

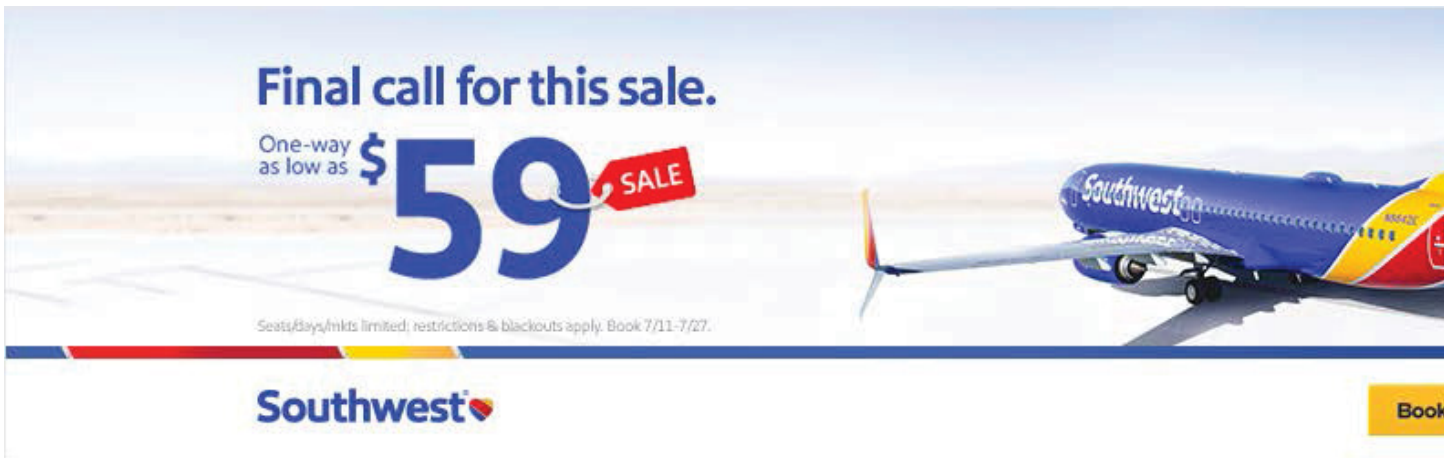
.st0{fill:#87E826;} .st1{fill:#FFFFFF;}

DATA CENTERS

## Understanding server load balancing

Getting your servers up to optimal stability and reliability is the most important issue of network administrators. Load balancing is one method of achieving a higher degree of efficiency, and Deb Shinder helps you to understand this concept.

By Deb Shinder | September 11, 2001, 12:00 AM PST



Final call for this sale.

One-way as low as **\$59** SALE

Seats/Days/mkts limited; restrictions & blackouts apply. Book 7/11-7/27.

Southwest

Book

0

Load balancing is a favorite buzzword (or in this case, buzz phrase) among IT professionals at the enterprise level and a favorite “feature” for selling new technologies. Nonetheless, many network administrators don’t really understand what it is and how it works. In this Daily Drill Down, I will provide an overview of how load balancing can increase the efficiency of your network servers and discuss

some of the options available for implementing load balancing on your network.

### What is load balancing?

The concept of load balancing is a simple one: spreading the work that a computer needs to do across multiple machines. However, the implementation of this idea can be quite complex. A number of vendors offer load balancing solutions that are implemented in different ways. You may have heard about Windows Load Balancing Service (WLBS), Network Load Balancing (NLB), Component Load Balancing (CLB), and other similar terms. In this Daily Drill Down, I will address the broad term *server load balancing*, which can encompass all of the above and more.

### Load balancing and server clustering

One way to distribute the workload is to use *server clustering*. A server cluster consists of two or more servers that operate and are managed as if they were a single entity. The servers must be able to access one another's disk data. Special software (such as MSCS, Microsoft Cluster Server) is used to manage the systems, automatically detect the failure of one system, and provide failover/recovery.

Server clusters are sometimes called *server farms*. In some implementations, the servers have individual operating systems, while in others they share an operating system. Large Web sites such as Yahoo use multiple servers in a *Web farm* to handle the huge volume of traffic.

### Hardware vs. software implementations

Operating systems such as Windows NT/2000 and Red Hat's High Availability Linux Server provide software-based load balancing, and there are also software packages such as Resonate. Many vendors also make hardware devices based on switching technology that include load balancing functionality.

Load balancing switches and routers, such as those made by Cisco, Radware, Foundary, Alteon, and other vendors, use a variety of algorithms to distribute TCP/IP requests among a group of servers. Load balancing switches are also often referred to as *content switches* and *content directors*.

Early load balancing solutions used a DNS round-robin algorithm; more recent methods include *least connections* and *fastest response* algorithms.

---

### Loadbalancing.net

For links to resources and vendors for both software and hardware load balancing solutions, as well as numerous articles on specific implementations, see [Loadbalancing.net](http://Loadbalancing.net).

## Advantages of server load balancing and clustering

Load balancing and clustering are part of High Availability (HA) strategy. Having two or more computers handle the workload increases performance (speeding up the process), and the redundancy also provides fault tolerance; if one of the machines goes down, the other can still continue to function. Clients see the group of servers as a single virtual server, with one IP address.

There are three big advantages of clustering servers to provide load balancing:

- Easier and more flexible management: With clustering software, administrators can move the workload onto particular servers within a cluster (for example, to update a server without impacting accessibility of data and services to clients).
- Uninterrupted availability and fault tolerance: If a server fails, clustering software detects the failure and fails over to a remaining server.
- Better scalability: Load balancing can be scaled across multiple servers in a cluster. Applications that are written to run on server clusters can perform dynamic load balancing.

Load balancing and server clustering technologies are important to enterprise-level networks because of the mission-critical nature of servers such as those that provide a Web presence (and often, secure transaction services and database access) on the Internet or those that provide applications and data on the corporate intranet. Load balancing ensures high availability and little or no downtime for Web, proxy, terminal, and VPN servers.

Load balancing servers in a cluster allow companies to scale their network services in conjunction with rapid growth so that additional servers can be added to the cluster as network traffic increases. Load balancing is usually implemented in conjunction with server clustering. A load balancing cluster distributes the load of incoming TCP/IP traffic, while a server cluster provides fault tolerance.

## How does server clustering work?

The servers that are members of a cluster are called *hosts* or *nodes* (depending on the vendor of the clustering technology). The cluster members are physically connected via network cables and *programmatically* connected via the clustering software.

The clustering software provides:

- A means by which the cluster members can have common access to disk data.
- A means of detecting when a server or application fails.

A means of recovering from a failure by shifting the work to the remaining server(s) or restarting the application.

- An interface through which the servers in the cluster can be managed as one entity, presenting a “single system image.”

It is also useful if the cluster administration software allows you to remotely manage the server cluster.

#### Sharing disk data access between servers

There are several different methods that can be used to allow more than one server to have access to disk data. These include:

- Shared disk method
- Mirrored disk method
- “Shared nothing” method

The shared disk method was used with the first implementations of server clustering. Software called Distributed Lock Manager (DLM) was used to give all servers in the cluster access to all physical disks. Shared disk clustering requires SCSI disks (or special cabling and switches) and applications that are modified to be aware of the disk sharing. Oracle’s Parallel Server uses shared disks.

With the mirrored disk method, each of the servers in the cluster has its own disks. The data is mirrored (an exact copy is written) to the disks on other servers. This requires special software such as that made by Veritas, NSI, and Octopus.

“Shared nothing” is a clustering method in which each server has its own disk resources. The clustering software transfers the ownership of a disk from one server to another if the server that owns the disk fails. An advantage of this method is that applications do not have to be modified. Microsoft Cluster Services (MSCS) uses this method.

#### Detecting server or application failure

When server clusters are used to provide fault tolerance, there must be a way for the cluster to detect when one of its member servers fails (or when an application on the cluster fails). One way, used by Microsoft in their clustering solutions, is with software “heartbeats”—messages that are sent on a regular basis between nodes. If a server fails, it will cease to emit the periodic heartbeat message, and the software will redistribute the workload among the remaining servers.

#### Understanding failover and failback

When a failure is detected, recovery involves *failover*, which is the transition that occurs when a server fails and another server(s) picks up its load. In many cases, this transition is transparent to the clients, as the applications, file shares, and other resources are restarted by the clustering software at the same IP address. If the client is browsing the Web or using some other “stateless” connection type, the user may not be aware of the failure at all. The application automatically reconnects after a failover. With some client applications, the user may receive a message that the server is unavailable and may be required to log back on. When the failed server comes back online, clustering software detects its presence and allows it to automatically rejoin the cluster.

After the failed server rejoins the cluster, *fallback* is the process that automatically redistributes the workload again to include the newly rejoined server.

### Balancing the load

Depending upon the load balancing implementations, administrators may be able to specify how much of the load each host should handle (the weight) or spread the load equally among all hosts in the load balancing cluster.

Load balancing can be integrated with other network services such as network address translation (NAT). The load sharing network address translation (LSNAT) technology allows for a router to intercept client requests directed to a server and select a node in the server pool to which the request will be sent, based on the load sharing algorithm.

---

### RFC 2391

Load sharing using NAT (LSNAT) is discussed in [RFC 2391](#).

---

Hardware load balancing devices tend to be expensive; you may need to purchase two devices to avoid the problem of having a single point of failure, with the second device remaining passive unless a failure occurs. Software solutions may be based on a “dispatcher” model in which all incoming requests go through one server, the “dispatcher server,” and are then distributed to other servers in the cluster. Other software solutions are fully distributed, avoiding the bottleneck that can result from the dispatcher model.

### Single system image

The single system image is the user interface that allows administrators to manage all of the cluster resources from a central location.

# Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

## Real-Time Litigation Alerts



Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

## Advanced Docket Research



With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

## Analytics At Your Fingertips



Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

## API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

## LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

## FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

## E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.