Amazon EC2 Auto Scaling FAQs

aws.amazon.com (https://aws.amazon.com/ec2/autoscaling/faqs/)

General

Q: What is Amazon EC2 Auto Scaling?

Amazon EC2 Auto Scaling is a fully managed service designed to launch or terminate Amazon EC2 instances automatically to help ensure you have the correct number of Amazon EC2 instances available to handle the load for your application. Amazon EC2 Auto Scaling helps you maintain application availability through fleet management for EC2 instances, which detects and replaces unhealthy instances, and by scaling your Amazon EC2 capacity up or down automatically according to conditions you define. You can use Amazon EC2 Auto Scaling to automatically increase the number of Amazon EC2 instances during demand spikes to maintain performance and decrease capacity during lulls to reduce costs.

Q. When should I use Amazon EC2 Auto Scaling vs. AWS Auto Scaling?

You should use AWS Auto Scaling to manage scaling for multiple resources across multiple services. AWS Auto Scaling lets you define dynamic scaling policies for multiple EC2 Auto Scaling groups or other resources using predefined scaling strategies. Using AWS Auto Scaling to configure scaling policies for all of the scalable resources in your application is faster than managing scaling policies for each resource via its individual service console. It's also easier, as AWS Auto Scaling includes predefined scaling strategies that simplify the setup of scaling policies. You should also use AWS Auto Scaling if you want to create predictive scaling for EC2 resources.

You should use EC2 Auto Scaling if you only need to scale Amazon EC2 Auto Scaling groups, or if you are only interested in maintaining the health of your EC2 fleet. You should also use EC2 Auto Scaling if you need to create or configure Amazon EC2 Auto Scaling groups, or if you need to set up scheduled or step scaling policies (as AWS Auto Scaling supports only target tracking scaling policies).

EC2 Auto Scaling groups must be created and configured outside of AWS Auto Scaling, such as through the EC2 console, Auto Scaling API or via CloudFormation. AWS Auto Scaling can help you configure dynamic scaling policies for your existing EC2 Auto Scaling groups.

Q: What are the benefits of using Amazon EC2 Auto Scaling?

Amazon EC2 Auto Scaling helps to maintain your Amazon EC2 instance availability. Whether you are running one Amazon EC2 instance or thousands, you can use Amazon EC2 Auto Scaling to detect impaired Amazon EC2

instances, and replace the instances without intervention. This ensures that your application has the compute capacity that you expect. You can use Amazon EC2 Auto Scaling to automatically scale your Amazon EC2 fleet by following the demand curve for your applications, reducing the need to manually provision Amazon EC2 capacity in advance. For example, you can set a condition to add new Amazon EC2 instances in increments to the ASG when the average utilization of your Amazon EC2 fleet is high; and similarly, you can set a condition to remove instances in increments when CPU utilization is low. You can also use Amazon CloudWatch to send alarms to trigger scaling activities and Elastic Load Balancing (ELB) to distribute traffic to your instances within the ASG. If you have predictable load changes, you can set a schedule through Amazon EC2 Auto Scaling to plan your scaling activities. Amazon EC2 Auto Scaling enables you to run your Amazon EC2 fleet at optimal utilization.

Q: What is fleet management and how is it different from dynamic scaling?

If your application runs on Amazon EC2 instances, then you have what's referred to as a 'fleet'. *Fleet management* refers to the functionality that automatically replaces unhealthy instances and maintains your fleet at the desired capacity. Amazon EC2 Auto Scaling fleet management ensures that your application is able to receive traffic and that the instances themselves are working properly. When Auto Scaling detects a failed health check (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/monitoring-system-instance-status-check.html), it can replace the instance automatically.

The *dynamic scaling* capabilities of Amazon EC2 Auto Scaling refers to the functionality that automatically increases or decreases capacity based on load or other metrics. For example, if your CPU spikes above 80% (and you have an alarm setup) Amazon EC2 Auto Scaling can add a new instance dynamically.

Q: What is target tracking?

Target tracking is a new type of scaling policy that you can use to set up dynamic scaling for your application in just a few simple steps. With target tracking, you select a load metric for your application, such as CPU utilization or request count, set the target value, and Amazon EC2 Auto Scaling adjusts the number of EC2 instances in your ASG as needed to maintain that target. It acts like a home thermostat, automatically adjusting the system to keep the environment at your desired temperature. For example, you can configure target tracking to keep CPU utilization for your fleet of web servers at 50%. From there, Amazon EC2 Auto Scaling launches or terminates EC2 instances as required to keep the average CPU utilization at 50%.

Q: What is an EC2 Auto Scaling group (ASG)?

An Amazon EC2 Auto Scaling group (ASG) contains a collection of EC2 instances that share similar characteristics and are treated as a logical grouping for the purposes of fleet management and dynamic scaling. For example, if a single application operates across multiple instances, you might want to increase the number of instances in that group to improve the performance of the application, or decrease the number of instances to reduce costs when demand is low. Amazon EC2 Auto Scaling will automaticallly adjust the number of instances in the group to maintain a fixed number of instances even if a instance becomes unhealthy, or based on criteria that you specify. You can find more information about ASG in the Amazon EC2 Auto Scaling User Guide (http://docs.aws.amazon.com/autoscaling

/latest/userguide/AutoScalingGroup.html).

Q: What happens to my Amazon EC2 instances if I delete my ASG?

If you have an EC2 Auto Scaling group (ASG) with running instances and you choose to delete the ASG, the instances will be terminated and the ASG will be deleted.

Q: How do I know when EC2 Auto Scaling is launching or terminating the EC2 instances in an EC2 Auto Scaling group?

When you use Amazon EC2 Auto Scaling to scale your applications automatically, it is useful to know when EC2 Auto Scaling is launching or terminating the EC2 instances in your EC2 Auto Scaling group. Amazon SNS (https://aws.amazon.com/sns/) coordinates and manages the delivery or sending of notifications to subscribing clients or endpoints. You can configure EC2 Auto Scaling to send an SNS notification whenever your EC2 Auto Scaling group scales. Amazon SNS can deliver notifications as HTTP or HTTPS POST, email (SMTP, either plain-text or in JSON format), or as a message posted to an Amazon SQS queue. For example, if you configure your EC2 Auto Scaling group to use the autoscaling: EC2_INSTANCE_TERMINATE notification type, and your EC2 Auto Scaling group terminates an instance, it sends an email notification. This email contains the details of the terminated instance, such as the instance ID and the reason that the instance was terminated.

For more information read Getting SNS Notifications when your EC2 Auto Scaling Group Scales (http://docs.aws.amazon.com/autoscaling/latest/userguide/ASGettingNotifications.html).

Q: What is a launch configuration?

A launch configuration is a template that an EC2 Auto Scaling group uses to launch EC2 instances. When you create a launch configuration, you specify information for the instances such as the ID of the Amazon Machine Image (AMI), the instance type, a key pair, one or more security groups, and a block device mapping. If you've launched an EC2 instance before, you specified the same information in order to launch the instance. When you create an EC2 Auto Scaling group, you must specify a launch configuration. You can specify your launch configuration with multiple EC2 Auto Scaling groups. However, you can only specify one launch configuration for an EC2 Auto Scaling group at a time, and you can't modify a launch configuration after you've created it. Therefore, if you want to change the launch configuration for your EC2 Auto Scaling group, you must create a launch configuration and then update your EC2 Auto Scaling group with the new launch configuration. When you change the launch configuration for your EC2 Auto Scaling group, any new instances are launched using the new configuration parameters, but existing instances are not affected. You can see the launch configurations (https://docs.aws.amazon.com/autoscaling/ec2/userguide/LaunchConfiguration.html) section of the EC2 Auto Scaling User Guide for more details.

Q: How many instances can an EC2 Auto Scaling group have?

You can have as many instances in your EC2 Auto Scaling group as your EC2 quota allows.

Q: What happens if a scaling activity causes me to reach my Amazon EC2 limit of instances?

Amazon EC2 Auto Scaling cannot scale past the Amazon EC2 limit of instances that you can run. If you need more Amazon EC2 instances, complete the Amazon EC2 instance request form (https://aws.amazon.com/contact-us/ec2-request/).

Q: Can EC2 Auto Scaling groups span multiple AWS regions?

EC2 Auto Scaling groups are regional constructs. They can span Availability Zones, but not AWS regions.

Q: How can I implement changes across multiple instances in an EC2 Auto Scaling group?

You can use AWS CodeDeploy or CloudFormation to orchestrate code changes to multiple instances in your EC2 Auto Scaling group.

Q: If I have data installed in an EC2 Auto Scaling group, and a new instance is dynamically created later, is the data copied over to the new instances?

Data is not automatically copied from existing instances to new instances. You can use lifecycle hooks (http://docs.aws.amazon.com/autoscaling/latest/userguide/lifecycle-hooks.html) to copy the data, or an Amazon RDS (https://aws.amazon.com/rds/) database including replicas.

Q: When I create an EC2 Auto Scaling group from an existing instance, does it create a new AMI (Amazon Machine Image)?

When you create an Auto Scaling group from an existing instance, it does not create a new AMI. For more information see Creating an Auto Scaling Group Using an EC2 Instance (http://docs.aws.amazon.com/autoscaling/latest/userguide/create-asg-from-instance.html).

Q: How does Amazon EC2 Auto Scaling balance capacity?

Balancing resources across Availability Zones (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-regions-availability-zones.html) is a best practice for well-architected applications, as this greatly increases aggregate system availability. Amazon EC2 Auto Scaling automatically balances EC2 instances across zones when you configure multiple zones (http://docs.aws.amazon.com/autoscaling/latest/userguide/as-add-availability-zone.html) in your EC2 Auto Scaling group settings. Amazon EC2 Auto Scaling always launches new instances such that they are balanced between zones as evenly as possible across the entire fleet. What's more, Amazon EC2 Auto Scaling only launches into Availability Zones in which there is available capacity for the requested instance type.

Q: What are lifecycle hooks?

Lifecycle hooks let you take action before an instance goes into service or before it gets terminated. This can be

especially useful if you are not baking your software environment into an Amazon Machine Image (AMI). For example, launch hooks can perform software configuration on an instance to ensure that it's fully prepared to handle traffic before Amazon EC2 Auto Scaling proceeds to connect it to your load balancer. One way to do this is by connecting the launch hook to an AWS Lambda function that invokes RunCommand on the instance. Terminate hooks can be useful for collecting important data from an instance before it goes away. For example, you could use a terminate hook to preserve your fleet's log files by copying them to an Amazon S3 bucket when instances go out of service.

Visit lifecycle hooks (https://docs.aws.amazon.com/autoscaling/ec2/userguide/lifecycle-hooks.html) in our Amazon EC2 Auto Scaling User Guide for more information.

Q: What are the characteristics of an "unhealthy" instance?

An unhealthy instance is one where the hardware has become impaired for some reason (bad disk, etc.), or it is not passing a user-configured ELB health check. Amazon EC2 Auto Scaling performs health checks (http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/monitoring-system-instance-status-check.html) on each individual EC2 instance at regular intervals, and if the instance is connected to an Elastic Load Balancing load balancer, it can also perform ELB health checks (http://docs.aws.amazon.com/elasticloadbalancing/latest/classic/elb-healthchecks.html).

Q: Can I customize a health check?

Yes, there is an API called *SetInstanceHealth* that allows you to change an instance's state to UNHEALTHY, which will then result in a termination and replacement.

Q: Can I suspend health checks (for example, to evaluate unhealthy instances)?

Yes, you can temporarily suspend Amazon EC2 Auto Scaling health checks by using the SuspendProcesses API. You can use the ResumeProcesses API to resume automatic health checks.

Q: Which health check type should I select?

If you are using Elastic Load Balancing (ELB) with your group, you should select an ELB health check. If you're not using ELB with your group, you should select the EC2 health check.

Q: Can I use Amazon EC2 Auto Scaling for health checks and to replace unhealthy instances if I'm not using Elastic Load Balancing (ELB)?

You don't have to use ELB to use Auto Scaling. You can use the EC2 health check to identify and replace unhealthy instances.

Q: Do the Elastic Load Balancing (ELB) health checks work with Application Load Balancers and Network Load

Balancers? Will an instance be marked as unhealthy if any target group associated with it becomes unhealthy?

Yes, Amazon EC2 Auto Scaling works with Application Load Balancers and Network Load Balancers including their health check feature.

Q: Is there any way to use Amazon EC2 Auto Scaling to only add a volume without adding an instance?

A volume is attached to a new instance when it is added. Amazon EC2 Auto Scaling doesn't automatically add a volume when the existing one is approaching capacity. You can use the EC2 API to add a volume to an existing instance.

Q: What does the term “stateful instances” refer to?

When we refer to a stateful instance, we mean an instance that has data on it, which exists only on that instance. In general, terminating a stateful instance means that the data (or state information) on the instance is lost. You may want to consider using lifecycle hooks to copy the data off of a stateful instance before it’s terminated, or enable instance protection to prevent Amazon EC2 Auto Scaling from terminating it.

Replacing Impaired Instances

Q: How does Amazon EC2 Auto Scaling replace an impaired instance?

When an impaired instance fails a health check, Amazon EC2 Auto Scaling automatically terminates it and replaces it with a new one. If you’re using an Elastic Load Balancing load balancer, Amazon EC2 Auto Scaling gracefully detaches the impaired instance from the load balancer before provisioning a new one and attaching it to the load balancer. This is all done automatically, so you don’t need to respond manually when an instance needs replacing.

Q: How do I control which instances Amazon EC2 Auto Scaling terminates when scaling in, and how do protect data on an instance?

With each Amazon EC2 Auto Scaling group, you control when Amazon EC2 Auto Scaling adds instances (referred to as scaling out) or remove instances (referred to as scaling in) from your group. You can scale the size of your group manually by attaching and detaching instances, or you can automate the process through the use of a scaling policy. When you have Amazon EC2 Auto Scaling automatically scale in, you must decide which instances Amazon EC2 Auto Scaling should terminate first. You can configure this through the use of a termination policy. You can also use instance protection to prevent Amazon EC2 Auto Scaling from selecting specific instances for termination when scaling in. If you have data on an instance, and you need that data to be persistent even if your instance is scaled in, then you can use a service like S3, RDS, or DynamoDB, to make sure that it is stored off the instance.

Q: How long is the turn-around time for Amazon EC2 Auto Scaling to spin up a new instance at inService state after detecting an unhealthy server?

The turnaround time is within minutes. The majority of replacements happen within less than 5 minutes, and on

average it is significantly less than 5 minutes. It depends on a variety of factors, including how long it takes to boot up the AMI of your instance.

Q: If Elastic Load Balancing (ELB) determines that an instance is unhealthy, and moved offline, will the previous requests sent to the failed instance be queued and rerouted to other instances within the group?

When ELB notices that the instance is unhealthy, it will stop routing requests to it. However, prior to discovering that the instance is unhealthy, some requests to that instance will fail.

Q: If you don't use Elastic Load Balancing (ELB) how would users be directed to the other servers in a group if there was a failure?

You can integrate with Route53 (which Amazon EC2 Auto Scaling does not currently support out of the box, but many customers use). You can also use your own reverse proxy, or for internal microservices, can use service discovery solutions.

Security

Q: How do I control access to Amazon EC2 Auto Scaling resources?

Amazon EC2 Auto Scaling integrates with AWS Identity and Access Management (https://aws.amazon.com/iam/) (IAM), a service that enables you to do the following:

- Create users and groups under your organization's AWS account
- Assign unique security credentials to each user under your AWS account
- Control each user's permissions to perform tasks using AWS resources
- Allow the users in another AWS account to share your AWS resources
- Create roles for your AWS account and define the users or services that can assume them
- Use existing identities for your enterprise to grant permissions to perform tasks using AWS resources

For example, you could create an IAM policy that grants the Managers group permission to use only the *DescribeAutoScalingGroups*, *DescribeLaunchConfigurations*, *DescribeScalingActivities*, and *DescribePolicies* API operations. Users in the Managers group could then use those operations with any Amazon EC2 Auto Scaling groups and launch configurations. With Amazon EC2 Auto Scaling resource-level permissions, you can restrict access to a particular EC2 Auto Scaling group or launch configuration.

For more information, see the Controlling Access to Your Auto Scaling Resources (http://docs.aws.amazon.com/autoscaling/latest/userguide/control-access-using-iam.html) section of the Amazon EC2 Auto Scaling user guide.

Q: Can you define a default admin password on Windows instances with Amazon EC2 Auto Scaling?

You can use the Key Name parameter to CreateLaunchConfiguration to associate a key pair with your instance. You can then use the *GetPasswordData* API in EC2. This is also possible through the AWS Management Console.

Q: Are CloudWatch agents automatically installed on EC2 instances when you create an Amazon EC2 Auto Scaling group?

If your AMI contains a CloudWatch agent, it's automatically installed on EC2 instances when you create an EC2 Auto Scaling group. With the stock Amazon Linux AMI, you need to install it (recommended, via yum).

Cost Optimization

Q: Can I create a single ASG to scale instances across different purchase options?

Yes. You can provision and automatically scale EC2 capacity across different EC2 instance types, Availability Zones, and On-Demand, RIs and Spot purchase options in a single Auto Scaling Group. You have the option to define the desired split between On-Demand and Spot capacity, select which instance types work for your application, and specify preference for how EC2 Auto Scaling should distribute the ASG capacity within each purchasing model.

Q: Can I use ASGs to launch and manage just Spot Instances or just On-Demand instances and RIs?

Yes. You can configure your ASG specifying all capacity to be only Spot instances or all capacity to be only On-Demand instances and RIs.

Q: Can I have a base capacity with On-Demand instances and RIs, and scale my ASG out on Spot instances?

Yes. When setting up an ASG to combine purchasing models, you can specify the base capacity of the group to be fulfilled by On-Demand instances. As the ASG scales in or scale out, EC2 Auto Scaling ensures the base capacity be fulfilled with On-Demand instances and anything beyond that be fulfilled with either only Spot instances or a specified percentage mix of On-Demand or Spot instances.

Q: Can I modify the configuration of an ASG to update the different properties pertaining to combining purchasing models and specifying multiple instance types?

Yes. Similar to other ASG parameters, customers can update an existing ASG to modify one or all parameters pertaining to combining purchasing models and specifying multiple instance types, including instance types, prioritization order for On-Demand instances, percentage split between On-Demand and Spot instances, and allocation strategy.

Q: Can I use RI discounts with On-Demand Instances in an ASG?

Yes. For example, if you have RIs for C4 instances and EC2 Auto Scaling launches a C4 you will receive your RI pricing for On-Demand Instances.

Q: Can I specify instances of different sizes (CPU cores, memory) in my Auto Scaling group?

Yes. You can specify any instance type available in a region. Note that all instance types will be treated as the same

weight.

Q: What if the instance types I like are not available in an Availability Zone?

If none of the specified instance types are available in an Availability Zone, Auto Scaling will retarget the launches in other Availability Zones associated with the Auto Scaling group. Auto Scaling will always prefer keeping your compute balanced across Availability Zones and retarget if all instance types are not available in an Availability Zone.

Pricing

Q: What are the costs for using Amazon EC2 Auto Scaling?

Amazon EC2 Auto Scaling fleet managment for EC2 instances carries no additional fees. The dynamic scaling capabilities of Amazon EC2 Auto Scaling are enabled by Amazon CloudWatch and also carry no additional fees. Amazon EC2 and Amazon CloudWatch service fees apply and are billed separately.

Learn more about Amazon EC2 Auto Scaling pricing

Visit the pricing page
Ready to get started?
Sign up (https://portal.aws.amazon.com/gp/aws/developer/registration/index.html)
Have more questions?
Contact us (https://aws.amazon.com/contact-us/)
Page Content
General Replacing Impaired Instances Security Cost Optimization Pricing

aws.amazon.com (https://aws.amazon.com/ec2/autoscaling/faqs/)