



Dynamic Logical Partitioning in IBM @server pSeries

First Edition (October 8, 2002)

Before using this information and the product it supports, read the information in “Notices” on page 8.

© **International Business Machines Corporation, 2002. All rights reserved.** Note to U.S. Government Users
Restricted Rights - Use, duplication, or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Introduction	1
Version Dependencies	1
Dynamic LPAR Resources	1
Dynamic versus Automated	1
Dynamic LPAR's Value	2
Dynamic LPAR Management	2
Moving LPAR Resources	2
Dynamic LPAR and Security	3
Dynamic LPAR and CUoD	4
AIX Support of Dynamic LPAR	4
Binary Compatibility	5
Exploiting DLPAR	5
Flexible Memory Assignment	5
User Interactions	6
Performance Implications	6
Monitoring DLPAR Operations	7
Build Support in AIX 5L 5.1	7
The IBM @server LPAR Family	7
Summary	8
Special Notices	8

Introduction

The introduction of logical partitioning (LPAR) technology to IBM @server pSeries™ systems has greatly expanded the options for deploying applications and workloads onto server hardware. Logical partitioning is a server design feature that provides more end-user flexibility by making it possible to run multiple, independent operating system images concurrently on a single server. IBM is adding to that LPAR capability with the introduction of dynamic LPAR (DLPAR), in which partition resources can be moved from one partition to another without requiring a reboot of the system or affected partitions. This paper will describe these new capabilities and explain the benefits of their use.

Version Dependencies

To fully utilize dynamic LPAR capabilities, the following are required:

- A POWER4™ processor-based pSeries system, such as the p690, p670, p630, or follow-on
- The October 2002 (or later) system microcode update
- A Hardware Management Console (HMC) at version R3V1.0 or later
- AIX® 5L™, Version 5.2 or later

Not all partitions on a system must be migrated to AIX 5.2 for dynamic LPAR to work, but only those partitions running AIX 5.2 or later will be able to participate in dynamic LPAR operations. These new versions of system microcode and HMC will continue to support AIX 5.1 at its current level of LPAR functionality.

Dynamic LPAR Resources

In an LPAR configuration, individual processor, 256MB memory region, and I/O

adapter slot resources are placed under the exclusive control of a given logical partition. One of the main advantages of the LPAR implementation is that it gives fine-grained allocation control over these individual resources, allowing them to be combined in almost any quantity and combination to create a logical partition.

Dynamic LPAR extends these capabilities by allowing this fine-grained resource allocation to occur not only when activating a logical partition, but also while the partitions are running. Individual processors, memory regions, and I/O adapter slots can be released into a “free pool,” acquired from that free pool, or moved directly from one partition to another— again, in almost any quantity or combination.

Resources moved by DLPAR operations have the same full set of capabilities that they would have if assigned to the partition at boot time. For example, a moved processor has full access to all of the partition’s memory, I/O address space, and I/O interrupts, and so can participate fully in supporting that partition’s workload.

Users of the affinity partition configuration option, which allocates CPU and memory resources in fixed patterns based on multi-chip module (MCM) boundaries, should note that only the I/O adapter resources can be dynamically reconfigured while in that mode.

Dynamic versus Automated

While this introductory release of dynamic LPAR does provide the full capability of moving resources between running partitions, it does not mean that such LPAR resource movements will be occurring in a spontaneous or unexpected fashion. Rather, it means that LPAR resource movements are fully *dynamic* (performed non-disruptively while partitions continue to run) but not necessarily *automated*

(driven independently by some internal policy or condition/response). As will be described later, the products are enabled so that scripts can be written to drive any sort of DLPAR automation. This capability also enables future offerings in server automation and workload management.

Dynamic LPAR's Value

Dynamic LPAR provides even more flexibility in dealing with changing workload demands and server deployments. Some obvious examples include:

- Move processors from a test partition to a production partition in periods of peak demand, then move them back again as demand decreases.
- Move memory to a partition that is doing excessive paging.
- Move an infrequently used I/O device between partitions, such as a CD-ROM for installations, or a tape drive for backups.
- Release a set of processor, memory, and I/O resources into the “free pool,” so that a new partition can be created from those resources.
- Configure a set of minimal LPARs on a single system to act as “failover” backup servers to some primary servers, and also keep some set of resources free. If one of the associated primaries fails, then assign free resources to that backup LPAR so that it can pick up the workload.

As can be seen by these examples, DLPAR opens a whole new set of possibilities for improving operational efficiency and getting more value from server hardware investments.

Dynamic LPAR Management

Dynamic LPAR operations in pSeries are enabled by three major components:

- System Firmware, including the LPAR hypervisor, provides services that can add or remove resources within a running partition.
- AIX 5L Version 5.2 provides commands and kernel services that allow AIX resources to be dynamically acquired and released.
- The IBM Hardware Management Console for pSeries provides a choice of graphical user interface or command line to control the movement of resources.

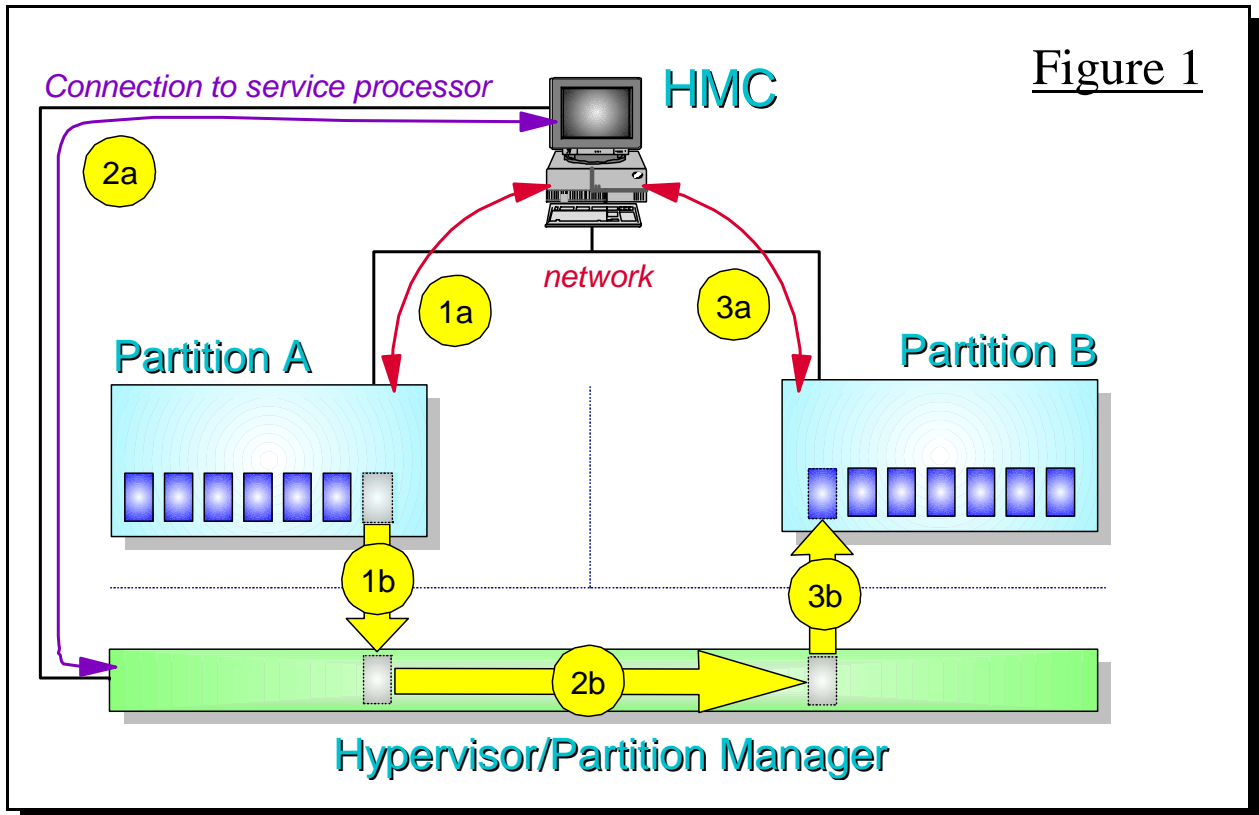
These three components, working together, enable seamless, integrated management of DLPAR resource movements.

Moving LPAR Resources

The movement of an LPAR resource from one partition to another actually requires a sequence of three discrete transactions, as illustrated in Figure 1. These transactions are:

1. The HMC sends a network request to AIX in Partition A, asking it to release a resource and put it into a quiesced state. The resource is stopped, and placed under control of the hypervisor.
2. The HMC sends a command to the hypervisor, asking it to reallocate the resource from Partition A to Partition B.
3. The HMC sends a network request to AIX in Partition B, asking it to acquire the resource from the hypervisor and configure it for use.

These requests can be initiated either through selections made on the graphical user interface of the HMC, which can be accessed locally or remotely from a WebSM client, or by issuing an HMC command-line function. The command-line functions may be initiated from any other operating system environment on the network – provided there is proper authorization – by connecting into the HMC



through either a *rexec* or *OpenSSH* client. For example, DLPAR commands could be initiated from a script running in one of the partitions, or from a centralized system management application.

In addition to the option to *move* resources directly between partitions, the HMC also provides options to *remove* resources from a partition and place them in an unassigned “free pool,” or to take resources from the unassigned free pool and *add* them to a partition.

This automation of DLPAR operations requires a network connection between the HMC and the partitions. DLPAR operations can be performed without the network connection, but this requires the three steps to be performed independently and manually. Because the HMC also uses this same network connection for collecting hardware error events and inventory data for the Service Focal Point

application on the HMC, a network connection is highly recommended.

Dynamic LPAR and Security

The ability to move resources between partitions does not compromise the basic security of the LPAR environment. An operating system within a partition does not have any visibility to other partitions or to any resources currently outside of that partition – not even the unassigned resources in the “free pool.” The operating system does see a set of virtual resource *connectors*, where processors, memory, or I/O can potentially be configured. After the HMC adds a resource to a partition, it sends a message to that operating system, requesting it to vary on one or more of those virtual connectors. If the operating system attempts this without the HMC and hypervisor first adding a resource to that partition, the operating system will simply receive an error on

the operation, indicating that no resource is present.

When resources are moved between the partitions, the hypervisor reinitializes the device and ensures that no residual data remains. For example, when a memory region is moved from one partition to another, it is reinitialized and filled with all 0 values.

The network connections from the HMC to the partitions use secure communication mechanisms based on IBM's Reliable Scalable Cluster Technology (RSCT), so that only the HMC can issue these DLPAR requests. Each user of the HMC is assigned a specific *role* that controls whether they are permitted to perform such actions.

Dynamic LPAR and CUoD

Dynamic LPAR also works with the keyed Capacity Upgrade on Demand (CUoD) offering. Through the use of a license key, CUoD provides on-demand hardware upgrades by activating resources that are already physically installed, but kept in a dormant state. For example, if a system is upgraded by entering a license key for additional processors, then those processor resources are activated and put into an unassigned state, while the system continues to run. DLPAR operations can then be used to dynamically and selectively add those newly available processor resources to the desired partitions.

Dynamic LPAR also works with CUoD to provide a *dynamic sparing* capability for processors, in which an unlicensed processor will be automatically activated and assigned to a partition in which one of the processors has reported a predictive failure. A *predictive failure* is one where the system observes an increasing trend of soft errors that have been successfully corrected or retried, and proactively requests AIX to remove that

processor from use because it has reached an error threshold. If enabled, AIX will automatically vary the new spare processor on, and then vary off the processor for which the predictive failure was reported. With AIX 5.2, this dynamic sparing capability is supported even in single-processor partitions.

AIX Support of Dynamic LPAR

AIX 5L Version 5.2 supports the dynamic reconfiguration of processors, memory, and PCI I/O slots in a non-disruptive manner; because the support for DLPAR has been added without impacting the programming model for applications and kernel extensions. Most programs need not be changed to function properly in a DLPAR-enabled environment.

The risk associated with the dynamic addition and removal of memory has been mitigated by changing the AIX kernel, so that it runs almost entirely in virtual mode. This design has the effect of insulating the system from the effects of removing physical memory, because applications, kernel extensions, and indeed most of the kernel, use only virtual memory.

For several years, AIX has supported the dynamic *removal* of processors that have been predicted to fail (Dynamic Processor Deallocation), so DLPAR is not really introducing anything new in this area. Processor *addition* is new, but carries little risk, because most applications are not internally aware of the number of online processors. Applications that are aware of the number of online processors typically use this information to determine the number of threads to create, or to segregate jobs, but the addition of a processor will not impact their operation.

The dynamic reconfiguration of PCI I/O slots is also not really new. It is an extension of the AIX PCI hot-plug capabilities, which were introduced several years ago. The movement flow clearly illustrates this point. First, the administrator logs into the partition and uses the AIX PCI hot-plug procedures to ensure that the associated adapters and devices are not in use. Second, the administrator uses the DLPAR procedures at the HMC to move the slot to the target partition. Finally, the administrator logs into the target partition and again uses the AIX PCI hot-plug procedures to configure the associated adapters and devices. This implementation ensures that there are no concurrent accesses to the slot while it is being reassigned.

Binary Compatibility

AIX 5L Version 5.2 maintains binary compatibility with earlier levels of AIX 5L, as well as 32-bit application compatibility for most applications from prior versions. Binaries built on earlier versions of AIX 5L will continue to run on AIX 5.2 without modification. DLPAR should not impact any applications or kernel extensions; however, it is recommended that vendors retest their applications and middleware on AIX 5.2 when utilizing DLPAR.

Exploiting DLPAR

While most applications and middleware are not impacted by DLPAR, it was also recognized that a more complete competitive advantage could be achieved if the entire software stack could be made to dynamically adjust its use of system resources in conjunction with the DLPAR event. To this end, the system architecture has been extended so that applications and middleware can also participate in the DLPAR event, and thereby expand and contract with the base operating system as capacity is added and removed.

AIX 5.2 provides two mechanisms for notifying applications and middleware that a DLPAR request is in progress. Vendors may use either DLPAR scripts or APIs to dynamically resize their subsystems. The former may also be used by system administrators who want to define their own policies regarding DLPAR. For example, a system administrator might want to halt an application if it is known to cause DLPAR requests to fail, and if the service provided by the application is not considered critical to the operation of the system.

Some examples of programs that are being made DLPAR-aware to exploit this new technology are enterprise level databases, workload managers, performance tools, and license managers.

Flexible Memory Assignment

One of the key design goals for this capability was to significantly reduce the amount of real memory required by the operating system. Previously, with AIX 5L Version 5.1, the hypervisor had to allocate an initial block of contiguous physical memory that scaled with the total amount of logical memory that was assigned to the logical partition. In some cases, this could result in a situation where it was not possible to activate a partition, particularly large ones, because there was not enough contiguous memory to satisfy the operating system requirement, even though there were sufficient unassigned resources in the system to accommodate the partition requirement.

To overcome this problem, the user had to change the size of the partition, shut down other partitions, and even in extreme cases, reboot the system. With AIX 5.2, this problem has been solved. Unassigned resources can now be drawn from anywhere in the system. To enable this feature, the administrator must select the new **Small Real Mode Address**

Region option when creating or editing a partition profile on the HMC.

User Interactions

The reconfiguration of processors and memory is well integrated into AIX; that is, the administrator need take no action in most cases to reconfigure these resources. The reconfiguration of PCI I/O slots is a bit more complicated in that the administrator must also log into the partition and follow the AIX PCI hot-plug procedures, which are provided through the SMIT GUI.

The reason for this manual intervention in the case of PCI I/O slots is that the availability of the associated resources and the services provided by them is not well understood by the base operating system. The decision to remove a slot may involve the reconfiguration of physical and logical devices, as well as applications, and is therefore better left to system administrators. Although the automation for I/O slot reconfiguration is somewhat limited in this release, this is not viewed as a critical shortcoming, because dynamic reconfiguration of I/O between partitions is rarely done. The critical resources that need to be handled smoothly are processors and memory.

Performance Implications

Most DLPAR operations can be performed in a couple of minutes with the exception of memory removal, which scales with the amount of memory that is specified by the user. In general, it takes 1 to 2 minutes to remove 4 GBs of memory, depending on the state of memory in the partition.

When DLPAR operations are in progress, the performance of the operating system may suffer slightly as resources are being examined and rebalanced. When a resource is added, it is

immediately made available for use, in the same way as if the operating system had booted with the resource. In general, the performance benefit associated with resource addition and removal will scale proportionally to the change in resources. This is particularly true for processors, which in general do not have secondary software impacts associated with their use in the way that memory does. For example, to preserve the contents of memory that is being removed, it is necessary to migrate it to a new location or to copy it to paging space.

As a consequence, the impact of memory removal on the performance of the system is a bit more complicated to measure, because it is more sensitive to the workload. For example, when memory is removed, the file system buffer pools are shrunk, which may cast out an i-node that may need to be reread from the disk. These sorts of secondary impacts should be automatically handled by the system as the working set is reestablished. However, if the partition is left with insufficient memory, it can lead to excessive paging. It is therefore important to understand the workload requirements in the partition, particularly in the case of memory, to ensure that the partition has adequate resources.

In short, while the operating system is designed to recognize and adjust its use of system resources, a systemwide approach including middleware and applications is required to achieve the best solution. Even so, it should be noted that the system in most cases will scale well without specific application and middleware support for DLPAR, because the base operating system contains significant built-in support for DLPAR. In particular, the AIX JFS2 File System and Networking subsystems have been enhanced to work with DLPAR so that the operating system is well

balanced with respect to the underlying physical configuration.

Monitoring DLPAR Operations

DLPAR operations can be monitored in a variety of ways. By default, AIX presents information about the current DLPAR request on the operator panel in the form of LED codes and short text messages, which can be seen in the partition view on the HMC user interface. The user may also request a more detailed report by specifying a detail level in the HMC user dialog when making a DLPAR request. This report includes progress information for the base operating system, and it may also include information about specific applications and middleware. This information is presented to the user at the HMC. The default is no detailed data.

Detailed reports may also be generated through the AIX **syslog** command, which is a more flexible logging facility. If this method is desired, the administrator must also configure the syslog facility itself. This report is not enabled by default.

If a DLPAR operation fails, the user should check the AIX error log to determine the actual cause of the failure. In general, the error log is used to capture error conditions caused by kernel extensions and applications, which in many cases can be corrected by the user. For example, if a processor attachment is preventing a processor from being taken offline, an error log entry is generated, identifying the specific process that needs to be reconfigured.

Build Support in AIX 5L 5.1

The DLPAR APIs for applications have also been added to AIX 5.1, which allows vendors to ship a single binary for AIX 5.1 and 5.2. Programmers should develop and test their code on AIX 5.2 systems under real DLPAR

conditions, and then recompile on an AIX 5.1 system to generate a common binary.

To upgrade your AIX 5.1 system to contain the DLPAR APIs, order APAR IY34097. If these interfaces are invoked on a system installed with AIX 5.1, they simply indicate that there is no DLPAR operation in progress

The IBM @server LPAR Family

IBM @server pSeries logical partitioning is based on hardware enablement in its family of PowerPC® microprocessors. This same family of microprocessors is the core technology for IBM @server iSeries™, which has offered the logical partition feature since 1999.

Those familiar with IBM @server zSeries™ and S/390® logical partitioning will recognize strong similarities with their LPAR and dynamic LPAR features, which have been available for many years. While the dynamic LPAR features being introduced on pSeries do not yet match the full set of capabilities on zSeries, the fundamental design has had the benefit of the extensive experience in this area on zSeries systems.

Summary

The pSeries dynamic LPAR offering provides increased flexibility and adaptability of LPAR configurations by enabling non-disruptive resource movement between running partitions. This paper has explored how this capability works, through a combination of architecture, system firmware, and software, including:

- **Enablement in the LPAR hypervisor** for secure resource movement between running partitions.
- **Enablement in AIX 5.2** to dynamically vary processor, memory, and I/O resource on and off while the operating system and applications continue their operation.
- **Applications in the Hardware Management Console** that coordinate the activities of the AIX images and the hypervisor, to provide a simple and seamless control over dynamic LPAR resource movement.
- **Enablement of middleware and applications in the operating system** to take advantage of dynamic resource changes.

In conclusion, the IBM @server pSeries dynamic LPAR offering can be used with confidence as part of a flexible and adaptable server consolidation strategy, doing so with minimal disruption of operation, and thereby providing better support for competitive and challenging business environments.

Special Notices

© Copyright IBM Corporation 2002

IBM Corporation
Marketing Communications
Server Group
Route 100
Somers, New York 10589

Printed in the united States of America
10-02
All rights Reserved

This document was produced in the United States. IBM may not offer the products, programs, services or features discussed herein in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the products, programs, services, and features available in your area. Any reference to an IBM product, program, service or feature is not intended to state or imply that only IBM's product, program, service or feature may be used. Any functionally equivalent product, program, service or feature that does not infringe on any of IBM's intellectual property rights may be used instead of the IBM product, program, service or feature.

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

IBM Director of Licensing
IBM Corporation
New Castle Drive
Armonk, NY 10504-1785 USA.

The information contained in this document has not been submitted to any formal IBM test and is distributed "AS IS". While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. The use of this information or the implementation of any techniques described herein is a customer responsibility and depends on the customer's ability

to evaluate and integrate them into the customer's operational environment. Customers attempting to adapt these techniques to their own environments do so at their own risk.

IBM is not responsible for printing errors in this publication that result in pricing or information inaccuracies.

The information contained in this document represents the current views of IBM on the issues discussed as of the date of publication. IBM cannot guarantee the accuracy of any information presented after the date of publication.

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

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

The [e(logo) server] brand consists of the established IBM e-business logo followed by the descriptive term "server."

AIX, AIX 5L, PowerPC, POWER4, pSeries, iSeries, zSeries, and S/390. A full list of U.S. trademarks owned by IBM may be found at : <http://iplswww.nas.ibm.com/wpts/trademarks/trademar.htm>.

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