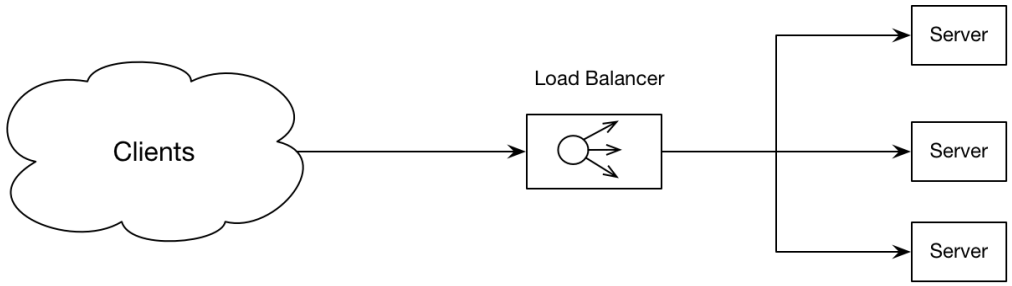
Kenny R N Wilson

# Load Balancing

A special router known as a load balancer routes requests across a cluster of servers to improve system availability, responsiveness and scalability. Load balancers enable horizontal scaling by adding more servers. The load balancer will avoid sending requests to a server that is overloaded, not responding or in error. In this way no server in the cluster becomes a single point of failure.



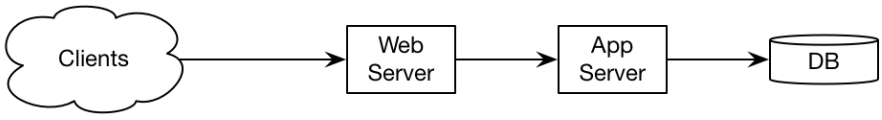

## Benefits

The benefits of load balancing are

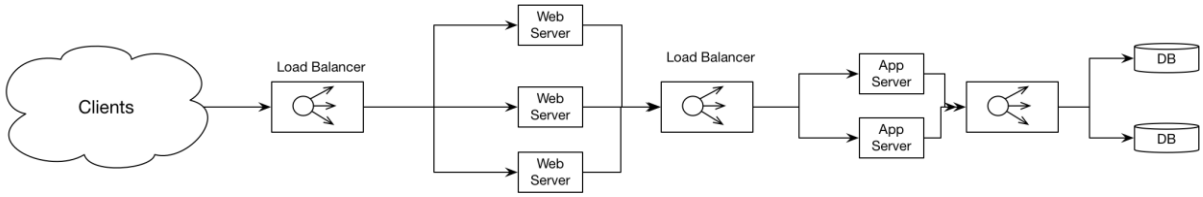
- Faster response time
- Increased availability
- Increased scalability

## Example

Consider a classic three tier web application.

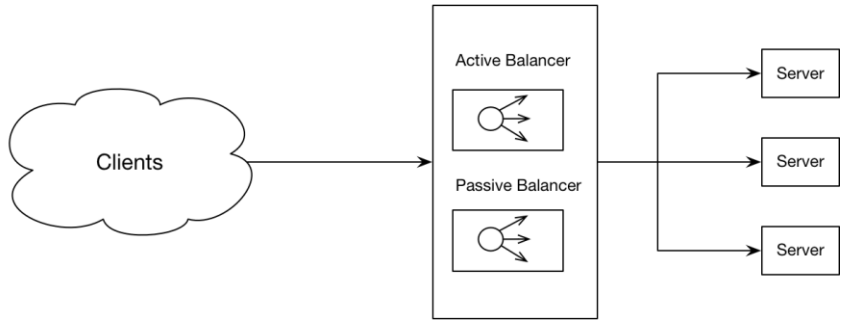



We can maximise scalability and availability by using a load balancer at every tier.




What if the load balancer fails? It is the single point of failure in the system. We can prevent this by using a pair of load balancer that communicate with each other. One can be active and

one passive. If the passive load balancer detects the active load balancer is down it can start processing requests.




## Routing Algorithms

All the below routing algorithms only choose servers actively processing requests. Heart beating or similar techniques can be used to determine which systems are running.

| | |
|---|---|
| **Round Robin** | The simplest algorithm. Can be enhanced by assigning weights to servers. Servers with more capacity get higher weights and get assigned more requests. |
| **IP Hashing** | Hash the IP address of the client to determine which server to assign to |
| **Least Connections** | Assign to the server currently processing the fewest active connections. |
| **Lowest Response Time** | Assign to the server with the lowest average response time |
| **Lowest Bandwidth** | Assign to the server handling the lowest bandwidth. |