LOADBALANCER

WHITE PAPER

An introduction to load balancing

With no technical jargon, this whitepaper describes load balancers in terms everyone can understand and explains how they improve the performance of web sites, mobile apps and key business systems.

> *"Challenges arise when large numbers of users all try to access the same applications over the Internet at the same time."*

Uncompromising expectations

We have all developed uncompromising expectations. Whether we are shopping in online sales, looking for directions on our mobile phones, accessing corporate file storage systems at work or looking up customer/ patient records, we expect to be able to obtain the information we need instantly. We expect every website to be up and running all of the time and demand access to our business systems from any device, in any location.

Load balancers make this possible. Hidden behind the scenes, in data centers all around the world, there are millions of load balancers that keep all kinds of business, mobile, web and cloud-based applications operational 24/7. They help to ensure that all users get the same high quality experience, even throughout exceptional peaks in usage, and deliver the fast, responsive application performance we now take for granted.

How do load balancers work?

To fully understand how load balancers work, it is important to first understand how applications themselves take instructions from users and respond with the required information or service.

At the simplest level, applications are software programs that have been designed to help people perform tasks. Today's applications enable us to do everything from creating documents to searching for a new home and accessing our patient records online. While most applications used to reside on individual PCs and laptops, now most applications are accessed over the Internet using a web browser (web apps), from mobile devices (mobile apps) or via corporate networks.

These web apps, mobile apps and shared business systems, like email and print management, all rely on centralized computers called servers, which deliver the functionality for the connected devices over the Internet. When someone (the 'user') opens a mobile app, accesses a web application via a web browser or logs into a corporate application, he or she triggers a request via the Internet to the central web servers or application servers. The servers perform the required task – such as retrieving information or processing data – and send the information back to the user's mobile, desktop computer or tablet.

Challenges arise when large numbers of users all try to access the same applications over the Internet at the same time. The central server can become overwhelmed with simultaneous requests and slow down or even shut down ('crash'), resulting in poor performance and a bad experience for the user. To address this, most organizations will run their applications on multiple servers, known as a cluster of servers, giving them more computing power and the capacity to meet the needs of large numbers of users.

What do load balancers do?

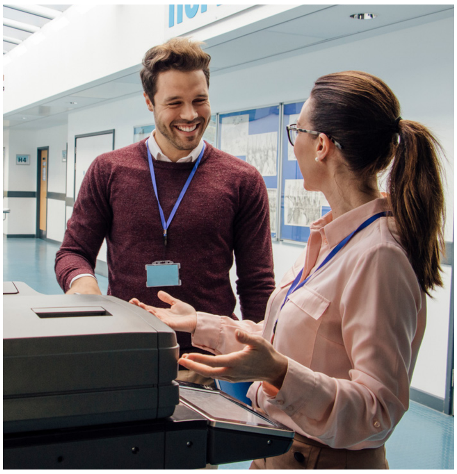
Load balancers sit in between the Internet and clusters of servers and distribute all the requests from users across the available servers. They do this in a number of different ways, the most basic of which is 'Round Robin', whereby each new user request is allocated to the next server in the sequence. There is also the 'Least Connection' approach, in which user requests are directed to the server in the cluster that currently has the fewest connections.

Alternatively, in the 'Lowest Response Time' method, the load balancer automatically directs traffic to the server that will respond to the user request in the fastest time.

> *Offering more sophisticated functionality than the embedded vendor solutions, third party load balancers distribute traffic across storage nodes in the scale-out NAS environment more logically.*

A new generation of intelligent load balancing solutions is now available that allows organizations to make more nuanced decisions about how to best balance user requests (or 'traffic') across their servers to improve application performance. Such solutions include functionality for routing user requests not just based on the number of connections per server, but on the type of requests and the current load on the server.

Organizations can also use load balancers to identify the individual device (the user's IP address) and route each user to a specific server or to keep sending users to the same server (known as 'persistence'). This function is particularly useful in the case of ecommerce websites, as it can be used to ensure that users don't repeatedly lose their shopping carts every time they are directed to a different server.



Example 1

Load balancers with print management and workflow systems

Under the hood of every organization is an intricate web of document and information-sharing, both physical and digital. This is orchestrated by print management and document workflow applications. Whether submitting expense forms for approval, printing reports for distribution, managing customer engagement journeys, or securing classified documentation – for organizations with large numbers of employees, there's no better way to effectively manage document sharing, administer the security of sensitive information, or keep printing costs down. Performing such an essential role means it's critical that these applications are working, and working optimally, at all times – and enterprise vendors are now recognising the fundamental benefits of load balancers to make these business-critical environments resilient and high-performing.

Load balancers themselves come in different forms. They can be hardware devices that physically sit in datacenters or at organizations' premises; equally they can be software-only solutions that are installed on physical or virtual servers in datacenters, on premise or in the cloud. Some vendors combine load balancing functionality with additional capabilities, such as security features.

These multi-faceted solutions (known as Application Delivery Controllers) offer a wide range of sophisticated, customizable capabilities that go far beyond just load balancing, but can be complex to manage and costly. For those organizations that simply need load balancing capabilities, standard hardware and software-based load balancers are generally inexpensive and easy to install and manage.

> *"For those organizations that simply need load balancing capabilities, standard hardware and software based load balancers are generally inexpensive and easy to install and manage."*

What are the benefits of load balancing?

Using load balancers enables organizations to:

Ensure apps are always available

Load balancers help organizations to keep their critical applications up and running 99.999% of the time. If one server in a cluster breaks down, the load balancers will automatically and instantly direct user requests to an alternative server instead. This significantly reduces the risk of downtime, which, at best, is deeply annoying for users and, at worst, could have a significant impact on an organization's revenues and reputation. If a medical facility were to experience downtime in a vital diagnostic system, the loss of the application could impede doctors' ability to save lives.

Deliver a consistently good user experience

By directing user requests to the servers that are best able to respond quickly, load balancers can provide consistent and fast app performance for thousands of concurrent users. Indeed, some web apps today handle thousands of user requests a second and rely on load balancers to return information, images or video instantly. Being able to deliver fast response times, enables organizations to ensure that their employees can work productively and their customers can have an enjoyable online experience.

Cope with surges in application usage

Some peaks in demand can be anticipated, such as increased ecommerce traffic on Black Monday. However, organizations also need to be prepared for unexpected surges in demand and plan for future growth. When they have load balancers installed to manage their application traffic, organizations can dynamically add more servers to respond to growing user numbers – and they can do this quickly and easily, without having to take down their apps temporarily or reconfigure the other servers in the cluster.

Maintain servers and apps cost effectively

Load balancers give IT managers the flexibility to manage and maintain their web servers and application servers more easily. They can disconnect, patch, upgrade and reconnect one server, without impacting the operation of the remaining servers and, importantly, without causing any disruption to app performance for users. Server maintenance activities can also be more cost effective, as IT teams do not have to be paid overtime to undertake routine maintenance outside of usual business hours or at weekends.

Example 2

Load balancers with storage solutions

Large organizations commonly have centralized IT systems for storing all documents and data created and used by all employees. These storage solutions are intelligently designed to recover or rebuild following an unexpected failure, but they don’t account for the needs of the users. Load balancers are therefore required to help ensure that all users have consistent, fast access to the documents and information they need to do their jobs. No matter where in the world the employee or the storage system is located, the load balancers will direct user requests to the most appropriate server to ensure that the employee can work productively. Even during peak periods and with heavy loads, the load balancers will ensure that users’ performance expectations are met.

What is the difference between load balancing and link balancing?

> “Load balancers deliver two vitally important benefits: zero downtime and optimum application performance.”

Link balancers (sometimes called link load balancers) are often confused with load balancers but, in fact, fulfill an entirely different role. Link balancers sit between a corporate network (the local area network or LAN) and larger geographic networks or telecommunications networks (knows as the wide area network or WAN). Link balancers then distribute requests from PCs, laptops and other devices connected to the LAN, across multiple Internet links to the best available destination.

In summary then, link balancers control how user requests reach the internet, while load balancers control how user requests are handled once they have travelled through the internet and have reached the destination applications.

Example 3

Load balancers with Healthcare solutions

In the medical industry, it is absolutely critical for key IT systems, such as healthcare imaging solutions, to be available and performing optimally all of the time. Doctors and clinicians need to have instant access to medical images, such as x-rays and CT scans, 24 hours a day, every day, to enable them to diagnose conditions, prescribe treatments and save lives. Load balancers are used to balance traffic to healthcare imaging systems and seamlessly divert user requests to alternate servers, in the event that one server fails, preventing downtime. Healthcare imaging systems are replaced infrequently, so load balancers also play a key role in helping to ensure that these vital applications can carry on delivering good performance for users over time, as the volume of archived images and usage grows.

About Loadbalancer.org

Loadbalancer.org's mission is to ensure that its clients' businesses are never interrupted. The load balancer experts ask the right questions to get to the heart of what matters, bringing a depth of understanding to each deployment. Experience enables Loadbalancer.org engineers to design less complex, unbreakable solutions - and to provide exceptional personalized support.

LOADBALANCER

Visit us: www.loadbalancer.org

Phone us: +44 (0)330 380 1064

Phone us: +1 833 274 2566

Email us: info@loadbalancer.org

Follow us: @loadbalancer.org