Module 10: Automatic Scaling and Monitoring

Module overview

Topics

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

☑ **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

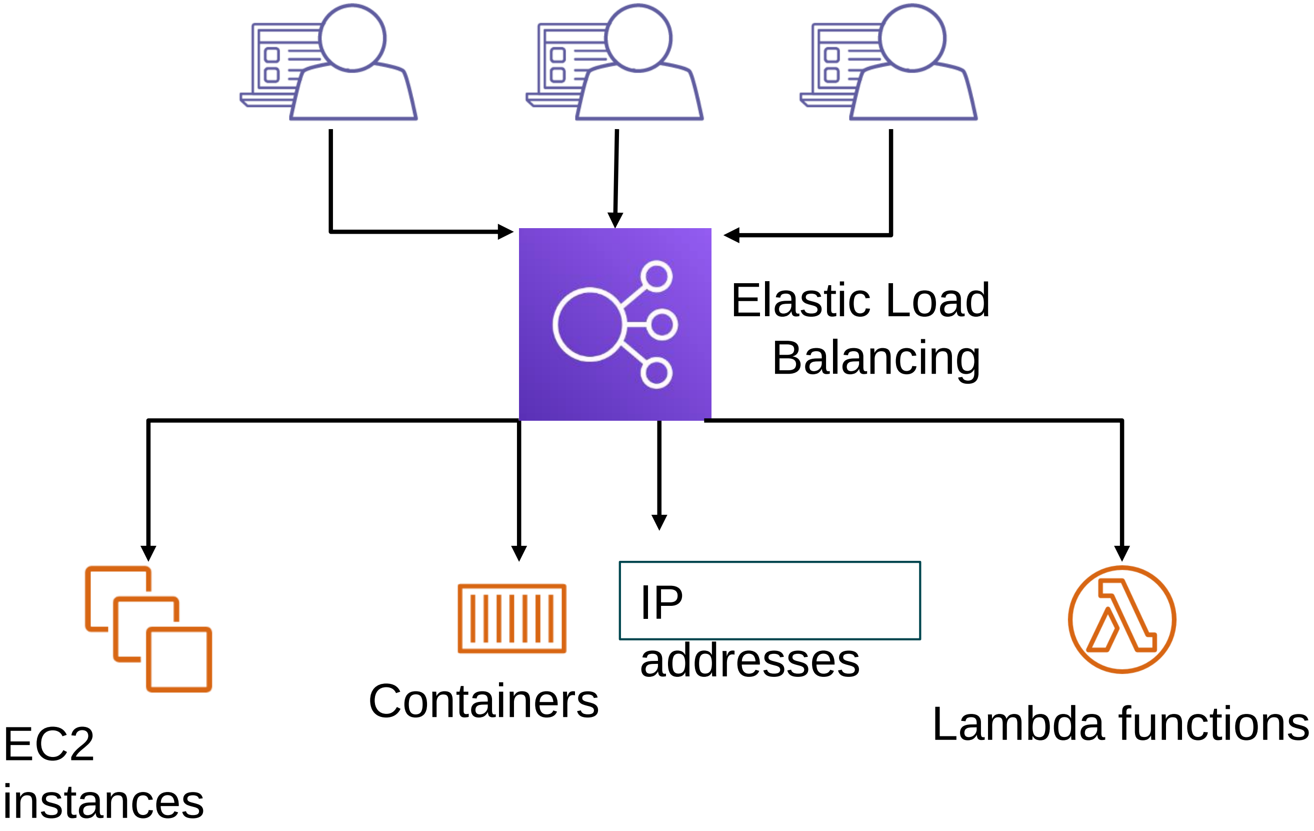
Section 1: Elastic Load Balancing

Module 10: Automatic Scaling and Monitoring

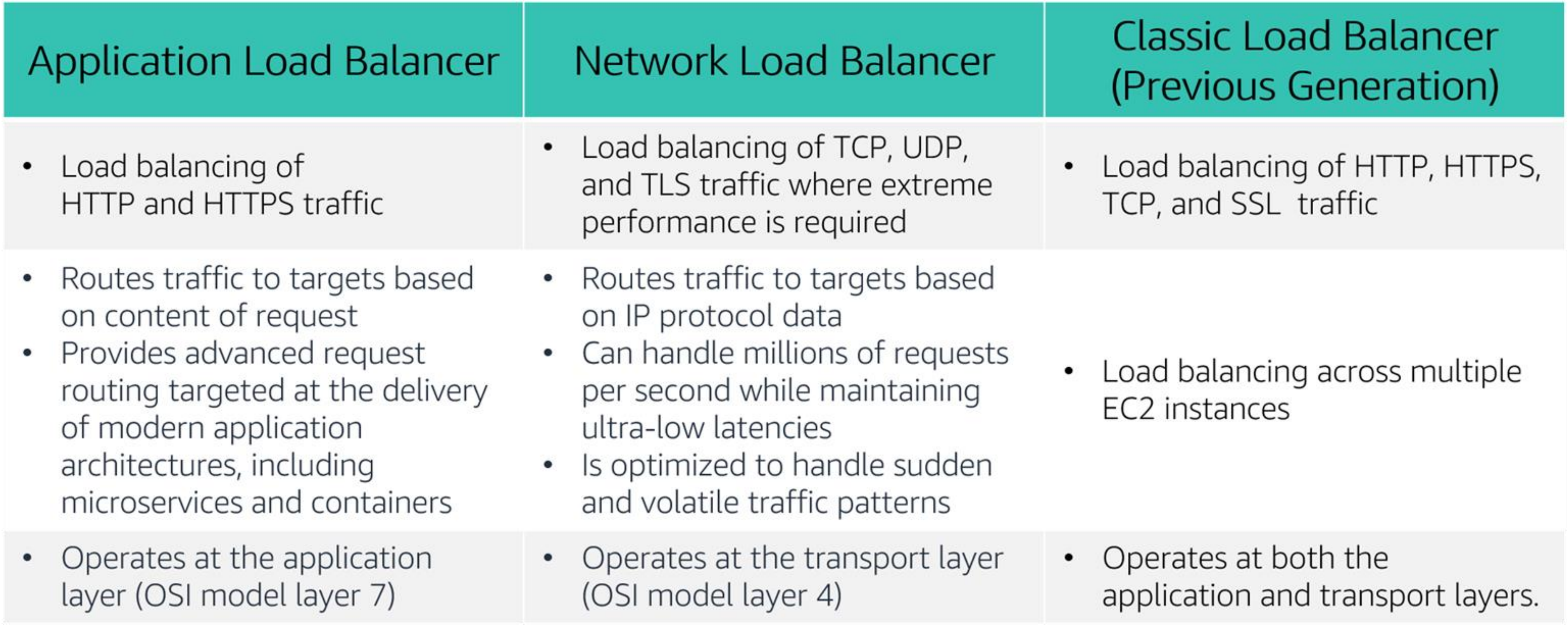
Elastic Load Balancing

- 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.



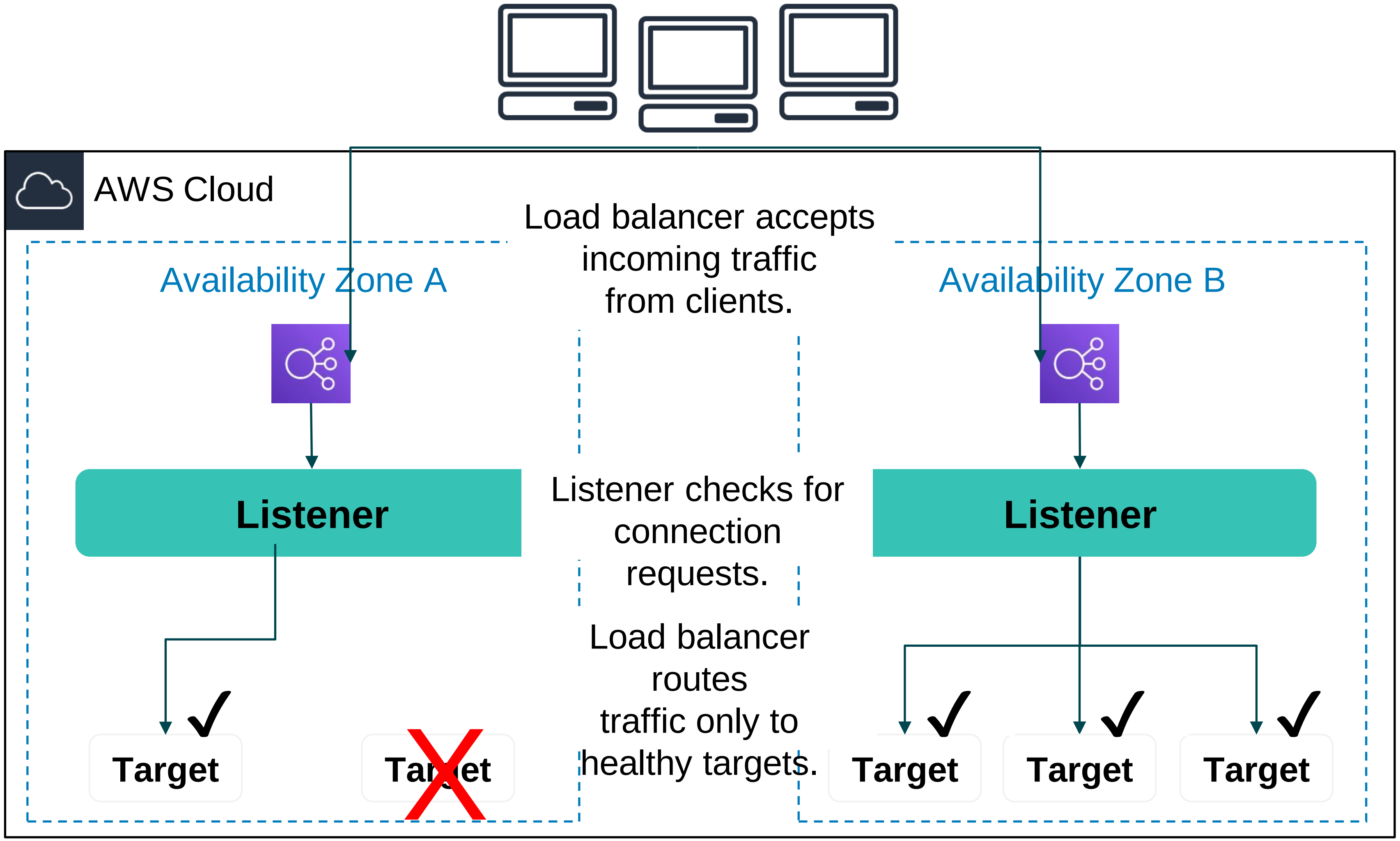
Types of load balancers



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.


AWS Cloud
Availability Zone A
Load balancer accepts incoming traffic from clients.
Availability Zone B
Listener
Listener checks for connection requests.
Listener
Load balancer routes traffic only to healthy targets.
Target ✓
Target ✗
Target ✓
Target ✓
Target ✓

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.

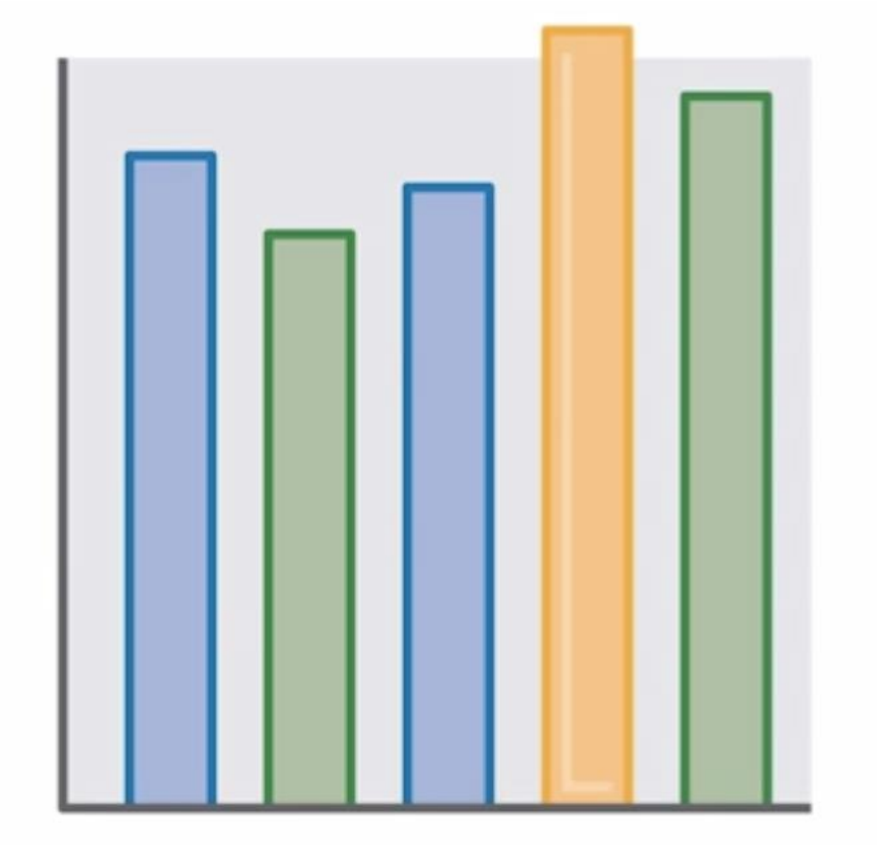
You need a load balancer that can handle millions of requests per second while maintaining low latencies.

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

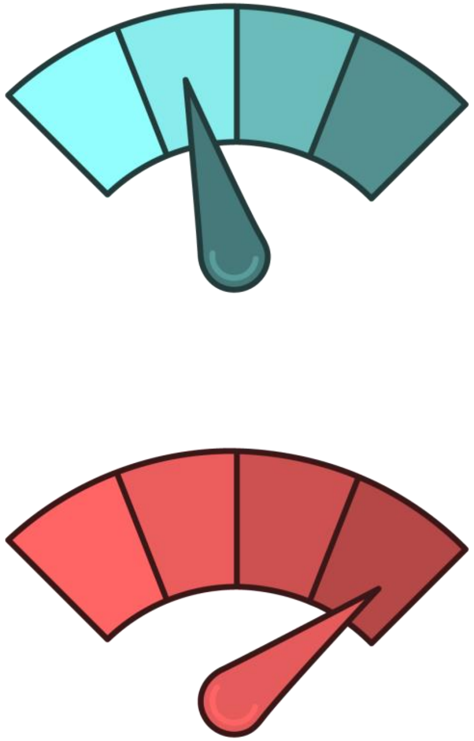
Monitoring AWS resources

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

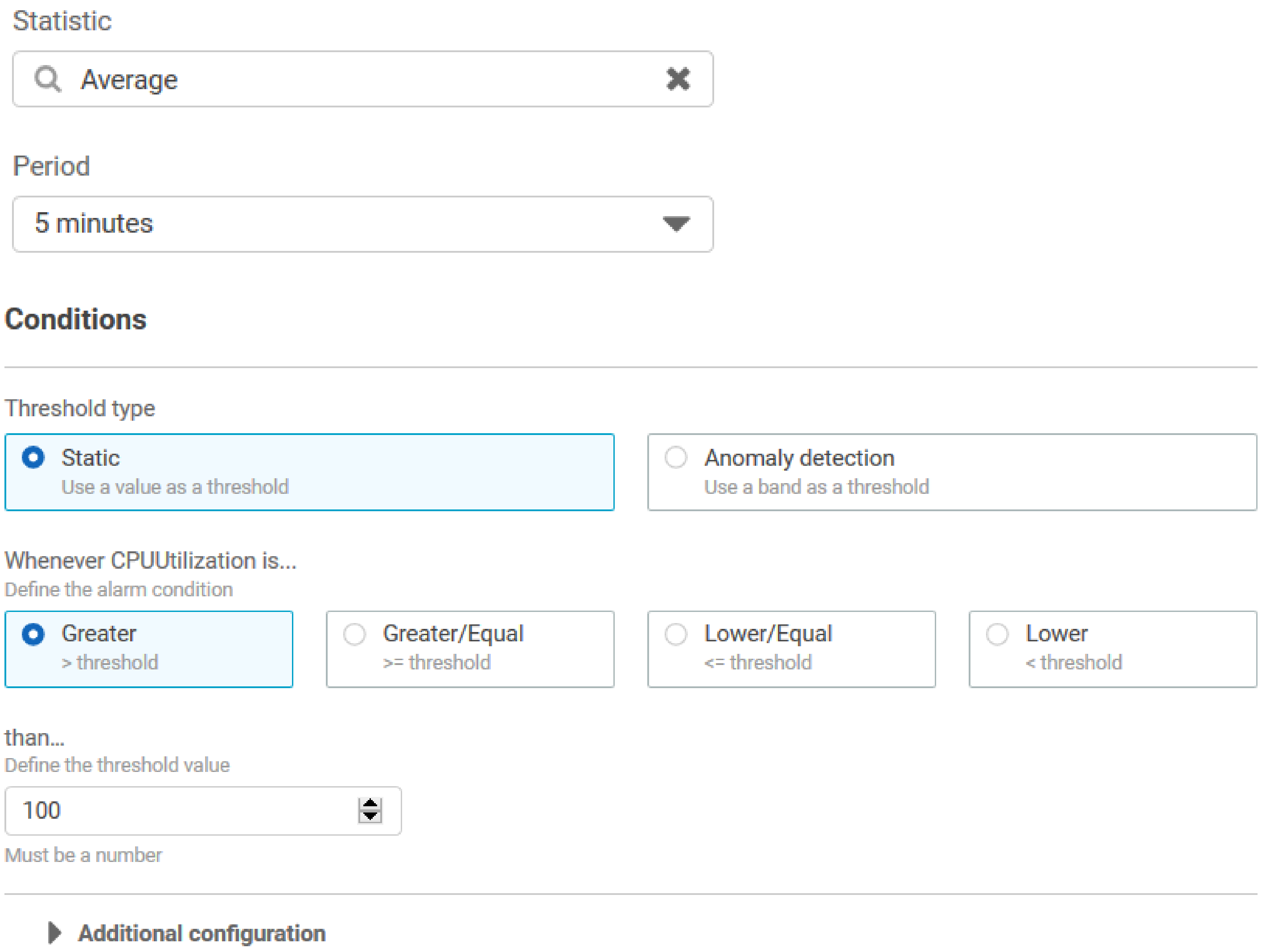
Amazon CloudWatch



- 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*).

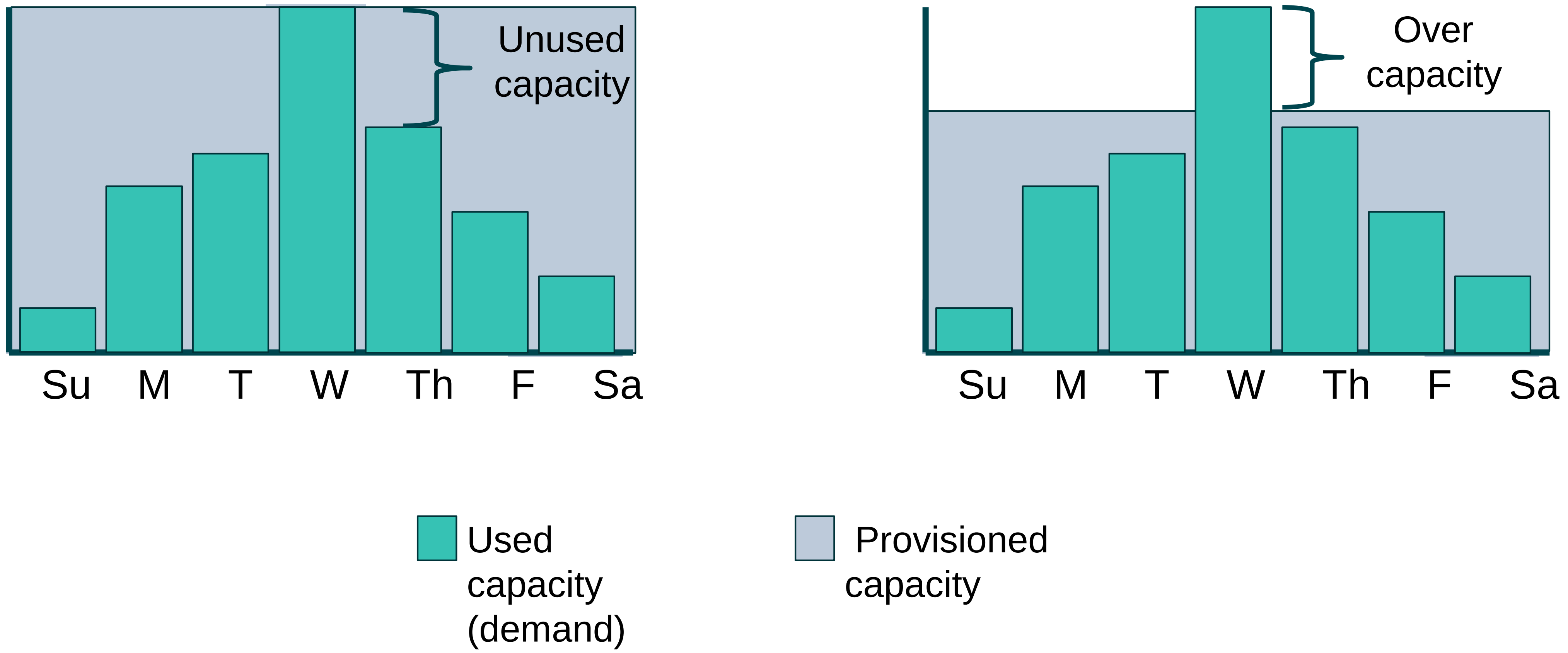
Section 2 key takeaways

- 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.

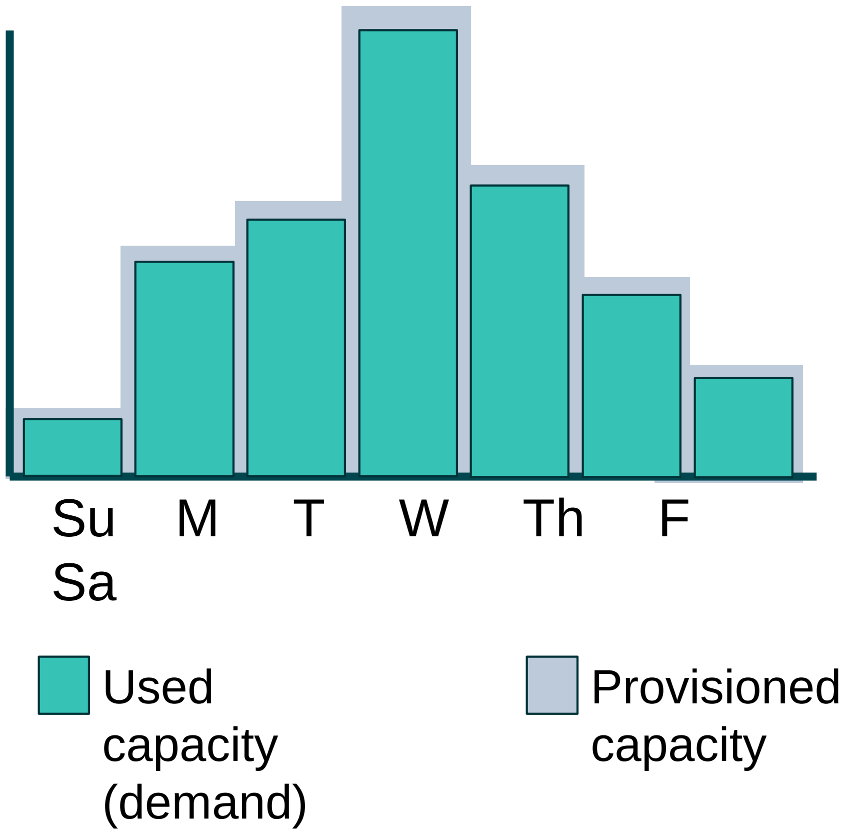
Section 3: Amazon EC2 Auto Scaling

Module 10: Automatic Scaling and Monitoring

Why is scaling important?

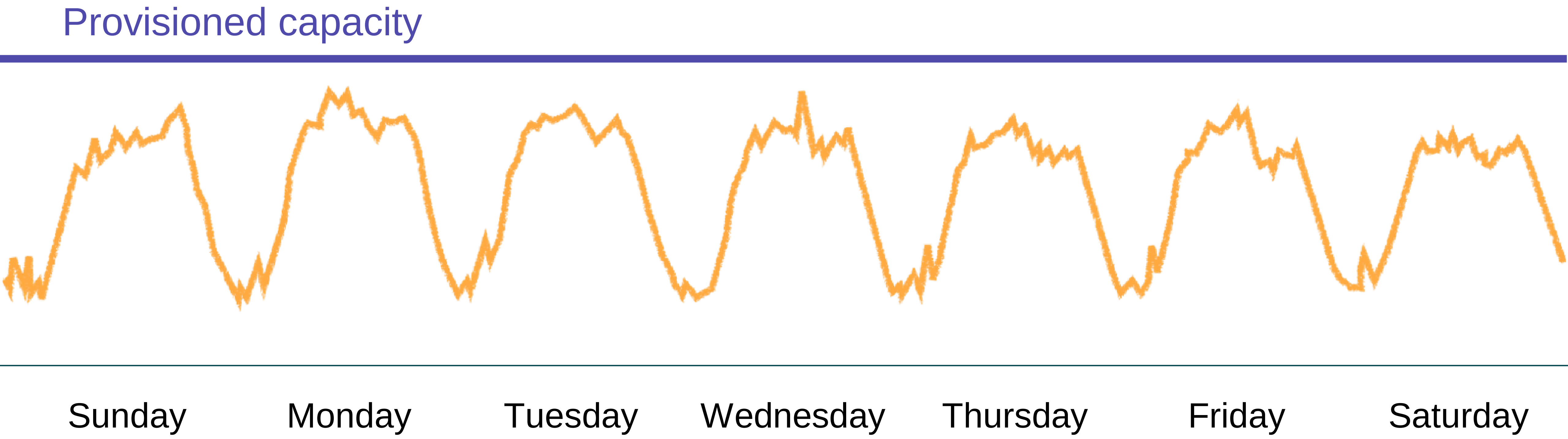


Amazon EC2 Auto Scaling

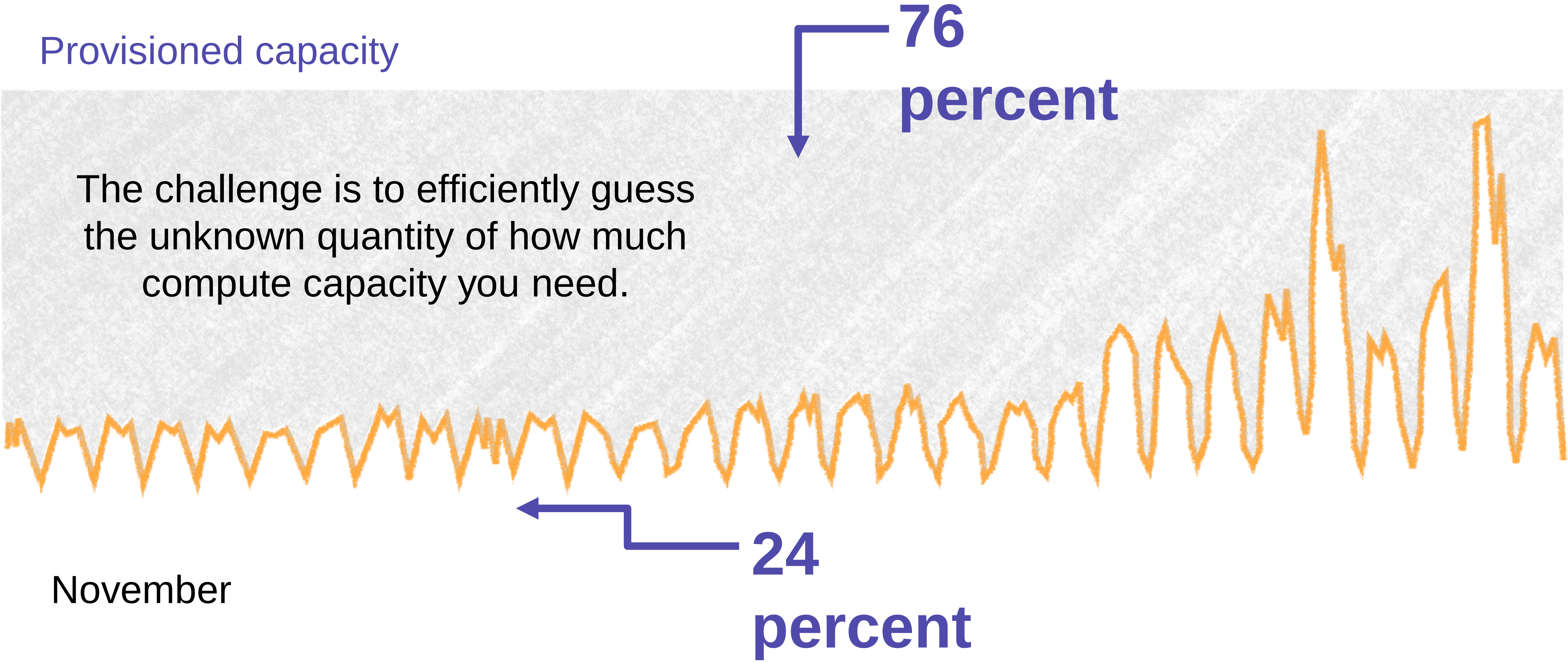


- Helps you maintain application availability
- Enables you to automatically add or remove EC2 instances according to conditions that you define
- Detects impaired EC2 instances and unhealthy applications, and replaces the instances without your intervention
- Provides several scaling options – Manual,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

Typical weekly traffic at Amazon.com



November traffic to Amazon.com



Provisioned capacity

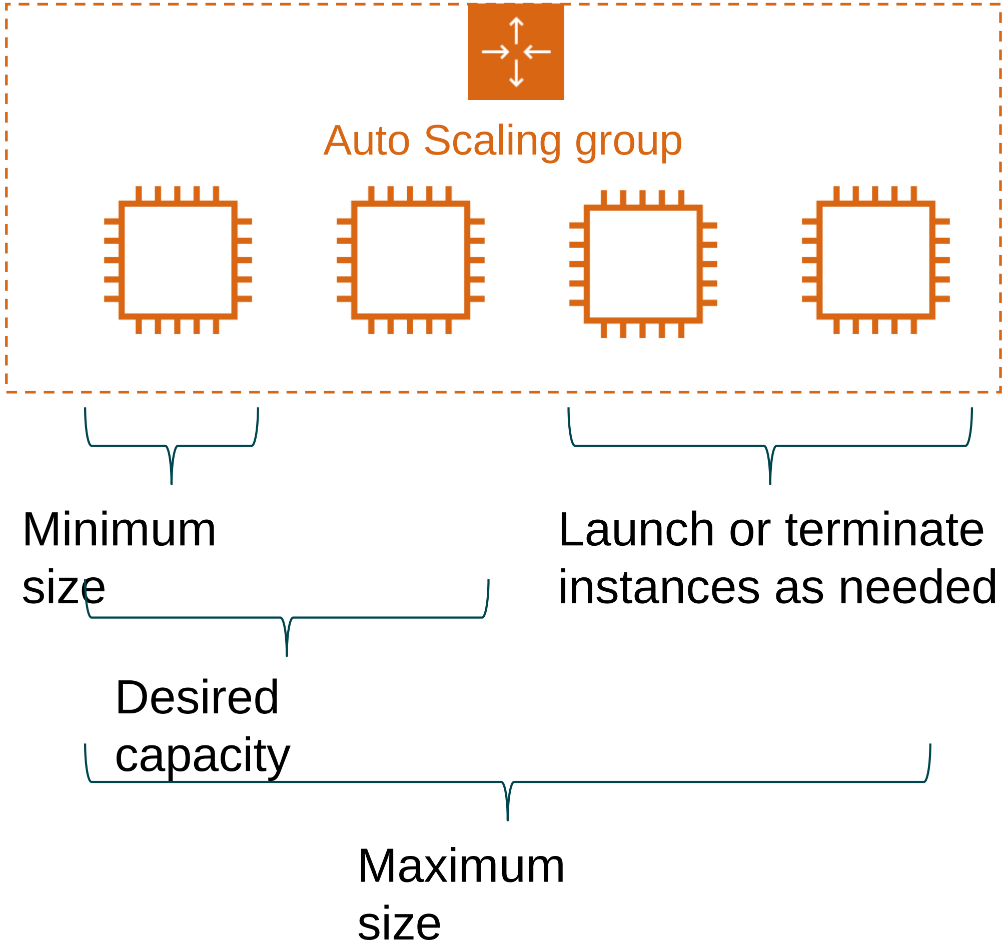
76 percent

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

24 percent

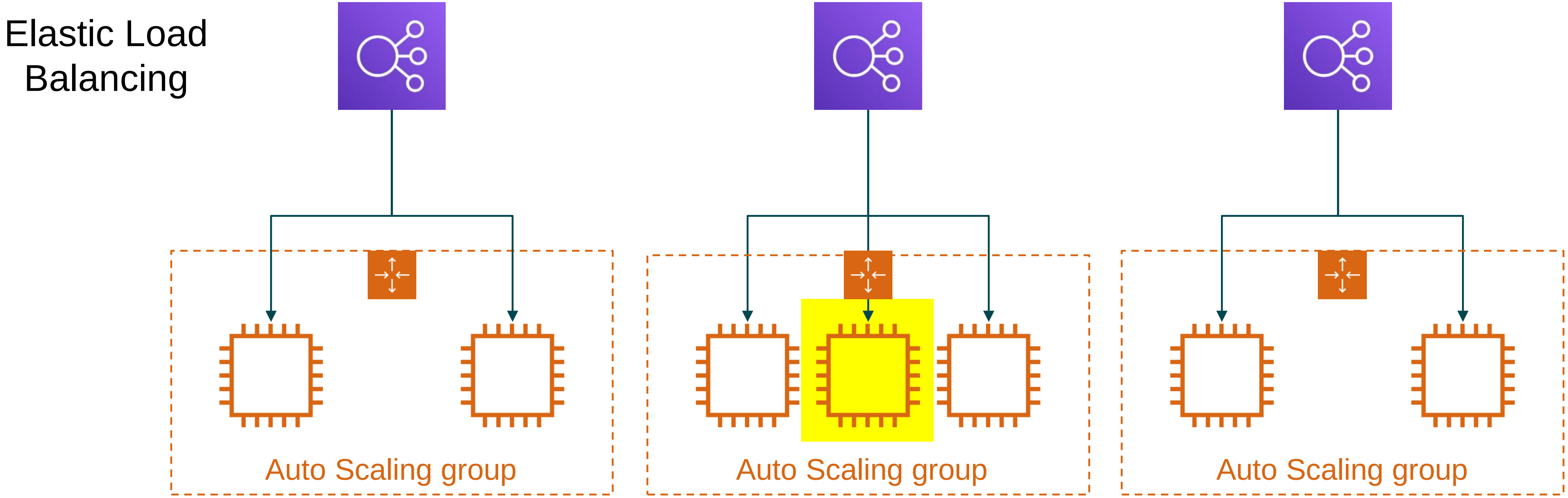
November

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

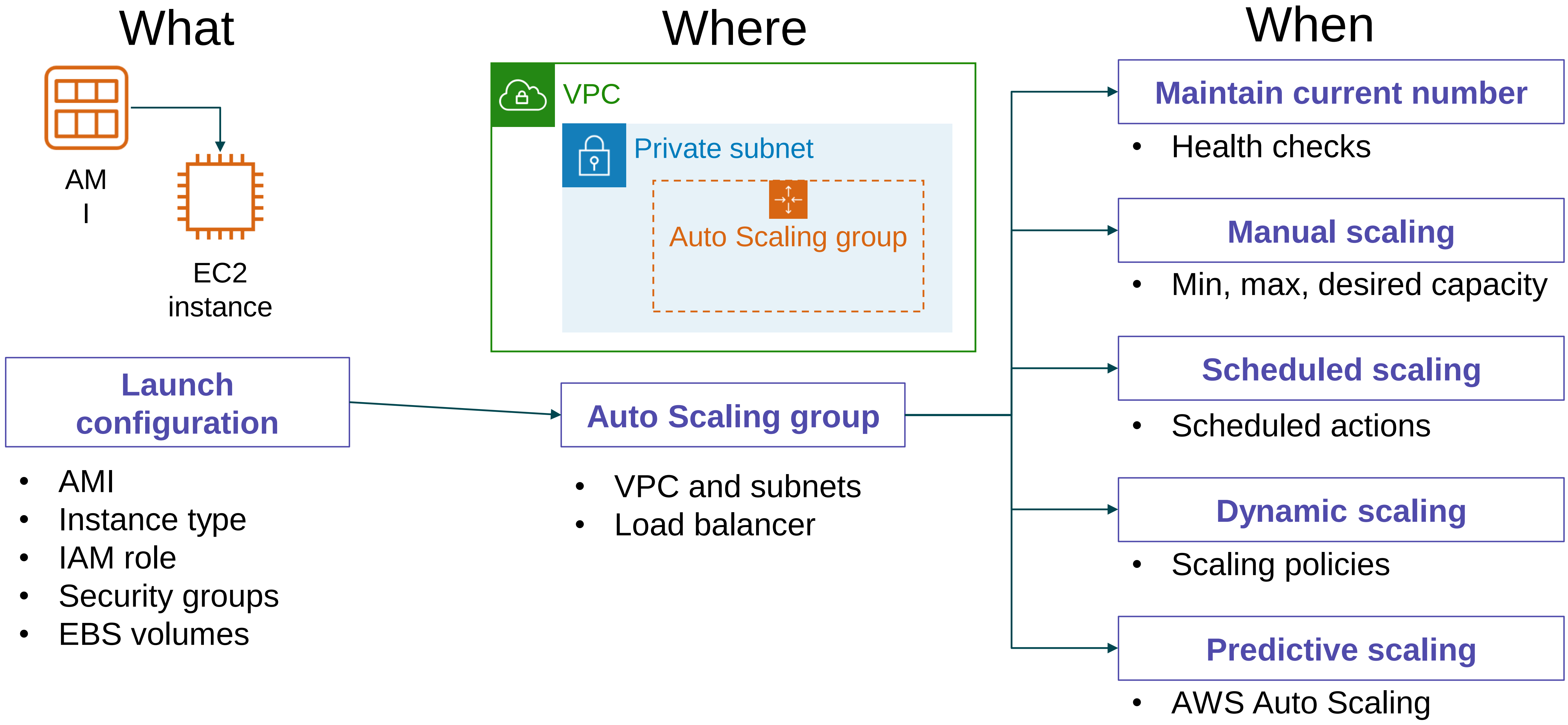


Base configuration

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

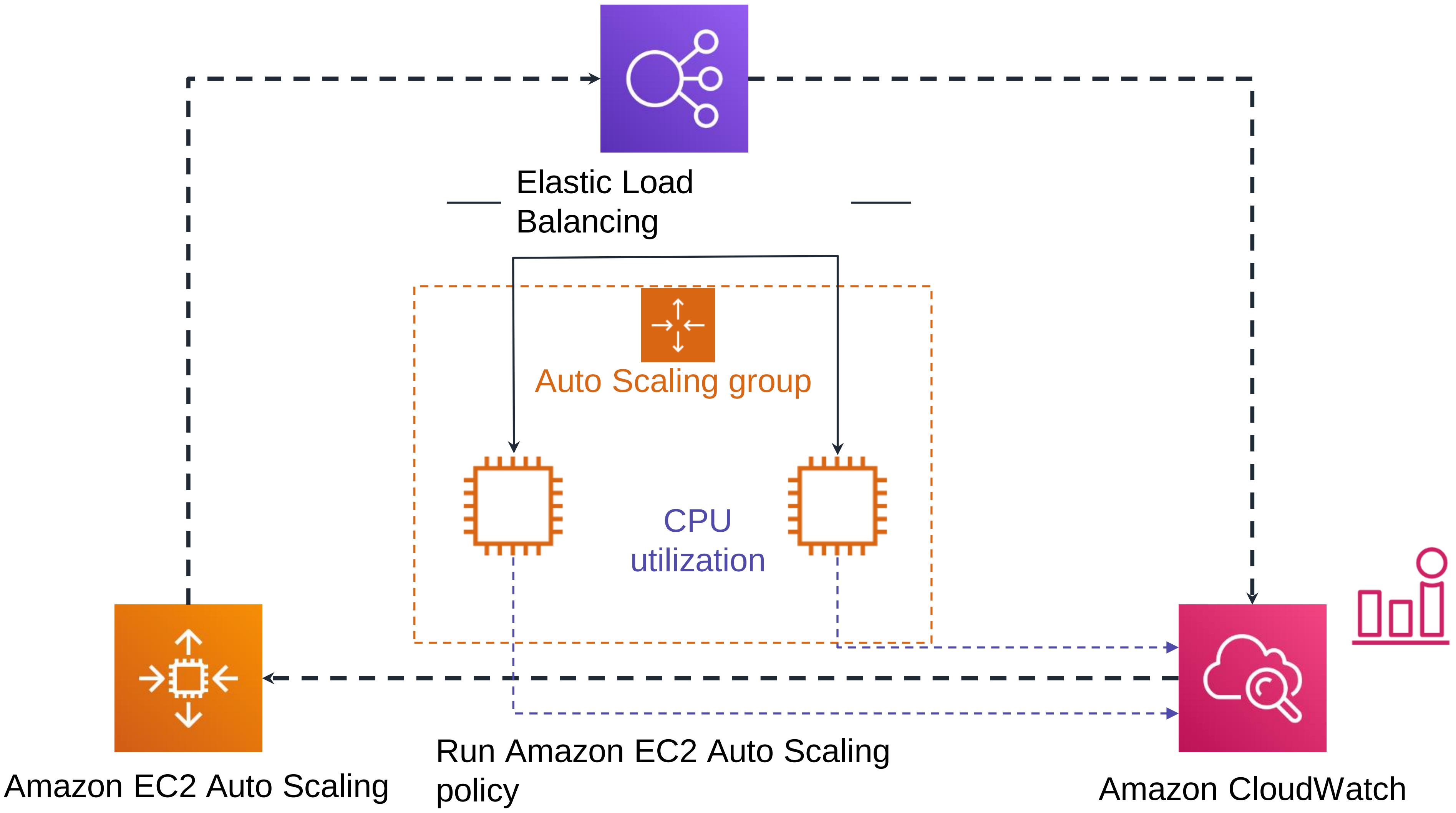
Dynamic scaling

- 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

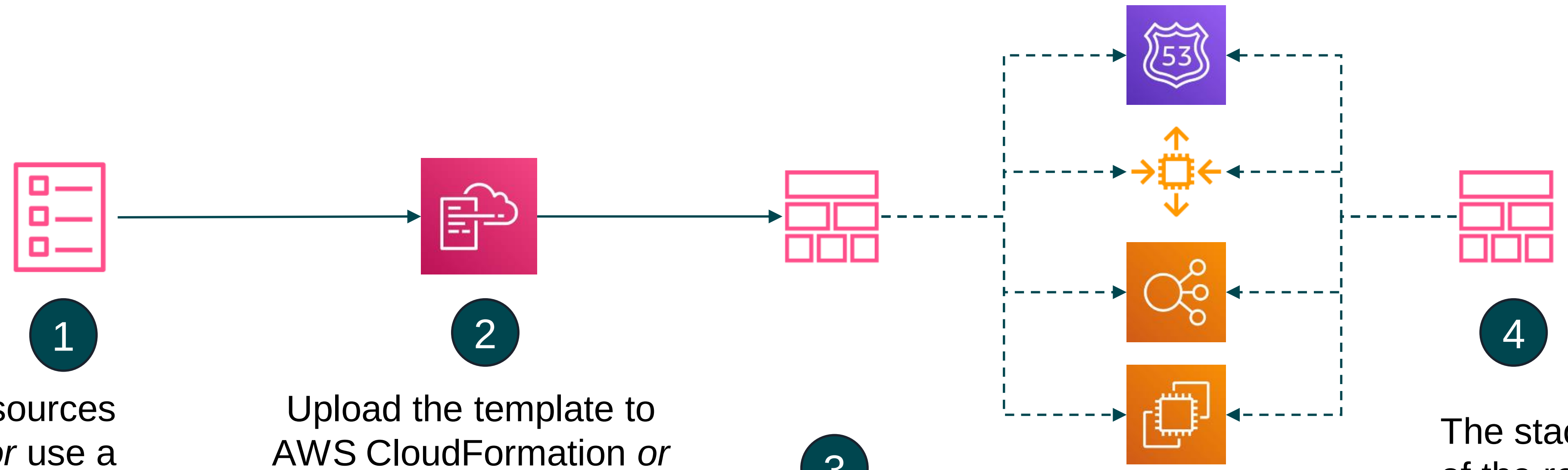
- Scaling enables you to respond quickly to changes in resource needs.
- Amazon EC2 Auto Scaling maintains application availability by automatically adding or removing EC2 instances.
- An Auto Scaling group is a collection of EC2 instances.
- A launch configuration is an instance configuration template.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

Automating your infrastructure

AWS CloudFormation

- AWS CloudFormation provides a simplified way to model, create, and manage a collection of AWS resources
  - Collection of resources is called an AWS CloudFormation stack
  - No extra charge (pay only for resources you create)
- Can create, update, and delete stacks
- Enables orderly and predictable provisioning and updating of resources
- Enables version control of AWS resource deployments

AWS CloudFormation overview



1

Define your resources in a **template** *or* use a pre-built template.

2

Upload the template to AWS CloudFormation *or* point to a template stored in an S3 bucket.

3

Run a **create stack** action. Resources are created across multiple services in your AWS account as a running environment.

4

The stack retains control of the resources that are created. You can later **update stack**, **detect drift** or **delete stack**.

AWS Quick Starts

AWS Quick Starts

AWS CloudFormation templates built by AWS solutions architects

- AWS Quick Starts provide AWS CloudFormation templates. The Quick Starts are built by AWS solutions architects and partners to help you deploy popular solutions on AWS, based on AWS best practices for security and high availability.
- Are gold-standard deployments
- Are based on AWS best practices for security and high availability
- Can be used create entire architectures with one click in less than an hour
- Can be used for experimentation and as the basis for your own architectures

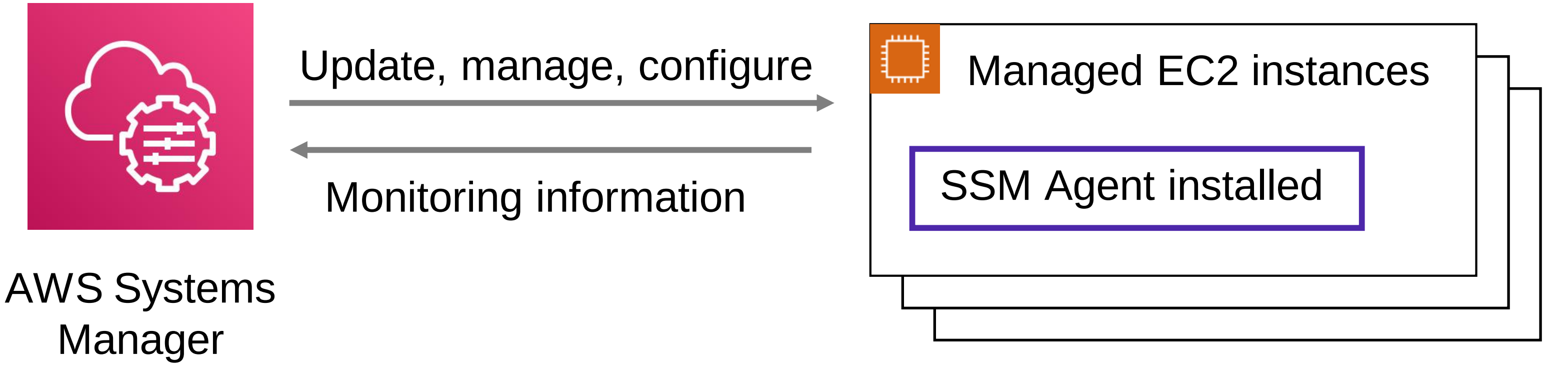
AWS Systems Manager

Gain operational insights and take action on AWS resources.

AWS Systems Manager

- Automates operational tasks
  - Example: Apply OS patches and software upgrades across a fleet of EC2 instances
- Simplifies resource and application management
  - Manage software inventory
  - View detailed system configurations across the fleet
- Manages servers on-premises and in the cloud

Systems Manager capabilities



AWS Systems Manager

Update, manage, configure

Monitoring information

Managed EC2 instances

SSM Agent installed

- Run Command
- Patch Manager
- Session Manager
- Maintenance Windows
- State Manager
- Inventory
- Parameter Store
- Automation
- Documents

You can install an *AWS Systems Manager Agent (SSM Agent)* on an EC2 instance, or even on an on-premises server, or a virtual machine (VM). Once the SSM Agent is installed, it will be possible for Systems Manager to update, manage, and configure the server on which it is installed. The agent processes requests from Systems Manager and then runs them in accordance with the specification provided in the request. The agent then sends status and relevant information back to Systems Manager.

Amazon API Gateway

Amazon API Gateway

- Enables you to create, publish, maintain, monitor, and secure APIs that act as entry points to backend resources for your applications
- Handles up to hundreds of thousands of concurrent API calls
- Can handle workloads that run on –
  - Amazon EC2
  - Lambda
  - Any web application
  - Real-time communication applications
- Can host and use multiple versions and stages of your APIs

Amazon API Gateway security

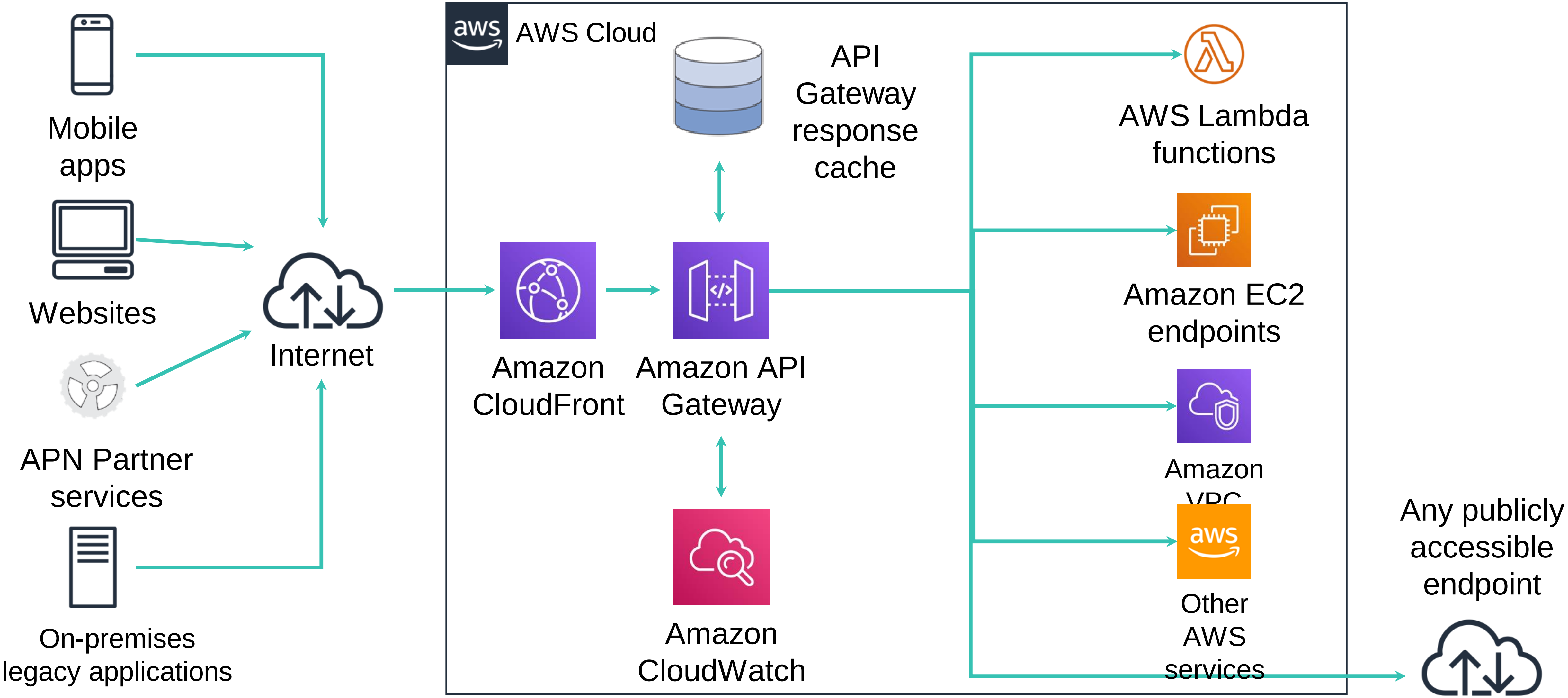
Require authorization

Apply resource policies

Throttling settings

Protection from Distributed Denial of Service (DDoS) and injection attacks

Amazon API Gateway: Common architecture example



Amazon SQS

Amazon Simple Queue Service (Amazon SQS)

- Fully managed message queueing service
- Uses a pull mechanism
- Messages are encrypted and stored until they are processed and deleted
- Acts as a buffer between producers and consumers

Achieve loose coupling with Amazon SQS

With Amazon SQS, you can:

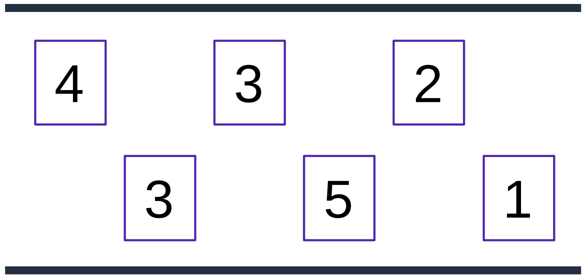
Use asynchronous processing to get your responses from each step quickly

Handle performance and service requirements by increasing the number of job instances

Easily recover from failed steps because messages will remain in the queue

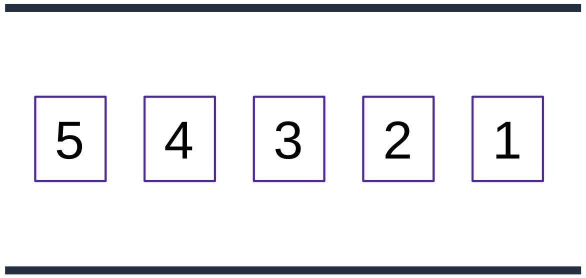
Queue types

Standard queues



- At-least-once delivery
- Best-effort ordering
- Nearly unlimited throughput

First in, first out (FIFO) queues



- First-in-first-out delivery
- Exactly once processing
- High throughput

Amazon SNS

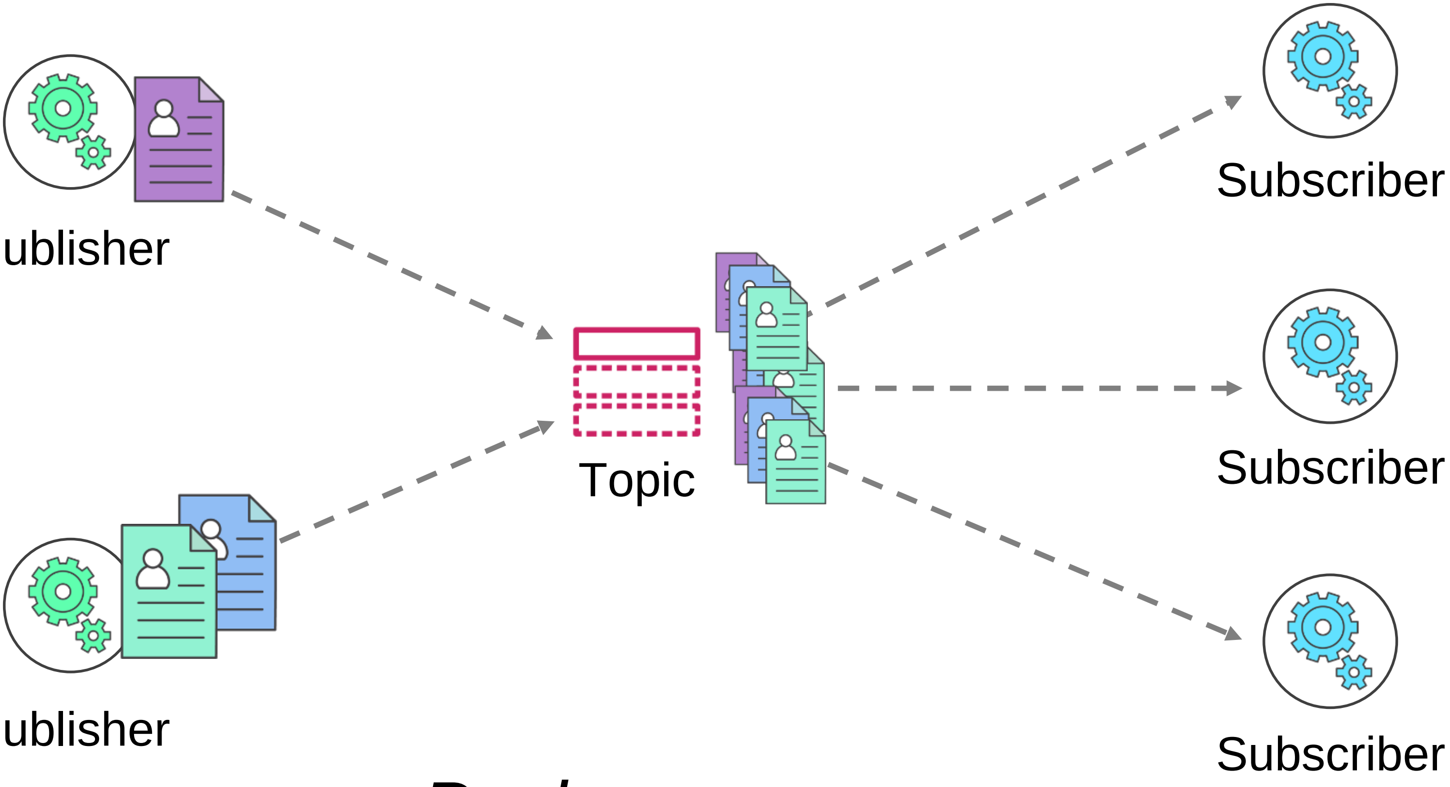
Amazon Simple Notification Service (Amazon SNS)

- Is a highly available, durable, secure, and fully managed pub/sub messaging service
- Uses a push mechanism
- Supports encrypted topics using customer master keys (CMKs)

Pub/sub messaging

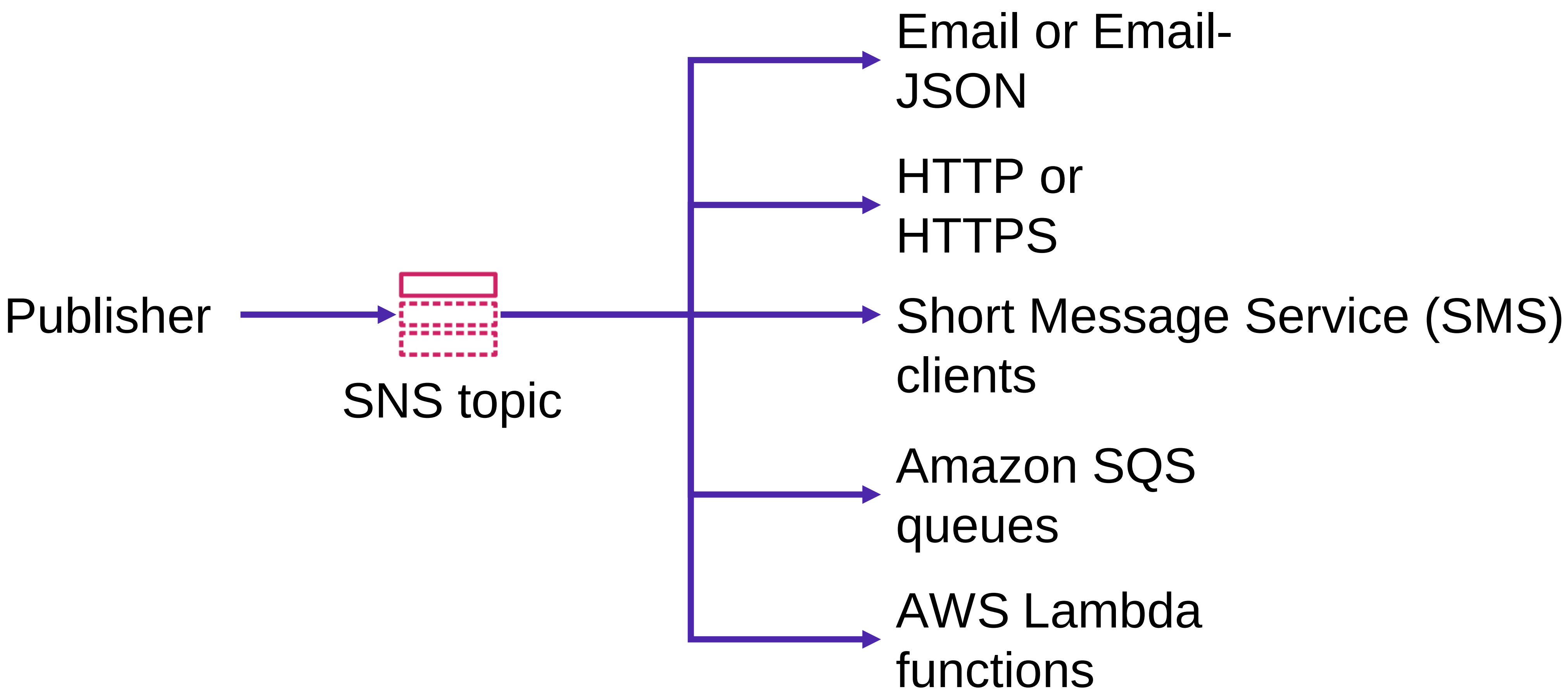
Publisher – Component that pushes a message to a topic

Subscriber – Component that subscribes to a topic



Push mechanism

Supported transport protocols



Amazon SNS versus Amazon SQS

Feature	**Amazon SNS** (Publisher/Subscriber)	**Amazon SQS** (Producer/Consumer)
Producer/consumer	Publish/subscribe	Send/receive
Delivery mechanism	Push (passive)	Poll (active)
Distribution model	Many to many	One to one
Message persistence	No	Yes

Amazon MQ

Amazon MQ

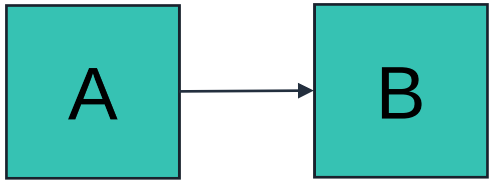
- Is a managed **message broker** service for Apache ActiveMQ
- Manages the provisioning, setup, and maintenance of ActiveMQ
- Simplifies message migration to the cloud
- Is compatible with open-standard APIs and protocols
  - JMS, NMS, AMQP, STOMP, MQTT, and WebSockets

AWS Step Functions

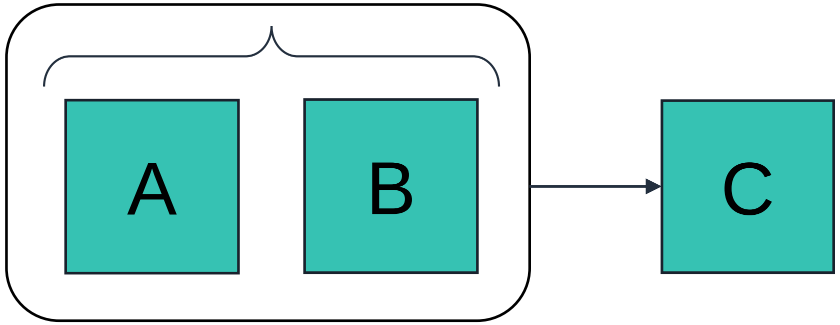
AWS Step Functions

- Coordinates microservices by using visual workflows
- Enables you to step through the functions of your application
- Automatically triggers and tracks each step
- Provides simple error catching and logging if a step fails

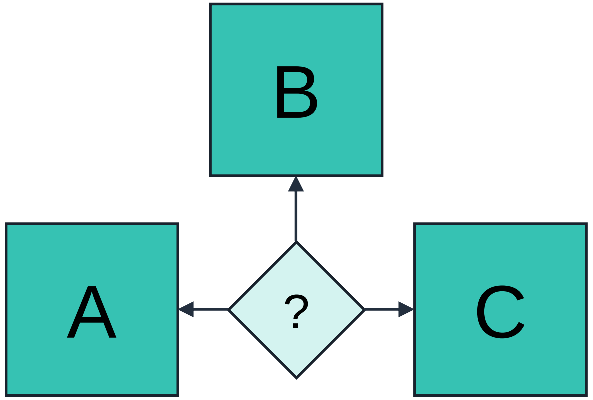
Workflow coordination



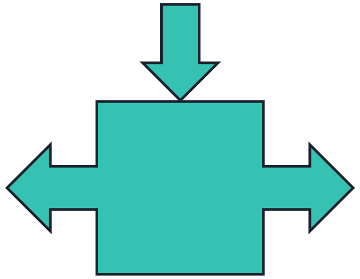
Run tasks in sequence



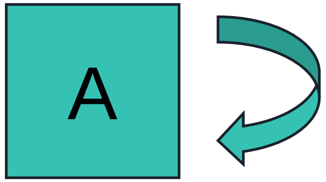
Run tasks in parallel



Select task based on data



Manage try-catch-finally behavior



Retry failed tasks

AWS CodeBuild

What is AWS CodeBuild?

Amazon codeBuild is a fully managed continuous integration service that helps developers to build and run test code rapidly. This service also provides test artifacts with great efficiency. This gives an advantage to developers/DevOps engineers not to wait in long build queues, scaling, configuring, and maintaining build service. AWS CodeBuild runs continuously and avoids the wait time for concurrent jobs. Additionally, users pay only for the build time they use. In other words, the ideal time won’t be counted in billing time for the user.

Features:

1. Easy to Set up – Amazon CodeBuild is easy to set up for the developers. Either they can build their code build environment or use one of the preconfigured environments.
2. It works with existing tools such as Jenkins plugin, GIT, etc.
3. It can scale automatically with several concurrent builds.
4. Automated Build: Developers need to configure builds once and whenever there is a code change, CodeBuild service will automatically run and generate the test results.
5. Pay as you go which means developers are only charged for the time it takes to complete the build, idle time won’t be considered in billing.

Pricing: Pricing will be computed based on the build minutes. The first 100 minutes of the build are free and then the rest will be charged based on each instance type usage.

AWS CodeCommit

What Is AWS CodeCommit?

AWS CodeCommit is a version control service hosted by Amazon Web Services that users can use privately to store and manage assets (such as documents, source code, and binary files) in the cloud.

In other words, AWS provides the service which allows you to just store the code or any assets without worrying about the version control like other version control tools such as Bitbucket, GitHub, etc. AWS manages everything on its own and takes full responsibility for scaling its infrastructure.

Features:

- Ensures High secure code (encrypted) with any type of code.
- Collaborative work (users can access the same piece of code with different IAM users and different security groups).
- Easy scalability.
- Easy to integrate with third-party groups.

Pricing:

AWS CodeCommit provides free tier term for the first 5 active users for the below configurations:

First 5 active user receives:

- Unlimited repositories
- 50 GB storage
- 10K Git requests/month

Additional active user beyond the first 5 users:

- Unlimited repositories
- 10 GB storage
- 2K Git requests/month

There will be additional charges for additional storage or an increase in GIT requests if increased.

AWS CodeDeploy

What is AWS CodeDeploy?

AWS CodeDeploy is a deployment service that automates application deployments to Amazon EC2 instances, on-premises instances, serverless Lambda functions, or Amazon ECS services.

Below type of deployments can be done using AWS CodeDeploy service:

- Code, Serverless Lambda Functions.
- Web & Configuration Files
- Executables and Packages.
- Scripts and Multimedia.

Following are components that concerns AWS CodeDeploy Service:

- Compute Platform
- Deployment Types & Groups.
- IAM & Service Roles
- Applications.

Features:

- Help to release new features rapidly.
- It supports avoiding downtime during application deployment by maximizing application availability and handles all application complexity.
- It allows easy launch and tracking of application status.

Pricing:

- Free code deployment to Amazon EC2 or AWS Lambda.
- $0.02 charges per on-premises instance deployment.

AWS CodePipeline - Code Artifact

AWS CodePipeline
AWS CodePipeline is a continuous integration and continuous delivery service for fast and reliable application and infrastructure updates. CodePipeline builds, tests, and deploys your code every time there is a code change, based on the release process models you define.

AWS CodeArtifact provides on-demand downloads of AWS security and compliance documents, such as AWS ISO certifications, Payment Card Industry (PCI), and Service Organization Control (SOC) reports. You can submit the security and compliance documents (also known as audit artifacts) to your auditors or regulators to demonstrate the security and compliance of the AWS infrastructure and services that you use.

You can also use these documents as guidelines to evaluate your own cloud architecture and assess the effectiveness of your company’s internal controls. AWS Artifact provides documents about AWS only. AWS customers are responsible for developing or obtaining documents that demonstrate the security and compliance of their companies.

AWS EventBridge

What is Amazon EventBridge?

A serverless event bus service for Software-as-a-Service (SAAS) and AWS services. In simple words, Amazon EventBridge provides an easy solution to integrate SAAS, custom-build applications with more than 17+ AWS services with the delivery of real-time data from different event sources. Users can easily set up the routing rules to determine the target web-service, and multiple target locations (such as AWS Lambda or AWS SNS) can be selected at once.

It is a fully managed service that takes care of event ingestion, delivery, security, authorization, error handling, and required infrastructure management tasks to set up and run a highly scalable serverless event bus.

EventBridge was formerly called

Amazon CloudWatch Events, and it uses the same CloudWatch Event API.

Features:

- Fully managed, pay-as-you-go.
- Native integration with SaaS providers.
- 90+ AWS services as sources.
- 17 AWS services as targets.
- $1 per million events put into the bus.
- No additional cost for delivery.
- Multiple target locations for delivery.
- Easy to scale and manage.

AWS Sagemaker

What is Sagemaker?

Amazon SageMaker is a cloud service that allows developers to prepare, build, train, deploy and manage machine learning models.

Amazon SageMaker Data Wrangler

A faster, visual way to aggregate and prepare data for ML

❖ *It provides a secure and scalable environment to deploy a model using SageMaker Studio or the SageMaker console.*

❖ *It has pre-installed machine learning algorithms to optimize and deliver 10x performance SageMaker console.*

❖ *It scales up to petabytes level to train models and manages all the underlying infrastructure.*

❖ *Amazon SageMaker notebook instances are created using Jupyter notebooks to write code to train and validate the models.*

❖ *Amazon SageMaker gets billed in seconds based on the amount of time required to build, train, and deploy machine learning models.*

AWS Sumerian

What is Sumerian?

Amazon Sumerian provides tools and resources that allows anyone to create and run augmented reality (AR), virtual reality (VR), and 3D applications with ease. With Sumerian, you can build multi-platform experiences that run on hardware like the Oculus, HTC Vive, and iOS devices using WebVR compatible browsers and with support for ARCore on Android devices coming soon.

This exciting new service, currently in preview, delivers features to allow you to design highly immersive and interactive 3D experiences from your browser. Some of these features are:

- **Editor:** A web-based editor for constructing 3D scenes, importing assets, scripting interactions and special effects, with cross-platform publishing.
- **Object Library:** a library of pre-built objects and templates.
- **Asset Import:** Upload 3D assets to use in your scene. Sumerian supports importing FBX, OBJ, and coming soon Unity projects.
- **Scripting Library:** provides a JavaScript scripting library via its 3D engine for advanced scripting capabilities.
- **Hosts:** animated, lifelike 3D characters that can be customized for gender, voice, and language.
- **AWS Services Integration:** baked in integration with Amazon Polly and Amazon Lex to add speech and natural language to into Sumerian hosts. Additionally, the scripting library can be used with AWS Lambda allowing use of the full range of AWS services.

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 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

Corrections, feedback, or other questions?
Contact us at https://support.aws.amazon.com/#/contacts/aws-academy.
All trademarks are the property of their owners.