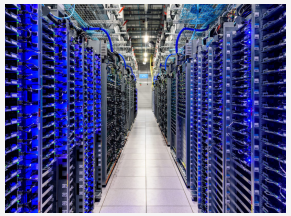
CS 436
Cloud Computing

21.03.2024

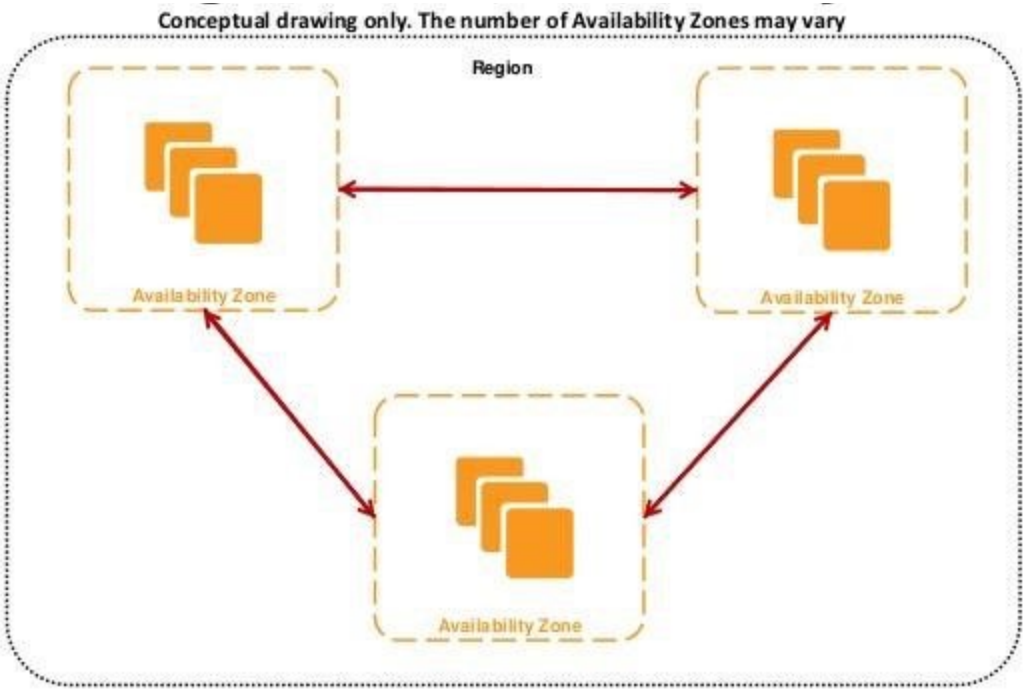
Cloud Computing Infrastructure over the globe



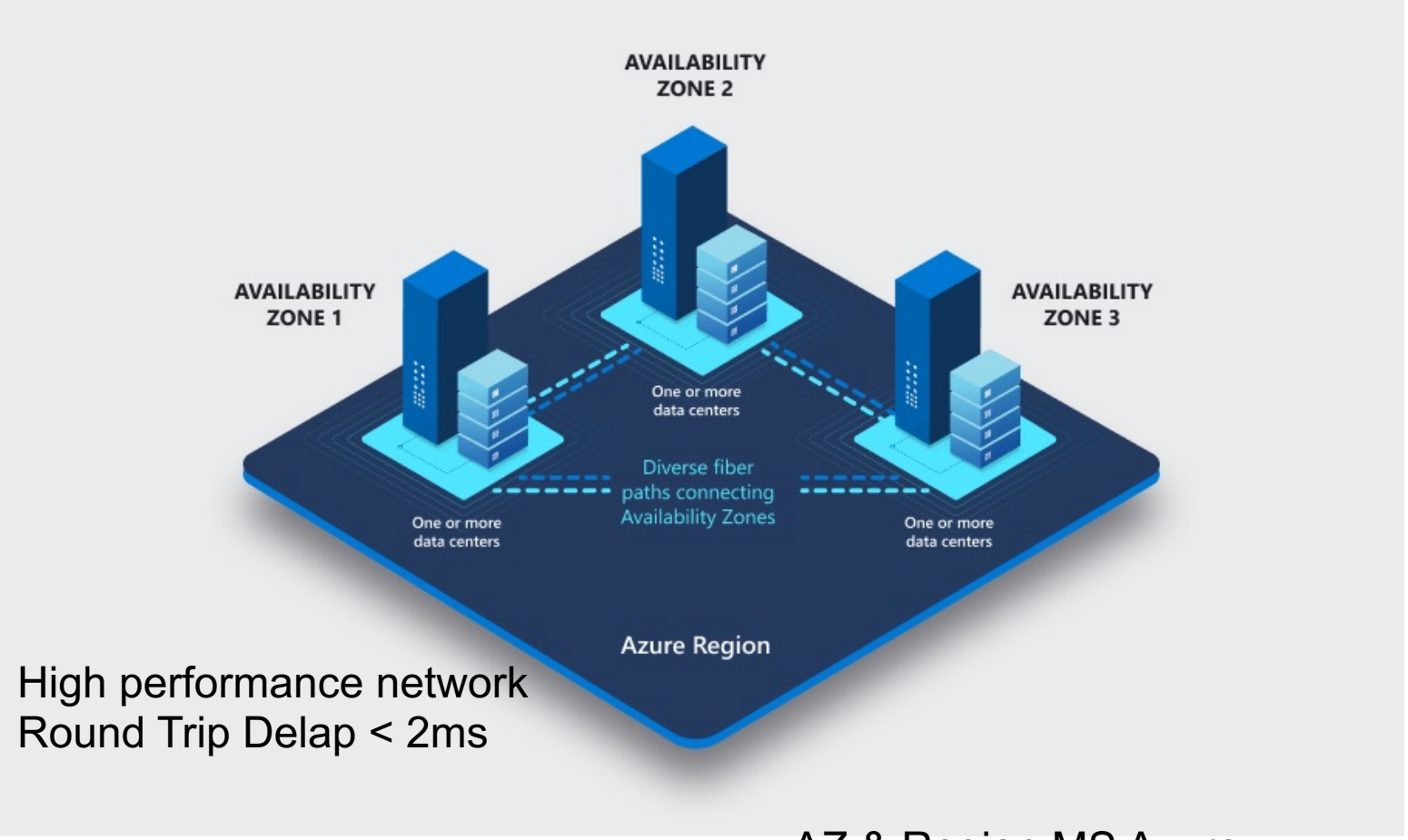
GCP Regions

Cloud Computing Infrastructure over the globe

Availability Zone vs Region



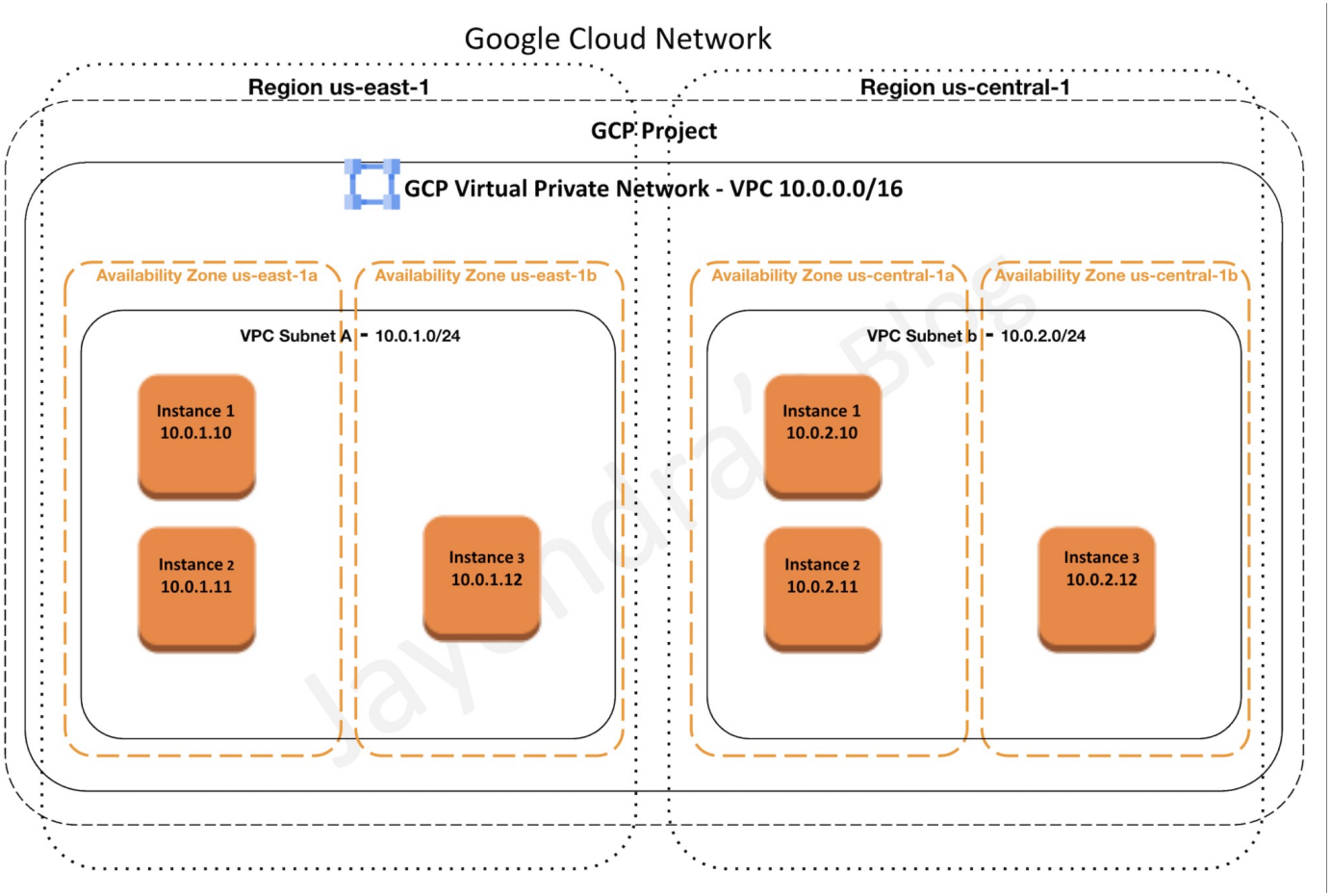
Availability Zone



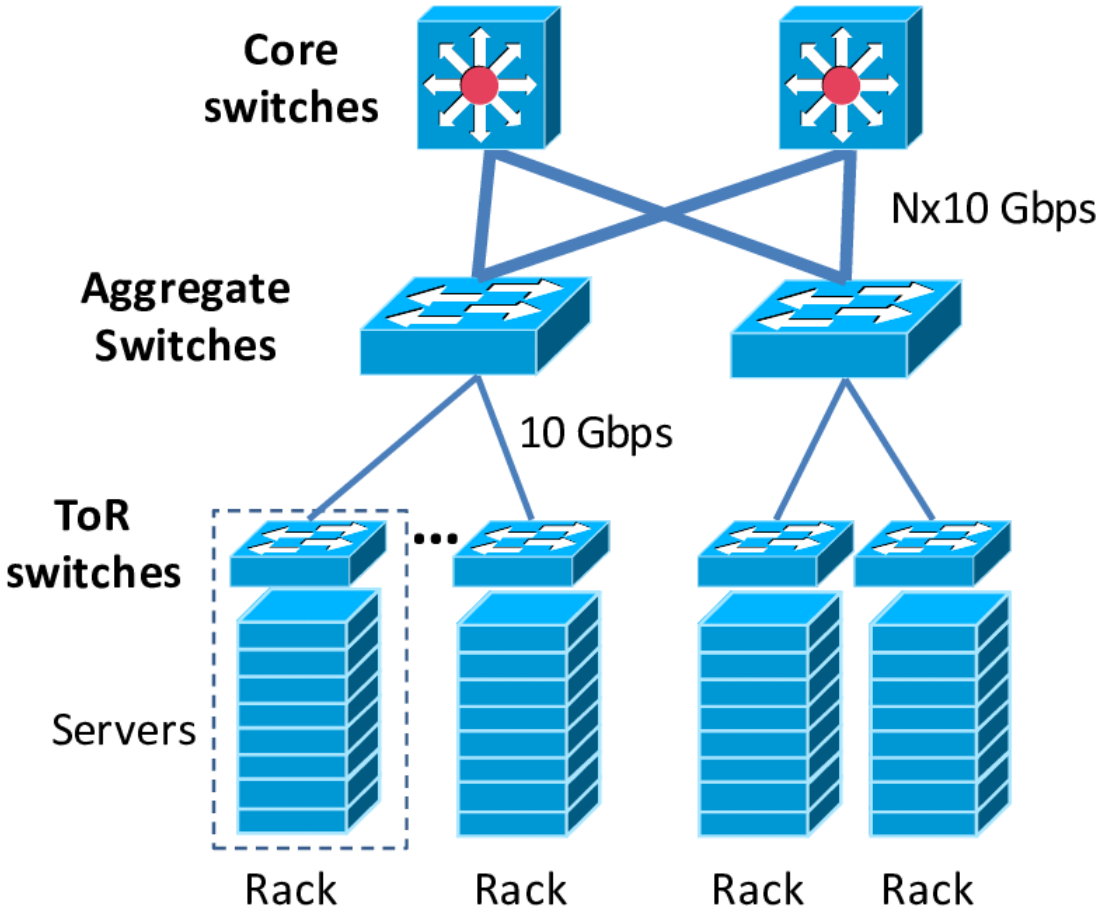
High performance network
Round Trip Delap < 2ms

AZ & Region MS Azure

https://docs.microsoft.com/en-us/azure/availability-zones/az-overview


Google Cloud Network
Region us-east-1
Region us-central-1
GCP Project
GCP Virtual Private Network - VPC 10.0.0.0/16
Availability Zone us-east-1a
Availability Zone us-east-1b
Availability Zone us-central-1a
Availability Zone us-central-1b
VPC Subnet A - 10.0.1.0/24
VPC Subnet b - 10.0.2.0/24
Instance 1
10.0.1.10
Instance 2
10.0.1.11
Instance 3
10.0.1.12
Instance 1
10.0.2.10
Instance 2
10.0.2.11
Instance 3
10.0.2.12

Networking inside a Datacenter



Ref:https://www.researchgate.net/figure/Typical-intra-data-center-network-architecture_fig1_258963909

Magic Quadrant for Cloud Infrastructure and Platform Services

- A sector report on Cloud Service Providers
- As of 2021 7 big vendors
  - Amazon Web Services
  - Google Cloud Platform
  - Microsoft Azure
  - Alibaba Cloud
  - Oracle
  - IBM
  - Tencent Cloud
- Last Magic Quadrant report published on 27 July 2021

SLA Agreement Example

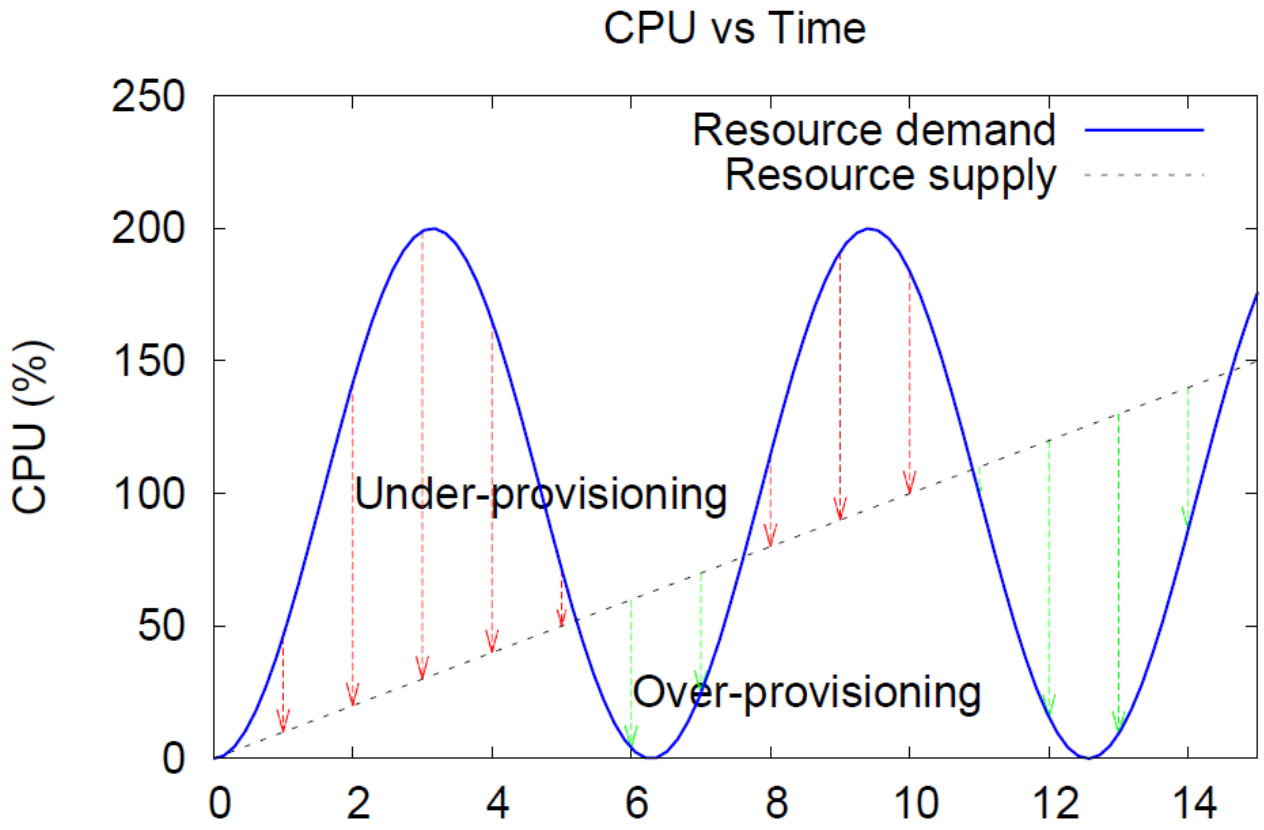
Compute Engine Service Level Agreement (SLA)

During the Term of the agreement under which Google has agreed to provide Google Cloud Platform to Customer (as applicable, the "Agreement"), the Covered Service will provide a Monthly Uptime Percentage to Customer as follows (the "Service Level Objective" or "SLO"):

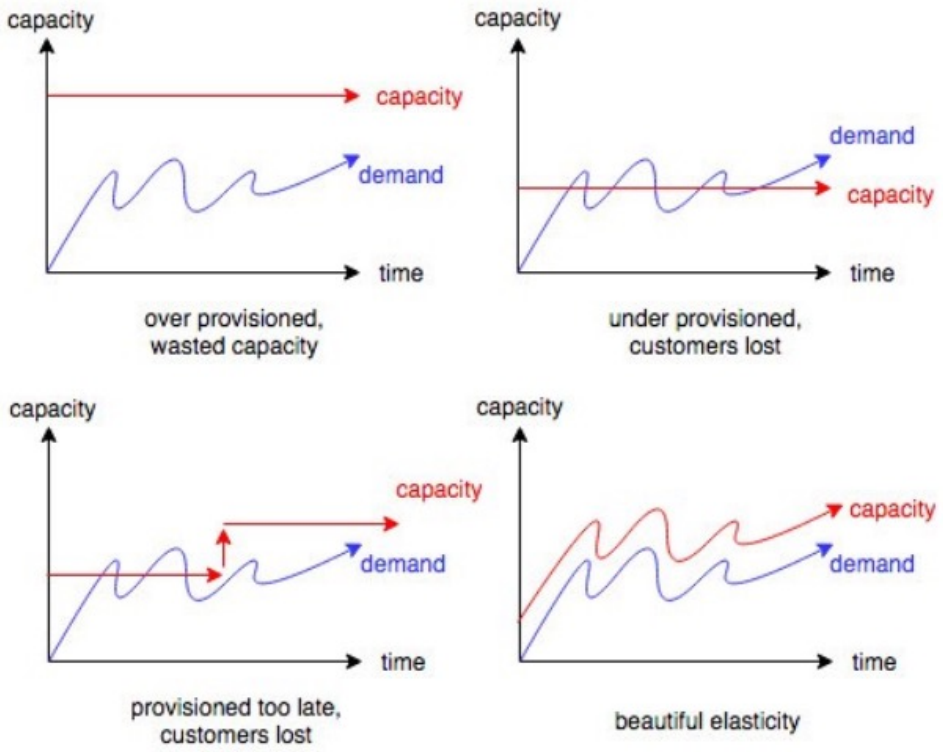
Covered Service	Monthly Uptime Percentage
Instances in Multiple Zones	>= 99.99%
A Single Instance	>= 99.5%
Load balancing	>= 99.99%

If Google does not meet the SLO, and if Customer meets its obligations under this SLA, Customer will be eligible to receive the Financial Credits described below. Monthly Uptime Percentage and Financial Credit are determined on a calendar month basis per Project or, for a Single Instance, per instance. This SLA states Customer's sole and exclusive remedy for any failure by Google to meet the SLO. Capitalized terms used in this SLA, but not defined in this SLA, have the meaning set forth in the Agreement. If the Agreement authorizes the resale or supply of Google Cloud Platform under a Google Cloud partner or reseller program, then all references to Customer in this SLA mean Partner or Reseller (as applicable), and any Financial Credit(s) will only apply for impacted Partner or Reseller order(s) under the Agreement.

Elasticity of Cloud Performance

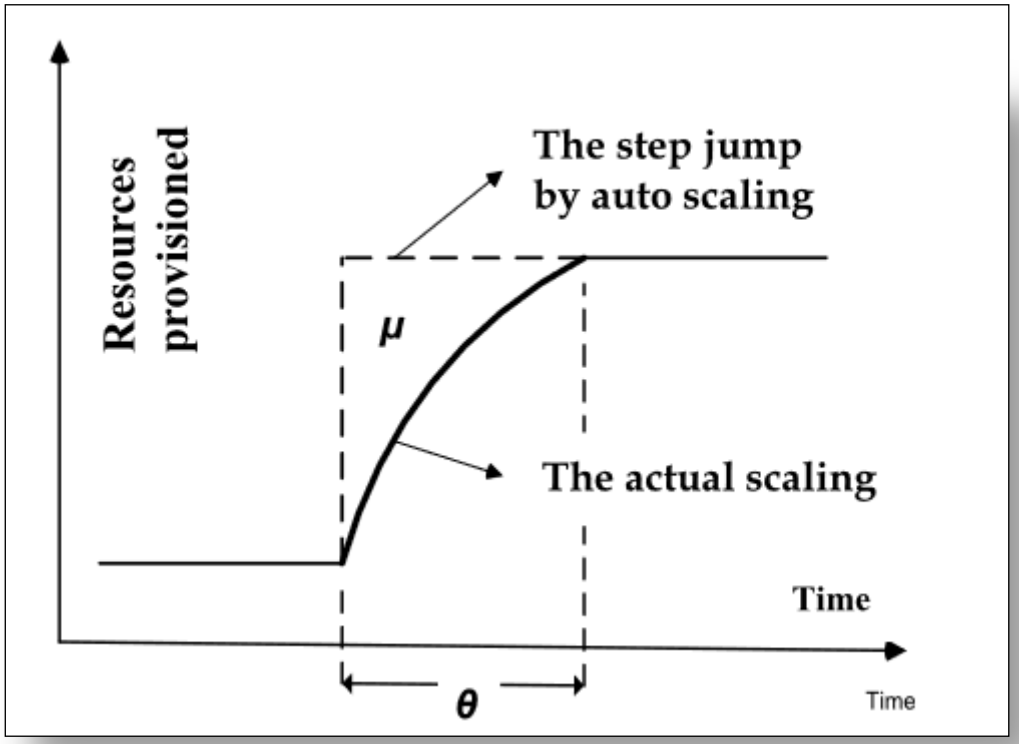


Elasticity vs Scalability



REF: https://pablo-iorio.medium.com/elasticity-does-not-equal-scalability-246bd9b3c128

Elasticity of Cloud Performance

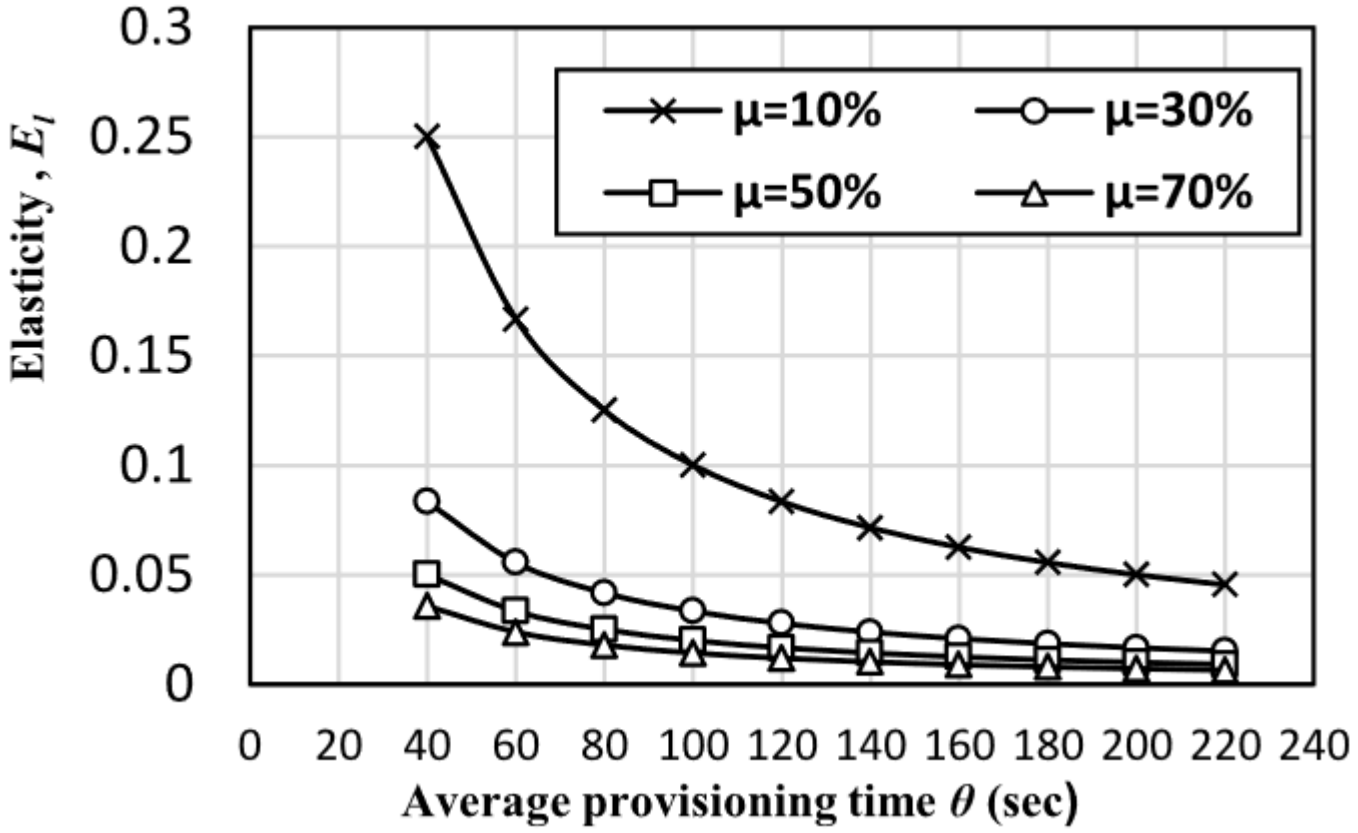


Elasticity defines the degree to which a system is able to adapt to workload changes by provisioning and deprovisioning resources in an autonomic manner, such that at each time the available resources match the current demand as closely as possible.

$$E_l = 1/(\theta \times \mu)$$

REF: Hwang, K., Bai, X., Shi, Y., Li, M., Chen, W. G., & Wu, Y. (2015). Cloud performance modeling with benchmark evaluation of elastic scaling strategies. *IEEE Transactions on parallel and distributed systems*, *27*(1), 130-143.

Elasticity of Cloud Performance



REF: Hwang, K., Bai, X., Shi, Y., Li, M., Chen, W. G., & Wu, Y. (2015). Cloud performance modeling with benchmark evaluation of elastic scaling strategies. *IEEE Transactions on parallel and distributed systems*, *27*(1), 130-143.