Introduction To Load Balancing

In this lesson, we will understand load balancing & 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 the increase in the traffic load, as well as stay highly available. Load balancing is facilitated by load balancers, that 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 convergence of all the traffic on the service to a single or a few machines in the cluster.

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

Load balancing helps us avoid all this mess. Amidst 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 thus enabling the service as a whole to stay available.

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

Load Balancing

Backend Servers

End Users

8bitmen.com
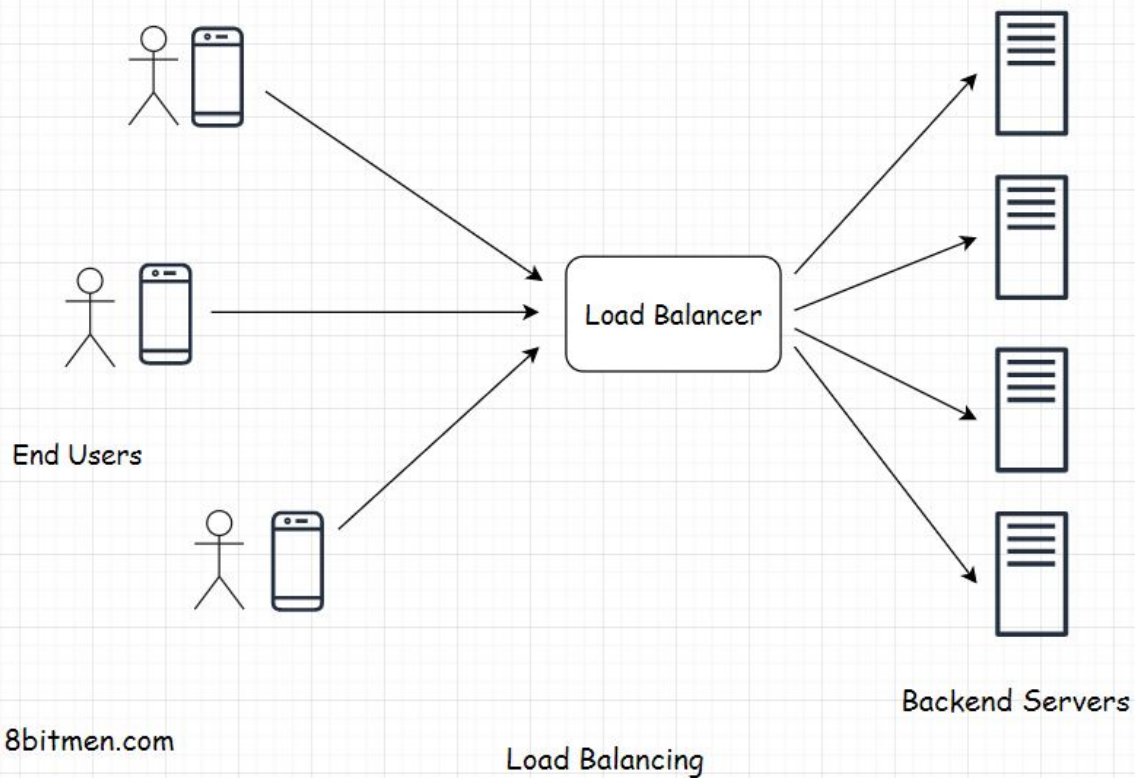
Load balancers can also be setup to efficiently manage traffic directed towards any component of the application be it the *backend application server, database component, message queue* or any other. This is done to uniformly spread the request load across the machines in the clusters powering that respective component.

End Users

Backend Servers

Database Servers

Load Balancing At Different Components Of The Application

8bitmen.com
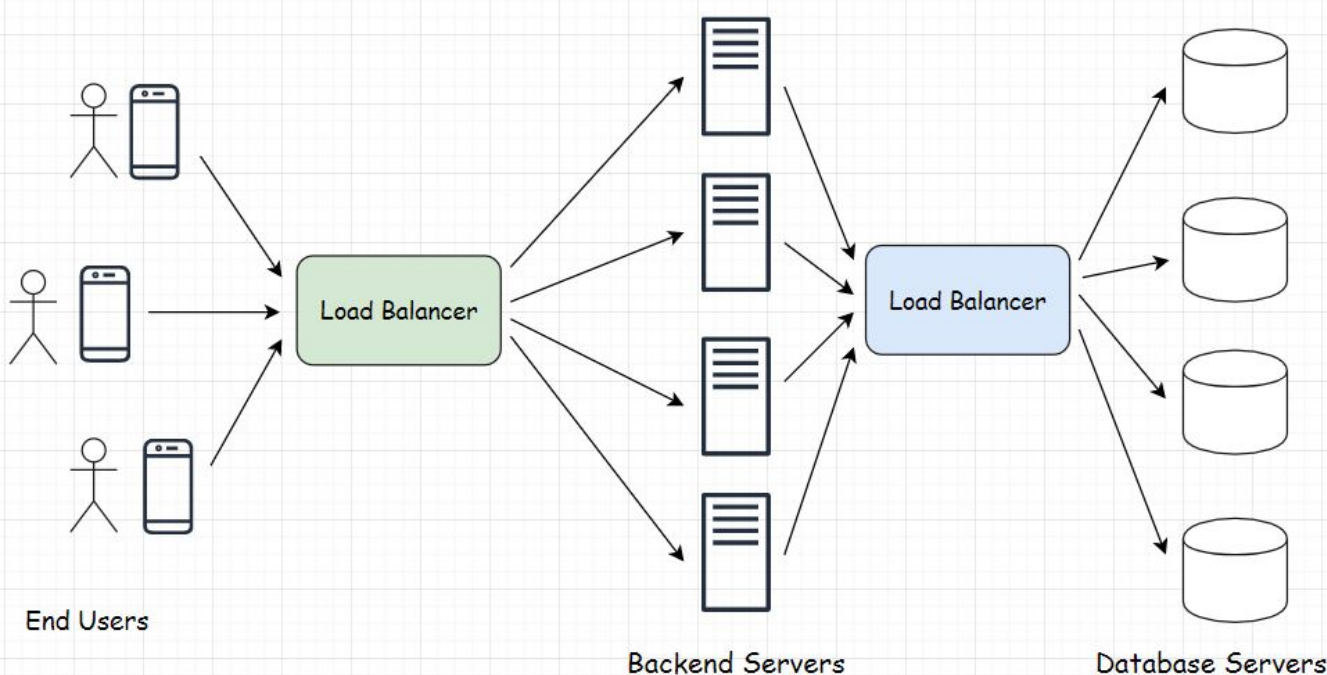
Performing Health Checks Of The Servers With Load Balancers

In order to intelligently route all the user requests to the running servers in the cluster, a load balancer should be well aware of their 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 of the machines in the cluster.
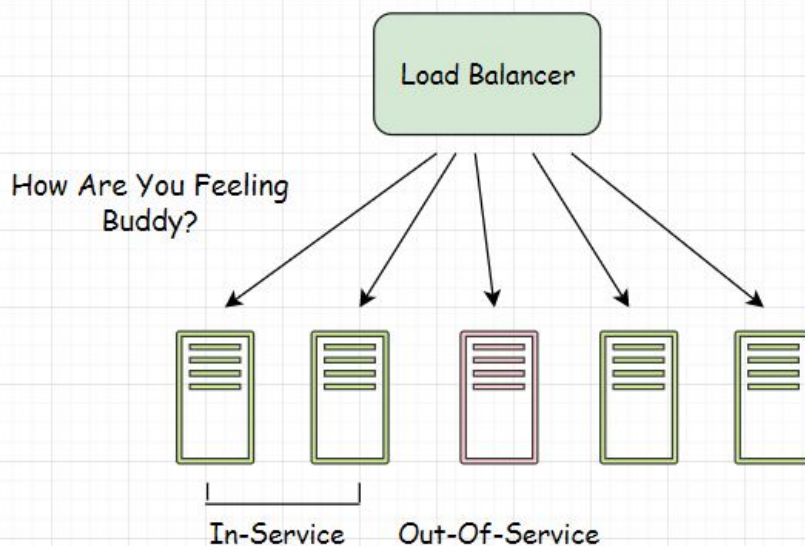
Ideally, a load balancer maintains a list of machines that are up and running in the cluster in real-time & the user requests are forwarded to only those machines that are in service. If a machine goes down it is removed from the list.

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, Machine* they all mean the same thing & can be used interchangeably.

After the out of service instance comes back online and becomes in service, the load

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, let's understand how load balancers work and to understand that well, we need to first understand the *DNS Domain Name System.*

We'll do that in the lesson up next.