Elastic Load Balancers

Now, when making complex architectures we may often find ourselves operating multiple versions of a resource to meet various scaling and operational needs. When one EC2 instance is not sufficient to satisfy all the requests for example, we often spin up and add more EC2 instances to the VPC using Auto-Scaling Groups as discussed before.

Now, in the above scenario, how do we ensure that all the EC2 instances are having the network traffic distributed properly among them, and that certain instances are being overburdened with work while others lay relatively dormant? The answer is **Load Balancers**.

Load balancers are systems that distribute network or application traffic across multiple servers, ensuring that no single server is overwhelmed with too much traffic, which can lead to performance degradation or outages.

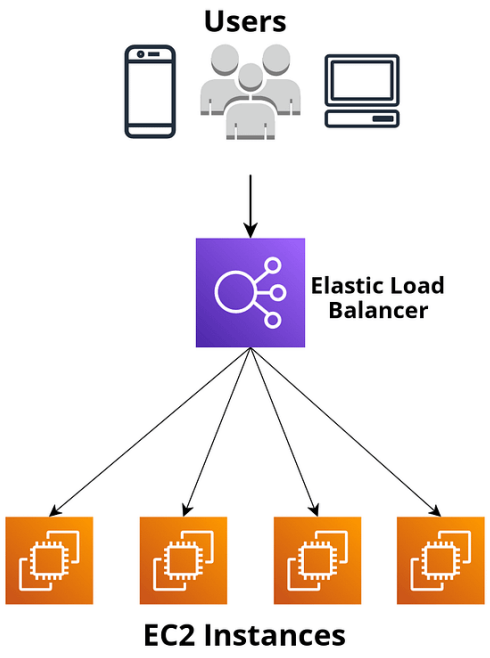
Load Balancers serve several critical purposes:

- **Traffic Distribution**: They balance incoming requests to prevent any single server from becoming a bottleneck.
- **High Availability**: Load balancers enhance the availability of applications by routing traffic to healthy instances. If one server goes down, the load balancer reroutes requests to other operational servers.
- **Scalability**: As demand increases, load balancers make it easier to add more servers, ensuring that applications can handle varying traffic loads.
- **Performance Optimization**: They can help optimize response times by directing requests to the most capable or least loaded servers.

In summary, load balancers allow us to combat **Single points of failure**, **Traffic spikes** and **Resource management**.

Elastic Load balancers (ELB) is the package provided to us by AWS that allows us to implement load balancers, distributing incoming application traffic across multiple AWS resources, such as EC2 instances, containers and IP addresses, enhancing the availability and fault tolerance of your applications by balancing the load and ensuring traffic is directed to healthy instances as mentioned above.

Though there are different types of Elastic Load Balancers (will be discussed in the next section), they all occupy similar positions in AWS network architecture, usually serving as a middleman between user agents and a swathe of compute or database instances as shown below:



Simple architecture diagram showcasing an Elastic Load Balancer

Finally, it is worth noting that Elastic Load Balancers, just like subnets within a VPC, can only be connected to Availability Zones in a single region.