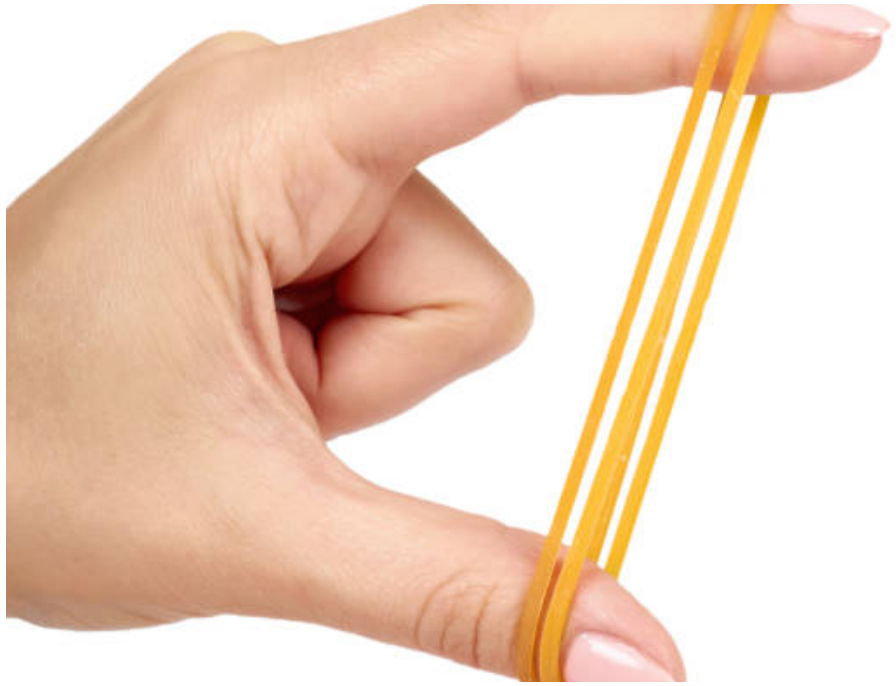
Agenda

- **What is Auto Scaling?**
- **Understanding Launch Configuration**
- **Understanding Auto Scaling Group**
- **Creating Notification (SNS Service)**
- **Creating Alarm (Cloud Watch Service)**
- **Add Auto Scaling Policies**

Elasticity & Scalability



Elasticity

Scalability

What is Auto-Scaling

- AWS Auto Scaling monitors your applications and adjusts capacity automatically to ensure consistent, predictable performance at the lowest possible cost.
- Automatically launch or terminate Amazon Elastic Compute Cloud (EC2) instances based on:
  - Health status checks
  - User-defined policies that are driven by Amazon CloudWatch
  - Schedules
  - Other criteria (for example, programmatically)
  - - Manually using set desired capacity
  - Scale out to meet demand, scale in to reduce costs

Amazon EC2 Auto Scaling in action

What to launch

Launch configuration, launch template

Instance configuration to be launched

Specify:
- Amazon Machine Image (AMI)
- Instance type
- Security group
- Instance key pair
- Storage
- AWS Identity and Access Management (IAM) roles
- User data

Only one active launch configuration at a time

How to manage

Amazon EC2 Auto Scaling Group

Logical group of EC2 instances

Automatically scale between:
- Minimum
- Desired (optional)
- Maximum

Integration with Elastic Load Balancing (optional)

Health checks to maintain group size

Distribute and balance instances across Availability Zones

When to scale

Amazon EC2 Auto Scaling Policy

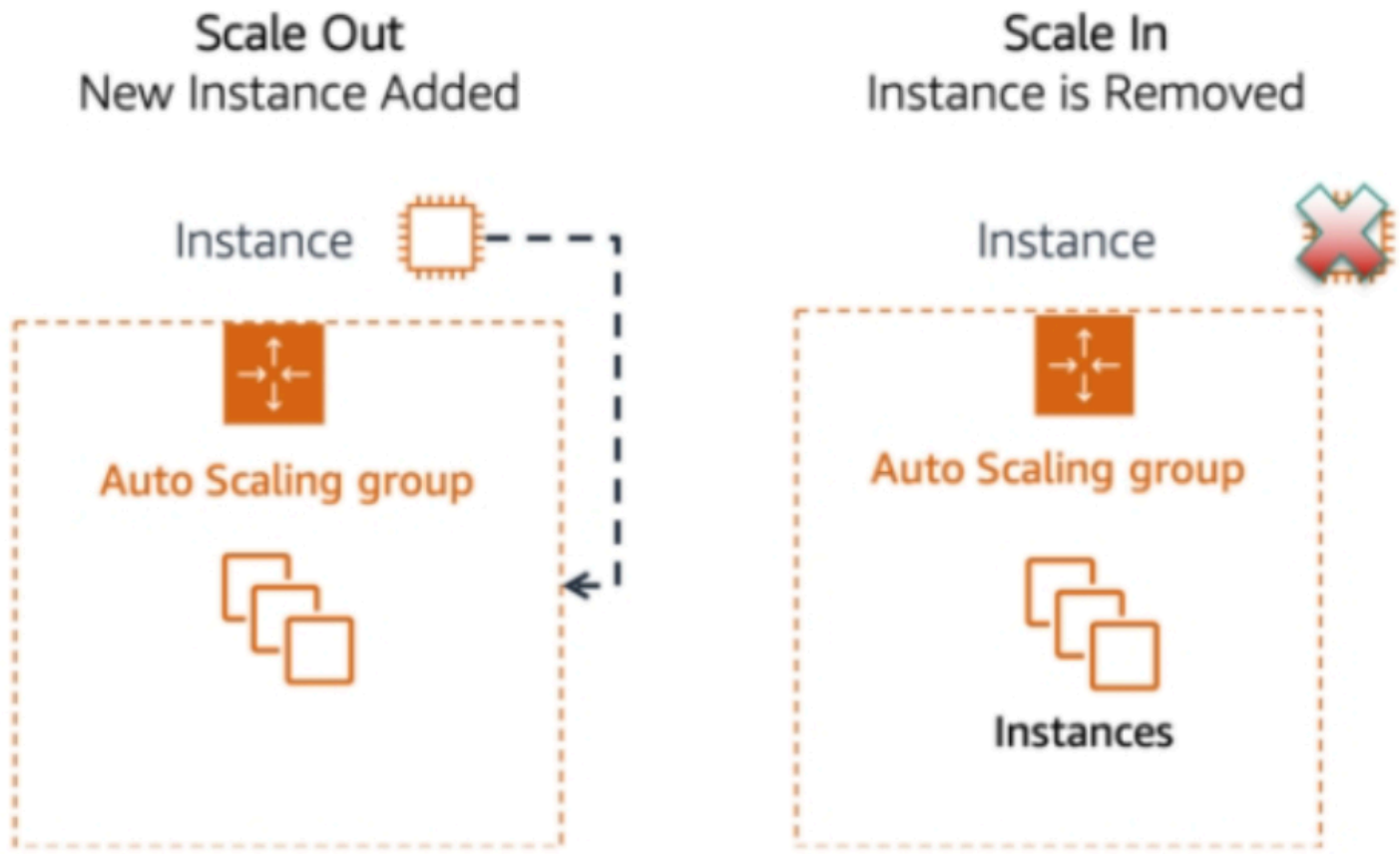
Parameters for performing an Amazon EC2 Auto Scaling action

How to trigger policies?
- Amazon CloudWatch
- Instance failure (health check)
- Scheduled
- Manually

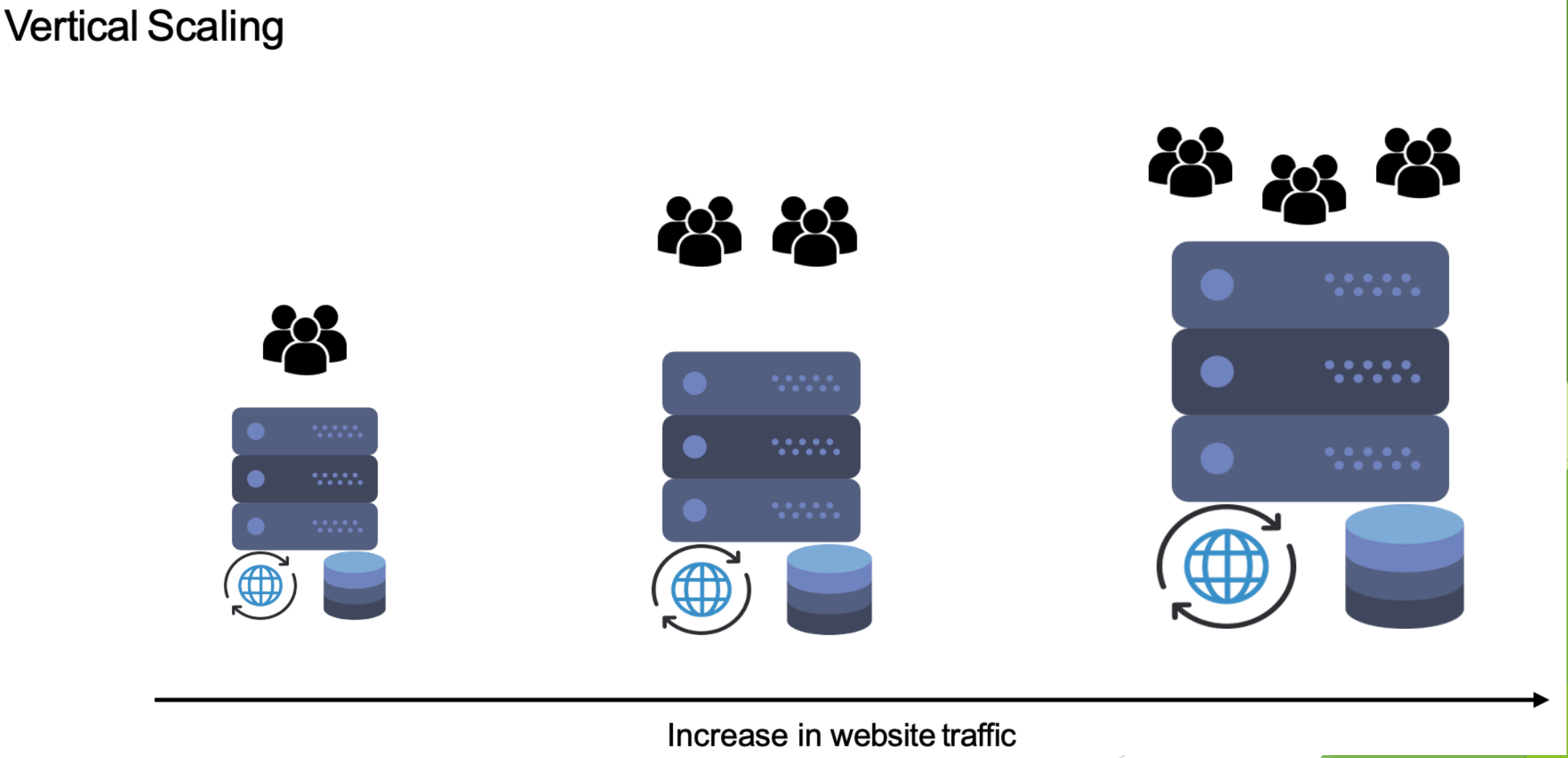
Scale out or in, and by how much:
- ChangeInCapacity (+/- #)
- ExactCapacity(#)
- ChangeInPercent (+/- %)

- Cooldown period (simple scaling)
- Warmup period (step scaling)

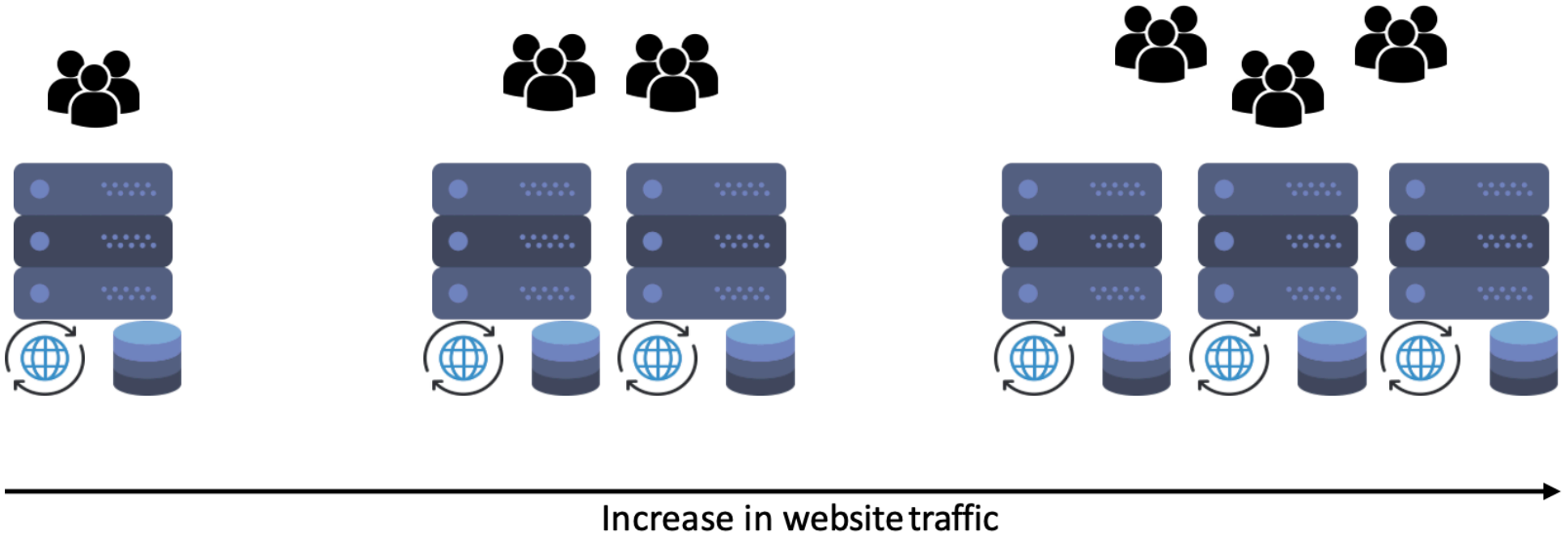
Amazon EC2 Auto Scaling termination policy



Vertical & Horizontal Scaling



Horizontal Scaling



What is Amazon CloudWatch?

Amazon CloudWatch

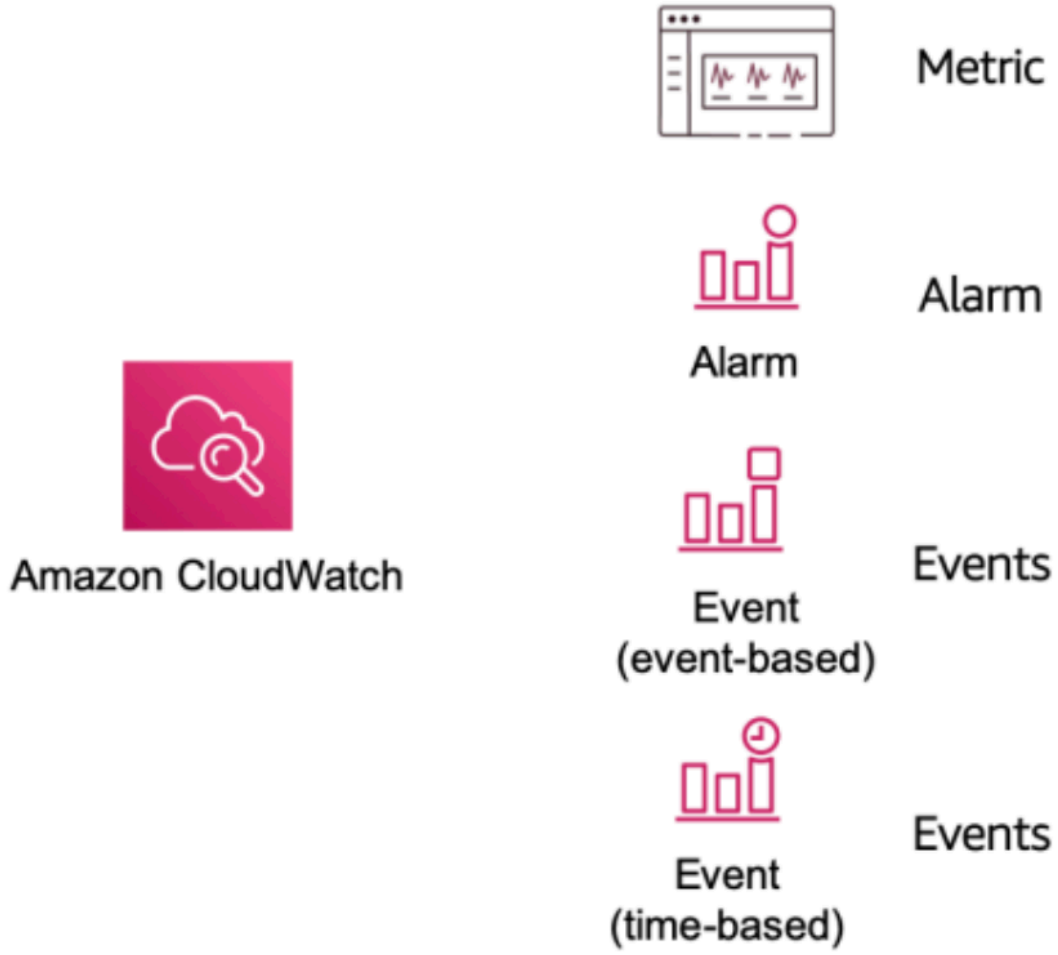
Monitors the state and utilization of most resources that you can manage under AWS

- Key concepts:
  - Standard metrics
  - Custom metrics
  - Alarms
  - Notifications

CloudWatch agent collects system-level metrics:

- EC2 instances
- On-premises servers

Amazon CloudWatch Terms



CloudWatch

- The primary function of Amazon CloudWatch to monitor the performance and health of your AWS resources and applications.
- You can also use CloudWatch to collect and monitor log files from EC2 instances, AWS CloudTrail, Amazon Route 53, and other sources.
- **Basic Monitoring for Amazon EC2 instances :** Seven pre-selected metrics at a 5-minute frequency and three status check metrics at a 1-minute frequency, for no additional charge.
- **Detailed Monitoring for Amazon EC2 instances :** All metrics that are available to basic monitoring at 1-Minute frequency, for and additional charge.

Amazon CloudWatch alarms

- Test a selected metric against a specific threshold (greater than or equal to, less than or equal to)
- The **ALARM** state is not necessarily an emergency condition

Alarms have three possible states:

Threshold not exceeded

Threshold exceeded

Alarm has just started, metric is not available, or insufficient data

SNS – (Simple Notification Service)

- Amazon Simple Notification Service (Amazon SNS) is a managed service that provides message delivery from publishers to subscribers (also known as *producers* and *consumers*). Publishers communicate asynchronously with subscribers by sending messages to a *topic*, which is a logical access point and communication channel. Clients can subscribe to the SNS topic and receive published messages using a supported endpoint type, such as Amazon Kinesis Data Firehose, Amazon SQS, AWS Lambda, HTTP, email, mobile push notifications, and mobile text messages (SMS).