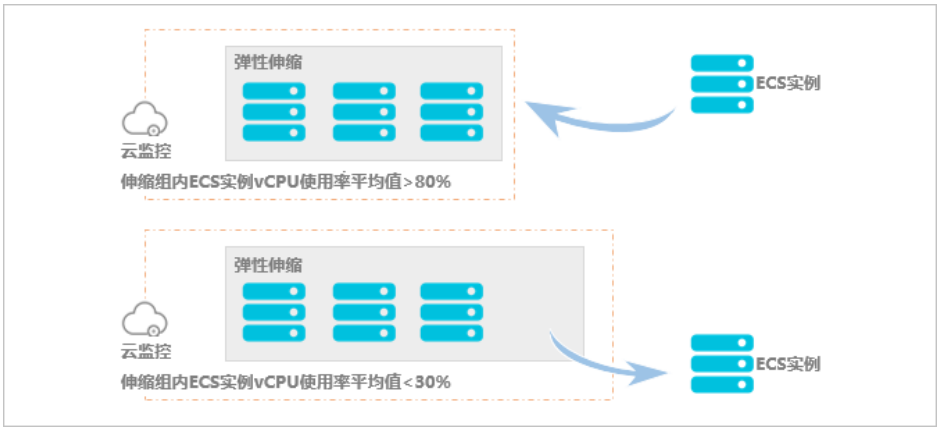
弹性伸缩

产品简介

产品简介

什么是弹性伸缩

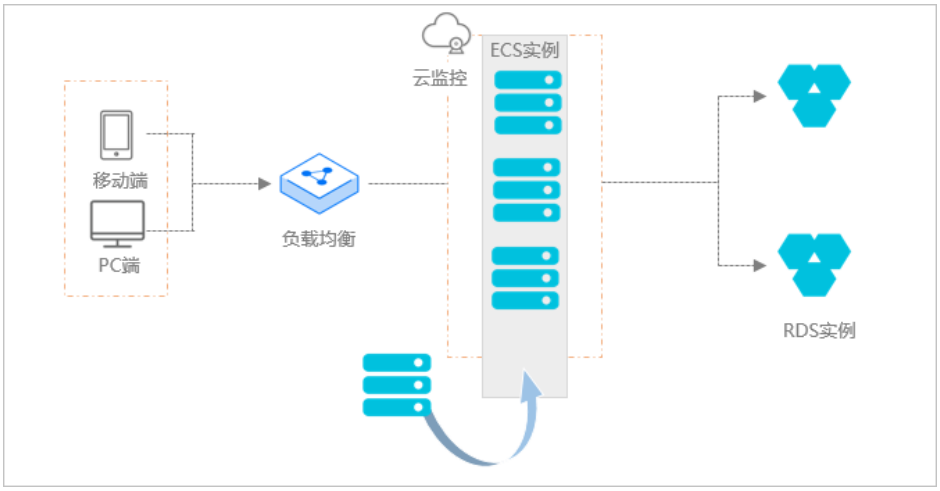
弹性伸缩自动为您调整弹性计算资源大小，以满足您业务需求的变化。弹性伸缩根据您设置的伸缩规则，在业务需求增长时自动为您增加ECS实例以保证计算能力，在业务需求下降时自动减少ECS实例以节约成本。



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。

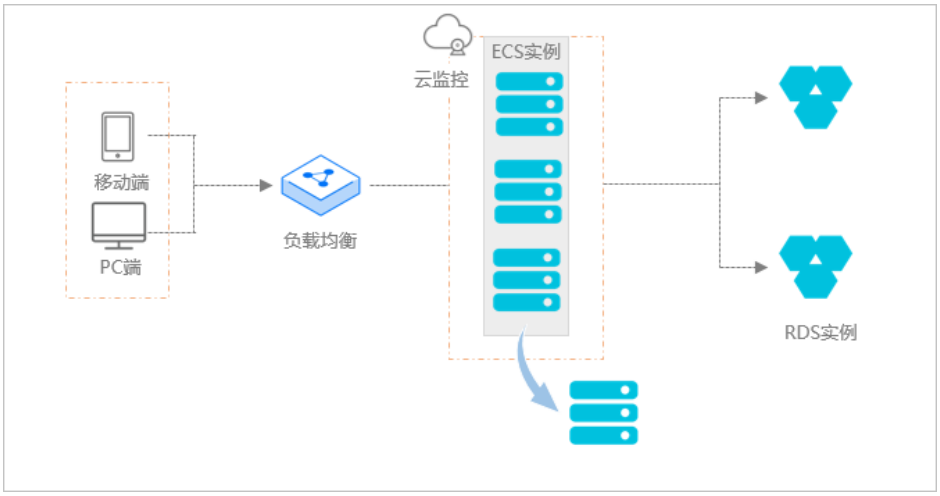
您可以配置云监控实时关注您的ECS实例使用情况。当云监控检测到伸缩组内的ECS实例vCPU使用率突破80%时，弹性伸缩根据您配置的伸缩规则弹性扩张ECS资源，自动创建合适数量的ECS实例，并自动添加ECS实例到负载均衡实例和RDS实例的访问白名单中。更多详情，请参阅 **创建伸缩组** 和 *云监控* **监控弹性伸缩**。



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。

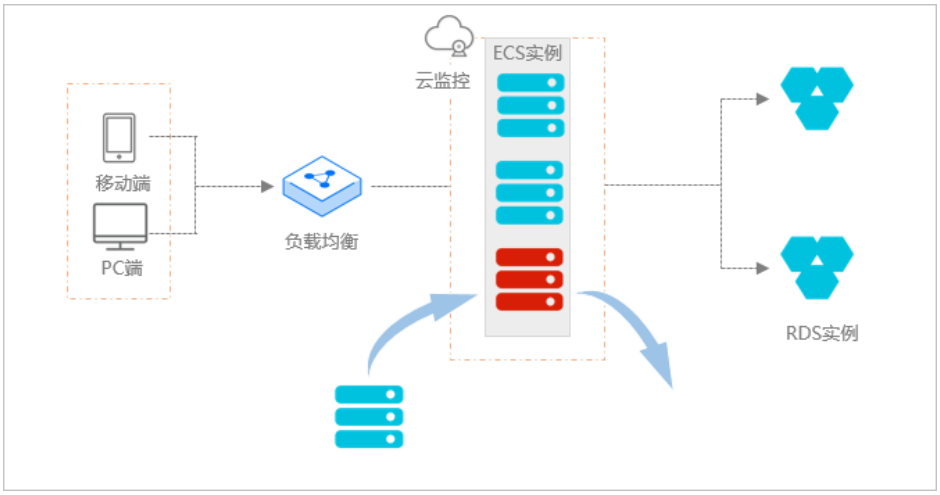
您可以配置云监控实时关注您的ECS实例使用情况。当云监控检测到伸缩组内的ECS实例vCPU使用率低于30%时，弹性伸缩根据您配置的伸缩规则弹性收缩ECS资源，自动释放合适数量的ECS实例，并自动从负载均衡实例和RDS实例的访问白名单中移除ECS实例。更多详情，请参阅 移出策略 和 *云监控* 监控弹性伸缩。



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和RDS实例的访问白名单中。更多详情，请参阅 移出不健康ECS实例。



相关链接

- 什么是云服务器ECS
- 什么是RDS
- 什么是负载均衡
- 什么是云监控

产品优势

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。

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。

限制条件

伸缩组内部署在 ECS 实例的应用必须无状态并且可横向扩展。

弹性伸缩会自动释放ECS实例，所以建议伸缩组内ECS实例不要保存应用状态信息和相关数据等信息，例如会话记录（Session）、数据库或者日志等。如果有需要，您可以保存状态信息到独立的状态云服务器ECS、保存数据库到 云数据库RDS 或者集中日志存储到 日志服务。

弹性伸缩无法自动添加ECS实例到开放缓存 Memcache 实例访问白名单，需要您自行添加。

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规格、内存和带宽等配置。

您能创建的伸缩组、伸缩配置、伸缩规则、ECS实例、定时任务有一定的限制数量。

发展历史

2015年8月27日 弹性伸缩产品上线开放

2014年10月15日 弹性伸缩产品内测

报警任务

弹性伸缩自定义监控项报警任务

弹性伸缩自定义监控项报警任务的监控对象为用户自主上报到云监控中的监控指标。在一些场景下，系统监控项可能不包含您所需要的监控指标，您可能拥有自己的一套监控系统，并且关心的是与您特定业务相关的某些指标,自定义监控报警任务，为您自有的监控系统，或者与业务相关的自有监控指标提供了设置报警任务的接入点。

弹性伸缩自定义监控报警任务是针对阿里云云监控服务的自定义监控项设置报警的，用户在使用弹性伸缩自定义监控报警任务之前需要首先向云监控上报自定义监控数据，即自定义监控项。云监控自定义监控是提供给用户自由定义监控项及报警规则的一项服务，通过此服务，用户可以针对自己关心的业务指标进行监控，将采集到监控数据上报至云监控，由云监控来进行数据的处理，并可以对其设置报警规则。

上报监控数据到云监控

云监控的自定义监控服务为您提供了上报监控数据的方式，您可以将自己采集到的时序数据上报到云监控，这样的数据称作时间序列。云监控提供了OpenAPI、Java SDK 和阿里云命令行工具（CLI）三种方式上报数据，这里我们将主要关注如何使用Java SDK的方式上报监控数据。更多详细的信息您可以查看文档 上报监控数据 。

使用Java sdk之前您需要首先在项目中引入相应的jar包，如果您使用maven管理项目，您只需要在项目中加入以下依赖：

```
<dependency>
<groupId>com.aliyun</groupId>
<artifactId>aliyun-java-sdk-core</artifactId>
<version>3.2.6</version>
</dependency>
<dependency>
<groupId>com.aliyun.openservices</groupId>
<artifactId>aliyun-cms</artifactId>
<version>0.2.4</version>
</dependency>
```

您可以按照如下的方式向云监控上报自定义监控项：

```
static String endPoint     = "https://metrichub-cms-cn-hangzhou.aliyuncs.com";
CMSClient cmsClient = new CMSClient(endPoint, accAutoScalingKey, accAutoScalingSecret);
CustomMetricUploadRequest request = CustomMetricUploadRequest.builder()
.append(CustomMetric.builder()
.setMetricName("myCustomMetric")//自定义指标名
.setGroupId(54504L)//设置分组id
.setTime(new Date())//时间
.setType(CustomMetric.TYPE_VALUE)//类型为原始值,
.appendValue(MetricAttribute.VALUE, number)//原始值，key只能为这个
.appendDimension("key1", "value1")//添加维度
.appendDimension("key2","value2")
.build())
.build();
CustomMetricUploadResponse response = cmsClient.putCustomMetric(request);//上报
```

在上述的代码片段中，我们上报了一个数据点到云监控。上报时，必须指定groupId参数，即云监控应用分组id，该分组id可以是您在云监控中已创建的应用分组，也可以是一个不存在的应用分组。您可以在 云监控应用分组 创建和查看应用应用分组信息。您上报的自定义监控项（时间序列），可以在云监控控制的 自定义监控 中查看。

我们建议您向一个 云监控中已经存在的应用分组中推送自定义监控数据，云监控的应用分组是针对多种云产品的逻辑分组，向一个已存在的应用分组推送自定义监控数据将使您在需要使用云监控等相关功能时保留扩展的能力。当然，您也可以完全不必理会应用分组的概念，选择向任意一个分组id推送数据。

您上报到云监控的监控数据，云监控会自动帮您进行聚合，当然，当您需要推送的数据量太大时，您也可以选择在本地聚合之后再推送到云监控。详细信息可以参考 上报监控数据。

注意事项

云监控对用户上报监控数据设置了以下限制：

- 单云账号QPS限制为100。
- 单次最多上报100条数据，body最大为256KB。
- “metricName”字段只支持字母、数字、下划线。需要以字母开头，非字母开头会替换为大写“A”，非法字符替换为“_”。

- “dimensions”字段不支持“=”、“&”、“,” ，非法字符会被替换为“_”。
- metricName 和dimensions的Key-value 最大均为64字节，超过64字节会被截断。
- 其他限制请关注 计量计费 说明。

弹性伸缩系统监控报警任务

弹性伸缩系统监控报警任务的监控指标是云监控为用户采集的 ECS 实例的相关数据指标，比如CPU，负载等。用户在弹性伸缩中设置的系统监控的报警任务是以伸缩组作为监控粒度的，即以伸缩组内的所有实例的监控指标的统计平均值作为伸缩组的指标值，当伸缩组内实例数量发生变化时，监控指标也会同时进行更新。

支持的监控指标

弹性伸缩系统监控报警任务目前支持的监控项，其中红色字体代表升级版新增支持的监控项：

监控项	单位
CPU	%
内存	%
系统平均负载	无
内网出流量	KB/min
内网入流量	KB/min
外网出流量	KB/min
外网入流量	KB/min
系统盘写bps	Byte/s
系统盘读bps	Byte/s
系统盘写iops	个/s
系统盘读iops	个/s
外网网卡发包数（经典网络）	个/s
外网网卡收包数（经典网络）	个/s
内网网卡发包数	个/s
内网网卡收包数	个/s
TCP总连接数	个
TCP已建立连接数	个

注意事项

- 伸缩组在同一时刻只能执行一个伸缩活动，当伸缩组内存在正在执行的伸缩活动是，由报警任务触发伸缩规则产生的伸缩活动将被拒绝。
- 弹性伸缩报警任务在触发伸缩规则时受伸缩规则冷却时间影响，伸缩规则在冷却时间内时，将拒绝执行伸缩规则。伸缩组内新增加的 ECS 实例从加入伸缩组到完成系统启动配置，部署用户业务，获取到监控数据需要一定的时间（通常需要几分钟），因此您应该根据具体的业务场景，设置合适的冷却时间，防止在新增实例的监控数据缺失的这段时间中，重复触发伸缩规则。
- 弹性伸缩报警任务，默认设置了一分钟的沉默时间，即触发报警之后，一分钟内不会再次触发伸缩规则。
- 部分系统监控项（内存、负载、网卡发包数、TCP连接数）指标的采集需要为您安装云监控客户端。默认情况下，当您针对需要云监控客户端采集的监控项设置报警任务时，将为报警任务关联的伸缩组内的所有实例安装云监控客户端，同时，将为您在云监控控制台开启 新购ECS自动安装云监控，为您所有新购的 ECS 实例安装云监控客户端。

弹性伸缩报警任务

弹性伸缩（Auto Scaling）报警任务是 弹性伸缩 与 云监控服务（CMS） 深度合作，提供的一种动态管理伸缩组的方式，类似于弹性伸缩定时任务，弹性伸缩报警任务通过触发您指定的伸缩规则来执行伸缩活动，达到调整伸缩组内实例个数的目的。

定时任务可以在您指定的时间执行您指定的伸缩规则，当业务场景在时间上可预料时，能够提前做出响应，但是，在面对突发或者时间上不可预料的业务场景时，定时任务就显得捉襟见肘，此时，就需要报警任务来提供更灵活的触发伸缩规则的方式，在业务高峰期增加伸缩组内实例数量来缓解业务压力，在业务低谷时释放伸缩组内实例，减小生产成本。

报警任务通过监控特定的监控指标，对数据指标进行实时的统计，当统计值满足您指定的报警条件时，触发报警，执行您指定的伸缩规则。使用报警任务，您可以实时的根据业务的变化来不断调整伸缩组内的实例数量，保证您监控的指标维持在您期望的范围内。

弹性伸缩报警任务为您提供了一种通过监控特定监控指标来动态调整伸缩组内实例数量的方法，让您能够根据业务的变化实时的执行指定的伸缩规则，调整伸缩组内的实例数量。

弹性伸缩报警任务升级版

弹性伸缩报警任务已经全新升级，从监控范围、监控方式、监控响应速度等三个方面做出了全方位的优化，升级后的弹性伸缩报警任务将您提供一种更全面，更可靠的利用报警任务动态管理伸缩组的方式。

升级内容主要包括：

- 增加了对系统磁盘，网卡，TCP连接数等监控指标的报警任务支持。
- 报警任务最小统计周期升级到 1分钟，提供更灵敏的监控报警。

- 增加自定义监控，为用户自有监控系统接入弹性伸缩报警任务提供标准化方式。

弹性伸缩报警任务升级版扩展了原有的监控指标，并在原有监控项的基础上，支持用户接入自定义的监控项，提供定制化的报警任务，大大增强了弹性伸缩报警任务的可用性和实用性，满足用户具体的，多样化的需求。

名词解释

弹性伸缩

弹性伸缩是根据用户的业务需求和策略，自动调整其弹性计算资源的管理服务。其能够在业务增长时自动增加 ECS 实例，并在业务下降时自动减少 ECS 实例。

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 ECS 实例数的最大值、最小值及其相关联的负载均衡实例和 RDS 实例等属性。

伸缩配置

伸缩配置定义了用于弹性伸缩的 ECS 实例的配置信息。

伸缩规则

伸缩规则定义了具体的扩展或收缩操作，例如加入或移出 N 个 ECS 实例。

伸缩活动

伸缩规则成功触发后，就会产生一条伸缩活动。伸缩活动主要用来描述伸缩组内 ECS 实例的变化情况。

伸缩触发任务

用于触发伸缩规则的任务，如定时任务、云监控的报警任务。

冷却时间

冷却时间是指，在同一伸缩组内，一个伸缩活动执行完成后的一段锁定时间。在这段锁定时间内，该伸缩组不执行其他的伸缩活动。

备注

- 伸缩组包含伸缩配置、伸缩规则、伸缩活动。
- 伸缩配置、伸缩规则、伸缩活动依赖伸缩组的生命周期管理，删除伸缩组的同时会删除与伸缩组相关联的伸缩配置、伸缩规则和伸缩活动。
- 伸缩触发任务有定时任务、云监控报警任务等类型。
- 定时任务独立于伸缩组存在，不依赖伸缩组的生命周期管理，删除伸缩组不会删除定时任务。
- 云监控报警任务独立于伸缩组存在，不依赖伸缩组的生命周期管理，删除伸缩组不会删除报警任务。