云容器引擎
Autopilot
快速入门

文档版本 01

发布日期 2024-12-18

华为云计算技术有限公司

商标声明


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。

华为云计算技术有限公司

地址： 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址： https://www.huaweicloud.com/

目 录

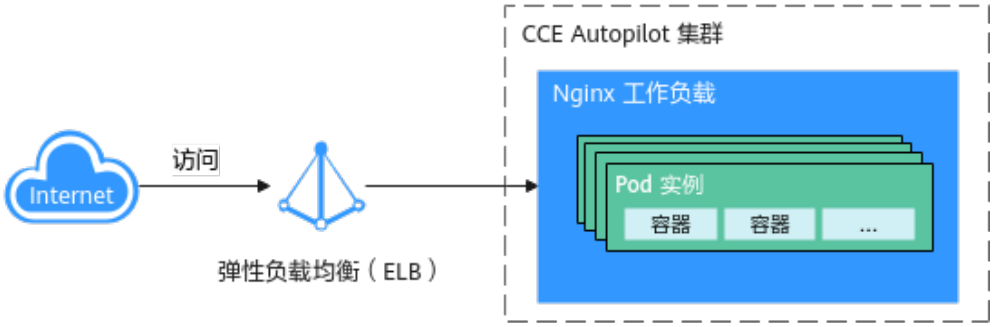
1 在 CCE Autopilot 集群中部署 Nginx 工作负载

CCE Autopilot集群是云容器引擎服务推出的Serverless版集群，支持智能化版本升级、漏洞自动修复和智能调参等技术，能够为您提供更智能的使用体验。同时，CCE Autopilot集群采用底层统一资源池技术，您无需管理和维护底层资源设施的分配和扩展，可以有效降低运维成本。底层统一资源池技术支持故障快速隔离和修复，能够确保应用的持续稳定运行，提升应用的可靠性。

Nginx是一款高性能的开源HTTP服务器和反向代理服务器，广泛用于处理高并发、负载均衡和静态资源服务。在集群中部署Nginx工作负载，可以通过负载均衡和反向代理有效地分配流量，确保高可用性和容错性。它还能够简化微服务架构中的流量管理、安全控制和API网关功能，提升系统的灵活性和扩展性。

本示例以部署Nginx工作负载为例，帮助您了解如何创建CCE Autopilot集群以及如何在集群中部署工作负载，整体架构请参见**图1 方案架构**。

图 1-1 方案架构



操作流程

表 1-1 部署 Nginx 的操作流程

操作步骤	步骤说明	费用说明
准备工作	您需要注册华为账号，并为账号充值。	不涉及费用。

操作步骤	步骤说明	费用说明
步骤一：首次开通CCE并进行授权	当您的账号在当前区域中首次使用CCE时，您需要为CCE进行授权。	不涉及费用。
步骤二：创建CCE Autopilot集群	在CCE服务中创建CCE Autopilot集群，以提供更简化的Kubernetes服务。	涉及集群管理和终端节点等费用，具体请参见集群计费说明。
步骤三：部署并访问Nginx工作负载	在集群中创建工作负载，并为其创建负载均衡类型的服务，然后您就可以从公网访问您的工作负载。	涉及Pod和弹性负载均衡（ELB）费用，具体请参见集群计费说明和ELB计费说明。
后续操作：释放资源	如果您在完成实践后不需要继续使用集群，请及时清理资源以免产生额外扣费。	不涉及费用。

准备工作

- 您需要注册华为账号并完成实名认证，详情请参见注册华为账号并开通华为云和个人实名认证。
- 您需要确保账号有足够的资金，以免创建集群失败，具体操作请参见账号充值。

步骤一：首次开通 CCE 并进行授权

由于CCE在运行中对计算、存储、网络以及监控等各类云服务资源都存在依赖关系，因此当您首次登录CCE控制台时，CCE将自动请求获取当前区域下的云资源权限，从而更好地为您提供服务。如果您在当前区域已完成授权，可忽略本步骤。

步骤1 使用华为账号登录CCE控制台。

步骤2 单击控制台左上角的  ，选择“区域”，本示例使用区域为“华东-上海一”。

步骤3 在首次登录某个区域的CCE控制台时将跳出“授权说明”，请您在仔细阅读后单击“确定”。

当您同意授权后，CCE将在IAM中创建名为“cce_admin_trust”委托，统一对您的其他云服务资源进行操作，并且授予其Tenant Administrator权限。Tenant Administrator拥有除IAM管理外的全部云服务管理员权限，用于对CCE所依赖的其他云服务资源进行调用，且该授权仅在当前区域生效。

须知

CCE对其他云服务存在依赖，如果没有Tenant Administrator权限，可能会因为某服务权限不足而无法正常使用。因此，在使用CCE服务期间，请不要自行删除或者修改“cce_admin_trust”委托。

----结束

步骤二：创建 CCE Autopilot 集群

在CCE服务中创建CCE Autopilot集群，以提供更简化的Kubernetes服务。本示例中仅解释必要参数，其他参数保留默认值。关于其他参数的详细说明，请参见**购买Autopilot集群**。

步骤1 进入**CCE控制台**。

- 如果您的账号在当前区域未创建过集群，请在当前页面单击“购买集群”或“购买CCE Autopilot集群”，进入购买页。
- 如果您的账号在当前区域已创建过集群，请在左侧菜单栏选择集群管理，单击右上角“购买集群”，进入购买页。

步骤2 配置集群基础信息，具体的参数示例请参见**图1-2**和**表1-2**。

图 1-2 集群基本信息

表 1-2 集群基础信息

参数	示例	参数说明
集群类型	**CCE Autopilot集群**	CCE支持创建多种类型集群，满足各种业务需求，提供高可靠、安全的商业级容器集群服务。 • CCE Standard集群：标准版本集群，提供高可靠、安全的商业级容器集群服务。 • CCE Turbo集群：拥有更高性能的云原生网络，提供云原生混部调度能力，可实现更高的资源利用率和更广的全场景覆盖。 • CCE Autopilot集群：Serverless版集群，提供免运维的容器服务，可以大幅降低运维成本，提高应用程序的可靠性和可扩展性。 了解集群类型的更多内容，请参见**集群对比**。
集群名称	**autopilot-example**	新建集群的名称。 集群名称长度范围为4-128个字符，以小写字母开头，支持小写字母、数字和中划线(-)，不能以中划线(-)结尾。
企业项目	**default**	该参数仅对开通企业项目的企业客户账号显示，不显示时请忽略。 企业项目是一种资源管理单位，可跨区域归类资源，方便企业按部门或项目组集中管理。了解企业项目的更多内容，请参见**项目管理**。 请根据需要选择适合的企业项目，如果没有特殊要求，可以选择default。

参数	示例	参数说明
集群版本	**v1.28**	集群安装的Kubernetes软件版本，建议选择最新版本。

步骤3 配置集群网络信息，具体的参数示例请参见**图1-3**和**表1-3**。

图 1-3 集群网络信息

表 1-3 集群网络信息

参数	示例	参数说明
虚拟私有云	**vpc-autopilot**	选择集群所在的虚拟私有云（VPC）。如果没有可选项，单击右侧“新建虚拟私有云”创建，具体请参见**创建虚拟私有云和子网**。集群创建后，VPC不支持修改。
容器子网	**subnet-502f**	选择容器所在子网。每个Pod需要唯一的IP地址，容器子网内IP地址的数量决定了集群中Pod的数量上限，进而决定容器的数量上限，集群创建后支持新增子网。 如果没有可选项，请单击右侧“新建子网”创建，具体请参见**创建虚拟私有云和子网**。
服务网段	**10.247.0.0/16**	同一集群下容器互相访问时使用的Service资源网段，决定Service资源数量的上限。集群创建后，服务网段不支持修改。
配置SNAT	**开启**	默认开启，开启后您的集群可以通过NAT网关访问公网。默认使用所选VPC中已有的NAT网关，若VPC没有NAT网关，系统将会为您自动创建一个默认规格的NAT网关并绑定弹性公网IP，同时自动配置SNAT规则。 使用NAT网关将产生一定费用，详情请参见**NAT网关计费说明**。

步骤4 单击“下一步：插件选择”，选择创建集群时需要安装的插件。

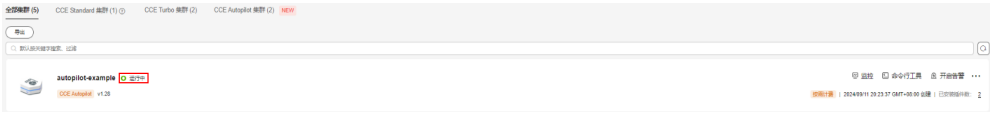
本示例中，仅选择默认安装插件，即CoreDNS域名解析和Kubernetes Metrics Server插件。

步骤5 单击“下一步：插件配置”，对选择的插件进行配置，默认插件无需配置。

步骤6 单击“下一步：确认配置”，显示集群资源清单，确认无误后，单击“提交”。

创建集群预计需要5-10分钟，请耐心等待。创建成功后，集群管理中对应集群的状态为运行中。

图 1-4 集群运行中



----结束

步骤三：部署并访问 Nginx 工作负载

在集群中创建Nginx工作负载，将应用程序或服务部署到容器环境中，实现资源的高效利用和自动化管理。同时，为该工作负载创建负载均衡类型的服务，使您能够从公网访问Nginx。本节介绍两种方式部署并访问Nginx工作负载，包括控制台方式和kubectl命令行方式。

使用控制台方式

步骤1 单击新建的集群名称，进入集群控制台。

步骤2 在左侧菜单栏中选择“工作负载”，单击右上角“创建工作负载”，进入创建页。

步骤3 配置工作负载基本信息，具体参数示例请参见图1-5和表1-4。

本示例中仅解释必要参数，其他参数保留默认值。关于其他参数的详细说明，请参见**创建工作负载**，您可以根据工作负载类型选择适合的参考文档。

图 1-5 工作负载基本信息

表 1-4 工作负载基本信息

参数	示例	参数说明
负载类型	**无状态负载** **Deployment**	工作负载是一种对Pod的抽象管理方式，用于定义和控制Pod的创建、运行和生命周期。通过工作负载，您可以批量管理和自动化控制多个Pod的行为，如伸缩、更新和恢复。 • 无状态负载（Deployment）：管理无状态应用，支持上线部署、滚动升级、创建副本和恢复上线。 • 有状态负载（StatefulSet）：管理有状态应用，确保每个Pod能够拥有独立的持久化状态，并能够在Pod重启或迁移时恢复其数据，以保障应用的可靠性和一致性。 • 普通任务（Job）：一次性任务，完成后Pod自动删除。 • 定时任务（CronJob）：基于时间的Job，指定时间周期内运行指定的Job。 了解工作负载的更多内容，请参见**工作负载概述**。 本示例将Nginx部署为无状态负载，原因在于Nginx主要用于处理请求转发、负载均衡和静态内容分发，不需要在本地保存任何持久性数据。
负载名称	**nginx**	请填写工作负载的名称。 工作负载名称长度范围为1-63个字符，可以包含小写英文字母、数字和和中划线(-)，并以小写英文字母开头，小写英文字母或数字结尾。
命名空间	**default**	命名空间是Kubernetes集群中的抽象概念，可以将集群中的资源或对象划分为一个组，且不同命名空间中的数据彼此隔离，您可以根据需要创建并使用命名空间。 集群创建后会默认生成default命名空间，如果没有特殊要求，可以直接选择default命名空间。
实例数量	**1**	工作负载中Pod实例的数量。Pod实例数量的设置策略： • 高可用性：如果您需要保证工作负载的高可用性，则实例数量至少设置为2，避免单点故障。 • 性能要求：您需要根据工作负载的流量和资源需求设置实例数量，避免过载或资源浪费。 本示例仅做演示，实例数量设置为1。

步骤4 配置工作负载容器信息，具体参数示例请参见**图1-6**和**表1-5**。

本示例中仅解释必要参数，其他参数保留默认值。关于其他参数的详细说明，请参见**创建工作负载**，您可以根据工作负载类型选择适合的参考文档。

图 1-6 工作负载容器信息

表 1-5 工作负载容器信息

参数	示例	参数说明
镜像名称	**nginx**	单击“选择镜像”，在弹出的窗口中切换至“镜像中心”，选择公共镜像。
镜像版本	**latest**	选择需要部署的镜像版本。
CPU配额	**0.25cores**	CPU资源限制值，即允许容器使用的CPU最大值，防止占用过多资源，默认0.25cores。
内存配额	**512MiB**	内存资源限制值，即允许容器使用的内存最大值。如果超过，容器会被终止，默认512MiB。

步骤5 单击“服务配置”下的＋，进入创建服务页面，配置工作负载服务信息，具体参数示例请参见**图1-7**和**表1-6**。

本示例仅解释必要参数，其他参数保留默认值。关于其他参数的详细说明，请参见**服务（Service）**，您可以根据服务类型选择适合的参考文档。

图 1-7 工作负载服务信息

表 1-6 工作负载服务信息

参数	示例	参数说明
Service名称	**nginx**	请填写服务的名称。 服务名称长度范围为1-63个字符，可以包含小写英文字母、数字和和中划线(-)，并以小写英文字母开头，小写英文字母或数字结尾。
访问类型	**负载均衡**	选择服务类型，即服务访问的方式。 ● 集群内访问：通过集群的内部IP暴露服务，只能够在集群内部访问。 ● 负载均衡：通过弹性负载均衡（ELB）对外部提供服务，即能够从公网访问到工作负载。 本示例中需要外部访问Nginx，所以访问类型设置为负载均衡。了解服务类型的更多内容，请参见**服务（Service）**。

参数	示例	参数说明
负载均衡器	• **独享型** • **网络型(TCP/UDP)** • **选择已有** • **elb-nginx**	• 如果已有弹性负载均衡（ELB）实例，可以选择已有ELB实例。 **说明** 使用已有的ELB时，ELB实例需要具备3个条件： – 与集群属于同一VPC。 – 实例类型为独享型。 – 网络类型必须支持私网（存在私有地址）。 • 如果没有弹性负载均衡（ELB）实例，请选择“自动创建”创建一个负载均衡器并绑定弹性公网IP，具体操作请参见**创建负载均衡类型的服务**。
端口配置	**协议：TCP**	负载均衡监听器端口协议。
	容器端口：80	容器中应用启动监听的端口，该容器端口需和应用对外提供的监听端口一致。 使用nginx镜像时容器端口需设置为80，原因在于Nginx默认使用80端口提供HTTP服务。
	服务端口：8080	ELB将会使用该端口创建监听器，提供外部流量访问入口，可自定义。

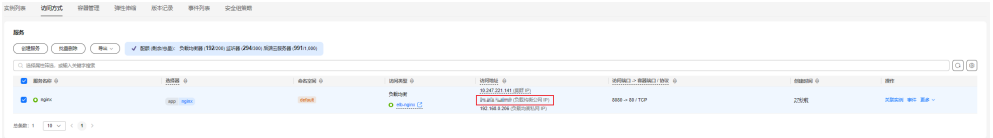
步骤6 单击右下角“创建工作负载”。

创建成功后，无状态工作负载列表中对应工作负载的状态为运行中。

图 1-8 工作负载运行中

步骤7 单击Nginx工作负载名称，进入工作负载详情页，获取Nginx的外部访问地址。“访问方式”页签中，“负载均衡公网IP：访问端口”即为外部访问地址。其中，“负载均衡公网IP”为**步骤5**中ELB实例的公网地址，“访问端口”为**步骤5**中的“服务端口”。

图 1-9 访问方式



步骤8 在浏览器中输入“负载均衡公网IP：访问端口”，即可成功访问应用。

图 1-10 访问应用

----结束

使用 kubectl 命令行方式

该步骤涉及命令行操作，您可以使用以下两种方式进行相关操作。

- **通过集群内命令行工具进行操作**：该命令行工具已经配置kubectl命令，并已连接集群，更多信息请参见**通过CloudShell连接集群**。
- **通过ECS虚拟机进行操作**：您需要购买一台Linux系统的ECS虚拟机，该ECS需与集群处于同一VPC，并绑定弹性公网IP，具体操作请参见**快速购买和使用Linux ECS**。如果已有满足条件的ECS，则无需购买。此外，您还需要安装kubectl命令，并**通过kubectl连接集群**。

以第一种方法为例，介绍如何使用kubectl命令行方式创建Nginx工作负载。

步骤1 单击新建的集群名称，进入集群控制台。

步骤2 单击右上角“命令行工具”，进入CloudShell页面。

说明

目前，只有部分区域支持CloudShell连接集群，具体情况请以控制台为准。如果区域不支持，请通过ECS虚拟机进行操作。

图 1-11 CloudShell

步骤3 执行以下命令，创建YAML文件nginx-deployment.yaml，用于配置nginx工作负载，文件名称可自定义。

说明

Linux文件命名支持字母、数字、下划线和连字符，但不能包含斜杠（/）和空字符（\0）。文件名区分大小写，建议避免使用特殊字符，如空格、问号（?）和星号（*）等，以提高兼容性。

```
vim nginx-deployment.yaml
```

文件内容如下：

```
apiVersion: apps/v1
kind: Deployment
metadata:
  name: nginx     # 工作负载名称
spec:
  replicas: 1     # 实例数量
  selector:
    matchLabels:  # 选择器，用于选择带有特定标签的资源，需与步骤6中负载均衡类型服务YAML文件中的
```

```
selector取值保持一致
      app: nginx
  template:
    metadata:
      labels:     # 标签，
        app: nginx
    spec:
      containers:
      - image: nginx:latest   # 镜像名称：镜像版本
        name: nginx
      imagePullSecrets:
      - name: default-secret
```

输入完成后，按**Esc键**退出编辑，输入**:wq**保存。

步骤4 执行以下命令，创建工作负载。

```
kubectl create -f nginx-deployment.yaml
```

回显如下，表示已经开始创建工作负载。

```
deployment.apps/nginx created
```

步骤5 执行以下命令，查看工作负载状态。

```
kubectl get deployment
```

回显如下，如果工作负载创建的Pod皆为可用状态，则表示创建成功。

```
NAME    READY   UP-TO-DATE   AVAILABLE   AGE
nginx   1/1     1            1           4m59s
```

回显内容的参数说明如下：

表 1-7 回显内容参数说明

参数	示例	参数说明
NAME	**nginx**	工作负载的名称。
READY	**1/1**	表示工作负载的可用状态，显示为“可用Pod个数/期望Pod个数”。
UP-TO-DATE	**1**	指当前工作负载已经完成更新的Pod数。
AVAILABLE	**1**	工作负载可用的Pod个数。
AGE	**4m59s**	工作负载已经运行的时间。

步骤6 执行以下命令，创建YAML文件nginx-elb-svc.yaml，用于配置负载均衡服务并关联已创建的工作负载nginx，文件名称可自定义。

本示例基于已有的弹性负载均衡（ELB）实例创建服务，如果您需要自动创建ELB请参考**通过kubectl命令行创建-自动创建ELB**。

```
vim nginx-elb-svc.yaml
```

文件内容如下：

```
apiVersion: v1
kind: Service
metadata:
  name: nginx     # 服务的名称
  annotations:
```



```
    kubernetes.io/elb.id: <your_elb_id>          # ELB ID，替换为实际值
    kubernetes.io/elb.class: performance          # 负载均衡器类型
spec:
  selector:       # selector取值需与步骤3中工作负载YAML文件中matchLabels参数取值一致
    app: nginx
  ports:
  - name: service0
    port: 8080
    protocol: TCP
    targetPort: 80
  type: LoadBalancer
```

输入完成后，按**Esc键**退出编辑，输入**:wq**保存。

表 1-8 使用已有 ELB 参数说明

参数	示例	参数说明
kubernetes.io/elb.id	**405ef586-0397-45c3-bfc4-xxx**	已有的ELB的ID。 **说明** 使用已有的ELB时，ELB实例需要具备3个条件： • 与集群属于同一VPC。 • 实例类型为独享型。 • 网络类型必须支持私网（存在私有地址）。 **获取方式：**进入网络控制台，在左上方选择集群所在区域，在左侧导航栏中选择“弹性负载均衡 > 我的ELB”，找到对应的ELB实例名称，名称下方即为对应ID。同时，单击ELB实例名称，在“基本信息”页签中验证该ELB是否满足上述条件。
kubernetes.io/elb.class	**performance**	负载均衡器类型，仅支持performance类型，即独享型负载均衡。
selector	**app: nginx**	选择器，服务将流量发送给对应标签的Pod。
ports.port	**8080**	弹性负载均衡（ELB）将会使用该端口创建监听器，提供外部流量访问入口，可自定义。
ports.protocol	**TCP**	负载均衡监听器端口协议。
ports.targetPort	**80**	Service访问目标容器的端口，此端口与容器中运行的应用强相关。 使用nginx镜像请设置为80。

步骤7 执行以下命令，创建服务。

```
kubectl create -f nginx-elb-svc.yaml
```

回显如下，表示服务已创建。

```
service/nginx created
```

步骤8 执行以下命令，查看服务。

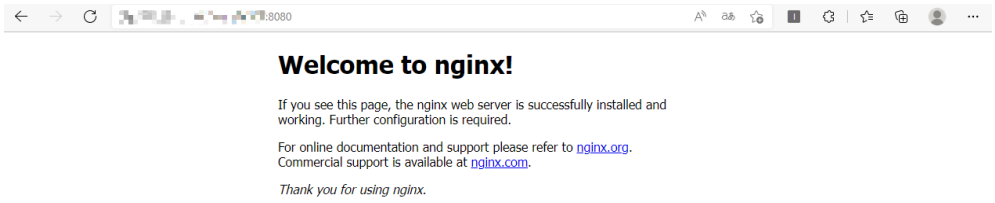
```
kubectl get svc
```

回显如下，表示工作负载访问方式已设置成功。


```
NAME         TYPE           CLUSTER-IP     EXTERNAL-IP                PORT(S)          AGE
kubernetes   ClusterIP      10.247.0.1     <none>                     443/TCP          18h
nginx        LoadBalancer   10.247.56.18   xx.xx.xx.xx,xx.xx.xx.xx    8080:30581/TCP   5m8s
```

步骤9 在浏览器中输入“外部访问地址：服务端口”，即可成功访问应用。其中外部访问地址为EXTERNAL-IP对应的第一个IP地址，服务端口为8080。

图 1-12 访问应用



----结束

后续操作：释放资源

如果您无需继续使用集群，请及时释放资源，避免产生额外的费用。

须知

- 删除集群会删除集群下工作负载和服务，相关业务将无法恢复。
- 集群关联创建的VPC级别的资源（如终端节点、NAT网关和SNAT出网EIP等），删除集群时默认保留，请确认其他集群或者地方没有重用该资源，再进行删除操作。

步骤1 进入CCE控制台，在左侧导航栏中选择“集群管理”。

步骤2 找到需要删除的集群，单击集群卡片右上角的“...”，并单击“删除集群”。

步骤3 在弹出的“删除集群”窗口中，根据页面提示删除相关资源。

步骤4 在确认框中输入“DELETE”，单击“是”，开始执行删除集群操作。

删除集群需要花费1-3分钟，请耐心等候。集群列表中对应集群名称消失，则说明删除集群成功。

----结束

2 使用 HPA 策略实现工作负载弹性伸缩

企业应用的流量呈现波动性，存在高峰期和低谷期。如果始终维持足够的计算资源以应对高峰流量，将导致较高的成本。为解决这一问题，您可以制定工作负载弹性伸缩策略。该策略可以根据流量变化或资源占用情况自动调整工作负载中Pod实例的数量，进而达到降低成本的目的。工作负载弹性伸缩策略可以在CCE Standard集群、CCE Turbo集群以及CCE Autopilot集群中进行配置，具体区别请参见**表2-1**。

本示例主要介绍CCE Autopilot集群如何通过HPA策略实现工作负载弹性伸缩。在CCE Autopilot集群中，您无需对节点的部署、管理和安全性进行维护，只需设置HPA策略，即可按需自动调整Pod实例数量，简化了资源管理和运维流程。此外，CCE Autopilot集群具有高性能的Serverless容器资源底座和多级资源池预热技术，可以实现秒级弹性伸缩，帮助您快速上线新应用，灵活应对市场变化。

表 2-1 配置工作负载弹性伸缩策略对比

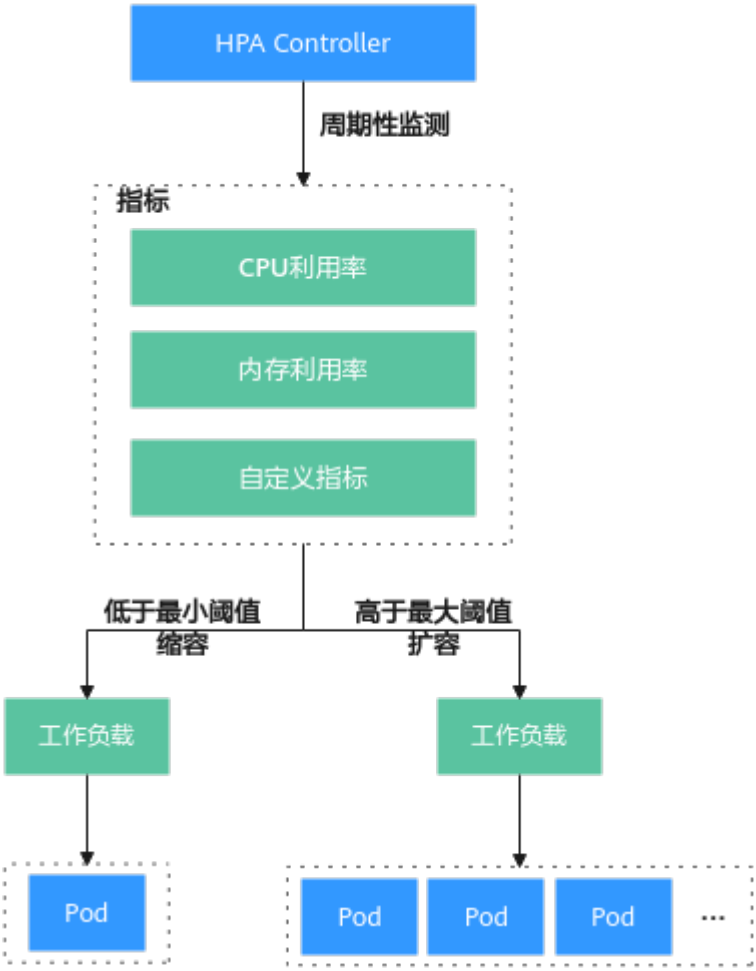
集群类型	配置策略	区别
CCE Standard/ Turbo集群	需要同时制定节点和工作负载的弹性策略，即CA（Cluster AutoScaling）策略和HPA（Horizontal Pod Autoscaling）。 • CA策略：负责节点弹性伸缩，以避免资源不足而导致工作负载创建失败。 • HPA策略：负责工作负载弹性伸缩，根据策略指标自动调整工作负载中Pod实例的数量。	• 涉及节点和工作负载的伸缩，需要同时配置CA策略和HPA策略。 • 可以实现分钟级弹性伸缩。
CCE Autopilot集群	只需要制定工作负载的弹性策略，即HPA策略。	• 涉及工作负载伸缩，需要配置HPA策略。 • 可以实现秒级弹性伸缩。

本示例只需要配置HPA策略即可，HPA策略的工作原理如下：

如**图2-1**所示，HPA策略通过周期性监测指定指标（如CPU利用率、内存利用率或其他自定义指标）实现工作负载中Pod实例数量的动态调整。以CPU利用率为例，当CPU利

用率高于最大阈值时，HPA策略将会增大Pod实例的数量以降低工作负载的计算压力；当CPU利用率低于最小阈值时，HPA策略将会减少Pod实例的数量以节约成本。了解HPA策略的更多信息，请参见**HPA工作原理**。

图 2-1 HPA 策略工作原理



操作流程

表 2-2 配置弹性伸缩的操作步骤

操作步骤	步骤说明	费用说明
准备工作	您需要注册华为账号，并为账号充值。	不涉及费用。
步骤一：首次开通CCE并进行授权	当您的账号在当前区域中首次使用CCE时，您需要为CCE进行授权。	不涉及费用。
步骤二：创建CCE Autopilot集群	在CCE服务中创建CCE Autopilot集群，以提供更简化的Kubernetes服务。	涉及集群管理和终端节点等费用，具体请参见**集群计费说明**。

操作步骤	步骤说明	费用说明
步骤三：创建算力密集型应用并上传SWR	创建算力密集型应用，用于压力测试。创建完成的应用需要上传至容器镜像服务（SWR），以便在集群中进行部署和管理。	涉及创建ECS实例，会产生云服务器和弹性公网IP等费用，具体请参见**ECS计费说明**。
步骤四：使用hpa-example镜像创建工作负载	使用构建的hpa-example镜像创建工作负载，并为该工作负载创建负载均衡类型服务，使您能够从公网访问应用。	涉及Pod和弹性负载均衡（ELB）费用，具体请参见**集群计费说明**和**ELB计费说明**。
步骤五：创建HPA策略	创建HPA策略并与工作负载关联，从而控制工作负载创建的Pod实例的数量。	不涉及费用。
步骤六：验证工作负载能否自动进行弹性伸缩	验证配置的HPA策略是否有效，即工作负载hpa-example能否自动进行弹性伸缩。	工作负载弹性伸缩时，会涉及Pod实例数量的增加与减少，会涉及Pod费用的变更，具体请参见**集群计费说明**。
后续操作：释放资源	如果您在完成实践后不需要继续使用集群，请及时清理资源以免产生额外扣费。	不涉及费用。

准备工作

- 您需要注册华为账号并完成实名认证，详情请参见**注册华为账号并开通华为云**和**个人实名认证**。
- 您需要确保账号有足够的资金，以免创建集群失败，具体操作请参见**账号充值**。

步骤一：首次开通 CCE 并进行授权

由于CCE在运行中对计算、存储、网络以及监控等各类云服务资源都存在依赖关系，因此当您首次登录CCE控制台时，CCE将自动请求获取当前区域下的云资源权限，从而更好地为您提供服务。如果您在当前区域已完成授权，可忽略本步骤。

步骤1 使用华为账号登录**CCE控制台**。

步骤2 单击控制台左上角的  ，选择“区域”，本示例使用区域为“华东-上海一”。

步骤3 在首次登录某个区域的CCE控制台时将跳出“授权说明”，请您在仔细阅读后单击“确定”。

当您同意授权后，CCE将在IAM中创建名为“cce_admin_trust”委托，统一对您的其他云服务资源进行操作，并且授予其Tenant Administrator权限。Tenant Administrator拥有除IAM管理外的全部云服务管理员权限，用于对CCE所依赖的其他云服务资源进行调用，且该授权仅在当前区域生效。

须知

CCE对其他云服务存在依赖，如果没有Tenant Administrator权限，可能会因为某服务权限不足而无法正常使用。因此，在使用CCE服务期间，请不要自行删除或者修改“cce_admin_trust”委托。

----结束

步骤二：创建 CCE Autopilot 集群

在CCE服务中创建CCE Autopilot集群，以提供更简化的Kubernetes服务。本示例中仅解释必要参数，其他参数保留默认值。关于其他参数的详细说明，请参见**购买Autopilot集群**。

步骤1 进入**CCE控制台**。

- 如果您的账号在当前区域未创建过集群，请在当前页面单击“购买集群”或“购买CCE Autopilot集群”，进入购买页。
- 如果您的账号在当前区域已创建过集群，请在左侧菜单栏选择集群管理，单击右上角“购买集群”，进入购买页。

步骤2 配置集群基础信息，具体的参数示例请参见**图2-2**和**表2-3**。

图 2-2 集群基本信息

表 2-3 集群基础信息

参数	示例	参数说明
集群类型	**CCE Autopilot集群**	CCE支持创建多种类型集群，满足各种业务需求，提供高可靠、安全的商业级容器集群服务。 • CCE Standard集群：标准版本集群，提供高可靠、安全的商业级容器集群服务。 • CCE Turbo集群：拥有更高性能的云原生网络，提供云原生混部调度能力，可实现更高的资源利用率和更广的全场景覆盖。 • CCE Autopilot集群：Serverless版集群，提供免运维的容器服务，可以大幅降低运维成本，提高应用程序的可靠性和可扩展性。 了解集群类型的更多内容，请参见**集群对比**。

参数	示例	参数说明
集群名称	**autopilot-example**	新建集群的名称。 集群名称长度范围为4-128个字符，以小写字母开头，支持小写字母、数字和中划线(-)，不能以中划线(-)结尾。
企业项目	**default**	该参数仅对开通企业项目的企业客户账号显示，不显示时请忽略。 企业项目是一种资源管理单位，可跨区域归类资源，方便企业按部门或项目组集中管理。了解企业项目的更多内容，请参见**项目管理**。 请根据需要选择适合的企业项目，如果没有特殊要求，可以选择default。
集群版本	**v1.28**	集群安装的Kubernetes软件版本，建议选择最新版本。

步骤3 配置集群网络信息，具体的参数示例请参见**图2-3**和**表2-4**。

图 2-3 集群网络信息

表 2-4 集群网络信息

参数	示例	参数说明
虚拟私有云	**vpc-autopilot**	选择集群所在的虚拟私有云（VPC）。如果没有可选项，单击右侧“新建虚拟私有云”创建，具体请参见**创建虚拟私有云和子网**。集群创建后，VPC不支持修改。
容器子网	**subnet-502f**	选择容器所在子网。每个Pod需要唯一的IP地址，容器子网内IP地址的数量决定了集群中Pod的数量上限，进而决定容器的数量上限，集群创建后支持新增子网。 如果没有可选项，请单击右侧“新建子网”创建，具体请参见**创建虚拟私有云和子网**。
服务网段	**10.247.0.0/16**	同一集群下容器互相访问时使用的Service资源网段，决定Service资源数量的上限。集群创建后，服务网段不支持修改。

参数	示例	参数说明
配置SNAT	**开启**	默认开启，开启后您的集群可以通过NAT网关访问公网。默认使用所选VPC中已有的NAT网关，若VPC没有NAT网关，系统将会为您自动创建一个默认规格的NAT网关并绑定弹性公网IP，同时自动配置SNAT规则。 使用NAT网关将产生一定费用，详情请参见**NAT网关计费说明**。

步骤4 单击“下一步：插件选择”，选择创建集群时需要安装的插件。

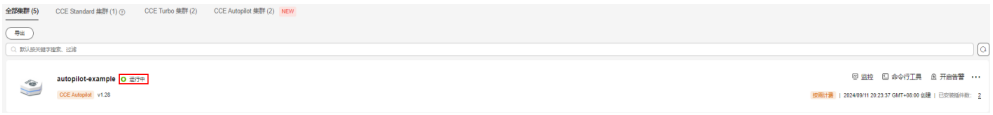
本示例中，仅选择默认安装插件，即CoreDNS域名解析和Kubernetes Metrics Server插件。

步骤5 单击“下一步：插件配置”，对选择的插件进行配置，默认插件无需配置。

步骤6 单击“下一步：确认配置”，显示集群资源清单，确认无误后，单击“提交”。

创建集群预计需要5-10分钟，请耐心等待。创建成功后，集群管理中对应集群的状态为运行中。

图 2-4 集群运行中



----结束

步骤三：创建算力密集型应用并上传 SWR

创建算力密集型应用主要用于压力测试，以验证部署的HPA策略是否能够根据指定指标自动调整Pod实例的数量。创建完成的应用需要上传至容器镜像服务（SWR），以便在集群中进行部署和管理。

步骤1 创建一台Linux系统的ECS虚拟机，该ECS与**步骤二：创建CCE Autopilot集群**中的集群处于同一VPC，并绑定弹性公网IP，具体操作请参见**快速购买和使用Linux ECS**。如果您已有符合要求的ECS，可以跳过本步骤。

本示例以“CentOS 7.9 64bit(40GiB)”操作系统为例，为您演示如何基于PHP创建算力密集型应用并上传SWR。

步骤2 登录ECS虚拟机，具体操作请参见**通过CloudShell登录Linux ECS**。

步骤3 安装Docker。

1. 执行以下命令，检查ECS是否安装Docker。如果已安装可以跳过安装Docker的步骤，否则请继续执行下面的安装、运行和检查的步骤。
   ```
   docker --version
   ```
   回显如下，则说明未安装Docker。
   ```
   -bash: docker: command not found
   ```
2. 执行以下命令，安装并运行Docker。
   ```
   yum install docker
   ```

设置Docker在系统启动时自动启动，并立即开始运行。

```
systemctl enable docker
systemctl start docker
```

3. 执行以下命令，检查安装结果。

```
docker --version
```

回显如下，则说明Docker安装成功。

```
Docker version 1.13.1, build 7d71120/1.13.1
```

步骤4 创建算力密集型应用，并制作镜像。

1. 执行以下命令，创建一个名为index.php的PHP文件，文件名称可自定义。该文件描述的含义为：在接收到用户请求时，先进行1000000次平方根循环计算，即“0.0001+0.01+0.1+...”，然后返回“OK!”。

说明

Linux文件命名支持字母、数字、下划线和连字符，但不能包含斜杠（/）和空字符（\0）。文件名区分大小写，建议避免使用特殊字符，如空格、问号（?）和星号（*）等，以提高兼容性。

```
vim index.php
```

文件内容如下：

```
<?php
  $x = 0.0001;
  for ($i = 0; $i <= 1000000; $i++)
    $x += sqrt($x);
  }
  echo "OK!";
?>
```

输入完成后，按**Esc键**退出编辑，输入**:wq**保存。

2. 执行以下命令，编写Dockerfile制作镜像。

```
vim Dockerfile
```

文件内容如下：

```
# 使用PHP的Docker官方镜像
FROM php:5-apache
# 将本地的index.php文件复制到容器指定目录，这个文件将被用作默认的网页
COPY index.php /var/www/html/index.php
# 修改index.php文件的权限，使其对所有用户可读和可执行，确保文件在Web服务器上能够被正确访问
RUN chmod a+rx index.php
```

3. 执行如下命令构建镜像，镜像名称为hpa-example，版本为latest，名称和版本可自定义。

```
docker build -t hpa-example:latest .
```

回显内容如下，则说明镜像构建成功。

```
Sending build context to Docker daemon  158.1MB
Step 1/3 : FROM php:5-apache
...
Successfully built xxx
Successfully tagged hpa-example:latest
```

步骤5 将创建的hpa-example镜像上传至SWR。

1. 登录**SWR控制台**，在右侧单击“组织管理”，单击页面右上角“创建组织”，创建镜像组织。在创建组织界面输入“组织名称”，单击“确定”。若已有组织，该步骤可以跳过。

图 2-5 创建组织

2. 在左侧导航栏中选择“我的镜像”，在右侧单击“客户端上传”。在弹出的页面中单击“生成登录指令”，单击，复制“临时登录指令”，用于ECS登录SWR。

   在ECS执行复制的登录指令，如下所示。

   ```
   docker login -u cn-east-3xxx swr.cn-east-3.myhuaweicloud.com
   ```

   回显如下，说明登录成功。

   ```
   Login Succeeded
   ```

3. 为hpa-example镜像添加标签，代码结构如下：

   docker tag {镜像名称1:版本名称1} {镜像仓库地址}/{组织名称}/{镜像名称2:版本名称2}

   具体示例：

   ```
   docker tag hpa-example:latest swr.cn-east-3.myhuaweicloud.com/test/hpa-example:latest
   ```

   表 2-5 镜像添加标签参数说明

参数	示例	参数说明
{镜像名称1:版本名称1}	**hpa-example:latest**	请替换为本地所要上传的实际镜像的名称和版本名称。
{镜像仓库地址}	**swr.cn-east-3.myhuaweicloud.com**	请替换为**登录指令**中的末尾域名，该域名即为镜像仓库地址。
{组织名称}	**test**	请替换为**已创建的镜像组织**。
{镜像名称2:版本名称2}	**hpa-example:latest**	请替换为SWR镜像仓库中需要显示的镜像名称和镜像版本，您可以自定义。

4. 上传镜像至镜像仓库，代码结构如下：

   docker push {镜像仓库地址}/{组织名称}/{镜像名称2:版本名称2}

   具体示例：

   ```
   docker push swr.cn-east-3.myhuaweicloud.com/test/hpa-example:latest
   ```

回显结果如下，则说明上传镜像成功。

```
6d6b9812c8ae: Pushed
...
fe4c16cbf7a4: Pushed
latest: digest: sha256:eb7e3bbdxxx size: xxx
```

返回**SWR控制台**，在“我的镜像”页面，执行刷新操作后可查看到对应的镜像信息。

----结束

步骤四：使用 hpa-example 镜像创建工作负载

使用构建的hpa-example镜像创建工作负载，并为该工作负载创建负载均衡类型服务，使您能够从公网访问应用。本节介绍两种方式创建该工作负载，包括控制台方式和kubectl命令行方式。

使用控制台方式

步骤1 返回**CCE控制台**。单击新建的集群名称，进入集群控制台。

步骤2 在左侧菜单栏中选择“工作负载”，单击右上角“创建工作负载”，进入创建页。

步骤3 配置工作负载基本信息，具体参数示例请参见**图2-6**和**表2-6**。

本示例中仅解释必要参数，其他参数保留默认值。关于其他参数的详细说明，请参见**创建工作负载**，您可以根据工作负载类型选择适合的参考文档。

图 2-6 工作负载基本信息

表 2-6 工作负载基本信息

参数	示例	参数说明
负载类型	**无状态负载** **Deployment**	工作负载是一种对Pod的抽象管理方式，用于定义和控制Pod的创建、运行和生命周期。通过工作负载，您可以批量管理和自动化控制多个Pod的行为，如伸缩、更新和恢复。 • 无状态负载（Deployment）：管理无状态应用，支持上线部署、滚动升级、创建副本和恢复上线。 • 有状态负载（StatefulSet）：管理有状态应用，确保每个Pod能够拥有独立的持久化状态，并能够在Pod重启或迁移时恢复其数据，以保障应用的可靠性和一致性。 • 普通任务（Job）：一次性任务，完成后Pod自动删除。 • 定时任务（CronJob）：基于时间的Job，指定时间周期内运行指定的Job。 了解工作负载的更多内容，请参见**工作负载概述**。 hpa-example主要用于压力测试，不需要在本地保存任何持久性数据，因此本示例将hpa-example部署为无状态负载。
负载名称	**hpa-example**	请填写工作负载的名称。 工作负载名称长度范围为1-63个字符，可以包含小写英文字母、数字和和中划线(-)，并以小写英文字母开头，小写英文字母或数字结尾。
命名空间	**default**	命名空间是Kubernetes集群中的抽象概念，可以将集群中的资源或对象划分为一个组，且不同命名空间中的数据彼此隔离，您可以根据需要创建并使用命名空间。 集群创建后会默认生成default命名空间，如果没有特殊要求，可以直接选择default命名空间。
实例数量	**1**	工作负载中Pod的数量。Pod实例数量的设置策略： • 高可用性：如果您需要保证工作负载的高可用性，则实例数量至少设置为2，避免单点故障。 • 性能要求：您需要根据工作负载的流量和资源需求设置实例数量，避免过载或资源浪费。 本示例仅做演示，实例数量设置为1。

步骤4 配置工作负载容器信息，具体参数示例请参见**图2-7**和**表2-7**。

本示例中仅解释必要参数，其他参数保留默认值。关于其他参数的详细说明，请参见**创建工作负载**，您可以根据工作负载类型选择适合的参考文档。

图 2-7 工作负载容器信息

表 2-7 工作负载容器信息

参数	示例	参数说明
镜像名称	**hpa-example**	单击“选择镜像”，在弹出的窗口中切换至“我的镜像”，选择上传的hpa-example镜像。
镜像版本	**latest**	选择需要部署的镜像版本。
CPU配额	**0.25cores**	CPU资源限制值，即允许容器使用的CPU最大值，防止占用过多资源，默认0.25cores。
内存配额	**512MiB**	内存资源限制值，即允许容器使用的内存最大值。如果超过，容器会被终止，默认512MiB。

步骤5 单击“服务配置”下的＋，进入创建服务页面，配置工作负载服务信息，具体参数示例请参见**图2-8**和**表2-8**。

本示例仅解释必要参数，其他参数保留默认值。关于其他参数的详细说明，请参见**服务（Service）**，您可以根据服务类型选择适合的参考文档。

图 2-8 工作负载服务信息

表 2-8 工作负载服务信息

参数	示例	参数说明
Service名称	**hpa-example**	请填写服务的名称。 服务名称长度范围为1-63个字符，可以包含小写英文字母、数字和和中划线(-)，并以小写英文字母开头，小写英文字母或数字结尾。
访问类型	**负载均衡**	选择服务类型，即服务访问的方式。 • 集群内访问：通过集群的内部IP暴露服务，只能够在集群内部访问。 • 负载均衡：通过弹性负载均衡（ELB）对外部提供服务，即能够从公网访问到工作负载。 本示例中需要外部访问hpa-example，所以访问类型设置为负载均衡。
负载均衡器	• **独享型** • **网络型(TCP/UDP)&应用型(HTTP/HTTPS)** • **选择已有** • **elb-hpa-example**	• 如果已有弹性负载均衡（ELB）实例，可以选择已有ELB实例。 **说明** 使用已有的ELB时，ELB实例需要具备3个条件： – 与集群属于同一VPC。 – 实例类型为独享型。 – 网络类型必须支持私网（存在私有地址）。 • 如果没有弹性负载均衡（ELB）实例，请选择“自动创建”，创建一个负载均衡器，并绑定弹性公网IP，具体操作请参见**创建负载均衡类型的服务**。
端口配置	**协议：TCP**	负载均衡监听器端口协议。
	容器端口：80	容器中应用启动监听的端口，该容器端口需和应用对外提供的监听端口一致。
	服务端口：80	ELB将会使用该端口创建监听器，提供外部流量访问入口，可自定义。

步骤6 单击右下角“创建工作负载”。

创建成功后，无状态工作负载列表中对应工作负载的状态为运行中。

图 2-9 工作负载运行中

步骤7 单击hpa-example工作负载名称，进入工作负载详情页，获取hpa-example的外部访问地址。在“访问方式”页签下可以看到hpa-example的IP地址，“负载均衡公网IP:

访问端口”即为外部访问地址。其中，“负载均衡公网IP”为**步骤5**中ELB实例的公网地址，“访问端口”为**步骤5**中的“服务端口”。

图 2-10 访问方式

----结束

使用 kubectl 命令行方式

步骤1 在ECS中安装kubectl命令行工具。您可以尝试执行**kubectl version**命令判断是否已安装kubectl，如果已经安装kubectl，则可跳过此步骤。

1. 执行以下命令，下载kubectl。
   ```
   cd /home
   curl -LO https://dl.k8s.io/release/{v1.28.0}/bin/linux/amd64/kubectl
   ```
   其中{v1.28.0}为指定的版本号，请根据集群版本进行替换。
2. 执行以下命令，安装kubectl。
   ```
   chmod +x kubectl
   mv -f kubectl /usr/local/bin
   ```

步骤2 为kubectl命令行工具配置访问Kubernetes集群的凭证。

1. 返回**CCE控制台**，并单击集群名称进入集群总览页。
2. 在集群总览页中找到“连接信息”模块。单击kubectl后的“配置”，查看kubectl的连接信息。
3. 在弹出页面中选择“内网访问”，然后下载对应的配置文件。
4. 登录已安装kubectl工具的ECS，将上一步中下载的配置文件（以kubeconfig.yaml为例）复制到/home目录下。
5. 将kubectl认证文件保持至$HOME/.kube目录下的config文件中。
   ```
   cd /home
   mkdir -p $HOME/.kube
   mv -f kubeconfig.yaml $HOME/.kube/config
   ```
6. 执行kubectl命令验证集群的连通性。

   以查看集群信息为例，执行以下命令。
   ```
   kubectl cluster-info
   ```
   回显内容如下，则说明集群连接成功。
   ```
   Kubernetes control plane is running at https://xx.xx.xx.xx:5443
   CoreDNS is running at https://xx.xx.xx.xx:5443/api/v1/namespaces/kube-system/services/coredns:dns/proxy
   To further debug and diagnose cluster problems, use 'kubectl cluster-info dump'.
   ```

步骤3 执行以下命令，创建YAML文件hpa-example.yaml，用于配置hpa-example工作负载，文件名称可自定义。
```
vim hpa-example.yaml
```

文件内容如下：
```
apiVersion: apps/v1
kind: Deployment
```



```
metadata:
  name: hpa-example     # 工作负载名称
spec:
  replicas: 1     # 实例数量
  selector:
    matchLabels:  # 选择器，用于选择带有特定标签的资源
      app: hpa-example
  template:
    metadata:
      labels:     # 标签
        app: hpa-example
    spec:
      containers:
      - name: container-1
        image: 'swr.cn-east-3.myhuaweicloud.com/test/hpa-example:latest' # 替换为您上传到SWR的镜像地址
        resources:
          limits:  # limits与requests建议取值保持一致，避免扩缩容过程中出现震荡
            cpu: 250m
            memory: 512Mi
          requests:
            cpu: 250m
            memory: 512Mi
      imagePullSecrets:
      - name: default-secret
```

说明

“swr.cn-east-3.myhuaweicloud.com/test/hpa-example:latest”需替换为您上传到SWR的镜像地址，镜像地址即为步骤**上传镜像至镜像仓库**中docker push后内容。

输入完成后，按**Esc键**退出编辑，输入**:wq**保存。

步骤4 执行以下命令，创建工作负载。

```
kubectl create -f hpa-example.yaml
```

回显如下，表示已经开始创建工作负载。

```
deployment.apps/hpa-example created
```

步骤5 执行以下命令，查看工作负载状态。

```
kubectl get deployment
```

回显如下，如果READY值为1/1，则说明工作负载创建的Pod皆为可用状态，即创建成功。

```
NAME          READY   UP-TO-DATE   AVAILABLE   AGE
hpa-example   1/1     1            1           4m59s
```

回显内容的参数说明如下：

表 2-9 回显内容参数说明

参数	示例	参数说明
NAME	**hpa-example**	工作负载的名称。
READY	**1/1**	表示工作负载的可用状态，显示为“可用Pod个数/期望Pod个数”。
UP-TO-DATE	**1**	指当前工作负载已经完成更新的Pod数。

参数	示例	参数说明
AVAILABLE	**1**	工作负载可用的Pod个数。
AGE	**4m59s**	工作负载已经运行的时间。

步骤6 执行以下命令，创建YAML文件nginx-elb-svc.yaml，用于配置负载均衡服务并关联已创建的工作负载hpa-example，文件名称可自定义。

本示例基于已有的弹性负载均衡（ELB）实例创建服务，如果您需要自动创建ELB请参考**通过kubectl命令行创建-自动创建ELB**。

```
vim hpa-example-elb-svc.yaml
```

文件内容如下：

```
apiVersion: v1
kind: Service
metadata:
  name: hpa-example      # 服务的名称
  annotations:
    kubernetes.io/elb.id: <your_elb_id>          # ELB ID，替换为实际值
    kubernetes.io/elb.class: performance          # 负载均衡器类型
spec:
  selector:             # selector取值需与步骤3中工作负载YAML文件的matchLabels参数取值一致
    app: hpa-example
  ports:
  - name: service0
    port: 80
    protocol: TCP
    targetPort: 80
  type: LoadBalancer
```

输入完成后，按**Esc键**退出编辑，输入**:wq**保存。

表 2-10 使用已有 ELB 参数说明

参数	示例	参数说明
kubernetes.io/elb.id	**405ef586-0397-45c3-bfc4-xxx**	已有的ELB的ID。 **说明** 使用已有的ELB时，ELB实例需要具备3个条件： • 与集群属于同一VPC。 • 实例类型为独享型。 • 网络类型必须支持私网（存在私有地址）。 **获取方式：**进入网络控制台，选择“弹性负载均衡 > 我的ELB”，找到对应的ELB名称，名称下方即为对应ID。同时，单击ELB实例名称，在“基本信息”页签中验证该ELB是否满足上述条件。
kubernetes.io/elb.class	**performance**	负载均衡器类型，仅支持performance类型，即独享型负载均衡。
selector	**app: hpa-example**	选择器，服务将流量发送给对应标签的Pod。

参数	示例	参数说明
ports.port	**8080**	弹性负载均衡（ELB）将会使用该端口创建监听器，提供外部流量访问入口，可自定义。
ports.protocol	**TCP**	负载均衡监听器端口协议。
ports.targetPort	**80**	Service访问目标容器的端口，此端口与容器中运行的应用强相关。

步骤7 执行以下命令创建服务。

```
kubectl create -f hpa-example-elb-svc.yaml
```

回显如下，表示服务已创建。

```
service/hpa-example created
```

步骤8 执行以下命令查看服务。

```
kubectl get svc
```

回显内容如下，表示工作负载访问方式已设置成功。您可以通过“外部访问地址：服务端口”访问该工作负载，其中外部访问地址为EXTERNAL-IP对应的第一个IP地址的，服务端口为8080。

```
NAME          TYPE           CLUSTER-IP     EXTERNAL-IP                PORT(S)          AGE
kubernetes    ClusterIP      10.247.0.1     <none>                     443/TCP          18h
hpa-example   LoadBalancer   10.247.56.18   xx.xx.xx.xx,xx.xx.xx.xx    8080:30581/TCP   5m8s
```

----结束

步骤五：创建 HPA 策略

HPA策略（Horizontal Pod Autoscaler）主要用来控制Pod的水平伸缩，通过周期性监测Pod的相关指标，计算实现HPA策略中资源目标值所需的副本数量，进而调整工作负载创建的Pod实例数量。本节介绍两种方式创建HPA策略，包括控制台方式和kubectl命令行方式。

使用控制台方式

本示例中仅解释必要参数，其他参数保留默认值。关于默认参数的详细说明，请参见**Pod水平自动扩缩**。

步骤1 返回集群总览页，左侧菜单栏选择“策略”，单击右上角“创建HPA策略”，进入创建页。

步骤2 填写策略基础信息，具体参数说明请参见**表2-11**。

图 2-11 策略基本信息

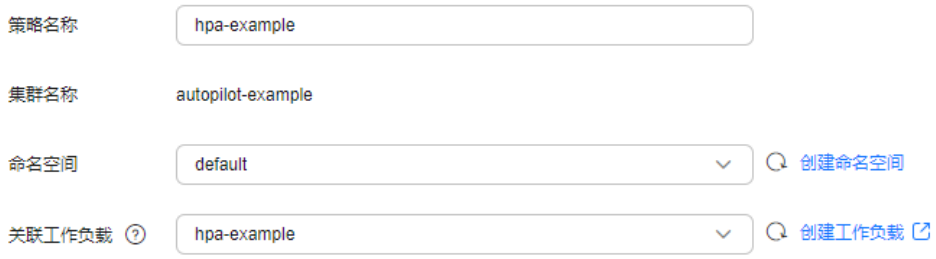


表 2-11 策略基本信息

参数	示例	参数说明
策略名称	**hpa-example**	请填写策略的名称。
命名空间	**default**	命名空间是Kubernetes集群中的抽象概念，可以将集群中的资源或对象划分为一个组，且不同命名空间中的数据彼此隔离，您可以根据需要创建并使用命名空间。 集群创建后会默认生成default命名空间，如果没有特殊要求，可以直接选择default命名空间。
关联工作负载	**hpa-example**	请选择**步骤四：使用hpa-example镜像创建工作负载**中创建的工作负载。 **说明** 弹性伸缩策略会根据名称匹配关联工作负载，若相同命名空间下的多个工作负载中app的label相同，会导致弹性伸缩策略不符合预期。

步骤3 进行“策略配置”，具体参数示例请参见**表2-12**。

图 2-12 策略配置

表 2-12 策略配置

参数	示例	参数说明
实例范围	**1-10**	表示工作负载创建的Pod实例的伸缩范围。
伸缩配置	**系统默认**	用于选择伸缩配置策略，有以下两种配置策略： ● 系统默认：稳定窗口和步长等策略直接采用K8S社区推荐的默认行为，更多信息请参见**Pod水平自动扩缩**。 ● 自定义：用户可以自定义稳定窗口和步长等策略实现更灵活的配置，未配置的参数将采用社区推荐的默认值，更多信息请参见**Pod水平自动扩缩**。

参数	示例	参数说明
系统策略	**指标：CPU利用率**	HPA可以通过跟踪相关指标的资源用量，触发Pod实例的扩缩操作。 • CPU利用率：工作负载容器组（Pod）的实际使用量/申请量。 • 内存利用率：工作负载容器组（Pod）的实际使用量/申请量。
	期望值：50%	表示所选指标的期望值，可以用来计算需要伸缩的实例数，即“需要伸缩的实例数=（当前指标值/期望值）×当前实例数”。
	容忍范围：30%-70%	表示伸缩的触发范围，当指标处于范围内时不会触发伸缩，可以避免轻微波动引发的频繁调整，保证系统的稳定性。需要注意的是，期望值必须在容忍范围内。

步骤4 单击“创建”。您可以在“HPA策略”列表中看到已创建的策略，最新状态为“已启动”。

图 2-13 HPA 策略已启动

----结束

使用 kubectl 命令行方式

步骤1 执行以下命令，创建YAML文件hpa-policy.yaml，用于配置HPA策略，文件名称可自定义。

```
vim hpa-policy.yaml
```

文件内容如下：

```
apiVersion: autoscaling/v2
kind: HorizontalPodAutoscaler
metadata:
  name: hpa-example   # HPA策略的名称
  namespace: default  # 命名空间
  annotations:
    # 定义伸缩的触发条件：CPU利用率在30%到70%之间时，不会进行伸缩操作，反之，则会进行伸缩操作。
    extendedhpa.metrics: '[{"type":"Resource","name":"cpu","targetType":"Utilization","targetRange":
{"low":"30","high":"70"}}]'
spec:
  scaleTargetRef:     # 关联的工作负载信息
    kind: Deployment
    name: hpa-example
    apiVersion: apps/v1
  minReplicas: 1     # 工作负载内Pod数量的最小阈值
  maxReplicas: 10   # 工作负载内Pod数量的最大阈值
```



```
metrics:
  - type: Resource  # 设置监控的资源
    resource:
      name: cpu     # 设置资源为cpu，还可以设置为memory
      target:
        type: Utilization  # 设置指标为利用率
        averageUtilization: 50  # HPA控制器会维持伸缩目标中的Pod的平均资源利用率在50%
```

输入完成后，按**Esc键**退出编辑，输入**:wq**保存。

步骤2 执行以下命令创建HPA策略。

```
kubectl create -f hpa-policy.yaml
```

回显如下，表示策略已创建。

```
horizontalpodautoscaler.autoscaling/hpa-example created
```

步骤3 执行以下命令查看创建的HPA策略。

```
kubectl get hpa
```

回显结果如下：

```
NAME          REFERENCE                TARGETS         MINPODS   MAXPODS   REPLICAS   AGE
hpa-example   Deployment/hpa-example   <unknown>/50%   1         10        0          10s
```

----结束

步骤六：验证工作负载能否自动进行弹性伸缩

本节介绍两种方式检查工作负载hpa-example能否自动进行弹性伸缩，包括控制台方式和kubectl命令行方式。

使用控制台方式

步骤1 登录**步骤三：创建算力密集型应用并上传SWR**中使用的ECS。

步骤2 执行以下命令循环访问工作负载，其中“ip:port”为工作负载的访问地址，与**步骤7**中的“负载均衡公网IP：访问端口”一致。

```
while true;do wget -q -O- http://ip:port; done
```

回显结果如下：

```
OK!
OK!
...
```

步骤3 返回**CCE控制台**，单击新建的集群名称，进入集群控制台。左侧导航栏单击“工作负载”，单击“工作负载名称”。单击搜索栏右侧，观察工作负载自动扩容过程，可以发现工作负载hpa-example创建的Pod实例个数由1个扩容至2个。

图 2-14 工作负载自动扩容

步骤4 返回ECS，利用**Ctrl+c**停止访问工作负载，观察工作负载自动缩容过程。

返回控制台界面，单击搜索栏右侧 。工作负载缩容的稳定窗口默认为300秒，即缩容策略成功触发后，不会立即对目标进行缩容，需要先冷却300秒，以防止短期波动造成影响

图 2-15 工作负载缩容

----结束

使用 kubectl 命令行方式

步骤1 执行以下命令循环访问工作负载，其中“ip:port”为工作负载的访问地址，与**步骤8**中的“外部访问地址：服务端口”相对应。

```
while true;do wget -q -O- http://ip:port; done
```

回显结果如下：

```
OK!
OK!
...
```

步骤2 在页面左上方单击“终端 > 新建会话”，新建一个服务器页面，执行以下命令观察工作负载自动扩容过程。

```
kubectl get hpa hpa-policy --watch
```

回显结果如下：

```
NAME          REFERENCE                TARGETS    MINPODS   MAXPODS   REPLICAS   AGE
hpa-example   Deployment/hpa-example   24%/50%    1         10        1          17m
hpa-example   Deployment/hpa-example   100%/50%   1         10        1          17m
hpa-example   Deployment/hpa-example   100%/50%   1         10        2          18m
hpa-example   Deployment/hpa-example   100%/50%   1         10        2          18m
hpa-example   Deployment/hpa-example   57%/50%    1         10        2          19m
hpa-example   Deployment/hpa-example   57%/50%    1         10        2          20m
...
```

表 2-13 回显结果参数说明

参数	示例	说明
NAME	**hpa-example**	HPA策略的名称。 本示例中HPA策略名称为hpa-example。
REFERENCE	**Deployment/hpa-example**	HPA策略管理的对象。 本示例中为无状态工作负载hpa-example。
TARGETS	**24%/50%**	HPA策略监控指标的当前值和期望值。 本示例中监控指标为CPU利用率，24%表示CPU利用率的当前值，50%表示CPU利用率的期望值。

参数	示例	说明
MINPODS	**1**	HPA策略允许的最小Pod实例数量。 本示例中允许的最小Pod实例数量为1，当Pod实例数量为1时，即使当前CPU利用率低于容忍范围，也不会进行缩容。
MAXPODS	**10**	HPA策略允许的最大Pod实例数量。 本示例中允许的最大Pod实例数量为10，当Pod实例数量为10时，即使当前CPU利用率高于容忍范围，也不会进行扩容。
REPLICAS	**1**	当前实际运行的Pod实例数量。
AGE	**17m**	HPA策略已存在的时间。

- 第2条记录中，工作负载的CPU利用率为100%，超过70%，触发弹性伸缩。
- 第3条记录中，工作负载将Pod实例数量由1扩容至2。
- 第5条记录中，工作负载的CPU利用率降至57%，属于30%~70%，并未触发弹性伸缩。

步骤3 在**步骤1**所在的服务器页面中利用**Ctrl+c**停止访问工作负载，观察负载自动缩容过程。

在**步骤2**所在的服务器页面中，继续观察回显结果。值得注意的是，工作负载缩容的稳定窗口默认为300秒，即缩容策略成功触发后，不会立即对目标进行缩容，需要先冷却300秒，以防止短期波动造成影响。

```
NAME          REFERENCE                TARGETS   MINPODS   MAXPODS   REPLICAS   AGE
...
hpa-example   Deployment/hpa-example   19%/50%   1         10        2          30m
hpa-example   Deployment/hpa-example   0%/50%    1         10        2          31m
hpa-example   Deployment/hpa-example   0%/50%    1         10        2          35m
hpa-example   Deployment/hpa-example   0%/50%    1         10        1          36m
...
```

- 第1条记录中，工作负载的CPU利用率为19%，低于30%，触发弹性伸缩。但此时处于冷却时间，因此工作负载并未立即缩容。
- 第4条记录中，工作负载将Pod实例数量由2缩容至1，达到最低实例数量，不再触发弹性伸缩。

----结束

后续操作：释放资源

如果您无需继续使用集群和ECS，请及时释放资源，避免产生额外的费用。

须知

- 删除集群会删除集群下工作负载和服务，相关业务将无法恢复。
- 集群关联创建的VPC级别的资源（如终端节点、NAT网关和SNAT出网EIP等），删除集群时默认保留，请确认其他集群或者地方没有重用该资源，再进行删除操作。

步骤1 进入**CCE控制台**，在左侧导航栏中选择“集群管理”。

步骤2 找到需要删除的集群，单击集群卡片右上角的“...”，并单击“删除集群”。

步骤3 在弹出的“删除集群”窗口中，根据页面提示删除相关资源。

在确认框中输入“DELETE”，单击“是”，开始执行删除集群操作。

删除集群需要花费1-3分钟，请耐心等候。集群列表中对应集群名称消失，则说明删除集群成功。

步骤4 进入**云服务器控制台**，左侧导航栏单击“弹性云服务器”，找到对应的ECS，右侧单击“更多”，单击“删除”。

在删除界面中勾选“删除云服务器绑定的弹性公网IP地址”和“删除云服务器挂载的数据盘”，单击“下一步”。

图 2-16 删除 ECS

在确认框中输入“DELETE”，单击“确定”，开始执行删除ECS操作。

删除ECS需要花费0.5-1分钟，请耐心等候。ECS列表中对应ECS名称消失，则说明删除ECS成功。

图 2-17 删除 ECS 确认

----结束