I D C   E X E C U T I V E   B R I E F

Using Availability and Clustering Software to Maintain Business Continuity in the Era of Virtualization

November 2008

Adapted from *Worldwide Clustering and Availability Software 2007–2011 Forecast* by Jean S. Bozman, IDC #208632 and *Worldwide and U.S. High-Availability Server 2007–2011 Forecast and Analysis* by Stephen L. Josselyn and Jean S. Bozman, IDC #210096

Sponsored by Symantec

Introduction

The market for highly available servers and high availability is growing, reflecting enterprise demand for solutions that strive to maintain business continuity. These solutions are becoming more important for operational viability, especially as organizations increase their adoption of virtualized computing solutions. This Executive Brief describes high-availability options, particularly availability and clustering software (ACS), and how they are invaluable in virtual computing environments. This paper also describes the growing market for ACS solutions, high-availability software options to ensure business continuity, and what to look for when selecting high-availability solutions.

High-Availability Software Market Growth and Drivers

Minimizing the cost of downtime to ensure operational viability is a critical element for business and financial stability as organizations continue to be held accountable for the availability of their data and applications. While all businesses are affected by failure or extended periods of downtime, they also experience slowdowns in performance when some of their distributed systems go offline, making overall business operations less efficient.

The need to preserve data and application availability remains a demanding one for IT managers and business organizations. Without it, business continuity would be difficult to maintain, and preparation for unexpected downtime would threaten the high availability of applications.

IDC_718

In a virtualized IT world, where many virtual machines are running workloads on physical servers, high availability is more important than ever. The increasing adoption of virtualization on servers in the datacenter requires new approaches to protecting applications that are running inside virtual machines (VMs) — and has become an essential element of business continuity planning. Companies implement high-availability solutions for many reasons, including to minimize planned downtime for maintenance and upgrades, to enable rapid recovery from unplanned downtime — and also to meet compliance requirements that ensure continued end-user access to important business applications.

As a result, the market for ACS is expanding and is expected to grow to $2 billion by 2009. In addition, the emergence of virtualization software, and the twin needs to simplify and to integrate functionality that is time-consuming to install and manage, is further driving the market in terms of innovation, new products, and new approaches to attain highly available applications and data.

On a worldwide basis, businesses want to gain greater flexibility within existing IT infrastructure to align IT infrastructure processes better with the business processes in their organizations. That quest for improved IT flexibility is a key driver for ACS revenue growth in the Americas, which represented 51% of the market in 2007; EMEA, which held a 34% market share in 2007; and Asia/Pacific, which held a 14% share in 2007 (see Figure 1).

F i g u r e   1

Worldwide Availability and Clustering Software Revenue by Region, 2007 and 2012
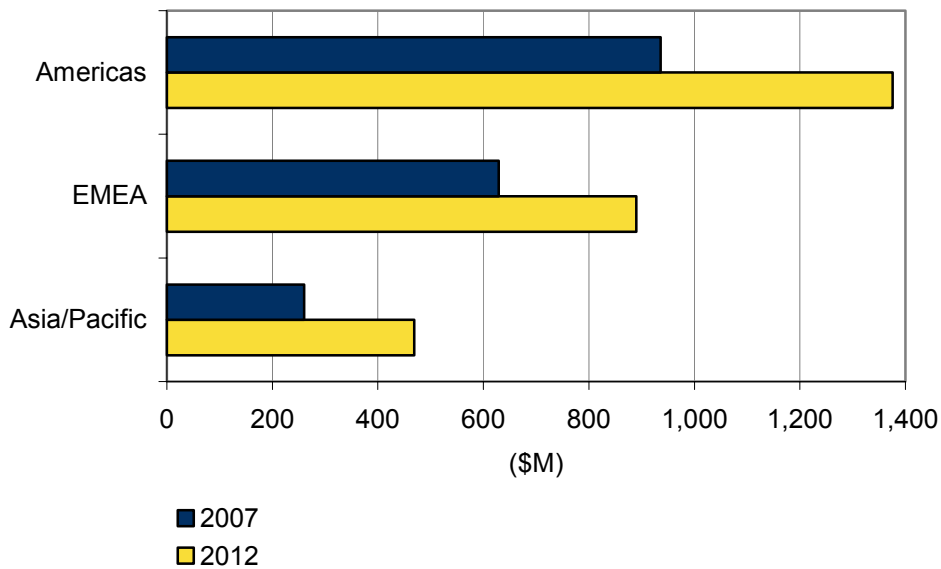
Source: IDC, 2008

Key drivers for ACS market growth include the following:

- Customers' high-availability requirements for enterprise applications

- The need for business continuance in the event of outages

- The need for flexible operations that can tap available applications running across multiple computing tiers

This growth rate in the ACS space exceeded the growth rate for the worldwide server market, which was more than 3% in 2007. Growth for the ACS market also was seen across operating environments — Windows, Linux, and Unix. The continued strong growth stems from continuing use in the enterprise space to protect both mission-critical workloads and new deployments in the scale-out, virtualized world. ACS is designed to ensure that alternate resources are available as computing needs grow, spurring the use of new server resources — or as outages affect some servers, requiring the use of other servers.

High-Availability Solutions

High-availability software is evolving into a broad portfolio of products, many of which work with physical servers or virtual servers — or both.

Clusters

A standard method of maintaining high availability is clustering. Clustering and associated failover and availability software virtualizes system services of multiple computing resources so they appear as a single computing resource. Software for clustering includes a cluster file system and software that manages the cluster — allowing workloads to move from one server node to another. Two or more servers are linked through a "heartbeat" that allows each to know when the other is offline. If that heartbeat should stop, then one of the other servers within the cluster takes actions to restart the applications, leveraging information on a shared storage resource (e.g., disk drive or SAN).

The combination of servers, storage, and clustering software to achieve highly available data and applications provides a broad range of solutions for customers who would otherwise need to integrate these solutions by themselves, leveraging IT skill sets at their site.

Workload Balancing

Another method of maintaining high availability is workload balancing software, which stands between the user request and the processors or systems supporting applications or services. Workload balancing software, which is often integrated into other software products (e.g., operating systems), automatically routes processing requests to the processor or system that has the most available capacity. The performance parameters that can be set to trigger the workload balancing typically are determined by the service-level agreements (SLAs) with the business units in the enterprise.

Data Replication

Data replication enables companies to guard against a "data unavailable" situation. Although not having data available when it is wanted or needed is indeed problematic and disconcerting, data still is recoverable using restoration, restart, recovery, and/or copy tools. Having a second replica available for users or their applications to recover mitigates the damages caused by data loss and guards against manmade and natural disasters. That said, data replication solutions themselves do not provide application-level high availability or disaster recovery.

Data replication has many uses — including protection of shared data, availability of data at multiple sites within an organization, and data protection for purposes of disaster recovery. For virtualization scenarios, data replication ensures that production data will be available for use by applications running inside VMs, especially when those VMs have been "migrated" to new locations for high-availability/disaster recovery purposes.

The full spectrum of availability levels shows the need for a spectrum of high-availability solutions, including fault-tolerant solutions (for uninterrupted processing), clustering failover solutions for highly available applications, workload balancing to match incoming requests to available computing resources, and data replication to protect the data that will support the full range of high-availability solutions.

In addition, scaled-out, virtualized computing environments in many enterprises bring new availability needs — and new approaches to achieving high availability for applications and data. As enterprise workloads increasingly are supported in this scale-out environment, IT needs to look to a spectrum of high-availability software options to identify which workloads should be protected most — and which software products should be used to ensure that protection.

Business Considerations for High-Availability Software

Because availability software is now evolving into a broad portfolio of products, many of which work with physical servers or virtual servers, or both, enterprises have many options. Deciding which options to employ, based on each organization's specific needs, requires not only a technical decision but also a business decision.

Enterprises that are evaluating high-availability software solutions should consider both the IT needs and the business needs of their organization. The high-availability solution should do the following:

- **Provide alternate resources, either physical or virtual.** For both planned and unplanned downtime, end users must be able to access computing resources somewhere within the network. Otherwise, business revenue would be lost, and customers would become unhappy with unavailable Web sites and business services.

- **Ensure continuous or near-continuous data processing at scaled-out computing sites.** Companies that are using scale-out computing need availability just as much as companies that have large servers within their datacenters. The focus must be on the applications and their ability to restart on other servers (physical or virtual) to ensure business continuity.

- **Support and protect applications running on both physical and virtual machines.** The rapid adoption of virtualization software in the x86 server space means that x86 servers — like Unix servers and mainframes in the datacenter — are now virtualized resources. This allows improved resource utilization for the servers in the datacenter, and it supports workload consolidation on the physical servers. However, in the event that any physical server goes offline, that outage will affect many more end users than it would if the server were not virtualized. Therefore, now that the average number of VMs running on each physical server is higher than ever, it is more important than ever to protect applications running on VMs with detailed application monitoring and automated failover.

- **Scale resources as business needs change, and as applications and databases grow.** Businesses see peaks and valleys in terms of demand for use of computing resources. As demand for computing grows, more capacity must be added or existing capacity must be repurposed. When demand decreases, workloads could be moved around to repurpose online servers for other tasks or to take some servers offline for repair, upgrades, and updates (e.g., using planned downtime for maintenance). High-availability solutions must be able to support both scenarios (adding/removing or repurposing servers) for ongoing cost-effective business operations.

- **Reduce system administration and user complexity challenges.** IT skill sets are not evenly distributed; large enterprises may employ more of the IT staffers who have the skills needed to optimize, tune, or rescript restart procedures. Automating more of these software "steps" allows restarts of applications and data to happen more easily and normal processing to be restored more quickly.

- **Automate disaster recovery to minimize issues associated with manual recovery by IT staff.** Difficulty in recovery may be due to a number of factors, including lack of access to a given office or datacenter, power outages, network outages, or the inability of IT staffers to restart applications and to access production data.

- **Proactively test systems for high-availability/disaster recovery scenarios and notify administrators of problems or potential problems.** Many organizations have disaster recovery plans, but they do not test them adequately, often until an outage actually occurs. Some products offer built-in high-availability testing that allows testing without the disruption to production workloads.

In addition, high-availability solutions must be easy to install and maintain. IDC research shows that a lack of IT skills related to installing and managing clustering and high availability is a concern in enterprises.

Of particular importance is the challenge of evolving the way platforms are supported, especially in virtualized environments, because many of the software tools that were originally designed to manage the physical server environment are not designed well for some of the features of computing in the virtualized world, such as support for live migration of VMs. The next generation of management tools must manage in an environment populated by both physical and virtual servers to provide a holistic solution across the enterprise.

Adoption Considerations

Adopting high-availability solutions is not without its challenges, however. A critical challenge is having adequate IT skill sets in-house to deploy and maintain clustered and highly available software solutions. However, for enterprises large and small, increasing pressure on IT budgets means that simplified administration, single-pane-of-glass management, and increased automation are all becoming more important to next-generation high-availability solutions.

Another concern is managing new pricing models, often based on virtualization rather than on the number of processors (or processor cores). This is an area that is prompting ISVs to consider their pricing models, and to match the pricing more closely to usage patterns, whether by site licenses, number of end users accessing the software, or number of server "nodes" within the cluster.

In addition, IT managers also face the prospect of virtual machine "sprawl," which, when combined with server sprawl for physical servers, means that administrators must manage more software-defined "objects" than they did just a few short years ago. Ultimately, most organizations will need a unified management console that supports both high-availability software and virtualization software and that helps them discover, manage, and visualize both physical and virtual servers. A unified management approach can save enterprises operating and training costs while providing the high-availability requirements that mission-critical applications need.

Conclusion

Failure or downtime in mission-critical systems affects businesses of all types and sizes. But the needs of each organization are different, which is why a spectrum of high-availability solutions is available in the marketplace. High-availability solutions must be evaluated for their ability to address the business and technical needs of the organization in which they will be deployed.

High-availability solutions are essential to maintain business continuity to avoid downtime that impacts user productivity and end-customer access to important applications. Choosing the right approach to high-availability will depend on the organization and on the level of availability needed. As more organizations adopt virtualization, software-based availability and clustering solutions for those virtualized platforms will be more important than ever. These solutions can reduce operational costs, save IT staff time, and minimize the potential expense of data recovery. In addition, IDC notes that organizations will need products that take a unified management approach to supporting physical and virtual servers across the enterprise.