Microsoft®

Microsoft® Windows NT™ Server Cluster Strategy:

High Availability and Scalability with Industry-Standard Hardware

A White Paper
from the Business Systems Division

Microsoft®

Microsoft Windows NT Server Cluster Strategy:

High Availability and Scalability with Industry-Standard Hardware

A White Paper from the Business Systems Division

Abstract

The following paper details Microsoft's vision for enhancing Windows NT™ Server and BackOffice™ through clustering to provide greater availability and scalability. Clustering technology, combined with Windows NT Server, will bring data center capabilities and performance to a wider range of customer installations. Windows NT Server clustering will take advantage of the economics of industry-standard hardware and extend the ease of configuration and maintenance in Windows® to clustering.

The Microsoft Business Systems Division series of white papers is designed to educate information technology (IT) professionals about Windows NT and the Microsoft BackOffice family of products. While current technologies used in Microsoft products are often covered, the real purpose of these papers is to give readers an idea of how major technologies are evolving, how Microsoft is using those technologies, and how this information affects technology planners.

For the latest information on Windows NT Server, check out our World Wide Web site at ***http://www.microsoft.com/backoffice*** *or the Windows NT Server Forum on the Microsoft Network (****GO WORD: MSNTS****).*

Legal Notice

The information contained in this document represents the current view of Microsoft Corporation on the issues discussed as of the date of publication. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information presented after the date of publication.

This document is for informational purposes only. MICROSOFT MAKES NO WARRANTIES, EXPRESS OR IMPLIED, IN THIS DOCUMENT.

© 1995 Microsoft Corporation. All rights reserved.

Microsoft and Windows are registered trademarks, and Windows NT, SQL Server, and BackOffice are trademarks of Microsoft Corporation.

Alpha AXP is a trademark of Digital Equipment Corporation.
DEC is a trademark of Digital Equipment Corporation.
Intel is a registered trademark of Intel Corporation.
IBM is a registered trademark of International Business Machines Corporation.
PowerPC is a trademark of International Business Machines Corporation.
MIPS is a registered trademark of MIPS Computer Systems, Inc.

1195
Part No. 098-63018

Table of Contents

Microsoft is committed to providing a computing platform for the demands of today's enterprise business applications, as well as anticipating the needs of the most demanding information systems of tomorrow. In order to meet these needs, Microsoft is working with leading technology vendors to develop systems that provide greater availability and scalability through clustering technology.

Many large customers have used clustering technology to provide greater availability and scalability for their high-end mission-critical applications. However, these clustering solutions were complex, difficult to configure, and were built using expensive proprietary hardware. Microsoft, in conjunction with the industry, intends to bring the benefits of clustering technology to the mainstream of client/server computing. Microsoft will deliver on this vision by developing clustering technology for the Microsoft® Windows NT™ Server-based operating system based on open specifications, industry-standard hardware, and the ease-of-use customers have come to expect from Microsoft products.

Microsoft plans to deliver clustering in two phases. The first phase will allow one server to automatically "fail-over" to another server, creating a high-availability Windows NT Server environment. The second phase will extend clustering by adding dynamic scalability to the availability features offered in phase 1. The first design preview of this architecture is planned for the first half of 1996. Microsoft will also provide tools that allow application developers to create and manage enterprise applications that take advantage of clustering for Windows NT Server.

Clustering Defined

In broad terms, a cluster is a group of independent systems working together as a single system. A client interacts with a cluster as though it were a single server. Cluster configurations are used to address both availability and scalability.

Availability. When a system in the cluster fails, the cluster software will respond by dispersing the work from the failed system to the remaining systems in the cluster.

Scalability. When the overall load exceeds the capabilities of the systems in the cluster, additional systems may be added to the cluster. Formerly, customers that desired future system expansion capability needed to make up front commitments to expensive, high-end servers that provided space for additional CPUs, drives, and memory. With clustering, customers can add systems as needed to meet overall processing power requirements.

Why is Clustering Important?

It is estimated that system "downtime" costs U.S. businesses $4.0 billion per year[1]. The average downtime event results in a $140,000 loss in the retail industry and a $450,000 loss in the securities industry[2]. Clustering promises to minimize downtime by providing an architecture that keeps systems running in the event of a single system failure. Clustering also affords organizations the ability to aggregate separate servers into a single computing facility, allowing IT organizations the flexibility to grow installations beyond a single machine.

[1] SVP Strategic Research Division Report, 1992
[2] *ibid.*

Clustering Illustration—Retail Industry/Data Availability

In a retail operation, the point-of-sales system is the heartbeat of the business. Cashiers require ongoing access to the store's database of products, codes, names, and prices to keep the business logging sales. If the point-of-sale system fails, sales cannot be logged and the operation loses money, customers, as well as its reputation for quality service.

In this case, clustering technology can be utilized to deliver system availability. The clustering solution would allow a pair of servers to access the multi-port storage devices (Disk Array) on which the database resides. In the event of a server failure[3] on Server 1, the backup system (Server 2) is automatically brought online and end users are switched over to the new server—without operator intervention. Thus, downtime is minimized. The fault-tolerant disk technology built into Windows NT Server 3.51 (striping, duplexing, etc.) would protect the disk array, but with the addition of clustering technology, the overall system would remain online.

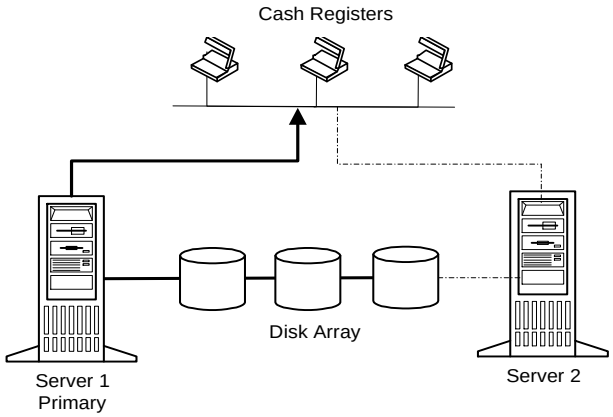

Figure 1: Retail branch before failure

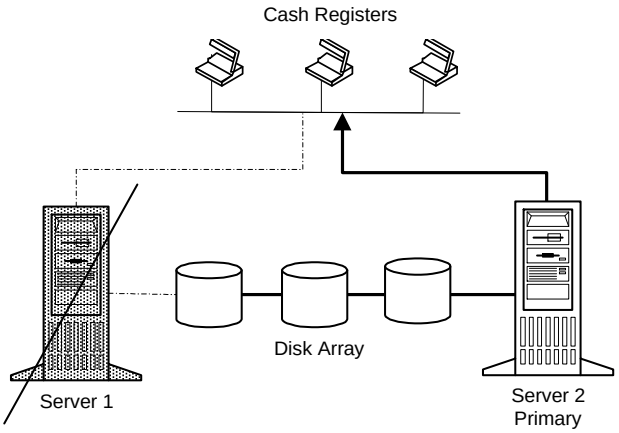

Figure 2: Retail branch during failure

[3] Server failure due to hardware (CPU/Motherboard, storage adapter, network card, etc.), application failure, or operator error

Clustering Illustration—Financial Services Industry/Scalability

It has been said, "the two greatest fears of an Chief Information Officer (CIO) are system success and failure." If a system fails, the CIO's staff is inundated with complaints. Conversely, if a system is successful, usage demands may outstrip the needs of the system as it grows. Clustering, as has been discussed, can greatly minimize system downtime. In addition, clustering can also help Information Technology (IT) departments design systems that can grow with the demands of an organization.

For example, in recent years billions of dollars have poured into mutual funds companies. Although this type of growth is positive in financial terms, the technological burden of managing information systems growth can be overwhelming. As a result, Chief Information Officers and their staffs are faced with developing systems that not only meet current systems demand, but also provide for future systems growth. Formerly, the system choices were rather limited: extremely expensive mainframes and minicomputers.

Windows NT Server clustering can provide a competitive advantage for the IT department, allowing faster system deployment, automatic re-tasking, and easier maintenance with smaller staffs—all while using inexpensive PC components. These components, available from multiple sources, not only help ensure competitive pricing but also help ensure parts availability. Consequently, IT departments can expand their hardware as needed without the burden of single supplier shortages.

Clustering technology also gives IT departments greater flexibility. With clustering, multiple servers can be tied together in one system, and additional servers can be integrated into the system as usage requirements dictate. Windows NT Server clustering gives a choice to system architects that they have never enjoyed before—availability and scalability on inexpensive, mainstream platforms.

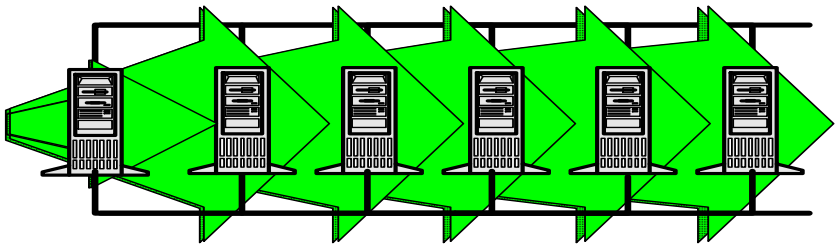

Figure 3: For increased system performance, clusters will allow customers to scale their information systems incrementally, adding processing power as needed.

Traditional Architectures for High-Availability

Today, several architectures are commonly used for achieving increased availability in computer systems. One traditional hardware structure for achieving high availability is duplicate systems with fully replicated components. The traditional software model for utilizing this hardware is one in which one system runs the application while the other sits idle, acting as a standby to take over when the primary system fails. The drawbacks of this approach include increased hardware costs, with no improvement in system throughput, and the lack of protection from intermittent application failures.

Traditional Architectures for Scalability

Several different architectures have been used to enhance scalability. One hardware structure for achieving scalability beyond a single processor is the symmetric multiprocessor (SMP) system. In an SMP system, several processors share a global memory and I/O subsystem. The traditional SMP software model, known as the "shared memory model," runs a single copy of the operating system with application processes running as if they were on a single processor system. If non-data-sharing applications are run on a SMP system, the systems will provide high scalability.

At the hardware level, the major drawback to SMP systems is that they encounter physical limitations in bus and memory speed that are expensive to overcome. As microprocessor speeds increase, shared memory multiprocessors become increasingly expensive. Today there are large price "steps" as a customer needs to scale from one processor to 2-4 processors, and especially when scaling beyond 8 processors. Finally, neither the SMP hardware structure nor its traditional software model provide inherent availability benefits over single processor systems.

There is one architecture that has proven itself for availability and scalability in business-critical computing applications: the cluster.

Cluster Architecture

A cluster is a set of loosely coupled, independent computer systems that behave as a single system. Clients see a cluster as if it were a single high-performance, highly reliable server. System managers see a cluster much as they see a single server. Cluster technology is readily adaptable to low cost, industry-standard computer technology and interconnects.

Clustering can take many forms. A cluster may be nothing more than a set of standard desktop personal computers interconnected by an Ethernet. At the high end of the spectrum, the hardware structure may consist of high-performance SMP systems interconnected via a high-performance communications and I/O bus. In both cases, adding processing power is done in small steps by the addition of another commodity system. To a client, the cluster provides the illusion of a single server, or *single-system image*, even though it may be composed of many systems. Additional systems can be added to the cluster as needed to process more complex or an increasing number of requests from the clients. If one system in a cluster fails, its workload can be automatically dispersed among the surviving systems. Frequently, this is transparent to the client.

Shared Disk Model

Two principal software models are used in clustering today: shared disk and shared nothing. In the *shared disk* model, software running on any system in the cluster may access any resource (e.g., a disk) connected to any system in the cluster. If two systems need to see the same data, the data must either be read twice from the disk or copied from one system to another. As in an SMP system, the application must synchronize and serialize its access to shared data. Typically a Distributed Lock Manager (DLM) is used to help with this synchronization. A DLM is a service provided to applications to track references to resources throughout the cluster. If more than one system attempts to reference a single resource, the Lock Manager will recognize and resolve the potential conflict. DLM coordination, however, may cause additional message traffic and reduce performance due to the associated serialized access to additional systems. One approach to reducing these problems is the *shared nothing* software model.

Shared Nothing Model

In the *shared nothing* software model, each system within the cluster owns a subset of the resources of the cluster. Only one system may own and access a particular resource at a time, although, on a failure, another dynamically determined system may take ownership of the resource. In addition, requests from clients are automatically routed to the system that owns the resource.

For example, if a client request requires access to resources owned by multiple systems, one system is chosen to host the request. The host system analyzes the client request and ships sub-requests to the appropriate systems. Each system executes the sub-request and returns only the required response to the host system. The host system assembles a final response and sends it to the client.

A single system request on the host system describes a high-level function (e.g., a multiple data record retrieve) that generates a great deal of system activity (e.g., multiple disk reads) and the associated traffic does not appear on the cluster interconnect until the final desired data is found. By utilizing an application that is distributed over multiple clustered systems, such as a database, overall system performance is not limited by a single machine's hardware limitations.

The shared disk and shared nothing models can be supported within the same cluster. Some software can most easily exploit the capabilities of the cluster through the shared disk model. This software includes applications and services which require only modest (and read intensive) shared access to data as well as applications or workloads that are very difficult to partition. Applications which require maximum scalability should use the cluster's shared nothing support.

Cluster Application Servers

While clusters bring availability and scalability to all server-based software, "cluster-aware" applications can take full advantage of the benefits of a cluster environment. Database server software must be enhanced either to coordinate access to shared data in a shared disk cluster, or to partition a SQL request into a set of sub-requests in a shared nothing cluster. In a shared nothing cluster, the database server may want to take further advantage of the partitioned data by making intelligent, parallel queries for execution across the cluster. The application server software may also be enhanced to detect component failures and initiate fast recovery via cluster APIs.

Windows NT Server Today

Windows NT Server provides all the components necessary to support mission-critical applications. The system was built on a fully 32-bit microkernel foundation. It is multithreaded, offers preemptive multitasking, and provides memory protection for both applications and the operating system itself. It scales to run on hardware with up to 32 processors, 4 GB of RAM, and 17 million terabytes of disk space. In addition, Windows NT Server supports Intel® X86, MIPS® R4x00, DEC™ Alpha AXP™, and IBM® PowerPC™ processors.

Microsoft's Vision

Microsoft's vision is to enhance the Windows NT Server platform to support clustering for a broad base of customers who would benefit from a cost-effective method of delivering increased availability and scalability. Microsoft believes the following factors must be provided to enable broad market acceptance:

- **Industry standard APIs.** Microsoft, in conjunction with technology partners, will work to establish industry standards for clustering Application Programming Interfaces (APIs). The cluster APIs will be developed to expose specific cluster features for software developers to use in developing high-availability applications, and in the future, more scaleable applications. File, print and database servers, transaction processing monitors, and other software will be able to use the cluster APIs to exploit fully the capabilities of the Windows NT cluster.
- **Industry standard hardware**. Windows NT Server clusters takes advantaage of today's industry-standard PC platforms and existing network technology. The Windows NT layered driver model will allow Microsoft to add support quickly for special purpose high-performance clustering technology (e.g., low-latency interconnects) as hardware vendors bring solutions to market.
- **Server application support.** Microsoft BackOffice products will be enhanced to use the clustering API and take full advantage of the scalability and availability characteristics of clusters. Of course, Microsoft will encourage other vendors to leverage Windows NT Server clusters.
- **Cluster enhancement without enterprise disruption.** Because Windows NT Server already implements a cluster-compatible security and user administration model, businesses can easily augment a current Windows NT Server installation with clustering without user disruption. In addition, cluster administration will be exposed through enhancements to existing Windows NT Server administration.

- **Ease of configuration and maintenance**. Clusters must be simple to configure and maintain with non-dedicated support staff. Windows NT Server Clustering will take advantage of the existing central management capabilities of Windows NT Server. Once Windows NT Server cluster is installed, cluster management will be performed with a series of graphical cluster and network management tools included with the operating system.

Windows NT Server Clusters

Windows NT Server 3.51 already contains many of the basic components for constructing a clustered system. The single-logon capability inherent in Windows NT domains, the multi-system monitoring capability of the administration tools and the performance monitor, and the ability to route requests via the redirector, are all examples of basic cluster features.

Windows NT Server cluster enhancements represent a spectrum of technologies that will continue to be phased into the Windows NT Server and BackOffice products over time. Microsoft has prioritized additional clustering features based on customer requirements.

Deploying Cluster Technology

Microsoft is developing cluster Application Programming Interfaces (APIs) that will allow applications to take advantage of Windows NT Server in a clustered environment. The company will deliver clustering products in two phases:

Phase 1: Fail-over solution

A fail-over solution improves data availability by allowing two servers to share the same hard disks within a cluster. When a system in the cluster fails, the cluster software will recover and disperse the work from the failed system to another server within the cluster. As a result, the failure of a system in the cluster will not affect the other systems, and in most cases, the client applications will be completely unaware of the failure. This means high server availability for the users.

Phase 2: Multiple node solution

Phase 2 will enable more than two servers to be connected together for higher performance and reliability. As a result, when the overall load exceeds the capabilities of the systems in the cluster, additional systems may be added to the cluster. This incremental growth enables customers to add processing power as needed. Formerly, customers had to make up-front commitments to expensive, high-end servers that provided space for additional CPUs, drives, and memory.

The design will support multiple nodes, but the initial product may be limited to two nodes. All Windows NT Server (x86, MIPS, ALPHA, PPC) processor architectures will be supported. Later versions of the product will increase the number of nodes supported and improve scalability.

Cluster features will be integrated into existing Windows NT Server system management tools to enable system administrators who are already familiar with Windows NT Server systems to easily setup and configure their clusters. The initial product will include base operating system support for clusters including components to configure, maintain, and monitor membership in the cluster, support for a cluster-wide name-space, communication, and failover support. Additional services will support the two primary cluster software models. As with domain management, ease of setup and cluster management tools will be a very high priority.

Microsoft BackOffice

The Microsoft BackOffice™ family of components, such as Microsoft SQL Server™ and Microsoft Exchange Server, will initially be enhanced to support the cluster's failover capability. Later releases of BackOffice are intended to support the scalability aspects of the cluster. Microsoft SQL Server is planned to support a *partitioned data model* and parallel execution to take full advantage of the shared nothing environment.

Cluster Application Development

Microsoft development tools will be enhanced to support the easy creation of cluster-aware applications. Facilities will be provided for automatic failover of applications. It is also important to note that not all server applications will need to be cluster-aware to take advantage of cluster benefits. Applications that build on top of cluster-aware core applications, such as large commercial database packages (e.g., an accounting or financial database application on top of SQL Server), will benefit automatically from cluster enhancements made to the underlying application (e.g., SQL Server). Many server applications that leverage database services, client/server connection interaction, and file and print services will benefit from clustering technology without application changes.

Future Steps

Microsoft is committed to providing open industry specifications for clustering technology. Through public APIs and driver models, Microsoft and the computer hardware and software industries will work together to provide robust economical standardized clustering solutions.

Specifically, Microsoft will host an open process forum in 1996 to help build industry consensus on clustering requirements that meet a broad base of customer needs. Open process is the mechanism to drive industry standard interfaces that result in multi-vendor compatibility, customer supplier independence, and competition that can bring clustering technology to a larger business data center audience. Participation will be from Independent Hardware Vendors (IHVs), Independent Software Vendors (ISVs) and computer Original Equipment Manufacturers (OEMs). In addition, hardware details on clustering requirements will be presented at the Windows Hardware Engineering Conference (WinHEC) in April 1996.

Summary

Microsoft will work with the industry to deliver clustering for the Microsoft Windows NT Server network operating system and the Microsoft BackOffice. Clustering technology will enable customers to connect a group of servers together to improve data availability/fault tolerance and performance, using industry-standard hardware components. The goal is to continue to build upon the strengths of Windows NT Server as an enterprise server, and to offer customers the greatest flexibility to design, develop, and implement systems for the most demanding business needs of the future.

---

*For the latest information on Windows NT Server, check out our World Wide Web site at **http://www.microsoft.com/backoffice** or the Windows NT Server Forum on the Microsoft Network (**GO WORD: MSNTS**).*