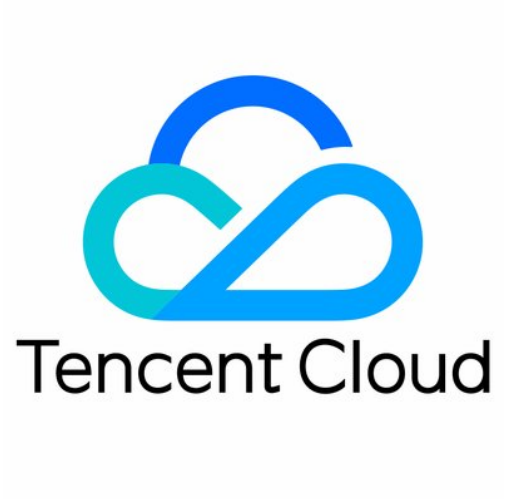
Auto Scaling

What is auto-scaling?

Product Introduction



Contents

What is auto-scaling? Product Introduction

Last updated：2017-11-28 10:29:01

What is Auto Scaling (AS)?

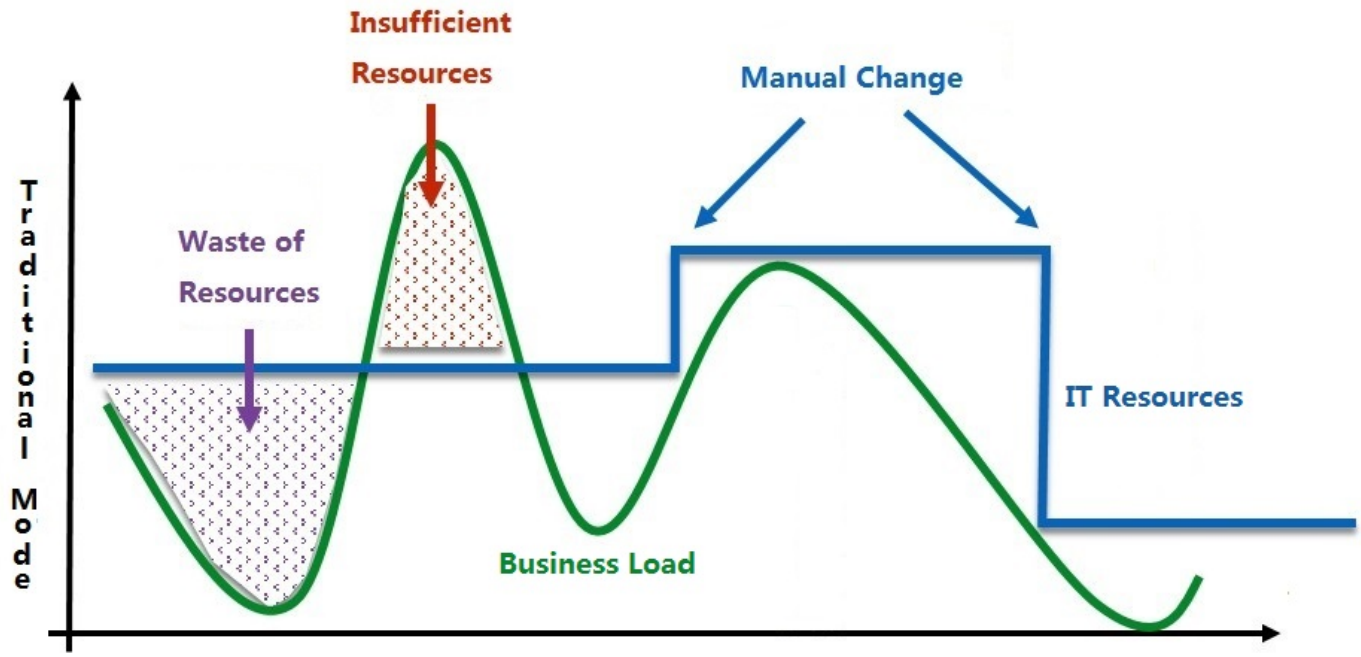
Auto Scaling (AS) can automatically adjust CVM computing resources according to your business needs and policies to ensure that you have an appropriate number of CVM instances to handle your application load. For Web applications, intelligent scaling can help control cost and manage resources. When requests increase, more servers are added to handle additional load. When the requests reduce, unnecessary servers are removed.

You only need to set policies for expending and reducing capacity. When the expanding policy is trigger, AS automatically increase servers to maintain the performance. When the demand decreases, AS reduces servers according to your reducing policy, so as to minimize your cost.

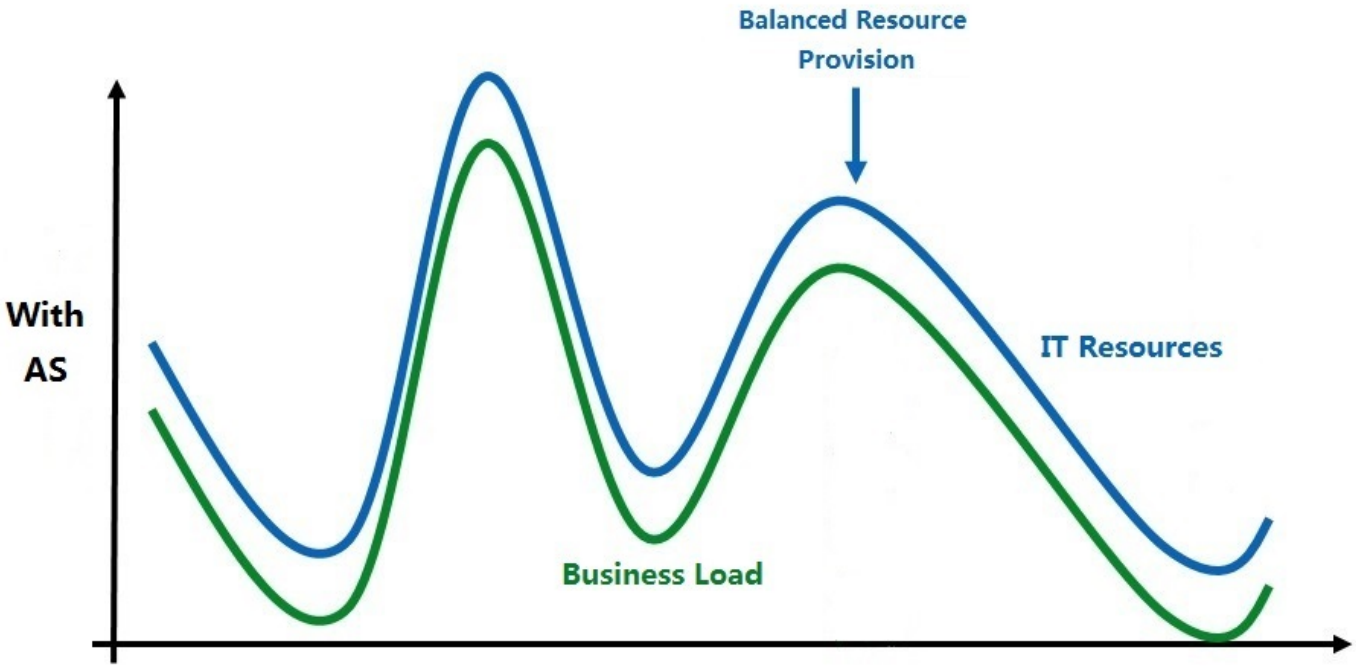
As shown in the figures below, by using AS, your cluster can always keep an appropriate number of resources and stay healthy. You will get rid of the following troubles in the traditional model:

- Insufficient machines due to a surge in business or a CC attack, resulting in no response from your service
- Estimating resources based on peak traffic while the traffic is rarely peaked, causing a waste of resources
- Personal surveillance and frequent handling of capacity alarms, which require multiple manual changes

Cluster maintenance in the traditional model:



Effects after using AS:



How AS Works

In common Web application services, your cluster usually runs multiple copies of an application to meet client traffic. For example, the frontend server cluster at the access layer, the application server cluster at the logical layer, and the backend cache server cluster. Every instance can process client requests.

The instances are similar or identical and are usually quantity adjustable. You can add these similar or identical machines to one scaling group for management:

- You can specify the minimum instances in each scaling group, and AS will ensure that the instances in the group will never be less than the minimum number.
- You can specify the maximum instances in each scaling group, and AS will ensure that the instances in the group will never be more than the maximum number.
- You can specify a scaling policy, and AS will start or terminate the instances when the demands for an application increase or decrease. There are two kinds of scaling policies:
  a) Alarm trigger policy: expand capacity dynamically according to specified conditions (for example, when CPU utilization of a server in the scaling group is over than 60%)
  b) Scheduled scaling policy: expand capacity at a specified time (for example, 21:00 every night)
- After setting the policy, you can also set scaling activity notification. When a scaling activity occurs, AS will inform you via email, SMS and internal message. You only need to check the notifications from AS instead of focusing on the changes of your business request volume all the time.
- You can also specify the number of machines required via one click at any time, or add existing machines to the scaling group for joint management.

Basic Concepts of AS

AS products have the following basic concepts:

- Scaling group
- Scaling configuration
- Scaling policy
- Cooldown period

1. Scaling Groups

A scaling group is a collection of CVM instances following the same rules and serving the same scenario. A scaling group defines attributes such as the maximum and minimum numbers of CVM instances, and its associated load balancer instances.

2. Scaling Configuration

Scaling configuration is a template for automatic creation of CVM. It contains image ID, CVM instance type, system disk/data disk types and capacities, key pair, security group, etc.

Scaling configuration must be specified when the scaling group is created. Once the scaling configuration is created, its attributes cannot be edited.

3. Scaling Policies

A scaling policy defines the conditions for executing a scaling action. The trigger condition can be a time point or an alarm of cloud monitoring, and the action can be removing or adding a CVM.
There are two scaling policies:

- **Scheduled scaling policy**
  CVM instances will be automatically increased or reduced at a fixed time point, which can be repeated periodically.

- **Alarm scaling**
  CVM instances will be automatically increased or reduced based on cloud monitoring metrics such as CPU, memory and network traffic.

4. Cooldown Period

Cooldown period refers to a period of time when the corresponding scaling group is locked after a scaling activity (adding or removing CVM instances) is completed. During this period, no scaling activities are performed with the scaling group. The cooldown period can be specified within 0-999,999 (seconds).

Product Advantages

Last updated：2017-04-14 15:19:19

Benefits	With Auto Scaling (AS)	Without AS
Automation	**Automatically scale instances without human intervention** Auto Scaling can automatically and dynamically create and release CVM instances based on your business load to help you ensure that your application always has the right amount of capacity to handle the current traffic demands. No human intervention is needed throughout the process, freeing you from the burden of manual deployment. For example, you can set a scaling policy to add new CVM instances to the scaling group when the CPU utilization is high, and the added CVM instances will be charged by seconds. Similarly, you can also set a policy to remove instances from the scaling group when the CPU utilization is low. If your load changes are predictable, you can set a scheduled task to plan your scaling activities. The added instances can also directly be associated with the existing cloud load balancing (CLB) to allow the added instances in the scaling group to share the distributed traffic and to improve service availability. You can also send an alarm to the administrator to keep an eye on any abnormalities for you.	**Cumbersome manual operation** Manually create and terminate resources, and the cloud load balance needs to be configured manually; Manual operation is prone to error, which has impact on the business.
Costs effectiveness	**Appropriate scaling of instances to save costs** Auto Scaling helps you to cope with business situations with the most appropriate number of instances. When the demand increases, it can seamlessly and automatically add an appropriate amount of CVM instances, and when the demand decreases, it can automatically reduce the unnecessary CVM instances, which improves device utilization, and saves the costs of deployment and instance.	**Idle resources resulting in waste** Extra CVMs need to be reserved to ensure the application always has enough capacity to meet demand.

Benefits	With Auto Scaling (AS)	Without AS
Fault Tolerance	**Automatic detection of the system, timely fault tolerance** Auto Scaling automatically detects the health of instances. When an instance is detected to be unhealthy, Auto Scaling can create a healthy one to replace it, so that your application has the desired computing capacity to keep your business up and running.	**Inability of timely fault tolerance** Usually, an unhealthy instance is not replaced until a business interruption is discovered, which compromises the business availability.

Application Scenario

Last updated：2017-04-14 15:19:29

1. Deploying Capacity Scaling in Advance

If the user knows when capacity scaling is needed, he/she can configure Auto Scaling schedule policy in advance. By the configured time, the system will automatically increase or decrease the number of CVM instances without the need to wait.

2. Coping with Business Volume Surge with Low Cost

When the customer is faced with access peak, he/she will need to prepare servers in advance and prevent server overload caused by the sudden surge in CPU usage. The customer may decrease the number of servers according to the situation when the surge has passed. The user can configure Auto Scaling monitor policy in advance and the system will automatically determine whether CVM scale-out is needed according to the business monitoring metrics that are already configured. The system will automatically increase or decrease the number of CVM instances and complete load balancer configurations when the monitoring metric reaches certain thresholds. For customers, this not only saves cost, but also saves the effort to be constantly prepared for manual capacity scaling.

3. Replacing Unhealthy CVMs Automatically

Users need to constantly monitor the operation statuses of CVMs and take actions on unhealthy CVMs in time to prevent them from affecting their business. With Auto Scaling, the system will regularly perform health check on CVMs. When the system detects an abnormal instance, it will automatically create a new instance as replacement. This operation will be logged for users to view later.

Service Limits

Last updated：2018-01-25 09:59:48

- For now， AS is available in Beijing, Shanghai, Guangzhou, Hong Kong, Toronto and Singapore.
- Each user is able to create up to 20 scaling configurations for each region.
- Each user is able to create up to 20 scaling groups.
- A scaling group can only correspond to one scaling configuration.
- For all regions and scaling groups, each user can configure auto scaling for up to 30 CVM instances.
- Up to 100 scaling policies and 10 scheduled tasks can be created in each scaling group.
- The number of sub machines in scaling group cannot exceed the number of IPs that the VPC subnet is able to provide.
- Currently, auto scaling does not support configuration upgrade/degrade of CVMs (increasing/reducing CPU, memory and bandwidth).
- Auto scaling and scaling configuration are regional concepts, which means they can only enable/terminate CVM instances in the same region.