

Alibaba Cloud

弹性伸缩 ESS 产品简介

文档版本：20220511

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 <code>cd /d C:/window</code> 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	<code>bae log list --instanceid</code> <code>Instance_ID</code>
[] 或者 [a b]	表示可选项，至多选择一个。	<code>ipconfig [-all -t]</code>
{ } 或者 {a b}	表示必选项，至多选择一个。	<code>switch {active stand}</code>

目录

1.什么是弹性伸缩Auto Scaling	05
2.产品优势	08
3.应用场景	10
4.工作原理	11
5.使用限制	15
6.基本概念	17

1.什么是弹性伸缩Auto Scaling

弹性伸缩（Auto Scaling）是根据业务需求和策略自动调整计算能力（即实例数量）的服务。您可以指定实例的类型，即ECS实例或ECI实例。在业务需求增长时，弹性伸缩自动增加指定类型的实例，来保证计算能力；在业务需求下降时，弹性伸缩自动减少指定类型的实例，来节约成本。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。

本视频以ECS实例为例介绍弹性伸缩产品。

为什么选择弹性伸缩

当您的业务需求有波动时，弹性伸缩能够帮助您自动调整指定类型的实例数量，满足业务需求。弹性伸缩能为您提供以下优势：

- 自动化：
 - 弹性扩张时，自动创建指定类型的实例，并为该实例关联负载均衡服务。如果指定类型是ECS实例，还支持为该实例自动关联RDS实例。
 - 弹性收缩时，自动移出指定类型的实例，并为该实例取消关联负载均衡服务。如果指定类型是ECS实例，还支持为该实例自动取消关联RDS实例。
- 降成本：无需投入大量人力来调整计算资源，无需提前预备计算资源，也无需担心不能及时释放冗余资源。弹性伸缩在适当的时间进行伸缩任务，降低资源拥有成本。
- 高可用：无需担心ECS实例或ECI实例的运行状态。弹性伸缩提供健康检查功能，在ECS实例或ECI实例不健康，即实例未处于运行中状态时，自动增加相应类型的实例替换不健康的实例。
- 灵活智能：
 - 支持指定提供计算能力的实例类型，即ECS实例或ECI实例。
 - 支持多伸缩模式兼容，可灵活调度应对各种复杂场景。伸缩模式包括固定数量、健康、定时、动态、自定义等，其中动态模式支持通过API对接外部的监控系统。
 - 支持灵活的实例模板，提高创建实例的成功率。
 - 支持丰富的扩缩容策略，可灵活适用于各种业务场景。
- 易审计：弹性伸缩支持记录每个伸缩活动，提供伸缩组监控功能，有助于您快速定位问题根源。

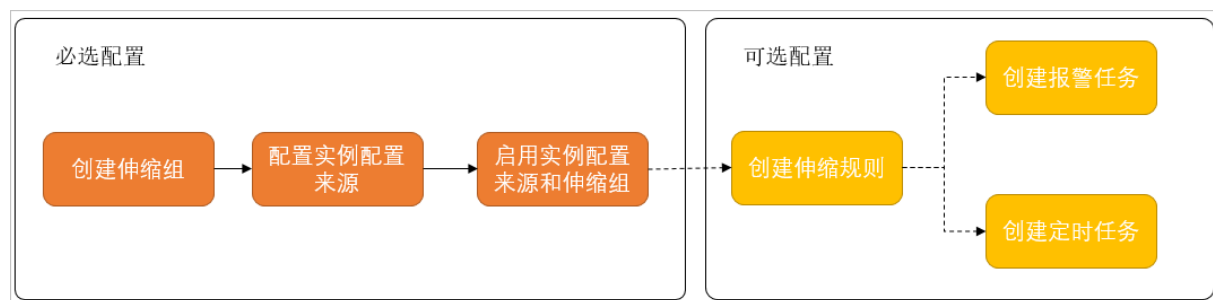
更多信息，请参见[产品优势](#)。

功能说明

弹性伸缩可以根据业务需求，自动创建或者移出ECS实例或ECI实例。您需要配置以下组件：

- **伸缩组**：用来管理一组具有相同应用场景、相同实例类型的实例。您需要指定伸缩组类型（即ECS实例或ECI实例，用于指定提供计算能力的实例类型）、实例配置来源、边界值（即最大实例数和最小实例数）或负载均衡（CLB实例或ALB服务器组）等。如果您有多个应用场景，您可以创建多个伸缩组。弹性伸缩按照您的配置为每个伸缩组分别调整计算能力。
- **实例配置来源**：用来管理ECS实例或ECI实例使用的模板信息。在弹性扩张时，弹性伸缩使用ECS类型的模板信息创建ECS实例，使用ECI类型的模板信息创建ECI实例。
- **伸缩规则**：主要用来触发伸缩活动，比如增加1台ECS实例。您可以手动执行伸缩规则，或者通过报警任务或定时任务执行伸缩规则。伸缩规则还支持智能设置伸缩组的边界值（即最大实例数和最小实例数）。
- **报警任务**：通过云监控系统，实时监测伸缩组的各项指标，在指标满足配置的阈值条件时，执行相应的伸缩规则。
- **定时任务**：指定时间执行相应的伸缩规则。

弹性伸缩必须配置并启用了伸缩组和组内实例配置来源，其他组件可以按需配置。弹性伸缩的使用流程如下图所示。



弹性伸缩还为您提供更多功能，来实现您不同场景的需求：

- 在伸缩活动成功、失败或者被拒绝时，弹性伸缩支持通过以下方式发送通知信息：
 - 消息通知**：支持通过短信、站内信和邮件发送消息通知。
 - 事件通知**：支持发送消息到云监控系统事件或消息服务。消息服务包括MNS主题和MNS队列两种服务模式。消息服务涉及计费，更多计费详情，请参见[价格说明](#)。
- 如果伸缩组类型为ECS实例，即您指定由ECS实例提供计算能力时，还支持以下功能：
 - 生命周期挂钩**：管理伸缩组内ECS实例生命周期的工具。弹性伸缩自动触发扩缩容活动，并触发生命周期挂钩使伸缩活动中的ECS实例处于挂起中的状态（即等待的状态），为您保留一段自定义操作的时间，直至生命周期挂钩超时结束。
 - 自定义方式**：手动向伸缩组添加或移出ECS实例等。
 - 滚动升级**：通过任务形式批量更新ECS实例配置。通过滚动升级，您可以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。

更多信息，请参见[工作原理](#)。

应用场景

弹性伸缩为您提供丰富的伸缩功能，适用各种业务量有变化的场景：

- 业务量变化有规律。比如，每周五20:00热门节目来临时，某视频公司的业务量激增，您可以创建定时任务，在每周五20:00自动增加1台ECS实例或ECI实例。
- 业务量变化无规律。比如，某视频直播公司的日常业务量难以预测，您可以创建报警任务，在CPU使用率大于60%时自动增加1台ECS实例或ECI实例。

更多信息，请参见[应用场景](#)。

产品计费

弹性伸缩本身不收取任何费用，但使用ECS实例、ECI实例、RDS实例、负载均衡服务（CLB实例或ALB服务器组）、消息服务等其他产品的资源时涉及收费。更多信息，请参见[费用说明](#)。

使用方式

- 弹性伸缩控制台**：具有交互式操作的Web服务页面。
- API：支持GET和POST请求的RPC风格API。关于API的更多信息，请参见[API概览](#)。以下为调用弹性伸缩API的常用开发者工具：
 - 命令行工具CLI**：基于阿里云API建立的灵活且易于扩展的管理工具。您可基于命令行工具封装阿里云的原生API，扩展出您需要的功能。
 - OpenAPI Explorer**：提供快速检索接口、在线调用API和动态生成SDK示例代码等服务。

相关服务

- [云服务器ECS简介](#)
- [弹性容器实例简介](#)
- [云数据库RDS简介](#)
- [负载均衡SLB产品家族简介](#)
- [云监控CloudMonitor简介](#)
- [消息服务MNS简介](#)

2. 产品优势

弹性伸缩具有自动化、降成本、高可用、灵活智能以及易审计等优势。

自动化

根据您预设的配置信息，弹性伸缩能够自动化实现以下功能，无需您人工干预，避免因手动操作而可能引入的低错。

- 弹性扩张时：
 - 自动创建指定数量、指定类型的实例（即ECS实例或ECI实例），确保伸缩内所有实例的计算能力能满足业务需求。
 - 如果伸缩组关联了负载均衡，自动为创建的ECS实例或ECI实例关联负载均衡。负载均衡按需将访问请求分发给该ECS实例或ECI实例。
 - 如果伸缩组是ECS类型，且关联了RDS数据库，自动将创建的ECS实例IP添加到RDS访问白名单。该ECS实例可以将应用数据保存到RDS数据库。
- 弹性收缩时：
 - 自动移出指定数量、指定类型的实例（即ECS实例或ECI实例），确保冗余的资源及时得到释放。
 - 如果伸缩组关联了负载均衡，自动为移出的ECS实例或ECI实例取消关联负载均衡。负载均衡不再给该ECS实例或ECI实例分发访问请求。
 - 如果伸缩组是ECS类型，且关联了RDS数据库，自动从RDS访问白名单中移出ECS实例IP。该ECS实例不再保存应用数据到RDS数据库。

降成本

弹性伸缩按需取用，自动释放，提高资源的利用率，有效降低成本。

- 无需提前准备冗余的ECS实例或ECI实例，来防止业务高峰期受到影响；无需担心不能及时释放冗余资源，造成成本浪费。弹性伸缩能够适时调整计算能力，降低了资源的拥有成本。
- 无需投入大量人力来调整计算资源，节约了人力成本和时间成本。

高可用

弹性伸缩支持监测ECS实例或ECI实例的健康状况（即运行状况）。如果发现一台ECS实例或ECI实例未处于运行中状态，则弹性伸缩判定为该ECS实例或ECI实例不健康，并及时自动增加ECS实例替换不健康的ECS实例，或者及时自动增加ECI实例替换不健康的ECI实例，来确保业务的高可用。弹性伸缩可以有效避免因不能及时发现ECS实例或ECI实例的不健康状态，而导致业务连续性受到影响的情况。

灵活智能

弹性伸缩的功能丰富、灵活智能、高可用，可以有效降低手动配置的复杂度，提高操作效率。

- 伸缩模式：支持多模式兼容，可同时配置固定数量、健康、定时、动态、自定义模式。其中，动态模式支持对接云监控服务，可以通过API对接外部的监控系统。更多信息，请参见[伸缩模式](#)。
- 实例配置来源（实例使用的模板）：
 - 支持多种配置方式。例如，从已有实例创建伸缩配置，或者新建伸缩配置。如果是ECS实例使用的模板，还支持指定启动模板。
 - 支持匹配多个ECS实例或ECI实例规格，有效增加模板的灵活性，提高扩容的成功率。比如，ECS实例的模板支持配置多实例规格、多磁盘类型等；ECI实例模板支持指定vCPU和内存来匹配多个实例规格。更多信息，请参见[实例配置来源概述](#)。

- 弹性伸缩策略：如果伸缩组是ECS类型，弹性伸缩支持丰富的扩缩容策略。如果伸缩组是ECI类型，伸缩策略是默认的。
 - 优先级策略：弹性伸缩会在优先级高的可用区扩缩容。如果无法扩缩容，则自动在下一优先级的可用区进行扩缩容。
 - 均衡分布策略：弹性伸缩在多个可用区均衡分布ECS实例，提高可用性。
 - 成本优化策略：当您在组内实例来源中指定了多实例规格时，优先创建vCPU单价最低的ECS实例，优先移出伸缩组内vCPU单价最高的ECS实例。
 - 实例移出策略：您可以选择移出最早伸缩配置对应的实例，或者最早、最新创建的ECS实例。
 - 实例回收模式：移出ECS实例时，您可以直接释放ECS实例，也可以保留部分资源，节省付费。

具体操作，请参见[创建伸缩组](#)。

易审计

弹性伸缩自动记录每一个伸缩活动的详细信息，有助于您快速定位问题根源，降低了排查难度。

弹性伸缩还提供伸缩组监控功能，可以通过云监控查看伸缩组内的实例运行状态。您无需多次查看多台ECS实例或ECI实例的运行状态，有助于您快速了解整体的业务供给能力。

3. 应用场景

弹性伸缩有广泛的应用场景，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。本文为您介绍弹性伸缩的典型应用场景，该场景均适用于ECS实例和ECI实例。

无规律的业务量波动

某新闻网站播出了热点新闻，访问量突增，新闻的时效性降低后，访问量回落。由于该新闻网站的业务量波动无规律，访问量突增和回落的具体时间难以预测，所以手动调整实例很难做到及时性，而且调整数量也不确定。

您可以利用弹性伸缩的报警任务，由阿里云自动根据CPU使用率等衡量指标进行弹性伸缩。

- 示例一：您可以设置两个报警任务，报警任务执行的伸缩规则配置为简单规则类型。一个报警任务用于在实例的CPU使用率超过70%时，自动为您增加3台实例；另一个报警任务用于在实例的CPU使用率低于30%时，自动为您减少3台实例。
- 示例二：您可以设置一个报警任务，报警任务执行的伸缩规则配置为目标追踪规则类型，使实例的CPU使用率一直维持在50%左右。

有规律的业务量波动

某游戏公司每天18:00业务需求急速增长进入高峰期，到22:00业务需求降低，高峰期结束。该游戏公司的业务量波动有规律，但是每天手动调整计算能力浪费人力和时间成本。

您可以利用弹性伸缩的定时任务，由阿里云定时自动进行弹性伸缩。您可以设置两个定时任务，报警任务执行的伸缩规则是简单规则类型。一个定时任务用于在每天17:55自动为您增加3台实例，另一个定时任务用于在每天22:05自动为您减少3台实例。该方式可以很好地应对每天18:00~22:00高峰期的业务量，且在高峰期结束后及时释放实例，不浪费多余的实例资源和成本。

无明显的业务量波动

某通信公司的业务支撑系统需要全天运作，业务量一段时间内无明显波动。如果现有计算资源突然出现故障，会导致业务受到影响，很难及时进行故障修复或者替换。

您可以利用弹性伸缩的高可用优势，开启健康检查模式。阿里云会自动检查实例的健康状态，当发现存在实例不健康时，自动增加实例替换不健康的实例，确保故障的计算资源及时得到修复。而且伸缩组必须设置最小实例数，确保无论在何种情况下，伸缩组内的实例数量都至少等于下限，确保业务可以运作。

混合型的业务场景

如果某公司的业务场景比较复杂，日常业务量波动不明显，且在某个时间段内，业务量是在一定基础上波动的，您已经订购了一部分包年包月的实例，只是想针对波动的业务量合理调整实例数量。

您可以手动将已订购的包年包月实例加入伸缩组，再结合弹性伸缩的报警任务，由阿里云自动根据CPU使用率等衡量指标进行弹性伸缩，更经济、稳定地管理业务的计算能力。

除手动调整实例数量和报警任务，弹性伸缩还支持定时任务、健康检查等。您可以根据业务场景灵活组合以上功能，从而在使用弹性伸缩的时候获得更丰富灵活的使用体验。

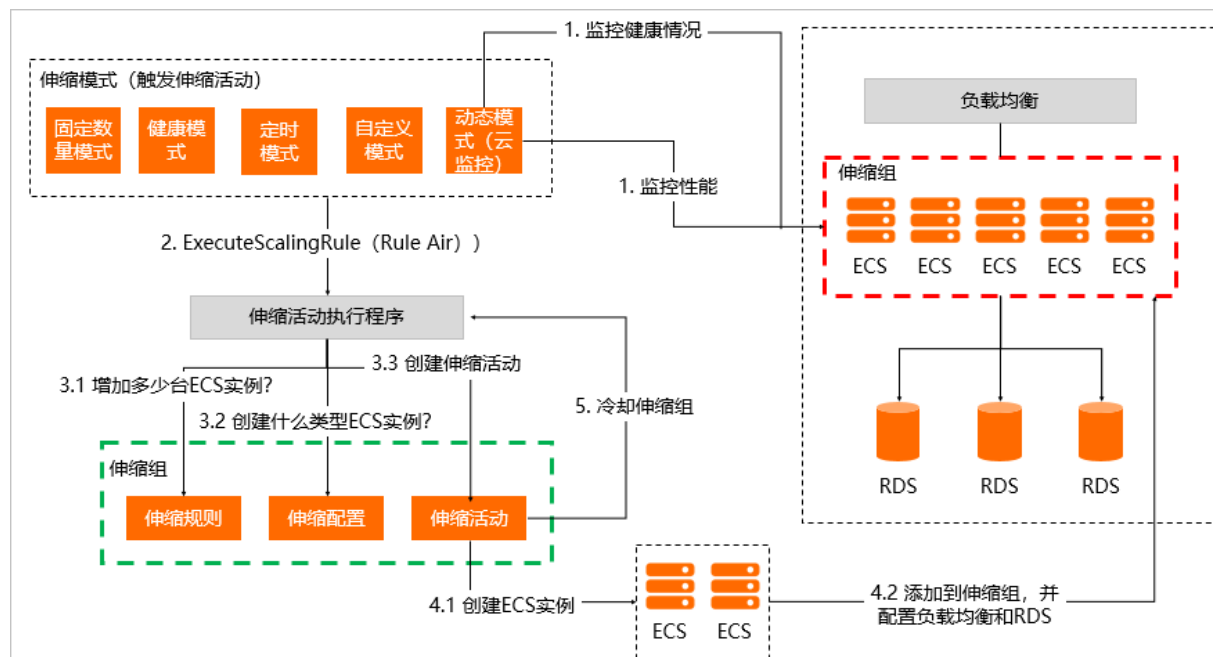
4.工作原理

本文介绍弹性伸缩的工作流程、伸缩模式的配置和弹性伸缩示意图。

本文以伸缩组内的ECS实例为例，介绍弹性伸缩的工作原理。如果您选择伸缩组内的ECI实例，除伸缩组不支持关联RDS数据库和不支持手动加入或删除已有ECI实例外，其他内容与ECS实例一致。

工作流程

弹性伸缩增加ECS实例的工作流程如下所示：



如果某Web应用使用ECS实例处理业务请求，系统架构分为三层（如上图右侧虚线框所示）。其中，最上层的负载均衡负责将客户端的请求转发给伸缩组内的ECS实例，中间层的伸缩组内的ECS实例负责处理客户端请求，最下层的RDS数据库负责存储来自ECS实例的业务数据。

您可以通过弹性伸缩调整中间层的ECS实例数量，从而自动调整处理业务请求的能力。具体流程说明如下所示：

1. 弹性伸缩在符合各伸缩模式的触发条件时自动触发伸缩活动，伸缩模式如下所示。关于伸缩模式的配置方式，请参见[伸缩模式的配置](#)。
 - 固定数量模式：
 - 您在伸缩组设置了**最小实例数**，当伸缩组的ECS实例数量低于最小实例数时，伸缩组会自动添加ECS实例，使得伸缩组内的ECS实例数量等于最小实例数。
 - 您在伸缩组设置了**最大实例数**，当伸缩组的ECS实例数量超过最大实例数时，伸缩组会自动移出ECS实例，使得伸缩组内的ECS实例数量等于最大实例数。
 - 您可以在创建伸缩组时设置**期望实例数**，伸缩组会自动将ECS实例数量维持在期望实例数。
 - 健康模式：您可以在伸缩组开启健康检查功能，伸缩组会定期检查ECS实例的运行状态，如果发现一台ECS实例未处于运行中状态，则判定为该ECS实例不健康并移出。
 - 定时模式：您可以创建定时任务，在指定时间执行指定伸缩规则。
 - 自定义模式：您可以手动进行弹性伸缩，包括手动执行伸缩规则，或者手动添加、移出或者删除已有的ECS实例。

- 动态模式：您可以基于云监控性能指标（例如CPU使用率）创建报警任务，当伸缩组的指标数据满足您指定的报警条件（例如伸缩组内所有ECS实例的CPU平均值大于60%）时，触发报警并执行您指定的伸缩规则。

 **说明** 以上所有模式都可以组合配置，即多模式并行。例如，在每天中午12点开始，业务需求明显增加，您可以设置定时任务，在每天12点创建20台ECS实例以应对业务高峰。但创建的ECS实例台数不一定能满足需求，则您可以选择其他伸缩模式，如动态模式、自定义模式等，与定时模式组合起来使用。

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唯一标识符，例如示例值为 `ari:acs:ess:cn-hangzhou:140692647406****:scalingrule/asr-bp1dvirgwkoowxk7****`。

 **说明** 如果动态模式使用的云监控指标是由您自有的监控系统上报给云监控的，您需要在自己的程序中调用ExecuteScalingRule接口。

- 如果您通过弹性伸缩控制台操作，您可以创建伸缩规则后，在伸缩规则列表下找到对应的伸缩规则，单击**伸缩规则ID/名称**列下的伸缩规则ID（例如asr-bp14u7kzh8442w9z****），在弹出的页面即可查询到该标识符的值。关于如何创建伸缩规则的具体操作，请参见[创建伸缩规则](#)。
 - 如果您通过API方式进行操作，关于如何获取伸缩组下的伸缩规则唯一标识符（ScalingRuleAri），您可以调用[DescribeScalingRules](#)接口，在返回数据中即可查询到该标识符的值。
3. 根据上述步骤传入的伸缩规则唯一标识符（或ScalingRuleAri）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. 通过伸缩规则唯一标识符（或ScalingRuleAri）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、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，并将创建好的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。

伸缩模式的配置

弹性伸缩会自动根据配置适时触发伸缩活动，增加或移出伸缩组内的ECS实例。如果您需要弹性伸缩通过某个伸缩模式触发伸缩活动，您可以选择如下对应的配置方式来实现。

伸缩模式	配置方式	说明
------	------	----

伸缩模式	配置方式	说明
固定数量模式	伸缩组+实例配置来源 ^①	该模式的伸缩效果由伸缩组以下配置项决定： • 最小实例数 • 最大实例数 • （可选）期望实例数
健康模式	伸缩组+实例配置来源 ^①	该模式需要开启伸缩组中实例的健康检查配置项。
定时模式	伸缩组+实例配置来源+伸缩规则+定时任务 ^②	该模式的伸缩效果由定时任务决定。
动态模式	伸缩组+实例配置来源+伸缩规则+报警任务 ^③	该模式的伸缩效果由报警任务决定。
自定义模式	任意伸缩模式的配置方式	在任意伸缩模式下，您可以手动添加、移出或者删除已有的ECS实例。如果您配置了伸缩规则，您还可以手动执行伸缩规则。
多模式并行	组合各伸缩模式的配置方式	根据选用的伸缩模式不同，生效的配置项也不同。并且各伸缩模式是互相独立的，各伸缩模式的伸缩效果没有优先级区分，即伸缩组先触发哪个伸缩模式就先执行哪个伸缩模式对应的伸缩配置项。 例如，为了满足业务新需求，您同时选用了定时模式和动态模式，需要同时配置定时任务和报警任务。如果伸缩组在某个时间先触发定时任务配置项，则先执行定时任务，再执行报警任务。

各配置方式具体说明如下：

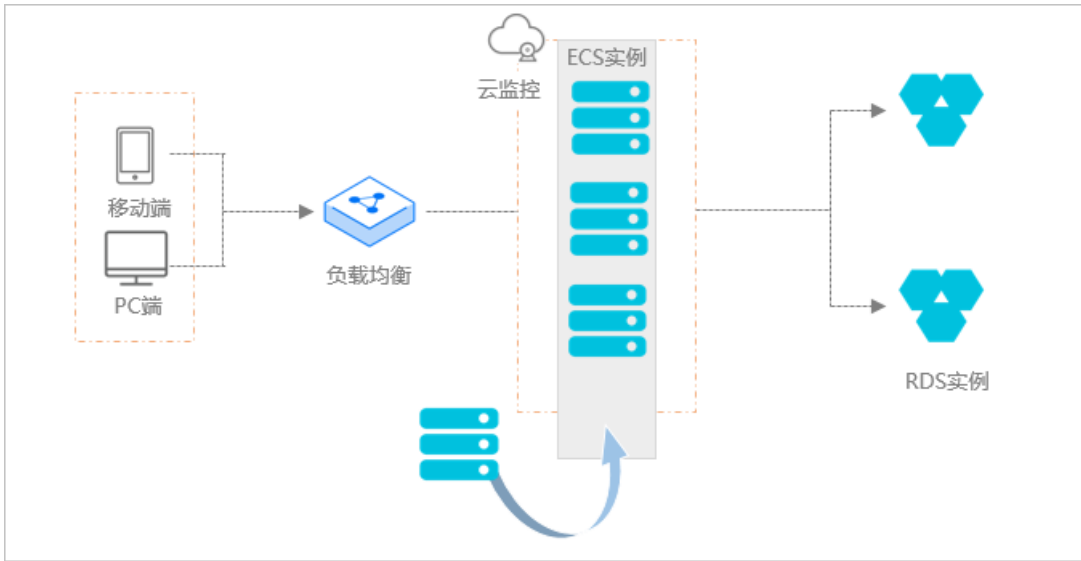
- ① 伸缩组+实例配置来源：您需要先创建伸缩组，再为伸缩组配置实例配置来源，然后启用实例配置来源和伸缩组。完成以上操作后，弹性伸缩才可以自动进行弹性扩张和收缩。该配置方式是必选的基础配置方式，是最小的配置单元。
- ② 伸缩组+实例配置来源+伸缩规则+定时任务：在基础配置方式上（即①），您需要创建伸缩规则，再创建定时任务。弹性伸缩通过定时任务来自动执行伸缩规则。
- ③ 伸缩组+实例配置来源+伸缩规则+报警任务：在基础配置方式上（即①），您需要创建伸缩规则，再创建报警任务。弹性伸缩通过报警任务来自动执行伸缩规则。

弹性伸缩示意图

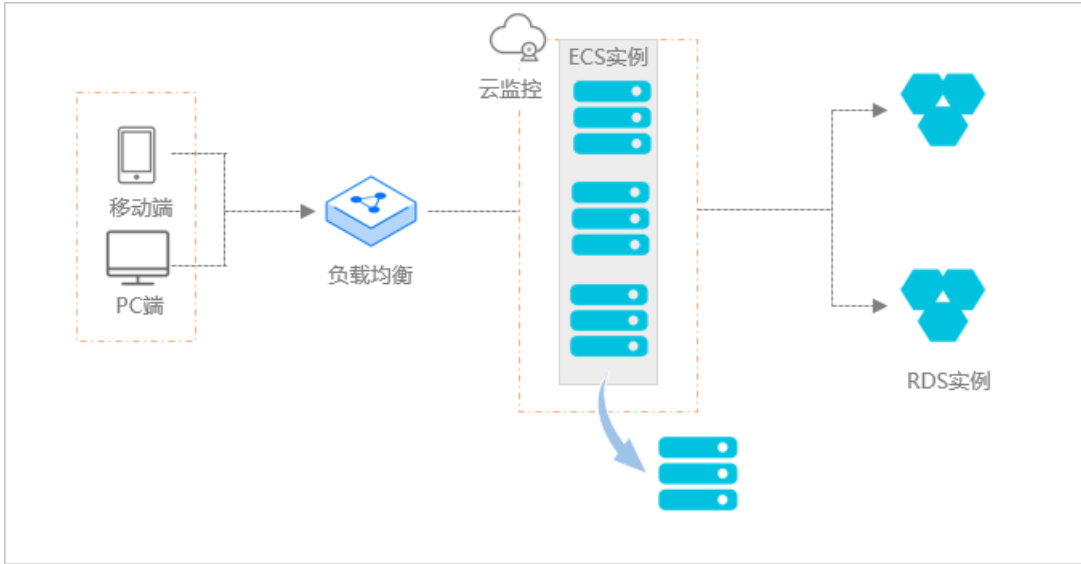
弹性伸缩支持为伸缩组关联负载均衡和RDS实例。当您从终端（例如移动端或PC端）发起请求时，负载均衡负责将您的请求转发给伸缩组内的某台ECS实例，ECS实例接收并处理请求，并将应用数据保存在云数据库RDS实例上。

弹性伸缩根据业务需求和配置方式自动调整伸缩组内的ECS实例数量，以下为您分别提供扩容、缩容和弹性自愈（即健康检查）场景下弹性伸缩的示意图。

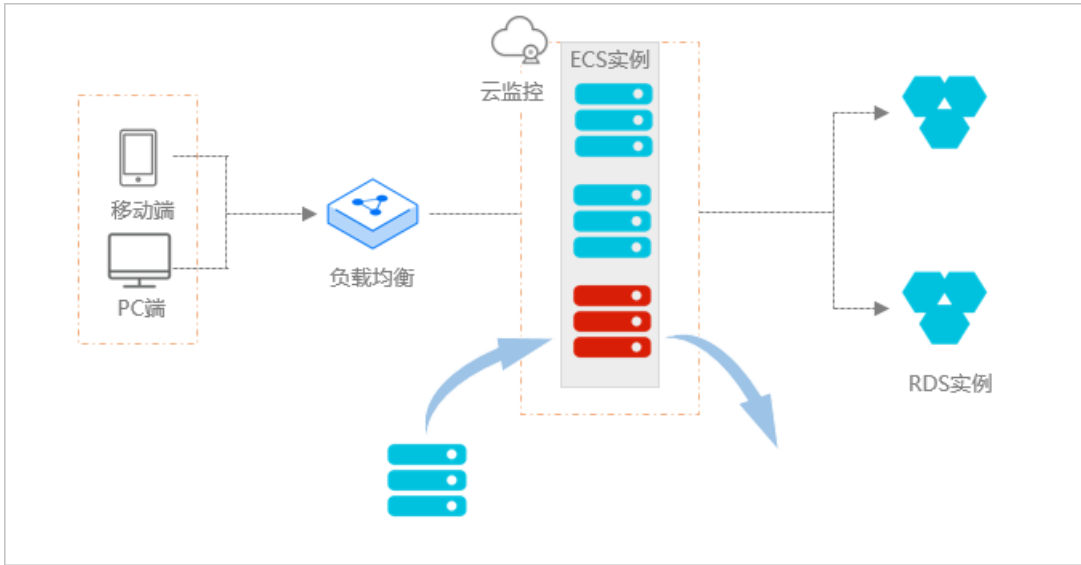
弹性扩容示意图



弹性扩容示意图



弹性自愈示意图



5.使用限制

本文为您介绍使用弹性伸缩时在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：

- 部署在伸缩组内ECS实例或ECI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实例或ECI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，并将应用数据保存到云数据库RDS，将日志存储到日志服务。更多信息，请参见[云数据库RDS简介](#)和[什么是日志服务](#)。
- 弹性伸缩不支持自动将ECS实例或ECI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](#)。
- 如果某个伸缩组关联的RDS实例、ALB服务器组、CLB实例或者CLB实例的后端服务器组被删除，则伸缩组自动解除与该资源的关联。
- 如果某个伸缩组内自动触发的伸缩活动连续失败超过30天，弹性伸缩系统巡检会暂停该伸缩组自动触发伸缩活动的功能，并通过短信或者邮件向您发送通知。

 **说明** 伸缩活动连续失败时，建议您及时处理该问题。您可以登录[弹性伸缩控制台](#)处理，或者请[提交工单](#)联系售后工程师进行协助处理。

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ALB服务器组总数	
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实例或ECI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

配额项	配额值
单个伸缩组内的生命周期挂钩总数	6

6.基本概念

本文列举并说明使用弹性伸缩过程中常见的概念。

常见概念

概念	说明	相关文档
弹性伸缩	弹性伸缩是根据业务需求和策略自动调整计算能力（即实例数量）的云服务。您可以指定实例的类型，即ECS或ECI实例。在业务需求增长时，弹性伸缩自动增加指定类型的实例以保证计算能力，在业务需求下降时自动减少指定类型的实例以节约成本。	什么是弹性伸缩Auto Scaling
伸缩组	伸缩组是具有相同应用场景、相同实例类型的实例集合。伸缩组定义了组内可容纳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
弹性容器实例（ECI实例）	弹性容器实例ECI是阿里云结合容器和Serverless技术为您提供的容器运行服务。	什么是弹性容器实例
负载均衡实例（SLB实例）	负载均衡SLB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。	负载均衡SLB产品家族介绍
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。	云数据库RDS简介
伸缩模式	伸缩模式对应不同的增加、减少实例的操作，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实例配置信息，并使用这些配置信息创建实例。组内实例配置信息来源支持伸缩配置和实例启动模板。	实例配置来源概述
伸缩配置	伸缩配置是一种组内实例配置信息来源，包含了实例的配置信息。	创建伸缩配置（ECS实例）
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

概念	说明	相关文档
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
伸缩活动	伸缩活动用于记录伸缩组内实例数、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 暂停和恢复伸缩组流程

概念	说明	相关文档
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
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