计算云服务 - 弹性伸缩

前言

- 本章主要介绍华为云计算服务产品中的弹性伸缩服务。介绍了弹性伸缩服务的基本功能、应用场景及使用方法。

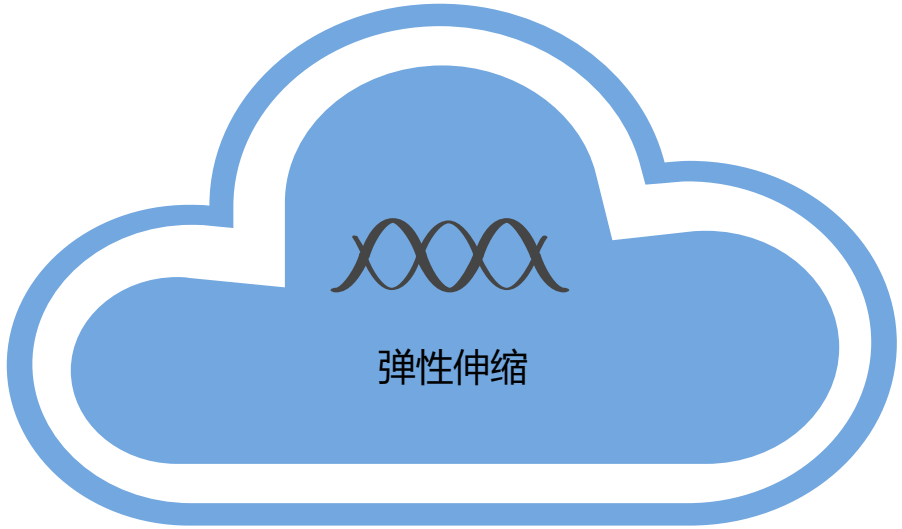
目标

- 学完本课程后，您将能够：
  - 熟悉弹性伸缩服务的相关概念、主要功能和应用场景
  - 掌握弹性伸缩组和伸缩带宽的创建和管理方法

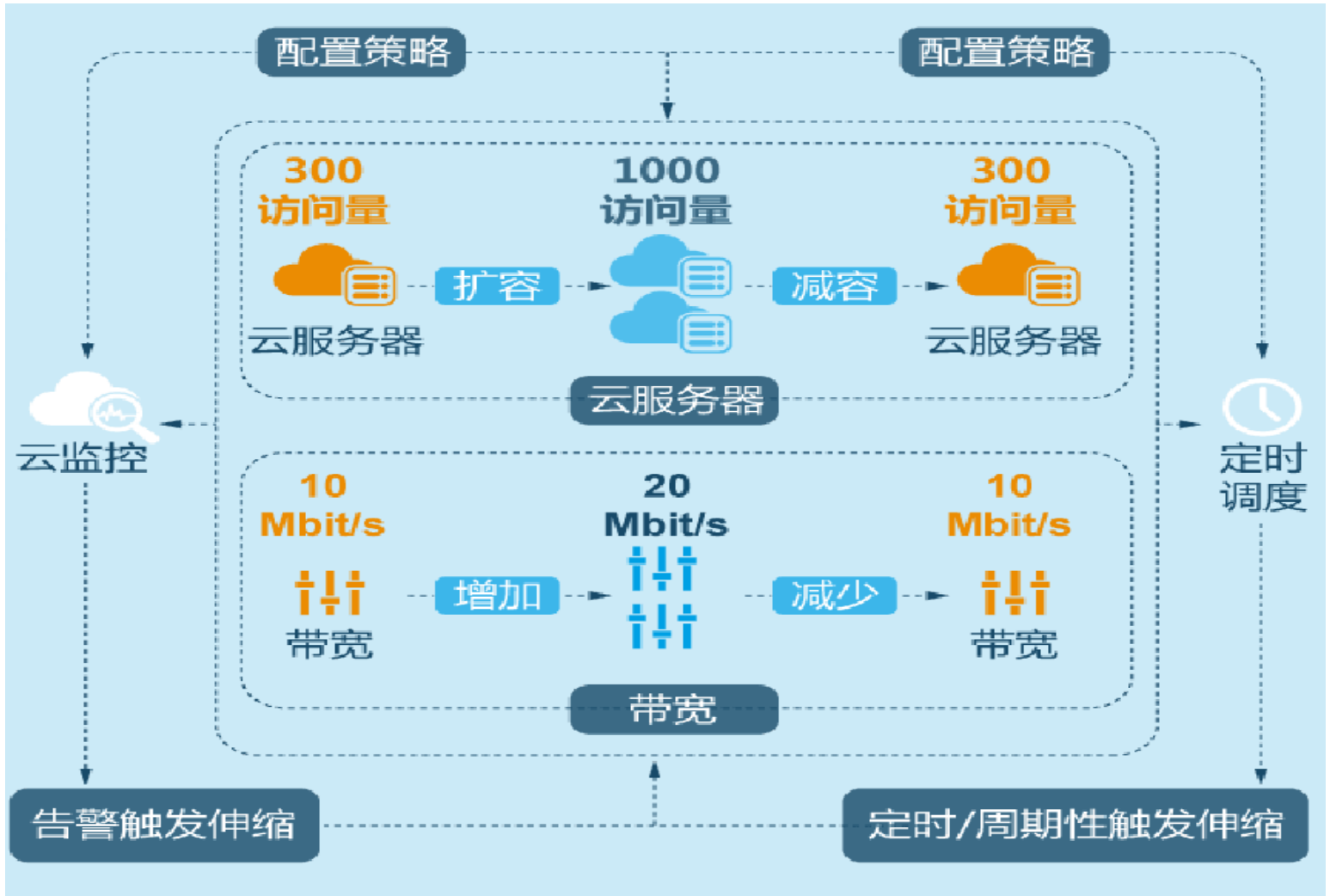
目录

AS概念

- 弹性伸缩（Auto Scaling）是根据用户的业务需求，通过策略自动调整其业务资源的服务。您可以根据业务需求自行定义伸缩配置和伸缩策略，降低人为反复调整资源以应对业务变化和高峰压力的工作量，帮助您节约资源和人力成本。

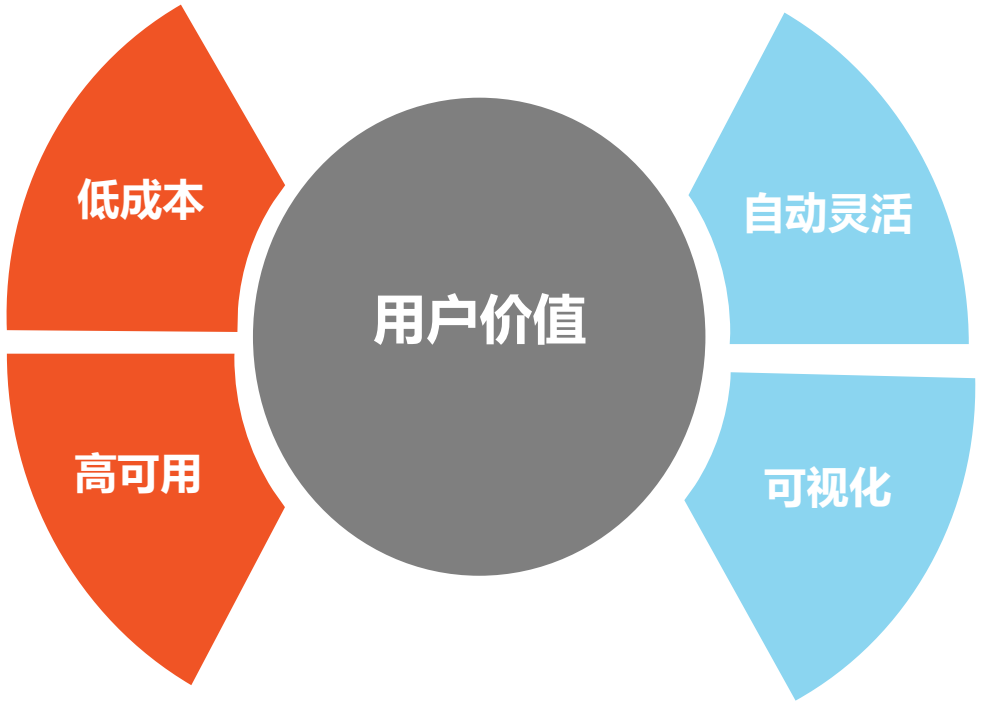


AS产品架构



AS功能特性

AS产品优势



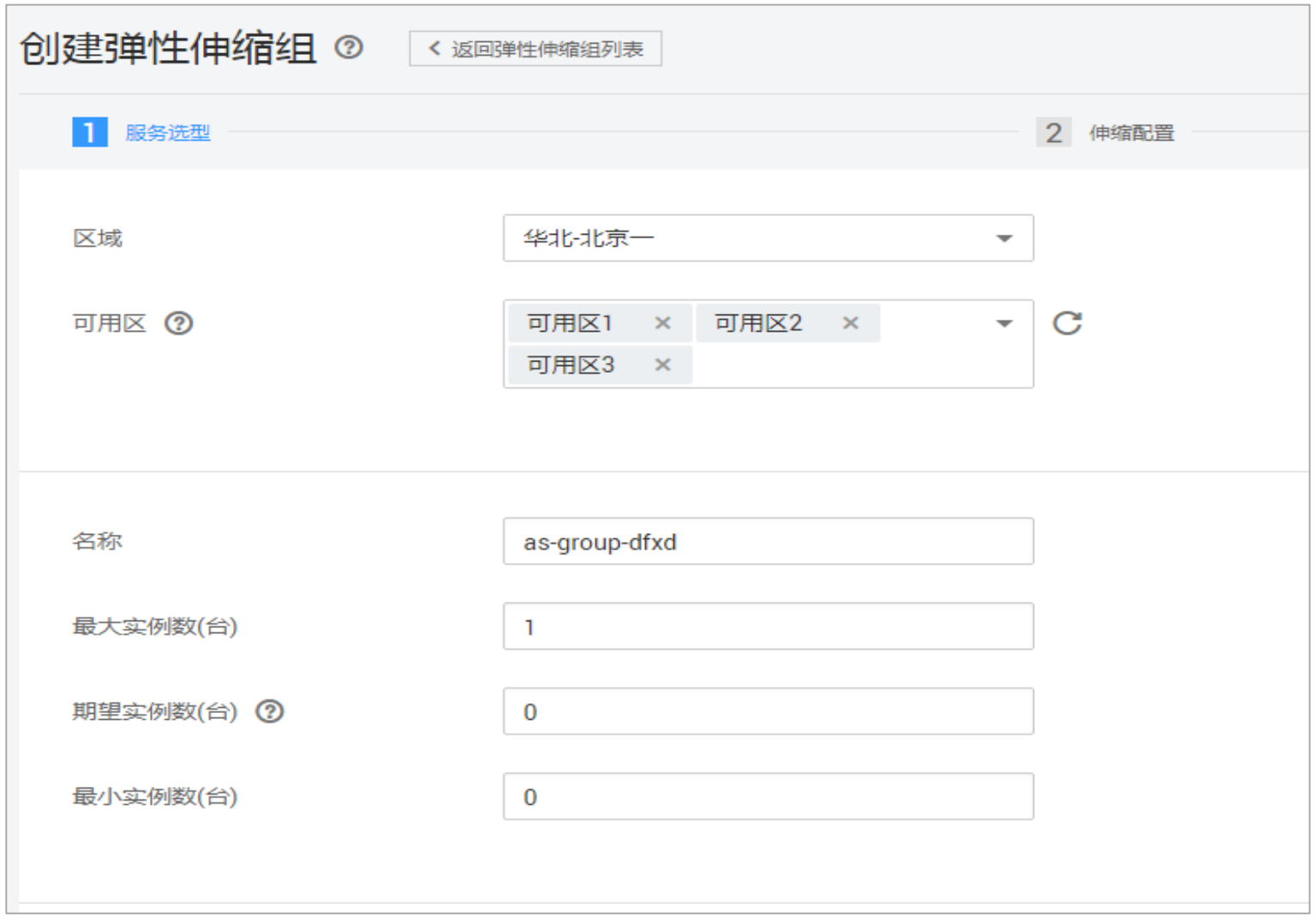
AS应用场景

典型应用场景	场景描述
Web应用服务	常见 Web 服务的逻辑层服务器扩缩容。 如企业网站、电商、视频网站、在线教育、移动应用等，客户端的请求通过负载均衡到达应用服务器。当访问量快速变化时，弹性伸缩服务可根据请求量弹性扩缩应用服务器的数量。若您使用了伸缩带宽功能，弹性伸缩服务也可根据访问流量多少自动调整IP公网带宽大小。
高性能计算集群部署	常见 Web 服务的分布式后台扩缩容。 如分布式大数据计算的计算节点、数据检索服务器等后端计算集群，根据计算量大小实时调整集群服务器数量。
请求类服务器部署	用于发送请求或收集数据的服务器集群的部署。 此类服务有明显的时效性，可依靠弹性伸缩服务快速完成请求服务器的创建部署和容量的扩大或缩小。

目录

伸缩组向导式创建流程式

配置参数 – 创建伸缩组_1

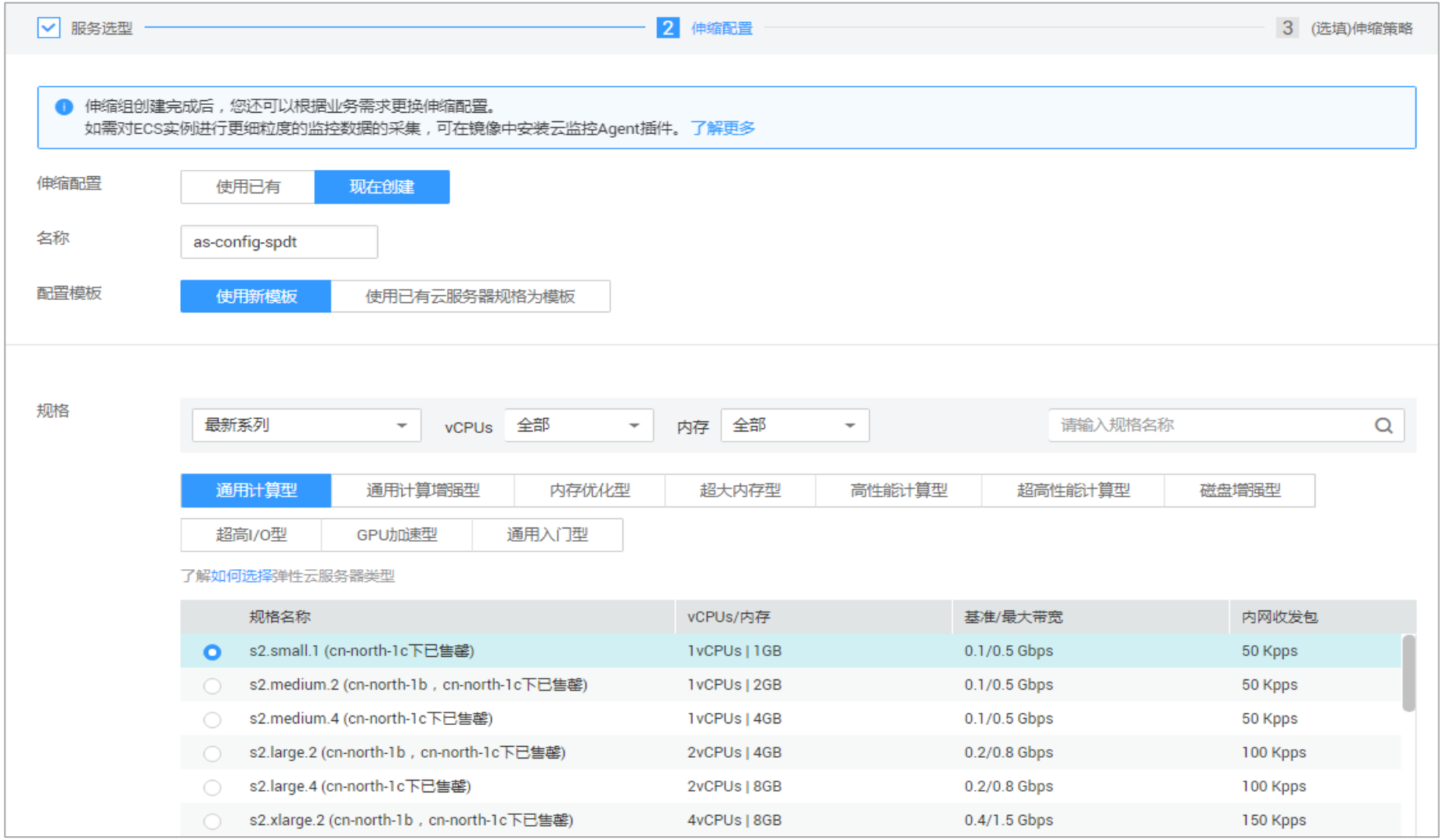


配置参数 – 创建伸缩组_2

参数说明 – 创建伸缩组参数

参数	解释	取值样例
最大/最小实例数	最大/最小实例数是指伸缩组中云服务器个数的最大值/最小值。	10/5台
期望实例数	期望实例数是指伸缩组中期望的云服务器数量。	6台
可用分区	指在同一地域下，电力、网络隔离的物理区域，可用分区之内内网互通，不同可用分区之间物理隔离。	-
VPC	弹性云服务器使用的网络是虚拟私有云（VPC）提供的。同一伸缩组内的弹性云服务器均属于该VPC。	-
子网	默认情况下，一个VPC子网内的弹性云服务器均可以进行通信，不同VPC子网内的弹性云服务器不能进行通信。	-
安全组	用户可以在安全组中定义各种访问规则，当弹性云服务器加入该安全组后，即受到这些访问规则的保护。	-
负载均衡	可选参数。选择使用负载均衡器后，访问流量将自动分发到伸缩组内的所有弹性云服务器，扩展应用系统对外的服务能力，实现更高水平的应用程序容错性能。	-
健康检查方式	健康检查会将异常的云服务器从伸缩组中移除，并重新创建新的云服务器，伸缩组的健康检查方式包括以下两种：云服务器健康检查、弹性负载均衡健康检查。	-
健康检查间隔	伸缩组执行健康检查的周期。	5分钟
实例移除策略	实例优先被移除的策略。当满足条件时，会触发实例移除活动。	-
移除实例时是否释放弹性IP	若伸缩组的伸缩配置使用了弹性IP，在进行伸的活动时，会给创建出来的云服务器绑定一个弹性IP。若勾选“是”，当进行缩的活动时，会将云服务器上的弹性IP释放，否则仅做解绑定操作，保留弹性IP资源。	-

配置参数 – 创建伸缩配置_1



配置参数 – 创建伸缩配置_2

配置参数 – 创建伸缩配置参数

参数	解释	取值样例
配置名称	创建伸缩配置的名称。	-
配置模板	选择“使用新模板” 重新选择云服务器类型、vCPU、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	公有云提供了多种类型的弹性云服务器，针对不同的应用场景，可以选择不同规格的弹性云服务器。	内存优化型
镜像	公共镜像、私有镜像、共享镜像	公共镜像
磁盘	也称云硬盘，包括系统盘和数据盘。 硬盘类型：普通IO、高IO、超高IO	“系统盘”选为“普通IO”
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，加强弹性云服务器云主机的安全保护。	-
弹性公网IP	弹性公网IP是指将公网IP地址和路由网络中关联的弹性云服务器绑定，以实现虚拟私有云内的弹性云服务器通过固定的公网IP地址对外提供访问服务。根据实际情况选择以下两种方式： 不使用：弹性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者集群所需弹性云服务器进行使用。 自动分配：自动为每台弹性云服务器分配独享带宽的弹性IP，带宽值可以由您设定。	自动分配
登录方式	使用密钥对作为弹性云服务器的鉴权方式，请在密钥对页面先创建或导入密钥对。	秘钥对
高级配置	高级配置可对文件注入、用户数据注入和云服务器组进行配置。可选择“暂不配置”和“现在配置”。	-

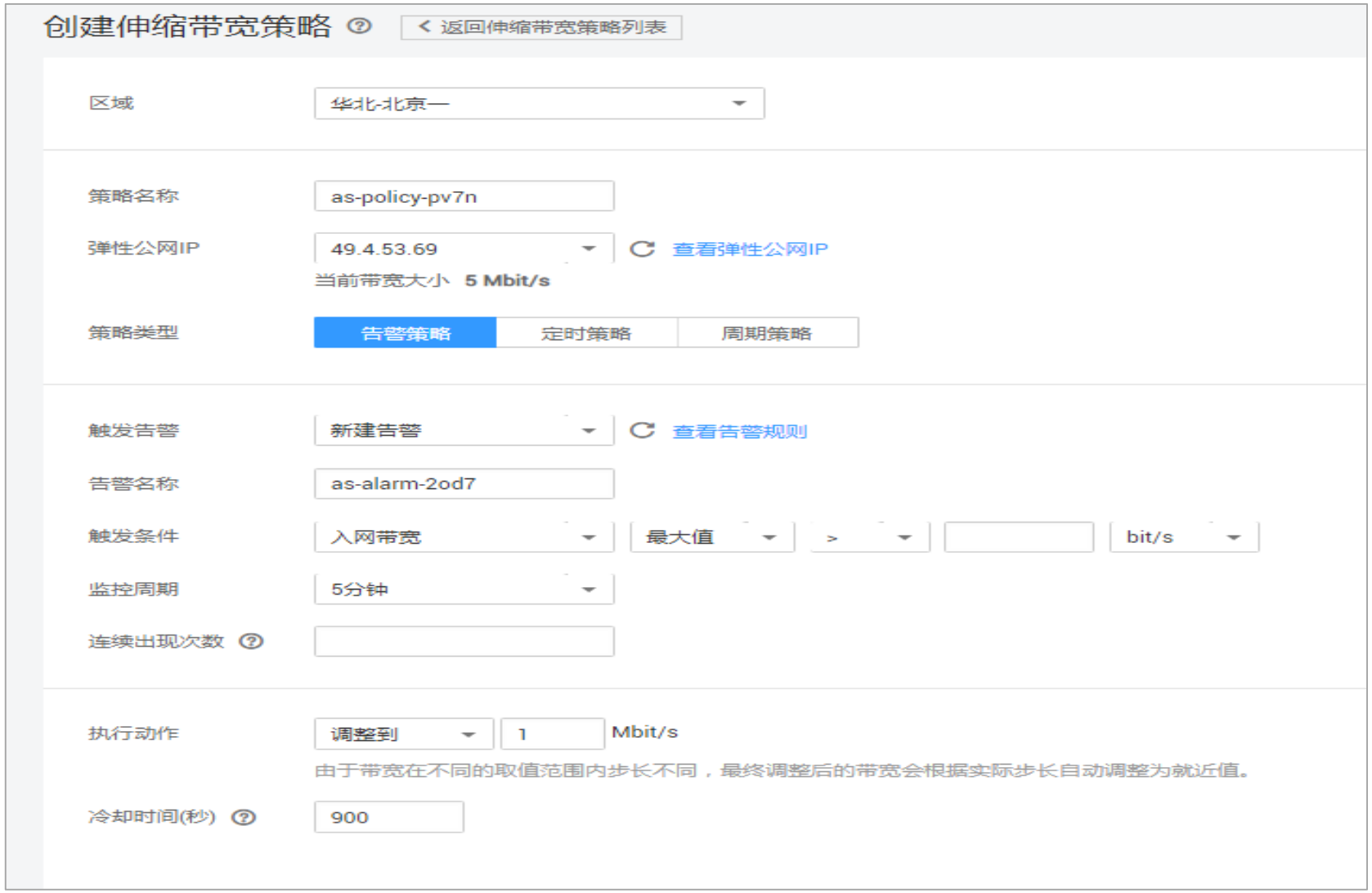
配置参数 – 添加伸缩策略

配置参数 – 伸缩策略参数说明

参数	解释	取值样例
策略名称	创建伸缩策略的名称。	as-policy-p6g5
策略类型	告警策略、定时策略、周期策略	告警策略
监控类型	设置告警的监控类型，可选择系统监控或自定义监控。	系统监控
触发告警	可选择已有的告警和新建告警。若选择新建告警，需配置如下参数： 告警名称：新建告警的名称，例如as-alarm-7o1u。 触发条件：选择弹性伸缩支持的监控指标及对监控质保设定的条件，例如CPU利用率最大值>70%。 监控周期：设定对弹性伸缩支持的监控指标监控的周期，例如5分钟。 连续出现次数：在监控周期内，连续达到触发条件几次后，开始执行伸缩活动，例如1次。	-
执行动作	设置伸缩活动执行动作及实例的个数。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	冷却时间是指冷却伸缩活动的时间，在每次伸缩活动完成之后，系统开始计算冷却时间。 伸缩组在冷却时间内，会拒绝由告警策略触发的伸缩活动，其他类型的伸缩策略（如定时策略和周期策略等）触发的伸缩活动不受限制，但会重新开始计算冷却时间，单位为秒。	900秒

目录

配置参数 – 创建伸缩带宽策略



配置参数 – 伸缩带宽策略参数说明

参数	解释	取值样例
区域	创建的伸缩组所在的区域。	-
策略名称	创建伸缩带宽策略的名称。	-
弹性公网IP	需要进行伸缩管理的公网IP。	-
策略类型	告警策略、定时策略、周期策略。	告警策略
触发告警	可选择已有告警和新建告警。 若选择新建告警，需配置如下参数：告警名称 新建告警规则的名称，例如as-alarm-7o1u。 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，例如CPU利用率最大值>70%。 监控周期 告警规则刷新告警状态的周期，例如5分钟。 连续出现次数 触发告警时的采样点数目。	新建告警
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 减少 设置为	增加一个实例
限制值	设置带宽可自动调整的上限和下限，单位Mbit/s。	2000Mbit/s
冷却时间	冷却时间是指冷却伸缩活动的时间，在每次伸缩活动完成之后，系统开始计算冷却时间。在冷却时间内，会拒绝由告警策略触发的伸缩活动，其他类型的伸缩策略（如定时策略和周期策略）触发的伸缩活动不受限制，但会重新开始计算冷却时间，单位为秒。	900秒

目录

AS管理概览

- 伸缩组
- 伸缩配置
- 伸缩活动
- 伸缩带宽
- 伸缩组和实例的监控
- 使用限制

伸缩组

- 伸缩组是具有相同属性和应用场景的云服务器和伸缩策略的集合，是启停伸缩策略和进行伸缩活动的基本单位。
  - 创建伸缩组
  - 添加负载均衡器到伸缩组
  - 为伸缩组添加/更换伸缩配置
  - 启用伸缩组
  - 停用伸缩组
  - 修改伸缩组
  - 删除伸缩组

伸缩配置

- 伸缩配置用于定义伸缩组内待添加的云服务器的规格数据，也就是定义了资源扩展时的云服务器的规格。
  - 使用已有云服务器创建伸缩配置
  - 使用新模板创建伸缩配置
  - 复制伸缩配置
  - 删除伸缩配置

伸缩活动 - 资源扩展

- 当业务需求增大时，需要通过伸缩活动实现资源扩展。
- 资源扩展方式：
  - 动态扩展资源
  - 按计划扩展资源
  - 手动扩展资源

伸缩活动 – 实例移出策略

- 当您的伸缩组自动移除实例时，可以根据下面策略进行实例移除：
  - 根据较早创建的配置较早创建的实例
  - 根据较早创建的配置较晚创建的实例
  - 较早创建的实例
  - 较晚创建的实例

伸缩活动 – 伸缩活动的查询

- 在伸缩组基本信息页面中，在“监控”页签中，可通过选择“图形”和“表格”两种方式查看伸缩活动的日志。以下截图为“图形” 的方式。

伸缩活动 – 生命周期挂钩

- 添加生命周期挂钩后，当伸缩组进行伸缩活动时，正在加入或正在移出伸缩组的实例将被挂钩挂起并置于等待状态，您能够在实例保持等待状态的时间内执行自定义操作。例如，您可以在新启动的实例上安装或配置软件，也可以在实例终止前从实例中下载日志文件。
  - 添加挂钩
  - 修改挂钩
  - 删除挂钩
  - 进行回调操作

伸缩活动 – 管理伸缩策略

- 伸缩策略是触发伸缩活动的条件和执行的动作，当满足条件时，会触发一次伸缩活动。AS支持对伸缩策略进行以下操作。
  - 创建伸缩策略
  - 修改伸缩策略
  - 删除伸缩策略
  - 启用伸缩策略
  - 停用伸缩策略
  - 立即执行伸缩策略

伸缩组和实例的监控

健康检查会将异常的实例从伸缩组中移除，伸缩组会重新创建新的实例以维持伸缩组的期望实例数和当前实例数保持一致，伸缩组的健康检查方式主要包括以下两种。

- 云服务器健康检查：是指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。伸缩组会自动将异常状态的云服务器移出伸缩组。
- 弹性负载均衡健康检查：是指根据ELB对服务器的健康检查结果进行的检查。在您将多个弹性负载均衡器添加到伸缩组时，只要有一个负载均衡器检测到云服务器状态异常，伸缩组会将该云服务器移出伸缩组。

使用限制

在向应用系统中添加弹性伸缩后，使用限制如下所示：

- 弹性伸缩的云服务器中运行的应用需要是无状态、可横向扩展的。
- 弹性伸缩对用户的资源数量或容量做的配额限制如下表所示。

类别	描述	默认值
弹性伸缩组	用户可以创建的最多伸缩组个数。	10
弹性伸缩配置	用户可以创建的最多伸缩配置个数。	100
弹性伸缩策略	某个弹性伸缩组下可以创建的最多伸缩策略个数。	10
弹性伸缩实例	某个弹性伸缩组下可以创建的最多实例个数。	300
伸缩带宽策略	用户最多可以创建的伸缩带宽策略个数。	50

目录

与AS关联的服务

- 弹性云服务器服务
- 虚拟私有云服务
- 弹性负载均衡服务
- 消息通知服务
- 云审计服务
- 云监控服务

思考题

1. 弹性伸缩服务中实现资源扩展有哪些方式?

   A. 动态扩展资源

   B. 计划扩展资源

   C. 手工扩展资源

   D. 自动扩展资源

2. 弹性伸缩支持哪几种伸缩策略?

   A. 告警策略

   B. 定时策略

   C. 周期策略

   D. 监控策略

本章总结

- 讲解了AS的概念、主要功能和应用场景。
- 讲解了AS伸缩组和伸缩带宽的创建和管理。

缩略语

缩写	全称
AS	Auto Scaling
ELB	Elastic Load Balace


谢 谢
www.huawei.com