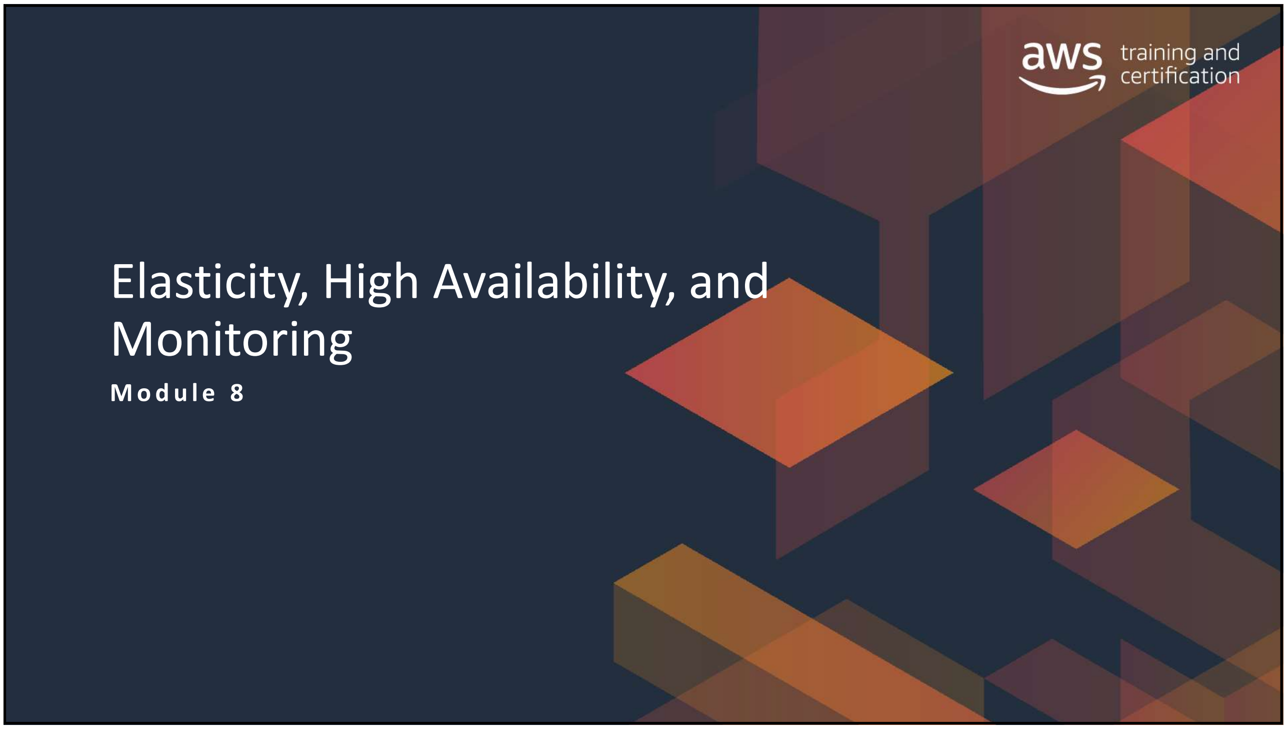
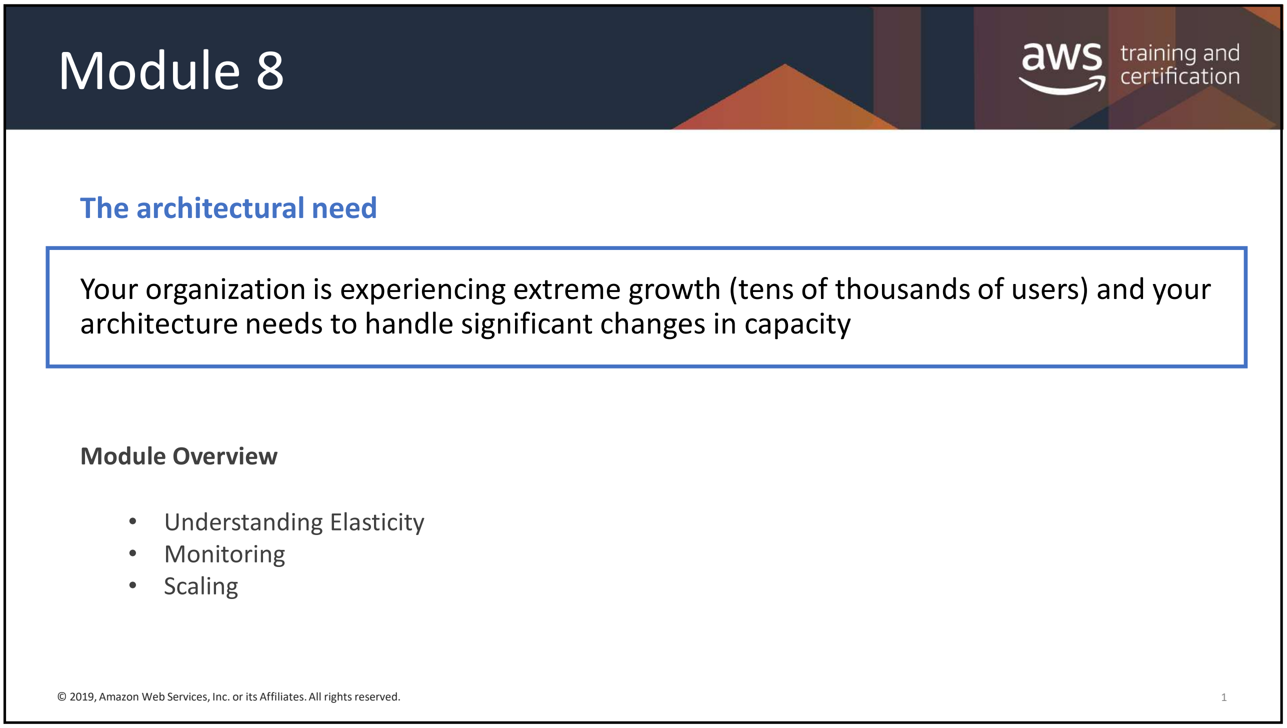


Elasticity, High Availability, and Monitoring

Module 8

0



Module 8

The architectural need

Your organization is experiencing extreme growth (tens of thousands of users) and your architecture needs to handle significant changes in capacity

Module Overview

- Understanding Elasticity
- Monitoring
- Scaling

1

High Availability Factors

aws training and certification

Fault tolerance:

The **built-in redundancy** of an application's components

Scalability:

The ability of an application to **accommodate growth** without changing design

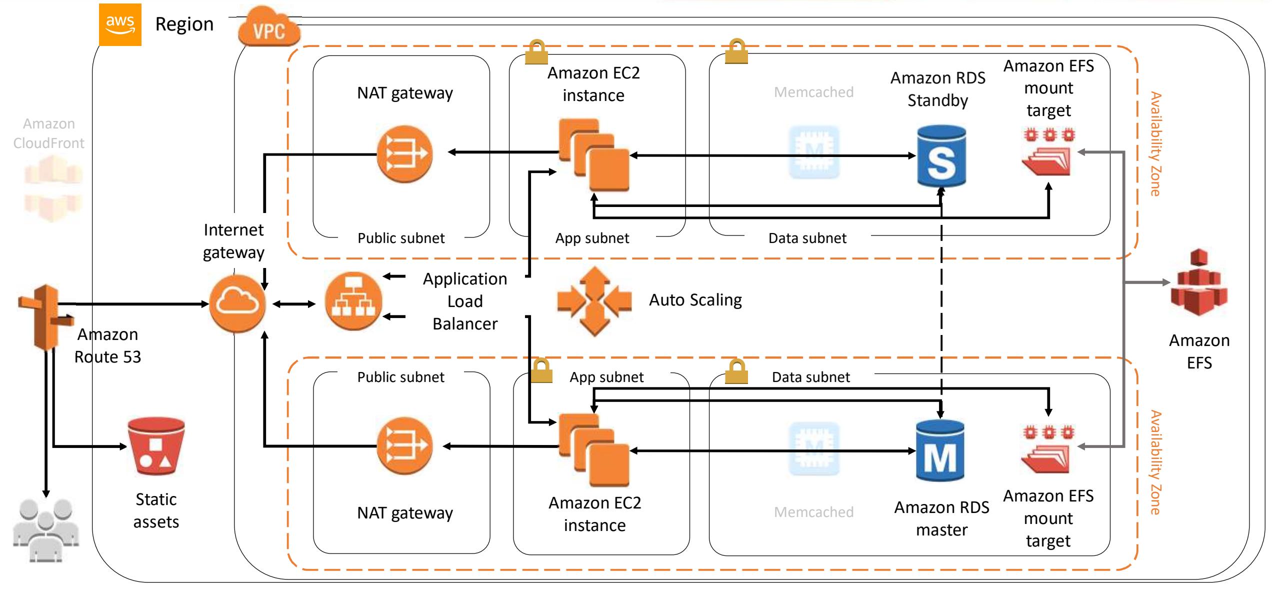
Recoverability:

The process, policies, and procedures related to **restoring service** after a catastrophic event

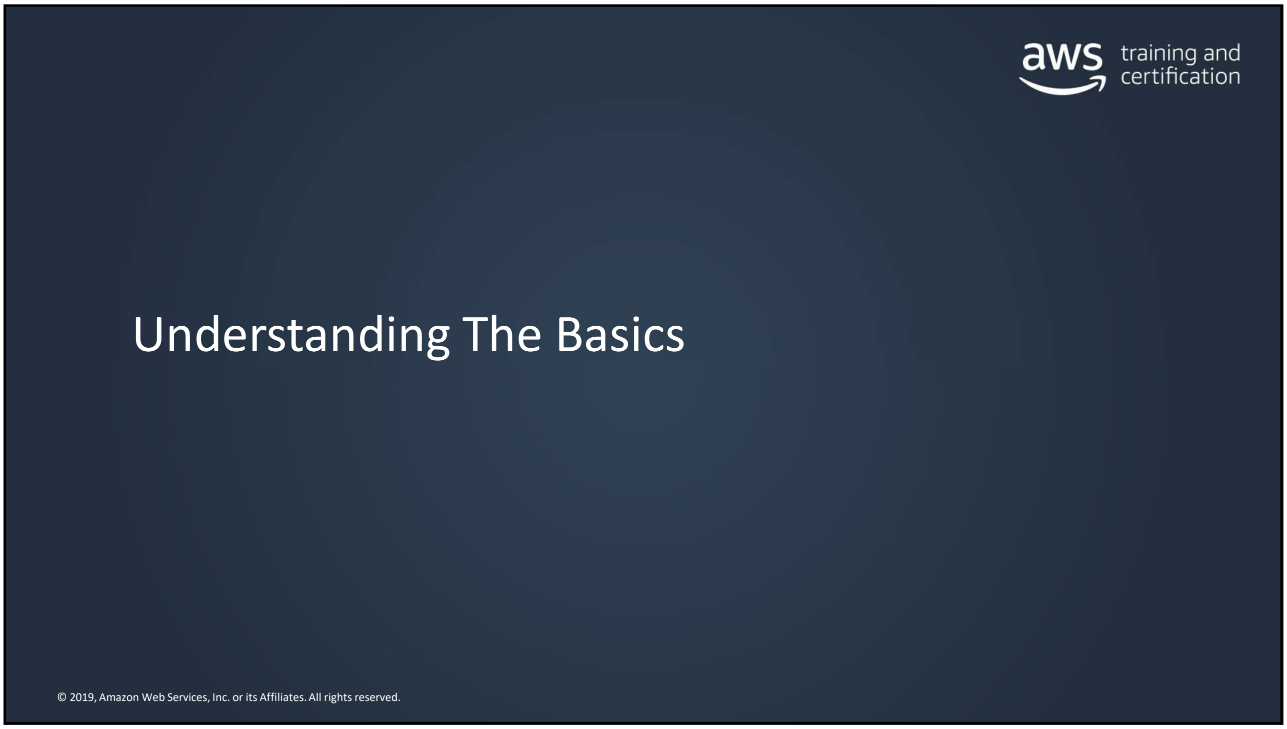
2

Scaling

aws training and certification



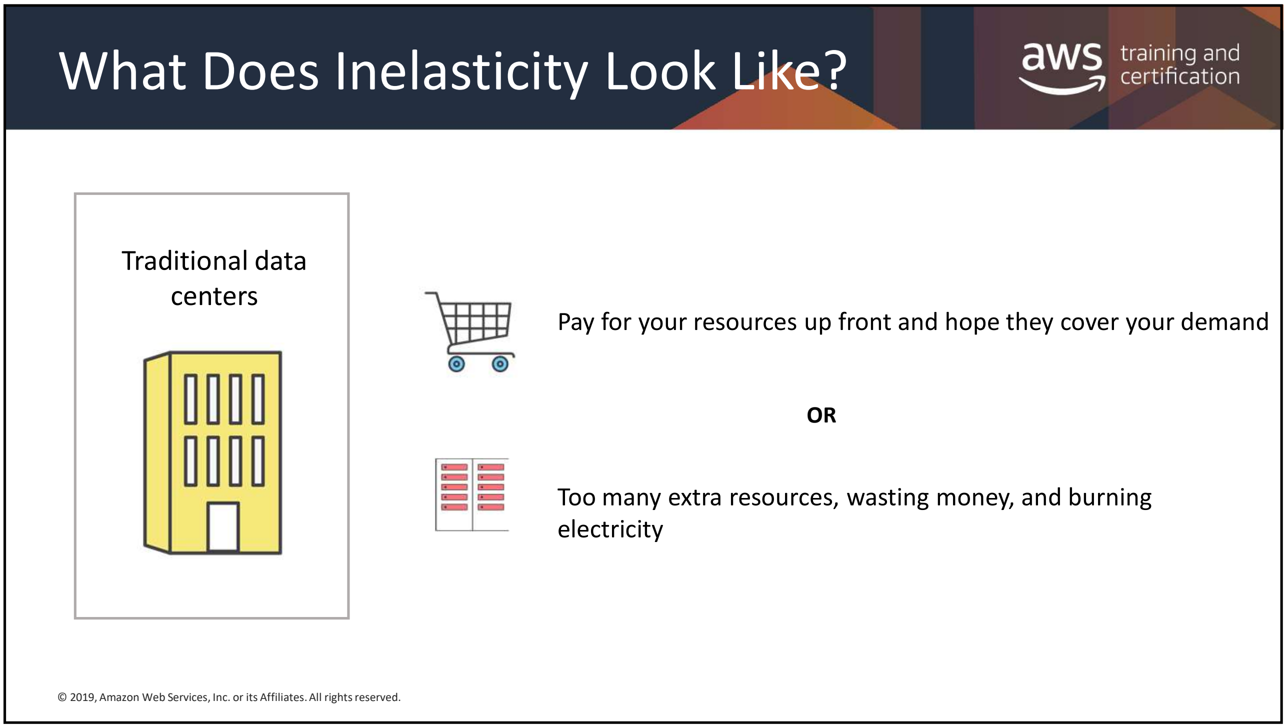
3



Understanding The Basics

aws training and certification

4



What Does Inelasticity Look Like?

aws training and certification

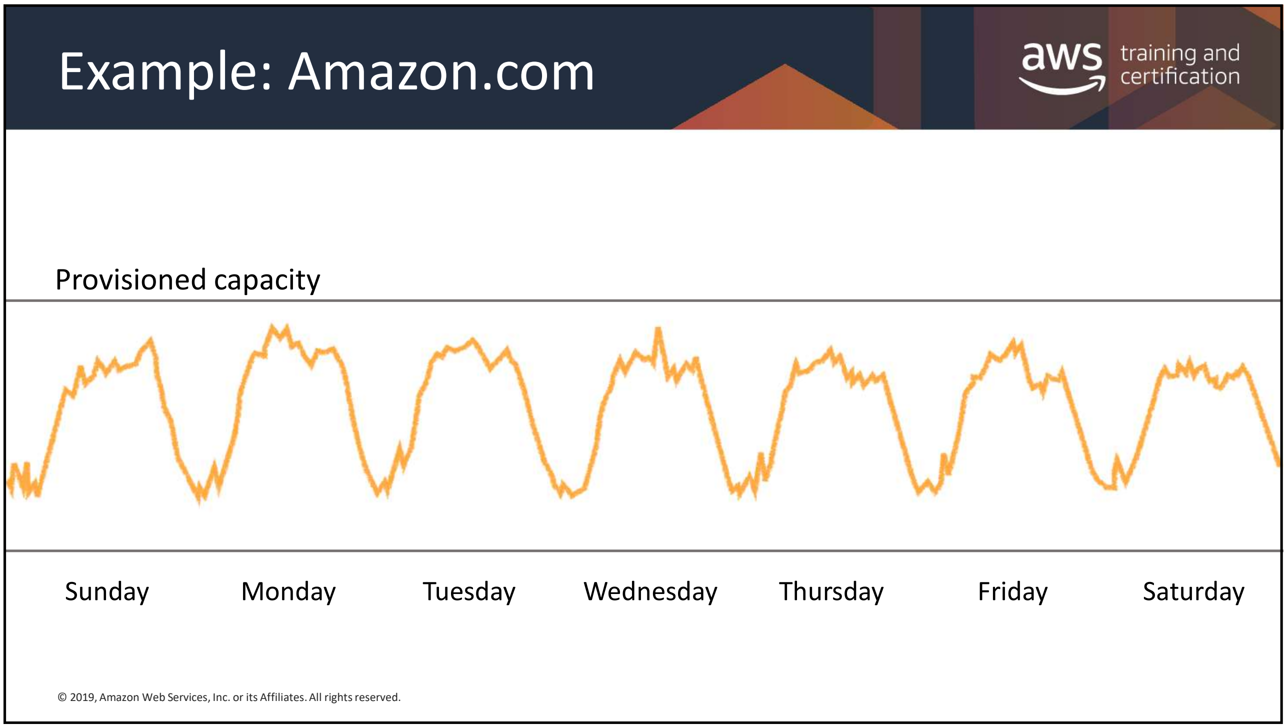
Traditional data centers

Pay for your resources up front and hope they cover your demand

OR

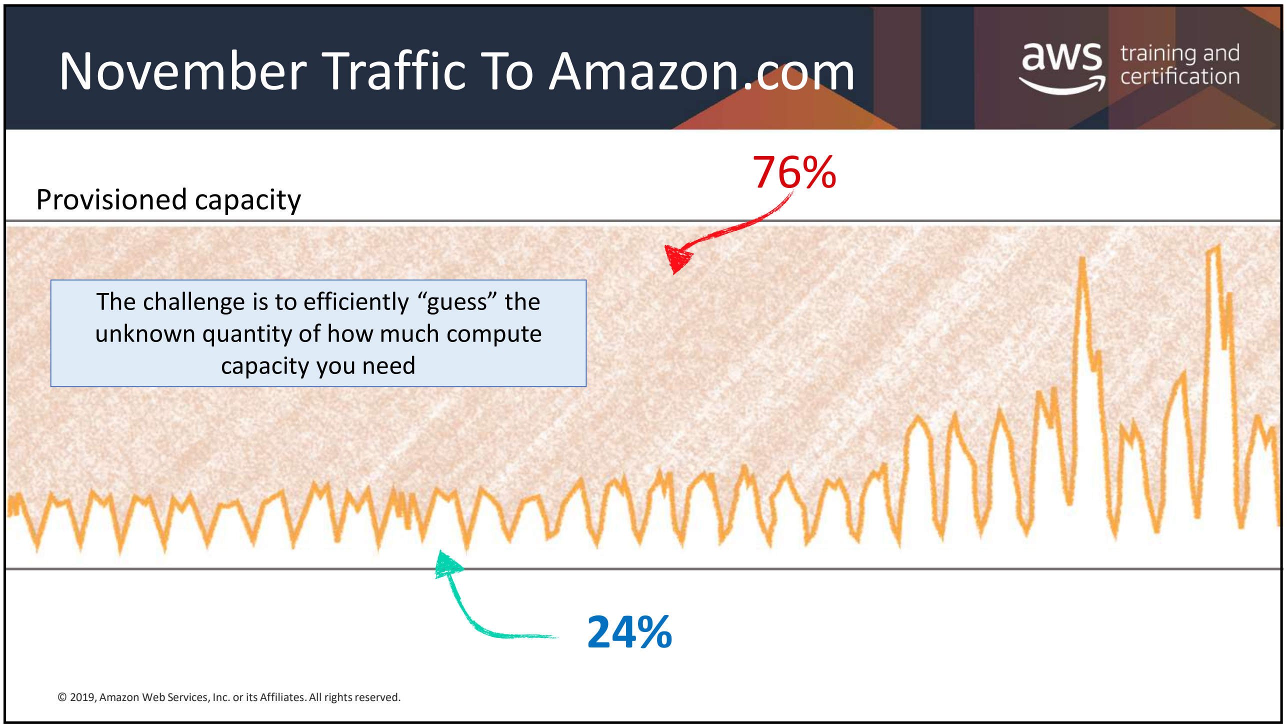
Too many extra resources, wasting money, and burning electricity

5



Example: Amazon.com

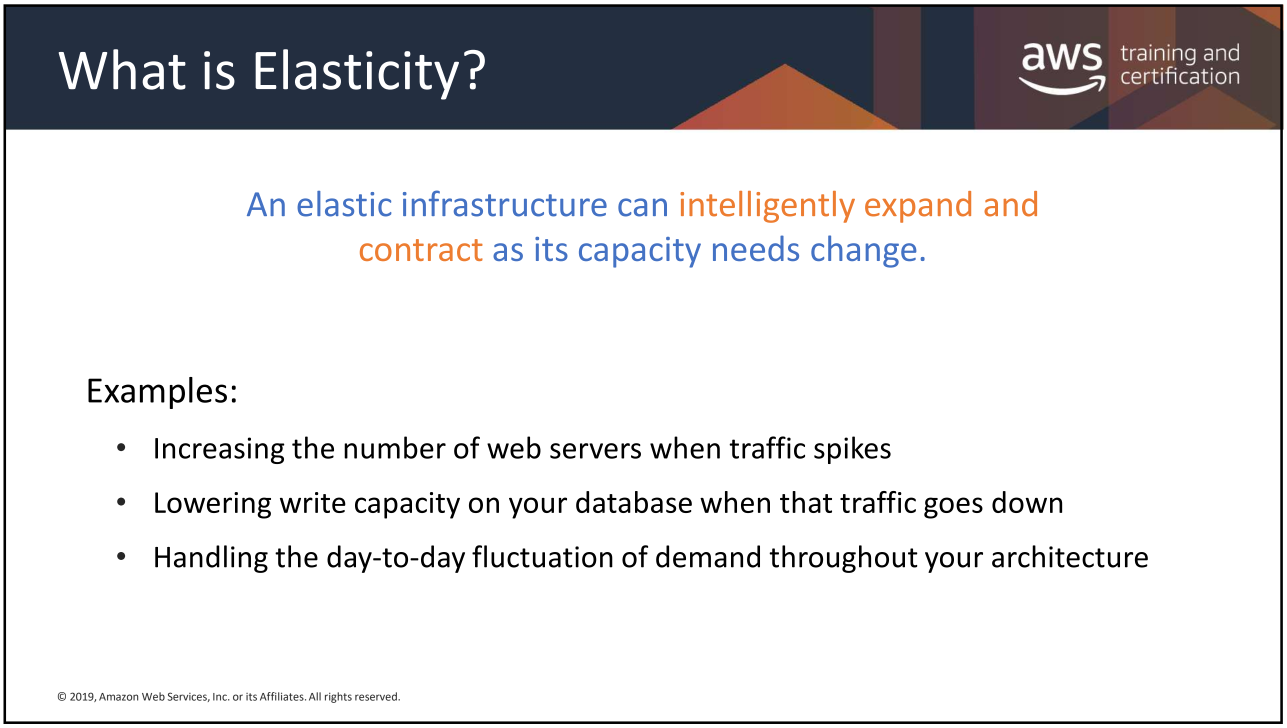
6



November Traffic To Amazon.com

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

7



What is Elasticity?

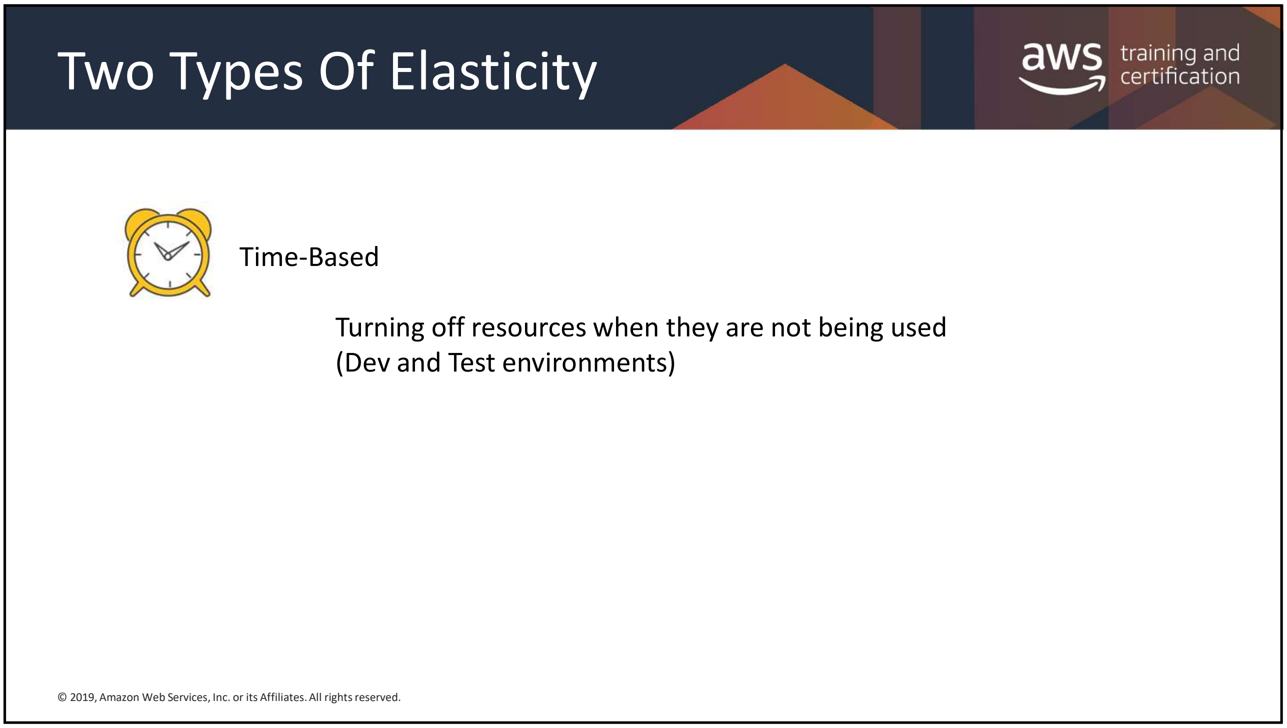
aws training and certification

An elastic infrastructure can intelligently expand and contract as its capacity needs change.

Examples:

- Increasing the number of web servers when traffic spikes
- Lowering write capacity on your database when that traffic goes down
- Handling the day-to-day fluctuation of demand throughout your architecture

8



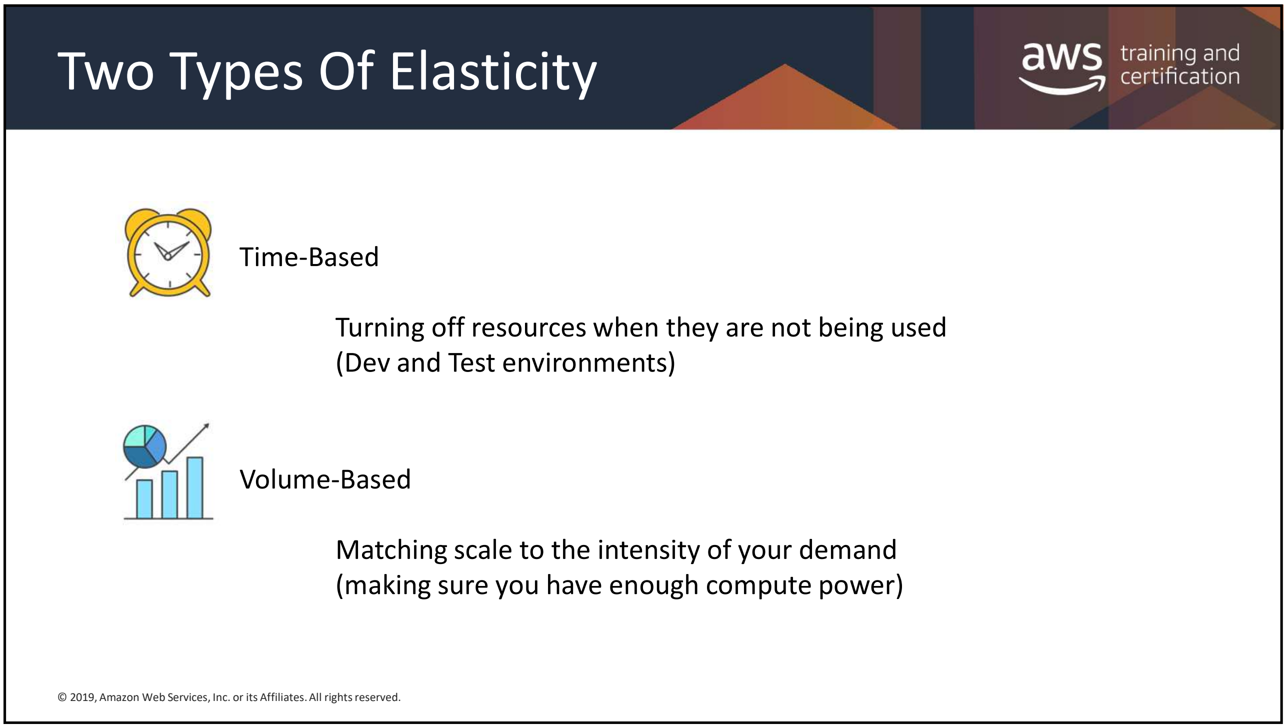
Two Types Of Elasticity

aws training and certification

Time-Based

Turning off resources when they are not being used
(Dev and Test environments)

9



Two Types Of Elasticity

aws training and certification

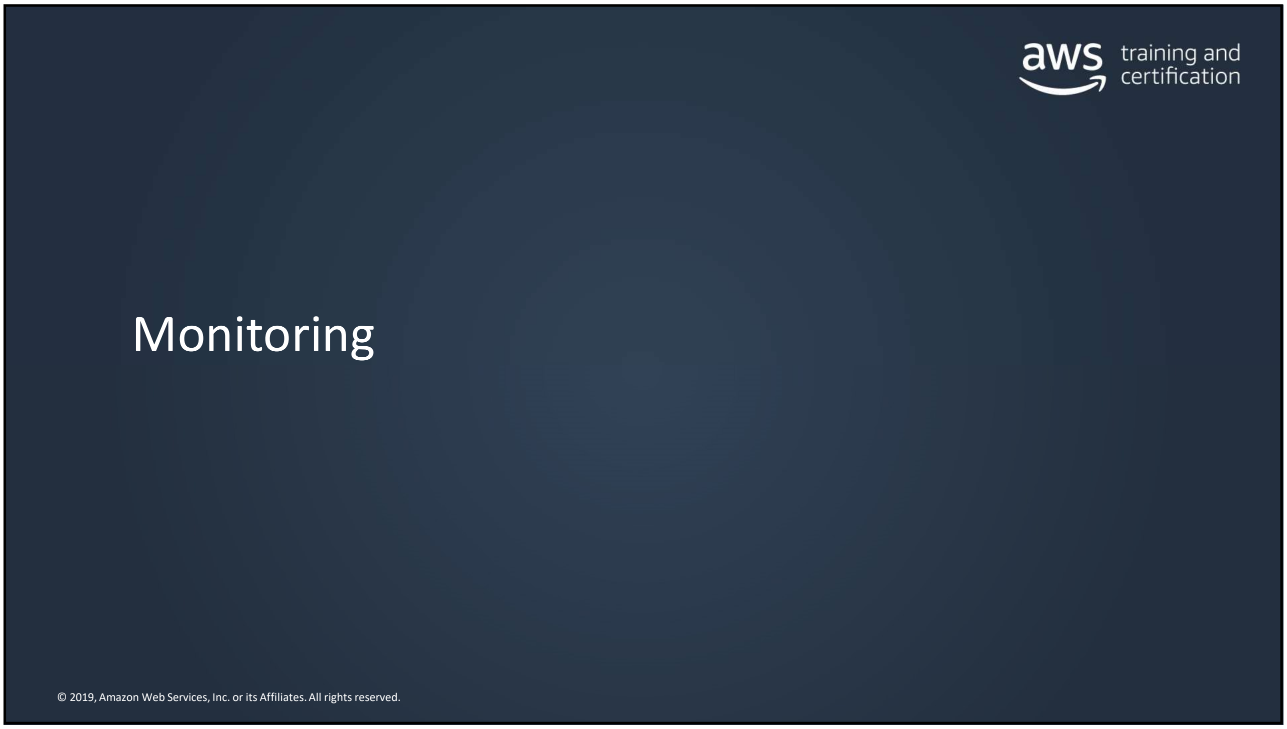
Time-Based

Turning off resources when they are not being used
(Dev and Test environments)

Volume-Based

Matching scale to the intensity of your demand
(making sure you have enough compute power)

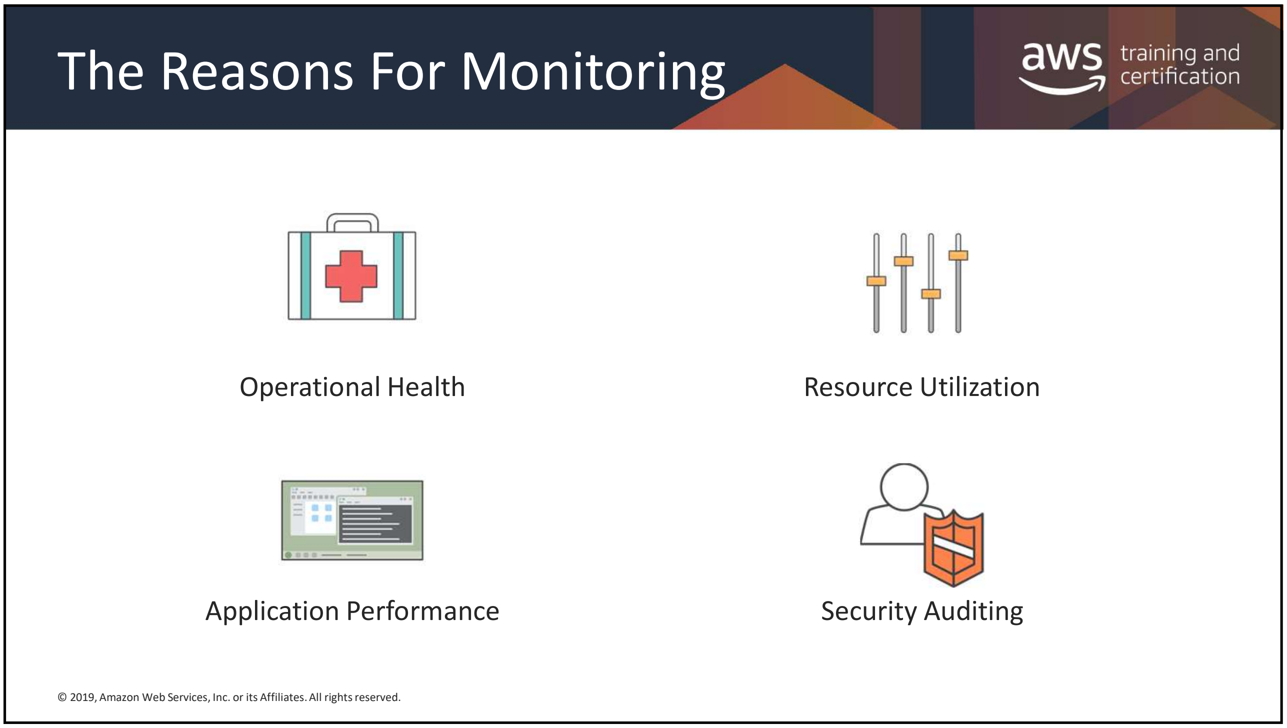
10



aws training and certification

Monitoring

11



The Reasons For Monitoring

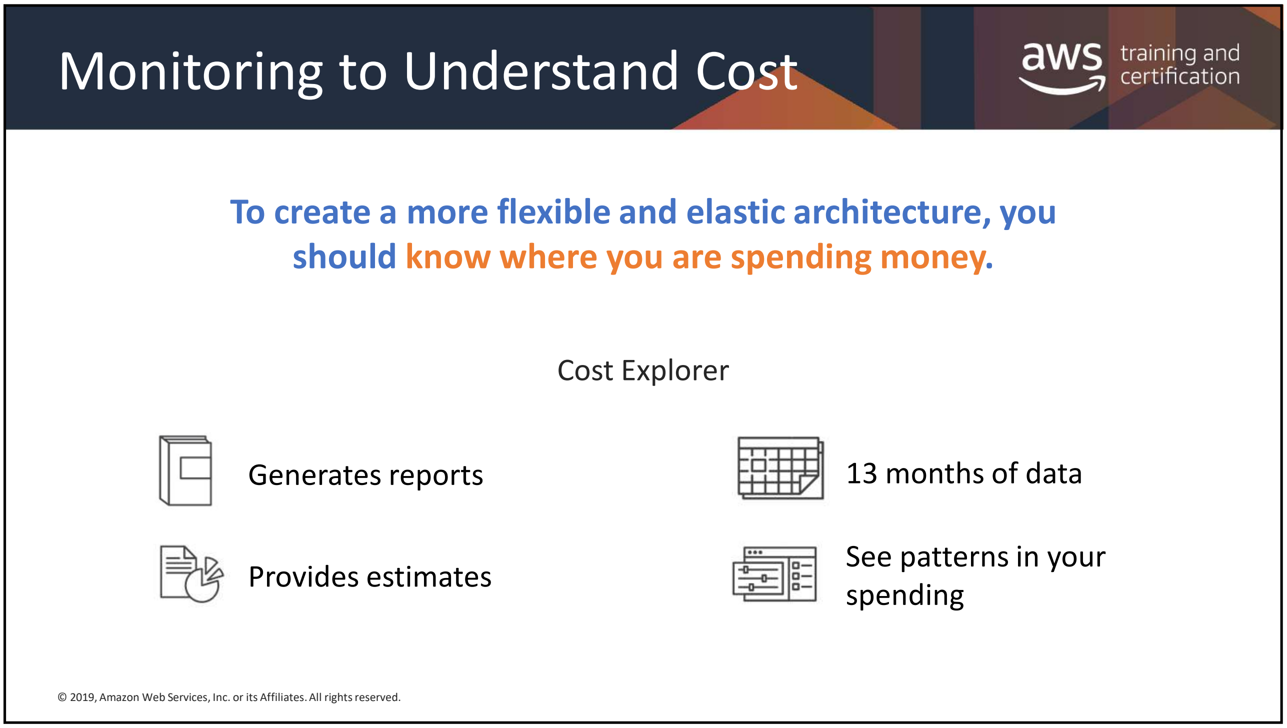
Operational Health

Resource Utilization

Application Performance

Security Auditing

12



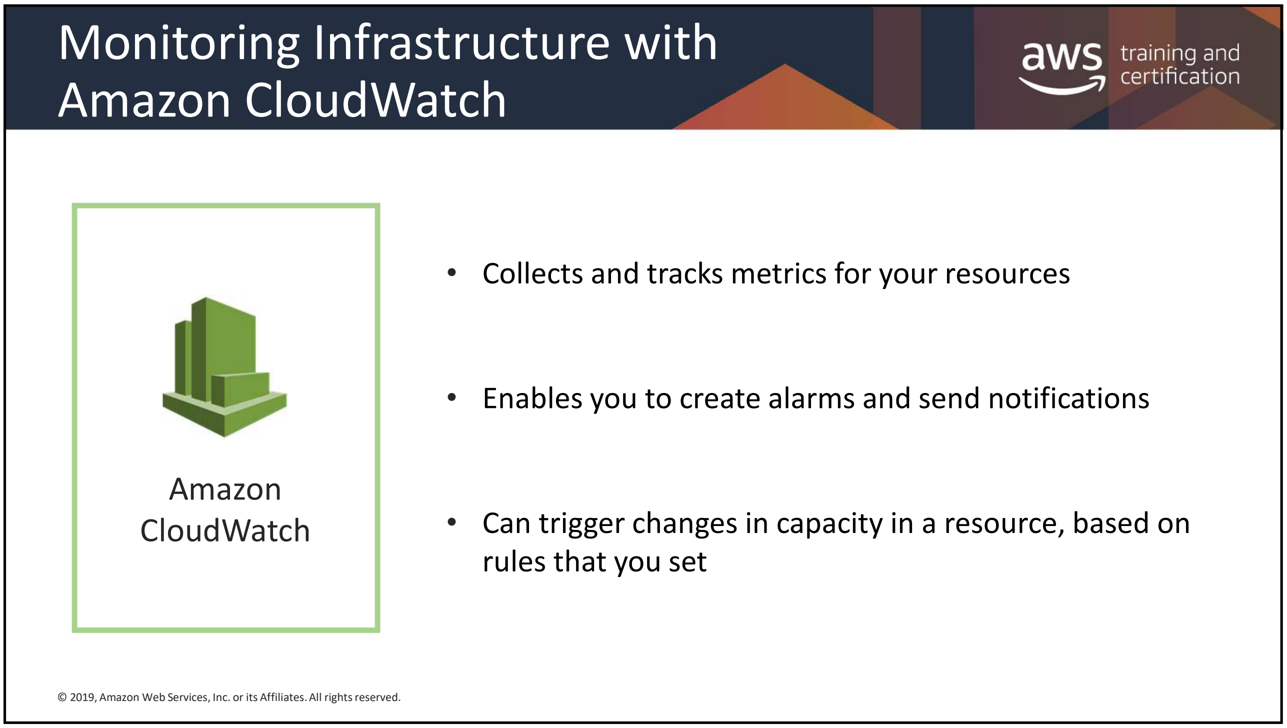
Monitoring to Understand Cost

To create a more flexible and elastic architecture, you should know where you are spending money.

Cost Explorer

- Generates reports
- 13 months of data
- Provides estimates
- See patterns in your spending

13



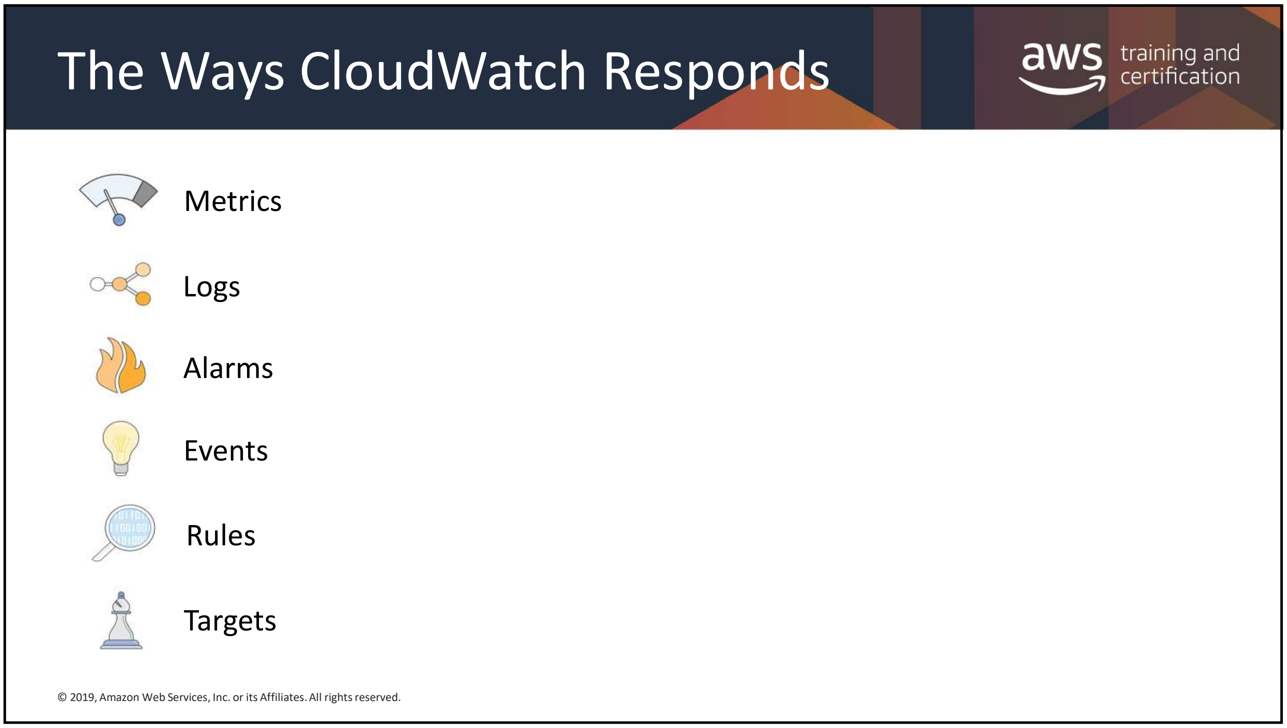
Monitoring Infrastructure with Amazon CloudWatch

aws training and certification

Amazon CloudWatch

- Collects and tracks metrics for your resources
- Enables you to create alarms and send notifications
- Can trigger changes in capacity in a resource, based on rules that you set

14



The Ways CloudWatch Responds

aws training and certification

Metrics

Logs

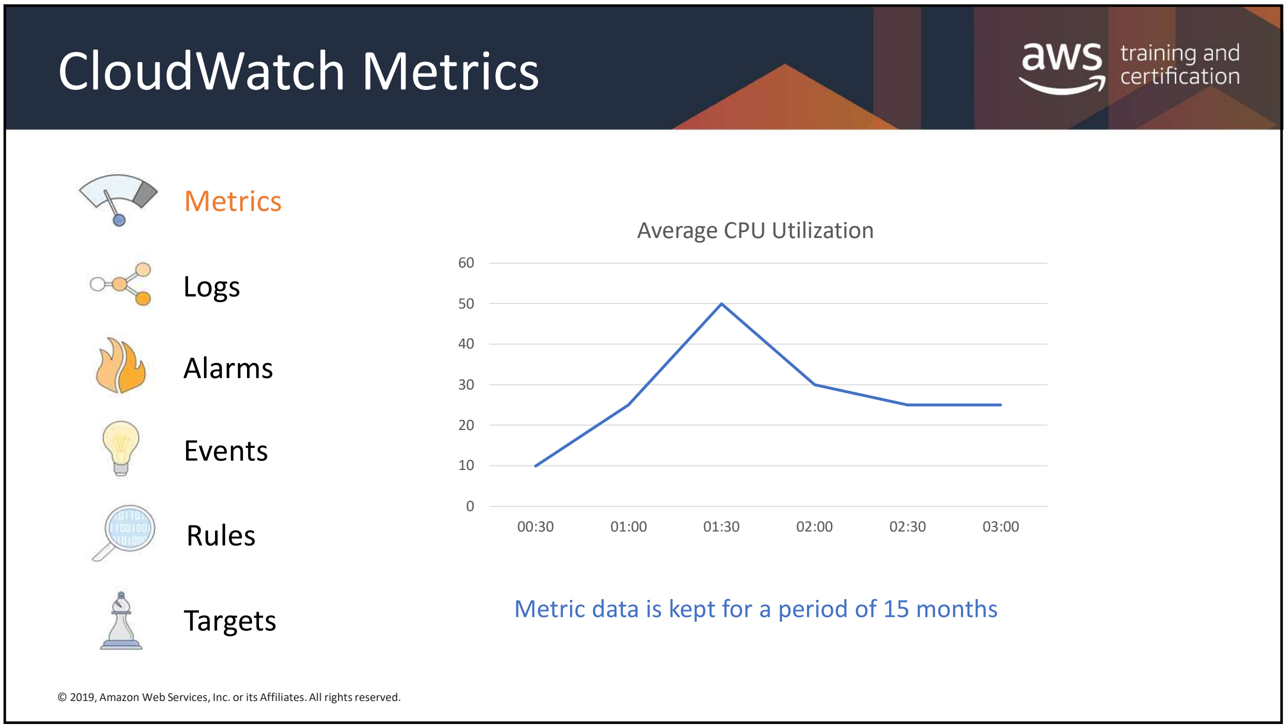
Alarms

Events

Rules

Targets

15

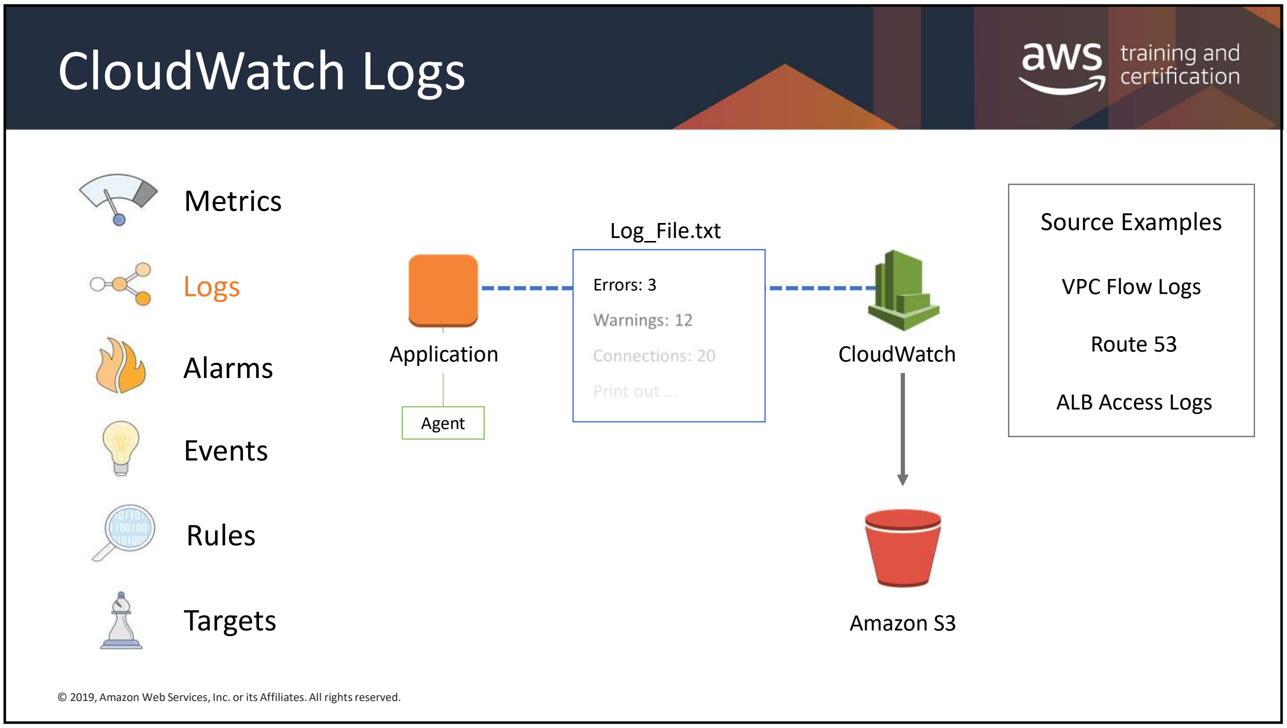


CloudWatch Metrics

- Metrics
- Logs
- Alarms
- Events
- Rules
- Targets

Metric data is kept for a period of 15 months

16



CloudWatch Logs

- Metrics
- Logs
- Alarms
- Events
- Rules
- Targets

Log_File.txt

Errors: 3
Warnings: 12
Connections: 20
Print out ...

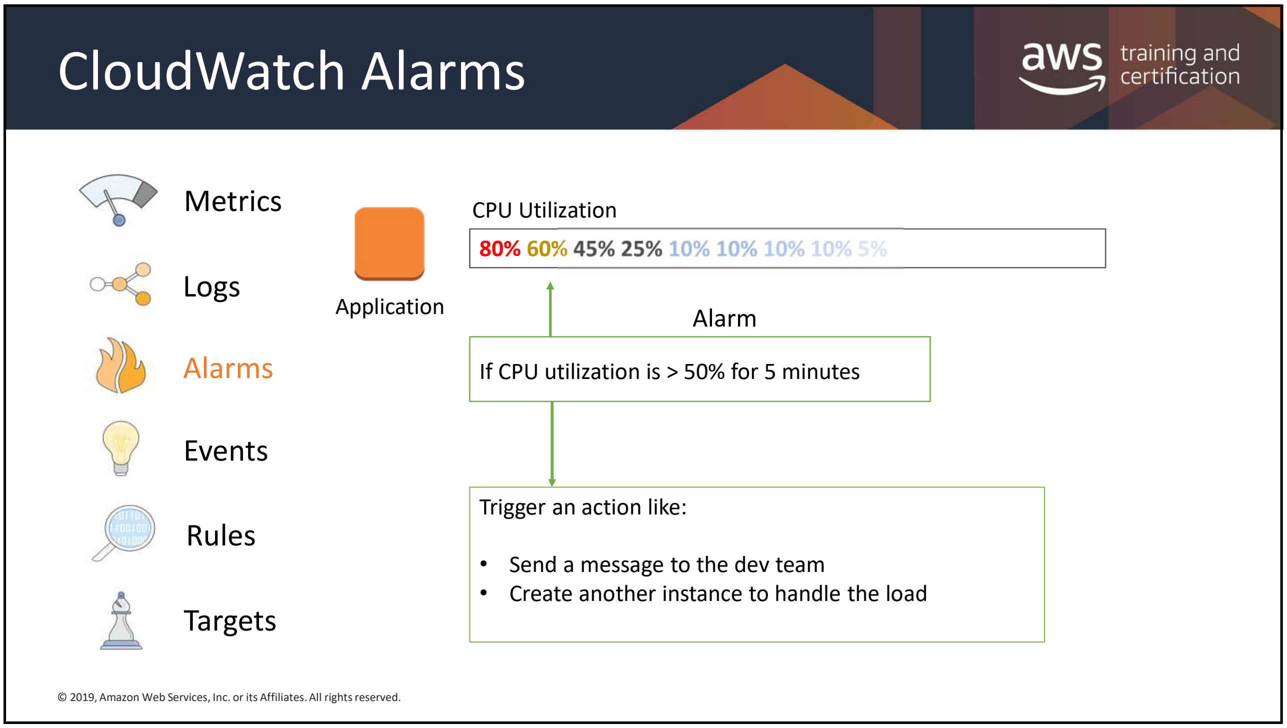
Source Examples

VPC Flow Logs

Route 53

ALB Access Logs

17



CloudWatch Alarms

- Metrics
- Logs
- Alarms
- Events
- Rules
- Targets

Application

CPU Utilization

80% 60% 45% 25% 10% 10% 10% 10% 5%

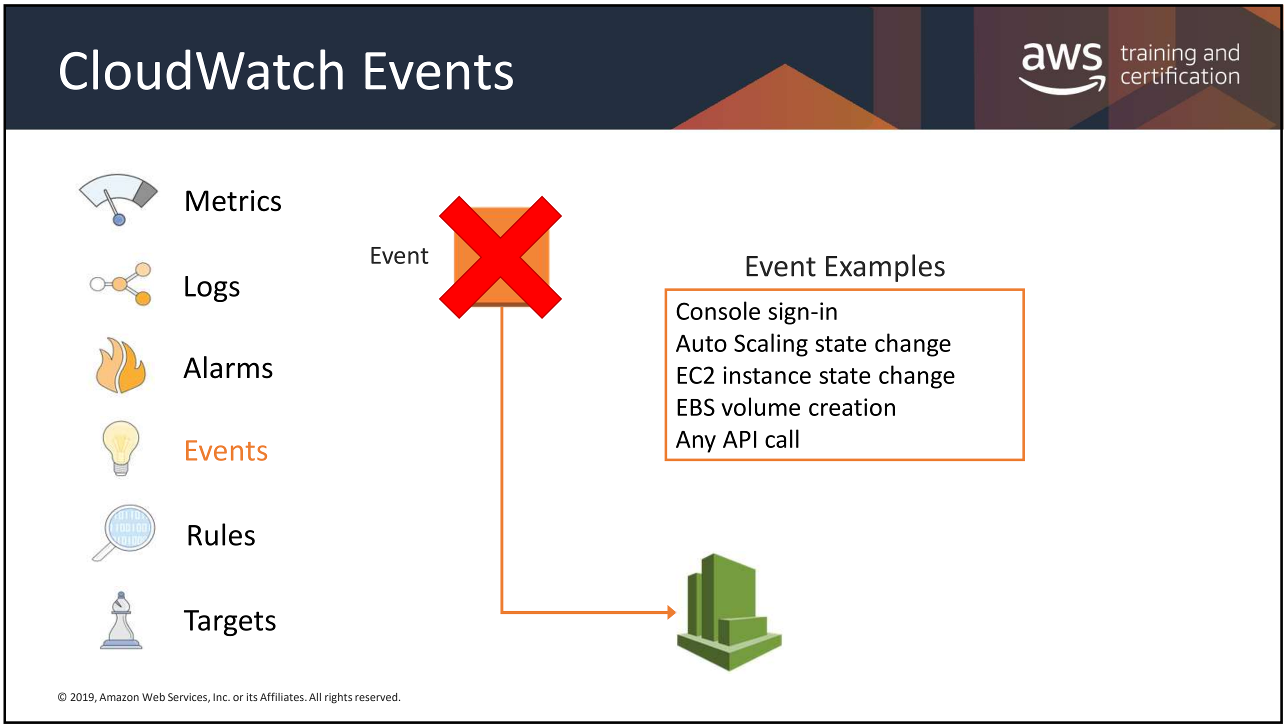
Alarm

If CPU utilization is > 50% for 5 minutes

Trigger an action like:

- Send a message to the dev team
- Create another instance to handle the load

18



CloudWatch Events

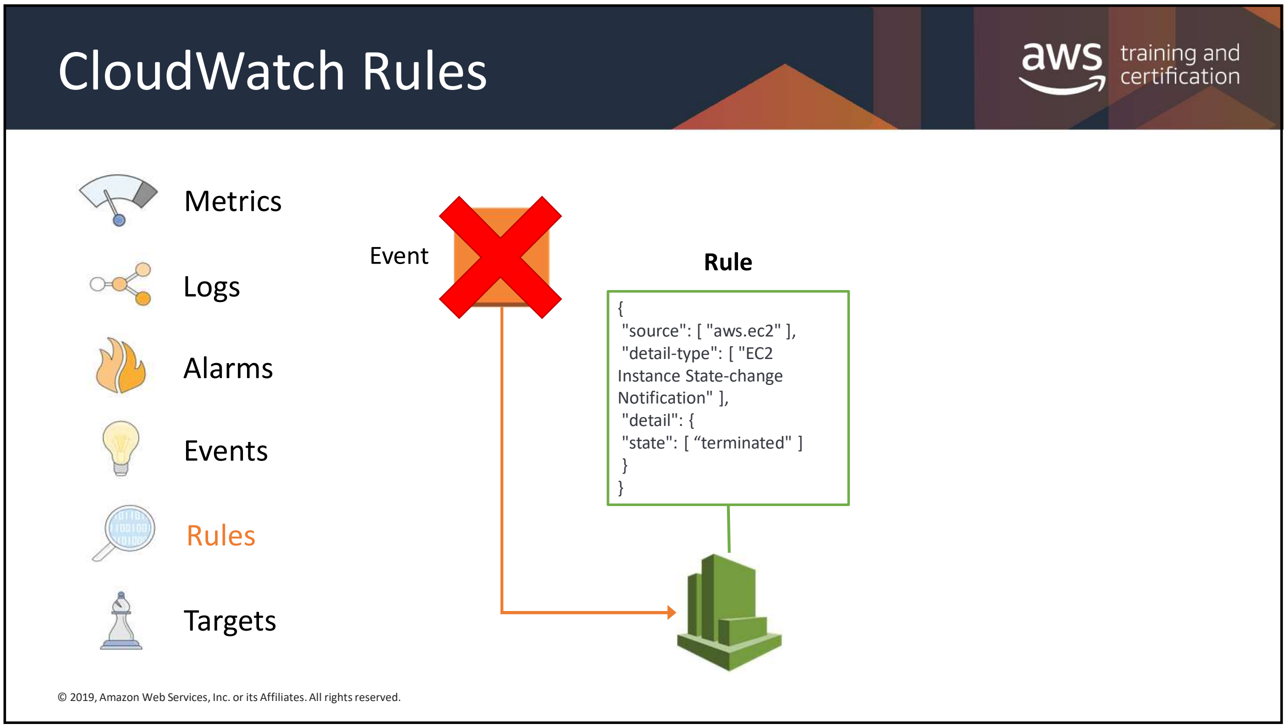
- Metrics
- Logs
- Alarms
- Events
- Rules
- Targets

Event

Event Examples

Console sign-in
Auto Scaling state change
EC2 instance state change
EBS volume creation
Any API call

19



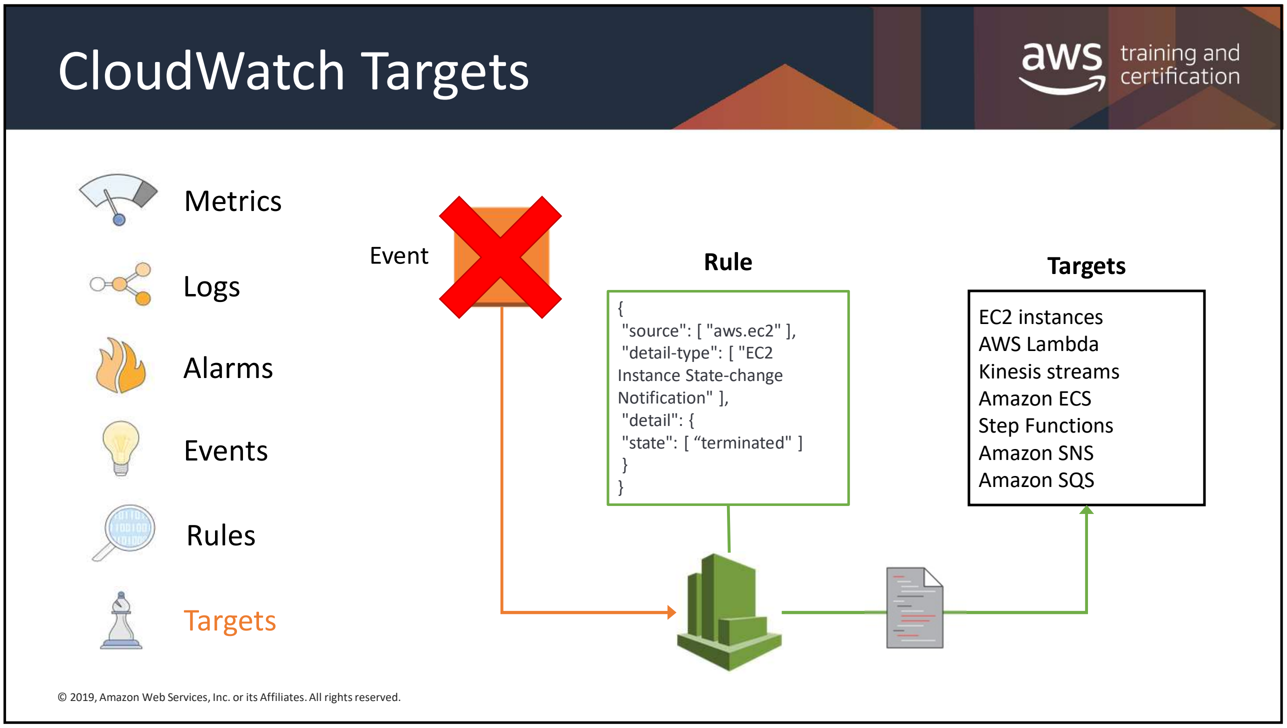
CloudWatch Rules

- Metrics
- Logs
- Alarms
- Events
- Rules
- Targets

Rule

```
{
 "source": [ "aws.ec2" ],
 "detail-type": [ "EC2
Instance State-change
Notification" ],
 "detail": {
 "state": [ “terminated" ]
 }
}
```

20



CloudWatch Targets

- Metrics
- Logs
- Alarms
- Events
- Rules
- Targets

Rule

```
{
 "source": [ "aws.ec2" ],
 "detail-type": [ "EC2
Instance State-change
Notification" ],
 "detail": {
 "state": [ “terminated" ]
 }
}
```

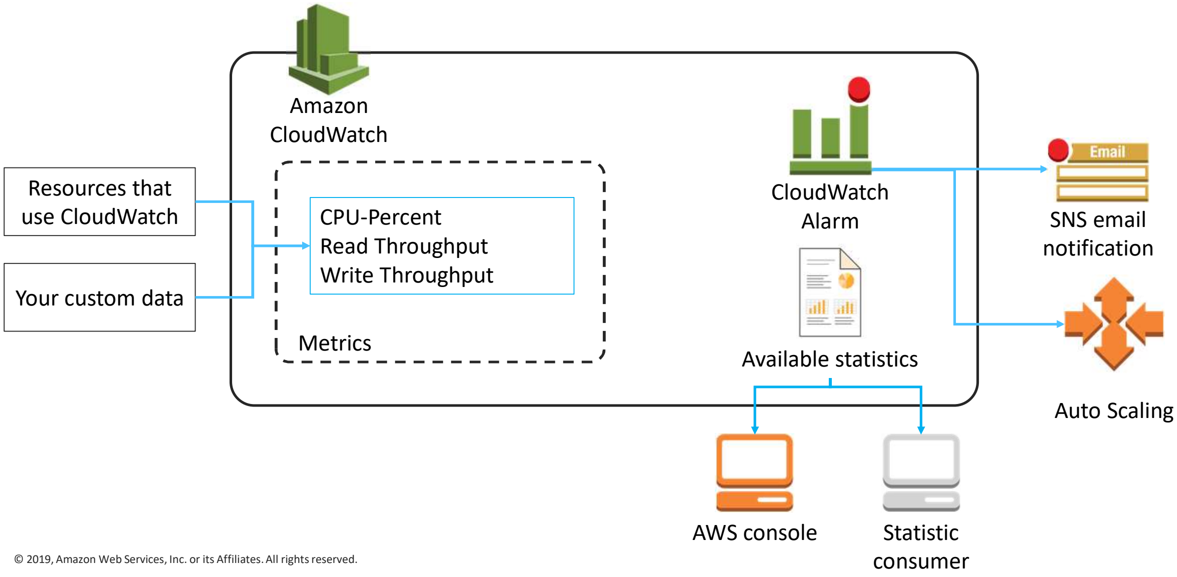
Targets

EC2 instances
AWS Lambda
Kinesis streams
Amazon ECS
Step Functions
Amazon SNS
Amazon SQS

21

Visualizing CloudWatch

aws training and certification



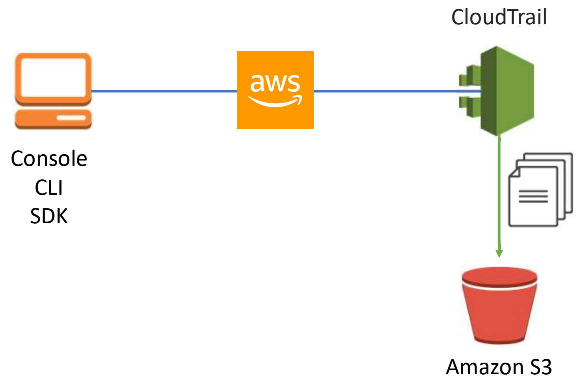
22

Monitoring Your Users with AWS CloudTrail

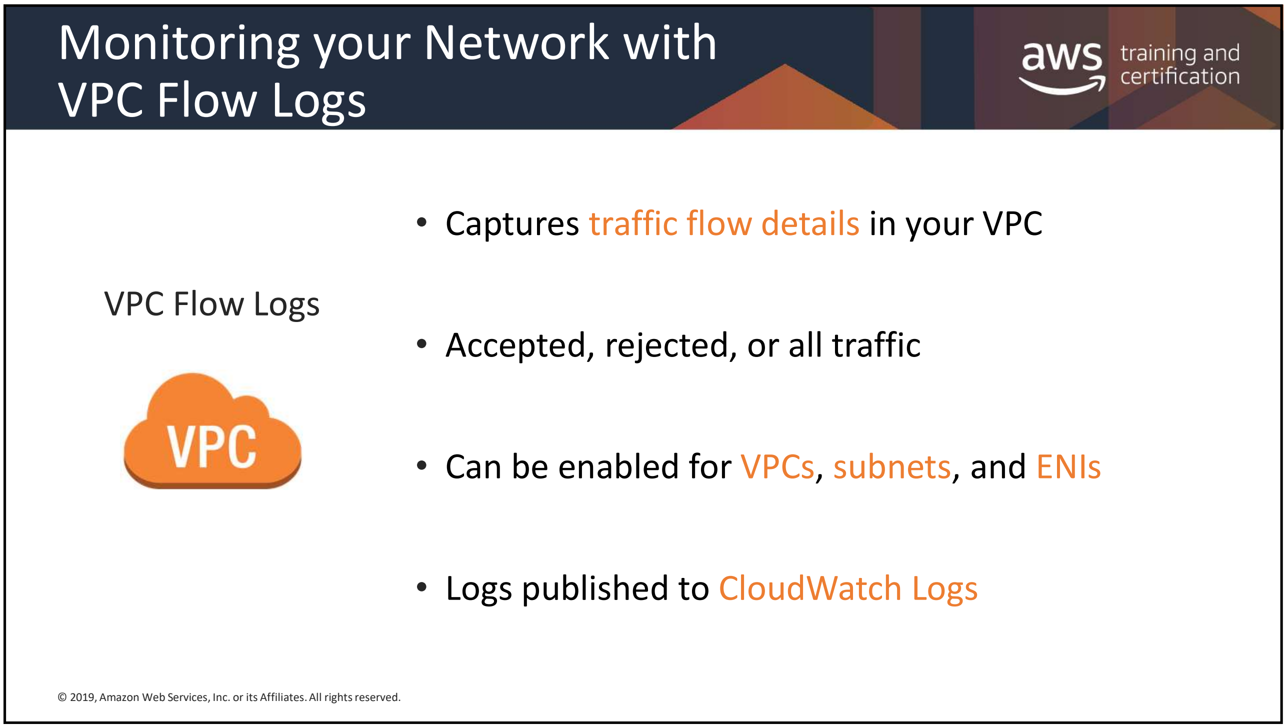
aws training and certification

AWS CloudTrail

CloudTrail records all API calls made in your account and saves logs in your designated Amazon S3 bucket.



23



Monitoring your Network with VPC Flow Logs

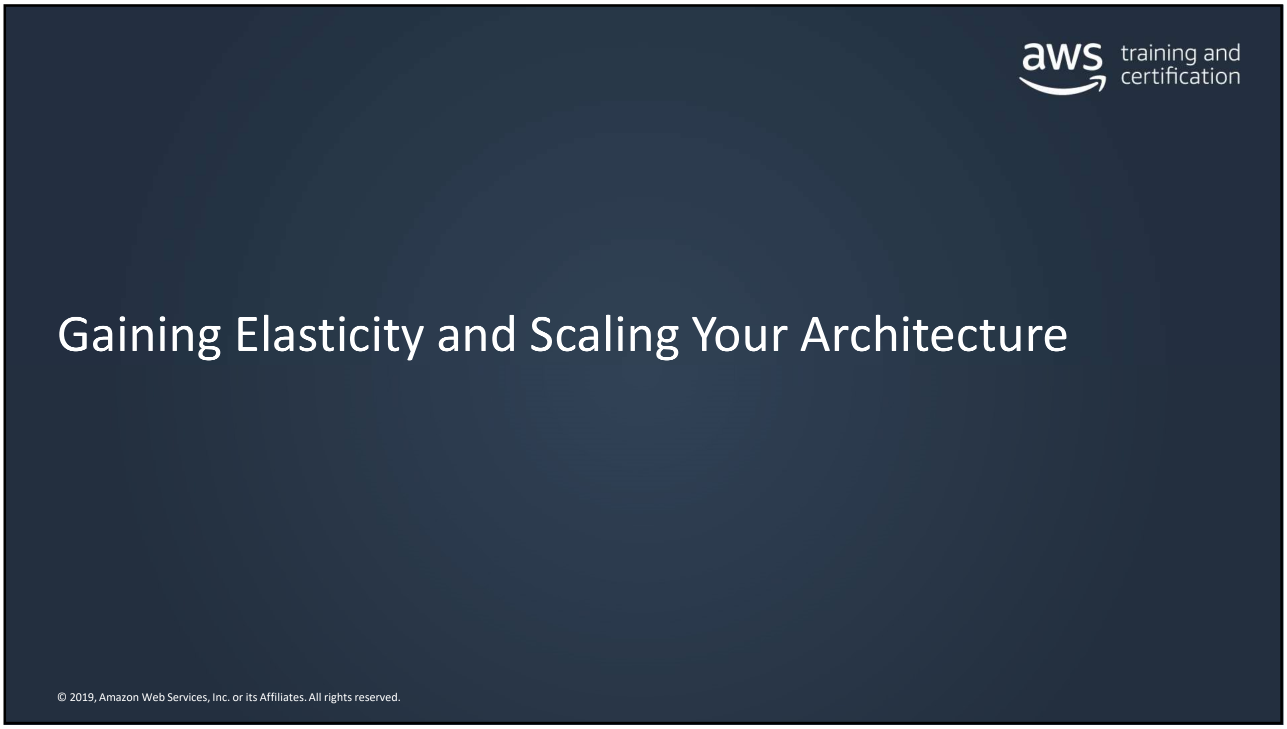
aws training and certification

VPC Flow Logs

VPC

- Captures traffic flow details in your VPC
- Accepted, rejected, or all traffic
- Can be enabled for VPCs, subnets, and ENIs
- Logs published to CloudWatch Logs

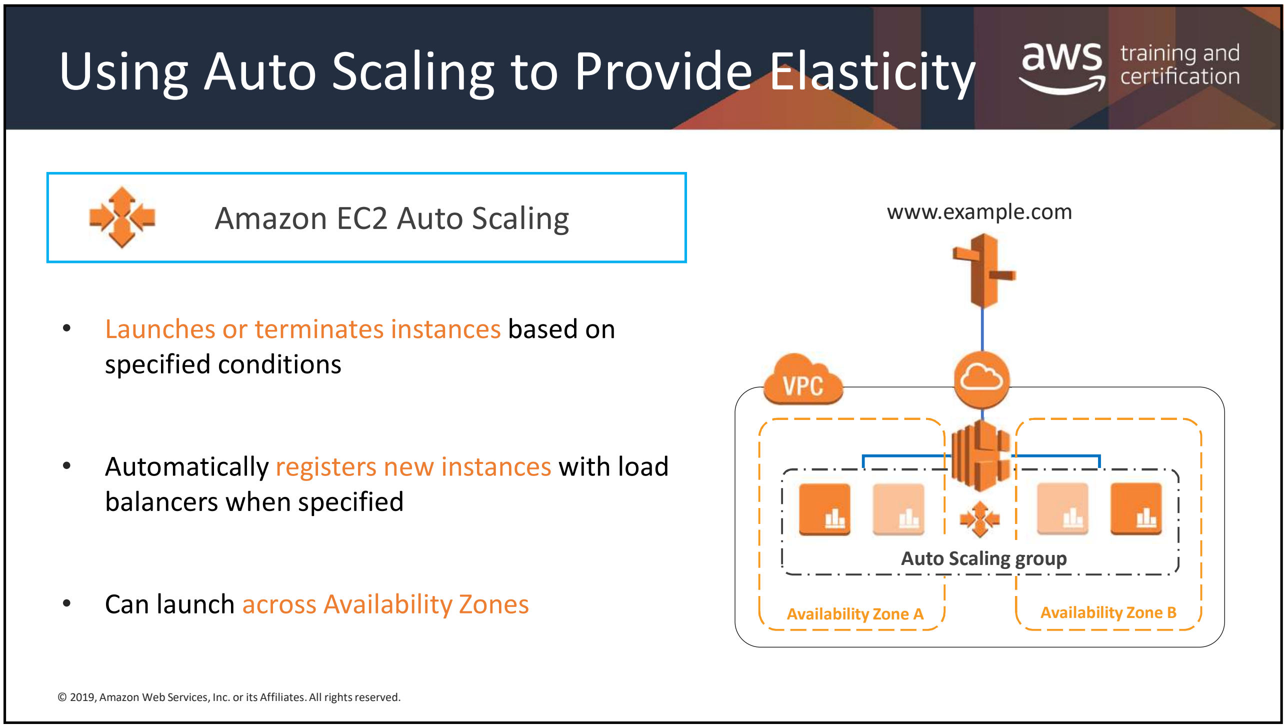
24



aws training and certification

Gaining Elasticity and Scaling Your Architecture

25



Using Auto Scaling to Provide Elasticity

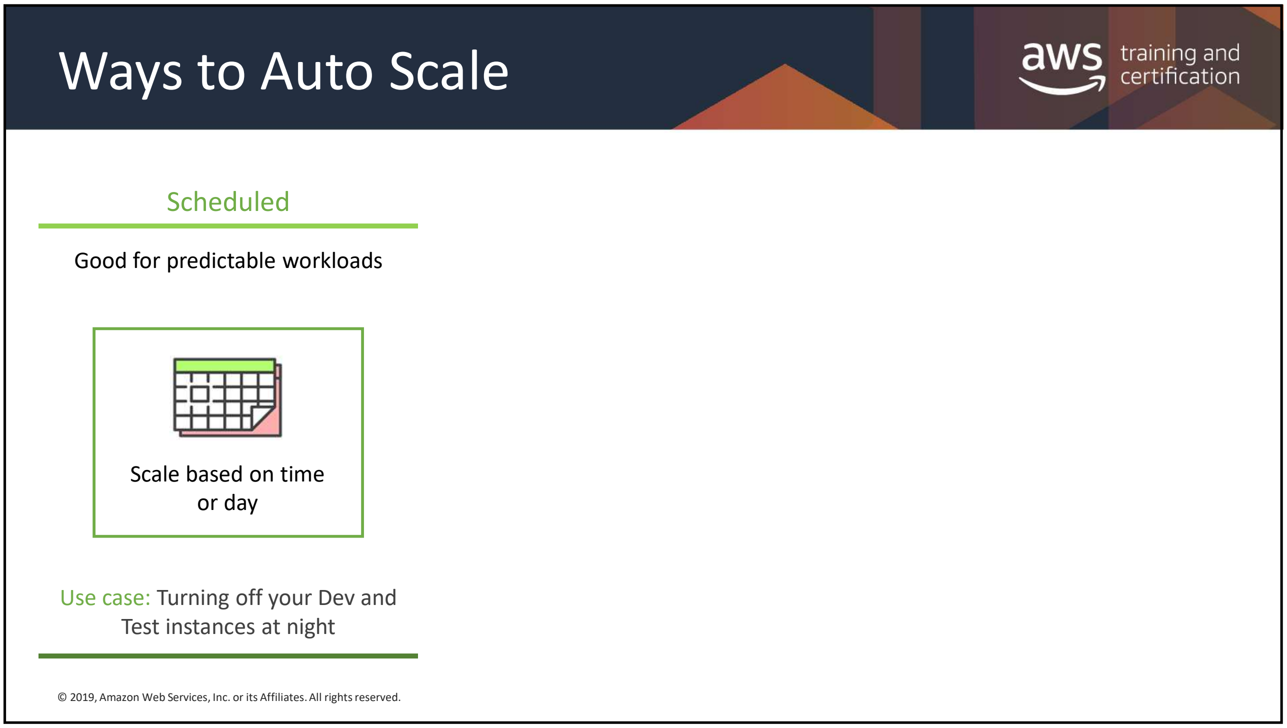
aws training and certification

Amazon EC2 Auto Scaling

- Launches or terminates instances based on specified conditions
- Automatically registers new instances with load balancers when specified
- Can launch across Availability Zones

www.example.com
VPC
Auto Scaling group
Availability Zone A
Availability Zone B

26



Ways to Auto Scale

aws training and certification

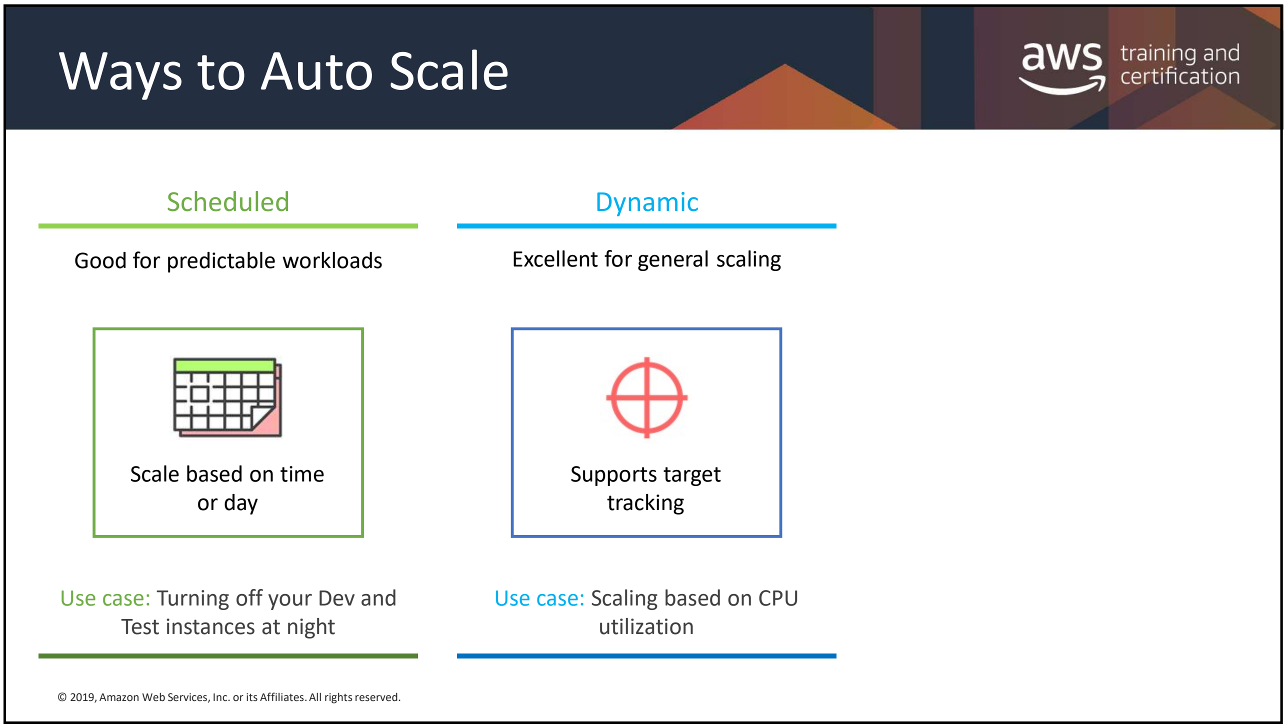
Scheduled

Good for predictable workloads

Scale based on time or day

Use case: Turning off your Dev and Test instances at night

27



Ways to Auto Scale

aws training and certification

Scheduled

Good for predictable workloads

Scale based on time or day

Use case: Turning off your Dev and Test instances at night

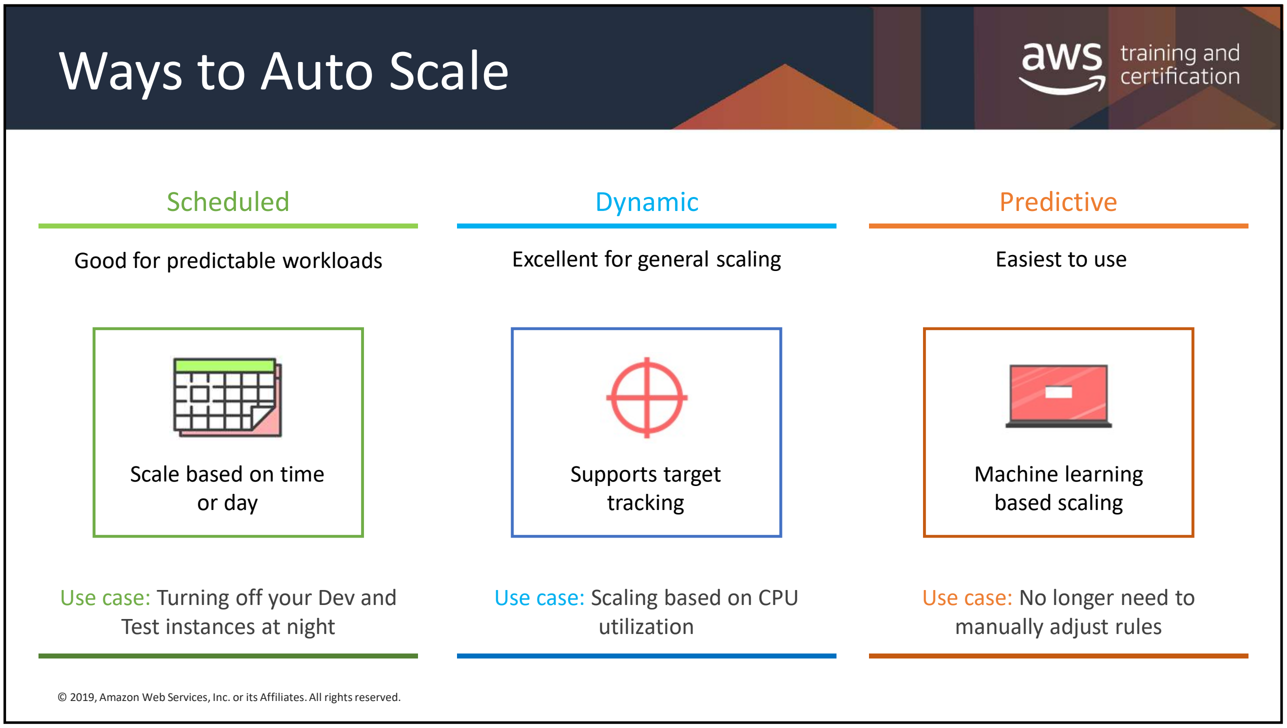
Dynamic

Excellent for general scaling

Supports target tracking

Use case: Scaling based on CPU utilization

28



Ways to Auto Scale

aws training and certification

Scheduled

Good for predictable workloads

Scale based on time or day

Use case: Turning off your Dev and Test instances at night

Dynamic

Excellent for general scaling

Supports target tracking

Use case: Scaling based on CPU utilization

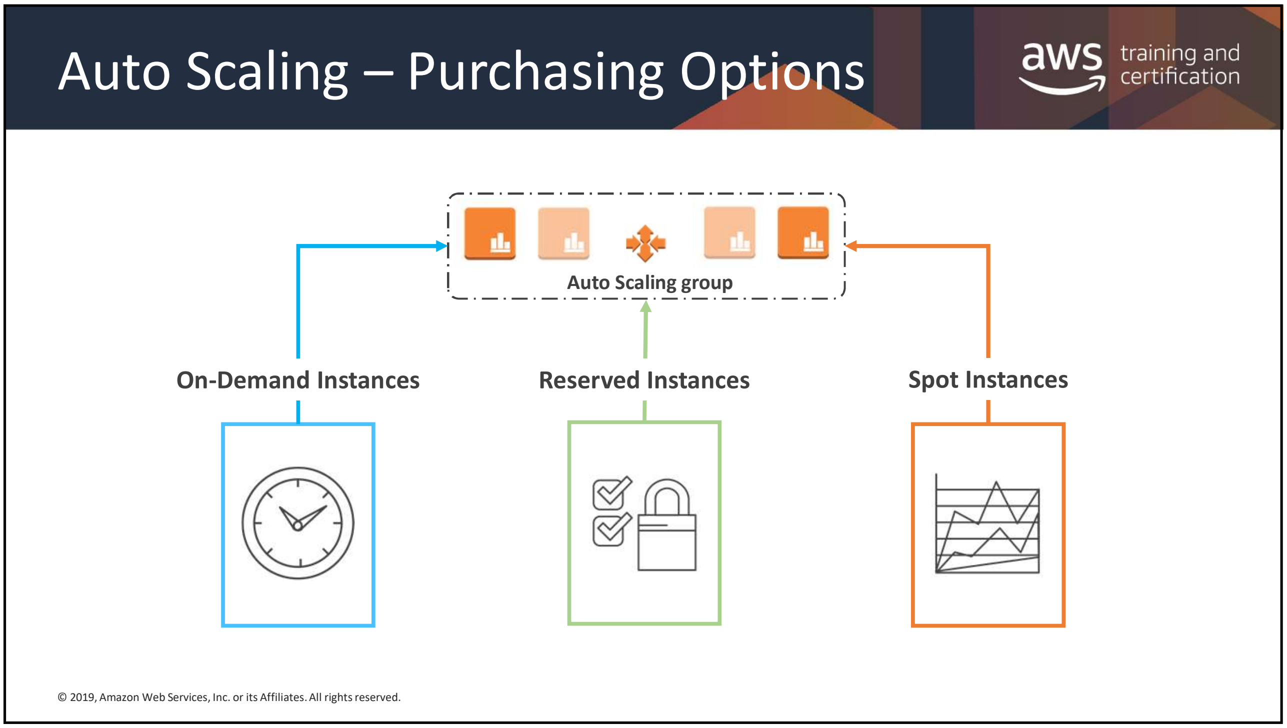
Predictive

Easiest to use

Machine learning based scaling

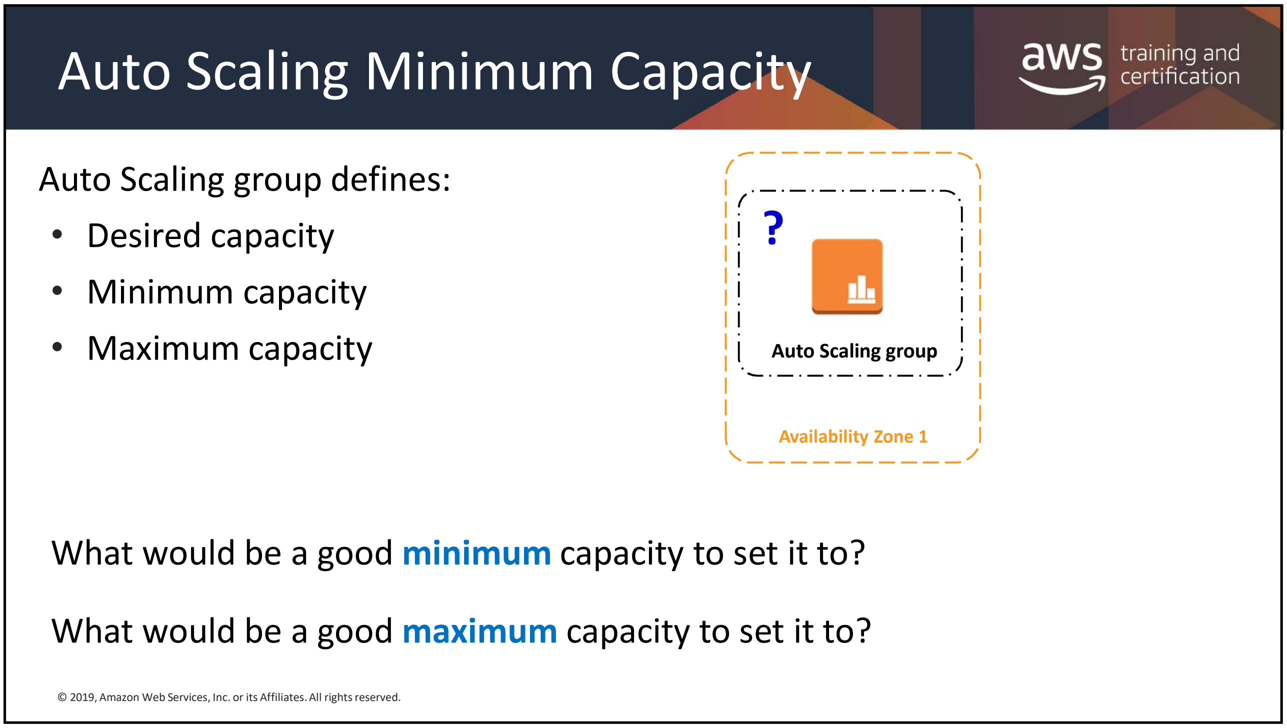
Use case: No longer need to manually adjust rules

29



Auto Scaling – Purchasing Options

30



Auto Scaling Minimum Capacity

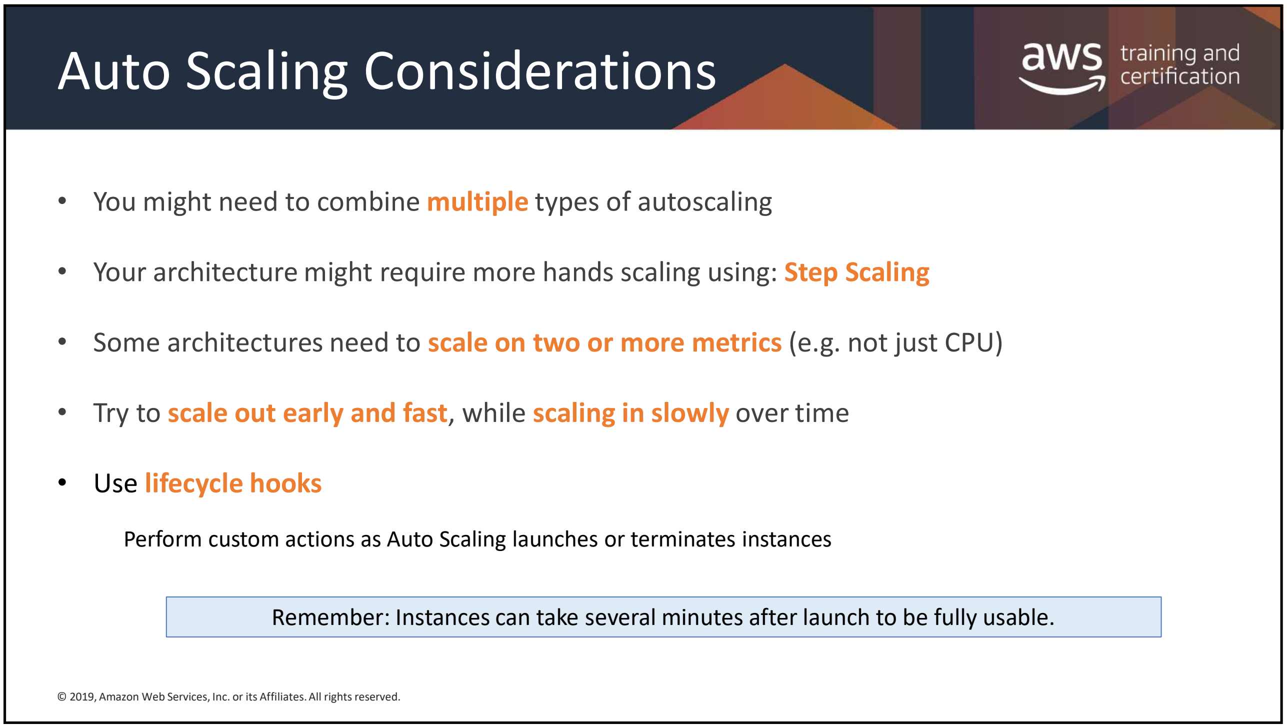
Auto Scaling group defines:

- 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?

31



Auto Scaling Considerations

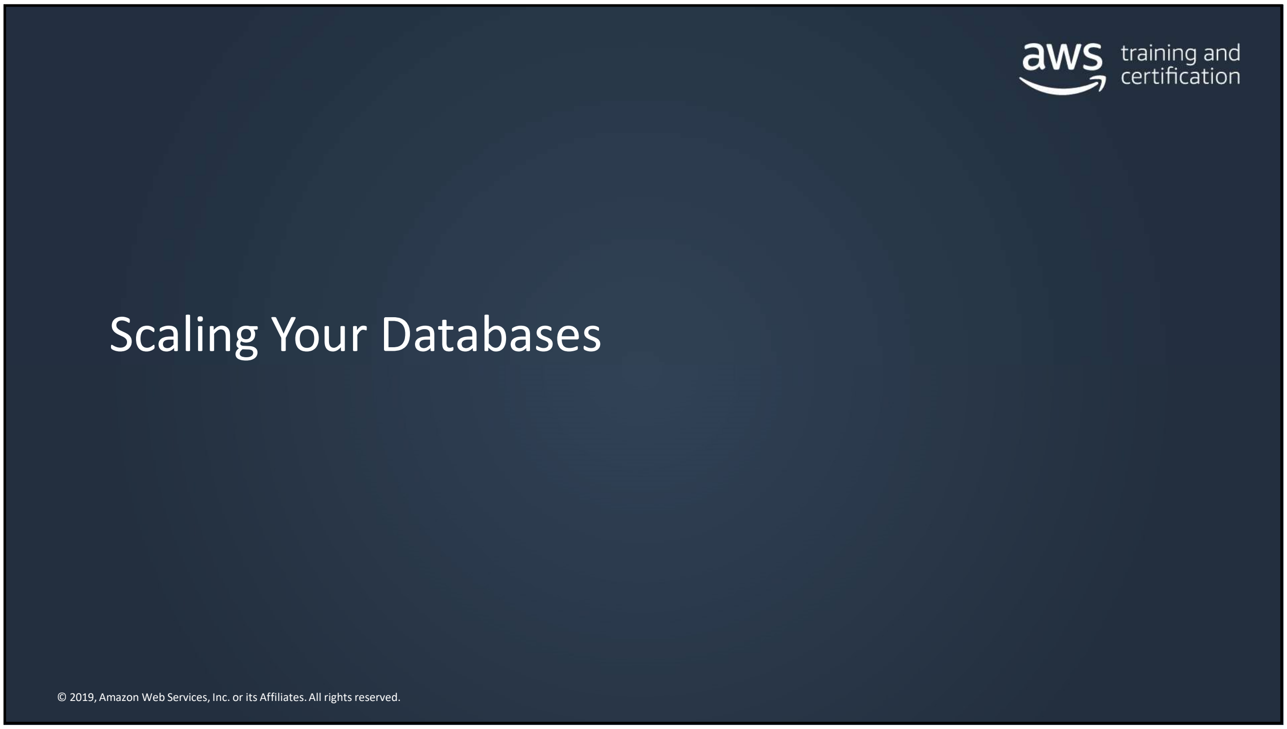
aws training and certification

- You might need to combine **multiple** types of autoscaling
- Your architecture might require more hands scaling using: **Step Scaling**
- Some architectures need to **scale on two or more metrics** (e.g. not just CPU)
- Try to **scale out early and fast**, while **scaling in slowly** over time
- Use **lifecycle hooks**

  Perform custom actions as Auto Scaling launches or terminates instances

Remember: Instances can take several minutes after launch to be fully usable.

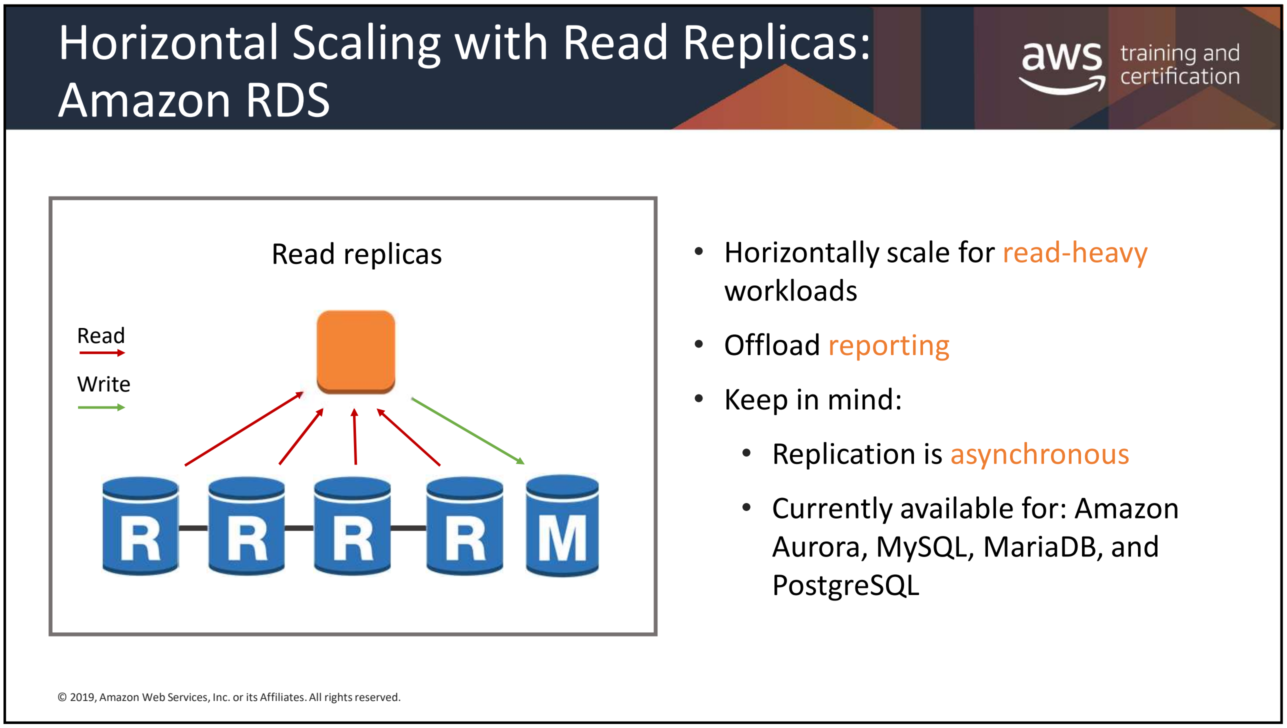
32



aws training and certification

Scaling Your Databases

33

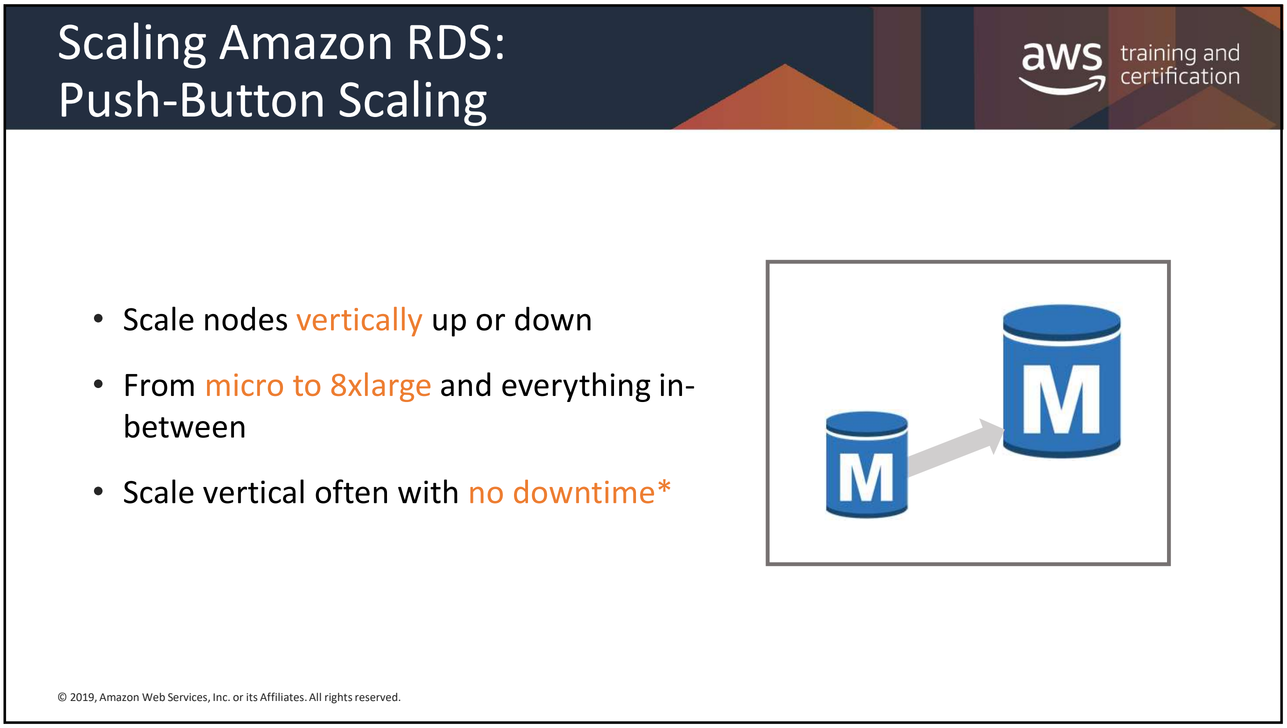


Horizontal Scaling with Read Replicas: Amazon RDS

Read replicas
Read
Write
R
R
R
R
M

- Horizontally scale for read-heavy workloads
- Offload reporting
- Keep in mind:
  - Replication is asynchronous
  - Currently available for: Amazon Aurora, MySQL, MariaDB, and PostgreSQL

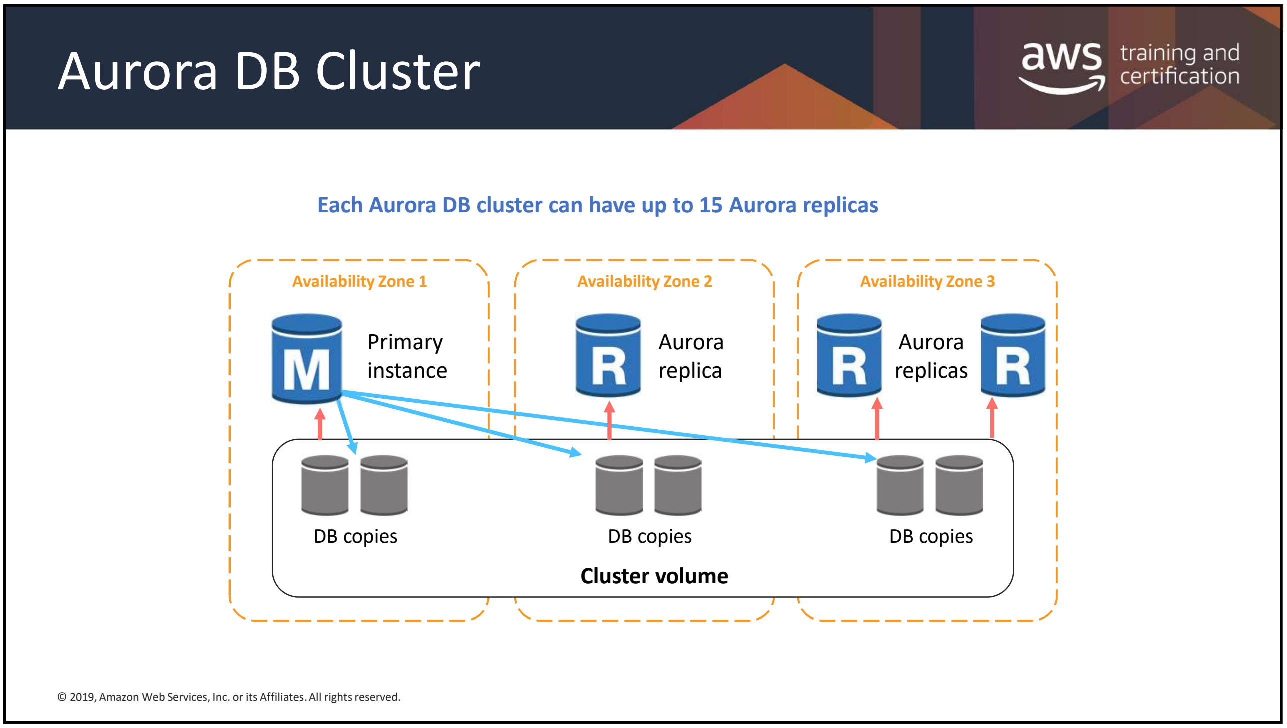
34



Scaling Amazon RDS: Push-Button Scaling

- Scale nodes vertically up or down
- From micro to 8xlarge and everything in-between
- Scale vertical often with no downtime*

35



Aurora DB Cluster

aws training and certification

Each Aurora DB cluster can have up to 15 Aurora replicas

Availability Zone 1

Primary instance

DB copies

Availability Zone 2

Aurora replica

DB copies

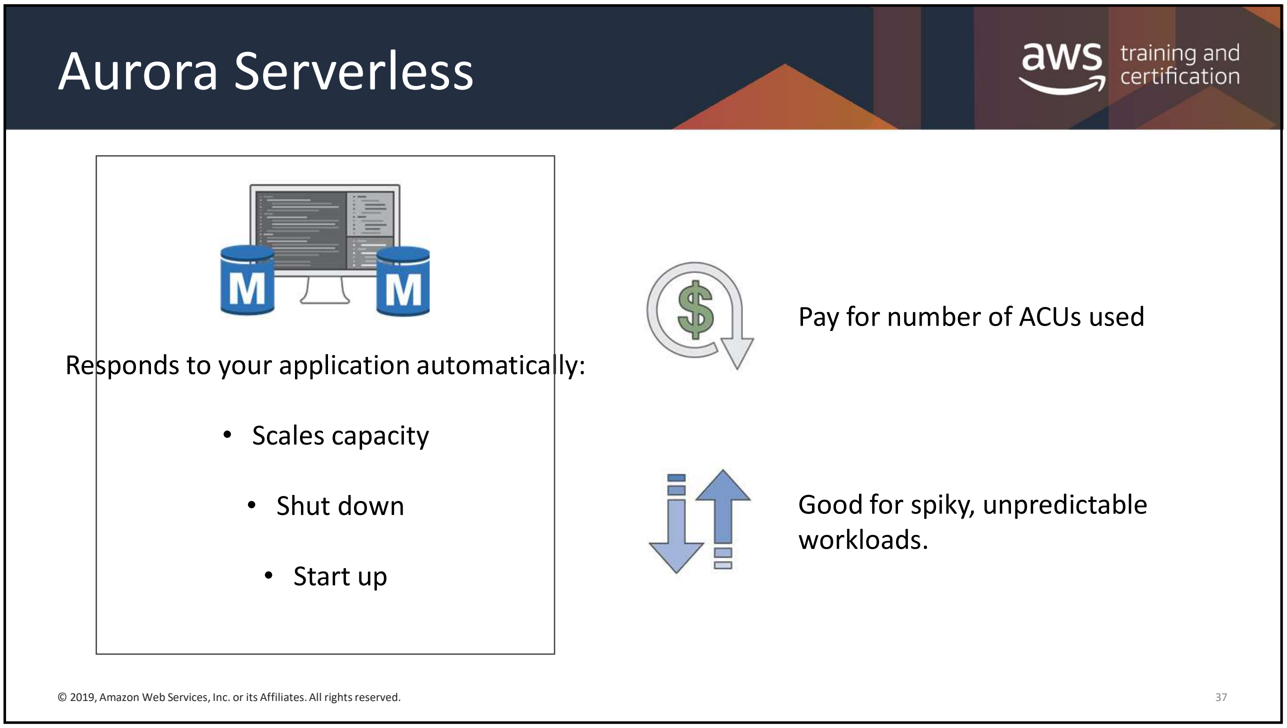
Availability Zone 3

Aurora replicas

DB copies

Cluster volume

36



Aurora Serverless

aws training and certification

Responds to your application automatically:

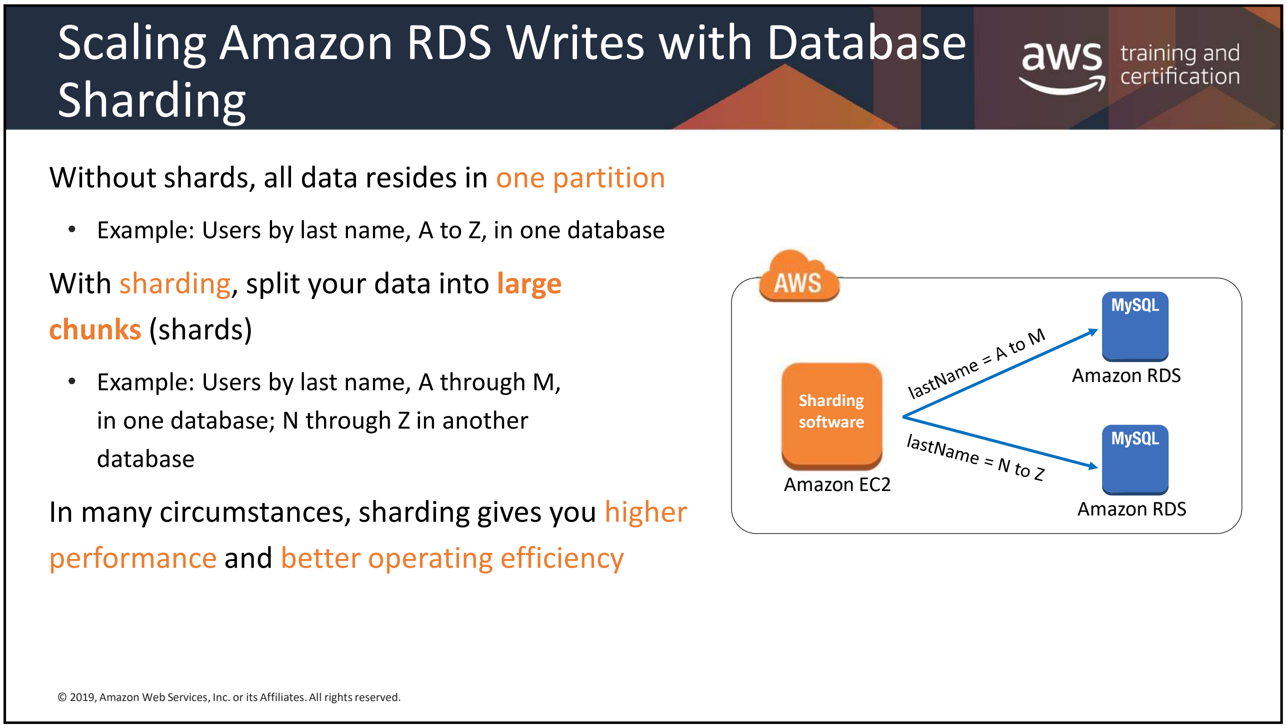
- Scales capacity
- Shut down
- Start up

Pay for number of ACUs used

Good for spiky, unpredictable workloads.

37

37



Scaling Amazon RDS Writes with Database Sharding

Without shards, all data resides in one partition

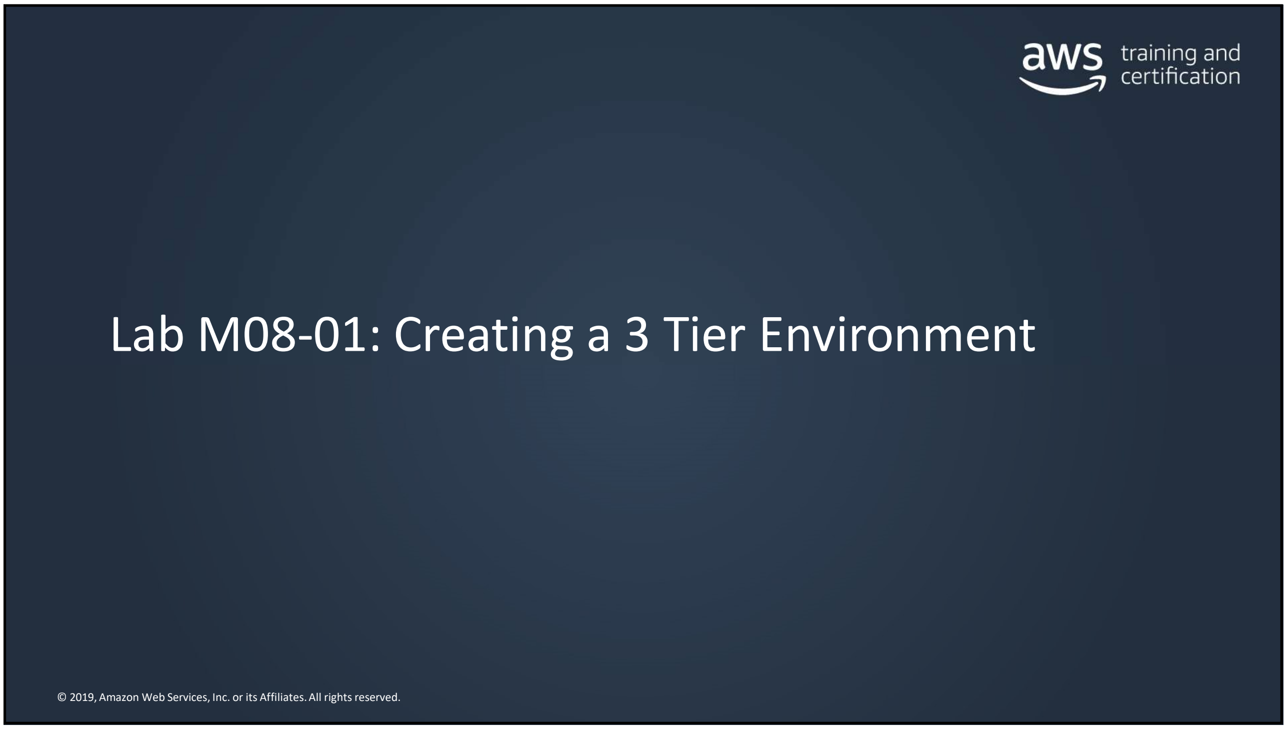
- Example: Users by last name, A to Z, in one database

With sharding, split your data into **large chunks** (shards)

- Example: Users by last name, A through M, in one database; N through Z in another database

In many circumstances, sharding gives you higher performance and better operating efficiency

38



Lab M08-01: Creating a 3 Tier Environment

39

Lab M08-01: Creating a 3 Tier Environment

aws training and certification

"I want 3-tier infrastructure."

Technologies used:

- Amazon VPC
- Application Load Balancer
- Amazon EC2 Auto Scaling group
- Amazon RDS
- Amazon Route 53

40

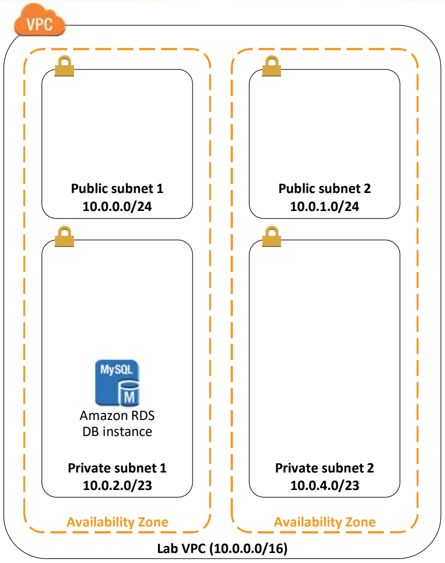
Lab M08-01: Creating a 3 Tier Environment

aws training and certification

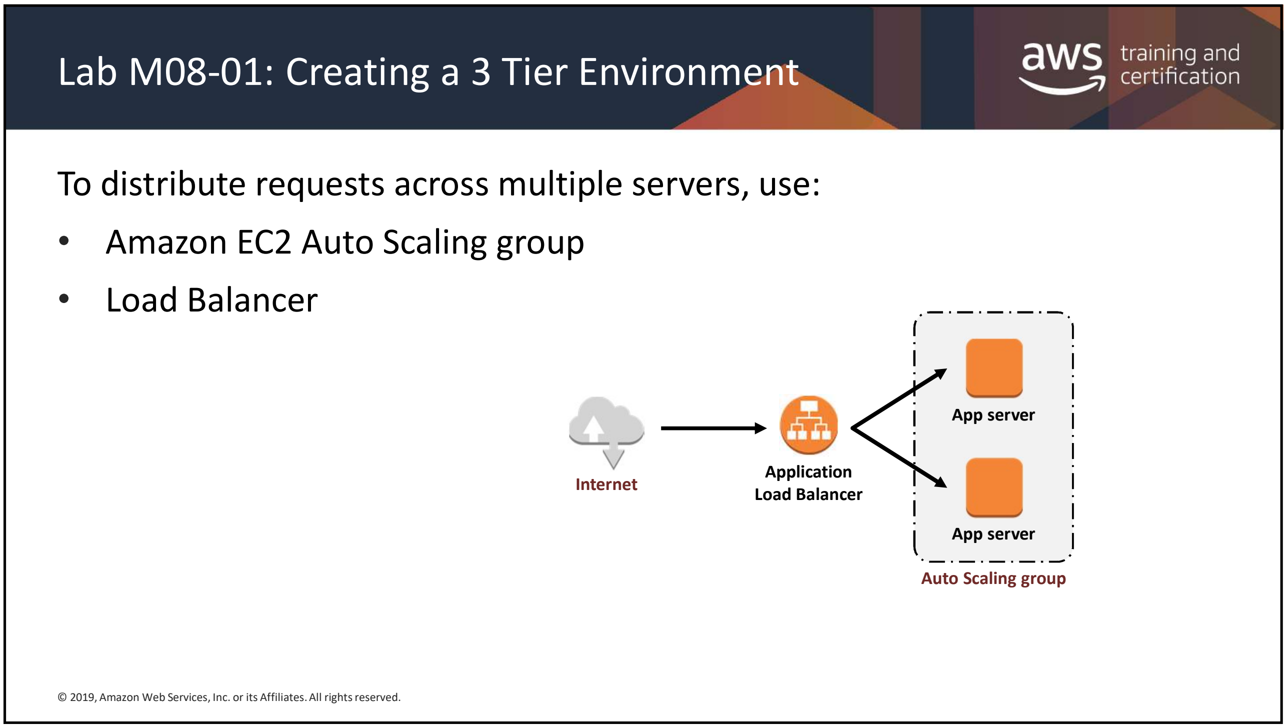
Provided at start of lab:

- VPC across two Availability Zones
- 2 x Public subnets
- 2 x DB private subnets
- Amazon RDS DB instance

You will make this highly available!



41



Lab M08-01: Creating a 3 Tier Environment

aws training and certification

To distribute requests across multiple servers, use:

- Amazon EC2 Auto Scaling group
- Load Balancer

Internet

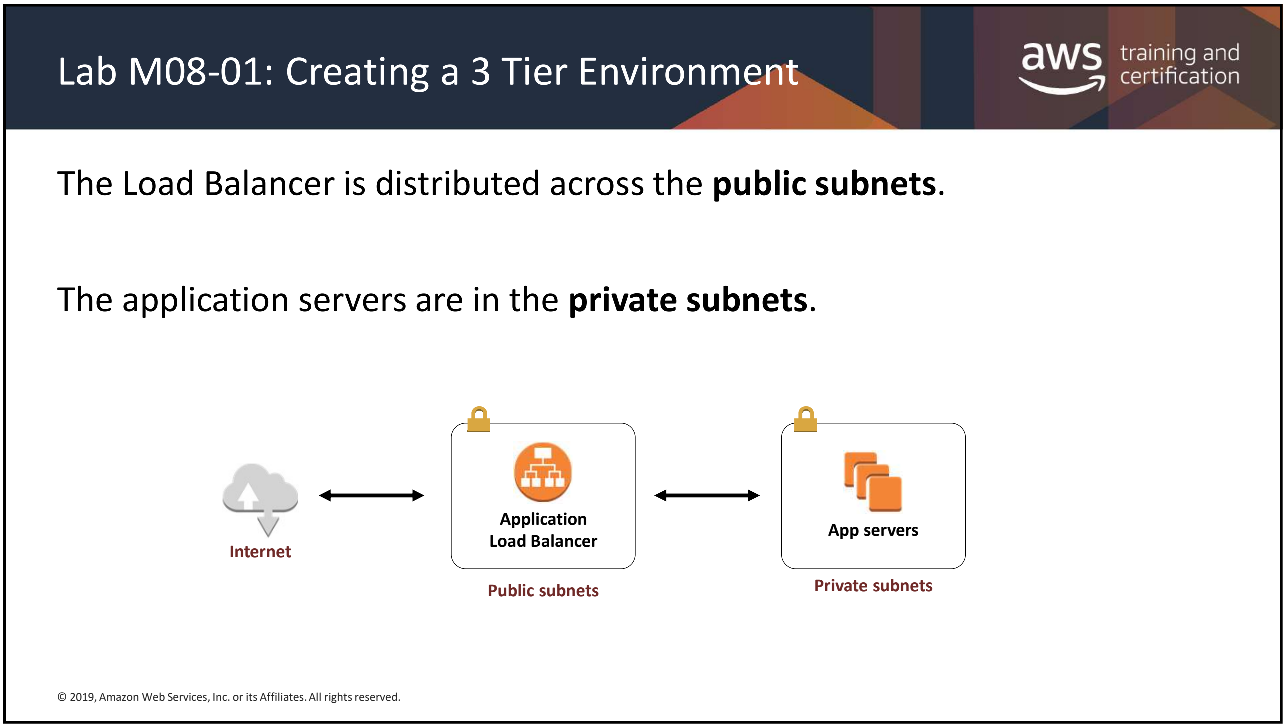
Application Load Balancer

App server

App server

Auto Scaling group

42



Lab M08-01: Creating a 3 Tier Environment

aws training and certification

The Load Balancer is distributed across the **public subnets**.

The application servers are in the **private subnets**.

Internet

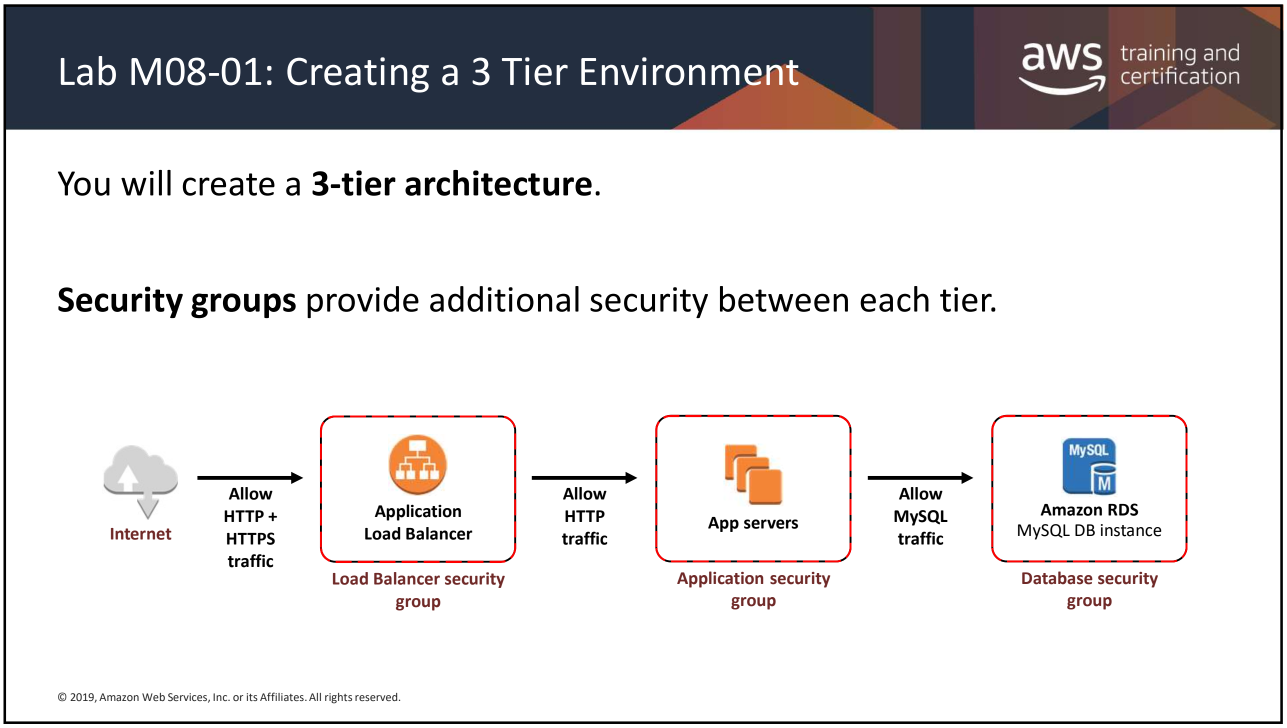
Application Load Balancer

Public subnets

App servers

Private subnets

43



Lab M08-01: Creating a 3 Tier Environment

aws training and certification

You will create a **3-tier architecture**.

Security groups provide additional security between each tier.

Internet

Allow HTTP + HTTPS traffic

Application Load Balancer

Load Balancer security group

Allow HTTP traffic

App servers

Application security group

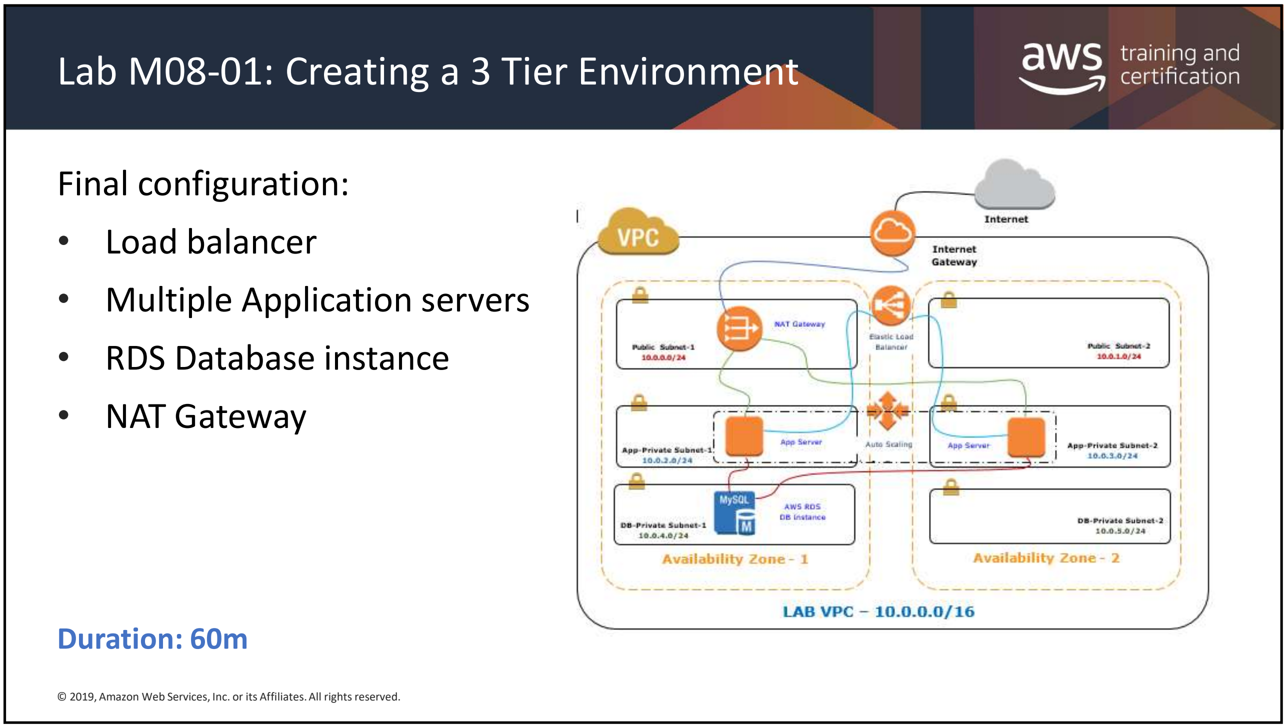
Allow MySQL traffic

MySQL

Amazon RDS MySQL DB instance

Database security group

44



Lab M08-01: Creating a 3 Tier Environment

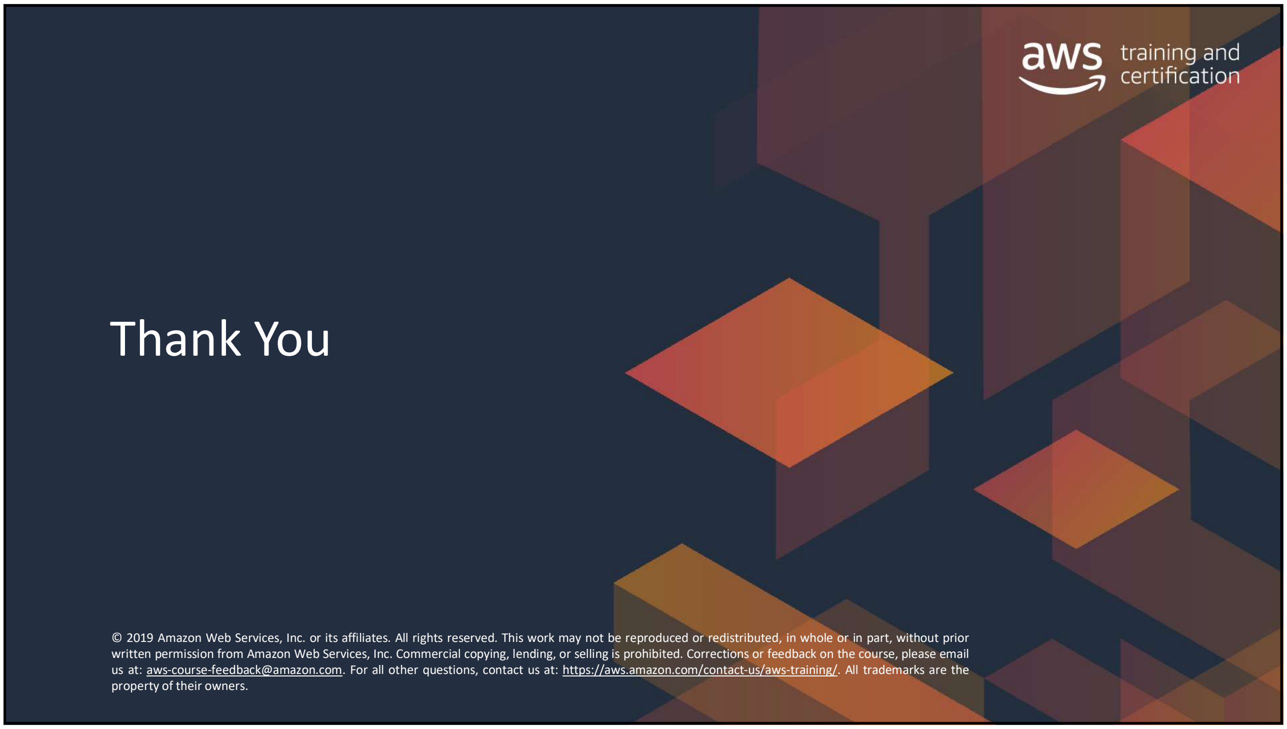
aws training and certification

Final configuration:

- Load balancer
- Multiple Application servers
- RDS Database instance
- NAT Gateway

Duration: 60m

45



aws training and certification

Thank You

46