Assignment 9

LoadBalancing, AutoScaling and Serverless Computing

Nandeep Nelagondanahalli
Nithin Kartha
Vaishak P Dinesh

LoadBalancing

Load balancing refers to efficiently distributing incoming network traffic across a group of backend servers, also known as a server farm or server pool. ... In this manner, a load balancer performs the following functions: Distributes client requests or network load efficiently across multiple servers.

GCP

- HTTP(S) Load Balancing
- Stackdriver Logging
- TCP/SSL Load Balancing
- SSL Offload
- UDP Load Balancing
- Cloud CDN Integration
- Affinity
- High Fidelity Health Checks
- Seamless Autoscaling

Azure

- **Port forwarding**
- **Application agnostic and transparent**
- **Automatic reconfiguration**
- **Health probes**
- **Outbound connections (Source NAT)**
- **Internal load balancer**
- **Backend pool**
- **HA Ports**

AWS

- **High Availability**
- **Health Checks**
- **Security Features**
- **TLS Termination**
- **Layer 4 or Layer 7 Load Balancing**
- **Operational Monitoring**
- **HTTPS Support**
- **Server Name Indication (SNI)**
- **Content-Based Routing**
- **Containerized Application Support**
- **Web Application Firewall**
- **Request Tracing**
- **Logging**

AutoScaling

AutoScaling is a method used in cloud computing, whereby the amount of computational resources in a server farm, typically measured in terms of the number of active servers, scales automatically based on the load on the farm. It is closely related to, and builds upon, the idea of load balancing.

GCP

- **Reducing Infrastructure Cost**
- **Reducing Operational Cost**
- **Avoiding long synchronous call chains**

Azure

- Integrated Development Experience (IDE)
- Continuous integration and continuous delivery
- Monitoring, logging, and diagnostics
- SDKs

AWS

- Automatically maintain performance
- Pay only for what you need
- Unified scaling
- Automatic resource discovery
- Built-in scaling strategies
- Predictable scaling
- Smart scaling policies
- Fully-managed

Serverless Computing

Serverless computing is a cloud computing execution model in which the cloud provider dynamically manages the allocation of machine resources. Pricing is based on the actual amount of resources consumed by an application, rather than on pre-purchased units of capacity.

GCP

- H

Azure

- P

AWS

- No server management
- Flexible scaling
- Automated high availabilty
- Cloud logic layer
- Orchestration & state management
- Responsive data sources
- Application modeling framework
- Developer ecosystem
- Application & integrations library
- Security & access control
- Reliability & performance
- Global scale & reach

Thank You !