Introduction to Load Balancing

In this lesson, you will understand load balancing and the need for it in web applications

We'll cover the following

- What is load balancing?
- Performing health checks of the servers with load balancers

What is load balancing?#

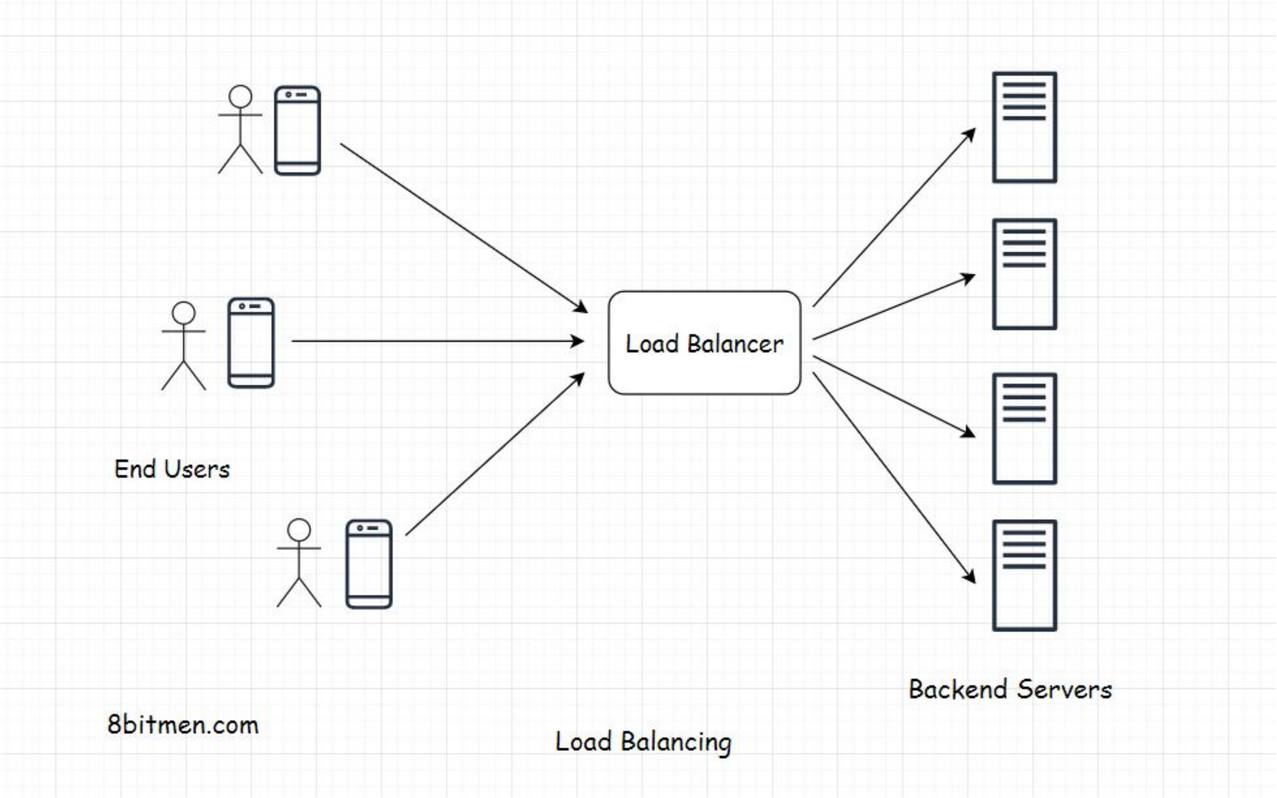
Load balancing is vital in enabling our service to scale well with an increase in traffic load, as well as stay highly available. Load balancing is facilitated by load balancers, which makes them a key component in the web application architecture.

Load balancers distribute heavy traffic load across the servers running in the cluster based on several different algorithms. This averts the risks of all the traffic converging on the service to a single or a few machines in the cluster.

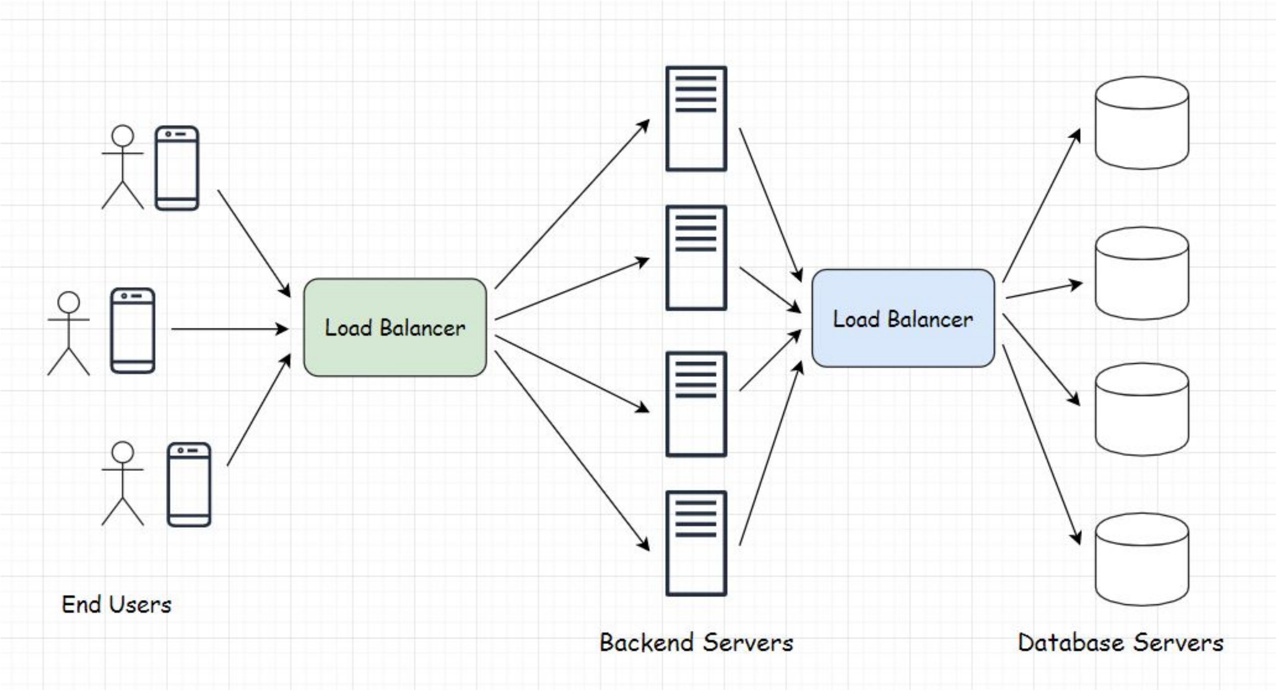
If the entire traffic on the service converges only to a few machines, this will not only overload them resulting in the increase in the latency of the application and killing its performance, but it will also eventually bring them down.

Load balancing helps us avoid all this mess. While processing a user request, if a server goes down, the load balancer automatically routes the future requests to other up and running servers in the cluster. This enables the service as a whole to stay available.

Load balancers act as a single point of contact for all the client requests.



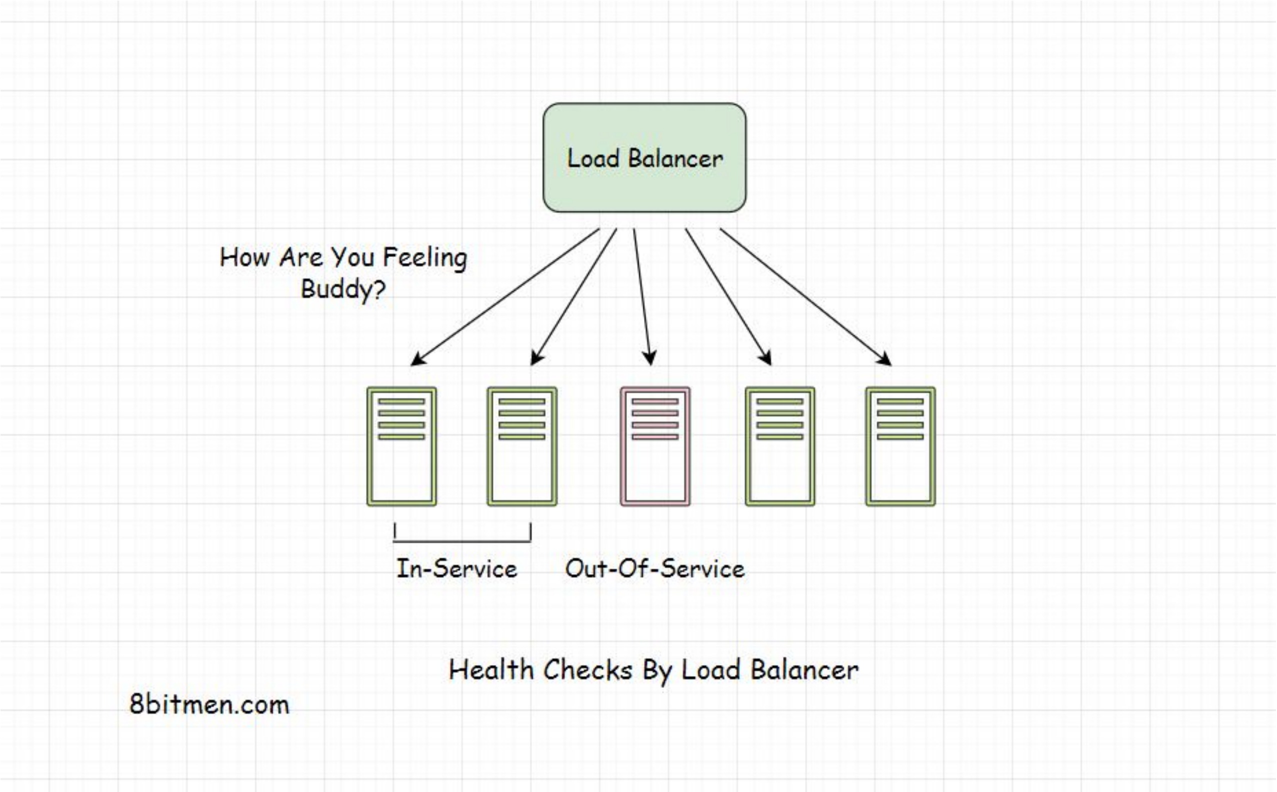
Load balancers can also be set up to efficiently manage traffic directed towards any component of the application, be it the _backend application server, database component, message queue, or any other component. This is done to uniformly spread the request load across the machines in the clusters powering that respective component.



Performing health checks of the servers with load balancers#

In order to intelligently route all the user requests to the running servers in the cluster, a load balancer should be well aware of its running status.

To ensure that the user request is always routed to the machine that is up and running, load balancers regularly perform health checks on the machines in the cluster.



Ideally, a load balancer maintains a list of machines that are up and running in the cluster in real-time, and the user requests are forwarded to only those machines that are in service. If a machine goes down it is removed from the list.

Machines that are up and running in the cluster are known as *in-service* machines, and the servers that are down are known as *out of service* instances.

> Just for the record, *Node, Server, Server Node, Instance, and Machine* all mean the same thing and can be used interchangeably.

After the *out of service* instance comes back online and becomes *in-service*, the load balancer updates its list and starts routing the future requests to that particular instance all over again.

Alright!! In the next few lessons, you will discover how load balancers work. To understand that well, you need to first understand the *Domain Name System (DNS)*.

We will discuss the DNS in the next lesson.

← Back

High Availability Quiz

Next →

Understanding DNS – Part 1

☑ Mark as Completed

Report an Issue