# Horizontal & Vertical Scaling

Systems design a procedure by which we define the architecture of a system to satisfy given requirements. It is a technique by which the required amounts of scalability, reliability, performance and consistency are satisfied in real world systems.

The first concept in designing a system is scalability. We discuss the two main approaches to solve this problem: Horizontal scaling and vertical scaling.

Horizontal scaling is adding more machines to deal with increasing requirements. These machines handle requests in parallel to improve user experience.

Vertical scaling is replacing the current machines with more advanced machines to improve throughput and hence response time. The techniques are used in conjunction in real world systems.

|  |  |
| --- | --- |
| Horizontal Scaling | Vertical Scaling |
| Load balancing is required to distribute the load between different machines. | NA |
| The communication between systems is slow since RPC calls. | Communication is fast since inter process communication. |
| Data inconsistency | Consistent |
| Resilient. If one server fails the other server can respond. | Single point of failure. |
| Scales well with increase in number of users. | Has hardware limitation. |