

弹性伸缩服务

最佳实践

文档版本

02

发布日期

2025-08-20



版权所有 © 华为云计算技术有限公司 2025。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 搭建可自动伸缩的 Discuz!论坛网站.....	1
2 扩缩容时为 ECS 实例自动绑定和释放 EIP.....	12
3 设置扩缩容策略和多实例规格降低成本.....	15
4 使用均衡分布策略部署高可用计算集群.....	19
5 使用标签鉴权精细化管理弹性伸缩.....	22

1 搭建可自动伸缩的 Discuz!论坛网站

应用场景

Discuz!论坛是全球成熟度最高、覆盖率最大的论坛软件系统之一。用户对论坛的访问可分为高峰期和平峰期，若论坛采用多服务器部署模式且满足高峰时期的负载需求，平峰期必有部分服务器处于闲置状态，增加了不必要的成本，也造成了资源浪费。

弹性伸缩可帮助您解决以上问题。当您在论坛的服务器系统中应用弹性伸缩后，弹性伸缩可以根据您设定的策略，自动地增加或减少服务器的数量，在保证您的网站正常运转的同时节约成本。本实践以搭建可自动伸缩的Discuz!论坛为例，介绍了如何使用弹性伸缩服务搭建一个可自动增加或减少弹性云服务器数量的Web服务。

方案介绍

为了实现创建可自动伸缩的Discuz!论坛，您需要按照表1-1中的步骤进行网站的搭建，本文重点介绍创建弹性伸缩实现云服务器自动伸缩的过程。当网站的负载增加时云服务器的CPU使用率会增大，负载降低时CPU使用率会降低。配置两条监控CPU使用率的告警策略，分别在CPU使用率高于70%时增加一台云服务器，在CPU使用率低于30%时减少一台云服务器，保证Discuz!论坛始终有合适数量的云服务器，实现自动伸缩云服务器的功能。

表 1-1 搭建 Discuz!论坛步骤

任务	分类	子任务描述	说明
搭建网站	申请服务	申请虚拟私有云	申请为云服务器提供网络服务的虚拟私有云vpc-DISCUZ。
		购买弹性公网IP	需申请使云服务器和互联网互通的弹性公网IP。
		创建安全组并添加规则	为了保证论坛的网络安全，需要设置安全组对网络访问进行控制。创建的安全组sg-DISCUZ。

任务	分类	子任务描述	说明
	配置服务器	购买弹性云服务器	需要购买两台弹性云服务器，云服务器discuz01用于部署论坛数据库，discuz02用于部署论坛业务。购买云服务器discuz01时绑定之前购买的弹性公网IP，discuz02暂不绑定弹性公网IP。
		在discuz01上搭建数据库	在discuz01上安装云数据库 RDS for MySQL，启动RDS for MySQL，设置开机自启动。
		在discuz02上部署网站代码	先将discuz01上的弹性公网IP解绑，再绑定至discuz02，在discuz02上部署Web环境和网站代码。
	配置特性	释放弹性公网IP	为了节省弹性公网IP资源，使用负载均衡服务前请先释放discuz02绑定的弹性公网IP。
		创建弹性负载均衡	为了在伸缩组中均衡访问网站的流量，需要购买独享型负载均衡监听器elb-DISCUZ。
		制作镜像	为了后续增加的云服务器可以自动搭建Web环境和部署网站代码，需要制作discuz02的镜像discuz_centos6.5(40GB)，该镜像在创建伸缩配置时作为私有镜像使用。
创建弹性伸缩	-	创建伸缩配置	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。创建伸缩配置as-config-discuz。
		创建伸缩组	伸缩组是云服务器进行伸缩的基本单位，伸缩活动将会以伸缩组为单位进行。创建弹性伸缩组as-group-discuz。
		创建伸缩策略	伸缩策略能够触发伸缩活动，配置两条监控CPU使用率的告警策略，在业务负载增加时增加云服务器数量，在业务负载减少时减少云服务器数量。
		手动移入实例	为保证discuz02可以和后续移入伸缩组中的服务器共同承载论坛业务，需要将discuz02手动移入伸缩组。
		修改最小实例数	最小实例数定义了伸缩组中云服务的最少数量，修改最小实例数为1后，伸缩组至少会保证有一台云服务器。discuz02是手动移入，在实例移除策略中被移出的优先级最低，故修改最小实例数可以保证discuz02在伸缩组中不被移出。
访问网站	验证配置结果	验证网站是否可以正常访问	获取负载均衡服务的弹性公网IP地址，在浏览器中输入 http://弹性公网IP地址/forum.php 进行验证。若可以访问则说明各项配置已生效。

前期准备

请您参考《[搭建Discuz!论坛网站](#)》完成[表1-1](#)中搭建网站部分的任务。

创建伸缩配置

伸缩配置定义了移入伸缩组的云服务器的规格，为了移入伸缩组的云服务器能自动承载业务，使用镜像discuz_centos6.5(40GB)，并使伸缩配置中的参数和discuz02保持一致。

1. 登录管理控制台，选择“计算 > 弹性伸缩”。
 2. 在“伸缩实例”页面，单击“创建伸缩配置”。
- 参考[表1-2](#)进行关键参数配置，未列出的参数选择默认值即可。

图 1-1 伸缩配置参数

< | 创建伸缩配置

★ 计费模式

按量计费 ⓘ

★ 区域

华北-北京一

不同区域的云服务产品之间内网互不相通，请就近选择靠近您业务的区域，可减少网络时延，提高访问速度。

★ 名称

使用该配置创建的云服务器名称为伸缩配置名称加八位随机码。

★ 配置模板

使用新模板

使用已有云服务器规格为模板

★ 规格

最新系列

vCPUs

全部

内存(GiB)

全部

规格名称

请输入规格名称

通用计算型

通用计算增强型

内存优化型

超大内存型

高性能计算型

磁盘增强型

超高IO型

GPU加速型

通用入门型

了解如何选择弹性云服务器类型

实例类型	规格名称	vCPUs 内存(GiB)	CPU	基准 / 最大带宽	内网收发包
☐	通用计算型s3	s3.small.1	1vCPUs 1GiB	Intel SkyLake	0.1/0.5 Gbit/s
☒	通用计算型s3	s3.medium.2	1vCPUs 2GiB	Intel SkyLake	0.1/0.5 Gbit/s
☐	通用计算型s3	s3.medium.4	1vCPUs 4GiB	Intel SkyLake	0.1/0.5 Gbit/s
☐	通用计算型s3	s3.large.2	2vCPUs 4GiB	Intel SkyLake	0.2/0.8 Gbit/s
☐	通用计算型s3	s3.large.4	2vCPUs 8GiB	Intel SkyLake	0.2/0.8 Gbit/s
☐	通用计算型s3	s3.xlarge.2	4vCPUs 8GiB	Intel SkyLake	0.4/1.5 Gbit/s
☐	通用计算型s3	s3.xlarge.4	4vCPUs 16GiB	Intel SkyLake	0.4/1.5 Gbit/s
☒	通用计算型s3	s3.2xlarge.2	8vCPUs 16GiB	Intel SkyLake	0.8/3 Gbit/s

已选规格 优先使用当前选中的规格进行伸缩，您可以点击已选规格查看规格信息。您还可以选择0个规格。

s3.2xlarge.2

s3.medium.2

通用计算型 | s3.2xlarge.2 | 8vCPUs | 16GiB

规格使用优先策略 ⓘ ☒ 选择优先 ☐ 成本优化

★ 镜像

公共镜像

私有镜像

共享镜像

市场镜像

★ 磁盘

云硬盘

系统盘

高IO

100

GiB

IOPS上限2,600, IOPS突发上限5,000 ⓘ

数据盘

高IO

100

GiB

IOPS上限2,600, IOPS突发上限5,000

数量

1

收起 ^

☐ 加密 ⓘ [用数据盘镜像创建磁盘 ⓘ](#)

ⓘ 增加一块数据盘 您还可以增加 22 块磁盘（云硬盘）。

★ 安全组

弹性伸缩安全组

新建安全组 ⓘ

安全组类似防火墙功能，是一个逻辑上的分组，用于设置网络访问控制。[如何配置安全组？](#)

入方向TCP | 出方向-

弹性公网IP

不使用

自动分配 ⓘ

不使用弹性公网IP的云服务器不能与互联网互通，仅可作为私有网络中部署业务或集群所需云服务器使用。

★ 登录方式

密钥对

密码

用户名

root

★ 密码

请牢记密码，如忘记密码可登录ECS控制台重置密码。

★ 确认密码

高级配置

暂不配置

现在配置

表 1-2 伸缩配置关键参数

参数	解释	取值样例
配置模板	选择“使用新模板”，重新选择云服务器类型、vCPUs、内存、镜像、磁盘等参数信息，创建新的弹性伸缩配置。	使用新模板
规格	可以选择多个规格，避免在伸缩时规格售罄的风险。规格使用优先策略包括“选择优先”和“成本优先”，请根据需要进行选择。	s3.medium.2 s3.large.2
镜像	为伸缩组中移入的实例提供软件和系统应用配置的模板，选择私有镜像discuz_centos6.5(40GB)。	私有镜像 discuz_centos6.5(40GB)
磁盘	为伸缩组中的移入的实例提供存储和存储管理功能。	系统盘 高IO 40GB 数据盘 高IO 100GB
安全组	安全组是一个逻辑上的分组，用来实现安全组内和组间弹性云服务器的访问控制，加强弹性云服务器的安全保护。选择安全组sg-DISCUZ。	sg-DISCUZ
弹性公网IP	伸缩组中已经添加了负载均衡后，伸缩配置可以不配置弹性公网IP。系统会自动将加入伸缩组的实例添加到负载均衡上，伸缩组中的实例统一通过负载均衡绑定的弹性公网IP对外提供服务。	不使用

3. 伸缩配置参数配置完成后，单击“立即创建”。

创建伸缩组

1. 单击“创建弹性伸缩组”。
- 参考表1-3进行关键参数配置，未列出的参数选择默认值即可。

图 1-2 设置伸缩组参数

<

创建弹性伸缩组

*

区域

华北-北京四

不同区域的云服务产品之间内网互不相通；请就近选择靠近您业务的区域，可减少网络时延，提高访问速度。

*

可用区

可用区1 × 可用区2 × 可用区3 ×

?

*

多可用区扩展策略

均衡分布

选择优先

①

*

名称

as-group-discuz

*

最大实例数(台)

50

*

期望实例数(台)

0

①

*

最小实例数(台)

0

选择伸缩配置作为您新建的伸缩组内伸缩实例的模板；选择子网后将向伸缩组中的每个实例分配IP地址。

*

伸缩配置

+

*

虚拟私有云

?

新建虚拟私有云

*

子网

本子网作为云服务器的主网卡

√

跨目的检查

①

⊕ 增加一个子网

您还可以增加 4 个子网

新建子网

负载均衡

不使用

使用弹性负载均衡

新建弹性负载均衡

伸缩组中的云服务器会自动挂载到您关联的负载均衡下。
将IPv6地址加入负载均衡后该服务器组需要保证实例具备IPv6地址，同时子网和负载均衡IPv6相关功能启用。

负载均衡器

后端云服务器组

后端端口

权重

1

IP协议版本

IPv4

* 新增一个负载均衡器

您还可以增加6个负载均衡器。

*

实例移除策略

根据较早创建的配置较早创建的实例

弹性公网IP

释放

不释放

若选择“释放”，在伸缩组进行缩的活动时，则会将云服务器上的弹性公网IP释放，否则仅做解绑定操作，保留弹性公网IP资源。

数据盘

删除

不删除

若选择“删除”，在伸缩组进行缩的活动时，则会将云服务器上的数据盘删除，否则仅做解绑定操作，保留数据盘资源。

*

健康检查方式

负载均衡健康检查

①

受保护的实例状态异常时，会被健康检查移除，并重新创建新的实例。
实例所在安全规则则需要配置放行100.125.0.0/16，并配置负载均衡用于健康检查的协议和端口，否则会导致健康检查失败。了解更多

*

健康检查间隔

5分钟

②

*

健康状况检查宽限期(秒)

600

②

*

企业项目

default

?

标签

如果您需要使用同一标签标识多种云资源，即所有服务均可在标签输入框下拉选择同一标签，建议在TMS中创建预定义标签。查看预定义

标签键

标签值

您还可以添加10个标签。

委托

--请选择--

?

新建委托

*

协议

√

我已经阅读并同意《弹性伸缩免责声明》

表 1-3 伸缩组关键参数

参数	解释	取值样例
最大实例数	伸缩组中弹性云服务器数量的最大值。	50
期望实例数	伸缩组中期望的云服务器数量，本实践中要将搭建Discuz!论坛的云服务器手动移入，为避免移入前发生伸缩活动，将期望实例数设置为0。	0
最小实例数	伸缩组中弹性云服务器数量的最小值。	0
虚拟私有云	为伸缩组中的实例提供所使用的网络。必须和云服务器discuz02属于同一VPC。	VPC-DISCUZ
子网	子网可以方便您管理vpc中的网络。选择中申请虚拟私有云时创建的子网。	vpc-test
负载均衡	为伸缩组中的实例均分流量，选择独享型负载均衡器elb-DISCUZ。后端端口配置为需要监听的业务端口，示例中配置为80，权重为1。	使用独享型
健康检查方式	健康检查方式选择“负载均衡健康检查”，负载均衡健康检查是通过系统向后端云服务器发起心跳检查的方式来实现的，推荐使用该方式。	负载均衡健康检查

2. 参数配置完后，单击“立即创建”。
3. 返回弹性伸缩组列表，若伸缩组为“已启用”状态，说明伸缩组创建成功。

创建伸缩策略

为了能实现云服务器的自动伸缩，配置两条监控CPU使用率的告警策略，在业务负载上升时增加云服务器数量的策略as-policy-discuz01，在业务负载降低时减少云服务器数量的策略as-policy-discuz02。

1. 在已创建的弹性伸缩组“as-group-discuz”所在行，单击操作列的“查看伸缩策略”。
2. 单击“添加伸缩策略”。

参考表1-4配置伸缩策略as-policy-discuz01的参数，当系统连续3次监控到CPU使用率超过70%时，触发伸缩策略as-policy-discuz01，伸缩组会增加一台弹性云服务器。

图 1-3 伸缩策略 as-policy-discuz01 参数

添加伸缩策略

策略名称

as-policy-discuz01

策略类型

告警策略

定时策略

周期策略

伸缩策略受告警规则状态影响，中途停用或处于停用状态下的告警规则会导致该伸缩策略失效。[查询已有告警状态](#)

告警规则

现在创建

使用已有

告警规则名称

as-alarm-cpu-01

监控类型

系统监控

自定义监控

触发条件

CPU使用率

最大值

>

70

%

如要使用Agent监控指标，请确认伸缩组中实例均已安装了Agent插件。[如何安装插件](#)
了解更多弹性伸缩监控指标信息，请点击[监控指标说明](#)
不同操作系统的监控指标有所不同。[了解更多](#)

监控周期

5分钟

连续出现次数

3

?

企业项目

default

?

告警规则所属企业项目，非实例所属企业项目。

告警策略类型

简单策略

区间策略

执行动作

增加

1

个实例

冷却时间(秒)

900

?

表 1-4 伸缩策略 as-policy-discuz01 关键参数

参数	解释	取值样例
策略名称	创建伸缩策略的名称。	as-policy-discuz01
策略类型	选择“告警策略”。	告警策略
告警规则	可选择“现在创建”或“使用已有”。	现在创建
告警规则名称	新建告警规则的名称。	as-alarm-cpu-01
监控类型	定义监控指标的类型，是系统支持的或是自定义的。选择“系统监控”。	系统监控

参数	解释	取值样例
触发条件	选择弹性伸缩支持的监控指标并对监控指标设定告警条件。	CPU使用率最大值 > 70%
监控周期	告警规则刷新告警状态的周期。	5分钟
连续出现次数	触发告警时的采样点数目。	3
执行动作	设置伸缩活动执行动作及实例的个数或实例百分比。 执行动作包括： ● 增加 当执行伸缩活动时，向伸缩组增加实例。 ● 减少 当执行伸缩活动时，从伸缩组中减少实例。 ● 设置为 将伸缩组中的期望实例数设置为固定值。	增加1个实例
冷却时间	为了避免告警策略频繁触发，必须设置冷却时间。	900

3. 单击“确定”。
4. 再次单击“添加伸缩策略”，配置伸缩策略as-policy-discuz02的参数，当系统连续3次监控到CPU使用率低于30%时，触发伸缩策略as-policy-discuz02，伸缩组会减少一台弹性云服务器。

图 1-4 伸缩策略 as-policy-discuz02 参数

添加伸缩策略

策略名称

as-policy-discuz02

策略类型

告警策略

定时策略

周期策略

伸缩策略受告警规则状态影响，中途停用或处于停用状态下的告警规则会导致该伸缩策略失效。[查询已有告警状态](#)

告警规则

现在创建

使用已有

告警规则名称

as-alarm-cpu-02

监控类型

系统监控

自定义监控

触发条件

CPU使用率

最小值

>

30

%

如要使用Agent监控指标，请确认伸缩组中实例均已安装了Agent插件。[如何安装插件](#)
了解更多弹性伸缩监控指标信息，请点击[监控指标说明](#)
不同操作系统的监控指标有所不同。[了解更多](#)

监控周期

5分钟

连续出现次数

3

企业项目

default

告警规则所属企业项目，非实例所属企业项目。

告警策略类型

简单策略

区间策略

执行动作

增加

1

个实例

冷却时间(秒)

900

5. 单击“确定”。
6. 返回伸缩策略列表页面，若伸缩策略为“已启用”状态，说明伸缩策略创建成功。

手动移入实例

手动将云服务器discuz02移入伸缩组。

1. 单击伸缩组as-group-discuz名称进入伸缩组详情页面。
2. 切换到“伸缩实例”页签，将discuz02手动移入伸缩组中。

修改最小实例数

为保证discuz02不被伸缩活动移出伸缩组，需修改伸缩组的最小实例数。

1. 单击伸缩组as-group-discuz名称，进入伸缩组详情页面。
2. 单击页面右上角的“修改伸缩组”。修改最小实例数为1。

图 1-5 修改最小实例数

★ 名称	as-group-discuz
★ 最大实例数(台)	50
★ 期望实例数(台)	1
★ 最小实例数(台)	1
★ 冷却时间(秒)	300
★ 可用区	可用区1 × 可用区2 × ▼

3. 修改完成后，单击“确定”。

结果验证

若论坛可以正常使用，当伸缩组中的云服务器CPU使用率持续高于70%（在伸缩组的“监控”页签可对监控指标进行观察），伸缩组会自动增加一台云服务器（在伸缩组的“活动历史”页签可对伸缩活动历史进行查看）。当伸缩组中的云服务器CPU使用率持续低于30%，且伸缩组中至少存在两台云服务器时，伸缩组会自动减少一台云服务器，则本次实践是成功的。若不然，请联系技术支持定位伸缩组不能正常进行伸缩活动的原因。

实践扩展

- 当应用场景有变化，需要在云服务器上部署新的软件时，可使用弹性伸缩的生命周期挂钩功能，在实例加入和移出伸缩组时进行自定义操作，灵活地管理加入或移出弹性伸缩组的实例。具体操作可参见[生命周期挂钩](#)。
- 当所需的弹性云服务器的规格变更时，可创建新的伸缩配置，操作可参考[使用新模板创建伸缩配置](#)。创建完成后，可参考[为伸缩组更换伸缩配置](#)，即可改变伸缩组新加入实例的规格。

2 扩缩容时为 ECS 实例自动绑定和释放 EIP

背景

- 伸缩组中扩容的ECS实例如果未分配公网IP地址时，则无法对外提供公网访问能力，您可以在扩容时为新的ECS实例自动绑定EIP以保持服务的可访问性。
- 伸缩组中的ECS实例如果已绑定了EIP但您需要及时释放ECS资源（例如业务进入低谷期），或不再需要公网访问能力，您可以在缩容时自动释放不再需要的EIP以优化资源使用并降低成本。

 说明

EIP是可以独立购买和持有公网IP地址资源，可以灵活与ECS实例进行绑定或释放。更多信息，请参见[什么是弹性公网 IP](#)。

准备工作

已创建伸缩组，且该伸缩组处于启用状态。具体操作请参见[创建伸缩组](#)。

扩容时为 ECS 实例自动绑定 EIP

- 修改伸缩配置。
 - 登录华为云控制台
 - 在控制台页面中选择“计算 > 弹性伸缩 > 伸缩实例”。
 - 在伸缩实例页面，单击“更多” > “修改伸缩配置”，按[表1](#)填写主要配置项信息。

 说明

本示例中采用以下配置，未提及的配置项可保持默认。更多关于伸缩配置的配置项说明，请参见使用[新模板创建伸缩配置](#)。

表 2-1 主要配置项信息

配置项	示例	说明
伸缩配置名称	as-config-test	输入伸缩配置的名称。格式要求请参考界面提示。

配置项	示例	说明
付费模式	按需计费	后付费模式，按弹性云服务器的实际使用时长计费，可以随时创建/删除弹性云服务器
规格	s3.small.1	实例规格最多不超过10个。 建议您按照vCPU、内存、内网带宽或者内网收发包等维度选择多个相近的实例规格。
选择镜像	test	选择用于部署ECS实例的镜像。 实际业务场景请选用您已部署应用的自定义镜像。
系统盘	通用型SSD 100GiB	为ECS实例选择系统盘。请根据业务需要进行设置。
安全组	sg-bp18kz60mefsicfg****	选择您提前创建好的安全组。如需创建安全组，请参见 创建安全组 。
弹性公网IP	自动分配	自动为每台弹性云服务器分配独享带宽的弹性公网IP，带宽值可以由您设定

- d. 单击“确定”。
2. 设置扩容策略，触发扩容活动，具体操作请参见[创建伸缩策略](#)。
3. 查看扩容的ECS实例是否已自动绑定EIP。
在伸缩组的伸缩实例列表中，单击自动创建的ECS实例对应的云服务器ID，进入ECS实例详情页，查看弹性公网IP，会显示已绑定的EIP。

缩容时为 ECS 实例自动释放 EIP

1. 修改伸缩组。
- a. 登录华为云控制台
- b. 在控制台页面中选择“计算 > 弹性伸缩 > 伸缩实例”。
- c. 在伸缩实例页面，单击“更多” > “修改伸缩组”，按[表2](#)填写主要配置项信息。

 说明

本示例采用以下配置，未提及的配置项保持默认。更多关于伸缩组的配置项说明，请参见[创建伸缩组](#)。

表 2-2 主要配置项信息

配置项	示例	说明
弹性公网 IP	释放	若选择“释放”，当进行缩容活动时，会将云服务器上的弹性公网IP释放，否则仅做解绑定操作，保留弹性公网IP资源。

d. 单击“确定”

- 2. 设置缩容策略，触发缩容活动。具体操作请参见[创建伸缩策略](#)。
- 3. 查看缩容的ECS实例是否已自动释放EIP。

您可以前往弹性公网IP列表，查看对应的EIP，如果EIP不存在，则说明已成功释放了EIP。

3 设置扩缩容策略和多实例规格降低成本

- 在创建伸缩组时，可配置多个可用区的扩缩容策略，确保某个可用区资源紧张或发生故障时，伸缩组可以在其他可用区进行资源调度，保证业务的高可用性。
- 在创建伸缩配置时，可设置多个实例规格，避免因单个实例规格库存不足而导致伸缩活动执行失败。

说明

• 扩缩容策略

弹性伸缩支持在创建伸缩配置时指定多个规格，在一个规格所在的可用区库存不足时，弹性伸缩会自动尝试在其他可用区创建实例，确保伸缩活动可以顺利执行。您可以根据业务部署情况配置对应的扩缩容策略（例如成本优化策略），灵活满足业务需求。使用成本优化策略的伸缩组会按规格单价从低到高尝试创建ECS实例。

• 多实例规格

弹性伸缩支持在伸缩配置中指定多个实例规格，在无法创建高优先级规格的实例时，弹性伸缩会自动尝试下一优先级规格的实例，直至成功创建实例。多实例规格可以有效应对单个实例规格库存不足的情况，保证伸缩活动可以顺利执行。例如，在业务高峰时，您可能需要争分夺秒地生成高配实例规格承载业务流量，更关注性能，而不局限于一种特定规格，此时多实例规格可以更灵活地应对不同场景的需求。

本文通过设置多可用区的应用成本优化策略和多实例规格，可有效提高弹性伸缩的扩缩容成功率，并降低成本。具体操作请参考本章节。

前提条件

已创建专有网络，具体操作请参见[创建虚拟私有云和子网](#)。

操作步骤

1. 创建伸缩配置

- a. 登录华为云控制台。
- b. 在控制台页面中选择“计算 > 弹性伸缩 > 伸缩实例”。
- c. 单击“创建伸缩配置”，填写配置项信息。

说明

本示例中采用以下配置，未提及的配置项可保持默认。更多关于伸缩配置的配置项说明，请参见使用[新模板创建伸缩配置](#)。

表 3-1 主要配置项信息

配置项	示例	说明
伸缩配置名称	as-config-test	输入伸缩配置的名称。
付费模式	按需计费	按弹性云服务器器的实际使用时长计费，可以随时创建/删除弹性云服务器。
规格	s3.small.1， s3.medium.2，s3.large.2 (每种实例规格单价依次递增)	选择多个实例规格（最多不超过10个）。 建议您按照vCPU、内存、内网带宽或者内网收发包等维度选择多个相近的实例规格。
规格使用 优先策略	成本优化	使用成本优化策略的伸缩组会按规格单价从低到高尝试创建ECS实例。
选择镜像	test	选择用于部署ECS实例的镜像。 实际业务场景请选用您已部署应用的自定义镜像。
系统盘	通用型SSD 100GiB	为ECS实例选择系统盘。请根据业务需要进行设置。
安全组	sg-bp18kz60mefsicfg****	选择您提前创建好的安全组。如需创建安全组，请参见 创建安全组 。

- d. 单击“立即创建”。
2. 创建伸缩组。

a. 在伸缩实例页面，单击“创建弹性伸缩组”。

b. 填写配置项信息。

 说明

本示例采用以下配置，未提及的配置项保持默认。更多关于伸缩组的配置说明，请参见[创建伸缩组](#)。

表 3-2 主要配置项信息

配置项	示例	说明
伸缩组名称	test	输入伸缩组名称。格式要求请参考界面提示。
可用区	可用区1、可用区2、可用区3	选择ECS想要分布的可用区。

配置项	示例	说明
多可用区扩展策略	均衡分布	均衡分布：云服务器扩容时优先保证选择的可用区列表中各可用区下云服务器数量均衡，当无法在目标可用区下完成云服务器扩容时，按照选择优先原则选择其他可用区。
组内最小实例数	0	设置伸缩组内实例数量的下限。当前实例数量低于下限时，伸缩组会自动添加实例，使得伸缩组内的实例数量等于下限。
组内最大实例数	120	设置伸缩组内实例数量的上限。当前实例数量超过上限时，伸缩组会自动移出实例，使得伸缩组内的实例数量等于上限。
期望实例数	0	期望实例数是指伸缩组中期望的云服务器数量。创建后可以手工修改该值，修改该值就会触发一次弹性伸缩活动。期望实例数必须大于或等于最小实例数，必须小于或等于最大实例数。
伸缩配置	as-config-test	为伸缩组选择所需的伸缩配置。
虚拟私有云	Vpc-test	同一伸缩组内的弹性云服务器均属于该VPC
子网	Subnet-test	您最多可以选择五个子网，伸缩组会自动为创建的实例绑定所有网卡。您选择的第一个子网默认作为云服务器的主网卡，其它子网作为云服务器的扩展网卡。
健康检查方式	云服务器健康检查	对云服务器的运行状态进行检查，如关机、删除都是云服务器异常状态。
健康检查间隔	5分钟	伸缩组执行健康检查的周期。您可以根据实际情况设置合理的健康检查间隔（10秒、1分钟、5分钟、15分钟、1小时、3小时）。

- c. 参数配置完后，单击“立即创建”。

d. 返回弹性伸缩组列表，若伸缩组为“已启用”状态，说明伸缩组创建成功。
3. 设置扩缩容策略，触发扩容伸缩活动，具体操作请参见[创建伸缩策略](#)。

效果验证

- 期望效果：**当伸缩组发生扩容伸缩活动时，伸缩组会优先按s3.small.1创建ECS实例，如果在可用1、可用区2、可用区3下均无法创建出实例，则依次尝试s3.medium.2，s3.large.2。

- **效果验证：**执行伸缩规则触发伸缩活动后，向伸缩组添加1台ECS实例。在伸缩实例页签下，单击自动创建的ECS实例名称，跳转到云服务器控制台的实例详情页签下，可查看实例规格。若与预期结果一样，优先使用s3.small.1成功创建了ECS实例，由此可见，弹性伸缩设置扩缩容策略和多实例规格后，可以明显降低成本。

4 使用均衡分布策略部署高可用计算集群

应用场景

对于分布式大数据计算、人工智能训练等场景，在业务高峰期可能会临时需要大量的计算资源，需要快速创建一个高可用的计算集群。弹性伸缩可实现自动批量创建ECS实例，同时使用均衡分布策略自动将ECS实例均衡分散在多个可用区。并实时检测ECS实例的运行状况，确保集群的高可用性。

方案优势

- 零运维成本
使用弹性伸缩可以实现自动扩缩容ECS实例，整个过程无需人工干预。
- 天然高可用
均衡分布策略可以实现自动分散部署ECS实例，避免单一可用区因库存不足等原因导致扩容失败，同时默认开启的实例健康检查功能可以确保伸缩组内的ECS实例均处于可用状态。

本教程介绍如何使用弹性伸缩实现均衡分布ECS实例，部署高可用计算集群。

步骤一：创建伸缩配置

须知

- 请根据您的业务架构评估业务模块，为需要部署高可用集群的业务模块创建伸缩组。
- 请提前准备好已部署业务应用的自定义镜像，并在配置伸缩配置时选择对应的自定义镜像，确保自动创建出的ECS实例符合要求。具体操作，请参见[通过云服务器创建整机镜像](#)。

1. 登录管理控制台。
2. 选择“计算 > 弹性伸缩 > 伸缩实例”。
3. 单击“创建伸缩配置”，填写配置项信息。

说明

本示例中采用以下配置，未提及的配置项保持默认。更多关于伸缩配置的配置项说明，请参见使用[新模板创建伸缩配置](#)。

表 4-1 主要配置项信息

配置项	示例	说明
伸缩配置名称	as-config-test	输入伸缩配置的名称。格式要求请参考界面提示。
付费模式	按需计费	后付费模式，按弹性云服务器器的实际使用时长计费，可以随时创建/删除弹性云服务器
规格	s3.small.1	设置ECS实例的vCPU和内存需求。
选择镜像	test	选择用于部署ECS实例的镜像。实际业务场景请选用您已部署应用的自定义镜像。
系统盘	通用型SSD 100GiB	为ECS实例选择系统盘。请根据业务需要进行设置。
安全组	sg-bp18kz60mefsicfg****	选择您提前创建好的安全组。如需创建安全组，请参见 创建安全组 。

4. 单击“立即创建”。

步骤二：创建弹性伸缩组

- 1. 在伸缩实例页面，单击“创建弹性伸缩组”。
- 2. 填写配置项信息。

说明

本示例采用以下配置，未提及的配置项保持默认。更多关于伸缩组的配置说明，请参见[创建伸缩组](#)。

表 4-2 主要配置项信息

配置项	示例	说明
伸缩组名称	test	输入伸缩组名称。格式要求请参考界面提示。
可用区	可用区1可用区2可用区3	选择ECS想要分布的可用区。
多可用区扩展策略	均衡分布	当选择两个及以上可用区时，才需要配置该选项。
组内最小实例数	10	设置伸缩组内实例数量的下限。当前实例数量低于下限时，伸缩组会自动添加实例，使得伸缩组内的实例数量等于下限。

配置项	示例	说明
组内最大实例数	120	设置伸缩组内实例数量的上限。当前实例数量超过上限时，伸缩组会自动移出实例，使得伸缩组内的实例数量等于上限。
期望实例数	10	期望实例数是指伸缩组中期望的云服务器数量。创建后可以手工修改该值，修改该值就会触发一次弹性伸缩活动。期望实例数必须大于或等于最小实例数，必须小于或等于最大实例数。
伸缩配置	as-config-test	为伸缩组选择所需的伸缩配置。
虚拟私有云	Vpc-test	同一伸缩组内的弹性云服务器均属于该VPC。
子网	Subnet-test	您最多可以选择五个子网，伸缩组会自动为创建的实例绑定所有网卡。 您选择的第一个子网默认作为云服务器的主网卡，其它子网作为云服务器的扩展网卡。
健康检查方式	云服务器健康检查	对云服务器的运行状态进行检查，如关机、删除都是云服务器异常状态
健康检查间隔	5分钟	伸缩组执行健康检查的周期。您可以根据实际情况设置合理的健康检查间隔（10秒、1分钟、5分钟、15分钟、1小时、3小时）。

3. 参数配置完后，单击“立即创建”。
4. 返回弹性伸缩组列表，若伸缩组为“已启用”状态，说明伸缩组创建成功。

执行结果

由于伸缩组的最小实例数配置为10，因此启用伸缩组后，伸缩组会自动在所选可用区中，均衡部署10台ECS实例来组成计算集群，并从以下方面来保证集群的高可用性，降低资源成本。

- 当单可用区中因库存不足等原因引发问题时，会自动尝试在其他可用区中部署ECS实例，对业务应用的影响有限。
- 伸缩组会自动移出不健康状态的ECS实例并创建新的ECS实例。

5 使用标签鉴权精细化管理弹性伸缩

背景

通过标签和访问控制IAM结合，并将标签作为权限策略的匹配条件，您可以实现对弹性伸缩的精细化管理。

- 标签是云资源的标识，可以帮助您从不同维度对具有相同特征的云资源进行分类、搜索和聚合。
- IAM提供权限管理能力，可基于权限策略，管理用户身份，控制云资源的访问和操作。更多信息，请参见[什么IAM](#)。

本文为您介绍如何通过标签鉴权控制IAM用户的权限，实现不同用户能够拥有不同的访问和控制权限，提高管理效率，同时降低信息泄露风险。

示例场景

假设已创建2个伸缩组，且各伸缩组标签（Tag）包含项目维度，即team，详情如下：

表 5-1 伸缩组及伸缩组标签

伸缩组	伸缩组名称	伸缩组标签
伸缩组1	asg-001	标签：team:game1，其中标签键为team，标签值为game1。
伸缩组2	asg-002	标签：team:game2，其中标签键为team，标签值为game2。

您可以通过标签鉴权方式控制某IAM用户对以上伸缩组拥有不同的特定权限，示例场景如下：

- **场景一：**不允许创建不带标签的伸缩组1。创建伸缩组1时须为伸缩组1绑定team:game1标签后，才可以创建成功。
- **场景二：**只允许查询绑定team:game1标签的伸缩组1。
- **场景三：**只允许操作伸缩组1，不允许操作伸缩组2。伸缩组1绑定了team:game1标签，伸缩组2绑定了team:game2标签，只允许操作绑定team:game1标签的伸缩组功能，不允许操作绑定team:game2标签的伸缩组功能。

操作步骤

须知

操作前请您确保已创建IAM用户且赋予ECS、VPC、IMS的只读权限。若未创建请您先创建IAM用户，具体操作，请参见[创建IAM用户](#)。

- 1. 创建2个伸缩组。具体操作，请参见[创建伸缩组](#)。
- 2. 登录IAM控制台。
- 3. 创建自定义策略。具体操作，请参见[创建自定义策略](#)。

您可以在策略（Condition）中为云资源设置多个标签条件来限制对弹性伸缩资源的操作权限。支持的标签鉴权条件如下所示：

表 5-2 标签鉴权条件

标签鉴权条件	说明
g:RequestTag	限制在请求中必须传入特定的标签。 如果API请求中没有标签参数，则不能使用g:RequestTag，否则会导致鉴权失败。
g:ResourceTag	限制指定的资源必须包含特定的标签。 如果API请求中没有资源ID参数，则不能使用g:ResourceTag，否则会导致鉴权失败。

- **场景一：不允许创建不带标签的伸缩组1。**
创建伸缩组1时须为伸缩组1绑定team:game1标签后，才可以创建成功。
对应的自定义权限策略如下所示：

```
{
  "Version": "5.0",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "as:scalingGroup:create"
      ],
      "Condition": {
        "StringEquals": {
          "g:RequestTag/team": [
            "game1"
          ]
        }
      }
    },
    {
      "Effect": "Allow",
      "Action": [
        "as:scalingConfig:list"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "as:listQuotas"
      ]
    }
  ]
}
```

```
]
}
]
}
```

- **场景二：只允许查询绑定标签的伸缩组1。**

伸缩组1绑定team:game1标签后，查询伸缩组时只能查询到伸缩组1的资源。

对应的自定义权限策略如下所示：

```
{
  "Version": "5.0",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "as:scalingGroup:get"
      ],
      "Condition": {
        "StringEquals": {
          "g:ResourceTag/team": [
            "game1"
          ]
        }
      }
    },
    {
      "Effect": "Allow",
      "Action": [
        "as:scalingConfig:list"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "as:listQuotas"
      ]
    }
  ]
}
```

- **场景三：只允许操作伸缩组1，不允许操作伸缩组2。**

伸缩组1绑定了team:game1标签，伸缩组2绑定team:game2标签，只允许操作绑定team:game1标签的伸缩组功能，不允许操作绑定了team:game2标签的伸缩组功能。

对应的自定义权限策略如下所示：

```
{
  "Version": "5.0",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "as:scalingConfig:list"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "as:listQuotas"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "as:scalingGroup:update"
      ]
    }
  ],
}
```

```
{
  "Effect": "Allow",
  "Action": [
    "as:scalingGroup:resume",
    "as:scalingGroup:pause"
  ],
  "Condition": {
    "StringEquals": {
      "g:ResourceTag/team": [
        "game1"
      ]
    }
  }
}
```

4. 将自定义策略授权给您希望控制访问的IAM用户。具体操作，请参见[给IAM用户授权](#)。

效果验证

验证权限策略是否生效。

- **场景一：不允许创建不带标签的伸缩组1。**
伸缩组1已绑定了标签team:game1，创建伸缩组1可成功创建。若没有设置标签team:game1或者设置了其他标签，则提示无权限创建。
- **场景二：只允许查询绑定标签的伸缩组1。**
指定查询伸缩组1（已绑定标签team:game1），但查询时未筛选标签选项，则可以查询到该伸缩组信息。
指定查询除伸缩组1外的某个未绑定标签team:game1的伸缩组，则查询结果为空。
- **场景三：只允许操作伸缩组1，不允许操作伸缩组2。**
修改的伸缩组1（已绑定标签team:game1），有修改伸缩组的权限。
修改的伸缩组2（未绑定标签team:game1）或设置了其他标签，则提示无权限修改该伸缩组。