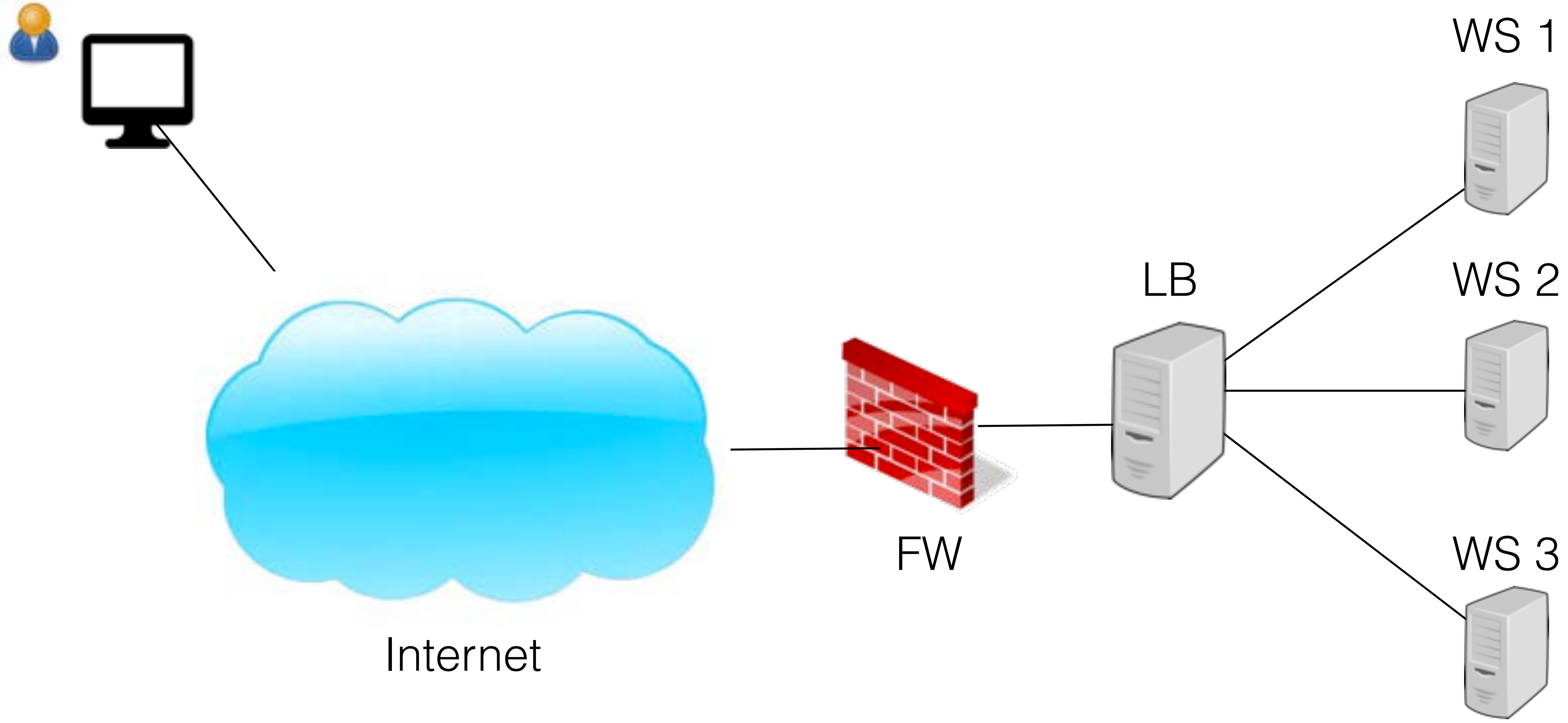
Load Balancer

Load Balancing

- Important characteristics of cloud is scalability
- Cloud computing resources can scale on demand
- Load Balancer distributes workload across multiple servers
- Helps achieve optimum utilization of resources
  - Achieve high availability and reliability.
  - In the even of a resource failure, the LB can reroute traffic

Deployment



Algorithms

- Load balancer can be programmed to distribute traffic in a variety of ways.
- Following are some of the algorithms (not exhaustive)
- Round Robin
  - Servers are selected one by one to serve the incoming requests
  - Each server gets a request in a circular fashion
  - All servers have same priority.
- Weighted Round Robin
  - Servers are assigned some weight
  - Incoming requested are directed proportion to the weight.
  - Each server will not receive same number of requests.

Algorithms

- Low Latency
  - Load balancer monitors the latency of each server
  - Incoming requests are routed to server with the lowest latency
- Priority
  - Each server is assigned a priority
  - Incoming request is routed to the highest priority server that is available
  - Lower priority server gets traffic when high priority servers are busy

Implementation

- Can be implemented in hardware or software

- Software based LB runs on OS
  - Easily run virtualized

- Hardware based solutions use specialized hardware to distribute traffic.