What Is Scalability?

This lesson is an introduction to scalability.

We'll cover the following

I am pretty sure, being in the software development universe, you've come across this word a lot many times. *Scalability*. What is it? Why is it so important? Why is everyone talking about it? Is it important to scale systems? What are your plans or contingencies to scale when your app or the platform experiences significant traffic growth?

This chapter is a deep dive into scalability. It covers all the frequently asked questions on it such as: what does scalability mean in the context of web applications, distributed systems or cloud computing? Etc.

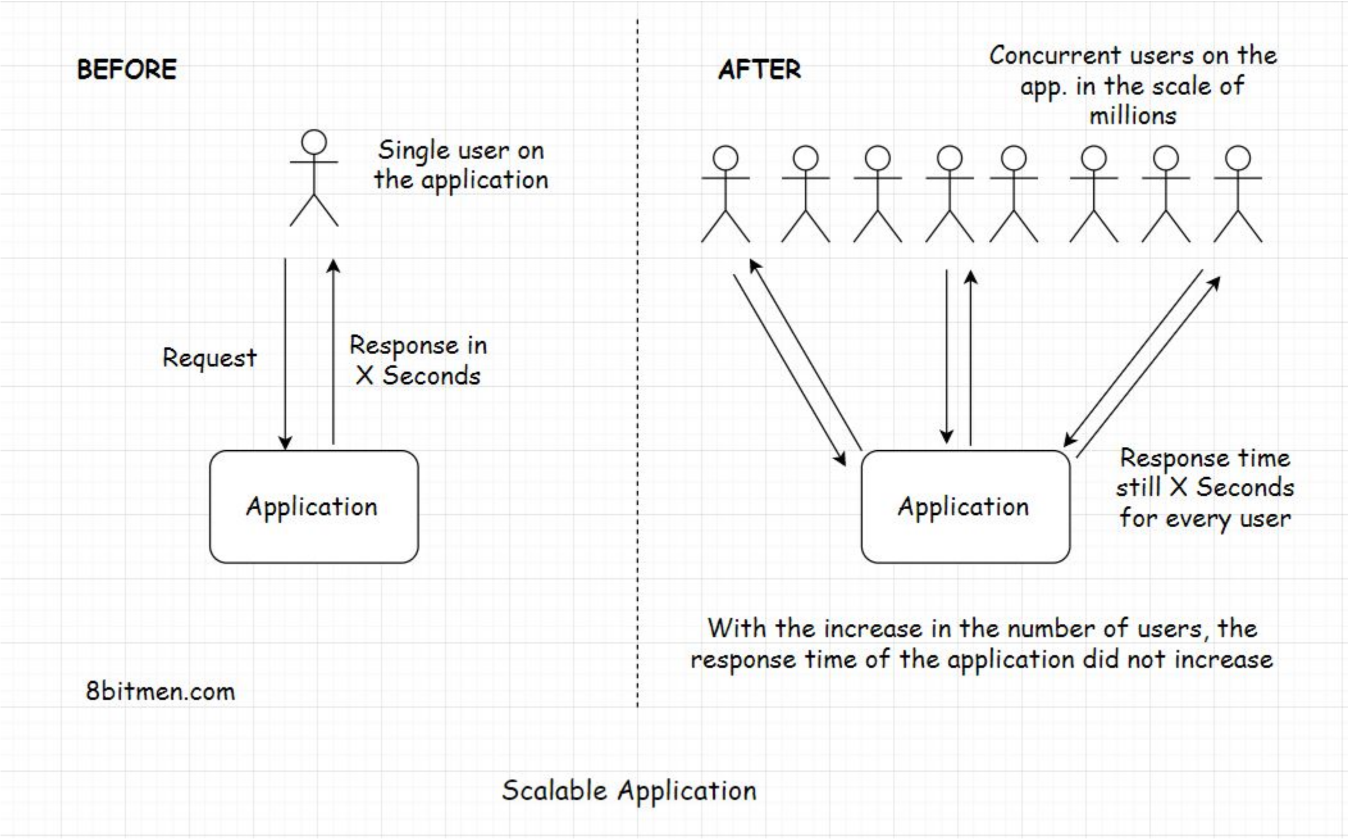
So, without further ado. Let's get started.

What is Scalability?

Scalability means the ability of the application to handle & withstand increased workload without sacrificing the latency.

For instance, if your app takes x seconds to respond to a user request. It should take the same x seconds to respond to each of the million concurrent user requests on your app.

The backend infrastructure of the app should not crumble under a load of a million concurrent requests. It should scale well when subjected to a heavy traffic load & should maintain the latency of the system.


BEFORE
Single user on the application
Request
Response in X Seconds
Application
8bitmen.com
AFTER
Concurrent users on the app. in the scale of millions
Application
Response time still X Seconds for every user
With the increase in the number of users, the response time of the application did not increase
Scalable Application

What Is Latency?

Latency is the amount of time a system takes to respond to a user request. Let's say you send a request to an app to fetch an image & the system takes 2 seconds to respond to your request. The latency of the system is 2 seconds.

Minimum latency is what efficient software systems strive for. No matter how much the traffic load on a system builds up, the latency should not go up. This is what scalability is.

If the latency remains the same, we can say yeah, the application scaled well with the increased load & is highly scalable.

Let's think of scalability in terms of *Big-O notation*. Ideally, the complexity of a system or an algorithm should be *O(1)* which is *constant time* like in a key-value database.

A program with the complexity of *O(n^2)* where n is the size of the data set is not scalable. As the size of the data set increases the system will need more computational power to process the tasks.

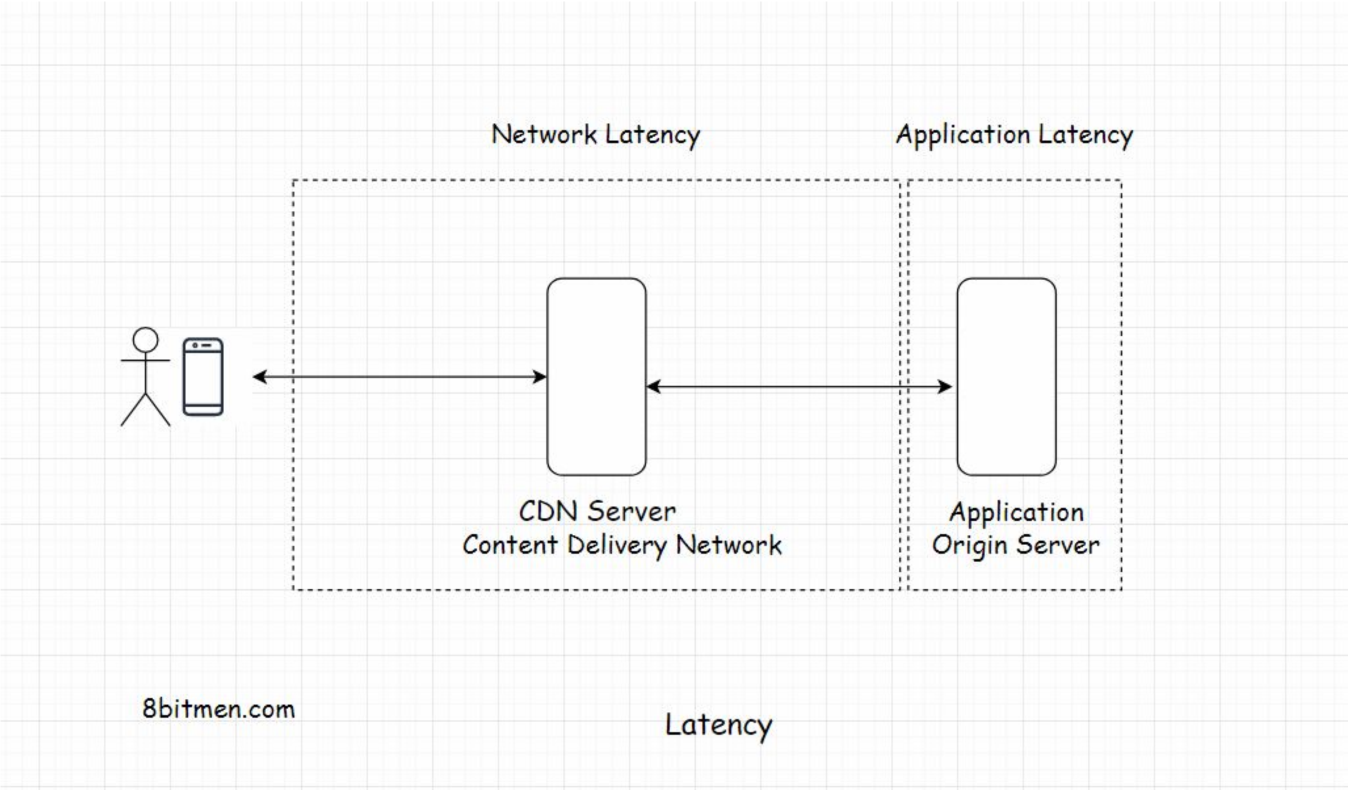
So, how do we measure latency?

Measuring Latency

Latency is measured as the time difference between the action that the user takes on the website, it can be an event like the click of a button, & the system response in reaction to that event.

This latency is generally divided into two parts:

1. Network Latency
2. Application Latency



Network Latency

Network Latency is the amount of time that the network takes for sending a data packet from point A to point B. The network should be efficient enough to handle the increased traffic load on the website. To cut down the network latency, businesses use CDN & try to deploy their servers across the globe as close to the end-user as possible.

Application Latency

Application Latency is the amount of time the application takes to process a user request. There are more than a few ways to cut down the application latency. The first step is to run stress & load tests on the application & scan for the bottlenecks that slow down the system as a whole. I've talked more about it in the upcoming lesson.

Why Is Low Latency So Important For Online Services?

Latency plays a major role in determining if an online business wins or loses a customer. Nobody likes to wait for a response on a website. There is a well-known saying if you want to test a person’s patience, give him a slow internet connection 😋

If the visitor gets the response within a stipulated time, great or he bounces off to another website.

There are numerous market researches that present the fact that high latency in applications is a big factor in customers bouncing off a website. If there is money involved, zero latency is what businesses want, only if that was possible.

Think of massive multiplayer online MMO games, a slight lag in an in-game event ruins the whole experience. A gamer with a high latency internet connection will have a slow response time despite having the best reaction time of all the players in an arena.

Algorithmic trading services need to process events within milliseconds. Fintech companies have dedicated networks to run low latency trading. The regular network just won’t cut it.

We can realize the importance of low latency by the fact that Huawei & Hibernia Atlantic in the year 2011 started laying a fibre-optic link cable across the Atlantic Ocean between London & Newyork, that was estimated having a cost of approx. $300M, just to save traders 6 milliseconds of latency.