®

Microsoft®

Windows NT® Server, Enterprise Edition

Server Operating System

White Paper

Microsoft Windows NT Load Balancing Service

Abstract

Microsoft® Windows NT® Load Balancing Service (WLBS), a feature of Windows NT Server 4.0, Enterprise Edition, provides load balancing and clustering for mission critical enterprise-class applications. WLBS dynamically distributes IP traffic across multiple cluster nodes, and provides automatic failover in case of node failure. WLBS also provides multihomed server and rolling upgrade support, ease of use and controllability. WLBS runs on economical, industry standard PC platforms and interoperates with both Microsoft and third-pary clustering products to provide a complete three-tier solution.

Microsoft Corporation • One Microsoft Way • Redmond, WA 98052-6399 • USA

CONTENTS

INTRODUCTION

Microsoft® Windows NT® Load Balancing Service (WLBS), a feature of Windows NT Server 4.0, Enterprise Edition, provides scalability and high availability to enterprise-wide Transmission Control Protocol/Internet Protocol (TCP/IP) services, such as Web, proxy, Virtual Private Networking (VPN), and streaming media services. WLBS brings special value to enterprises deploying intranets and other TCP/IP services to link clients with transaction applications and back-end databases. WLBS is based on the award-winning Convoy Cluster Software by Valence Research, Inc., a recent Microsoft acquisition.

WLBS servers (also called *nodes*) in a cluster communicate among themselves to provide for key benefits, including:

- High Availability – WLBS cluster servers emit a “heartbeat” to other nodes in the cluster, and listen for the heartbeat of other nodes. If a server in a cluster fails, the remaining nodes adjust and take over the workload.
- Load balancing – WLBS servers use a distributed algorithm to statistically map workload between the nodes of the cluster. When a WLBS cluster is started, or when a cluster node fails, and remaining nodes react to the absence of its heartbeat, the cluster nodes perform a process called “convergence” in which they communicate to determine the status of the cluster and determine which nodes are available for load balancing.
- Scalability – WLBS scales to meet the demands of the service. As traffic increases, just add another server into the cluster, with up to 32 servers possible in any one cluster.

WLBS allows incoming IP traffic to be dynamically distributed across the cluster nodes. Performance is scaled by running a copy of the TCP/IP service, such as a Web server, on each host within the cluster. WLBS transparently distributes the client requests among the hosts and lets the clients access the cluster using one or more “virtual” IP addresses. From the clients’ point of view, the cluster appears to be a single server that answers these IP addresses. As enterprise traffic increases, network administrators can simply plug another computer into the cluster. Because WLBS has no proprietary hardware requirements, any industry standard compatible computer can be used. This provides significant savings when compared to proprietary hardware load balancing solutions.

WLBS is superior to other software solutions such as round robin DNS, in which workload is distributed across servers but no process for server availability is implemented. With RRDNS, if a server within the node fails, it will continue to be sent work until a network administrator detects the failure and removes the server from the DNS, resulting in service disruption for clients. WLBS also is an important improvement over other load balancing solutions – both hardware- and software-based -- which introduce single points of failure or performance bottlenecks by using a centralized dispatcher.

WLBS’s unique, fully distributed, software architecture enables it to deliver the industry's best load balancing and high-availability performance. WLBS provides an economical solution for the reliability and performance that Web servers and other mission-critical IP services demand.

DEFINING CLUSTERING AND LOAD BALANCING

Internet server programs supporting mission-critical applications such as financial transactions, database access, corporate intranets, and other key functions must run 24 hours a day, seven days a week. And networks need the ability to scale performance to handle large volumes of client requests without creating unwanted delays.

For these reasons, clustering is of wide interest to the enterprise, and much has been written about it, but there is often confusion about exactly what is meant by terms such as *clustering* and *load balancing.*

Clustering

A “server cluster” is a group of independent servers managed as a single system for higher availability, easier manageability, and greater scalability. The minimum requirements for a server cluster are:

- Two (or more) servers connected by a network.
- Cluster management software, such as WLBS or Microsoft Cluster Service (MSCS).

The cluster management software provides services such as failure detection, recovery, load balancing, and the ability to manage the servers as a single system.

Server clustering provides a number of important benefits, including improved availability, easier manageability, and cost-effective scalability. For example, WLBS provides high availability by automatically detecting the failure of a server, allowing users to experience continuous service without interruption.

Load Balancing

“Load balancing” is a technique that allows the performance of a server-based program, such as a Web server, to be scaled by distributing its client requests across multiple servers within the cluster. With WLBS, each incoming IP packet is “seen” by each node, but only accepted by the intended recipient. Each WLBS host can specify the load percentage that it will handle, or the load can be equally distributed across all of the hosts. If a host fails, the load balancing mechanism dynamically redistributes the load among the remaining hosts.

Load balancing is used to enhance scalability, which boosts throughput while keeping response times low. By using WLBS, more IP traffic can be handled by simply by adding computers to the cluster as necessary. While WLBS is used for load balancing IP traffic, Microsoft COM+ (available in the Windows® 2000 platform) provides load balancing for applications that use COM objects.

WLBS, MSCS and COM+

To provide enterprise networks with a more complete set of clustering solutions, Microsoft provides both WLBS and MSCS for Windows NT Server 4.0, Enterprise Edition. A third clustering service, COM+, will be introduced with the Windows 2000 operating system. WLBS, COM+, and MSCS can be used separately or combined into powerful enterprise solutions.

- WLBS is intended primarily to load-balance incoming TCP/IP traffic. WLBS is most commonly used for running Web server programs.
- COM+ provides dynamic load balancing for applications that support COM+ by allocating COM+ application objects to the cluster node with the lightest processing load. COM+ is ideal for scaling mission critical applications across clustered servers.
- MSCS, formerly known by its code name “Wolfpack,” is ideal for clustering database services such as SQL Server™ and other data-intensive applications that require high availability.

Providing Three-Tier Clustering

Microsoft’s three forms of clustering can be used individually or grouped together, according to an organization’s needs. This is a powerful way to combine back-end database and transactional systems with a Web-based front end, to deliver the scalability, availability and robustness demanded by enterprise class customers. The three types of clustering provide ideal support for three-tier applications. Using the example of a Web-based retail business, the clustering could be used as follows:

- WLBS could be used on the first tier, load balancing access to the service by clustering Web servers.
- COM+ could be used on the second tier, load balancing the shopping basket logic across a cluster of application servers.
- MSCS could be used on the third tier, providing high availability for the billing and ordering databases.

WLBS is used to provide scalability and availability for the front end client UI presentation layer, by balancing and distributing client TCP/IP connections over multiple servers.

Intermediate (logic) tier load balancing is greatly facilitated by the use of Microsoft COM+, which uses a COM+ Router, through which all communication to an application cluster passes. The COM+ Router provides load balancing by determining where an instance of a COM Object must be created, based on server factors such as CPU response time, and creating an instance of the needed object on that server. Applications must be designed to allow COM+ objects to cooperate, by using standard methods to control the sequence of events.

Microsoft Cluster Service can provide two-node failover for the back-end application and data service layers of a three-tier application, creating a reliable platform for database, messaging, and similar application services.

Stateless and Stateful Servers

Under various conditions, networking clients may use either "stateless" or “stateful” services. Stateless servers do not update a server-side database based on a client request. Examples include standard Web servers with static Web pages, terminal, proxy, and read-only file servers. Because multiple stateless servers can independently handle portions of the client load, WLBS can transparently load-balance these servers, providing an elegant scalability solution.

In contrast, stateful servers typically update a server-side database when serving a client request. When multiple stateful servers are used, they must coordinate updates to avoid conflicts and keep shared data consistent. Examples include Web servers that use active server pages to update files, database servers, email servers, and read/write file servers. WLBS generally cannot be used to directly load-balance client requests across stateful servers. Instead, WLBS can be used to load-balance stateless, first tier services, such as Web servers, which in turn call a single, third tier database server. Electronic commerce applications typically use this server architecture. Most of the performance bottleneck typically is in the Web server. WLBS scales the performance of this first tier service and ensures that it provides continuous service to clients. In addition, the back-end database server can be made highly available by using Microsoft Cluster Service.

Comparing Other Load balancing Methods

A number of load balancing methods exist, and WLBS has unique advantages relative to each of them. A good way to appreciate the power of WLBS is to compare it to other methods.

Round Robin DNS

Round Robin DNS (RRDNS) is a popular solution for enabling a limited form of TCP/IP load balancing for Internet server farms. However, it does not function effectively as a high availability solution. RRDNS uses DNS to map incoming IP requests to a defined set of servers in a round robin fashion. In the event of a server failure, RRDNS continues to route requests to the failed server until it is manually removed from DNS, and even then many users must wait for DNS to time out their connections before being able to successfully access the target Web site.

Hardware-Based Load Balancing

WLBS and hardware load-balancers typically offer very similar functionality; both services transparently direct client requests for a single IP address to multiple hosts within a cluster. Hardware load-balancers typically use a technique called Network Address Translation (NAT), which exposes one or more “virtual” IP addresses to clients and forwards data for the hosts behind it by translating IP addresses and resending network packets. This technique introduces a single point of failure between the cluster and the clients. To provide high availability, a backup load-balancer is needed. Address translation and retransmission also imposes relatively higher overhead and delivers lower bandwidth than does WLBS’s filtering technique.

WLBS has three distinct advantages over hardware-based load-balancers:

- WLBS has significantly higher throughput. WLBS has been demonstrated at over 250 megabits per second (Mbps).
- Unlike hardware load-balancers, which have limited maximum performance, WLBS’s performance automatically improves with the speed of the cluster

hosts and Local Area Network (LAN), which ensures that it will never become a bottleneck as LAN and host speeds increase. WLBS can easily accommodate a switch to Gigabit Ethernet without requiring any upgrade.

- WLBS avoids a single point of failure, since it runs in parallel on all hosts in the cluster. Most hardware load-balancers require two hardware units to avoid a single point of failure, with the second equally expensive hardware component operating in a passive mode.

"Dispatcher" Software Load Balancing

Software-based load balancing products that employ various "dispatcher" models for load balancing face many of the disadvantages found in hardware load balancing products. Dispatching techniques, whether implemented by NAT or other methods (such as HTTP redirects), introduce higher overhead than does WLBS. This limits throughput and restricts performance. Also, the entire cluster's throughput is limited by the speed and processing power of the "dispatch" server.

These products require all incoming connections to be handled by one "dispatch" server, where they are then either handled by that system or farmed out to other servers in the network. This introduces a single point of failure which must be eliminated by moving the dispatching function to a second computer after a failure occurs. This recovery technique adds complexity and recovery time to the solution. Only WLBS's fully distributed software load balancing model avoids this problem by removing any dependency on a single server or node.

WLBS FEATURES

WLBS operates in a fully transparent manner to both server applications and to TCP/IP clients. WLBS lets users employ off-the-shelf software components, such as existing WWW, FTP, or proxy servers and other popular Internet applications. WLBS enhances fault-tolerance and scales performance transparently to the TCP/IP protocol, to server applications, and to clients.

WLBS gives the user flexible, fine-grained control over its operations. It lets the system administrator configure network traffic handling parameters for individual load-balanced TCP/UDP ports and for sequential groups of ports (called port ranges). Individual servers can also be commanded to dynamically (without re-booting the server) leave and join the cluster by issuing simple commands.

Scalable Performance

WLBS load-balances incoming TCP/IP traffic across all the hosts in a cluster, to scale performance. Running a copy of the server program on each load-balanced host enables the load to be partitioned among them in any manner an administrator chooses. WLBS transparently distributes the client requests among the hosts and lets the clients access the cluster using one or more “virtual” IP addresses. Up to 32 hosts may operate in each cluster, and hosts may be added transparently to a cluster to handle increased load. WLBS can also direct all traffic to a designated single host, called the default host. In addition, WLBS may optionally load-balance multiple requests from a single client (see “Affinity and Session Support” below).

The robust, fully pipelined implementation of WLBS ensures high performance and low overhead, through the use of an efficient synchronization technique. Each WLBS host broadcasts a packet approximately every second to tell other hosts its status, leaving almost all network bandwidth available for client/server communications. This unique, fully distributed, software architecture avoids the use of a centralized dispatcher, which enables WLBS to deliver the industry's best performance, in addition to fault-tolerance, as described below.

High Availability

WLBS manages TCP/IP traffic to maintain high availability for IP-based services. When a host fails or goes offline, WLBS automatically reconfigures the cluster to direct client requests to the remaining computers. In addition, for load-balanced ports, the load is automatically redistributed among the computers still operating, and ports with a single server have their traffic redirected to a specific host. Such redistribution of the workload typically takes less than ten seconds. While connections to the failed or offline server are lost, once the necessary maintenance is completed, the offline computer can transparently rejoin the cluster and regain its share of the workload. In addition, WLBS handles inadvertent subnetting and rejoining of the cluster network. This robust fault tolerance is enabled by a unique distributed architecture, which avoids the single points of failure or performance bottlenecks of other load balancing solutions.

Controllability

WLBS offers exceptional controllability, because load balancing can be specified for a single IP port or group of ports using straightforward port management rules that tailor the workload for each service. In addition, optional support for client sessions can be enabled, as well as optional port rules to let all client requests be directed to a single host to further refine load balancing among different applications. Undesired network access can also be blocked to certain IP ports, and WLBS logs all actions and cluster changes to the Windows NT event log.

Remote Controllability

Administrators can remotely start, stop, and control WLBS actions from any networked Windows NT-based computer using console commands or scripts. All cluster hosts can be controlled with one command, or controlled individually. The control program has fully encrypted password protection to prevent unauthorized access.

Ease of Use

WLBS is easy to use because it installs as a standard Windows NT networking driver component and requires no hardware changes to install and run. No special hardware is needed to interconnect cluster hosts; the cluster exchanges status messages over a single local area network using Ethernet (10, 100, or gigabit) or FDDI adapter cards. WLBS also lets clients access the cluster with a single Internet logical name and IP address, while retaining individual names for each computer. In addition, server applications need not be modified to run in a WLBS cluster, and all operations, including recovery, require no human intervention. Computers can also be taken offline for preventive maintenance without disturbing cluster operations.

Affinity and Session Support

WLBS supports client sessions and Secure Sockets Layer (SSL). If a server application (such as a Web server) maintains state information about a client session that spans multiple TCP connections, it is important that all TCP connections for this client be directed to the same cluster host. Should a server or network failure occur during a "stateful" client session, a new logon may be required to re-authenticate the client and re-establish session state.

WLBS also allows modification of session support to direct all client requests from a TCP/IP Class C address range to a single cluster host. This feature ensures that clients which use multiple proxy servers to access the cluster will have their TCP connections directed to the same cluster host. The use of multiple proxy servers at the client’s site causes requests from a single client to appear to originate from different systems. Assuming that all of the client’s proxy servers are located within the same 256 host Class C address range, WLBS ensures that client sessions are properly handled with minimum impact on load distribution among the cluster hosts.

Supports Rolling Upgrades

WLBS supports rolling upgrades to allow software or hardware upgrades without shutting down the cluster or disrupting service. Upgrades can be individually applied to each server, which immediately rejoins the cluster.

WLBS ARCHITECTURE

WLBS installs as a standard Windows NT networking driver and runs on an organization's existing LAN. All servers within a cluster are placed on a single subnet. Once installed, WLBS operates in a fully transparent manner to both the server applications and to TCP/IP clients. Clients can access the cluster as if it is a single computer by using one IP address.

The clustered nodes in Figure 1 below work together to service network traffic from the Internet. Each server runs a copy of the World Wide Web (WWW) server, or other application, and WLBS distributes the networking workload among them. This speeds up normal processing so that Internet clients see faster turnaround on their requests. The back-end application (database) may operate on a cluster running Microsoft Cluster Service.

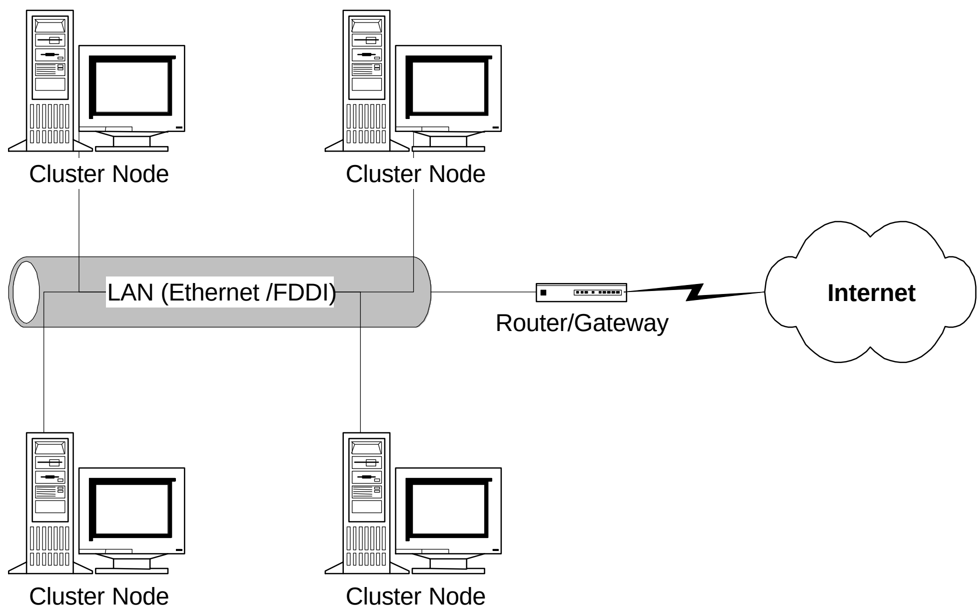

Figure 1. A four node cluster works as a single virtual server to handle network traffic. Each node runs its own copy of the server, with WLBS load balancing the work between the four nodes.

If one server should fail or go offline, WLBS automatically detects the problem and redirects network traffic to the remaining servers, thereby ensuring that service is not interrupted while at least one server is available. Recovery requires less than 10 seconds. Active connections to a failed server are lost when the server goes offline; all other connections are unaffected.

Parallel Filter Architecture

WLBS uses Level 2 multicast (or optional promiscuous reception with unicast MAC addresses) to enable all cluster hosts on a single subnet to simultaneously listen to incoming network traffic for the cluster's primary "virtual" IP address (and for additional IP addresses on multi-homed hosts). On each cluster host, the WLBS driver acts as a filter between the cluster NIC driver and the TCP/IP protocol stack to allow a portion of the incoming network traffic to be received by the local host.

Distributed Algorithm

WLBS employs a fully distributed algorithm to statistically map incoming clients to the cluster hosts based on their IP address, port, and other information. When inspecting an arriving packet, all hosts simultaneously perform this mapping to quickly determine which host should handle the packet. The mapping remains unchanged unless the number of cluster hosts changes. WLBS's filtering algorithm is much more efficient in its packet handling than centralized load-balancers, which must modify and retransmit packets. This enables it to offer much higher aggregate bandwidth. By running directly on the cluster hosts, WLBS's performance is not limited by a specific generation of processor and/or network technology.

Expansion of the cluster does not affect ongoing cluster operations and is achieved transparently to both Internet clients and to server programs. However, it may affect some client sessions that span multiple TCP connections when client affinity is selected because a subset of clients may be remapped to different cluster hosts between connections.

Convergence

To coordinate their actions, WLBS hosts periodically exchange multicast or broadcast messages within the cluster. This allows them to monitor the status of the cluster. When the state of the cluster changes (when hosts fail, leave, or join the cluster), WLBS invokes a process known as "convergence," in which the hosts exchange messages to determine a new, consistent state of the cluster and to elect the host with the highest host priority, known as the default host. When all cluster hosts have reached consensus on the correct new state of the cluster, they record the completion of convergence in the Windows NT event log.

During convergence, the available hosts continue to handle incoming network traffic. At the completion of convergence, the traffic for a failed host is redistributed to the remaining hosts. Load-balanced traffic is repartitioned among the remaining hosts to achieve the best possible new load-balance for specific TCP or User Datagram Protocol (UDP) ports. If a host is added to the cluster, convergence allows this host to take over handling ports for which it has the highest priority and to receive its share of the load-balanced traffic.

WLBS assumes that a host is functioning properly within the cluster as long as it participates in the normal "heartbeat" message exchange among the cluster hosts. If other hosts do not hear from any member for several periods (or seconds) of message exchange, they initiate convergence to redistribute the load previously handled by the failed host. The system administrator can control both the message exchange period and the number of missed messages required to initiate convergence. Their default values are respectively set to 1 second and 5 missed periods.

Using a Second NIC

To ensure maximum network performance, WLBS normally uses a second network interface card (NIC) to handle cluster traffic (including client requests, server responses, and convergence messages), while other network traffic to the server goes through a separate NIC. A second network card is only required if level 2 multicast is not selected and the clustered servers need to communicate with each other. However, using a second NIC typically boosts overall performance, especially when the servers access a networked, back end database or replicate files within the cluster.

Supporting Application Protocols

WLBS controls the distribution and partitioning of TCP and UDP traffic from Internet clients to selected hosts within a cluster as follows. After WLBS has been configured, incoming client requests to the cluster IP addresses are received by all hosts within the cluster. WLBS filters incoming datagrams to specified TCP and UDP ports before these datagrams reach the TCP/IP protocol software. WLBS manages only the TCP and UDP protocols within TCP/IP, and its actions are controlled on a per-port basis.

WLBS does not control any incoming IP traffic other than TCP and UDP traffic for specified ports. It does not filter Internet Control Message Protocol (ICMP), Internet Group Membership Protocol (IGMP), Address Resolution Protocol (ARP) (except as described below), or other IP protocols. All such traffic is passed unchanged to the TCP/IP protocol software on all of the hosts within the cluster. Because of the robustness of TCP/IP and its ability to deal with replicated datagrams, other protocols behave correctly in the clustered environment. However, duplicate responses should be expected from certain point-to-point TCP/IP programs (such as Ping) when the cluster IP address is used. These programs can use the dedicated IP address for each host to avoid this behavior.

WLBS runs as a network driver without knowledge of higher-level application protocols, such as HTTP and FTP, that use TCP or UDP for transport. WLBS does not directly monitor server applications, such as Web servers, for continuous and correct operation. Instead, it provides the mechanisms needed by application monitors to control cluster operations, for example, to remove a node from the cluster if an application fails or displays erratic behavior. Monitoring services are widely available for most client/server applications.

WLBS PERFORMANCE

Performance measurements have validated that WLBS's efficient, software implementation introduces negligible overhead to network traffic handling and delivers scaled performance very close to theoretical limits. For typical Web server loads and today's CPU speeds, four- to eight-way clusters are likely to fully utilize a fast Ethernet hub and DS3 Internet connection. Further throughput gains can be obtained by using larger clusters along with fast Ethernet switches or gigabit Ethernet.

The unique architecture of WLBS as a fully distributed, software-only solution with no centralized bottlenecks offers several important advantages. It eliminates any single points of failure in the WLBS implementation, and it lets the performance of WLBS scale without predefined limits as faster host CPU's and cluster networks are employed. These characteristics make WLBS the solution of choice for both high performance and high availability.

Performance Measurements

An identical copy of WLBS runs on each host as a Windows NT intermediate driver sandwiched between the TCP/IP protocol and the network adapter's Network Driver Interface Specification (NDIS) driver. Because WLBS runs beneath higher-level application protocols, such as Hypertext Transfer Protocol (HTTP), it makes simple and fast load balancing decisions that introduce very little overhead to packet reception. Its performance impact on the cluster hosts is most easily measured by observing its effect on CPU load. Measurements were made of the average CPU load per cluster host versus the average throughput per host using a simulated Web load of 100 byte requests and 10 kilobyte (KB) replies with minimum server application delay. The tests used single processor, 200 megahertz (MHz) Pentium systems interconnected over a 100 Mbps LAN.

Performance measurements showed WLBS had a negligible impact on CPU load, and in fact, WLBS running on a single host actually slightly reduced CPU load in the tests. This improvement occurred because WLBS fully pipelines the network traffic between the NIC driver and the protocol stack, which increases the aggregate throughput of non-pipelined NIC drivers.

For WLBS to scale application performance, it must not introduce bottlenecks as hosts are added that would constrain throughput. The overall throughput should increase linearly as hosts are added to the cluster and join in serving the client load. To measure throughput scaling, a second host was added to the configuration described above. Performance measurements showed that a two-host cluster delivered approximately 1.96 times the throughput of a single host, which is very close to the theoretical maximum two-fold increase. Each host handled 130 simulated Web hits per second and 10.5 Mbps network bandwidth. These measurements confirm that WLBS can scale to well beyond two hosts.

Performance Projections

Because WLBS is implemented as software running on the cluster hosts, it has no specific throughput limitation. The effective throughput limit is determined by the speed of the servers and of the LAN hardware interconnecting the cluster hosts. The LAN bottleneck is principally due to the bandwidth required to send replies to the clients (10 KB each in the tests), and this bottleneck can be relieved by using a switch connected to multiple outbound pipes to the Internet.

In Web server applications, fast Ethernet hubs will support up to eight-way clusters in typical configurations. For example, in the test configuration that was used two hosts in the above measurements consumed 21 percent of the capacity of a 100 Mbps hub at 260 simulated Web hits per second (or 21 Mbps) total throughput. Four- and eight-host clusters would respectively deliver approximately 520 and 1,040 Web hits per second and would use 42 percent and 84 percent of the capacity of a 100 Mbps hub. Note that a DS3 (T3) Internet connection at 45 Mbps very nicely matches the throughput of a four-way cluster.

To support higher throughput rates, a fast Ethernet switch can be used to increase LAN capacity for outbound traffic. If the above (two-way cluster) test configuration had used a fast Ethernet switch, each host would have consumed less than 10.6 percent of its switch port's capacity to generate 130 hits per second per host. A 16-host cluster would deliver about 2,080 Web hits per second (168 Mbps) and would consume only 12 percent of the capacity of each host's switch link. This throughput rate approximately matches the capacity of a 155 Mbps OC-3 Internet link. As fast Ethernet reaches capacity, WLBS-based solutions can immediately migrate to gigabit Ethernet technology to further increase LAN bandwidth. For example, WLBS has demonstrated 250 Mbps on gigabit Ethernet in performance tests.

SUMMARY

Internet technology has been widely embraced, serving as the foundation for delivering enterprise-wide and frequently mission-critical applications such as Web, proxy, and e-mail servers. WLBS provides an ideal solution for organizations because it provides load balancing as well as high availability.

WLBS uses a distributed algorithm to statistically map workload between the nodes of the cluster, while using a heartbeat signal so that all members of a cluster can monitor the presence of other nodes. Should a node be lost or a new one brought on line, the workload is automatically and transparently redistributed among the cluster nodes.

All of this is accomplished without the expense and support burden of using special purpose hardware or software. Organizations can simply network together industry standard computers and then configure and launch the service from Windows NT Server 4.0, Enterprise Edition, or using the forthcoming Windows 2000. No special development work is needed for existing applications to take full advantage of WLBS load balancing.

WLBS is easy to manage, giving network administrators excellent controllability, including the ability to remotely manage (with password protection) the cluster from any point on the network. This ease of use is matched by a high level of flexibility, as WLBS gives network administrators the ability to tailor clusters to specific services, with control defined on a port by port level. Servers can be added or subtracted to a cluster without stopping the service. In addition, software on cluster nodes can be upgraded using the rolling upgrade capabilities of WLBS, again without stopping services to clients.

Microsoft is committed to providing the best and most complete set of clustering solutions for the enterprise. In addition to WLBS, which is ideal for providing load balancing incoming TCP/IP traffic, organizations can use MSCS to provide failover service from one computer to the other, which is ideal for running mission critical programs such as databases that require high availability. COM+, which ships as part of the forthcoming Windows 2000, provides dynamic load balancing for applications deployed across clustered servers.

As part of the Windows NT Server Enterprise Edition high availability infrastructure, WLBS provides a cost-effective solution for both scalability and load balancing for Internet-based services, giving a powerful competitive advantage to organizations that deploy it now and in the future.

FOR MORE INFORMATION

For the latest information on Windows NT Server, Enterprise Edition, check out our World Wide Web site at http://www.microsoft.com/ntserver

For additional information on WLBS, refer to the WLBS Web site at http://www.microsoft.com/ntserver/ntserverenterprise/exec/feature/wlbs.