Server Consolidation Technologies – A Practical Guide

A Leostream White Paper

Contents

Executive summary

Server consolidation describes a variety of ways of reducing capital and operating expenses associated with running servers. Gartner research divides consolidation projects into three categories: *logical consolidation* means implementing common processes and management across a range of server applications; *physical consolidation* means collocating servers in fewer locations; *rationalized consolidation* means implementing multiple applications on fewer, more powerful platforms.

The main reasons why companies undertake server consolidation are to *simplify management* by reducing complexity and eliminating "server sprawl"; *reduce costs*, particularly staff costs but also hardware, software and facilities costs; and to *improve service*. Data on the ROI of server consolidation projects is hard to come by but anecdotal evidence from big companies indicates that typical savings run into millions of dollars. A server consolidation project may also provide the opportunity to improve scalability and resilience (including disaster recovery) and consolidate storage.

Rationalized consolidation usually involves *workload management* – a set of techniques that enable several applications to run together on a single instance of an operating system, or *partitioning* – the division of a server into several smaller systems each of which runs its own copy of an operating system. Partitioning is a more effective technique for server consolidation involving "badly-behaved" applications, i.e. those which do not co-exist well with other applications on a single operating system. Many common business applications fall into this category.

Workload management depends on *processor binding* – forcing different applications to run on different processors within a multi-processor server or *software-based resource allocation* – the use of special programs or features within the operating system which allow resources such as processing power, memory and I/O bandwidth to be allocated on a priority basis. The latter approach provides much greater granularity and a more dynamic response to changing workloads.

There is no standard implementation of workload management for UNIX so different vendors have implemented their own solutions. IBM, HP and Sun all provide software-

based resource allocation as well as processor binding on their UNIX ranges. Microsoft provides processor binding as part of Windows 2000 Server and companies such as IBM, HP and Aurema have developed software-based resource allocation packages for Windows 2000 but these are less sophisticated than the tools available for UNIX (in general, workload management on Windows is less successful because the operating system was not designed with this objective in mind). There is also a lack of advanced workload management tools.

Partitioning can occur at three different levels within a server: the *hardware*, *logical* and *software* levels. In hardware partitioning, each partition has one or more processors and a block of memory but all the partitions share, to some extent, the I/O components. Hardware partitions are electrically isolated from each other so a fault in one partition does not affect any of the others. On UNIX servers from major vendors, such as IBM, HP and Sun, the hardware can be partitioned dynamically (i.e. without stopping the operating systems) but Windows and Linux do not yet support his feature.

Blade servers offer an alternative approach to hardware partitioning. A blade server comprises several thin server modules that sit side by side in a chassis. The chassis provides high speed I/O capabilities for all the modules and so reduces the amount of cabling required in the data center. Blade servers are also supplied with management software that simplifies a number of server administration tasks. Each manufacturer's blade system is proprietary but the major vendors are expected to launch blades for each architecture that they sell so, for example, it may be possible to run IBM xSeries, pSeries, iSeries and zSeries servers together in a single chassis.

Logical partitioning uses a layer of hardware microcode or firmware to enable a single processor to run more than one partition. Intel's architecture does not readily support logical partitioning so this technology is only available from vendors like IBM and HP who have servers based on their own chip technology.

Software partitioning was originally developed in the 1960s to permit timesharing on mainframes but commercial products are now available for Intel-based servers from companies like VMware and Connectix. These products use Windows or Linux as a *host* operating system running on the physical server and allow several *guest* operating systems to run on top of the host in *virtual machines*. It is possible to run several different guest operating systems (e.g. different versions of Windows, Linux, FreeBSD, Novell, OS/2 and Solaris) simultaneously on the same server. Software partitioning creates more overhead than other partitioning techniques, particularly when a host operating system is used, but it has the benefits of being inexpensive and easy to implement while providing dynamic resource management.

The use of data center automation tools makes server consolidation more scalable, and delivers other benefits, such as more effective disaster recovery, improved system availability and higher peak load capacity. The Leostream Server Controller 1000 is one such tool designed for consolidation of Intel-based platforms. It provides a central management, access control and monitoring system for systems which use VMware and Connectix software as well as blade servers.

Features of the product include *cataloguing* of system images, *cloning* of images between local and remote servers, *customization* of cloned images, *monitoring* of applications, host and guest operating systems, *access control* so that guest machines can be managed by their owners while the underlying host machines are managed by a central administrator and *fail-over* in the event that guest or host machines stop responding. With the Leostream Server Controller 1000 it is possible to manage hundreds of virtual machines and look after thousands of system images. It is even possible to store data relating to a complete network of servers, enabling it to be restored in minutes for disaster recovery purposes.

Looking to the near future the key developments affecting server consolidation will be the advent of 64-bit processors from Intel and AMD, the launch of the first products resulting from Microsoft's .NET strategy, including Windows .NET Server 2003 (the successor to Windows 2000 Server) and the availability of a wider variety of blade servers from an increasing number of vendors. All of these developments are expected to improve the business case for server consolidation, particularly in the so-called Wintel environment. The ultimate goal of server consolidation is *autonomic computing* in which a server network becomes a self-configuring, self-regulating, self-protecting system which mimics the human body's autonomic nervous system. This ambitious goal will not be attained for many years, but companies can take practical steps today to start moving towards it.

What is server consolidation?

There is no commonly agreed definition of server consolidation. Instead it is used as an umbrella term to describe a variety of ways of reducing capital and operating expenses associated with running servers. Gartner Research divides consolidation projects into three different types, with progressively greater operational savings, return on investment and end-user benefits, but also progressively greater risks:

- *Logical consolidation*, when there is no physical relocation of servers and the goal is to implement common processes and enable standard systems management procedures across the server applications. Often this involves putting a single, centralized department in charge of all the servers;
- *Physical consolidation*, which entails the collocation of multiple platforms at fewer locations (i.e. reduction in the number of data centers through centralization without altering the number of actual servers). Physical consolidation is referred to by Microsoft and others as location consolidation;
- *Rationalized consolidation*, which means implementing multiple applications on fewer, more powerful platforms, usually through workload management and partitioning.

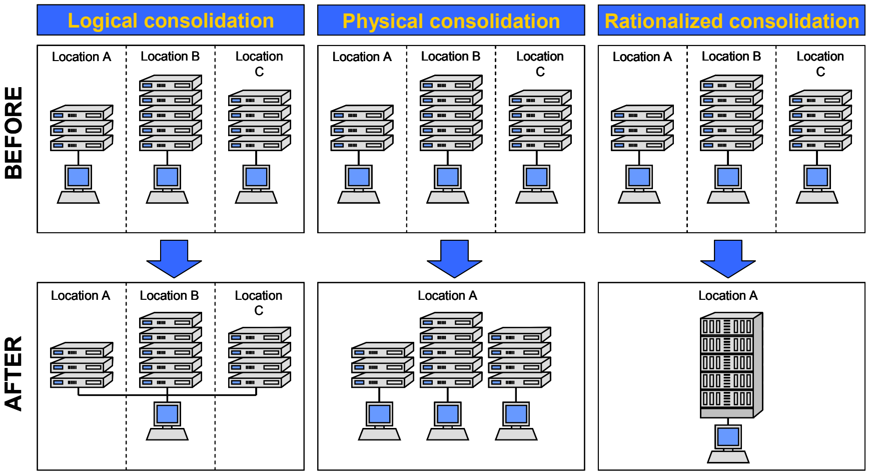


Figure 1. Three types of server consolidation

Organizations may reap significant benefits from logical and physical consolidation but these are essentially non-technological approaches. The main focus of this white paper is therefore rationalized consolidation, which is used here to include:

- *Replacing a number of small servers with fewer larger servers.* For example, a company with separate file and print servers for each department might decide to consolidate them into two large file and print servers which all the departments share. Dell calls this workload consolidation and Microsoft and others call it physical consolidation (note that this meaning of the term physical consolidation is different from Gartner's);

- *Replacing existing servers with higher density configurations*. For example, a collection of servers housed in PC towers could be replaced with 1U high rack-mounted equivalents, or a collection of 1U rack-mounted servers could be replaced with blade servers;
- *Combining several instances of the same application on a single server*. For example, four 4-processor SQL servers could be consolidated on a single 16-processor server or, if the original SQL servers are under-utilised, perhaps a single 8-processor server;
- *Combining several different applications on a single server*. For example, a company may have several applications that are only required by a small number of users, making it difficult to justify the cost of maintaining a separate server for each. In some cases it is possible to combine applications that run on different operating systems (e.g. Windows and Linux) on the same server.

A detailed discussion of data or storage consolidation is beyond the scope of this white paper but it should be noted that this will often be a feature of physical consolidation and rationalized consolidation projects. For example, if a number of servers are brought together on one site, a NAS (Network Attached Storage) device based on RAID (Redundant Array of Inexpensive Disks) technology may be used to reduce the risk of data loss for all applications. A larger consolidation project may involve the installation of a high performance, high availability storage back end based on SAN (Storage Area Network) technology.

Finally, server consolidation projects can focus on existing applications (backward consolidation), new applications (forward consolidation) or both. When planning the introduction of any new application it is a good idea to consider the opportunity for forward consolidation and for rationalizing the new application with other applications.

Why consolidate?

In a survey of Fortune 1000 companies by Forrester Research the top three benefits of server consolidation cited by respondents were simplified management, lower costs and improved service. Let's look at each of those in turn.

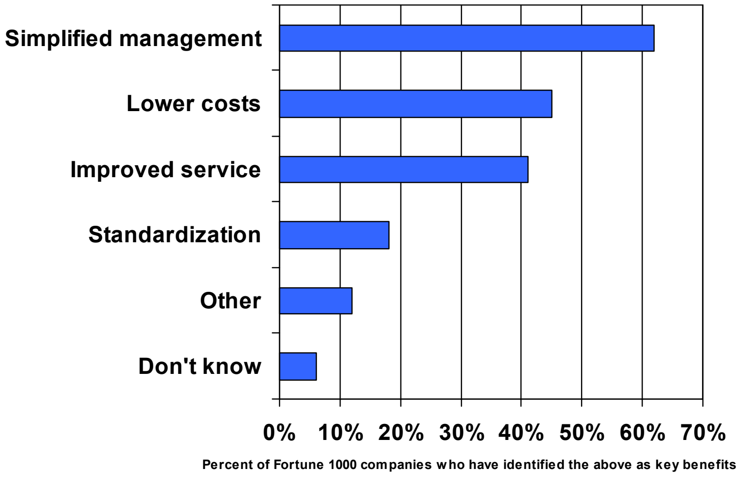


Figure 2: Reasons for undertaking server consolidation

Simplified management

In the late 1990s many enterprises expanded their IT infrastructure rapidly. Often this led to "server sprawl" – a huge proliferation of print servers, file servers, email servers, development servers and test servers throughout the organization. Server sprawl creates major challenges for IT management, particularly with regard to maintaining security, performing regular back-ups and keeping the software up to date.

Implementing consistent management tools and processes across the organization, standardizing on fewer server platforms, reducing the number of servers and concentrating them in fewer locations all help to reduce complexity. This may result in the need for fewer support staff or, alternatively, allow more projects to be undertaken and service levels to be improved without the need for additional staff. For global organizations, standardization may also permit support to be provided on a 24 x 7 basis using world-wide resources.

Consolidation also provides an opportunity to address management issues such as resilience and scalability, storage consolidation and disaster recovery (which may be impossible if servers are dispersed throughout the organization).

The introduction of fewer, more powerful servers can provide increased headroom for growth through pooling of excess capacity. Reducing the number of servers should also result in a simpler network structure that is easier to manage.

Lower costs

In the 1970s and early 1980s when enterprises ran centralized mainframe data centers, a common rule of thumb was that 80% of costs were capital and 20% were operating. Today studies by Gartner and others indicate that the ratio is more like 30% capital and 70% operating.

Server consolidation can reduce costs in a number of ways:

- *Staff costs*. Support staff are often recruited on the basis of how many servers an organisation has (with Windows NT, one member of staff per twenty servers is often used as a benchmark). Most server consolidation projects aim to reduce costs by freeing staff from mundane server maintenance tasks. Gartner suggests that more than 70% of the potential savings from a typical project will come from reduced staffing requirements, but they caution that this is usually the hardest area in which to pre-quantify savings, especially since displaced support staff often move to new posts elsewhere in the same information services organization.;
- *Hardware costs*. Consolidation can reduce costs through better server utilization, by reducing the total requirement for storage and by enabling the use of more cost-effective back-up/restore mechanisms (including in-house disaster recovery). Centralized purchasing may also enable better discounts to be negotiated with hardware suppliers. On the other hand remember that many Intel-based platforms, including 1U rack-mounted servers, are already very cost-effective because of cut-throat competition. Moving from a generic Intel-based platform to a single source platform, even one that is much higher powered, may increase hardware costs;
- *Software costs*. Consolidation may also reduce the total number of licenses needed while standardizing on fewer applications may allow better deals to be negotiated with suppliers. With many (but not all) applications the incremental cost of software licenses decreases as the number of users increases;
- *Facilities costs*. Server consolidation can reduce the amount of floor space needed for data centers. This is a particular benefit if one or more existing locations are full. Bear in mind however, that greater power density and cooling capacity are required – a 42U cabinet filled with low-power blade servers might consume three times as much power, and hence generate three times as much heat, as the same cabinet filled with 1U rack-mounted servers.

Improved service

Improved service should be a natural consequence of simplified management and the deployment of more resilient server, storage and back-up technology. It should also be possible to deploy new applications more quickly if they do not have to be loaded onto a large number of physically dispersed servers, while platform standardization should reduce the number of unforeseen delays which occur during deployment. Application development and customization may also be speeded up through server consolidation. Finally it is quite common to improve service levels by consolidating help desk and network support activities as part of a server consolidation project, although they can equally well be tackled separately.

The bottom line

Battery Ventures recently carried out research on the server consolidation market. They found little empirical data of hard ROI savings but uncovered some interesting anecdotal examples of savings, including:

- A Fortune 500 company has 300 servers in their IT organization handling development, test and QA functions. They believe that through rationalized consolidation of their servers they can cut this number in half, saving $5-10 million over the next five years;
- Another Fortune 500 company has 175 single processor Intel servers handling print function. Through rationalized consolidation they believe they can reduce the number of servers to about 40. They think this will save them over $5 million in capital and operating expenses over the next few years;
- A Fortune 50 company believes that one division can eliminate 600 development/test/QA servers from their server farm, saving them $2.4 million a year in operating expense;
- A multi-billion dollar software firm is consolidating 300 pre-sales servers down to 80 and believes they can save millions of dollars per year.

Technologies for rationalized server consolidation

The definition of rationalized consolidation given at the start of this white paper said that it usually involved workload management and partitioning. Like server consolidation these are rather loose terms which different vendors use in different ways. Here are the meanings that we give to them:

- *Workload management* describes a set of techniques which enable different applications to run together on a single instance of an operating system. The techniques aim to balance the resource demands that each of the applications places on the system so that all of them can co-exist. Note that, rather confusingly, some vendors refer to workload management as resource partitioning or soft partitioning;
- *Partitioning* involves the division of a server, which might ordinarily run a single instance of an operating system, into several smaller systems each of which can run its own copy of an operating system. Note that all the copies run simultaneously – this is not the same as partitioning a PC hard disk so that you can select which operating system runs when the machine boots up.

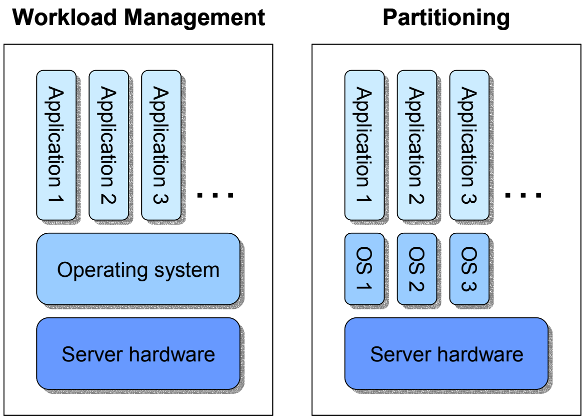


Figure 3: The difference between workload management and partitioning

In short, workload management allows several applications to run on one operating system, partitioning allows several operating systems to run on one machine. Strictly speaking, the techniques are not mutually exclusive but partitioning is commonly carried out because badly-behaved applications will not co-exist on a single operating system, i.e. when workload management cannot deliver. The problem commonly manifests itself as one application hogging all the processor cycles, or repeatedly causing the operating system to crash and thus bringing down the other applications.

The following sections discuss the workload management and partitioning solutions offered by a number of vendors. We focus particularly on solutions for Windows, UNIX and Linux, since these operating systems are the most common targets for server consolidation projects.

Workload management

Workload management techniques fall into two categories:

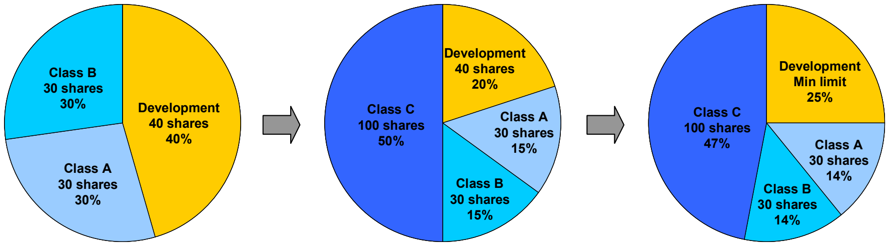
- *Processor binding*. In a server which contains several processors running the same instance of the operating system (i.e. a Symmetric MultiProcessing or SMP environment), applications can be forced to run on a particular processor or subset of processors. This is a relatively crude technique with limited granularity, since each application must be allocated a whole number of processors. It can be effective where the processing requirements of the applications are well understood but it can also lead to low processor utilization;
- *Software-based resource allocation*. In this technique software is used to allocate resources such as processing power, memory and I/O bandwidth to applications and users on a priority basis. Some implementations allow the priorities to be set in terms of service level objectives. This approach is more sophisticated and provides much greater granularity and a more dynamic response to changing workloads.

Workload management for UNIX

There is no standard implementation of workload management for UNIX so different vendors have implemented their own solutions. IBM, HP and Sun all provide software-based resource allocation as well as processor binding on their UNIX ranges.

IBM's product for its pSeries of servers running AIX (IBM's implementation of UNIX) is called AIX Workload Manager (WLM) and is supplied as part of the AIX operating system. WLM allows the system administrator to create different *classes* of service for jobs and to specify attributes for those classes. Jobs are automatically placed in a class according to the *user*, the user's *group* and the *application*. For example, when the user Tom from the group Engineers runs the CAD application, he might be placed in a class called Development. The administrator can allocate Development to one of ten *tiers* which determines the relative priority of the class (e.g. classes in Tier 0 have priority over classes in Tier 1 which in turn have priority over classes in Tier 2 etc).

Within each tier the administrator can allocate processor, memory and I/O bandwidth resources by means of *limits* and *shares*. For example Development might be allocated 40 processor shares. If all of the active classes in its tier have 100 shares between them then Development will be given 40% of the processor time. However, if another class with 100 shares becomes active Development will then only be given 20% of the processor time. The system administrator can also set limits to ensure, for example, Development is given a minimum of 25% and a maximum of 50% of the processor time. If limits are used they take precedence over shares. From AIX version 5 onwards, WLM also provides processor binding.



3 active classes.
Development has
40% of processor time

New class becomes active.
Development reduced to
20% of processor time

Development guaranteed
minimum of 25%.
Other classes scaled back

Figure 4: Example of shares and limits

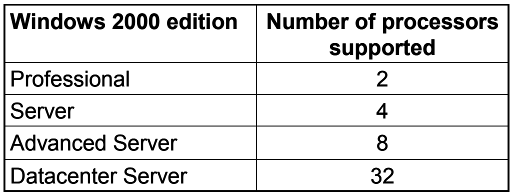
HP's mechanism for binding applications to processors is called Processor Sets (Psets) and is available on all multi-processor models in the HP 9000 server series running HP-UX11i, the latest version of HP's UNIX implementation. It allows dynamic reconfiguration of the processors in a Pset and dynamic binding of applications to a Pset or a specific processor within a Pset.

For software-based resource allocation, HP provides two packages called Workload Manager (WLM, not to be confused with IBM's product of the same name) and Process Resource Manager (PRM). PRM reserves processing power, memory and disk bandwidth within a partition for up to 64 separate applications. WLM allows the administrator to set service level objectives (e.g. response time for transactions, completion time for batch jobs) for different applications in priority order. WLM then adjusts the PRM processor settings dynamically as the workload varies to try to achieve the service level objectives.

Sun's workload management functionality is built into its Solaris operating system. Sun has offered processor binding and software-based resource allocation for some time but with the launch of Solaris 9, Sun integrates them into a new system called Solaris Containers. So far the only component of Containers that has been released is Solaris Resource Manager and this currently only supports processor management. Sun says it will introduce physical memory, swap space and I/O bandwidth management to Resource Manager shortly. The full implementation of Containers is intended to provide fault and security isolation between different applications running on the same operating system. Sun claims that Containers will thus provide many of the benefits of virtual machines (see section on Software partitioning below) with lower overhead.

Workload management for Windows and Linux

Depending on the edition (see table below) Windows 2000 will support up to 32 processors. Microsoft allows processes (instances of executing applications) to be bound to particular processors using a feature called Processor Affinity, which can be assigned in advance or set using the Windows Task Manager if the application is running. Microsoft also allows processes to be given different priority levels.



Windows 2000 edition	Number of processors supported
Professional	2
Server	4
Advanced Server	8
Datacenter Server	32

Figure 5: Multi-processor support in Windows 2000

IBM developed a more sophisticated workload management tool called Process Control which was provided to Microsoft for inclusion in Windows 2000 Datacenter Server but is only available from IBM for the Windows 2000 Server and Advanced Server products. Process Control integrates the standard Microsoft features and adds control over real and virtual memory consumption, limits on the number of copies of an application that can run concurrently, and processor time limits (so an application doesn't get stuck in a loop and endlessly eat processor cycles). Microsoft itself plans to introduce a greatly enhanced version of Process Control called Windows System Resource Manager with the higher end versions of Windows .NET Server 2003.

HP has an alternative product called Workload Management Pack (WMP) which works on all versions of Windows Server and on other vendors' hardware. WMP allows the administrator to create so-called Resource Partitions containing processors and memory and to allocate processes to the Resource Partitions. A dynamic rules engine then allows the resources within a partition to be altered depending on utilization or time of day.

Finally a company called Aurema, whose previous product forms the basis of Solaris Resource Manager, has developed a new product called Active Resource Management Technology (ARMTech). ARMTech implements the concepts of shares and limits (see discussion of IBM's AIX WLM above) for processors and memory (but not I/O bandwidth) on servers running any version of Windows 2000 Server. Microsoft has licensed ARMTech for distribution with Windows 2000 Datacenter Server and it can be purchased for use with Windows 2000 Server and Advanced Server. Aurema has also developed a version of ARMTech for Linux but it is not being actively marketed at present.

Partitioning

IBM mainframes have supported partitioning techniques for over 30 years so that multiple operating systems of different flavors could run simultaneously on the same system. Over time partitioning technology has trickled down on to smaller systems and it is now available for PCs. Partitioning can occur at three different levels within the server:

- *Hardware (or physical) partitioning* is a technique which can only be applied to servers with multiple processors. Each partition has one or more processors and a block of memory dedicated to it but the partitions share, to some degree, the disk and the I/O components of the server. Hardware partitions are electrically isolated from each other so a fault in one partition does not affect any of the others. In most cases, the allocation of resources (i.e. memory, processors and I/O paths) to hardware partitions can only be altered when the operating systems are off line. Of the three techniques, hardware partitioning generally creates the least overhead on the system;
- *Logical partitioning* uses a layer of hardware microcode or firmware (and sometimes software as well) to enable a single processor to run more than one

partition. The way in which resources are allocated to logical partitions can usually be altered without stopping the operating systems. The microcode or firmware is platform-specific and logical partitioning is only available on high end servers;

- *Software partitioning* achieves the same effect as logical partitioning using Virtual Machine (VM) software rather than microcode or firmware. The VM software acts as the master operating system, supporting the operating systems used by the applications as guests. Software partitioning is usually quite easy to implement but it tends to create more overhead on the system than the other two techniques – typically absorbing between 10% and 20% of the processor power.

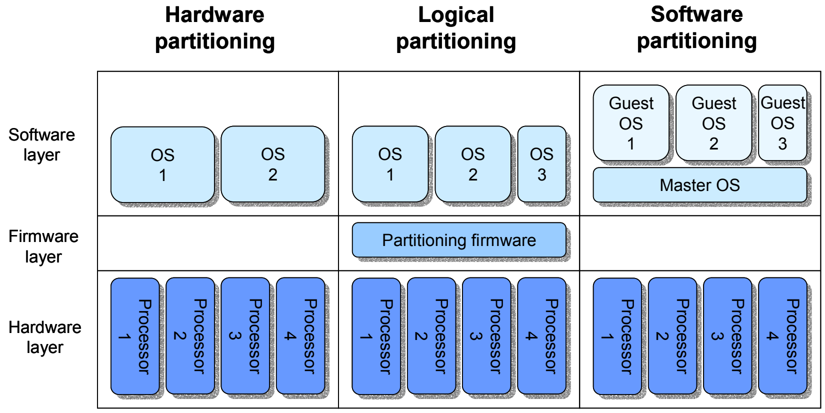


Figure 6: Three types of partitioning

Hardware partitioning

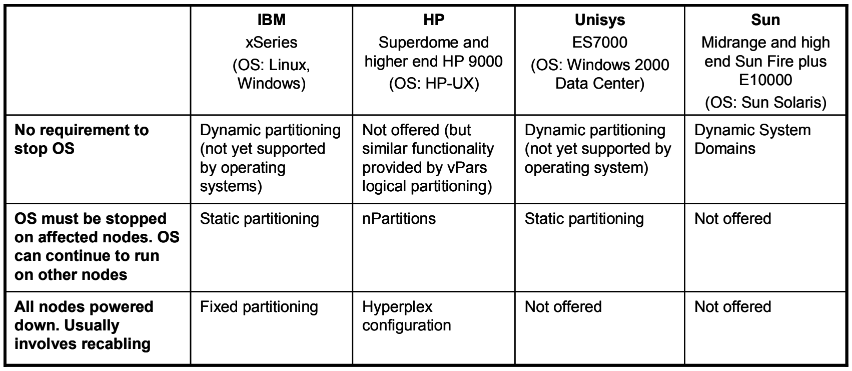
IBM supports three types of hardware partitioning on higher-end models within the Intel-based xSeries range of servers. IBM uses a four-processor node as its basic building block. Hardware partitions must coincide with node boundaries, so two or more nodes may act as a single partition but a node server may not be subdivided.

Fixed partitioning is carried out 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. Static partitioning requires the nodes being modified to be taken off line so that they can be accessed using systems management software, but the remaining nodes are unaffected. Dynamic partitioning allows changes to be made without stopping the operating system. However, while this technique is supported by the hardware it is not yet supported by Windows or Linux so it is of little practical benefit today.

HP offers two forms of hardware partitioning on its latest Superdome architecture and on other HP 9000 servers designed to run HP-UX. HP 9000 servers will work together in a configuration known as a Hyperplex. Hardware partitions within the Hyperplex consist of one or more nodes (i.e. servers). This form of partitioning is comparable to IBM's fixed partitioning. On Superdome and higher-end HP 9000 servers a technology called nPartitions is also provided. These servers are built out of cells containing up to four processors, memory and optional I/O resources and a Superdome server can comprise up to 16 cells. With nPartitions, a hardware partition consists of one or more cells. Cells are moved between nPartitions using the systems management interface. The affected partitions need to be taken off line. HP's nPartitions is comparable to IBM's static partitioning.

Unisys provides static partitioning on its ES7000 "Windows mainframe", an enterprise server designed to run Windows 2000 Datacenter Server. The processor building block in the ES7000 is known as a sub-pod and consists of four Intel processors, a third-level cache and memory. A single ES7000 can contain a maximum of eight sub-pods. A static partition can comprise any number of sub-pods, but sub-pods cannot be split between partitions. In order to move resources between static partitions, the affected partitions need to be taken off line and subsequently rebooted. Unisys also supports a feature that the company calls soft partitioning but which is, in reality, a workload management feature enabling the processors within a static partition to be assigned to different applications. ES7000 soft partitioning makes use of the processor affinity feature in Windows 2000 Datacenter Server. Like IBM's xSeries, the ES7000 will also permit dynamic partitioning once this is supported by the operating system.

Sun's hardware partitioning technology is called Dynamic System Domains (DSDs) and is available on Sun Fire "Midframe" and high end servers and the Sun Enterprise 10000 server. As the name suggests, DSDs deliver dynamic hardware partitioning today but only with Sun's Solaris operating system. The building block for DSDs is Sun's Uniboard processor and memory board which is configured with two or four Sun UltraSPARC processors. There are however, additional constraints on the number of DSDs that a particular server will support. For example, the Sun Fire 6800 server will take up to six Uniboards but the number of DSDs is limited to four.

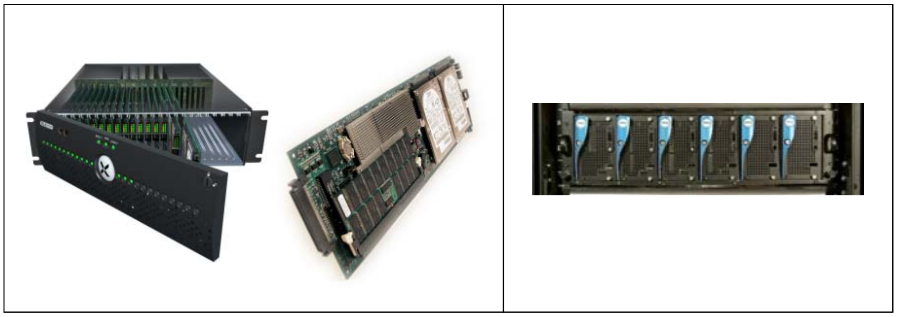


	IBM xSeries (OS: Linux, Windows)	**HP** Superdome and higher end HP 9000 (OS: HP-UX)	**Unisys** ES7000 (OS: Windows 2000 Data Center)	**Sun** Midrange and high end Sun Fire plus E10000 (OS: Sun Solaris)
No requirement to stop OS	Dynamic partitioning (not yet supported by operating systems)	Not offered (but similar functionality provided by vPars logical partitioning)	Dynamic partitioning (not yet supported by operating system)	Dynamic System Domains
OS must be stopped on affected nodes. OS can continue to run on other nodes	Static partitioning	nPartitions	Static partitioning	Not offered
All nodes powered down. Usually involves recabling	Fixed partitioning	Hyperplex configuration	Not offered	Not offered

Figure 7: Hardware partitioning terminology and availability

Blade servers offer an alternative approach to hardware partitioning. At present, they cannot do anything as sophisticated as the systems described above, but the technology is evolving rapidly and more sophisticated techniques are likely to emerge.

The first blade servers were launched by start-ups like RLX Technologies in late 2000 and consist of open, single board computers complete with memory and often one or two hard disk drives that slot into a chassis. The chassis provides high speed I/O capabilities that are shared between the blades. Blade servers were originally promoted for their high packing density (RLX can fit 24 blades in a 3U chassis) and low power consumption (the first RLX blades used Transmeta Crusoe processors for this reason) but it turned out that customers were more interested in the servers' management features and the reduction in cabling. Consequently, more recent offerings from mainstream vendors like HP, IBM and Dell use standard Intel processors and have lower packing densities but more robust mechanical designs (i.e. the blades are enclosed in metal cases). Each vendor's blade system is proprietary which locks customers in and, it could be argued, allows vendors to obtain higher margins than they could on standard Intel-based rack-mounted servers.



RLX System 300ex and ServerBlade 800i

Dell PowerEdge 1655MC

Photos courtesy of RLX Technologies, Inc. and Dell Computer Corporation

Figure 8: Comparison of first and second generation blade server designs

The management systems for blade servers allow each blade to be stopped, started and rebooted from a central location. Some systems have a hard disk on each blade so that the system image is tightly linked to that blade, but others allow the used of shared disk space so that any system image can be run on any blade. This gives system administrators great flexibility – a server can change its role from email server to web server in a matter of minutes and if one blade fails, the management system can restart the application on another blade. At present blade servers only support Windows and Linux operating systems, but Sun has announced a blade server that will run Solaris and HP is expected to introduce HP-UX support on its higher end pClass system.

Logical partitioning

Intel's architecture does not readily support logical partitioning so this technology is only available from vendors like IBM and HP who have servers based on their own chip technology.

IBM's product is called LPAR (simply an abbreviation of Logical PARtioning) and is available on the company's iSeries (formerly AS/400) and pSeries (formerly RS/6000). On the iSeries, the base partition must run IBM's OS/400 operating system but the other partitions can run OS/400 or Linux. Processors can be shared between partitions (including the base partition) but I/O paths cannot. The latest version of LPAR for the iSeries is dynamic, i.e. it allows resources such as processor, memory and interactive performance to be moved between partitions without taking their operating systems off line. On the pSeries LPAR is less sophisticated: processors cannot be shared between partitions and although dynamic LPAR is available in partitions running new versions of AIX, Linux partitions must be taken off line to perform logical partitioning on them.

HP's equivalent of LPAR is called vPars (short for virtual partitions) and is available on its medium to high end HP-UX (i.e. UNIX) servers. In reality, vPars is a technology that straddles the boundary between logical partitioning and software partitioning since the partitions are created and managed by virtual partition monitor software. Processors and I/O paths cannot be shared between vPars but the technology is dynamic so resources can be moved without having to reboot the partitions affected.

Software partitioning

Software partitioning technology was originally developed in the mid 1960s as a way of allowing many users to share the power and resources of a mainframe computer without affecting each other. The first commercial implementation was the VM operating system for IBM mainframes, which was the result of research projects carried out at MIT in Cambridge. VM has proved to be remarkably durable – 35 years on the latest incarnation, z/VM, is still widely used on IBM zSeries mainframes. In the last few years researchers have examined how software partitioning techniques can be applied to PC operating systems. This has resulted in commercial products for Intel-based servers from companies like VMware and Connectix.

The name VM was derived from Virtual Machine because the system gives each user the impression that they have access to the underlying physical machine. Here's how it works. The heart of any virtual machine technology is a special piece of software called the Control Program or Virtual Machine Monitor (VMM). The VMM controls the system hardware and determines what instructions are executed and when. It also manages the creation and maintenance of virtual machines each of which can run an application or a guest operating system.

The application or guest operating system executes a set of instructions that are categorized as *privileged* or *non-privileged*. Privileged instructions are those that could affect users of other virtual machines, e.g. instructions that attempt to alter the contents of a hardware register. The VMM allows non-privileged instructions to be executed directly on the hardware but it intercepts privileged instructions and executes them itself or emulates the results and returns them to the virtual machine that issued them.

Processors that are designed to be virtualizable, such as those on mainframes, can operate in distinct privileged and non-privileged (also called user) modes. The VMM runs in privileged mode and the virtual machines run in user mode. The processor only allows privileged instructions to be executed in privileged mode so if a virtual machine issues a privileged instruction it is automatically trapped and control of the processor is passed back to the VMM.

Unfortunately, Intel's x86 architecture is not fully virtualizable so a VMM running directly on PC hardware would not be able to trap all the privileged instructions. VMware and Connectix solve this issue by running their VMMs on a host operating system. Connectix products use various versions of Windows as the host operating system. VMware uses Windows or Linux as the host. In the case of VMware's top-of-the-range ESX Server the package contains the VMM embedded in a Linux kernel (where it executes with less overhead) so there is no need to install a separate host operating system.

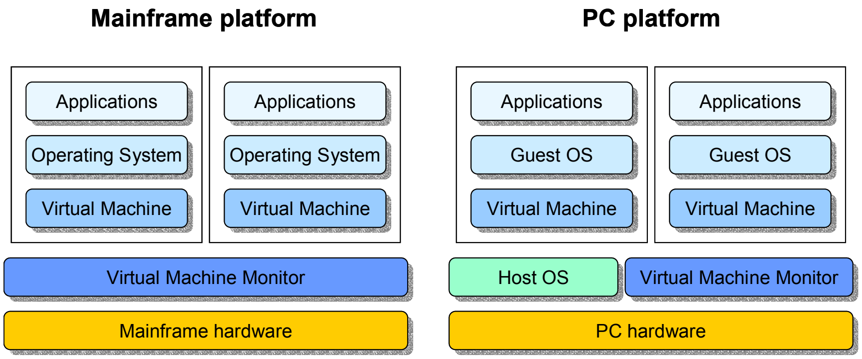


Figure 9: Comparison of virtual machine software on a mainframe and a PC

Using a host operating system to run a VMM on a PC solves another problem – how to provide drivers for the huge variety of hardware and I/O devices that are available for the PC platform. The virtual machine provides the guest operating system with a series of virtual interfaces for common devices. The VMM takes data from the virtual interface and sends it to the real interface at the appropriate time using the host operating system's hardware driver. One difference between VMware and Connectix is that VMware virtual interfaces are based on generic virtual hardware while Connectix emulates real hardware.

There is no reason why the virtual and real interfaces need be the same. This means that the virtual machine environment can be constant and independent of the underlying hardware, which brings two benefits. Firstly, system images (comprising operating systems and applications) can simply be copied between VMs, rather than requiring installation. Secondly, the virtual machine environment can emulate peripherals for obsolete operating systems, such as IBM's OS/2, enabling OS/2 to run on modern hardware even though there are no OS/2 drivers available for the hardware.

The following is a more detailed explanation of virtual machine software for Intel platforms. The description is based on VMware Workstation but the other products work in similar ways. When Workstation is installed it creates three components – the VMX driver, the VMM and the VMware Application. The VMX driver is installed within the host operating system, thereby obtaining the high privilege levels that are permitted to drivers and that the VMM requires. When the VMware Application is executed it uses the VMX driver to load the VMM into the memory used by privileged applications, such as the operating system. The host operating system is aware of the VMX driver and the VMware Application but it is ignorant of the VMM. The system now contains two "worlds" – the host world and the VMM world. When the guest operating systems are running purely computational programs, the VMM world communicates directly with the processor and memory. When an I/O function (such as disk access) needs to be performed, the VMM intercepts the request and switches to the host world. The VMware Application then carries out the request using standard host operating system calls and returns the data to the guest operating system through the VMX driver, the VMware Application and the VMM.

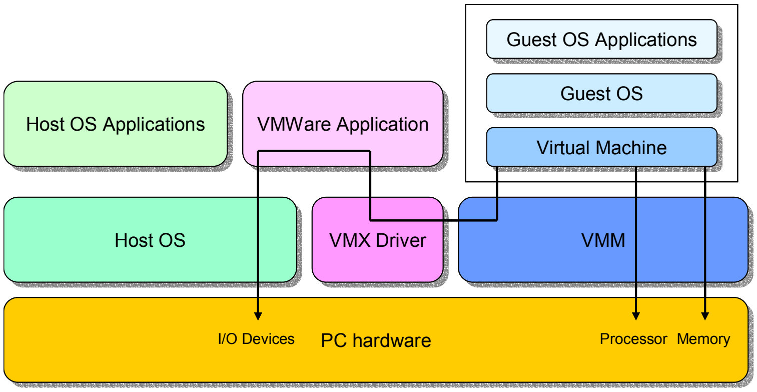


Figure 10: VMware Workstation architecture

Each time a world switch is performed all the user and processor state information needs to be stored, so performance is lower than it would be with a real system or a mainframe VMM. How much lower depends on the I/O intensity of the application, and the utilization of the server. Workstation's VMM tries to minimize the loss of performance by analyzing the I/O requests and determining which are actually moving data and which are simply checking the status of I/O ports, and emulating the latter. Performance is also affected by the fact that the VMware Application and guest operating systems run as applications under the host operating system and therefore run the risk of being swapped out of memory at inconvenient times.

Despite these limitations, virtual machine software for Intel-based servers is proving to be a valuable method of server consolidation. The major benefits are:

- *Low cost.* The software itself is inexpensive and commodity hardware can be used (or existing Intel-based server hardware re-used);

- *Wide range of guest operating systems*. Support is available for virtually every operating system that can run on a PC including Windows server software from NT3.1 onwards, Linux, FreeBSD, Novell, OS/2 and Solaris;
- *Quick and easy to install*. No recabling is required. Existing applications and operating systems are not touched;
- *Dynamic resource management*. Once installed virtual machine software permits dynamic reconfiguration of processors and memory to virtual machines with a fine level of granularity.

Scalable server consolidation

Converting a hundred physical servers into a hundred virtual machines running on a dozen or so physical hosts will reduce your facilities costs and may reduce management costs as well if the project includes logical or physical consolidation, but you will still have a hundred operating systems to manage separately. Similarly, replacing a hundred physical servers with a hundred blades in a single rack will reduce the amount of floor space you need and enable you to use the blade vendor's tools to simplify management but it won't alter the fundamental fact that there are still a hundred servers to be looked after.

To maximize the ongoing benefits of server consolidation you really need to consider data center automation tools as part of the project, to address day to day management issues and also help to deliver wider benefits such as more effective disaster recovery, improved system availability and higher peak load capacity.

The key benefit of virtual machines and, to a lesser extent, blade servers is that they separate the operating system from the underlying hardware. The hardware then becomes a resource pool that can be assigned to particular applications as needed. A new application draws resources from the pool, and retiring an obsolete application returns resources to the pool. Having a hardware pool makes it much easier to tackle issues like system availability and peak load capacity. The pool approach can also be extended to include a disaster recovery centre, where the available machines become another pool of resources that can be assigned to the transplanted applications.

If you only have a small number of physical servers running, say, 10-15 virtual machines it is quite feasible to manage the entire system using the tools provided by virtual machine software vendor, combined with a few custom-designed scripts to copy virtual machines from one physical server to another, and Microsoft's SysPrep tool to customize duplicated Windows images. Experience shows that this manual approach frequently breaks down if the initial small-scale system is successful and a decision is take to expand the deployment. Adding more host and guest machines, especially if some of them are sited remotely, significantly increases the management effort.

What is needed is a single management system that permits monitoring and control of the physical servers, the host operating system and the virtual machines, i.e. logical consolidation of the virtual machine environment. If however, the consolidated servers belong to different divisions, it is often necessary to leave server control with the divisional IT team. At the same time it is highly desirable to manage the underlying hardware, backup, spare capacity, and disaster recovery centrally. This implies that the management system need to offer some form of role-based access control that regulates access according to the user's role. This is no different from the user management systems in mainframes which have traditionally permitted sharing between departments.

Frequent system configuration changes may also strain a manual system of control to breaking point. Many training, QA, support and development situations require several system changes a day, and it may be necessary to co-ordinate the changes across hundreds of machines at a time. At this point automation tools become essential.

Training, QA, support and development would often like to archive thousands of system images so that a particular image can be built once and reused many times. This requires a sophisticated image cataloguing system. But image cataloguing need not be limited to individual servers. By storing both the images and the interconnection information for a network of servers it is possible to restore a completely pre-configured system in minutes.

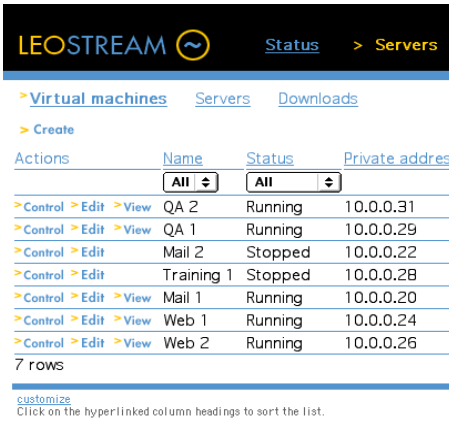
This idea of storing entire systems can be applied to Disaster Recovery, so at regular intervals the key line of business servers are backed up to tape. Recreating the entire data center in a disaster recovery facility then becomes a matter of loading virtual machine software onto the disaster recovery servers and copying the images from the tape in order to start the virtual machines.

Data center automation tools can also help to improve system reliability and peak load capability. System reliability is a key issue because consolidation increases the business impact of a failure in a host server or a blade rack. Instead of affecting a single system, a dozen or so systems may need to be recovered simultaneously. Data center automation tools can track the health of guest and host systems and automatically move the guest systems to another host if the primary host crashes. They can also provide information on available capacity system-wide, so it can be assigned when and where required.

How Leostream can help

Leostream has developed an integrated system for data center automation that makes server consolidation scaleable. The Leostream Server Controller provides a central management, access control and monitoring system for all virtual machine software supplied by VMware and Connectix.

The Leostream Server Controller performs six key functions: cataloguing, cloning, customization, monitoring, access control and fail-over.



Actions	Name	Status	Private addres
	All	All	
>Control >Edit >View	QA 2	Running	10.0.0.31
>Control >Edit >View	QA 1	Running	10.0.0.29
>Control >Edit	Mail 2	Stopped	10.0.0.22
>Control >Edit	Training 1	Stopped	10.0.0.28
>Control >Edit >View	Mail 1	Running	10.0.0.20
>Control >Edit >View	Web 1	Running	10.0.0.24
>Control >Edit >View	Web 2	Running	10.0.0.26

7 rows

customize
Click on the hyperlinked column headings to sort the list.

Figure 11: Screen shot from Leostream Server Controller management interface

Cataloguing

The Controller provides a central index of all system images spread across multiple host machines in different geographic locations.

The Leostream Host Agent runs on the host machine. When started up, it searches for system images on that particular machine and sends both the system information and a screen shot of the desktop (taken when it was last running) back to the Controller. The Leostream Guest Agent runs on the virtual machine and sends information such as the Windows Machine Name and its Domain to the Controller.

The Controller consolidates information on up to 10,000 images and allows the resulting list to be sorted and filtered according to a wide range of variables including the creator, the application and the owner.

Cloning

Virtual machine software provides the tools to copy images locally within a physical server but not between servers. The Leostream Server Controller enables images to be moved between local and remote servers. Since the images are distributed across the various host machines, it is possible to schedule many simultaneous transfers without overwhelming a central file server.

Customization

As the name implies, cloning a virtual machine results in an identical copy. For hot fail-over this is essential but for many other applications it is necessary to customize the copy so that it becomes a unique machine. The Leostream Server Controller allows customization of several properties, including Windows Machine Names and Windows Workgroups or Domains.

Monitoring

The Leostream Server Controller monitors the applications, the guest and the host machines. Should they fail to respond for a preset period of time the Controller generates an alert via email, XML-RPC or SNMP. It can also execute a recovery strategy, transferring all guest machines to another host if the host has failed, restarting a crashed guest machine and providing hot fail-over between a main and a standby guest machine.

Access control

The Leostream Server Controller provides role-based access control to the guest and host machines, so that guest machines can be managed by their owners, and underlying host machines can be managed by an overall system manager.

Fail-over

The Leostream Server Controller monitors the applications, guest and host machines. Should they fail to respond for a preset period of time the Controller can execute a recovery strategy, transferring all guest machines to another host should the host fail, restarting a crashed guest machine, or providing hot fail-over between a main and standby guest machine.

More information about the Leostream Server Controller 1000 can be found in our product sheet, available on line at www.leostream.com/product.html.

The future of server consolidation

We believe that three key technology developments will affect server consolidation over the next 12-18 months. They are 64-bit processors from Intel and AMD, advances in blade servers, Microsoft’s .NET initiative. We describe what we believe the impact will be in the following paragraphs. We also touch on a longer term vendor ambition, autonomic computing, which in many ways represents the ultimate goal for server consolidation.

64-bit processors from Intel and AMD

Intel's Itanium and AMD's Opteron are designed to compete head to head with Sun's 64-bit SPARC processors. Moving to a larger computer instruction size increases the amount of memory that a single processor can address and allows for a more powerful instruction set. There are 64-bit versions of both the Microsoft Windows .NET Server and Linux.

While the Intel Itanium processor is not backwards compatible with the current 32-bit Intel processors the AMD offering is compatible and therefore offers an easier migration path. At present however, AMD has a tiny share of the market for processors used in servers and major server vendors have yet to announce any forthcoming Opteron models.

The move by Intel and AMD to 64-bit processors will have two effects on server consolidation. Firstly, it will make the virtual machine approach more attractive by enabling more guest machines per host, and allowing all guest machines to have access to a greater peak load capacity. Secondly, servers using these processors will take market share away from Sun, and companies will have less reason to operate a mixed computing environment. Once again this will make server consolidation easier.

Advances in blade servers

One of the key advantages of blade servers over conventional servers should be that they allow the consolidation of different computer architectures into a single system. The blade servers on the market today are all based on Intel architectures but IBM, HP and Sun are all expected to offer their other processor architectures in blade form alongside their own Intel offerings. This should allow them to enjoy high margins while still delivering value to their customers by providing unified infrastructure.

The market for pure Intel-based server blade is likely to be very different. We expect that this will rapidly become a commodity business driven by low cost vendors like Dell and the Taiwanese manufacturers. Some differentiation will be provided by management systems supplied with the blades but the winners in the blade management market may well end up being third party software companies.

Microsoft .NET

Microsoft Windows .NET Server 2003, the successor to the Windows 2000 Server family, is expected to include support for dynamic hardware partitioning and far more sophisticated support for workload management, both of which will assist server consolidation in a Windows environment.

More broadly, Microsoft's .NET initiative aims to give mainstream developers in corporations the tools to write applications that work across multiple systems and geographic locations as easily as they write Visual Basic applications. This is likely to mean that a lot more complex systems will be built – and a lot more servers will be deployed. This in turn will accelerate the need for serve consolidation and for tools to manage the deployment of distributed components.

Autonomic computing

The term autonomic computing is intended to present an analogy with the human body's autonomic nervous system - the system which controls essential processes like blood flow and breathing without requiring any conscious recognition or effort. IBM describes an autonomic computing system as one that possesses eight key elements:

- It must "know itself" and comprise components that also possess a system identity;
- It must configure and reconfigure itself under varying and unpredictable conditions;

- It must always look for ways to optimize its workings;
- It must be able to recover from routine and extraordinary events that may cause some of its parts to malfunction;
- It must be an expert in self-protection;
- It must know its environment and the context surrounding its activity, and act accordingly;
- It must function in a heterogeneous world and implement open standards;
- It must anticipate the optimized resourced needed while keeping its complexity hidden.

Clearly all of this is some way off, though IBM and other companies are investing real money in autonomic computing research. Nevertheless better tools to supplement the current manual approach to server and software management represent a small step along the road to autonomic computing. Such tools have been available for some time, but companies have, by and large, not deployed them because there has not been a strong pressure to improve data center efficiency. The current recession has changed that, and faced with a hiring freeze IT directors are turning to data center automation tools to make their current teams much more efficient.

Leostream
7 New England Executive Park
2nd Floor
Burlington MA01803
USA

Tel: 617 718 1880
Fax: 617 718 1886
info@leostream.com

© Leostream, Inc. 2002. All Rights Reserved.

Leostream, the Leostream logo and roundel are trademarks of Leostream, Inc in the United States and/or other countries.

*AMD and Opteron are trademarks or registered trademarks of Advanced Micro Devices, Inc.
Aurema and ARMTech are trademarks or registered trademarks of Aurema.
Connectix is a trademark or registered trademark of Connectix Corporation.
Dell is a trademark or registered trademark of Dell Computer Corporation.
Forrester Research is a trademark or registered trademark of Forrester Research, Inc.
Gartner Research is a trademark or registered trademark of Gartner, Inc.
IBM, iSeries, OS/2, OS/400, pSeries, xSeries, z/VM and zSeries are trademarks of IBM.
HP, Hyperplex, pClass and Superdome are trademarks or registered trademarks of HP.
Intel and Itanium are registered trademarks of Intel Corporation.
Linux is a registered trademark of Linus Torvalds.
Microsoft, Visual Basic, Windows and Windows NT are either trademarks or registered trademarks of Microsoft Corporation in the United States and/or other countries .
Novell is a trademark or registered trademark of Novell, Inc.gRLX is a trademark or registered hyademark of RLX Technologies.
Sun, Solaris, SPARC, Sun Fire, UltraSPARC and Uniboard are trademarks or registered trademarks of Sun Microsystems, Inc.
Transmeta and Crusoe are trademarks or registered trademarks of Transmeta Corporation
Unisys is a trademark or registered trademark of Unisys.
UNIX is a registered trade mark of The Open Group.
VMware, GSX Server and ESX Server are trademarks of VMware, Inc.*

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