Auto Scaling

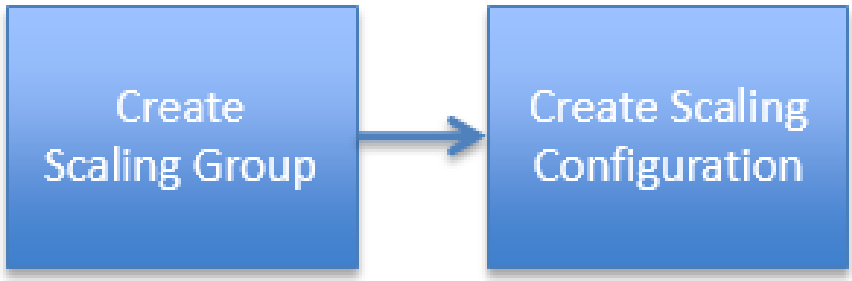
Quick Start

Quick Start

Create a scaling solution by two steps

This quick start guide explains how to use a simple scaling solution in Auto Scaling to create an ECS instance, automatically add it to a Server Load Balancer instance, and automatically add the instance IP address to the RDS access whitelist.

To create an Auto Scaling solution, follow the two steps as shown:



Create a scaling group

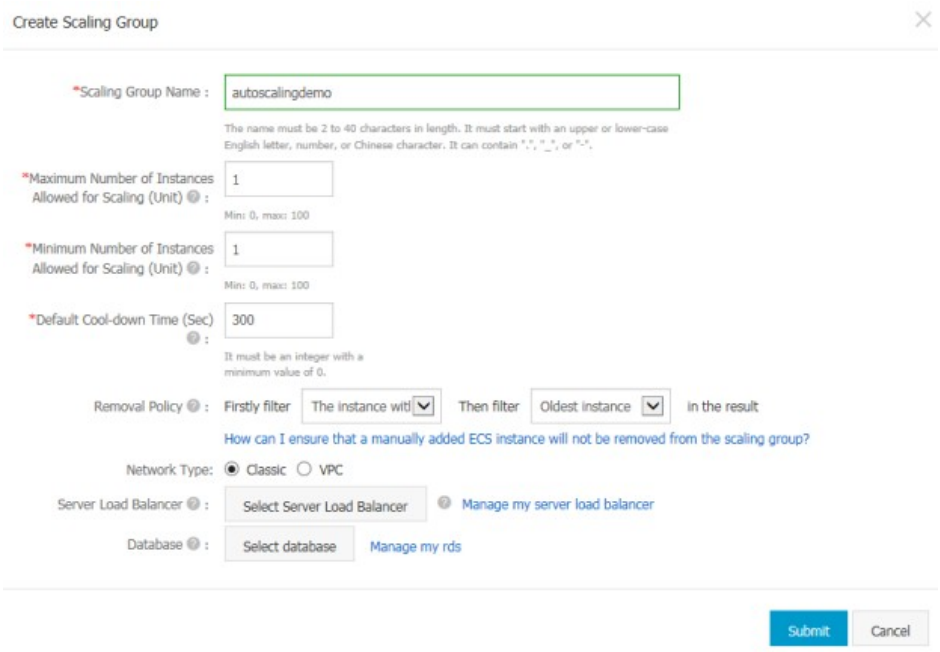
A scaling group is a collection of ECS instances with the same configuration deployed in an application scenario. A scaling group defines the minimum and maximum number of ECS instances, and associated Server Load Balancer instances, RDS instances, and other attributes.

On the **Scaling Group List** page, click **Create Scaling Group**.

Select the **Region** for the scaling group and enter the **Scaling Group Name**.

Set **Maximum Number of Instances Allowed for Scaling and Minimum Number of Instances Allowed for Scaling** to **1**. This automatically creates one ECS instance after the overall scaling solution is created.

Select the **Server Load Balancer** instance. Health check must be enabled for all listener ports configured for the specified Server Load Balancer instance.



Select a **Database** instance.

After making your desired RDS instance selections, click **Submit**.

Create scaling configuration

Scaling configuration defines the ECS instances information used for Auto Scaling. When automatically adding ECS instances to a scaling group, Auto Scaling will create ECS instances based on the scaling configuration.

On the **Scaling Group List** page, click **Manage** next to the desired scaling group.

On the **Scaling Configuration** page, click **Create** > **Scaling Configuration**.

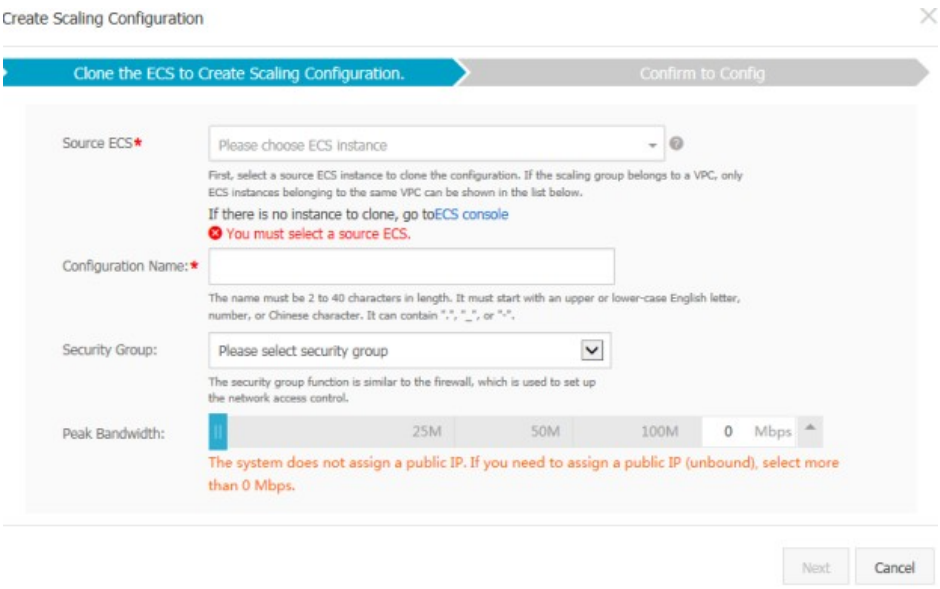
Choose the required ECS template.

Enter the scaling configuration name.

Choose a security group.

Choose the needed bandwidth.

Click **Next** and go to **Confirm to Config**.



Create a scaling solution by five steps

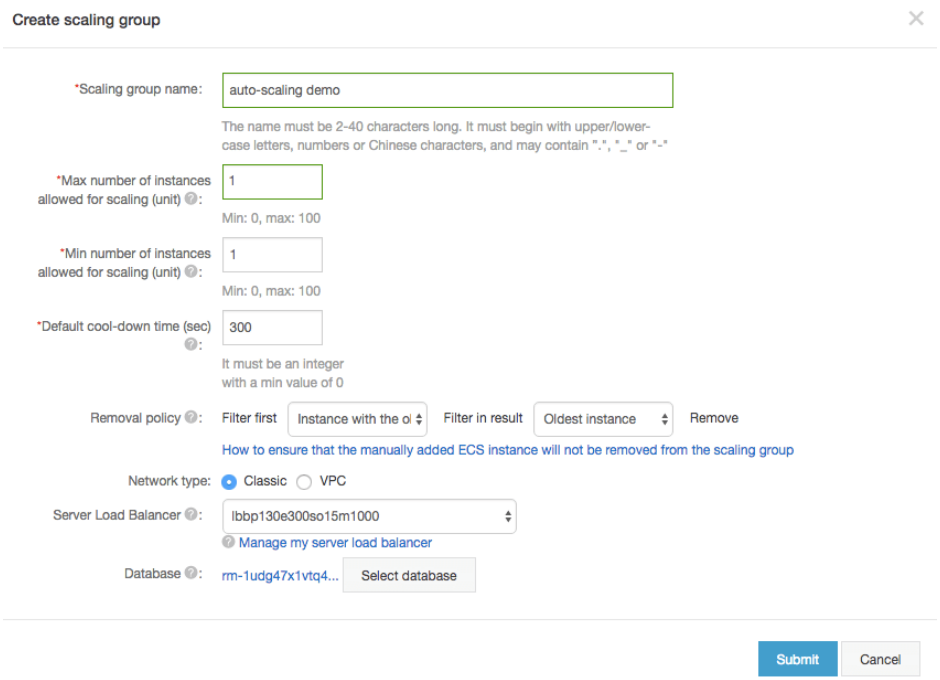
This guide explains how to create and configure an overall scaling solution, including timing, dynamic, custom, and fixed modes.

To create an Auto Scaling solution, follow the five steps as follows:

1. Create a scaling group

A scaling group is a collection of ECS instances with the same configuration deployed in an application scenario. A scaling group defines the minimum and maximum number of ECS instances, associated Server Load Balancer instances, RDS instances, and other attributes.

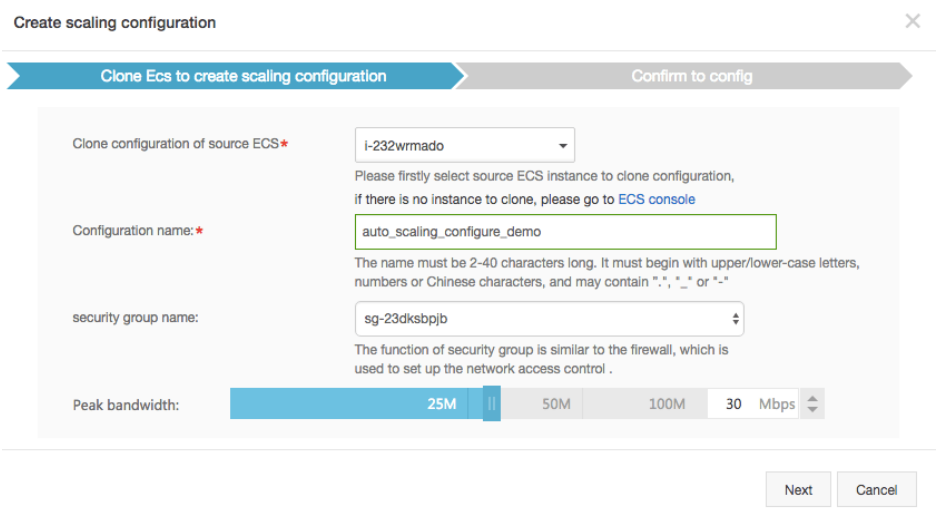
1. On the **Scaling group management** page, click **Create scaling group**.
2. Select the **Region** for the scaling group and enter the **Scaling group name**.
3. Set **Max number of instances allowed for scaling** and **Min number of instances allowed for scaling** to **1**. This automatically creates one ECS instance after the overall scaling solution is created.
4. Select the **Sever Load Balancer** instance. Health check must be enabled for all listener ports configured for the specified Server Load Balancer instance.
5. Select a **Database** instance to display the **Select RDS database** dialog box.
6. After configuration, click **Submit**.



2. Create scaling configuration

Scaling configuration defines the ECS instance information used for Auto Scaling. When automatically adding ECS instances to a scaling group, Auto Scaling will create ECS instances based on the scaling configuration.

1. On the **Scaling group management** page, click **Manage** next to the desired scaling group.
2. On the **Scaling configuration** page, click **Create scaling configuration**.
3. Choose the source ECS instance used to clone configuration.
4. Enter a name for the scaling configuration. For example, auto_scaling_configure_demo.
5. Select a security group for network access control.
6. Select the needed peak bandwidth.
7. Click **Next**.

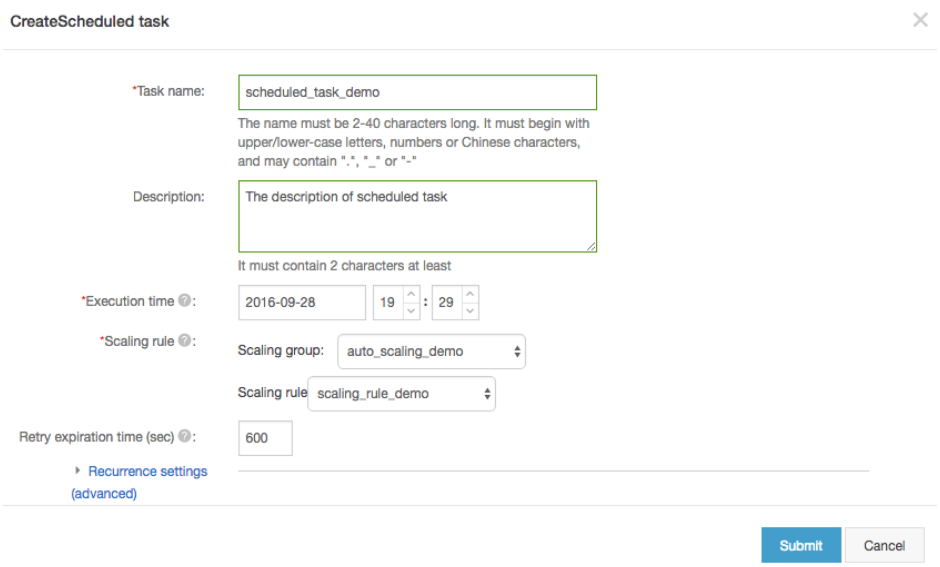


3. Add an existing ECS instance

1. On the **Scaling group management** page, click **Manage** next to the desired scaling group.
2. On the **ECS instance list** page, click **Add existing list**.
3. Select the ECS instance to be added, and click **Add**.

4. Create a scheduled task

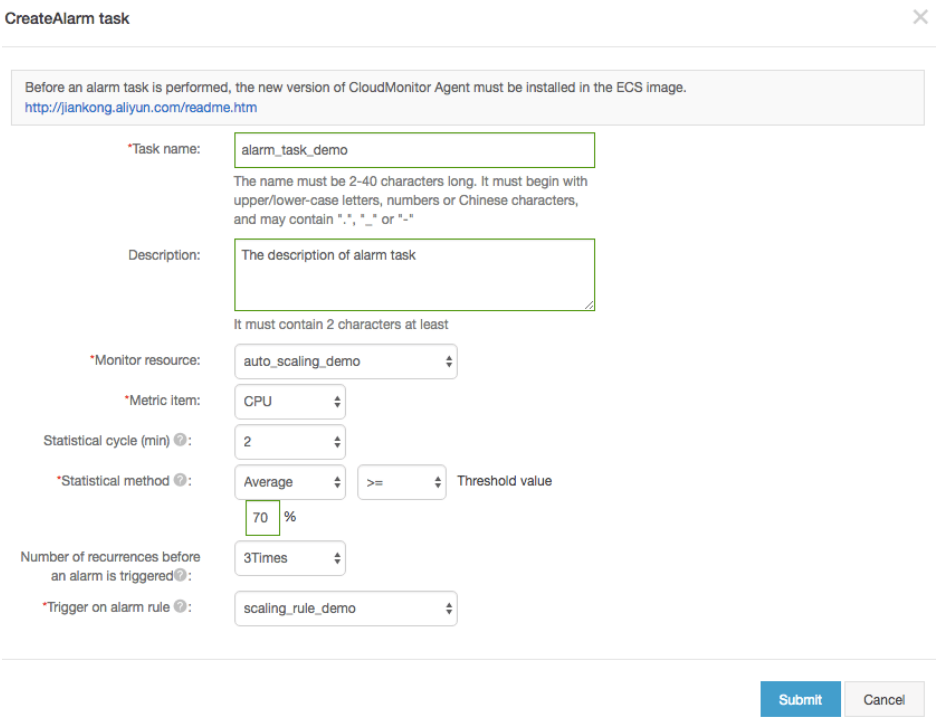
1. Click **Scheduled task** under **Auto-trigger task management** to display the **Scheduled Task** page.
2. Click **Create scheduled task** to display the **Create scheduled task** dialog box.
3. Enter the task name.
4. Enter the execution time. If recurrence is not set, the task is only executed once at the designated date and time. Otherwise, the task is executed periodically at the specified time.
5. Enter the recurrence.
6. Select a scaling group and the scaling rule to be triggered by the scheduled task.
7. Click **Submit**. The scheduled task is displayed on the **Scheduled Task** page.



5. Create an alarm task

An alarm task can be created to automatically add or remove ECS instances according to CloudMonitor performance indicators, such as CPU and memory usage. For alarms to be triggered, the latest version of CloudMonitor Agent must be installed in the ECS image.

1. Click **Alarm task** under **Auto-trigger task management** to display the alarm task list page.
2. Click **Create alarm task** to display the alarm task creation dialog box.
3. Enter the task name.
4. Select the scaling group to be monitored.
5. Select the item to be monitored.
6. Enter the statistical period. The finer the granularity of the statistical cycle, the more sensitive the alarm trigger mechanism will be.
7. Enter the statistical method.
8. Enter the number of recurrences before an alarm is triggered.
9. Select the scaling rule triggered by the alarm.
10. Click **Submit**. The alarm task will now be displayed on the **Alarm Task** page.

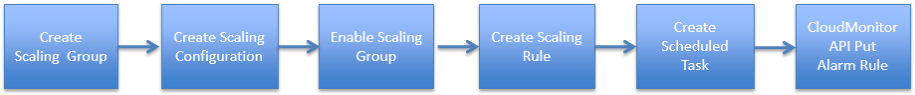


API quick start

Process introduction

This example explains how to use OpenAPIs to create and configure overall scaling solutions, including timed, dynamic, custom, and fixed mode scaling solutions.

To create a complete Auto Scaling solution, follow the steps shown in the following diagram. The first three steps are used to create a simple scaling solution:



Create a scaling group. Configure the minimum (Min Size) and maximum (Max Size) number of ECS instances for scaling, and select the associated Server Load Balancer and RDS instances.

Create scaling configuration. Specify the ECS instances attributes for Auto Scaling, such as Image ID and Instance Type.

Enable the scaling group with the scaling configuration created in Step 2.

Create a scaling rule. For example, add n (number) ECS instances.

Create a scheduled task. For example, trigger the scaling rule created in Step 4 at 12:00 AM.

Create an alarm task (CloudMonitor API PutAlarmRule). For example, add 1 ECS instance when the CPU usages are greater than or equal to 80%.

Create a scaling group

In the CreateScalingGroup operation, configure the minimum value (MinSize) and maximum value (MaxSize) of scaling resources and associate the necessary Server Load Balancer and RDS instances.

Request example

```
http://ess.aliyuncs.com/?Action=CreateScalingGroup
&RegionId=cn-qingdao
&MaxSize=20
&MinSize=2
&LoadBalancerId=147b46d767c-cn-qingdao-cm5-a01
&DBInstanceId.1=rdszzzyyunybaeu
&DBInstanceId.2=rdsia3u3yia3u3y
&<Public Request Parameters>
```

Return example

```
<CreateScalingGroupResponse>
<ScalingGroupId>dP8VqSd9ENXPc0ciVmbcrBT1</ScalingGroupId>
<RequestId>536E9CAD-DB30-4647-AC87-AA5CC38C5382</RequestId>
</CreateScalingGroupResponse>
```

Create scaling configuration

To create scaling configuration, specify the ECS instances attributes for Auto Scaling, such as Image ID and Instance Type. For requests, the ScalingGroupId returned in Step 1 must be specified.

Request example

```
http://ess.aliyuncs.com/?Action=CreateScalingConfiguration
&ScalingGroupId=dP8VqSd9ENXPc0ciVmbcrBT1
&SecurityGroupId=sg-280ih3w4b
&ImageId=centos6u5_64_20G_aliaegis_20140703.vhd
&InstanceType=ecs.t1.xsmall
&<Public Request Parameters>
```

Return example

```
<CreateScalingConfigurationResponse>
<ScalingConfigurationId>eOs27Kb0oXvQcUYjEGelJqUy</ScalingConfigurationId>
<RequestId>5CC0AD41-08ED-4559-A683-6F56355FE068</RequestId>
</CreateScalingConfigurationResponse>
```

Enable a scaling group

Use the scaling configuration created in Step 2 to perform the EnableScalingGroup operation. An existing ECS instance can also be added in this step.

Request example

```
http://ess.aliyuncs.com/?Action=EnableScalingGroup
&ScalingGroupId=dP8VqSd9ENXPc0ciVmbcrBT1
&ActiveScalingConfigurationId=eOs27Kb0oXvQcUYjEGelJqUy
&InstanceId.1=i-283vvyytn
&<Public Request Parameters>
```

Return example

```
< EnableScalingGroupResponse>
<RequestId>6469DCD0-13AC-487E-85A0-CE4922908FDE</RequestId>
</ EnableScalingGroupResponse>
```

Create a scaling rule

In the CreateScalingRule operation, create scaling rules, such as **Add 1 ECS instance**. For requests, the ScalingGroupId returned in Step 1 must be specified.

Request example

```
http://ess.aliyuncs.com/?Action=CreateScalingRule
&ScalingGroupId=dP8VqSd9ENXPc0ciVmbcrBT1
&AdjustmentType=QuantityChangeInCapacity
&AdjustmentValue=1
&<Public Request Parameters>
```

Return example

```
<CreateScalingRuleResponse>
<ScalingRuleAri>
ari:acs:ess:cn-qingdao:1344371:scalingrule/eMKWG8SRNb9dBLAjweNI1Ik
</ScalingRuleAri>
<ScalingRuleId>eMKWG8SRNb9dBLAjweNI1Ik</ScalingRuleId>
<RequestId>570C84F4-A434-488A-AFA1-1E3213682B33</RequestId>
</CreateScalingRuleResponse>
```

Create a scheduled task

Create a scheduled task, for example, trigger the scaling rule created in Step 4 at 12:00 AM. For requests, the ScalingRuleAri returned in Step 4 must be specified.

Request example

```
http://ess.aliyuncs.com/?Action=CreateScheduledTask
&RegionId=cn-qingdao
&LaunchTime=2014-08-17T12:00Z
&RecurrenceType=Daily
&RecurrenceValue=1
&RecurrenceEndTime=2014-09-17T16:55Z
&ScheduledAction=ari:acs:ess:cn-qingdao:1344371:scalingrule/eMKWG8SRNb9dBLAjweNI1Ik
&<Public Request Parameters>
```

Return example

```
<CreateScheduledTaskResponse>
<ScheduledTaskId>edRtShc57WGXdt8TlPbrjsnV</ScheduledTaskId>
<RequestId>0F02D931-2B12-44D7-A0E9-39925C13D15E</RequestId>
</CreateScheduledTaskResponse>
```

Limits

An application deployed in an ECS instance for Auto Scaling must be stateless and horizontally scalable.

ECS instances for Auto Scaling cannot be used to save application status (for example, session) or related data (for example, databases and logs) because Auto Scaling automatically releases ECS instances. Status information can be saved to an independent state server, database (for example, RDS), or centralized log storage (for example, SLS).

An ECS instance for Auto Scaling is not automatically added to the OCS whitelist. You can manually add it if necessary.

Auto Scaling does not support vertical scaling. It cannot automatically upgrade or downgrade the CPU, memory, or bandwidth for an ECS instance.