弹性伸缩 使用教程

产品版本：ZStack 3.10.0

文档版本：V3.10.0

ZStack

版权声明

商标说明

ZStack商标和其他云轴科技商标均为上海云轴信息科技有限公司的商标。

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、服务或特性等应受云轴科技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，云轴科技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 概述

ZStack提供基于负载均衡的云主机弹性伸缩，可根据用户业务的负载变化，按照预定义的策略，自动调整伸缩组内云主机的数量，提高云平台资源的使用效率，降低运维成本，保证业务平稳运行。目前支持KVM云主机的弹性伸缩。

伸缩模式

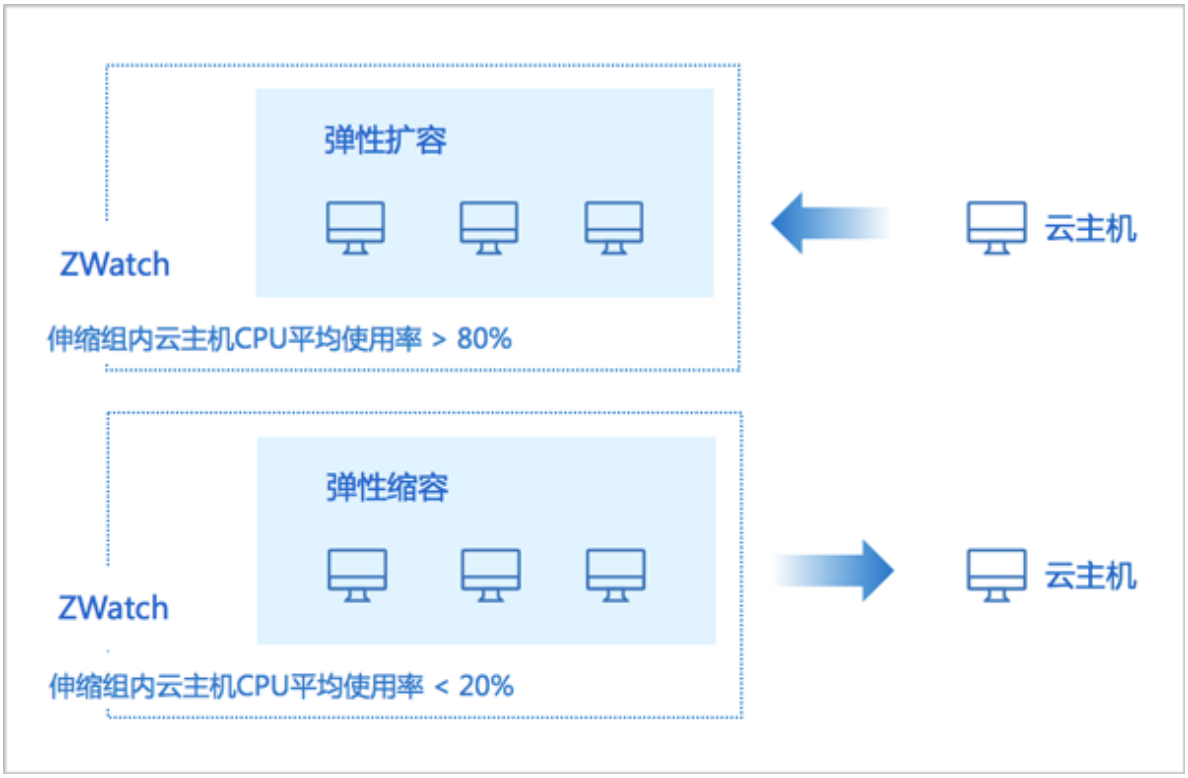
支持以下两种伸缩模式：

1. **弹性伸缩**

- 弹性伸缩包括：弹性扩容、弹性缩容，前者在业务增长时自动增加云主机，后者在业务下降时自动减少云主机；
- 提供ZWatch监控报警触发弹性伸缩，可自定义接收端类型，包括系统/邮箱/钉钉/HTTP应用/短信/Microsoft Teams。

如图 *1:* 弹性伸缩所示：

图 1: 弹性伸缩



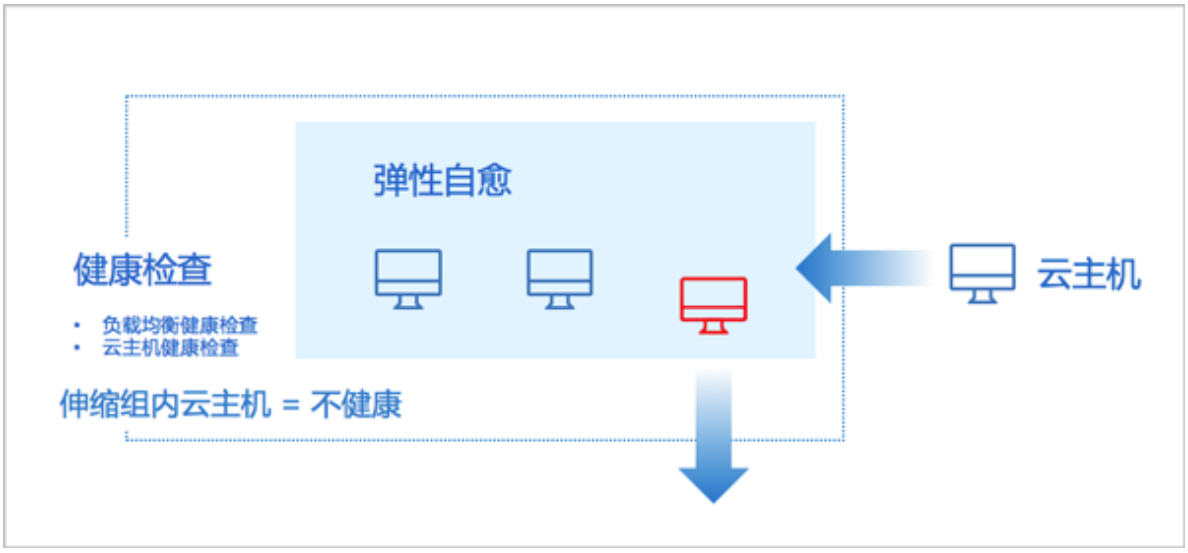
2. **弹性自愈**

- 弹性自愈：通过监控伸缩组内云主机的健康状态，自动移除不健康云主机并创建新的云主机，确保组内健康云主机数不低于设置的最小值；

- 提供两种健康检查机制触发弹性自愈：负载均衡健康检查、云主机健康检查，若伸缩组配置了负载均衡功能，建议选择负载均衡器自带的健康检查机制。

如图 *2:* 弹性自愈所示：

图 2: 弹性自愈



弹性伸缩组触发机制

弹性扩容触发条件：

1. 伸缩组内云主机负载超过阈值时触发弹性扩容
   - 若伸缩组处于冷却时间内，则不执行弹性扩容。
   - 若伸缩组内云主机数量已达上限，则不再新增云主机。
   - 伸缩组内云主机的数量不会大于最大云主机数量。
2. 伸缩组内云主机数量低于最小云主机数量时触发弹性扩容
   - 若伸缩组内云主机数量已达上限，则不再新增云主机。
   - 伸缩组内云主机的数量不会大于最大云主机数量。

弹性缩容触发条件：

1. 伸缩组内云主机负载低于阈值时触发弹性缩容
   - 若伸缩组处于冷却时间内，则不执行弹性缩容。
   - 若伸缩组内云主机数量已达下限，则不再减少云主机。
   - 伸缩组内云主机的数量不会小于最小云主机数量。

2. 伸缩组内云主机数量高于最大云主机数量时触发弹性缩容

   - 若伸缩组内云主机数量已达下限，则不再减少云主机。
   - 伸缩组内云主机的数量不会小于最小云主机数量。

弹性自愈触发条件：云主机为不健康状态

- 云平台将删除不健康的云主机，若删除后伸缩组内云主机数量小于最小云主机数量，则执行弹性扩容策略，自动添加云主机。

典型应用场景

以下介绍弹性伸缩服务的三种典型应用场景：

- 弹性扩容：

  某电商公司在双十一、春节等大型节日期间发起抢红包、秒杀等促销活动，负载激增，需及时、自动增加云主机进行扩容，避免访问延时和资源超负荷运行。

- 弹性缩容：

  当节假日过后，该电商公司的业务负载明显回落，需及时、自动减少云主机进行缩容，避免资源浪费。

- 弹性自愈：

  为保障该电商公司核心业务的正常运作，要求处于健康运行的云主机数量不能低于某个阈值。

2 准备工作

admin请提前安装最新版本的ZStack，并部署完成创建云主机必要的资源。详情可参考《用户手册》的安装部署章节。

本教程将详细介绍弹性伸缩服务的使用方法。

3 快速使用流程

弹性伸缩服务的快速使用流程如下：

1. 创建弹性伸缩组。如何创建弹性伸缩组，请参考弹性伸缩组章节。
2. 管理弹性伸缩组。如何管理弹性伸缩组，请参考弹性伸缩组章节。

4 弹性伸缩组

弹性伸缩组是一组具有相同应用场景的云主机集合，可根据用户业务变化自动实现弹性伸缩或弹性自愈。

弹性伸缩组支持以下操作：

- 创建弹性伸缩组
- 查看弹性伸缩组详情
- 启用弹性伸缩组
- 停用弹性伸缩组
- 删除弹性伸缩组

创建弹性伸缩组

在ZStack私有云主菜单，点击**云资源池 > 弹性伸缩组**，进入**弹性伸缩组**界面，点击**创建弹性伸缩组**，弹出**创建弹性伸缩组**界面。

创建弹性伸缩组分为以下三步：

1. 设置基本信息。可参考以下示例输入相应内容：
   - **伸缩组名称**：设置伸缩组名称
   - **简介**：可选项，可留空不填
   - **最小云主机数量**：设置最小云主机数量
     - 弹性缩容时组内云主机数不能低于设置的最小值；
     - 需输入整数，单位：台，取值范围：1~1000，请结合实际情况设置。
   - **最大云主机数量**：设置最大云主机数量
     - 弹性扩容时组内云主机数不能高于设置的最大值；
     - 需输入整数，单位：台，取值范围：1~1000，请结合实际情况设置。
   - **起始云主机数量**：设置起始云主机数量
     - 初次创建伸缩组时组内云主机数为设置的初始值；
     - 需输入整数，单位：台，取值范围：1~1000，且需在最小云主机数量与最大云主机数量之间，请结合实际情况设置。
   - 若选择配置负载均衡功能（推荐）：
     - **负载均衡器**：选择某一负载均衡器
       - 需提前创建负载均衡器，并绑定一个或多个监听器；
       - 如何使用负载均衡服务，请参考《用户手册》的负载均衡章节。

- **监听器**：选择负载均衡器后，必须选择监听器
  - 待选列表显示了该负载均衡器绑定的全部监听器；
  - 若选多个监听器，将分别通过不同端口对同一组云主机进行监听。
- **三层网络**：选择三层网络，用于创建云主机
  - 若未使用负载均衡服务，支持选择VPC网络/云路由网络/扁平网络。其中，VPC网络需加载到VPC路由器。
  - 若使用负载均衡服务，支持选择负载均衡服务可用的网络。
    - 若所选负载均衡器使用公有网络创建的虚拟IP提供负载均衡服务，支持以下3种场景：
      - 场景一：此网络可选择同一VPC路由器上加载的VPC网络，此时，需确保创建VPC路由器规格中的三层网络与虚拟IP所在公有网络相同。
      - 场景二：此网络可选择某个云路由网络，此时，需确保该云路由网络加载的云路由规格中的三层网络与虚拟IP所在公有网络相同。
      - 场景三：此网络可选择某个扁平网络，此时，需确保该扁平网络加载的云路由规格中的三层网络与虚拟IP所在公有网络相同。
    - 若所选负载均衡器使用VPC网络创建的虚拟IP提供负载均衡服务，此网络支持选择同一VPC路由器上加载的VPC网络，此时，所选VPC网络必须与提供虚拟IP的VPC网络加载到同一VPC路由器。
    - 若所选负载均衡器使用扁平网络创建的虚拟IP提供负载均衡服务，支持以下2个场景：
      - 场景一：此网络可选择创建虚拟IP的扁平网络。此时，该网络加载云路由规格中的三层网络不限。
      - 场景二：此网络可选择其他扁平网络。此时，需确保该网络加载云路由规格中的三层网络与虚拟IP所在扁平网络相同。

注: 若所选监听器已绑定云主机网卡，此网络和监听器已选网络必须属于同一路由器。

- **健康检查**：建议选择负载均衡健康检查
  - 负载均衡健康检查：负载均衡器自带的健康检查机制。
  - 关于负载均衡健康检查机制的详细介绍，请参考《用户手册》的负载均衡章节。
- **健康检查宽限时间**：选择负载均衡健康检查后，必须设置健康检查宽限时间

- 健康检查宽限时间：伸缩组内云主机创建启动后的一段时间，在该时间内，云主机相关应用服务可能仍在启动中，伸缩组不进行负载均衡健康检查，超过该时间，将基于负载均衡健康检查机制监控云主机健康状态；
- 需输入大于10的整数，单位：秒/分/小时，请结合实际情况设置。

如图 3: 伸缩组配置负载均衡(推荐)所示：

图 3: 伸缩组配置负载均衡(推荐)

- 若选择不配置负载均衡功能：
  - **负载均衡器**：留空不选
  - **监听器**：留空不选
  - **三层网络**：选择三层私有网络

    注: 目前支持云路由网络/VPC网络场景的云主机弹性伸缩。
  - **健康检查**：默认显示云主机健康检查

- 云主机健康检查：实时检查云主机健康状态，若检测到云主机处于不健康状态（包括：停止状态、未知状态、已删除状态），将自动移除不健康云主机并创建新的云主机，确保组内健康云主机数不低于设置的最小值。

如图 *4:* 伸缩组不配置负载均衡所示：

图 4: 伸缩组不配置负载均衡

- **启用报警通知**：伸缩组支持配置ZWatch监控报警机制触发弹性伸缩，选择是否启用报警通知
  - 默认不启用，相关报警消息可进入消息中心查看；
  - 若启用，必须指定接收端

    接收端：指定一个或多个接收端

    - 可选择系统类型接收端（默认提供），也可选择自定义类型接收端，包括邮箱/钉钉/HTTP应用/短信/Microsoft Teams；
    - 如何创建接收端，请参考《用户手册》的接收端章节。
- **创建后立即启用**：选择是否创建后立即启用伸缩组，默认不勾选

如图 *5: Step1* 设置基本信息所示：

图 5: Step1 设置基本信息


下一步(1/3)
取消
创建弹性伸缩组: 设置基本信息
区域
ZONE-1
伸缩组名称 *
伸缩组-业务A
简介
最小云主机数量 *
5
最大云主机数量 *
10
起始云主机数量 *
5

2. 配置伸缩云主机。伸缩配置定义了伸缩组内云主机的模板配置信息，可参考以下示例输入相应内容：

   - **云主机名称**：设置云主机名称
     - 组内云主机统一命名规则：**asg-伸缩组名称-云主机名称-云主机UUID前5位**，其中asg是autoscaling group的缩写。
   - **云主机简介**：可选项，可留空不填
   - **计算规格**：选择云主机的计算规格
   - **镜像**：选择创建云主机的镜像

注:

- 云主机镜像目前支持添加qcow2格式和raw格式；
- 如需使用内部监控条目，请选择已安装agent的镜像，或使用User Data脚本方式手动安装agent；
- 创建完成后修改镜像，新镜像仅对后续新生成的云主机生效，原有云主机镜像不变。

- **三层网络**：默认显示上一步中已设置的三层私有网络
- **高级**：可对高级选项进行设置
  - **数据云盘规格**：选择数据云盘规格，可为云主机直接创建并挂载数据云盘
  - **安全组**：选择安全组，组内云主机将共享相同的安全组规则
  - **控制台密码**：设置控制台密码（VNC密码），长度为6-18位
  - **SSH 公钥**：注入SSH公钥，云主机可SSH免密登录
    - SSH公钥注入需镜像提前安装cloud-init，且cloud-init推荐版本为：0.7.9、17.1、19.4、19.4以后版本；
    - 关于SSH公钥的详细介绍，请参考《用户手册》的*SSH KEY管理*章节。
  - **User Data**：导入User Data，通过上传自定义的参数或脚本，对云主机做定制化配置或完成特定任务
    - Linux云主机导入User Data，云主机镜像需提前安装cloud-init。

      Linux云主机导入User Data样例如下：

```
#cloud-config
users:
 - name: test
   shell: /bin/bash
   groups: users
   sudo: ['ALL=(ALL) NOPASSWD:ALL']
   ssh-authorized-keys:
       - ssh-rsa AAAAB3NzaC1LXCJfjroD1IT root@10-0-0-18
bootcmd:
 - mkdir /tmp/temp
write_files:
 - path: /tmp/ZStack_config
   content: |
       Hello,world!
   permissions: '0755'
hostname: Perf-test
disable_root: false
ssh_pwauth: yes
chpasswd:
  list: |
      root:word
```

```
  expire: False
runcmd:
 - echo ls -l / >/root/list.sh
```

上述样例脚本实现以下功能：

1. 创建云主机时，创建用户test，使用ssh-key；
2. 开机写入文件*/etc/hosts*，创建*/tmp/temp*目录，并创建文件写入内容；
3. 设置hostname，开启root用户，允许ssh密码登录，修改root密码；
4. 执行echo ls -l /命令。

如需使用User Data安装agent内部监控，请参考*agent*安装章节。

- Windows云主机导入User Data，云主机镜像需提前安装Cloudbase-Init，具体安装方法可参考*Cloudbase*官方文档。

  Windows云主机导入User Data样例如下：

```
#cloud-config
write_files:
  -   encoding: b64
      content: NDI=
      path: C:\b64
      permissions: '0644'
  -   encoding: base64
      content: NDI=
      path: C:\b64_1
      permissions: '0644'
  -   encoding: gzip
      content: !!binary |
          H4sIAGUfoFQC/zMxAgCIsCQyAgAAAA==
      path: C:\gzip
      permissions: '0644'
```

  上述样例脚本实现以下功能：云主机启动过程中，在c盘下创建**b64**、**b64_1**、**gzip**三个文件。

  注: 使用User Data时，一个二层网络仅支持配置一个三层网络。

如图 *6: Step2* 配置伸缩云主机所示：

图 6: Step2 配置伸缩云主机


上一步
下一步(2/3)
取消
创建弹性伸缩组: 配置伸缩云主机
云主机名称 *
VM
云主机简介
计算规格 *
InstanceOffering-1
镜像 *
Image-1
如需使用内部监控条目，请选择已安装agent的镜像，或使用User Data脚本方式手动安装agent。

注:

若将模板配置中的资源删除（例如计算规格、镜像、网络等），将导致伸缩组创建失败，请谨慎操作。

3. 配置伸缩策略。伸缩策略包括：扩容策略、缩容策略。

   - 扩容策略：

     - 在业务增长时，伸缩组自动增加云主机进行扩容，避免访问延时和资源超负荷运行；
     - 通过设置ZWatch监控报警机制触发弹性扩容；

       例如，当监测到伸缩组内全部云主机的平均内存使用率在一段时间内持续突破80%，将自动创建合适数量的云主机，使伸缩组重新达到合理的负载均衡。

     配置扩容策略，可参考以下示例输入相应内容：

     - **触发条目**：选择触发条目，包括：
       - 云主机CPU平均使用率：伸缩组内单个云主机CPU使用率之和/伸缩组内云主机总数量
       - 云主机内存平均使用率：伸缩组内单个云主机内存使用率之和/伸缩组内云主机总数量

- 云主机CPU平均使用率（需安装agent）：伸缩组内单个云主机CPU使用率之和/伸缩组内云主机总数量
- 云主机内存平均使用率（需安装agent）：伸缩组内单个云主机内存使用率之和/伸缩组内云主机总数量

注:

- 推荐使用agent内部监控监测云主机内存平均使用率，数据更为准确；
- 若选择需安装agent的触发条目，请选择已安装agent的镜像；
- Linux云主机支持使用User Data脚本方式手动安装agent，请参考本章节**User Data**部分内容；
- 若未安装agent内部监控，仍然选择需安装agent的触发条目，则弹性伸缩组无法生效。

- **触发条件**：设置触发条件，可选择设置：大于某值、大于等于某值；
  - 需输入1~100的整数，单位：%，请结合实际情况设置。
- **持续时间**：设置阈值持续时间
  - 需输入大于0的整数，单位：秒/分/小时，请结合实际情况设置。
- **冷却时间**：设置冷却时间
  - 冷却时间：伸缩组执行完成一次伸缩活动后的一段时间，在该时间内，伸缩组处于锁定状态，不执行其它伸缩活动；
  - 需输入大于0的整数，单位：秒/分，请结合实际情况设置。
- **每次增加云主机数量**：伸缩组执行一次扩容活动允许增加的云主机数量

注: 伸缩组每次扩容最小允许增加1台云主机，该数值设置过大可能导致扩容活动失败。

如图 *7: 配置扩容策略*所示：

图 7: 配置扩容策略

- 缩容策略：
  - 在业务下降时，伸缩组自动减少云主机进行缩容，避免资源浪费；
  - 通过设置ZWatch监控报警机制触发弹性缩容；

    例如，当监测到伸缩组内全部云主机的平均内存使用率在一段时间内持续低于20%，将自动移除合适数量的云主机，使伸缩组重新达到合理的负载均衡。

  配置缩容策略，可参考以下示例输入相应内容：

  - **触发条目**：设置缩容策略时，触发条目不可选择，与扩容时触发条目保持一致
  - **触发条件**：设置触发条件，可选择设置：小于某值、小于等于某值；
    - 需输入1~100的整数，单位：%，且不能与扩容策略触发条件冲突，请结合实际情况设置。
  - **持续时间**：设置阈值持续时间
    - 需输入大于0的整数，单位：秒/分/小时，请结合实际情况设置。
  - **冷却时间**：设置冷却时间

  - 冷却时间：伸缩组执行完成一次伸缩活动后的一段时间，在该时间内，伸缩组处于锁定状态，不执行其它伸缩活动；
  - 需输入大于0的整数，单位：秒/分，请结合实际情况设置。
- **移除策略**：选择移除策略，包括：
  - 最新创建的云主机（默认）：当伸缩组开始执行缩容活动时，将从最新一次创建的云主机开始逐台移除；
  - 最早创建的云主机：当伸缩组开始执行缩容活动时，将从最早创建的云主机开始逐台移除；
  - CPU利用率最低的云主机：当伸缩组开始执行缩容活动时，将从CPU利用率最低的云主机开始逐台移除；
  - 内存利用率最低的云主机：当伸缩组开始执行缩容活动时，将从内存利用率最低的云主机开始逐台移除。
- **每次减少云主机数量**：伸缩组执行一次缩容活动允许减少的云主机数量

  注: 伸缩组每次缩容最小允许减少1台云主机，该数值设置过大可能导致缩容活动失败。

如图 *8:* 配置缩容策略所示：

图 8: 配置缩容策略

如图 *9: Step3* 配置伸缩策略所示：

图 9: Step3 配置伸缩策略


上一步
确定
取消
创建弹性伸缩组: 配置伸缩策略
扩容策略
触发条目 *
云主机内存平均使用率
触发条件 *
>
80
%
持续时间 *
180
秒
冷却时间 *
300
秒
每次增加云主机数量 *
2

查看弹性伸缩组详情

在**弹性伸缩组**界面，选择某一伸缩组，展开其详情页，可查看当前创建的伸缩组状态和信息，包括：基本属性、云主机、伸缩记录、审计。

- 基本信息：
  - 展示伸缩组当前状态、名称和简介、基本配置（包括：最小云主机数量、最大云主机数量、当前云主机数量、起始云主机数量、健康检查机制、负载均衡信息、报警通知的接收端等）、伸缩配置（伸缩组内云主机的模板配置信息）、伸缩策略（包括：扩容策略、缩容策略）等信息；
  - 其中，名称和简介、扩容策略、缩容策略支持修改。
- 云主机：
  - 展示伸缩组内当前健康运行的云主机列表；

- 伸缩组基于健康检查机制（推荐负载均衡健康检查）监控云主机的健康状态，若检测到不健康云主机，将自动将其移除并创建新的云主机，确保组内健康云主机数不低于设置的最小值。
- 展示弹性伸缩组的监控状态，包括以下两种
  - 采集正常：弹性伸缩组可以正常采集云主机监控数据；
  - 数据不足：弹性伸缩组无法正常采集云主机监控数据，可能原因如下：
    1. 新创建的云主机，采集监控数据需等待一段时间；
    2. 云主机未安装agent，无法采集数据，请进入云主机详情页安装“性能优化工具”；
    3. 云主机状态异常，请检查云主机状态。

- 伸缩记录：展示伸缩组进行伸缩活动的记录信息，可调整合适的时间段进行搜索。
- 审计：查看此伸缩组的相关操作。

启用/停用弹性伸缩组

- 启用弹性伸缩组：将已停用的伸缩组启用。
- 停用弹性伸缩组：将伸缩组停用。

注:

- 若伸缩组已触发伸缩活动，停用伸缩组，正在执行的伸缩活动不受影响，该伸缩活动执行完成后，将停止触发新的伸缩活动；
- 若伸缩组处于检测中，停用伸缩组，ZWatch/健康检查机制将立即停止检测伸缩组，并停止触发新的伸缩活动。

删除弹性伸缩组

删除弹性伸缩组，将一并删除组内全部云主机，请谨慎操作。

补充说明

- 弹性伸缩组内云主机上运行的业务应用必须无状态并且可横向扩展。
- 弹性伸缩会自动释放云主机，建议不要对伸缩组内云主机手动挂载云盘、网卡、安全组等，若组内云主机保存有状态信息，相关数据将会丢失！
- 弹性伸缩组无法纵向扩展，即：无法自动扩缩云主机的计算规格、网络带宽等配置。
- 若需要修改外部监控触发条目为内部监控使用条目，请删除弹性伸缩组并重新创建。
- 关于弹性伸缩服务，提供以下全局设置：

- 当伸缩组使用负载均衡健康检查机制时，可设置云主机在负载均衡中健康状态检查的时间间隔。

  设置方法：进入**设置** > **全局设置** > **高级设置**，设置**负载均衡云主机健康检查间隔**即可，默认为10，单位为秒，最小值不能低于10秒，最大值不能高于1000秒。
- 当伸缩组使用负载均衡健康检查机制时，可设置云主机在负载均衡中健康状态检查的线程数。

  设置方法：进入**设置** > **全局设置** > **高级设置**，设置**负载均衡云主机健康检查线程数**即可，默认为10，最小值不能低于10线程，最大值不能高于1000线程。
- 当伸缩组使用云主机健康检查机制时，可设置删除组内不健康云主机的时间间隔。

  设置方法：进入**设置** > **全局设置** > **高级设置**，设置**组内不健康实例删除间隔**即可，默认为30，单位为秒，最小值不能低于10秒，最大值不能高于1000秒。
- 当伸缩组使用云主机健康检查机制时，可设置删除组内不健康云主机的线程数。

  设置方法：进入**设置** > **全局设置** > **高级设置**，设置**组内不健康实例删除线程数**即可，默认为10，最小值不能低于10线程，最大值不能高于1000线程。
- 可设置伸缩组内云主机数量检查的时间间隔。

  设置方法：进入**设置** > **全局设置** > **高级设置**，设置**组内实例数量检查间隔**即可，默认为20，单位为秒，最小值不能低于10秒，最大值不能高于1000秒。

注意事项

- 若弹性伸缩组重复执行伸缩策略，如不断创建和删除云主机，可能为以下原因：
  - 新创建的云主机无法在容忍时间内达到健康状态，云平台触发弹性自愈策略，删除不健康的云主机并重新创建，造成循环，需检查云主机健康检查策略或更改健康检查机制。
  - 缩容阈值或扩容阈值设置不合理。例如：设置触发条件为CPU低于40%缩容，CPU高于45%扩容，若伸缩组只有一台云主机，组内云主机平均CPU负载为60%，触发扩容后增加至两台云主机，组内云主机平均CPU负载降为30%，造成循环，需设置扩容缩容策略至合理的阈值范围。
- 若伸缩组未执行伸缩策略，但不断触发报警，可能为以下原因：
  - 最大云主机数量和扩容触发条件设置不合理，当伸缩组云主机数量已达上限，组内云主机平均负载仍然高于扩容阈值，则不断触发报警，需设置最大云主机数量和扩容触发条件至合理的阈值范围。

5 典型场景实践

背景信息

场景设定：假定某电商公司已部署一套最新的ZStack私有云环境，由于业务需要，现要部署一套业务云主机，并使用弹性伸缩组提供基于负载均衡的云主机弹性伸缩服务。

弹性伸缩组支持使用agent内部监控触发弹性伸缩服务，本场景以外部监控为例介绍具体操作流程。

具体实践流程如下：

1. 创建弹性伸缩组；
2. 功能验证：弹性自愈、弹性扩容、弹性缩容。

操作步骤

1. 创建弹性伸缩组。

   在ZStack私有云主菜单，点击**云资源池** > **弹性伸缩组**，进入**弹性伸缩组**界面，点击**创建弹性伸缩组**，弹出**创建弹性伸缩组**界面。

   创建弹性伸缩组分为以下三步：

   a) 设置基本信息。

   可参考以下示例输入相应内容：

   - **伸缩组名称**：设置伸缩组名称，例如：伸缩组-业务A
   - **简介**：可选项，可留空不填
   - **最小云主机数量**：5
   - **最大云主机数量**：10
   - **起始云主机数量**：5
   - **负载均衡器**：选择已创建的负载均衡器
   - **监听器**：选择监听器
   - **三层网络**：选择可用的三层网络
   - **健康检查**：选择负载均衡健康检查
   - **健康检查宽限时间**：300秒
   - **启用报警通知**：勾选，启用报警通知
   - **接收端**：指定已创建的自定义接收端

- **创建后立即启用**：勾选，创建后立即启用伸缩组

图 10: Step1 设置基本信息

b) 配置伸缩云主机。

可参考以下示例输入相应内容：

- **云主机名称**：设置云主机名称，例如VM
- **云主机简介**：可选项，可留空不填
- **计算规格**：选择云主机的计算规格
- **镜像**：选择创建云主机的镜像
- **三层网络**：默认显示上一步中已设置的云路由网络

- **高级**：可按需对云主机进行高级设置

图 11: Step2 配置伸缩云主机

c) 配置伸缩策略。

配置扩容策略，可参考以下示例输入相应内容：

- **触发条目**：选择触发条目，例如：云主机内存平均使用率
- **触发条件**：大于80%
- **持续时间**：180秒
- **冷却时间**：300秒
- **每次增加云主机数量**：2

配置缩容策略，可参考以下示例输入相应内容：

- **触发条目**：默认显示云主机内存平均使用率
- **触发条件**：小于20%
- **持续时间**：180秒
- **冷却时间**：300秒
- **移除策略**：选择移除策略，例如：最早创建的云主机
- **每次减少云主机数量**：2

如图 *12: Step3* 配置伸缩策略所示：

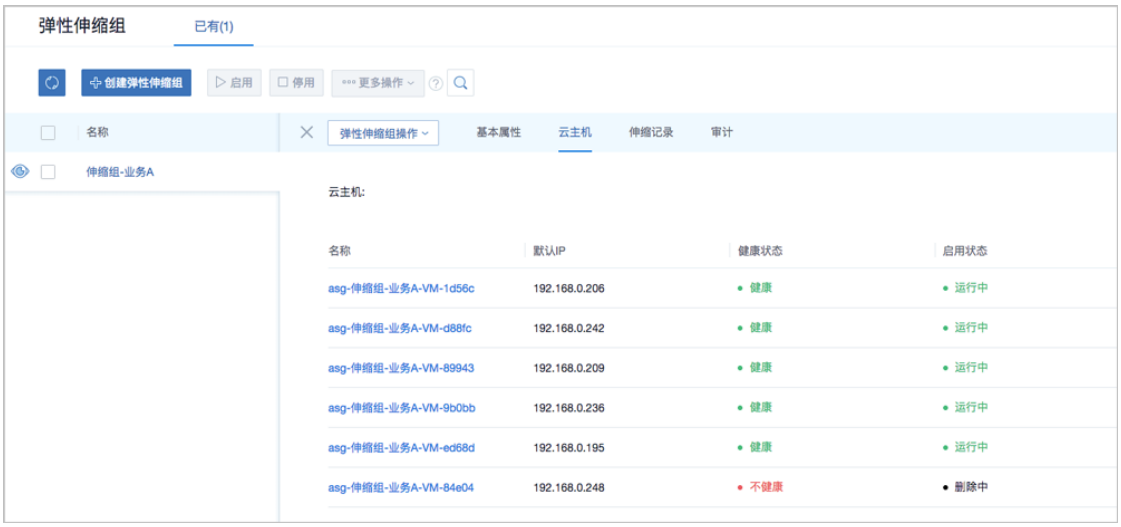
图 12: Step3 配置伸缩策略

2. 功能验证。

   - 弹性自愈：伸缩组处于健康运行的云主机数量持续保持在5台或以上，保障业务正常运作。

     如图 *13:* 删除不健康云主机所示：

     图 13: 删除不健康云主机



- 弹性扩容：在双十一、春节等大型节日期间，业务负载激增，伸缩组自动增加云主机进行扩容，高峰时云主机数量可达到10台，有效避免访问延时和资源超负荷运行。

  如图 *14:* 弹性扩容所示：

图 14: 弹性扩容

- 弹性缩容：节假日过后，业务负载明显回落，伸缩组自动减少云主机进行缩容，避免资源浪费。

  如图 *15:* 弹性缩容所示：

图 15: 弹性缩容

后续操作

至此，弹性伸缩使用方法介绍完毕。

术语表

区域（Zone）

ZStack中最大的一个资源定义，包括集群、二层网络、主存储等资源。

集群（Cluster）

一个集群是类似物理主机（Host）组成的逻辑组。在同一个集群中的物理主机必须安装相同的操作系统（虚拟机管理程序，Hypervisor），拥有相同的二层网络连接，可以访问相同的主存储。在实际的数据中心，一个集群通常对应一个机架（Rack）。

管理节点（Management Node）

安装系统的物理主机，提供UI管理、云平台部署功能。

计算节点（Compute Node）

也称之为物理主机（或物理机），为云主机实例提供计算、网络、存储等资源的物理主机。

主存储（Primary Storage）

用于存储云主机磁盘文件的存储服务器。支持本地存储、NFS、 Ceph、Shared Mount Point、Shared Block类型。

镜像服务器（Backup Storage）

也称之为备份存储服务器，主要用于保存镜像模板文件。建议单独部署镜像服务器。支持ImageStore、Sftp（社区版）、Ceph类型。

镜像仓库（Image Store）

镜像服务器的一种类型，可以为正在运行的云主机快速创建镜像，高效管理云主机镜像的版本变迁以及发布，实现快速上传、下载镜像，镜像快照，以及导出镜像的操作。

云主机（VM Instance）

运行在物理机上的虚拟机实例，具有独立的IP地址，可以访问公共网络，运行应用服务。

镜像（Image）

云主机或云盘使用的镜像模板文件，镜像模板包括系统云盘镜像和数据云盘镜像。

云盘（Volume）

云主机的数据盘，给云主机提供额外的存储空间，共享云盘可挂载到一个或多个云主机共同使用。

计算规格（Instance Offering）

启动云主机涉及到的CPU数量、内存、网络设置等规格定义。

云盘规格（Disk Offering）

创建云盘容量大小的规格定义。

二层网络（L2 Network）

二层网络对应于一个二层广播域，进行二层相关的隔离。一般用物理网络的设备名称标识。

三层网络（L3 Network）

云主机使用的网络配置，包括IP地址范围、网关、DNS等。

公有网络（Public Network）

由因特网信息中心分配的公有IP地址或者可以连接到外部互联网的IP地址。

私有网络（Private Network）

云主机连接和使用的内部网络。

L2NoVlanNetwork

物理主机的网络连接不采用Vlan设置。

L2VlanNetwork

物理主机节点的网络连接采用Vlan设置，Vlan需要在交换机端提前进行设置。

VXLAN网络池（VXLAN Network Pool）

VXLAN网络中的 Underlay 网络，一个 VXLAN 网络池可以创建多个 VXLAN Overlay 网络（即 VXLAN 网络），这些 Overlay 网络运行在同一组 Underlay 网络设施上。

VXLAN网络（VXLAN）

使用 VXLAN 协议封装的二层网络，单个 VXLAN 网络需从属于一个大的 VXLAN 网络池，不同 VXLAN 网络间相互二层隔离。

云路由（vRouter）

云路由通过定制的Linux云主机来实现的多种网络服务。

安全组（Security Group）

针对云主机进行第三层网络的防火墙控制，对IP地址、网络包类型或网络包流向等可以设置不同的安全规则。

弹性IP（EIP）

公有网络接入到私有网络的IP地址。

快照（Snapshot）

某一时间点某一磁盘的数据状态文件。包括手动快照和自动快照两种类型。