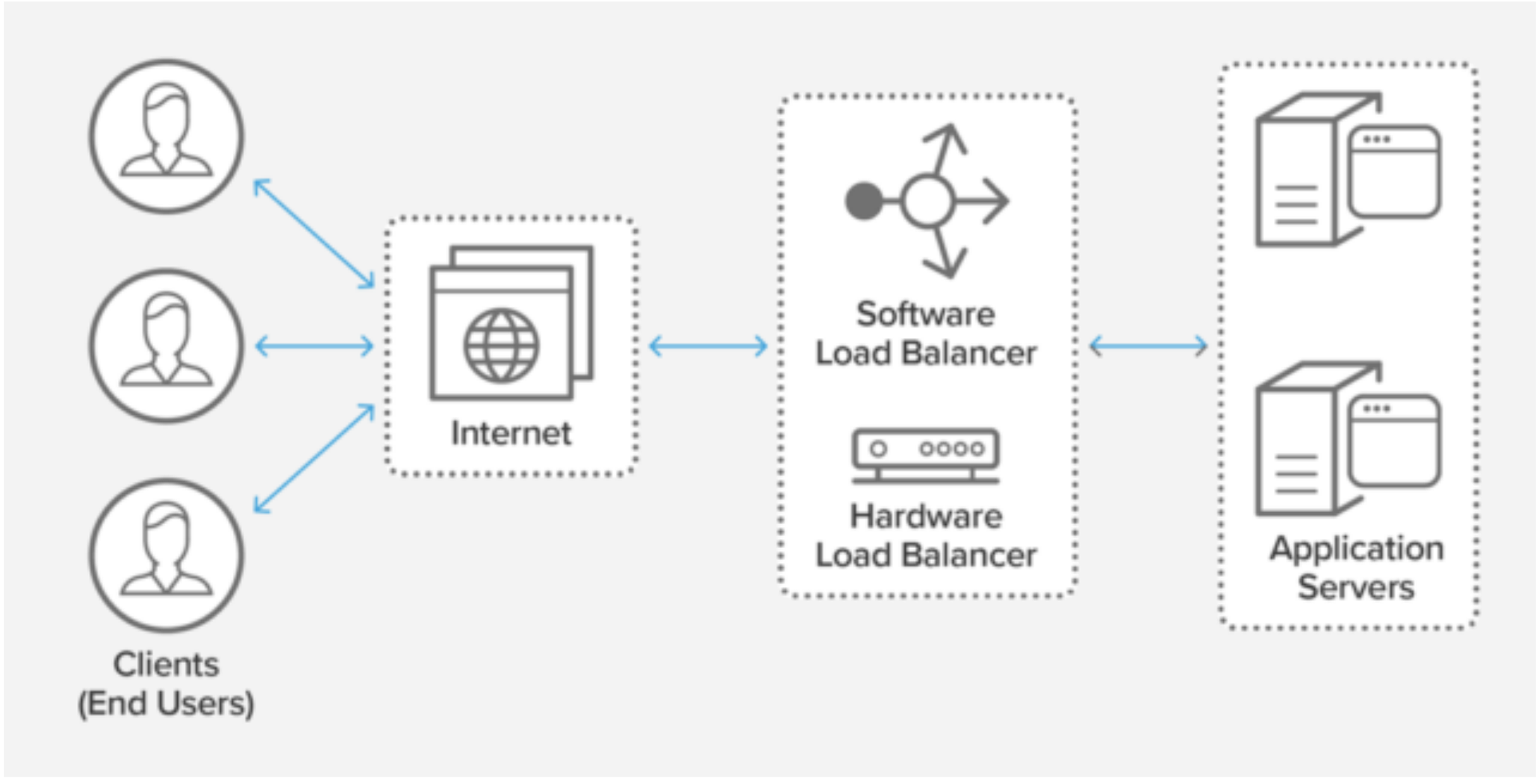
NGINX: Load Balancer Introduction

NGINX : Web-Server & Load Balancer

- 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.
- **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

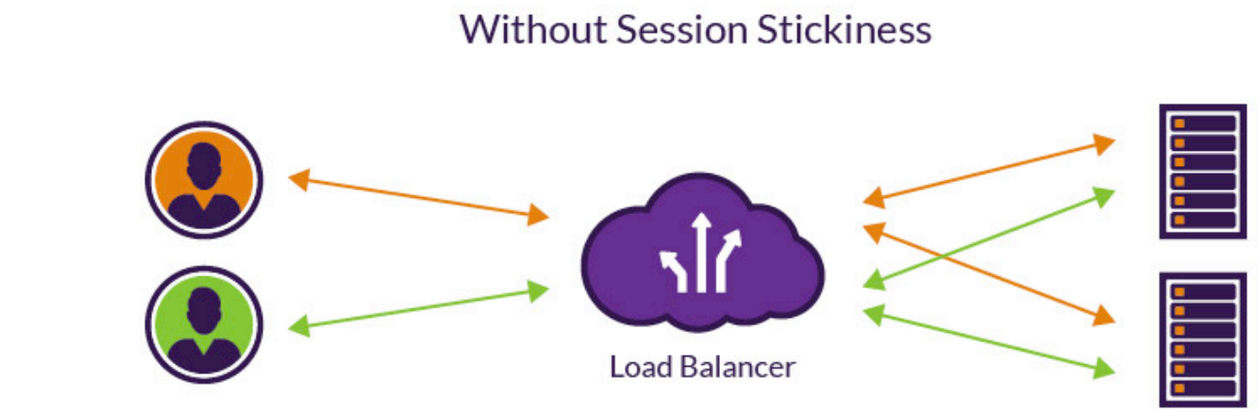
NGINX : Web-Server & Load Balancer

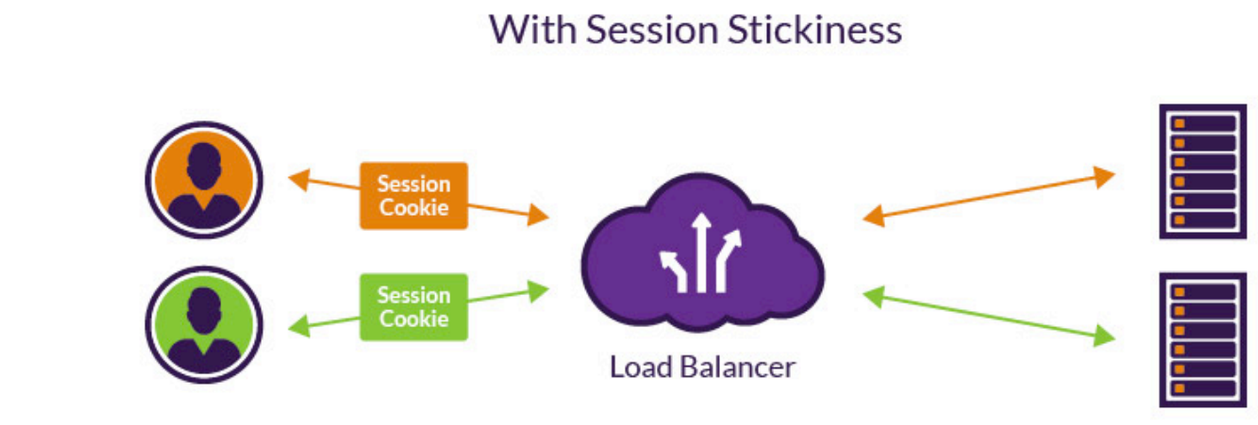


- **Load Balancing Algorithms -**
- **Round Robin -** Requests are distributed across the group of servers sequentially. It is easy for load balancers to implement, but does don't take into account the load already on a server.
- **Least Connection Method** - A new request is sent to the server with the fewest current connections to clients. Whereas round robin does not account for the current load on a server (only its place in the rotation), the least connection method does make this evaluation and, as a result, it usually delivers superior performance.
- **Least Response Time Method** - Sends requests to the server selected by a formula that combines the fastest response time and fewest active connections. More sophisticated than the least connection method and Exclusively used by **NGINX Plus**.

- **Load Balancing Algorithms -**
- **Least Bandwidth Method -** A relatively simple algorithm, the least bandwidth method looks for the server currently serving the least amount of traffic as measured in megabits per second (Mbps).
- **Hashing Methods** - Distributes requests based on a key you define, such as the client IP address or the request URL. NGINX Plus can optionally apply a consistent hash to minimize redistribution of loads if the set of upstream servers changes.
- **IP Hash** - The IP address of the client is used to determine which server receives the request.

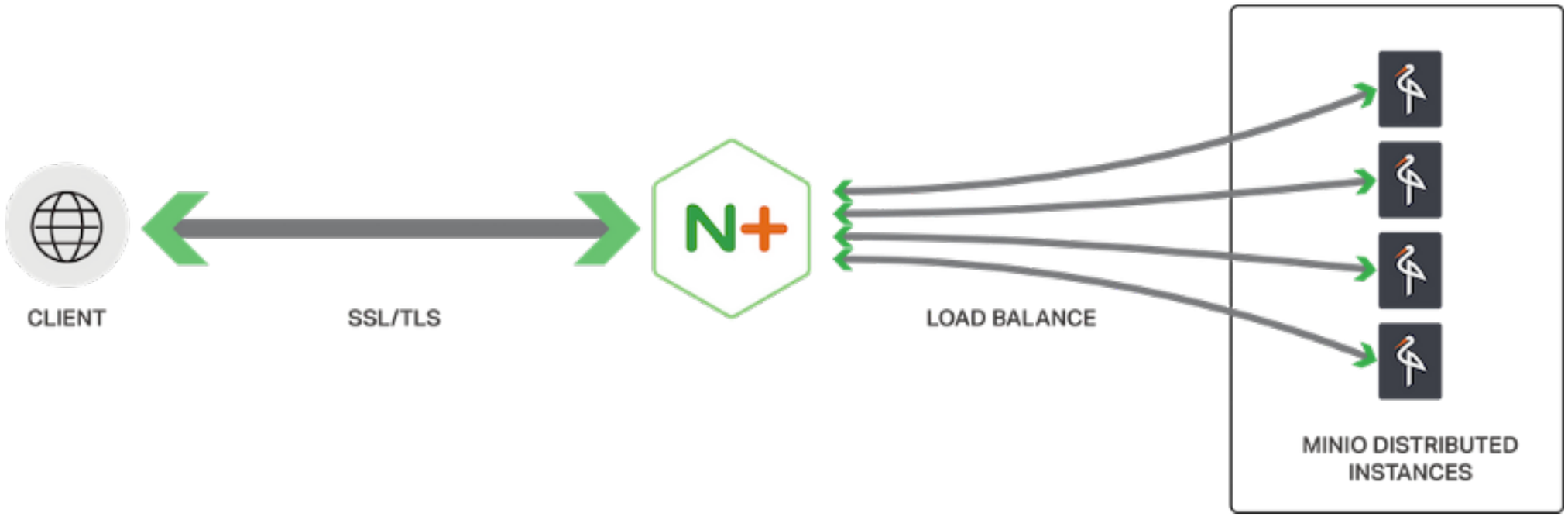
- Related Fundaments with Load Balancing -
- Session Persistence -





NGINX : Web-Server & Load Balancer

- **Related Fundaments with Load Balancing -**
- Dynamic Configuration of Server Groups(Auto Scaling)
- Hardware vs. Software Load Balancing
- NGINX As Load Balancer



Will see you in Next Lecture…

See you in next lecture …