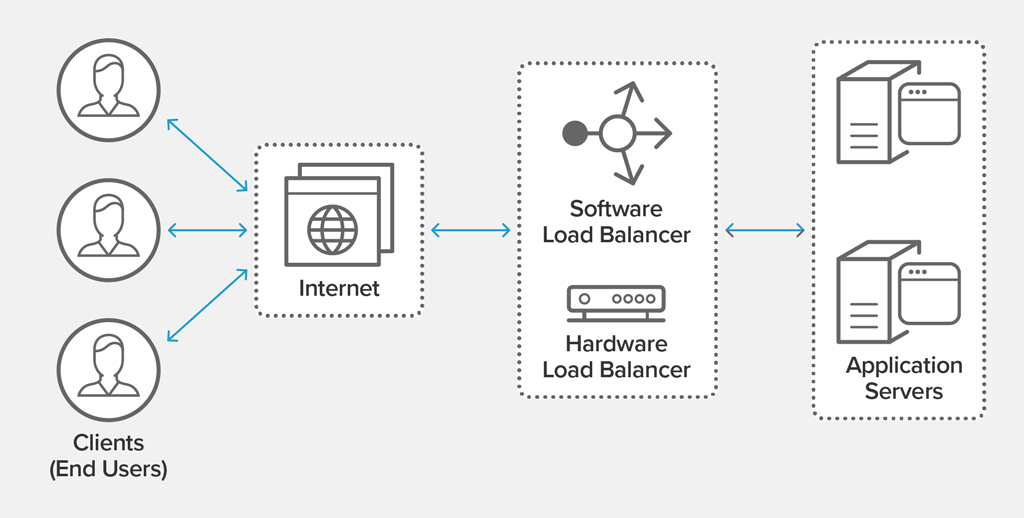
**Load balancing** refers to efficiently distributing incoming network traffic across a group of backend servers, also known as a *server farm* or *server pool*.

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](https://www.nginx.com/solutions/adc) 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

load balancing diagram

### Load Balancing Algorithms

Different load balancing algorithms provide different benefits; the choice of load balancing method depends on your needs:

* **Round Robin** – Requests are distributed across the group of servers sequentially.
* **Least Connections** – A new request is sent to the server with the fewest current connections to clients. The relative computing capacity of each server is factored into determining which one has the least connections.
* **Least Time** – Sends requests to the server selected by a formula that combines the  
  fastest response time and fewest active connections. Exclusive to NGINX Plus..
* **IP Hash** – The IP address of the client is used to determine which server receives the request.
* **Random with Two Choices** – Picks two servers at random and sends the request to the  
  one that is selected by then applying the Least Connections algorithm (or for NGINX Plus  
  the Least Time algorithm, if so configured)

Benefits of Load Balancing

* Reduced Downtime
* Scalable
* Redundancy
* Flexibility
* Efficiency
* Global Server Load Balancing

**Application Server**

Application server contains Web and EJB containers. It can be used for servlet, jsp, struts, jsf, ejb etc. It is a component based product that lies in the middle-tier of a server centric architecture.

It provides the middleware services for state maintenance and security, along with persistence and data access. It is a type of server designed to install, operate and host associated services and applications for the IT services, end users and organizations.

The block diagram representation of Application Server is shown below:



The Example of Application Servers are:

1. **JBoss:** Open-source server from JBoss community.
2. **Glassfish:** Provided by Sun Microsystem. Now acquired by Oracle.
3. **Weblogic:** Provided by Oracle. It more secured.
4. **Websphere:** Provided by IBM.