Effective server consolidation and resource management with system partitioning
December 2001

System Partitioning on IBM *e*server xSeries Servers

by Mark T. Chapman
IBM Server Group

Executive Summary

What is system partitioning, and why should you care? Simply put, it is the ability for a server to run multiple operating systems simultaneously. Each operating system instance runs in a separate partition and each partition is isolated and protected from the others, so that if one operating system freezes the other partitions are unaffected.

Until now, system partitioning has been the domain of mainframes and other large, sophisticated systems. However, with Enterprise X-Architecture™ technology IBM extends partitioning to industry-standard servers as well.

IBM @server X-Architecture technology[1] is a blueprint for extending the benefits of advanced mainframe technologies to our Intel® processor-based servers. These benefits are in the areas of availability, scalability, systems management, service and support. IBM has been delivering on the promise of the X-Architecture model since 1998 with such innovative technologies as Active™ PCI, C2T Interconnect™ cabling, Chipkill™ memory, Predictive Failure Analysis® (PFA), Light Path Diagnostics™ and IBM Director Software Rejuvenation, to name a few.

IBM continues to build on the X-Architecture blueprint with a number of Enterprise X-Architecture technologies[2]. One of the key scalability technologies introduced with Enterprise X-Architecture servers is hardware-based system partitioning. In addition, ESX Server™ from VMware™—a company that provides mainframe-class virtual machine technology to IBM servers—offers a *software* partitioning solution that complements Enterprise X-Architecture hardware partitioning and adds needed partitioning capabilities to other selected servers.

The benefits of system partitioning include:

- Server consolidation
- High availability
- Software migration and coexistence
- Version control
- Development, testing and maintenance
- Workload isolation
- Independent backup and recovery on a partition basis

This paper not only explores the various tools available to manage system partitioning, but also how you can take advantage of partitioning to help address your business and IT needs.

[1] See the white paper entitled "IBM X-Architecture Technology" at http://**ibm.com**/eserver/xseries for more information. From the xSeries home page, select **Library** for links to the different types of documentation available.

[2] Go to http://**ibm.com**/eserver/enterprise_xarchitecture for the "Introducing Enterprise X-Architecture Technology" white paper.

Table of Contents

What is System Partitioning?

System partitioning is one of the many mainframe capabilities that Enterprise X-Architecture technology brings to Intel architecture servers. Anyone who has worked with mainframes and other large systems, such as IBM @server iSeries and pSeries servers, is familiar with the concept of partitioning: System resources, including processor, memory, I/O and storage, are virtualized so that all concurrent users appear to have complete access to the system. Yet each user is actually segmented—and protected—from the actions of all other users. If one virtual partition were to freeze up, it would not affect the others. In a mission-critical environment, such as a worldwide airline reservation system, for example, it would be disastrous if one such errant partition could lock up the entire mainframe. Each partition is also protected from viruses and other security exposures that might have affected another partition. In many client-server environments, this level of system availability is no less essential.

Uses for System Partitioning

System partitioning may provide you with better security from viruses and software crashes, but can you do anything more *proactive* with it? Absolutely. If you're like most IT organizations these days, you are looking for ways to cut costs and improve IT efficiency and security via server consolidation, streamlined system administration (including remote management and software deployment), business/disaster recovery, squeezing more use out of existing servers and so on. Here are just a few of the ways that system partitioning can help you reach many of your business and IT goals:

- *Server hardware consolidation*— Consolidate many underused, underpowered and unnecessary servers into a few productive ones. Reduce the number of current servers and buy fewer servers in the future.
- *Increased server utilization* — Divide a processor into multiple partitions versus wasting an entire processor (or an entire 2-way server) on one low-throughput application; turn a multiprocessor server into a "virtual blade server."
- *Simplified server management* — Manage fewer servers centrally rather than many of them individually in multiple locations. Have fewer servers, cables, operating systems and applications to deal with.
- *Low-cost clustering/failover* — Create clusters of partitions among hardware nodes or software partitions instead of using separate servers. Set up N+1 failover *within* a server, or have several different servers failover to *partitions* in one server, locally or remotely.
- *Storage virtualization* — Instead of having fixed storage capacity for individual servers, simply (re)allocate resources (within a server or on a SAN) dynamically to partitions as needs change.
- *Simplified application deployment* — Once you have tested and qualified a specific hardware platform for use with a particular operating system and application combination you can deploy software images on multiple partitions—rather than having to requalify the software on another hardware platform.

These are all desirable goals and ones that you may be actively pursuing already, yet there are many possible obstacles that may prevent you from achieving them:

- Perhaps your entire engineering division is running Red Hat Linux®, the accounting group is using Microsoft® Windows® 2000 Advanced Server and the worldwide sales staff is on Windows NT® 4.0. Each group of users may span dozens of servers in different geographies. How could you possibly consolidate all those users on one server—or even a few?
- You have two departments running different *versions* of one application (which won't run together on the same server) for legitimate business reasons.

- You may even have several applications that you *think* will run correctly together; however, you don’t have the resources to test every possible software interaction, so you give each application its own server just to be sure.
- Your applications don’t scale effectively beyond one or two processors. There’s no point in running them on 4-way (or larger) servers, so you’re stuck with hordes of uniprocessor and 2-way servers.
- Another obstacle is security. You wouldn’t want your payroll application running alongside your user applications. Or, if you are a service provider hosting multiple commercial customers you would need absolute separation of customers’ data. And, of course, you wouldn’t want a virus to be able to affect the entire system.
- As a service provider (with internal *or* external customers), you also need to be able to commit to a level of service as defined in a service level agreement (SLA). How can you do that if one application can gobble up all the available processor cycles or memory at the expense of other applications?
- If you are billing for computer resources, you would need a way to be able to charge for discrete resource usage when multiple users are sharing a server.
- What about the dozens of ancient servers scattered around your organization that are too feeble to run more than one light-duty application? You’d like to eliminate most of those servers, but don’t have a way to consolidate them into one system.
- You may have some legacy applications that require out-of-date operating systems, or that are “ill-behaved” and will not tolerate any other software running on the servers; perhaps they have a problem with memory leaks or intermittent crashes, or each requires a different version of the same Windows dynamic link library (DLL). Let’s face it, many an application presumes that it’s the only one running on the server, and usually the safest course is to let it *have* an entire system all to itself. Of course, this approach contributes to server proliferation.

These situations occur in every organization to one extent or another, resulting in many underused servers. Perhaps you have uniprocessor or 2-way servers running at 40% utilization, or high-function, high-price 8-way servers with only four processors installed—because you don’t have enough users of the application it serves to justify adding more processors. So the extra capacity is being wasted. If some of the applications hosted by the underutilized servers could be consolidated onto one server, you could save a small fortune in systems management expenses, postponed server purchases, user training and other specific costs. Unfortunately, because of all the reasons listed previously, there doesn’t seem to be any way to accomplish this sort of consolidation.

What you need is a way to isolate all of those legacy, ill-behaved, “leaky” and crash-prone applications from one another, as well as to host simultaneous instances of multiple application versions and operating systems. This is exactly what system partitioning offers: It gives you a flexible way to minimize your hardware requirements by allowing you to run all that software together, while simultaneously eliminating the need to test every possible software interaction and simplifying and centralizing your system administration, application deployment and problem determination and resolution.

Today, there are two complementary approaches to system partitioning on industry-standard servers: hardware-based partitioning for enterprise-class servers—enabled by Enterprise X-Architecture technology, and a software solution that not only adds more granular partitioning capabilities to enterprise-class servers, but which also offers much of the same functionality on less advanced servers.

Hardware-based System Partitioning

Starting in 2002, IBM @server xSeries servers that offer hardware-based system partitioning will take advantage of an Enterprise X-Architecture innovation called *XpandOnDemand™ scalability*[3]. New levels of scalability for industry-standard servers are achieved with the Enterprise X-Architecture platform using enhanced, high-performance symmetrical multiprocessing (SMP) system building blocks that allow effective scalability beyond 4-way SMP. These technologies provide scalability from 4-way to 8-way to 12-way—and even to 16-way systems—using 4-way "scalable enterprise nodes."

A scalable enterprise node contains processors, memory, I/O support, storage and other devices and can operate independently like other computers. Each node may run an operating system (OS) different from the other nodes, or if desired multiple nodes can be assigned to one OS image via system partitioning. Nodes are attached to one another through dedicated high-speed interconnections, called *SMP Expansion Ports*, sharing resources for unmatched performance. This gives you the adaptability to run several nodes as either a single large "complex" or as two or more smaller units—and even to rearrange the configurations later as needed.

The SMP Expansion Ports allow nodes to talk to one another at up to **3.2 giga*bytes*** per second in bidirectional mode (roughly equivalent to a **32 giga*bit*** per second network connection, assuming eight bits of data and two bits for network overhead) *per connection*, with each node supporting *up to three* connections to other nodes. 3.2GB is ***32 times*** what is currently available from even Gigabit Ethernet connectivity (or 16 times as much as Gigabit Ethernet configured for bidirectional operation through a LAN switch). Plus, your infrastructure doesn't have to be redesigned to support Gigabit Ethernet for just those few boxes—not to mention the cost savings of not needing Gigabit Ethernet hubs, routers, switches or adapters for just those clustered servers. If you don't use Gigabit Ethernet, consider that the SMP Expansion Ports are ***320 times*** as fast as 100Mbps Ethernet. And the ports are *in addition* to any Ethernet ports installed on the server, leaving those ports available for normal network connectivity. (In other words, you don't have to tie up slots with Ethernet adapters in each server simply to connect to the other three servers.)

Enterprise X-Architecture system partitioning includes two types of partitioning for these multinode systems: *physical* partitioning (enabled today) and *logical* partitioning (coming in the future). With physical partitioning, a single multinode server can *simultaneously* run multiple instances of one or more operating system in separate partitions, as well as multiple *versions* of an OS—to perform as one virtual, flexible server, optimized for the users it supports. The server can have up to four nodes. (This is *not* the same as having multiple *hard disk* partitions loaded with different operating systems, where changing operating systems requires restarting the server from another partition. The operating systems are running *simultaneously* in different nodes on the same server. A partition can also span nodes—even to the point of having all four nodes serving one OS.) Each node can be managed independently by software.

For example, a server can continue to run an operating system in one node while you install and test *another* version of that operating system, or a different operating system entirely, in another node on that server—all without having to take the entire server offline. Multiple operating systems can function on the same server without interfering with one another.

Physical partitioning includes three modes: *Fixed*, *static* and *dynamic*.

- *Fixed* partitioning is done while the system is powered off, and involves the cabling together (or uncabling) of two or more physical nodes to modify the partitioning. After recabling, the operating system must be restarted.

[3] For much more information about XpandOnDemand scalability, refer to the "Introducing Enterprise X-Architecture Technology" white paper mentioned previously.

- *Static* partitioning requires only the nodes being adjusted to be taken offline. The remaining nodes in the server remain unaffected and continue to operate normally. Static partitioning is performed on node or system boundaries. This means that a partition must have the hardware to function independently (processor, memory, I/O, etc.). It also means that one node can't be subdivided into multiple partitions, but one partition can consist of multiple nodes. Partitioning is done by accessing the offline server from a remote system running systems management software (such as IBM Director) before restarting the operating system. Due to the lack of support for more flexible (dynamic) partitioning in current operating systems, this is the type of partitioning that will be available initially on Enterprise X-Architecture based servers.
- *Dynamic* partitioning has the same hardware boundaries as static partitioning. However, it permits hardware reconfiguring (adding or removing hardware) while the partition's operating system is still running. Servers based on the Enterprise X-Architecture design provide hardware support for dynamic partitioning. This capability requires extensive operating system modifications to support online insertion and removal of resources (essentially, plug-and-play for 4-way processor complexes and individual nodes). The Windows and Linux operating systems do not yet support this capability.

Figure 1 illustrates some sample partitioning implementations at the node level using Enterprise X-Architecture scalable servers. Other configurations are possible.

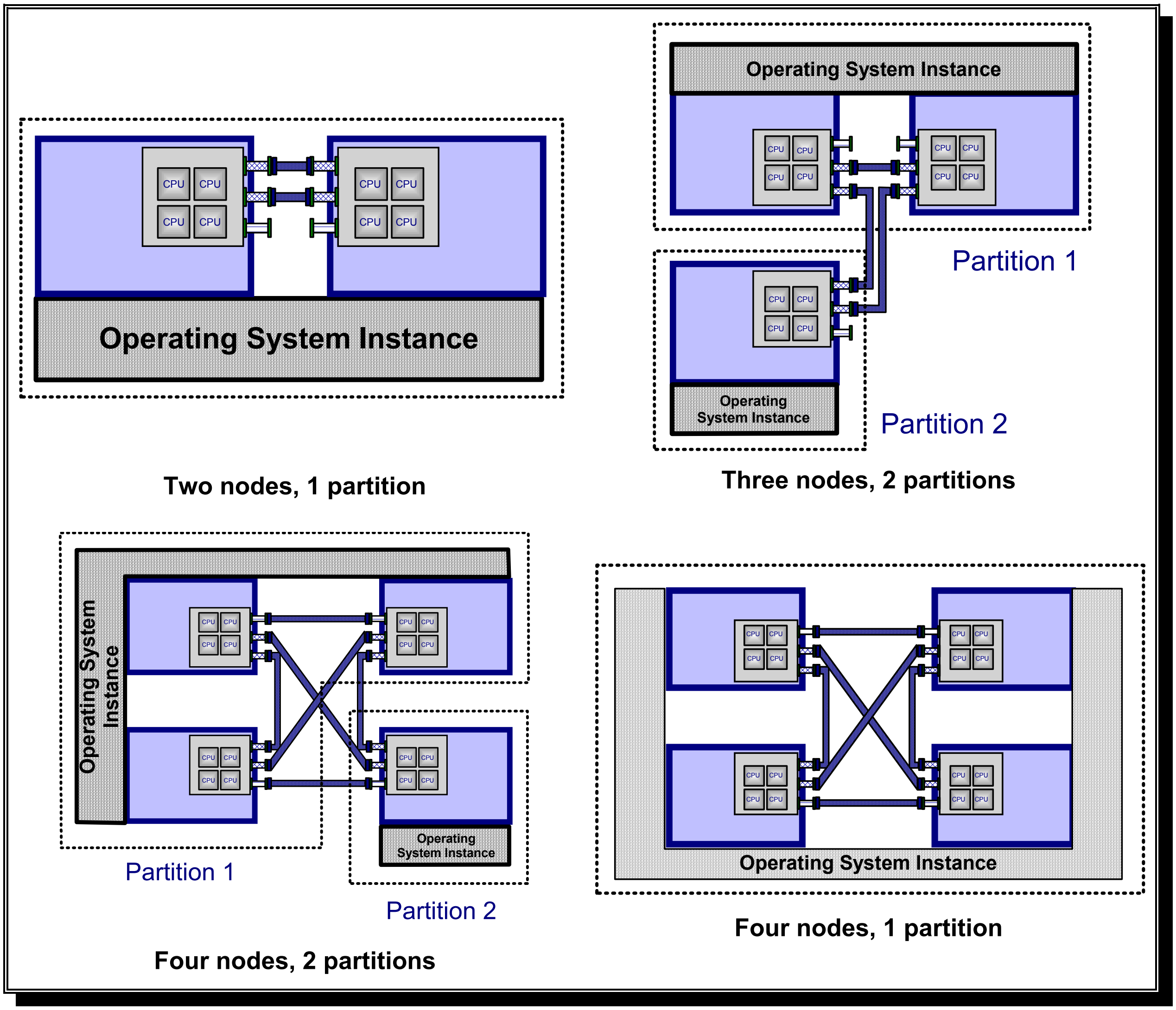

CPU
Operating System Instance
Two nodes, 1 partition
Operating System Instance
Partition 1
Operating System Instance
Partition 2
Three nodes, 2 partitions
Operating System Instance
Operating System Instance
Partition 1
Partition 2
Four nodes, 2 partitions
Operating System Instance
Four nodes, 1 partition

Figure 1. *Partitioning examples*

And that's just the beginning. Enterprise X-Architecture servers running ESX Server today, or future versions of Windows or Linux, can also implement *logical* partitioning, with even higher levels of flexibility and granularity in running concurrent operating systems than physical partitioning provides. (Logical partitioning is the type used on IBM @server zSeries—formerly S/390®—mainframes.)

- Using logical partitioning, administrators can easily partition a multinode complex at the individual processor level (with associated memory, I/O and other required resources) or even lower—*multiple partitions per processor*—without shutting down and restarting the hardware and software. More hardware can be added, or removed for maintenance, without powering the system off. When workload demands change, you can also reassign resources from one logical partition to another without having to shut down and restart the system. A simple user interface guides you through the appropriate steps.

Once your partitions have been created and the software installed, you are afforded an even greater degree of administrative control using IBM Process Control, which allows you to assign tasks to specific processors. Using the VMware software on an Enterprise X-Architecture server gives you the benefits of both hardware-enabled and software-based partitioning.

If you intend to consolidate servers, system partitioning offers many benefits. As stated earlier, with system partitioning the multiple operating systems previously used by multiple servers could all be running simultaneously on one server in one location, rather than being scattered across an organization. (Or many older, slower, single-processor and 2-way servers could be replaced by a few high-speed 4-way scalable enterprise nodes.) System partitioning allows you to eliminate the need to use multiple servers to support different operating systems within your business. One server with system partitioning could act as an application server, both for your marketing department that runs on Windows and for your engineering department that requires a Linux server. Or you can have most users hosted by Microsoft Windows 2000 Advanced Server, while you have a small development group beta testing another operating system in another partition. Or the full power of all the nodes could be applied to one operating system instance.

To illustrate, compare the results of having four separate 4-way servers in a cluster versus one system consisting of four 4-way nodes (assuming that the individual servers and the nodes are configured similarly): Let's say that each node and each stand-alone server has four processors, 8GB of memory, six adapter slots, two hard disk drives, one floppy drive, one CD-ROM drive, one SCSI controller, one Gigabit Ethernet controller, one systems management processor, one copy of an OS and one copy of whatever applications and utility software are installed.

In the case of stand-alone servers clustered together, there would still be four separate servers, with the capability for data sharing and failover—each with its own memory, adapters, software, etc. By comparison, the four-node Enterprise X-Architecture server could look like *one* 16-way server running *one* copy of the OS and applications, with access to 32GB of available memory, 24 adapter slots, eight hard disk drives, four floppy drives, four CD-ROM drives, four SCSI controllers, four Ethernet controllers, and four systems management controllers. In other words, it would be one megasystem with incredible expandability (not even counting the remote I/O capability offered by RXE-100 Remote Expansion Enclosures[4] attached to Enterprise X-Architecture systems). Alternatively, those four nodes could be configured as two 8-way systems (with 16GB of memory, 12 adapter slots, four hard disk drives, etc., apiece), with failover. You have the freedom to start out with the nodes configured one way, and then change the configuration later as needed. Some simple cabling changes (and software installation/deinstallation) and you have a whole new server configuration. Remember, each node can be a separate partition, or a partition can include several nodes.

[4] An RXE-100 expansion unit includes either six or 12 high-speed PCI-X adapter slots. For more information, go to http://**ibm.com**/eserver/xseries for the "Enterprise X-Architecture and Remote I/O on @server xSeries Servers" white paper. From the xSeries home page, select **Library** for links to the different types of documentation available.

If you prefer a clustered solution, the Enterprise X-Architecture design works well in that environment too. You can start out with two 4-way nodes in a cluster, perhaps running Microsoft Cluster Service (MSCS), in a high-availability failover configuration. Each node would have its own operating system and applications. Over time your requirements might increase to the point where you need to add more computing power to the cluster. Just as in the integrated multinode Enterprise X-Architecture configurations, you can cluster a series of 4-way nodes in various combinations and still get the benefits of the 3.2GBps SMP Expansion Port throughput. For example:

- If you are in a high volume, high-availability technical/scientific computing environment, you might prefer a four-node, 4-way, *four-partition* cluster for a small Beowulf-class supercomputer configuration to perform massively parallel computing operations.
- A database environment running IBM DB2® Universal Database™ or Oracle might use a configuration of four 4-way nodes under a *single* partition working as a single 16-way server.
- Another alternative might be a three-node, 12-way single partition running the database as a back end server, clustered with a one-node (4-way) server running a front-end application.

Enterprise X-Architecture technology gives you the ability to configure your system to best match current demands, while providing unprecedented flexibility to accommodate future needs.

Software Tools for System Partitioning

As seen in the preceding sections, hardware-based partitioning offers great functionality for enterprise-class servers. To simplify the management of Enterprise X-Architecture system partitioning, IBM offers IBM Director with System Partition Manager and IBM Process Control.

But what about all the servers you might have in your organization that *don't* offer these hardware capabilities? Are you out of luck? Absolutely not. Not only can VMware ESX Server be used in conjunction with hardware-based partitioning to combine the best features of both partitioning methods, it also offers much of the same flexibility for less powerful servers. Likewise, IBM Process Control can be used with both partitioning methods to control how applications access system resources within partitions.

IBM Director

(In January 2002, IBM will strengthen IBM Director's industry-leading systems management capabilities by adding Linux support for many of its previously Windows-only features, as well as many other enhancements.) With IBM Director[5] 3.1, xSeries servers provide you with the most sophisticated and easy-to-use local and remote systems management tools. IBM Director, built upon industry standards, is a powerful suite of tools and utilities that is included with xSeries servers. IBM Director is designed to manage servers in the Intel environment and support a variety of operating systems, including Microsoft Windows NT, Windows 2000, Windows XP, Windows 98 and Millennium Edition, IBM OS/2, Novell NetWare, Linux and SCO UnixWare. Director supports a veritable alphabet soup of industry standards such as DMI, CIM, WMI, SNMP, TCP/IP, IPX, SNA, NetBIOS, SLIP, XML and HTTP, among others.

IBM Director offers a graphical user interface for easy local and remote access and control and smooth integration into higher levels of workgroup or enterprise management tools, including :

- BMC Patrol (new in IBM Director 3.1)
- Computer Associates Unicenter
- HP OpenView

[5] Read the IBM Director white paper at http://**ibm.com**/eserver/xseries for more information. From the xSeries home page, select **Library** for a list of the types of documentation available.

- Intel LANDesk™ Management Suite
- Microsoft System Management Server (SMS)
- NetIQ (new in IBM Director 3.1)
- Tivoli® management software
- TNG Framework

By letting IT administrators view the hardware configuration of remote systems in detail and monitor the usage and performance of critical components, such as processors, disks and memory, IBM Director can help you manage your server with ease and efficiency. More importantly, it can help you control many of the hidden costs of operation.

IBM Director 3.1 adds support for IBM Enterprise X-Architecture[6] remote I/O via the IBM RXE-100 Remote Expansion Enclosure. In addition, several CIM-related enhancements include:

- CIM instrumentation for Linux
- Mass configuration of client CIM properties — Saves time by setting up and configuring multiple remote systems as a group, rather than having to touch each system individually
- Hardware instrumentation via CIM — Enables RAID and systems management hardware information and alerts to be passed up to higher-level management packages as part of the IBM Director upward integration modules (UIMs)

Another enhancement is configurable system updates. This enables customizable (i.e., destination drive) distribution of updates to IBM Director. And, of course, IBM Director 3.1 adds support for the latest xSeries, IntelliStation®, NetVista™ and ThinkPad® hardware.

IBM Director consists of a management server, the management console and agent and a portfolio of tools for advanced server management:

Management server

The management server is the heart of IBM Director, providing the application logic and current system-related management information stored in an integrated, centralized SQL database for easy access, even when the system in question is not available. The management server provides discovery of remote systems, presence checking, security and authentication, management console support and a persistent store of inventory information in its built-in SQL database. If preferred, the management server can be configured to use any of the following database products:

- IBM DB2® Universal Database™ 5.2 (or later)
- Microsoft SQL Server 6.5 or 7.0, SQL Server 2000, Access 2000 or Microsoft Data Engine (MSDE) 1.0
- Oracle 7.3.4 through 8.1.7

The management server runs on Windows NT, Windows 2000 or Windows XP Professional.

Management console and agent

The management console is a Java™-based graphical user interface for performing administrative tasks. It provides comprehensive hardware management via drag-and-drop or a single click. A scrolling "ticker tape" status bar on the bottom of the console window allows the monitoring of system attributes without the user having to open a separate console. (If desired, multiple management consoles can be opened simultaneously.) IBM Director 3.1 adds a color-coded system health status monitor. This provides rapid, at-a-glance management to ascertain the health status of managed systems. IBM Director 3.1 also adds agents for Linux distributions

[6] Go to http://**ibm.com**/eserver/enterprise_xarchitecture for the "Introducing Enterprise X-Architecture Technology" white paper.

based on the v2.4 kernel. This allows you to manage, monitor and receive alerts from systems running Caldera, Red Hat, SuSE or TurboLinux.

All system-specific data gathered by the management console is stored in the management server SQL database. The management console runs on Windows NT, Windows 2000, Windows 98 and Windows XP Professional.

Server Extensions

IBM Director server extensions prolong the manageability of your server hardware throughout its life cycle to help administrators configure, deploy, manage and maintain IBM @server xSeries servers easily and effectively. IBM Director Server Extensions include Capacity Manager, Cluster Manager, Management Processor Assistant, Rack Manager, Software Rejuvenation, RAID Manager and System Availability. IBM Director 3.1 adds Linux support for all server extensions except Cluster Manager.

- **Capacity Manager** — Capacity Manager monitors critical server resources such as processor utilization, disk capacity, memory usage and network traffic. Using advanced artificial intelligence, it identifies bottlenecks for an individual system, a group of systems (new in IBM Director 3.1) or a cluster and recommends upgrades to prevent diminished performance or downtime. Capacity Manager can even identify latent bottlenecks and make recommendations for preventive action.

  For example, Capacity Manager can predict hard disk drive and memory shortages that might cause problems for your systems. Because Capacity Manager features can help you predict problems *before* they occur, the administrator can perform proactive planning and—if necessary—schedule service and upgrades before potential problems degrade performance.
- **Cluster Manager** — Cluster Manager allows an administrator to easily identify, configure and manage clustered servers using one graphical tool. Administrators can be alerted via pager or e-mail about cluster events (in hardware, the operating system and Microsoft Cluster Service [MSCS]). Alternatively, Cluster Manager can trigger recovery programs or others automatically.
- **Management Processor Assistant** — The Management Processor Assistant (MPA) tool, formerly called Advanced System Management, offers exceptional control of systems by letting you monitor critical subsystems as well as restart and troubleshoot servers, even if a server has suffered a fatal error or is powered off. This utility works in concert with the management processors and adapters included with most xSeries servers or available as an option. IBM Director 3.1 adds management support for the RXE-100 remote I/O unit.
- **Rack Manager** — Rack Manager offers a drag-and-drop interface for easily configuring and monitoring rack components using a realistic visual representation of the rack and its components. It also provides detailed health status information for the rack and its elements. IBM Director 3.1 adds the ability to drag-and-drop objects *between* racks.
- **RAID Manager** — RAID Manager lets an administrator configure, monitor and manage local and remote SCSI and IDE RAID subsystems without taking the server(s) offline, avoiding costly downtime. IBM Director 3.1 adds FRU (field replaceable unit) number reporting in alerts for RAID components and hard disk drives. This can reduce labor and service costs by providing replacement part information in the alert message so that the correct part is sent with the service call.
- **Software Rejuvenation** — In networked servers software often exhibits an increasing failure rate over time, due to programming errors, data corruption, numerical error accumulation, etc. These errors can spawn threads or processes that are never terminated, or they can result in memory leaks or file systems that fill up over time. These effects constitute a phenomenon known as "software aging," which can lead to unplanned server outages.

  Advanced IBM analytical techniques allow IBM Director Software Rejuvenation to monitor trends and *predict* system outages based on the experience of system outages on a given server. Alerts of this sort act as Predictive Failure Analysis for software, giving an administrator the opportunity to schedule servicing (rejuvenation) at a convenient time in advance of an actual failure and *avoid* costly downtime. Software Rejuvenation can be scheduled to reset all or part of the software system with no need for operator intervention. When Software Rejuvenation reinitializes a server, the server's software failure rate returns to its initial lower level because resources have been freed up and the cumulative effects of numerical errors have been removed.

When Software Rejuvenation is invoked within a clustered environment, cluster management failover services (such as Microsoft Cluster Services and Microsoft Datacenter Server) may be used to gracefully stop the offending subsystem and restart it on the same or another node in the cluster in a controlled manner. In a clustered environment, xSeries servers can be set to failover to another server, then be reset by IBM Director without downtime.

IBM Director 3.1 adds a Trend Viewer feature to graphically monitor the software aging process.

- **System Availability** — System Availability accurately measures uptime/downtime for individual servers or groups of servers, and provides a variety of graphical views of this information. This enables you to track the improvements in your server availability in order to validate the benefits of the systems management processes and tools. IBM Director 3.1 adds the ability to distinguish between planned vs. unplanned outages.

IBM Director 3.2, planned to be available later in 2002, will add the ability to distinguish between a hardware node and a partition, so that systems management hardware alerts can be pinpointed by both partition and node. (*Figure 2* shows how the IBM Director 3.2 console may look with the Enterprise X-Architecture System Partition Manager. Note the system partitioning tasks.)

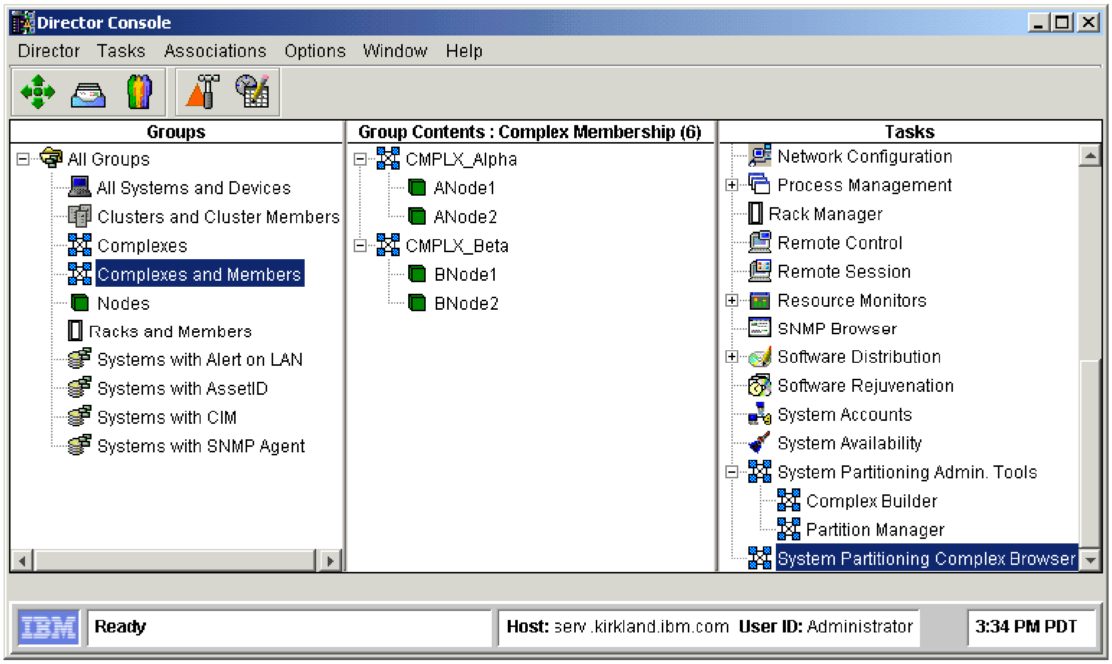
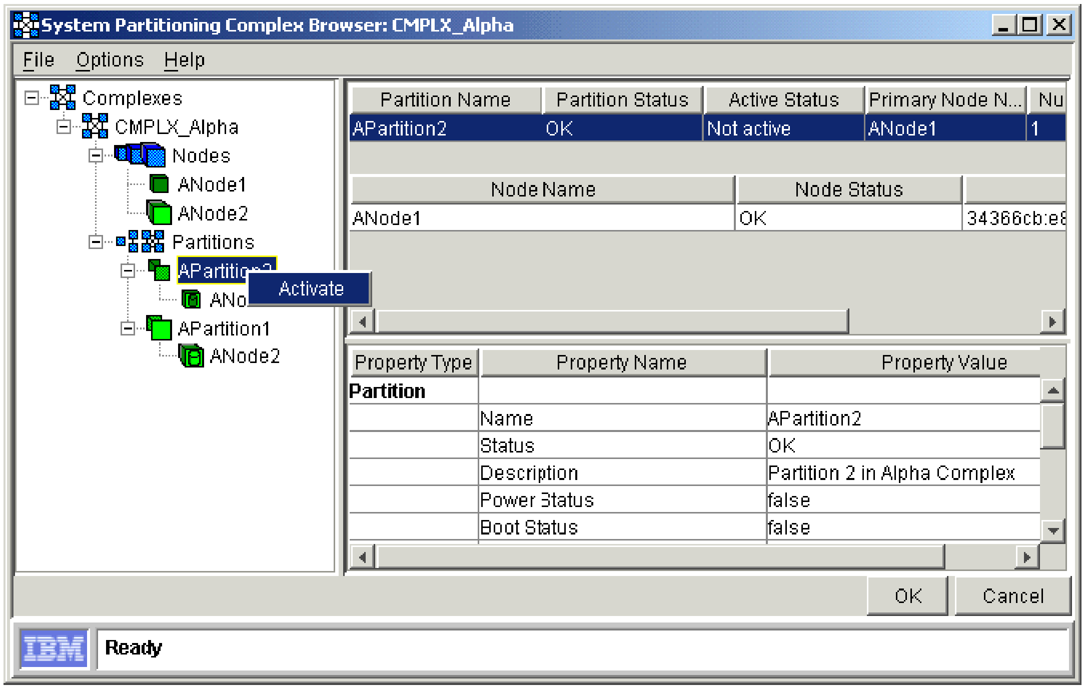


Figure 2. *IBM Director console*

IBM Director 3.2 will also add a new server extension to simplify the management of Enterprise X-Architecture hardware partitioning:

System Partition Manager

System Partition Manager provides a graphical interface for creating hardware partitions (initially static partitions, with other types to come in time). It allows an administrator to configure a specific server (while it is offline) from a remote system, prior to starting the OS. System Partition Manager uses the network link to the onboard systems management processor or adapter to establish the relationships among nodes. These relationships are maintained in a persistent database and can be recalled and activated at any time using the graphical interface. Because System Partition Manager integrates with IBM Director, it is part of a common management infrastructure that is used to manage a running partition. (*Figure 3* shows an early version of System Partition Manager.)



Partition Name	Partition Status	Active Status	Primary Node N...	Nu
APartition2	OK	Not active	ANode1	1

Node Name	Node Status	
ANode1	OK	34366cb:e8

Property Type	Property Name	Property Value
Partition		
	Name	APartition2
	Status	OK
	Description	Partition 2 in Alpha Complex
	Power Status	false
	Boot Status	false

Figure 3. *System Partition Manager*

VMware ESX Server

As we have shown in the preceding sections, hardware partitioning offers a lot of flexibility in server implementation and consolidation. But what if you need more partition granularity than hardware partitioning provides (currently at the node level), or if you have—or are considering—IBM servers that don't offer Enterprise X-Architecture scalability? This is where software-based partitioning comes into play.

VMware ESX Server is virtual machine software for consolidating and partitioning servers. It is a cost-effective, highly scalable virtual machine platform with advanced resource management capabilities. Ideally suited for high-performance environments, such as corporate IT and service provider data centers, VMware ESX Server is used to minimize the total cost of ownership (TCO) of server infrastructure by maximizing server manageability, flexibility and efficiency across the enterprise.

Dynamic logical partitioning

ESX Server works by letting you transform physical computers into a pool of logical computing resources. Physical servers are partitioned into secure virtual servers. Operating systems and applications are isolated in these multiple virtual servers that reside on a single piece of hardware. These resources can then be distributed to any operating system or application as needed, when needed.

VMware ESX Server provides dynamic logical partitioning. It runs directly on your hardware to partition and isolate server resources, using advanced resource management controls to let you remotely manage, automate and standardize these server resources. VMware ESX Server gives you mainframe-class control of your server infrastructure through dynamic logical partitioning, which involves:

- **Partitioning sever resources** — ESX Server acts as the host operating system, provides dynamic logical partitions to hold other operating systems and virtualizes most system resources, including processors, memory, network capacity and disk controllers.
- **Isolating server resources** — With ESX Server, each hosted OS *thinks* it owns the entire computer, yet it sees only the resources that the administrator (through ESX Server) assigns to it. (As shown in *Figure 4*, ESX Server resides between the hardware and the various operating systems and applications.) Because each partition is completely isolated from every other one on the system, it's—from a software standpoint—a completely separate server. Any viruses or other security violations that might affect one partition will have no effect on the others. If an OS in one partition locks up, the others are unaffected. Partitions can be administered remotely, even down to the BIOS level, just as individual servers are.
- **Managing server resources** — ESX Server's advanced resource management controls allow you to guarantee service levels. Processor cycles can be allotted on a time-share basis. Memory can be assigned dynamically based on partition workloads and defined minimums. If the allocated amount is insufficient in one partition, ESX Server can temporarily borrow memory from one partition and lend it to another, and then restore it to the original partition when needed. Network sharing is determined by token allocation or consumption based on the average or maximum bandwidth requirements for a partition.

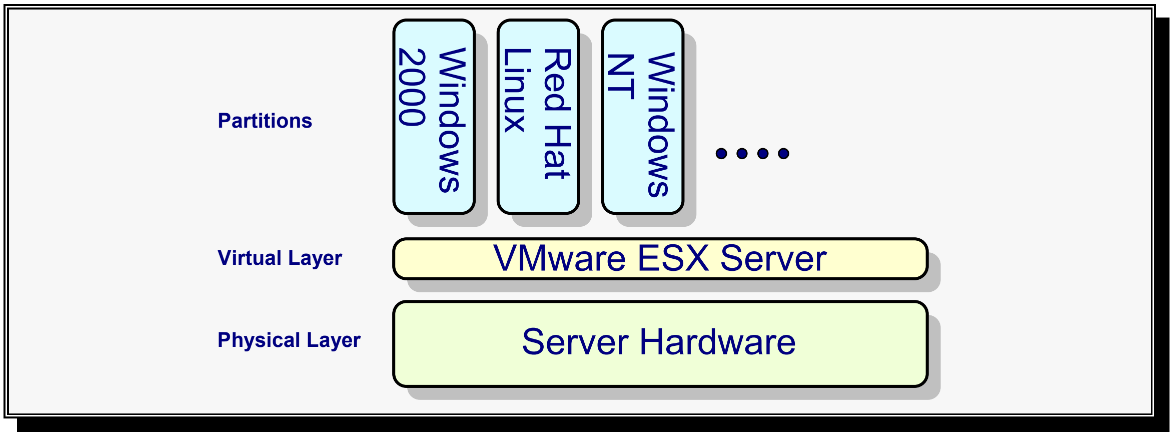


Figure 4. *ESX Server resides between your server hardware and your server resources, allowing you to partition, isolate, and manage server resources with optimal flexibility and scalability.*

Creating a uniform platform

VMware ESX Server works by running directly on your hardware to provide a secure, uniform platform for deploying, easily managing and remotely controlling more servers at a lower cost while saving space.

One of the administrative costs of having a variety of servers is maintaining a number of standard software "images," one for each server model, to simplify the setting up of new servers. This maintenance is time-consuming and expensive. Each image has to include the appropriate video device drivers and other system-specific software features unique to a server model, even if the operating system and application suite are identical on all servers. When you buy a new server tomorrow, even if it is from the same vendor as the last one, you may need to create a new software image for it. If instead you simply create a new partition on an existing server, the image needed will be identical because the hardware will be unchanged.

ESX Server creates a uniform platform from which to partition and provision virtual machine servers. When created in virtual machines, the needed software images will be identical because the hardware will be unchanged. ESX Server allows you to copy entire virtual machines (partitions) to quickly and easily provision new ones—even across a network. You can even *move* a partition image to another server. Because ESX Server virtualizes the system hardware, the operating system images can be identical on *all* servers (that is, one image per operating system type and version), rather than requiring a unique software image for each *server* type and model. When considered in the light of potentially adding hundreds of servers over the course of a year, the cost- and time-saving benefits of ESX Server can be tremendous.

Used on an Enterprise X-Architecture server with, for example, four 4-way nodes (forming a 16-way server), the updated version of ESX Server (available in the second half of 2002) will be able to span nodes. This would allow you to create a 7-way partition, a 6-way partition and a 3-way partition if that's what you need. Later you could reduce those partitions to a 6-way, a 5-way and a 3-way, and add a 2-way partition—or combine them into two 8-ways or many other possible configurations. (*Figure 5* shows how hardware and software partitioning work together.)

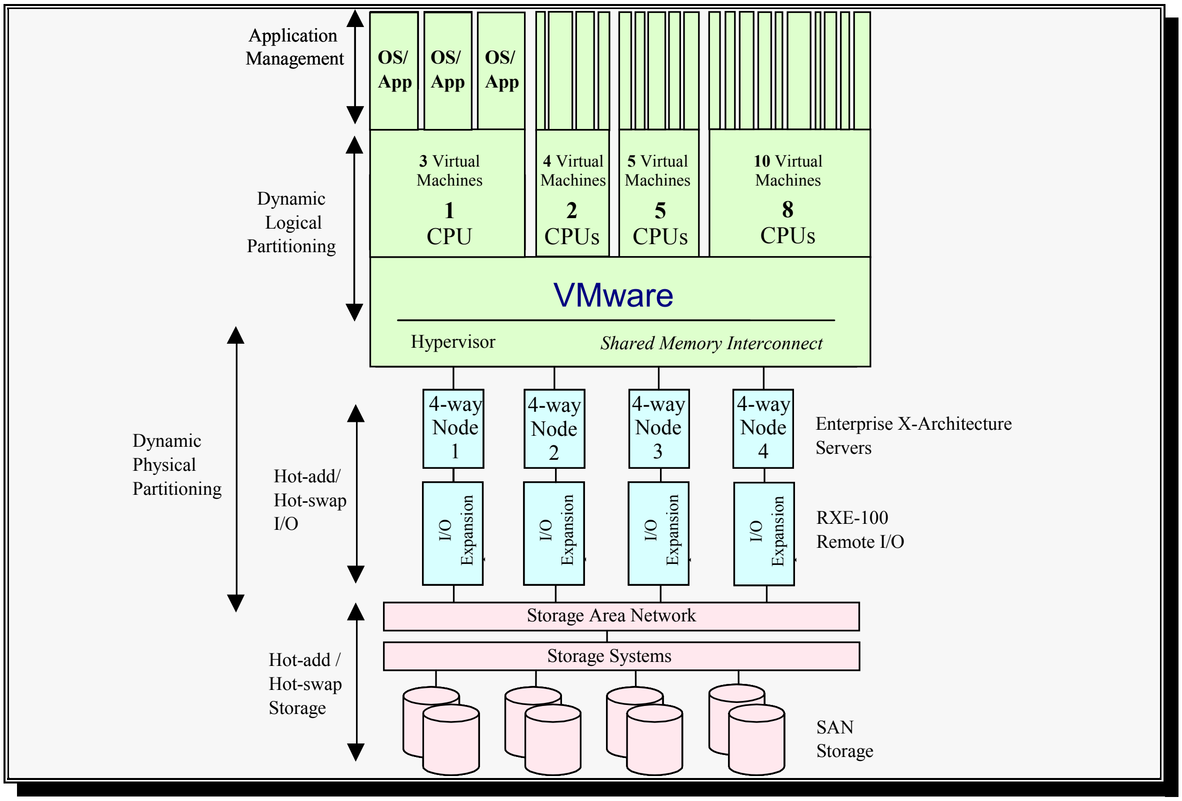


Figure 5. *ESX Server software partitioning running atop Enterprise X-Architecture hardware partitioning*

Implementing server consolidation

You would like to consolidate all those underutilized uniprocessor servers you have scattered around your enterprise onto fewer highly scalable, highly reliable enterprise-class servers. However, these servers need to be partitioned in order to consolidate their workloads. ESX Server allows administrators to set up multiple partitions per *processor*, rather than the current Enterprise X-Architecture hardware maximum of one partition per 4-way *node*, providing much more flexibility in server configurations.

For example, if you have eight uniprocessor servers whose peak usage per server rarely exceeds 25% of the processor's capabilities, you could set up one 2-way server with eight partitions (four per processor) running the same applications as before and utilizing 70-100% of the processor cycles. Or you might use a 4-way server and consolidate perhaps 16 uniprocessor servers into one.

The benefits of having one centrally located server instead of eight or 16 geographically diverse systems include:

- Potentially lower costs for hardware, or better hardware for the same price
- Reduced power consumption and air conditioning
- Less floor or rack space needed
- Fewer cables to deal with
- Simpler administration

ESX Server's dynamic resource management capabilities allow you to host any combination of the following operating systems in separate partitions:

- Microsoft Windows NT Server, Terminal Server and Enterprise Server 4.0 (with ServicePack 3 or later)
- Windows 2000 Server and Advanced Server
- Red Hat Linux 6.x and 7.x

Support for Red Hat Linux 7.2 and FreeBSD 4.3 is coming in early 2002, in ESX Server 2.0.

Any combination of these operating systems can be running concurrently on one system. For example, say you have an 8-way x370 server using 900MHz Intel Pentium III Xeon™ processors. You could partition each processor into two virtual machines, allowing you to run 16 separate partitions with all of the following workloads:

- Six partitions to running Windows 2000 Advanced server in support of your programming staff (allowing them to debug six programs at once on the same server)
- Five more for running legacy applications with Windows NT Server
- Three for running Linux applications
- Two for beta testing applications on Windows 2000 Server

With less taxing applications, such as print and fax serving, you might be able to carve out 20 or more partitions without running out of processor cycles and memory. Plus, if your application requirements change, you can dynamically increase or decrease the size of the processor "slice" allocated to a partition, or add and delete partitions, as needed.

If you use Siebel software, you know that it requires three separate servers: one as a Siebel server, one for a Siebel gateway, and one as a database server. If, for example, you are using dual-processor 1U servers (and therefore an aggregate of six processors in 3U of space), you could substitute a single 3U server, such as the 4-way x360, partitioned into three or more virtual servers. Similarly, if you are running Microsoft IIS on six 1U uniprocessor servers, you could replace them with one 4-way 3U server (x360) with six partitions, saving yourself 3U of rack space.

Delivering high availability

ESX Server allows you to deliver services and deploy new solutions faster and more efficiently from a stable, uniform platform that protects critical data in secure, isolated virtual machines. ESX Server offers a number of features that promote high availability:

- Because a frozen partition can be taken down and restarted without affecting the other partitions, ESX Server can reduce or eliminate server downtime caused by software lockups—unlike a stand-alone server, which would be offline for a time.

- Clusters of partitions can be formed, just as clusters of servers are. This is called *one-box clustering*.
- Multiple “hot standby” virtual machines can be created to provide N+1 failover.

Partition failover is a much more cost-effective solution than 1-1 hardware failover, because it makes better use of server resources. This is especially useful when budgets are tight. For example, instead of failing over from six servers to six others that are standing idle until needed, those six servers could failover to six virtual machine partitions on *one* server. This would eliminate the need for five servers or free up five additional servers to use (perhaps with four of them failing over to four partitions on the fifth server). The failover server could be located nearby or offsite. Or, you could configure a four-node 16-way Enterprise X-Architecture server with dozens of partitions and have all of them failover to one or several small servers with multiple partitions set up for just that reason. You can use inexpensive models (which may be lower-performing than the primary servers) for the partition failover, because they will be used only for a short period of time, while the server with the failing partition is being serviced. Additional servers can be gradually added into the mix as they become available, along with the failover partitions, if 1-1 failover is the ultimate goal.

For a *very* low-cost solution, one partition could failover to a second partition in the *same* server, allowing the server to continue running if the operating system in the first partition freezes. The first partition can then be restarted, or even reloaded if necessary, without having to take down the server. Of course, this will not provide failover protection in the event of a hardware failure involving a processor or memory, but at least it will provide a measure of protection for software failures.

A final high-availability advantage of both VMware and Enterprise X-Architecture partitioning is that by consolidating many low-cost and older servers into a few high-performance, high-function systems you can achieve the benefits of high-availability hardware and firmware technologies that aren’t available in the older and low-end servers, with features like:

- Chipkill memory
- Memory ProteXion™
- Memory mirroring
- Active PCI or Active™ PCI-X I/O
- Light Path Diagnostics
- Real Time Diagnostics
- More hot-swap components and additional PFA-enabled parts

When combined with software technologies such as IBM Director Software Rejuvenation, Capacity Manager and IBM Process Control, these hardware features contribute to extremely high availability. But even if a partition freezes or a server goes down, the clustering/failover capabilities of Enterprise X-Architecture servers and VMware can keep your operation up and running. These features take IBM customers closer to OnForever™ levels of high availability.

The benefits of virtual blade servers

If you are a service provider, either commercially or for in-house users, you may be considering implementing blade servers. The term “blade server” refers to a server that contains a number of cards (blades), each of which holds one or more processors and associated memory, disk storage and network controllers. Each blade within a system cabinet functions as a stand-alone server, although all the blades share common power supplies, I/O slots and system ports. This ultradense server design allows a standard 42U rack to hold hundreds of processors. The advantages to this design include floor/rack space savings compared to even a 1U server design, reduced power usage and heat output, as well as a potentially lower acquisition cost versus buying individual servers. While blade servers offer high processor density relative to rack space, they require an investment in all new hardware, and server blades are still a fairly new technology and thus somewhat untested in a high-volume enterprise environment.

By contrast, ESX Server allows you to configure a conventional server into a “virtual blade server.” You can take standard servers and partition the processors, memory and other system resources into virtual blades, which give you many of the same benefits as blade servers, but without requiring you to invest a lot of money in specialized blade server hardware.

Instead of buying a blade server with 16 processors, you can use a 3U server, such as the x360, and partition its four 1.5GHz or 1.6GHz Xeon Processor MP chips into (for example) 16 virtual blades—four per processor, for the equivalent of 224 low-power processors in a 42U rack. (Apportioning each processor into *six* partitions would give you 24 per server or 336 per rack of 14 x360 servers). Or, you might take an existing rack of 42 1U x330 2-way servers and partition them into eight virtual blades per server, for an effective 336 servers. (Of course, subdividing the two 1.26GHz Intel Pentium III processors used in the x330 four ways each may not produce the same results as, for example, eight 800MHz Transmeta Crusoe processors in a blade server would in a high-performance environment, but for many tasks those 800MHz processors would be overkill anyway.)

Of course, the virtual blade approach may not produce the effective processor density or offer as much maximum processing bandwidth as an actual server blade configuration. (In other words, 84 processors divided into 336 partitions won’t provide as much raw computing horsepower as 336 individual processors would.) However, it does offer several important advantages:

- You can use a mixture of new and existing conventional servers—which reduces your initial outlay—while adding blade servers incrementally, if desired.
- You have the flexibility to dynamically turn standard servers into virtual blade servers and back again as your needs change.
- Provisioning new partitions can be accomplished quickly and dynamically, without taking the system down. Partitions can be replicated quickly, even between different systems. When your needs change these partitions can be reconfigured dynamically.
- Even though the blade servers use low-power processors, they use a lot of them. The power usage and heat output of a conventional server, with few processors logically partitioned into many, may be lower.
- To keep the size of the blades as small as possible, blade servers are typically limited to one or two 2.5” notebook-type hard disk drives. This limits the maximum capacity of a blade to under 100GB of storage (using two drives) and means a maximum of 5,400rpm speeds on the drives. Because a standard server can have access to *hundreds* of high-speed (10,000-15,000rpm), high-capacity 3.5” or 5.25” drives through SCSI RAID arrays, Fibre Channel storage area networks (SANs) and network attached storage (NAS) servers, each virtual blade could have virtualized access to hundreds of gigabytes of storage. Plus, standard 3.5” drives tend to be much less expensive than 2.5” drives for the same capacity.
- If you’re a service provider who has many customers (internal or external) with simple computing needs, having hundreds of “fractional processor” partitions in a rack can be a much more cost-effective means of servicing those customers.

System requirements

In order to perform its partitioning mastery, ESX Server requires about a 10% performance overhead (mostly for I/O, which isn’t virtualized) and additional memory. This may sound high at first, but consider that ESX Server may enable you to raise the effective utilization of a server from less than 50% up to 90% or more. Viewed in that light, ESX Server’s overhead is trivial. Go to http://www.vmware.com/products/server/esx_specs.html for detailed specifications. In the current version of ESX Server (V1.1), a partition may use an entire processor or a portion thereof. The 2.0 version of ESX Server, shipping in early 2002, supports up to 4GB of RAM and one processor per partition (up to 64GB of RAM and eight processors per *server*). A version of ESX

Server that will support up to *four* processors per partition and 64 processors per server is planned to ship in the second half of 2002. Currently, ServerProven® testing for ESX Server has been completed successfully for the x350 and x370; testing for the x330 and x360 is under way. Other servers are planned to be added to the ESX compatibility list over time.

A Customer Perspective

For Agilera Inc., VMware ESX Server running on IBM servers has proved to be a potent combination. "Our xSeries servers have been extremely reliable, so we were looking for a way to make better use of them," says Brent Adam, Windows Systems Engineer Team Lead. Based in Englewood, Colorado, with its primary data center in Columbia, South Carolina, Agilera is an enterprise application service provider (ASP), providing application management and managed services (hosting) for applications including J.D. Edwards, Lawson, Microsoft Exchange, Oracle, PeopleSoft, SAP and Siebel.

Agilera sought a way to reduce hardware costs and increase server utilization, while still being able to provide customers with their own domain controllers and meet their service level agreements (SLAs). It seemed like a tall order. The solution? Have servers do double duty as both domain controllers and application servers. By using ESX Server on IBM @server xSeries systems running several versions of Windows, Agilera can combine multiple tasks on the same server with outstanding results. Several domain controllers are configured to failover to multiple partitions on one server for high availability—an essential feature, given Agilera's 24/7 operation.

Agilera also found that partitioning allows multiple in-house software developers to concurrently use the same server as a testing lab for all of their development efforts. This reduces development costs and simplifies the configuring of test beds. They can set up a J.D. Edwards lab in one partition, a test database in another and have separate partitions for various prerelease application versions—all running on one server.

Agilera is especially pleased with VMware's dynamic partitioning features ease of making changes. Agilera is able to bring up a new server from scratch in less than an hour, and is better able to meet the SLAs using ESX Server. "Any ASP that doesn't use VMware partitioning with IBM servers is crazy," concludes Adam.

IBM Process Control

Developed by IBM, Process Control will enable an administrator to control how multiple applications access a server's resources. (Process Control was provided to Microsoft for inclusion in Windows 2000 *Datacenter* Server; however, it is *available only from IBM* for the Windows 2000 *Server* and *Advanced Server* products.) Rather than letting applications demand as much memory and processor cycles as they want, administrators can set specific limitations on these and other system resources. (IBM Process Control will be available in the second half of 2002.)

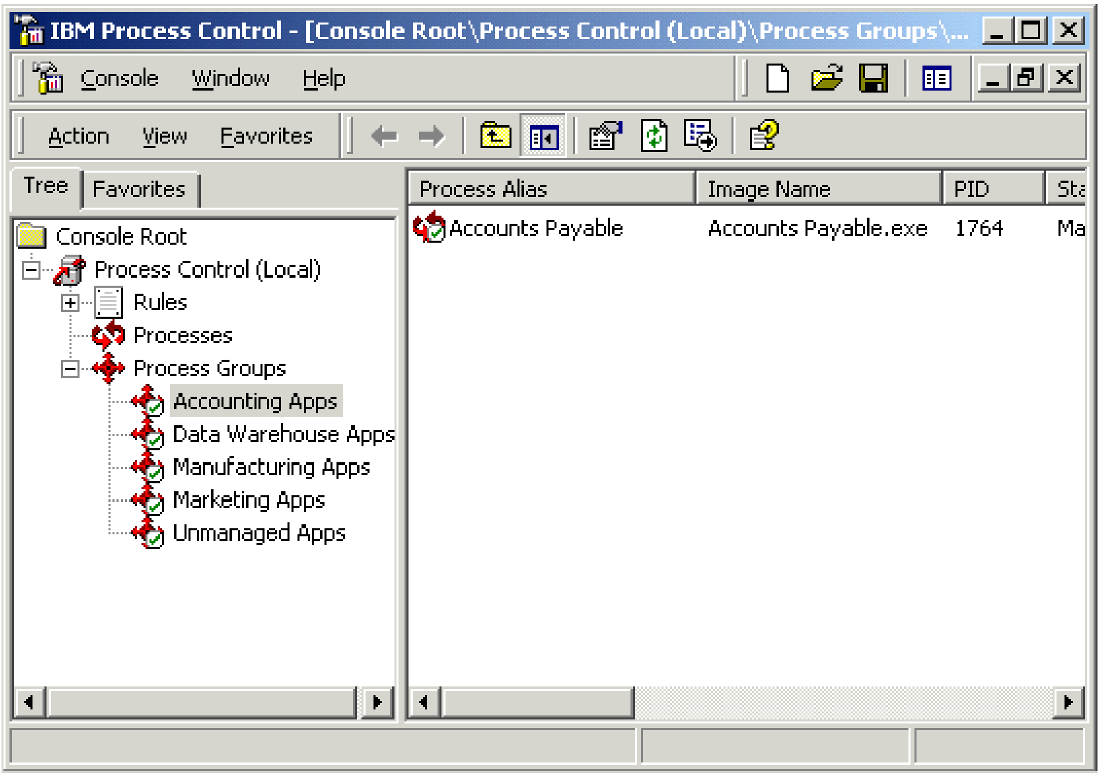
By preventing greedy applications from dominating server resources, IBM Process Control can help improve performance both for the server overall and for application users in general. Unfortunately, application vendors have no incentive to impose restrictions on their software. Consequently, some of them can be real resource hogs. One application may have its priority set unnecessarily high, to the detriment of all other applications. Or it may lock down the serial port or other resources so that no other application can use them. Or two applications may each try to use all available memory, causing contention. These types of ill-behaved applications make it virtually impossible to run many applications concurrently on a conventional server.

In response to this situation, IBM Process Control provides administrators with tools that allow the application of fairness rules, based on business needs, to programs running on a server by:

- Assigning affinities to achieve server partitioning

- Assigning priorities to rank applications relative to one another
- Assigning scheduling classes to differentiate within priorities
- Enforcing processor time limits to kill runaway processes
- Enforcing memory commitments to limit real and virtual memory consumption
- Limiting the number of processes running concurrently

Process Control is a no-charge snap-in for Microsoft Management Console (MMC) that offers both command line and graphical interfaces. (*Figure 6* shows the graphical interface.) It complements the Windows 2000 Task Manager and the System Monitor (without trying to replace either), as well as IBM Director; future versions will integrate IBM Director and IBM Process Control.



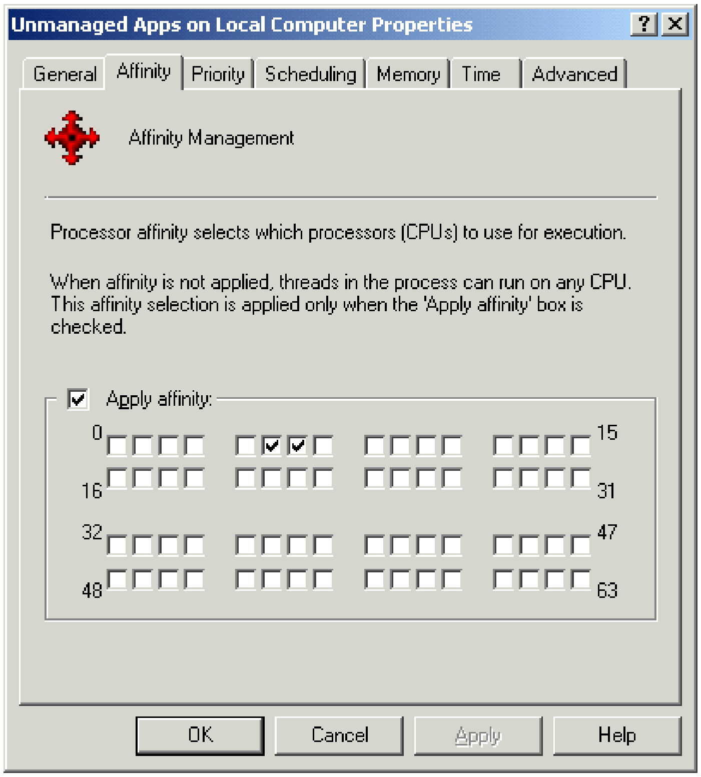
Process Alias	Image Name	PID	Sta
Accounts Payable	Accounts Payable.exe	1764	Ma

Figure 6. *IBM Process Control console*

Here are some sample uses for Process Control:

- Consolidate two applications on a 4-way server and set up Process Control so that each program is assigned to two processors.
- Assign a higher priority to the current production version of an application, while a newer version runs at a lower priority during limited testing prior to deployment. This allows you to test the new version without hurting the performance of the server for the production users.
- Limit the execution time of a buggy application so that it can't get stuck in a loop and tie up processor cycles. (At the same time, you can configure Process Control to automatically send an alert to IBM Director that the process has terminated and set up an IBM Director event action plan to automatically restart the application.)
- Limit the number of concurrent copies of an application to two when more instances of that program results in excessive memory paging and disk thrashing.

Figure 7 illustrates the simplicity of assigning application-to-processor affinity:



Affinity Management

Processor affinity selects which processors (CPUs) to use for execution.

When affinity is not applied, threads in the process can run on any CPU. This affinity selection is applied only when the 'Apply affinity' box is checked.

☑ Apply affinity:

0 □□□□ □☑☑□ □□□□ □□□□ 15
16 □□□□ □□□□ □□□□ □□□□ 31
32 □□□□ □□□□ □□□□ □□□□ 47
48 □□□□ □□□□ □□□□ □□□□ 63

Figure 7. *IBM Process Control affinity management*

Conclusion

At the top end of industry-standard servers, the combination of IBM Enterprise X-Architecture hardware-based partitioning and software partitioning enabled by VMware ESX Server offers the best of both worlds: the scalability of being able to add processors and other resources to a server as needed, the flexibility of dynamic logical partitioning, and the manageability advantages of being able to apply the same software partitioning to both high-end Enterprise X-Architecture servers and lower-function servers. Beyond that, IBM Director and IBM Process Control provide manageability and fairness rules within each partition for greater control of each multipartition server.

The combination of hardware and software partitioning with Process Control offers a number of time- and money-saving capabilities for enterprises and service providers alike, including the ability to:

- Increase capacity without buying new servers (by recovering unused processor cycles)
- Reduce floor/rack-space requirements by consolidating many uniprocessor and 2-way servers into fewer 4-way and larger servers
- Reduce hardware costs long-term by using fewer servers overall (a few large servers instead of scores of small servers), or by finding new uses for old servers by running multiple applications on a formerly single-use server
- Increase system availability by using a few highly available enterprise-class servers as opposed to many low-cost, lower-function servers

- Further increase availability by using Process Control and partition segregation, as well as hardware and partition clustering and hot-spare partition failover (among partitions on one system or on several)
- Increase availability still further by using IBM Director Software Rejuvenation to predict and prevent software failures, and IBM Director Capacity Manager to predict and prevent hard disk drive and memory shortages
- Solve Windows and Linux software scalability problems by using an 8-way server as eight (or more) uniprocessor systems
- Reduce floor and rack space requirements
- Reduce power and air conditioning expenses
- Reduce management and administration costs by using a few centralized servers rather than multitudes of small ones scattered about the organization
- Facilitate multiple operating system coexistence and increased security through partition segregation
- Dynamically reallocate processor cycles, memory and other system resources to partitions as needs change
- Dynamically assign software to specific processors, set processor and memory limits, and set program execution scheduling and priorities by using Process Control
- Manage and track server resources used and bill users accordingly
- Guarantee levels of service to users

For an IT organization that is a) trying to control expenses, b) trying to rein in server proliferation c) in need of cost-effective onsite or offsite failover, or for a service provider looking to improve customer services, the trinity of hardware partitioning, software partitioning and systems management tools is an unbeatable combination.

Additional Information

Visit our Web site at http://**ibm.com**/eserver/xseries (or call **1-888-SHOPIBM**) for more information on IBM @server xSeries server direction, products and services. From the xSeries home page, select **Library** for a list of the types of documentation available. For more information about VMware ESX Server, visit *http://www.vmware.com/products/server/esx_features* (or call 1-877-4VMWARE). Go to *http://www.agilera.com* for more information about Agilera products and services.

IBM Server Group
Department 4VNV
Research Triangle Park NC 27709

Produced in the USA
12-01
All rights reserved

IBM reserves the right to change specifications or other product information without notice. IBM makes no representations or warranties regarding third-party products or services. References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates. IBM PROVIDES THIS PUBLICATION “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions; therefore, this statement may not apply to you.

IBM @server xSeries servers are assembled in the U.S., Great Britain, Japan, Australia and Brazil and are composed of U.S. and non-U.S. parts.

This publication may contain links to third party sites that are not under the control of or maintained by IBM. Access to any such third party site is at the user's own risk. IBM is not responsible for the accuracy or reliability of any information, data, opinions, advice or statements made on these sites. IBM provides these links merely as a convenience and the inclusion of such links does not imply an endorsement.