Cloud Computing Architecture

Auto Scaling

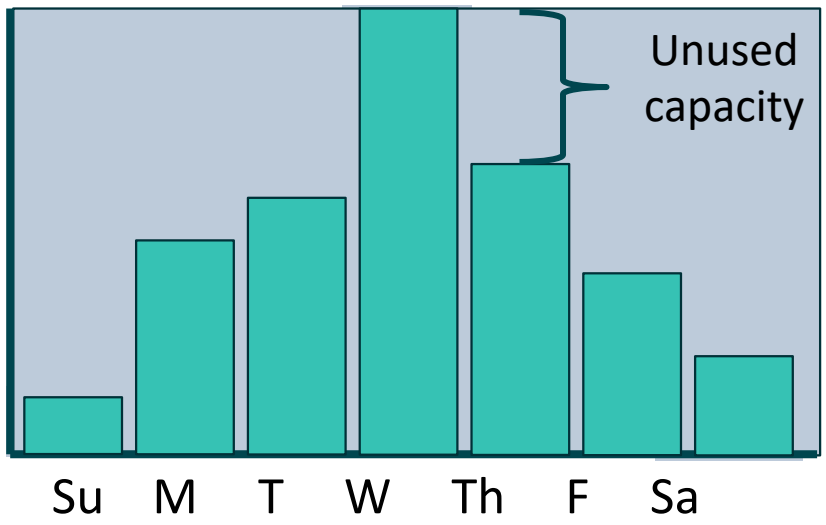
SWINBURNE UNIVERSITY OF TECHNOLOGY

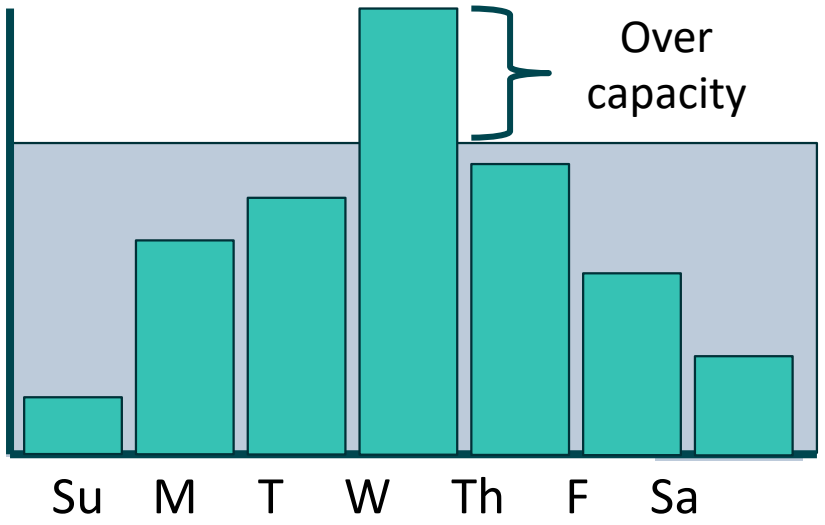
EC2 Autoscaling

This presentation:

- Scaling to Demand
  - Elastic Load Balancing
  - Cloud Watch
  - Auto-scaling

Why is scaling important?

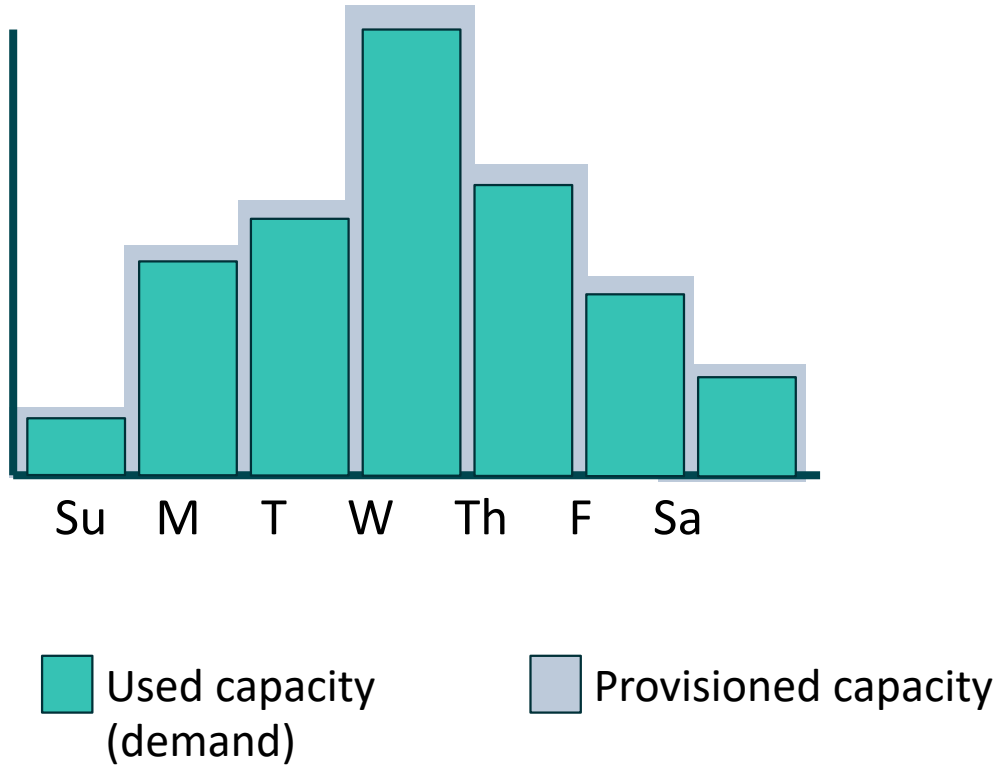




Used capacity (demand)

Provisioned capacity

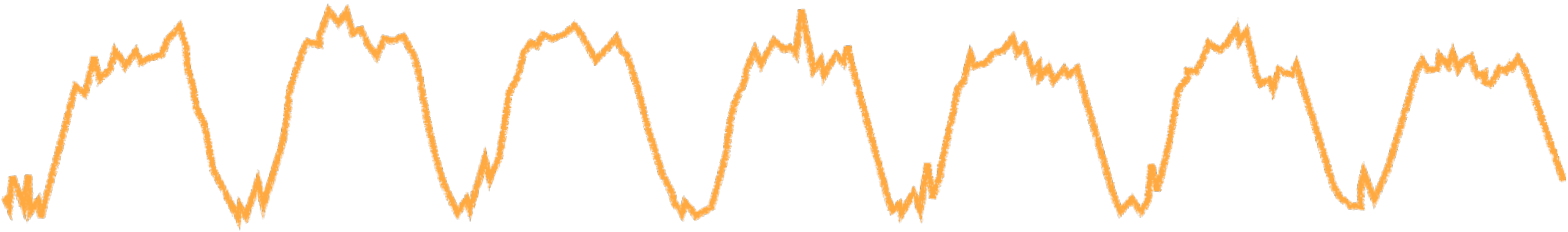
Amazon EC2 Auto Scaling



- Helps you maintain application availability
- Enables you to automatically add or remove EC2 instances according to conditions that you define
- Detects impaired EC2 instances and unhealthy applications, and replaces the instances without your intervention
- Provides several scaling options – Manual, scheduled, dynamic or on-demand, and predictive

Typical weekly traffic at Amazon.com

Provisioned capacity



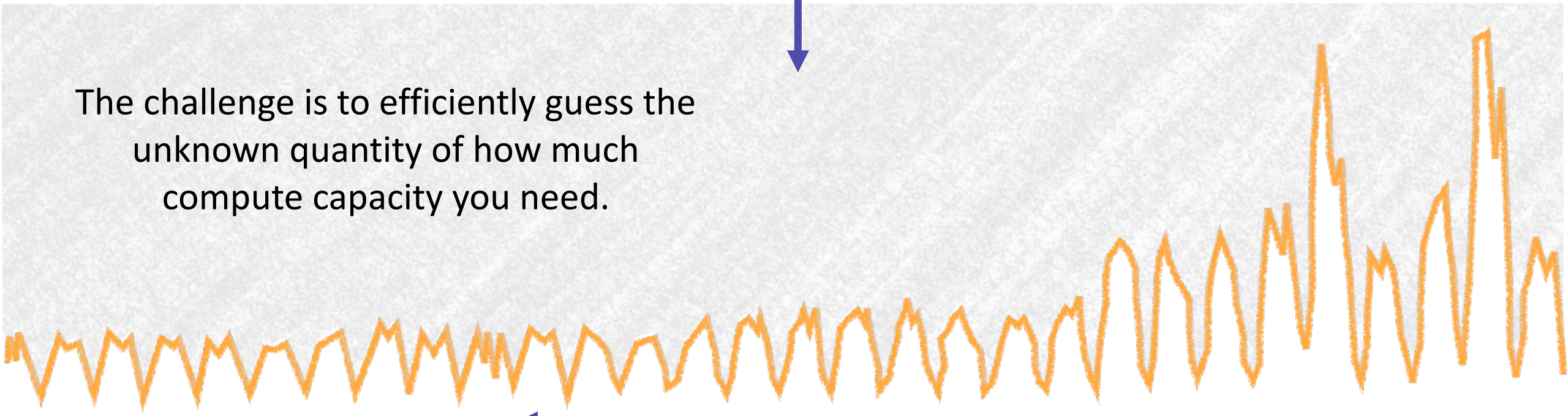
Sunday Monday Tuesday Wednesday Thursday Friday Saturday

November traffic to Amazon.com

Provisioned capacity

76 percent

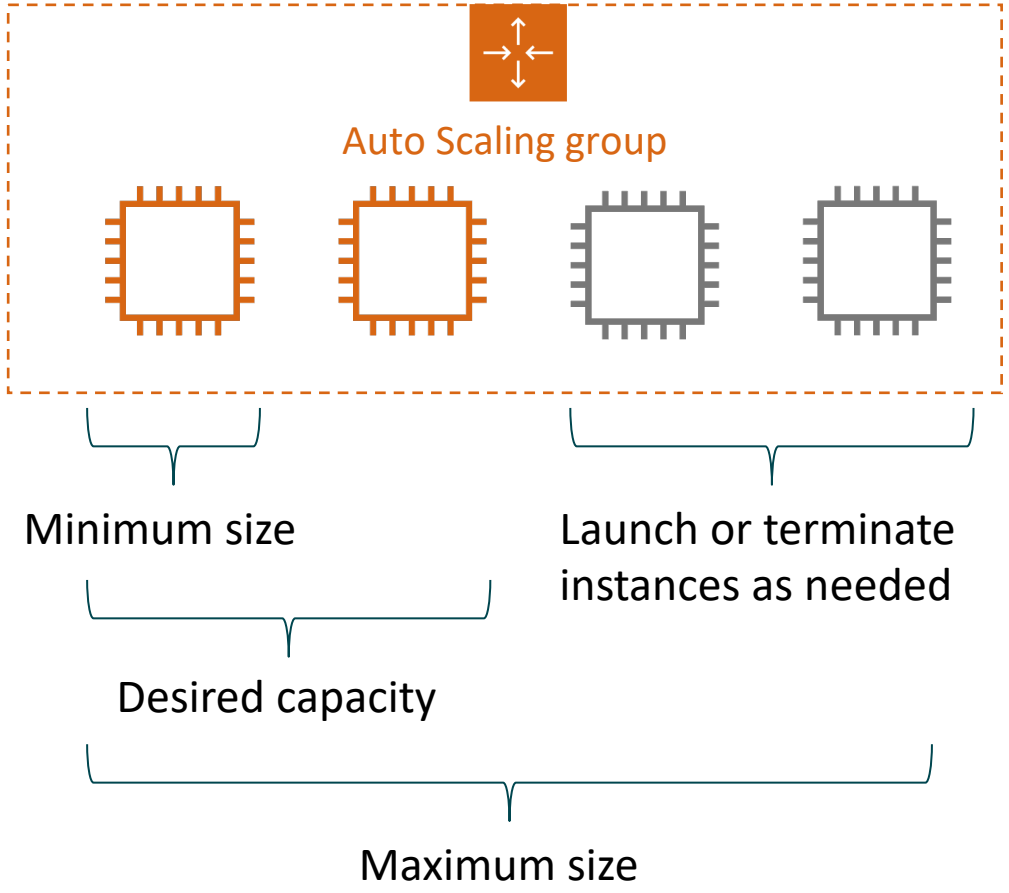
The challenge is to efficiently guess the unknown quantity of how much compute capacity you need.



24 percent

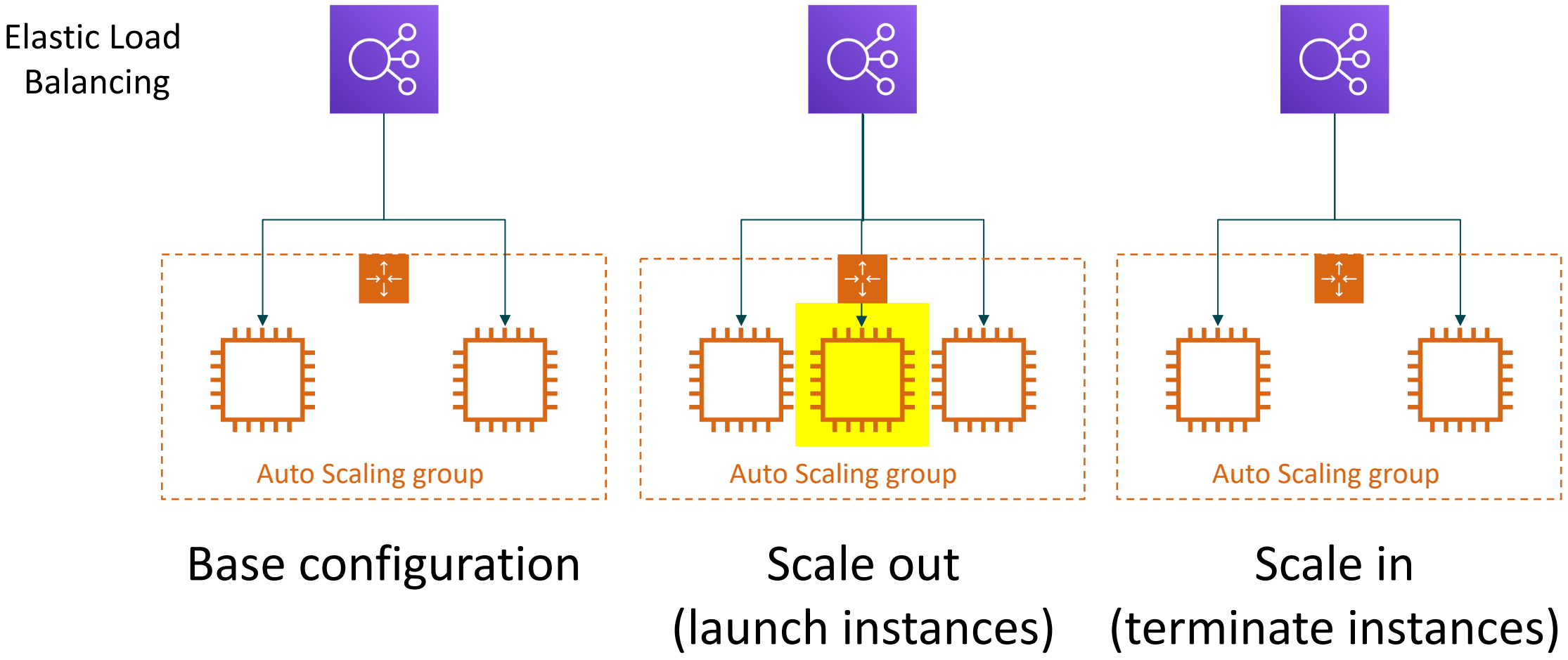
November

Auto Scaling groups



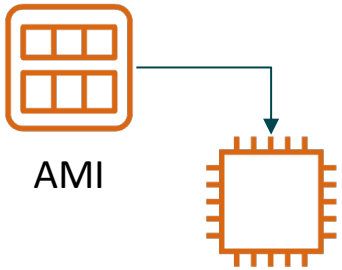
An **Auto Scaling group** is a collection of EC2 instances that are treated as a logical grouping for the purposes of automatic scaling and management.

Scaling out versus scaling in



How Amazon EC2 Auto Scaling works

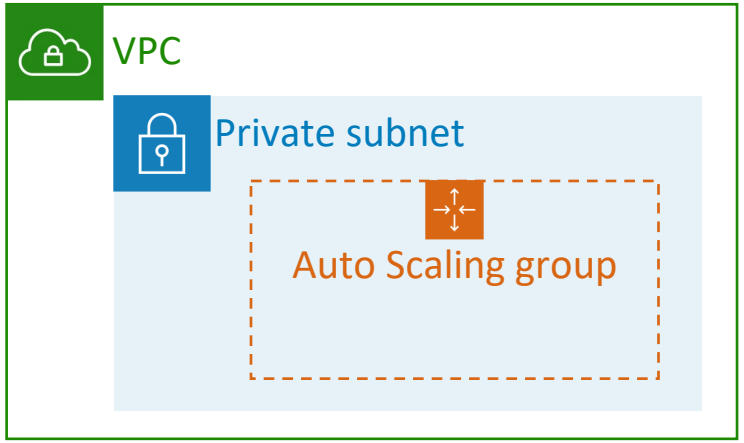
What



Launch configuration

- AMI
- Instance type
- IAM role
- Security groups
- EBS volumes

Where



Auto Scaling group

- VPC and subnets
- Load balancer

When

Maintain current number

- Health checks

Manual scaling

- Min, max, desired capacity

Scheduled scaling

- Scheduled actions

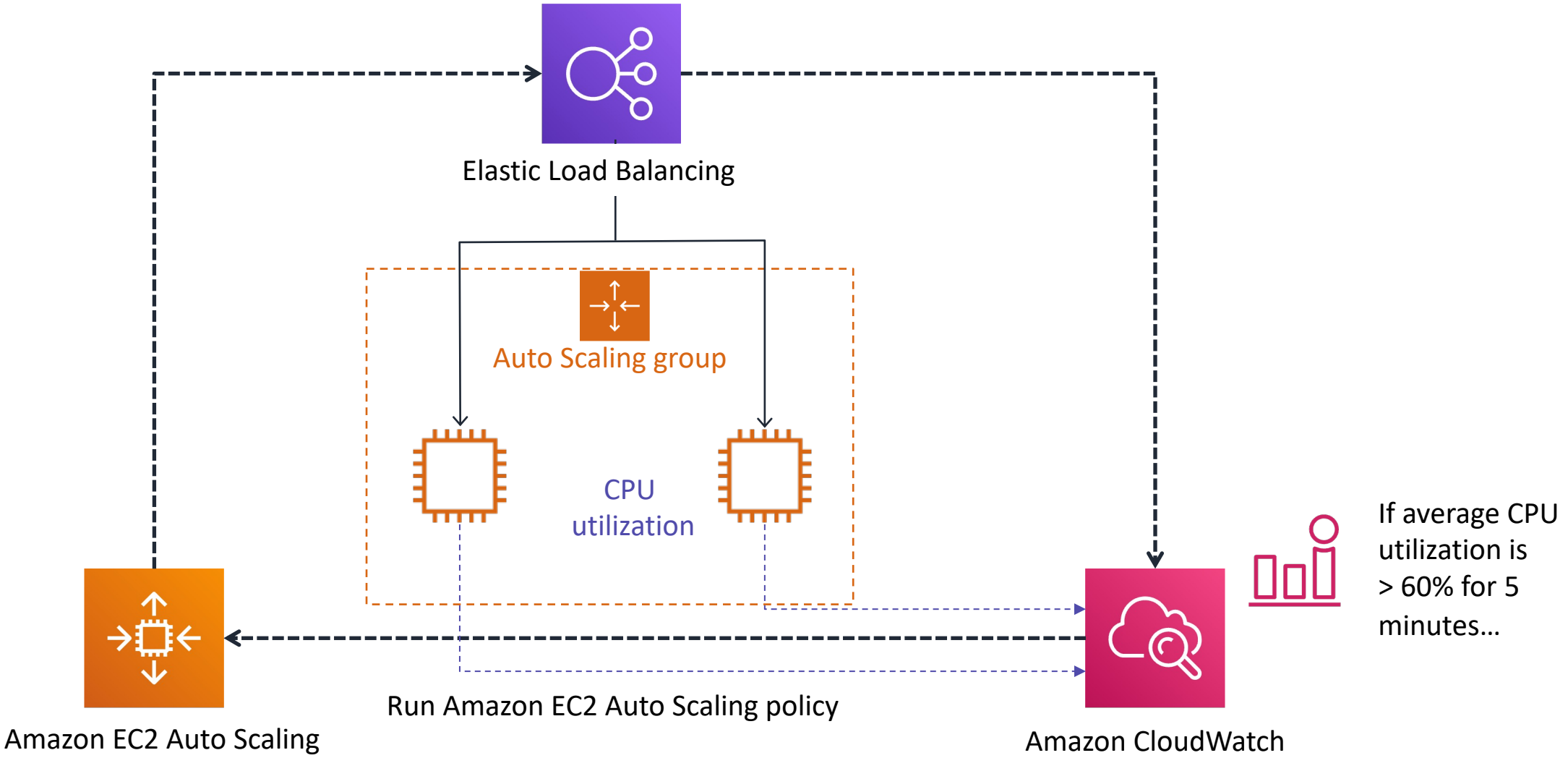
Dynamic scaling

- Scaling policies

Predictive scaling

- AWS Auto Scaling

Implementing dynamic scaling



AWS Auto Scaling

AWS Auto Scaling

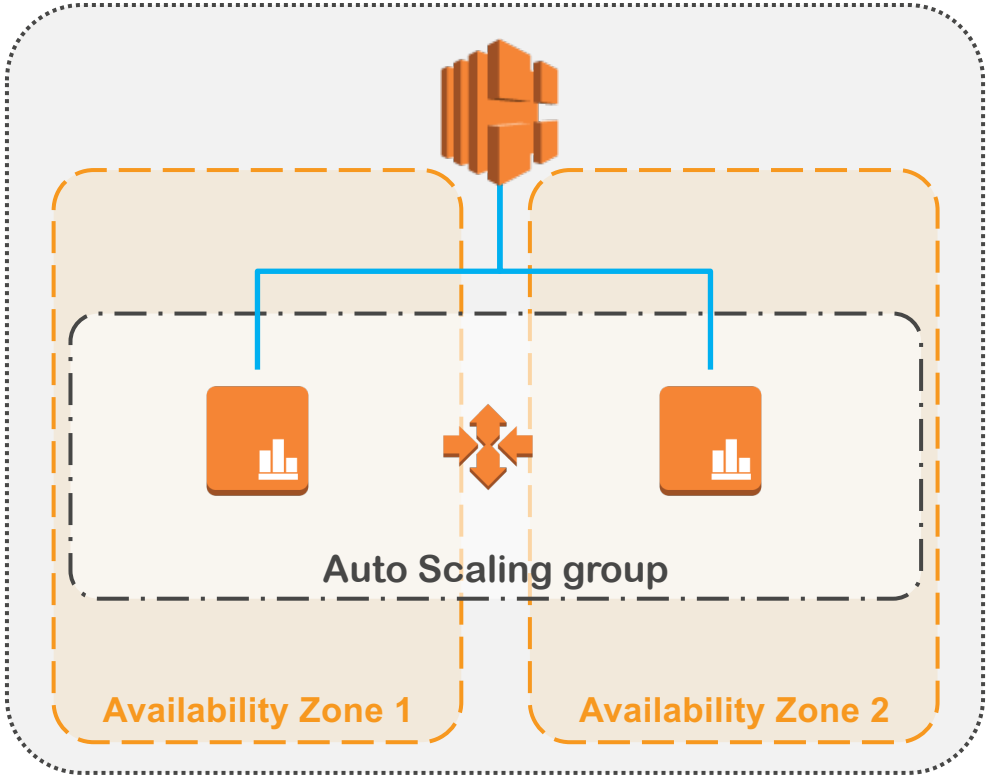
- Monitors your applications and automatically adjusts capacity to maintain steady, predictable performance at the lowest possible cost
- Provides a simple, powerful user interface that enables you to build scaling plans for resources, including –
  - Amazon EC2 instances and Spot Fleets
  - Amazon Elastic Container Service (Amazon ECS) Tasks
  - Amazon DynamoDB tables and indexes
  - Amazon Aurora Replicas

How Do You Decide On Minimum Capacity Size?

- Auto Scaling group defines:
  - Desired capacity
  - Minimum capacity
  - Maximum capacity
- What would be a good **minimum** capacity to set it to?
- What would be a good **maximum** capacity to set it to?

How Do You Decide On Minimum Capacity Size?

What about high availability?



Minimum = two instances (# of AZs)

Desired capacity = two instances (Min.)

Maximum Capacity Size And Auto Scaling

Scenario:

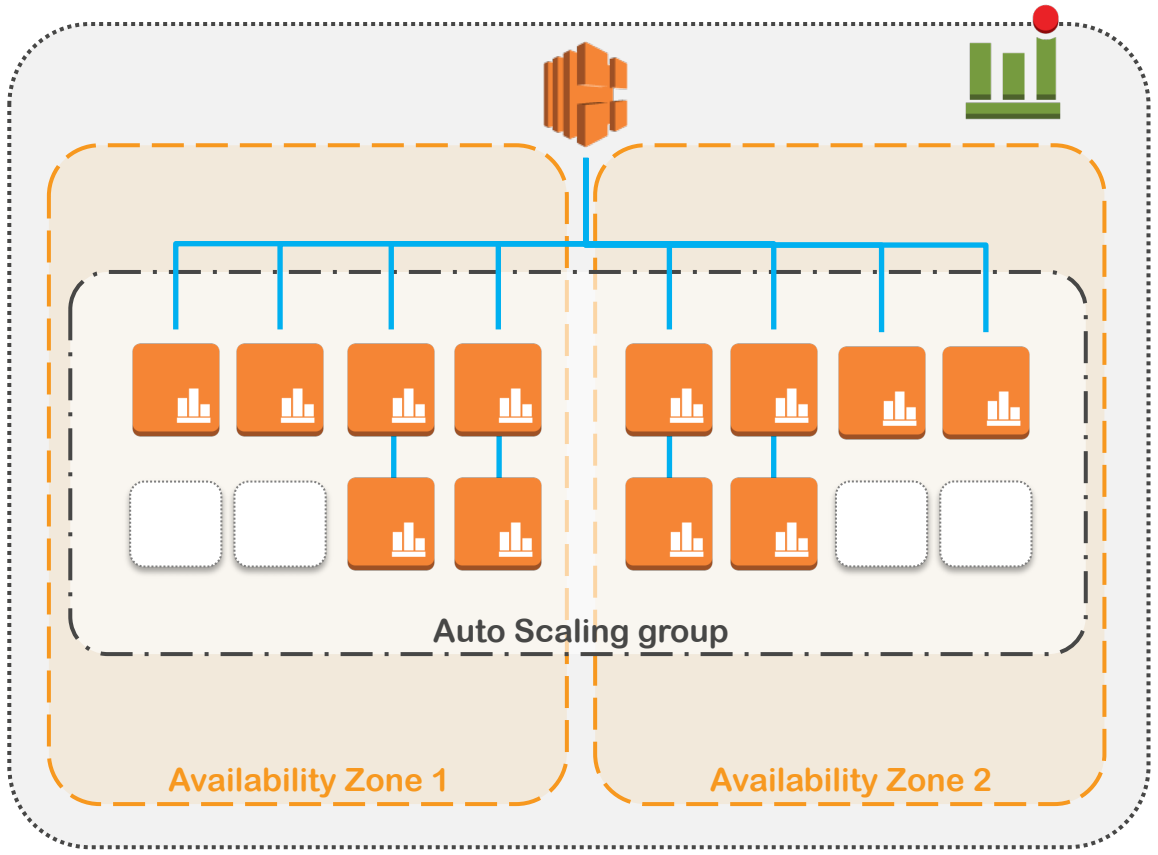
Auto Scaling group

- Minimum = 2
- Maximum = **12**

Auto Scaling policy

- When CPU utilization is greater than 60%
- Add 100% of group = **double the capacity**

CPU utilization triggers the alarm: capacity is doubled until CPU utilization drops below 60% or max capacity is reached.



Section 3 key takeaways

- Scaling enables you to respond quickly to changes in resource needs.
- Amazon EC2 Auto Scaling maintains application availability by automatically adding or removing EC2 instances.
- An Auto Scaling group is a collection of EC2 instances.
- A launch configuration is an instance configuration template.
- Dynamic scaling uses Amazon EC2 Auto Scaling, CloudWatch, and Elastic Load Balancing.
- AWS Auto Scaling is a separate service from Amazon EC2 Auto Scaling.

Lecture References

Recommend Viewing

Swinburne Lecture – High Level Overview

AWS Academy – Deeper dive

ACF Module 10