ALIBABA CLOUD

阿里云

弹性伸缩 ESS

产品简介

文档版本：20201231

阿里云

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。

2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。

3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。

4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。

5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。

6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
危险	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	危险 重置操作将丢失用户配置数据。
警告	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	警告 重启操作将导致业务中断，恢复业务时间约十分钟。
注意	用于警示信息、补充说明等，是用户必须了解的内容。	注意 权重设置为0，该服务器不会再接受新请求。
说明	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	单击**设置**> **网络**> **设置网络类型**。
粗体	表示按键、菜单、页面名称等UI元素。	在**结果确认**页面，单击**确定**。
Courier字体	命令或代码。	执行 `cd /d C:/window` 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[] 或者 [a|b]	表示可选项，至多选择一个。	`ipconfig [-all|-t]`
{} 或者 {a|b}	表示必选项，至多选择一个。	`switch {active|stand}`

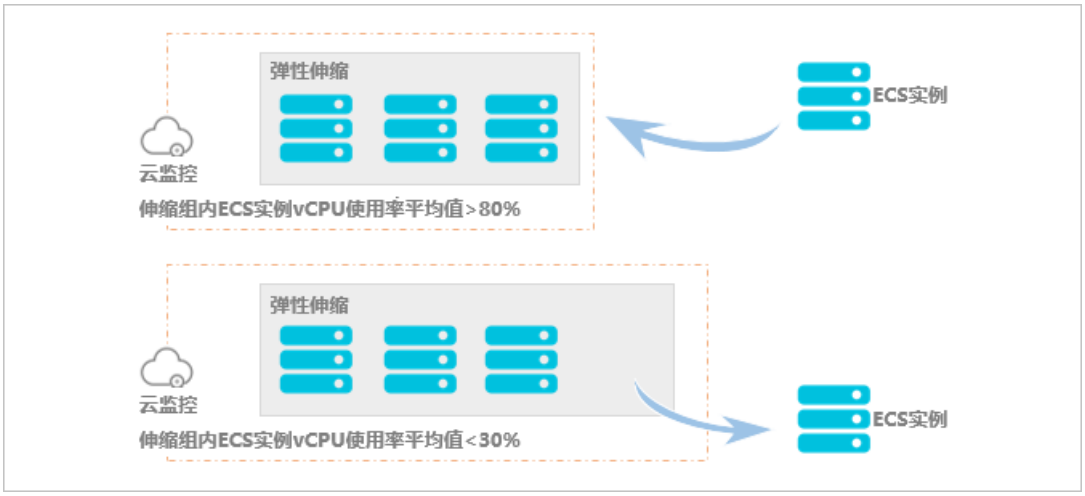
目录

1.什么是弹性伸缩Auto Scaling

使用弹性伸缩（Auto Scaling），您可以根据业务需求和策略设置伸缩规则，在业务需求增长时自动为您增加ECS实例以保证计算能力，在业务需求下降时自动减少ECS实例以节约成本。弹性伸缩不仅适合业务量不断波动的应用程序，同时也适合业务量稳定的应用程序。

弹性伸缩效果示例

您需要提前设置触发弹性伸缩的条件。下图中，监控项为伸缩组内ECS实例的vCPU使用率平均值，并假设触发弹性扩张的阈值为80%，触发弹性收缩的阈值为30%。

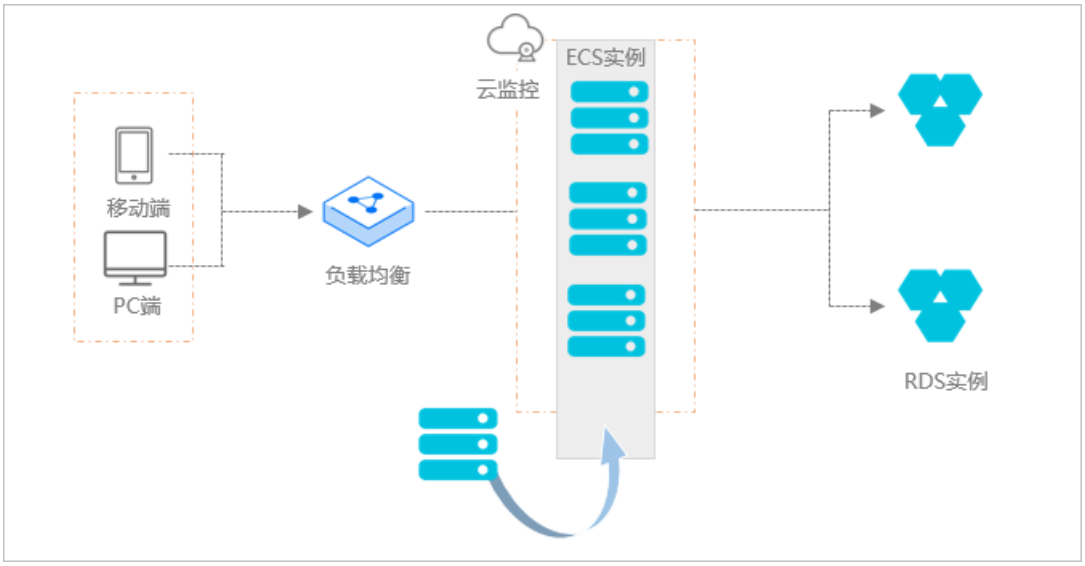


弹性扩张

当您的业务升级时，弹性伸缩为您自动完成底层资源升级，避免访问延时和资源超负荷运行。

您可以配置云监控实时关注您的ECS实例使用情况。例如，当云监控检测到伸缩组内的ECS实例vCPU使用率突破80%时，弹性伸缩根据您配置的伸缩规则弹性扩张ECS资源，自动创建合适数量的ECS实例，并自动添加ECS实例到负载均衡实例的后端服务器和RDS实例的访问白名单中。

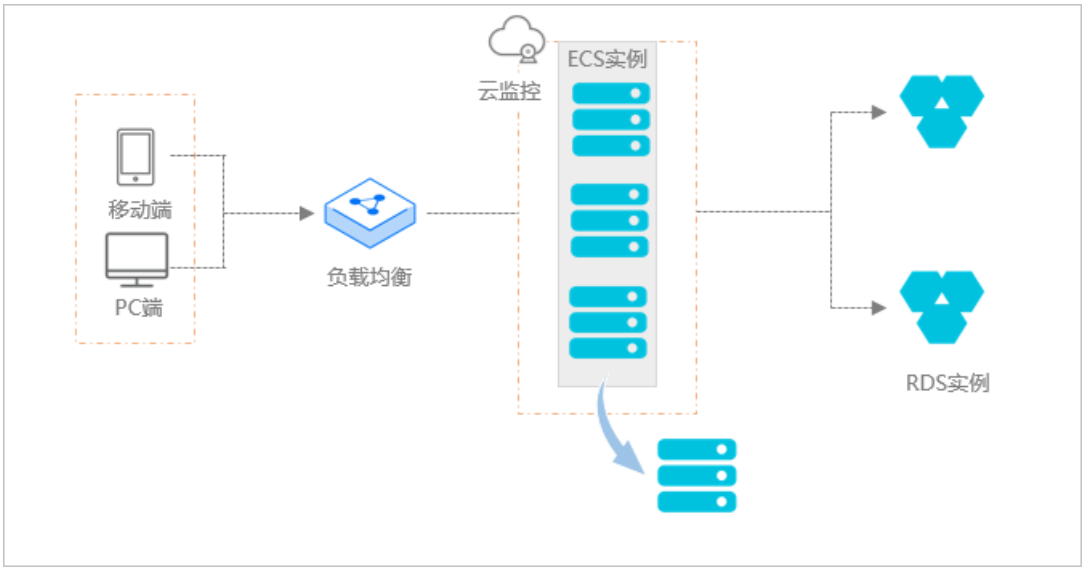
弹性扩张时，弹性伸缩使用伸缩组的组内实例配置信息自动创建ECS实例，实例配置信息支持实例的规格、操作系统、用户自定义数据等，更多说明请参见组内实例配置信息来源概述。您可以登录ECS管理控制台启动、停止已创建的ECS实例，也可以远程登录ECS实例修改系统配置。



弹性收缩

当您的业务需求下降时，弹性伸缩为您自动完成底层资源释放，避免资源浪费。

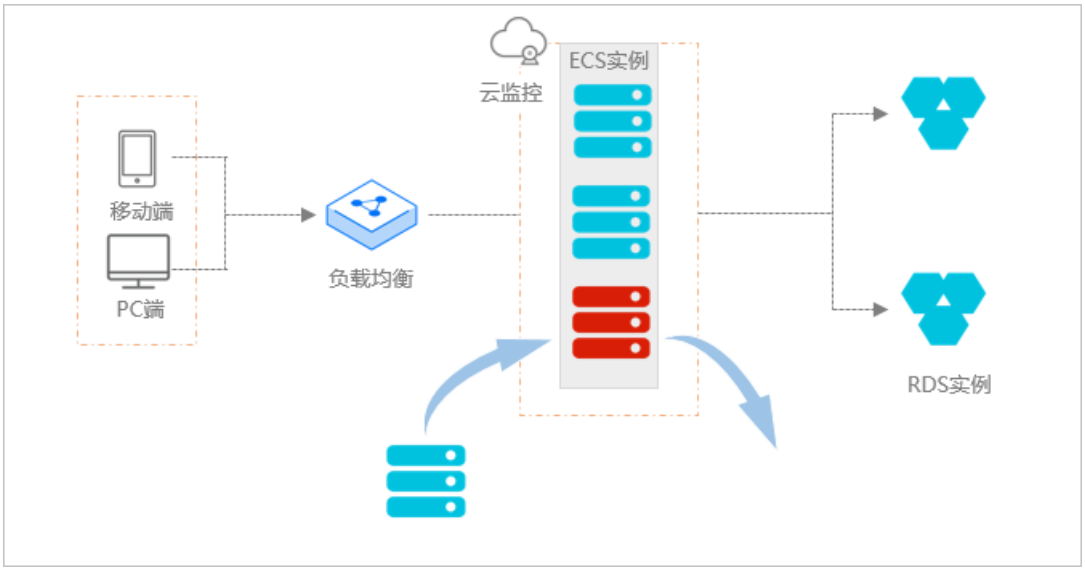
您可以配置云监控实时关注您的ECS实例使用情况。例如，当云监控检测到伸缩组内的ECS实例vCPU使用率低于30%时，弹性伸缩根据您配置的伸缩规则弹性收缩ECS资源，自动释放合适数量的ECS实例，并自动从负载均衡实例的后端服务器和RDS实例的访问白名单中移除ECS实例。



弹性自愈

弹性伸缩提供健康检查功能，自动监控伸缩组内的ECS实例的健康状态，避免伸缩组内健康ECS实例低于您设置的最小值。

当检测到某台ECS实例处于不健康状态时。弹性伸缩自动释放不健康ECS实例并创建新的ECS实例，自动添加新ECS实例到负载均衡实例的后端服务器和RDS实例的访问白名单中。



相关链接

- 云服务器ECS简介
- 云数据库RDS简介

- 负载均衡SLB简介
- 云监控CloudMonitor简介

2.产品优势

本文主要介绍弹性伸缩的功能、产品特点及应用场景。

功能概述

- 根据客户业务需求自动调整ECS实例数量。
- 自动向负载均衡的后端服务器组中添加或移除相应的ECS实例。
- 自动向RDS访问白名单中添加或移除ECS实例的IP。

产品特点

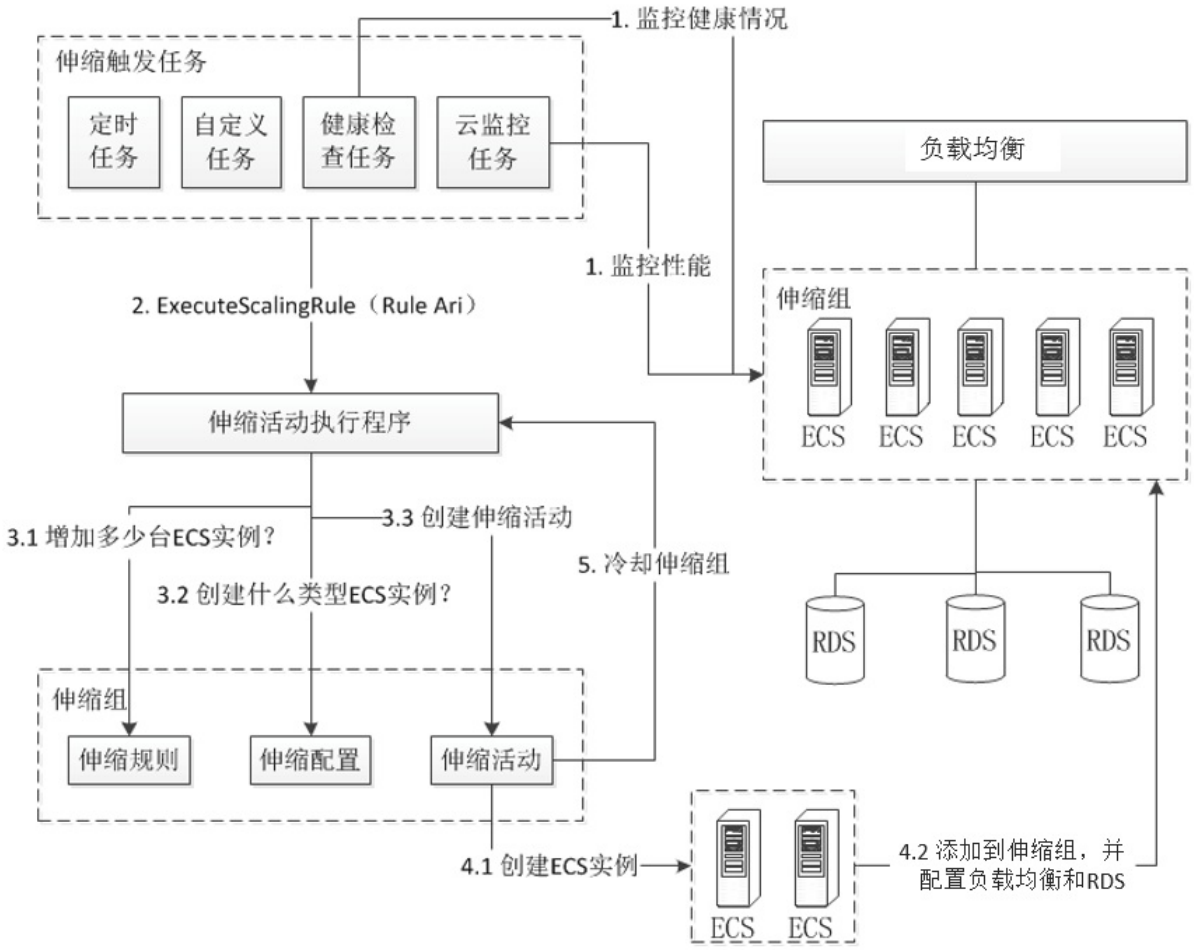
- 随需应变：根据需求“恰到好处”地分配资源，无需您提前预测需求变化，实时应对需求突增。
- 自动化：无需人工干预，自动创建和释放ECS实例，自动配置负载均衡和RDS访问白名单。
- 伸缩模式丰富：多模式兼容，可同时配置定时、动态、自定义、固定、健康模式，可通过API对接外在监控系统。
- 智能：智能调度云计算资源，应对各种复杂场景。

应用场景

- 某视频公司：春晚或每周五热门节目来临时，负载激增，需及时、自动扩展云计算资源。
- 某视频直播公司：业务负载变化难以预测，需要阿里云自动根据CPU利用率、应用负载、带宽利用率作为衡量指标进行弹性伸缩。
- 某游戏公司：每天中午12点及晚上6点到9点间需求增长，需要定时扩容。
- 某电商：在大促中，临时激增大量云服务器需求，需要在数分钟内实现从创建到可用。

3.工作流程

本文主要介绍弹性伸缩的工作流程。



创建好伸缩组、伸缩配置、伸缩规则、伸缩触发任务后，系统会自动化执行以下流程（以增加 ECS 实例为例）：

1. 伸缩触发任务会按照各自触发生效的条件来触发伸缩活动。
   - 云监控任务会实时监控伸缩组内 ECS 实例的性能，并根据用户配置的报警规则（如伸缩组内所有 ECS 实例的 CPU 平均值大于 60%）触发执行伸缩规则请求。
   - 定时任务会根据用户配置的时间来触发执行伸缩规则请求。
   - 您可以根据自己的监控系统及相应的报警规则（如在线人数、作业队列）来触发执行伸缩规则请求。
   - 健康检查任务会定期检查伸缩组和 ECS 实例的健康情况，如发现有不健康的 ECS 实例（如 ECS 为非 **Running** 状态）会触发执行 **移出该 ECS 实例** 的请求。
2. 系统自动通过 ExecuteScalingRule 接口触发伸缩活动，并在该接口中指定需要执行的伸缩规则的阿里云资源唯一标识符（Ari）。

   如果是用户自定义的任务，则需要用户在自己的程序中调用 ExecuteScalingRule 接口来实现。
3. 根据步骤 2 传入的伸缩规则 Ari（Rule Ari）获取伸缩规则、伸缩组、伸缩配置的相关信息，并创建伸缩活动。
   i. 通过伸缩规则 Ari 查询伸缩规则以及相应的伸缩组信息，计算出需要增加的 ECS 实例数量，并获得需要配置的负载均衡和 RDS 信息。

ii. 通过伸缩组查询到相应的伸缩配置信息，即获得了需要创建的ECS实例的配置信息（CPU、内存、带宽等）。

iii. 根据需要增加的 ECS 实例数量、ECS 实例配置信息、需要配置的负载均衡实例和 RDS 实例创建伸缩活动。

4. 在伸缩活动中，自动创建 ECS 实例并配置负载均衡和 RDS。

   i. 按照实例配置信息创建指定数量的 ECS 实例。

   ii. 将创建好的 ECS 实例的内网 IP 添加到指定的 RDS 实例的访问白名单当中，将创建好的 ECS 实例添加到指定的负载均衡实例当中。

5. 伸缩活动完成后，启动伸缩组的冷却功能。待冷却时间完成后，该伸缩组才能接收新的执行伸缩规则请求。

4.伸缩模式

本文主要介绍弹性伸缩的伸缩模式。

- 定时模式：您自定义自动伸缩发生的时间和频率，如每天 13：00增加 ECS 实例。
- 动态模式：基于云监控性能指标（如 CPU 利用率），自动增加或减少 ECS 实例。
- 固定数量模式：通过设置 **最小实例数** （MinSize），即健康运行的ECS 实例最小数量，以保证可用性。
- 自定义模式：通过 API 调用您的自有监控系统，您可以执行手工伸缩。
  - 手工执行伸缩规则。
  - 手工添加或移出既有的 ECS 实例。
  - 自定义 MinSize、MaxSize，弹性伸缩会自动创建或释放 ECS 实例，将当前 ECS 实例数维持在 MinSize 与 MaxSize 之间。
- 健康模式：如 ECS 实例为非 **Running** 状态，弹性伸缩将自动移出或释放不健康的 ECS 实例。
- 多模式并行：以上所有模式都可以组合配置。例如设置了每天 13：00 ~ 14：00 创建 20 个 ECS 实例以应对业务高峰，但实际需求有可能需要多于20个实例， 则您可以选择其他伸缩模式，与定时模式配合一起使用。

5.使用限制

本章节介绍使用弹性伸缩时功能和数量上的限制。

功能限制

使用弹性伸缩时具有以下功能限制：

- 弹性伸缩只支持自动增加或减少伸缩组内ECS实例的数量，不支持自动提升或降低单台ECS实例的vCPU、内存、带宽等配置。
- 部署在伸缩组内ECS实例上的应用必须是无状态并且可横向扩展的。
- 伸缩组内ECS实例可能会被自动释放，因此不适合保存会话记录、应用数据、日志等信息。如有需要，您可以将会话记录等状态信息保存到独立的状态服务器，将应用数据保存到云数据库RDS，将日志存储到日志服务，更多说明请参见什么是云数据库RDS和什么是日志服务。
- 弹性伸缩不支持自动将ECS实例添加到Memcache实例的访问白名单，您需要自行添加，具体操作请参见设置IP白名单。

数量限制

单个账号使用弹性伸缩时的数量限制如下表所示。

配额项	配额值
单个地域下的伸缩组总数	和弹性伸缩使用情况有关，请前往配额中心查看配额值。 **说明** 支持手动申请提升配额值。
单个伸缩组内的伸缩配置总数	
单个伸缩组内的伸缩规则总数	
单个伸缩组可以关联的RDS实例总数	
单个伸缩组可以关联的负载均衡实例总数	
单个伸缩组可以关联的虚拟服务器组总数	
单个伸缩组可以设置的组内最大实例数	
单个地域下的定时任务总数	
单次自动扩缩容可加入或删除的ECS实例总数	500
单个伸缩配置中的多实例规格总数	10
单个伸缩组内的事件通知总数	6
单个伸缩组内的生命周期挂钩总数	6

6.常见概念和操作

本文列举并说明使用弹性伸缩过程中常见的概念和操作。

常见概念

概念	说明	相关文档
弹性伸缩	弹性伸缩是自动调整计算能力（即ECS实例）的服务。您可以根据业务需求进行相关设置，实现在业务需求增长时自动增加ECS实例以保证计算能力，在业务需求下降时自动减少ECS实例以节约成本。	什么是弹性伸缩Auto Scaling
伸缩组	伸缩组是具有相同应用场景的ECS实例的集合。伸缩组定义了组内可容纳ECS实例数的最大最小值、关联负载均衡实例、关联RDS实例等属性。	伸缩组概述
ECS实例	ECS实例等同于一台虚拟服务器，内含CPU、内存、操作系统、网络配置、磁盘等基础的计算组件。云服务器ECS免去了您采购IT硬件的前期准备，让您像使用水、电、天然气等公共资源一样便捷、高效地使用服务器，实现计算资源的即开即用和弹性伸缩。	什么是云服务器ECS
负载均衡实例、SLB实例	负载均衡（SLB）服务是一种对流量进行按需分发的服务，通过将流量分发到不同的后端服务来扩展应用系统的服务吞吐能力，并且可以消除系统中的单点故障，提升应用系统的可用性。	什么是负载均衡
RDS实例	云数据库RDS服务是一种稳定可靠、可弹性伸缩的在线数据库服务，支持主流数据库引擎，并提供容灾、备份、恢复、监控、迁移等方面的全套解决方案。	什么是云数据库RDS
伸缩模式	伸缩模式对应不同的增加、减少ECS实例的操作，包括定时模式、动态模式、固定数量模式、自定义模式、健康模式、多模式并行。	伸缩模式
组内实例配置信息来源	弹性伸缩从您选择的组内实例配置信息来源获取ECS实例配置信息，并使用这些配置信息创建ECS实例。组内实例配置信息来源支持伸缩配置和实例启动模板。	组内实例配置信息来源概述
伸缩配置	伸缩配置是一种组内实例配置信息来源，包含了ECS实例的配置信息。	创建伸缩配置
伸缩规则	• 步进规则、目标追踪规则、简单规则：用于在触发伸缩活动时控制增加、减少ECS实例的数量。 • 预测规则：基于历史监控数据预测未来的指标值，并智能设置伸缩组边界值。	伸缩规则概述
自动触发任务	自动触发任务分为定时任务和报警任务。定时任务可以在指定的时间扩缩容。报警任务基于指定的监控指标动态扩缩容。	• 创建定时任务 • 报警任务概述

概念	说明	相关文档
伸缩活动	伸缩活动用于记录伸缩组内ECS实例数、伸缩组边界值、期望实例数等数量的变化情况。执行伸缩规则、修改伸缩组边界值、修改期望实例数等操作均会触发伸缩活动。	查看伸缩活动详情
期望实例数	为伸缩组开启期望实例数功能后，伸缩组会自动将ECS实例数量维持在期望实例数，无须人工干预。 **说明** 仅支持在创建伸缩组时开启该功能，已经开启该功能的伸缩组支持修改期望实例数。	期望实例数
并行伸缩活动	通过以下方式触发的伸缩活动为并行伸缩活动： • 手动执行伸缩规则、通过定时任务执行伸缩规则 • 手动添加ECS实例、手动移出ECS实例 • 期望实例数检查任务、实例健康检查任务、最大最小值检查任务 如果有执行中的并行伸缩活动，可以再触发其它并行伸缩活动。 **说明** 开启期望实例数功能后，才区分并行伸缩活动和非并行伸缩活动。否则，正在执行伸缩活动时均不能执行其他伸缩活动。	期望实例数
非并行伸缩活动	并行伸缩活动以外的伸缩活动均属于非并行伸缩活动。如果有执行中的非并行伸缩活动，不能再触发其它伸缩活动。 **说明** 开启期望实例数功能后，才区分并行伸缩活动和非并行伸缩活动。否则，正在执行伸缩活动时均不能执行其他伸缩活动。	期望实例数
稳态实例	稳态实例指伸缩组中处于服务中、保护中和备用中状态的ECS实例。	伸缩组内ECS实例的生命周期
伸缩组流程	伸缩组流程指您可以手动暂停、恢复的流程，包括扩容流程、缩容流程、健康检查、定时任务、报警任务，用于更精细地控制伸缩组流程级别的动作。	• 暂停伸缩组流程 • 恢复伸缩组流程
伸缩组内ECS实例的生命周期	伸缩组内ECS实例的生命周期指伸缩组内ECS实例从创建开始到释放结束的过程，ECS实例的生命周期管理方式和创建类型有关： • 弹性伸缩自动创建的ECS实例：由伸缩组管理。 • 您手动创建的ECS实例：如果已托管给伸缩组，由伸缩组管理。如果未托管给伸缩组，由您自行管理。	伸缩组内ECS实例的生命周期

概念	说明	相关文档
生命周期挂钩	生命周期挂钩用于挂起加入或移出中的ECS实例，您可以在挂起期间对ECS实例进行自定义操作。例如，在创建ECS实例后延迟一段时间，测试服务正常后再挂载到 SLB 实例接收流量。	创建生命周期挂钩
冷却时间	冷却时间是指同一伸缩组内成功完成一个伸缩活动后的一段锁定时间。在冷却时间内，伸缩组会拒绝云监控报警任务触发伸缩活动的请求，避免因监控指标值波动导致频繁触发伸缩活动。	冷却时间

常见操作

操作	说明	相关文档
创建伸缩组	创建伸缩组用于管理有相同应用场景的ECS实例。	创建伸缩组
创建伸缩配置	创建伸缩配置用于指定自动创建ECS实例时的配置信息。	创建伸缩配置
创建伸缩规则	创建伸缩规则用于在触发伸缩活动时控制增加、减少ECS实例的数量，或者智能设置伸缩组边界值。	创建伸缩规则
创建定时任务	创建定时任务用于在指定的时间扩缩容。	创建定时任务
创建报警任务	创建报警任务用于基于指定的监控指标动态扩缩容。	报警任务概述
执行伸缩规则	执行已经创建的伸缩规则，支持手动执行、通过定时任务执行、通过报警任务执行。	执行伸缩规则
添加已有ECS实例至伸缩组	将您已经创建的ECS实例手动添加至伸缩组。 **说明** 支持将包年包月实例添加至伸缩组，但不支持托管包年包月实例。	手动添加ECS实例
滚动升级	批量更新伸缩组内ECS实例的配置，支持为伸缩组内处于服务中状态的ECS实例批量更新镜像、执行脚本或者安装OOS软件包。	滚动升级
更新镜像任务	选择一台ECS实例，并使用该ECS实例的镜像替换指定伸缩配置中的镜像，从而在后续扩容出的ECS实例中使用新的镜像。 **说明** 如果需要更新伸缩组内已有ECS实例的镜像，请使用滚动升级功能。	批量修改伸缩配置的镜像
暂停伸缩组流程	主动暂停伸缩组流程，方便您在暂停指定流程之后再执行某些操作。例如，暂停健康检查流程后再去停止ECS实例，避免ECS实例被判定为不健康而自动移出伸缩组。	暂停伸缩组流程

操作	说明	相关文档
恢复伸缩组流程	恢复被暂停的伸缩组流程，由伸缩组继续按功能逻辑执行相关操作。例如，恢复健康检查流程，继续自动检查伸缩组内ECS实例是否健康，并及时移出不健康的ECS实例。	恢复伸缩组流程
删除实例、移除实例	将ECS实例移出伸缩组，并释放ECS实例。	手动移出或删除ECS实例
移出实例	将ECS实例移出伸缩组，但不释放ECS实例。	手动移出或删除ECS实例