Designing for Scalability on AWS

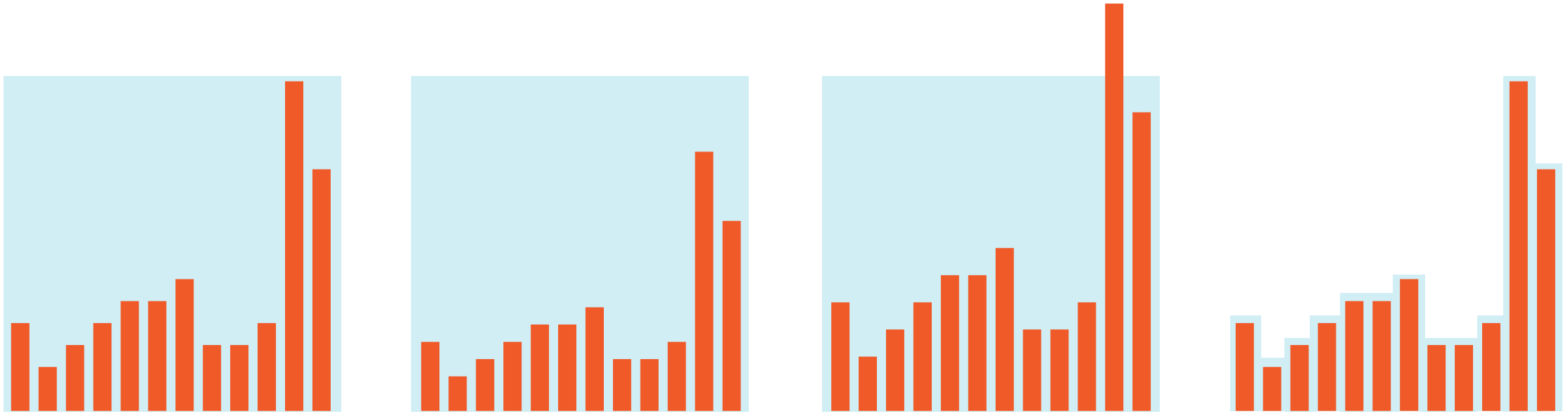
Danny Jessee
AWS CERTIFIED DEVOPS ENGINEER - PROFESSIONAL

@dannyjessee

Overview

- Architecting scalable solutions
- Launch Configurations and Launch Templates
- Create a new Auto Scaling Group
- Auto Scaling policies
- Auto Scaling lifecycle hooks
- Scaling on a schedule
- Self-healing nature of Auto Scaling Groups

Capacity Planning



Expectation

Too much capacity

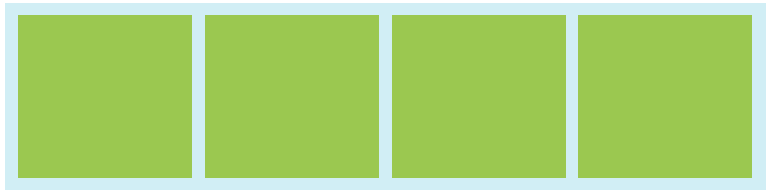
Not enough capacity

Auto scaling

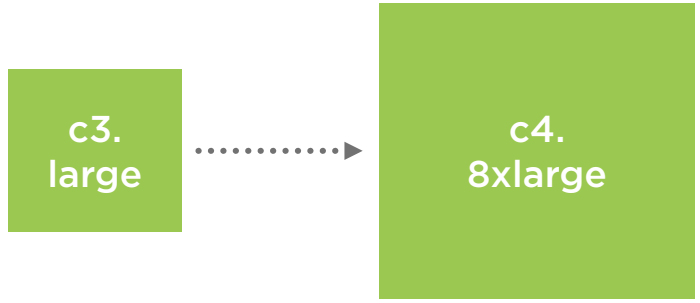
Demand

Available capacity

Horizontal and Vertical Scaling



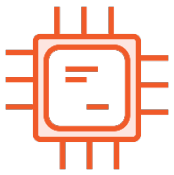
Horizontal scaling
"Scaling out"



Vertical scaling
"Scaling up"

Launch Configurations

Can be defined from scratch

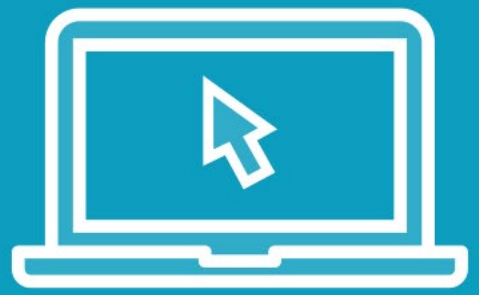


Can be created from a running EC2 instance

Can be reused across multiple Auto Scaling Groups

Are immutable

Demo



Create a new Launch Configuration

Spot instances and their benefits

Launch Templates

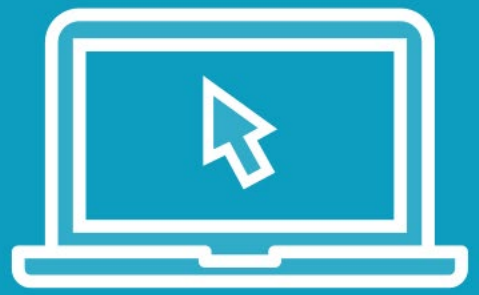
Recommended by AWS over Launch Configurations

Can be changed and have multiple versions

Allow for more advanced configuration options

Allow for mixing instance types and classes

Demo

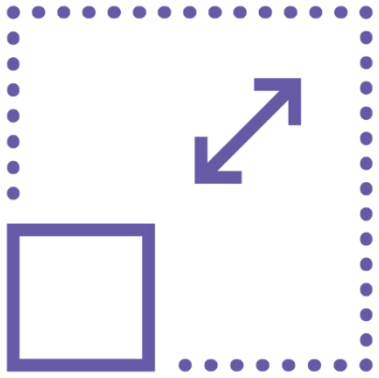


Create a new Launch Template

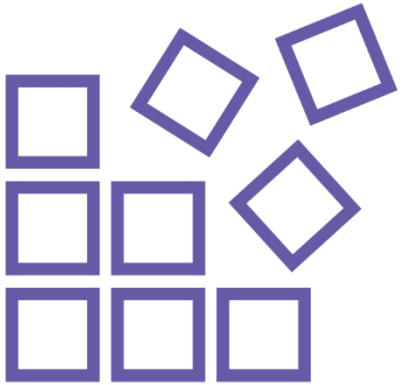
Benefits of using Launch Templates

Create a Launch Template from a Launch Configuration

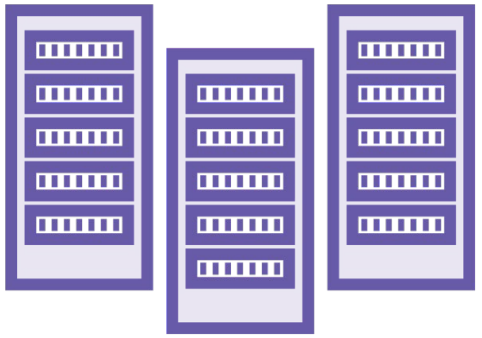
Auto Scaling



Right-sizing capacity needs



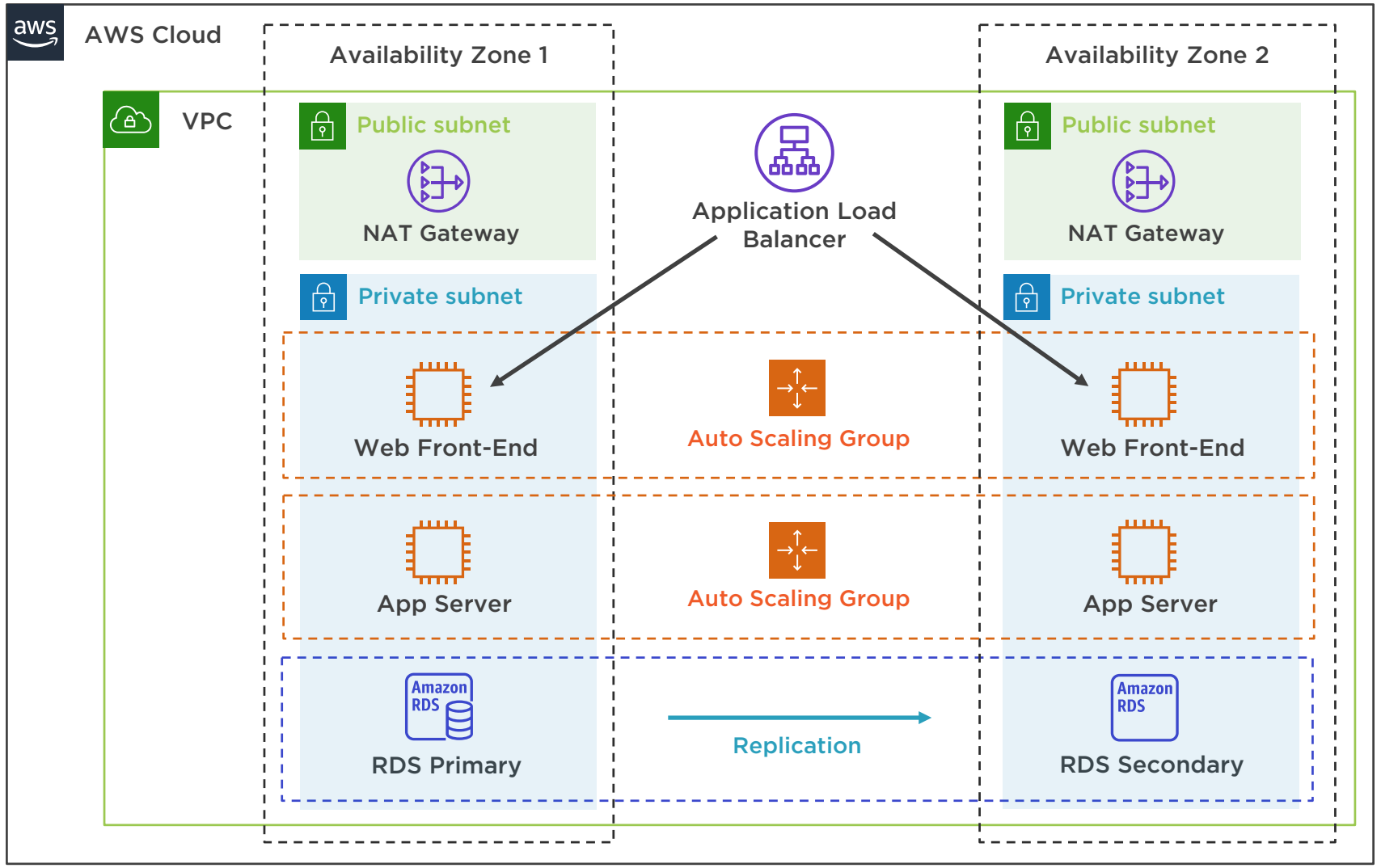
Logical collection of instances



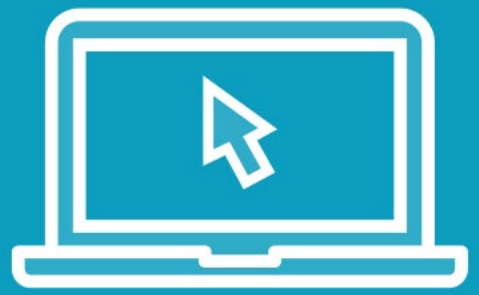
Spread across Availability Zones

Performs health checks on instances

Elastic Load Balancing and Auto Scaling



Demo



Create a new Auto Scaling Group

Associate an Auto Scaling Group with an Application Load Balancer

Auto Scaling Policies

Manually adjust minimum, maximum, or desired capacity values

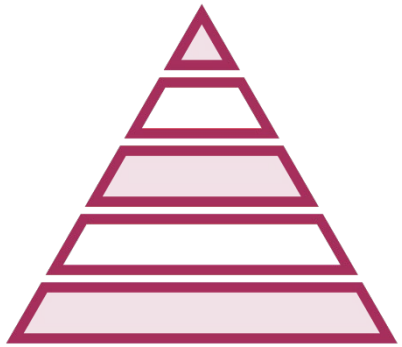
Automatically adjust based on metric values

Target Tracking

Auto Scaling will maintain target value

Simple

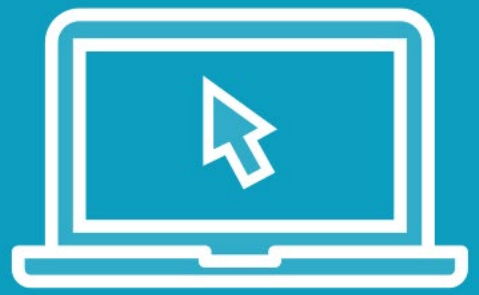
Auto Scaling acts based on CloudWatch Alarms



Step Scaling

Auto Scaling acts based on banded value ranges

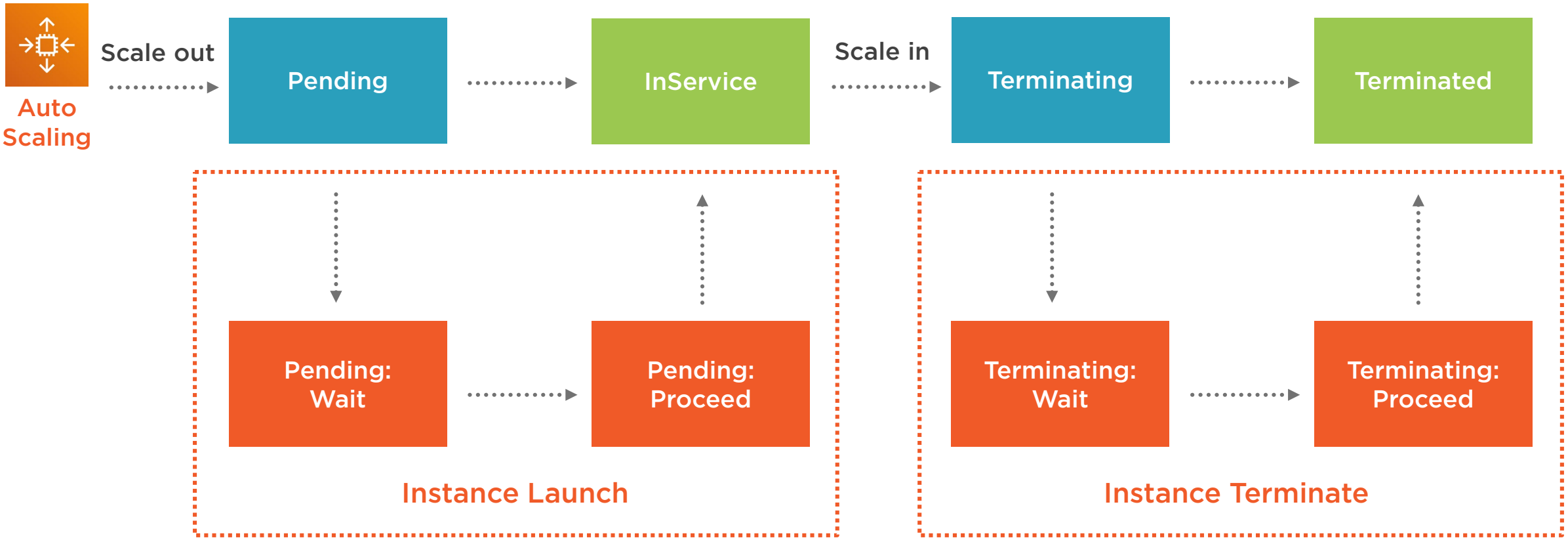
Demo



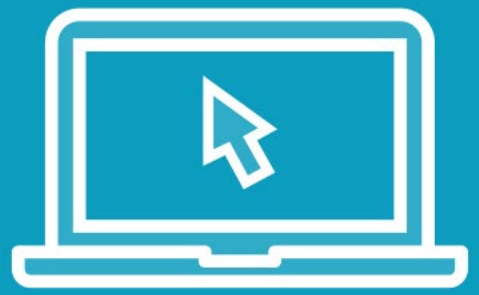
Configure scaling policies for an Auto Scaling Group

Target Tracking, Simple, and Step Scaling policies

Lifecycle Hooks



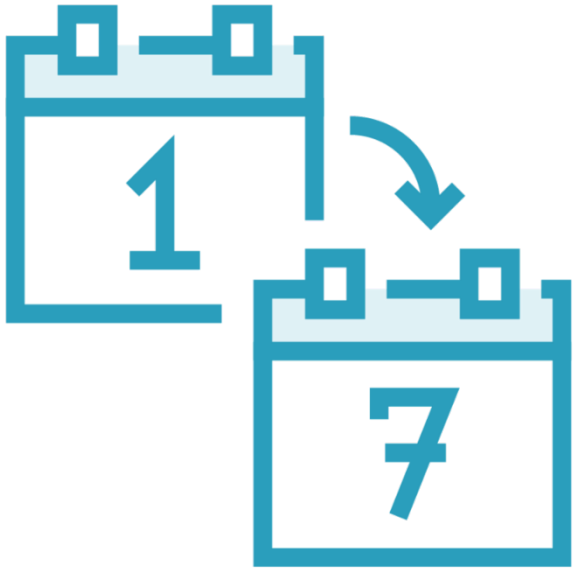
Demo



Create a lifecycle hook

CloudWatch Events rule and targets

Dynamically Scaling on a Schedule



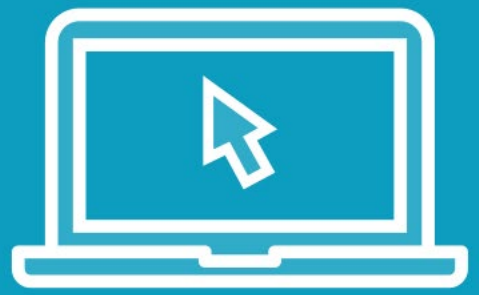
Set a fixed value for minimum, maximum, or desired capacity

Can be one-time or recurring

Cron expressions for recurrence pattern

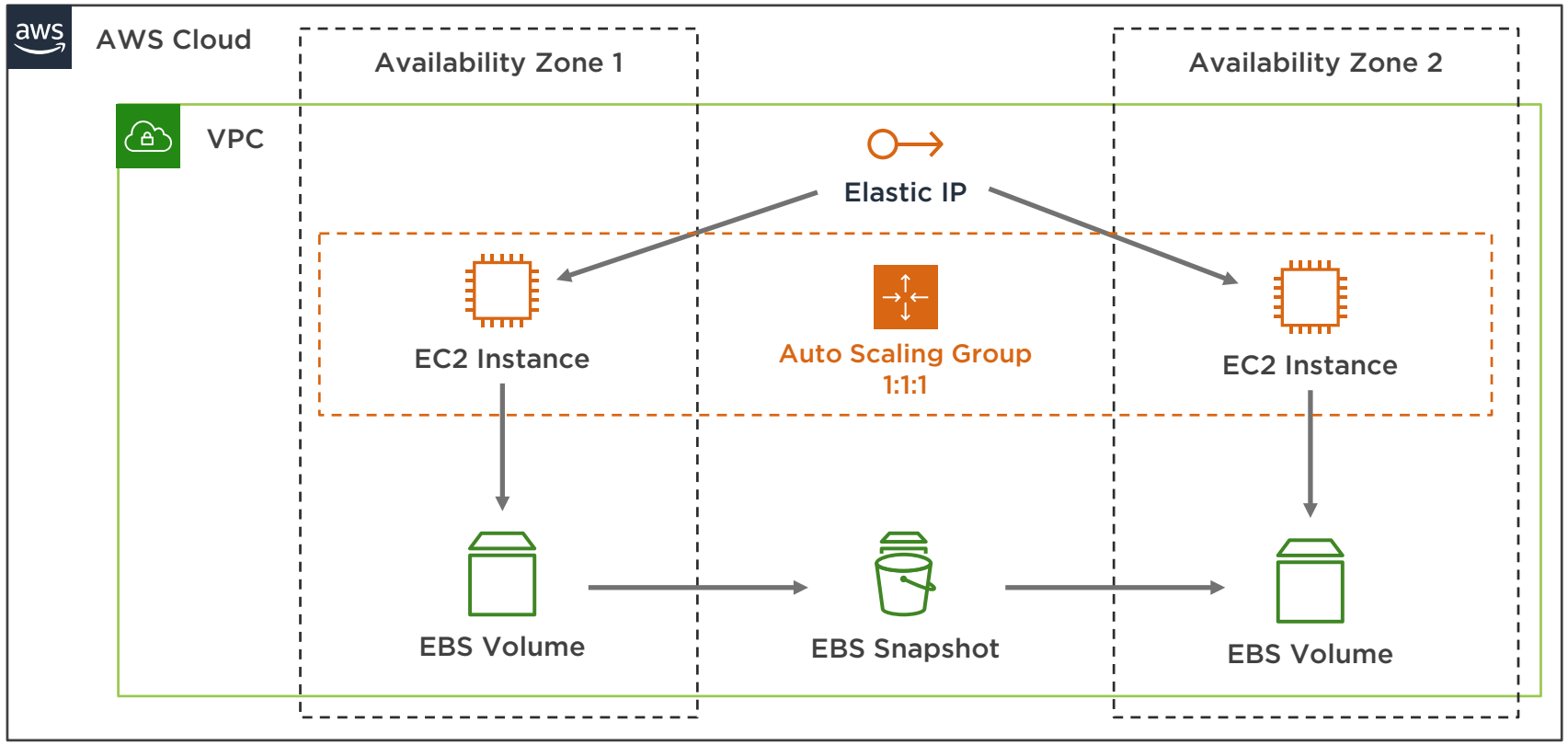
Start and end date

Demo



Create a Scheduled Action

Self-Healing Instance with Auto Scaling



Review

- AWS Auto Scaling
- Launch Configurations and Launch Templates
- Auto Scaling policies
- Auto Scaling lifecycle hooks
- Scheduled actions
- Self-healing with Auto Scaling Groups

Up Next:
Designing Highly Available and Scalable Databases