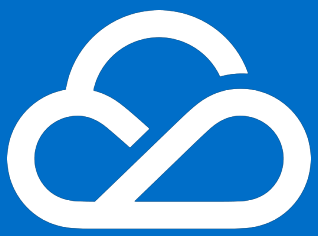
腾讯云弹性伸缩

什么是弹性伸缩 AS

产品文档



腾讯云

【版权声明】

【商标声明】

【服务声明】

本文档意在向客户介绍腾讯云全部或部分产品、服务的当时的整体概况，部分产品、服务的内容可能有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或模式的承诺或保证。

文档目录

什么是弹性伸缩 AS

产品概述

什么是弹性伸缩 AS？

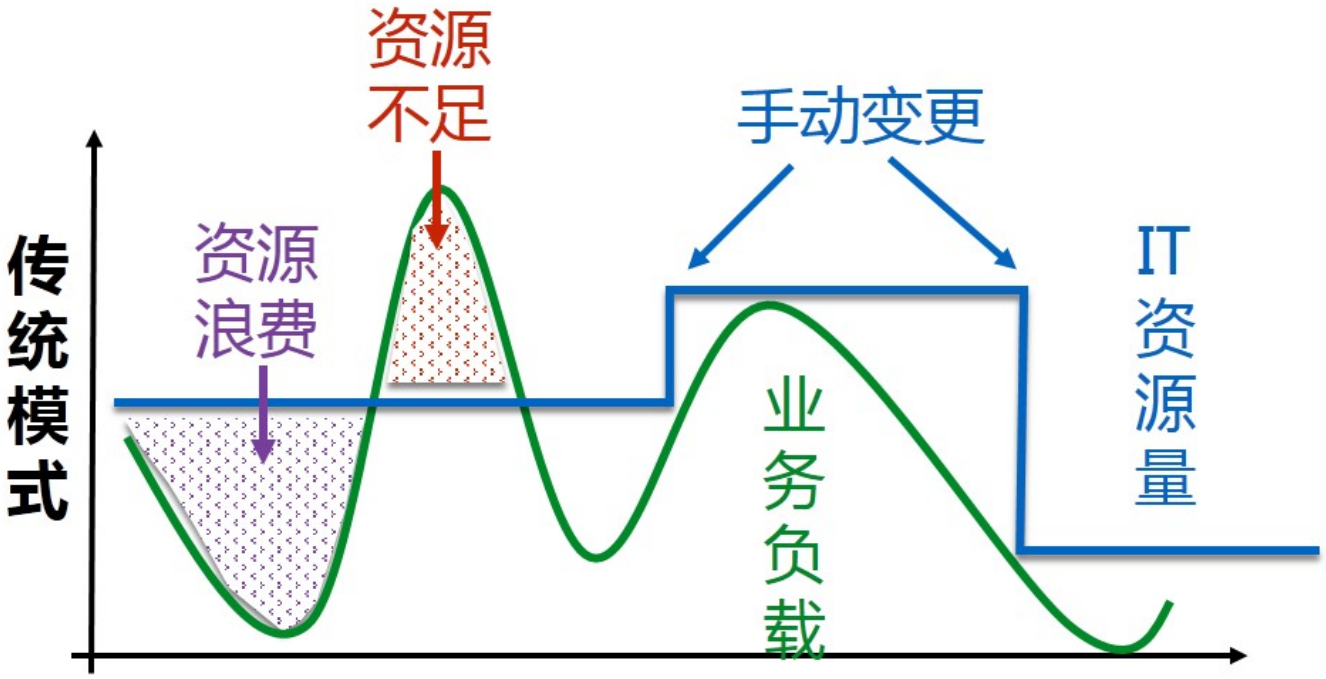
弹性伸缩 AS（Auto Scaling）可以根据您的业务需求和策略，自动调整 CVM 计算资源，确保您拥有适量的 CVM 实例来处理您的应用程序负载。对于您的 Web
服务而言，智能的扩展和收缩是成本控制和资源管理的重要组成部分。Web
应用程序开始获得更多请求流量时，您将添加更多的服务器来应对额外负载。同时，当 Web
应用程序的流量开始减少时，您将终止未充分利用的服务器。

如果使用 AS 进行容量调整，您只需事先设置好扩容条件及缩容条件。AS
会在达到条件时自动增加使用的服务器数量以维护性能；在需求下降时，AS
会根据您的缩容条件减少服务器数量，最大限度地帮助您降低成本。

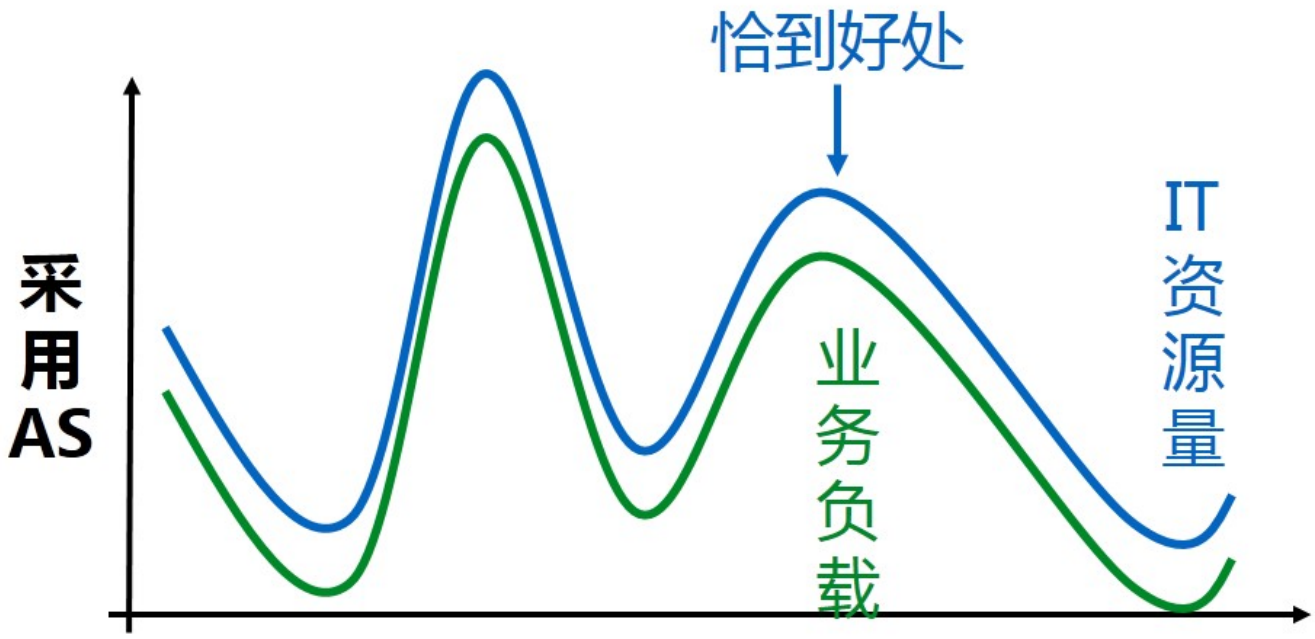
如下图对比所示，通过使用弹性伸缩
AS，您的集群可以永远保留恰到好处的资源量，并处于健康状态。您将告别传统模式下的多种烦恼：

- 业务突增或 CC 攻击导致机器数量不足，以致您的服务无响应
- 按高峰访问量预估资源，而平时访问量很少达到高峰，造成投入资源浪费
- 人工守护及频繁处理容量告警，需要多次手动变更

传统模式下的集群维护：



采用AS后的效果：



AS的工作方式

在常见的 Web 应用服务中，您的集群通常运行应用程序的多个副本来满足客户流量。比如接入层的前端服务器集群、逻辑层的应用服务器集群、后端的缓存服务器集群。每个实例都可以处理客户请求。

这些类似或相同的实例，数量通常是可调节的。您可以将这些相同或类似的机器归到一个伸缩组中管理起来：

- 您可以指定每个伸缩组中最少的实例数量，AS 会确保伸缩组中的实例永远不会低于这个数量；
- 您可以指定每个伸缩组中最大的实例数量，AS 会确保伸缩组中的实例永远不会高于这个数量；
- 您可以指定伸缩策略，则 AS 会在应用程序需求增加或降低时启动或终止实例。伸缩策略有两类：
  a) 告警触发策略：根据指定条件动态扩展（例如：伸缩组的机器的CPU 利用率超过60%时扩展）
  b) 定时伸缩策略：根据指定的时间扩展（例如：每晚21：00扩展）
- 设置完策略后，您还可以设置伸缩活动通知。AS 会在发生伸缩活动时通过邮件、短信、站内信方式告知您。您不需要时刻关注您的业务请求量变化，只需要留意 AS 的通知即可。
- 您也可以在任何时候一键指定所需要的机器数量，或者把已有的机器加入到伸缩组中一起管理。

AS的基本概念

弹性伸缩产品有以下基本概念：

- 伸缩组
- 启动配置
- 伸缩策略
- 冷却时间

1. 伸缩组

伸缩组是遵循相同规则、面向同一场景的云服务器实例的集合。伸缩组定义了组内 CVM 实例数的最大值、最小值及其相关联的负载均衡实例等属性。

2. 启动配置

启动配置是自动创建云服务器的模版，其中包括镜像ID、云服务器实例类型、系统盘及数据盘类型和容量、密钥对、安全组等。

创建伸缩组时必须指定启动配置，启动配置一经创建后其属性将不能编辑。

3. 伸缩策略

即执行伸缩动作的条件。触发条件可以是时间或云监控的报警，动作可以是移出或加入 CVM。
伸缩策略有以下两种：

- 定时伸缩策略
  到达某个固定时间点，自动增加或减少 CVM 实例，支持周期性重复。

- 告警伸缩
  基于云监控指标（如CPU、内存、网络流量），自动增加或减少 CVM 实例。

4. 冷却时间

冷却时间是指在同一个伸缩组内，一个伸缩活动（添加或移出 CVM
实例）执行完成后的一段锁定时间。在这段时间内，该伸缩组不执行伸缩活动。冷却时间可指定范围为
0-999999（秒）。

产品优势

优势	使用弹性伸缩 AS	不使用弹性伸缩 AS
自动化		

应用场景

1. 提前部署扩缩容

用户明确何时需要扩缩容，则可提前设置Auto Scaling定时策略。到相应时间时，系统将自动添加或减少CVM实例，无需人工等待。

2. 低成本应对业务浪涌

当客户面临访问峰值，需要提前准备服务器，预防CPU增长造成的服务器压力过大；待压力过去后再根据实际负载缩减服务器。客户可提前设置Auto Scaling监控策略，系统将根据设定好的业务监控指标自动判定是否需要CVM平行扩展。如果监控指标达到阈值，则实时自动增加或减少CVM实例，并自动完成负载均衡配置。既节约了成本，也无需客户时刻为手动扩容作准备。

3. 自动替换不健康CVM

为避免不健康的云服务器继续运行对业务造成影响，用户需要时刻关注系统中CVM的运行情况，并随时准备处理。使用Auto Scaling，系统将定时对CVM进行健康检查，若扫描出运行异常的CVM实例，则自动平行扩展一台实例替换异常实例。该操作记录将被保留供用户查看。

使用限制

- 目前地域有北京、上海、广州、香港、多伦多、新加坡。
- 用户在每个地域最多可创建20个启动配置。
- 一个用户最多可创建20个伸缩组。
- 一个伸缩组只能对应1个启动配置。
- 一个用户在一个地域下最多能弹性伸缩500台CVM实例。
- 一个伸缩组内最多能弹性伸缩200台CVM实例。
- 一个伸缩组内最多可创建100条伸缩策略，且最多可创建10个定时任务。
- 伸缩组的子机数量不能超过VPC子网能提供的IP上限。
- 弹性伸缩目前不支持纵向扩展，即无法自动升降CVM的CPU、内存和带宽。
- 弹性伸缩、启动配置均为地域概念，仅能在同一地域下启动/销毁CVM实例。