loadbalancer.org

GUIDE

Load Balancer Selection Guide

Helping you understand ADC performance terminology

Purpose of this guide

Finding the right fit for your needs

Load balancers are a core component of any IT infrastructure, providing a host of benefits ranging from **resilience** and **redundancy** to increased **security**. Selecting the right load balancer appliance for your needs is however essential.

This guide has been created to highlight some key considerations that may assist you in your decision making, including an explanation of the relevant technologies and terms involved, enabling you to confidently manage the demand and experience of your internal or external customers, without over-provisioning.

Should you wish to discuss your specific needs in more detail, our sales team are here to guide you, if/when you're ready.

Terminology

Demystifying the tech

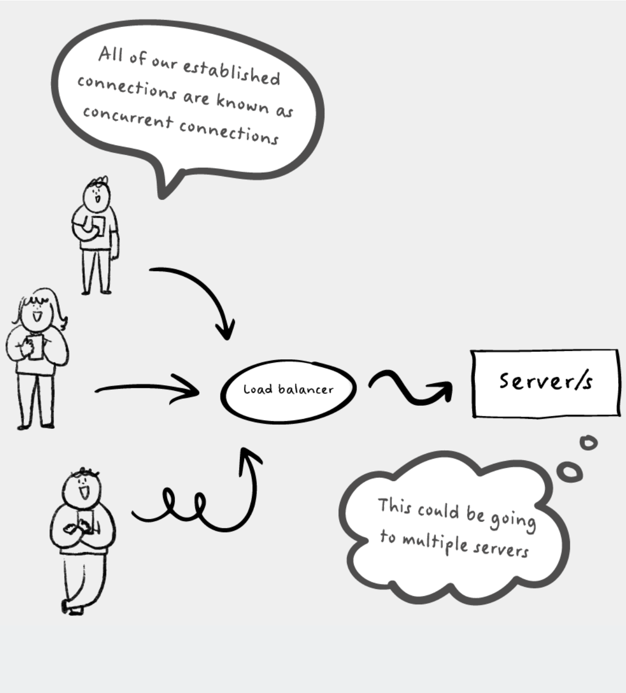
Sites, servers, services, data and users grow at an ever increasing rate, and as such the systems that support them need to be managed effectively and efficiently. Let's start by demystifying some of the terms frequently used when considering load balancer capacity and performance, along with some of the technologies they relate to.

Connections

A connection is, typically, an established communication channel between a client (a user using a web browser for instance) and a server or site, although in reality this is almost always a load balancer. Where web based traffic is concerned, a client browser will typically open multiple, parallel connections to the same site (Chrome will open up to six for instance) in order to improve performance - one user rarely equals one connection.

Concurrent Connections

This is the number of established connections between one or more users and a single server or load balancer. As we know the most common browser could open up to six connections to a site, we can infer a few things if we know the number of concurrent connections a VIP has.


All of our established connections are known as concurrent connections
Load balancer
Server/s
This could be going to multiple servers

EXAMPLE

If we take 1000 concurrent connections as an example...

The **minimum possible number of users** is roughly 166 (166 users using Chrome which opens six connections to the site: 166 x 6 = 996). Note however that one user doesn't always equal one connection - a single user may in fact result in several connections.

The **maximum possible number of users** is 1000 (1000 users using Chrome which opens one connection to the site - although this is very unlikely).

Equally, if we know we have 1000 active users we can calculate the **minimum and maximum possible number of concurrent** connections our VIP needs to support:

- If each user's browser only opens 1 connection, our VIP will see 1000 concurrent connections.
- If each user's browser opens all six possible connections to our site, the VIP will see 6000 concurrent connections.

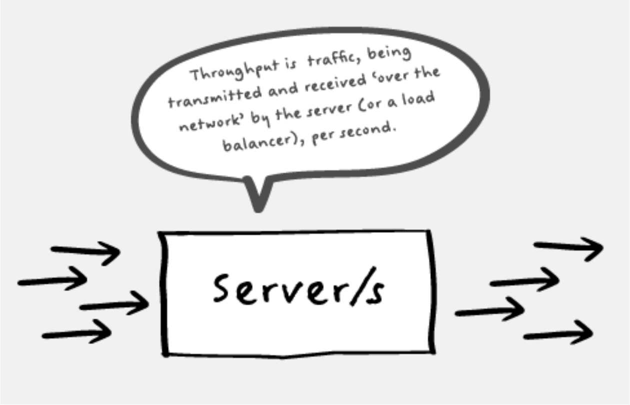
These are theoretical extremes because in most cases the number of connections a user's browser will have open will vary (and vary over time) based on factors such as:

- How active a user is
- The composition of the page(s) they are visiting
- Whether they are filling in a form, and so on.

Regardless, this is a great way to understand the absolute maximum possible number of concurrent connections your load balancer needs to support, particularly if you have a large number of users logging in at a particular time, or need to consider seasonal or promotional peaks.

Throughput

The amount of data, usually referred to as traffic, being (or capable of being) transmitted and received 'over the network' by a device such as a load balancer or server, per second. You may also be able to monitor the throughput of a VIP but this will always be limited by the capabilities of the device it exists on. Typically the maximum throughput of a load balancer is determined by the maximum throughput (or speed) of all of its network interfaces combined. Our Enterprise 1G hardware appliance has four interfaces with a speed of 1Gbps so its maximum possible throughput is 4Gb.


Throughput is traffic, being transmitted and received 'over the network' by the server (or a load balancer), per second.
Server/s

N.B. Protocol overheads mean you're only likely to actually achieve around 80% of the maximum theoretical throughput of an interface. It's also worth keeping in mind that cloud based host virtual network interface throughput may be limited by factors other than the reported speed of an interface. So while the operating system may be presented with a virtual 1Gb network interface, its actual maximum possible throughput may be far lower.

HTTP

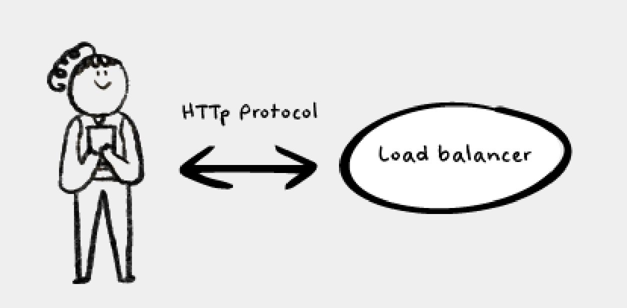
Hypertext Transfer Protocol is a surprisingly simple, text based request/response protocol which forms the basis for most web-based traffic. The web and most mobile apps use HTTP. As clients typically make very small requests for content (such as web pages and images) and responses containing that content are much larger, response throughput (towards the clients) is typically much higher than request (receive) throughput.

HTTP Persistent Connections (aka HTTP keep-alive)

In earlier versions of HTTP, when a user opened a web page, each request to the site opened a new connection and once the response was received, the connection was closed. This is highly inefficient as the connection setup and tear down process introduces delay and consumes CPU resources, especially if SSL/TLS (see below) is in use. To overcome this, persistent connections were introduced. These keep a connection open allowing it to be used for multiple requests and responses. If a persistent connection is no longer used, it will be closed after a period of idle time, usually specified by the server or load balancer.

HTTP RPS

The number of incoming HTTP Requests Per Second that can be processed by a load balancer or server. As small as they usually are in comparison to responses, requests contain a lot of information that must be parsed to ensure they are valid, do not present a security threat and to make load balancing decisions based upon any number of components within the request. For this reason, the maximum number of RPS is significantly lower than the maximum number of connections a load balancer supports.



SSL/TLS

SSL is an acronym for Secure Sockets Layer which is a protocol that provides security for connections to websites (among other things). It does this in two ways: 1) by verifying the identity of a website using keys and certificates and 2) by encrypting the data sent to and from that website. SSL was actually replaced by TLS, Transport Layer Security in 1999 but it's still more commonly referred to as SSL than TLS (or as SSL/TLS).

SSL/TLS TPS

SSL/TLS Transactions Per Second. Establishing SSL/TLS security for a new connection is a CPU intensive operation and the number of new connections that can be handled at any one time is limited by the power of a device's CPU(s). The key size used has a significant impact (using contemporary ECC keys, which are much smaller, will result in the highest TPS). Once a connection is secured and established things become far less taxing. A device can support hundreds of thousands or even millions of active SSL/TLS secured connections, but its capacity to establish new ones is limited by this figure.

Session Persistence (aka Affinity or Stickiness)

Session persistence is a load balancing feature that ensures that connections from the same client or user are sent to the same real server (aka RIP or backend) so that session data relating to that user which is only present on that server (authentication information, products in a basket, form input etc.) is maintained and available. This is particularly important where a browser opens multiple parallel connections to the site. Without session persistence it would be impossible for sites to offer features such as login, shopping baskets, payments pages, multi-page forms or anything else that relied upon the site 'remembering' who the user is and what they have done at any point in time. Information used by a load balancer to support persistence is usually stored in simple records held in memory.

Weighing it all up

What you might want to consider

Now you've got a better understanding of some of the relevant terminology and technologies involved, let's explore the considerations that will influence the appliance specification. As you can see, there are a lot of things to take into account. There's no easy formula you can use to reach a decision but the better you understand the dimensions that can influence it, the more judgment you can make.

1 How many applications do you have?

The more applications a load balancer needs to support the greater the resources required. This is particularly relevant if your users are likely to be using more than one of these applications as part of their work. The use of persistent connections will likely mean a higher connection count needs to be supported and if TLS is used, a higher number of TLS TPS.

2 How many concurrent users does each application support?

The higher the number, the more significant the answers to the those earlier questions become, particularly if TLS is being used.

3 Could you use Direct Server Return (DSR)?

DSR is a method of load balancing where inbound requests are load balanced but return traffic is sent directly to the client bypassing the load balancer. This greatly reduces the throughput demands on the load balancer and allows for a far greater number of clients to be supported. Consider using DSR where an application has high throughput demands and you do not need Layer 7 processing (such as TLS termination or cookie persistence) or WAF features.

4 What types of application?

Complex web sites and storage services are very likely to require greater resources. A complex website will likely require more connections, session persistence records, HTTPS RPS, TLS TPS and throughput than a simple one. A storage service simply has higher throughput.

5 What types of VIP?

Layfer 4 VIPs consume far few resources than Layer 7 ones and HTTPS RPS, TLS TPS and other factors can be ignored. Connections and throughput cannot.

6 What is the throughput required?

How much traffic is going to be passing through the load balancer? This may require some analysis and monitoring to determine. If you are migrating from an existing load balancer it probably won't be too hard to find out what the peak overall throughput it currently supports is. If you are launching a new service you'll likely need to do some testing based on the anticipated peak number of concurrent users.

7 Are you migrating from hardware to a virtual appliance?

As a rough guide, the overheads of virtualisation usually result in an approximate performance loss or overhead of around 10%. Aside from that the specification for network interfaces, CPU and memory can be matched between the two. It's worth noting that our unique Freedom License allows you to migrate across different platforms easily, quickly and without any financial penalties - and also back again if you need.

8 What are your peaks?

Don't rely on averages for any of the throughput or connection metrics you might rely upon as you need to support the inevitable peaks that occur over time, be it daily, weekly or something else.

- Do a large number of users login in the morning or after lunch,
- Are users likely to make greater use of the application as a deadline approaches? Perhaps completing a weekly time-sheet or tasks related to month end or even a seasonal holiday?
- Could a marketing campaign, a new feature or a change in company policy result in a spike of users?

9 Are you using the Web Application Firewall?

This feature adds a significant performance load that you should accommodate for when considering CPU and memory requirements. How much will depend on your specific configuration.

10 What growth do you expect?

Always factor in the likelihood that the number of users of an application will grow and that their active use of an application may also increase. How you anticipate this growth will depend on many factors such as customer growth, company and staff growth, sales and marketing activity and more.

Reducing load and increasing performance

Helping you support more users

There are a number of things you can do to reduce the load upon your load balancer, increase its performance, and likely improve the speed of your applications from a user's point of view. A number of these suggestions can help a device support a much greater number of users and applications which may be a factor in your selection.

- Enabling TLS Renegotiation can reduce TLS related CPU load. Due to DDoS vulnerabilities introduced by this feature, this isn't recommended for public websites.
- Backend connection reuse can reduce the number of connections between your load balancer and real servers and will likely slightly increase response times.
- A Layer 4 VIP will have a much lower load on your device than a Layer 7 one, although many advanced features will be unavailable.
- Using DSR typically allows for a far higher number of clients to be supported by a device and avoids the need for response traffic to be processed by the load balancer.
- Using smaller TLS keys greatly increases the maximum possible SSL/TLS TPS. Elliptic Curve Cryptography (ECC) keys, whilst very small, provide security equivalent to much larger RSA keys.
- WAF Inspection features have a high overhead - use them sparingly.
- HTTP Compression can be used to reduce the size of text-based response data resulting in lower throughput, allowing for a greater number of clients to be supported at the cost of an increase in memory use.
- Content caching can be used to reduce the load on your real servers by serving frequently requested content on their behalf.

Outside of the load balancer itself, it's helpful to have:

- Good network connectivity and capacity to both your clients (internet based or otherwise) and to your real servers.
- Performant real servers and other systems your site or servers rely upon such as databases, message buses and queues and authentication systems.

Next steps

Here if you want to discuss specifics

As always, our sales engineers are more than happy to answer any questions you may have, and to provide advice on your specific requirements, should you wish. Contact them at: sales@loadbalancer.org.

Alternatively, feel free to check out our detailed resources for deployment guides and our admin manual, for more information.

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 personalised support.

Email us: info@loadbalancer.org

Visit us: www.loadbalancer.org

Call us: +44(0)330 380 1064

Call us: +1 833 273 2566

Follow us: @loadbalancer.org