Module 10 - Automatic Scaling and Monitoring

Slides

Objectives / Topics

- Indicate how to distribute traffic across Amazon Elastic Compute Cloud (Amazon EC2) instances by using Elastic Load Balancing
- Identify how Amazon CloudWatch enables you to monitor AWS resources and applications in real time
- Explain how Amazon EC2 Auto Scaling launches and releases servers in response to workload changes
- Perform scaling and load balancing tasks to improve an architecture

Labs / Activities

- Knowledge Check
- Lab: Scale and Load Balance Your Architecture --- Lab Instructions

Section 1: Elastic Load Balancing

Elastic Load Balancing distributes incoming application or network traffic across multiple targets in a single Availability Zone or across multiple Availability Zones. It also scales your load balancer as traffic to your application changes over time. Monitoring is done via Amazon CloudWatch, access logs, and AWS CloudTrail logs

Types of Load Balancers

Application Load Balancer	Network Load Balancer	Classic Load Balancer (Previous Generation)
Load balancing of HTTP and HTTPS traffic	Load balancing of TCP, UDP, and TLS traffic where extreme performance is required	Load balancing of HTTP, HTTPS, TCP, and SSL traffic
- Routes traffic to targets based on content of request - Provides advanced request routing to targeted at the delivery of modern application architectures, including microservices and containers	- Routes traffic to targets based on IP protocol data - Can handle millions of requests per second while maintaining ultra-low latencies - Optimized to handle sudden and volatile traffic patterns	Load balancing across multiple EC2 instances
Operates at the application layer (OSI model layer 7)	Operate at the transport layer (OSI model layer 4)	Operates at both the application and transport layers

- With Application Load Balancers and Network Load Balancers, you register targets in target groups, and route traffic to the target groups.
- With Classic Load Balancers, you register instances with the load balancer.

Section 2: Amazon CloudWatch

Monitors: AWS resources and applications that run on AWS

Collects and tracks: Standard and custom metrics

Alarms: Send notifications to an Amazon SNS topic and perform Amazon EC2 Auto Scaling or Amazon EC2 actions

- Create alarms based on:
- Static threshold
- Anomaly detection
- Metric math expression
- Specify:
- Name space
- Metric

- Statistic
- Period
- Conditions
- Additional configuration
- Actions

Events: Define rules to match changes in AWS environment and route these events to one or more target functions or streams for processing

Section 3: Amazon EC2 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
- 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
- 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.
- Scale out (launch instances), Scale in (terminate instances)

Lab: Scale and Load Balance Your Architecture --- Lab Instructions

Knowledge Check

AWS Lambda Functions and Autoscaling Video --- Walkthrough Instructions