

目录

目录	1
启动配置概述	2
创建启动配置	2
操作步骤	2
查看启动配置列表	2
更改伸缩组启动配置	3
检测异常	3
伸缩组概述	3
创建伸缩组	3
查看伸缩组列表	4
修改伸缩组	4
修改伸缩组绑定的云主机	4
将负载均衡与伸缩组结合	4
添加伸缩组负载均衡	4
删除伸缩组的负载均衡	4
删除伸缩组	4
管理告警触发策略	5
创建告警策略需指定条件和动作	5
实例健康检查	5
手动扩容	6
查看伸缩活动日志	6
暂停及恢复扩缩容	6
指定云服务器免于缩容	7
伸缩活动失败	7
冷却时间说明	7
云服务器健康检查	8
创建伸缩活动通知	8
监测告警指标	8

启动配置概述

启动配置是自动创建云服务器的模版，其中包括自定义镜像ID、云服务器实例类型、系统盘及数据盘类型和容量、配置名称、密码等。

创建启动配置的流程和创建KEC实例类似，但创建启动配置并非直接创建KEC实例。您需要触发弹性扩容活动，由弹性伸缩使用启动配置作为模板自动创建KEC实例，例如通过定时任务、报警任务触发弹性扩容活动。

创建伸缩组时必须指定启动配置，启动配置绑定伸缩组且处于触发伸缩活动中时，其属性将不能编辑。

创建启动配置

启动配置定义了用于弹性伸缩的云计算服务器的配置信息，包括云计算服务器的自定义镜像、类型、存储和其他配置信息。

操作步骤

1. 登录[弹性伸缩AS控制台](#)，点击导航条中的启动配置。

2. 选择区域

区域的选择限制了伸缩组可以自动添加的云服务器和伸缩组可绑定的负载均衡。

如果启动配置的区域选择了华北1（北京），那么伸缩组里自动添加的就是华北1（北京）的云服务器。只能绑定华北1（北京）下的负载均衡。

3. 点击**新建**，在弹出页面填写启动配置信息。

- 选择机型：选择伸缩组扩容出的云计算服务器的机型。注：启动配置支持选择多种实例类型和套餐，触发伸缩活动时，会优先选择排序靠上的实例规格。支持多选实例规格排序优先级的上移、下移与删除。具体有关于可选机型的配置见[启动配置支持多实例规格相关限制](#)。

- 选择镜像：支持标准镜像、自定义镜像、共享镜像和镜像市场
- 选择硬盘：配置系统盘与数据盘的类型与容量，有关SSD云硬盘的收费标准，请参见[云硬盘价格](#)
- 根据需要选择购买新的弹性IP或稍后购买
- 设置基本信息：在页面填写配置名称、设置开机密码、传入自定义数据、增加标签等。
- 确认配置信息

在该页面中确认已选择的配置信息，若有需要，可直接返回修改

- 确认信息后点击**立即购买** 完成配置后，此条目将显示在页面的启动配置列表中，示例如下：

启动配置支持多实例规格相关限制	限制	说明
一个启动配置最多支持的实例类型与规格数量上限 10种（例：同种机型的两种套餐规格计为2种）		
计费方式限制	不支持竞价型实例	
同一启动配置不支持同时选择本地盘和云盘机型	详见 云服务器类型	
启动配置不支持的实例规格	<ul style="list-style-type: none"> x86架构:I4、D4 所有GPU架构机型 所有ARM架构机型 	
启动配置中本地数据盘限制	本地数据盘大小上限以选择的多种实例规格中最小的上限为限制	
启动配置下方的订单价格估算限制	以第一优先级的实例规格为准，进行估算	

查看启动配置列表

启动配置是自动创建云服务器的模版，其中包括镜像ID、云服务器配置及类型、系统盘及数据盘类型的容量等。

创建伸缩组时必须指定启动配置：

打开[弹性伸缩AS控制台](#)，选择导航条的启动配置即可查看列表。

如需删除启动配置，请点击相应启动配置条目的删除。

注意：已绑定伸缩组的启动配置无法删除。



更改伸缩组启动配置

需要为伸缩组更换启动配置，请按以下步骤操作：

1. 新建一个启动配置；
2. 更换对应伸缩组的启动配置，在伸缩组的详情页面，点击**编辑**按钮，在弹出窗口的启动配置选项中选择目标启动配置。如图：



检测异常

删除镜像，会导致无法正常扩容云服务器。弹性伸缩会在触发伸缩活动的一瞬间，检测到此类异常，并发出警告。

您可直接在启动配置列表中查看，如下图所示，若有效显示为失效，说明您的启动配置中镜像已被删除，导致不可用。



伸缩组概述

伸缩组是遵循相同规则、面向同一场景的云服务器的集合。伸缩组定义了组内云服务器的最大值、最小值、期望云服务器数、移除策略及其相关联的负载均衡等属性。在伸缩组中配置好实例配置信息来源和网络属性等后启用伸缩组，您就可以通过伸缩规则自动伸缩云服务器，也可以手动添加已有的云服务器。

创建伸缩组

打开[弹性伸缩AS控制台](#)，选择导航条中的**伸缩组**。

1. 选择区域

区域的选择限制了伸缩组可以手动添加的云服务器和伸缩组可绑定的负载均衡。

如果启动配置的区域选择了华北1区（北京），那么伸缩组里自动添加的就是华北1区（北京）的云服务器。只能绑定华北1区（北京）下的负载均衡。

2. 配置信息

点击**新建**按钮，定义伸缩组的属性：

- 伸缩组名称：用于标示这个伸缩组。
- 最小伸缩组：指定伸缩组中最少的实例数量。
- 期望云服务器数：指定伸缩组开始时自动生产的实例数量。伸缩组创建后会生产对应数量的实例。
- 最大伸缩数：指定伸缩组中最大的实例数量。
- 移出策略：当伸缩组要减少实例且有多个选择时，将根据移出策略来选择移出哪个云服务器。
- 启动配置：指定创建好的启动配置，扩容时会按照启动配置来创建扩容云服务器。
- 关联VPC、关联子网、安全组（防火墙）：选择的是扩容出来的机器的网络属性，即扩容出来的机器在某个私有网络（VPC）、子网、安全组中的。
- 多子网扩展策略：当伸缩组需增加实例且指定多个子网在不同可用区时，将根据该策略增加实例。可选择均衡分布或选择优先

- 负载均衡、监听器、服务端口：指定一个负载均衡，扩容出来的机器会自动挂载到该负载均衡下的指定监听器以及相应的服务器端口。当伸缩组未关联负载均衡时，伸缩组内实例的健康状态判断规则为：stopped和error的实例为不健康。

查看伸缩组列表

打开[弹性伸缩AS控制台](#)，单击导航条中的**伸缩组**即可查看列表。



修改伸缩组

1. 打开[弹性伸缩AS控制台](#)，选择导航条中的**伸缩组**。
2. 选择要修改的伸缩组，点击**伸缩组管理**进入伸缩组基本信息页面。



3. 点击**编辑**，可修改伸缩组名称，调整最小、最大伸缩数，修改云服务器移出策略等。



修改伸缩组绑定的云主机

打开[弹性伸缩AS控制台](#)，选择导航条中的**伸缩组**。

选择要修改的伸缩组，点击**伸缩组管理**进入伸缩组基本信息页面。



在该页面中可查看该伸缩组所绑定的云主机列表。



- 如需手动添加云服务器到伸缩组，点击**添加云主机**，选择要添加的云服务器，然后点击**确定**；
- 如需解绑某个云服务器机，在相应的云服务器条目后点击**移出**。

对自动生产的机器，移出后会销毁。

对手动加入的机器，移出后不会销毁，只会从伸缩组中移出，以及解绑负载均衡。

将负载均衡与伸缩组结合

伸缩组附加负载均衡器后，负载均衡自动注册伸缩组中的云服务器，并将流量转发到这些云服务器上。

添加伸缩组负载均衡

方式一：新建伸缩组时关联负载均衡 在[弹性伸缩AS控制台](#)，选择**新建**伸缩组，具体操作参考[创建伸缩组](#)，第二步中选择您需要的负载均衡。如果您没有事先创建好，需要创建新的负载均衡。



方式二：修改伸缩组关联负载均衡

- 在[弹性伸缩AS控制台](#)，选择伸缩组，点击**管理**进入需要修改的伸缩组的详情页面，
- **负载均衡**信息下，点击**新增负载均衡**，选择您需要的负载均衡，如果您没有事先创建好，需要创建新的负载均衡。



注：伸缩组关联的负载均衡实例必须与伸缩组在同一个网络环境中。

删除伸缩组的负载均衡

1. 只能在网络控制台对负载均衡进行删除。
2. 删除后伸缩组中的机器也会自动与被删除的负载均衡解绑定。

删除伸缩组

打开[弹性伸缩AS控制台](#)，选择导航条中的**伸缩组**。

在伸缩组列表中，每个伸缩组的最后都有**删除**按钮。注：您需要将伸缩组关联的云服务器移出后，才能删除该伸缩组。

管理告警触发策略

支持根据监测指标动态扩展伸缩组中的云服务器数量。您需定义告警触发策略，即触发扩展的监测指标状态以及如何按照需求变化进行扩展。

创建告警策略需指定条件和动作

条件格式 监测指标+阈值+周期+连续达到阈值的周期数。即指标在连续N个周期都达到了阈值。

监测指标

- CPU利用率
- 内存利用率
- 网卡出流量
- 网卡入流量
- GPU利用率
- GPU显存利用率
- 监听器出流量
- 监听器入流量

伸缩组活动 发送通知 + 增加/减少 指定数量或者指定百分比的云服务器或者直接指定云服务器数量，并选择冷却时间。

操作步骤

1. 打开[弹性伸缩AS控制台](#)，选择导航条中的**伸缩组**。
2. 选择要修改的伸缩组，单击伸缩组ID进入伸缩组基本信息页面。
3. 在上方的导航条中选择**告警触发策略**，在该页面管理与伸缩组相关联的告警触发策略。

4. 单击**新建**可添加新的告警触发策略，按照**创建告警策略需指定条件和动作**；

5. 单击**修改**可需改该条告警触发策略内容；

6. 单击**删除**可删除该条告警触发策略；

实例健康检查

在新建伸缩组时需要指定期望云服务器的数量，伸缩组将新建与期望云服务器相等的云服务器，同时，伸缩组会确保运行的云服务器大于等于最小伸缩数，小于等于最大伸缩数。

注意：

- **最小伸缩数**：伸缩组中允许的云服务器最小数量。当伸缩组的云服务器数量小于最小伸缩数时，弹性伸缩会增加云服务器，使得伸缩组当前云服务器数匹配最小伸缩数。
 - **期望云服务器数**：伸缩组刚创建时的云服务器数量，后续也可调整作为伸缩组稳定保持的云服务器数量。
 - **最大伸缩数**：伸缩组中允许的云服务器最大数量。当伸缩组的云服务器数量大于最大伸缩数时，弹性伸缩会移出云服务器，使得伸缩组当前云服务器数匹配最大伸缩数。
2. 为了保持伸缩组中的云服务器正常运行，弹性伸缩会对伸缩组内云服务器的运行状况执行定期检查。如果发现云服务器运行状况不佳，它将终止该云服务器，并启动一台新的云服务器。
 - **替换不健康云服务器**

不健康的云服务器被标记为不健康后，伸缩组将立即启动新的云服务器对它进行替换（设置了“移出保护”的云服务器除外）。

手动扩容

弹性伸缩除支持自动扩容外，还支持手动添加云服务器，达到快速手动扩缩容的效果。

1、将已有云服务器添加到伸缩组中

伸缩组提供手动添加云服务器的功能，可通过将一个或多个云服务器添加到现有伸缩组，与伸缩组的其他机器一起观察负载和管理。

2、手动加入伸缩组的云服务器的条件？

手动加入伸缩组的云服务器必须满足以下要求：

- (1) 非试用实例；
- (2) 实例处于运行状态；
- (3) 实例与伸缩组位于同一地域；
- (4) 实例的网络属性必须与伸缩组一样，即属于同一个VPC和子网。

3、手动添加云服务器后的处理

- 弹性伸缩会将该组的已有云服务器数量与手动添加的云服务器数量相加。比如您伸缩组目前已有云服务器是2，手动增加2台云服务器后，您伸缩组的现有云服务器数会变为4。（如果手动添加的云服务器的数量加上现有云服务器数量超过伸缩组的最大伸缩数，请求将失败）
- 手动加入的云服务器会自动添加到伸缩组的弹性伸缩中。
- 伸缩组缩容时会先移出自动创建的云服务器，没有自动创建的云服务器时，才会选择移出手动加入的云服务器。伸缩组移出手动加入的云服务器时，只是移出伸缩组并解绑负载均衡，不会销毁您的云服务器。

4、使用控制台手动添加云服务器示例

点击您要管理的伸缩组管理，进入伸缩组详情页，选择关联云服务器标签，点击添加云服务器，在对话框中勾选对应的云服务器，然后点击确定即可。



查看伸缩活动日志

1. 打开[弹性伸缩AS控制台](#)，选择导航条中的伸缩组。选择要查看的伸缩组，点击伸缩组管理进入伸缩组基本信息页面。



2. 在上方的导航条中选择伸缩活动日志，在该页面可查看该伸缩组根据伸缩策略已执行过的伸缩活动日志。




暂停及恢复扩缩容

使用场景：如果您需要排查配置或与应用相关的其他问题（例如关机重置密码、升级业务等），希望在不触发自动伸缩流程的前提下对应用进行更改，那么您可以暂停伸缩组，完成后再恢复。

1. 打开[弹性伸缩AS控制台](#)，选择导航条中的伸缩组，在伸缩组列表的右侧点击禁用。



2. 设置完后，在状态列中可看到禁用字样。 

注意事项

禁用伸缩组后，伸缩组扩缩容自动触发活动不会进行，但是伸缩组的限制仍然生效。

(1) 禁用后不会触发以下活动：

- 定时任务；
- 告警伸缩；
- 手动修改期望云服务器数；
- 健康检查；

(2) 禁用后仍会触发以下活动：

- 若手动移出时，伸缩组内实例数小于伸缩组最小实例数时，不允许移出；
- 若手动加入时，伸缩组内实例数超过伸缩组最大实例数时，不允许加入；
- 伸缩组禁用后，当前实例将继续收费和运行，只是不进行伸缩活动。

恢复伸缩组

如您已完成暂停伸缩组活动期间的问题排查或操作，您可为您的业务恢复自动伸缩设置。

打开[弹性伸缩AS控制台](#)，选择导航条中的**伸缩组**，在伸缩组列表的右侧点击**启用**即可。



指定云服务器免于缩容

在伸缩组中，当指定某台云服务器在缩容活动时不被缩容时，当缩容活动时，弹性伸缩不会对该云服务器进行删除。

控制台操作步骤：

1. 打开[弹性伸缩AS控制台](#)，选择导航条中的**伸缩组**，选择要修改的伸缩组，点击**管理**进入伸缩组基本信息页面。



2. 在上方的导航条中选择**关联云服务器**，在需要免于缩容的云服务器设置**移除保护**，点击操作中的**设置移出保护**：



3. 点击**确定**该云服务器即可免于缩容。

伸缩活动失败

1. 如何查看失败的伸缩活动？

- (1) 可在伸缩组中查看伸缩活动日志。
- (2) 可设置通知，第一时间知道伸缩活动失败。

2. 为什么会发生失败的伸缩活动？

请查阅[伸缩组扩容失败](#)

冷却时间说明

什么是冷却时间

冷却时间是伸缩组的一个可配置设置，设置冷却时间，可以确保在上一扩展活动生效前弹性伸缩不会启动或终止其他云服务器。

手动扩展伸缩组和弹性伸缩替换运行状况不佳的云服务器，默认不等待冷却时间。

为什么需要冷却时间

机器加入伸缩组后，需要一段时间才能将负载降下来。如果没有冷却时间，系统会在负载降下来前不断扩容，新加入的机器接管业务后，发现负载过低，然后又缩容。

示例场景

业务出现流量高峰，导致告警策略的警报触发。该警报触发时，弹性伸缩会启动一个云服务器来帮助处理增加的需求。但是存在一个问题：该云服务器需要几分钟的时间才能启动，并且启动后需要时间逐渐从负载均衡接收请求。在此期间，监测警报可能会继续触发，从而导致弹性伸缩在警报每次出现时都另外启动一个云服务器。

但若您设置了冷却时间，弹性伸缩在启动一个云服务器后，将暂停所有简单扩展策略或手动扩展引起的扩展活动，直至经过了该指定时间量（默认值为120秒）。这样，新启动的云服务器有时间开始处理应用程序流量。

冷却时间过后，所有暂停的扩展操作都会恢复。如果警报再次触发，则弹性伸缩将启动另一个云服务器，而冷却时间也会再次生效。不过，如果新增的云服务器足以将CPU使用率降为正常水平，则该组会保持其当前大小。

设置冷却时间

默认的冷却时间为120秒。

如需修改，请按以下步骤进行：

- 打开伸缩组的详情页；
- 单击**告警触发策略**，选择要设置的告警伸缩策略，选择**修改**，在修改框下方指定冷却时间的时长（可设置为 120-99999秒）

云服务器健康检查

弹性伸缩定期对伸缩组中的云服务器运行状态进行检查，如果发现云服务器的运行状态为不健康，会标识出伸缩组内所有运行状况不健康的云服务器，并会替换这些云服务器。具体可参考[实例监控检查](#)。

创建伸缩活动通知

在伸缩组的管理页中，点击**通知设置**，然后创建通知策略。

通知内容：扩容成功、扩容失败、缩容成功、缩容失败、替换不健康云服务器成功、替换不健康云服务器失败

通知方式：主账户的站内信

监测告警指标

利用金山云的云平台监测能力，按一组有序的时间序列数据（称为指标）来检索统计数据。您可使用这些指标来验证您的系统是否按预期运行，如果超过了阈值，则会进行扩缩容。

AS指标：

- CPU利用率
- 内存利用率
- 网卡出流量
- 网卡入流量
- GPU利用率
- GPU显存利用率
- 监听器出流量
- 监听器入流量

每个指标可以支持以下维度：

- 最大值
- 最小值
- 平均值

指标聚合方法

金山云弹性伸缩是对云服务器集群进行监测，这会涉及到多个云服务器以及这些云服务器在时间周期内产生的多个监测数据，

这些数据会先进行聚合，再根据用户配置策略进行操作。

统计的基本策略是每个周期对每台云服务器的设定监测项进行1分钟取值（每分钟取一个值），若取到的值连续多个周期都符合设定的规则（周期数用户可自定义），则会触发告警伸缩行为。

例如：

某伸缩组中有3台云服务器，定义的告警伸缩策略是：CPU利用率在5分钟内的最大/最小/平均值大于50%，发生3次。

弹性伸缩采集监测数据和策略判断，步骤如下

步骤1：系统会每分钟对每台云服务器取1个值，一个周期（当前设置为5分钟）里取了15个CPU使用率的值。

步骤2：根据配置是最大值/最小值/平均值结合策略进行判断是否符合告警规则。

最大值：如果这15个值中的最大值有超过阈值（50%）的，该周期符合告警伸缩规则。

最小值：如果这15个值中的最小值有超过阈值（50%）的，该周期符合告警伸缩规则。

平均值：如果这15个值的平均值有超过阈值（50%）的，该周期符合告警伸缩规则。

步骤3：如果连续3个周期（共15分钟，每5分钟判断当前周期）都符合此规则，则会触发伸缩行为。