Load balancing:

Load balancing refers to efficiently distributing incoming network traffic across a group of backend servers, also known as a *server farm* or *server pool*.

Modern high-traffic websites must serve hundreds of thousands, if not millions, of concurrent requests from users or clients and return the correct text, images, video, or application data, all in a fast and reliable manner. To cost-effectively scale to meet these high volumes, modern computing best practice generally requires adding more servers.

A load balancer acts as the “traffic cop” sitting in front of your servers and routing client requests across all servers capable of fulfilling those requests in a manner that maximizes speed and capacity utilization and ensures that no one server is overworked, which could degrade performance. If a single server goes down, the load balancer redirects traffic to the remaining online servers. When a new server is added to the server group, the load balancer automatically starts to send requests to it.

In this manner, a load balancer performs the following functions:

- Distributes client requests or network load efficiently across multiple servers.
- Ensures high availability and reliability by sending requests only to servers that are online
- Provides the flexibility to add or subtract servers as demand dictates.

Ref**:** https://www.nginx.com/resources/glossary/load-balancing/

Scalability:

Scalability is the ability of a system to expand to meet your business needs. You scale a system by adding extra hardware or by upgrading the existing hardware without changing much of the application.

Latency:

In web performance circles, "**latency**" is the amount of time it takes for the host server to receive and process a request for a page object. The amount of latency depends largely on how far away the user is from the server.

Reliability:

The term reliability refers to the ability of a computer-related hardware or software component to consistently perform according to its specifications. In theory, a reliable product is totally free of technical errors.

In other words, *Reliability* can be defined as the probability that a system will produce correct outputs up to some given time *t*.

Availability:

Availability means the probability that a system is operational at a given time, i.e. the amount of time a device is actually operating as the percentage of total time it should be operating.