Optimizing the Performance Of Multiple Server Web Site Hosting With Load Balancing

Optimizing Internet Server Performance with Load Balancing

Many businesses operating on the Internet experience a tremendous period of growth at some point in their development. Growth of any Internet business is a matter of increasing web site traffic.

Traffic often increases exponentially... from hundreds of visitors per month to perhaps a few hundred or even thousands of page visits a day. Many web site visits involve more than just web pages. File downloads and streaming audio/video increase the data transfer volume of a typical online transaction.

Often a business that started small with shared hosting will soon justify the expense of a dedicated server. Besides being able to handle increased traffic, a dedicated server offers greater capability and can be customized to handle a company's unique business requirements. With further growth in site traffic, even a dedicated server will show signs of limitations.

Limitations of Single Dedicated Server Web Site Hosting

A single server can handle only so many transactions per second. The number of transactions a single server can handle is determined by the server load created by each transaction (memory, processor, file IO, etc.) and the capabilities of the server.

Overloading a server often results in slower download times for web pages and other file transfers. Eventually web pages fail to load, especially if the stress on the server continues to grow.

Such difficulties could result in lost business, either directly or from links to other sites posted on your page by advertisers, as is often the case for information services (news, weather, TV listings, almanacs, etc.).

No matter how powerful the server - the most memory, the fastest CPU clock speed, there will always be a limit to how many transactions per second the server can handle.

Multiple Server Hosting

The logical solution would be to have a web site hosted on multiple servers containing the same content. This way a server's ability to handle requests for service from the Internet would be multiplied.

Multi-server hosting also provides for redundancy in the event that one of the servers fails.

As the current number of servers in the group become overwhelmed by increasing traffic volume, additional servers can be seamlessly added to the group.

Failure of any one of the servers would have little, if any effect on the web site since the other servers in the group would be performing the same services.

Problems with connecting a multiple server web site to the Internet.

The question now is... how will this group of identical servers interface with the Internet? There would be the need for a routing device to distribute Internet traffic among the servers in the group. This would assure each server would handle its share of the traffic and that no one server would be overwhelmed with requests for service.

The simplest distribution method would be to send each page visit to each subsequent server in the array, eventually going back to the first server and continuing in a circular manner, an algorithm known as 'round robin'.

The problem here is that not all web transactions are the same in terms of data transfer volume. Online transactions can vary from displaying web pages to collecting customer data from web page forms to downloading files. Downloading an audio file or transferring streaming media can keep a server engaged for several minutes or even a few hours.

What's needed is a traffic routing device that monitors the amount of traffic going to and from each server, and directing any incoming traffic to the server handling the least traffic. Such a routing device is called a load balancer.

Load Balancing for Multiple Server Web Hosting

Distribution of Web Traffic to the Servers Based on Which Server is Least Engaged in Handling Requests for Service

For any given request to the web site, the load balancer directs incoming traffic to the servers processing the least amount of data transfer at the moment, be it requests for web pages or output of streaming media.

A load balancer has one or more internet ports and numerous server network connections. Although it physically resembles a network hub, it is a router that distributes internet traffic to the servers in the group according to one of several protocols.

Each load balanced URL is assigned to an IP address on the load balancer. Doing this allows for all requests for that URL to be processed by the load balancer rather than the individual servers directly. The load balancer then forwards the request on to the real server chosen by the load balancing algorithm for that URL. The real server will then process the request and either respond directly to the client or forward the response back to the load balancer for further processing depending on the configuration of the cluster.

Typical Setup

Here is a diagram of two popular typical single array setups. Fig. 1a shows the basic configuration where the server cluster is connected directly to the Load Balancer. In this case, the IP address associated with the domain name by the DNS would be assigned to the load balancer.

Fig. 1b shows two load balancers are connected to the group of servers. Both load balancers are each connected to the server group. Though each load balancer has its own IP address, there is a third IP address shared by the two load balancers known as the virtual IP. The virtual IP can move from one load balancer to the other in the event of a failure of the primary load balancer in order to allow for continued service.

Basic Load Balanced Server Groups

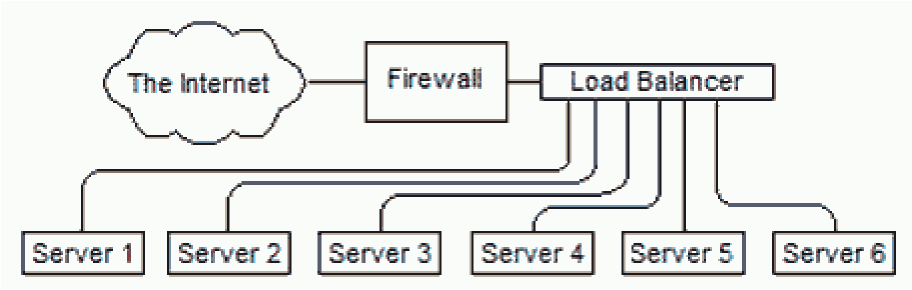

Fig. 1a, Load balancer connected to an array of six servers.

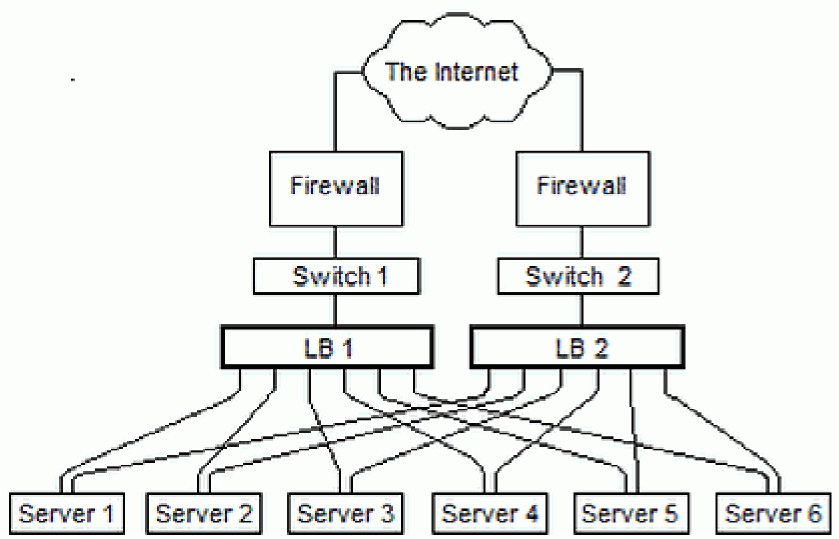

Fig. 1b, two or more load balancers with crossover connections to the group of servers.

Redundancy

Looking at the diagrams above, it is clear to see how load-balanced server arrays allow for redundancy and scalability. In both examples above, should any server fail, traffic is routed to the other servers in the group. Should any server require routine maintenance, it can be removed from the group, usually without any noticeable loss of function.

The arrangement shown in Fig. 1b extends system redundancy to the load balancing hardware. As previously mentioned, the two load balancers share an IP address known as a virtual IP which is associated with the URL being load balanced.

At any moment, only one of the two load balancers is distributing web traffic to the server group. Should one load balancer fail, the other would automatically take over the virtual IP and pass requests to the group of real servers.

Geographic Load Balancing

Redundancy and guaranteed 100% uptime are taken a step further by having multiple hosting server groups in different data center locations. Fig. 1c shows two identical multi-server hosting systems at two different geographic locations.

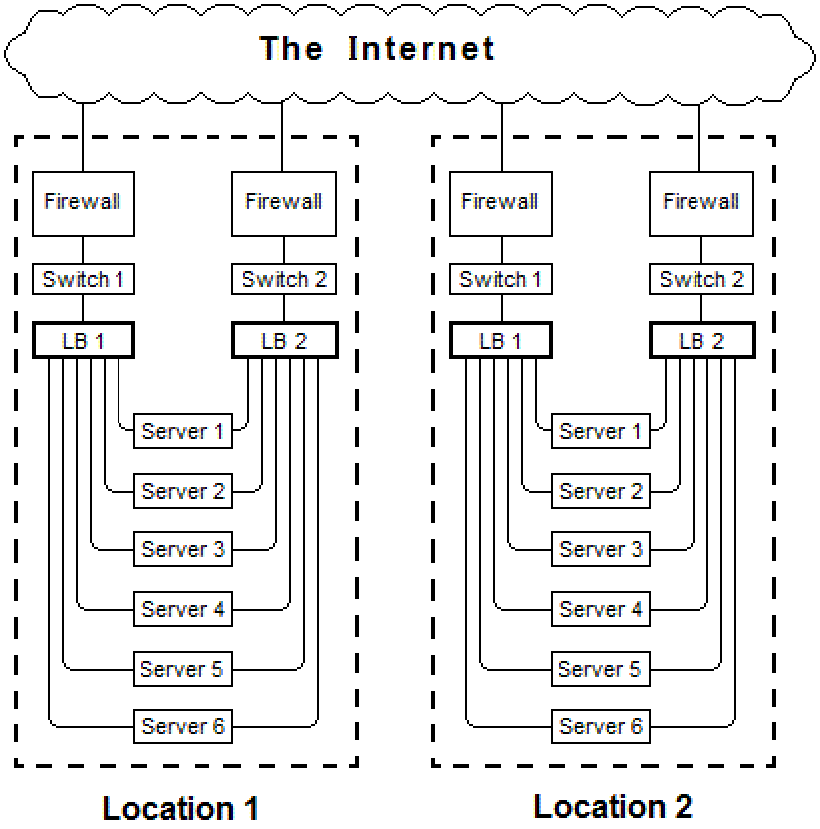

Fig. 1c, Geographic Load Balancing

Ideally the two identical multi-server groups would be in two cities sufficiently apart to where any destructive force, be it weather or earthquake, would affect only one hosting server system, and the web site would remain functional by means of the remote server group.

Geographic Load Balancing, as the name implies also functions to distribute traffic according to geographic origin of the traffic. Traffic from any point on the globe would likely be directed to the nearest data center location.

The two systems, connected only through the Internet, would communicate with each other as to operability and traffic volume.

Scalability

Multiple server hosting also allows for scalability. As web site traffic increases to where service would be inadequate, additional servers can be added to the group. An additional server can be installed without shutting down the system. The load balancer will immediately recognize the new server and start sending traffic to it.

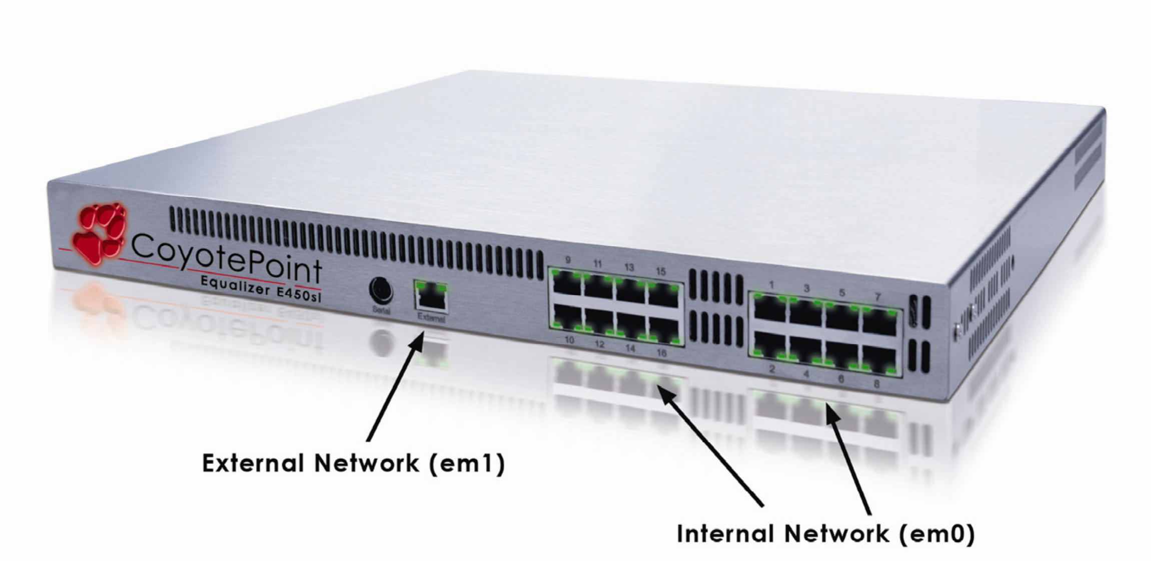

Fig. 2, Coyote Point Equalizer E450sl

This load balancer has sixteen Ethernet (T45) internal network connectors servers that can each be uplinked to a switch to support hundreds of real servers. Scalability is simply a matter of plugging in another server and configuring the added server to the load balancer. This task consists of registering the server's IP address and copying the contents from one of the other servers in the group onto the hard drive of the newly added server.

Distribution Policies
How traffic is distributed to the servers in the group.

There are four common types of distribution policies for load balancing a group of servers. Which to choose depends on server capabilities and what types of online services are performed by the web site.

Round Robin – Most common when all servers are identical in their capacity to handle basic HTML traffic. Each visitor is sent to the first, second, third,… last then back to the first.

Weighted Round Robin – Used when servers have different load handling capabilities. Servers with greater capacity will receive more web traffic.

Least Connections - New visitors are sent to the server with the least number of users connected.   This provides the most efficient traffic handling service.

Fastest Response - New visitors are sent to the server that responds most quickly.

Least connections and Fastest Response allow for web sites that also download files and deliver streaming media (audio and video) since these services tend to keep a server engaged in a session much longer than simple html document delivery.

Systems for Database-Driven Web Sites

So far, we have looked at internet servers performing functions involving a connection with Internet users like displaying web pages, and gathering user information and downloading files and media. Many web sites have more complex internal systems. This is particularly true with database-driven web sites.

For web operations involving large and multiple databases, it is common practice to have a separate server for databases. These too are often set up as multiple identical servers to provide for redundancy. Below is a typical two-server database-driven web site.

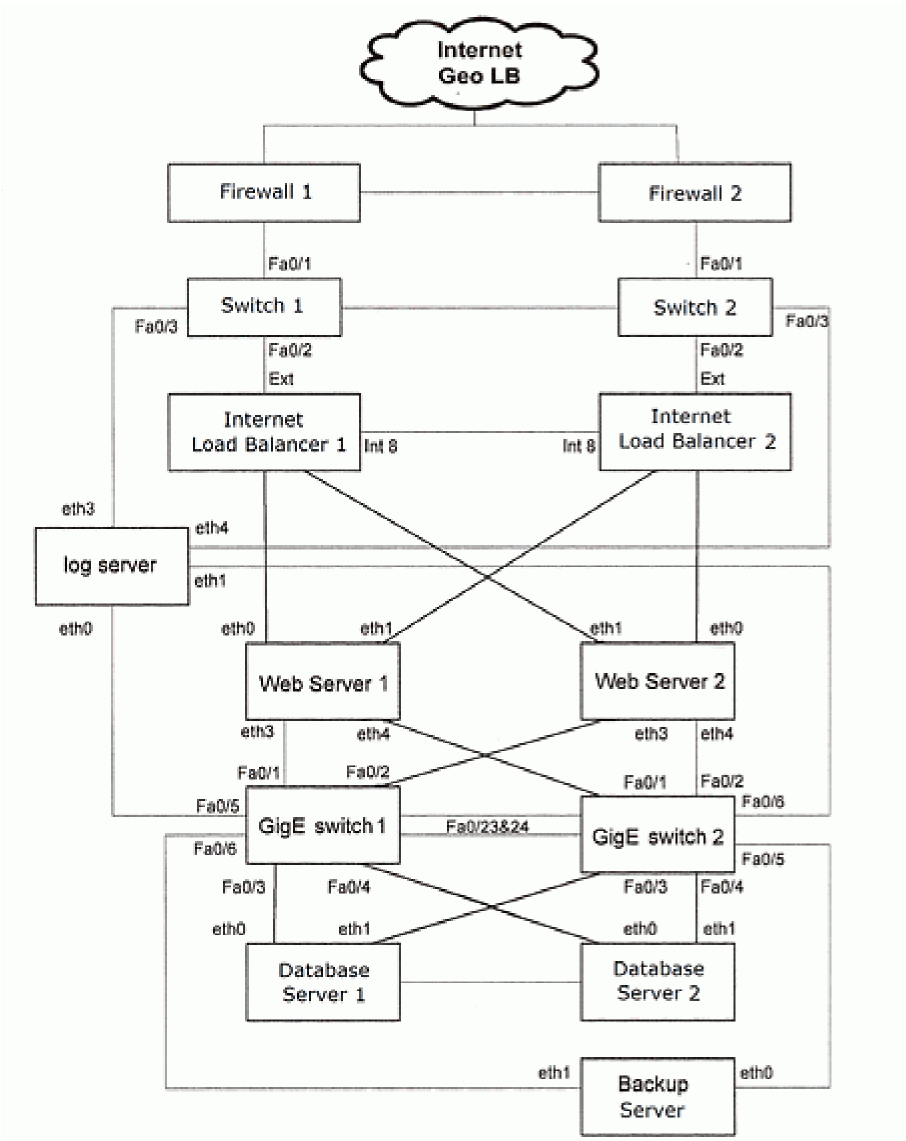

Fig. 3, Dedicated Load Balanced Web Site Server Group with Database Servers

Separate database servers are used because of large storage requirements. Also the process of querying databases is better done by dedicated processors unencumbered by the mundane tasks of serving web pages.

Among companies that use databases are employment, real estate personals and other membership sites where member objects are added, modified and deleted on a regular basis. Also many online retailers catalog and auction sites require several databases. For example a catalog site (nutritional supplements) would require a product database, an ezine subscriber database and a customer database.

Customer databases usually contain contact information and records of purchases made by each customer.

Setting Up a Load Balanced, Multi Server Web Site

For most growing Internet businesses, a hosting service providing multiple server systems with load balancing is the ideal way to provide the best customer experience with 100% uptime.

The major players in the Internet business world like Amazon and EBay, as well as major banks and insurance companies have their own server setups on their premises. Major banks have multiple data centers allowing for redundancy and optimal distribution of traffic.

For smaller businesses, however, using a hosting service eliminates the need for a employing a complete 'Server Department' staff as well as having to purchase and maintain expensive equipment and have a secure portion of a building to house the facility.

Superb Hosting offers complete turnkey hosting solutions for large, high traffic Internet enterprises

Superb Hosting employs knowledgeable staff and systems are monitored 24/7. Client companies need only employ a webmaster and web designer. These employees could also be the system administrator managing the company's internal computer network (LAN) if applicable.

Our hosting servers usually have access via multiple Gigabits per second (Gb/s) connectivity to the Internet Backbones allowing for the fastest transfer of massive amounts of data over the Internet.

Getting Started

Getting down to the actual business of setting up your multi-server system, once your order is placed, the staff at Superb Hosting will…

- physically install the web and database servers, switches, load balancers, and firewalls.
- connect all the components together with the proper cables.
- Configure the components of the system and setting the IP addresses of the components.
- change the web site IP address to the virtual cluster.
- provide any necessary assistance in installing server software specified by the client.
- Add and configure additional servers to the group to handle increasing online traffic to your site as your business grows.

All the client would have to do is prepare and upload web document files and server scripts. All web servers are configured to propagate all additions, deletions and changes of web server content throughout the entire array of Internet servers.

Since every server has its own IP address, each can be accessed directly for purposes of uploading files and viewing web documents (e.g. ftp://192.168.1.0 and http://192.168.1.0 respectively), and for installing server software.

This guide was prepared by Superb Internet.
Providers of Superb Hosting High Availability Server Load Balancing Service
http://www.superbhosting.net/hosting-solutions/server-load-balancing.php