

弹性伸缩服务

产品介绍

文档版本 06
发布日期 2026-02-11



版权所有 © 华为云计算技术有限公司 2026。保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明



HUAWEI和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

目录

1 什么是弹性伸缩?	1
2 弹性伸缩的优势	3
3 生命周期	7
4 产品功能	11
5 使用限制	15
6 区域和可用区	17
7 计费标准	19
8 安全	20
8.1 身份认证与访问控制	20
8.2 数据保护技术	21
8.3 审计与日志	21
8.4 监控安全风险	22
8.5 认证证书	22
9 与其他服务的关系	24
10 权限管理	27
11 基本概念	31

1 什么是弹性伸缩?

弹性伸缩简介

弹性伸缩（Auto Scaling，以下简称AS）是根据用户的业务需求，通过设置伸缩规则来自动增加/缩减业务资源。当业务需求增长时，AS自动为您增加弹性云服务器（ECS）实例或带宽资源，以保证业务能力；当业务需求下降时，AS自动为您缩减弹性云服务器（ECS）实例或带宽资源，以节约成本。AS支持自动调整弹性云服务器和带宽资源。

图 1-1 弹性伸缩图例



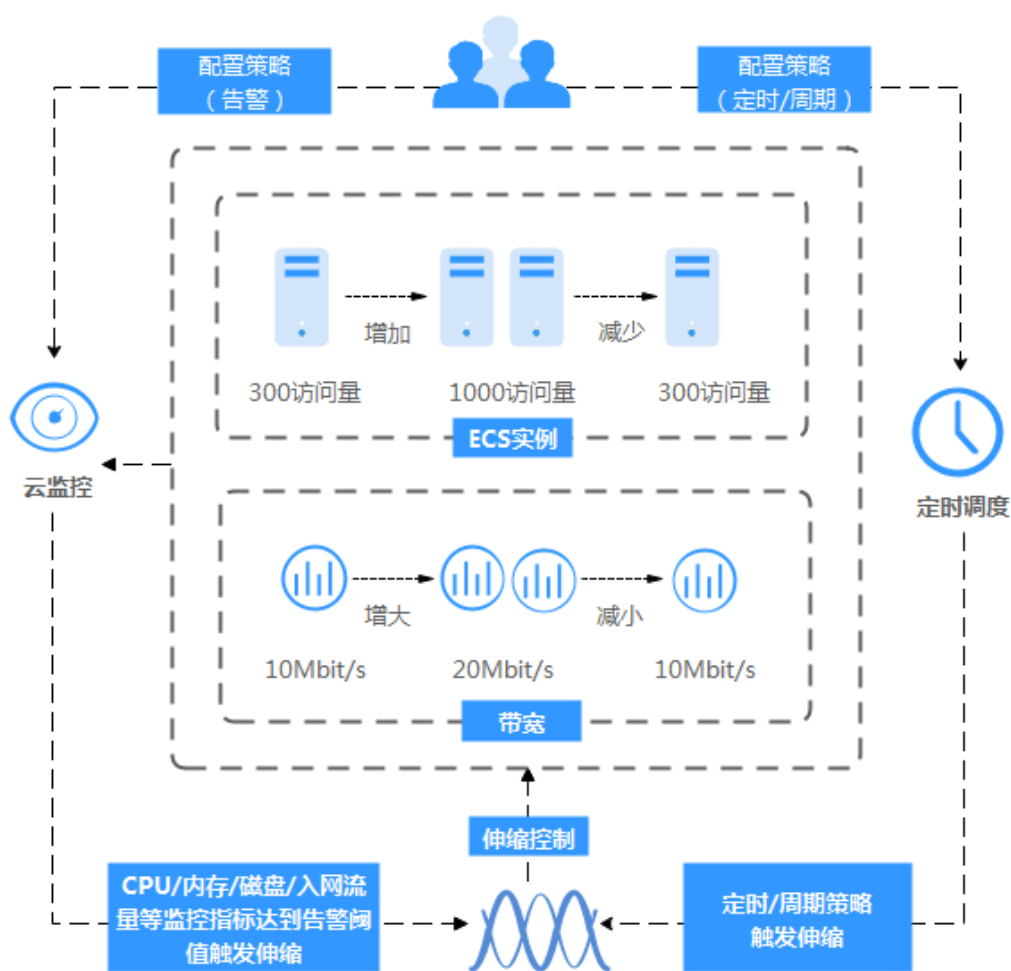
产品架构

通过伸缩控制可以实现弹性云服务器（ECS）实例伸缩和带宽伸缩：

- 伸缩控制：配置策略设置指标阈值/伸缩活动执行的时间，通过云监控监控指标是否达到阈值，通过定时调度，实现伸缩控制。
- 配置策略：可以根据业务需求，配置告警策略/定时策略/周期策略。
- 配置告警策略：可配置CPU、内存、磁盘、入网流量等监控指标。

- 配置定时策略：通过配置触发时间可以配置定时策略。
- 配置周期策略：通过配置重复周期、触发时间、生效时间可以配置周期策略。
- 云监控监控到所配置的告警策略中的某些指标达到告警阈值，从而触发伸缩活动，实现ECS实例的增加/减少或带宽的增大/减小。
- 到达所配置的触发时间时，触发伸缩活动，实现ECS实例的增加/减少或带宽的增大/减小。

图 1-2 弹性伸缩产品架构图



访问方式

公有云提供了Web化的服务管理平台，即管理控制台和基于HTTPS请求的API（Application programming interface）管理方式。

- API方式
如果用户需要将公有云平台上的弹性伸缩服务集成到第三方系统，用于二次开发，请使用API方式访问弹性伸缩服务，具体操作请参见《[弹性伸缩API参考](#)》。
- 控制台方式
其他相关操作，请使用管理控制台方式访问弹性伸缩服务。
如果用户已注册公有云，可直接登录[AS控制台](#)”。

2 弹性伸缩的优势

弹性伸缩服务可根据用户的业务需求，通过策略自动调整其业务的资源。具有自动调整资源、节约成本开支、提高可用性和容错能力的优势。适用以下场景：

- 访问流量较大的论坛网站，业务负载变化难以预测，需要根据实时监控到的云服务器CPU使用率、内存使用率等指标对云服务器数量进行动态调整。
- 电商网站，在进行大型促销活动时，需要定时增加云服务器数量和带宽大小，以保证促销活动顺利进行。
- 视频直播网站，每天14:00~16:00播出热门节目，每天都需要在该时段增加云服务器数量，增大带宽大小，保证业务的平稳运行。

自动调整资源

弹性伸缩能够实现应用系统自动按需调整资源，即在业务增长时能够实现自动增加实例数量和带宽大小，以满足业务需求，业务下降时能够实现应用系统自动扩容，保障业务平稳运行。

- 按需调整云服务器资源
向应用系统中添加弹性伸缩，能够实现按需调整资源，即能够在业务增长时增加实例，业务下降时减少实例，这样加强了应用系统的成本管理。调整资源主要包括以下几种方式：
 - 动态调整资源
动态调整资源是通过告警策略的触发来调整资源。
 - 计划调整资源
计划调整资源是通过定时策略或周期策略的触发来调整资源。
 - 手工调整资源
通过修改期望实例数或手动移入、移出实例来调整资源。

例如，运行在公有云上的基本Web应用程序。此应用程序允许乘客购买火车票。在每年中期时段，人员流动性较低，此应用程序的使用率较低。每年年底和年初，人员流动性较高，因此对此应用程序的需求会显著提高。一般系统会采用添加足够多的服务器，如[图2-1](#)所示，或添加处理应用程序平均需求所需的容量，如[图2-2](#)所示，来满足业务需求。但这两种方案会造成资源浪费或无法满足高峰期的需求。当您给应用程序中添加弹性伸缩后，弹性伸缩会自动根据需求调整服务器的数量，如[图2-3](#)所示，为您节约成本并且满足高峰期的需求。

图 2-1 服务器资源冗余

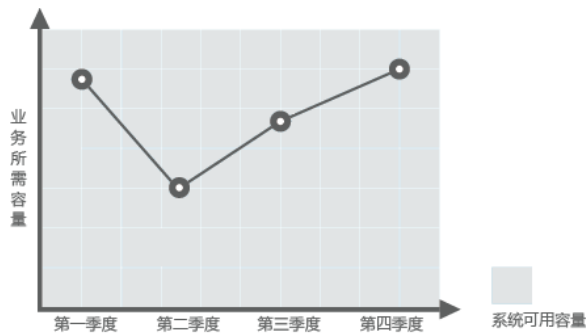


图 2-2 服务器资源不足

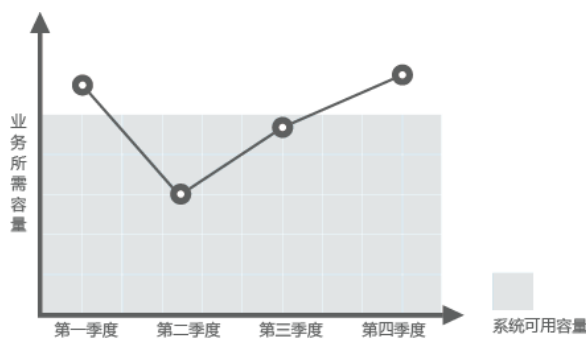
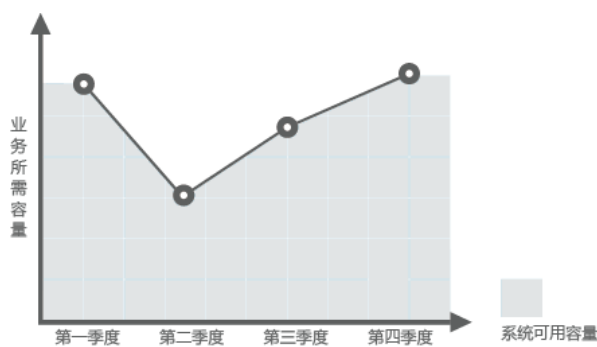


图 2-3 向应用程序中添加弹性伸缩



- 按需调整带宽资源

弹性伸缩能够实现按需调整带宽，即能够在业务增长时扩大带宽，业务下降时减小带宽，加强了应用系统的成本管理。

您可以根据实际情况选择如下伸缩带宽策略来实现按需调整IP带宽：

- 告警策略

可设置出网流量、出网带宽等告警触发条件，系统检测到触发条件满足时，会自动调整带宽的大小。

- 定时策略

系统可根据定时策略在固定的时间自动将带宽增大、减小或者调整到固定的值。

- 周期策略

系统可根据周期策略周期性的调整带宽大小，减少了人工重复设置带宽的工作量。

以告警策略的使用为例说明如下：

某视频直播网站，在不同时间段业务负载变化难以预测，需要根据出网流量、入网流量等指标在10Mbit/s到30Mbit/s之间动态调整带宽资源。弹性伸缩可以实现自动按需调整带宽，很好的解决这个问题。您只需选择需要调整的弹性公网IP，同时创建两个告警策略，一个策略设置在出网流量大于XXXbyte时，增加2Mbit/s，限制值为30Mbit/s；另一个策略在出网流量小于XXXbyte时，减少2Mbit/s，限制值为10Mbit/s。

● 按可用区均匀分配实例

按可用区均匀分配实例是指尽可能地将实例均匀分布在不同的可用区中，来降低电力、网络等可能出现的故障对整个系统稳定性的影响。

区域指弹性云服务器云主机所在的物理位置。每个区域包含许多不同的称为“可用区”的位置，即在同一区域下，电力、网络隔离的物理区域，可用区之间内网互通，不同可用区之间物理隔离。每个可用区都被设计成不受其他可用区故障影响的模式，并提供低价、低延迟的网络连接，以连接到同一地区其他可用区。

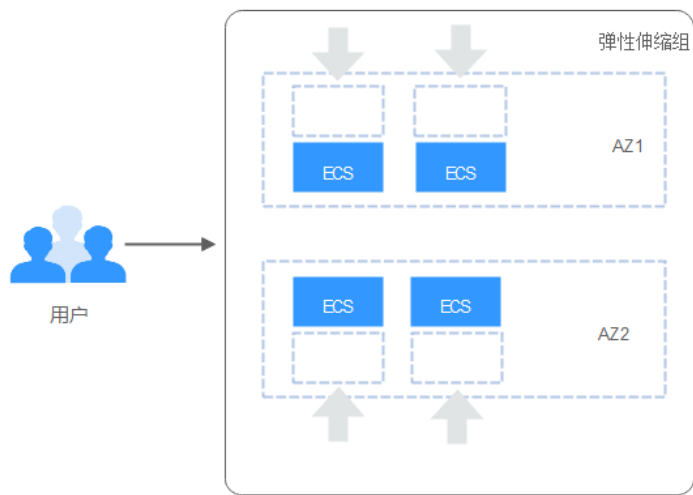
伸缩组可以包含来自同一区域的一个或多个可用区的实例。在资源调整时，弹性伸缩会通过实例分配和再均衡两种方法尽可能地将实例均匀分配到可用区中。

实例分配

弹性伸缩尝试在为伸缩组使用的可用区之间均匀分配实例。弹性伸缩通过尝试向实例最少的可用区中移入新实例来实现此目标。

例如，伸缩组目前有四个实例均匀分布在两个可用区内，若该伸缩组下一个伸缩活动增加四个实例时，会在两个可用区内分别增加两个实例，以实现可用区之间均匀分配实例。

图 2-4 均匀实例分配

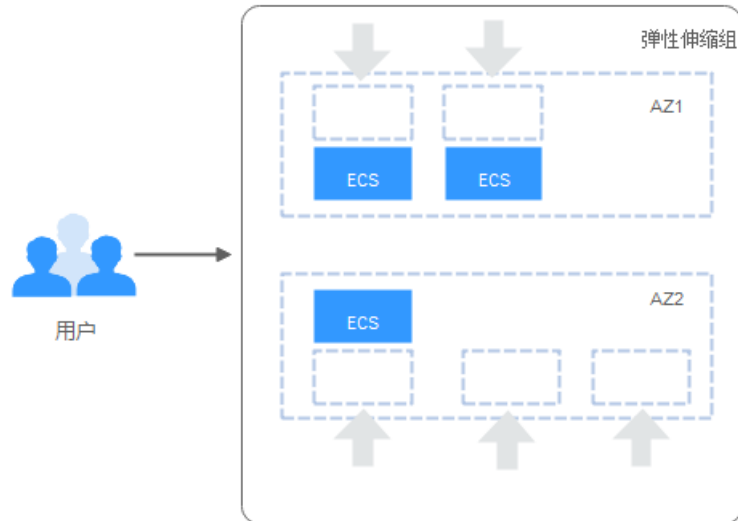


再均衡

手工加入或移出实例后，伸缩组中的实例没有均匀分配在可用区时，后续进行的伸缩活动会优先在可用区内均匀分配实例。

例如，伸缩组中目前有三个实例分布在两个可用区内，若该伸缩组下一个伸缩活动增加五个实例时，会在有两个实例的可用区内增加两个实例，在有一个实例的可用区增加三个实例，以实现可用区之间均匀分配实例。

图 2-5 再均衡



加强成本管理

弹性伸缩能够实现按需使用实例和带宽，并自动调整系统中的资源，节省了资源和人为调整资源带来的损耗，为您极大程度节约了成本。

提高可用性

弹性伸缩可确保应用系统始终拥有合适的容量以满足当前流量需求。

弹性伸缩和负载均衡结合使用

当您在使用弹性伸缩时，业务增长时应用系统自动扩容，业务下降时应用系统自动缩容，在伸缩组添加和删除实例时，须确保所有实例均可分配到应用程序的流量。弹性伸缩和负载均衡结合使用可以解决这个问题。

使用负载均衡后，伸缩组会自动地将加入伸缩组的实例绑定负载均衡监听器。访问流量将通过负载均衡监听器自动分发到伸缩组内的所有实例，提高了应用系统的可用性。若伸缩组中的实例上部署了多个业务，还可以添加多个负载均衡监听器到伸缩组，同时监听多个业务，从而提高业务的可扩展性。

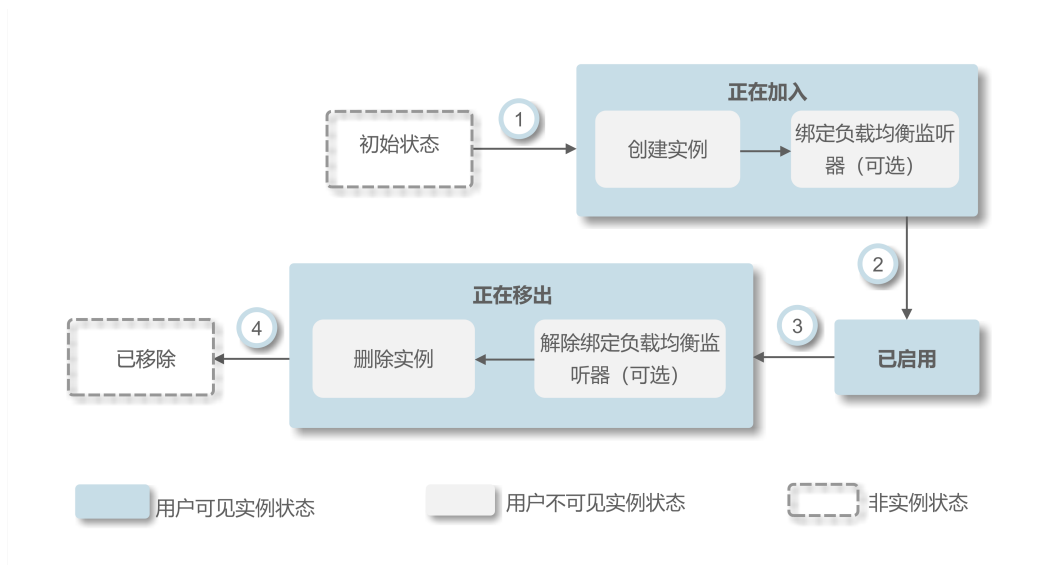
提高容错能力

弹性伸缩可以检测到应用系统中实例的运行状况，并启用新实例以替换运行状况不佳的实例。

3 生命周期

伸缩组中的实例生命周期，从创建实例开始，到该实例从伸缩组中移除结束。
伸缩组中未添加生命周期挂钩时，实例生命周期内状态之间的过渡如图3-1所示。

图 3-1 实例生命周期内状态之间的过渡



触发条件②和④表示系统自发触发实例状态的改变。

表 3-1 实例的状态

实例所处状态	子状态	实例状态含义	触发条件
初始状态	-	即实例还没状态。	触发条件①包括有两种情况，只要其中一种情况就能够触发实例进入“正在加入”状态。 <ul style="list-style-type: none"> • 手动修改期望实例数或满足伸缩策略的条件时，触发伸缩活动，进行扩容。 • 手动添加已有实例至伸缩组。
正在加入	创建实例	在触发条件①的作用下，伸缩组开始扩容，创建实例。	
	绑定负载均衡监听器（可选）	在触发条件①的作用下，创建实例完成后，实例绑定负载均衡监听器。	
已启用	-	实例进入伸缩组，开始接受处理业务流量。	触发条件③包括有三种情况，只要其中一种情况就能够触发实例从“已启用”状态到“正在移出”状态： <ul style="list-style-type: none"> • 手动修改期望实例数或满足伸缩策略的条件时，触发伸缩活动，进行扩容。 • 实例进入启用状态后，开始健康检查，健康检查失败后，移出实例。 • 手动将实例移出伸缩组。
正在移出	解除绑定负载均衡监听器（可选）	在触发条件③的作用下，伸缩组开始缩容，实例解除绑定负载均衡监听器。	
	删除实例	实例解除绑定负载均衡监听器后，从伸缩组中移出。	
已移除	-	实例在伸缩组中的生命周期已结束，即实例没有状态。	-

通过手动添加实例和伸缩活动向伸缩组添加实例，实例经过正在加入、已启用和正在移出状态后，实例将移出伸缩组。

伸缩组中已添加生命周期挂钩后，实例生命周期内状态之间的过渡如图3-2所示。当伸缩组进行伸缩活动时，实例将被生命周期挂钩挂起并置于等待状态，实例将保持此状态直至超时时间结束或者用户手动回调。用户能够在实例保持等待状态的时间段内执行自定义操作，例如，用户可以在新移入的实例上安装或配置软件，也可以在实例终止前从实例中下载日志文件。

图 3-2 实例生命周期内状态之间的过渡



触发条件②、④、⑥、⑧表示系统自发触发实例状态的改变。

表 3-2 实例状态

实例所处状态	子状态	实例状态含义	触发条件含义
初始状态	-	即实例还没状态。	触发条件①包括有两种情况，只要其中一种情况就能够触发实例进入“正在加入”状态。 <ul style="list-style-type: none"> • 手动修改期望实例数或满足伸缩策略的条件时，触发伸缩活动，进行扩容。 • 手动添加已有实例至伸缩组。
正在加入	创建实例	在触发条件①的作用下，伸缩组开始扩容，创建实例。	
加入挂起	-	正在加入伸缩组的实例被生命周期挂钩挂起，将实例置于等待的状态。	触发条件③包括有两种情况，只要其中一种情况就能够触发实例从“加入挂起”到“正在加入”状态。 <ul style="list-style-type: none"> • 默认回调操作 • 手动回调操作
正在加入	绑定负载均衡监听器（可选）	在触发条件③的作用下，实例将继续正在加入伸缩组，绑定负载均衡监听器。	

实例所处状态	子状态	实例状态含义	触发条件含义
已启用	-	实例进入伸缩组，开始接受处理业务流量。	触发条件⑤包括有三种情况，只要其中一种情况就能够触发实例从“已启用”状态到“正在移出”状态：
正在移出	解除绑定负载均衡监听器（可选）	在触发条件⑤的作用下，伸缩组开始缩容，实例解除绑定负载均衡监听器。	<ul style="list-style-type: none"> • 手动修改期望实例数或满足伸缩策略的条件时，触发伸缩活动，进行缩容。 • 实例进入启用状态后，开始健康检查，健康检查失败后，移出实例。 • 手动将实例移出伸缩组。
移出挂起	-	正在移出伸缩组的实例被生命周期挂钩挂起，将实例置于等待的状态。	触发条件⑦包括有两种情况，只要其中一种情况就能够触发实例从“移出挂起”到“正在移出”状态。
正在移出	删除实例	在触发条件⑦的作用下，实例将继续正在移出伸缩组，删除实例。	<ul style="list-style-type: none"> • 默认回调操作 • 手动回调操作
已移除	-	实例在伸缩组中的生命周期已结束，即实例没有状态。	-

通过手动添加实例和伸缩活动向伸缩组添加实例，实例经过正在加入、加入挂起、正在加入、已启用、正在移出、移出挂起和正在移出状态后，实例将移出伸缩组。

4 产品功能

以下为弹性伸缩服务AS提供的常用功能特性。在使用弹性伸缩服务之前，建议您先了解弹性伸缩服务AS的[基本概念](#)，以便更好地理解弹性伸缩服务AS提供的各项功能。

记录弹性伸缩

弹性伸缩支持使用云审计记录服务资源操作。云审计记录的操作类型有三种，通过云平台帐户登录管理控制台执行的操作，通过云服务支持的API执行的操作，以及系统内部触发的操作。如果用户开通了云审计，AS服务的API被调用时，调用信息将会上报到云审计，云审计会将操作信息定时的转储到用户指定的对象存储桶。通过云审计服务，您可以记录与弹性伸缩相关的操作事件，便于日后的查询、审计和回溯。

更多信息，请参考[记录弹性伸缩](#)

伸缩组

伸缩组是具有相同属性和应用场景的云服务器和伸缩策略的集合，是启停伸缩策略和进行伸缩活动的基本单位。您可以使用伸缩策略设定的条件自动增加、减少伸缩组中的实例数量，或维持伸缩组中固定的实例数量。创建伸缩组，需要配置最大实例数、最小实例数、期望实例数和负载均衡器等参数。

更多信息，请参考[伸缩组](#)

伸缩配置

伸缩配置用于定义伸缩组资源扩展时的云服务器的规格。包括云服务器的操作系统镜像、系统盘大小等。

更多信息，请参考[伸缩配置](#)

伸缩策略

伸缩策略可以触发伸缩活动，是对伸缩组中实例数量或带宽进行调整的一种方式。伸缩策略规定了伸缩活动触发需要满足的条件及需要执行的操作，当满足伸缩条件时，系统会自动触发一次伸缩活动。

目前系统支持以下三种策略。

- **告警策略：**基于云监控系统告警数据（例如CPU使用率），自动增加、减少或设置指定数量的云服务器。

- 定时策略：基于配置的某个时间点，自动增加、减少或设置指定数量的云服务器。
- 周期策略：按照配置周期（按天、按周、按月），周期性地增加、减少或设置指定数量的云服务器。

更多信息，请参考[伸缩策略](#)

动态扩展资源

弹性伸缩进行伸缩活动时，需定义如何按照不断变化的需求进行伸缩活动，即动态扩展资源。当业务需求变化频繁且无固定规律时，可通过配置告警策略实现动态扩缩资源的目的。当满足伸缩策略的条件时，系统自动修改期望实例数，从而触发伸缩活动进行资源的扩张或收缩。

更多信息，请参考[使用弹性伸缩动态扩展云服务器资源](#)

按计划扩展资源

弹性伸缩进行伸缩活动时，应对需求有规律变化的场景，需按照规律定期或者周期性地伸缩活动，可通过配置定时策略和周期策略有计划的来调整资源。

更多信息，请参考[使用弹性伸缩定时扩展云服务器资源](#)

手动扩展资源

通过手动将实例移入到伸缩组、手动将实例移出伸缩组或手动修改期望实例数，扩展资源。

更多信息，请参考[手动移入或移除资源](#)

实例移除策略

当您的伸缩组自动移除实例时，如果伸缩组内存在不属于当前配置的可用区的实例，移除实例时，会优先移除这些实例。其次，再按照您配置的实例移除策略移除实例。

更多信息，请参考[创建伸缩组](#)

生命周期挂钩

生命周期挂钩功能提供灵活控制伸缩组内ECS实例创建和移出过程的能力，以便用户灵活管理ECS实例的生命周期。

添加生命周期挂钩后，实例生命周期状态如图所示。



更多信息，请参考[生命周期挂钩](#)

实例保护

如果您希望伸缩组中特定的实例不被自动移出伸缩组，请使用实例保护。您可以对伸缩组中一个或多个正常状态的实例启用实例保护设置，当伸缩组发生缩容活动时，设置了实例保护的实例将不会被移出伸缩组。

更多信息，请参考[实例保护](#)

实例备用

若您需要伸缩组中的部分实例暂时停止承担业务流量且不被移出伸缩组，您可以使用弹性伸缩提供的实例备用功能。您可以对伸缩组中的一个或多个实例设置实例备用，实例备用能在保证实例不被移出伸缩组的同时对实例进行关机、重启等操作。

更多信息，请参考[实例备用](#)

伸缩带宽策略

用户可以通过伸缩带宽策略对购买的弹性公网IP带宽和共享带宽进行自动调整。创建伸缩带宽策略时，需要配置对应的基本信息，包括配置策略名称、资源类型、策略类型、触发条件等。系统支持告警策略、定时策略、周期策略三种伸缩带宽策略。

更多信息，请参考[伸缩带宽策略](#)

为伸缩组配置通知

用户可通过消息通知服务提供的功能，将伸缩组的扩容成功、扩容失败、减容成功、减容失败和异常等情况及时推送给用户，以使用户能够及时了解伸缩组的各种状态。每个伸缩组最多可以配置5个通知。给弹性伸缩组配置通知，需配置一个通知事件和通知主题，通知主题由用户先在消息通知服务界面创建，当通知主题对应的通知场景出现时，伸缩组会向用户发送通知。

更多信息，请参考[为伸缩组配置通知](#)

查看监控指标数据

为使用户更好地掌握自己的弹性云服务器运行状态，云平台提供了云监控，您可以了解如何查看伸缩组的监控指标详情，更好地了解弹性云服务器的各项性能指标。

更多信息，请参考[查看监控指标数据](#)、[监控指标说明](#)

设置监控告警规则

通过设置弹性云服务器告警规则，用户可自定义监控目标与通知策略，及时了解弹性云服务器运行状况，从而起到预警作用。

更多信息，请参考[设置监控告警规则](#)

权限管理

如果您需要对华为云上购买的弹性伸缩（Auto Scaling，简称AS）资源，为企业中的员工设置不同的访问权限，以达到不同员工之间的权限隔离，您可以使用统一身份认证服务（Identity and Access Management，简称IAM）进行精细的权限管理。该服务提供用户身份认证、权限分配、访问控制等功能，可以帮助您安全的控制华为云资源的访问。

更多信息，请参考[权限管理](#)

API

通过使用弹性伸缩所提供的接口，您可以完整的使用弹性伸缩的所有功能。

更多信息，请参考[API](#)

SDK

借助弹性伸缩AS的SDK开发包，您可以很容易的调用AS的API接口，创建基于华为云的互联网应用。目前SDK的语言支持：Java、Python。您可以使用API或其他任意一种熟知的SDK。

更多信息，请参考[SDK](#)

5 使用限制

功能限制

在应用系统中添加弹性伸缩后，使用时有以下功能限制：

- 弹性伸缩的云服务器中运行的应用需要是无状态、可横向扩展的。

📖 说明

- 无状态：**关于应用的既往事务，没有任何记录和参考，每项事务处理均是从头开始。无状态应用运行的实例不会在本地存储需要持久化的数据。
例如：可以将无状态事务看作一台自动售货机：一个请求对应一个响应。
- 有状态：**是可以周而复始、反复发生的应用和流程，操作是在之前的事务背景下执行的，当前事务可能会受到之前事务的影响。
有状态应用运行的实例会在本地存储需要持久化的数据。
例如：可以将有状态事务看作网上银行或电子邮件，有上下文记录。
- 弹性伸缩会自动释放云服务器，所以弹性伸缩组内的云服务器不可以保存应用的状态信息（例如session）和相关数据（如数据库、日志等）。如果应用中需要云服务器保存状态或日志信息，可以考虑把相关信息保存到独立的服务器中。
- 弹性伸缩无法纵向扩展，即弹性伸缩无法自动升降ECS实例的vCPU和内存等配置。

配额限制

弹性伸缩对用户的资源数量或容量做的配额限制如[表5-1](#)所示。

表 5-1 配额一览表

类别	描述	默认值
弹性伸缩组	用户可以创建的最多伸缩组个数。	10
弹性伸缩配置	用户可以创建的最多伸缩配置个数。	100
弹性伸缩策略	某个弹性伸缩组下可以创建的最多伸缩策略个数。	10

类别	描述	默认值
弹性伸缩实例	某个弹性伸缩组下可以创建的最多实例个数。	300
伸缩带宽策略	用户最多可以创建的伸缩带宽策略个数。	10
生命周期挂钩	某个弹性伸缩组内最多可添加的生命周期挂钩个数。	5
通知	某个弹性伸缩组最多可以配置的通知个数。	5
标签	某个弹性伸缩组最多可以添加的标签个数。	10

6 区域和可用区

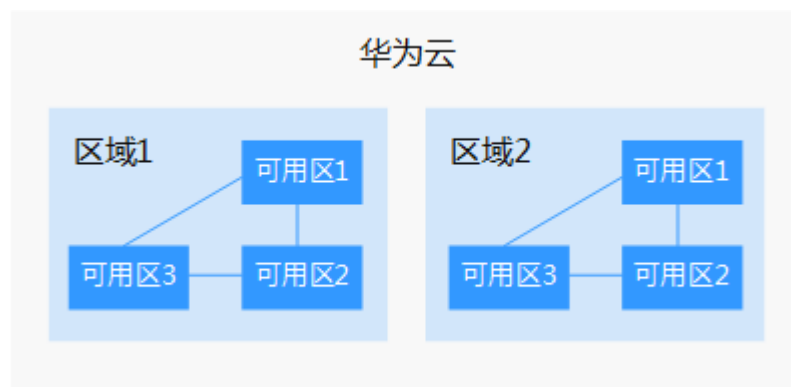
什么是区域、可用区？

区域和可用区用来描述数据中心的位置，您可以在特定的区域、可用区创建资源。

- 区域（Region）：从地理位置和网络时延维度划分，同一个Region内共享弹性计算、块存储、对象存储、VPC网络、弹性公网IP、镜像等公共服务。Region分为通用Region和专属Region，通用Region指面向公共租户提供通用云服务的Region；专属Region指只承载同一类业务或只面向特定租户提供业务服务的专用Region。
- 可用区（AZ，Availability Zone）：一个AZ是一个或多个物理数据中心的集合，有独立的风火水电，AZ内逻辑上再将计算、网络、存储等资源划分成多个集群。一个Region中的多个AZ间通过高速光纤相连，以满足用户跨AZ构建高可用性系统的需求。

图6-1阐明了区域和可用区之间的关系。

图 6-1 区域和可用区



目前，华为云已在全球多个地域开放云服务，您可以根据需求选择适合自己的区域和可用区。更多信息请参见[华为云全球站点](#)。

如何选择区域？

选择区域时，您需要考虑以下几个因素：

- 地理位置

一般情况下，建议就近选择靠近您或者您的目标用户的区域，这样可以减少网络时延，提高访问速度。

- 在除中国大陆以外的亚太地区有业务的用户，可以选择“中国-香港”、“亚太-曼谷”或“亚太-新加坡”区域。
- 在非洲地区有业务的用户，可以选择“非洲-约翰内斯堡”区域。
- 在拉丁美洲地区有业务的用户，可以选择“拉美-圣地亚哥”区域。

 说明

“拉美-圣地亚哥”区域位于智利。

- 资源的价格

不同区域的资源价格可能有差异，请参见华为云服务价格详情。

如何选择可用区？

是否将资源放在同一可用区内，主要取决于您对容灾能力和网络时延的要求。

- 如果您的应用需要较高的容灾能力，建议您将资源部署在同一区域的不同可用区内。
- 如果您的应用要求实例之间的网络延时较低，则建议您将资源创建在同一可用区内。

区域和终端节点

当您通过API使用资源时，您必须指定其区域终端节点。有关华为云的区域和终端节点的更多信息，请参阅[地区和终端节点](#)。

7 计费标准

弹性伸缩服务本身不收取费用，但伸缩组自动创建的按需付费实例需要支付相应的费用，实例的计费标准请参见[计费说明](#)。实例使用的弹性公网IP也需支付相应的费用，弹性公网IP的计费标准请参见[计费说明](#)。

伸缩组进行减容时，自动创建的实例会被移出伸缩组并删除，删除后将不再收取费用。而之前通过手动移入的实例只会被移出伸缩组，系统仍会收取该实例的使用费用。若您不再需要使用该实例，请自行在ECS页面进行退订。

例如，弹性伸缩进行扩容活动创建了两台实例，使用一个小时后，进行了缩容活动，这两台实例被移出伸缩组并删除了，则系统只收取这两台实例使用一小时产生的费用。

8 安全

8.1 身份认证与访问控制

身份认证

统一身份认证服务（Identity and Access Management，简称IAM）提供用户身份认证、权限分配、访问控制等功能，可以帮助您安全的控制华为云资源的访问。

通过IAM，您可以在账号中给员工创建IAM用户，并授权控制他们对华为云资源的访问范围。例如您的员工中有负责软件开发的人员，您希望他们拥有AS的使用权限，但是不希望他们拥有删除伸缩组等高危操作的权限，那么您可以使用IAM为开发人员创建用户，通过授予仅能使用伸缩组，但是不允许删除伸缩组的权限策略，控制他们对AS资源的使用范围。

访问控制

AS支持通过权限控制（IAM权限）、项目和企业项目、敏感操作、安全组进行访问控制。

表 8-1 AS 访问控制

访问控制方式	简要说明	详细介绍
权限控制（IAM权限）	默认情况下，管理员创建的IAM用户没有任何权限，需要将其加入用户组，并给用户组授予策略或角色，才能使得用户组中的用户获得对应的权限，这一过程称为授权。授权后，用户就可以基于被授予的权限对云服务进行操作。	权限管理

访问控制方式	简要说明	详细介绍
项目和企业项目	项目和企业项目都可以授权给一个或者多个用户组进行管理，管理企业项目的用户归属于用户组。通过给用户组授予策略，用户组中的用户就能在所属项目/企业项目中获得策略中定义的权限。	管理项目和企业项目
敏感操作	当您开启操作保护后，进行删除伸缩组操作时，需要进行身份认证。	敏感操作
安全组	安全组是一个逻辑上的分组，为具有相同安全保护需求并相互信任的云服务器提供访问策略。安全组创建后，用户可以在安全组中定义各种访问规则，当云服务器加入该安全组后，即受到这些访问规则的保护。 系统会为每个用户默认创建一个默认安全组，默认安全组的规则是在出方向上的数据报文全部放行，入方向访问受限，安全组内的云服务器无需添加规则即可互相访问。	配置安全组规则

8.2 数据保护技术

用户加密，是指用户通过提供的加密特性，对弹性云服务器资源进行加密，从而提升数据的安全性。

如果使用加密的弹性云服务器创建弹性伸缩配置，那么创建出来的伸缩配置，加密方式与原云服务器保持一致。

常见的加密方式有镜像加密，云硬盘加密，详情请参见[用户加密](#)。

8.3 审计与日志

云审计服务（Cloud Trace Service, CTS），是华为云安全解决方案中专业的日志审计服务，提供对各种云资源操作记录的收集、存储和查询功能，可用于支撑安全分析、合规审计、资源跟踪和问题定位等常见应用场景。

在您的应用系统中启用云审计服务后，将在日志文件记录对弹性伸缩执行的API调用的操作，并为您保存最近7天的操作记录。如果需要保存7天之前的操作记录，您可以通过对象存储服务（Object Storage Service，以下简称OBS），将操作记录实时同步保存至OBS。

- CTS的详细介绍和开通配置方法请参见[CTS快速入门](#)。
- 云审计服务支持的AS操作列表请参见[记录弹性伸缩](#)。

- 查看云审计日志的方法请参见[查看审计日志](#)。

8.4 监控安全风险

监控指标

弹性伸缩支持云监控的监控指标，用户可以通过云监控检索弹性伸缩服务产生的监控指标和告警信息。

查看弹性伸缩支持的监控指标请参见[监控指标说明](#)。

健康度检查

为使用户更好地掌握自己的弹性云服务器运行状态，云平台提供了云监控。您可以查看伸缩组的监控指标详情，更好地了解弹性云服务器的各项性能指标。详情请参考[查看监控指标数据](#)。

8.5 认证证书

合规证书

华为云服务及平台通过了多项国内外权威机构（ISO/SOC/PCI等）的安全合规认证，用户可自行[申请下载](#)合规资质证书。

图 8-1 合规证书下载

The screenshot displays a webpage titled "合规证书下载" (Compliance Certificate Download). At the top, there is a search bar with the placeholder text "请输入关键字搜索". Below the search bar, there are six cards, each representing a different certification or report. Each card includes a logo, the name of the certification, a brief description, and a "下载" (Download) button.

认证/报告名称	描述
BS 10012:2017	BS 10012为个人信息管理体系提供了一个符合欧盟GDPR原则的最佳实践框架。它概述了组织在收集、存储、处理、保留或处理与个人相关的个人记录时需要考虑的核心需求。保留或处理与个人相关的个人记录时需要考虑的核心需求。
CSA STAR认证	CSA STAR认证是由标准研发机构BSI（英国标准协会）和CSA（云安全联盟）合作推出的国际范围内的针对云安全水平的权威认证，旨在应对与云安全相关的特定问题，协助云计算服务商展现其服务成熟的解决方案。
ISO 20000-1:2018	ISO 20000是针对信息技术服务管理领域的国际标准，提供设计、建立、实施、运行、监控、评审、维护和改进服务管理体系的模型以保证服务提供商可提供有效的IT服务来满足客户和业务的需求。
SOC 1 类型II 报告 2022.04.01-2023.03.31	华为云每年滚动发布两期SOC1报告，均涵盖1年的时期（每年的4月1日至次年3月31日，以及每年10月1日至次年9月30日），报告分别在6月初和12月初发布。本期报告涵盖期间为2022.04.01-2023.03.31。SOC审计报告是由第三方审计机构根据美国注册会计师协会（AICPA）制定的相关准则，针对外包服务商的系统 and 内部控制情况出具的独立审计报告。SOC 1报告着重于评估与财务报告流程有关的控制，通常使用者为云客户和具独立审计师。
SOC 1 类型II 报告 2022.10.01-2023.09.30	华为云每年滚动发布两期SOC1报告，均涵盖1年的时期（每年的4月1日至次年3月31日，以及每年10月1日至次年9月30日），报告分别在6月初和12月初发布。本期报告涵盖期间为2022.10.01-2023.09.30。SOC审计报告是由第三方审计机构根据美国注册会计师协会（AICPA）制定的相关准则，针对外包服务商的系统 and 内部控制情况出具的独立审计报告。SOC 1报告着重于评估与财务报告流程有关的控制，通常使用者为云客户和具独立审计师。
SOC 2 类型II 报告 2022.04.01-2023.03.31	华为云每年滚动发布两期SOC2报告，均涵盖1年的时期（每年的4月1日至次年3月31日，以及每年10月1日至次年9月30日），报告分别在6月初和12月初发布。本期报告涵盖期间为2022.04.01-2023.03.31。SOC审计报告是由第三方审计机构根据美国注册会计师协会（AICPA）制定的相关准则，针对外包服务商的系统 and 内部控制情况出具的独立审计报告。SOC 2报告着重于组织的内部运作与合规，包括安全性、可用性、进程完整性、保密性、隐私性五大控制属性。

资源中心

华为云还提供以下资源来帮助用户满足合规性要求，具体请查看[资源中心](#)。

图 8-2 资源中心



合规资质证书

华为云安全服务提供了网络安全专用产品安全检测证书、软件著作权等证书，供用户下载和参考。具体请查看[合规资质证书](#)。

图 8-3 网络安全专用产品安全检测证书&软件著作权证书



9 与其他服务的关系

除直接使用弹性伸缩提供的对资源进行调整的功能外，若您同时购买了云服务中的其他产品，可以结合其他产品一起使用，能满足您多种场景下对云产品的需求。

弹性伸缩服务与周边服务的依赖关系如图9-1所示。

图 9-1 弹性伸缩服务与其他服务的关系示意图



表 9-1 弹性伸缩与其他服务的关系

服务名称	说明	交互功能	相关内容
弹性负载均衡 (Elastic Load Balance)	当配置了负载均衡服务后，弹性伸缩组在添加或移除云服务器时，自动会为云服务器绑定或解绑负载均衡监听器。 AS支持ELB的前提是：弹性伸缩组和负载均衡器必须处于同一VPC内。	使伸缩组中每一个实例均可分配到应用程序流量	创建伸缩组
云监控服务 (Cloud Eye)	弹性伸缩配置了告警触发策略时，会根据云监控的告警条件触发弹性伸缩活动。	通过监控伸缩组内实例的状态指标调节资源。	弹性伸缩支持的监控指标
弹性云服务器 (Elastic Cloud Server)	弹性伸缩活动中添加的云服务器可以通过弹性云服务器进行管理和维护。	自动调整弹性云服务器数量	动态扩展资源
虚拟私有云 (Virtual Private Cloud)	弹性伸缩支持自动调整虚拟私有云中创建的弹性公网IP带宽或共享带宽大小。	自动调整带宽大小	创建伸缩带宽策略
消息通知服务 (Simple Message Notification)	用户使用消息通知功能后，系统会将伸缩组的多种情况及时推送给用户，便于用户及时了解伸缩组的状态。	消息通知	为伸缩组配置通知
云审计服务 (Cloud Trace Service)	开通云审计服务后，可以记录弹性伸缩相关的操作事件，便于日后的查询、审计和回溯。	日志审计	记录弹性伸缩

服务名称	说明	交互功能	相关内容
标签管理服务 (Tag Management Service)	当您具有许多相同类型的弹性伸缩资源时，标签可以为您提供灵活的资源管理能力。	标签	标记伸缩组和实例

10 权限管理

如果您需要对华为云上购买的弹性伸缩（Auto Scaling，简称AS）资源，为企业中的员工设置不同的访问权限，以达到不同员工之间的权限隔离，您可以使用统一身份认证服务（Identity and Access Management，简称IAM）进行精细的权限管理。该服务提供用户身份认证、权限分配、访问控制等功能，可以帮助您安全的控制华为云资源的访问。如果华为账号已经能满足您的要求，不需要通过IAM对用户进行权限管理，您可以跳过本章节，不影响您使用AS服务的其它功能。

IAM是华为云提供权限管理的基础服务，无需付费即可使用，您只需要为您账号中的资源进行付费。

通过IAM，您可以通过授权控制他们对华为云资源的访问范围。例如您的员工中有负责软件开发的人员，您希望他们拥有弹性伸缩（Auto Scaling，简称AS）的使用权限，但是不希望他们拥有删除AS等高危操作的权限，那么您可以使用IAM进行权限分配，通过授予用户仅能使用AS，但是不允许删除AS的权限，控制他们对AS资源的使用范围。

目前IAM支持两类授权，一类是角色与策略授权，另一类为身份策略授权。

两者有如下的区别和关系：

表 10-1 两类授权的区别

名称	核心关系	涉及的权限	授权方式	适用场景
角色与策略授权	用户-权限-授权范围	<ul style="list-style-type: none">系统角色系统策略自定义策略	为主体授予角色或策略	核心关系为“用户-权限-授权范围”，每个用户根据所需权限和所需授权范围进行授权，无法直接给用户授权，需要维护更多的用户组，且支持的条件键较少，难以满足细粒度精确权限控制需求，更适用于对细粒度权限管控要求较低的中小企业用户。

名称	核心关系	涉及的权限	授权方式	适用场景
身份策略授权	用户-策略	<ul style="list-style-type: none"> 系统策略 自定义身份策略 	<ul style="list-style-type: none"> 为主体授予身份策略 身份策略附加至主体 	核心关系为“用户-策略”，管理员可根据业务需求定制不同的访问控制策略，能够做到更细粒度更灵活的权限控制，新增资源时，对比角色与策略授权，基于身份策略的授权模型可以更快地直接给用户授权，灵活性更强，更方便，但相对应的，整体权限管控模型构建更加复杂，对相关人员专业能力要求更高，因此更适用于中大型企业。

例如：如果需要对IAM用户授予可以创建华北-北京四区域的ECS和华南-广州区域的OBS的权限，基于角色与策略授权的场景中，管理员需要创建两个自定义策略，并且为IAM用户同时授予这两个自定义策略才可以实现权限控制。在基于身份策略授权的场景中，管理员仅需要创建一个自定义身份策略，在身份策略中通过条件键“g:RequestedRegion”的配置即可达到身份策略对于授权区域的控制。将身份策略附加主体或为主体授予该身份策略即可获得相应权限，权限配置方式更细粒度更灵活。

两种授权场景下的策略/身份策略、授权项等并不互通，推荐使用身份策略进行授权。[角色与策略权限管理](#)和[身份策略权限管理](#)分别介绍两种模型的系统权限。

关于IAM的详细介绍，请参见[IAM产品介绍](#)。

角色与策略权限管理

AS服务支持角色与策略授权。默认情况下，管理员创建的IAM用户没有任何权限，需要将其加入用户组，并给用户组授予策略或角色，才能使得用户组中的用户获得对应的权限，这一过程称为授权。授权后，用户就可以基于被授予的权限对云服务进行操作。

AS部署时通过物理区域划分，为项目级服务。授权时，“授权范围”需要选择“指定区域项目资源”，然后在指定区域（如华北-北京1）对应的项目（cn-north-1）中设置相关权限，并且该权限仅对此项目生效；如果“授权范围”选择“所有资源”，则该权限在所有区域项目中都生效。访问AS时，需要先切换至授权区域。

如表10-2所示，包括了AS的所有系统权限。角色与策略授权场景的系统策略与身份策略授权场景的并不互通。

表 10-2 AS 系统权限

系统角色/策略名称	描述	类别	依赖关系
AutoScalingFullAccess	对弹性伸缩全部资源的所有执行权限。	系统策略	无
AutoScalingReadOnlyAccess	弹性伸缩只读权限，对弹性伸缩全部资源的只读权限。	系统策略	无

系统角色/策略名称	描述	类别	依赖关系
AutoScaling Administrator	对弹性伸缩全部资源的所有执行权限。	系统角色	依赖ELB Administrator、CES Administrator、Server Administrator和Tenant Administrator角色，在同项目中勾选依赖的角色。

表10-3列出了AS常用操作与系统权限的授权关系，您可以参照该表选择合适的系统权限。

表 10-3 常用操作与系统权限的关系

操作	AutoScalingFullAccess	AutoScalingReadOnlyAccess	AutoScaling Administrator
创建伸缩组	√	x	√
修改伸缩组	√	x	√
查询伸缩组详情	√	√	√
删除伸缩组	√	x	√
创建伸缩配置	√	x	√
创建伸缩策略	√	x	√
创建伸缩带宽策略	√	x	√

身份策略权限管理

AS服务支持身份策略授权。如表10-4所示，包括了身份策略中的所有系统身份策略。身份策略授权场景的系统身份策略与角色与策略授权场景的并不互通。

表 10-4 AS 系统身份策略

系统身份策略名称	描述	策略类别
ASFullPolicy	弹性伸缩管理员权限，拥有该权限的用户可以操作并使用弹性伸缩的全部资源。	系统身份策略

系统身份策略名称	描述	策略类别
ASReadOnlyPolicy	弹性伸缩只读权限，拥有该权限的用户仅能查看弹性伸缩资源。	系统身份策略

表10-5列出了AS常用操作与系统身份策略的授权关系，您可以参照该表选择合适的系统身份策略。

表 10-5 常用操作与系统身份策略的关系

操作	ASFullPolicy	ASReadOnlyPolicy
创建伸缩组	√	x
修改伸缩组	√	x
查询伸缩组详情	√	√
删除伸缩组	√	x
创建伸缩配置	√	x
创建伸缩策略	√	x
创建伸缩带宽策略	√	x

相关链接

- [IAM产品介绍](#)

11 基本概念

伸缩组

伸缩组是具有相同应用场景的实例的集合，是启停伸缩策略和进行伸缩活动的基本单位。

伸缩配置

伸缩配置是伸缩组内实例（弹性云服务器）的模板，定义了伸缩组内待添加的实例的规格数据。包括云服务器类型、vCPU、内存、镜像、磁盘、登录方式等。

伸缩策略

伸缩策略可以触发伸缩活动，是对伸缩组中实例数量进行调整的一种方式。伸缩策略规定了伸缩活动触发需要满足的条件及需要执行的操作，当满足伸缩条件时，系统会自动触发一次伸缩活动。

伸缩活动

伸缩组中增加或减少实例的过程称为伸缩活动。伸缩活动的目的是使应用系统中当前实例数和期望实例数保持一致，或达到已设置的伸缩策略触发条件时，执行增加或减少实例数量的操作，保证业务正常运行。

冷却时间

为了避免告警策略频繁触发，必须设置冷却时间。冷却时间是指冷却期内限制伸缩活动触发的时间，单位为秒。在每次伸缩活动完成之后，系统开始计算冷却时间。伸缩组在冷却时间内，会拒绝告警策略的触发，其他类型的伸缩策略（如定时策略和周期策略）及手动触发不受限制。

例如：冷却时间设置为300秒，定时策略设置了10:32进行伸缩活动，10:30告警触发的伸缩活动结束，则在10:30-10:35时间内，伸缩组会拒绝新告警触发的伸缩活动，但不会拒绝在10:32时定时策略触发的伸缩活动；若10:36定时策略触发的伸缩活动结束，则冷却时间为10:36-10:41。

伸缩带宽

伸缩带宽可以根据用户配置的伸缩带宽策略自动调整带宽资源。弹性伸缩仅支持对按需购买的弹性公网IP带宽和共享带宽进行调整，不支持对包年包月的带宽进行调整。