Cloud Computing Architecture

Introduction to High Availability

Image licensed under creative commons

Introduction to High Availability

This presentation:

– Introduction to:
  – High Availability
  – Scaling; and
  – Fault Tolerance

Images licensed under creative commons.

"Everything fails, all the time."

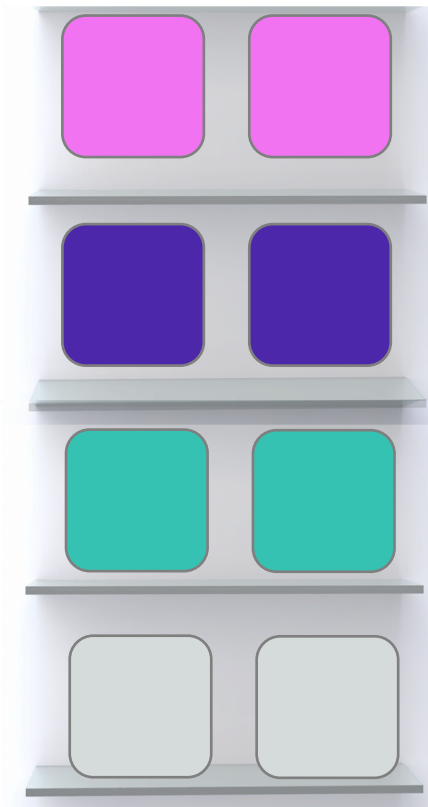Werner Vogels, CTO, Amazon.com

Highly available systems

- Can withstand some measure of degradation while remaining available
- Have minimized downtime
- Require minimal human intervention
- Recover from failure or roll over to secondary source in an acceptable amount of degraded performance time

Percentage of Uptime	Maximum Downtime Per Year	Equivalent Downtime Per Day
90%	36.5 days	2.4 hours
99%	3.65 days	14 minutes
99.9%	8.76 hours	86 seconds
99.99%	52.6 minutes	8.6 seconds
99.999%	5.25 minutes	0.86 seconds

Availability tiers

Availability	Max Disruption (per year)	Application Category
99%	3 days 15 hours	Batch processing, data extraction, transfer, and load jobs
99.9%	8 hours 45 minutes	Internal tools like knowledge management, project tracking
99.95%	4 hours 22 minutes	Online commerce, point of sale
99.99%	52 minutes	Video delivery, broadcast systems
99.999%	5 minutes	ATM transactions, telecommunications systems

Amazon S3 and Amazon S3 Glacier storage classes

Amazon S3 Intelligent Tiering

Automatically moves your objects between storage classes based on data access patterns.

S3 Standard:
Frequently accessed data

S3 Standard IA:
Long-lived, infrequently accessed data

S3 One Zone IA:
Long-lived, infrequently accessed, non-critical data

Amazon S3 Glacier or Deep Archive:
Archiving rarely accessed data

Amazon S3 Storage Class details

Factors that influence availability

Fault tolerance

- The built-in redundancy of an application's components and its ability to remain operational.

Scalability

- The ability of an application to accommodate increases in capacity needs without changing design.

Recoverability

- The process, policies, and procedures that are related to restoring service after a catastrophic event.
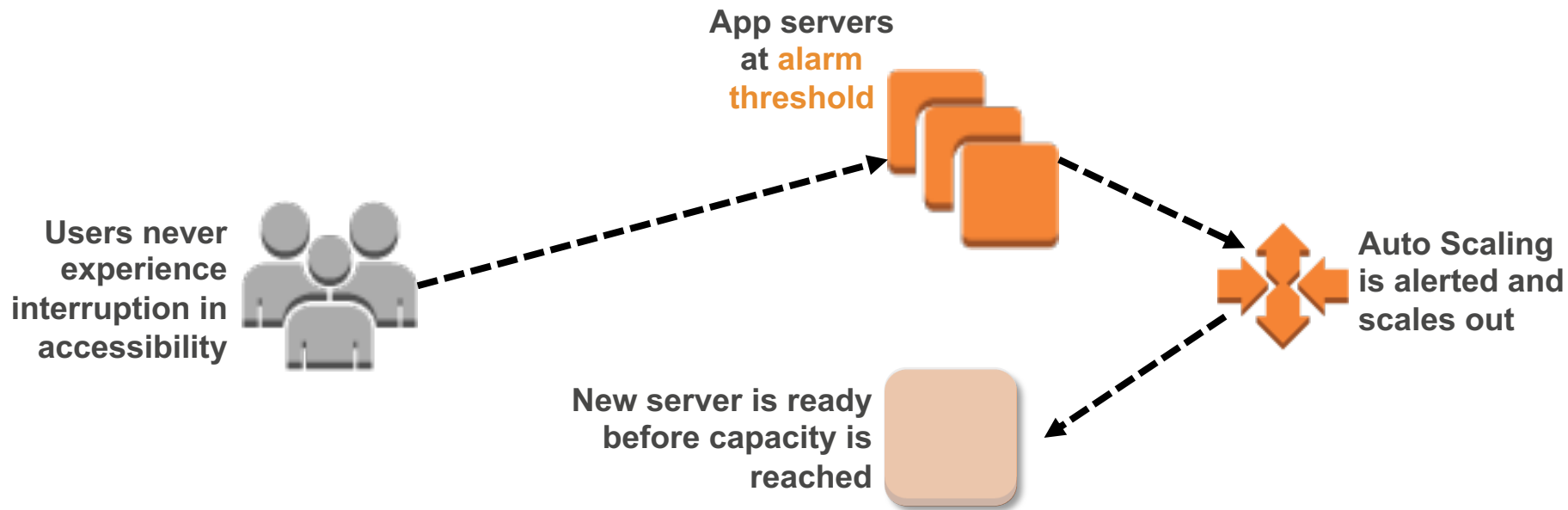
Services for Highly Available Web Application
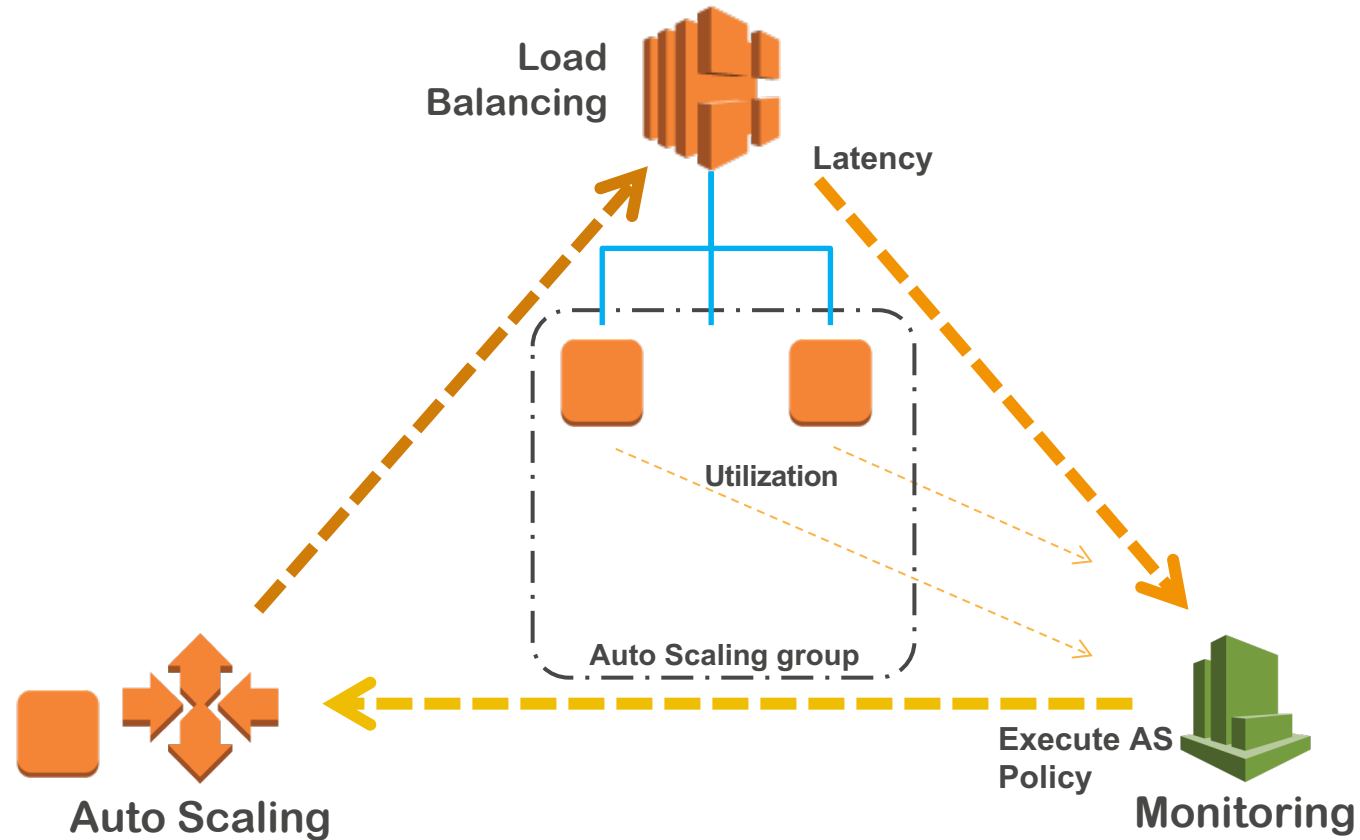
Best Practice: Enable Automated Scalability

Ensure that your architecture can handle changes in demand.

A key advantage of a cloud-based infrastructure is how quickly you can respond to changes in resource needs.

Users never experience interruption in accessibility

App servers at alarm threshold

Auto Scaling is alerted and scales out

New server is ready before capacity is reached

Best practice

amazon web services | Academy

Lecture References

Recommend Viewing

Swinburne Lecture – High Level Overview

AWS Academy – Deeper dive

ACF Module 9
ACA Module 9