AUTO SCALING

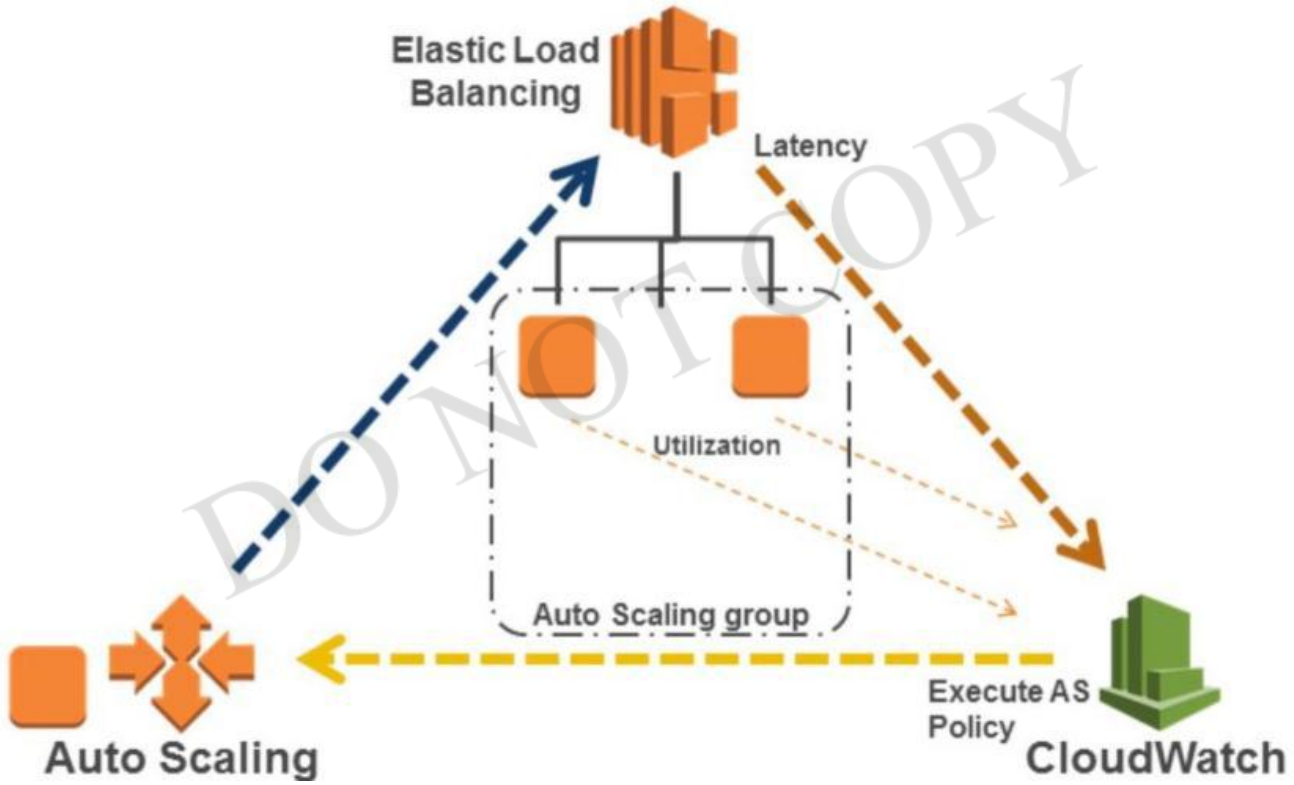
AUTO SCALING

- ✓ Scale your Amazon EC2 capacity automatically
- ✓ Well-suited for applications that experience variability in usage
- ✓ Available at no additional charge

Understand Auto Scaling concepts including:

- ✓ Launch Configurations
- ✓ Auto Scaling Groups
- ✓ Scaling Plans
- ✓ Auto Scaling Lifecycle
- ✓ Auto Scaling Limits

Trio of Services



Auto Scaling works as a triad of services working in sync. Elastic Load Balancing and EC2 instances feed metrics to Amazon CloudWatch. Auto Scaling defines a group with launch configurations and Auto Scaling policies. Amazon CloudWatch alarms execute Auto Scaling policies to affect the size of your fleet. All of these services work well individually, but together they become more powerful and increase the control and flexibility our customers demand.

Auto Scaling Benefits

Adding Auto Scaling to your application architecture is one way to maximize the benefits of the AWS cloud. When you use Auto Scaling, your applications gain the following benefits:

Better fault tolerance: Auto Scaling can detect when an instance is unhealthy, terminate it, and launch an instance to replace it. You can also configure Auto Scaling to use multiple Availability Zones. If one Availability Zone becomes unavailable, Auto Scaling can launch instances in another one to compensate.

Better availability: Auto Scaling can help you ensure that your application always has the right amount of capacity to handle the current traffic demands.

Better cost management: Auto Scaling can dynamically increase and decrease capacity as needed. Because you pay for the EC2 instances you use, you save money by launching instances when they are actually needed and terminating them when they aren't needed.

Launch Configurations

A launch configuration is a template that an Auto Scaling group uses to launch EC2 instances.

- ✓ When you create a launch configuration, you can specify:
- ✓ AMI ID
- ✓ Instance type
- ✓ Key pair
- ✓ Security groups
- ✓ Block device mapping
- ✓ User data

When you create an Auto Scaling group, you must specify a launch configuration. You can specify your launch configuration with multiple Auto Scaling groups. However, you can only specify one launch configuration for an Auto Scaling group at a time, and you can't modify a launch configuration after you've created it. If you want to change the launch configuration for your Auto Scaling group, you must create a new launch configuration and then update your Auto Scaling group with the new launch configuration. When you change the launch configuration for your Auto Scaling group, any new instances are launched using the new configuration parameters, but existing instances are not affected.

Auto Scaling Groups

- ✓ Contain a collection of EC2 instances that share similar characteristics.
- ✓ Instances in an Auto Scaling group are treated as a logical grouping for the purpose of instance scaling and management.

You can create collections of EC2 instances, called Auto Scaling groups. You can specify the minimum number of instances in each Auto Scaling group, and Auto Scaling ensures that your group never goes below this size. You can specify the maximum number of instances in each Auto Scaling group, and Auto Scaling ensures that your group never goes above this size. If you specify the desired capacity, either when you create the group or at any time thereafter, Auto Scaling ensures that your group has this many instances. If you specify scaling policies, then Auto Scaling can launch or terminate instances as demand on your application increases or decreases.

Dynamic Scaling

You can create a scaling policy that uses CloudWatch alarms to determine:

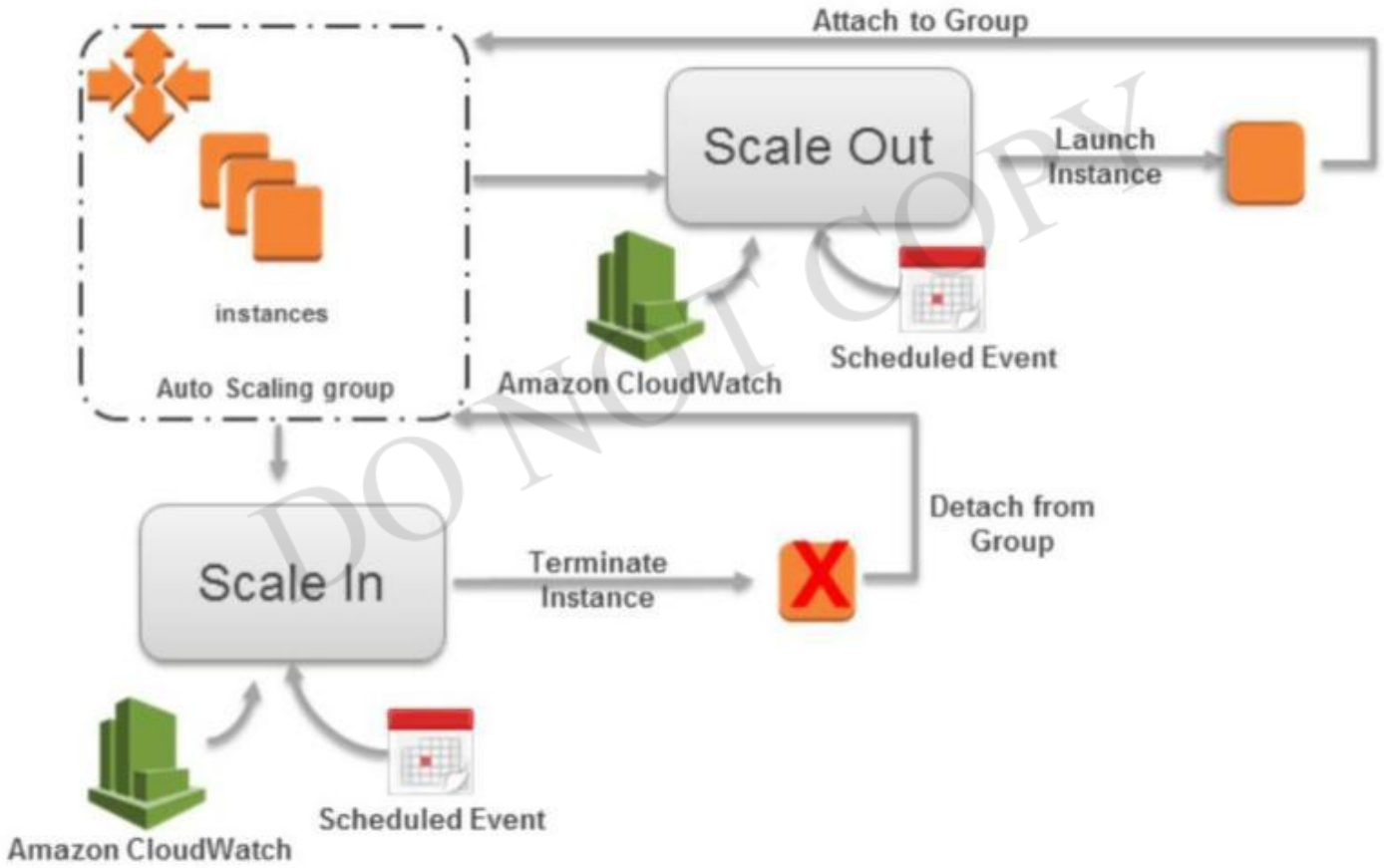
- ✓ When your Auto Scaling group should scale out.
- ✓ When your Auto Scaling group should scale in.

You can use alarms to monitor:

- ✓ Any of the metrics that AWS services send to Amazon CloudWatch.
- ✓ Your own custom metrics.

Each CloudWatch alarm watches a single metric and sends messages to Auto Scaling when the metric breaches a threshold that you specify in your policy.

Auto Scaling Basic Lifecycle



The basic lifecycle of instances within an Auto Scaling Group.

- ✓ The Scaling Group has a desired capacity of three instances.
- ✓ A Cloud Watch alarm trigger scaling events and policies scale the group at specific dates and times.
- ✓ The scaling policy launches an instance and attaches it to the Auto Scaling Group.
- ✓ A health check fails and triggers an alarm similar to scaling out.
- ✓ The instance is terminated.
- ✓ The instance is detached from the Auto Scaling Group.

THANK YOU