The High Availability challenge: 24x7 in a Microsoft® environment

White paper

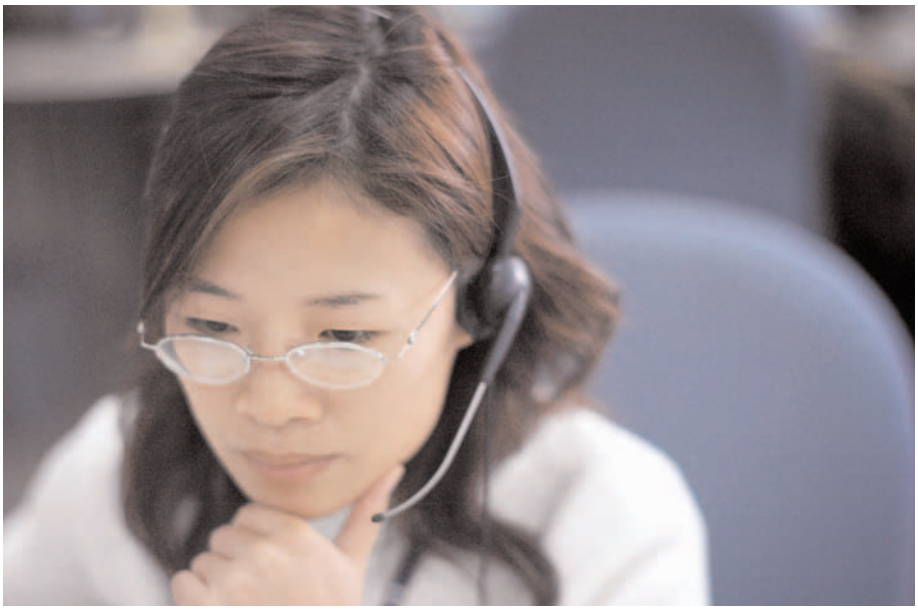


Table of contents

invent

The complexity of today's IT infrastructures and applications makes managing systems to high levels of availability difficult.

Executive summary

Increasingly, today's enterprises need to maintain continuous availability for growing numbers of applications in a Microsoft® environment. This need presents major IT and process challenges. Continuous application availability is achieved only with a carefully executed strategy and corresponding planning work. To achieve success, an organization must first establish a solid high-availability strategy, and only then look at technology that will support that strategy. There are four key steps in this process: analyze your requirements, select your strategy, design your architecture and manage your infrastructure. This process puts an organization onto a path to solutions that enable 24x7 application availability in a Microsoft environment.

The availability challenge

Meeting the demands of a 24x7 world

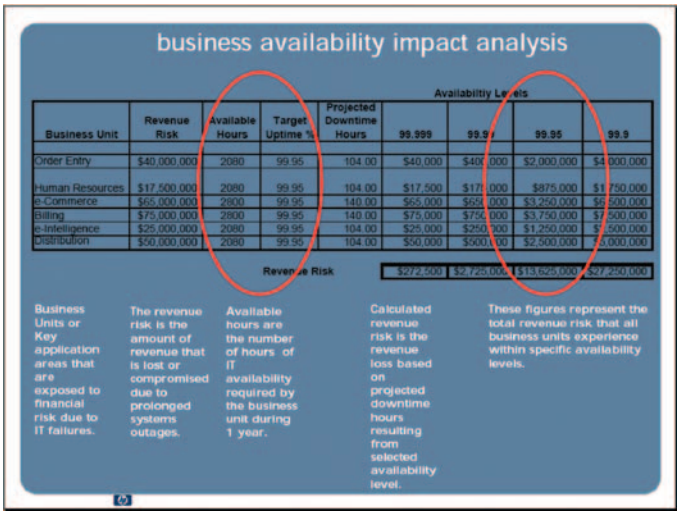
For today's enterprises, application availability has emerged as a formidable challenge. The days of the 9-to-5 business-hours application are fading into a 24x7 world. Increasing numbers of business applications must now be up and running around the clock. This is a need driven by multiple forces, including the widespread use of the Internet, the growth of direct business-to-business (B2B) and business-to-consumer (B2C) commerce, and increased real-time application integration.

In a large enterprise, business never stops. Customers, business partners and internal staff have come to expect 24x7 access to business services. And even at that, expectations continue to rise. Today, for instance, it is not good enough to simply enable customers to order goods over the Internet. Customers also expect to see inventory levels, check the availability of goods and determine anticipated ship dates.

This shift in expectations has major implications for IT operations. Many of the back-end applications that are integrated with front-end B2B and B2C applications now require expanded hours of operation, moving toward 24x7 accessibility. This need creates perplexing business issues. Maintaining an around-the-clock application can cost more than three times the price of maintaining a standard application. What's more, in many cases, a 24x7 requirement necessitates that vendors and partners either redesign their products to accommodate online upgrades and maintenance or architect their solutions for complete redundancy.

The complexity of today's IT infrastructures and applications makes managing systems to high levels of availability difficult. What's more, meeting the need for high availability requires more than just technology. People or process failures directly cause 80 percent of mission-critical application service downtime. Given these realities, achieving continuous availability requires a multi-pronged strategy that addresses and mitigates risks of failures and planned maintenance and upgrades. Continuous availability must be designed into not just the application but the overall business environment. It requires substantial levels of cross-organizational coordination of people, processes and technology.

Figure 1. Business availability impact analysis.



All companies have specific availability requirements for the applications used to operate their particular businesses. These requirements fall into different levels:

- Non-critical
- Standard business processing
- Mission-critical

In addition, investments must be made to achieve the level of application availability required in the areas of:

- Application development
- Technology infrastructure
- Process implementation
- Support services

Ultimately, achieving 24x7 application availability doesn't happen by chance, nor is it the result of buying the right point solutions. Continuous application availability is achieved only with a carefully executed strategy and corresponding planning work. To achieve success, and meet unique needs, an organization must first establish a solid high-availability strategy, and only then look at technology that will support that strategy.

A user perspective

The term "high availability" refers to the user experience, not the IT components. A highly available IT service provides users with access to applications and data for a minimum of 99 percent of scheduled time, despite unscheduled incidents. This typically implies the ability to eliminate or avoid unscheduled outages, via error detection, circumvention, correction and recovery, or at least to minimize downtime via capabilities such as rapid restart.

The term "high availability" also implies application and data integrity, as well as acceptable application performance, as defined by the user. This is because, as noted, high availability must be measured from a user's perspective. If a user can't access an IT service during scheduled hours, or if the service is slow and unresponsive, the application is unavailable in the eyes of the user.

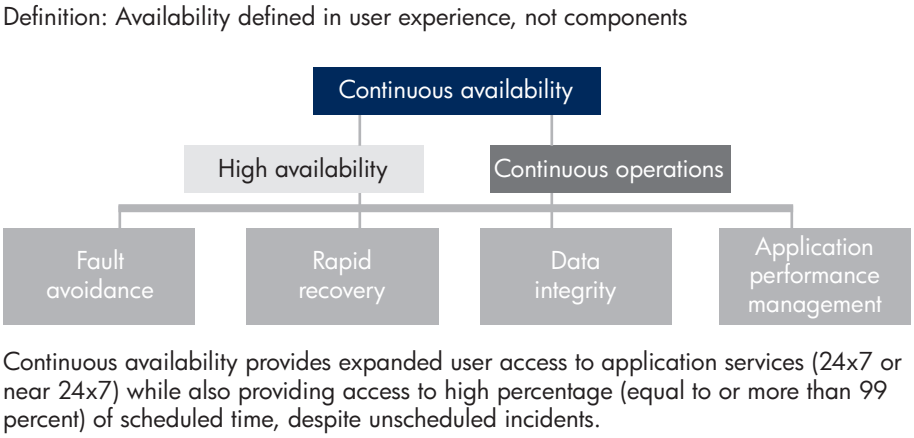
Most business-critical applications—including B2B, B2C and enterprise applications—have some planned or scheduled downtime for maintenance. Scheduled downtime should be negotiated with users to avoid times of peak usage and increased seasonal business demand. Further, scheduled downtime should clearly be communicated to users to set expectations and avoid the dissatisfaction that occurs when users try to access a site or a service that is unavailable.

In many cases, the goal is to have a continuously operable site to enable access during expanded hours, often on a 24x7 basis or a near-24x7 basis. Continuous availability is the combination of high availability and continuous operations. Continuous availability enables expanded hours of user access for a high percentage of the time (99 percent or more), even when taking into account planned maintenance.

Shifting the focus

As an organization seeks to increase its ability to manage business-critical applications and achieve high availability, it undergoes a shift in focus. The focus is less on technology components and more on the integrated management of processes, people and services. Figure 3 depicts how the focus changes as an organization progresses upward along the ladder to higher availability.

Figure 2. Availability is defined by the user experience, not the back-end IT components. Source: Gartner[1]



Continuous availability provides expanded user access to application services (24x7 or near 24x7) while also providing access to high percentage (equal to or more than 99 percent) of scheduled time, despite unscheduled incidents.

At the lowest levels of maturity and capability, organizations focus on hardware components and individual technology features for components. In the middle tier, the focus shifts to integrated management. And at the highest level, organizations place greater emphasis on IT processes to enable optimum business application service levels and operational excellence. Although technologies can help make all of this happen, the optimal, adaptive end-state for the business cannot occur without an IT service level management focus.

Starting from this point and perspective, the process of developing a high-availability strategy follows a series of progressive steps:

- Analyze your requirements.
- Select your strategy.
- Design your architecture.
- Manage your infrastructure.

Step1: Analyze your requirements

The high-availability journey begins with a close analysis of your availability needs and business requirements. Figure 4 illustrates the linkages between the various key elements and disciplines required to actually achieve these goals. High availability includes, but is not limited to:

- Clustering
- Redundancy of hardware components
- Dual or load balancing networks

In addition, high availability requires a discipline of designing end-to-end availability through the application integration and infrastructure stack. In parallel, an organization should embrace appropriate IT management best practices to better manage service levels and to mitigate the risk of interruption, whether from external or internal causes.

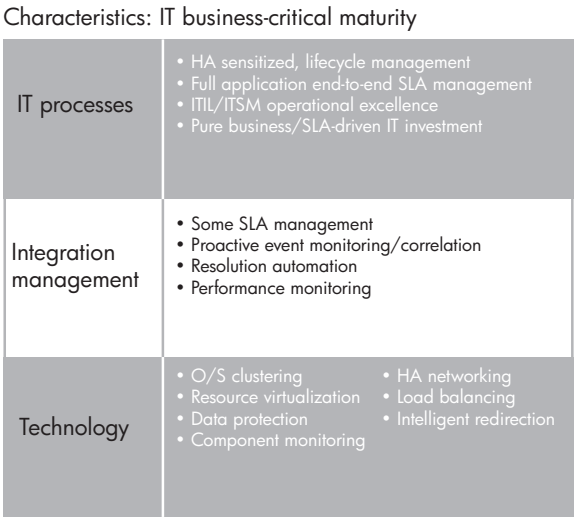
At the highest levels of the availability continuum, service level management capabilities are indicative of the most mature and trouble-free high availability solutions. At this stage, an organization is beyond the "construction phase" of the application design and infrastructure deployments. The focus here is on deploying solutions and transitioning the organization to actually operate an environment at "continuous availability" levels.

At this level of maturity, IT processes follow the widely accepted IT Infrastructure Library (ITIL) standard for excellence. In addition, organizations use event management and service level management tools to enable higher availability and to allow staff to continuously measure availability and performance levels.

Step 2: Select your strategy

Once you have gained a clear understanding of your needs and requirements, it's time to put a strategy in place. To achieve the best balance, organizations need to clarify the precise business requirements and to map those requirements into a broader strategy. This strategy then drives the selection of technologies and partner integrations that span the availability continuum and address unique challenges.

Figure 3. As an organization's business-critical environment matures, the integration of processes and people rises in importance.



To develop the right strategy, an organization needs a complete view of its availability needs over time. This view should include scalability requirements—both upward and outward—as well as the lifecycle management impact on availability as applications are upgraded.

Some businesses never require more than "reliable" solutions. These businesses can suffice with the quality and disciplines of fairly basic solutions. Other organizations must have "continuously available" applications to meet business needs. Identifying the appropriate cost/benefit balance for each situation requires careful consideration of all variables, not just the technology components. The decisions made here can have substantial cost implications. In general, the higher the availability, the higher the cost.

Business continuity solutions

Depending on business requirements and recovery time objectives, a full range of business continuity solutions and strategies should be considered. Figure 6 illustrates the range of business continuity options.

Infrastructure and tape backup provide the foundation of a business continuity solution. Tape backup can be automated to reduce operator media management. In addition, there are ways to gain cost-effective, predictable recovery through professionally managed outsourcing of backup and restore. Network and storage area network (SAN) backup are ideal for providing high availability and allowing sharing of tape library resources on a network.

Rapid backup solutions address one of the most important enterprise concerns: reducing or eliminating the downtime of an application during backup. In these deployments, a dedicated server is used to back up data from disk to tape within a LAN or a SAN. Direct backup solutions leverage this capability. They enable backup from disk directly to tape, bypassing a backup server within a SAN.

Rapid recovery solutions refer to using local disk mirrors or snapshots in conjunction with tape libraries. If the local mirror reflects the desired recovery point, it is instantaneously swapped with the live data. Otherwise data is recovered from tape.

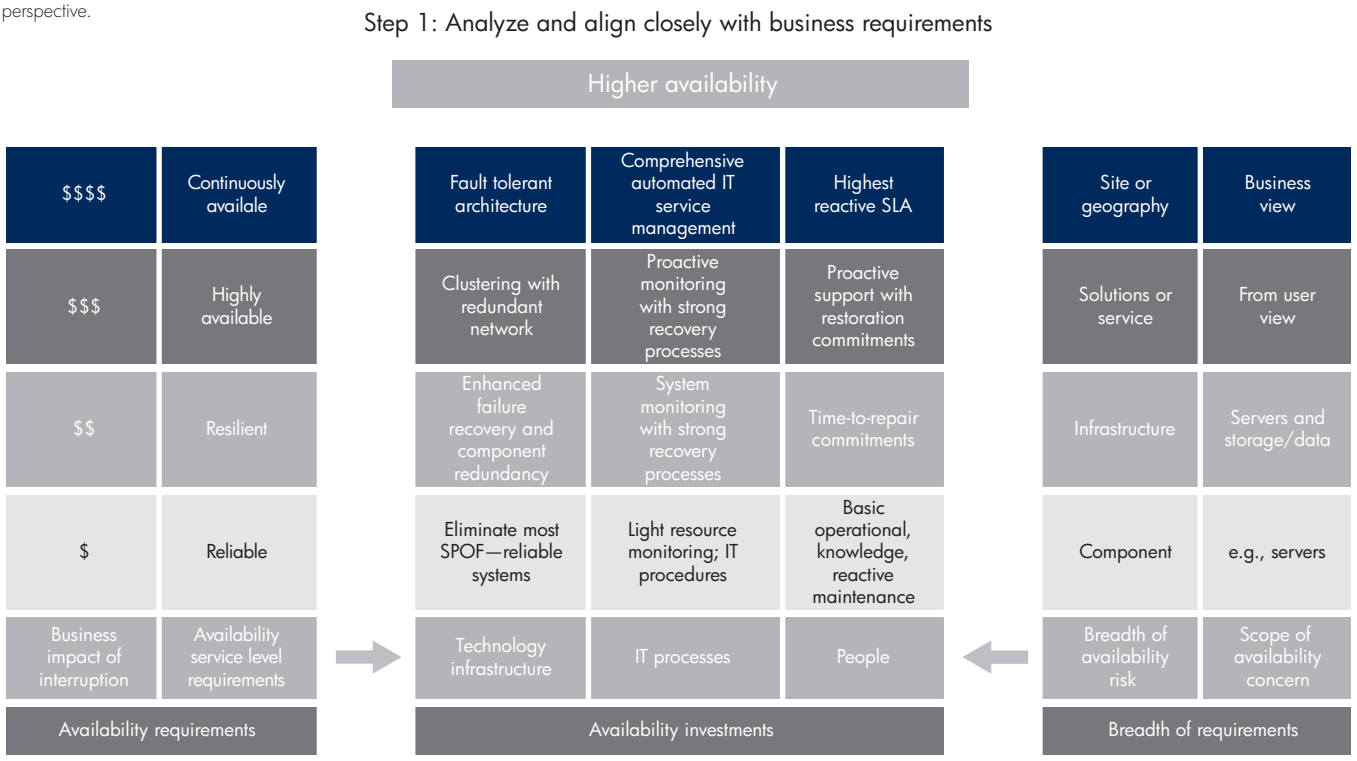
The next level is remote mirroring. This allows recovery from a site disaster. Careful thought must be given to provisioning systems and networking at the remote site for this type of recovery. Configurations must be installed, and users configured unless a full hotsite is in place.

In addition to remote mirroring, remote clustering can be used to keep critical applications running through a site disaster. These can be fully automated or, in cases of long distances, manual failover can be utilized. Each data center can have mirrored and clustered applications to the other; therefore, they can protect each other.

State-of-the-art hardware and software are only two pillars of business continuity. The third pillar is services and processes to ensure that the operating environment is designed for business continuity and that equipment is kept in optimal condition.

Higher deployment costs are often associated with more advanced solutions that can recover more quickly. When the total cost of downtime is factored in, it may be that they are actually less costly in the long run.

Figure 4. A high availability design requires an end-to-end perspective.



Disaster-tolerant design considerations
In developing a disaster tolerant strategy, four considerations are key:

- Protection level
- Data currency
- Failover time
- Performance requirements

Protection level involves distance considerations. There are a wide variety of interconnect options. A strategy can enable regional or wide-area protection, as well as support for local to global disaster-tolerant solutions.

Data currency considers recovery point objectives. Synchronous and asynchronous options are available. With either approach, data consistency should always be assured.

Failover time takes into account recovery time objectives. A solution can be designed to failover manually to a secondary site. Or a solution can enable fully automated failover with geographically dispersed clusters of Windows®, HP-UX, HP OpenVMS, Solaris, AIX or Linux systems.

Performance requirements are another important area of consideration. Asynchronous continuous access provides minimum latency across extended distances. Performance also hinges on bandwidth to remote data centers.

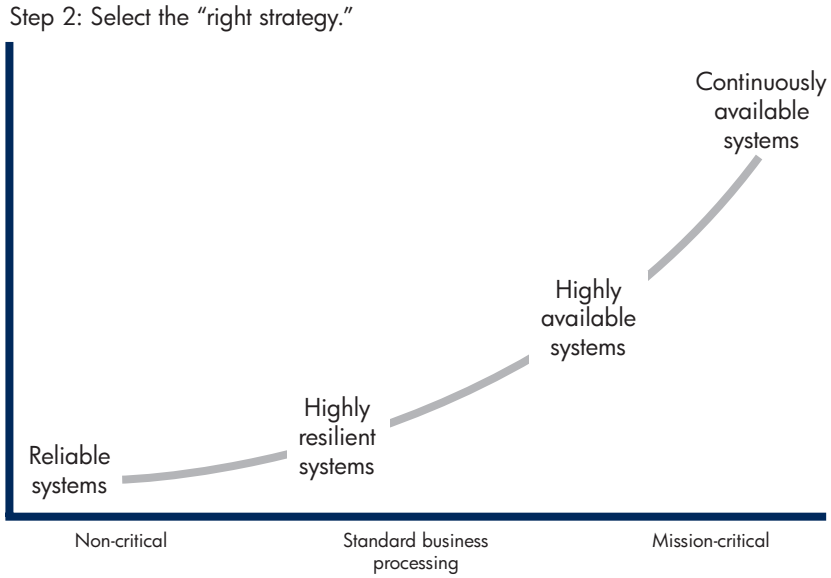
Step 3: Design your architecture

Once your requirements are fully understood and you have put a sound strategy in place, it is time to push forward with the design of your high availability architecture.

For application services with continuous availability requirements, including short recovery time objectives (RTOs) and recovery point objectives (RPOs), multi-site architectures are used. Often, a new real-time enterprise (RTE) application service starts with a single-site architecture and migrates to multiple sites as risks grow. Multiple sites complicate the design of application architectures. Multiple sites, for instance, introduce issues related to load balancing, database partitioning, database replication and site synchronization.

High availability solutions from Microsoft and HP address these challenges. The combination of Windows Server 2003 technologies, such as network-load balancing, clustering and data replication, work in tandem with HP software, hardware and services to allow organizations to design high availability multi-site architectures. Architectures that leverage Microsoft and HP components can address all continuous availability requirements, including short RTOs and RPOs.

Figure 5. In selecting a high availability solution, an organization should closely consider costs and benefits. In general, the higher the level of availability, the higher the cost of the solution.



Windows Network Load Balancing (NLB) service enhances the availability and scalability of applications, such as those used on web servers and other mission-critical servers. A single computer running Windows 2003 can provide a limited level of server reliability and scalable performance. However, by combining the resources of two or more computers running Windows Server 2003 into a single cluster, Network Load Balancing can deliver the increased reliability and performance that web servers and other mission-critical servers need. Network Load Balancing allows all of the computers in a cluster to be addressed by the same set of cluster IP addresses, yet also maintains their existing unique, dedicated IP addresses.

For load-balanced applications, when a host fails or goes offline, the load is automatically redistributed among the computers still operating. Applications with a single server have their traffic redirected to a specific host. When a computer fails or goes offline unexpectedly, active connections to the failed or offline server are lost. However, if you bring a host down intentionally, Windows 2003 allows the administrator to redirect all new traffic to available hosts. When ready, the offline computer can transparently rejoin the cluster and regain its share of the workload.

Windows Server 2003 Cluster can be used to provide server redundancy at the local data site or geographically dispersed data centers. Windows Server 2003 Cluster is a dramatically improved version of the Microsoft Cluster Service (MSCS) component included with Windows 2000 Advanced Server and Windows 2000 Datacenter Server. Server Cluster can be configured to work with two to eight servers that act as nodes in the cluster. Cluster resources available to an application include network names, IP addresses, applications, services and disk drives.

Clusters provide protection only at the host level. They cannot guarantee service availability when a site disaster occurs or there is a storage failure. Such a major disaster might require a full restore from backup tapes, hence the need for tapes to be stored offsite.

Virtual Disk Service and Volume Shadow Copy, along with HP storage solutions, provide sophisticated tools for the design of data replication solutions. Virtual Disk Service (VDS) makes management of storage subsystems independent of the hardware subsystem and the interconnections. Volume Shadow Copy services, known in the industry as "snapshots," allow for the creation of data backup copies at a remote location. Snapshots can also be used with data mining applications.

Solution building blocks for Windows infrastructures

Microsoft Windows Server 2003 Cluster

The building blocks of an HP-Microsoft high availability solution begin with Windows Server 2003 Cluster, which provides advanced capabilities for high availability architectures.

A server cluster is a group of computer systems (nodes) that work together as a single computer system to increase the availability of applications and resources. Clustering software enables the nodes of a cluster to exchange specific messages that trigger the transfer of resource operations at the appropriate times. Server clusters can be set up as one of three different cluster model configurations:

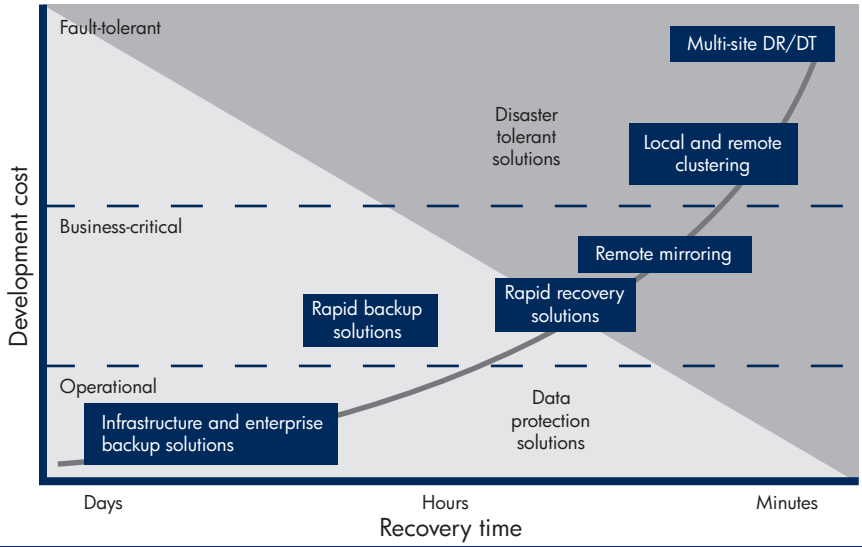


Figure 6. Depending on business requirements, and recovery time objectives, a full range of business continuity solutions and strategies should be considered.

- **Single node** server clusters can be configured with or without external cluster storage devices. For single node clusters without an external cluster storage device, the local disk is configured as the cluster storage device.
- **Single quorum** device server clusters have two or more nodes and are configured so that every node is attached to one or more cluster storage devices. The cluster configuration data is stored on a single cluster storage device.
- **Majority node** set server clusters have two or more nodes but the nodes may or may not be attached to one or more cluster storage devices. The cluster configuration data is stored on multiple disks across the cluster and the cluster service makes sure that this data is kept consistent across the different disks.

Most clustered applications, and their associated resources, are assigned to one cluster node at a time. If Server Cluster detects the failure of the primary node for a clustered application, or if that node is taken offline for maintenance, the clustered application is started on a backup cluster node. Client requests are immediately redirected to the backup cluster node to minimize the impact of the failure. Though most clustered services run on only one node at a time, a cluster can run many services simultaneously to optimize hardware utilization. Some clustered applications may run on multiple Server Cluster nodes simultaneously, such as Microsoft SQL Server. Server Cluster utilizes a filter driver to coordinate access to shared disks for the nodes in a cluster. Only one node in the cluster owns a disk at a single point in time.

Clustering solutions, such as the ones described above, protect against server failures. Nodes are typically located in the same building. This means clustering solutions do not provide protection against:

- Site disaster
- Storage failure
- Core infrastructure failure

A major disaster could require full restore from tape, so tapes should therefore be stored offsite.

Windows Server 2003 provides services for the data protection/replication at the software level. Advanced storage solutions, such as those from HP, allow you to design similar solutions where additional levels of availability, replication and scalability are achieved. These advanced solutions are tightly integrated with Windows 2003.

Among these software offerings is HP Cluster Extension XP/EVA, which allows data protection at extended distances. Cluster Extension XP/EVA enables disaster recovery and protects against the risk of downtime and data unavailability, whether planned or unplanned.

Leveraging the robust remote mirroring capability of HP Continuous Access XP/EVA, Cluster Extension XP/EVA will confirm that data will be available at a remote location up to metropolitan-wide or global distances, to enable business continuity without downtime or performance impact.

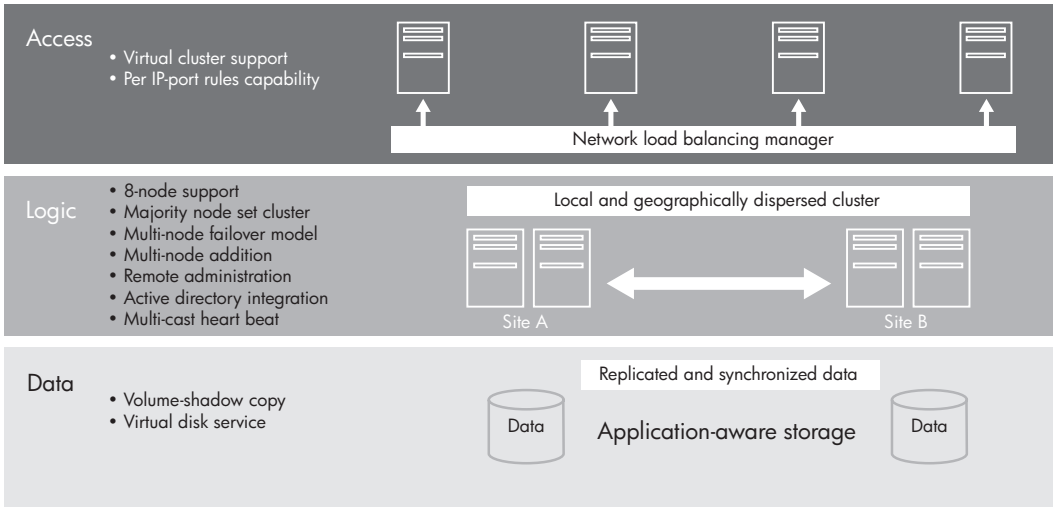


Figure 7. Windows Server 2003 technologies enable an architecture designed for high-availability and scalability.

The key word here is confirm. Without this confirmation, cluster failover will occur independently from the Continuous Access remote mirroring capability, thus requiring a labor-intensive, manual process to monitor the cluster failover and the remote mirroring process. In addition, the information provided by Continuous Access helps administrators know what recovery processes to follow in different disaster scenarios.

Viewed from this perspective, Cluster Extension XP/EVA is an integral link that brings together the capabilities of cluster software and Continuous Access remote mirroring to offer a true disaster recovery solution.

Step 4: Manage your infrastructure

Application and infrastructure availability and recovery solutions require more than hardware and software. A critical component for maintaining mission-critical environments is the streamlining of processes and the selection of a management platform that addresses the widespread maintenance and management requirements of an enterprise.

The HP IT Service Management (ITSM) Reference Model provides a platform that addresses these needs. ITSM focuses on delivering and supporting IT services that are appropriate to the business requirements of the organization. It achieves this by leveraging IT Infrastructure Library (ITIL) best practices that promote business effectiveness and efficiency. The HP ITSM Reference Model provides a tool to help develop a roadmap for transforming a corporate IT organization.

The HP ITSM Reference Model functions as a high-level, fully integrated IT process relationship map. It has proven to be invaluable to companies seeking to understand their entire IT environments, including their people, process and technology, and possible alternatives to their most critical challenges. And as a reference tool, the model provides a coherent representation of IT processes and a common language, making it useful in initiating a meaningful dialogue between all parties involved in IT process requirements and solutions.

The HP ITSM Reference Model leverages the HP portfolio of infrastructure management software and solutions. This portfolio stretches from the heterogeneous management and control provided by HP OpenView software to the systems level management, characterized by products such as HP Systems Insight Manager for the Intel® server platform.

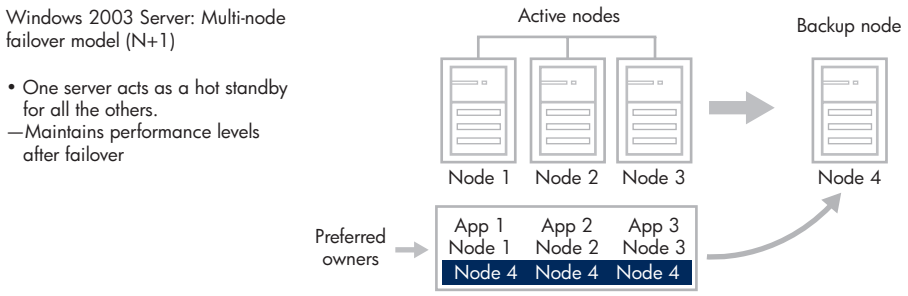
HP is moving into a new phase in the systems level management with a strategy that encompasses:

- Differentiated value delivered with the server
- A common system level management platform for HP's strategic server platform architectures and operating systems.
- Common training for administrators across multiple OS environments on IA-32 and IA-64 platforms.
- Optional plug-in value-added components to enable additional functionality in HP platforms

Additionally, HP strives to add value in platform management via:

- Virtual presence and control (iLO/RILOE)
- Availability monitoring and self healing
- Workload management
- Automated provisioning
- Trusted platform technologies

Figure 8. Windows Server 2003 enables multi-node failover for higher availability.



Getting started

HP Services

HP Business Continuity and Availability (BC&A) Program Services provide expert assistance and a structured methodology for assessing needs, analyzing options and designing solutions for Microsoft environments. These services draw on HP's extensive Microsoft skill set and years of experience with Microsoft platforms to help organizations put together high-performance business continuity and high availability solutions.

BC&A Workshops:

- Provide a streamlined methodology to help you quickly define critical requirements for business applications and continuity in a Microsoft environment
- Clearly link end-to-end application availability to supporting infrastructure requirements
- Capture organizational vision and strategy and correlate to requirements for prioritization and scheduling roadmaps
- Provide a proven starting point for a logical and justifiable process in implementing BC&A

BC&A Analysis services help you:

- Analyze business impact and the importance of availability
- Evaluate technology options against requirements and select an appropriate BC&A solution for your Microsoft environment
- Explore key cost vs. benefit sensitivities for your organization to assist in driving solution choices
- Provide a clear justification of availability investment based on business requirements

BC&A Design services:

- Enable design for a flexible high availability architecture with all the required and embedded availability capabilities for the business, such as scalability, security and lifecycle management
- Quickly leverage proven integrations and best practices
- Provide rapid skill and knowledge transfer to enable more effective ramp-up and enhance quality
- Lessen the business risk associated with targets and goals

Achieving the benefits

Around the world, private enterprises and public organizations are putting Microsoft and HP solutions to work to enable end-to-end IT service and application availability. That's the case at organizations as diverse as the London Stock Exchange and the City of Springfield, Illinois.

London Stock Exchange

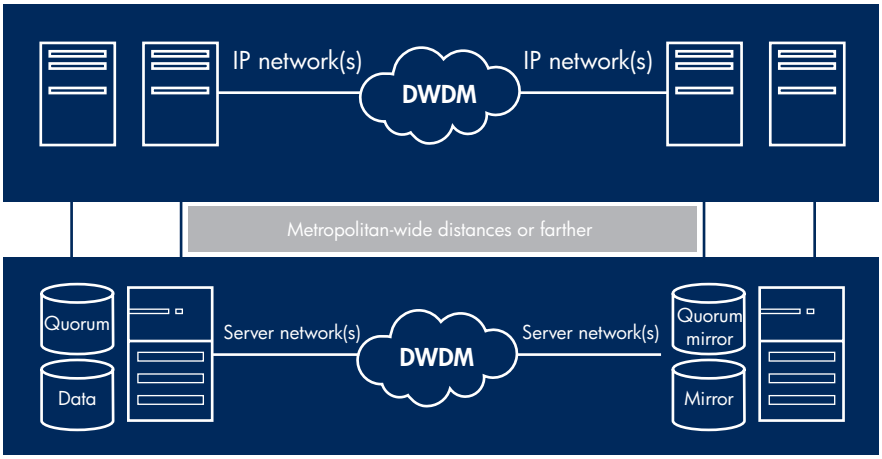
The London Stock Exchange required a system that could deliver time-critical information with lightening-fast speed. Specifically, it sought a system that would instantaneously process hundreds of messages per second with near 100 percent reliability.

The Stock Exchange found its answer in a solution that leverages the capabilities of Microsoft Windows Server 2003 Enterprise Edition, Visual Studio .NET 2003 and SQL Server 2000. The results have been dramatic. The solution can process more than 3,000 transactions per second with 300 millisecond latency. High availability is maintained with sub-second hot failover between servers with no client disruption.

Figure 9. HP Cluster Extension XP/EVA brings together the capabilities of cluster software and Continuous Access XP/EVA remote mirroring to enable a true disaster recovery solution.

HP Cluster Extension XP/EVA—benefits

- Disaster recovery to protect against the risk of downtime, whether planned or unplanned
- Automatic failover/failback to reduce the complexity involved in a disaster recovery situation.
- Ensure the highest standards in data integrity by leveraging the inherent advantages of XP disk array remote mirroring
- CLX quorum services provides market's only proven solution for quorum resource and full protection against 'split-brain' situations using arbitrator technology.
- Microsoft-certified: see Hardware Certification List (HCL)



"The system is designed for the demanding real-time environment in which market professionals work," said David Lester, CIO of the London Stock Exchange. "We are confident that the Microsoft solution has all the reliability, performance and scalability needed for many years to come."

City Water, Power & Light, Springfield, Illinois

In the United States, the City of Springfield in Illinois chose HP and Microsoft for a business-critical function—the creation of a disaster-tolerant SQL solution. The solution met the City's requirements to have much improved application availability and disaster preparedness. Availability requirements included online expansion and growth for the complete lifecycle management of the business application, not just the initial implementation and build-out.

HP and Microsoft met the need by providing application integration and infrastructure design services from HP to link two separate sites and leverage the clustering technologies of Microsoft.

"We established a reliable business-continuity plan using products and services from HP," said Scott Dragoo, GIS Systems Coordinator, CWLP, for Springfield, Illinois.

To learn more

For more information on Microsoft and HP solutions for enabling high availability and business continuity, visit:

www.hp.com/go/hpandmsbuscon

http://h71028.www7.hp.com/enterprise/cache/120851-0-0-0-121.html

For more information, go to www.hp.com/go/hpandmsbuscon

4AA0-3147ENA, November 2005