Analyze the Future

WHITE PAPER

Achieving High Availability Through IT Optimization

Sponsored by: IBM

Jean S. Bozman
January 2008

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com

EXECUTIVE SUMMARY

High availability of enterprise systems is a prerequisite for business continuity — and for sustaining services to an organization's end users and end customers. Achieving high availability (HA) during a period of rapid technology change can be challenging for many customers, who see too many obstacles to ensuring high availability across many servers — and across the enterprise.

The process of IT transformation brings new opportunities to improve high availability, especially for end-to-end applications that span the enterprise and that leverage the computing power of many servers across the network. That is because IT transformation opens the door to doing things differently — breaking down the information silos that prevented a deeper integration across business units — and a unified view of all networked servers. In so doing, there is also an opportunity to reduce server footprints via workload consolidation — resulting in more efficient computing and in reduced power/cooling costs. In the process of IT transformation, IT infrastructure is optimized so that workloads run on the platforms that support them with the best performance and the greatest efficiency.

Businesses that want to ensure that end users are able to access key business systems on a 24 x 7 x 365 basis, with little or no perceptible downtime, are studying ways to protect important applications by applying reliable server hardware and HA software to the workloads being deployed. Efficiency in operating these systems is essential to holding down operational costs (opex) associated with IT staff time, system downtime, and power/cooling for deployed systems.

IBM offers hardware and software in a comprehensive approach to ensuring high availability across a range of servers and storage systems and a range of services to protect end-to-end applications with high availability. Specifically, IBM sells hardware with reliability, availability, and serviceability (RAS) features built into the platform while selling software (directly or through partners) that provides virtualization, consolidation, and automation. Importantly, IBM provides software with advanced end-to-end systems management capability to support end-to-end applications that tap many servers across the network. This approach is aimed at ensuring that end users within the organization, and end customers, can continue to access data services across the IT infrastructure, without interruption, providing business resilience and business continuity, even in the event of outages in platforms or networks.

INTRODUCTION

High availability for systems and data is a high-priority goal for business. The business itself is solidly focused on business processes — the processes directly related to providing goods and services to end customers — and the ability to access worldwide business systems anytime, anywhere. This paper describes the relationship between that goal and the IT infrastructure that supports a spectrum of availability, as appropriate, for a wide range of applications and business systems.

In today's networked, Internet-enabled world, the bar has been raised for business expectations about what the IT infrastructure can deliver, in terms of its ability to change as business changes. And yet, the recognition of what it takes to ensure business continuity for business processes is not often considered. As advanced features for Web-enabled systems grow, supporting access from cell phones, personal digital assistants (PDAs), and PCs, one could ask: How is this end-to-end business process being supported?

This paper outlines the relationship between the business processes we see every day and the high-availability requirements for the IT systems that deliver the data and business systems on which those business processes depend. The "how-to" of assembling those IT systems is not the focus of this paper, because each organization deploys its own pattern of IT systems, based on the IT skill sets that are present within the organization, and the organization's preference for leveraging a set of IT technologies.

This paper is aimed at business managers and IT managers who are considering making changes in their current IT infrastructures, with the idea that barriers between today's islands of automation can be removed and the underlying data and systems can be linked. As this happens, it is important to ensure that these systems are protected by a combination of hardware, software, and services that ensure high availability so that the systems can be accessed whenever needed, from anywhere in the network.

Business Considerations for High Availability

Globalization, time to market, and the pressures to reduce operational costs are reasons why business needs to access IT systems in a highly reliable, available way. It can no longer be said that business considerations are not connected to, or affected by, IT considerations. Already, IDC has observed that organizations are including business leaders and IT leaders on the same system-procurement committees, in recognition that computers are the engines that enable a new business initiative, once the decision has been taken to drive business in a new direction. To that end, a new generation of applications is being written that can "tap" the data that formerly resided within specific business units throughout the organization. In effect, that data is often trapped within the IT systems that support it, with limited connectivity to systems that also need to use that data.

Today, a new series of application development technologies, including those aimed at service-oriented architecture (SOA), produce end-to-end applications that leverage data across organizations. One example of this is being able to tap Web servers that bring in Internet data, applications servers that run line-of-business (LOB) applications (e.g., ERP, CRM, HR), and database servers that hold customer data, inventory data, business intelligence (BI) data, and transactional data. This approach, spanning the network, provides end-to-end infrastructure that supports mission-critical applications in a new way — and it shows the way that IT transformation is taking place in today's enterprises. When these servers are linked together — via software, the corporate network, and the Internet — the engine for today's Web-enabled organization is ready for business.

Once this kind of end-to-end application is deployed, protecting it from any kind of interruption becomes vital. Some reasons why are as follows:

- Interruptions, or outages, in end-to-end applications and continuing access to data can disrupt operations impacting revenue

- Delayed orders, and deferred revenue, impacting profitability

- Possible loss of end customers, who may decide to switch providers

- Inaccurate "snapshot" of the current state of an organization's business

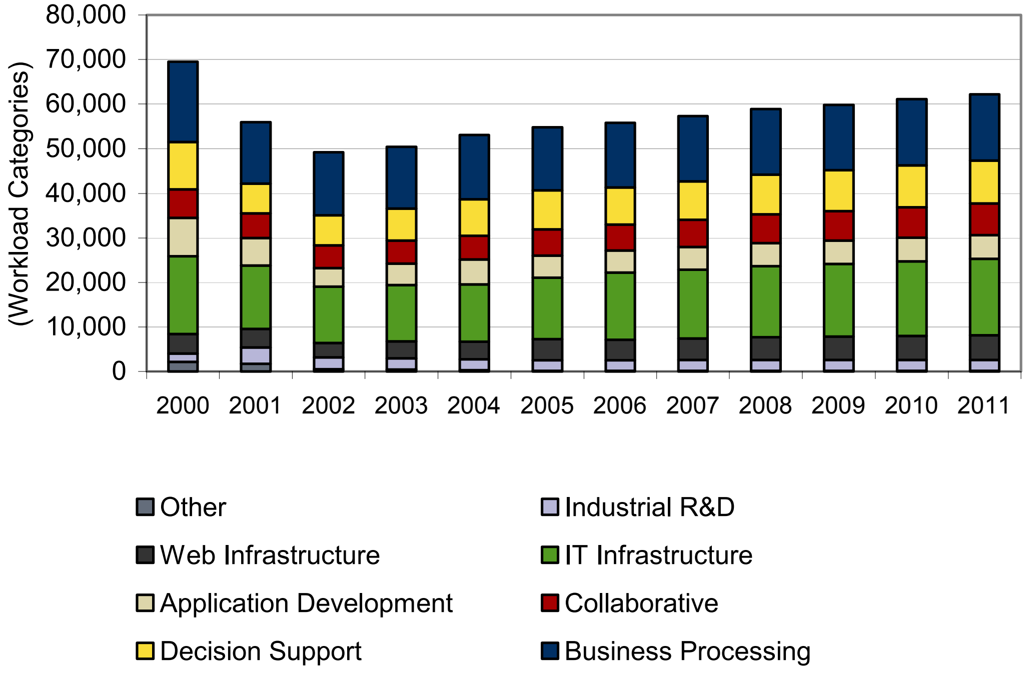
The ability to ensure business continuity — avoiding the impact of interruptions caused by outages — is the theme of business resilience, which is supported by IT ensuring that business processes, including end-to-end applications that span the enterprise, will be highly available, and accessible, to end users within the business — and to end customers of the business. Business operations can be heavily impacted if the applications and data that support critical processes are not available, or if the performance is degraded to the point that it impacts end users.

A number of mission-critical workloads are easy to identify, including systems that support banking, manufacturing, and retail operations. Without these systems, the business being transacted stops. However, as business processes are integrated with the delivery of IT services over the enterprise network and Internet, applications and data are increasingly moving into the business-critical or mission-critical category. For example, email is considered by many businesses to be a business-critical workload. That was not the case 15 years ago — but today, it is a mission-critical component of business because customers send purchase orders via email or make decisions to buy based on continuing email discussion.

IDC data regarding workload types shows that a range of systems supports modern organizations. This research, conducted annually since 1999, is based on surveys of 1,000 or more IT managers, providing a picture of what kinds of tasks are supported by modern organizations to run their business (see Figure 1).

FIGURE 1

Server Workloads Revenue Forecast, 2000–2011



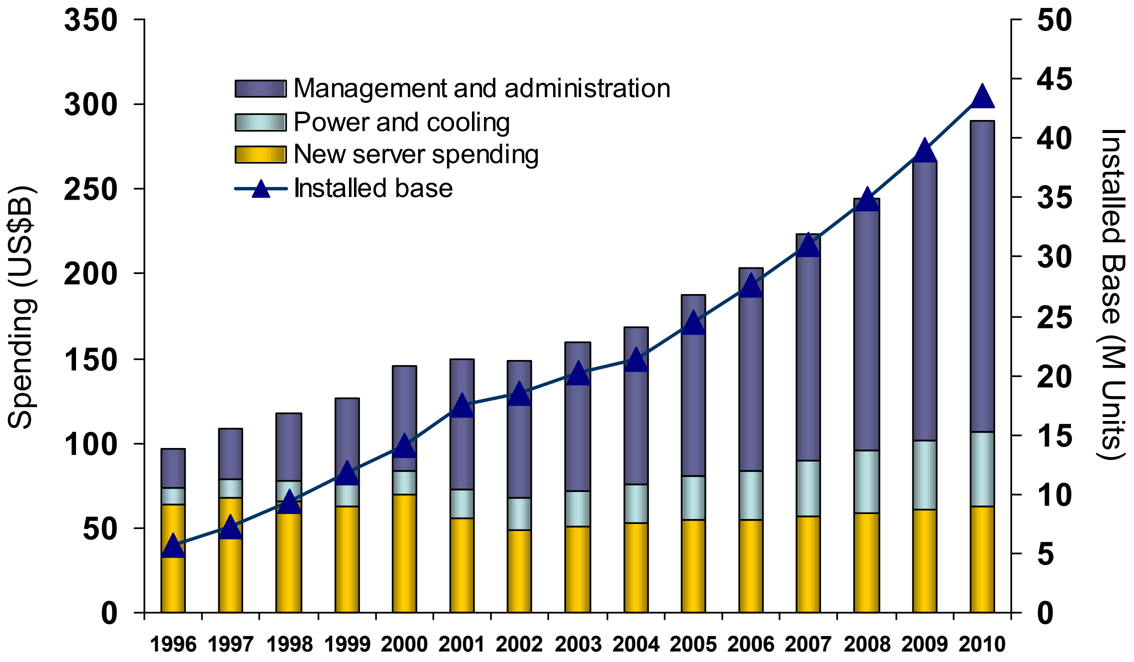
Source: IDC, 2008

The need to have responsive business applications is creating new demand for highly available systems. When IT services, applications, and data are interrupted, this disruption can slow, or stop, business operations. End customers frustrated by slow response times, or by downtime of these mission-critical and business-critical systems, may decide to do business elsewhere. Any kind of downtime, when systems are unavailable, causes operational costs to rise.

As seen in Figure 2, IT operational costs have been rising since 1996, increasing dramatically with the proliferation of small servers, beginning in 2001. Today, more than 30 million servers are installed worldwide — a big jump from the 4 million+ installed in 1996. Not only are there more servers, but many of them are located close to the end user, and close to the customer.

FIGURE 2

Worldwide Server Installed Base and Costs for Acquisition, Management, and Power and Cooling



Source: IDC, 2008

Companies must take action to reduce these operational costs in their IT infrastructures, and they are doing so in a number of ways, including:

- Consolidation of workloads onto fewer servers, reducing power/cooling costs and IT management costs
- Use of virtualization technologies, which support consolidation by allowing more work to be done on any single server platform
- Protection of these workloads through the use of high-availability software

The rapid growth, or proliferation, of volume servers (small servers priced less than $25,000) in IT infrastructure began in the 1990s when Web servers were first deployed widely. It accelerated in 2001, when concerns about capital expenditures (capex) focused on deploying workloads on small servers, rather than midrange enterprise and high-end enterprise servers, to reduce purchasing costs. This proliferation had an unintended consequence, however, because IT staff costs associated with system administration and management of large numbers of small servers were higher than those for larger, more scalable servers offering more advanced system management and control — and more RAS features to ensure uptime for applications.

Looking for Ways to Reduce Operational Costs

In recent years, the focus has changed: Operational expenditures are growing more rapidly than the costs of the servers themselves. Companies are looking for ways to reduce complexity in their IT deployments in order to drive down costs. Some approaches to doing this include the following:

- **Reducing power/cooling costs.** For example, IDC estimates that for every $1.00 spent on new servers today, an additional $0.50 is spent on power and cooling. In 2010, that ratio is expected to be $0.70 per $1.00 spent for new servers.

- **Reducing complexity.** Customers now have the option of consolidating multiple operating system images onto fewer servers — and managing them more effectively. To the degree that this consolidation occurs, it would be possible to reduce operational risk and operational costs that are linked to managing so many images on separate, physical servers — each of which could present interruptions in service down to hardware component failures. Importantly, clustering will benefit by having the option of "failing over" workloads to virtual servers, reducing the operational costs of deploying standby physical server machines that run in a "passive" mode rather than an "active" mode. Taken together, these approaches to reducing complexity in the infrastructure improve the responsiveness of IT systems and of the people who access them across the business, ensuring high levels of availability and reducing business risk and opex.

- **Improving management of physical and virtual servers.** Reducing the total number of systems under management simplifies IT operations and impacts the IT staffing requirements, along with the staff time devoted to applying software updates, security patches, and other forms of routine maintenance. Importantly, downtime is impacted by having fewer individual points of management. For example, bladed server systems manage more than a dozen individual blades in a unified way, avoiding the need to wire, or cable, each blade separately. This approach reduces operational costs and improves uptime for all of the blade servers within a chassis enclosure.

- **Going green across the infrastructure.** At the same time, business is being encouraged to "go green" in many dimensions: in the packaged products that a manufacturing business produces, in the end-of-life recycling for those products, and in the energy efficiency throughout the organization, to name just a few examples. The process of IT transformation brings the opportunity to change the IT infrastructure, supporting "go green" initiatives by reallocating workloads to the sets of server and storage devices on which they can run most efficiently and reducing total server footprints through workload consolidation.

By viewing the availability of IT data services to the end users and end customers as a business need, companies can consider how to reshape their current datacenters to be more energy efficient and, at the same time, how to build HA capabilities into the hardware and software that are being put into place as a result of "green IT." Workload consolidation, for example, brings more workloads to fewer servers, which has the effect of reducing the number of machines that must be managed — and of reducing the number of machines requiring power and cooling within the datacenter itself. The process of IT optimization, of placing IT resources where they can best be used, and mapping workloads to appropriate system resources reduces complexity in IT deployments and impacts opex.

Applying High Availability Where It Is Needed

How can businesses, and the IT organizations that support those businesses, get started on the path to improved high availability and business resiliency? The best answer is to take one step at a time because achieving high availability throughout the enterprise is not accomplished in a single project or at a single site.

Rather, it is the result of many steps taken, over time, to improve the availability of systems throughout the enterprise, using a variety of HA technologies, each appropriate to the type of workload being supported and to the end-user expectations about availability for those workloads.

Different workloads require different levels of availability. Some may require a five-nines level of availability, with just a few minutes of downtime expected per year. Others could easily withstand many minutes of downtime without causing irreparable harm to the business. There is a spectrum of HA solutions — and each solution must be mapped appropriately to the applications it supports.

Making changes in IT infrastructure requires careful consideration of the work entailed in making those changes — and of the consequences when efficiency in IT is not addressed. While "rip and replace" strategies are to be avoided, approaches that correctly identify which systems need to be updated right away can be most effective. Some aspects of IT inefficiency can be addressed best by breaking through the "islands of automation," and linking them, without replacing all of the hardware and software that took many years to deploy throughout the organization.

Evaluations and workshops are available from system vendors and services partners that will drive a workshop that helps a business evaluate how business priorities link to high-availability/disaster recovery (HA/DR) requirements — and how best to leverage existing assets across a complex, multivendor heterogeneous environment. This helps customers to build a plan or road map for delivering the availability required.

Once a plan for modernizing the IT infrastructure is in place, then steps must be taken to apply high-availability technology to the range of servers, storage, and software systems that are already in place. There is a spectrum of high-availability technologies, because some systems need to run without any kind of interruption (continuous operations), while in other cases, workload balancing and clustering between multiple servers, and data replication to ensure data availability, will provide adequate levels of availability for business systems.

IT Considerations for High Availability

As business processes are integrated with the delivery of IT services over the enterprise network and Internet, applications and data are increasingly moving from the noncritical category to the business-critical category. The need to have responsive business applications is creating new demand for highly available systems.

When IT services, applications, and data are interrupted, this disruption can slow, or stop, business operations. End customers frustrated by slow response times, or by downtime, may decide to do business elsewhere. As a result, the concept of service-level agreements (SLAs) must be updated to reflect the realities of real-time services being provided, via the Internet, to end customers. Business is no longer taking place within the confines of brick-and-mortar offices alone; it is taking place around the clock, through online banking services and online retail "storefronts." In these environments, a lack of responsiveness to the customer may result in the temporary or permanent loss of customer business, impacting revenue and profits.

Business resiliency is a term that refers to the ability of the business to make changes to its business systems while improving IT flexibility (the way in which applications are deployed) and availability of these important applications on a 24 x 7 x 365 basis to end customers worldwide. Efficiency in server deployment is allowing IT organizations to achieve cost efficiencies in a number of dimensions. A number of IT approaches can help IT organizations achieve this goal, including:

- Consolidation of workloads, driven by cost or energy issues, means that the remaining systems and infrastructure must be highly available to ensure access by end users and end customers. In recent years, many workloads have been deployed on rack-dense servers, including rack-optimized servers and bladed servers based on multicore processors. The increasing density of small servers is allowing more enterprise applications to be hosted on them, but it also increases the need for high-availability software to protect those applications. There is also the opportunity to consolidate workloads from many smaller servers to run on more scalable systems that already support high availability.

- Redeployment of some workloads onto scalable servers can also be an important component of workload consolidation. Scale-out heterogeneous, multivendor infrastructures can be very complex, especially in large organizations — and many enterprises may not have the IT skill sets or resources to design and to build the best solution for their business needs. There is also the opportunity to consolidate workloads from many smaller servers to run on more scalable systems that already support high availability.

- Virtualization, which allows applications to be moved across the IT infrastructure — and enables individual servers to support more applications per server "footprint" — means that IT investments can be more fully utilized. In many cases, small servers (volume servers priced less than $25,000) are used 15% of the time or less. Some midrange enterprise servers (servers priced from $25,000 to $500,000) are often used 40% of the time or less. High-end servers (servers priced at $500,000 or more), including mainframes, support utilization rates up to 90%, providing extremely high levels of resource utilization for mission-critical workloads. Virtualization in mainframes and Unix servers is already well-established (see Technology Drivers for Change sidebar) — and virtualization in x86 servers is accelerating, as customers seek improved utilization of already-installed servers. To the extent that virtualization leads to higher concentrations of applications or workloads running on fewer server footprints, then support of high availability becomes even more important to ensure business resiliency: the ability to ensure that business processes will continue to run, even in the event of outages in hardware, software, or the network itself.

- Ability to scale up resources on demand. As workloads increase, applications need more resources — processors, memory, and I/O. On scalable systems, additional capacity can be added, on demand, to support the peak demands of workloads, based on seasonality, time of day, or the amount of online access to the machine, especially for Web-enabled applications (e.g., online banking, online orders for consumer products and retail goods). In scale-out deployments, additional capacity can be added through clustering or workload balancing. This is usually enabled by the use of clustering and availability software that links multiple physical servers together via high-speed interconnects and shared storage. Alternatively, clustering or workload balancing software can be used to connect multiple physical or virtual servers to scale the resources supporting specific applications or data services being accessed across the enterprise. Another approach that is gaining in adoption is deploying workloads on bladed server systems — and adding blades as applications demand more processing power.

- Advanced system management that manages all computing tiers, across the IT infrastructure, both locally and across the enterprise. Enterprisewide management of workloads is effective at multiple levels — it leverages existing investments to manage workloads running on local servers, via "agent" software, and it allows organizations to manage end-to-end workloads across the business. It provides improved "visualization" of all managed workloads — of physical servers and virtual servers — across multiple locations, providing a bird's-eye view of the network and its applications. This ability to visualize all servers, including physical servers and virtual servers, on the network enhances availability on a daily basis — and it also enhances the ability to support disaster recovery procedures as systems are brought back online and data services are restored. Without it, the prospect of virtual server sprawl, as virtualization becomes more widely adopted throughout the enterprise, could become increasingly challenging for system administrators. By managing the virtualized environment more effectively, customers have the options of failover to alternate resources (physical or virtual) and of allocating new virtual resources when workload demand requires additional capacity. System management software needs to place a high priority on presenting a unified view of physical and virtual servers running on the network — providing "virtualization with visualization" to support system administrators and to allow applications to continue to meet performance goals and SLAs with business units. This capability is important, as server and storage resources are pooled together, to support workloads across the IT infrastructure.

- Data replication. To be highly available, end-to-end applications need access to enterprise data. If work moves from one server to another, through clustering of multiple servers or through workload balancing across the IT infrastructure, the data must be available to allow work to proceed. That is why attention must be given to replicating, or mirroring, data in more than one location. If one source of data storage is not available, then another can take its place. This ensures continued availability for business systems.

- Provisioning server workloads across the enterprise. New technologies are making it easier than ever before to provision workloads — to move them — across the IT infrastructure, where compute resources are available to run them. Examples of improvements in this area include live migration capabilities, which allow workloads to move from server to server, from blade to blade — or from partition to partition within a scalable server. Workload management software gives IT a control point from which to monitor where workloads are running — and whether they need additional resources (processors, memory, or I/O) to continue running with response times that address agreed-upon SLAs for end users of the systems being accessed.

Technology Drivers for Change

Technology trends are driving change in the datacenter. These technologies lead to initiatives to reimagine the datacenter as a place where energy efficiency and improved management of workloads lead to reductions in operational expenses. Following are some of the top technology trends within the worldwide server market — as reported by IDC supply-side and demand-side (customer-based) research — and how they affect customer deployments of new IT infrastructure. The process of IT transformation provides customers with an opportunity to do things differently, leveraging existing technologies, and adding to them, to optimize next-generation datacenters while reducing costs associated with IT staff time, unplanned downtime, and ongoing systems management.

Virtualization

Virtualization, which has been supported on mainframes for four decades, and on Unix servers for nearly two decades, allows system resources to be more fully utilized, preserving the IT investment in these systems while allowing workloads to run on available computing resources.

More recently, a wave of virtualization activity has been taking place in the x86 server world, based on ISV products from VMware, XenSource, and Microsoft that place "hypervisors" on x86 servers and support multiple operating system images to run on the same server platform. This process has the effect of placing more applications and workloads on the same servers, increasing the need for high availability on each server platform that is virtualized. On other types of architectures, the system platforms are already highly virtualized (e.g., RISC-based servers and mainframes) because the logical view of the system has already been abstracted away from the hardware itself. To the extent that virtualization has taken place, provisioning workloads is an easier task so that workloads can be moved from one area of the system to another. New technologies are also allowing workloads to be moved from one system to another, allowing provisioning and reallocation of workloads to take place on a datacenter level.

Consolidation

Growing costs and complexity required to sustain an ad hoc infrastructure are increasing operational expenditures, and when combined with new virtualization technologies, these consolidation efforts will enable significant ROI for customers. As mentioned earlier, the degree to which applications or workloads are packed into rack-dense server deployments only increases the need to ensure that these workloads are protected — and can be restored on alternate resources, if needed. The ability to consolidate on highly reliable hardware platforms, which are protected by RAS features as well as by HA software solutions, is part of a holistic, integrated approach to high availability.

Going Green

At the same time, business is being encouraged to "go green" in many dimensions: in the packaged products that a manufacturing business produces, in the end-of-life recycling for those products, and in the energy efficiency throughout the organization, to name just a few examples. By viewing the availability of IT services as a business requirement, companies can consider how to reshape their current datacenters to be more energy efficient and, at the same time, how to build HA capabilities into the hardware and software that is being put into place as a result of "green IT." Workload consolidation, for example, brings more workloads to fewer servers, which has the effect of reducing the number of machines that must be managed — and of reducing the number of machines requiring power and cooling within the datacenter itself.

As IT datacenters are transformed, with replacement of "islands of automation" by end-to-end systems across the computing tiers, the ability to consolidate workloads onto fewer server footprints is giving businesses the opportunity to "go green" by consolidating onto more energy-efficient platforms. The reduction in footprints is also an energy saver — for example, when workloads run on blades within a managed bladed server chassis. Redundant cabling is replaced by a chassis with electrical connections to all blades, and cooling elements within the bladed server chassis reduce overall power/cooling costs, compared with banks of rack-optimized servers, each wired and cabled separately.

IBM PRODUCTS AND SERVICES

IBM Systems and Technology Group (STG) products — System x x86 servers, the POWER-based systems (System p, System I, and System z mainframes) — offer a wide range of server and storage solutions that can be deployed across the enterprise to meet a wide variety of systems needs. This is important due to the variation of customer requirements, to differences in customer preference regarding use of IT technologies and IT skill sets, and to the needs of a range of computing workloads, which run in different operating environments across the enterprise. Although the server platforms and software stacks vary across deployments within the enterprise, there is a need to consistently address the availability requirements for each application or workload supported — and to meet the service levels specified by the business units across the enterprise.

This section lists IBM's hardware, software, and services offerings that are designed to address availability needs across this range of business and IT requirements:

- The x86 servers — IBM System x — are the product line for which IBM partners most closely with leading ISVs to provide off-the-shelf high-availability solutions to run on IBM server platforms. These HA solutions, from Microsoft, Oracle, Symantec, IBM Tivoli, and others, run on System x hardware, and can be linked to storage (SANs and NAS) consisting of highly available devices with RAS features built into the hardware platform itself. Regarding the software stack, the combination of the packaged software and IBM's system software and middleware creates a multilayer HA solution for robust protection of applications and data. IBM has taken elements of design from its mainframe and scalable server experience and engineered them into its System x and BladeCenter offerings, including the addition of RAS features for data integrity, security firmware and software, and system management that optimize throughput for System x servers.

- IBM POWER systems (System p and System i) are RISC-based systems that support IBM AIX Unix and several Linux distributions on IBM POWER processors and support the IBM i5/OS for System i applications. The scalable POWER systems for the enterprise and System i for SMBs bring high levels of RAS, security, and highly granular system management to customers. These systems have built-in server virtualization, which supports logical partitions (LPARs) and workload partitions (WPARs) and allows workloads to be moved to alternate resources, if needed. Now, with the release of IBM AIX 6 for POWER systems, customers will be able to move workloads, via WPARs, within a system or to another system to access system resources to run the workloads. This ability to move workloads to alternate resources is a key element in ensuring high availability for applications in the event that a hardware or software component fails. Importantly, this ability is also critical to address the effects of "planned maintenance," often a key inhibitor to achieving continuous operations. In the event of another type of outage caused by a natural disaster, an electrical outage, a network failure, or a man-made cause, clustering and availability software from IBM (e.g., HACMP, IBM Cluster software, and IBM Tivoli System Automation [SA] Base) links multiple server nodes together for failover support. In the case of IBM HAGEO, the ability to support geographic clusters, with server nodes separated by geographic distance, is also provided.

- IBM System z mainframes support Parallel Sysplex clusters, combining multiple System z servers into a single cluster that, through the use of the Sysplex Timer, achieves nearly continuous switchover to alternate resources within the Sysplex, resulting in uptime that is six-nines or more (less than 5 minutes of downtime per year). This places the Sysplex deployments at the top of the IDC Availability Spectrum, in Availability Level 4 (AL4), along with fault-tolerant servers, as there is no appreciable interruption in data services accessed by end users. IBM Geographically Dispersed Parallel Sysplex (GDPS) configurations of System z mainframes are an important element of supporting end-to-end workloads that span the enterprise — and the network. GDPS supports multisite or end-to-end application availability. It works with the IBM TotalStorage Enterprise Storage Server to automate Parallel Sysplex operation tasks and perform failure recovery from a single point of control. This capability supports near-continuous data operations, and ensures business continuity across multiple sites within the enterprise infrastructure.

- IBM Storage systems support virtualized storage so that multiple servers can write into "virtual storage" spaces that are logically, rather than physically, defined. That means that these storage devices can be the shared storage resources for a wide range of servers, across the enterprise, ensuring that data can be recovered when applications are restarted while preserving the logical partitioning on each shared server resource.

- IBM TotalStorage Productivity Center. This offering enables the integrated provisioning of servers and storage, across the enterprise, for rapid deployment of new computing capability. The combination of shared workflows and a common toolset for administrators makes provisioning easier and more efficient than before. TotalStorage Productivity Center leverages the capabilities of IBM Tivoli systems management software to control allocation of data, fabric, and disk. It takes on the role of the former IBM Storage Resource Manager and IBM Tivoli SAN Manager products to create a unified management framework for storage. Its automation capabilities reduce errors that can be caused by manual intervention in storage management, and its workflow automation speeds provisioning of storage capacity.

- IBM System Software. IBM offers a number of server operating systems, including IBM z/OS, z/VSE, and z/VM for System z mainframes and IBM AIX and IBM i5/OS for the IBM POWER-based systems. In addition, System z supports specialty processors, including the Integrated Facility for Linux (IFL) that supports Linux on the mainframe system for use with databases (e.g., IBM DB2 and Oracle Database 10g and 11g). This IFL capability means that Linux is now supported across all IBM systems (POWER, x, and z), allowing Linux applications to be written for the enterprise and to run on servers in a networked, Web-enabled computing environment. IBM WebSphere for application serving and IBM Lotus Notes/Domino can run on Linux running on the full range of IBM servers.

- IBM Management Software. IBM offers management software that addresses both the hardware layer and the software layer of the infrastructure. IBM Director is software that is used to manage the hardware components on x86 server systems, and it is working with IBM PowerExecutive to reduce power/cooling requirements "close to the hardware." IBM Director support is not limited to x86; it supports other IBM platforms, including POWER-based systems. IBM Director can be linked with the IBM Tivoli enterprise management framework to coordinate system management throughout the organization. It provides systems-level visibility to the entire enterprise and its IT infrastructure. Working across the enterprise, IBM Tivoli software supports end-to-end management, linking systems within the datacenter and between geographically separated datacenters. This capability supports geographically dispersed clusters, supporting failover of workloads across corporate campuses, or across network links between datacenters. IBM offers a number of IBM Tivoli products to support end-to-end applications. These products include an IBM Tivoli workload management portfolio, IBM Tivoli system automation portfolios, and the IBM Tivoli provisioning manager, and they are key offerings that help manage application availability.

- IBM Services. Key offerings from IBM's Global Technology Services (GTS) include the following: IBM High Availability assessment workshops designed to evaluate specific business requirements for availability, to assess current capabilities, and to develop a plan that delivers on requirements while leveraging existing assets across a multivendor heterogeneous environment; IBM Geographically Dispersed Open Clusters, which provide capabilities similar to those of GDPS, but in an open systems environment that supports hardware from multiple vendors; and IBM Business Continuity and Resiliency Services.

CHALLENGES AND OPPORTUNITIES

IBM has deep experience in providing hardware, software, and services that support high availability for data and applications. These high-availability solutions are offered for the full range of IBM servers and storage products, and they are supported on a global basis. In addition, IBM offers software, including Tivoli SA Base (clustering capability) and Tivoli SA End-to-End (supporting end-to-end applications that span the enterprise), that provides HA support across a wide variety of servers and clusters. The ability to orchestrate this HA support on an end-to-end basis will grow increasingly important as more SOA and end-to-end supply chain applications are deployed by customers.

Given the rapidly changing market dynamics within the IT industry, new technologies are continually being introduced into the marketplace. For example, virtualization software for x86 servers is being widely adopted, and it is offering support for high availability, especially for planned downtime and disaster recovery. But many of the scale-out, virtualized systems being deployed are not outfitted with built-in support for unplanned uptime — and some software providers in the x86 virtualization space are expanding their solution sets by partnering with server OEMs such as IBM and its server competitors, and with ISVs, to provide HA protection that covers the full spectrum of HA solutions for planned and unplanned downtime.

The challenge for IBM is to show how its support for a spectrum of high-availability solutions can be mapped into the range of enterprise solutions built on volume servers, midrange servers, and high-end servers, and do so in a holistic way that promotes simplified management across the enterprise. The opportunities are to leverage the advanced virtualization features of its server and storage systems and to show customers how scale-out virtualization can be combined with high-availability solutions to achieve business continuity and business resiliency. IBM has demonstrated that it understands the dimensions of this challenge, and it is bringing an array of products and services to the marketplace to provide a more unified approach to ensuring high availability across the technology stack.

CONCLUSION

The datacenter is being transformed in an effort to reduce operational costs associated with IT staffing, power/cooling, and downtime for server and storage systems. This transformation is marked by a new approach to IT infrastructure, in which that infrastructure is optimized so that workloads run on the platforms that support them with the best performance and the greatest efficiency.

IBM offers hardware, software, and services that allow customers to take a comprehensive approach to ensuring high availability across a range of servers and storage systems, using hardware with RAS features built in to the platform; software supporting virtualization, consolidation, and automation; and advanced end-to-end systems management to support end-to-end applications that tap many servers. This approach should benefit end users within the organization and end customers accessing data services across IT infrastructure, on a local, campuswide, or global basis.

Copyright Notice

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2008 IDC. Reproduction without written permission is completely forbidden.