PES UNIVERSITY

CELEBRATING 50 YEARS

CLOUD SERVICES

S Thenmozhi

Department of Computer Applications

CLOUD SERVICES

Cloud Computing Essentials

S Thenmozhi

Department of Computer Applications

Load Balancing

Load balancing is the practice of evenly distributing traffic, workloads, and client requests across multiple servers

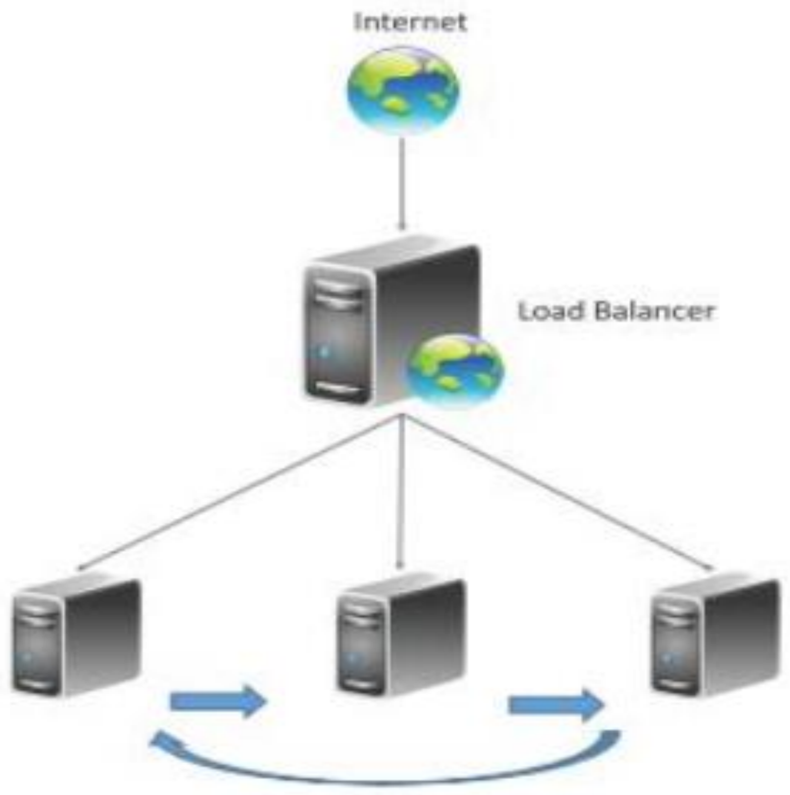
This is done to ensure that no single resource is overburdened.

Goals:

- Achieve maximum utilization of resources
- Minimizing the response times
- Maximizing throughput

CLOUD SERVICES

Load Balancing



Load Balancing

Load Balancing could be

- Network Load Balancing
- Application Load Balancing
- Database Load Balancing

Load Balancing Algorithms

Round Robin Load Balancing

- Servers are selected one by one to serve incoming requests
- Non-hierarchical circular fashion with no priority

Weighted Round Robin Load Balancing

- Servers are assigned some weights
- Incoming requests are proportionally routed using static or dynamic ratio of respective weights

Load Balancing Algorithms

Low Latency Load Balancing

- Load balancer monitors the latency of each server
- Incoming request is routed to low latency server

Least Connections Load Balancing

- Requests are routed to server with least number of connections

Load Balancing Algorithms

Priority Load Balancing

- Each server is assigned a priority
- Requests are routed to server with highest priority as long as the server is available
- When it fails, then incoming traffic is routed to next priority server

Overflow Load Balancing

- Similar to priority load balancing, routed to low priority server when the higher priority server overflows

Load Balancing Persistence Approaches

- Load balancing can route successive requests from a user
- Maintaining the state or the information of the session is important
- Persistence Approaches
  - Sticky Sessions
  - Session Database
  - Browser Cookies
  - URL re-writing

Load Balancing Persistence Approaches

Sticky Sessions

- All requests belonging to a user is routed to same server
- Session management is simple
- If server fails, all sessions belonging to that server is lost
- No automatic failover is possible

Session Database

- Session information is stored separately in a session database
- It is often replicated to avoid single point failure
- Allows automatic failover

Load Balancing Persistence Approaches

Browser Cookies

- Session information is stored in the client side
- Session management is easy
- Least amount of overhead for the load balancer

URL re-writing

- Stores the session information by modifying the URL’s on the client side
- The amount of session information that can be stored is limited
- Applications that require larger amounts of session information , this will not work

Load Balancers

Examples

Load Balancer	Type
NginX	Software
HA Proxy	Software
Pound	Software
Varish	Software
Cisco Systems Catalyst 6500	Hardware
Coyote Point Equalizer	Hardware
F5 Networks BIG-IP LTM	Hardware
Barracuda Load Balancer	Hardware

PES UNIVERSITY

CELEBRATING 50 YEARS

THANK YOU

S Thenmozhi

Department of Computer Applications

thenmozhis@pes.edu

+91 80 6666 3333 Extn 393