Clusters and High Availability - Part 1

This lesson provides insight into server clustering.

We'll cover the following

- Overview
- What is a cluster?
- What is high availability?
- How important is high availability to online services?

Overview

In this lesson, we will talk about the server cluster. Though, generally, as a developer or an architect, we aren't expected to manage clusters; they are taken care of by the *Ops* teams. However, to acquire an understanding of cloud computing and distributed systems, we should be aware of things like:

- *What is a cluster? How does it function?*
- *How do nodes talk to each other in a cluster?*
- *How do scalable services run in a distributed environment?*
- *How does a distributed node coordination work and so on?*

We may not have to manage the clusters our service runs on, but we do have to monitor them using open source tools, like *Grafana, Prometheus, CAdvisor*, and those provided by the cloud provider, to understand the memory footprint of the code, resource consumption of the service, system bottlenecks, and so on.

Running and managing clusters that are distributed across the globe is something that is not trivial; it requires domain expertise. This is the whole reason everyone, right from the indie devs to the tech giants, prefers to run their services on the cloud and be worry-free of anything and everything related to physical infrastructure management and maintenance.

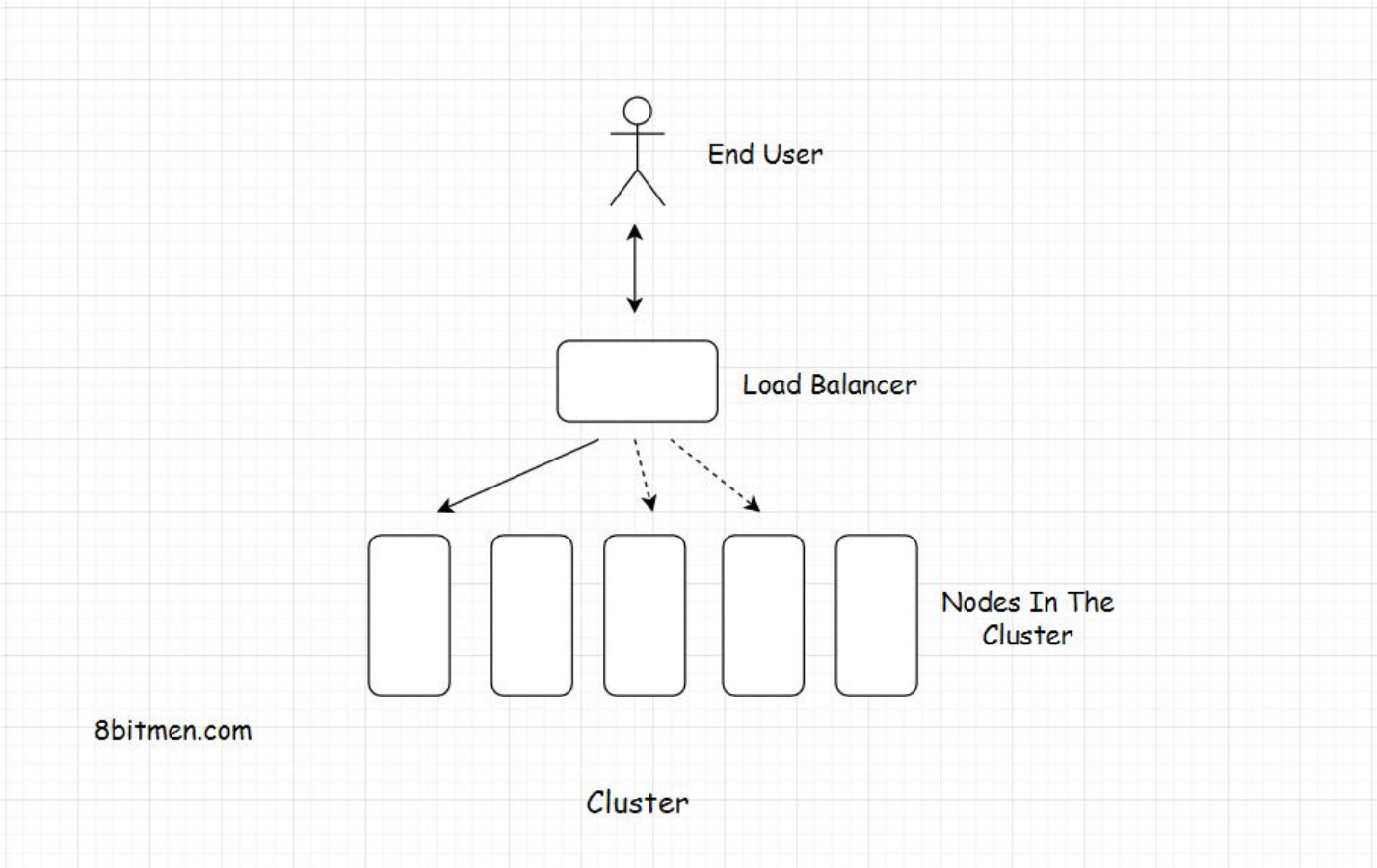
In this chapter, I discuss concepts that I think are essential for software developers

and everyone in general to acquire a good understanding of cloud computing.

So, without further ado, let's get started.

What is a cluster?

A cluster is a group of servers put together to make the system *highly available, fault-tolerant*, and *scalable*. Multiple servers running in conjunction also facilitate parallel processing of tasks.



Every large-scale internet service runs on single or multiple clusters to ensure minimum downtime. Before I discuss more clustering it's important to understand *high availability*

What is high availability?

High availability, also known as *HA,* is the system's ability to stay online despite having failures at the infrastructural level in real-time.

High availability ensures the uptime of the service is much more than the normal time. It improves the reliability of the system and ensures minimum downtime.

The sole mission of highly available systems is to stay online and stay connected. A very basic example of this is having back-up generators to ensure continuous

power supply in case of any power outages.

In the industry, HA is often expressed as a percentage. For instance, when the system is *99.99999%* highly available, it simply means *99.99999%* of the total hosting time the service will be up. You might often see this in the _ Service Level Agreements (SLA)_ of cloud platforms.

How important is high availability to online services?

It might not impact businesses that much if social applications go down for a bit and then bounce back. However, there are mission-critical systems like aircraft systems, spacecraft, mining machines, hospital servers, and finance stock market systems that just cannot afford to go down at any time. After all, lives depend on it.

The smooth functioning of the mission-critical systems relies on continual connectivity with their networks/servers.

These are the instances when we just cannot do without super highly available infrastructures. Besides, no service likes to go down, critical or not.

To meet the high availability requirements, systems are designed to be fault-tolerant, and their components are made redundant.

Let’s discuss fault-tolerance in the next lesson.