on how much you can compress the data that is used in the application. For example, an SAP ERP sample workload showed up to 100% expansion. An estimator tool and free trial are available.

AME differs from Active Memory Sharing (AMS). Active Memory Sharing moves memory from one partition to another partition. AMS is the best fit when one partition is not busy when another partition is busy, and it is supported on all AIX, IBM i, and Linux partitions.

A number of commands are available to monitor the AME configuration of a partition. These commands include **amepat**, **topas**, **vmstat**, and **lpartstat**.

The **amepat** command provides a summary of the AME configuration and can be used for monitoring and fine-tuning the configuration. The **amepat** command shows the current configuration, as well as the statistics of the system resource utilization over the monitoring period. For more information about the AME performance monitor, refer to 7.8.7, "Monitoring Active Memory Expansion (AME) statistics" on page 317.

7.2.6 Power management's effect on system performance

All power management modes can affect certain aspects of performance, depending on the system configuration and how performance is measured. Consider these issues before turning on any power management mode or feature:

- ▶ Systems running at low utilization (and consequently, low frequency) might maintain processor throughput. However, response time to a particular task might be affected. Also, the reaction time to an incoming workload can be affected.
- ▶ Any system setup that limits the amount of processing allowed, such as running with capped partitions, can cause the frequency to be reduced. Even though a partition might be running at 100% of its entitled capacity, the system as a whole might not be heavily utilized.

- ▶ Using virtual shared processor pools also can limit the overall system utilization and cause lower processor frequencies to be set.
- ▶ Certain external workload managers also have the effect of limiting system processing by adjusting workloads to a point where frequency is lowered.
- ▶ Because the processor frequency is variable, performance monitoring tools can be affected.

As shown in Figure 7-4 and Figure 7-5 on page 254 from a representative POWER7 system, enabling the various power savings modes can directly affect power consumption as a workload varies. For example, Dynamic Power Saver mode can deliver higher workload throughput than either of the other modes at the expense of system power consumption. At less than peak utilization, Dynamic Power Saver mode delivers power savings, and it might still deliver adequate workload throughput. It is important to note that trade-offs must be made between energy consumption, workload response times, and throughput. For additional details about these issues, refer to a companion white paper, *EnergyScale Performance Characteristics for IBM Power Systems*.

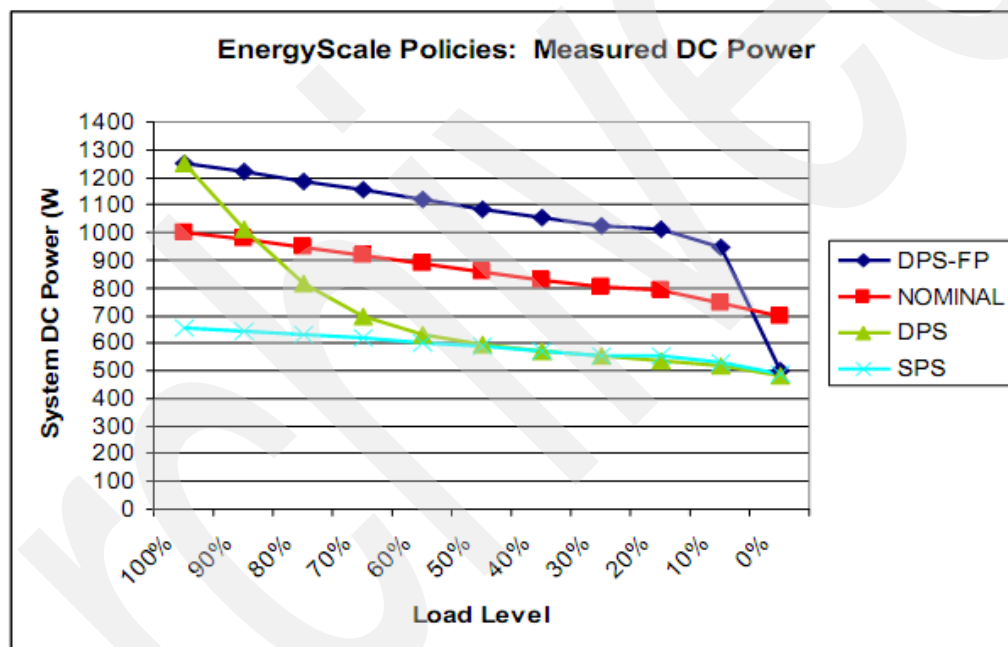


Figure 7-4 System energy consumption trends: System load level and average processor frequency

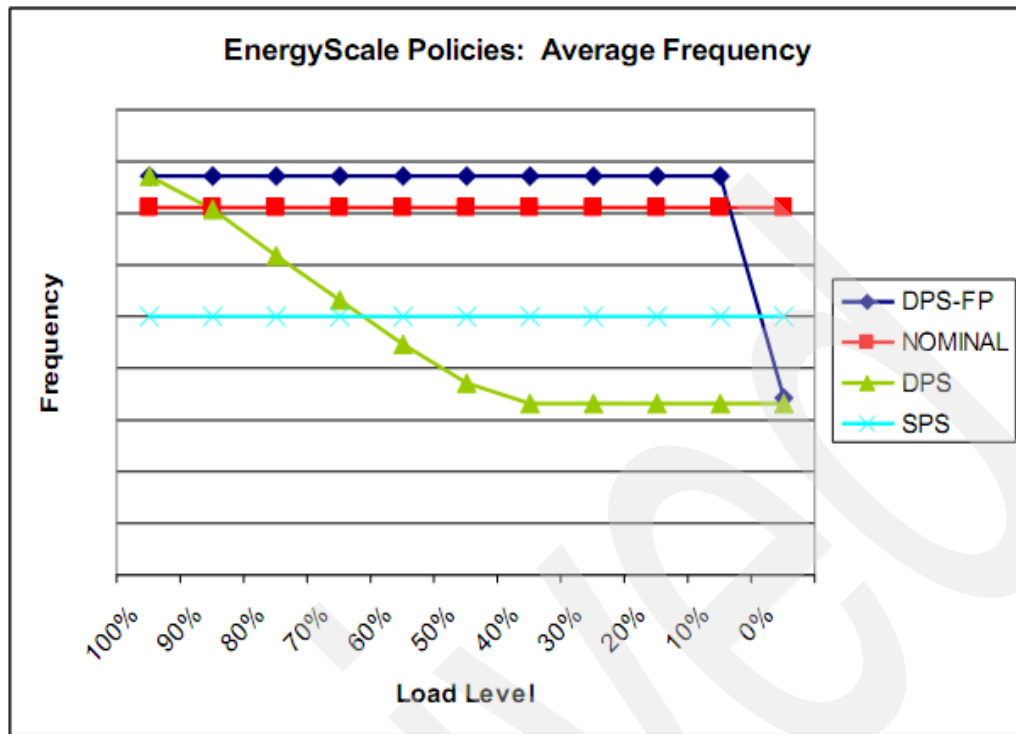


Figure 7-5 Nominal and static power save (SPS) modes

If you want to know about power management features and differences in Dynamic Power Saver from POWER6 to POWER7, refer to 2.5, “Power management” on page 36.

7.3 POWER7 Servers performance considerations

In this section, we introduce performance considerations with the POWER7 server hardware features. This section consists of the following topics:

- ▶ Processor compatibility mode
- ▶ TurboCore and MaxCore modes
- ▶ Active Memory Expansion
- ▶ Logical Memory Block size
- ▶ System huge-page memory

7.3.1 Processor compatibility mode

Processor compatibility modes enable you to move logical partitions (LPARs) between servers that have separate processor types without upgrading the operating environments that are installed in the LPARs. In certain cases, the options with this feature result in varying performance.

Regarding POWER7 servers, you have four options for choosing processor compatibility mode:

- ▶ POWER6

This execution mode is compatible with Version 2.05 of the Power Instruction Set Architecture (ISA)¹

► POWER6+

This mode is similar to POWER6 with eight additional storage protection keys.

► POWER7

The POWER7 mode is the native mode for POWER7 processors, implementing the V2.06 of the Power Instruction Set Architecture²

► Default

The Power hypervisor determines the current mode for the LPAR.

Each LPAR running on a POWER7-based system can run in one of these modes. Also, LPARs on the same system can run in separate modes. This mode setting is controlled by the partition definition when the LPAR is created.

For more information about processor compatibility mode, refer to the following link:

<http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7hc3/iphc3pcmdefs.htm>

Comparisons among the processor compatibility modes

Table 7-1 shows each processor compatibility mode and the servers on which the LPARs that use each processor compatibility mode can successfully operate.

Table 7-1 The definition and supported servers with separate processor compatibility mode

Processor compatibility mode	Description	Supported servers
POWER6	The POWER6 processor compatibility mode allows you to run operating-system versions that use all the standard features of the POWER6 processor.	POWER6 POWER6+ POWER7 processor-based servers
POWER6+	The POWER6+ processor compatibility mode allows you to run operating-system versions that use all the standard features of the POWER6+ processor.	POWER6+ POWER7 processor-based servers
POWER7	The POWER7 processor compatibility mode allows you to run operating-system versions that use all the standard features of the POWER7 processor.	POWER7 processor-based servers

¹ If you want to know detailed information about Power Instruction Set V2.05, refer to the following website:
http://www.power.org/resources/reading/PowerISA_V2.05.pdf

² If you want to know detailed information about Power Instruction Set V2.06, refer to the following website:
http://www.power.org/resources/downloads/PowerISA_V2.06_PUBLIC.pdf

Processor compatibility mode	Description	Supported servers
Default	The preferred processor compatibility mode enables the Power hypervisor to determine the current mode for the LPAR. When the preferred mode is set to default, the Power hypervisor sets the current mode to the most fully featured mode supported by the operating environment. In most cases, this mode is the processor type of the server on which the LPAR is activated.	Dependent on the current processor compatibility mode of the LPAR. For example, if the Power hypervisor determines that the current mode is POWER7, the LPAR can run on POWER7 processor-based servers.

In addition to the description and supported server differences between POWER6/6+ and POWER7 modes, the operating system requirement for the various modes also differs. For detailed information about the minimal operating system requirements, refer to the IBM Offering information website:

<http://www-01.ibm.com/common/ssi/index.wss>

In addition, many functional differences exist between POWER6/6+ mode and POWER7 mode. Table 7-2 lists the differences with regard to performance and RAS.

Table 7-2 Functional differences between POWER6/POWER6+ and POWER7 modes

POWER6 and POWER6+ mode	POWER7 mode	Comment
Dual-threaded (SMT2) Single-threaded (ST)	Quad-threaded (SMT4) Dual-threaded (SMT2) Single-threaded (ST)	Refer to “Simultaneous Multithreading Mode”
Vector Multimedia Extension (VMX) or AltiVec	VMX or AltiVec VSX (Vector Scalar Extension)	Refer to “Single Instruction Multiple Data” on page 257
64-core/128-thread Scaling	64-core/256-thread Scaling 256-core/1024-thread Scaling (Only AIX7 support)	Refer to “Large scale-up capability” on page 257
8/16 Storage Protection Keys ^a	32 Storage Protection Keys	Refer to “Storage protection keys” on page 258

a. POWER6+ mode provides 16 storage protection keys.

Simultaneous Multithreading Mode

Simultaneous Multithreading Mode (SMT) technology can help to increase the processor’s utilization by improving cache misses and instruction dependency delay issues³. Enabling SMT (2 or 4) mode provides concurrent execution of the instruction stream by multiple threads on the same core. Table 7-3 on page 257 lists the various SMT modes that are supported by the Power servers.

³ For the description of cache misses and instruction dependency delay issues, refer to the IBM white paper:
http://www-03.ibm.com/systems/resources/pwrsysperf_SMT40nP7.pdf

Table 7-3 SMT modes supported with different Power servers

Power server	Supported SMT mode	Number of logical CPUs per core
POWER5	ST	1
POWER5	SMT2	2
POWER6/6+	ST	1
POWER6/6+	SMT2	2
POWER7	ST	1
POWER7	SMT2	2
POWER7	SMT4	4

From Table 7-3, you can see that POWER7 servers now support SMT4 mode⁴. In general, because there are four threads running on one core concurrently, it can provide higher total performance and throughput than other SMT modes.

Multithreaded and multiple process applications typically benefit more by running in SMT2 or SMT4 mode. In AIX, ST, SMT2, or SMT4 mode can be set through the `smtctl` command dynamically. For detailed information about SMT tuning on AIX, refer to 7.6.6, “Simultaneous multithreading (SMT)” on page 294.

Linux for Power also has a similar command (`ppc64_cpu`) to control SMT modes. For detailed information about this command, refer to the following website (you need to register first):

<http://www.ibm.com/developerworks/wikis/display/LinuxP/Performance%20FAQs>

Single Instruction Multiple Data

Single Instruction Multiple Data (SIMD), also called *vector*, instructions provide a concise and efficient way to express data-level parallelism (DLP⁵), which is explained in the footnote. With SIMD instructions, fewer instructions are required to perform the same data computation, resulting in lower fetch, decode, and dispatch bandwidth and consequently higher power efficiency.

The POWER7 processor adds another SIMD instruction called Vector Scalar Extension (VSX), which is based on the Vector Media Extensions (VMX) instruction. This technology helps to improve POWER7’s performance, especially in High Performance Computing (HPC) projects.

For more information about SIMD, refer to the IBM research paper, “IBM Power Architecture,” at the following website:

[http://domino.research.ibm.com/library/cyberdig.nsf/papers/8DF8C243E7B01D948525787300574C77/\\$File/rc25146.pdf](http://domino.research.ibm.com/library/cyberdig.nsf/papers/8DF8C243E7B01D948525787300574C77/$File/rc25146.pdf)

Large scale-up capability

The IBM Power 795 can provide 256 cores in one LPAR⁶. It provides high performance and scalability for a large scale-up single system image, from which many workloads get benefit (for example, online transaction processing (OLTP), ERP scale-up, and so forth).

⁴ Regarding AIX OS, AIX 5.3 does not support SMT4, and AIX5.3 does not support POWER7 mode.

⁵ Data level parallelism (DLP) consists of simultaneously performing the same type of operations on separate data values, using multiple functional units, with a single instruction.

Storage protection keys

Power storage protection keys provide hardware-enforced access mechanisms for memory regions. Only programs that use the correct key are allowed to read or write to protected memory locations. The POWER7 mode provides 16 more storage protection keys than the POWER6+ mode.⁷

For more information about POWER7 storage protection keys, refer to the IBM white paper, *POWER7 System RAS*, at the following website:

<ftp://public.dhe.ibm.com/common/ssi/ecm/en/pow03056usen/POW03056USEN.PDF>

Processor compatibility mode performance considerations

From Table 7-2 on page 256, if you select POWER7 mode, it supports SMT4 mode and also enables other features to improve performance. If there is not necessarily a requirement⁸ to choose POWER6/6+ mode, we suggest choosing POWER7 mode. POWER7 mode is flexible, and you can choose an appropriate SMT mode to get the best performance. Applications that are single process and single threaded might benefit from running in ST mode. ST mode can be beneficial in the case of a multi-process application where the number of application processes is smaller than the number of cores assigned to the partition. Applications that do not scale with a larger number of CPUs might also benefit from running in SMT2 or ST mode instead of SMT4, because the lower number of SMT threads means a lower number of logical CPUs. For detailed information about SMT, refer to 7.6.6, “Simultaneous multithreading (SMT)” on page 294.

Configuration for processor compatibility mode

Processor compatibility mode is one feature of an LPAR that we can configure by editing the LPAR's profile. Figure 7-6 on page 259 shows the configuration window from the HMC.

⁶ At the time of writing, if you want to configure more than 128 cores in one LPAR with the FC4700 processor, you need to purchase software key FC1256 and install it in the server. The name of this code is “AIX Enablement for 256-cores LPAR”.

⁷ In POWER6+ and POWER7, the hypervisor reserves one storage protection key. So, the number of maximum available storage protection keys for OS is 15 (for POWER6+) and 31 (for POWER7).

⁸ If your AIX Version is AIX 5.3, you cannot choose POWER7 mode. If the LPAR is in the Live Partition Mobility (LPM) environment and POWER7 mode does not support partition mobility, you cannot choose POWER7 mode, too. For detailed information, refer to the IBM information center link:

<http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7hc3/iphc3pcm.htm>

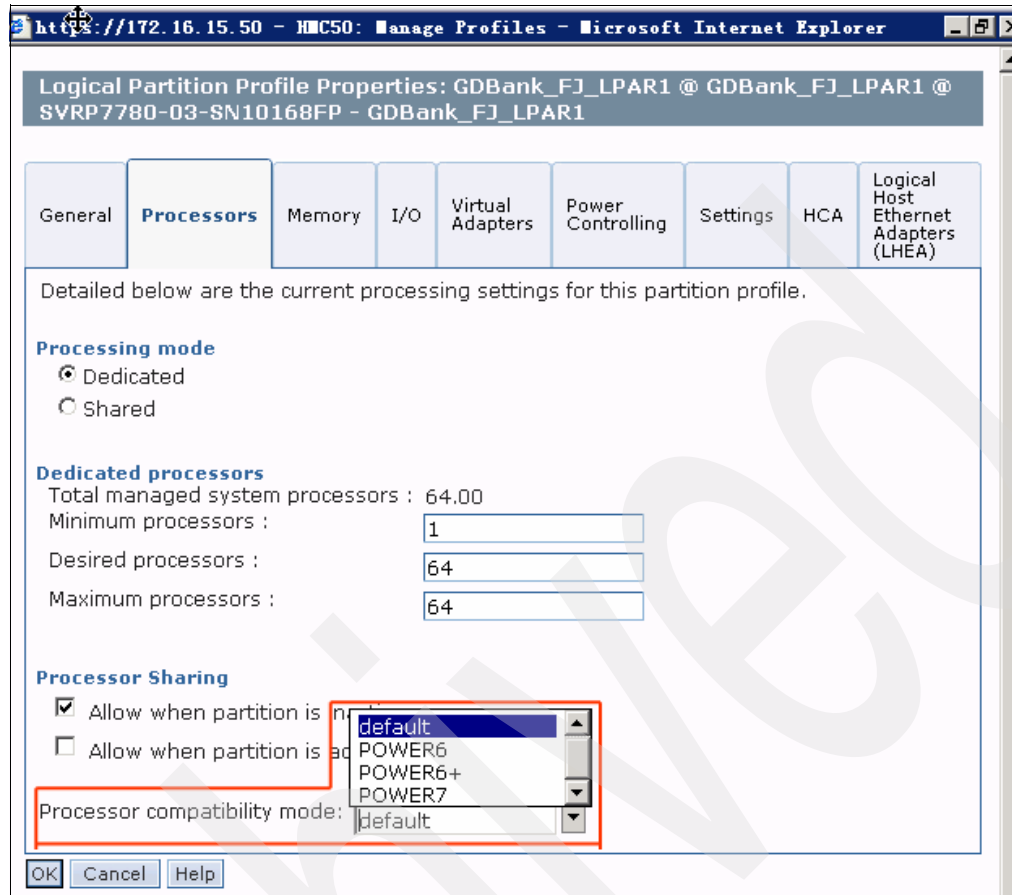


Figure 7-6 Configuring the processor compatibility mode

7.3.2 TurboCore and MaxCore modes

POWER7 high-end servers (780 and 795) offer two processor running modes:

- ▶ **MaxCore (default)**

The MaxCore mode allows for all processor cores in the system to be activated.

- ▶ **TurboCore**

The TurboCore mode allows for half of the processor cores in the system to be activated, but the cores run at a higher speed and have access to the entire L3 cache on the chip.

The TurboCore mode allows the processor cores to execute at a higher frequency (about 7.25% higher) and to have more processor cache per core. In general, higher frequency and more cache often provide better performance. Refer to Figure 7-7 on page 260.

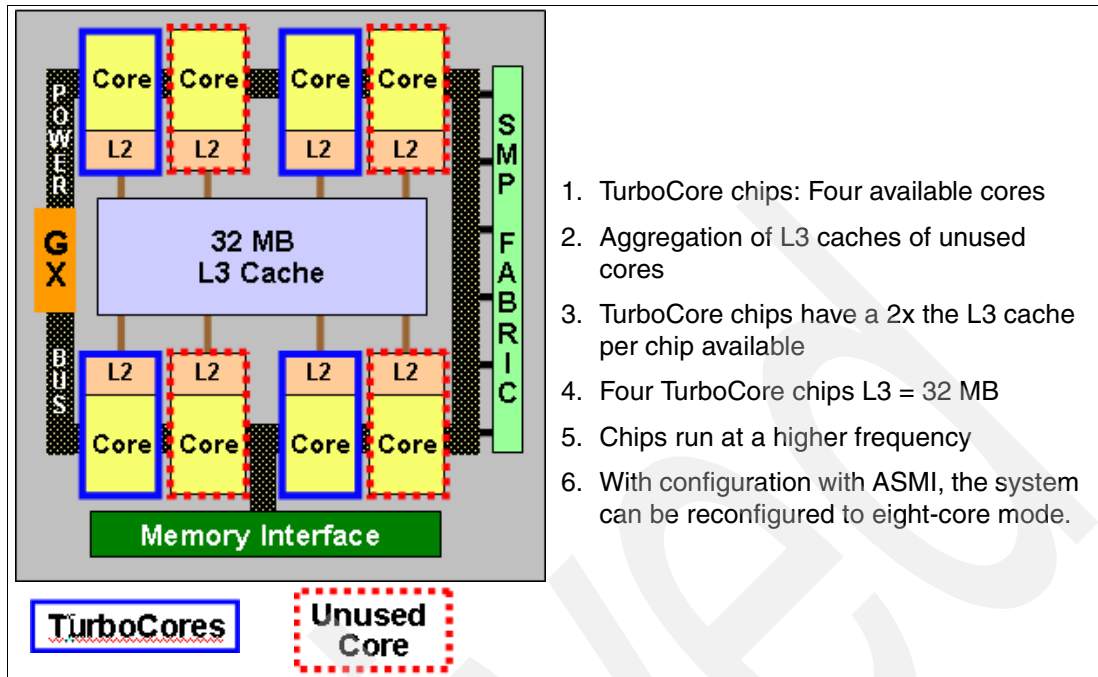


Figure 7-7 TurboCore design

Table 7-4 shows the processor differences between the MaxCore and TurboCore modes.

Table 7-4 Processor difference between MaxCore and TurboCore

Mode	Process feature of Power 780	Process feature of Power 795
MaxCore	4982/3.86 GHz	4700/4.0 GHz
TurboCore	4982/4.14 GHz	4700/4.25 GHz

Considerations with TurboCore mode

TurboCore provides a higher frequency core and more cache per core, which are normally good for performance. We also know that to make use of these positive attributes, the system's active core placement needed to change from having eight cores per chip to having four cores per chip. So, using TurboCore can also mean a change in the number of chips and perhaps in the number of drawers.

Here is an example to compare the performance between MaxCore and TurboCore. After switching to TurboCore, the performance increases about 20%. Refer to Figure 7-8 on page 261.

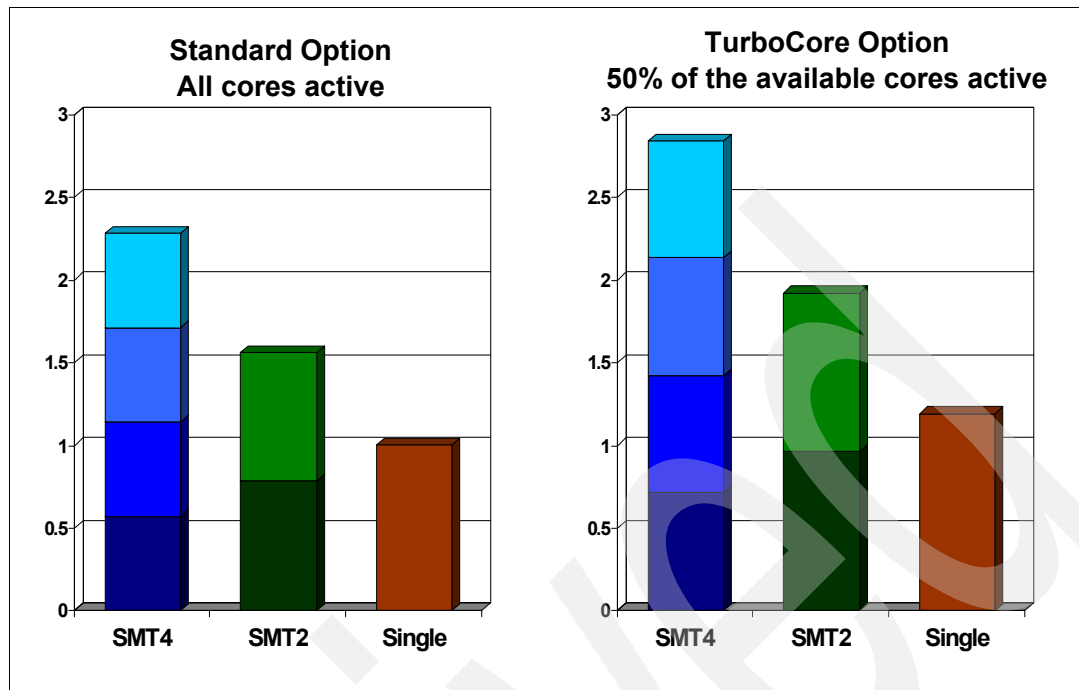


Figure 7-8 Comparing the performance difference between MaxCore and TurboCore

Enabling TurboCore in one server, for a given partition, means an increase in the probability of longer access latencies to memory and cache. But, partition placement and workload type can influence these probabilities. As a result, the positive benefits of TurboCore's higher frequency and increased cache also have the potential of being offset to various extents by these longer latency storage accesses.

Configuration of TurboCore

For information about enabling and disabling TurboCore mode, refer to 2.3, "TurboCore and MaxCore technology" on page 28.

Case study

There are case studies that relate to POWER7 TurboCore performance. See the IBM white paper *Performance Implications of POWER7 Model 780's TurboCore Mode*, which is available at the following site:

http://www-03.ibm.com/systems/resources/systems_i_pwrsysperf_turbocore.pdf

7.3.3 Active Memory Expansion (AME)

Active Memory Expansion is an innovative POWER7 technology that allows the effective maximum memory capacity to be up to 100% larger than the true physical memory maximum for AIX 6.1 and later partitions.

AME relies on the compression of in-memory data to increase the amount of data that can be placed into memory and thus expand the effective memory capacity of a POWER7 system. The in-memory data compression is managed by the system, and this compression is transparent to applications and users. Figure 7-9 on page 262 shows the memory structure change after applying AME.

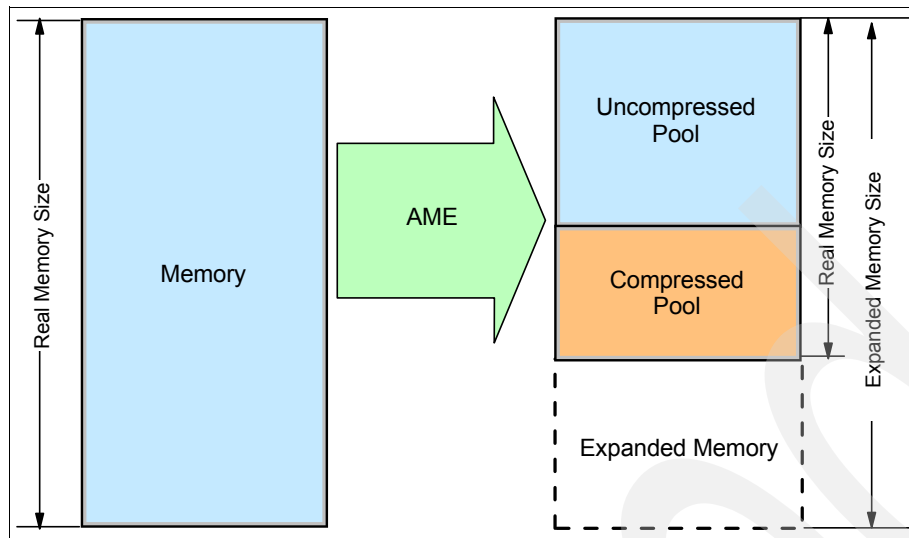


Figure 7-9 Memory structure change after applying AME

The AME feature can bring these benefits to clients:

- ▶ AME increases the system's effective memory capacity.
- ▶ AME enables a system to process more work by increasing the system's effective memory capacity.

Because AME relies on memory compression, additional CPU utilization is consumed when AME is in use. The amount of additional CPU utilization needed for AME varies based on the workload and the level of memory expansion being used.

For more information about AME, see the IBM white paper, *Active Memory Expansion: Overview and Usage Guide*, which is available at the following website:

<ftp://public.dhe.ibm.com/common/ssi/en/pow03037usen/POW03037USEN.PDF>

Active Memory Expansion considerations

Application performance in an AME environment depends on multiple factors, such as the memory expansion factor, application response time sensitivity, and how compressible the data is.

Figure 7-10 on page 263 illustrates the general relationship between application response time, application throughput, CPU utilization, and the percentage of memory expansion. The CPU utilization increases with a larger percentage memory expansion due to more compression and decompression activity, which impacts the application response time. An increase in application response time often results in less application throughput.

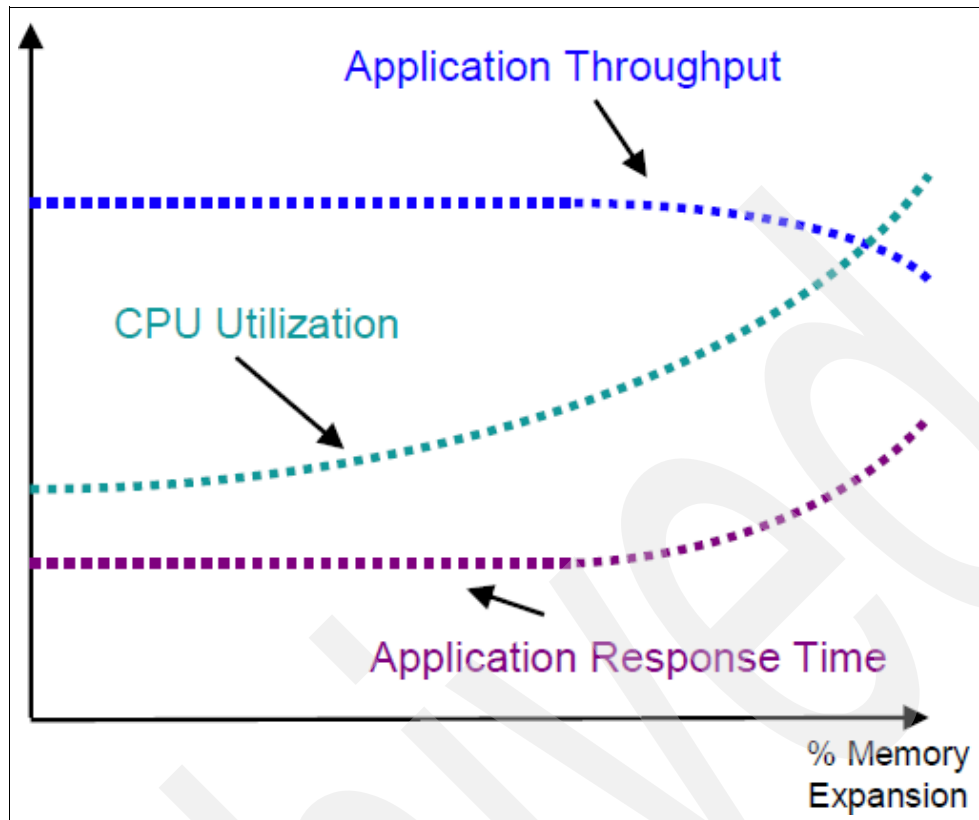


Figure 7-10 Generation performance factors relationship with AME

In AIX (from 6.1.0.4 SP2 or 7.1), the `amepat` command is useful for sizing if you want to apply AME technology. This command reports AME information and statistics, as well as provides an advisory report that assists in planning the use of AME for existing workloads. See 7.8.7, “Monitoring Active Memory Expansion (AME) statistics” on page 317 to get more information about the command.

Various kinds of applications result in various behaviors after enabling the AME function, for example, the SAP ABAP application can save more memory after you turn on AME, but the SAP Java application might gain less benefit from AME. Figure 7-11 on page 264 shows the difference between them.

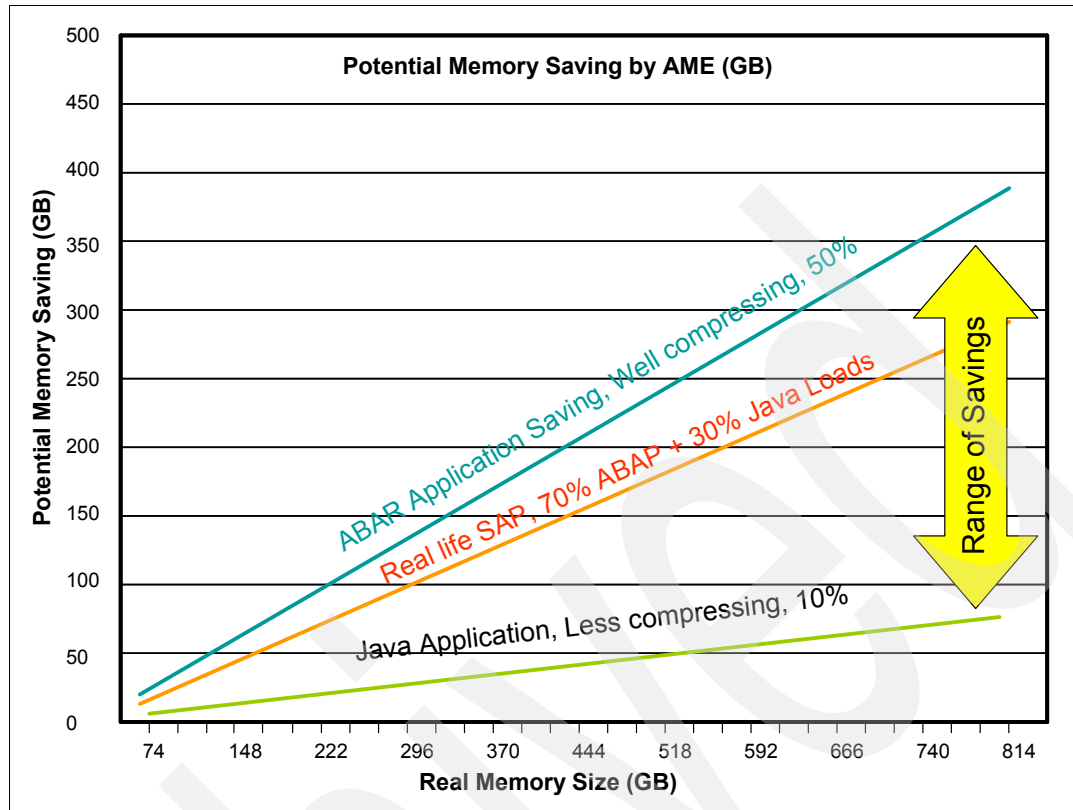


Figure 7-11 Potential real memory saving by AME for SAP

For more information about the details of Figure 7-11, refer to the following website:

[https://www-927.ibm.com/servers/eserver/storageplaza/bert.nsf/files/2010CSIIelective-presentations/\\$File/E08%20Improving%20SAP%20flexibility%20with%20POWER.pdf](https://www-927.ibm.com/servers/eserver/storageplaza/bert.nsf/files/2010CSIIelective-presentations/$File/E08%20Improving%20SAP%20flexibility%20with%20POWER.pdf)

For more performance considerations about AME, refer to the IBM white paper, *Active Memory Expansion Performance*, which is available at the following website:

<ftp://ftp.software.ibm.com/common/ssi/sa/wh/n/pow03038usen/POW03038USEN.PDF>

Configuring Active Memory Expansion

Regarding planning for AME and configuring AME, refer to “Active Memory Expansion (AME) configuration” on page 233.

Case study

For detailed testing and measuring information about how to apply the AME function with SAP ABAP application, refer to Chapter 3 of the IBM white paper, “*Active Memory Expansion Performance*”. This paper describes the ERP workload of performance measurements, single partition throughput, and the server and is available at the following website:

<ftp://ftp.software.ibm.com/common/ssi/sa/wh/n/pow03038usen/POW03038USEN.PDF>

7.3.4 Logical memory block size

Processors use memory to temporarily hold information. Memory requirements for LPARs depend on the LPAR configuration, assigned I/O resources, and applications used.

Logical memory block (LMB) size can be assigned in increments of 16 MB, 32 MB, 64 MB, 128 MB, and 256 MB. The default memory block size varies according to the amount of configurable memory in the system.

Table 7-5 shows the default logical memory block size in various systems.

Table 7-5 Default memory block size used for varying amounts of configurable memory

Amount of configurable memory	Default logic memory block size
Less than 4 GB	16 MB
Greater than 4 GB up to 8 GB	32 MB
Greater than 8 GB up to 16 GB	64 MB
Greater than 16 GB up to 32 GB	128 MB
Greater than 32 GB	256 MB

For more information about the logical memory block size, refer to the following website:

<http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7hat/iphat1parmemory.htm>

Considerations with logical memory block size

To select a reasonable logical block size for your system, consider both the desired performance and the physical memory size. Use the following guidelines when selecting logical block sizes:

- ▶ On systems with a small amount of memory installed (2 GB or less), a large logical memory block size results in the firmware consuming an excessive amount of memory. Firmware must consume at least 1 logical memory block. As a general rule, select the logical memory block size to be no greater than 1/8th the size of the system's physical memory.
- ▶ On systems with a large amount of installed memory, small logical memory block sizes result in a large number of logical memory blocks. Because each logical memory block must be managed during the system boot, a large number of logical memory blocks can cause boot performance problems.
- ▶ Ensure that the logical memory block (LMB) size is the same on the source and destination systems during Live Partition Mobility (LPM).

Configuring logical memory block size

The memory block size can be changed by using the Integrated Virtualization Manager (IVM), the Systems Director Management Console (SDMC) command-line interface, or the Logical Memory Block Size option in the Advanced System Management Interface (ASMI).

System restart: The logical memory block size can be changed at run time, but the change does not take effect until the system is restarted.

In this section, we introduce how to change the logical memory block size via the ASMI.

To perform this operation, you must have one of the following authority levels:

- ▶ Administrator
- ▶ Authorized service provider

To configure the logical memory block size, perform the following steps:

1. On the ASMI Welcome pane, specify your user ID and password, and click **Log In**.
2. In the navigation area, expand **Performance Setup**.
3. Select **Logical Memory Block Size**.
4. In the right pane, select the logical memory block size and click **Save Settings**, as shown in Figure 7-12.



Figure 7-12 Configuring the Logical Memory Block size

Remember: You must shut down and restart your managed system for the change to take effect.

7.3.5 System huge-page memory

IBM POWER6 servers or later can support 4 KB, 64 KB, 16 MB, and 16 GB page sizes. In this topic, we introduce the 16 GB page size, which is called the *huge-page memory* size.

Using a larger virtual memory page size, such as 16 GB, for an application's memory can significantly improve the application's performance and throughput due to the hardware efficiencies that are associated with larger page sizes.

To use huge-page memory, a huge-page memory pool needs to be created when the managed system is in the powered-off state. After a managed system has been configured with a 16 GB huge-page pool, a system administrator can assign 16 GB huge pages to partitions by changing a partition's profile.

Before specifying the value for huge-page memory, you must determine which applications might benefit from this feature.

For detailed information about how to determine huge-page memory requirements, and considerations for calculating the huge-page values, refer to the following website:

http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/ipha1_p5/calculatinghugepgmemory.htm

Huge-page memory considerations

Consider these factors when using the huge-page memory feature:

- ▶ Huge-page memory is intended to only be used in high-performance environments, because it can improve performance in specific environments that require a high degree of parallelism, for example, DB2 databases. You can specify the huge-page memory that can be used for the shared-memory buffer pools in DB2.
- ▶ The huge-page memory allocation cannot be changed dynamically.
- ▶ At the time of writing this book, huge-page memory was not supported when suspending an LPAR. Also, huge-page memory was not supported with active LPM. Huge-page memory is supported in inactive partition mobility solutions.
- ▶ After setting huge-page memory for your server and LPARs, you can monitor it from the HMC, SDMC, or the operating system.

Configuring the system huge-page memory pool

Follow this example of using the ASMI to configure a system huge-page memory pool.

To set up your system with larger memory pages, perform the following steps:

1. On the ASMI Welcome pane, specify your user ID and password, and click **Log In**.
2. In the navigation area, expand **Performance Setup**.
3. Select **System Memory Page Setup**.
4. In the right pane, select the settings that you want.
5. Click **Save Settings** and power on the server. Refer to Figure 7-13.

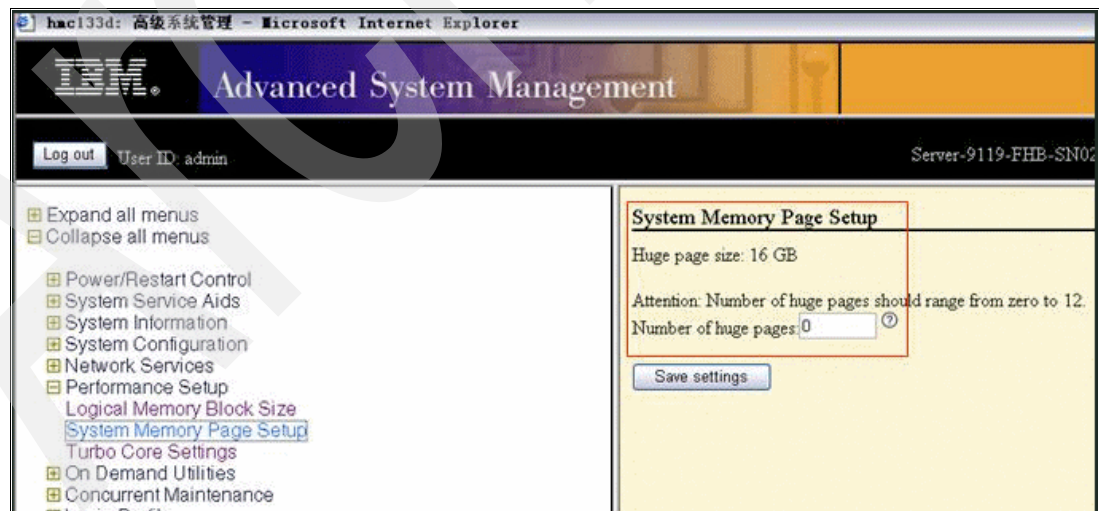


Figure 7-13 Configuration of system memory huge-page setup

Configuring huge-page memory for LPARs

To configure huge-page memory values for an LPAR, you can use the HMC, SDMC, and IVM. This section introduces the configuration method via the HMC:

1. In the navigation pane, expand **Systems Management** → **Servers**.
2. Select the server that has the LPAR that you want to configure.
3. In the work pane, select the LPAR for which you want to set huge-page memory values.
4. Select **Configuration** → **Manage profiles**. The Managed Profiles window opens.
5. Select the LPAR that you want to configure.
6. Select **Actions** → **Edit**. The Logical Partition Profile Properties window opens.
7. Click the **Memory** tab.
8. Assign the Huge Page memory for this partition profile, and Click **OK**.

Figure 7-14 shows the configuration window of the huge page size for an LPAR.

The screenshot shows a web browser window with the address bar displaying `https://hmc29.pba.ihost.com - hmc29: Manage Profiles -`. The main title of the window is **Logical Partition Profile Properties: p29n01 @ Server-9179-MHB-SN105**. Below the title, there are several tabs: **General**, **Processors**, **Memory** (selected), **I/O**, **Virtual Adapters**, and **Power Controlling**. The **Memory** tab is active, showing detailed memory settings. The text "Detailed below are the current memory settings for this partition profile" is displayed. Under the heading **Dedicated Memory**, the following information is shown: "Installed memory (MB): 1048576" and "Current memory available for partition usage (MB) : 961280". Below this, there are three rows of memory settings: "Minimum memory : 2 GB 0 MB", "Desired memory : 4 GB 0 MB", and "Maximum memory : 16 GB 0 MB". Each row has a text input field for the value in GB and a spinner control for the value in MB. Below these settings, there is a section for the Barrier Synchronization Register (BSR) with the text "Specify the Barrier Synchronization Register BSR for this profile". It shows "Available BSR arrays: 256" and "BSR arrays for this profile: 0". At the bottom of the window, there is a red-bordered box containing the **Huge Page Memory** settings: "Page size (in GB) : 16", "Configurable pages : 5", "Minimum pages : 1", "Desired pages : 6", and "Maximum pages : 7". Each of these settings has a corresponding spinner control. At the very bottom of the window, there are three buttons: **OK**, **Cancel**, and **Help**.

Figure 7-14 Configuring the huge page memory for an LPAR

Monitoring for huge-page memory from the HMC

To monitor huge-page memory from the HMC, Select **Properties** for the managed server. Click the **Advanced** tab. Look at the current server's huge-page memory state, as shown in Figure 7-15.

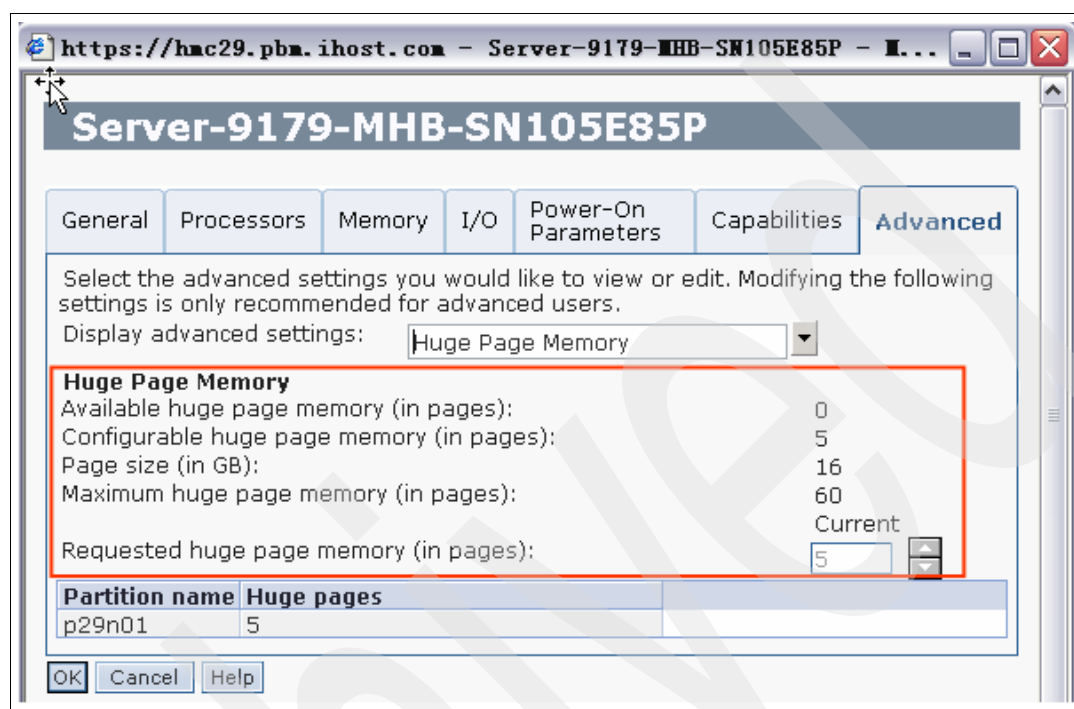


Figure 7-15 Monitoring a server's huge-page memory state from the HMC

To monitor the state of an LPAR's huge-page memory state from the HMC, select **Properties** of a Logical Partition. Click the **Hardware** tab. Click the **Memory** tab. Refer to Figure 7-16.

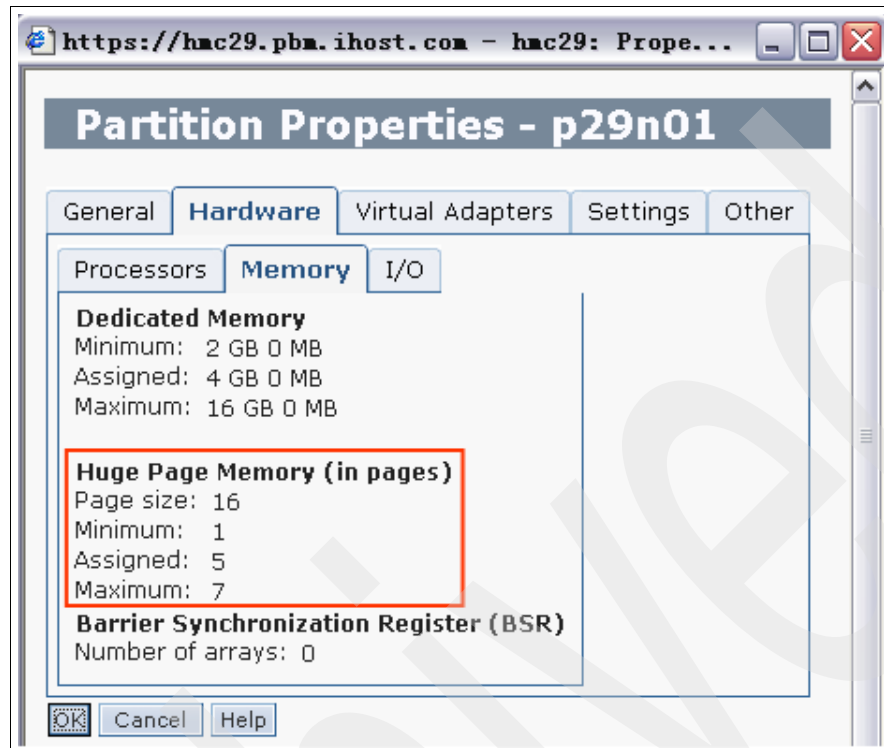


Figure 7-16 Monitoring an LPAR's huge-page memory state from the HMC

Monitoring huge-page memory from AIX

We can see the current state of an LPAR's huge-page memory by executing the **svmon** command in the AIX environment, as shown in Example 7-1.

Example 7-1 Monitoring the huge-page memory from AIX

```
p29n01:/ # svmon -G
```

	size	inuse	free	pin	virtual	mmode
memory	22020096	21468793	551303	21334339	445183	Ded
pg space	1048576	3474				
	work	pers	clnt	other		
pin	259427	0	0	103392		
in use	445183	0	52090			
PageSize	PoolSize	inuse	pgsp	pin	virtual	
s 4 KB	-	308073	3474	212275	255983	
m 64 KB	-	11825	0	9409	11825	
S 16 GB	5	0	0	5	0	

Case study

You can read about one test case, which showed performance improvement after using huge-page memory, at the following website:

<http://www.ibm.com/developerworks/data/library/techarticle/dm-0606kamath/index.html>

7.4 Performance considerations with hardware RAS features

In February 2010, IBM announced the first models in a new generation of Power servers based on the POWER7 microprocessor. POWER7 servers provide an extensive set of hardware features related to reliability, availability, and serviceability (RAS). At the same time, POWER7 servers deliver industry-leading performance through their architectural design, including modularity, timing closure, and efficiency.

The POWER7 RAS architecture is intended to work with the hardware, independently of any operating system. By using a dedicated service processor, there is usually no reason to turn off the features or tune them for performance.

Most recoverable errors are handled during run time by the dedicated service processor. This processor, independently of any system processor, has dedicated access to detailed error information from various processor and memory components that can be accessed and assessed during run time without affecting the performance of the system.

Typically, transient recoverable errors can be handled quickly within the hardware and can have no effect on performance.

Frequently occurring, but recoverable faults can be eliminated by using the built-in redundancy capabilities of a system. For example, customized dual inline memory modules (DIMMs) that are used in high-end systems have a spare memory module for each rank, which can be substituted for a faulty memory module.

If the handling of a recoverable fault causes a measurable effect on performance, it is the system design goal to report the fault through the error-reporting structure and, as needed, request repair. Until the repair is completed, performance might be affected. For example, a processor core has been determined to be unable to continue processing instructions. The RAS feature that is known as *Alternate Processor Recovery* can seamlessly migrate the workload that is being run on that core to another processor core. If the processor core in question was unlicensed in the system at time (referred for later capacity update), the operation does not affect the current system performance. Otherwise, the overall performance of the system is affected by the temporary deallocation of one core.

However, the RAS feature, Active Memory Mirroring for the hypervisor, might have an effect on performance even in a system that runs well.

7.4.1 Active Memory Mirroring for the hypervisor

Active Memory Mirroring for the hypervisor is a new RAS feature being introduced on the Power 795⁹ that is designed to eliminate the potential for a complete system outage as a result of an uncorrectable error in memory. Active Memory Mirroring requires that in each node of a Power 795 system at least one processor module must be fully configured with eight DIMMs. When Active Memory Mirroring for the hypervisor is enabled (default), the Power 795 system maintains two identical copies of the system hypervisor in memory at all times. Both copies are simultaneously updated with any changes. This design might result in a minor memory performance effect, and less memory might be available for partitions. If you want to disable Active Memory Mirroring for the hypervisor, refer to 2.1.1, “Active Memory Mirroring for the hypervisor on Power 795” on page 13.

⁹ In the IBM POWER7 product line announcement of October 2011, the Active Memory Mirroring feature is introduced in the Power 780 (9179-MHC) as a standard feature and in the Power 770 (9117-MMC) as an optional feature.

7.5 Performance considerations with Power virtualization features

IBM PowerVM on Power Systems servers can be consider virtualization without limits. Businesses are turning to PowerVM virtualization to consolidate multiple workloads into fewer systems, increasing server utilization to ultimately help reduce cost. PowerVM provides a secure and scalable virtualization environment for AIX, IBM i, and Linux applications that is built upon the advanced RAS features and leading performance of the Power Systems platform.

In this section, we introduce performance considerations when implementing PowerVM features. We discuss the following topics:

- ▶ Dynamic logical partitioning (DLPAR)
- ▶ Micro-partitioning
- ▶ Linux on Power (LX86)
- ▶ Virtual I/O server
- ▶ Active Memory Sharing (AMS)
- ▶ Live Partition Mobility (LPM)

7.5.1 Dynamic logical partitioning (DLPAR)

Dynamic logical partitioning is available on POWER4-based System p systems with microcode updates that are dated October 2002 or later. DLPAR increases the flexibility of logically partitioned systems by allowing you to dynamically add and remove processors, memory, I/O slots, and I/O drawers from active LPARs.

You can perform the following operations with DLPAR:

- ▶ Move a resource from one partition to another partition
- ▶ Remove a resource from a partition
- ▶ Add a resource to a partition

The resource includes processors, memory, and I/O slots.

For detailed information about how to use dynamic LPAR, refer to the IBM Redbooks publication, *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590, available at the following website:

<http://www.redbooks.ibm.com/abstracts/sg247590.html>

DLPAR considerations

You need to be aware of several considerations when using dynamic LPAR:

- ▶ When removing or adding memory from a partition, the time that it takes to complete a DLPAR operation is relative to the number of memory chunks being removed.
- ▶ The affinity logical partitioning configuration allocates CPU and memory resources in fixed patterns based on multi-chip module (MCM) boundaries. The HMC does not provide dynamic reconfiguration (DR) of processor or memory support on affinity partitions. Only the I/O adapter resources can be dynamically reconfigured when you run affinity logical partitioning.
- ▶ When you remove memory from a partition, the DR operation succeeds even if there is not enough free physical memory available to absorb outgoing memory, provided there is enough paging space available instead of physical memory. Therefore, it is important to monitor the paging statistics of the partition before and after a DR memory removal. The

virtual memory manager is equipped to handle paging; however, excessive paging can lead to performance degradations.

- In certain cases, the DLPAR operation breaks memory affinity with the processor, which affects performance.

Important: Dynamically partitioning large memory pages (16 MB page size) is not supported. A memory region that contains a large page cannot be removed.

There are tools that support DR operations. These tools are designed to recognize configuration changes and adjust their reporting accordingly. The following tools provide DLPAR support: **topas**, **sar**, **vmstat**, **iostat**, and **rmss**.

For detailed information about monitor tools, refer to the IBM Redbooks publication, *AIX 5L Performance Tools Handbook*, SG24-6039, which is located at the following website:

<http://www.redbooks.ibm.com/abstracts/sg246039.html?Open>

For more information about DLPAR performance considerations, refer to the information center website:

http://publib.boulder.ibm.com/infocenter/aix/v7r1/index.jsp?topic=/com.ibm.aix.prftungd/doc/prftungd/dyn_log_part.htm

7.5.2 Micro-partitioning

An LPAR that utilizes processor resources from a shared processor pool is known as a *micro-partition LPAR*. Micro-partitioning support provides flexible and efficient use of system hardware resources by allowing physical processors to be shared (time-sliced) between micro-partitions.

In a client's production environment, in general, there are multiple system images within one Power server, and each of these system images runs on a separate hardware system with sufficient capacity to handle spikes in processor requirements, but each system is underutilized.

Micro-partitioning support can allow these system images to be consolidated on a single set of physical processors, allowing more efficient use of hardware resources, and providing a reduction in the physical footprint required by multiple systems. Micro-partitioning with uncapped processing provides more granularity for CPU resource balancing and allows idle CPU cycles to be recovered and used by other partitions.

You can take advantage of the following advantages with micro-partitioning:

- Optimal resource utilization
- Rapid deployment of new servers
- Application isolation

For detailed information about concepts and how to use micro-partitioning, refer to the IBM Redbooks publication, *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940, which is located at the following website:

<http://www.redbooks.ibm.com/abstracts/sg247940.html>

Micro-partitioning considerations

The overall benefit of micro-partitioning is the increased utilization of system resources by applying only the required amount of processor resource that is needed to each partition. To ensure that the hypervisor's memory pages keep track of all virtual partitions, consider the capacity requirements carefully when choosing values for the attributes of the virtual partitions.

CPU-intensive applications, such as high-performance computing applications, might not be suitable for a micro-partitioning environment. If an application uses most of its entitled processing capacity during execution, use a dedicated processor partition to handle the demands of the application.

Tips when you deploy micro-partitioning

Consider these tips when implementing micro-partitioning:

- ▶ Correctly determine the micro-partition processor allocation. Sizing the partition too small can significantly increase response times. In addition, the processor efficiency is affected more with smaller partitions or more partitions.
- ▶ On POWER7 systems, consider using uncapped processing to better utilize idle processor cycles in other partitions in the shared pool.
- ▶ Limit the number of micro-partitions that are active at any one time. Workloads that are cache sensitive or have response time criteria might suffer with the increased contention that micro-partitioning places on shared resources.
- ▶ Balance the memory DIMMs and I/O across the modules. Use the same size memory DIMMs on the modules, whenever possible, to help reduce latencies caused by remote references and avoid "hot spots".
- ▶ On POWER7 systems, the hypervisor attempts to optimize memory allocations at full system startup. If after the system has started, you change a partition's memory allocation on a multi-module system, you can introduce more remote memory references as memory is "reallocated" from its initial optimal allocation to another module. If you suspect that this situation has happened, another full system startup re-optimizes the memory allocations.

Configuring virtual processors in a shared partition environment

Consider this additional information when configuring virtual processors in a shared partition environment:

- ▶ When creating a capped partition, for maximum processor efficiency and partition CPU capacity, the number of desired virtual processors needs to be the minimum that can consume the desired entitled capacity, for example:
 - If the desired entitled capacity is 3.6 processing units, the number of desired virtual processors must be 4.
 - If the desired entitled capacity is 0.75 processing units, the number of desired virtual processors must be 1.
- ▶ When creating an uncapped partition, for maximum processor efficiency and partition CPU capacity, follow these guidelines:
 - Do not make the number of virtual processors for a partition greater than the number of processors in the shared processor pool. The number of processors in the shared pool is the maximum number of physical processors that a partition can use concurrently.
 - Do not set the number of virtual processors for a partition to a number that is greater than the number of available processing units. Other shared partitions are using their entitled processing units, so, in many cases, the entire shared pool size is not available for a single partition to use.

- Where possible, set the partition's entitled processing units as close to the anticipated CPU processing requirements as possible. The more CPU that processing partitions use as uncapped (for example, beyond their entitlement), the greater the processor efficiency effects that are caused by increased virtual processor switching.
- When attempting to take advantage of unused shared pool resources, set the number of virtual processors close to the expected capacity that you are trying to achieve for that partition.
- ▶ Setting virtual processors to higher values usually results in reduced processor efficiency and can result in decreased performance from increased contention.

For detailed information about micro-partition performance with Power Systems servers, refer to following IBM white paper:

<http://www-03.ibm.com/systems/resources/lparperf.pdf>

For more information about IBM PowerVM and micro-partition, refer to following information center website:

http://publib.boulder.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.prftungd/doc/prftungd/micro_part.htm

Case studies

The following case studies were performed while implementing IBM PowerVM virtualization:

Case 1

This IBM Redpaper introduces an actual application (life sciences application) benchmark test with Power servers. The publication includes three applications in life sciences and how the availability of a pool of virtual processors improves the time to the solution. A partition that has exhausted its resources can take advantage of a pool of shared virtual processors, provided that the shared virtual processors are not required by other partitions. For more information, refer to the IBM Redpaper, *Life Sciences Applications on IBM POWER5 and AIX 5L Version 5.3: Virtualization Exploitation through Micro-Partitioning Implementation*, REDP-3966, which is located at the following website:

<http://www.redbooks.ibm.com/redpapers/pdfs/redp3966.pdf>

Case 2

This IBM white paper describes how micro-partitioning can be deployed on IBM Power servers for consolidation. It includes an example consolidation scenario and explores the performance robustness of micro-partitioning in a demanding transactional environment. For more information about this case, refer to the following website:

<ftp://ftp.software.ibm.com/software/uk/itsolutions/datacentreooptimisation/virtualization-consolidation/server/ibm-system-p5-570-server-consolidation-using-power5-virtualization.pdf>

7.5.3 PowerVM Lx86

PowerVM Lx86 supports migrating most 32-bit x86 Linux applications to any Power Systems or BladeCenter model with POWER7 or POWER6 processors, or with IBM Power architecture technology-based blade servers. Best of all, no native porting or application upgrade is required for running most x86 Linux applications. PowerVM Lx86 offers these advantages:

- ▶ Exceptional performance and scalability, allowing much greater consolidation of workloads
- ▶ Improved service quality through leadership availability and security features

- The ability to dynamically optimize the mix of processor, memory, disk, and network resources with optional IBM PowerVM virtualization technology

For more information about PowerVM Lx86 and how to set up the Lx86 environment, refer to the IBM Redpaper publication, *Getting Started with PowerVM Lx86*, REDP-4298, which is located at the following website:

<http://www.redbooks.ibm.com/abstracts/redp4298.html>

Lx86 considerations

PowerVM Lx86 runs most x86 Linux applications, but PowerVM Lx86 cannot run applications with these characteristics:

- Directly access hardware, for example, 3D graphics adapters.
- Require nonstandard kernel module access or use kernel modules that are not provided by the Linux on Power operating system distribution.
- Do not use only the Intel IA-32 instruction set architecture as defined by the *1997 Intel Architecture Software Developer's Manual consisting of Basic Architecture*, 243190, *Instruction Set Reference Manual*, 243191, and the *System Programming Guide*, 243192, dated 1997.
- Do not run correctly on RHEL 4 starting with Version 4.3 or Novell SLES 9 starting with Version SP3 or Novell SLES 10.
- Require RHEL 5, a Linux distribution currently unsupported by PowerVM Lx86, to run.
- Are Linux/x86-specific system administration or configuration tools.
- Require x86 real mode.

Regarding performance, Figure 7-17 on page 277 shows the PowerVM Lx86 application translation process. The translation is a three-stage process:

1. **Decoding:** x86 binary instructions from the x86 binary are decoded as the application requests them.
2. **Optimization:** The optimization is iterative, so more optimization is done on frequently used code.
3. **Generation of Power code:** Frequently used code is stored in memory, so it does not need to be translated again the next time that it runs.

From the translation process, we can find when an x86 application is executed in the PowerVM Lx86 environment, because more CPU and memory resources are needed than in a pure x86 environment. So, there are performance issues when migrating to the PowerVM Lx86 environment.

The performance of certain x86 Linux applications running on PowerVM Lx86 might significantly vary from the performance obtained when these applications run as a native port. There are various architectural differences between x86 and Power architecture that can affect the performance of translated applications. For example, translating dynamically generated code, such as Java byte code, is an ongoing translation process, which can affect the performance of x86 Java applications using an x86 Java virtual machine.

Floating-point applications running under x86 have a separate default precision level from Power architecture, so translating between these levels can have additional performance penalties. Also, translating and protecting multithreaded applications can incur an additional performance overhead as the translator works to manage shared memory accesses. IBM suggests that clients carefully consider these performance characteristics when selecting the best method for enabling applications for their environment.

For detailed information about PowerVM Lx86 performance, refer to the following website:

<http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=ca&infotype=an&appname=iSource&supplier=897&letternum=ENUS208-010>

Considerations: PowerVM Lx86 is not recommended with applications that are highly computational in nature, highly performance sensitive, or make heavy use of Java.

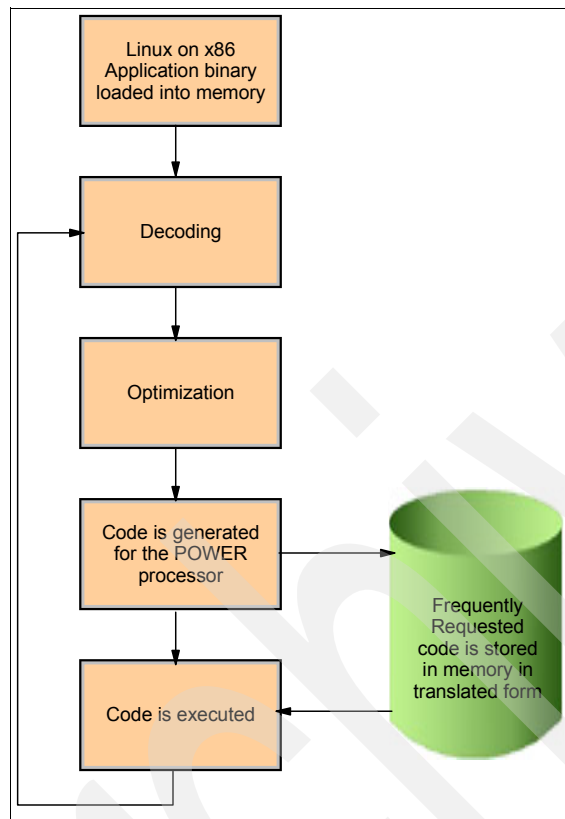


Figure 7-17 PowerVM Lx86 translation process

Case study

This IBM white paper provides performance comparisons between PowerVM Lx86 and x86-based environments and is available at the following website:

http://www-03.ibm.com/systems/power/software/virtualization/Whitepapers/powervm_x86.html

7.5.4 Virtual I/O server

The virtual I/O server is part of the PowerVM Editions. The virtual I/O server is software that is located in an LPAR. This software facilitates the sharing of physical I/O resources between client LPARs within the server.

As a result, you can perform the following functions on client LPARs:

- ▶ Share SCSI devices, Ethernet adapters, and FC adapters
- ▶ Expand the amount of memory available to LPARs and suspend and resume LPAR operations by using paging space devices

For more information about recent virtual I/O server features, refer to *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940, which is located at the following site:

<http://www.redbooks.ibm.com/abstracts/sg247940.html>

The virtual I/O server has an extremely important role in the IBM PowerVM solution, because it is the foundation of many advanced PowerVM features, such as AMS and LPM.

In the next section, we introduce performance considerations about CPU configuration, virtual SCSI, virtual network, and virtual FC.

Considerations with virtual SCSI

Using virtual SCSI (vSCSI), client LPARs can share disk storage and tape or optical devices that are assigned to the virtual I/O server LPAR. We list several performance considerations about vSCSI:

- ▶ A RAID card can be used by either (or both) the virtual I/O server and virtual I/O clients (VIOC) disk.
- ▶ For performance reasons, logical volumes within the virtual I/O servers that are exported as vSCSI devices must not be striped or mirrored, span multiple physical drives, or have bad block relocation enabled.
- ▶ SCSI reserves have to be turned off whenever you share disks across two virtual I/O servers.
- ▶ Set vSCSI Queue depth to match the underlying real devices.
- ▶ Do not configure a large number of vSCSI adapters per client; four vSCSI adapters are typically sufficient.
- ▶ If you use the FC Multi-Path I/O (MPIO) on the virtual I/O server, set the following fscsi device values (requires switch attachment and support by the switch):
 - a. `dyntrk=yes` (Dynamic Tracking of FC Devices)
 - b. `fc_err_recov= fast_fail` (FC Fabric Event Error Recovery Policy)
- ▶ If you use the MPIO on the VIOC, set the following hdisk device values:
 - `hcheck_interval=60` (Health Check Interval)

For more information: For detailed information about how to tune `queue_depth` and `qdepth_enable`, refer to the IBM white paper, *AIX disk queue depth tuning for performance*, at this website:

<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD105745>

Shared Ethernet Adapter (SEA) considerations

The following performance considerations about virtual networks are important:

- ▶ You need to know the network workload types: Transmission Control Protocol (TCP) streaming or TCP request and response.
- ▶ If the network workload type is TCP streaming, we suggest that you set maximum transmission unit (MTU) to 9000 (jumbo frames), which saves CPU resources.
- ▶ If the network workload type is TCP request and response, you need to monitor most network package sizes. If the size ≤ 64 bytes, it is a small package. If the size ≥ 1024 bytes, it is a large package. If it is a large package, we suggest setting MTU to 9000 (jumbo frames). If it is a small package, you do not need to change the MTU size; keep the default

size (1500). If most network package sizes are between small and large, a test to decide which MTU size to use must be performed.

- ▶ If there is only SEA running in a virtual I/O server environment (without vSCSI in the same LPAR), we suggest that you disable the SEA adapter's threading option. If there is vSCSI or virtual FC, we suggest that you keep the thread attribute to the default value (1). For example, the following command disables threading for the Shared Ethernet Adapter ent1:

```
mkvdev -sea ent1 -vadapter ent5 -default ent5 -defaultid 1 -attr thread=0
```

- ▶ If you enable the `largesend` attribute on the SEA, the client partition can transmit large data, which gets segmented by the real adapter to fit its MTU and saves the partition's CPU resource. It needs enabling on the physical adapter first before creating the SEA device, and it needs enabling on the client partition. The following commands show how to enable it on these devices:

- a. Enable on the physical adapter:

```
chdev -dev ent0 -attr large_send=1
```

- b. Enable on the SEA device after creation:

```
chdev -dev ent3 -attr largesend=1
```

- c. Enable on the client partition:

```
ifconfig en0 largesend
```

- d. Disable on the client partition:

```
ifconfig en0 -largesend
```

Considerations with Virtual Fibre Channel

With N_Port ID Virtualization (NPIV), you can configure the managed system so that multiple LPARs can access independent physical storage through the same physical FC adapter. Here are two performance considerations about Virtual FC:

- ▶ To increase the performance of the FC adapters, you sometimes need to modify the `max_xfer_size` and `num_cmd_elems` parameters. Each SAN vendor has a recommended setting.
- ▶ These suggestions are general suggestions for the parameters. For a production environment, you need to monitor the I/O activity and make an assessment if the parameters need to change. For detailed guidance, refer to the following IBM white paper at this website:

<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD105745>

Considerations with CPUs

The following considerations relate to processors (CPUs):

- ▶ Configure an uncapped micro-partition with enough virtual processors (VP) to manage peak loads, especially when it has high network activity through the SEA or plenty of I/O activity. You need to monitor first and determine the VP number.
- ▶ Configure it with a higher weight (priority) than its clients if they are also uncapped.
- ▶ IBM provides a simple formula to size the virtual I/O server CPU resource for SEA. It includes many factors, such as cycles per byte (CPB), type of streaming, size of transaction, MTU size, SEA threading option, and so on. For more information about this sizing formula, refer to the following website:

http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/iphb1_vios_planning_sea_procs.htm

Case studies

The following case studies provide additional information about network and I/O considerations.

Case 1

The following technical IBM white paper discusses performance characterization and configuration preferred practices of IBM i in a virtual I/O high-end external storage environment with IBM System Storage DS8000® attached natively through the IBM PowerVM virtual I/O server and through virtual I/O server and the IBM System Storage SAN Volume Controller (SVC). Refer to the following website:

[http://www-03.ibm.com/support/techdocs/atsmastr.nsf/5cb5ed706d254a8186256c71006d2e0a/af88e63ccded8f6e86257563002680e1/\\$FILE/IBM%20i%20Virtual%20I%20SAN%20Storage%20Performance.pdf](http://www-03.ibm.com/support/techdocs/atsmastr.nsf/5cb5ed706d254a8186256c71006d2e0a/af88e63ccded8f6e86257563002680e1/$FILE/IBM%20i%20Virtual%20I%20SAN%20Storage%20Performance.pdf)

Case 2

The following IBM white paper compares two similar configurations running the same SAS benchmark: one configuration uses directly-attached disks and the other configuration uses virtual I/O server-attached disks to handle high I/O activity needed by the serial-attached SCSI (SAS) computational back-end servers. Refer to the following website:

<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101664>

Case 3

The following document explains a typical network configuration scenario using virtual I/O server in a production environment to provide better network performance and redundancy. This document was created based on a recent experience in setting up a consolidated production environment using dual virtual I/O servers with a redundant network configuration. Refer to the following website:

<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101766>

7.5.5 Active Memory Sharing

Active Memory Sharing (AMS) is one feature of PowerVM from POWER6. AMS provides optimized memory utilization similar to the micro-partition's CPU optimization. The PowerVM Hypervisor manages real memory across multiple AMS-enabled partitions, distributing memory to partitions based on their workload demand. Several new concepts are involved in AMS, such as shared memory pool, paging device, shared memory partition, I/O-entitled memory, memory weight, virtualization control point (VCP), shared memory manager, paging virtual I/O server, collaborative memory manager (CMM), and so on.

For more information about AMS concepts and how to set up, monitor, and manage AMS, refer to *PowerVM Virtualization Active Memory Sharing*, REDP-4470, which is located at the following website:

<http://www.redbooks.ibm.com/redpieces/abstracts/redp4470.html?open>

Considerations with AMS

Performance tuning with AMS is complex, because there are many components related to performance. The major components are the shared memory pool, virtual I/O server, paging device, and shared memory partition. The key for tuning is how to reduce the number of paging activities on the paging devices and how to accelerate the access speed to the paging device.

Shared memory pool

There are two performance considerations for the shared memory pool:

- ▶ If the system has unallocated memory, it is better to add the memory to the memory pool (it can be added dynamically).
- ▶ If you need to reduce the size of the shared memory pool, we suggest that you reduce the size when the load on the shared memory partitions is low.

Virtual I/O server

The performance considerations in 7.5.4, “Virtual I/O server” on page 277 are also suitable for AMS.

Support exists to assign up to two paging virtual I/O server partitions to a shared memory pool to provide multi-path access to the paging devices. This redundant paging virtual I/O server configuration improves the availability of the shared memory partitions in the event of a planned or unplanned virtual I/O server outage. When you configure paging devices that are accessed redundantly by two paging virtual I/O server devices, the devices must meet the following requirements:

- ▶ Physical volumes
- ▶ Located on SAN
- ▶ Must be accessible to both paging virtual I/O server partitions

Paging device

Consider these performance factors for paging devices:

- ▶ Spread the I/O load across as many physical disks as possible.
- ▶ Use disk caches to improve performance. Due to the random access nature of paging, write caches provide benefits. Read caches might not have an overall benefit. If you can use solid state disks (SSD), it is better.
- ▶ Use a striped configuration, if possible with a 4 KB stripe size, which is ideal for a paging environment. Because the virtual I/O server cannot provide striped disk access, striping must be provided by a storage subsystem.
- ▶ Disk redundancy is recommended. When using a storage subsystem, a configuration using mirroring or RAID5 is appropriate. The virtual I/O server cannot provide redundancy.

Shared memory partition

The following performance considerations are for the shared memory partition:

- ▶ The AMS partition only supports the 4 KB page size and does not support 64 KB, 16 MB, or 16 GB page size.
- ▶ Logical memory must be sized based on the maximum quantity of memory that the LPAR is expected to use during the peak time.
- ▶ When logical memory size is reduced dynamically, we suggest that it is done during non-peak hours.
- ▶ When logical memory size is increased dynamically, it does not mean that the LPAR has more physical memory pages.
- ▶ Change the memory weight carefully to balance the priority among all the shared memory partitions to receive physical memory.
- ▶ Keep enough physical memory for I/O-entitled memory.
- ▶ The number of virtual processors on a shared memory partition must be calculated so that when a high page fault rate occurs, the number of running virtual processors is able to sustain the workload.

- Keep the parameter (`ams_loan_policy`) as the default value (1) for most production workloads.

Important: Memory weight is merely one of the parameters used by the hypervisor to decide how many physical pages are assigned to the shared memory partitions.

For more information about performance tuning with AMS, refer to *PowerVM Virtualization Active Memory Sharing*, REDP-4470, which is located at the following website:

<http://www.redbooks.ibm.com/redpieces/abstracts/redp4470.html?Open>

Or, refer to the IBM white paper, *IBM PowerVM Active Memory Sharing Performance*, which is located at the following website:

<http://public.dhe.ibm.com/common/ssi/ecm/en/pow03017usen/POW03017USEN.PDF>

Case study

This paper describes the performance of IBM WebSphere® MQ when running in an AMS environment and how WebSphere MQ benefits from AMS technology. Refer to the following website:

http://www-304.ibm.com/partnerworld/wps/servlet/ContentHandler/Whitepaper/power/aix/v6r1_power6/performance

7.5.6 Live Partition Mobility

Live Partition Mobility (LPM) is one of the PowerVM features that provides the ability to move AIX and Linux LPARs from one system to another system. The mobility process transfers the system environment, including the processor state, memory, attached virtual devices, and connected users.

There are two types of LPM: Active Partition Mobility and Inactive Partition Mobility.

For more information about the mechanism, planning, and configuration with AMS, refer to *IBM PowerVM Live Partition Mobility*, SG24-7460, which is located at the following website:

<http://www.redbooks.ibm.com/abstracts/sg247460.html?Open>

Considerations with LPM

LPM, combined with other virtualization technologies, such as micro-partitioning, virtual I/O server, AMS, and so on, provides a fully virtualized computing platform that helps provide the infrastructure flexibility that is required by today's production data centers.

For performance considerations with micro-partitioning, virtual I/O server, and AMS, refer to 7.5.2, "Micro-partitioning" on page 273, 7.5.4, "Virtual I/O server" on page 277, and 7.5.5, "Active Memory Sharing" on page 280.

There are two considerations about LPM performance:

- The network performance between the source and the destination systems, which is used for transferring the system's state, is important for the elapse time during the active partition mobility process. We suggest using a dedicated network for the state transfer, and the bandwidth must be at least 1 Gbps.

For instructions to set up the network to improve partition mobility performance, refer to the following website:

<https://www-304.ibm.com/support/docview.wss?uid=isg3T7000117>

- The CPU performance of the virtual I/O server is important for the application suspend time during the active partition mobility process. We recommend to test and monitor the CPU performance to get an appropriate performance for the production environment.

To describe the importance of the virtual I/O server' processor resource, Table 7-6 shows testing results performed with various configurations. The suspend time depends on the network bandwidth and the virtual I/O server CPU performance. During the time of active LPM, the VIOC keeps a high workload (CPU usage is 100% and has a lot of memory for the operation) and the MTU size of the SEA is 1500. Additional virtual I/O server information is provided:

- Virtual I/O server configuration
The virtual I/O server partition is a micro-partition under capped mode.
- Network throughput
The network throughput during the active LPM on the virtual I/O server partition.
- Elapsed time
The time between the start migration and end migration.
- Suspend time
The time that the application is suspended during the operation. The transaction's connection is not broken, and the client's operation can continue after the suspended state finishes.

Table 7-6 Testing results for an active LPM operation

Virtual I/O server configuration	Virtual I/O client configuration	Network throughput	Elapsed time	Suspend time
0.5C/512 MB	0.2C/8 G	50 MB/s	6m14s	2s
0.5C/512 MB	0.2C/16 G	50 MB/s	8m30s	34s
1C/512 MB	0.2C/16 G	77 MB/s	6m	2s
1.5C/512 MB	0.2C/16 G	100 MB/s	4m46s	2s

In this testing, the processor number of the virtual I/O server affects the network performance and suspended time. Note that our environment is an testing environment. For a production environment, we suggest that you perform a similar test to find the best configuration for the virtual I/O server.

Case studies

The following case studies provide additional information about how to set up the environment to get the best results when performing Live Partition Mobility operations.

Case 1

This paper documents the findings of a performance test using the PowerVM Live Partition Mobility feature to migrate an active SAP system with various workload levels. The workload tests documented in this paper show that the LPM operations succeed even at high system utilization levels. The amount of time that it takes to complete a migration is dependent on a number of factors, such as the memory size of the migrating partition, the amount of memory changes in the partition, and the sustainable network bandwidth between the two VIO servers performing the migration. The suspend/resume phase during a migration affects the application throughput and user response times for a few minutes. The overall effect and the amount of time that is required to reach normal processing levels again increases with the

active load on the system. This document provides suggestions when managing SAP within an LPM environment. See the following website for detailed information:

<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101917>

Case 2

This article provided detailed analysis for DB2 9.5 running with the PowerVM Enterprise Edition Live Partition Mobility feature. The DB2 9.5 instance is hosting an OLTP workload. The DB2 instance is servicing a multitude of clients, a throughput of as many as 2000 transactions per second (TPS) is migrated to another system. The client network connections remained intact, and the application observed a blip for only a few seconds. Although the environment is implemented on an IBM POWER6 system with AIX 5.3 TL7, it can also be deployed on a POWER7 environment. Refer to the publication at the following website:

<http://www.ibm.com/developerworks/aix/library/au-db2andpower/index.html>

Case 3

This paper explains how to set up a complex PowerVM solution environment, including dual virtual I/O servers, AMS, and LPM. Using Active Memory Sharing with advanced PowerVM configurations, including dual virtual I/O servers and Live Partition Mobility, provides benefits from the high availability and flexibility for your virtualized Power Systems environment. Refer to the following website:

<http://www.ibm.com/developerworks/aix/library/au-activemem/index.html?ca=drs->

7.6 Performance considerations with AIX

In this section, we introduce performance considerations and tuning methods with AIX when deploying or using POWER7 servers. This section includes the following topics:

- ▶ Olson and POSIX time zones
- ▶ Large page size
- ▶ One Terabyte (TB) segments aliasing
- ▶ Memory affinity
- ▶ Hardware memory prefetch
- ▶ Simultaneous multithreading
- ▶ New features of XL C/C++ V11.1 to support POWER7
- ▶ How to deal with unbalanced core and memory placement

Also, we introduce web resources about AIX performance tuning on Power servers.

AIX performance tuning: AIX performance tuning deals with many areas. In this section, we mention a few tuning methods about POWER7. For more detailed information about AIX tuning, refer to 7.6.9, “AIX performance tuning web resources” on page 300.

7.6.1 Olson and POSIX time zones

In AIX 5.3 or earlier versions, the default time zone is POSIX. In AIX 6.1, the default time zone is replaced with the Olson time zone.

In AIX 6.1, the Olson time zone is implemented with the AIX International Components for Unicode (ICU) library and provides additional features:

- It is natural for an user to know a city name than to know the POSIX format of a time zone.
- It is easy to maintain the changes of daylight saving time (DST) over history, although historically DST changes are not an important factor for most users.

In AIX 7.1, the default time zone is still Olson, but the implementation changed over the native code and avoids using the ICU library.

In AIX 6.1 and AIX 7.1, the POSIX time zone is still supported. Refer to the following IBM information center website:

http://publib.boulder.ibm.com/infocenter/aix/v7r1/topic/com.ibm.aix.baseadm/doc/baseadmdita/olson_time_zone.htm?resultof=%22%70%6f%73%69%78%22%20%22%74%69%6d%65%22%20%22%7a%6f%6e%65%22%20

How to know the time zone that is used in the current AIX environment

Check the output of the command (`env | grep TZ`). If the output format is similar to “TZ=Asia/Shanghai”, it is an Olson time zone format. If the output format is similar to “TZ=BEIST-8”, it is a POSIX time zone format.

Performance difference between Olson and POSIX time zones

Although time zone is an environmental variable, it provides a value and does not involve other functions. But some local time-related subroutines, for example, *localtime()* and *gettimeofday()*, use time zone values. Using other time zone values might cause these functions' response times to differ.

In AIX 6.1, when enabling the Olson time zone, the implementation of those subroutines relies on the ICU library, and the arithmetic of the ICU library, which has been introduced by the International Components for Unicode, is more complex than the implementation of the POSIX time zone. Normally, the POSIX time zone value has better performance than the Olson value.

The Olson time zone penalty might not be a concern in cases where the application looks up the local time only one time or occasionally, but in cases where the local time-related subroutines are going to be repeatedly called, or if an application is performance sensitive, it is a much better option to continue using the POSIX time zone format in AIX 6.1.

In AIX 7.1, the implementation of the Olson time zone does not rely on the ICU library; it uses native code. So, the performance gains improvements. The performance is similar between the POSIX and Olson time zones in AIX 7.1.

Setting the POSIX time zone in AIX 6.1 or later

Refer to the following steps to set the POSIX time zone in AIX 6.1 or later:

1. Log in with the root user and edit the `/etc/environment`:

```
vi /etc/environment
```
2. Change the TZ environment variable to POSIX time zone format and save the file, for example:

```
TZ=BEIST-8
```
3. Then, reboot the system.

Testing: For Independent Software Vendor (ISV) applications, we suggest that you perform proof of concept (POC) testing to verify the feasibility and stability before changing the time zone in a production environment.

Case study

In order to provide a timeout mechanism with the DB2 client, DB2 provides one environment variable (DB2TCP_CLIENT_RCVTIMEOUT) to enable it. For example, when you set DB2TCP_CLIENT_RCVTIMEOUT=10, the timeout value is 10 seconds. After enabling this function, DB2 adds invoking subroutines for each transaction, it includes localtime() and gettimeofday() and others. If the response time of the transaction is extremely short, for example, 0.00005s, and you enable this environment variable, the performance difference is obvious (sometimes more than 10%) between the Olson and POSIX time zones in the AIX 6.1 environment.

For more information about the DB2 registry variable (DB2TCP_CLIENT_RCVTIMEOUT), refer to the following IBM website and search DB2TCP_CLIENT_RCVTIMEOUT:

<http://publib.boulder.ibm.com/infocenter/db2luw/v9/index.jsp?topic=/com.ibm.db2.udb.admin.doc/doc/r0005660.htm>

7.6.2 Large page size

AIX supports the 16 MB page size, which is also known as large pages. The use of large pages reduces the number of translation lookaside buffer (TLB) misses and therefore improves performance. Applications can use large pages for their text, data, heap, and shared memory regions.

Important: The 16 MB page size is only for high-performance environments, 64 KB pages are considered general purpose, and most workloads will likely see a benefit from using 64 KB pages rather than 4 KB pages. For more information about the 64 KB page size, refer to the IBM information center website (search on multiple page size support):

http://publib.boulder.ibm.com/infocenter/aix/v7r1/index.jsp?topic=/com.ibm.aix.prftungd/doc/prftungd/multiple_page_size_support.htm

Starting from AIX 5.3 TL8 or AIX 6.1 TL1, AIX supports specifying the page size to use for a process's shared memory with the SHMPSIZE environment variable. For detailed information about the SHMPSIZE environment variable, refer to the IBM information center website:

http://publib.boulder.ibm.com/infocenter/aix/v6r1/index.jsp?topic=/com.ibm.aix.prftungd/doc/prftungd/multiple_page_size_app_support.htm

Considerations with large pages

The major benefit of larger page sizes is improved performance for applications that repeatedly access large amounts of memory. Performance improvements from larger page sizes result from reducing the overhead of translating an application page address to a page address that is understood by the computer's memory subsystem.

Large page support is a special-purpose performance improvement feature and is not recommended for general use. Note that not all applications benefit from using large pages. In fact, certain applications, such as applications that perform a large number of *fork()* functions, are prone to performance degradation when using large pages.

Rather than using the *LDR_CNTRL* environment variable, consider marking specific executable files to use large pages, which limits the large page usage to the specific application that benefits from large page usage.

If you are considering using large pages, think about the overall performance effect on your system. Certain applications might benefit from large page usage, but you might see performance degradation in the overall system performance due to the reduction of 4 KB page storage available on the system. If your system has sufficient physical memory so that reducing the number of 4 KB pages does not significantly hinder the performance of the system, you might consider using large pages.

AMS partitions: An AMS partition does not support large page size. It only supports 4 KB page size.

Enabling large pages on AIX

To enable large pages on AIX, follow these steps:

1. Compute the number of 16 MB pages needed from the memory size requirements of the application, including text, data, heap, and shared memory region:

```
vmo -p -o lgpg_regions=<number of largepages> -o lgpg_size=16777216
```

Input calculation: $\text{number_of_large_pages} = \text{INT}[(\text{Share Memory Size} - 1) / 16 \text{ MB}] + 1$

2. Turn on the *v_pinshm* parameter to allow pinning of shared memory segments:

```
vmo -p -o v_pinshm=1
```

Leave *maxpin%* at the default of 80.

3. Change the attribute profile of the Oracle user to allow using large pages:

```
chuser capabilities=CAP_BYPASS_RAC_VMM,CAP_PROPAGATE <non root user>
```

Enabling large pages for DB2

To enable large pages, use the **db2set** command to set the *DB2_LARGE_PAGE_MEM* registry variable to DB:

```
db2set DB2_LARGE_PAGE_MEM=DB
```

For detailed information about how to enable large pages on DB2, refer to this website:

<http://publib.boulder.ibm.com/infocenter/db2luw/v9r5/index.jsp?topic=/com.ibm.db2.luw.admin.dboobj.doc/doc/t0010405.html>

Enabling large pages for Informix (from IDS 11.50)

To enable large pages, change the profile of the Oracle user to allow the use of large pages:

```
export IFX_LARGE_PAGES=1
```

For detailed information about how to enable large pages on Informix®, refer to this website:

http://publib.boulder.ibm.com/infocenter/idshelp/v115/index.jsp?topic=/com.ibm.gsg.doc/ids_rel_188.htm

Enabling large pages for Oracle

To enable large page usage, refer to the following steps:

1. Modify the XCOFF executable file header of the Oracle binary file:

```
ldedit -b lpdata <oracle binary>
```

2. Change the Oracle initialization parameter, so that Oracle requests large pages when allocating shared memory:

```
LOCK_SGA=true
```

For detailed information about how to enable large pages on Oracle, refer to this website:

<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP100883>

When these steps are complete, we can start the database to start using large page memory for the database shared memory region.

How to monitor large page utilization

To see how many large pages are in use, use the **vmstat** command with the flags that are shown in Table 7-7. There are three parameters to show large pages.

Table 7-7 Description of vmstat command option for monitoring large page utilization

vmstat option	Description
-l	Displays an additional "large-page" section with the active large pages and free large pages columns
-P	Displays only the VMM statistics, which are relevant for the specified page size
-p	Appends the VMM statistics for the specified page size to the regular vmstat output

Example 7-2 introduces how to use the command, and it shows that there are 32 active large pages (alp) and 16 free large pages (flp), for a total of 48 large pages. There are 32 large pages that can be used for the client's application (including databases).

Example 7-2 Monitoring large pages using vmstat

```
#vmstat -l 2
```

System configuration: lcpu=4 mem=8192MB

kthr		memory		page						faults				cpu				large-page	
r	b	avm	fre	re	pi	po	fr	sr	cy	in	sy	cs	us	sy	id	wa	alp	flp	
0	0	1359405	169493	0	0	0	0	0	0	0	31	396	382	0	0	99	0	32	16
0	0	1359405	169493	0	0	0	0	0	0	0	22	125	348	0	0	99	1	32	16
0	0	1359405	169493	0	0	0	0	0	0	0	22	189	359	0	0	99	1	32	16

```
#vmstat -P ALL
```

System configuration: mem=8192MB

pgsz	memory		page							
	siz	avm	fre	re	pi	po	fr	sr	cy	
4K	882928	276631	103579	0	0	0	56	110	0	
64K	63601	59481	4120	0	0	0	0	0	0	
16M	48	32	16	0	0	0	0	0	0	

```
#vmstat -p ALL
```

System configuration: lcpu=4 mem=8192MB

kthr		memory		page						faults				cpu			
r	b	avm	fre	re	pi	po	fr	sr	cy	in	sy	cs	us	sy	id	wa	
0	0	1359405	169493	0	0	0	0	0	0	0	31	396	382	0	0	99	

```
1 2 1359403 169495 0 0 0 56 110 0 90 4244 1612 3 1 96 0
```

```
psz   avm   fre re  pi  po  fr  sr  cy   siz
4K 276636 103574 0 0 0 56 110 0 882928
64K 59481 4120 0 0 0 0 0 0 63601
16M 32 16 0 0 0 0 0 0 48
```

You also can use the **svmon** command to display information about the memory state of the operating system or specified process.

See the following documentation for additional information about large pages:

http://publib.boulder.ibm.com/infocenter/aix/v7r1/index.jsp?topic=/com.ibm.aix.prftungd/doc/prftungd/large_page_ovw.htm

7.6.3 One TB segment aliasing

AIX supports application segments that are one terabyte (TB) in size starting from AIX 6.1 TL6 or AIX 7.1. Prior AIX versions supported only 256 MB segments. With the TB segment support, we nearly eliminate a huge amount of Segment Lookaside Buffer (SLB) misses, and therefore SLB reload overhead in the kernel¹⁰. Applications using a large amount of shared memory segments (for example, if one process needs 270 GB, leading to over 1300 256 MB segments) incur an increased number of SLB hardware faults, because the data referenced is scattered across all of these segments. This problem is alleviated with TB segment support.

There are two types of 1 TB segment aliasing:

► Shared aliases:

- A single shared memory region large enough on its own to trigger aliasing, by default, at least 3 GB in size.
- TB aliases used by the entire system are “shared” by processes.
- Aliasing triggered at `shmat()` time (shared memory attach).
- AIX does not place other attachments into the terabyte region, unless address space pressure is present.

► Unshared aliases:

- Multiple small, homogeneous shared memory regions grouped into a single 1 TB region.
- Collectively large enough to trigger aliasing. By default, they must exceed 64 GB in size.
- TB aliases are private to the process.
- Aliasing is triggered at the `shmat` of the region that crosses the threshold.
- Unshared aliasing is expensive. And `shmat` requires that the unshared alias is invalidated and removed to avoid access via page table entries (PTEs) to stale regions.

See the following documentation for additional information about 1 TB segment aliasing:

http://publib.boulder.ibm.com/infocenter/aix/v7r1/index.jsp?topic=/com.ibm.aix.prftungd/doc/prftungd/1TB_segment_aliasing.htm

¹⁰ POWER7 core provides 12 SLB for user processes, and it will yield 3 GB of accessible memory without SLB faults. But if the shared memory size exceeds 3 GB, there is a potential risk for a performance decrease.

Considerations with 1 Terabyte segments

This section provides considerations when using the 1 TB segment.

mmap considerations

For the mmap, there are two considerations:

- ▶ Aliasing is not supported.
- ▶ A new address space layout is in use (if 1 TB segment aliasing is active).

Unshared aliases can have performance issues

Here are a few details about performance when using unshared aliases:

- ▶ Every address space removal can cause an unshared alias removal.
- ▶ A 1 TB unshared alias is not reusable until a background kernel process clears all its possible PTEs up to a high-water mark that has been established at run time.

Determine how to configure 1 TB segment aliasing in a production environment after sufficient testing with the client application.

Implementing 1 TB segment aliasing

There are four key AIX **vmo** parameters to control the 1 TB segment aliasing:

- ▶ **esid_allocator**, **VMM_CNTRL=ESID_ALLOCATOR=[0,1]**
The default is off (0) in AIX 6.1 TL6, and default is on (1) in AIX 7.1. When on, it indicates that 1 TB segment aliasing is in use, including aliasing capabilities and address space layout changes. This parameter can be changed dynamically.
- ▶ **shm_1tb_shared**, **VMM_CNTRL=SHM_1TB_SHARED=[0,4096]**
The default is set to 12 on POWER7 and set to 44 on POWER6 and earlier. It controls the threshold, which is the “trigger value” at which a shared memory attachment gets a shared alias (3 GB, 11 GB). The unit is in 256 MB segments.
- ▶ **shm_1tb_unshared**, **VMM_CNTRL=SHM_1TB_UNSHARED=[0,4096]**
The default is set to 256 MB (up to 64 GB). It controls the threshold at which multiple homogeneous small shared memory regions are promoted to an unshared alias. Conservatively, it is set high, because unshared aliases are compute-intensive to initially establish and to destroy. The unit is in 256 MB segments.
- ▶ **shm_1tb_unsh_enable**
The default is set to on; it determines whether unshared aliases are used. Unshared aliases can be “expensive”.

Important: The four **vmo** parameters are restricted tunables and must not be changed unless recommended by IBM support. Any 32-bit applications are not affected by these parameters.

How to verify 1 TB segment aliasing (LSA) usage

To determine if LSA is active for a specific process, you need to utilize the AIX kernel debugger **kdb**, which is run as user root. To quit **kdb**, enter quit.

Using kdb: Use caution when using **kdb** and follow the described procedure carefully. You might terminate (kill) AIX from within **kdb** if you use the wrong commands.

The following IBM white paper describes how to verify 1 TB segment aliasing (LSA) usage for the Oracle process with other testing scenarios. It also provides experiences with tuning Oracle SGA with 1 TB segment aliasing.

The method to verify other applications, for example DB2, Informix, and so on, is similar to the method that is mentioned in this document:

[http://www-03.ibm.com/support/techdocs/atsmastr.nsf/5cb5ed706d254a8186256c71006d2e0a/a121c7a8a780d41a8625787200654e3a/\\$FILE/Oracle_DB_and_Large_Segment_Aliasing_v1.0.pdf](http://www-03.ibm.com/support/techdocs/atsmastr.nsf/5cb5ed706d254a8186256c71006d2e0a/a121c7a8a780d41a8625787200654e3a/$FILE/Oracle_DB_and_Large_Segment_Aliasing_v1.0.pdf)

7.6.4 Memory affinity

AIX added enhanced affinity (*memory affinity*) support for POWER7 systems, because the POWER7 architecture is extremely sensitive to application memory affinity due to 4x the number of cores on a chip compared to the POWER6 chip. There is a **vmo** tunable `enhanced_affinity_private` that can be tuned from 0 to 100 to improve application memory locality.

The **vmo** tunable `enhanced_affinity_private` is set to 40 by default on AIX 6.1 TL6 or AIX 7.1 and beyond. It is set to 20 on AIX 6.1 TL5. A value of 40 indicates the percentage of application data that is allocated locally (or “affinitized”) on the memory behind its home POWER7 socket. The rest of the memory is striped across behind all the sockets in its partition.

In order to decide the need to implement this tuning parameter, an analysis of the AIX kernel performance statistics and an analysis of the AIX kernel trace are required to determine if tuning changes to the enhanced affinity tunable parameters can help to improve application performance.

Considerations with memory affinity

Increasing the degree of memory locality can negatively affect application performance on LPARs configured with low memory or with workloads that consist of multiple applications with various memory locality requirements.

Determine how to configure `enhanced_affinity_private` or other memory affinity parameters in a production environment after you have performed sufficient testing with the application.

Important: The `enhanced_affinity_private` and `enhanced_affinity_vmpool_limit` **vmo** tunables are restricted tunables and must not be changed unless recommended by IBM support.

How to monitor memory affinity

Certain AIX commands, `lssrad`, `mpstat`, and `svmon`, are enhanced to retrieve POWER7 memory affinity statistics. For detailed information about how to use these commands, refer to 7.8.8, “Monitoring memory affinity statistics” on page 323.

Description of key memory affinity parameters

This section provides information about the key memory affinity parameters.

enhanced_affinity_vmpool_limit

Example 7-3 on page 292 shows a description of the `enhanced_affinity_vmpool_limit` with the **vmo** command.

Example 7-3 Description of enhanced_affinity_vmpool_limit vmo command parameter

```
# vmo -h enhanced_affinity_vmpool_limit
```

Help for tunable enhanced_affinity_vmpool_limit:

Purpose:

Specifies percentage difference of affinitized memory allocations in a vmpool relative to the average across all vmpools.

Values:

Default: 10

Range: -1 - 100

Type: Dynamic

Unit: numeric

Tuning:

Affinitized memory allocations in a vmpool are converted to balanced allocations if the affinitized percentage difference between the vmpool and the average across all vmpools is greater than this limit.

enhanced_affinity_private

Example 7-4 shows a description of the enhanced_affinity_private parameter with the **vmo** command.

Example 7-4 Description of enhanced_affinity_private vmo command parameter

```
# vmo -h enhanced_affinity_private
```

Help for tunable enhanced_affinity_private:

Purpose:

Specifies percentage of process private memory allocations that are affinitized by default.

Values:

Default: 40

Range: 0 - 100

Type: Dynamic

Unit: numeric

Tuning:

This tunable limits the default amount of affinitized memory allocated for process private memory. Affinitizing private memory may improve process performance. However, too much affinitized memory in a vmpool can cause paging and impact overall system performance.

7.6.5 Hardware memory prefetch

Hardware memory prefetch helps to improve the performance of applications that reference memory sequentially by prefetching memory.

Hardware memory prefetch considerations

There might be adverse performance effects, depending on the workload characteristics. Determine whether to disable hardware memory prefetch in a production environment after sufficient testing with the application.

Important: The default value of the hardware memory prefetch must not be changed unless recommended by IBM AIX support.

How to monitor and set prefetch settings

The **dscrctl** command is used for the system administrator to read the current settings for the hardware streams mechanism and to set a system-wide value for the Data Stream Control Register (DSCR). Table 7-8 shows the options and descriptions of the command.

Table 7-8 Descriptions of the *dscrctl* command options

dscrctl option	Description
-q	Query: This option displays the number of hardware streams supported by the platform and the values of the firmware and operating system default prefetch depth.
-c	Cancel: This option cancels a permanent setting of the system default prefetch depth at boot time by removing the dscrctl command from the <code>/etc/inittab</code> file.
-n	Now: When used in conjunction with the <code>-s</code> flag, this option changes the runtime value of the operating system default prefetch depth. The change is not persistent across boots.
-b	Boot: When used in conjunction with the <code>-s</code> flag, this option makes the change persistent across boots by adding a dscrctl command to the <code>/etc/inittab</code> file.
-s dscr_value	Set: This option defines the value for the new operating system default prefetch depth. The value is treated as a decimal number unless it starts with <code>0x</code> , in which case it is treated as hexadecimal. The default is <code>0x0</code> . The value <code>0x1</code> means disable it.

For detailed information about the **dscrctl** command, refer to the IBM information center website and search for it:

<http://publib.boulder.ibm.com/infocenter/aix/v7r1/index.jsp?topic=/com.ibm.aix.ntl/RELNOTES/GI11-9815-00.htm>

Example 7-5 shows how to query the characteristics of the hardware streams on the system.

Example 7-5 To query the characteristics of the hardware streams on the system

```
# dscrctl -q
```

Current DSCR settings:

```
Data Streams Version = V2.06
number_of_streams    = 16
platform_default_pd  = 0x5 (DPFD_DEEP)
os_default_pd        = 0x0 (DPFD_DEFAULT)
```

Example 7-6 shows how to disable the hardware default prefetch depth, effective immediately. But after a reboot, it is restored to the default value (0).

Example 7-6 Disable hardware default prefetch depth

```
# dscrctl -n -s 1
```

```
# dscrctl -q
```

Current DSCR settings:

```
Data Streams Version = V2.06
number_of_streams    = 16
platform_default_pd  = 0x5 (DPFD_DEEP)
```

```
os_default_pd          = 0x1 (DPFD_NONE)
```

Example 7-7 shows how to disable across a reboot using the **-b** flag. This command creates an entry in the `/etc/inittab` file.

Example 7-7 Disable hardware default prefetch depth with the -b flag

```
# dscrctl -n -b -s 1
```

The value of the DSCR OS default will be modified on subsequent boots

```
# dscrctl -q
```

Current DSCR settings:

```
Data Streams Version = V2.06
number_of_streams    = 16
platform_default_pd  = 0x5 (DPFD_DEEP)
os_default_pd        = 0x1 (DPFD_NONE)
```

```
# cat /etc/inittab|grep dscrset
dscrset:2:once:/usr/sbin/dscrctl -n -s 1 >/dev/null 2>/dev/console
```

For most JAVA applications, we suggest that you turn off the hardware memory prefetch with the AIX command **dscrctl -n -s 1**. For detailed information, refer to page 19 of the IBM white paper, *Java performance on POWER7 - Best Practice*, at the following website:

<http://public.dhe.ibm.com/common/ssi/ecm/en/pow03066usen/POW03066USEN.PDF>

7.6.6 Simultaneous multithreading (SMT)

Simultaneous multithreading (SMT) is a hardware multithreading technology, which enables the execution of multiple instructions from multiple code paths, or hardware threads, in a single CPU clock cycle. POWER5 was the first IBM Power series processor to implement this technology with the capability of using either one or two hardware threads per processor core. With the IBM POWER7 generation and AIX 6.1, an additional two hardware threads are available.

Although SMT is implemented in physical hardware, its use is enabled at the operating system layer, requiring operating system software awareness of this feature. AIX 5.3 can recognize up to two threads per core; AIX Version 6.1 or higher utilizes all four threads on POWER7. In addition to AIX, four SMT threads can be used with IBM i 6.1.1, SUSE SLES 11 SP1, and Red Hat RHEL 6.

POWER7 offers three types of SMT: 1-way, 2-way, and 4-way. With 4-way SMT, you can increase the number of instruction streams that your system can run on the same processor core.

In 2-way SMT (SMT2), the number of the logical processors that the operating system sees is double the number of the physical processor cores in the system. That number of logical processors becomes quadrupled with the 4-way SMT (SMT4). Consequently, the overall system capacity increases as the number of instruction streams increases.

In order to detect the need to implement this tuning, an analysis of AIX kernel performance statistics and an analysis of the AIX kernel trace are required to determine if tuning changes to SMT can help to improve application performance.

Considerations with SMT

Simultaneous multithreading is primarily beneficial in commercial environments where the speed of an individual transaction is not as important as the total number of transactions that are performed. Simultaneous multithreading is expected to increase the throughput of workloads with large or frequently changing working sets, such as database servers and web servers.

With SMT (2/4) on, POWER7 can deliver more total capacity as more tasks are accomplished in parallel. The higher CPU utilization the application gets, the higher relative improvement you get of transaction rates, response times, and even CPU-intensive calculation batch jobs.

Figure 7-18¹¹ shows one compute-intensive workload with POWER7's SMT2 and SMT4 mode, which provides better throughput than Single Thread (ST) mode with the same CPU resources.

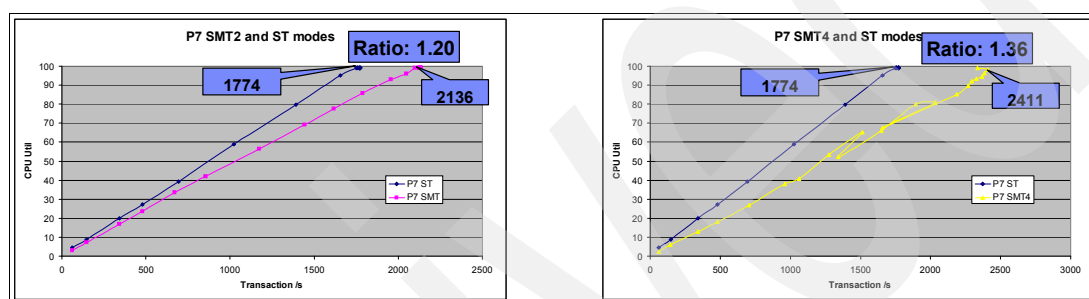


Figure 7-18 ST, SMT, and SMT4 efficiency with compute-intensive workloads

Figure 7-19¹² shows one Java workload with POWER7's SMT2 and SMT4 mode, which also provides better performance than ST with the same CPU resources.

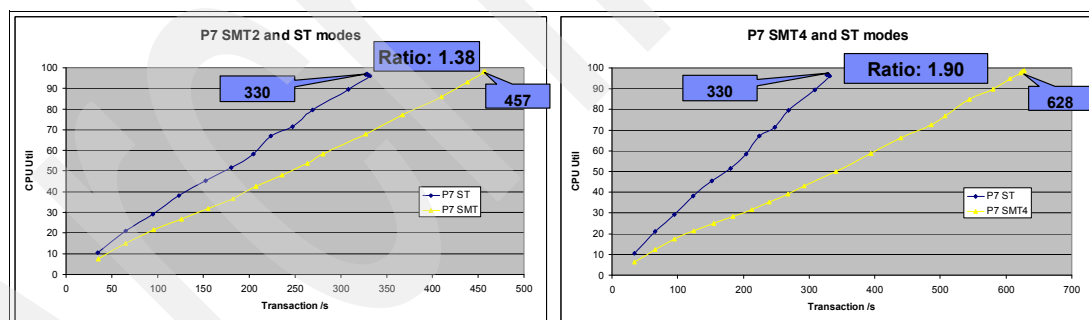


Figure 7-19 ST, SMT, and SMT4 efficiency with Java workloads

Applications that are single process or single threaded might benefit from running in ST mode. ST mode can be beneficial in the case of a multi-process application where the number of application processes is smaller than the number of cores assigned to the partition. Certain workloads do not benefit much from SMT, mostly they have the majority of individual software threads using a large amount of resources in the processor or memory. For example, workloads that are floating-point intensive are likely to gain little from SMT, and they are the ones most likely to lose performance. These workloads heavily use either the floating-point units or the memory bandwidth.

Determine how to configure SMT in a production environment after sufficient testing with the client application.

¹¹ The testing data is provided by the IBM STG Performance team.

¹² The testing data is provided by the IBM STG Performance team.

APARs: We suggest that you apply AIX APAR IZ97088, which enhances the SMT4 performance, and AIX APAR IZ96303, which resolves AIX crash issues when the processor number exceed 32 and you want to switch SMT mode from 2 to 4.

The **smtctl** command controls the enabling (0, 2, or 4) and disabling of processor SMT mode. The **smtctl** command uses this syntax:

```
smtctl [ -m off | on [ -w boot | now ]]
smtctl [-t #SMT [-w boot | now ]]
```

Table 7-9 shows the description of each option.

Table 7-9 Descriptions of the smtctl command options

Option	Description
-m off	This option sets simultaneous multithreading mode to disabled. This option cannot be used with the -t flag.
-m on	This option sets simultaneous multithreading mode to enabled. By using the -m flag, the maximum number of threads supported per processor is enabled. This option cannot be used with the -t flag.
-t #SMT	This option sets the number of the simultaneous threads per processor. The value can be set to one to disable simultaneous multithreading. The value can be set to two for systems that support 2-way simultaneous multi-threading. The value can be set to four for the systems that support 4-way simultaneous multithreading. This option cannot be used with the -m flag.
-w boot	This option makes the simultaneous multithreading mode change effective on the next and subsequent reboots if you run the bosboot command before the next system reboot.
-w now	This option makes the simultaneous multithreading mode change immediately but it does not persist across reboot. If the -w boot option is not specified or if the -w now option is not specified, the mode change is made immediately. It persists across subsequent reboots, if you run the bosboot command before the next system reboot.

For more information, refer to the **smtctl** command man manual (**man smtctl**).

Case study

This paper illustrates the use of SMT for a client's Oracle Data Warehouse workload running on AIX. The parallel degree that is used to execute a particular query can have a significant effect on the run time for that query. Oracle's **CPU_COUNT** parameter, determined by the number of virtual processors and the SMT value, is the key determining factor for Oracle's default parallel degree. For this client's workload, changing from SMT2 to SMT4 appears to increase the total run time of their jobs by over 10%.

<http://w3-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101824>

7.6.7 New features of XL C/C++ V11.1

IBM XL C for AIX is a standards-based, high-performance C compiler with advanced optimizing and debugging features. IBM XL C for AIX 11.1 introduces enhancements for exploitation of the latest POWER7 architecture:

- ▶ Vector unit and vector scalar extension (VSX) instruction support
- ▶ Mathematical Acceleration Subsystem (MASS) libraries support
- ▶ Power instruction direct control support
- ▶ New arch and tune compiler options

Vector unit and vector scalar extension (VSX) instruction support

IBM XL C for AIX V11.1 supports the VSX instruction set in the POWER7 processor. New data types and built-in functions are introduced to support the VSX instruction allowing you to efficiently manipulate vector operations in your applications. The advanced compiler optimizer can also automatically take advantage of these vector facilities to help automatically parallelize your applications.

Mathematical Acceleration Subsystem (MASS) libraries support

The highly tuned MASS libraries are enhanced to support the POWER7 processor:

- ▶ The vector functions within the vector MASS library are tuned for the POWER7 architecture. The functions can be used in either 32-bit or 64-bit mode.
- ▶ New functions, such as `exp2`, `exp2m1`, `log21p`, and `log2`, are added in both single-precision and double-precision functional groups. In addition, functions that support the previous Power processors are enhanced to support POWER7 processors.
- ▶ A new MASS single-instruction, multiple-data (SIMD) library that is tuned for the POWER7 processor is provided. It contains an accelerated set of frequently used mathematical functions.

Power instruction direct control support

New built-in functions unlock POWER7 processor instructions to enable you to take direct control at the application level:

- ▶ POWER7 prefetch extensions and cache control instructions
- ▶ POWER7 hardware instructions

New arch and tune compiler options

New arch and tune compiler options are added to specify code generation for the POWER7 processor architecture:

- ▶ `-qarch=pwr7` instructs the compiler to produce code that can fully exploit the POWER7 hardware architecture.
- ▶ `-qtune=pwr7` enables optimizations that are specifically tuned for the POWER7 hardware platforms.

For more information about IBM XL C/C++ V11.1 for AIX, refer to the IBM website:

http://www-947.ibm.com/support/entry/portal/Documentation/Software/Rational/XL_C_C++_for_AIX

7.6.8 How to deal with unbalanced core and memory placement

Unbalanced core and memory placement happens when you are frequently shutting down and activating partitions using various amounts of resources (cores and memory) or using the DLPAR operation to change the resource (cores and memory) frequently. This configuration results in a performance decrease that is greater than you might originally expect.

How to detect unbalanced core and memory placement

To display core and memory placement on an LPAR, use the `lssrad` command. The REF1 column in the output of Example 7-8 is the first hardware-provided reference point that identifies sets of resources that are near each other. SRAD is the Scheduler Resource Allocation Domain. Cores and memory must be allocated from the same REF1 and SRAD.

The `lssrad` output that is shown in Example 7-8 shows an undesirable placement of an LPAR with 32 cores and 256 GB of memory. Pinned memory, such as large pages, is not reflected in the output.

Example 7-8 The output of the lssrad command

```
# lssrad -va
```

REF1	SRAD	MEM	CPU
0	0	64287.00	0-31
1	1	8945.56	32-47
	2	0.00	48-79
2	3	0.00	80-99
	4	1893.00	100-127
3	5	74339.00	

How to handle unbalanced core and memory placement

In many cases, unbalanced core and memory placement does not affect performance. It is unnecessary to fix this situation.

The performance of certain applications, especially memory-sensitive applications, might be affected by unbalanced core and memory placement, for example, > 10%. If CPU and memory resources are insufficient in the server, and you have to optimize the core and memory placement, refer to the following steps to release the resources manually.

Deactivation: Deactivating an existing LPAR does not free its resources, so you cannot optimize the core and memory placement merely by deactivating and rebooting LPARs.

Follow these steps:

1. Shut down all the partitions in the server.
2. Log in the HMC as the hscroot user and execute the following commands for every LPAR:

```
chhwres -r mem -m <machine_name> -o r -q <size_in MBytes> --id <partition_id>
```

```
chhwres -r proc -m <machine_name> -o r --procs <number_of_cores> --id  
<partition_id>
```

Commands: The first command frees all the memory, and the second command frees all the cores.

3. Activate the LPARs in the order of performance priority.

For example, we have one Power 780 server. There are six LPARs in the server: one LPAR for the DB partition, two LPARs for the application partitions, and three LPARs for the web partitions. Figure 7-20 shows the HMC view.

Select	Name	ID	Status	Pr... Uni...	Memory ...	Active Profile	Environment
<input type="checkbox"/>	DB_lpar1	1	Running	8	32	lpar1	AIX or Linux
<input type="checkbox"/>	APP_lpar2	2	Running	4	24	lpar2	AIX or Linux
<input type="checkbox"/>	WEB_lpar7	3	Running	16	24	lpar7	AIX or Linux
<input type="checkbox"/>	APP_lpar4	4	Running	4	24	lpar4	AIX or Linux
<input type="checkbox"/>	WEB_lpar5	5	Running	16	24	lpar5	AIX or Linux
<input type="checkbox"/>	WEB_lpar6	6	Running	16	24	lpar6	AIX or Linux

Max Page Size: 50 Total: 6 Filtered: 6 Selected: 0

Figure 7-20 LPAR configuration of the Power 780 server

After running dynamic LPAR operations with the DB partition, the core and memory placement is not as good, as shown in Example 7-9.

Example 7-9 Checking the CPU and memory placement

```
# lssrad -av
```

REF1	SRAD	MEM	CPU
0			
	0	31714.00	0-7
1			
	1	0.00	8-19
2			
	2	0.00	20-27
	3	0.00	28-31

Then, shut down all the LPARs through the AIX command line or the HMC GUI. Then, we log in to the HMC command line as the hscroot user and execute the following commands, as shown in Example 7-10.

Example 7-10 Release LPAR resources with the chhwres command manually

```
hscroot@HMC50:~> chhwres -r mem -m SVRP7780-04-SN0661F4P -o r --id 1 -q 32768
hscroot@HMC50:~> chhwres -r mem -m SVRP7780-04-SN0661F4P -o r --id 2 -q 24576
hscroot@HMC50:~> chhwres -r mem -m SVRP7780-04-SN0661F4P -o r --id 3 -q 24576
hscroot@HMC50:~> chhwres -r mem -m SVRP7780-04-SN0661F4P -o r --id 4 -q 24576
hscroot@HMC50:~> chhwres -r mem -m SVRP7780-04-SN0661F4P -o r --id 5 -q 24576
hscroot@HMC50:~> chhwres -r mem -m SVRP7780-04-SN0661F4P -o r --id 6 -q 24576
```

After the commands execute successfully, we can see that the values of the Processing Units column and Memory column in the HMC view, as shown in Figure 7-21, have changed to 0. This value means that the resource has been released.

Systems Management > Servers > SVRP7780-04-SN0661F4P									
<div> </div> <div>Filter</div> <div>Tasks ▾ Views ▾</div>									
Select	Name	ID	Status	Processing Units	Memory (G...	Active Profile	Environment	Reference Code	
<input type="checkbox"/>	DB_lpar1	1	Not Activated	0	0	lpar1	AIX or Linux	00000000	
<input type="checkbox"/>	APP_lpar2	2	Not Activated	0	0	lpar2	AIX or Linux	00000000	
<input type="checkbox"/>	WEB_lpar7	3	Not Activated	0	0	lpar7	AIX or Linux	00000000	
<input type="checkbox"/>	APP_lpar4	4	Not Activated	0	0	lpar4	AIX or Linux	00000000	
<input type="checkbox"/>	WEB_lpar5	5	Not Activated	0	0	lpar5	AIX or Linux	00000000	
<input type="checkbox"/>	WEB_lpar6	6	Not Activated	0	0	lpar6	AIX or Linux	00000000	
<div>Max Page Size: 50</div> <div>Total: 6 Filtered: 6 Selected: 0</div>									

Figure 7-21 The HMC view after releasing the resource manually

Then, we activate the LPARs in order of their performance priority level, as shown in Example 7-11.

Example 7-11 HMC commands to active LPARs

```
hscroot@HMC50:~> chsysstate -r lpar -m SVRP7780-04-SN0661F4P -o on --id 1 -f lpar1
hscroot@HMC50:~> chsysstate -r lpar -m SVRP7780-04-SN0661F4P -o on --id 2 -f lpar2
hscroot@HMC50:~> chsysstate -r lpar -m SVRP7780-04-SN0661F4P -o on --id 3 -f lpar7
hscroot@HMC50:~> chsysstate -r lpar -m SVRP7780-04-SN0661F4P -o on --id 4 -f lpar4
hscroot@HMC50:~> chsysstate -r lpar -m SVRP7780-04-SN0661F4P -o on --id 5 -f lpar5
hscroot@HMC50:~> chsysstate -r lpar -m SVRP7780-04-SN0661F4P -o on --id 6 -f lpar6
```

After all the LPARs are started, we check the CPU and memory placement again with the **lssrad** command, as shown in Example 7-12.

Example 7-12 Checking core and memory placement

```
# lssrad -av
REF1  SRAD      MEM      CPU
0
      0  31714.00    0-31
```

Now, the core and memory placement has changed to optimized. If this method cannot get optimal placement, reboot the CEC to fix it.

Unbalanced core and memory placement considerations

To avoid this bad situation, at the time of writing this book, we decided that it is better to plan the partitions' configurations carefully before activating them and performing the DLPAR operation. IBM intends to fix this situation in a future release of the system firmware.

7.6.9 AIX performance tuning web resources

This section provides web resources for AIX performance tuning:

- ▶ AIX 7.1 Information Center performance management and tuning

This IBM information center topic contains links to information about managing and tuning the performance of your AIX system: Performance management, Performance Tools Guide, and Performance Toolbox Version 2 and 3 Guide and Reference. The first link (Performance management) provides application programmers, service support representatives (SSRs), system engineers, system administrators, experienced users,

and system programmers with complete information about how to perform tasks, such as assessing and tuning the performance of processors, file systems, memory, disk I/O, Network File System (NFS), Java, and communications I/O:

<http://publib.boulder.ibm.com/infocenter/aix/v7r1/index.jsp?topic=/com.ibm.aix.doc/doc/base/performance.htm>

- AIX on Power performance FAQ

This IBM white paper is intended to address the most frequently asked questions concerning AIX performance on Power Systems and to provide guidelines for the most commonly seen performance issues:

ftp://ftp.software.ibm.com/common/ssi/rep_wh/n/POW03049USEN/POW03049USEN.PDF

- Database performance tuning on AIX

This IBM Redbooks publication provides information about the database's life cycle, in the planning and sizing stage, during implementation, and when running a productive database system. It also describes many tuning experiences for databases on AIX. The databases include DB2, Informix, and Oracle:

<http://www.redbooks.ibm.com/abstracts/sg245511.html?Open>

7.7 IBM i performance considerations

In this section, we provide performance considerations for IBM i.

7.7.1 Overview

IBM i has excellent scalability features and uses the POWER7 architecture without any extra and special tuning. Due to the multithreading nature of IBM i, all applications automatically take advantage of the underlying operating system, microcode, and hardware. For example, IBM i exploits the four SMT threads per processor, as well as the many processors available in POWER7.

The new Level 3 cache design (the cache is on the processor chip and shared among all eight cores on the chip (refer to Figure 7-22 on page 302)), as well as the capability to “lateral cast out” instructions and data elements to the other seven remaining processors’ caches on the chip, helps to free up level 2 and 3 caches. This design helps to dramatically improve the performance and throughput of highly interactive applications, whether they are “green screen” or web-based applications. Sharing the commonly used data in the high-speed Level 3 cache among eight processors reduces the need drastically to fetch the data from real memory. In addition, because the memory bus is much wider and faster on POWER7 than before, applications, which consume a lot of memory and process a large amount of database information, gain performance when migrated to POWER7.

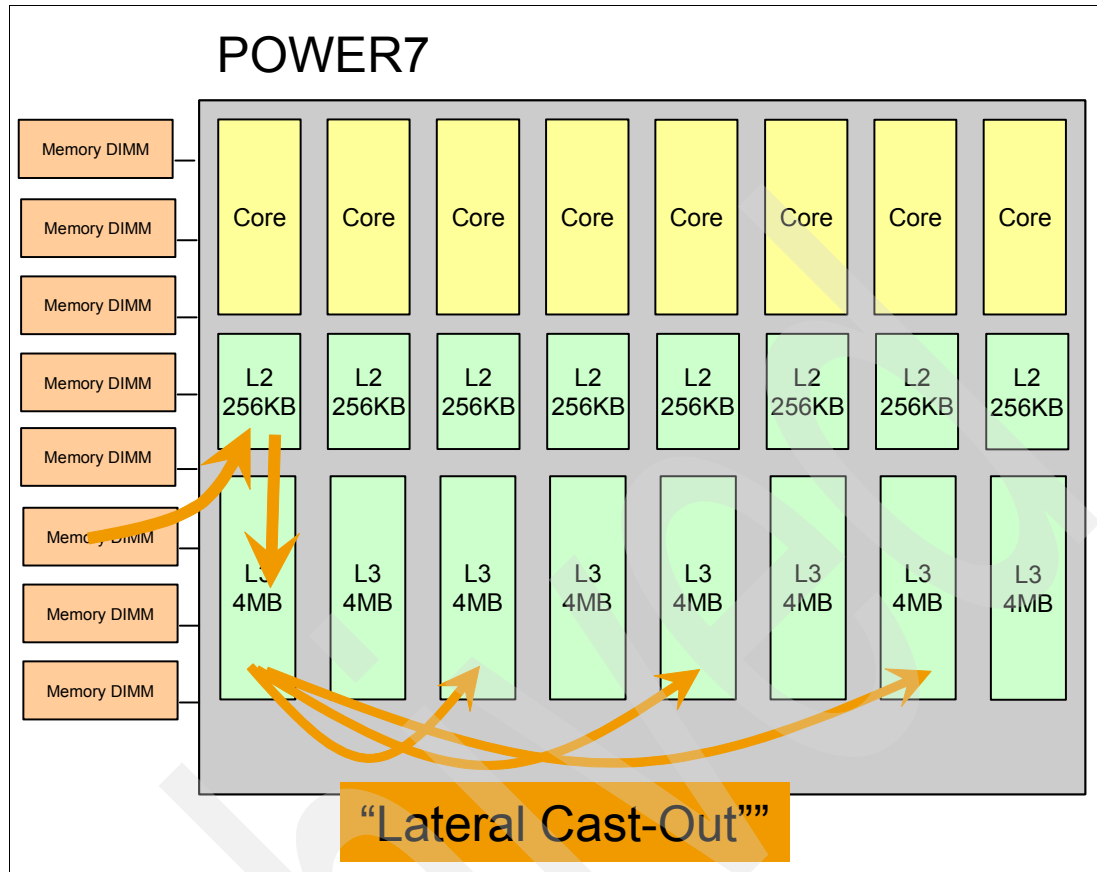


Figure 7-22 The new Level 3 cache design of POWER7

Processor chips are tightly interconnected with high-speed fabrics to ensure greater performance when application workloads span multiple processor chips at the same time. Figure 7-23 shows a conceptual picture of the processor interconnection in a Power 750.

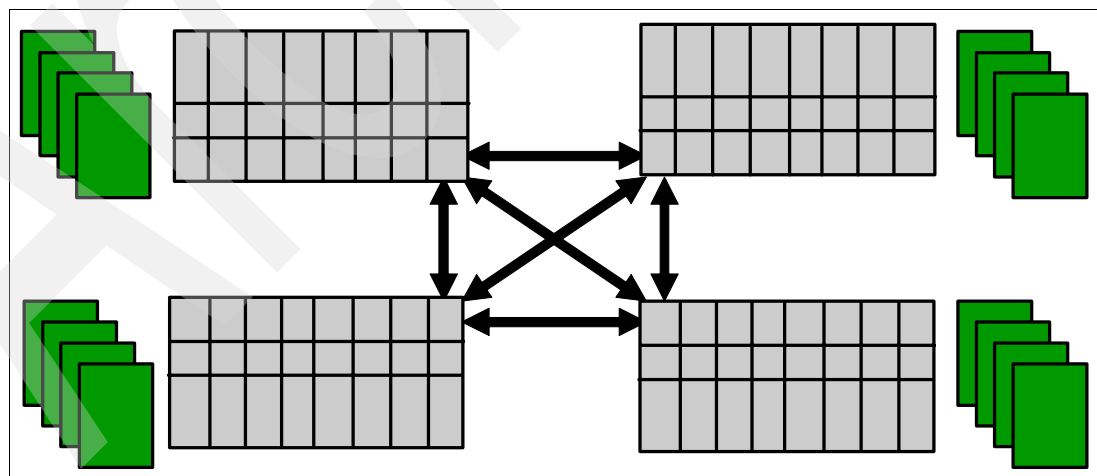


Figure 7-23 The processor interconnection in POWER7 (750)

The Power 780 has the following conceptual interconnection fabric design, as shown in Figure 7-24.

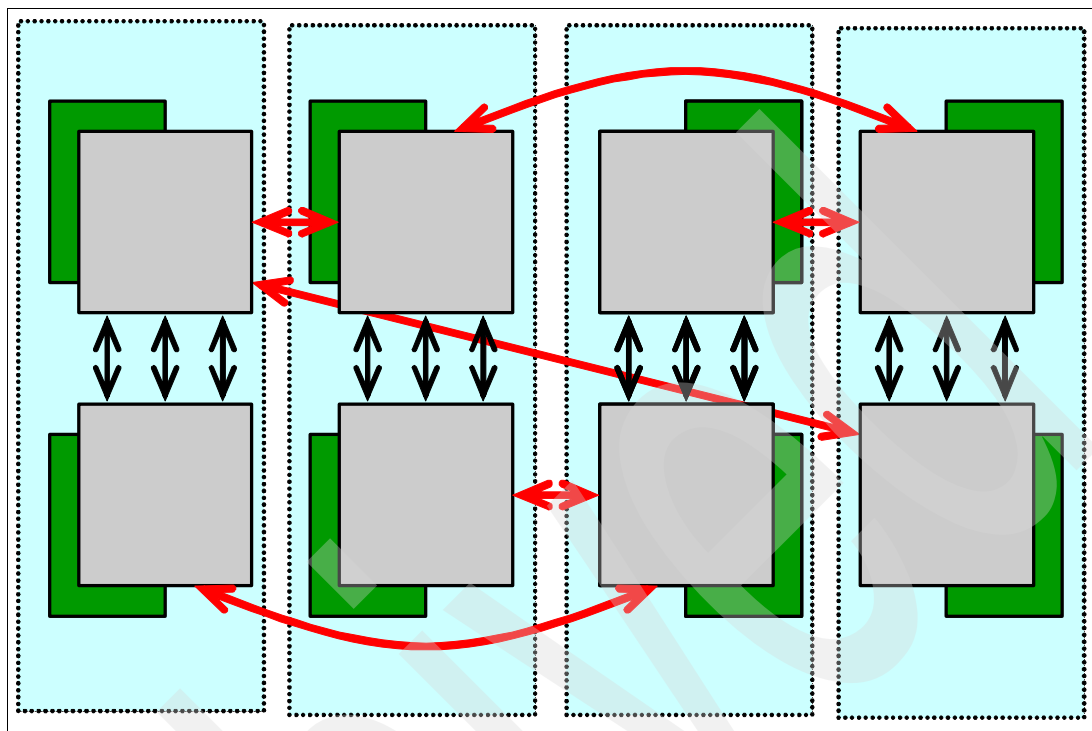


Figure 7-24 The interconnection fabric design of Power 780

The hypervisor in POWER7 ensures that processors and memory resources are as close as possible to each other to use both processor and memory affinity. However, you can improve the affinity further during the setup of partitions by considering the eight cores and the amount of memory when partitioning the system. A good example is to design an eight processor partition even though there might be a need for a ninth processor eventually. The hypervisor takes the eight core request and attempts to place these eight cores on a single processor chip if possible and if they are not in conflict with other previously defined partitions.

7.7.2 Optimizing POWER7 performance through tuning system resources

All known IBM i tuning practices today can be applied to POWER7, as well. For example, providing enough processors and memory to each of the IBM i partitions is a good base for excellent performance. A good assumption is to use about 8 GB of memory for each of the processors being allocated in the partition to fully exploit the processing capabilities. Having enough disk arms available (not to be confused with enough disk capacity) is essential to sustain good throughput when running any commercial application.

There are a few tuning options in IBM i, which become more important and sensitive with POWER7 than with prior systems. Processors on POWER7 systems can be shared or dedicated to a partition. In addition, there can be multiple processor pools that are defined, each with a certain use and characteristics. Shared processors, in general, are highly effective in multithreaded applications, such as interactive 5250 applications, web-based applications using HTTP, application and database servers with hundreds or thousands of clients connected to the system, each with a separate thread.

Typically, these workloads are tuned for the best possible throughput. A good setup strategy and design for the best performance is to round up or round down processor fractions to full processors, if possible. For example, 0.9 or 1.1 processors need to be one processor, 1.9 or 2.1 processors can be better defined as 2 processors, and so on. This approach also provides a good, even relationship to the number of virtual processors for the assigned physical processors.

Dedicated processors are used best for ensuring that the processors are instantaneously available to the partition to guarantee good response times. Also, dedicated processors are best for applications that are sensitive to stolen level 2 and 3 cache information, as well as processor cycles being used by other partitions. Batch applications typically benefit most from dedicated processors. Also, other transaction-oriented applications with a high degree of response time requirements also benefit from dedicated processors.

POWER7 provides simultaneous multithreading (SMT), which consists of SMT4 (four processor threads concurrently) and automatic switching between single thread (ST), SMT2, and SMT4 mode. SMT is not always beneficial for all workloads. Single-threaded processes work more efficiently when executing in single-threaded mode rather than SMT2 or SMT4 mode. Because the system itself can determine what mode is best for the workload currently running, it relieves system administrators from having to make decisions and trade-offs for the best performance and highest optimization levels.

The POWER7 system can run in either POWER6 (compatibility mode) or in POWER7 mode. The mode is determined and set in the HMC (refer to Figure 7-25). Although mostly ignored with IBM i, there are a few slight consequences in terms of SMT. POWER6 supports only SMT2, and therefore, a POWER7 system in POWER6 mode only runs in SMT2 mode.

Logical Partition Profile Properties: P4 @ P7sys1 @ Server-8233-E8B-SN10026EP - P7sys1

General Processors Memory I/O Virtual Adapters Power Controlling Settings Logical Host Ethernet Adapters (LHEA) Tagged I/O OptiConnect

Detailed below are the current processing settings for this partition profile.

Processing mode

☐ Dedicated
☒ Shared

Processing units

Total managed system processing units : 16.00
 Minimum processing units : 0.1
 Desired processing units : 0.1
 Maximum processing units : 1.0
 Shared processor pool: DefaultPool (0)

Virtual processors

Minimum processing units required for each virtual processor : 0.10
 Minimum virtual processors : 1.0
 Desired virtual processors : 1.0
 Maximum virtual processors : 1.0

Sharing mode

☒ Uncapped Weight : 255
 Processor compatibility mode: POWER7

OK Cancel Help

Figure 7-25 Configure processor compatibility mode

IBM i versions 6.1 and 7.1 are supported on POWER7; however, there are differences in terms of the number of hardware threads that can be active at any single point in time. The maximum value for threads supported by IBM i 6.1 is 128. The IBM i 7.1 supports 256 hardware threads on POWER7 in a single partition. You can run up to 64 processors in either version, but 6.1 supports only SMT2 when 64 processors are used. IBM i 7.1 supports 64 processors with SMT4, up to a maximum of 256 hardware threads.

In case you need more than 64 processors in one IBM i partition, you can request special support from IBM Lab Services by going to the following website:

<http://www.ibm.com/systems/services/labservices>

7.8 Enhanced performance tools of AIX for POWER7

In this section, we introduce new features with AIX commands on the POWER7 server platform:

- ▶ Monitoring POWER7 processor utilization
- ▶ Monitoring power saving modes
- ▶ Monitoring CPU frequency using the **lparstat** command
- ▶ Monitoring hypervisor statistics
- ▶ Capabilities for 1024 CPU support
- ▶ Monitoring block I/O statistics
- ▶ Monitoring Active Memory Expansion (AME) statistics
- ▶ Monitoring memory affinity statistics
- ▶ Monitoring the available CPU units in a processor pool
- ▶ Monitoring the remote node statistics using the perfstat library in a clustered AIX environment

7.8.1 Monitoring POWER7 processor utilization

POWER7 introduces improved reporting of the consumed capacity of a processor. This section explains the difference in processor utilization reporting between POWER5, POWER6, and POWER7.

Figure 7-26 on page 306 illustrates how processor utilization is reported on POWER5, POWER6, and POWER7. On POWER5 and POWER6, when one of the two hardware threads in SMT2 mode is busy (T0) while the other hardware thread is idle (T1), the utilization of the processor is 100%. On POWER7, the utilization of the processor in SMT2 mode is around 69%, providing a better view about how much capacity is available.

In SMT4 mode, with one hardware thread busy (T0) and the other three hardware threads idle (T1, T2, and T3), the utilization of the processor is around 63%. The processor's utilization in SMT4 mode is less than in SMT2 mode, because it has more capacity available through the additional two hardware threads.

POWER5/POWER6 SMT2		POWER7 SMT2		POWER7 SMT4	
T0	busy 100% busy	T0	busy ~69% busy	T0	busy ~63% busy
T1	idle	T1	idle	T1	idle
				T2	idle
				T3	idle

Figure 7-26 Differing processor utilization among POWER5, POWER6, POWER7 SMT2, and SMT4

The following examples demonstrate the CPU utilization for a single-threaded program running in SMT4, SMT2, and ST modes. All of these measurements were taken on a POWER7 LPAR with two physical processors.

The first example (Example 7-13) demonstrates the CPU utilization, as reported by the `sar` command, when running a single-threaded application in SMT4 mode. It shows that the single-threaded program consumed an entire logical CPU (cpu4) but not the entire capacity of the processor.

Example 7-13 CPU utilization when the single-thread process is running with SMT4

```
# sar -P ALL 1 20
```

```
System configuration: 1cpu=8 mode=Capped
17:49:05 cpu    %usr    %sys    %wio    %idle    physc
17:49:07  0         0        2         0        98        0.25
           1         0        0         0       100        0.25
           2         0        0         0       100        0.25
           3         0        0         0       100        0.25
           4        100        0         0         0        0.63
           5         0        0         0        99        0.12
           6         0        0         0       100        0.12
           7         0        0         0       100        0.12
           -        32         0         0        68        2.00
```

Example 7-14 shows the output after switching the SMT mode from 4 to 2 (`smtctl -t 2`). The same program is running on logical cpu5, and it consumes an entire logical CPU but now it is consuming 69% of the processor's capacity.

Example 7-14 CPU utilization when the single-thread process is running with SMT2

```
# sar -P ALL 2 10
```

```
System configuration: 1cpu=4 mode=Capped
17:48:18 cpu    %usr    %sys    %wio    %idle    physc
17:48:20  0         0        2         0        98        0.50
           1         0        0         0       100        0.50
           4         0        0         0       100        0.31
           5        100        0         0         0        0.69
           -        35         0         0        65        2.00
```

Example 7-15 shows the output after switching SMT mode from 2 to single thread (`smtctl -t 1`). The same program is running on logical cpu4, and it consumes the entire capacity of a processor because there are no other hardware threads available to execute code.

Example 7-15 CPU utilization when single-thread process is running with ST (single thread) mode

```
# sar -P ALL 2 10
```

```
System configuration: 1cpu=2 mode=Capped
17:47:29 cpu    %usr    %sys    %wio    %idle
17:47:31  0        0        1        0        99
           4       100        0        0        0
           -       50        1        0        50
```

If you want to learn more about processor utilization on Power Systems, refer to the IBM developerWorks article, “Understanding Processor Utilization on Power Systems - AIX”. This article covers in detail how processor utilization is computed in AIX and what changes it has undergone in the past decade in sync with the IBM Power processor technology changes:

<http://www.ibm.com/developerworks/wikis/display/WikiPtype/Understanding+Processor+Utilization+on+POWER+Systems+-+AIX>

7.8.2 Monitoring power saving modes

In AIX 6.1 TL6 or AIX 7.1, there are enhanced features with the `lparstat`, `topas`, and `topas_nmon` commands. They can display power saver modes now.

There are four values of the power saving mode, as showed in Table 7-10.

Table 7-10 Description of power saving modes

Value	Description
Disabled	Power saver mode is disabled
Static	Static power savings
Dynamic-performance	Dynamic power savings favoring performance
Dynamic-power	Dynamic power savings favoring power

Example 7-16, Example 7-17, and Example 7-18 on page 308 show the power saving mode features.

Example 7-16 Monitoring the power saving mode using lparstat -i

```
# lparstat -i | grep Power
```

```
Power Saving Mode : Static Power Savings
```

Example 7-17 Monitoring the power saving mode using topas -L

```
Interval:2          Logical Partition: p29n01          Tue May 24 20:38:01 2011
Psize:    64.0      Shared SMT          4          Online Memory:    96.00G
                                Power Saving: Static
Ent: 2.00          Mode: Capped          Online Logical CPUs:16
Partition CPU Utilization          Online Virtual CPUs:4
```

%usr	%sys	%wait	%idle	physc	%entc	app	vcs	phint	%lbusy	%hypv	hcalls
0.0	0.4	0.0	99.6	0.01	0.69	63.98	264	0	0.2	0.0	0

LCPU	MINPF	MAJPF	INTR	CSW	ICSW	RUNQ	LPA	SCALLS	USER	KERN	WAIT	IDLE	PHYSC	LCSW
0	69.0	0	275	308	121	0	100	128.00	4.0	87.1	0.0	8.9	0.01	227
2	0	0	17.0	0	0	0	0	0	0.0	1.9	0.0	98.1	0.00	17.0
3	0	0	10.0	0	0	0	0	0	0.0	0.9	0.0	99.1	0.00	10.0
1	0	0	10.0	0	0	0	0	0	0.0	0.8	0.0	99.2	0.00	10.0

Example 7-18 Monitoring the power saving mode using nmon and entering “r”

```
+--topas_nmon--h=Help-----Host=p29n01-----Refresh=2 secs---20:38.56
| Resources -----
| OS has 16 PowerPC_POWER7 (64 bit) CPUs with 16 CPUs active SMT=4
| CPU Speed 3864.0 MHz          SerialNumber=105E85P MachineType=IBM,9179-MHB
| Logical partition=Dynamic    HMC-LPAR-Number&Name=1,p29n01
| AIX Version=7.1.0.2 TL00 Kernel=64 bit Multi-Processor
| Power Saving=Static
| Hardware-Type(NIM)=CHRP=Common H/W Reference Platform Bus-Type=PCI
| CPU Architecture =PowerPC Implementation=POWER7
| CPU Level 1 Cache is Combined Instruction=32768 bytes & Data=32768 bytes
|   Level 2 Cache size=not available   Node=p29n01
| Event= 0 ---      ---      SerialNo Old=---      Current=F65E85 When=---
```

Tip: The Power Saving Mode field shows “-” with the `lparstat -i` command and shows “Unknown” with the `topas -L` and `nmon` commands when the power modes are not supported.

7.8.3 Monitoring CPU frequency using lparstat

The IBM energy saving features let the user modify the CPU frequency. The frequency can be set to any selected value (static power saver mode) or can be set to vary dynamically (dynamic power saver mode). In AIX 5.3 TL11 and AIX 6.1 TL4, the `lparstat` command provides new options to monitor the CPU frequency:

► **-E**

It shows both the actual and normalized CPU utilization metrics.

► **-W**

It works with the `-E` flag to provide longer output.

Example 7-19 shows the report when executing the command in one LPAR in static power saving mode. The actual frequency is 3.1 GHz. The nominal frequency is 3.864 GHz.

Example 7-19 Monitoring processor frequency using lparstat -E

```
#lparstat -E 2
```

System configuration: type=Shared mode=Capped smt=4 lcpu=16 mem=180224MB ent=2.00

Power=Static

Physical Processor Utilisation:

-----Actual-----					-----Normalised-----			
user	sys	wait	idle	freq	user	sys	wait	idle


```

-----
1.926 0.006 0.000 0.068 3.1GHz [ 81%] 1.565 0.005 0.000 0.430
1.926 0.005 0.000 0.069 3.1GHz [ 81%] 1.565 0.004 0.000 0.431
1.891 0.009 0.000 0.099 3.1GHz [ 81%] 1.537 0.008 0.000 0.456
-----

```

When the LPAR is running, there are 16 logical CPUs running in 4-way SMT mode, the processing entitled capacity is 2 CPUs, and the Virtual Processor number is 4. The actual metrics use PURR counters, and the normalized metrics use the SPURR counters¹³.

The values that are shown in each mode are the actual physical processors that are used in each mode. Adding up all the values (user, sys, idle, and wait) equal the total entitlement of the partition in both the actual and normalized view. The current idle capacity is shown by PURR, and the idle value that is shown by SPURR is what the idle capacity is (approximately) if the CPU is run at the nominal frequency.

Example 7-20 shows the report of the **lparstat -Ew** command with its long output.

Example 7-20 Monitoring the processor frequency using lparstat -Ew

```

#lparstat -Ew 2

System configuration: type=Shared mode=Capped smt=4 lcpu=16 mem=180224MB ent=2.00
Power=Static
Physical Processor Utilisation:
-----Actual-----
-----Normalised-----
      user   sys      wait   idle      freq      user      sys
wait      idle
-----
1.8968[95%] 0.0046[0%] 0.0000[0%] 0.0985[5%] 3.1GHz[81%] 1.5412[77%] 0.0038[0%]
0.0000[0%] 0.4551[23%]
1.8650[93%] 0.0100[1%] 0.0031[0%] 0.1218[6%] 3.1GHz[81%] 1.5153[76%] 0.0082[0%]
0.0025[0%] 0.4740[24%]
1.8944[95%] 0.0047[0%] 0.0000[0%] 0.1009[5%] 3.1GHz[81%] 1.5392[77%] 0.0038[0%]
0.0000[0%] 0.4570[23%]
1.8576[93%] 0.0057[0%] 0.0017[0%] 0.1349[7%] 3.1GHz[81%] 1.5093[75%] 0.0047[0%]
0.0014[0%] 0.4846[24%]
-----

```

Clarification: In shared uncapped mode, the result of this command, as shown in Example 7-20, might not be true, because the actual processor consumption can exceed the entitlement. So, in this case, adding these values might not be equal.

In Example 7-20, the Power field does not show when power modes are not supported.

7.8.4 Monitoring hypervisor statistics

The power hypervisor (PHYP) is the most important component to power virtualization technology. It is a firmware that resides in flash memory. Sometimes, we need to monitor its activities.

The AIX commands **topas** and **lparstat** show the hypervisor statistics.

¹³ For detailed information about PURR and SPURR, refer to the following website:
<http://www.ibm.com/developerworks/wikis/display/WikiPtype/CPU+frequency+monitoring+using+lparstat>

topas

Two hypervisor statistics are displayed in a report of the **topas -L** command:

- **hpi**

The aggregate number of hypervisor page faults that have occurred for all of the LPARs in the pool.

- **hpit**

The aggregate of time that is spent in waiting for hypervisor page-ins by all of the LPARs in the pool in milliseconds.

Example 7-21 shows the report after running **topas -L**.

Example 7-21 Monitoring hypervisor statistics using topas -L

Interval:2

Logical Partition: lpar1

Tue May 24 21:10:32 2011

Psize: -

Shared SMT 4

Online Memory: 6.00G

Power Saving: Disabled

Ent: 0.10

Mode: Un-Capped

Online Logical CPUs: 4

Mmode: Shared

IOME: 668.00

Online Virtual CPUs: 1

Partition CPU Utilization

%usr	%sys	%wait	%idle	physc	%entc	app	vcs	phint	hpi	hpit	pmem	iomu
0.3	6.3	0.0	93.4	0.02	15.44	-	278	0	2.00	0	1.78	13.05

=====

LCPU	MINPF	MAJPF	INTR	CSW	ICSW	RUNQ	LPA	SCALLS	USER	KERN	WAIT	IDLE	PHYS	LCSW
0	55.0	0	278	158	74.0	0	100	121.00	3.9	85.9	0.0	10.2	0.01	229
1	0	0	14.0	16.0	8.00	0	100	4.00	0.6	4.1	0.0	95.3	0.00	19.0
3	0	0	18.0	0	0	0	0	0	0.0	1.6	0.0	98.4	0.00	18.0
2	0	0	11.0	0	0	0	0	0	0.0	1.0	0.0	99.0	0.00	11.0

lparstat -h

The following statistics are displayed when the **-h** flag is specified:

- **%hypv**

This column indicates the percentage of physical processor consumption spent making hypervisor calls.

- **hcalls**

This column indicates the average number of hypervisor calls that are started.

Example 7-22 shows the report after running the **lparstat -h** command.

Example 7-22 Monitoring hypervisor statistics using lparstat -h

# lparstat -h 2													
System configuration: type=Shared mode=Capped smt=4 lcpu=16 mem=180224MB psize=64 ent=2.00													
%user	%sys	%wait	%idle	physc	%entc	lbusy	app	vcs	phint	%hypv	hcalls	%nsp	
0.0	0.4	0.0	99.6	0.02	0.8	1.3	64.00	874	0	86.3	963	81	
0.0	0.3	0.0	99.6	0.01	0.6	0.0	64.00	694	0	99.4	790	81	
0.0	0.5	0.0	99.5	0.02	0.8	1.9	63.99	92	0	82.5	108	81	

lparstat -H

The following statistics are displayed when the **-H** flag is specified.

The report of the **lparstat -H** command provides detailed hypervisor information. This option displays the statistics for each of the hypervisor calls. We list the various hypervisor statistics that are displayed in columns by this option for each of the hypervisor calls and the description of that statistic:

- ▶ **Number of calls**
Number of hypervisor calls made
- ▶ **%Total Time Spent**
Percentage of total time spent in this type of call
- ▶ **%Hypervisor Time Spent**
Percentage of hypervisor time spent in this type of call
- ▶ **Avg Call Time (ns)**
Average call time for this type of call in nanoseconds
- ▶ **Max Call Time (ns)**
Maximum call time for this type of call in nanoseconds

Example 7-23 shows the output of the **lparstat -H** command.

Example 7-23 Monitoring hypervisor statistics using lparstat -H

```
lparstat -H 2
```

System configuration: type=Shared mode=Capped smt=4 lcpu=16 mem=180224MB psize=64 ent=2.00

Detailed information on Hypervisor Calls					
Hypervisor Call	Number of Calls	%Total Time Spent	%Hypervisor Time Spent	Avg Call Time(ns)	Max Call Time(ns)
remove	0	0.0	0.0	0	1250
read	0	0.0	0.0	0	0
nclear_mod	0	0.0	0.0	0	0
page_init	11	0.0	0.0	480	1937
clear_ref	0	0.0	0.0	0	0
protect	0	0.0	0.0	0	0
put_tce	0	0.0	0.0	0	0
xirr	2	0.0	0.0	1140	1593
eoi	2	0.0	0.0	546	1375
ipi	0	0.0	0.0	0	0
cppr	2	0.0	0.0	312	500
asr	0	0.0	0.0	0	0
others	0	0.0	0.0	0	3875
enter	11	0.0	0.0	360	2812
cede	518	75.8	99.8	54441	2595109
migrate_dma	0	0.0	0.0	0	0
put_rtce	0	0.0	0.0	0	0
confer	0	0.0	0.0	0	0
prod	16	0.0	0.0	480	2000
get_ppp	1	0.0	0.0	3343	4031
set_ppp	0	0.0	0.0	0	0
purr	0	0.0	0.0	0	0

pic	1	0.1	0.1	31750	31750
bulk_remove	0	0.0	0.0	0	0
send_crq	0	0.0	0.0	0	0
copy_rdma	0	0.0	0.0	0	0
get_tce	0	0.0	0.0	0	0
send_logical_lan	0	0.0	0.0	0	0
add_logical_lan_buf	0	0.0	0.0	0	0

More tools: There are many other useful trace tools to monitor the hypervisor activities, for example, the CPU utilization reporting tool (**curt**). Refer to the IBM information center website or man manual for more information.

7.8.5 Capabilities for 1024 CPU support

On the POWER7 server, the maximum logical CPU¹⁴ is 1024 (256 cores)¹⁵. IBM announced that AIX 7.1 supports 1024 logical CPUs in one LPAR. So, many tools in AIX 7.1 were enhanced to support this feature. The modified tools are **mpstat**, **sar**, and **topas**. The tools include an additional capability to generate XML reports, which enables applications to consume the performance data and generate the required reports. The tools that support XML output are **sar**, **mpstat**, **vmstat**, **iostat**, and **lparstat**. In this section, we introduce several of these tools.

mpstat

The **mpstat** tool has this syntax:

mpstat -0 (option for sorting and filtering)

mpstat [{ -d | -i | -s | -a | -h }] [-w] [-0 Options] [-@ wparname] [interval [count]]

There are three values for the **-0** option for the **mpstat** command to sort and filter data. The following options are supported:

- ▶ **sortcolumn** = The name of the metrics in the **mpstat** command output.
- ▶ **sortorder** = [asc|desc]. The default value of **sortorder** is **desc**.
- ▶ **topcount** = The number of CPUs to be displayed in the **mpstat** command sorted output.

To see the list of the top 10 users of the CPU, enter the following command, which is shown in Example 7-24 on page 313:

```
mpstat -w -0 sortcolumn=us,sortorder=desc,topcount=10 2
```

¹⁴ The definition of a logical processor is that it is the basic unit of processor hardware that allows the operating system to dispatch a task or execute a thread. Intelligent thread technology dynamically switches the processor threading mode (SMT) between 1, 2, and 4 threads per processor core to deliver optimal performance to your applications. Each logical processor can execute only one thread context at a time.

¹⁵ At the time of writing this book, if you want to configure more than 128 cores in one LPAR with the FC4700 processor, you need to purchase software key FC1256 and install it in the server. The name of this code is the "AIX Enablement for 256-cores LPAR".

Example 7-24 Example of the sorting and filtering function of the mpstat command

```
# mpstat -w -0 sortcolumn=us,sortorder=desc,topcount=10 2
```

System configuration: lcpu=16 ent=2.0 mode=Capped

cpu	min	maj	mpc	int	cs	ics	rq	mig	lpa	sycs	us	sy	wa	id	pc	%ec	lcs
15	0	0	0	100	0	0	0	0	-	0	100.0	0.0	0.0	0.0	0.12	6.2	100
14	0	0	0	100	0	0	0	0	-	0	100.0	0.0	0.0	0.0	0.12	6.2	100
7	0	0	0	100	0	0	0	0	-	0	99.9	0.1	0.0	0.0	0.13	6.6	100
11	0	0	0	100	0	0	0	0	-	0	99.9	0.1	0.0	0.0	0.12	6.2	100
6	0	0	0	100	0	0	0	0	-	0	99.9	0.1	0.0	0.0	0.13	6.7	100
13	0	0	0	100	6	3	0	0	100.0	0	99.9	0.1	0.0	0.0	0.12	6.2	100
3	0	0	0	100	0	0	0	0	-	0	99.9	0.1	0.0	0.0	0.12	6.2	100
5	0	0	0	100	0	0	0	0	100.0	0	99.9	0.1	0.0	0.0	0.15	7.7	100
9	0	0	0	100	70	34	0	0	100.0	1	99.9	0.1	0.0	0.0	0.13	6.3	100
12	0	0	0	100	31	15	0	0	100.0	1	99.9	0.1	0.0	0.0	0.12	6.2	100
ALL	0	0	0	1760	125820	125655	0	0	0.0	194	95.8	0.2	0.0	4.0	2.00	99.9	1602

sar

The **sar** command has this syntax:

sar -0 (option for sorting and filtering)

```
/usr/sbin/sar [ { -A [ -M ] | [ -a ] [ -b ] [ -c ] [ -d ] [ -k ] [ -m ] [ -q ] [ -r ] [ -u ] [ -v ] [ -w ] [ -y ] [ -M ] } ] [ -P processoridentifier, ... | ALL | RST ] [-0 {sortcolumn=col_name[,sortorder={asc|desc}[,topcount=n}]]] [ [ -@ wparname ] [ -e[YYYYMMDD]hh [:mm [:ss ] ] ] [ -ffile ] [ -iseconds ] [ -ofile ] [ -s[YYYYMMDD]hh [:mm [:ss ] ] ] [-x] [ Interval [ Number ] ]
```

There are three values for the **-0** option of the **sar** command to realize sorting and filtering. The following options are supported:

- ▶ **sortcolumn** = The name of the metrics in the **sar** command output.
- ▶ **sortorder** = [asc|desc] The default value of **sortorder** is **desc**.
- ▶ **topcount** = The number of CPUs to be displayed in the **sar** command sorted output.

To list the top 10 CPUs, which are sorted on the **scall/s** column, enter the following command, as shown in Example 7-25:

```
sar -c -0 sortcolumn=scall/s,sortorder=desc,topcount=10 -P ALL 1
```

Example 7-25 Example of the sorting and filtering function of the sar command

```
# sar -c -0 sortcolumn=scall/s,sortorder=desc,topcount=10 -P ALL 1
```

System configuration: lcpu=16 ent=2.00 mode=Capped

11:44:38	cpu	scall/s	sread/s	swrit/s	fork/s	exec/s	rchar/s	wchar/s
11:44:39	4	653	176	180	0.00	0.00	12545	8651
	0	9	0	0	0.00	0.00	0	0
	1	5	0	0	0.00	0.00	0	0
	8	2	0	0	0.00	0.00	0	0
	9	1	0	0	0.00	0.00	0	0
	12	1	0	0	0.00	0.00	0	0
	7	0	0	0	0.00	0.00	0	0
	2	0	0	0	0.00	0.00	0	0
	5	0	0	0	0.00	0.00	0	0
	3	0	0	0	0.00	0.00	0	0
	-	1016	180	180	0.99	0.99	17685	8646

topas

There are two enhanced features for the **topas** command:

- Implement **topas** panel freezing. Use the spacebar key as a toggle for freezing.

The spacebar key on the keyboard acts as a toggle for freezing the **topas** panel. If frozen, **topas** stops data collection and continues to display the data from the previous iteration. You can move around the panel and sort the data based on the selected column. In the frozen state, if you move between panels, certain panels might not display the data. In this case, press the spacebar key to unfreeze the **topas** panel.

- Implement **topas** panel scrolling and sorting. Use the Page Up and Page Down keys for scrolling.

If the amount of data is more than the **topas** window size, you use the Page Up and Page Down keys to scroll through the data. The data is sorted based on the selected column.

Table 7-11 lists the freezing and scrolling properties of the **topas** command.

Table 7-11 Freezing and scrolling properties of the **topas** command

Panel	Freezing	Scrolling
Process Panel	Y	Y
Logical Partition	Y	Y
Tape Panel	Y	Y
Disk Panel	Y	Y
Tape Panel	Y	Y
SRAD Panel	Y	Y
Volume Group	Y	Y
File System	Y	Y
WLM	Y	Y
WPAR	Y	Y
CEC	N	N
Cluster	N	N
Adapter	N	N
Virtual I/O server	N	N

XML output commands

The following XML output commands are for **lparstat**, **vmstat**, **iostat**, **mpstat**, and **sar**:

- **iostat** [-X [-o filename]] [interval[count]]
- **vmstat** [-X [-o filename]] [interval [count]]
- **lpartstat** [-X [-o filename]] [interval[count]]
- **mpstat** [-X [-o filename]] [interval[count]]
- **sar** [-X [-o filename]] [interval[count]]

The following features are for the XML output commands:

- ▶ The default output file name is *command_DDMMYYHHMM.xml* and is generated in the current directory.
- ▶ The user can specify the output file name and the directory using the **-o** flag:

```
lparstat -X -o /tmp/lparstat_data.xml
```
- ▶ These XML schema files are shipped with the base operating system under */usr/lib/perf*:
 - iostat_schema.xsd
 - lparstat_schema.xsd
 - mpstat_schema.xsd
 - sar_schema.xsd
 - vmstat_schema.xsd
- ▶ Currently, the XML output that is generated by these commands is not validated as per the schema. It is up to the application to perform this validation.

7.8.6 Monitoring block IO statistics

In AIX 6.1 TL6 or AIX 7.1, there are enhanced features with the **iostat** command. The new **-b** option was added to capture data to help identify I/O performance issues and correct the problem more quickly:

- ▶ The **-b** option provides the block device utilization report, which shows detailed I/O statistics for block devices.
- ▶ The block I/O stats collection has been turned off by default.
- ▶ The root user can turn it on with the **raso** tunable command **raso -o biostat=1**.
- ▶ The **-b** option can be used by the root user, as well as a non-root user.
- ▶ The minimum value that can be specified for the interval is 2 seconds.
- ▶ Syntax:

```
iostat -b [block Device1 [block Device [...]]] Interval [Sample]
```

Table 7-12 shows the column names and descriptions of the output report of the **iostat -b** command.

Table 7-12 The column names and descriptions of the output report from the **iostat -b** command

Column name	Description
device	Name of the device
bread	Indicates the number of bytes read over the monitoring interval. The default unit is bytes; a suffix is appended if required (1024=K, 1024K=M).
bwrite	Indicates the number of bytes written over the monitoring interval. The default unit is bytes; a suffix is appended if required (1024=K, 1024K=M).
rserv	Indicates the read service time per read over the monitoring interval. The default unit is millisecond.
wserv	Indicates the write service time per write over the monitoring interval. The default unit is millisecond.
rerr	Indicates the number of read errors over the monitoring interval. The default unit is numbers; a suffix is appended if required (1000 = K, 1000K = M, and 1000M = G).

Column name	Description
werr	Indicates the number of write errors over the monitoring interval. The default unit is numbers; a suffix is appended if required (1000 = K, 1000K = M, and 1000M = G).
reads	Indicates the number of read requests over the monitoring interval. The default unit is numbers; a suffix is appended if required (1000=K, 1000K=M, and 1000M=G).
writes	Indicates the number of write requests over the monitoring interval. The default unit is numbers; a suffix is appended if required (1000=K, 1000K=M, and 1000M=G).

Example 7-26 shows the description of the **raso** command's tunable parameter (**biostat**).

Example 7-26 The description of the biostat parameter of raso

```
# raso -h biostat
```

Help for tunable biostat:

Purpose:

Specifies whether block I/O device statistics collection should be enabled or not.

Values:

Default: 0

Range: 0, 1

Type: Dynamic

Unit: boolean

Tuning:

This tunable is useful in analyzing performance/utilization of various block I/O devices. If this tunable is enabled, we can use **iostat -b** to show I/O statistics for various block I/O devices.

Possible Value:

1 : Enabled

0 : Disabled

Examples

Example 7-27 shows turning on the **biostat**.

Example 7-27 Enable the analysis of the performance and utilization of various block I/O devices

```
# raso -o biostat=1
```

Setting biostat to 1

Example 7-28 shows monitoring block I/O devices using **iostat -b**. It shows that there are I/O activities on the **hdisk1** device.

Example 7-28 Monitor block I/O devices using iostat -b

```
# iostat -b 2
```

System configuration: lcpu=16 drives=3 vdisks=0

Block Devices :6

device	reads	writes	bread	bwrite	rserv	wserv	rerr	werr
hdisk0	0.00	0.00	0.000	0.000	0.00	0.00	0.00	0.00
hdisk1	319.00	0.00	319.000M	0.000	6.00	0.00	0.00	0.00

hd4	0.00	0.00	0.000	0.000	0.00	0.00	0.00	0.00
hd8	0.00	0.00	0.000	0.000	0.00	0.00	0.00	0.00
hd9var	0.00	0.00	0.000	0.000	0.00	0.00	0.00	0.00
hd2	0.00	0.00	0.000	0.000	0.00	0.00	0.00	0.00

7.8.7 Monitoring Active Memory Expansion (AME) statistics

AIX 6.1 with the 6100-04 TL SP2 release or AIX 7.1 introduced the new **amepat** command, which is an Active Memory Expansion (AME) planning and advisory tool.

amepat

The **amepat** command reports AME information and statistics, as well as provides an advisory report that assists you in planning the use of AME for existing workloads.

The AME planning and advisory tool **amepat** serves two key functions:

- ▶ Workload planning

You can run the **amepat** command to determine a workload that will benefit from AME and also to provide a list of possible AME configurations for a workload.

- ▶ Monitoring

When AME is enabled, the **amepat** tool is used to monitor the workload and AME performance statistics.

You can invoke the **amepat** command in two modes:

- ▶ Recording mode

In this mode, **amepat** records the system configuration and various performance statistics and places them into a user-specified recording file.

- ▶ Reporting mode

In this mode, **amepat** analyzes the system configuration and performance statistics, which were collected in real time or from the user-specified recording file to generate workload utilization and planning reports.

You can invoke the **amepat** command by using the System Management Interface Tool (SMIT). For example, you can use the **smit amepat** fast path to run this command.

Note: When you invoke **amepat** without specifying the duration or interval, the utilization statistics (system and AME) do not display any average, minimum, or maximum values. The utilization statistics only display the current values. The CPU utilization only displays the average from the system boot time.

In 6.5.4, “Active Memory Sharing configuration” on page 223, there is one testing scenario that introduces how to use the **amepat** command to analyze memory behavior and get the AME recommendation on a running AIX LPAR environment with a workload. It explains how to configure the AME attribution based on its recommendation. It also explains how to monitor the AME performance statistics with the **topas** command to see the benefit after enabling AME.

In fact, **amepat** can also be run in LPARs in which AME is already enabled. When used in this mode, **amepat** provides a report of other possible AME configurations for the workload.

Example 7-29 on page 318 shows one **amepat** report after AME has been enabled in this LPAR.

Example 7-29 The amepat report when AME is enabled

```
# amepat 1

Command Invoked           : amepat 1

Date/Time of invocation   : Fri May 27 17:07:15 EDT 2011
Total Monitored time      : 3 mins 32 secs
Total Samples Collected  : 1

System Configuration:
-----
Partition Name            : lpar2_p780
Processor Implementation Mode : POWER7
Number Of Logical CPUs      : 4
Processor Entitled Capacity : 1.00
Processor Max. Capacity    : 1.00
True Memory                : 4.00 GB
SMT Threads                : 4
Shared Processor Mode      : Enabled-Uncapped
Active Memory Sharing      : Enabled
Active Memory Expansion    : Enabled
Target Expanded Memory Size : 6.00 GB
Target Memory Expansion factor : 1.50


System Resource Statistics:
-----
Current

CPU Util (Phys. Processors)      : 0.98 [ 98%]
Virtual Memory Size (MB)            : 5115 [ 83%]
True Memory In-Use (MB)              : 4092 [100%]
Pinned Memory (MB)                  : 1240 [ 30%]
File Cache Size (MB)                 : 2 [ 0%]
Available Memory (MB)                : 1066 [ 17%]


AME Statistics:
-----
Current

AME CPU Usage (Phy. Proc Units)    : 0.89 [ 89%]
Compressed Memory (MB)             : 1648 [ 27%]
Compression Ratio                  : 2.61


Active Memory Expansion Modeled Statistics:
-----

Modeled Expanded Memory Size      : 6.00 GB
Average Compression Ratio        : 2.61


Expansion  Modeled True  Modeled  CPU Usage
Factor     Memory Size   Memory Gain  Estimate
-----
1.03       5.88 GB      128.00 MB [ 2%] 0.00 [ 0%]
1.10       5.50 GB      512.00 MB [ 9%] 0.00 [ 0%]
1.15       5.25 GB      768.00 MB [14%] 0.00 [ 0%]
1.20       5.00 GB       1.00 GB [20%] 0.13 [13%]
```

1.27	4.75 GB	1.25 GB [26%]	0.43 [43%]
1.34	4.50 GB	1.50 GB [33%]	0.73 [73%]
1.38	4.38 GB	1.62 GB [37%]	0.88 [88%]
1.50	4.00 GB	2.00 GB [50%]	0.89 [89%] << CURRENT CONFIG

Active Memory Expansion Recommendation:

The recommended AME configuration for this workload is to configure the LPAR with a memory size of 5.00 GB and to configure a memory expansion factor of 1.20. This will result in a memory gain of 20%. With this configuration, the estimated CPU usage due to AME is approximately 0.13 physical processors, and the estimated overall peak CPU resource required for the LPAR is 0.22 physical processors.

NOTE: amepat's recommendations are based on the workload's utilization level during the monitored period. If there is a change in the workload's utilization level or a change in workload itself, amepat should be run again.

The modeled Active Memory Expansion CPU usage reported by amepat is just an estimate. The actual CPU usage used for Active Memory Expansion may be lower or higher depending on the workload.

In one AME-enabled environment, several existing tools, including **vmstat -c**, **lparstat -c**, **svmon -O summary=ame**, and **topas**, have been enhanced to monitor AME statistics.

vmstat -c

The following new statistics are displayed when executing the **vmstat -c** command:

- **csz**
Current compressed pool size in 4 KB page units.
- **cfr**
Free pages available in compressed pool in 4 KB page units.
- **dxm**
Deficit in the expanded memory size in 4 KB page units.

Example 7-30 shows an example of the **vmstat -c** command.

Example 7-30 Monitor AME statistics using vmstat -c

```
# vmstat -c 2

System Configuration: lcpu=4 mem=6144MB tmem=4096MB ent=1.00 mmode=shared-E mpsz=24.00GB
kthr      memory          page        faults          cpu
-----
r  b    avm    fre csz cfr dxm  ci   co   pi po in   sy  cs  us sy id wa   pc   ec
17 17 1192182 389787 47789 2227 0  11981 34223 0 0 61 176 4849 54 43 1 3 1.00 100.0
12 15 1221972 359918 51888 2335 0  23081 39926 0 0 42 348 1650 52 44 1 3 1.00 100.3
12 15 1242037 340443 56074 4204 0  21501 33590 0 0 10 285 2849 56 39 2 3 1.00 99.7
23  0 1262541 320006 58567 3988 0  30338 41675 0 0 83 277 2204 52 45 0 3 1.00 100.0
12  7 1275417 306433 62665 3494 0  27048 35895 0 0 195 229 2802 49 47 0 4 1.00 100.0
```

lparstat -c

The following statistics are displayed only when the **-c** flag is specified. Refer to Example 7-31 on page 320.

- **%xcpu**
Indicates the percentage of CPU utilization for the Active Memory Expansion activity.
- **xphysc**
Indicates the number of physical processors used for the Active Memory Expansion activity.
- **dxm**
Indicates the size of the expanded memory deficit for the LPAR in MB.

Example 7-31 Monitoring AME statistics using lparstat -c

```
# lparstat -c 2
```

System configuration: type=Shared mode=Uncapped mmode=Shar-E smt=4 lcpu=4
mem=6144MB tmem=4096MB psize=14 ent=1.00

%user	%sys	%wait	%idle	physc	%entc	lbusy	vcsu	phint	%xcpu	xphysc	dxm
48.3	51.7	0.0	0.0	1.00	100.1	99.2	0	5	89.6	0.8968	0
41.1	55.2	3.8	0.0	1.00	100.1	92.9	67	2	88.3	0.8842	0
40.0	56.2	3.7	0.0	1.00	100.0	92.2	53	0	88.7	0.8863	0
43.6	54.1	2.2	0.0	1.00	100.2	95.5	44	2	72.4	0.7248	0
39.2	54.7	6.0	0.1	1.00	99.7	84.6	154	4	50.7	0.5049	0

svmon -O summary

The **svmon** command provides two options for **svmon -O summary**. Refer to Example 7-32.

- **ame**
Displays the Active Memory Expansion information (in an Active Memory Expansion-enabled system).
- **longreal**
Displays the Active Memory Expansion information (in an Active Memory Expansion-enabled system) in a long format.

Example 7-32 Monitor AME statistics using svmon -O

```
# svmon -O summary=ame
Unit: page
```

	size	inuse	free	pin	virtual	available	loaned	mmode
memory	1572864	1302546	270318	317841	1312425	269358	0	Shar-E
ucomprsd	-	865851	-					
comprsd	-	436695	-					
pg space	131072	4973						
	work	pers	clnt	other				
pin	215054	0	0	102787				
in use	1301468	0	1078					
ucomprsd	864773							
comprsd	436695							

True Memory: 104857

	CurSz	%Cur	TgtSz	%Tgt	MaxSz	%Max	CRatio
ucomprsd	866689	82.65	691918	65.99	-	-	-

comprsd 181887 17.35 356658 34.01 693230 66.11 2.53

 txf cxf dxf dxm
 AME 1.50 1.50 0.00 0

topas

The topas tool displays memory compression statistics in an Active Memory Expansion-enabled system with the **topas** command. Refer to Example 7-33. The following data is reported:

- ▶ **TMEM,MB**
True memory size, in megabytes
- ▶ **CMEM,MB**
Compressed pool size, in megabytes
- ▶ **EF[T/A]**
Expansion factors: Target and Actual
- ▶ **CI**
Compressed pool page-ins
- ▶ **CO**
Compressed pool page-outs

Example 7-33 Monitoring AME statistics using topas

Topas Monitor for host:lpnr2							EVENTS/QUEUES		FILE/TTY	
Fri May 27 18:14:33 2011 Interval:FROZEN							Cswitch	1388	Readch	4824
							Syscall	459	Writech	579
CPU	User%	Kern%	Wait%	Idle%	Physc	Entc%	Reads	52	Rawin	0
Total	49.3	48.7	2.0	0.0	1.00	99.88	Writes	1	Ttyout	579
							Forks	0	Igets	0
Network	BPS	I-Pkts	O-Pkts	B-In	B-Out	Execs	0	Namei	38	
Total	1.76K	25.02	1.00	1.13K	645.5	Runqueue	16.01	Dirblk	0	
							Waitqueue	0.0		
Disk	Busy%	BPS	TPS	B-Read	B-Writ					
Total	0.0	0	0	0	0					
FileSystem		BPS	TPS	B-Read	B-Writ					
Total		4.71K	52.04	4.71K	0					
Name	PID	CPU%	PgSp	Owner						
cmemd	655380	32.0	120K	root						
lrud	262152	16.3	76.0K	root						
nmem64	8519768	13.8	255M	root						
nmem64	5767306	6.5	255M	root						
nmem64	9895952	4.6	255M	root						
nmem64	5898280	4.6	255M	root						
nmem64	8716434	2.5	255M	root						
nmem64	5374162	2.4	255M	root						
nmem64	10420454	2.3	255M	root						
nmem64	9699524	2.2	255M	root						
nmem64	7995508	2.1	255M	root						
nmem64	9175286	2.1	255M	root						
nmem64	10223870	2.1	255M	root						
							PAGING		Real,MB	6144
							Faults	39717K	% Comp	84
							Steals	39567K	% Noncomp	0
							PgspIn	0	% Client	0
							PgspOut	0		
							PageIn	0	PAGING SPACE	
							PageOut	0	Size,MB	512
							Sios	0	% Used	4
									% Free	96
							AME			
							TMEM	4.00G	WPAR Activ	0
							CMEM	482.81M	WPAR Total	0
							EF[T/A]	1.5/1.5	Press: "h"-help	
							CI:39.7	KCO:38.4K	"q"-quit	

nmem64	8257664	1.9	255M	root
nmem64	10027162	1.4	255M	root
nmem64	7405742	1.0	255M	root
topas	10944586	0.5	1.93M	root
topas	3473408	0.1	3.70M	root
java	8323080	0.0	78.7M	root
java	9437240	0.0	48.6M	pconsole

nmon (topas_nmon)

The **nmon** command records the AME statistics in the nmon recording file. The file format for the nmon recording file is *.nmon, for example, lpar2c_p780_110603_1028.nmon.

Start the nmon recording by running **nmon -f** on the command line. For detailed information about how to use the **nmon** command, refer to the man manual (**man nmon**). The nmon recording file is created on the current directory or in the directory that you defined in the command line. After the nmon file is generated, download it and use the nmon analyzer tool to analyze and generate one .xls file, for example, lpar2c_p780_110603_1028.nmon.xls.

The following tags in the xls file have AME statistics.

MEM tag

The MEM tag includes the following details that relate to AME. The details are recorded if AME is enabled in the partition:

- ▶ Size of the compressed pool in MB
- ▶ Size of true memory in MB
- ▶ Expanded memory size in MB
- ▶ Size of the uncompressed pool in MB

Figure 7-27 shows the output of MEM tag in the lpar2c_p780_110603_1028.nmon.xls file.

Size of the Compressed pool (MB)	Size of true memory (MB)	Size of Expanded memory size (MB)	Size of the Uncompressed pool (MB)
276.7	4096	6144	3819.3
356.9	4096	6144	3739.1
364.9	4096	6144	3731.1
405.2	4096	6144	3690.8
405.2	4096	6144	3690.8

Figure 7-27 AME statistics in MEM tag of nmon file

MEMNEW tag

The MEMNEW tag includes the following detail that relates to AME, as shown in Figure 7-28. Figure 7-28 shows the percentage of total memory used for the compressed pool.

Memory New	Process%	FS cache%	System%	Free%	Pinned%	User%	Compressed Pool%
10:28:49	69.6	0.1	7.8	22.5	14.4	47.9	15.4
10:28:54	70.2	0.1	7.8	22	14.4	46.5	16.7
10:28:59	69.7	0.1	7.8	22.4	14.4	46.4	16.8
10:29:04	69.7	0.1	7.8	22.4	14.4	45.8	17.5

Figure 7-28 AME statistics in MEMNEW tag of nmon file

PAGE tag

The PAGE tag of the nmon file includes the following details that relate to AME:

- ▶ Compressed pool page-ins
Other tools, such as **topas**, call this CI.
- ▶ Compressed pool page-outs
Other tools, such as **topas**, this call CO.

Fix: Unfortunately, at the time of writing this book, the last two tags appear in Excel columns that are used by the nmon analyzer and therefore got overwritten. There is one temporary fix to resolve it (refer to the following website):

https://www.ibm.com/developerworks/mydeveloperworks/blogs/aixpert/entry/quick_temporary_fix_for_nmon_s_analyser_ame_stats37?lang=zh

Examples: The previous examples are only for illustrating how to use the commands and tools. The AME factor that is used with the workload in the examples is not the best factor.

7.8.8 Monitoring memory affinity statistics

IBM POWER7 processor-based systems contain modules that are capable of supporting multiple processor chips depending on the particular system. Each module contains multiple processors, and the system memory is attached to these modules. Although any processor can access all of the memory in the system, a processor has faster access and higher bandwidth when addressing memory that is attached to its own module rather than memory that is attached to the other modules in the system.

Several AIX commands have been enhanced to retrieve POWER7 memory affinity statistics, including **lssrad**, **mpstat**, and **svmon**.

lssrad

This is a new tool to display core and memory placement on an LPAR. The REF1 column in the output is the first hardware-provided reference point that identifies sets of resources that are near each other. The SRAD is the Scheduler Resource Allocation Domain column. You need to allocate cores and memory from the same REF1 and SRAD.

We explain the following new terminology:

- ▶ Resource allocation domain (RAD)
A collection of system resources (CPUs and memory)
- ▶ Scheduler Resource Allocation Domain (SRAD)
A collection of system resources that are the basis for most of the resource allocation and scheduling activities that are performed by the kernel

- Resource affinity structures:
 - Various hierarchies:
 - Two-tier (local/remote)
 - Three-tier (local/near/far)
 - AIX Topology Service – System detail level (SDL):
 - SRAD SDL
This affinity structure is used to identify local resources.
 - REF1 SDL (first hardware provided reference point)
This affinity structure is used to identify near/far memory boundaries.

Figure 7-29 shows the relationship among system topology, RAD topology, and the `lssrad` command output.

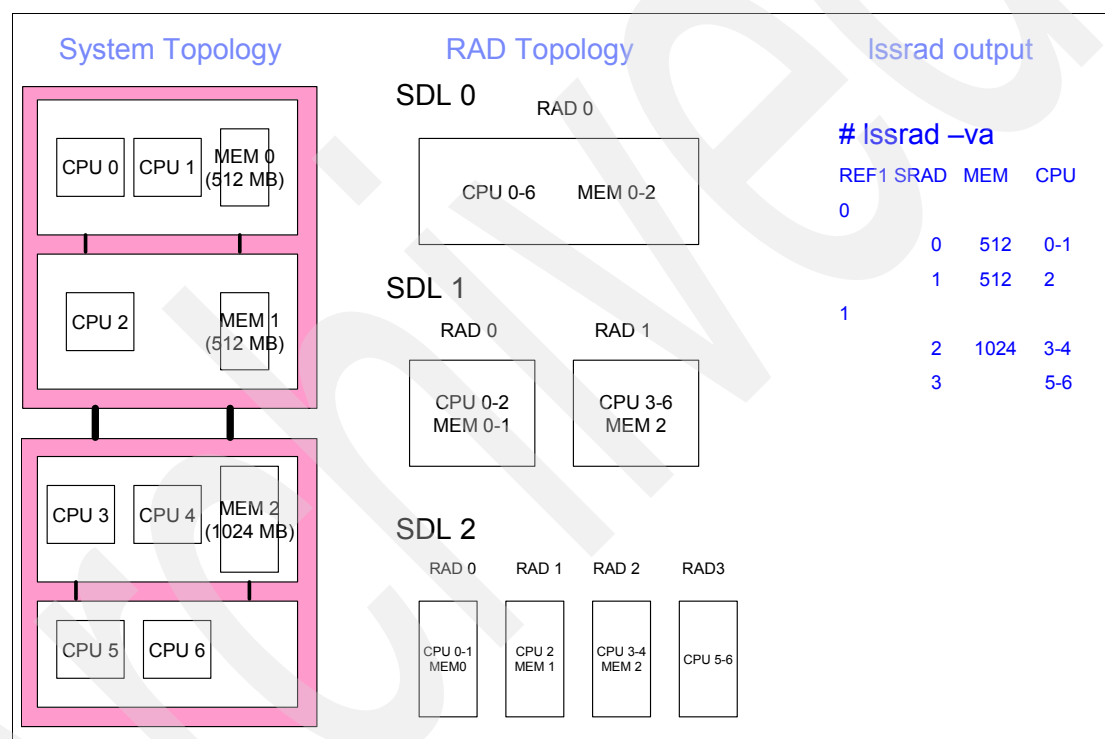


Figure 7-29 Relationship among system topology, RAD topology, and `lssrad` command output

Another example of `lssrad` (refer to Example 7-34) shows the placement of an LPAR with 12 cores (SMT2) and 48 GB memory.

Example 7-34 One example of `lssrad` command

```
# lssrad -av
```

REF1	SRAD	MEM	CPU
0			
	0	19900.44	0-1 4-5 8-9 12-13 16-17
	1	27519.19	20-21 24-25 28-29 32-33 36-37 40-41 44-45

mpstat

There is an enhanced feature for the **mpstat -d** command to display per logical CPU SRAD affinity. It adds three columns to the report. Refer to Table 7-13.

Table 7-13 Description of three new columns of the **mpstat -d** command report

Column name	Description
S3hrd	(-a, -d flag) The percentage of local thread dispatches on this logical processor
S4hrd	(-a, -d flag) The percentage of near thread dispatches on this logical processor
S5hrd	(-a, -d flag) The percentage of far thread dispatches on this logical processor

Example 7-35 shows the output of the **mpstat -d** command.

Example 7-35 Monitoring memory affinity statistics using **mpstat -d**

```
# mpstat -d 2

System configuration: lcpu=16 ent=2.0 mode=Capped
cpu   cs    ics  ...S2rd S3rd S4rd S5rd  ilcs  vlcs  S3hrd S4hrd S5hrd
0     709   286  ... 0.0  0.0  0.0  0.0    1   641  41.0  0.0  59.0
1      14     7  ... 0.0  0.0  0.0  0.0    0    36   0.0  0.0 100.0
2       0     0  ... -    -    -    -    0    49    -    -    -
3       0     0  ... -    -    -    -    0    29    -    -    -
14      0     0  ... -    -    -    -    0     4    -    -    -
ALL    723   293  ... 0.0  0.0  0.0  0.0    1   759  40.4  0.0  59.6
-----
0     994   404  ... 0.0  0.0  0.0  0.0    0   886  40.1  0.0  59.9
1      16     8  ... 0.0  0.0  0.0  0.0    0   48  0.0  0.0 100.0
2       0     0  ... -    -    -    -    0   68    -    -    -
3       0     0  ... -    -    -    -    0    40    -    -    -
ALL   1010   412  ... 0.0  0.0  0.0  0.0    0  1042  39.6  0.0  60.4
```

svmon

The **svmon** command provides options to display the memory affinity at the process level, segment level, or cpu level SRAD allocation for each thread:

- ▶ The affinity domains are represented based on SRADID and provide this information:
 - Memory information of each SRAD (total, used, free, and filecache)
 - Logical CPUs in each SRAD
- ▶ Display home SRAD affinity statistics for the threads of a process
- ▶ Provide application's memory placement policies

Table 7-14 on page 326 and Example 7-36 on page 326 show descriptions and examples.

Table 7-14 *svmon options, values, and descriptions*

svmon option	Value	Description
affinity	on	Displays memory affinity at the process level
	detail	Displays memory affinity at the segment level
	off (default)	Does not display the memory affinity
threadaffinity	on	Displays the cpu level SRAD allocation for each thread
	off (default)	Does not display the cpu level SRAD allocation for each thread

Example 7-36 *Monitoring memory affinity using svmon*

```
# svmon -O affinity=detail,threadaffinity=on -P 4522046
```

Unit: page

```
-----
      Pid Command      Inuse      Pin      Pgps Virtual
4522046 topasrec      26118    11476      0    25931
      Tid HomeSRAD LocalDisp NearDisp FarDisp
15401157      0      272      0      0
      Text Stack Data SHMNamed SHMAnon MapFile UnmapFil EarlyLRU
      Default Default Default Default Default Default Default N
Domain affinity Npages Percent lcpus
0      15494 59.8 200 0 1 4 5 8 9 12 13 16 17
1      10437 40.2 136 20 21 24 25 28 29 32 33 36 37 40 41 44 45
-----
```

7.8.9 Monitoring the available CPU units in a processor pool

In a micro-partition environment, you can use the **lparstat** command to monitor the current available physical processors in the shared pool, but you need to turn on “Allow processor pool utilization authority” through the HMC or SDMC. This LPAR property is on the processor Configuration tab. Figure 7-30 on page 327 shows the window where you enable processor pool utilization authority.

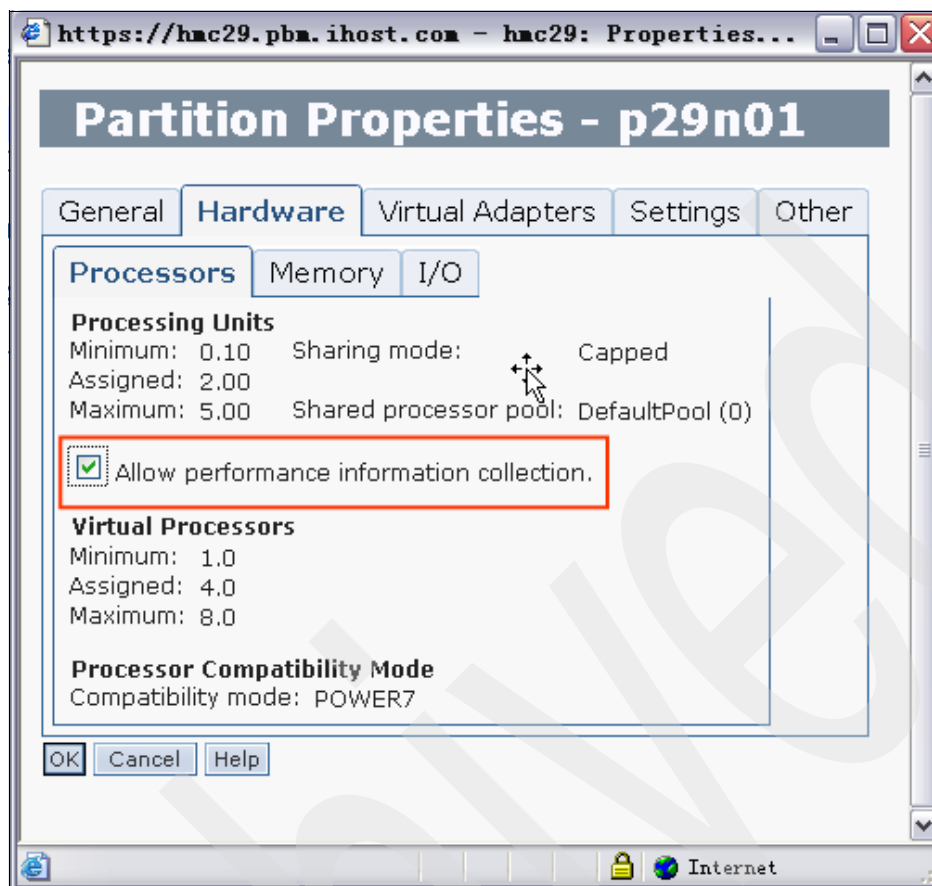


Figure 7-30 Turn on “Allow performance information collection” through the HMC

After selecting “Allow processor pool utilization authority” and starting the LPAR, from the AIX command, we can see the current available physical processors in the shared pool, as shown in Example 7-37.

Example 7-37 Monitor processor pool's available CPU units

```
# lparstat 2
```

```
System configuration: type=Shared mode=Capped smt=4 lcpu=16 mem=180224MB psize=64
ent=2.00
```

%user	%sys	%wait	%idle	physc	%entc	lbusy	app	vcs	phint
0.0	0.4	0.0	99.5	0.02	0.8	1.0	64.00	44	0
0.0	0.3	0.0	99.7	0.01	0.5	0.9	64.00	495	0
0.0	0.3	0.0	99.7	0.01	0.5	1.7	62.74	608	0

The app statistics column: In Example 7-37, the app statistics column takes affect immediately after you turn the “Allow processor pool utilization authority” option on or off from the HMC or SDMC.

7.8.10 Monitoring remote node statistics in a clustered AIX environment

In AIX 7.1 and AIX 6.1 TL6, the existing *perfstat* library is enhanced to support performance data collection and analysis for a single node or multiple nodes in a cluster. The enhanced *perfstat* library provides application programming interfaces (APIs) to obtain performance metrics that relate to processor, memory, I/O, and others to provide performance statistics about a node in a cluster.

The *perfstat* API is a collection of C programming language subroutines that execute in the user space and use the *perfstat* kernel extension to extract various AIX performance metrics. The *perfstat* library provides three kinds of interfaces for clients to use in their programs to monitor remote node statistics in a clustered AIX Environment:

- ▶ Node interface

Node interfaces report metrics related to a set of components or to the individual components of a remote node in the cluster. The components include processors or memory, and individual components include a processor, network interface, or memory page of the remote node in the cluster.

- ▶ Cluster interface

The *perfstat_cluster_total* interface is used to retrieve cluster statistics from the *perfstat_cluster_total_t* structure, which is defined in the *libperfstat.h* file.

- ▶ Node list interface

The *perfstat_node_list* interface is used to retrieve the list of nodes in the *perfstat_node_t* structure, which is defined in the *libperfstat.h* file.

For detailed information about the *perfstat* library and how to program these APIs in the client application, refer to the IBM information center website:

<http://publib.boulder.ibm.com/infocenter/aix/v7r1/index.jsp?topic=/com.ibm.aix.doc/doc/base/technicalreferences.htm>

7.9 Performance Management for Power Systems

Performance Management (PM) for Power is a tool to help you manage the growth and performance of your IBM Power Systems. It provides a comprehensive and secure performance management and capacity planning capability that can help ensure that your IT system is ready to meet your business opportunities and challenges.

PM for Power Systems supports the latest release of AIX 7.1 and the entire family of POWER7 processor-based systems, plus the Power processor blades in a BladeCenter.

PM for Power Systems presents a broad array of benefits to you, whether you are running IBM i or AIX. As a systems management tool, it helps you ensure the most performance from your system by continuously measuring growth and performance. These measurements allow you, your IBM Business Partner, or IBM to more quickly diagnose existing performance or capacity problems, identify potential resource constraints, and plan for future system growth. Measuring your system's performance and utilization trends and their effect on future resource requirements can help you make better informed and cost-effective decisions when planning for future system needs. In addition, PM for Power Systems provides the partition-level statistics that are needed to help you evaluate the benefits of increasing total system utilization through IBM consolidation and virtualization capabilities.

System and workload management tasks are an important aspect of a system administrator's role. The administrator has the responsibility to monitor and maintain the system, gather performance data, summarize results, and manage growth. PM for Power Systems offerings are designed to help you manage the performance of the IBM i and AIX systems in your enterprise.

Whether you have a single server with one LPAR or multiple servers with multiple LPARs, PM for Power Systems can save you time. These tools allow you to be proactive in monitoring your system performance, help you identify system problems, and help you plan for future capacity needs.

PM for Power Systems is an easy to use, automated, and self-managing offering. A collection agent that is specifically for PM for Power Systems is integrated into the current releases of IBM i and AIX. *This agent automatically gathers non-proprietary performance data from your system* and allows you to choose if you want to send it to IBM, at your discretion, on a daily or weekly basis. In return, you receive access to reports, tables, and graphs on the Internet that show your specific partition's (or total system if no partitioning is used) utilization, growth, and performance calculations.

For more information: More information about all facets of PM for Power Systems is available at this website:

<http://www.ibm.com/systems/power/support/perfmgmt>

7.9.1 Levels of support available within PM for Power Systems

There are two levels of support that are available within the PM for Power Systems service.

No additional charge summary-level service

If your IBM Power System server is still under warranty or if it is covered under an IBM hardware maintenance agreement, you receive the benefit of the management summary graph at no additional charge. This level provides an easy to implement and use process with interactive reporting that provides summary-level capacity trend information and performance management parameters for your IBM i-based or AIX-based systems and LPARs. Users are also allowed access to the Workload Estimator (WLE) for no additional charge to size future requirements based on the collected data.

After you register the partition with the registration key that IBM provides, you can view the performance management summary reports via a standard web browser. These reports are referred to as the management summary graphs (MSG). You can monitor the CPU and disk attributes of the system, measure capacity trends, and anticipate requirements. You can also merge the previously collected PM historical data with the IBM Systems Workload Estimator to size needed upgrades and so on. Flexibility is also provided so that you can arrange the information about specific partitions or systems in groups in the viewing tool, to make the information more meaningful to your operation.

Full-detail level (fee service)

Available as either an extension of the IBM Enhanced Technical Support offering or as a stand-alone offering (depending on your country location), this full detail level is a web-based fee service that provides access to many more detailed reports and graphs, again through a secure Internet connection. Additionally, detailed access to many of the graphs is provided using the interactive function that allows various time frame views of the data, including data transmitted as recently as the previous day. The detailed reports provide current information about resource constraints, resources approaching maximum capacity, disk files, processor

utilization, and memory utilization. Like the no additional charge summary graph offering, users are allowed unlimited access to the WLE for sizing future requirements.

There are several functions that are available to both the “no additional charge” and “fee” service options:

- ▶ Monthly .pdf of detailed reports
A .pdf of your entire report package is generated monthly by partition. It is viewable and downloadable from the website via your personalized secure password.
- ▶ Customizable graphs
PM for Power Systems provides the user with the ability to customize the reports and graphs by changing the time period. For example, instead of looking at a 30-day view, you can drill down to a seven day view, or even a daily view of the same information.

Additionally, the user has access to up to 24 months of history to redraw the same graph from a historical perspective, providing that the system or partition has been transmitting PM for Power Systems data consistently.

PM for Power Systems uses performance information and capacity information from your system. This data includes system utilization information, performance information, and hardware configuration information. After the data is collected, PM for Power Systems processes the data and prepares it for transmission to IBM for future analysis and report generation. Within IBM, the PM data will be used to prepare the reports and graphs that are delivered as part of the PM for Power Systems offering.

7.9.2 Benefits of PM for Power Systems

PM for Power Systems capabilities are automated, self-maintaining tools for single or multiple partition systems. IBM stores the data input that is collected by PM for Power Systems for you and helps you to perform these tasks:

- ▶ Identify performance bottlenecks before they affect your performance
- ▶ Identify resource-intensive applications
- ▶ Maximize the return on your current and future hardware investments
- ▶ Plan and manage consistent service levels
- ▶ Forecast data processing growth that is based on trends

The management of your system is simplified with the ability to stay abreast of utilization. This is true even if you run multiple partitions in separate time zones with a mixture of IBM i and AIX. If your system is logically partitioned (LPAR), we suggest that you enable PM Agent on all partitions.

Releases that are supported include those releases that have not reached their End of Program Support date. information about releases of AIX and IBM i that have not reached their End of Program Support dates is available at this website:

- ▶ IBM i
<http://www-947.ibm.com/systems/support/i/planning/upgrade/suptschedule.html>
- ▶ AIX
<http://www-01.ibm.com/software/support/systemsp/lifecycle/>

The PM for Power team diligently maintains start-up instructions by OS release on the PM for Power website, as shown on the Getting started page that is shown in Figure 7-31 on page 331.



Figure 7-31 PM for Power Getting started web page

Getting started is extremely easy and there is nothing to order or install on the Power server or operating system from a PM collection agent perspective.

7.9.3 Data collection

The collection of performance data using the PM agent is automated, self-managing, and done on a partition boundary. The Electronic Service Agent (ESA) automatically triggers the collection of non-proprietary performance data and automatically transmits the data to IBM based on the parameters that you have defined. From a performance standpoint, the collection programs use less than 1% of your processor unit.

The data is encrypted and sent to a secure IBM site. IBM automatically formats the raw data into reports, tables, and graphs that are easy to understand and interpret. You can review your performance data as often as you choose. Your information is updated daily from the previous collection if you transmit daily. It is updated weekly if you transmit weekly.

The automated collection mechanism relieves the system administrator of the time-consuming tasks that are associated with starting and stopping performance collections, and getting raw data into a readable format that can easily be analyzed. After the ESA is configured and data is transmitted, collection and reporting are self-managing. Data continues to be collected, transmitted to IBM, and then deleted from your system to minimize storage requirements. Your reports, graphs, and tables are available for analysis on the website to view at your convenience.

A self-maintained, automated approach allows you to be proactive in the analysis of your system performance. It provides a mechanism to avoid potential resource constraints and to plan for future capacity requirements.

It is important that systems and partitions transmit the PM for Power Systems data to IBM on a consistent basis. The collection agent and IBM Service Agent are designed to automatically continuously collect and periodically transmit the data to IBM; however, the possibility exists that transmissions might not occur for a variety of reasons.

To help make it easy for you to monitor whether transmissions are getting through to IBM on a consistent basis, an icon on the server information panel portrays a 90-day calendar of successful or unsuccessful transmissions, as seen in Figure 7-32.

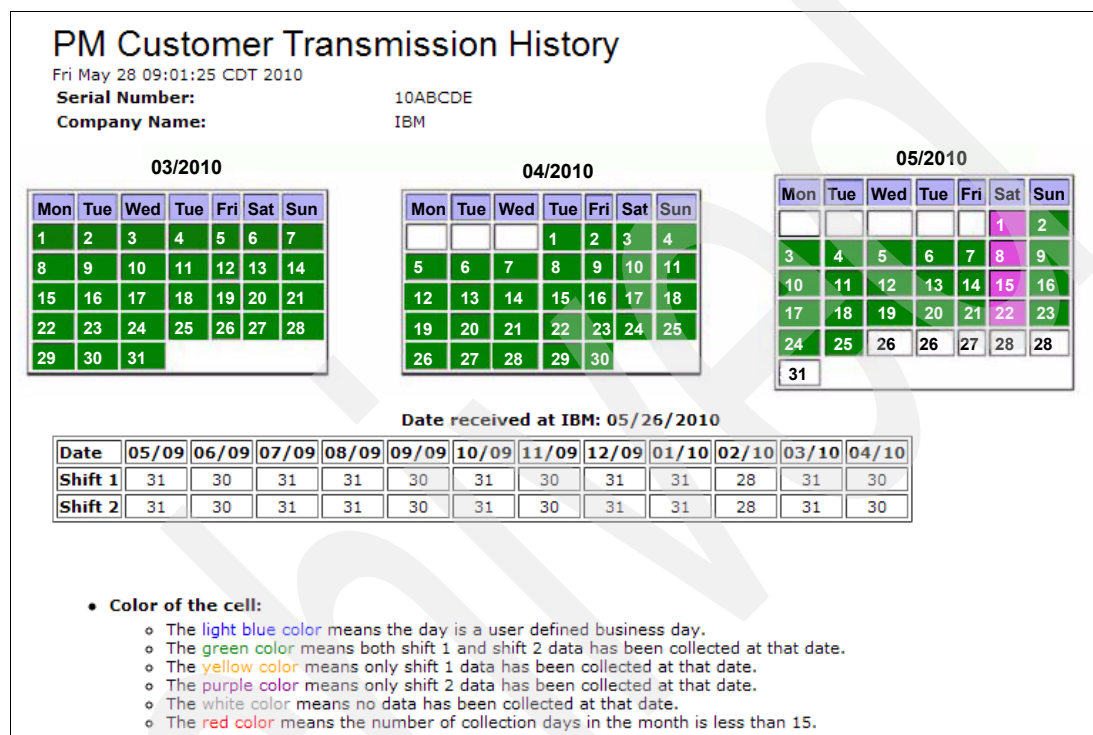


Figure 7-32 Customer Transmission History

7.9.4 Accessing the PM for Power Systems website

General information about PM for Power Systems, including any setup instructions, is available at the PM for Power Systems home page:

<http://www-03.ibm.com/systems/power/support/perfmgmt/index.html>

On the right side of that page is a call-out box for PM for Power Systems reports, with a link to the actual login page for report access:

<https://pmeserver.rochester.ibm.com/PMServerInfo/loginPage.jsp>

It is at this site that the clients must indicate that they are clients and enter their IBM Web IDs. Figure 7-33 on page 333 shows the login page.

IBM Performance Management for Power Systems United States [change] [Terms of use](#)

[Home](#) | [Products](#) | [Services & solutions](#) | [Support & downloads](#) | [My account](#)

Performance Management for Power Systems

Related links

- PM for Power Systems Home
- News and Announcements
- PM for Power Systems FAQs
- Country Contact Information
- PM for Power Systems Redbook
- Get Adobe® Reader®

Please select user login type, Then enter Userid and Password:

☐ **Customers login with:**
• IBM Web ID

☐ **Business Partners login with:**
• IBM Web ID

☐ **IBM Employees login with:**
• Intranet ID

Login ID:
Password:

Sign in

Hot Topics

- April 28, 2011 - Shared processor pool reporting for IBM i
- April 28, 2011 - Updated IBM PM for Power Systems: Graph Reference Document
- March 21, 2011 - New flyovers available
- January 26, 2011 - PM AIX activation streamlined
- November 19, 2010 - IBM Systems Director to transmit PM data
- October 12, 2010 - New File System detail report available
- July 16, 2010 - New detail charts in support of IBM i
- May 26, 2010 - Guideline changes for Disk Capacity
- Dec 11, 2009 - Important setup information regarding the new integrated PM AIX Collection Agent
- Jul 1, 2009 - PM data transmission via SNA not available

[About IBM](#) | [Privacy](#) | [Contact](#)

Figure 7-33 PM for Power Systems login page

After an LPAR or system is registered, the user is presented with the Server Information Panel (SIP) when signing on with the IBM Web ID. From this window, you can select the group of systems that you want to view. All systems and partitions are in one group if you do not differentiate them into separate groups when you register the system or partition.

The SIP provides information about each partition for both first and second shift. There are icons at the left of the window to use for the interactive graphing function or for requesting a .pdf of either the full service detail report set (fee) or the summary level (no additional charge) report.

Figure 7-34 on page 334 is an example of an SIP showing the icons to access to view reports. We also show the definitions of the icons for authorizing an IBM Business Partner to access your graphs and for checking the status of PM data transmission.





































Actions	Company Name	Serial Number ▲	LPAR	Mach Type	Model	Shift Nbr	O/S	GTS Cont
    	IBM	06ABCDE		9117	MMA	1	i5/OS	
    	IBM	06ABCDE		9117	MMA	2	i5/OS	
    	IBM	10ABCDE		9117	MMA	1	i5/OS	
    	IBM	10ABCDE		9117	MMA	2	i5/OS	
    	IBM	6512345	0001	9113	550	1	AIX	
    	IBM	6512345	0001	9113	550	2	AIX	

Figure 7-34 Server information panel in PM for Power

Important: For a full description of the PM for Power Systems process and an explanation of the detail level reports and graphs that are available, view the graph reference document at this website:

<http://www.ibm.com/systems/support/perfmgmt>

This website shows you all the latest information and documentation about PM for Power.

PowerCare Services offerings for Power Enterprise Servers

IBM Power Enterprise Servers are the most powerful and scalable members of our Power Systems family. They have been designed to provide clients the most cost-effective IT infrastructure. Power systems provide exceptional performance, massive scalability, and energy-efficient processing to meet the highest levels of computing requirements.

This chapter identifies the PowerCare Services offerings available to our Power Enterprise Server clients that have purchased Power Model 780 or 795 systems. IBM has integrated the PowerCare Services offering with your Power server. The services are independent of the operating system that you have chosen to deploy. This suite of service options, which is provided to you at no additional charge, offers you technical leadership and consulting resources to ensure that you effectively deploy and utilize your Power Systems environment, and the many reliability, availability, and serviceability (RAS) and virtualization features described within this IBM Redbooks publication.

We discuss the following topics in this chapter:

- ▶ PowerCare highlights
- ▶ PowerCare Services offerings

For more information: You can find the latest information and offerings on our PowerCare website at this website:

<http://www-03.ibm.com/systems/power/support/powercare/>

8.1 PowerCare highlights

We think that in order for our clients to concentrate on their core business issues, it is essential to provide them with world-class IT services that complement our world-class IT solutions. Our Power development teams have delivered outstanding RAS and virtualization technologies that can provide high levels of system availability in today's on-demand world. Keeping your business up and running is about managing and minimizing risks along with optimizing and maximizing the availability of your systems.

IBM Power Systems PowerCare Services bring skills and expertise to help you increase the business value from your Power Systems investments, independent of whether you use the IBM i or AIX operating system, or both. Every IBM Power 780 or 795 system is entitled to receive one of these PowerCare service choices per serial number at no additional charge:

- ▶ Availability optimization
- ▶ IBM Systems Director and VMControl enablement
- ▶ IBM Systems Director Active Energy Manager (AEM) enablement
- ▶ Security assessment
- ▶ Performance optimization
- ▶ Power Flex enablement
- ▶ Power 795 upgrade implementation service
- ▶ Technical training

Important: Every IBM Power 780 or 795 system is entitled to receive *one* PowerCare service choice per serial number at no additional charge.

8.2 PowerCare Services offerings

IBM Power Systems PowerCare Services help you use your installation to deliver real business value. The services are delivered by IBM Systems Lab Services and Training personnel around the globe to provide expertise in all aspects of managing Power Systems environments. This service includes your choice of one of the following services, which are described in the following sections, at no additional charge. These services are designed to assist you in taking advantage of emerging technologies on your Power Systems platform by bringing the skills and resources of the development lab to your enterprise via on-site consulting.

PowerCare benefit: PowerCare clients, who complete a PowerCare engagement and return the client feedback request, receive one complimentary admission to attend any IBM Power Systems conference worldwide. One tuition waiver per company for each qualifying serial number is eligible for this special offer.

The tuition waiver provides us an opportunity to say thank you and to provide you the opportunity to explore the latest in Power Systems technology, and learn from IBM product developers and experts, as well as network with your peers.

Our Power Systems Technical University offers hundreds of sessions on a wide range of Power topics. Sessions include from beginner to advanced training levels, preferred practices, and certification testing. You hear details behind all the latest POWER7 announcements and have an opportunity to see all the latest Power System products and solutions in our solution center. Technical Universities are held across the globe in a location that is convenient to you.

If you have questions, contact the IBM PowerCare team at pwrcare@us.ibm.com.

8.2.1 Availability optimization services

With the availability optimization service, you can choose one of two options to fit your requirements. Choices include either an analysis of your Power Systems availability technologies or a system health check of both hardware and operating system software. Our consultants bring the skills and experience of the development lab to you for your choice of on-site consulting, which assists you in taking advantage of emerging technologies, RAS features, and virtualization on your Power platform.

Availability optimization assessment

The availability optimization assessment (AOA) is designed to be an analysis of your Power Systems infrastructure based on your specific availability requirements. This independent availability assessment provides a high-level review of the overall IBM Power Systems environment availability readiness. The AOA reviews your current Power Systems environment and aims to proactively identify system exposures that might affect overall availability. Additionally, the assessment reviews the current system management processes that support the overall environment and captures any concerns and issues that you might have. The reviewer assists you in understanding how the new availability and virtualization features, such as concurrent maintenance, PowerVM Live Partition Mobility, and PowerHA, can strengthen the overall system's availability.

The AOA provides a summary of findings and observations, along with specific recommendations to meet your specific availability requirements, and supporting information that acts as education and backup to the recommendations. Recommendations are made based on preferred practices for availability that have been developed and learned through many client deployments of our enterprise servers worldwide. IBM recognizes that every client's environment is unique, as well as its business requirements, so preferred practice-based recommendations are tailored to meet your specific availability requirements.

There are twelve availability indicators that have been identified and each one is addressed during your AOA. Refer to Figure 8-1 on page 338.

Power Systems Availability Indicators

1. High Availability
2. Disaster Recovery
3. Single Points of Failure & Hardware Configuration
4. HMC, FSP and LPAR Configuration
5. Operating System Software Configuration
6. Patch and fix maintenance
7. Security and authority settings
8. Database
9. Backup and Recovery strategies
10. Communications
11. Data Center
12. Systems Management and Service Delivery processes

Four (R,O,Y,G) Assessment Metrics

Color	Description
Red	High risk of not meeting requirements Availability exposures exist that are likely to cause system outage and failure to meet business requirements.
Orange	Medium-High risk of not meeting requirements Need focus on Availability exposures which have capability to cause outage and high potential of not meet business requirements
Yellow	Medium risk of not meeting requirements Partially exploiting availability functions and features but remaining vulnerable to outages with potential to not meet business requirements.
Green	Low risk of not meeting requirements Fully exploiting availability functions and features. Likely to meet business requirements.

1

©2011 IBM Corporation

Figure 8-1 PowerCare availability optimization assessment indicators

Power Systems health check optimization

The Power Systems health check optimization is designed to assess the overall health of your Power Systems environment. The assessment aims to proactively identify hardware single system points of failure (SPOF) and other problems. The assessment aims to identify areas of exposure quickly before they can affect critical operations, including hardware, IBM software, and setup. With this option, you gain access to IBM preferred practices and technology that can maximize your resource utilization to improve your return on investment. With help from the IBM PowerCare team, you can gain immediate insights into how to optimize your Power systems environment and identify opportunities to strengthen overall system availability.

The end result of the health check is a comprehensive presentation, outlining system health status and trends, areas for improvement, and recommended actions. These actions can be performed by your staff or by IBM System Technology Group (STG) Lab Services in a separate engagement.

The IBM PowerCare team has identified 10 individual health check indicators, and each indicator is addressed during your PowerCare Health Check. Figure 8-2 on page 339 outlines the health check indicators.

PowerCare Healthcheck Assessment Indicators

Helping to keep critical systems up and running, enabling end user access 24x7

1. Single Points of Failure & Hardware configuration
2. HMC, FSP and LPAR Configuration
3. Operating System Software Configuration
4. Patch and fix maintenance
5. High Availability/Disaster Recovery
6. Security and authority settings
7. Journaling & Database
8. Backup and Recovery strategies
9. Communications
10. Systems Management and Service Delivery processes

Four (R,O,Y,G) Assessment Metrics

Color	Description
Red	URGENT - High risk Availability exposures exist that are likely to cause system outage and failure to meet business requirements.
Orange	WARNING - Medium-High risk Need focus on Availability exposures which have capability to cause outage and high potential of to not meet business requirements
Yellow	CAUTION - Medium risk Partially exploiting availability functions and features but remaining vulnerable to outages with potential to not meet business requirements.
Green	Low risk Fully exploiting availability functions and features. Likely to meet business requirements.

Figure 8-2 Health check indicators

8.2.2 Systems Director and VMControl enablement

IBM introduced IBM Systems Director VMControl in the first half of 2009, and IBM has significantly enhanced its capabilities since its introduction. VMControl is designed to automate the management of a virtualized infrastructure, to improve workload resiliency, and to help reduce deployment time for new virtual servers. VMControl is a plug-in for the IBM Systems Director, which is the IBM enterprise-wide management platform for servers, storage, networks, and software. After you install VMControl, it seamlessly integrates into Systems Director's browser-based interface, and VMControl can be used with systems that are already under Systems Director management.

IBM Systems Director and VMControl enablement provide simplified virtualization management, which enables faster problem-solving and helps you more effectively utilize your virtual environment. VMControl enables you to deploy virtual appliances quickly and easily and provides centralized management of virtual appliance assets, thus helping to increase systems administrator productivity.

For more information: Refer to one of the following websites for more information about IBM Systems Director and VMControl.

For IBM Systems Director:

<http://www.ibm.com/systems/management/director>

For IBM Systems Director VMControl:

<http://www.ibm.com/systems/management/director/plugins/vmcontrol>

IBM Systems Director VMControl Standard Edition utilizes a workload-optimized approach to decrease infrastructure costs and improve service levels. VMControl Standard Edition captures workloads from active systems and stores them into a repository as reusable system images, which are also referred to as *virtual appliances*. The VMControl Standard Edition also provides support to manage virtual appliances and automate the deployment of virtual appliances from a centralized repository.

The VMControl Standard Edition enables the creation and management of images to use when creating and deploying virtual workloads using AIX (by using Network Installation Management (NIM)) and Linux on Power. The definition of these system images is based on the Distributed Management Task Force (DMTF) Open Virtualization Format (OVF) specifications. The open design and support of industry standards enable IBM Systems Director and VMControl to provide heterogeneous physical and virtual management of multiple platforms and operating systems, which helps protect client IT investments.

Terminology differences: VMControl refers to a logical partition (LPAR) as a *virtual server*.

VMControl offers these benefits:

- ▶ Eliminates the installation, configuration, and maintenance costs that are associated with running complex stacks of software
- ▶ Reduces operational and infrastructure costs due to increased efficiency in using IT resources
- ▶ Manages a library of ready-to-deploy or customized system templates that meet specific hardware and software requirements
- ▶ Stores, copies, and customizes existing images to reuse them within system templates for creating virtual servers

The intent of the PowerCare Systems Director and VMControl enablement service is to set up a proof-of-concept (POC) environment in your Power environment. The service provides a highly skilled Lab Services and Training resource team at your location to help install, configure, and exploit the capabilities of IBM Systems Director and PowerVMControl. It is not intended to be the foundation of a production-ready solution, but to provide you with the ability to learn how to best use the features and functions that are important to you and to observe the configuration behavior in a safe and non-disruptive environment.

The IBM team works with the client's team to identify Power platform management requirements, issues, and strategies for your Power Systems environment. The team helps you to develop the foundation of your IBM Systems Director solution that addresses your specific objectives.

The strategies and objectives incorporate many features of IBM Systems Director VMControl and might include these features:

- ▶ Managing your virtualized environments
- ▶ Creating and managing virtual servers
- ▶ Managing a cross-platform environment
- ▶ Monitoring system resources and alerting with automation plans
- ▶ Updating management
- ▶ Discovering inventory and devices

There are several activities that must be completed before your on-site service can be performed. The IBM team needs to understand your current IT environment and your systems management objectives to identify what to include in the implementation. There are also

hardware, software, and network requirements for the installation of IBM Systems Director and VMControl that must be met to have the overall readiness of the environment.

At the end of the service, you receive a document that contains your identified Power platform management requirements and issues and an overview of your IBM Systems Director installation and configuration, along with a summary of the key functions that were implemented.

8.2.3 Systems Director Active Energy Manager enablement

The IBM Systems Director Active Energy Manager (AEM) measures, monitors, and manages the energy components that are built into IBM Power Systems, enabling a cross-platform management solution. AEM extends the scope of energy management to include facility providers to enable a more complete view of energy consumption within the data center.

AEM is an IBM Director extension that supports the following endpoints: IBM BladeCenter, Power Systems, System x, and System z servers. IBM storage systems and non-IBM platforms can be monitored through protocol data unit (PDU+) support. In addition, AEM can collect information from selected facility providers, including Liebert SiteScan from Emerson Network Power and SynapSense.

The AEM server can run on the following platforms: Windows on System x, Linux on System x, Linux on System p, and Linux on System z. AEM uses agent-less technology, and therefore no agents are required on the endpoints.

The objective of the AEM enablement service is to help you install, configure, and exploit the capabilities of IBM Systems Director AEM along with the core features of IBM Systems Director. The IBM team works with your team to understand your data center platform, as well as your energy and thermal management requirements. The IBM team works with the data center staff to show them how to use AEM to manage actual the power consumption that affects the thermal load that your IBM servers are placing on your data center. By providing hands-on skills transfer throughout the engagement, your staff learns how to best use the features and functions that are important within your environment.

Your PowerCare AEM enablement summary document contains this information:

- ▶ Your data center platform, energy, and thermal management requirements and issues
- ▶ An overview of your Systems Director and AEM installation and configuration
- ▶ A summary of the key functions that are implemented
- ▶ Next steps

8.2.4 IBM Systems Director Management Console

The PowerCare SDMC services are designed to set up a POC with the SDMC in your data center. The POC is intended to provide your staff with how to best use the SDMC features and functions that are important to your environment, and to observe the behavior of the features and functions in a non-disruptive setting. The IBM consultants can create a plan to extend this solution to a production environment if you want.

On-site PowerCare SDMC activities include these tasks:

1. Review your current environment and overall readiness for implementing SDMC.
2. Discuss your systems management objectives and clarify any new objectives.
3. Examine the features and capabilities of the SDMC for a Power Systems environment.
4. Develop a tactical plan to address the systems management objectives identified.

5. Install IBM Systems Director Common Agents on target managed systems for enabling SDMC management control of the systems.
6. Configuring and activating one or more of the SDMC functions might be included in the implementation of the platform management objectives.

Informal hands-on skills transfer is provided throughout the engagement for the Power staff to have a complete understanding of the SDMC.

At the end of your PowerCare services engagement, we develop the IBM SDMC engagement summary document that summarizes your systems and platform management requirements that were identified, along with the identification of managed endpoints that were included in the implementation.

8.2.5 Security assessment

In our current economic climate, many security and IT experts are warning enterprises that the threat for data theft is growing and can prove devastating for many enterprises. Good security is like an onion. Layered security provides the best protection, because it does not rely solely on the integrity of any one element. Multiple layers of security that cost potential outside intruder's time and dollars, because they must deal with and defeat successive layers of barriers. In the real world, multiple layers of security often cause the malicious attacker to get frustrated, or to simply run out of time and options before an actual vulnerability occurs.

The PowerCare security service offers you a choice of several options to meet your specific security needs:

- ▶ PowerCare security assessment for AIX or IBM i
- ▶ Single sign-on (SSO) for IBM i implementation assistance
- ▶ Lightweight Directory Access Protocol (LDAP)-Microsoft Active Directory (MSAD) AIX implementation workshop
- ▶ Role-based access control (RBAC) AIX implementation workshop

Security assessment

The PowerCare security assessment evaluates the security configuration of your Power Systems, identifying vulnerabilities before they can be exploited. New laws and regulations in the United States, such as Payment Card Industry (PCI), Sarbanes-Oxley (SOX), and the Health Insurance Portability and Accountability Act (HIPAA), are forcing organizations to be compliant with security and privacy requirements. Our security professionals can help you to adapt to new ways of working, and new ways of thinking about security and the regulatory requirements:

- ▶ Sarbanes-Oxley Act of 2002 (SOX), which requires compliance for software security and application security based around digital data integrity and accountability.
- ▶ Health Insurance Portability and Accountability Act of 1996 (HIPAA), which protects the privacy of individually identifiable health information. It sets national standards for the security of electronic protected health information and the confidentiality provisions of the patient safety rule, which protect identifiable information from being used to analyze patient safety events and improve patient safety.
- ▶ Payment Card Industry Data Security Standard (PCI DSS) standards include twelve requirements for any business that stores, processes, or transmits payment cardholder data. These requirements specify the framework for a secure payments environment.

Thorough documentation of the results and specific recommendations for mitigating the identified vulnerabilities and improving overall security posture are provided at the conclusion of the service. Do not consider the review a complete and comprehensive report on every aspect of security on your Power System. You cannot capture each and every exposure that might exist. We think that the findings and recommendations are an add-on to complement your well-thought-out existing security policy. The review can be used as a status check on many of the settings and configurations that exist at the time of the assessment. It can further be used as a measurement of how your system is configured in relationship to your security policy and other regulatory requirements that exist.

Every client's environment differs, so final recommendations are created that take into account the uniqueness of your particular environment. The security assessments focus on IBM AIX Version 5.3 or later and on IBM i 5.4 or later.

The security assessment proactively looks at several areas within AIX:

- ▶ System installation and configuration
- ▶ Auditing and logging
- ▶ File and directory permissions
- ▶ Login controls
- ▶ Security architecture
- ▶ AIX-specific security feature usage by release
- ▶ User and group accounts
- ▶ Passwords
- ▶ TCP/IP security
- ▶ Network File System (NFS) security

The security assessment on IBM i addresses the following security areas:

- ▶ System installation and configuration
- ▶ Auditing and logging
- ▶ Resource security
- ▶ Integrated file system (IFS) security
- ▶ NetServer security
- ▶ Network security, including TCP/IP
- ▶ Users and group controls

At the completion of the assessment, feedback is provided on how you can improve your current security implementation. It also includes advice on taking your security implementation to the next level to enhance your overall security strategy and to take advantage of the features that are provided with IBM Power Systems. Preferred practices can significantly help reduce security exposures in all environments.

Figure 8-3 on page 344 shows one example of the type of information that your assessment provides.

Quick Summary		
Area Reviewed	Potential Problems	Value Retrieved
Profiles with *ALLOBJ Special Authority	*YES	182
Profiles with *JOBCTL Special Authority	*YES	165
Profiles with *SPLCTL Special Authority	*YES	156
Profiles with Default Passwords	*YES	70
Profiles with Passwords that Never Expire (*NOMAX)	*YES	50
Group Profiles with Passwords	*YES	12
*ALLOBJ Special Authority through Group Profile	*TBD	5
*JOBCTL Special Authority through Group Profile	*TBD	7
*SPLCTL Special Authority through Group Profile	*TBD	2
Profiles with *PUBLICly Authorized Profiles	*YES	1
Audit Journal	*TBD	Yes
DDM Password Requirements	*YES	*NO
Does the *SYSTEM Store Exist	*TBD	*RW
ROOT (/) is Shared	*YES	Yes
ROOT (/) *PUBLIC Authority is *RWX	*YES	*RWX
Subsystems with *PUBLIC not *USE or *EXCLUDE	*YES	47
Job Descriptions with *PUBLIC not *USE or *EXCLUDE	*YES	161
Output Queues with *PUBLIC not *USE or *EXCLUDE	*YES	21
Job Queues with *PUBLIC not *USE or *EXCLUDE	*YES	15
*IBM Libraries with *PUBLIC not *USE or *EXCLUDE	*YES	10
USER Libraries with *PUBLIC not *USE or *EXCLUDE	*YES	922
QSECOFR Adoption in USER Libraries	*YES	635
AUTH Lists with *PUBLIC not *USE or *EXCLUDE	*YES	6
Allow Change to System Values	*YES	Yes
QSECURITY - System security level	*YES	30

Figure 8-3 Shows a quick summary of security-related exposures on IBM i

Preferred practice: Perform a security assessment yearly to provide a health check of your security effectiveness.

Single sign-on (SSO) for IBM

The objective of SSO is to have one single logon per user that allows users access to all applications and all systems that they require. Its goal is to provide a unified mechanism to manage the authentication of users and implement business rules that determine user access to applications and data. Without a unified sign-on strategy, developers re-implement custom security for each application, which can limit scalability and create on-going maintenance issues.

We implemented SSO technology within IBM i in V5R2 (4 June 2002). The implementation is designed where network users can access a Network Authentication Service (NAS) to automatically authenticate and authorize themselves to sign on to IBM i applications without entering a user profile and password.

The Enterprise Identity Mapping (EIM) table maps a network user's Microsoft Windows domain identity to specific user profiles for each i partition to which the user is authorized to sign on. The EIM table is maintained on an i partition and accessed through an LDAP server. You can obtain additional information about SSO in *Windows-based Single Sign-on and the EIM Framework on the IBM eServer iSeries Server*, SG24-6975.

There are many benefits that SSO can provide to your organization:

- ▶ Improved user production because users no longer have to remember multiple user IDs and passwords.
- ▶ Fewer requests to your help desk to reset forgotten passwords or locked accounts.
- ▶ Simplified administration because managing user accounts is much easier, but applications can still require additional user-specific attributes within the application.

You might choose to have IBM assist you with the implementation of SSO for your IBM i and show you how to avoid several of the pitfalls that slow down your deployment. Your PowerCare SSO service includes these tasks:

- ▶ Integrating up to four IBM i systems with your Microsoft Windows Server 2003 domain/active directory authentication environment.
- ▶ Configure your IBM i operating system correctly to participate in the environment.
- ▶ Configure one instance on an EIM domain to be used by all four instances of the IBM i operating system.
- ▶ Enable up to five user IDs to use the password elimination.
- ▶ Assist with the analysis of your critical applications relating to the ability and effort to allow participation in the password elimination/SSO environment.

IBM Lab Services EIM Populator utility: The PowerCare service includes the use of the IBM Lab Services EIM Populator utility during this engagement to assist with the loading of identity mapping information into EIM.

LDAP-MSAD AIX implementation workshop

Lightweight Directory Access Protocol (LDAP) is an open standard that provides a central mechanism for maintaining system configuration and policy information. This standard allows you to configure and manage multiple systems with a single set of user identity configuration information, which simplifies system administration tasks.

Today, most of your users begin their computing sessions by logging into a Microsoft Windows Active Directory (MSAD) domain. The ability to reuse their MSAD credentials is a significant benefit for users when their end destination is a Power server that requires additional login authentication. You can find more information about AIX and LDAP in *Integrating AIX into Heterogeneous LDAP Environments*, SG24-7165.

Your LDAP-MSAD implementation workshop provides an on-site demonstration using your AIX LDAP client partition and your MSAD server to show the capability of a user logging onto an AIX partition using a Windows account and password combination that has been properly enabled. Your IBM Services Specialist provides on-site training and skills transfer to explain the fundamentals of how LDAP is used with AIX to provide centralized user management and

how to use Microsoft's Identity Management with UNIX to enable Windows user and group accounts to be used by your Power AIX partitions.

8.2.6 Performance optimization assessment

Performance is a concern for IT worldwide. Dependencies on hardware, software applications, and distributed computing enhance the need for these systems to perform well. By properly balancing system resources, jobs can run at their optimal levels with minimal resource conflicts. To optimize your system's performance, you need to fully understand the business requirements that your Power System is addressing and be able to translate these business needs into performance objectives.

Changing performance objectives: Remember that as your business needs evolve and change, your performance objectives must also evolve and change.

The performance optimization assessment is for clients who want to obtain information that assists with their Power System performance optimization. The performance services offering is three-fold:

- ▶ Provide guidance with the usage of virtualization technologies to identify where consolidation and workload balancing apply. The major focus of server virtualization is to reduce IT complexity and total cost of ownership, while maximizing resource usage. Virtualization enables the consolidation of many disparate and potentially underutilized servers into fewer physical systems, which can result in reduced system management, software licensing, and hardware costs.
- ▶ Identify areas where machine consolidation applies.
- ▶ Provide a system health check with a focus on performance optimization, inspect the operating system running on specific LPARs for up-to-date fix levels and drivers, and provide memory, disk, and swap space system parameters.

IBM i and AIX: Both IBM i and AIX operating systems are included in the performance assessment.

The IBM Power Systems performance assessment is designed to help you optimize your Power Systems environment using virtualization, and to increase the efficiency of one or more of your critical partitions. The number of LPARs that can be analyzed depends on the complexity of your environment. The IBM Power Systems performance assessment aims to identify the areas where the performance of specific partitions can be fine-tuned by changing, where applicable certain settings:

- ▶ Micro-partitioning settings
- ▶ Partitioning and resource allocation
- ▶ Virtual I/O server setup
- ▶ Virtual Ethernet
- ▶ Virtual Fibre Channel (FC)
- ▶ Virtual storage
- ▶ Subsystem setup
- ▶ Journal configuration
- ▶ General system settings
- ▶ Additional configuration objects

Figure 8-4 on page 347 shows an example of a performance assessment summary.

Indicator	Score	Observations
Virtualization	Red	<ul style="list-style-type: none"> • Basic Virtualization implemented-LPARs. • Intermediate Virtualization features like shared processors, VIO Servers not implemented, but in plans. • Advanced Virtualization features like Live Partition Mobility are not possible due to lack of pre-requisites.
Consolidation	Yellow	<ul style="list-style-type: none"> • Grouping of production and development / test workloads achieved. • DR servers are also used for development and test workloads. During DR, non-business critical workloads will be shutdown to make way for production-in DR. • Additional consolidation can happen based on resource usage /VIO Servers. • Additional processors will be added support newer workloads.
Performance Review		
CPU performance	Green	<ul style="list-style-type: none"> • CPUs utilization is maximum for Singapore server. • Other workloads are running at optimal utilization.
Memory performance	Green	<ul style="list-style-type: none"> • Overall memory utilization is optimal. • A few best practice changes for memory management are recommended. • Memory management tuning recommendations can potentially reduce the possibility of paging during peak workload periods.
IO performance	Green	<ul style="list-style-type: none"> • Some servers running with visibly very high IO Waits. • At times, IO waits are higher than the actual system utilization at that time. • IOs pending due to non-availability of LVM, JFS2 and Device driver buffers increase daily.
Availability Review	Green	<ul style="list-style-type: none"> • FLRT and MDS reports indicate recommended upgrades. • Some of the updates needed are HIPER & PE.

Figure 8-4 Performance assessment summary

Performance data is collected from your identified Power System partition and analyzed using the appropriate operating system tools. Although performance improvements cannot be guaranteed, the intended result of the assessment is to provide you with recommendations to improve your overall system performance. If time permits, modifications to the system might be tested under the IBM team's supervision.

8.2.7 Power Flex enablement

Power Flex is a new and exciting capability on our high-end IBM POWER7 Power 795 Enterprise Server that allows more flexible use of purchased processor and memory activations across a pool of Power 795 systems to help increase the utility of these resources and to enhance your application availability.

A Power Flex infrastructure on the Power 795 can deliver unprecedented performance, capacity, and seamless growth for your AIX, IBM i, and Linux applications today and into the future. Flex Capacity Upgrade on Demand provides new options for resource sharing in support of large-scale workload consolidation and changing business and application requirements. Employing the strength and innovation of the IBM Power 795 server, PowerVM virtualization, and Capacity on Demand technology, Power Flex can enable your organization to more affordably deploy applications across two to four enterprise systems to enhance application availability. At the same time, Power Flex allows vital virtual processor and memory resources to be deployed precisely where you need them, dynamically adjusting capacity and even reallocating it beyond the boundaries of a single Power 795 system.

The Power Flex enablement service is centered around high availability and is intended for clients who want to implement Power Flex, and take advantage of all the capabilities that it provides. In addition, the offering takes a system health check in both hardware and software areas with the goal of improving system availability. The Power Flex offering is limited to analyzing up to four eligible IBM Power servers as part of a Power Flex pool.

IBM PowerCare experts assist you with the enablement of your Power Flex environment, which consists of two to four Power 795 systems, Flex Capacity on Demand option, On/Off Capacity on Demand, and optionally a PowerVM Live Partition Mobility. The Power Flex solution enables you to shift licensed capacity from one system to another system to perform scheduled system maintenance. It can also help you balance workloads and help handle peaks in demand.

The Power Systems Power Flex enablement and consulting workshop focus on your Power Systems hardware and software configuration and provide solution options that relate to Power Flex and high availability.

This offering includes these functions:

- ▶ Power Flex solution offering and design
- ▶ Hardware configuration options
- ▶ Power Flex licensing options
- ▶ Operational processes and expectations
- ▶ Software licensing
- ▶ Resource balancing scenarios
- ▶ Planned maintenance scenarios
- ▶ Capacity Upgrade on Demand and On/Off Capacity on Demand
- ▶ Planned outages utilizing Live Partition Mobility and Capacity Upgrade on Demand

The following availability components are addressed in conjunction with Power Flex:

- ▶ PowerHA solution scenarios and options
- ▶ Unplanned outages utilizing PowerHA and Capacity Upgrade on Demand
- ▶ Single system points of failure and hardware configuration
- ▶ Hardware Management Console (HMC), Fibre Channel Service Protocol (FSP), and LPAR configuration
- ▶ Operating system software levels and support
- ▶ Technology Levels and fix maintenance prerequisites

Figure 8-5 on page 349 shows a sample Power Flex enablement summary.



PowerCare Power Flex Enablement Summary

Power Flex Components	Status	Observations
Power Flex Licensing & Keys	✓	Power Systems Capacity on Demand Project Office has not received signed documents as of this enablement engagement
Advanced Planning	✓	Will use Advanced Planning to eliminate semi-annual maintenance window outages. They are aware of the 2 day lead-time requirement.
Resource re-balancing	✓	Planning to move activations when needed and can predict peak workload times
On/Off Capacity on Demand	✓	Currently using On/Off Capacity on Demand processor days to solve performance issues rather than to provide utility computing for short-term projects or workload spikes
Trial On/Off Capacity on Demand	✓	The client has already used 30-day trial capacity
Live Partition Mobility	✓	Actively using LPM on selected partitions outside of Power Flex
PowerHA System Mirror	✓	Very familiar with using PowerHA (HACMP) for failovers. Current version is V5.4.1 goes out of support 9/2011 and they are developing migration plan to V6.1
Single Points of failure & Hardware Configuration	✓	Each system within the Power Flexpool can handle 100% of the additional CPU requirements at this time. Hardware has excellent redundancy configuration
HMC, FSP and LPAR Configuration	✓	HMC code current; FSP code levels supported and configuration supports Power Flex PowerVM 2.2.0.10 (sp24) Newest fix pack installed.
Operating System Software Configuration	✓	Standardized on AIX 5.3 TL12 and AIX 6.1 TL6 across enterprise.
Technology Levels and fix maintenance	✓	Excellent fix level currency to support Power Flex

1

© 2011 IBM Corporation

Figure 8-5 Power Flex enablement summary

8.2.8 Power 795 upgrade implementation services

Our consultants bring the experience and expertise of IBM to assist you in the planning and implementation of an upgrade to a Power 795 environment. Our experience with other Power client upgrades worldwide can help you to identify and mitigate potential risks that might affect both the time and cost of your migration project. The PowerCare Power 795 upgrade implementation service works with you during the planning and deployment of a POWER6 595 to Power 795 upgrade. This service helps you to ensure that the LPARs, virtual I/O servers, and AIX configurations are prepared and backed up so that they can be quickly deployed after the hardware upgrade is complete. The PowerCare team has developed several tools to automate the data gathering and backup requirements and to ensure that the data is captured and readily available when you need to restore it. Data is captured from both AIX and IBM i operating system partitions.

Skills transfer: The IBM team works side-by-side with your team to develop and execute the project steps to deliver a successful Power 795 upgrade.

Upgrading the POWER6 595 to the Power 795 is a complex process, which requires careful coordination and planning. The Power 795 requires the current levels of operating systems and firmware on the HMC, virtual I/O servers, and all the LPARs that will be in place before the upgrade can be successfully completed. The PowerCare upgrade services deal strictly

with the technical work that is necessary to move workloads from your existing Power system to your new POWER7 system. Several aspects of the system configuration must be captured before the upgrade to ensure that the system can be fully returned to an operating state after the upgrade. Figure 8-6 shows a sample of the detailed hardware planning that is documented.

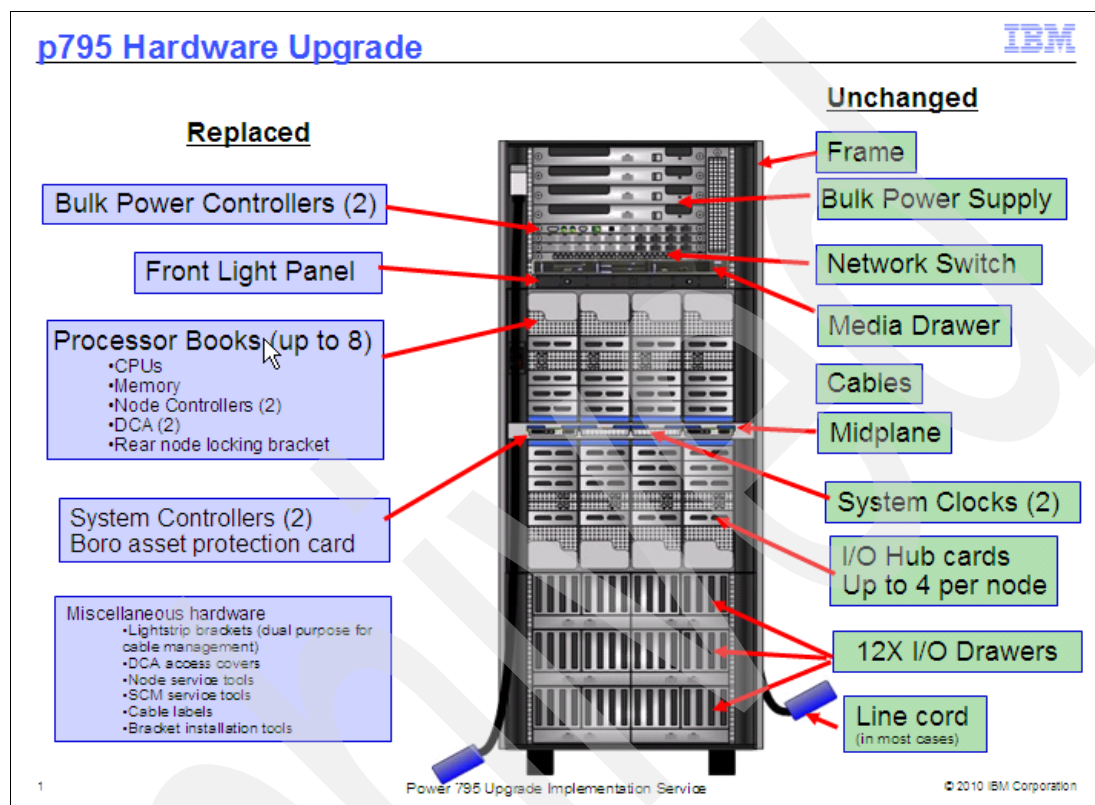


Figure 8-6 Hardware upgrade planning

Your Power technology services specialist from our STG Lab Services consulting organization provides support during the pre-upgrade/migration, upgrade/migration, and post-upgrade/migration phases for ensuring that the LPARs, virtual I/O servers, and AIX configurations, and software/firmware prerequisites are set up as required. This support includes the remote assessment of the source and target environments prior to the migration, on-site support to the IBM migration team, and post-migration support to assist with the reconfiguration and setup of the target environment LPARs, virtual I/O servers, and AIX, if required.

After your hardware upgrade has been completed, your PowerCare services specialist remains on-site while you boot all partitions and validate that the upgrade was successful.

8.2.9 PowerCare technical training

The IBM Systems Lab Services and Consulting team provides PowerCare service options to our Power 780 and 795 clients with technical leadership and consulting at no charge. We can help optimize the utilization of your Power System solutions. We are focused on the new technologies that are emerging from the IBM product development labs and the delivery and training of new and important technologies to our clients. The IBM Systems Lab Services and Training Power Services team is fully experienced in technology implementation in Power Systems environments. This highly trained part of the IBM Power Systems development lab

offers comprehensive knowledge of and experience with the products and solutions that can help you get more from your Power Systems investment.

Our newest PowerCare service option, PowerCare Technical Training, brings a subject matter expert (SME) to your site to help you increase the knowledge of your technical team on one of several Power course topics. You can choose from a selection of Power course topics and solution areas (Refer to Table 8-1):

- ▶ Availability
- ▶ Systems Director VMControl
- ▶ Systems Director/AEM
- ▶ Security
- ▶ Performance optimization

The PowerCare training courses consists of content that maps closely to the skills that are required to perform other PowerCare services options. These on-site training services provide lab exercises to complement the learning objectives. You can ask questions and learn about processes that are specific to your Power environment from experts in the industry. Your training also includes consultation with you on the next steps to address your training requirements through our training road maps.

Table 8-1 PowerCare course catalog

Course code	Course title	Operating system
AP21	Availability for Power Systems	AIX
AP23	Security for Power Systems	AIX
AP24	Security for IBM i	IBM i
AP25	Performance for Power Systems	AIX
AP26	Performance for IBM i	IBM i
AP27	Systems Director for Power Systems	AIX and IBM i

For more information: You can obtain more information about the IBM PowerCare education choices at this website:

<http://www-03.ibm.com/services/learning/ites.wss/ph/en?pageType=page&contentID=a0000775>

Archived

Administration concepts

This appendix provides IBM Power Systems administration with AIX concepts that relate to the testing that we performed and the examples in this publication.

We discuss the following topics:

- ▶ Making a root volume group (rootvg) easier to manage
- ▶ Example importing non-root volume group
- ▶ A dynamic LPAR operation using the HMC
- ▶ Setting up Secure Shell keys between two management consoles
- ▶ Simple cluster installation
- ▶ Installing and configuring PowerHA

Making a root volume group (rootvg) easier to manage

If at all possible, do not put non-system file systems in the rootvg. Only system data must reside in rootvg.

For instance, an administrator can install WebSphere in /usr/WebSphere or DB2 in /opt/IBM. Both /usr and /opt are operating system file systems. You end up with a large rootvg. Consider using a completely separate path. If the application is hard-coded to use only a certain path, consider creating a separate file system on another disk with the required path.

Non-system data makes it difficult to manage the root (rootvg) location configuration. Application configuration files, data, and log files increase exponentially and can make the rootvg large. Unless the /etc/exclude.rootvg file is created and updated with files that must not be backed up, all file systems on the rootvg are backed up with the **mksysb** command.

The **mksysb** backups, as well as the alternate disk clones, become very large and time-consuming to create and restore.

Example A-1 shows which of the file systems must be included in the rootvg. Any other file systems must be on separate disks (physical volumes), thus another volume group.

Example A-1 Rootvg without non-system files

```
# hostname
rflpar20
# lsvg -l rootvg
rootvg:
LV NAME                TYPE      LPs      PPs      PVs      LV STATE      MOUNT POINT
hd5                    boot      2        2        2        open/syncd    1
closed/syncd          N/A
hd6                    paging    32       32       1        open/syncd    N/A
hd8                    jfs2log   1        1        1        open/syncd    N/A
hd4                    jfs2      12       12       1        open/syncd    /
hd2                    jfs2      114      114      1        open/syncd    /usr
hd9var                jfs2      15       15       1        open/syncd    /var
hd3                    jfs2      4        4        1        open/syncd    /tmp
hd1                    jfs2      1        1        1        open/syncd    /home
hd10opt               jfs2      22       22       1        open/syncd    /opt
hd11admin             jfs2      8        8        1        open/syncd    /admin
lg_dump1v             sysdump   64       64       1        open/syncd    N/A
livedump              jfs2      16       16       1        open/syncd
/var/adm/ras/livedump
```

Important: Administrators might put only configurations and binaries (installation) on file systems other than the rootvg file system. Make sure that log files and data are *excluded* from the root volume group file systems. Restoring rootvg restores both the binaries and the base operating system (BOS).

For more information about volume groups and file system creation, refer to the *AIX Logical Volume Manager from A to Z: Introduction and Concepts*, SG24-5432.

Example importing non-root volume group

After a new and complete overwrite or a **mksysb** restore, it might be necessary to re-import the data volume groups. If the data resided on the root volume group, it might be lost. If the data and applications resided on a disk other than the rootvg disk, they can be recovered by importing the volume groups.

Example A-2 shows how to confirm the content of a disk after recovering the volume group.

Example A-2 Displaying how to import a volume group

After a system creation. df might show only system file systems

```
# df
Filesystem      512-blocks      Free %Used    Iused %Iused Mounted on
/dev/hd4         393216         70456   83%    11875   57% /
/dev/hd2         3735552        247072   94%    34250   53% /usr
/dev/hd9var       491520        139432   72%     6501   29% /var
/dev/hd3         163840        160536    3%         37    1% /tmp
/dev/hd1          32768         32064    3%         5    1% /home
/dev/hd11admin    262144        261416   1%         5    1% /admin
/proc            -              -      -         -     - /proc
/dev/hd10opt      720896        242064   67%     8144   23% /opt
/dev/livedump     524288        523552   1%         4    1% /var/adm/ras/livedump
#
```

This shows only root volume group volumes.

To check if there are any other disks run

```
# lspv
hdisk0          00c1f170c2c44e75          rootvg          active
```

This example shows only hdisk0 as being available.

Confirm if there are no other attached or lost disks by running lsdev.

```
# lsdev -Cc disk
hdisk0 Available   Virtual SCSI Disk Drive
```

The example still shows one disk

Run cfgmgr to see if any other devices are added but not in the systems Configuration (CuDv). And then, run lspv if the number of disks changes.

```
# cfgmgr
# lspv
hdisk0          00c1f170c2c44e75          rootvg          active
hdisk1          00f69af6dbccc5ed          None
hdisk2          00f69af6dbccc57f          None
#
```

Now, the system shows two more disks, but one volume group.

Confirm the contents of the disks before assuming they are not in use.

```
# lqueryvg -Atp hdisk1
0516-320 lqueryvg: Physical volume hdisk1 is not assigned to
a volume group.
0516-066 lqueryvg: Physical volume is not a volume group member.
Check the physical volume name specified.
#
# lqueryvg -Atp hdisk2
```

0516-320 lqueryvg: Physical volume hdisk2 is not assigned to a volume group.

```
Max LVs:      256
PP Size:      25
Free PPs:     311
LV count:     6
PV count:     1
Total VGDA's: 2
Conc Allowed: 0
MAX PPs per PV 32768
MAX PVs:      1024
Quorum (disk): 1
Quorum (dd):   1
Auto Varyon ? : 0
Conc Autovaryo 0
Varied on Conc 0
Logical:      00c1f17000004c0000000130095dd271.1 db2bin 1
              00c1f17000004c0000000130095dd271.2 loglv00 1
              00c1f17000004c0000000130095dd271.3 data1lv 1
              00c1f17000004c0000000130095dd271.4 data2lv 1
              00c1f17000004c0000000130095dd271.5 data1lg 1
              00c1f17000004c0000000130095dd271.6 data1tmp 1
Physical:     00f69af6dbccc57f      2 0
Total PPs:    317
LTG size:     128
HOT SPARE:    0
AUTO SYNC:    0
VG PERMISSION: 0
SNAPSHOT VG:  0
IS_PRIMARY VG: 0
PSNFSTPP:     140288
VARYON MODE:   ???????
VG Type:       2
Max PPs:      32768
Mirror Pool St n
```

Hdisk2 shows that it has some data, but no volume group. You can now import the volume. If you have documentation of the volume group name and Major number, You can specify it. PowerHA has Major number requirements not discussed in this topic

```
Import the volume
# importvg -y datavg hdisk2
datavg
# lspv
hdisk0      00c1f170c2c44e75      rootvg      active
hdisk1      00f69af6dbccc5ed      None
hdisk2      00f69af6dbccc57f      datavg      active
```

Confirm the contents of datavg

```
# lsvg -l datavg
datavg:
LV NAME      TYPE      LPs      PPs      PVs  LV STATE  MOUNT POINT
db2bin       jfs2      1        1        1    closed/syncd /opt/IBM/db2
loglv00      jfs2log   1        1        1    closed/syncd N/A
data1lv      jfs2      1        1        1    closed/syncd /userdata/dat1
```


data2lv	jfs2	1	1	1	closed/syncd	/userdata/dat2
data1lg	jfs2	1	1	1	closed/syncd	/userdata/dblog
data1tmp	jfs2	1	1	1	closed/syncd	/userdata/dbtmp

```
Mount the filesystems
# mount /opt/IBM/db2
# mount /userdata/dat1
# mount /userdata/dat2
# mount /userdata/dblog/
# mount /userdata/dblog
# mount /userdata/dbtmp
```

Rerun df and compare the results with that in the beginning of this example.

```
# df
Filesystem      512-blocks      Free %Used    Iused %Iused Mounted on
/dev/hd4         393216          70408   83%     11894   57% /
/dev/hd2         3735552         247072   94%    34250   53% /usr
/dev/hd9var       491520          139432   72%     6501   29% /var
/dev/hd3          163840          160536    3%         37    1% /tmp
/dev/hd1           32768           32064    3%         5    1% /home
/dev/hd11admin    262144          261416    1%         5    1% /admin
/proc            -                -         -         -         - /proc
/dev/hd10opt      720896          242064   67%     8144   23% /opt
/dev/livedump     524288          523552    1%         4    1% /var/adm/ras/livedump
/dev/db2bin       65536           64864    2%         4    1% /opt/IBM/db2
/dev/data1lv      65536           64864    2%         4    1% /userdata/dat1
/dev/data2lv      65536           64864    2%         4    1% /userdata/dat2
/dev/data1lg      65536           64864    2%         4    1% /userdata/dblog
/dev/data1tmp     65536           64864    2%         4    1% /userdata/dbtmp
```

A dynamic LPAR operation using the HMC

In this example, we remove an adapter dynamically from a running logical partition (LPAR) by using an HMC. We added this adapter in 2.10.2, “Dynamically changing the LPAR configurations (DLAR)” on page 58 using the Systems Director Management Console (SDMC). Follow these steps:

1. Run the **lsdev** command to list the devices on the LPAR, as shown in Example A-3.

Example A-3 Executing the lsdev command

```
# lsdev -Cc adapter
ent0 Available Virtual I/O Ethernet Adapter (1-lan)
fcs0 Available 20-T1 Virtual Fibre Channel Client Adapter
fcs1 Available 21-T1 Virtual Fibre Channel Client Adapter
fcs2 Available 22-T1 Virtual Fibre Channel Client Adapter
fcs3 Available 23-T1 Virtual Fibre Channel Client Adapter
vsa0 Available LPAR Virtual Serial Adapter
vscsi0 Defined Virtual SCSI Client Adapter
```

- Remove the adapter configuration from the operating system by issuing the following commands:

```
# rmdev -dl vscsi0
vscsi0 deleted
```

Confirm that the adapter has been removed, as shown in Example A-4.

Example A-4 Confirmation of the removal of the adapter

```
# lsdev -Cc adapter
ent0 Available Virtual I/O Ethernet Adapter (1-lan)
fcs0 Available 20-T1 Virtual Fibre Channel Client Adapter
fcs1 Available 21-T1 Virtual Fibre Channel Client Adapter
fcs2 Available 22-T1 Virtual Fibre Channel Client Adapter
fcs3 Available 23-T1 Virtual Fibre Channel Client Adapter
vsa0 Available LPAR Virtual Serial Adapter
#
```

- Log on to the HMC. Select **Systems Management** → **Servers**. Select the server. Select the required managed server. See Figure A-1.

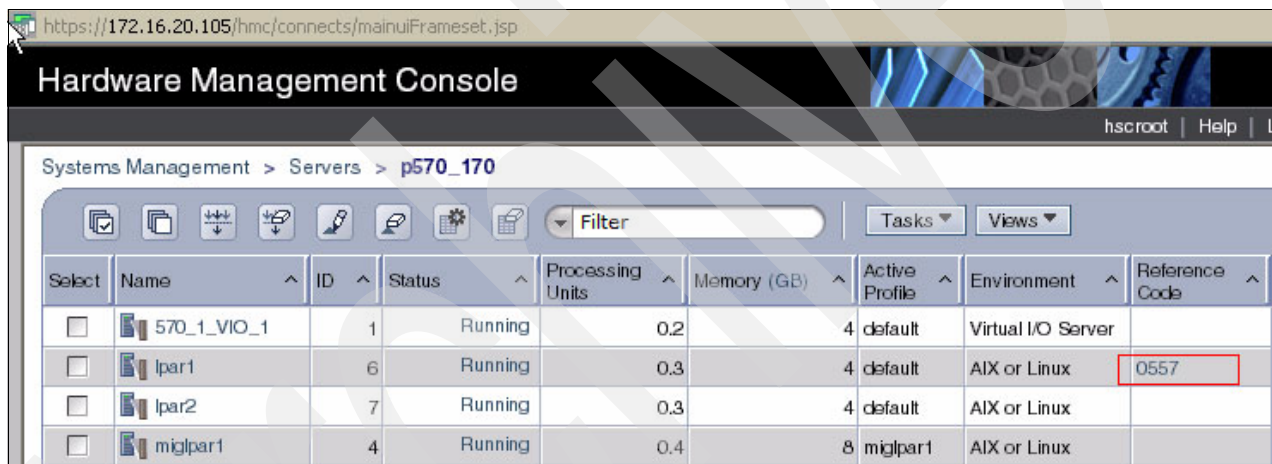


Figure A-1 Selecting a managed server

4. Select the LPAR. Click **Tasks** → **Dynamic Logical Partitioning** → **Virtual Adapters**. See Figure A-2.

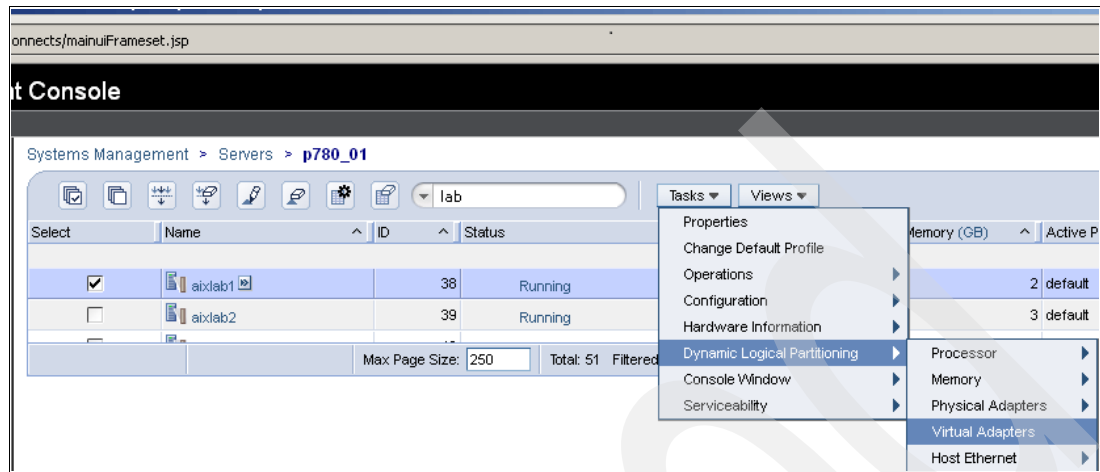


Figure A-2 Selecting virtual adapters

5. Highlight the adapter. Click **Actions** → **Delete** → **Confirm when prompted**, as shown in Figure A-3.

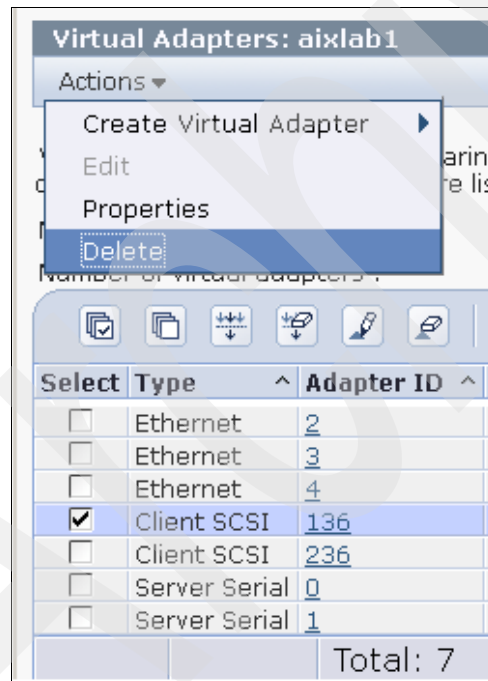


Figure A-3 Dynamic LPAR operation to delete an adapter

6. You can run the **cfgmgr** and **lsdev** commands again to confirm that the adapter is no longer available.

Setting up Secure Shell keys between two management consoles

In this example, we use the SDMC and the HMC. To be able to send remote commands between an HMC and an SDMC, perform these steps:

1. Log on to the HMC and confirm that you have a connection between the consoles, as shown in Example A-5.

Example A-5 Pinging the HMC for a connection

```
sysadmin@sdmc1:~> ping -c 2 -i 2 172.16.20.109
PING 172.16.20.109 (172.16.20.109) 56(84) bytes of data.
64 bytes from 172.16.20.109: icmp_seq=1 ttl=64 time=0.165 ms
64 bytes from 172.16.20.109: icmp_seq=2 ttl=64 time=0.185 ms

--- 172.16.20.109 ping statistics ---
2 packets transmitted, 2 received, 0% packet loss, time 1999ms
rtt min/avg/max/mdev = 0.165/0.175/0.185/0.010 ms
```

2. Set up Secure Shell (ssh) keys between the HMC and the SDMC. From the source HMC, run the **mkauthkeys** command, as shown in Example A-6.

Remote command: If the destination manager is an HMC, you have to set up remote command by selecting **HMC Management** and then selecting **Set up Remote command**. This step is done by default on the SDMC.

Example A-6 Running the mkauthkeys

```
hscroot@hmc4:~> mkauthkeys --ip 172.16.20.22 -u sysadmin -t rsa
Enter the password for user sysadmin on the remote host 172.16.20.22:
hscroot@hmc4:~>
```

In Example A-6:

- `--ip` is the destination server HMC (the SDMC in this example).
- `--u` is the user for this migration, typically, `sysadmin` or `hscroot`.

This ssh connection is required for remote executions, such as *Remote Live Partition mobility*.

Simple cluster installation

In this section, we show you the PowerHA installation that we performed for this book. For a comprehensive installation of PowerHA, refer to *IBM PowerHA SystemMirror 7.1 for AIX*, SG24-7845.

System planning is a requirement. Figure A-4 on page 361 shows a planning diagram for our environment.

Prior to the installation, read the product installation guide to confirm that you have met all the prerequisites. The prerequisites influence the success or failure of the installation.

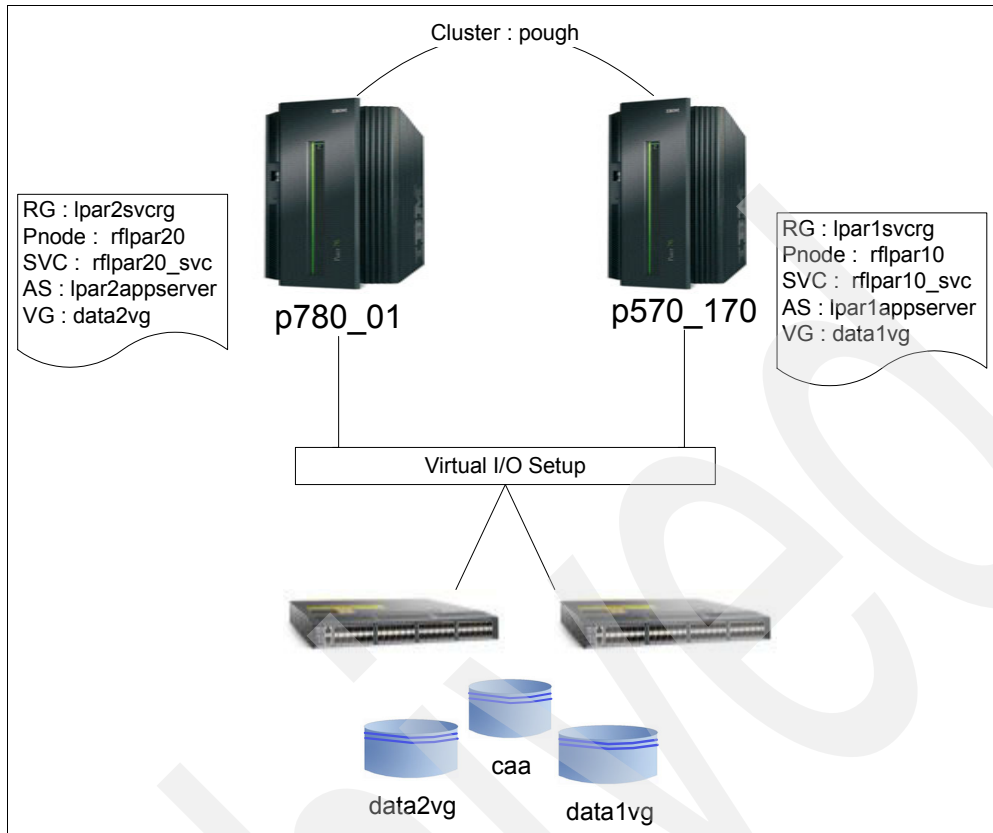


Figure A-4 Simple cluster configuration

We performed the following steps on each client. At this point, you do not have PowerHA configured. Follow these steps:

1. Mount the installation media on the two LPARs that are going to form part of the cluster. The media can be CD ROM or a file system. In Example A-7, we used a Network File System (NFS) mount file system.

Example A-7 Mounting the file system

```
# mount nimres1:/bigfs /mnt
```

2. Run **installp Preview** to see if all prerequisites have been met. Although you can run **installp -agXY -d -p** to confirm the prerequisites, we recommend that you use smitty. Select **Smitty install** → **Install and Update Software** → **Install Software**. Then, enter the INPUT device/directory for the software. You can change the directory to the installation directory, but it is not necessary.

Under the input directory, specify the directory where the installation media is located. In our example, we used /mnt/HA71. This action opens an installation window, as shown in Figure A-5 on page 362.

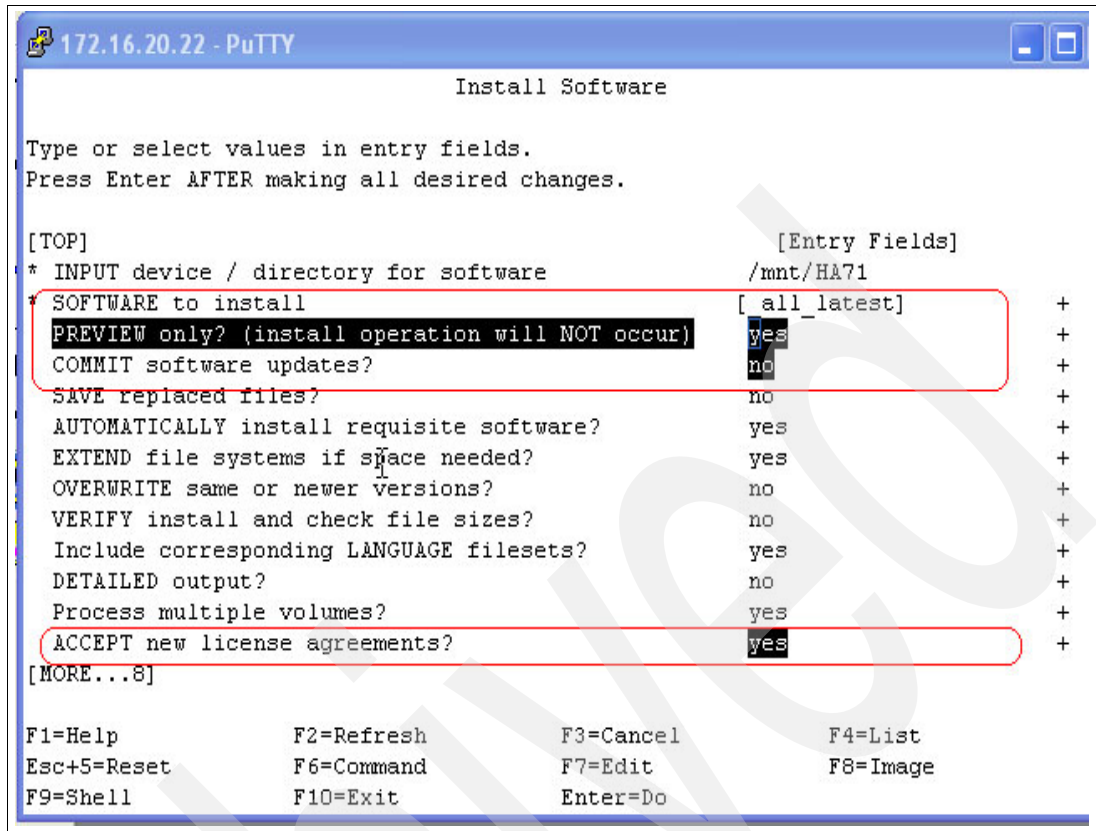


Figure A-5 Installation selection menu

3. The following selections are possible:

- **SOFTWARE to install**

Press F4 to manually select the product that you want to install, as shown in Figure A-6 on page 363.

- **PREVIEW only? (install operation will NOT occur)**

To perform a preview only, select **yes**.

- **COMMIT software updates?**

Select **no**. Committed installations cannot be rolled back using the smitty reject fast path. All new installations have an auto-commit.

- **ACCEPT new license agreements?**

Select **yes**.

Select **Continue** when prompted and check for any preview errors. If the result is OK, change the Preview only field to **no** and press Enter to continue. Follow the same procedure for all of the LPARs that participate in the cluster.

Updates: Make sure that you install all of the available updates.

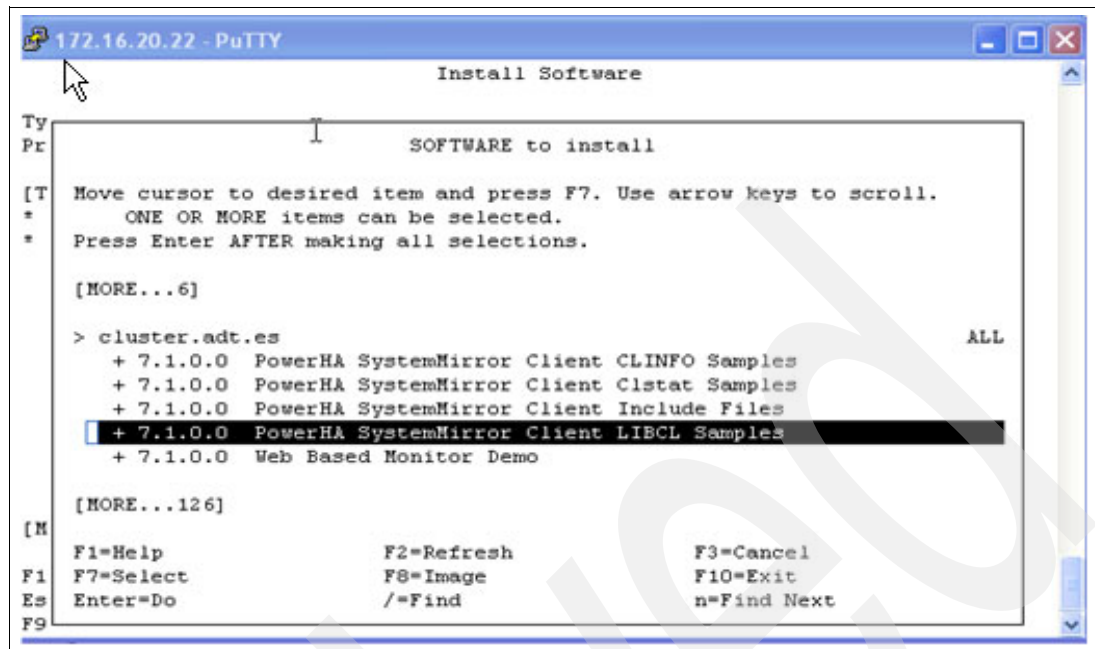


Figure A-6 Manual installation product selection

Installing and configuring PowerHA

We show the initial setup of a cluster. PowerHA 7.1 depends on Cluster Aware AIX (CAA). You do not have to set up the CAA cluster up front, because this task can be done by PowerHA when creating a cluster. We use PowerHA 7.1 terminology. Refer to *IBM PowerHA SystemMirror 7.1 for AIX*, SG24-7845, for list of the changes from the previous versions of PowerHA.

If you have never set up PowerHA, using the typical setup helps make the configuration simple. You can use the custom setup after the cluster is configured to make custom changes. Follow this procedure:

1. Plan and prepare the cluster for communication.

In this step, you prepare your LPARs for the cluster. The book *IBM PowerHA SystemMirror 7.1 for AIX*, SG24-7845 explains the options in this section in detail. You have the following options:

- Planning the cluster topology, resource groups, and resources
- Preparing networks and storage
- Preparing CAA repository disks
- Preparing the rhost file for communication between the LPARs

2. Set up the initial cluster. Type `smitty sysmirror`. Click **Cluster Nodes and Networks** → **Initial Cluster Setup (Typical)**. This path takes you to the initial step-by-step setup of a cluster. The PowerHA 7.1 menu was designed for ease of use. If you follow the menu items, as shown in Example A-8 on page 364 sequentially, you create a simple yet complete cluster. We follow the sequence logically in this example.

Example A-8 Initial cluster setup example

Initial Cluster Setup (Typical)

Move cursor to desired item and press Enter.

Setup a Cluster, Nodes and Networks
Define Repository Disk and Cluster IP Address

What are the repository disk and cluster IP address ?

F9=Shell F10=Exit Enter=Do

3. Select **Setup a Cluster, Nodes and Networks** and enter the initial details. Our details differ from the details in your environment. Refer to Example A-9.

Example A-9 Setting up the cluster and nodes

Setup a Cluster, Nodes and Networks

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

Cluster Name
Currently Configured Node(s)

[Entry Fields]
pough
rflpar10 rflpar20

F9=Shell F10=Exit Enter=Do

4. We then set up CAA. Select **Define Repository and Cluster IP Address**. Notice that hdisk2 is selected for the repository disk in Example A-10.

Example A-10 Defining CAA parameters

Define Repository and Cluster IP Address

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

* Cluster Name
* Repository Disk
Cluster IP Address

[Entry Fields]
pough
[hdisk2] +
[]

F9=Shell F10=Exit Enter=Do

5. Example A-11 shows the disk that was created for the repository, and it shows that hdisk2 has changed to caa_private:

caa_private0	00f69af6dbccc5ed	caavg_private	active
--------------	------------------	---------------	--------

Example A-11 Repository disk setup

# lspv			
hdisk0	00c1f170c2c44e75	rootvg	active
hdisk1	00f69af6dbccc57f	None	
caa_private0	00f69af6dbccc5ed	caavg_private	active
hdisk3	00c1f170c59a0939	None	

Example A-12 shows the multicast address.

Example A-12 Multicast address

Define Repository Disk and Cluster IP Address

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

Cluster Name
Repository Disk
Cluster IP Address

[Entry Fields]
pough
caa_private0
228.16.21.36

F1=Help
Esc+5=Reset
F9=Shell

F2=Refresh
F6=Command
F10=Exit

F3=Cancel
F7=Edit
Enter=Do

F4=List
F8=Image

6. At this point, you can synchronize the cluster. But first, in this example, we also add persistent IP addresses before synchronizing the cluster. On the Cluster, Nodes and Networks menu, select **Manage Nodes** → **Configure Persistent Node IP Label/Addresses** → **Add a Persistent Node IP Label/Address**. Select a node and press Enter. Figure A-7 shows a Configure Persistent Node IP Label/Addresses window.

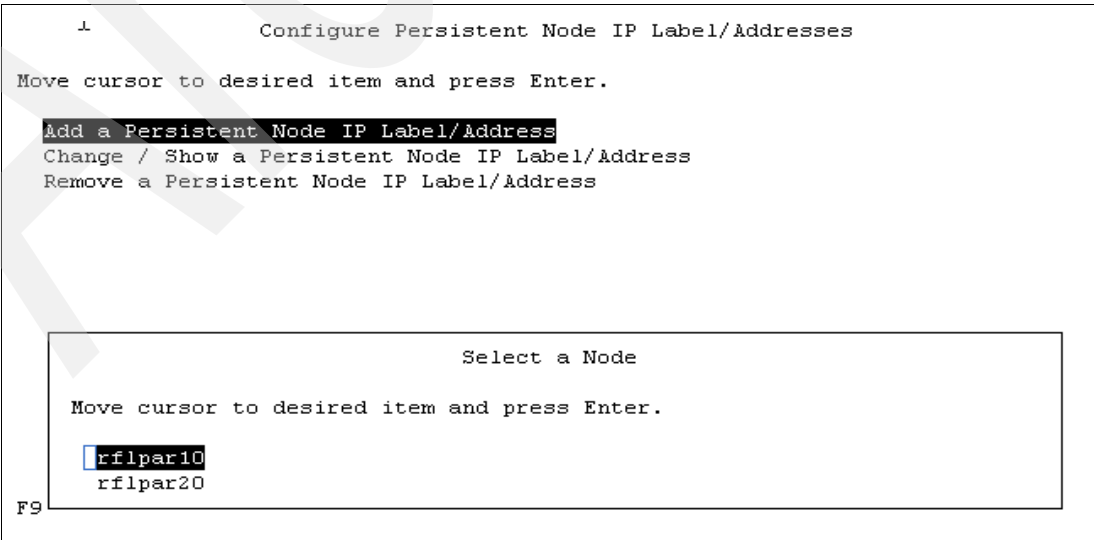


Figure A-7 Node selection dialog

7. Enter the necessary details, as shown in Figure A-8.

```

                                Add a Persistent Node IP Label/Address

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

* Node Name                                [Entry Fields]
* Network Name                            rflpar10
* Node IP Label/Address                    [net_ether_01] +
                                           [rflpar10_prs] +
                                           [255.255.252.0]
                                           Netmask(IPv4)/Prefix Length(IPv6)

F9=Shell      F10=Exit      Enter=Do
```

Figure A-8 Setting up the persistent IP addresses for nodes

Smitty: Where possible, use the smitty selection feature to avoid typing values. A plus sign (+) next to a text box indicates that the available options that can be selected.

8. At this point, we verify the cluster. After the cluster verification is *successful*, you can start the cluster. After this stage, most configurations can use the dynamic cluster reconfiguration DARE facility.

9. You can use the **clstart** shortcut to get to the Start Cluster Services window, which is shown in Example A-13. Most of the details in the window that is shown in the example are selected with the F4 or ESC+4 selection facility.

Example A-13 Starting the cluster services

Start Cluster Services

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

* Start now, on system restart or both
Start Cluster Services on these nodes

* Manage Resource Groups
BROADCAST message at startup?
Startup Cluster Information Daemon?
Ignore verification errors?
Automatically correct errors found during
cluster start?

[Entry Fields]

now
[rflpar10,rflpar20]
Automatically
true
true
false
Interactively

+

+

+

+

+

+

+

F1=Help
Esc+5=Reset
F9=Shell

F2=Refresh
F6=Command
F10=Exit

F3=Cancel
F7=Edit
Enter=Do

F4=List
F8=Image

10. On a successful startup, you can observe the following characteristics:
- An OK status, as shown in Example A-14 on page 368.
 - The Persistent IP address, as shown in Example A-15 on page 368.
 - Startup details in the `/var/hacmp/log/hacmp.out` file.
 - The **clstat** utility must show the status of the nodes (Example A-16 on page 369).

Example A-14 Successful cluster startup

COMMAND STATUS

Command: OK stdout: yes stderr: no

Before command completion, additional instructions may appear below.

[TOP]

WARNING: Multiple communication interfaces are recommended for networks that use IP aliasing in order to prevent the communication interface from becoming a single point of failure. There are fewer than the recommended number of communication interfaces defined on the following node(s) for the given network(s):

Node:	Network:
-----	-----

WARNING: Network option "routerevalidate" is set to 0 on the following nodes:
[MORE...128]

F1=Help	F2=Refresh	F3=Cancel	F6=Command
F8=Image	F9=Shell	F10=Exit	/=Find
n=Find Next			

Example A-15 Persistent IP address

```
# ifconfig -a
en0:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64
BIT,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
    inet 172.16.21.36 netmask 0xfffffc00 broadcast 172.16.23.255
    inet 172.16.21.40 netmask 0xfffffc00 broadcast 172.16.23.255
    tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
lo0:
flags=e08084b,c0<UP,BROADCAST,LOOPBACK,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT,
LARGESEND,CHAIN>
    inet 127.0.0.1 netmask 0xff000000 broadcast 127.255.255.255
    inet6 ::1%1/0
    tcp_sendspace 131072 tcp_recvspace 131072 rfc1323 1
```

Example A-16 Clstat facility to show cluster status

```
clstat - PowerHA SystemMirror Cluster Status Monitor
-----

Cluster: pough (1111817142)
Thu Jun  2 10:34:18 EDT 2011
      State: UP                Nodes: 2
      SubState: STABLE

      Node: rflpar10           State: UP
      Interface: rflpar10 (0)   Address: 172.16.21.36
                                State:  UP

      Node: rflpar20           State: UP
      Interface: rflpar20 (0)   Address: 172.16.21.35
                                State:  UP

***** f/forward, b/back, r/refresh, q/quit
*****
```

Setting up resources

In this example, we use mutual takeover. The IP address and volume groups are failed over between two LPARs. We created two resource groups with the following information:

- ▶ Resource group poures1:
 - Volume group data1vg
 - Service Address rflpar10_svc
- ▶ Resource group poures2:
 - Volume group data2vg
 - Service Address rflpar20_svc

Plan and identify all resource components that are used in the cluster. The steps that follow are valid for our example. You might choose other options. Refer to *IBM PowerHA SystemMirror 7.1 for AIX, SG24-7845*, for possible configurations. Follow these steps:

1. Set up service IP. Type `smitty sysmirror`. Select **Cluster Applications and Resources** → **Resources** → **Configure Service IP Labels/Addresses** → **Add a Service IP Label/Address** → **Select Network** · **net_ether_01 (172.16.20.0/22)** → Select required service address and press Enter, as shown in Figure A-9.

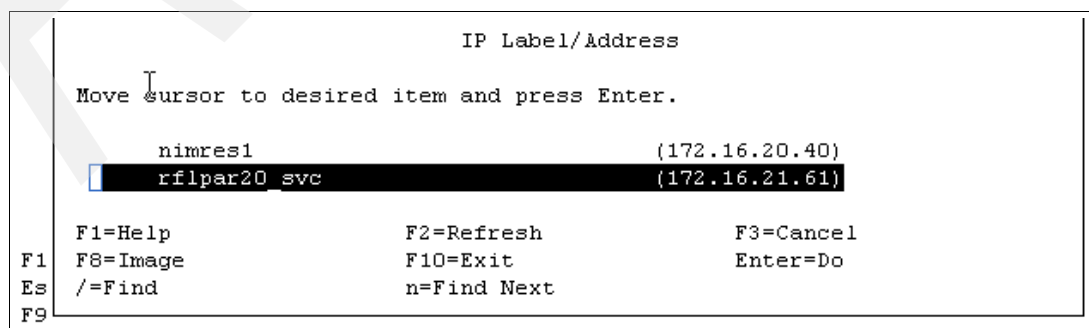


Figure A-9 Service IP address selection

2. Set up application control scripts. This step was called application servers in previous versions. Select **Cluster Applications and Resources** → **Resources** → **Configure User Applications (Scripts and Monitors)** → **Application Controller Scripts** → **Add Application Controller Scripts**. See Example A-17. The scripts must exist on both nodes with the execute bit set. You must test these scripts before you include them in PowerHA.

Example A-17 Start and stop scripts for application servers

Add Application Controller Scripts

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

* Application Contoller Name

* Start Script

* Stop Script

Application Monitor Name(s)

[Entry Fields]

[lparlappserver]

[/hascripts/startlpar1>

[/hascripts/stoplpar1.>

+

F1=Help

F2=Refresh

F3=Cancel

F4=List

Esc+5=Reset

F6=Command

F7=Edit

F8=Image

F9=Shell

F10=Exit

Enter=Do

3. Create the resource groups and include the application controller and service IP address in the resource groups. This resource group becomes a unit that is controlled by PowerHA. To add a resource group, select **Cluster Applications and Resources** → **Resource Groups** → **Add a Resource Group**. See Figure A-10. The book, *IBM PowerHA SystemMirror 7.1 for AIX*, SG24-7845, explains the startup, fallover, and fallback policies in detail.

Add a Resource Group

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

* Resource Group Name

* Participating Nodes (Default Node Priority)

Startup Policy

Fallover Policy

Fallback Policy

[Entry Fields]

[lparlsvcrg]

[rflpar10 rflpar20]

+

Online On Home Node O>

Fallover To Next Prio>

Never Fallback

+

F9=Shell

F10=Exit

Enter=Do

Figure A-10 Adding a resource group

4. Populate the resource group with the highly available resources. Select **Cluster Applications and Resources → Resource Groups → Change/Show Resources and Attributes for a Resource Group**. Select the Resource Group Name from the pop-up menu. Add the Application Controller, associated Volume Groups, and Service IP Labels/Addresses, as shown in Figure A-11.

Change/Show All Resources and Attributes for a Resource Group

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

[TOP]	[Entry Fields]
Resource Group Name	lpar1svcrg
Participating Nodes (Default Node Priority)	rflpar10 rflpar20
Startup Policy	Online On Home Node O>
Fallover Policy	Fallover To Next Prio>
Fallback Policy	Never Fallback
Service IP Labels/Addresses	[rflpar10_svc] +
Application Controllers	[lpar1appserver] +
Volume Groups	[data1vg] +
Use forced varyon of volume groups, if necessary	false +
Automatically Import Volume Groups	false +
[MORE...24]	

F1=Help

F2=Refresh

F3=Cancel

F4=List

Esc+5=Reset

F6=Command

F7=Edit

F8=Image

F9=Shell

F10=Exit

Enter=Do

Figure A-11 Populating a resource group

5. Verify and synchronize the cluster configuration. Notice that the clstat - PowerHA SystemMirror Cluster Status Monitor window changes. Compare the clstat in Figure A-12 on page 372 and the previous clstat in Example A-16 on page 369. Notice the availability of the resource group.

Figure A-12 *clstat* with resource groups

- ▶ The **lspv** command shows that the volume groups are varied on in concurrent mode.
- ▶ The **lsvg** command on the two-cluster volume groups shows the logical volumes in each volume group.
- ▶ The **data2vg** command shows closed logical volumes, because they are open on the primary node.
- ▶ The **df** command does not show the data2vg file systems.

```
# lspv
hdisk0          00f69af6dbccc621          rootvg          active
hdisk1          00f69af6dbccc57f          data1vg         concurrent
caa_private0    00f69af6dbccc5ed          caavg_private   active
hdisk3          00c1f170c59a0939          data2vg         concurrent
# lsvg -l data1vg
data1vg:
LV NAME          TYPE          LPs          PPs          PVs  LV STATE      MOUNT POINT
db2bin           jfs2          96           96           1    open/syncd    /opt/IBM/db2
loglv00          jfs2log       1            1            1    open/syncd    N/A
data1lv          jfs2          1            1            1    open/syncd    /userdata/dat1
data1lg          jfs2          1            1            1    open/syncd    /userdata/dblog
data1tmp         jfs2          1            1            1    open/syncd    /userdata/dbtmp
swlv             jfs2          96           96           1    open/syncd    /somesoftware
db2rep           jfs2          96           96           1    open/syncd    /db2repos
# lsvg -l data2vg
data2vg:
LV NAME          TYPE          LPs          PPs          PVs  LV STATE      MOUNT POINT
```



```

data2lv          jfs2          1          1          1  closed/syncd  N/A
loglv01          jfs2log       1          1          1  closed/syncd  N/A
# df
Filesystem      512-blocks      Free %Used      Iused %Iused Mounted on
/dev/hd4         458752         31936  94%       10793   69% /
/dev/hd2         5668864        11904 100%       48285   88% /usr
/dev/hd9var       786432         87992  89%        6007   36% /var
/dev/hd3         262144        256208   3%         144    1% /tmp
/dev/hd1         32768         32064   3%          5    1% /home
/dev/hd11admin   262144        261384   1%          5    1% /admin
/proc            -              -      -          -    - /proc
/dev/hd10opt     720896        345592  53%       6986   16% /opt
/dev/livedump    524288        523552   1%          4    1% /var/adm/ras/livedump
/aha             -              -      -          43    1% /aha
/dev/fslv00      524288        509552   3%         17    1% /clrepos_privatel
/dev/db2rep      6291456       5850240   8%          59    1% /db2repos
/dev/db2bin      6291456       3273864  48%       8381    3% /opt/IBM/db2
/dev/swlv        6291456       3182352  50%       4519    2% /somesoftware
/dev/data1lv     65536         64864   2%          4    1% /userdata/dat1
/dev/data1lg     65536         64864   2%          4    1% /userdata/dblog
/dev/data1tmp    65536         64864   2%          4    1% /userdata/dbtmp
#

```

Archived

Performance concepts

This appendix provides an overview of systems performance concepts, benchmarks, and specific POWER7 Enterprise Server performance data.

We discuss the following topics:

- ▶ Performance concepts
- ▶ Throughput versus response time
- ▶ Performance and computing resources

Performance concepts

One of the key tasks of IT Architects and solution designers is to develop “design-for-performance” as part of the overall system design. This type of design aims to produce a solution that meets particular performance measurements, for example, to enable the solution or the system to deliver a particular quality of service (QoS). Therefore, system performance can be defined as “The ability of a system to meet a particular QoS”.

However, the measure of the QoS varies depending on the type of the workload, and the role of the person defining or measuring the QoS. Sometimes, QoS can be qualitative as opposed to being quantitative.

The QoS can mean user response time, throughput, system utilization, or returning to operations. QoS can also mean on broader terms, system availability or power efficiency.

Performance measurement is usually defined from one of two perspectives:

- ▶ **System perspective:** This perspective is typically based on throughput, which is the average of items, such as transactions or processes per particular measured unit of time, and utilization, which is the percentage of time that a particular resource is busy.
- ▶ **User perspective:** This perspective is typically based on response time, which is the average elapsed time from the initiation of the task to the point where the user or the application receives the first response. The response time often is seen as a critical aspect of performance because of its potential visibility to users or customers.

An example is a computer system that is running a web server with an online store. The response time here is the elapsed time between clicking the submit button to place an order, and the beginning of receiving the order confirmation.

Throughput versus response time

Throughput and response time are related. In many cases, a higher throughput comes at the cost of poorer response time or slower response time. Better response time comes at the cost of lower throughput. We often see response time graphs, such as the graph that is shown in Figure B-1 on page 377, which represents the right half of an inverse parabola.

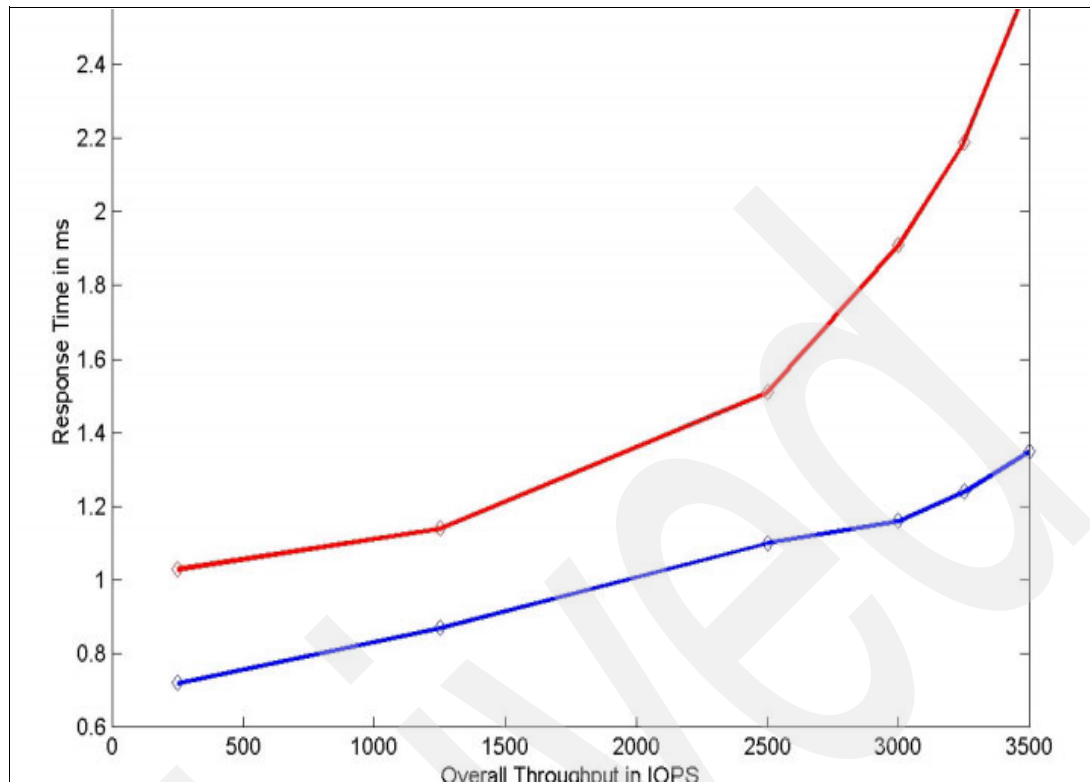


Figure B-1 Example of a typical response time versus throughput graph

For example, we can compare the performance of a 32-bit but relatively faster CPU (higher GHz) to a 64-bit but relatively slower speed. The 32-bit relatively faster CPU provides better response time but with less throughput than the 64-bit relatively slower CPU. The response time, or the user perspective performance measurement, can be more qualitative than quantitative, and it can vary considerably, because there are many variables involved.

It is critical to the performance modeler to precisely define what is measured in what configuration and in what circumstances. The output is usually expressed in average values. Benchmarks have been developed to address this issue. We discuss benchmarks in the next section.

Performance and computing resources

In this section, we discuss the computing resources that affect the entire server performance:

- ▶ CPU architecture
- ▶ Multi-core architecture
- ▶ Memory architecture
- ▶ I/O storage architecture
- ▶ I/O networking architecture

Central processing unit

The CPU has a central role in the performance of a computer system. The following CPU architecture parameters are significant in defining the server performance:

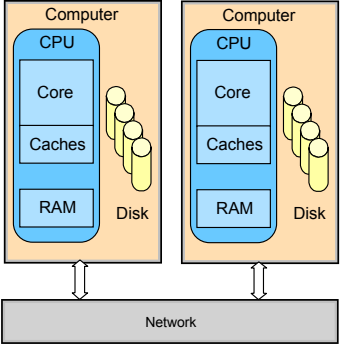
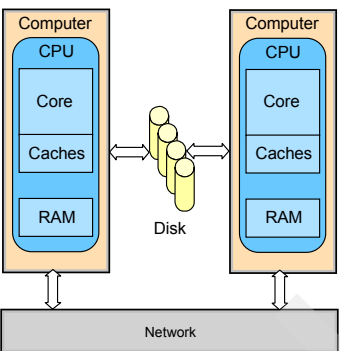
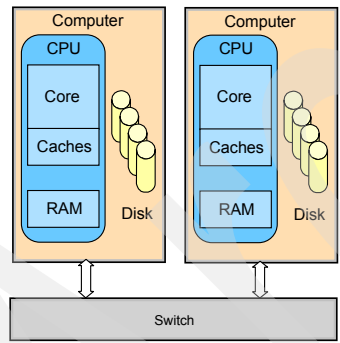
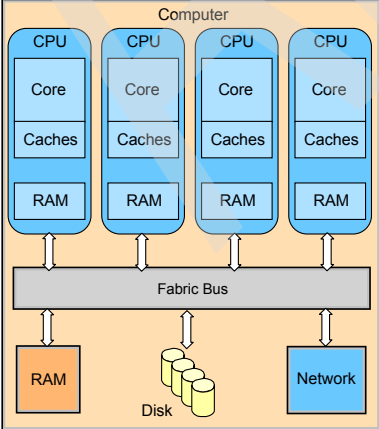
- ▶ The CPU clock frequency: A *CPU cycle* is the interval of time that is needed to perform one operation. Modern CPUs can perform multiple operations during one CPU cycle. A CPU cycle is not equal to a CPU instruction, which often requires multiple CPU cycles to complete.
- ▶ CPU instruction-set architecture: A *CPU instruction* is a machine or assembly language instruction that specifies the operation that is given to a CPU. Examples for CPU instructions are load, add, and store. IBM Power Systems use the Reduced Instruction Set Computing (RISC) instruction-set technology, which usually has fixed lengths, requires more-or-less fixed CPU cycles to execute, and uses a high number of CPU registers. Other processors are based on the Complex Instruction Set Computing technology (CISC), where the CPU instructions are more complex than others and therefore require more CPU cycles to complete.
- ▶ The path length: The number of instructions that it takes to complete a certain task.
- ▶ Multithreading: Cache misses can delay the execution of instructions in a processor for many cycles during which no other instruction can be executed. Multithreading addresses this issue by simultaneously holding the state of two or more threads. When one thread becomes stalled, due to a cache miss for example, the processor switches to the state and attempts to execute the instructions of another thread. Hardware multithreading originally was introduced with the models M80 and p680. Newer processors, such as POWER5, POWER6, and POWER7, support Simultaneous Multithreading (SMT), which allows both hardware threads to execute instructions at the same time. The POWER5 and POWER6 cores support single thread mode (ST) and simultaneous multithreading with two SMT threads (SMT2). Each SMT thread is represented as a logical CPU in AIX. When running in SMT2 mode, a system with a single POWER5 or POWER6 core has two logical CPUs. The POWER7 core supports single thread mode (ST) and simultaneous multithreading with two SMT threads (SMT2) and four SMT threads (SMT4). When running in SMT4 mode, a system with a single POWER7 core has four logical CPUs. To fully benefit from the throughput improvement of SMT, the applications need to use all the SMT threads of the processors.
- ▶ Processor virtualization: In a virtualized environment, physical processors are represented as virtual processors that can be shared across multiple partitions. The hypervisor assigns physical processors to shared partitions (SPLPARs), which are also known as micro-partitions, based on the capacity configurations and resource consumption of the partitions. Virtual processor management, which is also known as *processor folding*, dynamically increases and reduces the number of virtual processors of a shared partition based on the instantaneous load of the partition. Many workloads benefit from virtual processor management due to its higher degree of processor affinity.

Multiple core systems

Multiple processor systems are more complex than single processor systems, because access to shared resources, such as memory, needs to be serialized, and the data needs to be kept synchronized across the caches of the individual CPUs. It is important to note that the server performance is not directly proportional to the number of CPUs. For example, the performance of the system does not increase linearly with the number of CPUs.

There are various types of multiple CPU systems, as shown in Table B-1 on page 379.

Table B-1 Types of multiple CPU systems

	<p>Cluster (Shared Nothing)</p> <ul style="list-style-type: none"> ▶ Each processor is a stand-alone machine. ▶ Each processor has its own copy of the operating system. ▶ No resources shared communication through the network.
	<p>Share Disk MP</p> <ul style="list-style-type: none"> ▶ Each processor has its own memory and cache. ▶ Each processor has its own copy of the operating system. ▶ Processors run in parallel. ▶ Processors share disks. ▶ Communication through network.
	<p>Shared Memory Cluster (SMC)</p> <ul style="list-style-type: none"> ▶ All processors are in a shared memory cluster. ▶ Each processor has its own resources. ▶ Each processor has its own copy of the operating system. ▶ Processors are tightly coupled. ▶ Connected through switch.
	<ul style="list-style-type: none"> ▶ All processors are tightly coupled; all processors are inside the same box with a high-speed bus or switch. ▶ Processors share memory, disks, and I/O devices. ▶ There is one copy of the operating system. ▶ Multi-threaded operating system.

Memory architecture

The memory of a computer system is divided into several layers of various speeds and sizes. Typically, faster memory is more expensive to implement and, therefore, smaller in size.

Figure B-2 is a high-level representation of the memory hierarchy based on the location of the memory in a computer system. The implementation of the individual memory layer might differ depending on the implementation of a specific architecture.

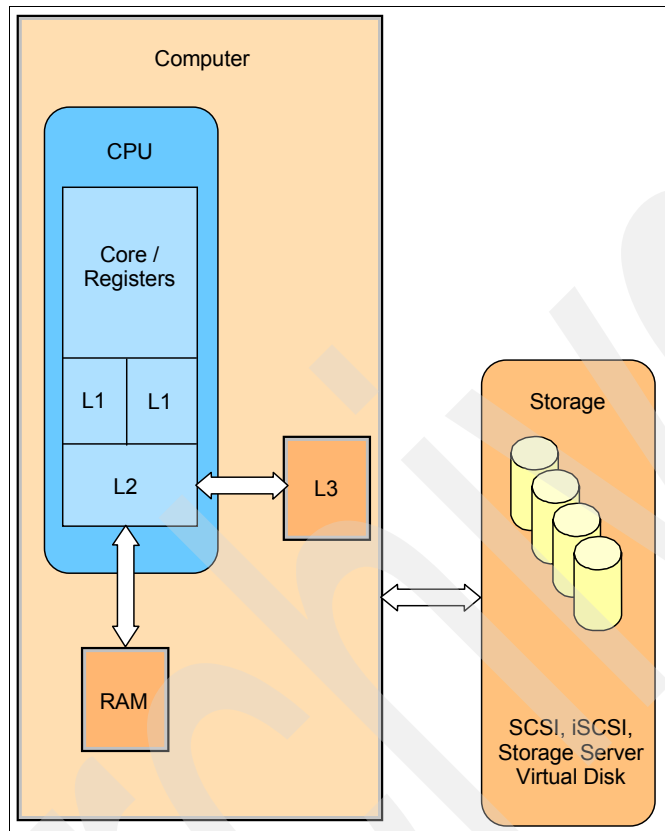


Figure B-2 Memory hierarchy

Registers

CPU registers are the fastest memory and are the top layer of the memory hierarchy of a computer system. A CPU has a limited number of registers, which can be used for integer and floating-point operations.

Caches

Modern processors have multiple layers of caches. The fastest cache that is closest to the registers is the Level 1 cache. It is the smallest cache and is often divided into a Level 1 instruction and Level 1 data cache.

The next level of cache is the Level 2 cache, which often holds instructions and data. The Level 2 cache has higher access latency than the Level 1 cache, but it has the advantage that it can be several megabytes in size.

Certain processors have a third level of cache, which either can be on the same chip as the processor or external; in the later case, the processor has a cache controller.

Cache coherency

Cache coherency becomes an important factor in symmetric multiprocessor (SMP) systems when each processor has its own cache. A coherency problem can occur when two or more processors have a copy of the same data in their caches. To keep the data consistent, a processor uses snooping logic to broadcast a message over the bus each time that its cache has been modified. When a processor receives a message from another processor and detects that another processor changed a value for an address that exists in its own cache, it invalidates its own copy of the data, which is called *cross-invalidate*.

Cross invalidate and snooping affect the performance and scalability of SMT systems due to the increased number of cache misses and increased bus traffic.

Random access memory

The next level in the memory hierarchy is the random access memory (RAM). It is much slower than the caches but also much cheaper to produce. The size of the RAM in a computer system can vary from several hundred megabytes on a small workstation to several terabytes on high-end servers. A processor accesses RAM either through integrated memory controllers or through bus systems, which connect it to an external memory controller.

Virtual memory is a method that allows the operating system to address more memory than a computer system actually has in real memory. Virtual memory consists of real memory and physical disk space that is used for working storage and file pages. On AIX, virtual memory is managed by the Virtual Memory Manager (VMM).

VMM virtual segments

The AIX virtual memory is partitioned into virtual segments. Each virtual segment is a continuous address space of 256 MB (default segment size) or 1 TB (super segment) and further divided into pages. Pages can have multiple sizes and are not necessarily contiguous in physical memory.

The VMM virtual segment type defines the type of pages for which the segment is being used, for example, for working pages or file pages. Table B-2 lists the most commonly used VMM virtual segment types.

Table B-2 VMM virtual segment types

Segment type	Purpose
Computational	Processes private segments Shared segments Paging space
Client	Enhanced journaled file system 2 (JFS2) file and executables Network File System (NFS) files and executables CD-ROM, DVD file system Compressed JFS files and executables
Persistent	JFS files and executables

Real memory is divided into page frames. The page frame size depends on the version of AIX and the platform on which it is running. On existing systems that do not support page sizes larger than 4 KB, the real memory is divided into 4 KB frames. Platforms and AIX versions that support larger page sizes divide the memory into frames with multiple page sizes.

AIX 5.3 and later dynamically manage pools of 4 KB and 64 KB page sizes. Starting with AIX 6.1 on POWER6, individual segments can have 4 KB and 64 KB page sizes.

Address translation

Applications that are running on AIX, 32-bit or 64-bit, have their own address space starting from address 0 to the highest possible address. Shared segments, such as the shared library segment, are mapped into the address space of the application.

When an application accesses memory, the effective address that is used by the application is translated into a real memory address. The effective-to-real-address translation is done by the processor, which maintains an effective-to-real-address (ERAT) translation table. When a processor does not have the necessary information for the address translation, it tries to walk the page frame table and access the translation lookaside buffer (TLB) first.

Memory affinity

Memory affinity is an approach to allocate memory that is closest to the processor on which a process caused a page fault. The AIX memory affinity support allows user memory allocation in a first-touch or round-robin (default) scheduling policy. The scheduling policy can be specified for individual memory types, such as data, mapped files, shared memory, stack, text, and unmapped files.

An efficient use of memory affinity requires an appropriate degree of processor affinity to assure that application threads that are interrupted are re-dispatched to the processors from which their memory was allocated.

Processor affinity

The goal of *processor affinity* is to reduce the number of cache misses by re-dispatching an interrupted thread to the same processor on which it previously was running. The efficiency of processor affinity mainly depends on the contents of the processor's cache. In the best case, the processor's cache contains sufficient data from the thread, and the thread's execution can continue without any waits to resolve cache misses. In the worst case, the processor's cache has been depleted, and the thread will experience a series of cache misses.

Server I/O storage

Figure B-3 on page 383 demonstrates the various I/O paths that applications can use when accessing data that is located on a local storage device.

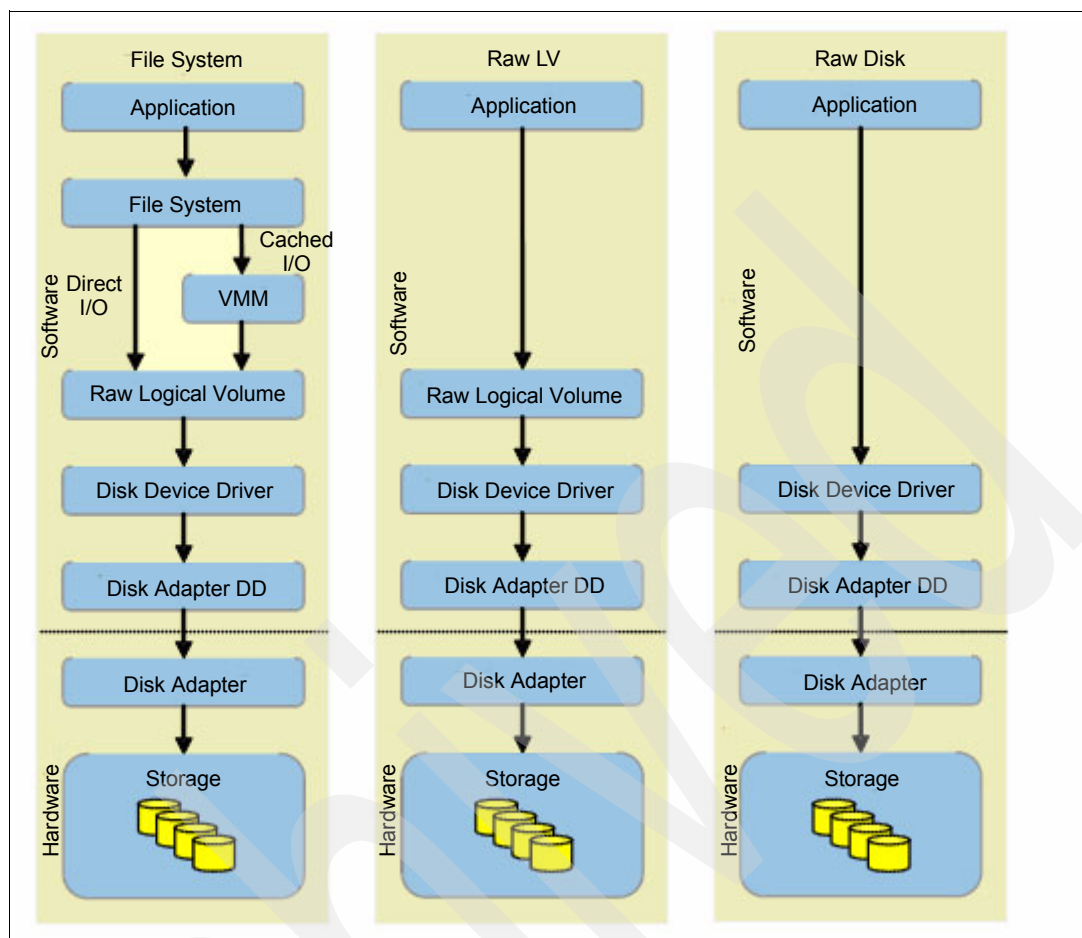


Figure B-3 Types of I/O paths

The most commonly used I/O path is the file system I/O where applications read or write the data that is managed by the file system. Applications can specify through the open flags whether the data of a file must be cached in VMM (default) or directly accessed bypassing VMM. Refer to Table B-3.

Table B-3 File access modes

File access mode	Description
Non-synchronous I/O	Regular cached I/O (default unless specified otherwise); data is flushed out to disk through write behind or syncd; the file system reads pages into memory ahead of time when the sequential read access pattern is determined.
Synchronous I/O	Cached I/O and writes to files do not return until the data has been written to disk; the file system reads pages into memory ahead of time when the sequential read access pattern is determined.
Direct I/O	Regular cached I/O (default unless specified otherwise); data is flushed out to disk through write behind or syncd; the file system reads pages into memory ahead of time when the sequential read access pattern is determined.
Concurrent I/O	Same as direct I/O but without inode lock serialization.
Asynchronous I/O	I/O is serviced asynchronously by the AIX kernel subsystem.

Certain applications, typically database applications, bypass the file system and VMM layers and access the logical volumes directly. Bypassing the file system and VMM layers usually is done to improve performance by reducing the path length.

Applications can bypass Logical Volume Manager (LVM) altogether by accessing the raw disks directly. Similar to raw logical volumes, this method typically is done to improve performance by reducing the path length.

Similar to the I/O path for local storage, the data of a Network File System (NFS) can be cached by VMM (default) or accessed without caching by using the direct I/O mount option. The concurrent I/O option can also be used, which results in access similar to direct I/O, but without the rnode lock serialization. Any operation on an NFS is handled by the NFS client that communicates with the NFS server using the User Datagram Protocol (UDP) or TCP network protocol.

The server networking I/O

Figure B-4 demonstrates the traditional network hierarchy for computer systems that are using the TCP/IP network protocol as a communication vehicle.

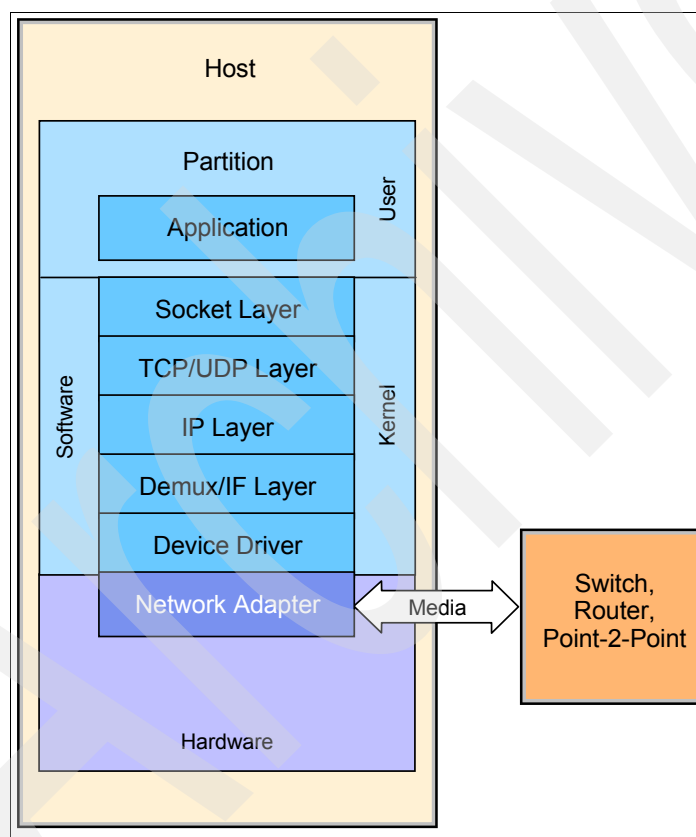


Figure B-4 The traditional network hierarchy of a stand-alone server

Most applications that communicate across networks use the sockets API as the communication channel. A *socket* is an endpoint of a two-way communication channel. When an application creates a socket, it specifies the address family, the socket type, and the protocol. A new socket is “unnamed”, which means that it does not have any association to a local or remote address. In this state, the socket cannot be used for a two-way communication.

In a virtualized environment, physical network adapters are replaced by virtual adapters that communicate with other systems through the hypervisor instead of a physical media. Figure B-5 demonstrates the network hierarchy in a virtualized environment.

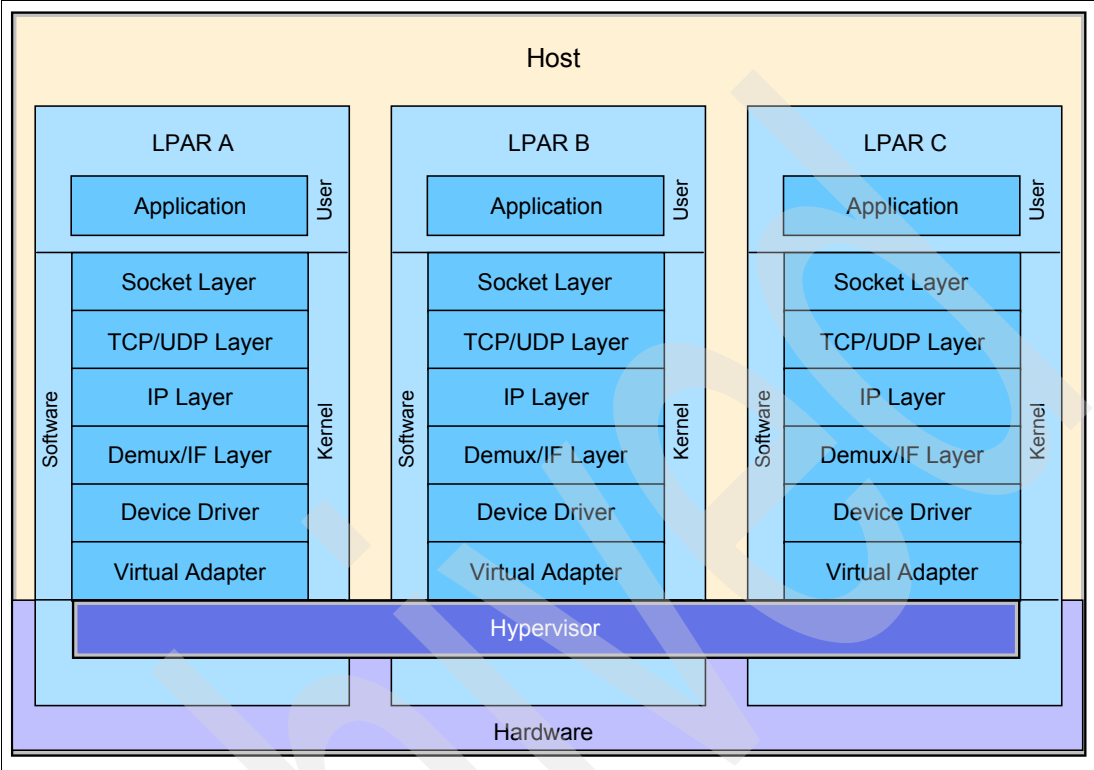


Figure B-5 The network hierarchy in a virtualized environment

Figure B-6 on page 386 demonstrates the Shared Ethernet Adapter (SEA) bridging the virtual LAN (VLAN) for LPAR A and LPAR B to a physical Ethernet.

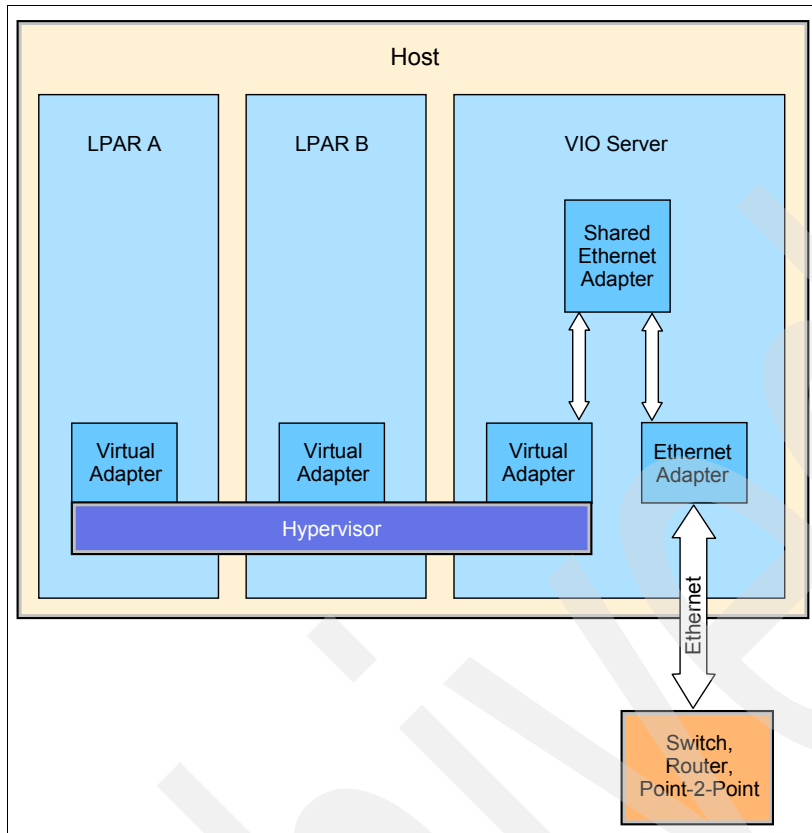


Figure B-6 SEA VLAN bridging

POWER6-based machines provide the Host Ethernet Adapter (HEA) feature that is also known as the Integrated Virtual Ethernet adapter (IVE), which allows the sharing of physical Ethernet adapters across multiple logical partitions (LPARs). An HEA connects directly to the GX+ bus and offers high throughput and low latency.

LPARs connect directly to HEAs and can access external networks through the HEA without going through an SEA or another LPAR.

Figure B-7 on page 387 demonstrates the network hierarchy for LPARs that communicate to an external network through an HEA.

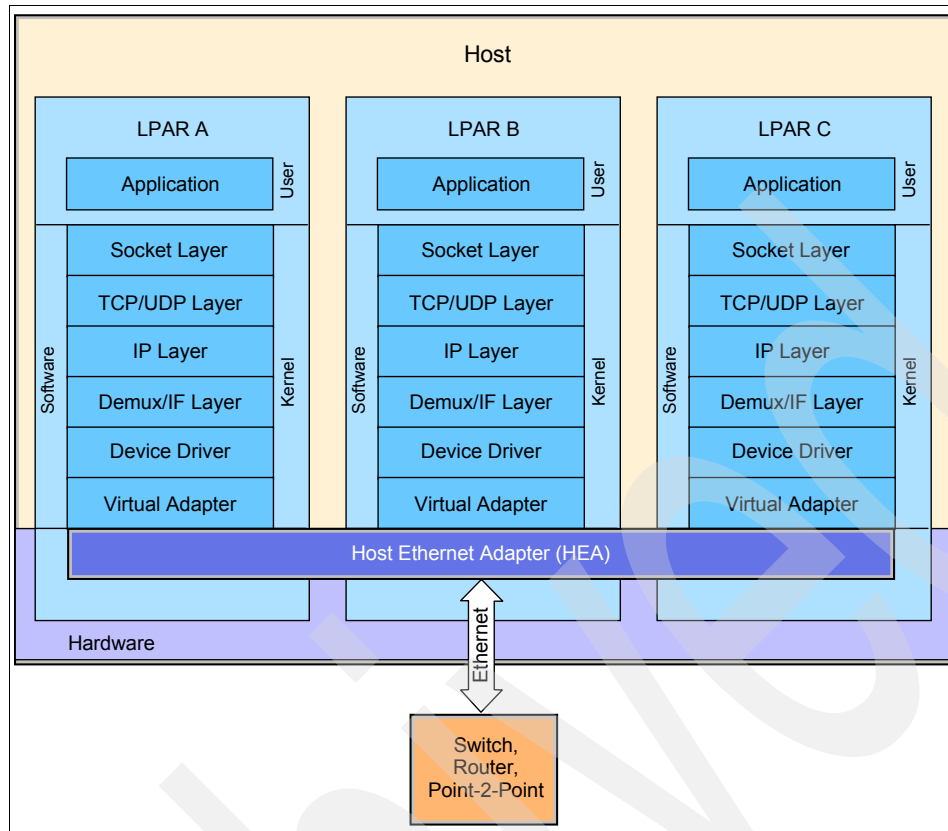


Figure B-7 LPARs' network hierarchy

Performance metrics

We prefer to call the computer performance metrics “*utilization metrics*” because they mainly measure the utilization of a particular resource. We refer to the output that the computer actually delivers as “*performance metrics*”. The utilization metrics of computer systems are mainly measured through the use of analysis tools.

The following high-level overview shows the major performance metrics for various system components:

- ▶ CPU:
 - %user, %system, %idle, and %wait
 - Physical consumed, entitlement
 - Number of context switches, interrupts, and system calls
 - Length of the run queue
- ▶ Memory:
 - Virtual memory paging statistics
 - Paging rate of computational pages
 - Paging rate of file pages
 - Page replacement page scanning and freeing statistics
 - Address translation faults
 - Cache miss rates

- ▶ Disk I/O:
 - Amount of data read/written per second (KB, MB, or GB per second)
 - Transactions per second
 - Elapsed time for I/O to complete
 - Queuing time
- ▶ Network I/O:
 - Amount of data transmitted/received per second (KB or MB per second)
 - Number of packets transmitted/received per second
 - Network CPU utilization
 - Network memory utilization
 - Network statistics, errors, and retransmissions

Performance benchmarks

Performance benchmarks are well defined problems or tests that serve as a basis to evaluate and compare the performance of computer systems. Performance benchmark tests use representative sets of programs and data that are designed to evaluate the performance of computer hardware and software in a particular configuration.

There are industry standard benchmarks, such the TPC and SPEC benchmarks, and non-industry benchmarks, such the IBM rPerf, and the IDEAS' performance metric that is called the Relative Performance Estimate v2 (RPE2).

IBM rPerf

Workloads have shifted over the last eight years, and IBM is committed to providing clients with a relative system performance metric that reflects those changes. IBM publishes the rPerf relative performance metric for the IBM Power Systems family of UNIX servers. This metric replaced ROLTP which was withdrawn.

rPerf is a combination of several measurements of total systems commercial performance that takes into account the demands on a server in today's environment. It is derived from an IBM analytical model, which uses characteristics from IBM internal workloads and Transaction Processing Council (TPC) and Standard Performance Evaluation Corporation (SPEC) benchmarks.

The rPerf model is not intended to represent any specific public benchmark results and must not be reasonably used in that way. The model simulates certain system operations, such as CPU, cache, and memory. However, the model does not simulate disk or network I/O operations.

The IBM eServer™ pSeries® 640 is the baseline reference system and has a value of 1.0. Although rPerf can be used to compare estimated IBM UNIX commercial processing performance, actual system performance might vary and depends on many factors, including system hardware configuration and software design and configuration.

IDEAS' RPE2

RPE2 is a methodology that uses public domain material in conjunction with IDEAS' own research and analysis to calculate a number that represents an estimate of relative performance for a specific processor type/number of processors combination.

SPEC benchmarks

SPEC provides a standardized set of benchmarks to evaluate the performance of the newest generation of high-performance computers. The Standard Performance Evaluation Corporation (SPEC) is a non-profit corporation that was formed to establish, maintain, and endorse a standardized set of relevant benchmarks that can be applied to the newest generation of high-performance computers. SPEC develops benchmark suites and also reviews and publishes submitted results from our member organizations and other benchmark licensees.

SPEC provides a standardized set of benchmarks to evaluate the performance of the newest generation of high-performance computers:

- ▶ SPEC CPU2006 is an industry-standard benchmark that is designed to provide performance measurements that can be used to compare compute-intensive workloads on separate computer systems, SPEC CPU2006 contains two benchmark suites: CINT2006 for measuring and comparing compute-intensive integer performance, and CFP2006 for measuring and comparing compute-intensive floating point performance.

For more information, see this website:

<http://www.spec.org/cpu2006/>

- ▶ SPECpower_ssj2008 is the first industry-standard SPEC benchmark that evaluates the power and performance characteristics of volume server class computers. SPEC has designed SPECpower_ssj2008 to be used as both a benchmark to compare power and performance among various servers and as a toolset to improve server efficiency.

The benchmark workload represents typical server-side Java business applications. The workload is scalable, multi-threaded, and portable across a wide range of operating environments, and economical to run. It exercises the CPUs, caches, memory hierarchy, and scalability of shared memory processors (SMPs), as well as the implementations of the Java virtual machine (JVM), Just-In-Time (JIT) compiler, garbage collection, threads, and certain aspects of the operating system.

- ▶ SPECjbb2005 is an industry-standard benchmark that is designed to measure the server-side performance of the Java runtime environment (JRE).
- ▶ SPECjAppServer2004 (Java Application Server) is a multi-tiered benchmark for measuring the performance of a Java 2 Enterprise Edition (J2EE) technology-based application server.
- ▶ The SPECweb2005 benchmark includes workloads to measure banking, e-commerce, and support web server performance using HTTP (non-secure), HTTPS (secure), and a mix of secure and non-secure HTTP connections.
- ▶ The SPECsfs97_R1 benchmark includes workloads to measure both NFS V2 and NFS V3 server performance over UDP and TCP. Due to NFS V2 and UDP becoming less prevalent in client environments, the primary workload receiving the most focus is SPECsfs97_R1.v3 over TCP. The metrics for this benchmark include peak throughput (in NFS ops/sec) and response time (in msec/op).

TPC benchmarks

The TPC is a non-profit corporation that was founded to define transaction processing and database benchmarks and to disseminate objective, verifiable TPC performance data to the industry.

In this section, we provide a general description of the TPC benchmarks. The purpose of these database benchmarks is to provide performance data to the industry.

All the TPC results must comply with standard TPC disclosure policies and be reviewed by a TPC auditor. A Full Disclosure Report and Executive Summary must be submitted to the TPC before a result can be announced.

For further details about the TPC benchmarks and announced results, refer to the TPC website:

<http://www.tpc.org>

The TPC-C benchmark emulates a moderately complex online transaction processing (OLTP) environment. It simulates a wholesale supplier with a number of geographically distributed sales districts and associated warehouses, managing orders where a population of users executes transactions against a database.

The workload consists of five types of transactions:

- ▶ New order: Enters a new order from a customer
- ▶ Payment: Updates a customer's balance (recording payment)
- ▶ Order status: Retrieves the status of a customer's most recent orders
- ▶ Delivery: Deliver orders (queued for deferred execution)
- ▶ Stock level: Monitors the stock (inventory) level

The TPC-H benchmark models a decision support system by executing ad hoc queries and concurrent updates against a standard database under controlled conditions. The purpose of the benchmark is to “provide relevant, objective performance data to industry users” according to the specifications and all implementations of the benchmark. In addition to adhering to the specifications, the benchmark must be relevant to real-world (that is, client) implementations.

TPC-H represents the information analysis of an industry, which must manage, sell, or distribute a product worldwide. The 22 queries answer questions in areas, such as pricing and promotions, supply and demand management, profit and revenue management, client satisfaction, market share, and shipping management. The refresh functions are not meant to represent concurrent OLTP; they are meant to reflect the need to periodically update the database.

The TPC-E benchmark simulates the OLTP workload of a brokerage firm. The focus of the benchmark is the central database that executes transactions related to the firm's customer accounts. Although the underlying business model of TPC-E is a brokerage firm, the database schema, data population, transactions, and implementation rules have been designed to be broadly representative of modern OLTP systems.

Benchmark results

When performing a custom benchmark or Proof of Concept (PoC), it is important that you construct the test to simulate the production environment. This simulation is especially important as the hardware continues to evolve into the multi-core era and more time is being invested in the cache/memory hierarchy.

The most common pitfall when running a custom benchmark or a PoC is that the benchmark test does not simulate the real production environment and the benchmark result does not represent the performance that the system will achieve in the production environment. The achieved benchmark result might be much better for the benchmark test than for the real production workload, which most likely will lead to performance problems later when running the real workload. It also can happen the other way, potentially causing delays or the failure of the PoC.

Comparing benchmark results

When comparing performance benchmark results, it is important to compare the results of the same performance benchmark tests. The result of one benchmark test often does not represent the performance of a computer system for another workload.

For example, the result of a floating point-intensive benchmark test does not provide any information about the performance of the same computer running an integer-intensive benchmark or an OLTP workload and vice versa.

A common pitfall in setting the wrong performance expectations is to look at the results of one performance benchmark and apply it to another workload. For example, comparing the benchmark results of two computer systems running an OLTP workload that shows that machine A is 50% faster than machine B and expect that machine A will also be 50% faster for a workload that was not measured.

Archived



ITSO Power Systems testing environment

In this appendix, we explain the ITSO Power Systems testing environment.

Austin environment

Figure C-1 shows the ITSO Power System environment at the IBM Austin labs.



Figure C-1 Austin environment

Poughkeepsie benchmark center environment

Figure C-2 on page 395 shows the Poughkeepsie benchmark center Power System environment.



Figure C-2 Poughkeepsie benchmark center Power System environment

ITSO Poughkeepsie environment

Figure C-3 on page 396 shows the ITSO Power Systems environment.

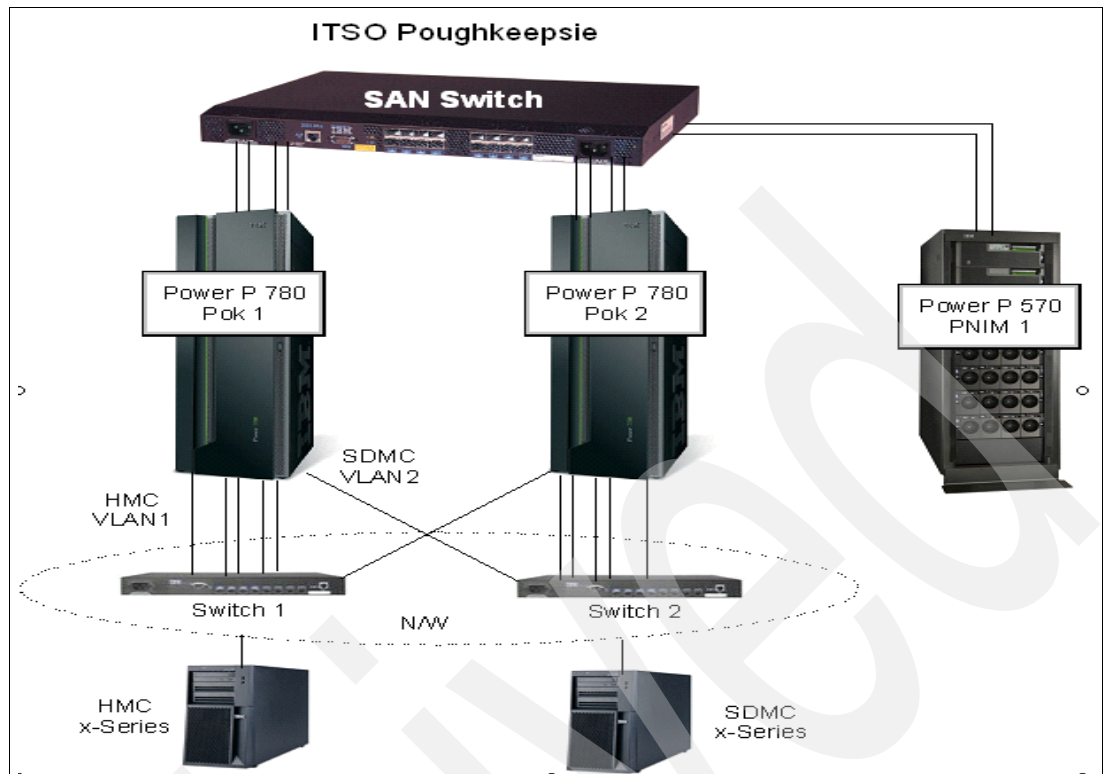


Figure C-3 ITSO Poughkeepsie Power Systems environment

Related publications

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

IBM Redbooks publications

The following IBM Redbooks publications provide additional information about the topic in this document. Note that several publications referenced in this list might be available in softcopy only.

- ▶ *IBM System z Personal Development Tool Volume 4 Coupling and Parallel Sysplex*, SG24-7859
- ▶ *Windows-based Single Sign-on and the EIM Framework on the IBM eServer iSeries Server*, SG24-6975
- ▶ *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590
- ▶ *Integrating AIX into Heterogeneous LDAP environments*, SG24-7165
- ▶ *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940-04
- ▶ *IBM Power 770 and 780 Technical Overview and Introduction*, REDP-4639
- ▶ *IBM Power 795 Technical Overview and Introduction*, REDP-4640
- ▶ *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590-02
- ▶ *Exploiting IBM AIX Workload Partitions*, SG24-7955
- ▶ *Integrated Virtual Ethernet Adapter Technical Overview and Introduction*, REDP-4340
- ▶ *Hardware Management Console V7 Handbook*, SG24-7491
- ▶ *IBM Systems Director Management Console: Introduction and Overview*, SG24-7860
- ▶ *IBM PowerVM Live Partition Mobility*, SG24-7460
- ▶ *IBM PowerHA SystemMirror 7.1 for AIX*, SG24-7845
- ▶ *PowerHA for AIX Cookbook*, SG24-7739
- ▶ *IBM AIX Version 6.1 Differences Guide*, SG24-7559
- ▶ *IBM Electronic Services Support using Automation and Web Tools*, SG24-6323
- ▶ *NIM from A to Z in AIX 5L*, SG24-7296
- ▶ *PowerVM Migration from Physical to Virtual Storage*, SG24-7825
- ▶ *AIX 5L Performance Tools Handbook*, SG24-6039
- ▶ *Getting Started with PowerVM Lx86*, REDP-4298
- ▶ *AIX Logical Volume Manager from A to Z: Introduction and Concepts*, SG24-5432

You can search for, view, download or order these documents and other IBM Redbooks publications, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Other publications

These publications are also relevant as further information sources:

- ▶ *POWER7 RAS features see POWER7 System RAS Key Aspects of Power Systems Reliability, Availability, and Serviceability*
<http://www-03.ibm.com/systems/power/hardware/whitepapers/ras7.html>
- ▶ *IBM Power Platform Reliability, Availability, and Serviceability (RAS) - Highly Available IBM Power Systems Servers for Business-Critical Applications, POW03003*
- ▶ IBM Power Architecture
[http://domino.research.ibm.com/library/cyberdig.nsf/papers/8DF8C243E7B01D948525787300574C77/\\$File/rc25146.pdf](http://domino.research.ibm.com/library/cyberdig.nsf/papers/8DF8C243E7B01D948525787300574C77/$File/rc25146.pdf)
- ▶ *PowerVM Virtualization Active Memory Sharing, REDP-4470*

Online resources

These websites are also relevant as further information sources:

- ▶ Power Systems Information Center
<https://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp>
- ▶ System Planning Tool
<http://www.ibm.com/systems/support/tools/systemplanningtool>
- ▶ IBM PowerVM Editions
<http://www.ibm.com/systems/power/software/virtualization/editions/index.html>
- ▶ To obtain a license for AME or to request a free 60 day trial
https://www-912.ibm.com/tcod_reg.nsf/TrialCod?OpenForm
- ▶ IBM Power Systems firmware
<http://www14.software.ibm.com/webapp/set2/sas/f/power5cm/power7.html>
- ▶ IBM Storage System drivers
<http://www-03.ibm.com/system/support/storage/ssic/interoperability.wss>
- ▶ For the latest HMC software
<http://www.ibm.com/support/fixcentral>
- ▶ IBM Fix Level Recommendation Tool (FLRT)
<http://www14.software.ibm.com/webapp/set2/flrt/home>
- ▶ Information about the virtual I/O server and the latest downloads
<http://www14.software.ibm.com/webapp/set2/sas/f/vios/download/home.html>
- ▶ Linux kernel security updates and downloads
<http://www14.software.ibm.com/webapp/set2/sas/f/pm/component.html>
- ▶ Virtual I/O server data sheet
<http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/datasheet.html>

- ▶ IBM Live Partition Mobility (LPM) information
<http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7hc3/iphc3whatsnew.htm>
- ▶ IBM Power Systems Hardware Information Center
http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7ed3/p7ed3cm_matrix_mmb.htm
- ▶ Director download link
<http://www.ibm.com/systems/software/director/resources.html>
- ▶ VMControl installation download
<http://www-03.ibm.com/systems/software/director/downloads/plugins.html>
- ▶ AEM plug-in installation download
<http://www-03.ibm.com/systems/software/director/downloads/plugins.html>
- ▶ Processor compatibility mode information
<http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7hc3/iphc3pcmdefs.htm>
- ▶ IBM PowerCare
<http://www-03.ibm.com/systems/power/support/PowerCare/>

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Archived

Index

Numerics

2-way SMT (SMT2) 294

4-way SMT (SMT4) 294

A

Active Energy Management (AEM) 183

Active Energy Manager 37, 341

Active Energy Manager (AEM) 336

Active Memory Expansion 8, 13

Active Memory Expansion (AME) 8, 13, 19, 248, 252, 305

Active Memory Mirroring 13, 271

Active Memory Sharing 6–8, 13, 45, 107, 217

Active Memory Sharing (AMS) 46, 252, 280

Active migration 66, 75

Active Partition Mobility 115

Advanced Energy Manager (AEM) 158

Advanced planning event 84

Advanced System Management Interface (ASMI) 265

Alternate Processor Recovery 271

Alternate processor recovery algorithm 23

Application WPAR 70

Autonomic Health Advisor FileSystem (AHAFS) 90

Autonomic Health Advisory File System API (AHAFS) 87

Availability 3

Availability optimization assessment (AOA) 337

Availability optimization service 337

B

Barrier synchronization registers (BSR) 67

Base operating system (BOS) 354

BM Systems Director VMControl Standard Edition 340

C

Capacity BackUp 87

Capacity on Demand 7

Capacity on Demand (CoD) 86

Capacity on demand (CoD) 24

CEC Concurrent Maintenance (CCM) 121

CEC Hot Add Repair Maintenance (CHARM) 17, 121

CEC hot add repair maintenance (CHARM) 196

Central Electronic Complex (CEC) 16, 121, 196

Check point/restart 75

Chipkill 19

Cluster Aware AIX (CAA) 81, 87

Cluster interface 328

Cluster type 82

Command

./lwiupdatemgr.sh 171

/opt/ibm/director/bin/srmstart 184

/opt/ibm/director/bin/srmstop 184

amepat 235, 252, 317

bosboot 296

cfgassist 222

cfgdev 223, 230

cfgmgr 61, 143, 359

chdev -dev ent0 -attr large_send=1 279

chdev -dev ent3 -attr largesend=1 279

chfs 82

chgrp 82

chhwres 15

chhwres -r mem -m -o r -q --id 298

chhwres -r proc -m -o r --procs --id 298

chlv 82

chuser capabilities

ties=CAP_BYPASS_RAC_VMM,CAP_PROPAGATE 287

clcmd 90

clstat 244

data2vg 372

db2set 287

db2set DB2_LARGE_PAGE_MEM=DB 287

df 142, 372

dscrcf 293

export IFX_LARGE_PAGES=1 287

gzip -cd SysDir_VMControl__Linux/AIX.tar.gz | tar -xvf - 168

ifconfig en0 -largesend 279

ifconfig en0 largesend 279

installp -agXY -d -p 361

installp Preview 361

iostat 273, 312, 315

iostat -b 315

ldedit -b lpdata 287

loadopt 40

lparstat 305, 307, 309, 312, 326

lparstat -c 319

lparstat -Ew 309

lparstat -H 311

lparstat -h 310

lparstat -i 308

lparstat -X -o /tmp/lparstat_data.xml 315

lpartstat 252

lscfg -vl 106

lsconf 149, 215–216

lsconfig -V 127

lsdev 40, 139, 357, 359

lshwres 15, 160

lslparmigr 66

lslparmigr -r manager 211

lsmap 40, 232

lsmemopt -m 15

lsnports 231

lspartition -dlpar 217

lspv 372

lsrep 39

lssp 39

lssrad 291, 298, 323
 lsvg 139, 372
 migratepv 139
 mkauthkeys 360
 mkdvd 136
 mkrep 39
 mksysb 136, 354
 mksysplan 154
 mkvdev 39, 56, 221
 mkvdev -sea ent1 -vadaptor ent5 -default ent5 -defaultid 1 -attr thread=0 279
 mkvopt 39
 mount ahafs /aha /aha 91
 mpstat 291, 312, 323
 mpstat -d 325
 mpstat -w -O sortcolumn=us,sortorder=desc,topcount=10 2 312
 nmon 322
 nohup ./nmem64 -m 2000 -s 3000 239
 optmem -x -o start -t mirror -q 15
 optmem -x -o stop 15
 pcpath query device 139
 ppc64_cpu 257
 raso 315
 raso -o biostat=1 315
 replacepv 139
 rmdev 205
 rmdev -dl vscsi0 358
 rmss 273
 sar 273, 306, 312–313
 sar -c -O sortcolumn=scall/s,sortorder=desc,topcount=10 -P ALL 1 313
 shutdown -Fr 143
 smcli lsmemopt -m 15
 smcli optmem -x -o start -t mirror -q 15
 smcli optmem -x -o stop 15
 smitty ffdc 23
 smstatus 184
 smctl 251, 257, 296
 smctl -t 1 307
 smctl -t 2 306
 svmon 270, 289, 291, 323
 svmon -O summary=ame 319
 topas 252, 273, 307, 309, 312, 314, 319
 topas -L 308
 topas_nmon 307
 vfcmap 232
 vmo 127, 290–291
 vmo -p -o lgpg_regions= -o lgpg_size=16777216 287
 vmo -p -o v_pinshm=1 287
 vmstat 252, 273, 288, 312
 vmstat -c 319
 Common agent 170
 Complex Instruction Set Computing technology (CISC) 378
 Component Trace (CT) 23
 Concurrent Firmware maintenance (CFM) 35
 Concurrent GX adapter add 123
 Customer Specified Placement (CSP) 154

D

Data backups 2
 Data conversions 2
 Data Stream Control Register (DSCR) 293
 Dedicated processor partitions 34
 Deferred updates 35
 DRAM sparing 19
 Dual-threaded (SMT2) 256
 Dynamic Automatic Reconfiguration (DARE) 80
 Dynamic Logical Partitioning 6
 Dynamic logical partitioning (DLPAR) 113
 Dynamic Power Saver 37
 Dynamic power saver 37
 Dynamic power saving 188
 Dynamic processor deallocation 23
 Dynamic processor sparing 23

E

Electronic Service Agent 125, 331
 Electronic service agent 92
 Electronic Service Agent (ESA) 14
 Electronic service agent (ESA) 92
 Electronic services 92
 Energy Optimized Fans 37
 EnergyScale for I/O 37
 Enterprise Identity Mapping (EIM) 345
 ESA (Electronic Service Agent) 17
 Etherchannel 82
 Event infrastructure 90

F

Favor Performance 37
 Favor Power 37
 Favour energy mode 188
 Favour performance mode 188
 File
 /etc/exclude.rootvg 138
 /etc/inittab 294
 Fileset
 bos.ahafs 91
 First Failure Data Capture (FFDC) 23
 Flex Capacity Upgrade on Demand 85

G

Global Environment 69
 Group Services 88
 GX Add 129
 GX Repair 129

H

Hardware Management Console (HMC) 7, 25, 93, 125, 158, 203, 206
 Hardware management Console (HMC) 151
 Hardware page table (HPT) 68
 Hardware scrubbing 25
 Hardware upgrades 2
 HBA (Host Bus Adapter) 105

- Health Insurance Portability and Accountability Act of 1996 (HIPPA) 342
- High Performance Computing (HPC) 257
- Host Based Adapters (HBAs) 138
- Host channel adapter (HCA) 16
- Host Ethernet Adapter (HEA) 48, 386
- Host node repair 205
- Hot GX adapter add/repair 13
- Hot node add 28, 121
- Hot node repair 28
- Hot node upgrade 28
- Hot node upgrade (memory) 121
- Hot upgrade 203

I

- IBM PowerVM 5
- Image repository 166
- Inactive memory units 86
- Inactive migration 66, 68
- Inactive mobility 75
- Inactive Partition Mobility 115
- Inactive processor cores 86
- Integrated Virtual Ethernet (IVE) 48, 55
- Integrated Virtual Ethernet Adapter (IVE) 8, 217
- Integrated Virtual Ethernet adapter (IVE) 386
- Integrated Virtualization Manager 265
- Integrated Virtualization Manager (IVM) 7, 41, 108, 158, 229

L

- Lightweight Directory Access Protocol (LDAP) 345
- Lightweight Directory Access Protocol server (LDAP) 345
- Lightweight Memory Trace (LMT) 23
- Link aggregation 102
- Link Aggregation (LA) 82
- Live Application Mobility 75
- Live Dump 23
- Live Partition Mobility 8, 64, 113, 159
- Live Partition Mobility 6
- Live relocation 173
- Local Host Ethernet Adapters (LHEA) 105
- Logical Memory Block (LMB) 117
- Logical memory block (LMB) 265
- Logical memory block size (LMB) 67

M

- Managed server 69
- Mathematical Acceleration Subsystem (MASS) 297
- MaxCore 259
- MaxCore mode 28
- Memory defragmentation 14
- Memory page deallocation 25
- Mover Service Partition 68
- Mover service partition 222
- Multiple Shared Processor Pool 6
- Multiple Shared Processor Pools 7

N

- N_port ID Virtualization 6
- N_Port ID Virtualization (NPIV) 279
- Network Authentication Service (NAS) 345
- Network File System (NFS) 384
- Network interface backup 82
- Network Interface Backup (NIB) 221
- Networks 81
- Node Add 129
- Node evacuation 204
- Node interface 328
- Node list interface 328
- Node Repair 129
- Node Upgrade 129
- Nodes 81
- Non-volatile RAM (NVRAM) 68

O

- Off peak schedule 17
- On/Off Capacity on Demand 87
- Open Virtualization Format (OVF) 340

P

- Paging devices 47
- Partition availability priority 24
- Partition Power Management 37
- Payment Card Industry Data Security Standard (PCI DSS) 342
- Power Capping 37
- Power capping 187
 - Absolute value 187
 - Percentage value 187
- Power distribution units (PDU) 100
- Power Flex 82
- POWER hypervisor (PHYP) 33
- Power Instruction Set Architecture (ISA) 254
- Power saving 188
- Power storage protection keys 258
- Power Trending 37
- PowerVM Enterprise Edition 6
- PowerVM Hypervisor 6
- PowerVM Lx86 6
- PowerVM Standard Edition 6
- Prepare for Hot Repair or Upgrade utility (PHRU) 204
- Prepare for Hot Repair/Upgrade (PHRU) 17
- Processor Core Nap 37
- Processor Folding 37
- Processor folding 378
- Processor instruction retry algorithms 23

Q

- Quad-threaded (SMT4) 256
- Quality of Service (QoS) 376

R

- Random Access Memory (RAM) 381
- Redbooks website 397

- Contact us xiii
- Reduced Instruction Set Computing (RISC) 378
- Redundant I/O 17
- Redundant service processor 25
- Release Level 35–36
- Reliability 3
- Reliability, Availability and Serviceability (RAS) 1, 121, 271
- Reliable Scalable Cluster Technology (RSCT) 88
- Relocation policies 166
 - Automatic relocation 166
 - Manual relocation 166
 - Policy based relocation 166
- Resource allocation domain (RAD) 323
- Resource group 82
- Resource Monitoring and Control (RMC) 64, 88, 93
- RSCT Peer Domains (RPD) 89
- Run-Time Error Checking (RTEC) 23

S

- Sarbanes-Oxley Act of 2002 (SOX) 342
- Scheduler Resource Allocation Domain (SRAD) 323
- Script
 - dbstart 233
 - dbstart.sh 239
 - nmem64 233
 - webstart 233
- Segment Lookaside Buffer (SLB) 289
- Server Power Down 37
- Service Pack 35
- Service packs 36
- Serviceability 3
- Shared aliases 289
- Shared Dedicated Capacity 7
- Shared Ethernet Adapter 82
- Shared Ethernet Adapter (SEA) 55, 221, 386
- Shared memory pool 46
- Shared processor partitions 34
- Shared Storage Pools 6
- Simultaneous multi-thread (SMT2) 251
- Simultaneous Multi-Threading (SMT) 7
- Simultaneous Multithreading (SMT) 294, 378
- Simultaneous Multithreading Mode (SMT) 256
- Single Instruction Multiple Data (SIMD) 257
- single points of failures (SPOFs) 101
- Single processor checkpoint 24
- Single thread (ST) 251
- Single thread mode (ST) 378
- Single-chip module (SCM) 30
- Single-threaded (ST) 256
- Software upgrades 2
- Solid State Disk (SSD) 281
- Static power saving 188
- Static Power Server 37
- Static relocation 173
- Symmetrical Multi-Processor (SMP) 381
- System Director Management Console (SDMC) 7, 25, 64, 102, 206, 265, 357
- System firmware 35
- System firmware mirroring 13

- System migrations 2
- System Planning Tool (SPT) 103, 151
- System pool 165
- System WPAR 70
- Systems Director Management Console (SDMC) 93, 151

T

- Technical and Delivery Assessment (TDA) 150
- Technology Independent Machine Interface (TIMI) 35
- Thermal Reporting 37
- Time of day (ToD) 68
- Tool
 - amepat 21
- Topology Services 88
- Translation Control Entry (TCE) 123
- Transport Layer Security (TLS) protocol 94
- Trial PowerVM Live Partition Mobility 9
- TurboCore 252, 259
- TurboCore mode 28

U

- Uninterruptable power supply (UPS) 101
- Unshared aliases 289
- Utility Capacity on Demand 87

V

- Vector Media Extensions (VMX) 257
- Vector Multimedia Extension (VME) 146
- Vector Scalar Extension (VSX) 146, 257
- Vector scalar extension (VSX) 297
- Virtual appliance 165
- Virtual Asynchronous Service interface (VASI) 65
- Virtual Ethernet 8
- Virtual farm 166
- Virtual I/O Server 6–7
- Virtual Memory Manager (VMM) 381
- Virtual SCSI 8
- Virtual Server 69
- Virtual server 165
- Virtual Servers 50
- Virtual servers 159

W

- Workload 165
- Workload Partition (WPAR) 69
- Workload Partitions 8
- World Wide Port Name (WWPN) 43



Power Systems Enterprise Servers with PowerVM Virtualization and RAS

(0.5" spine)
0.475" <-> 0.873"
250 <-> 459 pages



Power Systems Enterprise Servers with PowerVM Virtualization and RAS



Redbooks®

**Unleash the IBM
Power Systems
virtualization features**

**Understand reliability,
availability, and
serviceability**

**Learn about various
deployment case
scenarios**

This IBM Redbooks publication illustrates implementation, testing, and helpful scenarios with IBM Power Systems 780 and 795 using the comprehensive set of the Power virtualization features. We focus on the Power Systems functional improvements, in particular, highlighting the reliability, availability, and serviceability (RAS) features of the enterprise servers.

This document highlights IBM Power Systems Enterprise Server features, such as system scalability, virtualization features, and logical partitioning among others. This book provides a documented deployment model for Power 780 and Power 795 within a virtualized environment, which allows clients to plan a foundation for exploiting and using the latest features of the IBM Power Systems Enterprise Servers.

The target audience for this book includes technical professionals (IT consultants, technical support staff, IT Architects, and IT Specialists) responsible for providing IBM Power Systems solutions and support.

**INTERNATIONAL
TECHNICAL
SUPPORT
ORGANIZATION**

**BUILDING TECHNICAL
INFORMATION BASED ON
PRACTICAL EXPERIENCE**

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

**For more information:
ibm.com/redbooks**

SG24-7965-00

ISBN 0738436267