DNS Load Balancing

In this lesson, we will understand DNS Load balancing

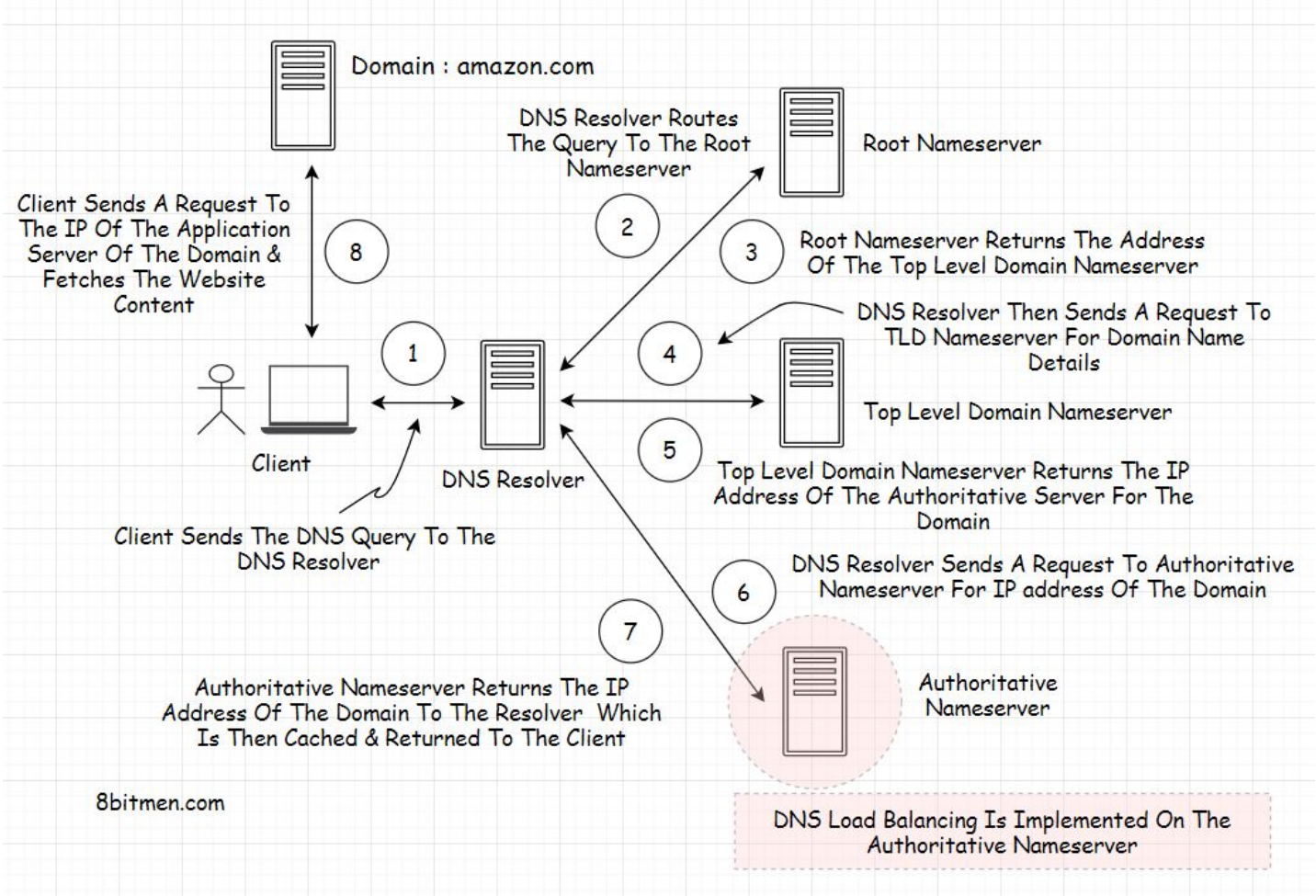
We'll cover the following

DNS Load Balancing

In the previous lesson, we understood how the *DNS query lookup process* works and the role of different servers in the domain name system. The final end server, in the lookup chain, is the *authoritative server* that returns the *IP address* of the domain.

When a large-scale service such as *amazon.com* runs, it needs way more than a single machine to run its services. A service as big as *amazon.com* is deployed across multiple data centers in different geographical locations across the globe.

To spread the user traffic across different clusters in different data centers. There are different ways to setup load balancing. In this lesson, we will discuss *DNS load balancing* that is setup at the *DNS* level on the *authoritative server*.



DNS load balancing enables the *authoritative server* to return different *IP addresses* of a certain domain to the clients. Every time it receives a query for an *IP*, it returns a list of *IP addresses* of a domain to the client.

With every request, the *authoritative server* changes the order of the *IP addresses* in the list in a *round-robin* fashion.

As the client receives the list, it sends out a request to the first *IP address* on the list to fetch the data from the website. The reason for returning a list of *IP addresses* to the client is to enable it to use other *IP addresses* in the list in case the first doesn't return a response within a stipulated time.

When another client sends out a request for an *IP address* to the *authoritative server*, it re-orders the list and puts another *IP address* on the top of the list following the *round-robin algorithm*.

Also, when the client hits an *IP* it may not necessarily hit an application server, it may hit another load balancer implemented at the data center level that manages the clusters of application servers.

Limitations Of DNS Load Balancing

Limitations Of DNS Load Balancing

DNS load balancing is largely used by companies to distribute traffic across multiple data centers that the application runs in. Though this approach has several limitations, for instance, it doesn't take into account the existing load on the servers, the content they hold, their request processing time, their *in-service* status and so on.

Also, since these *IP addresses* are cached by the client's machine and the *DNS Resolver*, there is always a possibility of a request being routed to a machine that is out of service.

DNS load balancing despite its limitations is preferred by companies because it's an easy and less expensive way of setting up load balancing on their service.

Recommended Read – Round Robin DNS