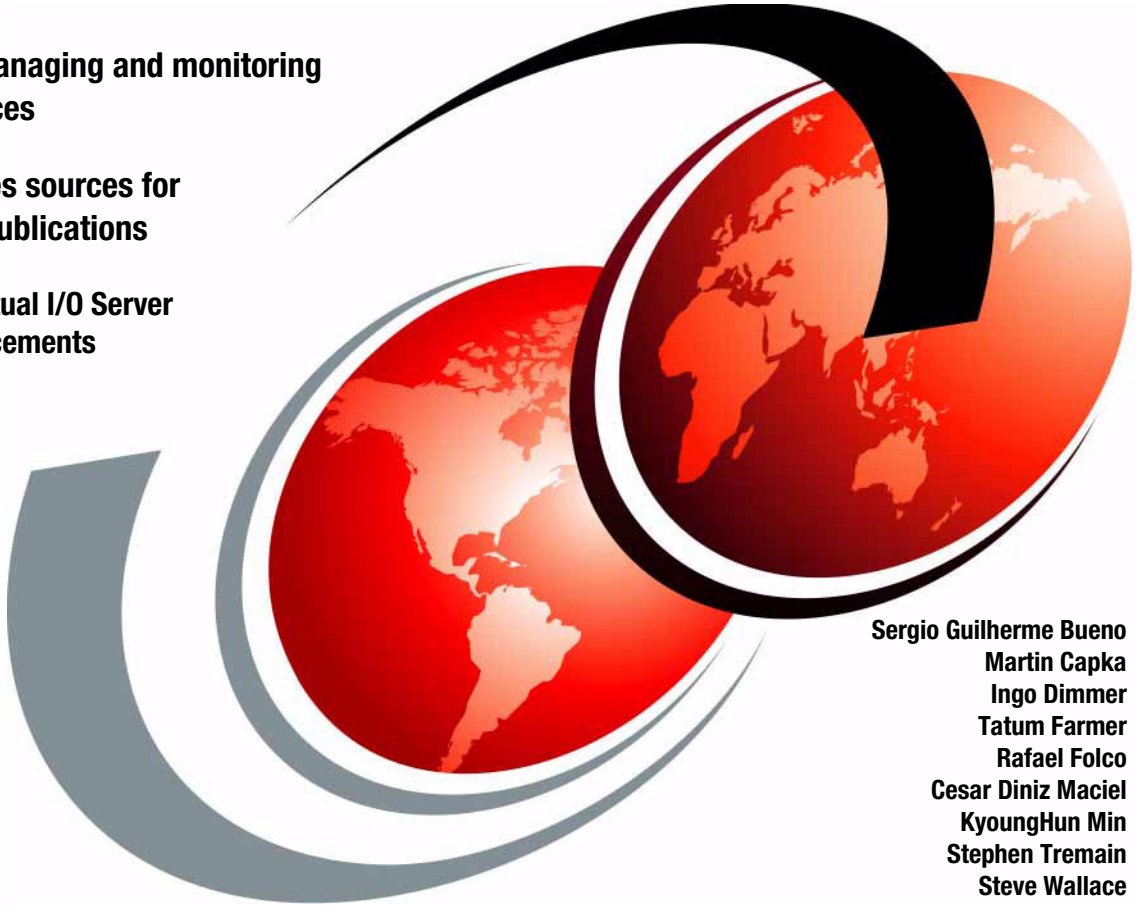
IBM PowerVM Virtualization Managing and Monitoring

- Provides managing and monitoring best practices
- Consolidates sources for PowerVM publications
- Includes Virtual I/O Server 2.2.2 enhancements



Sergio Guilherme Bueno
Martin Capka
Ingo Dimmer
Tatum Farmer
Rafael Folco
Cesar Diniz Maciel
KyoungHun Min
Stephen Tremain
Steve Wallace

Redbooks

ibm.com/redbooks

International Technical Support Organization

IBM PowerVM Virtualization Managing and Monitoring

June 2013

SG24-7590-04

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

Fifth Edition (June 2013)

This edition applies to:
Version 7, Release 1 of AIX (product number 5765-G98)
Version 7, Release 1 of IBM i (product number 5770-SS1)
Version 2, Release 2, Modification 2, Fixpack 26, Service pack 1 of the Virtual I/O Server
Version 7, Release 7, Modification 6 of the HMC
Version EM350, release 132 of the POWER6 System Firmware
Version AM760, release 051of the POWER7 System Firmware.

Contents

Figures

Tables

Examples

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 grant 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.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

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. 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 IBM's application programming interfaces.

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™
AIX®
BladeCenter®
DB2®
DS4000®
DS8000®
EnergyScale™
Enterprise Storage Server®
Focal Point™
GDPS®
Geographically Dispersed Parallel Sysplex™
Global Business Services®
GPFS™
HACMP™
i5/OS™
IBM®
IBM Systems Director Active Energy Manager™
Micro-Partitioning®
Netcool®
Parallel Sysplex®
POWER®
POWER Hypervisor™
Power Systems™
POWER6®
POWER6+™
POWER7®
POWER7 Systems™
POWER7+™
PowerHA®
PowerLinux™
PowerVM®
PureFlex™
Redbooks®
Redbooks (logo) ®
RS/6000®
Storwize®
System i®
System p®
System Storage®
SystemMirror®
Systems Director VMControl™
Tivoli®

The following terms are trademarks of other companies:

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

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.

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

Preface

IBM® PowerVM® virtualization technology is a combination of hardware and software that supports and manages the virtual environments on POWER5-, POWER5+, IBM POWER6®, and IBM POWER7®-based systems.

PowerVM is available on IBM Power Systems™, and IBM BladeCenter® servers as optional Editions, and is supported by the IBM AIX®, IBM i, and Linux operating systems. You can use this set of comprehensive systems technologies and services to aggregate and manage resources by using a consolidated, logical view. Deploying PowerVM virtualization and IBM Power Systems offers you the following benefits:

- Lower energy costs through server consolidation
- Reduced cost of your existing infrastructure
- Better management of the growth, complexity, and risk of your infrastructure

To achieve this goal, PowerVM virtualization provides the following technologies:

- Virtual Ethernet
- Shared Ethernet Adapter
- Virtual SCSI
- IBM Micro-Partitioning® technology
- Multiple Shared-Processor Pools
- N_Port Identifier Virtualization
- PowerVM Live Partition Mobility
- IBM Active Memory™ Sharing
- Active Memory Expansion
- Active Memory Mirroring
- Active Memory Deduplication
- Partition Suspend and Resume
- Shared Storage Pools
- Dynamic Platform Optimizer

This IBM Redbooks® publication is an extension of *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940. It provides an organized view of best practices for managing and monitoring your PowerVM environment concerning virtualized resources managed by the Virtual I/O Server.

This publication is divided by technology into five parts:

1. Processor virtualization
2. Memory virtualization
3. I/O virtualization

4. Virtual I/O Server
5. Managed system virtualization

Authors

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

Sergio Guilherme Bueno is an IBM Senior IT Specialist working in the ITS Delivery in IBM Brazil. He has 14 years of experience in AIX System Administration. He is an IBM Certified Advanced Technical Expert for Power Systems, and has a degree in Data Processing Technology from Faculdade de Administracao e Informatica (FAI) - Santa Rita do Sapucai, Brazil. His expertise includes IBM PowerHA®, server consolidation, and IBM Enterprise Disk Storage Subsystems. He was a coauthor of the *AIX Version 4.3 to 5L Migration Guide*.

Martin Capka is a senior IBM consultant currently working for IBM Business Partner GC system in the Czech Republic. He has 10 years of experience with a broad range of IBM technologies. His areas of expertise include IBM POWER® Systems, AIX, PowerVM, PowerHA, and IBM Tivoli® Storage Manager. He is an IBM Certified Advanced Technical Expert for Power Systems, an IBM Certified Systems Expert for Virtualization, and an IBM Certified Systems Expert for High Availability. In his current job, he is responsible for the architecture, design, and implementation of virtualized solutions with IBM Power Systems for IBM strategic customers in the Czech Republic.

Ingo Dimmer is an IBM Consulting IT Specialist for IBM i, and a PMI Project Management Professional working in the IBM STG ATS Europe storage support organization in Mainz, Germany. He has thirteen years of experience in enterprise storage support from working in IBM post-sales and pre-sales support. He holds a degree in Electrical Engineering from the Gerhard-Mercator University Duisburg. His areas of expertise include IBM i external disk and tape storage solutions, PowerVM virtualization, IBM PureFlex™ systems, I/O performance, and high availability. He has authored several White Papers and IBM Redbooks publications on these subjects.

Tatum Farmer is the UNIX Team Lead at Independence Blue Cross in Philadelphia supporting AIX, Linux, and Solaris systems. A graduate of Drexel University and a former IBMer, he has 20 years of experience in the IT industry. His areas of expertise range from systems programmer to administrator, and include Power Systems, PowerVM, and AIX.

Rafael Folco is a Software Engineer at the STG Linux Technology Center, IBM Brazil. He has ten years of experience with Linux and seven years at IBM. He is a

Gold Redbooks Author who has written four PowerVM publications. He holds a postgraduate degree in Software Engineering, and his areas of expertise include Power Systems high availability and performance.

Cesar Diniz Maciel is an Executive IT Specialist with IBM in the United States. He joined IBM in 1996 as Presales Technical Support for the IBM RS/6000® family of UNIX servers in Brazil, and came to IBM United States in 2005. He is part of the Global Techline team, working on presales consulting for Latin America. He holds a degree in Electrical Engineering from Universidade Federal de Minas Gerais (UFMG) in Brazil. His areas of expertise include Power Systems, AIX, and IBM POWER Virtualization. He has written extensively on Power Systems and related products. This is his eighth ITSO residency.

KyoungHun Min is a senior System Service Representative working for IBM Korea in Gumi. He has 10 years of experience in IT industry. His areas of expertise include Power Systems, AIX, System performance and tuning, PowerVM, PowerHA, IBM GPFS™, and Storage subsystems. He is an IBM Certified Advanced Technical Expert for Power Systems.

Stephen Tremain Stephen is a former resident, who has been with IBM for six years, and currently works as a Software Engineer at the IBM Security Systems Lab on the Gold Coast in Queensland, Australia. Before joining IBM, Stephen worked as a UNIX System Administrator with an investment bank for 10 years, and also worked in the education and research sectors. Stephen graduated from the University of New England in Australia, with a B.Sc. and a Dip. Sci. Agriculture.

Steve Wallace has been with IBM UK for 15 years and works as an Infrastructure Architect for IBM Global Business Services® (GBS). Steve works onsite with large public sector clients in the UK, designing and implementing virtualization-based solutions and overseeing major POWER based infrastructure projects.

The project that produced this publication was managed by:
Scott Vetter, PMP

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

Syed R Ahmed, Suman Batchu, Bob Battista, David Bennin, Ping Chen, Richard M. Conway, Joeseph Czap, Linda Flanders, Maria Garza, Robert Jennings, Bob Kovacs, Yiwei Li, Ann Lund, Neal Marion, P Scott McCord, Francisco Moraes, Nidugala Muralikrishna, Steve Nasypany, Terrence Nixa, Paul F. Olsen, Ed Prosser, Steven E Royer, Manash Sarma, Ron Schmerbauch, Alfred Schwab, Vasu Vallabhaneni, Steve Wallace, Kristopher Whitney, Bradley Vette
IBM US

Bruno Blanchard
IBM France

Nigel Griffiths
IBM UK

The authors of the first edition are:

Tomas Baublys	IBM Germany
Damien Faure	Bull France
Jackson Alfonso Krainer	IBM Brazil
Michael Reed	IBM US

The authors of the second edition are:

Ingo Dimmer	IBM Germany
Volker Haug	IBM Germany
Thierry Huché	IBM France
Anil K Singh	IBM India
Morten Vågmo	IBM Norway

The authors of the third edition are:

Stuart Devenish	IBM Australia
Ingo Dimmer	IBM Germany
Rafael Folco	IBM Brazil
Mark Roy	Sysarb, Inc. Australia
Stephane Saleur	IBM France
Oliver Stadler	IBM Switzerland
Naoya Takizawa	IBM Japan

The authors of the fourth edition are:

Nicolas Guerin	IBM France
Jimi Inge	Tieto Sweden
Narutsugu Itoh	IBM Japan
Robert Miciovici	IBM Romania
Rajendra Patel	IBM US
Arthur Török	IBM Hungary

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 customer 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

Summary of changes

This section describes the technical changes made in this edition of the book and in previous editions. This edition might also include minor corrections and editorial changes that are not identified.

Summary of Changes
for SG24-7590-04
for IBM PowerVM Virtualization Managing and Monitoring
as created or updated on June 27, 2014.

June 2013, Fifth Edition

This revision reflects the addition, deletion, or modification of new and changed information described below.

New information

The following information was sourced from other publications and updated to reflect the latest enhancements:

- Chapter 6, “Active Memory Deduplication” on page 131
- Chapter 13, “Partition Suspend and Resume” on page 517
- Chapter 14, “Live Partition Mobility” on page 541
- Chapter 15, “Dynamic Platform Optimizer” on page 599
- Chapter 16, “Active System Optimizer and Dynamic System Optimizer for AIX” on page 611

Changed information

- Reorganized the contents of the book into parts that reflect the overall areas of processor, memory, I/O, Virtual I/O Server, Managed Systems, and management tools. Within those parts, separated managing and monitoring activities.
- The following sections have been updated to include POWER7 based offerings:
  - Chapter 1, “PowerVM introduction” on page 1
  - Chapter 2, “Shared Processor Pool” on page 11
  - Chapter 3, “Multiple Shared Processor Pools” on page 65
  - Chapter 10, “Shared storage pools” on page 285
  - Chapter 11, “Virtual I/O Server” on page 323
  - Chapter 12, “Dynamic logical partitioning” on page 457

1

PowerVM introduction

Businesses are turning to PowerVM virtualization to consolidate multiple workloads onto fewer systems, increasing server utilization and reducing cost. PowerVM technology provides a secure and scalable virtualization environment for AIX, IBM i, and Linux applications. It is built on the advanced reliability, availability, and serviceability features and the leading performance of the Power Systems platform.

This book targets clients who are experienced in virtualization. It is split into six parts: PowerVM Introduction, Processor virtualization, Memory virtualization, I/O virtualization, Virtual I/O Server, and Managed systems virtualization.

This chapter describes the available PowerVM Editions. It also provides an overview of the new Virtual I/O Server Version 2.2.2 features and PowerVM enhancements.

This chapter includes the following sections:

- Management and monitoring strategy
- PowerVM
- PowerVM editions
- New PowerVM version 2.2.2 features

1.1 Management and monitoring strategy

A management and monitoring strategy is important in any computing environment. Consider guidelines for updates and changes before you go into production. When you manage a complex virtualization environment on a Power Systems server that runs dozens of partitions with varying applications and workloads, develop a strategic plan for both management and monitoring.

A single strategy for every enterprise and situation does not exist. Your strategic plans must be tailored to accommodate the uniqueness of your organization and computing environment.

PowerVM offers features that allow you to reduce outages, planned or unplanned, in your environment. For example, Live Partition Mobility can be used to move a workload from one managed system to another without interruption. Dual Virtual I/O Server configuration allows Virtual I/O Server maintenance without disrupting partitions. By combining Power Systems virtualization and SAN technologies, you can create a flexible and responsive implementation in which any hardware or software can be exchanged and upgraded.

Cross-platform tools such as IBM Systems Director, Tivoli, and Extreme Cloud Administration Toolkit (xCAT) offer single management and monitor interfaces for multiple physical and virtual systems. You can use the Hardware Management Console to manage and monitor multiple virtual systems on Power Systems. The Integrated Virtualization Manager manages virtual systems on a single server.

The advantages that are provided by virtualization (infrastructure simplification, energy savings, flexibility, and responsiveness) also include manageability. A virtualized environment replaces not a single server, but dozens and sometimes hundreds of hard-to-manage, under utilized, stand-alone servers.

1.2 PowerVM

PowerVM is the industrial-strength virtualization solution for IBM Power Systems servers and blades. This solution provides workload consolidation that helps you control costs by improving overall performance, availability, flexibility, and energy efficiency. PowerVM is a combination of hardware enablement and added value software.

The term PowerVM usually refers to the features and technologies listed in Table 1-1.

Table 1-1 PowerVM features and technologies

Features and Technologies	Function Provided by
Active Memory Deduplication	Hypervisor
Active Memory Mirroring[a]	Hypervisor
Dynamic Logical Partitioning	Hypervisor
Dynamic Platform Optimization[a]	Hypervisor
Integrated Virtualization Manager	Hypervisor, VIOS
Live Partition Mobility	Hypervisor, VIOS
Logical Partitioning	Hypervisor
Micro-Partitioning	Hypervisor
Partition Suspend/Resume	Hypervisor, VIOS
PowerVM Hypervisor	Platform
Shared Processor Pools	Hypervisor
Share Storage Pools	Hypervisor, VIOS
Virtual Fibre Channel[b]	Hypervisor, VIOS
Virtual I/O Server	Hypervisor
Virtual optical device & tape	Hypervisor, VIOS
Virtual SCSI	Hypervisor, VIOS

a. Requires mid-tier and large-tier POWER7 or later Power System hardware support, including p770, p780, and p795.
b. Often referred to as NPIV.

The technologies in Table 1-2 are not part of the PowerVM family, but they are frequently mentioned with PowerVM. They are addressed because they are complementary technologies that are important in a virtualized environment.

Table 1-2 Complementary technologies

Features and Technologies	Function Provided by
POWER processor modes	Hypervisor
Capacity on Demand	Hypervisor

Features and Technologies	Function Provided by
Simultaneous multithreading	Hardware, AIX, IBM i, Linux
Active Memory Expansion	Hardware[a], AIX
Workload Partition	AIX[b]
System Planning Tool (SPT)	SPT
Integrated Virtualization Subsystems	IBM i[c]

a. Requires POWER7 processor-based systems or later
b. Requires AIX Version 6.1 or later
c. Subsystems are provided by IBM i

PowerVM has three editions, each with a specific feature set for clients. For more information about the licensed features of the PowerVM editions, see 1.3, “PowerVM editions” on page 4.

1.3 PowerVM editions

This section provides information about the virtualization capabilities of PowerVM. There are three versions of PowerVM, suited for various purposes:

- PowerVM Express Edition

  PowerVM Express Edition is designed for clients who are looking for an introduction to virtualization features at a highly affordable price.

- PowerVM Standard Edition

  PowerVM Standard Edition provides the most complete virtualization functionality for AIX, IBM i, and Linux operating systems in the industry. PowerVM Standard Edition is supported on Power Systems servers and includes features that are designed to allow businesses to increase system utilization.

- PowerVM Enterprise Edition

  PowerVM Enterprise Edition includes all the features of PowerVM Standard Edition plus three new industry-leading capabilities called Active Memory Sharing, Live Partition Mobility, and Dynamic Platform Optimization.

You can upgrade from the Express Edition to the Standard or Enterprise Edition, and from the Standard to the Enterprise Edition. Table 1-3 outlines the functional elements of the PowerVM editions.

Table 1-3 Overview of PowerVM capabilities by edition

PowerVM capability	PowerVM Express Edition	PowerVM Standard Edition	PowerVM Enterprise Edition
Maximum VMs	3 / Server	1000 / Server	1000 / Server
Micro-partitions[a]	Yes	Yes	Yes
Virtual I/O Server	Yes (Single)	Yes (Multiple)	Yes (Multiple)
Management	VMControl, IVM	VMControl, IVM[b], HMC	VMControl, IVM[b], HMC
Shared Dedicated Capacity	Yes	Yes	Yes
Multiple Shared-Processor Pools[c]	No	Yes	Yes
Live Partition Mobility	No	No	Yes
Active Memory Sharing[c]	No	No	Yes
Active Memory Deduplication[d]	No	No	Yes
Suspend/Resume	No	Yes	Yes
Virtual Fibre Channel[e]	Yes	Yes	Yes
Shared Storage Pools (SSP)	No	Yes	Yes
SSP Thin Provisioning	No	Yes	Yes
SSP Thick Provisioning	No	Yes	Yes

a. When the firmware is at level 7.6, or later, Micro-Partitioning can be defined as small as 0.05 of a processor instead of 0.1 of a processor.
b. IVM supports only a single Virtual I/O Server.
c. Requires IBM POWER6 processor-based systems or later.
d. Requires IBM POWER7 processor-based systems with firmware at level 7.4 or later.
e. Often referred to as NPIV.

For an overview of the availability of the PowerVM features by Power Systems models, see this website:

http://www.ibm.com/systems/power/software/virtualization/editions/features.html

The PowerVM feature is a combination of hardware enablement and software that are available together as a single priced feature. It is charged as one unit for each activated processor, including software maintenance. The software maintenance can be ordered for a one-year or three-year period, and is charged for each active processor on the managed system.

When the hardware feature is specified with the initial system order, the managed system is shipped with the firmware activated to support the PowerVM features.

For an HMC-attached system with the PowerVM Standard Edition or the PowerVM Enterprise Edition, the processor-based license enables you to install several Virtual I/O Server partitions on a single physical server. This configuration provides redundancy and spreads the I/O workload across several Virtual I/O Server partitions.

Virtual Ethernet and dedicated processor partitions are available without the PowerVM feature for servers that are attached to an HMC.

1.4 New PowerVM version 2.2.2 features

Power Systems servers coupled with PowerVM technology are designed to help clients build a dynamic infrastructure, reducing costs, managing risk, and improving service levels.

IBM PowerVM V2.2.2 includes VIOS 2.2.2.1-FP26, HMC V7R7.6 and Power Systems Firmware level 760, and contains the following enhancements for managing a PowerVM virtualization environment:

- Supports for up to 20 partitions per processor, doubling the number of partitions that are supported per processor. This improvement provides more flexibility by reducing the minimum processor entitlement to 5% of a processor.
- Dynamic LPAR add or remove of virtual I/O adapters to or from a Virtual I/O Server partition. HMC V7R7.6 or later now automatically runs the add/remove commands (cfgdev/rmdev) on the Virtual I/O Server for the user. Before this enhancement, you had to manually run these commands on the Virtual I/O Server.

- Ability for the user to specify the destination Fibre Channel port for any or all virtual Fibre Channel adapters.
- Improved Virtual I/O Server setup, tuning, and validation by using the Runtime Expert.
- Live Partition Mobility supports up to 16 concurrent LPM activities.
- Shared Storage Pools create pools of storage for virtualized workloads. These pools can improve storage usage, simplify administration, and reduce SAN infrastructure costs. The enhanced capabilities enable 16 nodes to participate in a shared storage pool configuration, which can improve efficiency, agility, scalability, flexibility, and availability.

  Shared Storage Pools provide these flexibility and availability improvements:

  - IPv6 and VLAN tagging (IEEE 802.1Q) support for intermodal shared storage pools communication.
  - Cluster reliability and availability improvements.
  - Improved storage utilization statistics and reporting.
  - Nondisruptive rolling upgrades for applying service.
  - Advanced features that accelerate partition deployment, optimize storage usage, and improve availability through automation.
- New VIOS Performance Advisor analyzes Virtual I/O Server performance, and makes recommendations for performance optimization.
- PowerVM has the following new advanced features that are enabled by VMControl that accelerate partition deployment, optimize storage utilization, and improve availability through automation:

  - Linked clones allow for sharing of partition images, which greatly accelerates partition deployment and reduces the storage usage.
  - System pool management for IBM i workloads provides increased flexibility and resource utilization. For more information about the appropriate IBM Systems Director VMControl™ release, see:

    http://www.ibm.com/systems/software/director/vmcontrol/

Part 1

Processor virtualization

This part describes the processor virtualization considerations for managing and monitoring virtual processors assigned to the logical partitions in your managed system. Different approaches and techniques for processor virtualization are addressed.

This part contains the following chapters:

- Shared Processor Pool
- Multiple Shared Processor Pools
- Multiple Shared Processor Pools
- POWER processor compatibility modes

2

Shared Processor Pool

Shared Processor Pools have been available since the introduction of IBM POWER5 processor-based managed systems. Using shared processor pools, processor resources can be used more efficiently and overall system utilization is significantly increased.

IBM POWER5 processor-based managed systems support one default shared processor pool, whereas IBM POWER6 processor-based and later managed systems support multiple shared processor pools.

2.1 Managing Shared Processor Pools

The default Shared Processor Pool is Pool ID 0. This pool and its default configuration values cannot be changed. Managed systems capable of Multiple Shared Processor Pools allow the activation and modification of predefined Shared Processor Pools. For more information about managing of Multiple Shared Processor Pools, see Chapter 3, “Multiple Shared Processor Pools” on page 65.

2.1.1 Micro-Partitioning

Shared Processor Pools are used to manage the processor capacity of a group of micro-partitions. Although dedicated processor partitions can donate processor cycles to Shared Processor Pools, they cannot be assigned to a Shared Processor Pool.

Changing dedicated processor partitions to micro-partitions

If you have partitions that are running with dedicated processors, you can convert them to micro-partitions (also known as shared processor partitions). You must edit the partition profile, and then deactivate and reactivate the partition. To change the partitions, complete the following steps:

1. Edit the profile for the partition you want to change the processing mode. Click **Configuration** → **Manage Profiles** and select the profile that you want to edit.

2. On the new panel, click the **Processor** tab, then select **Shared** as shown in Figure 2-1. Click **OK** to save the changes.

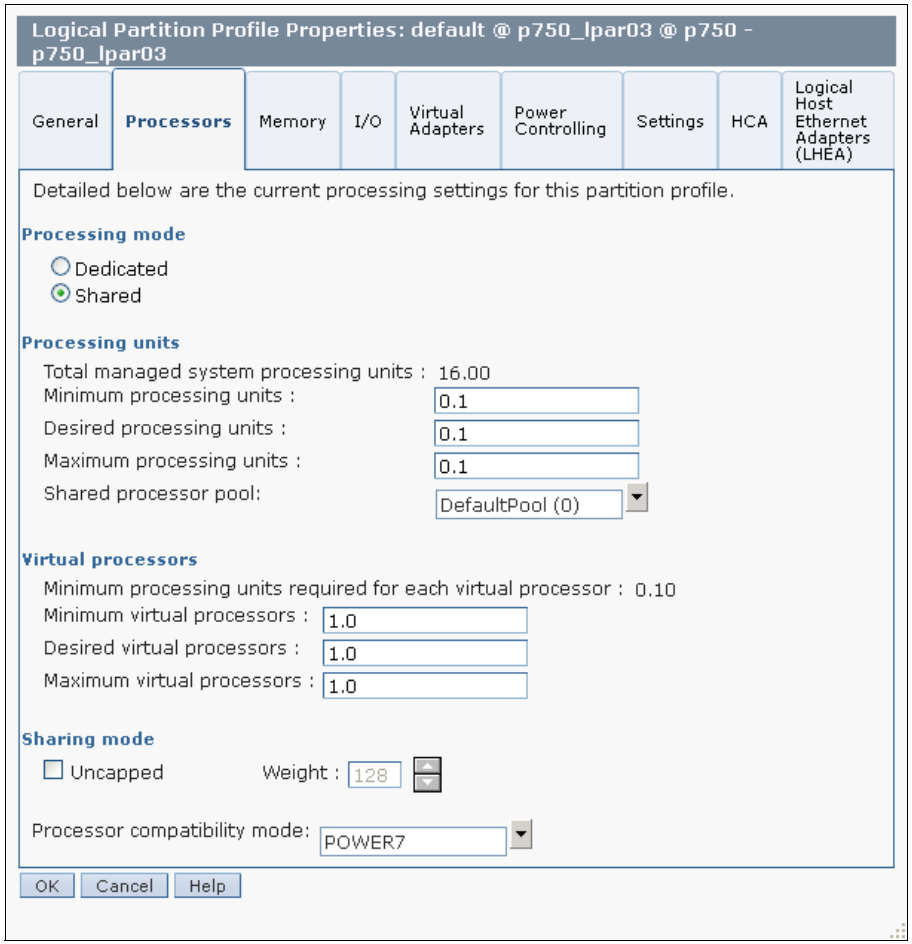


Figure 2-1 Changing dedicated processor partitions to micro-partitions

3. If the partition is running, deactivate and then reactivate it. If it is powered off, activate the partition. It is now using processing resources from the shared processor pool.

Modifying attributes of a micro-partition

You can perform several actions on a micro-partition to change its attributes. Some attributes can be changed through a dynamic LPAR operation. Others require a change in the partition profile, which requires deactivating and reactivating the partition.

With processing resources, you can dynamically change the processor entitlement, number of virtual processors, the sharing mode (either capped or uncapped), and the uncapped weight. For more information about how to change attributes dynamically, see Chapter 12, “Dynamic logical partitioning” on page 457.

To change the minimum, wanted, and maximum values for processing units and virtual processors, you must change the partition profile. For more information about how to access the panel to change these attributes, see “Changing dedicated processor partitions to micro-partitions” on page 12.

Retrieving partition information using the HMC

You can retrieve partition information from the HMC by using either the GUI or the command line. The **Properties** panel on the HMC shows information about the active configuration for the partition. It can be different from the partition profile information if the administrator ran a dynamic LPAR operation on this partition. To access the information, select the partition and select the **Properties** option on the menu. You can then navigate among the different tabs to see the details on the running configuration, as shown in Figure 2-2.

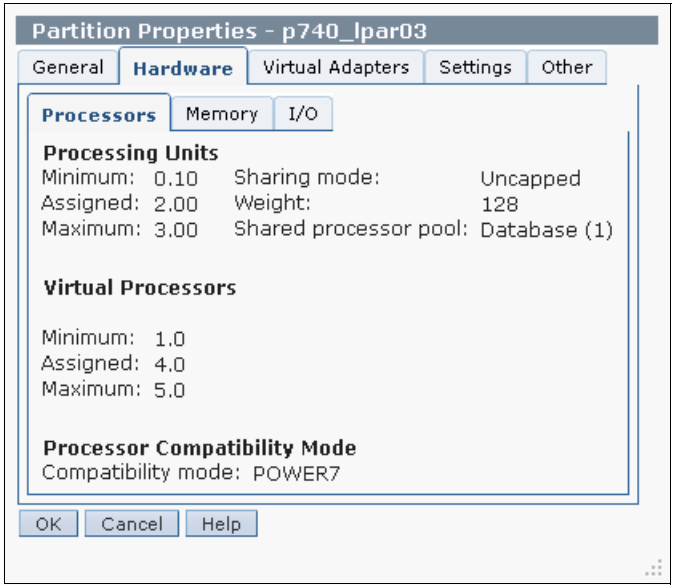


Figure 2-2 Partition properties panel that shows the memory configuration

You can also retrieve the partition profile information and current configuration from the HMC by using the command line, as shown in Example 2-1.

Example 2-1 Listing current partition configuration from the HMC command line

```
hscroot@hmc8:~> lshwres -m p740 --level lpar -r proc -F
lpar_name,lpar_id,curr_proc_mode,curr_min_proc_units,curr_proc_units,cu
rr_max_proc_units,curr_min_procs,curr_procs,curr_max_procs,curr_sharing
_mode,curr_uncap_weight --filter "lpar_names=p740_lpar03"
p740_lpar03,5,shared,0.1,2.0,3.0,1,4,5,uncap,128
```

Example 2-2 showing listing the partition profile configuration.

Example 2-2 Listing partition profile configuration from the HMC command line

```
hscroot@hmc8:~> lssyscfg -r prof -m p740 --filter
"lpar_names=p740_lpar03" -F
name,lpar_name,lpar_id,proc_mode,min_proc_units,desired_proc_units,max_
proc_units,min_procs,desired_procs,max_procs,sharing_mode,uncap_weight
default,p740_lpar03,5,shared,0.1,1.5,3.0,1,2,5,uncap,128
```

Retrieving partition information using the command line

The **lparstat** command lists the partition profile information and current configuration as shown in Example 2-3.

Example 2-3 Showing partition information using the lparstat command

```
p740_lpar03:/ # lparstat -i
Node Name                                  : p740_lpar03
Partition Name                             : p740_lpar03
Partition Number                           : 5
Type                                       : Shared-SMT-4
Mode                                       : Uncapped
Entitled Capacity                          : 2.00
Partition Group-ID                         : 32773
Shared Pool ID                             : 1
Online Virtual CPUs                        : 4
Maximum Virtual CPUs                       : 5
Minimum Virtual CPUs                       : 1
Online Memory                              : 12288 MB
Maximum Memory                             : 15360 MB
Minimum Memory                             : 512 MB
Variable Capacity Weight                   : 128
Minimum Capacity                           : 0.10
Maximum Capacity                           : 3.00
Capacity Increment                         : 0.01
```

```
Maximum Physical CPUs in system           : 16
Active Physical CPUs in system            : 16
Active CPUs in Pool                       : 5
Shared Physical CPUs in system            : 16
Maximum Capacity of Pool                  : 500
Entitled Capacity of Pool                 : 250
Unallocated Capacity                      : 0.00
Physical CPU Percentage                   : 50.00%
Unallocated Weight                        : 0
Memory Mode                               : Dedicated
Total I/O Memory Entitlement              : -
Variable Memory Capacity Weight           : -
Memory Pool ID                            : -
Physical Memory in the Pool               : -
Hypervisor Page Size                      : -
Unallocated Variable Memory Capacity Weight: -
Unallocated I/O Memory entitlement        : -
Memory Group ID of LPAR                   : -
Desired Virtual CPUs                      : 4
Desired Memory                            : 12288 MB
Desired Variable Capacity Weight          : 128
Desired Capacity                          : 2.00
Target Memory Expansion Factor            : -
Target Memory Expansion Size              : -
Power Saving Mode                         : Static Power Savings
```

The **nmon** command can also be used to retrieve partition information by using option **p** as shown in Figure 2-3.

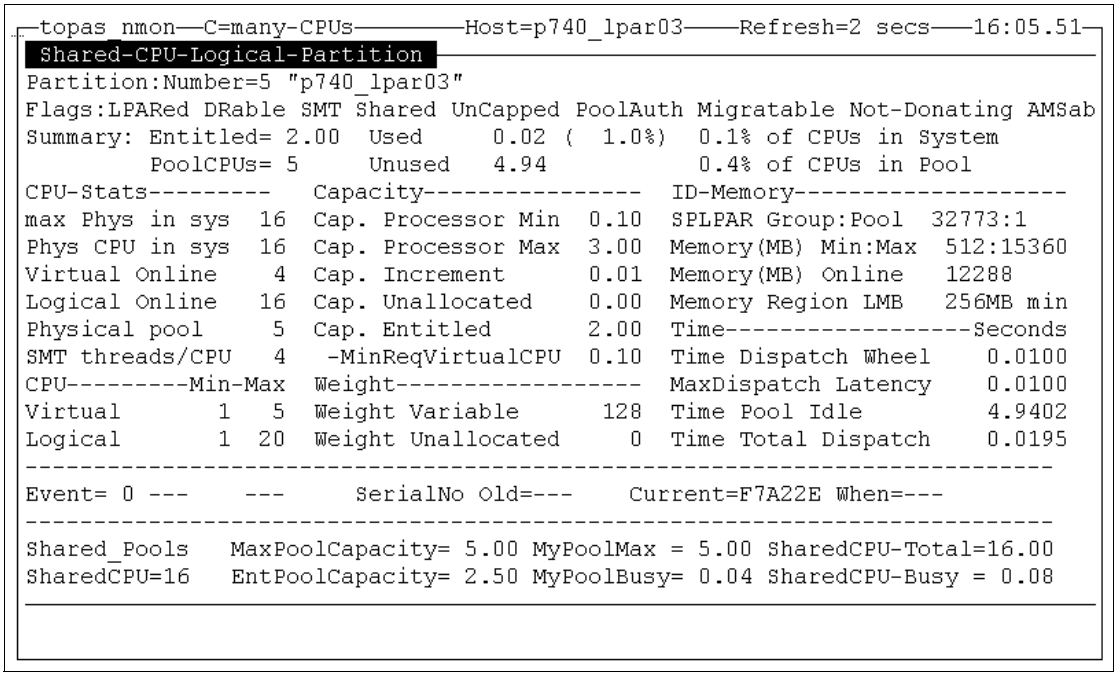


Figure 2-3 nmon command showing partition information

Note: Linux also provides the **lparstat** command as part of the **powerpc-utils** package available for Linux running on Power Systems. The **nmon** command is also available on Linux as part of the IBM Service and Productivity tools, available at:

https://www14.software.ibm.com/webapp/set2/sas/f/lopdiags/home.html.

2.2 Monitoring Shared Processor Pools

This section describes how to monitor Shared Processor Pools, processing units, simultaneous multithreading, and PowerVM processor terminologies. It also explains cross-partition monitoring and other processor monitoring tools that are used to monitor Virtual I/O Servers and virtual I/O clients with AIX, IBM i, and Linux operating systems.

2.2.1 Processor-related terminology and metrics

The IBM POWER5 and POWER6 processor-based managed systems introduced a high-level abstraction of processor consumption for sharing between multiple partitions as described in the following sections.

Common to POWER5 or later systems

On IBM POWER5 processor-based or later managed system, you can assign either dedicated processors to a logical partition or processing unit parts of the *Global Shared Processor Pool*.

The Global Shared Processor Pool consists of the processors that are not already assigned to running partitions with dedicated processors.

Consider the example that is illustrated in Figure 2-4 on page 19. The managed system has 16 processors. Two partitions (A and B) are running. Each partition is using three dedicated processors. The Shared Processor Pool therefore contains 16 - (2 * 3) = 10 processors.

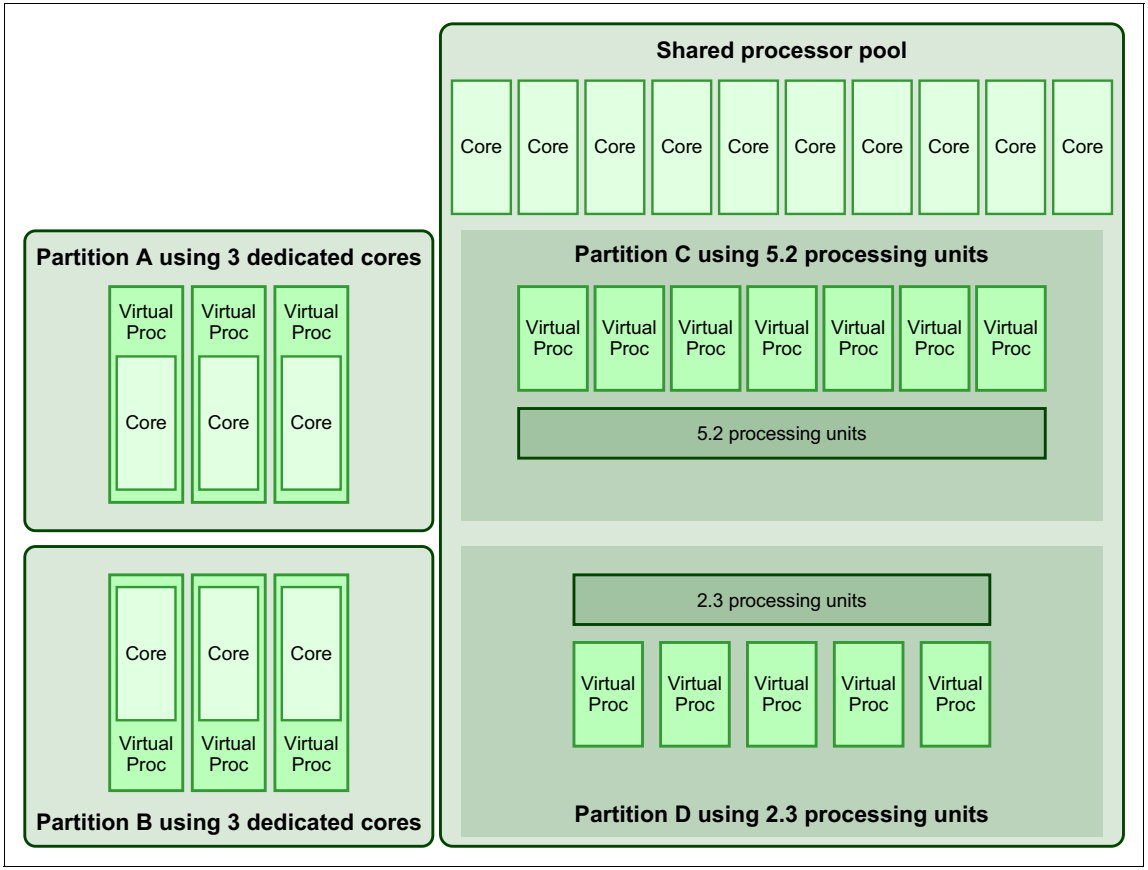


Figure 2-4 16-core system with dedicated and shared processors

Partitions that use the Shared Processor Pool are allocated processing capacity that is based on the minimum processor capacity that can be assigned to the partition. The minimum processing capacity that can be allocated per partition is based on the firmware of the managed system. Managed systems running firmware before 760 can allocate a minimum processing capacity of one-tenth (0.10) of a processing unit per partition. Managed systems running firmware 760 and later can allocate a minimum processing capacity of one-twentieth (0.05) of a processing unit per partition.

Note: The examples that follow are based on a minimum entitlement capacity of 0.1 processing units per processor.

In the example, the Shared Processor Pool contains 10 processors.

Partitions that use the Shared Processor Pool are also assigned some virtual processors. This represents the number of processors that are seen by the operating system running in the partition. The processing capacity is distributed among the virtual processors.

In the same example, define two partitions by using the Shared Processor Pool. The first partition (C) has 5.2 processing units and seven virtual processors. The second partition (D) has 2.3 processing units and five virtual processors. Note that 2.5 processing units are not allocated.

Another feature that was introduced with the IBM POWER5 processor-based managed systems is the ability of a partition to allocate more processing capacity than was originally assigned to it. Note the following points:

- *Capped* shared partitions cannot use more processing capacity than originally assigned.
- *Uncapped* shared partitions can use more processing capacity than originally assigned if it is available in the shared processor pool.

The number of processing units that are currently allocated to a partition represents its *entitlement*. If a partition is an uncapped shared partition, it might use more processing capacity than allocated. It can also use less processing capacity than allocated. This is why the metrics named *physical consumption* and *entitlement consumption* are defined as follows:

- Physical consumption represents the amount of processing capacity currently used.
- Entitlement consumption represents the percentage of processing capacity currently used compared to the entitle processing capacity allocated to the partition. Consequently, uncapped shared partitions can have an entitlement consumption that exceeds 100%.

An extra feature that is related to processor utilization was introduced with the IBM POWER5 architecture. The *simultaneous multithreading* (SMT) function can be activated in the operating system of a partition. When active, it allows the partition to run two, and on IBM POWER7 processor-based managed systems four, simultaneous threads on a single virtual processor. A virtual processor is then seen as two or four logical processors. When SMT is not enabled, a virtual processor is displayed as a single logical processor.

Table 2-1 provides a summary of the processor terminology and the metrics that are defined to monitor utilization.

Table 2-1 IBM POWER5 processor-based terminology and metrics

Term	Description	Related metrics
Dedicated processor	Processor that is directly allocated to a partition. Other partitions cannot use it.	Standard processor consumption metrics (user, sys, idle, wait, and so on)
Shared Processor	Processor part of the shared processor pool	Physical consumption
Processing unit	Power resource of a tenth of the combined processors in a shared processor pool	Physical consumption
Virtual processor	Processor as seen by a partition	► Simultaneous multithreading state ► Logical processor count
Logical processor	Processor as seen by a partition when simultaneous multithreading is on	► simultaneous multithreading state ► Logical processor count
Simultaneous multithreading	Capacity for a partition to run two threads simultaneously on a virtual processor	► Simultaneous multithreading state ► Logical processor count

Specific to IBM POWER6 processor-based or later systems

Several enhancements that were provided in the IBM POWER6 architecture introduced new functions and the related metrics.

Two reasons for using several Shared Processor Pools are subcapacity licensing and reserving processing units. Define maximum processing units for a Shared Processor Pool to ensure that the total processing unit consumption of the partitions does not exceed the maximum processing unit value. This value is used to compute the processor value unit (PVU) license fees for IBM products.

You can also reserve processing units for a Multiple Shared Processor Pool. This configuration ensures that processing units are available to be distributed among the uncapped shared partitions in the pool.

Specific to IBM POWER7+ processor-based or later systems

Defining the unit reserve is illustrated in Figure 2-5. In this figure, you can see that an extra shared processor pool is defined. This pool can use up to three processors. Two partitions (E and F) are defined in this Shared Processor Pool. Partition E has an entitlement of 0.8 processors. Partition F has an entitlement of 0.3 processors.

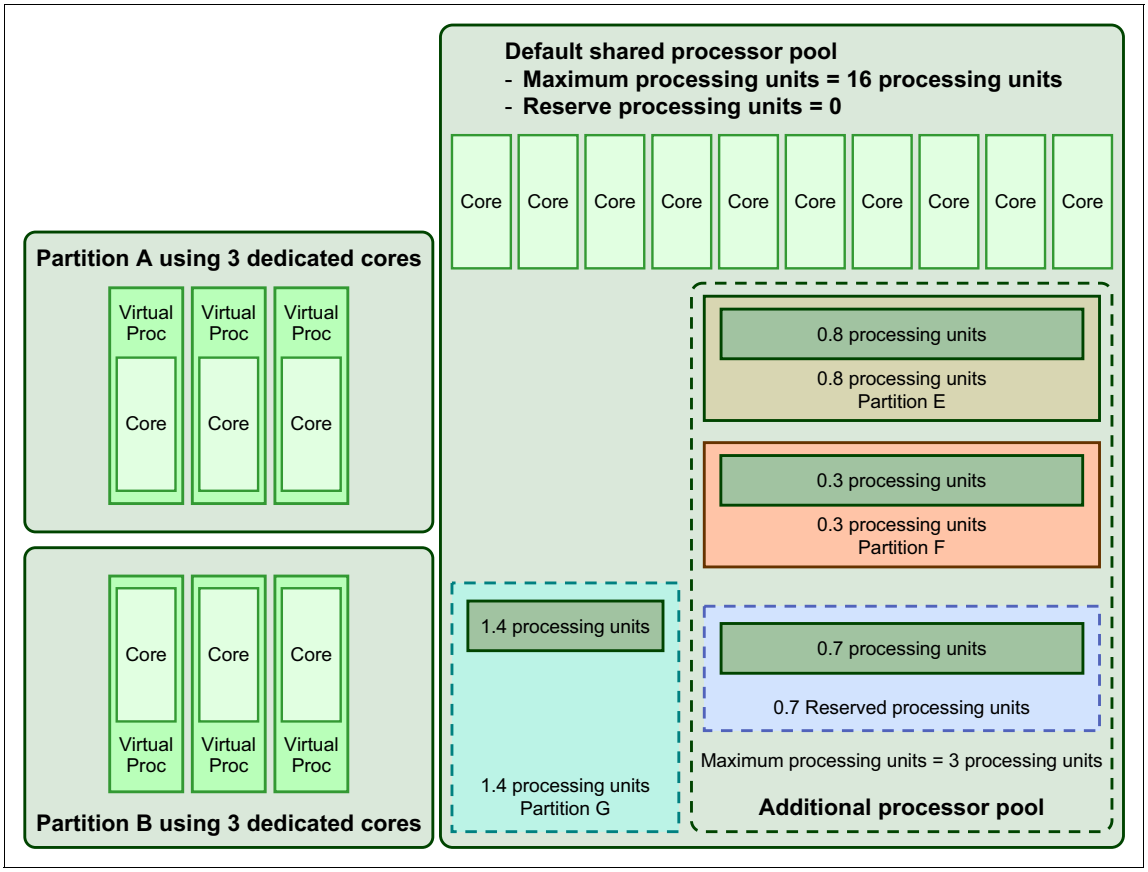


Figure 2-5 A Multiple Shared Processor Pool example on POWER6

The *Maximum processing units* value does not ensure the processing capacity, but the partitions in the Shared Processor Pool cannot exceed the maximum value.

The *Reserved processing units* value ensures the processing capacity for all partitions in the Shared Processor Pool. The *reserved processing units* cannot be configured for other partitions (Figure 2-6).

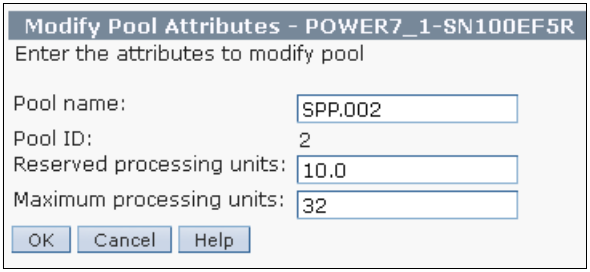


Figure 2-6 Shared Processor Pool attributes

Another enhancement that was introduced with IBM POWER6 architecture is Shared Dedicated Capacity for partitions that use dedicated processors. Upon activation, if a partition does not use 100% of its dedicated processor resources, then unused processor resources are ceded to the Shared Processor Pools. These processors are called Donating Dedicated processors.

The additional terminologies and related metrics for IBM POWER6 processor-based or later managed systems are shown in Table 2-2.

Table 2-2 IBM POWER6 or later system-specific terminology and metrics

Term	Related metric
Multiple shared processor pool	Pool size Maximum capacity Pool entitlement Reserve = bucket size Available capacity
Dedicated shared processors	Donation mode Donated cycles count

There are many tools that can provide processor consumption metrics on the Virtual I/O Server and its clients. Depending on the requirements, the tools that are described in subsequent sections can be used.

2.2.2 Processor metrics computation

The previous mechanism of calculating performance that was based on sampling is not accurate in the case of virtual processor resources. This inaccuracy is because one processor can be shared among more than one partition. It also

fails for SMT-enabled processors. The uncapping feature also causes difficulties because a partition can go beyond its capacity. This makes it difficult to determine when a processor is 100% busy, unlike with capped partitions.

The POWER5 family of processors addressed this issue by implementing a new performance-specific register called the Processor Utilization Resource Register (PURR). PURR tracks real processor resource usage on a per-thread or per-partition basis. AIX performance tools have been updated in AIX V5.3 to show these new statistics.

Traditional performance measurements were based on a 100-Hz sample rate (each sample corresponded to a 10-ms tick). Each sample was sorted into either a user, system, iowait, or idle category based on the code it was running when interrupted. This sampling-based approach does not work in a virtualized environment because the dispatch cycle of each virtual processor is no longer the same (which was the assumption in traditional performance measurement). A similar issue exists with SMT: If one thread is using 100% of the time on a physical processor, sample-based calculations report the system as 50% busy (one processor at 100% and another at 0%). To preserve binary compatibility, the traditional mechanism has not been changed.

Processor Utilization Resource Register (PURR)

PURR is a 64-bit counter with the same units for the timebase and decrementer registers that provides per-thread processing time statistics. Figure 2-7 on page 25 shows the relationship between PURR registers in a single POWER5 processor and the two hardware threads and two logical processors. With SMT enabled, each hardware thread is seen as a logical processor.

The *timebase* register shown in Figure 2-7 is incremented at each tick. The *decrementer* register provides periodic interrupts.

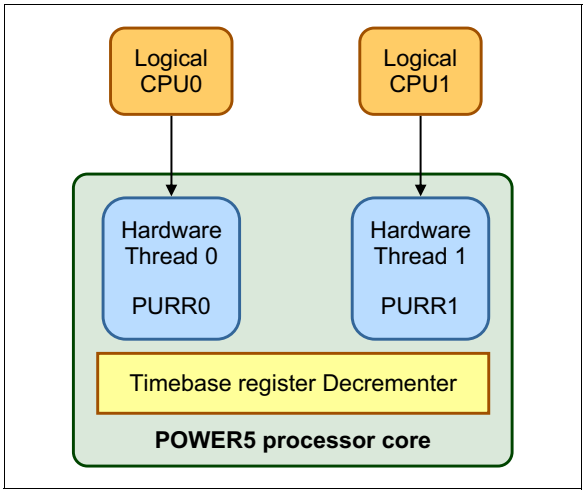


Figure 2-7 Per-thread PURR

At each processor clock cycle, the PURR that last ran (or is running) the instruction is incremented. The sum of two PURRs equals the value of timebase register. This approach is an approximation because SMT allows both threads to run in parallel. Therefore, it cannot be evolved to distinguish the performance difference between SMT on mode and SMT off mode.

PURR-based metrics

The PURR registers provide statistics that might be helpful when you look at the output of performance tools.

Processor statistics in SMT environment

The ratio of *(delta PURR)/(delta timebase)* over an interval indicates the fraction of physical processor that is used by a logical processor. This is the value that is returned by the **`sar -P ALL`** and **`mpstat`** commands, as shown in subsequent sections.

The value of *(delta PURR/delta timebase)*100* over an interval provides the percentage of physical processor that is used by a logical processor. This value is returned by the **`mpstat -s`** command, which shows the SMT statistics as explained in subsequent sections.

Processor statistics in shared processor partitions

In a shared processor environment, the PURR measures the time that a virtual processor runs on a physical processor. With SMT on, virtual time base is the

sum of the two PURRs. With SMT off, virtual time base is the value that is stored in the PURR. The PURR calculation differs on capped and uncapped shared processors:

- Capped shared processors

  For capped shared processors, the *entitled PURR* over an interval is given as *entitlement * time base*.

  To calculate %user time over an interval is (and %sys and %iowait):

  %user = (delta PURR in user mode/`entitled PURR`*) * 100*

- Uncapped shared processors

  For uncapped shared processors, the calculations take the variable capacity into account. The *entitled PURR* in the preceding formula is replaced by the *consumed PURR* whenever the latter is greater than the entitlement. So the calculation of %user time over an interval is:

  %user = (delta PURR in user mode/`consumed PURR`*) * 100*

Physical processor consumption for a shared processor

A partition's physical processor consumption is the sum of all its logical processor consumption:

SUM(delta PURR/delta timebase)

Partition entitlement consumption

A partition's entitlement consumption is the ratio of its physical processor consumption (PPC) to its entitlement expressed as a percentage:

*(Physical processor consumption/entitlement) * 100*

Shared processor pool spare capacity

Unused cycles in a shared processor pool are spent in the IBM POWER Hypervisor™'s idle loop. The POWER Hypervisor enters this loop when all partition entitlements are satisfied and there are no partitions to dispatch. The time spent in Hypervisor's idle loop, which is measured in ticks, is called the Pool Idle Count. The Shared processor pool spare capacity over an interval is expressed as:

(delta Pool idle count/delta timebase)

This statistic is measured in numbers of processors. Only partitions with shared processor pool authority are able to display this figure.

Logical processor utilization

Logical processor utilization is the sum of traditional 10 ms tick-based sampling of the time spent in %sys and %user. If it starts approaching 100%, it indicates that the partition can use more virtual processors.

System-wide tools that are modified for virtualization

The AIX tools **topas**, **lparstat**, **vmstat**, **sar**, and **mpstat** now add two extra columns of information when they are run in a shared-processor partition:

- Physical processor consumed by the partition, which is shown as `pc` or `%physc`.
- Percentage of entitled capacity consumed by the partition, which is shown as `ec` or `%entc`.

IBM i can display the current virtual processors and processing capacity by using the WRKSYSACT command as shown in Figure 2-15 on page 58.

The logical processor tools are the **mpstat** and **sar -P ALL** commands. When running in a partition with SMT enabled, these commands add the column Physical Processor Fraction Consumed (PPFC), shown as `physc` and calculated by *(delta PURR/delta timebase)*. This shows the relative split of physical processor time for each of the two logical processors.

When running in a shared processor partition, these commands add a column, Percentage of Entitlement Consumed *((PPFC/ENT)*100)* shown as `%entc`. It gives relative entitlement consumption for each logical processor expressed as a percentage.

Scaled Processor Utilization Resource Register (SPURR)

The IBM POWER6 and later microprocessor chips support advanced, dynamic power management solutions for managing not just the chip but the managed system. This design facilitates a programmable power management solution for greater flexibility and integration into system- and data center-wide management solutions.

The design of these microprocessors provides real-time access to detailed and accurate information about power, temperature, and performance. Their sensing, actuation, and management support is known as the IBM EnergyScale™ architecture. It enables higher performance, greater energy efficiency, and power management capabilities such as power and thermal capping, and power savings with explicit performance control.

The EnergyScale architecture and statistics computations

The EnergyScale implementation is primarily an out-of-band power management design. However, managing system-level power and temperature has effects on in-band software. Basically, power management results in performance variability. This implies that, as the power management implementation operates, it can change the effective speed of the processor.

To account for below-nominal-frequency usage of a processor by a program because of power management actions, an extra special-purpose register for each hardware thread known as the Scaled Processor Utilization Resource Register (SPURR) was introduced.

The SPURR is used to compensate for the effects of performance variability on the operating systems. The hypervisor virtualizes the SPURR for each hardware thread so that each OS obtains accurate readings that reflect only the portion of the SPURR count associated with its partition. The implementation of virtualization for the SPURR is the same as that for the PURR.

The POWER hypervisor provides functions on which operating systems can build and use SPURR to do accurate accounting similar to POWER5 processor-based managed systems. With the EnergyScale architecture for POWER6 and newer processor-based managed systems, not all timebase ticks have the same computational value. Some of them represent more usable processor cycles than others. The SPURR provides a scaled count of the number of timebase ticks assigned to a hardware thread. The scaling reflects the speed of the processor (taking into account frequency changes and throttling) relative to its nominal speed.

System-wide tools modified for variable processor frequency

The EnergyScale architecture can affect performance tools and metrics that are built with the user-visible performance counters. Many of these tools count processor cycles. Because the number of cycles per unit time is variable, the values reported by unmodified performance monitors must be interpreted.

The **lparstat** command has been updated in AIX Version 6.1 to display statistics if the processor is not running at nominal speed. The *%nsp* metric shows the current average processor speed as a percentage of nominal speed. This field is also displayed by the updated version of the **mpstat** command.

Note: The *%nsp* metric is seen only when Power Management is enabled on the managed system.

The **lparstat** command also displays statistics if Turbo mode accounting is disabled (the default) and the processor is running above nominal speed. The *%utcyc* field reflects the total percentage of unaccounted turbo cycles.

Finally, the **lparstat -d** command displays unaccounted turbo cycles in user, kernel, idle, and I/O wait modes if Turbo mode accounting is disabled and the processor is running above nominal speed.

All of the other stat commands automatically switch to use SPURR-based metrics. Percentages below entitlement become relative to scaled (up and down) entitlement unless Turbo mode accounting is off.

2.2.3 Cross-partition processor monitoring

This section describes cross-partition processor monitoring from AIX, Virtual I/O Server, and IBM i partitions.

Monitoring from AIX and Virtual I/O Server

Recent versions of the **topas** tool available on AIX and the Virtual I/O Server provide monitoring of cross-partition processor consumption. This command sees only partitions that are running AIX V5.3 TL3 or later. Virtual I/O Servers with Version 1.3 or later are also reported. Cross-partition processor monitoring requires the `perfagent.tools` and `bos.perf.tools` file sets, which are installed by default on a base AIX or Virtual I/O Server installation. For more information about cross-partition monitoring with **topas**, see the Information Center website at:

http://pic.dhe.ibm.com/infocenter/aix/v6r1/topic/com.ibm.aix.prftungd/doc/prftungd/view_cross-partition_panel.htm

To see a cross-partition report from the Virtual I/O Server, run **topas -cecdisp** as shown in Example 2-4.

Example 2-4 topas -cecdisp command on Virtual I/O Server

```
Topas CEC Monitor             Interval:  10             Mon Dec 3 14:39:18 2012
Partitions Memory (GB)           Processors
Shr:  3    Mon:11.5  InUse: 3.2  Shr:1.5  PSz:  3   Don: 0.0 Shr_PhysB  0.21
Ded:  1    Avl:   -              Ded:  1  APP:  2.8 Stl: 0.0 Ded_PhysB  0.00

Host         OS  M Mem InU Lp  Us Sy Wa Id  PhysB  Vcsw Ent  %EntC PhI
-----------------------------------shared-------------------------------------
NIM_server   A61 C 2.0 1.1  4  97  1  0  0   0.20  386  0.20  99.6   0
DB_server    A61 U 4.5 0.8  4   0  0  0 99   0.01  403  1.00   0.6   2
VIO_Server1  A53 U 1.0 0.4  2   0  0  0 99   0.00  212  0.30   1.2   0
```

```
Host         OS  M Mem InU Lp  Us Sy Wa Id  PhysB  Vcsw  %istl %bstl
-------------------------------------dedicated-----------------------------------
Apps_server  A61 S 4.0 0.8  2   0  0  0 99   0.00  236    0.00  0.00
```

After the **topas** command runs, keyboard shortcuts behave as in a standard AIX partition.

Run **topas -C** as shown in Example 2-5 to view a cross-partition report from the virtual I/O AIX client.

Example 2-5 topas -C command on virtual I/O client

```
Topas CEC Monitor             Interval:  10             Mon Dec 3 14:40:07 2012
Partitions Memory (GB)           Processors
Shr:  4    Mon:11.5  InUse: 4.7  Shr:1.5  PSz:  3   Don: 0.0 Shr_PhysB  1.17
Ded:  1    Avl:   -              Ded:  1  APP:  1.8 Stl: 0.0 Ded_PhysB  0.00

Host         OS  M Mem InU Lp  Us Sy Wa Id  PhysB  Vcsw Ent  %EntC PhI
-------------------------------------shared------------------------------------
DB_server    A61 C 4.5 0.9  4  99  0  0  0   1.00  405  1.00  99.7  38
NIM_server   A61 C 2.0 2.0  4  50 27  2 19   0.16  969  0.20  80.9  11
VIO_Server1  A53 U 1.0 1.0  2   0  3  0 96   0.01  667  0.30   4.5   0
VIO_Server2  A53 U 1.0 1.0  2   0  1  0 98   0.01  358  0.30   2.7   0
Host         OS  M Mem InU Lp  Us Sy Wa Id  PhysB  Vcsw  %istl %bstl
-------------------------------------dedicated-----------------------------------
Apps_server  A61 S 4.0 0.9  2   0  0  0 99   0.00  332    0.00  0.00
```

You can obtain extensive information about a system and its partitions from this report. The report shown in Example 2-5 contains the following fields:

- `Partitions Shr=4, Ded=1`

  The system has five active partitions, and four use the shared processor pools.

- `Processors Psz = 3`

  There are three processors in the same pool as this partition.

- `Processors Shr = 1.5`

  Within the shared processor pool, 1.5 processing units are currently allocated to the partitions.

- `Processors Ded = 1`

  There is one processor that is used for dedicated partitions.

- `Processors APP = 1.8`

  Out of three processors in the shared processor pool, 1.8 are considered to be idle. Because Shr = 1.5 and 3 - 1.8 < 1.5, you can be conclude that the shared partitions globally do not use all of their entitlements (allocated processing units).

- `Processors Shr_PhysB = 1.17`

  This represents the number of processors that are busy for all shared partitions. It confirms what is observed with the APP field. The shared processor pool processor consumption is lower than the entitlement of the shared partitions.

- `Processors Ded_PhysB = 0.00`

  The dedicated partition is currently idle.

- `Processors Don = 0.0`

  This is related to a feature introduced with IBM POWER6 architecture. It represents the number of idle processor resources from the dedicated partitions that are donated to the shared processor pools.

- `Stl = 0`

  Sum of stolen processor cycles from all partitions, reported as a number of processors.

- `%istl = 0.00`

  This shows the percentage of physical processors that is used while idle cycles are being stolen by the hypervisor. This metric is applicable only to dedicated partitions.

- `%bstl = 0.00`

  This shows the percentage of physical processors that is used while busy cycles are being stolen by the hypervisor. This metric is applicable only to dedicated partitions.

To activate idle processor resources donation for dedicated processor partitions, select **Processor Sharing** in the Hardware tab from the partition properties on the HMC as shown in Figure 2-8. For IBM POWER5 processor-based managed systems, processor sharing is only possible for inactive partitions. For IBM POWER6 processor-based or later managed systems, processor sharing is also possible for active partitions.

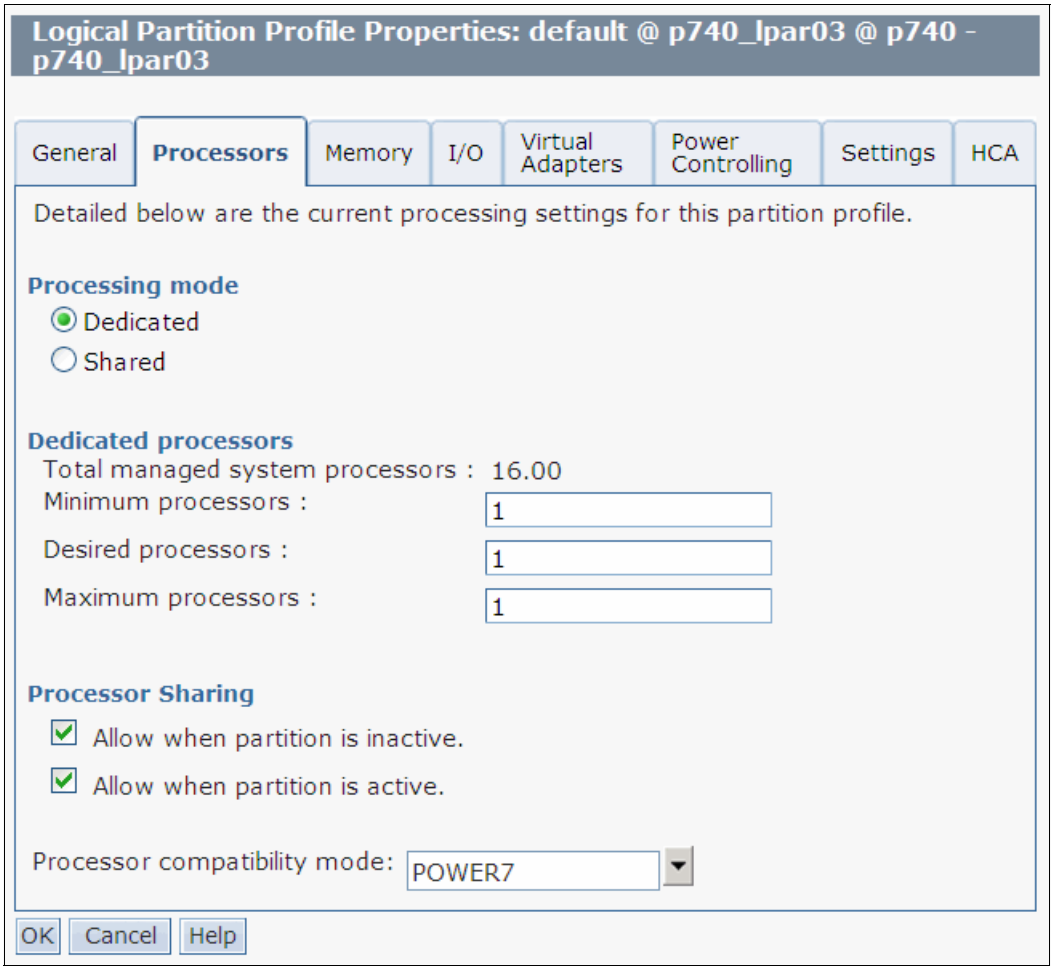


Figure 2-8 Dedicated partition's Processor Sharing properties

The partition section lists all the partitions that **topas** command can find in the CEC.

- The OS column indicates the type of operating system. In Example 2-5 on page 30, A61 indicates AIX Version 6.1.
- The M column shows the partition mode.
  - For shared partitions:
    - C: SMT enabled and capped
    - c: SMT disabled and capped
    - U: SMT enabled and uncapped
    - u: SMT disabled and uncapped
  - For dedicated partitions:
    - S: SMT enabled
    - d: SMT disabled and donating
    - D: SMT enabled and donating
    - '' (blank): SMT disabled and not donating
- The other values are equivalent to those provided in the standard view.

Pressing the **g** key in the **topas** command window expands global information, as shown in Example 2-6. A number of fields are not completed in this example, such as the total available memory. These fields are reserved for an update to this command that allows **topas** to interrogate the HMC to determine their values. It is possible to manually specify some of these values on the command line.

Example 2-6 topas -C command global

```
Topas CEC Monitor            Interval:  10             Mon Dec 3 15:10:19 2012
Partition Info    Memory (GB)        Processor         Virtual Pools :     0
Monitored  :  4   Monitored  :12.5   Monitored  :2.7   Avail Pool Proc:    0.1
UnMonitored:  -   UnMonitored:   -   UnMonitored:  -   Shr Physical Busy:  3.95
Shared     :  3   Available  :   -   Available  :  -   Ded Physical Busy:  0.00
Uncapped   :  2   UnAllocated:   -   UnAllocated:  -   Donated Phys. CPUs  1.00
Capped     :  2   Consumed   : 5.7   Shared     :1.7   Stolen Phys. CPUs : 0.00
Dedicated  :  1                     Dedicated  :  1   Hypervisor
Donating   :  1                     Donated    :  1   Virt. Context Switch:1026
2                                   Pool Size  :  4   Phantom Interrupts  :  11

Host         OS  M Mem InU Lp  Us Sy Wa Id  PhysB  Vcsw Ent  %EntC PhI
-------------------------------------shared-------------------------------------
DB_server    A61 U 4.5 0.9  8  99  0  0  0   3.93 9308  1.00 393.4  11
NIM_server   A61 C 3.0 2.9  4   0  1  0 98   0.01  556  0.40   2.4   0
VIO_Server1  A53 U 1.0 1.0  2   0  0  0 99   0.00  168  0.30   0.9   0
```

```
Host         OS  M Mem InU Lp  Us Sy Wa Id  PhysB  Vcsw  %istl %bstl %bdon %idon
-------------------------------------dedicated-----------------------------------
Apps_server  A61 D 4.0 0.8  2   0  0  0 99   0.00  230    0.01  0.00  0.00 99.69
```

Tip: **`topas -C`** might not be able to locate partitions on other subnets. To circumvent this, create a `$HOME/Rsi.hosts` file that contains the fully qualified host names for each partition (including domains), one host per line.

The cross-partition monitoring feature is not provided by other processor performance monitoring tools like **`lparstat`**, **`vmstat`**, **`sar`**, and **`mpstat`**.

Cross-partition processor monitoring from IBM i

Collection Services on IBM i 6.1 or later can collect high-level performance data for all logical partitions on a single physical server, regardless if they are running an AIX, IBM i, or Linux operating system. An IBM i partition retrieves this information by directly communicating with the POWER Hypervisor. It requests partition configuration and utilization data at each Collection Services interval, and stores it in a management collection object. Performance data can be converted into a structured database file, similar to other Collection Services data, from the management collection object. The cross-partition performance data is stored in the QAPMLPARH performance database file.

Consideration: Enabling cross-partition processor performance data collection on IBM i 6.1 or later requires a POWER6 processor-based managed system or later, running firmware xx340_061 or later.

You only need to collect data on one partition per system, and it must be an IBM i partition.

Use the **`CFGPFRCOL`** command to change the collection interval.

Select **Allow performance information collection** to enable collection of cross-partition processor performance data as shown in Figure 2-9.

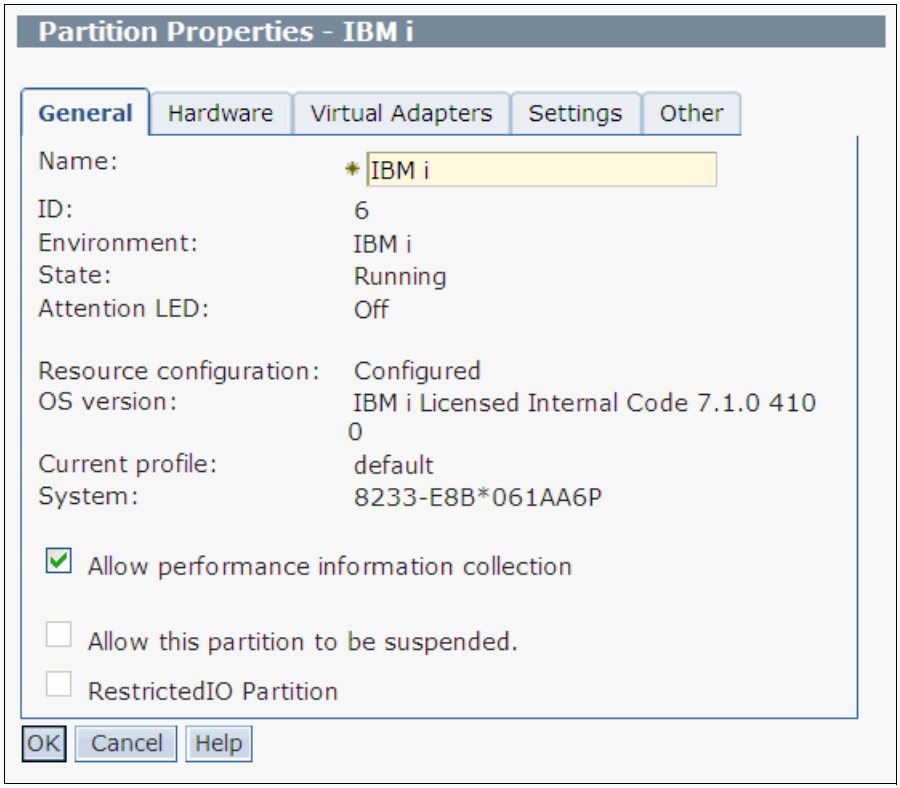


Figure 2-9 Allow performance information collection on IBM i

The collection of processor performance data from multiple partitions that are provided by the POWER Hypervisor is a powerful feature for granular monitoring of the total server processor utilization. None of the other operating systems currently have a similar ability to collect and investigate this data. This restriction is especially of concern if you require inclusion of IBM i partition processor performance data.

For more information about the cross-partition processor performance data that are collected in the QAPMLPARH database file, see the IBM i Information Center at:

http://pic.dhe.ibm.com/infocenter/iseries/v6r1m0/topic/rzahx/rzahxqapmlparh.htm

The data from the QAPMLPARH file can either be analyzed by user-written SQL queries or investigated by using the IBM Systems Director Navigator for i GUI. To

use the GUI, click **Performance** → **Investigate Data**, which provides the following perspectives under **Collection Services** → **Physical System**:

- Logical Partitions Overview
- Donated Processor Time by Logical Partition
- Uncapped Processor Time Used by Logical Partition
- Virtual Shared Processor Pool Utilization
- Physical Processors Utilization by Physical Processor
- Dedicated Processors Utilization by Logical Partition
- Physical Processors Utilization by Processor Status Overview
- Physical Processors Utilization by Processor Status Detail
- Shared Memory Overview

Figure 2-10 shows an example of the Logical Partitions Overview window that shows configuration data and processor utilization for all logical partitions on the system. The system can be running IBM i, AIX, Virtual I/O Server, or Linux.

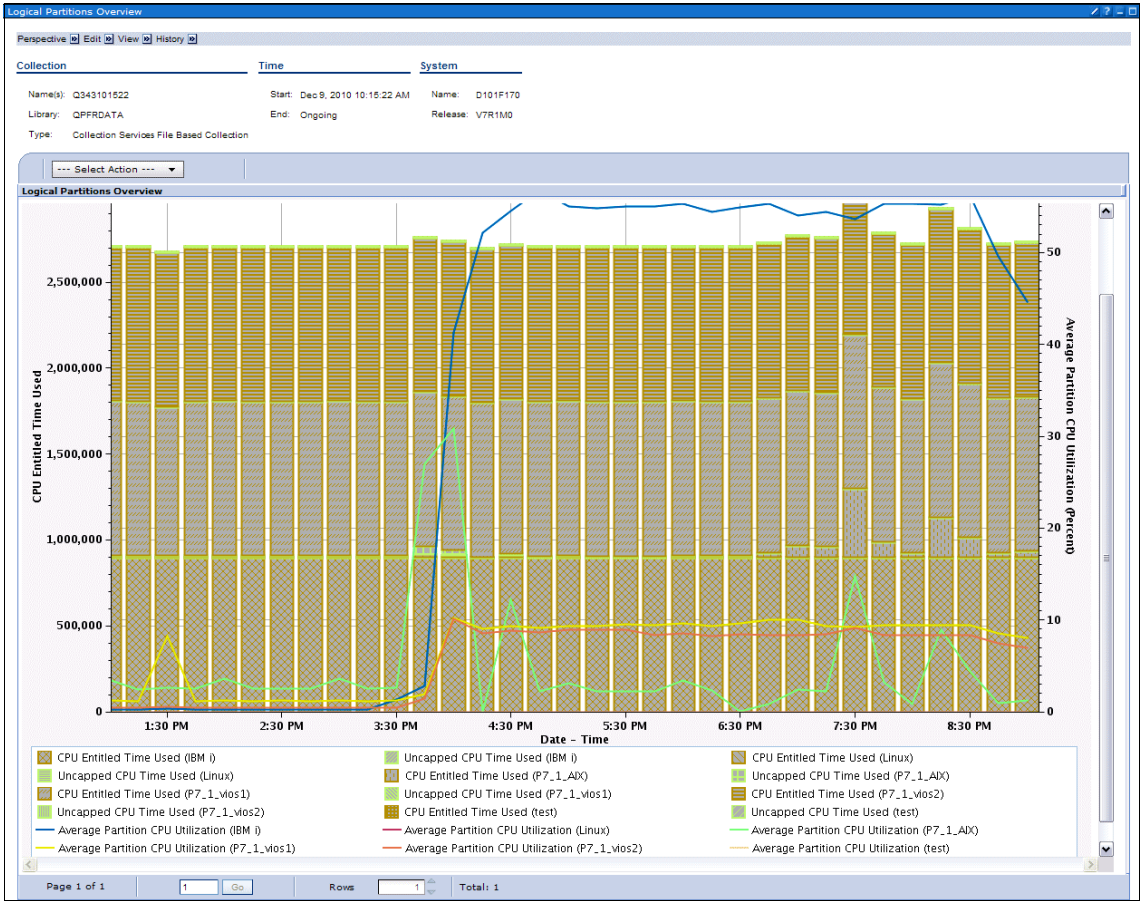


Figure 2-10 IBM Systems Director Navigator for i Logical Partitions Overview

2.2.4 AIX and Virtual I/O Server processor monitoring

There are several commands for monitoring processor utilization for the Virtual I/O Server and virtual I/O client (AIX). Each command is covered separately. For examples that use the Virtual I/O Server, it is assumed that the command is running from the restricted shell, not from the root shell.

Monitoring using topas

The **topas** command gathers general information from metrics on a partition. The major advantage of this tool is that you can see all of the important performance metrics of the partition in real time.

Basic display

In basic display mode, **topas** is run without any argument as shown in Example 2-7.

Example 2-7 Basic topas monitoring

```
Topas Monitor for host:    DB_server           EVENTS/QUEUES    FILE/TTY
Fri Dec  7 10:27:10 2012 Interval:  2         Cswitch     590  Readch     2735
                                               Syscall  971.0K  Writech   72691
CPU  User%  Kern%  Wait%  Idle%  Physc   Entc   Reads         2  Rawin         0
ALL   18.2   44.9    0.0   36.9   1.01  99.9    Writes    71274  Ttyout      176
                                               Forks         0  Igets         0
Network  KBPS   I-Pack  O-Pack   KB-In  KB-Out  Execs         0  Namei     71271
Total    47.0     47.5    10.0    35.0    12.0  Runqueue    1.0  Dirblk        0
                                               Waitqueue   0.0
Disk    Busy%     KBPS     TPS KB-Read KB-Writ                   MEMORY
Total     0.0      3.5     1.0     2.5     1.0  PAGING           Real,MB    4096
                                               Faults        0  % Comp       28
FileSystem        KBPS     TPS KB-Read KB-Writ  Steals        0  % Noncomp    5
Total              0.1     0.5    0.1    0.0    PgspIn        0  % Client     5
                                               PgspOut       0
Name            PID  CPU%  PgSp Owner           PageIn        0  PAGING SPACE
ksh         8650922  62.3   0.5 root            PageOut       0  Size,MB    1536
topas       9699434   7.6   1.8 padmin          Sios          0  % Used       1
vmmd         458766   0.2   1.2 root                             % Free      99
lock_rcv    3342452   0.0   0.4 root            NFS (calls/sec)
cimserve    7799026   0.0  53.9 root            SerV2         0  WPAR Activ
SNKP        2424938   0.0   0.5 root            CliV2         0  WPAR Total
NCKP        3014754   0.0   0.4 root            SerV3         0  Press: "h"-help
xmtopas     9109612   0.0   1.0 root            CliV3         0         "q"-quit
Fri Dec  7 10:21:06 2012 Interval:  2         Cswitch     623  Readch     2754
                                               Syscall     247  Writech     966
```

You immediately get the most important processor information for this partition:

- `Entc` = 99.9 and remains steady upon refresh (every 2 seconds)

  This partition therefore uses all its entitlement.

- `Physc` = 1.0

  This partition is consuming one physical processor.

- `Kern%` = 44.9%

  Most of the computation is spent in kernel mode.

- `Runqueue` = 1

  On average over 2s, only one thread is ready to run.

- `Waitqueue` = 0

  On average over 2s, no thread is waiting for paging to complete.

Logical Partition display

Example 2-8 assume a process is running with a processor utilization of 49.7%. The other processes on the managed system are idle, and `Physc = 1.00`. In this scenario, `Simultaneous Multi-Threading` is probably enabled and the **`ksh`** process is not multi-threaded. You can check this by pressing **`L`** to toggle to the logical partition display, as shown in Example 2-8. Alternatively, this view can also be accessed by running **`topas -L`**.

The upper section shows a subset of the **`lparstat`** command statistics. The lower section shows a sorted list of logical processors with a number of the **`mpstat`** command figures.

Example 2-8 Logical partition information report in topas (press L)

```
Interval:    2    Logical Partition: DB_server         Mon Dec 3 16:27:01 2012
Psize:       -               Shared SMT  ON            Online Memory:   4608.0
Ent: 1.00                     Mode: Capped             Online Logical CPUs:   4
Partition CPU Utilization                              Online Virtual CPUs:   2
%usr %sys %wait %idle physc  %entc %lbusy    app    vcsw    phint   %hypv   hcalls
  99    0     0     0   1.0  99.90  25.00      -     418       26     0.4      397
================================================================================
LCPU  minpf majpf  intr    csw   icsw runq lpa  scalls usr sys _wt idl   pc   lcsw
Cpu0      0     0   175     28     14    0 100      13   5  50   0  46 0.00    121
Cpu1      0     0    94      4      2    0 100       3   4  33   0  63 0.00     94
Cpu2      0     0    93     26     18    0 100      13   9  27   0  64 0.00    102
Cpu3      0     0   128      2      2    0 100       0 100   0   0   0 1.00    100
```

This additional report confirms the initial assumptions:

- `Shared SMT` = ON

  This confirms that SMT is active on this partition.

- For the two virtual processors that are allocated to this partition, the SMT provides four logical processors.

- `%lbusy` = 25%

  Only a quarter of the logical processors are effectively in use.

- The detailed logical processor activity listing displayed at the bottom of the report also shows that only one of the logical processors is in use. This confirms that the process in the scenario is probably single-threaded. Use the **ps** command to get a deeper analysis of this process consumption.

Processor subsection display

If you are interested in individual logical processor metrics, pressing C twice in the main report window displays the individual processor activity report as shown in Example 2-9.

Example 2-9 Upper part of topas busiest processor report

```
Topas Monitor for host:    DB_server           EVENTS/QUEUES    FILE/TTY
Fri Dec  7 10:34:34 2012 Interval:  2         Cswitch      61  Readch        2
                                              Syscall      35  Writech     115
CPU      User%   Kern%   Wait%   Idle%  Physc  Reads         1  Rawin         0
cpu2      99.9     0.1     0.0     0.0   1.00  Writes        1  Ttyout      113
cpu1       4.6    35.4     0.0    60.1   0.00  Forks         0  Igets         0
cpu0       3.6    53.8     0.0    42.6   0.00  Execs         0  Namei         2
cpu3       0.0    17.8     0.0    82.2   0.00  Runqueue    1.0  Dirblk        0
```

In this example, a single logical processor is handling the whole partition load. By observing the values, you can pinpoint performance bottlenecks. As shown in Example 2-7 on page 38, `Entc` is about 100% and does not exceed this value. You can then make the following suppositions:

- Because this partition also has `Physc` = 1.0, it can run on a dedicated processor.
- If it runs in a shared processor pool, it might be capped.
- If it runs in a shared processor pool, it might be uncapped but only have a single virtual processor defined. A virtual processor cannot consume more processing units than a physical processor.
- If it runs in a shared processor pool, it might be uncapped but have reached the maximum processing units for this partition as defined in its active profile.
- If it runs in a shared processor pool, it might be uncapped but have reached the maximum processing units for the shared processor pool.

To move to the logical partition report, press Shift+L. In Example 2-8 on page 39, the partition is capped and thus running in shared mode.

In Example 2-8 on page 39, the `Psize` field is not filled in. This value represents the global shared pool size on the managed system, which is the number of processors that are not used for running dedicated partitions. To get this value, you must modify the partition properties from the HMC or the IVM. Then, click the

Hardware tab and select **Allow performance information collection** and validate the change. The `Psize` value is then displayed by **`topas`**.

Monitoring using nmon

Starting with the Virtual I/O Server release 2.1, AIX 5.3 TL9, and AIX 6.1 TL2, **`nmon`** is included in the default installation. To use **`nmon`**, in the Virtual I/O Server complete these steps:

1. Run the **`topas`** command at the shell prompt (Example 2-10).

Example 2-10 Topas basic panel

```
Topas Monitor for host:nimres2                  EVENTS/QUEUES    FILE/TTY
Tue Dec  4 09:04:53 2012 Interval:2           Cswitch     258  Readch     9818
                                                Syscall     744  Writech    2008
CPU     User% Kern% Wait% Idle%                 Reads         9  Rawin         0
Total     0.1   0.2   0.0  99.7                 Writes        1  Ttyout      994
                                                Forks         0  Igets         0
Network    BPS  I-Pkts  O-Pkts    B-In   B-Out  Execs         0  Namei        35
Total    1.39K    3.54    1.52   306.6   1.09K  Runqueue      0  Dirblk        0
                                                Waitqueue   0.0
Disk    Busy%      BPS     TPS  B-Read  B-Writ                    MEMORY
Total     0.0        0       0       0       0  PAGING           Real,MB    8192
                                                Faults       72  % Comp       28
FileSystem          BPS    TPS  B-Read  B-Writ  Steals        0  % Noncomp    70
Total             12.6K  41.49   12.6K       0  PgspIn        0  % Client     70
                                                PgspOut       0
Name           PID  CPU%  PgSp Owner           PageIn        0  PAGING SPACE
topas       9109556  0.0 1.82M root            PageOut       0  Size,MB      512
getty      11862072  0.0  572K root            Sios          0  % Used       2
ksh        10813518  0.0  560K root                            % Free      98
syncd       2097250  0.0  572K root            NFS (calls/sec)
nfsd        5374128  0.0 1.82M root            SerV2         0  WPAR Activ     0
ksh         6881496  0.0  532K root            CliV2         0  WPAR Total     0
lrud         262152  0.0 72.0K root            SerV3         0  Press: "h"-help
java        6422732  0.0 78.1M root            CliV3         0         "q"-quit
```

2. When the **`topas`** window is displayed, press ~ (tilde). The window shown in Example 2-11 is displayed.

Example 2-11 Initial window of the NMON application

```
.                                                                    .
.  ----------------------------                                      .
.  N    N  M    M   0000   N    N   For online help type: h          .
.  NN   N  MM  MM  0    0  NN   N   For command line option help:    .
.  N N  N  M MM M  0    0  N N  N      quick-hint  nmon -?           .
.  N  N N  M    M  0    0  N  N N    full-details  nmon -h           .
```

```
·  N   NN  M    M  O    O  N   NN   To start nmon the same way every time?     ·
·  N    N  M    M   OOOO   N    N    set NMON ksh variable, for example:       ·
·  ----------------------------     export NMON=cmt                           ·
·    TOPAS_NMON                                                               ·
·                                4 - CPUs currently                           ·
·                                4 - CPUs configured                          ·
·                             1654 - MHz CPU clock rate                       ·
·                   PowerPC_POWER7 - Processor                                ·
·                           64 bit - Hardware                                 ·
·                           64 bit - Kernel                                   ·
·                       1,nimres2 - Logical Partition                         ·
·                     7.1.2.1 TL02 - AIX Kernel Version                       ·
·                          nimres2 - Hostname                                 ·
·                          nimres2 - Node/WPAR Name                           ·
·                          10479OE - Serial Number                            ·
·                       IBM,9113-55O - Machine Type                           ·
·····························································································
```

Tip: To switch between the **nmon** and **topas** application displays, press ~(tilde).

3. To obtain help information about monitoring the available system resources, press ? (question mark). The help window is displayed as shown in Example 2-12.

Example 2-12 Display of command help for monitoring system resources

```
··HELP·····most-keys-toggle-on/off·····································
·h = Help information     q = Quit nmon             O = reset peak counts      ·
·+ = double refresh time  - = half refresh          r = ResourcesCPU/HW/MHz/AIX·
·c = CPU by processor     C=upto 128 CPUs           p = LPAR Stats (if LPAR)   ·
·l = CPU avg longer term  k = Kernel Internal       # = PhysicalCPU if SPLPAR  ·
·m = Memory & Paging      M = Multiple Page Sizes  P = Paging Space            ·
·d = DiskI/O Graphs       D = DiskIO +Service times o = Disks %Busy Map        ·
·a = Disk Adapter         e = ESS vpath stats       V = Volume Group stats     ·
·^ = FC Adapter (fcstat)  O = VIOS SEA (entstat)    v = Verbose=OK/Warn/Danger ·
·n = Network stats        N=NFS stats (NN for v4)   j = JFS Usage stats        ·
·A = Async I/O Servers    w = see AIX wait procs   "="= Net/Disk KB<-->MB      ·
·b = black&white mode     g = User-Defined-Disk-Groups (see cmdline -g)        ·
·t = Top-Process --->     1=basic 2=CPU-Use 3=CPU(default) 4=Size 5=Disk-I/O   ·
·u = Top+cmd arguments    U = Top+WLM Classes       . = only busy disks & procs·
·W = WLM Section          S = WLM SubClasses        @=Workload Partition(WPAR) ·
·[ = Start ODR            ] = Stop ODR                                         ·
```

```
·~ = Switch to topas screen                                                  ·
·Need more details?  Then stop nmon and use: nmon -?                         ·
```

4. Use the letters to monitor the system resources. For example, to monitor processor activity, press c. Example 2-13 shows the command output.

Example 2-13 Monitoring processor activity with nmon

```
··topas_nmon··2=Top-Child-CPU····Host=nimres2·····Refresh=2 secs···08:53.14··
· CPU-Utilisation-Small-View ·····················································
·                          0----------25----------50----------75----------100·
·CPU User%  Sys% Wait% Idle%|          |           |          |           |·
·  0   0.0   0.0   0.0 100.0|>                                             |·
·  1   0.0   0.0   0.0 100.0|>                                             |·
·  2   0.0   0.0   0.0 100.0|  >                                           |·
·  3   0.0   0.0   0.0 100.0|>                                             |·
·Physical Averages          +----------|-----------|----------|-----------+·
·All   0.0   0.2   0.0  99.8|   >                                          |·
·                           +----------|-----------|----------|-----------+·
·····························································································
```

5. To monitor resources other than the one that you are currently monitoring, type the appropriate letter against that system resource (see the help window in Example 2-12 on page 42). For example, in addition to monitoring the processor activity, you can monitor the network resources by typing n. This adds network statistics to the existing display, as shown in Example 2-14.

Example 2-14 NMON monitoring of processor and network resources

```
··topas_nmon··p=Partitions·····Host=nimres2·····Refresh=2 secs···09:08.18··
· CPU-Utilisation-Small-View ·····················································
·                          0----------25----------50----------75----------100·
·CPU User%  Sys% Wait% Idle%|          |           |          |           |·
·  0   0.0   0.0   0.0 100.0|>                                             |·
·  1   0.0   0.0   0.0 100.0|>                                             |·
·  2   0.0   4.8   0.0  95.2|ss>                                           |·
·  3   0.0   0.0   0.0 100.0|>                                             |·
·Physical Averages          +----------|-----------|----------|-----------+·
·All   0.3   0.7   0.0  99.0|>                                            |·
·                           +----------|-----------|----------|-----------+·
· Network ···················································· Size/Speed ·
·I/F Name Recv=KB/s Trans=KB/s packin packout insize outsize Peak->Recv TransKB·
·    en0       0.1       0.3      1.5      0.5   85.3  570.0        0.3     0.8·
·    lo0       0.0       0.0      0.0      0.0    0.0    0.0        0.0     0.0·
·  Total       0.0       0.0 in Mbytes/second   Overflow=0                    ·
·I/F Name  MTU  ierror oerror collision Mbits/s Description                   ·
```

```
·    en0   1500      0      0      0   1024 Standard Ethernet Network Interface·
·    lo0  16896      0      0      0      0 Loopback Network Interface         ·
·····························································································
```

6. To quit from the nmon/topas application, press q. This returns you to the shell prompt.

Monitoring using vmstat

To focus only on processor and memory consumption, use the **vmstat** command because you can see summarized performance metrics at regular intervals.

This command works the same on both the Virtual I/O Server and the virtual I/O client running AIX.

The **vmstat** command shows trends for processor utilization by partition.

In Example 2-15, the values are consistent during the observation period. This reflects stable processor utilization by the partition.

In this example, utilization is reported every 5 seconds until Ctrl+C is pressed. If you want to limit the number of reports (10, for example), use the **vmstat -wI 5 10** command.

Example 2-15 Monitoring with the vmstat command

```
# vmstat -wI 5

System configuration: lcpu=8 mem=4608MB ent=1.00

   kthr           memory                          page                    faults                cpu
----------- --------------------- ------------------------------------ ------------------ -----------------------
  r   b   p        avm        fre    fi    fo    pi    po    fr    sr    in     sy    cs us sy id wa    pc    ec
  8   0   0     247142     943896     0     0     0     0     0     0     7     95   171 99  0  0  0  3.93 393.0
  8   0   0     247141     943897     0     0     0     0     0     0     5     42   167 99  0  0  0  3.85 384.6
  8   0   0     247141     943897     0     0     0     0     0     0     6     29   161 99  0  0  0  3.94 393.8
  8   0   0     247141     943897     0     0     0     0     0     0     9     37   165 99  0  0  0  3.92 391.6
  8   0   0     247141     943897     0     0     0     0     0     0     3     34   164 99  0  0  0  3.93 392.6
  8   0   0     247141     943897     0     0     0     0     0     0     6     34   161 99  0  0  0  3.78 378.1
  8   0   0     247141     943897     0     0     0     0     0     0     5     28   162 99  0  0  0  3.73 372.7
```

Most of the information can be displayed by using the **topas** command as well, as shown in Example 2-7 on page 38 and Example 2-8 on page 39. The comparable fields are named as follows:

- Physical processor consumption: *pc* in **vmstat**, *PhysC* in **topas**
- Percentage of entitlement consumed: *ec* in **vmstat**, *EntC* in **topas**
- Consumption type: *us*/*sy*/*id*/*wa* in **vmstat**, *User*/*Kernel*/*Idle*/*Wait* in **topas**
- Logical processors: *lcpu* in **vmstat**, *Online Logical CPUs* in **topas**

- Partition entitlement: *ent* in **vmstat**, *Ent* in **topas**
- Runqueue: Field *r* in the *kthr* column in **vmstat**, *RunQueue* in **topas**.

  This field denotes the number of threads that are runnable, which includes threads that are running and threads that are waiting for the processor.
- Waitqueue: Field *b* in the *kthr* column in **vmstat**, *WaitQueue* in **topas**.

  This field denotes the average number of threads that are blocked either because they are waiting for a file system I/O or they were suspended because of memory load control.
- Threadqueue for I/O raw devices: Field *p* in the *kthr* column in **vmstat** (there is no corresponding field in **topas**).

  This field shows the number of threads that are waiting on I/O to raw devices per second. This does not include threads that are waiting on I/O to file systems.

As with the **topas** tool, focus on the *r* field value to determine the number of virtual processors that can be used at a particular time.

Monitoring using lparstat

The **lparstat** command provides an easy way to determine whether processor utilization is optimized. This command does not work for Virtual I/O Server. Example 2-16 on page 46 shows how the **lparstat** command can be used to display processor utilization metrics at an interval of 1 second for two iterations. Similar to **vmstat**, processor utilization metrics can be reported continuously until pressing Ctrl+C by running **lparstat -h 1**.

Most of the fields that are shown in Example 2-16 on page 46 are explained for the **topas** and **vmstat** commands in previous sections. There are two extra fields:

- *lbusy* shows the percentage of logical processor utilization that occurs while running in user and system mode. If this value approaches 100%, it might indicate that the partition can use more virtual processors. In this example, it is about 25%. Because `lcpu` is 4 and *%entc* is about 100%, the partition is probably running a single-threaded process that consumes a considerable amount of the processor.
- *app* shows the number of available processors in the shared pool. In this example *psize* = 2, *ent* = 0.4, and *%entc* =100. Therefore, the remaining processing resource on the managed system is approximately 1.6. Because the *app* field is 1.59, you can conclude that the other partitions on the managed system consume almost no processing resources.

Consideration: The app field is only available when the `Allow performance information collection` item is selected for the current partition properties.

Example 2-16 Monitoring using the lparstat command

```
# lparstat -h 1 2
System configuration: type=Shared mode=Capped smt=On lcpu=4 mem=4096 psize=2 ent=0.40
%user  %sys  %wait  %idle physc %entc  lbusy   app  vcsw phint  %hypv hcalls
-----  ----  -----  ----- ----- ----- ------   ---  ---- -----  ----- ------
 84.9   2.0    0.2   12.9  0.40  99.9   27.5  1.59   521     2   13.5   2093
 86.5   0.3    0.0   13.1  0.40  99.9   25.0  1.59   518     1   13.1    490
```

You also have access to these metrics by running **topas -L**. However, by using **lparstat**, you can see the short-term evolution of these values. For more information about **topas -L**, see “Monitoring using topas” on page 38.

Monitoring variable processor frequency

There are two main reasons for processor frequency to vary, as described in “Scaled Processor Utilization Resource Register (SPURR)” on page 27. It can vary in the following ways:

- Down: Used to control power consumption or fix a heat problem
- Up: Used to boost performance

The impact on the system can be monitored by using the **lparstat** command. When the processor is not running at nominal speed, the `%nsp` field is displayed showing the current average processor speed as a percentage of nominal speed. This is illustrated in Example 2-17.

Example 2-17 Variable processor frequency view with lparstat

```
# lparstat
 System configuration: type=Shared mode=Uncapped smt=On lcpu=2 mem=432 psize=2 ent=0.50
 %user  %sys  %wait  %idle   physc  %entc   lbusy   vcsw   phint %nsp  %utcyc
 ----------------------------------------------------
  80.5  10.2    0.0    9.3    0.90    0.5    90.5 911944  434236 110     9.1
 # lparstat –d
 System configuration: type=Shared mode=Uncapped smt=On lcpu=2 mem=432 psize=2 ent=0.50
 %user  %sys  %wait  %idle  physc %entc  %nsp  %utuser  %utsys %utidle  %utwait
 -----------------------------------------------------------
  70.0  20.2    0.0    9.7    0.5   100   110     5.01    1.70   2.30     0.09
```

In this example, you can tell that the processor is running above nominal speed because %nsp > 100.

Accounting is disabled by default in turbo mode. It can be enabled by using SMIT. If accounting is disabled and the processor is running above nominal speed, the %utcyc is displayed. This statistic represents the total percentage of unaccounted turbo cycles. In this example, 9.1% of the cycles were unaccounted. If turbo accounting mode is disabled, the processor utilization statistics are capped to the PURR values.

In Example 2-17 on page 46, these extra metrics are displayed by **lparstat -d** because turbo mode accounting is disabled and the processor is running above nominal speed:

- %utuser shows the percentage of unaccounted turbo cycles in user mode.
- %utsys shows the percentage of unaccounted turbo cycles in kernel mode.
- %utidle shows the percentage of unaccounted turbo cycles in idle state.
- %utwait shows the percentage of unaccounted turbo cycles in I/O wait state.

Monitoring using sar

The **sar** command provides statistics for every logical processor. It can be used in two ways, as described in this section.

Real-time processor utilization metrics

The **sar** command can show real-time processor utilization metrics that are sampled at an interval for a specified number of iterations. Example 2-18 shows two iterations of processor utilization for all processors on a 3 second interval. It is not possible to show continuous metrics until Ctrl+C is pressed, unlike other commands.

Example 2-18 Individual processor monitoring using the sar command

```
# sar -P ALL 3 2

AIX p740_lpar03 1 7 00F7A22E4C00 12/04/12

System configuration: lcpu=8 ent=1.50 mode=Uncapped

11:33:28 cpu    %usr    %sys    %wio   %idle   physc   %entc
11:33:31  0        3      85       0      12    0.01     2.8
          1        0      36       0      63    0.00     0.4
          U        -       -       0      96    0.29    96.8
          -        0       3       0      97    0.01     3.2
11:33:34  0        4      63       0      33    0.00     1.1
          1        0      35       0      65    0.00     0.4
          U        -       -       0      98    0.30    98.5
          -        0       1       0      99    0.00     1.5

Average   0        3      79       0      18    0.01     1.9
          1        0      36       0      64    0.00     0.4
```

```
        U       -       -       0      97    0.29    97.6
        -       0       2       0      98    0.01     2.4
```

Example 2-18 on page 47 shows that the activity of individual logical processors is reported. The `U` line shows the unused capacity of the virtual processor.

Information that was shown previously by using the **topas** command is also shown here:

- `physcc` in **sar** = *pc* in **mpstat** = *PhysC* in **topas** = physical processor consumption
- `%entc` in **sar** = %*ec* in **mpstat** = *EntC* in **topas** = percentage of entitlement consumed

Processor utilization metrics from a file

The **sar** command can extract and show processor utilization metrics that were saved in a file (`/var/adm/sa/sa`*dd*, where *dd* refers to the current day). The system utilization information is saved by two shell scripts (`/usr/lib/sa/sa1` and `/usr/lib/sa/sa2`) running in the background. These shell scripts are started by the cron daemon by using crontab file `/var/spool/cron/crontabs/adm`.

Tip: The crontab entries for sar are commented out by default. To enable them, run the command **crontab -e adm** and uncomment the adm users crontab file.

Collecting data in this manner is a useful way to characterize system usage over time and determine peak usage hours.

Example 2-19 shows the **sar** command working on a previously saved file.

Example 2-19 The sar command working a previously saved file

```
# ls -l /usr/adm/sa/
total 112
-rw-r--r--    1 root     system        21978 Nov 03 10:25 sa03
-rw-r--r--    1 root     system        26060 Oct 30 17:04 sa30
-rw-r--r--    1 root     system          780 Nov 03 10:25 sar03
# sar -f /usr/adm/sa/sa03

AIX aix61 1 6 00C1F1704C00    11/03/08

System configuration: lcpu=2 ent=0.50 mode=Uncapped

10:25:09    %usr    %sys    %wio   %idle   physc   %entc
10:25:12       1       1       0      98    0.01     2.3
10:25:13       0       1       0      99    0.01     1.1
```

```
10:25:14       0       0       0     100    0.00     0.7
... omitted lines ...
10:25:20       0       0       0     100    0.00     0.7
10:25:21       0       0       0     100    0.00     0.7

Average        0       0       0      99    0.01     1.1
# sar -P ALL -f /usr/adm/sa/sa03

AIX aix61 1 6 00C1F1704C00    11/03/08

System configuration: lcpu=2 ent=0.50 mode=Uncapped

10:25:09 cpu    %usr    %sys    %wio   %idle   physc   %entc
10:25:12  0       55      35       0      11    0.01     2.0
          1        0      50       0      50    0.00     0.3
          U        -       -       0      98    0.49    97.7
          -        1       1       0      98    0.01     2.3
10:25:13  0       18      57       0      25    0.00     0.8
          1        0      49       0      51    0.00     0.3
          U        -       -       0      99    0.49    98.9
          -        0       1       0      99    0.01     1.1
10:25:14  0        5      50       0      44    0.00     0.5
          1        0      47       0      53    0.00     0.3
          U        -       -       0      99    0.50    99.3
          -        0       0       0     100    0.00     0.7
... omitted lines ...
Average   0       32      43       0      25    0.00     0.8
          1        0      48       0      52    0.00     0.3
          U        -       -       0      99    0.49    98.9
          -        0       0       0      99    0.01     1.1
```

The **sar** command does not work for Virtual I/O Server.

Monitoring using mpstat

The **mpstat** command provides the same information as **sar**, but it also provides more information about the run queue, page faults, interrupts, and context switches. If you do not need this level of detail, the **sar** command is sufficient.

Both **sar** and **mpstat** provide parameter flags to display more system information. See the man page of each command for more information about these flags.

The **mpstat** command displays these specific metrics:

- *mig*: The number of thread migrations to another logical processor.
- *lpa*: The number of redispatches within affinity domain 3.
- *ics*: Involuntary context switches.

- *mpc: The* number of mpc interrupts. These are proactive interrupts that are used to ensure rapid response to an inter-processor preemption request when the preempting thread is considered a “real time” thread.
- *lcs*: The logical processor context switches

Example 2-20 shows that there are standard logical processor context switches. However, no thread was forced to migrate to another logical processor. The output from this command is displayed endlessly without stopping. To use mpstat for only a specified number of iterations (10, for example), run **mpstat 3 10**.

Example 2-20 Individual processor monitoring using the mpstat command

```
# mpstat 3

System configuration: lcpu=8 ent=1.5 mode=Uncapped

cpu  min  maj  mpc  int   cs  ics   rq  mig lpa sysc us sy wa id   pc  %ec  lcs
  0    0    0    0  141  117   61    0    0 100   45  3 64  0 33 0.00  1.0  145
  1    0    0    0  182    0    0    0    0   -    0  0 33  0 67 0.00  0.4  130
  U    -    -    -    -    -    -    -    -   -    -  -  -  0 98 0.29 98.0    -
ALL    0    0    0  323  117   61    0    0 100   45  0  1  0 99 0.00  1.4  275
------------------------------------------------------------------------------
  0    1    0    0  158  172   88    0    0 100   53  5 63  0 32 0.00  1.2  169
  1    0    0    0  194    0    0    0    0   -    0  0 34  0 66 0.00  0.4  140
  U    -    -    -    -    -    -    -    -   -    -  -  -  0 98 0.29 97.8    -
ALL    1    0    0  352  172   88    0    0 100   53  0  1  0 99 0.00  1.6  309
------------------------------------------------------------------------------
```

The **mpstat** command does not work for Virtual I/O Server.

Report generation for processor utilization

Report generation was not an option provided for Virtual I/O Server through restricted shell. To start report generation, enable it from the root shell. Report generation works on virtual I/O client with AIX.

Continuous processor monitoring using topas

Data collection in **topas** for continuous monitoring can be enabled so that statistics can be collected to provide an idea of the load on the system. This function was introduced in AIX V5.3 TL 4.

This data collection involves several commands. The **xmwlm** command is started with the inittab. It records data such as processor usage, memory, disk, network, and kernel statistics. The **topasout** command generates reports based on the statistics that are recorded by the **xmwlm** command. The **topas -R** command can also be used to monitor cross-partition activities.

The easiest way to quickly set up data collection with **topas** is to use the associated SMIT interface.

To create a report, you must start recording first (unless it is already started) as shown in Figure 2-11 using **smitty topas**.

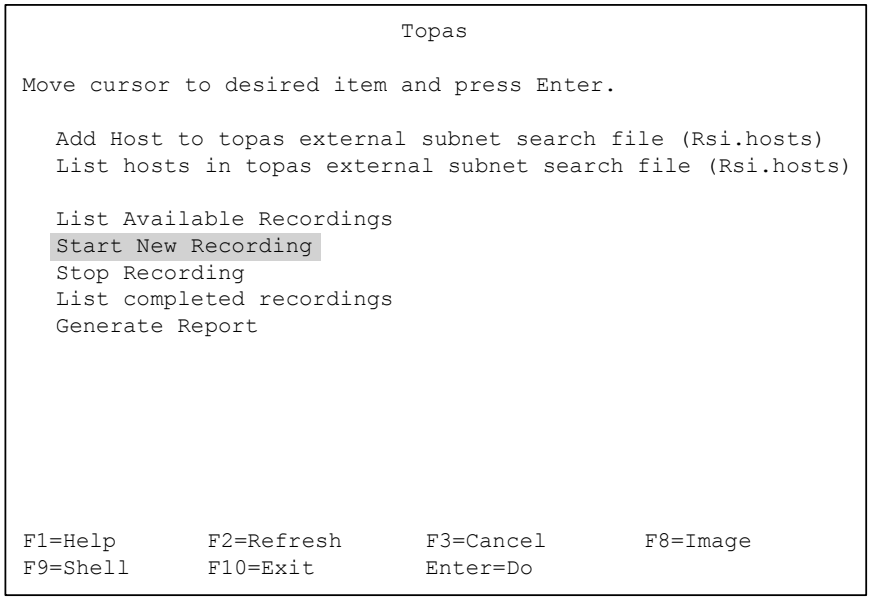


Figure 2-11 Using smitty topas for processor utilization reporting

From there, select the kind of recording you want for report generation. In this example, CEC recording is selected.

You must specify the path where the recording output file will be saved in the **Output Path** field as shown in Figure 2-12.

```
                    Start Local CEC recording

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

                                                [Entry Fields]
  Type of Recording                            ced
  Length of Recording                          hour
*  Recording intervals in seconds              [60]                #
*  Number of Samples                           [60]                #
  Output Path                                  [/etc/perf/daily]



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

Figure 2-12 Local CEC recording attributes window

You can select the following types of recordings through the interface:

CEC Recording This monitors cross-partition activity (only available on AIX 5.3 TL5 and later). It starts the **topas -R** daemon. These reports are only available when **Allow performance information collection** is selected for the partition on the HMC.

Local Recording This monitors the current partition activity. It starts the **xmwlm -L** daemon. This daemon is active by default.

After the process successfully completes, you can start generating the report from **smitty topas** by selecting Generate Report → Filename or Printer and specifying the path as shown in Figure 2-13. You might have to wait until data is logged in to the report file.

```
                  Path to locate the recording File

Type or select a value for the entry field.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
* Path to Locate the recording file                  [/etc/perf/daily]   +









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

Figure 2-13 Report generation

Specify the Reporting Format on the next panel. Press F4 to see all of the available formats as shown in Figure 2-14.

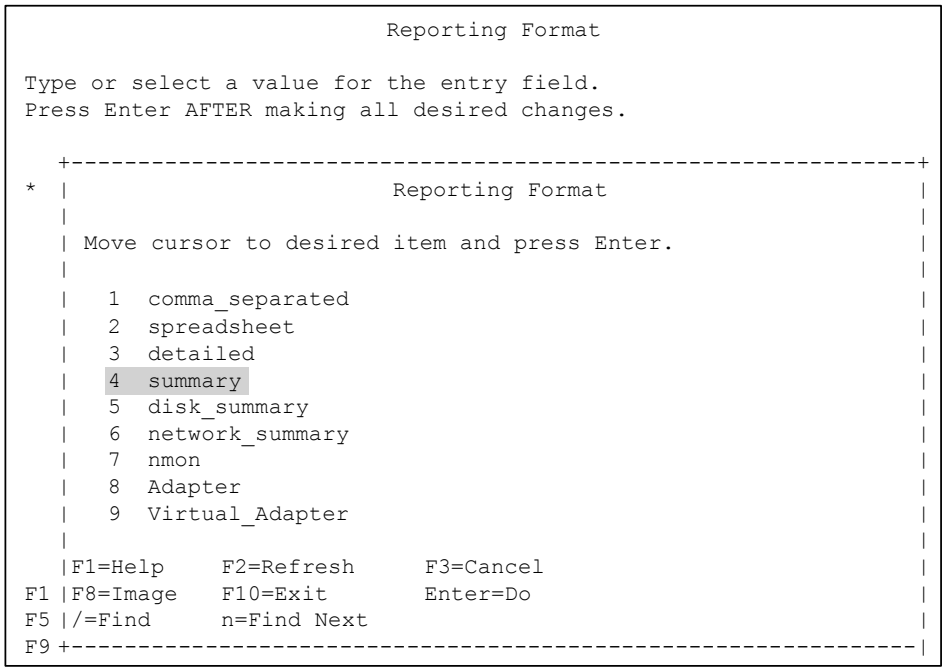


Figure 2-14 Reporting Format panel

Formats of various reports

The following are a sample of the various reports available:

- comma_separated:

```
#Monitor: xmtrend recording--- hostname: Apps_server ValueType: mean
Time="2008/14/10 17:25:33", CPU/gluser=19.62
Time="2008/14/10 17:25:33", CPU/glkern=2.48
Time="2008/14/10 17:25:33", CPU/glwait=12.83
. . .
```

- spreadsheet: Produces a file that can be opened by spreadsheet software.

```
"#Monitor: xmtrend recording --- hostname: Apps_server" ValueType: mean
"Timestamp" "/Apps_server/CPU/gluser" "/Apps_server/CPU/glkern" . . . .
"2008/14/10 17:25:33" 19.62  2.48  12.83  65.07 1.00  6.10  0.00  . . .
"2008/14/10 17:30:33" 0.04  0.24  0.06  99.66  1.00  6.10 0.00 . . .
. . .
```

- detailed: Provides a detailed view of the system metrics.

```
#Report: System Detailed --- hostname: Apps_server
version:1.2

Start:14/10/08 17:25:33 Stop:14/10/08 23:50:26   Int:  5 Min    Range:  384
Min

Time: 17:30:32
------------------------------------------------------------

CONFIG        CPU             MEMORY        PAGING
Mode     Ded  Kern     2.5    Sz,GB    4.0  Sz,GB   0.5
LP       2.0  User    19.6    InU      0.8  InU     0.0
SMT       ON  Wait    12.8    %Comp   19.2  Flt    2382
Ent      0.0  Idle    65.1    %NonC    2.8  Pg-I    404
Poolid    -   PhyB    22.1    %Clnt    2.8  Pg-O      4
              Entc     0.0

PHYP          EVENTS/QUEUES    NFS
Bdon     0.0  Cswth      602   SrvV2      0
Idon     0.0  Syscl     6553   CltV2      0
Bstl     0.0  RunQ         1   SrvV3      0
Istl     0.0  WtQ          2   CltV3      0
Vcsw   456.4
Phint    0.0

Network    KBPS   I-Pack  O-Pack    KB-I    KB-O
en0         1.6     10.4     5.1     1.0     0.6
lo0         0.2      0.8     0.9     0.1     0.1
. . .
```

- summary: Presents a consolidated view of system information.

```
Report: System Summary  --- hostname: Apps_server                   version:1.2
Start:14/10/08 17:25:33   Stop:14/10/08 23:50:26  Int: 5 Min  Range: 384 Min
Mem:  4.0 GB  Dedicated SMT: ON   Logical CPUs:  2
Time       InU Us Sy Wa  Id  PhysB RunQ  WtQ CSwitch Syscall PgFault  %don  %stl
---------  ------------------------------- ----------------------  ----------
17:30:32   0.8 20  2 13  65  22.10  1.2  1.7     602    6553    2382   0.0   0.0
17:35:32   0.8  0  0  0 100   0.28  0.7  0.2     164      46       6   0.0   0.0
17:40:32   0.8  0  0  0 100   0.36  1.2  0.0     167      74      25   0.0   0.0
. . .
```

- disk_summary: Provides information about the amount of data that is read or written to disks.

```
Report: Total Disk I/O Summary  --- hostname: Apps_server
version:1.1
Start:14/10/08 17:25:33   Stop:14/10/08 23:50:26  Int: 5 Min  Range: 384
Min
```

```
Mem:  4.0 GB   Dedicated SMT: ON   Logical CPUs:  2
Time       InU   PhysB   MBPS    TPS   MB-R   MB-W
17:30:32  0.8    22.1    1.1  132.1    1.0    0.1
17:35:32  0.8     0.3    0.0    1.1    0.0    0.0
17:40:32  0.8     0.4    0.0    0.6    0.0    0.0
. . .
```

- network_summary: Provides information about the amount of data that is received or sent by the network interfaces.

```
#Report: System LAN Summary  --- hostname: Apps_server                    version:1.1
Start:14/10/08 17:25:33 Stop:14/10/08 23:50:26   Int:  5 Min   Range:  384 Min
Mem:  4.0 GB  Dedicated SMT: ON  Logical CPUs:  2
Time       InU   PhysB   MBPS  MB-I  MB-O Rcvdrp Xmtdrp
17:30:33   0.8    22.1    0.0   0.0   0.0      0      0
17:35:33   0.8     0.3    0.0   0.0   0.0      0      0
17:40:33   0.8     0.4    0.0   0.0   0.0      0      0
. . .
```

- nmon: Generates a nmon analyzer report that can be viewed with the nmon analyzer, as described in “Monitoring using nmon” on page 41.

```
CPU_ALL,CPU Total ,User%,Sys%,Wait%,Idle%,CPUs,
CPU01,CPU Total ,User%,Sys%,Wait%,Idle%,
CPU00,CPU Total ,User%,Sys%,Wait%,Idle%,
CPU03,CPU Total ,User%,Sys%,Wait%,Idle%,
CPU02,CPU Total ,User%,Sys%,Wait%,Idle%,
DISKBUSY,Disk %Busy ,hdisk0,cd0,cd1,hdisk2,hdisk1,cd0,hdisk2,hdisk1,
DISKREAD,Disk Read kb/s ,hdisk0,cd0,cd1,hdisk2,hdisk1,cd0,hdisk2,hdisk1,
. . .
```

You can also use the **topasout** command to directly generate the reports. For more information, see its man page.

Remember: The **topas -R** and **xmwlm** daemons store their recordings in the /etc/perf directory by default. **topas -R** stores its reports in files with the topas_cec.*YYMMDD* name format. **xmwlm** uses the daily/xmwlm.*YYMMDD* file name format.

Recordings cover single-day periods (24 hours) and are retained for two days before they are automatically deleted in AIX V5.3 TL4. This period was extended to seven days in AIX V5.3 TL5. Therefore, a week’s worth of data is retained on the system at all times.

The **topasout** command can generate an output file that can be transmitted to the **nmon** analyzer tool. It uses the file that is generated by the **xmwlm** daemon as input.

To generate output that can be transmitted to the **nmon** analyzer tool, run the following commands:

- To process a **topas** cross-partition report, run **topasout -a** with the report file:

  ```
  # topasout -a /etc/perf/topas_cec.071205
  ```

  The result file is then stored in `/etc/perf/topas_cec.071205.csv`.

- To process a local **xmwlm** report, run **topasout -a** with the report file:

  ```
  # topasout -a /etc/perf/daily/xmwlm.071201
  ```

  The results are stored in `/etc/perf/daily/xmwlm.071201.csv`. This file can be opened with a nmon analyzer to graphically display the data.

FTP the resulting csv file to your station that runs Microsoft Excel using the ASCII or TEXT options. Then, open the nmon_analyser spreadsheet. Select **Analyze nmon data** and select the csv file you transferred.

The program generates several graphs ready for you to study or use for a performance report.

Continuous processor monitoring using sar

You can use the **sar** command to display processor utilization information in a report format. This report information is saved by two shell scripts that are started in the background by a cron job. For more information, see “Processor utilization metrics from a file” on page 48.

2.2.5 IBM i processor monitoring

This section provides examples of IBM i processor monitoring. You must use native IBM i tools to monitor processor allocations and usage. These metrics are not visible to Virtual I/O Server cross-partition monitoring tools like **topas**.

Real-time processor monitoring on IBM i

The **WRKSYSACT** command can be used for real-time processor monitoring on IBM i, as shown in Figure 2-15.

```
                          Work with System Activity                   P71I04
                                                              12/05/11  17:49:55
Automatic refresh in seconds  . . . . . . . . . . . . . . . . . . .     5
Job/Task CPU filter . . . . . . . . . . . . . . . . . . . . . . . .    .10
Elapsed time . . . . . . :   00:00:02    Average CPU util . . . . :     196.2
Virtual Processors . . . . :      32     Maximum CPU util . . . . . :   197.2
Overall DB CPU util  . . . :      .0     Minimum CPU util . . . . . :   194.9
Average CPU rate . . . . . :    68.8     Current processing capacity:     5.00

Type options, press Enter.
  1=Monitor job   5=Work with job
                                                             Total  Total   DB
     Job or                                                   CPU   Sync  Async   CPU
Opt  Task         User        Number  Thread    Pty   Util   I/O    I/O   Util
     ZJCPULOOP    QSECOFR     007204  00000002   50    1.6     0      0    .0
     ZDEV_9_AP    QSECOFR     007222  00000002   50    1.6     0      0    .0
     ZDEV_2_AP    QSECOFR     007175  00000002   50    1.6     0      0    .0
     ZJCPULOOP    QSECOFR     007215  00000002   50    1.6     0      0    .0
     ZJI2_MI_P    QSECOFR     007064  00000001   50    1.6     0      0    .0
     ZDEV_7_AP    QSECOFR     007210  00000002   50    1.6     0      0    .0
                                                                       More...
F3=Exit   F10=Update list   F11=View 2   F12=Cancel   F19=Automatic refresh
F24=More keys
(C) COPYRIGHT IBM CORP. 1980, 2009.
```

Figure 2-15 IBM i WRKSYSACT command output

The Work with System Activity panel shows the current physical and virtual processor allocation to the IBM i partition. It shows processor utilization percentages for average, maximum, and minimum utilizations, and the processor utilization by all jobs performing database processing work within the `Overall DB CPU util` value. Additionally, a list of jobs or system tasks is displayed with the ones consuming the most processing time listed at the top.

Selecting **F10=Update list** refreshes the display with current information. Set the refresh interval to 5 seconds or longer so the sampled data is statistically meaningful. Use the **F19=Automatic refresh** function to automatically refresh the display at the interval duration specified in **Automatic refresh in seconds**. Doing so allows you to easily identify jobs that are consuming significant processing time because they appear repeatedly at the top of the list.

Consideration: The processor utilization can actually exceed 100% for an *uncapped* IBM i partition using a shared processor pool, and reach up to the percentage value for the number of virtual processors.

For further information about IBM i performance management tools, see *IBM eServer iSeries Performance Management Tools*, REDP-4026.

Long-term processor monitoring on IBM i

For long-term monitoring of processor usage, the IBM Performance Tools for IBM i licensed program (5770-PT1) can be used. Using it, you can generate various reports from QAPM* performance database files that are created from Collection Services data.

IBM Performance Tools for IBM i functions are accessible on IBM i 5250 sessions through a menu by using the **STRPFRT** or **GO PERFORM** command. They are also accessible through native CL commands like **PRTSYSRPT**, **PRTCPTRPT**, **PRTACTRPT**, and so on.

Example 2-21 shows a spool file output from a *component report* for *component interval activity* that was created with the following command:

```
PRTCPTRPT MBR(Q339155623) TYPE(*INTERVAL)
```

Example 2-21 IBM i component report for component interval activity

```
*...+....1....+....2....+....3....+....4....+....5....+....6....+....7....+....8....+....9....+....0....+....1....+....2....+....3
                                                  Component Report                                                 120511   18:20:1
                                              Component Interval Activity                                                  Page
 Member  . . . : Q339155623 Model/Serial  . : E8B/10-0EF5R       Main storage . . :   20.0  GB Started . . . . :  12/05/11 15:56:2
   Library . . : QPFRDATA   System name . . :P71IO4              Version/Release  :    7/ 1.0  Stopped . . . . :  12/05/11 18:20:1
 Partition ID  : 004        Feature Code  . :EPA1-EPA1           Int Threshold  . :    .00 %
 Virtual Processors:  32    Processor Units :   5.00
                                                 Uncap Int   Int   DB    ----- Disk I/O ------     High         Pool        Excp
  Itv     Tns     Rsp     DDM   -CPU Utilization- CPU   Feat  CPU   Cpb   ----- Per Second ----  -- Util --  -Faults/Sec-     per
  End    /Hour   /Tns     I/O   Total Inter Batch Avail Util  >Thld Util     Sync      Async     Disk  Unit  Mch User  ID   Second
 ----- --------- ----- -------- ----- ----- ----- ----- ----- ----- ----- ---------- ----------  ----  ----  --- ----  --  --------
 15:56      1800   .00        0 181.1    .0 181.1 623.9    .0     0    .0       67.5       22.8     5  0001    0   26  02       6.0
 15:56       960   .00        0 177.2    .0 177.2 616.5    .0     0    .0         .6        1.5     1  0001    0    0  03        .0
 15:57       480   .00        0 169.9    .0 169.9 630.8    .0     0    .0        2.3        2.0     1  0001    0    0  00        .2
 15:57       960   .00        0 169.9    .0 169.9 630.2    .0     0    .0        1.0        1.8     1  0001    0    0  02        .1
----- Cont ----
 18:19       720   .00        0 196.2   1.8 194.3 200.0    .0    28    .0       52.6      163.8    46  0001    0    2  02        .2
 18:19       720   .00        0 196.2   1.8 194.4 200.0    .0    27    .0       16.8      196.6    55  0001    0    2  02        .2
 18:19      2640   .00        0 196.2   1.8 194.3 200.0    .0    28    .0       59.4      200.2    47  0001    0    0  02        .8
 18:19       480 99.90        0 196.2   1.7 194.4 200.0    .0    27    .0       20.2      272.2    49  0001    0    0  02        .2
 18:20       480 99.90        0 196.2    .3 195.8 200.0    .0   123    .0       19.8      264.2    44  0001    0    0  02        .2
 18:20       720 99.90        0 194.1    .2 193.8 200.0    .0    12    .0       36.0      252.3    75  0001    0    1  02        .1

                                                                                                                          Bottom
 F3=Exit   F12=Cancel   F19=Left   F20=Right   F24=More keys
```

Regularly monitor processor utilization so you can take proactive measures like ending jobs or adding processor allocations to prevent slow response time for users.

As a general performance rule, the IBM i average processor utilization for high priority jobs (RUNPTY of 25 or less), should be below the percentage values shown in Table 2-3. The percentages depend on the number of available physical processors as derived from queuing theory.

Table 2-3 IBM i processor utilization guidelines

Number of processors	Processor utilization guideline
1	70
2	76
3	79
4	81
6	83
8	85
12	87
16	89
18	90
24	91
32	93

The IBM Performance Tools for IBM i system report for resource utilization expansion can be used to monitor the cumulative processor utilization according to job type run priority values as shown in Example 2-22. It is available through the IBM Performance Tools for IBM i Manager Feature, 5761-PT1 option 1. This report was created using the following command:

```
PRTSYSRPT MBR(Q339155623) TYPE(*RSCEXPN)
```

This resource utilization expansion report shows all high run priority jobs with a run priority equal or less than 25 consumed 7.9% processor utilization. All jobs and system threads, including the lower priority ones, used 44.8% of available processing time on average for the selected report time frame.

Example 2-22 IBM i System Report for Resource Utilization Expansion

```
                                              System Report                                          12/05/11 18:20:2
                                     Resource Utilization Expansion                                          Page 000
Member  . . . : Q339155623 Model/Serial  . : E8B/10-0EF5R       Main storage . . :   20.0 GB  Started  . . . . : 12/05/11 15:56:2
  Library . . : QPFRDATA   System name . . : P71I04             Version/Release  :   7/ 1.0   Stopped  . . . . : 12/05/11 18:20:1
Partition ID  : 004        Feature Code  . : EPA1-EPA1          Int Threshold  . :    .00 %
Virtual Processors:  32    Processor Units :   5.00
```

```
                Job          CPU    Cum                    ---------- Disk I/O ---------  ---- CPU Per I/O ----  ----- DIO /Sec ----
Pty             Type        Util   Util      Faults          Sync           Async         Sync       Async      Sync      Async
---  ------------------  -----  -----  --------------  -------------  -------------  ---------- ----------  --------  --------
000  System                .3     .3          23,536         49,112         35,095       .0030      .0043       5.6       4.0
001  PassThru              .0     .3              62             76              4       .0007      .0137        .0        .0
     Batch                 .0     .3           1,687          4,193         16,134       .0008      .0002        .4       1.8
009  System                .0     .3               0             12              3       .0000      .0002        .0        .0
010  Interactive           .0     .3               0              0              0       .0000      .0000        .0        .0
     Batch                 .0     .3             482            525             18       .0041      .1208        .0        .0
015  System                .0     .3             226            300            216       .0013      .0019        .0        .0
016  System                .0     .3               0              0              0       .0000      .0000        .0        .0
020  PassThru             3.4    3.8           4,580      1,218,557      1,508,268       .0012      .0009      141.1     174.7
     Batch                 .0    3.8              86            151             56       .0001      .0004        .0        .0
     AutoStart             .0    3.8               0              0              0       .0000      .0000        .0        .0
     IBM i Access-Bch      .0    3.8              96            227             83       .0019      .0053        .0        .0
     System                .0    3.8          14,225         69,198         11,915       .0000      .0001       8.0       1.3
025  Batch                 .0    3.8             458            793            103       .0409      .3151        .0        .0
     AutoStart             .0    3.8               0              0              0       .0000      .0000        .0        .0
027  Batch                 .0    3.8               0              0              0       .0000      .0000        .0        .0
030  Batch                 .0    3.8               0              0              0       .0000      .0000        .0        .0
                                                                                                                      More...
F3=Exit    F12=Cancel    F19=Left    F20=Right    F24=More keys
```

For more information about IBM Performance Tools for IBM i, see *IBM eServer iSeries Performance Tools for iSeries*, SC41-5340.

You can use the IBM Systems Director Navigator for i GUI, which is accessed using `http://`*`IBM_i_server_IP_address`*`:2001`, to graphically display long-term monitoring data such as processor utilization.

The processor utilization and waits overview graph that is shown in Figure 2-16, was generated by selecting **i5/OS™ Management** → **Performance** → **Collections**. Select a collection services DB file, and then **Investigate Data** from the menu. Finally, select **CPU Utilization and Waits Overview** from the **Select Action** menu of the Resource Utilization Rates diagram.

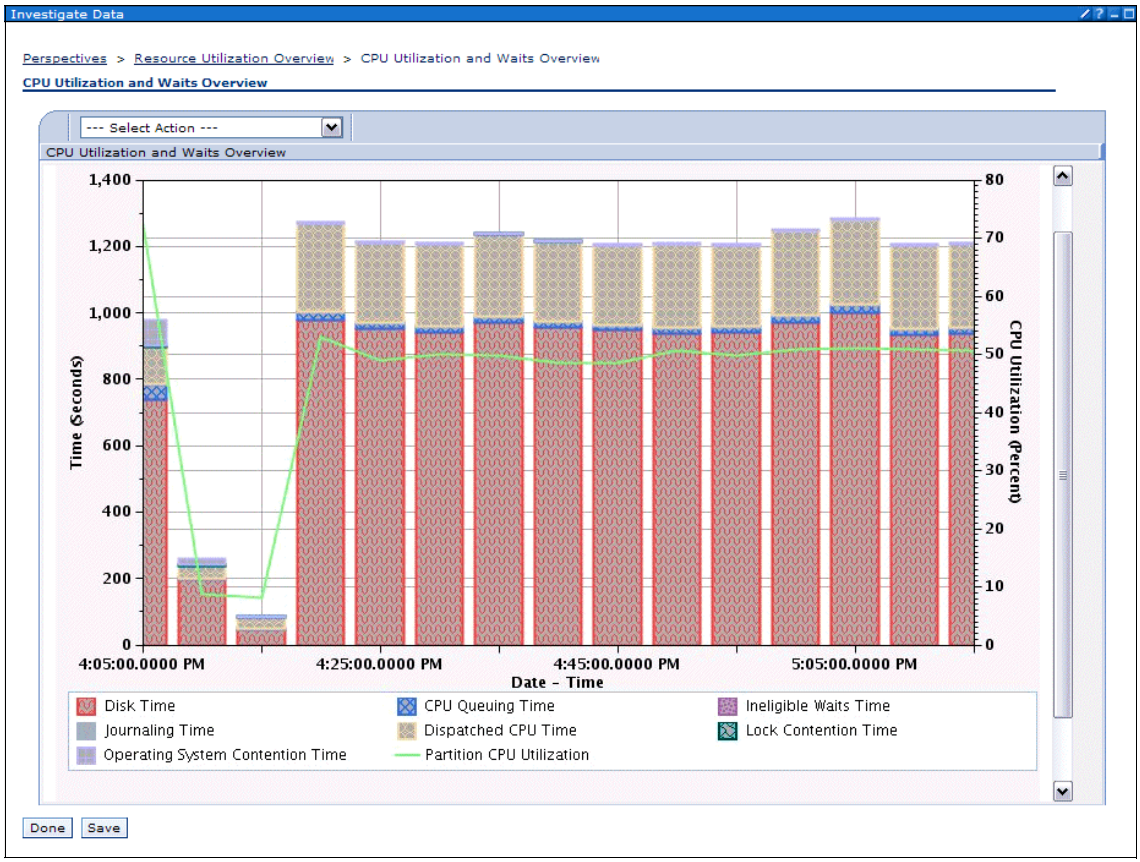


Figure 2-16 IBM i CPU Utilization and Waits Overview

As shown in the chart, you can see the processor utilization over time stabilized at around 50% as represented by the green line. Looking at where the jobs spent the most time, you can see that the system was running a disk I/O-intensive workload with jobs having spent most of their time for disk I/O and only a fraction of about 20% on processor utilization (dispatched CPU time). If you are experiencing performance problems, examining where your jobs spend most of their time reveals the area (for example, disk versus processor resource, or lock contention) where improvements will be most helpful.

Another approach for long-term processor monitoring for IBM i, which also allows IBM i cross-partition monitoring, is using the Navigator for its Management Central system monitors function. You can define a threshold for processor utilization of an IBM i system or group of systems which, when exceeded for a specified number of intervals, can trigger an event such as sending a system operator message.

For more information about using Navigator for i for performance monitoring, see *Managing OS/400 with Operations Navigator V5R1 Volume 5: Performance Management*, SG24-6565.

2.2.6 Linux processor monitoring

You can monitor processor activity by using the **iostat**, **mpstat**, and **sar** commands in the sysstat utility. For more information about the sysstat utility, see “sysstat utility” on page 452.

You can also use the **top** command to monitor the dynamic real-time view of a running system. Linux includes the top utility in the default installation. **nmon** is another widely used monitoring tool for UNIX systems.

The **mpstat** command reports processor activities for each available processor. The first processor is represented as processor 0. An example of the output is shown in Example 2-23. The figure also shows the **mpstat** command that can be used with a monitoring interval (in secs) and count for dynamic monitoring.

Example 2-23 The mpstat command output

```
[root@p750_lpar02 /]# mpstat
Linux 2.6.32-279.el6.ppc64 (p750_lpar02)        12/11/2012      _ppc64_ (4 CPU)

12:12:13 PM  CPU    %usr   %nice    %sys %iowait    %irq   %soft  %steal  %guest   %idle
12:12:13 PM  all   37.28    0.30   16.59    0.00    1.18    1.19   43.47    0.00    0.00
[root@p750_lpar02 /]# mpstat 5 2
Linux 2.6.32-279.el6.ppc64 (p750_lpar02)        12/11/2012      _ppc64_ (4 CPU)

12:12:19 PM  CPU    %usr   %nice    %sys %iowait    %irq   %soft  %steal  %guest   %idle
12:12:24 PM  all    0.00    0.00  100.00    0.00    0.00    0.00    0.00    0.00    0.00
12:12:29 PM  all   50.00    0.00    0.00    0.00    0.00    0.00   50.00    0.00    0.00
Average:     all   33.33    0.00   33.33    0.00    0.00    0.00   33.33    0.00    0.00

[root@p750_lpar02 /]#
```

The **iostat** command with the **-c** flag displays a processor activity report. The use of interval and count parameters is similar to other sysstat utility components. Example 2-24 shows an example of the processor activity output.

Example 2-24 Using iostat for processor monitoring

```
[root@p750_lpar02 ~]# iostat -c 5 2
Linux 2.6.32-279.el6.ppc64 (p750_lpar02)        12/07/2012      _ppc64_
(4 CPU)

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
          18.04    0.68   51.12    0.00   30.01    0.14

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           0.00    0.00  100.00    0.00    0.00    0.00
```

3

Multiple Shared Processor Pools

Multiple Shared-Processor Pools (MSPP) is a capability that is supported on IBM POWER6 processor-based and later managed systems. This capability allows a system administrator to create a *set of micro-partitions* with the purpose of controlling their combined physical shared processor pool consumption.

This chapter includes the following sections:

- Managing Multiple Shared Processor Pools
- Monitoring Multiple Shared Processor Pools

3.1 Managing Multiple Shared Processor Pools

You can dynamically modify existing Multiple Shared Processor Pools and create new ones through the HMC.

IBM POWER6 and POWER7 processor-based managed systems can define MSPPs and assign shared partitions to any of these pools.

Complete these steps to set up a shared processor pool (SPP):

1. In the HMC navigation panel, open **Systems Management** and click **Servers**.
2. Select the managed system of the shared processor pool you want to configure. Click **Task** and select **Configuration** → **Virtual Resources** →

Shared Processor Pool Management. The defined shared processor pools are displayed as shown in Figure 3-1.

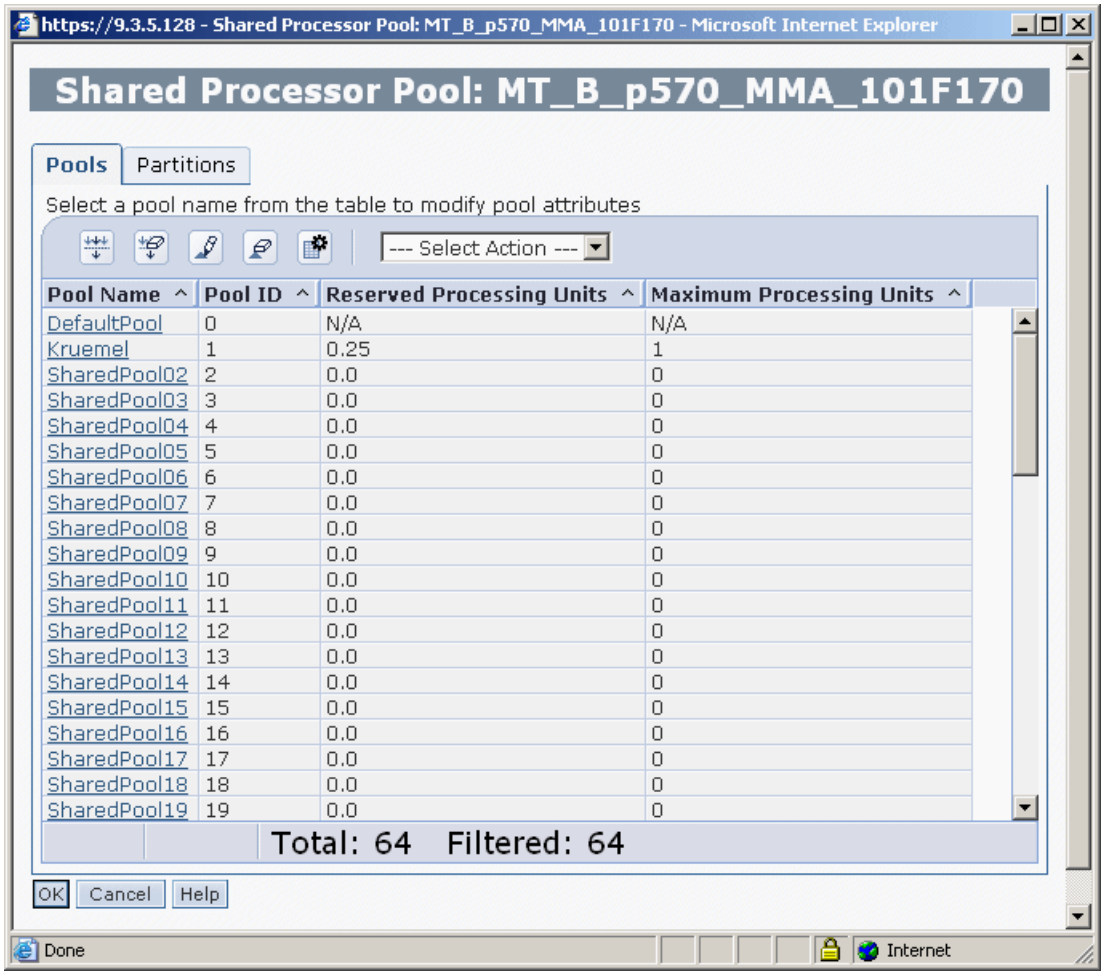


Figure 3-1 Shared Processor Pool

3. Click the name of the shared processor pool that you want to configure.

4. Enter the maximum number of processing units that you want the logical partitions in the shared processor pool to use in the **Maximum processing units** field. If wanted, change the name of the shared processor pool in the **Pool name** field, and enter the number of processing units that you want to reserve for uncapped logical partitions in the shared processor pool in the **Reserved processing units** field (Figure 3-2).

   Requirement: The name of the shared processor pool must be unique on the managed system.

   When you are done, click **OK**.

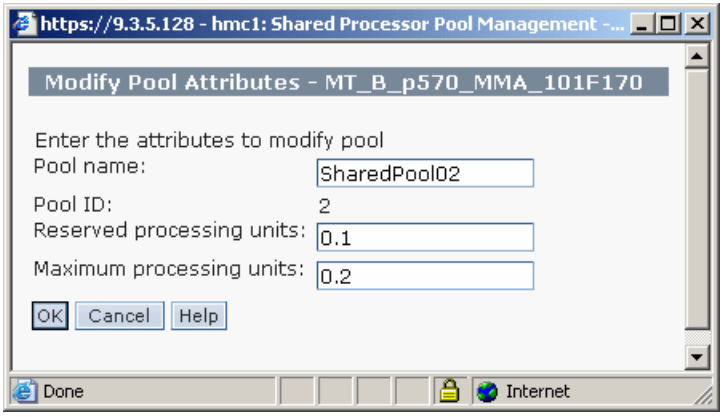


Figure 3-2 Modifying the shared processor pool attributes

5. Repeat steps 3 and 4 for any other shared processor pools that you want to configure.
6. Click **OK**.

After you modify the processor pool attributes, assign logical partitions to the configured shared processor pools. You can assign a logical partition to a shared processor pool when you create a logical partition.

Alternatively, you can reassign existing logical partitions from their current shared processor pools to the (new) shared processor pools that you configured using the following procedure:

1. Click the **Partitions** tab and select the partition name as shown in Figure 3-3.

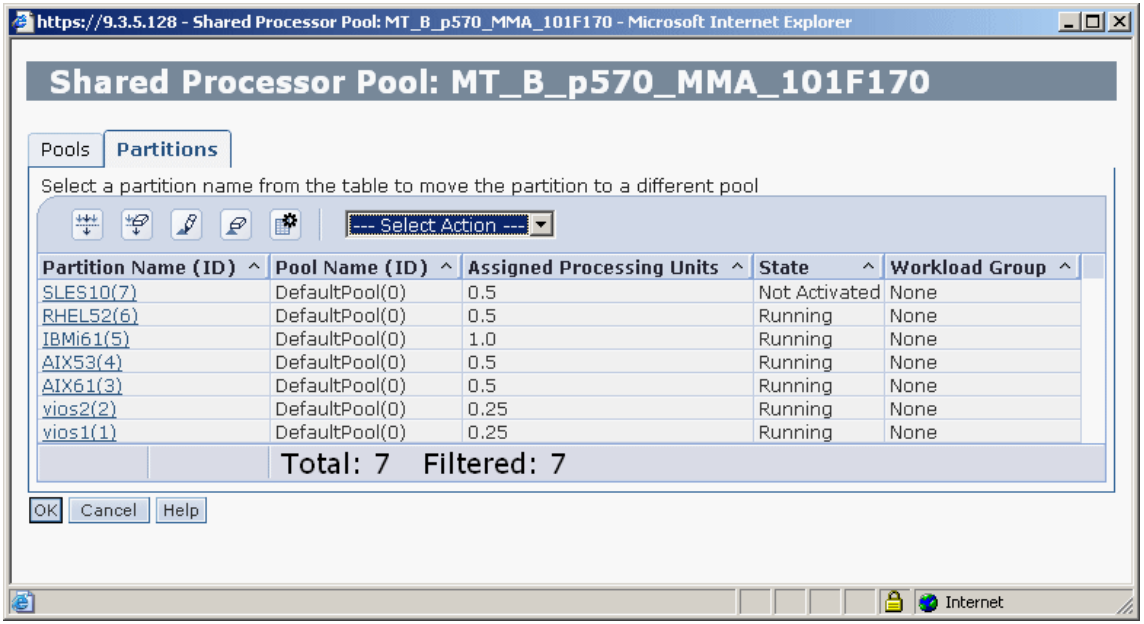


Figure 3-3 Partition assignment to Multiple Shared Processor Pools

2. Select a shared storage pool as shown in Figure 3-4.

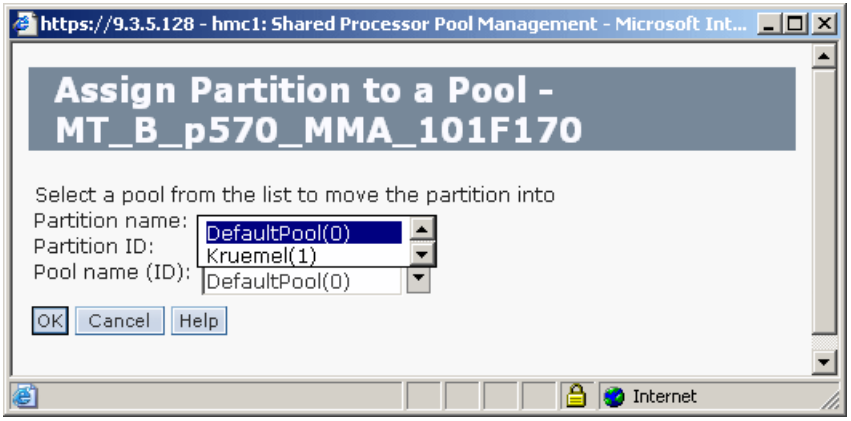


Figure 3-4 Assign a partition to a Shared Processor Pool

Remember: The default Shared Processor Pool is Pool ID 0. This pool and its default configuration values cannot be changed.

When you no longer want to use a Shared Processor Pool, you can unconfigure the shared processor pool. Do so by reassigning all logical partitions that use the shared processor pool to another shared processor pool and setting the maximum and reserved processing units to 0.

3.1.1 Calibrating the shared partitions' weight

Pay attention to the values that you provide for the partition weight when you define shared partitions. Unused processor capacity is distributed to uncapped partitions based on the weight of each partition in the managed system regardless of the shared processor pools. The partitions with the highest weight get more processor resources (Figure 3-5).

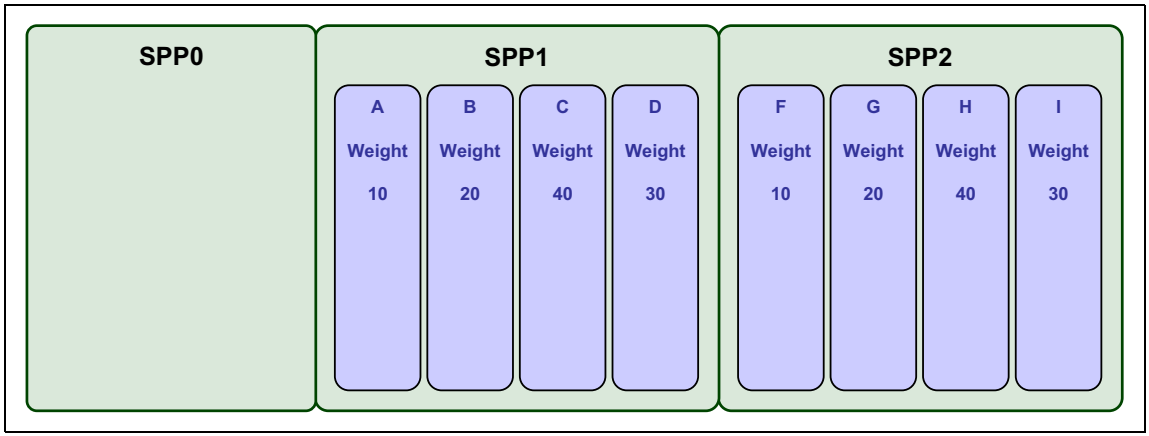


Figure 3-5 Comparing partition weights from different Shared Processor Pools

In the example shown in Figure 3-5, if the partitions C, D, and H require extra processing resources, they are distributed based on their weight value even though they are not all in the same shared processor pool.

Based on the weight value, partition C and H get most of the available shared resources (of equal amounts), and partition D receives a lesser share. In situations where your partition requires more resources, consider changing the weight of partitions in other shared processor pools.

In summary, if several partitions from different shared processor pools are competing for extra resources, the partitions with the highest weight are served

first. You must therefore define a partition's weight based on the weight of partitions in other shared processor pools.

For more information, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

3.2 Monitoring Multiple Shared Processor Pools

MSPP utilization can be monitored by using the **topas -C** (**topas -cecdisp** for Virtual I/O Server) command. Press **p** to see the processor pools section, as shown in Example 3-1.

Attention: You can monitor the consumption of other partitions by using **topas** only when at least one AIX or Virtual I/O partition to be monitored has the **Allow performance information collection** setting enabled in the partition's profile.

Example 3-1 Monitoring processor pools with topas -C

```
Topas CEC Monitor             Interval:  10             Mon Oct 13 14:09:07 2008
Partitions Memory (GB)           Processors
Shr:  0    Mon:12.5  InUse: 6.1  Shr:1.7  PSz:  4   Don: 1.0 Shr_PhysB  0.57
Ded:  4    Avl:   -              Ded:  1  APP:  3.4 Stl: 0.0 Ded_PhysB  0.00
pool  psize  entc  maxc  physb  app   mem   muse
------------------------------------------------
0     4.0    230.0 400.0 0.0    0.00  1.0   1.0
1     2.0    140.0 200.0 0.6    1.52  7.5   3.9

Host         OS  M Mem InU Lp  Us Sy Wa Id  PhysB  Vcsw  %istl %bstl %bdon %idon
-------------------------------------dedicated----------------------------------
Apps_server  A61 D 4.0 1.2  2   0  0  0 99   0.00  231    0.00  0.00  0.00 99.66
```

The report contains the following metrics. For more information about them, see "Specific to IBM POWER6 processor-based or later systems" on page 21.

- `maxc` represents the maximum pool capacity for each pool.
- `app` represents the number of available processors in the shared processor pool.
- `physb` represents the summation of physical busy of processors in shared partitions of an SPP.

- `mem` represents the sum of monitored memory for all shared partitions in the SPP.
- `muse` represents the sum of memory that is consumed for all shared partitions in the SPP.

Tip: The pool with identifier `0` is the default SPP.

When the MSPP section is displayed, you can select a pool using the up and down arrow keys. Press the **f** key to display the metrics of the shared partitions that belong to the selected pool as shown in Example 3-2.

Example 3-2 Shared pool partitions listing in topas

```
Topas CEC Monitor             Interval:  10             Mon Oct 13 14:34:05 2009
Partitions Memory (GB)           Processors
Shr:  2    Mon:12.5  InUse: 6.1  Shr:1.7  PSz:  4   Don: 1.0 Shr_PhysB  0.02
Ded:  2    Avl:   -              Ded:  1  APP:  4.0 Stl: 0.0 Ded_PhysB  0.00
pool  psize  entc  maxc  physb  app   mem   muse
------------------------------------------------
0     4.0    230.0 400.0 0.0    0.00  1.0   1.0
1     2.0    140.0 200.0 0.0    1.98  7.5   3.8

Host         OS  M Mem InU Lp  Us Sy Wa Id  PhysB  Vcsw Ent  %EntC PhI
-----------------------------------shared-----------------------------------
NIM_server   A61 C 3.0 2.9  4   0  0  0 98   0.01  518  0.40   1.9   0
DB_server    A61 U 4.5 0.9  4   0  0  0 99   0.01  401  1.00   0.5   0

Host         OS  M Mem InU Lp  Us Sy Wa Id  PhysB  Vcsw  %istl %bstl %bdon %idon
---------------------------------dedicated----------------------------------
Apps_server  A61 D 4.0 1.2  2   0  0  0 99   0.00  225    0.00  0.00  0.00 99.65
```

4

POWER processor compatibility modes

A processor compatibility mode is a value that is assigned to a logical partition by the hypervisor. It specifies the processor environment on which the logical partition can successfully operate. It is not a feature of PowerVM, but advanced PowerVM features such as Live Partition Mobility (LPM) depend on processor compatibility mode when moving partitions between servers with different processor technology. This chapter describes how to check the processor compatibility mode, and how to change it when needed.

For more information about processor compatibility mode, see *IBM PowerVM Virtualization: Introduction and Configuration*, SG24-7940, and the Information Center at:

http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7hc3/iphc3pcmdefs.htm

This chapter includes the following sections:

- Processor compatibility mode management
- Checking the compatibility mode

4.1 Processor compatibility mode management

Partitions running on POWER7 processor-based systems and later can run in three different compatibility modes: POWER6, IBM POWER6+™, and POWER7. You can select one of the compatibility modes when you create the partition, or you can modify the partition profile later. The default processor compatibility mode when you start the partition is the processor architecture of the system where the partition is running, if the operating system supports it. If the operating system does not fully support the processor architecture, the partition runs in a compatibility mode.

You can manually specify the compatibility mode for the partition. This is required if you plan to move the partition to a system with an older processor architecture, such as moving from POWER7 to POWER6, by using LPM. Also, when you move a partition from an older processor architecture to a newer processor architecture using LPM, the processor compatibility mode on the destination system is set to the older architecture. The system remains in the older processor compatibility mode even if you deactivate and reactivate it.

To change the processor compatibility mode, you must change the partition profile, and deactivate and reactivate the partition so that the profile is reloaded. To change the mode, complete these steps:

1. Edit the profile for the partition you want to change the processing mode. Click **Configuration** → **Manage Profiles** and select the profile that you want to edit.

2. Click the **Processor** tab. Select the processor compatibility mode in the menu at the bottom of the window as shown in Figure 4-1.

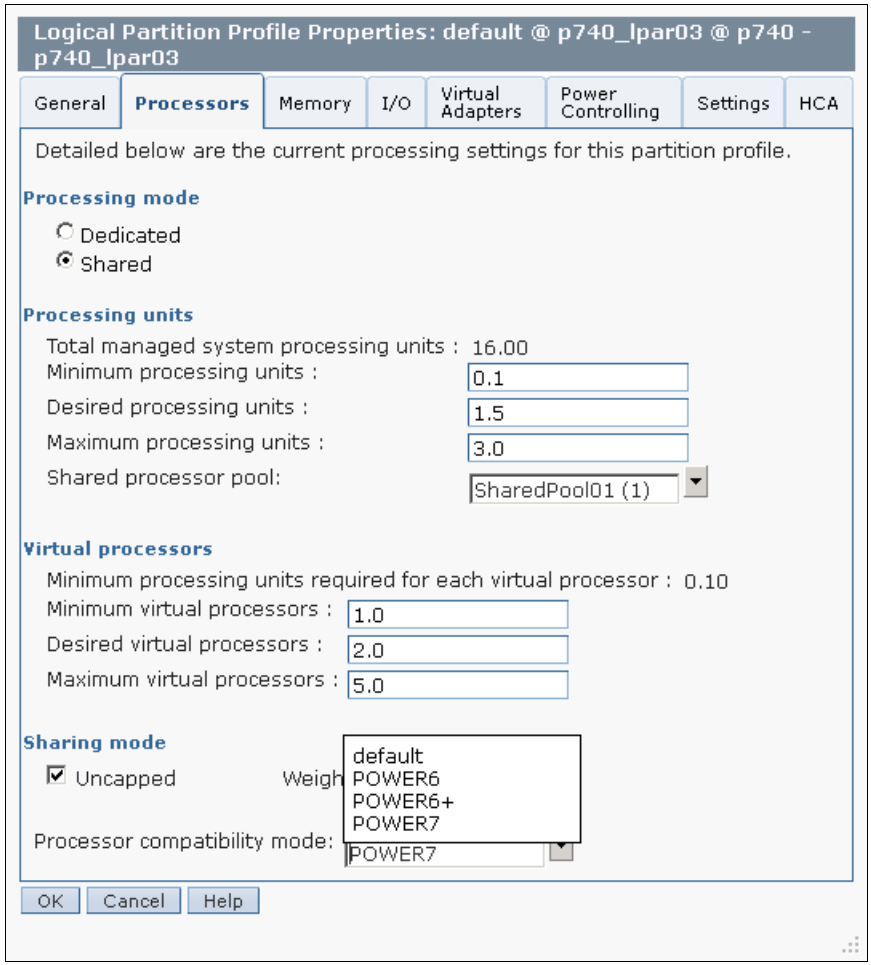


Figure 4-1 Changing POWER processor compatibility mode

3. Change to the value that you want and click **OK** when done.
4. Deactivate (shut down the partition) and then activate the partition to apply the selected mode.

4.2 Checking the compatibility mode

You can find which processor mode the partition is running using the HMC by accessing the GUI through the following steps:

1. Log in to HMC and select the system you want. On the right pane, select the partition that you want to check.
2. In the content panel, click **Task** and select **Properties.**
3. Click **Hardware**. You can then find the running mode in the Processors tab as shown in Figure 4-2.

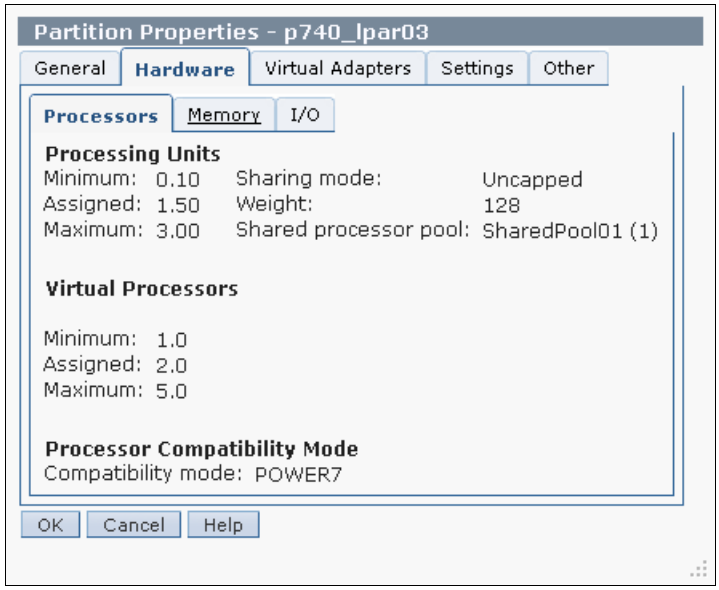


Figure 4-2 Checking the processor compatibility mode using the HMC

You can also check the compatibility mode by using the command line as shown in Example 4-1.

Example 4-1 Listing partitions and processor compatibility mode

```
hscroot@hmc8:~> lssyscfg -r lpar -m p740 -F
name,state,curr_lpar_proc_compat_mode --header

name,state,curr_lpar_proc_compat_mode
p740_lpar03,Running,POWER6+
p740_lpar02,Running,POWER7
p740_vios04,Running,POWER7
p740_vios03,Running,POWER7
```

```
p740_lpar01,Running,POWER7
```

If the partition is running AIX, you can also check the compatibility mode with the **prtconf** command as shown in Example 4-2.

Example 4-2 POWER7 mode with prtconf | grep Processor

```
p750_lpar01:/ # prtconf | grep Processor
Processor Type: PowerPC_POWER7
Processor Implementation Mode: POWER 7
Processor Version: PV_7_Compat
Number Of Processors: 1
Processor Clock Speed: 3000 MHz
```

However, it is not possible to differentiate between POWER6 and POWER6+ compatibility modes in **prtconf** as shown in Example 4-3.

Example 4-3 POWER6 or POWER6+ mode with prtconf | grep Processor

```
p740_lpar03:/ # prtconf | grep Processor
Processor Type: PowerPC_POWER7
Processor Implementation Mode: POWER 6
Processor Version: PV_6_Compat
Number Of Processors: 1
Processor Clock Speed: 3550 MHz
```

Part 2

Memory virtualization

This part addresses the managing and monitoring tasks for memory virtualization. The following topics are covered:

- Active Memory Sharing
- Active Memory Expansion
- Active Memory Deduplication

5

Active Memory Sharing

The goal of Active Memory Sharing (AMS) is to improve the overall utilization of physical memory. The hypervisor dynamically allocates physical memory from a shared pool to virtual servers, based on workload demands. AMS allows over-commitment of logical memory, which enables more virtual servers to run on a memory footprint. It can also allow virtual servers to run with larger memory configurations.

This chapter shows how to maintain and monitor Active Memory Sharing, and describes tuning configurations to improve performance by optimizing memory utilization.

This chapter includes the following sections:

- Managing
- Monitoring Active Memory Sharing

5.1 Managing

This section describes how the components of Active Memory Sharing are managed. The following topics are covered:

- Management of paging devices
- Management of the shared memory pool
- Management of shared memory partitions
- Dual Virtual I/O Server considerations

5.1.1 Requirements

Active Memory Sharing has the following requirements:

- POWER6, POWER7 hardware (firmware level 340_75 or later for POWER6).
- Virtual I/O Server v2.1.1.10-FP21 or later.
- HMC v7.3.4 with service pack 2 or later, or IVM.
- Minimum supported client operating systems:
  - AIX v6.1 TL3 on POWER6
  - AIX v6.1 TL4 on POWER7
  - IBM i V6R1 with PTF SI32798
  - SUSE Linux Enterprise Server 11 (minimum level 2.6.27.25-0.1-ppc64)
  - RHEL6.0
- All client virtual server I/O must be virtualized. No dedicated physical adapters can be assigned to the client partition.
- Only 4-KB memory page size is supported.
- Logical partitions must use shared processors only.

5.1.2 Paging device assignment

There is no fixed relationship between a paging device and a shared memory partition when a system is managed using the Management Console. The smallest suitable paging device is automatically selected when the shared memory partition is activated for the first time. The selected paging device must be at least as large as the maximum memory value in the partition profile for AIX and Linux virtual servers. For IBM i, the paging device must be at least as large as the maximum memory setting in the partition profile, plus 8 KB per MB of maximum memory.

After a paging device is selected for a partition, this device is used again if it is available when the partition is activated. However, if the paging device is unavailable, for example because it has been unconfigured or is in use by another partition, a new suitable paging device is selected.

Regarding IVM, the link is persistent and no implicit stealing of paging devices is done. When using IVM, if you want to reuse a paging device for another partition, you must explicitly remove it from the old partition.

The IVM CLI allows you to specify which paging device to use for each partition. By default, IVM picks one automatically, but you can override this selection manually.

Remember that when using IVM, you must have paging devices for every shared memory partition, whether activated or not (unless you are manually configuring using the IVM CLI).

To see which partition is using which paging device, select **Shared Memory Pool Management** from the Management Console, then click the Paging Space Device(s) tab as shown in Figure 5-1.

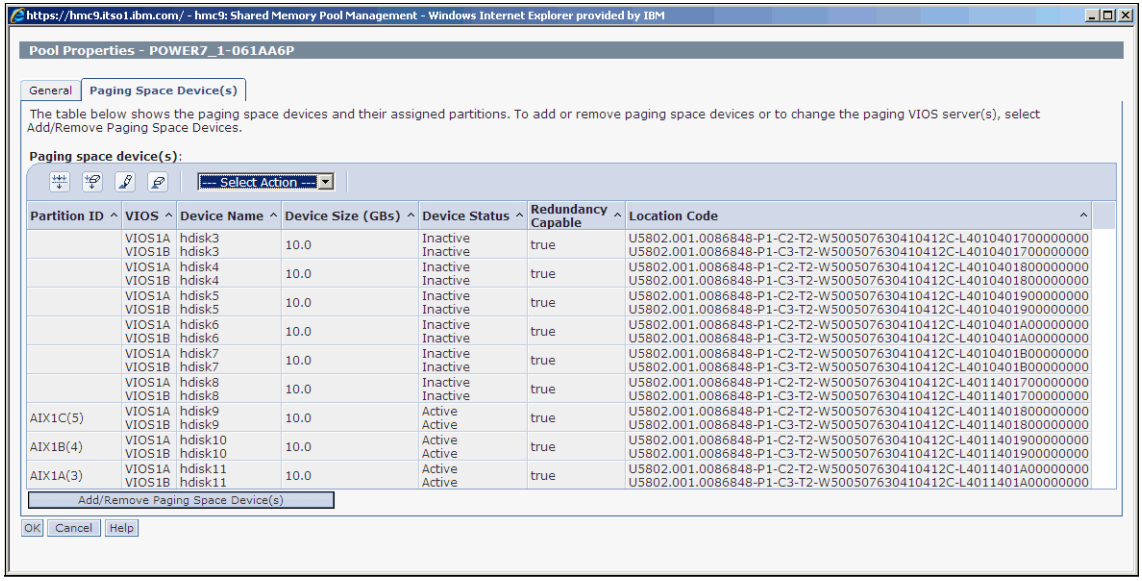


Figure 5-1 Pool properties settings

You can also use the **lshwres** command to display the paging device assignment, as shown in Example 5-1.

Example 5-1 Displaying paging devices by using lshwres

```
hscroot@hmc9:~> lshwres -m POWER7_1-061AA6P -r mempool --rsubtype pgdev
device_name=hdisk9,paging_vios_name=VIOS1A,paging_vios_id=1,size=1024,type=phys,state=Active,phys_loc=U58
02.001.0086848-P1-C2-T2-W500507630410412C-L4011401800000000,is_redundant=1,redundant_device_name=hdisk9,re
dundant_paging_vios_name=VIOS1B,redundant_paging_vios_id=2,redundant_state=Active,redundant_phys_loc=U5802
.001.0086848-P1-C3-T2-W500507630410412C-L4011401800000000,lpar_name=AIX1C,lpar_id=5
device_name=hdisk10,paging_vios_name=VIOS1A,paging_vios_id=1,size=1024,type=phys,state=Active,phys_loc=U5
802.001.0086848-P1-C2-T2-W500507630410412C-L4011401900000000,is_redundant=1,redundant_device_name=hdisk10,
redundant_paging_vios_name=VIOS1B,redundant_paging_vios_id=2,redundant_state=Active,redundant_phys_loc=U58
02.001.0086848-P1-C3-T2-W500507630410412C-L4011401900000000,lpar_name=AIX1B,lpar_id=4
device_name=hdisk11,paging_vios_name=VIOS1A,paging_vios_id=1,size=1024,type=phys,state=Active,phys_loc=U5
802.001.0086848-P1-C2-T2-W500507630410412C-L4011401A00000000,is_redundant=1,redundant_device_name=hdisk11,
redundant_paging_vios_name=VIOS1B,redundant_paging_vios_id=2,redundant_state=Active,redundant_phys_loc=U58
02.001.0086848-P1-C3-T2-W500507630410412C-L4011401A00000000,lpar_name=AIX1A,lpar_id=3
```

5.1.3 Adding paging devices

To add a paging device, first provision the physical disk or logical volume on the Virtual I/O Server, and add the paging devices to the pool. For more information, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940. Instead of the Management Console GUI you can also use the Management Console command line and enter the **chhwres** command, as shown in Example 5-2 on page 85.

Exception: Logical volumes or physical disks that have been mapped to a virtual SCSI adapter are not displayed in the selection list. However, it is possible to map a storage device to a virtual SCSI adapter even if it is defined as a paging device to the shared memory pool. No warning message is issued when the **mkvdev** command is run on the Virtual I/O Server.

5.1.4 Removing paging devices

To remove a paging device, deactivate the shared memory partition that is using the paging device. If a paging device is in use by a shared memory partition, it cannot be removed. Remove the paging device from the pool by using the Management Console.

After the paging device is removed, unconfigure the corresponding storage devices from the Virtual I/O Server. This can be done by using the Management Console GUI or by issuing the **chhwres** command on the Management Console command line, as shown in Example 5-2 on page 85.

When using IVM, if the paging device was already used by a partition, you must use the CLI to remove it.

5.1.5 Changing the size of a paging device

It is not possible to increase or decrease the size of an existing paging device. You must add a paging device with the size that you want, and then remove the old device. Before you remove the old paging device, deactivate the shared memory partition using it.

Adding and removing a paging device can be done by using the Management Console GUI or the Management Console command line. Example 5-2 shows how an existing paging device of 5 GB size is removed and a new 10 GB paging device is added by using the **chhwres** command.

Example 5-2 Removing and adding paging devices using chhwres

```
hscroot@hmc9:~> lshwres -m POWER7_1-061AA6P -r mempool --rsubtype pgdev
device_name=amspaging01,paging_vios_name=VIOS1A,paging_vios_id=1,size=5120,type=logical,state=Inactive,is_
redundant=0,lpar_id=none
device_name=amspaging02,paging_vios_name=VIOS1A,paging_vios_id=1,size=5120,type=logical,state=Inactive,is_
redundant=0,lpar_id=none

hscroot@hmc9:~> chhwres -m POWER7_1-061AA6P -r mempool --rsubtype pgdev -o r --device amspaging02 -p
VIOS1A
hscroot@hmc9:~> chhwres -m POWER7_1-061AA6P -r mempool --rsubtype pgdev -o a --device amspaging03 -p
VIOS1A

hscroot@hmc9:~> lshwres -m POWER7_1-061AA6P -r mempool --rsubtype pgdev
device_name=amspaging01,paging_vios_name=VIOS1A,paging_vios_id=1,size=5120,type=logical,state=Inactive,is_
redundant=0,lpar_id=none
device_name=amspaging03,paging_vios_name=VIOS1A,paging_vios_id=1,size=10240,type=logical,state=Inactive,is
_redundant=0,lpar_id=none
```

5.1.6 Managing the shared memory pool size

The size of the shared memory pool can be increased and decreased dynamically. To increase the shared memory pool, unused memory must be available. If you want to increase the shared memory pool beyond the maximum pool size, increase the maximum pool size to a value that is greater than or equal to the pool size you want. The maximum pool size can be increased dynamically.

Considerations:

- Keep the maximum pool size within reasonable limits. For each 16 GB of pool memory, the hypervisor reserves 256 MB of memory. Although this is not a requirement, consider reducing the maximum pool size when you reduce the pool size.
- Reducing the size of the shared memory pool might take more time if there is not enough free memory available. In such a case, the hypervisor first must page out memory. If you display the pool size during the time when the pool is being decreased, the figure displayed will be changing.

5.1.7 Deleting the shared memory pool

A shared memory pool cannot be deleted while any shared memory partitions are configured to use the pool. The partitions must be removed or changed to dedicated memory partitions before the shared memory pool is deleted.

Using the Management Console, when a partition is changed from shared memory to dedicated memory in the partition profile, you must activate it so that the partition picks up the new memory mode. If the partition is not activated in dedicated memory mode, the shared memory pool cannot be deleted. This is also true if the client virtual servers that use the pool are inactive, but the Virtual I/O Server managing that pool is active.

When using the IVM, the change from shared memory to dedicated memory is immediate.

To delete the shared memory pool, select **Shared Memory Pool Management** from the Management Console and then click **Delete Pool**, as shown in Figure 5-2.

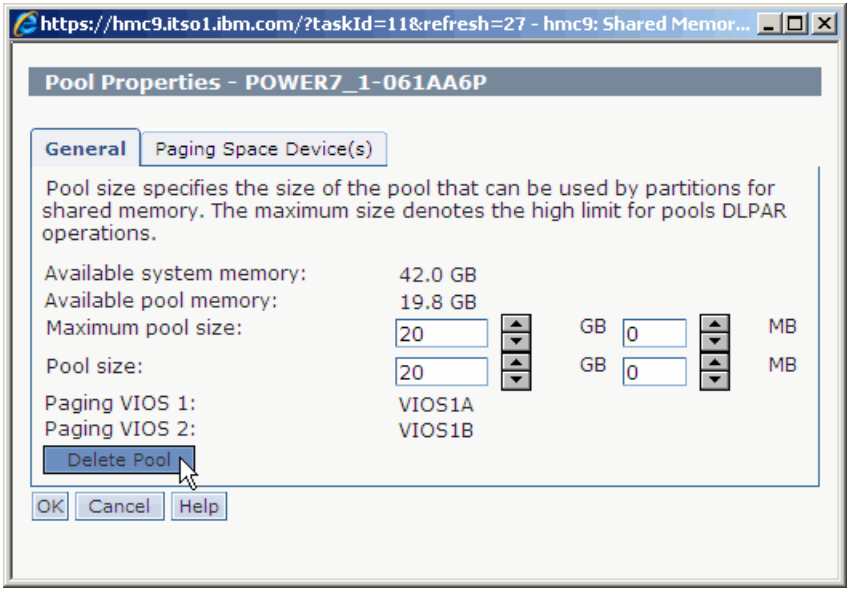


Figure 5-2 Shared memory pool deletion

After the shared memory pool is deleted, delete the storage pool that was automatically created for the paging devices if you want to reuse them. Select **Reserved Storage Device Pool Management** on the Management Console, then click **Delete Pool** as shown in Figure 5-3.

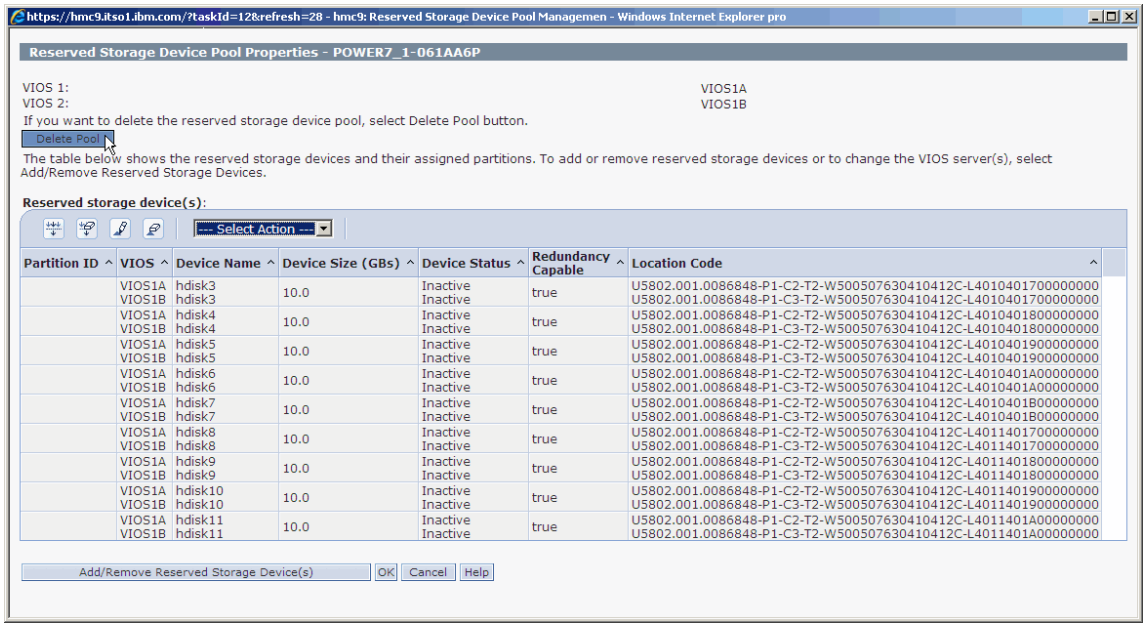


Figure 5-3 Paging devices reserved storage pool deletion

5.1.8 Dynamic operations for shared memory partitions

The amount of logical memory of a shared memory partition can be changed within the minimum and maximum boundaries that are defined in the partition profile. Increasing the logical memory in a shared memory partition does not mean that the amount of physical memory that is assigned through the hypervisor is changed. The amount of physical memory that a shared memory partition actually gets depends on the availability of free memory in the shared memory pool.

The memory weight of a shared memory partition can also be changed dynamically.

When you add or remove virtual SCSI, virtual Ethernet, or virtual FC adapters, the I/O entitled memory is automatically adjusted. The values used are based on default values and normally do not need to be changed. After they are changed, the values will remain at their new setting.

5.1.9 Switching between dedicated and shared memory

To change a partition from using shared memory to dedicated memory or back again, the memory mode setting in the partition profile must be changed. When you switch from dedicated to shared, you must not have any physical I/O devices configured. A warning window is displayed in this case. When you click **Yes,** all physical I/O assignments are removed.

Requirement: To activate the changed memory mode, the partition must be deactivated and reactivated. When using IVM, the partition status must be deactivated for the change to be allowed.

5.1.10 Starting and stopping the Virtual I/O Server

Before you activate a shared memory partition, the Virtual I/O Server that is used for paging must be active.

Note: The shared memory partitions cannot be activated until the Management Console has established an RMC connection with the Virtual I/O Server.

Before you shut down a Virtual I/O Server assigned as a paging space partition that has active shared memory partitions as clients, you must shut down the clients. Otherwise, the shared memory partitions lose their paging resources, and system paging errors occur (as happens with native paging devices). When you use the Management Console to deactivate a Virtual I/O Server with active shared memory partitions as clients, a warning message assists you in identifying this condition.

5.1.11 Dual Virtual I/O Server considerations

When using a Dual Virtual I/O Server configuration, you can configure your shared memory pool with two paging Virtual I/O Server. After you define a shared memory partition using this shared memory pool, you can assign two paging Virtual I/O Servers for the shared memory partitions: A primary and a secondary. The primary paging Virtual I/O Server is selected for AMS paging activity at partition activation. For more information, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

Only physical volumes are supported when you configure redundant access to paging devices from dual-VIO Servers.

When assigning paging devices to more than one VIO Server, the *reserve_policy* of the paging device is automatically set to *no_reserve*.

Failover and load balancing

In a Virtual I/O Server outage, affected partitions switch to their secondary paging Virtual I/O Server. Unless the system is rebooted, the partitions remain on the switched Virtual I/O Server. There is no failback to the primary paging Virtual I/O Server, unless the current paging Virtual I/O Server has an outage. You can, however, force a failover to the other Virtual I/O Server by using the Management Console command shown in Example 5-3.

Example 5-3 Switching the paging Virtual I/O Server for a partition

```
hscroot@hmc9:~> lshwres -r mem -m POWER7_1-061AA6P --level lpar -F
lpar_name curr_paging_vios_name
AIX1C VIOS1A
AIX1B VIOS1A
AIX1A VIOS1B
hscroot@hmc9:~> chhwres -r mem -m POWER7_1-061AA6P -p AIX1C -o so
hscroot@hmc9:~> lshwres -r mem -m POWER7_1-061AA6P --level lpar -F
lpar_name curr_paging_vios_name
AIX1C VIOS1B
AIX1B VIOS1A
AIX1A VIOS1B
```

Because the primary and secondary paging Virtual I/O Server are defined at the partition level, you can evenly assign paging Virtual I/O Server as primary and secondary to your shared memory partitions to load balance AMS paging activity. Keep in mind that if Virtual I/O Server maintenance or failure occurs, all your partitions will be on the same remaining Virtual I/O Server. You must then force a failover for some of your partitions when the updated/failing Virtual I/O Server is back online. To do so, use the Management Console command line as shown in Example 5-3 to evenly spread your partitions over the two paging Virtual I/O Servers.

Paging devices

With a dual Virtual I/O Server configuration, if your partition has a primary and a secondary paging Virtual I/O Server defined, both Virtual I/O Servers must access a common paging device with a size greater than or equal to the partition maximum memory. Otherwise, the partition fails to activate.

You can force the activation of a partition using a non-redundant paging device by using the **chsysstate** command with the **--force** flag on the Management Console as shown on Example 5-4. However, the partition will not support failover.

Example 5-4 Forcing the activation of a partition with non-redundant paging device

```
hscroot@hmc9:~> chsysstate -r lpar -m POWER7_1-061AA6P -o on -n AIX1C -f AMS
HSCLA47C Partition AIX1C cannot be activated with the paging Virtual I/O Server
(VIOS) partition configuration specified in the profile because one of the
paging VIOS partitions is not available, or a paging space device that can be
used with that paging VIOS configuration is not available. However, this
partition can be activated with a different paging VIOS partition configuration
now. If this partition is configured to use redundant paging VIOS partitions,
then this partition can be activated to use a non-redundant paging VIOS
partition. If this partition is configured to use non-redundant paging VIOS
partitions, then this partition can be activated to use a different paging VIOS
partition than the one specified in the profile. If you want to activate this
partition with the paging VIOS configuration that is available now, then run
the chsysstate command with the --force option to activate this partition.

hscroot@hmc9:~> chsysstate -r lpar -m POWER7_1-061AA6P -o on -n AIX1C -f AMS
--force
```

5.1.12 Tuning

This section describes how to tune Active Memory Sharing. The optimum configuration can be obtained by monitoring and then optimizing two main areas:

- The shared memory pool
- Each single shared memory partition

Shared memory pool tuning

The performance of a shared memory configuration depends on the number of page faults managed by the hypervisor plus the time required to run paging activity on the paging devices.

Memory sharing is designed to reduce the physical memory size that is required to run system workload, and to automatically reallocate memory where it is most needed. Memory reduction is accomplished by allowing over-commitment of real memory. Page stealing and page faults are required to keep the shared memory partitions running in this environment, and cannot be completely avoided. A balance between virtual server performance and physical memory optimization is the goal of tuning.

The number of acceptable page faults depends on the difference between the forecasted and effective over-commitment of memory. This value varies from case to case. Use the tools that are described in 5.2, “Monitoring Active Memory Sharing” on page 102 to track changes. Then, modify the configuration accordingly while tuning a system for production workloads.

Shared memory pool configuration

A shared memory pool has one global configuration value: The size of the pool. The size can be dynamically changed based on pool usage and the number of page faults.

If the system has deallocated memory, it can be added to the memory pool. By increasing the size of the pool, you affect the over-commitment of the pool by reducing the need to page out the memory to the Virtual I/O Server-managed paging devices. The hypervisor assigns the additional memory pages to the logical partitions with higher memory demand. If page faults and paging activity are present during pool enlargement, they might not decrease immediately because some memory pages still must be read from the disk. However, increasing the size of the shared memory pool is immediate, provided that sufficient free memory is available. The pool maximum size can also be increased dynamically.

Pool size reduction can be done if memory is required for dedicated memory partitions. The removal of memory from the pool must be carefully planned. The hypervisor must free physical memory, and might need to page out memory to disks. In this case, shared memory partition memory access time might be affected. Memory is removed from the shared memory pool in logical memory blocks (LMBs). Each LMB can have in-use pages or have pages that are pinned by I/O adapters. Therefore, the system free memory might only change by small increments. It can take a considerable amount of time for the hypervisor to reduce the pool by the full amount of memory requested.

Note: Generally, perform pool size reduction when the load on the shared memory partitions is low.

Virtual I/O Server configuration

The Virtual I/O Server is involved in shared memory pool tuning because it runs the disk paging activities on behalf of the hypervisor.

Disk access time contributes to memory access time only when the hypervisor requires paging activity. This can be monitored by using the page-in delay statistics on the operating system on the shared memory partition, or by the Management Console or IVM performance data.

Paging device selection

The memory pool's paging devices can be backed either by a disk device or a logical volume provided by the Logical Volume Manager (LVM) of the Virtual I/O Server.

The LVM can split a single device into logical volumes, allowing the device to support multiple paging devices. It is *not* recommended, though possible, to use the physical storage that is occupied by the Virtual I/O Server operating system. However, logical volumes that belong to other storage pools can be used.

The Virtual I/O Server's physical volumes are preferred for paging devices because their usage provides a slightly shorter instruction path length to the physical storage. They also simplify the performance monitoring of the device because it is a unique paging device. Select a logical volume when you cannot use the storage system to split the disk space in LUNs, or if disk resources are scarce.

To improve paging device performance, also consider keeping the client virtual disk devices provided by the Virtual I/O Server separate from the hypervisor paging devices. If both workloads share devices, they might compete for disk access. A possible solution is to have two Virtual I/O Servers, one handling paging devices and the other virtual disks, but this setup is not a requirement.

Paging device assignment to shared memory partitions is made at logical partition startup by the Management Console or IVM. the assignment is based on the size of the logical partition's maximum logical memory and the size of each defined paging device. The smallest possible paging device is chosen. If available, the same device is used at each logical partition activation. Otherwise a new one is selected.

Because the selection of a paging device is automatically assigned by the Management Console or IVM, generally create all paging devices on storage with similar performance characteristics. Although it is possible to have a mixture of device types (for example, logical volumes on internal SAS disks and LUNs on an IBM DS8300 subsystem), you have no control over which device is assigned to which logical partition. Using the IVM CLI, you can manually assign paging devices to partitions.

Storage configuration

Disk paging activity has the following characteristics:

- Very random I/O operations
- Typically performed in blocks of 4 KB in size

Paging activity can be improved by using performance-oriented disk storage devices that are configured to match as closely as possible the I/O patterns requested by the Virtual I/O Server. Disk availability is also important to avoid disk failures that might lead to unavailability of memory pages and shared memory partition failure.

Consider the following suggestions for disk paging setup:

- Spread the I/O load across as many physical disks as possible.
- Use disk caches to improve performance. Because of the random access nature of paging, write caches provide benefits, whereas read caches might not have an overall benefit.
- 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.
- Use disk redundancy. For a storage subsystem, a configuration that uses mirroring or RAID5 is appropriate. The Virtual I/O Server cannot provide redundancy.
- Perform the usual I/O tuning on the Virtual I/O Server to improve disk access time. This tuning includes queue depth values and specific parameters that are provided by the storage provider.

Processor configuration

Paging activity by the Virtual I/O Server is run using processor resources. Although the amount of processor consumption might not be a concern in many environments, monitor it to avoid performance effects. This is especially important if the Virtual I/O Server provides other services such as shared Ethernet adapters. The usual monitoring and tuning apply, as with any new environment.

As a general guideline, configure the Virtual I/O Server as an uncapped shared processor partition with enough virtual processors to manage peak loads, and a high processor weight.

Shared memory partition tuning

In all shared resource environments, it is important to evaluate either global resource consumption or the behavior of each resource user. When multiple logical partitions compete for memory access, compare performance values from all logical partitions. Be aware that changes to one logical partition's configuration might affect the others.

Memory sharing policy and memory load factor

AMS allows memory pages to flow between partitions based on hypervisor's perception of the relative memory loads.

Periodically, the hypervisor computes a memory load factor for each AMS partition. This calculation is based on a combination of the following metrics:

- Accumulated running average of hypervisor page faults that are incurred by the partition
- Accumulated running average of page table faults that are incurred by the partition
- Accumulated running average of OS memory load reported by the operating system of the partition
- Memory efficiency of the partition (proportion of used memory pages to unused pages that are assigned to the partition)
- User-assigned memory weight

Each of these contributing metrics is normalized in a fashion that shows their relationship to the other partitions. The calculated load factors of each partition are then compared and analyzed so the hypervisor can decide how to share the memory.

Apart from the memory weight factor, there is no possible tuning here. However, keeping in mind how AMS works helps you better fine-tune your infrastructure.

The flow of memory between partitions is not instantaneous. The hypervisor must see sustained demand from a partition before giving it significantly more memory at the expense of other partitions.

By default, the hypervisor does share the full amount of physical memory from the shared pool between the shared-memory partitions if this is not needed. Each partition receives only the physical memory that it needs according to the memory allocation it does. Therefore, if the sum of the working set of each partition does not exceed the shared memory pool size, you will see free memory in the pool.

Logical partition

The following parameters define a shared memory partition:

- The logical memory size
- The memory weight
- The I/O entitled memory

All three parameters can be dynamically modified, as explained here in the following sections.

Logical memory size

The logical memory size is the quantity of memory that the operating system can address. Depending on how much logical memory is used and the memory demand on the entire memory pool, the logical memory can use either physical memory or a paging device.

If the shared memory partition requires more memory than the provided logical memory, the operating system can provide virtual memory by using its own paging space. If an operating system runs local paging, it does not affect the shared memory pool performance.

Size the logical memory based on the maximum quantity of memory that the logical partition is expected to use during peak hours, or the maximum physical memory that you want the logical partition to allocate. Monitor the usage of physical memory to check whether over-commitment occurs.

Logical memory can be resized. When it is reduced, the operating system must first free the logical memory by using its local paging space to keep providing the same amount of virtual memory in use by the applications.

Logical memory reduction of selected shared memory partitions can be performed to reduce the load on the memory pool. The paging activity is then moved from the pool to each operating system.

If logical memory is increased, the logical partition's addressable memory is increased. However, the effective quantity of physical memory that is assigned depends on the global load on the pool and on the logical partition's access to logical memory. Increasing the logical memory does *not* imply that the logical partition has more physical memory pages assigned to it.

Memory weight

When there is physical memory over-commitment, the hypervisor must decide how much memory to assign to each logical partition and which pages must be copied on the paging devices. The choice is made by evaluating global memory usage, global memory load, and the memory weight that are assigned by all active shared memory partitions.

If the page fault rate for one logical partition becomes too high, you can increase the memory weight assigned to it to improve the amount of physical memory it receives. The weight change is immediately applied, but you must monitor the effects over a short interval because hypervisor paging might be required to rebalance memory assignment.

Note: Memory weight is just one of the parameters that are used by the hypervisor to decide how many physical pages are assigned to the shared memory partitions.

I/O entitled memory

The I/O entitled memory value represents the memory that is allowed to be permanently kept in the memory pool to handle I/O activity.

When a shared memory partition is started, the I/O entitled memory is automatically configured based on the number and the type of adapters defined for the logical partition. This value is defined to satisfy I/O requests for most configurations. If the operating system shows that some I/O has been delayed, you can increase the I/O entitlement.

I/O entitlements can be added by running a dynamic LPAR reconfiguration on the logical partition's memory configuration.

Note: After you begin to manually tune entitled memory, the Management Console and IVM will no longer automatically manage it for the partition. For example, if you add or remove a virtual adapter, the entitled memory that is assigned to the partition is not adjusted by these applications.

Processor configuration

The effect of Active Shared Memory on computing time is minimal, and is primarily tied to extra address translation. Some additional processor cycles are required to manage the paging activity that is managed by the Virtual I/O Server. However, they are similar to the cycles that are required in case the over-commitment of memory must be managed on a dedicated memory partition by the operating system that is issuing paging activity on its own paging device.

When a logical partition accesses a logical memory page that is not in the physical memory pool, the virtual processor performing the access cannot proceed. It is no longer dispatched on a physical processor until the logical memory page is available. The temporary unavailability of that virtual processor shifts the load of the logical partition onto the remaining virtual processors.

Calculate the number of configured virtual processors on a shared memory partition so that when a high page fault rate occurs, the running virtual processors can sustain the workload. For example, if your sizing requires 3.0 processing units and six virtual processors to handle all the load peaks, configure an extra three virtual processors.

Note: AIX memory tuning described in “Loaning” on page 98 can be used to shift paging activity from the hypervisor to AIX. Reduced hypervisor paging lowers the need for more virtual processors.

AIX operating system

There are specific tuning capabilities for AIX that can be used for Active Memory Sharing. They are related to loaning and I/O memory.

Loaning

AIX interacts with the hypervisor for shared memory handling by classifying the importance of logical memory pages and receiving the request to loan them from the hypervisor. Loaned logical memory pages are kept free by AIX for a longer time, and are not shown in the free statistics of commands such as **vmstat** or **lparstat**. AIX can reclaim loaned memory if needed.

When the hypervisor must reduce the number of physical memory pages that are assigned to a logical partition, it first selects loaned memory pages, with no effect on logical partition’s memory access time. If more memory pages are required, the hypervisor first selects logical free memory and then logical used memory. Both of these might cause hypervisor page faults if referenced by AIX.

Used logical memory pages can be used to cache file data or to store working storage (process code and data). AIX tries to maximize its usage of logical memory by caching as much file data as possible. In a shared memory partition, excessive file caching might cause unwanted memory pool consumption.

When AIX starts to loan logical memory pages, it first selects pages that are used to cache file data.

It is possible to tune the loaning process by modifying the ams_loan_policy attribute by using the **vmo** command, as follows.

vmo -a ams_loan_policy=0	Disable page loaning.
vmo -a ams_loan_policy=1	Default value. Only select file cache pages for loaning.
vmo -a ams_loan_policy=2	Aggressive loaning: when file cache is depleted, continue loaning by paging out working storage pages.

When page loaning is disabled, AIX stops adding pages in the loaning state, even if requested by the hypervisor. When the hypervisor must reduce the logical partition’s memory footprint and free pages are already selected, either file cache pages or working storage can be moved into the hypervisor’s paging space.

With the default loaning configuration, AIX performs loaning and targets only pages that are used for file caching. The working storage pages are never selected for loaning. When the hypervisor must reduce the number of physical memory pages that are assigned to the logical partition, it first selects loaned pages, and then free and used memory pages. The effect of loaning is to reduce the number of hypervisor page faults because AIX reduces the number of active logical pages and classifies them as loaned.

Note: The default page loaning configuration is suitable for most production workloads, but some environments can take advantage of a different loaning setup. For example, some applications run direct I/O and cache data themselves, and so AIX file cache is minimal and default loaning has minimal effect.

Monitor the loaning changes to check the results.

The aggressive loaning policy allows AIX to loan a higher number of logical memory pages by using pages owned by applications when all file cache pages have been loaned. When AIX selects a working storage page, it is first copied to the local paging space. This setup reduces the effort of the hypervisor by moving paging activity to AIX.

Note: Aggressive loaning can cause extra AIX paging activity. To minimize the effect of loading, increase the size of the operating system's paging space and place it on fast disk devices. Start with a paging space size equal to the size of logical memory that is assigned to the logical partition.

Figure 5-4 illustrates the effect of the three page loaning configurations. The upper half of the figure shows the status of the logical memory as provided by AIX. At the beginning, all logical memory has physical memory assigned to it, except for loaned pages, as shown by the vertical bar near logical memory. AIX is using most of the active logical pages to cache file data.

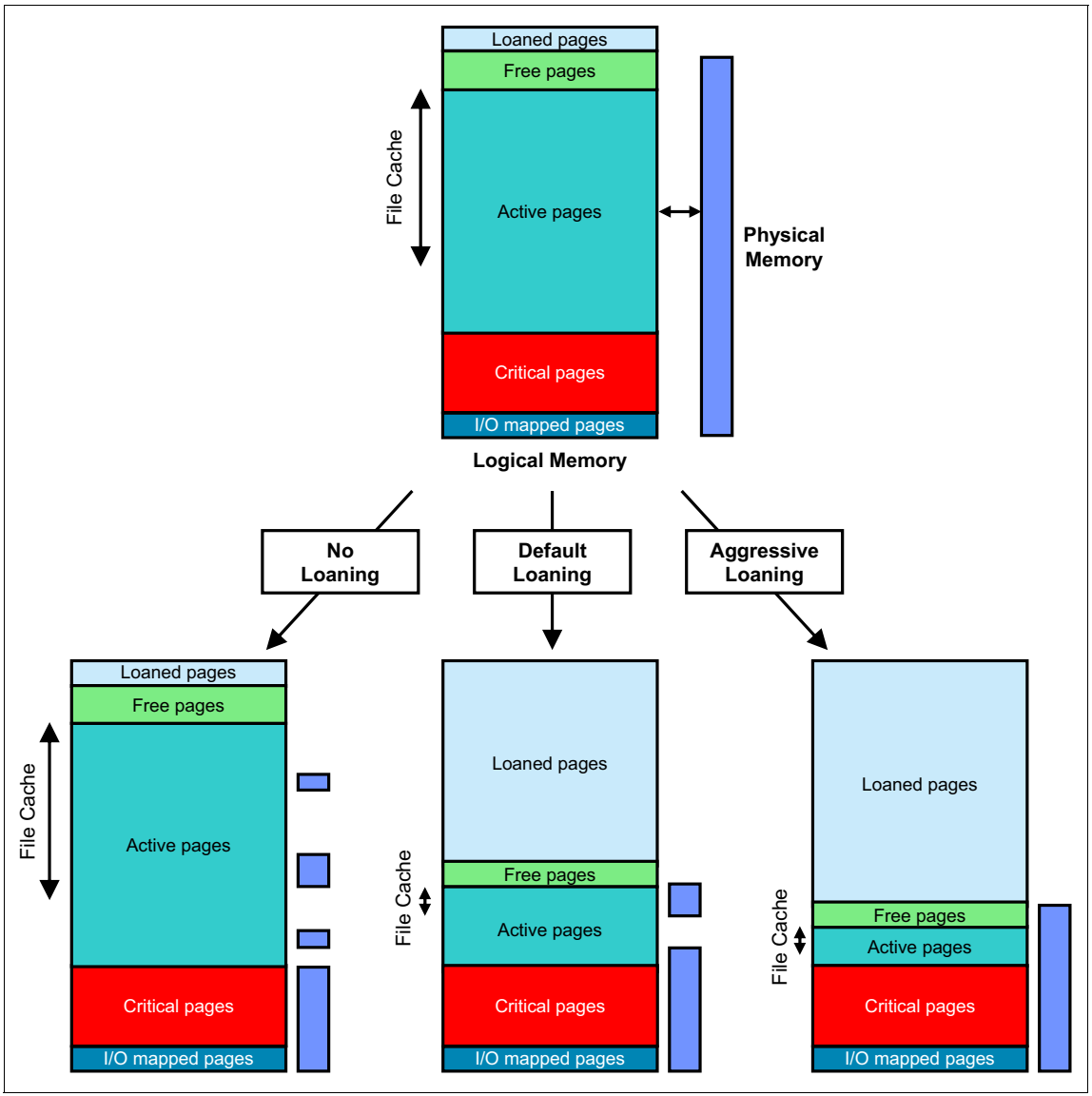


Figure 5-4 AIX loaning tuning example

When the hypervisor must reduce the physical memory that is assigned to the logical partition, it starts by asking for loans. If loans do not fill its requests, it begins stealing logical memory pages. The bottom half of the figure shows three possible scenarios, depending on the loaning configuration of AIX.

Loaning disabled

When loaning is disabled, AIX performs no action and the hypervisor steals first free and then active memory pages. Both file cache and working set pages are sent to the hypervisor's paging space because the hypervisor cannot differentiate how the pages are used by the operating system.

Default loaning

With the default loaning configuration, AIX first reduces the number of logical pages that are assigned to the file cache and loans them. The hypervisor steals first the loaned pages and then free pages, and finally copies a few working storage pages into its paging space.

Aggressive loaning

If page loaning is set to aggressive, AIX either reduces the file cache or frees more working storage pages by copying them into the AIX paging space. The number of loaned pages is greater than is the case with default loaning. The hypervisor uses the loaned pages, and might not need to run any activity on its paging space.

I/O memory pages

I/O memory pages are used by AIX to run I/O operations. They must be kept in the shared memory pool to avoid delays in I/O.

I/O entitled memory is defined when the logical partition is started. It represents the maximum number of I/O memory pages that AIX can use. AIX normally uses only a subset of such entitlement. The default values are suited for most configurations.

You can monitor actual memory entitlement usage by using the **lparstat** command as shown in Example 5-5. You can then decide whether to dynamically increase the I/O entitlement of the logical partition.

Example 5-5 I/O memory entitlement monitoring

```
# lparstat -m 2

System configuration: lcpu=8 mem=10240MB mpsz=20.00GB iome=334.00MB
iomp=10 ent=1.00

physb   hpi  hpit  pmem  iomin   iomu   iomf  iohwm iomaf %entc  vcsw
```

```
----- ----- ----- ----- ------ ------ ------ ------ ----- ----- -----
 1.64     0     0  2.88  287.7   12.4   34.3   14.6     0   2.6   752
 1.80     0     0  2.88  287.7   12.4   34.3   14.6     0   3.0   484
```

The `iomaf` value displays the number of times that the operating system has attempted to get a page for I/O and failed. If the value is not zero (0), increase the I/O entitlement. The `iomu` and `iomf` values represent the I/O entitlement used and the free I/O entitlement. If the `iomu` value is near the logical partition's I/O entitlement, or if the `iomf` value is constantly low, increase the I/O entitlement.

I/O entitlements can be added by running a dynamic LPAR reconfiguration on the logical partition's memory configuration.

IBM i operating system

There are several special considerations when you are tuning ASM on the IBM i operating system.

Loaning

On IBM i, you cannot change the behavior of the loaning process. Everything is managed by the hypervisor.

I/O memory pages

You can monitor actual memory entitlement usage by checking the IBM i Collection Services QAPMSHRMP table. The fields SMENTIOC, SMMINIOC, SMOPTIOC, SMIOCUSE, SMIOCMAX, and SMIOMDLY display I/O dependencies for the actual configuration. If the number of Partition I/O mapping delays is not zero, increase the I/O entitlement.

I/O entitlements can be added by running a dynamic LPAR reconfiguration on the logical partition's memory configuration. This parameter is set to auto adjustment by default.

5.2 Monitoring Active Memory Sharing

Metrics for monitoring Active Memory Sharing are available on:

- Management Console
- AIX or IBM i shared memory partitions
- Virtual I/O Server
- Linux

This section describes the commands and tools that are available.

5.2.1 Management Console

The Management Console provides Active Memory Sharing statistics through a utilization data collection feature. A mempool resource type has been added to the `lslparutil` command. Some fields have also been added to the lpar resource type.

The following fields are available for the mempool resource type:

curr_pool_mem Current amount of physical memory that is allocated to the shared memory pool.

lpar_curr_io_entitled_mem Current amount of physical memory that is defined as I/O entitled memory by all shared memory partitions.

lpar_mapped_io_entitled_mem Current amount of physical memory that is actually used as I/O entitled memory by all shared memory partitions.

page_faults Number of page faults by all shared memory partitions since the system was booted.

lpar_run_mem Current amount of physical memory that is used by all shared memory partitions.

page_in_delay Page-in delay for all shared memory partitions since the system was booted.

sys_firmware_pool_mem The amount of memory in the shared pool that is reserved for use by the firmware.

The following fields are available for the lpar resource type:

mem_mode Memory mode of the partition (`ded` for dedicated memory partitions, `shared` for shared memory partitions).

curr_mem Current amount of logical memory for the partition.

curr_io_entitled_mem Current amount of logical memory that is defined as I/O entitled memory for the partition.

mapped_io_entitled_mem Current amount of logical memory that is actually mapped as I/O entitled memory for the partition.

phys_run_mem Current amount of physical memory that is allocated to the partition by the hypervisor.

run_mem_weight	Current memory weight for the partition.
mem_overage_cooperation	Current amount of logical memory that is not backed by physical memory. This is a negative value because it is calculated as **phys_run_mem** minus **curr_mem**. At the time of writing, this resource type is not part of the IVM.

The performance data is taken in a timely manner by the management console. The default sampling is 5 minutes, but you can change the value to better meet your requirements by using the **chlparutil** command. In Example 5-6, the **lslparutil** command is used to list the current `sample_rate` for both systems. The **chlparutil** command is then used to change the sampling rate to 1 minute only in the POWER7_1-061AA6P system.

Example 5-6 Listing and changing the sample rate

```
hscroot@hmc9:~> lslparutil -r config
type_model_serial_num=8233-E8B*061AA6P,name=POWER7_1-061AA6P,sample_rat
e=300
type_model_serial_num=8233-E8B*061AB2P,name=POWER7_2-061AB2P,sample_rat
e=300

hscroot@hmc9:~> chlparutil -r config -m POWER7_2-061AB2P -s 60

hscroot@hmc9:~> lslparutil -r config
type_model_serial_num=8233-E8B*061AA6P,name=POWER7_1-061AA6P,sample_rat
e=60
type_model_serial_num=8233-E8B*061AB2P,name=POWER7_2-061AB2P,sample_rat
e=300
```

Example 5-7 shows how to use the **lslparutil** command to get information based on the performance data from the last 24 hours. It is also possible to check the last days (**-d**), minutes (**--minutes**), and others. For more information, see the **lslparutil** man pages. The output of the following example as been truncated for readability purposes.

Example 5-7 Using lslparutil to monitor memory utilization on LPARs running AMS

```
hscroot@hmc9:~> lslparutil -m POWER7_2-061AB2P -d 7 -r lpar -F
time,lpar_name,phys_run_mem

04/28/2011 11:46:31,IBMi2A,2048
04/28/2011 11:46:31,IBMi2B,488
04/28/2011 11:46:31,LINUX2A,699
```

```
04/28/2011 11:46:31,LINUX2B,473
04/28/2011 11:46:31,AIX2A,1120
04/28/2011 11:46:31,AIX2B,1156
[...]
04/27/2011 12:10:05,IBMi2A,2048
04/27/2011 12:10:05,IBMi2B,2076
04/27/2011 12:10:05,LINUX2A,2641
04/27/2011 12:10:05,LINUX2B,5282
04/27/2011 15:46:05,AIX2A,1431
04/27/2011 15:46:05,AIX2B,1373
```

When you display the usage rates with the Management Console GUI, extra values are calculated from the collected data, such as memory overcommitment. Figure 5-5 shows an example. To display the utilization data with the Management Console GUI, select a managed system and click **Operations** → **Utilization Data** → **View**. Utilization data collection must be enabled for the data to be available.

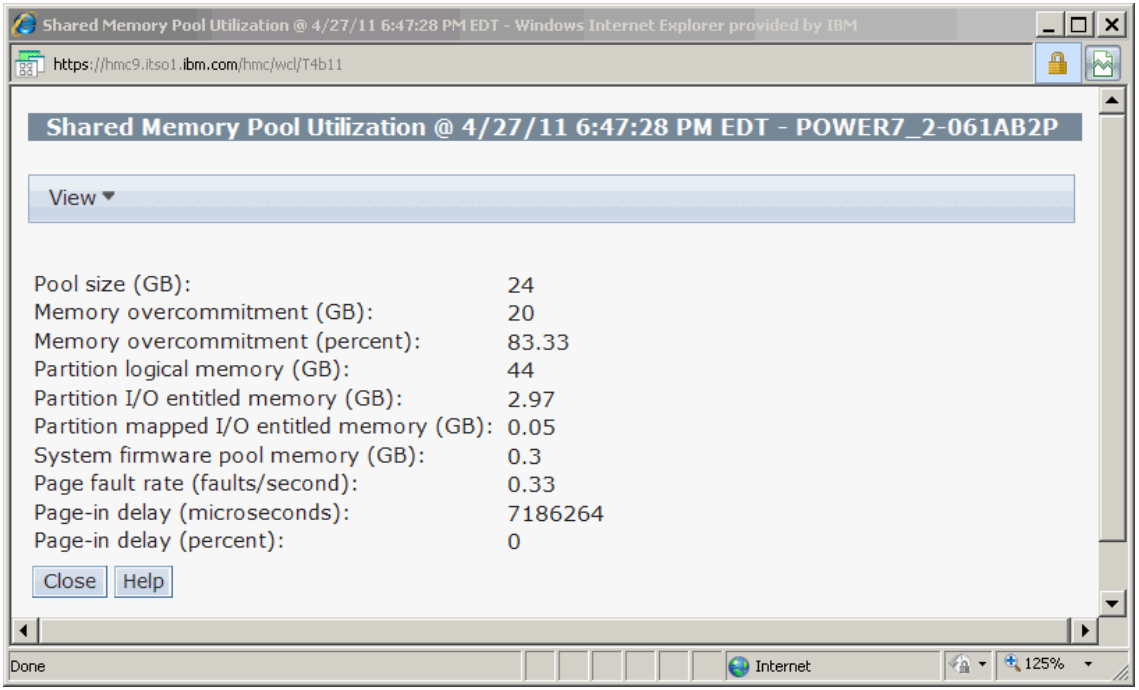


Figure 5-5 Displaying shared memory pool utilization

As shown in Figure 5-6, you can also check the memory distribution per partition from the Management Console. Select a managed system, then click **Operations** → **Utilization Data** → **View**. Select a snapshot and click **View** → **Partition** → **Memory**.

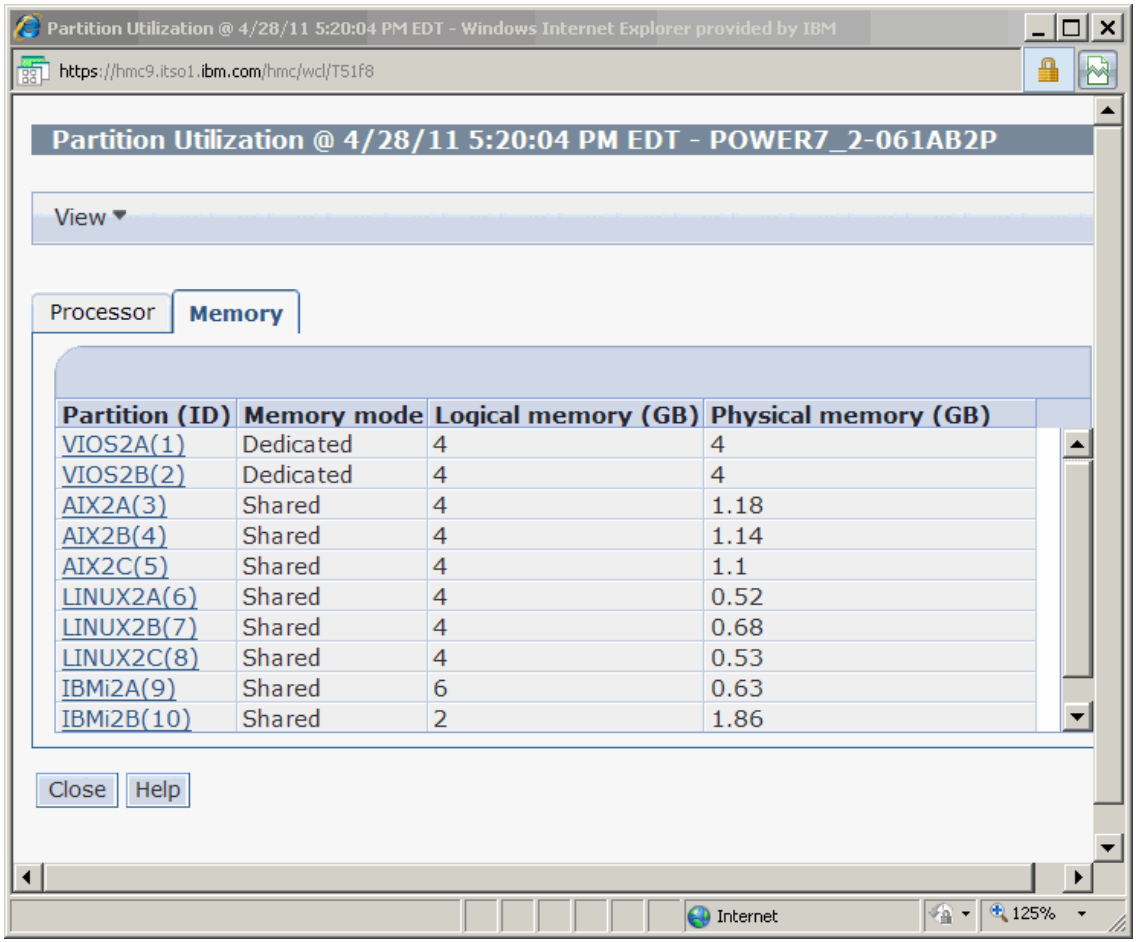


Figure 5-6 Memory utilization per partition

The I/O entitled memory for a specific shared memory partition can be displayed using the partition properties. Select a partition and click **Properties** → **Hardware** → **Memory**. Click **Memory Statistics** to display the window, as shown in Figure 5-8 on page 108. The following values are displayed:

Assigned I/O Entitled Memory — I/O entitled memory that is assigned to the partition by the Management Console or IVM.

Minimum I/O Entitled Memory Usage	Minimum I/O entitled memory reported by the OS, or through the Management Console or IVM by using the **`lshwres`** command.
Optimal I/O Entitled Memory Usage	Optimal amount of I/O entitled memory reported by the OS, Management Console, or IVM.
Maximum I/O Entitled Memory Usage	High water mark of used I/O entitled memory reported by the OS. The high water mark can be reset by clicking **Reset Statistics**, or by using the **`chhwres`** command on the Management Console or IVM interface.

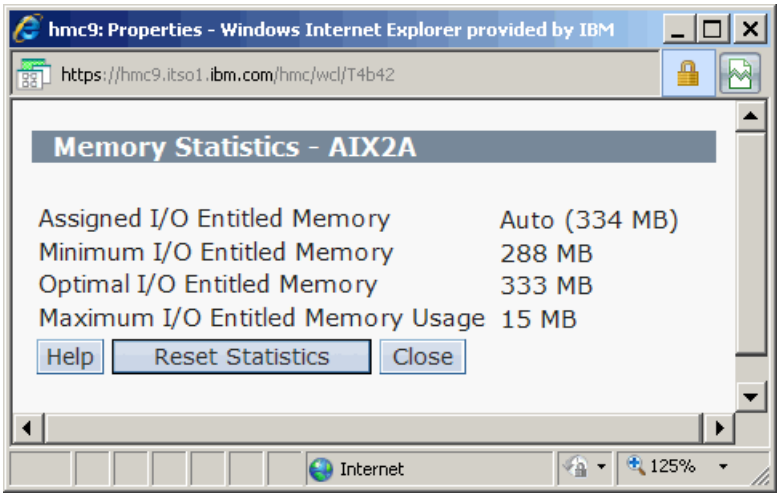


Figure 5-7 I/O entitled memory statistics

5.2.2 Virtual I/O Server monitoring

The I/O performance of the physical disks or logical volumes that serve as paging devices on the Virtual I/O Server must be monitored. The Virtual I/O Server provides the **`viostat`** and **`topas`** commands for displaying I/O performance data.

In AMS deployments, Virtual I/O Server has the critical role of providing paging devices to the shared pool. Therefore, the most important aspects to monitor in a Virtual I/O Server are processors and disk I/O utilization for the paging devices.

To monitor processing, you can run the **topas** command to see how the Virtual I/O Server is performing, as shown in the Figure 5-8.

```
Topas Monitor for host:    VIOS2A              EVENTS/QUEUES    FILE/TTY
Thu Apr 28 09:36:54 2011   Interval:  2        Cswitch     369  Readch     2781
                                               Syscall     263  Writech     245
CPU  User%  Kern%  Wait%  Idle%  Physc   Entc   Reads        31  Rawin         0
ALL    0.1    4.0    0.0   95.9   0.04    7.0   Writes        3  Ttyout       85
                                               Forks         0  Igets         0
Network  KBPS   I-Pack  O-Pack   KB-In  KB-Out  Execs         0  Namei        53
Total    36.4     27.4     2.0    36.1     0.3  Runqueue    1.0  Dirblk        0
                                               Waitqueue   0.0
Disk    Busy%     KBPS     TPS KB-Read KB-Writ                   MEMORY
Total     0.0      0.0     0.0     0.0     0.0  PAGING           Real,MB    4096
                                               Faults        0  % Comp       24
FileSystem        KBPS     TPS KB-Read KB-Writ  Steals        0  % Noncomp     5
Total              2.7    30.4    2.7     0.0   PgspIn        0  % Client      5
                                               PgspOut       0
Name            PID  CPU%  PgSp Owner           PageIn        0  PAGING SPACE
pager0      2490458   2.8   1.0 root            PageOut       0  Size,MB    1536
vmmd         458766   0.6   1.2 root            Sios          0  % Used        1
topas       5308632   0.2   1.6 padmin                           % Free       99
xmgc         851994   0.2   0.4 root            NFS (calls/sec)
seaproc     7274572   0.1   1.0 padmin          SerV2         0  WPAR Activ    0
```

Figure 5-8 Monitoring VIOS by using topas

To identify which disks are being used as paging devices at the moment, use **lshwres** on the Management Console, as shown in Example 5-8.

Example 5-8 Disks being used as paging devices

```
hscroot@hmc9:~> lshwres -r mempool --rsubtype pgdev -m POWER7_2-061AB2P
-F paging_vios_name,device_name,lpar_name
VIOS2A,hdisk3,AIX2C
VIOS2A,hdisk4,AIX2B
VIOS2A,hdisk5,AIX2A
VIOS2A,hdisk6,LINUX2C
VIOS2A,hdisk7,LINUX2A
VIOS2A,hdisk8,LINUX2B
VIOS2A,hdisk9,IBMi3C
VIOS2A,hdisk10,IBMi2B
VIOS2A,hdisk11,IBMi2A
```

Paging activity can be monitored by checking I/O activity on disks used as paging devices by using the **viostat** command (Example 5-9).

Example 5-9 Using viostat to check disk performance

```
$ viostat

System configuration: lcpu=8 drives=12 ent=0.50 paths=39 vdisks=26
tty:      tin         tout    avg-cpu: % user % sys % idle % iowait physc % entc
          0.0          0.3                0.0   0.1   99.9      0.0   0.0    0.2

Disks:        % tm_act     Kbps      tps    Kb_read   Kb_wrtn
hdisk0           0.0      11.4       0.5    3951847   2740372
hdisk1           0.0      75.7       1.6   17345601  27235122
hdisk2           0.0       0.0       0.0        185         0
hdisk3           0.0       0.8       0.2      37764    447392
hdisk4           0.0       0.7       0.1      17824    386204
hdisk5           0.0       1.3       0.3     185004    584404
hdisk6           0.0       0.4       0.1      33948    186540
hdisk7           0.0       0.0       0.0         37         0
hdisk8           0.0       0.0       0.0         37        36
hdisk9           0.0       0.3       0.1      36309    126086
hdisk10          0.0       0.8       0.2      98251    346357
hdisk11          0.0       0.9       0.2     120018    385034
```

In cases where logical volumes are being used as paging devices, you can check how these devices are performing by using the **lvmstat** command (Example 5-10).

Example 5-10 Checking paging device performance

```
$ oem_setup_env
#  lvmstat -v amspaging
Logical Volume       iocnt   Kb_read   Kb_wrtn      Kbps
  amspaging03        64372    257488         0      0.98
```

Note: LVM statistics collection must be enabled for you to use **lvmstat**. Enable this feature with either of these commands:

```
lvmstat -e -v <vg_name>
lvmstat -e -l <lv_name>
```

5.2.3 Monitoring AIX

The following AIX commands have been enhanced to provide information for Active Memory Sharing:

- vmstat
- **lparstat**
- **topas**

The vmstat command

Hypervisor paging information is displayed when the **vmstat** command is issued with the **-h** option, as shown in Example 5-11.

Example 5-11 Displaying hypervisor paging information by using vmstat -h

```
# vmstat -h 10

System configuration: lcpu=8 mem=10240MB ent=1.00 mmode=shared mpsz=20.00GB

kthr    memory              page              faults              cpu                   hypv-page
----- ----------- ------------------------ ------------ ----------------------- ------------------------
 r  b   avm    fre     re  pi  po  fr  sr  cy  in  sy  cs  us sy id wa    pc  ec    hpi    hpit  pmem  loan
 3  0 1245552 115568   0   0   0   0   0   0   7  742 219 11 10 79  0  0.25  25.5   326      6  3.97  4.64
 4  0 1247125 111240   0   0   0   0   0   0   2  543 199 11  4 85  0  0.18  17.7   274      9  3.97  4.65
 4  0 1248872 108790   0   0   0   0   0   0   4  593 199 13  7 81  0  0.21  20.8   284      9  3.98  4.65
 4  0 1251021 101579   0   0   0   0   0   0   1  608 252 15  7 79  0  0.24  24.1   246     10  3.98  4.67
 2  0 1252582 93616    0   0   0   0   0   0   2  691 193 12  7 82  0  0.21  20.9   241      9  3.98  4.69
 2  0 1255414 86513    0   0   0   0   0   0   1  754 223 27  8 65  0  0.39  39.1   237      7  3.99  4.71
```

The following fields that are highlighted in Example 5-11 have been added for Active Memory Sharing:

mmode Shows `shared` if the partition is running in shared memory mode. This field is not present on dedicated memory partitions.

mpsz Size of the shared memory pool.

hpi Number of hypervisor page-ins for the partition. A hypervisor page-in occurs if a page is being referenced that is not available in real memory because it has been paged out by the hypervisor previously.

hpit Time that is spent in hypervisor paging for the partition, in milliseconds.

pmem Amount of physical memory that is backing the logical memory, in gigabytes.

loan Amount of the logical memory, in gigabytes, that is loaned to the hypervisor. The amount of loaned memory can be influenced through the **vmo ams_loan_policy** tunable. For more information, see 5.1.12, “Tuning” on page 91.

As highlighted in Example 5-12, the **vmstat -v -h** command shows the number of AMS memory faults and the time spent, in milliseconds, for hypervisor paging since boot time. It also shows the number of 4-KB pages that AIX has loaned to the hypervisor, and the percentage of partition logical memory that has been loaned.

Example 5-12 Displaying hypervisor paging information by using vmstat -v -h

```
# vmstat -v -h
              2621440 memory pages
              2407164 lruable pages
              1699043 free pages
                    1 memory pools
               365286 pinned pages
                 90.0 maxpin percentage
                  3.0 minperm percentage
                 90.0 maxperm percentage
                  2.3 numperm percentage
                56726 file pages
                  0.0 compressed percentage
                    0 compressed pages
                  2.3 numclient percentage
                 90.0 maxclient percentage
                56726 client pages
                    0 remote pageouts scheduled
                    0 pending disk I/Os blocked with no pbuf
                14159 paging space I/Os blocked with no psbuf
                 2228 filesystem I/Os blocked with no fsbuf
                    0 client filesystem I/Os blocked with no fsbuf
                    0 external pager filesystem I/Os blocked with no
fsbuf
               421058 Virtualized Partition Memory Page Faults
              3331187 Time resolving virtualized partition memory page
faults
               488154 Number of 4k page frames loaned
                   18 Percentage of partition memory loaned
                 33.0 percentage of memory used for computational pages
```

If the Virtualized Partition Memory Faults are increasing, it means that hypervisor paging has occurred at some time. This might be an indicator that the physical memory is overcommitted.

Impact of Active Memory Sharing on vmstat metrics

When you are using Active Memory Sharing, some of the metrics that are displayed by the **vmstat** command do not represent the same information displayed for a dedicated memory partition. A similar effect occurs in the percentage used processor metrics when shared processor partitions are used.

The **mem** field shows the amount of available logical memory. Unlike in a dedicated memory partition, where the logical memory is always backed by physical memory, this is not the case in a shared memory partition. The partition in Example 5-11 on page 110 has 10 GB of logical memory. This does not mean that there is actually this amount of logical memory available to the partition. To see how much physical memory the partition currently has assigned, look at the `pmem` column. In this case, the partition has only 3.99 GB of physical memory assigned.

The `fre` column shows the number of free logical pages. As Example 5-13 shows, this is not necessarily true in a shared memory partition, where the fre column shows the amount of free logical memory. Although this column shows 8.4 GB (2203204 4-KB pages) as free, the physical memory that is actually assigned is only 5.39 GB. This means that the hypervisor has stolen memory from the partition and provided it to another partition. The partition shown in Example 5-13 has the `ams_loan_policy` set to 0. Therefore, no memory has been loaned.

Example 5-13 Shared memory partition with free memory not backed by physical

```
# vmstat -h 2

System configuration: lcpu=8 mem=10240MB ent=1.00 mmode=shared mpsz=20.00GB

kthr    memory              page              faults              cpu                   hypv-page
----- ----------- ------------------------ ------------ ----------------------- -------------------------
 r  b   avm   fre  re  pi  po  fr   sr  cy  in   sy  cs us sy id wa    pc    ec   hpi  hpit   pmem   loan
 0  0 444694 2203204   0   0   0   0    0   0   7 3711 552  1  1 98  0  0.03   2.9     0     0   5.39   0.00
 0  0 444699 2203199   0   0   0   0    0   0   4 3187 539  0  1 98  0  0.03   2.5     0     0   5.39   0.00
 0  0 444701 2203197   0   0   0   0    0   0   3 3290 541  1  1 98  0  0.03   2.7     0     0   5.39   0.00
 0  0 444699 2203199   0   0   0   0    0   0   0  725 113  1  1 98  0  0.03   3.2     0     0   5.39   0.00
```

When loaning is enabled (`ams_loan_policy` is set to `1` or `2`), AIX loans pages when the hypervisor initiates a request. AIX removes free pages that are loaned to the hypervisor from the free list.

Example 5-14 shows a partition that has a logical memory size of 10 GB. It has also been assigned 9.94 GB of physical memory. Of this assigned 9.94 GB, 8.3 GB (2183045 4-KB pages) is free because there is no activity in the partition.

Example 5-14 Shared memory partition not loaning memory

```
# vmstat -h 2

System configuration: lcpu=8 mem=10240MB ent=1.00 mmode=shared mpsz=20.00GB

kthr    memory              page              faults              cpu                   hypv-page
----- ----------- ------------------------ ------------ ----------------------- -------------------------
 r  b    avm   fre  re  pi  po  fr   sr  cy  in   sy  cs us sy id wa    pc    ec   hpi  hpit   pmem   loan
 1  0 442321 2183045   0   0   0   0    0   0   4  652 119  1  1 98  0  0.03   3.4     0     0   9.94   0.00
 1  0 442320 2183046   0   0   0   0    0   0  12 3365 527  1  1 98  0  0.03   3.1     0     0   9.94   0.00
 1  1 442320 2182925   0  42   0   0    0   0  83 1810 392  1  2 97  0  0.04   4.1     0     0   9.94   0.00
```

Example 5-15 shows the same partition a few minutes later. In the intervening time, the hypervisor requested memory and the partition loaned 3.30 GB to the hypervisor. AIX has removed the free pages that it has loaned from the free list. The free list is therefore reduced by 875634 4-KB pages.

Example 5-15 Shared memory partition loaning memory

```
# vmstat -h 2

System configuration: lcpu=8 mem=10240MB ent=1.00 mmode=shared mpsz=20.00GB

kthr    memory              page              faults              cpu                   hypv-page
----- ----------- ------------------------ ------------ ----------------------- -------------------------
 r  b    avm   fre  re  pi  po  fr   sr  cy  in   sy  cs us sy id wa    pc    ec   hpi  hpit   pmem   loan
 0  0 442355 1307291   0   0   0   0    0   0  20 3416 574  1  2 98  0  0.04   3.7     0     0   6.70   3.30
 1  0 442351 1307295   0   0   0   0    0   0  17 3418 588  1  1 98  0  0.03   2.7     0     0   6.70   3.30
 2  0 442353 1307293   0   0   0   0    0   0  15 3405 558  1  1 98  0  0.03   2.7     0     0   6.70   3.30
```

From a performance perspective, it is important to monitor the number of hypervisor page-ins. If there are non-zero values, the partition is waiting for pages to be brought into real memory by the hypervisor. Because these pages must be read from a paging device, the page-ins have an impact on application performance.

Example 5-11 on page 110 shows a situation where pages for a shared memory partition are being paged in by the hypervisor. The `hpi` column shows activity.

AIX paging and hypervisor paging

When you use Active Memory Sharing, paging can occur on the AIX level or on the hypervisor level. When you see non-zero values in the `pi` or `po` column of vmstat, it means that AIX is running paging activities.

In a shared memory partition, AIX paging occurs not only when the working set exceeds the size of the logical memory, as in a dedicated partition. This can happen even if the LPAR has less physical memory than logical memory. AIX is dependent on the amount of logical memory available. So, if an LPAR is configured with 4 GB of logical memory but is running a workload that uses 4.5 GB of memory, the workload will page to the AIX paging spaces regardless of how much physical memory the hypervisor has allocated to the LPAR.

Another reason is that AIX is freeing memory pages to loan them to the hypervisor. If the loaned pages are used pages, AIX must save the content to its paging space before loaning them to the hypervisor. This behavior is especially frequent if you select an aggressive loaning policy (`ams_loan_policy=2`).

The lparstat command

The **`lparstat`** command has been enhanced to display statistics about shared memory. Most of the metrics show the I/O entitled memory statistics.

Note: The I/O memory entitlement does not have to be changed unless you encounter I/O performance problems. If so, check the `iomaf` column for failed I/O allocation requests for I/O memory entitlement.

When you use the **`lparstat -m`** command, the following attributes are displayed.

mpsz	Size of the memory pool in GB.
iome	I/O memory entitlement in MB.
iomp	Number of I/O memory entitlement pools. Details about each pool can be displayed by using the **`lparstat -me`** command.
hpi	Number of hypervisor page-ins.
hpit	Time spent in hypervisor paging in milliseconds.
pmem	Allocated physical memory in GB.
iomin	Minimum I/O memory entitlement for the pool.
iomu	Used I/O memory entitlement in MB.
iomf	Free I/O memory entitlement in MB.
iohwm	High water mark of I/O memory entitlement usage in MB.
iomaf	Number of failed allocation requests for I/O memory entitlement.

Example 5-16 shows the output of the **lparstat -m** command.

Example 5-16 The lparstat -m command

```
# lparstat -m 1

System configuration: lcpu=8 mem=10240MB mpsz=20.00GB iome=334.00MB
iomp=10 ent=1.00

physb   hpi  hpit  pmem  iomin   iomu   iomf  iohwm iomaf %entc  vcsw
----- ----- ----- ----- ------ ------ ------ ------ ----- ----- -----
 1.82     0     0  8.95  287.7   12.4   34.3   18.0     0   3.1   496
 1.62     0     0  8.95  287.7   12.4   34.3   18.0     0   2.6   596
 1.64     0     0  8.95  287.7   12.4   34.3   18.0     0   2.6   592
 1.60     0     0  8.95  287.7   12.4   34.3   18.0     0   2.5   601
 1.64     0     0  8.95  287.7   12.4   34.3   18.0     0   2.5   598
```

When you use the **lparstat -me** command, the I/O memory entitlement details for the partition are displayed as shown in Example 5-17.

Example 5-17 The lparstat -me command

```
# lparstat -me

System configuration: lcpu=8 mem=10240MB mpsz=20.00GB iome=334.00MB iomp=10 ent=1.00

physb   hpi  hpit  pmem  iomin   iomu   iomf  iohwm iomaf %entc  vcsw
----- ----- ----- ----- ------ ------ ------ ------ ----- ----- -----
 0.11 1055380 10218330  8.95  287.7   12.4   34.3   18.0     0   0.2 67360383

              iompn: iomin  iodes   iomu  iores  iohwm  iomaf
         ent0.txpool  2.12  16.00   2.00   2.12   2.00      0
      ent0.rxpool__4  4.00  16.00   3.50   4.00   3.50      0
      ent0.rxpool__3  4.00  16.00   2.00  16.00   2.70      0
      ent0.rxpool__2  2.50   5.00   2.00   2.50   2.00      0
      ent0.rxpool__1  0.84   2.25   0.75   0.84   0.75      0
      ent0.rxpool__0  1.59   4.25   1.50   1.59   1.50      0
        ent0.phypmem  0.10   0.10   0.09   0.10   0.09      0
                fcs1 136.25  136.25   0.30  136.25   2.80      0
                fcs0 136.25  136.25   0.30  136.25   2.69      0
                sys0  0.00   0.00   0.00   0.00   0.00      0
```

iodes Wanted entitlement of the I/O memory pool in MB

iores Reserved entitlement of the I/O memory pool in MB

The topas command

When you use the **topas -L** command, the logical partition view with the Active Memory Sharing statistics highlighted in bold in Example 5-18 is displayed. The IOME field shows the I/O memory entitlement that is configured for the partition, whereas the iomu column shows the I/O memory entitlement in use. For a description of the other fields, see “The vmstat command” on page 110.

Example 5-18 The topas -L command

```
Interval:2        Logical Partition: AIX1C             Fri Apr 29 11:08:02 2011
Psize:     16.0            Shared SMT       4        Online Memory:     10.00G
                        Power Saving: Disabled
Ent: 1.00                       Mode: Un-Capped       Online Logical CPUs:  8
Mmode: Shared                   IOME: 334.00          Online Virtual CPUs:  2
Partition CPU Utilization
 %usr  %sys %wait %idle  physc %entc    app vcsw phint  hpi   hpit  pmem  iomu
  0.5   1.1   0.0  98.4   0.03  2.55  13.92  394     0     0      0   8.95 12.45
================================================================================
LCPU MINPF MAJPF INTR  CSW ICSW RUNQ  LPA SCALLS USER KERN WAIT IDLE  PHYSC LCSW
1       411     0  118  121 56.0    0  101  1.52K 27.6 60.7  0.0 11.7   0.01  137
4         0     0 23.0    0    0    0    0      0  0.0 53.5  0.0 46.5   0.00 19.0
0       317     0  256  150 84.0    0  100 115.00 25.5 46.3  0.0 28.2   0.01  215
3         0     0 9.00    0    0    0    0      0  0.0  0.6  0.0 99.4   0.00 9.00
2         0     0 9.00    0    0    0    0      0  0.0  0.6  0.0 99.4   0.00 9.00
```

Press the E key to view details about the I/O memory entitlement, as shown in Example 5-19.

Example 5-19 Displaying I/O memory entitlement by using topas

```
Interval:2        Logical Partition: AIX1C             Fri Apr 29 11:10:28 2011
Psize:     16.0            Shared SMT       4        Online Memory:     10.00G
                        Power Saving: Disabled
Ent: 1.00                       Mode: Un-Capped       Online Logical CPUs:  8
Mmode: Shared                   IOME: 334.00          Online Virtual CPUs:   2
Partition CPU Utilization
physb %entc vcsw hpi    hpit   pmem  iomu  iomf  iohwm iomaf
  0.0   3.0 364  15     65     8.95 12.45 34.34  18.03  0.00
================================================================================
iompn          iomin  iodes  iomu   iores  iohwm  iomaf
fcs1           136.2  136.2   0.3   136.2    2.8    0.0
fcs0           136.2  136.2   0.3   136.2    2.7    0.0
ent0.rxpool__3   4.0   16.0   2.0    16.0    2.7    0.0
ent0.rxpool__4   4.0   16.0   3.5     4.0    3.5    0.0
ent0.rxpool__2   2.5    5.0   2.0     2.5    2.0    0.0
ent0.txpool      2.1   16.0   2.0     2.1    2.0    0.0
```

```
ent0.rxpool__0   1.6    4.2   1.5     1.6    1.5    0.0
ent0.rxpool__1   0.8    2.2   0.8     0.8    0.8    0.0
ent0.phypmem     0.1    0.1   0.1     0.1    0.1    0.0
sys0             0.0    0.0   0.0     0.0    0.0    0.0
```

Monitoring multiple partitions with topas

The **`topas -C`** command displays the CEC view, which is especially useful for monitoring multiple partitions on a managed system. The Active Memory Sharing statistics that are highlighted in bold in Example 5-20 on page 118 are displayed. The following values are shown:

pmem Amount of physical memory that is assigned to each individual partition.

Mode The mode in which the partition is running. The mode is displayed in a set of three characters that are explained in Table 5-1.

Table 5-1 Partition mode

The first character indicates the processor mode in the partition	The second character indicates the memory mode of the partition	The third character indicates the energy state of the partition
► For shared partitions: – **C** SMT enabled and capped – **c** SMT disabled and capped – **U** SMT enabled and uncapped – **u** SMT disabled and uncapped ► For dedicated partitions: – **S** SMT enabled and not donating – **d** SMT disabled and donating – **D** SMT enabled and donating – **-** SMT disabled and not donating	► **M** In shared memory mode (for shared partitions) and AME disabled ► **-** Not in shared memory mode and AME disabled ► **E** In shared memory mode and AME enabled ► **e** Not in shared memory mode and AME enabled	► **S** Static power save mode is enabled ► **d** Power save mode is disabled ► **D** Dynamic power save mode is enabled ► **-** Unknown / Undefined

Note: Monitoring multiple partitions with **`topas`** is not supported for Linux partitions.

The Active Memory Sharing statistics are shown in Example 5-20.

Example 5-20 The topas -C command

```
Topas CEC Monitor             Interval:  10             Fri Apr 29 14:59:04 2011
Partitions Memory (GB)           Processors
Shr:  5    Mon:38.0  InUse:13.0  Shr:  4  PSz: 16   Don: 0.0 Shr_PhysB  0.11
Ded:  0    Avl:   -              Ded:  0  APP: 31.7 Stl: 0.0 Ded_PhysB  0.00

Host       OS  Mod Mem InU Lp  Us Sy Wa Id  PhysB  Vcsw Ent  %EntC PhI  pmem
-------------------------------------shared-------------------------------------
VIOS1A     A61 Ued 4.0 1.1  8   1  1  0 96   0.03    0  0.50   5.8   0     -
AIX1A      A71 UEd  10 4.7  8   0  1  0 98   0.03  966  1.00   2.6   0  1.89
AIX1C      A71 UEd  10 1.8  8   0  1  0 98   0.03    0  1.00   2.6   0  4.69
VIOS1B     A61 Ued 4.0 1.1  8   0  1  0 98   0.02    0  0.50   3.6   0     -
AIX1B      A71 UEd  10 4.8  8   0  0  0 99   0.02    0  1.00   1.6   0  3.55

Host       OS  Mod Mem InU Lp  Us Sy Wa Id  PhysB  Vcsw  %istl %bstl
-----------------------------------dedicated----------------------------------
```

Press the M key in the CEC view to display the attributes of the shared memory pool (Example 5-21).

Example 5-21 Displaying shared memory pool attributes by using topas

```
Topas CEC Monitor             Interval:  10             Fri Apr 29 15:00:24 2011
Partitions Memory (GB)           Memory Pool(GB)          I/O Memory(GB)
Mshr:  3   Mon: 38.0 InUse: 13.5 MPSz: 20.0 MPUse: 10.1  Entl: 1002.0se: 37.3
Mded:  2   Avl: 24.5             Pools: 1
mpid   mpsz  mpus   mem  memu  iome  iomu     hpi    hpit
------------------------------------------------------------------------------
0     20.00 10.13 30.00 11.41 1002.0  37.3       0      0
```

5.2.4 Monitoring IBM i

Information about the Active Memory Sharing configuration and paging activity that is provided by the POWER Hypervisor is available above the Machine Interface layer to IBM i. The IBM i Collection Services have been enhanced and new data structures have been added to obtain that information.

The fields that are identified in Table 5-2 are useful for evaluating partition performance behaviors because of real memory paging activity by the hypervisor.

Table 5-2 QAPMSHRMP field details

Field name	Description
INTNUM	Interval number. The nth sample database interval is based on the start time that is specified in the Create Performance Data (CRTPFRDTA) command.
DATETIME	Interval date and time. The date and time of the sample interval.
INTSEC	Elapsed interval seconds. The number of seconds since the last sample interval.
SMPOOLID	Shared memory pool identifier. The identifier of the shared memory pool that this partition is using.
SMWEIGHT	Memory weight. Indicates the variable memory capacity weight that is assigned to the partition. Valid values are hex 0 - 255. The larger the value, the less likely this partition is to lose memory.
SMREALUSE	Physical real memory used. The amount of shared physical real memory, in bytes, that was being used by partition memory at sample time.
SMACCDLY	Real memory access delays. The number of partition processor waits that have occurred because of page faults on logical real memory.
SMACCWAIT	Real memory access wait time. The amount of time, in milliseconds, that partition processors have waited for real memory page faults to be satisfied.
SMOVRCAP	Partition memory processor usage capacity. The maximum amount of memory, in bytes, that the partition is allowed to assign to data areas shared between the partition operating system and the firmware.
SMENTIOC	Entitled memory capacity for I/O. The amount of memory, in bytes, currently assigned to the partition for use by I/O requests.
SMMINIOC	Minimum entitled memory capacity for I/O. The minimum amount of entitled memory, in bytes, needed to function with the current I/O configuration.
SMOPTIOC	Optimal entitled memory capacity for I/O. The amount of entitled memory, in bytes, that allows the current I/O configuration to function without any I/O memory mapping delays.

Field name	Description
SMIOCUSE	Current I/O memory capacity in use. The amount of I/O memory, in bytes, currently mapped by I/O requests.
SMIOCMAX	Maximum I/O memory capacity used. The maximum amount of I/O memory, in bytes, that has been mapped by I/O requests since the partition was last IPLed or the value was reset by an explicit request.
SMIOMDLY	I/O memory mapping delays. The cumulative number of delays that have occurred because insufficient entitled memory was available to map an I/O request since the partition was last IPLed.
MPACCDLY	Pool real memory access delays. The number of virtual partition memory page faults within the shared memory pool for all partitions.
MPACCWAIT	Pool real memory access wait time. The amount of time, in milliseconds, that all partitions' processors have spent waiting for page faults to be satisfied within the shared memory pool.
MPPHYMEM	Pool physical memory. The total amount of physical memory, in bytes, assigned to the shared memory pool.
MPLOGMEM	Pool logical memory. The summation, in bytes, of the logical real memory of all active partitions that are served by the shared memory pool.
MPENTIOC	Pool entitled I/O memory. The summation, in bytes, of the I/O entitlement of all active partitions that are served by the shared memory pool.
MPIOCUSE	Pool entitled I/O memory in use. The summation, in bytes, of I/O memory that is mapped by I/O requests from all active partitions that are served by the shared memory pool.

Note: Active Memory Sharing data is only populated in partitions that are part of the shared memory pool.

The expected consumer of this information for the initial release of the Active Memory Sharing support is the collection services, which are intended to log as much data as possible in existing DB fields. Other potential consumers of the information gathered are programs that are written specifically to retrieve this information by using service/performance APIs.

Checking the QAPMSHRMP table using SQL

Example 5-22 shows how to obtain Active Memory Sharing data by using the Start Structured Query Language (STRSQL) command.

Example 5-22 Sample query to gather QAPMSHRMP data

```
SELECT * FROM QPFRDATA/QAPMSHRMP
```

Figure 5-9 shows the command's output.

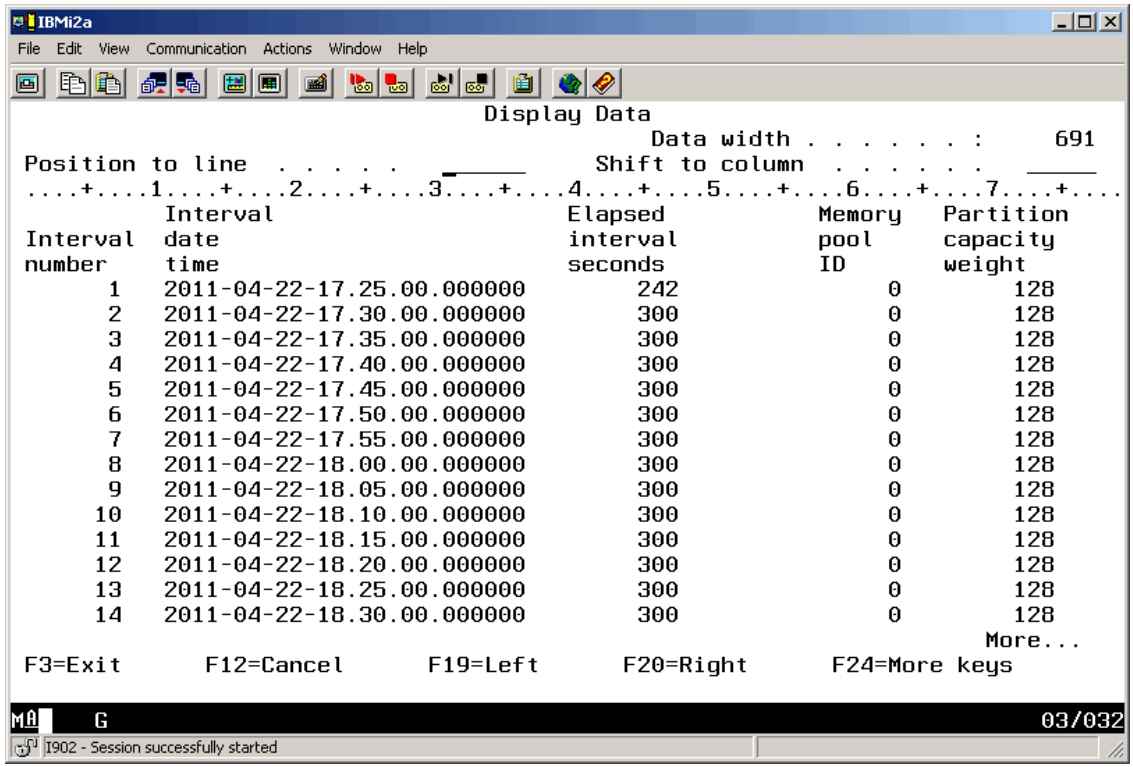


Figure 5-9 Sample query output

Checking the QAPMSHRMP table using IBM Systems Director Navigator for i

This section describes the method to obtain Active Memory Sharing data in graphic format by using the IBM System Director Navigator for i. For more information about setup and requirements, see *IBM Systems Director Navigator for i*, SG24-7789.

Use the following steps to display performance data using the IBM System Director Navigator for i:

1. Start your browser and go to `http://<systemname>:2001`.
2. Click **Performance** → **Investigate Data** → **Collection Services** → **Collection Services Database Files** → **select QAPMSHRMP**. Select the Collection Library and Collection Name and click **Display,** as shown in Figure 5-10.

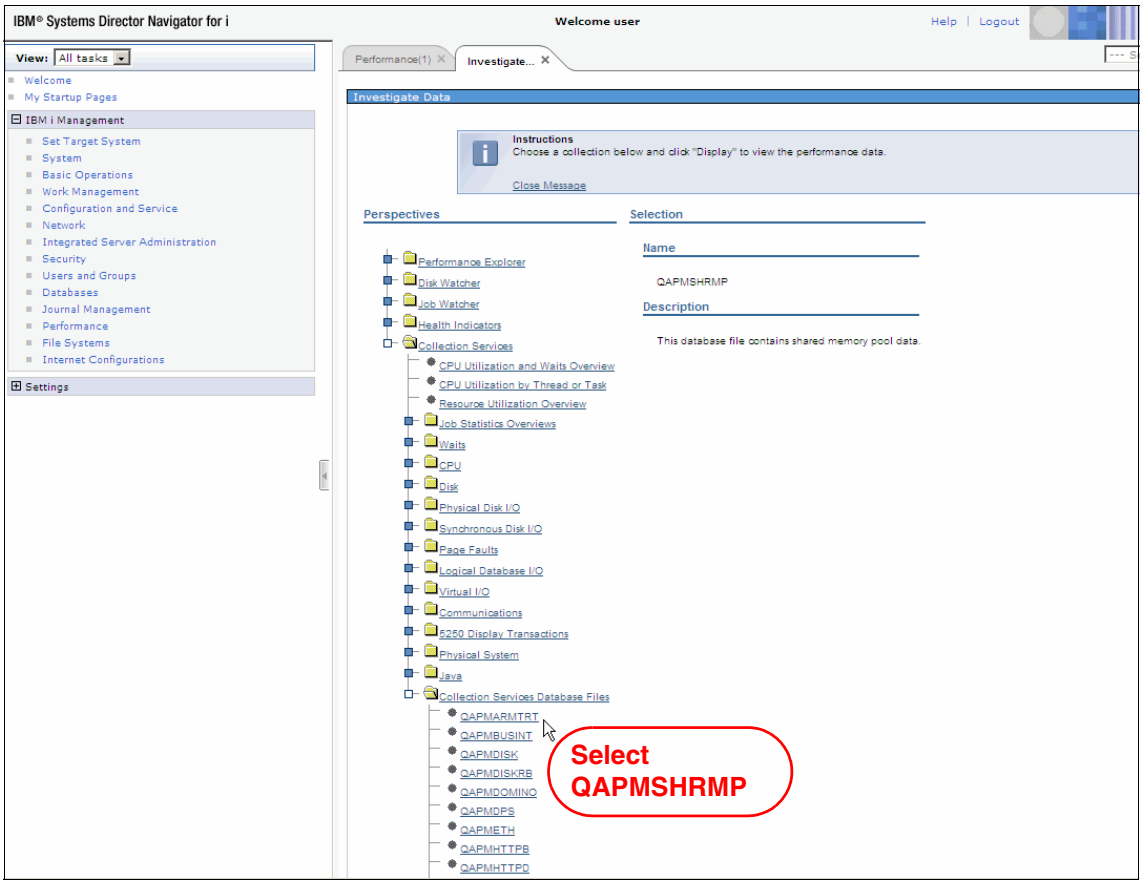


Figure 5-10 Systems Director Navigator for i performance collection data

3. As shown in Figure 5-11, you can select some or all collection intervals, and then click **Done**.

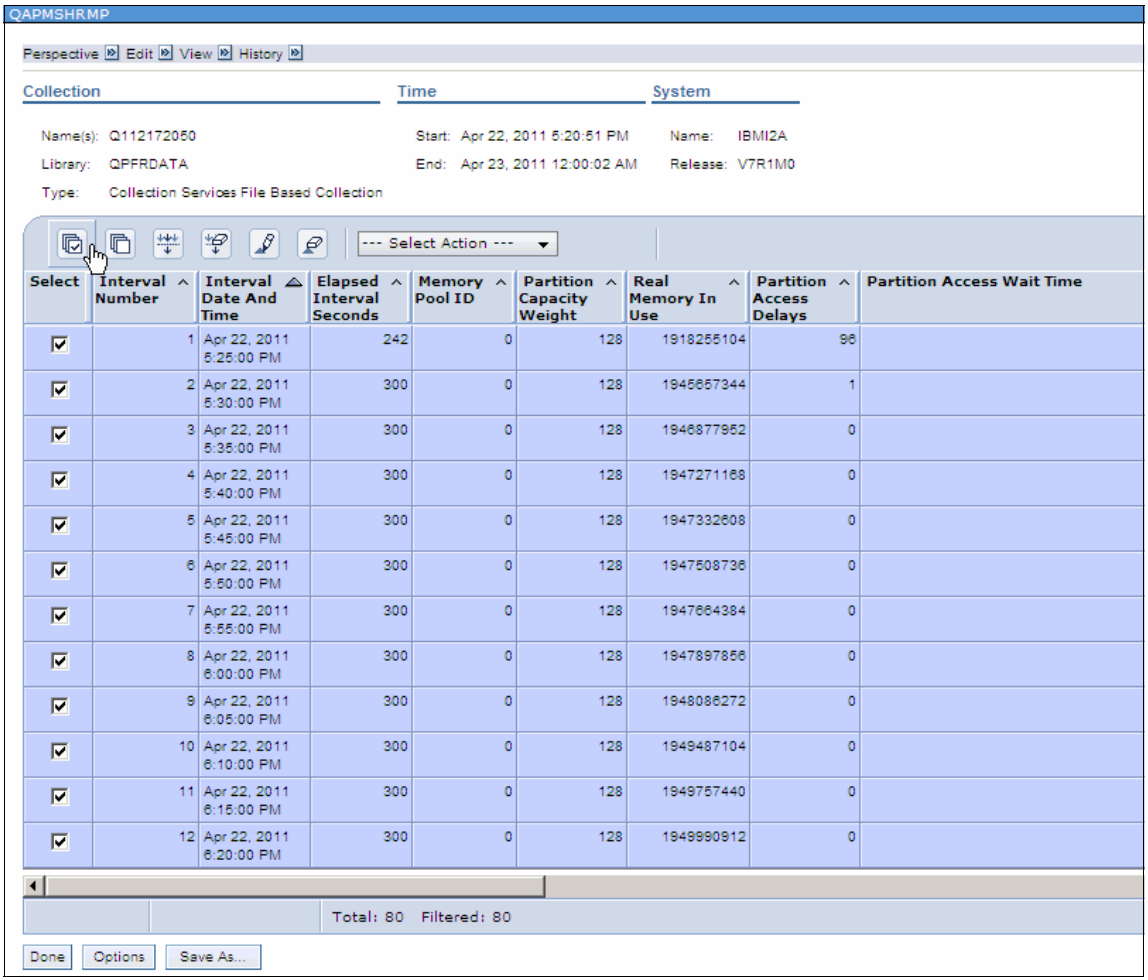


Figure 5-11 Collection intervals

4. After you select intervals, click **Select Action** and select **Show as a chart**, as shown in Figure 5-12.

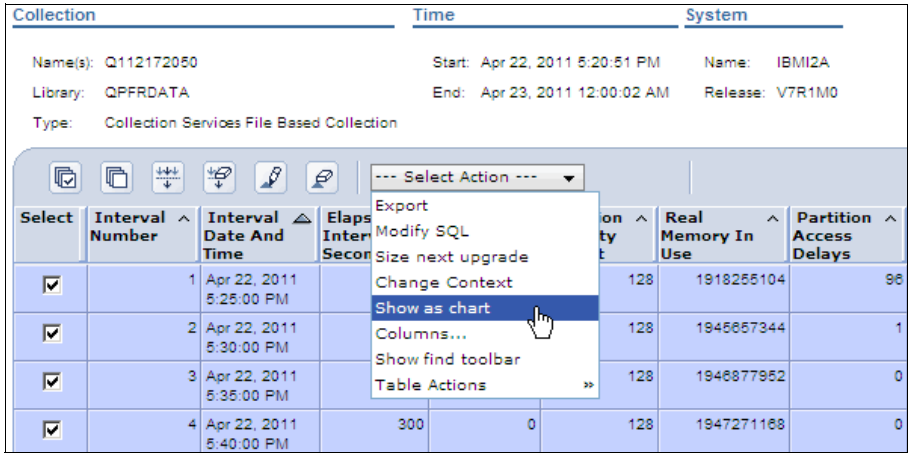


Figure 5-12 Selecting Show as chart to produce the graph

5. As shown on Figure 5-13, you can specify the category that you want to present on the graph. Select **Real Memory in Use** to display memory usage, then click **Add** and **OK**.

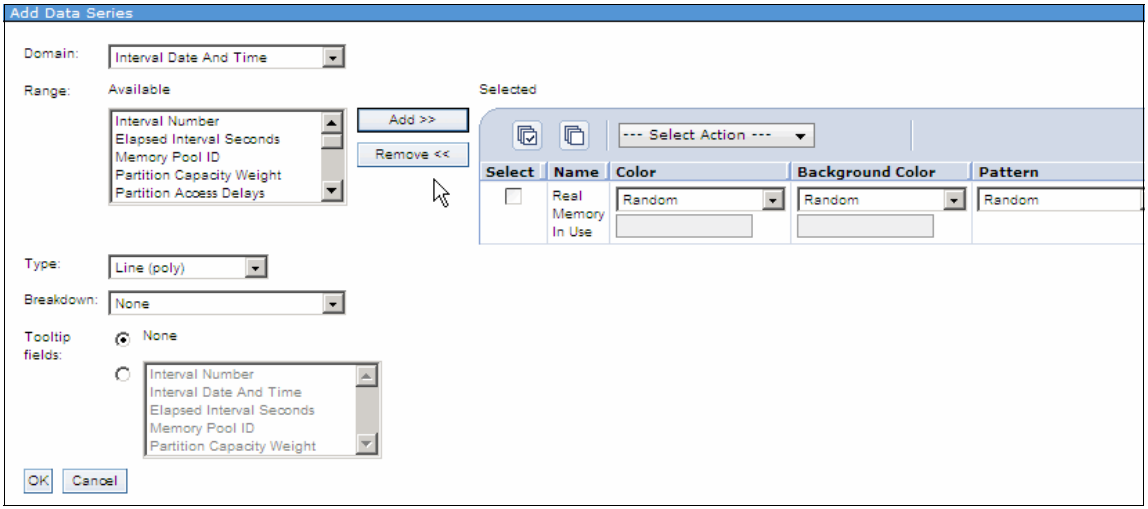


Figure 5-13 Selecting the category

The result is a graph that shows real memory usage in the specified intervals (Figure 5-14).

Figure 5-14 Real memory in use

You can navigate through values as you usually do in the interactive query environment. You can also run this query by using IBM System i® Navigator or any query tool that uses a database protocol that connects to the IBM i database.

5.2.5 Monitoring Linux

Performance information about Linux is available on /proc and /sys file systems. However, the most effective way to monitor AMS performance is by using the **amsstat** tool available from the powerpc-utils project at sourceforge: http://sourceforge.net/projects/powerpc-utils. This tool consolidates all AMS performance data, as shown in Example 5-23.

Example 5-23 Using amsstat to monitor AMS performance

```
LINUX2A:~ #  amsstat
Tue Apr 26 16:32:25 EDT 2011
System Memory Statistics:
     MemTotal:                      2901568 kB
     MemFree:                       2676096 kB
     Buffers:                       22464 kB
     Cached:                        61952 kB
     Inactive:                      91392 kB
     Inactive(anon):                33920 kB
     Inactive(file):                57472 kB
     SwapTotal:                     4194176 kB
     SwapFree:                      4145728 kB
     DesMem:                        8192 MB
Entitlement Information:
     entitled_memory:               368050176
     mapped_entitled_memory:        4739072
     entitled_memory_group_number:  32774
     entitled_memory_pool_number:   0
     entitled_memory_weight:        0
     entitled_memory_pool_size:     21474836480
     entitled_memory_loan_request:  997240832
     backing_memory:                2294312960
     cmo_enabled:                   1
     cmo_faults:                    370096
     cmo_fault_time_usec:           489855309
     cmo_primary_psp:               1
     cmo_secondary_psp:             2
CMM Statistics:
     disable:                       0
     debug:                         0
     min_mem_mb:                    256
     oom_kb:                        1024
     delay:                         1
     loaned_kb:                     5174272
     loaned_target_kb:              5174272
```

```
     oom_freed_kb:                 973824
VIO Bus Statistics:
     cmo_entitled:                 368050176
     cmo_reserve_size:             24832000
     cmo_excess_size:              343218176
     cmo_excess_free:              343218176
     cmo_spare:                    1562624
     cmo_min:                      7813120
     cmo_desired:                  24832000
     cmo_curr:                     5361664
     cmo_high:                     20631552
VIO Device Statistics:
   l-lan@30000002:
     cmo_desired:                  4243456
     cmo_entitled:                 4243456
     cmo_allocated:                4239360
     cmo_allocs_failed:            0
   vfc-client@30000003:
     cmo_desired:                  8450048
     cmo_entitled:                 8450048
     cmo_allocated:                393216
     cmo_allocs_failed:            0
   vfc-client@30000004:
     cmo_desired:                  8450048
     cmo_entitled:                 8450048
     cmo_allocated:                589824
     cmo_allocs_failed:            0
   v-scsi@30000005:
     cmo_desired:                  2125824
     cmo_entitled:                 2125824
     cmo_allocated:                139264
     cmo_allocs_failed:            0
```

Table 5-3 describes the most relevant fields from the **amsstat** output. Consult the man pages for a complete list.

Table 5-3 Definitions of amsstat command output fields

Field name	Description
System memory statistics	
Memtotal	Maximum memory available to the system.
Memfree	Unused memory

Field name	Description
Inactive	Free memory from buffer or cache that is available to be used because it was not recently used and can be reclaimed for other purposes.
Swaptotal	Available swap memory.
Swapfree	Free swap memory.
Desmem	Desired memory from LPAR profile.
Entitlement information	
Entitled_memory	Memory that is given to the operating system and available to be mapped for IO operations.
Mapped_entitled_memory	Memory that is currently mapped for IO operations.
Entitled_memory_weight	The weighting that is used by firmware to help prioritize partitions for memory loaning.
Entitled_memory_pool_size	Total amount size of memory in the memory pool that the LPAR belongs to.
Entitled_memory_loan_request	Amount of memory that firmware wants to give using the CMM. A positive value means the amount of memory that is expected from partition to loan. A negative value means the operating system can take back that amount for its own use.
Backing_memory	Physical memory that is currently reserved for access by the partition.
Cmo_enabled	If set to 1, indicates partition is running in shared mode. A zero value indicates the partition is running in dedicated mode.
Cmo_faults	The number of page faults since the operating system was booted. Increases in this value indicate memory contention between partitions that are running in the shared pool.
Cmo_fault_time_usec	The amount of time since the boot time, in microseconds, that the operating system has been suspended by the platform firmware to process cmo_faults.

Field name	Description
CMM statistics	
Disable	Indicates whether CMM mode is disabled. If it is set to 1, it is disabled. A value of 0 means it is enabled. If CMM is disabled, no loaning will occur.
Oom_kb	The number of kilobytes of memory taken back from firmware by the CMM module for the operating system when an out of memory signal from the kernel is caught by CMM.
Delay	The number of seconds that CMM waits between requests to firmware for the number of pages that firmware requests the operating system to loan.
Loaned_kb	The amount of memory, in kilobytes, that the operating system has given back to the platform firmware to be used by other partitions. This value fluctuates to meet the demands of all of the partitions in the shared memory pool of an AMS environment.
Loaned_target_kb	The amount of memory, in kilobytes, that the firmware requests the operating system to loan for use by other partitions. This value can be greater than `loaned_kb` if firmware would like extra pages to be loaned. It can be less than `loaned_kb` if the firmware is providing extra pages to the operating system.
Oom_freed_kb	The amount of memory, in kilobytes, that is no longer being loaned by CMM as a result of out-or-memory kernel signals.
Vio bus statistics	
Cmo_desired	The amount of memory, in kilobytes, that the device has requested from the bus to provide optimal performance.
Cmo_entitled	The amount of memory, in kilobytes, that the device is ensured that it can map for IO operations.
Cmo_allocated	The amount of memory, in kilobytes, that the device has currently mapped for IO operations.

Field name	Description
Cmo_allocs_failed	When the amount of memory allocated (`cmo_allocated`) has exhausted both the entitled memory (`cmo_entitled`) and the bus excess pool, memory mapping failures occur. This field is a cumulative counter. Large changes in this value indicate resource contention that might require system tuning.

Note: On Linux systems that run in shared memory mode, commands like **`top`**, **`vmstat`**, and **`free`** can show total memory values that change over time. This is an expected behavior that is caused by the hypervisor assigning memory to the partition based on the current system workload.

6

Active Memory Deduplication

Active Memory Deduplication is a feature of the PowerVM Active Memory Sharing technology (AMS). It detects and removes duplicated memory pages to optimize memory usage in partitions that use a shared memory pool.

This chapter describes how to manage and monitor the Active Memory Deduplication on the shared memory partitions. For more information about how to enable memory deduplication in an Active Memory Sharing setup, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

This chapter includes the following sections:

- Managing Active Memory Deduplication
- Monitoring Active Memory Deduplication

6.1 Managing Active Memory Deduplication

When Active Memory Deduplication is enabled in a system, all partitions that use the shared memory pool have their memory pages inspected for deduplication. The tunables are set based on the size of the AMS pool. This section presents the parameters that can be tuned within Active Memory Deduplication.

6.1.1 Tunable parameters

The *deduplication table ratio* is the single parameter that can be tuned for Active Memory Deduplication. It defines the size of the deduplication table.

Active Memory Deduplication uses the deduplication table to store information about which pages are deduplicated and potential deduplication candidate pages. When the deduplication table is enlarged, Active Memory Deduplication has more memory space to store the information it needs to maintain deduplicated pages. Also, the potential to deduplicate memory pages within the shared memory pool increases.

However, increasing the size of the deduplication table requires more memory to be used by the hypervisor. For example, on a shared memory pool with the maximum size set to 1 TB, if the deduplication table ratio is set to 1:256, 4 GB of extra memory is set aside by the hypervisor for the deduplication table. It is not, however, a large amount when compared to the size of the memory pool.

The *deduplication table size* is calculated as the product of the maximum AMS pool size and the deduplication table ratio. Both parameters can be retrieved from the Hardware Management Console (HMC).

There are other factors that have influence on, but are not explicit tuning parameters for, Active Memory Deduplication. Maximum memory pool size can affect the deduplication table size because its size is calculated based on the maximum memory pool size. Changing this value changes the deduplication table. However, if you want to change the deduplication table size, change the deduplication table ratio instead of changing the maximum memory pool size.

To create the memory page signatures for deduplication, the hypervisor needs processing resources. It uses processor cycles that are taken from the Virtual I/O Server to build and compare the page signatures, and the memory pages. The amount of processing resources that are required by the hypervisor is small. Generally, configure the Virtual I/O Server as an uncapped partition with the partition weight at 255 (the highest value). Also, round up the processing capacity that was defined for the Virtual I/O Server.

6.1.2 Tuning the ratio of the deduplication table

The current system configuration for the deduplication table ratio can be obtained from the HMC command-line interface, as shown in Example 6-1.

Example 6-1 Listing the value of the deduplication table ratio using the HMC

```
hscroot@hmc8:~> lshwres -r mempool -m p740
curr_pool_mem=24576,curr_avail_pool_mem=23969,curr_max_pool_mem=40960,p
end_pool_mem=24576,pend_avail_pool_mem=23969,pend_max_pool_mem=40960,sy
s_firmware_pool_mem=282,paging_vios_names=p740_vios04,paging_vios_ids=2
,mem_dedup=1,dedup_table_ratio=1:1024
```

The command **lshwres** with these parameters retrieves the memory pool configuration information, and the **dedup_table_ratio** parameter shows the deduplication table ratio. The **curr_max_pool_mem** parameter shows the AMS maximum pool size.

In this example, **dedup_table_ratio** is 1/1024 and **curr_max_pool_mem** is 40960 (value in MB). With these values, the following is the size of the deduplication table in this example:

Deduplication table size = 40960 MB * 1/1024 = 40 MB

This amount of memory is reserved by the hypervisor for Active Memory Deduplication use only, and is taken from the shared memory pool. Users planning to use Active Memory Deduplication must take this value into consideration when they implement this technology into their systems.

You can check the supported deduplication table ratios for your system by using the HMC as shown in Example 6-2. The default value is 1:1024.

Example 6-2 Listing the possible values for the deduplication table ratio parameter

```
hscroot@hmc8:~> lshwres -r mem -m p740 --level sys -F
possible_dedup_table_ratios
"1:256,1:512,1:1024,1:2048,1:4096,1:8192"
```

To change the deduplication table, use the **chhwres** command on the HMC command-line interface as shown in Example 6-3. Changing the value of the deduplication table ratio is a dynamic operation and effective immediately after the command is run.

Example 6-3 Changing the value of the deduplication table ratio

```
hscroot@hmc6:~> chhwres -r mempool -m p740 -o s -a
"dedup_table_ratio=1:256"
```

Whenever a deduplication table ratio value changes, all of the deduplicated memory pages are broken out into separate unique physical pages, and another deduplication table is created. The process of deduplicating memory pages is then restarted. This process also happens when the maximum size of the memory pool changes.

6.2 Monitoring Active Memory Deduplication

This section explains how to monitor memory resources in a virtualized IBM Power server with Active Memory Deduplication enabled.

6.2.1 Statistics

Memory page coalescing is a transparent operation in which the hypervisor detects duplicate pages and directs all the user read pages to a single copy. The hypervisor then reclaims the other duplicate physical memory pages.

Memory page coalescing efficiency is measured in physical page savings as opposed to coalescing processor requirements. The overall page savings depend on the workloads that are running on the partitions. The more similar they are, the greater the savings.

You can inspect the results of memory deduplication in your system by using the HMC or operating system commands. Some of the relevant parameters that you can inspect are described as follows.

Pool coalesced memory

Pool coalesced memory refers to the total amount of physical memory that is coalesced among all of the partitions during the Active Memory Deduplication activity. This global metric is available to all partitions that are authorized to collect performance statistics. This value represents the current snapshot for pool coalesced pages.

The pool coalesced memory counter increases for each deduplicated page that is stored in the pool, no matter how many virtual pages among all partitions reference it. In other words, having more or less partitions that share a page does not change the counter.

Logical partition coalesced memory

Logical partition coalesced memory indicates how much memory is being deduplicated for a partition. It shows how much memory is being made available by the deduplication process that otherwise would be used by the partition storing pages that are identical to other pages in the system.

Figure 6-1 on page 136 shows an example of pool coalesced count and LPAR coalesced count. Three partitions and three memory pages are duplicated in the system, which are shown as A, B, and C. These pages get deduplicated and stored only once in the shared pool memory.

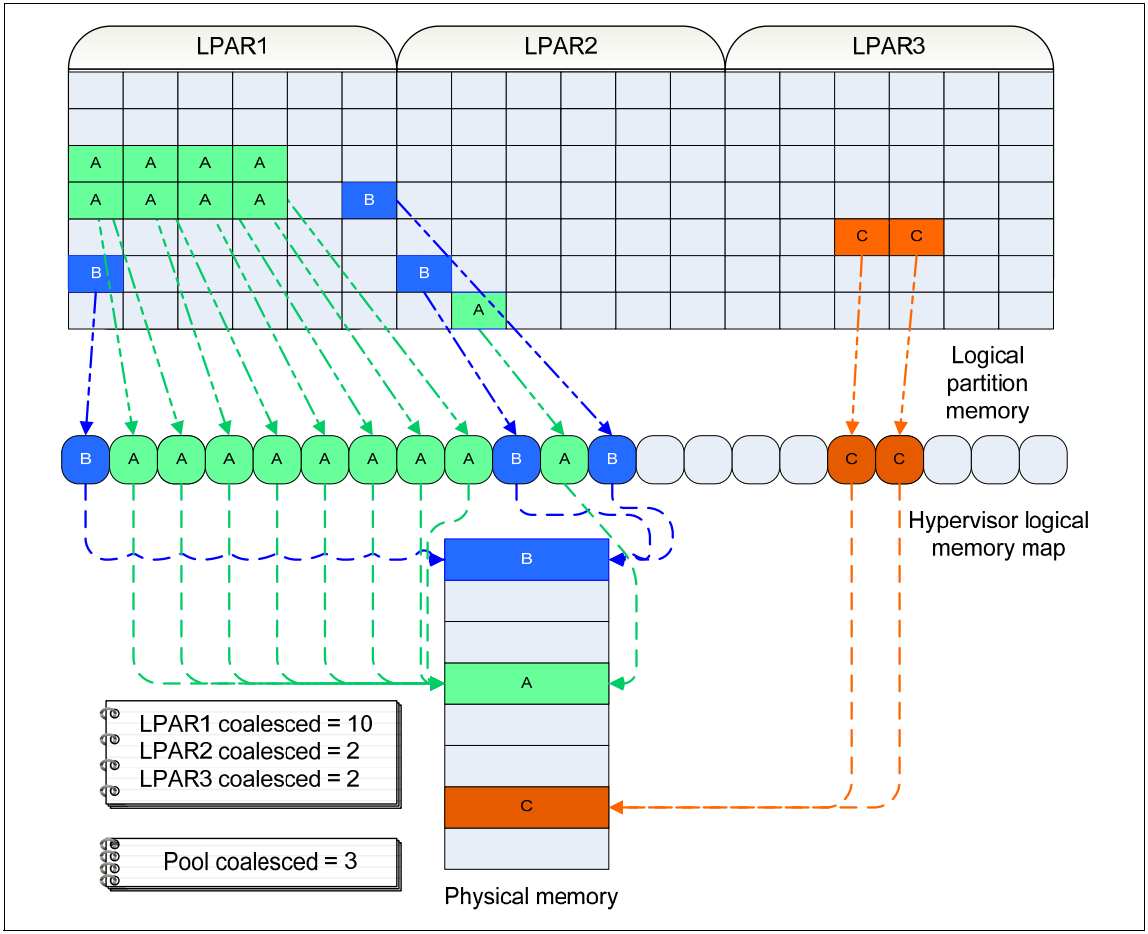


Figure 6-1 Memory coalescing counters

The pool coalesced memory counter increases by three pages, whereas the LPAR1 coalesced memory counter increases by 10 pages (eight A pages plus two B pages). LPAR2 coalesced memory counter increases by two pages (one B page plus one A page). Finally, the coalesced memory counter for LPAR3 increases by two pages (two C pages).

The amount of memory that is saved by the deduplication process is the sum of the logical partition coalesced memory for all the partitions using shared memory, minus the pool coalesced memory. You can retrieve the coalesced memory information by using commands in AIX, Linux, or by using the HMC. For more information, see 6.2.2, “Monitoring tools” on page 137.

6.2.2 Monitoring tools

This section shows the tools that are available in the platform and in the operating systems to monitor Active Memory Deduplication.

Monitoring in AIX

Monitoring coalesced memory in an AIX partition is simple. The **lparstat** command has been enhanced to display the number of coalesced pages and the Virtual I/O Server processor cycles used for page coalescing. These statistics are shown in Example 6-4.

Example 6-4 Checking the deduplication information using the lparstat command

```
# lparstat -mpw 1

System configuration: lcpu=2 mem=10240MB mpsz=24.00GB iome=77.00MB iomp=9 ent=2.00

physb   hpi  hpit  pmem  iomin   iomu   iomf  iohwm iomaf  pgcol mpgcol ccol %entc  vcsw
----- ----- ----- ----- ------ ------ ------ ------ -----  ------ ------ ---- ----- -----
 0.64     0     0  4.74   23.7   12.0   53.3   13.5     0 1094.2 379.2   0.0   0.4   153
 0.23     0     0  4.74   23.7   12.0   53.3   13.5     0 1094.2 379.2   0.0   0.2   156
 2.27     0     0  4.74   23.7   12.0   53.3   13.5     0 1094.2 379.2   0.0   1.2   149
 1.54     0     0  4.75   23.7   12.0   53.3   13.5     0 1094.2 379.2   0.0   0.8   150
 0.25     0     0  4.75   23.7   12.0   53.3   13.5     0 1094.2 379.2   0.0   0.2   157
 0.24     0     0  4.75   23.7   12.0   53.3   13.5     0 1094.2 379.2   0.0   0.2   154
 0.35     0     0  4.75   23.7   12.0   53.3   13.5     0 1094.2 379.2   0.0   0.3   232
```

In this report, you see these Active Memory Deduplication statistics using the **lparstat** command.

pmem Indicates the physical memory, in GB, that is allocated to the LPAR by the hypervisor.

pgcol Indicates the amount of LPAR memory, in MB, that is coalesced because of Active Memory Deduplication activity.

mpgcol Indicates the amount of coalesced memory, in MB, in the entire shared memory pool because of Active Memory Deduplication activity. If the partition is not authorized to access pool-wide statistics, this metric displays zero.

ccol Indicates the amount of processor power, measured in number of processing units, that is used for coalescing pages during Active Memory Deduplication activity. If the partition is not authorized to access pool-wide statistics, this metric displays zero.

Monitoring in IBM i

At the time of writing, monitoring the amount of memory that is coalesced on IBM i partitions is not possible.

Monitoring in Linux

In Linux, memory deduplication statistics are shown by **amsstat** command, which is provided by the *powerpc-utils* package. RedHat Enterprise Linux and SuSE Linux Enterprise Server include the powerpc-utils package within their Power server releases.

The **amsstat** command captures memory statistics relevant to an Active Memory Sharing environment. You can run this command once, or set it to run repeatedly for a specified number of times at a particular interval in seconds.

The standard output for **amsstat** is shown in Example 6-5. This command gathers more information for AMS. The example shows only the section relevant to memory deduplication.

Example 6-5 Monitoring memory coalescing in Linux with amsstat

```
[root@RH63 ~]# amsstat
Mon Dec 17 11:15:34 EST 2012

...

Entitlement Information:

     entitled_memory:               80740352
     mapped_entitled_memory:        2609152
     entitled_memory_group_number:  32774
     entitled_memory_pool_number:   0
     entitled_memory_weight:        128
     entitled_memory_pool_size:     25769803776
     entitled_memory_loan_request:  -290082816
     backing_memory:                4015951872
     coalesced_bytes:               100892672
     pool_coalesced_bytes:          323366912
     cmo_enabled:                   1
     cmo_faults:                    0
     cmo_fault_time_usec:           0
     cmo_primary_psp:               2
     cmo_secondary_psp:             65535

...
```

The following statistics are relevant to Active Memory Deduplication:

backing_memory Amount of physical memory, in bytes, that is assigned to the partition. This value changes over time based on the load of all of the partitions in the shared memory pool.

coalesced_bytes Number of bytes assigned to the LPAR that have been coalesced with other identical pages either from within the LPAR, or from another LPAR.

pool_coalesced_bytes Number of bytes in the system shared memory pool that have been coalesced with identical pages.

Refer to the **amsstat** man page for a more detailed explanation of the statistics available.

Monitoring through the Hardware Management Console

The HMC can be used to monitor Active Memory Deduplication through its command-line interface, or through its GUI. You must enable the collection of partition utilization data on the HMC to see the statistical reports.

The command **lslparutil** provides deduplication information, as shown in Example 6-6.

Example 6-6 Output of the lslparutil command showing deduplication statistics

```
hscroot@hmc8:~> lslparutil -r mempool -m p740
time=12/14/2012
15:02:22,event_type=sample,resource_type=mempool,sys_time=12/14/2012
15:04:59,curr_pool_mem=24576,lpar_curr_io_entitled_mem=308,lpar_mapped_
io_entitled_mem=24,lpar_run_mem=34816,sys_firmware_pool_mem=282,page_fa
ults=5425430,page_in_delay=99565925172,mem_dedup=1,dedup_pool_mem=352.0
664,lpar_dedup_mem=3269.6641,dedup_cycles=268954112996
```

After you set up statistics collection, you can view the results of Active Memory Deduplication in action. The following statistics about memory deduplication are reported by HMC CLI:

mem_dedup Indicates whether Active Memory Deduplication is enabled for the shared memory pool.

dedup_pool_mem The amount of deduplicated memory (in megabytes) in the shared memory pool.

lpar_dedup_mem The total amount of partition logical memory (in megabytes) that has been deduplicated.

dedup_cycles The number of processing cycles spent deduplicating data since Active Memory Deduplication was enabled for the shared memory pool.

6.2.3 Test scenarios

To evaluate the benefits of Active Memory Deduplication, some tests were run on a system and the deduplication statistics collected. The tools used to collect the performance statistics are described in 6.2.2, “Monitoring tools” on page 137.

The test scenarios were run on an IBM POWER7 processor-based system that was configured with two Virtual I/O Servers and 30 partitions divided into three groups of 10 partitions each. One group ran AIX, one group ran Linux, and one group ran IBM i.

Table 6-1 shows the general system configuration used for all of the tests that are described in this section. More test-specific setup is shown within the particular test sections.

Table 6-1 The system configuration that was used in scenarios

Component	Value
System model	POWER7 processor-based server
System memory	128 GB
System processor	16 core 3.0 GHz
Partitions	10 Linux 10 AIX 10 IBM i Two Virtual I/O Servers
Partition memory entitlement (AIX, Linux, IBM i)	3 GB
Partition processor entitlement (AIX, Linux, IBM i)	0.5 entitled processing units One virtual processor
Partition storage (AIX, Linux, IBM i)	20 GB in SAN IBM DS4000® drives

Comparing workloads and results: The results that are presented do not represent absolute performance numbers. They provide a way to compare workloads independent of factors that influence the raw performance of a system, such as processor throughput, processor count, and I/O access speeds. Always test a production workload on a test system to ensure that the performance meets your needs.

Scenario 1: Multiple partitions, same workload

In this scenario, the baseline was generated by running a similar workload on all partitions, without activating deduplication. During the run, Active Memory Deduplication was enabled. Table 6-2 details the configuration used.

Table 6-2 Scenario 1: Same workload on all partitions

Setting	Value
Active Memory Deduplication	Enabled
Active Memory Sharing pool size	30 GB
Active Memory Sharing maximum pool size	95 GB
Operating system running on the LPARs	Linux (10 partitions)
Workload	Same workload on all partitions
Deduplication table ratio	1/1024 (default)

Figure 6-2 shows the amount of memory that was coalesced within the partition and the amount that was coalesced in the shared memory pool.

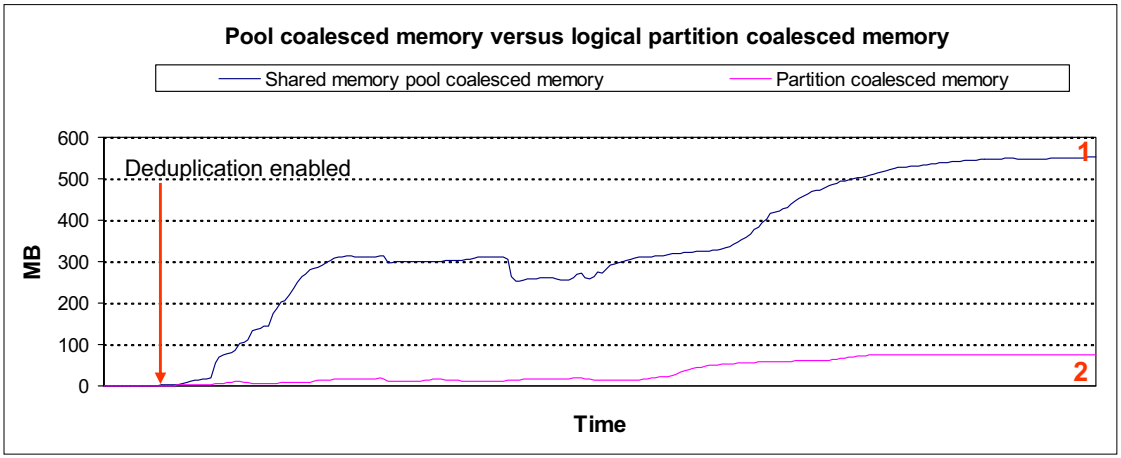


Figure 6-2 Scenario 1 coalesced memory

Line 1 represents the number of coalesced bytes for the pool. This line indicates the total amount of physical memory that was freed to the Active Memory Sharing pool by all of the Active Memory Sharing partitions.

Line 2 indicates the number of coalesced bytes within the partition where data was collected. This value is the total logical memory that was deduplicated in that partition.

Scenario 2: Different workloads

In this scenario, the 10 partitions run different workloads. Table 6-3 summarizes the configuration for this test scenario:

Table 6-3 Scenario 2: Different workloads running

Setting	Value
Active Memory Deduplication	Enabled
Active Memory Sharing pool size	30 GB
Active Memory Sharing maximum pool size	95 GB
Operating system running on the partitions	Linux (10 partitions)

Setting	Value
Workload	Five partitions running a Java workload Three partitions running a network-intensive workload Two partitions running a file system-intensive workload
Deduplication table ratio	1/1024 (default)

When the partitions run a similar workload, it is more likely that identical pages exist in main memory. However, when workloads are different, finding identical memory pages happens less frequently, and therefore the gains with memory deduplication tend to be smaller.

Figure 6-3 shows the amount of deduplicated memory in the shared memory pool for scenario 1 (same workload on all partitions), and the amount for scenario 2 (different workloads).

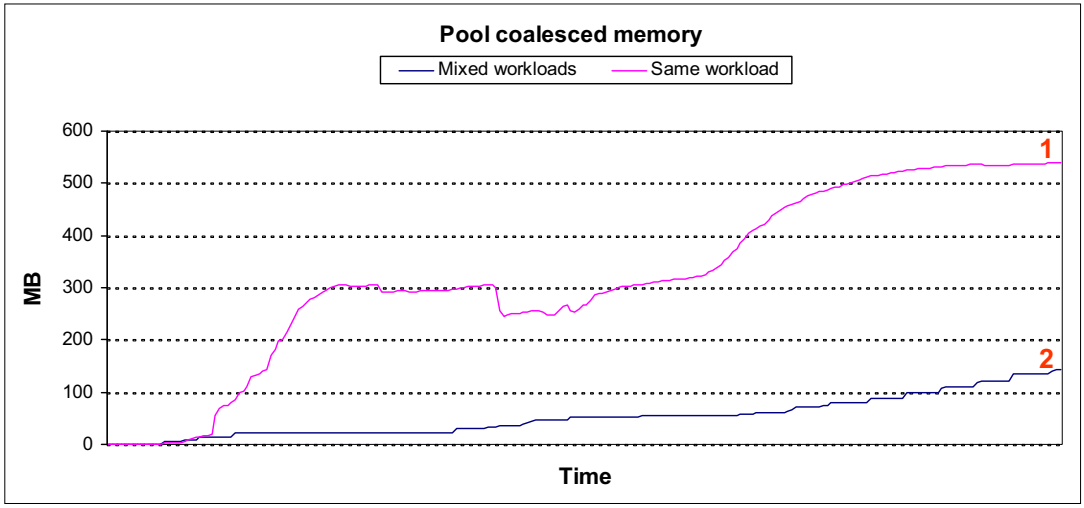


Figure 6-3 Memory deduplication for scenarios 1 and 2

You can see that the number of coalesced pages depends directly on the workload. The more similar the workloads among the partitions, the more memory is deduplicated and made available. In this test, the deduplicated memory for the similar workloads was almost four times larger than for different workloads.

Scenario 3: Multiple OSs, with memory overcommitment

Active Memory Deduplication results in better use of the shared memory pool when it is configured to have logical memory overcommitment, which is when the sum of the wanted memory of the partitions in the shared memory pool exceeds the pool size. If memory is not overcommitted, although there are savings by using deduplication, these savings are not being used by the partitions because they already have the memory that they require.

This scenario shows a typical production environment, running partitions with different operating systems, all sharing the memory pool. Even when different operating systems run on a shared memory environment, pages can still be identical among them, and the platform can benefit from deduplication.

Table 6-4 shows the configuration details that were considered for this test scenario.

Table 6-4 Configuration: Multiple OS types with memory overcommitment

Setting	Value
Active Memory Deduplication	Enabled
Active Memory Sharing pool size	75 GB
Active Memory Sharing maximum pool size	100 GB
Operating system running on the LPARs	AIX (10 partitions), Linux (10 partitions), IBM i (10 partitions)
Workload	Same workload for each operating system (set of 10 partitions)
Deduplication table ratio	1/1024 (default)

Each of the 30 partitions has 3 GB of memory, amounting to a total of 90 GB. The shared memory pool size is 75 GB, creating a memory overcommitment of 15 GB.

The first run is with deduplication disabled. The second run is with deduplication enabled. Figure 6-4 shows the number of memory pages that are loaned to the hypervisor in both cases.

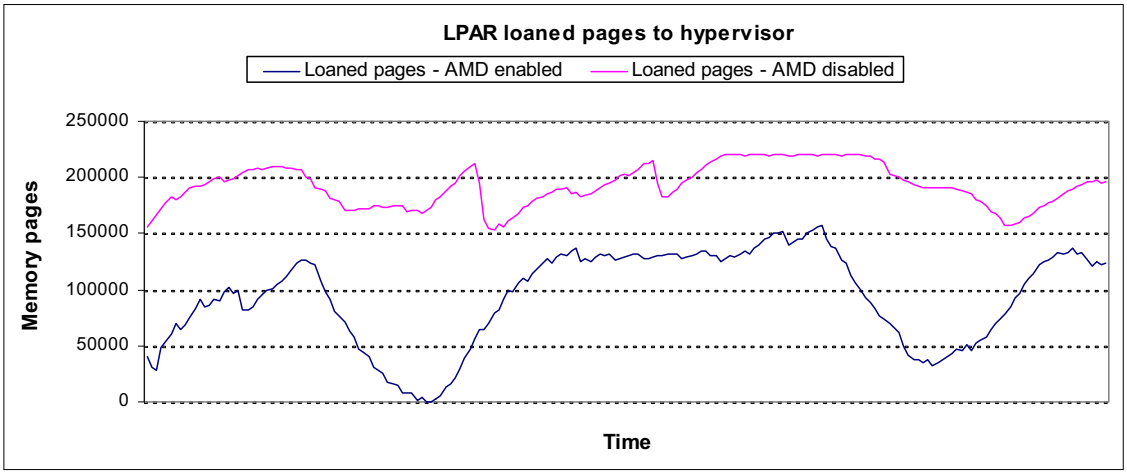


Figure 6-4 Loaned pages with and without deduplication

You can see the benefits of using Active Memory Deduplication in an overcommitted environment by observing that the number of memory pages the partition loans to the hypervisor decreases. This means a smaller overcommitment of the shared pool logical memory, and a potential reduction of partition pages being stolen by the hypervisor and generating paging on the AMS paging devices.

Virtual I/O Server processor utilization

To evaluate the effects of the Virtual I/O Server processing capacity in Active Memory Deduplication, two tests were run with 10 AIX partitions running the same workload, using a shared memory pool. One test had the Virtual I/O Server configured with the entitlement of 0.1 processing units in a capped partition. The second test had the entitlement set for 1.0 processing units in an uncapped partition. All the other configuration settings were kept the same, as shown in Table 6-5.

Table 6-5 Test configuration settings

Setting	Value
Active Memory Deduplication	Enabled
Active Memory Sharing pool size	30 GB

Setting	Value
Active Memory Sharing maximum pool size	95 GB
Operating system running on the partitions	AIX (10 partitions)
Workload	Same workload on all partitions
Deduplication table ratio	1/1024 (default)
Virtual I/O Server processing resources for Test 1	Processing units: 0.1 Virtual processors: 1 Partition mode: capped
Virtual I/O Server processing resources for Test 2	Processing units: 1.0 Virtual processors: 4 Partition mode: uncapped (weight=255)

Figure 6-5 shows the number of coalesced pages in the memory pool during both tests.

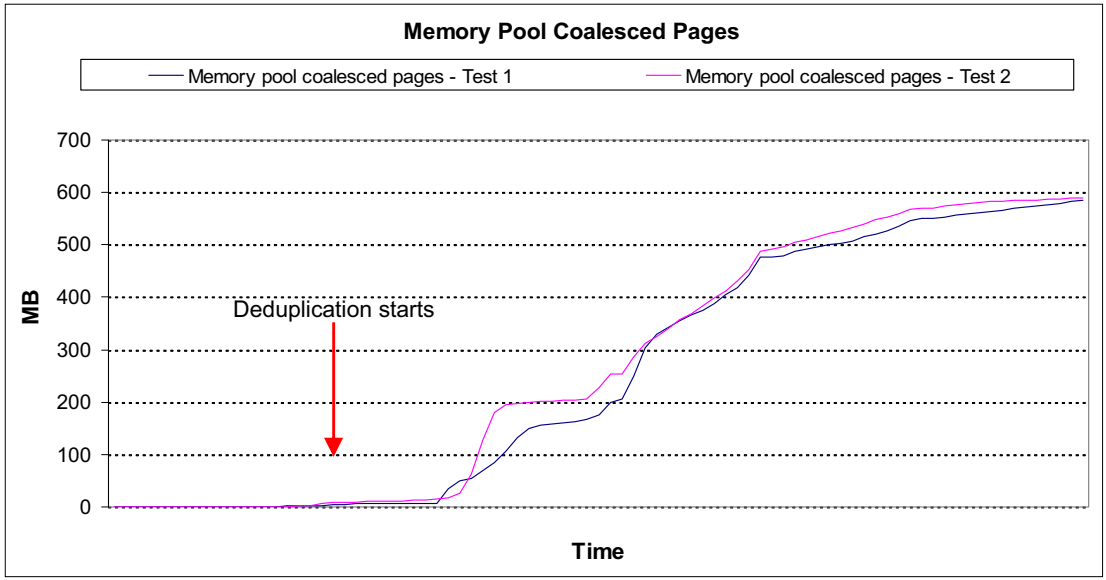


Figure 6-5 Effects of VIOS processing resources on memory deduplication

After some time, the amount of deduplicated memory in the shared pool is practically the same in both tests, differing by less than 5%. This suggests that the amount of processing resources on the Virtual I/O Server, although important

for the deduplication process, is not a critical factor. If you have spare processing resources available on the Virtual I/O Server, the deduplication process runs efficiently.

6.2.4 Troubleshooting

This section targets problems that you might face when you enable Active Memory Deduplication in your environment.

Memory pool performance statistics are not displayed

To retrieve performance information from the shared memory pool, the partition must be authorized to collect performance data from the system. This is enabled on the partition properties on the HMC, as shown in Figure 6-6.

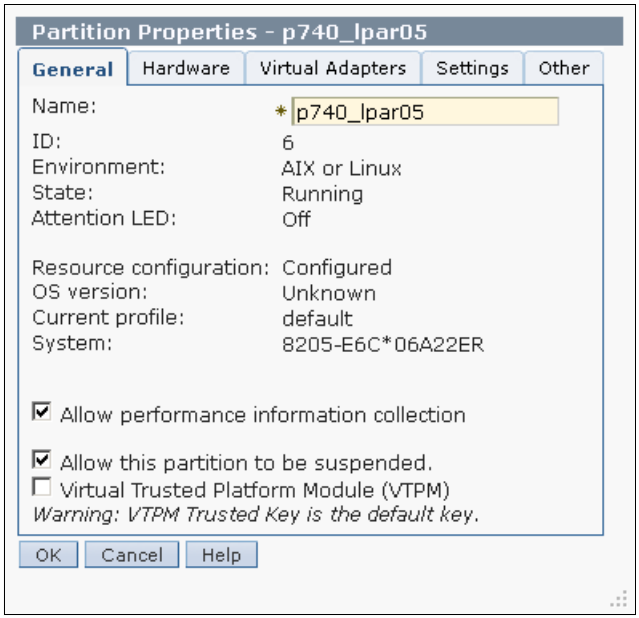


Figure 6-6 Enabling the partition to collect performance data

The shared memory pool coalesced bytes value is not reported by the corresponding operating system command.

If the partition does not have the authority to collect performance information, the commands to retrieve performance statistics for deduplication do not display memory pool statistics.

In Linux, the **amsstat** command does not display the **pool_coalesced_bytes** field. In AIX, the **mpgcol** column displays zero, as shown in Example 6-7.

Example 6-7 AIX mpgcol column with a zero value

```
root@aix1dedup /root # lparstat -mp

System configuration: lcpu=4 mem=3072MB mpsz=40.00GB iome=111.00MB
iomp=10 ent=0.50
physb   hpi  hpit  pmem  pgcol mpgcol ccol %entc  vcsw
----- ----- ----- ----- ------ ------ ---- ----- -----
 0.00   214    55  0.82  654.4   0.0   0.0   0.0 301550
```

Generally, configure all partitions in the system with the ability to collect performance information.

amsstat: command not found message on Linux

In Linux, if the user is trying to monitor the coalesced bytes and the **amsstat** command is not found, it means that the **powerpc-utils** package is not installed. This package must be installed to use the command.

Error code HSCLA4F3 when changing the deduplication table

If you are changing the deduplication table ratio and Active Memory Deduplication is not enabled on the system, the following error is returned:

```
HSCLA4F3 The deduplication table ratio cannot be specified when Active
Memory Deduplication is disabled for the shared memory pool.
```

Turn on Active Memory Deduplication before you change the ratio. For more information about how to turn on Active Memory Deduplication in your system, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

Error code HSCLA4F2 when disabling Active Memory Deduplication

If you get the error code HSCLA4F2 when you disable Active Memory Deduplication, it means that Active Memory Deduplication is already disabled for the shared memory pool.

Error code HSCLA4F1 when enabling Active Memory Deduplication

If you get the error code HSCLA4F1 when you enable Active Memory Deduplication, it means that Active Memory Deduplication is already enabled for the shared memory pool.

Error code HSCLA4F0 when enabling Active Memory Deduplication

If the system returns the error code HSCLA4F0 when you try to enable Active Memory Deduplication, it means that the managed system does not support Active Memory Deduplication. Check the requirements for Active Memory Deduplication in *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

7

Active Memory Expansion

Active Memory Expansion provides compression of in-memory data to expand the effective memory capacity of an IBM POWER7 or POWER7+™ system. It is configurable on a per-partition basis. This chapter describes the maintenance and monitoring tasks that are related to Active Memory Expansion.

This chapter includes the following sections:

- Monitoring your workload to size the Active Memory Expansion
- Memory monitoring tools

7.1 Managing Active Memory Expansion

Active Memory Expansion (AME) is currently only supported on AIX partitions on POWER7 and POWER7+ systems. A minimum level of AIX V6.1 TL4 SP2 is required.

When you consider enabling Active Memory Expansion for an existing workload, use the **amepat** command to provide guidance on possible Active Memory Expansion configurations for the workload. For more information about the basic usage of **amepat**, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940. This section explores more options for running **amepat** in a workload where Active Memory Expansion is already enabled.

Important: To get the best possible results from Active Memory Expansion, run the **amepat** tool during the period of peak usage of the workload. It ensures that the tool captures peak of utilization and memory information of the workload.

Use **amepat** with no arguments to display the current Active Memory Expansion settings. An example output is shown in Example 7-1.

Example 7-1 Displaying Current AME Configuration with the amepat command

```
System Configuration:
---------------------
Partition Name                  : p740_lpar01
Processor Implementation Mode   : POWER7
Number Of Logical CPUs          : 8
Processor Entitled Capacity     : 0.20
Processor Max. Capacity         : 2.00
True Memory                     : 1.50 GB
SMT Threads                     : 4
Shared Processor Mode           : Enabled-Uncapped
Active Memory Sharing           : Disabled
Active Memory Expansion         : Enabled
Target Expanded Memory Size     : 2.00 GB
Target Memory Expansion factor : 1.34
```

You can also check the current Active Memory Expansion configuration on the Hardware Management Console (HMC) in the Memory tab of the partition's profile, as shown in Figure 7-1.

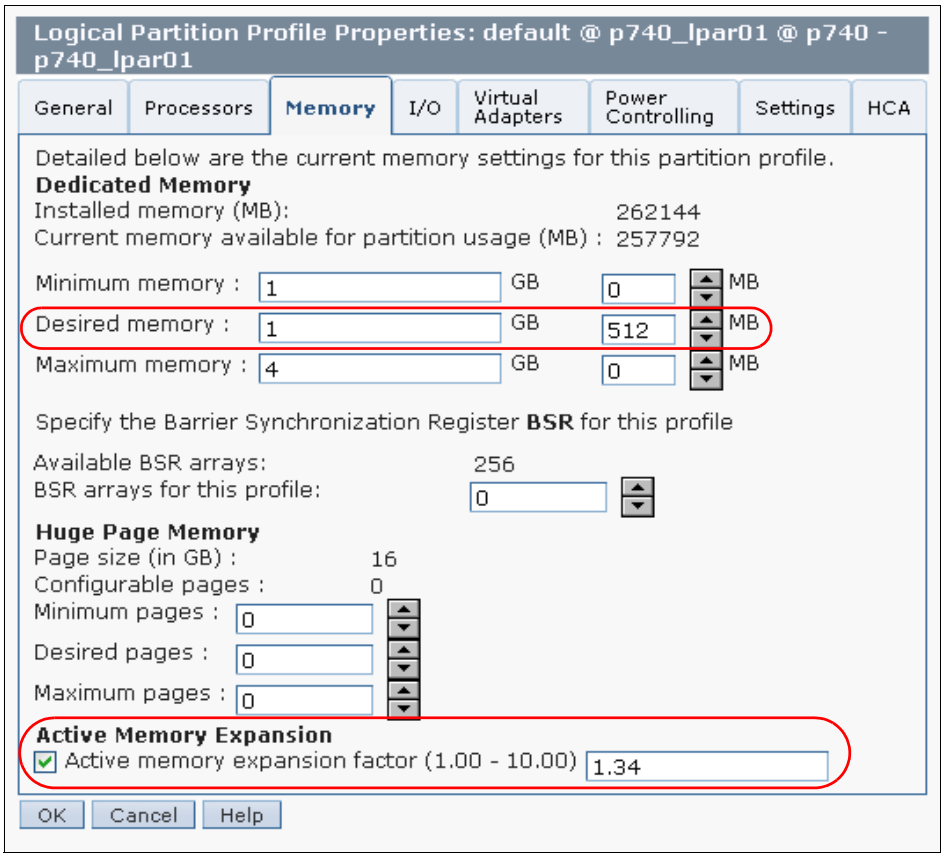


Figure 7-1 Current Active Memory Expansion configuration

To monitor your workload for a duration of 10 minutes with 5 minute sampling interval and 2 samples, use the following command:

```
amepat 5 2
```

You might also want to limit the modeled memory expansion factor to refine the report. The following command tells the monitoring tool to use memory expansion factors between 2.5 and 7.5 at 0.25 incremental factor with a duration of 10 minutes:

```
amepat -e 2.5:7.5:0.25 10
```

If you want to monitor your workload for 5 minutes and limit modeled AME processor usage to 10%, run the following command:

```
amepat -c 10 5
```

These options are useful when you want to check whether your current configuration for Active Memory Expansion can be improved in terms of memory gain and processor usage. For more options and detailed information, see the **amepat** online help.

Example 7-2 shows the current configuration compared to the suggested configurations.

Example 7-2 Output for suggested memory expansion configurations

```
[truncated output]
...
Expansion    Modeled True      Modeled             CPU Usage
Factor       Memory Size       Memory Gain         Estimate
---------    -------------     ------------------  -----------
     1.15          1.75 GB       256.00 MB [ 14%]  0.00 [  0%]
     1.34          1.50 GB       512.00 MB [ 33%]  0.00 [  0%] <<
CURRENT CONFIG
     1.60          1.25 GB       768.00 MB [ 60%]  0.01 [  1%]
     2.00          1.00 GB         1.00 GB [100%]  0.01 [  1%]
     2.67        768.00 MB         1.25 GB [167%]  0.01 [  1%]
     4.00        512.00 MB         1.50 GB [300%]  0.01 [  1%]

...
[truncated output]
```

After you retrieve the suggested configuration from the Active Memory Expansion monitoring tool, you can change the memory settings for your AIX logical partition as shown in Figure 7-2.

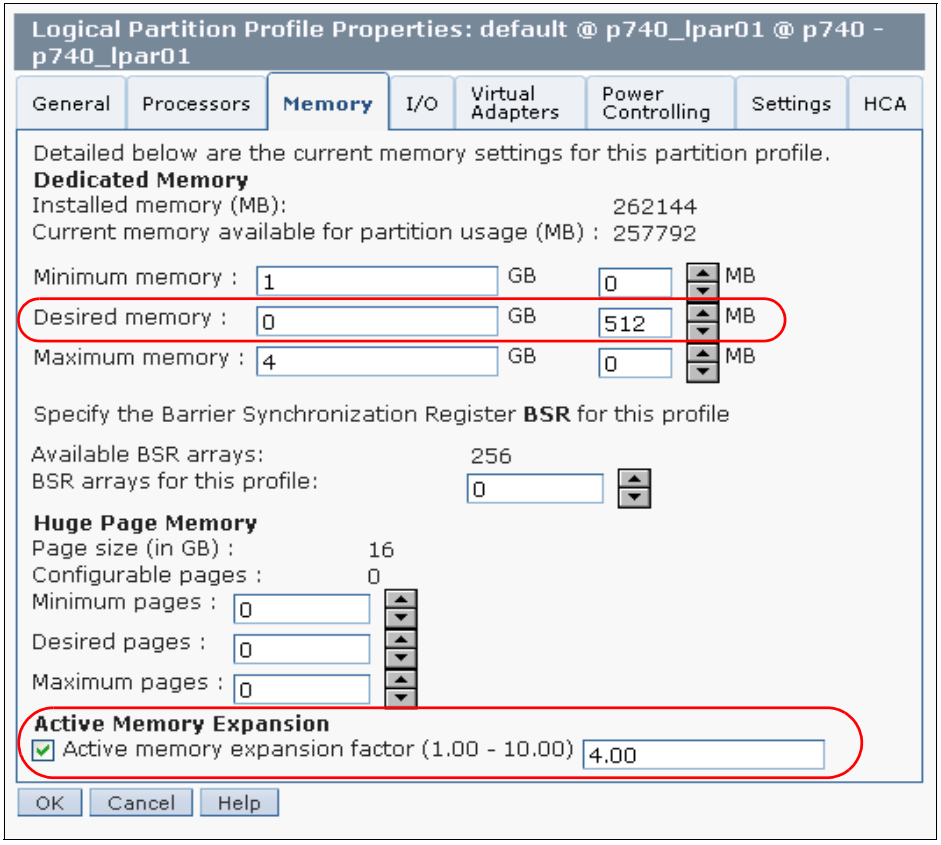


Figure 7-2 Changing the Active Memory Expansion configuration

7.2 Monitoring Active Memory Expansion

A number of commands are available to monitor the Active Memory Expansion configuration of an AIX partition. The following sections describe these commands in more detail.

Tip: POWER7+ processors include an AME compression accelerator that reduces the amount of partition processing that is required for AME compression and decompression operations. No additional configuration is required. AME automatically offloads the compression work to the accelerator on systems with POWER7+ processors.

7.2.1 The amepat command

The **amepat** command provides a summary of the Active Memory Expansion configuration, and can be used for monitoring and fine-tuning the configuration. It shows the current configuration and statistics of the system resource usage over the monitoring period. It can collect metrics over time while a workload is running.

Example 7-3 shows sample output. This partition has an expansion factor of 5, which is caused when the workload that the partition is running is too high. The **amepat** command shows a memory deficit in the AME Statistics section. This means the workload cannot be compressed enough to achieve the target memory size of 10 GB. Also, note the large amount of processor capacity that is used for compressing and decompressing memory pages.

To eliminate the memory deficit, two options are available:

- Reducing the expansion factor. This solution also reduces the expanded memory size for the partition, which might then be too small for the intended workload.
- Adding more physical memory to the partition.

Both options can be run dynamically without requiring a shutdown of the partition.

Example 7-3 shows sample output from the **amepat** command.

Example 7-3 Monitoring Active Memory Expansion with the amepat command

```
# amepat 1 5

Command Invoked                : amepat 1 5
```

```
Date/Time of invocation        : Wed Dec  1 15:46:42 EST 2010
Total Monitored time           : 5 mins 14 secs
Total Samples Collected        : 5

System Configuration:
---------------------
Partition Name                 : P7_1_AIX
Processor Implementation Mode  : POWER7
Number Of Logical CPUs         : 8
Processor Entitled Capacity    : 0.20
Processor Max. Capacity        : 2.00
True Memory                    : 2.00 GB
SMT Threads                    : 4
Shared Processor Mode          : Enabled-Uncapped
Active Memory Sharing          : Disabled
Active Memory Expansion        : Enabled
Target Expanded Memory Size    : 10.00 GB
Target Memory Expansion factor : 5.00


System Resource Statistics:               Average             Min             Max
---------------------------             -----------       -----------     -----------
CPU Util (Phys. Processors)             0.69 [ 35%]       0.34 [ 17%]     1.10 [ 55%]
Virtual Memory Size (MB)                4269 [ 42%]        967 [  9%]     5350 [ 52%]
True Memory In-Use (MB)                 1828 [ 89%]        962 [ 47%]     2046 [100%]
Pinned Memory (MB)                       927 [ 45%]        915 [ 45%]      938 [ 46%]
File Cache Size (MB)                       1 [  0%]          0 [  0%]        6 [  0%]
Available Memory (MB)                   2776 [ 27%]       1208 [ 12%]     3873 [ 38%]

AME Statistics:                           Average             Min             Max
---------------                         -----------       -----------     -----------
AME CPU Usage (Phy. Proc Units)         0.43 [ 22%]       0.22 [ 11%]     0.64 [ 32%]
Compressed Memory (MB)                  2549 [ 25%]        243 [  2%]     4812 [ 47%]
Compression Ratio                       6.60              5.45            7.90
Deficit Memory Size (MB)                3016 [ 29%]       1839 [ 18%]     3982 [ 39%]

Active Memory Expansion Modeled Statistics        :
------------------------------------------
Modeled Expanded Memory Size :  10.00 GB
Achievable Compression ratio   :6.60

Expansion    Modeled True      Modeled             CPU Usage
Factor       Memory Size       Memory Gain         Estimate
---------    -------------     -----------------   -----------
     1.00         10.00 GB         0.00 KB [  0%]   0.00 [  0%]
     1.49          6.75 GB         3.25 GB [ 48%]   0.00 [  0%]
     2.00          5.00 GB         5.00 GB [100%]   0.00 [  0%]
     2.50          4.00 GB         6.00 GB [150%]   0.00 [  0%]
     2.86          3.50 GB         6.50 GB [186%]   0.00 [  0%]
```

```
     3.34          3.00 GB         7.00 GB [233%]   0.50 [ 25%]
     4.00          2.50 GB         7.50 GB [300%]   1.03 [ 52%]

Active Memory Expansion Recommendation:
---------------------------------------
WARNING: This LPAR currently has a memory deficit of 3016 MB.
A memory deficit is caused by a memory expansion factor that is too
high for the current workload.  It is recommended that you reconfigure
the LPAR to eliminate this memory deficit.  Reconfiguring the LPAR
with one of the recommended configurations in the above table should
eliminate this memory deficit.

The recommended AME configuration for this workload is to configure the LPAR
with a memory size of 3.50 GB and to configure a memory expansion factor
of 2.86.  This will result in a memory gain of 186%. With this
configuration, the estimated CPU usage due to AME is approximately 0.00
physical processors, and the estimated overall peak CPU resource required for
the LPAR is 0.46 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.
```

7.2.2 The topas command

The **topas** command shows the following Active Memory Expansion metrics on the default panel when started with no options:

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 7-4 shows an example **topas** panel.

Example 7-4 Monitoring Active Memory Expansion with the topas command

```
Topas Monitor for host:P7_1_AIX                EVENTS/QUEUES    FILE/TTY
Wed Dec  1 15:23:41 2010   Interval:2          Cswitch   21033  Readch     2465
                                               Syscall     259  Writech     457
CPU     User% Kern% Wait% Idle%   Physc  Entc%  Reads        27  Rawin         0
```

```
Total    45.3  41.3   2.3  11.1    1.27 637.45  Writes        1  Ttyout      457
                                                 Forks         0  Igets         0
Network    BPS  I-Pkts  O-Pkts    B-In   B-Out  Execs         0  Namei         8
Total    650.3    3.01    1.00   138.4   512.0  Runqueue   1.00  Dirblk        0
                                                 Waitqueue   2.0
Disk    Busy%      BPS      TPS  B-Read  B-Writ                    MEMORY
Total     0.3     142K   27.58    142K       0  PAGING            Real,MB   10240
                                                 Faults    75549  % Comp       80
FileSystem           BPS    TPS  B-Read  B-Writ  Steals    75565  % Noncomp     0
Total              2.41K  27.08   2.41K       0  PgspIn        0  % Client      0
                                                 PgspOut       0
Name            PID  CPU%  PgSp Owner           PageIn       59  PAGING SPACE
cmemd        655380 17.4  124K root            PageOut       0  Size,MB    5120
lrud         262152 15.2 76.0K root            Sios         59  % Used        0
nmem        9371680 13.1 64.2M root                               % Free      100
nmem        9240604 13.0 64.2M root            AME
nmem        9306142  9.1 64.2M root            TMEM,MB   512.00KWPAR Activ    0
nmem        8323072  1.9 64.2M root            CMEM,MB   231.37MWPAR Total    0
nmem        6946956  1.6 64.2M root            EF[T/A]    1.89   Press: "h"-help
nmem        8192252  1.5 64.2M root            CI:65.0KCO:73.1K        "q"-quit
nmem        8716300  1.4 64.2M root
nmem        6291484  1.4 64.2M root
nmem        8847376  1.3 64.2M root
nmem        8912914  1.3 64.2M root
nmem        9043990  1.0 64.2M root
```

7.2.3 The vmstat command

The **vmstat -c** command shows the following Active Memory Expansion metrics:

csz Current compressed pool size, in 4-K page units.

cfr Free pages available in the compressed pool, in 4-K page units.

dxm Deficit in Expanded Memory Size, in 4-K page units.

ci Number of page-ins per second from the compressed pool.

co Number of page-outs per second to the compressed pool.

Example 7-5 shows example output of the **vmstat** command.

Example 7-5 Monitoring Active Memory Expansion with the vmstat command

```
# vmstat -c 1

System configuration: lcpu=8 mem=10240MB tmem=2048MB ent=0.20 mmode=dedicated-E
```

```
 kthr                          memory                                  page
faults                cpu
------- ------------------------------------------------- ------------------------
------------------ -----------------------
  r   b        avm        fre        csz        cfr        dxm    ci    co wa    pc    ec
  0   0     709350     502692      30970       3112    1425972 10656 79587 1  1.96 981.0
 54   0     760296     449496      37700       2539    1426418 13224 70464 3  1.84 922.1
 55   0     810819     400238      45343       2946    1426546  9835 68154 4  1.82 908.2
 56   0     870987     519460      53548       2588    1248134 14391 82160 1  1.98 992.3
```

7.2.4 The lparstat command

The **lparstat -c** command shows the following Active Memory Sharing metrics:

%xcpu Indicates the percentage of utilization relative to the overall processor consumption by the logical partition for the AME activity.

xphysc Indicates the number of physical processors that are used for the Active Memory Expansion activity.

dxm Indicates the size of the expanded memory deficit for the LPAR.

Example 7-6 shows example output of the **lparstat** command.

Example 7-6 Monitoring Active Memory Expansion with the lparstat command

```
# lparstat -c 1

System configuration: type=Shared mode=Uncapped mmode=Ded-E smt=4 lcpu=8 mem=10240MB
tmem=2048MB psize=14 ent=0.20

%user  %sys  %wait  %idle physc %entc  lbusy  vcsw phint  %xcpu xphysc    dxm
----- ----- ------ ------ ----- ----- ------ ----- ----- ------ ------ ------
 87.7  12.3    0.0    0.0  1.99 995.7  100.0   795     0   12.1 0.2402   6136
 78.4  18.1    3.5    0.0  1.82 910.1   83.4   931     0   17.7 0.3230   6136
 87.4  12.6    0.0    0.0  2.00 999.9  100.0   800     0   12.8 0.2552   6136
 79.3  17.5    3.2    0.0  1.83 916.7   86.0   863     0   17.4 0.3183   6136
```

7.2.5 The svmon command

The **svmon -O summary=ame** command displays a summary of Active Memory Expansion statistics as shown in Example 7-7.

Example 7-7 Monitoring Active Memory Expansion with the svmon command

```
# svmon -O summary=ame
Unit: page
---------------------------------------------------------------------------------------
               size       inuse        free         pin     virtual  available   mmode
memory      2621440     1231515      960972      237879     1237317     960012   Ded-E
 ucomprsd         -      408137           -
 comprsd          -      823378           -
pg space    1310720        4313

               work        pers        clnt       other
pin          144677           0           0       93202
in use      1231114           0         401
 ucomprsd    407736
 comprsd     823378
---------------------------------------------------------------------------------------
True Memory: 524288

              CurSz        %Cur       TgtSz        %Tgt       MaxSz       %Max     CRatio
ucomprsd     409190       78.05      194548       37.11           -          -          -
comprsd      115098       21.95      329740       62.89      261816      49.94       7.38

               txf         cxf         dxf         dxm
AME           5.00        4.17        0.83      432003
```

Part 3

I/O virtualization

This part describes guidelines for managing and monitoring your virtual storage and network devices.

This part includes the following chapters:

- Network virtualization
- Storage virtualization
- Shared Storage Pools

8

Network virtualization

Network connectivity in the virtual environment is extremely flexible. This section describes how to run common configuration tasks and give guidelines on how to monitor your virtualized network.

It is assumed you are well-versed in setting up a virtual network environment. For more information about this task, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

This chapter includes the following sections:

- Managing network virtualization
- Monitoring network virtualization

8.1 Managing network virtualization

This section covers how to change the IP or the VLAN in a virtualized environment, along with mapping management and tuning packet sizes for best performance.

This section contains the following sections:

- Modifying IP addresses
- Modifying VLANs
- Modifying MAC addresses
- Managing the mapping of network devices
- SEA threading on the Virtual I/O Server

8.1.1 Modifying IP addresses

Many hosts, regardless of operating system, have numerous IP addresses, depending on the number of network interfaces that are configured. Generally, one of these IP addresses is considered to be the primary IP address and is registered against the host name of the system. This is the interface that is used for most administrative tasks, and is the address that is registered in any directory services such as DNS. The remaining IP addresses are often used for specific tasks such as point-to-point connections or access to other networks.

This section describes how to change the IP addresses assigned to the Virtual I/O Server and client partitions, and the implications of these changes.

Important: If the IP address you are modifying is the address that is used for RMC connectivity between the HMC and your partition, verify that this connectivity still exists after the address changes. Otherwise, dynamic LPAR operations will be disabled.

Virtual I/O Server

The primary IP address of the Virtual I/O Server is used for these purposes:

- RMC communication for dynamic LPAR operations on the Virtual I/O Server.
- Remote access to the Virtual I/O Server through telnet or Secure Shell (SSH).
- NIM operations.

This address can be configured in one of these ways:

- On a stand-alone interface that is dedicated solely to system administration.
- On top of a Shared Ethernet Adapter (SEA) device.

Using either method, the IP address is transparent to client partitions, and can be changed without affecting their operation.

For example, if the IP address must be changed on en5 from 9.3.5.108 to 9.3.5.109 and the host name must be changed from VIO_Server1 to VIO_Server2, use the following command:

```
mktcpip -hostname VIO_Server2 -inetaddr 9.3.5.109 -interface en5
```

If you want to change only the IP address or the gateway of a network interface, use the **chtcpip** command:

```
chtcpip -interface en5 -inetaddr 9.3.5.109
```

To change the adapter at the same time, such as from en5 to en8, complete these steps:

1. Delete the TCP/IP definitions on en5 by using the **rmtcpip** command.
2. Run **mktcpip** on en8.

You can also possible to make these changes using the **cfgassist** menu system.

Important: If the IP address you are modifying is configured on an SEA device, take care not to modify or remove the layer-2 device (the ent device). Doing so disrupts any traffic that is serviced by the SEA. Only changes to the layer-3 device (the en device) are transparent to clients of the SEA.

Client partitions

Client partition IP addresses can be changed like a physical environment. There is no specific requirement to modify any configuration on the Virtual I/O Server if there is no requirement to change the VLAN configuration to support the new IP address. VLAN modifications are covered in 8.1.2, “Modifying VLANs” on page 169.

AIX

The primary interface on an AIX partition is used for the same tasks as on a Virtual I/O Server. The process to modify IP addresses is identical.

For an AIX virtual I/O client, to change the IP address on a virtual Ethernet adapter use SMIT or the **mktcpip** command.

This example involves changing the IP address from 9.3.5.113 to 9.3.5.112 and the host name from lpar03 to lpar02. The virtual Ethernet adapter can be modified in the same way you modify a physical adapter, using the following command:

```
mktcpip -h lpar02 -a 9.3.5.112 -i en0
```

IBM i

For an IBM i virtual I/O client, change the IP address on a physical or virtual Ethernet adapter by using the following procedure:

1. Add a TCP/IP interface with the new IP address (9.3.5.123) to an existing Ethernet line description (ETH01) by using the ADDTCPIFC command as follows:

   ```
   ADDTCPIFC INTNETADR('9.3.5.123') LIND(ETH01) SUBNETMASK('255.255.254.0')
   ```

2. Start the new TCP/IP interface by using the STRTCPIFC command:

   ```
   STRTCPIFC INTNETADR('9.3.5.123')
   ```

3. The TCP/IP interface with the old IP address (9.3.5.119) can now be ended and removed by using the ENDTCPIFC and RMVTCPIFC commands:

   ```
   ENDTCPIFC INTNETADR('9.3.5.119')
   RMVTCPIFC INTNETADR('9.3.5.119')
   ```

Alternatively, you can use the CFGTCP command. Selecting the option `1. Work with TCP/IP interfaces` allows a menu-based change of TCP/IP interfaces.

To change the host name for an IBM i virtual I/O client, use the CHGTCPDMN command.

Linux

Changing the IP address of an interface using the **`ifconfig`** command will not persist through a reboot of the operating system.

For systems that run Red Hat Enterprise Linux, the fastest method is to use the **`system-config-network`** application to make the changes. On SUSE systems, use the **`yast`** or **`yast2`** applications. Both are menu driven systems that work in the command shell and update the necessary configuration files.

Note: Red Hat Enterprise Linux Version 6 has deprecated the **`system-config-network`** package, and it is no longer installed by default. You have these options for network configuration:

- Install the optional **`system-config-network-tui`** package.
- Install the **`NetworkManager`** package.
- Edit the interface configuration files in `/etc/sysconfig/network-scripts/` manually.

8.1.2 Modifying VLANs

Changing the VLAN configuration of a partition can be a more in-depth process than changing the IP address of an interface. Because VLANs are configured lower in the networking model than IP, changes are likely to be required in more than one place to ensure that the connectivity is achieved. In addition, there are many ways to configure VLANs in a system to suit your network infrastructure.

It is important to have a strong understanding of VLANs before reading this section. For an introduction and more in-depth description of VLANs, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

Process overview

Regardless of the operating system, there are generally these places where VLAN modifications to a partition must be made:

- The Hardware Management Console. Changes to the virtual Ethernet adapter configuration for partitions are performed in the HMC. VLAN IDs are added and removed here.
- On the Shared Ethernet Adapter in the Virtual I/O Server. The SEA is the bridge between the internal hypervisor networks and external networks. If a partition must participate in a VLAN that is external to the hypervisor, access to the external network is provided by an SEA.
- Within the operating system. Extra configuration within the target operating system might be required for the changes to be effective.

In the HMC, the following methods are available for configuring a VLAN in a partition:

- Add an extra virtual Ethernet adapter that has the Port VLAN ID (PVID) set to the required VLAN ID. This method requires no extra configuration within the operating system. All Ethernet traffic that originates from this interface uses the PVID assigned to the adapter.
- Add an extra virtual Ethernet Adapter to the partition with the required VLAN included in the 802.1q Additional VLANs field. This method requires an operating system that is capable of 802.1q VLAN tagging. Typically, a pseudo interface is configured within the operating system on top of the base adapter to tag traffic with the required VLAN ID.
- Modify the list of 802.1q VLANs on an existing virtual Ethernet adapter to include the required VLAN. As with the previous method, this method requires an operating system that is capable of 802.1q VLAN tagging. The ability to dynamically modify an existing virtual Ethernet adapter using the dynamic LPAR function is dependent on the target operating system and system

firmware. If your environment does not support dynamic VLAN modifications, perform one of the following tasks:

- Unconfigure and remove the adapter from the partition, then add it again with the correct list of VLANs included. This process disrupts traffic that is using the interface.
- Modify the adapter in the partition profile and reactivate the partition.

Table 8-1 shows the current supported versions of components that are required to run dynamic VLAN modifications.

Table 8-1 Required versions for dynamic VLAN modifications

Component	Required version for dynamic VLAN
HMC	7.7.2
System Firmware	efw7.2
Virtual I/O Server	2.2.0.10 FP24
AIX	AIX V6.1 TL 6
IBM i	Not supported
Linux	Not supported

Reminder: The virtual Ethernet adapter supports 20 VLAN IDs in addition to the PVID. If you require more than 20 VLANs, more virtual Ethernet adapters are needed.

IBM i V7R1 TR3 or later supports link aggregation, also referred to as IEEE 802.3ad, IEEE 802.1ax, Etherchannel, or Link Aggregation Control Protocol (LACP). For configuration details, see *IBM i 7.1 Technical Overview with Technology Refresh Updates*, SG24-7858.

Hardware Management Console

All operations can be run through the HMC graphical interface or on the command line. Only examples of dynamic VLAN modifications are shown in this section.

The following examples show how to dynamically modify the virtual Ethernet adapter in slot `5` of the partition named `p750_lpar02`, running on the managed system `p750` to include VLAN ID `200`.

Dynamic VLAN modification in the GUI

To modify the VLAN by using the GUI, complete the following steps:

1. Navigate to the partition and select **Dynamic Logical Partitioning** → **Virtual Adapters** as depicted in Figure 8-1.

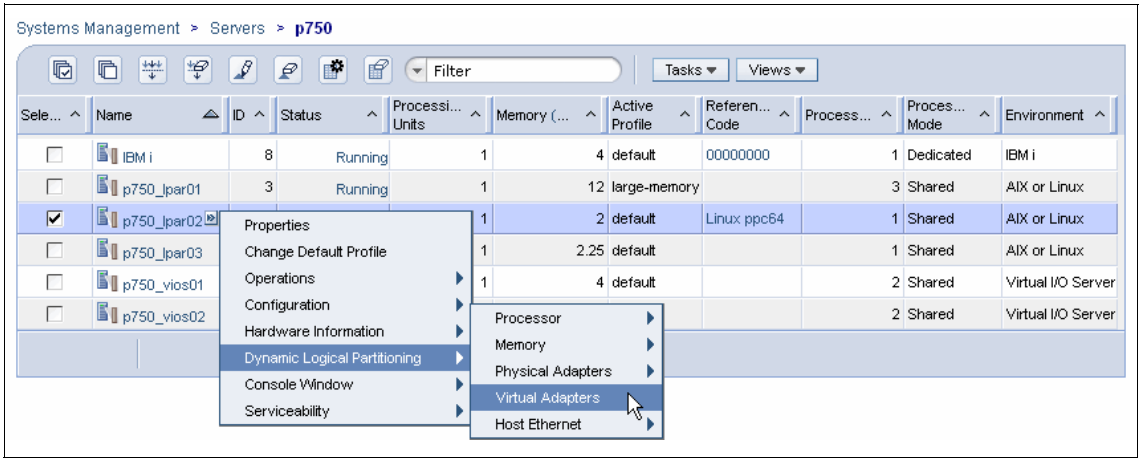


Figure 8-1 Dynamically adding a virtual adapter to a partition

2. Select the virtual Ethernet adapter to be modified and click **File** → **Edit** as shown in Figure 8-2.

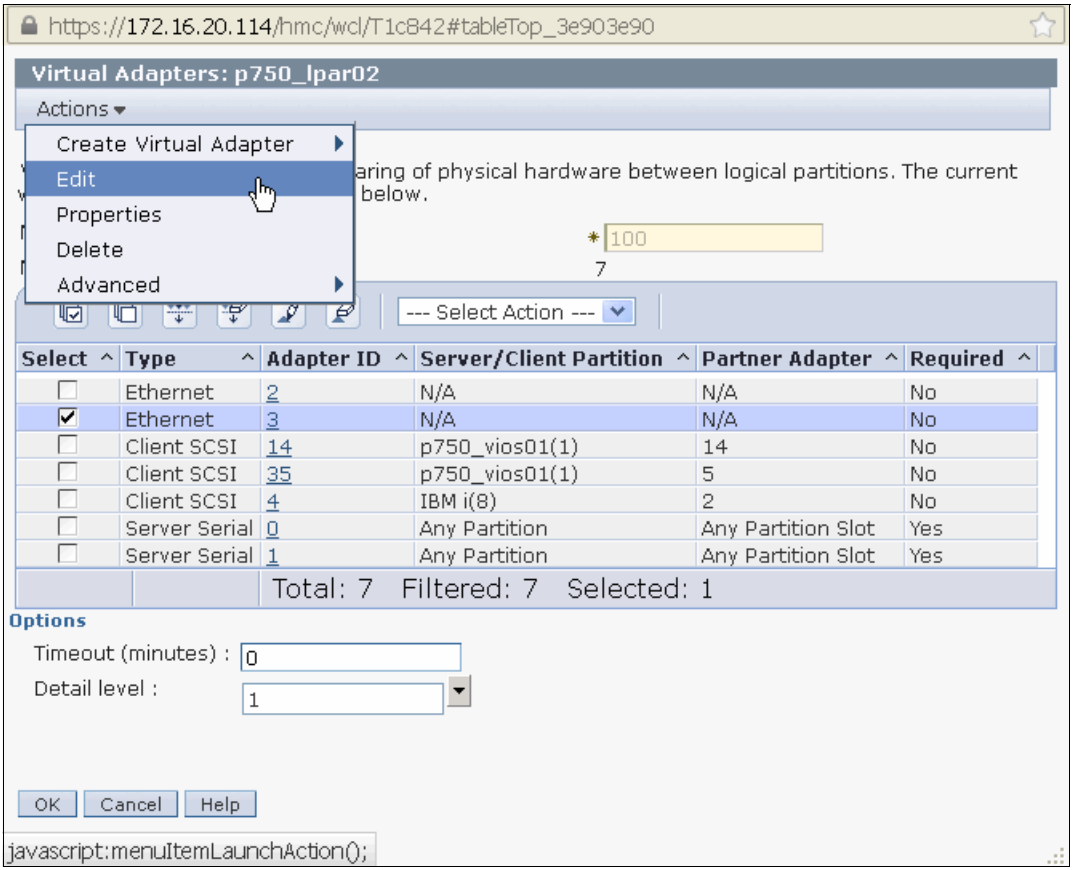


Figure 8-2 Modifying an existing adapter

3. Modify the list of VLANs to include the required VLANs. In the example in Figure 8-3, VLAN ID 200 is being added. More than one ID can be added by supplying a comma-separated list in the **New VLAN ID** field. To remove VLANs, select them in the **Additional VLANs** list and click **Remove**.

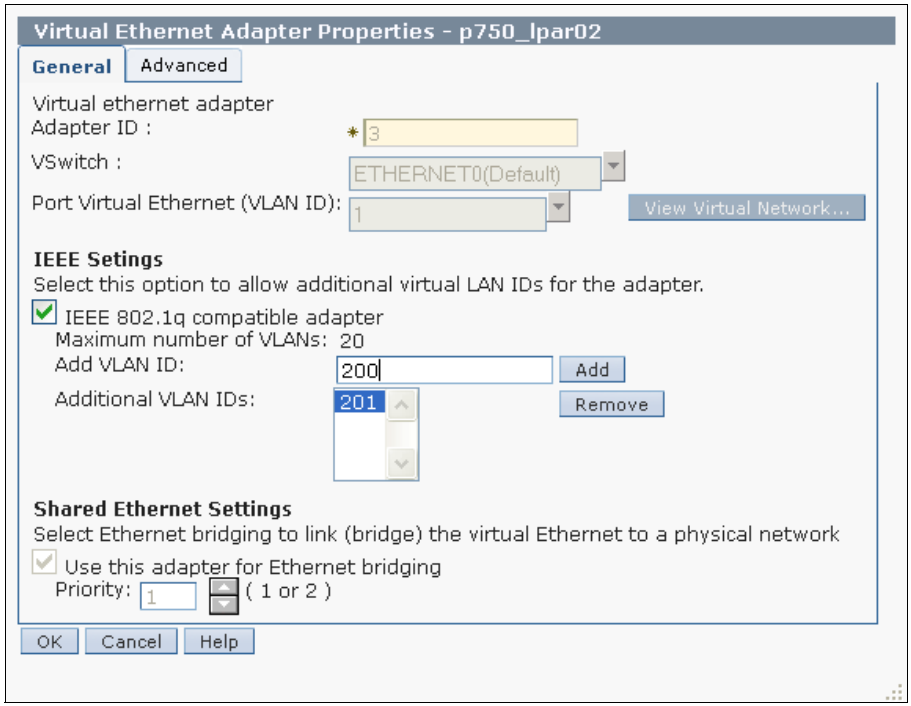


Figure 8-3 Adding VLAN 200 to the additional VLANs field

After the dynamic VLAN operation completes, the new VLAN is available on the virtual adapter.

Dynamic VLAN modification in the CLI

The following examples show the HMC CLI method of modifying VLANs using the same systems as the previous section.

Example 8-1 shows the command to modify the virtual Ethernet adapter in slot 5 of the partition named P7_2_vios2, running on the managed system POWER7_2-061AB2P. The operation is adding the VLAN ID 200 to the additional VLANs field.

Example 8-1 Dynamically modifying the additional VLANs field

```
hscroot@hmc9:~> chhwres -r virtualio --rsubtype eth -m POWER7_2-061AB2P -o s -p
P7_2_vios2 -s 5 -a "addl_vlan_ids+=200"
```

The command in Example 8-2 is an extension of the previous example and slightly more complicated. The operation is enabling the IEEE 802.1q capability and adding the VLAN ID 200 to the additional VLANs field in a single operation.

Example 8-2 Dynamically modifying VLANs field and setting the IEEE 802.1q flag

```
chhwres -r virtualio --rsubtype eth -m POWER7_2-061AB2P -o s -p P7_2_vios2 -s 5
-a "ieee_virtual_eth=1,addl_vlan_ids+=200"
```

The command in Example 8-3 demonstrates removing VLAN ID 200 from the configuration.

Example 8-3 Dynamically removing the VLAN ID

```
chhwres -r virtualio --rsubtype eth -m POWER7_2-061AB2P -o s -p P7_2_vios2 -s 5
-a "ieee_virtual_eth=1,addl_vlan_ids-=200"
```

Integrated Virtualization Manager (IVM)

The IVM does not provide advanced options for creating and managing VLAN tagged interfaces. You can create SEAs with 802.1q compatible tagging, but you must use the CLI commands **chhwres** and **chsyscfg**.

For more information about using these commands with an IVM managed partition, see the technote at:

http://www-01.ibm.com/support/docview.wss?uid=isg3T1010975

Virtual I/O Server

Modifying VLANs in the Virtual I/O Server environment is generally required for one of the following reasons:

- The Virtual I/O Server is required to participate in a particular VLAN.
- A client partition requires access to a particular VLAN through an SEA.
- A combination of these scenarios.

If the Virtual I/O Server is required to participate in the VLAN, use one of the following methods:

- Add a new virtual Ethernet adapter with the PVID field set to the required VLAN ID. All traffic that originates on this adapter uses the PVID. If access to external networks is required for this VLAN, at least one Virtual I/O Server on the managed system must have an SEA capable of bridging the VLAN.
- Add or modify an existing adapter such that the required VLAN is listed in the 802.1q Additional VLAN fields. Then, use the **mkvdev** command to create a VLAN tagged interface over the base adapter. This method also works if the virtual Ethernet adapter is a member of an SEA. The **mkvdev** syntax is demonstrated in Example 8-4.

Important: You cannot dynamically modify extra VLAN ID fields to add a VLAN ID that exists on another trunk adapter within the same virtual switch on a Virtual I/O Server. This configuration might cause unpredictable behavior.

If a client partition requires access to a particular VLAN through an SEA, use one of the following methods to enable the VLAN on the SEA:

- If your system is not capable of dynamic VLAN modifications, or if you have reached the limit of 20 extra VLANs per virtual Ethernet adapter, add an extra virtual Ethernet Adapter with the required VLAN listed in the 802.1q Additional VLAN field. Then, add the new adapter into an existing SEA configuration by using the **chdev** command to modify the *virt_adapters* field of the SEA device. The SEA immediately begins bridging the new VLAN ID without interruption to existing traffic.
- If your system is capable of dynamic VLAN modifications, select the virtual Ethernet adapter that is a member of the SEA and modify the list of 802.1q Additional VLANs to include (or exclude) the required VLAN. The changes take effect immediately without affecting existing traffic on the SEA.

Example 8-4 demonstrates the use of the **mkvdev** command to create a VLAN tagged interface over the ent9 interface. In this example, ent9 is a Shared Ethernet Adapter.

Example 8-4 Creating the VLAN tagged interface

```
$ lsdev -dev ent9
name             status      description
ent9             Available   Shared Ethernet Adapter

$ mkvdev -vlan ent9 -tagid 200
ent10 Available
en10
et10
```

```
$ lsdev -dev ent10 -attr
attribute     value description       user_settable

base_adapter  ent9  VLAN Base Adapter True
vlan_priority 0     VLAN Priority     True
vlan_tag_id   200   VLAN Tag ID       True
```

Important: If your system does not support dynamic VLAN modifications and you are modifying the VLAN list of a virtual Ethernet adapter that is configured in a SEA with *ha_mode* enabled, the HMC will not allow you to reconfigure the list of VLANs on that interface. You must add an extra virtual Ethernet adapter and modify the virt_adapters list of the SEA, or modify the profile of both Virtual I/O Servers and reactivate both Virtual I/O Servers at the same time.

Client partitions

The process of modifying the VLAN configuration in client partitions is similar to modifying VLANs in the Virtual I/O Server, but without the complexity of the SEA.

AIX

To enable an AIX partition to participate in a particular VLAN, use one of the following methods:

- Add a new virtual Ethernet adapter with the PVID field set to the required VLAN ID. All traffic that originates on this adapter will use the PVID.
- Add or modify an existing adapter such that the required VLAN is listed in the 802.1q Additional VLAN fields. Then, use either the **smitty vlan** fast path or the **mkdev** command to create a VLAN tagged interface over the base adapter. The **mkdev** command syntax is demonstrated in Example 8-5.

Example 8-5 demonstrates the use of the mkdev command on AIX. This example creates a VLAN tagged interface for VLAN `200`, using `ent1` as the base adapter.

Example 8-5 Creating the VLAN tagged interface

```
P7_2_AIX:/ # mkdev -c adapter -s vlan -t eth -a base_adapter='ent1' -a
vlan_tag_id='200'
ent2 Available
P7_2_AIX:/ # /usr/lib/methods/defif
en2
et2
```

IBM i

To enable an IBM i partition to participate in a particular VLAN, an extra adapter must be added to the profile that has the PVID field set to the required VLAN.

IBM i does not support 802.1q VLAN tagging. For an IBM i operating system to participate in multiple VLANs, you must configure an adapter for each VLAN.

Linux

To enable a Linux partition to participate in a particular VLAN, use one of the following methods:

- Add a new virtual Ethernet adapter with the PVID field set to the required VLAN ID. All traffic that originates on this adapter uses the PVID.
- Add or modify an existing adapter such that the required VLAN is listed in the 802.1q Additional VLAN fields. Then, use the **`vconfig`** command to create a VLAN tagged interface over the base adapter. The **`vconfig`** command syntax is demonstrated in Example 8-6.

The command in Example 8-6 creates an interface on top of `eth0` that tags frames with VLAN `200`. The resulting device is `eth0.200`.

Example 8-6 Creating a VLAN tagged interface on Linux

```
[root@P7-1-RHEL ~]# vconfig add eth0 200
Added VLAN with VID == 200 to IF -:eth0:-
[root@P7-1-RHEL ~]# ifconfig eth0.200
eth0.200  Link encap:Ethernet  HWaddr 22:5C:2A:1A:23:02
          BROADCAST MULTICAST  MTU:1500  Metric:1
          RX packets:0 errors:0 dropped:0 overruns:0 frame:0
          TX packets:0 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:0
          RX bytes:0 (0.0 b)  TX bytes:0 (0.0 b)
```

To remove this device, use the **`vconfig`** command as shown in Example 8-7.

Example 8-7 Removing a VLAN tagged interface on Linux

```
[root@P7-1-RHEL ~]# vconfig rem eth0.200
Removed VLAN -:eth0.200:-
```

If you receive an error about the `8021q` module not being loaded, use the **`modprobe`** command to load the `8021q` module as shown in Example 8-8.

Example 8-8 Loading the 8021q module into the kernel

```
[root@P7-1-RHEL ~]# modprobe 8021q
```

Consideration: **`vconfig`** is deprecated in Red Hat Enterprise Linux Version 6. Equivalent functionality is provided by the **`iproute`** package, or by editing the interface configuration files in `/etc/sysconfig/network-scripts/` manually.

8.1.3 Modifying MAC addresses

In a Power Systems server, the hardware MAC address of a virtual Ethernet adapter is automatically generated by the HMC when it is defined. Enhancements that were introduced in POWER7 servers allow the LPAR administrator to perform these tasks:

- Specify the hardware MAC address of the virtual Ethernet adapter at creation time.
- Restrict the range of addresses that are allowed to be configured by the operating system within the LPAR.

These features further improve the flexibility and security of the PowerVM networking stack.

The following examples show how to configure these new features.

Hardware Management Console (HMC)

When you create a virtual Ethernet adapter, the Advanced tab of the Create Virtual Ethernet page contains the settings that are related to custom MAC addresses.

Defining a custom MAC address

The MAC Address field displays a value of `auto-assigned` unless you choose to specify a custom MAC address by selecting **Override**.

Figure 8-4 shows the Create Virtual Ethernet Adapter window with the **Override** option selected, and the custom MAC address set to `06:00:00:00:00:AA`.

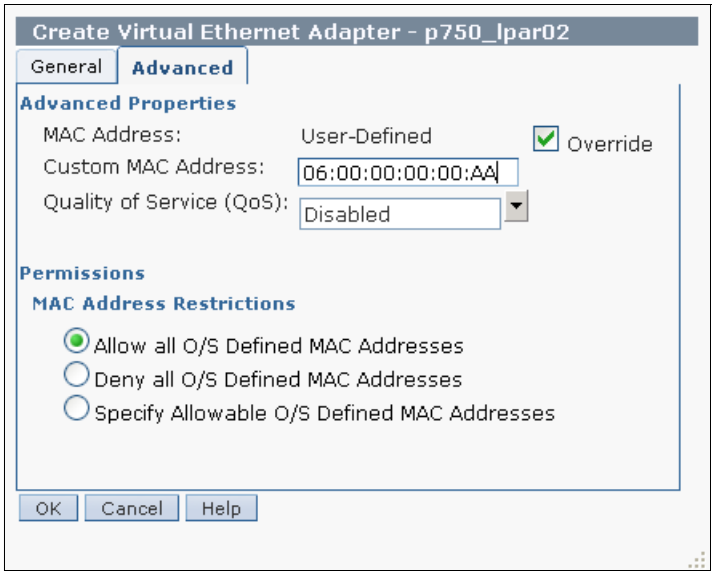


Figure 8-4 Defining a custom MAC address

The MAC address of an Ethernet adapter is a twelve-character (6 bytes) hexadecimal string. The valid characters are from 0 to 9 and A to F, and the characters are not case-sensitive.

There are two rules for custom MAC addresses:

- Bit 1 of byte zero of the MAC address is reserved for Ethernet multicasting, and must always be 0.
- Bit 2 of byte zero of the MAC address must always be 1 because it indicates that the MAC is a locally administered address.

Figure 8-5 shows the MAC address format rules.

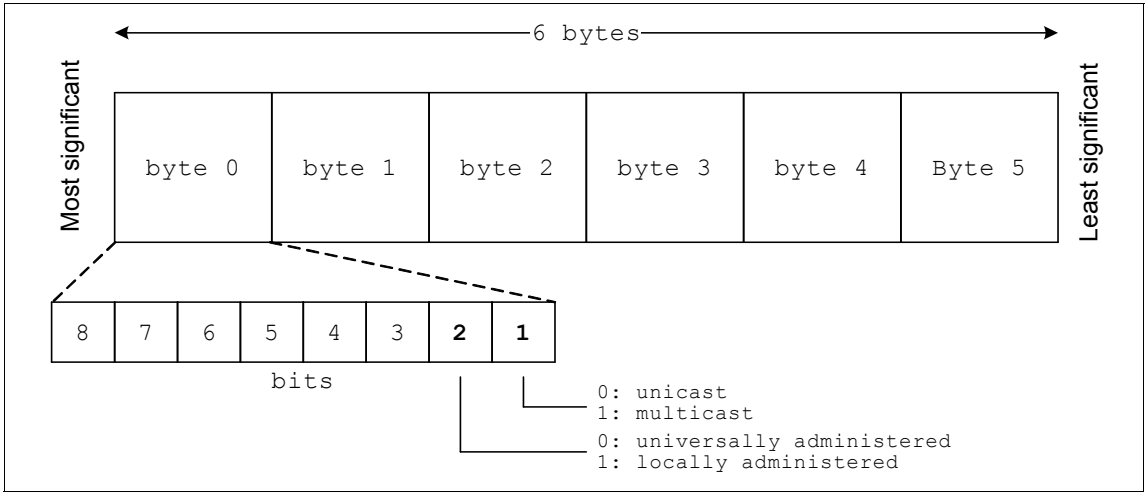


Figure 8-5 MAC address format

These rules mean that valid MAC addresses must conform to the following format, where x is a hexadecimal value (0-9 or A-F).

```
x2:xx:xx:xx:xx:xx
x6:xx:xx:xx:xx:xx
xA:xx:xx:xx:xx:xx
xE:xx:xx:xx:xx:xx
```

Restricting allowable MAC addresses

The MAC Address Restriction option allows the administrator to fine-tune the operating system's ability to override the hardware MAC address of the adapter.

These options are independent of the choice to override the auto-assignment of the hardware MAC address. The restrictions can be used with an auto-assigned MAC address or with a custom MAC address. You have the following options:

- **Allow all O/S Defined MAC Addresses**

  Allows the operating system to override the hardware MAC address with any valid MAC address.

- **Deny all O/S Defined MAC Addresses**

  Denies any modifications to the MAC address of this adapter by the operating system.

- **Specify Allowable O/S Defined MAC Addresses**

  Allows the administrator to define up to four addresses that are allowed to be used in an override by the operating system on this adapter. If you are configuring a custom hardware MAC address on this adapter, it does not need to be included in this list.

Remember: The rules in the previous section about custom hardware MAC address do not apply to addresses defined by the operating system. If the **Allow all O/S Defined MAC Addresses** option is used, the operating system can define any hexadecimal value as the MAC address.

Operating system MAC modifications

This section gives examples on how to perform the following tasks on AIX, IBM i, and Linux:

- Display the current MAC address of an adapter.
- Modify the MAC address within the operating system, and the result of this operation if the MAC address modification is not permitted.

For all examples, an adapter was configured with a custom hardware MAC address of `06:00:00:00:00:AA` in the HMC, and then later changed in the operating system to `06:00:00:00:00:BB`.

AIX

To show the current MAC address of an Ethernet adapter in AIX, use the **`entstat`** or **`lscfg`** command as shown in Example 8-9.

Example 8-9 Listing an adapter MAC address within AIX

```
P7_2_AIX:/ # entstat -d ent2 | grep "Hardware Address"
Hardware Address: 06:00:00:00:00:aa
P7_2_AIX:/ # lscfg -vl ent2 | grep "Network Address"
        Network Address.............0600000000AA
```

Using either command, it is not possible to determine whether the MAC address is the true hardware address or an operating system defined address. This is still true when the output is not filtered as in the previous example.

Modifications to an Ethernet adapter MAC address by AIX are controlled by two parameters on the layer-2 (ent) device:

use_alt_addr This parameter enables the alternate address. Valid values are yes and no.

alt_addr The address to use, represented as a hex value. As an example, the MAC address 06:00:00:00:00:BB would be entered as 0x0600000000BB.

If the value of use_alt_addr is no, the adapter is using the hardware MAC address. If the value is yes, the address that is specified in the alt_addr parameter is in use.

Example 8-10 shows changing the MAC from the hardware address of 06:00:00:00:00:AA to the operating system defined address of 06:00:00:00:00:BB and back again. You can tell that 06:00:00:00:00:AA is the true hardware address because the use_alt_addr field is initially set to no.

The device in this example is ent2. It has been configured in the HMC to **Allow all O/S Defined MAC addresses**.

Example 8-10 Changing an adapter MAC address within AIX

```
P7_2_AIX:/ # lsattr -El ent2 -a use_alt_addr -a alt_addr
use_alt_addr no             Enable Alternate Ethernet Address True
alt_addr     0x000000000000 Alternate Ethernet Address        True

P7_2_AIX:/ # entstat -d ent2 | grep "Hardware Address"
Hardware Address: 06:00:00:00:00:aa

P7_2_AIX:/ # chdev -l ent2 -a use_alt_addr=yes -a alt_addr=0x0600000000BB
ent2 changed

P7_2_AIX:/ # entstat -d ent2 | grep "Hardware Address"
Hardware Address: 06:00:00:00:00:bb

P7_2_AIX:/ # chdev -l ent2 -a use_alt_addr=no
ent2 changed

P7_2_AIX:/ # entstat -d ent2 | grep "Hardware Address"
Hardware Address: 06:00:00:00:00:aa
```

In Example 8-11, the same adapter is redefined with the **Deny all O/S Defined MAC addresses** option. The example shows the layer-2 device (ent2) is still allowed to be modified. However, the layer-3 device (en2) has been deconfigured by the AIX kernel and is not available for use.

Example 8-11 Failed changing of an adapter MAC address within AIX

```
P7_2_AIX:/ # entstat -d ent2 | grep "Hardware Address"
Hardware Address: 06:00:00:00:00:aa

P7_2_AIX:/ # chdev -l ent2 -a use_alt_addr=yes -a alt_addr=0x0600000000BB
ent2 changed

P7_2_AIX:/ # entstat -d ent2

entstat: 0909-003 Unable to connect to device ent2, errno = 22

P7_2_AIX:/ # lsdev | egrep "en2|ent2"
en2        Defined         Standard Ethernet Network Interface
ent2       Available       Virtual I/O Ethernet Adapter (l-lan)
```

IBM i

To display the current MAC address of an Ethernet adapter in IBM i, use the DSPLIND *line_description* CL command as shown in Figure 8-6.

```
Line description . . . . . :   ETH02
Option . . . . . :   *BASIC
Category of line . . . . . :   *ELAN

Resource name  . . . . . :   CMN05
Online at IPL  . . . . . :   *YES
Vary on wait . . . . . :   *NOWAIT
Network controller . . . . . :   ETH02NET
Local adapter address  . . . . . :   0600000000AA
Exchange identifier  . . . . . :   056E970F
Ethernet standard  . . . . . :   *ETHV2
Line speed . . . . . :   *AUTO
Current line speed . . . . . :   1G
Duplex . . . . . :   *AUTO
Current duplex . . . . . :   *FULL
Serviceability options . . . . . :   *NONE
Maximum frame size . . . . . :   1496
```

Figure 8-6 IBM i displaying the line description

To change an Ethernet adapter's MAC address in IBM i, complete these steps:

1. End any TCP interfaces associated with the Ethernet line description.
2. Vary off the line description.
3. Change the adapter's address.
4. Vary on the line description again.
5. Start any associated TCP interfaces again.

Example 8-12 shows the CL commands that were used to change the MAC address of the (virtual) Ethernet adapter in IBM i.

Example 8-12 Changing an Ethernet adapter MAC address within IBM i

```
ENDTCPIFC INTNETADR('172.16.20.196')
VRYCFG CFGOBJ(ETH02) CFGTYPE(*LIN) STATUS(*OFF)
CHGLINETH LIND(ETH02) ADPTADR(060000000OBB)
VRYCFG CFGOBJ(ETH02) CFGTYPE(*LIN) STATUS(*ON)
STRTCPIFC INTNETADR('172.16.20.196')
```

When trying to change the MAC address of a virtual Ethernet adapter that was defined with the Deny all O/S Defined MAC addresses option on the HMC, changing the MAC address in the line description still works. However, trying to vary on the line description fails with a CPI59F1 message `Line ETH02 failed. Internal system failure.`

Linux

To see the current hardware address of an Ethernet adapter in Linux, use the **`ifconfig`** command as shown in Example 8-13.

Example 8-13 Displaying an adapter MAC address within Linux

```
[root@Power7-2-RHEL /]# ifconfig eth1 | grep "HWaddr"
eth1      Link encap:Ethernet  HWaddr 06:00:00:00:00:AA
```

The **`ifconfig`** command can also be used to modify the address. Example 8-14 shows changing the hardware address from `06:00:00:00:00:AA` to `06:00:00:00:00:BB`. The device in this example is `eth1`, and it has been configured in the HMC to **Allow all O/S Defined MAC addresses**.

Example 8-14 Changing an adapter MAC address within Linux

```
[root@Power7-2-RHEL /]# ifconfig eth1 | grep "HWaddr"
eth1      Link encap:Ethernet  HWaddr 06:00:00:00:00:AA

[root@Power7-2-RHEL /]# ifconfig eth1 hw ether 06:00:00:00:00:BB
```

```
[root@Power7-2-RHEL /]# ifconfig eth1 | grep "HWaddr"
eth1      Link encap:Ethernet  HWaddr 06:00:00:00:00:BB
```

The **ifconfig** command can only show you the current MAC address. When using **ifconfig**, you cannot determine whether this is the true hardware MAC address or an operating-system-defined MAC address. This is true even when the output is not filtered as in the previous example. On Power Systems servers, you can determine this information from the /proc file system as shown in Example 8-15. In this example, the adapter was configured in slot 3. Therefore, the details in the 30000003 file are relevant to the adapter (the least significant digits are the hex value of the adapter slot).

Example 8-15 Displaying an adapter firmware MAC address within Linux

```
[root@Power7-2-RHEL /]# grep MAC /proc/net/ibmveth/*
/proc/net/ibmveth/30000002:Current MAC:     6E:8D:DA:FD:46:02
/proc/net/ibmveth/30000002:Firmware MAC:    6E:8D:DA:FD:46:02
/proc/net/ibmveth/30000003:Current MAC:     06:00:00:00:00:BB
/proc/net/ibmveth/30000003:Firmware MAC:    06:00:00:00:00:AA
```

In Example 8-16, the same adapter is redefined with the **Deny all O/S Defined MAC addresses** option. Similar to AIX, the MAC address change appears to work on both the eth1 interface and in the /proc file. However, when you attempt to enable the interface with an IP configuration, it fails. When you revert it back to the original hardware MAC, the command succeeds.

Example 8-16 Failed changing of an adapter MAC address in Linux

```
[root@Power7-2-RHEL /]# ifconfig eth1 | grep "HWaddr"
eth1      Link encap:Ethernet  HWaddr 06:00:00:00:00:AA

[root@Power7-2-RHEL /]# ifconfig eth1 hw ether 06:00:00:00:00:BB

[root@Power7-2-RHEL /]# ifconfig eth1 | grep "HWaddr"
eth1      Link encap:Ethernet  HWaddr 06:00:00:00:00:BB

[root@Power7-2-RHEL /]# grep MAC /proc/net/ibmveth/*
/proc/net/ibmveth/30000002:Current MAC:     6E:8D:DA:FD:46:02
/proc/net/ibmveth/30000002:Firmware MAC:    6E:8D:DA:FD:46:02
/proc/net/ibmveth/30000003:Current MAC:     06:00:00:00:00:BB
/proc/net/ibmveth/30000003:Firmware MAC:    06:00:00:00:00:AA

[root@Power7-2-RHEL /]# ifconfig eth1 172.200.0.100 netmask 255.255.255.0 up
SIOCSIFFLAGS: Machine is not on the network
SIOCSIFFLAGS: Machine is not on the network
```

```
[root@Power7-2-RHEL /]# ifconfig eth1 hw ether 06:00:00:00:00:AA

[root@Power7-2-RHEL /]# ifconfig eth1 172.200.0.100 netmask 255.255.255.0 up
```

8.1.4 Managing the mapping of network devices

One of the keys to managing a virtual environment is tracking what virtual objects correspond to what physical objects. In the network area, this can involve physical and virtual network adapters, and VLANs that span across hosts and switches. This mapping is critical for managing performance and to understand what systems will be affected by hardware maintenance.

In environments that require redundant network connectivity, this section focuses on the SEA failover method rather than the Network Interface Backup method of providing redundancy.

Depending on whether you choose to use 802.1q tagged VLANs, you might need to track the following information:

- For the Virtual I/O Server:
  - Server host name
  - Physical adapter device name
  - Switch port
  - SEA device name
  - Virtual adapter device name
  - Virtual adapter slot number
  - Port virtual LAN ID (in tagged and untagged usages)
  - Extra virtual LAN IDs
- For the virtual I/O client:
  - Client host name
  - Virtual adapter device name
  - Virtual adapter slot number
  - Port virtual LAN ID (in tagged and untagged usages)
  - Extra virtual LAN IDs

Because of the number of fields to be tracked, use a spreadsheet or database program to record this information. Record the data when the system is installed, and track it over time as the configuration changes.

Virtual network adapters and VLANs

Virtual network adapters operate at memory speed. In many cases where more physical adapters are needed, there is no need for more virtual adapters.

The POWER Hypervisor supports tagged VLANs that can be used to separate traffic in the system. Separate adapters can be used to accomplish the same goal. Select the method, or a combination of both, based on common networking practice in your data center.

Virtual device slot numbers

Virtual storage and virtual network devices have a unique slot number. In complex systems, there tend to be far more storage devices than network devices because each virtual SCSI device can communicate with only one server or client. Reserve slot numbers through 20 for network devices on all LPARs to keep the network devices grouped. In some complex network environments with many adapters, more slots might be required for networking.

Increase the maximum number of virtual adapter slots per LPAR above the default value of 10 when you create an LPAR. The appropriate number for your environment depends on the number of LPARs and adapters that you expect on each system. Each unused virtual adapter slot uses a small amount of memory, so balance the allocation with expected requirements. To plan memory requirements for your system configuration, use the IBM System Planning Tool, which is available at:

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

AIX Virtual Ethernet configuration tracing

For an AIX virtual I/O client partition with multiple virtual network adapters, the slot number of each adapter can be determined by using the adapter physical location from the **lscfg** command. For virtual adapters, this field includes the card slot following the letter C, as shown in the Example 8-17.

Example 8-17 Virtual Ethernet adapter slot number

```
# lscfg -l ent*
ent0         U9117.MMA.101F170-V3-C2-T1  Virtual I/O Ethernet Adapter (l-lan)
```

You can use the slot numbers from the physical location field to trace back through the HMC Virtual Network Management option and determine what connectivity and VLAN tags are available on each adapter:

1. From the HMC, click **Systems Management** → **Servers**.
2. Select your Power Systems server and select **Configuration** → **Virtual Resources** → **Virtual Network Management** as shown in Figure 8-7.

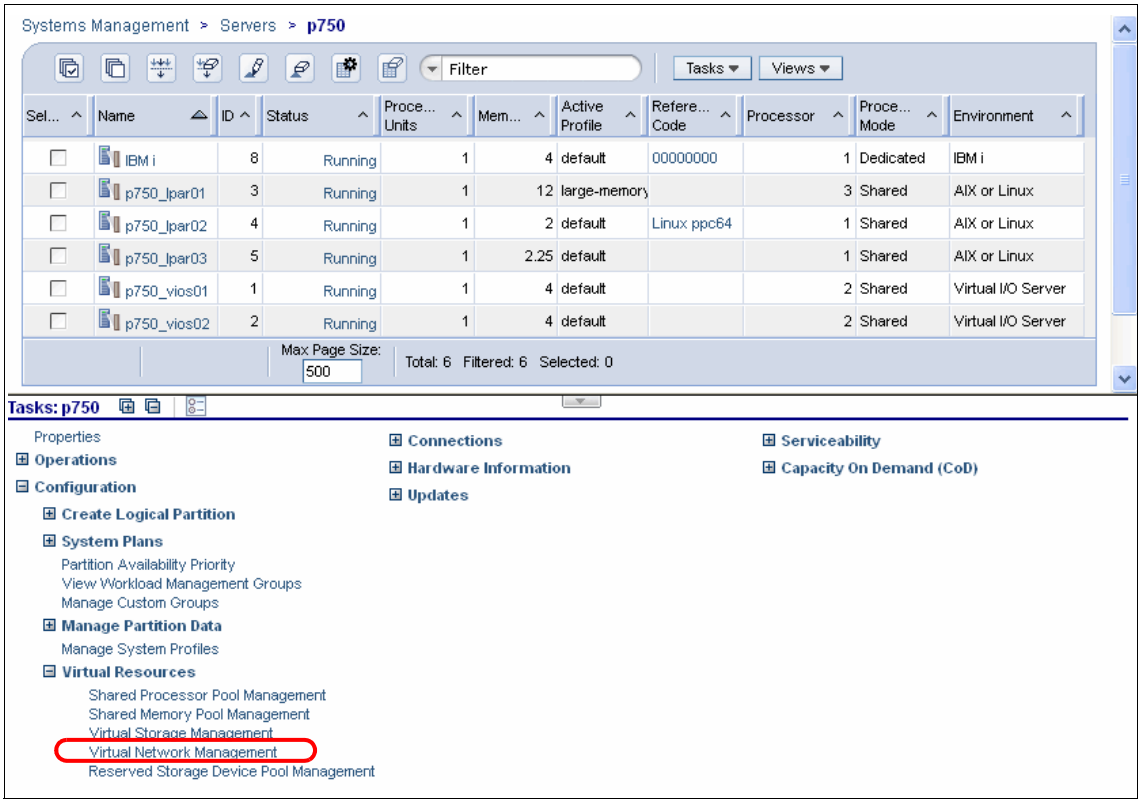


Figure 8-7 HMC Virtual Network Management

3. To determine where your AIX virtual I/O client Ethernet adapter is connected, select a VLAN. In the example, the `p750_lpar01` LPAR virtual I/O client adapter `ent0` in slot `2` is on VLAN1 as shown in Figure 8-8.

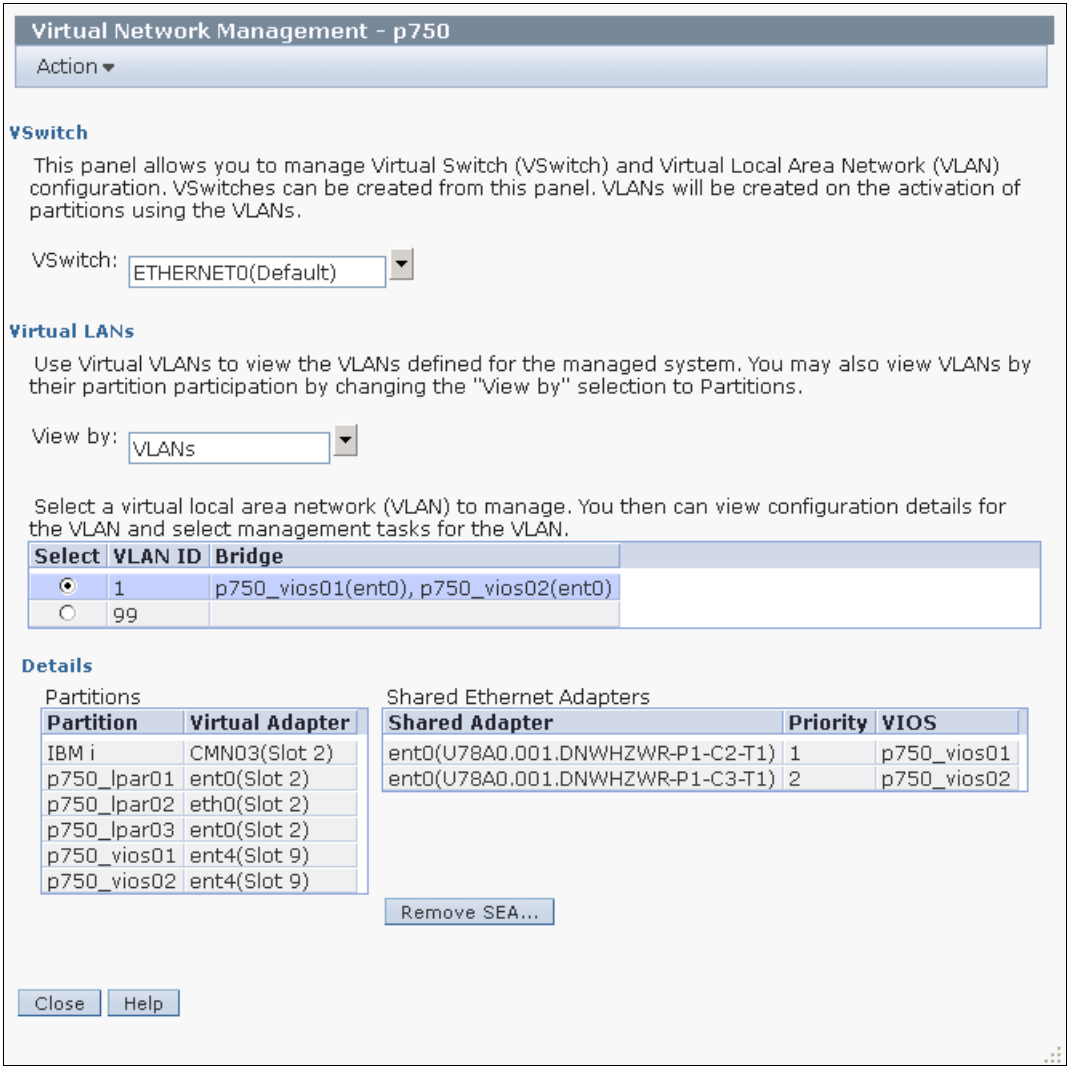


Figure 8-8 Virtual Ethernet adapter slot assignments

IBM i Virtual Ethernet configuration tracing

For an IBM i virtual I/O client partition with virtual Ethernet adapters, the slot number of each adapter can be determined by using the adapter location information:

1. To display the adapter location information, use the WRKHDWRSC *CMN command.
2. Select option **7=Display resource detail** for the virtual Ethernet adapter (type 268C) as shown in Figure 8-9.

```
Type options, press Enter.
 5=Work with configuration descriptions  7=Display resource detail

Opt  Resource       Type  Status              Text
     CMB06          6B03  Operational         Comm Processor
       LIN03        6B03  Operational         Comm Adapter
         CMN02      6B03  Operational         Comm Port
     CMB07          6B03  Operational         Comm Processor
       LIN01        6B03  Operational         Comm Adapter
         CMN03      6B03  Operational         Comm Port
     CMB08          268C  Operational         Comm Processor
       LIN02        268C  Operational         LAN Adapter
 7       CMN01       268C  Operational             Ethernet Port
```

Figure 8-9 IBM i Work with Communication Resources panel

The location field includes the card slot after the letter C as shown with slot 2 in Figure 8-10.

```
 Display Resource Detail
                                                                System:E101F170
 Resource name  . . . . . :   CMN01
 Text . . . . . . . . . . :   Ethernet Port
 Type-model . . . . . . . :   268C-002
 Serial number  . . . . . :   00-00000
 Part number  . . . . . . :


 Location :   U9117.MMA.101F170-V5-C2-T1


 Logical address:
  SPD bus:
    System bus                  255
    System board                128

More...
```

Figure 8-10 IBM i Display Resource Details panel

You can use this slot number from the physical location field to trace back through the HMC partition properties and determine what connectivity and VLAN tags are available on each adapter.

3. From the HMC, click **Systems Management** → **Servers**.
4. Select your Power Systems server.
5. Select your IBM i partition and select **Properties** to open the partition properties window as shown in Figure 8-11.

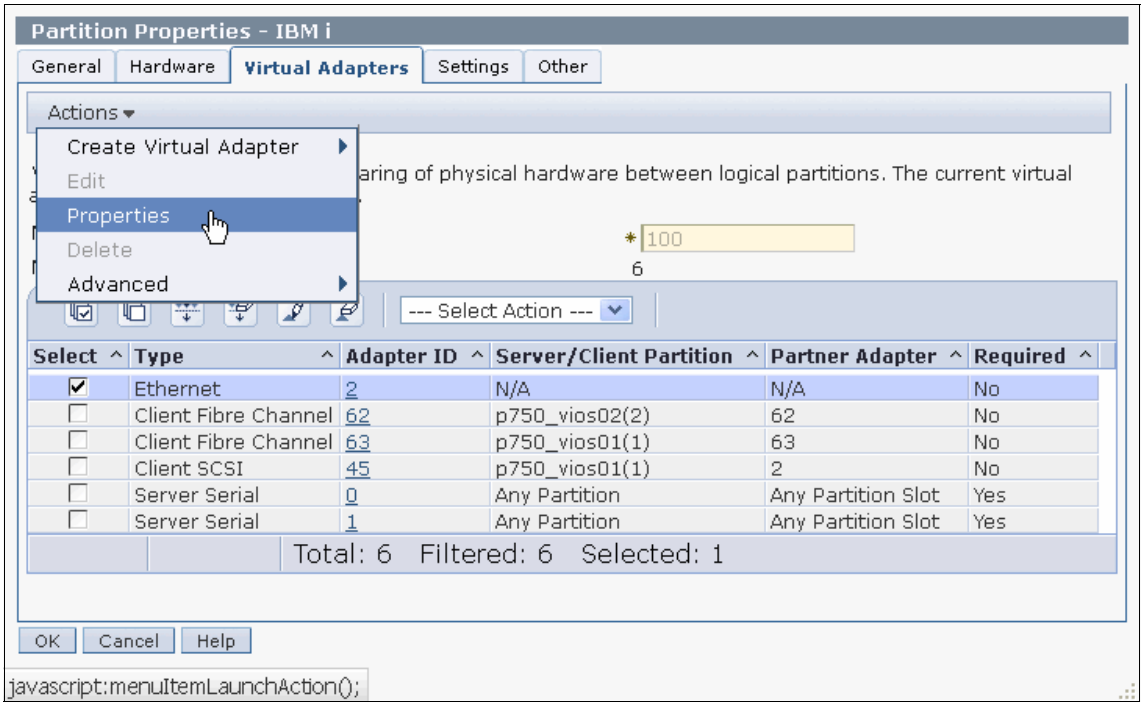


Figure 8-11 HMC IBM i partition properties panel

6. Selecting the IBM i client virtual Ethernet adapter and selecting **Actions** → **Properties** shows, in the example, that the IBM i client virtual Ethernet adapter in slot 2 is on VLAN1, as shown in Figure 8-12.

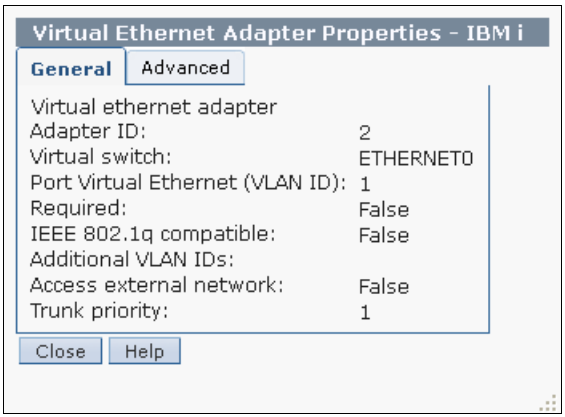


Figure 8-12 HMC Virtual Ethernet Adapter Properties panel

8.1.5 SEA threading on the Virtual I/O Server

The Virtual I/O Server enables you to virtualize both disk and network traffic for AIX, IBM i, and Linux operating system-based clients. The main difference between these types of traffic is their persistence. If the Virtual I/O Server must move network data around, it must do so immediately because network data has no persistent storage. For this reason, the network services that are provided by the Virtual I/O Server (such as the Shared Ethernet Adapter) run with the highest priority. Disk data for virtual SCSI devices is run at a lower priority than the network because the data is stored on the disk. Therefore, there is less of a danger of losing data because of timeouts. The devices are also normally slower.

The shared Ethernet process of the Virtual I/O Server before Version 1.3 runs at an interrupt level that was optimized for high performance. With this approach, it ran with a higher priority than the virtual SCSI if there was high network traffic. If the Virtual I/O Server did not provide enough processor resources for both, the virtual SCSI performance could experience a degradation of service.

Starting with Virtual I/O Server Version 1.3, the Shared Ethernet function is implemented by using *kernel threads*. This configuration allows an even distribution of the processing power between virtual disk and network.

This threading can be turned on and off per SEA by changing the thread attribute. This setting can be changed while the SEA is operating without any

interruption to service. A value of 1 indicates that threading is to be used, and 0 indicates the original interrupt method:

```
$ lsdev -dev ent2 -attr thread
value
0
$ chdev -dev ent2 -attr thread=1
ent2 changed
$ lsdev -dev ent2 -attr thread
value
1
```

Using threading requires a minimal increase of processor usage for the same network throughput. However, with the burst nature of network traffic, enabling threading (this is now the default) is generally a good idea. Network traffic usually comes in spikes, as users log on or as web pages load, for example. These spikes might coincide with disk access. For example, a user logging on to a system can generate a network activity spike because during the logon process some form of password database stored on the disk is accessed or the user profile read.

Consider disabling threading when you have a Virtual I/O Server that is dedicated for network and another dedicated for disk (whether the Virtual I/O server is in a VSCSI, NPIV or SSP environment). This is only a good configuration when you mix extreme disk and network loads together on a processor-constricted server.

Usually the network processor requirements are higher than those for disk. In addition, you will probably have the disk Virtual I/O Server set up to provide a network backup with SEA failover if you want to remove the other Virtual I/O Server from the configuration for scheduled maintenance. In this case, both disk and network will run through the same Virtual I/O Server, so use threading.

8.2 Monitoring network virtualization

After you set up a virtualized environment, verify that all contingencies are in place to enable network connectivity to function during a failure.

This section includes the following topics:

- Monitoring the Virtual I/O Server
- Virtual I/O Server networking monitoring
- AIX client network monitoring
- IBM i client network monitoring
- Linux network monitoring

For more information about monitoring, see 11.1, “Managing Virtual I/O Servers” on page 324.

8.2.1 Monitoring the Virtual I/O Server

You can monitor the Virtual I/O Server by using error logs or by using IBM Tivoli Monitoring.

Error logs

AIX, IBM i, and Linux client logical partitions log errors against failing I/O operations. Hardware errors on the client logical partitions that are associated with virtual devices usually have corresponding errors logged on the Virtual I/O Server. The error log on the Virtual I/O Server is displayed by using the **errlog** command.

However, if the failure is within the client partition, there are typically no errors logged on the Virtual I/O Server. Also, on Linux client logical partitions, if the algorithm for trying SCSI temporary errors again is different from the algorithm that is used by AIX, the errors might not be recorded on the server.

IBM Tivoli Monitoring

You can configure and start the IBM Tivoli Monitoring System agents on the Virtual I/O Server. Doing so enables you to monitor the health and availability of multiple IBM Power Systems servers (including the Virtual I/O Server) from the Tivoli Enterprise Portal. It gathers the following network-related data from Virtual I/O Server:

- Network adapter details
- IBM Tivoli Monitoring Network adapter utilization
- Network interfaces
- Network protocol views
- Shared Ethernet
- Shared Ethernet adapter high availability details
- Shared Ethernet bridging details

You can monitor these metrics from the Tivoli Enterprise Portal Client. For more information about network monitoring using the Tivoli Enterprise Portal, see 18.1.2, “IBM Tivoli Usage and Accounting Manager agent” on page 660.

Testing your configuration

Test your environment configuration periodically to help to avoid errors. The tests can be run in the network configuration and the virtual SCSI configuration.

This section explains how to test a Shared Ethernet Adapter Failover and a Network Interface Backup (NIB). It is assumed that the SEA and NIB are already configured in an environment with two Virtual I/O Servers. For more information about how to configure this environment, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

Testing the SEA Failover

Complete these steps to test whether the SEA Failover configuration works as expected:

1. Open a remote session from a system on the external network to any AIX, IBM i, or Linux client partition. If your session becomes disconnected during any of the tests, it means that your configuration is not highly available. You might want to run a continuous **ping** command to verify that you have connectivity.
2. On Virtual I/O Server 1, check whether the primary adapter is active by using the **entstat** command. In this example, the primary adapter is `ent4`.

   ```
   $ entstat -all ent4 | grep Active
     Priority: 1  Active: True
   ```

3. Perform a manual failover by using the **chdev** command to switch to the standby adapter:

   ```
   chdev -dev ent4 -attr ha_mode=standby
   ```

4. Check whether the SEA Failover was successful by using the **entstat** command. On Virtual I/O Server 1, the entstat output should look as follows:

   ```
   $ entstat -all ent4 | grep Active
     Priority: 1  Active: False
   ```

   You should also see the following entry in the errorlog when you issue the **errlog** command:

   ```
   40D97644   1205135007 I H ent4           BECOME BACKUP
   ```

   On Virtual I/O Server 2, the **entstat** command output should look as follows:

   ```
   $ entstat -all ent4 | grep Active
     Priority: 2  Active: True
   ```

   You should see the following entry in the errorlog when you issue the **errlog** command:

   ```
   E136EAFA   1205135007 I H ent4           BECOME PRIMARY
   ```

   Tip: You might experience up to a 30-second delay in failover when you use SEA Failover. The behavior depends in the network switch and the spanning tree settings. Consider enabling PortFast on your network switches, where possible, to reduce this delay.

5. On Virtual I/O Server 1, switch back to the primary adapter and verify that the primary adapter is active by using these commands:

```
$ chdev -dev ent4 -attr ha_mode=auto
ent4 changed
$ entstat -all ent4 | grep Active
  Priority: 1  Active: True
```

6. Unplug the link of the physical adapter on Virtual I/O Server 1. Use the **entstat** command to check whether the SEA has failed over to the standby adapter.
7. Replug the link of the physical adapter on Virtual I/O Server 1 and verify that the SEA has switched back to the primary.

Testing NIB

Follow these steps to verify that the Network Interface Backup (NIB) or Etherchannel configuration works as expected for an AIX virtual I/O client partition.

Consideration: NIB and Etherchannel are not available on IBM i. A similar concept on IBM i that uses virtual IP address (VIPA) failover is not supported for use with virtual Ethernet adapters.

These steps demonstrate how failover works when the network connection of a Virtual I/O Server is disconnected. The failover works in the same fashion when a Virtual I/O Server is rebooted. You can test this process by rebooting Virtual I/O Server 1 instead of unplugging the network cable in step 2.

1. Perform a remote login to the client partition by using telnet or SSH.
1. Verify whether the primary channel that is connected to Virtual I/O Server 1 is active by using the **entstat** command (Example 8-18).

Example 8-18 Verifying the active channel in an Etherchannel

```
# entstat -d ent2 | grep Active
Active channel: primary channel
```

2. Unplug the network cable from the physical network adapter that is connected to Virtual I/O Server 1.

3. As soon as the Etherchannel notices that it has lost its connection, it switches to the backup adapter. The message as shown in Example 8-19 is displayed in the error log.

Important: Do not disconnect your telnet or SSH connection. If it is disconnected, your configuration will fail the High Availability criterion.

Example 8-19 Errorlog message when the primary channel fails

```
LABEL:          ECH_PING_FAIL_PRMRY
IDENTIFIER:     9F7B0FA6

Date/Time:       Fri Oct 17 19:53:35 CST 2008
Sequence Number: 141
Machine Id:      00C1F1704C00
Node Id:         NIM_server
Class:           H
Type:            INFO
WPAR:            Global
Resource Name:   ent2
Resource Class:  adapter
Resource Type:   ibm_ech
Location:

Description
PING TO REMOTE HOST FAILED

Probable Causes
CABLE
SWITCH
ADAPTER

Failure Causes
CABLES AND CONNECTIONS

        Recommended Actions
        CHECK CABLE AND ITS CONNECTIONS
        IF ERROR PERSISTS, REPLACE ADAPTER CARD.

Detail Data
FAILING ADAPTER
PRIMARY
SWITCHING TO ADAPTER
ent1
Unable to reach remote host through primary adapter: switching over to
backup adapter
```

As shown in Example 8-20, the **entstat** command also shows that the backup adapter is active.

Example 8-20 Verifying the active channel in an Etherchannel

```
# entstat -d ent2 | grep Active
Active channel: backup adapter
```

4. Reconnect the physical adapter in Virtual I/O Server 1.
5. The Etherchannel does not automatically switch back to the primary channel. The SMIT menu contains the option `Automatically Recover to Main Channel`. This is set to `Yes` by default, which is the behavior when using physical adapters.

   However, virtual adapters do not adhere to this. Instead, the backup channel is used until it fails, at which time the system switches back to the primary channel. Manually switch back using the **/usr/lib/methods/ethchan_config -f** command as shown in Example 8-21. A message is displayed in the error log when the Etherchannel recovers to the primary channel.

Example 8-21 Manually switching to primary channel using entstat

```
# /usr/lib/methods/ethchan_config -f ent2
#  entstat -d ent2 | grep Active
Active channel: primary channel
# errpt
8650BE3F   1123195807 I H ent2           Etherchannel
RECOVERY
```

Tip: You can use the **dsh** command to automate the check or switch back if you have many client partitions. Alternatively, you can use the script in Appendix A, “AIX disk and NIB network checking and recovery script” on page 699.

Testing the bonding device configuration on Linux

Complete these steps to verify that the bonding device configuration on Linux is working as expected. Example 8-22 shows how to check for the link failure count in an interface.

Example 8-22 Checking for the link failure count

```
# cat /proc/net/bonding/bond0
Ethernet Channel Bonding Driver: v2.6.3-rh (June 8, 2005)

Bonding Mode: fault-tolerance (active-backup)
Primary Slave: None
Currently Active Slave: eth0
```

```
MII Status: up
MII Polling Interval (ms): 0
Up Delay (ms): 0
Down Delay (ms): 0

Slave Interface: eth0
MII Status: up
Link Failure Count: 0
Permanent HW addr: ba:d3:f0:00:40:02

Slave Interface: eth1
MII Status: up
Link Failure Count: 0

Permanent HW addr: ba:d3:f0:00:40:03
```

8.2.2 Virtual I/O Server networking monitoring

This section presents a monitoring scenario in which the intention is to identify which adapter from a Link Aggregation (or Etherchannel) was used to transfer data. A certain amount of data was transferred through FTP from a server through the Virtual I/O Server, and Link Aggregation was used as an SEA backing device.

Describing the scenario

The scenario to be analyzed is presented in Figure 8-13. In this scenario, there is a Virtual I/O Server, a logical partition, and an external Linux server. Data is transferred from the Linux server to the logical partition through the Virtual I/O Server using FTP. The interface that is used to transfer this data on the Virtual I/O Server will be identified.

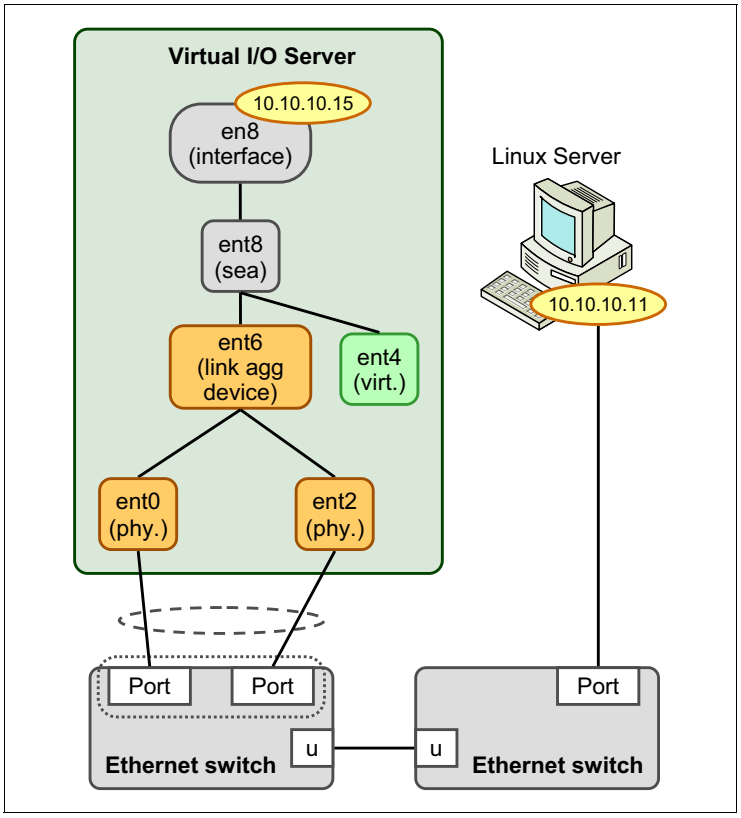


Figure 8-13 Network monitoring testing scenario

To set up the environment, the Virtual I/O Server was connected to a switch that supports Link Aggregation. The example uses a 4-port Ethernet card where port0 and port2 were connected to the switch. These ports were recognized as ent0 and ent2 on the Virtual I/O Server.

The Link Aggregation device (ent6) was created by using this command:

```
$ mkvdev -lnagg ent0,ent2 -attr mode=8023ad hash_mode=src_dsc_port
ent6 Available
```

When operating under IEEE 802.3ad mode, as defined by the **mode=8023ad** flag, the **hash_mode** attribute determines how the outgoing adapter for each packet is chosen. In this case, the `src_dsc_port` was chosen, which means that both the source and the destination TCP or UDP ports are used for that connection to determine the outgoing adapter.

The hash_mode attribute was introduced in IY45289 (devices.common.IBM.ethernet.ret 5.2.0.13). The hash attribute can also be set to the default, src_port and dst_port. Use this attribute to define that HTTP traffic goes through one specific adapter and FTP traffic goes through another adapter.

After the Link Aggregation device is created, the Shared Ethernet Adapter can be configured. Create an SEA by using the following command:

```
$ mkvdev -sea ent6 -vadapter ent4 -default ent4 -defaultid 1
ent8 Available
```

Next, configure the IP address on the SEA by running the following command:

```
$ mktcpip  -hostname 'VIO_Server1' -inetaddr '10.10.10.15' -netmask '255.0.0.0'
-interface 'en8'
```

The Virtual I/O Server is now ready to run file transfer tests.

Before you start the transfer tests, reset all the statistics for the adapters on the Virtual I/O Server. One way to do that is by using a simple “for” loop:

```
$ for i in 0 1 2 3 4 5 6 7 8
> do
> entstat -reset ent$i
> done
```

Check the statistics of adapter ent8 for the first time as shown in Example 8-23.

Example 8-23 Output of entstat on SEA

```
$ entstat ent8
-------------------------------------------------------------
ETHERNET STATISTICS (ent8) :
Device Type: Shared Ethernet Adapter
Hardware Address: 00:11:25:cc:80:38
Elapsed Time: 0 days 0 hours 0 minutes 10 seconds

Transmit Statistics:                          Receive Statistics:
--------------------                          -------------------
Packets: 9                                    Packets: 10
Bytes: 788                                    Bytes: 830
Interrupts: 0                                 Interrupts: 10
Transmit Errors: 0                            Receive Errors: 0
Packets Dropped: 0                            Packets Dropped: 0
```

```
                                                Bad Packets: 0
Max Packets on S/W Transmit Queue: 1
S/W Transmit Queue Overflow: 0
Current S/W+H/W Transmit Queue Length: 1

Elapsed Time: 0 days 0 hours 0 minutes 0 seconds
Broadcast Packets: 1                            Broadcast Packets: 1
Multicast Packets: 9                            Multicast Packets: 9
No Carrier Sense: 0                             CRC Errors: 0
DMA Underrun: 0                                 DMA Overrun: 0
Lost CTS Errors: 0                              Alignment Errors: 0
Max Collision Errors: 0                         No Resource Errors: 0
Late Collision Errors: 0                        Receive Collision Errors: 0
Deferred: 0                                     Packet Too Short Errors: 0
SQE Test: 0                                     Packet Too Long Errors: 0
Timeout Errors: 0                               Packets Discarded by Adapter: 0
Single Collision Count: 0                       Receiver Start Count: 0
Multiple Collision Count: 0
Current HW Transmit Queue Length: 1

General Statistics:
-------------------
No mbuf Errors: 0
Adapter Reset Count: 0
Adapter Data Rate: 0
Driver Flags: Up Broadcast Running
        Simplex 64BitSupport ChecksumOffload
        DataRateSet
```

The **entstat -all** command can be used to provide the information about ent8 and all the adapters that are integrated to it as shown in Example 8-24.

Example 8-24 entstat -all command on SEA

```
$ entstat -all ent8 |grep  -E "Packets:|ETHERNET"
ETHERNET STATISTICS (ent8) :
Packets: 111                                    Packets: 101
                                                Bad Packets: 0
Broadcast Packets: 8                            Broadcast Packets: 8
Multicast Packets: 103                          Multicast Packets: 93
ETHERNET STATISTICS (ent6) :
Packets: 18                                     Packets: 93
                                                Bad Packets: 0
Broadcast Packets: 8                            Broadcast Packets: 0
Multicast Packets: 10                           Multicast Packets: 93
ETHERNET STATISTICS (ent0) :
Packets: 5                                      Packets: 87
                                                Bad Packets: 0
Broadcast Packets: 0                            Broadcast Packets: 0
```

```
Multicast Packets: 5                          Multicast Packets: 87
ETHERNET STATISTICS (ent2) :
Packets: 13                                   Packets: 6
                                              Bad Packets: 0
Broadcast Packets: 8                          Broadcast Packets: 0
Multicast Packets: 5                          Multicast Packets: 6
ETHERNET STATISTICS (ent4) :
Packets: 93                                   Packets: 8
                                              Bad Packets: 0
Broadcast Packets: 0                          Broadcast Packets: 8
Multicast Packets: 93                         Multicast Packets: 0
Invalid VLAN ID Packets: 0
Switch ID: ETHERNET0
```

Tip: Because the Link Aggregation Control Protocol (LACP) is being used in this example, you can see packets flowing as the switch and the logical partitions negotiate configurations.

You can see the statistics of the SEA (ent8), the Link Aggregation device (ent6), the physical devices (ent0 and ent2), and the virtual Ethernet adapter (ent4).

Now, initiate the first data transfer and check the statistics. Log in to the Linux server box and run an FTP to the Virtual I/O Server by running these commands:

```
ftp> put "| dd if=/dev/zero bs=1M count=100" /dev/zero
local: | dd if=/dev/zero bs=1M count=100 remote: /dev/zero
229 Entering Extended Passive Mode (|||32851|)
150 Opening data connection for /dev/zero.
100+0 records in
100+0 records out
104857600 bytes (105 MB) copied, 8.85929 seconds, 11.8 MB/s
226 Transfer complete.
104857600 bytes sent in 00:08 (11.28 MB/s)
```

This operation transferred 100 MB from the Linux server to the Virtual I/O Server. No file was transferred because the **dd** command was used to create 100 packets of 1 MB each, fill them up with zeros, and transfer them to `/dev/zero` on the Virtual I/O Server.

You can check which adapter was used to transfer the data by reusing the **enstat** command, as shown in Example 8-25. Compared to the number of packets that are shown in Example 8-24 on page 203, you can see that the number increased after the first file transfer.

Example 8-25 entstat -all command after file transfer attempt 1

```
$ entstat -all ent8 |grep  -E "Packets:|ETHERNET"
ETHERNET STATISTICS (ent8) :
```

```
Packets: 35485                          Packets: 74936
                                        Bad Packets: 0
Broadcast Packets: 23                   Broadcast Packets: 23
Multicast Packets: 238                  Multicast Packets: 214
ETHERNET STATISTICS (ent6) :
Packets: 35270                          Packets: 74914
                                        Bad Packets: 0
Broadcast Packets: 22                   Broadcast Packets: 1
Multicast Packets: 24                   Multicast Packets: 214
ETHERNET STATISTICS (ent0) :
Packets: 14                             Packets: 74901
                                        Bad Packets: 0
Broadcast Packets: 0                    Broadcast Packets: 1
Multicast Packets: 12                   Multicast Packets: 201
ETHERNET STATISTICS (ent2) :
Packets: 35256                          Packets: 13
                                        Bad Packets: 0
Broadcast Packets: 22                   Broadcast Packets: 0
Multicast Packets: 12                   Multicast Packets: 13
ETHERNET STATISTICS (ent4) :
Packets: 215                            Packets: 22
                                        Bad Packets: 0
Broadcast Packets: 1                    Broadcast Packets: 22
Multicast Packets: 214                  Multicast Packets: 0
Invalid VLAN ID Packets: 0
Switch ID: ETHERNET0
```

Note the packet count for the ent2 interface is now 35256. This means that physical adapter ent2 was chosen to transfer the file.

Go back to the FTP session, transfer another 100 MB, and check again for the adapter statistics.

The FTP session now look like this:

```
ftp> put "| dd if=/dev/zero bs=1M count=100" /dev/zero
local: | dd if=/dev/zero bs=1M count=100 remote: /dev/zero
229 Entering Extended Passive Mode (|||32855|)
150 Opening data connection for /dev/zero.
100+0 records in
100+0 records out
104857600 bytes (105 MB) copied, 8.84978 seconds, 11.8 MB/s
226 Transfer complete.
104857600 bytes sent in 00:08 (11.29 MB/s)
ftp> quit
221 Goodbye.
```

Check the adapter statistics again as shown in Example 8-26.

Example 8-26 entstat -all command after file transfer attempt 2

```
$ entstat -all ent8 |grep  -E "Packets:|ETHERNET"
ETHERNET STATISTICS (ent8) :
Packets: 70767                                Packets: 149681
                                              Bad Packets: 0
Broadcast Packets: 37                         Broadcast Packets: 37
Multicast Packets: 294                        Multicast Packets: 264
ETHERNET STATISTICS (ent6) :
Packets: 70502                                Packets: 149645
                                              Bad Packets: 0
Broadcast Packets: 36                         Broadcast Packets: 1
Multicast Packets: 30                         Multicast Packets: 264
ETHERNET STATISTICS (ent0) :
Packets: 17                                   Packets: 149629
                                              Bad Packets: 0
Broadcast Packets: 0                          Broadcast Packets: 1
Multicast Packets: 15                         Multicast Packets: 248
ETHERNET STATISTICS (ent2) :
Packets: 70485                                Packets: 16
                                              Bad Packets: 0
Broadcast Packets: 36                         Broadcast Packets: 0
Multicast Packets: 15                         Multicast Packets: 16
ETHERNET STATISTICS (ent4) :
Packets: 265                                  Packets: 36
                                              Bad Packets: 0
Broadcast Packets: 1                          Broadcast Packets: 36
Multicast Packets: 264                        Multicast Packets: 0
Invalid VLAN ID Packets: 0
Switch ID: ETHERNET0
```

The ent2 interface was used again, and has increased to 70485.

Open a new FTP session to the Virtual I/O Server, transfer more data, and verify which interface was used.

On the Linux server, open a new FTP session and transfer the data:

```
ftp> put "| dd if=/dev/zero bs=1M count=100" /dev/zero
local: | dd if=/dev/zero bs=1M count=100 remote: /dev/zero
229 Entering Extended Passive Mode (|||32858|)
150 Opening data connection for /dev/zero.
100+0 records in
100+0 records out
104857600 bytes (105 MB) copied, 8.85152 seconds, 11.8 MB/s
226 Transfer complete.
104857600 bytes sent in 00:08 (11.28 MB/s)
```

On the Virtual I/O Server, check the interface statistics to identify which interface was used to transfer the data this time, as shown in Example 8-27.

Example 8-27 entstat -all command after file transfer attempt 3

```
$ entstat -all ent8 |grep  -E "Packets:|ETHERNET"
ETHERNET STATISTICS (ent8) :
Packets: 106003                                 Packets: 224403
                                                Bad Packets: 0
Broadcast Packets: 44                           Broadcast Packets: 44
Multicast Packets: 319                          Multicast Packets: 287
ETHERNET STATISTICS (ent6) :
Packets: 105715                                 Packets: 224360
                                                Bad Packets: 0
Broadcast Packets: 43                           Broadcast Packets: 1
Multicast Packets: 32                           Multicast Packets: 287
ETHERNET STATISTICS (ent0) :
Packets: 35219                                  Packets: 224343
                                                Bad Packets: 0
Broadcast Packets: 0                            Broadcast Packets: 1
Multicast Packets: 16                           Multicast Packets: 270
ETHERNET STATISTICS (ent2) :
Packets: 70496                                  Packets: 17
                                                Bad Packets: 0
Broadcast Packets: 43                           Broadcast Packets: 0
Multicast Packets: 16                           Multicast Packets: 17
ETHERNET STATISTICS (ent4) :
Packets: 288                                    Packets: 43
                                                Bad Packets: 0
Broadcast Packets: 1                            Broadcast Packets: 43
Multicast Packets: 287                          Multicast Packets: 0
Invalid VLAN ID Packets: 0
Switch ID: ETHERNET0
```

This time, the ent0 adapter was used to transfer the data.

If you open multiple sessions from the Linux server to the Virtual I/O Server, divide the traffic between the adapters because each FTP session uses a separate port number.

Open two FTP connections to the Virtual I/O Server and verify their usage.

First, reset all the statistics of the adapters as shown here:

```
$ for i in 1 2 3 4 5 6 7 8
> do
> entstat -reset ent$i
> done
```

The adapter statistics should be similar to those shown in Example 8-28.

Example 8-28 entstat -all command after reset of Ethernet adapters

```
$ entstat -all ent8 |grep  -E "Packets:|ETHERNET"
ETHERNET STATISTICS (ent8) :
Packets: 1                                  Packets: 1
                                            Bad Packets: 0
Broadcast Packets: 0                        Broadcast Packets: 0
Multicast Packets: 1                        Multicast Packets: 1
ETHERNET STATISTICS (ent6) :
Packets: 0                                  Packets: 1
                                            Bad Packets: 0
Broadcast Packets: 0                        Broadcast Packets: 0
Multicast Packets: 0                        Multicast Packets: 1
ETHERNET STATISTICS (ent0) :
Packets: 0                                  Packets: 1
                                            Bad Packets: 0
Broadcast Packets: 0                        Broadcast Packets: 0
Multicast Packets: 0                        Multicast Packets: 1
ETHERNET STATISTICS (ent2) :
Packets: 0                                  Packets: 0
                                            Bad Packets: 0
Broadcast Packets: 0                        Broadcast Packets: 0
Multicast Packets: 0                        Multicast Packets: 0
ETHERNET STATISTICS (ent4) :
Packets: 1                                  Packets: 0
                                            Bad Packets: 0
Broadcast Packets: 0                        Broadcast Packets: 0
Multicast Packets: 1                        Multicast Packets: 0
Invalid VLAN ID Packets: 0
Switch ID: ETHERNET0
```

On the first terminal, open an FTP session from the Linux server to the Virtual I/O Server. Use a larger amount of data to give yourself time to open a second FTP session:

```
server1:~ # ftp 10.10.10.15
Connected to 10.10.10.15.
220 VIO_Server1 FTP server (Version 4.2 Fri Oct 17 07:20:05 CDT 2008) ready.
Name (10.10.10.15:root): padmin
331 Password required for padmin.
Password:
230-Last unsuccessful login: Thu Oct 16 20:26:56 CST 2008 on ftp from
::ffff:10.10.10.11
230-Last login: Thu Oct 16 20:27:02 CST 2008 on ftp from ::ffff:10.10.10.11
230 User padmin logged in.
Remote system type is UNIX.
Using binary mode to transfer files.
```

```
ftp> put "| dd if=/dev/zero bs=1M count=1000" /dev/zero
local: | dd if=/dev/zero bs=1M count=1000 remote: /dev/zero
229 Entering Extended Passive Mode (|||33038|)
150 Opening data connection for /dev/zero.
```

Check the statistics of the adapter as shown in Example 8-29.

Example 8-29 entstat -all command after opening one FTP session

```
$ entstat -all ent8 |grep  -E "Packets:|ETHERNET"
ETHERNET STATISTICS (ent8) :
Packets: 41261                                 Packets: 87496
                                               Bad Packets: 0
Broadcast Packets: 11                          Broadcast Packets: 11
Multicast Packets: 38                          Multicast Packets: 34
ETHERNET STATISTICS (ent6) :
Packets: 41241                                 Packets: 87521
                                               Bad Packets: 0
Broadcast Packets: 11                          Broadcast Packets: 0
Multicast Packets: 4                           Multicast Packets: 34
ETHERNET STATISTICS (ent0) :
Packets: 41235                                 Packets: 87561
                                               Bad Packets: 0
Broadcast Packets: 0                           Broadcast Packets: 0
Multicast Packets: 2                           Multicast Packets: 32
ETHERNET STATISTICS (ent2) :
Packets: 21                                    Packets: 2
                                               Bad Packets: 0
Broadcast Packets: 11                          Broadcast Packets: 0
Multicast Packets: 2                           Multicast Packets: 2
ETHERNET STATISTICS (ent4) :
Packets: 34                                    Packets: 11
                                               Bad Packets: 0
Broadcast Packets: 0                           Broadcast Packets: 11
Multicast Packets: 34                          Multicast Packets: 0
Invalid VLAN ID Packets: 0
Switch ID: ETHERNET0
```

Notice that the first FTP session is using the physical adapter ent0 to transfer the data. Now open a second terminal and a new FTP session to the Virtual I/O Server:

```
server1:~ # ftp 10.10.10.15
Connected to 10.10.10.15.
220 VIO_Server1 FTP server (Version 4.2 Fri Oct 17 07:20:05 CDT 2008) ready.
Name (10.10.10.15:root): padmin
331 Password required for padmin.
Password:
```

```
230-Last unsuccessful login: Thu Oct 16 20:26:56 CST 2008 on ftp from
::ffff:10.10.10.11
230-Last login: Thu Oct 16 20:29:57 CST 2008 on ftp from ::ffff:10.10.10.11
230 User padmin logged in.
Remote system type is UNIX.
Using binary mode to transfer files.
ftp> put "| dd if=/dev/zero bs=1M count=1000" /dev/null
local: | dd if=/dev/zero bs=1M count=1000 remote: /dev/null
229 Entering Extended Passive Mode (|||33041|)
150 Opening data connection for /dev/null.
1000+0 records in
1000+0 records out
1048576000 bytes (1.0 GB) copied, 154.686 seconds, 6.8 MB/s
226 Transfer complete.
1048576000 bytes sent in 02:34 (6.46 MB/s)
```

Remember: In the second FTP session, the device `/dev/null` was used because the device `/dev/zero` was already used by the first FTP session.

Now both FTP transfers should be completed. Check the adapter statistics again as shown in Example 8-30.

Example 8-30 entstat -all command after opening two FTP sessions

```
$ entstat -all ent8 |grep  -E "Packets:|ETHERNET"
ETHERNET STATISTICS (ent8) :
Packets: 704780                                 Packets: 1493888
                                                Bad Packets: 0
Broadcast Packets: 108                          Broadcast Packets: 108
Multicast Packets: 437                          Multicast Packets: 391
ETHERNET STATISTICS (ent6) :
Packets: 704389                                 Packets: 1493780
                                                Bad Packets: 0
Broadcast Packets: 108                          Broadcast Packets: 0
Multicast Packets: 46                           Multicast Packets: 391
ETHERNET STATISTICS (ent0) :
Packets: 352118                                 Packets: 1493757
                                                Bad Packets: 0
Broadcast Packets: 0                            Broadcast Packets: 0
Multicast Packets: 23                           Multicast Packets: 368
ETHERNET STATISTICS (ent2) :
Packets: 352271                                 Packets: 23
                                                Bad Packets: 0
Broadcast Packets: 108                          Broadcast Packets: 0
Multicast Packets: 23                           Multicast Packets: 23
ETHERNET STATISTICS (ent4) :
Packets: 391                                    Packets: 108
                                                Bad Packets: 0
```

```
Broadcast Packets: 0                        Broadcast Packets: 108
Multicast Packets: 391                      Multicast Packets: 0
Invalid VLAN ID Packets: 0
Switch ID: ETHERNET0
```

Both adapters were used to transfer approximately the same amount of data. This shows that the traffic of the FTP sessions was spread across both adapters.

This example illustrates how network utilization can be monitored and improved by tuning a parameter on the Shared Ethernet Adapter. It can also be used as the basis for new configurations and monitoring.

Advanced SEA monitoring

The advanced tool **seastat** can be used on a Virtual I/O Server to track the number of packets and bytes received by and sent from each MAC address. It can also monitor traffic for the Virtual I/O Server and individual virtual I/O clients. Users can monitor the statistics based on a number of parameters that include MAC address, IP address, and host name.

Using seastat

To use **seastat**, you must first enable it as shown in Example 8-31. In this example, ent5 is an SEA.

The **seastat** command has the following format:

```
seastat -d <device_name> -c [-n | -s search_criterion=value]
```

Where:

`<device_name>`	The shared adapter device whose statistics is sought
`-c`	Clears all per-client SEA statistics
`-n`	Displays name resolution on the IP addresses

Example 8-31 Enabling advanced SEA monitoring

```
$ seastat -d ent5
Device ent5 has accounting disabled

$ lsdev -dev ent5 -attr
accounting    disabled Enable per-client accounting of network statistics              True
ctl_chan      ent3     Control Channel adapter for SEA failover                        True
gvrp          no       Enable GARP VLAN Registration Protocol (GVRP)                   True
ha_mode       auto     High Availability Mode                                          True
jumbo_frames  no       Enable Gigabit Ethernet Jumbo Frames                            True
large_receive no       Enable receive TCP segment aggregation                          True
largesend     0        Enable Hardware Transmit TCP Resegmentation                     True
```

```
netaddr       0        Address to ping                                                 True
pvid          1        PVID to use for the SEA device                                  True
pvid_adapter  ent2     Default virtual adapter to use for non-VLAN-tagged packets      True
_mode      disabled N/A                                                              True
real_adapter  ent0     Physical adapter associated with the SEA                        True
thread        1        Thread mode enabled (1) or disabled (0)                         True
virt_adapters ent2     List of virtual adapters associated with the SEA (comma separated) True

$ chdev -dev ent5  -attr accounting=enabled
ent5 changed

$ lsdev -dev ent5 -attr
accounting    enabled  Enable per-client accounting of network statistics             True
ctl_chan      ent3     Control Channel adapter for SEA failover                       True
gvrp          no       Enable GARP VLAN Registration Protocol (GVRP)                  True
ha_mode       auto     High Availability Mode                                         True
jumbo_frames  no       Enable Gigabit Ethernet Jumbo Frames                           True
large_receive no       Enable receive TCP segment aggregation                         True
largesend     0        Enable Hardware Transmit TCP Resegmentation                    True
netaddr       0        Address to ping                                                True
pvid          1        PVID to use for the SEA device                                 True
pvid_adapter  ent2     Default virtual adapter to use for non-VLAN-tagged packets     True
_mode      disabled N/A                                                              True
real_adapter  ent0     Physical adapter associated with the SEA                       True
thread        1        Thread mode enabled (1) or disabled (0)                        True
virt_adapters ent2     List of virtual adapters associated with the SEA (comma separated) True
```

Example 8-32 shows SEA statistics without any search criterion. The output displays statistics for all clients that this Virtual I/O Server is serving.

Example 8-32 Sample seastat statistics

```
$ seastat -d ent5
================================================================================
Advanced Statistics for SEA
Device Name: ent5
================================================================================
MAC: 6A:88:82:AA:9B:02
----------------------
VLAN: None
VLAN Priority: None
Transmit Statistics:                    Receive Statistics:
--------------------                    -------------------
Packets: 7                              Packets: 2752
Bytes: 420                              Bytes: 185869
================================================================================
MAC: 6A:88:82:AA:9B:02
----------------------
```

```
VLAN: None
VLAN Priority: None
IP: 9.3.5.115
Transmit Statistics:                  Receive Statistics:
--------------------                  -------------------
Packets: 125                          Packets: 3260
Bytes: 117242                         Bytes: 228575
================================================================================
MAC: 6A:88:85:BF:16:02
----------------------
VLAN: None
VLAN Priority: None
Transmit Statistics:                  Receive Statistics:
--------------------                  -------------------
Packets: 1                            Packets: 1792
Bytes: 42                             Bytes: 121443
================================================================================
MAC: 6A:88:86:26:F1:02
----------------------
VLAN: None
VLAN Priority: None
IP: 9.3.5.119
Transmit Statistics:                  Receive Statistics:
--------------------                  -------------------
Packets: 2                            Packets: 1573
Bytes: 190                            Bytes: 107535
================================================================================
MAC: 6A:88:86:26:F1:02
----------------------
VLAN: None
VLAN Priority: None
Transmit Statistics:                  Receive Statistics:
--------------------                  -------------------
Packets: 1                            Packets: 2575
Bytes: 42                             Bytes: 173561
================================================================================
MAC: 6A:88:8D:E7:80:0D
----------------------
VLAN: None
VLAN Priority: None
Hostname: vios1
IP: 9.3.5.111
Transmit Statistics:                  Receive Statistics:
--------------------                  -------------------
Packets: 747                          Packets: 3841
Bytes: 364199                         Bytes: 327541
================================================================================
```

```
MAC: 6A:88:8D:E7:80:0D
----------------------
VLAN: None
VLAN Priority: None
Transmit Statistics:                    Receive Statistics:
--------------------                    -------------------
Packets: 10                             Packets: 3166
Bytes: 600                              Bytes: 214863
================================================================================
MAC: 6A:88:8F:36:34:02
----------------------
VLAN: None
VLAN Priority: None
Transmit Statistics:                    Receive Statistics:
--------------------                    -------------------
Packets: 9                              Packets: 3149
Bytes: 540                              Bytes: 213843
================================================================================
MAC: 6A:88:8F:36:34:02
----------------------
VLAN: None
VLAN Priority: None
IP: 9.3.5.121
Transmit Statistics:                    Receive Statistics:
--------------------                    -------------------
Packets: 125                            Packets: 3256
Bytes: 117242                           Bytes: 229103
================================================================================
MAC: 6A:88:8F:ED:33:0D
----------------------
VLAN: None
VLAN Priority: None
Transmit Statistics:                    Receive Statistics:
--------------------                    -------------------
Packets: 10                             Packets: 3189
Bytes: 600                              Bytes: 216243
================================================================================
MAC: 6A:88:8F:ED:33:0D
----------------------
VLAN: None
VLAN Priority: None
IP: 9.3.5.112
Transmit Statistics:                    Receive Statistics:
--------------------                    -------------------
Packets: 330                            Packets: 3641
```

```
Bytes: 194098                          Bytes: 309419
================================================================================
```

This command shows an entry for each pair of VLANs, VLAN priority, IP address, and MAC address. In Example 8-32 on page 212, there are two entries for several MAC addresses. One entry is for MAC address and the other one is for the IP address that is configured for that MAC address.

Statistics based on search criterion

The **seastat** tool can also print statistics based on search criteria. Currently the following search criteria are supported:

- MAC address (mac)
- Priority
- VLAN ID (vlan)
- IP address (ip)
- Hostname (host)
- Greater than bytes sent (gbs)
- Greater than bytes recv (gbr)
- Greater than packets sent (gps)
- Greater than packets recv (gpr)
- Smaller than bytes sent (sbs)
- Smaller than bytes recv (sbr)
- Smaller than packets sent (sps)
- Smaller than packets recv (spr)

To use a search criterion, you must specify it in the following format:

<search_criteria>=<value>

Example 8-33 shows statistics based on the search criterion IP address, which is specified as `ip=9.3.5.121`.

Example 8-33 seastat statistics using search criterion

```
$ seastat -d ent5 -n
================================================================================
Advanced Statistics for SEA
Device Name: ent5
================================================================================
MAC: 6A:88:8D:E7:80:0D
----------------------
VLAN: None
VLAN Priority: None
IP: 9.3.5.111
Transmit Statistics:                    Receive Statistics:
--------------------                    -------------------
```

```
Packets: 13                             Packets: 81
Bytes: 2065                             Bytes: 5390
=============================================================================
MAC: 6A:88:8F:36:34:02
---------------------
VLAN: None
VLAN Priority: None
IP: 9.3.5.121
Transmit Statistics:                    Receive Statistics:
--------------------                    -------------------
Packets: 1                              Packets: 23
Bytes: 130                              Bytes: 1666
=============================================================================
$ seastat -d ent5 -s  ip=9.3.5.121
=============================================================================
Advanced Statistics for SEA
Device Name: ent5
=============================================================================
MAC: 6A:88:8F:36:34:02
---------------------
VLAN: None
VLAN Priority: None
IP: 9.3.5.121
Transmit Statistics:                    Receive Statistics:
--------------------                    -------------------
Packets: 115                            Packets: 8542
Bytes: 14278                            Bytes: 588424
=============================================================================
```

Using topas

In Version 2.1 of Virtual I/O Server, the **topas** performance monitoring tool was enhanced to provide visibility of the performance of Shared Ethernet Adapters.

The SEA monitor is available by pressing E after **topas** is running as shown in Example 8-34.

Example 8-34 topas Shared Ethernet Adapter monitor

```
Topas Monitor for host:    P7_1_vios1  Interval:   2   Wed Dec 15 10:09:13 2010
================================================================================
Network                               KBPS   I-Pack   O-Pack    KB-In   KB-Out
ent6 (SEA PRIM)                       38.7      5.0     29.0      1.8     36.9
  |\--ent0 (PHYS)                     19.6      4.0     14.0      1.7     17.9
  |\--ent5 (VETH)                     19.2      1.0     15.0      0.1     19.0
   \--ent4 (VETH CTRL)                0.1      0.0      3.5      0.0      0.1
lo0                                   2.7     14.0     14.0      1.3      1.3
```

For this tool to work on a Shared Ethernet Adapter, the state of the layer-3 device (`en`) cannot be in the `defined` state. If you are not using the layer-3 device on the SEA, the easiest way to change the state of the device is to change one of its parameters. The following command will change the state of a Shared Ethernet Adapter's layer-3 device without affecting bridging:

```
chdev -l <sea_en_device> -a state=down
```

Another useful command is `topas -cecdisp`, which shows statistics from Virtual I/O servers and their Virtual I/O clients.

8.2.3 AIX client network monitoring

On the AIX virtual I/O client, you can use the **`entstat`** command to monitor a virtual Ethernet adapter as shown in the preceding examples. It can also be used to monitor a physical Ethernet adapter. **topas** (press n for network) and **nmon** (press n for network) are other options for monitoring virtual Ethernet adapters, as shown in Example 8-35.

Example 8-35 AIX nmon network monitoring

```
··topas_nmon··#=PURR Stats·····Host=p750_lpar01····Refresh=2 secs···17:11.51··
· Network ·····································································
·I/F Name Recv=KB/s Trans=KB/s packin packout insize outsize Peak->Recv TransKB·
·    en0       0.1       0.1      1.5      0.5   46.0  266.0        0.1     0.5·
·    lo0       0.6       0.6      6.5      6.5   89.0   89.0        0.6     0.6·
·  Total       0.0       0.0 in Mbytes/second   Overflow=0                    ·
·I/F Name  MTU  ierror oerror collision Mbits/s Description                    ·
·    en0   1500      0      0      0  10240 Standard Ethernet Network Interface·
·    lo0  16896      0      0      0      0 Loopback Network Interface         ·
·····································································
·                                                                             ·
```

8.2.4 IBM i client network monitoring

This section explains how to monitor network health and performance on an IBM i virtual I/O client.

Checking network health on the IBM i virtual I/O client

To verify an active working Internet Protocol network configuration on an IBM i virtual I/O client, check the following items:

1. Verify the TCP/IP interface is in an `Active` state by running WRKTCPSTS *IFC as shown in Figure 8-14.

   If the interface is in an Inactive state, start it by using option **9=Start**. Otherwise, proceed with the next step.

```
Work with TCP/IP Interface Status
                                                          System:   P71I05
Type options, press Enter.
  5=Display details   8=Display associated routes   9=Start   10=End
  12=Work with configuration status   14=Display multicast groups

     Internet         Network             Line      Interface
Opt  Address          Address          Description  Status
     127.0.0.1        127.0.0.0        *LOOPBACK    Active
     172.16.21.108    172.16.20.0      ETHLIN001    Active
```

Figure 8-14 IBM i Work with TCP/IP Interface Status panel

2. The corresponding Ethernet line description is ACTIVE (varied on).

   Run WRKCFGSTS *LIN as shown in Figure 8-15.

   If the line description is VARIED OFF, vary it on by using option **1=Vary on**. Otherwise, proceed to the next step.

```
Work with Configuration Status
                                                          System:   P71I05
Position to  . . . . .               Starting characters
Opt  Description       Status                -------------Job--------------
     ETHLIN001         ACTIVE
       ETHLINET        ACTIVE
         ETHLITCP      ACTIVE                QTCPWRK     QSYS        002337
     QESLINE           VARIED OFF
     QTILINE           VARIED OFF
```

Figure 8-15 IBM i Work with Configuration Status panel

3. The corresponding virtual Ethernet adapter resource CMN*xx* (type 268C) is `Operational`.

   Run WRKHDWRSC *CMN as shown in Figure 8-16.

   If the virtual Ethernet adapter resource is in an `Inoperational` state, reset the virtual IOP resource for recovery.

   If it is in a `Not connected` state, check the IBM i and Virtual I/O Server virtual Ethernet adapter partition configuration.

```
Work with Communication Resources
                                                              System:   P71I05
Type options, press Enter.
  5=Work with configuration descriptions   7=Display resource detail

Opt  Resource        Type  Status               Text
     CMB05           268C  Operational          Combined function IOP
       LIN04         6B26  Operational          Comm Adapter
     CMB06           6B03  Operational          Comm Processor
       LIN02         6B03  Operational          Comm Adapter
         CMN03       6B03  Operational          Comm Port
     CMB07           6B03  Operational          Comm Processor
       LIN01         6B03  Operational          Comm Adapter
         CMN02       6B03  Operational          Comm Port
     CMB08           268C  Operational          Comm Processor
       LIN03         268C  Operational          LAN Adapter
         CMN01       268C  Operational          Ethernet Port
```

Figure 8-16 IBM i Work with Communication Resources panel

Monitoring network performance on an IBM i virtual I/O client

This section provides examples of IBM i network monitoring using IBM Performance Tools for IBM i.

Example 8-36 shows a System Report for TCP/IP Summary created with the following command:

```
PRTSYSRPT MBR(Q342145852) TYPE(*TCPIP)
```

Example 8-36 IBM i System Report for TCP/IP Summary

```
                                        System Report                                 120811   15:03:4
                                      TCP/IP Summary                                         Page 000
Member  . . . : Q342145852 Model/Serial  . : E8B/10-0EF5R       Main storage . . :   16.0 GB  Started  . . . . : 12/08/11 14:58:5
  Library . . : QPFRDATA   System name . . : P71I05            Version/Release  :   7/ 1.0   Stopped  . . . . : 12/08/11 15:03:3
Partition ID  : 005        Feature Code  . : EPA1-EPA1          Int Threshold  . :    .00 %
```

```
Virtual Processors:   4    Processor Units :   4.00
                MTU         KB      ------------- Packets Received --------------     KB      ----------- Packets Sent -----------
Line Type/      Size      Received                                    Number   Pct  Transmitted                                    Pct
  Line Name    (bytes)    /Second       Unicast       Non-Unicast    Error  Error   /Second       Unicast       Non-Unicast   Erro
------------ --------- ------------ ---------------- ---------------- ------- ----- ----------- ---------------- ---------------- ----
                   576
  *LOOPBACK                      0                0                0       0   .00           0                0                0   .0
ETHERNET         1,500
  ETHLIN001                  1,552          316,206               55       0   .00          55          296,988                2   .0
Line Type/Line Name     --  The type and name of the line description used by the interface.
MTU Size (bytes)        --  Maximum Transmission Unit (MTU) size in bytes for interface
KB Received/Second      --  Number of kilobytes (1024 bytes) received on interface per second
Unicast Packets Rcvd    --  Number of unicast packets received
Non-Unicast Packet Rcvd --  Number of non-unicast packets received
                                                                                                                          More...
```

Example 8-37 shows a Component Report for TCP/IP Activity that was created with the following command:

```
PRTCPTRPT MBR(Q342145852) TYPE(*TCPIP)
```

Example 8-37 IBM i resource report for disk utilization

```
                                                    Component Report                                                 12/08/11 15:00:5
                                                    TCP/IP Activity                                                         Page
Member  . . . : Q342145852 Model/Serial  . : E8B/10-0EF5R       Main storage . . :   16.0 GB  Started . . . . :  12/08/11 14:58:5
  Library . . : QPFRDATA   System name . . :P71I05              Version/Release  :    7/ 1.0  Stopped . . . . :  12/08/11 15:00:4
Partition ID  : 005        Feature Code  . :EPA1-EPA1           Int Threshold  . :    .00 %
Virtual Processors:   4    Processor Units :   4.00
System TCP/IP
      ----- Datagrams -----  Datagrams Requested  --- TCP Segments ---- ------- UDP Datagrams ------- ------- ICMP Messages ------
 Itv                   Pct  - for Transmission -- - per Second --  Pct                           Pct                           Pct
 End     Received     Error      Total      Dscrd  Rcvd    Sent   Rtrns  Received      Sent     Error  Received     Sent     Error
----- ---------- ----- ---------- ----- ------- ------- ----- ---------- ---------- ----- ---------- ---------- -----
14:59       22,488   .00     21,340   .00   3,746   3,555   .00          0          0   .00          6          6   .00
14:59       57,427   .00     50,222   .00   3,827   3,347   .00          4          0   .00         15         15   .00
14:59       68,742   .00     61,303   .00   4,581   4,085   .00          4          0   .00         15         15   .00
14:59       64,854   .00     63,516   .00   4,322   4,233   .00          0          1   .00         15         15   .00
15:00       68,764   .00     67,326   .00   4,583   4,487   .00          2          0   .00         15         15   .00
15:00       33,719   .00     33,047   .00   2,243   2,198   .00          2          0   .00         15         15   .00
15:00           23   .00         17   .00       0       0   .00          8          0   .00         15         15   .00
15:00           29   .00         29   .00       0       0   .00          0          0   .00         15         15   .00
                                                                                                                  More...
```

These system and component reports for TCP/IP can help you obtain an overview of IBM i network usage and performance by providing information about network I/O throughput and the percentage of packet errors.

For more information about IBM Performance Tools for IBM i, see *IBM eServer iSeries Performance Tools for iSeries*, SC41-5340.

Another approach for long-term network monitoring for IBM i that also allows IBM i cross-partition monitoring is to use the Navigator for i Management Central monitors function. You can define a threshold based on average network utilization to notify you about unusually high network utilization so you can address potential network performance problems.

For more information about using Navigator for i for performance monitoring, see *Managing OS/400 with Operations Navigator V5R1 Volume 5: Performance Management*, SG24-6565.

8.2.5 Linux network monitoring

Linux distributions include the **netstat** utility in the default installation. Using the **netstat** command with the **-p** option provides you with information about socket and process IDs. The output that you receive when using the **-p** option provides information about processes using network resources. Therefore, bandwidth usage can be collected by analyzing the throughput on a per-adapter basis from the **netstat** command run.

In addition to the **netstat** utility, you can use the **tcpdump** utility that is included in most distributions. You can use the **tcpdump** tool for network monitoring, protocol debugging, and data acquisition. The tool sees all network traffic, which can be used to create statistical monitoring scripts. A major drawback of **tcpdump** is the large size of the flat file that contains the text output. For more information about the availability of this tool in the current Linux release, see the distributor's documentation.

8.2.6 Tuning network throughput

This section looks at tuning the amount of data that can be transmitted through the network stack in a Power Systems server. The details are not specific to Power Systems servers. They are standard networking protocols and concepts.

In most cases in IP networking, you want to transmit the largest sized network payloads possible to maximize bandwidth and reduce protocol processor usage. There are numerous places where this setting can be tuned. This section covers information about jumbo frames, the maximum transfer unit (MTU), and maximum segment size (MSS). It also describes the path MTU discovery changes in AIX Version 5.3 and later, and other performance variables.

Operating system device configuration

All operating systems configure network devices in a different manner. Some differentiate between layers, whereas others do not. This section provides a reference for each operating system that runs on the Power platform.

AIX

AIX differentiates between layer-2 and layer-3 devices in these ways:

en Layer-3 device for Ethernet version 2. This is the most commonly used layer-3 interface.

et Layer-3 device for IEEE 802.3 Ethernet. Not used as often as the en device.

ent Layer-2 device.

IBM i

The IBM i operating system uses a single interface for both layers:

ETH Layer-2 and layer-3 device.

Linux

The Linux operating system uses a single interface for both layers:

eth Layer-2 and layer-3 device.

Tuning network payloads

Tuning network payloads can result in significant throughput increases especially in configurations that use network-attached storage (NAS).

For this section, it is important to understand the differentiation between jumbo frames and the maximum transmission unit. They are not the same thing, although they are closely related. Often the terms are used interchangeably, and it can be difficult to describe one in the absence of the other.

The following definitions are used in this chapter:

Jumbo frames Refers to the Ethernet payload size. Configured at layer-2 of the OSI networking model. Generally refers to a 9000-byte payload.

Maximum transfer unit The maximum size of the IP datagram. Configured at layer-3 of the OSI networking model.

Maximum segment size The maximum segment refers to the size of the payload of the Transmission Control Protocol (TCP) packet. Configured at layer-4 of the OSI networking model.

Jumbo frames

Traditional Ethernet specifications define a payload of 1500 bytes. This is the amount of data that can be delivered by every Ethernet frame on the wire. The term *jumbo frames* refers to increasing the payload area of the frame beyond the 1500-byte standard. Technically speaking, the term is accurate for any size

beyond 1500 bytes. However, manufacturers have generally agreed on a 9000-byte payload as a standard. Jumbo Frames on Power based systems use a 9000-byte payload.

In an Internet Protocol network, the IP datagram is encapsulated within the payload area of the Ethernet frame. The maximum size of the IP datagram, including all headers, data, and padding, is called the MTU. For more information about the MTU, see “Maximum transfer unit (MTU)” on page 223.

Enabling Jumbo Frames on a physical Ethernet adapter under AIX requires the `jumbo_frames` parameter on the layer-2 (`ent`) device to be modified.

It is important to point out that there is no attribute for jumbo frames on a virtual Ethernet adapter. Traffic through a virtual Ethernet adapter is handled by the Power hypervisor through memory buffers. This virtualized implementation does not interact with traditional layer-1 mediums such as cable. Therefore, many Ethernet specific attributes such as jumbo frames have limited relevance to a virtual adapter.

If a virtual Ethernet adapter must communicate with an external network, the Shared Ethernet Adapter on the Virtual I/O Server handles the framing of the traffic. Jumbo frames must be enabled on the SEA.

Shared Ethernet Adapter and jumbo frames

The primary purpose of the SEA is to bridge network communication between the virtual I/O clients and an external network. The SEA can bridging traffic that require jumbo frames, by enabling the `jumbo_frames` attribute on both the physical adapter layer-2 device and the Shared Ethernet Adapter layer-2 device.

Although the SEA can bridge jumbo frames, it cannot generate them from its own layer-3 interface. The layer-3 interfaces that are associated with the SEA are meant for administration traffic.

For more information about how to set the jumbo frames parameters, see “Payload tuning examples” on page 227.

Important: Before you enable jumbo frames on a physical adapter or SEA, ensure the other devices in your network (or VLAN) are also configured for jumbo frames. Errors will occur if your system attempts to transmit Ethernet frames into a layer 2 network that is not configured to handle them.

Maximum transfer unit (MTU)

The MTU value of the Internet Protocol denotes the maximum size of an IP datagram. Ideally the value of the MTU is one that fits within the payload area of the underlying layer-2 protocol, in this case Ethernet. If the MTU is greater than

the payload size of the layer-2 protocol, the IP datagram must be fragmented to be delivered.

The MTU size can affect the network performance between source and target systems. The use of large MTU sizes allows the operating system to send fewer packets of a larger size to reach the same network throughput. The larger packets reduce the processing that is required in the operating system because each packet requires the same amount of resources but delivers a greater payload. However, incorrectly configuring the MTU results in fragmentation and potentially undeliverable packets. This can reduce performance significantly so take care to ensure that the correct value is chosen.

If the workload sends only small messages, the larger MTU size might not result in an increase in performance, though it generally will not decrease performance.

For more information about how to set the MTU, see “Payload tuning examples” on page 227.

Path MTU discovery

Every network link has a maximum packet size described by the MTU. The datagrams can be transferred from one system to another through many links with different MTU values. If the source and destination system have different MTU values, it can cause fragmentation or dropping of packets when the smallest MTU for the link is selected. The smallest MTU for all the links in a path is called the path MTU. The process of determining the smallest MTU along the entire path from the source to the destination is called path MTU discovery (PMTUD).

For AIX Version 5.2 or earlier, the Internet Control Message Protocol (ICMP) echo request and ICMP echo reply packets are used to discover the path MTU using IPv4. The basic procedure is simple. When one system tries to optimize its transmissions by discovering the path MTU, it sends packets of its maximum size. If these do not fit through one of the links between the two systems, a notification from this link is sent back saying what maximum size this link supports. The notifications return an ICMP `Destination Unreachable` message to the source of the IP datagram, with a code that indicates the “fragmentation needed and DF set” (type 3, type 4).

When the source receives the ICMP message, it lowers the send MSS and tries again using this lower value. This process is repeated until the maximum value for all of the link steps is found.

The path MTU discovery procedure can have these possible outcomes:

- The packet can get across all the links to the destination system without being fragmented.
- The source system can get an ICMP message from any hop along the path to the destination system, indicating that the MSS is too large and not supported by this link.

This ICMP echo request and reply procedure has a few considerations. Some system administrators do not use path MTU discovery because they believe that there is a risk of denial-of-service (DoS) attacks.

Also, if you already use the path MTU discovery, routers or firewalls can block the ICMP messages from being returned to the source system. In this case, the source system does not have any messages from the network environment and sets the default MSS value, which might not be supported across all links.

The discovered MTU value is stored in the routing table by using a cloning mechanism in AIX Version 5.2 or earlier. Therefore, it cannot be used for multi-path routing. This is because the cloned route is always used instead of alternating between the two multi-path network routes. For this reason, you can see the discovered MTU value by using the **netstat -rn** command.

Beginning with AIX Version 5.3, there are changes in the procedure for path MTU discovery. The ICMP echo reply and request packets are not used anymore. AIX Version 5.3 uses TCP packets and UDP datagrams rather than ICMP echo reply and request packets. In addition, the discovered MTU are not stored in the routing table. Therefore, it is possible to enable multi-path routing to work with path MTU discovery.

When one system tries to optimize its transmissions by discovering the path MTU, a pmtu entry is created in a Path MTU (PMTU) table. You can display this table by using the **pmtu display** command as shown in Example 8-38. To avoid the accumulation of pmtu entries, unused pmtu entries will expire and be deleted when the **pmtu_expire** time is exceeded.

Example 8-38 Path MTU display

```
# pmtu display
dst             gw              If    pmtu      refcnt   redisc_t      exp
--------------------------------------------------------------------------
9.3.4.148       9.3.5.197         en0   1500        1         22        0
9.3.4.151       9.3.5.197         en0   1500        1          5        0
9.3.4.154       9.3.5.197         en0   1500        3          6        0
9.3.5.128       9.3.5.197         en0   1500       15          1        0
9.3.5.129       9.3.5.197         en0   1500        5          4        0
9.3.5.171       9.3.5.197         en0   1500        1          1        0
```

```
9.3.5.197         127.0.0.1         lo0   16896      18         2        0
192.168.0.1       9.3.4.1           en0    1500       0         1        0
192.168.128.1     9.3.4.1           en0    1500       0        25        5
9.3.5.230         9.3.5.197         en0    1500       2         4        0
9.3.5.231         9.3.5.197         en0    1500       0         6        4
127.0.0.1         127.0.0.1         lo0   16896      10         2        0
```

Path MTU table entry expiration is controlled by the **pmtu_expire** option of the **no** command. The pmtu_expire option is set to 10 minutes by default.

For IBM i, path MTU discovery is enabled by default for negotiation of larger frame transfers. To change the IBM i path MTU discovery setting, use the CHGTCPA command.

IPv6 never sends ICMPv6 packets to detect the PMTU. The first packet of a connection always starts the process. In addition, IPv6 routers are designed to never fragment packets and always return an ICMPv6 `Packet too big` message if they are unable to forward a packet because of a smaller outgoing MTU. Therefore, for IPv6, no changes are necessary to make PMTU discovery work with multi-path routing.

TCP MSS

The maximum segment size (MSS) corresponds to the payload area of the TCP packet. This is the IP MTU size minus IP and TCP header information. The MSS is the largest data or payload that the TCP layer can send to the destination system. When a connection is established, each system announces an MSS value. If one system does not receive an MSS from the other system, it uses the default MSS value.

In AIX Version 5.2 and earlier, the default MSS value was 512 bytes. Starting with AIX Version 5.3, 1460 bytes is supported as the default value.

The **no -a** command displays the value of the default MSS as `tcp_mssdflt`. On AIX, you receive the information that is shown in Example 8-39.

Example 8-39 The default MSS value in AIX 6.1

```
# no -a |grep tcp
       tcp_bad_port_limit = 0
                  tcp_ecn = 0
       tcp_ephemeral_high = 65535
        tcp_ephemeral_low = 32768
             tcp_finwait2 = 1200
          tcp_icmpsecure = 0
        tcp_init_window = 0
    tcp_inpcb_hashtab_siz = 24499
              tcp_keepcnt = 8
```

```
          tcp_keepidle = 14400
          tcp_keepinit = 150
         tcp_keepintvl = 150
tcp_limited_transmit = 1
           tcp_low_rto = 0
          tcp_maxburst = 0
           tcp_mssdflt = 1460
       tcp_nagle_limit = 65535
     tcp_nagleoverride = 0
            tcp_ndebug = 100
          tcp_newreno = 1
       tcp_nodelayack = 0
     tcp_pmtu_discover = 1
         tcp_recvspace = 16384
         tcp_sendspace = 16384
         tcp_tcpsecure = 0
          tcp_timewait = 1
               tcp_ttl = 60
        tcprexmtthresh = 3
```

For IBM i, the default MTU size that is specified by default in the Ethernet line description's maximum frame size parameter is 1496 bytes. This size means 1500 bytes for non-encapsulated TCP/IP packets.

If the source network does not receive an MSS when the connection is first established, the system uses the default MSS value. Most network environments are Ethernet, which can support at least a 1500-byte MTU.

Payload tuning examples

The following examples show how to configure these options within the operating systems that run on Power hardware.

AIX

AIX differentiates between layer-2 and layer-3 devices. To configure your system to use an MTU of 9000 with jumbo frames when you are using a physical adapter, you need configure both devices. If you are using a virtual Ethernet adapter, you only need to configure the MTU of the layer-3 device to enable the larger packet size to work on inter-partition networks within the hypervisor. To send the larger sized packets outside the managed system without fragmentation, the associated SEA on the Virtual I/O Server must be configured to bridge jumbo frames.

To configure an MTU of 9000 on a layer-3 interface, use the following command:

```
$chdev -l <en_device> -a mtu=9000
```

To configure jumbo frames on a layer-2 interface, use the following command:

```
$chdev -l <ent_device> -a jumbo_frames=yes
```

In AIX, network settings can be configured in these places:

- Globally using the **no** command.
- Per interface by using the **chdev** command on the specific device. This is the most common approach.
- Using the **ifconfig** command in the same manner as on most UNIX operating systems. However, these changes do not persist through a reboot.

IBM i

For an IBM i virtual I/O client with the default setting of the MTU size defined in the Ethernet line description, use the following procedure to increase it to MTU 9000:

1. End the TCP/IP interface for the virtual Ethernet adapter by using the ENDTCPIFC command:

   ```
   ENDTCPIFC INTNETADR('9.3.5.119')
   ```

2. Vary off the virtual Ethernet adapter line description by using the VRYCFG command:

   ```
   VRYCFG CFGOBJ(ETH01) CFGTYPE(*LIN) STATUS(*OFF)
   ```

3. Change the corresponding virtual Ethernet adapter line description by using the CHGLIND command:

   ```
   CHGLINETH LIND(ETH01) MAXFRAME(8996)
   ```

4. Vary on the virtual Ethernet adapter line description again by using the VRYCFG command:

   ```
   VRYCFG CFGOBJ(ETH01) CFGTYPE(*LIN) STATUS(*ON)
   ```

5. Start the TCP/IP interface again by using the STRTCPIFC command:

   ```
   STRTCPIFC INTNETADR('9.3.5.119')
   ```

6. Verify that jumbo frames are enabled on the IBM i virtual I/O client by using the WRKTCPSTS *IFC command and selecting `F11=Display interface status`, as shown in Figure 8-17.

```
     Internet        Subnet          Type of              Line
 Opt  Address         Mask            Service        MTU  Type
     9.3.5.119       255.255.254.0   *NORMAL        8992  *ELAN
     127.0.0.1       255.0.0.0       *NORMAL         576  *NONE
```

Figure 8-17 IBM i Work with TCP/IP Interface Status panel

Remember: Using the default setting of *ALL for the Ethernet standard parameter allows for a maximum frame size of 1496 or 8996 when protocols like SNA or TCP are required.

Setting the line description Ethernet standard to *ETHV2 allows the system to use the full Ethernet MTU size of 1500 or 9000.

Linux

To configure MTU 9000 on any interface in Linux, use the **`ifconfig`** command. Doing so also sets the driver to use jumbo frames, although this is transparent to the administrator.

```
[root@Power7-2-RHEL ~]# ifconfig eth0 mtu 9000 up
[root@Power7-2-RHEL ~]# ifconfig eth0
eth0      Link encap:Ethernet  HWaddr 6E:8D:DA:FD:46:02
          inet addr:172.16.20.174  Bcast:172.16.23.255  Mask:255.255.252.0
          UP BROADCAST RUNNING MULTICAST  MTU:9000  Metric:1
          RX packets:1479508 errors:0 dropped:0 overruns:0 frame:0
          TX packets:386984 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:323734168 (308.7 MiB)  TX bytes:38172890 (36.4 MiB)
          Interrupt:18
```

Virtual I/O Server

To configure a Shared Ethernet Adapter to bridge jumbo frames, first ensure that the physical adapter has been configured correctly by using the **`chdev`** command:

```
$ chdev -dev <physical_adapter> -attr jumbo_frames=yes
```

Then, create the Shared Ethernet Adapter by using the **jumbo_frames=yes** parameter. Substitute the appropriate adapter values for your system.

```
$ mkvdev -sea ent1 -vadapter ent8 -default ent8 -defaultid 2  -attr
ctl_chan=ent7 ha_mode=auto accounting=enabled jumbo_frames=yes
```

You can also use the **chdev** command to modify the SEA device after creation if required:

```
$ chdev -dev <sea_ent> -attr jumbo_frames=yes
```

Payload tuning verification

After you have configured your payloads at the various layers of the networking stack, you must ensure that the configuration works as intended by testing fragmentation. Each operating system has tools to check this configuration. The basic idea is to send a packet to a remote host and make sure that the packet is not fragmented in transit. Programs generally use ICMP packets and set the Do Not Fragment flag to measure this.

AIX and Virtual I/O Server

The **traceroute** command on AIX and Virtual I/O Server can help determine whether fragmentation is occurring along a path. Example 8-40 shows a trace between two AIX systems though an interface configured with an MTU of 1500. Note there are no messages about fragmentation.

Example 8-40 Example of no fragmentation using AIX

```
# traceroute 172.16.20.172
trying to get source for 172.16.20.172
source should be 172.16.20.92
traceroute to 172.16.20.172 (172.16.20.172) from 172.16.20.92 (172.16.20.92),
30 hops max
outgoing MTU = 1500
 1  172.16.20.172 (172.16.20.172)  1 ms  0 ms  0 ms
```

The systems in the previous example are on separate managed systems. In Example 8-41, the MTU of the source system is increased to 9000. No checks were made on other network devices such as SEAs or the network switches to see whether they were configured for jumbo frames. Notice the fragmentation that occurs when attempting to trace this path. In this example, the Shared Ethernet Adapters were not configured correctly, which caused this behavior.

Example 8-41 Example of fragmentation using AIX

```
# chdev -l en0 -a mtu=9000
en0 changed
# traceroute 172.16.20.172
trying to get source for 172.16.20.172
```

```
source should be 172.16.20.92
traceroute to 172.16.20.172 (172.16.20.172) from 172.16.20.92 (172.16.20.92),
30 hops max
outgoing MTU = 8166
 1  P7_1_AIX (172.16.20.92)  0 ms
fragmentation required, trying new MTU = 8146
 1   0 ms
fragmentation required, trying new MTU = 4464
 1   0 ms
fragmentation required, trying new MTU = 4444
 1   0 ms
fragmentation required, trying new MTU = 4352
 1   0 ms
fragmentation required, trying new MTU = 4332
 1   0 ms
fragmentation required, trying new MTU = 2048
 1   0 ms
fragmentation required, trying new MTU = 2028
 1   0 ms
fragmentation required, trying new MTU = 2002
 1   0 ms
fragmentation required, trying new MTU = 1982
 1   0 ms
fragmentation required, trying new MTU = 1536
 1   0 ms
fragmentation required, trying new MTU = 1516
 1   0 ms
fragmentation required, trying new MTU = 1500
 1  172.16.20.172 (172.16.20.172)  0 ms  0 ms  0 ms
```

IBM i

The Trace TCP/IP Route command, TRCTCPRTE, in IBM i can help determine whether fragmentation is occurring along a path.

Example 8-42 on page 232 shows a trace between two systems though an interface configured with an MTU of 1500 on the source system. The TRCTCPRTE command has these parameters:

PKTLEN(1500)	Sets the packet length
FRAGMENT(*NO)	Sets the Do Not Fragment option in the IP header of the probe packet

There are no messages that indicate errors in the results that are returned by the TRCTCPRTE command. This indicates that all components in the network can handle a packet size of 1500.

Example 8-42 Example of no fragmentation using IBM i

```
TRCTCPRTE RMTSYS('172.16.20.90') PKTLEN(1500) FRAGMENT(*NO)
Probing possible routes to 172.16.20.90 using *ANY interface.
1 172.16.20.90    0.101 0.064 0.068
```

In Example 8-43, a packet is sent that is larger than the MTU size configured on the source system. Notice the error message that indicates the frame was not sent successfully.

Example 8-43 Example of exceeding MTU size on IBM i

```
TRCTCPRTE RMTSYS('172.16.20.90') PKTLEN(9000) FRAGMENT(*NO)
Probing possible routes to 172.16.20.90 using *ANY interface.
*RAWSEND socket operation code 2 failed. Error number 3432, 'Message size
  out of range.'.
```

Selecting the error message and pressing F1 reveals that it was a Send Data error, suggesting the frame did not leave the IBM i partition, as shown in Figure 8-18.

```
Message ID . . . . . :   TCP3263       Severity . . . . . :   40

Message . . . . :   *RAWSEND socket operation code 2 failed. Error number
  3432, 'Message size out of range.'.
Cause . . . . . :   The socket operation codes are:
  1 - Create socket.
  2 - Send data.
  3 - Receive data.
  4 - Bind socket to port 0, IP address 172.16.20.90.
  5 - Listen operation.
  6 - Connect to destination socket.
  7 - Accept incoming connection.
Recovery  . . . :   Correct the error and try the request again. If the
  problem persists, contact service.

Bottom
```

Figure 8-18 Send data error

Diagnosing fragmentation in the network using only IBM i is more difficult. When you send packets that are smaller than the MTU defined in IBM i, but larger than the MTU size configured in an external network component, the error shown in Example 8-44 occurs. However, the same error is issued for other reasons, such as the target blocking ICMP.

Example 8-44 No response from TRCTCPRTE

```
TRCTCPRTE RMTSYS('172.16.20.90') RANGE(1) PKTLEN(7000) FRAGMENT(*NO)
Probing possible routes to 172.16.20.90 using *ANY interface.
1 * * *
2 * * *
3 * * *
4 * * *
... omitted lines ...
29 * * *
30 * * *
```

IBM i cannot determine what that real cause of the error was. To eliminate other causes of the problem, you can try and send a small packet. It is unlikely that small MTU sizes would be configured in any network. If sending a small packet succeeds but larger packets fail, the likely cause of the failure is an MTU setting somewhere in the network.

Linux

The **tracepath** command traces a network path and displays the MTU value of each hop. Example 8-45 shows the **tracepath** command failing with an MTU of 9000 configured. The MTU is then changed to 1500 and the trace works.

Example 8-45 The tracepath command on Linux

```
[root@Power7-2-RHEL ~]# tracepath 172.16.20.90
 1:  172.16.20.174 (172.16.20.174)                          0.056ms pmtu 9000
 1:  no reply
 2:  no reply
 3:  no reply
 4:  no reply
<interupted>

[root@Power7-2-RHEL ~]# ifconfig eth0 mtu 1500 up
[root@Power7-2-RHEL ~]# tracepath 172.16.20.90
 1:  172.16.20.174 (172.16.20.174)                          0.075ms pmtu 1500
 1:  172.16.20.90 (172.16.20.90)                            0.270ms reached
     Resume: pmtu 1500 hops 1 back 1
```

The **ping** command can also be used with the **-s** and **-M** options on Linux to determine the MTU between hosts.

TCP checksum offload

The TCP checksum offload option enables the network adapter to verify the TCP checksum when transmitting and receiving, which saves the host processor from having to compute the checksum. This feature is used to detect a corruption of data in the packet during transmission.

This option is enabled by default on virtual Ethernet adapters. It can be enabled or disabled by using the attribute **chksum_offload** of the adapter. PCI-X Gigabit Ethernet Adapters can operate at wire speed with the option set so it is enabled by default.

Largesend option

The gigabit or higher Ethernet adapters for IBM Power Systems support TCP segmentation offload (also called *largesend*). In largesend environments, TCP sends a large *chunk* of data to the adapter when TCP knows that the adapter supports large send. A physical Ethernet adapter breaks up this large TCP packet into multiple smaller TCP packets that fit the outgoing MTU size of the adapter. This process saves system processor load and increases network throughput.

You can apply this TCP large send capability on virtual Ethernet adapters and SEAs. It helps reduce the processor utilization of VIOSs significantly.

The TCP large send capability is extended from the Virtual I/O client all the way down to the physical adapter of VIOS. The TCP stack on the Virtual I/O client determines whether the Virtual I/O Server supports large send. If the Virtual I/O Server supports TCP large send, the Virtual I/O client sends a large TCP packet directly to the Virtual I/O Server.

You can enable or disable the **largesend** option on the SEA by using the VIOS CLI **chdev** command. To enable it, use the **-attr largesend=1** option as shown in Example 8-46. To disable it, use the **-attr largesend=0** option. As of Virtual I/O Server Version 2.2, the **largesend** option is not set on the SEA by default, and you must enable it explicitly.

Example 8-46 largesend option for Shared Ethernet Adapter

```
$ chdev -dev ent6 -attr largesend=1
ent6 changed

$ lsdev -dev ent6 -attr
attribute     value     description                                          user_settable

accounting    disabled  Enable per-client accounting of network statistics       True
ctl_chan      ent5      Control Channel adapter for SEA failover                 True
gvrp          no        Enable GARP VLAN Registration Protocol (GVRP)            True
```

```
ha_mode       auto      High Availability Mode                                    True
jumbo_frames  no        Enable Gigabit Ethernet Jumbo Frames                      True
large_receive no        Enable receive TCP segment aggregation                    True
largesend     1         Enable Hardware Transmit TCP Resegmentation
True
lldpsvc       no        Enable IEEE 802.1qbg services                             True
netaddr       0         Address to ping                                           True
pvid          10        PVID to use for the SEA device                            True
pvid_adapter  ent4      Default virtual adapter to use for non-VLAN-tagged packets  True
_mode      disabled  N/A                                                        True
real_adapter  ent1      Physical adapter associated with the SEA                  True
thread        1         Thread mode enabled (1) or disabled (0)                   True
virt_adapters ent4      List of virtual adapters associated with the SEA (comma separated) True
```

The physical adapter in the SEA must also be enabled for TCP large send for the segmentation offload from the Virtual I/O client to the SEA to work. The **large_send** option for a physical Ethernet adapter is enabled by default. You can check the status of a physical adapter by using **lsdev** command on a VIOS, and also change the attribute by using **chdev**.

```
$ lsdev -dev ent1 -attr | grep large
large_send   yes      Enable hardware TX TCP resegmentation       True
```

You can also enable the **largesend** option for virtual Ethernet adapters on an AIX virtual I/O client by using **ifconfig** command on the interface (*not* on the adapter):

```
# ifconfig en0 largesend
```

To check whether the **largesend** option is enabled on the interface, use the following command:

```
# ifconfig en0
en0:
flags=1e080863,4c0<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GR
OUPRT,64BIT,CHECKSUM_OFFLOAD(ACTIVE),LARGESEND,CHAIN>
        inet 172.16.21.104 netmask 0xfffffc00 broadcast 172.16.23.255
         tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
```

It can be disabled by using the following command:

```
# ifconfig en0 -largesend
```

Note: Enabling the **largesend** option on an interface by using the **ifconfig** command is not persistent across reboots. To make the **largesend** option on a virtual Ethernet adapter persistent, add an **ifconfig** command entry into an initial file such as `/etc/rc.net`.

Starting with AIX 6.1 TL7 SP1, and AIX 7.1 SP1, the operating system also supports the **mtu_bypass** attribute for shared Ethernet adapters to provide a persistent way to enable largesend. You can enable **mtu_bypass** with the **chdev** command.

For an IBM i client partition, nothing needs to be configured by the user because TCP large send offload is used automatically by IBM i.

Large receive option

Similar to large send, a large receive option is available for Ethernet adapters that support *TCP large receive offload* to prevent TCP segmentation for packages up to 64 KB. Enabling large receive helps reduce processor usage for TCP workload processing, and, especially for 10-Gb Ethernet, helps to improve single stream I/O throughput performance.

By default, the large_receive option is enabled for the physical adapter on the Virtual I/O Server but disabled for the SEA to account for Linux client partitions that typically do not support TCP packet sizes larger than their MTU size.

Note: For IBM i, TCP large receive support is available with IBM i 7.1 TR5 or later.

To enable TCP large receive for the SEA on the Virtual I/O Server, use the **chdev -dev <*SEA_adapter*> -attr large_receive=yes** command as shown in Example 8-47.

Example 8-47 Enabling large_receive on the SEA

```
$ chdev -dev ent6 -attr large_receive=yes
ent6 changed
$ lsdev -dev ent6 -attr
attribute     value       description                                          user_settable
accounting    disabled    Enable per-client accounting of network statistics             True
ctl_chan      ent5        Control Channel adapter for SEA failover                       True
gvrp          no          Enable GARP VLAN Registration Protocol (GVRP)                  True
ha_mode       auto        High Availability Mode                                         True
hash_algo     0           Hash algorithm used to select a SEA thread                     True
jumbo_frames  no          Enable Gigabit Ethernet Jumbo Frames                           True
large_receive yes         Enable receive TCP segment aggregation                         True
```

```
largesend     1           Enable Hardware Transmit TCP Resegmentation                    True
lldpsvc       no          Enable IEEE 802.1qbg services                                  True
netaddr       172.16.20.1 Address to ping                                                True
nthreads      7           Number of SEA threads in Thread mode                           True
pvid          1           PVID to use for the SEA device                                 True
pvid_adapter  ent4        Default virtual adapter to use for non-VLAN-tagged packets     True
qos_mode      disabled    N/A                                                            True
queue_size    8192        Queue size for a SEA thread                                    True
real_adapter  ent0        Physical adapter associated with the SEA                       True
thread        1           Thread mode enabled (1) or disabled (0)                        True
virt_adapters ent4        List of virtual adapters associated with the SEA (comma separated)
True
```

9

Storage virtualization

This chapter describes common storage managing and monitoring tasks in a PowerVM virtual storage environment. For more information about how to plan and setup storage virtualization, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

This chapter includes the following sections:

- Moving a virtual optical device to another partition
- Moving a virtual tape device to another partition
- Virtual storage configuration tracing
- Virtual storage monitoring

9.1 Moving a virtual optical device to another partition

The Virtual I/O Server support for virtual optical devices allows sharing of a physical CD or DVD drive assigned to the Virtual I/O Server between multiple AIX, IBM i, and Linux client partitions.

A shared optical device can be accessed by only one virtual I/O client partition at a time. If you want the shared optical device to be used by another virtual I/O client partition, it must first be deallocated from the client partition accessing it. It can then be allocated to another virtual I/O client partition.

The following sections describe how to allocate and deallocate a shared optical device to/from a client partition by using OS functions:

- "Allocating and deallocating a virtual optical device on AIX" on page 240
- "Allocating and deallocating a virtual optical device on IBM i" on page 242
- "Allocating and deallocating a virtual optical device on Linux" on page 246
- "Allocating and deallocating an optical device" on page 247

9.1.1 Allocating and deallocating a virtual optical device on AIX

This section describes how to allocate and deallocate a shared optical drive to or from an AIX client partition.

Remember: In IVM, the optical device is moved by using the graphical user interface.

Allocating a shared optical device on AIX

In the AIX client partition, run the **cfgmgr** command to assign the virtual optical drive to it. If the drive is already assigned to another partition, you will get an error message. In this case, you must release the drive from the partition that is holding it.

Optical drive: The virtual optical drive can also be used to install an AIX partition when selected in the SMS startup menu. This can be done only if the drive is not assigned to another LPAR.

Deallocating a shared optical device on AIX

Use the **rmdev -Rl vscsi*n*** command to change the virtual SCSI adapter and the optical device to a defined state in the AIX partition that holds the device.

If you do not know the virtual SCSI adapter number, find it with the **lscfg|grep C*n*** command, where *n* is the slot number of the virtual SCSI client adapter from the Hardware Management Console (HMC).

You can use the **dsh** command to find the AIX partition that currently holds the drive, as shown in Example 9-1. Starting with AIX 7.1, **dsh** is no longer installed by default. You must manually install it with the DSM package file sets dsm.core and dsm.dsh. You can use **dsh** with **rsh**, **ssh**, or Kerberos authentication if **dsh** can run commands without being prompted for a password.

Consideration: Set the DSH_REMOTE_CMD=/usr/bin/ssh variable if you use SSH for authentication:

```
# export DSH_REMOTE_CMD=/usr/bin/ssh
# export DSH_LIST=<file listing lpars>
# dsh lsdev -Cc cdrom|dshbak
```

The DSH_LIST and DSH_REMOTE_CMD definitions can also be added to the .profile on your admin server. You can change the file that contains the names of target LPARs without redefining DSH_LIST.

Example 9-1 Finding which LPAR is holding the optical drive by using dsh

```
# dsh lsdev -Cc cdrom|dshbak
HOST: DB_server
--------------
cd0 Available  Virtual SCSI Optical Served by VIO Server
HOST: NIM_server
---------
cd0 Defined  Virtual SCSI Optical Served by VIO Server
```

Tip: If some partitions are not displayed in the list, it is usually because the drive has never been assigned to the partition or was completely removed with the **-d** option.

You can also use the **ssh** command to query the optical device status for other AIX LPARs as shown in Example 9-2.

Example 9-2 Finding which LPAR is holding the optical drive by using ssh

```
# for i in NIM_server DB_server
> do
> echo $i; ssh $i lsdev -Cc cdrom
> done
NIM_server
```

```
cd0 Defined  Virtual SCSI Optical Served by VIO Server
DB_server
cd0 Available  Virtual SCSI Optical Served by VIO Server
```

Tips:

- If you have IBM i or/and Linux client partitions, find the partition ID of the partition that holds the drive by using the **`lsmap -all`** command on the Virtual I/O Server.
- AIX 6.1 or later offers a new graphical interface to system management called IBM Systems Console for AIX. This has a menu setup for **`dsh`**.

9.1.2 Allocating and deallocating a virtual optical device on IBM i

The following sections show how to dynamically allocate or deallocate an optical device on IBM i virtualized by the Virtual I/O Server and shared between multiple IBM i client partitions. This process eliminates the need to use dynamic LPAR to move around any physical adapter resources.

Important: An active IBM i partition will, by default, automatically configure an accessible optical device. This process makes it unavailable for use by other partitions unless the IBM i virtual IOP is disabled by using an IOP reset or removed with dynamic LPAR operation. For this reason, the IOP must remain disabled when not using the DVD.

Make sure to select the correct IOP to be reset. Otherwise, if redundancy for disk devices has been lost, the IBM i partition might lose access to its disk storage and become inaccessible.

Allocating a shared optical device on IBM i

To allocate a shared optical device, complete these steps:

1. Use the WRKHDWRSC *STG command to verify that the IBM i virtual IOP (type 290A) for the optical device is *operational*.

   If it is *inoperational* as shown here in Figure 9-1, locate the logical resource for the virtual IOP in SST Hardware Service Manager and re-IPL the virtual IOP. Use the **I/O debug** and **IPL I/O processor** options as shown in Figure 9-2 on page 244 and Figure 9-3 on page 245.

```
                          Work with Storage Resources
                                                                System:E101F170
Type options, press Enter.
  7=Display resource detail   9=Work with resource

Opt  Resource        Type-model  Status               Text
     CMB01           290A-001    Operational          Storage Controller
       DC01          290A-001    Operational          Storage Controller
     CMB02           290A-001    Operational          Storage Controller
       DC02          290A-001    Operational          Storage Controller
     CMB03           290A-001    Inoperative          Storage Controller
       DC03          290A-001    Inoperative          Storage Controller
     CMB05           268C-001    Operational          Storage Controller
       DC05          6B02-001    Operational          Storage Controller

Bottom
 F3=Exit   F5=Refresh   F6=Print   F12=Cancel
```

Figure 9-1 IBM i Work with Storage Resources panel

Figure 9-2 shows the I/O debug option.

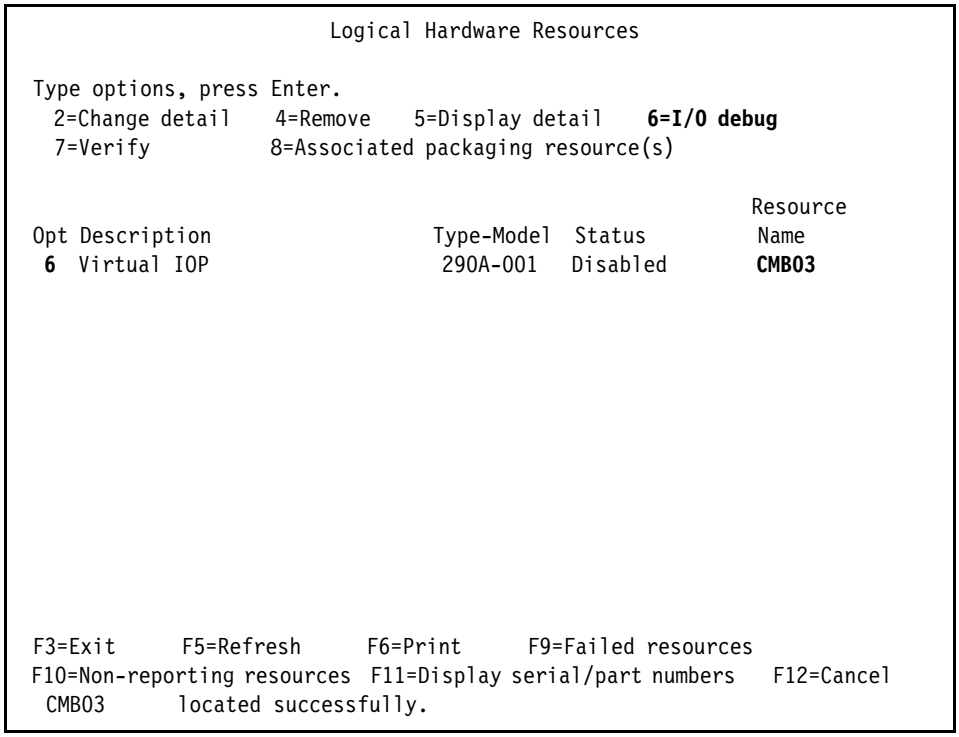


Figure 9-2 IBM i Logical Hardware Resources panel: I/O debug option

Figure 9-3 shows the IPL I/O processor option.

```
                    Select IOP Debug Function

Resource name . . . . . . . . . :  CMB03
Dump type . . . . . . . . . . . :  Normal

Select one of the following:

     1. Read/Write I/O processor data
     2. Dump I/O processor data
     3. Reset I/O processor
     4. IPL I/O processor
     5. Enable I/O processor trace
     6. Disable I/O processor trace




Selection
     -

F3=Exit      F12=Cancel
F8=Disable I/O processor reset     F9=Disable I/O processor IPL
Re-IPL of IOP was successful.
```

Figure 9-3 IBM i Select IOP Debug Function panel: IPL I/O processor option

2. After the IOP is operational, vary on the optical drive by using this command:

   ```
   VRYCFG CFGOBJ(OPT01) CFGTYPE(*DEV) STATUS(*ON)
   ```

   Tip: As an alternative, you can use the WRKCFGSTS *DEV command to access the options on the Work with Configuration Status panel.

Deallocating a shared virtual optical device on IBM i

To deallocate a shared virtual optical device, complete these steps:

1. Use the following VRYCFG command to vary off the optical device from IBM i:

   ```
   VRYCFG CFGOBJ(OPT01) CFGTYPE(*DEV) STATUS(*OFF)
   ```

2. To release the optical device with its virtual SCSI connection to the virtual SCSI server adapter for use by another Virtual I/O Server client partition, disable its virtual IOP from the SST Hardware Service Manager. Locate the logical resource for the virtual IOP, and select the **I/O debug** option.
3. Select the **Reset I/O processor** option, as shown in Figure 9-4.

```
                    Select IOP Debug Function

Resource name . . . . . . . . . :  CMB03
Dump type . . . . . . . . . . . :  Normal

Select one of the following:

     1. Read/Write I/O processor data
     2. Dump I/O processor data
     3. Reset I/O processor
     4. IPL I/O processor
     5. Enable I/O processor trace
     6. Disable I/O processor trace




Selection
     -

F3=Exit      F12=Cancel
F8=Disable I/O processor reset     F9=Disable I/O processor IPL
Reset of IOP was successful.
```

Figure 9-4 IBM i Select IOP Debug Function panel: Reset I/O processor option

9.1.3 Allocating and deallocating a virtual optical device on Linux

A virtual optical device can be assigned to a Red Hat or Novell SuSE Linux partition. However, the partition must be rebooted to be able to assign the free drive. Likewise, the partition must be shut down to release the drive.

Linux automatically detects virtual optical devices that are provided by the Virtual I/O Server. Usually virtual optical devices are named as /dev/sr<ID>.

9.1.4 Allocating and deallocating an optical device

A physical optical device can be assigned to a Virtual I/O Server that is virtualized for use by its client partition. If it is used in the Virtual I/O Server itself for Virtual I/O Server maintenance or local backups, it must be unconfigured to make the physical device accessible for the Virtual I/O Server. To do so, complete the following steps:

1. Release the drive from the client partition that holds it.
2. Unconfigure the virtual optical device in the Virtual I/O Server by using the **rmdev -dev <*virtual_optical_device*> -ucfg** command.

   Attention: Take care not to unconfigure recursively when production disks share an adapter with the optical drive.

   If a CD is in the drive, the **rmdev** command will fail because the device is allocated (ready).

3. When finished using the drive locally, use the **cfgdev** command in the Virtual I/O Server to configure the drive as a virtual drive again.

Complete the following steps to unconfigure the virtual optical device in one Virtual I/O Server when it is going to be moved *physically* to another Virtual /O Server partition, and to move it back:

1. Release the drive from the client partition that holds it.
2. Unconfigure the virtual device in the Virtual I/O Server.
3. Unconfigure the PCI or SAS adapter recursively.

   Tip: Use the **lsdev -dev <*drive*> -parent** command to find the correct parent adapter for the drive to be removed.

4. Use the HMC to move the adapter to the target partition.
5. Run the **cfgdev** command on the other Virtual I/O Server partition to configure the drive. You can also use the **mkvdev** command to virtualize it for use by virtual I/O client partitions.
6. When finished, remove the PCI adapter recursively.
7. Use the HMC to move the adapter back.

8. Run the **cfgdev** command on the original Virtual I/O Server partition to reconfigure the drive to make it available again as a virtual optical device.

Example 9-3 shows the commands that are used for unconfiguring and configuring the drive (disregard the error message from the test system).

Example 9-3 Unconfiguring and reconfiguring the DVD drive

```
$ rmdev -dev vcd -ucfg
vcd Defined
$ lsdev -slots
# Slot                     Description       Device(s)
U787B.001.DNW108F-P1-C1    Logical I/O Slot  pci3 ent0
U787B.001.DNW108F-P1-C3    Logical I/O Slot  pci4 fcs0
U787B.001.DNW108F-P1-C4    Logical I/O Slot  pci2 sisioa0
U787B.001.DNW108F-P1-T16   Logical I/O Slot  pci5 ide0
U9113.550.105E9DE-V1-C0    Virtual I/O Slot  vsa0
U9113.550.105E9DE-V1-C2    Virtual I/O Slot  ent1
U9113.550.105E9DE-V1-C3    Virtual I/O Slot  ent2
U9113.550.105E9DE-V1-C4    Virtual I/O Slot  vhost0
U9113.550.105E9DE-V1-C20   Virtual I/O Slot  vhost1
U9113.550.105E9DE-V1-C22   Virtual I/O Slot  vhost6
U9113.550.105E9DE-V1-C30   Virtual I/O Slot  vhost2
U9113.550.105E9DE-V1-C40   Virtual I/O Slot  vhost3
U9113.550.105E9DE-V1-C50   Virtual I/O Slot  vhost4
$ rmdev -dev pci5 -recursive -ucfg
cd0 Defined
ide0 Defined
pci5 Defined

$ cfgdev
Method error (/usr/lib/methods/cfg_vt_optical -l vcd ):
$ lsdev -virtual
name            status
description
ent1            Available  Virtual I/O Ethernet Adapter (l-lan)
ent2            Available  Virtual I/O Ethernet Adapter (l-lan)
vhost0          Available  Virtual SCSI Server Adapter
vhost1          Available  Virtual SCSI Server Adapter
vhost2          Available  Virtual SCSI Server Adapter
vhost3          Available  Virtual SCSI Server Adapter
vhost4          Available  Virtual SCSI Server Adapter
vhost6          Available  Virtual SCSI Server Adapter
vsa0            Available  LPAR Virtual Serial Adapter
apps_rootvg     Available  Virtual Target Device - Disk
db_rootvg       Available  Virtual Target Device - Disk
```

```
linux_lvm       Available  Virtual Target Device - Disk
nim_rootvg      Available  Virtual Target Device - Disk
vcd             Available  Virtual Target Device - Optical Media
vtscsi0         Available  Virtual Target Device - Logical Volume
ent3            Available  Shared Ethernet Adapter
```

9.2 Moving a virtual tape device to another partition

The Virtual I/O Server support for virtual tape devices allows sharing of a physical tape drive that is assigned to the Virtual I/O Server between multiple AIX, IBM i, and Linux client partitions.

A shared tape device can be accessed only by one virtual I/O client partition at a time. If the shared tape device is to be used by another virtual I/O client partition, it first must be deallocated from the client partition currently accessing it. It can then be allocated to another virtual I/O client partition.

The following sections describe how to allocate and deallocate a shared tape device to/from a client partition:

- "Allocating and deallocating a virtual tape device on AIX" on page 249
- "Allocating and deallocating a virtual tape device on IBM i" on page 250
- "Allocating and deallocating a virtual tape device on Linux" on page 250
- "Allocating and deallocating a tape device" on page 251

Tip: Moving the virtual tape device is similar to the scenarios for moving the virtual optical device as described in 9.1, "Moving a virtual optical device to another partition" on page 240.

9.2.1 Allocating and deallocating a virtual tape device on AIX

This section describes how to allocate and deallocate a shared tape drive to or from an AIX client partition.

Allocating a shared virtual tape device on AIX

Using the **cfgmgr** command in the AIX target LPAR makes the drive available.

Remember: If the tape drive is not assigned to another LPAR, the drive is displayed as an install device in the SMS menu.

Deallocating a shared virtual tape device on AIX

To deallocate a shared virtual tape device, complete these steps:

1. If you do not know the vscsi adapter number, determine it by using the **lscfg|grep C*n*** command, where *n* is the slot number of the virtual SCSI client adapter from the HMC.
2. Use the **rmdev -Rl vscsi*n*** command to change the vscsi adapter and the tape drive to a defined state in the AIX client partition that holds the drive. Adding the **-d** option also removes the adapter from the ODM.

You can use the **dsh** command in combination with the **lsdev -Cc tape** command to find if another AIX client is holding the tape drive. The procedure is similar to the one documented for virtual optical devices in 9.1.1, “Allocating and deallocating a virtual optical device on AIX” on page 240.

9.2.2 Allocating and deallocating a virtual tape device on IBM i

Allocating or deallocating a shared virtual tape drive on IBM i works the same way as for a shared virtual optical device as described in 9.1.2, “Allocating and deallocating a virtual optical device on IBM i” on page 242.

To deallocate a shared virtual device on the IBM i client partition, vary it off first. Then release it for use by another Virtual I/O Server client partition by resetting the IBM i virtual IOP.

To allocate a shared virtual device, if the virtual IOP is in *inoperative* state, the virtual IOP must be re-IPLed first. When the IOP is operational, the shared virtual device can be varied on.

9.2.3 Allocating and deallocating a virtual tape device on Linux

If you do not know the vscsi adapter number, determined it by using the **lscfg|grep C*n*** command, where *n* is the slot number of the virtual SCSI client adapter from the HMC.

Important: If the virtual SCSI server adapter on the Virtual I/O Server is configured with the `Any client partition can connect` option and shared among client partitions, the virtual tape drive cannot be removed from the running Linux client partition. This is because the virtual SCSI adapter on the Linux client connecting to the virtual SCSI server adapter cannot be removed by Linux commands.

If you want to remove the virtual SCSI adapter on the Linux, remove the virtual SCSI adapter by using the dynamic LPAR operation on the HMC. You can also shut down the Linux client partition that holds the virtual tape drive.

If you want to move the virtual tape drive on the Linux client partition without the dynamic LPAR or shutdown operation, configure dedicated virtual SCSI server-client pairs for each client partition without the `Any client partition can connect` option.

To allocate and deallocate a virtual tape device, complete these steps:

1. Enter **`echo 1 > /sys/block/st0/device/delete`** to remove the tape drive from the Linux client partition that holds the drive.
2. On the Virtual I/O Server, remove the virtual target device for the virtual tape drive and map the physical tape drive to the target client partition.
3. Enter **`echo “- - -” > /sys/class/scsi_host/host`*`X`*`/scan`** to recognize the tape drive on the target LPAR, where ***`X`*** stands for the SCSI bus you want to scan.

9.2.4 Allocating and deallocating a tape device

A physical tape device assigned to the Virtual I/O Server can be virtualized for use by its client partition to be used in the Virtual I/O Server itself for local restores or backups. If so, its corresponding virtual tape device must be unconfigured first to make the physical device accessible for the Virtual I/O Server.

If this physical tape device is to be used as a physical device by another partition, it must be further deconfigured together with its parent PCI or SAS adapter on the Virtual I/O Server currently owning it. It can then be moved like a dynamic LPAR to another partition.

For more information about allocating and deallocating a virtual tape device on the Virtual I/O Server, see 9.1.4, “Allocating and deallocating an optical device” on page 247. The procedure is essentially the same as for a virtual optical device.

9.3 Virtual storage configuration tracing

An important aspect of managing a virtual environment is tracking which virtual objects correspond to which physical objects. This is particularly challenging in the storage arena, where individual virtual servers can have hundreds of virtual disks.

Important: Understanding and documentation of an end-to-end virtual to physical device mapping is critical to managing performance and understanding which systems will be affected by hardware maintenance.

As illustrated in Figure 9-5, virtual disks can be mapped to physical disks as *physical volumes* or as *logical volumes*. Logical volumes can be mapped from *volume groups* or *storage pools*.

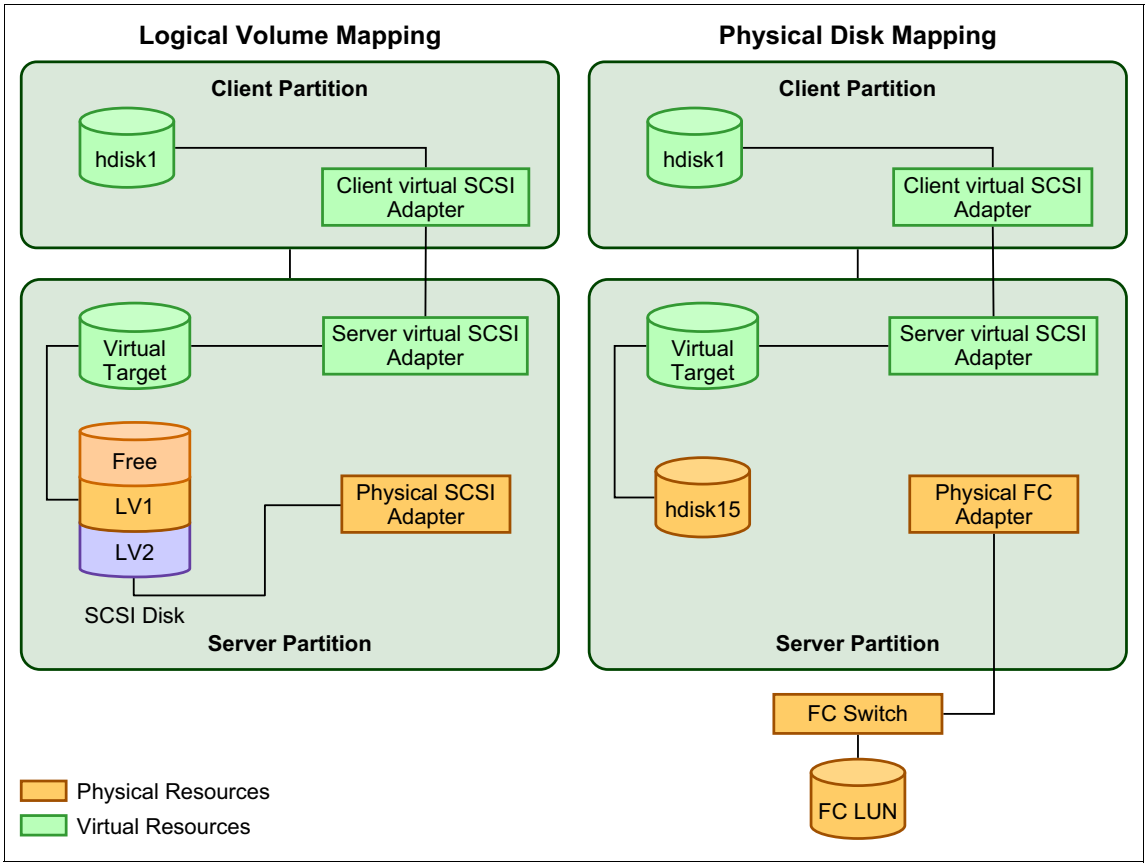


Figure 9-5 Logical versus physical drive mapping

Depending on which storage provisioning method you choose, you might need to track the following information:

- Virtual I/O Server:
  - Server host name
  - Physical disk location
  - Physical adapter device name
  - Physical hdisk device name
  - Cluster name (for shared storage pool backed devices only)
  - Volume group or storage pool name (for logical volume or storage pool backed devices only)
  - Logical volume or storage pool backing device name (for logical volume or storage pool backed devices only)
  - Virtual SCSI adapter slot
  - Virtual SCSI adapter device name
  - Virtual target device
- Virtual I/O client:
  - Client host name
  - Virtual SCSI adapter slot
  - Virtual SCSI adapter device name
  - Virtual disk device name

When tracking information that pertains to the partition profile, including virtual adapter IDs, you can use the System Planning Tool (SPT) for planning and documenting your configuration. The SPT system plan can be deployed through an HMC or the Integrated Virtualization Manager (IVM). It ensures correct naming and numbering.

For more information about creating and deploying an SPT system plan, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

The following sections cover how to manually trace virtual disks from the virtual I/O client back to the physical hardware for the available client operating systems:

- “AIX virtual storage configuration tracing” on page 254
- “IBM i virtual storage configuration tracing” on page 255
- “Linux virtual storage configuration tracing” on page 267

9.3.1 AIX virtual storage configuration tracing

AIX virtual storage (including NPIV and logical units from a shared storage pool) can be traced from Virtual I/O Server by using the **lsmap** command.

Example 9-4 illustrates tracing *virtual SCSI* storage from the Virtual I/O Server.

Example 9-4 Tracing virtual SCSI storage from Virtual I/O Server

```
$ lsmap -all
SVSA            Physloc                                     Client Partition ID
--------------- ------------------------------------------- ------------------
vhost0          U9117.MMA.101F170-V1-C21                    0x00000003

VTD                   aix61_rvg
Status                Available
LUN                   0x8100000000000000
Backing device        hdisk7
Physloc               U789D.001.DQDYKYW-P1-C2-T2-W201300A0B811A662-L1000000000000

SVSA            Physloc                                     Client Partition ID
--------------- ------------------------------------------- ------------------
vhost1          U9117.MMA.101F170-V1-C22                    0x00000004

VTD                   NO VIRTUAL TARGET DEVICE FOUND
```

Example 9-5 illustrates how to trace *NPIV* storage devices from the Virtual I/O Server by using the **lsmap -npiv -all** command. `ClntID` shows the LPAR ID as seen from the HMC. `ClntName` is the host name.

Example 9-5 Tracing NPIV virtual storage from the Virtual I/O Server

```
$ lsmap -npiv -all
Name          Physloc                            ClntID ClntName       ClntOS
============= ================================== ====== ============== =======
vfchost0      U9117.MMA.101F170-V1-C31                 3 AIX61          AIX

Status:LOGGED_IN
FC name:fcs3                    FC loc code:U789D.001.DQDYKYW-P1-C6-T2
Ports logged in:2
Flags:a<LOGGED_IN,STRIP_MERGE>
VFC client name:fcs2            VFC client DRC:U9117.MMA.101F170-V3-C31-T1

Name          Physloc                            ClntID ClntName       ClntOS
============= ================================== ====== ============== =======
vfchost1      U9117.MMA.101F170-V1-C32                 4

Status:NOT_LOGGED_IN
FC name:fcs3                    FC loc code:U789D.001.DQDYKYW-P1-C6-T2
```

```
Ports logged in:0
Flags:4<NOT_LOGGED>
VFC client name:                VFC client DRC:
```

Example 9-6 shows how to trace storage that is assigned from a *shared storage pool*.

Example 9-6 Listing all disk mappings in a cluster

```
$ lsmap -clustername ssp_cluster -all
Physloc                                          Client Partition ID
------------------------------------------------ ------------------
U8233.E8B.061AA6P-V33-C136                       0x00000024

VTD                  vtscsi0
LUN                  0x8100000000000000
Backing device       sspdisk06.f578cbb7b4d930ccbe3abc27f8f62376

Physloc                                          Client Partition ID
------------------------------------------------ ------------------
U8233.E8B.10OEF5R-V1-C104                        0x00000004

VTD                  vtscsi0
LUN                  0x8100000000000000
Backing device       sspdisk01.198d854abebe7e965214d8360eae60fe
```

For more information about tracing storage from a shared storage pool, see 10.2, “Monitoring shared storage pools” on page 310.

The IBM Systems Hardware Information Center contains a guide to tracing virtual disks, which is available at:

http://publib.boulder.ibm.com/infocenter/systems/scope/hw/index.jsp?topic=/iphb1/iphb1_vios_managing_mapping.htm

9.3.2 IBM i virtual storage configuration tracing

This section describes how to trace the configuration of virtual SCSI and virtual Fibre Channel LUNs from the IBM i client perspective down to the physical storage resource.

IBM i virtual SCSI disk configuration tracing

Virtual SCSI LUNs *always* show up as disk units with device type *6B22 model 050* on the IBM i client, regardless of which storage subsystem ultimately provides the backing physical storage.

Figure 9-6 shows an example of IBM i System Service Tools output retrieved by using STRSST → **3. Work with disk units** → **1. Display disk configuration** → **1. Display disk configuration status**. The output shows the disk configuration from a new IBM i client after SLIC installation that is set up for mirroring its disk units across two Virtual I/O Servers.

```
                    Display Disk Configuration Status

          Serial                      Resource
ASP Unit  Number          Type Model Name       Status
  1                                             Mirrored
      1   Y3WUTVVQMM4G    6B22  050  DD001      Active
      1   YYUUH3U9UELD    6B22  050  DD004      Resume Pending
      2   YD598QUY5XR8    6B22  050  DD003      Active
      2   YTM3C79KY4XF    6B22  050  DD002      Resume Pending


Press Enter to continue.

F3=Exit       F5=Refresh        F9=Display disk unit details
F11=Disk configuration capacity   F12=Cancel
```

Figure 9-6 IBM i SST Display Disk Configuration Status panel

To trace the IBM i disk units to the corresponding SCSI devices on the Virtual I/O Server, select **F9=Display disk unit details** to display the disk unit details information as shown in Figure 9-7.

```
                          Display Disk Unit Details

Type option, press Enter.
  5=Display hardware resource information details

              Serial          Sys  Sys    I/O   I/O
OPT ASP Unit  Number          Bus  Card Adapter Bus   Ctl  Dev  Compressed
      1    1  Y3WUTVVQMM4G    255   21            0     1    0   No
      1    1  YYUUH3U9UELD    255   22            0     2    0   No
      1    2  YD598QUY5XR8    255   21            0     2    0   No
      1    2  YTM3C79KY4XF    255   22            0     1    0   No



F3=Exit          F9=Display disk units                F12=Cancel
```

Figure 9-7 IBM i SST Display Disk Unit Details panel

Tip: To trace down an IBM i virtual disk unit to the corresponding virtual target device (VTD) and backing hdisk on the Virtual I/O Server, use the provided system card `Sys Card` and controller `Ctl` information from the IBM i client:

- `Sys Card` shows the IBM i virtual SCSI client adapter slot as configured in the IBM i partition profile.
- `Ctl` XOR 0x80 corresponds to the virtual target device LUN information on the Virtual I/O Server.

Alternatively, the virtual SCSI client adapter slot and target device LUN information can also be retrieved from the location information of the DSPHDWRSC *STG or WRKHDWRSC *STG command output. After you run either of these CL commands, select `9=Display associated resource`. Then select `9=Work with resource` for a 290A virtual IOP, and `7=Display resource detail` for a 6B22 disk unit resource. The location information is displayed, such as the output for virtual SCSI client adapter slot `21` and virtual target device LUN `0x81` shown in Figure 9-8.

```
                          Display Resource Detail
                                                                System:   IFLEX
 Resource name  . . . . . . . :   DD001
 Text . . . . . . . . . . . . :
 Type-model . . . . . . . . . :   6B22-050
 Serial number  . . . . . . . :   Y3WUTVVQMM4G
 Part number  . . . . . . . . :

Location :   U7895.42X.103553B-V3-C21-T1-L810000000000000000

 Logical address:
  SPD bus:
    System bus                    255
    System board                  128
    System card                   21
                                                                      More...
  Storage:
    I/O bus                       0
    Controller                    1
    Device                        0
                                                                      Bottom
 Press Enter to continue.

 F3=Exit   F5=Refresh   F6=Print   F12=Cancel
```

Figure 9-8 IBM i WRKHDWRSC Display Resource Detail: 6B22 disk unit device

The following example illustrates how to trace the IBM i mirrored load source (disk unit 1) reported on IBM i at `Sys Card 21 Ctl 1` and `Sys Card 22 Ctl 2` down to the devices on the two Virtual I/O Servers and the SAN storage system.

1. Look at the virtual adapter mapping in the IBM i partition properties on the HMC, as shown in Figure 9-9. The Sys Card 21 and 22 information from the IBM i client corresponds to the virtual SCSI adapter slot numbers. Therefore, the partition properties information shows that the IBM i client virtual SCSI client adapters 21 and 22 connect to the virtual SCSI server adapters 23 and 23 of the Virtual I/O Server partitions vios1 and vios2.

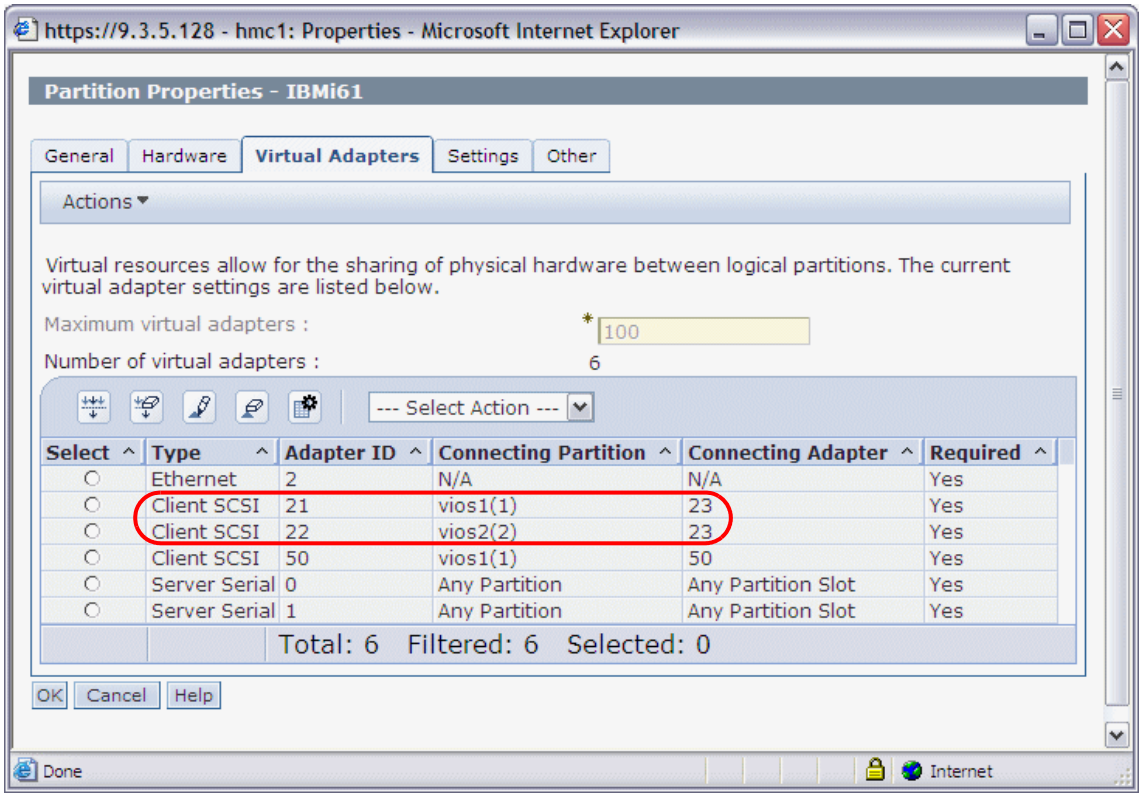


Figure 9-9 IBM i partition profile virtual adapters configuration

2. Knowing the corresponding virtual SCSI server adapter slots 23 and 23, look at the device mapping on the two Virtual I/O Servers. The **lsmap** command on Virtual I/O Server vios1 shows the device mapping between physical and virtual devices as shown in Example 9-7.

Example 9-7 Displaying the Virtual I/O Server device mapping

```
$ lsdev -slots
# Slot                     Description       Device(s)
U789D.001.DQDYKYW-P1-T1    Logical I/O Slot  pci4 usbhc0 usbhc1
U789D.001.DQDYKYW-P1-T3    Logical I/O Slot  pci3 sissas0
U9117.MMA.101F170-V1-C0    Virtual I/O Slot  vsa0
U9117.MMA.101F170-V1-C2    Virtual I/O Slot  vasi0
U9117.MMA.101F170-V1-C11   Virtual I/O Slot  ent2
U9117.MMA.101F170-V1-C12   Virtual I/O Slot  ent3
U9117.MMA.101F170-V1-C13   Virtual I/O Slot  ent4
U9117.MMA.101F170-V1-C21   Virtual I/O Slot  vhost0
U9117.MMA.101F170-V1-C22   Virtual I/O Slot  vhost1
U9117.MMA.101F170-V1-C23   Virtual I/O Slot  vhost2
U9117.MMA.101F170-V1-C24   Virtual I/O Slot  vhost3
U9117.MMA.101F170-V1-C25   Virtual I/O Slot  vhost4
U9117.MMA.101F170-V1-C50   Virtual I/O Slot  vhost5
U9117.MMA.101F170-V1-C60   Virtual I/O Slot  vhost6

$ lsmap -vadapter vhost2
SVSA            Physloc                                        Client Partition ID
--------------- ---------------------------------------------- ------------------
vhost2          U9117.MMA.101F170-V1-C23                       0x00000005

VTD                   IBMi61_0
Status                Available
LUN                   0x8100000000000000
Backing device        hdisk11
Physloc               U789D.001.DQDYKYW-P1-C2-T2-W201300A0B811A662-L5000000000000

VTD                   IBMi61_1
Status                Available
LUN                   0x8200000000000000
Backing device        hdisk12
Physloc               U789D.001.DQDYKYW-P1-C2-T2-W201300A0B811A662-L6000000000000
```

3. Because the IBM i client `disk unit 1` is connected to `Sys Card 21` and `Ctl 1`, you find the corresponding virtual target device LUN on the Virtual I/O Server: *`Ctl 1`* `XOR 0x80 = 0x81`. That is, LUN 0x81, which is backed by `hdisk11`, corresponds to the disk unit 1 of the IBM i client whose mirror side is connected to vios1.

4. To locate `hdisk11` and its physical disk and LUN on the SAN storage system, use the **`lsdev`** command to see its multipath device. After you determine that it is an MPIO device, use the **`mpio_get_config`** command as shown in Example 9-8. This process determines that IBM i disk unit 1 corresponds to LUN 5 on the DS4800 storage subsystem.

Example 9-8 Virtual I/O Server hdisk to LUN tracing

```
$ lsdev -dev hdisk11
name             status      description
hdisk11          Available   MPIO Other DS4K Array Disk

$ oem_setup_env
# mpio_get_config -Av
Frame id 0:
    Storage Subsystem worldwide name: 60ab800114632000048ed17e
    Controller count: 2
    Partition count: 1
    Partition 0:
    Storage Subsystem Name = 'ITSO_DS4800'
        hdisk      LUN #   Ownership          User Label
        hdisk6         0   A (preferred)      VIOS1
        hdisk7         1   A (preferred)      AIX61
        hdisk8         2   B (preferred)      AIX53
        hdisk9         3   A (preferred)      SLES10
        hdisk10        4   B (preferred)      RHEL52
        hdisk11        5   A (preferred)      IBMi61_0
        hdisk12        6   B (preferred)      IBMi61_1
        hdisk13        7   A (preferred)      IBMi61_0m
        hdisk14        8   B (preferred)      IBMi61_1m
```

IBM i virtual Fibre Channel disk configuration tracing

Virtual Fibre Channel LUNs from NPIV-attached IBM System Storage® DS8000® storage systems are displayed as disk units with their native device type 2107 and model A*xx* as configured on the DS8000. They report in under a virtual IOP/IOA device type-model *6B25-001* for the virtual Fibre Channel client

adapter as can be seen from the following IBM i system service tools (SST) panel from Hardware Service Manager for Logical Hardware Resources shown in Figure 9-10.

```
                Logical Hardware Resources Associated with IOP

Type options, press Enter.
  2=Change detail    4=Remove    5=Display detail    6=I/O debug
  7=Verify           8=Associated packaging resource(s)

                                                                 Resource
Opt Description                    Type-Model  Status            Name
    Virtual IOP                     6B25-001   Operational       CMB09
     Virtual Storage IOA            6B25-001   Operational       DC04
      Disk Unit                     2107-A02   Operational       DD002
      Disk Unit                     2107-A02   Operational       DD003
      Disk Unit                     2107-A02   Operational       DD004
      Disk Unit                     2107-A02   Operational       DD012




F3=Exit    F5=Refresh    F6=Print    F8=Include non-reporting resources
F9=Failed resources      F10=Non-reporting resources
F11=Display serial/part numbers      F12=Cancel
```

Figure 9-10 IBM i SST Logical Hardware Resources Associated with IOP

To trace down the IBM i disk units to the corresponding DS8000 volume IDs, select `F11=Display serial/part numbers` to display the disk unit serial numbers as shown in Figure 9-11. The IBM i disk unit serial number 50-*XXXXYYY* includes the 4-digit DS8000 volume ID *XXXX* followed by a 3-digit suffix *YYY.* The suffix is by default composed of the last three digits from the DS8000 worldwide node name (WWNN) or, on older DS8000 systems, the default 001 or a user-defined number.

```
                Logical Hardware Resources Associated with IOP

Type options, press Enter.
  2=Change detail    4=Remove    5=Display detail    6=I/O debug
  7=Verify           8=Associated packaging resource(s)

                                                Serial          Part
Opt Description                   Type-Model    Number          Number
    Virtual IOP                    6B25-001     00-00000
     Virtual Storage IOA           6B25-001     00-00000
      Disk Unit                    2107-A02     50-1000001
      Disk Unit                    2107-A02     50-1001001
      Disk Unit                    2107-A02     50-1100001
      Disk Unit                    2107-A02     50-1101001









F3=Exit    F5=Refresh    F6=Print    F8=Include non-reporting resources
F9=Failed resources      F10=Non-reporting resources
F11=Display logical address          F12=Cancel
```

Figure 9-11 IBM i SST Logical Hardware Resources disk unit serial numbers

Selecting `5=Display detail` for the virtual storage IOA displays the IBM i virtual Fibre Channel client adapter's worldwide port name (WWPN) `C05076030398000E` and slot number `41` that were configured for the IBM i client partition on the HMC as shown in Figure 9-12.

```
                  Auxiliary Storage Hardware Resource Detail

Description . . . . . :   Virtual Storage IOA
Type-model  . . . . . :   6B25-001
Status  . . . . . :   Operational
Serial number . . . . . :   00-00000
Part number . . . . . :
Resource name . . . . . :   DC04
Port  . . . . . :   0
  Worldwide port name . . . . . :   C05076030398000E
Physical location . . . . . :   U8233.E8B.061AA6P-V6-C41
SPD bus . . . . . :
  System bus  . . . . . :   255
  System board  . . . . . :   128
  System card . . . . . :   41
Storage . . . . . :
  I/O adapter . . . . . :
  I/O bus . . . . . :   127
  Controller  . . . . . :
                                                                  More...

F3=Exit      F5=Refresh      F6=Print
F9=Change detail     F11=Display additional port information       F12=Cancel
```

Figure 9-12 IBM i SST Auxiliary Storage Hardware Resource Detail

Note: Like virtual SCSI configuration tracing, the virtual Fibre Channel client adapter slot information, LUN ID, and virtual Fibre Channel adapter WWPN information can also be displayed by using the DSPHDWRSC *STG or WRKHDWRSC *STG command. Select `7=Display resource detail` for the 6B25 virtual IOP to view its corresponding disk unit resources.

Knowing the IBM i partition's virtual Fibre Channel adapter resource location `U8233.E8B.061AA6P-V6-C41`, that is, slot `41`, or resource name `DC04`, you can trace the corresponding physical Fibre Channel adapter with its location and WWPN (shown in the network address field) used on the Virtual I/O Server by using the **`lsmap -all -npiv`** and **`lsdev -dev fcs`*`X`*` -vpd`** command (Example 9-9).

Example 9-9 Virtual I/O Server virtual to physical Fibre Channel adapter mapping

```
$ lsmap -all -npiv
Name          Physloc                            ClntID ClntName       ClntOS
```

```
------------ -------------------------------- ------ -------------- -------
vfchost0     U8233.E8B.061AA6P-V1-C36               7 P7_2_AIX       AIX

Status:LOGGED_IN
FC name:fcs0                    FC loc code:U5802.001.0086848-P1-C2-T1
Ports logged in:2
Flags:a<LOGGED_IN,STRIP_MERGE>
VFC client name:fcs0            VFC client DRC:U8233.E8B.061AB2P-V3-C36-T1

Name         Physloc                            ClntID ClntName       ClntOS
------------ -------------------------------- ------ -------------- -------
vfchost1     U8233.E8B.061AA6P-V1-C41               6 IBM i          IBM i

Status:LOGGED_IN
FC name:fcs0                    FC loc code:U5802.001.0086848-P1-C2-T1
Ports logged in:1
Flags:a<LOGGED_IN,STRIP_MERGE>
VFC client name:DC04            VFC client DRC:U8233.E8B.061AA6P-V6-C41

$ lsdev -dev fcs0 -vpd
  fcs0             U5802.001.0086848-P1-C2-T1  8Gb PCI Express Dual Port FC Adapter
(df1000f114108a03)

        Part Number.................10N9824
        Serial Number...............1B02104269
        Manufacturer................001B
        EC Level....................D76482B
        Customer Card ID Number.....577D
        FRU Number..................10N9824
        Device Specific.(ZM)........3
        Network Address.............10000000C99FC71E
        ROS Level and ID............02781174
        Device Specific.(Z0)........31004549
        Device Specific.(Z1)........00000000
        Device Specific.(Z2)........00000000
        Device Specific.(Z3)........09030909
        Device Specific.(Z4)........FF781116
        Device Specific.(Z5)........02781174
        Device Specific.(Z6)........07731174
        Device Specific.(Z7)........0B7C1174
        Device Specific.(Z8)........20000000C99FC71E
        Device Specific.(Z9)........US1.11X4
        Device Specific.(ZA)........U2D1.11X4
        Device Specific.(ZB)........U3K1.11X4
        Device Specific.(ZC)........00000000
        Hardware Location Code......U5802.001.0086848-P1-C2-T1

  PLATFORM SPECIFIC
```

```
Name:  fibre-channel
  Model:  10N9824
  Node:  fibre-channel@0
  Device Type:  fcp
  Physical Location: U5802.001.0086848-P1-C2-T1
```

To locate the virtual Fibre Channel client adapter, log in to the SAN switch and look at the name server registrations on the switch as shown with the **nsshow** command for a Brocade FOS SAN switch in Example 9-10. The *Pid* column information shows the Fibre Channel address. The address is in the form *DDPPNN* with *DD*=switch domain in hex, *PP* = switch port in hex, *NN* = sequential number for the physical or virtual FC port connected to the switch port in hex. The IBM i virtual Fibre Channel adapter with WWPN C05076030398000E is logged in to switch domain 01, port 00 as NPIV port number 02 of the physical adapter WWPN 10000000C99FC71E.

Example 9-10 Brocade SAN switch name server registration information

```
itso-aus-san-01:admin> nsshow
{
 Type Pid    COS     PortName                NodeName                 TT
 N    010000;    2,3;10:00:00:00:c9:9f:c7:1e;20:00:00:00:c9:9f:c7:1e; na
    Fabric Port Name: 20:00:00:05:1e:02:aa:c1
    Permanent Port Name: 10:00:00:00:c9:9f:c7:1e
    Port Index: 0
    Share Area: No
    Device Shared in Other AD: No
    Redirect: No
 N    010001;    2,3;c0:50:76:03:03:9e:00:0a;c0:50:76:03:03:9e:00:0a; na
    Fabric Port Name: 20:00:00:05:1e:02:aa:c1
    Permanent Port Name: 10:00:00:00:c9:9f:c7:1e
    Port Index: 0
    Share Area: No
    Device Shared in Other AD: No
    Redirect: No
 N    010002;    2,3;c0:50:76:03:03:98:00:0e;c0:50:76:03:03:98:00:0e; na
    Fabric Port Name: 20:00:00:05:1e:02:aa:c1
    Permanent Port Name: 10:00:00:00:c9:9f:c7:1e
    Port Index: 0
    Share Area: No
    Device Shared in Other AD: No
    Redirect: No
 N    010100;    2,3;10:00:00:00:c9:9f:c7:1f;20:00:00:00:c9:9f:c7:1f; na
    Fabric Port Name: 20:01:00:05:1e:02:aa:c1
    Permanent Port Name: 10:00:00:00:c9:9f:c7:1f
    Port Index: 1
    Share Area: No
    Device Shared in Other AD: No
```

```
    Redirect: No
...
```

Use the **lshostconnect -login** command from the DS8000 DS CLI to see that the IBM i virtual Fibre Channel adapter with WWPN C05076030398000E is logged in to DS8000 host adapter port I0201 as shown in Example 9-11.

Example 9-11 DS8000 DSCLI displaying the logged in host initiators

```
dscli> lshostconnect -login
Date/Time: 17. Dezember 2010 17:23:07 CET IBM DSCLI Version: 6.5.15.19 DS:
IBM.2107-75BALB1
WWNN             WWPN             ESSIOport LoginType Name          ID
===============================================================================
...
C0507603039E000A C0507603039E000A I0201     SCSI      AIX_NPIV_1    000F
C05076030398000E C05076030398000E I0201     SCSI      IBMi_NPIV     0014
20000000C99FC3F6 10000000C99FC3F6 I0201     SCSI      P7_2_vios1_1  0004
...
```

9.3.3 Linux virtual storage configuration tracing

Virtual SCSI disks in a Linux client partition can be traced by completing the following steps:

1. On a Linux client partition, use the **lsscsi** command to display the information about virtual SCSI disks as shown in Example 9-12. In this example, [1:0:1:1] means that sda is Host: scsi1, Channel: 00, Target: 01, and LUN: 00.

   Example 9-12 List of SCSI disks

```
[root@Power7-2-RHEL ~]# lsscsi -v
[1:0:1:0]    disk    AIX      VDASD            0001  /dev/sda
  dir: /sys/bus/scsi/devices/1:0:1:0
[/sys/devices/vio/30000036/host1/target1:0:1/1:0:1:0]
[2:0:1:0]    disk    AIX      VDASD            0001  /dev/sdb
  dir: /sys/bus/scsi/devices/2:0:1:0
[/sys/devices/vio/30000037/host2/target2:0:1/2:0:1:0]
[3:0:0:0]    disk    IBM      2107900          .278  /dev/sdc
  dir: /sys/bus/scsi/devices/3:0:0:0
[/sys/devices/vio/30000038/host3/rport-3:0-0/target3:0:0/3:0:0:0]
[4:0:0:0]    disk    IBM      2107900          .278  /dev/sdd
  dir: /sys/bus/scsi/devices/4:0:0:0
[/sys/devices/vio/30000039/host4/rport-4:0-0/target4:0:0/4:0:0:0]
[root@Power7-2-RHEL ~]# lsscsi -c
Attached devices:
```

```
Host: scsi1 Channel: 00 Target: 01 Lun: 00
  Vendor: AIX      Model: VDASD            Rev: 0001
     Type:   Direct-Access                    ANSI SCSI revision: 03
Host: scsi2 Channel: 00 Target: 01 Lun: 00
  Vendor: AIX      Model: VDASD            Rev: 0001
  Type:   Direct-Access                    ANSI SCSI revision: 03
Host: scsi3 Channel: 00 Target: 00 Lun: 00
  Vendor: IBM      Model: 2107900          Rev: .278
  Type:   Direct-Access                    ANSI SCSI revision: 05
Host: scsi4 Channel: 00 Target: 00 Lun: 00
  Vendor: IBM      Model: 2107900          Rev: .278
  Type:   Direct-Access                    ANSI SCSI revision: 05
```

a. Example 9-13 shows how to display the information of the location code of the virtual SCSI adapter corresponding to Host: scsi1. This information includes the partition name of the Virtual I/O Server that has the corresponding virtual SCSI host adapter, and the vhost name of the virtual SCSI host adapter.

Example 9-13 Information of scsi1 adapter

```
[root@Power7-2-RHEL ~]# cat /sys/class/scsi_host/host1/vhost_loc
U8233.E8B.061AB2P-V5-C54-T1
[root@Power7-2-RHEL ~]# cat /sys/class/scsi_host/host1/partition_name
P7_2_vios1
[root@Power7-2-RHEL ~]# cat /sys/class/scsi_host/host1/vhost_name
vhost1
```

2. On the Virtual I/O Server, use the **lsmap** command to display the device mapping between physical and virtual devices as shown in Example 9-14. In the example, the backing device of vhost1 has LUN 0x81 and 0x81 corresponds Target: 01 of the sda information in step1. Therefore, the `sda` on the Linux client partition corresponds to `hdisk19` of the Virtual I/O Server.

Example 9-14 Device mapping information

```
$ lsmap -vadapter vhost1
SVSA            Physloc                                     Client Partition ID
--------------- ------------------------------------------- ------------------
vhost1          U8233.E8B.061AB2P-V1-C54                    0x00000001

VTD                   rhel_hd19
Status                Available
LUN                   0x8100000000000000
Backing device        hdisk19
Physloc               U5802.001.0087356-P1-C2-T1-W500507630410412C-L4011401600000000
Mirrored              false
```

For NPIV devices, enter **c**`at /sys/class/scsi_host/hostX/port_loc_code` to display the physical location code of the backing physical Fibre Channel adapter, where the ***X*** stands for SCSI adapter number in step 2.

9.4 Virtual storage monitoring

This section describes how to check and monitor virtual storage health and performance for Virtual I/O Server and virtual I/O client partitions.

This chapter includes the following sections:

9.4.1 Virtual I/O Server storage monitoring

This section explains how to check the storage health and performance on the Virtual I/O Server.

Checking storage health on the Virtual I/O Server

On the Virtual I/O, use the following commands to check its disk status:

`lsvg rootvg`	Check for stale PPs and no stale PV.
`lsvg -pv rootvg`	Check for missing disks.
`lsdev -type disk`	Check all expected disks are available.
`errlog`	Check for disk-related errors in the error log.

If you are using IBM SAN storage with the Virtual I/O Server use these commands:

`lspath`	Check for missing paths. If paths became missing because of a configuration change, clean them up using the **`rmpath`** command.
`mpio_get_config -Av`	Check all SAN storage using MPIO multipathing to ensure that LUNs are detected and paths are established
`pcmpath query device`	Check all SAN storage using SDDPCM multipathing to ensure that LUNs are detected and paths are established. For example, in `OPEN` state, note that SDDPCM will not close the last available path for a missing device.

Example 9-15 shows an excerpt from the SDDPCM command output from a Virtual I/O Server with 23 SAN LUNs attached from an IBM Storwize® V7000 storage system using four active paths to each device that the non-preferred paths marked as *.

Example 9-15 pcmpath query device output for an attached IBM Storwize V7000

```
$ oem_setup_env
# pcmpath query device

Total Dual Active and Active/Asymmetric Devices : 23


DEV#:   0  DEVICE NAME: hdisk0  TYPE: 2145  ALGORITHM:  Load Balance
SERIAL: 60050768028086EDD000000000000000
==========================================================================
Path#      Adapter/Path Name          State     Mode     Select     Errors
    0*           fscsi0/path0           OPEN   NORMAL         40          0
    1            fscsi0/path1           OPEN   NORMAL     490474          0
    2*           fscsi1/path2           OPEN   NORMAL         41          0
    3            fscsi1/path3           OPEN   NORMAL     504215          0

DEV#:   1  DEVICE NAME: hdisk1  TYPE: 2145  ALGORITHM:  Load Balance
SERIAL: 60050768028086EDD000000000000004
==========================================================================
Path#      Adapter/Path Name          State     Mode     Select     Errors
    0*           fscsi0/path0           OPEN   NORMAL         28          0
    1            fscsi0/path1           OPEN   NORMAL       3285          0
    2*           fscsi1/path2           OPEN   NORMAL         28          0
    3            fscsi1/path3           OPEN   NORMAL       4930          0
...
```

Monitoring storage performance on the Virtual I/O Server

The **viostat** command can be helpful in tracing system activity with regards to I/O workload. It allows for relatively fine-grained measurements of different types of adapters and attached disks, and the usage of paths to redundant attached disks, including virtual adapters and virtual disks and their backing devices by using the **-adapter** flag.

Example 9-16 displays the output of a measurement for extended disk I/O performance statistics while disk I/O occurred on one client with a virtual disk.

Example 9-16 Monitoring I/O performance with viostat

```
$ viostat -extdisk
System configuration: lcpu=2 drives=15 paths=22 vdisks=23

hdisk8          xfer:  %tm_act      bps      tps      bread      bwrtn
                         94.7      9.0M    65.9       14.7       9.0M
                read:      rps  avgserv  minserv  maxserv   timeouts      fails
                           0.0      1.6      0.1      5.0           0          0
               write:      wps  avgserv  minserv  maxserv   timeouts      fails
                          65.8     27.0      0.2     3.8S           0          0
               queue:  avgtime  mintime  maxtime  avgwqsz    avgsqsz     sqfull
                           0.0      0.0     1.8S      0.0        0.9        0.0
hdisk11         xfer:  %tm_act      bps      tps      bread      bwrtn
                         94.8      9.0M    64.7       14.7       9.0M
                read:      rps  avgserv  minserv  maxserv   timeouts      fails
                           0.0      2.0      0.1     10.6           0          0
               write:      wps  avgserv  minserv  maxserv   timeouts      fails
                          64.7     27.7      0.2     3.5S           0          0
               queue:  avgtime  mintime  maxtime  avgwqsz    avgsqsz     sqfull
                           0.0      0.0    263.5      0.0        0.9        0.0
hdisk9          xfer:  %tm_act      bps      tps      bread      bwrtn
                           0.2      0.0      0.0        0.0        0.0
                read:      rps  avgserv  minserv  maxserv   timeouts      fails
                           0.0      0.0      0.0      0.0           0          0
               write:      wps  avgserv  minserv  maxserv   timeouts      fails
                           0.0      0.0      0.0      0.0           0          0
               queue:  avgtime  mintime  maxtime  avgwqsz    avgsqsz     sqfull
                           0.0      0.0      0.0      0.0        0.0        0.0
hdisk6          xfer:  %tm_act      bps      tps      bread      bwrtn
                           2.6    134.6K     8.9      118.5K      16.1K
                read:      rps  avgserv  minserv  maxserv   timeouts      fails
                           6.5      6.6      0.1     1.9S           0          0
               write:      wps  avgserv  minserv  maxserv   timeouts      fails
                           2.3      0.9      0.2    220.5           0          0
               queue:  avgtime  mintime  maxtime  avgwqsz    avgsqsz     sqfull
                           0.2      0.0     44.2      0.0        0.0        0.7
```

9.4.2 AIX virtual I/O client storage monitoring

This section explains how to check the storage health and performance on the AIX virtual I/O client.

Checking storage health on the AIX virtual I/O client

This section differentiates between checking the storage health on an AIX virtual I/O client in an MPIO and an LVM mirroring environment.

AIX virtual I/O client MPIO environment

The following procedure applies to a dual Virtual I/O Server environment by using MPIO on the AIX virtual I/O client partition, as shown in Figure 9-13.

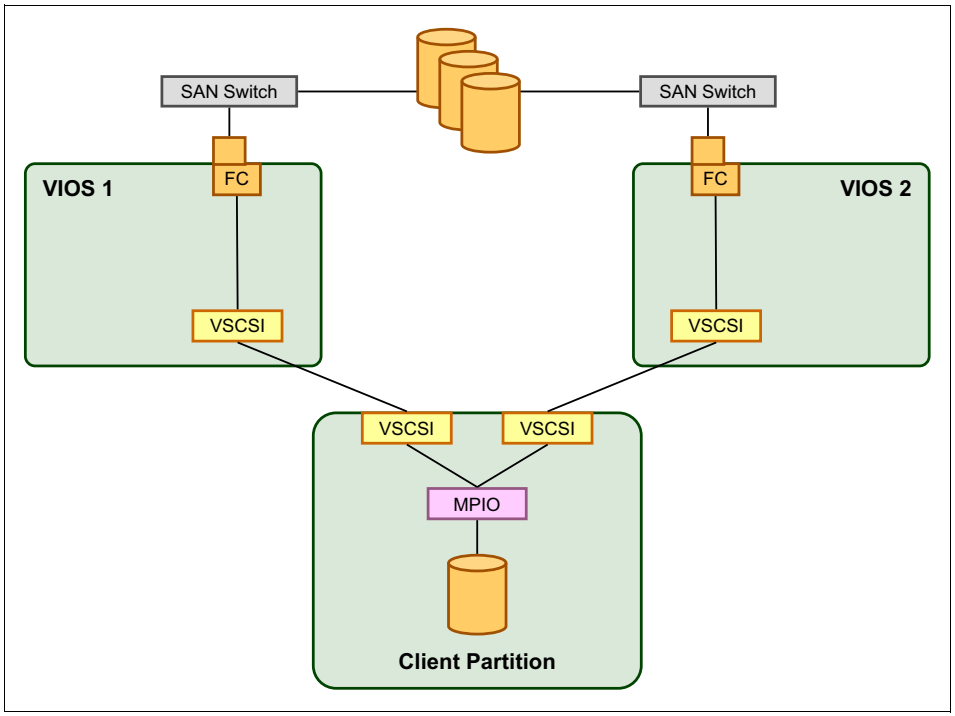


Figure 9-13 AIX virtual I/O client using MPIO

Run the following commands on the AIX virtual I/O client to check its storage health:

`lspath` Check all the paths to the disks to make sure they are all in the enabled state as shown in Example 9-17.

Example 9-17 AIX lspath command output

```
# lspath
Enabled hdisk0 vscsi0
Enabled hdisk0 vscsi1
```

lsattr -El hdisk0 Verify the MPIO heartbeat for hdisk0. The attribute hcheck_mode must be set to nonactive, and hcheck_interval is 60. If you run IBM SAN storage, check that reserve_policy is no_reserve. Other storage vendors might require other values for reserve_policy. Example 9-18 shows the output of the **lsattr** command.

Example 9-18 AIX client lsattr command that shows hdisk attributes

```
# lsattr -El hdisk0
PCM             PCM/friend/vscsi                 Path Control Module        False
algorithm       fail_over                        Algorithm                  True
hcheck_cmd      test_unit_rdy                    Health Check Command       True
hcheck_interval 60                               Health Check Interval      True
hcheck_mode     nonactive                        Health Check Mode          True
max_transfer    0x40000                          Maximum TRANSFER Size      True
pvid            00c1f170e327afa70000000000000000 Physical volume identifier False
queue_depth     3                                Queue DEPTH                True
reserve_policy  no_reserve                       Reserve Policy             True
```

If the Virtual I/O Server was rebooted earlier and the health_check attribute is not set, you might need to enable it. There are instances when the path shows up as failed even though the path to the Virtual I/O Server is up.

The way to correct this is to set the hcheck_interval and hcheck_mode attributes by using the **chdev** command as shown in Example 9-19.

Example 9-19 Using the chdev command to set hdisk recovery parameters

```
# chdev -l hdisk0 -a hcheck_interval=60 -a hcheck_mode=nonactive -P
```

lsvg -p rootvg Check for a missing hdisk.

AIX virtual I/O client LVM mirroring environment

This section addresses storage health monitoring for an AIX virtual I/O client LVM software mirroring environment as shown in Figure 9-14.

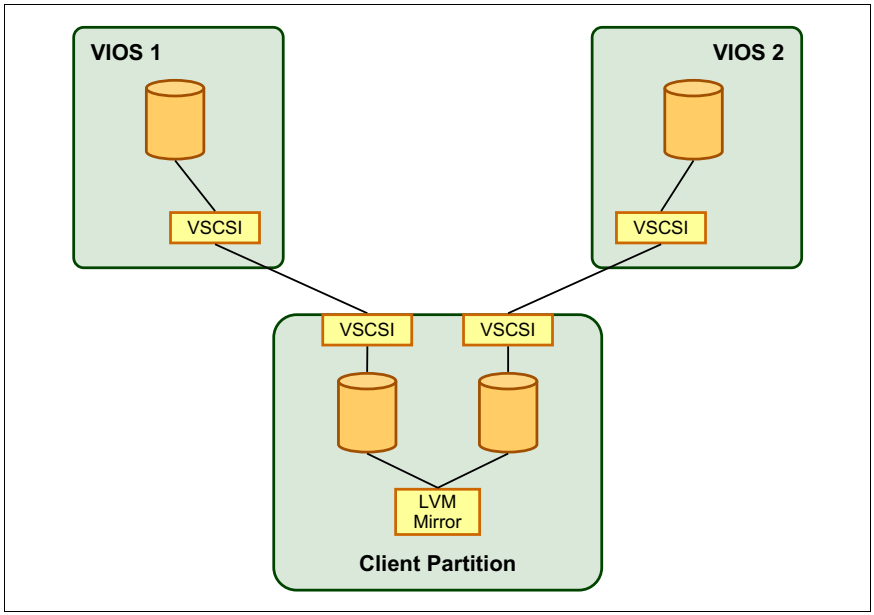


Figure 9-14 AIX virtual I/O client using LVM mirroring

Check for any missing disks by using the **`lsvg -p rootvg`** command as shown in Example 9-20.

Example 9-20 Checking for any missing disks

```
# lsvg -p rootvg
rootvg:
PV_NAME           PV STATE          TOTAL PPs   FREE PPs    FREE
DISTRIBUTION
hdisk0            active            511         488
102..94..88..102..102
hdisk1            missing           511         488
102..94..88..102..102
```

If one disk is missing, there will be stale partitions that can be verified by using the **`lsvg rootvg`** command. If there are stale partitions and all disks are available, resynchronize the mirror by using the **`varyonvg`** command and the **`syncvg -v`** command on the volume groups that use virtual disks from the Virtual I/O Server environment.

Example 9-21 shows the output of these commands.

Example 9-21 AIX command to recover from stale partitions

```
# varyonvg rootvg
# syncvg -v rootvg
# lsvg -p rootvg
rootvg:
PV_NAME           PV STATE          TOTAL PPs   FREE PPs    FREE DISTRIBUTION
hdisk0            active            511         488         102..94..88..102..102
hdisk1            active            511         488         102..94..88..102..102
# lsvg rootvg
VOLUME GROUP:       rootvg                   VG IDENTIFIER:  00c478de00004c00000
00006b8b6c15e
VG STATE:           active                   PP SIZE:        64 megabyte(s)
VG PERMISSION:      read/write               TOTAL PPs:      1022 (65408 megabytes)
MAX LVs:            256                      FREE PPs:       976 (62464 megabytes)
LVs:                9                        USED PPs:       46 (2944 megabytes)
OPEN LVs:           8                        QUORUM:         1
TOTAL PVs:          2                        VG DESCRIPTORS: 3
STALE PVs:          0                        STALE PPs:      0
ACTIVE PVs:         2                        AUTO ON:        yes
MAX PPs per VG:     32512
MAX PPs per PV:     1016                     MAX PVs:        32
LTG size (Dynamic): 256 kilobyte(s)          AUTO SYNC:      no
HOT SPARE:          no                       BB POLICY:      relocatable
```

For a configuration with many AIX virtual I/O client partitions, it is time-consuming and error-prone to check the storage health individually.

For this reason, a sample script that uses a distributed shell for automating the health checking and recovery for a group of AIX virtual I/O client partitions is provided in Appendix A, “AIX disk and NIB network checking and recovery script” on page 699.

Monitoring storage performance on the AIX virtual I/O client

The **iostat** command can help trace system activity with regards to I/O workload. It allows for relatively fine-grained measurements of different types of adapters and attached disks, and the usage of paths to redundant attached disks. The output of a measurement while disk I/O occurred on hdisk0 disk is shown in Example 9-22.

Example 9-22 Monitoring disk performance with iostat

```
# iostat -d hdisk0 2

System configuration: lcpu=2 drives=2 paths=3 vdisks=3
```

```
Disks:        % tm_act     Kbps      tps    Kb_read   Kb_wrtn
hdisk0           0.0       0.0       0.0          0         0
hdisk0           0.0       0.0       0.0          0         0
hdisk0          85.2     538.8     589.5        614       512
hdisk0          97.4     744.9     709.2        692       768
hdisk0          90.7     673.2     672.2        693       724
hdisk0          92.9     723.1     704.6        718       768
hdisk0         100.0     654.5     674.0        669       640
hdisk0         100.0     669.5     704.0        699       640
```

9.4.3 IBM i virtual I/O client storage monitoring

This section describes how to monitor the storage health and performance from an IBM i virtual I/O client.

Checking storage health on the IBM i virtual I/O client

There are two things to check on the IBM i virtual I/O client about storage health:

- Verify that the used capacity does not reach the auxiliary storage pool (ASP) limit.
- Verify that all disk units and paths are available.

To check the used storage capacity in the system ASP (ASP 1), use the WRKSYSSTS command to check the value that is displayed for `% system ASP used`. For user ASPs, the used capacity can be displayed from SST by running STRSST and selecting `3. Work with disk units` → `1. Display disk configuration` → `2. Display disk configuration capacity`.

By default, the ASP usage threshold is 90%. If this threshold is reached, it causes a system operator message `CPF0907` notification. User ASPs reaching 100% used capacity will, by default, overflow into the system ASP. Independent ASPs (IASPs) running out of capacity do not overflow into SYSBAS, but might get varied off.

A system ASP becoming filled up is likely to cause a system crash. Therefore, it is important to monitor IBM i storage usage and take preventive measures like data housekeeping and adding more storage.

The RTVDSKINF command submitted as a batch job collects diverse detailed disk capacity usage information about system and user data in the QAEZDISK database file. This file which can be printed as a disk space report for libraries and object owners by using the PRTDSKINF command.

For more information, see *Using the RTVDSKINF command to avoid disk storage disasters on IBM i* at:

https://www.ibm.com/developerworks/ibmi/library/i-using-rtvdskinf-command/index.html

To check whether all disk units are available, run STRSST to log in to System Service Tools. Select `3. Work with disk units` → `1. Display disk configuration` → `1. Display disk configuration status`. Check whether all disk units are shown with a status of *Active* as shown in Figure 9-15. The ASP status is either reported as *Unprotected* when not using IBM i mirroring, or *Mirrored* when you use IBM i mirroring.

```
                    Display Disk Configuration Status

          Serial                        Resource                    Hot Spare
ASP Unit  Number          Type Model    Name       Status           Protection
  1                                               Unprotected
      1   50-1003001      2107  A04     DMP001     RAID 5/Active         N
      2   50-1004001      2107  A04     DMP002     RAID 5/Active         N
      3   50-1102001      2107  A04     DMP003     RAID 5/Active         N
      4   50-1103001      2107  A04     DMP004     RAID 5/Active         N



Press Enter to continue.

F3=Exit        F5=Refresh         F9=Display disk unit details
F11=Disk configuration capacity   F12=Cancel
```

Figure 9-15 IBM i SST Display Disk Configuration Status panel

If there are units that are shown as `suspended`, `not connected`, or `missing from the configuration`, resolve the problems:

- If you are using virtual SCSI attachment, check whether the corresponding volumes are available and accessible on the Virtual I/O Server.
- If you are using NPIV attachment, verify the SAN switch zoning and ensure that the volumes are in good status on the storage system.

  For more information about how to identify the corresponding IBM i volumes on the Virtual I/O Server and storage system, see 9.3.2, “IBM i virtual storage configuration tracing” on page 255.
- For suspended units, determine whether the IBM i mirroring state changes to resuming. If they do not, resume the units manually by using the SST function

`3. Work with disk units` → `3. Work with disk unit recovery` → `4. Resume mirrored protection`.

To check whether all disk unit *paths* are available, from SST, select `3. Work with disk units` → `1. Display disk configuration` → `9. Display disk path status`. Check if all disk units paths are *active* as shown in Figure 9-16. Correct any paths in *failed* status by verifying the connections from the IBM i client through the Virtual I/O Server to the backing storage devices. Reconnect paths in *missing* status or, if they were intentionally removed by a configuration change, remove them from the IBM i configuration by using the MULTIPATHRESETTER macro.

```
                        Display Disk Path Status

          Serial                          Resource   Path
ASP Unit  Number          Type Model Name       Status
  1    1  50-1003001      2107  A04  DMP001     Active
                                     DMP005     Active
  1    2  50-1004001      2107  A04  DMP002     Active
                                     DMP006     Active
  1    3  50-1102001      2107  A04  DMP003     Active
                                     DMP007     Active
  1    4  50-1103001      2107  A04  DMP004     Active
                                     DMP008     Active

Press Enter to continue.

F3=Exit         F5=Refresh          F9=Display disk unit details
F11=Display encryption status       F12=Cancel
```

Figure 9-16 IBM i SST Display Disk Path Status panel

Monitoring storage performance on the IBM i virtual I/O client

This section provides examples of IBM i storage monitoring using native IBM i tools.

Real-time storage monitoring on IBM i

For nearly real-time storage monitoring on IBM i, you can use the WRKDSKSTS command as shown in Figure 9-17.

```
                  Size    %     I/O   Request   Read  Write   Read  Write    %
Unit  Type         (M)  Used    Rqs  Size (K)    Rqs   Rqs     (K)   (K)  Busy
   1  6B22       19088  76.0  125.1       5.7   30.2  94.9    4.5   6.1     3
   1  6B22       19088  76.0  120.5       7.1   40.0  80.5    4.5   8.3     2
   2  6B22       19088  75.5  213.4       5.2  121.0  92.3    4.5   6.2     4
   2  6B22       19088  75.5  184.1       5.4  124.6  59.5    4.5   7.4     3
```

Figure 9-17 IBM i WRKDSKSTS command output

The current disk I/O workload statistics are shown for each IBM i disk unit. Select `F5=Refresh` to update the display with current information. Select `F10=Restart statistics` to restart the statistics from the last displayed time.

Collection Services on IBM i is the core component for system-wide performance data collection at specified intervals (the default is 15 minutes). It is enabled by default. To administer Collection Services, use the commands CFGPFRCOL, STRPFRCOL, and ENDPFRCOL. Alternatively, you can use the menu interface of the IBM Performance Tools for IBM i licensed program. To obtain information about IBM i disk performance statistics collection services, data from the QAPMDISK database file must be analyzed. This can be done by using either native SQL commands or, more easily, by using IBM Performance Tools for IBM i as described in the following section.

For more information about IBM i Performance Management tools, see *IBM eServer iSeries Performance Management Tools*, REDP-4026.

Long-term storage monitoring on IBM i

For long-term storage performance monitoring, you can use the IBM Performance Tools for IBM i licensed program (5770-PT1). You can use this program to generate various reports from QAPM* performance database files created from Collection Services data.

IBM Performance Tools for IBM i functions are accessible on IBM i 5250 sessions through a menu by using the STRPFRT or GO PERFORM commands. You can

also use the native CL commands like PRTSYSRPT and PRTRSCRPT. Example 9-23 shows a system report for disk utilization that was created by using the following command:

```
PRTSYSRPT MBR(Q306160006) TYPE(*DISK)
```

Example 9-23 IBM i System Report for Disk Utilization (PRTSYSRPT)

```
System Report                                    11/01/08 18:43:0
                                                    Disk Utilization                                                    Page 000
Member  . . . : Q306160006 Model/Serial  . : MMA/10-1F170       Main storage . . : 8192.0 MB  Started  . . . . : 11/01/08 16:00:0
  Library . . : QPFRDATA   System name . . : E101F170           Version/Release  :   6/ 1.0   Stopped  . . . . : 11/01/08 18:00:0
Partition ID  : 005        Feature Code  . : 5622-5622          Int Threshold  . :    .00 %
Virtual Processors:   2    Processor Units :   1.00
         Unit              Size  IOP  IOP        Dsk CPU --Percent--  Op Per    K Per    - Average Time Per I/O --
  Unit   Name       Type   (M)   Util Name       Util    Full   Util  Second     I/O     Service   Wait   Response
  ----   ---------- ---- -------- ---- ---------- ------- ----   ---- -------- ---------  -------  ------  --------
 ASP ID/ASP Rsc Name:  1/
  0001A  DD004      6B22  19,088   .0 CMB02           .0 94.1   54.1   157.34      14.8    .0052   .0000     .0052
  0001B  DD003      6B22  19,088   .0 CMB01           .0 94.1   59.8   251.37      10.5    .0035   .0000     .0035
  0002A  DD002      6B22  19,088   .0 CMB02           .0 94.1   59.1   168.70      14.1    .0055   .0000     .0055
  0002B  DD001      6B22  19,088   .0 CMB01           .0 94.1   63.3   260.78      10.3    .0038   .0000     .0038
Total for ASP ID:  1      76,352
Average                                                  94.1   59.0   205.28      12.1    .0043   .0000     .0043
Total                     76,352
Average                                                  94.1   59.0   205.28      12.1    .0043   .0000     .0043
```

Example 9-24 shows a resource report for disk utilization that was created by using the following command:

```
PRTRSCRPT MBR(Q306160006) TYPE(*DISK)
```

Example 9-24 IBM i Resource Report for Disk Utilization (PRTRSCRPT)

```
Resource Interval Report                         11/01/08 18:36:1
                                                 Disk Utilization Summary                                          Page
Member  . . . : Q306160006 Model/Serial  . : MMA/10-1F170       Main storage  . . : 8192.0 MB Started . . . : 11/01/08 16:00:07
  Library . . : QPFRDATA   System name . . : E101F170           Version/Release . :     6/1.0 Stopped . . . : 11/01/08 18:00:00
Partition ID  : 005        Feature Code  . : 5622-5622
                           Average      Average       Average                          High      High      High       Disk
 Itv       Average         Reads        Writes        K Per       Avg       High       Util      Srv       Srv        Space
 End       I/O /Sec        /Sec         /Sec          I/O         Util      Util       Unit      Time      Unit     Used (GB)
-----      ---------       --------     --------      -------     -----     -----      ----      -----     ----    -----------
16:05        1,005.7         886.6        119.0         8.5       59.7      65.4      0002B     .0047     0002A       66.327
16:10          769.1         510.6        258.4        16.8       71.9      77.4      0002B     .0115     0002A       66.018
16:15          656.1         630.2         25.9         7.3       51.8      53.0      0002B     .0042     0002A       66.634
16:20          560.3         479.7         80.5        11.0       55.9      61.1      0002B     .0065     0002A       66.912
16:25          691.0         469.7        221.3        15.8       70.0      74.7      0002B     .0118     0001A       66.110
16:30          722.5         679.8         42.7         9.1       52.9      54.7      0002B     .0042     0002A       66.954
16:35          728.0         649.6         78.4         9.4       55.4      60.6      0002B     .0050     0002A       66.905
16:40          666.1         478.2        187.8        20.7       63.2      70.2      0002B     .0140     0002A       66.940
16:45        1,032.1         972.6         59.5         6.4       53.0      55.8      0002B     .0030     0002A       66.517
16:50          765.3         596.8        168.5        16.1       61.5      69.4      0002B     .0105     0001A       65.983
                                                                                                                      More...
---------      --------      --------     -------      -----
Average:       795.4         677.9        117.4        12.1       59.0
```

For monitoring IBM i storage performance, it is useful to look at a system report and a resource report from the same time frame. The system report provides information about the average disk response time (= service time + wait time). The resource report shows the amount of I/O workload for the time frame.

If the resource report shows changing workload characteristics in terms of I/O throughput or I/O transfer size, generate a separate system report for each workload period to get a better view of the disk performance for the workloads. For example, it might show small I/O transfers typical for interactive workloads during the day, and large I/O transfers typical for batch or save workloads during the night.

For more information about IBM Performance Tools for IBM i, see *IBM eServer iSeries Performance Tools for iSeries*, SC41-5340.

The *IBM Systems Director Navigator for i* graphical user interface can be used to graphically display long-term monitoring data for disk performance. This GUI is available from the IBM i partition using `http://IBM_i_server_IP_address:2001`.

To generate the disk overview graph that is shown in Figure 9-18, click **IBM i Management** → **Performance** → **Collections**. Then, select a collection services DB file, and select **Investigate Data**. Then click **Select Action** → **Disk Overview for System Disk Pool** for the diagram.

In the example, the information for the I/O workload was added by using the **Add data series** option from the **Select Action** menu. Look at disk performance in context of the I/O workload.

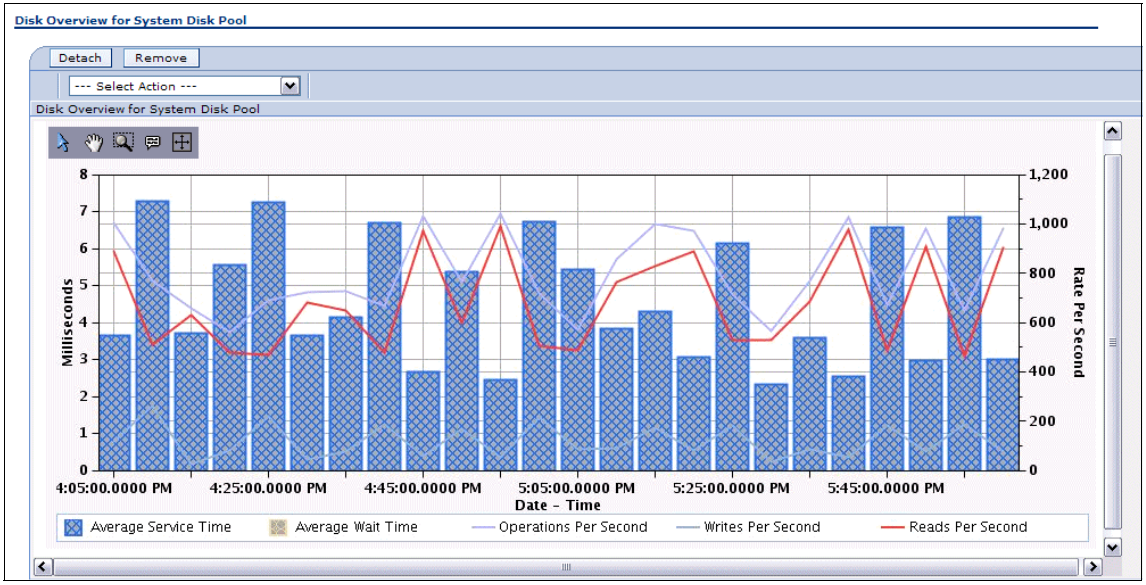


Figure 9-18 IBM i Navigator Disk Overview for System Disk Pool

Another approach for long-term disk monitoring for IBM i that also allows IBM i cross-partition monitoring is to use the System i Navigator's Management

Central monitors function. There is no metric for disk response time. However, based on the experienced average disk arm usage, you can define a threshold at which to be notified about unusually high utilization so that you can be alerted to potential disk performance problems.

For more information about using System i Navigator for performance monitoring, see *Managing OS/400 with Operations Navigator V5R1 Volume 5: Performance Management*, SG24-6565.

9.4.4 Linux virtual I/O client storage monitoring

I/O activity across internal and external disks can be monitored by using the **iostat** command, which is included in the sysstat package in Linux distributions. A brief description of the sysstat utility can be found in “sysstat utility” on page 452. You can obtain the I/O transfer rate and other I/O statistics by using the **iostat** command. Sample output is shown in Example 9-25.

Example 9-25 Linux iostat command output showing I/O activity

```
[root@p750_lpar02 ~]# iostat
Linux 2.6.32-279.el6.ppc64 (p750_lpar02)        12/19/2012      _ppc64_ (4 CPU)

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
          31.69    0.25   43.09    0.00   24.93    0.03

Device:            tps   Blk_read/s   Blk_wrtn/s   Blk_read   Blk_wrtn
scd0              0.00         0.00         0.00         64          0
sda               0.45         1.05         7.91     735764    5525546
sdb               0.00         0.01         0.00       8320          0
dm-0              1.02         1.00         7.91     697730    5525288
dm-1              0.00         0.01         0.00       7296          0
```

When you run the **iostat** command as shown in Example 9-25, the utility generates a report of accumulated I/O statistics on all disk devices since boot time.

You can also use the **-d** flag to monitor the I/O activity of specific disks within a specified interval as shown in Example 9-26.

Example 9-26 Linux iostat output with -d flag and a 5-second interval

```
[root@p750_lpar02 ~]# iostat -d sda 5 2
Linux 2.6.32-279.el6.ppc64 (p750_lpar02)        12/19/2012      _ppc64_ (4 CPU)

Device:            tps   Blk_read/s   Blk_wrtn/s   Blk_read   Blk_wrtn
sda               0.45         1.06         7.92     743404    5545146
```

```
Device:            tps   Blk_read/s   Blk_wrtn/s   Blk_read   Blk_wrtn
sda               0.40         0.00         4.80          0         24
```

10

Shared storage pools

This chapter covers enhancements and management and monitoring tasks for the shared storage pools feature in Virtual I/O Server Version 2.2.2.0 and later.

Shared storage pools are a feature in PowerVM Standard and Enterprise Editions that was introduced in Virtual I/O Server Version 2.2.0.11 Fix Pack 11 Service Pack 1. It is a server-based storage virtualization that provides distributed storage access to Virtual I/O Server partitions for their client partitions.

For more information about shared storage pools, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

This chapter includes the following sections:

- Managing shared storage pools
- Monitoring shared storage pools

10.1 Managing shared storage pools

Virtual I/O Server Version 2.2.2.0 has new enhancements that were not available in previous releases. The following enhancements address cluster outages:

- *Rolling updates* allow software updates to be applied to the Virtual I/O Server partitions without causing an outage in the entire cluster.
- *Repository resiliency* allows the replacement of the repository disk without causing an outage in the entire cluster.

Important: The new shared storage pool enhancements in Virtual I/O Server Version 2.2.2.0 cannot be used until all the Virtual I/O Server partitions being used in the cluster are at Virtual I/O Server Version 2.2.2.0 or later.

10.1.1 Scalability enhancements for shared storage pools

Table 10-1 lists the scalability of cluster components in Virtual I/O Server Version 2.2.2.0 and later.

Table 10-1 Scalability in Virtual I/O Server version 2.2.2.0

Feature	Minimum	Maximum
Number of VIOS nodes in cluster	1	16
Number of physical disks in pool	1	1024
Number of virtual disks (LUs) mappings in pool	1	8192
Number of client LPARs per VIOS node	1	200
Capacity of physical disks in pool	5 GB	16 TB
Storage capacity of storage pool	10 GB	512 TB
Capacity of a virtual disk (LU) in pool	1 GB	4 TB
Number of repository disks	1	1

10.1.2 Managing nodes in a cluster

This section describes cluster requirements and management tasks. To take advantage of the shared storage pool enhancements, you must be using Virtual I/O Server version 2.2.2.0 or later.

Cluster Requirements

The following list of minimum requirements must be met before either creating a cluster or adding a node to a cluster:

- Managed system:
  - POWER6 processor-based or later.
- Per node:
  - Processor: One core
  - Memory: 4 GB
  - Adapters: One Fibre Channel
  - Disks:
    - Storage area network (SAN) disk
    - RAID protected
    - One 10 GB repository disk
    - One 10 GB pool disk
  - Network:
    - The Virtual I/O Server partition must have uninterrupted network connectivity for shared storage pool operations. You can use an Integrated Virtual Ethernet (IVE) adapter. Virtual local area network (VLAN) tagging is supported in Virtual I/O Server Version 2.2.2.0 or later.
    - A cluster on Virtual I/O Server version 2.2.2.0 or later supports either IPV4 or IPV6, but not a combination of both.
    - The forward and reverse DNS lookups for a Virtual I/O Server partition host name must resolve to the same IP address.

Note: All storage devices to be allocated in the shared storage pool must be on a storage system that is configured with RAID protection for redundancy.

Creating a cluster

The physical volumes that are used in a cluster must not be in use by any other volume group or cluster. You can list the physical volumes available for use when creating the cluster by using the **`lspv -free`** command.

A shared storage pool cluster is created by running the **`cluster`** command and specifying the following attributes:

- Cluster name
- Pool name

- Repository physical volume
- Pool physical volumes

Example 10-1 shows creating a cluster with the following attributes:

- Cluster name: *ssp_cluster*
- Pool name: *ssp_pool*
- Repository physical volume: *hdisk12*
- Pool physical volumes: *hdisk8* and *hdisk9*

Example 10-1 Creating the cluster with one node

```
$ cluster -create -clustername ssp_cluster -repopvs hdisk12 -spname ssp_pool
-sppvs hdisk8 hdisk9 -hostname p71vios1
Cluster ssp_cluster has been created successfully.
```

The creation of the cluster can be verified by running the **cluster** command with the **-list** option, as shown in Example 10-2.

Example 10-2 Listing the cluster information

```
$ cluster -list
CLUSTER_NAME:    ssp_cluster
CLUSTER_ID:      a13bdaa23ff511e2bb4ef6dd51518384
```

For more information about creating a cluster, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

Adding a node to a cluster

A cluster node is added by running the **cluster -addnode** command on any node in the cluster, as shown in Example 10-3.

Example 10-3 Adding nodes to a cluster

```
$ cluster -addnode -clustername ssp_cluster -hostname p71vios2
Partition p71vios2 has been added to the ssp_cluster cluster.

$ cluster -addnode -clustername ssp_cluster -hostname p72vios1
Partition p72vios1 has been added to the ssp_cluster cluster.

$ cluster -addnode -clustername ssp_cluster -hostname p72vios2
Partition p72vios2 has been added to the ssp_cluster cluster.
```

You can check the status of the cluster and its nodes by running the **cluster** command with the **-status** option. The example cluster consists of four Virtual

I/O Server partitions running on two POWER7 managed systems, as shown in Example 10-4.

Example 10-4 Checking the status of the cluster

```
$ cluster -status -clustername ssp_cluster
Cluster Name         State
ssp_cluster          OK

    Node Name      MTM                   Partition Num    State    Pool State
    p71vios1       8233-E8B0210DD51P     1                OK       OK
    p71vios2       8233-E8B0210DD51P     7                OK       OK
    p72vios1       8233-E8B02061AB2P     1                OK       OK
    p72vios2       8233-E8B02061AB2P     2                OK       OK
```

The graphical representation of the example cluster is shown in Figure 10-1.

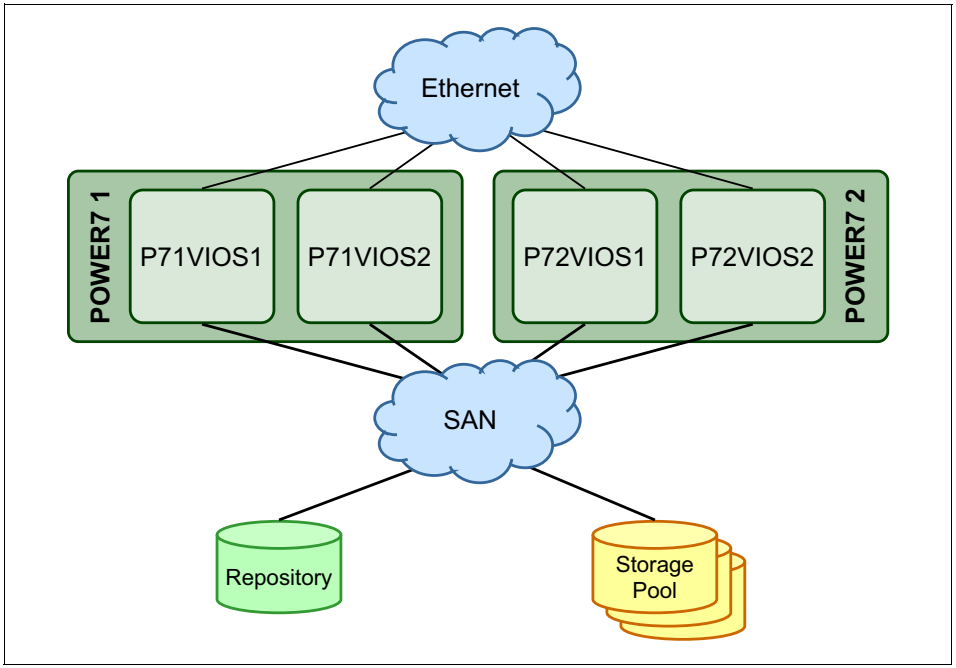


Figure 10-1 Abstraction of SSP cluster

Stopping and starting the cluster service

The cluster can be stopped or started by running the **clstartstop** command with the **-stop** or **-start** options.

You can stop or start the cluster service on all the nodes in the cluster by using the **-a** option.

You can stop or start the cluster service on a specific node in the cluster by using the **-m** option with Virtual I/O Server partition host name as shown in Example 10-5.

Example 10-5 Stopping and starting a node

```
$ cluster -status -clustername ssp_cluster
Cluster Name         State
ssp_cluster          OK

    Node Name     MTM                 Partition Num  State  Pool State
    p71vios1      8233-E8B0210DD51P   1              OK     OK
    p71vios2      8233-E8B0210DD51P   7              OK     OK
    p72vios1      8233-E8B02061AB2P   1              OK     OK
    p72vios2      8233-E8B02061AB2P   2              OK     OK

$ clstartstop -stop -n ssp_cluster -m p71vios1

$ cluster -status -clustername ssp_cluster
Cluster Name         State
ssp_cluster          DEGRADED

    Node Name     MTM                 Partition Num  State  Pool State
    p71vios1      8233-E8B0210DD51P   1              DOWN
    p71vios2      8233-E8B0210DD51P   7              OK     OK
    p72vios1      8233-E8B02061AB2P   1              OK     OK
    p72vios2      8233-E8B02061AB2P   2              OK     OK

$ clstartstop -start -n ssp_cluster -m p71vios1

$ cluster -status -clustername ssp_cluster
Cluster Name         State
ssp_cluster          OK

    Node Name     MTM                 Partition Num  State  Pool State
    p71vios1      8233-E8B0210DD51P   1              OK     OK
    p71vios2      8233-E8B0210DD51P   7              OK     OK
    p72vios1      8233-E8B02061AB2P   1              OK     OK
    p72vios2      8233-E8B02061AB2P   2              OK     OK
```

Removing a node from a cluster

A Virtual I/O Server partition cannot be removed from a cluster until the logical units mapped to it are removed. For more information about removing logical units, see 10.1.7, “Unmapping and removing logical units” on page 299.

You can list the logical unit mapping in the cluster by running the **lsmap -clustername** command. In Example 10-6, the logical unit *sspdisk01* is mapped to the client partition *ID3* on Virtual I/O Server partitions *ID1* and *ID7*.

Example 10-6 Listing the logical unit mapping in the cluster

```
$ lsmap -clustername ssp_cluster -all
Physloc                                                          Client
Partition ID
---------------------------------------------------------------  ------------
U8233.E8B.10DD51P-V1-C13                                         0x00000003

VTD                  vtscsi0
LUN                  0x8100000000000000
Backing device       sspdisk01.a28b5064df887005d85bdae34615a383

Physloc                                                          Client
Partition ID
---------------------------------------------------------------  ------------
U8233.E8B.10DD51P-V7-C23                                         0x00000003

VTD                  vtscsi0
LUN                  0x8100000000000000
Backing device       sspdisk01.a28b5064df887005d85bdae34615a383
```

After the mapped logical units for the node are removed, run the **cluster -rmnode** command to remove the Virtual I/O Server partition from the cluster. The **cluster -rmnode** command cannot be run on the node that is being removed. In Example 10-7 the node *p71vios1* is removed from the *ssp_cluster*.

Example 10-7 Removing a Virtual I/O Server partition from the cluster

```
$ cluster -rmnode -clustername ssp_cluster -hostname p71vios1
Partition p71vios1 has been removed from the ssp_cluster cluster

$ cluster -status -clustername ssp_cluster
Cluster Name         State
ssp_cluster          OK

    Node Name        MTM           Partition Num  State  Pool State
    p71vios2         8233-E8B0210DD51P         7  OK     OK
    p72vios1         8233-E8B02061AB2P         1  OK     OK
    p72vios2         8233-E8B02061AB2P         2  OK     OK
```

Important: You cannot restore a node if it is removed by using the **cluster** command with the **-rmnode** option.

Deleting a cluster

A cluster cannot be deleted until all logical units in the shared storage pool are deleted.

You can delete a cluster by running the **cluster -delete** command as shown in Example 10-8.

Example 10-8 Deleting a cluster

```
$ cluster -delete -clustername ssp_cluster
Cluster ssp_cluster has been removed successfully.
```

You can verify that the physical volumes are no longer assigned to the cluster by running the **lspv -free** command.

Important: You cannot restore a cluster if it is removed by using the **cluster** command with the **-delete** option.

10.1.3 Adding physical volumes to the shared storage pool

The following is a list of considerations when you are working with physical volumes:

- Shared storage pools can only contain physical volumes from a SAN storage system.
- The free space in a shared storage pool can be increased only by adding physical volumes or replacing an existing physical volume to the shared storage pool.
- A physical volume can be replaced only by a physical volume with the same or larger size.
- Physical volumes cannot be removed from the shared storage pool.
- Shared storage pools cannot be decreased in size.

Complete these steps to add physical volumes to a shared storage pool:

1. List the physical volumes that can be added to a shared storage pool, as shown in Example 10-9.

Example 10-9 Listing of physical volumes that can be added to an shared storage pool

```
$ lspv -clustername ssp_cluster -capable
PV NAME SIZE(MB)    PVUDID
```

```
hdisk10 51200       3E213600A0B8000291B0800003DFF09E30CF30F1815 FAStT03IBMfcp
hdisk11 51200       3E213600A0B800011463200005C5950C045EA0F1815 FAStT03IBMfcp
```

2. Verify that the physical volume being added is not in another volume group or shared storage pool.
3. Use the **chsp** command with the **-add** option to add a physical volume to the shared storage pool. Example 10-10 shows adding hdisk10 to the shared storage pool.

Example 10-10 Adding the physical volume to the shared storage pool

```
$ chsp -add -clustername ssp_cluster -sp ssp_pool hdisk10
Current request action progress: % 5
Current request action progress: % 5
Current request action progress: % 80
Current request action progress: % 100
```

4. Verify that the physical volume has been added to the shared storage pool by listing all physical volumes in the shared storage pool. You can do this by running the **lspv** command **-clustername** *cluster* and **-sp** *shared storage pool* options, as shown in Example 10-11.

Example 10-11 Listing of the physical volumes in the shared storage pool

```
$ lspv -clustername ssp_cluster -sp ssp_pool
PV NAME SIZE(MB)    PVUDID
hdisk9  51200       3E213600A0B800011463200005C5750C045AD0F1815 FAStT03IBMfcp
hdisk8  51200       3E213600A0B8000291B0800003DFC09E30CC60F1815 FAStT03IBMfcp
hdisk10 51200       3E213600A0B8000291B0800003DFF09E30CF30F1815 FAStT03IBMfcp
```

5. Verify the free space has increased for the shared storage pool by running the **lssp** command, as shown in Example 10-12.

Example 10-12 Listing the shared storage pool

```
$ lssp -clustername ssp_cluster
POOL_NAME:       ssp_pool
POOL_SIZE:       153216
FREE_SPACE:      150562
TOTAL_LU_SIZE:   0
OVERCOMMIT_SIZE: 0
TOTAL_LUS:       0
POOL_TYPE:       CLPOOL
POOL_ID:         FFFFFFFFAC10153C0000000050C11E7B
```

10.1.4 Replacing a disk in the shared storage pool

When you replace a physical volume, the new physical volume must be the same size or larger than the one being replaced. When the new physical volume is added to the shared storage pool, the old physical volume is removed. The **chsp** command is used to replace physical volumes, as shown in Example 10-13.

Example 10-13 Replacing a disk in the shared storage pool

```
$ chsp -replace -clustername ssp_cluster -sp ssp_pool -oldpv hdisk4 -newpv hdisk5
Current request action progress: % 5
Current request action progress: % 20
Current request action progress: % 40
Current request action progress: % 60
Current request action progress: % 80
Current request action progress: % 100
```

10.1.5 Repository resiliency

Repository resiliency is only available in Virtual I/O Server Version 2.2.2.0 and later. Repository resiliency allows the cluster to remain operational even though the repository disk has failed. While the repository disk is in a failed state, all requests for cluster configuration fail, but the cluster remains operational.

You can list the disks with a repository signature by running the command **/usr/lib/cluster/clras lsrepos** as root on a Virtual I/O Server in the cluster as shown in Example 10-14.

Example 10-14 Checking known disk repository signature

```
vios01:/home/padmin # /usr/lib/cluster/clras lsrepos
hdisk4 has a cluster repository signature.
Cycled 10 disks.
Found 1 cluster repository disk.
```

There are two repository modes in Virtual I/O Server Version 2.2.2.0 and later:

- ASSERT mode means the node will fail if it loses access to the repository disk.
- EVENT mode means the node will continue running if it loses access to the repository disk.

You can check the repository mode of your cluster by running the **cluster -status** command with **-verbose** option, as shown in Example 10-15.

Example 10-15 Checking repository mode

```
$ cluster -status -clustername ssp_cluster -verbose
Cluster Name:            ssp_cluster
Cluster Id:              6e1e70e043dc11e283fc661dbccf745a
Cluster State:           OK
Repository Mode:         EVENT
[...]
```

The repository disk can be replaced with another physical volume that is the same size or larger. Example 10-16 shows replacing the repository disk, hdisk11, with hdisk12 by running the **chrepos** command.

Example 10-16 Replacing a repository disk

```
chrepos -n ssp_cluster -r +hdisk11,-hdisk12
```

10.1.6 Creating and mapping logical units

A logical unit provides the backing storage for the virtual volume that is mapped to a client partition. A logical unit can be mapped by the Virtual I/O Server partition to one or more client partitions. Logical units can be assigned to AIX, IBM i, and Linux partitions.

There are two types of logical units:

- Thin:
  - The default logical unit type.
  - Allocation occurs on an as-needed basis.
  - The free space in the shared storage pool is decreased only by the amount of space that is used by the client partitions.
- Thick:
  - Allocation occurs when mapped to the client partition.
  - The free space in the shared storage pool decreases immediately by the size of the logical unit.

Important:

- Logical units cannot be used as paging devices for PowerVM Active Memory Sharing (AMS).
- IBM i 7.1 TR3 or later is required to support thin logical units. Otherwise, the thin logical units are allocated the same as thick logical units.

Complete these steps to create and map logical units to client partitions:

1. Run the **mkbdsp** command to create a logical unit. Thin logical units are created by default. To create a thick logical unit, specify the **-thick** parameter. Example 10-17 shows the creation of a thick logical unit.

Example 10-17 Creating a thick logical unit

```
$ mkbdsp -clustername ssp_cluster -sp ssp_pool 10G -bd sspdisk01
-thick
Lu Name:sspdisk01
Lu Udid:63f00cffbc0629f32b7927510a882e18

$ mkbdsp -clustername ssp_cluster -sp ssp_pool 10G -bd sspdisk02
-thick
Lu Name:sspdisk02
Lu Udid:a72d65e77bab81e115a07561b0e735f0
```

2. Verify the creation of the new logical units by running the **lssp** command, as shown in Example 10-18.

Example 10-18 Listing logical units

```
$ lssp -clustername ssp_cluster -sp ssp_pool -bd
Lu Name          Size(mb)    ProvisionType      %Used Unused(mb)  Lu Udid
sspdisk01        10240       THIN                  0% 10240   a28b5064df887005d85bdae34615a383
sspdisk01        10240       THICK               100% 0       63f00cffbc0629f32b7927510a882e18
sspdisk02        10240       THICK               100% 0       a72d65e77bab81e115a07561b0e735f0
```

Tip: Notice that **sspdisk01** is listed twice. Logical unit names do not have to be unique. However, generally use unique logical unit names for ease of administration. If you decide not to use unique logical unit names, use the Logical unit Unique device identifier (`Lu Udid`) to identify the logical unit.

3. Identify the vhost that is mapped to the client logical partition on the Virtual I/O Server by using the **lsmap -all** command, as shown in Example 10-19.

Example 10-19 Listing mapping locally on Virtual I/O Server partition

```
$ lsmap -all
SVSA            Physloc                                     Client Partition
ID
--------------- ------------------------------------------- ------------------
vhost0          U8233.E8B.10DD51P-V1-C13                    0x00000003

VTD                   NO VIRTUAL TARGET DEVICE FOUND

SVSA            Physloc                                     Client Partition
ID
--------------- ------------------------------------------- ------------------
vhost1          U8233.E8B.10DD51P-V1-C14                    0x00000004

VTD                   NO VIRTUAL TARGET DEVICE FOUND
```

4. Run the **mkbdsp** command on the Virtual I/O Server partition to map the existing logical unit to the client partition. In Example 10-20, an existing logical partition `sspdisk01` is mapped to `vhost0`, which is assigned to client partition ID 3.

Example 10-20 Mapping the logical unit to a vhost adapter

```
$ mkbdsp -clustername ssp_cluster -sp ssp_pool -bd sspdisk01 -vadapter vhost0
Assigning file "sspdisk01" as a backing device.
VTD:vtscsi0
```

The **mkbdsp** command can create logical units and map them to a client partition in one step. In Example 10-21, a new logical unit called `sspdisk03` is created and mapped to `vhost1`.

Example 10-21 Creating and mapping of a logical unit with one command

```
$ mkbdsp -clustername ssp_cluster -sp ssp_pool 20G -bd sspdisk03
-vadapter vhost1 -thick
Lu Name:sspdisk03
Lu Udid:a7d40fdb5238364833efb6ef9078dfbf

Assigning file "sspdisk03" as a backing device.
VTD:vtscsi1
```

5. Verify that the logical units are mapped to the client partition by running the **lsmap -all** command on the Virtual I/O Server partition, as shown in Example 10-22.

Example 10-22 Listing the logical units mapping

```
$ lsmap -all | more
SVSA            Physloc                                    Client Partition
ID
--------------- ------------------------------------------ ------------------
vhost0          U8233.E8B.10DD51P-V1-C13                   0x00000003

VTD                   vtscsi0
Status                Available
LUN                   0x8100000000000000
Backing device        sspdisk01.a28b5064df887005d85bdae34615a383
Physloc
Mirrored              N/A

SVSA            Physloc                                    Client Partition
ID
--------------- ------------------------------------------ ------------------
vhost1          U8233.E8B.10DD51P-V1-C14                   0x00000004

VTD                   vtscsi1
Status                Available
LUN                   0x8100000000000000
Backing device        sspdisk03.a7d40fdb5238364833efb6ef9078dfbf
Physloc
Mirrored              N/A
[...]
```

6. Verify that the cluster recognizes the mapping of the logical units to the client partition by running the **lsmap -cluster** command on another Virtual I/O Server partition in the cluster as shown in Example 10-23.

   Notice that the logical unit sspdisk01 is mapped to two Virtual I/O Server partition IDs V1 and V7 with serial number 10DD51P.

Example 10-23 Listing the logical units that are mapped to the VIOS on the cluster

```
$ lsmap -clustername ssp_cluster -all
Physloc                                                        Client
Partition ID
-------------------------------------------------------------  ------------
U8233.E8B.10DD51P-V1-C13                                       0x00000003

VTD                  vtscsi0
LUN                  0x8100000000000000
Backing device       sspdisk01.a28b5064df887005d85bdae34615a383
```

```
Physloc                                                          Client
Partition ID
---------------------------------------------------------------  ------------
U8233.E8B.10DD51P-V7-C23                                         0x00000003

VTD                  vtscsi0
LUN                  0x8100000000000000
Backing device       sspdisk01.a28b5064df887005d85bdae34615a383

Physloc                                                          Client
Partition ID
---------------------------------------------------------------  ------------
U8233.E8B.10DD51P-V1-C14                                         0x00000004

VTD                  vtscsi1
LUN                  0x8100000000000000
Backing device       sspdisk03.a7d40fdb5238364833efb6ef9078dfbf
```

7. Run the hardware discovery procedure specific to the operating system on the client partition.

   IBM i client partitions discover newly mapped devices automatically when you use the default system value setting QAUTCFG=1.

10.1.7 Unmapping and removing logical units

Logical units can be unmapped from their client partition and removed from the shared storage pool by using the Virtual I/O Server command-line interface.

Unmapping logical units

To unmap logical units, complete these steps:

1. Remove the virtual SCSI disk from the client partition.
2. Verify the virtual target device (VTD) to be unmapped with the **lsmap** command as shown in Example 10-24.

Example 10-24 Verifying the VTD device to be unmapped

```
$ lsmap -vadapter vhost2
SVSA            Physloc                                        Client Partition
ID
--------------- ---------------------------------------------- ------------------
vhost2          U8233.E8B.10DD51P-V1-C15                       0x00000005

VTD                   vtscsi1
Status                Available
LUN                   0x8100000000000000
Backing device        sspdisk02.a72d65e77bab81e115a07561b0e735f0
Physloc
Mirrored              N/A
```

3. Run the **rmbdsp** command with the **-vtd** option to unmap the logical unit from the client partition. Example 10-25 shows the unmapping of `sspdisk02` from `vhost2`, which is assigned to client logical partition `ID 5`.

Example 10-25 Unmapping a logical unit

```
$ rmbdsp -vtd vtscsi1
vtscsi1 deleted

$ lsmap -vadapter vhost2
SVSA            Physloc                                        Client Partition
ID
--------------- ---------------------------------------------- ------------------
vhost2          U8233.E8B.10DD51P-V1-C15                       0x00000005

VTD                   NO VIRTUAL TARGET DEVICE FOUND
```

Removing logical units from a shared storage pool

To remove logical units, complete these steps:

1. Verify the logical unit to be removed by running the **lssp** and **lsmap** commands as shown in Example 10-26.

Example 10-26 Verifying the logical unit to be removed

```
$ lssp -clustername ssp_cluster -sp ssp_pool -bd
Lu Name      Size(mb)  ProvisionType %Used Unused(mb)  Lu Udid
sspdisk01    10240     THIN          0%    10240       a28b5064df887005d85bdae34615a383
sspdisk02    10240     THICK         100%  0           a72d65e77bab81e115a07561b0e735f0
sspdisk03    20480     THICK         100%  0           a7d40fdb5238364833efb6ef9078dfbf

$ lsmap -clustername ssp_cluster -all
Physloc                                                          Client
Partition ID
---------------------------------------------------------------  ------------
U8233.E8B.10DD51P-V1-C13                                         0x00000003

VTD                  vtscsi0
LUN                  0x8100000000000000
Backing device       sspdisk01.a28b5064df887005d85bdae34615a383

Physloc                                                          Client
Partition ID
---------------------------------------------------------------  ------------
U8233.E8B.10DD51P-V7-C23                                         0x00000003

VTD                  vtscsi0
LUN                  0x8100000000000000
Backing device       sspdisk01.a28b5064df887005d85bdae34615a383

Physloc                                                          Client
Partition ID
---------------------------------------------------------------  ------------
U8233.E8B.10DD51P-V1-C14                                         0x00000004

VTD                  vtscsi1
LUN                  0x8100000000000000
Backing device       sspdisk03.a7d40fdb5238364833efb6ef9078dfbf
```

2. Run the **rmbdsp** command to remove the logical unit from the shared storage pool as shown in Example 10-27.

Example 10-27 Removing a logical unit by name

```
$ rmbdsp -clustername ssp_cluster -sp ssp_pool -bd sspdisk03
vtscsi1 deleted
```

```
Logical unit sspdisk03 with udid "a7d40fdb5238364833efb6ef9078dfbf" is
removed.
```

If the logical unit being removed is mapped to client partitions being serviced by multiple Virtual I/O Servers, the mapping of the VTD must be removed from the other Virtual I/O Servers before the **rmdbsp** command can successfully remove the logical unit. For more information, see “Unmapping logical units” on page 300.

Example 10-28 shows the error message that displays when you try to remove a logical unit that is mapped on more than one Virtual I/O Server partition.

Example 10-28 Error message when removing LU on multiple Virtual I/O Servers

```
$ rmbdsp -clustername ssp_cluster -sp ssp_pool -bd sspdisk01
Unable to remove specified logical unit due to mapping "vtscsi0" on remote
system "172.16.21.61".
```

Important: If the logical unit is only mapped to client partitions serviced by the same Virtual I/O Server, the mapping to all client partitions is removed. If you want to remove only the mapping to a particular client partition, use the **-vtd** option with the **rmbdsp** command as shown in Example 10-25 on page 300.

Logical units can also be removed based on their Lu Udid as shown in Example 10-29.

Example 10-29 Removing a logical unit by Lu Udid

```
$ rmbdsp -clustername ssp_cluster -sp ssp_pool -luudid
63f00cffbc0629f32b7927510a882e18
Logical unit  with udid "63f00cffbc0629f32b7927510a882e18" is
removed.
```

10.1.8 Changing the storage threshold

The shared storage pool space is used to store virtual client partition user data. You can monitor threshold alerts to verify whether the free space decreases to a value lower that the acceptable value.

You can change the threshold limit of the storage usage by using the Virtual I/O Server command-line interface.

Important: Free space must not be reduced to a value below 5% of the total space. If this occurs, I/O operations on the virtual client partition might fail. To avoid this failure, add physical volumes to the pool or delete data from the pool to create free space.

The threshold limit for alert generation is a percentage value. If the actual storage usage changes to a value that is either higher or lower than the threshold limit, an alert is generated. In addition, an entry is added to the Virtual I/O Server error log on the Primary Notification Node (PNN). If a PNN does not exist, the error log is created on the Database Node (DBN). To determine whether the Virtual I/O Server partition is a PNN or the DBN, run the **lssrc -ls vio_daemon** command.

The system error log is used to track threshold conditions. These conditions are propagated to the Hardware Management Console (HMC) connected to the Virtual I/O Server partition. The threshold limit can be changed to a value from 1% - 99%, with the percentage representing the amount of free space. The default threshold monitoring is set to alert when the free space decreases to a value lower than 35% of the total capacity.

For example, if the threshold limit is 20% and the amount of free space decreases to a value lower than 20%, an alert is generating indicating that the threshold limit was exceeded. Increasing the storage capacity of the shared storage pool such that the free space exceeds 20% generates another alert that indicates that the threshold is no longer exceeded. An optimum threshold limit depends on the administrative capability to respond to alerts and on how quickly storage is used.

The following list describes how to change the threshold limit, remove threshold alerts, and view threshold alerts:

- To change the threshold limit, run the **alert** command. In the following example, the threshold limit is changed to 10%. An exceeded alert is generated when the free space decreases to a value lower than 10% of the physical storage pool capacity.

  ```
  alert -set -clustername ssp_cluster -spname ssp_pool -type threshold
  -value 10
  ```

  Note: You can check threshold alerts in the Virtual I/O Server system error log.

- To remove the threshold alert from the storage pool, enter the **alert -unset** command:

  ```
  alert -unset -clustername ssp_cluster -spname ssp_pool -type
  threshold
  ```

  Note: If you disable the threshold alert notification feature, no threshold alerts will be generated. Threshold alerts are important when you use thin-provisioned logical units in shared storage pool.

- To view the threshold alert on the storage pool, enter the **alert -list** command:

  ```
  alert -list -clustername ssp_cluster -spname ssp_pool -type
  threshold
  ```

- To list the error log, enter the **errlog –ls | more** command. Look for log entries that contain the following information:

  ```
  Information messages
  VIO_ALERT_EVENT label
  Threshold Exceeded alert
  ```

  When the actual storage pool usage changes to a value that is either higher or lower than this limit, an alert is raised. In addition, an entry is made into the Virtual I/O Server error log. You can monitor the Virtual I/O Server by using the **errlog** command as shown in Example 10-30.

Example 10-30 Checking the alert in the Virtual I/O Server error log

```
$ errlog
IDENTIFIER TIMESTAMP  T C RESOURCE_NAME  DESCRIPTION
0FD4CF1A   1210121810 I O VIO1_268303150 Informational Message
$ errlog -ls
---------------------------------------------------------------------------
LABEL:          VIO_ALERT_EVENT
IDENTIFIER:     0FD4CF1A

Date/Time:       Fri Dec 10 12:18:03 EST 2010
Sequence Number: 44433
Machine Id:      00F61AA64C00
Node Id:         p71vios1
Class:           O
Type:            INFO
WPAR:            Global
Resource Name:   VIO1_26830315039NONE

Description
Informational Message
```

```
Probable Causes
Asynchronous Event Occurred

Failure Causes
PROCESSOR

        Recommended Actions
        Check Detail Data

Detail Data
Alert Event Message
25b8001
A Storage Pool Threshold alert event occurred on pool D_E_F_A_U_L_T_061310 pool
id FFFFFFFFAC10153C0000000050C7AA2D in cluster ssp_cluster cluster id
6e1e70e043dc11e283fc661dbccf745a The alert event received is:     Threshold
Exceeded.

Diagnostic Analysis
Diagnostic Log sequence number: 1579
Resource tested:        sysplanar0
Menu Number:            25B8001
Description:


A Storage Pool Threshold alert event occurred on pool D_E_F_A_U_L_T_061310 pool
id FFFFFFFFAC10153C0000000050C7AA2D in cluster clusterA cluster id
6e1e70e043dc11e283fc661dbccf745a The alert event received is:     Threshold
Exceeded.
```

Overcommitment is the sum of both capacity from *thin* and *thick* logical units minus the total physical volume capacity in the pool.

The following list describes how to change the overcommit limit of the storage pool, view alerts, and remove alerts:

- To change the overcommit limit of the storage pool, run the **`alert -set`** command:

  ```
  alert -set -clustername ssp_cluster -spname ssp_pool -type
  overcommit -value 80
  ```

- To view the alert on the storage pool, run the **`alert -list`** command:

  ```
  alert -list -clustername ssp_cluster -spname ssp_pool
  ```

Example 10-31 shows the alert setup.

Example 10-31 Listing the alert setup

```
$ alert -list -clustername ssp_cluster -spname ssp_pool
PoolName:          ssp_pool
PoolID:            FFFFFFFFAC10153C0000000050C7AA2D
ThresholdPercent:  10
OverCommitPercent: 80
```

- To remove the alert on the storage pool, run the **`alert -unset`** command:

```
alert -unset -clustername ClusterA -spname poolA -type overcommit
```

Important: You cannot restore a cluster or a node from a cluster if it is removed by using the **`cluster`** command with **`-delete`** or **`-rmnode`** options.

10.1.9 Rolling updates in a cluster

Virtual I/O Server Version 2.2.2.0 supports rolling updates for clusters. The rolling updates enhancement prevents requiring an outage for the entire cluster when you apply software updates to Virtual I/O Server partitions in the cluster.

Rolling updates are not supported in clusters with Virtual I/O Server partitions running versions earlier than 2.2.1.4. Rolling updates are only supported for Virtual I/O Server partitions running version 2.2.1.4 or 2.2.1.5 when they are being updated to version 2.2.2.0 or later. If your cluster contains Virtual I/O Server partitions running versions earlier than 2.2.1.4, see 10.1.10, “Upgrading a cluster configuration” on page 307.

The *capability* of a node refers to its ability to take advantage of the enhancements available in the highest Virtual I/O Server version running in the cluster. Virtual I/O Server partitions running version 2.2.20 or later can check the capability level of the nodes in the cluster relative to the capability of the cluster itself. A new field, `Node Upgrade Status`, has been added to the **`cluster -status -verbose`** command. This field can have three values:

- UP_LEVEL. The capability level of the node is higher than the capability of the cluster.
- ON_LEVEL. The capability level of the node is the same as the capability of the cluster.
- DOWN_LEVEL. The capability level of the node is lower than the capability of the cluster.

To upgrade the capability of the cluster, all nodes in the cluster must be active and running the Virtual I/O Server for the capability level wanted. The capability of the cluster cannot be upgraded until all nodes are active.

The DBN, in Virtual I/O Server Version 2.2.2.0 and later, uses the command **/usr/sbin/sspupgrade** to automatically upgrade the cluster configuration every 10 minutes by using cron. Only the DBN is allowed to initiate and coordinate the upgrades.

Note: The following cluster configuration operations are restricted when an upgrade is being performed:

- Adding a Virtual I/O Server partition to the cluster
- Adding a physical volume to the storage pool
- Replacing a physical volume in the storage pool

10.1.10 Upgrading a cluster configuration

This section addresses upgrading a Virtual I/O Server partition in a cluster from a version earlier than 2.2.1.4 to version 2.2.14. The upgrade is needed to take advantage of the new shared storage pool enhancements. To upgrade a cluster node to Virtual I/O Server Version 2.2.2.0 or later, the Virtual I/O Server must be running at least version 2.2.1.4.

When performing this task, you can restore the previous shared storage pool mappings with a new shared storage pool and database versions.

Use the following steps can be used to upgrade a Virtual I/O Server partition in a cluster from version 2.2.1.1 or 2.2.1.3 to Virtual I/O Server Version 2.2.1.4 or 2.2.1.5. With these steps, you can restore your shared storage pool mappings to the new shared storage pool and database versions:

1. Create a backup of the cluster. This command creates a backup of the cluster, `clusterA`, and stores it in a file named `oldCluster`. The backup file that is generated is `oldCluster.clusterA.tar.gz`.

   ```
   viosbr -backup -file oldCluster -clustername clusterA
   ```

2. Copy the backup file that is generated to a different system.
3. Reinstall the Virtual I/O Server partition with version 2.2.1.4.

Tip: Do not change the physical volumes that are used for the storage pool.

4. Upgrade the cluster configuration to version 2.2.1.4:

   ```
   viosbr -migrate -file oldCluster.clusterA.tar.gz
   ```

5. Clean the physical volume that will be used as the cluster repository disk:

   ```
   cleandisk -r hdisk9
   ```

6. Restore the network devices by using the migrated backup file:

   ```
   viosbr -restore -file oldCluster_MIGRATED.clusterA.tar.gz
   -clustername clusterA -repopvs hdisk9 -type net

   viosbr -restore -file oldCluster_MIGRATED.clusterA.tar.gz
   -clustername clusterA -subfile clusterAMTM9117-MMA02O6AB272P9.xml
   -type net
   ```

7. Restore the cluster by using the migrated backup file:

   ```
   viosbr -restore -file oldCluster_MIGRATED.clusterA.tar.gz
   -clustername clusterA -repopvs hdisk9

   viosbr -restore -file oldCluster_MIGRATED.clusterA.tar.gz
   -clustername clusterA -subfile clusterAMTM9117-MMA02O6AB272P9.xml
   ```

8. Verify that the cluster restored successfully by listing the status of the nodes in the cluster:

   ```
   cluster -status -clustername clusterA
   ```

9. Verify the storage mappings on the Virtual I/O Server:

   ```
   lsmap -all
   ```

10.1.11 Upgrading a cluster from IPv4 to IPv6

Starting from Virtual I/O Server Version 2.2.2.0, or later, you can upgrade an existing cluster from IPv4 to IPv6. The migration is possible only when each of the Virtual I/O Server partitions are updated to Version 2.2.2.0 or later.

Important: Do not change the Virtual I/O Server partition's IPv4 address configuration on the Virtual I/O Server partition or in your Domain Name System (DNS) until instructed to do so.

The following steps upgrade a cluster from IPv4 to IPv6. These steps must be performed on each node in the cluster.

1. Run the **mktcpip** command to add an IPv6 address.
2. Stop the cluster services by using this command:

   ```
   clstartstop -stop -n clustername -m node_hostname
   ```

3. Make the necessary changes in the network configuration, Neighbor Discovery protocol (NDP) daemon router, or DNS information so that the IPv6 address of the Virtual I/O Server partition resolves to the same host name that earlier resolved to the IPv4 address.

   Note: Ensure that both the forward and reverse DNS lookup for the host name resolves to the required IPv6 address.

4. Restart the cluster services on the Virtual I/O Server partition by using this command:

   ```
   clstartstop -start -n <clustername> -m <node_hostname>
   ```

5. Run the **rmtcpip** command to remove the IPv4 address.

10.1.12 Virtual I/O Server host name changes

In a cluster configuration, the host name of a Virtual I/O Server partition cannot be changed. The Virtual I/O Server partition must be removed from the cluster to change its host name.

Use the following steps to change the host name of a Virtual I/O Server that is a node in a cluster. If there are two or more Virtual I/O Server partitions in the cluster. complete these steps:

1. Remove the Virtual I/O Server partition from the cluster and change the host name.
2. Update `/etc/hostname` in all remaining cluster Virtual I/O Server partitions.
3. Read the Virtual I/O Server partition to the cluster.

If there is only one Virtual I/O Server partition in the cluster, complete these steps:

1. Remove the logical unit mapping and the logical units.
2. Delete the cluster.
3. Change the Virtual I/O Server partition host name.

4. Re-create the cluster, pool, logical units, and mapping.
5. If necessary, restore the client partition from a backup.

For more information, see 10.1.2, “Managing nodes in a cluster” on page 286.

10.2 Monitoring shared storage pools

This section describes how to troubleshoot shared storage pools issues, monitor cluster status, and display information about pools, logical units, and mappings.

10.2.1 Listing the cluster and node names

List the cluster information by running the **cluster -list** command, as shown in Example 10-32.

Example 10-32 Listing the cluster information

```
$ cluster -list
Cluster Name     Cluster ID
ssp_cluster      b2326428161811e1bc83001a64bb6948
```

Run the **cluster** command to check the status of the cluster and the nodes as shown in Example 10-33.

Example 10-33 Checking the status of the cluster

```
$ cluster -status -clustername ssp_cluster
Cluster Name         State
ssp_cluster          OK

    Node Name        MTM           Partition Num  State  Pool State
    p71vios1         8233-E8B0210OEF5R         1  OK     OK
    p71vios2         8233-E8B0210OEF5R         2  OK     OK
    p72vios1         8233-E8B02061AA6P        33  OK     OK
    p72vios2         8233-E8B02061AA6P        34  OK     OK
```

10.2.2 Verifying the cluster

This section describes how to correct common cluster problems.

Verifying node membership status

Many cluster issues can be diagnosed by reviewing the cluster state information.

The primary command to display the state of a cluster is **cluster -status**, which shows the health of the cluster configuration and pool. See Example 10-4 on page 289.

If any node shows `State of node: DOWN`, perform these troubleshooting steps:

- List the cluster node configuration information with the **lscluster -m** command, as shown in Example 10-34.

Example 10-34 Listing node status

```
$ lscluster -m
Calling node query for all nodes...
Node query number of nodes examined: 4

        Node name: p71vios1
        Cluster shorthand id for node: 1
        UUID for node: 6e212f74-43dc-11e2-83fc-661dbccf745a
        State of node: DOWN
        Smoothed rtt to node: 500
        Mean Deviation in network rtt to node: 1500
        Number of clusters node is a member in: 1
        CLUSTER NAME       SHID         UUID
        ssp_cluster        0            6e1e70e0-43dc-11e2-83fc-661dbccf745a
        SITE NAME          SHID         UUID
        LOCAL              1            51735173-5173-5173-5173-517351735173

        Points of contact for node: 2
        ----------------------------------------
        Interface     State  Protocol    Status
        ----------------------------------------
        dpcom         DOWN   none      RESTRICTED
        en4           DOWN   IPv4         none
```

If any nodes are down, check the HMC for the logical partition state and run the **ping** command. Ensure that all nodes can be pinged by both the IP address and host name specified at cluster creation time.

- Check the error logs for loss of connectivity or hardware errors.
- Check that the `/etc/hosts`, `/etc/resolv.conf`, `/etc/netsvc.conf`, and `/etc/irs.conf` files on all nodes are consistent with each other and with the cluster topology.
- If the node is offline due to a failure, attempt to resolve the failure. Then run the **clstartstop** command with **-start** option, as shown in Example 10-5 on page 290, to bring the node online.

Verifying physical volumes availability

If any node shows that the state of the pool is DOWN, perform these troubleshooting steps:

- List the disks in the pool with **lspv** command, as shown in Example 10-11 on page 293.
- Check the availability of storage by running the **lsdev** command.
- Run the **df** command to make sure that there is space available on all root file systems. Run the **chfs** command to add space if necessary.

Verifying error to create thick logical units

Insufficient storage in the pool might result in I/O errors on Virtual I/O Server clients, especially when you are using *thin* logical units.

You will not be able to create *thick* logical units until you have enough space in the pool, as shown on Example 10-35.

Example 10-35 Error to create a thick logical unit

```
$ mkbdsp -clustername ssp_cluster -sp ssp_pool 30G -bd sspdisk05 -thick
Storage Pool subsystem operation failed, unable to create LU.
Storage Pool subsystem operation failed, not enough space in the pool.
```

To list the space available in the pool, run the **lssp** command as shown in Example 10-36.

Example 10-36 Listing free space in the pool

```
$ lssp -clustername ssp_cluster
POOL_NAME:       ssp_pool
POOL_SIZE:       102144
FREE_SPACE:      8192
TOTAL_LU_SIZE:   194560
OVERCOMMIT_SIZE: 94242
TOTAL_LUS:       5
POOL_TYPE:       CLPOOL
POOL_ID:         FFFFFFFFAC10153C0000000050C7AA2D
```

Virtual I/O Server client cannot access mapped logical units

If the disks cannot be seen on the client partition or I/O to the mapped logical unit is consistently failing, verify that the Virtual I/O Server is up. Complete the following troubleshooting steps:

- List the logical units by running **lssp** command, as shown in Example 10-37.

Example 10-37 Listing logical units in the cluster

```
$ lssp -clustername ssp_cluster -sp ssp_pool -bd
Lu Name          Size(mb)    ProvisionType %Used Unused(mb)  Lu Udid
sspdisk01        30720       THICK         100%  0   bc95d534e394d483905d9d6bbe0c9c54
sspdisk02        30720       THICK         100%  0   60f4ac6a423aeb0d76ef7c749ec04d29
sspdisk03        51200       THIN          0% 51203  b0cf3978a09c36806f5f68e36aa03066
sspdisk03        51200       THIN          0% 51203  86f7e1051c38e2534db68a655f2c0b63
sspdisk04        30720       THICK          100%  0  c9c9cb88d7bf972919d597520b3c1696
```

- On Virtual I/O Server, list the mapping devices by using the **lsmap -all** command. You can also run the **lsmap -cluster** commando to list logical units that are mapped in the entire cluster, as shown in Example 10-23 on page 298.
- On Virtual I/O Server, check for error messages with the **errpt -a** command.
- On Virtual I/O Server, check for storage issues as detailed in “Verifying physical volumes availability” on page 312.

Listing the cluster storage interfaces

If the Virtual I/O Server loses access to the repository disk or cluster shared disk, it will be reflected in the **lscluster -d** command. You can check whether the disk is available, as shown in Example 10-38.

Example 10-38 Listing the cluster storage interfaces

```
 $ lscluster -d
Storage Interface Query

Cluster Name: ssp_cluster
Cluster UUID: 6e1e70e0-43dc-11e2-83fc-661dbccf745a
Number of nodes reporting = 4
Number of nodes expected = 4

Node p71vios1
Node UUID = 6e212f74-43dc-11e2-83fc-661dbccf745a
Number of disks discovered = 3
        hdisk10:
              State : UP
               uDid : 3E213600A0B8000291B08000003DFF09E30CF30F1815      FAStT03IBMfcp
```

```
            uUid : f77ce565-12f4-ecf1-4fae-bcba1006ae9a
      Site uUid : 51735173-5173-5173-5173-517351735173
            Type : CLUSDISK
      hdisk9:
           State : UP
            uDid : 3E213600A0B80001146320000 5C5750C045AD0F1815      FAStT03IBMfcp
            uUid : f80e20af-82b0-5b96-577f-5d21c3e84f2c
      Site uUid : 51735173-5173-5173-5173-517351735173
            Type : CLUSDISK
      hdisk12:
           State : DOWN
            uDid : 3E213600A0B8000291B0800003E0109E30D2D0F1815      FAStT03IBMfcp
            uUid : 5f66753e-01aa-1fb4-e703-cb63de5cf455
      Site uUid : 51735173-5173-5173-5173-517351735173
            Type : REPDISK
[...]
```

In this example, the repository disk is DOWN. For the disk that is listed as DOWN, check that it is available by running the **lsdev -dev** command on the same node.

For more information about replacing a repository disk, see 10.1.5, “Repository resiliency” on page 294.

Checking for Virtual I/O Server version mismatches

To use new functions that are introduced in a new Virtual I/O Server version, all nodes in the cluster must be at the same level.

You can verify the current level of Virtual I/O Server partitions in the cluster by running the **cluster** command with **-verbose** option as shown in Example 10-39.

Example 10-39 Listing the Virtual I/O Server version in the cluster

```
$ cluster -status -clustername ssp_cluster -verbose
Cluster Name:            ssp_cluster
Cluster Id:              6e1e70e043dc11e283fc661dbccf745a
Cluster State:           OK
Repository Mode:         EVENT
Number of Nodes:         4
Nodes OK:                4
Nodes DOWN:              0


    Node Name:           p71vios1
    Node Id:             6e212f7443dc11e283fc661dbccf745a
    Node MTM:            8233-E8B0210DD51P
```

```
Node Partition Num:   1
Node State:           OK
Node Repos State:     OK
Node Upgrade Status:   ON_LEVEL
Node Roles:
    Pool Name:        ssp_pool
    Pool Id:          FFFFFFFFAC10153C0000000050C7AA2D
    Pool State:       OK

Node Name:            p71vios2
Node Id:              a600b4a043dc11e2b5f0661dbccf745a
Node MTM:             8233-E8B0210DD51P
Node Partition Num:   7
Node State:           OK
Node Repos State:     OK
Node Upgrade Status:   ON_LEVEL
Node Roles:           DBN
    Pool Name:        ssp_pool
    Pool Id:          FFFFFFFFAC10153C0000000050C7AA2D
    Pool State:       OK

Node Name:            p72vios1
Node Id:              c523978043dc11e2a7e0661dbccf745a
Node MTM:             8233-E8B02061AB2P
Node Partition Num:   1
Node State:           OK
Node Repos State:     OK
Node Upgrade Status:   ON_LEVEL
Node Roles:
    Pool Name:        ssp_pool
    Pool Id:          FFFFFFFFAC10153C0000000050C7AA2D
    Pool State:       OK

Node Name:            p72vios2
Node Id:              e3e3c15e43dc11e29d98661dbccf745a
Node MTM:             8233-E8B02061AB2P
Node Partition Num:   2
Node State:           OK
Node Repos State:     OK
Node Upgrade Status:   ON_LEVEL
Node Roles:
    Pool Name:        ssp_pool
    Pool Id:          FFFFFFFFAC10153C0000000050C7AA2D
    Pool State:       OK
```

The *Node Upgrade Status* shows the current level that a node is running and it can be:

ON_LEVEL	Node software capabilities match current cluster level
UP_LEVEL	Node software capabilities are newer than current cluster level.
DOWN_LEVEL	Node software capabilities do not meet current cluster levels.

10.2.3 Displaying the physical volumes in the shared storage pool

To display the physical volumes in the shared storage pool, use the **lspv** command as shown in Example 10-40.

Example 10-40 Listing physical volumes in the shared storage pool

```
$ lspv -clustername ssp_cluster -sp ssp_pool
PV NAME           SIZE(MB)    PVUDID
hdisk9            51200       3E213600A0B800011463200005C5750C045AD0F1815      FAStT03IBMfcp
hdisk10           51200       3E213600A0B8000291B0800003DFF09E30CF30F1815      FAStT03IBMfcp
```

To display the size of the shared storage pool and free space, use the **lssp** command as shown in Example 10-41.

Example 10-41 Listing the shared storage pool

```
$ lssp -clustername ssp_cluster
POOL_NAME:       ssp_pool
POOL_SIZE:       102144
FREE_SPACE:      8192
TOTAL_LU_SIZE:   194560
OVERCOMMIT_SIZE: 94242
TOTAL_LUS:       5
POOL_TYPE:       CLPOOL
POOL_ID:         FFFFFFFFAC10153C0000000050C7AA2D
```

For logical units from a shared storage pool using thin provisioned devices, blocks on the physical disks in the shared storage pool are only allocated when the client partition starts writing to the disks. The **lssp** command displays the information that is required by the administrator to anticipate storage needs.

The **alert** command also allows you to list and set up a threshold alert to raise a message in the Virtual I/O Server partition error log. For more information how to set up the threshold alert, see 10.1.8, “Changing the storage threshold” on page 302.

To list the current alert configuration, issue **alert -list** command as shown in Example 10-42.

Example 10-42 Listing the alert configuration

```
$ alert -list -clustername ssp_cluster -spname ssp_pool
PoolName:          ssp_pool
PoolID:            FFFFFFFFAC10153C0000000050C7AA2D
ThresholdPercent:  10
OverCommitPercent: 80
```

10.2.4 Tracing logical units

To list the logical units in a shared storage pool, use the **lssp** command with **-bd** option, as shown in Example 10-43.

Example 10-43 Listing the logical units in a shared storage pool

```
$ lssp -clustername ssp_cluster -sp ssp_pool -bd
Lu Name          Size(mb)    ProvisionType      %Used Unused(mb)  Lu Udid
sspdisk01        30720       THICK               100% 0       bc95d534e394d483905d9d6bbe0c9c54
sspdisk02        30720       THICK               100% 0       60f4ac6a423aeb0d76ef7c749ec04d29
sspdisk03        51200       THIN                  0% 51203   b0cf3978a09c36806f5f68e36aa03066
sspdisk03        51200       THIN                  0% 51203   86f7e1051c38e2534db68a655f2c0b63
sspdisk04        30720       THICK               100% 0       c9c9cb88d7bf972919d597520b3c1696
```

Use **lsmap** to display the mapping of logical units to virtual adapters and partitions as shown in Example 10-44. Notice the name of the backing devices. The client partition ID is shown in hexadecimal. You must convert it to decimal to match it with the corresponding id on the HMC.

Example 10-44 Listing the mapping on a specific host

```
$ lsmap -clustername ssp_cluster -hostname p71vios1
Physloc                                          Client Partition ID
-----------------------------------------------  ------------------
U8233.E8B.10EF5R-V1-C104                         0x00000004

VTD                  vtscsi0
LUN                  0x8100000000000000
Backing device       sspdisk01.198d854abebe7e965214d8360eae60fe

Physloc                                          Client Partition ID
-----------------------------------------------  ------------------
U8233.E8B.10EF5R-V1-C105                         0x00000005

VTD                  vtscsi1
```

```
LUN                 0x8100000000000000
Backing device      sspdisk02.b64b1ad9f28fd2052f0355a4b3fd8481
```

Tip: You can display the mapping for all members of a cluster by changing the host name in this command. If you use the **-all** parameter, the output includes mappings for all members. When the Virtual I/O Servers in the cluster are on different physical servers, the client partition ID can be the same.

To display the detailed mapping information, use the **lsmap -all** command. If you just want to find the vhost adapter mapped to a partition, you can filter on the Partition ID as shown in Example 10-45.

Example 10-45 vhost adapters mapped to client partition 4

```
$ lsmap -all|grep 0x00000004
vhost1          U8233.E8B.100EF5R-V1-C104        0x00000004
vhost5          U8233.E8B.100EF5R-V1-C904        0x00000004
```

To get the details for a specific adapter, use the **lsmap** command with **-vadapter** option, as shown in Example 10-46.

Example 10-46 Mapping information of vhost1

```
$ lsmap -vadapter vhost1
SVSA            Physloc                      Client Partition ID
--------------- ---------------------------- ------------------
vhost1          U8233.E8B.100EF5R-V1-C104    0x00000004

VTD                   vtscsi0
Status                Available
LUN                   0x8100000000000000
Backing device        sspdisk01.198d854abebe7e965214d8360eae60fe
Physloc
Mirrored              N/A
```

You can also get a combined view from the **cfgassist** menu by selecting `Shared Storage Pools` → `Manage Logical Units in Storage Pool` → `List Logical Unit Map`**s**, as shown in Example 10-47.

Example 10-47 Abstract from cfgassist menu

```
SVSA(VHOST) Physloc         Client ID  VTD Name Backing dev(LU) Name
--------------------------  ---------- -------- ---------------
U8233.E8B.061AA6P-V33-C136  0x00000024 vtscsi0    sspdisk06
U8233.E8B.100EF5R-V1-C104   0x00000004 vtscsi0    sspdisk01
U8233.E8B.100EF5R-V2-C104   0x00000004 vtscsi0    sspdisk01
U8233.E8B.100EF5R-V1-C105   0x00000005 vtscsi1    sspdisk02
```

```
U8233.E8B.10OEF5R-V2-C106   0x00000006 vtscsi1    sspdisk03
U8233.E8B.10OEF5R-V1-C106   0x00000006 vtscsi2    sspdisk03
U8233.E8B.10OEF5R-V2-C103   0x00000003 vtscsi2    sspdisk04
U8233.E8B.10OEF5R-V1-C103   0x00000003 vtscsi3    sspdisk04
U8233.E8B.10OEF5R-V2-C105   0x00000005 vtscsi3    sspdisk02
U8233.E8B.10OEF5R-V1-C111   0x0000000b vtscsi4    sspdisk05
U8233.E8B.10OEF5R-V2-C111   0x0000000b vtscsi4    sspdisk05
U8233.E8B.10OEF5R-V1-C190   0x0000005a vtscsi7    sspdisk07
U8233.E8B.10OEF5R-V2-C190   0x0000005a vtscsi7    sspdisk07
U8233.E8B.10OEF5R-V1-C190   0x0000005a vtscsi9    sspdisk99
U8233.E8B.10OEF5R-V1-C103   0x00000003 vtscsi10   sspdisk09
```

Part 4

Virtual I/O Server

This part addresses guidelines for managing and monitoring the Virtual I/O Server.

This part includes the following chapter:

- Virtual I/O Server

11

Virtual I/O Server

The Virtual I/O Server is a software appliance that allows you to share physical resources among multiple logical partitions. It is included with all three editions of PowerVM. For more information about capabilities by PowerVM edition, see Table 1-3 on page 5. The Virtual I/O Server can use both virtualized storage and network adapters, using the virtual SCSI and virtual Ethernet facilities.

It is assumed you are well-versed in concept of PowerVM environment. To obtain detailed information, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

This chapter includes the following sections:

- Managing Virtual I/O Servers
- Monitoring Virtual I/O Servers

11.1 Managing Virtual I/O Servers

As with all other servers included in an enterprise's data recovery program, you must install, backup/restore, and update the Virtual I/O Server partition.

This section includes the following topics:

- Upgrading to a new Virtual I/O Server version 2.x
- Virtual I/O Server backup and restore strategy
- Backing up user-defined virtual devices
- Backing up using IBM Tivoli Storage Manager
- Planning backups of the Virtual I/O Server
- Restoring the Virtual I/O Server
- Rebuilding the Virtual I/O Server
- Updating the Virtual I/O Server
- Updating Virtual I/O Server adapter firmware
- Error logging on the Virtual I/O Server
- VM Storage Snapshots/Rollback
- Automated management
- Virtualization management tools

11.1.1 Upgrading to a new Virtual I/O Server version 2.x

If you are planning to upgrade a version of Virtual I/O Server lower than 2.1 to 2.2, you must first upgrade your Virtual I/O Server to version 2.1 using the Migration DVD. For more information, see "Upgrading Virtual I/O Server version 1.x to 2.1" on page 325. After the Virtual I/O Server version is on version 2.1, the Update Release can be applied to bring the Virtual I/O Server to the latest level.

The following sections explain in greater detail how to upgrade to Virtual I/O Server version 2.1.

Memory requirements

The minimum memory requirement for Virtual I/O Server version 2.2.1.0 varies based on the configuration.

A general rule for a minimum current memory requirement for Virtual I/O Server version 2.2.1.0 is 512 MB. This amount supports a configuration with few devices.

rootvg requirements for release 2.2.x and beyond

Virtual I/O Server update release 2.2.1.1 Fix Pack 25 adds new features and capabilities to the Virtual I/O Server. As a result of these additions, Virtual I/O Server now requires a minimum of 30 GB in the rootvg volume group. Ensure

that your rootvg contains at least 30 GB of available space before you upgrade to release 2.2.1.1 Fix Pack 25.

Host Ethernet Adapter (HEA) memory requirements

Configurations that contain one or more HEAs require more memory than the 512 MB minimum. Each logical HEA port that is configured requires an extra 102 MB of memory. The minimum memory requirement for configurations with one or more HEA ports, where n is the number of HEA ports, is 512 MB + n x 102 MB.

For more information, see the FP25 Release Notes at:

http://www-01.ibm.com/support/docview.wss?uid=isg400000800

Multi-path drivers

Multi-path drivers such as IBM SDD or IBM SDDPCM must be compatible with version AIX 6.1. If they are not, the multi-path drivers must be replaced after successful migration of the Virtual I/O Server to version 2.x. See your multi-path vendor's documentation for compatibility information and replacement procedures.

Upgrading Virtual I/O Server version 1.x to 2.1

This section describes the steps to upgrade a Virtual I/O Server version 1.x to 2.1. If your Virtual I/O Server is running in a version lower than Virtual I/O Server 1.5.2.6-FP-11.1 SP-02, you must update it before upgrading. To upgrade your Virtual I/O Server to version 2.1, you can choose one of the following options depending on how your system is managed:

- Migrating from a DVD that is managed by an HMC
- Upgrading from a DVD that is managed by an IVM

Before you begin a migration, back up your existing Virtual I/O Server installation and then follow the steps for the installation method that you chose.

You might already have received a Virtual I/O Server version 2.x Migration DVD shipped with the Virtual I/O Server version 2.x Installation DVD. If you do not have the Migration DVD, you can download it from Fix Central at:

http://www-933.ibm.com/support/fixcentral

Customers with a Software Maintenance Agreement (SWMA) can order both sets of Virtual I/O Server version 2.x media from the IBM Entitled Software Support website (login credentials required) at:

http://www.ibm.com/servers/eserver/ess/ProtectedServlet.wss

In a redundant Virtual I/O Server environment, you can upgrade one Virtual I/O Server at a time to avoid any interruption of service.

Tip: Check for release updates and fix packs on Fix Central at:

http://www-933.ibm.com/support/fixcentral/

Migrating from a DVD that is managed by an HMC

Before you begin migrating from a DVD, make sure that the following requirements are fulfilled:

- The HMC version is at a minimum level of V7R7.4.0 or later, and the server firmware is at the appropriate level.
- A DVD drive is assigned to the Virtual I/O Server partition, and you have the Virtual I/O Server version 2.x Migration DVD.
- The Virtual I/O Server version is at least on 1.5.2.6-FP-11.1 SP-02.
- Run the **backupios** command and save the `mksysb` image to a secure location.

Important: Do not use the **updateios** command to upgrade the Virtual I/O Server.

To start the Virtual I/O Server migration, complete these steps:

1. Insert the Virtual I/O Server version 2.x Migration DVD into the DVD drive assigned to the Virtual I/O Server partition.
2. Shut down the Virtual I/O Server partition by running the command **shutdown -force** and wait for the shutdown to complete.
3. Activate the Virtual I/O Server partition and boot it into the SMS menu by clicking **Tasks** → **Operations** → **Activate**.
4. Select the correct profile, select **Open a terminal window or console session**, and then select **SMS** as the Boot mode in the Advanced selection. Click **OK**.
5. A console window opens and the partition starts the SMS main menu. In the SMS menu, select option `5. Select Boot Options` and press Enter.
6. Select option `1. Select Install/Boot device` and press Enter.
7. Select option `3. CD/DVD` and press Enter.
8. Select option `6. List all devices` and press Enter.
9. Select the installation drive and press Enter.
10. Select option `2. Normal Mode Boot` and press Enter.
11. Select option `Yes` and press Enter.

12. The partition boots from the Migration DVD. Figure 11-1 shows the menu that appears after a few moments. Select the console that you want and press Enter.

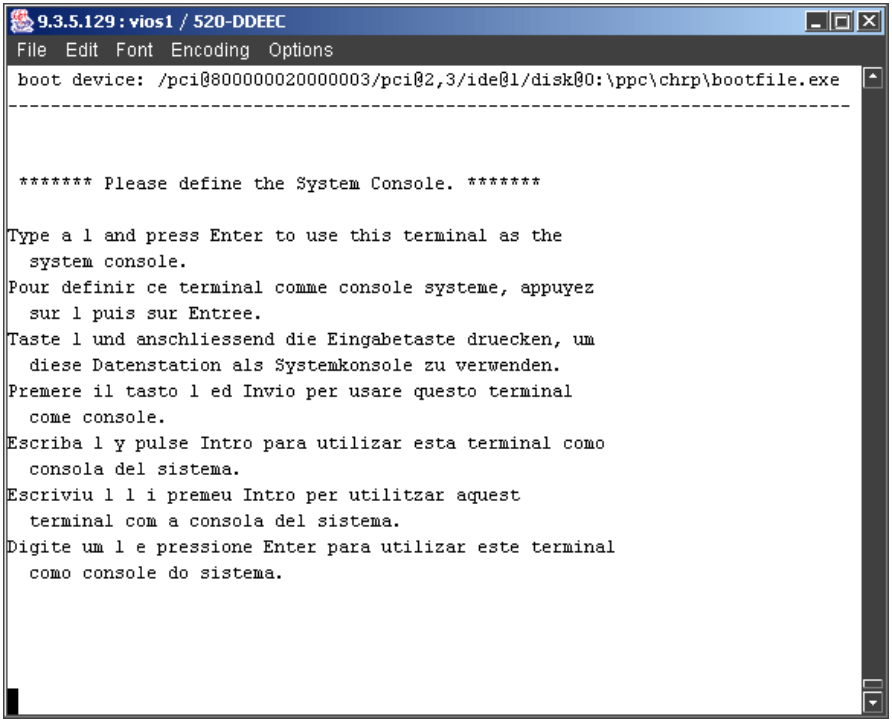


Figure 11-1 Defining the System Console

13. Type 1 in the next panel and press Enter to use English during the installation.

14. The migration proceeds and the main menu is displayed as shown in Figure 11-2.

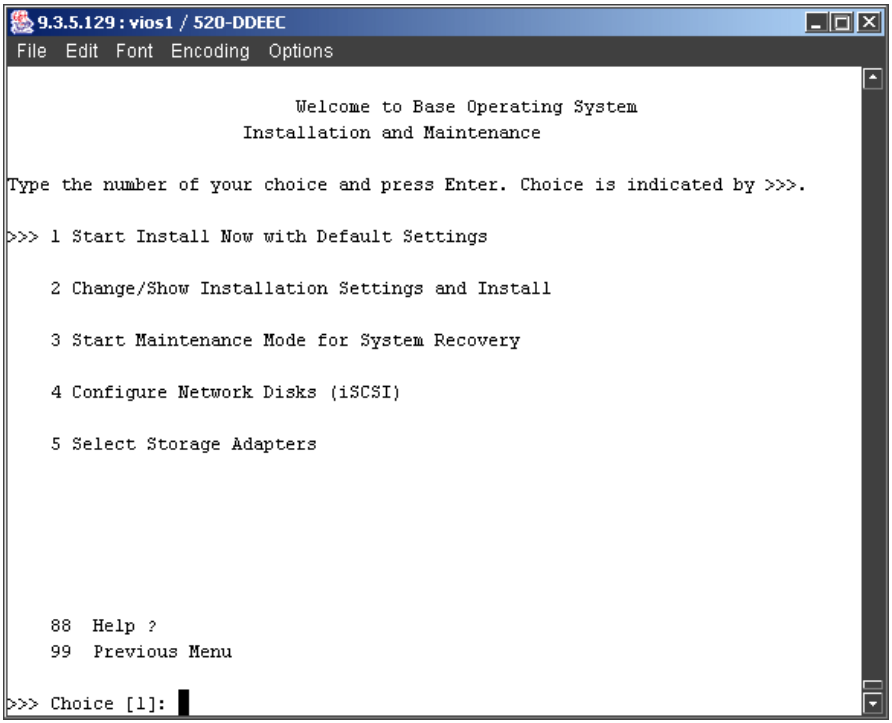


Figure 11-2 Installation and Maintenance main menu

15. Type `1` to select option `1 Start Install Now with Default Settings`, or verify the installation settings by selecting option `2 Change/Show Installation Settings and Install`. Then, press Enter.

16. Figure 11-3 shows the Virtual I/O Server Installation and Settings menu.

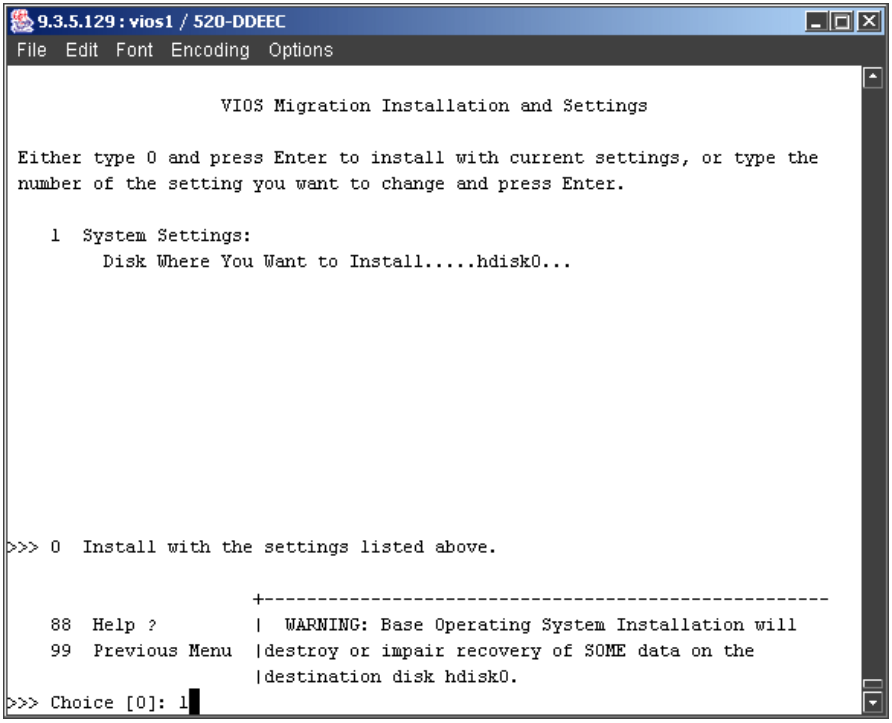


Figure 11-3 Virtual I/O Server Migration Installation and Settings

Type option 1 to verify the system settings.

17. Figure 11-4 shows the menu where you can select the disks for migration. The example uses a mirrored Virtual I/O Server version 1.5 environment, so option 1 is used.

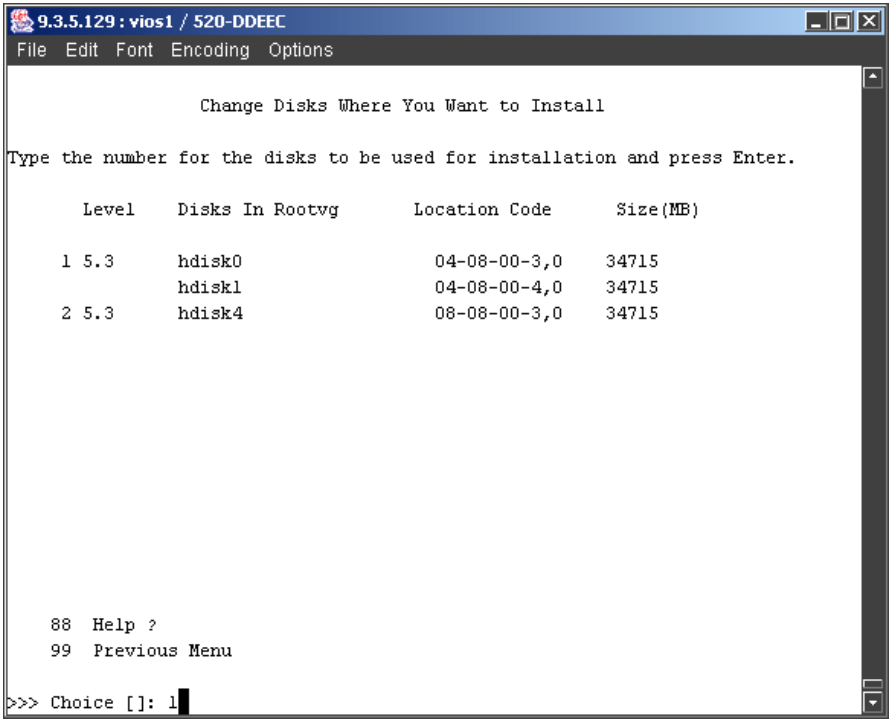


Figure 11-4 Change Disk Where You Want to Install

Tip: Here you can see that the existing Virtual I/O Server is reported as an AIX 5.3 system. Other disks that are not part of the Virtual I/O Server rootvg can also have AIX 5.3 installed.

The first two disks that are shown in Figure 11-4 are internal SAS disks where the existing Virtual I/O Server 1.x is located. hdisk4 is another Virtual I/O Server installation on a SAN LUN.

18. Type 0 to continue and start the migration as shown in Figure 11-5.

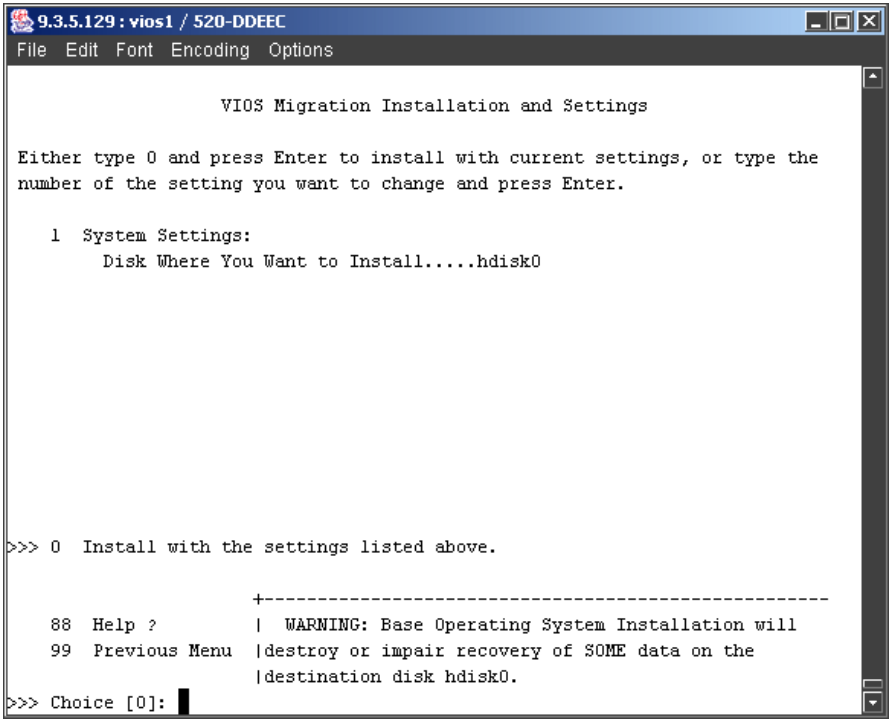


Figure 11-5 Virtual I/O Server Migration Installation and Settings: Starting migration

19. The migration starts and then prompts you for a final confirmation as shown in Figure 11-6. At this point, you can still stop the migration and boot up your existing Virtual I/O Server version 1.x environment.

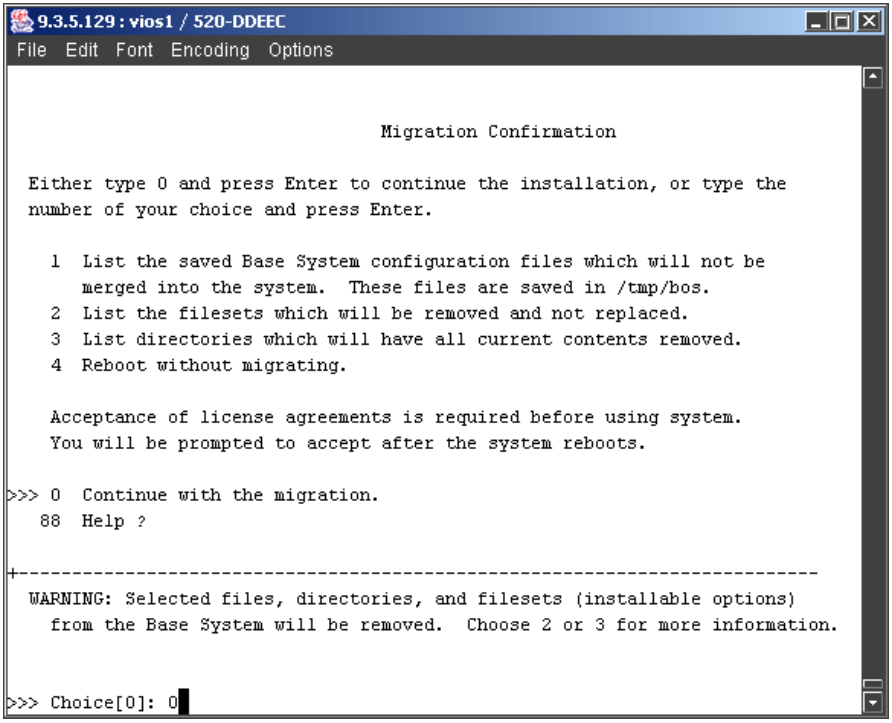


Figure 11-6 Migration Confirmation

20. Type 0 to continue with the migration. After a few seconds, the migration will start as shown in Figure 11-7.

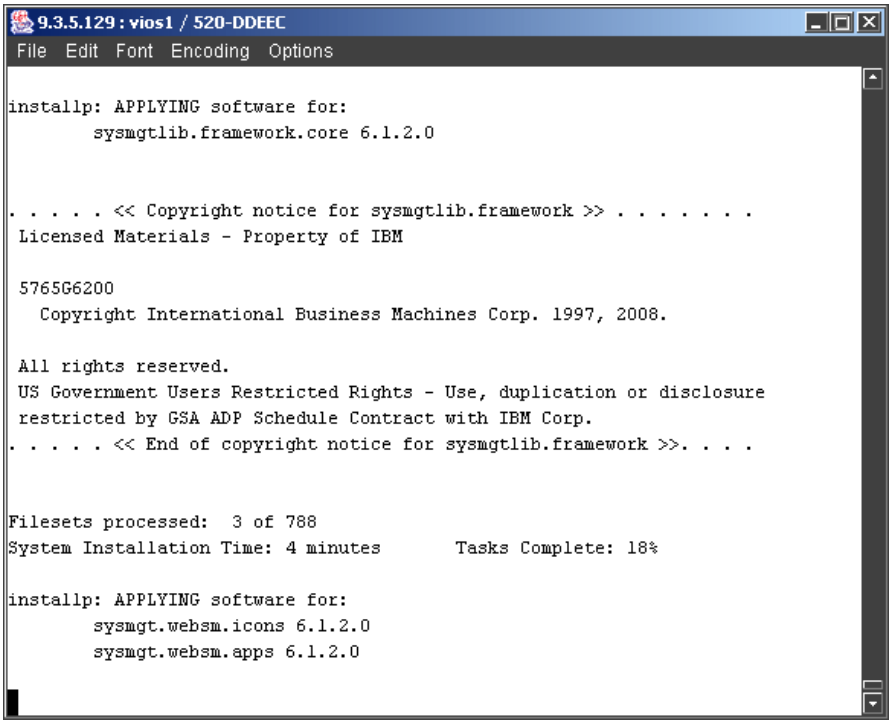


Figure 11-7 Running migration

The running migration process might take some time to complete.

21. After the migration completes, set the terminal type by entering `vt320` as shown in Figure 11-8.

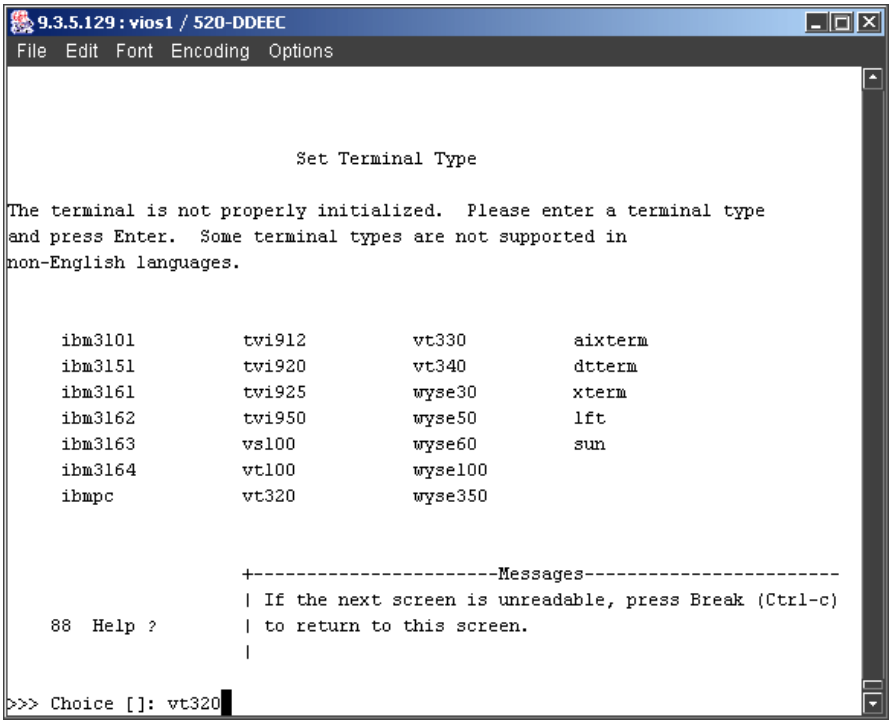


Figure 11-8 Set Terminal Type

22. Accept the license agreements.
23. Exit the menu by pressing F10 (or pressing Esc+0).
24. The Virtual I/O Server login panel is displayed. Log in as padmin and verify the new Virtual I/O Server version with the **`ioslevel`** command.
25. Check the configuration of all disks and Ethernet adapters on the Virtual I/O Server and the mapping of the virtual resources to the virtual I/O client partitions. Use the **`lsmap -all`** and **`lsdev -virtual`** commands.
26. Start the client partitions.

Verify the Virtual I/O Server environment, document the update, and create a new backup of your Virtual I/O Server.

Remember: After you successfully upgrade to Virtual I/O Server version 2.1, if you manually added multi-path drivers (such as IBM SDD or IBM SDDPCM) you must remove them. Then install the corresponding version for an AIX Version 6.1 kernel. See your multi-path driver vendor's documentation for the correct replacement procedure.

Upgrading from a DVD that is managed by an IVM

Before you begin the upgrade from a DVD that uses the Integrated Virtualization Manager (IVM), make sure that the following requirements are fulfilled:

- A DVD drive is assigned to the Virtual I/O Server partition, and you have the Virtual I/O Server version 2.x Migration DVD.
- The Virtual I/O Server version is at least on 1.5.2.6-FP-11.1 SP-02.
- The partition profile data for the management partition and its client partitions are backed up before you back up the Virtual I/O Server. Use the **bkprofdata** command to save the partition configuration data to a secure location.

  Attention: The IVM configuration in Virtual I/O Server 2.x is not compatible with an earlier version. If you want to revert to an earlier version of the Virtual I/O Server, you must restore the partition configuration data from the backup file.

- Run the **backupios** command, and save the `mksysb` image to a secure location.

To start the Virtual I/O Server migration, complete these steps:

1. This step is for a blade server environment only.

   Access the Virtual I/O Server partition by using the management module of the blade server:

   a. Verify that all logical partitions except the Virtual I/O Server partition are shut down.

   b. Insert the Virtual I/O Server Migration DVD into the DVD drive assigned to your Virtual I/O Server partition.

   c. Use telnet to connect to the management module of the blade server on which the Virtual I/O Server partition is located.

   d. Enter the command **env -T system:blade[*x*]**, where *x* is the specific number of the blade to be upgraded.

   e. Enter the **console** command.

   f. Log in into the Virtual I/O Server by using the padmin user.

g. Enter the **shutdown -restart** command.

h. When the system management services (SMS) logo is displayed, select `1` to enter the SMS menu.

i. Skip to step 3.

2. Step 2 is for a non-blade server environment only.

   Access the Virtual I/O Server partition by using the Advanced System Management Interface (ASMI) with a Power Systems server that is not managed by an HMC:

   a. Verify that all logical partitions except the Virtual I/O Server partition are shut down.

   b. Insert the Virtual I/O Server Migration DVD into the Virtual I/O Server partition.

   c. Log in to the ASCII terminal to communicate with the Virtual I/O Server. If you need assistance, see *Accessing the ASMI using an ASCII terminal* at:

      `http://publib.boulder.ibm.com/infocenter/systems/scope/hw/topic/iphby/ascii.htm`

   d. Sign on to the Virtual I/O Server using the padmin user.

   e. Enter the **shutdown -restart** command.

   f. When the SMS logo is displayed, select `1` to enter the SMS menu.

3. Select the boot device:

   a. Select option `5 Select Boot Options` and press Enter.

   b. Select option `1 Select Install/Boot Device` and press Enter.

   c. Select `IDE` and press Enter.

   d. Select the device number that corresponds to the DVD and press Enter. You can also select List all devices, select the device number from a list, and press Enter.

   e. Select `Normal mode boot`.

   f. Select `Yes` to exit SMS.

4. Install the Virtual I/O Server:

   Follow the steps that are described in “Migrating from a DVD that is managed by an HMC” on page 326, beginning with step 13.

11.1.2 Updating Virtual I/O Server version 2.1 to 2.2

After you are on Virtual I/O Server version 2.1, or later, use the **updateios** command to update the Virtual I/O Server.

If you are planning to update a Virtual I/O Server partition that is configured with Shared Storage Pool version 2.2.1.1 or 2.2.1.3, you must first update it to version 2.2.1.4 or 2.2.1.5. For more information about upgrading the cluster version, see 10.1.10, “Upgrading a cluster configuration” on page 307.

For shared storage pools on Virtual I/O Server version 2.2.2.0, the *rolling updates* enhancement allows you to apply updates to the Virtual I/O Server partitions in the cluster individually without causing an outage in the entire cluster. The updated logical partition cannot use the new capabilities until all logical partitions in the cluster are updated and the cluster is upgraded to use the new capabilities. You can check for cluster version mismatches as shown in “Checking for Virtual I/O Server version mismatches” on page 314.

Check for the latest update release, release notes, and installations instructions on the Virtual I/O Server web page at:

`http://www14.software.ibm.com/webapp/set2/sas/f/vios/home.html`

11.1.3 Virtual I/O Server backup and restore strategy

A backup strategy for the Virtual I/O Server must be tailored to your environment. Two strategies worth considering are version-based and schedule-based. Generally, use a combination of the two strategies in your environment.

- A versioning strategy backs up the Virtual I/O Server when a change to the configuration of the Virtual I/O Server occurs. The level of change can be as minor as an update to the Message Of The Day (MOTD) or as major as applying a fix pack.
- A scheduling strategy backs up the Virtual I/O Server regularly. Starting with the release of Virtual I/O Server version 1.3, you can schedule Virtual I/O Server backups through cron.

The Virtual I/O Server contains the following types of information that you must back up:

- The Virtual I/O Server operating system includes the base code, applied fix packs, custom device drivers to support disk subsystems, Kerberos, and LDAP client configurations. All of this information is backed up by using the **`backupios`** command.
- User-defined virtual devices include metadata, such as virtual device mappings, that define the relationship between the physical environment and the virtual environment. All of this information is backed up using the **`viosbr`** command. Also, backing up user-defined virtual device configuration can be done manually as shown in “Backing up user-defined virtual devices by using

backupios” on page 347. This is necessary if you restore the Virtual I/O Server to a different managed system.

Note: A Virtual I/O Server `mksysb` requires a separate backup for the user-defined virtual devices unless the Virtual I/O Server is being restored to the same managed system.

You must also back up the following components of your environment to fully recover your Virtual I/O Server configuration:

- Resources that are defined on the Hardware Management Console (HMC).
- Resources that are defined on the IVM

Attention: It is not possible to restore a suspended partition to a separate server. The only way to move a suspended partition to a different server is by using partition mobility. However, this process requires both the source and the target system to be available

Backing up HMC resources

If the Virtual I/O Server is managed by an HMC, the Virtual I/O Server partition profile must be backed up. The profile of the Virtual I/O Server partition on the HMC not only contains processor, memory, and adapter information, but also virtual device configuration.

Starting with HMC V7, you can save the current system configuration to an HMC system plan. The system plan can be redeployed to rebuild the HMC. The restore must be done on hardware that supports the system level of the backup.

Tip: Check that the system plan is valid by viewing the report. Look for a message that says that the system plan cannot be deployed (in red).

For more information, see the **`mksysplan`** command and the HMC interface.

Backing up IVM resources

If the system is managed by the Integrated Virtualization Manager, you must back up your partition profile data for the management partition and its client partitions before you back up the Virtual I/O Server operating system.

To do so, from the Service Management menu, click **Backup/Restore**. The Backup/Restore page is displayed. Then, click **Generate Backup**.

This operation can also be done from the Virtual I/O Server. To do so, enter this command:

```
bkprofdata -o backup -f /home/padmin/profile.bak
```

Backing up the Virtual I/O Server operating system

The Virtual I/O Server operating system consists of the base code, fix packs, custom device drivers to support disk subsystems, and user-defined customization. The user-defined customization can be as simple as the changing of the Message of the Day or the security settings.

The **backupios** command creates a backup of the Virtual I/O Server to a bootable tape, a DVD, or a file system (local or a remotely mounted Network File System).

Remember: Client data is not backed up. Also, the contents of the virtual media repository can be excluded by using the **-nomedialib** flag.

You can back up and restore the Virtual I/O Server by the methods that are listed in Table 11-1.

Table 11-1 Virtual I/O Server backup and restore methods

Backup method	Media	Restore method
To tape	Tape	From tape
To DVD	DVD-RAM	From DVD
To remote file system	`nim_resources.tar` image	From an HMC by using the Network Installation Management on Linux (NIMOL) facility and the **installios** command
To remote file system `mksysb` image	`mksysb` image	From an AIX NIM server and a standard `mksysb` system installation

Backing up to tape

You can back up the Virtual I/O Server base code, applied fix packs, custom device drivers to support disk subsystems, and some user-defined metadata to tape.

If the system is managed by the Integrated Virtualization Manager, back up your partition profile data for the management partition and its client partitions before

you back up the Virtual I/O Server. For more information, see “Backing up IVM resources” on page 338. To back up to tape, complete these steps:

1. Assign a tape drive to the Virtual I/O Server.
2. You can find the device name on the Virtual I/O Server by typing the following command:

   ```
   $ lsdev -type tape
   name            status     description
   rmt0            Available  Other SCSI Tape Drive
   ```

   If the device is in the `Defined` state, type the following command where *dev* is the name of your tape device:

   ```
   $ cfgdev -dev dev
   ```

3. Run the **backupios** command with the **-tape** option. Specify the path to the device. Use the **-accept** flag to automatically accept licenses. For example:

   ```
   backupios -tape /dev/rmt0 -accept
   ```

Example 11-1 illustrates a **backupios** command execution to back up the Virtual I/O Server on a tape.

Example 11-1 Backing up the Virtual I/O Server to tape

```
$ backupios -tape /dev/rmt0

Creating information file for volume group volgrp01.

Creating information file for volume group storage01.
Backup in progress.  This command can take a considerable amount of time
to complete, please be patient...


Creating information file (/image.data) for rootvg.

Creating tape boot image..............

Creating list of files to back up.
Backing up 44950 files.........................
44950 of 44950 files (100%)
0512-038 mksysb: Backup Completed Successfully.
```

Backing up to a DVD-RAM

You can back up the Virtual I/O Server base code, applied fix packs, custom device drivers to support disk subsystems, and some user-defined metadata to DVD.

If the system is managed by the Integrated Virtualization Manager, back up your partition profile data for the management partition and its client partitions before you back up the Virtual I/O Server. For more information, see “Backing up IVM resources” on page 338.

To back up the Virtual I/O Server to one or more DVDs, generally use DVD-RAM media. Vendor disk drives might support burning to more disk types, such as CD-RW and DVD-R. See the documentation for your drive to determine which disk types are supported.

DVD-RAM media can support both **-cdformat** and **-udf format** flags. DVD-R media supports only the **-cdformat** option.

The DVD device cannot be virtualized and assigned to a client partition when you use the **backupios** command. Remove the device mapping from the client before you proceed with the backup.To back up on DVD-RAM, complete these steps:

1. Assign an optical drive to the Virtual I/O Server.
2. You can find the device name on the Virtual I/O Server by typing the following command:

   ```
   $ lsdev -type optical
   name            status     description
   cd0             Available  SATA DVD-RAM Drive
   ```

   If the device is in the `Defined` state, type the following command where *dev* is the name of your CD or DVD device:

   cfgdev -dev *dev*
3. Run the **backupios** command with the **-cd** option. Specify the path to the device. Use the **-accept** flag to automatically accept licenses. For example:

   ```
   backupios -cd /dev/cd0 -accept
   ```

Example 11-2 illustrates a **backupios** command execution to back up the Virtual I/O Server on a DVD-RAM.

Example 11-2 Backing up the Virtual I/O Server to DVD-RAM

```
$ backupios -cd /dev/cd0 -udf -accept

Creating information file for volume group volgrp01.

Creating information file for volume group storage01.
Backup in progress.  This command can take a considerable amount of time
to complete, please be patient...

Initializing mkcd log: /var/adm/ras/mkcd.log...
Verifying command parameters...
```

```
Creating image.data file...
Creating temporary file system: /mkcd/mksysb_image...
Creating mksysb image...

Creating list of files to back up.
Backing up 44933 files.........
44933 of 44933 files (100%)
0512-038 mksysb: Backup Completed Successfully.
Populating the CD or DVD file system...
Copying backup to the CD or DVD file system...
.........................................
.........................................
Building chrp boot image...
```

Tip: If the Virtual I/O Server does not fit on one DVD, the **backupios** command provides instructions for disk replacement and removal until all the volumes have been created.

Backing up to a remote file

The major difference for this type of backup is that all of the previous commands resulted in a form of bootable media that can be used to directly recover the Virtual I/O Server.

Backing up the Virtual I/O Server base code, applied fix packs, custom device drivers to support disk subsystems, and some user-defined metadata to a file results in one of these file types:

- A `nim_resources.tar` file that contains all the information that is needed for a restore. This is the preferred solution if you intend to restore the Virtual I/O Server on the same system. This backup file can be restored by either the HMC or a Network Installation Management (NIM) server.
- A `mksysb` image. This solution is preferred if you intend to restore the Virtual I/O Server from a NIM server.

Tip: The `mksysb` backup of the Virtual I/O Server can be extracted from the `tar` file that is created in a full backup. Therefore, either method is appropriate if the restoration method uses a NIM server.

If the system is managed by the IVM, back up your partition profile data for the management partition and its client partitions before you back up the Virtual I/O Server. For more information, see “Backing up IVM resources” on page 338.

You can use the **backupios** command to write to a local file on the Virtual I/O Server. However, the more common scenario is to perform a backup to a remote

NFS-based storage. The ideal situation might be to use the NIM server as the destination because this server can be used to restore these backups. In the following example, a NIM server has a host name of nim_server and the Virtual I/O Server is vios1.

The first step is to set up the NFS-based storage export on the NIM server. Export a file system named `/export/ios_backup` by using the following commands:

```
$ mkdir /export/ios_backup
$ mknfsexp -d /export/ios_backup -B -S sys,krb5p,krb5i,krb5,dh -t rw -r vios1
$ grep vios1 /etc/exports
/export/ios_backup -sec=sys:krb5p:krb5i:krb5:dh,rw,root=vios1
```

Important: The NFS server must have the root access NFS attribute set on the file system that is exported to the Virtual I/O Server partition for the backup to succeed.

In addition, make sure that the name resolution is functioning from the NIM server to the Virtual I/O Server and back again (reverse resolution) for both the IP address and host name. To edit the name resolution on the Virtual I/O Server, use the **hostmap** command to manipulate `/etc/hosts` or the **cfgnamesrv** command to change the DNS parameters.

The backup of the Virtual I/O Server can be large, so ensure that the system `ulimits` parameter in the `/etc/security/limits` file on the NIM server is set to `-1` to allow for the creation of large files.

After the NFS export and name resolution are set up, the file system must be mounted on the Virtual I/O Server. You can use the **mount** command:

```
$ mkdir /mnt/backup
$ mount nim_server:/export/ios_backup /mnt/backup
```

Remember: Mount the remote file system automatically at bootup of the Virtual I/O Server to simplify the scheduling of regular backups.

Backing up to a nim_resources.tar file

After the remote file system is mounted, you can start the backup operation to the `nim_resources.tar` file.

Backing up the Virtual I/O Server to a remote file system creates the `nim_resources.tar` image in the directory that you specify. The `nim_resources.tar` file contains all the necessary resources to restore the Virtual I/O Server. These include the mksysb image, the `bosinst.data` file, the network boot image, and the Shared Product Object Tree (SPOT) resource.

The **backupios** command empties the target_disks_stanza section of bosinst.data, and sets RECOVER_DEVICES=Default. This process allows the mksysb file that is generated by the command to be cloned to another logical partition. If you plan to use the nim_resources.tar image to install to a specific disk, repopulate the target_disk_stanza section of bosinst.data, and replace this file in the nim_resources.tar image. All other parts of the nim_resources.tar image must remain unchanged.

Run the **backupios** command with the **-file** option to specify the target directory path. For example:

```
backupios -file /mnt/backup
```

Example 11-3 illustrates a **backupios** command execution to back up the Virtual I/O Server on a nim_resources.tar file.

Example 11-3 Backing up the Virtual I/O Server to the nim_resources.tar file

```
$ backupios -file /mnt/backup

Creating information file for volume group storage01.

Creating information file for volume group volgrp01.
Backup in progress.  This command can take a considerable amount of time
to complete, please be patient...
```

This command creates a nim_resources.tar file that you can use to restore the Virtual I/O Server from the HMC as described in “Restoring from a nim_resources.tar file with the HMC” on page 357.

Remember: The argument for the **backupios -file** command is a directory. The nim_resources.tar file is stored in this directory.

Backing up to a mksysb file

Alternatively, after the remote file system is mounted, you can start the backup operation to a mksysb file. The mksysb image is an installable image of the root volume group in a file.

Run the **backupios** command with the **-file** option to specify the target directory path and the **-mksysb** option. For example:

```
backupios -file /mnt/backup -mksysb
```

Example 11-4 illustrates a **backupios** command execution to back up the Virtual I/O Server on a mksysb file.

Example 11-4 Backing up the Virtual I/O Server to the mksysb image

```
$ backupios -file /mnt/VIOS_BACKUP_13Oct2008.mksysb -mksysb

/mnt/VIOS_BACKUP_13Oct2008.mksysb  doesn't exist.

Creating /mnt/VIOS_BACKUP_13Oct2008.mksysb

Creating information file for volume group storage01.

Creating information file for volume group volgrp01.
Backup in progress.  This command can take a considerable amount of time
to complete, please be patient...


Creating information file (/image.data) for rootvg.

Creating list of files to back up...
Backing up 45016 files.........................
45016 of 45016 files (100%)
0512-038 savevg: Backup Completed Successfully.
```

Remember: If you intend to use a NIM server for the restoration, it must be running a level of AIX that can support the Virtual I/O Server installation. For this reason, always run the NIM server with the latest technology level and service packs:

- For the restoration of any backups of a Virtual I/O Server version 2.1, your NIM server must be at the latest AIX Version 6.1 level.
- For a Virtual I/O Server 1.x environment, your NIM server must be at the latest AIX Version 5.3 level.

11.1.4 Backing up user-defined virtual devices

After you back up the Virtual I/O Server operating system, you also must back up the user-defined virtual devices:

- If you are restoring to the same server, the necessary data structures such as storage pools or volume groups and logical volumes might be held on non-rootvg disks.
- If you are restoring to new hardware, these devices cannot be automatically recovered because the disk structures do not exist.

- If the physical devices exist in the same location and structures such as logical volumes are intact, the virtual devices such as virtual target SCSI and Shared Ethernet Adapters are recovered during the restoration.

In a situation where the disk structures do not exist and adapters are at different location codes, the following must be backed up:

- Any user-defined disk structures such as storage pools or volume groups and logical volumes.
- The linking of the virtual device through to the physical devices.

User-defined virtual devices include metadata, such as virtual device mappings, that define the relationship between the physical and the virtual environment. You can back up this data in the following ways:

- Saving the configuration information manually to a location that is automatically backed up when the **backupios** command is run. This is required if you want to restore the configuration to a separate Virtual I/O Server.
- Using the **viosbr** command, you can save the user-defined virtual device configuration and restore it to the same Virtual I/O Server from where it was backed up.

The following sections describe both methods in more detail.

Backing up user-defined virtual devices by using viosbr

The **viosbr** command backs up all relevant data to recover a Virtual I/O Server after an installation:

Logical devices	For example, logical volume or file-backed storage pool, virtual media repository, and paging space device configurations
Virtual devices	For example, Shared Ethernet Adapter, virtual SCSI server, and virtual Fibre Channel server adapter configurations
Device attributes	For example, device attributes for disk, network, or Fibre Channel devices

Example 11-5 shows an example of the **viosbr** command. Using the **-file** option, you can specify the name of the backup file. By default, the file is written to the `cfgbackups` subdirectory in the home directory of the padmin user. You can also specify another location by specifying a directory and file name like **-file /tmp/backup**.

After the backup has finished, you can display the backup file by using the **viosbr -view -list** command

Example 11-5 Performing a backup by using the viosbr command

```
$ viosbr -backup -file backup_1
Backup of this node(P7_1_vios2) successfull
$ viosbr -view -list
backup_1.tar.gz
```

The backup of an SSP cluster can be done with **viosbr** as shown in Example 11-6.

Example 11-6 Backing up the SSP configuration

```
vios01:/home/padmin # viosbr -backup -clustername clusterA -file backup
Backup of node vios02 successful
Backup of node vios03 successful
Backup of node vios04 successful
Backup of this node (vios01) successful
```

The backup can be viewed by using the **viosbr -view -list** command, and is stored in the`/home/padmin/cfgbackups/backup.clusterA.tar.gz` file.

Scheduling regular backups by using the viosbr command

Using the **viosbr** command, regular backups of the user-defined virtual device configuration can be scheduled on a daily, weekly, or monthly basis. Example 11-7 shows how to schedule a daily backup where a rotation of 10 backups are kept.

Example 11-7 Scheduling regular backups by using the viosbr command

```
$ viosbr -backup -file backup -frequency daily -numfiles 10
Backup of this node(P7_1_vios2) successfull
```

Backing up user-defined virtual devices by using backupios

Restoring the user-defined virtual device configuration to another Virtual I/O Server requires the **backupios** command.

The following three categories of configuration data are required to rebuild the Virtual I/O Server configuration after a restore:

Disk structures These are used-defined disk structures like the volume group information.

Device mappings The mappings define the link between the physical devices and the virtual devices.

Extra information Virtual I/O Server configuration data like network routing information, tuning settings, and security settings.

Backing up disk structures with savevgstruct

Use the **savevgstruct** command to back up user-defined disk structures. This command writes a backup of the structure of a named volume group (and therefore the storage pool) to the `/home/ios/vgbackups` directory.

For example, assume that you have the following storage pools:

```
$ lssp
Pool            Size(mb)   Free(mb)   Alloc Size(mb)  BDs
rootvg          139776     107136     128             0
storage01       69888      69760      64              1
volgrp01        69888      69760      64              1
```

Run the **savevgstruct storage01** command to back up the structure in the storage01 volume group:

```
$ savevgstruct storage01

Creating information file for volume group storage01.

Creating list of files to back up.
Backing up 6 files
6 of 6 files (100%)
0512-038 savevg: Backup Completed Successfully.
```

The **savevgstruct** command is automatically called by the **backupios** command to back up all active non-rootvg volume groups or storage pools before it backs up rootvg on a Virtual I/O Server. These volume group structures are included in the system backup.

Remember: The **backupios** command will only back up active volume groups. Use the **lspv** command to list volume groups names, and the **lsvg** command to list volume group disk structure. To activate a volume group, use the **activatevg** command if necessary before you start the backup.

Backing up virtual device linking information

The last item to back up is the linking information. You can gather this information from the output of the **lsmap** command as shown in Example 11-8.

Example 11-8 Sample output from the lsmap command

```
$ lsmap -net -all
SVEA   Physloc
------ --------------------------------------------
ent2   U9117.MMA.101F170-V1-C11-T1

SEA                  ent5
Backing device       ent0
Status               Available
Physloc              U789D.001.DQDYKYW-P1-C4-T1

$ lsmap -all
SVSA            Physloc                                        Client Partition
ID
--------------- ---------------------------------------------- ------------------
vhost0          U9117.MMA.101F170-V1-C21                       0x00000003

VTD                  aix61_rvg
Status               Available
LUN                  0x8100000000000000
Backing device       hdisk7
Physloc
U789D.001.DQDYKYW-P1-C2-T2-W201300A0B811A662-L1000000000000

SVSA            Physloc                                        Client Partition
ID
--------------- ---------------------------------------------- ------------------
vhost1          U9117.MMA.101F170-V1-C22                       0x00000004

VTD                  aix53_rvg
Status               Available
LUN                  0x8100000000000000
Backing device       hdisk8
Physloc
U789D.001.DQDYKYW-P1-C2-T2-W201300A0B811A662-L2000000000000
```

The important information to note is the mapping of virtual devices to physical devices. The vhost and ent numbers are assigned sequentially by the Virtual I/O Server at initial discovery time or when the **cfgdev** command is run. Take caution when you rebuild user-defined linking devices based on the virtual device names. In this example, the important information for vhost0 is that a virtual SCSI device in slot 21 (the C21 value in the location code) is mapped to hdisk7, not that the device name is vhost0.

Consideration: The previous output does not gather information such as SEA control channels (for SEA failover), IP addresses to ping, and whether threading is enabled for the SEA devices. These settings and any other changes that have been made (for example MTU settings) must be documented separately, as explained later in this section.

Backing up shared storage pool information

If you are using a shared storage pool, collect the cluster, storage pool, and logical unit configurations. You can get this information by using the **lscluster** command and the **lssp** command as shown in Example 11-9. The **lscluster** command displays the cluster name and the repository and storage pool disks. The **lssp** command displays the pool name and the logical unit configuration. For more information about listing the logical units mapping, see 10.2.4, “Tracing logical units” on page 317.

Example 11-9 Displaying the shared storage pool information

```
$ lscluster -d
Storage Interface Query

Cluster Name:  clusterA
Cluster uuid:  8e167044-0155-11e0-a50f-f61aa6a64371
Number of nodes reporting = 1
Number of nodes expected = 1
Node P7_1_vios1
Node uuid = 3a783312-f36f-11df-b987-00145ee9e161
Number of disk discovered = 3
        cldisk2
          state : UP
          uDid  : 200B75BALB1111507210790003IBMfcp
          uUid  : 5d7eace7-5213-07d6-e080-bcc62ff95386
          type  : CLUSDISK
        cldisk1
          state : UP
          uDid  : 200B75BALB1111407210790003IBMfcp
          uUid  : 42499f99-5563-89ae-9453-07ff800a7e91
          type  : CLUSDISK
        caa_private0
          state : UP
          uDid  :
          uUid  : 1264a0af-4a7e-fb93-62d7-6a33d5f17f35
          type  : REPDISK
$ lssp -clustername clusterA
Pool                Size(mb)  Free(mb)  LUs       Type   PoolID
poolA               40704     40409     2         CLPOOL 2683031503901849366
```

```
$ lssp -clustername clusterA -sp poolA -bd
Lu(Disk) Name                           Size(MB)       Lu Udid
test1                                   10
6beea81a3d25723c9e3d1e72df34a296
test2                                   20
971d5586da279b2fa9a15d089c812514
```

Backing up extra information

Save the information about network settings, adapters, users, and security settings to the `/home/padmin` directory by running each command with the **`tee`** command as follows:

`command | tee /home/padmin/filename`

Where:

– `command` is the command that produces the information you want to save.

– `filename` is the name of the file to which you want to save the information.

The `/home/padmin` directory is backed up by using the **`backupios`** command. Therefore, it is a good location to collect configuration information before a backup. Table 11-2 provides a summary of the commands that help you to save the information.

Table 11-2 Commands to save information about Virtual I/O Server

Command	Information provided (and saved)
`cfgnamesrv -ls`	Shows all system configuration database entries that are related to the domain name server information used by local resolver routines.
`entstat -all devicename` devicename is the name of a device. Run this command for each device whose attributes or statistics you want to save.	Shows Ethernet driver and device statistics for the device specified.
`hostmap -ls`	Shows all entries in the system configuration database.
`ioslevel`	Shows the current maintenance level of the Virtual I/O Server.

Command	Information provided (and saved)
`lsdev -dev devicename -attr` devicename is the name of a device. Run this command for each device whose attributes or statistics you want to save. You generally want to save the customized devices attributes. Try to keep track of them when you are managing the Virtual I/O Server.	Shows the attributes of the device specified.
`lsdev -type adapter`	Shows information about physical and logical adapters.
`lsuser`	Shows a list of all attributes of all system users.
`netstat -routinfo`	Shows the routing tables, including the user-configured and current costs of each route.
`netstat -state`	Shows the state of the network, which includes errors, collisions, and packets transferred.
`optimizenet -list`	Shows characteristics of all network tuning parameters, including the current and reboot value, range, unit, type, and dependencies.
`viosecure -firewall view`	Shows a list of allowed ports.
`viosecure -view -nonint`	Shows all of the security level settings for non-interactive mode.

11.1.5 Backing up using IBM Tivoli Storage Manager

You can use the IBM Tivoli Storage Manager to automatically back up the Virtual I/O Server on regular intervals.

Configuring the IBM Tivoli Storage Manager client

Virtual I/O Server includes the IBM Tivoli Storage Manager agent. With Tivoli Storage Manager, you can protect your data from failures and other errors by storing backup and disaster recovery data in a hierarchy of auxiliary storage. Tivoli Storage Manager can help protect computers that run various operating environments, including the Virtual I/O Server, on various hardware, including IBM Power Systems servers. Configuring the Tivoli Storage Manager client on

the Virtual I/O Server enables you to include the Virtual I/O Server in your standard backup framework.

Complete these steps to configure the Tivoli Storage Manager agent:

1. List all the attributes that are associated with the agent configuration:

```
$ cfgsvc -ls TSM_base
SERVERNAME
SERVERIP
NODENAME
```

2. Configure the agent:

```
$ cfgsvc TSM_base -attr SERVERNAME=hostname
SERVERIP=name_or_address NODENAME=vios
```

Where:

- *hostname* is the host name of the Tivoli Storage Manager server with which the Tivoli Storage Manager client is associated.
- *name_or_address* is the IP address or domain name of the Tivoli Storage Manager server with which the Tivoli Storage Manager client is associated.
- *vios* is the name of the system on which the Tivoli Storage Manager client is installed. The name must match the name that is registered on the Tivoli Storage Manager server.

Ask the Tivoli Storage Manager administrator to register the client node, the Virtual I/O Server, with the Tivoli Storage Manager server. To determine what information you must provide to the Tivoli Storage Manager administrator, see the IBM Tivoli Storage Manager documentation at:

http://publib.boulder.ibm.com/infocenter/tivihelp/v1r1/index.jsp

After you are finished, you are ready to back up and restore the Virtual I/O Server by using Tivoli Storage Manager.

IBM Tivoli Storage Manager automated backup

You can automate backups of the Virtual I/O Server by using the **crontab** or and the Tivoli Storage Manager scheduler.

Before you start, complete the following tasks:

- Ensure that you configured the Tivoli Storage Manager client on the Virtual I/O Server.
- Ensure that you are logged in to the Virtual I/O Server as the administrator (padmin).

To automate backups of the Virtual I/O Server, complete the following steps:

1. Write a script that creates a `mksysb` image of the Virtual I/O Server and save it in a directory that is accessible to the padmin user ID. For example, create a script called `backup` and save it in the `/home/padmin` directory. If you plan to restore the Virtual I/O Server to a different system than the one from which it was backed up, ensure that your script includes commands for saving information about user-defined virtual devices. For more information, see the following tasks:
   - For more information about how to create a `mksysb` image, see “Backing up to a mksysb file” on page 344.
   - For more information about how to save user-defined virtual devices, see 11.1.4, “Backing up user-defined virtual devices” on page 345.
2. Create a `crontab` file entry that runs the backup script on a regular interval. For example, to create a `mksysb` image every Saturday at 2:00 a.m., enter the following commands:

   ```
   $ crontab -e
   0 2 * * 6 /home/padmin/backup
   ```

   Or work with the Tivoli Storage Manager administrator who can create a schedule on Tivoli Storage Manager server that runs the script `/home/padmin/backup` automatically for you. Then, start the client schedule scheduler for the Virtual I/O Server and add the following entry to the `/etc/inittab` file:

   ```
   itsm::once:/usr/bin/dsmc sched > /dev/null 2>&1 # TSM scheduler
   ```

Running IBM Tivoli Storage Manager backup from the CLI

You can back up the Virtual I/O Server at any time by running an incremental backup with the Tivoli Storage Manager.

Perform incremental backups in situations where the automated backup does not suit your needs. For example, before you upgrade the Virtual I/O Server, run an incremental backup to ensure that you have a backup of the current configuration. After you upgrade the Virtual I/O Server, run another incremental backup to ensure that you have a backup of the upgraded configuration.

Before you start, complete the following tasks:

- Ensure that you configured the Tivoli Storage Manager client on the Virtual I/O Server. For more information, see “Configuring the IBM Tivoli Storage Manager client” on page 352.
- Ensure that you have a `mksysb` image of the Virtual I/O Server. If you plan to restore the Virtual I/O Server to a separate system than the one from which it

was backed up, ensure that the `mksysb` includes information about the user-defined virtual devices.

To perform an incremental backup of the Virtual I/O Server, run the **`dsmc`** command. For example:

```
dsmc incremental sourcefilespec
```

where `sourcefilespec` is the directory path to where the `mksysb` file is located. For example, `/home/padmin/mksysb_image`.

11.1.6 Planning backups of the Virtual I/O Server

You can schedule regular backups of the Virtual I/O Server and user-defined virtual devices to ensure that your backup copy accurately reflects the current configuration.

To ensure that your backup of the Virtual I/O Server accurately reflects your current running Virtual I/O Server, back up the Virtual I/O Server each time that its configuration changes:

- Changing the Virtual I/O Server, such as by installing a fix pack.
- Adding, deleting, or changing the external device configuration, such as changing the SAN configuration.
- Adding, deleting, or changing resource allocations and assignments for the Virtual I/O Server, such as memory, processors, or virtual and physical devices.
- Adding, deleting, or changing user-defined virtual device configurations, such as virtual device mappings.

Back up your Virtual I/O Server manually after any of these modifications, or schedule regular backups by using the crontab function or the Tivoli Storage Manager scheduler.

11.1.7 Restoring the Virtual I/O Server

Once you have created a backup by using one of the techniques described, you can then rebuild the server from scratch. The example scenario used in this section involves a Virtual I/O Server hosting an AIX operating system-based client partition that is running on virtual disk and network. The restore is described from the uninstalled bare metal Virtual I/O Server upward, and guidelines are provided on where to use each backup strategy.

This complete end-to-end solution is only for this extreme disaster recovery scenario. If you must back up and restore a Virtual I/O Server onto the same server, the restoration of the operating system is probably of interest.

Restoring the HMC configuration

In the most extreme case, a natural or man-made disaster renders an entire data center unusable. In this case, systems must be restored to a disaster recovery site. Therefore, you need another HMC and server location to which to recover your settings. Also, have a disaster recovery server in place with your HMC profiles ready to start recovering your systems.

The details of this scenario are beyond the scope of this document but are, along with the following section, the first steps for a disaster recovery.

Restoring other IT infrastructure devices

All other IT infrastructure devices, such as network routers, switches, storage area networks, and DNS servers, also must be part of an overall IT disaster recovery solution. Not just the Virtual I/O Server, but the whole IT infrastructure relies on these common services for a successful recovery.

Restoring the Virtual I/O Server operating system

This section details how to restore the Virtual I/O Server from a complete disaster.

If you upgraded to a separate system and this system is managed by the Integrated Virtualization Manager, restore your partition profile data for the management partition and its client partitions before you restore the Virtual I/O Server.

To do so, click **Service Management** → **Backup/Restore**. The Backup/Restore page is displayed. Then, click **Restore Partition Configuration**.

Restoring from DVD backup

The backup procedures that are described in this chapter created bootable media that you can use to restore as stand-alone backups.

Insert the first DVD into the DVD drive and boot the Virtual I/O Server partition into SMS mode, making sure that the DVD drive is assigned to the partition. Using the SMS menus, select the option to install from the DVD drive and work through the usual installation procedure.

Consideration: If the DVD backup spanned multiple disks during the installation, you will be prompted to insert the next disk in the set with a message similar to the following:

```
Please remove volume 1, insert volume 2, and press the ENTER key.
```

Restoring from tape backup

The procedure for the tape is similar to the DVD procedure. Because it is a bootable media, place the backup media into the tape drive and boot the Virtual I/O Server partition into SMS mode. Select to install from the tape drive and follow the same procedure as previously described.

Restoring from a nim_resources.tar file with the HMC

If you made a full backup of the Virtual I/O Server to a `nim_resources.tar` file, use the HMC to restore it using the **`installios`** command.

To do so, the tar file must be located either on the HMC, an NFS-accessible directory, or a DVD. To make the `nim_resources.tar` file accessible for restore, complete the following steps:

1. Create a directory named `backup` by using the **`mkdir /home/padmin/backup`** command.
2. Check that the NFS server was exporting a file system with the **`showmount nfs_server`** command.
3. Mount the NFS-exported file system onto the `/home/padmin/backup` directory.
4. Copy the `tar` file that you created in “Backing up to a nim_resources.tar file” on page 343 to the NFS mounted directory by using the following command:

   ```
   $ cp /home/padmin/backup_loc/nim_resources.tar /home/padmin/backup
   ```

At this stage, the backup is ready to be restored to the Virtual I/O Server partition by using the **`installios`** command on the HMC or an AIX partition that is a NIM server. The restore procedure shuts down the Virtual I/O Server partition if it is still running. The following is an example of the command help:

```
hscroot@hmc1:~> installios -?
installios: usage: installios [-s managed_sys -S netmask -p partition
                          -r profile -i client_addr -d source_dir -m mac_addr
                          -g gateway [-P speed] [-D duplex] [-n] [-l language]]
                          | -u
```

When you use the **`installios`** command, the **`-s`** managed_sys option requires the HMC defined system name. The **`-p`** partition option requires the name of the Virtual I/O Server partition, and the **`-r`** profile option requires the partition profile that you want to use to boot the Virtual I/O Server partition during the recovery.

If you do not specify the **-m** flag and include the MAC address of the Virtual I/O Server being restored, the restore takes longer because the **installios** command shuts down the Virtual I/O Server and boots it in SMS to determine the MAC address. The following is an example of the use of this command:

```
hscroot@hmc1:~> installios -s MT_B_p570_MMA_101F170 -S 255.255.254.0 -p vios1
-r default -i 9.3.5.111 -d 9.3.5.5:/export_fs -m 00:02:55:d3:dc:34 -g 9.3.4.1
```

Tip: If you do not input a parameter, the **installios** command prompts you for one:

```
hscroot@hmc1:~> installios

The following objects of type "managed system" were found.  Please select
one:

1. MT_B_p570_MMA_101F170
2. MT_A_p570_MMA_100F6A0
3. p550-SN106629E

Enter a number (1-3): 1

The following objects of type "virtual I/O server partition" were found.
Please select one:

1. vios2
2. vios1

Enter a number (1-2):
```

Open a terminal console on the server to which you are restoring in case user input is required. Then, run the **installios** command as described previously.

NIMOL on the HMC then takes over the NIM process and mounts the exported file system to process the **backupios** `tar` file that was created on the Virtual I/O Server previously. NIMOL then installs the Virtual I/O Server, and a reboot of the partition completes the installation.

Tips:

- The configure client network setting must be set to `no` when prompted by the **installios** command. This is because the physical adapter you are installing the backup through might already be used by an SEA. In this case, the IP configuration fails. Log in and configure the IP if necessary after the installation using a console session.
- If the command seems to be taking a long time to restore, the reason is usually a speed or duplex misconfiguration in the network.

Restoring from a file with the NIM server

The **installios** command is also available on the NIM server, but currently only supports installations from the base media of the Virtual I/O Server. The method used in the example from the NIM server was to install the mksysb image. This can either be the mksysb image generated with the **-mksysb** flag in the **backupios** command, or you can extract the mksysb image from the nim_resources.tar file.

Note: To use a NIM, ensure that the NIM Master is at the appropriate level to support the Virtual I/O Server image.

Whatever method you use, after you store the mksysb file on the NIM server, create a NIM mksysb resource as shown:

```
# nim -o define -t mksysb -aserver=master
-alocation=/export/mksysb/VIOS_BACKUP_13Oct2008.mksysb VIOS_mksysb
# lsnim VIOS_mksysb
VIOS_mksysb     resources       mksysb
```

After NIM mksysb resource is successfully created, generate a SPOT from the NIM mksysb resource or use the SPOT available at the latest AIX technology and service pack level. To create the SPOT from the NIM mksysb resource, run the following command:

```
# nim -o define -t spot -a server=master -a location=/export/spot/ -a
source=VIOS_mksysb VIOS_SPOT

Creating SPOT in "/export/spot" on machine "master" from "VIOS_mksysb" ...
Restoring files from BOS image.  This may take several minutes ...
# lsnim VIOS_SPOT
VIOS_SPOT     resources       spot
```

With the SPOT and the mksysb resources defined to NIM, you can install the Virtual I/O Server from the backup. If the Virtual I/O Server partition you are installing is not defined to NIM, make sure that it is now defined as a machine. Then enter the **smitty nim_bosinst** fast path command. Select the NIM mksysb resource and SPOT defined previously.

Important: The Remain NIM client after the install field must be set to `no`. If it is not, the last step of the NIM installation is configuring an IP address onto the physical adapter through which the Virtual I/O Server was installed. This IP address is used to register with the NIM server. If this is the adapter used by an existing SEA, it causes error messages to be displayed.

If so, reboot the Virtual I/O Server if necessary, and then log in to it using a terminal session and remove any IP address information and the SEA. After doing so, re-create the SEA and configure the IP address back for the SEA interface.

After setting up the NIM server to push out the backup image, the Virtual I/O Server partition must have the remote IPL setup completed. For more information, see “Installing with Network Installation Management” under the *Installation and Migration* category of the IBM System p® and AIX Information Center at:

`http://publib16.boulder.ibm.com/pseries/index.htm`

Tip: One of the main causes of installation problems when using NIM is the NFS exports from the NIM server. Make sure that the `/etc/exports` file is correct on the NIM server.

The installation of the Virtual I/O Server completes, but here is a significant difference between restoring to the existing server and restoring to a new disaster recovery server. One of the NIM installation options is to preserve the NIM definitions for resources on the target. With this option, NIM attempts to restore any virtual devices that were defined in the original backup. This process depends on the devices being defined in the partition profile (virtual and physical) such that the location codes are not changed.

This means that virtual target SCSI devices and Shared Ethernet Adapters are all recovered without any need to re-create them, assuming the logical partition profile has not changed. If you are restoring to the same machine, the non-rootvg volume groups must be present to be imported, and any logical volume structures on them are intact.

To demonstrate this, the following scenario was tested: A Virtual I/O Server was booted from a diagnostics CD and the Virtual I/O Server operating system disks were formatted and certified, deleting all data. The other disks that contained the volume groups and storage pools were not touched.

Using a NIM server, the backup image was restored to the initial Virtual I/O Server operating system disks. Examining the virtual devices after the

installation, the virtual target devices and Shared Ethernet Adapters are all recovered as shown in Example 11-10.

Example 11-10 Restoration of Virtual I/O Server to the same logical partition

```
$ lsdev -virtual
name             status      description
ent2             Available   Virtual I/O Ethernet Adapter (l-lan)
ent3             Available   Virtual I/O Ethernet Adapter (l-lan)
ent4             Available   Virtual I/O Ethernet Adapter (l-lan)
ent6             Available   Virtual I/O Ethernet Adapter (l-lan)
vasi0            Available   Virtual Asynchronous Services Interface (VASI)
vbsd0            Available   Virtual Block Storage Device (VBSD)
vhost0           Available   Virtual SCSI Server Adapter
vhost1           Available   Virtual SCSI Server Adapter
vhost2           Available   Virtual SCSI Server Adapter
vhost3           Available   Virtual SCSI Server Adapter
vhost4           Available   Virtual SCSI Server Adapter
vsa0             Available   LPAR Virtual Serial Adapter
IBMi61_0         Available   Virtual Target Device - Disk
IBMi61_1         Available   Virtual Target Device - Disk
aix53_rvg        Available   Virtual Target Device - Disk
aix61_rvg        Available   Virtual Target Device - Disk
rhel52           Available   Virtual Target Device - Disk
sles10           Available   Virtual Target Device - Disk
vtopt0           Defined     Virtual Target Device - File-backed Optical
vtopt1           Defined     Virtual Target Device - File-backed Optical
vtopt2           Defined     Virtual Target Device - File-backed Optical
vtopt3           Available   Virtual Target Device - Optical Media
vtscsi0          Defined     Virtual Target Device - Disk
vtscsi1          Defined     Virtual Target Device - Logical Volume
ent5             Available   Shared Ethernet Adapter
ent7             Available   Shared Ethernet Adapter

$ lsmap -all
SVSA            Physloc                                     Client Partition
ID
--------------- ------------------------------------------- ------------------
vhost0          U9117.MMA.101F170-V1-C21                    0x00000003

VTD                   aix61_rvg
Status                Available
LUN                   0x8100000000000000
Backing device        hdisk7
Physloc U789D.001.DQDYKYW-P1-C2-T2-W201300A0B811A662-L1000000000000

SVSA            Physloc                                     Client Partition
ID
--------------- ------------------------------------------- ------------------
```

```
vhost1          U9117.MMA.101F17O-V1-C22                    0x00000004

VTD                   aix53_rvg
Status                Available
LUN                   0x8100000000000000
Backing device        hdisk8
Physloc U789D.001.DQDYKYW-P1-C2-T2-W201300A0B811A662-L2000000000000

SVSA            Physloc                                     Client Partition
ID
--------------- ------------------------------------------- ------------------
vhost2          U9117.MMA.101F17O-V1-C23                    0x00000005

VTD                   IBMi61_0
Status                Available
LUN                   0x8100000000000000
Backing device        hdisk11
Physloc U789D.001.DQDYKYW-P1-C2-T2-W201300A0B811A662-L5000000000000

VTD                   IBMi61_1
Status                Available
LUN                   0x8200000000000000
Backing device        hdisk12
Physloc U789D.001.DQDYKYW-P1-C2-T2-W201300A0B811A662-L6000000000000

SVSA            Physloc                                     Client Partition
ID
--------------- ------------------------------------------- ------------------
vhost3          U9117.MMA.101F17O-V1-C24                    0x00000006

VTD                   rhel52
Status                Available
LUN                   0x8100000000000000
Backing device        hdisk10
Physloc U789D.001.DQDYKYW-P1-C2-T2-W201300A0B811A662-L4000000000000

SVSA            Physloc                                     Client Partition
ID
--------------- ------------------------------------------- ------------------
vhost4          U9117.MMA.101F17O-V1-C60                    0x00000003

VTD                   NO VIRTUAL TARGET DEVICE FOUND

$ lsmap -net -all
SVEA   Physloc
------ --------------------------------------------
ent2   U9117.MMA.101F17O-V1-C11-T1

SEA                   ent5
```

```
Backing device        ent0
Status                Available
Physloc
U789D.001.DQDYKYW-P1-C4-T1
```

If you restore to a different logical partition where you defined similar virtual devices by using the HMC recovery step provided previously, there are no linking devices.

Consideration: The devices are always different between machines because the machine serial number is part of the virtual device location code for virtual devices. For example:

```
$ lsdev -dev ent4 -vpd
  ent4            U8204.E8A.10FE411-V1-C11-T1  Virtual I/O Ethernet Adapter
(l-lan)

        Network Address.............C21E4467D40B
        Displayable Message.........Virtual I/O Ethernet Adapter (l-lan)
        Hardware Location Code......U8204.E8A.10FE411-V1-C11-T1

  PLATFORM SPECIFIC

  Name:  l-lan
    Node:  l-lan@3000000b
    Device Type:  network
    Physical Location: U8204.E8A.10FE411-V1-C11-T1
```

This is because the backing devices are not present for the linking to occur. The physical location codes have changed, and thus the mapping fails.
Example 11-11 shows the same restore of the Virtual I/O Server originally running on a Power 570 onto a Power 550 that has the same virtual devices defined in the same slots.

Example 11-11 Devices recovered if restored to a different server

```
$ lsdev -virtual
name            status                                          description
ent2            Available  Virtual I/O Ethernet Adapter (l-lan)
vhost0          Available  Virtual SCSI Server Adapter
vsa0            Available  LPAR Virtual Serial Adapter

$ lsmap -all -net
SVEA   Physloc
------ --------------------------------------------
ent4   U8204.E8A.10FE411-V1-C11-T1
```

```
SEA                  ent6
Backing device       ent0
Status               Available
Physloc              U78A0.001.DNWGCV7-P1-C5-T1

$ lsmap -all
SVSA            Physloc                                    Client Partition
ID
--------------- ------------------------------------------ ------------------
vhost0          U9117.MMA.101F170-V1-C10                   0x00000003

VTD                  NO VIRTUAL TARGET DEVICE FOUND
```

You now need to recover the user-defined virtual devices and any backing disk structure.

Restoring with IBM Tivoli Storage Manager

You can use the IBM Tivoli Storage Manager to restore the `mksysb` image of the Virtual I/O Server.

Attention: The IBM Tivoli Storage Manager can restore the Virtual I/O Server only to the system from which it was backed up.

First, restore the `mksysb` image of the Virtual I/O Server by using the **`dsmc`** command on the Tivoli Storage Manager client. Restoring the `mksysb` image does not restore the Virtual I/O Server. You then must transfer the mksysb image to another system and convert the mksysb image to an installable format.

Before you start, complete the following tasks:

1. Ensure that the system to which you plan to transfer the mksysb image is running AIX.
2. Ensure that the system that is running AIX has a DVD-RW or CD-RW drive.
3. Ensure that AIX has the cdrecord and mkisofs RPMs downloaded and installed. To download and install the RPMs, see the AIX Toolbox for Linux Applications website at:

   `http://www.ibm.com/systems/p/os/aix/linux`

Attention: Interactive mode is not supported on the Virtual I/O Server. You can view session information by typing the **`dsmc`** command on the Virtual I/O Server command line.

To restore the Virtual I/O Server by using Tivoli Storage Manager, complete the following tasks:

1. Determine which file you want to restore by running the **dsmc** command to display the files that have been backed up to the Tivoli Storage Manager server:

   `dsmc query backup` *filespec*

   where *filespec* represents the path and file name that you want to query. For example, `/home/padmin/mksysb_image`.

2. Restore the mksysb image by using the **dsmc** command. For example:

   `dsmc restore` *filespec <destination>*

   If you do not specify a destination, the files are restored to their original location.

3. Transfer the `mksysb` image to a server with a DVD-RW or CD-RW drive by running the following File Transfer Protocol (FTP) commands:

   a. Run the following command to make sure that the FTP server is started on the Virtual I/O Server:

      `startnetsvc ftp`

   b. Open an FTP session to the server with the DVD-RW or CD-RW drive:

      `ftp` *server_hostname*

      where *server_hostname* is the host name of the server with the DVD-RW or CD-RW drive.

   c. At the FTP prompt, change the installation directory to the directory where you want to save the `mksysb` image.

   d. Set the transfer mode to binary by running the **binary** command.

   e. Turn off interactive prompting by using the **prompt** command.

   f. Transfer the `mksysb` image to the server by running the **mput mksysb_image** command.

   g. Close the FTP session after you transfer the `mksysb` image by entering the **quit** command.

4. Write the `mksysb` image to CD or DVD by using the **mkcd** or **mkdvd** commands.

Reinstall the Virtual I/O Server using the CD or DVD that you just created. For more information, see “Restoring from DVD backup” on page 356. Or reinstall the Virtual I/O Server from a NIM server. For more information, see “Restoring from a file with the NIM server” on page 359.

Recovering user-defined virtual devices and disk structure

The recovery of the user-defined virtual devices depends on the backup method that was used.

If you used the **viosbr** command to perform the backup, the **viosbr** command can be used to restore the configuration as described in “Restoring user-defined virtual devices by using viosbr” on page 366.

If the configuration data was saved as part of a **backupios** command operation, you can restore the user-defined virtual device configuration manually by using the saved information as described in “Manually restoring user-defined virtual devices” on page 369.

Restoring user-defined virtual devices by using viosbr

Before you restore a backup file that was generated using the viosbr command, display the content of the file by using the **-view** option of the **viosbr** command.

For example, the command **viosbr -view -file backup_1.tar.gz** shown in Example 11-12 displays the entities that were backed up in “Backing up user-defined virtual devices by using viosbr” on page 346. Using the **-mapping** option, the mappings to the virtual adapters can be displayed similar to the way they are displayed by the **lsmap** command.

Example 11-12 Using viosbr -view to display backup contents

```
$ viosbr -view -file backup_1.tar.gz

Details in:
------------------------------------------------------------

Controllers:
============
Name                Phys Loc
----                --------
iscsi0
sissas0             U5802.001.0086848-P1-C1-T1
pager0              U8233.E8B.061AA6P-V2-C32769-L0-L0
vasi0               U8233.E8B.061AA6P-V2-C32769
vbsd0               U8233.E8B.061AA6P-V2-C32769-L0
sata0               U5802.001.0086848-P1-C1-T1
fcs0                U5802.001.0086848-P1-C3-T1
fcs1                U5802.001.0086848-P1-C3-T2
.
. (Lines omitted for clarity)
.
Physical Volumes:
=================
```

```
Name                Phys Loc
----                --------
hdisk20 U5802.001.0086848-P1-C3-T2-W500507630419C12C-L4011401700000000
hdisk21 U5802.001.0086848-P1-C3-T2-W500507630419C12C-L4011401800000000
hdisk22 U5802.001.0086848-P1-C3-T1-W500507630414C12C-L4011401900000000
.
. (Lines omitted for clarity)
.

Optical Devices:
================
Name                Phys Loc
----                --------

Tape Devices:
=============
Name                Phys Loc
----                --------

Ethernet Interfaces:
====================
Name
----
en0
en1
en2
en3
en4
en5
en6
en7

Storage Pools:
==============
SP Name                         PV Name
-------                         -------
rootvg                          hdisk0
my_vg                           hdisk8
lv_pool                         hdisk7
caavg_private                   caa_private0

File Backed Storage Pools:
==========================
Name                            Parent SP
----                            ---------
fb_pool                         lv_pool

Optical Repository:
===================
```

```
Name                        Parent SP
----                        ---------
VMLibrary                   lv_pool

Shared Ethernet Adapters:
=========================
Name                Physical Adapter    Default Adapter     Virtual Adapters
----                ----------------    ---------------     ----------------
ent7                ent0                ent5                ent5


Virtual Server Adapters:
========================
SVSA                Phys Loc                      VTD
----                --------                      ---
vhost0              U8233.E8B.061AA6P-V2-C34      vtscsi0
                                                  vtscsi3
                                                  vtopt0

vhost1              U8233.E8B.061AA6P-V2-C54      vtscsi4
                                                  vtscsi5

vhost2              U8233.E8B.061AA6P-V2-C55
vhost3              U8233.E8B.061AA6P-V2-C64
vhost4              U8233.E8B.061AA6P-V2-C65

Cluster:
========
Cluster             State
-------             -----
cluster0            UP

Cluster Name        Cluster ID
------------        ---------------------------------

Attribute Name      Attribute Value
--------------      ---------------
node_uuid           6a0463ce-fc8b-11df-82df-00145ee9e395
clvdisk             1264a0af-4a7e-fb93-62d7-6a33d5f17f35
```

Run the restore by entering the command **viosbr -restore -file backup_1.tar.gz**. By adding the **-inter** option, the user is prompted before a configuration change is applied.

Note: **viosbr** cannot restore the cluster configuration if **cluster -delete** or **-rmnode** was issued.

Manually restoring user-defined virtual devices

In the example Virtual I/O Server partition, two extra disks are used in a non-rootvg volume group. If these were SAN disks or physical disks that were directly mapped to client partitions, you could simply restore the virtual device links.

However, if a logical volume or storage pool structure is on the disks, you must restore this structure first. To do this, use the volume group data files.

Save the volume group or storage pool data files as part of the backup process. These files should be in the `/home/ios/vgbackups` directory if you ran a full backup using the **savevgstruct** command. The following command lists all of the available backups:

```
$ restorevgstruct -ls
total 104
-rw-r--r--    1 root     staff         51200 Oct 21 14:22 extra_storage.data
```

The **restorevgstruct** command restores the volume group structure onto the empty disks. In Example 11-13, there are new blank disks and the same storage01 and datavg volume groups to restore.

Example 11-13 Disks and volume groups to restore

```
$ lspv
NAME            PVID                                VG               STATUS
hdisk0          00c1f170d7a97dec                    old_rootvg
hdisk1          00c1f170e170ae72                    clientvg         active
hdisk2          00c1f170e170c9cd                    clientvg         active
hdisk3          00c1f170e170dac6                    None
hdisk4          00c1f17093dc5a63                    None
hdisk5          00c1f170e170fbb2                    None
hdisk6          00c1f170de94e6ed                    rootvg           active
hdisk7          00c1f170e327afa7                    None
hdisk8          00c1f170e3716441                    None
hdisk9          none                                None
hdisk10         none                                None
hdisk11         none                                None
hdisk12         none                                None
hdisk13         none                                None
hdisk14         none                                None
hdisk15         00c1f17020d9bee9                    None

$ restorevgstruct -vg extra_storage hdisk15
hdisk15
extra_storage
testlv

Will create the Volume Group:   extra_storage
```

```
Target Disks:   Allocation Policy:
        Shrink Filesystems:     no
        Preserve Physical Partitions for each Logical Volume:   no
```

After you restore all of the logical volume structures, restore the virtual devices linking the physical backing device to the virtual. To restore these, use the **lsmap** outputs recorded from the backup steps in 11.1.4, “Backing up user-defined virtual devices” on page 345, or build documentation. As previously noted, it is important to use the slot numbers and backing devices when you restore these links.

The restoration of the Shared Ethernet Adapters requires the linking of the correct virtual Ethernet adapter to the correct physical adapter. Usually, the physical adapters are placed into a VLAN in the network infrastructure of the organization. Any network support team or switch configuration data can help with this task.

The disaster recovery restore involves a bit more manual re-creating of virtual linking devices (vtscsi and SEA) and relies on good user documentation. If there is no multipath setup on the Virtual I/O Server to preserve, another solution is a new installation of the Virtual I/O Server from the installation media. You can then restore from the build documentation.

After you run the **mkvdev** commands to re-create the mappings, the Virtual I/O Server will host virtual disks and networks that can be used to rebuild the AIX, IBM i, or Linux client partitions.

Restoring the Virtual I/O Server client operating system

After you have the Virtual I/O Server operational and all of the devices are re-created, you are ready to start restoring any AIX, IBM i, or Linux client partitions. The procedure for this is probably already be defined in your organization. It is usually identical to that for any server using dedicated disk and network resources. The method depends on the solution that is employed, and should be defined by you.

For AIX client partitions, this information is available in the IBM Systems Information Center at:

http://publib.boulder.ibm.com/infocenter/pseries/v5r3/index.jsp?topic=/com.ibm.aix.baseadmn/doc/baseadmndita/backmeth.htm

For IBM i client partitions, information about system backup and recovery is available in the IBM Systems Information Center at:

http://publib.boulder.ibm.com/infocenter/systems/scope/i5os/index.jsp?topic=/rzahg/rzahgbackup.htm

11.1.8 Rebuilding the Virtual I/O Server

This section describes what to do if there are no valid backup devices or backup images. In this case, you must install a new Virtual I/O Server.

The following description assumes that the partition definitions of the Virtual I/O Server and of all client partitions on the HMC are still available. It describes how to rebuild the example scenario's network and SCSI configurations.

It is useful to generate a System Plan on the HMC as documentation of partition profiles, settings, slot numbers, and so on.

Example 11-14 shows the command to create a System Plan for a managed system. The file name must have the extension `*.sysplan`.

Example 11-14 Creating an HMC system plan from the HMC command line

```
hscroot@hmc1:~> mksysplan -f p570.sysplan -m MT_B_p570_MMA_101F170
```

To view the System Plan, select **System Plans**. Then, select the System Plan that you want to see, and select **View System Plan**. A browser window is opened where you are prompted for the user name and password of the HMC. Figure 11-9 shows a System Plan generated from a managed system.

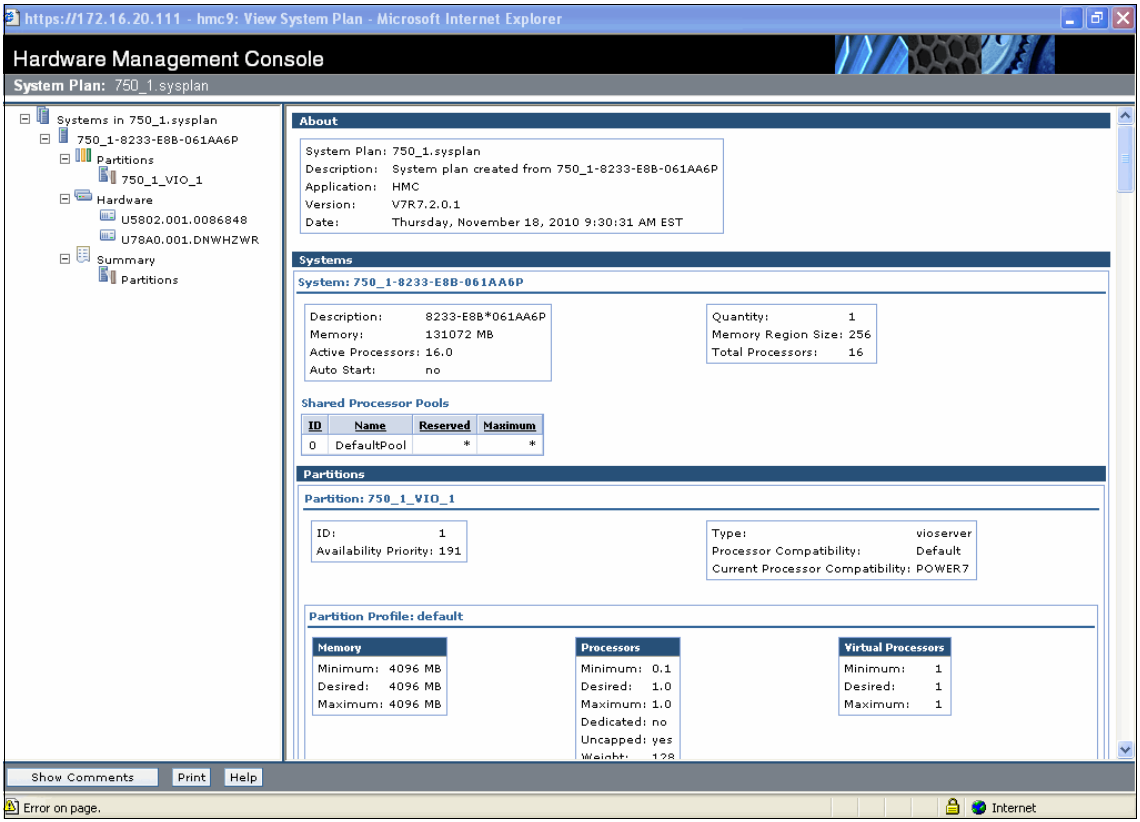


Figure 11-9 Example of a System Plan generated from a managed system

In addition to the regular backups using the **`backupios`** command, document the configuration of the following items by using the commands provided:

- Network settings:

```
netstat -state
netstat -routinfo
netstat -routtable
lsdev -dev Device -attr
cfgnamsrv -ls
hostmap -ls
optimizenet -list
entstat -all Device
```

- All physical and logical volumes, SCSI devices:

```
lspv
lsvg
lsvg -lv VolumeGroup
```

- All physical and logical adapters:

```
lsdev -type adapter
```

- The mapping between physical and logical devices and virtual devices:

```
lsmap -all
lsmap -all -net
```

- Code levels, users, and security:

```
ioslevel
viosecure -firewall view
viosecure -view -nonint
```

With this information, you can reconfigure your Virtual I/O Server manually. The following sections provide the commands you need to get the necessary information, and the commands to rebuild the configuration. The important information from the command outputs is highlighted. In your environment, the commands might differ from those shown as examples.

To start rebuilding the Virtual I/O Server, you must know which disks are used for the Virtual I/O Server itself and for any assigned volume groups for virtual I/O.

The **lspv** command lists the Virtual I/O Server that was installed on hdisk0. The first step is to install the new Virtual I/O Server from the installation media onto disk hdisk0.

```
$ lspv
hdisk0          00c0f6a0f8a49cd7                    rootvg          active
hdisk1          00c0f6a02c775268                    None
hdisk2          00c0f6a04ab4fd01                    None
hdisk3          00c0f6a04ab558cd                    None
hdisk4          00c0f6a0682ef9e0                    None
hdisk5          00c0f6a067b0a48c                    None
hdisk6          00c0f6a04ab5995b                    None
hdisk7          00c0f6a04ab66c3e                    None
hdisk8          00c0f6a04ab671fa                    None
hdisk9          00c0f6a04ab66fe6                    None
hdisk10         00c0f6a0a241e88d                    None
hdisk11         00c0f6a04ab67146                    None
hdisk12         00c0f6a04ab671fa                    None
hdisk13         00c0f6a04ab672aa                    None
hdisk14         00c0f6a077ed3ce5                    None
hdisk15         00c0f6a077ed5a83                    None
```

For more information about the installation procedure, see *PowerVM Virtualization on IBM System p: Introduction and Configuration*, SG24-7940. The further rebuild of the Virtual I/O Server is done in these steps:

1. Rebuilding the SCSI configuration.
2. Rebuilding the network configuration.

These steps are explained in greater detail in the following sections.

Rebuilding the SCSI configuration

The **lspv** command also shows us that there is an extra volume group that is on the Virtual I/O Server (datavg):

```
$ lspv
hdisk0          00c0f6a0f8a49cd7                    rootvg          active
hdisk1          00c0f6a02c775268                    None
hdisk2          00c0f6a04ab4fd01                    None
hdisk3          00c0f6a04ab558cd                    datavg          active
hdisk4          00c0f6a0682ef9e0                    None
hdisk5          00c0f6a067b0a48c                    None
hdisk6          00c0f6a04ab5995b                    None
hdisk7          00c0f6a04ab66c3e                    None
hdisk8          00c0f6a04ab671fa                    None
hdisk9          00c0f6a04ab66fe6                    None
hdisk10         00c0f6a0a241e88d                    None
hdisk11         00c0f6a04ab67146                    None
hdisk12         00c0f6a04ab671fa                    None
hdisk13         00c0f6a04ab672aa                    None
hdisk14         00c0f6a077ed3ce5                    None
hdisk15         00c0f6a077ed5a83                    None
```

The following command imports this information into the new Virtual I/O Server system's ODM:

```
importvg -vg datavg hdisk3
```

Example 11-15 shows the mapping between the logical and physical volumes, and the virtual SCSI server adapters.

Example 11-15 lsmap -all command

```
$ lsmap -all
SVSA            Physloc                                     Client Partition
ID
--------------- ------------------------------------------- ------------------
vhost0          U9117.MMA.100F6A0-V1-C15                    0x00000002

VTD                   vcd
Status                Available
```

```
LUN                  0x8100000000000000
Backing device       cd0
Physloc              U789D.001.DQDWWHY-P4-D1

SVSA            Physloc                                    Client Partition
ID
--------------- ------------------------------------------ ------------------
vhost1          U9117.MMA.100F6A0-V1-C20                   0x00000002

VTD                  vnim_rvg
Status               Available
LUN                  0x8100000000000000
Backing device       hdisk12
Physloc
U789D.001.DQDWWHY-P1-C2-T1-W200400A0B8110D0F-L12000000000000

VTD                  vnimvg
Status               Available
LUN                  0x8200000000000000
Backing device       hdisk13
Physloc
U789D.001.DQDWWHY-P1-C2-T1-W200400A0B8110D0F-L13000000000000

SVSA            Physloc                                    Client Partition
ID
--------------- ------------------------------------------ ------------------
vhost2          U9117.MMA.100F6A0-V1-C25                   0x00000003

VTD                  vdb_rvg
Status               Available
LUN                  0x8100000000000000
Backing device       hdisk8
Physloc
U789D.001.DQDWWHY-P1-C2-T1-W200400A0B8110D0F-LE000000000000

SVSA            Physloc                                    Client Partition
ID
--------------- ------------------------------------------ ------------------
vhost3          U9117.MMA.100F6A0-V1-C40                   0x00000004

VTD                  vapps_rvg
Status               Available
LUN                  0x8100000000000000
Backing device       hdisk6
Physloc
U789D.001.DQDWWHY-P1-C2-T1-W200400A0B8110D0F-LC000000000000

SVSA            Physloc                                    Client Partition
ID
```

```
-------------- ------------------------------------------ ------------------
vhost4         U9117.MMA.100F6A0-V1-C50                   0x00000005

VTD                   vlnx_rvg
Status                Available
LUN                   0x8100000000000000
Backing device        hdisk10
Physloc
U789D.001.DQDWWHY-P1-C2-T1-W200400A0B8110D0F-L10000000000000
```

Virtual SCSI server adapter `vhost0` (defined on slot 15 in HMC) has one Virtual Target Device `vcd`. It maps the optical device `cd0` to `vhost0`.

Virtual SCSI server adapter `vhost1` (defined on slot 20 in HMC) has two Virtual Target Devices, `vnim_rvg` and `vnimvg`. They are mapping the physical volumes `hdisk12` and `hdisk13` to `vhost1`.

Virtual SCSI server adapter `vhost2` (defined on slot 25 in HMC) has `vdb_rvg` as a Virtual Target Device. It is mapping the physical volume `hdisk8` to `vhost2`.

Virtual SCSI server adapter `vhost3` (defined on slot 40 in HMC) has `vapps_rvg` as a Virtual Target Device. It is mapping the physical volume `hdisk6` to `vhost3`.

Virtual SCSI server adapter `vhost4` (defined on slot 50 in HMC) has `vlnx_rvg` as a Virtual Target Device. It is mapping the physical volume `hdisk10` to `vhost4`.

These commands are used to create the Virtual Target Devices that are needed:

```
mkvdev -vdev cd0 -vadapter vhost0 -dev vcd
mkvdev -vdev hdisk12 -vadapter vhost1 -dev vnim_rvg
mkvdev -vdev hdisk13 -vadapter vhost1 -dev vnimvg
mkvdev -vdev hdisk8 -vadapter vhost2 -dev vdb_rvg
mkvdev -vdev hdisk6 -vadapter vhost3 -dev vnim_rvg
mkvdev -vdev hdisk10 -vadapter vhost4 -dev vlnx_rvg
```

Tip: The names of the Virtual Target Devices are generated automatically, except when you define a name by using the **-dev** flag of the **mkvdev** command.

Rebuilding the network configuration

After successfully rebuilding the SCSI configuration, rebuild the network configuration.

The **netstat -state** command shows that en4 is the only active network adapter:

```
$ netstat -state
Name  Mtu   Network     Address           ZoneID    Ipkts Ierrs    Opkts Oerrs  Coll
en4   1500  link#2      6a.88.8d.e7.80.d          4557344     0  1862620     0     0
en4   1500  9.3.4       vios1                     4557344     0  1862620     0     0
lo0   16896 link#1                                   4521     0     4634     0     0
lo0   16896 127         loopback                    4521     0     4634     0     0
lo0   16896 ::1                                  0     4521     0     4634     0     0
```

With the **lsmap -all -net** command, determine that ent5 is defined as a Shared Ethernet Adapter that maps physical adapter ent0 to virtual adapter ent2:

```
$ lsmap -all -net
SVEA   Physloc
------ --------------------------------------------
ent2   U9117.MMA.101F170-V1-C11-T1

SEA                  ent5
Backing device       ent0
Status               Available
Physloc              U789D.001.DQDYKYW-P1-C4-T1

SVEA   Physloc
------ --------------------------------------------
ent4   U9117.MMA.101F170-V1-C13-T1

SEA                 NO SHARED ETHERNET ADAPTER FOUND
```

The information for the default gateway address is provided by the **netstat -routinfo** command:

```
$ netstat -routinfo
Routing tables
Destination        Gateway           Flags     Wt  Policy  If   Cost
Config_Cost

Route Tree for Protocol Family 2 (Internet):
default            9.3.4.1           UG        1    -   en4       0    0
9.3.4.0            vios1             UHSb      1    -   en4       0    0 =>
9.3.4/23           vios1             U         1    -   en4       0    0
vios1              loopback          UGHS      1    -   lo0       0    0
9.3.5.255          vios1             UHSb      1    -   en4       0    0
127/8              loopback          U         1    -   lo0       0    0

Route Tree for Protocol Family 24 (Internet v6):
::1                ::1               UH        1    -   lo0       0    0
```

To list the subnet mask, use the **lsdev -dev en4 -attr** command:

```
$ lsdev -dev en4 -attr
attribute     value         description
user_settable

alias4                      IPv4 Alias including Subnet Mask           True
alias6                      IPv6 Alias including Prefix Length         True
arp           on            Address Resolution Protocol (ARP)          True
authority                   Authorized Users                           True
broadcast                   Broadcast Address                          True
mtu           1500          Maximum IP Packet Size for This Device     True
netaddr       9.3.5.111     Internet Address                           True
netaddr6                    IPv6 Internet Address                      True
netmask       255.255.254.0 Subnet Mask                                True
prefixlen                   Prefix Length for IPv6 Internet Address    True
remmtu        576           Maximum IP Packet Size for REMOTE Networks True
rfc1323                     Enable/Disable TCP RFC 1323 Window Scaling True
security      none          Security Level                             True
state         up            Current Interface Status                   True
tcp_mssdflt                 Set TCP Maximum Segment Size               True
tcp_nodelay                 Enable/Disable TCP_NODELAY Option          True
tcp_recvspace               Set Socket Buffer Space for Receiving      True
tcp_sendspace               Set Socket Buffer Space for Sending        True
```

The last information we need is the default virtual adapter and the default PVID for the Shared Ethernet Adapter. This is shown by the **lsdev -dev ent5 -attr** command:

```
$ lsdev -dev ent5 -attr
attribute     value    description
user_settable

accounting    disabled Enable per-client accounting of network statistics               True
ctl_chan      ent3     Control Channel adapter for SEA failover                         True
gvrp          no       Enable GARP VLAN Registration Protocol (GVRP)                    True
ha_mode       auto     High Availability Mode                                           True
jumbo_frames  no       Enable Gigabit Ethernet Jumbo Frames                             True
large_receive no       Enable receive TCP segment aggregation                           True
largesend     0        Enable Hardware Transmit TCP Resegmentation                      True
netaddr       0        Address to ping                                                  True
pvid          1        PVID to use for the SEA device                                   True
pvid_adapter  ent2     Default virtual adapter to use for non-VLAN-tagged packets       True
qos_mode      disabled N/A                                                              True
real_adapter  ent0     Physical adapter associated with the SEA                         True
thread        1        Thread mode enabled (1) or disabled (0)                          True
virt_adapters ent2     List of virtual adapters associated with the SEA (comma separated) True
```

Consideration: In this example, the IP of the Virtual I/O Server is not configured on the Shared Ethernet Adapter (`ent5`) but on another adapter (`ent4`). This configuration avoids network disruption between the Virtual I/O Server and any other partition on the same system when the physical card (`ent0`) used as the SEA is replaced.

The following commands re-created the network configuration:

```
$ mkvdev -sea ent0 -vadapter ent2 -default ent2 -defaultid 1
$ mktcpip -hostname vios1 -inetaddr 9.3.5.111 -interface en5 -start -netmask
255.255.254.0 -gateway 9.3.4.1
```

These steps complete the basic rebuilding of the Virtual I/O Server.

11.1.9 Updating the Virtual I/O Server

Three scenarios for updating a Virtual I/O Server are described in this section. A dual Virtual I/O Server environment (useful when you perform Virtual I/O Server software upgrades and service) provides a continuous connection of your client partitions to their virtual I/O resources. For client partitions using non-critical virtual resources, or when you have service windows that allow a Virtual I/O Server to be rebooted, you can use a single Virtual I/O Server scenario. The last scenario covers update of cluster configuration.

For the dual Virtual I/O Server scenario, if you are using SAN LUNs and MPIO or IBM i mirroring on the client partitions, the maintenance on the Virtual I/O Server will not cause extra work after the update on the client side.

Updating a single Virtual I/O Server environment

When you apply routine service that requires a reboot in a single Virtual I/O Server environment, plan downtime and shut down every client partition that uses virtual storage provided by this Virtual I/O Server.

Tip: Before you start an upgrade, create a backup of the Virtual I/O Server and the virtual I/O client partitions if a current backup is not available. To back up the Virtual I/O Server, use the **`backupios`** command. Also, document the virtual Ethernet and SCSI devices before the update.

To avoid complications during an upgrade or update, check the environment before you upgrade or update the Virtual I/O Server. The following is a list of useful commands for the virtual I/O client and Virtual I/O Server:

`lsvg rootvg`	On the Virtual I/O Server and AIX virtual I/O client, check for stale PPs and PVs.
`cat /proc/mdstat`	On the Linux client using mirroring, check for faulty disks.
`multipath -ll`	On the Linux client using MPIO, check the paths.
`lsvg -pv rootvg`	On the Virtual I/O Server, check for missing disks.
`netstat -cdlistats`	On the Virtual I/O Server, check that the Link status is Up on all used interfaces.
`errpt`	On the AIX virtual I/O client, check for processor, memory, disk, or Ethernet errors, and resolve them before you continue.
`dmesg, messages`	On the Linux virtual I/O client, check for processor, memory, disk, or Ethernet errors, and resolve them before you continue.
`netstat -v`	On the virtual I/O client, check that the Link status is Up on all used interfaces.

Running update on a single Virtual I/O Server

There are several options for downloading and installing a Virtual I/O Server update: download iso-images, packages, or installation from CD.

Tip: You can get the latest available updates for the Virtual I/O Server and check also the recent installation instructions, from Fix Central at:

`http://www-933.ibm.com/support/fixcentral/`

To update the Virtual I/O Server, complete these steps:

1. Perform a backup of the Virtual I/O Server.
2. Shut down the virtual I/O client partitions that are connected to the Virtual I/O Server, or disable any virtual resource that is in use.
3. If you use a Virtual Media Repository, unload all media images by using the **`lsvopt`** and **`unloadvopt`** commands.
4. If previous updates have been applied to the Virtual I/O Server, commit them with this command:

```
# updateios -commit
```

This command does not provide any progress information, but you can run:

```
$ tail -f install.log
```

In another terminal window, follow the progress. If the command hangs, interrupt it by pressing Ctrl + C and run it again until you see the following output:

```
$ updateios -commit
There are no uncommitted updates.
```

5. Apply the update with the **updateios** command. Use `/dev/cd0` for CD or any directory that contains the files. You can also mount an NFS directory with the mount command:

   ```
   $ mount <name_of_remote_server>:/software/AIX/VIO-Server /mnt
   $ updateios -dev /mnt -install -accept
   ```

6. Reboot the Virtual I/O Server when the update has finished:

   ```
   $ shutdown -restart
   ```

7. Verify the new level with the **ioslevel** command.
8. Check the configuration of all disks and Ethernet adapters on the Virtual I/O Server.
9. Start the client partitions.

Verify the Virtual I/O Server environment, document the update, and create a backup of your updated Virtual I/O Server.

Updating a dual Virtual I/O Server environment

When you apply an update to the Virtual I/O Server in a properly configured dual Virtual I/O Server environment, you can do so without downtime to the virtual I/O services and without disruption in continuous availability.

Tip: Back up the Virtual I/O Servers and the virtual I/O client partitions if a current backup is not available. Also, document the virtual Ethernet and SCSI device before the update. This reduces the time that is needed for a recovery scenario.

Checking network health

Check the virtual Ethernet and disk devices on the Virtual I/O Server and virtual I/O client before you start the update on either of the Virtual I/O Servers. Check the physical adapters to verify connections. As shown in Example 11-16, Figure 11-10 on page 383, Example 11-17 on page 383, and Example 11-18 on page 383, all the virtual adapters are up and running.

Example 11-16 The netstat -v command on the virtual I/O client

```
netstat -v
.
. (Lines omitted for clarity)
.
Virtual I/O Ethernet Adapter (l-lan) Specific Statistics:
---------------------------------------------------------
RQ Length: 4481
No Copy Buffers: 0
Filter MCast Mode: False
Filters: 255
  Enabled: 1  Queued: 0  Overflow: 0
LAN State: Operational
Hypervisor Send Failures: 0
  Receiver Failures: 0
  Send Errors: 0
Hypervisor Receive Failures: 0

ILLAN Attributes: 0000000000003002 [0000000000002000]

. (Lines omitted for clarity)
```

Figure 11-10 shows the Work with TCP/IP Interface Status panel.

```
 Work with TCP/IP Interface Status
                                                                System:E101F170
 Type options, press Enter.
   5=Display details   8=Display associated routes   9=Start   10=End
   12=Work with configuration status   14=Display multicast groups

      Internet         Network            Line      Interface
 Opt  Address          Address          Description  Status
      9.3.5.119        9.3.4.0          ETH01        Active
      127.0.0.1        127.0.0.0        *LOOPBACK    Active
```

Figure 11-10 IBM i Work with TCP/IP Interface Status panel

Example 11-17 shows the primary Virtual I/O Server being checked.

Example 11-17 The netstat -cdlistats command on the primary Virtual I/O Server

```
$ netstat -cdlistats
.
. (Lines omitted for clarity)
.
Virtual I/O Ethernet Adapter (l-lan) Specific Statistics:
---------------------------------------------------------
RQ Length: 4481
No Copy Buffers: 0
Trunk Adapter: True
  Priority: 1  Active: True
Filter MCast Mode: False
Filters: 255
  Enabled: 1  Queued: 0  Overflow: 0
LAN State: Operational
.
. (Lines omitted for clarity)
```

Example 11-18 shows the secondary Virtual I/O Server being checked.

Example 11-18 The netstat -cdlistats command on the secondary Virtual I/O Server

```
$ netstat -cdlistats
.
. (Lines omitted for clarity)
.
Virtual I/O Ethernet Adapter (l-lan) Specific Statistics:
---------------------------------------------------------
RQ Length: 4481
```

```
No Copy Buffers: 0
Trunk Adapter: True
  Priority: 2  Active: False
Filter MCast Mode: False
Filters: 255
  Enabled: 1  Queued: 0  Overflow: 0
LAN State: Operational
.
. (Lines omitted for clarity)
```

Checking storage health

Checking the disk status depends on how the disks are shared from the Virtual I/O Server.

Checking the storage health in the MPIO environment

If you have an MPIO setup on your Virtual I/O Server client partitions similar to Figure 11-11 on page 385, run these commands before and after the first Virtual I/O Server update to verify the disk path status:

`lspath`	On the AIX virtual I/O client, check all the paths to the disks. They should all be in the enabled state.
`multipath -ll`	Check the paths on the Linux client.
`lsattr -El hdisk0`	On the virtual I/O client, check the MPIO heartbeat for `hdisk0`. The attribute `hcheck_mode` should be set to `nonactive`, and that `hcheck_interval` should be `60`. If you run IBM SAN storage, check that reserve_policy is `no_reserve`. Other storage vendors might require other values for `reserve_policy`. Issue this command in all disks on the Virtual I/O Server.

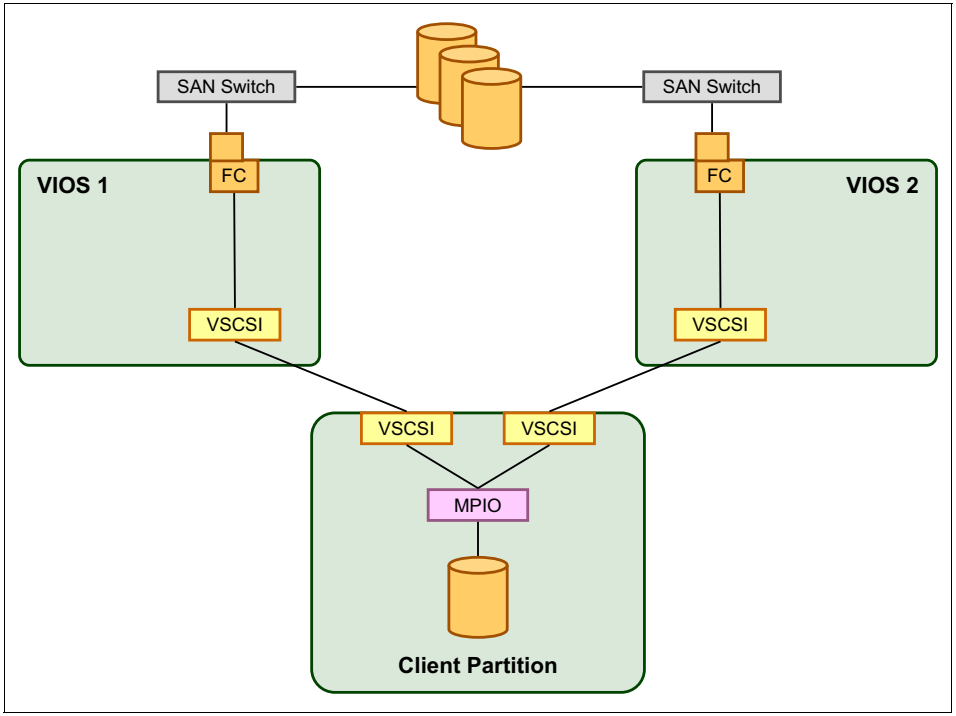


Figure 11-11 Virtual I/O client that is running MPIO

Checking storage health in the mirroring environment

Figure 11-12 shows the concept of a mirrored infrastructure.

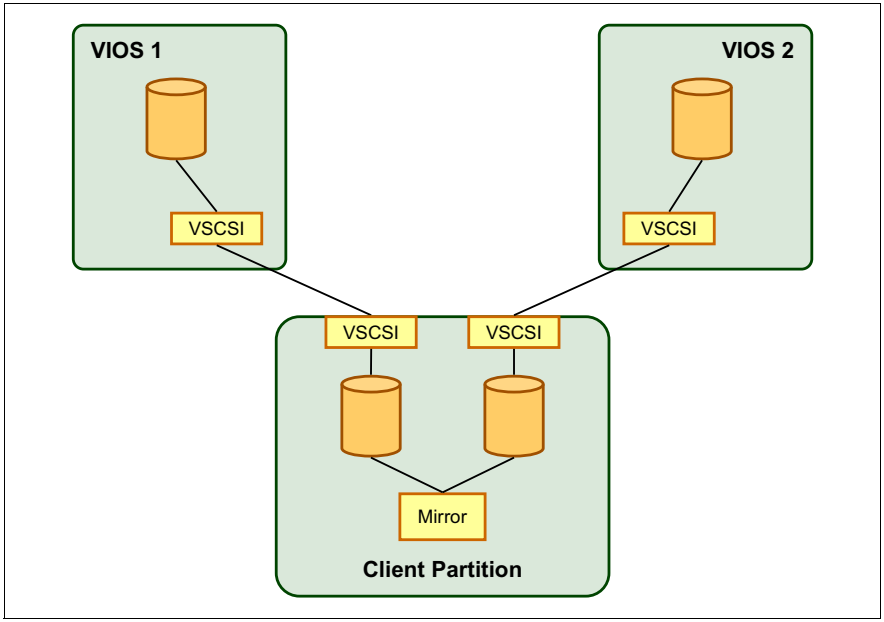


Figure 11-12 Virtual I/O client partition software mirroring

If you use mirroring on your Virtual I/O client partitions, verify a stable mirroring status for the disks that are shared from the Virtual I/O Server with the following procedures.

On the AIX virtual I/O client, run these commands:

`lsvg rootvg` Verify there are no stale PPs, and that the quorum is off.

`lsvg -p rootvg` Verify there is no missing hdisk.

Note: The `fixdualvios.ksh` script that is provided in Appendix A, “AIX disk and NIB network checking and recovery script” on page 699 is a useful tool for performing a health check.

On the IBM i virtual I/O client, complete these steps:

1. Run **STRSST** and log in to System Service Tools.
2. Select `3. Work with disk units` → `1. Display disk configuration` → `1. Display disk configuration status` and verify all virtual disk units (type 6B22) are in mirrored `Active` state as shown in Figure 11-13.

```
Display Disk Configuration Status

          Serial                      Resource
ASP Unit  Number         Type Model Name      Status
  1                                           Mirrored
      1  Y3WUTVVQMM4G    6B22  050  DD001     Active
      1  YYUUH3U9UELD    6B22  050  DD004     Active
      2  YD598QUY5XR8    6B22  050  DD003     Active
      2  YTM3C79KY4XF    6B22  050  DD002     Active
```

Figure 11-13 IBM i Display Disk Configuration Status panel

On the Linux virtual I/O client, run these commands:

cat /proc/mdstat Check the mirror status. A stable environment is shown in Example 11-19.

Example 11-19 The mdstat command showing a stable environment

```
cat /proc/mdstat
Personalities : [raid1]
md1 : active raid1 sdb3[1] sda3[0]
      1953728 blocks [2/2] [UU]
md2 : active raid1 sdb4[1] sda4[0]
      21794752 blocks [2/2] [UU]
md0 : active raid1 sdb2[1] sda2[0]
      98240 blocks [2/2] [UU]
```

After you check the environment and resolve any issues, back up the Virtual I/O Server and virtual I/O client if a current backup is not available.

Step-by-step update

To update a dual Virtual I/O Server environment, complete the following steps:

1. Find the standby Virtual I/O Server and run the **netstat** command. Locate the priority of the SEA and whether it is active. In this case, the standby adapter is not active, so you can begin upgrading this server.

```
$ netstat -cdlistats
.
```

```
. (Lines omitted for clarity)
.
Trunk Adapter: True
  Priority: 2  Active: False
Filter MCast Mode: False
Filters: 255
  Enabled: 1  Queued: 0  Overflow: 0
LAN State: Operational
.
. (Lines omitted for clarity)
```

If you must change the active adapter, use following command to put it in backup mode manually:

```
$ chdev -attr entXX ha_mode=standby
```

2. All Interim Fixes must be applied before the upgrade is removed. To remove the Interim Fixes, complete the following steps:

   a. Become root on Virtual I/O Server:

      ```
      $ oem_setup_env
      ```

   b. List all Interim Fixes installed:

      ```
      # emgr -P
      ```

   c. Remove each Interim Fix by label:

      ```
      # emgr -r -L <label name>
      ```

   d. Exit the root shell:

      ```
      # exit
      ```

3. Apply the update from DVD or a remote directory with the **updateios** command and press y to start the update.

   ```
   $ updateios -dev /mnt -install -accept
   . (Lines omitted for clarity)

   Continue the installation [y|n]?
   ```

4. Reboot the standby Virtual I/O Server when the update completes:

   ```
   $ shutdown -force -restart

   SHUTDOWN PROGRAM
   Mon Oct 13 21:57:23 CDT 2008

   Wait for 'Rebooting...' before stopping.
   Error reporting has stopped.
   ```

5. After the reboot, verify the software level:

```
$ ioslevel
1.5.2.1-FP-11.1
```

6. For an AIX MPIO environment, as shown in Figure 11-11 on page 385, run the **lspath** command on the virtual I/O client. Verify that all paths are enabled.

   For an AIX LVM mirroring environment, as shown in Figure 11-12 on page 386, run the **varyonvg** command as shown in Example 11-20. The volume group should begin to sync. If it does not, run the **syncvg -v *<VGname>*** command on the volume groups that used the virtual disk from the Virtual I/O Server environment to synchronize each volume group. ***<VGname>*** is the name of the Volume Group.

   For the IBM i client mirroring environment, you can proceed to the next step. No manual action is required on IBM i client side because IBM i automatically resumes the suspended mirrored disk units as soon as the updated Virtual I/O Server resumes operations.

   Consideration: IBM i tracks changes for a suspended mirrored disk unit for a limited time, allowing it to resynchronize changed pages only. During testing, IBM i did not require a full mirror resynchronize when rebooting the Virtual I/O Server. However, this might not be the case for any reboot that takes an extended amount of time.

Example 11-20 AIX LVM mirror resynchronization

```
# lsvg -p rootvg
rootvg:
PV_NAME           PV STATE          TOTAL PPs   FREE PPs    FREE
DISTRIBUTION
hdisk0            active            511         488
102..94..88..102..102
hdisk1            missing           511         488
102..94..88..102..102

# varyonvg rootvg

# lsvg -p rootvg
rootvg:
PV_NAME           PV STATE          TOTAL PPs   FREE PPs    FREE
DISTRIBUTION
hdisk0            active            511         488
102..94..88..102..102
hdisk1            active            511         488
102..94..88..102..102

# lsvg rootvg
```

```
VOLUME GROUP:       rootvg                  VG IDENTIFIER:
00c478de00004c00000
00006b8b6c15e
VG STATE:           active                PP SIZE:        64 megabyte(s)
VG PERMISSION:      read/write              TOTAL PPs:      1022 (65408
megabytes)
MAX LVs:            256                     FREE PPs:       976 (62464
megabytes)
LVs:                9                       USED PPs:       46 (2944
megabytes)
OPEN LVs:           8                     QUORUM:         1
TOTAL PVs:          2                     VG DESCRIPTORS: 3
STALE PVs:          0                     STALE PPs:      0
ACTIVE PVs:         2                     AUTO ON:        yes
MAX PPs per VG:     32512
MAX PPs per PV:     1016                  MAX PVs:        32
LTG size (Dynamic): 256 kilobyte(s)        AUTO SYNC:      no
HOT SPARE:          no                      BB POLICY:      relocatable
```

For a Linux client mirroring environment, complete these steps for every md-device (`md0`, `md1`, `md2`):

a. Set the disk `faulty` (repeat the steps for all mdx devices):

```
# mdadm --manage --set-faulty /dev/md2 /dev/sda4
```

b. Remove the device:

```
# mdadm --manage --remove /dev/md2 /dev/sda2
```

c. Rescan the device (select the corresponding path):

```
# echo 1 > /sys/class/scsi_device/0\:0\:1\:0/device/rescan
```

d. Hot-add the device to `mdadm`:

```
# mdadm --manage --add /dev/md2 /dev/sda4
```

e. Check the sync status and wait for it to complete:

```
# cat /proc/mdstat
   Personalities : [raid1]
   md1 : active raid1 sda3[0] sdb3[1]
         1953728 blocks [2/2] [UU]

   md2 : active raid1 sda4[2] sdb4[1]
         21794752 blocks [2/1] [_U]
         [=>..................]  recovery =  5.8%
   (1285600/21794752) finish=8.2min speed=41470K/sec
   md0 : active raid1 sda2[0] sdb2[1]
         98240 blocks [2/2] [UU]
```

7. If you use Shared Ethernet Adapter Failover, shift the standby and primary connections to the Virtual I/O Server with the **chdev** command. Check with the **netstat -cdlistats** command whether the state has changed, as shown in this example:

```
$ chdev -dev ent4 -attr ha_mode=standby
ent4 changed
$ netstat -cdlistats
.
. (Lines omitted for clarity)
.
Trunk Adapter: True
  Priority: 1  Active: False
Filter MCast Mode: False
.
. (Lines omitted for clarity)
```

   After you verify the network and storage health on all Virtual I/O Server active client partitions, update the other Virtual I/O Server as well.

8. Remove all interim fixes on the second Virtual I/O Server to be updated.
9. Apply the update to the second Virtual I/O Server, which is now the standby Virtual I/O Server, by using the **updateios** command.
10. Reboot the second Virtual I/O Server with the **shutdown -restart** command.
11. Check the new level with the **ioslevel** command.
12. For an AIX MPIO environment as shown in Figure 11-11 on page 385, run the **lspath** command on the virtual I/O client and verify that all paths are enabled.

    For an AIX LVM environment, as shown in Figure 11-12 on page 386, run the **varyonvg** command, and the volume group should begin to synchronize. If it does not, use the **syncvg -v <VGname>** command on the volume groups that used the virtual disk from the Virtual I/O Server environment to synchronize each volume group. **<VGname>** is the name of the Volume Group.

    For the IBM i client mirroring environment, proceed to the next step. No manual action is required on IBM i client side because IBM i automatically resumes the suspended mirrored disk units as soon as the updated Virtual I/O Server resumes operations.

    For the Linux mirroring environment, manually resynchronize the mirror again (see step 6).

13. If you use Shared Ethernet Adapter Failover, reset the Virtual I/O Server role to primary with the **chdev** command:

```
$ chdev -dev ent4 -attr ha_mode=auto
ent4 changed
```

14. After verifying the network and storage health again, create another backup, this time from both updated Virtual I/O Servers, before considering the update process complete.

Rolling updates in cluster

The Virtual I/O Server version 2.2.2.0 supports rolling updates for clusters.

You can use the rolling updates enhancement to apply software updates to the Virtual I/O Server partitions in the cluster individually without causing an outage in the entire cluster.

For more information about rolling updates, see 10.1.9, “Rolling updates in a cluster” on page 306.

11.1.10 Updating Virtual I/O Server adapter firmware

This section describes how to update the firmware for I/O adapters that are owned by the Virtual I/O Server.

Remember: When an IBM i partition is assigned an adapter device, the IBM i Licensed Internal Code (SLIC) maintains the firmware. When an adapter device is assigned to the Virtual I/O Server, the process is manual, even though IBM i is using the device. This can cause mismatches between IBM i and the firmware of adapters that are managed by the Virtual I/O Server.

Complete the following steps to check for and update to the latest available I/O adapter firmware:

1. Log in to the Virtual I/O Server and run the **`lsdev -type adapter`** command to list the installed adapters as shown in Example 11-21.

Example 11-21 lsdev -type adapter command

```
$ lsdev -type adapter
name             status      description
ent0             Available   2-Port 10/100/1000 Base-TX PCI-X Adapter (14108902)
ent1             Available   2-Port 10/100/1000 Base-TX PCI-X Adapter (14108902)
ent2             Available   Virtual I/O Ethernet Adapter (l-lan)
ent3             Available   Shared Ethernet Adapter
fcs0             Available   8Gb PCI Express Dual Port FC Adapter (df1000f114108a03)
fcs1             Available   8Gb PCI Express Dual Port FC Adapter (df1000f114108a03)
pager0           Available   Pager Kernel Extension
sissas0          Available   PCI-X266 Planar 3Gb SAS Adapter
vasi0            Available   Virtual Asynchronous Services Interface (VASI)
vbsd0            Available   Virtual Block Storage Device (VBSD)
vfchost0         Available   Virtual FC Server Adapter
```

```
vhost0          Available   Virtual SCSI Server Adapter
vhost1          Available   Virtual SCSI Server Adapter
vhost2          Available   Virtual SCSI Server Adapter
vhost3          Available   Virtual SCSI Server Adapter
vhost4          Available   Virtual SCSI Server Adapter
vsa0            Available   LPAR Virtual Serial Adapter
```

2. Run the **lsmcode -d *adapter_name*** command to list the currently installed adapter firmware as shown for the Fibre Channel adapter fcs0 in Example 11-22.

Example 11-22 lsmcode -d fcs0 command

```
$ oem_setup_env
# lsmcode -d fcs0

DISPLAY MICROCODE LEVEL
802111
fcs0    8Gb PCI Express Dual Port FC Adapter (df1000f114108a03)

The current microcode level for fcs0 is 110105.

Use Enter to continue.
```

3. Note the current microcode level for the adapter, and go to the IBM Fix Central support website to check whether there is a newer version available:

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

4. Select **Power** for Product Group, **Firmware and HMC** for Product, enter your corresponding Machine type-model, and click **Continue** as shown in Figure 11-14.

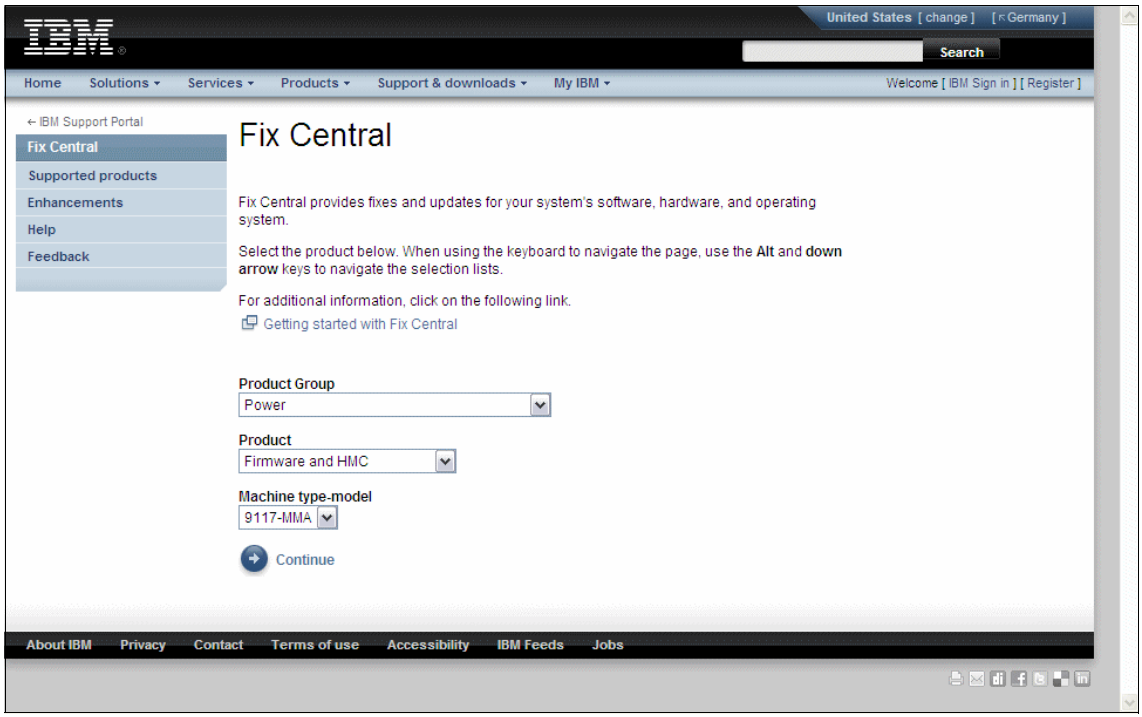


Figure 11-14 IBM Fix Central website

5. Select **Device Firmware** and click **Continue** as shown in Figure 11-15.

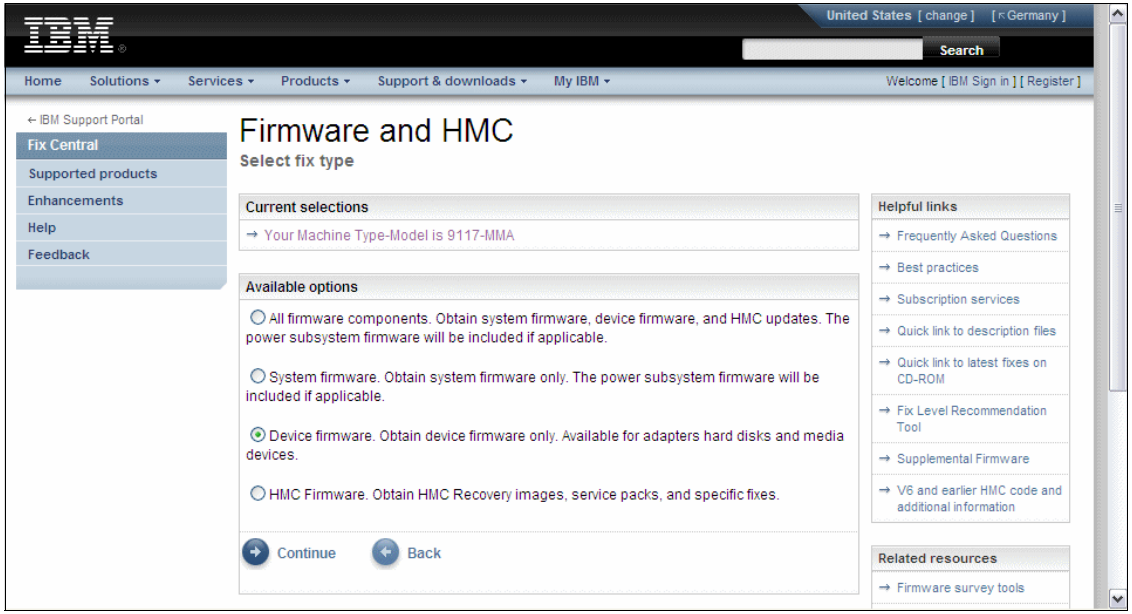


Figure 11-15 IBM Fix Central website: Firmware and HMC

6. Select **Select by feature code**, or if you do not know the adapter feature code, select **Select by device type** and click **Continue** as shown in Figure 11-16.

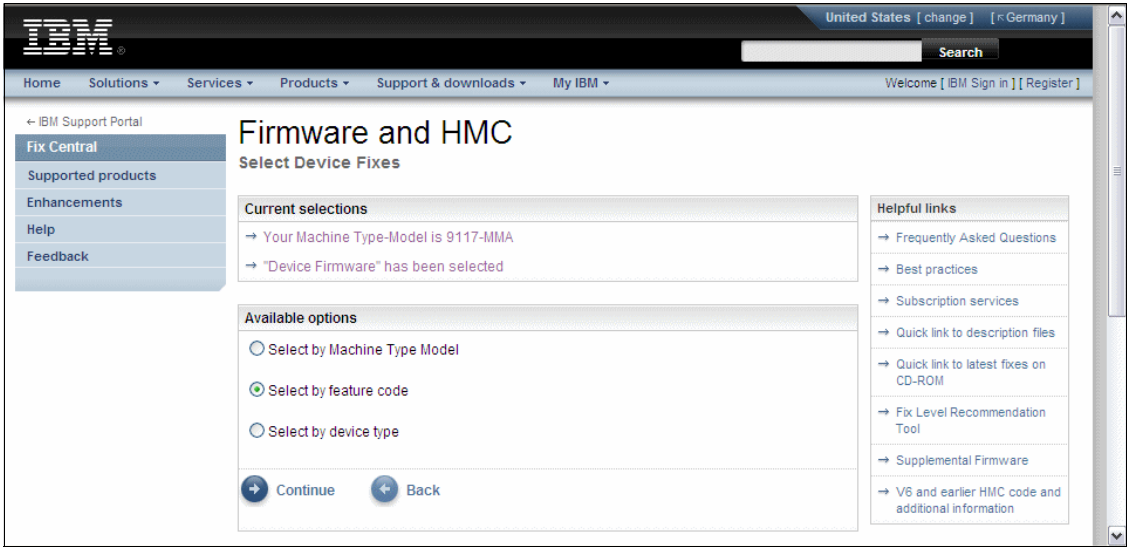


Figure 11-16 IBM Fix Central website: Select by feature code

7. Enter the adapter's feature code and click **Continue** as shown in Figure 11-17.

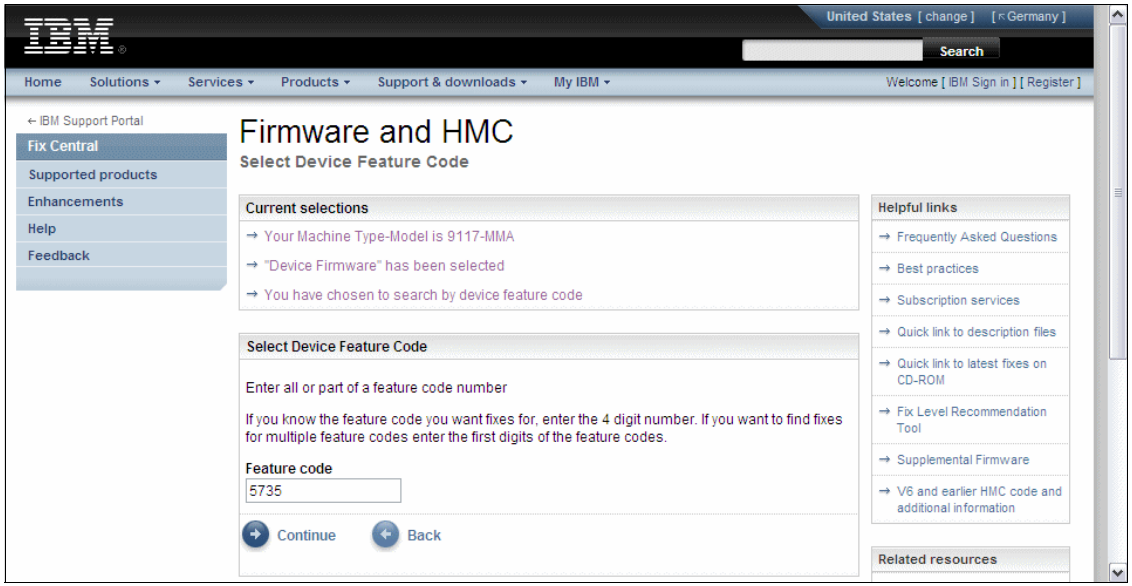


Figure 11-17 IBM Fix Central website: Select device feature code

8. Check the displayed version of the available latest RPM firmware package for the adapter. If it is newer than the currently installed one checked in step 2, click **Description** to display the Firmware Description File including installation instructions. Then click **Continue** as shown in Figure 11-18 to proceed with accepting the license agreement and starting its download.

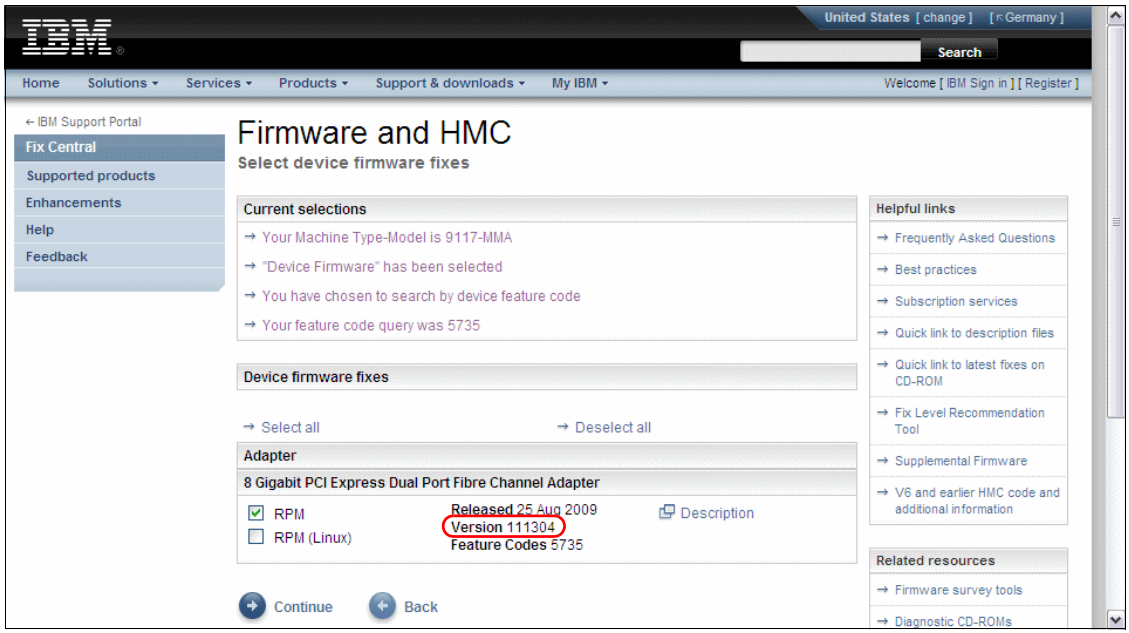


Figure 11-18 IBM Fix Central website: Select device firmware fixes

9. Transfer the downloaded adapter firmware package `*.aix.rpm` file to the Virtual I/O Server by using binary FTP as shown in Example 11-23.

Example 11-23 FTP transfer of adapter firmware to the Virtual I/O Server

```
D:\tmp>ftp 172.16.20.190
Connected to 172.16.20.190.
220 P6_1_vios1 FTP server (Version 4.2 Tue Sep 14 20:17:37 CDT 2010) ready.
User (172.16.20.190:(none)): padmin
331 Password required for padmin.
Password:
230-Last unsuccessful login: Sat Dec  4 10:06:27 EST 2010 on ssh from
172.16.254.34
230-Last login: Sat Dec  4 13:59:08 EST 2010 on ftp from
::ffff:172.16.254.34
230 User padmin logged in.
ftp> bin
200 Type set to I.
ftp> put df1000f114108a03-111304.aix.rpm
```

```
200 PORT command successful.
150 Opening data connection for df1000f114108a03-111304.aix.rpm.
226 Transfer complete.
ftp: 655470 bytes sent in 15,33Seconds 42,76Kbytes/sec.
ftp> bye
221 Goodbye.
```

10. On the Virtual I/O Server, unpack the adapter firmware package `*.aix.rpm` file by using the **rpm -ihv --ignoreos *package_name*** command as shown in Example 11-24. Doing so adds the adapter firmware file to the `/etc/microcode` directory.

Example 11-24 Unpacking the adapter firmware package on the Virtual I/O Server

```
$ oem_setup_env
# rpm -ihv --ignoreos df1000f114108a03-111304.aix.rpm
pci.df1000f114108a03
################################################
# ls -l /etc/microcode/df*
-rwxr-xr-x    1 root     system       920768 Jul 10 2009
/etc/microcode/df1000f114108a03.111304
```

11. Start the diagnostic service aids for updating the adapter's firmware by running the **diag** command, then press Enter to continue on the Diagnostic Operating Instructions panel as shown in Example 11-25.

Example 11-25 diag command

```
# diag

DIAGNOSTIC OPERATING INSTRUCTIONS  VERSION 6.1.6.2
801001


LICENSED MATERIAL and LICENSED INTERNAL CODE - PROPERTY OF IBM
(C) COPYRIGHTS BY IBM AND BY OTHERS 1982, 2010.
ALL RIGHTS RESERVED.

These programs contain diagnostics, service aids, and tasks for
the system.  These procedures should be used whenever problems
with the system occur which have not been corrected by any
software application procedures available.

In general, the procedures will run automatically.  However,
sometimes you will be required to select options, inform the
system when to continue, and do simple tasks.

Several keys are used to control the procedures:
- The Enter key continues the procedure or performs an action.
```

```
- The Backspace key allows keying errors to be corrected.
- The cursor keys are used to select an option.

Press the F3 key to exit or press Enter to continue.
```

12. Select **Task Selection** and press Enter as shown in Figure 11-19.

```
FUNCTION SELECTION
801002


Move cursor to selection, then press Enter.

  Diagnostic Routines
    This selection will test the machine hardware. Wrap plugs and
    other advanced functions will not be used.
  Advanced Diagnostics Routines
    This selection will test the machine hardware. Wrap plugs and
    other advanced functions will be used.
  Task Selection (Diagnostics, Advanced Diagnostics, Service Aids, etc.)
    This selection will list the tasks supported by these procedures.
    Once a task is selected, a resource menu may be presented showing
    all resources supported by the task.
  Resource Selection
    This selection will list the resources in the system that are supported
    by these procedures. Once a resource is selected, a task menu will
    be presented showing all tasks that can be run on the resource(s).




F1=Help                 F10=Exit                 F3=Previous Menu
```

Figure 11-19 Diagnostics aids: Task Selection

13. Select **Microcode Tasks** and press Enter as shown in Figure 11-20.

```
TASKS SELECTION LIST
801004


From the list below, select a task by moving the cursor to
the task and pressing 'Enter'.
To list the resources for the task highlighted, press 'List'.

[MORE...20]
  Display Service Hints
  Display Software Product Data
  Display or Change Bootlist
  Format Media
  Gather System Information
  Hot Plug Task
  Identify and Attention Indicators
  Local Area Network Analyzer
  Log Repair Action
  Microcode Tasks
  Periodic Diagnostics
  RAID Array Manager
[MORE...1]

F1=Help                F4=List                F10=Exit
Enter
F3=Previous Menu
```

Figure 11-20 Diagnostics aids: Microcode Tasks

14. Move the cursor to **Download Microcode** and press Enter as shown in Figure 11-21.

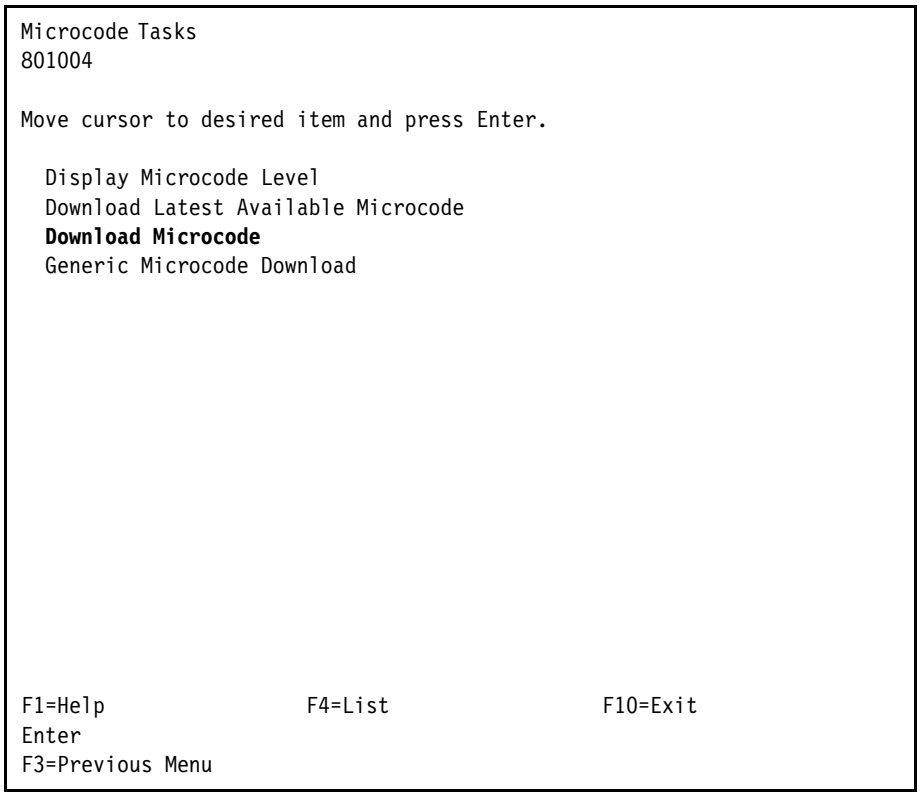


Figure 11-21 Diagnostics aids: Download Microcode

15. Move the cursor to each I/O adapter port to be updated and press Enter, then press **F7=Commit** to start the adapter firmware download as shown in Figure 11-22.

```
RESOURCE SELECTION LIST
801006

From the list below, select any number of resources by moving
the cursor to the resource and pressing 'Enter'.
To cancel the selection, press 'Enter' again.
To list the supported tasks for the resource highlighted, press 'List'.

Once all selections have been made, press 'Commit'.
To avoid selecting a resource, press 'Previous Menu'.


[TOP]
  All Resources
      This selection will select all the resources currently displayed.
                U789D.001.DQDYKYW-
+ fcs0             P1-C1-T1            8Gb PCI Express Dual Port FC Adapter
                                       (df1000f114108a03)
+ fcs1             P1-C1-T2            8Gb PCI Express Dual Port FC Adapter
                                       (df1000f114108a03)
  ent0             P1-C4-T1            2-Port 10/100/1000 Base-TX PCI-X
Adapter
[MORE...13]

F1=Help                 F4=List                  F7=Commit
F10=Exit
F3=Previous Menu
```

Figure 11-22 Diagnostics aids: Resource selection list

16. Read through the notice and press Enter to continue as shown in Figure 11-23.

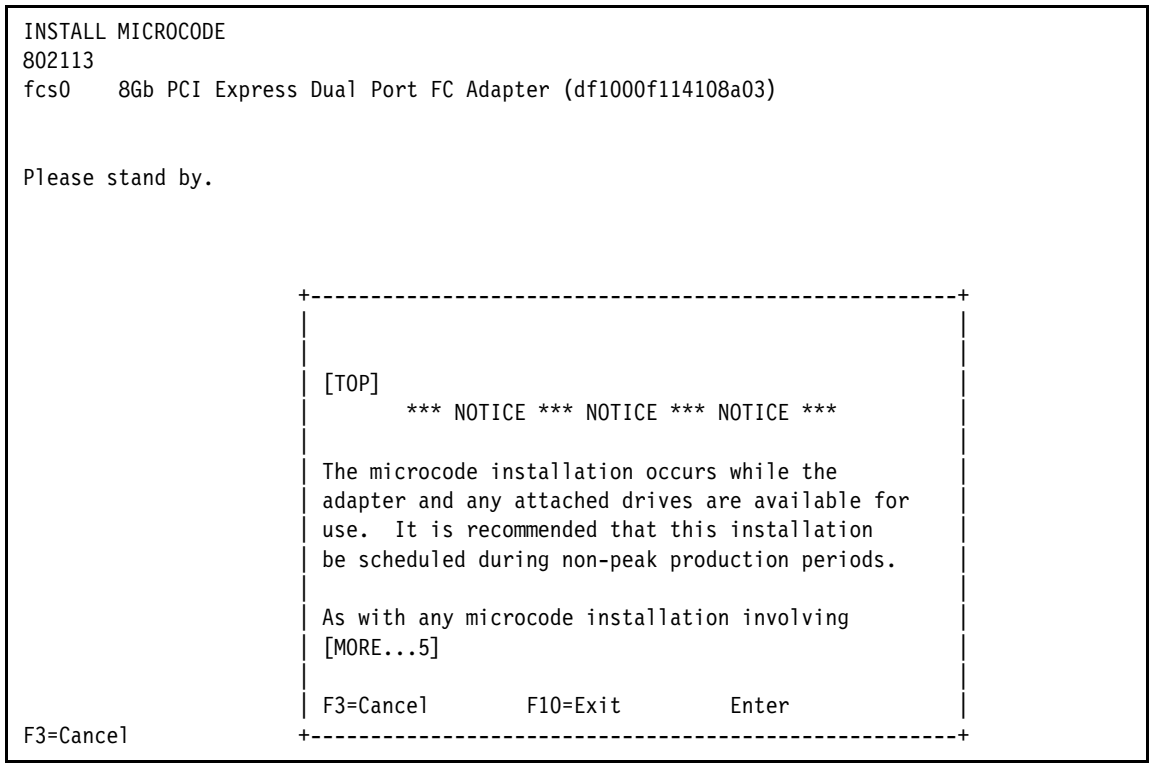


Figure 11-23 Diagnostic aids: Install microcode notice

17. Select /etc/microcode as the installation source and press Enter as shown in Figure 11-24.

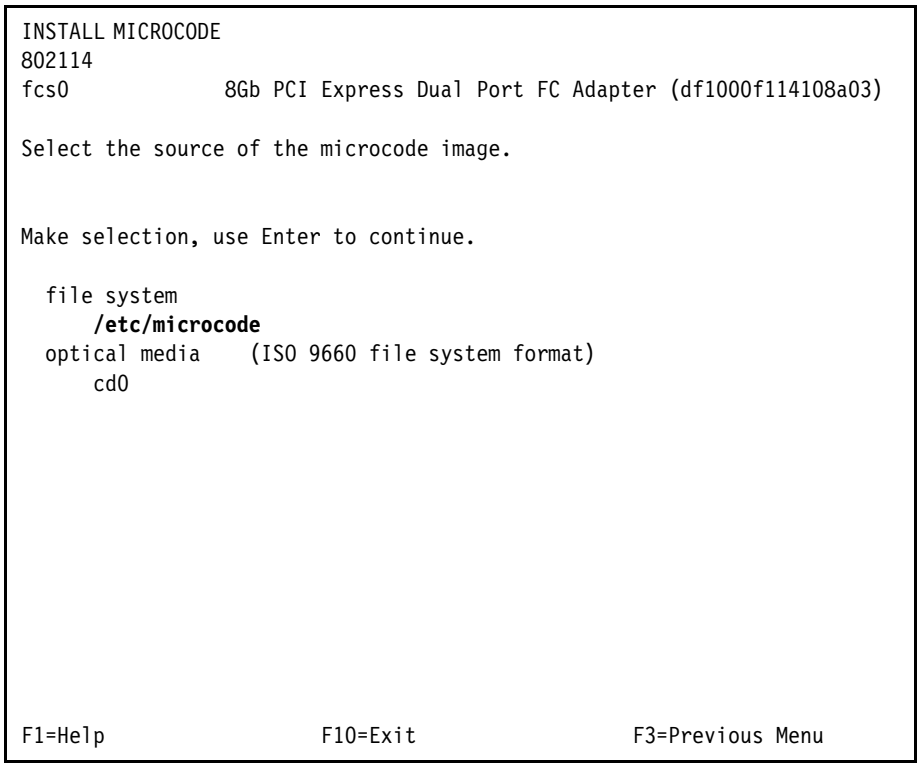


Figure 11-24 Diagnostics aids: Install microcode image source selection

18. Select the new adapter firmware microcode image for installation and press Enter to proceed as shown in Figure 11-25.

```
INSTALL MICROCODE
802116
fcs0    8Gb PCI Express Dual Port FC Adapter (df1000f114108a03)

The current microcode level for fcs0 is 110105.

Available levels to install are listed below.
Select the microcode level to be installed.

Use Help for explanations of "M", "L", "C"
and "P" .

Make selection, use Enter to continue.

  M  111304

F1=Help                 F10=Exit                 F3=Previous Menu
```

Figure 11-25 Diagnostics aids: Microcode level selection

19. The successful installation of the new adapter microcode is shown in Figure 11-26.

```
INSTALL MICROCODE
802118
fcs0    8Gb PCI Express Dual Port FC Adapter (df1000f114108a03)

Installation of the microcode has completed successfully.
The current microcode level for fcs0 is 111304.

Please run diagnostics on the adapter to ensure that it is
functioning properly.

Use Enter to continue.









F3=Cancel               F10=Exit                  Enter
```

Figure 11-26 Diagnostics aids: Install microcode success message

20. Repeat steps 16-19 for each adapter port selected. After all adapter ports are updated, exit the diagnostic aids by pressing **F10=Exit**.
21. As indicated on the Install Microcode panel, run diagnostic tests on each I/O adapter that was updated to ensure that it is functioning properly by making the following selections from the diagnostic aids main menu:
    a. Diagnostic Routines.
    b. System Verification.
    c. Select all updated adapter resources to test and press F7=Commit.

An example of a successful diagnostics test is shown in Figure 11-27.

```
TESTING COMPLETE on Sat Dec  4 15:51:36 EST 2010
801010

No trouble was found.

The resources tested were:

- sysplanar0                              System Planar
- fcs0                                    8Gb PCI Express Dual Port FC Adapter
                                          (df1000f114108a03)
  U789D.001.DQDYKYW-P1-C1-T1
- fcs1                                    8Gb PCI Express Dual Port FC Adapter
                                          (df1000f114108a03)
  U789D.001.DQDYKYW-P1-C1-T2

Use Enter to continue.









F3=Cancel               F10=Exit                  Enter
```

Figure 11-27 Diagnostic aids: Successful diagnostic test

11.1.11 Error logging on the Virtual I/O Server

Error logging on the Virtual I/O Server uses the same error logging facility as AIX. The error logging daemon is started with the **errdemon** command. This daemon reads error records from the `/dev/error` device and writes them to the error log in `/var/adm/ras/errlog`. **errdemon** also places specified notifications in the notification database at `/etc/objrepos/errnotify`.

The command to display binary error logs is **errlog**. Example 11-26 shows a short error listing.

Example 11-26 errlog short listing

```
$ errlog
IDENTIFIER TIMESTAMP  T C RESOURCE_NAME  DESCRIPTION
4FC8E358   1015104608 I O hdisk8         CACHED DATA WILL BE LOST IF CONTROLLER
B6267342   1014145208 P H hdisk12        DISK OPERATION ERROR
DF63A4FE   1014145208 T S vhost2         Virtual SCSI Host Adapter detected an
```

```
B6267342   1014145208 P H hdisk12        DISK OPERATION ERROR
DF63A4FE   1014145208 T S vhost2         Virtual SCSI Host Adapter detected an
B6267342   1014145208 P H hdisk11        DISK OPERATION ERROR
B6267342   1014145208 P H hdisk11        DISK OPERATION ERROR
C972F43B   1014111208 T S vhost4        Misbehaved Virtual SCSI ClientB6267342
B6267342   1014164108 P H hdisk14        DISK OPERATION ERROR
```

To obtain all the details listed for each event, you can use **errlog -ls** as shown in Example 11-27.

Example 11-27 Detailed error listing

```
$ errlog -ls |more
---------------------------------------------------------------------------
LABEL:          SC_DISK_PCM_ERR7
IDENTIFIER:     4FC8E358

Date/Time:       Wed Oct 15 10:46:33 CDT 2008
Sequence Number: 576
Machine Id:      00C1F1704C00
Node Id:         vios1
Class:           O
Type:            INFO
WPAR:            Global
Resource Name:   hdisk8

Description
CACHED DATA WILL BE LOST IF CONTROLLER FAILS

Probable Causes
USER DISABLED CACHE MIRRORING FOR THIS LUN

User Causes
CACHE MIRRORING DISABLED

        Recommended Actions
        ENABLE CACHE MIRRORING
...
```

All errors are divided into the classes that are listed in Table 11-3.

Table 11-3 Error log entry classes

Error log entry class	Description
H	Hardware error
S	Software error
O	Operator messages (logger)
U	Undetermined error class

Redirecting error logs to other servers

In certain cases, you might need to redirect error logs to one central instance. For example, you might need to be able to run automated error log analysis in one place. For the Virtual I/O Server, you must set up redirecting error logs to syslog first, and then assign the remote syslog host in the syslog configuration.

To redirect error logs to syslog, create the file `/tmp/syslog.add` with the content shown in Example 11-28.

Note: Before redirecting errors logs to syslog, you must first become the root user on the Virtual I/O Server by running this command:

```
$ oem_setup_env
```

Example 11-28 Content of /tmp/syslog.add file

```
errnotify:
en_pid = 0
en_name = "syslog"
en_persistenceflg = 1
en_method = "/usr/bin/errpt -a -l $1 |/usr/bin/fgrep -v 'ERROR_ID TIMESTAMP'|
/usr/bin/logger -t ERRDEMON -p local1.warn"
```

Use the **odmadd** command to add the configuration to the ODM:

```
# odmadd /tmp/syslog.add
```

In the syslog file, you can redirect all messages to any other server that is running **syslogd** and accepting remote logs. Simply add the following line to your /etc/syslog.conf file:

```
*.debug                 @9.3.5.115
```

Restart your syslog daemon by using the following command:

```
# stopsrc -s syslogd
0513-044 The syslogd Subsystem was requested to stop.
# startsrc -s syslogd
0513-059 The syslogd Subsystem has been started. Subsystem PID is 520236.
```

Troubleshooting error logs

If your error log becomes corrupted, you can always move the file, and a clean error log file is created as shown in Example 11-29.

Example 11-29 Creating an error log file

```
$ oem_setup_env
# /usr/lib/errstop
# mv /var/adm/ras/errlog /var/adm/ras/errlog.bak
# /usr/lib/errdemon
```

If you want to back up your error log to an alternate file and view it later, use the command that is shown in Example 11-30.

Example 11-30 Copy errlog and view it

```
$ oem_setup_env
# cp /var/adm/ras/errlog /tmp/errlog.save
# errpt -i /tmp/errlog.save
IDENTIFIER TIMESTAMP  T C RESOURCE_NAME  DESCRIPTION
4FC8E358   1015104608 I O hdisk8         CACHED DATA WILL BE LOST IF CONTROLLER
B6267342   1014145208 P H hdisk12        DISK OPERATION ERROR
DF63A4FE   1014145208 T S vhost2         Virtual SCSI Host Adapter detected an
B6267342   1014145208 P H hdisk12        DISK OPERATION ERROR
DF63A4FE   1014145208 T S vhost2         Virtual SCSI Host Adapter detected an
B6267342   1014145208 P H hdisk11        DISK OPERATION ERROR
B6267342   1014145208 P H hdisk11        DISK OPERATION ERROR
C972F43B   1014111208 T S vhost4         Misbehaved Virtual SCSI ClientB6267342
B6267342   1014164108 P H hdisk14        DISK OPERATION ERROR
```

11.1.12 VM Storage Snapshots/Rollback

VM Storage Snapshots/Rollback is a new function that allows multiple point-in-time snapshots of individual virtual machine storage. These point-in-time copies can be used to quickly roll back a virtual machine to a particular snapshot image. This function can be used to capture a VM image for cloning purposes or before you apply maintenance.

As a sample usage scenario, the SSR checks on the client partition for prerequisites that are needed for an upcoming hardware upgrade:

```
root@p71aix03 /root # lslpp -f devices.pci.2b102725.X11
  Fileset               File
  ----------------------------------------------------------------------
Path: /usr/lib/objrepos
  devices.pci.2b102725.X11 7.1.1.0
                        /usr/lpp/gai
                        /usr/lpp/gai/pci2b102725/loadddx
                        /usr/lpp/gai/pci2b102725
```

Have the Virtual I/O Server take snapshots of the **sspdisk04** logical unit during minimal I/O workload (every night at midnight) as shown in Example 11-31.

Example 11-31 snapshot create command

```
snapshot -clustername ssp_cluster -spname ssp_pool_1 -lu sspdisk04 -create
snap03_01
```

Tip: The snapshot must be created under minimal I/O load on the source LU.

Because of a power outage, the data on the partition's `rootvg` gets corrupted. The SSR notices that the prerequisite file set is no longer present:

```
root@p71aix03 /root # lslpp -f devices.pci.2b102725.X11
lslpp: Fileset devices.pci.2b102725.X11 not installed.
root@p71aix03 /root # lslpp -l devices.pci.2b102725.X11
lslpp: Fileset devices.pci.2b102725.X11 not installed.
 ls -l /usr/lpp/gai/pci2b102725/loadddx
/usr/lpp/gai/pci2b102725/loadddx not found
```

The SSR shuts down the client partition to restore the snapshot as shown in Example 11-32.

Example 11-32 snapshot rollback

```
snapshot -clustername ssp_cluster -spname ssp_pool_1 -rollback snap03_01 -lu
sspdisk04
```

Attention: You cannot run a rollback operation on a snapshot while the LU is in use. Therefore, when the client partition is in *Running* state, to roll back you must change the client partition to the *Not Activated* state.

Check file set integrity after the snapshot rollback on the client partition:

```
root@p71aix03 /root # lslpp -f devices.pci.2b102725.X11
  Fileset               File
  ----------------------------------------------------------------------------
Path: /usr/lib/objrepos
  devices.pci.2b102725.X11 7.1.1.0
                        /usr/lpp/gai
                        /usr/lpp/gai/pci2b102725/loadddx
                        /usr/lpp/gai/pci2b102725
root@p71aix03 /root # ls -l /usr/lpp/gai/pci2b102725/loadddx
-rw-r--r--    1 bin      bin          371219 Jul 25 23:46
/usr/lpp/gai/pci2b102725/loadddx
```

11.1.13 Automated management

The following topics explains how to automate and streamline the management of partitions with the HMC:

- Using System profiles
- Automating remote operations by using the HMC command-line interface
- Scheduling jobs on a Virtual I/O Server

The number of commands that are available on the HMC makes it impractical to document them all here. This chapter is designed to help you configure the interface and get you started with common tasks.

Using system profiles

You can make the startup of the partitions easier by placing them in a system profile and starting this system profile. System profiles contain logical partitions and an associated partition profile to use.

The menu to create a system profile is shown in Figure 11-28. To access the menu select a Managed System and click **Configuration** → **Manage System Profiles**. In this menu, you can select **New** to add a system profile. Give the system profile a name and select the partitions to be included in the profile.

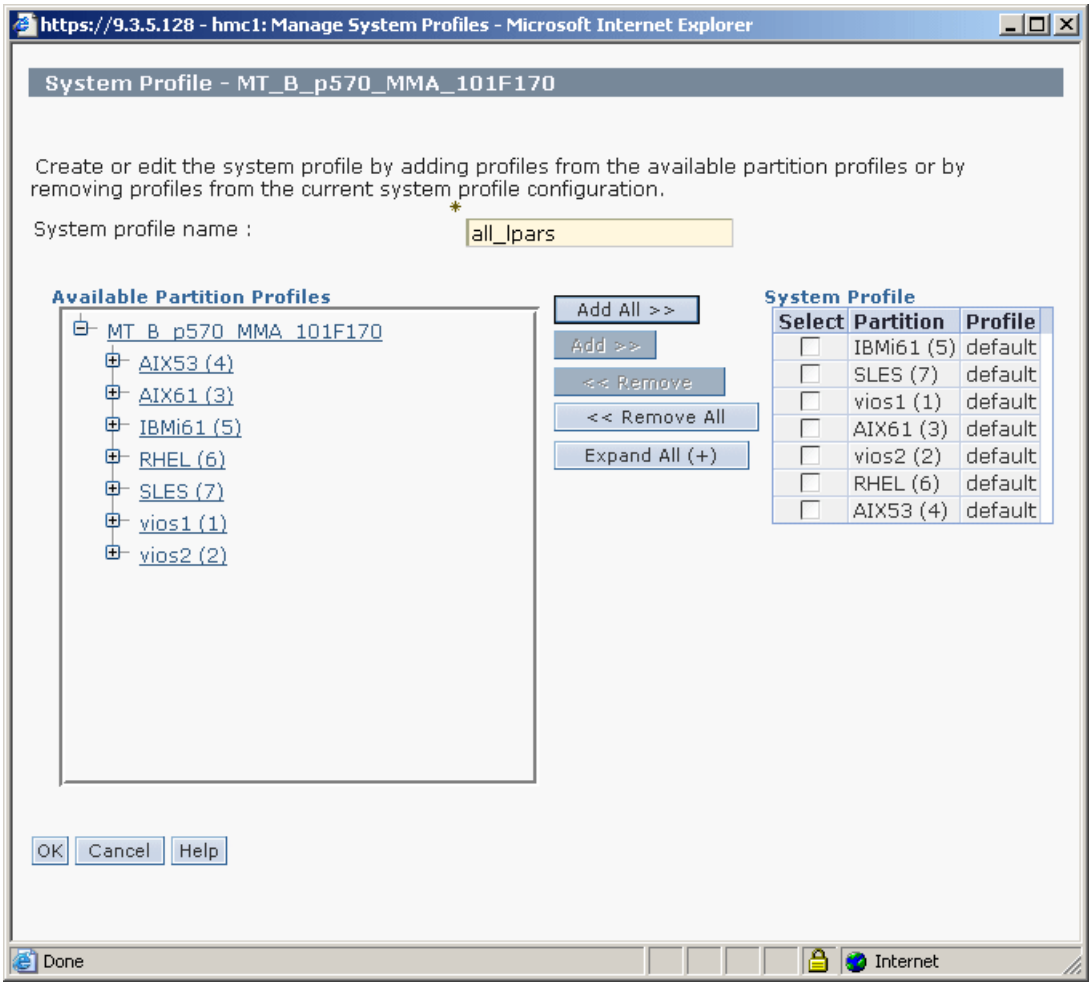


Figure 11-28 Creating a system profile on the HMC

Remember: When a Virtual I/O Server is part of a system profile, the system automatically starts this partition first.

For more information about system profiles, see the IBM Systems Hardware Information Center.

Using the HMC command-line interface

In an environment with numerous partitions or servers, it is convenient to perform operations using the HMC command-line interface (CLI) instead of the graphical user interface (GUI). You can perform almost all operations from the command line that can be done in the graphical interface. Using the CLI enables scripting, can reduce human error when you make numerous changes, and allows for the efficient repetition of common tasks.

From a system with an SSH client, you can either log in interactively to the HMC, or issue remote commands to the HMC in the same manner as accessing a remote UNIX or Linux server.

Configuring the Secure Shell interface

You must use the **ssh** protocol to access the HMC command-line interface remotely. In addition, remote command execution must be enabled on the HMC. It is found in the HMC Management panel as Remote Command Execution. Select **Enable remote command execution using the ssh facility** as shown in Figure 11-29.

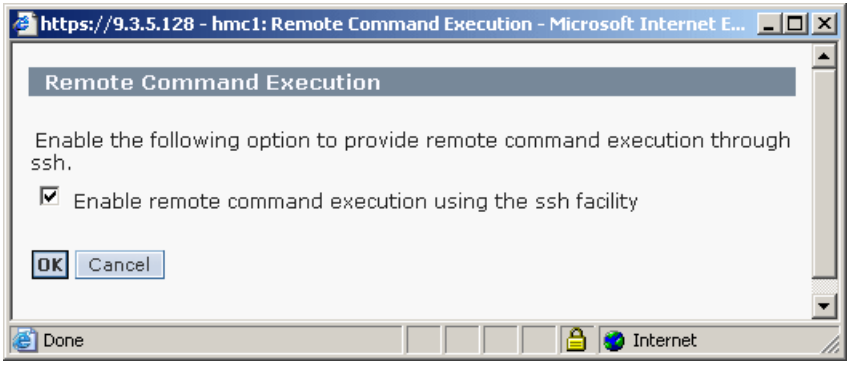


Figure 11-29 The HMC Remote Command Execution menu

Client configuration

Any SSH client program can be used to connect to the HMC. This section provides a few tips that can make remote administration easier when using the OpenSSH client from a UNIX or Linux based system.

Host-specific client customization

It is unlikely that you will be logged in to your workstation or central server with the same username that you use to access the HMC. This means that every time

you start an ssh connection to the HMC, you must specify the user name to use on the command line using the **-l** flag or `username@host` syntax. Failure to do so results in the SSH client trying to use your current user name on the HMC.

Example 11-33 shows the default behavior of the SSH client. The connection is attempted using the user name of the local user. In this case, the user is **root**. The second connection uses the `username@host` syntax.

Example 11-33 The default behavior of ssh

```
[root@Power7-2-RHEL ~]# ssh hmc9
root@hmc9's password:

[root@Power7-2-RHEL ~]# ssh hscroot@hmc9
hscroot@hmc9's password:
```

You can specify host-specific parameters in the `~/.ssh/ssh_config` file to tailor this behavior. Example 11-34 shows that the configuration file is configured so that all connections to systems that have host names that begin with *hmc* are to use the user name `hscroot`. Now the user name does not need to be specified every time that a connection is attempted.

Example 11-34 Using host-specific options

```
[root@Power7-2-RHEL ~]# cat ~/.ssh/config
Host hmc*
user hscroot
[root@Power7-2-RHEL ~]#

[root@Power7-2-RHEL ~]# ssh hmc9
Password:
Last login: Thu Dec  9 21:35:17 2010 from 172.16.254.38
hscroot@hmc9:~>
```

See the `ssh_config` man page for more parameters.

Public key authentication

For scripting purposes or your own convenience, you might want to configure your SSH client to use SSH public key authentication. This configuration allows you to log in remotely without being prompted for a password, while still providing a strong level of security.

The procedure for this is available in the IBM Systems Hardware Information Center at:

http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/ipha1/settingupsecurescriptexecution.htm

Example 11-35 shows a working example of the configuration.

Example 11-35 Configuring the SSH public key authentication

```
[root@Power7-2-RHEL ~]# ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/root/.ssh/id_rsa):
Created directory '/root/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /root/.ssh/id_rsa.
Your public key has been saved in /root/.ssh/id_rsa.pub.
The key fingerprint is:
e0:a6:74:65:84:49:e6:a9:ab:31:3a:a0:6c:5f:9d:fb root@Power7-2-RHEL
[root@Power7-2-RHEL ~]#

[root@Power7-2-RHEL ~]# scp .ssh/id_rsa.pub hmc9:
Password:
id_rsa.pub
100%  400     0.4KB/s   00:00

[root@Power7-2-RHEL ~]# ssh hmc9
Password:
Last login: Fri Dec 10 10:30:32 2010 from 172.16.20.174
hscroot@hmc9:~> KEY=`cat id_rsa.pub`
hscroot@hmc9:~> mkauthkeys -a "$KEY"
hscroot@hmc9:~> exit
exit
Connection to hmc9 closed.
[root@Power7-2-RHEL ~]#

[root@Power7-2-RHEL ~]# ssh hmc9 'ls -al'
total 56
drwxr-xr-x 5 hscroot hmc    4096 Dec 10 10:35 .
drwxr-xr-x 5 root    root   4096 Nov 29 16:37 ..
-rw------- 1 hscroot hmc   18111 Dec 10 10:35 .bash_history
-r-xr-xr-x 1 root    root     94 Nov 16 13:07 .bash_profile
-r-xr-xr-x 1 root    root    327 Nov 16 13:07 .bashrc
drwxr-xr-x 2 hscroot users  4096 Jun  2  2009 .fonts
drwxr-xr-x 2 hscroot users  4096 Jun  2  2009 .mozilla
drwxr-xr-x 2 root    hmc    4096 Jun  2  2009 .ssh
-rw-r--r-- 1 hscroot hmc      37 Nov 16 13:08 .version
-rw-r--r-- 1 hscroot hmc     400 Dec 10 10:35 id_rsa.pub
-rw-r--r-- 1 root    root      0 Dec  2 16:30 jobsMonitorThread.txt
```

Running non-interactive commands

The SSH client can pass commands to a remote SSH server in a non-interactive manner. This capability is useful for scripting. Example 11-36 shows a simple example of a non-interactive command.

Example 11-36 Running a non-interactive command

```
[root@Power7-2-RHEL ~]# /usr/bin/ssh hmc9 'lssyscfg -r sys -F name'
POWER7_2-061AB2P
POWER7_1-061AA6P
```

There are two points to notice in the previous example:

- The full path to the SSH client is used. This is generally a good practice to ensure that you are starting the correct client, and not an alias or similarly named file in your path.
- The remote command is enclosed in single quotation marks. This is sufficient for basic commands because the shell on the local system will not attempt to interpret anything between the quotation marks. When scripting, you might need to use double quotation marks or even escape quotation marks with the \ operator. This depends on which shell you want to interpret the commands. For more information, see any shell programming book or search for “shell programming” on the World Wide Web.

Initial login and shell conventions

After you have connected to the HMC, notice that the interface is a restricted version of the **bash** shell.

The following are a few tips for first time users:

- Pressing the Tab key twice shows all the commands that are available.
- Command names follow the regular IBM naming convention for UNIX interfaces. Commands that start with **ls** list information, and commands that start with **ch** change parameters.
- There are two methods to get help on the command line: either run the command with no parameters, or use the **man** page.
- Most of the **ls** commands provide a lot of information. The **-F** flag is useful in limiting the fields in the result set for readability.
- The **-F** flag allows you to use your own delimiter when you specify fields. If you use colons in your command, the output is delimited with semicolons. This is useful for scripting because some results include commas in the result data. This can make parsing difficult if you have used commas as your delimiter.

Basic reporting

The following are common commands to help get you started.

The following command lists all managed systems visible to the HMC:

```
lssyscfg -r sys
```

Most **ls** commands require the name of a managed system. You can use the **-F** flag to limit the output to just the name field. This is one command that is worth remembering because you will use it a lot. Where you see *<managed system name>* in further examples, a name from this output is required:

```
lssyscfg -r sys -F name
```

The following command lists all the partitions in a managed system:

```
lssyscfg -r lpar -m <managed system name>
```

To list all of the partitions visible to an HMC, use a simple `for` loop:

```
for i in `lssyscfg -r sys -F name` ; do lssyscfg -r lpar -m ${i} -F
name; done
```

The following command lists all the physical adapters in a managed system:

```
lshwres -m <managed system name> -r io --rsubtype slot -F
unit_phys_loc,phys_loc,description,lpar_name
```

Modifying the power state of partitions and systems

Changing the power state of partitions and servers is done by using the **chsysstate** command. In systems that support suspend and resume, the **chlparstate** command is used to manage these capabilities. It is possible to shut down a partition by using the **chlparstate** command.

To power on a system to partition standby, run the following command, where the managed system name is the name of the server as shown on the HMC:

```
chsysstate -m <managed system name> -o onstandby -r sys
```

To monitor the status of the server startup, use the **lsrefcode** command and check the LED status codes:

```
lsrefcode -r sys -m <managed system name> -F refcode
```

Run the following command to activate all partitions in the System Profile named `all_lpars`. When there are Virtual I/O Servers in the System Profile, these will automatically be activated before client partitions. If client partitions seem to be started first, they wait for the Virtual I/O Servers to be started.

```
chsysstate -m <managed system name> -o on -r sysprof -n all_lpars
```

Run the following command to shut down a partition immediately:

```
chsysstate -m <managed system> -r lpar -o shutdown -n <lpar name>
--immed
```

Modifying profiles

Modifications to profiles are done with the **chsyscfg** command. Commands to modify profiles can become quite long. However, when you are familiar with the syntax, it is efficient to use the command line, especially if multiple updates are required. The **man** page provides more examples and a detailed syntax guide.

Example 11-37 shows how to make the following changes on the *Normal* profile on partition 9:

- Decrease the minimum processing units by 0.1.
- Set the desired processing units to 0.2.
- Increase the maximum processing units by 0.2.

Example 11-37 Profile modification

```
chsyscfg -r prof -m POWER7_2-061AB2P -i
"name=Normal,lpar_id=9,min_proc_units-=0.1,desired_proc_units=0.2,max_proc_unit
s+=0.2"
```

Dynamic LPAR operations

The **chhwres** command is used to run dynamic LPAR operations. The **man** page has many examples of how to use the **chhwres** command. A few are listed here.

Example 11-38 shows how to increase by 128 MB the memory of the partition with ID 1, and time out after 10 minutes.

Example 11-38 Memory dynamic operation

```
chhwres -r mem -m <managed system name> -o a --id 1 -q 128 -w 10
```

Example 11-39 shows how to add a virtual Ethernet adapter with a port VLAN ID of 4 to the partition with ID 3. The adapter also has the trunk flag enabled and trunks VLAN 5 and 6 with a priority of 1.

Example 11-39 Virtual adapter dynamic operation

```
chhwres -r virtualio -m POWER7_2-061AB2P -o a --id 3 --rsubtype eth -a
"ieee_virtual_eth=1,port_vlan_id=4,"addl_vlan_ids=5,6",is_trunk=1,trunk_priorit
y=1"
```

Scheduling jobs on the Virtual I/O Server

Starting with Virtual I/O Server version 1.3, the **crontab** command is available to enable you to submit, edit, list, and remove cron jobs. A cron job is a command that is run by the cron daemon at regularly scheduled intervals, such as system tasks, nightly security checks, analysis reports, and backups.

With the Virtual I/O Server, a cron job can be submitted by specifying the **crontab** command with the **-e** flag. The **crontab** command starts an editing session to modify the padmin users' crontab file, and create entries for each cron job in this file.

Tip: When scheduling jobs, use the padmin user's crontab file. You cannot create or edit other users' crontab files.

When you finish creating entries and exit the file, the **crontab** command copies it into the `/var/spool/cron/crontabs` directory and places it in the `padmin` file.

The following syntax is available to the **crontab** command:

```
crontab [-e padmin | -l padmin | -r padmin | -v padmin]
```

-e padmin This edits a copy of the padmin's crontab file. When editing is complete, the file is copied into the crontab directory as the padmin's crontab file.

-l padmin This lists the padmin's crontab file.

-r padmin This removes the padmin's crontab file from the crontab directory.

-v padmin This lists the status of the padmin's cron jobs.

11.1.14 Virtualization management tools

This section describes different virtualization management tools that can be used to manage a POWER virtualized environment

Virtual I/O Server Performance Advisor

The Virtual I/O Server Performance Advisor tool provides advisory reports. These reports are based on the key performance metrics on various partition resources that are collected from the Virtual I/O Server environment.

Starting with Virtual I/O Server version 2.2.2.0, you can use the Virtual I/O Server Performance Advisor tool. Use this tool to provide health reports that have proposals for making configurational changes to the Virtual I/O Server environment, and to identify areas to investigate further. On the Virtual I/O Server

command line, enter the **part** command to start the Virtual I/O Server Performance Advisor tool.

You can start the Virtual I/O Server Performance Advisor tool in the following modes:

- On-demand monitoring mode
- Postprocessing mode

When you start the Virtual I/O Server Performance Advisor tool in the on-demand monitoring mode, provide the duration to monitor the system for in minutes. The duration that you provide must be 10 - 60 minutes, at the end of which the tool generates the reports. During this time, samples are collected at regular intervals of 15 seconds. For example, to monitor the system for 30 minutes and generate a report, enter the following command:

```
part –i 30
```

Reports for the on-demand monitoring mode are generated in the `ic43_120228_06_15_20.tar` file.

The output that is generated by the **part** command is saved in a `.tar` file, which is created in the current working directory. The naming convention for files in the on-demand monitoring mode is `short-hostname_yymmdd_hhmmss.tar`. In the postprocessing mode, the file name is that of the input file with the file name extension changed from a *`.nmon`* file to a *`.tar`* file.

When you start the Virtual I/O Server Performance Advisor tool in the postprocessing mode, you must provide a file as the input. The tool tries to extract as much data as possible from the file that you provide, and then generates reports. If the file does not have the required data for the tool to generate reports, an `Insufficient Data` message is added to the relevant fields. For example, to generate a report that is based on the data available in the `ic43_120206_1511.nmon` file, enter the following command:

```
part -f ic43_120206_1511.nmon
```

Reports for the postprocessing mode are successfully generated in the `ic43_120206_1511.tar file`.

Note: The size of the input file in the postprocessing mode must be less than 100 MB. The postprocessing of huge data results in more time to generate the reports. For example, if the size of a file is 100 MB and the Virtual I/O Server has 255 disks that are configured, with more than 4000 samples, it might take 2 minutes to generate the reports.

For Virtual I/O Servers before version 2.2.2.0, see this online reference:

http://www.ibm.com/developerworks/wikis/display/WikiPtype/VIOS+Advisor

Virtual I/O Server Performance Advisor reports

The Virtual I/O Server Performance Advisor tool provides advisory reports that are related to performance of various subsystems in the Virtual I/O Server environment. New Virtual I/O Server Performance Advisor analyzes Virtual I/O Server performance, and makes recommendations for performance optimization.

The output that is generated by the **part** command is saved in a `.tar` file that is created in the current working directory. The `vios_advisor.xml` report is present in the output `.tar` file with the other supporting files. To view the generated report, complete the following steps:

1. Transfer the generated `.tar` file to a system that has a browser and a `.tar` file extractor installed.
2. Extract the `.tar` file.
3. Open the `vios_advisor.xml` file that is in the extracted directory.

The `vios_advisor.xml` file structure is based on an XML Schema Definition (XSD) in the `/usr/perf/analysis/vios_advisor.xsd` file.

Each report is shown in a tabular form. The descriptions of all of the columns are provided in Table 11-4.

Table 11-4 Performance metrics

Performance metrics	Description
Measured Value	This metric displays the values that are related to the performance metrics collected over a period.
Recommended Value	This metric displays all the suggested values when the performance metrics pass the critical thresholds.
First Observed	This metric displays the time stamp when the measured value is first observed.
Last Observed	This metric displays the time stamp when the measured value is last observed.
Risk	If either the warning or the critical thresholds are passed, the risk factor is indicated on a scale of 1 - 5, with 1 being the lowest value and 5 being the highest value.

Performance metrics	Description
Impact	If either the warning or critical thresholds are passed, the impact is indicated on a scale of 1 - 5, with 1 being the lowest value and 5 being the highest value.

The following are the types of advisory reports that are generated by the Virtual I/O Server Performance Advisor tool:

- System configuration advisory report
- CPU (central processing unit) advisory report
- Memory advisory report
- Disk advisory report
- Disk adapter advisory report
- I/O activities (disk and network) advisory report

The system configuration advisory report consists of the information that is related to the Virtual I/O Server configuration. This information includes processor family, server model, number of cores, frequency at which the cores are running, and the Virtual I/O Server version. The output is similar to that shown in Figure 11-30.

SYSTEM - CONFIGURATION

	Name	Value
	Processor Family	POWER7
	Server Model	IBM,9117-MMC
	Server Frequency	3.920 GHz
	Server - Online CPUs	16 cores
	Server - Maximum Supported CPUs	64 cores
	VIOS Level	2.2.1.0
	VIOS Advisor Release	081711A

Figure 11-30 System - Configuration

The CPU advisory report consists of the information that is related to the processor resources. This information includes the number of cores assigned to the Virtual I/O Server, processor consumption during the monitoring interval, and shared processor pool capacity for shared partitions. The output is similar to that shown in Figure 11-31.

VIOS - CPU

	Name	Measured Value	Recommended Value	First Observed	Last Observed	Risk 1=lowest 5=highest	Impact 1=lowest 5=highest
✓	CPU Capacity	4.0 ent	-	08/17 13:25:13	-	n/a	n/a
i	CPU Consumption	avg:27.1% (cores:1.1) high:27.4% (cores:1.1)	-	-	-	n/a	n/a
i	Processing Mode	Shared CPU, (UnCapped)	-	08/17 13:25:13	-	n/a	n/a
⚠	Variable Capacity Weight	128	129-255	08/17 13:25:13	-	1	5
✓	Virtual Processors	4	-	08/17 13:25:13	-	n/a	n/a
✓	SMT Mode	SMT4	-	08/17 13:25:13	-	n/a	n/a

SYSTEM - SHARED PROCESSING POOL

	Name	Measured Value	Recommended Value	First Observed	Last Observed	Risk 1=lowest 5=highest	Impact 1=lowest 5=highest
✓	Shared Pool Monitoring	enabled	-	08/17 13:25:13	-	n/a	n/a
i	Shared Processing Pool Capacity	16.0 ent.	-	08/17 13:25:13	-	n/a	n/a
✓	Free CPU Capacity	avg_free:14.9 ent. lowest_free:14.8 ent.	-	-	-	n/a	n/a

Figure 11-31 Virtual I/O Server -CPU

Note: In the Virtual I/O Server - CPU table, the status of the variable capacity weight is marked with the **Warning** icon because the best practice is for the Virtual I/O Server to have an increased priority of 129 - 255 when in uncapped shared processor mode. See Table 11-5 on page 428 for the definition of the **Warning** icon.

The memory advisory report consists of the information that is related to the memory resources. This information includes the available free memory, paging space that is allocated, paging rate, and pinned memory. The output is similar to that shown in Figure 11-32.

VIOS - MEMORY

	Name	Measured Value	Recommended Value	First Observed	Last Observed	Risk 1=lowest 5=highest	Impact 1=lowest 5=highest
	Real Memory	4.000 GB	7.000 GB	08/17 13:25:13	-	1	5
	Available Memory	0.571 GB	1.5 GB Avail.	08/17 13:25:33	08/17 13:29:30	n/a	n/a
	Paging Rate	163.8 MB/s pg rate	No Paging	08/17 13:25:33	08/17 13:30:00	n/a	n/a
	Paging Space Size	1.500 GB	-	08/17 13:25:13	-	n/a	n/a
	Free Paging Space	1.491 GBfree	-	-	-	n/a	n/a
	Pinned Memory	0.748 GB pinned	-	-	-	n/a	n/a

Figure 11-32 Virtual I/O Server - Memory

Note: In this report, the status of the real memory is marked with the **Critical** icon because the available memory is less than the 1.5 GB limit that is specified in the Recommended Value column of the available memory. See Table 11-5 on page 428 for the definition of the **Critical** icon.

The disk advisory report consists of the information that is related to the disks attached to the Virtual I/O Server. This information includes the I/O activities that are getting blocked and I/O latencies. The output is similar to that shown in Figure 11-33.

VIOS - DISK DRIVES

	Name	Measured Value	Recommended Value	First Observed	Last Observed	Risk 1=lowest 5=highest	Impact 1=lowest 5=highest
	Physical Drive Count	13	-	08/17 13:25:13	-	n/a	n/a
	I/Os Blocked (hdisk0)	high:9.1% I/Os blocked	5.0% or less	08/17 13:25:45	08/17 13:28:45	n/a	n/a
	Long I/O Latency	pass	-	-	-	n/a	n/a

Figure 11-33 Virtual I/O Server - Disk Drives

The disk adapter advisory report consists of information that is related to the Fibre Channel adapters that are connected to the Virtual I/O Server. This report illustrates the information that is based on the average I/O operations per second, adapter utilization, and running speed. The output is similar to that shown in Figure 11-34.

VIOS - DISK ADAPTERS

	Name	Measured Value	Recommended Value	First Observed	Last Observed	Risk 1=lowest 5=highest	Impact 1=lowest 5=highest
	FC Adapter Count	2	-	08/17 13:25:13	-	n/a	n/a
	FC Avg IOps	avg: 827 iops @ 3KB	-	08/17 13:25:13	08/17 13:30:13	n/a	n/a
	FC Idle Port: (fcs1)	idle	-	08/17 13:25:13	08/17 13:30:13	4	4
	FC Adapter Utilization	pass	-	-	-	n/a	n/a
	FC Port Speeds	running at speed	-	-	-	n/a	n/a

Figure 11-34 Virtual I/O Server - Disk Adapters

Note: In this report, the status of the Fibre Channel idle port is marked with the **Investigate** icon because the tool identifies a Fibre Channel adapter that is not used often. See Table 11-5 on page 428 for the definition of the **Investigate** icon.

The I/O activity advisory report consists of the following information:

- Disk I/O activity, such as average and peak I/O operations per second
- Network I/O activity, such as average and peak inflow and outflow I/O per second

The output is similar to that shown in Figure 11-35.

VIOS - I/O ACTIVITY

	Name	Value
	Disk I/O Activity	avg: 1906 iops @ 103KB peak: 1893 iops @ 57KB
	Network I/O Activity	[avgSend: 9641 iops 0.6MBps , avgRcv: 75914 iops 97.7MBps] [peakSend: 9956 iops 0.6MBps , peakRcv: 78668 iops 112.5MBps]

Figure 11-35 Virtual I/O Server - I/O Activity

The details related to these advisory reports can also be obtained by clicking the respective report fields from the browser. The following details are available for all the advisory reports:

- What Is This: Brief description of the advisory field
- Why Important: Significance of the particular advisory field
- How to Modify: Details related to the configuration steps that you can use to modify the parameters that are related to the particular advisory field

For example, to know more about the processor capacity, you can click the corresponding row in the Virtual I/O Server - CPU table. The information is then displayed.

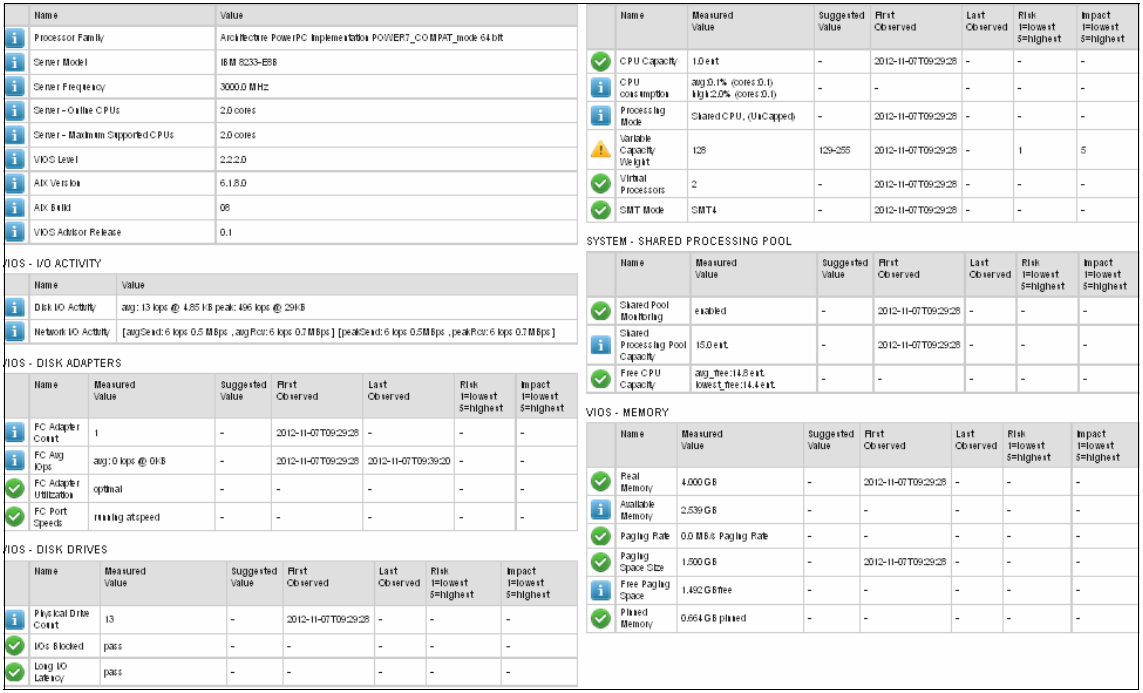
Note: The suggested values are based on the behavior during the monitoring period. Therefore, the values can be used only as a guideline.

Table 11-5 describes the icon definitions.

Table 11-5 Icon definitions

Icons	Definitions
	Information that is related to configuration parameters
	Values acceptable in most cases
	Possible performance problem
	Severe performance problem
	Investigation required

The final output that you see in your browser is shown in Figure 11-36.



	Name	Value
	Processor Family	Architecture PowerPC Implementation POWER7_COMPAT_mode 64-bit
	Server Model	IBM 8233-E8B
	Server Frequency	3000.0 MHz
	Server - Online CPUs	2.0 cores
	Server - Maximum Supported CPUs	2.0 cores
	VIOS Level	2.2.2.0
	ADK Version	6.1.8.0
	ADK Build	08
	VIOS Advisor Release	0.1

VIOS - I/O ACTIVITY

	Name	Value
	Disk I/O Activity	avg: 13 Iops @ 4.85 KB peak: 496 Iops @ 29KB
	Network I/O Activity	[avgSend: 6 Iops 0.5 MBps , avgRcv: 6 Iops 0.7 MBps] [peakSend: 6 Iops 0.5MBps , peakRcv: 6 Iops 0.7 MBps]

VIOS - DISK ADAPTERS

	Name	Measured Value	Suggested Value	First Observed	Last Observed	Risk 1=lowest 5=highest	Impact 1=lowest 5=highest
	FC Adapter Count	1	-	2012-11-07T09:29:28	-	-	-
	FC Avg Iops	avg: 0 Iops @ 0KB	-	2012-11-07T09:29:28	2012-11-07T09:39:20	-	-
	FC Adapter Utilization	optimal	-	-	-	-	-
	FC Port Speeds	running at speed	-	-	-	-	-

VIOS - DISK DRIVES

	Name	Measured Value	Suggested Value	First Observed	Last Observed	Risk 1=lowest 5=highest	Impact 1=lowest 5=highest
	Physical Drive Count	13	-	2012-11-07T09:29:28	-	-	-
	IOs Blocked	pass	-	-	-	-	-
	Long IO Latency	pass	-	-	-	-	-

	Name	Measured Value	Suggested Value	First Observed	Last Observed	Risk 1=lowest 5=highest	Impact 1=lowest 5=highest
	CPU Capacity	1.0 ent.	-	2012-11-07T09:29:28	-	-	-
	CPU consumption	avg:0.1% (cores:0.0) high:2.0% (cores:0.0)	-	-	-	-	-
	Processing Mode	Shared CPU, (UnCapped)	-	2012-11-07T09:29:28	-	-	-
	Variable Capacity Weight	128	129-255	2012-11-07T09:29:28	-	1	5
	Virtual Processors	2	-	2012-11-07T09:29:28	-	-	-
	SMT Mode	SMT4	-	2012-11-07T09:29:28	-	-	-

SYSTEM - SHARED PROCESSING POOL

	Name	Measured Value	Suggested Value	First Observed	Last Observed	Risk 1=lowest 5=highest	Impact 1=lowest 5=highest
	Shared Pool Monitoring	enabled	-	2012-11-07T09:29:28	-	-	-
	Shared Processing Pool Capacity	15.0 ent.	-	2012-11-07T09:29:28	-	-	-
	Free CPU Capacity	avg_free:14.8 ent. lowest_free:14.4 ent.	-	-	-	-	-

VIOS - MEMORY

	Name	Measured Value	Suggested Value	First Observed	Last Observed	Risk 1=lowest 5=highest	Impact 1=lowest 5=highest
	Real Memory	4.000 GB	-	2012-11-07T09:29:28	-	-	-
	Available Memory	2.539 GB	-	-	-	-	-
	Paging Rate	0.0 MB/s Paging Rate	-	-	-	-	-
	Paging Space Size	1.500 GB	-	2012-11-07T09:29:28	-	-	-
	Free Paging Space	1.492 GB free	-	-	-	-	-
	Pinned Memory	0.664 GB pinned	-	-	-	-	-

Figure 11-36 Overview picture of Virtual I/O Server advisor output

11.2 Monitoring Virtual I/O Servers

This section describes how to monitor your advanced PowerVM environment. An introduction to performance considerations is presented. Then, various monitoring techniques that are based on common situations are demonstrated and finally, the client monitoring tools are described.

This section includes the following topics:

- Overview of selected tools
- Monitoring global system resource allocations
- Monitoring commands on the Virtual I/O Server
- Third-party monitoring tools
- Other monitoring tools

11.2.1 Overview of selected tools

Table 11-6 provides an overview of selected tools that can be used for monitoring resources like processor, memory, storage, and network in a Virtual I/O Server virtualized environment that includes AIX, IBM i, and Linux virtual I/O client partitions.

Table 11-6 Tools for monitoring resources in a virtualized environment

Resource or Platform	Processor	Memory	Storage	Network
AIX	**topas** **nmon** PM for Power Systems	**topas** **nmon** PM for Power Systems	**topas** **nmon** **iostat** **fcstat** PM for Power Systems	**topas** **nmon** **entstat** PM for Power Systems
IBM i	**WRKSYSACT** IBM Performance Tools for IBM i, IBM Systems Director Navigator for i PM for Power Systems	**WRKSYSSTS, WRKSHRPOOL** IBM Performance Tools for IBM i, IBM Systems Director Navigator for i PM for Power Systems	**WRKDSKSTS** IBM Performance Tools for IBM i, IBM Systems Director Navigator for i PM for Power Systems	**WRKTCPSTS** IBM Performance Tools for IBM i, IBM Systems Director Navigator for i PM for Power Systems
Linux	**iostat** **sar**	/proc/meminfo	**iostat**	**netstat** **iptraf**

Resource or Platform	Processor	Memory	Storage	Network

Virtual I/O Server	`topas`
	`viostat`
		`topas`
		`vmstat`
		`svmon`
			`topas`
			`viostat`
			`fcstat`
				`topas`
				`entstat`

System-wide	IBM Director (all clients) `topas` (AIX, Virtual I/O Server), IBM System i Navigator (IBM i)	`topas` (AIX, Virtual I/O Server), Navigator for i (IBM i)	`topas` (AIX, Virtual I/O Server), Navigator for i (IBM i)	`topas` (AIX, Virtual I/O Server), Navigator for i (IBM i)

Notes:

- For IBM i, the IBM Director 6.1 has the IBM Systems Director Navigator for i that is fully integrated for all level 0 or higher IBM i 6.1 or IBM i 5.4 agents.
- For Linux systems, the sysstat RPM must be installed for resource performance monitoring.
- `topas` supports cross-partition information for AIX and Virtual I/O Server partitions only.

11.2.2 Monitoring global system resource allocations

To precisely monitor the consumption of various resources such as processor, memory, network adapters, and storage adapters, first have a clear understanding of the global resource allocations on the system. This chapter explains how to monitor those allocations.

Several tools provide information about resource allocations. The first tool is the Hardware Management Console (HMC), which is also used to manage resource allocations. If a system is managed by an Integrated Virtualization Manager (IVM) rather than an HMC, the IVM can also be used to monitor resource allocations. On a virtual I/O client, the `lparstat` command can run on any AIX partition and inspect its current resource allocations. A Linux alternative command is also available.

This section includes the following topics:

- Hardware Management Console (HMC) monitoring
- Integrated Virtualization Manager (IVM) monitoring
- Monitoring resource allocations from a partition

Hardware Management Console (HMC) monitoring

The HMC is generally used to set up system allocations and maintain the system. However, it is also useful for monitoring the current resource allocations.

To monitor current system allocations, you must log on to the HMC web interface. Open the login window by accessing the IP or DNS name of your HMC using https through a web browser.

You first see the global server allocations by selecting **Systems Management** → **Servers** from the left pane. The list of managed systems is then displayed in the main pane as shown in Figure 11-37.

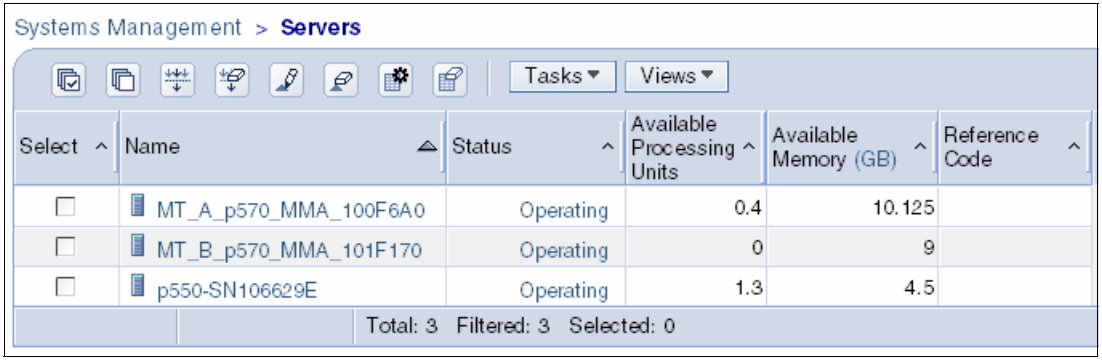


Figure 11-37 Available servers being managed by the HMC

The unallocated resources are directly visible, which simplifies the administrator's investigations. Moreover, it is possible to view extra information by clicking **Configure Columns** as shown in Figure 11-38.

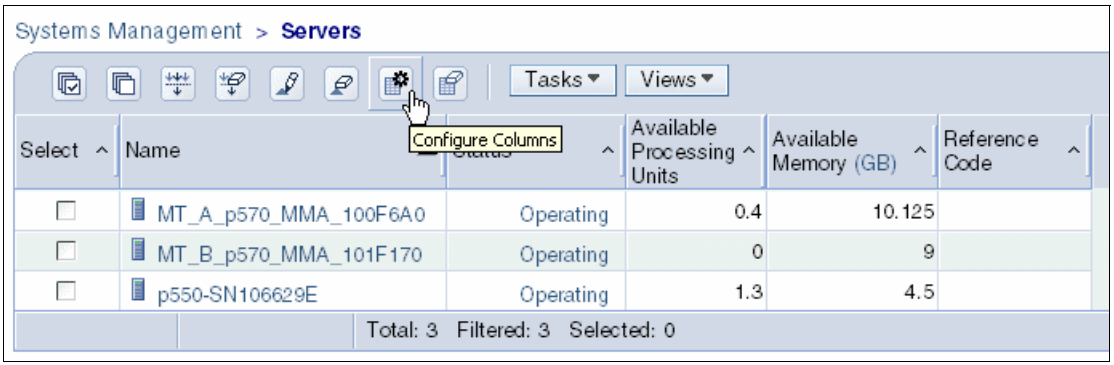


Figure 11-38 Configuring the displayed columns on the HMC

You can then track the resources allocation information that you are interested in.

Partition properties monitoring

To obtain detailed information for a partition, select its system name and then select the name of the partition. A new window opens and displays the partition properties. Navigating in the tabs shows the current resources allocations as shown in Figure 11-39.

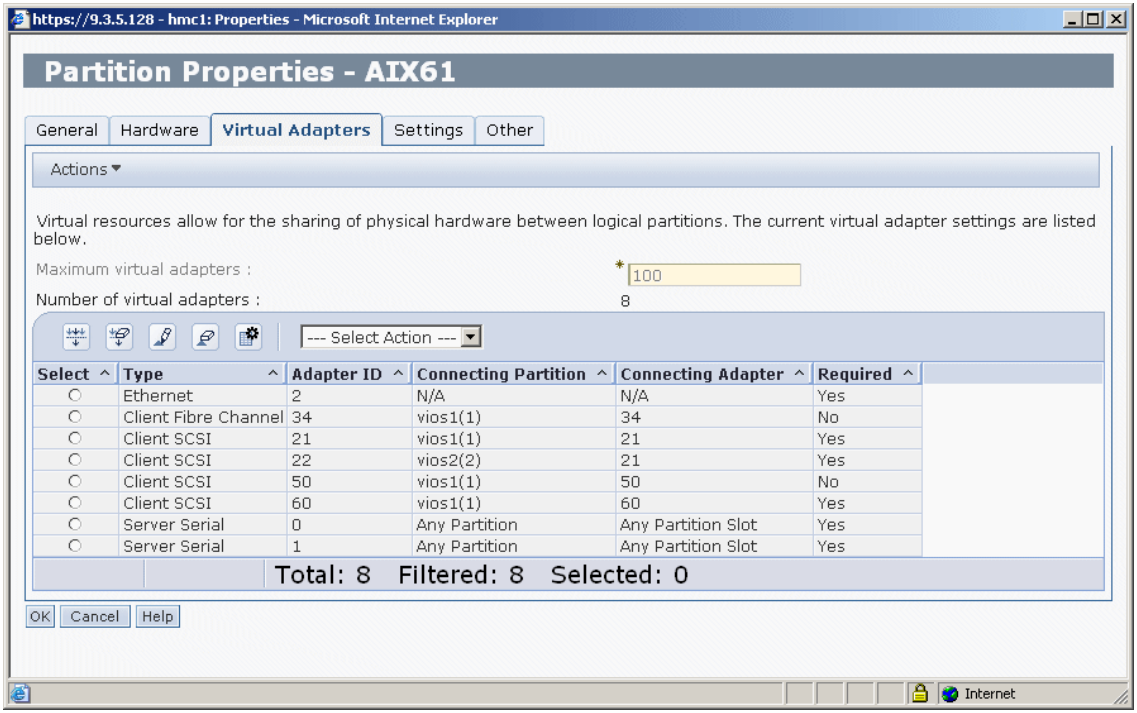


Figure 11-39 Virtual adapters configuration in the partition properties

HMC hardware information monitoring

Starting with HMC V7, you can now look at virtual SCSI and virtual LAN topologies of the Virtual I/O Server from the HMC.

Tip: The HMC has a feature to aid administrators in looking at virtual SCSI and virtual Ethernet topologies in the Virtual I/O Server.

To view these topologies, select the Virtual I/O Server partition where you want to see the topologies and select **Hardware Information** → **Virtual I/O Adapters** → **SCSI** as shown in Figure 11-40.

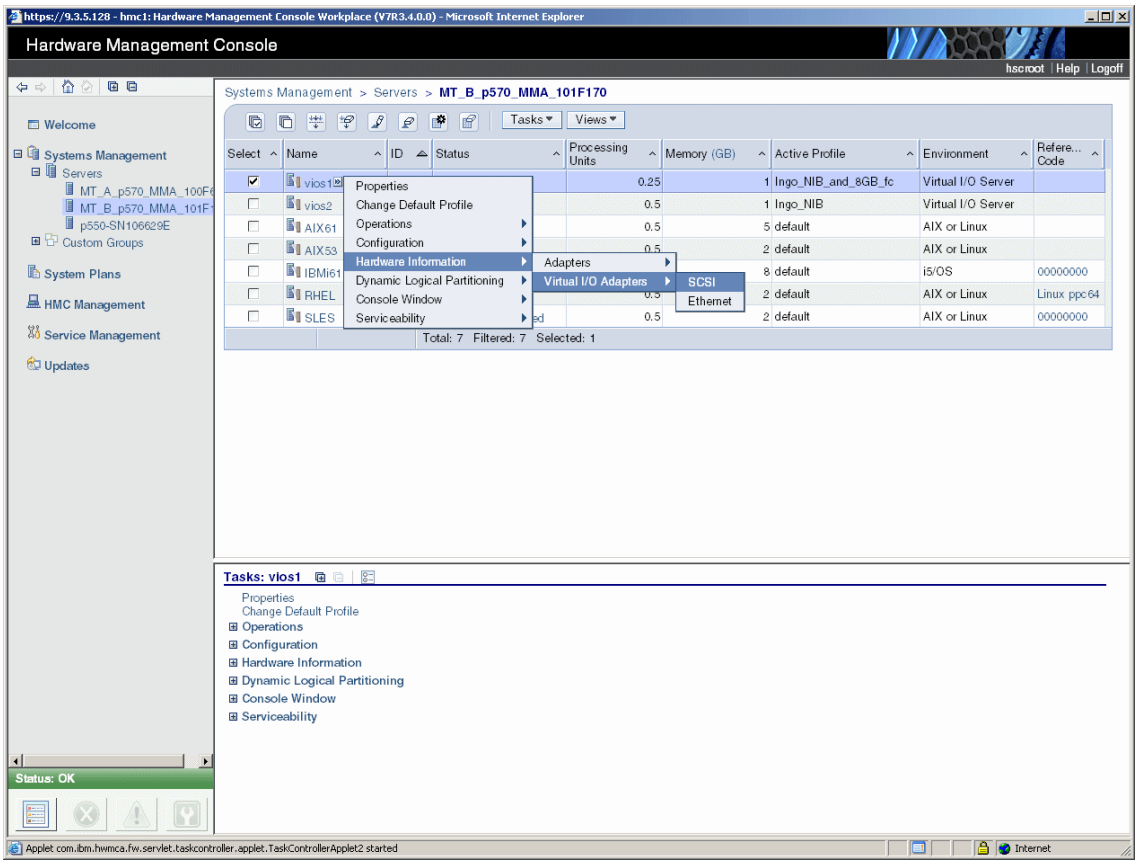


Figure 11-40 Virtual I/O Server hardware information menu

A window is then displayed that shows the virtual storage topology, as shown in Figure 11-41.

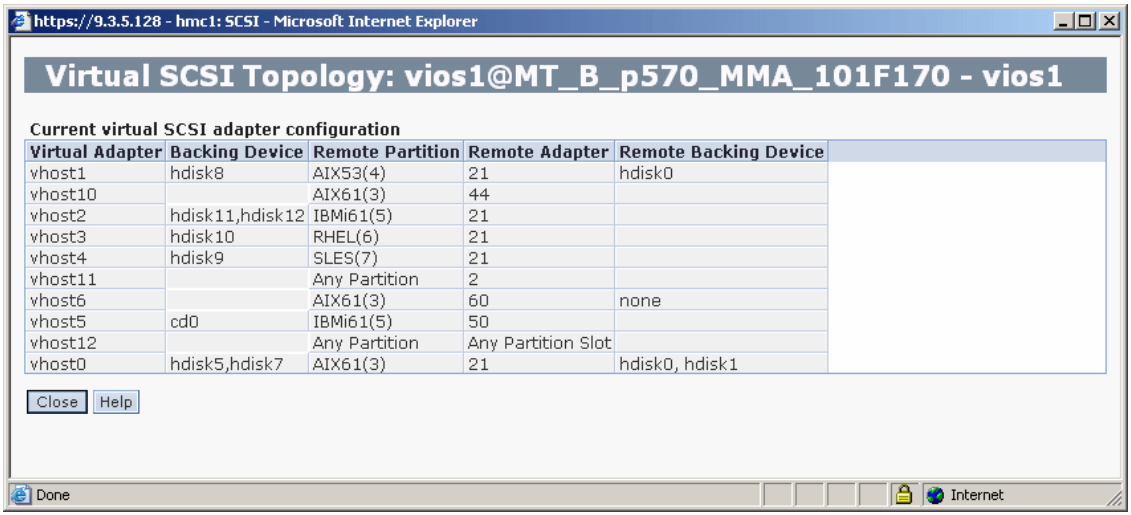


Figure 11-41 The Virtual I/O Server virtual SCSI topology window

HMC virtual storage monitoring

You can monitor the current storage allocations by using the HMC. The HMC shows you the storage allocations of physical disks and virtual disks, including shared storage pool information.

The shared storage pool improves storage utilization, reduces storage infrastructure cost, and reduces administration costs. For more information, see 10.1, “Managing shared storage pools” on page 286.

To obtain detailed information for the storage allocations, select the server where the Virtual I/O Server is that you want to check the mappings. Then, click **Tasks** → **Configuration** → **Virtual Resources** → **Virtual Storage Management**. Selecting the target Virtual I/O Server in a new window shows the information for storage allocations.

As of HMC Version 7 Release 7.4.0, you can select **Show shared storage pool storage** on the lower right corner, as shown in Figure 11-42.

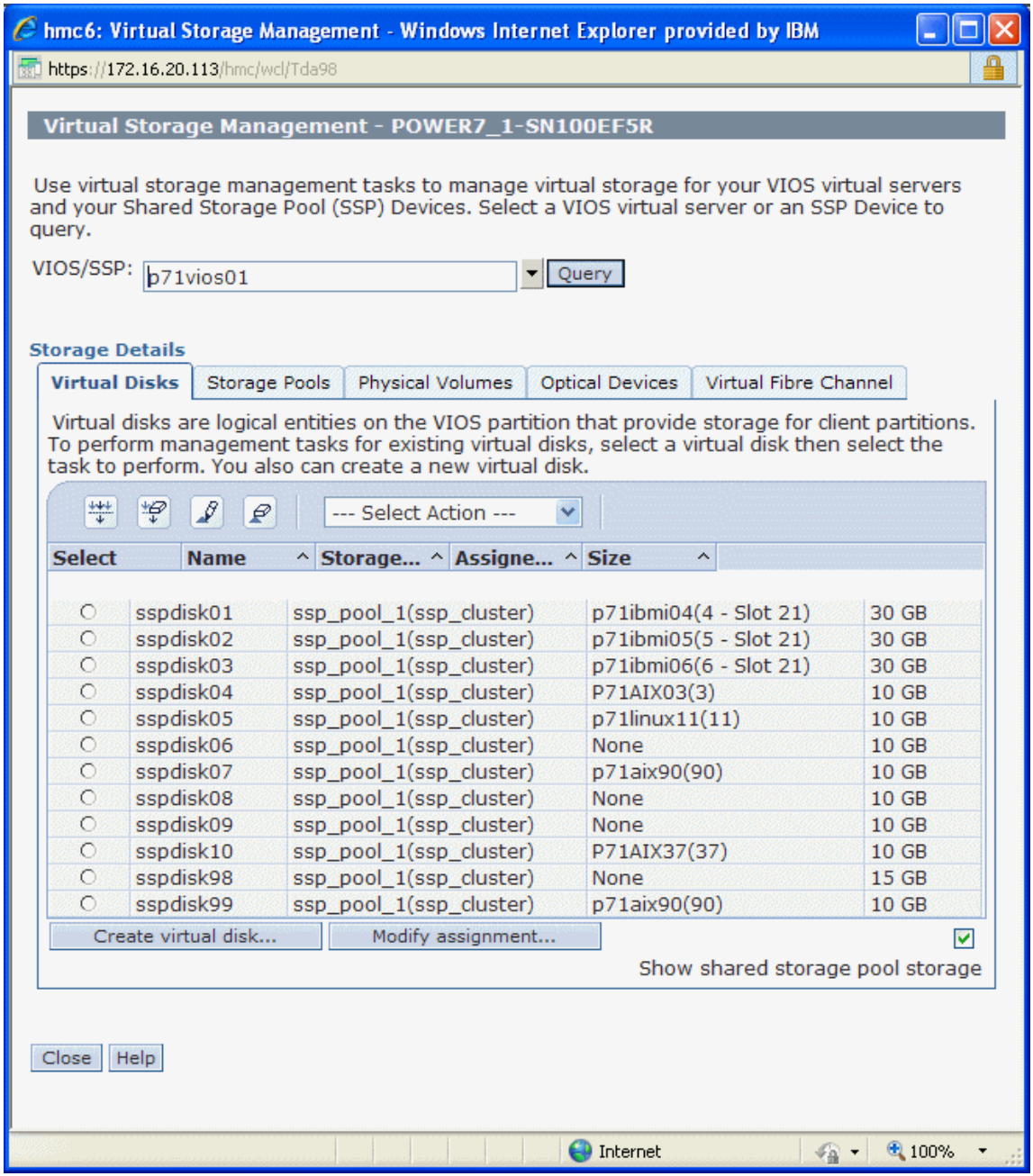


Figure 11-42 HMC Virtual Storage Management window

HMC virtual network monitoring

Starting with HMC V7, you can monitor the virtual network information for each server that is attached to the HMC.

To monitor this information, select the server where you want to monitor the available virtual networks, then click **Tasks** → **Configuration** → **Virtual Resources** → **Virtual Network Management** as shown in Figure 11-43.

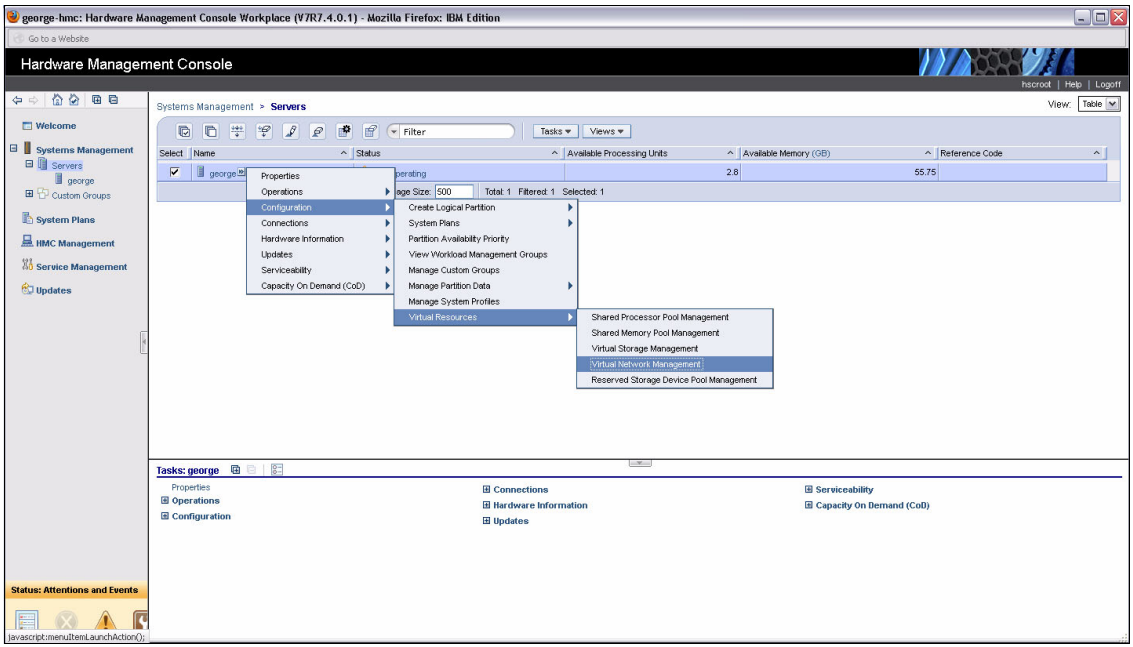


Figure 11-43 Virtual Network Management

A new window is displayed that provides information about the available virtual network topology. If you select a VLAN, you receive detailed information about the partitions assign to this VLAN as shown in Figure 11-44.

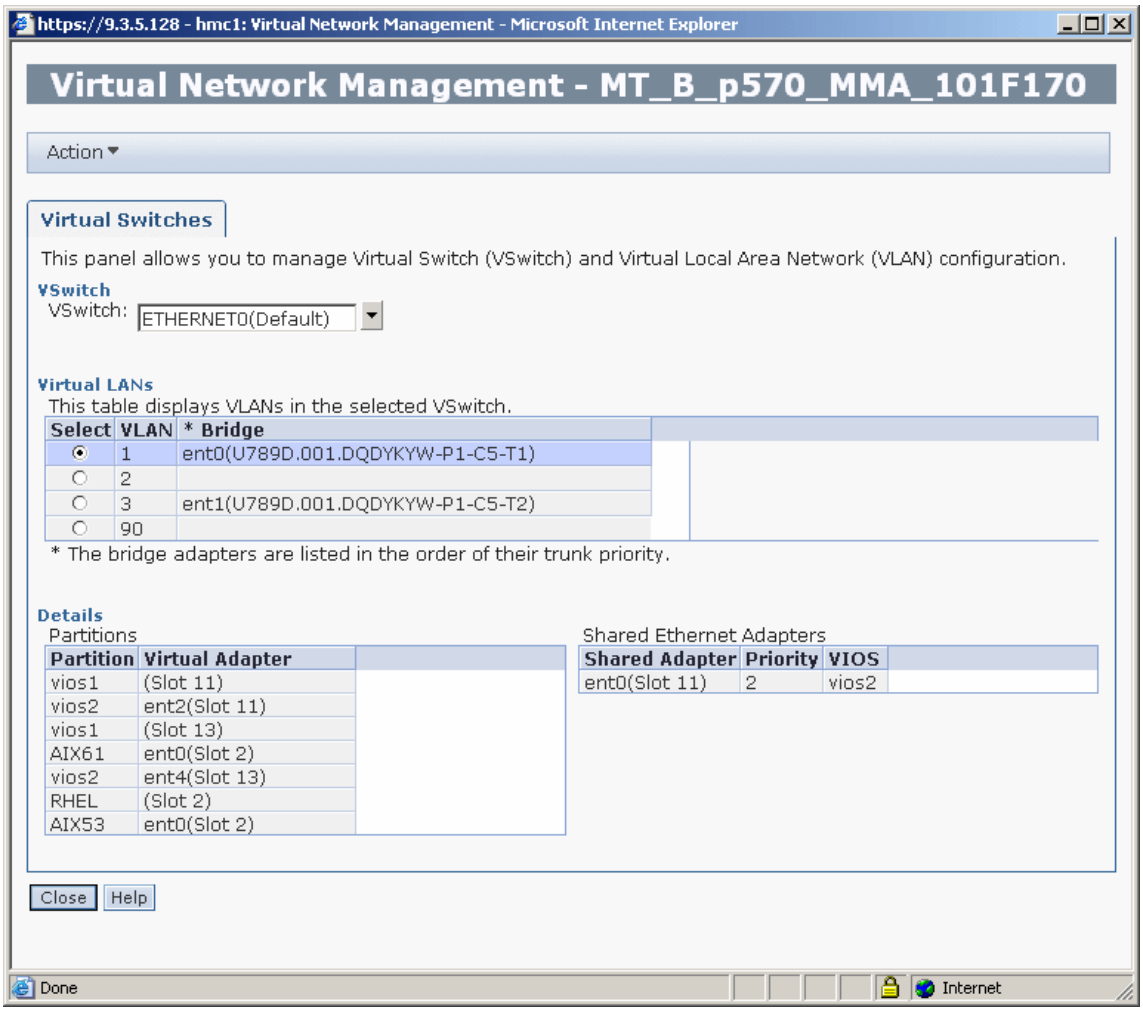


Figure 11-44 Virtual Network Management: Detailed information

HMC shell scripting

It is also possible to log on to the HMC by using the **ssh** command. You can then script the HMC commands to retrieve the information that is provided by the HMC interface.

For more information about the HMC V7, see the *Hardware Management Console V7 Handbook*, SG24-7491.

Integrated Virtualization Manager (IVM) monitoring

The IVM shows detailed partition information in the partitions management panel for servers that are not managed by an HMC.

Log on to the IVM web interface of your Virtual I/O Server. Access the login window by entering the IP address or the DNS name of your Virtual I/O Server in a web browser.

Click **View/Modify Partitions** in the left menu as shown in Figure 11-45. You see the amount of memory, the processing units that are allocated, and those that are available on the system.

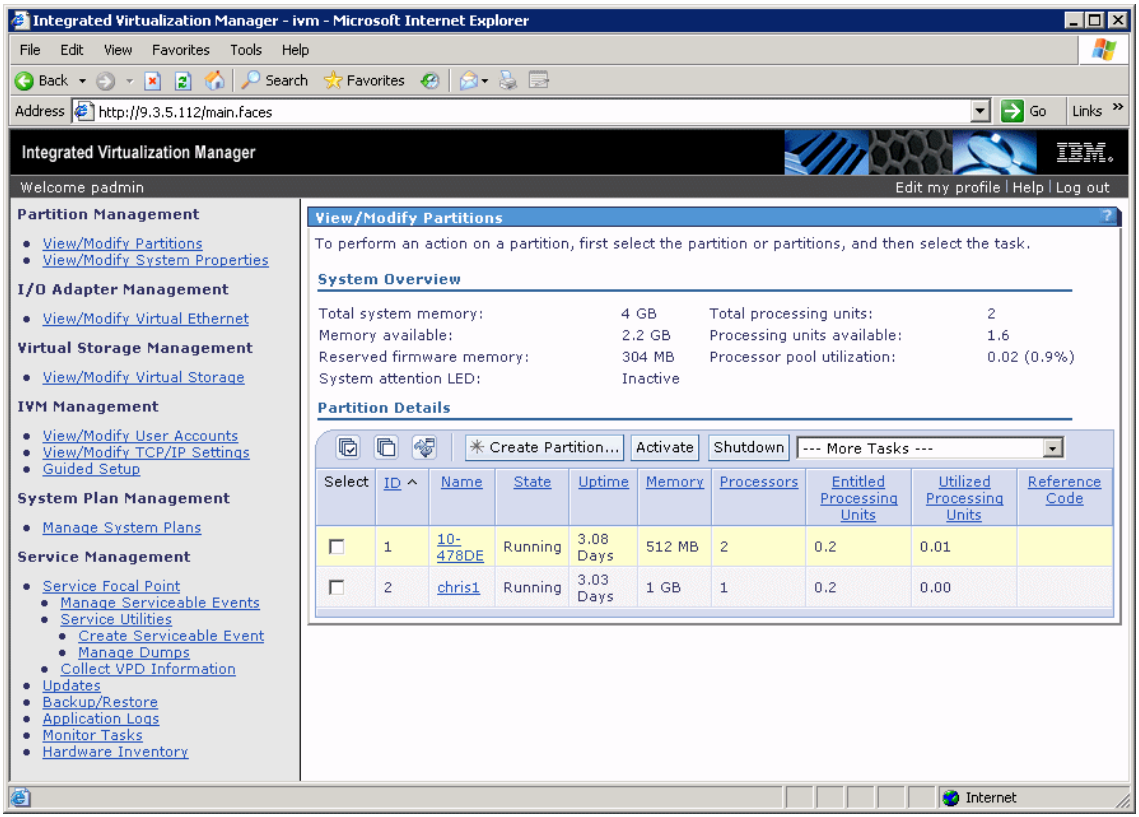


Figure 11-45 IVM partitions monitoring

You can also access the Virtual network configuration by clicking **View/Modify Virtual Ethernet** in the left pane.

Figure 11-46 illustrates the virtual Ethernet configuration monitoring.

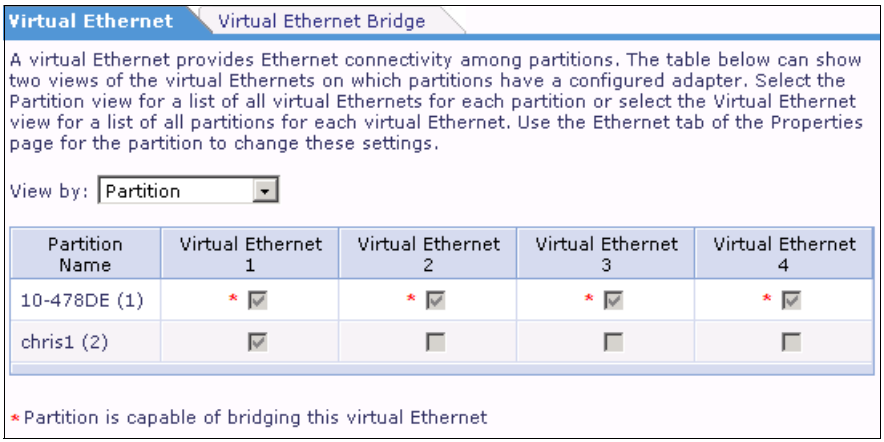


Figure 11-46 IVM virtual Ethernet configuration monitoring

Another useful configuration to monitor is the Virtual storage configuration. Click **View/Modify Virtual Storage** from the left menu.

Figure 11-47 illustrates the virtual Ethernet configuration monitoring.

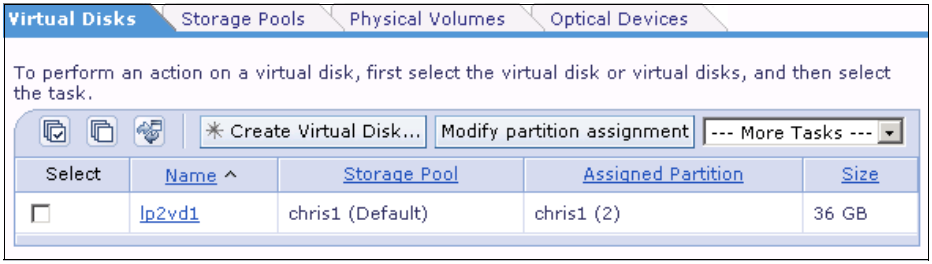


Figure 11-47 IVM virtual storage configuration monitoring

For more information about the IVM interface, see *Integrated Virtualization Manager on IBM System p5*, REDP-4061.

Monitoring resource allocations from a partition

When you are logged on a partition, you can use the command-line interface to display current partition and system-wide processor and memory resource allocations.

Commands are available on AIX or Linux partitions to display these resource allocations.

Tip: For IBM i system-wide resource allocation information, use the HMC interface.

Monitoring processor and memory allocations in AIX

In the AIX operating system, run the **`lparstat -i`** command. Example 11-40 illustrates the command output.

Example 11-40 lparstat -i command output on AIX

```
# lparstat -i
Node Name                                  : aix61
Partition Name                             : AIX61
Partition Number                           : 3
Type                                       : Shared-SMT
Mode                                       : Uncapped
Entitled Capacity                          : 0.50
Partition Group-ID                         : 32771
Shared Pool ID                             : 0
Online Virtual CPUs                        : 1
Maximum Virtual CPUs                       : 1
Minimum Virtual CPUs                       : 1
Online Memory                              : 2048 MB
Maximum Memory                             : 6016 MB
Minimum Memory                             : 256 MB
Variable Capacity Weight                   : 128
Minimum Capacity                           : 0.10
Maximum Capacity                           : 1.00
Capacity Increment                         : 0.01
Maximum Physical CPUs in system            : 16
Active Physical CPUs in system             : 4
Active CPUs in Pool                        : 4
Shared Physical CPUs in system             : 4
Maximum Capacity of Pool                   : 400
Entitled Capacity of Pool                  : 375
Unallocated Capacity                       : 0.00
Physical CPU Percentage                    : 50.00%
Unallocated Weight                         : 0
Memory Mode                                : Dedicated
Total I/O Memory Entitlement               : -
Variable Memory Capacity Weight            : -
Memory Pool ID                             : -
Physical Memory in the Pool                : -
Hypervisor Page Size                       : -
Unallocated Variable Memory Capacity Weight: -
Unallocated I/O Memory entitlement         : -
Memory Group ID of LPAR                    : -
```

Monitoring processor and memory allocations in Linux

Supported Linux distributions also provide a command-line interface to display the partition's processor and memory resource allocations. This interface uses the **/proc/ppc64/lparcfg** special device. The content is shown in Example 11-41.

Example 11-41 Listing partition resources on Linux

```
[root@VIOCRHEL52 ~]# cat /proc/ppc64/lparcfg
lparcfg 1.8
serial_number=IBM,02101F170
system_type=IBM,9117-MMA
partition_id=6
BoundThrds=1
CapInc=1
DisWheRotPer=5120000
MinEntCap=10
MinEntCapPerVP=10
MinMem=128
MinProcs=1
partition_max_entitled_capacity=100
system_potential_processors=16
DesEntCap=50
DesMem=2048
DesProcs=1
DesVarCapWt=128
DedDonMode=0

partition_entitled_capacity=50
group=32774
system_active_processors=4
pool=0
pool_capacity=400
pool_idle_time=0
pool_num_procs=0
unallocated_capacity_weight=0
capacity_weight=128
capped=0
unallocated_capacity=0
entitled_memory=2147483648
entitled_memory_group_number=32774
entitled_memory_pool_number=65535
entitled_memory_weight=0
unallocated_entitled_memory_weight=0
unallocated_io_mapping_entitlement=0
entitled_memory_loan_request=0
backing_memory=2147483648 bytes
cmo_enabled=0
purr=13825418184
```

```
partition_active_processors=1
partition_potential_processors=2
shared_processor_mode=1
```

11.2.3 Monitoring commands on the Virtual I/O Server

The Virtual I/O Server comes with several commands that can monitor its activity. Some look like the standard AIX commands, while others are a bit different. The command parameters are specific to the Virtual I/O Server system. It is therefore good practice to familiarize yourself with them. The Virtual I/O Server online help can be used with running the **help** command to display the available commands. Generally speaking, running the commands with the **-h** parameter displays the command syntax and provides enough information to correctly use the tool. If more information is required, **man** pages are available. You can also look at the online reference that can be found at:

http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/topic/iphcg/iphcg.pdf

This part presents the principal monitoring commands available on the Virtual I/O Server, along with practical usage examples.

This section includes the following topics:

- Global system monitoring
- Device inspection
- Storage monitoring and listing
- Shared storage pool monitoring
- Network monitoring

Global system monitoring

To get general system information, use the following commands:

topas	This command is similar to **topas** in AIX. It presents various system statistics such as processor, memory, network adapters, and disk usage.
sysstat	This command gives you an uptime for the system and a list of logged-on users.
svmon	This command captures and analyzes a snapshot of virtual memory.
vmstat	This command reports statistics about kernel threads, virtual memory, disks, traps, and processor activity.

wkldout This command provides post-processing of recordings that are made with **wkldagent**. The files are in the /home/ios/perf/wlm path.

lsgcl This command displays the contents of the global command log.

vasistat This command shows VASI device driver and device statistics (used for PowerVM Live Partition Mobility).

Certain commands are also useful for configuration inspection:

ioslevel This command gives the version of the Virtual I/O Server.

lssw This command lists the software that is installed.

lsfware This command displays microcode and firmware levels of the system, adapters, and devices.

lslparinfo This command displays the client partition number and name.

lssvc This command lists available agent names if the parameter is not given. It lists the agent's attributes and values if the agent's name is provided as a parameter.

oem_platform_level This command returns the operating system level of the OEM install and setup environment.

chlang This command is primarily for Japanese locales. Use this command to force messages on the left to be displayed in English. Without this option, messages during the boot sequence might be corrupted.

If you are not familiar with AIX monitoring and you want to know how much processor and memory your Virtual I/O Server is using, you can use the **topas** command as shown in Example 11-42 on page 446.

The processor statistics are on the left of the panel toward the top. User% shows the percentage of time the processor is running code in user mode. Kern% shows the percentage of time the processor is running code in system mode. Wait% shows the percentage that the processor is idle and waiting for I/O requests. Idle% shows the percentage that the processor is idle and waiting for work. In a shared uncapped processor environment, these values are relative to the entitlement capacity Entc, or physical processors consumed Physc. Entitlement capacity, Ent, and physical processors used, Physc, are displayed on the center of panel toward the top.

You can find the memory statistics in the middle of the right side of the panel. Real**,**MB shows the total amount of memory on the Virtual I/O Server. To see how much of that memory is used, look at the % Comp figure. It shows how much

memory is used as computational memory. Computational memory consists of pages that belong to working-storage segments or program text segments.

Example 11-42 Using topas to display processor and memory usage on the VIO

```
Topas Monitor for host:    P7_1_vios1           EVENTS/QUEUES    FILE/TTY
Tue Dec 14 08:57:27 2010   Interval:  2         Cswitch     604  Readch   375.6K
                                                Syscall    1239  Writech    2448
CPU  User%  Kern%  Wait%  Idle%  Physc   Entc  Reads       147  Rawin         0
ALL   60.5    1.3    0.0   38.2   0.98   97.5  Writes       25  Ttyout      333
                                                Forks         0  Igets         0
Network  KBPS   I-Pack  O-Pack   KB-In  KB-Out  Execs         0  Namei       175
Total     5.7     27.0    24.0     2.0     3.6  Runqueue    1.0  Dirblk        0
                                                Waitqueue   0.0
Disk    Busy%     KBPS     TPS KB-Read KB-Writ                  MEMORY
Total     1.0     81.0    17.0     9.0    72.0  PAGING           Real,MB    2048
                                                Faults        4  % Comp      68
FileSystem        KBPS     TPS KB-Read KB-Writ  Steals        0  % Noncomp   14
Total            375.8   144.0  374.6    1.2    PgspIn        0  % Client    14
                                                PgspOut       0
Name            PID  CPU%  PgSp Owner           PageIn        2  PAGING SPACE
padmin      9371864  60.0   0.1 root            PageOut       2  Size,MB    1536
vmmd         458766   0.5   1.2 root            Sios          4  % Used      1
cld         5767348   0.4   2.0 root                             % Free     99
topas       8782026   0.2   1.8 padmin          NFS (calls/sec)
xmgc         851994   0.1   0.4 root            SerV2         0  WPAR Activ    0
java        4849856   0.1  89.3 root            CliV2         0  WPAR Total    0
solidhac    8454180   0.1  31.7 root            SerV3         0  Press: "h"-help
cimserve    7143674   0.1  56.8 root            CliV3         0         "q"-quit
```

Device inspection

Certain commands provide device configuration information:

lsdev This command displays devices in the system and their characteristics.

lsmap This command displays the mapping between physical and virtual devices.

viosbr This command backups the virtual and logical configuration including the shared storage pool database, and displays and restores the configuration from a previous backup.

Storage monitoring and listing

Certain monitoring commands report various storage activities:

viostat	This command reports Central Processing Unit statistics, asynchronous input/output, input/output statistics for entire system, adapters, tty devices, disks, and CD-ROMs. The parameter **-extdisk** provides detailed performance statistics information for disk devices.
nmon	This command displays local system statistics such as system resources and processor usage. Users can use interactive or recording mode.
fcstat	This command displays statistics that are gathered by the specified Fibre Channel device driver.
lsvg	This command displays information about volume groups.
lslv	This command displays information about logical volumes.
lspv	This command displays information about physical volumes.
alert	This command sets, removes, and lists all the alerts for a cluster and storage pool.
lsvopt	This command lists and displays information about the system's virtual optical devices.
lsrep	This command lists and displays information about the Virtual Media Repository.
lspath	This command displays information about paths to MPIO-capable devices.

Shared storage pool monitoring

To monitor the shared storage spool, new commands such as **cluster** or **lscluster** are available and existing commands like **lssp**, **lsmap**, or **lspv** have new options.

Cluster information commands

cluster	This command allows operations on a cluster such as creation and removal, and also shows cluster information.
lscluster	This command lists the cluster configuration information.

Pool information commands

lssp	Lists and displays information about storage pools.
lspv	Lists physical volumes in a shared storage pool when used with **-clustername** option.
lsmap	Displays the mapping that is related to shared storage pools when used with **-clustername** option.

Network monitoring

Certain monitoring commands can be used for network monitoring:

netstat	This command displays active sockets for each protocol, routing table information, or displays the contents of a network data structure.
entstat	This command displays Ethernet device driver and device statistics.
seastat	This command generates a report to view, per client, the Shared Ethernet Adapter statistics.

Extra commands report network configuration information:

hostname	This command sets or displays the name of the current host system.
lsnetsvc	This command gives the status of a network service.
lstcpip	This command displays the TCP/IP settings.
optimizenet -list	This command lists the characteristics of one or all network tunables.
snmp_info -get|-next	This command requests the values of Management Information Base variables that are managed by a Simple Network Management Protocol agent.
traceroute	This command prints the route that IP packets take to a network host.
showmount	This command displays a list of all clients that have remotely mountable file systems.

11.2.4 Third-party monitoring tools

This section addresses monitoring tools for AIX and Linux partitions on Power Systems. Linux does not have all the monitoring commands that are available for AIX or PowerVM Virtualization. The `/proc/ppc64/lparcfg` special device provides significant information about the logical partition.

This information is used by the third-party tools that are addressed in this chapter, which can be used on both AIX and Linux operating systems. These tools are not IBM products. They are supported by their developers and user communities.

This section includes the following topics:

- The nmon utility
- sysstat utility
- Ganglia tool
- Other monitoring tools

The nmon utility

The **nmon** utility is a downloadable monitoring tool for AIX and Linux available for no extra fee. This tool provides a text-based summary, similar to the **topas** command, of key system metrics. Useful for both online immediate monitoring and for offline monitoring by saving the data file for later analysis, the **nmon** analyzer provides you with an easy way to transform the saved file data into graphs.

Tip: The **nmon** utility is included in AIX 6.1 TL2 and later releases, and in Virtual I/O Server version 2.1 and later releases. The information that is provided by **nmon** might be used if the system is running earlier versions of AIX or Virtual I/O Server.

To use the **nmon** utility at the shell prompt, run the command **topas** → ~ (tilde).

As of Version 11, the **nmon** command is simultaneous multithreading and partition-aware. **nmon** Version 12 now integrates processor donation statistics.

The **nmon** command is available at:

http://www.ibm.com/developerworks/wikis/display/WikiPtype/nmon

Extract and copy the **nmon** binary file to the partition you want to monitor, typically under /usr/bin/nmon. Optionally, you can change and verify the **iostat** flags to continuously maintain the disk I/O history by using the following commands:

```
# lsattr -E -l sys0 -a iostat
iostat false Continuously maintain DISK I/O history True
# chdev -l sys0 -a iostat=true
sys0 changed
# lsattr -E -l sys0 -a iostat
iostat true Continuously maintain DISK I/O history True
```

You might receive a warning if you do not set this flag to true. However, **nmon** continues to show non-null values for disk usage.

Important: If you have many disks (more than 200), setting `iostat` to `true` starts consuming processor time (around 2%). After the measurement campaign is completed, set the flag back to `false`.

Run the **nmon** command and then press P to show partition statistics. The result differs between the AIX and Linux systems.

nmon on AIX

On AIX 6.1 systems and later, **nmon** provides statistics that include nearly all that **topas** provides. It includes the following statistics and more:

- Processor
- Memory
- Paging
- Network
- Disks
- Logical volumes
- File systems
- NFS
- Async I/O
- Fibre Channel adapters
- SEA
- Kernel numbers
- Multiple page size stats
- WLM
- WPARs
- Top processes

Example 11-43 shows processor utilization and top processes.

Example 11-43 nmon output

```
··topas_nmon··= KB<-->MB·····Host=nimres1·····Refresh=2 secs···17:55.06··
· CPU-Utilisation-Small-View ·····················································
·                          0----------25----------50----------75----------100·
·CPU User%  Sys% Wait% Idle%|           |           |           |           |·
·  0   0.0   0.0   0.0 100.0|    >                                          |·
·  1   0.0   0.0   0.0 100.0|>                                              |·
·  2   0.0   0.0   0.0 100.0|       >                                       |·
·  3   0.0   0.0   0.0 100.0|>                                              |·
·Physical Averages          +-----------|-----------|-----------|-----------+·
·All   0.0   0.1   0.0  99.8| >                                             |·
·                          +-----------|-----------|-----------|-----------+·
· Top-Processes-(110) ·····Mode=3  [1=Basic 2=CPU 3=Perf 4=Size 5=I/O 6=Cmds]···
·  PID     %CPU  Size   Res   Res   Res  Char RAM      Paging       Command    ·
```

```
·          Used    KB   Set  Text  Data   I/O  Use io other repage            ·
·10027162   0.0  5040  4968   580  4388     0   0%    0     0     0 topas_nmon   ·
·  393228   0.0   120   120     0   120     0   0%    0     0     0 vmmd         ·
·  458766   0.0    56    56     0    56     0   0%    0     0     0 memgrdd      ·
·  524364   0.0   176   200    60   140     0   0%    0     0     0 shlap64      ·
·  589842   0.0    76    76     0    76     0   0%    0     0     0 vtiol        ·
·  655380   0.0    56    56     0    56     0   0%    0     0     0 devstatd     ·
·  720918   0.0    88    88     0    88     0   0%    0     0     0 pilegc       ·
·  262152   0.0    72    72     0    72     0   0%    0     0     0 lrud         ·
·····························································································
```

nmon on Linux

The **nmon** tool is available for IBM Power Systems running Red Hat or Novell SUSE Linux distributions. A comprehensive explanation about the usage of **nmon**, including the source files, can be found at:

http://www.ibm.com/developerworks/wikis/display/WikiPtype/nmon

On Linux systems, the number of PowerVM Virtualization-related metrics is restricted. **nmon** therefore shows fewer statistics in the logical partition section than on AIX 6.1 systems as shown in Figure 11-48.

```
 LPAR Stats
LPAR=5  SerialNumber=IBM,021O1F170  Type=IBM,9117-MMA
Flags:      Shared-CPU=true   Capped=false
Systems CPU Pool=  400.00          Active=    4.00    Total=   16.00
LPARs CPU    Min=    0.10     Entitlement=    0.80      Max=    4.00
Virtual CPU  Min=    1.00          VP Now=    1.00      Max=    4.00
Memory       Min= unknown             Now=  256.00      Max= 2048.00
Other     Weight=  128.00   UnallocWeight=    0.00 Capacity=    0.01
      BoundThrds=    1.00 UnallocCapacity=    0.00  Increment
Physical CPU use=   0.016                              [timebase=512000000]
```

Figure 11-48 The nmon LPAR statistics report for a Linux partition

Extra nmon statistics

The **nmon** command can also display other kinds of statistics, such as those related to disk, network, memory, and adapters. See the **nmon** documentation for more information about these topics. You can also press H while **nmon** is running to get a help summary, or use **nmon -h** for more information about specific options.

Recording with the nmon tool

You can record resource usage using **nmon** for subsequent analysis with the **nmon** analyzer tool, or other post-capture tools. This works on both standard AIX and Linux partitions:

```
# nmon -f -t [-s <seconds> -c <count>]
```

The **nmon** process runs in the background, and you can log off the partition if you want. For best results, do not have a count greater than 1,500. The command creates a file with a name in the following format:

```
<hostname>_<date>_<time>.nmon
```

After the recording process completes, transfer the file to a system that runs Microsoft Excel spreadsheet software to run the **nmon** analyzer tool.

You can find the nmon_analyser tool at:

http://www.ibm.com/developerworks/wikis/display/WikiPtype/nmonanalyser

sysstat utility

The **sysstat** package includes at least three groups of monitoring tools for Linux. This utility might be included in the Linux distributions. Users can also download the current version of this utility at:

http://sebastien.godard.pagesperso-orange.fr/features.html

The tools that are included are **sadc**, **sar**, **iostat**, **sadf**, **mpstat**, and **pidstat**. These tools can be used for obtaining various metrics on the host partition, such as processor statistics, memory, paging, swap space, interrupts, network activity, and task switching activity.

In addition to this wide array of system resource monitoring, the **sysstat** tool also offers the following benefits:

- Output can be saved to a file for analysis.
- Averages can be calculated over the sampling period.
- You can specify the duration of data collection.
- There is support for hotplugging in certain environments.
- There is support for 32-bit and 64-bit architectures.

Ganglia tool

Ganglia is a monitoring system that was initially designed for large, high performance computing clusters and grids. It uses lightweight agents and can

use multicast communication to save computing and network resources. Ganglia is available at:

http://ganglia.sourceforge.net/

With additional metrics added, Ganglia can visualize performance data that is specific to a virtualized Power Systems environment. For Power Systems monitoring, you can design your Power Systems server as a cluster and treat all the client partitions on your server as nodes in the same cluster. This way, the visualization allows you to see the summarized overall server load.

Ganglia can show all general processor use on a server and the amount of physical processor that is used by each partition in the last hour. This includes AIX, Virtual I/O Server, and Linux partitions. From these graphs, you can determine which partitions are using the shared processors most.

Ganglia, for example, can record shared processor pool for a whole week to determine which partitions are using processor cycles and when they are using them. Some workloads can be seen as constrained while others only run for a certain time each day.

Adapted Ganglia packages with additional Power Systems metrics for AIX and Linux and instructions for best practices are provided at:

http://www.perzl.org/ganglia/

General rules for the use of Ganglia can be found at:

http://www.ibm.com/developerworks/wikis/display/WikiPtype/ganglia

Lpar2rrd

Lpar2rrd is a tool capable of producing historical processor utilization graphs of logical partitions and shared processor usage.

It also collects complete physical (hardware) and logical configuration of all managed systems and logical partitions. This includes all changes in their state and configuration. This tool is not intended to be real-time monitoring. It is used only for HMC-based micro-partitioned systems with a shared processor pool. It creates charts that are based on utilization data that are collected on HMC by running the **lslparutil** command. It is agent-less, so you do not need to install any software on logical partitions. It uses ssh key based access to get all data from HMC. It supports operating systems AIX, Virtual I/O Server, Linux, and IBM i.

This tool is simple to install, configure, and use. Default graphs can provide up to a year of historical data if available at the HMC.

More information can be found here:

http://www.lpar2rrd.com

11.2.5 Other monitoring tools

You can find more tools and useful information on the Internet at the following links:

- Virtual I/O Server Monitoring wiki:

  http://www.ibm.com/developerworks/wikis/display/WikiPtype/VIOS_Monitoring

- Performance monitoring with non-AIX tools wiki:

  http://www.ibm.com/developerworks/wikis/display/WikiPtype/Other+Performance+Tools

- Partitioning monitoring using the **lparmon** command:

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

- Power Systems and virtualized environments can be monitored by Performance toolbox (PTX):

  http://www-03.ibm.com/systems/power/software/aix/ptx/perftoolbox.html

- Easy system monitoring using **sar**:

  http://www.ibm.com/developerworks/aix/library/au-unix-perfmonsar.html

- Performance counters for Linux using **perf**:

  https://perf.wiki.kernel.org/index.php/Main_Page

Part 5

Managed systems virtualization

This part describes the additional features of PowerVM related to managed systems virtualization.

This part includes the following chapters:

- Dynamic logical partitioning
- Partition Suspend and Resume
- Live Partition Mobility
- Dynamic Platform Optimizer
- Active System Optimizer and Dynamic System Optimizer for AIX

12

Dynamic logical partitioning

This section describes how to change resources dynamically, which can be useful to maintain a virtualized environment. It also addresses how to monitor the resources after dynamic LPAR operations. The focus is on the following operations and the corresponding resource monitoring valid for the Virtual I/O Server and for AIX, IBM i, and Linux operating systems:

- Addition of resources
- Movement of adapters between partitions
- Removal of resources
- Replacement of resource

This chapter includes the following sections:

- Managing dynamic LPAR operations
- Monitoring dynamic LPAR operations

12.1 Managing dynamic LPAR operations

This section includes the following sections:

12.1.1 Dynamic LPAR operations on AIX and IBM i

The following sections explain how to perform dynamic LPAR operations for AIX and IBM i.

Considerations:

- When using IBM i 6.1 dynamic LPAR operations with *virtual* adapters, make sure that SLIC PTFs MF45568 and MF45473 are applied or you are at cumulative PTF level C9111610 or later.
- For dynamic LPAR operations with AIX and Linux, the Hardware Management Console (HMC) communicates with partitions by using Resource Monitoring and Control (RMC). Therefore, you must ensure that the RMC port has not been restricted in firewall settings. For dynamic LPAR operations with IBM i, the partition directly communicates with the POWER hypervisor rather than using an RMC connection.

Adding and removing processors dynamically

Follow these steps to add or remove processors dynamically:

1. Select the logical partition where you want to initiate a dynamic LPAR operation, then select **Dynamic Logical Partitioning** → **Processor** → **Add or Remove** as shown in Figure 12-1.

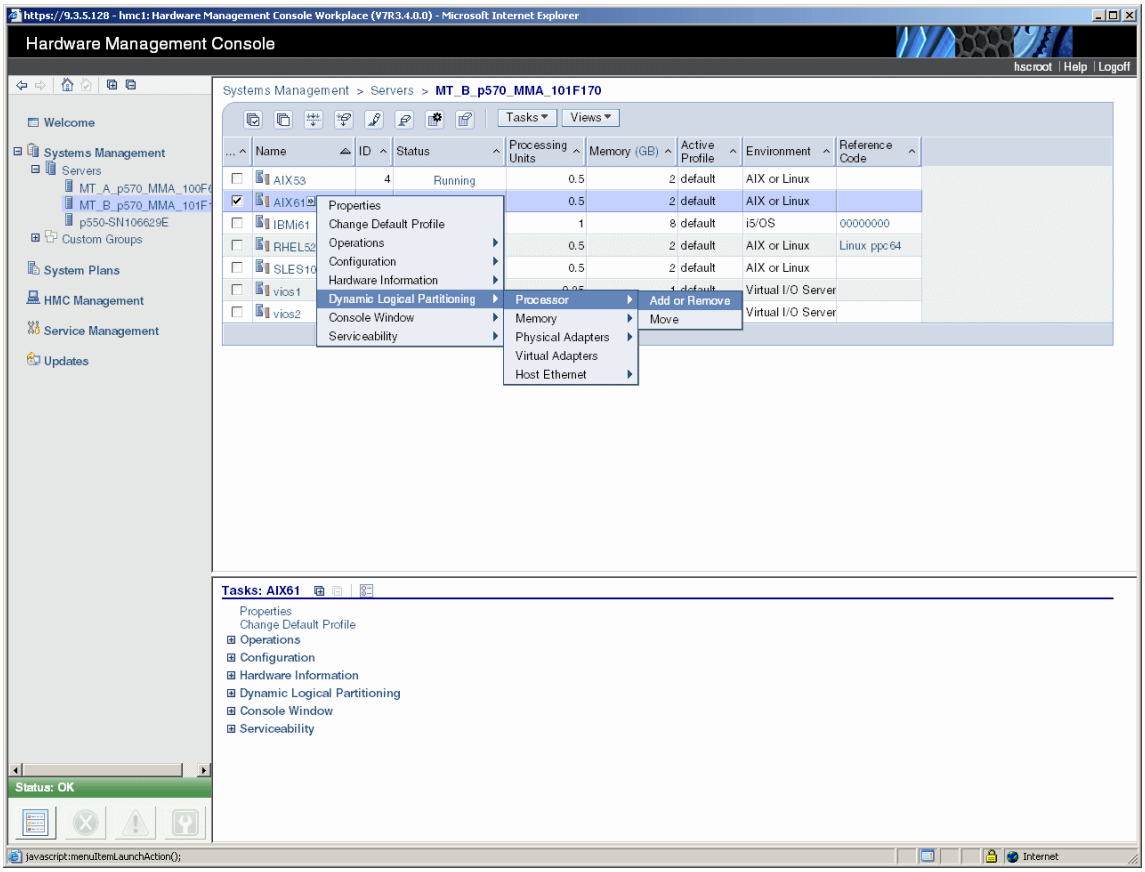


Figure 12-1 Add or remove processor operation

2. On HMC Version 7, you do not have to define a certain number of processors to be removed or added to the partition. Simply indicate the total number of processor units to be assigned to the partition. You can change processing units and the virtual processors of the partition to be more or less than the

current value. The values for these fields must be between the minimum and maximum values that are defined for them on the partition profile.

Figure 12-2 shows a partition being set with 0.5 processing units and one virtual processor.

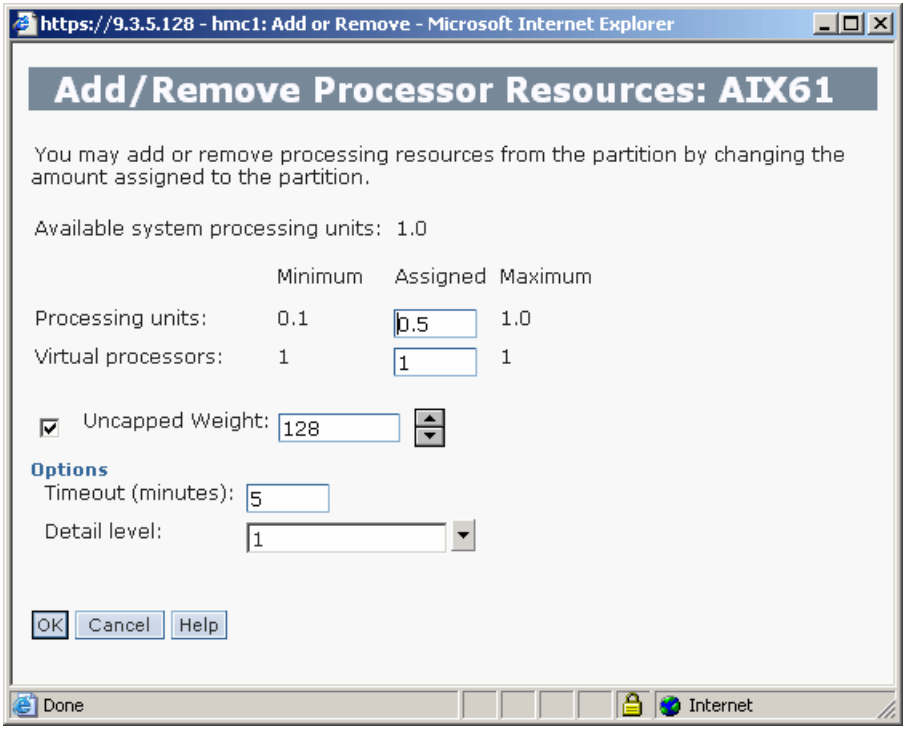


Figure 12-2 Defining the number of processing units for a partition

3. Click **OK** when done.

> **Tip:** In this example, a partition that employs Micro-partition technology was used. However, this process is also valid for dedicated processor partitions where you move dedicated processors.

Adding memory dynamically

To dynamically add more memory to the logical partition as shown in Figure 12-3, complete these steps:

1. Select the partition and then select **Dynamic Logical Partitioning** → **Memory** → **Add or Remove**.

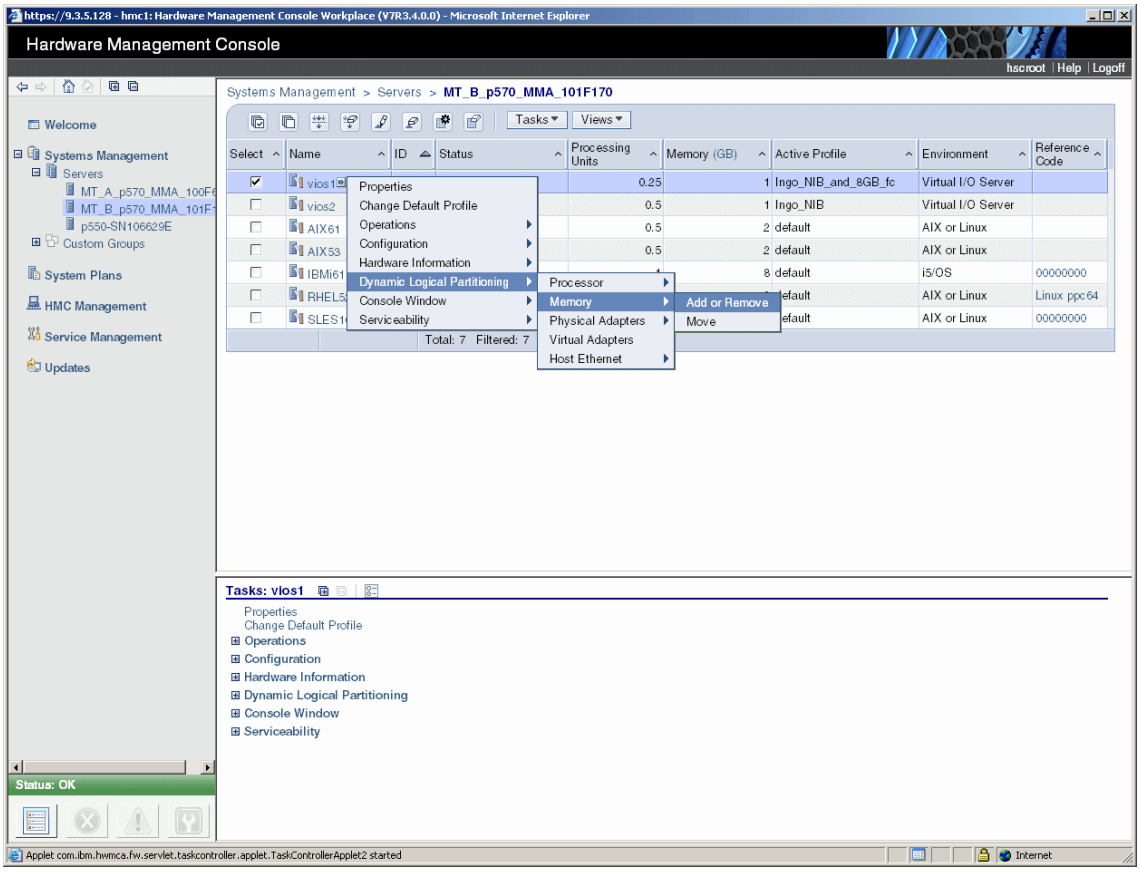


Figure 12-3 Add or remove memory operation

2. Change the total amount of memory to be assigned to the partition. On HMC Version 7, you do not provide the amount of extra memory that you want to add to the partition, but the total amount of memory that you want to assign to the partition. In Figure 12-4 the total amount of memory that is allocated to the partition was changed to 5 GB.

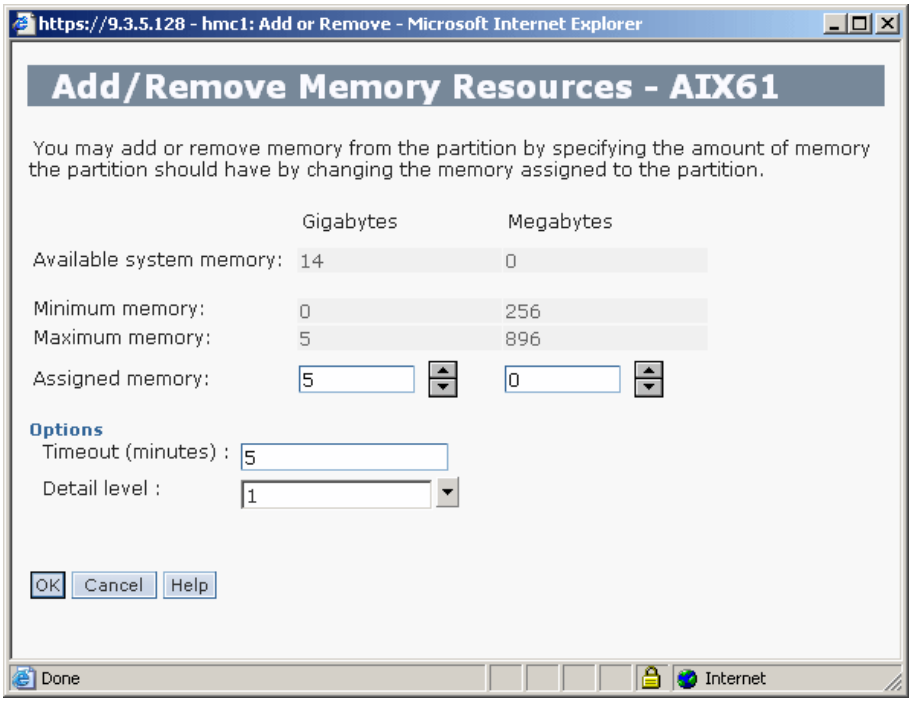


Figure 12-4 Changing the total amount of memory of the partition to 5 GB

3. Click **OK** when you are done. A status window as shown in Figure 12-5 displays.

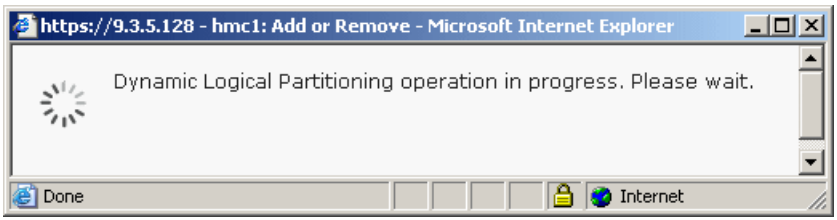


Figure 12-5 Dynamic LPAR operation in progress

Removing memory dynamically

Use the following steps for the dynamic removal of memory from a logical partition:

1. Select the logical partition where you want to initiate a dynamic LPAR operation. The first window in any dynamic operation is similar to Figure 12-6.

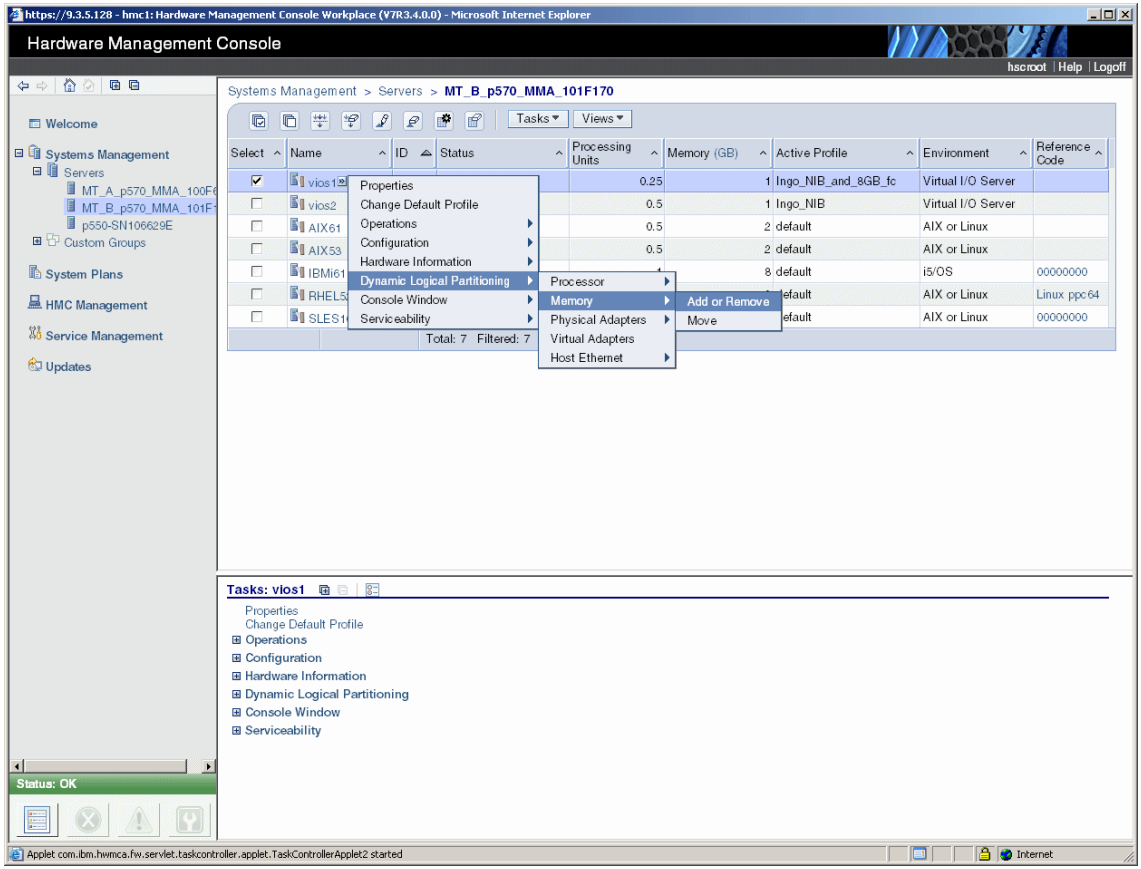


Figure 12-6 Add or remove memory operation

The graphical user interface to change the memory that is allocated to a partition is the same one used to add memory in Figure 12-4 on page 462. On HMC Version 7, you do not select the amount to remove from the partition as you did in the previous versions of HMC. Instead, you change the total amount of memory to be assigned to the partition. In the command output that is shown, the partition has 5 GB and you want to remove, for example,

1 GB from it. To do so, change the total amount of memory to 4 GB, as shown in Figure 12-7.

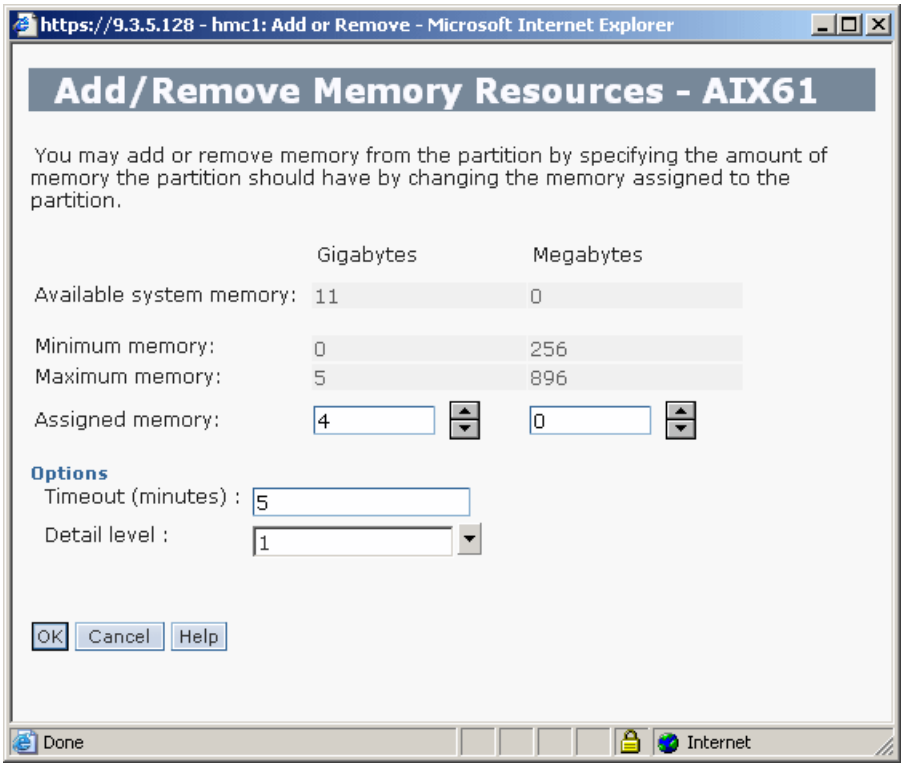


Figure 12-7 Dynamically reducing memory in a partition by 1 GB

2. Click **OK** when done.

Adding physical adapters dynamically

Restriction: Physical adapters cannot be added to the following types of partitions because they are not allowed to own physical adapters:

- Uses Active Memory Sharing (AMS)
- Suspend capable
- An IBM i restricted I/O partition

Complete these steps to add physical adapters dynamically:

1. Log in to HMC and select the system-managed name. On the right, select the partition where you want to run a dynamic LPAR operation, as shown in Figure 12-8.

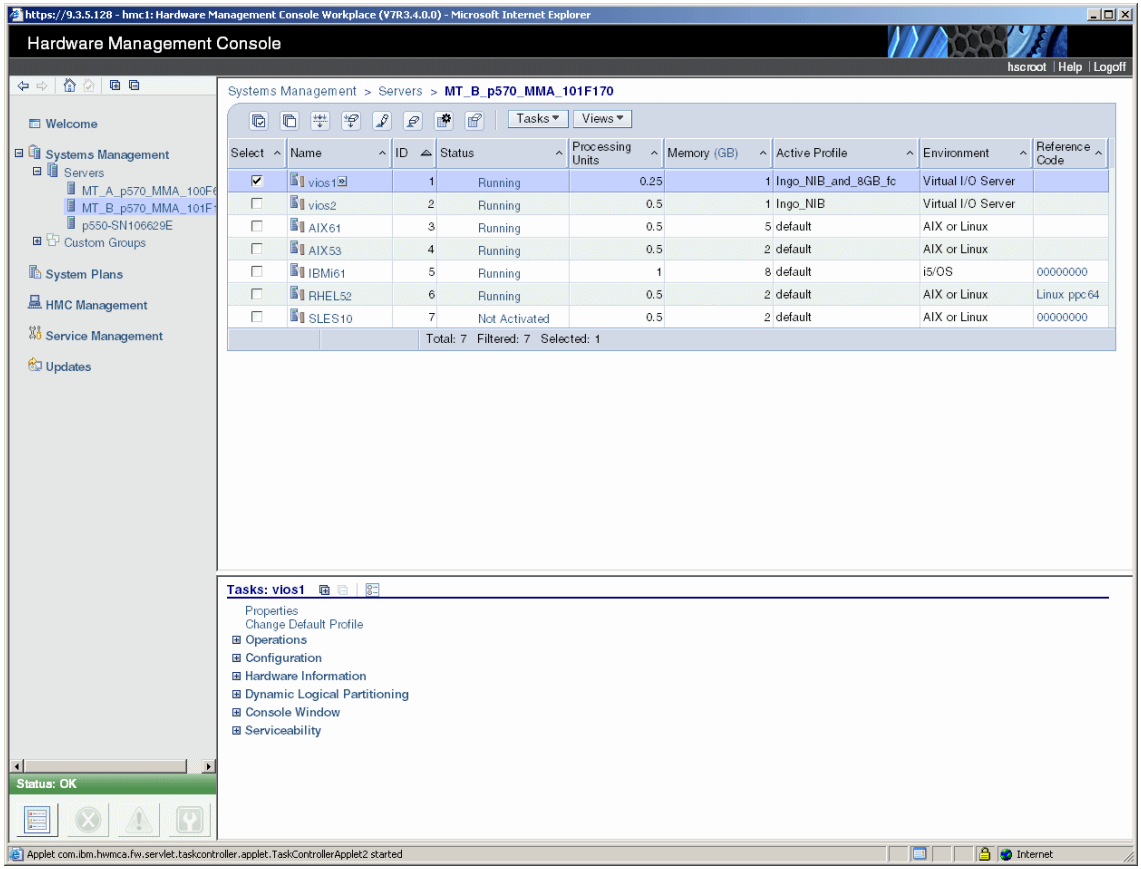


Figure 12-8 LPAR overview menu

2. On the **Tasks** menu on the right side of the window, select **Dynamic Logical Partitioning** → **Physical Adapters** → **Add** as shown in Figure 12-9.

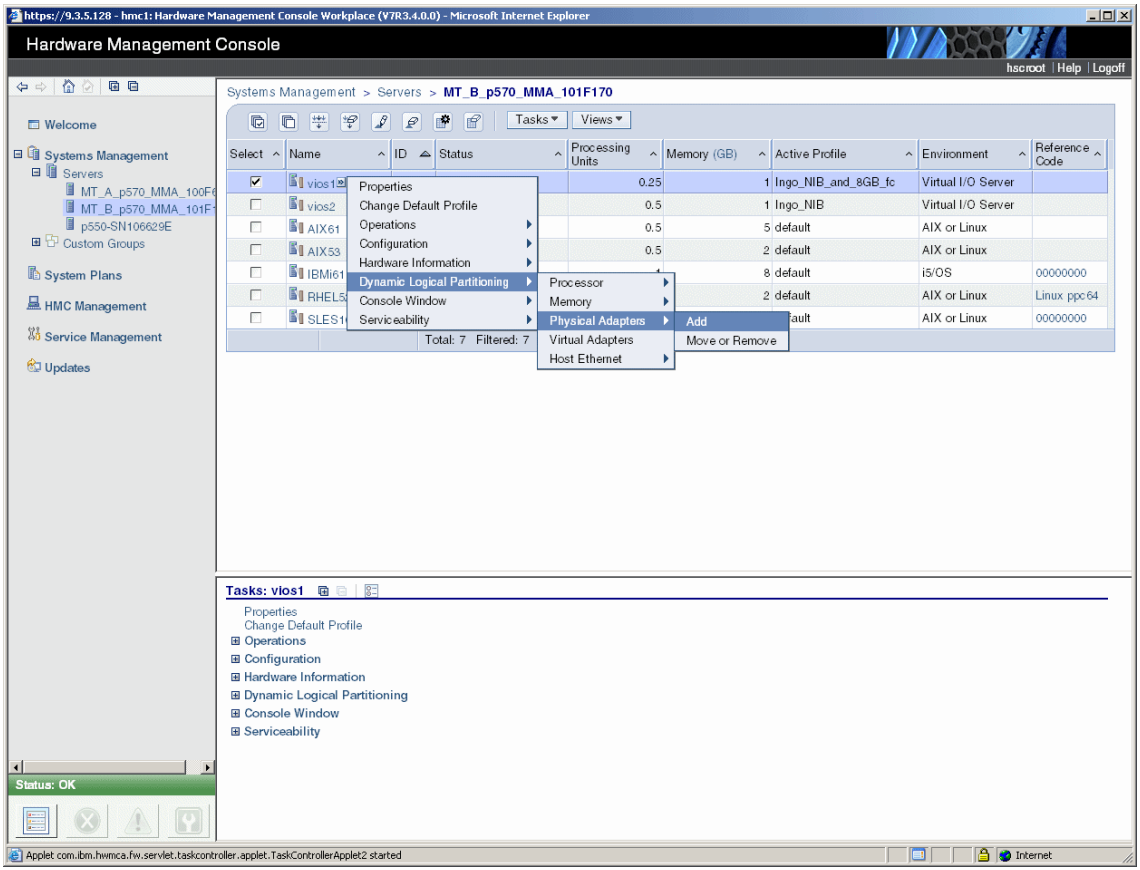


Figure 12-9 Add physical adapter operation

3. The next window will look like the one in Figure 12-10. Select the physical adapter that you want to add to the partition.

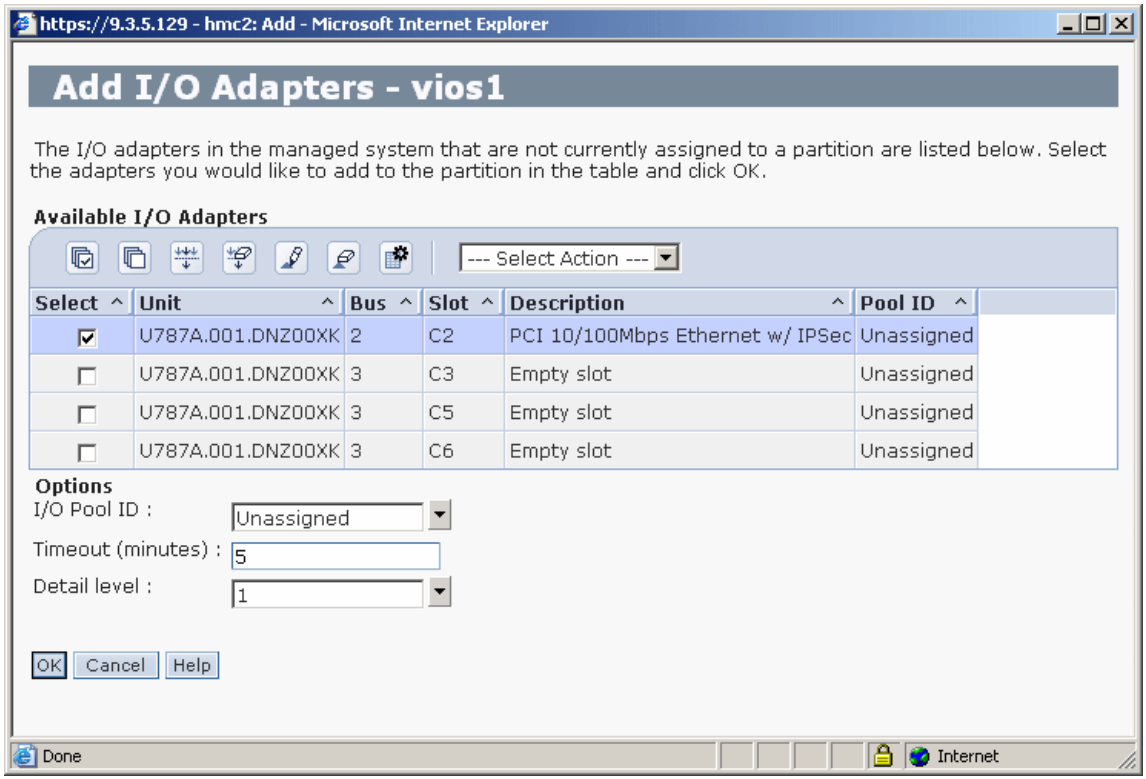


Figure 12-10 Selecting physical adapter to be added

4. Click **OK** when done.

Preparing the move or removal of adapters with dynamic LPAR

To move or remove a physical adapter, you first must release its usage in the partition that currently owns it:

1. Use the HMC to list which partition owns the adapter. In the left menu, select **Systems Management** and then click the system's name.
2. In the right menu, select **Properties**.

3. Select the **I/O** tab on the window that is displayed, as shown in Figure 12-11. You can see each I/O adapter for each partition.

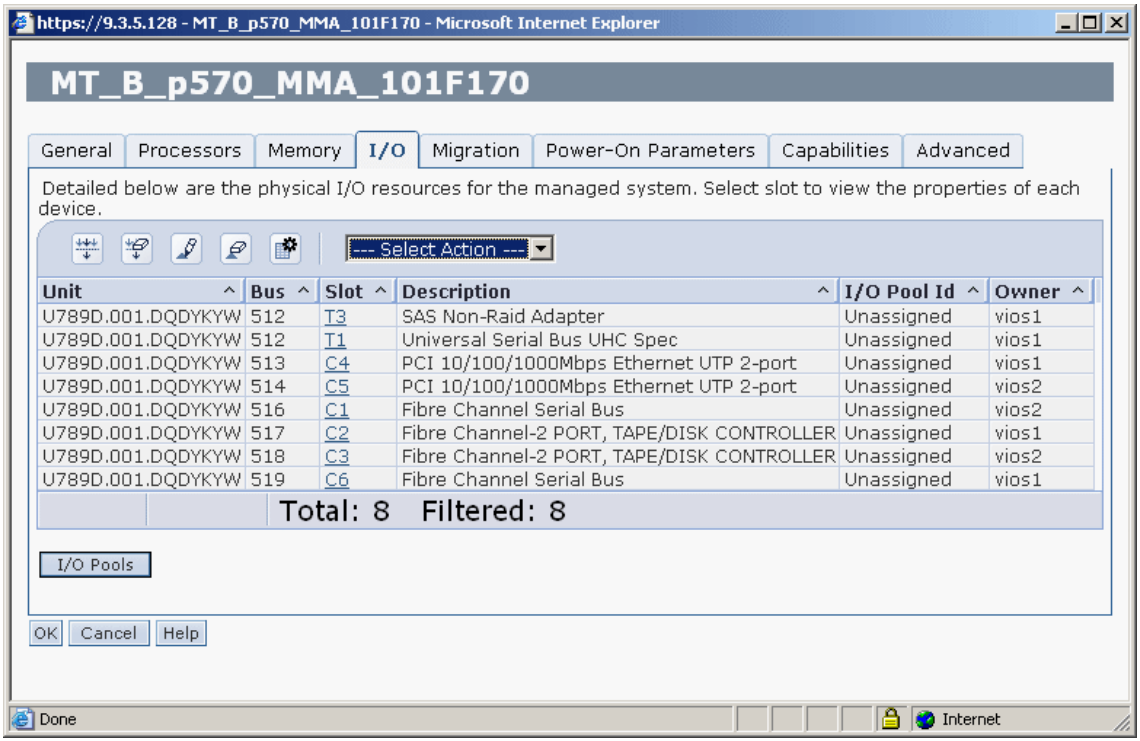


Figure 12-11 I/O adapter properties for a managed system

Remove devices that belong to the adapter, such as optical drives, as well.

The optical drive often must be moved to another partition. For an AIX partition, use the **`lsslot -c slot`** command as `root` user to list adapters and their members. In the Virtual I/O Server, use the **`lsdev -slots`** command as `padmin` user as follows:

```
$ lsdev -slots
# Slot                     Description       Device(s)
U789D.001.DQDYKYW-P1-T1    Logical I/O Slot  pci4 usbhc0 usbhc1
U789D.001.DQDYKYW-P1-T3    Logical I/O Slot  pci3 sissas0
U9117.MMA.101F170-V1-C0    Virtual I/O Slot  vsa0
U9117.MMA.101F170-V1-C2    Virtual I/O Slot  vasi0
U9117.MMA.101F170-V1-C11   Virtual I/O Slot  ent2
U9117.MMA.101F170-V1-C12   Virtual I/O Slot  ent3
U9117.MMA.101F170-V1-C13   Virtual I/O Slot  ent4
U9117.MMA.101F170-V1-C14   Virtual I/O Slot  ent6
U9117.MMA.101F170-V1-C21   Virtual I/O Slot  vhost0
```

```
U9117.MMA.101F170-V1-C22  Virtual I/O Slot  vhost1
U9117.MMA.101F170-V1-C23  Virtual I/O Slot  vhost2
U9117.MMA.101F170-V1-C24  Virtual I/O Slot  vhost3
U9117.MMA.101F170-V1-C25  Virtual I/O Slot  vhost4
U9117.MMA.101F170-V1-C50  Virtual I/O Slot  vhost5
U9117.MMA.101F170-V1-C60  Virtual I/O Slot  vhost6
```

For an AIX partition, use the **rmdev -l pi*X* -d -R** command to remove the adapter from the configuration. That is, release it to be able to move it to another partition. In the Virtual I/O Server, use the **rmdev -dev pci*X* -recursive** command, where ***X*** is the adapter number.

Example 12-1 shows how to remove a Fibre Channel adapter from an AIX partition that was virtualized and does not need this adapter any more.

Example 12-1 Removing the Fibre Channel adapter

```
# lsslot -c pci
# Slot                   Description                          Device(s)
U789D.001.DQDYKYW-P1-C2  PCI-E capable, Rev 1 slot with 8x lanes  fcs0 fcs1
U789D.001.DQDYKYW-P1-C4  PCI-X capable, 64 bit, 266MHz slot       ent0 ent1
# rmdev -dl fcs0 -R
fcnet0 deleted
fscsi0 deleted
fcs0 deleted
# rmdev -dl fcs1 -R
fcnet1 deleted
fscsi1 deleted
fcs1 deleted
```

For an IBM i partition, vary off any devices using the physical adapter before you remove it or move it to another partition. Use a VRYCFG command such as VRYCFG CFGOBJ(TAP02) CFGTYPE(*DEV) STATUS(*OFF) to release the tape drive from the physical adapter. To see which devices are attached to which adapter, use a WRKHDWRSC command such as WRKHDWRSC *STG for storage devices. Select option `7=Display resource detail` for an adapter resource to see its physical location (slot) information. Select option **9**=`Work with resources` to list the devices that are attached to it.

Moving physical adapters dynamically

After the adapter resource is released in the partition that owns it, the physical adapter can be moved to another partition with the HMC by completing the following steps:

1. Select the partition that currently holds the adapter and then select **Dynamic Logical Partitioning** → **Physical Adapters** → **Move or Remove** (Figure 12-12).

   The adapter must not be set as `required` in the profile. To change the setting from `required` to `desired`, you must update the profile.

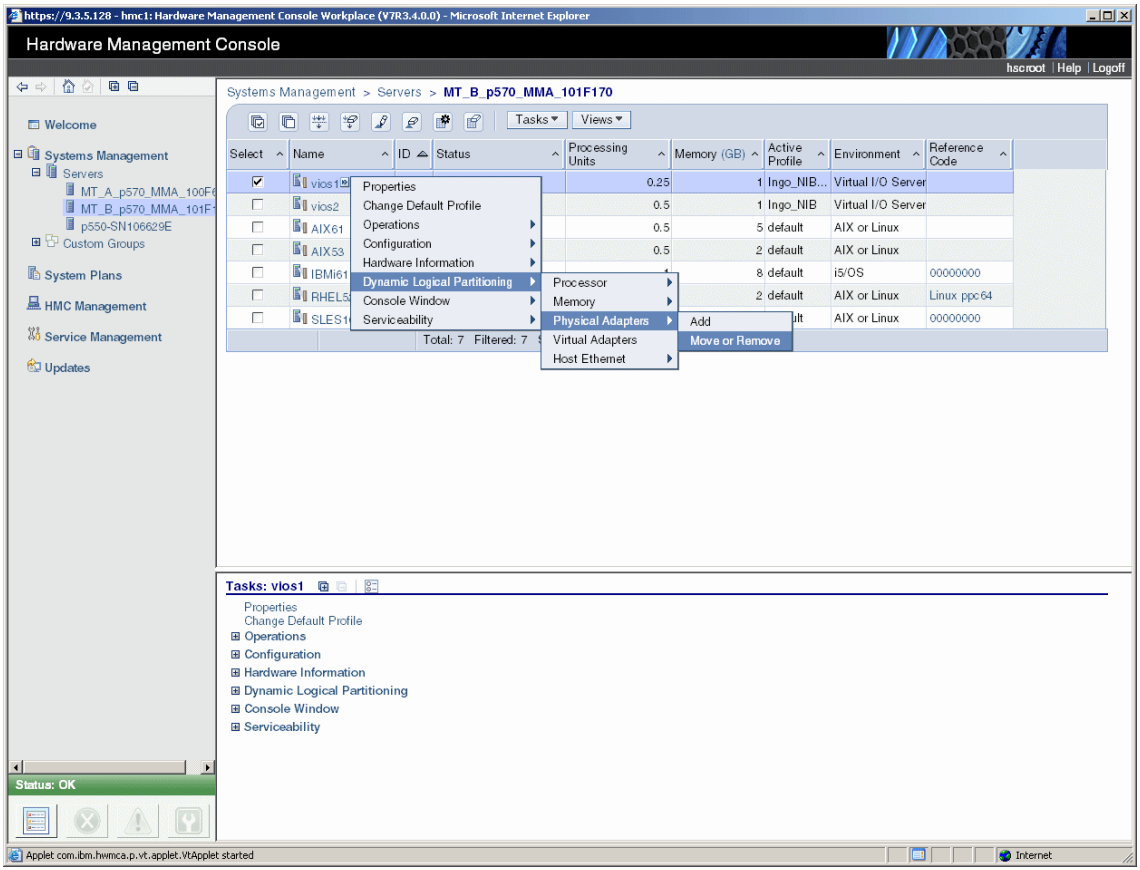


Figure 12-12 Move or remove physical adapter operation

2. Select the adapter to be moved and select the receiving partition as shown in Figure 12-13.

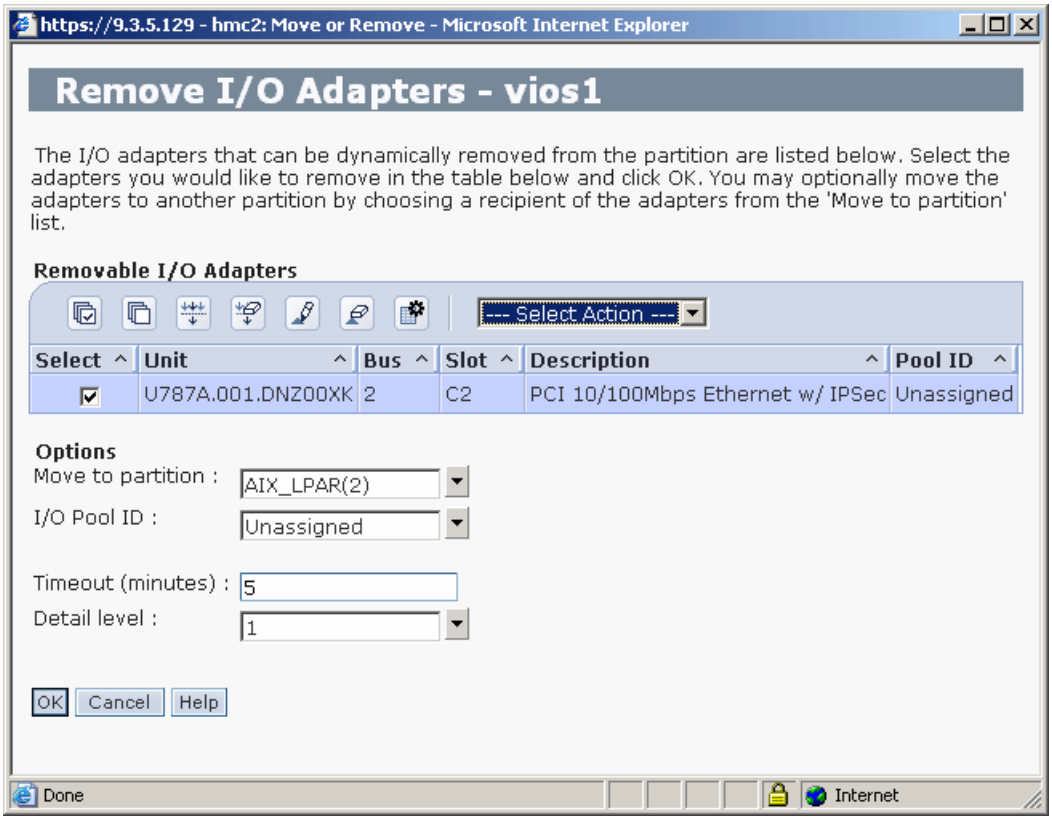


Figure 12-13 Selecting adapter in slot C2 to be moved to partition AIX_LPAR

3. Click **OK** to run the operation.
4. For an AIX partition, run the **cfgmgr** command (**cfgdev** in the Virtual I/O Server) in the receiving partition to make the adapter and its devices available.

   An IBM i partition, by default, automatically discovers and configures new devices that are attached to it if the system value QAUTOCFG is set to 1. Therefore, they only need to be varied on by using the VRYCFG command before they are used.

5. To reflect the change across restarts of the partitions, remember to update the profiles of both partitions.

   Alternatively, use the **Configuration** → **Save Current Configuration** option to save the changes to a profile as shown in Figure 12-14.

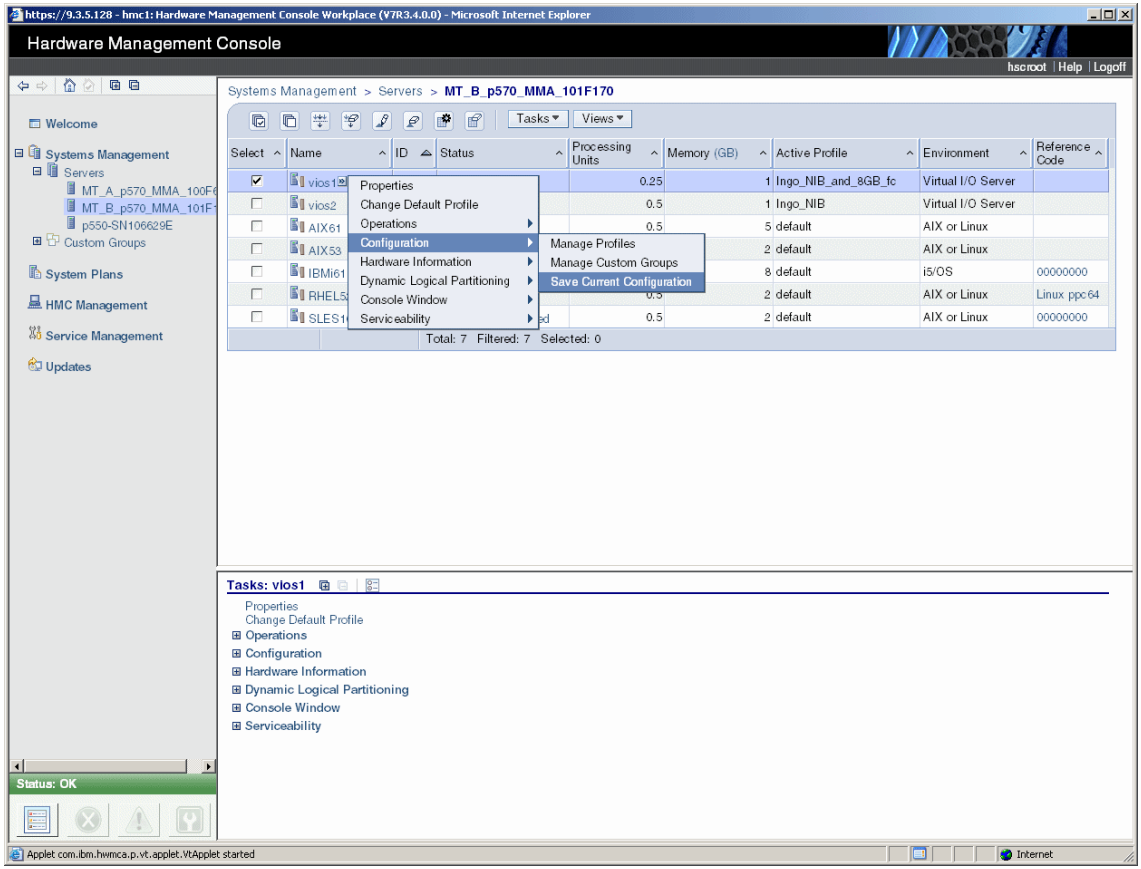


Figure 12-14 Save current configuration

Removing physical adapters dynamically

First, release the adapter resource to be removed from the partition that owns it as described in “Preparing the move or removal of adapters with dynamic LPAR”

on page 467. Then, to remove physical adapters from a partition dynamically, complete these steps:

1. On the HMC, select the partition to remove the adapter from and select **Dynamic Logical Partitioning** → **Physical Adapters** → **Move or Remove** (Figure 12-15).

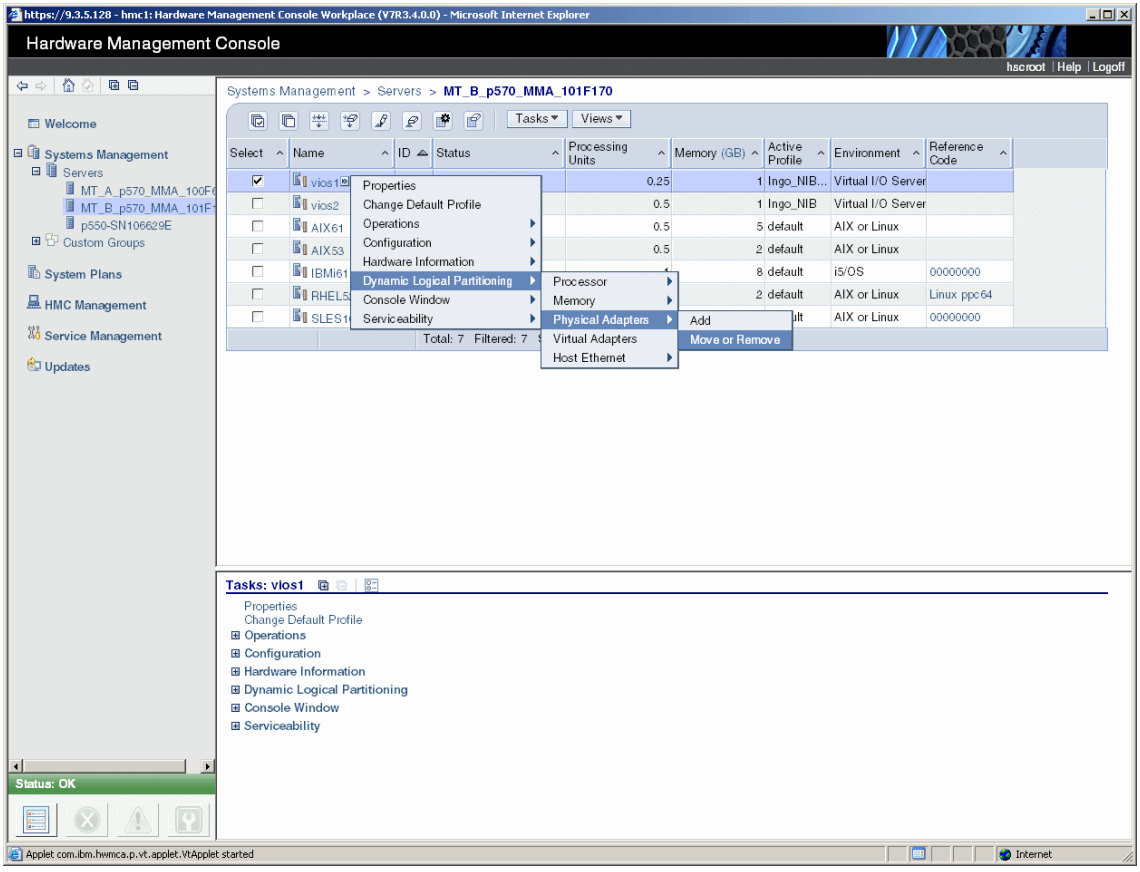


Figure 12-15 Remove physical adapter operation

2. Select the adapter that you want to delete and do not select any partition in **Move to partition** selection box as shown in Figure 12-16.

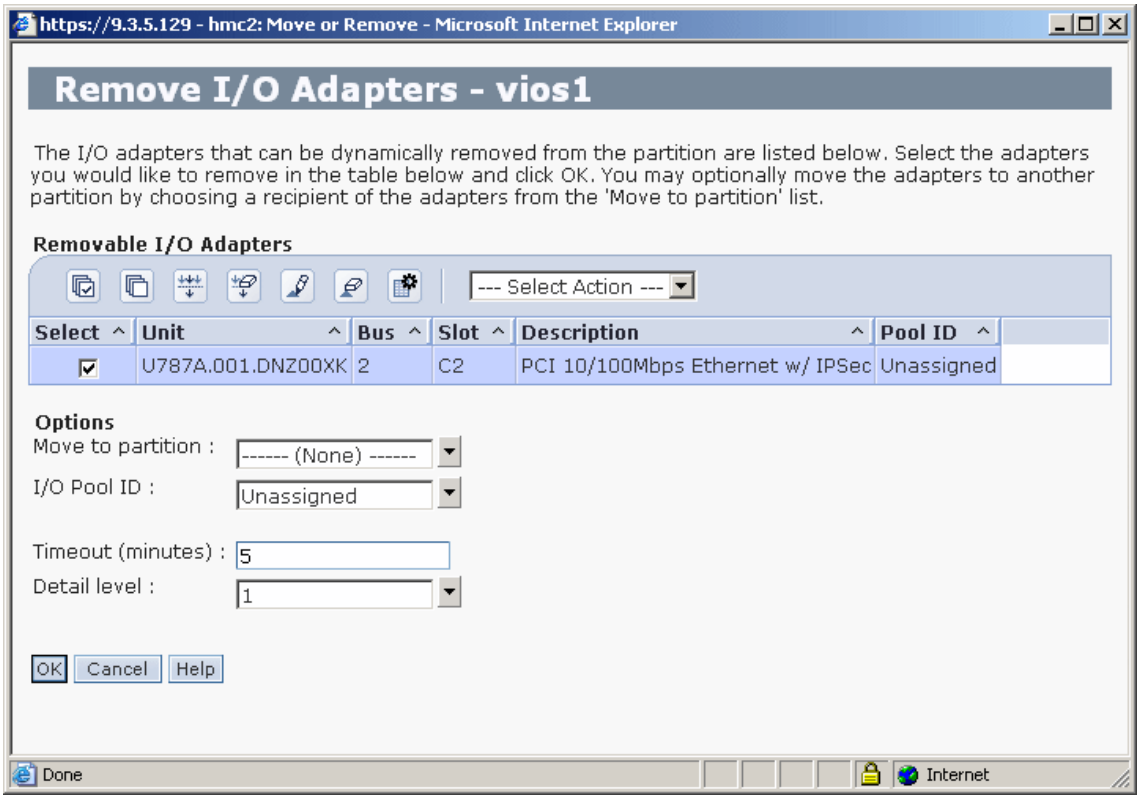


Figure 12-16 Selecting the physical adapter to be removed

3. Click **OK** when done.

Adding virtual adapters dynamically

The following steps illustrate one way to add virtual adapters dynamically:

1. Log in to HMC and then select the system-managed name. In the right window, select the partition where you want to perform a dynamic LPAR operation.

2. On the **Tasks** menu on the right side of the window, select **Dynamic Logical Partitioning** → **Virtual Adapters** as shown in Figure 12-17.

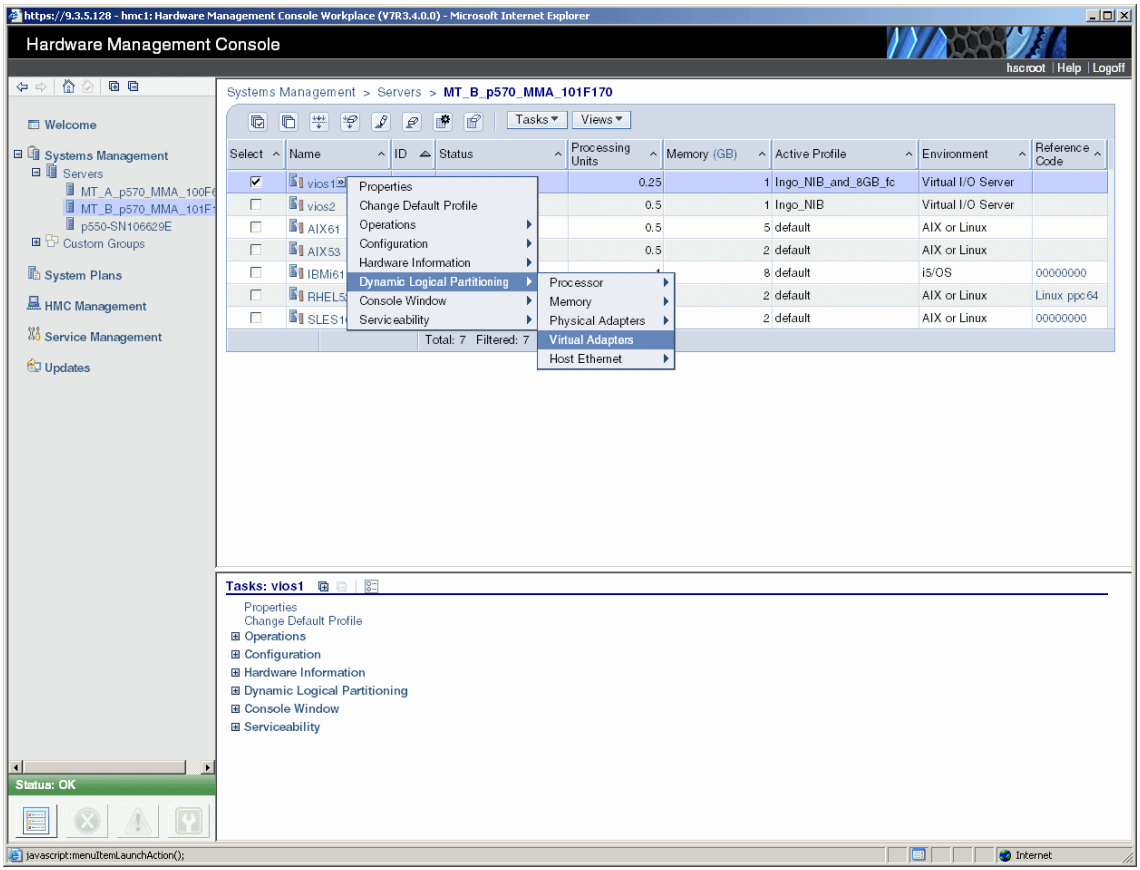


Figure 12-17 Add virtual adapter operation

3. Click **Actions** and select **Create** followed by the virtual adapter type that you want to add, such as **SCSI Adapter** as shown in Figure 12-18.

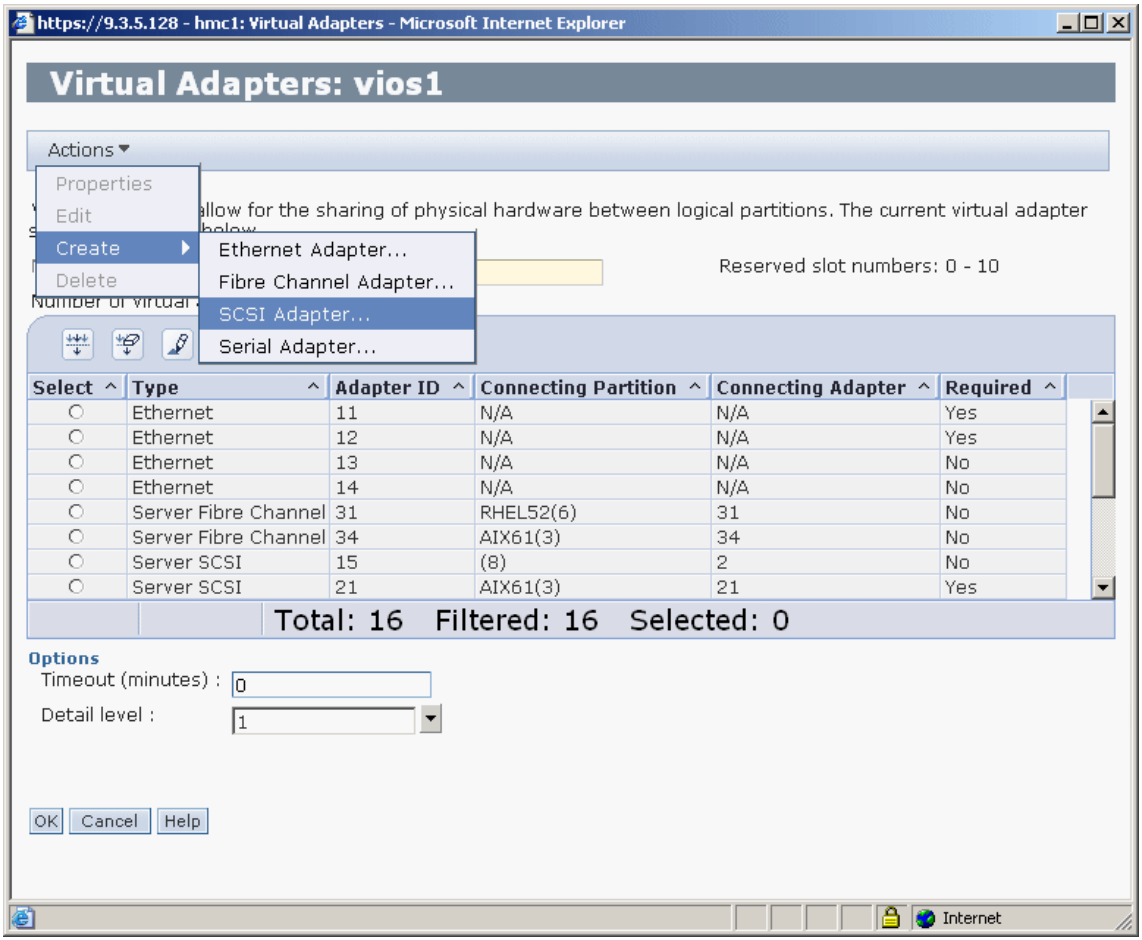


Figure 12-18 Dynamically adding a virtual SCSI adapter

4. Figure 12-19 shows the window after you select **SCSI Adapter**. Type the slot adapter number of the new virtual SCSI being created, then select whether this new SCSI adapter can be accessed by any client partition or only by a specific one. In this case, as an example, only the SCSI client adapter in slot 2 of the AIX61 partition is allowed to access it.

Consideration: This example uses a different slot numbering for the client and the server virtual SCSI adapter. Be sure to create an overall numbering scheme.

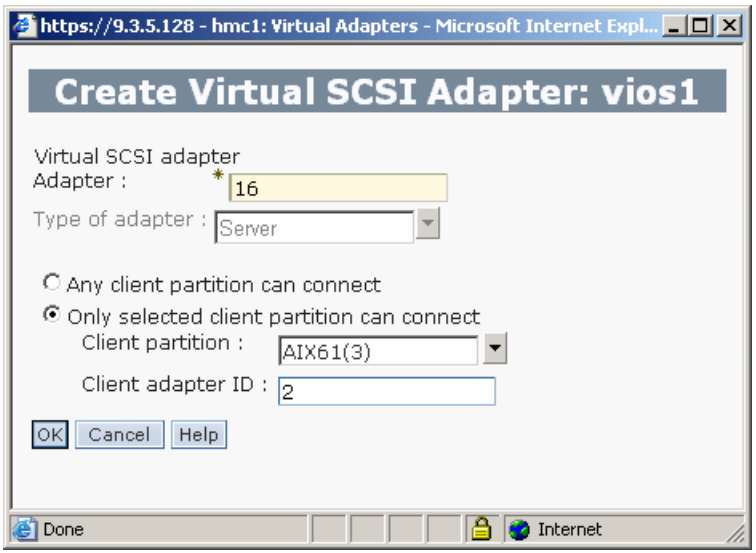


Figure 12-19 Virtual SCSI adapter properties

5. The newly created adapter is displayed in the adapters list as shown in Figure 12-20.

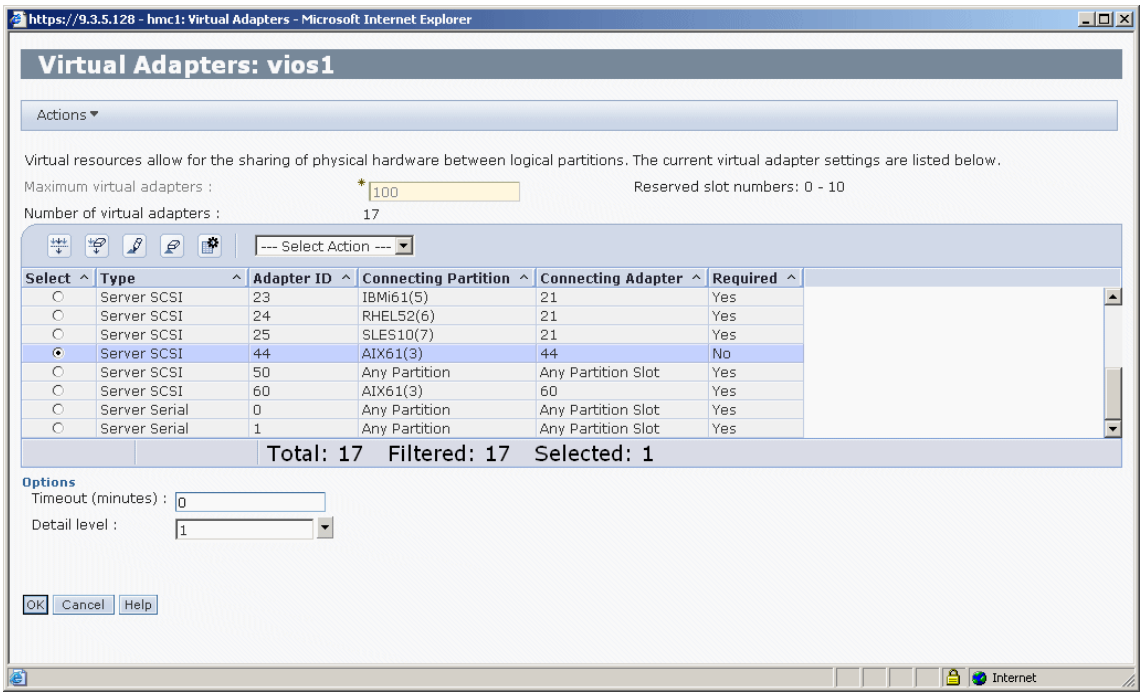


Figure 12-20 Virtual adapters for an LPAR

6. Click **OK** when done.

To reflect the change across restart of the partition, remember to update the profile of partition.

Removing virtual adapters dynamically

To remove virtual adapters from a partition dynamically, complete these steps:

1. For AIX, unconfigure the devices and the virtual adapter itself on the AIX. For IBM i, vary-off any devices that are attached to the virtual client adapter.

> **Remember:** First, remove all associated virtual *client* adapters in the virtual I/O clients before you remove a virtual *server* adapter in the Virtual I/O Server.

2. On the HMC, select the partition to remove the adapter from and click **Dynamic Logical Partitioning** → **Virtual Adapters** (Figure 12-21).

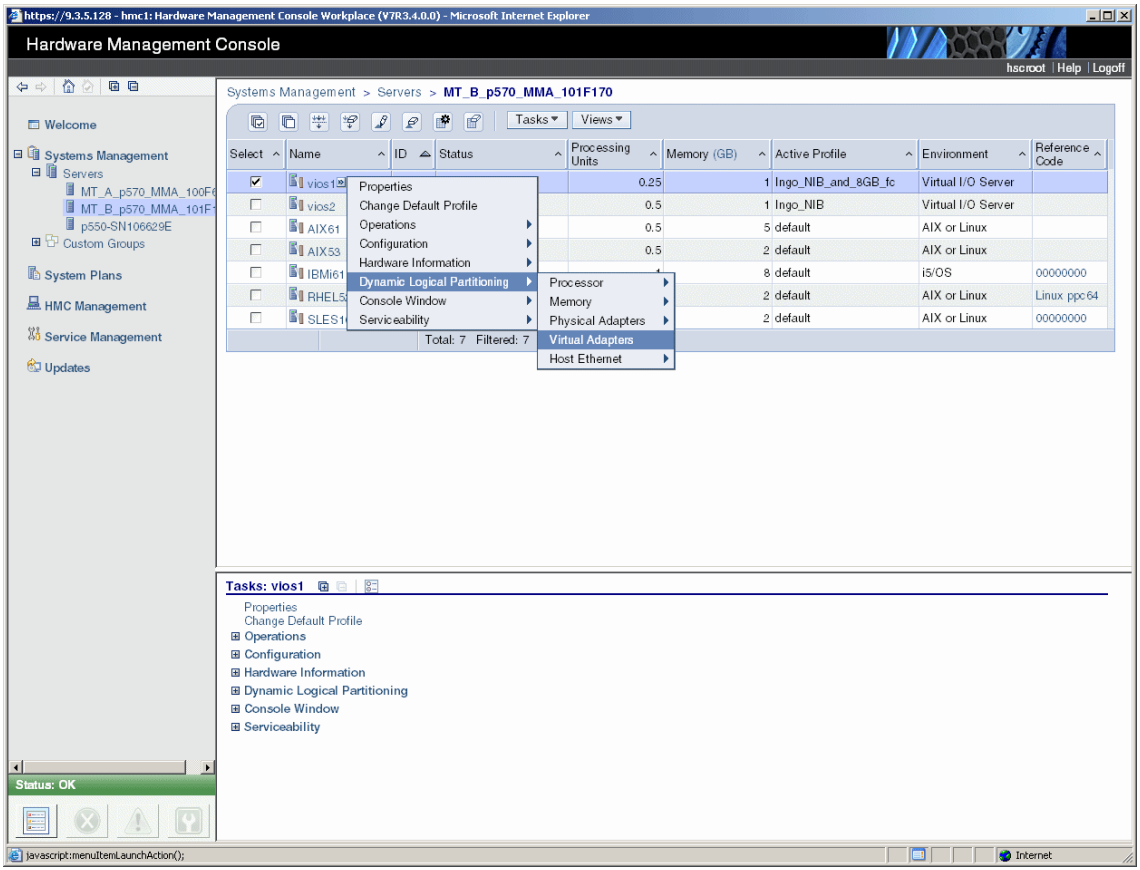


Figure 12-21 Remove virtual adapter operation

3. Select the adapter that you want to delete and click **Actions** → **Delete** (Figure 12-22).

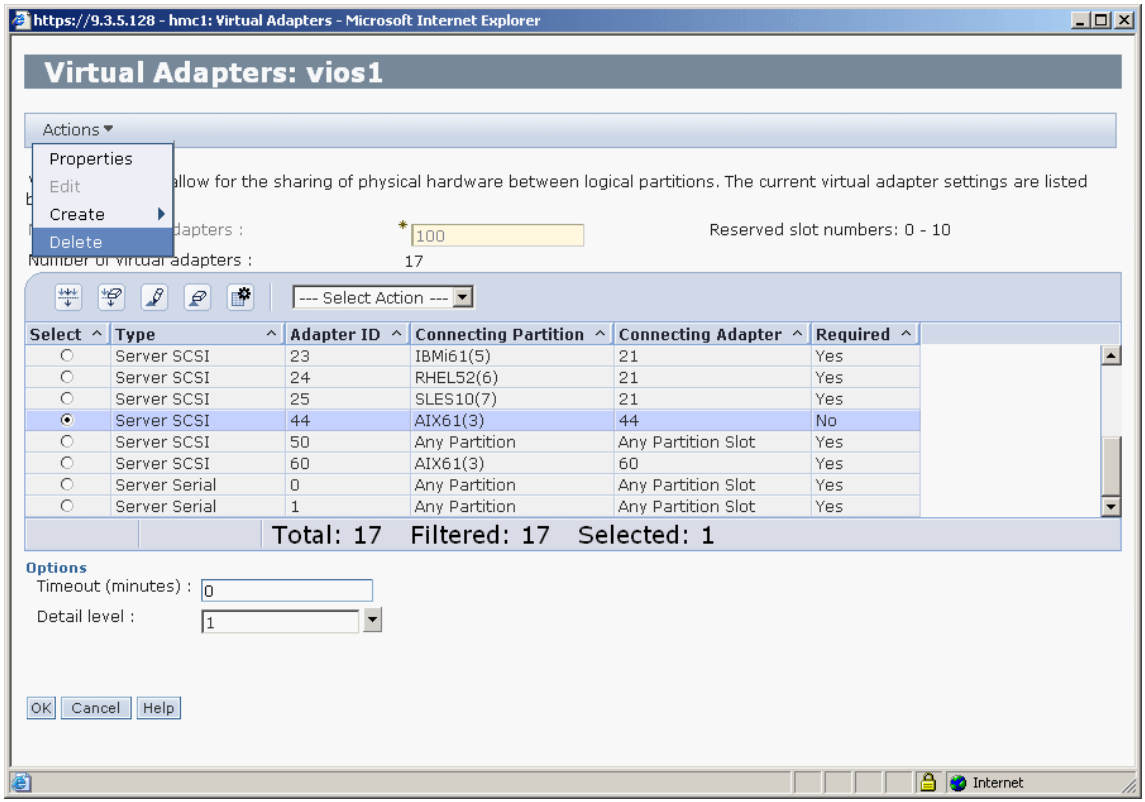


Figure 12-22 Delete virtual adapter

4. Click **OK** when done.

Removing or replacing a PCI Hot Plug adapter

The PCI Hot Plug feature enables you to remove host-based adapters without shutting down the partition. Replacing an adapter might be needed, for example, when you exchange 2 Gb Fibre Channel adapters for a 4 Gb Fibre Channel adapter. It might also be needed when you make configuration changes or updates.

For virtual Ethernet adapters in the virtual I/O client, redundancy must be enabled either through Shared Ethernet failover being enabled on the Virtual I/O Servers, or through Network Interface Backup being configured if continuous network connectivity is required. If there is no redundancy for Ethernet, the replace operation for a physical Ethernet adapter on the Virtual I/O Server can be

done while the virtual I/O client is still running. However, it loses network connectivity during replacement. For virtual I/O clients that have redundant paths to their virtual disks and are not mirroring these disks, you must shut them down while the adapter is being replaced.

On the Virtual I/O Server, in both cases child devices are connected to the adapter because the adapter was in use before. Therefore, the child devices and the adapter must be unconfigured before the adapter can be removed or replaced. Normally, you do not need to remove the child devices (for example disks and mapped disks, also known as Virtual Target Devices) during a Fibre Channel adapter replacement. However, they must be unconfigured (set to the defined state) before the adapter they rely on can be replaced.

12.1.2 Dynamic LPAR operations on Linux

This section explains how to run dynamic LPAR operations for Power Systems running Linux.

Service and productivity tools for Linux

Virtualization and hardware support in Linux for IBM Power Systems is realized through open source drivers included in the standard Linux Kernel for 64-bit POWER-based systems. However, IBM provides more tools for virtualization management. These tools are useful for using advanced features and hardware diagnostic tests. These tools are called *Service and productivity tools for POWER Linux servers*, and are provided as a no-cost download for all supported distributions and systems.

The tools include Reliable Scalable Cluster Technology (RSCT) daemons that are used for communication with the Hardware Management Console (HMC). Some packages are open source and are included on the distribution media. However, the website download offers the latest version, which is available at:

`http://www14.software.ibm.com/webapp/set2/sas/f/lopdiags`

Consideration: Package names and dependencies can vary between one Linux distribution and another. For more information about each Linux release, see the service and productivity tools website.

Table 12-1 provides details about each package.

Table 12-1 Service and productivity tools description

Tool name	Description
librtas	**Platform Enablement Library** (base tool) The librtas package contains a library that allows applications to access certain functions that are provided by platform firmware. This functionality is required by many of the other higher-level service and productivity tools. This package is open source and shipped by both Red Hat and Novell SUSE.
src	**SRC** src is a facility for managing daemons on a system. It provides a standard command interface for defining, undefining, starting, stopping, querying status, and controlling trace for daemons. This package is IBM proprietary.
rsct.core and rsct.core.utils	**Reliable Scalable Cluster Technology (RSCT) core and utilities** The RSCT packages provide the RMC functions and infrastructure that are needed to monitor and manage one or more Linux systems. RMC provides a flexible and extensible system for monitoring numerous aspects of a system. It also allows customized responses to detected events. This package is IBM proprietary.
csm.core and csm.client	**Cluster Systems Management (CSM) core and client** The CSM packages provide for the exchange of host-based authentication security keys. These tools also set up distributed RMC features on the HMC. This package is IBM proprietary.
devices.chrp.base .ServiceRM	**Service Resource Manager (ServiceRM)** Service Resource Manager is an RSCT resource manager that creates the Serviceable Events from the output of the Error Log Analysis Tool (diagela). ServiceRM then sends these events to the Service IBM Focal Point™ on the HMC. This package is IBM proprietary.

Tool name	Description
DynamicRM	**DynamicRM (Productivity tool)** Dynamic Resource Manager is an RSCT resource manager that allows an HMC to run the following tasks: ► Dynamically add or remove processors or I/O slots from a running partition ► Concurrently update system firmware ► Perform certain shutdown operations on a partition ► Show how virtual disks map to disk names in Linux and to virtual Ethernet interfaces ► Support migration from POWER6-based to POWER7-based servers. This package is IBM proprietary.
lsvpd / libvpd	**Hardware Inventory** The `lsvpd` package contains the **`lsvpd`**, **`lscfg`**, and **`lsmcode`** commands. These commands, along with a boot-time scanning script named `update-lsvpd-db`, constitute a hardware inventory system. The **`lsvpd`** command provides vital product data (VPD) about hardware components to higher-level serviceability tools. The **`lscfg`** command provides a more human-readable format of the VPD, and system-specific information. This package is open source, and shipped by both Red Hat and Novell SuSE.
servicelog	**Service Log** (service tool) The Service Log package creates a database to store system-generated events that might require service. The package includes tools for querying the database. This package is open source, and shipped by both Red Hat and Novell SuSE.
ppc64-diag	**Error Log Analysis** This tool provides automatic analysis and notification of errors that are reported by the platform firmware on IBM systems. This RPM analyzes errors that are written to `/var/log/platform`. If a corrective action is required, notification is sent to the Service Focal Point on the HMC, if so equipped, or to users subscribed for notification through the file `/etc/diagela/mail_list`. The Serviceable Event sent to the Service Focal Point and listed in the email notification might contain a Service Request Number. This number is listed in the Diagnostics manual *Information for Multiple Bus Systems*. This package is IBM proprietary.

Tool name	Description
powerpc-utils	**Service Aids** The utilities in the `powerpc-utils` and `powerpc-utils-papr` packages enable several RAS (reliability, availability, and serviceability) features. Among others, these utilities include the **`update_flash`** command for installing system firmware updates, the **`serv_command`** for modifying various serviceability policies, the **`usysident`** and **`usysattn`** utilities for manipulating system LEDs, the **`bootlist`** command for updating the list of devices from which the system will boot, and the **`snap`** command for capturing extended error data to aid analysis of intermittent errors. This package is open source.
IBMinvscout	**Inventory Scout** This tool surveys one or more systems for hardware and software information. The gathered data can be used by web services such as the Microcode Discovery Service, which generates a report indicating if installed microcode needs to be updated. This package is IBM proprietary.

IBM Installation Toolkit for PowerLinux

As an alternative to manually installing extra packages as described in “Installing the service and productivity tools” on page 485, you can use the *IBM Installation Toolkit for PowerLinux™*.

The IBM Installation Toolkit for PowerLinux is a bootable CD that provides access to the extra packages that you need to install to provide more capabilities for your server. You can also use it to set up an installation server to make your customized operating system installation files available for other server installations. Download the IBM Installation Toolkit for PowerLinux ISO image from:

`http://www14.software.ibm.com/webapp/set2/sas/f/lopdiags/installtools/home.html`

The IBM Installation Toolkit for PowerLinux simplifies the Linux installation by providing a wizard that helps you install and configure Linux for IBM Power Systems servers in just a few steps. It supports DVD and network-based installations by providing an application to create and manage network repositories that contain Linux and IBM packages.

The IBM Installation Toolkit includes these applications:

- The Welcome Center, the main toolkit application, which is a centralized user interface for system diagnostics, Linux, and RAS Tools installation; microcode update; and documentation

- System Tools, which is an application to create and manage network repositories for Linux and IBM RAS packages
- The POWER Advance Toolchain, which is a technology preview toolchain that provides decimal floating point support, Power architecture c-library optimizations, optimizations in the gcc compiler for POWER, and performance analysis tools
- Microcode packages
- More than 20 RAS tools packages
- More than 60 Linux user guides and manuals

Installing the service and productivity tools

Download the service and productivity tools at:

http://www14.software.ibm.com/webapp/set2/sas/f/lopdiags/home.html

Select your distribution and whether you have an HMC-connected system.

Download all the packages in one directory and run the **rpm -Uvh *<filename>*.rpm** command for each package individually. Depending on your Linux distribution, version, and installation choice, you are prompted for missing dependencies. Keep your software installation source for your distribution available and accessible.

Important: Install the packages in the same order as listed at the service and productivity tools website. This avoids installation and setup issues.

Service and Productivity tools examples

After the packages are installed, run the **vpdupdate** command to initialize the Vital Product Data database.

Now you can list hardware with the **lscfg** command and see correct location codes as shown in Example 12-2.

Example 12-2 lscfg command on Linux

```
[root@localhost ~]# lscfg
INSTALLED RESOURCE LIST

The following resources are installed on the machine.
+/- = Added or deleted from Resource List.
*   = Diagnostic support not available.

  Model Architecture: chrp
  Model Implementation: Multiple Processor, PCI Bus
```

```
+ sys0                                  System Object
+ sysplanar0                           System Planar
+ eth0             U9117.MMA.101F170-V5-C2-T1
                                        Interpartition Logical LAN
+ eth1             U9117.MMA.101F170-V5-C3-T1
                                        Interpartition Logical LAN
+ scsi0            U9117.MMA.101F170-V5-C21-T1
                                        Virtual SCSI I/O Controller
+ sda              U9117.MMA.101F170-V5-C21-T1-L1-L0
                                        Virtual SCSI Disk Drive (21400 MB)
+ scsi1            U9117.MMA.101F170-V5-C22-T1
                                        Virtual SCSI I/O Controller
+ sdb              U9117.MMA.101F170-V5-C22-T1-L1-L0
                                        Virtual SCSI Disk Drive (21400 MB)
+ mem0                                  Memory
+ proc0                                 Processor
```

You can use the **lsvpd** command to display vital product data (for example, the firmware level), as shown in Example 12-3.

Example 12-3 lsvpd command

```
[root@localhost ~]# lsvpd
*VC 5.0
*TM IBM,9117-MMA
*SE IBM,02101F170
*PI IBM,02101F170
*OS Linux 2.6.18-53.el5
```

To display virtual adapters, use the **lsvio** command as shown in Example 12-4.

Example 12-4 Displaying the virtual SCSI and network

```
[root@linuxlpar ~]# lsvio -s
scsi0 U9117.MMA.101F170-V5-C21-T1
scsi1 U9117.MMA.101F170-V5-C22-T1
[root@linuxlpar ~]# lsvio -e
eth0 U9117.MMA.101F170-V5-C2-T1
eth1 U9117.MMA.101F170-V5-C3-T1
```

Dynamic LPAR with Linux

Dynamic logical partitioning is needed to change the physical or virtual resources that are assigned to the partition without a reboot or other disruption. If you create or assign a new virtual adapter, a dynamic LPAR operation is needed to make the operating system aware of this change. On Linux-based systems, existing hot plug and udev mechanisms are used for dynamic LPAR operations.

Therefore, all the changes occur dynamically and there is no need to run a configuration manager.

Dynamic LPAR requires a working IP connection to the HMC (port 657) and the following extra packages installed on the Linux system as described in “Installing the service and productivity tools” on page 485:

```
librtas, src, rsct.core and rsct.core.utils, csm.core and csm.client,
powerpc-utils-papr, devices.chrp.base.ServiceRM, DynamicRM,
rpa-pci-hotplug, rpa-dlpar
```

If you are encountering problems with dynamic LPAR operations not working, check with your HMC whether the partition is connected to as shown in Example 12-5. Also, check whether you can ping the HMC from the Linux partition.

Example 12-5 Listing the management server

```
root@p750_lpar02 ~]# lsrsrc IBM.MCP
Resource Persistent Attributes for IBM.MCP
resource 1:
        MNName            = "172.16.21.126"
        NodeID            = 15025524784985301456
        KeyToken          = "hmc8"
        IPAddresses       = {"172.16.20.114"}
        ConnectivityNames = {"172.16.21.126"}
        HMCName           = "7042CR6*107627C"
        HMCIPAddr         = "172.16.20.114"
        HMCAddIPs         = "192.168.128.1"
        HMCAddIPv6s       = ""
        ActivePeerDomain  = ""
        NodeNameList      = {"p750_lpar02"}
```

For more information about diagnosing RMC connection problems, see the IBM Technote *The most common reasons for failures with Dynamic Logical Partitioning* at:

https://www-304.ibm.com/support/docview.wss?uid=isg3T1010615

Adding processor or memory resources dynamically

After the tools are installed, and depending on the available shared system resources, you can use the HMC to add (virtual) processors or memory to the desired partition shown in Figure 12-23.

Consideration: The dynamic LPAR changes made to the partition attributes (processor and memory) are not saved to the current active profile. Therefore, save the current partition configuration by selecting the partition and clicking **Configuration** → **Save Current Configuration** and then, during the next reboot, using the saved partition profile as the default profile.

An alternative method is to make the same changes in the default profile.

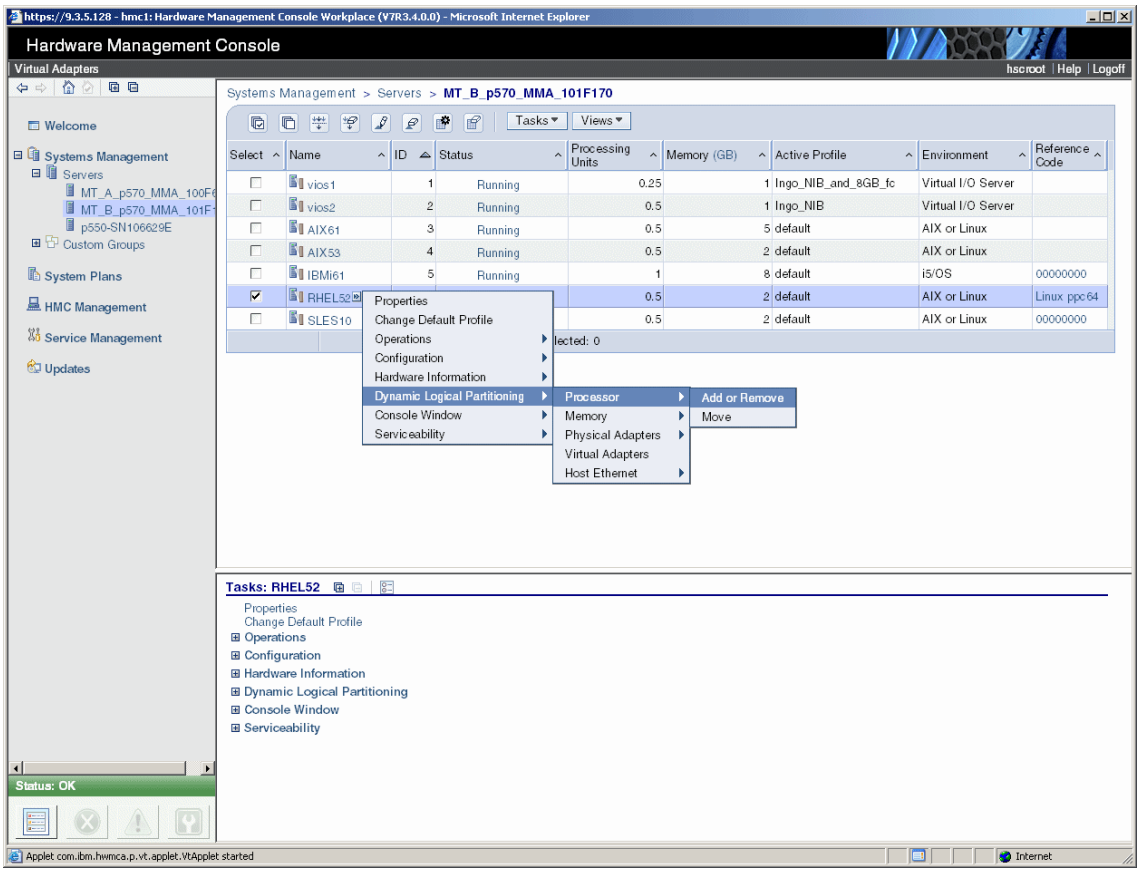


Figure 12-23 Adding a processor to a Linux partition

From the window shown in Figure 12-24, you can increase the number of processors.

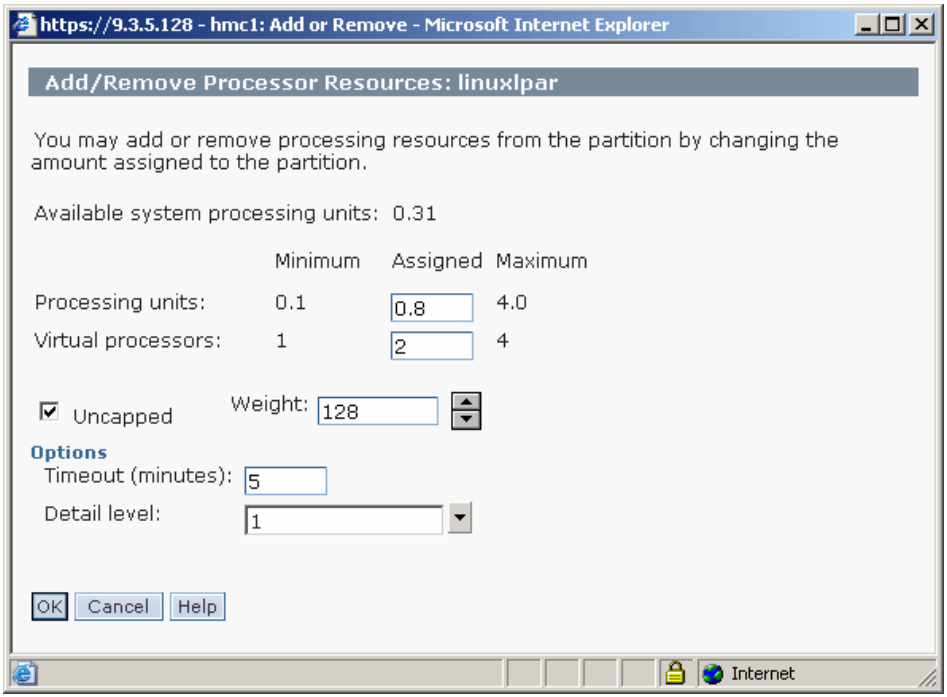


Figure 12-24 Increasing the number of virtual processors

Adding memory dynamically

Adding memory is supported by Red Hat Enterprise (RHEL5.0 or later) and Novell SUSE (SLES10 or later) Linux distributions.

Important: You must ensure that `powerpc-utils` RPM is installed. See the productivity download site for the release information.

From the Hardware Management Console of the partition, click **System Management** → your **Server** → **Dynamic Logical Partitioning** → **Memory** → **Add or Remove** (Figure 12-25).

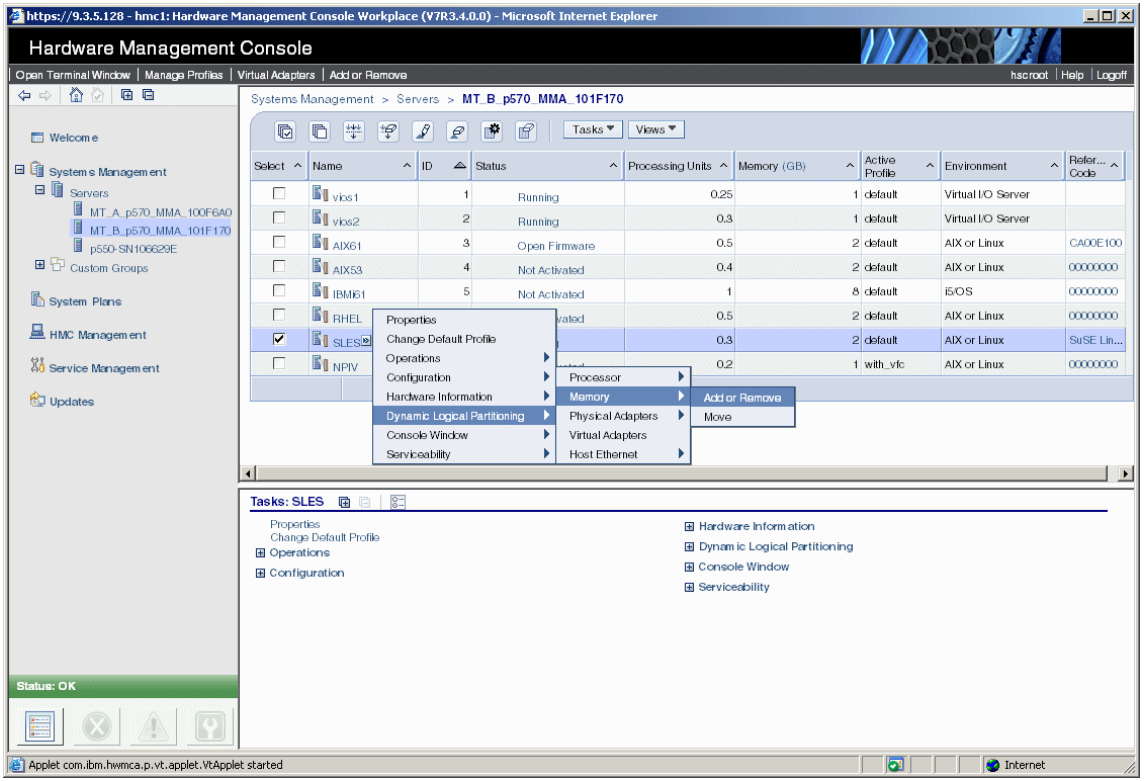


Figure 12-25 Dynamic LPAR add or remove memory

After you select **Add or Remove**, you see the display that is shown in Figure 12-26.

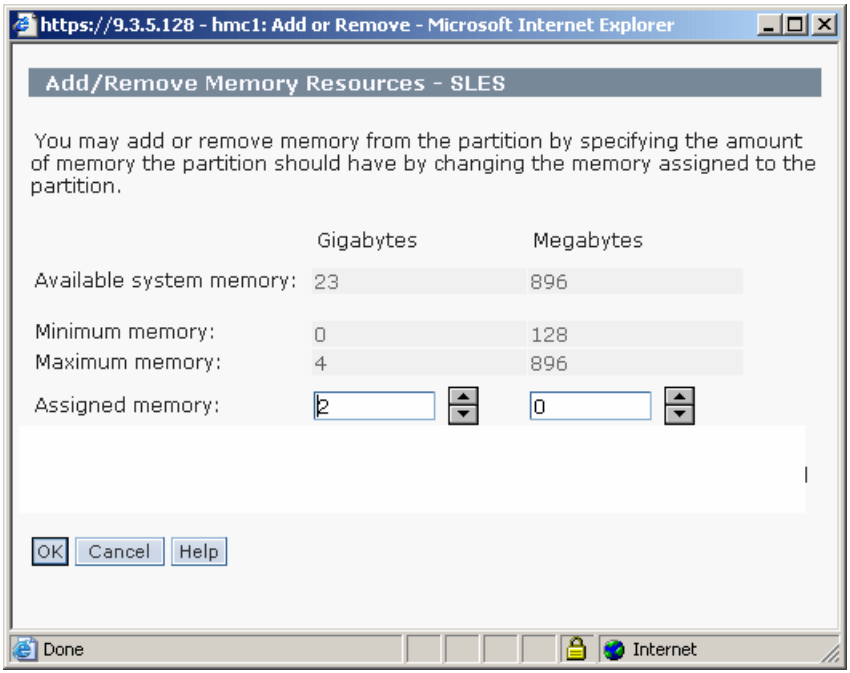


Figure 12-26 Dynamic LPAR adding 2 GB of memory

Enter the desired memory for the partition in the Assigned Memory box by increasing the value (in the example, it is increased by 1 GB), then click **OK**.

Managing virtual SCSI changes in Linux

If a new virtual SCSI adapter is added to a Linux partition and dynamic LPAR is functional, this adapter and any attached disks are immediately ready for use.

Sometimes you might need to add a virtual target device to an existing virtual SCSI adapter. In this case, the operation is not a dynamic LPAR operation. The adapter itself does not change. Instead, a new disk is attached to the same adapter. You must issue a **scan** command to recognize this new disk, and run the **dmesg** command to see the result as shown in Example 12-6.

Example 12-6 Rescanning a SCSI host adapter

```
# echo "- - -" > /sys/class/scsi_host/host0/scan
# dmesg
# SCSI device sdb: 585937500 512-byte hdwr sectors (300000 MB)
sdb: Write Protect is off
sdb: Mode Sense: 2f 00 00 08
```

```
sdb: cache data unavailable
sdb: assuming drive cache: write through
SCSI device sdb: 585937500 512-byte hdwr sectors (300000 MB)
sdb: Write Protect is off
sdb: Mode Sense: 2f 00 00 08
sdb: cache data unavailable
sdb: assuming drive cache: write through
 sdb: sdb1 sdb2
sd 0:0:2:0: Attached scsi disk sdb
sd 0:0:2:0: Attached scsi generic sg1 type 0
```

The added disk is recognized and ready to use as `/dev/sdb`.

If you are using software mirroring on the Linux client and one of the adapters was set *faulty* because of Virtual I/O Server maintenance, you might need to rescan the disk after both Virtual I/O Servers are available again. Use the following command to issue a disk rescan:

```
echo 1 > /sys/bus/scsi/drivers/sd/<SCSI-ID>/block/device/rescan
```

More examples and detailed information about SCSI scanning is provided in the PowerLinux IBM Wiki page at:

http://www-941.ibm.com/collaboration/wiki/display/LinuxP/SCSI+-+Hot+add%2C+remove%2C+rescan+of+SCSI+devices

12.1.3 Dynamic LPAR operations on the Virtual I/O Server

This section describes two maintenance tasks for a Virtual I/O Server partition:

- Replacing Ethernet adapters on the Virtual I/O Server
- Replacing a Fibre Channel adapter on the Virtual I/O Server

If you want to change any processor, memory, or I/O configuration, follow the steps that are described in 12.1.1, “Dynamic LPAR operations on AIX and IBM i” on page 458. The steps are similar to the ones required for the Virtual I/O Server.

Note: With HMC V7R7.6 or later, newly added/removed *virtual* adapter resources to/from the Virtual I/O Server by using dynamic LPAR are automatically configured/deconfigured.

Replacing Ethernet adapters on the Virtual I/O Server

You can perform the replace and remove functions by using the **diagmenu** command. Complete these steps:

1. Run the **diagmenu** command.
2. Read the Diagnostics Operating Instructions and press Enter to continue.
3. Select `Task Selection (Diagnostics, Advanced Diagnostics, Service Aids, etc.)` and press Enter.
4. Select `Hot Plug Task` and press Enter.
5. Select `PCI Hot Plug Manager` and press Enter.
6. Select `Replace/Remove a PCI Hot Plug Adapter` and press Enter.
7. Select the adapter that you want to replace and press Enter.
8. Select `replace` in the `Operation` field and press Enter.
9. Before a replace operation is performed, the adapter can be identified by a flashing LED at the adapter card. You see the following message:

   ```
   The visual indicator for the specified PCI slot has been set to the
   identify state. Press Enter to continue or enter x to exit.
   ```

If there are still devices connected to the adapter and a replace or remove operation is run on that device, the following error messages are displayed in **diagmenu**:

```
The visual indicator for the specified PCI slot has been set to the identify
state. Press Enter to continue or enter x to exit.

The specified slot contains device(s) that are currently
configured. Unconfigure the following device(s) and try again.

pci5
ent0
ent1
ent2
ent3
```

These messages mean that devices that are dependent on this adapter must be unconfigured first.

To replace a single Physical Ethernet adapter that is part of a Shared Ethernet Adapter, complete these steps:

1. Use the **diagmenu** command to unconfigure the Shared Ethernet Adapter by selecting **Task Selection** → **Hot Plug Task** → **PCI Hot Plug Manager** → **Unconfigure a Device**.

   You should get output similar to the following:

```
Device Name

Move cursor to desired item and press Enter. Use arrow keys to scroll.

[MORE...16]
  en7          Defined            Standard Ethernet Network Interface
  ent0         Available 05-20    4-Port 10/100/1000 Base-TX PCI-X Adapt
  ent1         Available 05-21    4-Port 10/100/1000 Base-TX PCI-X Adapt
  ent2         Available 05-30    4-Port 10/100/1000 Base-TX PCI-X Adapt
  ent3         Available 05-31    4-Port 10/100/1000 Base-TX PCI-X Adapt
  ent4         Available          Virtual I/O Ethernet Adapter (l-lan)
  ent5         Available          Virtual I/O Ethernet Adapter (l-lan)
  ent6         Available          Shared Ethernet Adapter
  et0          Defined   05-20    IEEE 802.3 Ethernet Network Interface
  et1          Defined   05-21    IEEE 802.3 Ethernet Network Interface
[MORE...90]
Select the Shared Ethernet Adapter (in this example, ent6), and in the
following dialogue choose to keep the information about the database:
Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
* Device Name                                        [ent6]
+
  Unconfigure any Child Devices                      no
+
  KEEP definition in database                        yes
+
```

   Press Enter to accept the changes. The system shows that the adapter is now defined:

```
ent6 Defined
```

2. Perform the same operation on the physical adapter (in this example `ent0`, `ent1`, `ent2`, `ent3`, and `pci5`). The difference is that now `Unconfigure any Child devices` must be set to `yes`.

3. Run the **diagmenu** command. Then, select `Task Selection` → `Hot Plug Task` → `PCI Hot Plug Manager` → `Replace/Remove a PCI Hot Plug adapter`, and select the physical adapter.

   You see an output panel similar to the following:

```
Command: running       stdout: yes           stderr: no

Before command completion, additional instructions may appear below.


The visual indicator for the specified PCI slot has
been set to the identify state. Press Enter to continue
or enter x to exit.
```

   Press Enter as directed and the next message will appear:

```
The visual indicator for the specified PCI slot has
been set to the action state. Replace the PCI card
in the identified slot and press Enter to continue.
Enter x to exit. Exiting now leaves the PCI slot
in the removed state.
```

4. Locate the flashing adapter, replace it, and press Enter. The window then displays the message `Replace Operation Complete`.
5. Run the **diagmenu** command. Then, select `Task Selection` → `Hot Plug Task` → `PCI Hot Plug Manager` → `Configure a Defined Device`. Select the physical Ethernet adapter `ent0` that was replaced.
6. Repeat the Configure operation for the Shared Ethernet Adapter.

This method changes if the physical Ethernet adapter is part of a Network Interface Backup Configuration or an IEE 802.3ad link aggregation.

Replacing a Fibre Channel adapter on the Virtual I/O Server

For Virtual I/O Servers, have at least two Fibre Channel adapters attached for redundant access to FC-attached disks. This configuration allows for concurrent maintenance because the multipathing driver of the attached storage subsystem is supposed to handle any outage of a single Fibre Channel adapter.

This section explains how to hot-plug replace a Fibre Channel adapter that is connected to an IBM DS4000 series storage device. Depending on the storage subsystem used and the multipathing driver installed, your results might be different.

If there are disks mapped to the virtual SCSI adapters, these devices must be unconfigured first. There is no automatic configuration method used to define them.

1. Use the **diagmenu** command to unconfigure devices that are dependent on the Fibre Channel adapter. Run **diagmenu** and then select `Task Selection` → `Hot Plug Task` → `PCI Hot Plug Manager` → `Unconfigure a device`.

   Select the disk (or disks) in question and set its state to `Defined`:

```
Unconfigure a Device

Device Name

Move cursor to desired item and press Enter. Use arrow keys to scroll.
[MORE...43]
hdisk6       Available 04-08-02      3542     (200) Disk Array Device
hdisk9       Defined   09-08-00-4,0  16 Bit LVD SCSI Disk Drive
inet0        Available               Internet Network Extension
iscsi0       Available               iSCSI Protocol Device
lg_dumplv    Defined                 Logical volume
lo0          Available               Loopback Network Interface
loglv00      Defined                 Logical volume
lpar1_rootvg Available               Virtual Target Device - Disk
lpar2_rootvg Available               Virtual Target Device - Disk
lvdd         Available               LVM Device Driver
[MORE...34]
```

2. Perform that task for every mapped disk (Virtual Target Device). Then, set the state of the Fibre Channel Adapter to *Defined* also:

```
Unconfigure a Device

Device Name                                 ?
Move cursor to desired item and press Enter. Use arrow keys to scroll.
[MORE...16]
et1          Defined   05-09         IEEE 802.3 Ethernet Network Inter
et2          Defined                 IEEE 802.3 Ethernet Network Inter
et3          Defined                 IEEE 802.3 Ethernet Network Inter
et4          Defined                 IEEE 802.3 Ethernet Network Inter
fcnet0       Defined   04-08-01      Fibre Channel Network Protocol De
fcnet1       Defined   06-08-01      Fibre Channel Network Protocol De
fcs0         Available 04-08         FC Adapter
fcs1         Available 06-08         FC Adapter?
fscsi0       Available 04-08-02      FC SCSI I/O Controller Protocol D
fscsi1       Available 06-08-02      FC SCSI I/O Controller Protocol D?
[MORE...61]
```

Be sure to set `Unconfigure any Child Devices` to `Yes`. This unconfigures the `fcnet0` and `fscsi0` devices, and the RDAC driver device `dac0` as shown:

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

                                                      [Entry Fields]
* Device Name                                         [fcs0]
  Unconfigure any Child Devices                        yes
  KEEP definition in database                          yes
```

Following is the output of that command, showing the other devices unconfigured:

```
COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

fcnet0 Defined
dac0 Defined
fscsi0 Defined
fcs0 Defined
```

3. Run **diagmenu**, and select `Task Selection` → `Hot Plug Task` → `PCI Hot Plug Manager` → `Replace/Remove a PCI Hot Plug Adapter`.
4. Select the adapter to be replaced. Set the operation to replace, then press Enter. You are presented with the following panel:

```
COMMAND STATUS

Command: running       stdout: yes           stderr: no

Before command completion, additional instructions may appear below.


The visual indicator for the specified PCI slot has
been set to the identify state. Press Enter to continue
or enter x to exit.
```

5. Press Enter as directed and the following message appears:

```
The visual indicator for the specified PCI slot has
been set to the action state. Replace the PCI card
in the identified slot and press Enter to continue.
Enter x to exit. Exiting now leaves the PCI slot
in the removed state.
```

6. Locate the flashing adapter, replace it, and press Enter. The system then displays the message `Replace Operation Complete`.

7. Run **diagmenu**, and select `Task Selection` → `Hot Plug Task` → `PCI Hot Plug Manager` → `Install/Configure Devices Added After IPL`.
8. Press Enter. This calls the **cfgdev** command internally and sets all previously unconfigured devices back to `Available`.
9. If a Fibre Channel adapter is replaced, the settings such as zoning on the Fibre Channel switch and the definition of the WWPN of the replaced adapter to the storage subsystem must be done before the replaced adapter can access the disks on the storage subsystem.

   For IBM DS4000 storage subsystems, switch the LUN mappings back to their original controllers because they might have been distributed to balance I/O load.

12.2 Monitoring dynamic LPAR operations

This section addresses how to monitor resources after dynamic LPAR operations:

- Monitoring dynamic LPAR operations on the Virtual I/O Server
- Monitoring dynamic LPAR operations on AIX
- Monitoring dynamic LPAR operations on IBM i
- Monitoring dynamic LPAR operations on Linux

12.2.1 Monitoring dynamic LPAR operations on the Virtual I/O Server

This section describes some basic monitoring capabilities on a Virtual I/O Server partition for resource changes by dynamic LPAR operations.

Processor and memory configuration monitoring on VIOS

An easy way to get a concise view of the current processor and memory configuration settings for a Virtual I/O Server partition is to use the **nmon** command's LPAR stats (**p**) panel with shows the minimum, desired (online) and

maximum (virtual) processor and memory configuration as shown in Figure 12-27.

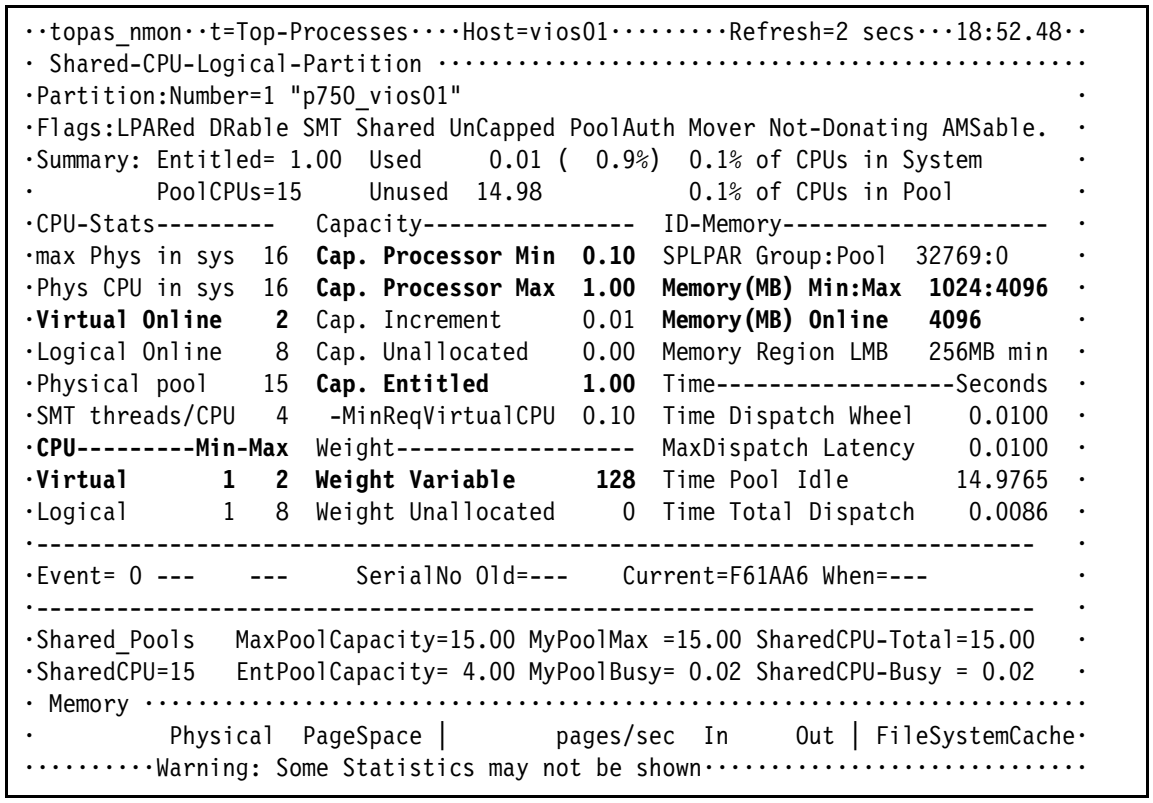


Figure 12-27 nmon LPAR Stats command output

Adapter configuration monitoring on VIOS

After adding or removing physical or virtual adapters to/from the Virtual I/O Server, the configured adapters can be displayed by using the **lsdev -type adapter** command. This command shows both physical and virtual adapters as shown in Example 12-7. You can also use the **lsdev -slots** command, which shows virtual adapters with their slot information only.

Example 12-7 Displaying configured adapters on the Virtual I/O Server

```
$ lsdev -type adapter | grep Adapter
ent0            Available   4-Port 10/100/1000 Base-TX PCI-Express
Adapter (14106803)
ent1            Available   4-Port 10/100/1000 Base-TX PCI-Express
Adapter (14106803)
```

```
ent2             Available   4-Port 10/100/1000 Base-TX PCI-Express
Adapter (14106803)
ent3             Available   4-Port 10/100/1000 Base-TX PCI-Express
Adapter (14106803)
ent4             Available   Virtual I/O Ethernet Adapter (l-lan)
ent5             Available   Virtual I/O Ethernet Adapter (l-lan)
ent6             Available   Shared Ethernet Adapter
fcs0             Available   8Gb PCI Express Dual Port FC Adapter
(df1000f114108a03)
fcs1             Available   8Gb PCI Express Dual Port FC Adapter
(df1000f114108a03)
sissas0          Available   PCI-X266 Planar 3Gb SAS RAID Adapter
vfchost0         Available   Virtual FC Server Adapter
vhost0           Available   Virtual SCSI Server Adapter
vhost1           Available   Virtual SCSI Server Adapter
vhost2           Available   Virtual SCSI Server Adapter
vhost3           Available   Virtual SCSI Server Adapter
vhost4           Available   Virtual SCSI Server Adapter
vsa0             Available   LPAR Virtual Serial Adapter
```

To display PCI slot information for a physical adapter on the Virtual I/O Server, use the **`lsdev -dev <adapter> -slot`** command as shown for the Fibre Channel adapter resource `fcs0` in Example 12-8.

Example 12-8 Displaying slot information for a physical adapter on the Virtual I/O Server

```
$ lsdev -dev fcs0 -slot
# Slot                   Description
Device(s)
U5802.001.0086848-P1-C2  PCI-E capable, Rev 1 slot with 8x lanes  fcs0
fcs1
```

12.2.2 Monitoring dynamic LPAR operations on AIX

This section describes some basic monitoring capabilities on an AIX partition for resource changes by dynamic LPAR operations.

Processor configuration monitoring on AIX

You can monitor the current processor configuration with its entitled capacity and virtual processors after dynamic LPAR operations by using the **lparstat** command as shown in Example 12-9. You can use these commands for both shared and dedicated processors.

Example 12-9 Checking entitled capacity and processors

```
p740_lpar01:/ # lparstat -i | grep "Entitled Capacity"
Entitled Capacity                          : 0.20
Entitled Capacity of Pool                  : 150

p740_lpar01:/ # lparstat -i | grep "Virtual CPUs"
Online Virtual CPUs                        : 2
Maximum Virtual CPUs                       : 5
Minimum Virtual CPUs                       : 1
Desired Virtual CPUs                       : 2
```

If you add 0.3 processing units and three virtual processors dynamically, the output is changed as shown in Example 12-10.

Example 12-10 Checking entitled capacity and processors after dynamic LPAR operation

```
p740_lpar01:/ # lparstat -i | grep "Entitled Capacity"
Entitled Capacity                          : 0.50
Entitled Capacity of Pool                  : 150

p740_lpar01:/ # lparstat -i | grep "Virtual CPUs"
Online Virtual CPUs                        : 5
Maximum Virtual CPUs                       : 5
Minimum Virtual CPUs                       : 1
Desired Virtual CPUs                       : 5
```

Memory configuration monitoring on AIX

The memory configuration settings for an AIX partition can be displayed after dynamic LPAR operations by using the **lparstat** command as shown in Example 12-11.

Example 12-11 Monitoring memory

```
p740_lpar01:/ # lparstat -i | grep Memory
Online Memory                              : 30976 MB
Maximum Memory                             : 32768 MB
Minimum Memory                             : 1024 MB
Memory Mode                                : Shared
```

If you remove 10 GB of memory dynamically, you see the configuration change in the output as shown in Example 12-12.

Example 12-12 Monitoring memory after a dynamic LPAR operation

```
p740_lpar01:/ # lparstat -i | grep Memory
Online Memory                              : 22272 MB
Maximum Memory                             : 32768 MB
Minimum Memory                             : 1024 MB
Memory Mode                                : Shared
```

Adapter configuration monitoring on AIX

After you add or remove physical adapters, you can check the status by using the **lsslot** command as shown in Example 12-13.

Example 12-13 Checking physical adapters on AIX

```
# lsslot -c pci
# Slot                    Description
Device(s)
U78A0.001.DNWHZWR-P1-C1  PCI-E capable, Rev 1 slot with 8x lanes  Empty
U78A0.001.DNWHZWR-P1-C4  PCI-X capable, 64 bit, 266MHz slot       ent1
ent2
U78A0.001.DNWHZWR-P1-C5  PCI-X capable, 64 bit, 266MHz slot       ent3
ent4
```

For servers like Power Systems 740 that do not support PCI hot-pluggable adapters on the main system board, these adapters are not shown with **lsslot -c pci**. Instead, use **lsslot -c phb** to list them as shown in Example 12-14.

Example 12-14 Checking physical adapters on a server with no hot-plug support

```
vios03:/home/padmin # lsslot -c pci

There are no PCI hot plug slots on this system.

vios03:/home/padmin # lsslot -c phb
PHB Name  Description              Device(s)
PHB 10    Logical PCI Host Bridge  pci0 sissas0
PHB 514   Logical PCI Host Bridge  pci1 fcs0 fcs1
PHB 517   Logical PCI Host Bridge  pci2 pci3 pci4 ent0 ent1 pci5 ent2
ent3
```

Virtual adapters on AIX can be found by using the **lsslot -c slot** command as shown in Example 12-15.

Example 12-15 Checking virtual adapters on AIX

```
p740_lpar01:/ # lsslot -c slot
# Slot                     Description       Device(s)
U8205.E6C.06A22ER-V3-C0    Virtual I/O Slot  vsa0
U8205.E6C.06A22ER-V3-C2    Virtual I/O Slot  ent0
U8205.E6C.06A22ER-V3-C33   Virtual I/O Slot  fcs0
U8205.E6C.06A22ER-V3-C43   Virtual I/O Slot  fcs1
```

12.2.3 Monitoring dynamic LPAR operations on IBM i

This section describes some basic monitoring capabilities on an IBM i partition for resource changes by dynamic LPAR operations.

Processor configuration monitoring on IBM i

On IBM i, you can display the current virtual processors and processing capacity by using the WRKSYSACT CL command as shown in Figure 12-28.

```
                        Work with System Activity                   I7PFE
                                                          12/06/12  01:25:30
Automatic refresh in seconds  . . . . . . . . . . . . . . . . . .     5
Job/Task CPU filter . . . . . . . . . . . . . . . . . . . . . . .    .10
Elapsed time . . . . . . :   00:00:02    Average CPU util . . . . :      19.6
Virtual Processors . . . . :       4     Maximum CPU util . . . . . :    27.8
Overall DB CPU util  . . . :      .0     Minimum CPU util . . . . . :     9.1
Average CPU rate . . . . . :   100.0     Current processing capacity:     3.50

Type options, press Enter.
  1=Monitor job   5=Work with job
                                                          Total  Total   DB
     Job or                                          CPU   Sync  Async   CPU
Opt  Task         User        Number  Thread    Pty   Util   I/O   I/O   Util
     ASP010002    QDEXUSER    075963  00000003   90    8.8   829  16187    .0
     ASP010001    QDEXUSER    075962  00000003    9    8.4   710  17304    .0
     QPADEV0001   IDIMMER     075896  00000005    1     .2    14      0    .0
     QGLDPUBA     QDIRSRV     075266  00000001   50     .0    12      2    .0


                                                                      Bottom
F3=Exit   F10=Update list   F11=View 2   F12=Cancel   F19=Automatic refresh
F24=More keys
(C) COPYRIGHT IBM CORP. 1980, 2009.
```

Figure 12-28 IBM i WRKSYSACT command output

Changes for the IBM i processor configuration using dynamic LPAR are logged in the history log (using the DSPLOG QHST CL command) by using message CPI098A as shown in Figure 12-29.

```
                        Additional Message Information

Message ID . . . . . :   CPI098A       Severity . . . . . :   00
Message type . . . . . :   Information
Date sent  . . . . . :   12/06/12      Time sent  . . . . . :   14:55:21

Message . . . . :   Number of processors changed to 5.
Cause . . . . . :   The number of system processors in this partition has
  changed from 4 to 5.



                                                                      Bottom
Press Enter to continue.

F3=Exit   F6=Print   F9=Display message details   F12=Cancel
F21=Select assistance level
```

Figure 12-29 IBM i history log entry for a dynamic LPAR processor change

Detailed logical partition configuration information for processor and memory and processor utilization data can be obtained by using IBM i Collection Services. Use them from the POWER hypervisor for all partitions that are running on the Power Systems server regardless if they are AIX, Virtual I/O Server, IBM i, or Linux. The data is stored in the Collection Services QAPMLPARH database file. To allow this cross-partition performance data collection by IBM i, **Allow**

performance information collection must be enabled with the HMC within the IBM i partition's properties as shown in Figure 12-30.

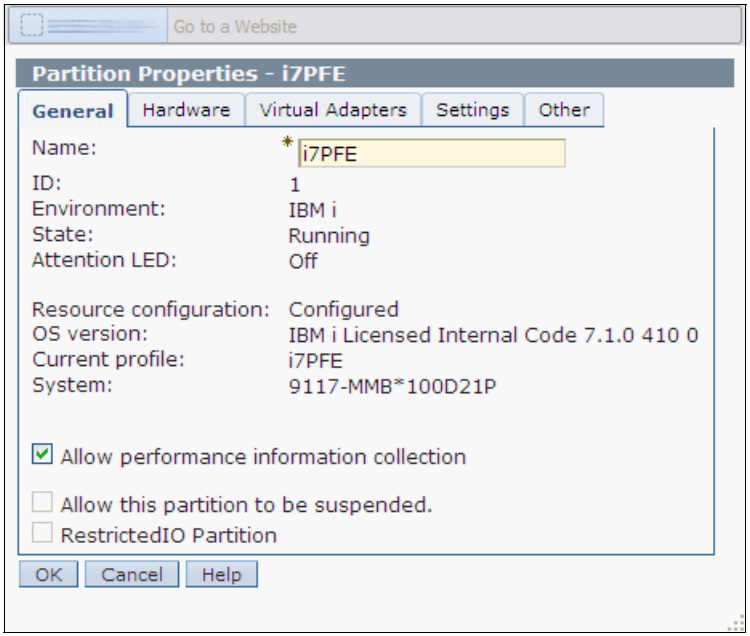


Figure 12-30 Allow performance information collection for IBM i

Using an SQL query against QAPMLPARH as shown in Example 12-16, information about a partition's configuration settings can be displayed.

Example 12-16 IBM i SQL query to display LPAR configuration details

```
SELECT hpshrf, hpvpid, hpvprl, hpvprc, hpvprh, hpprccl, hpprccc,
hpprcch, hpmeml, hpmemc, hpmemh FROM qpfrdata/qapmlparh WHERE
hppid=binary(x'0001') ORDER BY intnum DESC FETCH FIRST 1 ROWS ONLY
```

The output for selected partition ID 1 is shown in Figure 12-31.

```
                                Display Data
                                                Data width . . . . . :     175
Position to line  . . . . .               Shift to column  . . . . . .
....+....1....+....2....+....3....+....4....+....5....+....6....+....7....+....
Shared     Virtual   Minimum     Current     Maximum     Minimum     Current
processor  shared    virtual     virtual     virtual     processing  processing
flag       pool id   processors  processors  processors  capacity    capacity
    1            0           1           4          16         .50        3.50
8....+....9....+...10....+...11....+...12....+...13....+...14....+...15....+...
  Maximum                          Minimum                        Current
  processing                       memory                         memory
  capacity                         (MB)                           (MB)
     16.00                           8,192                          65,536
..10....+...11....+...12....+...13....+...14....+...15....+...16....+...17....+
               Minimum                          Current                    Maximum
               memory                           memory                     memory
               (MB)                             (MB)                       (MB)
                 8,192                            65,536                     262,144
```

Figure 12-31 IBM i QAPMLPARH SQL query output

For more information see *Collecting and displaying CPU utilization for all partitions* in the IBM i Information Center at:

http://publib.boulder.ibm.com/infocenter/iseries/v7r1m0/index.jsp?topic=%2Frzahx%2Frzahxcollectdisplaycpuforallpartitions.htm

Memory configuration monitoring on IBM i

On IBM i, the configured main memory is organized in storage pools for which the current configuration can be displayed by using the WRKSHRPOOL CL command as shown in Figure 12-32.

```
                          Work with Shared Pools
                                                              System:   I7PFE
Main storage size (M)  . :       63784.00

Type changes (if allowed), press Enter.

              Defined    Max   Allocated   Pool  -Paging Option--
Pool         Size (M)  Active  Size (M)     ID   Defined  Current
*MACHINE     27625.13   +++++    27625.13    1   *FIXED   *FIXED
*BASE        36157.86    1827    36157.86    2   *FIXED   *FIXED
*INTERACT    84907.96    4263                    *FIXED
*SPOOL            .25       5                    *FIXED
*SHRPOOL1         .00       0                    *FIXED
*SHRPOOL2         .00       0                    *FIXED
*SHRPOOL3         .00       0                    *FIXED
*SHRPOOL4         .00       0                    *FIXED
*SHRPOOL5         .00       0                    *FIXED
*SHRPOOL6         .00       0                    *FIXED
                                                                    More...
Command
===>
F3=Exit   F4=Prompt   F5=Refresh   F9=Retrieve   F11=Display tuning data
F12=Cancel
```

Figure 12-32 IBM i WRKSHRPOOL command output

Note: For an IBM i partition, dynamically added/removed memory is added to/removed from the *base* memory pool. In the example, this is system pool ID 2 as shown in the WRKSYSSTS or WRKSHRPOOL panel.

Added memory is dynamically distributed to other memory pools when you use the default automatic performance adjustment (QPFRADJ=2 system value setting).

Memory is removed up to the limit of the minimum amount of memory required in the base pool as determined by the base storage pool minimum size (QBASPOOL system value).

Changes for the IBM i memory configuration using dynamic LPAR are logged in the history log using messages CPI098B and CPI098C. Figure 12-33 shows an example of a dynamic LPAR memory increase from 64 GB to 80 GB.

```
Message ID . . . . . :   CPI098B       Severity . . . . . :   00
Message type . . . . . :   Information
Date sent  . . . . . :   12/06/12      Time sent  . . . . . :   20:01:05

Message . . . . :   Main storage size change to 81920.00M in progress.
Cause . . . . . :   A request was made to change the size of main storage for
  this partition to 81920.00 megabytes. The current size of main storage for
  this partition is 63848.00 megabytes.

Message ID . . . . . :   CPI098C       Severity . . . . . :   00
Message type . . . . . :   Information
Date sent  . . . . . :   12/06/12      Time sent  . . . . . :   20:01:07

Message . . . . :   Main storage size changed to 81920.00M.
Cause . . . . . :   The size of main storage in this partition has changed to
  81920.00 megabytes.
```

Figure 12-33 IBM i history log entry for a dynamic LPAR memory addition

Adapter configuration monitoring on IBM i

Adapter information about IBM i can be displayed for communication adapters like Ethernet using the WRKHDWRSC *CMN command, or for storage adapters using the WRKHDWRSC *STG command as shown in Figure 12-34.

```
                        Work with Storage Resources
                                                            System:   F001AA6P
Type options, press Enter.
  7=Display resource detail   9=Work with resource

Opt  Resource        Type-model  Status               Text
     CMB01            290A-001   Operational          Storage Controller
       DC01           290A-001   Operational          Storage Controller
     CMB02            6B25-001   Operational          Storage Controller
       DC02           6B25-001   Operational          Storage Controller
     CMB03            6B25-001   Operational          Storage Controller
       DC03           6B25-001   Operational          Storage Controller
     CMB04            268C-001   Operational          Combined function IOP
       DC04           6B02-001   Operational          Storage Controller



                                                                        Bottom
F3=Exit   F5=Refresh   F6=Print   F12=Cancel
```

Figure 12-34 IBM i WRKHDWRSC command output

Note: An IBM i adapter resource that has been removed from the partition shows up in the *Not connected* status. If that resource will not be used again, remove the corresponding logical resource for the adapter by using the IBM i System Service Tools' Hardware Service Manager.

Using option 7=Display resource detail shows detailed adapter information as shown in Figure 12-35. This example shows a virtual Fibre Channel client adapter (type 6B25) in slot 63 with WWPN C050760303980094 connected to an 8 Gb SAN fabric with one connected storage target device.

```
                          Display Resource Detail
                                                              System:   FOO1AA6P
 Resource name  . . . . . . . . :   DC03
 Text . . . . . . . . . . . . . :   Storage Controller
 Type-model . . . . . . . . . . :   6B25-001
 Serial number  . . . . . . . . :   00-00000
 Part number  . . . . . . . . . :

 Location :   U8233.E8B.061AA6P-V6-C63

 Logical address:
  SPD bus:
   System bus                      255
   System board                    128
   System card                     63

Location :   U8233.E8B.061AA6P-V6-C63

  Storage:
   I/O bus                         127
 Port                              0
  World wide port name             C050760303980094
  Port status                      Active
  Protocol                         Switched fabric
  Port speed (Gbps)                8
  Number of targets detected       1


                                                                        Bottom
 Press Enter to continue.

 F3=Exit   F5=Refresh   F6=Print   F12=Cancel
```

*Figure 12-35 IBM i WRKHDWRSC *STG displaying the adapter resource*

12.2.4 Monitoring dynamic LPAR operations on Linux

This section describes some basic monitoring capabilities on a Linux partition for resource changes by dynamic LPAR operations.

Processor configuration monitoring on Linux

When new processors are added by using dynamic LPAR to a Linux partition, you see messages like those shown in Example 12-17. This occurs on the Linux client when you run the **tail -f /var/log/messages** command.

Example 12-17 Linux finds new processors

```
Dec  2 11:26:08 linuxlpar : drmgr: /usr/sbin/drslot_chrp_cpu -c cpu -a -q 60 -p
ent_capacity -w 5 -d 1
Dec  2 11:26:08 linuxlpar : drmgr: /usr/sbin/drslot_chrp_cpu -c cpu -a -q 1 -w
5 -d 1
Dec  2 11:26:08 linuxlpar kernel: Processor 2 found.
Dec  2 11:26:09 linuxlpar kernel: Processor 3 found.
```

In addition to the messages in the log directory file, configuration changes by dynamic LPAR can be monitored by running **cat /proc/ppc64/lparcfg**. Example 12-18 shows that the partition had 0.5 processor as its entitled capacity.

Example 12-18 The lparcfg command before adding processors dynamically

```
lparcfg 1.7
serial_number=IBM,02101F170
system_type=IBM,9117-MMA
partition_id=7
R4=0x32
R5=0x0
R6=0x80070000
R7=0x800000040004
BoundThrds=1
CapInc=1
DisWheRotPer=5120000
MinEntCap=10
MinEntCapPerVP=10
MinMem=128
MinProcs=1
partition_max_entitled_capacity=100
system_potential_processors=16
DesEntCap=50
DesMem=2048
DesProcs=1
DesVarCapWt=128
DedDonMode=0

partition_entitled_capacity=50
group=32775
system_active_processors=4
pool=0
pool_capacity=400
```

```
pool_idle_time=0
pool_num_procs=0
unallocated_capacity_weight=0
capacity_weight=128
capped=0
unallocated_capacity=0
purr=19321795696
partition_active_processors=1
partition_potential_processors=2
shared_processor_mode=1
```

The LPAR configuration attributes that are associated with addition/deletion of processors are `partition_entitled_capacity` and `partition_ative_processor`. The change in the values of **`lparcfg`** attributes after the addition of 0.1 processor is shown in Example 12-19.

Example 12-19 The lparcfg command after adding 0.1 processor dynamically

```
lparcfg 1.7
serial_number=IBM,02101F170
system_type=IBM,9117-MMA
... omitted lines ...
partition_entitled_capacity=60
group=32775
system_active_processors=4
pool=0
pool_capacity=400
pool_idle_time=0
pool_num_procs=0
unallocated_capacity_weight=0
capacity_weight=128
capped=0
unallocated_capacity=0
purr=19666496864
partition_active_processors=2
partition_potential_processors=2
shared_processor_mode=1
```

A dynamic removal of processors on a Linux partition is logged as shown by the **`dmesg`** command is shown in Example 12-20.

Example 12-20 Ready to die message

```
IRQ 18 affinity broken off cpu 0
IRQ 21 affinity broken off cpu 0
```

```
cpu 0 (hwid 0) Ready to die...
cpu 1 (hwid 1) Ready to die...
```

Memory configuration monitoring on Linux

Before and after you dynamically add or remove memory to a Linux partition, you can obtain the partition's current memory information by running the **cat /proc/meminfo** command as shown in Example 12-21.

Example 12-21 Display of total memory in the partition before you add memory

```
[root@VIOCRHEL52 ~]# cat /proc/meminfo | head -3
MemTotal:      2057728 kB
MemFree:       1534720 kB
Buffers:        119232 kB
```

Adapter configuration monitoring on Linux

After you add or remove physical adapters, you can check the status by using the **lsslot** command as shown in Example 12-22

Example 12-22 Checking physical adapters on Linux

```
[root@p750_lpar02 ~]# lsslot -c pci
# Slot                     Description                     Device(s)
U5802.001.0086848-P1-C10  PCI-E capable, Rev 1, 8x lanes  Empty
```

Entry level Power Systems do not support PCI hot-pluggable adapters on the main system board. These adapters are not shown with **lsslot -c pci**. Instead, use **lsslot -c phb** to list them as shown in Example 12-23.

Example 12-23 Checking physical adapters on a server with no hot-plug support

```
[root@p740_lpar05 ~]# lsslot -c pci
lsslot: There are no PCI hot plug slots on this system.

[root@p740_lpar05 ~]# lsslot -c phb
PHB name  OFDT Name            Slot(s) Connected
PHB 513   /pci@800000020000201
```

Virtual adapters on Linux can be found by using the **lsslot -c slot** command as shown in Example 12-24.

Example 12-24 Checking virtual adapters on Linux

```
[root@p740_lpar05 ~]# lsslot -c slot
# Slot                     Description       Linux Name    Device(s)
```

```
U8205.E6C.06A22ER-V6-C0   Virtual I/O Slot  30000000      vty
U8205.E6C.06A22ER-V6-C2   Virtual I/O Slot  30000002      l-lan
U8205.E6C.06A22ER-V6-C88  Virtual I/O Slot  30000058      v-scsi
```

Tip: The **lsslot** command is part of the powerpc-utils RPM package. This package also contains the informative commands **lsdevinfo**, **lsvio**, **ls-veth**, **ls-vscsi**, and **ls-vdev**.

13

Partition Suspend and Resume

The Virtual I/O Server provides the Partition Suspend and Resume capability to client logical partitions within the IBM POWER7 Systems™. Suspend and Resume operations allow the partition's state to be suspended and resumed later.

A suspended logical partition indicates that it is in a hibernated state, and all of its resources can be used by other partitions. A resumed logical partition means that the partition's state has been successfully restored from a suspend operation. A partition's state is stored in a paging space on a persistent storage device.

This chapter includes the following sections:

- Managing Suspend and Resume
- Monitoring Suspend and Resume

13.1 Managing Suspend and Resume

This section describes listing, adding, or removing devices in the reserved storage device pool. It also addresses shutting down a suspended partition, and recovering a suspended or resumed partition. Common validation error messages are also described.

Tip: Partition Suspend and Resume are supported for IBM i V7R1 TR2 and HMC V7R7.3 or later.

For more information, see the *IBM Power Systems Hardware Information Center* for Suspend and Resume requirements and configuration:

http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7hat/iphatphibreqs.htm

Attention: A reserved storage device pool must be used for Partition Suspend and Resume capability in a PowerVM Standard Edition or Enterprise Edition environment. In a PowerVM Enterprise Edition environment, the configured reserved storage device pool can also be used for Active Memory Sharing.

This chapter includes the following sections:

- Reserved storage device pool management:
  - Listing volumes in the reserved storage device pool
  - Adding volumes to the reserved storage device pool
  - Removing a volume from the reserved storage device pool
- Suspend and resume operations:
  - Suspending a partition
  - Shutting down a suspended partition
  - Recovering a suspended or resumed partition
  - Correcting validation errors

13.1.1 Reserved storage device pool management

This section addresses the management of reserved storage device pools.

Listing volumes in the reserved storage device pool

To list physical volumes in the reserved storage device pool by using the Hardware Management Console (HMC), complete these steps:

1. On the HMC, select the managed system where the reserved storage device pool is located.
2. Click **Configuration** → **Virtual Resources** → **Reserved Storage Device Pool Management** as shown in Figure 13-1.

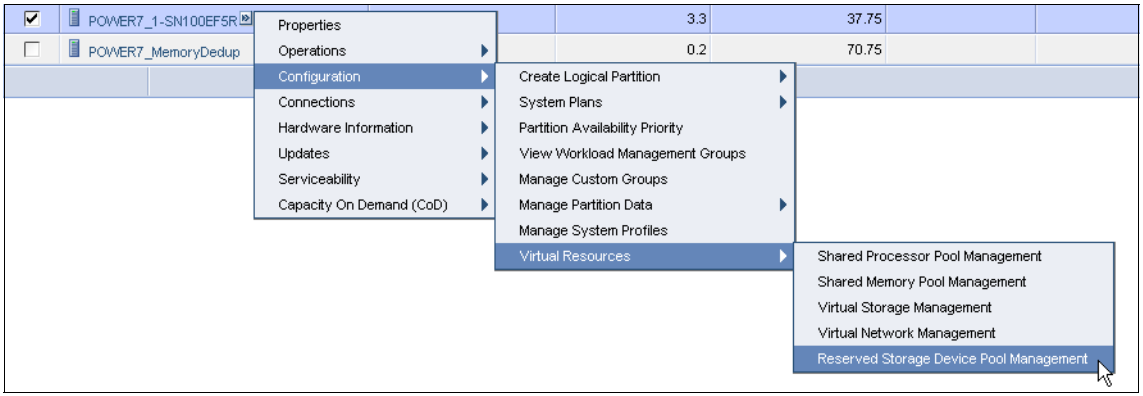


Figure 13-1 Reserved storage device pool management access menu

The list of devices in the reserved storage device pool is displayed as shown in Figure 13-2.

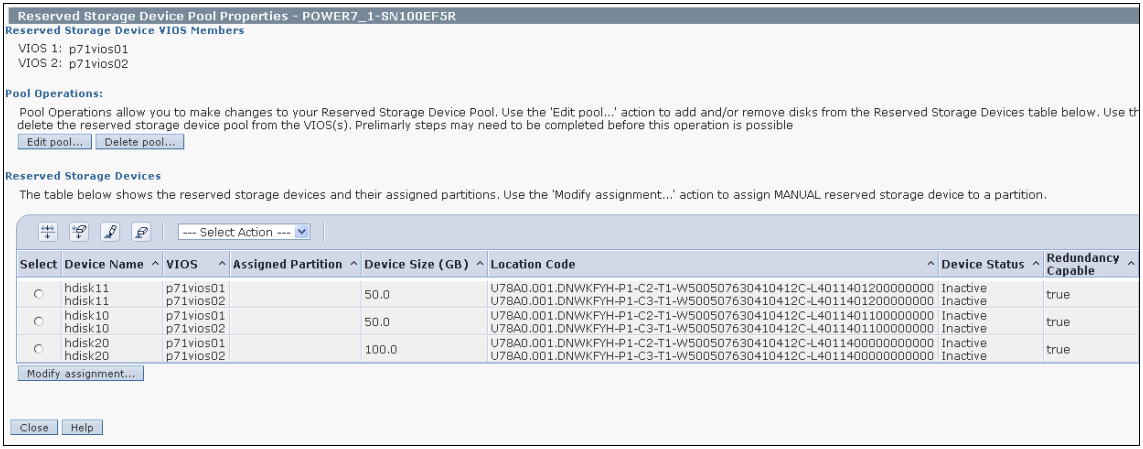


Figure 13-2 Reserved storage device pool device list

3. From the HMC command-line interface, run **lshwres** with the **--rsubtype rsdev** flag. This command shows the reserved storage device that is used to save suspension data for partition as shown in Example 13-1.

Example 13-1 lshwres output that shows reserved storage device properties

```
hscroot@hmc6:~> lshwres -r rspool -m POWER7_1-SN100EF5R --rsubtype rsdev
device_name=hdisk11,vios_name=p71vios01,vios_id=1,size=51200,type=phys,state=Ina
ctive,phys_loc=U78A0.001.DNWKFYH-P1-C2-T1-W500507630410412C-L4011401200000000,is
_redundant=1,redundant_device_name=hdisk11,redundant_vios_name=p71vios02,redunda
nt_vios_id=2,redundant_state=Inactive,redundant_phys_loc=U78A0.001.DNWKFYH-P1-C3
-T1-W500507630410412C-L4011401200000000,lpar_id=none,device_selection_type=auto
device_name=hdisk10,vios_name=p71vios01,vios_id=1,size=51200,type=phys,state=Ina
ctive,phys_loc=U78A0.001.DNWKFYH-P1-C2-T1-W500507630410412C-L4011401100000000,is
_redundant=1,redundant_device_name=hdisk10,redundant_vios_name=p71vios02,redunda
nt_vios_id=2,redundant_state=Inactive,redundant_phys_loc=U78A0.001.DNWKFYH-P1-C3
-T1-W500507630410412C-L4011401100000000,lpar_id=none,device_selection_type=auto
device_name=hdisk20,vios_name=p71vios01,vios_id=1,size=102400,type=phys,state=In
active,phys_loc=U78A0.001.DNWKFYH-P1-C2-T1-W500507630410412C-L4011400000000000,i
s_redundant=1,redundant_device_name=hdisk20,redundant_vios_name=p71vios02,redund
ant_vios_id=2,redundant_state=Inactive,redundant_phys_loc=U78A0.001.DNWKFYH-P1-C
3-T1-W500507630410412C-L4011400000000000,lpar_id=none,device_selection_type=auto
```

Adding volumes to the reserved storage device pool

To add a physical volume to the reserved storage device pool by using the HMC, complete these steps:

1. On the HMC, select the managed system where the reserved storage device pool is located.
2. Click **Configuration** → **Virtual Resources** → **Reserved Storage Device Pool Management** as shown in Figure 13-1 on page 519.

3. Click **Edit Pool** as shown in Figure 13-3.

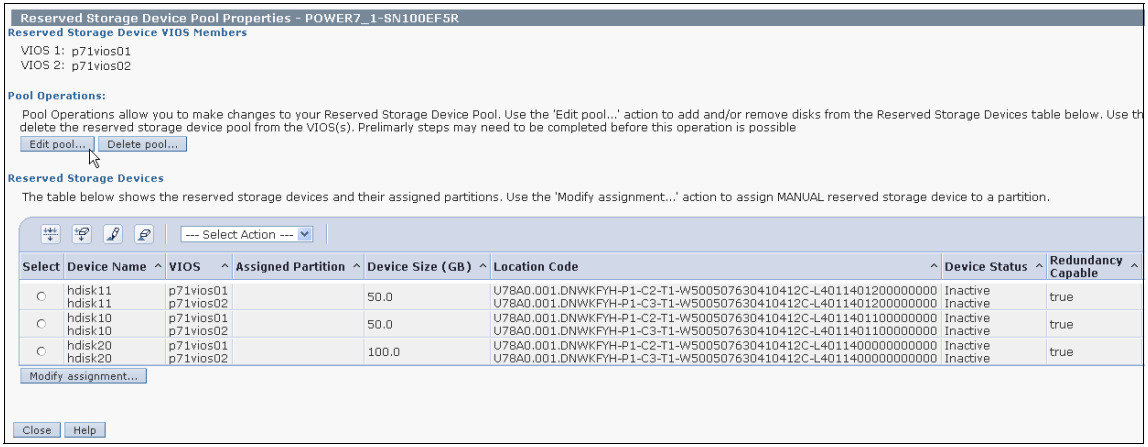


Figure 13-3 Edit pool operation

4. Click **Select Device(s)** as shown in Figure 13-4.

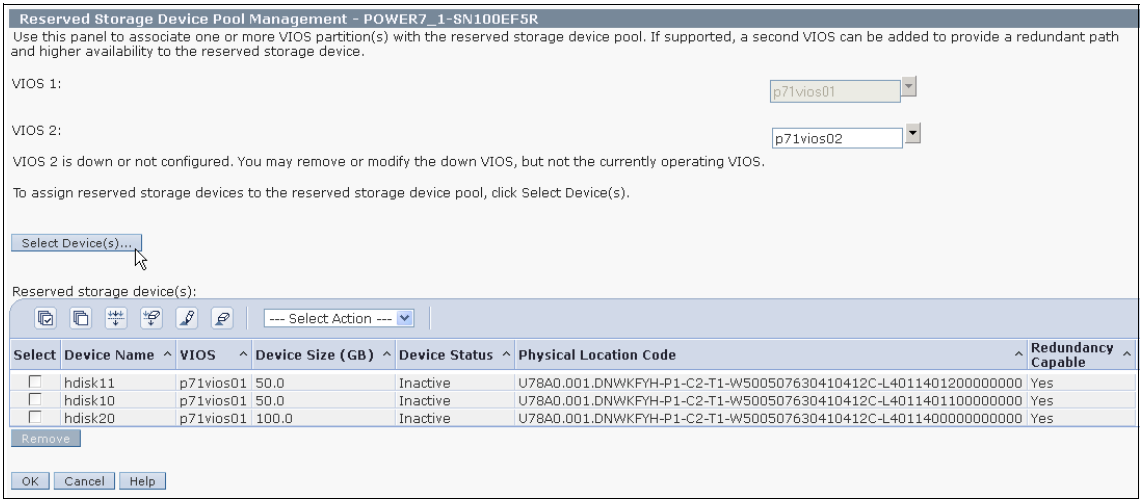


Figure 13-4 Reserved storage device pool management device

5. Select the device type, and optionally select the minimum and maximum devices size, as shown in Figure 13-5. Then, click **Refresh** to display the list of available devices.

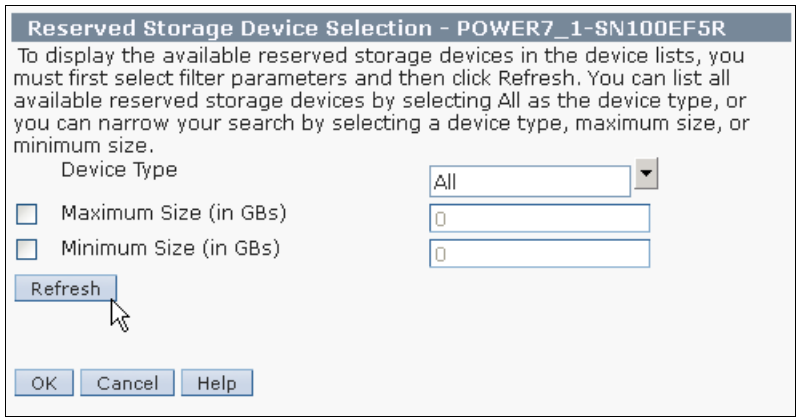


Figure 13-5 Reserved storage device pool management device list selection

Important: The size of the volume must be at least 110% of the maximum memory that is specified in the profile for the suspending partition.

6. Select the devices to add to the reserved storage device pool as shown in Figure 13-6, then click **OK**.

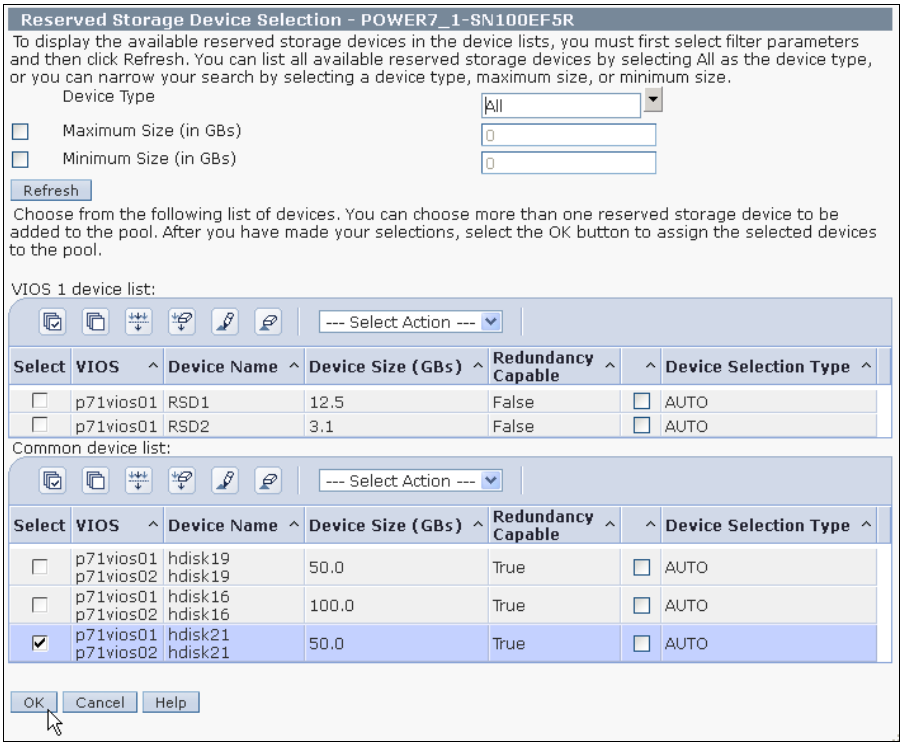


Figure 13-6 Reserved storage device pool management device selection

7. Review the list of devices in the reserved storage device pool as shown in Figure 13-7. Click **OK** to complete the device addition operation.

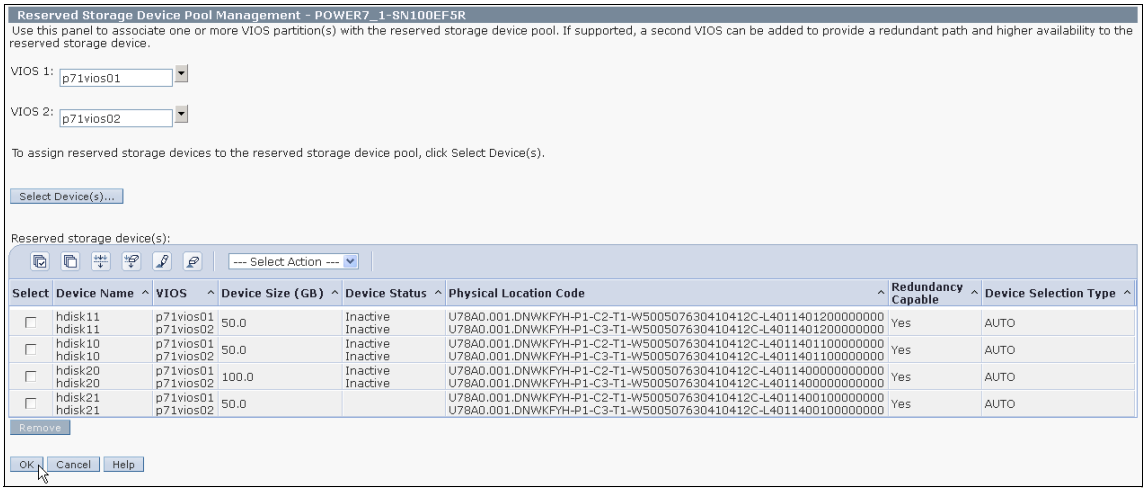


Figure 13-7 Adding a device to the reserved storage device pool validation

You can now list the devices as explained in “Listing volumes in the reserved storage device pool” on page 519.

Removing a volume from the reserved storage device pool

To remove a physical volume from the reserved storage device pool by using the HMC, complete these steps:

1. On the HMC, select the managed system where the reserved storage device pool is located.
2. Click **Configuration** → **Virtual Resources** → **Reserved Storage Device Pool Management** as shown in Figure 13-1 on page 519.

3. Click **Edit Pool** as shown in Figure 13-8.

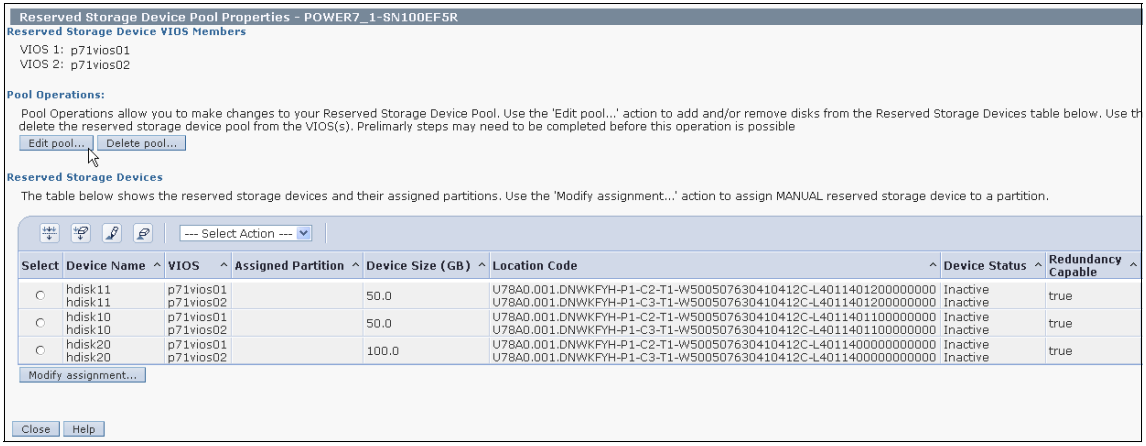


Figure 13-8 Reserved storage device pool management

4. Select the device that you want to remove, then click **Remove** as shown in Figure 13-9.

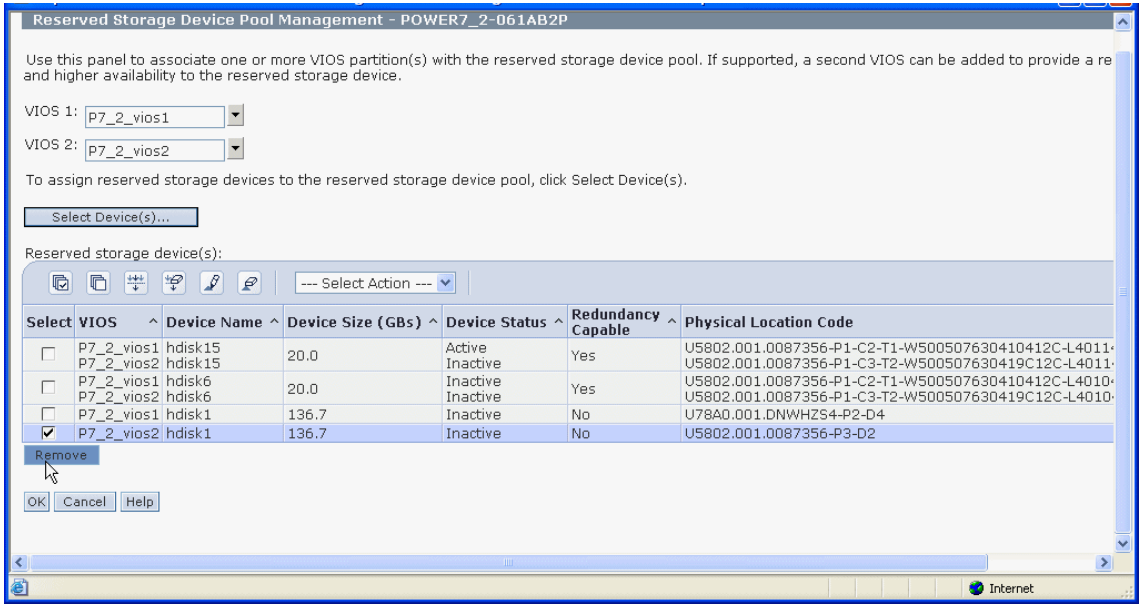


Figure 13-9 Reserved storage device pool management device

5. Review the list of devices in the reserved storage device pool as shown in Figure 13-10, then click **OK** to complete the device removal operation.

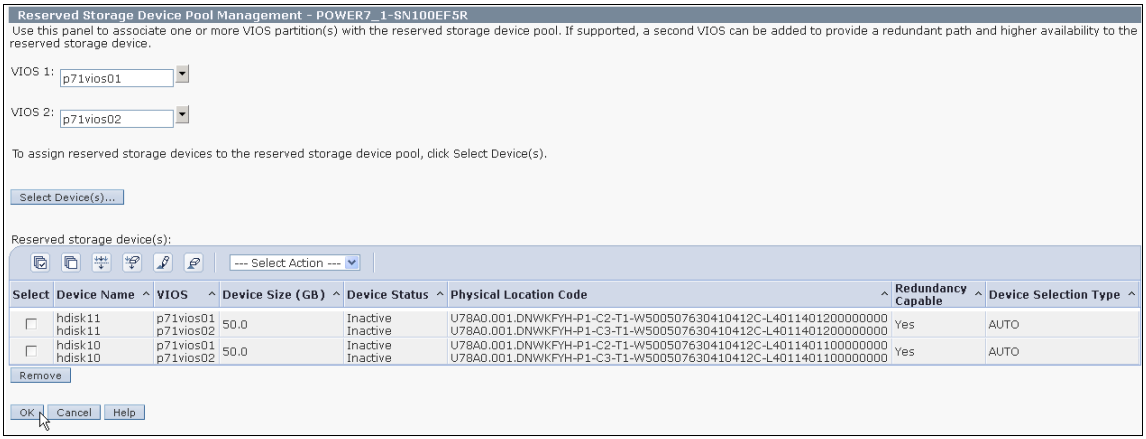


Figure 13-10 Removing a device from reserved storage device pool validation

You can now list the devices and volumes in the reserved storage device pool. For more information, see “Listing volumes in the reserved storage device pool” on page 519.

13.1.2 Suspend and resume operations

Suspending a partition

The suspend operation removes all virtual device mappings from the Virtual I/O Servers for the suspended partition. In addition, the suspend operation removes all virtual server adapters that are used by the suspended partition from the Virtual I/O Servers.

The physical volumes that are used as backing devices by the suspended partition are displayed as volumes that are available for use.

Important: Before suspending a partition, make sure to unconfigure any virtual SCSI optical or tape devices and vary off any NPIV tape devices from IBM i. Otherwise, the suspend validation will fail.

To suspend a running partition, complete these steps:

1. On the HMC, select the partition that you want to suspend.
2. Click **Operation** → **Suspend Operations** → **Suspend** as shown in Figure 13-11.

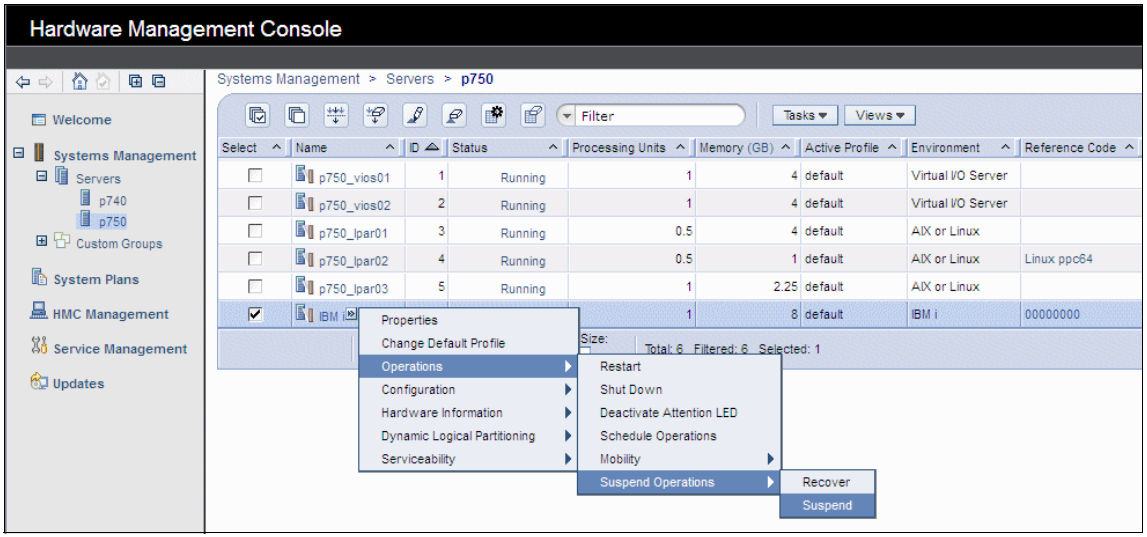


Figure 13-11 Selecting the Suspend operation

3. Select the Virtual I/O Servers and click **Suspend** as shown in Figure 13-12.

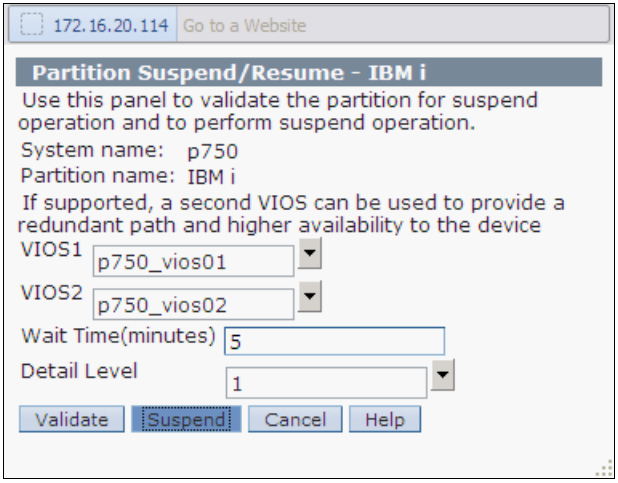


Figure 13-12 Options for partition validate and suspend

Status per activity is shown in a separate window. A validation is done during this phase. If the validation fails, see “Correcting validation errors” on page 535 for some guidance on how to correct typical validation errors. The Suspend status window is shown in Figure 13-13.

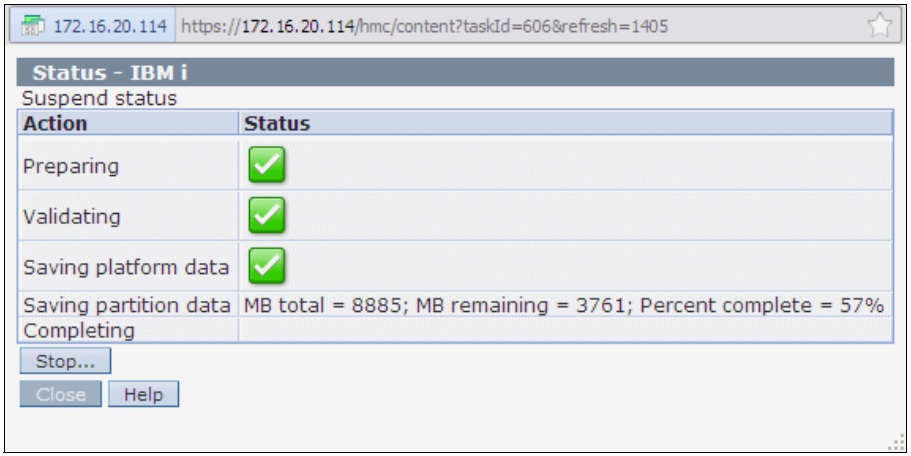


Figure 13-13 Activity status window

When all steps are completed, you receive a confirmation window as shown in Figure 13-14.

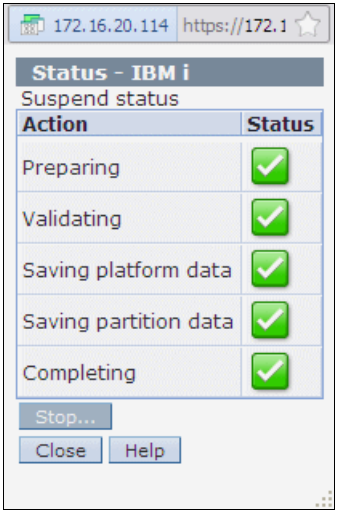


Figure 13-14 Suspend final status

The partition is now suspended and has a status of **Suspended** as shown in Figure 13-15.

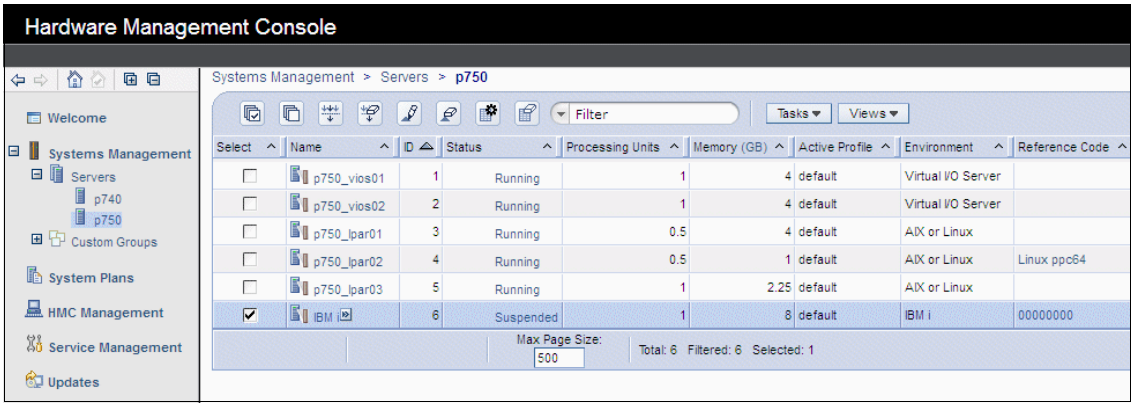


Figure 13-15 HMC suspended partition status

You can also perform the suspend operation from the HMC command line by completing the following steps:

1. Run the **`chlparstate`** command to suspend the partition `p71ibmi08` as shown in Example 13-2.

Example 13-2 Suspending partition p71ibmi08 from the HMC command line

```
hscroot@hmc6:~> chlparstate -o suspend -m p750 -p "IBM i"
```

2. Using the **`lssyscfg`** command, view the state of the partition `p71ibmi08` as shown in Example 13-3.

Example 13-3 Listing state for partition IBM i from the HMC command line

```
hscroot@hmc8:~> lssyscfg -r lpar -m p750 -F name,state --filter
"lpar_names=IBM i"
IBM i,Suspended
```

Resuming a suspended partition

To resume a suspended partition from the HMC, complete these steps:

1. On the HMC, select the partition that you want to suspend.
2. Click **Operation** → **Suspend Operations** → **Resume** as shown Figure 13-16.

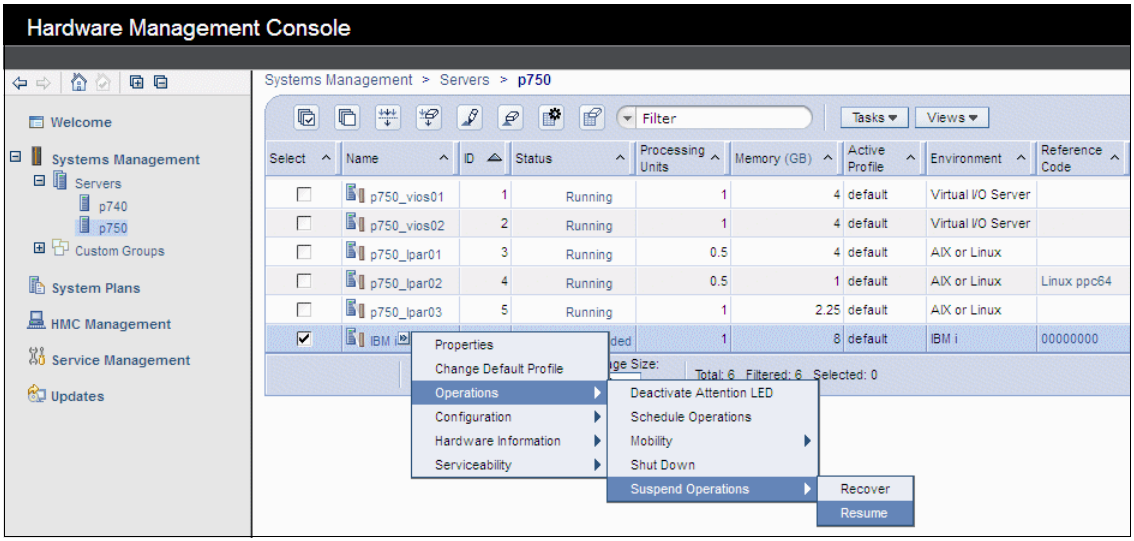


Figure 13-16 Selecting the Resume operation

3. Click **Resume** as shown in Figure 13-17

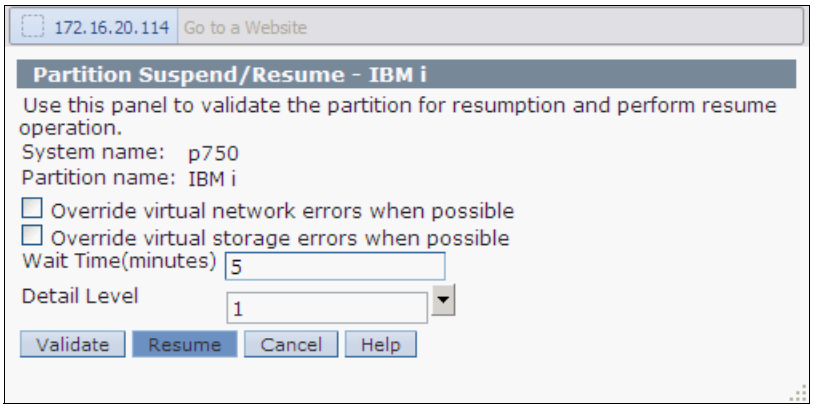


Figure 13-17 Running the Resume operation

The progress of the resume operation is shown in a separate window as shown in Figure 13-18.

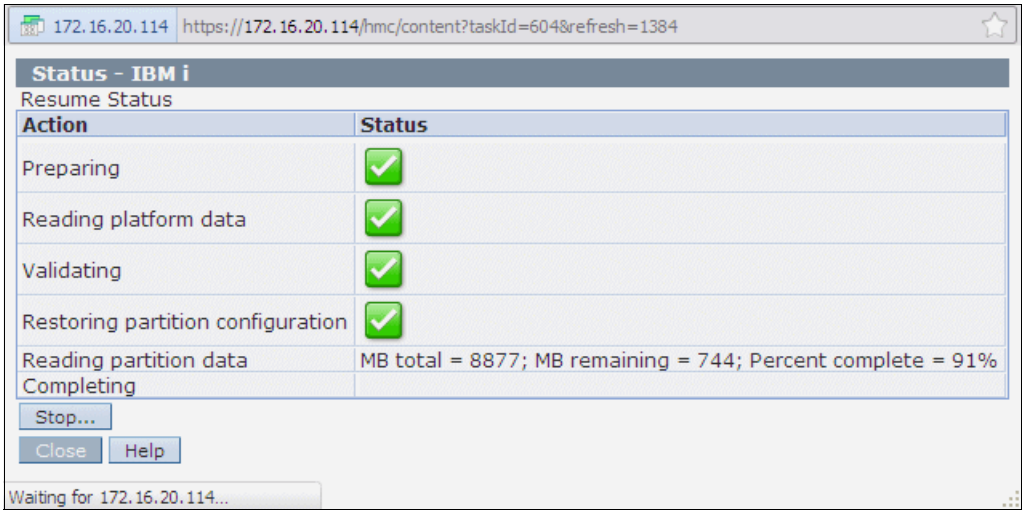


Figure 13-18 Resume operation progress

After the operation is complete, you receive a completion window as shown in Figure 13-19.

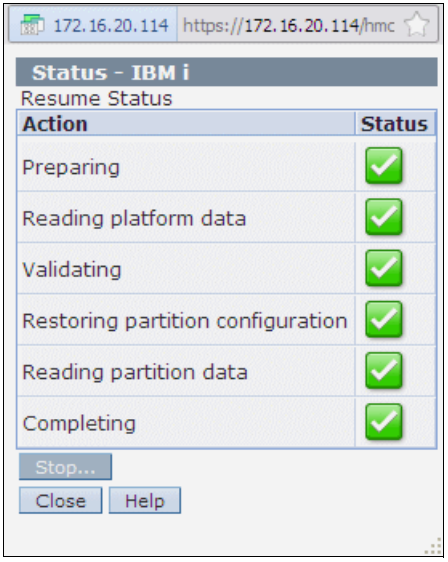


Figure 13-19 Resume operation completion

You can also perform the resume operation from the HMC command line by completing the following steps:

1. Run the **chlparstate** command to resume the partition `IBM i` as shown in Example 13-4.

Example 13-4 Resuming partition IBM i from the HMC command line

```
hscroot@hmc8:~> chlparstate -o resume -m p750 -p "IBM i"
```

2. Using the **lssyscfg** command**,** view the state of the partition `IBM i` as shown in Example 13-5.

Example 13-5 Listing state for partition IBM i from the HMC command line

```
hscroot@hmc8:~> lssyscfg -r lpar -m p750 -F name,state --filter
"lpar_names=IBM i"
IBM i,Running
```

Shutting down a suspended partition

After it is in a suspended state, a partition can be changed with one of these operations:

Resumed	Returns partition to the state it was in before it was suspended.
Shutdown	Invalidates the suspend state and moves the partition to a state of powered off. This operation has a similar impact as an immediate shutdown on the application hosted by the partition, and the partition itself. If the storage device that contains the partition state is available, all saved virtual server adapter configuration entries are restored.
Migrated	Allows a partition to be moved from one server to another. This function uses PowerVM Live Partition Mobility, which requires PowerVM Enterprise Edition.

This section describes how to shut down a suspended partition from the HMC command-line interface.

Important:

- Avoid shutting down a suspended partition. In this case, resume the partition, then perform an OS shutdown on the running partition if possible.
- When a partition is suspended, there are two types of shutdowns:

Normal	HMC reconfigures all virtual server adapters at the shutdown of the suspended partition.
Force	Available if virtual server adapter reconfiguration faces an unrecoverable error.

From HMC command-line interface, complete these steps:

1. Using the **chlparstate** command, shut down the suspended partition `IBM i` with the default `normal` option as shown in Example 13-6.

Example 13-6 Shutting down and suspending a partition

```
hscroot@hmc6:~> chlparstate -m p750 -o shutdown -p "IBM i"
```

2. Using **lssyscfg** command**,** view the state of the partition `p71ibmi08` as shown in Example 13-7.

Example 13-7 Verifying the state of the partition

```
hscroot@hmc6:~> lssyscfg -r lpar -m p750 -F name,state --filter "IBM i"
IBM i,Not Activated
```

Tips:

- The shutdown operation of a suspended partition with the `Normal` option recreates all virtual server adapters and all virtual device mappings used by the suspended partition on the Virtual I/O Servers.
- If the partition is not a shared memory partition, all devices used to store partition suspend data are released.

Recovering a suspended or resumed partition

This section describes how to recover a partition from a failed Suspend or Resume operation.

The recover operation might have to be issued in the following circumstances:

- Suspend or Resume is taking a long time and the user ends the operation abruptly.
- The user is not able to cancel a Suspend or Resume operation successfully.
- Initiating a Suspend or Resume operation results in an extended error that indicates that the partition state is not valid.

When issuing a recover, the HMC determines the last successful step in the previous operation from the progress data. Both suspend and resume store the operation progress in both the HMC and POWER Hypervisor.

Depending on the last successful step, the HMC either completes or rolls back the operation.

Exception: If no progress data is available, the recover operation must be run with the **`force`** option. The HMC will recover as much data as possible.

To run the recover operation on the HMC, complete the following steps:

1. Select the partition to recover.
2. Click **Operations** → **Suspend Operations** → **Recover** as shown in Figure 13-20.

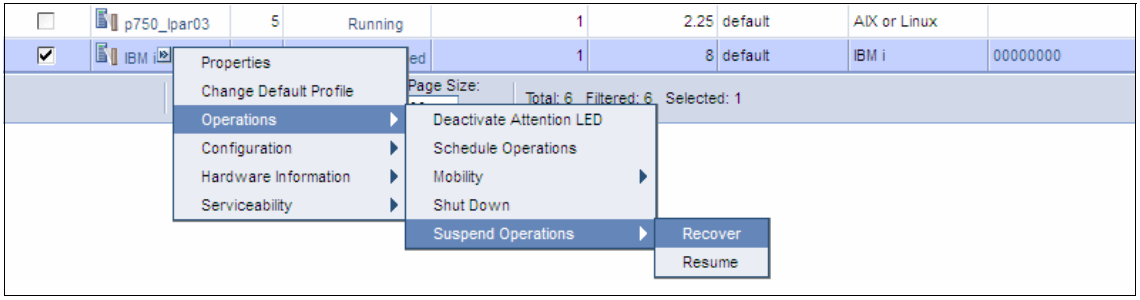


Figure 13-20 Recovering a suspended partition

3. Select the type of operation you want to recover from in the **Target Operation** menu and click **OK** as shown in Figure 13-21.

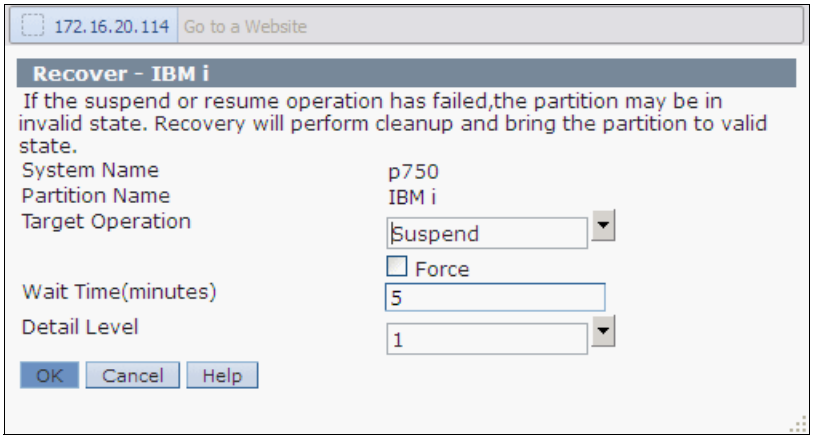


Figure 13-21 Partition recover operation

After completing the recover operation, the partition will return to a **Running** or a **Suspended** state. This state depends on the last successful step and the type of operation that you are recovering from.

Correcting validation errors

Table 13-1 shows a list of the most common validation errors and how to correct those errors.

Table 13-1 Common Suspend and Resume validation errors

Message ID	Validation error message	Correction
HSC0A929	There is no non-redundant device available in the reserved storage device pool that can be used by this partition. This partition requires a device with a size of at least 4318 MB. Add a device of at least that size to the reserved storage device pool, then try the operation again.	There is no device available in the reserved storage pool to run the Suspend operation. Add a physical volume with the required size to the reserved storage device pool.

Message ID	Validation error message	Correction
HSCLA930	There is no redundant device available in the reserved storage device pool that can be used by this partition. This partition requires a device with a size of at least 4318 MB. Add a device of at least that size to the reserved storage device pool, then try the operation again.	There is no redundant device available in the reserved storage pool to run the Suspend operation. Add a redundant physical volume that can be seen by both Virtual I/O Servers in the reserved storage device pool.
HSCLA27C	The operation to get the physical device location for adapter U8233.E8B.061AB2P-V1-C36 on the virtual I/O server partition P7_2_vios1 has failed. The partition command is: migmgr -f get_adapter -t vscsi -s U8233.E8B.061AB2P-V1-C36 -w 13857705808736681994 -W 13857705808736681995 -d 1 The partition standard error is: child process returned error	There can be several reasons for this error: ► A virtual optical device is configured for the client partition. Unconfigure the device using the **rmdev** command on the Virtual I/O Server. ► The `rootvg` is located on a logical volume that is a non-supported setup. ► The `rootvg` is backed by a device from a shared storage pool that is currently not supported ► The backing device is configured with a SCSI reserve policy. Unconfigure the corresponding virtual target device and use the **chdev** command to change the reserve policy to `no_reserve`. ► During a suspend operation in an NPIV environment, there is no SAN zoning of the virtual Fibre Channel WWPNs. Perform the SAN zoning of the virtual Fibre Channel WWPNs.

Message ID	Validation error message	Correction
HSCLA319	The migrating partition's virtual Fibre Channel client adapter 36 cannot be hosted by the existing Virtual I/O Server (VIOS) partitions on the destination managed system. To migrate the partition, set up the necessary VIOS host on the destination managed system, then try the operation again.	The resume operation in an NPIV environment requires proper zoning of the WWPNs. Ensure both WWPNs are in the same zone.

If the reason for the failed validation is still unclear, the configuration log on the Virtual I/O Server can be checked for any errors. Use the **alog** command as shown in Example 13-8.

Example 13-8 Virtual I/O Server configuration log

```
$ oem_setup_env
# alog -ot cfg > /tmp/cfglog
# more /tmp/cfglog

CS 21102768 10813540 /usr/sbin/migmgr -f get_adapter -t vscsi -s
U8205.E6C.06A22ER-V1-C36 -d 1
C4 21102768 Running method '/usr/lib/methods/mig_vscsi'
CS 21102768 10813540 12:06:11 mig_vscsi.c 620 /usr/sbin/migmgr -f
get_adapter -t vscsi -s U8205.E6C.06A22ER-V1-C36 -d 1
...
L4 21102768 12:06:11 fp_get_pci_slots.c 808 fp_find_of_nodes:
children_present=0
C0 21102768 12:06:11 vsmig_util.c 1826 original pipe_ctrl from RMC =0x1
C0 21102768 12:06:11 vscsi_get_adapter.c 1035 ERROR: cannot migrate
reserve type single_path
C0 21102768 12:06:11 mig_vscsi.c 695 leaving mig_vscsi fn= get_adapter,
rc= 80
```

In this example, the configuration log clearly shows that the validation error was caused by a device with `single_path` reservation policy. This policy needs to be changed by using the **chdev** command to `no_reserve` policy for the suspend (validate) operation to succeed.

13.2 Monitoring Suspend and Resume

This section addresses monitoring of Suspend and Resume operations.

13.2.1 Monitoring Suspend and Resume operations on the HMC

During Suspend and Resume operations, you can monitor the progress on the HMC as shown in the Figure 13-22.

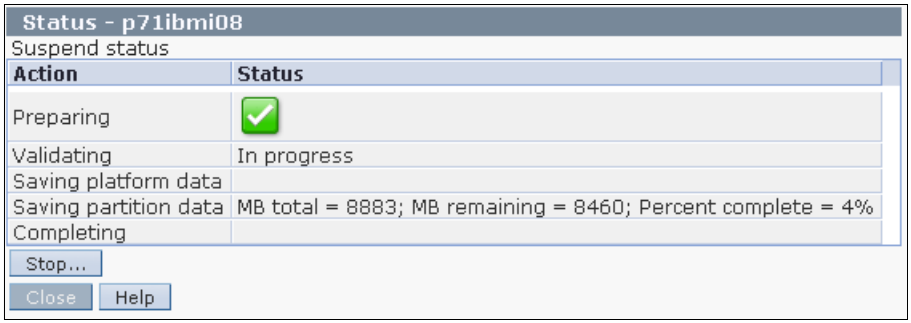


Figure 13-22 Progress status of Suspend and Resume

If errors are encountered, error messages like the one shown in Figure 13-23 are displayed.

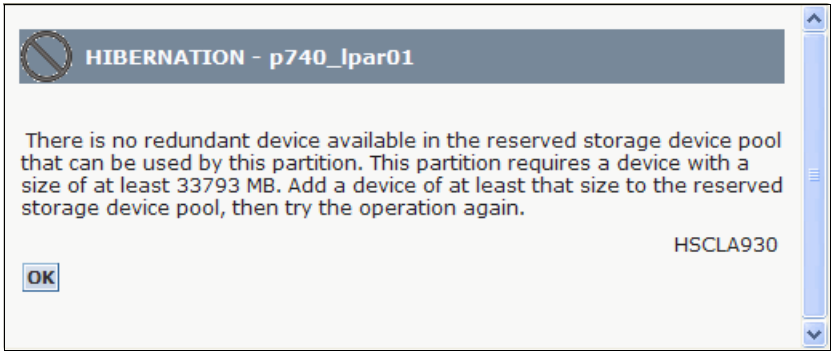


Figure 13-23 Error messages for Suspend and Resume operations

The actual step at which the error occurred is shown in Figure 13-24.

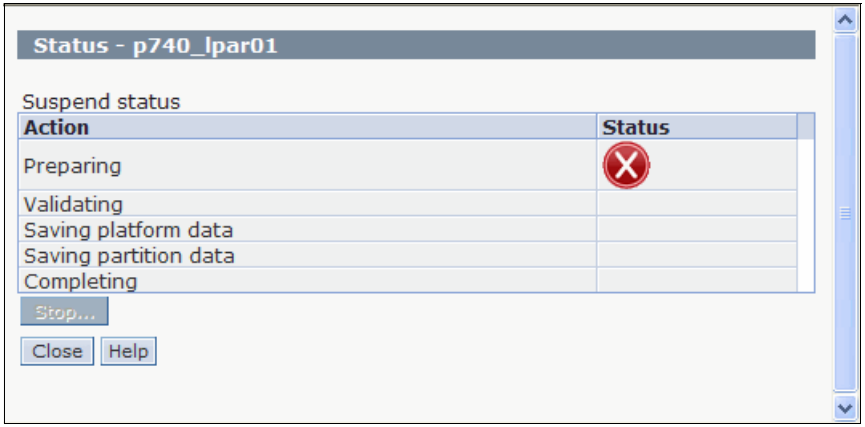


Figure 13-24 Error Box for Suspend and Resume

Carefully read and check the messages (Figure 13-25). After you fix the problems, run the validate operation again.

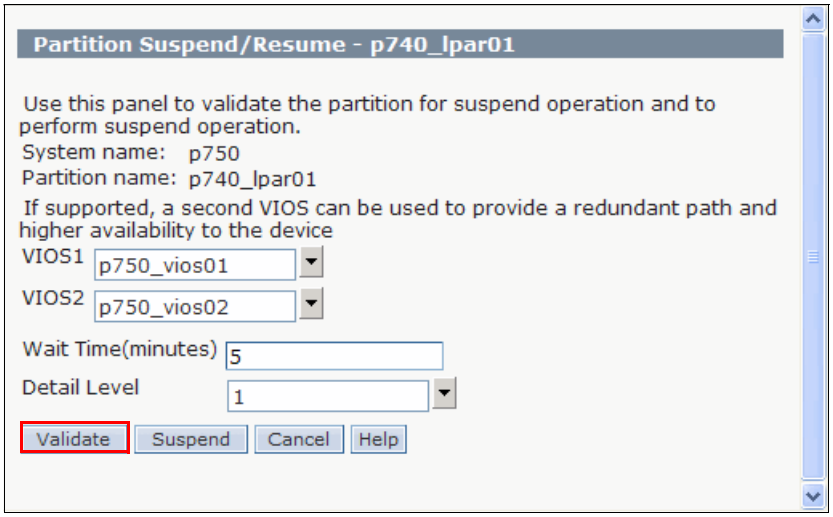


Figure 13-25 Validate Suspend and Resume operation

If the partition is ready for the operation, you get a successful validation result like that shown in Figure 13-26.

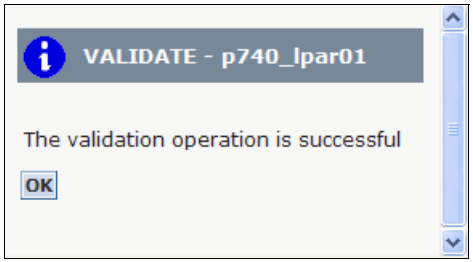


Figure 13-26 Successful validation result

13.2.2 Monitoring Suspend and Resume operations on IBM i

On an IBM i partition, the following messages are logged in the history log QHST for Suspend and Resume operations as shown in Figure 13-27.

```
                        Additional Message Information

Message ID . . . . . :   CPI09A5       Severity . . . . . :   00
Message type . . . . . :   Information
Date sent  . . . . . :   12/08/12      Time sent  . . . . . :   13:31:20

Message . . . . :   Partition suspend request in progress.
Cause . . . . . :   A request was made to suspend the partition.


                        Additional Message Information

Message ID . . . . . :   CPI09A9       Severity . . . . . :   00
Message type . . . . . :   Information
Date sent  . . . . . :   12/08/12      Time sent  . . . . . :   13:50:03

Message . . . . :   Partition resumed from hibernation.
Cause . . . . . :   The partition has resumed from hibernation.
```

Figure 13-27 IBM i history log messages for suspend and resume

14

Live Partition Mobility

PowerVM Live Partition Mobility allows for the movement of an active (running), suspended, or inactive (not activated) partition from one system to another with no application downtime.

PowerVM Live Partition Mobility requires systems with POWER6 or newer processors running PowerVM Enterprise Edition. It is supported for partitions that run AIX 5.3 TL7 or later, RedHat Enterprise Linux version 5 Update 1 or later, SuSE Linux Enterprise Server 10 Service Pack 4 or later, and IBM i 7.1 TR4 or later. LPM for IBM i is supported on POWER7 processors only.

For more information about the architecture, requirements and setup for Live Partition Mobility see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

This section describes how to manage and monitor a Live Partition Mobility environment.

This chapter includes the following sections:

- Managing Live Partition Mobility
- Monitoring Live Partition Mobility

14.1 Managing Live Partition Mobility

This part describes how to manage Live Partition Mobility, and includes the following sections:

- Migrating a logical partition
- HMC commands for Live Partition Mobility
- Making applications migration aware
- Making AIX applications migration aware using scripts
- Making AIX kernel extension migration aware
- Migration recovery
- Monitoring Live Partition Mobility

14.1.1 Migrating a logical partition

This section shows a simple configuration for Live Partition Mobility (LPM) using an Hardware Management Console (HMC) and virtual SCSI disks. A single Virtual I/O Server partition is configured on the source and destination systems. The configuration in Figure 14-1 on page 543 shows performing a Live Partition Mobility migration for a partition to be migrated, called the *mobile partition*, to another Power Systems server.

When attempting to do Live Partition Mobility (LPM) in your environment, use the Live Partition Mobility Preparation Checklist. The assumption for using this checklist is that you previously set up your LPM environment based on the Live Partition Mobility Setup Checklist or the *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940, an IBM® Redbooks® publication.

The checklist can be found at:

http://www.redbooks.ibm.com/abstracts/tips1185.html?Open

Note: Although this migration example is shown for a mobile partition using VSCSI with a single Virtual I/O Server, the migration process works similarly for a setup that uses NPIV with virtual Fibre Channel and a dual Virtual I/O Server configuration.

For detailed planning and setup information about Live Partition Mobility, see *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

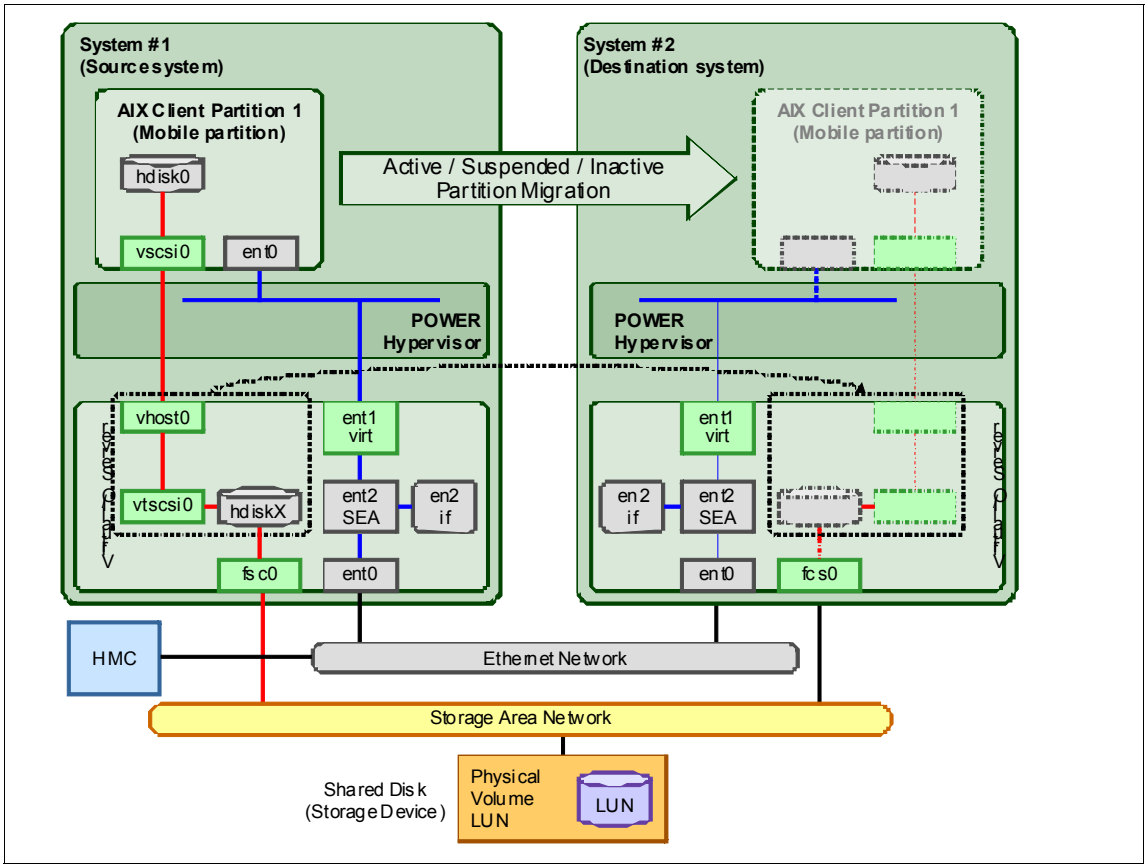


Figure 14-1 Basic Live Partition Mobility configuration

The example used in this chapter involves migrating a logical partition that is called a *mobile partition* from the source to the destination system by using the following main steps:

1. Performing validation for partition mobility
2. Migrating the mobile partition

Performing validation for partition mobility

Before you run a migration, complete the validation steps. These steps are optional, but can help to eliminate errors. You can perform the validation steps by

using the HMC GUI or CLI. This section covers the GUI steps. For information about the CLI, see “The migrlpar command” on page 568.

1. In the navigation pane, expand **Systems Management** → **Servers**, and select the source system.
2. In the contents pane (the upper right of the HMC Workplace), select the partition that you will migrate to the destination system.
3. Click the **View popup menu** icon and select **Operations** → **Mobility** → **Validate**, as shown in Figure 14-2, to start the validation process.

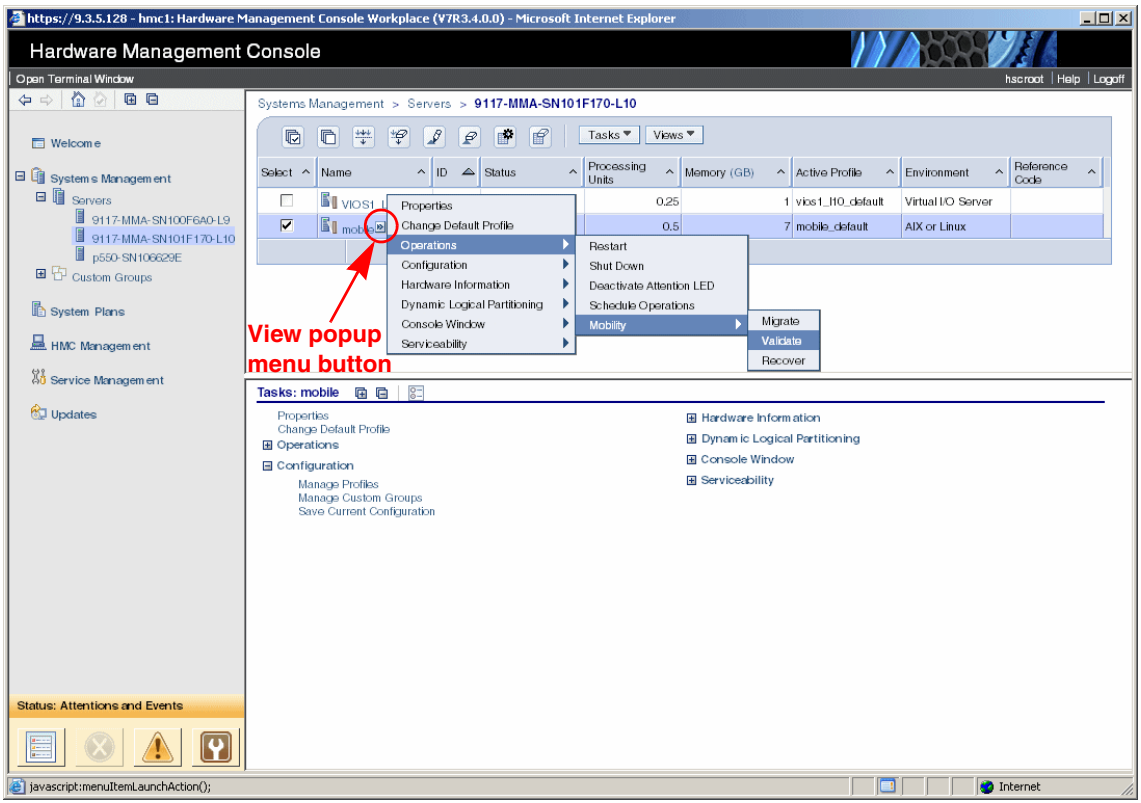


Figure 14-2 Partition mobility validate menu on the HMC

4. Select the destination system, specify **Destination profile name** and **Wait time**, and then click **Validate** (Figure 14-3).

> **Note:** Figure 14-3 shows the option of entering a remote HMC's information. This step applies only to a migration between systems that are managed by different HMCs. The example shows migration of a partition between systems that are managed by a single HMC.

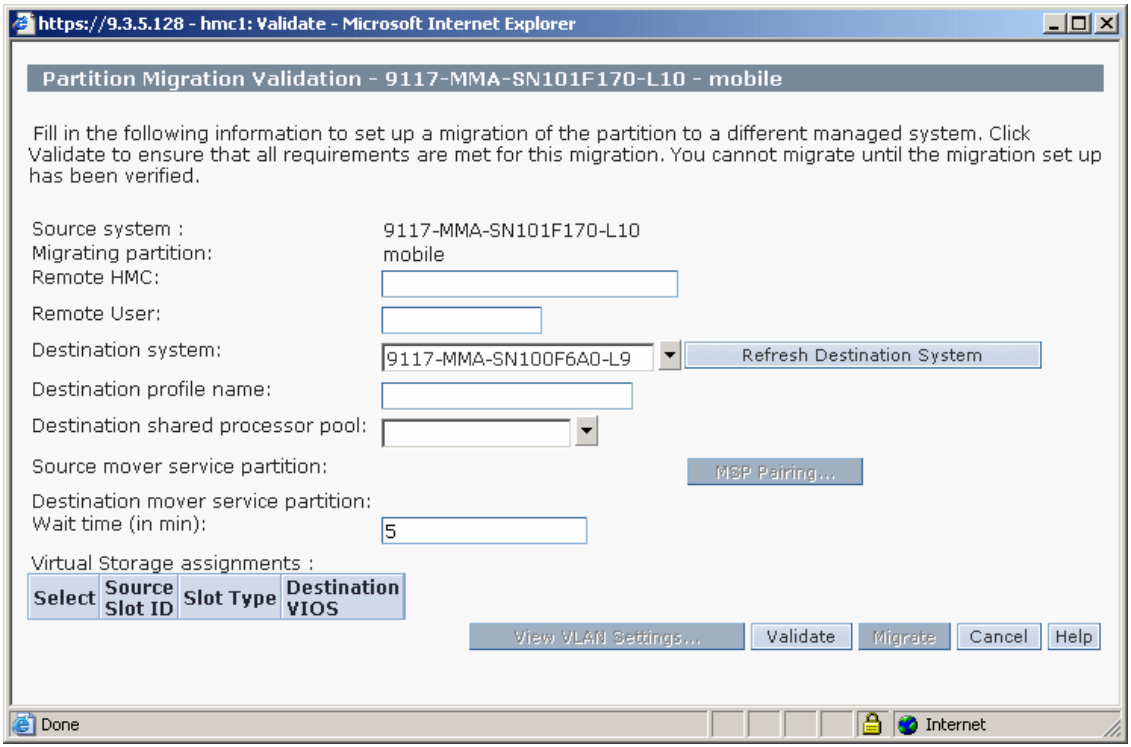


Figure 14-3 Selecting the Remote HMC and Destination System

When the mobile partition is in the **Not Activated** state, the **Destination** and **Source mover service partition** and **Wait time** fields are not shown. These are not required for inactive partition migration where no partition memory data must be moved.

5. Check for errors or warnings in the Partition Validation Errors/Warnings window, and eliminate any errors.

   If any errors occur, check the messages in the window and the prerequisites for the migration. You cannot complete the migration steps with any errors.

For example, if you are proceeding with the validation steps on the mobile partition with physical adapters in the **Running** state (active migration), you get the error shown in Figure 14-4.

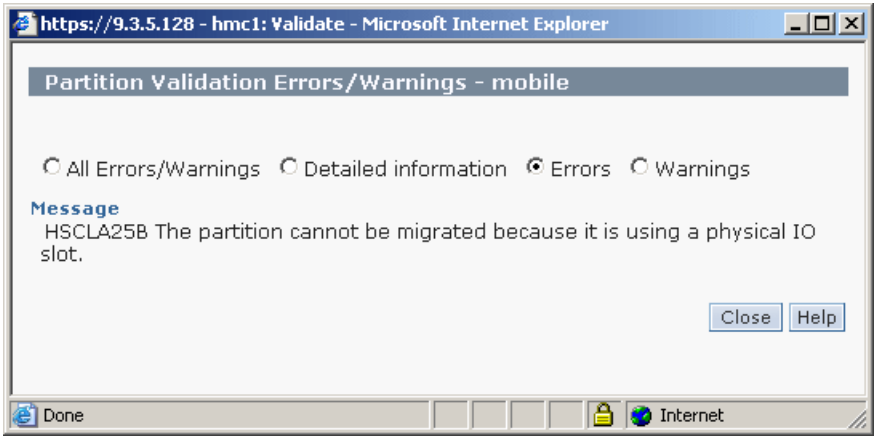


Figure 14-4 Partition Validation Errors

If the mobile partition is in the **Not Activated** state, a warning message is reported as shown in Figure 14-5.

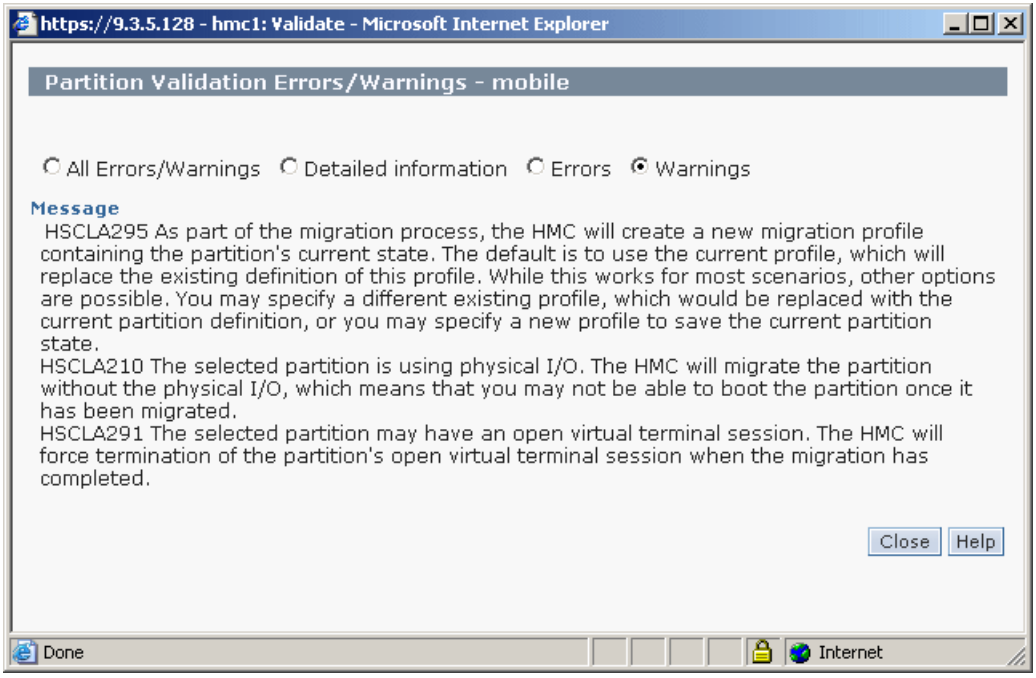


Figure 14-5 Partition Validation Warnings

6. After you close the Partition Validation Errors/Warnings window, the Partition Migration Validation window, as shown in Figure 14-6, opens again.

   If you had no errors in the previous step, you can now run the migration by clicking **Migrate**.

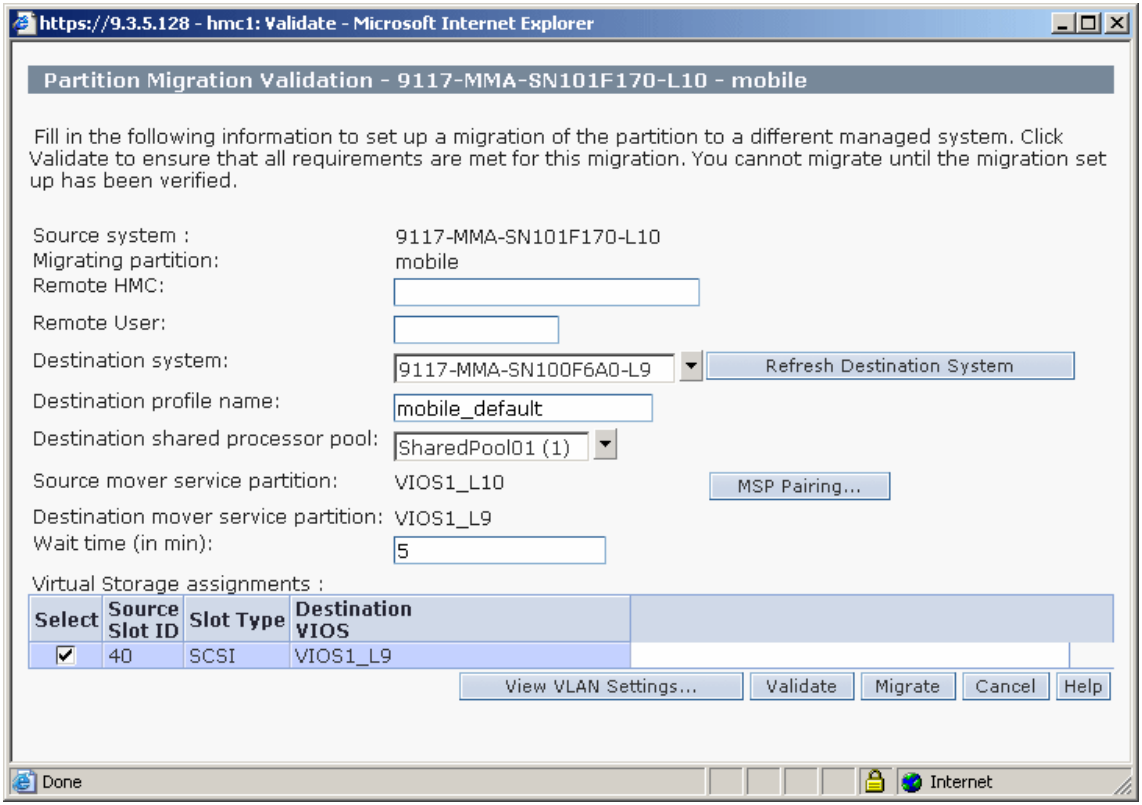


Figure 14-6 HMC Partition Migration Validation window after validation

Migrating a mobile partition

After you complete the validation steps, migrate the mobile partition from the source to the destination system. You can complete the migration steps by using the HMC GUI or CLI. For more information about the CLI, see “The migrlpar command” on page 568.

This scenario involves migrating a partition named *mobile* from the source system (9117-MMA-SN101F170-L10) to the destination system (9117-MMA-SN10F6A0-L9). To migrate a mobile partition, complete these steps:

1. In the navigation pane, expand **Systems Management** → **Servers**, and select the source system.

   You can now see that the mobile partition is on the source system, as shown in Figure 14-7.

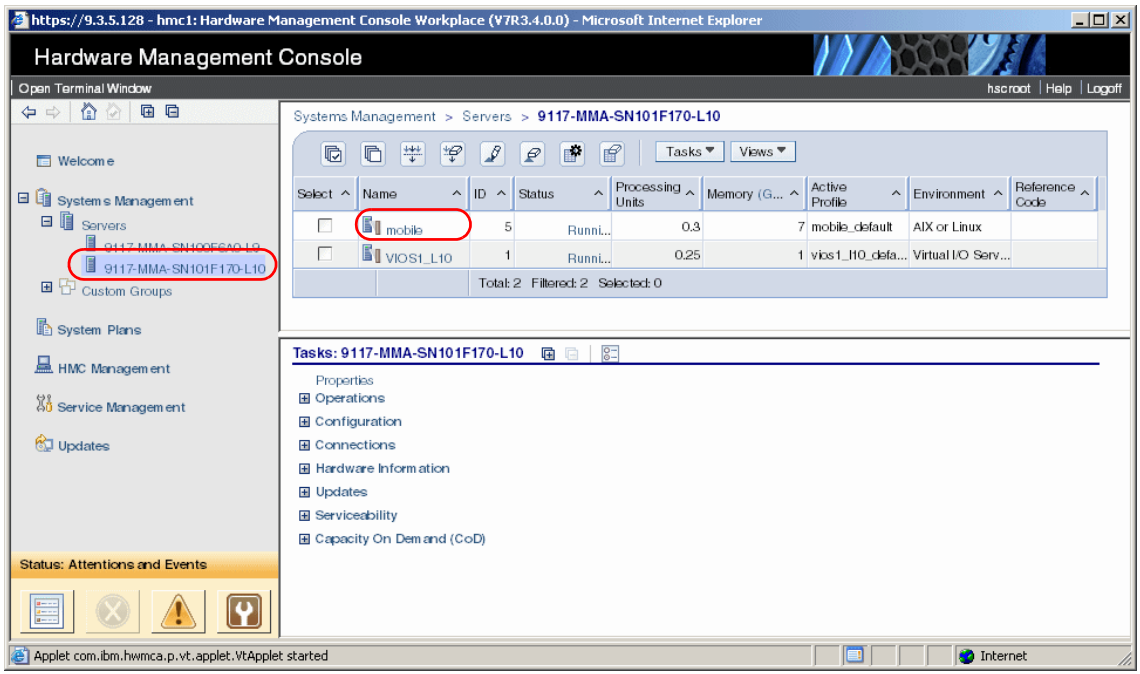


Figure 14-7 System environment before migration

2. In the contents pane, select the partition to migrate to the destination system, that is, the *mobile* partition.

3. Click **View popup menu** and select **Operations** → **Mobility** → **Migrate**, as shown in Figure 14-8, to start the Partition Migration wizard.

> **Note:** Alternatively, the `Validate` operation can be chosen to manually perform a validation first. After this operation completes successfully, you can start the migration from the same Partition Migration Validation window as shown at the end of this section.

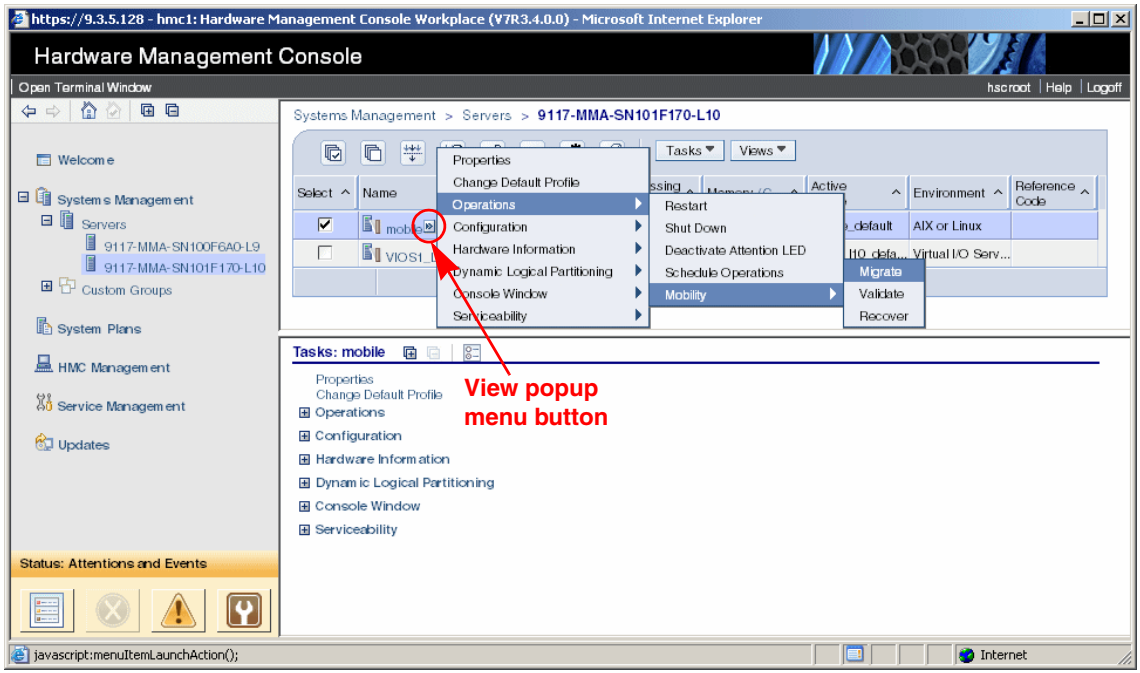


Figure 14-8 Partition mobility migrate menu on the HMC

4. Check the Migration Information of the mobile partition in the Partition Migration wizard.

   If the mobile partition is powered off, **Migration type** is **Inactive**. If the partition is suspended, **Migration type** is **Suspended**. If the partition is in the running state, it is **Active**, as shown in Figure 14-9 (new image 3/4/14).

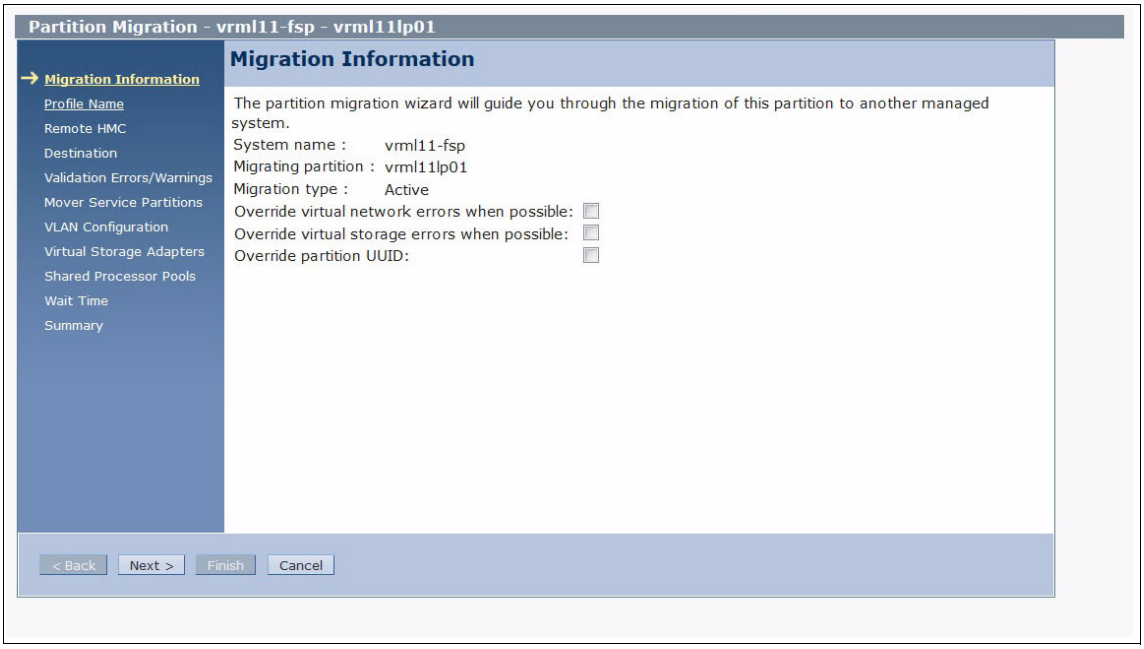


Figure 14-9 Partition migration information

5. You can specify the New destination profile name in the Profile Name window, as shown in Figure 14-10.

   If you leave the name blank or do not specify a unique profile name, the profile on the destination system is overwritten.

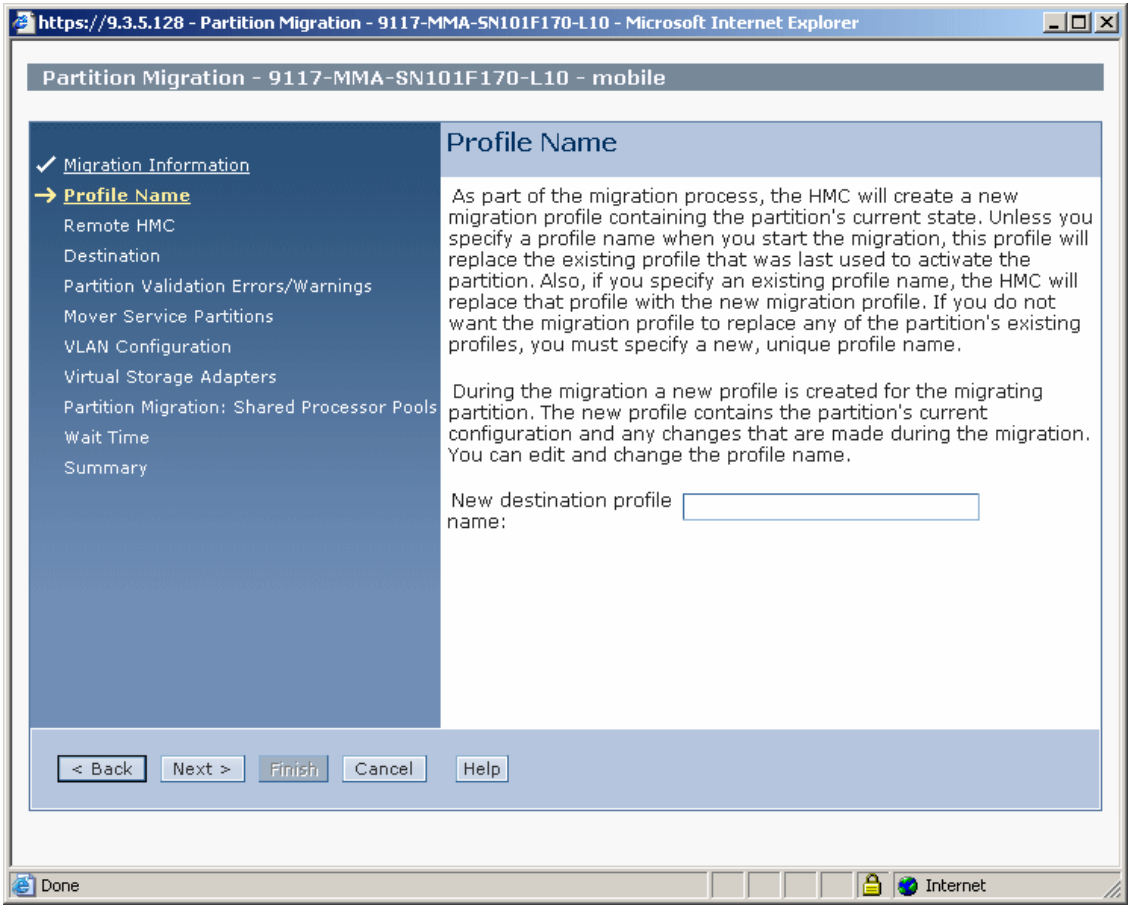


Figure 14-10 Specifying the profile name on the destination system

6. Optionally, for a remote migration to a destination server managed by a different HMC, enter the Remote HMC network address and Remote User. The example uses a single HMC, and therefore do not need Remote Migration as shown in Figure 14-11. Click **Next**.

> **Note:** Partition migration to a remote HMC is a function that became available starting in HMC Version 7 Release 3.4. It requires prior set up of SSH authentication between the local and remote HMC as described in the *Live Partition Mobility setup* section of *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

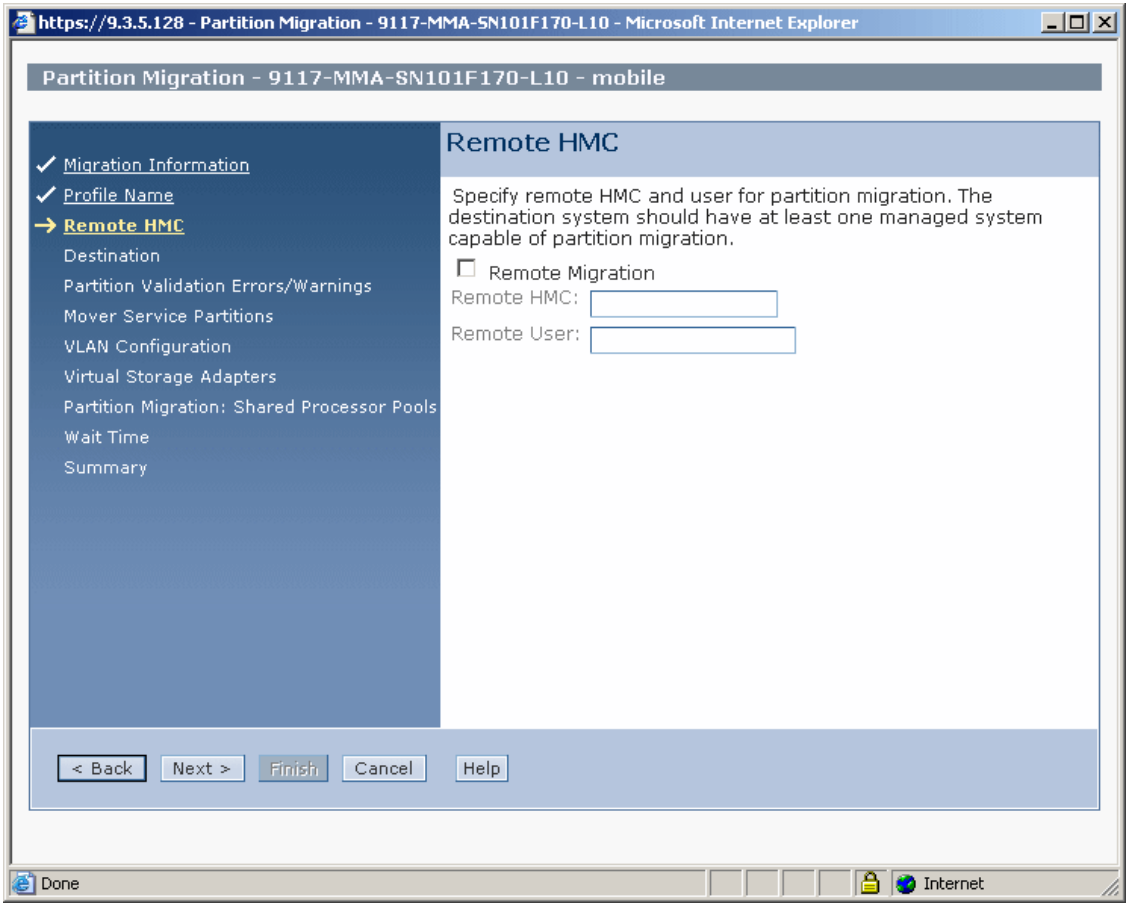


Figure 14-11 Optionally specifying the Remote HMC of the destination system

7. Select the destination system and click **Next** as shown in Figure 14-12.

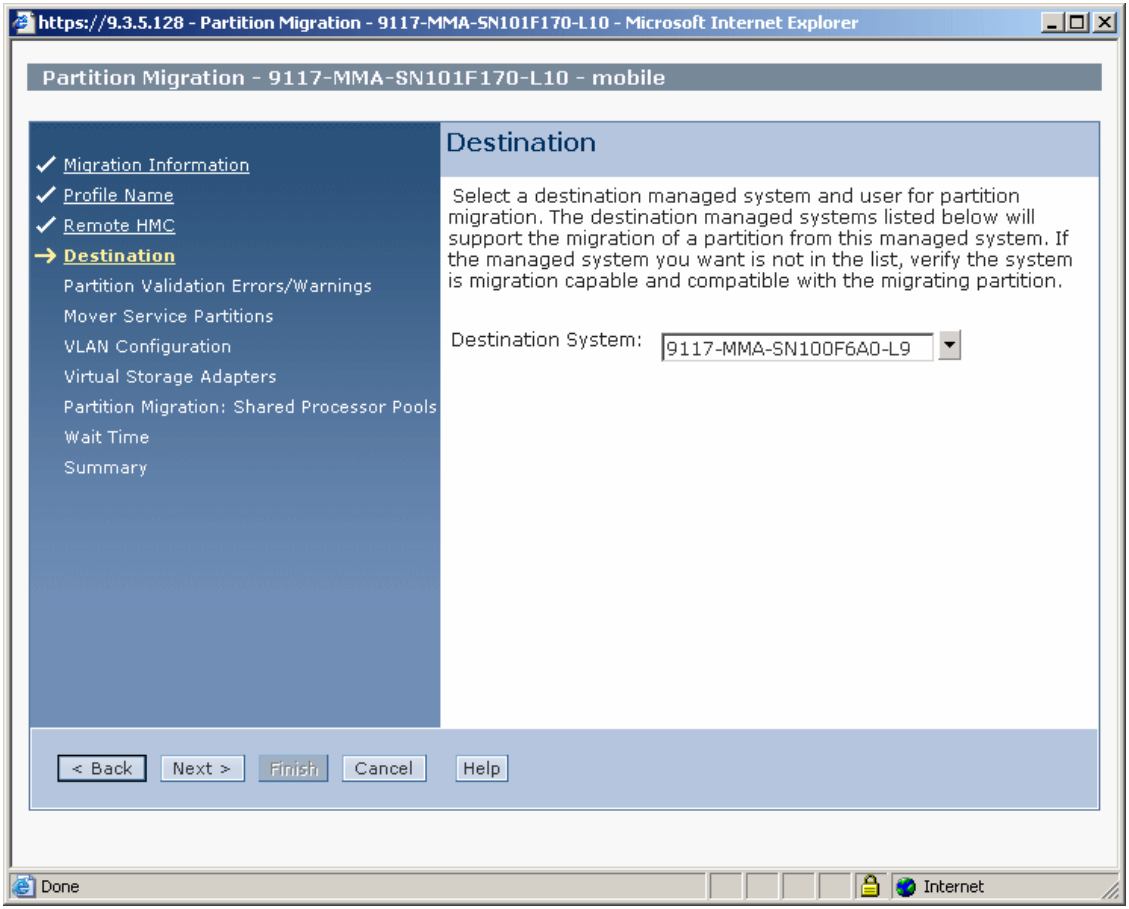


Figure 14-12 Selecting the destination system

The HMC then validates the partition migration environment.

8. Check errors or warnings in the Partition Validation Errors/Warnings panel shown in Figure 14-13 (updated 3/4/14) and eliminate any errors. If errors exist, you cannot proceed to the next step. If only warnings exist, you can proceed to the next step.

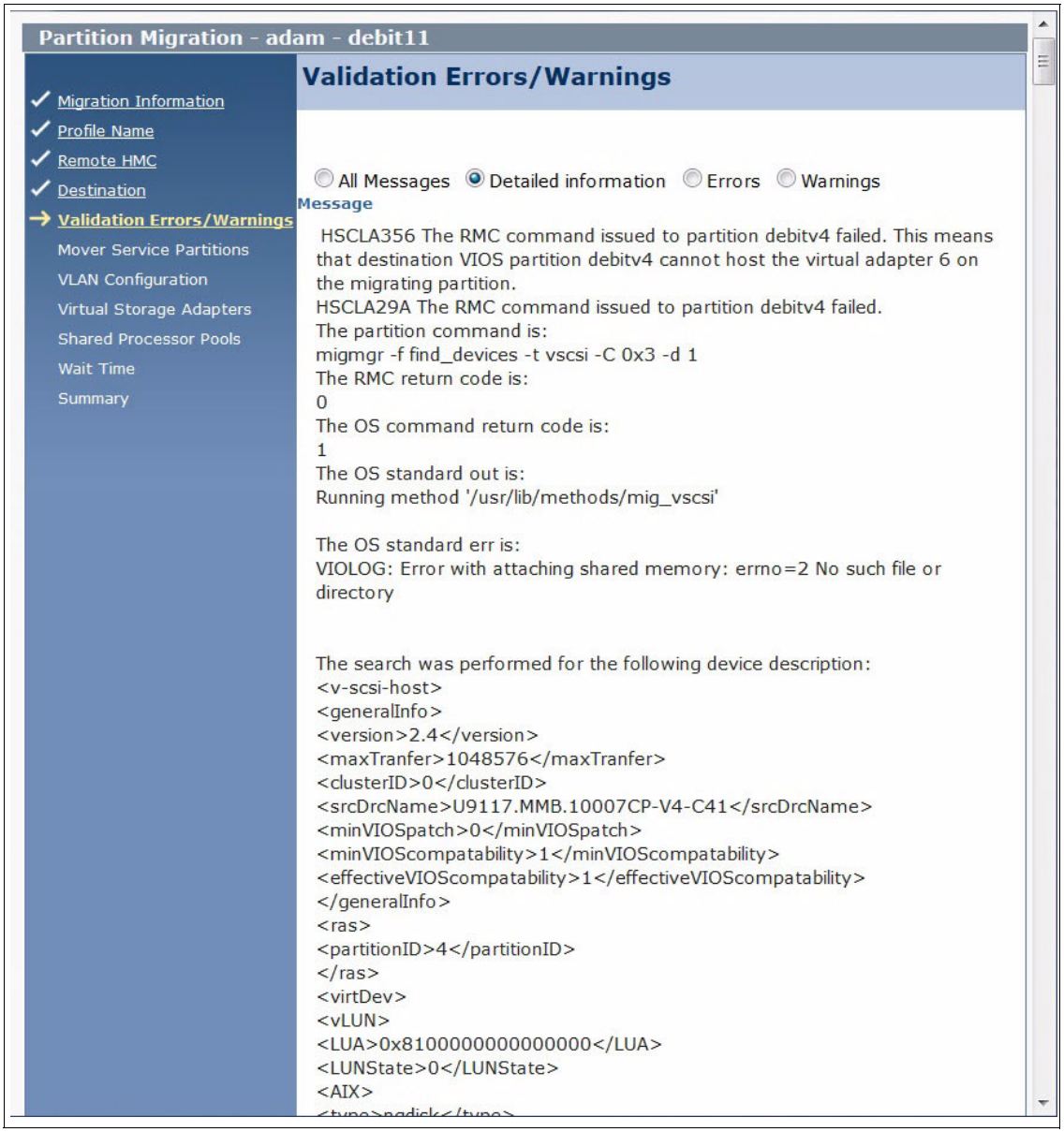


Figure 14-13 Sample of Partition Validation Errors/Warnings

The most common missing requirements can be treated as shown in Table 14-1.

Table 14-1 Missing requirements for PowerVM Live Partition Mobility

Partition validation error message reported	Correction
HSCLA2B7 The management console was unable to find a valid mover service partition (MSP) on managed system...	► Check that there is a Virtual I/O Server on the source and the destination system that has “Mover service partition” selected in its general properties. ► Check that both Virtual I/O Servers can communicate with each other through the network.
HSCLA27C The operation to get physical device location for adapter... has failed...	► The partition might have access to the CD/DVD drive through a virtual device. Temporally remove the mapping on the virtual CD/DVD drive on the Virtual I/O Server by using the **rmdev** command.

The migrating partition's virtual SCSI adapter *xx* cannot be hosted by the existing Virtual I/O Server (VIOS) partitions on the destination managed system.	Check the Detailed information tab:
	► If a message mentions “Missing Begin Tag reserve policy mismatched”, check that the moving storage disks reserve policy is set to `no_reserve` on the source and destination disks on the Virtual I/O Servers. You can use a command similar to:
	`lsdev -dev hdiskXXX -attr`
	You can fix it with the command:
	`chdev -dev hdisk`*`xxx`*` -attr reserve_policy=no_reserve`
	► If not, check in your SAN zoning that the destination Virtual I/O Server can access the same LUNs as the source. For more information, see 9.4.1, “Virtual I/O Server storage monitoring” on page 269.
	► Another possibility is the max_transfer sizes of the disks are different on each Virtual I/O Server. So on each Virtual I/O Server run a command similar to: lsdev -dev hdiskXXX -attr to see the current values. If the max_transfer size is different on each Virtual I/O Server then change the lower of the values to match that of the higher value. To change the value you can use a command similar to:
	`chdev -dev hdiskXXX -attr max_transfer=XXXX`

HSCLAA23 The partition `IBM i` cannot be migrated because IBM i restricted I/O mode is not enabled for the partition.	► Enable the partition properties setting “Restricted IO Partition.” This setting can be changed only when the IBM i partition is not activated.

9. If you are running an *inactive migration*, skip this step and go to step 10 on page 557.

   If you are running an *active or suspended migration*, select the source and the destination mover service partitions to be used for the migration as shown in Figure 14-14 (image updated 3/4/14).

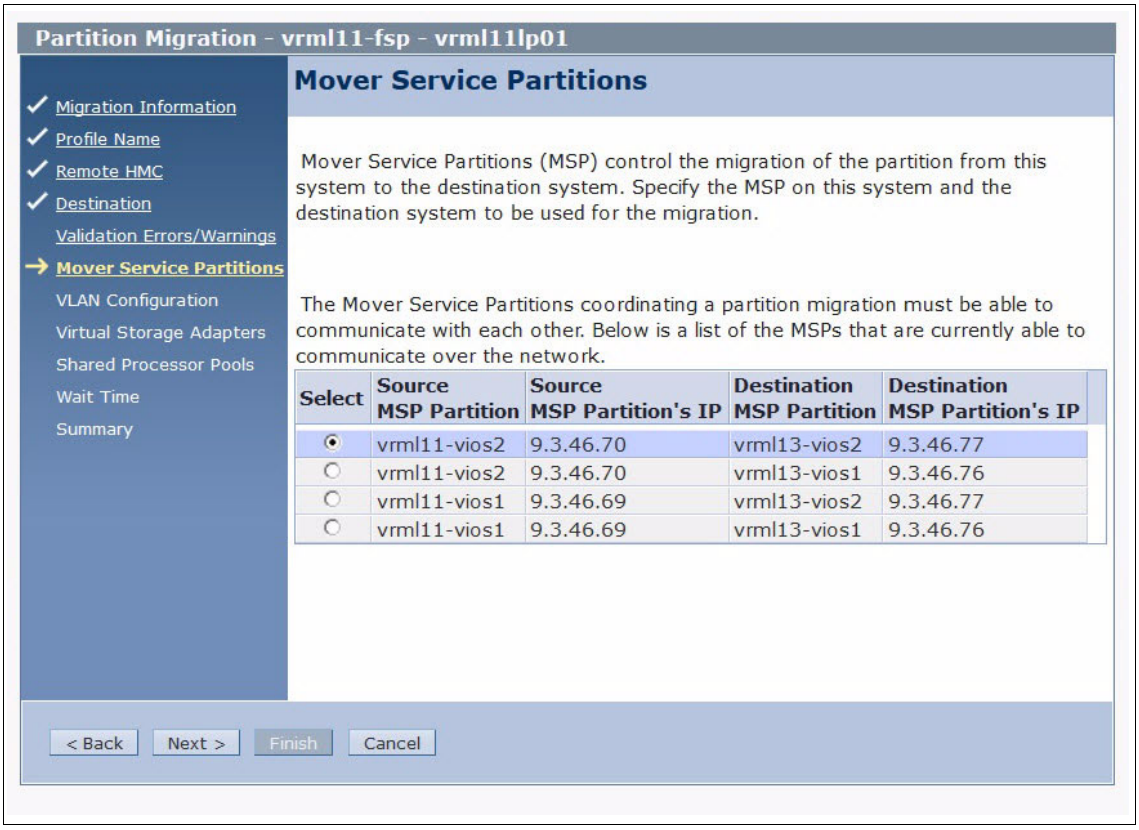


Figure 14-14 Selecting mover service partitions

In this basic scenario, one Virtual I/O Server partition is configured on the destination system, so the wizard window shows only one mover service partition candidate. If you have more than one Virtual I/O Server partition on the source or on the destination system, you can select which mover server partitions to use. If your MSP's are configured with multiple IP addresses you'll want to make sure to pick the MSP pairing that is appropriate. This is especially important if one network is faster and/or has less traffic on it than the other.

10. Select the VLAN configuration as shown in Figure 14-15.

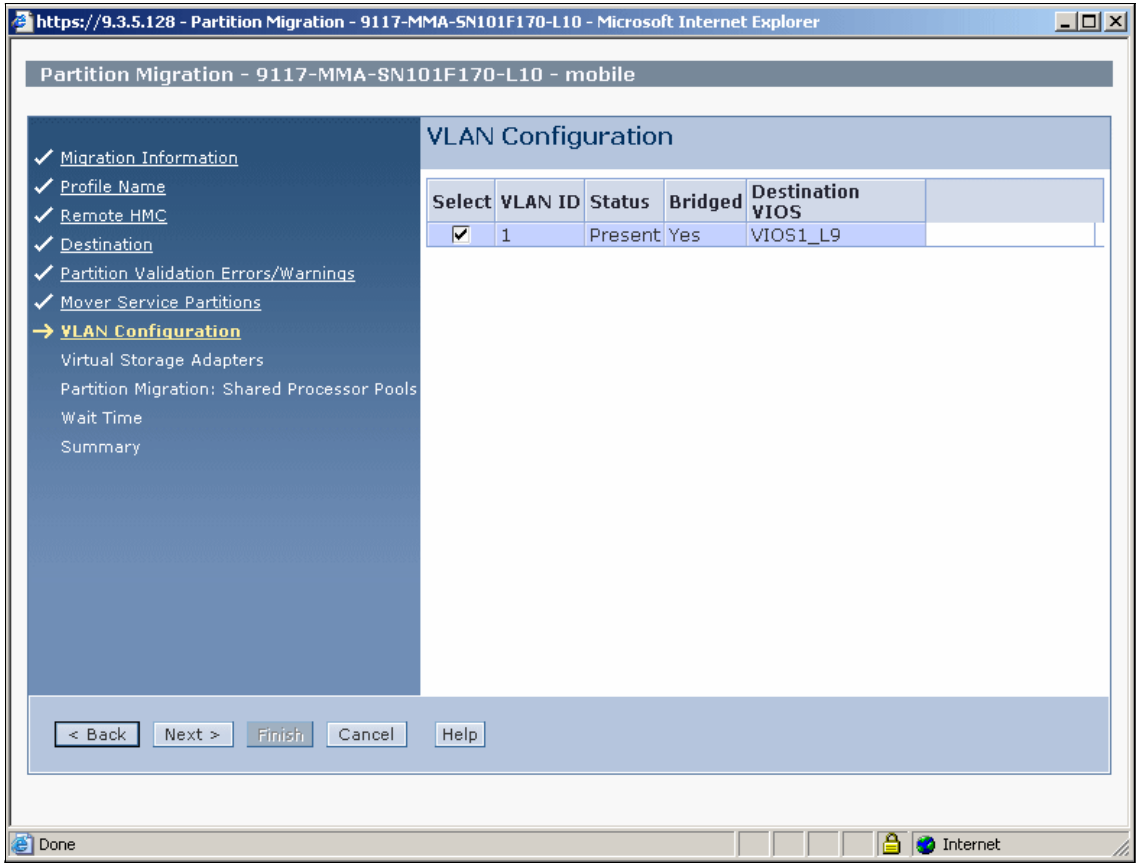


Figure 14-15 Selecting the VLAN configuration

11. Select the virtual storage adapter assignment as shown in Figure 14-16.

    In this case, one Virtual I/O Server partition is configured on each system, so this wizard window shows one candidate only. If you have more than one Virtual I/O Server partition on the destination system, you can choose which Virtual I/O Server to use as the destination.

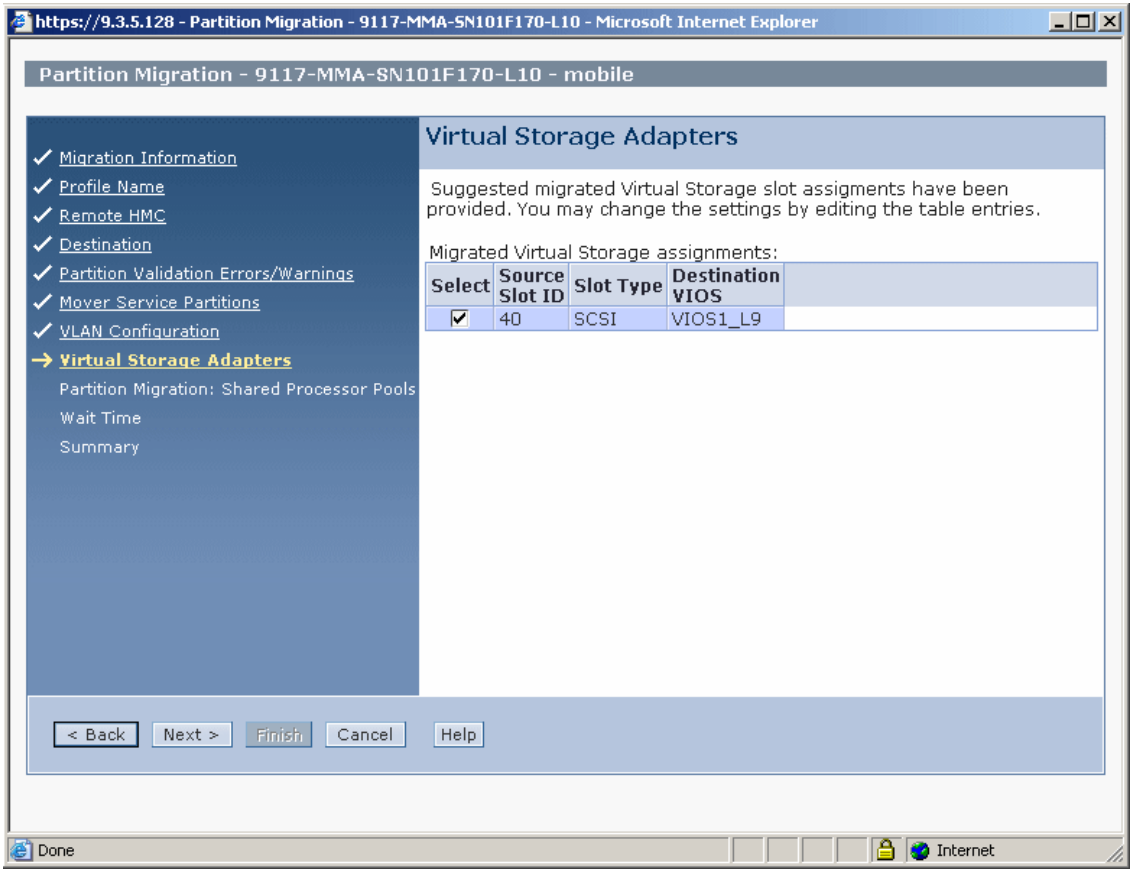


Figure 14-16 Selecting the virtual SCSI adapter

12. Select the destination shared processor pool from the list of shared processor pools that support the source partition's shared processor pool configuration as shown in Figure 14-17.

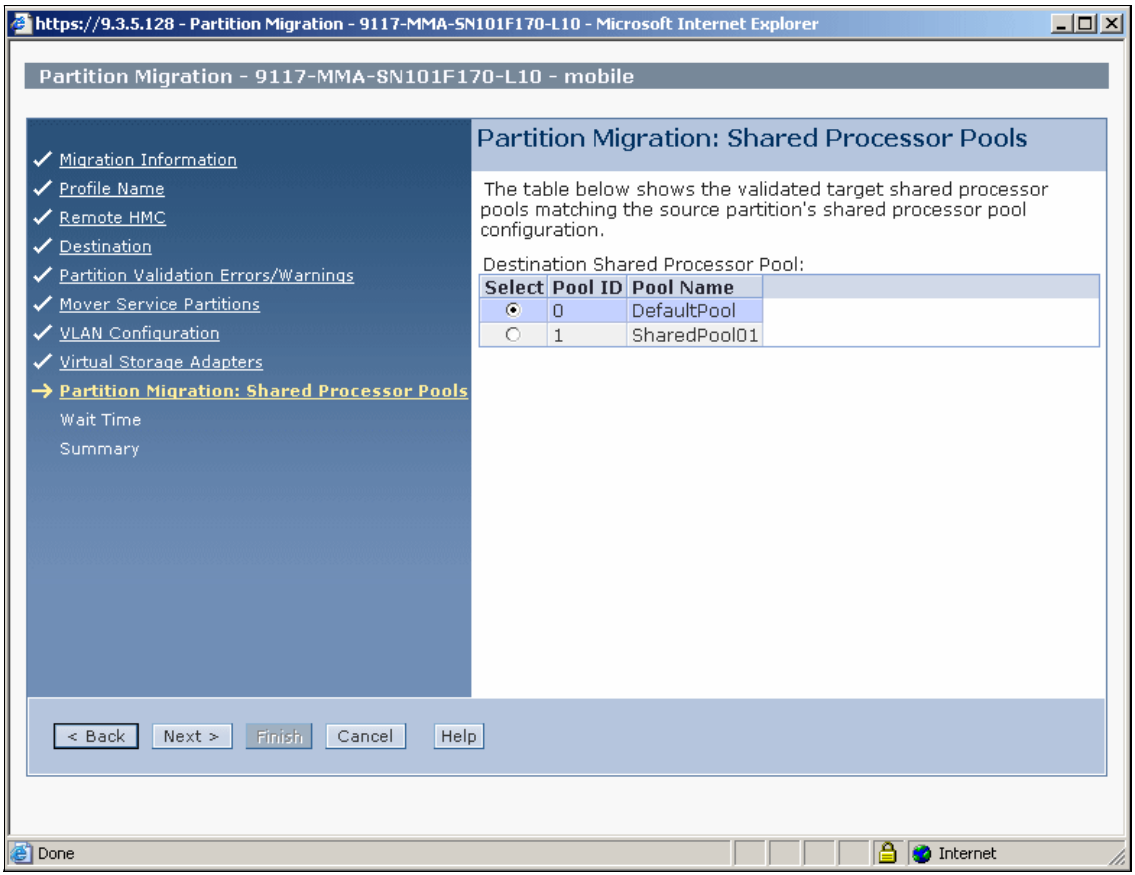


Figure 14-17 Specifying the shared processor pool

Note: If the mobile partition is using dedicated processors or there is only one shared processor pool on the destination system, this option might not be displayed.

13. Specify the wait time in minutes as shown in Figure 14-18.

    The wait time value is passed to the commands that are started on the HMC and that run migration-related operations on the relevant partitions using the Remote Monitoring and Control (RMC).

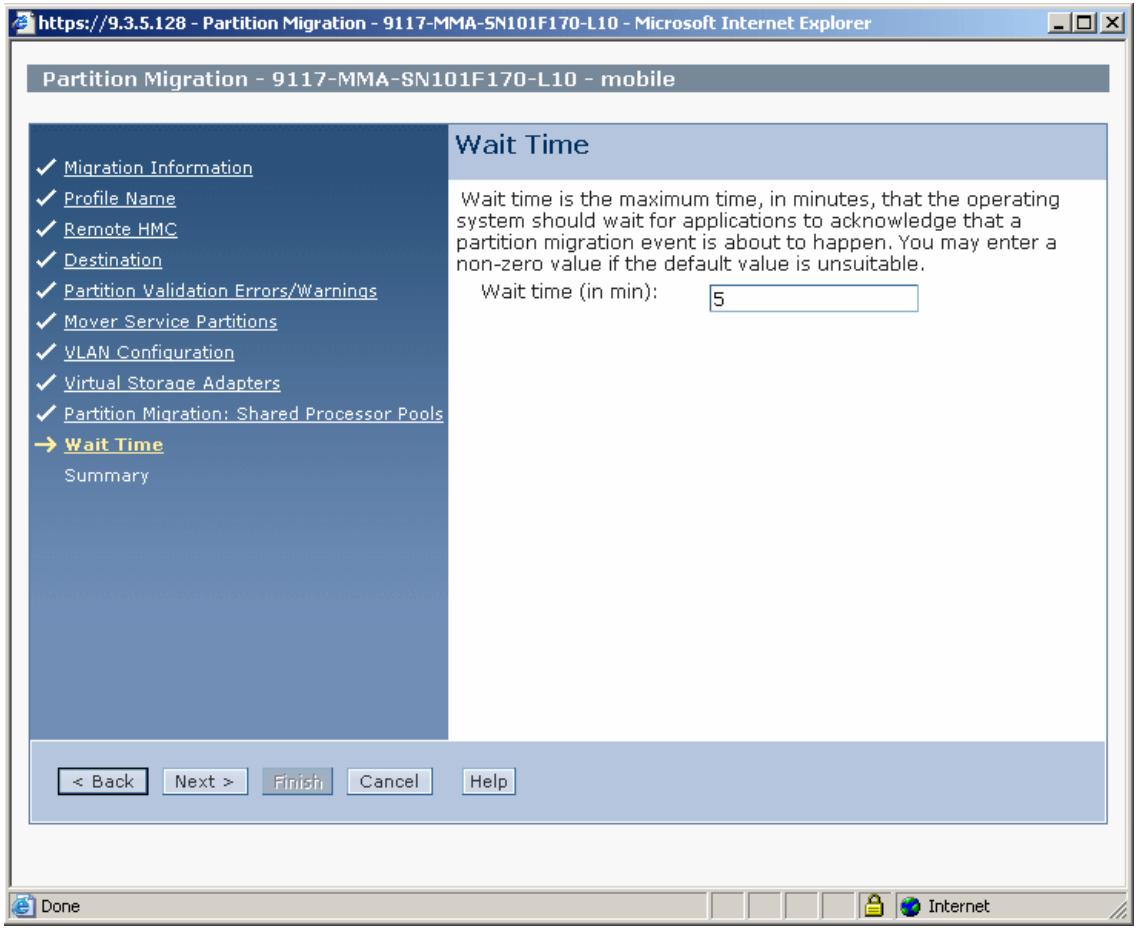


Figure 14-18 Specifying wait time

14. Check the settings that you have specified for this migration on the Summary window, then click **Finish** to begin the migration as shown in Figure 14-19.

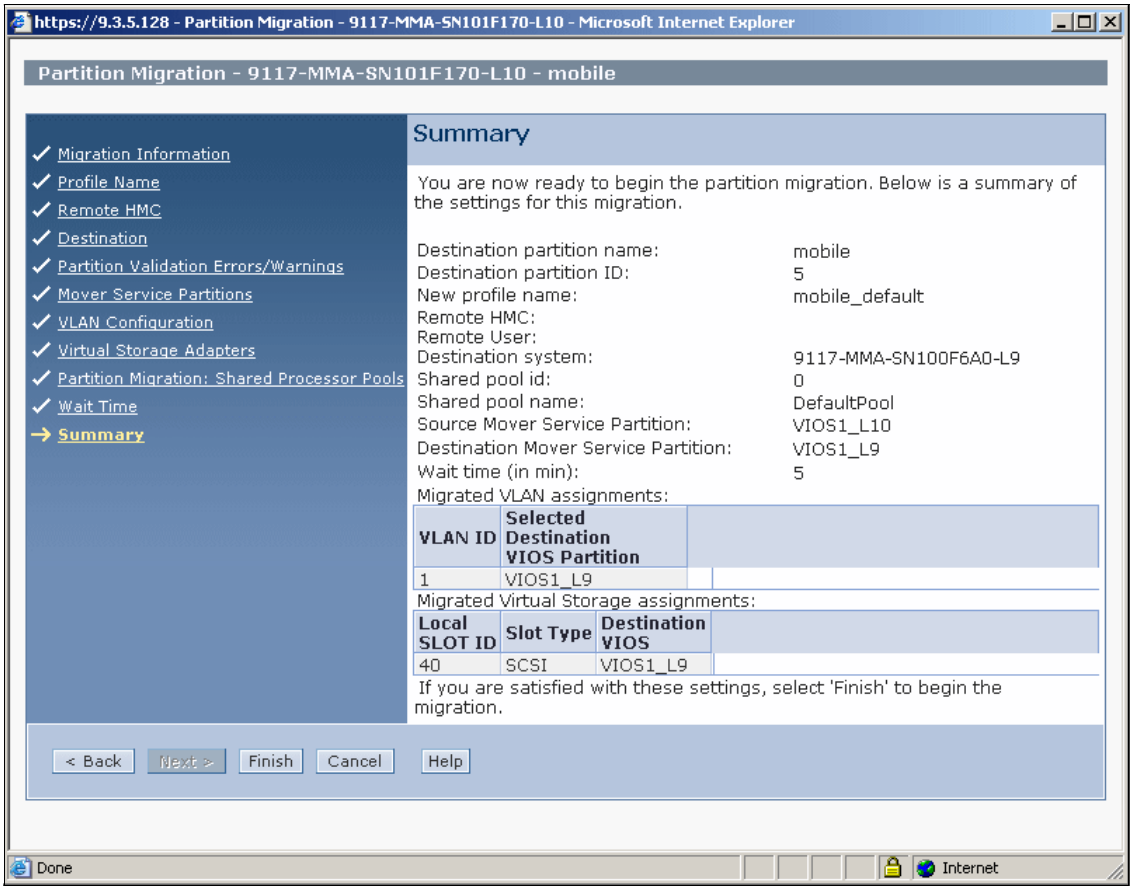


Figure 14-19 Partition Migration Summary window

15. The Migration status and Progress are shown in the Partition Migration Status window, as shown in Figure 14-20.

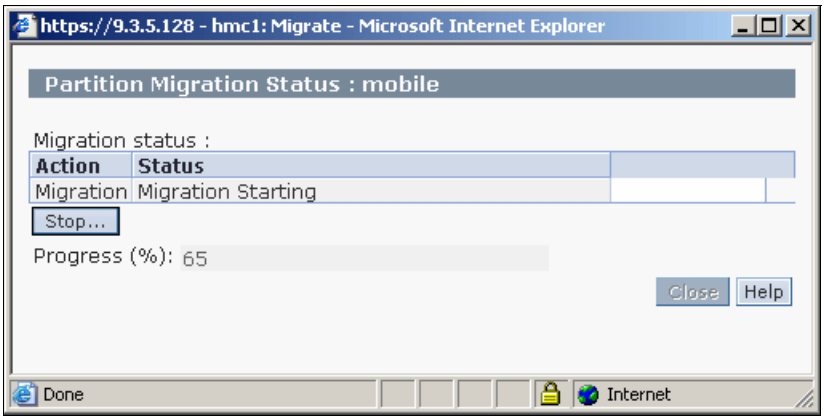


Figure 14-20 Partition Migration Status window

16. When the Partition Migration Status window indicates that the migration is 100% complete, verify that the mobile partition is in the Running state on the destination system. The mobile partition is on the destination system, as shown in Figure 14-21.

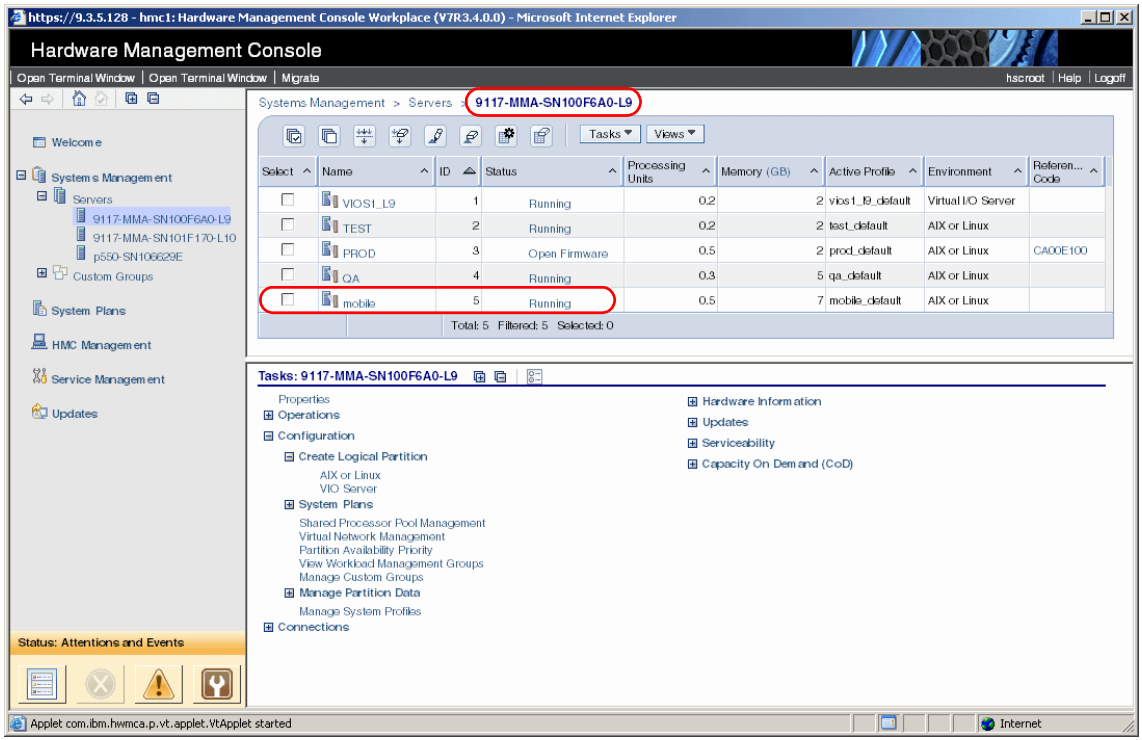


Figure 14-21 Migrated partition

Migration by using the Partition Migration Validation window

Rather than using the *Partition Migration* wizard, the more direct way of performing a partition migration using the HMC GUI is to use the Partition Migration Validation window.

This window can be started by choosing the mobile partition to be migrated and selecting **Operations** → **Mobility** → **Validate** as shown in Figure 14-22.

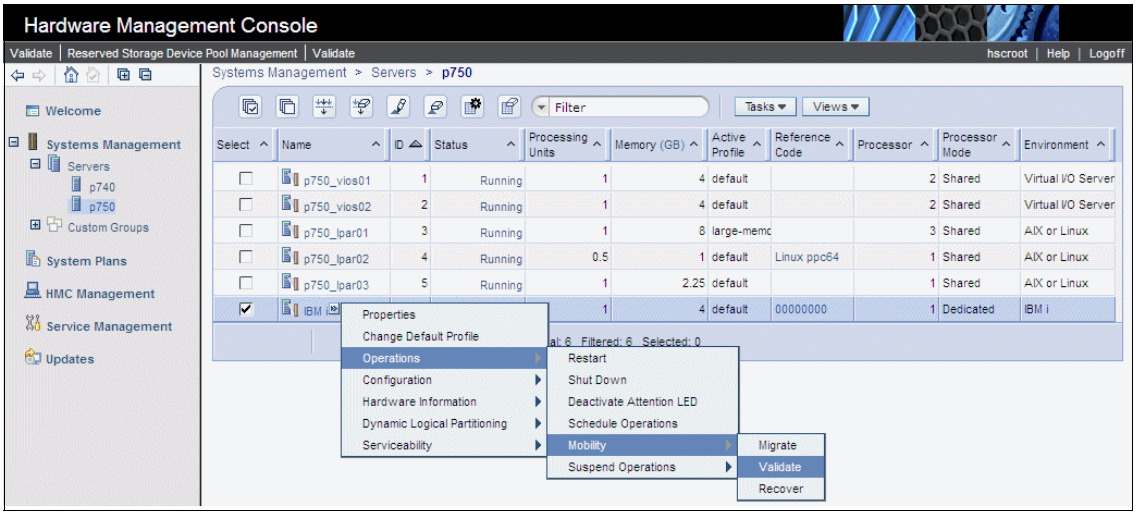


Figure 14-22 HMC partition mobility validate operation

The Partition Migration Validation window is displayed as shown in Figure 14-23 (updated 3/4/14).

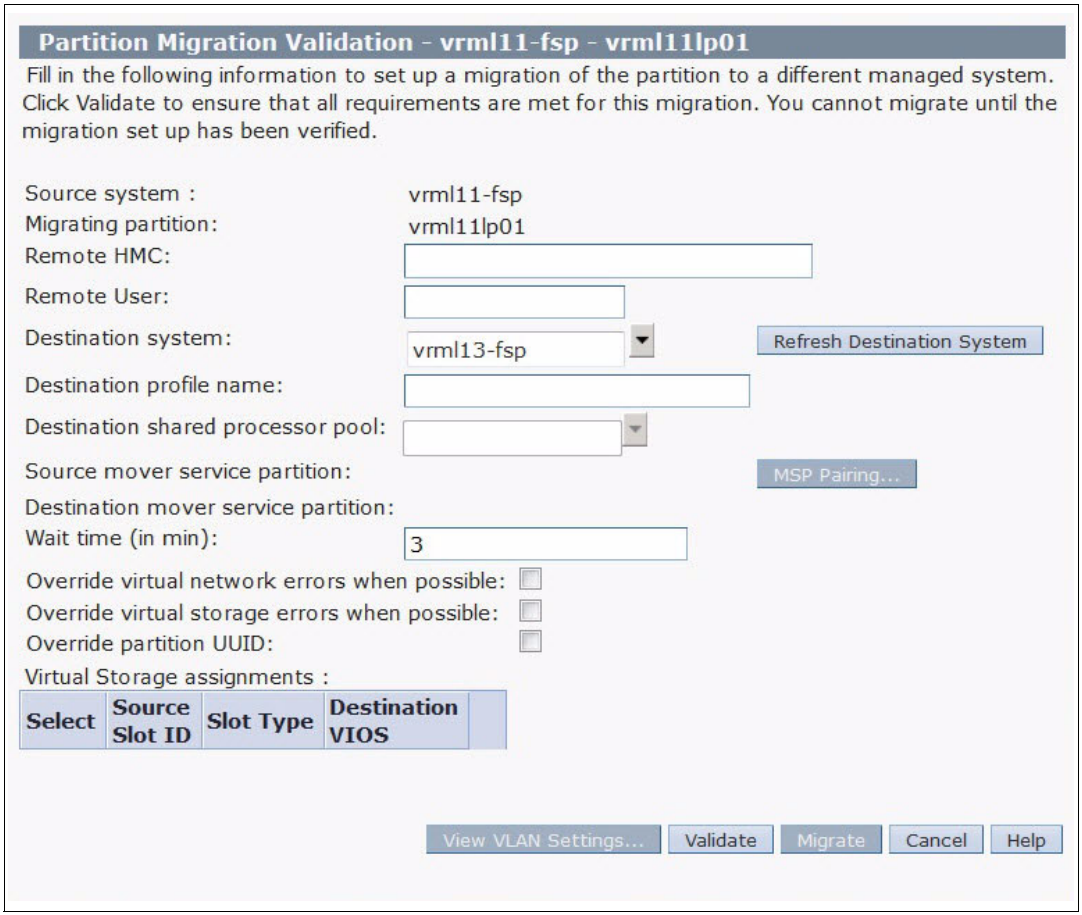


Figure 14-23 HMC partition migration validation

After you select the settings you want for the migration and click **Validate**, the **Migrate** button becomes selectable after a successful validation. Click this button to actually start the migration as shown in Figure 14-24.

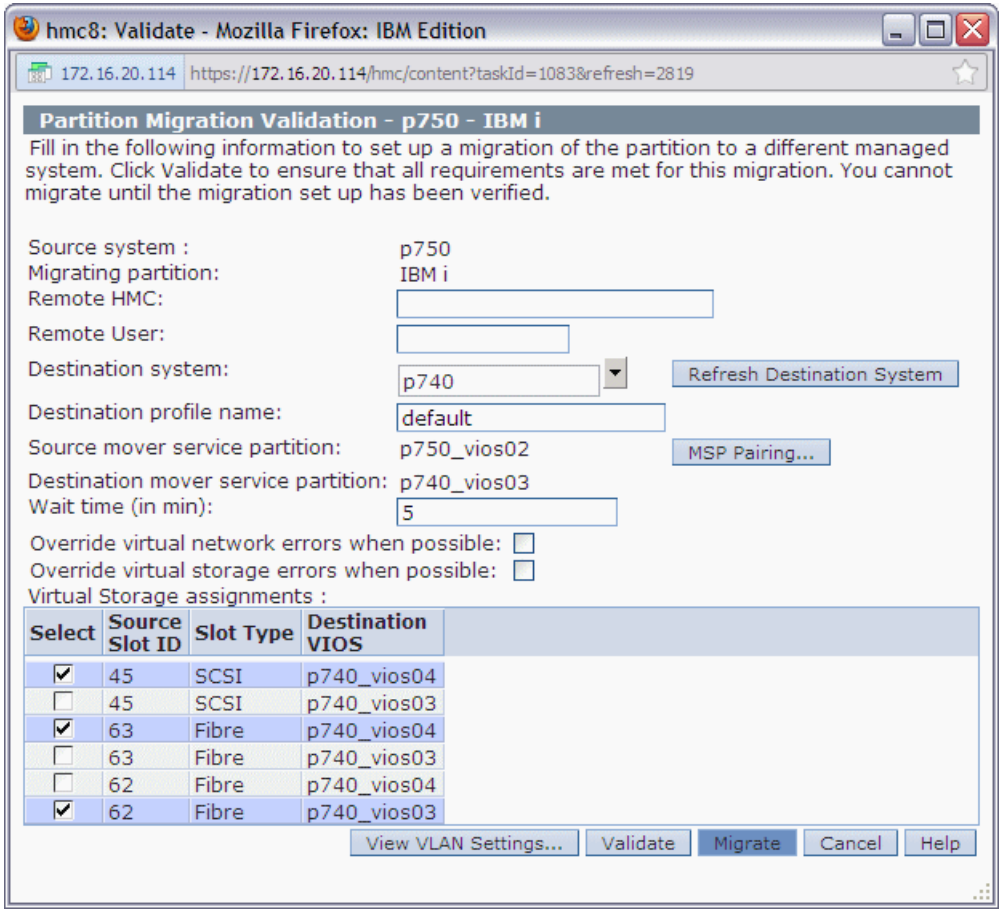


Figure 14-24 HMC partition migration validation migrate operation

The status of a successful migration for an IBM i partition is shown in Figure 14-25.

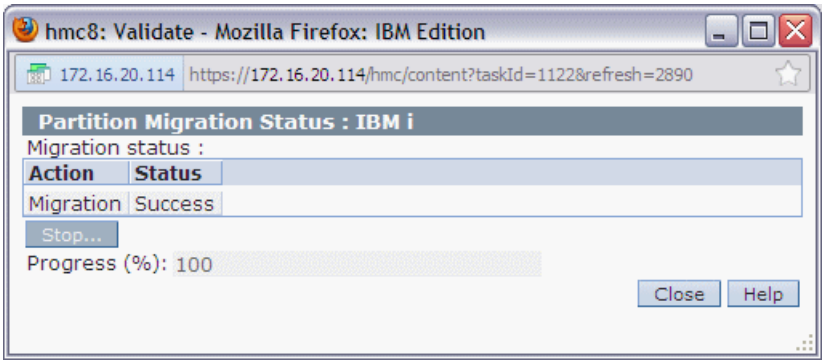


Figure 14-25 HMC message for successful partition migration

If you keep a record of the virtual I/O configuration of the partitions, check and record the migrating partition's configuration in the destination system. Although the migrating partition retains the same virtual client adapter slot numbers as on the source system, the server virtual adapter slot numbers can be different.

Note: When you migrating a partition using NPIV with virtual Fibre Channel adapters by using the HMC GUI, the mappings between the virtual FC server adapters and the physical FC adapters are done automatically on the destination system.

If you want to specify which physical FC port is used on the destination system, use the HMC command **`migrlpar`** with the **`-f`** option to specify a CSV file with attribute settings. Or you can change the mappings after the migration with the **`vfcmap`** command on the Virtual I/O Servers.

14.1.2 HMC commands for Live Partition Mobility

The HMC provides a command-line interface (CLI) and an easy-to-use GUI for Live Partition Mobility. The CLI allows you to script frequently performed operations to implement automation, which saves time and reduces the chance of errors.

The HMC commands can be started either locally on the HMC or remotely by using the **`ssh -l <hmc> <hmc_command>`** command.

Tip: Use the **ssh-keygen** command to create the public and private key-pair on your client. Then, add these keys to the HMC user's key-chain by using the **mkauthkeys --add** command on the HMC.

The following section briefly describes the HMC commands **migrlpar**, **lslparmigr**, and **lssyscfg** used for managing Live Partition Mobility, including a few usage examples.

For detailed command syntax information and further examples, see either the HMC command online help using the **man <command_name>** command, or the *HMC commands* topic in the *IBM Power Systems Hardware Information Center* at:

http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7edm/p7edm_kickoff.htm

The migrlpar command

This command is used to validate, initiate, stop, and recover a partition migration. The same command, syntax, and options are used for all types of migrations. The HMC determines which type of migration to run based on the state of the partition that is referenced in the command.

The following examples illustrate how the **migrlpar** command can be used:

- To migrate the partition myLPAR from the system srcSystem to the destSystem using the default MSPs and adapter maps, use the following command:

```
$ migrlpar -o m srcSystem -t destSystem -p myLPAR
```

- In an environment with multiple mover service partitions on the source and destination, you can specify which mover service partitions to use in a validation or migration operation. The following command validates the migration in the previous example with specific mover service partitions. You can use both partition names and partition IDs on the same command:

```
$ migrlpar -o v -m srcSystem -t destSystem -p myLPAR \
-i source_msp_id=2,dest_msp_name=VIOS2_L10
```

- When the destination system has multiple shared-processor pools, you can stipulate to which shared-processor pool the moving partition is assigned at the destination with either of the following commands:

```
$ migrlpar -o m  -m srcSystem -t destSystem -p myLPAR -i
"shared_proc_pool_id=1"
```

```
$ migrlpar -o m -m srcSystem -t destSystem -p myLPAR -i
"shared_proc_pool_name=DefaultPool"
```

The capacity of the chosen shared-processor pool must be sufficient to accommodate the migrating partition. Otherwise, the migration operation fails. If a shared-processor pool is not chosen the default pool will be used, even if the default pool is not being used on the source system.

- The syntax to stop a partition migration is:

```
$ migrlpar -o s -m srcSystem -p MyLPAR
```

- The syntax to recover a failed migration is:

```
$ migrlpar -o r -m srcSystem -p MyLPAR
```

- You can use the **--force** flag on the **recover** command. However, only do so when the partition migration fails, leaving the partition definition on both the source and destination systems.
- When migrating suspended partitions you must use the --protectstorage 2 option. This indicates that you are aware that once the partition is on the destination system the partitions virtual storage devices can be accidentally reassigned.

The lslparmigr command

Use the **lslparmigr** command to show the state of running migrations or to show managed mover service partitions, Virtual I/O Servers, and adapter mappings.

The following examples illustrate how the **lslparmigr** command can be used.

System migration information

To display the migration capabilities of a system, use the following syntax:

```
$ lslparmigr -r sys -m mySystem
```

This command produces this output:

```
inactive_lpar_mobility_capable=1,num_inactive_migrations_supported=40,n
um_inactive_migrations_in_progress=1,active_lpar_mobility_capable=1,num
_active_migrations_supported=40,num_active_migrations_in_progress=0
```

In this example, the system is capable of both active and inactive migration, and there is one inactive partition migration in progress. By using the **-F** flag, the same information is produced in a CSV format:

```
$ lslparmigr -r sys -m mySystem -F
```

This command produces this output:

```
1,40,1,1,40,0
```

These attribute values are the same as in the preceding example, without the attribute identifier. This format is appropriate for parsing or for importing into a spreadsheet. Adding the **--header** flag prints column headers on the first line:

If you are only interested in specific attributes, you can specify these as options to the **-F** flag. For example, if you want to know just the number of active and inactive migrations in progress, use the following command:

```
$ lslparmigr -r sys -m mySystem -F \
num_active_migrations_in_progress,num_inactive_migrations_in_progress
```

This command produces the following results, which indicate that there are no active migrations and one inactive migration running:

```
0,1
```

If you want a space instead of a comma to separate values, surround the attributes with double quotation marks.

Remote migration information

To show the remote migration information of the HMC, use the **-r manager** option:

```
$ lslparmigr -r manager
```

This option produces this output:

```
remote_lpar_mobility_capable=1
```

Here, the command supplies only one attribute for the user on the HMC from which it is run. The attribute `remote_lpar_mobility_capable` displays a value of `1` if the HMC can run migrations to a remote HMC. Conversely, a value of `0` indicates that the HMC is incapable of remote migrations.

You can also use the **-F** flag followed by the attribute to limit the output of the command to the value:

```
$ lslparmigr -r manager -F remote_lpar_mobility_capable
```

This command produces this output:

```
1
```

Partition migration information

To show the migration information of the logical partitions of a managed system, use the `-r lpar` option:

```
$ lslparmigr -r lpar -m mySystem
```

This option produces this output:

```
name=QA,lpar_id=4,migration_state=Not Migrating
name=VIOS1_L10,lpar_id=1,migration_state=Not Migrating
name=PROD,lpar_id=3,migration_state=Migration Starting,
migration_type=inactive,dest_sys_name=9117-MMA-SN100F6A0-L9,
dest_lpar_id=65535
```

Here, the system `mySystem` is hosting three partitions, `QA`, `VIOS1_L10`, and `PROD`. Of these, the `PROD` partition is in the `Starting` state of an inactive migration as indicated by the `migration_state` and `migration_type` attributes. When the command was run, the migration the ID of the destination partition was not chosen, as seen by the `65535` value for the `dest_lpar_id` parameter.

Use the **`-filter`** flag to limit the output to a certain set of partitions with either the `lpar_names` or the `lpar_ids` attributes:

```
$ lslparmigr -r lpar -m mySystem --filter lpar_ids=3
```

This flag produces this output:

```
name=PROD,lpar_id=3,migration_state=Migration Starting,
migration_type=inactive,dest_sys_name=9117-MMA-SN100F6A0-L9,
dest_lpar_id=7
```

Here, the output information is limited to the partition with ID=3, which is the one running the inactive migration. Note that the `dest_lpar_id` is chosen.

You can use the **`-F`** flag to generate the same information in CSV format or to limit the output:

```
$ lslparmigr -r lpar -m mySystem --filter lpar_ids=3 -F
```

This flag produces this output:

```
PROD,3,Migration Starting,inactive,9117-MMA-SN101F170-L10,
9117-MMA-SN100F6A0-L9,3,7,,unavailable,,unavailable
```

Here the **`-F`** flag, without extra parameters, has printed all the attributes. In the example, the last four fields of output pertain to the MSPs. Because the partition in question is undergoing an inactive migration, no MSPs are involved and these fields are empty. You can use the **`--header`** flag with the **`-F`** flag to print a line of column headers at the start of the output.

Mover service partition information

The **`-r msp`** option shows the possible mover service partitions for a migration:

```
$ lslparmigr -r msp -m vrml11-fsp -t vrml13-fsp --filter "lpar_names=vrml11lp01"
```

This option produces this output:

```
source_msp_name=vrml11-vios2,source_msp_id=2,source_msp_num_active_migr
ations_configured=16,source_msp_num_active_migrations_supported=16,sour
ce_msp_num_active_migrations_in_progress=0,source_msp_concurr_migration
_perf_level=3,"dest_msp_names=vrml13-vios2,vrml13-vios1","dest_msp_ids=
2,1","dest_msp_num_active_migrations_configured=vrml13-vios2/2/4,vrml13
-vios1/1/4","dest_msp_num_active_migrations_supported=vrml13-vios2/2/16
,vrml13-vios1/1/4","dest_msp_num_active_migrations_in_progress=vrml13-v
ios2/2/0,vrml13-vios1/1/0","dest_msp_concurr_migration_perf_level=vrml1
3-vios2/2/4,vrml13-vios1/1/unavailable","ipaddr_mappings=9.3.46.70//2/v
rml13-vios2/9.3.46.77/,9.3.46.70//1/vrml13-vios1/9.3.46.76/"
source_msp_name=vrml11-vios1,source_msp_id=1,source_msp_num_active_migr
ations_configured=4,source_msp_num_active_migrations_supported=4,source
_msp_num_active_migrations_in_progress=0,source_msp_concurr_migration_p
erf_level=unavailable,"dest_msp_names=vrml13-vios2,vrml13-vios1","dest_
msp_ids=2,1","dest_msp_num_active_migrations_configured=vrml13-vios2/2/
4,vrml13-vios1/1/4","dest_msp_num_active_migrations_supported=vrml13-vi
os2/2/16,vrml13-vios1/1/4","dest_msp_num_active_migrations_in_progress=
vrml13-vios2/2/0,vrml13-vios1/1/0","dest_msp_concurr_migration_perf_lev
el=vrml13-vios2/2/4,vrml13-vios1/1/unavailable","ipaddr_mappings=9.3.46
.69//2/vrml13-vios2/9.3.46.77/,9.3.46.69//1/vrml13-vios1/9.3.46.76/"
```

If you move the partition TEST from `srcSystem` to `destSystem`, the following are true:

- There is a mover service partition on the source (VIOS1_L9).
- There is a mover service partition on the destination (VIOS1_L10).
- If the migration uses VIOS1_L9 on the source, VIOS1_L10 can be used on the destination.

This approach gives one possible mover service partition combination for the migration.

Shared processor pool

Use the **`-r procpool`** option to display the shared processor pools capable of hosting the client partition on the destination server. The **`-F`** flag (optional) can be used to format or limit the output, as follows:

```
$ lslparmigr -r procpool -m srcSystem -t destSystem \
--filter lpar_names=TEST -F shared_proc_pool_ids
```

The flag produces this output:

```
1,0
```

This output indicates that processor pool IDs `1` and `0` can host the client partition called TEST. The command requires the **`-m`**, **`-t`**, and **`--filter`** flags. The

--filter flag requires that you use either the `lpar_ids` or `lpar_names` attributes to identify the client partition. You can specify only one client partition at a time.

The command can also be used to identify shared-processor pools available on remote HMC migrations with the **--ip** and **-u** flags to specify the remote HMC and remote user ID. Also, without the **-F** flag you are given detailed output of the attribute and values, as shown in the following example. You are also given remote HMC specification and using `lpar_ids` to specify the client partition:

```
$ lslparmigr -r procpool -m srcSystem --ip 9.3.5.180 \
-u hscroot -t destSystem --filter lpar_ids=2
```

This command shows that you are communicating with an HMC with IP address `9.3.5.180` using the HMC's user ID `hscroot`. The command then checks the remote HMC for the best system-managed system for possible processor pools. It produces the following output:

```
"shared_proc_pool_ids=1,0","shared_proc_pool_names=SharedPool01,Default
Pool"
```

Here, the system is showing that two shared-processor pools are possible destinations for the client partition.

Virtual I/O Server

Use the **-r virtualio** option to display the possible virtual adapter mappings for a migration. The **-F** flag can be used to format or limit the output, as follows:

```
$ lslparmigr -r virtualio -m srcSystem -t destSystem \
--filter lpar_names=TEST -F suggested_virtual_scsi_mappings
```

The flag produces this output:

```
40/VIOS1_L10/1
```

This output indicates that if you migrate the client partition called TEST from the `srcSystem` to `destSystem`, the suggested virtual SCSI adapter mapping connects the client virtual adapter in slot 40 to the Virtual I/O Server called `VIOS1_L10`, which has a partition ID of 1, on the destination system.

The lssyscfg command

The **lssyscfg** command is a native HMC command that supports Live Partition Mobility.

The **lssyscfg -r sys** command displays two attributes, `active_lpar_mobility_capable` and `inactive_lpar_mobility_capable`. These attributes can have a value of either `0` (incapable) or `1` (capable).

The l**ssyscfg -r lpar -m *<managed_system>*** command displays the `msp` and `time_ref` partition attributes on Virtual I/O Server partitions that can participate in active or suspended partition migrations. The `msp` attribute has a value of 1 when enabled as mover service partition, and 0 when it is not enabled.

The mkauthkeys command

The **mkauthkeys** command is used for retrieval, removal, and validation of SSH key authentication to a remote system. Examples

To get the remote HMC user's SSH public key, use the **-g** flag:

```
$ mkauthkeys --ip rmtHostName -u hscroot -g
```

You can also specify a preferred authentication method. To choose between DSA authentication and RSA authentication key usage, use the **-t** flag:

- `$ mkauthkeys --ip rmtHostName -u hscroot -t rsa`
- `$ mkauthkeys --ip rmtHostName -u hscroot -t dsa`

In some cases, you can choose to remove the authentication keys, which you can do by using the **mkauthkeys** command with the **-r** flag:

```
$ mkauthkeys -r ccfw@rmtHostName
```

The HMC stores the key as user called `ccfw`. It is not stored as the user ID that you specified in the steps to retrieve the authentication keys. Also, note that the remote HMC's host name must be specified in this command. If DNS is unable to resolve the host name and you used the IP address to configure the authentication, use the actual IP address in the place of `rmtHostName`. This is the only situation when you must use the actual IP address.

You can use the **--test** flag to check whether authentication is properly configured to the remote HMC:

```
$ mkauthkeys --ip rmtHostName -u hscroot --test
```

The command returns the following error if the keys were not configured properly:

```
HSCL3653 The Secure Shell (SSH) communication configuration between the
source and target Hardware Management Consoles has not been set up
properly for user hscroot. Please run the mkauthkeys command to set up
the SSH communication authentication keys.
```

Complex example of using LPM commands

Example 14-1 on page 576 provides a detailed example of how the Live Partition Mobility commands can be used. This script fragment moves all the migratable partitions from one system to another. The example assumes that two

environment variables, SRC_SERVER and DEST_SERVER, point to the system to empty and the system to load.

Important: Starting in HMC version 7 release 7.8.0 the **migrlpar** command allows you to migrate all partitions off a system with a single command. See *IBM PowerVM Enhancements What is new in 2013*, SG24-8198.

http://www.redbooks.ibm.com/abstracts/sg248198.html?Open

The algorithm starts by listing all the partitions on SRC_SERVER. It then filters out any Virtual I/O Server partitions and partitions that are already migrating. For each remaining partition, it starts a migration operation to DEST_SERVER. In this example, the migrations take place sequentially. Running them in parallel is acceptable if there are no more than four concurrent active and suspended migrations per mover service partition.

In Virtual I/O Server 2.2.3.0 the maximum number of concurrent migrations has been increased to 8.

How it works

The script starts by checking that both the source and destination systems are mobility capable. To do so, it uses the new attributes that are given in the **lssyscfg** command. It then uses the **lslparmigr** command to list all the partitions on the system. It uses this list as an outer loop for the rest of the script.

The program then performs a number of elementary checks:

- The source and destination must be capable of mobility.

  The **lssyscfg** command shows the mobility capability attribute.

- Only partitions of type aixlinux or os400 can be migrated. You cannot migrate vioserver type partitions.

  The script uses the **lssyscfg** command to ascertain the partition type.

- Determines whether to avoid migrating a partition that is already migrating.

  The script reuses the **lslparmigr** command for this.

- Validate the partition migration.

  The script uses the **migrlpar -v** and checks the return code.

If all the checks pass, the migration is started with the **migrlpar** command. The code snippet does some elementary error checking. If **migrlpar** returns a non-zero value, a recovery is attempted using the **migrlpar -o r** command.

Example 14-1 Script fragment to migrate all partitions on a system

```
#
# Get the mobility capabilities of the source and destination systems
#
SRC_CAP=$(lssyscfg -r sys -m $SRC_SERVER \
          -F active_lpar_mobility_capable,inactive_lpar_mobility_capable)
DEST_CAP=$(lssyscfg -r sys -m $DEST_SERVER \
          -F active_lpar_mobility_capable,inactive_lpar_mobility_capable)

#
#
# Make sure that they are both capapble of active and inactive migration
#
   if [ $SRC_CAP = $DEST_CAP ] && [ $SRC_CAP = "1,1" ]
   then
#
# List all the partitions on the source system
#
      for LPAR in $(lslparmigr -r lpar -m $SRC_SERVER -F name)
      do
#
# Only migrate “aixlinux” and “os400” partitions. VIO servers cannot be migrated
#
         LPAR_ENV=$(lssyscfg -r lpar -m $SRC_SERVER \
                    --filter lpar_names=$LPAR -F lpar_env)
         if [ $LPAR_ENV = "aixlinux" ] || [ $LPAR_ENV = "os400" ]
         then
#
# Make sure that the partition is not already migrating
#
            LPAR_STATE=$(lslparmigr -r lpar -m $SRC_SERVER --filter lpar_names=$LPAR -F
migration_state)
            if [ "$LPAR_STATE" = "Not Migrating" ]
            then
#
# Perform a validation to see if there’s a good chance of success
#
               migrlpar -o v -m $SRC_SERVER -t $DEST_SERVER -p $LPAR
               RC=$?
               if [ $RC -ne 0 ]
               then
                 echo "Validation failed. Cannont migrate partition $LPAR"
               else
#
# Everything looks good, let’s do it...
#
                  echo "migrating $LPAR from $SRC_SERVER to $DEST_SERVER"
                  migrlpar -o m -m $SRC_SERVER -t $DEST_SERVER -p $LPAR
```

```
                  RC=$?
                  if [ $RC -ne 0 ]
                  then
#
# Something went wrong, let's try to recover
#
                     echo "There was an error RC = $RC . Attempting recovery"
                     migrlpar -o r -m $SRC_SERVER -p $LPAR
                     break
                  fi
               fi
            fi
         fi
      done
fi
```

14.1.3 Making applications migration aware

Software applications might be designed to recognize and adapt to changes in the system hardware after being moved from one system to another.

Most software applications that run in AIX, IBM i, and Linux logical partitions do not require any changes to work correctly during active partition mobility. Some applications might have dependencies on characteristics that change between the source and destination servers, and other applications might need to adjust to support the migration.

IBM PowerHA SystemMirror® for AIX and PowerHA SystemMirror for i clustering software is aware of partition mobility. You can move a mobile partition, either an active/primary one, or a standby/backup partition that is running the PowerHA clustering software, to another server without restarting the PowerHA clustering software.

The following sections show different ways of making applications migration aware and registering user-defined programs with certain migration operations:

- Making AIX programs migration aware using APIs
- Making AIX applications migration aware using scripts
- Making AIX kernel extension migration aware
- Registering IBM i programs with migration exit points

Making AIX programs migration aware using APIs

Application programming interfaces are provided to make programs migration-aware. The SIGRECONFIG signal is sent to all applications at each of the three migration phases. Applications can watch (trap) this signal and use the

DLPAR-API system calls to learn more about the operation in progress. If your program does trap the SIGRECONFIG signal, it is notified of all dynamic-reconfiguration operations, not just Live Partition Mobility events.

Consideration: An application must not block the SIGRECONFIG signal, and the signal must be handled in a timely manner. The dynamic LPAR and Live Partition Mobility infrastructure wait a short time for a reply from applications. If no response occurs after this amount of time, the system assumes all is well and proceeds to the next phase. You can speed up a migration or dynamic reconfiguration operation by acknowledging the SIGRECONFIG event even if your application takes no action.

Applications must perform the following operations to be notified of a Live Partition Mobility operation:

1. Catch the SIGRECONFIG signal by using the sigaction() or sigwait() system calls. The default action is to ignore the signal.
2. Control the signal mask of at least one of the application's threads and the priority of the handling thread such that the signal can be delivered and handled promptly.
3. Uses the dr_reconfig() system call, through the signal handler, to determine the nature of the reconfiguration event and other pertinent information. For the check phase, the application passes DR_RECONFIG_DONE to accept a migration, or DR_EVENT_FAIL to refuse. Only applications with root authority can refuse a migration.

The dr_reconfig() system call has been modified to support partition migration. The returned dr_info structure includes the following bit fields:

- migrate
- partition

These fields are for the new migration action and the partition object that is the object of the action.

The code snippet in Example 14-2 shows how dr_reconfig() might be used. This code runs in a signal-handling thread.

Example 14-2 SIGRECONFIG signal-handling thread

```
#include <signal.h>
#include <sys/dr.h>
   :
   :
   struct dr_info  drInfo;       // For event-related information
   struct sigset_t signalSet;    // The signal set to wait on
```

```
   int             signalId;     // Identifies  signal was received
   int             reconfigFlag  // For accepting or refusing the DR
   int             rc;           // return code

   // Initialise the signal set
   SIGINITSET(signalSet);

   // Add the SIGRECONFIG to the signal set
   SIGADDSET(signalSet, SIGRECONFIG);

   // loop forever
   while (1) {
      // Wait on signals in signal set
      sigwait(&signalSet, &signalId);
      if (signalID == SIGRECONFIG) {
         if (rc = dr_reconfig(DR_QUERY, &drInfo)) {
            // handle the error
         } else {
            if (drInfo.migrate) {
               if {drInfo.check) {
                  /*
                   * If migration OK  reconfigFlag = DR_RECONFIG_DONE
                   * If migration NOK reconfigFlag = DR_EVENT_FAIL
                   */
                   rc = dr_reconfig(reconfigFlag, &drInfo);
                } else if (drInfo.pre) {
                  /*
                   * Prepare the application for migration
                   */
                   rc = dr_reconfig(DR_RECONFIG_DONE, &drInfo);
                } else if (drInfo.post) {
                   /*
                    * We're being woken up on the destination
                    * Check new environment and resume normal service
                    */
                } else {
                    // Handle the error cases
                }
             } else {
                // It's not a migration.  Handle or ignore the DR
         }
   }
}
```

You can use the sysconf() system call to check the system configuration on the destination system. The _system_configuration structure has been modified to include the following fields:

icache_size	Size of the L1 instruction cache
icache_asc	Associativity of the L1 instruction cache
dcache_size	Size of the L1 data cache
dcache_asc	Associativity of the L1 data cache
L2_cache_size	Size of the L2 cache
L2_cache_asc	Associativity of the L2 cache
itlb_size	Instruction translation look-aside buffer size
itlb_asc	Instruction translation look-aside buffer associativity
dtlb_size	Data translation look-aside buffer size
dtlb_asc	Data translation look-aside buffer associativity
tlb_attrib	Translation look-aside buffer attributes
slb_size	Segment look-aside buffer size

These fields are updated after the partition arrives at the destination system to reflect the underlying physical processor characteristics. In this fashion, applications that are moved from one processor architecture to another can dynamically adapt themselves to their running environment. All new processor features, such as the single-instruction multiple-data (SIMD) and decimal floating point instructions, are exposed through the _system_configuration structure and the lpar_get_info() system call.

The lpar_get_info() call returns two capabilities, which are defined in <sys/dr.h>:

LPAR_INFO1_MSP_CAPABLE	If the partition is a Virtual I/O Server partition, this capability indicates that the partition is also a mover service partition.
LPAR_INFO1_PMIG_CAPABLE	Indicates whether the partition is capable of migration.

Making AIX applications migration aware using scripts

Dynamic reconfiguration scripts allow you to cleanly quiesce and restart your applications over a migration. You can register your own scripts with the dynamic reconfiguration infrastructure by using the **drmgr** command. The command copies the scripts to a private repository, the default location of which is `/usr/lib/dr/scripts/all`.

The scripts can be implemented in any interpreted (scripted) or compiled language. The **drmgr** command is detailed in the IBM Information Center at:

http://publib.boulder.ibm.com/infocenter/pseries/v5r3/index.jsp?topic=/com.ibm.aix.cmds/doc/aixcmds2/drmgr.htm

The Dynamic reconfiguration scripts have the following syntax:

```
[env_variable1=value ...] scriptname command [param1 ...]
```

The input variables are set as environment variables on the command line, followed by the name of the script to be started and any additional parameters.

Live Partition Mobility introduces four commands, which are listed in Table 14-2.

Table 14-2 Dynamic reconfiguration script commands for migration

Command and parameter	Description
checkmigrate <resource>	Used to indicate whether the migration continues or not. A script might indicate that a migration should not continue if the application is dependent upon an invariable running environment. The script is called with this command at the check-migration phase.
premigrate <resource>	At this point the migration is initiated. The script can reconfigure or suspend an application to facilitate the migration process. The script is called with this command at the prepare-migration phase.
postmigrate <resource>	This command is called after migration is completed. The script can reconfigure or resume applications that were changed or suspended in the prepare phase. The script is called with this command in the post-migration phase.
undopremigrate <resource>	If an error is encountered during the check phase, the script is called with this command to roll back any actions that were taken in the **checkmigrate** command in preparation for the migration.

In addition to the script commands, a `pmig` resource type indicates a partition migration operation. The **register** command of your dynamic LPAR scripts can choose to handle this resource type. A script that supports partition migration writes out the name-value pair DR_RESOURCE=pmig when it is started with the **register** command. A dynamic LPAR script can be registered to support only partition migration. No new environment variables are passed to the dynamic LPAR scripts for Live Partition Mobility support.

The code in Example 14-3 shows a Korn shell script that detects the partition migration reconfiguration events. For this example, the script logs the called command to a file.

Example 14-3 Outline Korn shell dynamic LPAR script for Live Partition Mobility

```
#!/usr/bin/ksh

if [[ $# -eq 0 ]]
then
   echo "DR_ERROR=Script usage error"
   exit 1
fi

ret_code=0
command=$1

case $command in
   scriptinfo )
      echo "DR_VERSION=1.0"
      echo "DR_DATE=27032007"
      echo "DR_SCRIPTINFO=partition migration test script"
      echo "DR_VENDOR=IBM"
      echo "SCRIPTINFO" >> /tmp/migration.log;;

   usage )
      echo "DR_USAGE=$0 command [parameter]"
      echo "USAGE" >> /tmp/migration.log;;

   register )
      echo "DR_RESOURCE=pmig"
      echo "REGISTER" >> /tmp/migration.log;;

   checkmigrate )
      echo "CHECK_MIGRATE" >> /tmp/migration.log;;

   premigrate )
      echo "PRE_MIGRATE" >> /tmp/migration.log;;

   postmigrate )
       echo "POST_MIGRATE" >> /tmp/migration.log;;

   undopremigrate )
      echo "UNDO_CHECK_MIGRATE" >> /tmp/migration.log;;

   * )
      echo "*** UNSUPPORTED *** : $command" >> /tmp/migration.log
      ret_code=10;;
esac
```

```
exit $ret_code
```

If the file name of the script is `migrate.sh`, register it with the dynamic reconfiguration infrastructure by using the following command:

```
# drmgr -i ./migrate.sh
```

Use the **`drmgr -l`** command to confirm script registration, as shown in Example 14-4. In this example, you can see the output from the **`scriptinfo`**, **`register`**, and **`usage`** commands of the shell script.

Example 14-4 Listing the registered dynamic LPAR scripts

```
# drmgr -l
DR Install Root Directory: /usr/lib/dr/scripts
Syslog ID: DRMGR
------------------------------------------------------------
/usr/lib/dr/scripts/all/migrate.sh               partition migration test
script
        Vendor:IBM,     Version:1.0,    Date:27032007
        Script Timeout:10,      Admin Override Timeout:0
        Memory DR Percentage:100
        Resources Supported:
                Resource Name: pmig      Resource Usage:
/usr/lib/dr/scripts/all/migrate.sh command [parameter]
------------------------------------------------------------
```

Making AIX kernel extension migration aware

AIX kernel extensions can register to be notified of migration events. The notification mechanism uses the standard dynamic reconfiguration mechanism, which is the reconfig_register() kernel service. The following example shows the service interface signature:

```
int reconfig_register_ext(handler, actions, h_arg, h_token, name)
int (*handler)(void*, void*, long long action, void* dri);
long long actions;
void* h_arg;
ulong *h_token;
char* name;
```

The actions parameter supports the following values for mobility awareness:

- DR_MIGRATE_CHECK
- DR_MIGRATE_PRE
- DR_MIGRATE_POST
- DR_MIGRATE_POST_ERROR

The following is the interface to the handler:

```
int handler(void* event, void* h_arg, long long action, void*
resource_info);
```

The **action** parameter indicates the specific reconfiguration operation that is being performed, for example, DR_MIGRATE_PRE. The **resource_info** parameter maps to the following structure for partition migration:

```
struct dri_pmig {
          int version;
          int destination_lpid;
          long long streamid
}
```

The version number is changed if more parameters are added to this structure. The `destination_lpid` and `streamid` fields are not available for the check phase.

The interfaces to the reconfig_unregister() and reconfig_complete() kernel services are not changed by Live Partition Mobility.

Registering IBM i programs with migration exit points

On IBM i, the WRKREGINF command can be used to register user programs with the following exit points defined for migration and hibernation operations of an IBM i partition:

- QIBM_QWC_SUSPEND

  The *Suspend System* exit program is called before the system becomes temporarily unavailable because the partition is being migrated to another machine or the partition is being hibernated.

  For more information, see the IBM i Information Center at:

  http://publib.boulder.ibm.com/infocenter/iseries/v7r1m0/index.jsp?topic=%2Fapis%2Fxsuspend.htm

- QIBM_QWC_RESUME

  The *Resume System* exit program is called when the system becomes available again after the partition was migrated to another machine or the partition resumed from hibernation.

  For more information, see the IBM i Information Center at:

  http://publib.boulder.ibm.com/infocenter/iseries/v7r1m0/index.jsp?topic=%2Fapis%2Fxresume.htm

14.1.4 Migration recovery

This section describes topics that are related to recovery procedures to be followed when errors occur during migration of a logical partition.

Recovery

Live Partition Mobility is designed to verify whether a requested migration can be run, and to monitor all migration processes. If a running migration cannot be completed, a rollback procedure is performed to undo all configuration changes applied.

A partition migration might be prevented from running for two main reasons:

- The migration is not valid and does not meet prerequisites.
- An external event prevents a migration component from completing its job.

The migration validation that is described in “Performing validation for partition mobility” on page 543 checks all prerequisites. It can be explicitly run at any time, and does not affect the mobile partition.

Perform a validation before you request any migration. The migration process, however, performs another validation before starting any configuration changes.

After the migration begins, the HMC manages the configuration changes and monitors the status of all involved components. If any error occurs, recovery actions automatically begin.

When the HMC cannot run a recovery, administrator intervention is required to perform problem determination and issue final recovery steps. This situation might occur when the HMC cannot contact a migration component (for example, the mobile partition, a Virtual I/O Server, or a system service processor) because of a network problem or an operator error. After a timeout, an error message is provided that requests a recovery.

When a recovery is required, the mobile partition name can be displayed on both the source and the destination system. The partition is either powered down (inactive or suspended migration) or really working only on one of the two systems (active migration). Configuration cleanup is made during recovery.

Although a mobile partition requires a recovery, its configuration cannot be changed to prevent any attempt to modify its state before its state is returned to normal operation. Activating the same partition on two systems is not possible.

To start recovery, select the migrating partition on the source system and then selecting **Operations** → **Mobility** → **Recover**, as shown in Figure 14-26.

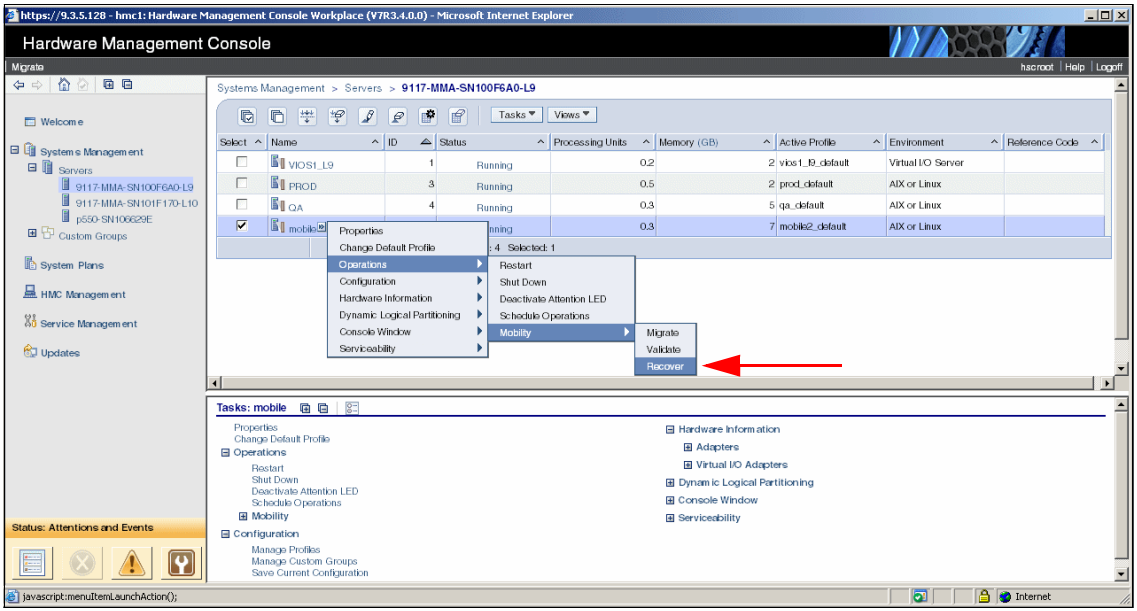


Figure 14-26 Recovery menu

A window opens, similar to the one shown in Figure 14-27, requesting recovery confirmation. Click **Recover** to start a recovery.

Note: The recovery should always be attempted from the source system. Only attempt the recovery from the destination system if the source system is down or when using remote HMCs and the HMC attached to source system is down.

Note: Use the **Force recover** check box only in these circumstances:

- The HMC cannot contact one of the migration components that requires a new configuration
- The migration was started by another HMC
- A normal recovery does not succeed

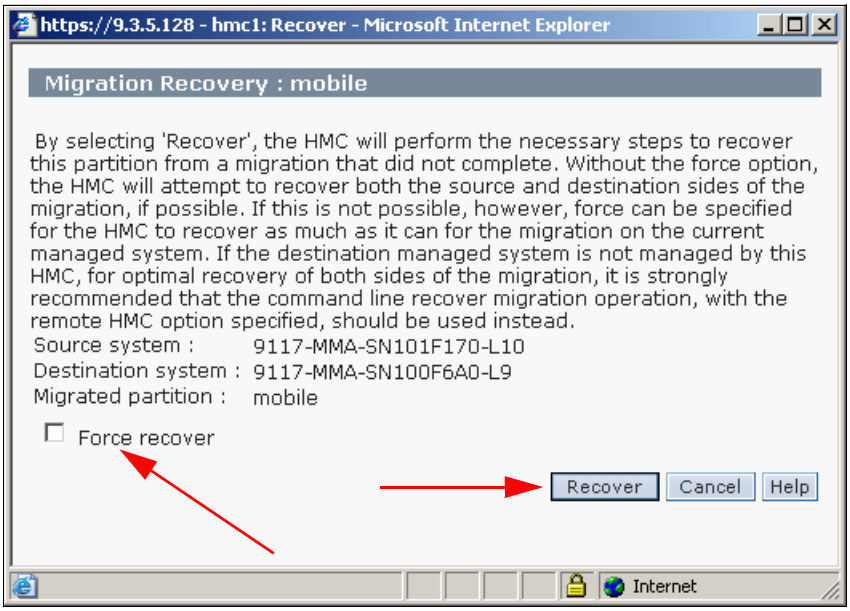


Figure 14-27 Recovery window

The same actions that are performed on the GUI can be performed with the **`migrlpar`** command on the HMC's command line. For more information, see 14.1.2, "HMC commands for Live Partition Mobility" on page 567.

After a successful recovery, the partition returns to normal operation state and changes to its configuration are then allowed. If the migration is run again, the validation phase detects the component that prevented the migration and selects alternate elements or provides a validation error.

A recovery example

As an example, a network outage occurs during an active partition migration.

During an active migration, there is a partition state transfer through the network between the source and destination mover service partitions. The mobile partition continues running on the source system while its state is copied on the destination system. Then, it is briefly suspended on the source and immediately reactivated on the destination.

The network connection of one mover service partition was unplugged in the middle of a state transfer. Several attempts were required to create this scenario because the migration on the partition (2 GB of memory) was extremely fast.

In the HMC GUI, the migration process fails and an error message is displayed.

Because the migration stopped in the middle of the state transfer, the partition configuration on the two involved systems is kept in the migrating status, waiting for the administrator to identify the problem and decide how to continue.

In the HMC, the status of the migrating partition, `mobile`, is present in both systems, while it is active only on the source system. On the destination system, only the shell of the partition is present. You can view the situation by expanding **Systems Management** → **Custom Groups** → **All partitions**. In the content area, a situation similar to Figure 14-28 is shown.

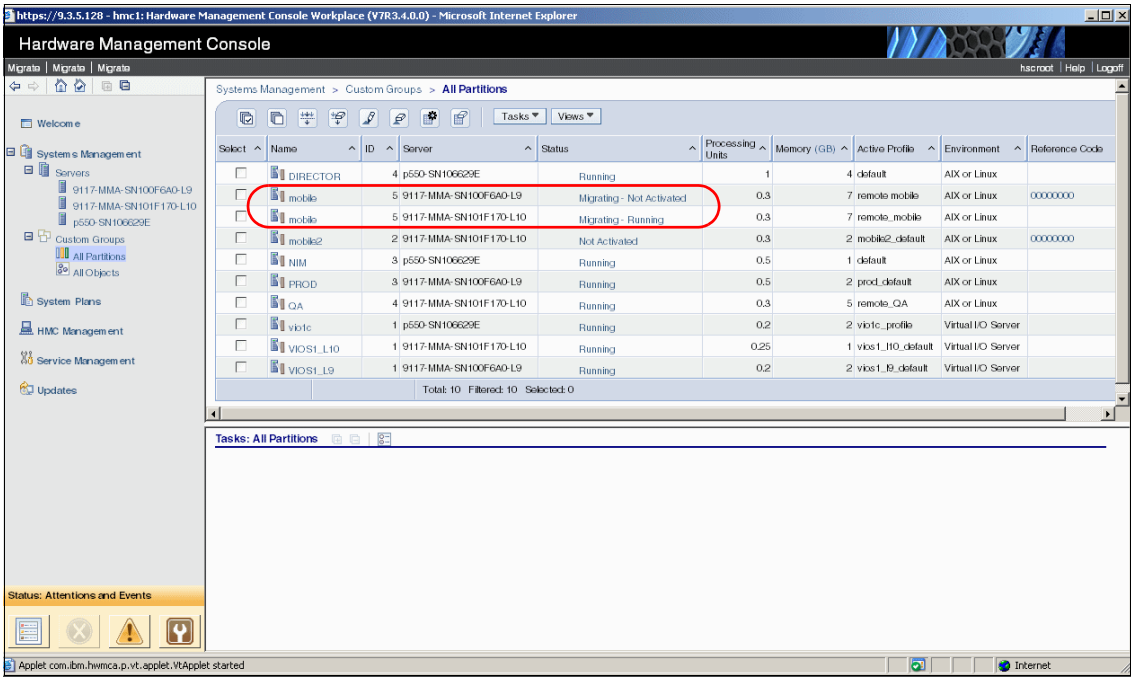


Figure 14-28 Interrupted active migration status

The applications that are running on the partition are not affected by the network outage, and are running on the source system. The only visible effect is on the partition's error log that shows the start and the abort of the migration, as shown in Example 14-5. No action is required on the partition.

Example 14-5 Migrating partition's error log after aborted migration

```
[mobile]# errpt
IDENTIFIER TIMESTAMP  T C RESOURCE_NAME  DESCRIPTION
5E075ADF   1118180308 I S pmig           Client Partition Migration Aborted
08917DC6   1118180208 I S pmig           Client Partition Migration Started
```

Both Virtual I/O Servers, using a single mover service partition, recorded the event in their error logs. On the Virtual I/O Server, where the cable was unplugged, you see both the physical network error and the mover service partition communication error, as indicated in Example 14-6.

Example 14-6 Mover service partition with network outage

```
$ errlog
IDENTIFIER TIMESTAMP  T C RESOURCE_NAME  DESCRIPTION
427E17BD   1118181908 P S Migration      Migration aborted: MSP-MSP connection do
0B41DD00   1118181708 I H ent4           ADAPTER FAILURE
...
```

The other Virtual I/O Server only shows the communication error of the mover service partition because no physical error was created, as indicated in Example 14-7.

Example 14-7 Mover service partition with communication error

```
$ errlog
IDENTIFIER TIMESTAMP  T C RESOURCE_NAME  DESCRIPTION
427E17BD   1118182108 P S Migration      Migration aborted: MSP-MSP connection do
...
```

To recover from an interrupted migration, select the mobile partition and select **Operations** → **Mobility** → **Recover**, as shown in Figure 14-26 on page 586.

A window similar to the one shown in Figure 14-27 on page 587 opens. Click **Recover**, and the partition state is cleaned up (normalized). The mobile partition is present only on the source system where it is running. It is removed on the destination system, where it has never been run.

After the network outage is resolved, the migration can be issued again. Wait for the RMC protocol to reset communication between the HMC and the Virtual I/O Server that had the network cable unplugged.

14.2 Monitoring Live Partition Mobility

This section describes basics how to monitor LPM. It includes the following topics:

- Monitoring migration from HMC GUI
- Monitoring migration from the HMC command line
- Monitoring migration from the partitions

14.2.1 Monitoring migration from HMC GUI

When you using the HMC GUI for partition migration, the actual state and progress of migration is shown in the *Partition Migration Status* window as shown in Figure 14-29 on page 591. The percentage indicates the completion of memory state transfer during an *active* or *suspended* migration. During an *inactive* migration, there is no memory management and the value is zero.

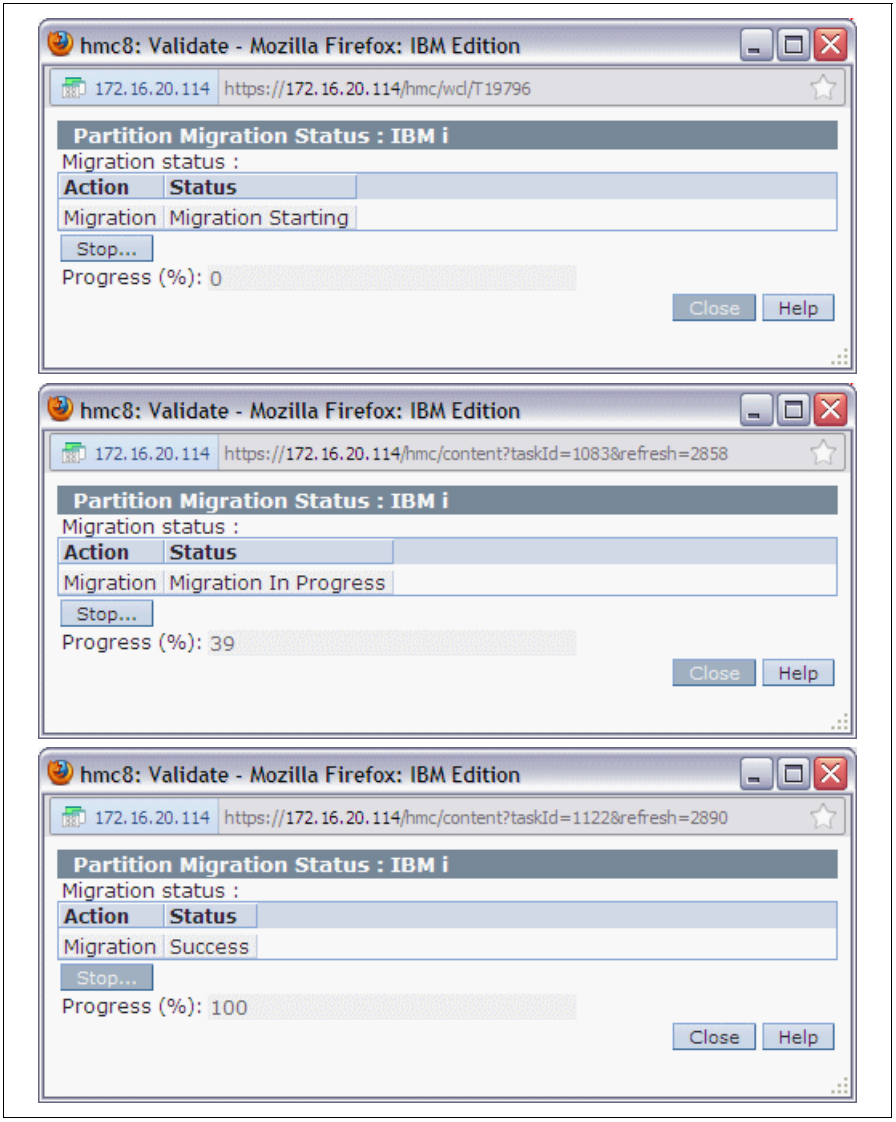


Figure 14-29 HMC GUI windows showing LPM statuses

Progress and reference code information

The migration status is also shown in the *Status* and *Reference Code* information for the partitions that are displayed when you select the managed system and then **Systems Management** → **Servers** as shown in Figure 14-30. The partitions `QA` and `mobile` are undergoing an active migration, and for both partitions the latest reference code is displayed.

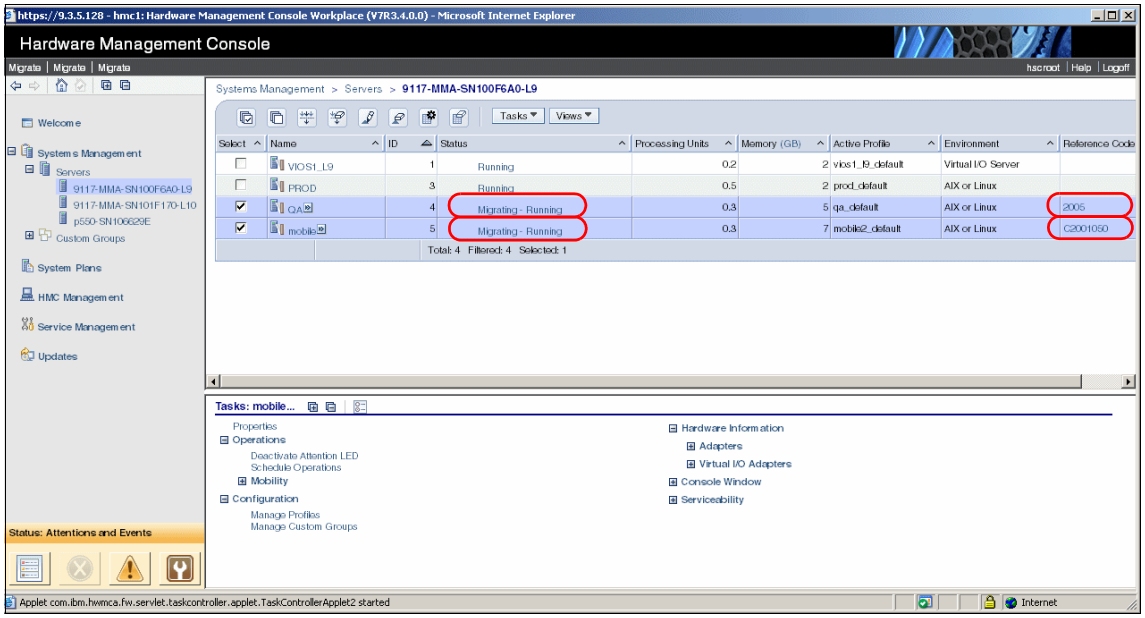


Figure 14-30 Partition reference codes

Reference codes indicate the progress and errors of the migration. When the reference code represents an error, a migration recovery procedure might be required. For more information, see 14.1.4, “Migration recovery” on page 585.

Table 14-3 lists the SRCs that indicate the current state of a partition migration.

Table 14-3 Progress SRCs

Code	Meaning
2005	Partition is running the **`drmgr`** command. The code is displayed on the source server while the partition is waiting to suspend, and on the destination server until the **`drmgr`** processing completes.
C2001020	Partition is the source of an inactive migration.
C2001030	Partition is the target of an inactive migration.

Code	Meaning
C2001040	Partition is the target of an active migration.
C2001080	Partition processors are stopped.
C2001082	Partition processors are restarted.
C20010FF	Migration is complete.
D200A250	Partition has requested to suspend, as part of an active migration.
D200AFFF	Partition migration was canceled.

Table 14-4 lists SRC codes that indicate problems with a partition migration.

Table 14-4 SRC error codes

Code	Meaning
B2001130	Partition migration readiness check failed.
B2001131	Resume of LpQueues failed.
B2001132	Allocated LpEvents failed.
B2001133	Failed to lock the partition configuration of the partition.
B2001134	Failed to unlock the partition configuration of the partition.
B2001140	Processing of transferred data failed.
B2001141	Processing of transferred data failed.
B2001142	Processing of transferred data failed.
B2001143	Processing of transferred data failed.
B2001144	Failed to suspend virtual I/O for the partition.
B2001145	Failed to resume virtual I/O for the partition.
B2001151	Partition attempted a memory dump during migration.
B2002210	Data import failure.
B2002220	Data import failure.
B2008160	PFDS build failure.

14.2.2 Monitoring migration from the HMC command line

The actual migration progress can be monitored by command **lslparmigr**, with attributes `name,migration_state,bytes_transmitted,bytes_remaining` as shown in Example 14-8 for partition `p740_lpar01`. This partition is migrating from machine `p750` to `p740`.

Example 14-8 Checking LPM progress

```
hscroot@hmc8:~> lslparmigr -r lpar -m p750 -F
name,migration_state,bytes_transmitted,bytes_remaining
p750_lpar03,Not Migrating
p750_lpar02,Not Migrating
p750_lpar01,Not Migrating
p750_vios02,Not Migrating
p750_vios01,Not Migrating
p740_lpar01,Migration Starting
10s later ...
hscroot@hmc8:~> lslparmigr -r lpar -m p750 -F
name,migration_state,bytes_transmitted,bytes_remaining
p750_lpar03,Not Migrating
p750_lpar02,Not Migrating
p750_lpar01,Not Migrating
p750_vios02,Not Migrating
p750_vios01,Not Migrating
p740_lpar01,Migration Starting,6754789747,26511859712
10s later
hscroot@hmc8:~> lslparmigr -r lpar -m p750 -F
name,migration_state,bytes_transmitted,bytes_remaining
p750_lpar03,Not Migrating
p750_lpar02,Not Migrating
p750_lpar01,Not Migrating
p750_vios02,Not Migrating
p750_vios01,Not Migrating
```

After a successful migration, check the target system to see `p740_lpar01` back on the target machine as shown in Example 14-9.

Example 14-9 Checking target machine after successful transfer

```
hscroot@hmc8:~> lslparmigr -r lpar -m p740
name=p740_lpar04,lpar_id=6,migration_state=Not Migrating
name=p740_lpar03,lpar_id=5,migration_state=Not Migrating
name=p740_lpar02,lpar_id=4,migration_state=Not Migrating
name=p740_vios04,lpar_id=2,migration_state=Not Migrating
```

```
name=p740_vios03,lpar_id=1,migration_state=Not Migrating
name=p740_lpar01,lpar_id=3,migration_state=Not Migrating
```

14.2.3 Monitoring migration from the partitions

An active migration requires the coordination of the mobile partition and the two Virtual I/O Servers that are selected as mover service partitions. All these objects record migration events in their error logs. These logs can also be used to check for possible events that will prevent the migration from succeeding, such as user interruption or network problems.

Virtual I/O Server migration messages

Migration information is recorded in the error log from the Virtual I/O Servers that acted as a mover service partition.

Example 14-10 shows using the **errlog** command to retrieve the data available on the source mover service partition. The first event in the log states when the mobile partition has been suspended on the source system and activated on the destination system. The second records the successful end of the migration.

Example 14-10 Migration log on source mover service partition

```
$ errlog
IDENTIFIER TIMESTAMP  T C RESOURCE_NAME  DESCRIPTION
3EB09F5A   1118164408 I S Migration      Migration completed successfully
6CB10B8D   1118164408 I S unspecified    Client partition suspend issued
...
```

On the destination mover service partition, the error log registers only the end of the migration, as shown in Example 14-11.

Example 14-11 Migration log on destination mover service partition

```
$ errlog
IDENTIFIER TIMESTAMP  T C RESOURCE_NAME  DESCRIPTION
3EB09F5A   1118164408 I S Migration      Migration completed successfully
...
```

For an active migration, the actual memory data transfer between the systems can also be monitored by using the **vasistat <*vasi_device*>** command for the Virtual Asynchronous Services Interface (VASI) used for migration like `vasi0` as shown in Example 14-12. Consider resetting the VASI statistics using the **-reset** parameter first and optionally using the **-interval <*interval_seconds*>** parameter to enable a periodic refresh for continuous monitoring.

Example 14-12 Using vasistat command to monitor memory data transfer

```
$ vasistat vasi0
-------------------------------------------------------------
VASI STATISTICS (vasi0) :
Elapsed Time: 7 days 21 hours 57 minutes 12 seconds

Transmit to PHYP Statistics:         Receive from PHYP Statistics:
----------------------------         -----------------------------
Packets: 2959834                             Packets: 4950814
Bytes: 23444547770                           Bytes: 588418992
No Buffers: 160                              No Buffers: 0
Transmit Errors: 0                           Receive Errors: 0
Bad Packets: 0                               Bad Packets: 0
Output Calls: 2959994                Interrupts: 6210795
                                     Maximum Buffers Per Interrupt: 76
                                     Average Buffers Per Interrupt: 0
                                             System Buffers: 0

Interrupt Processing Exceeded: 2
Offlevel Interrupt Scheduled: 421804

Average Time Spent in CRQ Send: 17 microseconds
Maximum Time Spent in CRQ Send: 20548 microseconds
Minimum Time Spent in CRQ Send: 0 microseconds

Driver Flags: Up Running 64BitSupport

Maximum Operations: 8
Maximum Receive Pools: 3
Active Operations: 0

DMA Channel: 1001000000000044
Bus ID: 90000340

Local DMA Window:
        Logical I/O Bus Number: 10010000
        TCE Base: 0000000000000000
```

```
        TCE Size: 0000000010000000

Remote DMA Window:
        Logical I/O Bus Number: 00000000
        TCE Base: 0000000000000000
        TCE Size: 0000000000000000

Supported Operation Types: Migration
```

Because the memory data of the mobile partition is transferred between the Virtual I/O Server mover service partitions, the destination Virtual I/O Server mover service partition shows the data transfer in the `Transmit to PHYP Statistics`. The source Virtual I/O Server mover service partition shows it in the `Receive from PHYP Statistics`.

For analysis of partition migration issues, the LPM log on the Virtual I/O Server can be retrieved by using the **`alog -ot lpm > output_filename`** command from `oem_setup_env`.

AIX migration messages

An AIX mobile partition records the start and the end of the migration process. You can extract the data by using the **`errpt`** command as shown in Example 14-13.

Example 14-13 Migration log on mobile partition

```
[mobile:/]# errpt
IDENTIFIER TIMESTAMP  T C RESOURCE_NAME  DESCRIPTION
A5E6DB96   1118164408 I S pmig           Client Partition Migration Completed
08917DC6   1118164408 I S pmig           Client Partition Migration Started
...
```

IBM i migration messages

An IBM i mobile partition logs suspend and resume operations as part of partition migration in the QHST history log and QSYSOPR message queue. Message ID CPI09A5 indicates a suspend as shown in Figure 14-31.

```
                      Additional Message Information

Message ID . . . . . :   CPI09A5       Severity . . . . . :   00
Message type . . . . . :   Information
Date sent  . . . . . :   12/18/12      Time sent  . . . . . :   14:57:12

Message . . . . :   Partition suspend request in progress.
Cause . . . . . :   A request was made to suspend the partition.



                                                                      Bottom
Press Enter to continue.

F3=Exit   F6=Print   F9=Display message details   F12=Cancel
F21=Select assistance level
```

Figure 14-31 IBM i message CPI09A5 for partition suspend operation

Message ID CPI09A8 indicates a resume after migration as shown in Figure 14-32.

```
                      Additional Message Information

Message ID . . . . . :   CPI09A8       Severity . . . . . :   00
Message type . . . . . :   Information
Date sent  . . . . . :   12/18/12      Time sent  . . . . . :   14:58:28

Message . . . . :   Partition resumed after migration.
Cause . . . . . :   The partition has been migrated to another machine.



                                                                      Bottom
Press Enter to continue.

F3=Exit   F6=Print   F9=Display message details   F12=Cancel
F21=Select assistance level
```

Figure 14-32 IBM i message CPI09A8 for partition migration resume operation

15

Dynamic Platform Optimizer

This section describes Dynamic Platform Optimizer (DPO), a PowerVM virtualization feature that enables users to improve partition memory and processor affinity across the logical partitions in a Power Systems server.

This chapter include the following sections:

- Dynamic Platform Optimizer overview
- Dynamic Platform Optimizer requirements
- Managing Dynamic Platform Optimizer
- Monitoring Dynamic Platform Optimizer

Important: For a list of recommended service regarding DPO, see:

http://www-912.ibm.com/s_dir/slkbase.NSF/DocNumber/669691974

15.1 Dynamic Platform Optimizer overview

The Dynamic Platform Optimizer (DPO) is a PowerVM virtualization feature that enables users to improve partition memory and processor placement (affinity) on Power servers. These servers must be running firmware level 760 or later. The user initiates DPO from the Hardware Management Console (HMC) command-line interface. DPO determines an optimal resource placement strategy for the server based on the partition configuration and hardware topology on the system. It then performs a sequence of memory and processor relocations to transform the existing server layout to the optimal layout. This process occurs dynamically while the partitions are running.

The HMC commands include capabilities to initiate DPO, cancel an ongoing DPO operation, monitor operation status, and compute a system-wide affinity score. These commands are discussed in 15.3, “Managing Dynamic Platform Optimizer” on page 602 and 15.4, “Monitoring Dynamic Platform Optimizer” on page 608.

Systems can become suboptimal in terms of processor and memory affinity when reassigning or moving resources. The following are some examples of virtualization features that can affect affinity:

- Dynamic LPAR add and remove of processors or memory
- Live Partition Mobility (LPM)
- Hibernation suspend or resume
- Dynamic creation or deletion of partitions
- Partition processor or memory configuration changes
- CEC Hot Add & Repair Maintenance (CHARM) node repair or add

On larger systems, resources relocation can improve the performance considerably depending on the configuration. DPO provides an affinity score to determine how optimal the system affinity is and what potential affinity can be achieved. These are describes in the following sections.

Note: Single-socket systems (such as the Power 710 Express) cannot take advantage of DPO.

Figure 15-1 illustrates how DPO moves resources to improve the best system affinity possible for a system configuration. The gray segments are unused system resources.

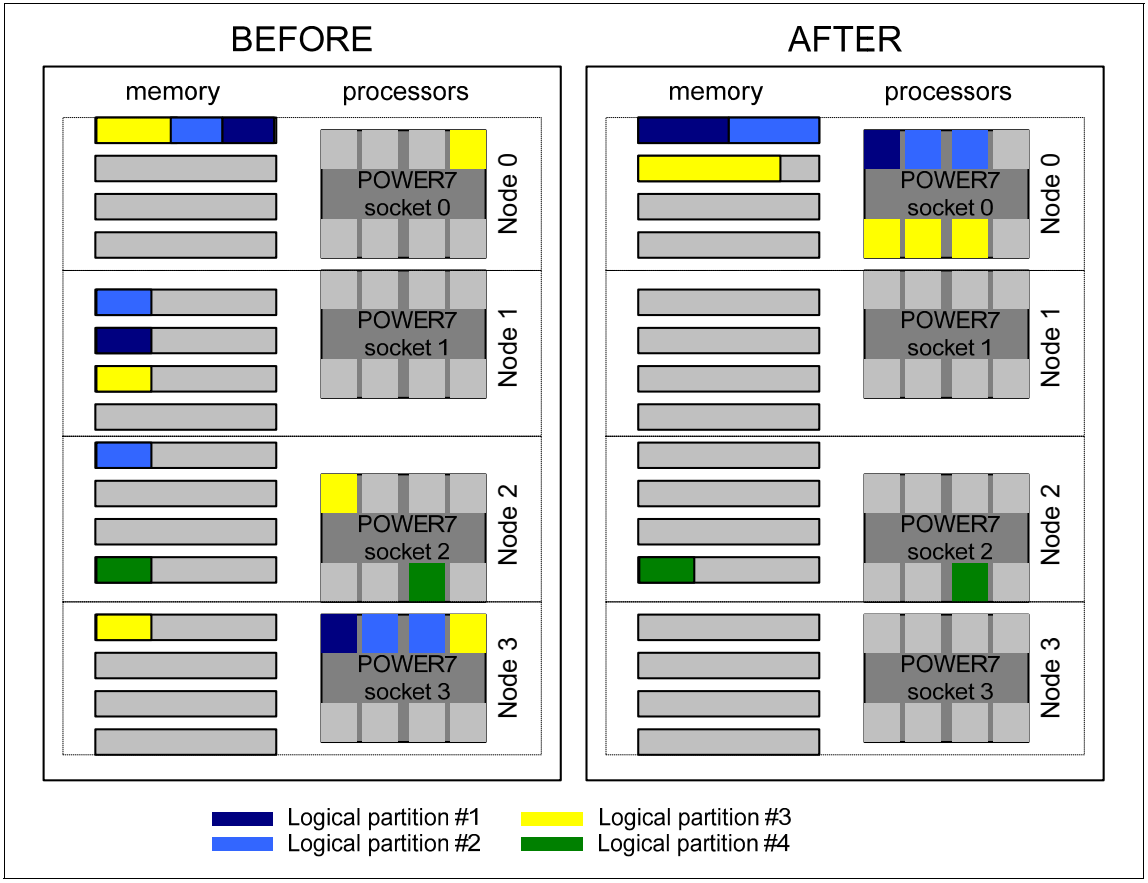


Figure 15-1 DPO in action

The following sections describe how the processor affinity can be optimized by running DPO.

15.2 Dynamic Platform Optimizer requirements

The Dynamic Platform Optimizer feature was introduced on Power Systems Firmware version 760, and requires HMC Version 7.6.0 or later. The following are required to run DPO:

- Power Systems Firmware level 760 or later
- HMC version 7.6.0 or later
- License code for DPO. Systems that ship with Power Systems Firmware version 760 or later have the Virtualization Engine Technology (VET) code installed during the manufacturing. If you upgrade an existing system to Power Systems Firmware version 760, you must acquire the VET code from IBM.
- At least one free Logical Memory Block (LMB) on the system

A DPO aware operating system is preferable, but not required. You can use older operating systems if you either reboot all the affected client logical partitions after you run DPO, or exclude them from participating in the DPO operation. You can exclude them by specifying those partitions in the optional protected partition list parameter of the **optmem** command.

The following are DPO-aware operating systems:

- AIX 6.1 TL8 or later
- AIX 7.1 TL2 or later
- IBM i 7.1 PTF MF56058
- VIOS 2.2.2.0 or later
- Red Hat Enterprise Linux 7
- SuSE Linux Enterprise Server 12

15.3 Managing Dynamic Platform Optimizer

This section describes how to initiate, halt, and display status of the optimizer y using the **lsmemopt** and **optmem** HMC commands.

DPO determines an optimal resource placement strategy for the server that is based on the partition configuration and hardware topology on the system. It performs a sequence of processor and memory relocations to transform the existing server layout into the optimal layout. This process occurs dynamically while the partitions are running.

15.3.1 Example of DPO interaction

Before you run the optimizer, check the current affinity state and what potential affinity you can achieve. The following sequence of HMC commands illustrates a typical interaction with DPO:

1. Confirm you have at least one LMB available using the HMC properties. The free LMB can either be licensed or unlicensed.
2. Retrieve the current affinity score:

   ```
   lsmemopt -m <system_name> -o currscore
   ```

3. Calculate the potential score that can be achieved:

   ```
   lsmemopt -m <system_name> -o calcscore
   ```

4. Run the affinity optimizer:

   ```
   optmem -m <system_name> -t affinity -o start
   ```

5. Check the optimizer status:

   ```
   lsmemopt -m <system_name>
   ```

6. Retrieve the current affinity score again after DPO completes:

   ```
   lsmemopt -m <system_name> -o currscore
   ```

All of these commands are explained in more detail later in this chapter. The HMC command-line interface also provides online help for each of these commands.

15.3.2 Requested and protected partition sets

When running DPO, you can optionally identify two special classes of partitions:

- Partitions that the DPO operation prioritizes for optimization, which are known as the *requested partitions*
- Partitions that the DPO operation does not modify, which are known as the *protected partitions*. The syntax of these options is documented in the help text for the **optmem** and **lsmemopt** commands.

The protected partition list, if specified, tells the DPO operation to avoid changing the processors or memory for the listed partitions. These partitions are not optimized, and their affinity is not changed by the DPO operation. This also ensures that the POWER Hypervisor does not steal any processor cycles from these partitions while optimizing. Because the resources of protected partitions are off-limits for the optimization, the quality of the optimization of the non-protected partitions can be impacted. The default setting for protected partitions is to protect no partitions.

The requested partition list, if specified, allows the user to explicitly enumerate a set of partitions to be prioritized for optimization. The DPO operation optimizes all non-protected partitions. By default, the hypervisor prioritizes partitions based on its internal algorithms for partition priority. The requested list allows the user to override this prioritization, with the requested partitions given higher priority than non-requested partitions. Within the requested set of partitions, and within the set of non-requested partitions, the hypervisor's internal prioritization still applies.

15.3.3 Partition operating system affinity

When the optimizer completes an optimization, it notifies the operating systems in the partitions that their physical memory and processors configurations have been changed. Partitions running on DPO-aware levels of their respective operating systems have support for this notification.

Restriction: For older operating system versions that do not support the notification, the dispatching, memory management, and tools that display the affinity will be incorrect. For some configurations, the performance of dispatching and memory management can actually degrade after you run the optimizer even though the partition has better affinity because the operating system is making decisions based on stale information.

For OS versions that are not DPO-aware, a reboot of the partition refreshes the affinity. Another option is to use the protected partition set option on the **`optmem`** command to not change the affinity of partitions with older operating system levels. However, doing so means the DPO operation has less scope to improve resource placement because the protected partition resources cannot take part in the optimization.

15.3.4 DPO performance considerations

To run the optimizer, there must be at least one free Logical Memory Block available on the system. For more information about Logical Memory Blocks, see Chapter 17 of *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.DPO can also use memory that is installed on the system that has not been activated (*unlicensed memory*). The busier the processors are on the system, the longer the optimization takes to complete. This delay is because the background optimizer tries to minimize its effect on running partitions. Partitions that are powered off can be optimized quickly because the contents of their memory do not need to be maintained.

While an optimization is running, the performance of the partitions is degraded. Overall, there is a higher demand placed on the processors because of

increased virtualization processor usage incurred by the hypervisor while the partition memory is being relocated. However, partitions that are explicitly protected from DPO are not affected. For more information, see 15.3.2, “Requested and protected partition sets” on page 603.

Important: Impacted partitions can have up to 20% of performance degradation while partition memory is relocated.

DPO spends most of its time moving partition memory. The actual amount of time that is spent depends on the processor workload, the amount of memory that must be moved, and on the system configuration.

Relocation of memory and processors is transparent to the logical partitions. It uses idle cycles that are periodically donated from the hypervisor dispatcher to run the optimization with minimal disruption to the active partitions on the system.

15.3.5 Estimating potential DPO affinity score

Before starting DPO, run the **lsmemopt** command with the **calcscore** option to estimate the potential affinity score after optimization (Example 15-1).

Example 15-1 Estimating the potential affinity score after optimization

```
hscroot@ftchmc5b:~> lsmemopt -m z2436ae -o calcscore
curr_sys_score=68,predicted_sys_score=94,"requested_lpar_names=none","r
equested_lpar_ids=none",protected_lpar_ids=none
```

15.3.6 Starting DPO

Use the HMC command-line interface to initiate the Dynamic Platform Optimizer by using the **optmem** command, as shown in Example 15-2.

Example 15-2 Initiating the Dynamic Platform Optimizer

```
hscroot@ftchmc5b:~> optmem -m z2436ae -t affinity -o start
```

This command initiates the optimization for all logical partitions on the entire server. The time the optimizer takes is dependant upon the placement of partitions, the overall amount of processor and memory that need to be moved, and the amount of processor cycles available.

Tip: You can exclude partitions from optimization (**optmem**) or potential affinity score calculations (**lsmemopt**) by using either the **-x** switch to exclude partitions by name, or the **--xid** switch to exclude partitions by partition ID.

15.3.7 Checking DPO status

Use the HMC command-line interface to check the status of the Dynamic Platform Optimizer through the **lsmemopt** command, as shown in Example 15-3.

Example 15-3 Checking the status of the Dynamic Platform Optimizer

```
hscroot@ftchmc5b:~> lsmemopt -m z2436ae
opt_id=22,in_progress=1,status=In
progress,type=affinity,total_mem=null,remaining_mem=null,elapsed_time=n
ull,progress=2,"requested_lpar_names=none","requested_lpar_ids=none",pr
otected_lpar_ids=none,impacted_lpar_ids=none
```

This command displays the status of the most recently requested optimization. If the optimization is in progress, it displays an estimate of the operation's progress as a percentage of completion, as shown in the example where progress is 2 percent completed.

Note: The estimated completion is a projection on the amount of resources that were moved and is not a estimating of completion time.

After the optimizer is complete, the **lsmemopt** command also reports the affected partitions, as shown in Example 15-4.

Example 15-4 Checking the status of the Dynamic Platform Optimizer completed

```
hscroot@ftchmc5b:~> lsmemopt -m z2436ae
opt_id=22,in_progress=0,status=Finished,type=affinity,total_mem=null,re
maining_mem=null,elapsed_time=null,progress=100,"requested_lpar_names=n
one","requested_lpar_ids=none",protected_lpar_ids=none,"impacted_lpar_n
ames=lpar3,lpar7,lpar8,lpar9,lpar10,lpar11,lpar12,lpar13","impacted_lpa
r_ids=3,7,8,9,10,11,12,13"
```

15.3.8 Stopping DPO

Use the HMC command-line interface to halt the Dynamic Platform Optimizer through the **optmem** command, as shown in Example 15-5.

Example 15-5 Stopping the Dynamic Platform Optimizer

```
hscroot@ftchmc5b:~> optmem -m z2436ae -o stop
```

This command stops an optimization that is in progress.

Note: Stopping a DPO operation before it completes all the movement of processors and memory can result in affinity being poor for some partitions that were not completed. For this reason, avoid using this command when possible.

15.3.9 Troubleshooting

HMC reports the error shown in Example 15-6 when the managed system is not DPO capable.

Example 15-6 The managed system does not support DPO error

```
hscroot@hmc8:~> lsmemopt -m p750
HSCL02D0 This operation is not allowed because the managed system does
not support Dynamic Platform Optimization.
```

The only reason that you will see this message is if you are running on FW760 and do not have the VET code installed. Acquire the VET code to make your system DPO-capable.

You can also check whether the managed system is DPO capable by looking at the system properties in the HMC GUI as shown in Figure 15-2.

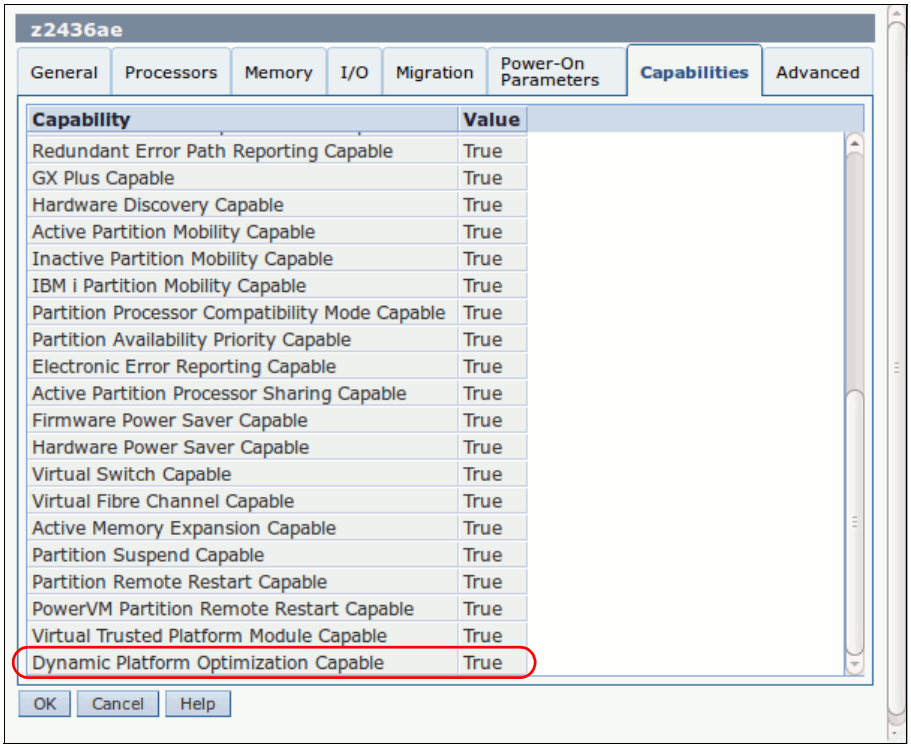


Figure 15-2 Checking if the managed system is DPO capable

15.4 Monitoring Dynamic Platform Optimizer

This section describes how to monitor DPO by using the HMC command-line interface and the `lsmemopt` command.

15.4.1 Computing the current affinity score

You can retrieve the current affinity score for the server before actually initiating the DPO. The score is a number in the range 0 - 100, where 0 means poor affinity and 100 means perfect affinity.

Note: The affinity score is based on hardware characteristics and partition configurations, so that a score of 100 might not be achievable.

To report the current affinity score for the server, issue the command **lsmemopt** with **-o currscore** argument, as shown in Example 15-7.

Example 15-7 Computing the current affinity score

```
hscroot@ftchmc5b:~> lsmemopt -m z2436ae -o currscore
curr_sys_score=68
```

Note: The affinity scores are only valid within the context of a specific server. Scores cannot be compared between different systems. The value of the scores is to enable comparison of the current affinity score with the potential affinity score, and with the current affinity score captured an earlier time.

15.4.2 Predicting an affinity score

The first thing when doing affinity optimization is to determine whether the server might benefit from running DPO. To do this, compare the current server affinity score with a score that is predicted to result if DPO were run. This gives a potential score that can be achieved by optimizing the system with the DPO. A calculated score of 100 might not be possible. The scoring is meant to provide a gauge to determine whether running the optimizer is likely to provide improved performance. For example, if the current score is 80 and the calculated score is 90, running the optimizer might not have a noticeable impact on the system performance. The amount of gain from doing an optimization is dependent on the applications that run within the various partitions.

To calculate the potential affinity score for the server, issue the command **lsmemopt** with **-o calcscore** argument, as shown in Example 15-8.

Example 15-8 Calculating the potential affinity score

```
hscroot@ftchmc5b:~> lsmemopt -m z2436ae -o calcscore
curr_sys_score=68,predicted_sys_score=94,"requested_lpar_names=none","r
equested_lpar_ids=none",protected_lpar_ids=none
```

16

Active System Optimizer and Dynamic System Optimizer for AIX

Active System Optimizer (ASO) is a new subsystem that is designed to automatically improve the performance of AIX workloads running on POWER7 Systems. ASO identifies and optimizes workloads and improves cache and memory affinity. Dynamic System Optimizer (DSO) is built on the ASO framework, and provides more optimizations. Improvements in partition level workload performance can be achieved because of the optimizations performed by the ASO subsystem.

IBM i includes this function as part of the operating system.

This chapter includes the following sections:

- Managing ASO/DSO
- Monitoring ASO/DSO

16.1 Managing ASO/DSO

This section addresses the prerequisites for using ASO and DSO, the types of optimization provided, and how to activate and manage the ASO subsystem.

16.1.1 ASO/DSO prerequisites

ASO/DSO optimizers have the following prerequisites:

- AIX V7.1 TL01 SP1 or later, or AIX V6.1 TL8 SP1 or later, for ASO
- AIX V7.1 TL02 SP1 or later, or AIX V6.1 TL8 SP1 or later, for DSO
- POWER7 or POWER7+ processor-based systems
- bos.aso file set for ASO
- dso.aso file set for DSO
- Partitions *cannot* be using Active Memory Sharing (AMS)
- Enhanced affinity must be enabled (default behavior for AIX V7.1)

If you attempt to run ASO on an unsupported configuration, ASO hibernates and no automatic tuning is attempted.

ASO is an AX feature that is available at no extra charge. DSO includes more charged-for optimizations through an enablement file set.

16.1.2 The ASO subsystem

ASO runs as a subsystem under the control of the System Resource Controller (SRC) on AIX. DSO features are built on top of ASO, and use the same subsystem.

Use the commands in Example 16-1 to list the status of the ASO subsystem, and if necessary start the ASO subsystem. Also, it is necessary to set the ASO tunable `aso_active` to ensure that ASO activation is permanent.

Example 16-1 Listing the current status of the ASO subsystem

```
# lssrc -s aso
Subsystem         Group            PID          Status
 aso                                            inoperative
# startsrc -s aso
0513-059 The aso Subsystem has been started. Subsystem PID is 3211410.
# lssrc -s aso
Subsystem         Group            PID          Status
 aso                               3211410      active
# asoo -a
```

```
aso_active = 0
# asoo -p -o aso_active=1
Setting aso_active to 1 in nextboot file
Setting aso_active to 1
```

16.1.3 Types of ASO optimization

ASO optimization has the Cache Affinity, Aggressive Cache Affinity, and Memory Affinity types available.

Cache Affinity

The aim of Cache Affinity optimization is to bind the threads from eligible workloads to the smallest processor affinity domain, while still meeting the workload's demands for processor and memory resources. Figure 16-1 shows two 8-core POWER7 processors (sockets). ASO has moved the running threads, which were previously communicating across sockets, to the same number of cores within the same socket. By localizing the processor resource set (RSet) for a process, ASO can minimize the amount of data (for example, L3 Cache) crossing affinity domains, resulting in a workload performance improvement.

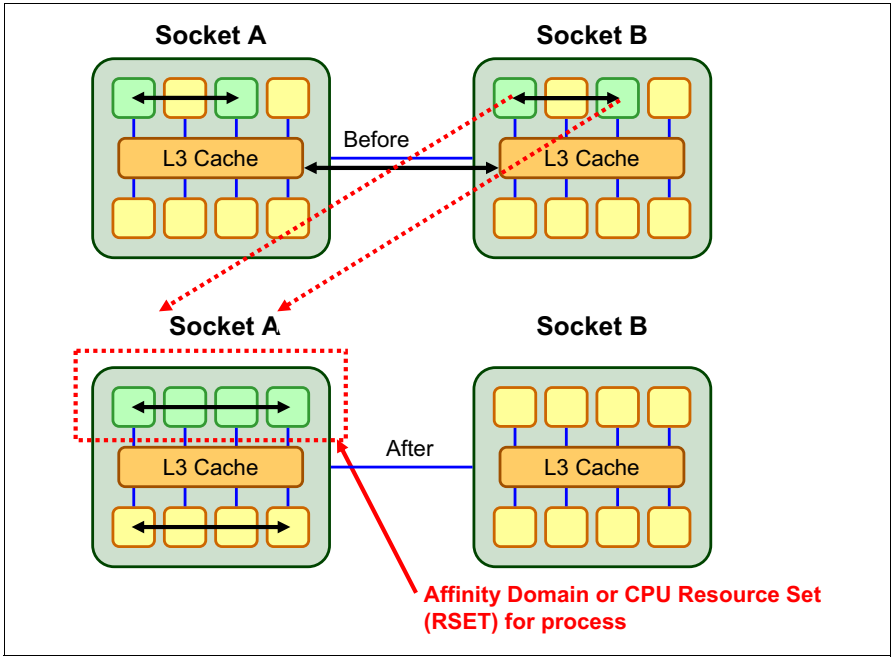


Figure 16-1 Cache Affinity

Aggressive Cache Affinity

If ASO performs Aggressive Cache Affinity, it consolidates workloads onto fewer cores to achieve greater cache affinity. Figure 16-2 shows that originally a workload was running on 10 cores across two POWER7 sockets. ASO consolidates the workload to run on eight cores on a single 8-core socket to provide a smaller affinity domain. ASO performs aggressive cache affinity only if it has sufficient evidence from its pre and post monitoring processes that performance improvements can, and have, been realized by consolidating the workload.

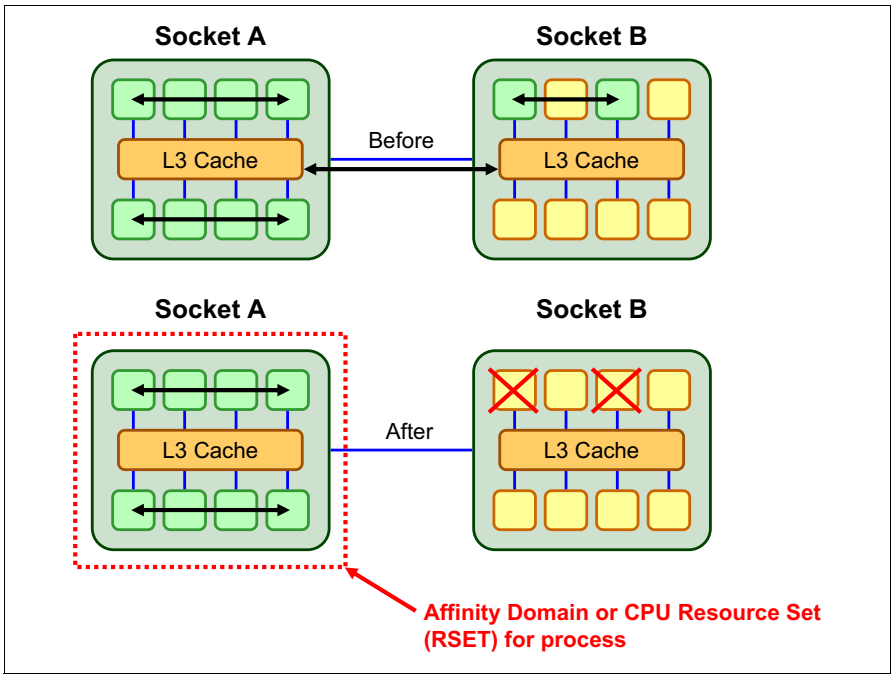


Figure 16-2 Aggressive Cache Affinity

Memory Affinity

When ASO implements Memory Affinity, it migrate frequently accessed memory pages that are attached to remote sockets to memory pages attached to the sockets where the process is running. For the scope of the affinity domain to be established, memory affinity can be applied only after a workload is optimized for cache affinity.

Multi-threaded workloads with 5+ minute periods of stability are best suited for cache and memory affinity. Workloads must fit within a single Scheduler Resource Affinity Domain (SRAD). An SRAD typically caps to a single POWER7 processor/socket.

With cache and memory affinity working together, the workloads and memory pages are relocated to a *Processor and Memory Resource Set* contained on a single socket and memory bank. This configuration reduces latency, and increases overall workload performance as shown in Figure 16-3.

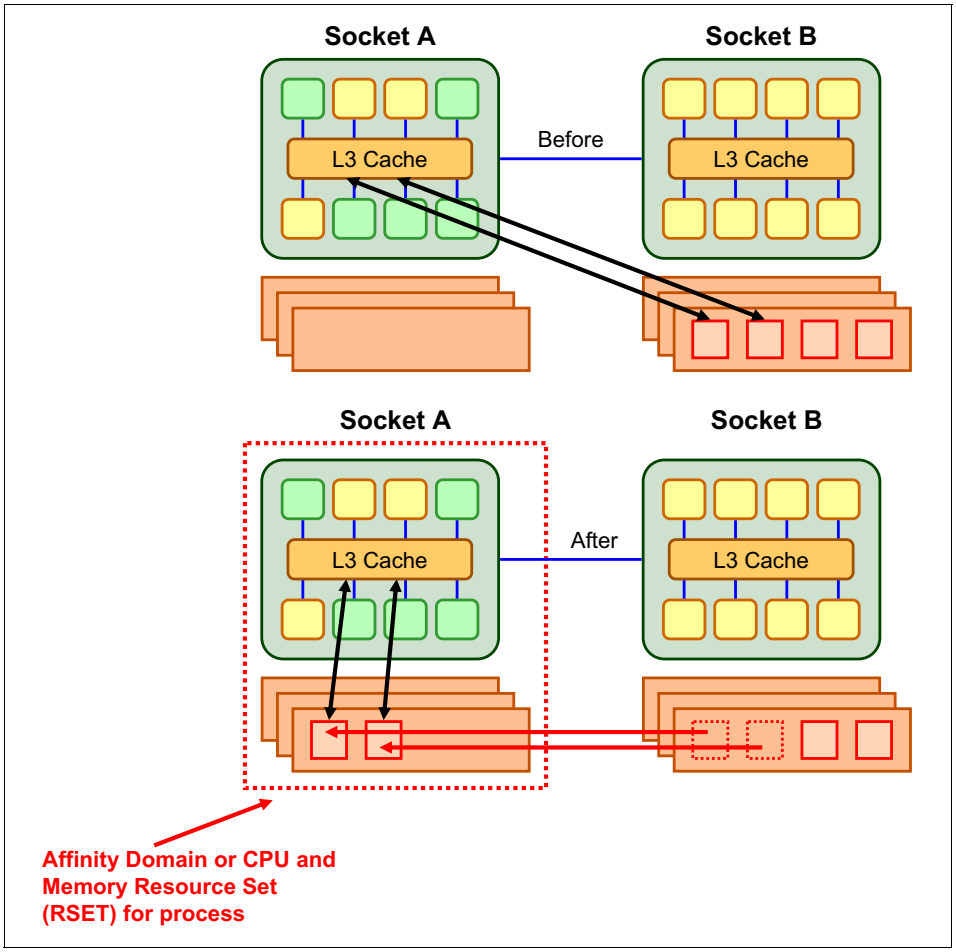


Figure 16-3 Memory Affinity

16.1.4 Types of DSO optimization

DSO optimization has the Multiple Page Segment Size (MPSS) and Memory pre-fetch types available.

Multiple Page Segment Size (MPSS)

With AIX V6.1 TL8 and AIX v7.1 TL2, 16 MB memory pages can be alongside 4 K and 64 K pages within a single 256 MB memory segment. Therefore, DSO is able to automatically convert smaller pages into larger 16 MB pages.

This is of particular benefit to workloads that use large memory regions (database buffers or JVMs using large heaps, by reducing the need to manage large numbers of smaller pages, and I/O operations. Workloads with a minimum of 16 GB of system memory are considered potential MPSS targets.

Previously, 16 MB page sizes were supported on AIX, but had to be manually tuned and pre-allocated (pinned) in memory. This limitation resulted in difficult sizing of large memory pages and possible waste of memory resources.

Memory pre-fetch

DSO collects information from the AIX kernel, Power Hypervisor performance utilities, and Process Counters to dynamically determine the optimal setting of the Data Stream Control Register (DSCR). The DSCR controls memory pre-fetching on POWER7 processor-based systems. Pre-fetch depth and stride values are adjusted in an attempt to avoid L3 cache misses. For example, a high depth value is desirable for sequential access to data and a low depth for random access patterns. If the DSCR is manually tuned using the **dscrctl** command, this optimization is disabled.

16.1.5 ASO configuration

ASO requires little user configuration when it is running. However, if necessary, you can control ASO/DSO at the process level by setting ASO shell variables. Set these before you run important processes to include or exclude the process from the different types of ASO optimization. To make these settings permanent and system wide, enter the following line into the `/etc/environment` file:

```
ASO_ENABLED=[ALWAYS|NEVER]
```

When this shell variable is set for a process, ASO skips some eligibility checks and prioritizes this process for optimization. Conversely, you can exclude the process from ASO optimization.

In addition to the ASO_ENABLED variable, one or more ASO_OPTIONS variables can be specified. When multiple options conflict, only the last defined setting takes effect.

```
ASO_OPTIONS=ALL=[ON|OFF]
```

This shell variable either enables ALL the affinity optimizations for a process, or, if set to OFF, allows individual optimizations to be specified:

```
ASO_OPTIONS=CACHE_AFFINITY=[ON|OFF]
```

Enables or disables cache affinity for a process:

```
ASO_OPTIONS=MEMORY_AFFINITY=[ON|OFF]
```

Enables or disables memory affinity for a process:

```
ASO_OPTIONS=LARGE_PAGE=[ON|OFF]
```

Enables or disables 16 MB Multiple Page Segment Size (MPSS). This sell variable is only available with DSO.

```
ASO_OPTIONS=MEMORY_PREFETCH[ON|OFF]
```

Enables or disables memory pre-fetch optimization. Only available with DSO.

On the example system, disable memory affinity but enable cache affinity for your shell and sub shells. Use the command that is shown in Example 16-2.

Example 16-2 Exporting ASO and DSO shell variables

```
# export ASO_OPTIONS=ALL=OFF,CACHE_AFFINITY=ON
```

16.2 Monitoring ASO/DSO

ASO/DSO activities are logged in the `/var/log/aso` directory on AIX as shown in Example 16-3.

Example 16-3 ASO/DSO log files

```
# cd /var/log/aso
# ls -o
total 192
-rw-r--r--    1 root           6869 Dec 19 10:50 aso.log
-rw-r--r--    1 root          88665 Dec 19 10:50 aso_process.log
```

The `aso.log` file provides you with information about the overall state of the ASO subsystem. The `aso_process.log` file provides information about optimizations that are performed on the process. Optimization activities are logged against the PID (all process threads are treated the same), so you can easily determine which optimizations have been attempted on your system.

An extract from the aso.log file is shown in Example 16-4. You can see that during periods of low activity, ASO hibernates. After the workloads reach minimum entitlement levels for the partition, ASO resumes.

Example 16-4 The aso.log file

```
Dec 18 16:08:11 p750_lpar01 aso:notice aso[3539088]: [HIB] Used
entitlement per unfolded vCPU is below threshold (1% of a cor
e).
Dec 18 16:08:11 p750_lpar01 aso:notice aso[3539088]: [HIB] ASO will
hibernate until used entitlement is at least 30% of a cor
e per unfolded vCPU
Dec 18 16:26:36 p750_lpar01 aso:notice aso[3539088]: [HIB] Resuming
from hibernation.
```

An example of the aso.process.log file is shown in Example 16-5. It is possible to monitor the behaviors of ASO and DSO optimizations by tracing your PID through this file.

Note: The date and time stamps have been removed, and line numbers added for ease of reading

In lines 1 and 2, you can see that two processes on this partition are being considered for optimization by ASO (PIDs 7012560 and 6226514). ASO monitors the current utilization and load for each PID before it takes any action. In lines 5 and 6, ASO moves PID 7012560 to 4 cores, on socket 0. It then checks utilization levels again for the PID in line 9. You can see that the utilization of this PID drops from 2.23 to 1.85 after it is relocated (compare line 1 and line 9).

Meanwhile, PID 6226514 is being monitored and ASO makes some predictions about the possible performance gain (lines 13 and 14). Because there is no performance advantage in its prediction, no relocation of this PID actually occurs.

Example 16-5 The aso.process.log file

```
1. [SC][7012560] Considering for optimisation (cmd='paraworms',
utilisation=2.23, attaching StabilityMonitorBasic)
2. [SC][6226514] Considering for optimisation (cmd='paraworms',
utilisation=1.17, attaching StabilityMonitorBasic)
3. [perf_info] system utilisation 4.71; total process load 9.96
4. attached(   7012560): cores=4, firstCpu=  0, srads={0}
5. [WP][7012560] Placing non-FP (norm load 3.20) on 4.00 node
6. [EF][sys_action][7012560] Attaching (load 3.20) to domain SRAD
(cores=4,firstCpu=0)
7. [perf_info] system utilisation 5.24; total process load 9.96
```

```
8. [perf_info] system utilisation 4.91; total process load 9.93
9. [SC][7012560] Considering for optimisation (cmd='paraworms',
utilisation=1.85, attaching StabilityMonitorAdvanced)
10. [EF][7012560] attaching strategy StabilityMonitorAdvanced
11. [perf_info] system utilisation 4.61; total process load 9.96
12. [EXP] Allowing domain System
13. [PRED][6226514] SRAD (4): -Cross: 0.00 +Compr: 0.00 Gain: 0.00  --
SCORE: 1.00
14. [PRED][6226514] Book (4): -Cross: 0.00 +Compr: 0.00 Gain: 0.00  --
SCORE: 1.00
15. [PRED][6226514] Recommending max domain SRAD of minimum size 4
16. [EXP][6226514] Predictor recommends trying SRAD (4)
17. [EXP] Allowing domain Book (4)
18. [EXP] Allowing domain SRAD (4)
19. attached(   1835312): [free]
```

Part 6

Enterprise management tools

This part describes some of the IBM management tools for a PowerVM virtualized environment. It includes IBM Systems Director suite of products, and IBM Tivoli products. These tools provide integrated and efficient management and monitoring capabilities to Power Systems and PowerVM.

This part includes the following chapters:

- IBM Systems Director
- Tivoli Systems Management integration

17

IBM Systems Director

IBM Systems Director is an integrated suite of tools that streamlines the managing and monitoring of resources across your PowerVM environment. IBM Systems Director can manage both simple and complex heterogeneous environments, and can scale to 5000 managed resources. It supports these technologies:

- HMC, IVM, and Virtual I/O Servers
- POWER Servers and POWER blades
- AIX, IBM i, and Linux operating systems

In addition to PowerVM resources, these virtualization technologies can also be managed from the IBM Systems Director management server:

- KVM
- VMware console and VMware guest operating systems
- Microsoft Hyper-V

This chapter is a quick-start guide to implementing a base IBM Systems Director configuration. For more information about developing base IBM Systems Director infrastructure, see the website at:

http://pic.dhe.ibm.com/infocenter/director/pubs/index.jsp?topic=%2Fcom

This chapter includes these sections:

- Managing IBM Systems Director
- Monitoring IBM Systems Director

17.1 Managing IBM Systems Director

This chapter highlights common IBM Systems Director tasks. The following topics are addressed with examples for different operating systems including Virtual I/O Server, AIX, IBM i, and Linux:

- IBM Systems Director installation
- Discovering, navigating, and visualizing resources in the network
- Collecting detailed inventory of both managed systems and logical partitions
- Setting up repositories for operating system fixes and distributing them to individual virtual servers or groups of virtual servers

17.1.1 IBM Systems Director installation overview

The IBM Systems Director management server installation includes the IBM Systems Director Server, DB2® RDBMS, and the Common Agent Manager. It can be installed on an AIX, Linux, or Windows operating system. The endpoint systems run the Systems Director *Common Agent* or *Platform Agent* (see “Installing the Common and Platform Agents” on page 627). The software for both the IBM Systems Director Server and the Agents software can be downloaded from the Systems Director download site (login credentials required) at:

`http://www-03.ibm.com/systems/software/director/downloads`

The example IBM Systems Director management server used in this chapter is composed of these components:

- AIX 7.1 TL1 SP5
- IBM Systems Director Version 6.3.1.1
- Embedded DB2 RDBMS

When you install the IBM Systems Director management server on an AIX logical partition, generally use the following minimum settings:

- Processor:
  - 2.0 processing units.
  - Three virtual processors.
  - Active memory expansion enabled. As a Java based product, good compression ratios can be expected. Run **`amepat`** to determine acceptable memory compression ratios after IBM Systems Director is up and running.

- Memory:
  - 6 GB.
  - 6 GB paging space.
- Storage:
  - / 1 GB.
  - /usr 4 GB.
  - /var 4 GB.
  - /tmp 4 GB.
  - /home 16 GB (default installation directory for DB2 RDMS).
  - /opt 64 GB (IBM Systems Director software and OS updates).
- Operating system:
  - Install the latest technology levels and service packs.
  - Install ssh.base and dsm.dsh.
  - Set the root fsize ulimit to -1 in the `/etc/security/limits` file.
  - Configure NTP to ensure consistent timing between the IBM Systems Director management server and agents.

17.1.2 Installing IBM Systems Director

To install the IBM System Director, you must download and install the program, and then download and install Common and Platform Agents.

IBM Systems Director Server

The software for the IBM Systems Director Server can be downloaded from the Systems Director download site (login credentials required) at:

`http://www-03.ibm.com/systems/software/director/downloads`

The IBM Systems Director Server software is contained in a `tar.gz` file. This file contains the IBM Systems Director Server installation script, `dirinstall.server`. When you run this script, you have the option of using an embedded or remote DB2 RDBMS. The embedded DB2 RDBMS provided with the product is sufficient for most environments and simplifies the installation of the product. Detailed installation instructions for all IBM Systems Director components can be obtained from the IBM Systems Director Information Center at:

`http://pic.dhe.ibm.com/infocenter/director/pubs/index.jsp?topic=%2Fcom.`

Tip: After installation, you might see a warning message that is posted to the IBM Systems Director Server console every 10 seconds. This is a known issue. Use the script available at the following URL to prevent this:

http://www-01.ibm.com/support/docview.wss?uid=nas77d16faf83372b0b386257997006f12b1

Generally, sign on to the IBM Systems Director server with dedicated user IDs and passwords. The users must be assigned to the `smadmin` group, as shown in Example 17-1.

Example 17-1 Creating a dedicated IBM Systems Director Server user

```
p740_lpar03:/ # mkuser fred
p740_lpar03:/ # passwd fred
Changing password for "fred"
fred's New password:
Enter the new password again:
p740_lpar03:/ # chuser groups=smadmin fred
p740_lpar03:/ # pwdadm -c fred
```

After installation, you can sign on to the IBM Systems Director web interface by using the dedicated user ID you created. You can access the IBM Systems Director GUI through your web browser at:

`https://HOSTNAME:8422/ibm/console`

The IBM Systems Director home window is displayed after you sign on to the IBM Systems Director Server as shown in Figure 17-1.

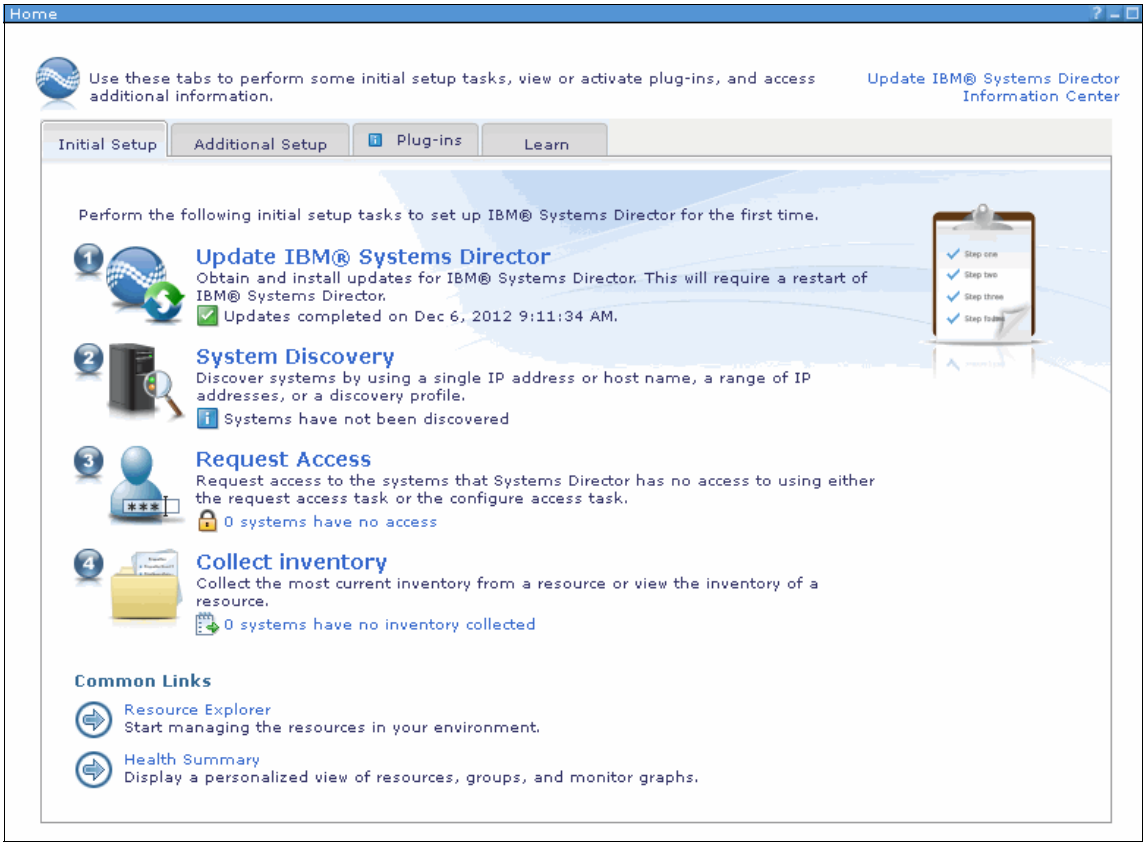


Figure 17-1 The IBM Systems Director home window

Installing the Common and Platform Agents

The Common Agent provides the full set of security and deployment functions, including discovery, monitoring health, and managing alerts. The Platform Agent provides a subset of the Common Agent functions for environments that require a smaller footprint. A full list of features that are supported by each type of agent can be found at:

`http://www-03.ibm.com/systems/software/director/downloads/agents.html`

The IBM Systems Director Common Agent is installed by default in AIX Version 5.3 TL10 (or later) and VIOS Version 2.1.1 (or later). The Common Agent software for IBM i V6R1 (or later) and Linux must be downloaded from this website.

After your virtual servers have the IBM Systems Director Common Agent or the IBM Systems Director Platform Agent installed and running, they can be managed by IBM Systems Director.

17.1.3 Updating IBM Systems Director Server

The IBM Systems Director Server can be updated by directly referencing files on the IBM Systems Director Support site. It can also be updated by referencing a local or NFS file system that is mounted on your IBM Systems Director Server.

To update the IBM Systems Director Server, complete these steps:

1. From the left navigation pane, select **Release Management** → **Updates** → **Update IBM Systems Director**. IBM Systems Director downloads the required updates in a compressed file from the IBM Systems Director support site.

   If your Systems Director Server does not have an external internet connection, click the **Stop** and provide the directory or path to the compressed file that you manually downloaded and stored on the IBM Systems Director Server.

2. IBM Systems Director extracts the maintenance compressed file, imports the fixes to the IBM Systems Director repository, and applies the fixes to the IBM Systems Director Server. You can monitor the status of the update job and the job logs in the Active and Scheduled Jobs pane.

3. After successfully updating the IBM Systems Director Server you must recycle the IBM Systems Director Management Server services by using the commands that are shown in Example 17-2.

Example 17-2 Recycling the IBM Systems Director Server

```
# cd /opt/ibm/director/bin
# ./smstop
Shutting down IBM Director...
# ./smstatus -r
Inactive
# ./smstart
Starting IBM Director...
The starting process may take a while. Please use smstatus to check
if the server is active.
# ./smstatus -r
Starting
Active#
```

4. After this is completed, you can view the applied level of code for the IBM Systems Director Server on the IBM Systems Director Welcome window as shown in Figure 17-2.

Figure 17-2 Welcome window showing the installed version of IBM Systems Director

Recycling Common Agents

You might have to recycle the Common Agents after you upgrade the IBM Systems Director management server. The commands shown in Example 17-3 can be used to recycle Common Agents on VIO Server.

Example 17-3 Recycling the VIO Server common agent

```
$ stopsvc director_agent
Stopping Director Common Agent...
Stopping dirsnmpd...
Stopping tier1slp...
Stopping cimlistener...
$ startsvc director_agent
Starting cimserver...
$
```

Use the commands shown in Example 17-4 to recycle Common Agents on AIX.

Example 17-4 Recycling the AIX common agent

```
# stopsrc -s cas_agent
0513-044 The cas_agent Subsystem was requested to stop.
# startsrc -s cas_agent
```

```
0513-059 The cas_agent Subsystem has been started. Subsystem PID is
14155912.
#
```

Similarly, use the commands in Example 17-5 to recycle agents on IBM i.

Example 17-5 Recycling the IBM i common agent

```
ENDTCPSVR SERVER(*HTTP) HTTPSVR(CAS)
STRTCPSVR SERVER(*HTTP) HTTPSVR(CAS)
```

Recycle agents on Linux using the commands in Example 17-6.

Example 17-6 Recycling the Linux common agent

```
# pwd
/opt/ibm/director/agent/runtime/agent/bin
# ./endpoint.sh stop
Stopping The LWI Nonstop Profile...
Stopped The LWI Nonstop Profile.
# ./endpoint.sh start
Starting The LWI Nonstop Profile...
The LWI Nonstop Profile succesfully started. Please refer to logs to
check the LWI status.
```

17.1.4 System Discovery

The IBM Systems Director Server uses System Discovery to establish a connection to manageable resources, such as HMCs, Virtual I/O Servers, and logical partitions, in your environment.

You can use System Discovery to manage a resource and view its hardware and software inventory. After the resource is discovered, access must be granted to bring the resource under IBM Systems Director management.

The System Discovery process can be started by selecting **Inventory** → **System Discovery** from the left navigation pane. You can then use the System Discovery window to search for resources in various ways by using the **Select a discovery option** menu. This example shows how to find a single resource using a known IP address, but it is possible to search using ranges of IP addresses.

System Discovery populates the **Discovered Manageable Systems** table when it connects to the Common or Platform Agent of the targeted resources, as shown in Figure 17-3.

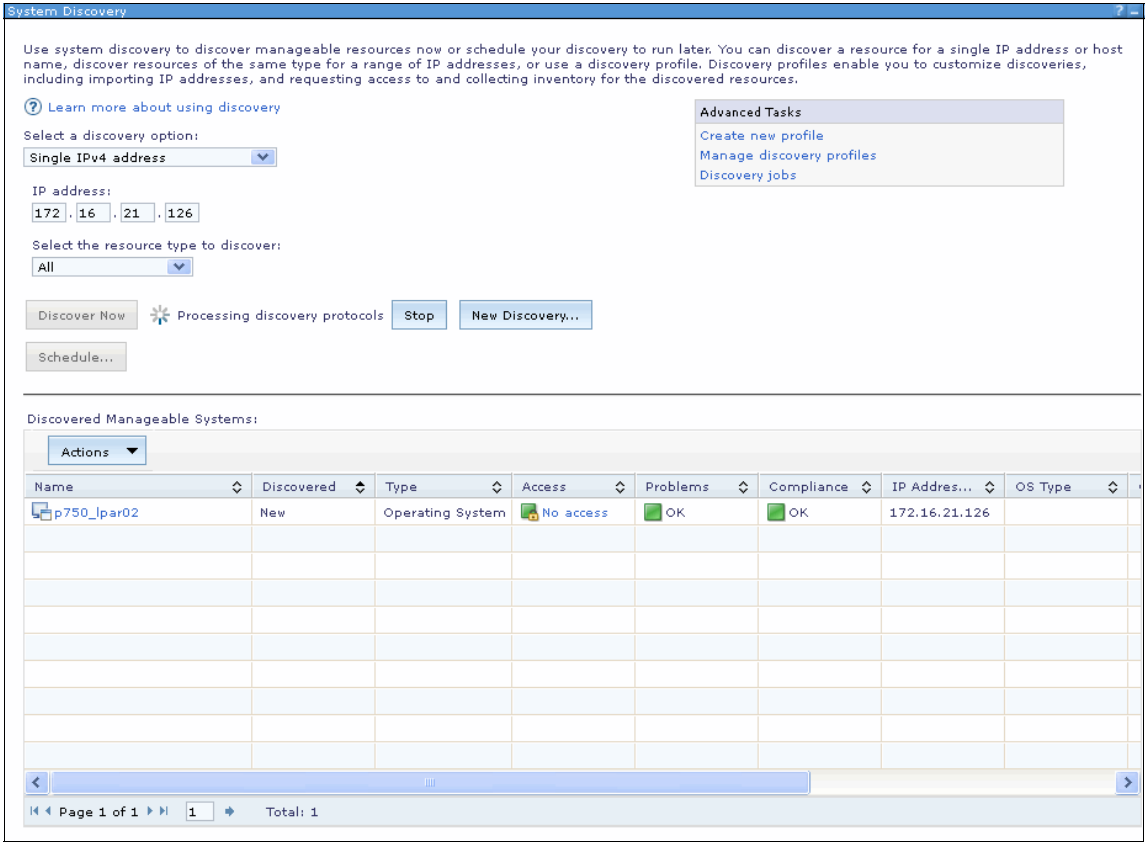


Figure 17-3 System Discovery by single IP address

The results of the search shows the **Access** field in the **Discovered Manageable Systems** table has a value `No access` for the **p750_lpar02** resource. You must authorize the IBM Systems Director management server to

manage this resource by clicking the **No access** link and providing the superuser account and password for **p750_lpar02** as shown in Figure 17-4.

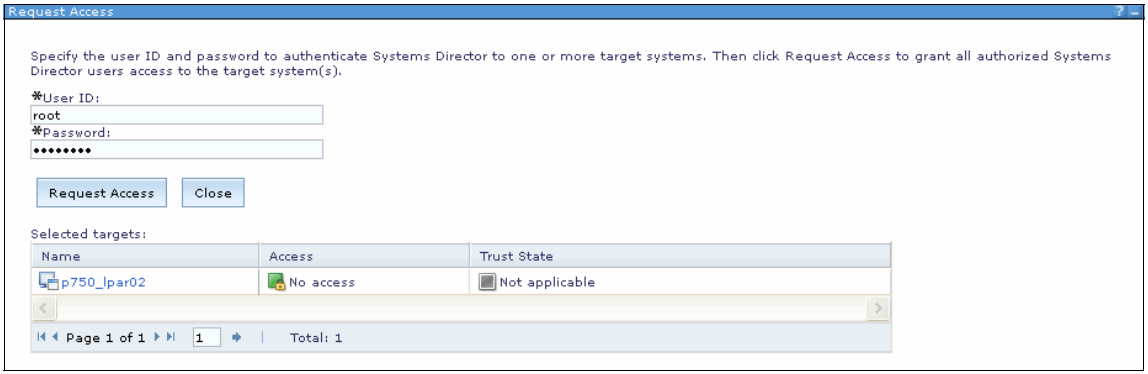


Figure 17-4 Providing authorization information for a discovered system

IBM Systems Director cannot manage discovered resources until the superuser credentials of the resource's operating system are entered in the management server. After authorization, the newly discovered system is displayed in the Resource Explorer discovered operating system group (Figure 17-5).

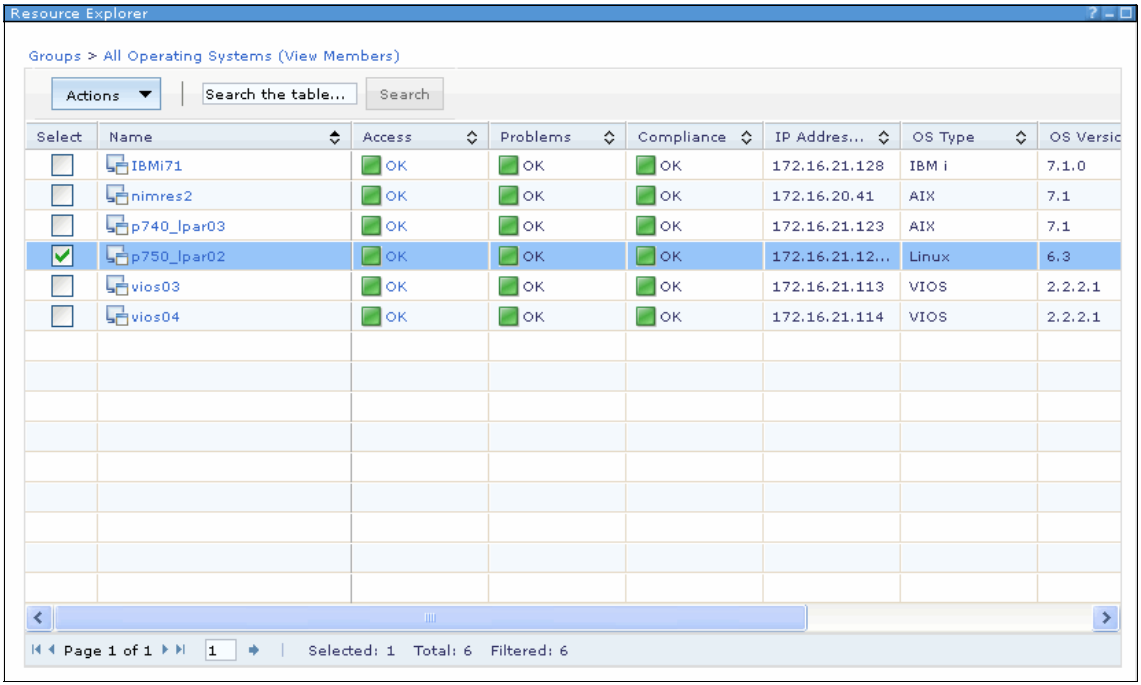


Figure 17-5 Resource Explorer: Discovered operating system group

The discovered resource is a Linux virtual server running v6.3 that is an RHEL distribution. Now that it has been brought the resource under IBM Systems Director management, you can perform a full hardware and software inventory. For more information, see 17.1.5, “Inventory collection” on page 633.

17.1.5 Inventory collection

Inventory collection establishes a connection to discovered resources and collects data about their installed hardware and software. Generally, discover resources in the following order:

1. HMCs
2. Virtual I/O Servers
3. Logical partitions

A full system inventory on a discovered resource can be started by selecting **Actions** → **Inventory** → **View and Collect Inventory** from the Resource Explorer window. This creates an inventory collection job that can be run immediately or on a schedule. Run inventory collection jobs during low utilization periods for the resource so that you do not interfere with production workloads. This is especially important in large IBM Systems Director configurations.

After the inventory collection job is submitted, you can monitor its progress on the Active and Scheduled Jobs window as shown in Figure 17-6.

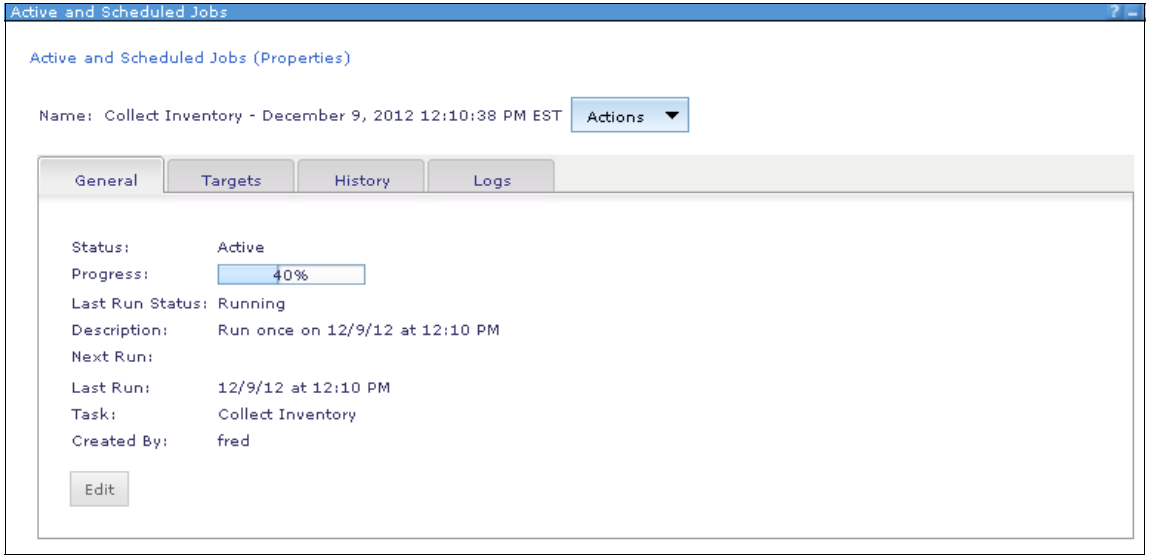


Figure 17-6 Monitoring an active job

After completion of the inventory collection job, you can view the collected information by selecting I**nventory** → **Collect and View Inventory** from the left navigation pane.

For a summary of the collected information for your system, select your target system and clicking **Refresh** as shown in Figure 17-7. You can also select **View Report** for the full, printable version of the report from the same window.

The inventory includes physical, logical, and virtual hardware; software, applications, operating systems, middleware, firmware, BIOS, and diagnostic information; network information; and system contained resources.

Inventory reports can also be exported to a comma-separated value (CSV), HTML, or XML file.

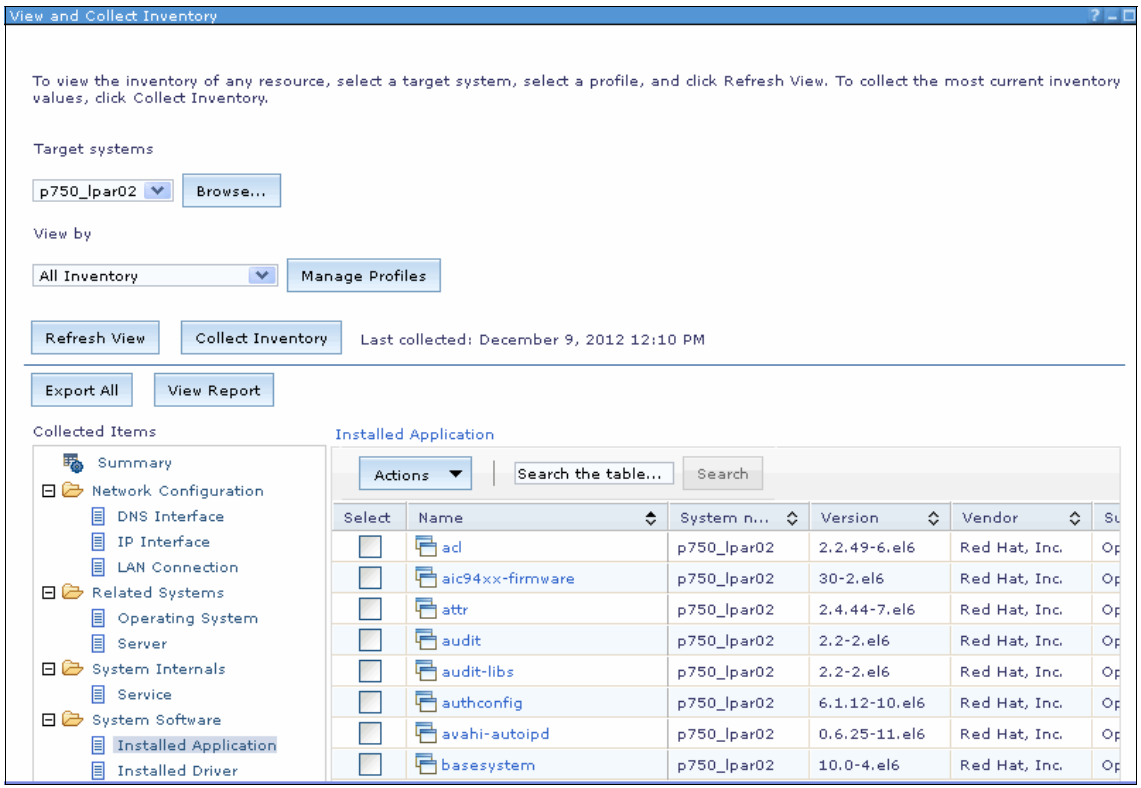


Figure 17-7 Inventory summary

17.1.6 Acquiring updates

IBM Systems Director can acquire, install, and manage updates for the systems that comprise your PowerVM environment. **Update Manager** helps keep your systems at the wanted level of software or firmware by comparing updates that are loaded into its repository against known inventories of specified systems to determine whether updates are required.

Obtaining updates

If an Internet connection is available to the IBM Systems Director Server, the acquire updates task automatically contacts the IBM fix repository and downloads the latest updates into **Update Manager**.

If no Internet connection is available, you can manually download supported updates and then import them into the update manager repository from a local or NFS mounted file system. Fixes can be downloaded from the IBM Fix Central at:

`http://www-933.ibm.com/support/fixcentral/`

Note: Update Manager cannot process updates without update descriptor files (`*.sdd`). If you use Fix Central to download the fix packs, ensure that you select the option to download the `.sdd` files for import into IBM Systems Director. It is currently not possible to import updates from physical media without `.sdd` files.

Configuring Update Manager

To configure Update Manager, complete these steps:

1. Select **Release Management** → **Updates** from the left navigation pane to access Update Manager.
2. Select **Configure Settings**.
3. Enter your Internet connection details on the Connection tab and use the **Test Connection** process to verify connection back to IBM.
4. Enter the path to the fix and update repository on the IBM Systems Director Server in the Location tab.

   This is where IBM Systems Director stores the fixes and updates, whether downloaded directly from the Internet or manually imported into the IBM Systems Director repositories. The default location for storing updates on an AIX-based IBM Systems Director Server is `/opt/ibm/director/data/updateslib`.

   Ensure that you have enough file system space to store your fixes and updates in the IBM Systems Director repositories. A single AIX Technology

level can be in excess of 4 GBs and a VIO fix pack can be 3 GB-4 GB. Size the file system appropriately.

A NIM server is required for AIX Technology Level updates and Virtual I/O Server upgrades (not updates). Define your discovered NIM Server in Update Manager on the AIX tab and VIO tab. These can be different NIM Servers if required.

Manually importing updates

After you download the fixes from IBM Fix Central with the relevant `.sdd` files, you can import the packages into IBM Systems Director with the following steps:

1. Select **Release Management** → **Updates** from the left navigation pane to access **Update Manager**.
2. Select **Acquire Updates** → **Import updates from the filesystem**.
3. Provide the local or NFS mounted path to the files.
4. Click **OK** to submit the import job.

   The import job copies the updates from the file system path you specified to the permanent IBM Systems Director update repository defined in “Configuring Update Manager” on page 635.

The progress of the import job can be monitored from the Active and Scheduled jobs window. Figure 17-8 shows an IBM i Group PTF Package successfully imported into IBM Systems Director's software repository. The same process can be followed for importing AIX, Virtual I/O Server, HMC, and Linux fix packs, as well as firmware updates, into your IBM Systems Director repository.

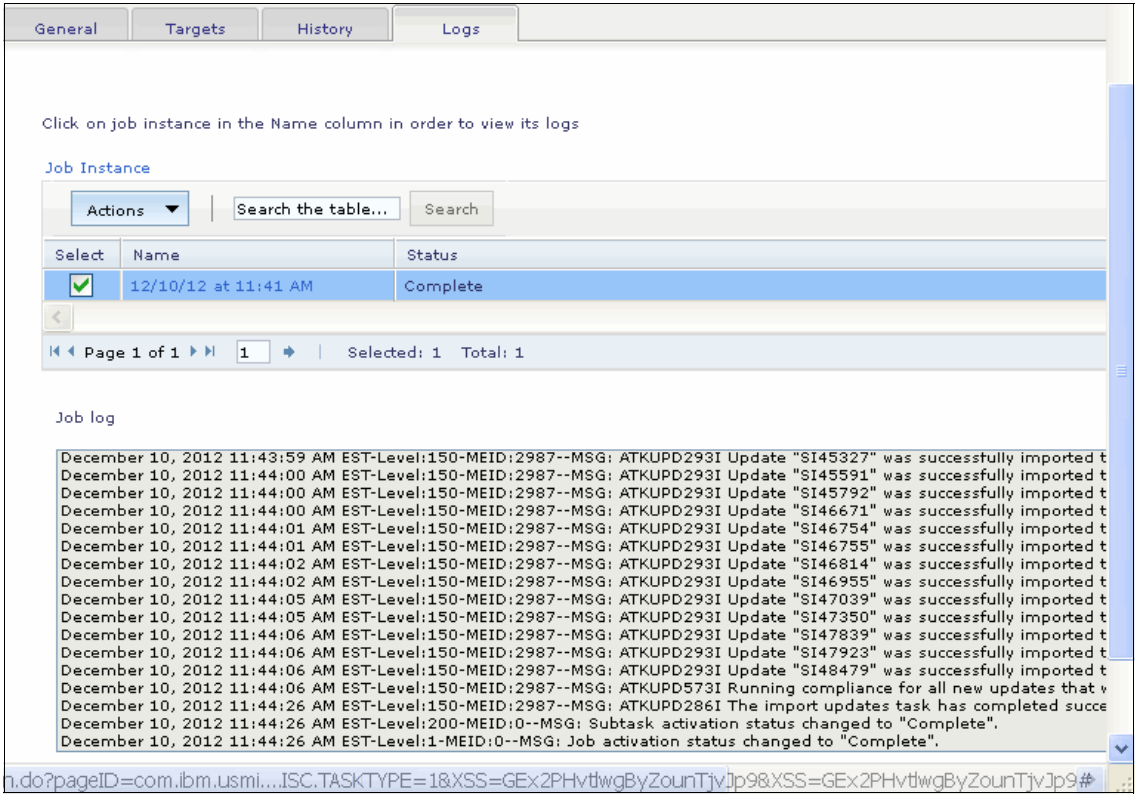


Figure 17-8 Monitoring an import job

17.1.7 Installing updates

To install the downloaded updates, complete these steps:

1. Select **Release Management** → **Updates** from the left navigation pane to access Update Manager.
2. Select **Show and Install Updates** and select the systems that you want to update.
3. Click **Show and Install Updates** to start the repository comparison.

IBM Systems Director compares the acquired fixes within its repository against the last known inventory of the selected virtual servers. IBM Systems Director shows a list of available fixes based on the operating system of each selected virtual server.

In Figure 17-9, IBM Systems Director is suggesting the installation of Virtual I/O Server Fixpack 26 for the selected Virtual I/O Server.

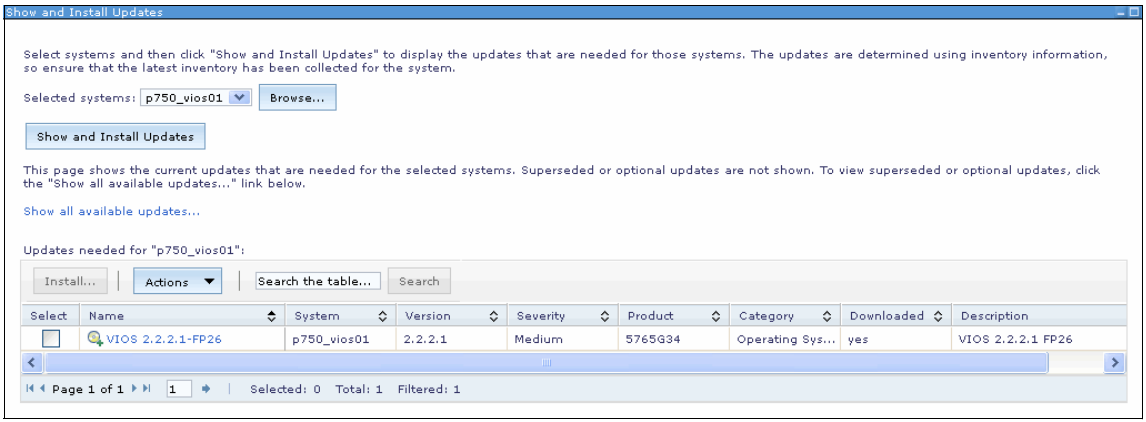


Figure 17-9 Suggested fixes for a VIOS resource

IBM Systems Director is suggesting the installation of IBM i Group PTF package SF99362 for the IBM i resource selected in Figure 17-10.

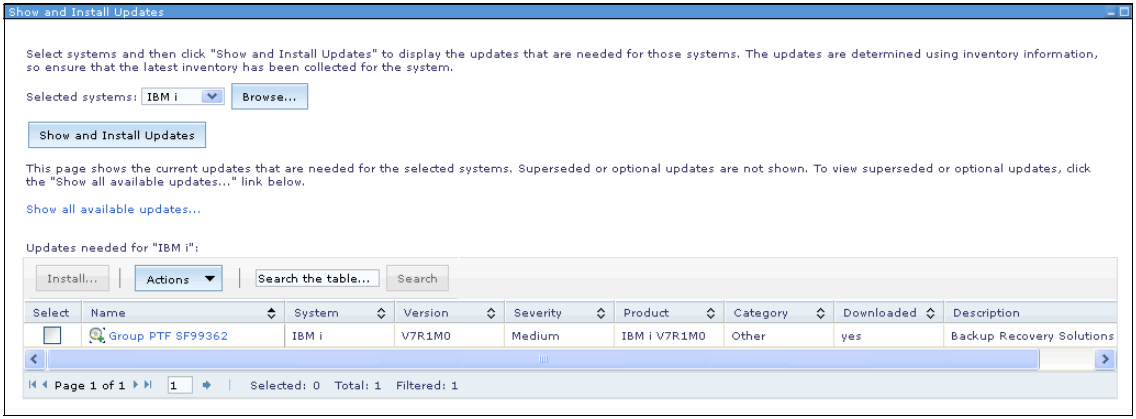


Figure 17-10 Suggested fixes for an IBM i resource

4. Select the fix package for your virtual server and click **Finish** to submit the installation job.

   The installation can be run immediately or scheduled to run later. The installation job can be monitored in **Active and Schedule Jobs** as shown Figure 17-11.

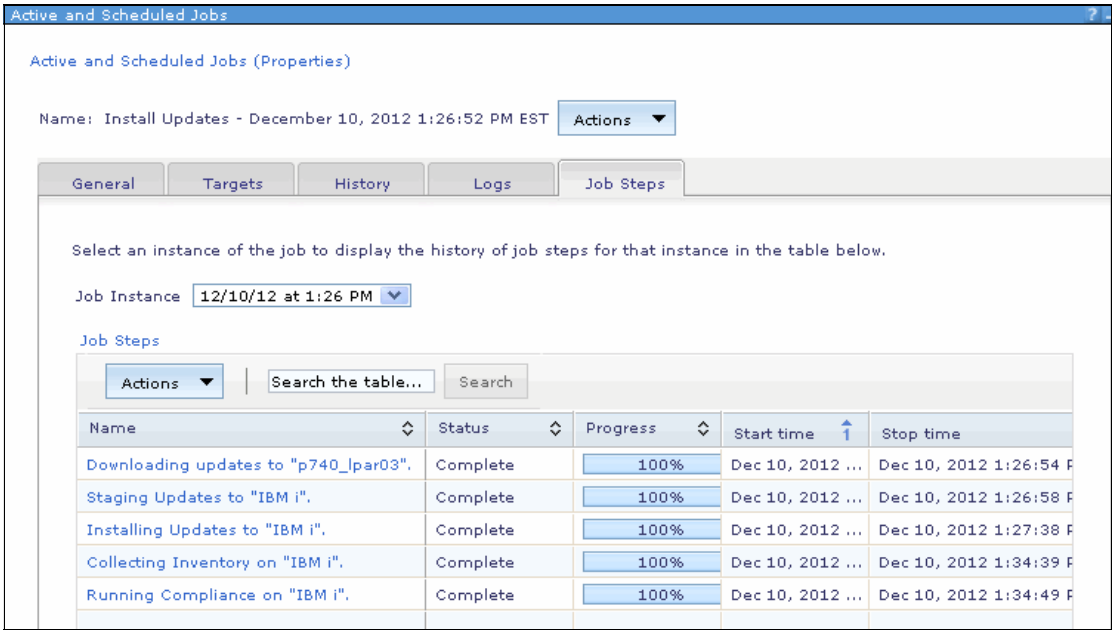


Figure 17-11 Monitoring an installation job

Installing updates on Virtual I/O Server

Virtual I/O Server Fixpacks are staged in the `/opt` file system on the target Virtual I/O Server before they are applied. The `/opt` file system must have at least 3 GB of free space available.

You can automatically reboot the Virtual I/O Server after the fixes are applied.

Important: If the Virtual I/O Server is a member of a shared storage pool, the cluster services on the Virtual I/O Server must be stopped before its upgrade.

Also, ensure that the VIO Server software levels in the shared storage pool cluster support the *rolling updates* feature. This feature allows cluster nodes to be upgraded independently of each other. Read the VIOS fix pack installation notes for more information.

Virtual I/O Server migration images are staged in the /export/um_lpp_source file system on the NIM master that is defined in “Configuring Update Manager” on page 635. The /export/um_lpp_source file system must have at least 16 GB of free space available. The file system is created in rootvg if it does not exist.

Installing updates on AIX

A NIM server is required for AIX updates as defined in “Configuring Update Manager” on page 635. If the target AIX resource does not exist as a NIM client for the defined NIM server, IBM Systems Director automatically adds the target system as a NIM client.

The updates are staged in the file system /export/um_lpp_source on the NIM server. This file system is automatically created by IBM Systems Director if it does not exist.

Updates can only be installed within a release of AIX, you cannot upgrade to a new version of AIX with Update Manager.

AIX 5.3 TL6 SP4 and later releases are supported by IBM Systems Director Update Manager.

Installing updates on IBM i

You must create a mapping of the IBM Systems Director user ID to the privileged IBM user ID on the target host. This mapping then performs the fix installation (for example, QSECOFR) before installing updates on IBM i virtual servers. The following steps create this mapping:

1. Select **Security** → **Credentials** → **Create** from the left navigation pane.
2. Select the IBM i system and supply the privileged user ID and password for the host (QSECOFR).
3. Select the created credential and click **Actions** → **Mappings**. You can then map the IBM Systems Director web interface user ID to the IBM i privileged user ID.

IBM Systems Director does not provide an option for automatic reboots for IBM i virtual servers. You must manually restart the IBM i virtual server after you apply the fixes,

IBM i v5r4 and later releases are supported by IBM Systems Director Update Manager.

Installing updates on Linux

IBM Systems Director only supports updates to Red Hat Enterprise Linux versions 5.x and 6.x, and SUSE Linux Enterprise Server 10 and 11.

The target Linux virtual server must have network connectivity to the IBM Systems Director management server AND either a direct or proxy connection to the Internet to obtain the updates from the Linux Distribution Partner.

The Linux virtual machine must be a Common Agent managed system.

zip, gunzip, rug (for SUSE 10), zypper (for SUSE 11), and yum (for RHEL5 and 6) must be installed on the Linux virtual servers.

Installing firmware updates on Power Systems

IBM Systems Director can import firmware updates after they are downloaded along with their `.sdd` files from IBM Fix Central. Firmware updates for all POWER6 and POWER7 managed systems along with a limited set for POWER5 managed systems can be obtained from IBM Fix Central.

System Discovery and inventory collection must be run on the target managed system and their HMCs or IVMs before you apply the firmware updates.

HMCs must meet the minimum required levels for the proposed version of POWER firmware before you apply the firmware updates.

Installing updates on Hardware Management Consoles

IBM Systems Director Version 6.3.1.1 does not support automatic Hardware Management Console (HMC) updates. HMCs must be updated manually.

17.2 Monitoring IBM Systems Director

This chapter covers some common tasks included with IBM Systems Director for monitoring your PowerVM environment:

- Viewing the real-time system status and health of discovered systems
- Viewing common performance monitors
- Setting and responding to alert thresholds in relation to warnings or critical messages.

17.2.1 IBM Systems Director monitors

The monitor task provides real-time status and performance statistics for resources in your PowerVM environment. You can observe changes in system resources, set thresholds for each monitor, graph the data, and establish automation plans to respond to an exceeded threshold.

Selecting and viewing monitors

In the IBM Systems Director Web interface navigation area, complete these steps to select and view monitors:

1. Select **System Status and Health** → **Monitors**.
2. Click **Browse** and **Add** discovered resources to be monitored.
3. Click **OK**.
4. Select the type of monitors that you want for the selected target systems, and click **Show Monitors**.

   IBM Systems Director retrieves the performance and status information for the target systems and displays the full list of available monitors and their real-time values.

A small sample of the available monitors for Virtual I/O Servers is shown in Figure 17-12.

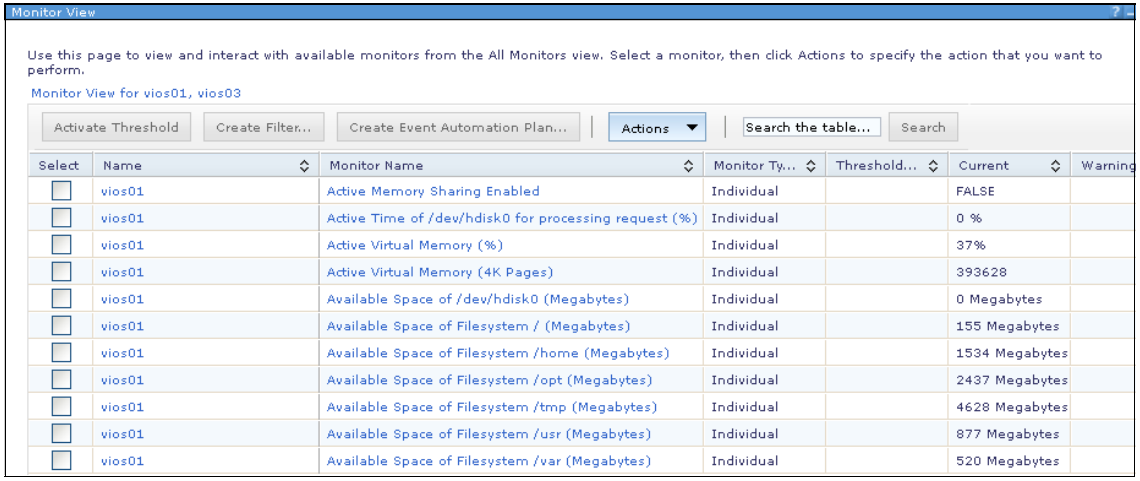


Figure 17-12 Sample of system health monitors

When the data is returned, you can view it in text format or graph the data for a visual representation. You can view a monitors real-time graph by selecting a monitor. For example, you can select CPU % utilization → **Action** → **Graph** as shown in Figure 17-13.

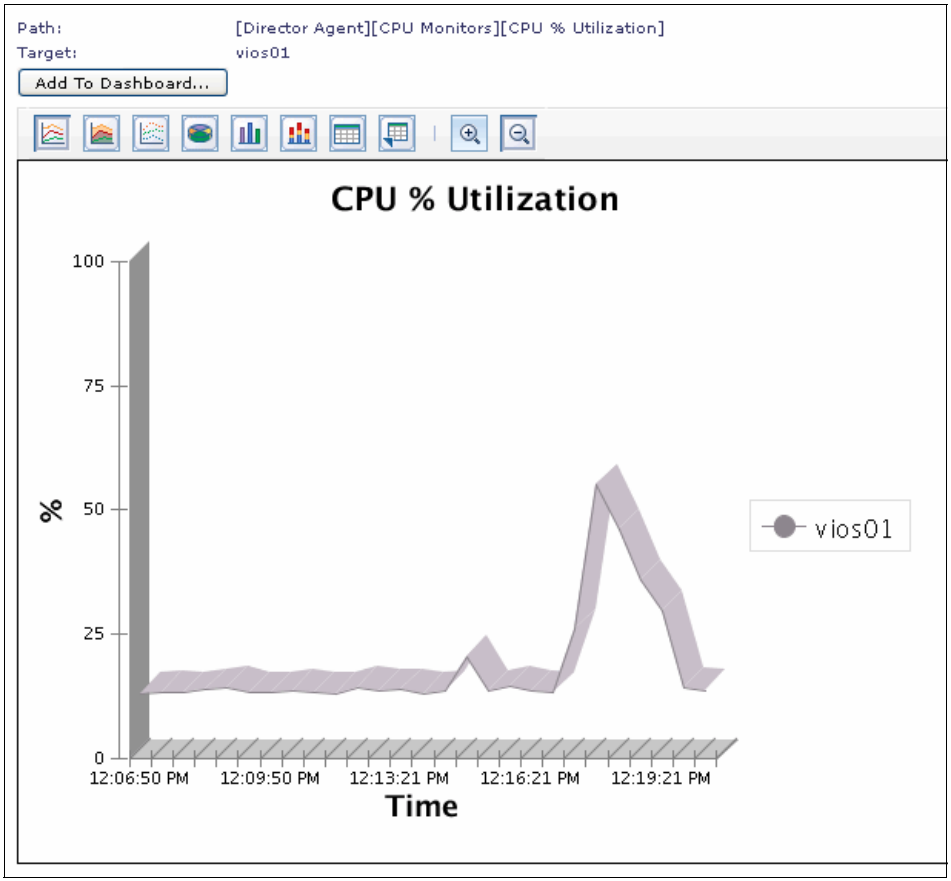


Figure 17-13 Graph of CPU utilization monitor

The monitor can be added to the IBM Systems Director *dashboard* from this window where it then remains available to the IBM Systems Director administrator for real-time monitoring.

The Health Summary Dashboard

The IBM Systems Director Dashboard can be displayed by selecting **System Status and Health** → **Health summary** in the IBM Systems Director Web interface navigation pane. The dashboard in Figure 17-14 shows graphical monitors for Virtual I/O Server entitled capacity utilization, Shared Ethernet Adapter throughput rates, memory steal rates, and the I/O throughput of a virtual SCSI adapter. You can add monitors to display the built-in performance metrics of different operating systems.

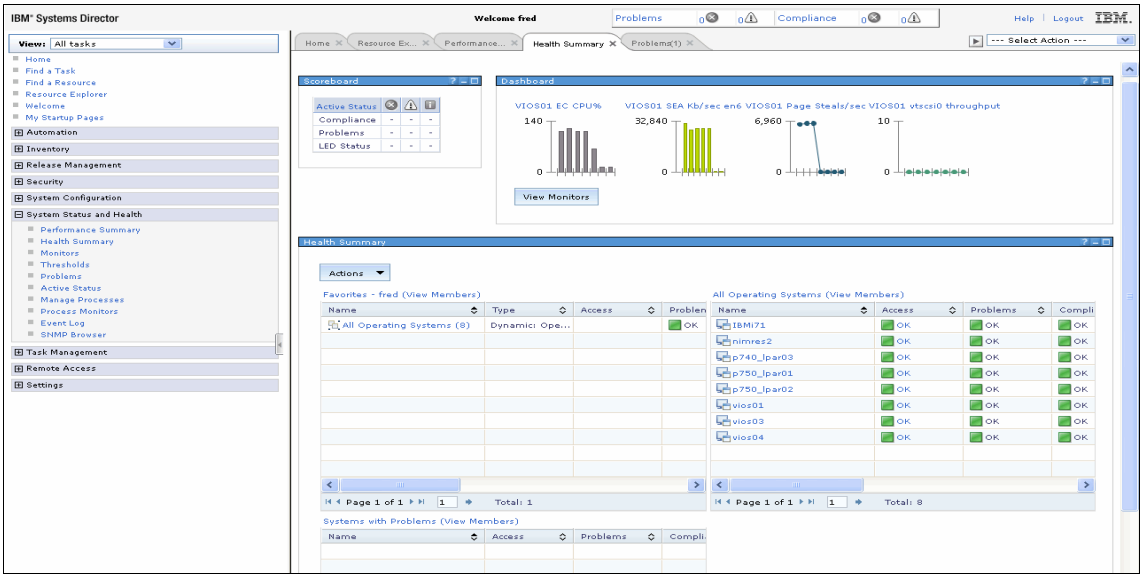


Figure 17-14 The Health Summary Dashboard

The dashboard also shows a summary scorecard for the total of all Critical, Warning, and Informational messages that must be responded to in the PowerVM environment.

Further down the dashboard is a list of the members of the All Operating Systems group, with traffic light indicators to show detected problems. Any systems that have reported problems are shown separately at the bottom of the dashboard, which is empty in Figure 17-14.

Setting monitor thresholds

You can set high and low monitor threshold limits in IBM Systems Director with the option of specifying warning and critical values. For example, the warning limit for the percentage used space on a file system may be 85%, and the critical limit 95%.

Activating the threshold includes setting options for generating an event when the threshold is exceeded, and determining the amount of time the threshold waits before resending the information. The following steps set monitor thresholds for target systems:

1. Select **System Status and Health** → **Monitors** in the IBM Systems Director Web interface navigation pane.
2. Click **Browse** and **Add** discovered resources to be monitored.
3. Click **OK**.
4. Select the type of monitors that you want for the selected target systems, and click **Activate Threshold**.

5. Set the warning and critical values, the amount of time the condition must be true before sending a warning, and the amount of time to wait before resending a warning or critical message for the Active Virtual Memory monitor as shown in Figure 17-15.

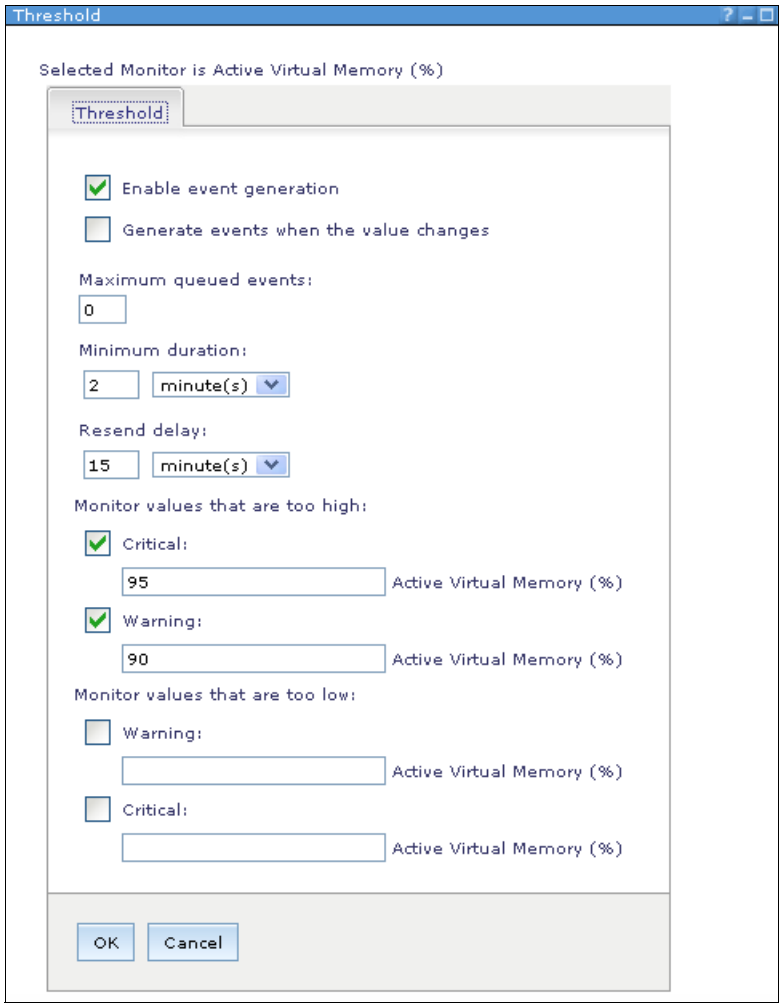


Figure 17-15 Setting monitor thresholds

The monitor shows ‘activated’ in the main monitor view.

17.2.2 Viewing and responding to warnings and critical messages

When threshold levels are exceeded, a warning or critical message is sent to the main **Health Summary dashboard** as displayed in Figure 17-16.

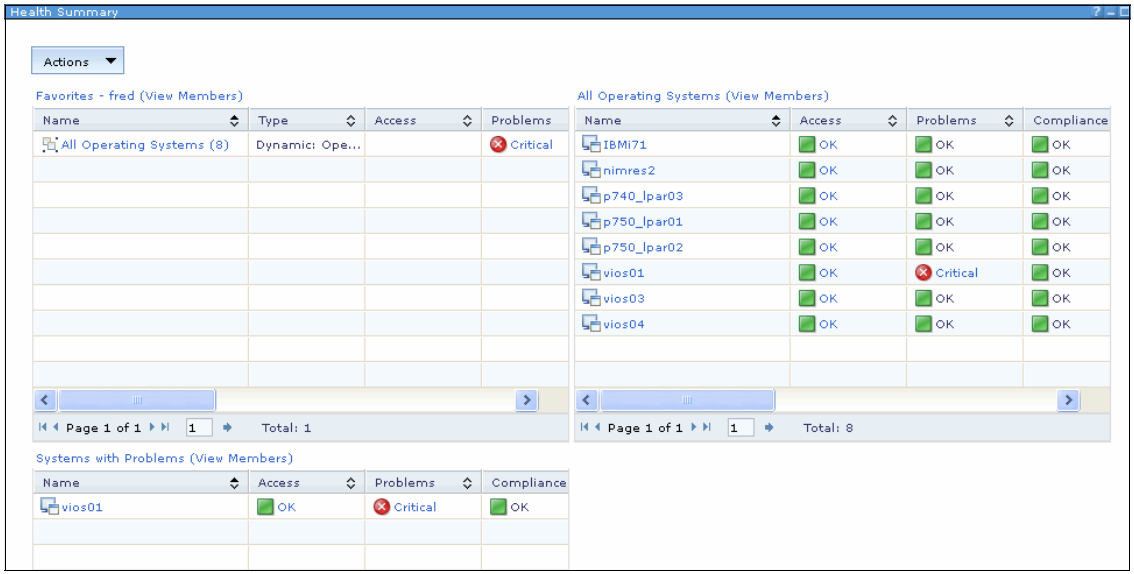


Figure 17-16 Viewing warning and critical messages

Select the **Critical** link to view the detail of the message in the **Event Log** as shown in Figure 17-17.

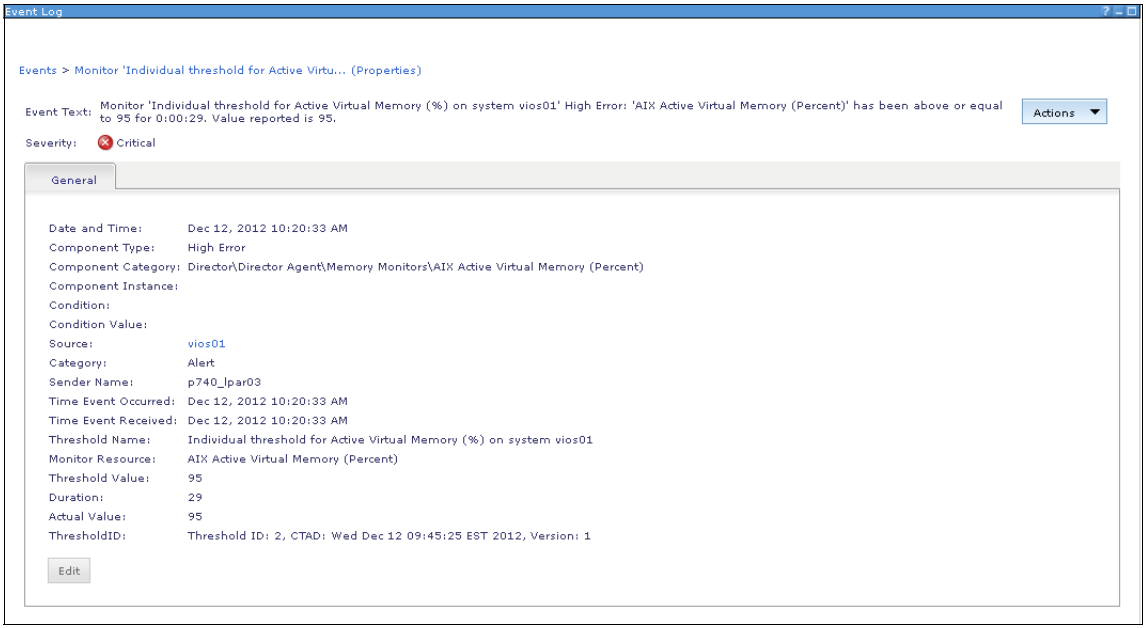


Figure 17-17 Critical message details

You can acknowledge receipt of the message from the critical message detail window. You can also set up event automation plans in response to receiving warning or critical messages. For example, you might choose to initiate a Dynamic LPAR operation to add more memory to a virtual server when you receive a critical message.

Performance monitoring

A number of real-time performance metrics can be selected in IBM Systems Director to monitor targeted systems. These real-time monitors are available by selecting **System Status and Health** → **Health summary** → **View Monitors.**

You cannot save performance data in IBM Systems Director for capacity planning or ad hoc performance reporting over time.

IBM Systems Director Enterprise Edition includes IBM Tivoli Monitoring, which saves performance data to an RDBMS. Generally, use Tivoli Monitoring for historical and capacity planning functions. For more information, see Chapter 18, “Tivoli Systems Management integration” on page 651.

Extending the Managing and Monitoring Capabilities of IBM Systems Director

After you implement your base IBM Systems Director infrastructure, you can extend the capabilities of the product by installing one or more plug-ins.

The base installation of IBM Systems Director includes a 30-day trial of IBM Systems Director VMControl plug-in, which can be activated from the IBM Systems Director welcome window. You can use VMControl to rapidly deploy virtual appliances to create virtual servers with the operating system, fixes, and software applications that you want.

Storage Control

The Storage Control plug-in extends the System Discovery, inventory, and monitoring capabilities of the IBM Systems Director management server to include IBM storage subsystems and Fibre Channel switches.

Network Control

The Network Control plug-in extends the System Discovery, inventory, and monitoring capabilities of the IBM Systems Director management server to include IBM and third-party network devices.

Active Energy Manager

The base installation of IBM Systems Director includes a 30-day trial of the IBM Systems Director Active Energy Manager™ plugin. You can use this plug-in to monitor the power and cooling requirements for your POWER servers and BladeCenter systems. Non-IBM systems can also be monitored.

18

Tivoli Systems Management integration

Virtual I/O Server includes a number of preinstalled Tivoli agents that you can use to integrate your Virtual I/O Servers, along with other managed Power Systems, into an existing Tivoli Systems Management infrastructure.

This chapter includes the following sections:

- Managing Tivoli Systems Management integration
- Monitoring using Tivoli management systems

18.1 Managing Tivoli Systems Management integration

This section addresses how to configure the preinstalled agents to allow integration with the following IBM Tivoli products:

- IBM Tivoli Monitoring
- IBM Tivoli Usage and Accounting Manager
- IBM Tivoli Storage Manager
- IBM Tivoli Storage Productivity Center
- IBM Tivoli Application Dependency Discovery Manager

18.1.1 IBM Tivoli Monitoring

IBM Tivoli Monitoring manages and monitors system and network applications on various operating systems, tracks the availability and performance of your systems, and provides reports to track trends and troubleshoot problems. You can use Tivoli Monitoring to monitor both the physical and logical resources of Power Systems, including Virtual I/O Server resources.

An overview of IBM Tivoli Monitoring for Power Systems is shown on Figure 18-1.

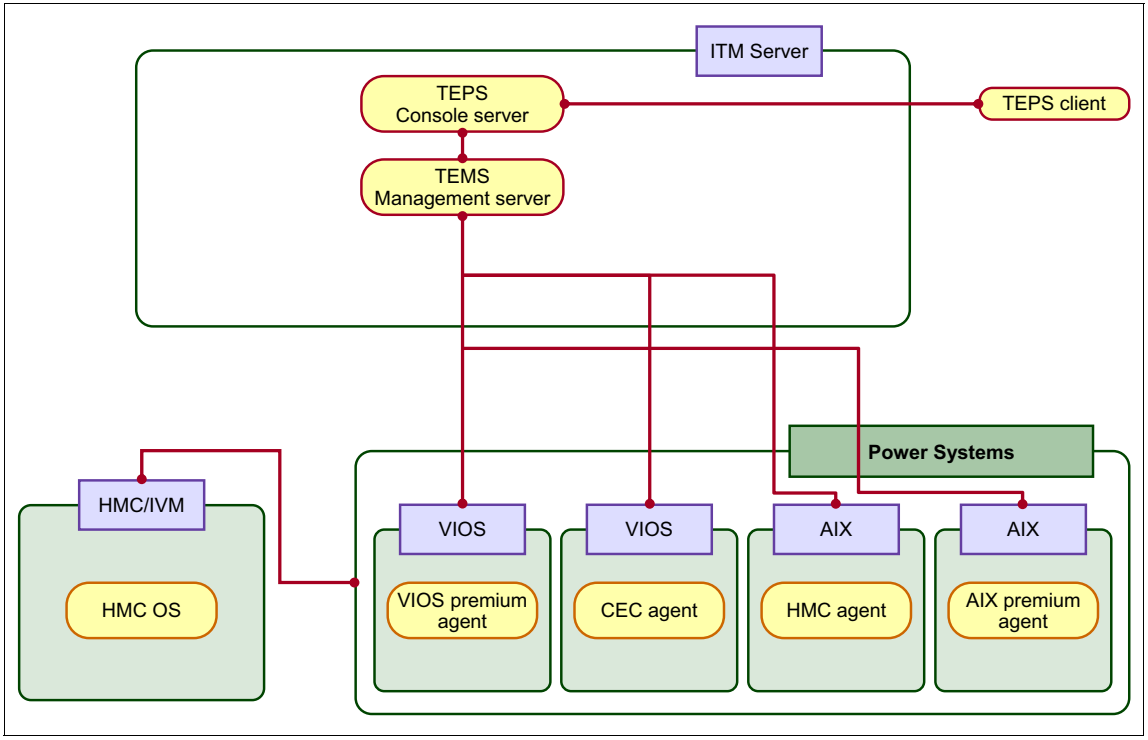


Figure 18-1 IBM Tivoli Monitoring: Power Systems overview

The basic installation of IBM Tivoli Monitoring requires the following components:

1. One or more Tivoli Enterprise Monitoring Servers, which act as a collection and control point for alerts that are received from the agents, and collect their performance and availability data. The monitoring server also manages the connection status of the agents.
2. A Tivoli Enterprise Portal Server, which provides the core presentation layer for retrieval, manipulation, analysis, and pre-formatting of data.
3. One or more Tivoli Enterprise Portal clients with Java-based user interface for viewing and monitoring your enterprise. There are two kinds of portal clients: Browser client and desktop client
4. Tivoli enterprise monitoring agents, which are installed on the systems you want to monitor. These agents collect data from monitored or managed

systems and distribute it to a monitoring server. Four different Power Systems agents are available:

a. Virtual I/O Server Premium Agent: Monitors the health of the VIOS, provides mapping of storage and network storage resources to client LPAR, and provides utilization statistics. IBM Tivoli Monitoring enables you to monitor the health and availability of multiple IBM Power System servers from the Tivoli Enterprise Portal. If you are already using IBM Tivoli Monitoring, you can integrate the Virtual I/O Server and client partition agents into your existing Tivoli Enterprise Monitoring Server. Virtual I/O Server has the Premium Agent installed.

b. CEC Agent: Provides overall processor and memory utilization of the frame for monitored partitions. The central electronic complex (CEC) agent is installed by default on VIOS since version 2.2.0.0 Fix Pack 24. The CEC agent can also be installed in a AIX partition, but running on a VIO server is the better placement.

c. AIX Premium Agent: The AIX Premium Agent provides statistics for each LPAR (entitled CPU, physical and logical CPUs), memory utilization, and disk and network performance. This agent provides usage statistics for WPARs as well.

d. HMC agent: Provides health and availability of the Hardware Management Console (HMC). HMC agent can be installed on any AIX partition, but requires a non-prompted SSH connection to the HMC.

At the time of writing, the current version of IBM Tivoli Monitoring version is 6.2.3 Fix Pack 1. For more information, visit this website:

http://www-01.ibm.com/software/tivoli/products/monitor/

The Tivoli Monitoring 6.2.3 Fix Pack 1 Information Center is available at:

http://publib.boulder.ibm.com/infocenter/tivihelp/v15r1/index.jsp?topic=/com.ibm.itm.doc_6.2.3fp1/welcome.htm

VIOS Premium Agent configuration

The IBM Tivoli Monitoring VIOS Premium Agent is installed by default on the Virtual I/O Server. To configure it, complete the following steps:

1. List all agents available on the Virtual I/O Server as shown in Example 18-1.

Example 18-1 Listing agents available on VIOS

```
$ lssvc
ITM_premium
ITM_cec
TSM_base
```

```
ITUAM_base
DIRECTOR_agent
perfmgr
ipsec_tunnel
ILMT
```

2. List all the attributes that are associated with the agent configuration, as shown in Example 18-2.

Example 18-2 Listing attributes available for ITM_premium agent

```
$ cfgsvc -ls ITM_premium
MANAGING_SYSTEM
HOSTNAME
RESTART_ON_REBOOT
MIRROR
```

3. Configure the agent, as shown in Example 18-3.

Example 18-3 Configuring the ITM_premium agent

```
$ cfgsvc ITM_premium -attr RESTART_ON_REBOOT=TRUE
HOSTNAME=p750_lpar01 MANAGING_SYSTEM=hmc8

Agent configuration started...
Agent configuration completed...
```

The RESTART_ON_REBOOT attribute that is set to TRUE specifies to restart the VIOS Premium Agent when the Virtual I/O Server is being rebooted. The HOSTNAME attribute specifies the Tivoli Enterprise Monitoring Server host name. The MANAGING_SYSTEM is the HMC host name. The MIRROR attribute is the host name or IP address of a secondary Tivoli Enterprise Monitoring Server server, and is optional.

4. Check the agent configuration, as shown in Example 18-4.

Example 18-4 Checking the ITM_premium agent configuration

```
$ lssvc ITM_premium
MANAGING_SYSTEM:hmc8
HOSTNAME:p750_lpar01
RESTART_ON_REBOOT:TRUE
MIRROR:
```

5. Prepare the **ssh** configuration.

   VIOS Premium Agent requires **ssh** communication between the Virtual I/O Server and the HMC to allow the monitoring agent on the VIO Server to gather extra information only accessible from the HMC.

   Display the **ssh** public key that is generated for a particular agent configuration, as shown in Example 18-5.

Example 18-5 Listing the VIOS host key to be used by the ITM_premium agent

```
$ cfgsvc ITM_premium -key
ssh-rsa
AAAAB3NzaC1yc2EAAAABIwAAAQEAvLNFaiXBDHe2pbZ7TSOHRmfdLAqCzT8PHn2fFlVf
V4S/hTaEO6FzkmTCJPfs9dB5ATnRuKC/WF9ZfOBU3hVE/Lm1qyCkK7wym1TKOOseaRu+
xKKZjGZkHf/+BDvz4M/nvAELUGFco5vE+e2pamv9jJyfYvvGMSI4hj6eisRvjTfckcGp
r9vBbJN27Mi7T6RVZ2scN+5rRK8Okj+bqvgPZZk4IldOMLboRossgTO1LURo3bGvuih9
Xd3rUIdObQdj8dJK8mVoA5T1O+RF/zvPiQU9GT6XvcjmQdKbxLm2mgdATcPaDN4vOSBv
XTaNSozpPnNhyxvpugidtZBohznBDQ== root@vios03
```

Tip: The **ssh** key can also be retrieved directly from the HMC with the **viosrvcmd** command:

```
viosvrcmd -m <machine> -p <lpar> -c "cfgsvc ITM_premium -key"
```

6. Log in to the HMC as hscroot, and add the **ssh** public key from the previous step, as shown in Example 18-6.

Example 18-6 Adding the VIOS host key to the HMC

```
hscroot@hmc1:~> mkauthkeys --add 'ssh-rsa
AAAAB3NzaC1yc2EAAAABIwAAAQEAvLNFaiXBDHe2pbZ7TSOHRmfdLAqCzT8PHn2fFlVf
V4S/hTaEO6FzkmTCJPfs9dB5ATnRuKC/WF9ZfOBU3hVE/Lm1qyCkK7wym1TKOOseaRu+
xKKZjGZkHf/+BDvz4M/nvAELUGFco5vE+e2pamv9jJyfYvvGMSI4hj6eisRvjTfckcGp
r9vBbJN27Mi7T6RVZ2scN+5rRK8Okj+bqvgPZZk4IldOMLboRossgTO1LURo3bGvuih9
Xd3rUIdObQdj8dJK8mVoA5T1O+RF/zvPiQU9GT6XvcjmQdKbxLm2mgdATcPaDN4vOSBv
XTaNSozpPnNhyxvpugidtZBohznBDQ== root@vios03'
```

7. Confirm that you have non-prompted **ssh** access from the VIOS, as root, to the HMC, and accept the HMC's host key, as shown in Example 18-7.

Example 18-7 Confirming non-prompted access from VIOS to managing HMC

```
$ oem_setup_env
# ssh hscroot@hmc8 lshmc -V
The authenticity of host 'hmc8 (172.16.20.114)' can't be
established.
```

```
RSA key fingerprint is
5c:13:ff:36:24:ef:8c:68:76:ce:38:5b:c6:1c:b0:aa.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'hmc8' (RSA) to the list of known hosts.
"version= Version: 7
 Release: 7.6.0
 Service Pack: 0
HMC Build level 20120828.1
HMC Driver AP06_1235A (0829) Rev 1.0
","base_version=V7R7.6.0"
```

8. Start the monitoring agent as shown in Example 18-8.

Example 18-8 Starting ITM_premium

```
$ startsvc ITM_premium
Starting Premium Monitoring Agent for VIOS ...
Premium Monitoring Agent for VIOS started
```

CEC Base Agent configuration

The IBM Tivoli Monitoring CEC Base Agent is now installed by default on the Virtual I/O Server. Use the following steps to configure it:

1. List all the attributes that are associated with the agent configuration, as shown in Example 18-9.

Example 18-9 List attributes available for ITM_cec agent

```
$ cfgsvc -ls ITM_cec
HOSTNAME
MANAGING_SYSTEM
SECOND_MANAGING_SYSTEM
CEC
DIRECTOR_HOST_ADDRESS
DIRECTOR_AUTHENTICATION
DIRECTOR_PORT_NUMBER
RESTART_ON_REBOOT
MIRROR
```

2. Configure the agent as shown in Example 18-10.

Example 18-10 Configuring ITM_cec agent

```
$ cfgsvc ITM_cec -attr RESTART_ON_REBOOT=TRUE HOSTNAME=p750_lpar01
MANAGING_SYSTEM=hmc8 cec=p740 DIRECTOR_HOST_ADDRESS=p740_lpar03
DIRECTOR_AUTHENTICATION=yes DIRECTOR_PORT_NUMBER=8422
```

```
Agent configuration started...
Agent configuration completed...
```

The RESTART_ON_REBOOT attribute that is set to TRUE specifies to restart the CEC Base Agent when the Virtual I/O Server is being rebooted. The HOSTNAME attribute specifies the Tivoli Enterprise Monitoring Server host name. The MANAGING_SYSTEM is the Hardware Management Console host name. The CEC attribute is the name of CEC to be monitored.

To get a list of CEC names, use the following command on HMC:

```
lssyscfg -r sys -F name
```

The `mirror` and `second_managing_system` attributes are optional, and are the host names of a second Tivoli Enterprise Monitoring Server server, and HMC.

The DIRECTOR_HOST_ADDRESS attribute specifies the Systems Director host name. The DIRECTOR_AUTHENTICATION attribute specifies a System Director authentication value. The DIRECTOR_PORT_NUMBER is a Systems Director port number. The DIRECTOR_HOST_ADDRESS, DIRECTOR_AUTHENTICATION, and DIRECTOR_PORT_NUMBER are optional attributes and must be provided only if you are interested in starting IBM Systems Director contextually from the CEC agent workspace.

3. Check the agent configuration, as shown in Example 18-11.

Example 18-11 Confirming the attribute settings for ITM_cec agent

```
$ lssvc ITM_cec
HOSTNAME:p750_lpar01
MANAGING_SYSTEM:hmc8
SECOND_MANAGING_SYSTEM:
CEC:p740
DIRECTOR_HOST_ADDRESS:p740_lpar03
DIRECTOR_AUTHENTICATION:Yes
DIRECTOR_PORT_NUMBER:8422
RESTART_ON_REBOOT:TRUE
MIRROR:
```

4. Prepare the **ssh** configuration. CEC Base Agent requires **ssh** communication between the Virtual I/O Server and the HMC to allow the monitoring agent on the VIO Server to gather extra information only accessible from the HMC.

Tip: If you previously added the VIOS **ssh** key to the HMC, for the ITM_premium agent, this step is not required if the ITM_cec agent is on the same VIOS as the ITM_premium agent.

Display the **ssh** public key that is generated for a particular agent configuration, as shown in Example 18-12.

Example 18-12 Listing the VIOS ssh host key used by ITM_cec agent

```
$ cfgsvc ITM_cec -key
ssh-rsa
AAAAB3NzaC1yc2EAAAABIwAAAQEAvLNFaiXBDHe2pbZ7TSOHRmfdLAqCzT8PHn2fFlVf
V4S/hTaEO6FzkmTCJPfs9dB5ATnRuKC/WF9ZfOBU3hVE/Lm1qyCkK7wym1TKOOseaRu+
xKKZjGZkHf/+BDvz4M/nvAELUGFco5vE+e2pamv9jJyfYvvGMSI4hj6eisRvjTfckcGp
r9vBbJN27Mi7T6RVZ2scN+5rRK8Okj+bqvgPZZk4IldOMLboRossgTO1LURo3bGvuih9
Xd3rUIdObQdj8dJK8mVoA5T1O+RF/zvPiQU9GT6XvcjmQdKbxLm2mgdATcPaDN4vOSBv
XTaNSozpPnNhyxvpugidtZBohznBDQ== root@vios03
```

5. Connect to the HMC and add the **ssh** public key, as shown in Example 18-13.

Example 18-13 Adding the VIOS host key to HMC

```
hscroot@hmc1:~> mkauthkeys --add 'ssh-rsa
AAAAB3NzaC1yc2EAAAABIwAAAQEAvLNFaiXBDHe2pbZ7TSOHRmfdLAqCzT8PHn2fFlVf
V4S/hTaEO6FzkmTCJPfs9dB5ATnRuKC/WF9ZfOBU3hVE/Lm1qyCkK7wym1TKOOseaRu+
xKKZjGZkHf/+BDvz4M/nvAELUGFco5vE+e2pamv9jJyfYvvGMSI4hj6eisRvjTfckcGp
r9vBbJN27Mi7T6RVZ2scN+5rRK8Okj+bqvgPZZk4IldOMLboRossgTO1LURo3bGvuih9
Xd3rUIdObQdj8dJK8mVoA5T1O+RF/zvPiQU9GT6XvcjmQdKbxLm2mgdATcPaDN4vOSBv
XTaNSozpPnNhyxvpugidtZBohznBDQ== root@vios03'
```

6. Confirm that you have non-prompted **ssh** access from the VIOS, as root, to the HMC, as shown in Example 18-14. Accept the HMC's host key if necessary.

Example 18-14 Confirming non-prompted access from VIOS to HMC

```
$ oem_setup_env
# ssh hscroot@hmc8 lshmc -V
"version= Version: 7
 Release: 7.6.0
 Service Pack: 0
HMC Build level 20120828.1
HMC Driver APO6_1235A (0829) Rev 1.0
","base_version=V7R7.6.0"
```

7. Start the monitoring agent as shown in Example 18-15.

Example 18-15 Starting the ITM_cec agent

```
$ startsvc ITM_cec
Starting Base Monitoring Agent for CEC ...
Base Monitoring Agent for CEC started
```

HMC Base Agent configuration

The HMC Base Agent is installed on a AIX LPAR or physical system, rather than on the Virtual I/O Server. For more information about installing and configuring the agent, see the *HMC Base Agent User's Guide* at:

http://pic.dhe.ibm.com/infocenter/tivihelp/v15r1/index.jsp?topic=%2Fcom.ibm.itm.doc_6.2.2%2Fphmc6221_user.htm

AIX Premium Agent configuration

The AIX Premium Agent is installed on a AIX LPAR or physical system, rather than on the Virtual I/O Server. For more information about installing and configuring the agent, see the *AIX Premium Agent User's Guide* at:

http://pic.dhe.ibm.com/infocenter/tivihelp/v15r1/index.jsp?topic=%2Fcom.ibm.itm.doc_6.2.2%2Fpaix6221_user.htm

18.1.2 IBM Tivoli Usage and Accounting Manager agent

IBM Tivoli Usage and Accounting Manager (ITUAM) helps you track, allocate, and invoice your IT costs by collecting, analyzing, and reporting on the resources that are used by entities such as cost centers, departments, and users. ITUAM can gather data from multi-tiered data centers including Windows, AIX, Virtual I/O Server, HP/UX Sun Solaris, Linux, IBM i, and VMware.

For more information about the capabilities of ITUAM and how to use it, see Tivoli Usage and Accounting Manager in the IBM Tivoli Systems Management Information Center at:

http://pic.dhe.ibm.com/infocenter/tivihelp/v3r1/index.jsp?topic=%2Fcom.ibm.ituam.doc_7.3%2Fwelcome.htm

You can configure and start the IBM Tivoli Usage and Accounting Manager agent on the Virtual I/O Server.

Complete these steps to configure the IBM Tivoli Usage and Accounting Manager agent:

1. List all the attributes that are associated with the agent configuration as shown in Example 18-16.

Example 18-16 Listing attributes for ITUAM_base agent

```
$cfgsvc –ls ITUAM_base
ACCT_DATA0
ACCT_DATA1
ISYSTEM
IPROCESS
```

2. Configure the agent as shown in Example 18-17.

Example 18-17 Configuring the ITUAM_base agent

```
$ cfgsvc ITUAM_base -attr ACCT_DATA0=value1 ACCT_DATA1=value2
ISYSTEM=value3 IPROCESS=value4
```

Where:

value1	This is the size (in MB) of the first data file that holds daily accounting information.
value2	This is the size (in MB) of the second data file that holds daily accounting information.
value3	This is the time (in minutes) when the agent generates system interval records.
value4	This is the time (in minutes) when the system generates aggregate process records.

3. Restart the agent.
4. Stop and restart the monitoring agent to use the new configuration by running the **stopsvc** and **startsvc** commands as shown in Example 18-18.

Example 18-18 Stopping and starting the ITUAM_base agent

```
$ stopsvc ITUAM_base
Stopping agent...
Agent stopped...
$ startsvc ITUAM_base
Starting agent...
Agent started...
```

After you start the IBM Tivoli Usage and Accounting Manager agent, it begins to collect data and generate log files. You can configure the ITUAM server to retrieve the log files, which are then processed by the IBM Tivoli Usage and Accounting Manager Processing Engine.

You can work with the data from the IBM Tivoli Usage and Accounting Manager Processing Engine as follows:

- You can generate customized reports, spreadsheets, and graphs. ITUAM provides full data access and reporting capabilities by integrating Microsoft SQL Server Reporting Services or Crystal Reports with a database management system (DBMS).
- You can view high-level and detailed cost and usage information.
- You can allocate, distribute, and charge IT costs to users, cost centers, and organizations in a manner that is fair, understandable, and reproducible.

For more information, see one of the following resources:

- If you are running the IBM Tivoli Usage and Accounting Manager Processing Engine on Windows, see the *IBM Tivoli Usage and Accounting Manager Data Collectors for Microsoft Windows User's Guide,* SC32-1557-02, at:

  `http://www.elink.ibmlink.ibm.com/public/applications/publications/cgibin/pbi.cgi?CTY=US&FNC=SRX&PBL=SC32-1557-02`

- If you are running the IBM Tivoli Usage and Accounting Manager Processing Engine on UNIX or Linux, see the *IBM Tivoli Usage and Accounting Manager Data Collectors for UNIX and Linux User's Guide,* SC32-1556-02, at:

  `http://www.elink.ibmlink.ibm.com/public/applications/publications/cgibin/pbi.cgi?CTY=US&FNC=SRX&PBL=SC32-1556-02`

An ITUAM window that shows the available report types is shown in Figure 18-2.

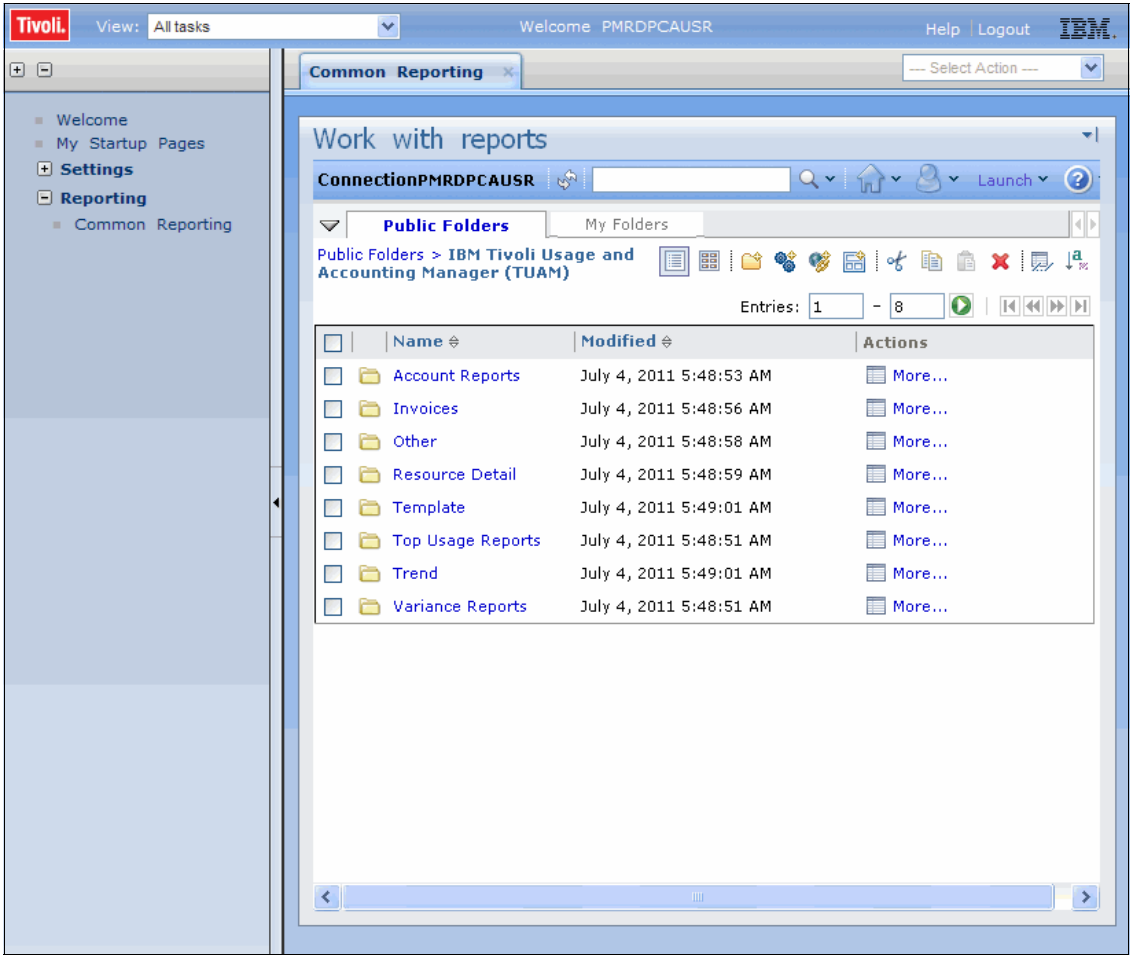


Figure 18-2 Usage and accounting reports available in ITUAM

The example output of an invoice report is shown in Figure 18-3.

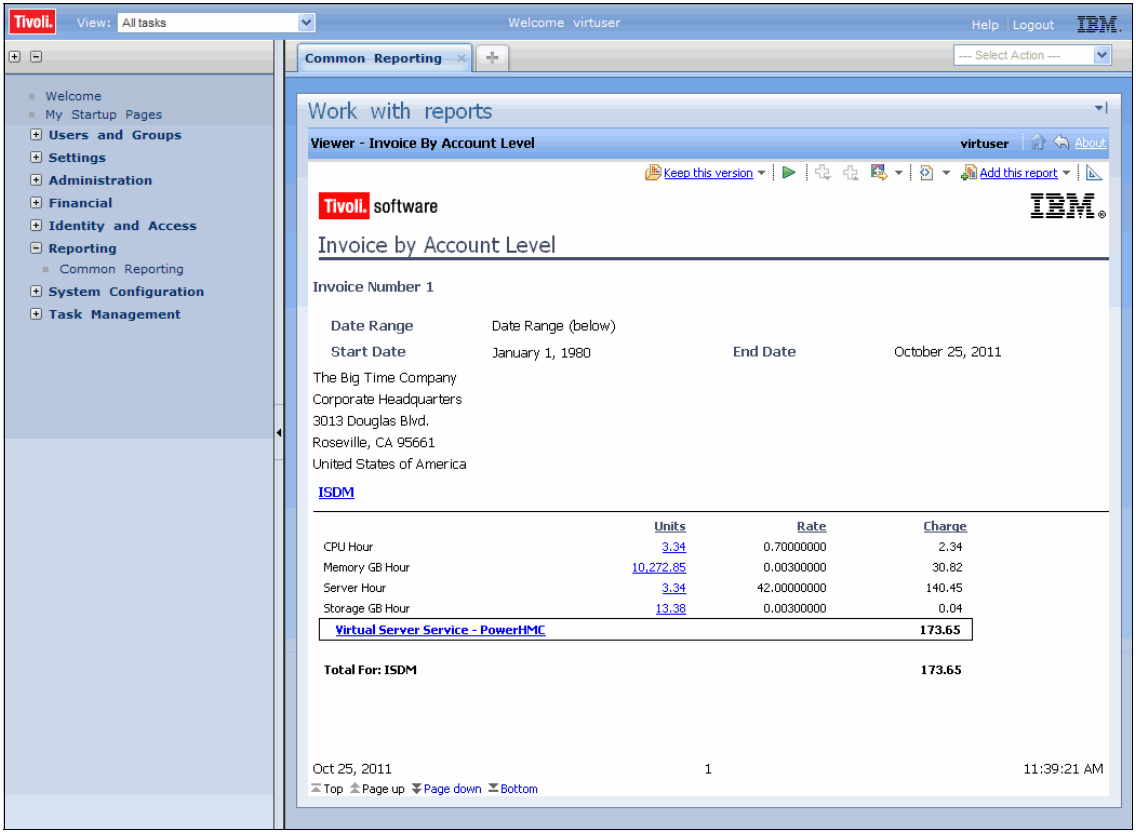


Figure 18-3 Sample output from ITUAM invoice report

18.1.3 IBM Tivoli Productivity Center

IBM Tivoli Storage Productivity Center is a storage infrastructure management suite. It is designed to help simplify and automate the management of storage devices, storage networks, and capacity utilization of file systems and databases. IBM Tivoli Productivity Center can help you perform the following activities:

- Manage capacity utilization of storage systems, file systems, and databases.
- Automate file system capacity provisioning.
- Perform device configuration and management of multiple devices from a single user interface.

- Tune and proactively manage the performance of storage devices on the storage area network (SAN).
- Manage, monitor, and control your SAN fabric.

> **Note:** IBM TotalStorage Productivity Center (TPC) was renamed Tivoli Storage Productivity Center in Version 4.1.

For more information about IBM Tivoli Storage Productivity Center, see the Information Center at:

http://publib.boulder.ibm.com/infocenter/tivihelp/v4r1/index.jsp?topic=%2Fcom.ibm.help.doc%2Ftpc_infocenter_home.htm

Complete these steps to configure the IBM Tivoli Productivity Center agent:

1. Use the **lssvc** command, as shown in Example 18-19, to confirm that the TPC agent is installed, and available for configuration. If the TPC agent is not listed, you can install it from the Virtual I/O Server media.

Example 18-19 List agents available for configuration

```
$ lssvc
ITM_premium
ITM_cec
TSM_base
ITUAM_base
DIRECTOR_agent
perfmgr
ipsec_tunnel
ILMT
TPC
```

2. To list all the attributes that are associated with an agent configuration, enter the command shown in Example 18-20.

Example 18-20 List the attributes that are supported by the TPC agent

```
$cfgsvc –ls TPC
A
S
devAuth
caPass
caPort
amRegPort
amPubPort
dataPort
```

```
devPort
newCA
oldCA
daScan
daScript
daInstall
faInstall
U
```

3. Configure the agent:

   The TPC agent is a Tivoli Storage Productivity Center agent. Agent names are case-sensitive. This agent requires that you specify the `S`, `A`, `devAuth`, and `caPass` attributes for configuration. By default, specifying this agent configures both the `TPC_data` and `TPC_fabric` agents, as provided in Table 18-1.

Table 18-1 TPC agent attributes, descriptions, and their values

Attributes	Description	Value
S	Provides the Tivoli Productivity Center agent with a Tivoli Productivity Center server host name.	Host name or IP address
A	Provides the Tivoli Productivity Center agent with an agent manager host name.	Host name or IP address
devAuth	Sets the Tivoli Productivity Center device server authentication password.	Password
caPass	Sets the CA authentication password.	Password
caPort	Sets the CA port. The default value is 9510.	Number
amRegPort	Specifies the agent manager registration port. The default value is 9511.	Number
amPubPort	Specifies the agent manager public port. The default value is 9513.	Number
dataPort	Specifies the Tivoli Productivity Center data server port. The default value is 9549.	Number
devPort	Specifies the Tivoli Productivity Center device server port. The default value is 9550.	Number

Attributes	Description	Value
newCA	The default value is true.	True or false
oldCA	The default value is true.	True or false
daScan	The default value is true.	True or false
daScript	The default value is true.	True or false
daInstall	The default value is true.	True or false
faInstall	The default value is true.	True or false
U	Specifies to uninstall the agent.	all | data | fabric

4. To configure the TPC agent with several attributes, run the cfgsvc command as shown in Example 18-21.

Example 18-21 Configuring the TPC agent

```
$ cfgsvc TPC -attr S=tpc_server_hostname A=agent_manager_hostname
devAuth=password caPass=password
```

5. The installation wizard is displayed. Type the number next to the language that you want to use for the installation and enter zero (0).

   The license agreement panel is displayed as shown in Example 18-22.

Example 18-22 Initializing the InstallShield Wizard

```
Initializing InstallShield Wizard........
Launching InstallShield Wizard........

------------------------------------------------------------------------
-----------
Select a language to be used for this wizard.

[ ] 1  - Czech
[X] 2  - English
[ ] 3  - French
[ ] 4  - German
[ ] 5  - Hungarian
[ ] 6  - Italian
[ ] 7  - Japanese
[ ] 8  - Korean
[ ] 9  - Polish
[ ] 10 - Portuguese (Brazil)
[ ] 11 - Russian
[ ] 12 - Simplified Chinese
```

```
[ ] 13 - Spanish
[ ] 14 - Traditional Chinese

To select an item enter its number, or 0 when you are finished: [0]
```

6. Read the license agreement panel. Type `1` to accept the terms of the license agreement. Then, type `0` to continue, as shown in Example 18-23. The agent is installed on the Virtual I/O Server according to the attributes specified in the **cfgsvc** command.

Example 18-23 Accepting the license agreement

```
Please choose from the following options:
[x] 1 - I accept the terms of the license agreement.
[ ] 2 - I do not accept the terms of the license agreement.

To select an item enter its number, or 0 when you are finished: [0]

Installing TPC Agents
Install Location: /opt/IBM/TPC
TPC Server Host: tpc_server_hostname
Agent Manager Host: agent_manager_hostname
```

7. Start the agent as shown in Example 18-24.

Example 18-24 Starting the agent

```
startsvc TPC_fabric
startsvc TPC_data
```

After the TPC agents are started, you can use the TPC user interface shown in Figure 18-4 to collect and view information about the Virtual I/O Server.

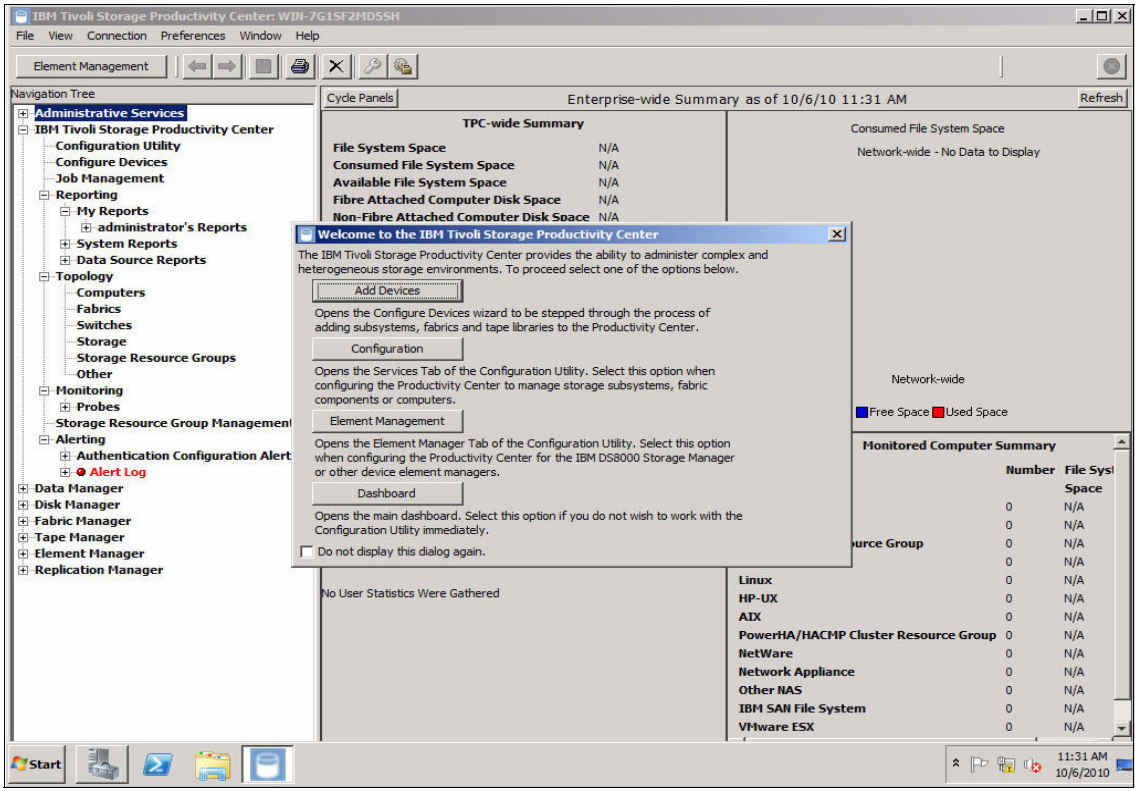


Figure 18-4 Tivoli Storage Productivity Center welcome panel

18.1.4 IBM Tivoli Application Dependency Discovery Manager

IBM Tivoli Application Dependency Discovery Manager discovers infrastructure elements that are found in the typical data center. These elements include application software; hosts; operating environments (including the Virtual I/O Server); network components such as routers, switches, load balancers, firewalls, and storage; and network services such as LDAP, NFS, and DNS.

Based on the data it collects, Tivoli Application Dependency Discovery Manager automatically creates and maintains application infrastructure maps that include runtime dependencies, configuration values, and change history. With this information, you can determine the interdependences between business applications, software applications, and physical components to help ensure and improve application availability in your environment.

For example, you can perform the following tasks:

- You can isolate configuration-related application problems.
- You can plan for application changes to minimize or eliminate unplanned disruptions.
- You can create a shared topological definition of applications for use by other management applications.
- You can determine the effect of a single configuration change on a business application or service.
- You can see what changes take place in the application environment, and where.

Tivoli Application Dependency Discovery Manager includes an agent-free discovery engine. This means that the Virtual I/O Server does not require that an agent or client be installed and configured to be discovered by Tivoli Application Dependency Discovery Manager. Instead, it uses discovery sensors that rely on open and secure protocols and access mechanisms to discover the data center components.

For more information, see the Tivoli Application Dependency Discovery Manager product website at:

http://www-01.ibm.com/software/tivoli/products/taddm/

18.2 Monitoring using Tivoli management systems

This section describes using Tivoli software tools to monitor the availability and performance of your Virtual I/O Server environment.

18.2.1 IBM Tivoli Monitoring

Using the Tivoli Enterprise Portal

This section demonstrates examples of using the Tivoli Enterprise Portal graphical interface to retrieve important information about the Virtual I/O Server that has the Tivoli Monitoring agent started:

- Virtual IO Mappings
- Top Resources
- System
- Storage
- Networking

- CEC Resources
- CEC Utilization

There are two kinds of Tivoli Enterprise Portal clients. The browser-based client requires a Java-enabled browser. The Tivoli Enterprise Portal Server desktop client can be installed from the Tivoli Monitoring base installation media for the platform that you want to install on. The Tivoli Enterprise Portal Server desktop client is only supported on Windows and Linux.

Starting the Tivoli Enterprise Portal Client browser application

Navigate to the host name or IP address of the IBM Tivoli Monitoring server followed by port 1920 as shown in the following example:

http://*<TEMS-server>***:1920**

Select **IBM Tivoli Enterprise Portal Web Client** from the web page as shown in Figure 18-5.

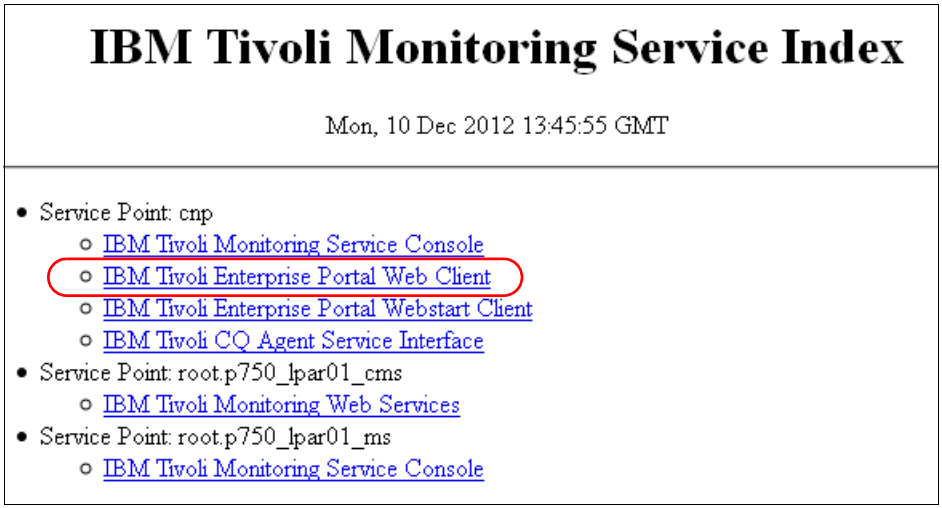


Figure 18-5 Tivoli Enterprise Portal login using web browser

You are asked for Authentication by a Java applet as shown in Figure 18-6 on page 672. Proceed with login to open the IBM Portal Client applet.

Starting the Tivoli Enterprise Portal Client desktop application

Start Tivoli Enterprise Portal Client (Figure 18-6) using **Start** → **Programs** → **IBM Tivoli Monitoring** → **Tivoli Enterprise Portal** for a Windows installation, or run **`itmcmd manage`** for a Linux installation.

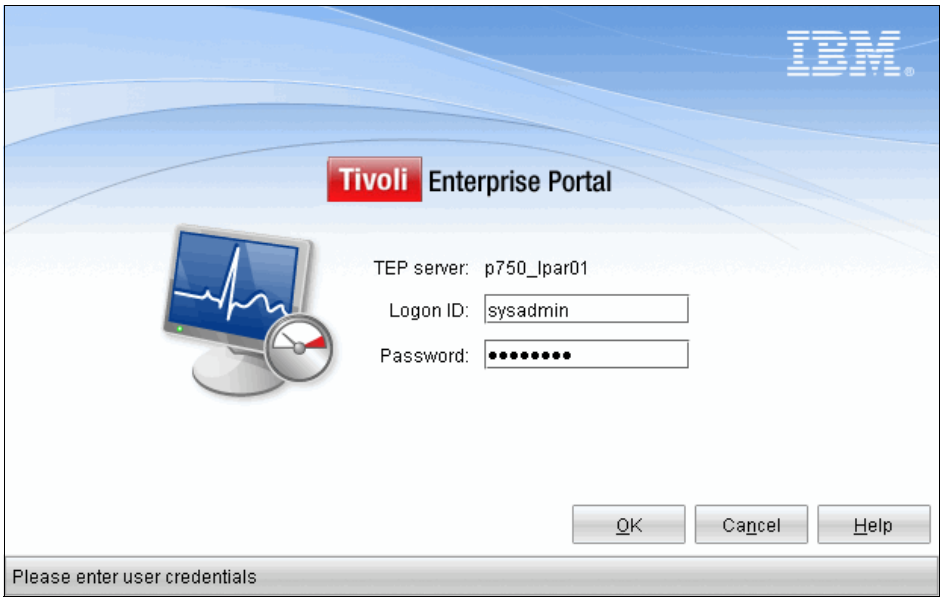


Figure 18-6 Tivoli Enterprise Portal login

Virtual IO mappings (VIOS Premium Agent)

Virtual IO mappings provide you with information about virtual SCSI or virtual network mappings that are based on the selected workspace. Figure 18-7 on page 673 shows how to select a workspace within the virtual IO mappings Navigator item.

Storage mappings (VIOS Premium Agent)

In the Navigator window, right-click **Virtual IO Mappings** and select **Workspace** → **Storage Mappings** as shown in Figure 18-7.

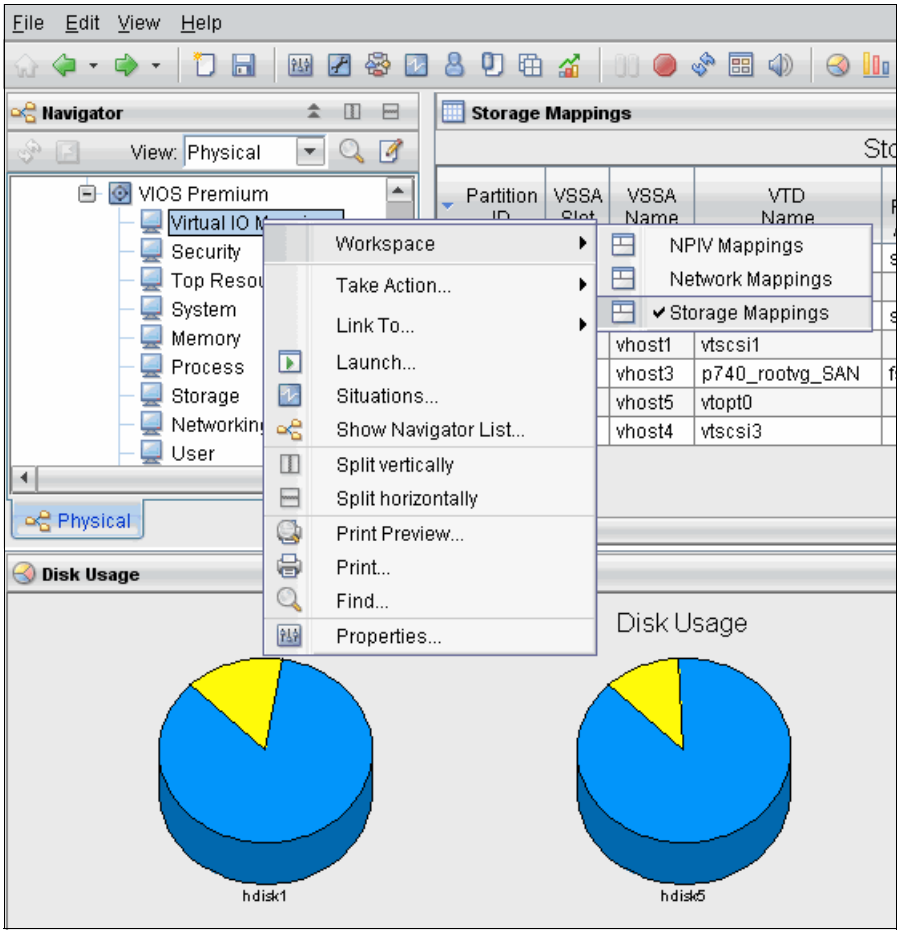


Figure 18-7 Storage Mappings workspace selection

Storage Mappings provide you with all information related to the virtual SCSI configuration. Figure 18-8 shows, for example, that the Virtual I/O Server `vios03` has a virtual SCSI server adapter `vhost1` with server slot number `11`, and that the client partition with partition ID `3` can access the physical disk that is on `vios03` by using a SCSI client adapter with client slot number `11`.

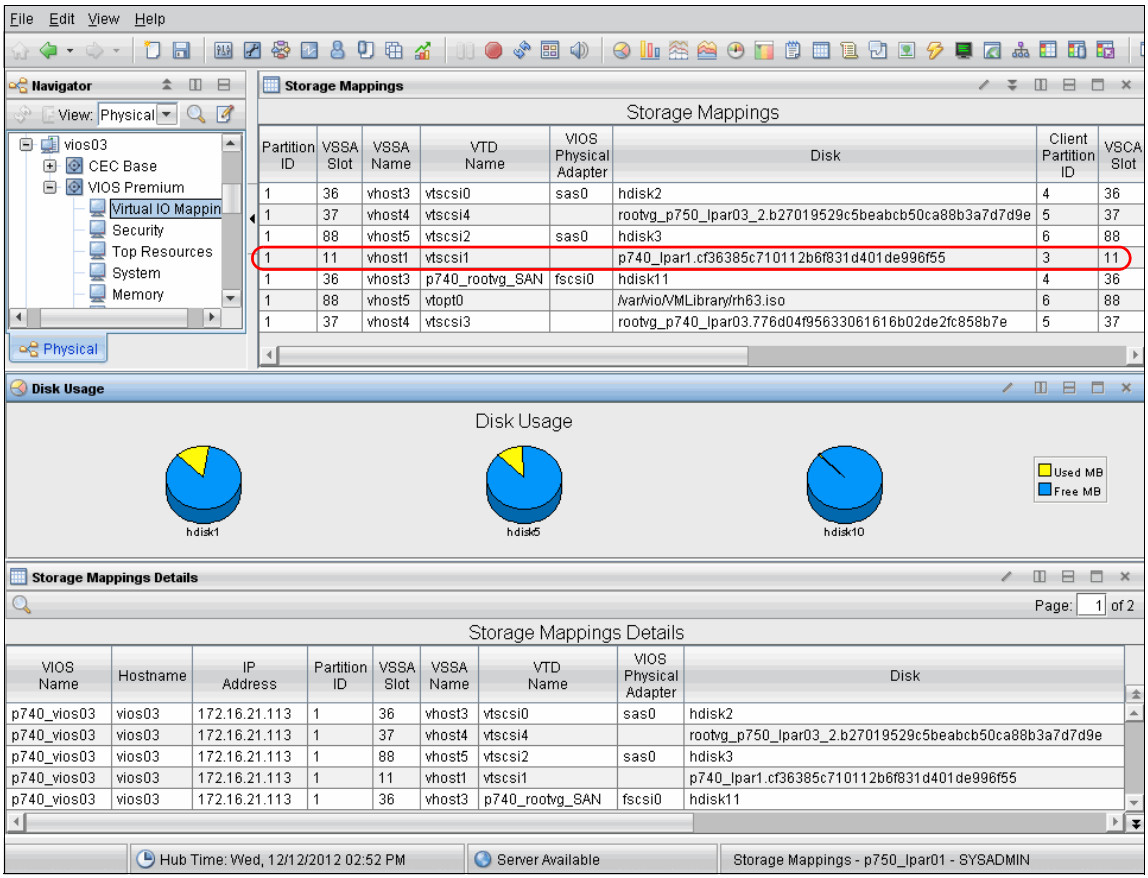


Figure 18-8 Tivoli Monitoring panel showing Storage Mappings

Network mappings (VIOS Premium Agent)

In the Navigator, right-click **Virtual IO Mappings** and then select **Workspace** → **Network Mappings**.

Network Mappings provide you with all information related to virtual network configuration. Figure 18-9 shows the physical and virtual adapters that are configured on the Virtual I/O Server `vios03`.

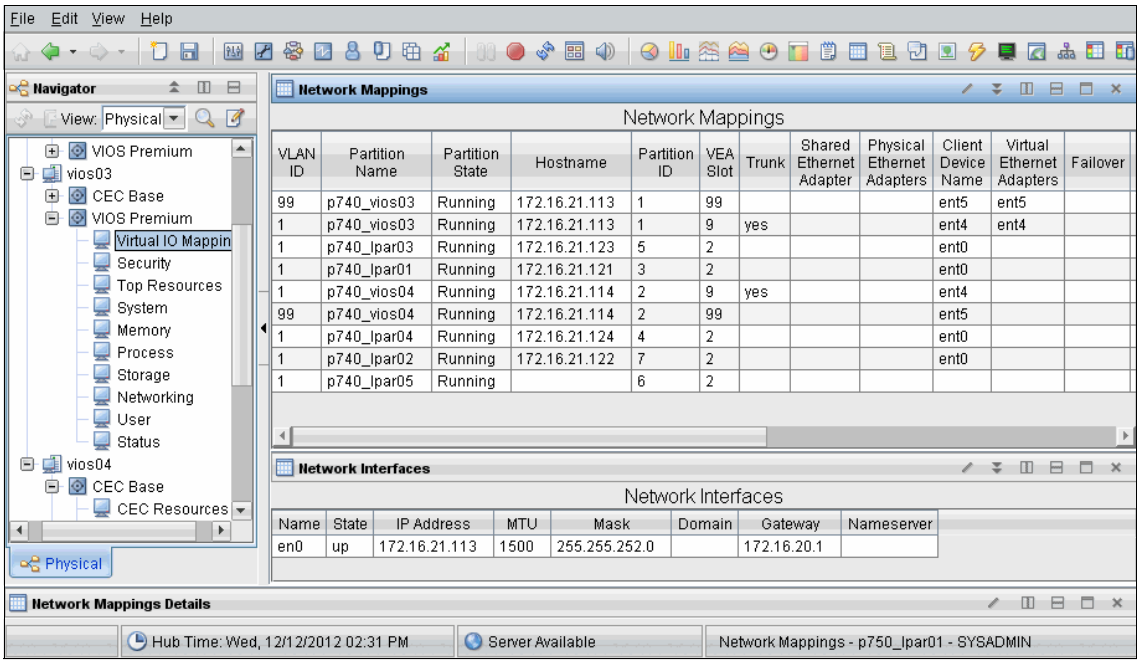


Figure 18-9 Tivoli Monitoring window that shows Network Mappings

There is also an NPIV Mapping workspace available.

Top Resources (VIOS Premium Agent)

In the Navigator, right-click **Top Resources** and then select **Workspace** → **Top Resources Usage**.

The Top Resources Usage window (Figure 18-10) provides you with information that is related to file systems metrics and processes' processor and memory consumption.

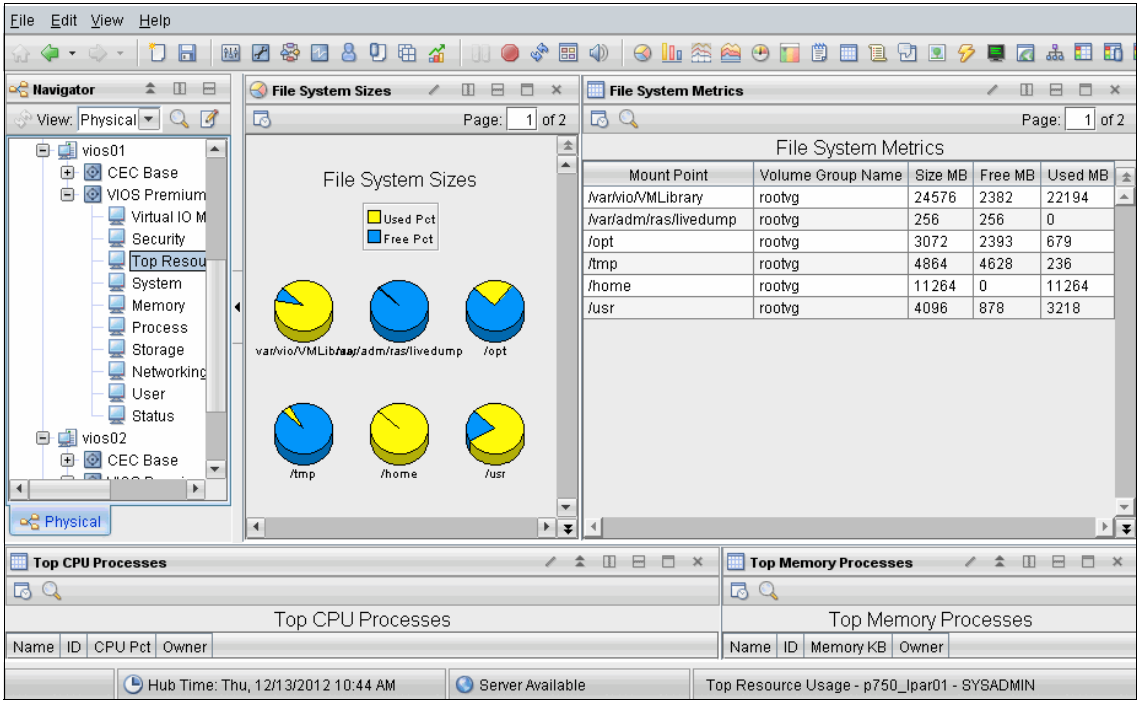


Figure 18-10 Tivoli Monitoring window that shows Top Resources Usage

System (VIOS Premium Agent)

In the Navigator, right-click **System** and then select **Workspace** → **CPU Utilization**.

The CPU Utilization window (Figure 18-11) provides you with real-time information related to user, idle, system, and IO wait CPU percentage utilization.

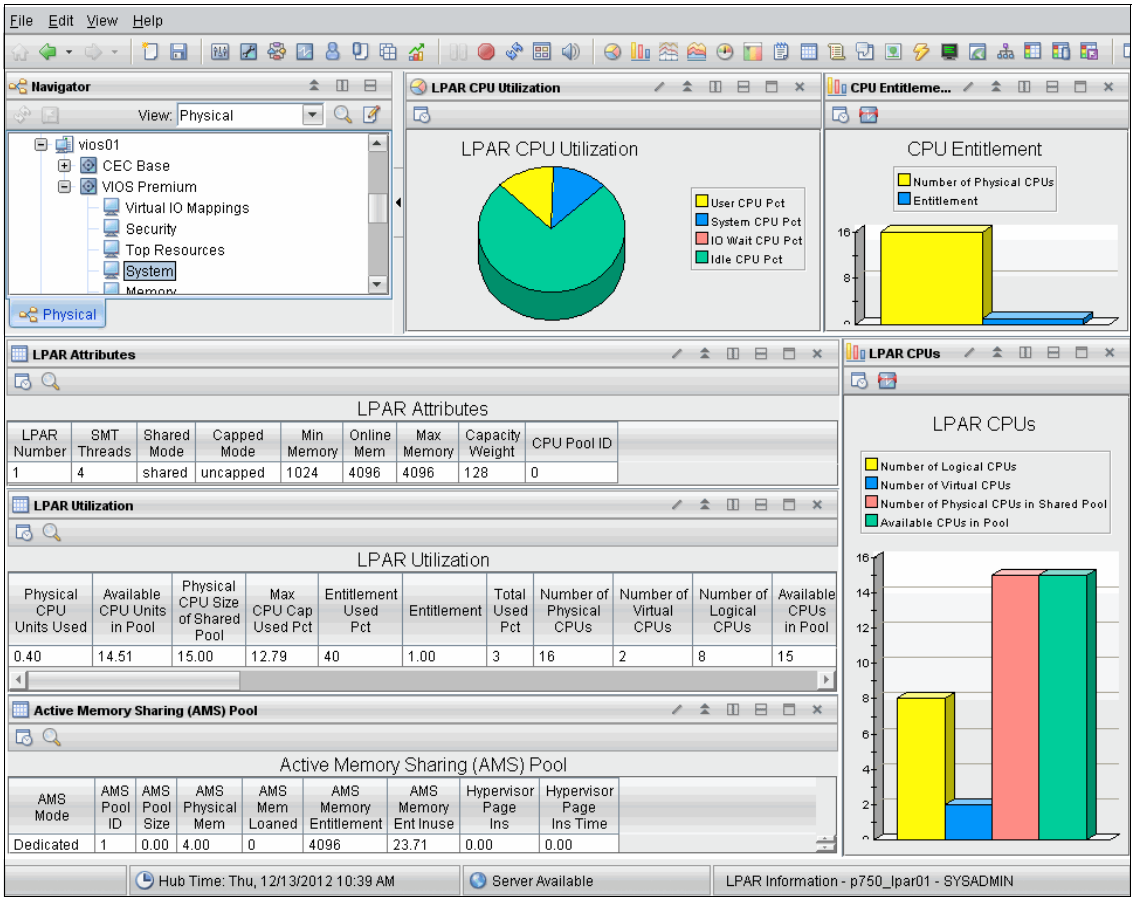


Figure 18-11 Tivoli Monitoring window that shows CPU Utilization

Storage (VIOS Premium Agent)

In the Navigator, right-click **Storage** and then select **Workspace** → **System Storage Information**.

The System Storage Information window (Figure 18-12) provides you with information related to disk activity.

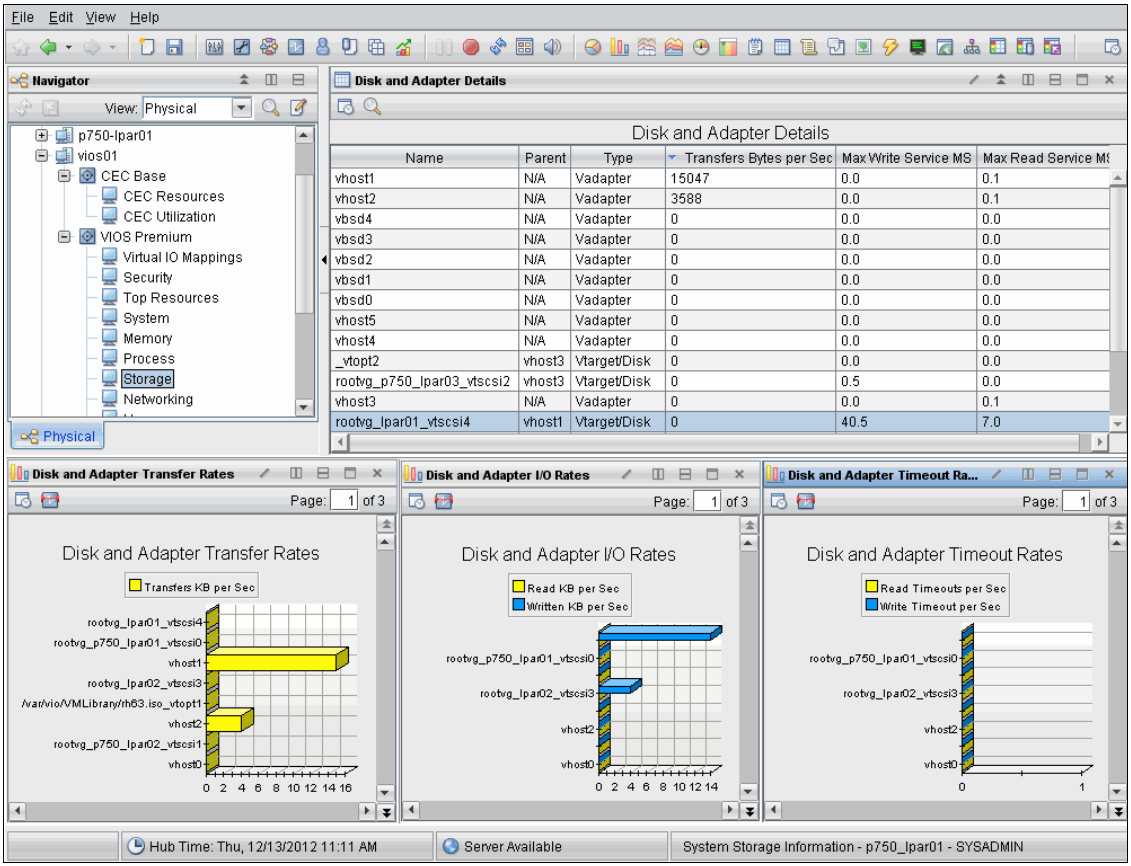


Figure 18-12 Tivoli Monitoring window that shows System Storage Information

A number of other storage workspaces might also be of interest, including File Systems, Logical Volume Details, Physical Storage Performance Details, Physical Volume Details, Virtual Storage Performance Details, Volume Group and Logical Volumes, MPIO Storage Information, and Fibre Channel.

Networking (VIOS Premium Agent)

In the Navigator, right-click **Networking** and then select **Workspace** → **Network Adapter Utilization**.

The Network Adapter Utilization window (Figure 18-13) provides you with information related to network adapter activity.

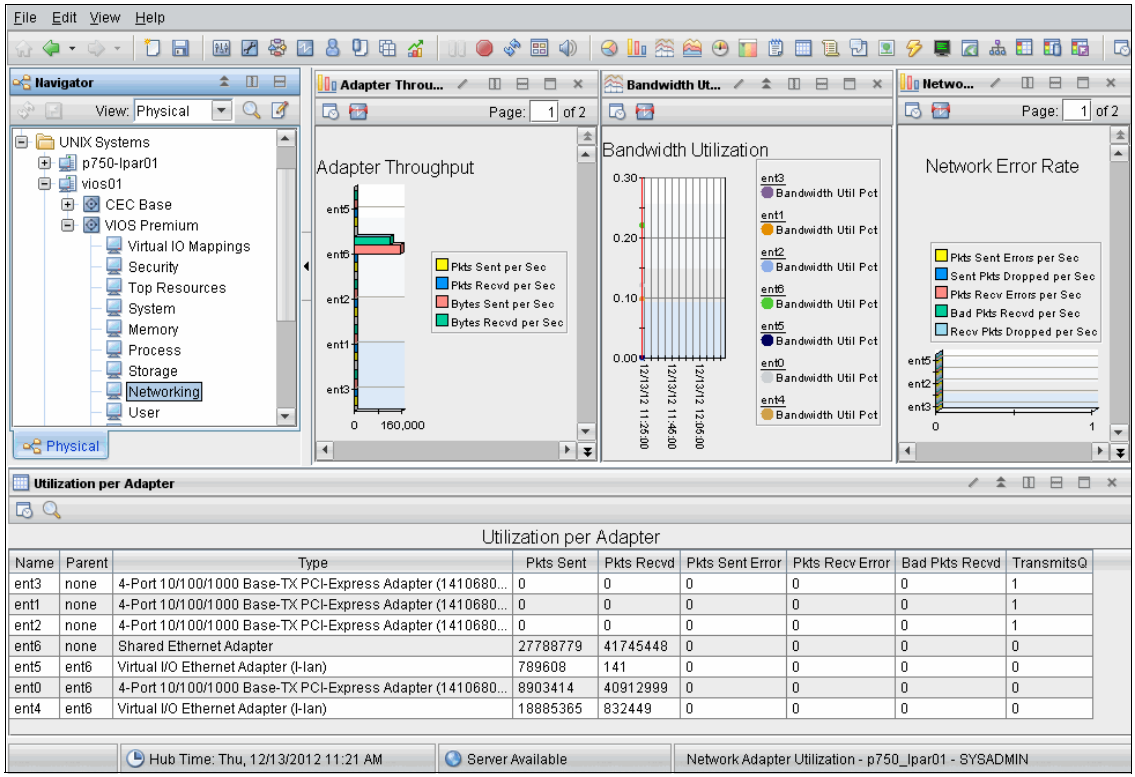


Figure 18-13 Tivoli Monitoring window that shows Network Adapter Utilization

A number of other network workspaces might also be of interest, including Network Adapter Details, Network Protocol View, Shared Ethernet Adapter High Availability Details, Shared Ethernet Bridging Details, Shared Ethernet, and Network Interfaces.

CEC Resources (CEC Base Agent)

In the Navigator, right-click **CEC Base** and then select **Workspace** → **CEC Resource Inventory**.

The CEC Resource Inventory window (Figure 18-14) provides you with an overview of information that is related to CEC resources.

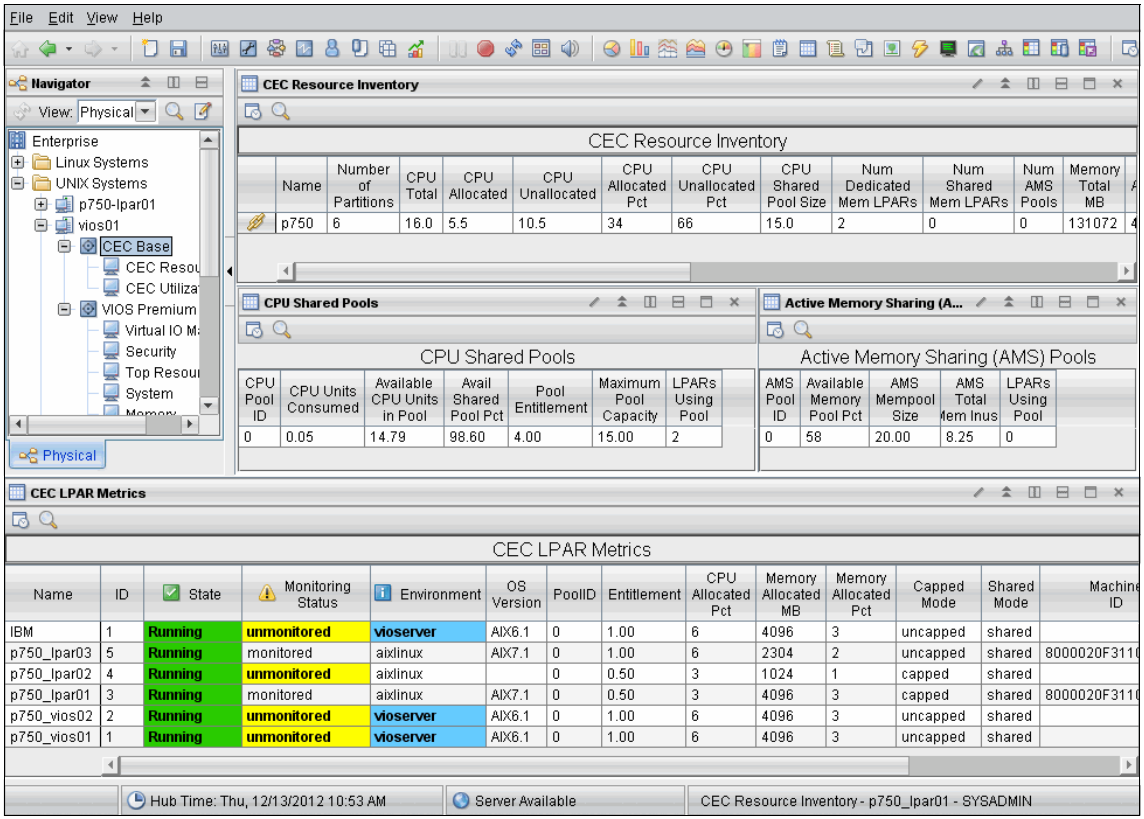


Figure 18-14 CEC Resource Inventory workspace

Note: LPARs that are listed as unmonitored are not running the **xmtopas** or **xmservd** daemons

CEC Utilization (CEC Base)

In the Navigator, right-click **CEC Utilization** and then select **Workspace** → **Active Memory Expansion**.

The Active Memory Expansion workspace (Figure 18-15) provides statistics on AME on the selected VIOS.

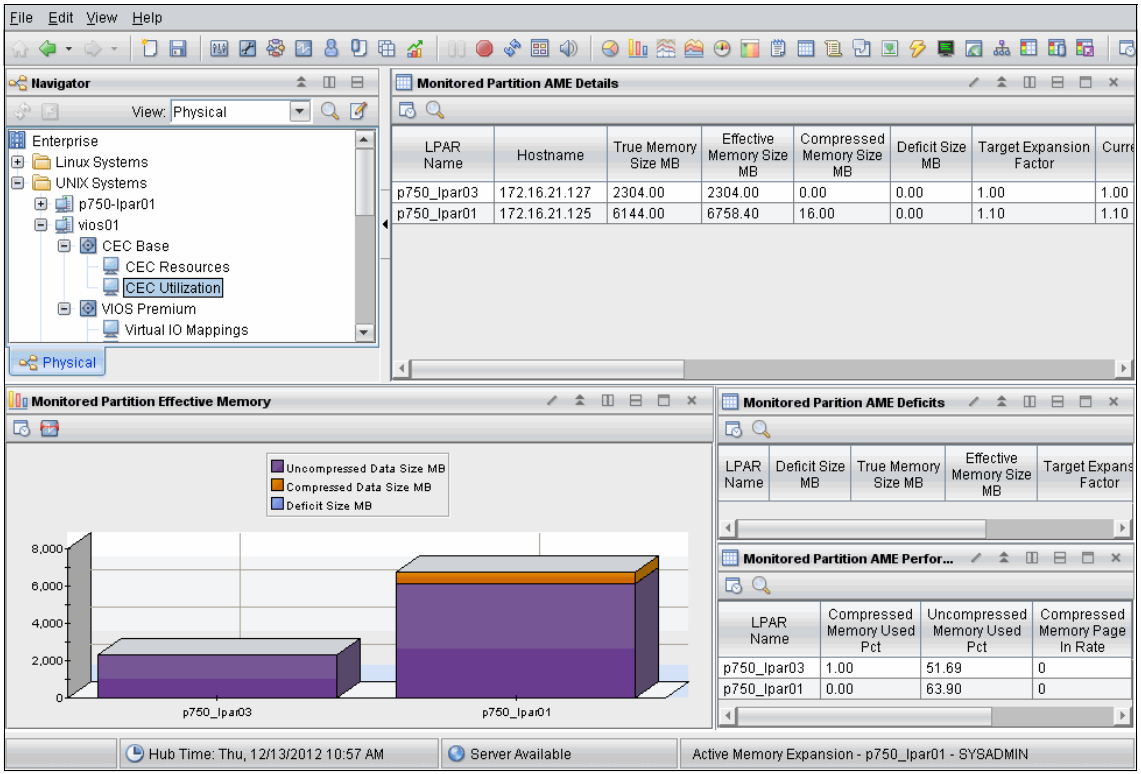


Figure 18-15 Tivoli Monitoring Active Memory Expansion workspace

Creating and modifying Tivoli Monitoring situations

In IBM Tivoli Monitoring, a situation is defined as a set of conditions that are measured according to criteria and evaluated to be true or false. For example, a file system with utilization greater than 95% or a network interface with transmission errors greater than 0. An action, such as running a command or sending a message, can be associated with situation. The action is automatically performed when the situation becomes true.

Tivoli Monitoring comes with a comprehensive set of situations and actions preconfigured. You can modify existing situations, or create new situations by using the Situation Editor as follows:

1. Select **Edit** → **Situation Editor** from the IBM Tivoli Monitoring Desktop Client, or use the Ctrl+E shortcut. The Situation editor launches and displays a list of available agents in the left pane, and some user assistance is displayed in the right pane, as shown in Figure 18-16.

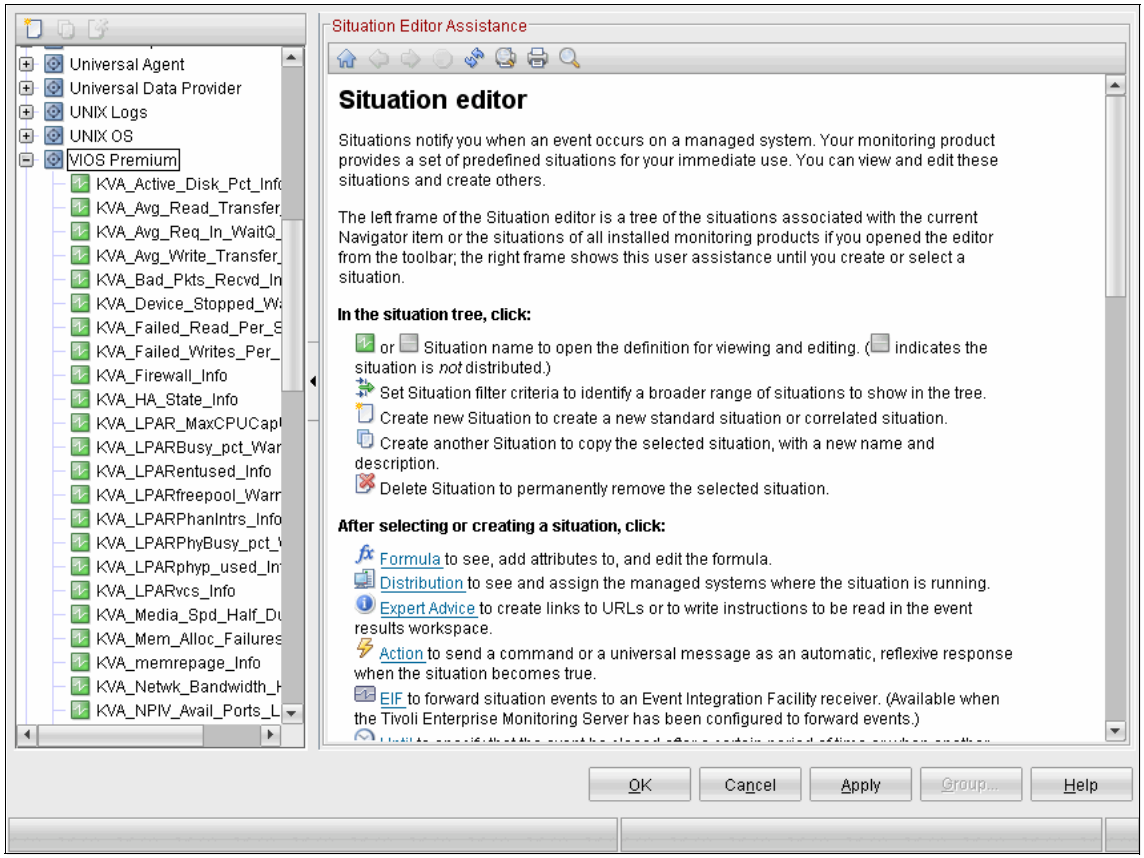


Figure 18-16 Assistance window for Tivoli Monitoring Situation editor

2. You can then change the threshold at which a situation becomes true by using the Formula tab as shown in Figure 18-17. You can use the Distribution tab to assign the situation to specific managed systems. The Expert Advice and Action tabs can be used to describe, or automatically perform, tasks that will resolve the situation.

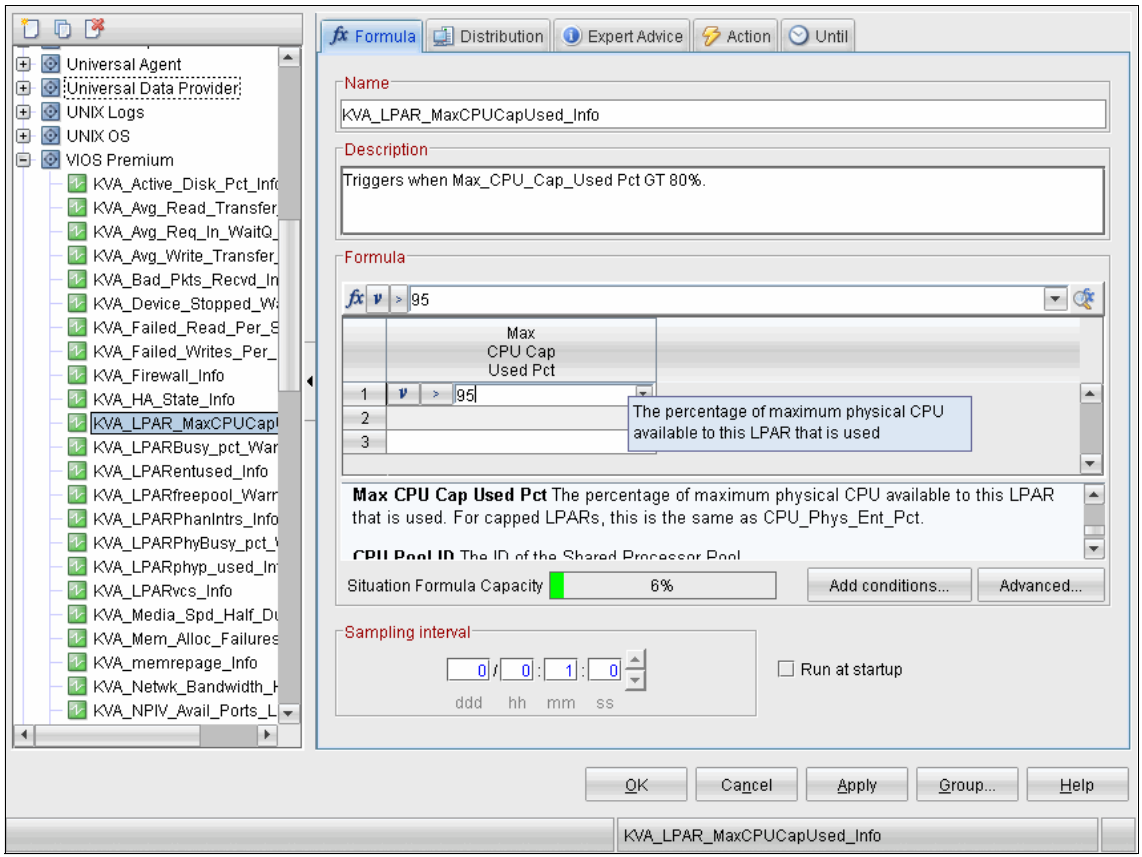


Figure 18-17 Changing the trigger threshold for a situation

When situations are triggered, the alert appears in the relevant Tivoli Enterprise Portal workspaces, as shown in Figure 18-18.

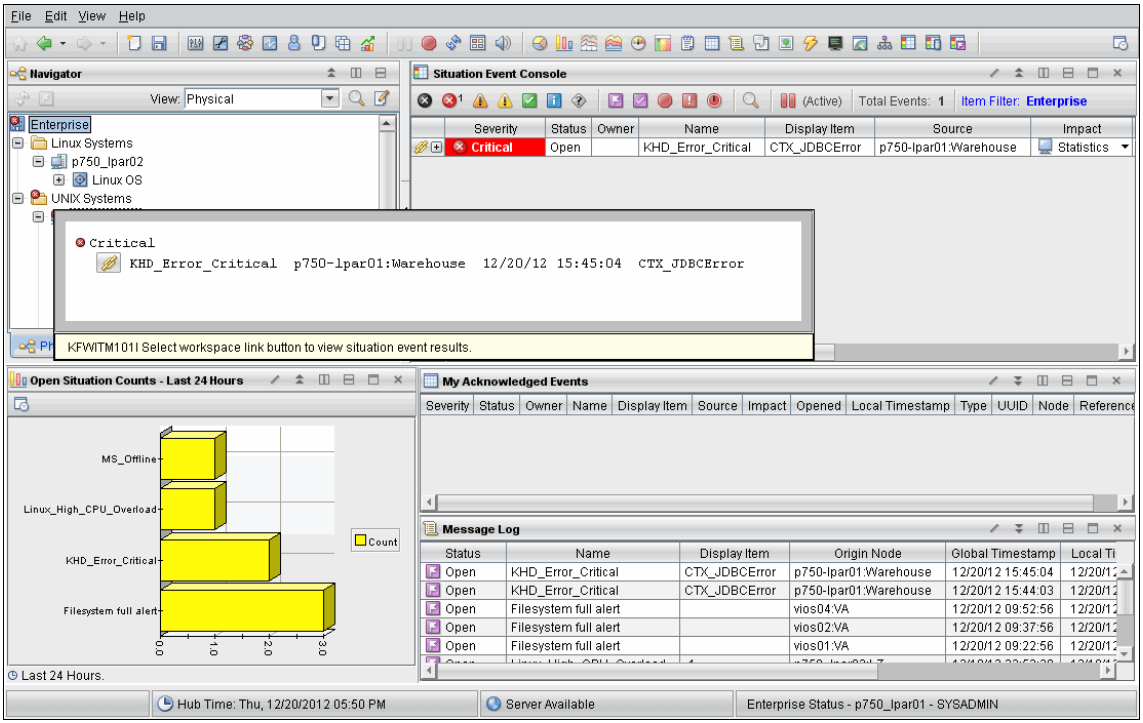


Figure 18-18 Tivoli Enterprise Portal workspace showing active situations

Hovering the mouse over the alert icons in the Tivoli Enterprise Portal Navigator displays more information. Clicking the chain icon displays a workspace with **Take Action** and **Expert Advice** panes. These panes can provide a shortcut to a command or script to resolve the situation, or provide information about how to resolve the situation.

Tivoli Data Warehouse

Historical data collection can be enabled in IBM Tivoli Monitoring, which stores long term monitoring agent data in the Tivoli Data Warehouse.

For more information about the initial setup of Tivoli Data Warehouse, see the IBM Tivoli Monitoring Information Center at:

http://publib.boulder.ibm.com/infocenter/tivihelp/v15r1/topic/com.ibm.itm.doc_6.2.2fp1/history_manage_intro.htm

After the Data Warehouse is created, and the Warehouse Proxy and Summarization and Pruning agents started, select which agents and attributes you want to collect historical data for using these steps:

1. Select **Edit** → **History Configuration** from the Tivoli Enterprise Portal Desktop Client, or use the Ctrl+H shortcut.
2. Specify the summarization and pruning controls as shown in Figure 18-19. Choose sensible retention periods for your historical data. Retaining too much data requires large amounts of disk space for your data warehouse.

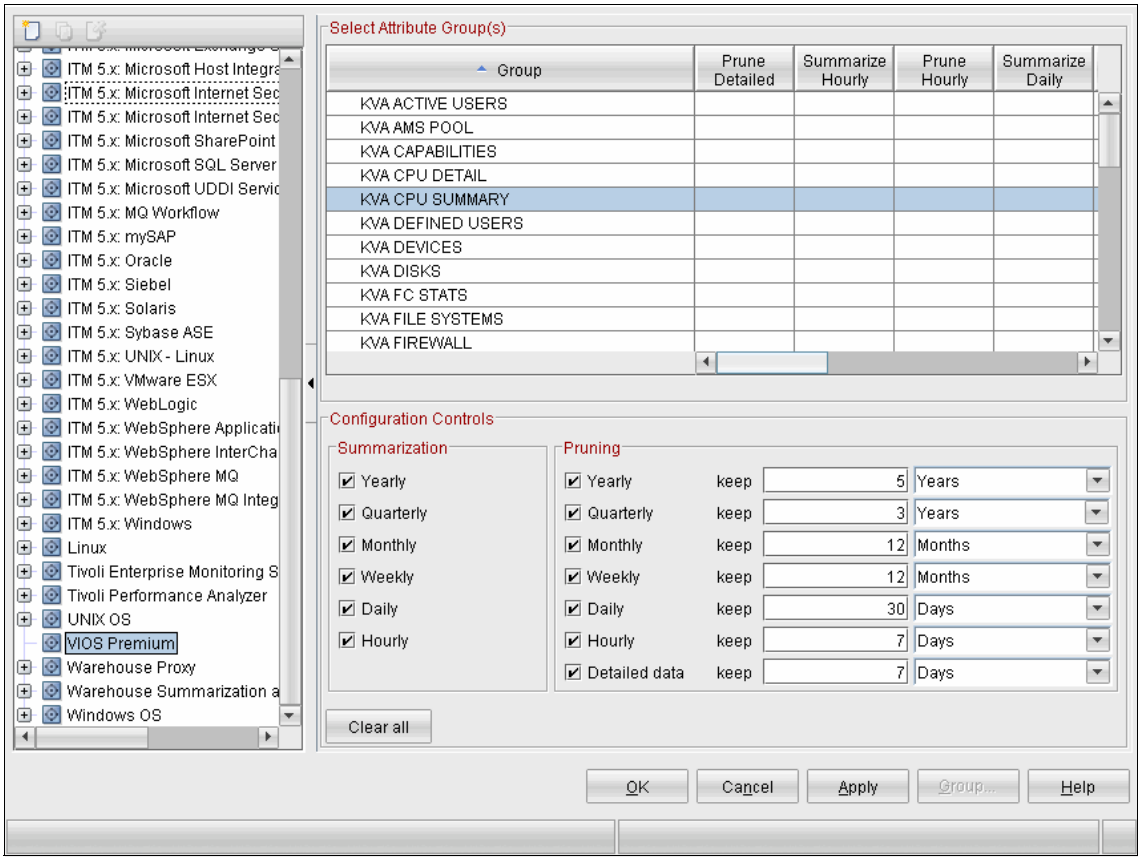


Figure 18-19 History Configuration summarization and pruning controls window

IBM provides a load projection spreadsheet that can be used to accurately estimate the required size of your data warehouse at:

https://www-304.ibm.com/software/brandcatalog/ismlibrary/details?catalog.label=1TW10TM1Y

Tip: You can select multiple attribute groups, and apply summarization and pruning controls to all of them at once.

3. Specify which attribute groups you want to collect data for, and create history collections for them, as shown in Figure 18-20. This window is opened by clicking the **Create new collection setting** icon in the upper left of the History configuration window.

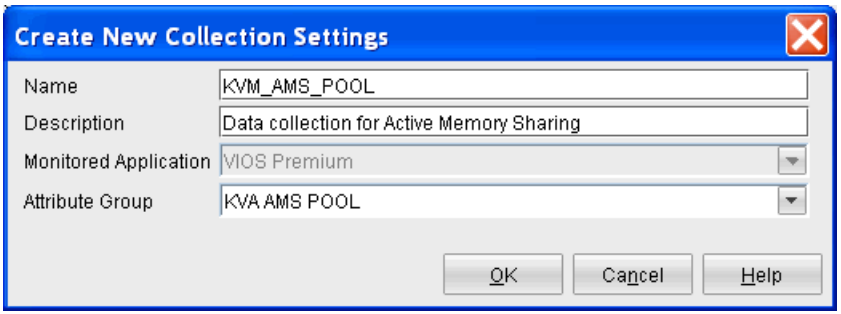


Figure 18-20 Create New History Collection window

4. Define the collection parameters on the Basics tab, as shown in Figure 18-21.

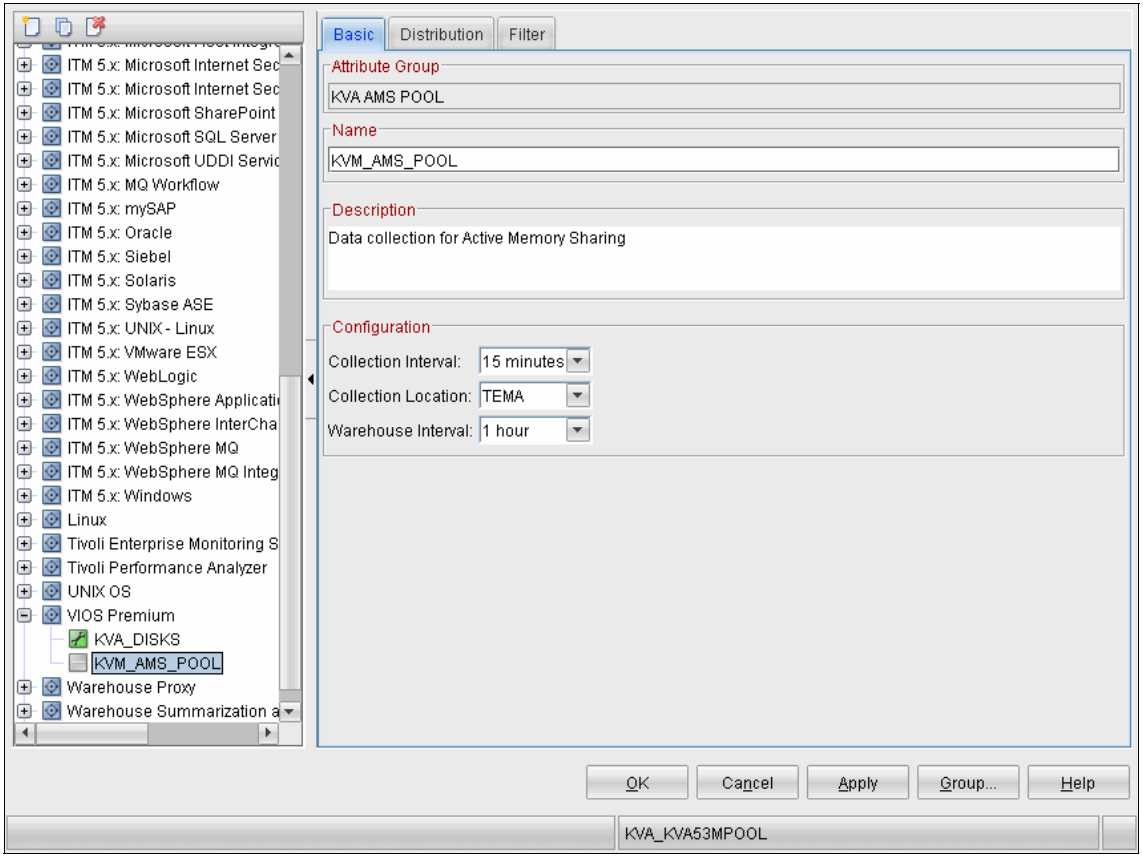


Figure 18-21 Basic information for a history collection

5. Define the nodes for the collection to be applied to in the Distribution tab, as in Figure 18-22.

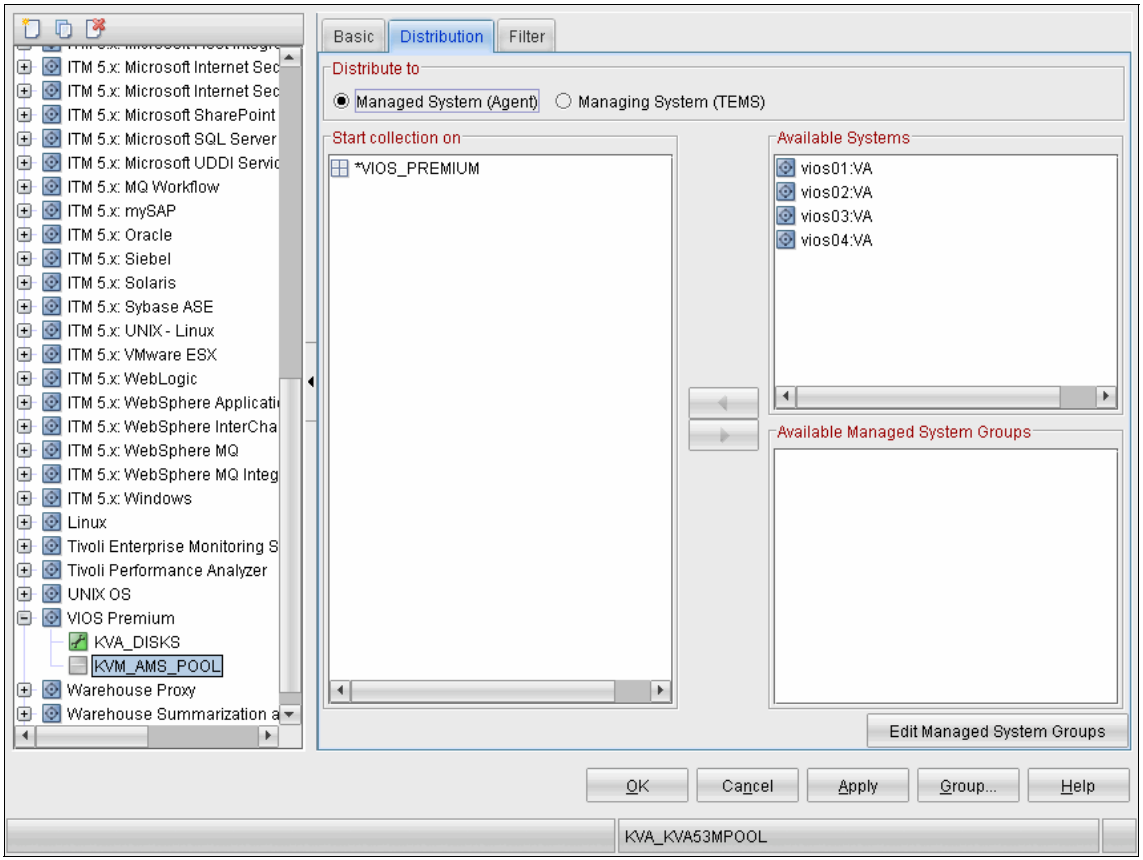


Figure 18-22 Specifying which managed systems to apply history collection to

6. Click **Apply**. The icon next to the Collection name turns to a green wrench when data collection is active. Data collection is not active until the collection is distributed to at least one agent. You can specify individual managed systems or customizable managed system groups.

An icon of a running person is shown next to attribute groups that have active data collections that are associated with them as shown in Figure 18-23.

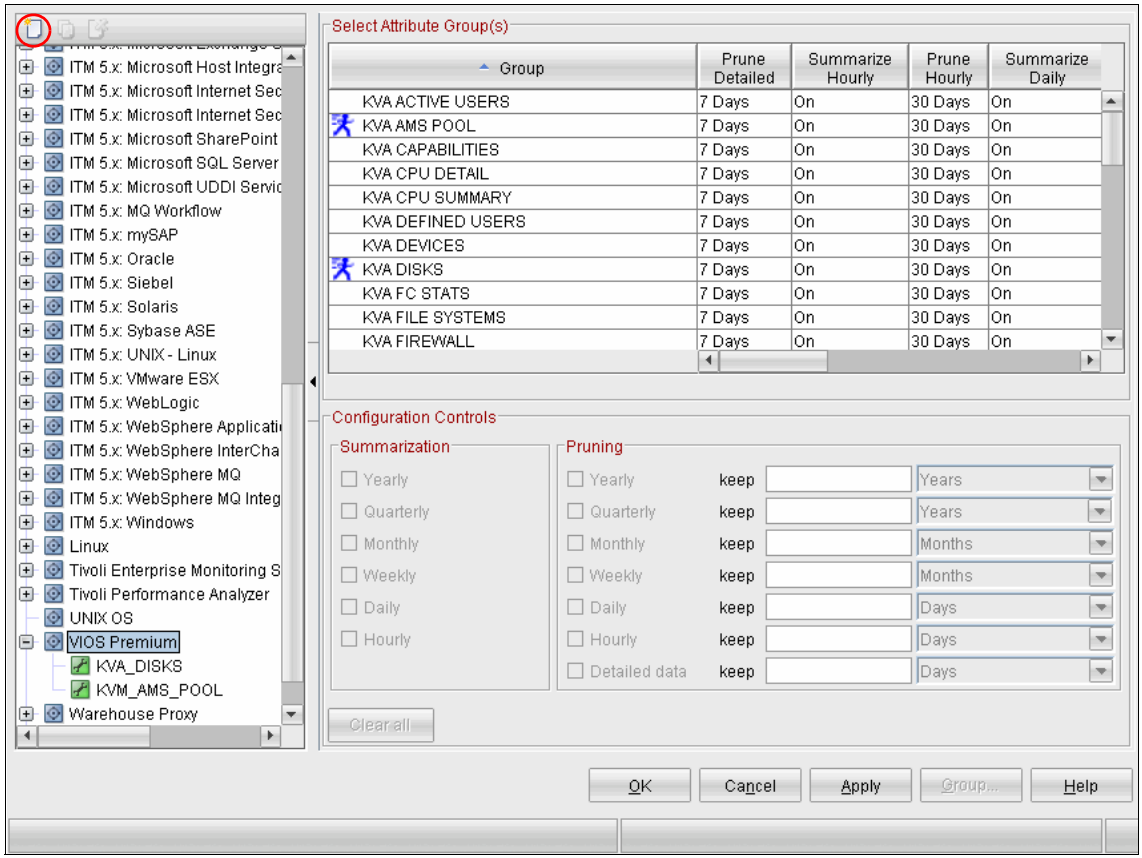


Figure 18-23 History configuration window that shows two active collections

Note: You can also use the Tivoli Monitoring command-line tools **`tacmd histCreateCollection`** and **`tacmd histStartCollection`** to create and activate history collections.

7. After data warehousing is enabled, you can specify longer time periods in charts and tables in Tivoli Enterprise Portal workspaces. To do this, first select the chart or table you want to display historical data for. Then, click the icon in the upper left of the pane that has the small clock on it, as shown in Figure 18-24.

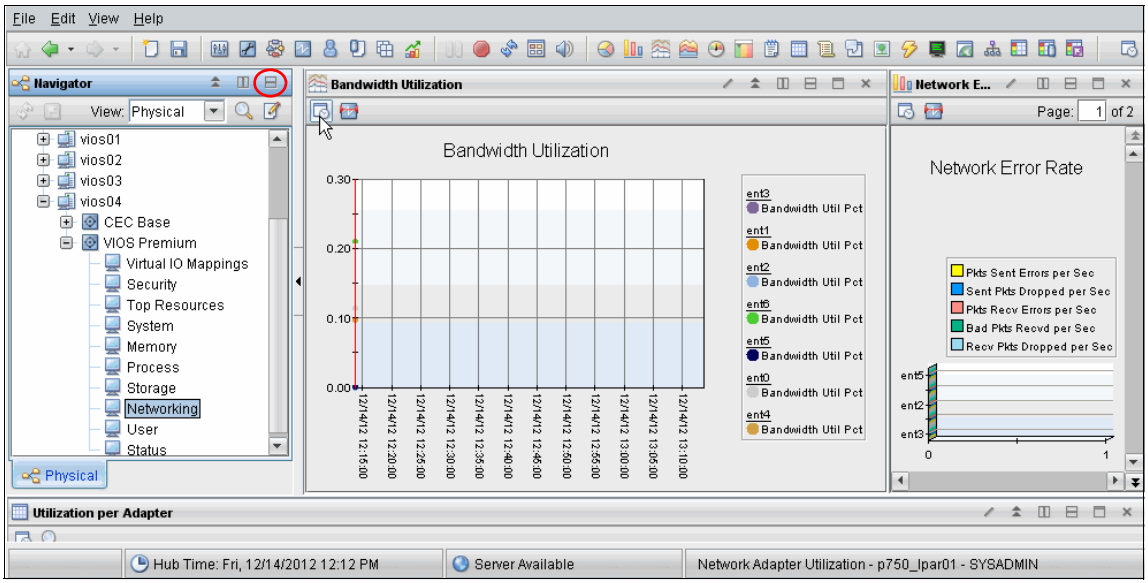


Figure 18-24 Displaying historical data for a chart by clicking the Clock icon

8. You can then select the time-frame for the historical data you want to display. Several predefined options are available, plus custom options, as shown in Figure 18-25.

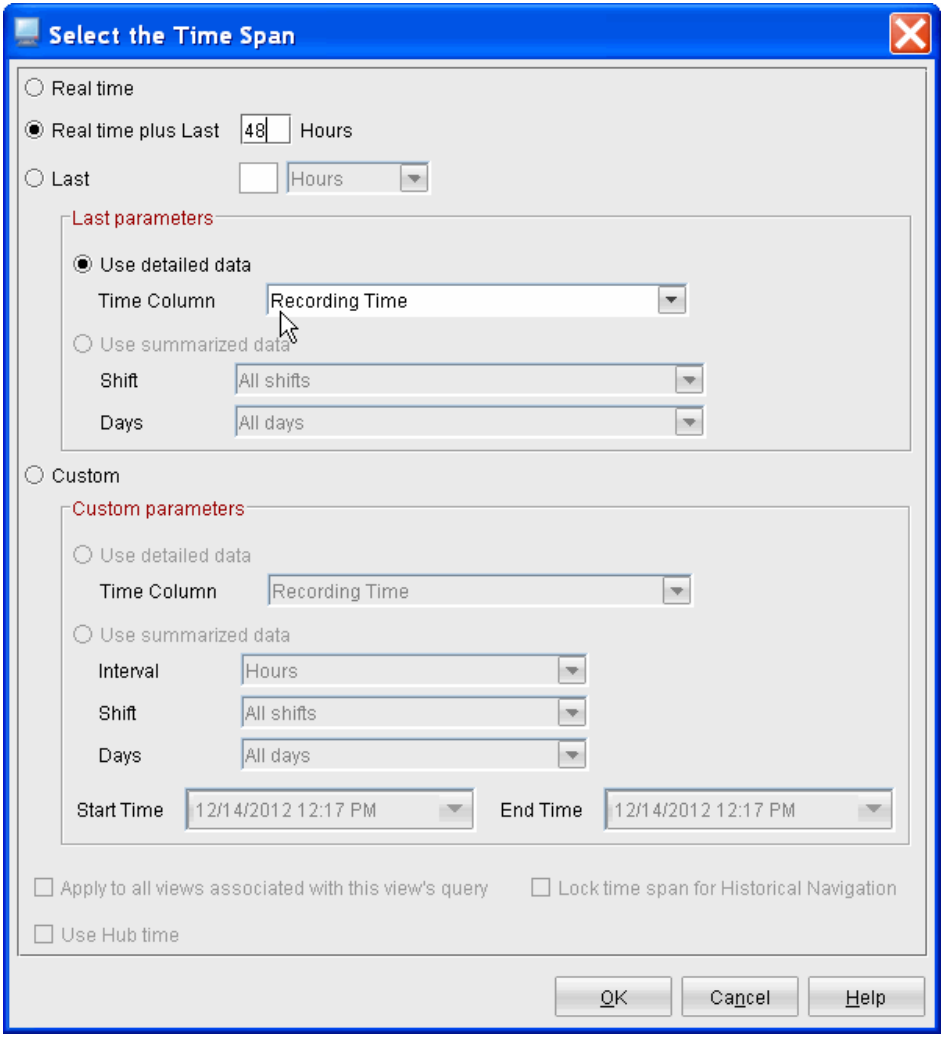


Figure 18-25 Time span selection window for historical data

9. Click **OK** when done, and the chart is updated with the historical data for the selected time span, as shown in Figure 18-26.

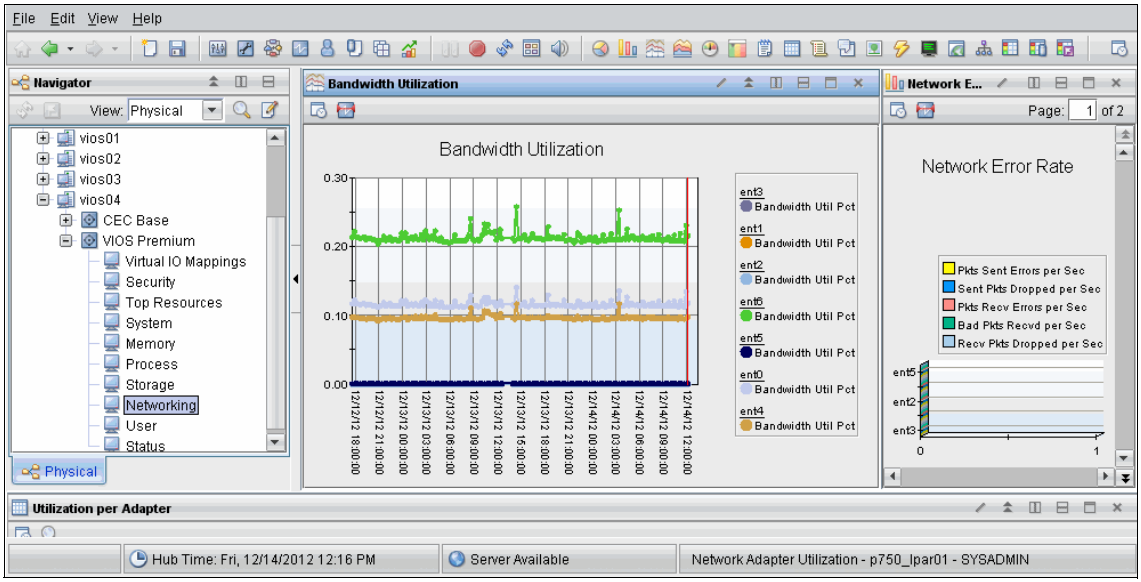


Figure 18-26 Historical data displayed for the bandwidth utilization chart

Tivoli Common Reporting

Tivoli Common Reporting provides a convenient way to run predefined or custom reports on any data metric that is stored in the Tivoli Data Warehouse. Tivoli Common Reporting can be used to gather, analyze, and report important trends in your managed environment.

For installation instructions, see *Chapter 8 Tivoli Common Reporting for the System p monitoring agents* in the VIOS Tivoli Monitoring agent documentation at:

`http://pic.dhe.ibm.com/infocenter/tivihelp/v15r1/topic/com.ibm.itm.doc_6.2.3/pviosagent6222_user.htm`

After Tivoli Common Reporting is installed, you can run any of the predefined reports as shown in Figure 18-27. You can also write your own custom reports using the Report Studio, and schedule reports to run at pre-determined times.

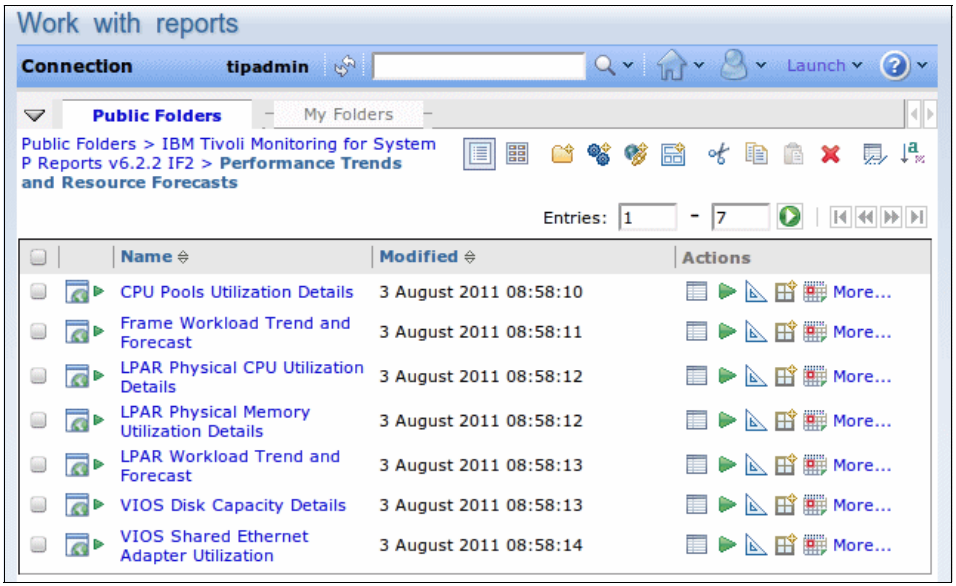


Figure 18-27 Examples of available Tivoli Monitoring for System p reports

When you run a report, you can choose several report output formats, including HTML (Figure 18-28), PDF, and CSV.

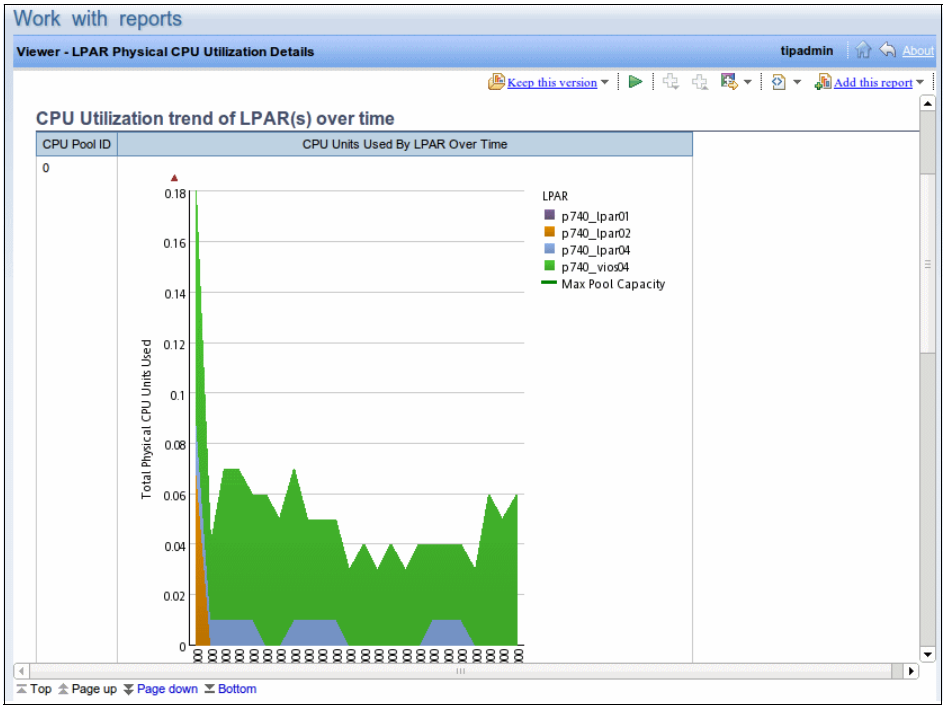


Figure 18-28 Sample output for a System p Tivoli Common Reporting report

Forwarding events to enterprise event consoles

IBM Tivoli Monitoring can be configured to forward situations to other enterprise level event consoles, including Tivoli Netcool/OMNIbus, and Tivoli Event Console. For more information, see the *Integrating event management systems* section of the IBM Tivoli Monitoring installation guide at:

`http://pic.dhe.ibm.com/infocenter/tivihelp/v15r1/index.jsp?topic=%2Fcom.ibm.itm.doc_6.2.3%2Fitm623_install876.htm`

After Tivoli Monitoring situations are being forwarded to your event console, you can customize a view of the event console to display all alerts forwarded from Tivoli Monitoring and events relevant to your managed systems generated from other sources.

Figure 18-29 shows a Netcool/OMNIbus filter that is written to display alerts forwarded to the event console by the Tivoli Monitoring Premium AIX agent (PX), Tivoli Monitoring CEC Base Agent (PK), Tivoli Monitoring VIOS agent (VA), and Tivoli Monitoring HMC agent (PH). The filter also displays events from other sources where the host name of the system matches the string ‘aix’.

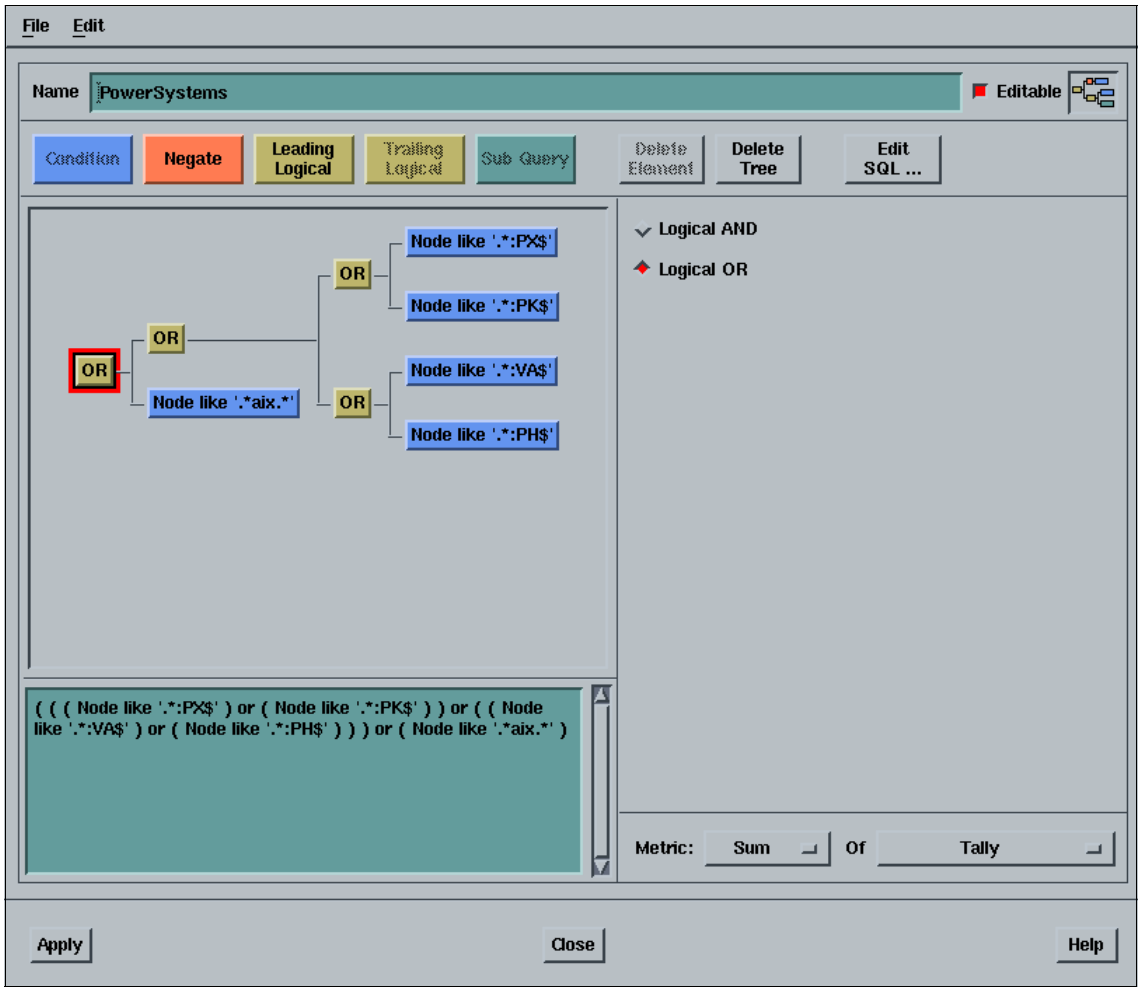


Figure 18-29 Creating a Netcool/OMNIbus filter to display Tivoli Monitoring forwarded events

Figure 18-30 shows the Netcool/OMNIbus event list showing events that match this filter. The events include syslog messages, SNMP traps, IBM Netcool® ping probe, and other monitoring tools. SNMP traps also are displayed here, if they are configured to be sent to the event manager.

Tip: Configure syslog messages to be sent direct to your event console by modifying */etc/syslog.conf* to send messages to a host running the Netcool/OMNIbus syslog probe.

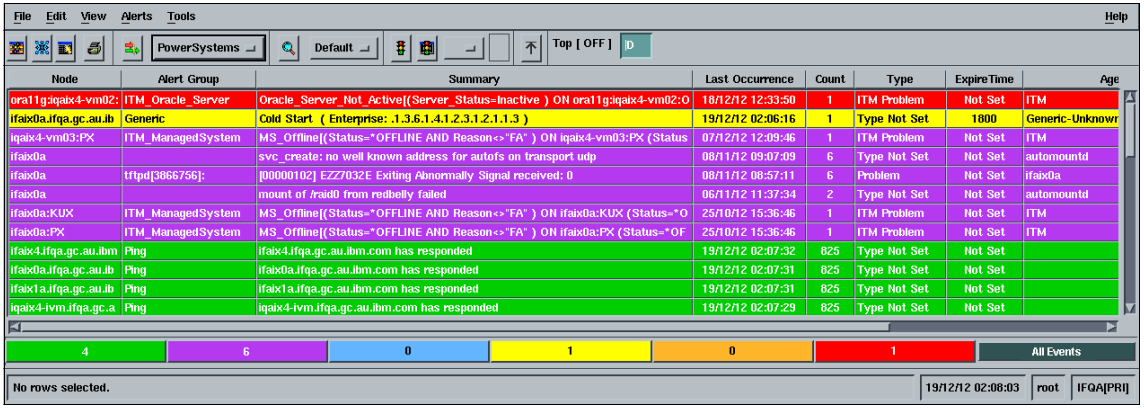


Figure 18-30 Netcool/OMNIbus event list showing forwarded Tivoli Monitoring situations

Part 7

Appendixes

This part contains additional reading material that may be helpful to you to better understand the content of this publication.

A

AIX disk and NIB network checking and recovery script

When using LVM mirroring between disks from two Virtual I/O Servers, rebooting one Virtual I/O Server changes the disk state to *missing* and requires the resynchronization of stale partitions by using the **varyonvg** command when the Virtual I/O Server is finished booting.

When using NIB for network redundancy, the backup adapter does not return the active channel to the primary adapter until the backup adapter fails. This is true for partitions that use virtual adapters and AIX V5.3-ML03 or later. Setting `Automatically Recover to Main Channel` to `Yes` does not resolve this issue. This is because a link-up event is not received during reactivation of the path because virtual Ethernet adapters are always up. If you configure NIB to do load balancing, you might want the NIB to be on the primary channel.

If the settings are correct, MPIO does not require any special attention when a Virtual I/O Server is restarted, but check a failed path anyway.

Checking and fixing these things in a managed system with many partitions is time-consuming and prone to errors.

"Listing of the fixdualvios.ksh script" on page 703 is a sample script that you can be tailor to your needs. The script checks for the following settings and fixes the configuration accordingly:

- Redundancy with dual Virtual I/O Servers and LVM mirroring
- Redundancy with dual Virtual I/O Servers, Fibre Channel SAN disks, and AIX MPIO
- Network redundancy using Network Interface Backup

Place the script locally on each partition. You can customize the script to account for any configuration uniqueness in the partition. You can also schedule the script to run at regular intervals by using **cron**. If you are using **dsh** (distributed shell), locate the script in the same directory on each partition.

Distributed shell, **dsh**, can be used to run the script on all required target partitions in parallel from a NIM or admin server.

Considerations:

- Before AIX 7.1, the **dsh** command is installed by default in AIX and is part of CSM. Use of the full function clustering that is offered by CSM requires a license. For **dsh** command information, see the AIX documentation.
- In AIX 7.1, Distributed Systems Management (DSM) replaces the Cluster Systems Management package (CSM). Commands such as **dcp** and **dsh** are not available without installing the DSM package. DSM is not installed by default, but is on the base installation media. The DSM package is in the file sets `dsm.core` and `dsm.dsh`.
- You can use **dsh** based on **rsh**, **ssh**, or Kerberos authentication if **dsh** can run commands without being prompted for a password.

Example 18-25 shows how to run `fixdualvios.ksh` in parallel on the partitions `dbserver`, `appserver`, and `nim`.

Example 18-25 Using a script to update partitions

```
# dsh -n dbserver,appserver,nim /tmp/fixdualvios.ksh | dshbak >\
/tmp/fixdualvios.out
```

Tips:

- Use the **DSH_LIST=<file listing lpars>** variable so you do not have to type in the names of the target partitions when using **dsh**.
- Use the **DSH_REMOTE_CMD=/usr/bin/ssh** variable if you use **ssh** for authentication.
- The output file /tmp/fixdualvios.out is on the system that runs the **dsh** command.

The **dshbak** command will group the output from each partition.

Example 18-26 shows how to run the script and the output listing from the sample partitions named dbserver, appserver, and nim.

Example 18-26 Running the script and listing output

```
# export DSH_REMOTE_CMD=/usr/bin/ssh
# export DSH_LIST=/root/nodes
# dsh /tmp/fixdualvios.ksh|dshbak > /tmp/fixdualvios.out

HOST: appserver
---------------
1 Checking if Redundancy with dual VIO Server and LVM mirroring is being used.
Redundancy with dual VIO Server and LVM mirroring is NOT used.
No disk has missing status in any volumegroup.

2 Checking if Redundancy with dual VIO Server, Fibre Channel SAN disks and AIX
MPIO is being used.
Status:
Enabled hdisk0 vscsi0
Enabled hdisk0 vscsi1
hdisk1 has vscsi0 with Failed status. Enabling path.
paths Changed
New status:
Enabled hdisk1 vscsi0
Enabled hdisk1 vscsi1

3 Checking if Network redundancy using Network interface backup is being used.
EtherChannel en2 is found.
Backup channel is being used. Switching back to primary.
Active channel: primary adapter

HOST: dbserver
--------------
1 Checking if Redundancy with dual VIO Server and LVM mirroring is being used.
Redundancy with dual VIO Server and LVM mirroring is NOT used.
No disk has missing status in any volumegroup.
```

```
2 Checking if Redundancy with dual VIO Server, Fibre Channel SAN disks and AIX
MPIO is being used.
hdisk0 has vscsi0 with Failed status. Enabling path.
paths Changed
New status:
Enabled hdisk0 vscsi0
Enabled hdisk0 vscsi1

3 Checking if Network redundancy using Network interface backup is being used.
EtherChannel en2 is found.
Backup channel is being used. Switching back to primary.
Active channel: primary adapter

HOST: nim
---------
1 Checking if Redundancy with dual VIO Server and LVM mirroring is being used.
Redundancy with dual VIO Server and LVM mirroring is being used.
Checking status.
No disk in rootvg has missing status.
No disk has missing status in any volumegroup.

2 Checking if Redundancy with dual VIO Server, Fibre Channel SAN disks and AIX
MPIO is being used.
Redundancy with dual VIO Server, Fibre Channel SAN disks and AIX MPIO
is NOT used.

3 Checking if Network redundancy using Network interface backup is being used.
EtherChannel en2 is found.
Backup channel is being used. Switching back to primary.
Active channel: primary adapter
```

Remember: The reason for the Failed status of the paths is that the **hcheck_interval** parameter was not set on the disks.

This script assumes that the system is using one or more of the following configurations:

1. Redundancy with dual VIO Server and LVM mirroring
2. Redundancy with dual VIO Server, Fibre Channel SAN disks, and AIX MPIO
3. Network redundancy using “Network interface backup”

Listing of the fixdualvios.ksh script

```
#!/bin/ksh
#set -x
#
# This script will check and restore the dual VIO Server
# configuration for partitions served from two VIO Servers after
# one VIO Server has been unavailable.
# The script must be tailored and TESTED to your needs.
#
# Disclaimer
# IBM DOES NOT WARRANT OR REPRESENT THAT THE CODE PROVIDED IS COMPLETE OR UP-TO-DATE.  IBM DOES
NOT WARRANT, REPRESENT OR IMPLY RELIABILITY, SERVICEABILITY OR FUNCTION OF THE CODE.  IBM IS
UNDER NO OBLIGATION TO UPDATE CONTENT NOR PROVIDE FURTHER SUPPORT.

# ALL CODE IS PROVIDED "AS IS," WITH NO WARRANTIES OR GUARANTEES WHATSOEVER.  IBM EXPRESSLY
DISCLAIMS TO THE FULLEST EXTENT PERMITTED BY LAW ALL EXPRESS, IMPLIED,  STATUTORY AND OTHER
WARRANTIES, GUARANTEES, OR REPRESENTATIONS, INCLUDING, WITHOUT LIMITATION, THE WARRANTIES OF
MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NON-INFRINGEMENT OF PROPRIETARY AND
INTELLECTUAL PROPERTY RIGHTS.  YOU UNDERSTAND AND AGREE THAT YOU USE THESE MATERIALS,
INFORMATION, PRODUCTS, SOFTWARE, PROGRAMS, AND SERVICES, AT YOUR OWN DISCRETION AND RISK AND
THAT YOU WILL BE SOLELY RESPONSIBLE FOR ANY DAMAGES THAT MAY RESULT, INCLUDING LOSS OF DATA OR
DAMAGE TO YOUR COMPUTER SYSTEM.

# IN NO EVENT WILL IBM BE LIABLE TO ANY PARTY FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL,
EXEMPLARY OR CONSEQUENTIAL DAMAGES OF ANY TYPE WHATSOEVER RELATED TO OR ARISING FROM USE OF THE
CODE FOUND HEREIN, WITHOUT LIMITATION, ANY LOST PROFITS, BUSINESS INTERRUPTION, LOST SAVINGS,
LOSS OF PROGRAMS OR OTHER DATA, EVEN IF IBM IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH
DAMAGES. THIS EXCLUSION AND WAIVER OF LIABILITY APPLIES TO ALL CAUSES OF ACTION, WHETHER BASED
ON CONTRACT, WARRANTY, TORT OR ANY OTHER LEGAL THEORIES.
#
# Assuming that the configuration may be using one or more of:
#   1 Redundancy with dual VIO Server and LVM mirroring.
#   2 Redundancy with dual VIO Server, Fiber Channel SAN disks and
#     AIX MPIO.
#   3 Network redundancy using "Network interface backup".
#
# Syntax: fixdualvio.ksh
#
#

#   1 Redundancy with dual VIO Server and LVM mirroring.
#
echo 1 Checking if "Redundancy with dual VIO Server and LVM mirroring" is being used.

# Check if / (hd4) has 2 copies
MIRROR=`lslv hd4|grep COPIES|awk '{print $2}'`
if [ $MIRROR -gt 1 ]
```

```
then
        # rootvg is most likely mirrored
        echo "Redundancy with dual VIO Server and LVM mirroring" is being used.
        echo Checking status.
        # Find disk in rootvg with missing status
        MISSING=`lsvg -p rootvg|grep missing|awk '{print $1}'`
        if [ "$MISSING" = "" ]
        then
                echo No disk in rootvg has missing status.
        else
                echo $MISSING has missing status.
#
# Restore active status and sync of rootvg
                echo Fixing rootvg.
                varyonvg rootvg
               syncvg -v rootvg

        fi
else
        echo "Redundancy with dual VIO Server and LVM mirroring" is NOT used.
fi
# Check now if ANY disk has missing status.
        ANYMISSING=`lsvg -o|lsvg -ip|grep missing|awk '{print $1}'`
        if [ "$ANYMISSING" = "" ]
        then
                echo No disk has missing status in any volumegroup.
        else
                echo $ANYMISSING has missing status. CHECK CAUSE!
        fi

#   2 Redundancy with dual VIO Server, Fiber Channel SAN disks and
#     AIX MPIO.
echo
echo 2 Checking if "Redundancy with dual VIO Server, Fiber Channel SAN disks and AIX MPIO" is
being used.
# Check if any of the disks have more than one path (listed twice)
MPIO=`lspath | awk '{print $2}' | uniq -d`
if [ $MPIO ]
then
        for n in $MPIO
        do
                # Check if this disk has a Failed path.
                STATUS=`lspath -l $n | grep Failed | awk '{print $1}'`
                if [ $STATUS ]
                then
                        ADAPTER=`lspath -l $n | grep Failed | awk '{print $3}'`
                        echo $n has $ADAPTER with Failed status. Enabling path.
                        chpath -s ena -l $n -p $ADAPTER
                        # Check new status
```

```
                        echo New status:
                        lspath -l $n
                else
                echo Status:
                        lspath -l $n
                fi

        done
else

echo "Redundancy with dual VIO Server, Fiber Channel SAN disks and AIX MPIO
"is NOT used.
fi


#   3 Network redundancy using "Network interface backup".
# Find out if this is being used and if so which interface number(s).

echo
echo 3 Checking if Network redundancy using "Network interface backup" is being used.

ECH=`lsdev -Cc adapter -s pseudo -t ibm_ech -F name | awk -F "ent" '{print $2}'`

if [ -z "$ECH" ]
then
echo No EtherChannel is defined.
else
# What is the status
        for i in $ECH
                do
                echo EtherChannel en$i is found.

                ETHCHSTATUS=`entstat -d en$i | grep Active | awk '{print $3}'`
                        if [ "$ETHCHSTATUS" = "backup" ]
                        then
                        # switch back to primary (requires AIX5.3-ML02 or higher)
                        echo Backup channel is being used. Switching back to primary.

                        /usr/lib/methods/ethchan_config -f en$i

                        # Check the new status
                        NEWSTATUS=`entstat -d en$i | grep Active | awk '{print $3}'`
                        echo Active channel: $NEWSTATUS adapter
                        #
                        else
                        echo Active channel: $ETHCHSTATUS adapter.
                        fi
#
                done
```

```
fi
exit
end
```

Abbreviations and acronyms

ABI	Application Binary Interface
AC	Alternating Current
ACL	Access Control List
AFPA	Adaptive Fast Path Architecture
AIO	Asynchronous I/O
AIX	Advanced Interactive Executive
APAR	Authorized Program Analysis Report
API	Application Programming Interface
ARP	Address Resolution Protocol
ASMI	Advanced System Management Interface
BFF	Backup File Format
BIND	Berkeley Internet Name Domain
BIST	Built-In Self-Test
BLV	Boot Logical Volume
BOOTP	Boot Protocol
BOS	Base Operating System
BSD	Berkeley Software Distribution
CA	Certificate Authority
CATE	Certified Advanced Technical Expert
CD	Compact Disk
CDE	Common Desktop Environment
CD-R	CD Recordable
CD-ROM	Compact Disk-Read Only Memory
CEC	Central Electronics Complex
CHRP	Common Hardware Reference Platform
CLI	Command Line Interface
CLVM	Concurrent LVM
CPU	Central Processing Unit
CRC	Cyclic Redundancy Check
CSM	Cluster Systems Management
CUoD	Capacity Upgrade on Demand
DCM	Dual Chip Module
DES	Data Encryption Standard
DGD	Dead Gateway Detection
DHCP	Dynamic Host Configuration Protocol
DLPAR	Dynamic LPAR
DMA	Direct Memory Access
DNS	Domain Naming System
DRM	Dynamic Reconfiguration Manager
DR	Dynamic Reconfiguration
DVD	Digital Versatile Disk
EC	EtherChannel
ECC	Error Checking and Correcting
EOF	End of File
EPOW	Environmental and Power Warning
ERRM	Event Response resource manager
ESS	IBM Enterprise Storage Server®
F/C	Feature Code
FC	Fibre Channel

FC_AL	Fibre Channel Arbitrated Loop
FDX	Full Duplex
FLOP	Floating Point Operation
FRU	Field Replaceable Unit
FTP	File Transfer Protocol
GDPS®	IBM Geographically Dispersed Parallel Sysplex™
GID	Group ID
GPFS	General Parallel File System
GUI	Graphical User Interface
HACMP™	High Availability Cluster Multiprocessing
HBA	Host Bus Adapter
HMC	Hardware Management Console
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
Hz	Hertz
I/O	Input/Output
IBM	International Business Machines
ID	Identification
IDE	Integrated Device Electronics
IEEE	Institute of Electrical and Electronics Engineers
IP	Internetwork Protocol
IPAT	IP Address Takeover
IPL	Initial Program Load
IPMP	IP Multipathing
ISV	Independent Software Vendor
ITSO	International Technical Support Organization
IVM	Integrated Virtualization Manager
JFS	Journaled File System
L1	Level 1
L2	Level 2
L3	Level 3
LA	Link Aggregation
LACP	Link Aggregation Control Protocol
LAN	Local Area Network
LDAP	Lightweight Directory Access Protocol
LED	Light Emitting Diode
LMB	Logical Memory Block
LPAR	Logical Partition
LPP	Licensed Program Product
LUN	Logical Unit Number
LV	Logical Volume
LVCB	Logical Volume Control Block
LVM	Logical Volume Manager
MAC	Media Access Control
Mbps	Megabits Per Second
MBps	Megabytes Per Second
MCM	Multichip Module
ML	Maintenance Level
MP	Multiprocessor
MPIO	Multipath I/O
MTU	Maximum Transmission Unit
NFS	Network File System
NIB	Network Interface Backup
NIM	Network Installation Management
NIMOL	NIM on Linux
N_PORT	Node Port
NPIV	N_Port Identifier Virtualization
NVRAM	Non-Volatile Random Access Memory
ODM	Object Data Manager
OS	Operating System
OSPF	Open Shortest Path First
PCI	Peripheral Component Interconnect

PCI-e	iPeripheral Component Interconnect Express
PIC	Pool Idle Count
PID	Process ID
PKI	Public Key Infrastructure
PLM	Partition Load Manager
POST	Power-On Self-test
POWER	Performance Optimization with Enhanced Risc (Architecture)
PPC	Physical Processor Consumption
PPFC	Physical Processor Fraction Consumed
PTF	Program Temporary Fix
PTX	Performance Toolbox
PURR	Processor Utilization Resource Register
PV	Physical Volume
PVID	Physical Volume Identifier
PVID	Port Virtual LAN Identifier
QoS	Quality of Service
RAID	Redundant Array of Independent Disks
RAM	Random Access Memory
RAS	Reliability, Availability, and Serviceability
RBAC	Role based access control
RCP	Remote Copy
RDAC	Redundant Disk Array Controller
RIO	Remote I/O
RIP	Routing Information Protocol
RISC	Reduced Instruction-Set Computer
RMC	Resource Monitoring and Control
RPC	Remote Procedure Call
RPL	Remote Program Loader
RPM	Red Hat Package Manager
RSA	Rivet, Shamir, Adelman
RSCT	Reliable Scalable Cluster Technology
RSH	Remote Shell
SAN	Storage Area Network
SCSI	Small Computer System Interface
SDD	Subsystem Device Driver
SDDPCM	MPIO Subsystem Device Driver
SMIT	System Management Interface Tool
SMP	Symmetric Multiprocessor
SMS	System Management Services
SMT	simultaneous multithreading
SP	Service Processor
SPOT	Shared Product Object Tree
SRC	System Resource Controller
SRN	Service Request Number
SSA	Serial Storage Architecture
SSH	Secure Shell
SSL	Secure Sockets Layer
SUID	Set User ID
SVC	SAN Virtualization Controller
TCP/IP	Transmission Control Protocol/Internet Protocol
TL	Technology Level
TSA	Tivoli System Automation
UDF	Universal Disk Format
UDID	Universal Disk Identification
VIPA	Virtual IP Address
VG	Volume Group
VGDA	Volume Group Descriptor Area

VGSA	Volume Group Status Area
VLAN	Virtual Local Area Network
VP	Virtual Processor
VPD	Vital Product Data
VPN	Virtual Private Network
VRRP	Virtual Router Redundancy Protocol
VSD	Virtual Shared Disk
WLM	Workload Manager
WWN	Worldwide Name
WWPN	Worldwide Port Name

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

For information about ordering these publications, see “How to get Redbooks” on page 715. Note that some of the documents referenced here might be available in softcopy only.

- *AIX 5L Practical Performance Tools and Tuning Guide*, SG24-6478
- *Effective System Management Using the IBM Hardware Management Console for pSeries*, SG24-7038
- *Hardware Management Console V7 Handbook*, SG24-7491
- *i5/OS on eServer p5 Models A Guide to Planning, Implementation, and Operation*, SG24-8001
- *IBM AIX Continuous Availability Features*, REDP-4367
- *IBM AIX Version 7.1 Differences Guide*, SG24-7910
- *IBM Director on System p5*, REDP-4219
- *IBM eServer iSeries Performance Management Tools*, REDP-4026
- *IBM i and Midrange External Storage*, SG24-7668
- *IBM PowerVM Getting Started Guide*, SG24-4815
- *IBM PowerVM Live Partition Mobility*, SG24-7460
- *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940
- *IBM System i and System p*, SG24-7487
- *IBM Systems Director VMControl Implementation Guide on IBM Power Systems*, SG24-7829
- *Implementing an IBM b-type SAN with 8 Gbps Directors and Switches*, SG24-6116
- *Integrated Virtual Ethernet Adapter Technical Overview and Introduction*, REDP-4340
- *Integrated Virtualization Manager on IBM System p5*, REDP-4061

- *Introduction to pSeries Provisioning*, SG24-6389
- *Introduction to Workload Partition Management in IBM AIX Version 6.1*, SG24-7431
- *Linux Applications on pSeries*, SG24-6033
- *Managing AIX Server Farms*, SG24-6606
- *Managing OS/400 with Operations Navigator V5R1 Volume 5: Performance Management*, SG24-6565
- *NIM from A to Z in AIX 5L*, SG24-7296
- *Partitioning Implementations for IBM eServer p5 Servers*, SG24-7039
- *Power Systems Memory Deduplication*, REDP-4827
- *PowerHA for AIX Cookbook*, SG24-7739
- *PowerVM Migration from Physical to Virtual Storage*, SG24-7825
- *PowerVM Virtualization Active Memory Sharing*, REDP-4470
- *A Practical Guide for Resource Monitoring and Control (RMC)*, SG24-6615
- *Virtualizing an Infrastructure with System p and Linux*, SG24-7499

Other publications

These publications are also relevant as further information sources:

- The following types of documentation are located on the Internet at:

  http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/index.jsp

  - User guides
  - System management guides
  - Application programmer guides
  - All commands reference volumes
  - Files reference
  - Technical reference volumes used by application programmers
- Detailed documentation about the PowerVM feature and the Virtual I/O Server is available at:

  https://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/home.html

- *IBM eServer iSeries Performance Tools for iSeries*, SC41-5340
- *IBM Tivoli Usage and Accounting Manager Data Collectors for UNIX and Linux User's Guide,* SC32-1556

Online resources

These websites are also relevant as further information sources:

- IBM System p Virtualization — the most complete virtualization offering for UNIX and Linux:

  http://www-01.ibm.com/cgi-bin/common/ssi/ssialias?infotype=an&subtype=ca&htmlfid=897/ENUS207-269&appname=usn&language=enus

- HMC interaction script:

  http://www.the-welters.com/professional/scripts/hmcMenu.txt

- IBM Redbooks:

  http://www.redbooks.ibm.com/

- IBM Systems information center: Power Systems Virtual I/O Server and Integrated Virtualization Manager commands:

  http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/topic/iphcg/iphcg.pdf

- IBM System Planning Tool:

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

- IBM wikis:

  http://www.ibm.com/developerworks/wikis/dashboard.action

  - AIX Wiki — Performance Monitoring Documentation

    http://www.ibm.com/developerworks/wikis/display/WikiPtype/Performance+Monitoring+Documentation

  - nmon analyzer tool

    http://www.ibm.com/developerworks/wikis/display/Wikiptype/nmonanalyser

- Virtual I/O Server monitoring wiki

  http://www.ibm.com/developerworks/wikis/display/WikiPtype/VIOS_Monitoring

- The nmon tool

  http://www.ibm.com/developerworks/wikis/display/WikiPtype/nmon

- IBM Systems Information Centers

  http://publib.boulder.ibm.com/eserver/?tocNode=int_17

- Virtual I/O Server 2.2 release notes

  https://www-304.ibm.com/support/docview.wss?uid=isg400000259

- pSeries and AIX information center - Installing and configuring the system for Kerberos integrated login using KRB5

  http://publib.boulder.ibm.com/infocenter/pseries/v5r3/index.jsp?topic=/com.ibm.aix.security/doc/security/kerberos_auth_only_load_module.htm

- Advanced Power Virtualization

  http://www-03.ibm.com/systems/power/software/virtualization/index.html

- Tivoli software information center

  http://publib.boulder.ibm.com/tividd/td/IdentityManager5.0.html

- Architecting for power management: The IBM POWER7 approach

  http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=5416627&queryText%3DArchitecting+for+power+management

- EnergyScale for IBM POWER6 microprocessor-based systems

  http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=5388630&queryText%3DEnergyScale+for+IBM+POWER6+microprocessor-based+systems

- System power management support in the IBM POWER6 microprocessor

  http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=5388627&queryText%3DSystem+power+management+support+in+the+IBM+POWER6+microprocessor

- IBM i 6.1 Information Center

  http://publib.boulder.ibm.com/infocenter/systems/scope/i5os/index.jsp

- IBM i Software Knowledge Base: Cross-Referencing (Device Mapping) IBM i Disks with VIOS Disks with IVM

  http://www-912.ibm.com/s_dir/slkbase.nsf/1ac66549a21402188625680b0002037e/23f1e308b41e40a486257408005aea5b?OpenDocument&Highlight=2,481468986

- SSD and Powerpath information

  http://www-01.ibm.com/support/search.wss?rs=540&tc=ST52G7&dc=DA480+DB100&dtm

- IBM Tivoli Monitoring

  http://www-306.ibm.com/software/tivoli/products/monitor-systemp/

- IBM Tivoli Monitoring Information Center

  http://publib.boulder.ibm.com/infocenter/tivihelp/v15r1/index.jsp?topic=/com.ibm.itm.doc/welcome.htm

- Tivoli Monitoring web page

  http://www-01.ibm.com/software/tivoli/products/monitor/

- IBM Tivoli Application Dependency Discovery Manager

  http://www-01.ibm.com/software/tivoli/products/taddm/

How to get Redbooks

You can search for, view, or download Redbooks, Redpapers, Technotes, draft publications and Additional materials, as well as order hardcopy Redbooks, at this website:

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Index

D

E

F

G

H

I

J

L

M

Q

R

S

T

U

V

W

X

IBM PowerVM Virtualization Managing and Monitoring

(1.0” spine)
0.875”<->1.498”
460 <-> 788 pages

IBM PowerVM Virtualization Managing and Monitoring

Provides managing and monitoring best practices

Consolidates sources for PowerVM publications

Includes Virtual I/O Server 2.2.2 enhancements

IBM PowerVM virtualization technology is a combination of hardware and software that supports and manages the virtual environments on POWER5-, POWER5+, IBM POWER6, and IBM POWER7-based systems.

PowerVM is available on IBM Power Systems, and IBM BladeCenter servers as optional Editions, and is supported by the IBM AIX, IBM i, and Linux operating systems. You can use this set of comprehensive systems technologies and services to aggregate and manage resources by using a consolidated, logical view. Deploying PowerVM virtualization and IBM Power Systems offers you the following benefits:

- Lower energy costs through server consolidation
- Reduced cost of your existing infrastructure
- Better management of the growth, complexity, and risk of your infrastructure

This IBM Redbooks publication is an extension of *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940. It provides an organized view of best practices for managing and monitoring your PowerVM environment concerning virtualized resources managed by the Virtual I/O Server.

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

SG24-7590-04 ISBN 0738438162