Module 10: Automatic Scaling and Monitoring

AWS Academy Cloud Foundations

Module overview

Topics

- Elastic Load Balancing
- Amazon CloudWatch
- Amazon EC2 Auto Scaling

Activities

- Elastic Load Balancing activity
- Amazon CloudWatch activity

Lab

- Scale and Load Balance Your Architecture

☑ **Knowledge check**

Module objectives

After completing this module, you should be able to:

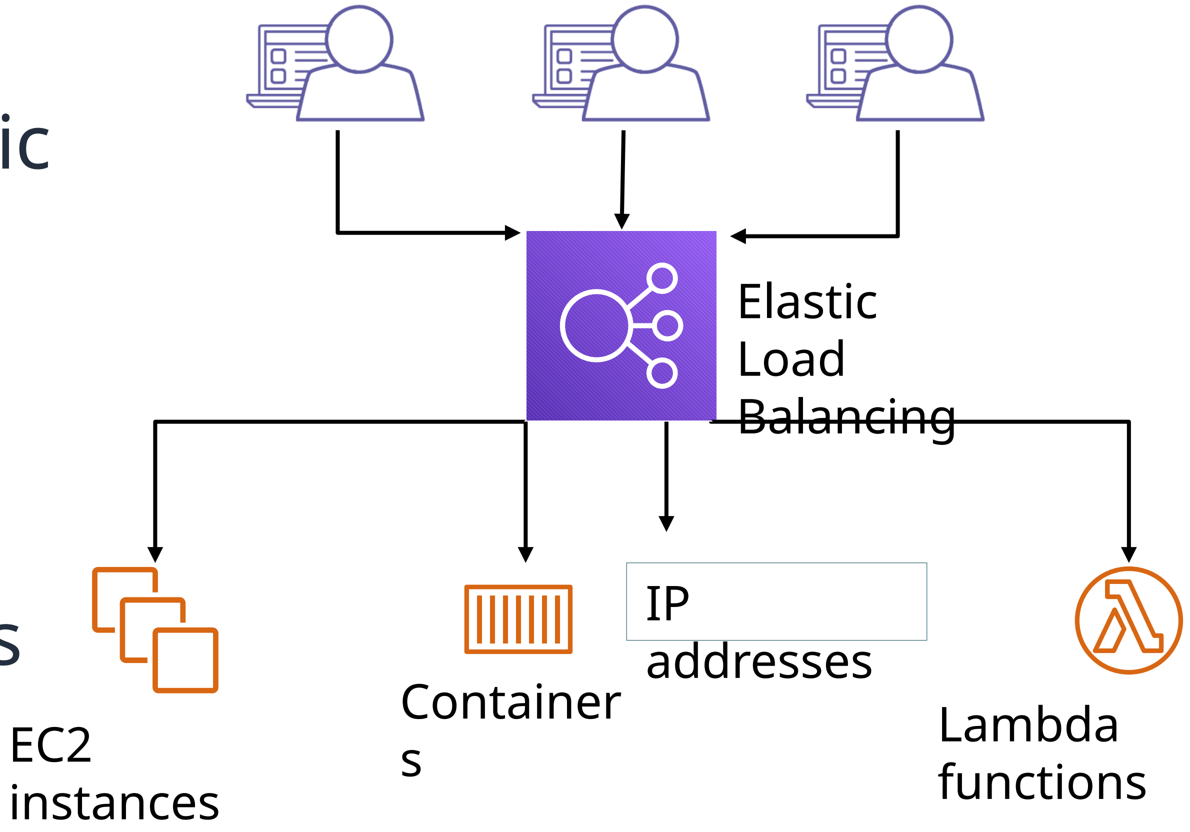
- Indicate how to distribute traffic across Amazon Elastic Compute Cloud (Amazon EC2) instances by using Elastic Load Balancing
- Identify how Amazon CloudWatch enables you to monitor AWS resources and applications in real time
- Explain how Amazon EC2 Auto Scaling launches and releases servers in response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

Section 1: Elastic Load Balancing

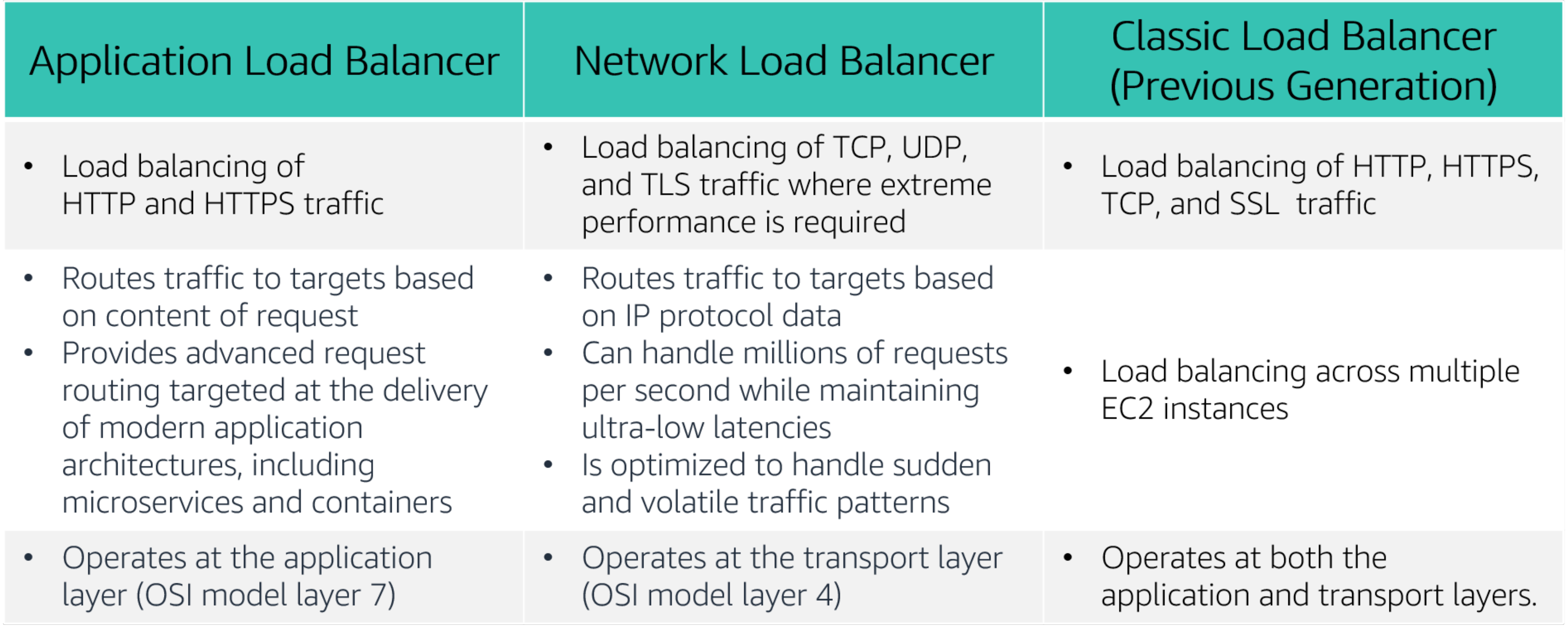
Module 10: Automatic Scaling and Monitoring

Elastic Load Balancing

- Distributes incoming application or network traffic across multiple targets in a single Availability Zone or across multiple Availability Zones.
- Scales your load balancer as traffic to your application changes over time.



Types of load balancers

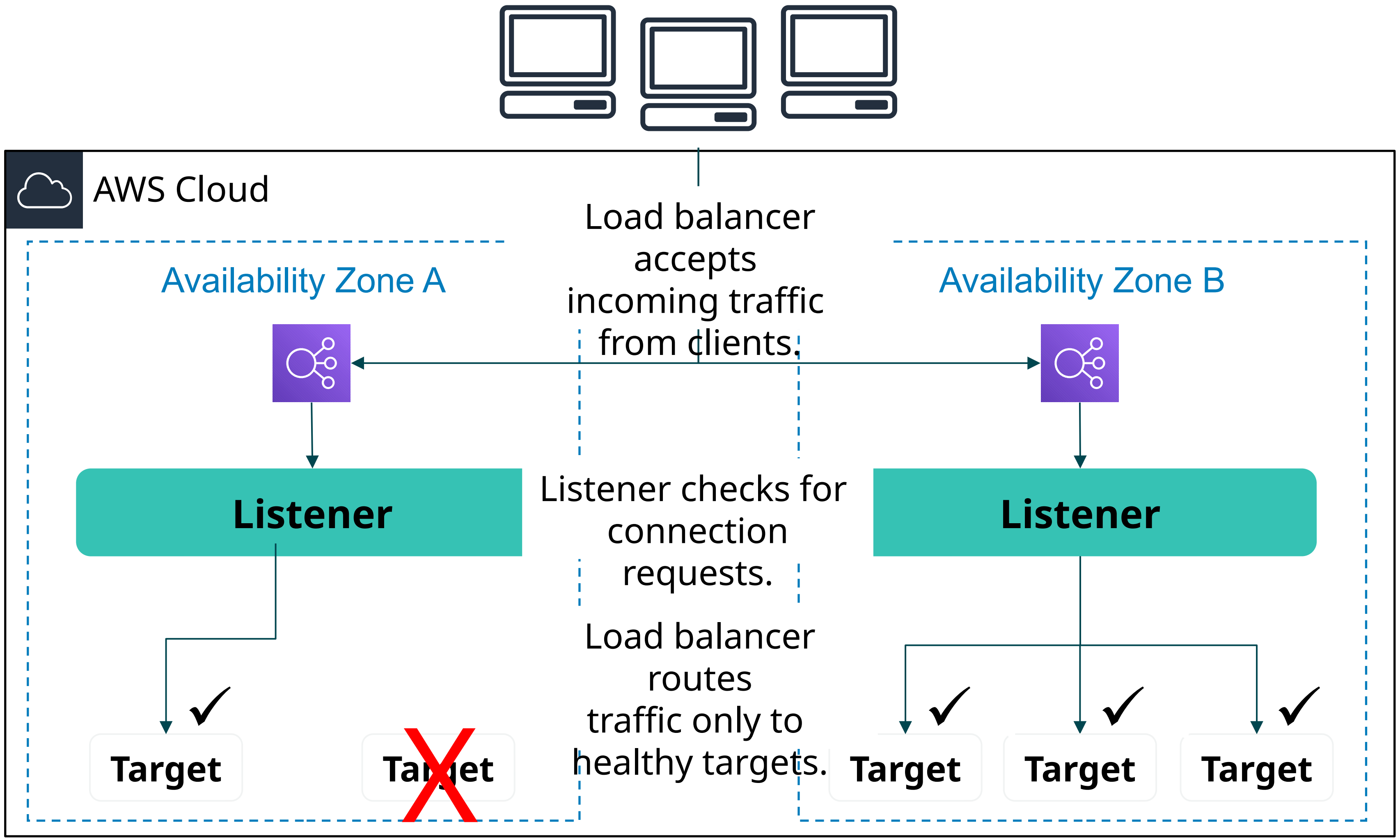


Application Load Balancer	Network Load Balancer	Classic Load Balancer (Previous Generation)
- Load balancing of HTTP and HTTPS traffic	- Load balancing of TCP, UDP, and TLS traffic where extreme performance is required	- Load balancing of HTTP, HTTPS, TCP, and SSL traffic
- Routes traffic to targets based on content of request - Provides advanced request routing 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 - Is optimized to handle sudden and volatile traffic patterns	- Load balancing across multiple EC2 instances
- Operates at the application layer (OSI model layer 7)	- Operates at the transport layer (OSI model layer 4)	- Operates at both the application and transport layers.

How Elastic Load Balancing works

- 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.



Load balancer accepts incoming traffic from clients.

Listener checks for connection requests.

Load balancer routes traffic only to healthy targets.

Load balancer performs health checks to monitor health of registered targets.

Elastic Load Balancing use cases

Highly available and fault-tolerant applications

Containerized applications

Elasticity and scalability

Virtual private cloud (VPC)

Hybrid environments

Invoke Lambda functions over HTTP(S)

Activity: Elastic Load Balancing

You must support traffic to a containerized application.

You have extremely spiky and unpredictable TCP traffic.

You need simple load balancing with multiple protocols.

You need to support a static or Elastic IP address, or an IP target outside a VPC.

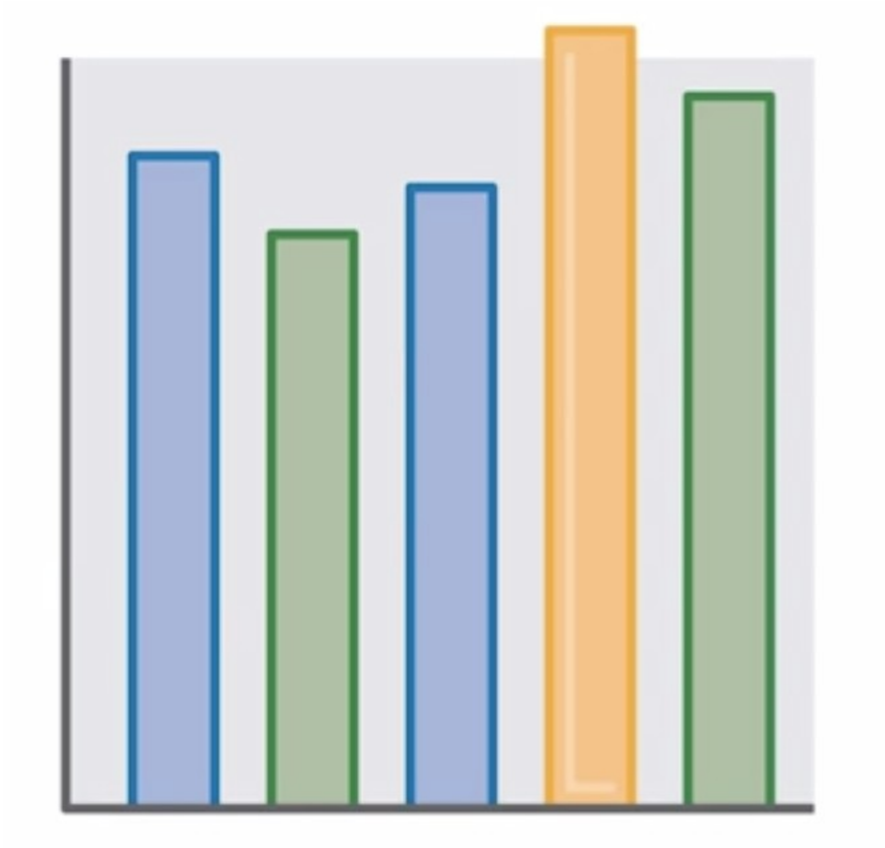
You need a load balancer that can handle millions of requests per second while maintaining low latencies.

You must support HTTPS requests.

Activity: Elastic Load Balancing Answers

You must support traffic to a containerized application.	Application Load Balancer
You have extremely spiky and unpredictable TCP traffic.	Network Load Balancer
You need simple load balancing with multiple protocols.	Classic Load Balancer
You need to support a static or Elastic IP address, or an IP target outside a VPC.	Network Load Balancer
You need a load balancer that can handle millions of requests per second while maintaining low latencies.	Network Load Balancer
You must support HTTPS requests.	Application Load Balancer

Load balancer monitoring



- **Amazon CloudWatch metrics** – Used to verify that the system is performing as expected and creates an alarm to initiate an action if a metric goes outside an acceptable range.
- **Access logs** – Capture detailed information about requests sent to your load balancer.
- **AWS CloudTrail logs** – Capture the who, what, when, and where of API interactions in AWS services.

Section 1 key takeaways

- Elastic Load Balancing distributes incoming application or network traffic across multiple targets in one or more Availability Zones.
- Elastic Load Balancing supports three types of load balancers:
  - Application Load Balancer
  - Network Load Balancer
  - Classic Load Balancer
- ELB offers instance health checks, security, and monitoring.

Section 2: Amazon CloudWatch

Module 10: Automatic Scaling and Monitoring

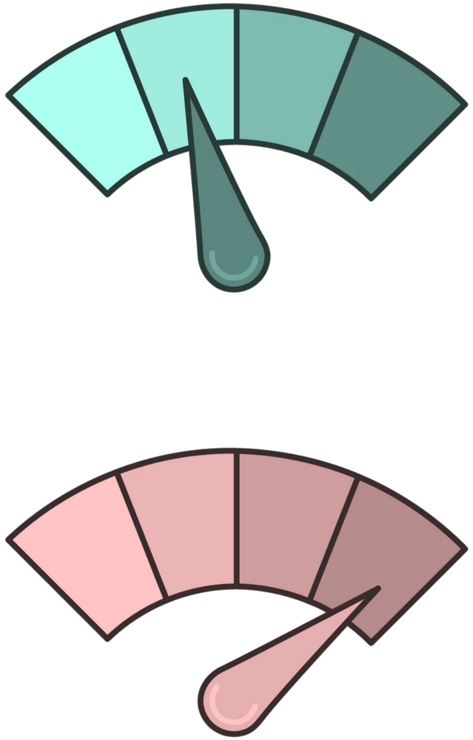
Monitoring AWS resources

To use AWS efficiently, you need insight into your AWS resources:

- How do you know when you should **launch more Amazon EC2 instances**?
- Is your **application's performance or availability** being affected by a lack of sufficient capacity?
- How much of your infrastructure is actually **being used**?

Amazon CloudWatch

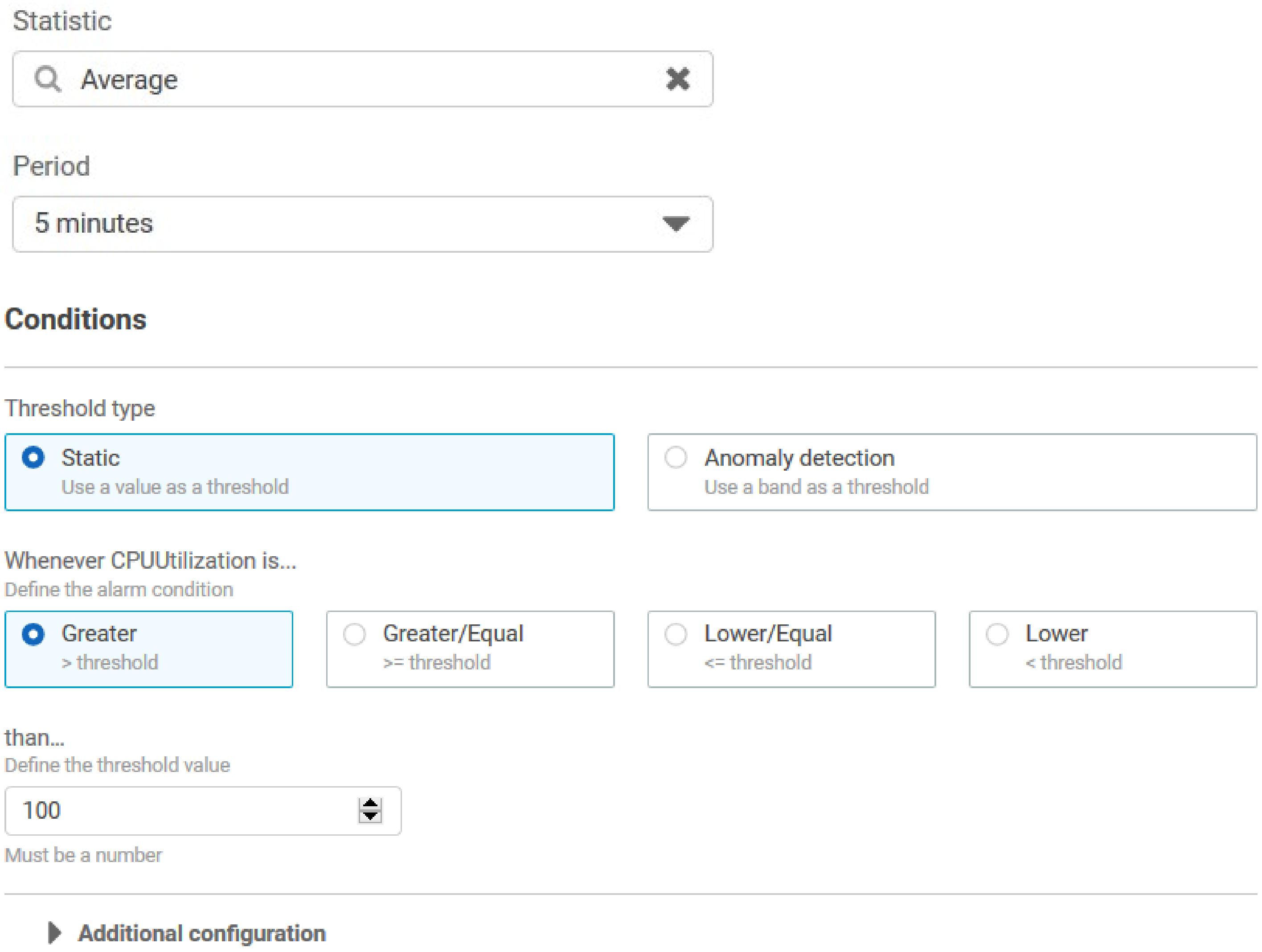
Amazon CloudWatch



- Monitors –
  - AWS resources
  - Applications that run on AWS
- Collects and tracks –
  - Standard metrics
  - Custom metrics
- Alarms –
  - Send notifications to an Amazon SNS topic
  - Perform Amazon EC2 Auto Scaling or Amazon EC2 actions
- Events –
  - Define rules to match changes in AWS environment and route these events to one or more target functions or streams for processing

CloudWatch alarms

- Create alarms based on –
  - Static threshold
  - Anomaly detection
  - Metric math expression
- Specify –
  - Namespace
  - Metric
  - Statistic
  - Period
  - Conditions
  - Additional configuration
  - Actions



Statistic

Average

Period

5 minutes

Conditions

Threshold type

● Static
Use a value as a threshold

○ Anomaly detection
Use a band as a threshold

Whenever CPUUtilization is...
Define the alarm condition

● Greater
> threshold

○ Greater/Equal
>= threshold

○ Lower/Equal
<= threshold

○ Lower
< threshold

than...
Define the threshold value

100

Must be a number

▶ **Additional configuration**

Activity: Amazon CloudWatch

Amazon EC2

If average CPU utilization is > 60% for 5 minutes…

Amazon RDS

If the number of simultaneous connections is > 10 for 1 minute…

Amazon S3

If the maximum bucket size in bytes is around 3 for 1 day…

Elastic Load Balancing

If the number of healthy hosts is < 5 for 10 minutes…

Amazon Elastic Block Store

If the volume of read operations is > 1,000 for 10 seconds…

Activity: Amazon CloudWatch Answers

Amazon EC2	If average CPU utilization is > 60% for 5 minutes…	Correct!
Amazon RDS	If the number of simultaneous connections is > 10 for 1 minute…	Correct!
Amazon S3	If the maximum bucket size in bytes is around 3 for 1 day…	Incorrect. *Around* is not a threshold option. You must specify a threshold of >, >=, <=, or <.
Elastic Load Balancing	If the number of healthy hosts is < 5 for 10 minutes…	Correct!
Amazon Elastic Block Store	If the volume of read operations is > 1,000 for 10 seconds…	Incorrect. You must specify a statistic (for example, *average volume*).

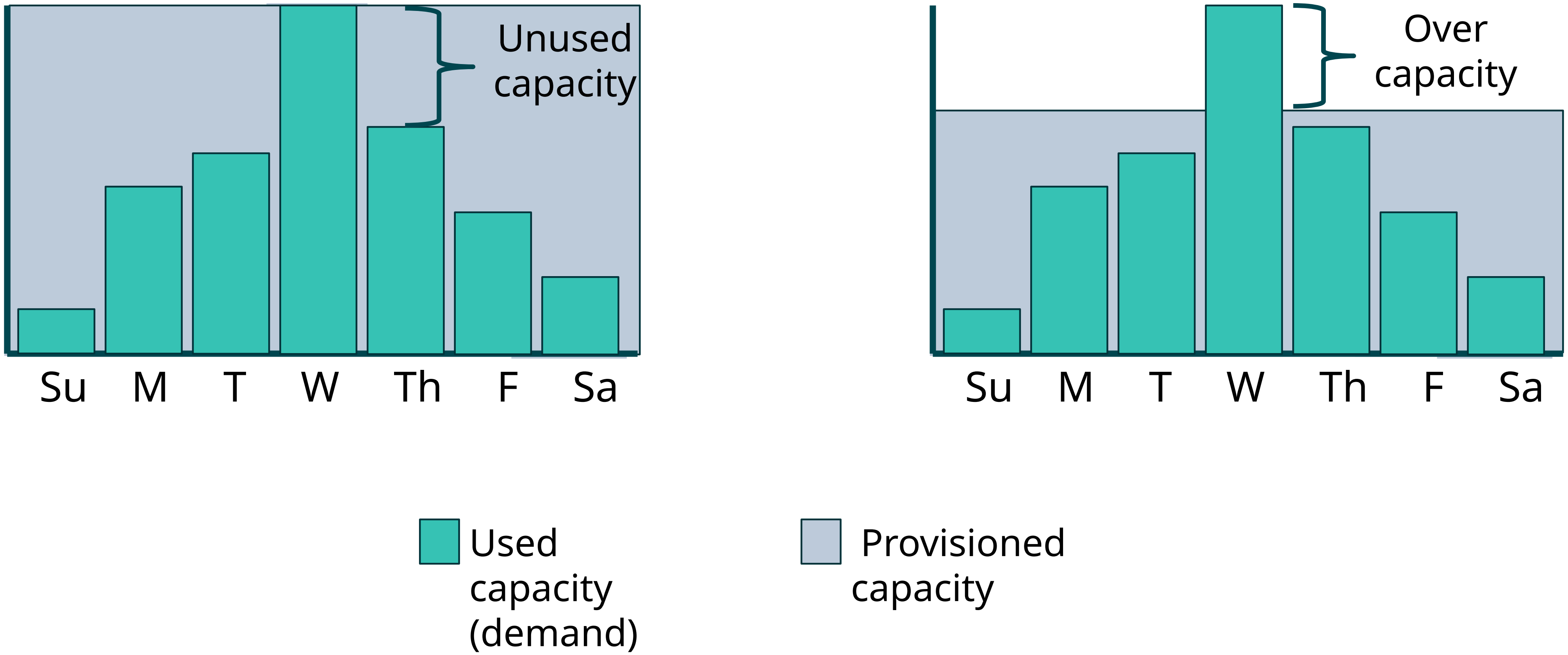
Section 2 key takeaways

- Amazon CloudWatch helps you monitor your AWS resources—and the applications that you run on AWS—in real time.
- CloudWatch enables you to –
  - Collect and track standard and custom metrics.
  - Set alarms to automatically send notifications to SNS topics, or perform Amazon EC2 Auto Scaling or Amazon EC2 actions.
  - Define rules that match changes in your AWS environment and route these events to targets for processing.

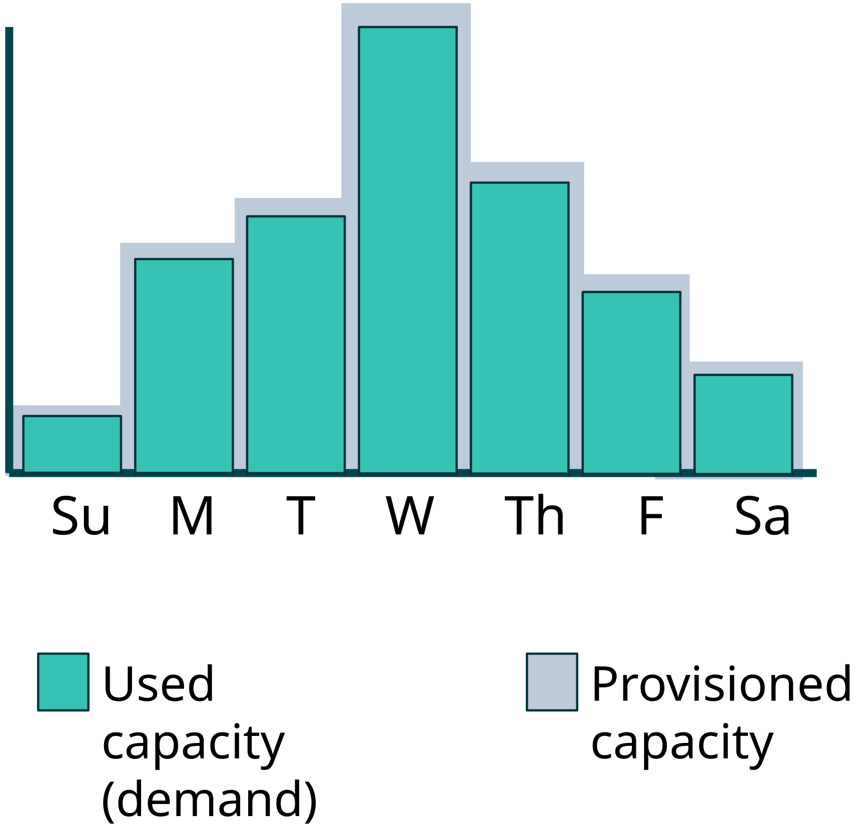
Section 3: Amazon EC2 Auto Scaling

Module 10: Automatic Scaling and Monitoring

Why is scaling important?

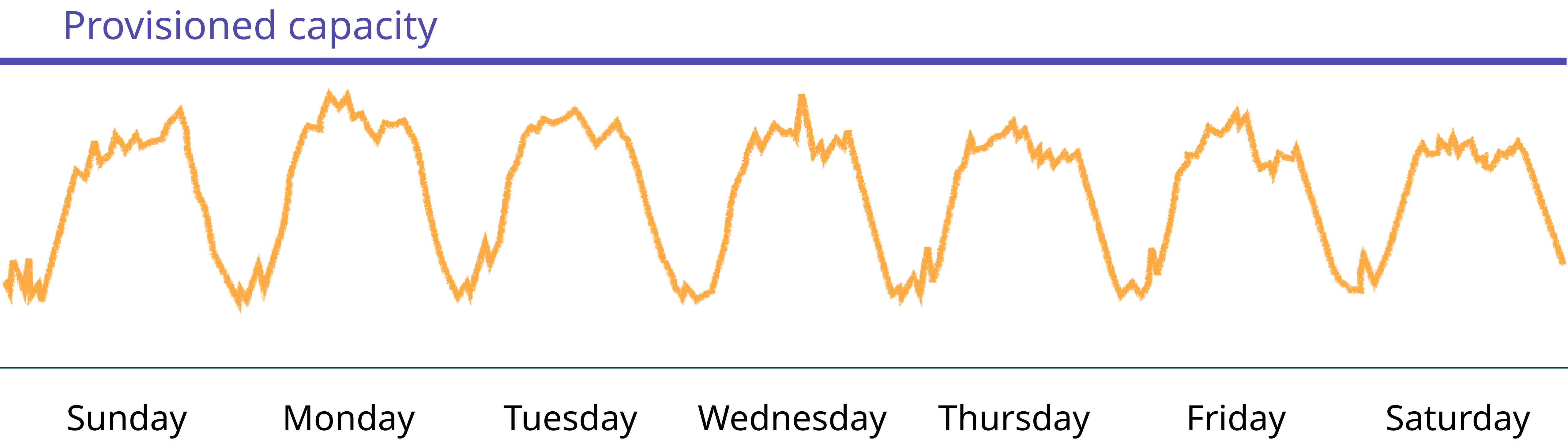


Amazon EC2 Auto Scaling



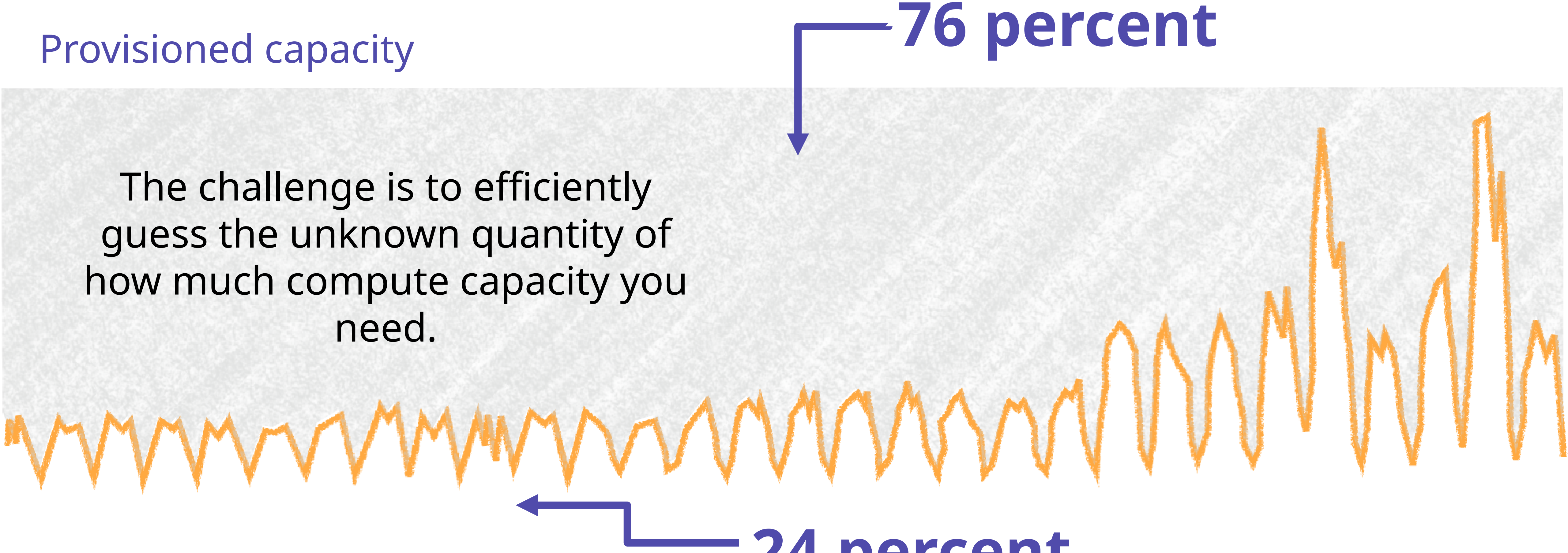
- Helps you maintain application availability
- Enables you to automatically add or remove EC2 instances according to conditions that you define
- Detects impaired EC2 instances and unhealthy applications, and replaces the instances without your intervention
- Provides several scaling options – Manual,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

Typical weekly traffic at Amazon.com



November traffic to Amazon.com

Provisioned capacity



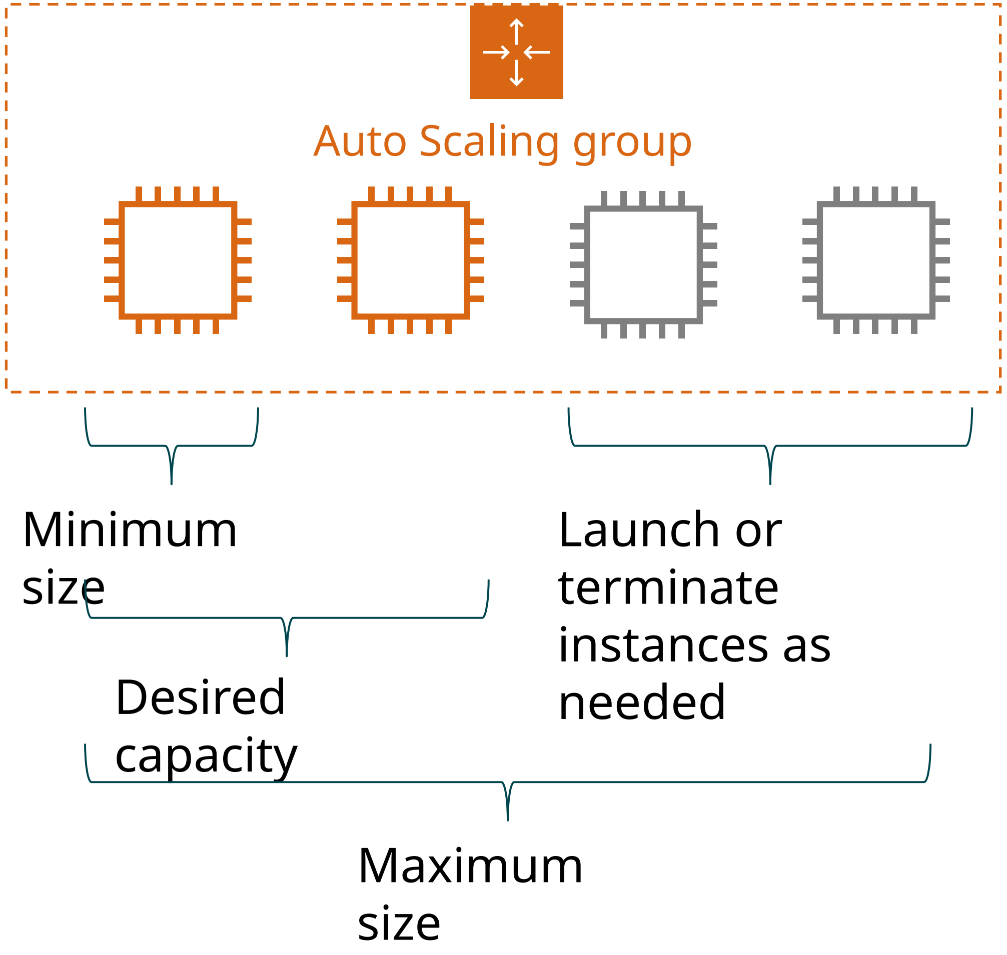
76 percent

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

24 percent

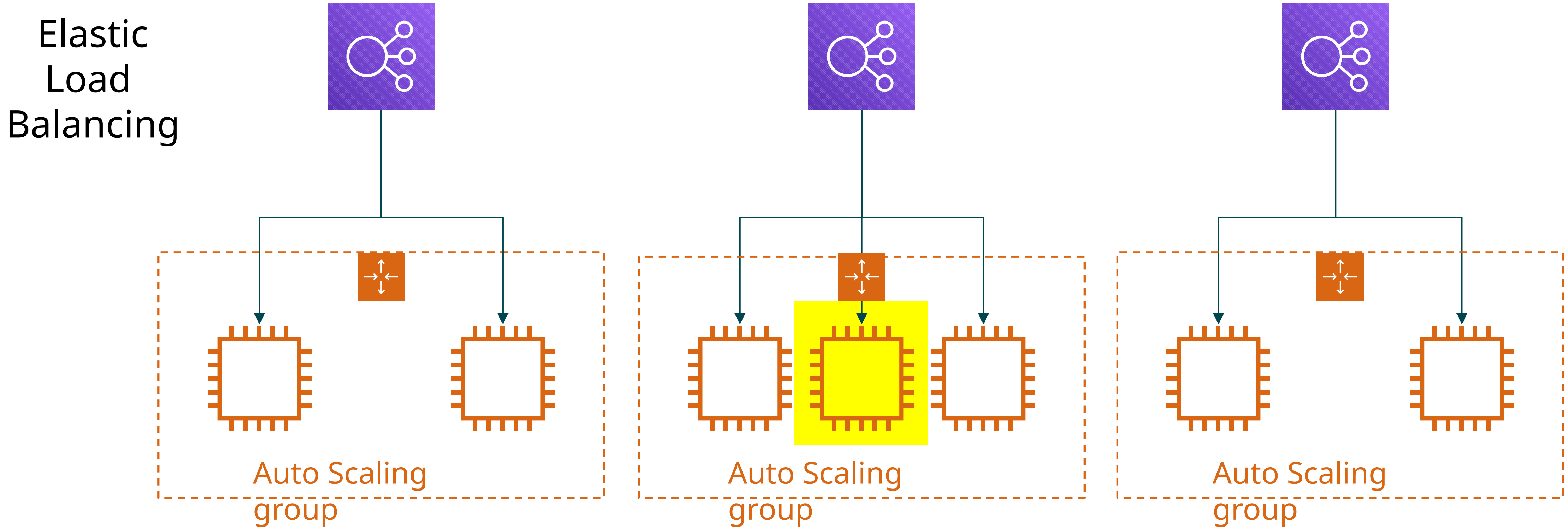
November

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



Elastic Load Balancing

Auto Scaling group

Auto Scaling group

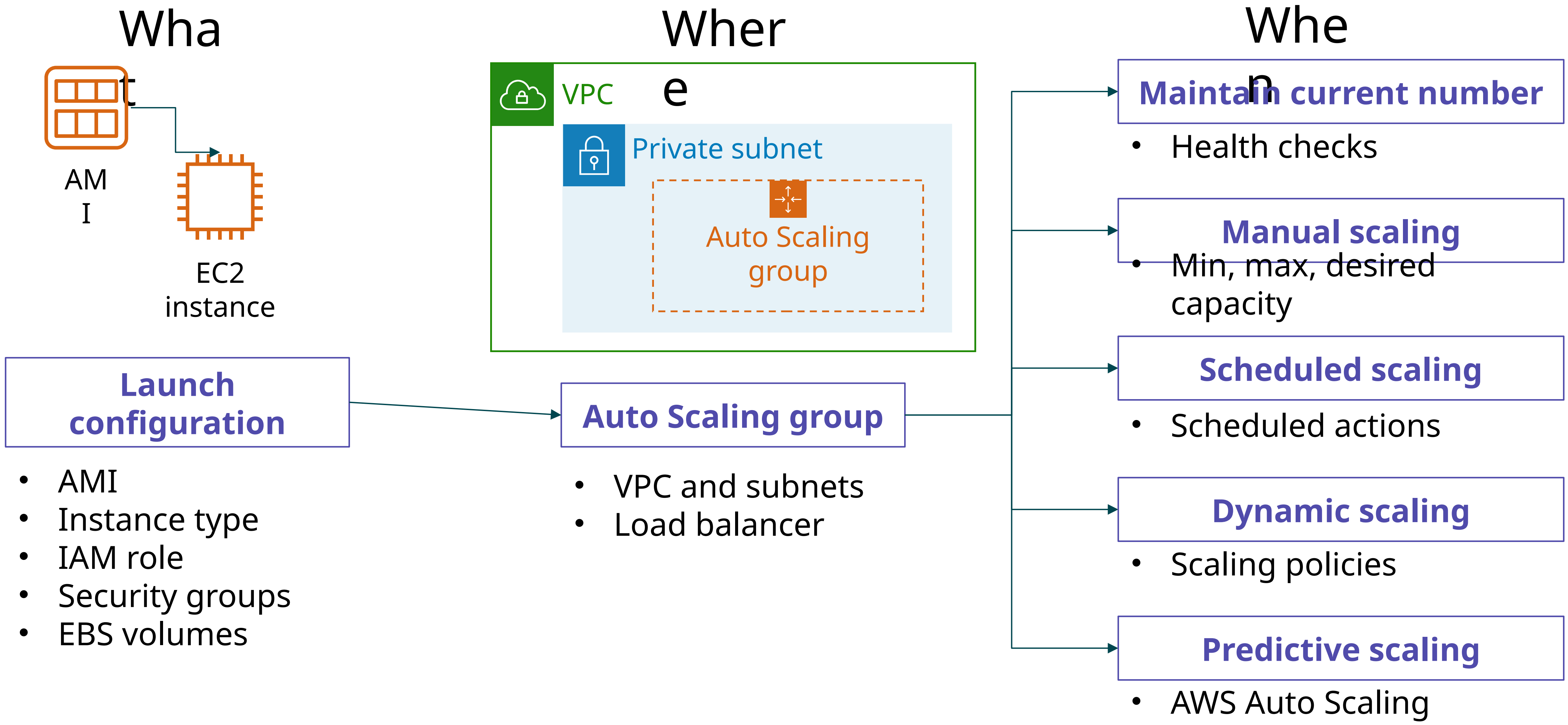
Auto Scaling group

Base configuration

Scale out (launch instances)

Scale in (terminate instances)

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

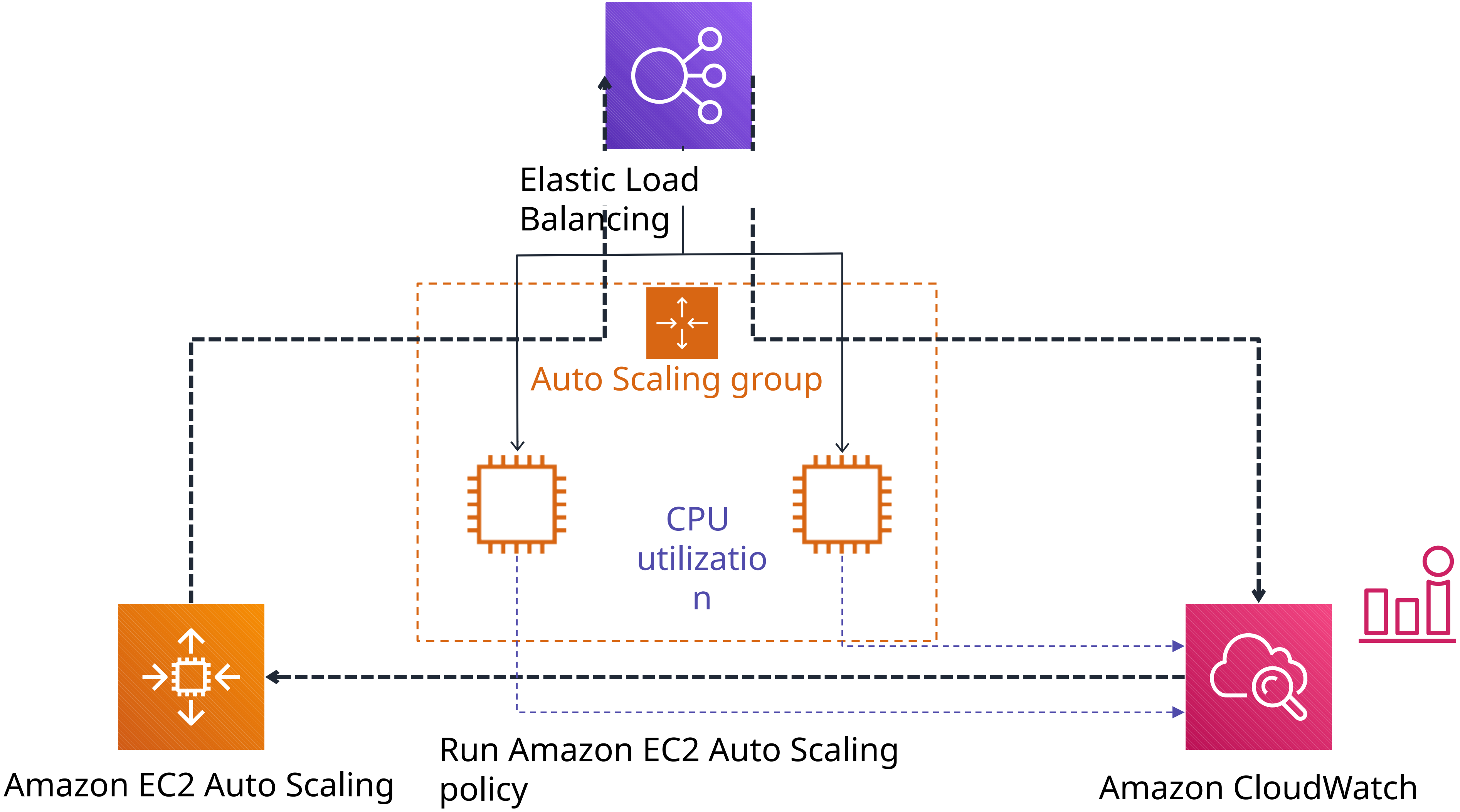
Dynamic scaling

- Scaling policies

Predictive scaling

- AWS Auto Scaling

Implementing dynamic scaling



If average CPU utilization is > 60% for 5 minutes…

AWS Auto Scaling

AWS Auto Scaling

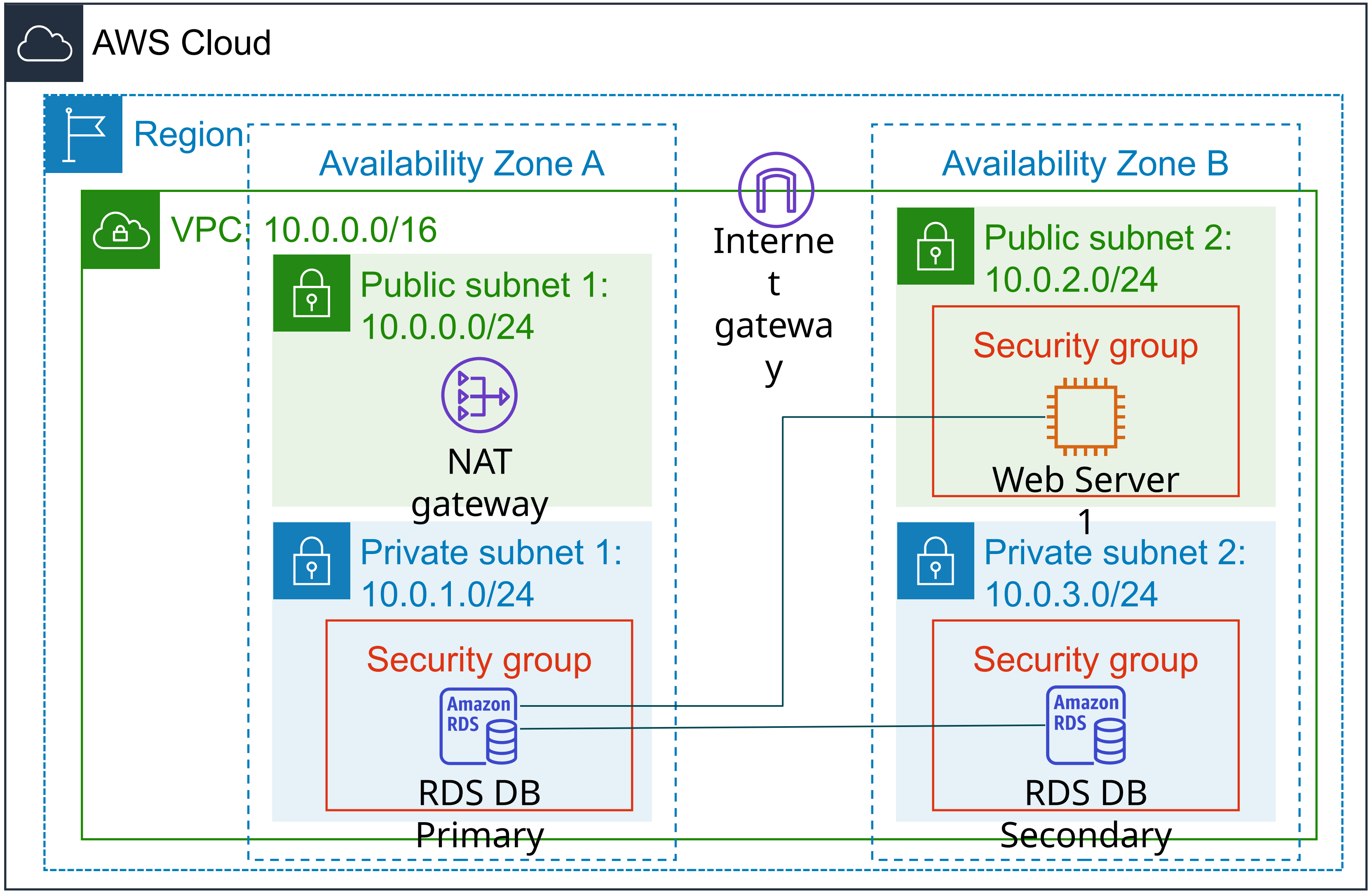
- Monitors your applications and automatically adjusts capacity to maintain steady, predictable performance at the lowest possible cost
- Provides a simple, powerful user interface that enables you to build scaling plans for resources, including –
  - Amazon EC2 instances and Spot Fleets
  - Amazon Elastic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

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.

Lab 6: Scale and Load Balance Your Architecture

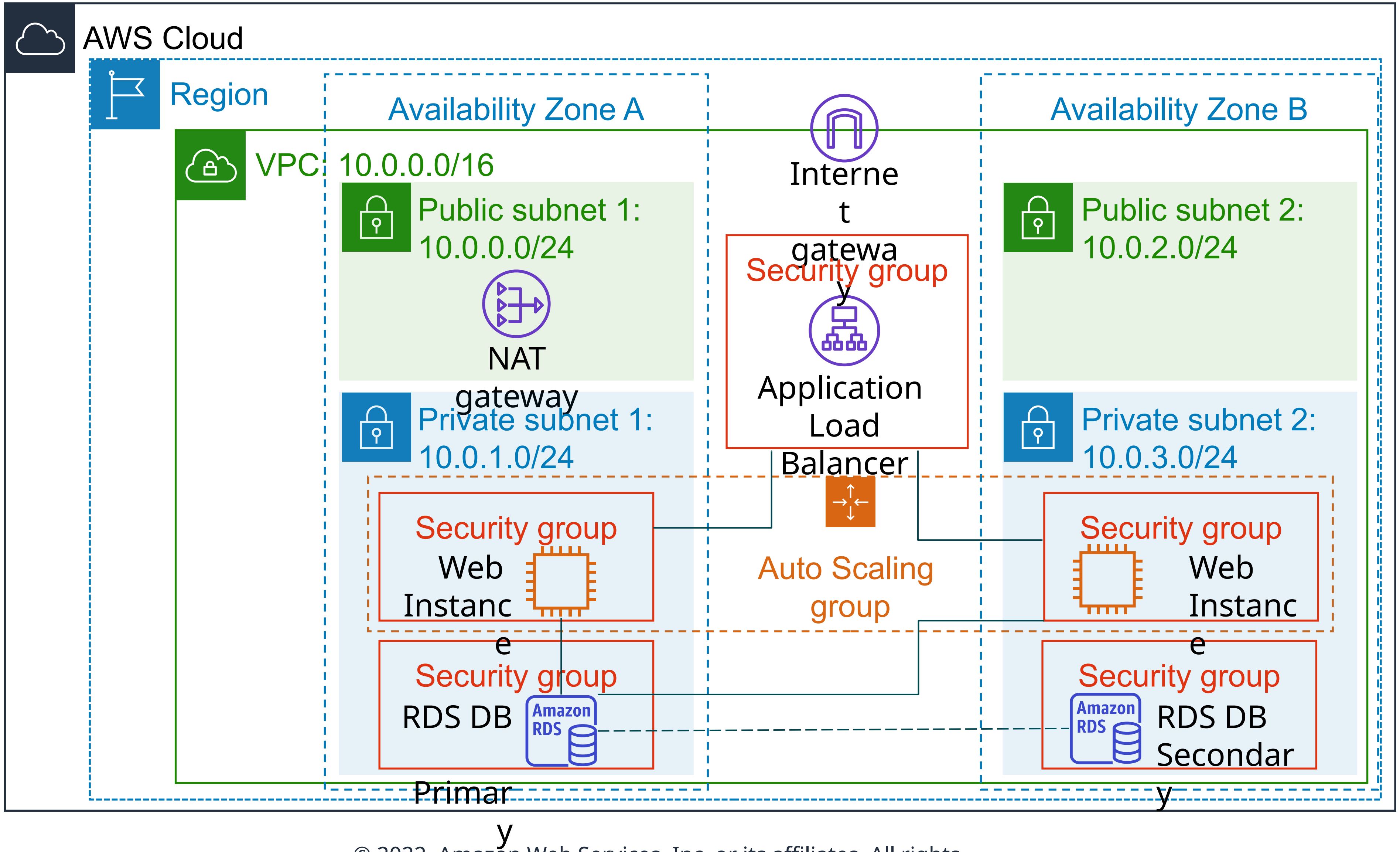
Lab 6: Scenario



Lab 6: Tasks

- Create an Amazon Machine Image (AMI) from a running instance.
- Create an Application Load Balancer.
- Create a launch configuration and an Auto Scaling group.
- Automatically scale new instances within a private subnet.
- Create Amazon CloudWatch alarms and monitor performance of your infrastructure.

Lab 6: Final product



~ 30 minutes

Begin Lab 6: Scale and Load Balance Your Architecture

Lab debrief: Key takeaways

Module wrap-up

Module 10: Automatic Scaling and Monitoring

Module summary

In summary, in this module you learned how to:

- Indicate how to distribute traffic across Amazon Elastic Compute Cloud (Amazon EC2) instances 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.

Complete the knowledge check

```
<div class="wrap">
    <div id="content">
        <header id="topnav">
            <nav>
                <ul>
                    <li class="active"><a class="scroll" href=
                        Home                           </a></li>
                    <li><a class="scroll" href="#service">
                        Service                        </a></li>
                    <li><a class="scroll" href="#product">
                        Products                             </a></li>
                    <li><a class="scroll" href="#portfolio">
                        Portfolio                            </a></li>
                    <li><a class="scroll" href="#team">
                        Team                           </a></li>
                    <li><a class="scroll" href="#contact">
                        Contact                        </a></li>
                    <div class="clear"></div>
                </ul>
            </nav>
            <div class="logo">
                <a href="./">
                    <img src="web/images/logo3.png" id="logo_l
```

Sample exam question

Which service would you use to send alerts based on Amazon CloudWatch alarms?

Choice	Response
A	Amazon Simple Notification Service
B	AWS CloudTrail
C	AWS Trusted Advisor
D	Amazon Route 53

Sample exam question answer

Which service would you use to send alerts based on Amazon CloudWatch alarms?

The correct answer is A.

The keywords in the question are “send alerts” and “Amazon CloudWatch alarms”.

Thank you

All trademarks are the property of their owners.