



What is Scalability?

This lesson introduces scalability.

We'll cover the following



- What is scalability?
- What is latency?
- Measuring latency
 - Network latency
 - Application latency
- Why is low latency so important for online services?

I am pretty sure, being in the software development universe, you've come across the word *scalability* many times. What is it? Why is it so important? Why is everyone talking about it? 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 about it including: 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?#

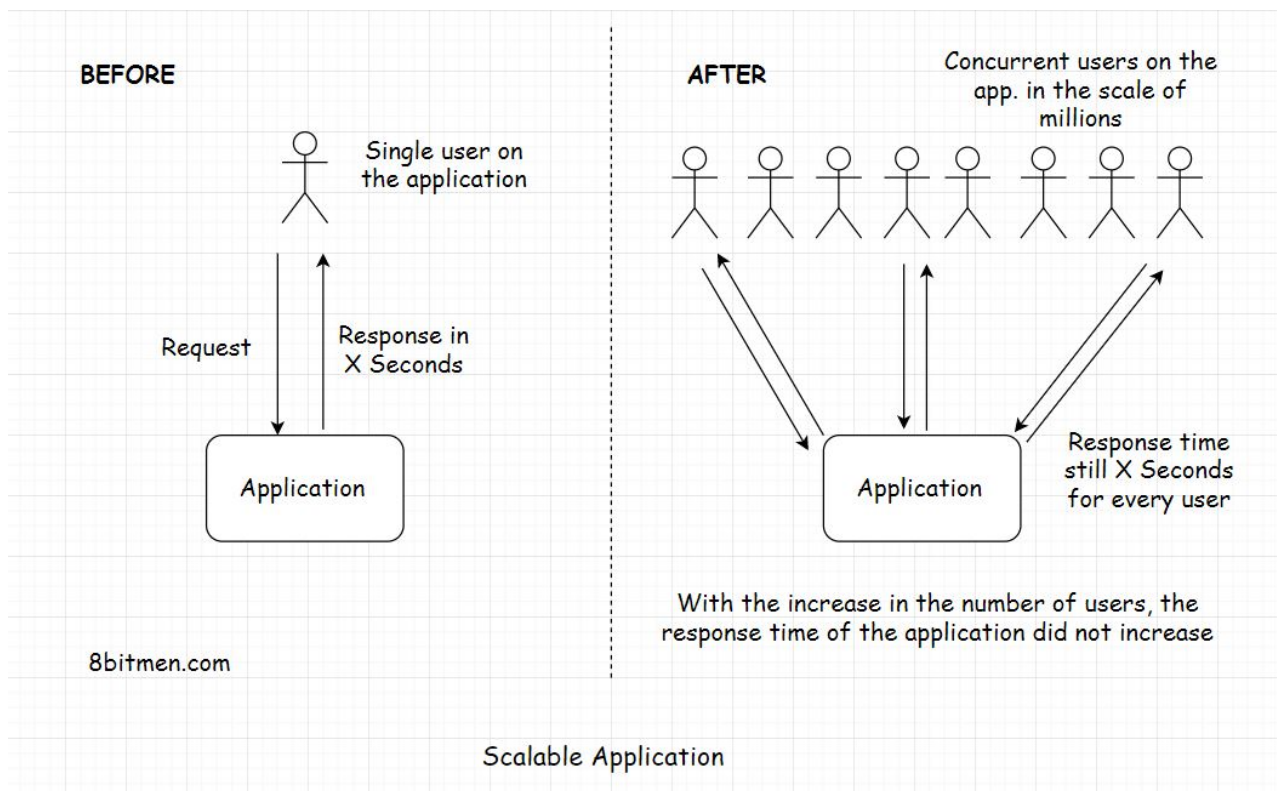
Scalability means the ability of the application to handle and withstand

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 and should maintain the latency of the system.



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 and 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 that the application scaled well with the increased load and 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.

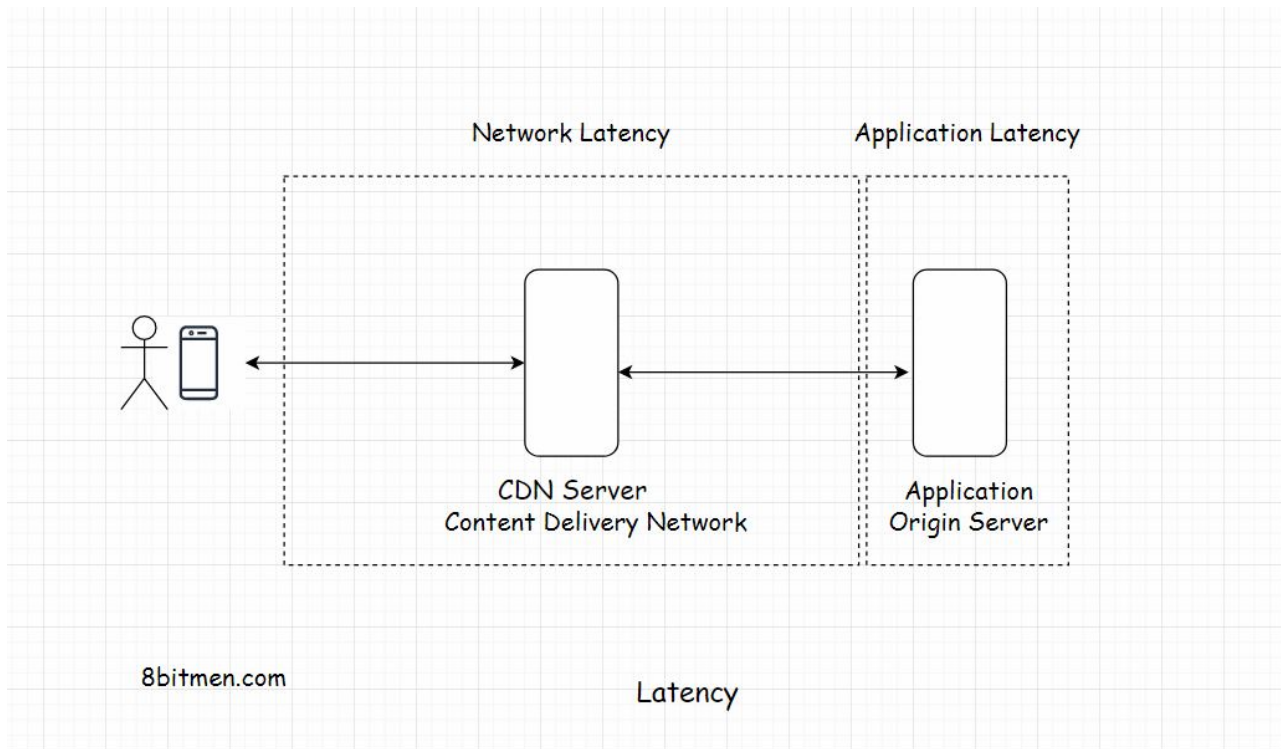
So, how do we measure latency?

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, and 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 to send 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 and 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 and load tests on the application and scan for the bottlenecks that slow down the system as a whole. We'll talk 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 them a slow internet connection.”

If the visitor gets the response within a stipulated time, great otherwise they’ll bounce off to another website.

There is ample market research that concludes 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. If only if this 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 in 2011 Huawei and Hibernia Atlantic started laying a fiber-optic link cable across the Atlantic Ocean between London and New York (<https://www.telegraph.co.uk/technology/news/8753784/The-300m-cable-that-will-save-traders-milliseconds.html>). This property was estimated to

cost approximately \$300M just to save traders 6 milliseconds of latency.



[← Back](#)

[Next →](#)

Web Architecture Quiz - Part 2

Types of Scalability

☒ Mark as Completed

Report an Issue