

云应用引擎

目录

- 产品介绍..... 5
 - 什么是云应用引擎..... 5
 - 产品架构..... 5
 - 产品功能..... 5
 - 产品计费..... 6
 - 产品优势..... 6
 - 应用全托管免运维..... 6
 - 精细化成本管理..... 6
 - 丰富的弹性策略..... 6
 - 低门槛容器化，全套微服务治理..... 7
 - 开放标准，自主无忧..... 7
 - 安全可靠，专注业务..... 7
 - 生态集成，开箱即用..... 7
 - 功能特性..... 7
 - CAE 应用托管..... 7
 - 开源自建对比..... 9
 - 基础场景..... 9
 - 应用场景..... 10
 - 应用托管..... 10
 - 任务托管..... 10
 - 使用限制..... 11
 - 配额限制..... 11
 - 基本概念..... 12
 - 基本概念总览..... 12
 - 命名空间..... 12
 - 应用..... 12
 - 应用实例..... 12
 - 微服务应用..... 12
 - 任务..... 12
 - 任务实例..... 13
 - 任务模板..... 13
 - 任务记录..... 13
 - 启动命令..... 13
 - 容器..... 13
 - 容器镜像..... 13
 - 开服地域..... 13
 - 公共云..... 14
 - 实例规格..... 14
- 快速入门..... 16
 - 新手指引..... 16
 - 背景信息..... 16
 - CAE使用流程..... 16
 - 准备工作..... 16
 - 步骤一：开通服务并授权（天翼云账号） 16
 - 步骤二：创建VPC..... 17

步骤三：创建命名空间.....	17
部署第一个 CAE 应用.....	18
部署应用.....	18
通过公网访问应用.....	19
清理资源.....	20
后续步骤.....	21
用户指南.....	22
命名空间.....	22
管理命名空间.....	22
对应用进行批量操作.....	24
应用管理.....	26
创建、查看与删除应用.....	26
应用部署.....	28
应用高级配置.....	36
复制应用.....	51
实例列表.....	52
变更记录.....	53
应用事件.....	54
日志管理.....	54
基础监控.....	55
回退历史版本.....	56
手动扩缩.....	57
弹性扩缩.....	58
启动应用.....	61
停止应用.....	61
为应用绑定ELB公网访问.....	62
配置管理.....	66
管理和使用配置项（ConfigMap）	66
管理和使用保密字典（Secret）	68
流量管理.....	71
网关路由.....	71
微服务治理.....	73
开启微服务治理.....	73
服务网格.....	74
微服务引擎.....	86
示例应用.....	105
JAVA-hellocae.....	105
golang-hellocae.....	106
JAVA-provider.....	107
JAVA-consumer.....	108
DotNet-hellocae.....	110
NodeJS-hellocae.....	110
python-hellocae.....	111
php-hellocae.....	112
常见问题.....	113
应用托管FAQ.....	113
如何处理常见的Pod问题？	113
CAE是否支持定时部署？	114

为什么创建应用后在控制台找不到应用?	114
如何修改应用部署方式, 例如将JAR包部署改为镜像部署?	114
如何在容器中安装一个ffmpeg命令?	114
微服务相关FAQ.....	114
CAE应用是否必须选择服务发现?	114
如何配置使用云应用引擎内置注册中心?	114
一个应用可以部署多个微服务吗?	115
为什么微服务上下线的时候流量有损?	115
最佳实践.....	116
使用CAE实现微服务应用注册发现与服务调用.....	116
方案概览.....	116
操作指引.....	117
部署应用.....	117
配置公网访问应用.....	118
验证结果.....	119
资源清理.....	120
通过命名空间实现开发环境、测试环境和生产环境的隔离.....	120
方案概览.....	120
操作指引.....	120
1. 创建命名空间.....	120
2. 部署应用.....	120
3. 批量启停应用.....	121

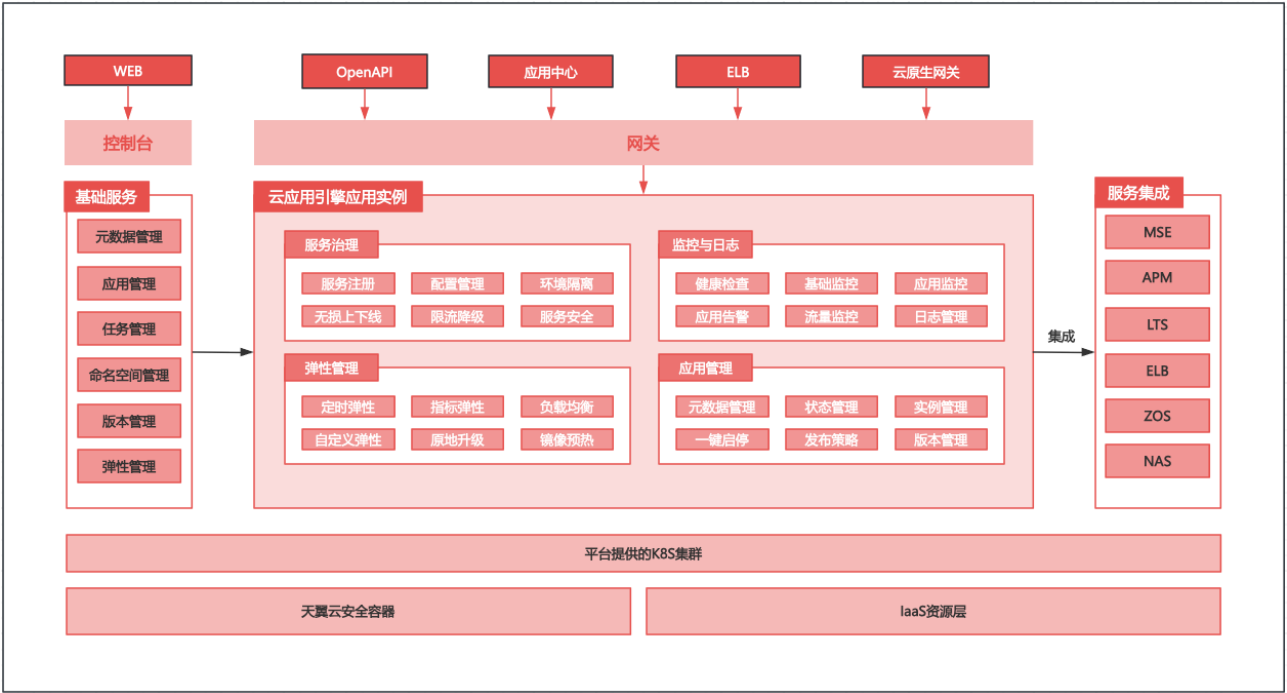
产品介绍

什么是云应用引擎

云应用引擎 CAE（Cloud App Engine）是一款面向容器应用的 Serverless 全托管平台，提供完整的微服务应用托管和治理能力。CAE 可帮助您将微服务应用便捷地容器化并托管，实现秒级部署与全自动弹性伸缩，按使用量计费，大幅提高业务交付效率和系统可用性。作为 Serverless 架构中承载复杂应用的核心组件，CAE 让您聚焦业务逻辑，免运维底层基础设施，开箱即用监控、日志、服务治理等企业级能力，全面提升应用的可管理性与弹性。

产品架构

CAE产品架构如下所示。更多信息，请参见 [基本概念](#)。



- 底层基于Kubernetes，实现了Serverless架构与微服务架构的完美结合。
- 支持Spring Cloud多种微服务框架、多种部署方式（ZIP包、镜像）和多种技术栈语言（Java、PHP、Python等）。

产品功能

功能	说明
应用全生命周期管理	提供从创建到运行的应用全生命周期管理服务，支持分批、灰度等多种发布策略，支持按流量秒级灰度。

功能	说明
无侵入应用监控	提供无侵入的应用监控和告警能力，支持任意语言和任意框架。
无侵入微服务治理	支持Spring Cloud零代码改造迁移至云应用引擎。提供服务注册与发现、环境隔离、配置管理、限流降级、应用无损上下线、服务鉴权、全链路灰度等能力。
一键启停开发测试环境	中大型企业包含多套环境，测试环境如果长期保有应用实例，会导致闲置浪费高。CAE提供逻辑隔离运行环境，通过一键启停开发测试环境能力，有效节省硬件成本。

产品计费

CAE支持按量计费的计费模式。关于CAE的计费详情，请参见 [计费概述](#)。

产品优势

云应用引擎 CAE（Cloud App Engine）是集成 IaaS + PaaS 的 Serverless PaaS 平台，支持多种方式部署应用和任务，支持主流微服务框架和开源任务框架。CAE采用最小化闲置资源原则，以秒级弹性伸缩性能为您打造高效的低成本服务。

应用全托管免运维

CAE采用一站式、全托管的应用管理界面，让您聚焦核心业务，专注于业务的逻辑开发，无需考虑底层资源，例如集群可用性、容量和性能等方面的问题，减少繁杂的运维管理。同时，CAE采用一站式、全托管的管理界面，为应用和任务提供开箱即用的生命周期管理、可观测等功能，帮您快速排查故障。

精细化成本管理

告别传统模式下为峰值流量而长期预留固定资源的低效做法。CAE让您转向按量付费，精准匹配应用的实际负载，并结合一键启停（如批量管理开发/测试环境），彻底消除资源闲置成本，实现企业级资源的精细管控与成本优化。

丰富的弹性策略

CAE 提供多层次、自动化的弹性策略，保障应用在负载波动下始终稳定运行、成本最优：

- 指标弹性：基于 CPU、内存等核心监控指标自动扩缩容，实时响应业务压力。
- 定时弹性：根据预设的日程（如工作日/大促）自动调整实例数，提前应对规律性流量。
- 混合弹性：可灵活组合定时与指标策略，实现对复杂场景的精细化、自动化容量管理。

低门槛容器化，全套微服务治理

CAE 让您免运维 K8s 基础设施，秒级完成从代码包、Docker 镜像部署任意语言的在线应用（如微服务、Web 服务），并自动弹性伸缩按量计费。平台内置免费 Nacos 注册中心，支持 Spring Cloud 无侵入迁移，开箱即用服务注册发现、限流降级、无损上下线及全链路灰度等全套微服务治理能力，极简助力业务容器化转型。

开放标准，自主无忧

基于标准容器与 Kubernetes 生态构建，避免厂商锁定。您的应用可平滑迁移至任意兼容 K8s 的平台，保障业务长期发展的灵活性与自主权。

安全可靠，专注业务

底层采用安全容器技术，网络层通过 VPC 实现租户级强隔离，为您的应用与数据提供从基础设施到网络的双重安全保障，让您安心专注于业务创新。

生态集成，开箱即用

深度集成天翼云负载均衡（ELB）、文件存储（SFS）、对象存储（ZOS）等 IaaS 服务，以及微服务引擎（MSE）、应用监控（ARMS）等 PaaS 生态，提供一站式、高内聚的完整解决方案，大幅提升开发和运维效率。

功能特性

CAE 应用托管

CAE 应用托管相关能力

功能集	功能	功能描述	参考文档
应用生命周期管理	创建/部署/更新/查看/启动/停止/删除应用	创建/部署/更新/查看/启动/停止/删除应用。	
微服务治理	无损上线	在应用启动过程中，无损上线为应用提供服务延迟注册、服务就绪检查和服务小流量预热的保护能力	
	无损下线	在应用执行部署、停止、回滚、缩容和重置时，通过无损下线来保证应用正常关闭	

功能集	功能	功能描述	参考文档
	金丝雀灰度	对于部署在 CAE 的 Spring Cloud 或 Dubbo 微服务应用，为了确保升级操作的安全性，您可以通过启用灰度发布（即金丝雀发布）的灰度规则进行小规模验证，验证通过后再全量升级	
	限流降级	CAE 支持使用微服务引擎 MSE 实现应用的限流降级，全面保障应用的可用性	
微服务注册中心	使用 CAE 内置注册中心	CAE 为用户提供免费的内置 Nacos 注册中心，在无需购买或自建注册中心的情况下即可部署微服务应用到 CAE 平台	
	使用自建 Nacos 注册中心	CAE 支持使用自建 Nacos 注册中心实现服务的注册与发现功能	
运维管理	基础监控	CAE 对应用所运行设备的 CPU、负载、内存、网络和磁盘进行数据采集与分析，并以动态图的方式展示，方便实时、直观地了解应用所运行设备地状态	
	应用监控	CAE 为多种语言和框架提供无侵入的应用监控能力	
	一键启停	CAE 为应用提供了一键启停按钮，方便用户进行运维操作，同时还支持批量启停操作	
	日志管理	日志管理提供应用实时日志查询功能。	
弹性伸缩	手动扩缩	在应用的实例负载过高时以手动方式添加新应用实例，在应用闲置时减少应用实例，能够高效利用应用资源、降低成本	
	自动扩缩	在分布式应用管理中，弹性伸缩能够感知应用内各个实例的状态，并根据实例状态自动增加或减少实例数量，即扩容或缩容	
应用访问	基于 ELB 实现应用公网及私网访问	在 CAE 中部署应用后，可以通过添加公网 ELB 实现公网访问应用，也可以添加私网 ELB 实现同 VPC 内私网访问应用	
配置管理	配置项 (ConfigMap)	配置项是一种存储应用所需配置信息地资源类型，它可以作为容器运行环境中的环境变量，便于应用部署后灵活变更容器配置，也可以通过挂载配置文件的方式向容器中注入配置信息	
	保密字典 (Secret)	保密字典是一种用于存储和管理密钥、证书等敏感信息的资源类型。为避免敏感数据暴露到镜像或应用与任务部署参数中，推荐您使用 CAE 命名空间级别的保密字典	
高级设置	设置启动命令	CAE 会根据预设的启动参数来启动容器	
	设置环境变量	应用在系统中运行更需要配置特定的环境变量	
	设置 Hosts 绑定	CAE 支持应用级别的实例，通过绑定 Hosts 对主机名进行解析，方便应用实例通过主机名进行访问	
	设置持久化日志	CAE 集成了云日志服务的日志收集功能，支持将业务文件日志（容器内日志文件）、容器标准输出日志（stdio）无限制行数地收集至 ALS，便于您聚合分析。	
	设置 NAS 存储	将 NAS 挂载至 CAE 应用实例，可以有效解决应用数据地持久化存储需求，并实现应用实例之间地数据共享	
	设置 OSS 存储	OSS 适用于读多写少地场景，例如挂载配置文件或者前端静态文件等	

功能集	功能	功能描述	参考文档
	设置应用生命周期管理	如果您精通 K8s，且需要在应用容器启动前或者关闭前执行相关操作，例如运行前部署资源或者停止前优雅下线应用，可以设置应用生命周期管理	
	设置配置项	配置项能够将环境配置信息和容器镜像解耦，方便您修改应用配置	
版本管理	版本回退	应用修改配置并部署后，会自动生成一个基于时间点的应用版本，您可以查看对应时间点应用版本的配置，也可以操作版本回退到指定时间点的应用版本	

开源自建对比

云应用引擎 CAE（Cloud App Engine）是面向应用的 Serverless PaaS 平台，提供成本更优、效率更高的一站式应用托管方案，容器新手也可以体验 Serverless、K8s 和微服务。您可以通过不同场景，对比 CAE 与开源自建产品在系统搭建、成本、功能管理、弹性效率和监控诊断等方面的优势。

基础场景

对比项	开源自建 SC/Dubbo + IaaS 集群	开源自建 SC/Dubbo + 自建 K8s 集群	CAE
IaaS 购买和系统搭建	需要。	需要。	不需要。CAE 内置 K8s 集群底座和微服务框架。
运维成本	需要。	需要。	不需要。CAE 免运维。
硬件成本	按峰值固定保有包年 IaaS，资源利用率低，有闲置浪费。	按峰值固定保有包年 IaaS，资源利用率低，有闲置浪费。	按需使用，按量付费，无闲置浪费。
应用管理	- 纯黑屏操作，效率低且易出错。 - 自行集成各个云服务，自研发布策略用于规避停机升级风险。	- Kubectrl 或 K8s Dashboard 操作，门槛高。 - 自行集成各个云服务，自研发布策略规避停机升级风险。	- 白屏化操作，根据应用健康自动调配流量。 - 可灰度、可观测和可回滚。 - 提供镜像仓库和代码源等。
弹性效率	分钟级。	分钟级。	秒级。
使用体验	- 创建应用前置步骤多（需购买机器、创建集群和初始化集群等）。 - 自行对接各个云服务，集成体验差。	- 创建应用前置步骤多（需购买机器、创建集群和初始化集群等）。 - 自行对接各个云服务，集成体验差。	- 直接创建应用，流程极简。 - 无缝对接各个云服务，一站式 PaaS 体验。

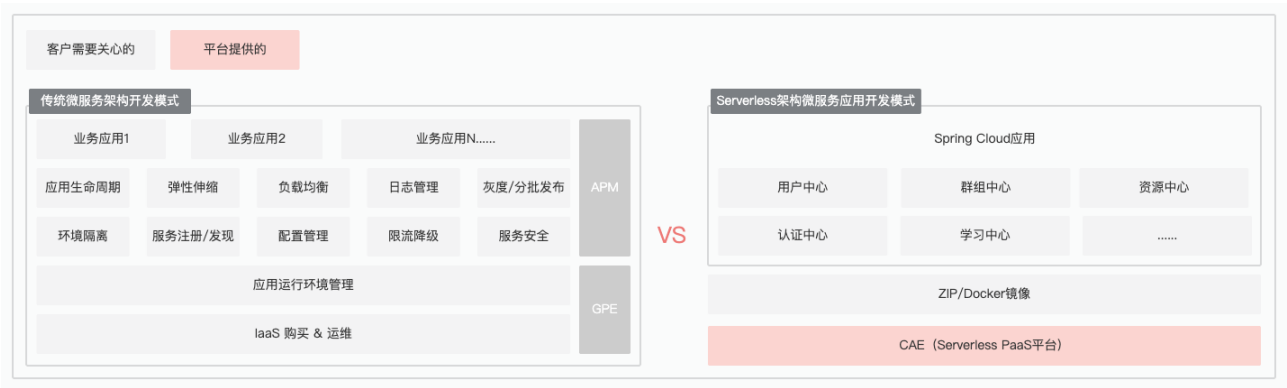
应用场景

云应用引擎 CAE（Cloud App Engine）具有广泛的应用场景，帮助您企业极速上云、从容应对突发性流量洪流和灵活启停应用环境，降低资源成本。

应用托管

在企业生产环境中，通过合理拆分微服务，将每个微服务应用压缩为 ZIP 包、Docker 镜像存储在天翼云镜像仓库。您只需基于 Spring Cloud 或 Dubbo 等框架开发规范迭代每个微服务应用，由 CAE 提供底层资源调度、部署、灰度发布、微服务治理和监控诊断等能力。同时提供丰富的高级应用配置项，实现业务快速迁移上云。

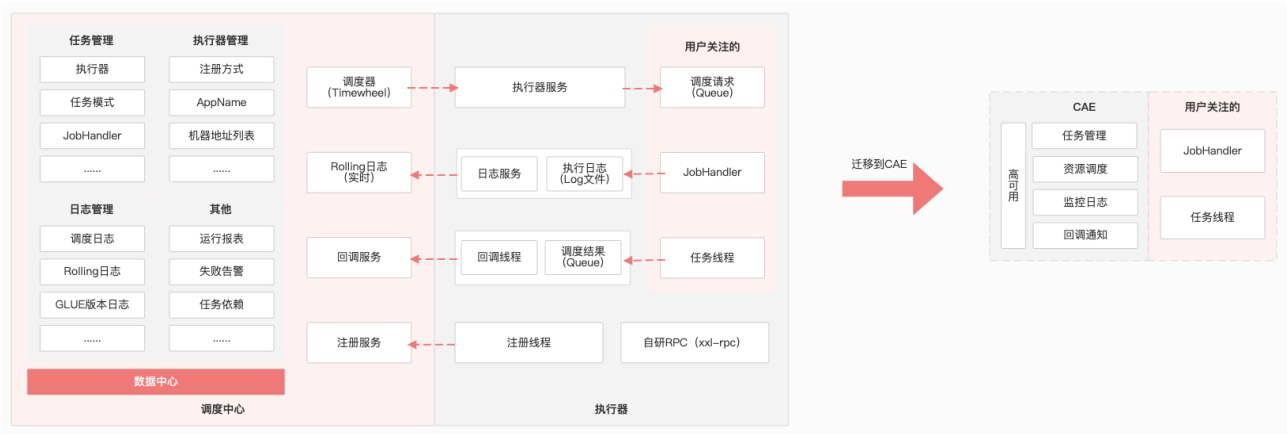
- 零改造：CAE 能够平滑迁移应用，零改造地完成 Spring Cloud 或 Dubbo 应用快速上云。
- 免运维：CAE 能够免运维底层基础设施，例如 IaaS、K8s、微服务组件和 APM 组件等，无需自建注册中心，极大降低开发运维成本。
- 低门槛：CAE 能够一站式开箱使用全套微服务能力，提供自动构建镜像、灰度发布、流量控制、环境隔离、应用监控等企业级高级特性。



任务托管

聚焦于泛互联网、新零售、电商、文化传媒、制造、IoT、物流、金融证券、医疗卫健和保险等行业。主要场景如下：

- 定时任务：定时拉取数据、爬虫。
- 批处理数据：数据清洗、转换、分析，对实时性要求低。
- 异步执行解耦：异步状态刷新以及离线查询。
- 微服务架构：与原有的微服务架构进行调用通信、流程解耦。



相比开源的分布式框架，其优点在于全托管免运维的用户体验，开箱即用的完备功能以及白屏化管控，任务运行完立即释放资源，不会浪费闲置资源成本。

使用限制

本文介绍使用云应用引擎时，应用的实例配额限制、CPU和内存限制，以及网络、容器、镜像等限制条件。

配额限制

限制项	限制值	说明
可创建应用的用户	实名认证，且账户内金额大于0元。	无
单应用实例数	20	单个应用的实例数量上限
总应用实例数	100	单Region下用户能够创建的全部应用实例数量
内存用量	10G	单Region下用户所有应用使用的内存大小上限
vCPU核数	2核	单Region下用户所有应用使用的vCPU核数上限
网络	专有网络VPC	不支持经典网络。专有网络VPC内部署的实例默认不可访问公网，如需访问公网，则需进行相关配置。
容器	Linux容器	不支持Windows容器。
系统盘磁盘空间	默认20GiB	不支持更改系统盘磁盘空间大小。为了保证您的业务连续性，在CAE控制台使用镜像部署应用时，请确保镜像大小不超过20GiB。

基本概念

在使用云应用引擎 CAE（Cloud App Engine）前，需了解该产品所涉及的概念。本文介绍与 CAE 产品相关的专有名词与基本概念。

基本概念总览

分类	资源	概念
命名空间	微服务应用	分布式配置管理、应用实例、应用健康检查、启动命令
	任务	任务实例、任务模板、任务记录、启动命令
基础信息	容器、容器镜像	

命名空间

逻辑隔离的运行环境。

包含 K8s 的命名空间和微服务的命名空间两层含义。从应用的服务调用的视角隔离不同的运行环境。例如开发环境、测试环境、生产环境。

应用

应用是 CAE 管理的基本单位。

一个应用内通常包含一个或多个实例。CAE 提供了完整的应用生命周期管理机制，可以完成应用从发布到运行过程的全面管理，包括应用创建、部署、启动、回滚、扩缩容等操作。

应用实例

应用运行的最小单位。

应用通常在一个或多个实例上运行，应用可以拥有多个部署版本，不同实例上运行着不同版本的应用程序。

微服务应用

提供服务注册发现和服务治理的能力，例如使用 Spring Cloud 以及 K8s Service 框架来开发和通信的应用。

任务

短时任务（Job）。

具备单机、广播、并行计算、分片运行、定时、失败自动重试、事件触发等特性。

任务实例

基于任务模板创建的执行业务逻辑的具体实例，CAE 单个任务中的单个实例是调度的最小单位。

任务模板

可以自动创建具体任务的模板。

可以对任务模板进行创建、修改、复制、停止和删除等操作。任务模板中可以对任务执行时间、实例规格、并发数等配置项进行配置。任务模板和任务的关系，类似面向对象程序设计中的类和对象的关系，任务模板是静态的概念，任务是动态的概念，可以根据单个任务模板批量创建多个任务。

任务记录

任务执行的记录。

可以查看和删除创建过的任务的执行记录，并支持通过状态筛选，状态分为：未执行、执行成功、执行失败、执行中。任务记录列表可以查看任务ID、执行状态、创建时间、完成时间、正在运行的实例数、成功运行的实例数和运行失败的实例数。

启动命令

设置容器启动和运行时所需要的命令。

启动容器就是启动主进程。在制作镜像时请在 Dockerfile 文件中设置 ENTRYPOINT 或 CMD。例如在 Dockerfile 中设置了 ENTRYPOINT ["top", "-b"]，则在容器启动时执行该命令。

容器

镜像运行时的实体。

可以对容器进行创建、启动、停止、删除和暂停等操作。镜像和容器的关系，类似面向对象程序设计中的类和实例的关系，镜像是静态的概念，而容器是动态的概念。实例与容器是一一对应的关系。

容器镜像

容器镜像是一种容器化标准交付件，用于打包应用程序及其依赖的环境。可以基于 Dockerfile 文件将应用构建为容器镜像并上传到容器镜像仓库中，然后您可以在测试或者生产环境中拉取容器镜像并启动容器。

开服地域

地域是指物理的数据中心，资源创建成功后不能更换地域。本文介绍云应用引擎 CAE（Cloud App Engine）支持的地域。

公共云

地域名称	RegionId
华东1	bb9fdb42056f11eda1610242ac110002

实例规格

云应用引擎 CAE（Cloud App Engine）的实例是应用或任务运行的最小单位。CAE 实例规格定义的属性包括 CPU 和内存，每种规格具有相应的网络基础带宽和网络收发包 PPS。请根据业务场景和业务特点选择应用的实例规格。

CAE提供了默认服务器资源类型，其支持的实例规格如下所示。

CPU	内存
0.5 核	1 GB
0.5 核	2 GB
1 核	1 GB
1 核	2 GB
1 核	4 GB
2 核	2 GB
2 核	4 GB
2 核	6 GB
2 核	8 GB
2 核	16 GB
4 核	4 GB
4 核	6 GB
4 核	8 GB
4 核	16 GB
4 核	32 GB
6 核	6 GB
6 核	12 GB
8 核	8 GB
8 核	16 GB
8 核	32 GB
8 核	64 GB
12 核	12 GB
12 核	24 GB

CPU	内存
12 核	48 GB
16 核	16 GB
16 核	32 GB
16 核	64 GB

说明

CAE 提供的 CPU 核数均指虚拟核（vCPU），非物理 CPU 内核。

快速入门

新手指引

云应用引擎 CAE（Cloud App Engine）是一款面向应用的 Serverless 全托管平台，能够帮助 PaaS 层用户免运维 IaaS、按需使用、按量计费，做到低门槛微服务、Java、PHP 等技术栈语言的应用上云。本文介绍如何使用 CAE，帮助您快速了解 CAE 以及各方面的实践教程。

背景信息

如果您初次接触 CAE，建议您阅读 [什么是云应用引擎](#)。

CAE 使用流程

- 在首次部署 CAE 应用前，需要完成规划 VPC、vSwitch、命名空间（区分测试、预发、生产等环境）的准备工作。具体操作，请参见 [准备工作](#)。
- 在 CAE 控制台部署应用。

具体操作，请参见 [应用部署](#)。除通过控制台部署应用外，CAE 还支持通过 OpenAPI 方式来部署应用。

说明

如果您是第一次部署应用到 CAE，需要在 CAE 控制台 创建应用。

- 通过以下方式访问 CAE 应用。
 - 方式一：绑定 ELB 访问，一个端口支持绑定一个应用。具体操作，请参见 [应用绑定 ELB 访问](#)。
 - 方式二：配置网关路由访问。具体操作，请参见 [网关路由](#)。
- 为 CAE 应用配置更多进阶功能
 - 例如，弹性（降本增效）、Java 微服务增强、高可用、存储等功能。

准备工作

您在使用云应用引擎托管应用之前，需要完成一些准备工作，包括开通服务、创建 VPC 和创建命名空间。VPC 为您构建隔离的网络环境，命名空间为您构建隔离的资源环境。

步骤一：开通服务并授权（天翼云账号）

使用 CAE 前，您需要开通 CAE 服务并对天翼云账号授权。CAE 支持免费开通服务，开通后可以按照实际使用量进行计费。CAE 计费方式为按量付费。详细说明，请参见 [计费概述](#)。

- 1. 登录 [CAE产品主页](#)。
- 2. 单击免费开通。
- 3. 在登录页面，输入阿里云账号和密码，单击登录。
- 4. 在云应用引擎页面，选中我已阅读并同意云应用引擎服务协议，并单击立即开通。
- 5. 在开通完成页面，单击管理控制台，在控制台首页弹出的欢迎使用云应用引擎 CAE 对话框，单击确认创建。

步骤二：创建VPC

VPC为您的应用提供了完全隔离的网络环境，处于同一VPC的应用可以互相访问，从网络层面提高应用的安全性。具体操作，请参见 [创建和管理专有网络](#)。

说明

如果您尚未开通专有网络 VPC 服务，请依据页面提示开通 VPC 服务。

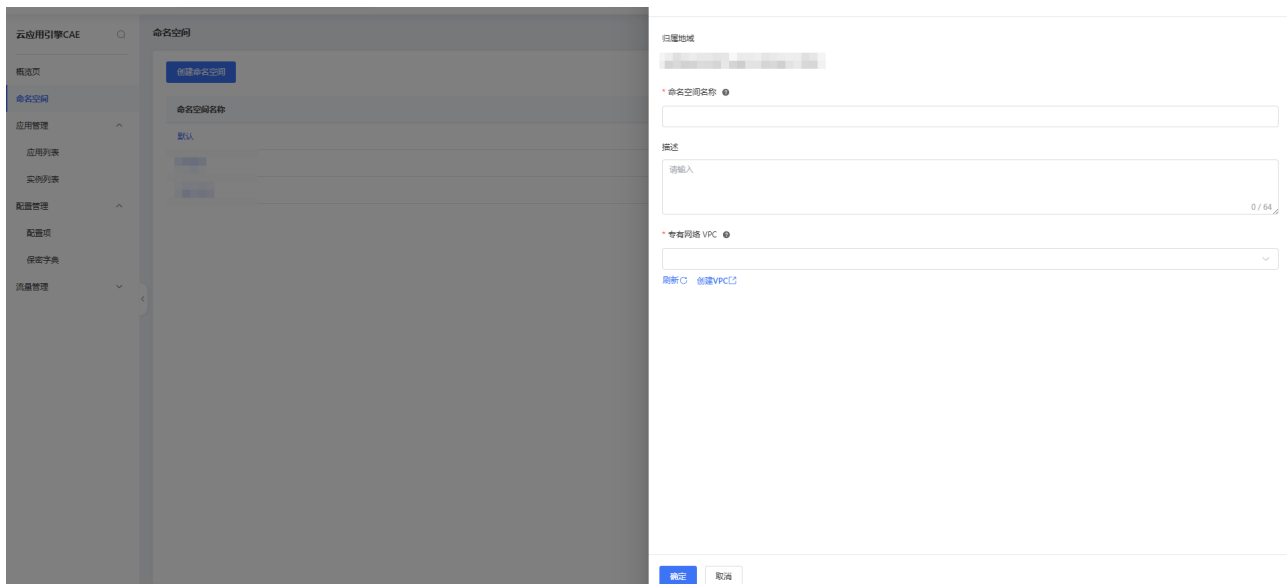
请提前规划 VPC 和命名空间需要创建的地域。创建 VPC 时选择的地域必须与步骤三：创建命名空间时选择的地域保持一致。

步骤三：创建命名空间

命名空间以应用的服务调用与分布式配置推送为视角，提供逻辑隔离的运行环境。如果您有开发环境、测试环境和生产环境等场景，建议使用命名空间，将应用逻辑隔离，便于您管理应用以及一键启停应用，提高应用的安全性。

- 1. 登录 [CAE控制台](#)，在顶部菜单栏选择目标地域，然后在左侧导航栏单击命名空间。
- 2. 在命名空间页面，单击创建命名空间。
- 3. 在创建命名空间面板，配置以下信息，单击确定。

配置项	说明	示例
命名空间名称	自定义命名空间名称。	测试
命名空间ID	命名空间ID不支持修改。	demo
描述	命名空间的描述信息，自定义。	测试
专有网络VPC	选择已创建的VPC。一个部署环境只能关联一个VPC，但一个VPC能被多个命名空间关联。命名空间创建成功后，您可以根据需求切换关联的VPC。 说明 如果您需要切换目标命名空间关联的 VPC，请先删除当前 VPC 内的所有应用，之后再行进行专有网络 VPC 的切换。	vpc-wz9i4zexkzbqceo5n****



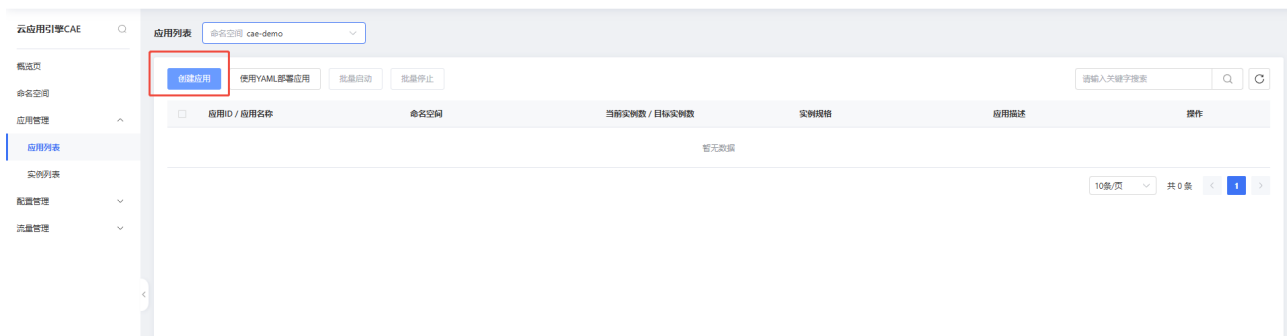
部署第一个 CAE 应用

本文介绍如何使用云应用引擎，帮助您快速了解云应用引擎以及各方面的实践教程。

您仅需将代码或镜像上传到 CAE，即可部署应用服务。CAE 会代您管理底层的计算资源，通过简单配置即可实现根据访问量自动扩缩容、跨可用区部署。

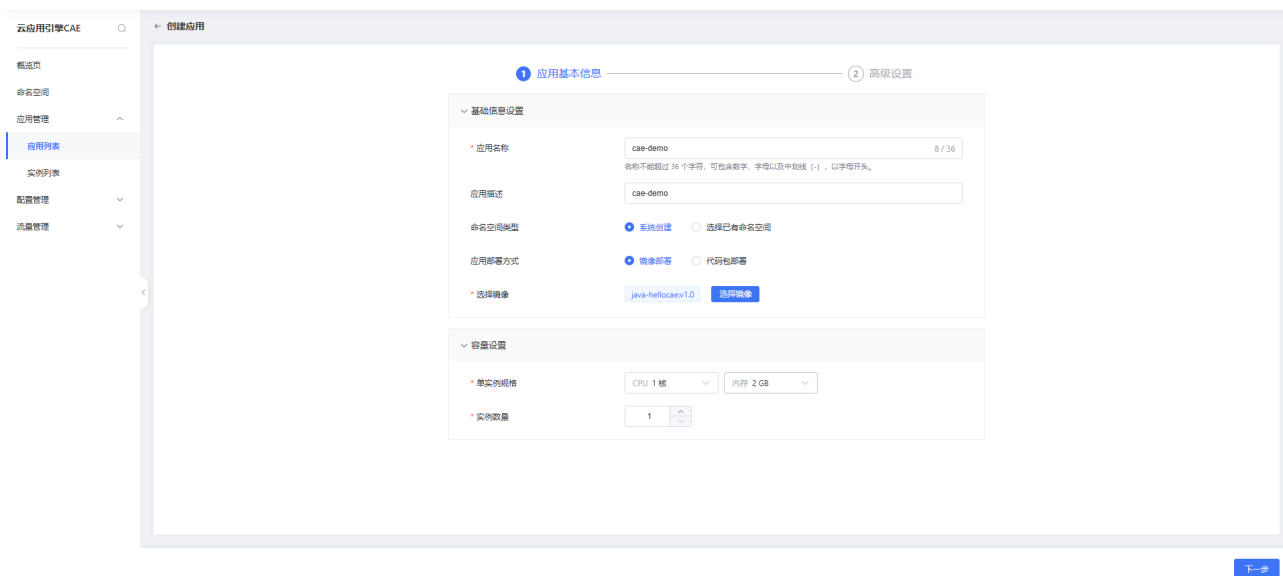
部署应用

1. 登录 [CAE 控制台](#)，在顶部选择应用部署的地域（本文以华东1为例），点击创建应用。

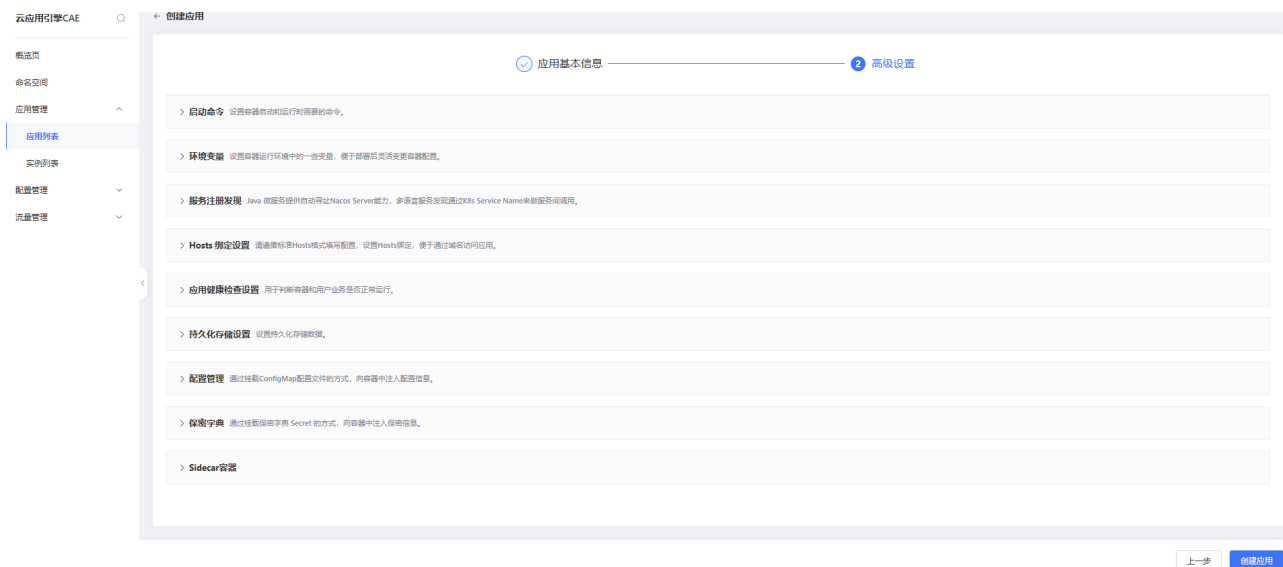


2. 在应用基本信息配置以下参数，其余保持默认。点击下一步。

- 自定义应用名称，例如cae-demo。
- 设置命名空间类型为####，表示应用将会创建于当前地域的默认命名空间，且自动创建并关联VPC、交换机、安全组等网络资源，无需用户关注。
- 设置应用部署方式为#####。点击选择镜像，在示例镜像标签页，找到java-hellocae，选择镜像版本为v1.0，点击确定。
- 在容量设置区域，自定义单实例规格和实例数，这决定了应用初始运行的实例数量、系统为每个实例分配多少计算资源。



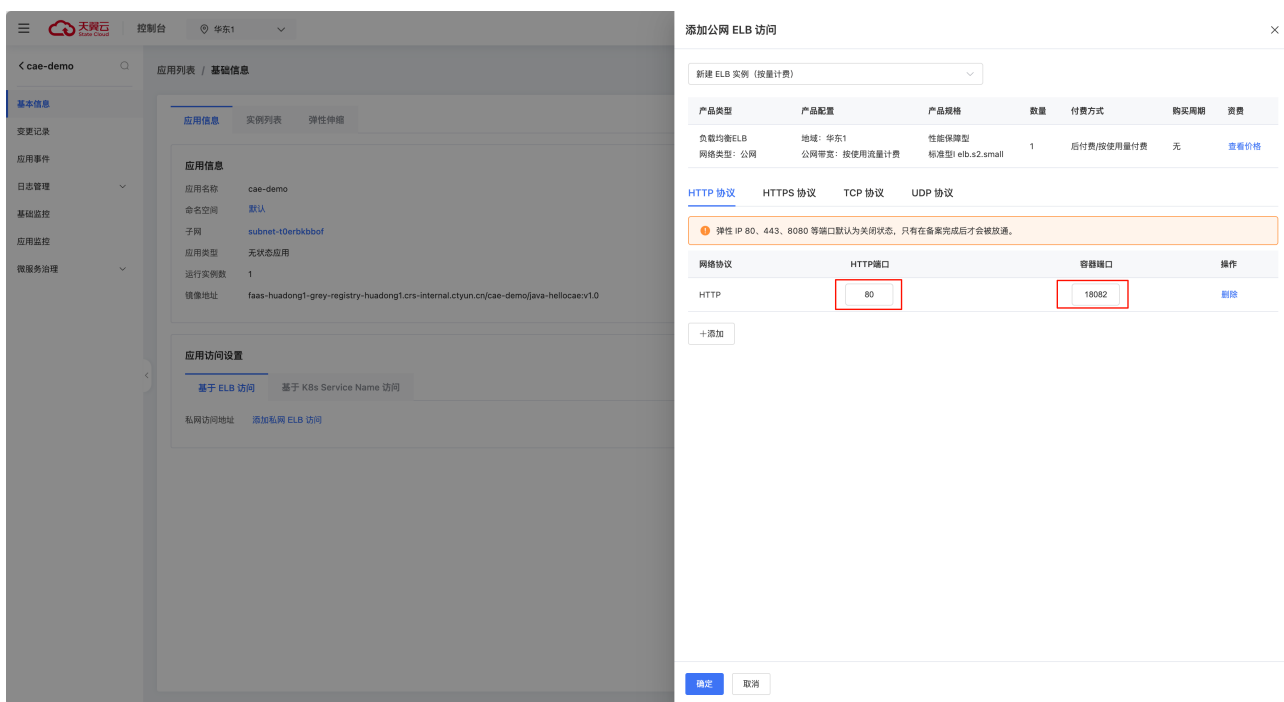
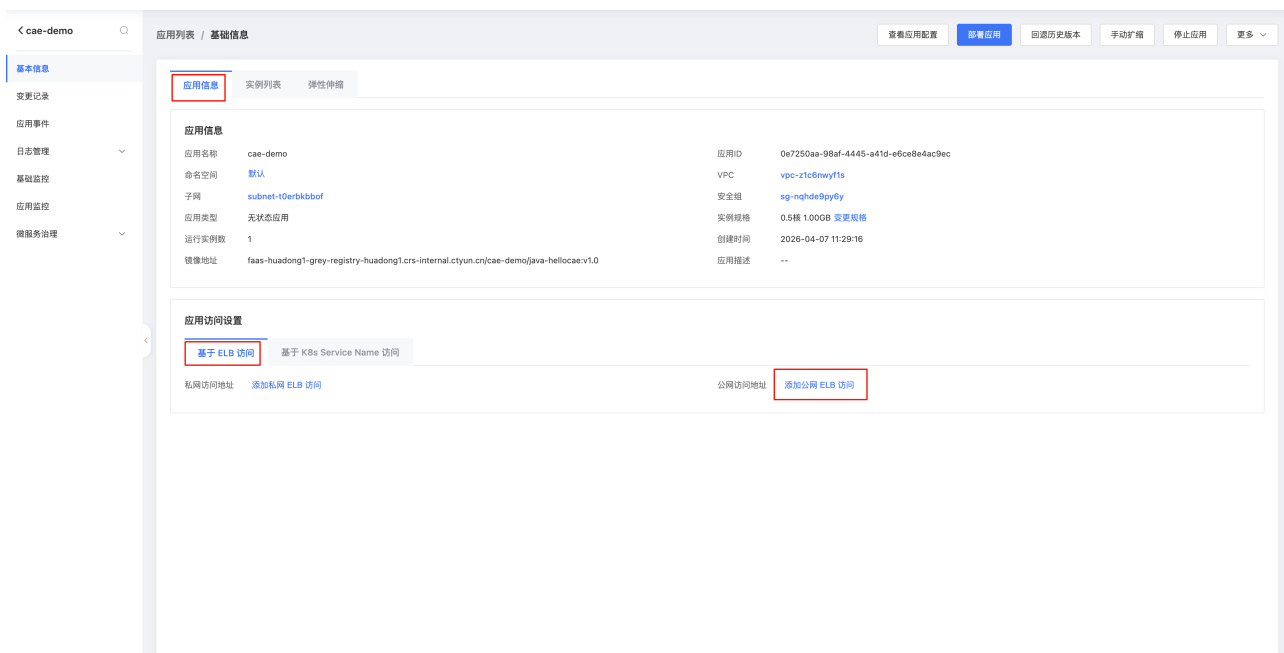
3. 在高级设置页面可以配置高阶的k8s功能，我们保持默认，点击创建应用，应用将开始创建。请耐心等待应用创建完成，大约需要1分钟。



通过公网访问应用

为了实现从公网访问该应用，一种简便的方式是为该应用绑定公网 ELB 实例。

1. 在应用列表中查看已创建的应用，点击该应用可以进入其详情页面。
2. 在应用信息页签中，在应用访问设置区域选择基于ELB访问。点击添加公网ELB访问，配置如下参数，然后点击确定。
 - a. 在ELB实例中，选择新建ELB实例。
 - b. 在HTTP协议页签中，设置HTTP端口为80，容器端口为18082。



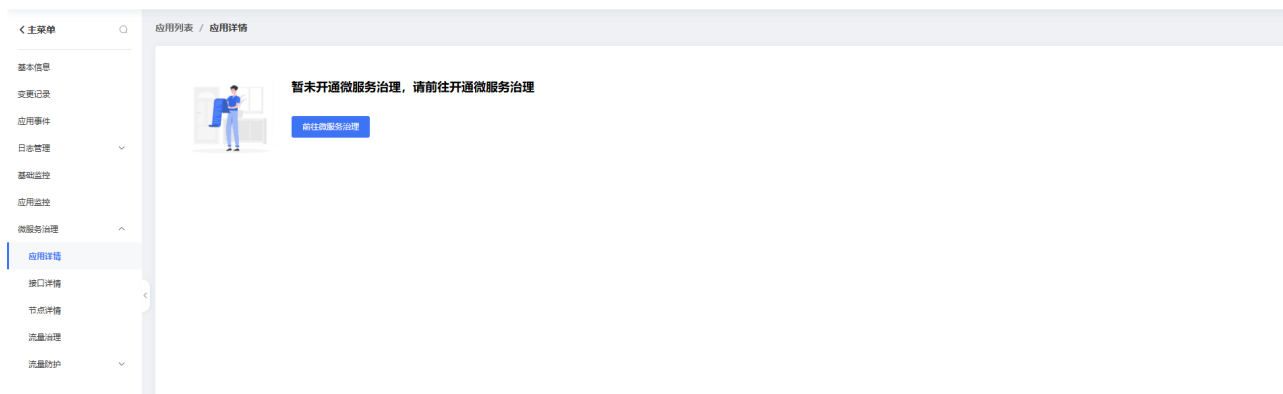
清理资源

在完成本教程后，建议清理相关资源，避免持续产生费用。

1. 删除应用：进入 cae-demo 应用的详情页面，点击更多 > 删除应用，并按照指引操作来删除上述应用。
2. （可选）删除网络资源：在部署应用的过程中，系统自动创建了VPC、交换机、安全组。
 - a. 查看VPC：在左侧导航栏选择命名空间，点击默认，在基础信息页面查看VPC，点击链接跳转到VPC详情页。
 - b. 查看交换机、安全组：在VPC的资源管理标签页点击链接跳转到相应资源的详情页。
 - c. 在各资源的详情页执行删除操作。

后续步骤

- 您已通过 Demo 镜像体验了部署应用的流程。CAE 为您提供以下方式部署实际的应用：
 - [使用镜像部署应用（推荐）](#)：您可以将任何应用制作成镜像，然后将其推送到镜像仓库，最后将其部署到 CAE。
 - [使用代码包部署应用](#)：CAE 提供 Java、Python、Go、Node.js、.NET Core 特定版本的运行环境，如果您的代码兼容，则可以使用代码包部署应用。您需要先制作 ZIP 格式压缩包，然后将其部署到 CAE。
- 本示例使用系统创建的默认命名空间来部署应用。您也可以通过自定义命名空间来实现不同应用之间、开发/测试/生产环境之间的隔离。每个命名空间需要绑定一个 VPC。部署应用时，需要为应用实例绑定该 VPC 中的交换机，选择不同可用区的交换机即可实现应用的跨可用区部署。
- 创建应用后，可以手动调整实例数量与实例规格，或通过配置弹性伸缩策略实现根据访问量、资源使用情况自动调整实例数量。
- 为应用启用更多进阶功能。例如，微服务治理、持久化日志、应用监控等功能。



用户指南

命名空间

管理命名空间

命名空间为应用提供相互独立的逻辑运行环境。通过命名空间，您可以更好地组织和隔离不同环境下的应用，确保服务调用更加清晰，配置管理更加高效。在典型的应用场景中，您可能需要区分 开发环境、测试环境和生产环境。

说明

在云应用引擎中，每个资源池都会自动创建一个默认命名空间，并为其分配一个专有网络（VPC）。您可以直接在该命名空间中部署和运行应用。需要注意的是，默认命名空间不支持更新或删除。如果默认命名空间无法满足您的业务需求，您可以根据实际情况新建并管理命名空间，实现更灵活的环境隔离与应用管理。

命名空间管理指引

查看命名空间

1. 登录云应用引擎控制台。
2. 在左侧导航栏中选择“命名空间”，进入命名空间管理页面，可以查看命名空间列表。
3. 单击目标命名空间，可以查看命名空间的基础信息。

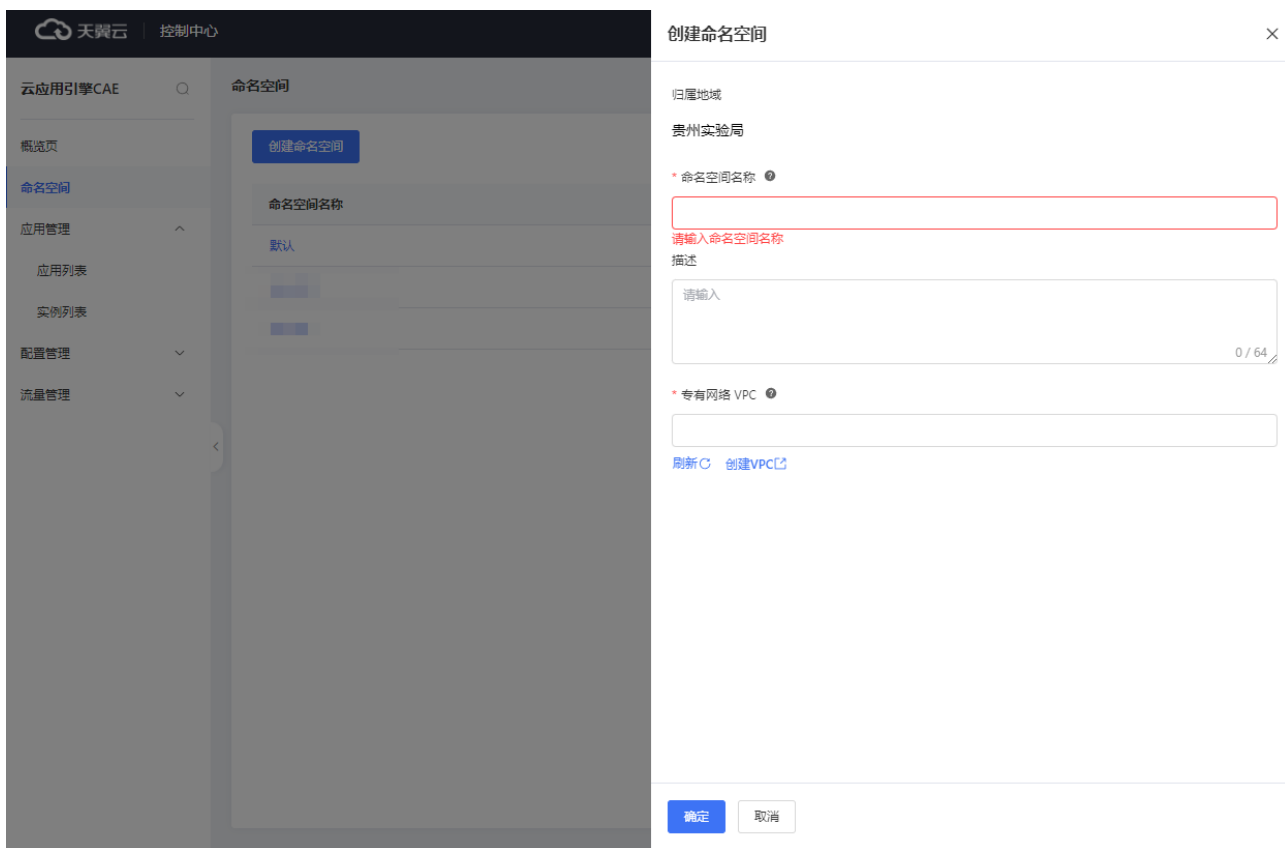




创建命名空间

1. 登录云应用引擎控制台。
2. 在左侧导航栏中选择“命名空间”，进入命名空间管理页面。
3. 单击“创建命名空间”。
4. 在弹出的 创建命名空间 面板中，根据下表说明配置相关信息，然后单击“确定”完成创建。

配置项	说明	示例
命名空间名称	为自定义命名空间设置名称，用于区分不同命名空间。在同一资源池内，命名空间名称 不得重复。	测试
描述	对命名空间进行说明或备注，便于管理	本命名空间用于测试环境
专有网络 VPC	选择已有 VPC 进行关联。一个命名空间只能绑定一个 VPC，但一个 VPC 可关联多个命名空间。创建后不可切换 VPC。	cae-auto-create-vpc



编辑命名空间

1. 登录云应用引擎控制台。
2. 在左侧导航栏中选择“命名空间”，进入命名空间管理页面。
3. 在目标命名空间所在行的操作列中，单击“编辑”。
4. 在弹出的编辑命名空间面板中修改命名空间信息，然后单击“确定”完成更新。

删除命名空间

1. 登录云应用引擎控制台。
2. 在左侧导航栏中选择“命名空间”，进入命名空间管理页面。
3. 在目标命名空间所在行的操作列中，单击“删除”。
4. 在弹出的提示框中单击“确定”完成删除。

对应用进行批量操作

在同一命名空间下，云应用引擎支持对多个应用进行批量操作，包括批量启动、批量停止、批量绑定标签、批量解绑标签等。本文主要介绍如何对同一命名空间下的应用进行批量操作，以及如何导出目标地域下所有应用的信息。

前提条件

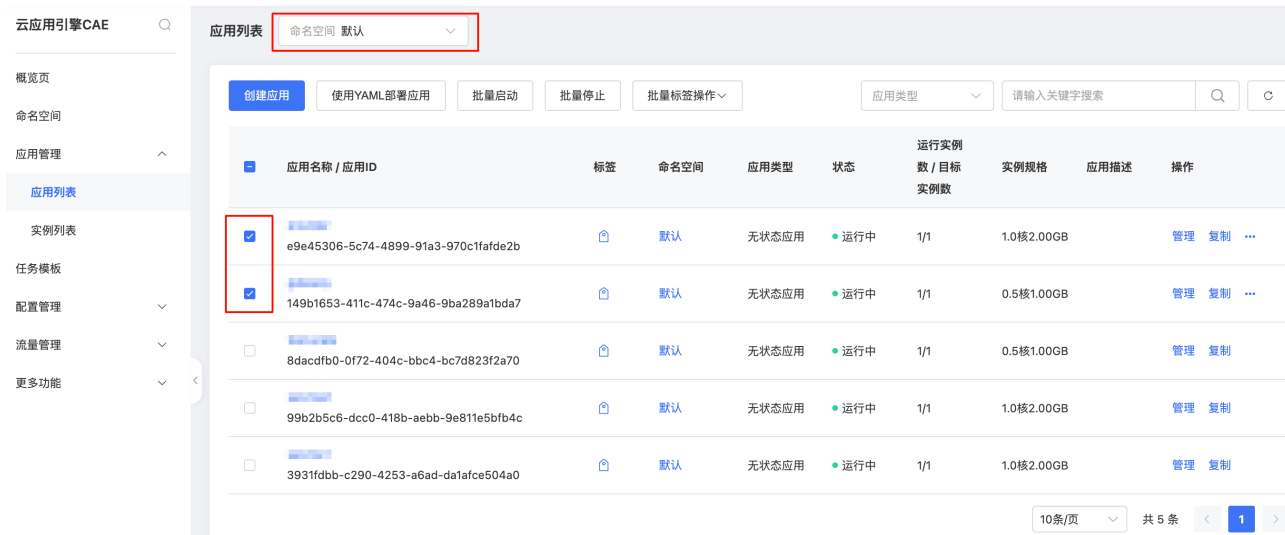
- 已创建命名空间。具体操作，请参见 [创建命名空间](#)。
- 在已创建成功的命名空间下创建了一个或多个应用。具体操作，请参见 [应用部署](#)。

限制条件

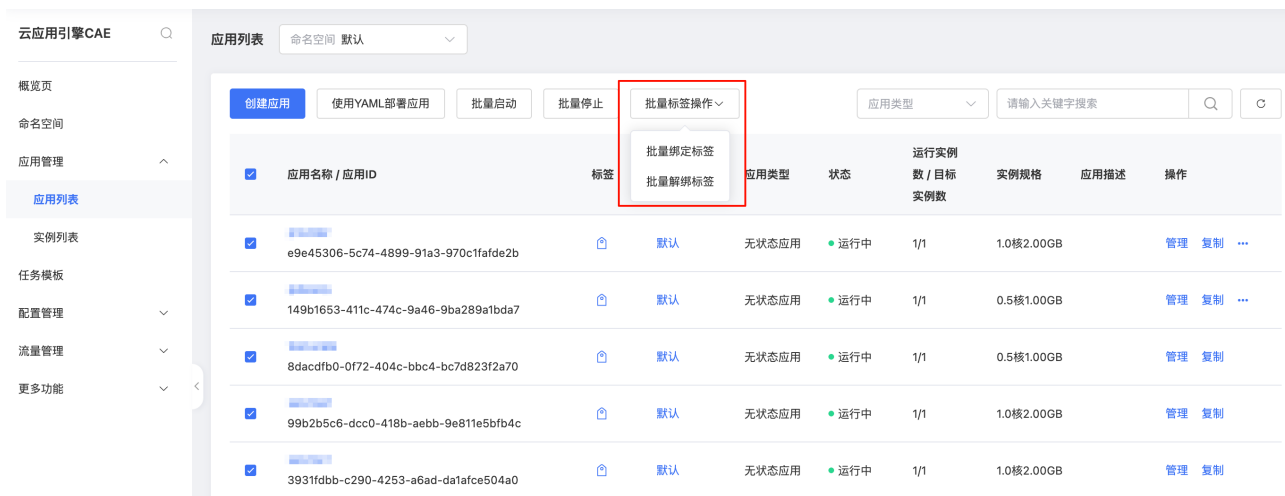
批量开启或关闭ARMS高级监控的功能仅适用于Java应用。

批量操作

1. 登录 [CAE控制台](#)，在左侧导航栏选择应用管理 > 应用列表，然后在顶部菜单栏选择目标地域。
2. 在应用列表页面，选择目标命名空间，然后勾选需要批量操作的应用。



3. 勾选好需要批量操作的应用后，单击需要批量操作的按钮（例如：批量停止应用）。



说明

如果您需要为应用批量绑定标签、批量解绑标签等，请单击批量标签操作，然后在下拉列表中单击您需要进行的批量操作。

应用管理

创建、查看与删除应用

创建应用

云应用引擎支持以下2种方式创建应用。

- 常规创建方式：用户能够通过表单方式进行应用配置，快速初始化应用实例
- YAML创建方式：通过内置的应用模板或者导入用户已有的YAML文件来快速创建应用

常规创建

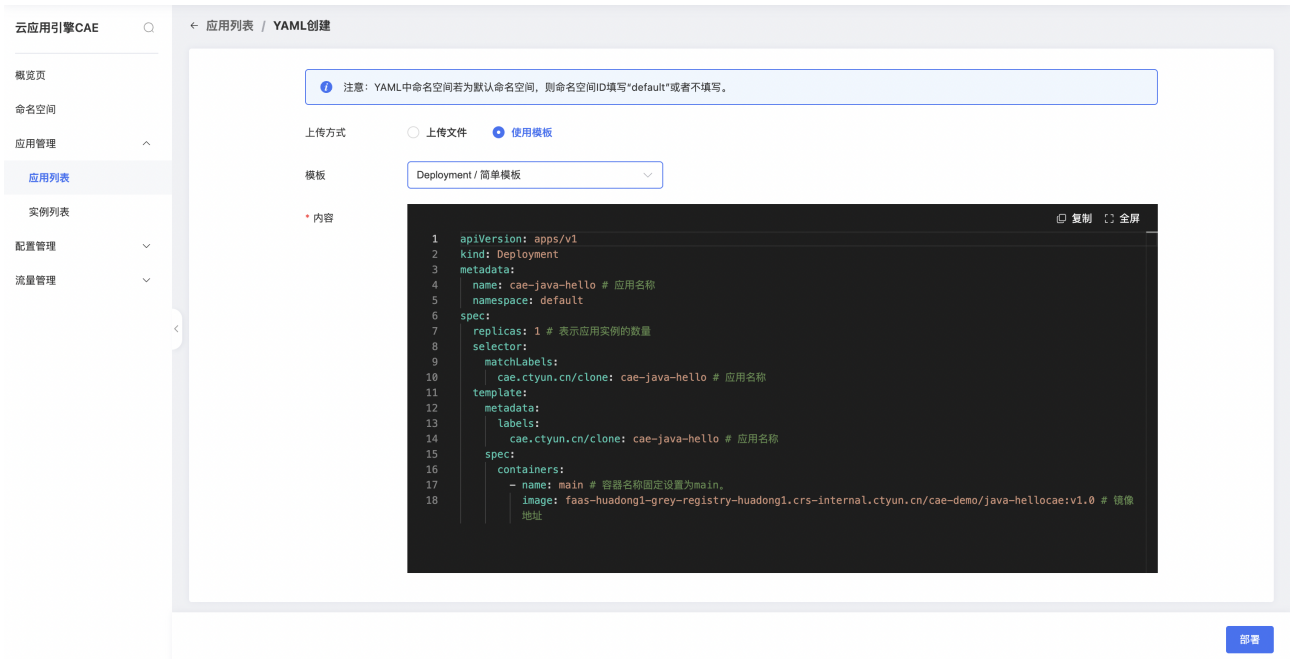
1. 在云应用引擎控制台，单击左侧导航栏选择 应用管理 > 应用列表，即可查看所有命名空间或指定命名空间下的应用列表。点击创建应用，打开应用创建基础信息页面。
2. 在应用创建基础信息引导页面，进行基础信息设置（如命名空间、部署方式、网络配置等）、容量设置（单实例规格、实例数量等）。点击下一步，进入高级设置页面。
3. 按需进行高级设置，也可以暂时跳过。单击创建应用按钮以创建应用实例。

YAML方式创建

在云应用引擎控制台，左侧导航栏中选择 应用管理 > 应用列表，单击使用YAML部署应用按钮。在页面中，可以基于提供的模板进行修改后部署，也可以上传您已有的K8s YAML文件进行部署。

说明

目前YAML方式创建仅支持namespace、configmap、secret以及deployment资源。

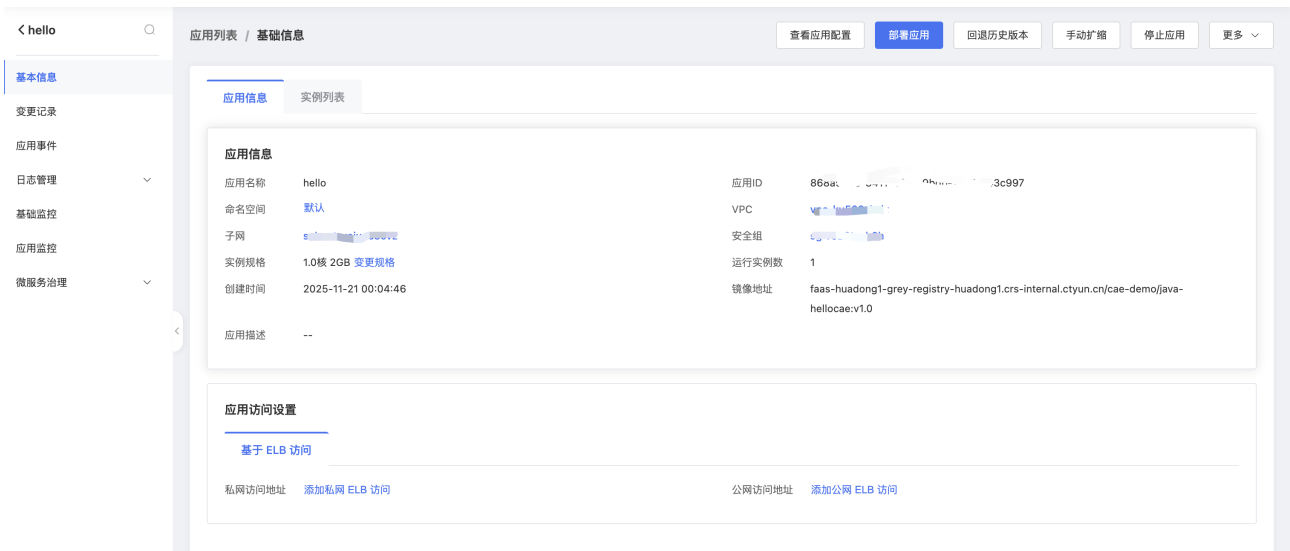


查看应用

在云应用引擎控制台，单击左侧导航栏选择 应用管理 > 应用列表，即可查看指所有/定命名空间下的应用列表。

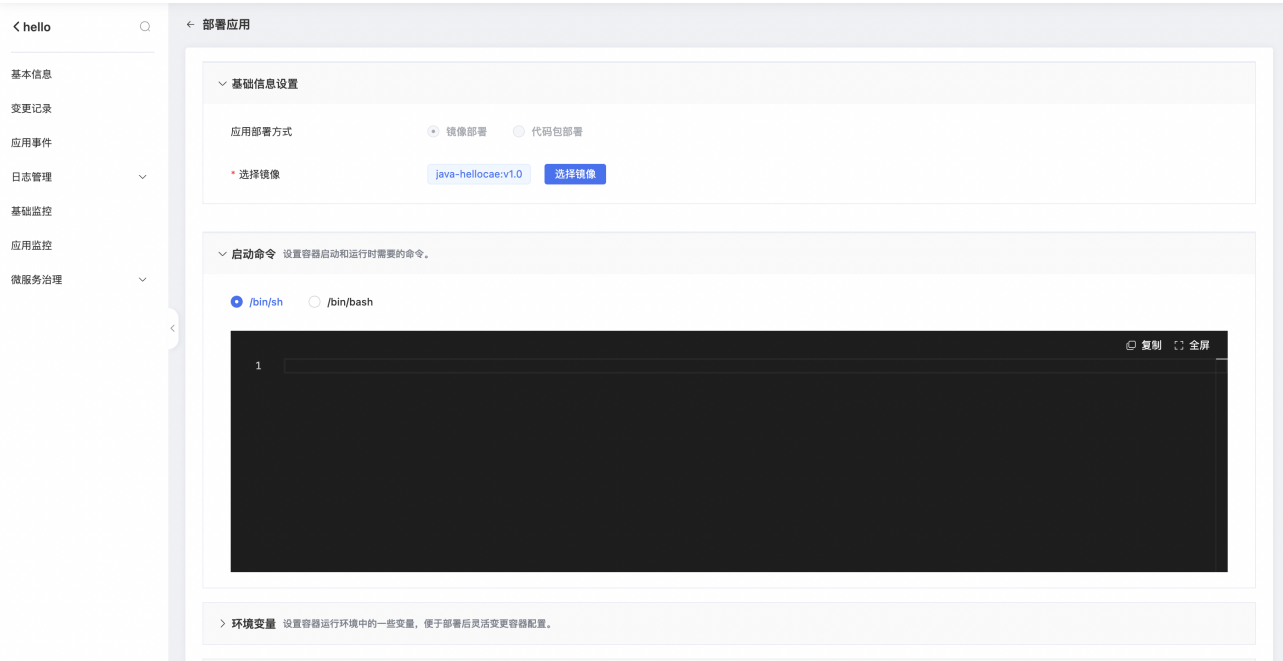


您可以在列表中单击目标应用名称，查看目标应用详情。



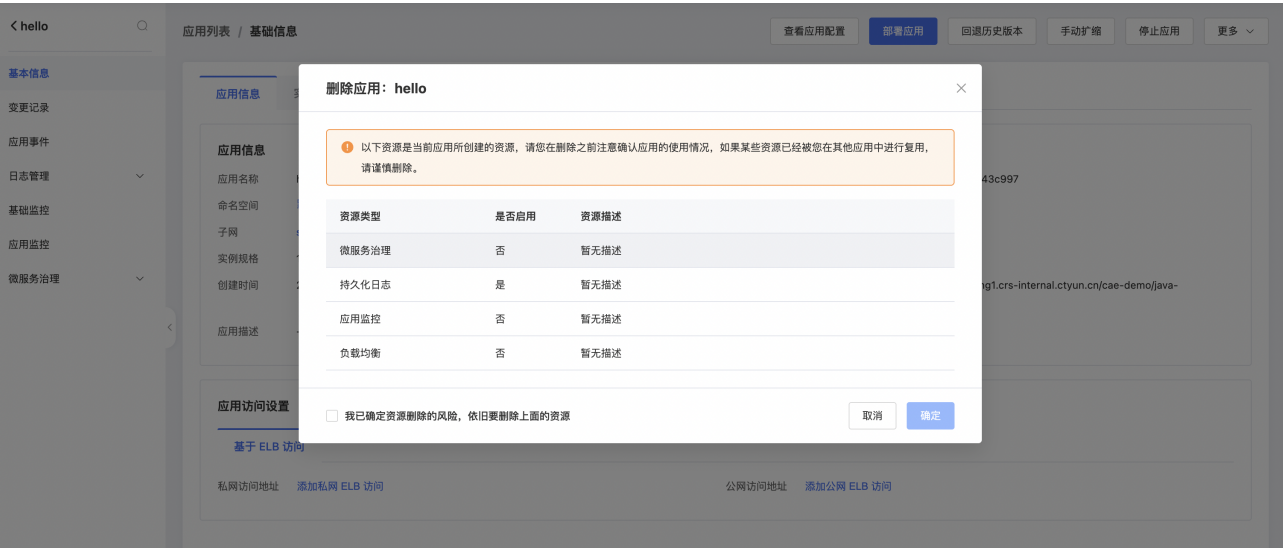
编辑应用

在云应用引擎控制台的应用详情页面，单击右上角部署应用按钮，您可以修改应用的配置信息。



删除应用

在云应用引擎控制台的应用详情页面，单击右上角 更多>删除 应用按钮，在弹出的确认窗口中，勾选下方“我已确定资源删除的风险，依旧要删除上面的资源”，单击确定，即可进行应用删除。



应用部署

使用镜像部署应用

您可以使用 CAE 提供的多语言Demo镜像，或者配置 Docker Hub、Harbor、GitHub Container Registry（GHCN）等常见的镜像地址，来自定义镜像，最后将其部署到 CAE。

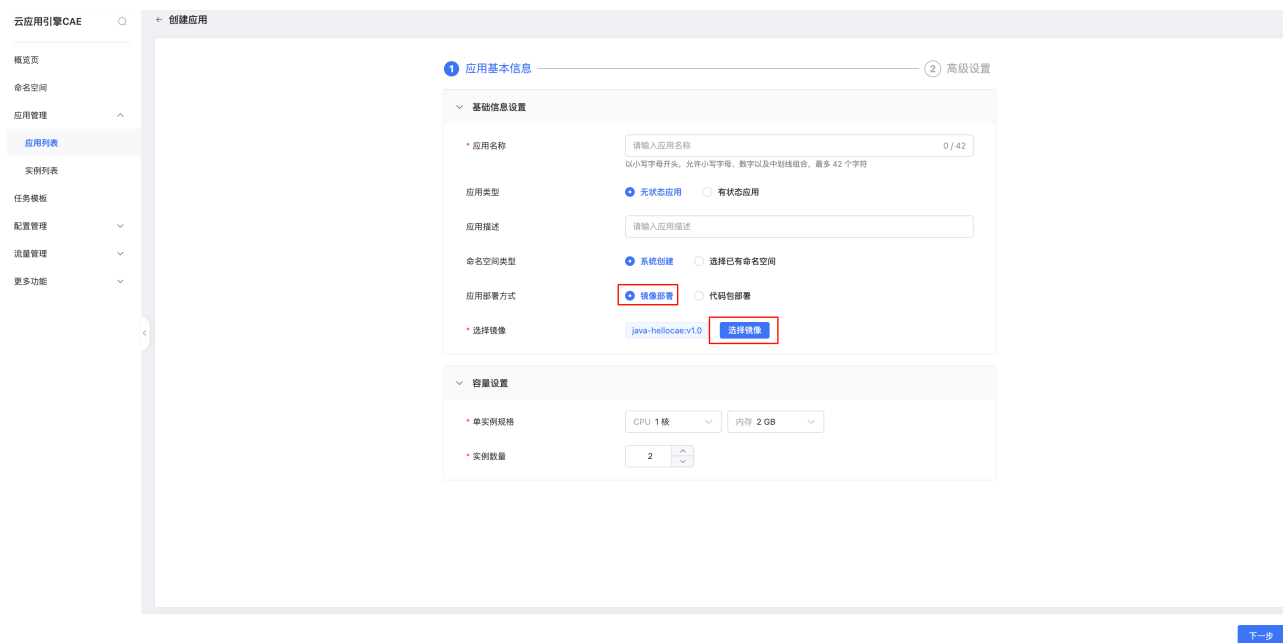
功能入口

创建应用

1. 在应用基础信息页面中，应用部署方式选择镜像部署，点击选择镜像。
2. 在弹出的设置镜像窗口，选择应用所属的技术栈语言，并根据需要选择示例镜像或者自定义镜像。
3. 当选择自定义镜像时，您需要填写镜像仓库地址以及配置用户名密码（可选）。

注意

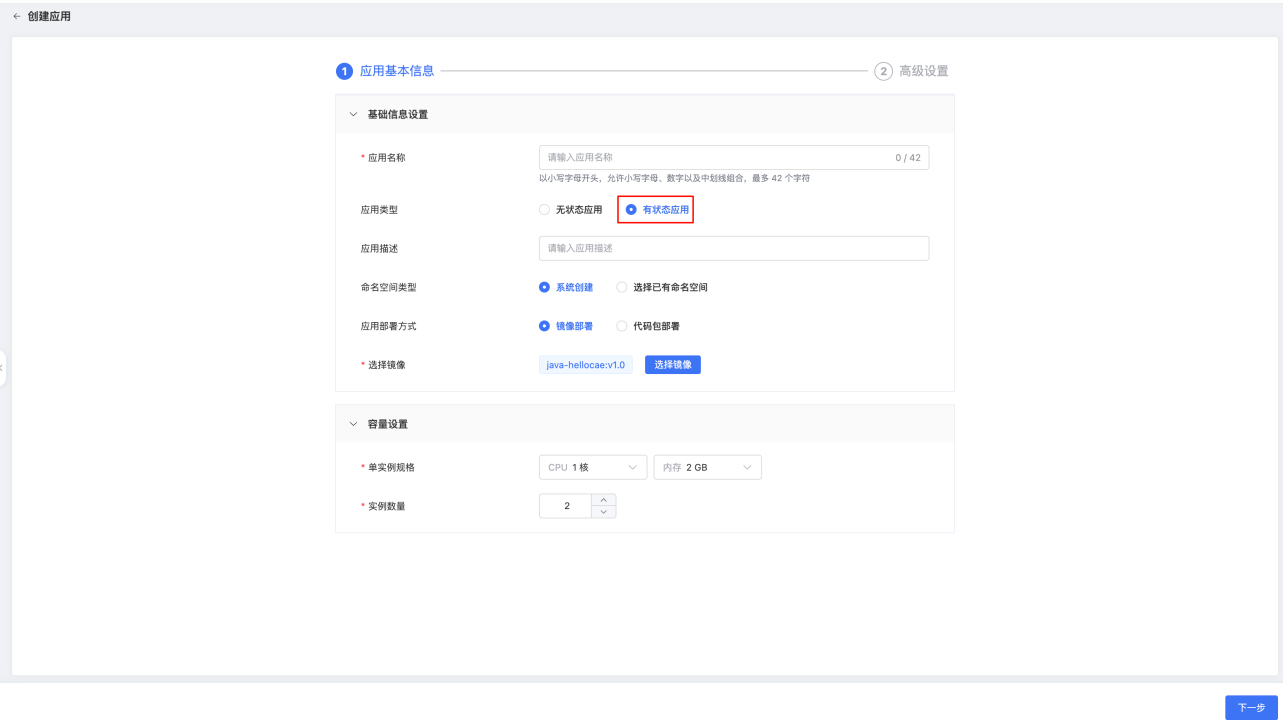
1. 当选择镜像网络为公网时，需要完成应用访问公网设置，确保应用可以访问公网
2. 当选择镜像网络为私网时，需要保证应用所在VPC网络与私有镜像仓库网络打通



创建有状态应用

创建有状态应用与创建无状态应用的操作基本一致，只需在应用基本信息页面中，应用类型选择有状态应用。

- 1. 在应用基础信息页面中，应用类型选择有状态应用。
- 2. 应用部署方式选择镜像部署，点击选择镜像。
- 3. 在弹出的设置镜像窗口，选择应用所属的技术栈语言，并根据需要选择示例镜像或者自定义镜像。
- 4. 当选择自定义镜像时，您需要填写镜像仓库地址以及配置用户名密码（可选）。



使用代码包部署应用

通过代码包部署应用时，云应用引擎支持的技术栈语言和运行环境如下表，部署前需要先制作符合云应用应用引擎要求的代码包。如果以下列表不包含您使用的技术栈语言和运行环境，则需要使用镜像部署应用。

配置项	说明	示例
运行环境	目前支持Java、Python特定版本的运行环境： • Java：支持 8、11、17、19 • Python：支持 3.6、3.9、3.10 • Go：支持1.x 如果您的代码兼容，则可以使用代码包部署应用。	Java8
代码包	目前只支持以zip包形式上传代码包，您需要将依赖一同打包到zip包。一次部署最多只能上传一个zip包文件。	Demo.zip
代码包版本	代码包版本会作为镜像版本，可使用时间戳自动生成，也可按需指定。	1756797405250

操作指引

1. 在 应用创建-基础信息设置 页面，应用部署方式选择 代码包部署 ，单击右侧的 选择代码包 。
2. 在弹出的 设置代码包 窗口，选择应用所属的技术栈语言以及应用所需的运行环境。
3. 上传含有应用执行程序的Zip包。
4. 设置代码包版本 ，并单击左下方的 确认 。
5. 在设置完应用基础信息后，在下一步的页面中，设置 启动命令 。

说明

启动命令执行目录与代码包解压所在目录相同。

例如，代码包中可执行文件路径为 target/cae-demo.jar ，则启动命令为 java -jar target/cae-demo.jar 。

可参考文本下方部署示例部分。

6. 单击 创建应用 。

填写效果

部署应用

基础信息设置

应用部署方式

镜像部署 代码包部署

* 选择代码包

hello-cae-java-offline.zip:0.0.7 选择代码包

启动命令 设置容器启动和运行时需要执行的命令。

镜像里必须存在 /bin/bash 或者 /bin/sh

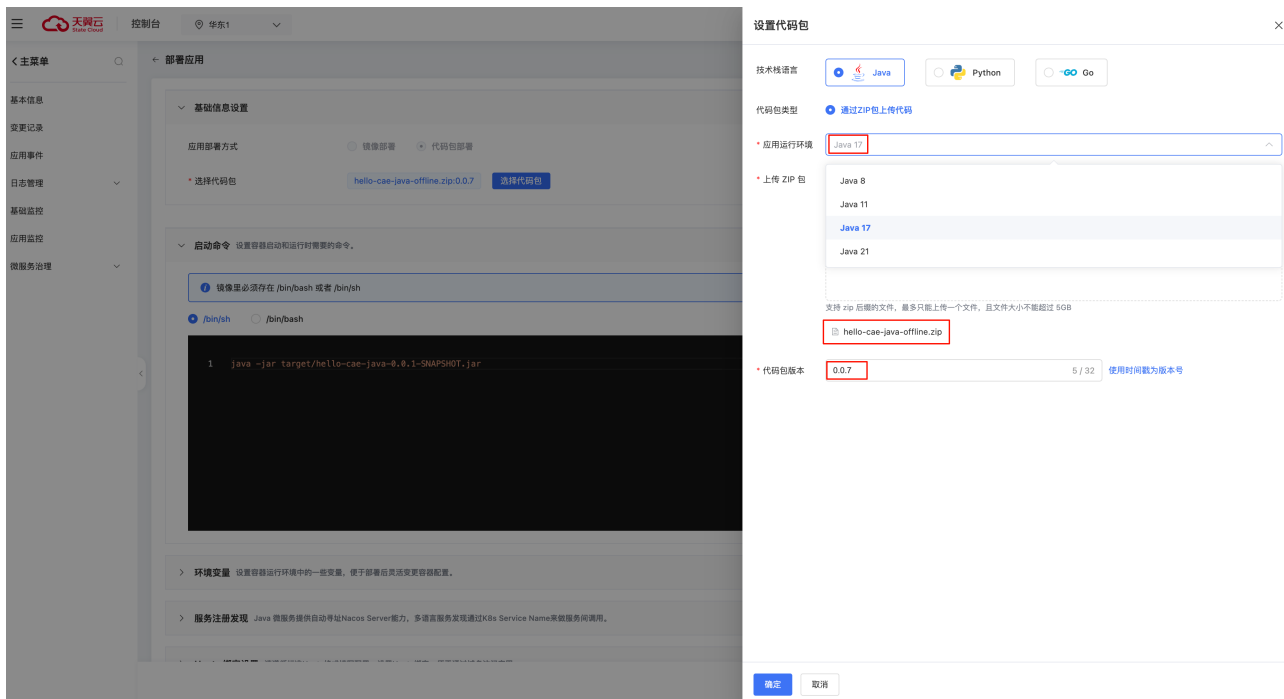
/bin/sh /bin/bash

```
1 java -jar target/hello-cae-java-0.0.1-SNAPSHOT.jar
```

环境变量 设置容器运行环境中的一些变量，便于部署后灵活调整容器配置。

服务注册发现 Java 微服务提供自动寻址Nacos Server能力，多语言服务发现通过K8s Service Name来做服务间调用。

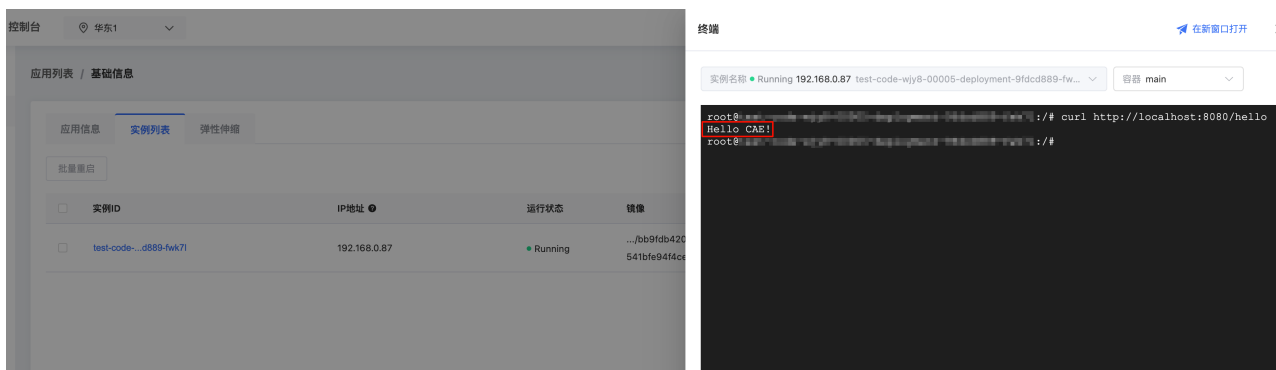
保存



部署示例

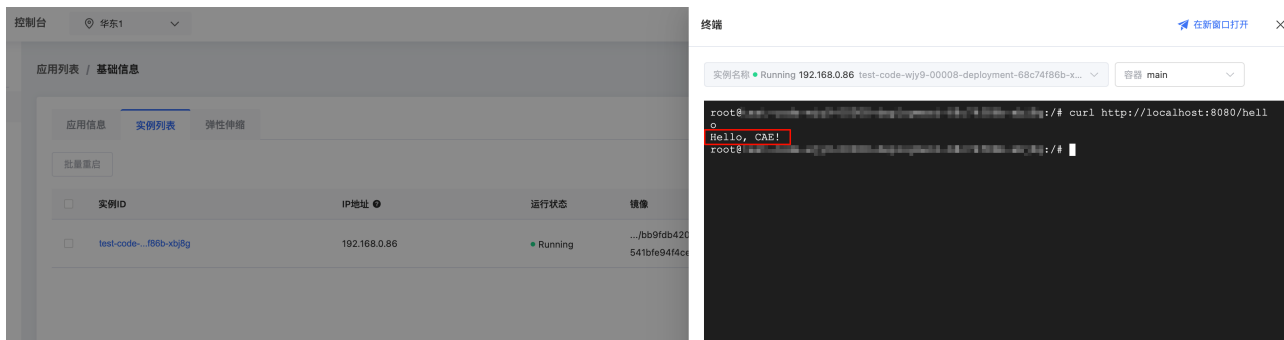
Java代码包部署示例

- 类型：Web 应用框架，企业级
- 框架：Spring Boot 4.0.2
- 运行环境：Java 17
- 启动命令：java -jar target/hello-cae-java-0.0.1-SNAPSHOT.jar
- 代码包：[hello-cae-java.zip](#)
- 验证：进入终端，执行 curl http://localhost:8080/hello



Python代码包部署示例

- 类型：Web 应用框架，轻量级
- 框架：Flask 2.2.0
- 运行环境：python 3.9
- 启动命令：python3 run.py
- 代码包：[hello-cae-python.zip](#)
- 验证：进入终端，执行 curl http://localhost:8080/hello

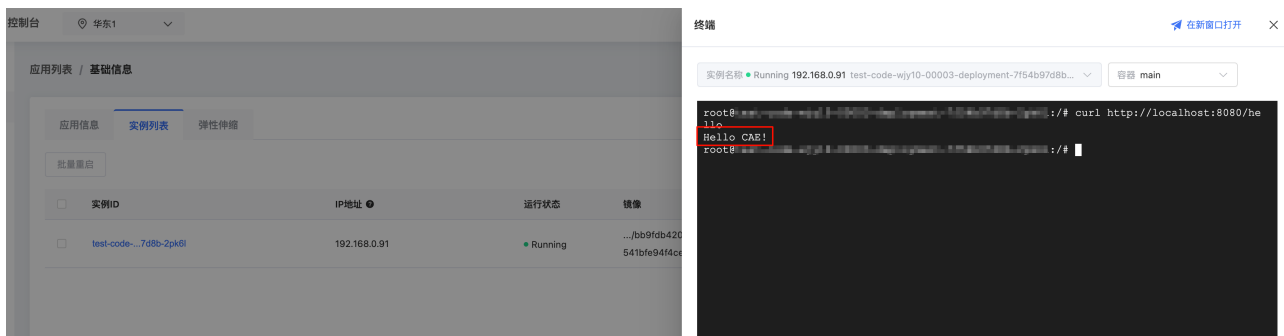


Go 代码包部署示例

- 类型：高性能 Web 应用框架，企业级
- 框架：Hertz 0.9.6
- 运行环境：go 1.24
- 启动命令：

```
chmod +x hello-cae-go-linux
./hello-cae-go-linux
```

- 代码包：[hello-cae-go.zip](#)
- 验证：进入终端，执行 `curl http://localhost:8080/hello`

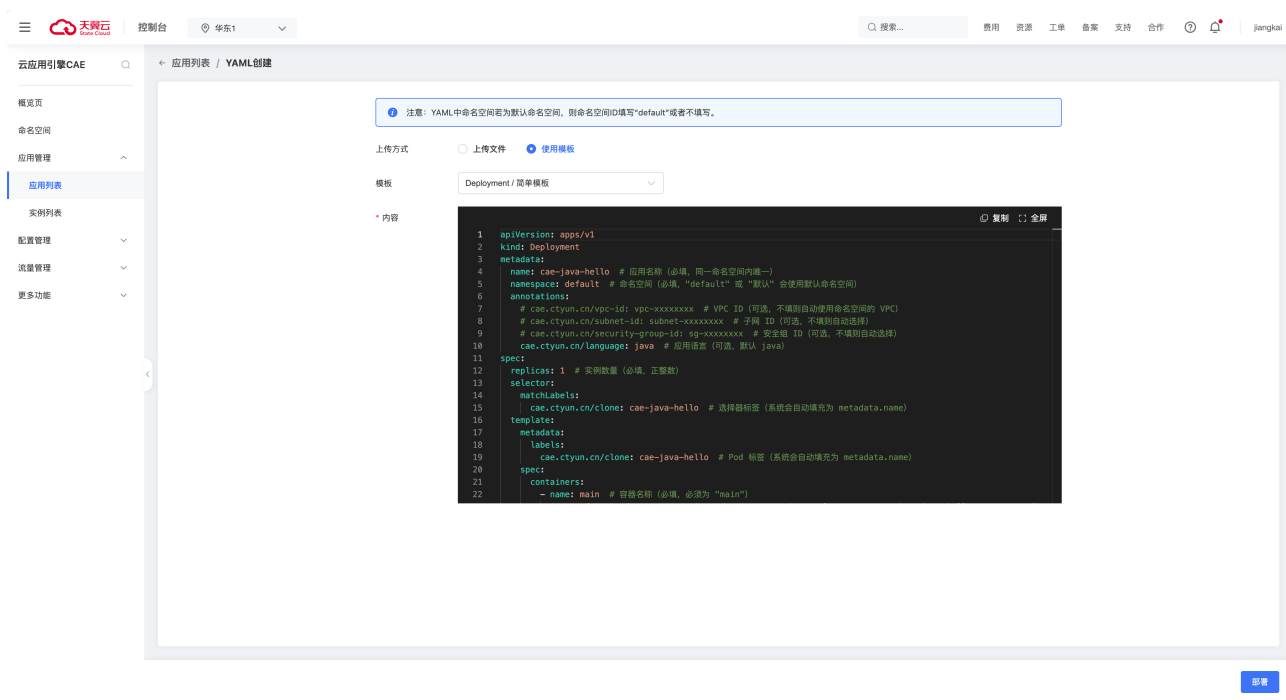


使用YAML部署应用

云应用引擎（CAE）支持通过 k8s YAML 文件部署应用和相关资源。

操作指引

1. 在应用列表页面中，点击使用 YAML 部署应用。
2. 选择上传方式：
 - 上传文件：本地上传 k8s YAML 文件。可用 CAE 参数可参考下方支持的配置参数，或页面提供的模板。
 - 使用模板：从模板下拉列表选择你所需的资源模板，即可查看模板内容，根据实际需要进行编辑调整。
3. 点击右下方部署，等待部署完成。



支持的配置参数

目前支持 Namespace、Deployment、Configmap 和 Secret 四种资源类型。其中，和 CAE 相关的 k8s YAML 配置项说明见下表。

Namespace

配置参数	说明	是否必填	示例值
metadata.name	命名空间名称。	是	test-namespace
metadata.annotations: cae.ctyun.cn/vpc-id	应用关联的 VPC ID。为空时自动取命名空间绑定的 VPC。	否	vpc-abcasd

Deployment

配置项	说明	是否必填	示例值
metadata.name	应用名称	是	my-app
metadata.namespace	应用所在的命名空间。为空时自动取 default，也就是默认命名空间。	否	default
cae.ctyun.cn/vpc-id	应用关联的 VPC ID。为空时自动取命名空间绑定的 VPC。	否	vpc-abcasd
cae.ctyun.cn/subnet-id	应用关联的子网 ID。为空时自动取VPC下的默认子网，或可用子网。	否	subnet-abcasd
cae.ctyun.cn/security-group-id	应用关联的安全组 ID。为空时自动取VPC下的默认安全组，或可用安全组。	否	sg-abcasd
cae.ctyun.cn/language	应用语言。为空时默认 Java。支持 Java、Golang、Python、PHP 等，不区分大小写。	否	Java

配置项	说明	是否必填	示例值
spec.replicas	容器实例数。	是	
spec.template.metadata.labels	应用标签	否	
spec.template.spec.container[i].env	环境变量	否	
spec.template.spec.container[i].envFrom	配置方式是从已有的 ConfigMap 中引用键值对	否	
spec.template.spec.containers[i].volumeMounts spec.template.spec.volumes	将 ConfigMap 配置项作为文件挂载到容器中。	否	
spec.template.spec.container[i].name	使用固定值 main，创建时必须指定此配置项。	是	
spec.template.spec.container[i].image	镜像地址。	是	
spec.template.spec.container[i].resources.limits.cpu	单个容器实例的 CPU 规格。	是	
spec.template.spec.container[i].resources.limits.memory	单个容器实例的内存规格。	是	

ConfigMap

配置项	说明	是否必须	示例值
metadata.name	ConfigMap 名称。	是	
metadata.namespace	ConfigMap 所在的命名空间。	是	
data	ConfigMap 键值对。	是	

Secret

配置项	说明	是否必须	示例值
metadata.name	Secret 名称。	是	
metadata.namespace	Secret 所在的命名空间。	是	
type	Secret 类型	是	
data	Secret 键值对。	是	

应用高级配置

设置启动命令

通过 CAE 控制台使用容器镜像或代码包方式部署应用时，CAE 会根据预设的启动参数来启动容器。如果需要容器启动执行特殊配置或不采用预设的启动参数，您可以通过 CAE 自定义容器的启动命令，以实现特殊配置或覆盖镜像的启动默认参数。

功能入口

场景不同，操作入口也有所不同。

创建应用

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击创建应用
2. 在应用基本信息向导页面进行配置后，单击下一步：高级设置

对正在运行的应用进行变更

警告

注意 重新部署应用后，该应用将会被重启。为避免业务中断等不可预知的错误，请在业务低峰期执行部署操作。

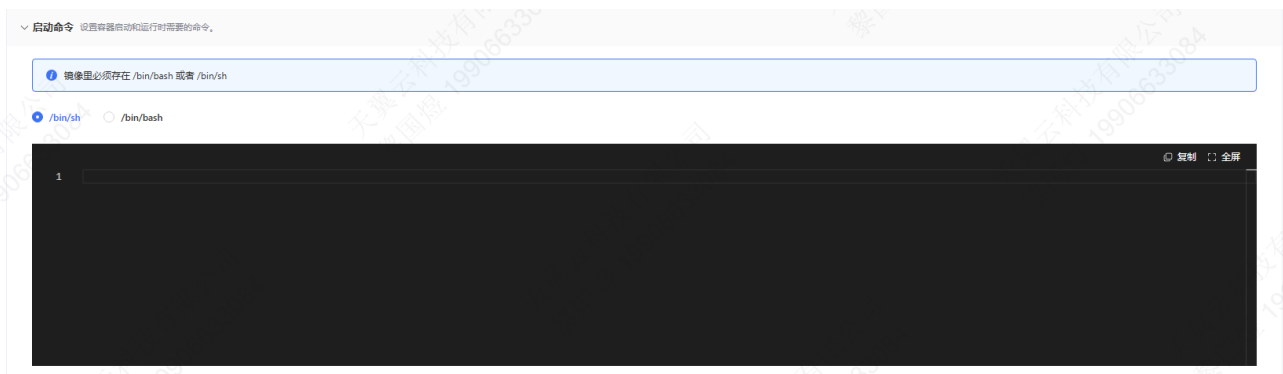
1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击目标应用名称
2. 在目标应用的基础信息页面，单击部署应用

对已停止的应用进行变更

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击目标应用名称
2. 在目标应用的基础信息页面，单击修改应用配置

设置应用的启动命令

在启动命令区域，选择脚本类型，然后配置容器的启动命令



设置环境变量

应用在系统中运行需要配置特定的环境变量，本文介绍如何在 CAE 控制台上为应用设置环境变量，包括自定义环境变量和从命名空间配置项、保密字典中引入的环境变量。

功能入口

场景不同，操作入口也有所不同。

创建应用

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击创建应用
2. 在应用基本信息向导页面进行配置后，单击下一步：高级设置

对正在运行的应用进行变更

警告

注意 重新部署应用后，该应用将会被重启。为避免业务中断等不可预知的错误，请在业务低峰期执行部署操作。

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击目标应用名称
2. 在目标应用的基础信息页面，单击部署应用

对已停止的应用进行变更

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击目标应用名称
2. 在目标应用的基础信息页面，单击修改应用配置

环境变量配置指引

添加自定义环境变量

从类型下拉列表，选择自定义，并输入变量名称和变量值。

The screenshot shows the 'Environment Variables' configuration page. At the top, there are two buttons: 'Create ConfigMap (ConfigMap)' and 'Create Secret (Secret)'. Below them, the 'Type' dropdown is set to 'Custom'. The 'Variable Name' field contains 'demo', and the 'Variable Value' field contains 'test'. There is an equals sign between the name and value fields. A '+ Add' button is at the bottom left.

引用配置项作为环境变量

从类型下拉列表，选择引用配置项，输入变量名称，并在配置项名称、键下拉框中选择已创建的配置项名称和该配置项中的键。支持配置单个键或者全部键。

The screenshot shows the 'Environment Variables' configuration page. At the top, there are two buttons: 'Create ConfigMap (ConfigMap)' and 'Create Secret (Secret)'. Below them, the 'Type' dropdown is set to 'Reference ConfigMap'. The 'Variable Name' field contains 'demo'. The 'ConfigMap Name' field contains 'other-config', and the 'Key' dropdown is set to 'cm_key3'. There is an equals sign between the name and configmap name fields. A '+ Add' button is at the bottom left.

说明

1. 如果您挂载了全部键，每个键的变量名称默认和所选配置项的键名称保持一致，不可自定义。
2. 环境变量名称的长度建议不超过256个字符。云应用引擎将环境变量作为属性存储在应用部署属性中，允许配置多个环境变量。

引用保密字典作为环境变量

从类型下拉列表，选择引用保密字典，输入变量名称，并在保密字典名称、键下拉框选择已创建的保密字典名称和该保密字典中的键。支持配置单个键或者全部键。当选择全部键时无需填写变量名称。

说明

1. 如果您挂载了全部键，每个键的变量名称默认和所选保密字典的键名称保持一致，不可自定义。
2. 环境变量名称的长度建议不超过256个字符。云应用引擎将环境变量作为属性存储在应用部署属性中，允许配置多个环境变量。

服务注册与发现

使用K8s Service配置应用服务访问

云应用引擎支持基于 K8s ServiceName 的多语言服务注册发现功能，通过 ServiceName 做服务间调用。您可以为应用配置一个云应用引擎集群内可以访问的固定域名，解决每次部署以后实例 IP 变化的问题。本文介绍如何通过云应用引擎控制台，为应用配置基于 K8s ServiceName 的服务访问。

背景信息

在分布式微服务框架下，当服务上下线变更时，通过服务注册与服务发现功能，实现服务端和调用端的正常通信。

- 服务注册：针对服务端。服务实例把自身的服务信息注册到注册中心后，调用端从注册中心读取并订阅相关信息。
- 服务发现：针对调用端。注册中心对服务下的实例地址和元数据进行探测，并以预先定义的接口提供给调用端进行查询。服务实例向注册中心发起请求，在获取依赖的服务信息后，再通过这些信息去请求提供的服务。

使用限制

- 该域名仅支持访问 CAE 应用内的实例，不支持其他实例或者集群通过该域名访问 CAE，例如 FC、ACK、ECS。如果其他产品需要访问 CAE，建议您绑定一个私网/公网 ELB 来提供访问。更多信息，请参见 [为应用绑定公网 ELB 访问](#)。
- 该域名不支持使用 ping 命令访问 K8s Service，您可以在代码里通过 K8s Service 以 TCP 或者 UDP 的方式进行访问。建议您使用 curl 命令，测试应用的 Pod IP 来排查网络是否联通。
- 应用 Pod 无法通过 K8s Service 访问到应用 Pod 自身。
- Job 不支持通过 K8s Service 的方式访问应用。
- 如果服务端应用与客户端应用想要进行互相调用，两者都需要先配置 K8s Service。

功能入口

场景不同，操作入口也有所不同。

创建应用

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击创建应用
2. 在应用基本信息向导页面进行配置后，单击下一步：高级设置

对正在运行的应用进行变更

警告

注意 重新部署应用后，该应用将会被重启。为避免业务中断等不可预知的错误，请在业务低峰期执行部署操作。

- 1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击目标应用名称
- 2. 在目标应用的基础信息页面，单击部署应用

对已停止的应用进行变更

- 1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击目标应用名称
- 2. 在目标应用的基础信息页面，单击修改应用配置

K8s Service Name配置指引

在高级设置配置页展开服务注册发现区域，打开启用基于 K8s Service 服务注册发现开关，配置相关信息。

配置项	说明
服务名称	自定义服务名称。必须保证同一用户地域内唯一，创建成功后不允许修改。 说明 说明 默认的服务名称为应用名称。
端口	输入端口，取值范围为[1,65535]。 说明 说明 最多支持添加 5 条端口与协议的配置信息。
容器端口	输入容器端口。
协议	在下拉列表选择目标协议。支持协议如下： • TCP • UDP

基于K8s Service服务注册发现

服务名称

test

端口和协议

端口	容器端口	协议	操作
8080	8081	TCP	删除

+添加

查看访问的域名

配置完成后，在目标应用的基础信息页面，在应用访问设置区域的基于K8s Service访问页签查看。

说明

说明 支持以下三种域名：

- 访问与该应用相同命名空间内的应用，使用服务名称（本示例中为demo）即可进行访问。
- 跨命名空间访问，需要使用以下域名进行访问，以下域名只作为示例说明。
 - demo.{namespaceID}
 - demo.{namespaceID}.svc.cluster.local

应用访问设置

基于 ELB 访问

基于 K8s Service Name 访问

通过域名test-create-svc可实现在相同命名空间内的应用访问，若需要跨命名空间访问应用，可使用以下几种域名：

1.

2.

服务名称

端口	容器端口	协议
8080	8080	TCP

有状态应用配置服务发现

部署有状态应用时，高级设置页面的服务注册发现区域，可以配置实例间服务发现。

实例间服务发现，仅可配置服务名称。必须保证同一用户地域内唯一，创建成功后不允许修改。

服务注册发现

Java 微服务提供自动寻址Nacos Server能力，多语言服务发现通过K8s Service Name来做服务间调用。

Nacos注册中心服务发现

实例间服务发现

服务名称

test-namespace-headless

应用间服务发现

服务名称

test-namespace-service

端口和协议

端口	容器端口	协议	操作
7615	7437	TCP	删除

+添加

配置完成后，可在应用信息的应用访问设置中查看。

说明

说明 支持以下三种域名：

- 访问与该应用相同命名空间内的应用，使用服务名称（本示例中为demo）即可进行访问。
- 跨命名空间访问，需要使用以下域名进行访问，以下域名只作为示例说明。
 - demo.{namespaceID}
 - demo.{namespaceID}.svc.cluster.local
- 如需访问指定实例，可以使用以下域名进行访问。
 - {实例ID}.demo.{namespaceID}
 - {实例ID}.demo.{namespaceID}.svc.cluster.local

应用访问设置

基于 ELB 访问

基于 K8s Service Name 访问

服务名称 test--service

端口	容器端口	协议
7615	7437	TCP

通过域名test--service可实现相同命名空间内的应用访问，若需要跨命名空间访问应用，可使用以下几种域名：

- test--service.cae-f084fe68-6751-4f42-98f4-9bebd6736cd7
- test--service.cae-f084fe68-6751-4f42-98f4-9bebd6736cd7.svc.cluster.local

服务名称(Headless Services) test--headless

通过域名test--headless可实现相同命名空间内的实例访问，若需要跨命名空间访问实例，可使用以下几种域名：

- {实例ID}.test--headless
- {实例ID}.test--headless.cae-f084fe68-6751-4f42-98f4-9bebd6736cd7
- {实例ID}.test--headless.cae-f084fe68-6751-4f42-98f4-9bebd6736cd7.svc.cluster.local

设置Hosts绑定

云应用引擎支持应用级别的实例，通过 Hosts 绑定对主机名进行解析，方便应用实例通过主机名进行访问。本文介绍如何在 CAE 控制台设置 Hosts 绑定。

功能入口

场景不同，操作入口也有所不同。

创建应用

- 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击创建应用
- 在应用基本信息向导页面进行配置后，单击下一步：高级设置

对正在运行的应用进行变更

警告

注意 重新部署应用后，该应用将会被重启。为避免业务中断等不可预知的错误，请在业务低峰期执行部署操作。

- 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击目标应用名称
- 在目标应用的基础信息页面，单击部署应用

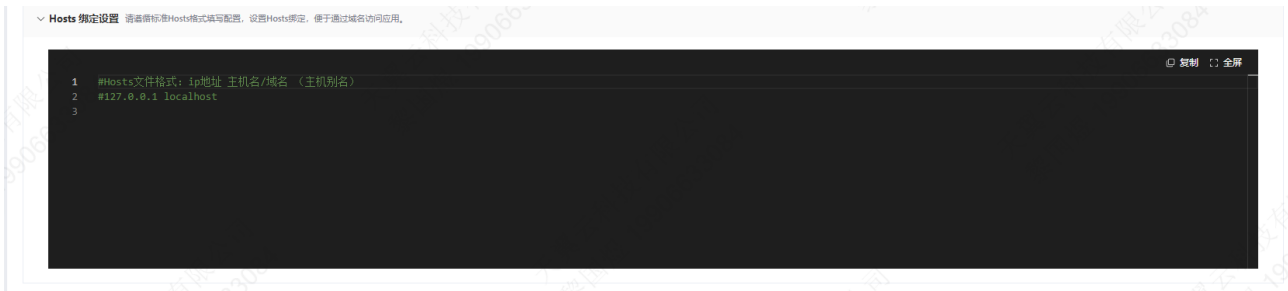
对已停止的应用进行变更

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击目标应用名称
2. 在目标应用的基础信息页面，单击修改应用配置

Hosts配置指引

展开Hosts绑定设置区域，输入Hosts配置项。参考如下示例：

```
# Hosts#####IP## ###/##  
127.0.0.1 localhost
```



持久化存储

设置NAS存储

通常，当容器被销毁时，其内部的数据也会随之丢失，这对生产环境来说可能会产生负面影响。NAS适用于高性能计算和数据共享场景。将NAS挂载至云应用引擎的应用实例，可以有效解决应用数据的持久化存储需求，并实现应用实例之间的数据共享。

功能入口

场景不同，操作入口也有所不同。

创建应用

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击创建应用
2. 在应用基本信息向导页面进行配置后，单击下一步：高级设置

对正在运行的应用进行变更

警告

注意 重新部署应用后，该应用将会被重启。为避免业务中断等不可预知的错误，请在业务低峰期执行部署操作。

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击目标应用名称
2. 在目标应用的基础信息页面，单击部署应用

对已停止的应用进行变更

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击目标应用名称
2. 在目标应用的基础信息页面，单击修改应用配置

NAS配置指引

找到并展开持久化存储区域，配置相关参数。打开启用NAS文件存储的开关。在NAS文件系统所在行的下拉列表，选择待挂载的NAS，并设置挂载源、挂载目录、容器路径、权限和挂载参数。

说明

- 挂载目录只能为根目录/或者非/开头的子目录
- 容器路径不能重复，也不能存在包含关系，例如/tmp和/tmp/nas
- 单击+添加增加挂载路径。最多支持10条挂载路径

持久化存储设置

设置持久化存储数据。

启用NAS文件存储

NAS 文件存储在吞吐和共享上有明显优势，特别适合高性能计算、以及数据共享场景。不建议使用 NAS 作为日志持久化工具，可能会出现多客户端并发导致冲突的问题。

文件系统	vpce 地址	挂载目录	容器目录	操作
		示例: / 或者 xxx/xxx	示例: /tmp/nas	删除

+添加

启用ZOS对象存储

ZOS 对象存储在海量吞吐上有明显优势，特别适合互联网图片、音视频海量文件处理场景，也特别适合网页/移动应用的静态和动态资源分离的场景。不建议使用ZOS作为日志持久化工具，可能导致数据延迟，占用大量内存。

设置OSS存储

与 NAS 相比，OSS 提供了便捷的工具以及控制台，支持可视化管理 Bucket，并在解决应用实例数据持久化和实例间数据分发问题的基础上，进一步降低成本。OSS适用于读多写少的场景，例如挂载配置文件或者前端静态文件等。

功能入口

场景不同，操作入口也有所不同。

创建应用

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击创建应用
2. 在应用基本信息向导页面进行配置后，单击下一步：高级设置

对正在运行的应用进行变更

警告

注意 重新部署应用后，该应用将会被重启。为避免业务中断等不可预知的错误，请在业务低峰期执行部署操作。

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击目标应用名称
2. 在目标应用的基础信息页面，单击部署应用

对已停止的应用进行变更

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击目标应用名称
2. 在目标应用的基础信息页面，单击修改应用配置

挂载OSS指引

挂载OSS

展开持久化存储设置区域，启用 OSS 对象存储。配置Bucket、挂载目录、容器路径以及读写权限等相关信息。如需添加多条信息，请点击添加。

取消挂载OSS

挂载 OSS 后，如果您不再使用 OSS 存储，可以取消挂载 OSS。在云应用引擎控制台取消挂载 OSS 后，您在 OSS 中所存储的数据仍然存在，不会被删除。具体操作为找到需要取消挂载的 OSS 配置条目，单击操作列的删除按钮。

持久化存储设置 设置持久化存储数据。

启用NAS文件存储

NAS 文件存储在吞吐和共享上有明显优势，特别适合高性能计算、以及数据共享场景。不建议使用 NAS 作为日志持久化工具，可能会出现多客户端并发导致冲突的问题。

启用ZOS对象存储

ZOS 对象存储在海量和吞吐上有明显优势，特别适合互联网图片、音视频海量文件处理场景，也特别适合网页/移动应用的静态和动态资源分离的场景。不建议使用ZOS作为日志持久化工具，可能导致数据延迟，占用大量内存。

Bucket	挂载目录	容器路径	权限	操作
	示例: /folderA/folderB/	示例: /tmp/nas		删除

+添加

配置项	说明	示例值
Bucket	已创建的 OSS Bucket。	bucketname
挂载目录	已创建的 OSS 目录或 OSS 对象。 如果 OSS 挂载目录不存在，会触发异常。	示例如下： / - tmp/oss-test/ - tmp/oss-demo.log
容器路径	CAE 的容器路径。如果路径已存在，会覆盖原有路径；如果路径不存在，会新建路径。	/home/admin/app/php/
权限	容器路径对挂载目录资源的权限，取值如下： - 只读 - 读写	只读

注入配置信息

将配置项（ConfigMap）注入容器可以将其转换为容器中的文件，应用可以读取注入的文件作为配置文件，这使得您可以灵活修改应用的配置而无需重新构建应用的镜像。

功能入口

场景不同，操作入口也有所不同。

创建应用

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击创建应用

2. 在应用基本信息向导页面进行配置后，单击下一步：高级设置

对正在运行的应用进行变更

警告

注意 重新部署应用后，该应用将会被重启。为避免业务中断等不可预知的错误，请在业务低峰期执行部署操作。

- 1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击目标应用名称
- 2. 在目标应用的基础信息页面，单击部署应用

对已停止的应用进行变更

- 1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击目标应用名称
- 2. 在目标应用的基础信息页面，单击修改应用配置

注入配置信息指引

注入配置项前，您需要先创建配置项。您可以在应用的命名空间中创建配置项，也可以在当前区域单击创建配置项（ConfigMap），在创建配置项面板进行创建，创建的配置项会同步到应用的命名空间。

展开配置管理区域，单击+添加按钮，您可以将特定配置项（ConfigMap）的单个键或全部键注入容器，以将其转换为容器中的文件。配置项的值将作为文件的内容，您可以自定义文件的挂载路径。

将单个键注入容器

选择已创建的配置项名称和该配置项中的键，并输入挂载路径。挂载路径是注入的文件在容器环境中的绝对路径（含文件名）。如果路径不存在将自动创建。如果路径下存在同名文件，则注入的文件会覆盖原有文件。注入的文件在容器中是只读的。

将全部建注入容器

选择已创建的配置项名称，从键下拉列表选择全部，并输入挂载路径。挂载路径是注入的文件在容器环境中的绝对路径（不含文件名，且不能以/结尾），注入的文件以键名作为文件名。如果路径不存在将自动创建。如果路径下存在同名文件，则注入的文件会覆盖原有文件。注入的文件在容器中是只读的。

卸载配置项信息

若需要卸载配置项信息，可在区域中找到需要卸载的配置项，点击右侧操作列的删除按钮，重新部署应用即可

注意

若应用在运行过程中重新部署，该应用将会被重启。为避免业务中断等不可预知的错误，请在业务低峰期执行部署操作。

配置管理 通过挂载ConfigMap配置文件的方式，向容器中注入配置信息。

创建配置项 (ConfigMap)

配置项名称	配置项键	挂载路径	权限	操作
my-config	cm_key1	/tmp/myconfig/cm_key2	只读	删除
other-config	全部键	/tmp/otherconfig	只读	删除

+添加

注入保密信息

保密字典能够将敏感信息与应用程序代码分离，在应用程序中安全地传递和使用敏感信息。本文介绍如何将云应用引擎命名空间中创建的保密字典以文件方式注入到容器中。

功能入口

场景不同，操作入口也有所不同。

创建应用

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击创建应用
2. 在应用基本信息向导页面进行配置后，单击下一步：高级设置

对正在运行的应用进行变更

警告

注意 重新部署应用后，该应用将会被重启。为避免业务中断等不可预知的错误，请在业务低峰期执行部署操作。

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击目标应用名称
2. 在目标应用的基础信息页面，单击部署应用

对已停止的应用进行变更

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击目标应用名称
2. 在目标应用的基础信息页面，单击修改应用配置

注入保密信息配置指引

挂载保密信息

说明

说明 您可以在命名空间页面提前创建好保密字典，也可以在当前区域单击创建保密字典（Secret），在创建保密字典面板进行创建。

展开保密字典区域，单击+添加按钮，您可以将特定保密字典（Secret）的单个键或全部键注入容器，以将其转换为容器中的文件。保密字典项的值将作为文件的内容，您可以自定义文件的挂载路径。

将单个键注入容器

选择已创建的保密字典名称和该字典中的键，并输入挂载路径。挂载路径是注入的文件在容器环境中的绝对路径（含文件名）。如果路径不存在将自动创建。如果路径下存在同名文件，则注入的文件会覆盖原有文件。注入的文件在容器中是只读的。

将全部键注入容器

选择已创建的保密字典名称，从键下拉列表选择全部，并输入挂载路径。挂载路径是注入的文件在容器环境中的绝对路径（不含文件名，且不能以/结尾），注入的文件以键名作为文件名。如果路径不存在将自动创建。如果路径下存在同名文件，则注入的文件会覆盖原有文件。注入的文件在容器中是只读的。

卸载保密信息

若需要卸载保密信息，可在区域中找到需要卸载的保密字典，点击右侧操作列的删除按钮，重新部署应用即可

警告

注意 若应用在运行过程中重新部署，该应用将会被重启。为避免业务中断等不可预知的错误，请在业务低峰期执行部署操作。

▼ 保密字典 通过挂载保密字典 Secret 的方式，向容器中注入保密信息。

创建保密字典 (Secret)

保密字典名称	键	挂载路径	保密字典类型	权限	操作
my-secret	全部键	/tmp/mysecret/secret_	Opaque	只读	删除
other-secret	全部键	/tmp/othersecret	Opaque	只读	删除

+添加

设置应用生命周期

如果您精通K8s，且需要在应用容器启动前或者关闭前执行相关操作，例如运行前部署资源或者停止前优雅下线应用，可以设置应用生命周期管理。本文主要介绍如何在CAE应用中配置PostStart、PreStop和TerminationGracePeriodSeconds。

功能介绍

- Pod hook（容器钩子）是由Kubernetes管理的kubelet组件发起的机制，用于在容器的生命周期的关键阶段执行自定义任务。这些阶段包括容器进程启动前（preStart）和终止前（preStop）。钩子的引入使Kubernetes能够更加灵活和精细地管理容器的生命周期。
 - 启动后处理（PostStart设置）：是容器生命周期中的一种钩子，该钩子在容器被创建后立刻触发，通知容器它已经被创建。该钩子不需要向其所对应的hook handler传入任何参数。如果该钩子对应的hook handler执行失败，则该容器会被关闭，并根据该容器的重启策略决定是否重启该容器。
 - 停止前处理（PreStop设置）：是容器生命周期中的一种钩子，该钩子在容器终止前触发，通常用于执行清理操作或优雅地关闭服务。其对应的hook handler必须在Kubernetes向容器发送终止信号（如SIGTERM）之前完成。无论hook handler的执行结果如何，一旦其完成，Kubernetes将向容器发送SIGTERM信号以终止容器，并根据配置等待一段时间后强制删除容器。
- 优雅下线超时设置（TerminationGracePeriodSeconds）：其主要作用是为容器的优雅终止提供时间保障，确保应用能够在被强制停止之前完成必要的清理和关闭操作。这不仅有助于提升系统的稳定性和可靠性，还能避免因强制终止导致的数据丢失、请求失败和服务不可用等问题。结合 PreStop钩子可以实现完整的优雅下线流程。

功能入口

操作入口在不同场景下有差异：

创建应用

- 登录CAE控制台，在左侧导航栏选择应用管理 > 应用列表，然后选择目标地域和目标命名空间，最后单击创建应用。

2. 在应用基本信息向导页面进行配置后，单击下一步跳转到高级设置。
3. 在高级设置页面，找到并展开应用生命周期管理区域。

对正在运行的应用进行变更

警告

注意

重新部署应用后，该应用将会被重启。为避免业务中断等不可预知的错误，请在业务低峰期执行部署操作。

1. 登控CAE制台，在左侧导航栏选择应用管理 > 应用列表，然后选择目标地域和目标命名空间，最后单击目标应用名称。
2. 在目标应用的基础信息页面，单击部署应用。
3. 在部署应用页面，找到并展开应用生命周期管理区域。

对已停止的应用进行变更

1. 登录CAE控制台，在左侧导航栏选择应用管理 > 应用列表，然后选择目标地域和目标命名空间，最后单击目标应用名称。
2. 在目标应用的基础信息页面，单击修改应用配置。
3. 在修改应用配置页面，找到并展开应用生命周期管理区域。

操作步骤

在应用生命周期管理区域，按需配置启动后处理（PostStart设置）、停止前处理（PreStop设置）和优雅下线超时设置（TerminationGracePeriodSeconds）。

启动后处理（PostStart设置）

在应用生命周期管理区域的启动后处理（PostStart设置）页签配置PostStart。

例如应用启动前，需要将"Hello from the postStart handler"内容写入/usr/share/message文件。

(在生产环境中，请根据实际业务场景进行配置。本文提供的示例仅为最小化验证配置功能可用性，不可直接用于生产环境。)

应用生命周期管理

生命周期脚本定义，管理应用容器在运行前和关闭前的一些动作，如环境准备、优雅下线等。

启动后处理 (PostStart设置)

☒

应用启动后执行的任务，用于执行初始化任务、通知其他系统等。

镜像里必须存在 /bin/bash 或者 /bin/sh

☒ /bin/sh

☐ /bin/bash

1 echo "Hello from the postStart handler" >> /usr/share/message

结果验证：

应用创建或重新部署完成后，使用Webshell进入应用实例的字符界面，查看/usr/share/message文件是否存在并写入了相应的内容。



停止前处理（PreStop设置）

在应用生命周期管理区域的停止前处理（PreStop设置）页签配置PreStop。

说明

CAE提供了两种不同的Shell解释器可供您选择，分别是：

- >_ /bin/sh
- >_ /bin/bash

例如在容器终止前执行Perform cleanup操作并将其写入到/cleanup.log文件中，然后等待100秒以便进行验证。

(在生产环境中，请根据实际业务场景进行配置。本文提供的示例仅为最小化验证配置功能可用性，不可直接用于生产环境。)



结果验证：

在应用的基础信息页面，单击停止应用，然后使用Webshell进入应用实例的字符界面，查看/cleanup.log文件是否存在并写入了相应的内容。



优雅下线超时设置（TerminationGracePeriodSeconds）

在应用生命周期管理区域的优雅下线超时设置（TerminationGracePeriodSeconds）页签配置优雅下线超时时间。

如果您设置的优雅下线超时时间小于容器优雅终止的时间（包括PreStop钩子执行耗时和容器处理SIGTERM耗时），则在优雅下线超时时间耗尽时强制终止（信号名称：SIGKILL，退出码：137）容器，导致优雅终止流程中断。例如容器优雅终止的时间为100秒，而设置的优雅超时时间为60秒，那么容器会在优雅终止之前被杀死。

添加Sidecar容器

云应用引擎支持在应用中添加 Sidecar 容器，以增强主应用容器的功能。Sidecar 容器可以用于实现监控和日志收集等非业务功能的解耦与标准化。通过这种方式，您可以将一些辅助性任务从主应用容器中分离出来，使主应用容器更加专注于其核心业务逻辑。

功能入口

场景不同，操作入口也有所不同。

创建应用

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击创建应用
2. 在应用基本信息向导页面进行配置后，单击下一步：高级设置

对正在运行的应用进行变更

警告

注意 重新部署应用后，该应用将会被重启。为避免业务中断等不可预知的错误，请在业务低峰期执行部署操作。

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击目标应用名称
2. 在目标应用的基础信息页面，单击部署应用

对已停止的应用进行变更

1. 登录 CAE 控制台，在左侧导航栏选择 应用管理 > 应用列表，然后单击目标应用名称
2. 在目标应用的基础信息页面，单击修改应用配置

添加Sidecar容器配置指引

展开 Sidecar 容器区域，单击+添加按钮，在弹出的 Sidecar 容器面板，按照以下步骤完成添加：

1. 自定义 Sidecar 容器名称
2. 配置拉取镜像的地址
3. 为 Sidecar 容器设置 CPU 资源上限和内存资源上限

说明

说明 Sidecar 容器与主应用容器共享 CPU 和内存资源。为了确保主容器的正常运行，请合理设置 Sidecar 容器的最大可使用资源上限，避免其过度占用资源。

4. （可选）单击高级设置，按需为 Sidecar 容器配置以下参数。
- a. 在启动命令区域为 Sidecar 容器配置启动命令。具体操作，请参见 [设置启动命令](#)。
 - b. 在环境变量区域为 Sidecar 容器配置容器环境中需要运行的变量。具体操作，请参见 [设置环境变量](#)。
 - c. 在配置管理区域通过挂载 ConfigMap 配置文件的方式，向 Sidecar 容器中注入所需的配置信息。具体操作，请参见 [注入配置信息](#)。
 - d. 在共享临时存储区域，设置临时存储目录，并将其挂载到主应用容器和 Sidecar 容器中。

说明

说明 临时存储名称只能包含小写字母、数字和短划线 (-)。



复制应用

如果您希望将当前应用的配置完全复制到其他命名空间，您可以在应用引擎控制台使用应用复制功能。

功能入口

1. 在 CAE 控制台左侧导航栏选择应用管理 > 应用列表。
2. 在应用列表页面，单击目标应用操作列的复制按钮，跳转至复制应用页面。
3. 复制应用页面，您可以按需修改命名空间以及配置信息，并按照界面提示创建当前应用。



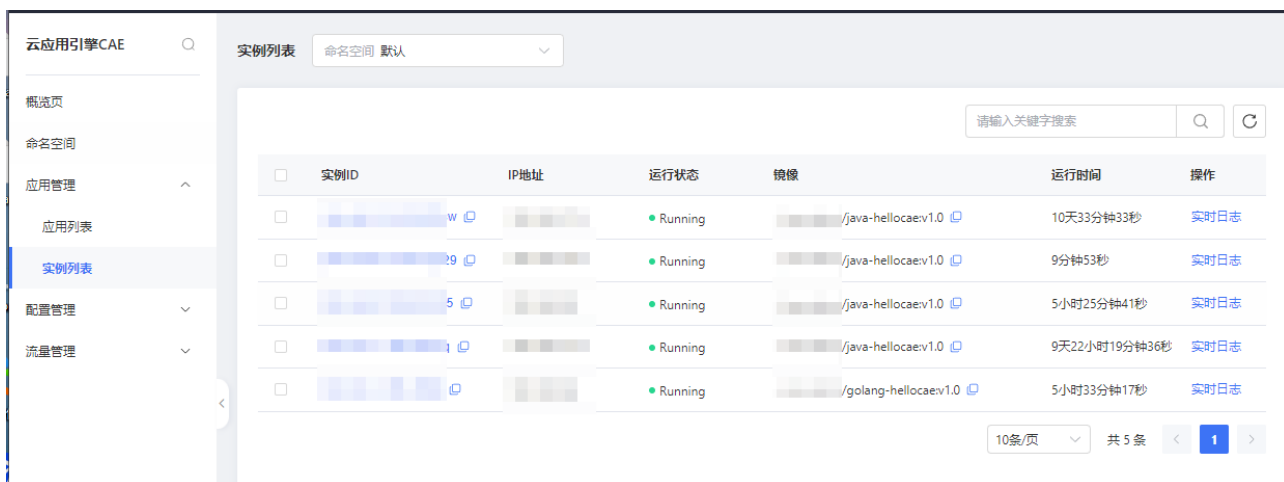
实例列表

使用云应用引擎部署应用后，可以查看应用的实例列表，观察实例和容器的启动状态。

功能入口

查看所有实例列表

1. 登录 CAE 控制台。
2. 在 CAE 控制台左侧导航栏选择应用管理 > 实例列表。
3. 可以筛选特定命名空间下的实例列表。



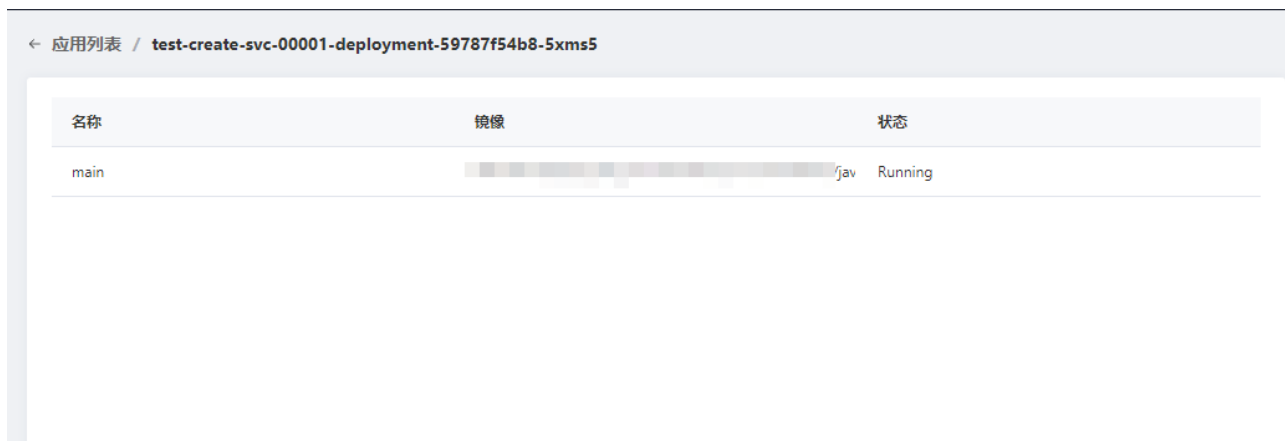
查看目标应用的实例列表

1. 在 CAE 控制台左侧导航栏选择应用管理 > 应用列表。
2. 在应用列表页面，单击目标应用名称，进入应用基础信息页。
3. 在应用基础信息页中选择实例列表页签。



查看实例容器列表

1. 进入实例列表页面。
2. 单击目标实例的名称。



变更记录

当您在云应用引擎上进行应用部署、启动、扩容/缩容等生命周期操作后，可以查看变更记录的创建时间、执行耗时、操作人等关键信息。

功能入口

1. 在 CAE 控制台左侧导航栏选择应用管理 > 应用列表。
2. 在应用列表页面，单击目标应用名称，进入应用基本信息页。
3. 在左侧导航栏中单击变更记录，即可查看应用的历史变更操作。

应用列表 / 变更记录

应用列表

变更记录

应用事件

日志管理

基础监控

应用监控

微服务治理

创建时间	结束时间	执行耗时 (秒)	变更类型	变更状态	描述	来源	操作人
2025-09-04 11:14:42	2025-09-04 11:14:58	16	扩容应用	● 执行成功		console	
2025-09-04 11:14:12	2025-09-04 11:14:13	1	变更日志配置	● 执行成功		console	
2025-09-03 15:44:20	2025-09-03 15:44:36	16	部署应用	● 执行成功		console	
2025-09-03 15:38:45	2025-09-03 15:38:46	1	部署应用	● 执行成功		console	
2025-09-03 15:37:44	2025-09-03 15:37:46	2	部署应用	● 执行成功		console	
2025-09-03 14:43:13	2025-09-03 14:43:15	2	部署应用	● 执行成功		console	
2025-09-03 14:39:40	2025-09-03 14:39:42	2	部署应用	● 执行成功		console	
2025-09-03 14:30:10	2025-09-03 14:30:10	0	关闭应用监控	● 执行成功		system	

应用事件

云应用引擎控制台支持查看K8s原生应用的事件，帮助您了解应用运行时的状态，方便快速聚焦问题。

功能入口

- 1. 在 CAE 控制台左侧导航栏选择应用管理 > 应用列表。
- 2. 在应用列表页面，单击目标应用名称，进入应用基本信息页。
- 3. 在左侧导航栏中单击应用事件，即可查看应用相关的 K8s 事件。

应用列表 / 应用事件

应用列表

应用事件

日志管理

基础监控

应用监控

微服务治理

事件等级	事件原因	来源类型	来源名	首次发生时间	最近发生时间
Normal	Started	Pod	test-svc-00011-deployment-664c5c88c7-rmmkx	2025-09-04 11:15:04	2025-09-04 11:15:04
Started container main					
Normal	Created	Pod	test-svc-00011-deployment-664c5c88c7-rmmkx	2025-09-04 11:15:04	2025-09-04 11:15:04
Normal	Pulled	Pod	test-svc-00011-deployment-664c5c88c7-rmmkx	2025-09-04 11:15:03	2025-09-04 11:15:03
Normal	Pulling	Pod	test-svc-00011-deployment-664c5c88c7-rmmkx	2025-09-04 11:15:03	2025-09-04 11:15:03
Normal	AddedInterface	Pod	test-svc-00011-deployment-664c5c88c7-rmmkx	2025-09-04 11:15:03	2025-09-04 11:15:03
Normal	Started	Pod	test-svc-00011-deployment-664c5c88c7-znsqc	2025-09-04 11:15:03	2025-09-04 11:15:03
Normal	Created	Pod	test-svc-00011-deployment-664c5c88c7-znsqc	2025-09-04 11:15:03	2025-09-04 11:15:03

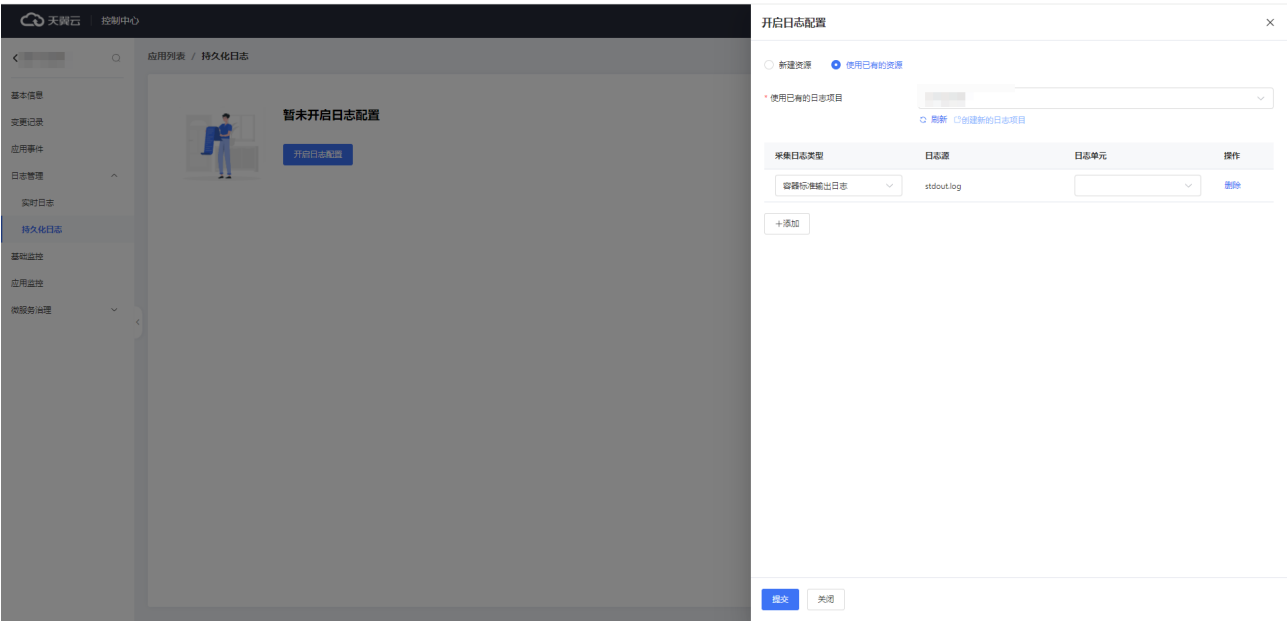
日志管理

当应用出现异常情况时，您可以通过 CAE 控制台查看 Pod 的实时日志或者持久化日志，以便定位问题。

查看持久化日志

创建或者升级应用时，如果您启用了持久化日志功能，CAE 会将业务文件日志收集并输入云日志服务中，您可以无限制行数的查看任意指定时间段的日志，具体步骤如下：

- 1. 登录 CAE 控制台，在左侧导航栏选择应用管理 > 应用列表，选择目标应用。
- 2. 在应用详情页面，在左侧导航栏选择日志管理 > 持久化日志。
- 3. 在持久化日志页面，您可以新建或者使用已有的日志项目，并配置采集相关的参数。
- 4. 单击提交。



基础监控

应用托管到云应用引擎 CAE（Cloud App Engine）后，CAE 对应用所运行设备的 CPU、负载、内存、网络和磁盘进行数据采集与分析，并以动态图的方式展示，方便您实时、直观地了解应用所运行设备的状态。所有监控均以应用为单位进行数据统计和处理。

CAE 基础监控指标

监控项	单位	统计值
应用 cpu	%	平均值 最大值 最小值
应用平均负载	无	平均值 最大值 最小值
应用内存使用率	%	平均值 最大值 最小值
应用内存使用量	字节数	平均值 最大值 最小值
应用磁盘使用率	%	平均值 最大值 最小值
应用磁盘使用量	字节数	平均值 最大值 最小值

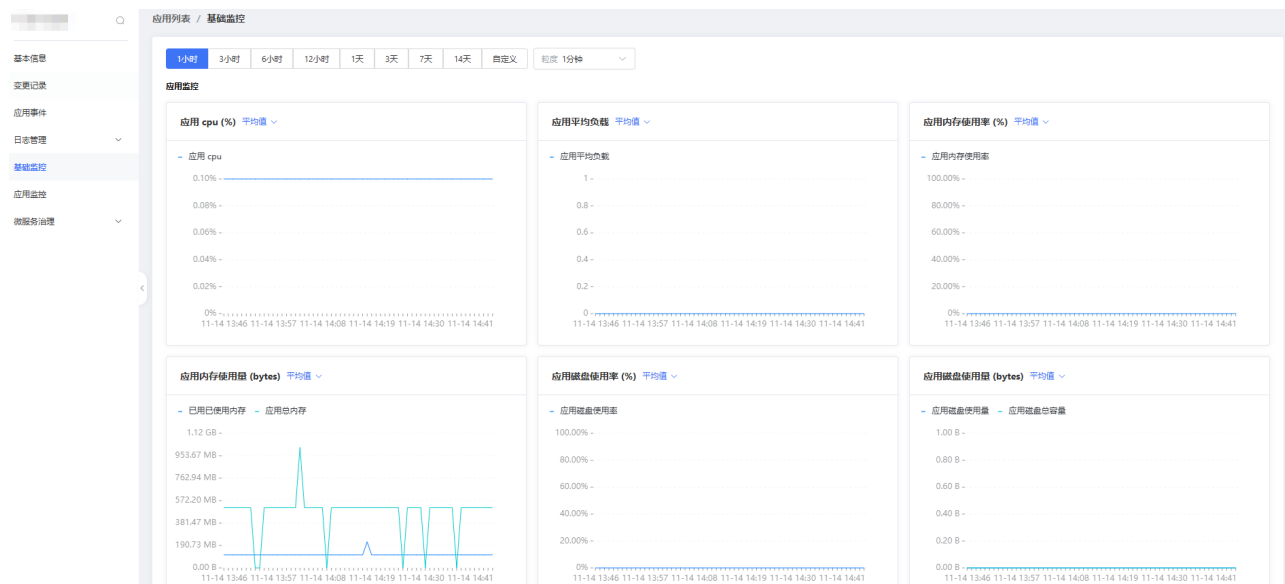
监控项	单位	统计值
网络流入流出速率	字节数/秒	平均值 最大值 最小值
网络数据包	包数量/秒	平均值 最大值 最小值
磁盘 IOPS	次数/秒	平均值 最大值 最小值
磁盘吞吐率	字节数/秒	平均值 最大值 最小值

注意事项

- 由于从数据采集到分析存在一定延时，时延约2分钟，因此 CAE 无法提供百分之百的实时监控视图。
- 如果托管在 CAE 上的应用由于升级或扩缩容导致实例发生变更，将会导致监控数据产生断点，不能保持连续。

查看基础监控数据

1. 登录 CAE 控制台，在左侧导航栏选择应用管理 > 应用列表，单击目标应用名称。
2. 在基础信息页面的左侧导航，单击基础监控。
3. 在基础监控页面，您可以按需选择统计的时间长度区间，并查看相关的监控数据。



回退历史版本

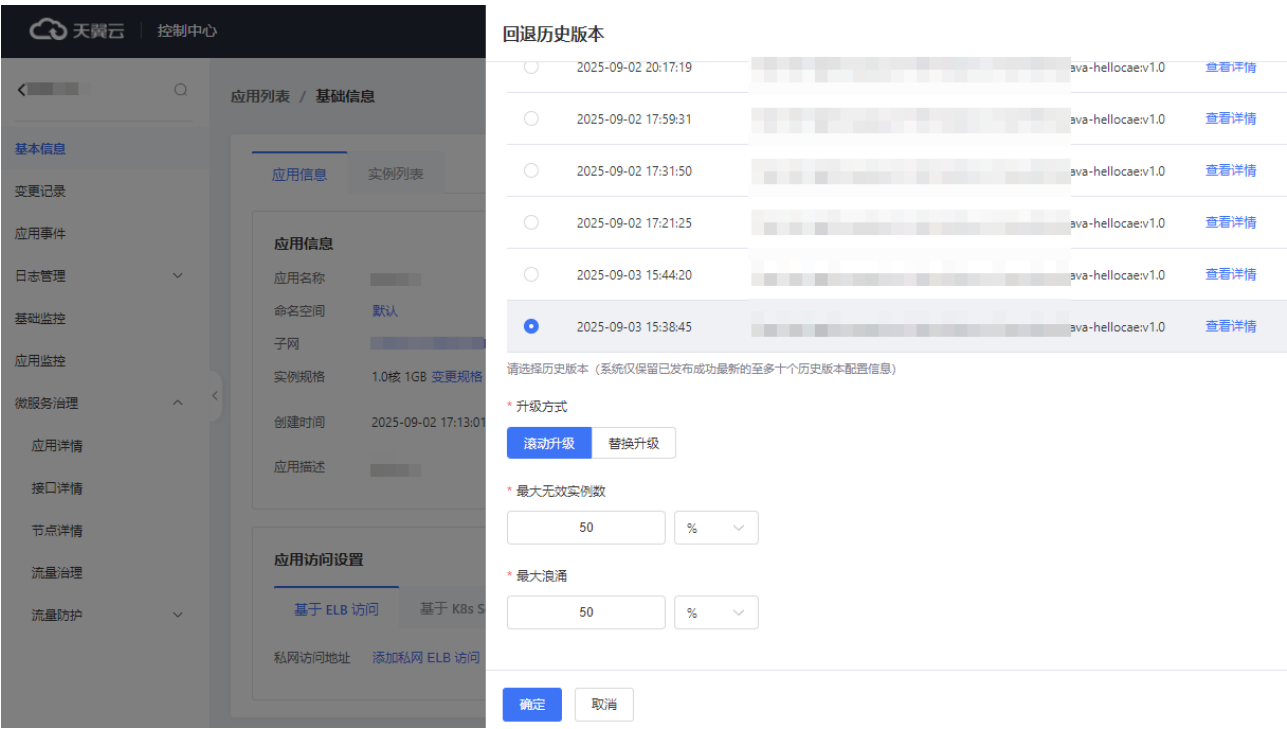
云应用引擎创建完成后，还需不断迭代升级，如果升级的版本出现问题，需要回退至所需的历史版本。

说明

系统仅保留已发布成功最新的至多十个历史版本配置信息

功能入口

- 1. 在 CAE 控制台左侧导航栏选择应用管理 > 应用列表。
- 2. 在应用列表页面，单击目标应用名称，进入应用基本信息页。
- 3. 在基础信息页面，单击回退历史版本。
- 4. 在回退历史版本页面，选择回退历史版本，并配置升级方式。
- 5. 单击确定，即可进行回退操作。



手动扩缩

应用扩缩是通过改变应用的实例数来增加或减少应用的计算容量。您可以在应用的实例负载过高时以手动方式添加新应用实例，在应用闲置时减少应用实例，能够高效利用应用资源、降低成本。

功能入口

- 1. 在 CAE 控制台左侧导航栏选择应用管理 > 应用列表。
- 2. 在应用列表页面，单击目标应用名称，进入应用基础信息页。
- 3. 在基础信息页面，单击手动扩缩。
- 4. 在应用扩缩对话框中，滑动指针或者单击+/-来设置目标实例数。

The screenshot shows a dialog box titled '应用扩缩' (Application Scaling). It contains two main sections. The first section, '应用目标实例数' (Application Target Instance Count), features a horizontal slider with a blue circle handle and a numeric input field showing '1'. To the right of the input are minus and plus buttons. The second section, '应用当前实例数' (Application Current Instance Count), shows the current value '1'. At the bottom right, there are two buttons: '取消' (Cancel) and '确定' (Confirm).

5. 单击确定即可进行手动扩缩操作。

弹性扩缩

功能介绍

弹性伸缩是云应用引擎为微服务应用提供的一项核心能力，用于根据业务负载变化自动调整应用实例数量。通过配置弹性伸缩策略，应用可以在高并发或突发流量场景下自动扩容，保障服务的稳定性和可用性；在业务负载下降时自动缩容，降低资源消耗和运行成本。CAE支持多种弹性伸缩策略，涵盖基于时间的定时伸缩以及基于运行时负载的指标伸缩。用户可根据业务特性灵活选择，实现更精细化的弹性调度能力。

定时策略

定时策略是一种基于时间触发的弹性伸缩方式。用户可通过配置指定的时间窗口，在预期的时间段内将应用实例扩容至设定的实例数，在时间窗口结束后自动恢复至原有规模。

适用场景

- 业务存在明显的周期性流量高峰，如白天/夜间、工作日/节假日
- 秒杀、促销、活动开始前的预扩容场景
- 定时任务、批处理作业在固定时间段集中运行
- 希望在流量到来前提前保障实例容量，提升用户体验

指标策略

指标策略是一种基于应用运行状态的弹性伸缩方式。CAE会持续监控应用的运行指标，并根据指标变化自动调整实例数量。当指标超过或低于设定阈值时，系统将触发扩容或缩容操作，实现按需弹性伸缩。指标策略能够实时响应业务负载变化，适用于流量波动不确定、峰值不可预估的业务场景。

适用场景

- 请求流量、并发量随业务变化不具备明显规律
- 微服务接口存在突发访问或瞬时流量激增
- 希望根据系统负载实现精细化、自动化扩缩容
- 对资源利用率和成本控制要求较高的在线服务

注意事项

- 1. 弹性策略将指标规则与定时规则合并为统一配置，同一时间仅允许启用一条策略。
- 2. 默认情况下，弹性策略包含一条指标规则（最多支持一条），且不预设定时规则。
- 3. 弹性策略最多可添加 5 个定时规则。
- 4. 弹性策略启用时，请勿进行手动管理应用生命周期操作，请停用弹性策略后，再执行。
- 5. 执行应用变更（如应用扩缩、变更规格等）时，无法添加弹性策略。
- 6. 单条定时策略内的两个邻近时间间隔须大于 5 分钟。
- 7. 指标规则的缩容稳定窗口默认为 300s，可在高级设置中修改；定时规则的缩容稳定窗口默认为 0，不支持修改。

功能入口

- 1. 登录CAE控制台，选择目标地域，在左侧导航栏选择应用管理 > 应用列表，然后选择目标命名空间，最后单击目标应用名称。
- 2. 在目标应用的基础信息页面，单击弹性伸缩页签，进入弹性伸缩功能页。

创建指标策略

- 1. 在弹性伸缩功能页，点击“添加弹性策略”。
- 2. 在弹窗中，把“指标规则”右侧的开关打开。
- 3. 填写指标策略相关配置，点击“确定”完成创建。
- 4. 在策略列表页，找到刚才创建的策略，点击“启用”，完成弹性策略的启用。

配置项	说明	示例
规则名称	自定义。	test-metrics
触发条件	CPU和内存指标扩缩容基于HPA（Horizontal Pod Autoscaler）机制，通过持续监控应用实例的CPU或内存使用率，当资源利用率超过或低于设定阈值时，系统将自动触发实例的扩容或缩容，从而保障应用性能并提升资源利用效率。 为避免因资源指标轻微波动导致频繁扩缩容，系统预设了扩缩容容忍度（Tolerance），当实际资源利用率与目标利用率偏差比例未超过容忍度时，不触发扩缩容操作。容忍区间可在弹性伸缩页面，指标监控中查看，具体的计算公式如下： 下限 = 目标利用率 * (1 - Tolerance), Tolerance = 0.32 上限 = 目标利用率 * (1 + Tolerance), Tolerance = 0.1	类型：CPU使用率 目标值：50%
最小实例数	扩缩容期间，保持的最小实例数。	1
最大实例数	扩缩容期间，保持的最大实例数。	10
弹性扩容步长	限制单次扩容时允许增加的实例数量，防止实例数量在短时间内快速膨胀。	3。

弹性扩容周期	限定扩容操作的时间窗口，在指定周期内最多只允许按步长进行扩容。	60 秒
扩容稳定窗口	在扩容决策前参考一段时间内的历史指标，避免因指标短时波动而频繁触发扩容。	0 秒
弹性缩容步长	限制单次缩容时允许减少的实例数量，避免实例数量在短时间内大幅下降。	3
弹性缩容周期	限定缩容操作的时间窗口，在指定周期内最多只允许按步长进行缩容。	60 秒
缩容稳定窗口	在缩容决策前参考一段时间内的历史指标，防止因指标短时波动导致频繁缩容或误缩容。	300 秒
禁止缩容	该策略的缩容规则失效。	关闭

创建定时策略

1. 在弹性伸缩功能页，点击“添加弹性策略”。
2. 在弹窗中，把“定时规则”右侧的开关打开。
3. 填写定时策略相关配置，点击“确定”完成创建。
4. 在策略列表页，找到刚才创建的策略，点击“启用”，完成弹性策略的启用。

配置项	说明	示例
规则名称	自定义。	test-cron
是否长期	根据业务需求选择。	短期
选择时间	策略生效的起始日期，短期时需要选择。	2026-02-03 至 2027-02-03
时区	选择策略生效的时区。	Asia/Shanghai
周期	策略生效周期。支持每天、每周、每月。	每天。
单天触发时间	有效期内、周期内，单天触发的扩缩容时间。	08:00 目标实例数：10

停用弹性策略

1. 在弹性策略列表页，找到需要停用的策略，点击“停用”，完成弹性策略的停用。

预览时间线

1. 当策略存在定时规则时，可在列表页中，点击“预览”按钮，可预览定时策略在最大 90 天内的扩缩容计划。

查看扩缩容事件

系统支持两种方式查看扩缩容事件，

方式1：

1. 在策略列表页，右侧操作项点击“更多-事件”按钮。

方式2：

1. 在弹性伸缩页面，弹性指标变化趋势图中查看。

启动应用

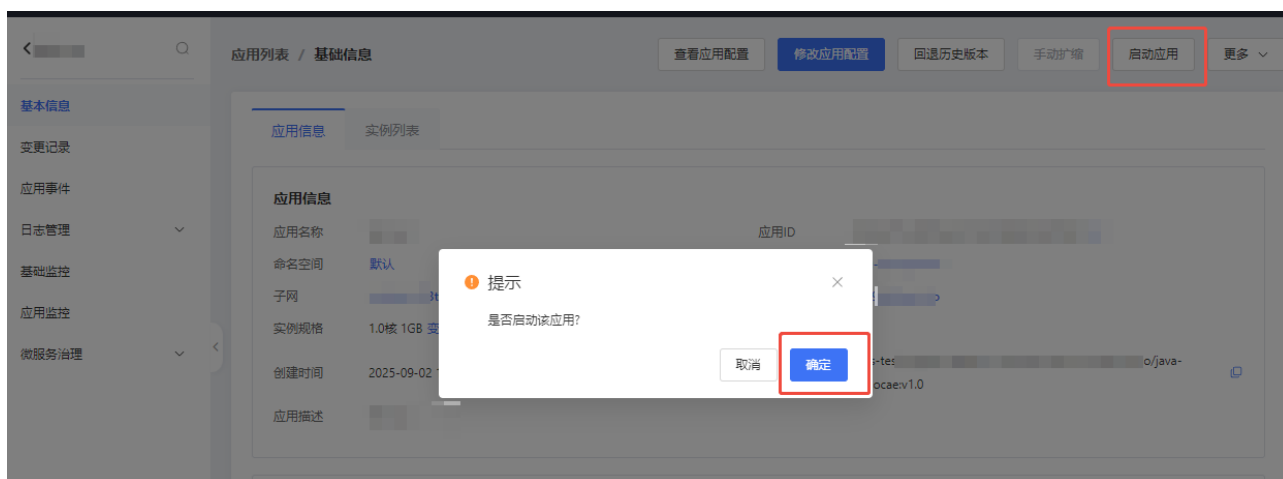
为了节省成本，您可以在业务需要时手动启用已停止的应用。

前提条件

应用处于停止状态。

功能入口

1. 在 CAE 控制台左侧导航栏选择应用管理 > 应用列表。
2. 在应用列表页面，单击目标应用名称，进入应用基础信息页。
3. 在基础信息页面，单击启动应用。
4. 弹出的对话框中进行二次确定即可启动。



停止应用

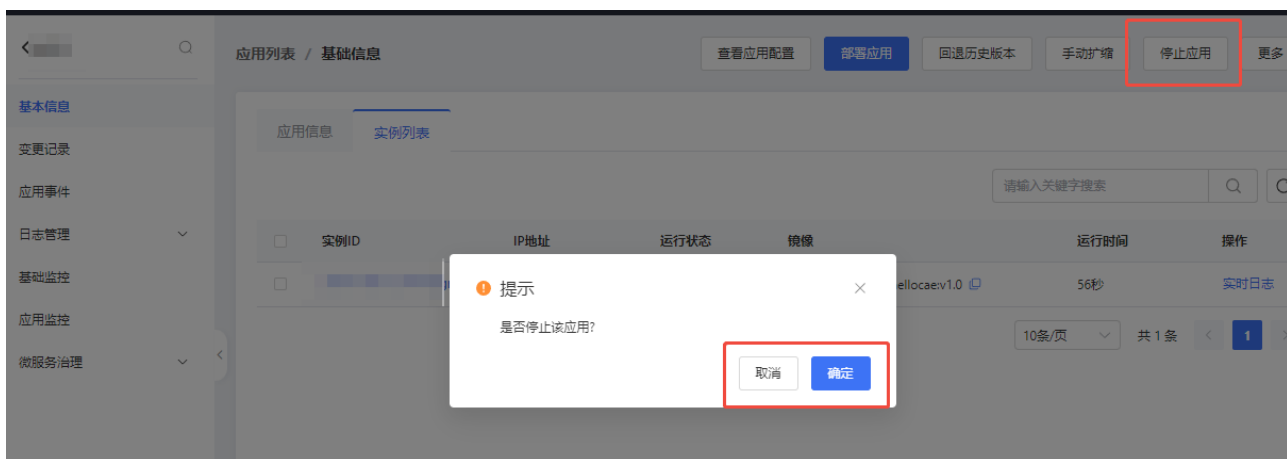
为了节省成本，您可以在业务低谷期时手动停止已启用的应用。处于停止状态的应用不产生计算资源的费用。

前提条件

应用处于启动状态。

功能入口

1. 在 CAE 控制台左侧导航栏选择应用管理 > 应用列表。
2. 在应用列表页面，单击目标应用名称，进入应用基础信息页。
3. 在基础信息页面，单击停止应用。
4. 弹出的对话框中进行二次确定进行停止。



为应用绑定ELB公网访问

在云应用引擎中部署应用后，您可以通过添加公网ELB实现公网访问应用，也可以添加私网ELB实现同VPC内所有应用间互相访问。

功能入口

1. 在 CAE 控制台左侧导航栏选择应用管理 > 应用列表
2. 在应用列表页面，单击目标应用名称，进入应用基础信息页
3. 在应用信息区域下方的应用访问设置进行配置

应用访问设置配置指引

添加私网ELB访问

在 CAE 部署应用后，在 VPC 内部虽然可以通过实例的私网 IP 访问应用，但应用变更或重启会导致实例的私网 IP 发生变化。为解决上述问题，您可以为应用绑定私网 ELB，实现通过一个固定的私网 IP 访问应用，并实现应用实例间的负载均衡。

1. 在应用访问设置区域中，选择基于 ELB 访问页签。
2. 单击添加私网 ELB 访问。
3. 为应用绑定私网 ELB 实例：您可以新建 ELB 实例或绑定已有 ELB 实例。



4. 参考以下说明，至少配置一个监听。如果您需要添加多条监听，请单击添加下一条监听。

配置项	说明	示例值
HTTP协议	- HTTP端口：提供私网访问应用的ELB端口，取值范围为[1,65535]。 - 容器端口：进程监听端口，由程序定义。	- HTTP端口：80 - 容器端口：8080（Web服务的默认端口）
HTTPS协议	- HTTPS端口：提供私网访问应用的ELB端口，取值范围为[1,65535]。 - SSL证书：SSL协议证书，在下拉列表中选择已创建的SSL证书。 - 容器端口：进程监听端口，由程序定义。	- HTTPS端口：443 - SSL证书：在下拉列表中选择已创建的SSL证书。 - 容器端口：8080（Web服务的默认端口）
TCP协议	- CLB端口：提供私网访问应用的ELB端口，取值范围为[1,65535]。 - 容器端口：进程监听端口，由程序定义。	- ELB端口：21 - 容器端口：8080（Web服务的默认端口）

配置项	说明	示例值
UDP协议	- CLB端口：提供私网访问应用的ELB端口，取值范围为[1,65535]。 - 容器端口：进程监听端口，由程序定义。	- ELB端口：49152 - 容器端口：8080（Web服务的默认端口）

- 完成配置后，单击确定。
- 等待 ELB 绑定完成，在私网访问地址区域，可以查看应用的 IP 地址和端口。

添加公网ELB访问

在云应用引擎部署应用后，默认无法从公网访问应用。为解决上述问题，您可以为应用绑定公网ELB，实现通过一个固定的公网IP访问应用，并实现应用实例间的负载均衡。

- 在应用访问设置区域中，选择基于 ELB 访问页签。
- 单击添加公网ELB访问。
- 为应用绑定公网ELB实例：您可以新建ELB实例或绑定已有ELB实例。



- 参考以下说明，至少配置一个监听。如果您需要添加多条监听，请单击添加下一条监听

配置项	说明	示例值
HTTP协议	- HTTP端口：提供公网访问应用的ELB端口，取值范围为[1,65535]。 - 容器端口：进程监听端口，由程序定义。	- HTTP端口：80 - 容器端口：8080（Web服务的默认端口）
HTTPS协议	- HTTPS端口：提供公网访问应用的ELB端口，取值范围为[1,65535]。 - SSL证书：SSL协议证书，在下拉列表中选择已创建的SSL证书。 - 容器端口：进程监听端口，由程序定义。	- HTTPS端口：443 - SSL证书：在下拉列表中选择已创建的SSL证书。 - 容器端口：8080（Web服务的默认端口）
TCP协议	- CLB端口：提供公网访问应用的ELB端口，取值范围为[1,65535]。 - 容器端口：进程监听端口，由程序定义。	- ELB端口：21 - 容器端口：8080（Web服务的默认端口）
UDP协议	- CLB端口：提供公网访问应用的ELB端口，取值范围为[1,65535]。 - 容器端口：进程监听端口，由程序定义。	- ELB端口：49152 - 容器端口：8080（Web服务的默认端口）

5. 完成配置后，单击确定。

6. 等待ELB绑定完成，在公网访问地址区域，可以查看应用的IP地址和端口。

删除ELB访问

1. 在基础信息页面的应用信息页签，找到应用访问设置区域的公网访问地址或私网访问地址，并根据需求单击编辑或删除链接。
2. 在弹出的对话框中根据提示操作，单击确认。

注意

- 在删除公网/私网ELB访问或直接删除应用后，会自动释放 CAE 代购的 ELB 实例。删除后将无法找回相同的 ELB 实例和访问地址，请谨慎操作。
- 删除后，建议您访问负载均衡控制台查看，以确保 ELB 实例已释放。如果 ELB 实例仍处于计费状态，且确认无需继续使用 ELB 实例，请手动释放 ELB 实例，避免持续产生费用。

配置管理

管理和使用配置项（ConfigMap）

命名空间配置项用于存储应用运行所需的各种配置信息。通过配置项，用户可以灵活管理应用在容器中的运行环境，支持多种注入方式：

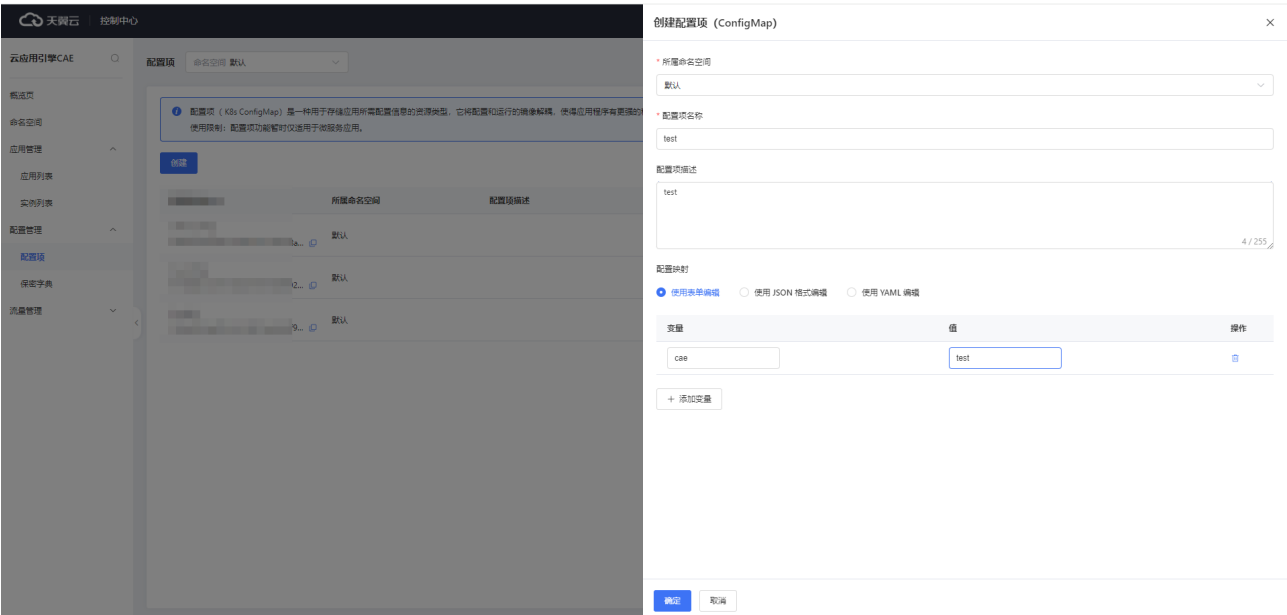
- 环境变量：配置项可在容器中作为环境变量使用，方便在应用部署后随时调整配置。
- 命令行参数：可通过启动命令将配置项传入应用，实现即时配置。
- 挂载配置文件：配置项可写入文件并挂载到容器中，便于集中管理和动态更新配置。本章节将详细说明如何在云应用引擎控制台中创建命名空间的配置项，并展示如何在应用中使用这些配置，实现高效的配置管理与灵活部署。

功能入口

1. 登录 CAE 控制台。
2. 在左侧导航栏选择配置管理 > 配置项，进入配置管理页面。
3. 单击创建，在弹出的创建配置项面板，根据下表说明完成参数配置，然后单击确认。

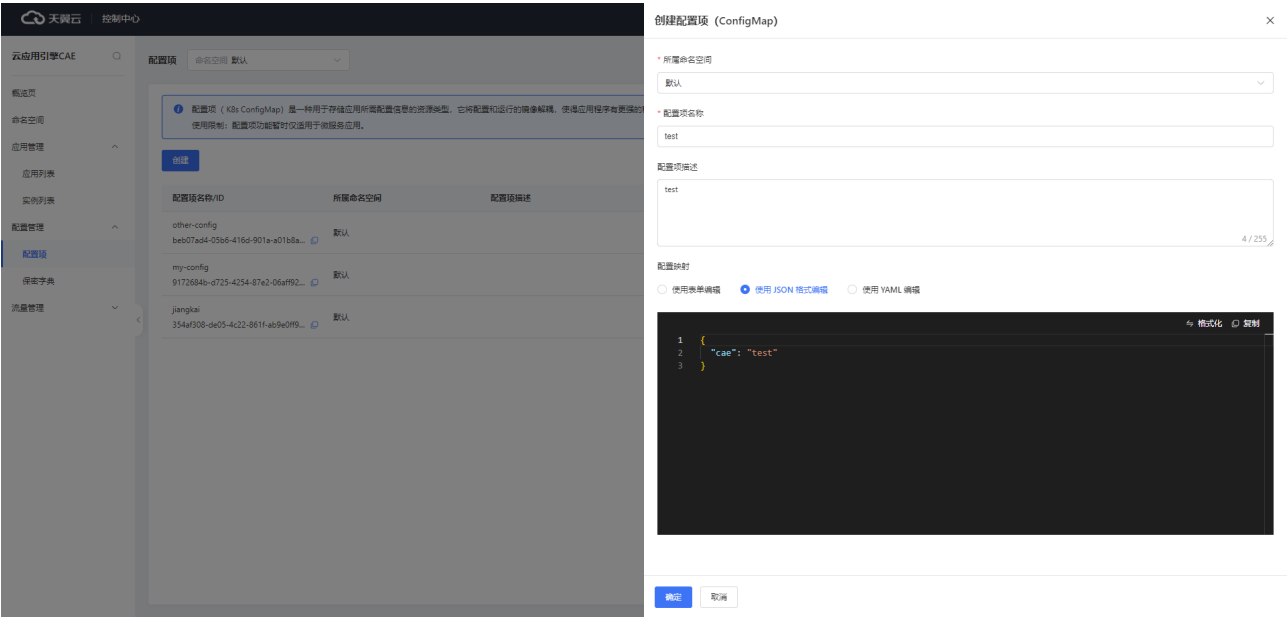
使用表单创建

配置项	说明	示例
所属命名空间	ConfigMap所属的命名空间，仅在该命名空间内有效	默认
配置项名称	自定义的配置项名称	test
配置项描述	对配置项的文字说明，便于识别和管理	这是一个测试 ConfigMap
配置映射	单击“使用表单添加变量”，在对话框中输入变量名和变量值	变量名：env 变量值：test



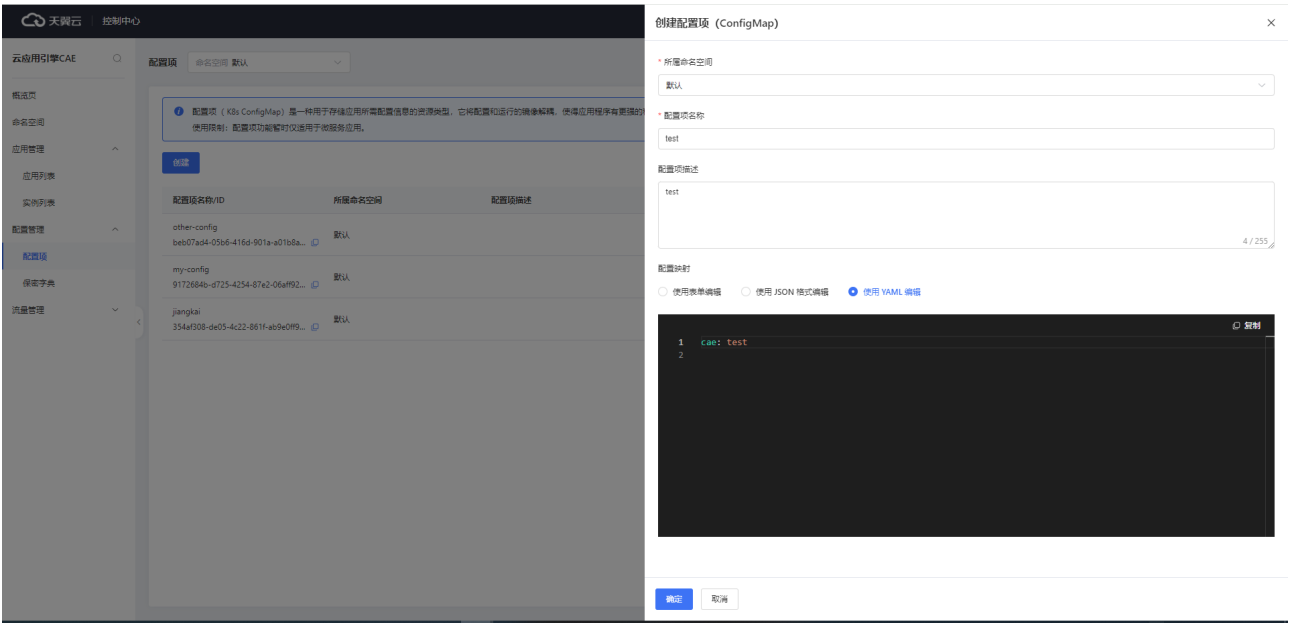
使用 JSON 格式编辑

配置项	说明	示例
所属命名空间	ConfigMap所属的命名空间，仅在该命名空间内有效	默认
配置项名称	自定义的配置项名称	test
配置项描述	对配置项的文字说明，便于识别和管理	这是一个测试 ConfigMap
配置映射	单击“使用JSON 格式编辑”	{ "nv_shell": "/bin/sh"}



使用YAML格式编辑

配置项	说明	示例
所述命名空间	所属命名空间内有效	默认
配置项名称	自定义配置项名称	test
配置项描述	描述配置项	这是一个测试 ConfigMap
配置映射	单击“使用 YAML 格式编辑”	



说明

- 1. 修改配置项后，生成的新版本仅对新创建的实例生效。
- 2. 删除配置项可能导致关联应用无法正常运行。请务必先在关联应用中解除该配置项的使用，再执行删除操作。

管理和使用保密字典（Secret）

命名空间保密字典用于存储应用运行所需的各种配置信息。通过配置项，用户可以灵活管理应用在容器中的运行环境，支持多种注入方式：

- 环境变量：配置项可在容器中作为环境变量使用，方便在应用部署后随时调整配置。
- 命令行参数：可通过启动命令将配置项传入应用，实现即时配置。
- 挂载配置文件：配置项可写入文件并挂载到容器中，便于集中管理和动态更新配置。 本章节将详细说明如何在 云应用引擎控制台 中创建命名空间的配置项，并展示如何在应用中使用这些配置，实现高效的配置管理与灵活部署。

创建保密字典功能入口

- 1. 登录 CAE 控制台。
- 2. 在左侧导航栏选择配置管理 > 保密字典，进入保密字典管理页面。
- 3. 单击创建，在弹出的创建保密字典面板，根据需求选择不同的保密字典类型，并根据下表说明完成参数配置，然后单击确认。

说明

说明 在创建保密字典时，您可以根据需要存放的信息选择不同的类型：

- Opaque类型：Opaque 是默认的 Secret 类型，适用于存储各类没有特定格式的敏感信息，如密码或API 密钥。当您需要存储一些自定义格式的敏感信息时，Opaque 类型提供了灵活性，让您可以以键值对的形式自由地保存所需的数据。
- TLS证书类型：这种类型专门用于存储 TLS 证书相关的数据，当您的服务需要启用 HTTPS 访问时，可以使用这种类型的 Secret 来保存和管理 TLS 证书，以确保在 CAE 环境中的 Pod 中使用，确保数据传输的安全性。
- 私有镜像仓库登录密钥类型：这种类型适用于存储访问私有容器镜像仓库的凭证，使得 Kubernetes 在拉取私有镜像时能够进行身份验证。如果您的应用使用的容器镜像存储在私有仓库中，您需要提供凭证才能访问这些镜像。在这种情况下，该类型的 Secret 用于存储仓库用户名和密码或访问令牌，确保 CAE 在部署应用时能够拉取到私有镜像。

Opaque类型

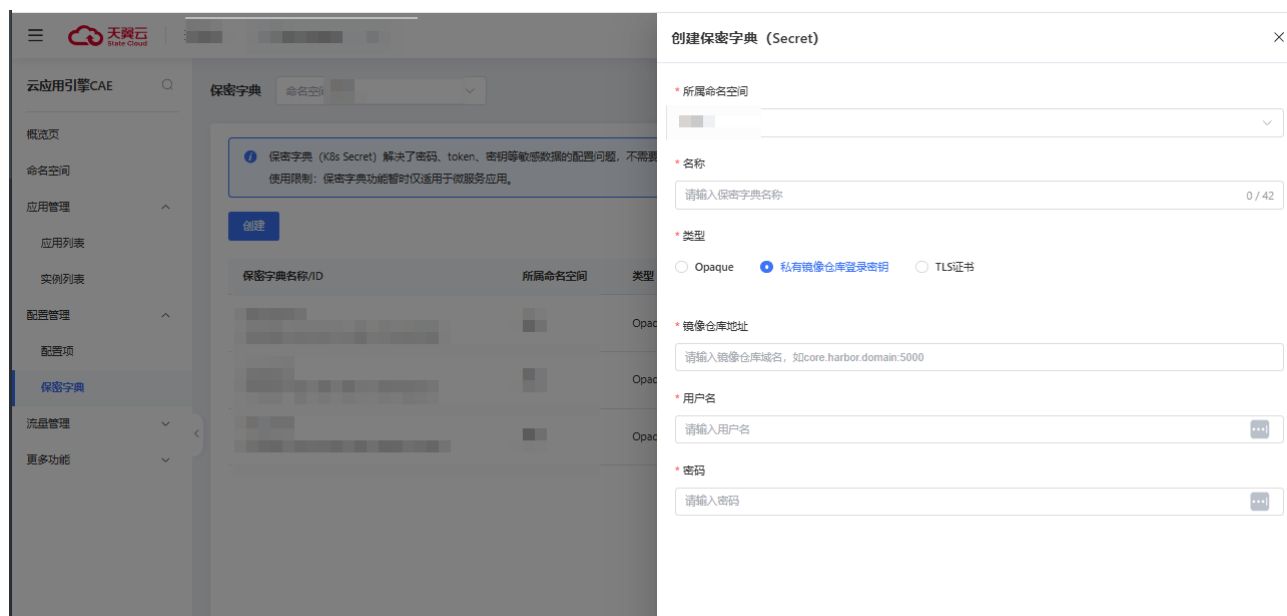
配置项	说明	示例
所属命名空间	Secret 所属的命名空间，仅在该命名空间内有效	默认
名称	自定义的保密字典名称	test
类型	选择 Opaque 类型	Opaque
配置映射	单击添加变量，在对话框中输入敏感信息的 Key 和 Value	变量名：env 变量值：test



TLS证书类型

配置项	说明	示例
所属命名空间	Secret 所属的命名空间，仅在该命名空间内有效	默认

配置项	说明	示例
名称	自定义的保密字典名称	test
类型	选择 TLS 证书类型	TLS证书
Cert	TLS 证书的公钥	-
Key	TLS 证书的私钥	-



私有镜像仓库登录密钥类型

配置项	说明	示例
所属命名空间	Secret 所属的命名空间，仅在该命名空间内有效	默认
名称	自定义的保密字典名称	test
类型	选择私有镜像仓库登录密钥类型	私有镜像仓库登录密钥
镜像仓库地址	镜像仓库地址	默认
用户名	镜像仓库的用户名	admin
密码	镜像仓库的密码	123456



- 说明
- 1. 修改保密字典后，生成的新版本仅对新创建的实例生效。
 - 2. 删除保密字典可能导致关联应用无法正常运行。请务必先在关联应用中解除该保密字典的使用，再执行删除操作。

创建的保密字典将显示在 保密字典 页面，您可以在保密字典列表的 操作 列进行 编辑、复制 和 删除 操作。

流量管理

网关路由

前提条件

- AGW
 - 您已创建云原生 API 网关。
- CAE
 - 您已开通 CAE 且创建了命名空间，并确保云原生 API 网关与 CAE 命名空间绑定同一个 VPC。

创建路由规则

- 1. 在 CAE 网关路由中，在顶部点击创建路由。
- 2. 在创建路由页面，配置相关信息，然后点击保存。

配置项	描述	示例
所属命名空间	需要流量分发的应用所处命名空间	demo
网络名称	路由规则名称	test-route

配置项	描述	示例
网络类型	选择 HTTP 类型	
网关类型	选择 API 网关	
网关实例	选择与命名空间使用同一 VPC 的网关实例。如果您需要创建新的网关实例，点击创建网关实例进行创建	demo-agw
域名	选择需要匹配的域名，支持选择多个域名。	www.app.com
路径	支持多种路径匹配规则 - 前缀：以前缀作为匹配条件。例如，Path 以/demo开头。 - 等价：即完全匹配。例如，Path 等于/demo。	- 条件：前缀 - 路径：/demo
请求方法	设置匹配 HTTP 请求中的 Method 参数	GET
请求头	设置匹配 HTTP 请求中的 Header 参数	- 请求头：user - 条件：等于 - 值：tony
请求参数	设置匹配 HTTP 请求中的 Query 参数	- 请求参数：name - 条件：等于 - 值：tony
Cookie	设置匹配 HTTP 请求中的 Cookie 参数	- Cookie：token - 条件：等于 - 值：test
服务来源	支持 K8s Service	
使用场景	选择当前路由的目标服务类型。 - 单服务：将请求分发到唯一一个后端服务，是最常使用的场景。 - 多服务：将请求按比例分发到多个后端服务，通常在切流和灰度场景下使用。	单服务
后端服务	选择后端应用、服务及其对应的端口、超时时间以及重试次数	
Fallback	开启 Fallback 开关，设置 Fallback 服务，您需要选择指定的服务。当路由指向的后端服务没有可用节点时，原请求会访问此处指定的 Fallback 服务。	

管理路由规则

路由规则新建成功后，可以在网关路由页面进行查看、编辑和上下线等操作。

微服务治理

开启微服务治理

功能介绍

CAE 提供两种微服务治理方案，您可根据应用技术栈及业务场景选择合适的治理模式。

服务网格

基于 Istio 开源框架构建，支持多语言应用无侵入接入，为应用提供统一的流量治理能力。

微服务引擎

仅支持 Java 应用开启，兼容 Spring Cloud、Dubbo 2.x 等主流微服务框架，无需改造业务代码即可接入。

警告

注意 使用全链路灰度功能时，同一泳道组内的所有应用必须启用相同的服务治理模式（服务网格或微服务引擎），否则可能导致灰度流量无法按照预期进行路由。

注意事项

1. CAE 提供服务网格和微服务引擎两种治理方案，同一应用仅支持开启其中一种。
2. 微服务引擎当前仅支持 Java 应用。
3. 开启服务网格后，系统会为应用实例自动注入 Sidecar 代理容器，因此会额外占用一定的 CPU 和内存资源。
4. 开启服务网格后，系统将在应用所在 VPC 内自动创建终端节点，用于与服务网格控制面通信。
5. 开启或关闭服务治理均会触发应用实例重启，请尽量选择业务低峰期进行操作。

功能入口

1. 登录CAE控制台，选择目标地域，在左侧导航栏选择应用管理 > 应用列表，然后选择目标命名空间，最后单击目标应用名称。
2. 在目标应用的左侧菜单栏，单击“微服务治理”，展开二级菜单。
3. 任意点击一个二级菜单，例如“流量治理”，打开服务治理功能介绍页面。

开启服务治理

1. 在服务治理功能介绍页面，选择需要开启的服务治理卡片，点击“开启治理”。
2. 在弹出的确认窗口中，点击“确定”。
3. 系统将自动创建相关资源，开启过程约需1~2分钟。完成后，页面将自动刷新并展示最新状态。

关闭服务治理

1. 进入任意一个服务治理的功能页面，点击页面右上方的“关闭治理”。
2. 在弹出的确认窗口中，点击“确定”。

警告

注意 关闭治理功能后，已有的服务治理规则将会删除，服务治理功能将无法使用，且关闭过程中会触发一次应用重启，请自行判断业务风险后再进行操作。

服务网格

应用详情

功能介绍

通过应用详情页面可查看应用的QPS数据、RT数据、流入流出流量情况、TOP列表等信息。

- 请求数据：展示应用在指定时间范围内的总请求数、通过请求数、异常请求数和拒绝请求数。
- QPS：展示应用在指定时间范围内的总QPS、通过QPS、异常QPS 和拒绝 QPS。
- RT：展示应用在指定时间范围内的P95、P99响应延时。
- 流入流量：以流入应用为统计维度，展示指定时间范围内的流量流入总量及成功率。仅开启服务治理（服务网格）的应用可参与统计。
- 流出流量：以流出应用为统计维度，展示指定时间范围内的流量流出总量及成功率。仅开启服务治理（服务网格）的应用可参与统计。
- 通过QPS TOP：展示通过QPS最高的接口，包含通过流量、拒绝流量、RT和成功率等指标。
- 防护拒绝QPS TOP：展示拒绝QPS最高的接口，包含通过流量、拒绝流量、RT和成功率等指标。
- 平均RT TOP：展示平均RT最高的接口，包含通过流量、拒绝流量、RT和成功率等指标。

前提条件

- 已部署应用。
- 应用已开启服务治理（服务网格）。

功能入口

1. 登录CAE控制台，选择目标地域，在左侧导航栏选择应用管理 > 应用列表，然后选择目标命名空间，最后单击目标应用名称。
2. 在目标应用的左侧菜单栏，单击“微服务治理”，展开二级菜单。
3. 点击“应用详情”。

接口详情

功能介绍

通过接口详情页面可查看接口最近5分钟的请求量、拒绝量、平均RT、成功率。同时，可以查看指定时间范围内的QPS、P95、P99等指标。

- QPS：展示应用在指定时间范围内的总QPS、通过QPS、异常QPS 和拒绝 QPS。
- RT：展示应用在指定时间范围内的P95、P99响应延时。

说明

说明 仅展示最近24小时有流量的接口

接口指标归一化

在微服务场景中，接口路径中常包含动态参数（如 ID、订单号等），容易导致接口维度数量急剧增加，形成接口高基数问题。高基数接口不仅会增加监控与统计的复杂度，还可能影响系统的性能与可观测性效果。

CAE提供了接口指标归一化功能，对接口路径进行归一化处理，将包含动态参数的接口统一映射为标准化路径，从而有效降低接口维度数量，提升统计分析的准确性与系统运行效率。

配置示例：

接口匹配规则	接口名称
^/user/[0-9]+\$	/user/{id}
^/order/[0-9]+\$	/order/{id}

配置完成后，/user/1、/user/2 等实际接口请求的指标数据，都会统一汇总并展示在 /user/{id} 接口下。

前提条件

- 已部署应用。
- 应用已开启服务治理（服务网格）。

功能入口

1. 登录 CAE 控制台，选择目标地域，在左侧导航栏选择应用管理 > 应用列表，然后选择目标命名空间，最后单击目标应用名称。
2. 在目标应用的左侧菜单栏，单击“微服务治理”，展开二级菜单。
3. 点击“接口详情”。

节点详情

功能介绍

通过节点详情页面可同时查看基线应用、灰度应用所有节点最近5分钟的请求量、拒绝量、平均RT、成功率。同时，可以查看指定时间范围内的QPS、P95、P99等指标。

- QPS：展示应用在指定时间范围内的总QPS、通过QPS、异常QPS 和拒绝 QPS。
- RT：展示应用在指定时间范围内的P95、P99响应延时。

说明

说明 灰度应用的节点，会在节点名称右侧展示灰度标签。

前提条件

- 已部署应用。
- 应用已开启服务治理（服务网格）。

功能入口

1. 登录CAE控制台，选择目标地域，在左侧导航栏选择应用管理 > 应用列表，然后选择目标命名空间，最后单击目标应用名称。
2. 在目标应用的左侧菜单栏，单击“微服务治理”，展开二级菜单。
3. 点击“节点详情”。

流量治理

灰度规则

功能介绍

灰度规则用于在业务发布或变更过程中，将部分流量按规则引导至新版本应用，实现平滑发布和风险可控。通过灰度规则，可在不影响全量业务的前提下验证新版本功能和稳定性，降低发布风险。

灰度规则支持按内容灰度和按比例灰度两种方式，满足不同业务场景下的灰度需求。

按内容灰度

按内容灰度基于请求内容进行流量匹配，将符合条件的请求定向到灰度版本。支持对URI、Scheme、Method、Port、Cookie、Header、Parameter 等一个或多个条件进行组合匹配，实现精细化的流量灰度控制。

适用场景：需要对指定用户、特定请求或固定业务流量进行灰度验证，且希望流量命中结果稳定可控的场景。

按比例灰度

按比例灰度根据设定的流量比例，将请求随机分配到灰度版本。

适用场景：适用于对整体流量进行逐步放量验证的新版本发布场景，便于快速评估新版本在真实流量下的整体表现。

注意事项

1. 灰度规则仅在请求方应用开启了服务治理（服务网格）时生效。
2. 请确保需要灰度的灰度应用已开启K8s Service 服务注册与发现功能。

前提条件

- 已部署应用。
- 应用已开启服务治理（服务网格）。

功能入口

1. 登录CAE控制台，选择目标地域，在左侧导航栏选择应用管理 > 应用列表，然后选择目标命名空间，最后单击目标应用名称。
2. 在目标应用的左侧菜单栏，单击“微服务治理”，展开二级菜单。
3. 单击“流量治理”。
4. 单击“灰度规则”页签，进入灰度规则管理页面。

创建灰度规则

在灰度规则页面，点击“创建规则”按钮，在新建规则面板，配置相关信息，然后点击“确定”。

配置项	说明
名称	规则名称。
描述	规则描述。
是否开启	规则开启状态。
灰度类型	按内容灰度：基于请求内容进行流量匹配。 按比例灰度：基于预设的比例进行流量匹配。
灰度场景	灰度规则匹配条件。 按内容灰度时需要填写。

配置项	说明
流量比例	预设的灰度比例。 按比例灰度时需要填写。
路由应用	命中灰度规则时，流量需要转发的应用。

编辑规则

在灰度规则页面，找到目标规则，在其操作列，点击“编辑”。

警告

注意 如果规则状态为启用，编辑后的规则会实时生效。

启用或停用规则

在灰度规则页面，找到目标规则，在其操作列，点击“启用”或“停用”。

删除规则

在灰度规则页面，找到目标规则，在其操作列，点击“删除”。

说明

说明 启用中的规则，无法删除，需要先停用。

警告

注意 删除后的规则无法恢复，请谨慎操作。

故障注入

功能介绍

故障注入用于在可控环境中主动引入异常情况，模拟真实故障场景，以验证微服务系统在异常条件下的稳定性与容错能力。通过故障注入，可提前发现系统潜在风险，提升服务的可靠性和鲁棒性。

故障注入支持延时注入和错误注入两种方式，满足不同故障演练和稳定性测试场景的需求。支持对 URI、Scheme、Method、Port、Cookie、Header、Parameter 等一个或多个条件进行组合匹配，实现精细化的注入。

延时注入：在请求链路中人为增加响应延迟，用于模拟网络抖动、下游服务响应变慢等场景，验证系统在高延时情况下的超时、重试和降级能力。

错误注入：在请求过程中主动返回指定错误，用于模拟服务异常、依赖故障等场景，验证系统的熔断、限流和容错处理能力。

前提条件

- 已部署应用。
- 应用已开启服务治理（服务网格）。

功能入口

1. 登录CAE控制台，选择目标地域，在左侧导航栏选择应用管理 > 应用列表，然后选择目标命名空间，最后单击目标应用名称。
2. 在目标应用的左侧菜单栏，单击“微服务治理”，展开二级菜单。

3. 点击“流量治理”。
4. 点击“故障注入”页签，进入故障注入管理页面。

创建故障注入规则

在故障注入规则页面，点击“创建规则”按钮，在新建规则面板，配置相关信息，然后点击“确定”。

配置项	说明
名称	规则名称。
描述	规则描述。
是否开启	规则开启状态。
注入场景	支持对URI、Scheme、Method、Port、Cookie、Header、Parameter 等一个或多个条件进行组合匹配。
延时注入状态	是否开启延时注入。
延时注入百分比	延时注入的比例。
延时时间	单位：秒 支持的区间：1-180秒
错误注入状态	是否开启错误注入
错误注入百分比	错误注入的比例。
错误码	触发错误后的响应状态码。

编辑规则

在故障注入规则页面，找到目标规则，在其操作列，点击“编辑”。

警告

注意 如果规则状态为启用，编辑后的规则会实时生效。

启用或停用规则

在故障注入规则页面，找到目标规则，在其操作列，点击“启用”或“停用”。

删除规则

在故障注入规则页面，找到目标规则，在其操作列，点击“删除”。

说明

说明 启用中的规则，无法删除，需要先停用。

警告

注意 删除后的规则无法恢复，请谨慎操作。

离群实例摘除

功能介绍

离群实例摘除用于基于请求异常情况自动识别并隔离异常应用实例。当实例在一定时间窗口内连续返回5xx错误时，系统将其判定为异常实例，并通过服务治理策略将该实例临时从负载均衡池中摘除，避免异常实例继续承载业务请求。

通过离群实例摘除能力，可有效降低异常实例对整体服务稳定性的影响，提升系统的容错能力和用户访问体验，同时减少人工干预成本。

注意事项

1. 离群实例摘除规则仅在请求方应用开启了服务治理（服务网格）时生效。

前提条件

- 已部署应用。
- 应用已开启服务治理（服务网格）。

功能入口

1. 登录CAE控制台，选择目标地域，在左侧导航栏选择应用管理 > 应用列表，然后选择目标命名空间，最后单击目标应用名称。
2. 在目标应用的左侧菜单栏，单击“微服务治理”，展开二级菜单。
3. 单击“流量治理”。
4. 单击“离群实例摘除”页签，进入管理页面。

创建离群实例摘除规则

在离群实例摘除规则管理页面，单击“创建规则”按钮，在新建规则面板，配置相关信息，然后单击“确定”。

配置项	说明
名称	规则名称。
描述	规则描述。
是否开启	规则开启状态。
连续错误次数阈值	用于指定实例在连续返回指定次数错误后，将被判定为异常实例并触发摘除。
最小健康实例比例	用于限制至少保留的健康实例比例，防止过多实例被同时摘除。
最大驱逐比例	用于限制单次或同一时间内允许被摘除的实例最大比例。
最短熔断时间	用于指定异常实例被摘除后，最短保持熔断状态的时间。
检测间隔	用于指定系统对实例异常状态进行检测和评估的时间间隔。

编辑规则

在离群实例摘除规则管理页面，找到目标规则，在其操作列，单击“编辑”。

警告

注意 如果规则状态为启用，编辑后的规则会实时生效。

启用或停用规则

在离群实例摘除规则管理页面，找到目标规则，在其操作列，点击“启用”或“停用”。

删除规则

在离群实例摘除规则管理页面，找到目标规则，在其操作列，点击“删除”。

说明

说明 启用中的规则，无法删除，需要先停用。

警告

注意 删除后的规则无法恢复，请谨慎操作。

流控规则**功能介绍**

流控规则用于对服务或接口的访问流量进行控制，通过限制请求速率（QPS）或并发请求数量，防止突发流量、恶意请求或资源竞争导致服务性能下降甚至不可用。当请求量超过设定阈值时，系统将按照配置规则对请求进行限流或拒绝处理，从而保障服务的稳定性和可用性。

通过合理配置流控规则，可有效保护核心业务服务，提升系统在高并发场景下的稳定性和用户访问体验。

流控类型**单机 QPS 限流**

单机 QPS 限流用于限制单个应用实例每秒可处理的请求数量。当请求速率超过实例设定阈值时，超出的请求将被限流或拒绝，以避免实例因瞬时流量过大而出现响应变慢、资源耗尽或服务异常等问题。支持针对指定接口单独配置限流策略，实现更精细化的流量控制。

适用场景：

适用于实例处理能力有限、需要对单个实例进行保护的场景，如接口突发访问、热点接口防护或实例资源受限的应用。

总并发限流

总并发限流用于限制服务整体能够同时处理的请求数量。当系统当前处理中的请求数达到设定阈值时，后续请求将被拒绝，直到已有请求处理完成并释放并发额度。与单机 QPS 限流不同，总并发限流更关注服务当前资源占用情况，能够有效防止因请求积压导致服务响应变慢或资源耗尽。

适用场景：

适用于下游依赖资源有限（数据库、缓存、中间件等）、请求处理耗时较长、需要限制系统整体负载、防止服务因高并发导致雪崩的应用。

前提条件

- 已部署应用。
- 应用已开启服务治理（服务网格）。

功能入口

1. 登录CAE控制台，选择目标地域，在左侧导航栏选择应用管理 > 应用列表，然后选择目标命名空间，最后单击目标应用名称。
2. 在目标应用的左侧菜单栏，单击“微服务治理”，展开二级菜单。
3. 单击“流量治理”。
4. 单击“流控规则”页签，进入流控规则管理页面。

创建单机 QPS 限流规则

在流控规则管理页面，切换到“单机 QPS 限流”页签，点击“创建规则”按钮，在新建规则面板，配置相关信息，然后点击“确定”。

配置项	说明
名称	规则名称。
描述	规则描述。
是否开启	规则开启状态。
单机阈值	单机限流的阈值。
接口限流	针对指定接口配置独立限流策略，支持精确匹配、前缀匹配和正则匹配。 警告 注意 命中接口限流规则的请求会同时消耗全局限流额度。
响应状态码	触发限流时的响应状态码。
响应头	触发限流时，需要设置的响应头。

创建总并发限流规则

在流控规则管理页面，切换到“总并发限流”页签，点击“创建规则”按钮，在新建规则面板，配置相关信息，然后点击“确定”。

配置项	说明
名称	规则名称。
描述	规则描述。
是否开启	规则开启状态。
总并发阈值	服务允许同时处理的最大请求数，超过阈值的新请求将被拒绝。
最大连接数	服务允许建立的最大连接数，超过限制后新连接将被拒绝。

编辑规则

在流控规则管理页面，找到目标规则，在其操作列，点击“编辑”。

警告

注意 如果规则状态为启用，编辑后的规则会实时生效。

启用或停用规则

在流控规则管理页面，找到目标规则，在其操作列，点击“启用”或“停用”。

删除规则

在流控规则管理页面，找到目标规则，在其操作列，点击“删除”。

说明

说明 启用中的规则，无法删除，需要先停用。

警告

注意 删除后的规则无法恢复，请谨慎操作。

熔断规则

功能介绍

熔断规则用于在服务调用异常时，自动中断对异常服务或接口的访问。当下游服务出现持续错误或响应异常时，系统将触发熔断机制，暂时阻断请求，避免故障持续扩散，从而保护上游服务的稳定运行。

通过配置熔断规则，可在异常场景下快速隔离故障服务，并在服务恢复后自动恢复流量，提升系统整体的容错能力和稳定性。

支持对URI、Scheme、Method、Port、Cookie、Header、Parameter等一个或多个条件进行组合匹配，实现精细化的熔断控制。

注意事项

熔断规则作用于被调用方，在服务A调用服务B时，需要在A的出站流量中配置针对服务B的熔断规则，用于控制对B异常实例的访问行为。

前提条件

- 已部署应用。
- 应用已开启服务治理（服务网格）。

功能入口

1. 登录CAE控制台，选择目标地域，在左侧导航栏选择应用管理 > 应用列表，然后选择目标命名空间，最后单击目标应用名称。
2. 在目标应用的左侧菜单栏，单击“微服务治理”，展开二级菜单。
3. 单击“流量治理”。
4. 单击“熔断规则”页签，进入熔断规则管理页面。

创建熔断规则

在熔断规则管理页面，点击“创建规则”按钮，在新建规则面板，配置相关信息，然后点击“确定”。

配置项	说明
名称	规则名称。

配置项	说明
描述	规则描述。
是否开启	规则开启状态。
熔断场景	支持对URI、Scheme、Method、Port、Cookie、Header、Parameter 等一个或多个条件进行组合匹配。
统计窗口时长	用于判断是否触发熔断的时间间隔，默认10秒，最大不超过12秒。
最小请求量	最少请求次数，防止请求量太少时，统计失真。
慢请求阈值	定义慢请求的响应延迟时间，超过该响应时长的请求则被认定为慢请求。
最大慢请求数	超过该次数的新请求会被熔断。
错误率阈值	失败率阈值，时间窗口内错误请求（5XX）比例超过该值，则触发熔断。
熔断时长	熔断时长，默认30s，最大不超过180s。
响应状态码	触发熔断时的响应状态码。
响应头	触发熔断时，需要设置的响应头。
响应体	触发熔断时的响应体。

编辑规则

在熔断规则管理页面，找到目标规则，在其操作列，点击“编辑”。

警告

注意 如果规则状态为启用，编辑后的规则会实时生效。

启用或停用规则

在熔断规则管理页面，找到目标规则，在其操作列，点击“启用”或“停用”。

删除规则

在熔断规则管理页面，找到目标规则，在其操作列，点击“删除”。

说明

说明 启用中的规则，无法删除，需要先停用。

警告

注意 删除后的规则无法恢复，请谨慎操作。

全链路灰度

基于云原生 API 网关实现全链路灰度

名词说明

- 云原生 API 网关：一款面向云原生场景设计的统一流量入口产品，兼容 Kubernetes Ingress 标准，集成了 API 管理、流量转发、微服务访问控制及安全防护等能力，帮助用户快速构建安全、稳定的访问体系。
- 泳道：用于为同一业务应用划分独立的运行隔离单元。只有满足灰度路由规则的请求，才会被引导至对应泳道中的标记应用实例。应用与泳道之间支持多对多关系，便于灵活组织复杂的灰度场景。
- 泳道组：用于对多个泳道进行统一归类和管理，通常用于区分不同团队、业务线或使用场景下的灰度环境。
- 基线环境：用于承载未参与灰度的应用实例，作为默认流量承载环境。当请求未命中任何灰度规则时，将自动路由至基线环境。
- 请求唯一标识：请求会携带此字段以标识请求唯一性，例如 x-trace-id、request-id。

场景示例

本文通过模拟真实的服务调用链路，演示如何基于云原生 API 网关实现全链路灰度。整个过程中无需修改业务代码，仅需在入口应用配置流量规则，相关流量标签即可在调用链路中自动透传。在服务调用过程中，命中规则请求将优先路由至对应的灰度应用；当链路中某一环节不存在匹配的灰度应用时，请求将自动回退至基线应用，确保业务的连续性和稳定性。

各测试应用如下所示：

创建sla、slb和slc三个应用，镜像都选择 Golang 的 cae-demo。sla 应用创建灰度应用 sla-g1，sla-g2，slb 应用创建灰度应用 slb-g2，slc 应用创建灰度应用 slc-g1，slc-g2。以上基线应用和灰度应用都开启 K8s Service 服务注册发现，都打开服务治理（服务网格）。

基线应用	灰度应用	灰度标签
sla	sla-g1	cae.service.tag:g1
	sla-g2	cae.service.tag:g2
slb	slb-g1	cae.service.tag:g1
slc	slc-g1	cae.service.tag:g1
	slc-g2	cae.service.tag:g2

正常流量调用链路：

云原生 API 网关 -> 应用 sla -> 应用 slb -> 应用 slc。

标签流量调用链路：

x-cae-traffic-tag=g1 流量：云原生 API 网关 -> 应用 sla-g1 -> 应用 slb-g1 -> 应用 slc-g1。

x-cae-traffic-tag=g2 流量：云原生 API 网关 -> 应用 sla-g2 -> 应用 slb -> 应用 slc-g2。

前提条件

- 已部署应用。
- 基线应用、灰度应用已开启服务治理（服务网格）。
- 已开通云原生 API 网关，云原生 API 网关实例和应用需要在同一 VPC 中。

操作步骤

部署基线应用并开通服务治理

创建 sla、slb、slc 三个基线应用。

1. 在 CAE 应用列表中，在顶部选择目标地域和命名空间，点击创建应用。
2. 在镜像选择中，选择 Golang -> cae-demo。
3. 在“服务注册发现”中，打开“基于K8s Service服务注册发现”，服务名称分别填写 sla、slb、slc，端口和容器端口都填 8080，协议选择 TCP。
4. 在“微服务治理中”中，打开“服务网格”。
5. 点击“创建应用”。

部署灰度应用

基于基线应用部署灰度应用：

- 其中基线应用 sla 部署两个灰度应用，分别为 sla-g1 应用和 sla-g2 应用。
- 其中基线应用 slb 部署一个灰度应用，为 slb-g1。
- 其中基线应用 slc 部署两个灰度应用，分别为 slc-g1 应用和 slc-g2 应用。

具体操作步骤，如下：

1. 在 CAE 应用列表中，找到刚才创建的基线应用，点击右侧的省略号，再点击“创建灰度应用”。
2. 根据实际情况，灰度标签填写 g1 或者 g2。
3. 在服务注册发现中，打开“基于 K8s Service 服务注册发现”，服务名称分别填写 sla-g1、sla-g2、slb-g1、slc-g1、slc-g2，端口和容器端口都填 8080，协议选择 TCP。
4. 灰度应用，默认会跟踪基线应用开启相应的微服务治理功能。
5. 点击“创建应用”。

创建云原生API网关路由

假设 sla 为入口应用，分别为 sla、sla-g1、sla-g2 创建路由。三条路由名称分别为 sla、sla-g1、sla-g2。实现全链路灰度流量的第一跳。

1. 在 CAE 控制台左侧菜单栏的流量管理中点击“网关路由”。
2. 在顶部选择目标地域和命名空间，点击“创建路由”。
3. 路径选择前缀，填写“/”。
4. 点击更多匹配规则，sla 路由的优先级设置为 0。sla-g1 路由的优先级设置为 1，请求头添加 x-cae-traffic-tag=g1。sla-g2 路由的优先级设置为 1，请求头添加 x-cae-traffic-tag=g2。
5. 使用场景选择“单服务”。
6. 后端服务，sla 路由选择 sla 应用、sla 服务、HTTP 协议、8080 服务端口。sla-g1 路由选择 sla-g1 应用、sla-g1 服务、HTTP 协议、8080 服务端口。sla-g2 路由选择 sla-g2 应用、sla-g2 服务、HTTP 协议、8080 服务端口。
7. 在列表页，三条路由分别操作上线。

创建泳道组

1. 在 CAE 控制台左侧菜单栏的流量管理中点击“全链路灰度”。
2. 在顶部选择目标地域和命名空间，点击“创建泳道组”。
3. 版本类型选择“服务网格”。
4. 入口类型选择“云原生API网关”。
5. 泳道组涉及应用，选择 sla、slb、slc。
6. 请求唯一标识，填写 request-id。

创建泳道

该示例，需要创建两条泳道 g1、g2。

1. 在 CAE 控制台左侧菜单栏的流量管理中点击“全链路灰度”。
2. 在顶部选择目标地域和命名空间，选择对应的泳道组。
3. 点击“创建泳道”。
4. 泳道名称分别填写 g1、g2。
5. 泳道标签选择 g1、g2，系统会自动匹配灰度应用。
6. 灰度模式，选择“按请求头灰度”。

结果验证

1. 可在云原生 API 网关控制台，在左侧菜单栏找到“API”。
2. 点击“cae-http-api”进入详情页，找到路由“sla”，在右侧操作选项中，点击“更多”、“调试”，进入调试页面。
3. 请求路径填写“/a”。
4. Request Headers，添加两个请求头：request-id=1，x-cae-traffic-tag=g1。
5. 点击发送请求。

微服务治理可观测

在网格服务治理页面，右侧的看板可以看到，x-cae-traffic-tag 在不传时，sla、slb、slc 会路由到基线应用。x-cae-traffic-tag 传 g1 时，sla、slb、slc 会路由到 g1 灰度应用。x-cae-traffic-tag 传 g2 时，sla、slc 会路由到 g2 灰度应用，而 slb 会路由到基线应用。

微服务引擎

应用详情

前提条件

- 已部署应用。
- 已开通 [MSE 微服务治理](#)。

说明

说明 使用 MSE 时会产生单独费用。MSE 的计费说明，请参见 [微服务治理计费概述](#)。

功能入口

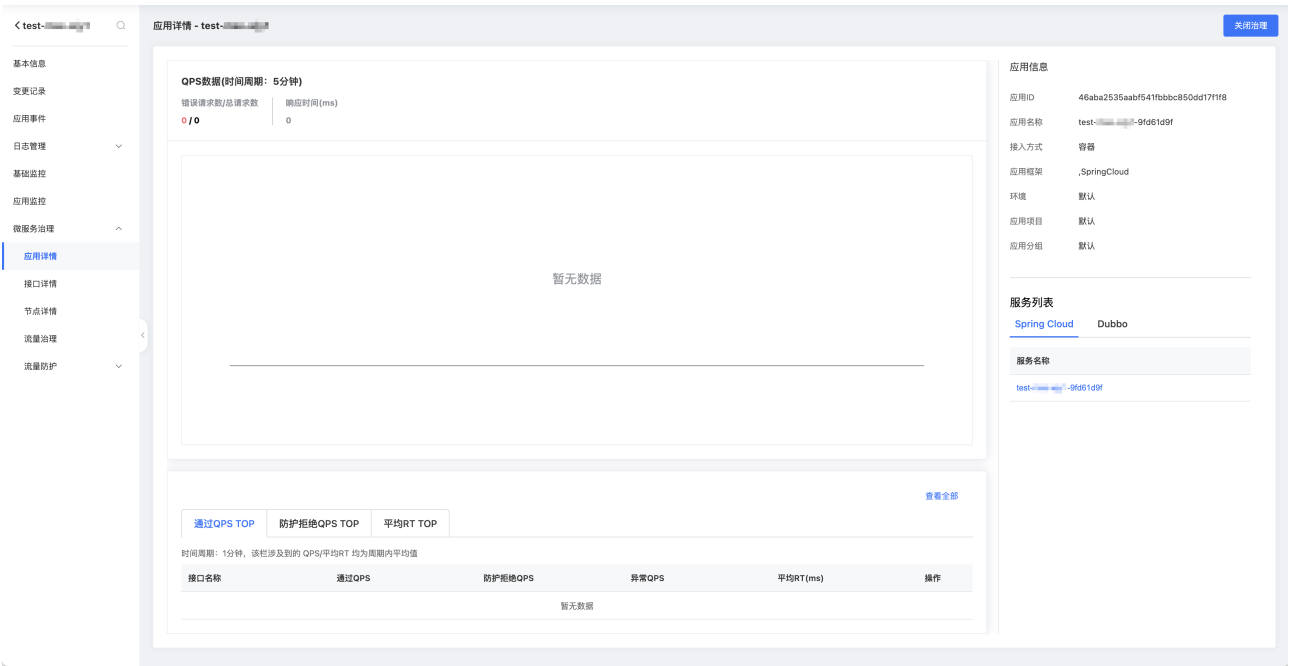
1. 登录 CAE 控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
2. 在左侧导航栏，选择微服务治理 > 应用详情。

功能介绍

应用详情页面会动态刷新，展示错误请求数、总请求数、平均响应时间、应用基础信息和服务列表等数据。

说明

说明 应用详情中涉及到的错误请求数、总请求数、平均响应时间，都是应用入口接口的统计，不包括应用内部方法调用的统计。



- 性能指标：统计应用QPS、RT(ms)等数据。
 - 请求数据：展示近 5 分钟通过实例数、通过请求数和拒绝请求数的走势图。
 - RT(ms)：响应时间，单位为 ms。展示了近 5 分钟响应时间的走势图。
- TOP列表及系统资源指标：包括通过QPS TOP、防护拒绝 QPS TOP、平均 RT TOP接口列表。
 - TOP 接口列表会动态刷新，通过 QPS TOP 页签下接口按照通过 QPS 排序，其他页签下接口按照各自的字段排序。
 - 单击接口名称或该区域右上角的查看全部，进入WEB 服务页签，查看所有接口的 QPS 数据（秒级）、RT 数据（ms）、最大并发（秒级）和事件的详细信息。
 - 单击目标接口操作列的流控或隔离，可为该资源配置相应规则。

接口详情

前提条件

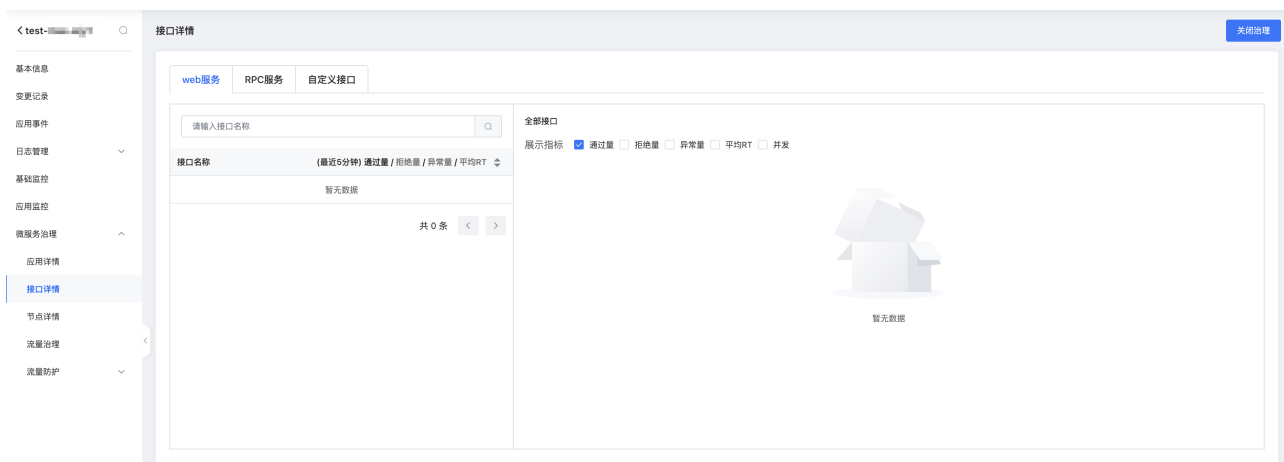
- 已部署应用。
- 已开通 [MSE 微服务治理](#)。

说明

说明 使用 MSE 时会产生单独费用。MSE 的计费说明，请参见 [微服务治理计费概述](#)。

功能入口

- 登录 CAE控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
- 在左侧导航栏，选择微服务治理 > 接口详情。



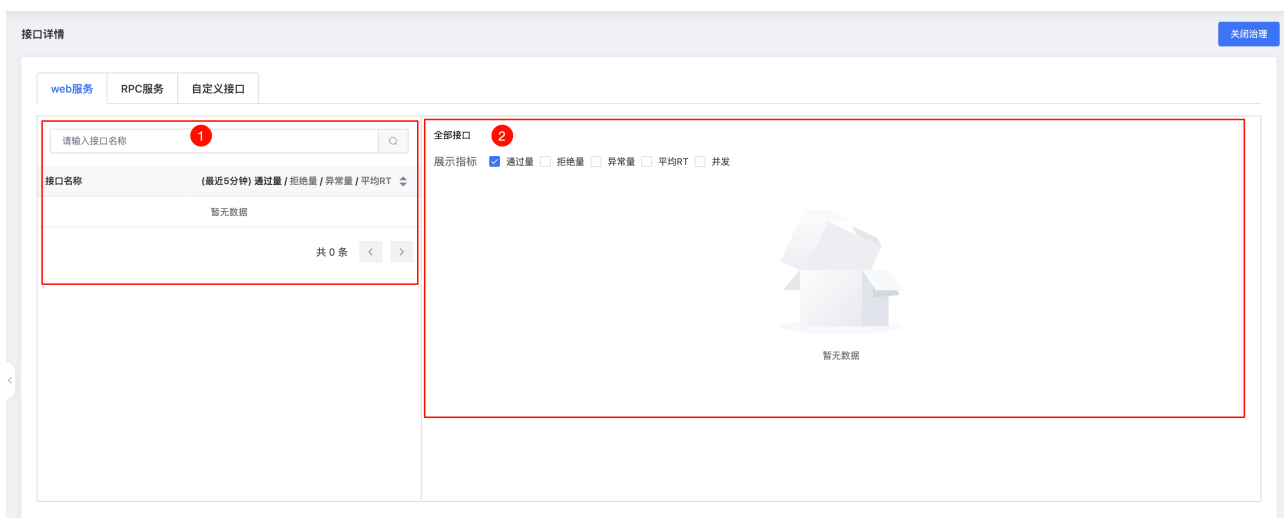
功能介绍

接口详情页面展示了应用的所有接口列表和当前的通过QPS、限流QPS、异常QPS、RT，同时，也展示了接口5分钟内的历史数据。

说明

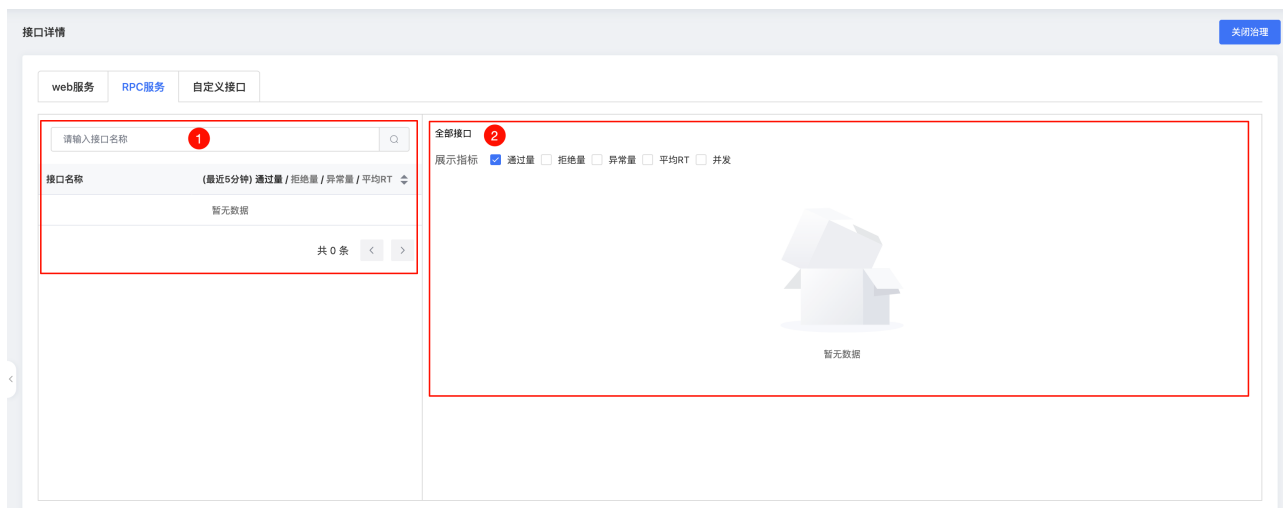
- WEB服务：展示应用 HTTP 服务的接口列表
- RPC服务：展示应用 RPC 调用的接口列表，如 Dubbo 调用

WEB服务



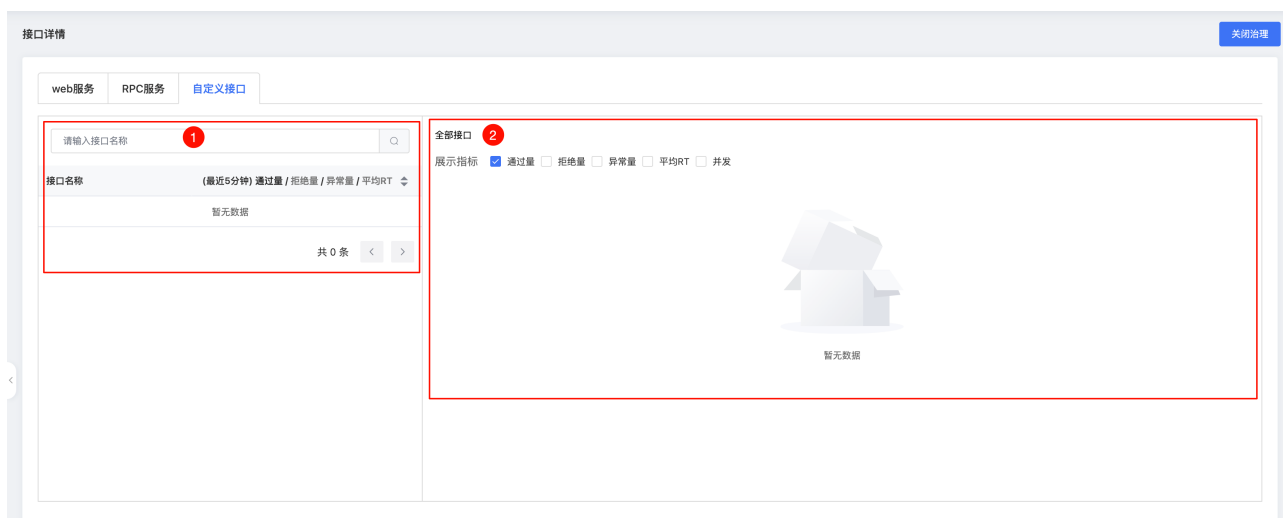
- （图标①）可以预览服务端以及客户端的 WEB 接口列表，以及每个接口最近五分钟的通过量、拒绝量、异常量、平均RT等信息，提供接口筛选能力以及通过量/拒绝量/异常量/平均RT排序。
- （图标②）默认全部接口页签下，展示相关的指标信息。

RPC服务



- （图标①）可以预览服务端以及客户端的 RPC 接口列表，以及每个接口最近五分钟的通过量、拒绝量、异常量、平均RT等信息，提供接口筛选能力以及通过量/拒绝量/异常量/平均RT排序。
- （图标②）默认全部接口页签下，展示相关的指标信息。

自定义接口



- （图标①）可以预览服务接口列表，以及每个接口最近五分钟的通过量、拒绝量、异常量、平均RT等信息，提供接口筛选能力以及通过量/拒绝量/异常量/平均RT排序。
- （图标②）默认全部接口页签下，展示相关的指标信息。

节点详情

前提条件

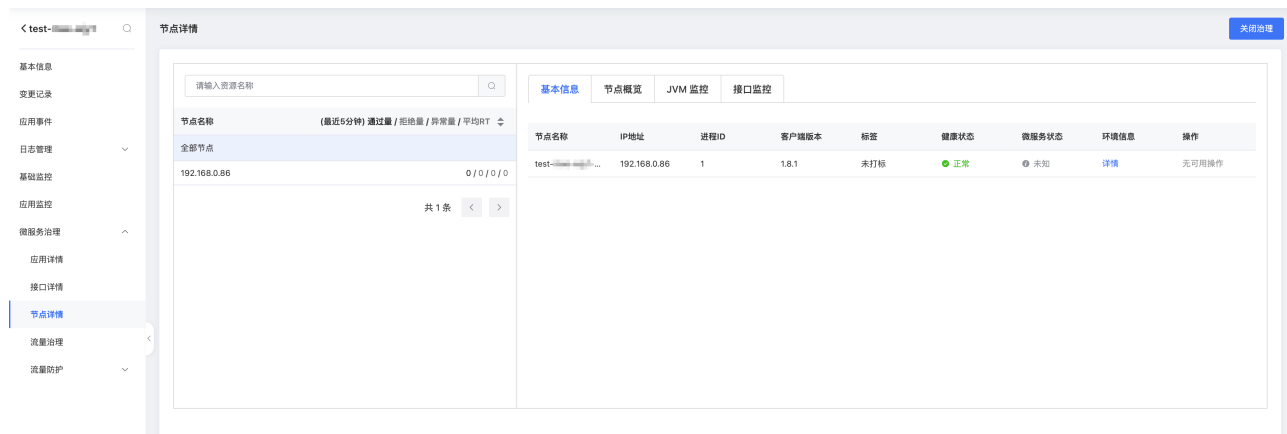
- 已部署应用。
- 已开通 [MSE 微服务治理](#)。

说明

说明 使用 MSE 时会产生单独费用。MSE 的计费说明，请参见 [微服务治理计费概述](#)。

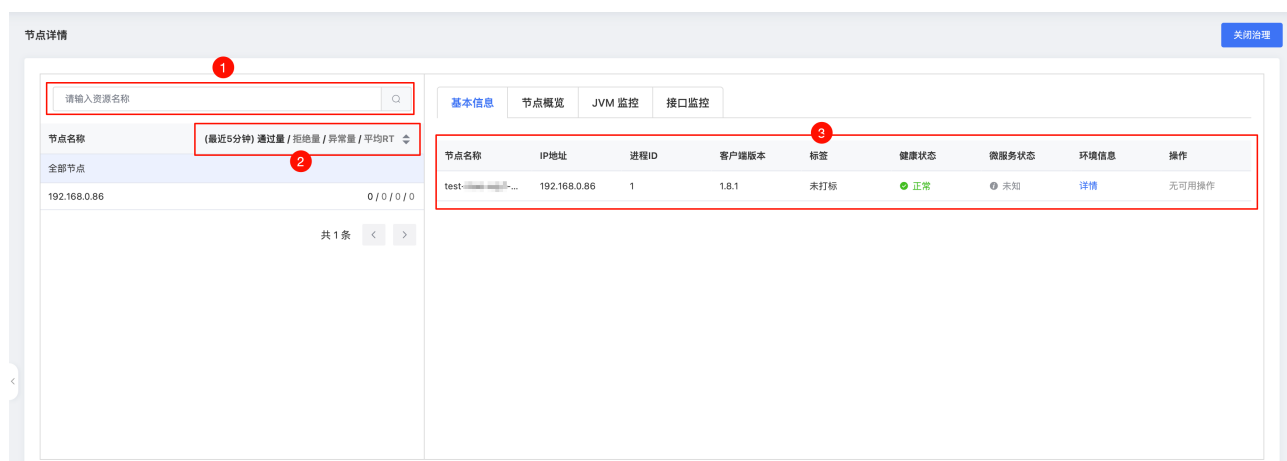
功能入口

1. 登录 CAE控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
2. 在左侧导航栏，选择微服务治理 > 节点详情。



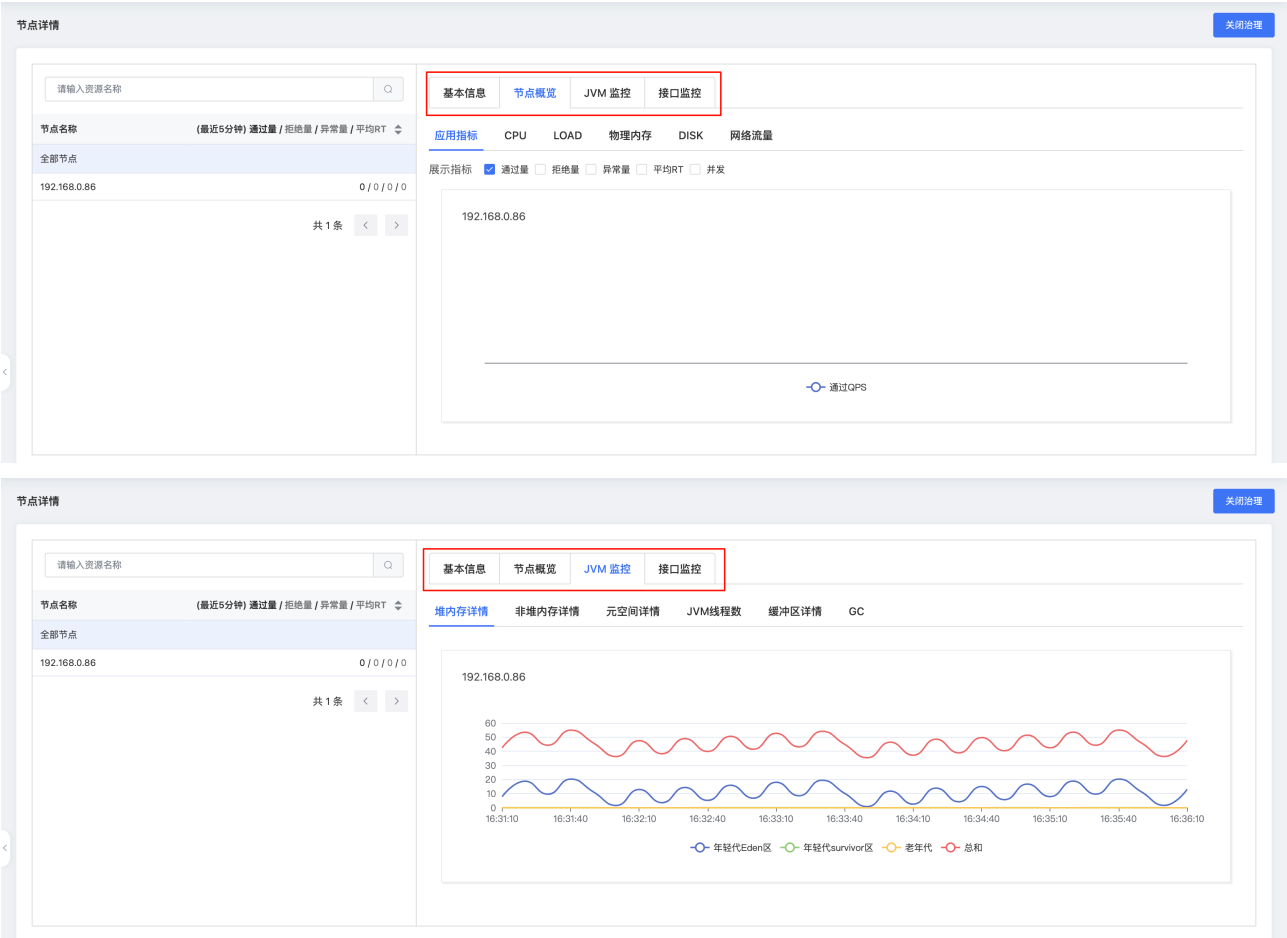
功能介绍

查看节点基础信息



- ①：可按IP或名称搜索节点。
- ②：可按最近五分钟的通过量、拒绝量、异常量与平均RT进行排序。
- ③：展示节点的基本信息。
 - 包括节点名称、IP地址、进程ID、客户端版本、标签、健康状态、微服务状态及环境信息。
 - 在操作列，可手动对当前的节点进行上下线操作。

查看节点详细信息



- 基本信息：可查看节点的基本信息，与节点概览展示的内容和方式相同。
- 节点概览：可查看应用指标、CPU 指标等。
- JVM监控：可查看JVM堆内存详情、元空间详情等。
- 接口监控：可查看当前节点下不同接口的监控指标。

流量治理

无损上下线

使用无损下线

默认情况下，当应用开启服务治理后即可享受无损下线功能，无需额外操作

前提条件

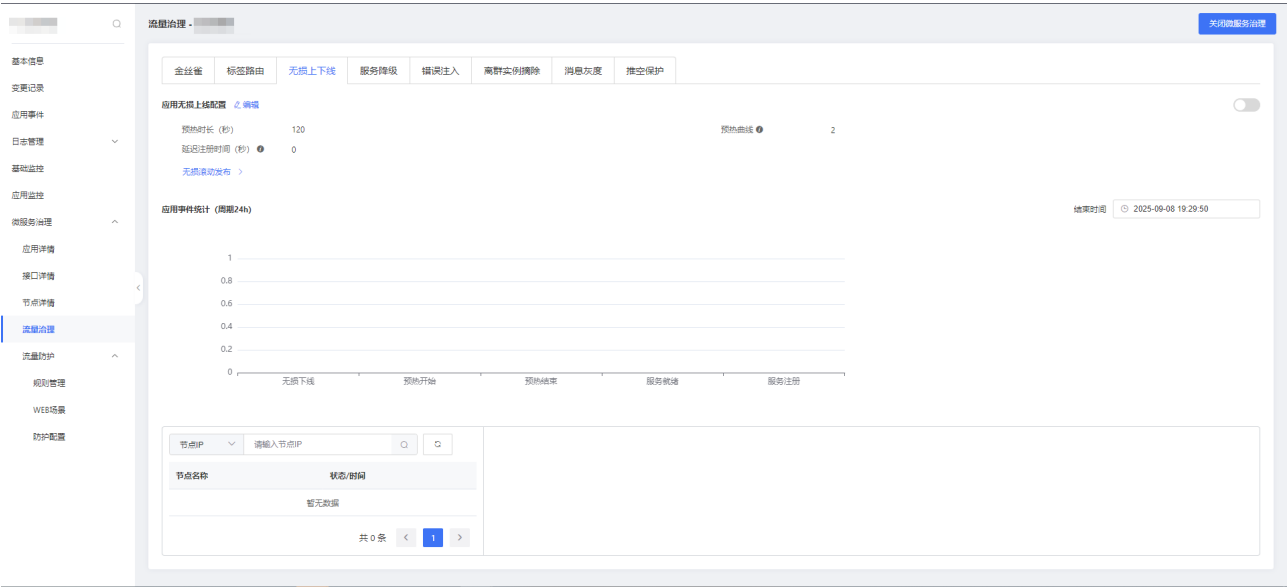
- 已部署应用。
- 已开通 [MSE 微服务治理](#)。

说明

说明 使用 MSE 时会产生单独费用。MSE 的计费说明，请参见 [微服务治理计费概述](#)。

功能入口

1. 登录 CAE控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
2. 在左侧导航栏，选择微服务治理 > 流量治理，单击无损上下线。



服务预热

应用在启动后，由于内部资源还未彻底初始化，直接处理大流量可能会导致出现请求阻塞或报错的情况。在应用刚启动的一段时间内，通过小流量预热的方式使应用完成内部资源的初始化，这样可以有效的避免应用上线后出现的抖动问题。

预热参数说明：

参数	说明	
预热时长（秒）	应用实例下一次启动的预热时间，默认预热时长为120秒。服务预热时长设置范围为0~86400秒（即24小时）。	
预热曲线	默认为2（适合于一般预热场景），表示在预热周期内服务提供者的流量接收曲线形状呈2次曲线形状。预热曲线设置范围为0~20。	

延迟注册

应用在启动时，会存在内部资源还未初始化完成，服务就被暴露到注册中心，被外部的消费者调用，这种情况可能会导致调用报错。通过设置延迟注册，可以在应用的内部资源初始化完成之后再服务注册到注册中心，供外部消费者调用。

参数	说明
延迟注册时间（秒）	延迟注册时间，时长设置范围为0~86400秒（即24小时）。

标签路由

前提条件

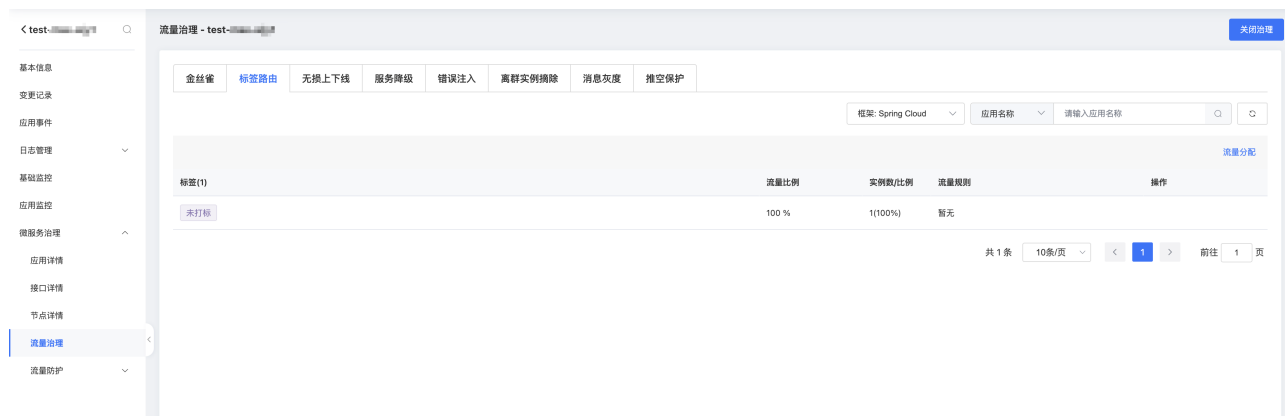
- 已部署应用。
- 已开通 MSE 微服务治理。

说明

说明 使用 MSE 时会产生单独费用。MSE 的计费说明，请参见 [微服务治理计费概述](#)。

功能入口

1. 登录 CAE控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
2. 在左侧导航栏，选择微服务治理 > 流量治理，单击标签路由。



使用标签路由

1. 在标签路由页单击流量分配，为标签按比例分配流量。
2. 在标签路由页的流量规则栏新增流量规则。

流量规则参数说明：

参数	说明
路由名称	路由规则的名称。
应用	所选的应用。
标签	设置的标签名。
应用实例	设置了该标签的实例ip。
是否链路传递	代表开启全链路流控。
框架类型	Spring Cloud Dubbo。
Path	SpringCloud为path路径，Dubbo为接口。
条件模式	满足一个条件，或者满足所有条件。
条件列表	可以设置Header、Cookie、Parameter和Body Content四种参数类型。
是否开启流量规则	流量规则开关。

服务降级

前提条件

- 已部署应用。
- 已开通 [MSE 微服务治理](#)。

说明

说明 使用 MSE 时会产生单独费用。MSE 的计费说明，请参见 [微服务治理计费概述](#)。

功能入口

1. 登录 CAE控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
2. 在左侧导航栏，选择微服务治理 > 流量治理，单击服务降级。

服务降级规则参数说明

参数	说明
规则名称	服务降级规则的名称。
描述	规则的详情描述。
服务提供者应用	服务提供者，被降级的应用。
降级应用	选择应用为待降级应用。
服务降级规则列表	--
框架类型	SpringCloud和Dubbo。
服务路径	服务的接口。
请求方法	GET/POST。
执行策略	所有请求生效/异常请求生效。
降级策略	降级策略分为四种，分别是返回Null值、返回Exception异常、返回自定义Json数据、自定义回调。
默认状态	默认打开或关闭规则。

错误注入

前提条件

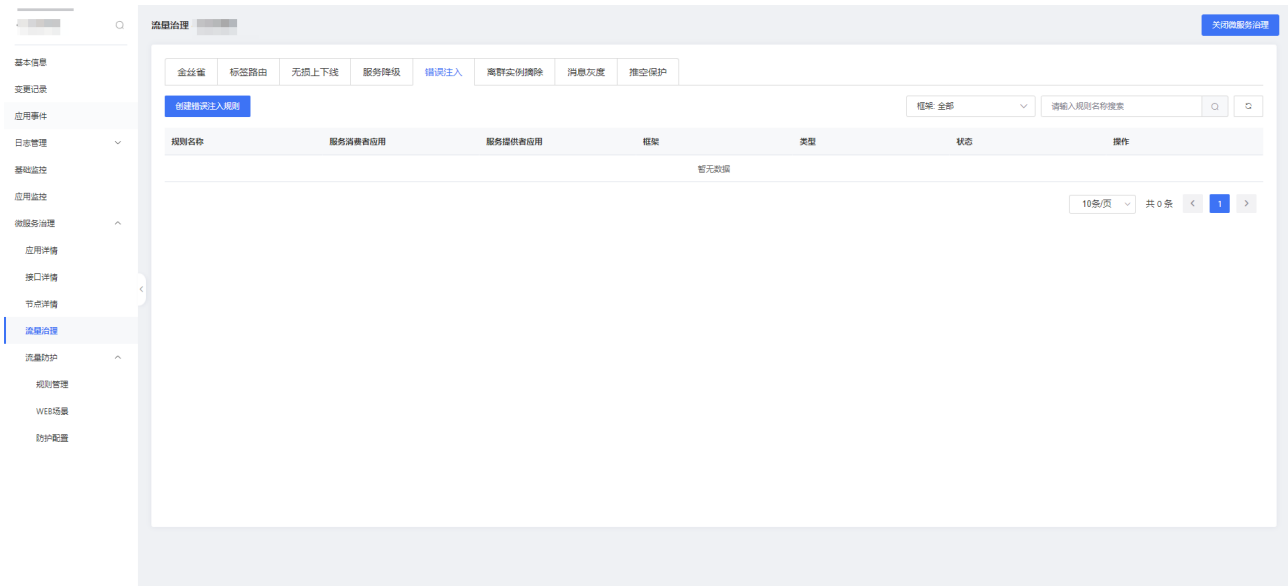
- 已部署应用。
- 已开通 [MSE 微服务治理](#)。

说明

说明 使用 MSE 时会产生单独费用。MSE 的计费说明，请参见 [微服务治理计费概述](#)。

功能入口

1. 登录 CAE控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
2. 在左侧导航栏，选择微服务治理 > 流量治理，单击错误注入。



错误注入规则参数说明：

参数	说明
规则名称	错误注入规则的名称
框架	支持SpringCloud和Dubbo框架
服务消费者应用	服务消费者对应的应用名称
服务提供者应用	服务提供者对应的应用名称
服务	被调用服务标识SpringCloud框架：服务注册名称Dubbo框架：服务名称/分组/版本信息，格式为{DubboService}:{group}:{version}，如 com.example.DemoService:dev:v1
请求路径	SpringCloud框架被调用接口的路径
请求方法	SpringCloud框架Http请求方法支持GET/POST/PUT/DELETE选择ALL匹配所有方法
服务方法	Dubbo框架被调用的方法，格式为{MethodName}:{ParamType...}如 sayHello:java.lang.String,java.lang.String
类型	延迟：模拟调用延迟，可自定义延迟时间错误：SpringCloud框架下模拟调用返回的Http状态码和响应体；Dubbo框架下模拟调用触发的异常，可输入异常类全路径和异常信息
触发概率	触发此规则相关模拟调用的概率

离群实例摘除

概述

在微服务场景中，当服务提供者的实例出现异常时，服务消费者无法感知到提供者出现异常，此时就可能出现异常调用。通过配置离群摘除功能可以实时监测下游实例的可用性，摘除异常实例，提升业务的可用性。

版本限制

框架	限制	详情
Spring Cloud	Spring Cloud Dalston及以上版本	--
Dubbo	2.x	支持 Apache Dubbo 不支持 Alibaba Dubbo
Jdk版本	1.8+	--

开通离群摘除

1. 登录 CAE控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
2. 在左侧导航栏，选择 微服务治理 > 流量治理，单击离群实例摘除。

离群摘除规则参数说明：

参数	说明
策略名称	离群摘除规则的名称。
被调用服务所用框架	Spring Cloud 或 Dubbo。
错误率下限	被调用的应用中某个应用实例的错误率高于设置的域值后，将摘除该实例。默认值为50%。例如该实例在统计时间窗口内被调用10次，有6次调用失败，错误率为60%，超过了配置的错误率域值（50%），则从应用中移除该实例。
异常类型	目前只支持网络异常+业务异常（HTTP 5xx）。
QPS下限	QPS按照统计时间窗口进行计算，默认为10秒。
摘除实例比例上限	摘除的异常实例比例上限，即达到阈值后，不再摘除异常实例。
恢复检测单位时间	摘除的异常实例比例上限，即达到阈值后，不再摘除异常实例。
未恢复累计次数上限	持续对异常实例进行检测，检测间隔随检测次数按恢复检测单位时间线性增加，当达到设置的检测次数上限后，会按最长时间间隔持续检测异常实例是否恢复。
默认状态	默认是否开启离群摘除规则。

消息灰度

前提条件

- 已部署应用。
- 已开通 [MSE 微服务治理](#)。

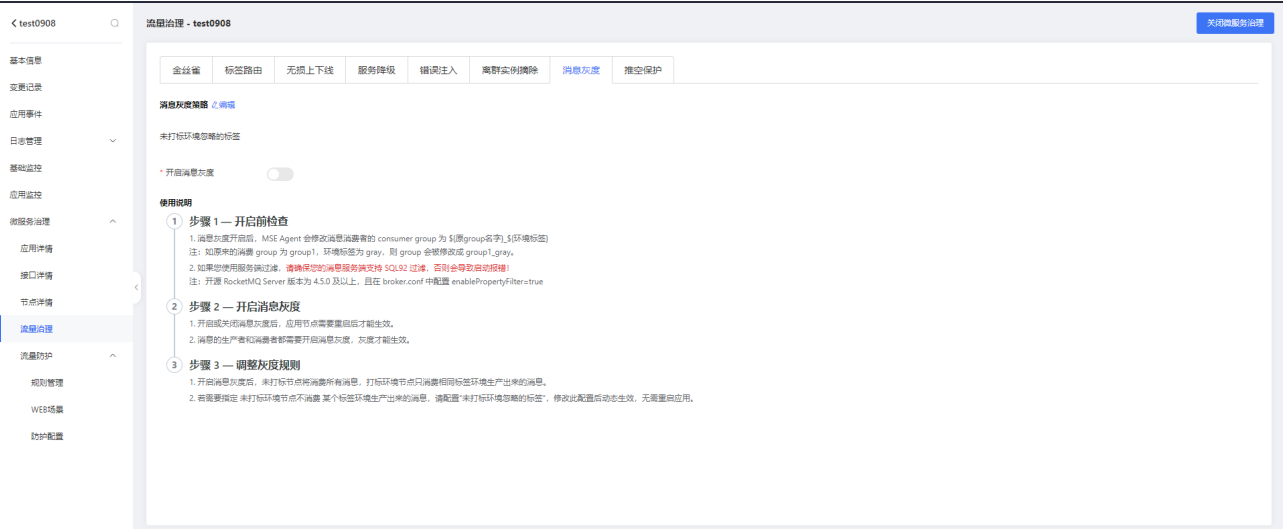
说明

说明 使用 MSE 时会产生单独费用。MSE 的计费说明，请参见 [微服务治理计费概述](#)。

功能入口

1. 登录 CAE控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。

2. 在左侧导航栏，选择微服务治理 > 流量治理，单击服务降级。



消息灰度参数说明：

参数	说明
未打标环境忽略标签	未打标环境默认会消费所有环境的消息。若配置“未打标环境忽略标签”，则未打标环境会忽略配置的标签。
开启消息灰度开关	消息灰度生效开关。
消息灰度过滤侧	客户端过滤：在消费端MSE Agent过滤，会拉取所有的消息，建议提前评估消息量。服务端过滤：在服务端通过SQL92方式过滤，需要RocketMQ服务端支持。

推空保护

前提条件

- 已部署应用。
- 已开通 [MSE 微服务治理](#)。

说明

说明 使用 MSE 时会产生单独费用。MSE 的计费说明，请参见 [微服务治理计费概述](#)。

功能入口

- 登录 CAE控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
- 在左侧导航栏，选择微服务治理 > 流量治理，单击服务降级。

开启推空保护

- 在推空保护页面打开开关。



查看推空保护事件

若发生推空保护，在推空保护页面即可查看推空保护事件。



流量防护

配置流控规则

配置流控规则的原理是监控应用或服务流量的QPS指标，当指标达到设定的阈值时立即拦截流量，避免应用被瞬时的流量高峰冲垮，从而保障应用高可用性。

功能入口

1. 登录 CAE 控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
2. 在左侧导航栏，选择微服务治理 > 流量治理 > 流量防护 > 流控规则。
3. 在流控规则页签单击新增流控规则，然后在弹出的对话框中配置规则信息，参数说明如下。
4. 设置完成单击新增。

页面参数说明如下：

参数	描述
接口名称	待流控的资源名称。
来源应用	该规则针对的来源应用，默认来源应用设为default，代表不区分来源应用。
统计维度	选择资源调用关系进行流控。 • 当前接口：直接控制来自来源应用中调用来源的访问流量，如果来源应用为default则不区分调用来源。通常应用于流量匀速通过的场景。 • 关联接口：控制当前资源的关联资源的流量。通常应用于资源争抢时，留足资源给优先级高接口的场景。 • 链路入口：控制该资源所在的调用链路的入口流量。选择链路入口后需要继续配置 入口资源，即该调用链路入口的上下文名称。通常应用于预热启动避免大流量冲击的场景。

参数	描述
单机QPS阈值	触发对流控接口的统计维度对象的QPS阈值。
流控效果	选择流控方式来处理被拦截的流量。 - 快速失败：达到阈值时，立即拦截请求。按照应用系统设置中的适配模块配置信息，进行内容返回。 - 预热启动：需设置具体的预热时间。如果系统在此之前长期处于空闲的状态，当流量突然增大的时候，该方式会让处理请求的速率缓慢增加，经过设置的预热时间以后，到达系统处理请求速率的设定值。默认会从设置的QPS阈值的1/3开始慢慢往上增加至设置的QPS值，多余请求会按照快速失败处理。 - 排队等待：请求匀速通过，允许排队等待，通常用于请求调用削峰填谷等场景。需设置具体的超时时间，达到超时时间后请求会快速失败。
是否开启	打开开关表示启用该规则，关闭开关表示禁用该规则。开关修改之后会立即生效。

使用场景

常用场景 1：当资源争抢时，需留足资源给优先级高的接口

为了确保提交的数据不丢失，我们将数据库读写分离，并给写库分配了更高的优先级。对于读取数据的请求，如果过于频繁，会对写入操作进行限流。

在新建流控防护规则对话框中配置以下规则信息：

1. 统计维度选择关联接口。
2. 流控效果选择快速失败。
3. 关联接口阈值为10。

当写库操作的QPS超过10之后，读库操作会被限流以保证留足资源给写库操作，避免写库操作数据丢失。

常用场景 2：预热启动避免大流量冲击

采用流量控制的方法，在流量入口处进行控制，以缓慢增加的方式让流量通过，在一定时间内达到阈值上限，以便系统能够预热。这种方法最适合应对突发流量的场景。

在新建流控防护规则对话框中配置以下规则信息：

1. 统计维度选择链路入口。
2. 流控效果选择预热启动。
3. 单机QPS阈值为90。
4. 预热时间为10s。

预热流控方式下，默认会从设置的QPS阈值的1/3开始慢慢往上增加至QPS设置值。本示例中，当入口的QPS超过30（即 $90 \div 3$ ）时，会在预热的10s内缓慢增长至90。

常用场景 3：削峰填谷，使流量匀速通过

请求流量具有波峰波谷的特点，流量管理的原理是将前期的高峰流量延迟到后期再处理，以最大化满足所有请求，并保证用户体验。

在新建流控防护规则对话框中配置以下规则信息：

1. 统计维度选择当前接口。

2. 流控效果选择排队等待。
3. 配置匀速模式下请求单机QPS阈值为5。
4. 等待时长为5s。

配置隔离规则

隔离规则，通过控制接口或调整依赖的并发线程数，以保证系统的稳定性。这种方法适用于应用程序内部或下游依赖出现不稳定的情况，例如慢SQL或下游应用响应时间变慢等。

功能入口

1. 登录 CAE 控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
2. 在左侧导航栏，选择微服务治理 > 流量治理 > 流量防护 > 隔离规则。
3. 在隔离规则页签单击新增隔离规则，然后在弹出的对话框中配置规则信息。
4. 设置完成单击新增。

使用场景

常用场景 1：保障自身资源充足

当运行该请求的响应时间变长，会导致线程的并发数变大。当并发数超过阈值以后，微服务治理将拒绝多余的请求，直到堆积的任务完成，并发线程数变少。达到将异常隔离，减小不稳定性的效果。例如某个SQL执行时间为20毫秒，预期该请求每秒有50个。

在新增隔离防护规则对话框中配置以下规则信息：

1. 填写接口名称和来源应用。
2. 统计维度选择当前接口。
3. 并发数阈值为50。

常用场景 2：有一定相关性的接口

当关联接口被来源应用调用QPS超过阈值时，会对当前接口来源应用的请求进行限流，有一定的相关性的方法来配置规则。例如数据库读操作和数据库写操作这两个资源分别代表数据库读写，数据库写操作接口优先级更高。

为保证读写资源争抢时，数据库写操作的接口可以留足资源，可在新增隔离防护规则对话框中配置以下规则信息：

1. 接口名称为数据库读操作。
2. 统计维度选择关联接口。
3. 关联接口名为数据库写操作。
4. 并发数阈值为10。

常用场景 3：针对入口链路来配置隔离规则

将入口处的资源进行分离，以确保更高优先级的入口。当调用callstack入口的请求数量超过阈值时，会对来自于来源应用的请求进行隔离流控。这样可以确保重要的请求能够得到及时处理，并且不会影响其他请求的处理。

在新增隔离防护规则对话框中配置以下规则信息：

1. 填写接口名称和来源应用。
2. 统计维度选择链路入口。
3. 并发数阈值设置为10。

配置熔断规则

熔断规则可以监控应用程序内部或下游依赖的响应时间或异常情况，并在达到指定的阈值时立即降低下游依赖的优先级。在指定的时间内，系统将不会调用该不稳定的资源，以确保应用程序的高可用性。当指定时间过后，再重新恢复对该资源的调用。这样可以保证应用程序的稳定性和可靠性。

功能入口

1. 登录 CAE 控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
2. 在左侧导航栏，选择微服务治理 > 流量治理 > 流量防护 > 熔断规则。
3. 在熔断规则页签单击新增熔断规则，然后在弹出的对话框中配置规则信息。
4. 设置完成单击新增。

使用场景

常用场景 1：慢调用熔断示例

为了保证当前接口的性能，如果出现第三方服务的响应时间过慢，则会对其进行熔断操作。

常用场景 2：异常熔断示例

为了保证更好的用户体验，如果第三方内容展示出现异常，且异常比例较高，则会对其进行熔断操作。

常用场景 3：熔断个数

如第三方异常个数达到指定数量，则会对其进行熔断操作。

配置主动降级规则

主动降级规则可以指定哪些接口需要进行降级，以及在何种情况下需要进行降级。被降级的接口会自动触发自定义的降级行为，以确保系统的稳定性和可靠性。

功能入口

1. 登录 CAE 控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
2. 在左侧导航栏，选择微服务治理 > 流量治理 > 流量防护 > 主动降级规则。
3. 在主动降级规则页签单击新增主动降级规则，然后在弹出的对话框中配置规则信息。
4. 设置完成单击新增。

使用场景

当某个资源发生异常，可以指定对接口进行降级，而不会执行原来的逻辑。

配置热点规则

为应用程序配置热点规则后，微服务治理中心会对系统中的资源调用次数进行分析，并根据配置的热点规则来限制包含热点参数的资源调用，以保证系统的稳定性。

功能入口

1. 登录 CAE 控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
2. 在左侧导航栏，选择微服务治理 > 流量治理 > 流量防护 > 热点规则。
3. 在热点规则页签单击新增热点规则，然后在弹出的对话框中配置规则信息。
4. 设置完成单击新增。

使用场景

常用场景 1：秒杀场景

为了保证系统稳定性，可以配置热点规则，当超过一定量的阈值后，系统会让满足热点规则的请求流量排队等待。例如，对于购买同一商品的请求，如果在1秒内调用次数超过100次，则其他请求将被等待处理。在新建热点规则对话框中，可以配置以下规则信息：

- 阈值：设置超过多少次请求会被限流。
- 等待时间：设置等待时间，单位为秒。
- 调用次数：设置调用次数的阈值。

例如购买同一商品，1s内调用超过100次请求后，则其余请求进行等待。在新建热点规则对话框中配置以下规则信息：

- 填写接口名称。
- 统计维度选择通过请求数。
- 统计周期时间设置为1s，单机阈值设置为100。
- 流控效果选择排队等待。
- 超时时间设置为30 ms。

常用场景 2：调用请求频繁，占用较多系统资源

例如当调用修改请求较多的时候，会占用了写数据库较多资源，则可以对其进行热点快速失败的处理，稍后再修改。在新建热点规则对话框中配置以下规则信息：

- 填写接口名称。
- 统计维度选择并发数。
- 统计周期时间设置为1s，单机阈值设置为100。
- 流控效果选择快速失败。

配置自适应流控

系统支持自动化流控或手动设置流控规则，以确保系统的资源分配和负载平衡。自动化流控是根据系统的CPU使用率动态调整应用程序的入口流量，以达到最大吞吐量状态下的系统稳定运行；手动设置流控规则则是从整体角度出发，人工设置规则来控制应用程序的入口流量，以保证系统的负载平衡。

功能入口

1. 登录 CAE 控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
2. 在左侧导航栏，选择微服务治理 > 流量治理 > 流量防护 > 自适应流控。
3. 在自适应流控页签单击新增系统保护规则，然后在弹出的对话框中配置规则信息。
4. 设置完成单击新增。

参数说明：

参数	描述	使用场景说明
CPU使用率	当系统CPU使用率超过阈值即触发系统保护，阈值设置范围为0.0~1.0（代表0%~100%）。	适用于设置基础资源水位的场景，比如需要保证一定的冗余水位。但系统水位不宜过高，需要留部分水位。
Load	当系统的Load1超过阈值，且系统当前的并发线程数超过系统容量时才会触发系统保护。系统容量由系统的 $\max Qps * \min Rt$ 计算得出。	适用于设置基础资源水位的场景，比如需要保证一定的冗余水位。但系统水位不宜过高，需要留部分水位。
线程数	当单台机器上所有入口流量的并发线程数达到阈值即触发系统保护。	适用于设置基础资源水位的场景，比如需要保证一定的冗余水位。但系统水位不宜过高，需要留部分水位。

参数	描述	使用场景说明
入口平均RT	当单台机器上所有入口流量的平均RT达到阈值即触发系统保护，单位是毫秒。	适用于衡量入口请求的场景。
入口总QPS	当单台机器上所有入口流量的QPS达到阈值即触发系统保护。	适用于衡量入口请求的场景。

WEB 场景防护

Web 场景防护功能是一种用于保护提供Web服务的应用程序的防护机制，可以对访问请求中的参数项进行精细化的流量控制。该功能可以用于处理使用主流 Web 框架（如Servlet容器、Spring Web、SpringBoot）的应用程序，并配置了API粒度的请求参数解析，以便对请求中的IP地址、主机名、头部信息和 URL 参数等参数维度进行流量控制。通过这种控制，可以保护应用程序和系统的稳定性。

背景信息

在提供 Web 服务的场景中，除了 API 维度的限流降级防护，对于来源 IP 和访问参数等资源调用的限流防护可以更好地保证业务应用的正常运行。在一些大流量的 Web 业务场景下，可能需要对多个接口进行限制，以保证系统的稳定性。通过对当前访问频次最高的来源 IP 或访问频次最高的商品序号进行有针对性的限制，可以有效地保护系统。比如：

- 对于一段时间内最频繁购买的商品序号，可以进行限制以防止击穿缓存导致大量请求到数据库。
- 对于一段时间内频繁大量访问的来源IP，可以进行限制以防止利用虚假信息恶意刷单。这样可以保证系统的稳定性和可靠性。

功能入口

1. 登录 CAE 控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
2. 在左侧导航栏，选择微服务治理 > 流量治理 > WEB 场景。
3. 在 WEB 场景页签单击新增 Web 防护规则，然后在弹出的对话框中配置规则信息。
4. 设置完成单击新增。

参数说明

参数	说明
接口名称	适用该规则的应用资源。
参数属性	针对所选接口的属性进行流量控制： • Client IP：请求端的 IP 地址。 • Remote Host：请求端的 Host。 • Header：HTTP 请求携带的 Header。 • URL 参数：HTTP 请求携带的 URL 参数。

条件模式	针对参数属性的匹配模式： - 单条件：针对单个参数属性进行热点流控或属性值匹配流控。当参数属性为 Client IP 时，针对 Client IP 的值进行热点限流；当参数属性为 Remote Host 时，针对 Remote Host 的值进行热点限流；当参数属性为 Header 时，输入 Header 名称，若不打开属性值匹配，则对 Header 的值进行热点限流，若打开属性值匹配，则根据 Header 值是否满足匹配串进行限流；当参数属性为 URL 参数时，输入 URL 参数名称，若不打开属性值匹配，则对 URL 参数的值进行热点限流，若打开属性值匹配，则根据 URL 参数值是否满足匹配串进行限流。 - 多条件：针对多个参数属性进行属性值匹配限流。
阈值	QPS 值，支持秒、分、小时和天四种维度。如阈值填写 10，统计时间选择秒，则表示每秒请求数大于 10 时会触发限流。
流控方式	- 快速失败：多余的请求会被立即拒绝。允许设置额外的 burstsize，即针对突发请求额外允许的请求数目。 - 排队等待：请求匀速通过，允许排队等待，可设置超时时间，即最长的排队等待时间，单位为 ms，超出此时间将直接被拒绝。

配置防护行为

在防护行为分析中，可以对 Web 或者 RPC 类型的埋点资源进行监控，以便在满足某种规则时触发自定义处理行为。例如，如果某个 Web 接口触发了流控规则，则可以返回“Blocked by Sentinel”的提示文本。

新增行为

1. 登录 CAE 控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
2. 在左侧导航栏，选择微服务治理 > 流量治理 > 防护配置。
3. 在防护配置页签单击新增行为，然后在弹出的对话框中配置行为信息。
4. 单机确定完成新增。

修改或删除行为

1. 登录 CAE 控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
2. 在左侧导航栏，选择微服务治理 > 流量治理 > 防护配置。
3. 行为列表页，您可以查看各个行为的具体描述，修改或删除行为。

关联行为

1. 登录 CAE 控制台，在左侧导航栏选择应用管理 > 应用列表，然后单击目标应用名称。
2. 在左侧导航栏，选择微服务治理 > 流量治理 > 流量防护 > 流控规则。
3. 在流控规则页签单击新增流控规则，然后在弹出的对话框中配置规则信息。
4. 完成选择防护场景和配置防护规则后，在配置限流行为区域，完成下列设置：

选择接口类型为 Web，在关联行为的下拉列表中选择目标行为进行关联，或单击新增行为来创建新的行为进行关联。

示例应用

JAVA-hellocae

应用介绍

一个简单的 Spring Boot Web 应用，提供多个 REST 接口用于演示 HTTP 请求/响应的基本概念，适合新手入门学习。启动类为 com.example.hello.HelloApplication，接口统一以 /api 为前缀。

前置条件

- 开通云应用引擎控制台

应用使用方式

选择示例镜像

1. 点击选择镜像-示例镜像- java-hellocae 镜像
2. 选择镜像版本 v1.0

应用访问

应用默认端口

8080

应用API文档

请求方式	接口API	请求参数	示例参数	含义
GET	/api/hello	无	无	返回 "Hello, World!", 验证服务是否正常运行
GET	/api/ip	无	无	返回应用所在服务器的 IP 地址
GET	/api/sleep	seconds (可选, 默认 3)	seconds=5	模拟延迟响应, 返回睡眠时长和起止时间
GET	/api/echo	任意 query 参数	name=test&age=18	回显请求的 method、URI、URL、请求头、参数等信息

请求方式	接口API	请求参数	示例参数	含义
POST	/api/echo	请求体 (JSON)	{"key": "value"}	回显 POST 请求的 Content-Type、Content-Length 和 请求体
GET	/api/exception	code (必填)	code=404	模拟抛出指定 HTTP 状态码的异常

应用访问示例

```
# #####
curl http://localhost:8080/api/hello

# ##### IP
curl http://localhost:8080/api/ip

# #####5##
curl "http://localhost:8080/api/sleep?seconds=5"

# GET #####
curl "http://localhost:8080/api/echo?name=test&age=18"

# POST #####
curl -X POST http://localhost:8080/api/echo \
  -H "Content-Type: application/json" \
  -d '{"key": "value"}'

# #####
curl "http://localhost:8080/api/exception?code=404"
```

golang-hellocae

应用介绍

本项目是一个基于 Gin 框架的 Golang 微服务 Demo 应用，用于演示 CAE（Cloud Application Engine）平台的灰度发布、链路追踪、健康检查、监控指标采集等能力。

前置条件

- 开通云应用引擎控制台

应用使用方式

选择示例镜像

1. 点击选择镜像-示例镜像- golang-hellocae 镜像
2. 选择镜像版本 v1.0

应用访问

应用默认端口

8080

应用API文档

请求方式	接口API	请求参数	示例请求	含义
GET	/	无	curl http://localhost:8080/	根路径，返回欢迎字符串 hello cae!
GET	/ping	无	curl http://localhost:8080/ping	健康检查接口，返回 pong!
GET	/ip	无	curl http://localhost:8080/ip	获取本机 IP、主机名和当前时间戳，返回 JSON
GET	/user:id	路径参数: id (用户ID)	curl http://localhost:8080/user:12345	根据用户ID查询用户信息
GET	/echo	查询参数: q (回显内容)	curl http://localhost:8080/echo?q=hello	回显接口，将 q 参数内容原样返回
GET	/error	无	curl http://localhost:8080/error	模拟错误响应，返回 502 Bad Gateway
GET	/slow-ping	无	curl http://localhost:8080/slow-ping	模拟慢请求，睡眠3秒后返回 `pong!` 及本机IP

JAVA-provider

应用介绍

Nacos 服务提供者示例，注册到 Nacos 注册中心（服务名 cae-provider），被 Consumer 项目通过 Feign 远程调用。用于演示微服务架构中的服务注册与发现。启动类为 org.example.NacosProviderApplication。

前置条件

- 开通云应用引擎控制台

应用使用方式

选择示例镜像

1. 点击选择镜像-示例镜像- java-provider 镜像
2. 选择镜像版本 v1.0

应用访问

应用默认端口

26170

应用API文档

请求方式	接口API	请求参数	示例参数	含义
GET	/callB	无	无	返回提供者的标签和IP，配合 Consumer 进行全链路灰度测试
GET	/degradeB	无	无	配合 Consumer 测试服务降级功能
GET	/outlierB	id (必填)	id=123	配合 Consumer 测试离群实例摘除；`id=999999` 时抛异常
GET	/flowRuleB	无	无	配合 Consumer 测试流量控制规则

应用访问示例

```
# #####
curl http://localhost:26170/callB

# #####
curl http://localhost:26170/degradeB

# #####
curl "http://localhost:26170/outlierB?id=123"

# #####
curl "http://localhost:26170/outlierB?id=999999"

# #####
curl http://localhost:26170/flowRuleB
```

JAVA-consumer

应用介绍

Nacos 服务消费者示例，通过 OpenFeign 远程调用 Provider (cae-provider) 的服务。用于演示微服务中的服务间调用、负载均衡、熔断降级、灰度发布等能力。启动类为 org.example.NacosConsumerApplication。

前置条件

- 开通云应用引擎控制台
- 同命名空间，已部署 JAVA provider 应用

应用使用方式

选择示例镜像

1. 点击选择镜像-示例镜像- java-consumer 镜像
2. 选择镜像版本 v1.0

应用访问

应用默认端口

26160

应用API文档

请求方式	接口API	请求参数	示例参数	含义
GET	/callA	无	无	调用 Provider 的 /callB, 返回 A # B 链路信息, 验证灰度发布
GET	/degradeA	无	无	调用 Provider 的 /degradeB, 验证服务降级策略
GET	/outlierA	id (必填)	id=123	调用 Provider 的 /outlierB, 验证离群实例摘除机制
GET	/flowRuleA	无	无	返回当前实例标签和 IP, 验证流量控制规则
POST	/postA	请求体 (JSON)	{"name": "test", "value": "hello"}	通过 Feign POST 调用 Provider 的 postB

应用访问示例

```
# #####
curl http://localhost:26160/callA

# #####
curl http://localhost:26160/degradeA

# #####
curl "http://localhost:26160/outlierA?id=123"

# #####
curl http://localhost:26160/flowRuleA

# POST #####
curl -X POST http://localhost:26160/postA \
-H "Content-Type: application/json" \
-d '{"name": "test", "value": "hello"}
```

DotNet-hellocae

应用介绍

本项目是一个极简 Demo 应用，用于演示 CAE（Cloud Application Engine）平台的 .net 部署能力。

前置条件

- 开通云应用引擎控制台

应用使用方式

选择示例镜像

1. 点击选择镜像-示例镜像- dotnet-hellocae 镜像
2. 选择镜像版本 v1.0

应用访问

应用默认端口

5000

应用API文档

请求方式	接口API	请求参数	示例请求	含义
GET	/	无	curl http://localhost:5000/	根路径，返回欢迎字符串 hello cae!

NodeJS-hellocae

应用介绍

本项目是一个极简 Demo 应用，用于演示 CAE（Cloud Application Engine）平台的 Node.js 部署能力。

前置条件

- 开通云应用引擎控制台

应用使用方式

选择示例镜像

1. 点击选择镜像-示例镜像- nodejs-hellocae 镜像
2. 选择镜像版本 v1.0

应用访问

应用默认端口

9000

应用API文档

请求方式	接口API	请求参数	示例请求	含义
GET	/	无	curl http://localhost:9000/	根路径，返回 HTML 祝贺页面

python-hellocae

应用介绍

本项目是一个基于 Flask 框架的 python web 服务 Demo 应用，用于演示 CAE（Cloud Application Engine）平台的 python web 程序部署能力。

前置条件

- 开通云应用引擎控制台

应用使用方式

选择示例镜像

1. 点击选择镜像-示例镜像- python-hellocae 镜像
2. 选择镜像版本 v1.0

应用访问

应用默认端口

5000

应用API文档

请求方式	接口API	请求参数	示例请求	含义
GET	/	无	curl http://localhost:5000/	根路径，返回欢迎字符串 hello cae!
GET	/hello	无	curl http://localhost:5000/hello	健康检查接口，返回 OK
POST	/post_hello	无	curl -X POST http://localhost:5000/post_hello	获取本机 IP、主机名和当前时间戳，返回 JSON

php-hellocae

应用介绍

本项目是一个 php web 服务 Demo 应用，用于演示 CAE（Cloud Application Engine）平台的 php web 程序部署能力。

前置条件

- 开通云应用引擎控制台

应用使用方式

选择示例镜像

1. 点击选择镜像-示例镜像- php-hellocae 镜像
2. 选择镜像版本 v1.0

应用访问

应用默认端口

80

应用API文档

请求方式	接口API	请求参数	示例请求	含义
GET	/	无	curl http://localhost/	根路径，返回欢迎字符串 Hello cae! cae handle get req.
POST	/hello	name	curl -X POST http://localhost/handler.php -d "name=Alice"	返回欢迎字符串 hello, Alice! cae handle post req.

常见问题

应用托管FAQ

如何处理常见的Pod问题？

部署在云应用引擎 CAE（Cloud App Engine）上的应用运行时，可能会出现 Pod 问题。本文介绍常见的 Pod 问题和解决方法。

ImagePullBackOff

当 CAE 无法获取到 Pod 中某个容器的镜像时，将出现此错误。

可能原因：

- 镜像名称无效，例如镜像名称拼写错误、镜像不存在。
- 镜像标签无效，例如标签拼写错误、标签不存在。
- 镜像属于私有仓库。

解决方案：

- 更正镜像名称与标签。
- 将镜像上传至天翼云镜像仓库。

CrashLoopBackOff

如果容器无法启动，出现此错误。

可能原因：

- 应用程序中存在错误，导致无法启动。
- 未正确配置容器。
- Liveness探针失败太多次。

解决方案：

您可以通过查看实时日志和事件分析失败原因，并修改相关配置。

处于未就绪状态的Pod

如果Pod正在运行但未就绪（not ready），则表示 Readiness 就绪探针失败。

可能原因：

当“就绪”探针失败时，Pod未连接到服务，并且没有流量转发到该实例。

解决方案：

登录Webshell，执行您自定义的就绪探针命令，查看 Pod 是否正常运行。如果不正常，您可以通过实时日志或事件对其进行分析。

CAE是否支持定时部署？

不支持。您可以通过调用部署接口 API，自行在代码中实现。

为什么创建应用后在控制台找不到应用？

本文介绍使用 CAE 时，如何排查在控制台找不到应用的情况。

问题现象

应用创建成功后，在 CAE 控制台的概览页可以查询到对应的应用数量，但是无法在具体的应用列表找到该应用。

可能原因

查询的地域、命名空间的范围有误。

解决方案

1. 登录 CAE 控制台。
2. 确认地域是否正确。在顶部菜单栏，选择地域。
3. 确认命名空间是否正确。在应用列表页面，选择命名空间，然后根据已知的应用信息，例如应用名称维度，搜索目标应用。

如何修改应用部署方式，例如将JAR包部署改为镜像部署？

应用创建成功后，不支持修改部署方式。您需要重新创建应用。

如何在容器中安装一个ffmpeg命令？

您可以获取一个 ffmpeg 镜像，并部署到 CAE。

微服务相关FAQ

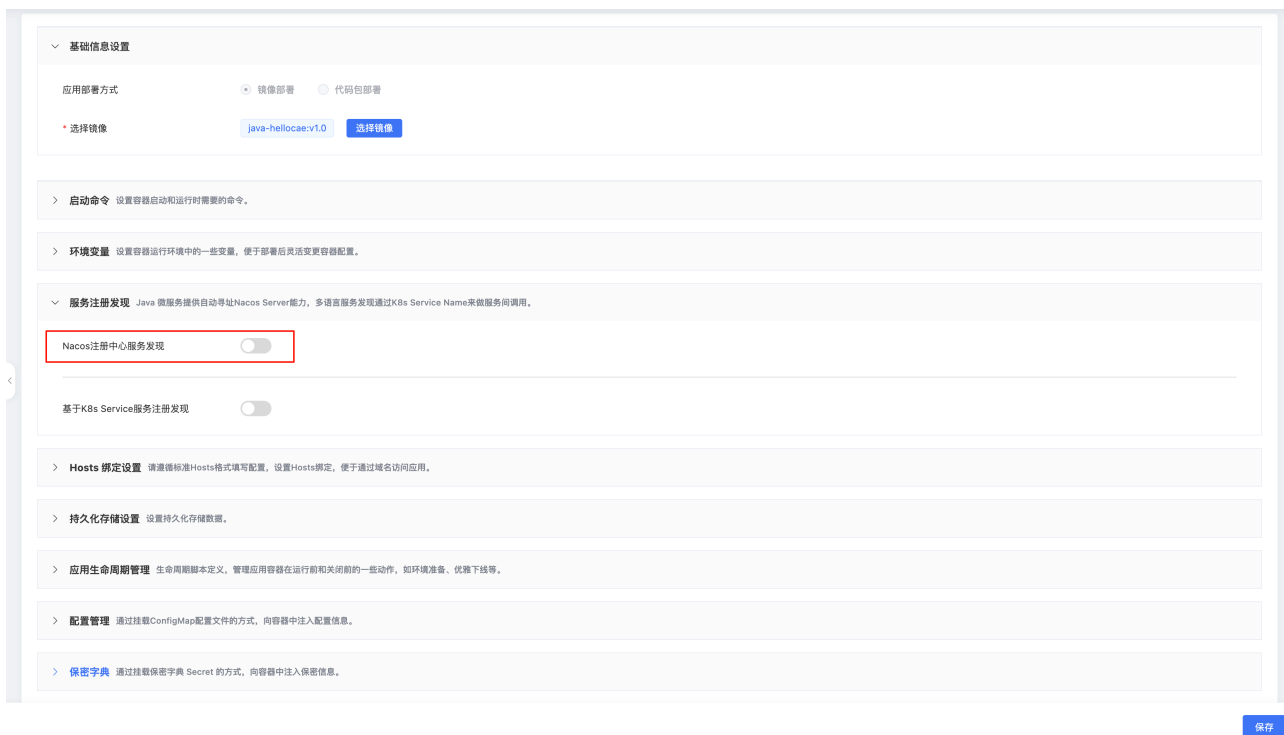
CAE应用是否必须选择服务发现？

CAE 应用不一定需要服务注册发现，您可以在创建应用或部署应用时，选择基于K8s Service服务注册发现，或者保持开关为关闭状态。

如何配置使用云应用引擎内置注册中心？

在创建和部署微服务应用时，您只需要选择云应用引擎的内置 Nacos，无需进行多余配置即可使用此功能。

您只需在部署应用时选择启用Nacos注册中心服务发现，云应用引擎会通过字节码注入的方式修改您程序的注册中心地址。



基础信息设置

应用部署方式 ☒ 镜像部署 ☐ 代码包部署

选择镜像 java-hellocae:v1.0 选择镜像

启动命令 设置容器启动和运行时需要的命令。

环境变量 设置容器运行环境中的一些变量，便于部署后灵活变更容器配置。

服务注册发现 Java 微服务提供自动寻址Nacos Server能力，多语言服务发现通过K8s Service Name来做服务间调用。

Nacos注册中心服务发现 ☒

基于K8s Service服务注册发现 ☐

Hosts 绑定设置 请遵循标准Hosts格式填写配置，设置Hosts绑定，便于通过域名访问应用。

持久化存储设置 设置持久化存储数据。

应用生命周期管理 生命周期脚本定义，管理应用在运行前和关闭前的一些动作，如环境准备、优雅下线等。

配置管理 通过挂载ConfigMap配置文件的方式，向容器中注入配置信息。

保密字典 通过挂载保密字典 Secret 的方式，向容器中注入保密信息。

保存

一个应用可以部署多个微服务吗？

不推荐。建议一个应用部署一个微服务，您可以完整地管理微服务的生命周期，包括部署、回滚等。如果您在一个容器中运行了多个进程，将无法对其进行隔离管理。

为什么微服务上下线的时候流量有损？

通常是客户端没有及时刷新服务列表的缓存，导致请求到下线中的服务器时出现流量有损。

- 使用 Spring Cloud、Dubbo 框架的用户

请确认是否开启 CAE 的微服务无损上下线。如果没有开启，请开启设置微服务无损下线和设置微服务无损上线。

- 使用其他微服务框架的用户

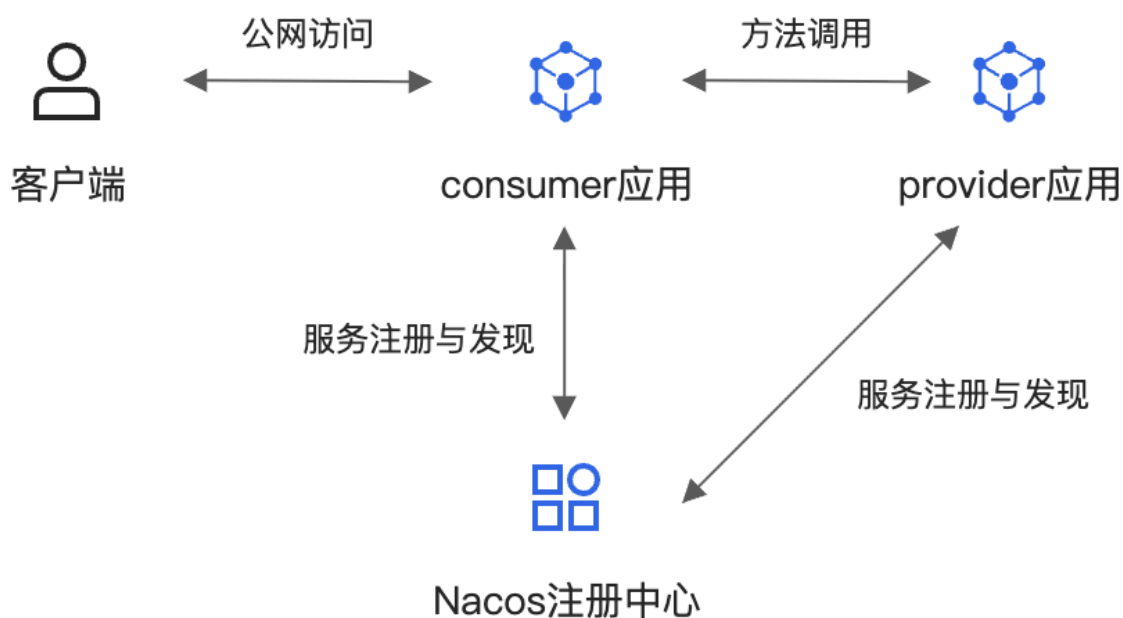
可以自行实现优雅下线接口，在该接口内实现从注册中心摘除并且等待一段时间的逻辑（比如10~30秒）。

最佳实践

使用CAE实现微服务应用注册发现与服务调用

方案概览

您将部署一组微服务应用，实现服务的注册与发现、服务之间的方法调用，并且通过公网来访问应用。示例应用的架构如下：



为实现本方案，您将：

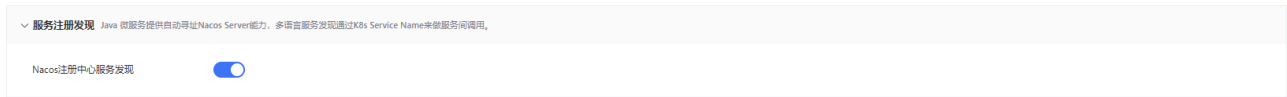
1. 部署应用：使用示例镜像部署 provider 应用与 consumer 应用，并通过 CAE 内置的 Nacos 实现服务注册与发现，使得 consumer 应用可以调用 provider 应用提供的服务。
2. 通过公网访问应用：为 consumer 应用配置公网访问地址并访问应用。

操作指引

部署应用

部署 provider 应用

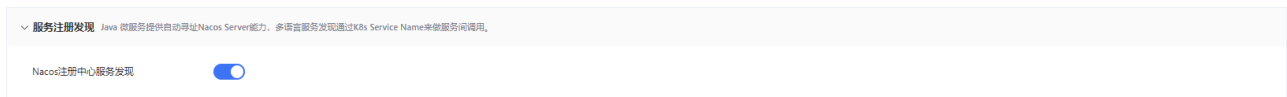
1. 登录云应用引擎控制台
2. 左侧导航栏选择应用管理 > 应用列表
3. 点击创建应用，配置以下参数，点击下一步进入高级设置
 - a. 自定义应用名称，例如 provider
 - b. 选择命名空间类型为系统创建，表示应用将会使用系统默认创建的命名空间，并关联 VPC、子网、安全组等网络资源
 - c. 设置应用部署方式为镜像部署。点击选择镜像，选择 Java 技术栈语言，在示例镜像标签页，找到java-provider镜像仓库，选择版本为v1.0，点击确定
 - d. 在容量设置区域，自定义单实例规格和实例数，这决定了应用初始运行的实例数量、系统为每个实例分配多少计算资源
4. 在高级设置页面中，展开服务注册发现区域，启用内置 Nacos 注册中心服务发现



5. 点击创建应用

部署 consumer 应用

1. 登录云应用引擎控制台
2. 左侧导航栏选择应用管理 > 应用列表
3. 点击创建应用，配置以下参数，点击下一步进入高级设置
 - a. 自定义应用名称，例如 consumer
 - b. 选择命名空间类型为系统创建，表示应用将会使用系统默认创建的命名空间，并关联 VPC、子网、安全组等网络资源
 - c. 设置应用部署方式为镜像部署。点击选择镜像，选择 Java 技术栈语言，在示例镜像标签页，找到 java-consumer 镜像仓库，选择版本为v1.0，点击确定
 - d. 在容量设置区域，自定义单实例规格和实例数，这决定了应用初始运行的实例数量、系统为每个实例分配多少计算资源
4. 在高级设置页面中，展开服务注册发现区域，启用内置 Nacos 注册中心服务发现



5. 点击创建应用

查看应用部署结果

请耐心等待应用创建完成，大约需要1分钟。

说明

- 您可以在应用列表中查看已创建的应用，点击其中某个应用可以进入其详情页面。
- 在左侧导航栏选择基础信息，在实例列表页签中，可以查看已创建的应用实例。

验证微服务应用调用能力

在 consumer 应用的基本信息页面，切换到实例列表，打开终端，输入访问命令：`curl http://localhost:26160/callA`

可以看到 consumer 应用调用 provider 应用的输出：

终端

[在新窗口打开](#) ✕

实例名称 ● Running 192.168.0.25 demo-consumer-57d9c46695-bpcqz

容器 main

demo-consumer-57d9c46695-bpcqz ● connected

```
root@demo-consumer-57d9c46695-bpcqz:/# curl http://localhost:26160/callA
a_cae-consumer_192.168.0.25 -> B_cae-provider_192.168.0.57
root@demo-consumer-57d9c46695-bpcqz:/#
```

配置公网访问应用

为了实现从公网访问 consumer 应用，您可以为应用绑定公网 ELB 实例。

进入 consumer 应用的详情页面，在应用信息页签中，在应用访问设置区域选择基于 ELB 访问。

点击添加公网 ELB 访问，在 HTTP 协议页签中，设置 HTTP 端口为 6677，容器端口为 26160，然后点击确定。

说明

说明 consumer 应用的端口是 26160，provider 应用的端口是 26170。这里是将公网 ELB 指定端口（6677）的流量转发到 consumer 应用的端口（26160）。

添加公网 ELB 访问



新建 ELB 实例（按量计费）

产品类型	产品配置	产品规格	数量	付费方式	购买周期	资费
负载均衡ELB	地域：华东1	性能保障型	1	后付费/按使用量付费	无	查看价格
网络类型：公网	公网带宽：按使用流量计费	标准型I elb.s2.small				

- HTTP 协议
- HTTPS 协议
- TCP 协议
- UDP 协议

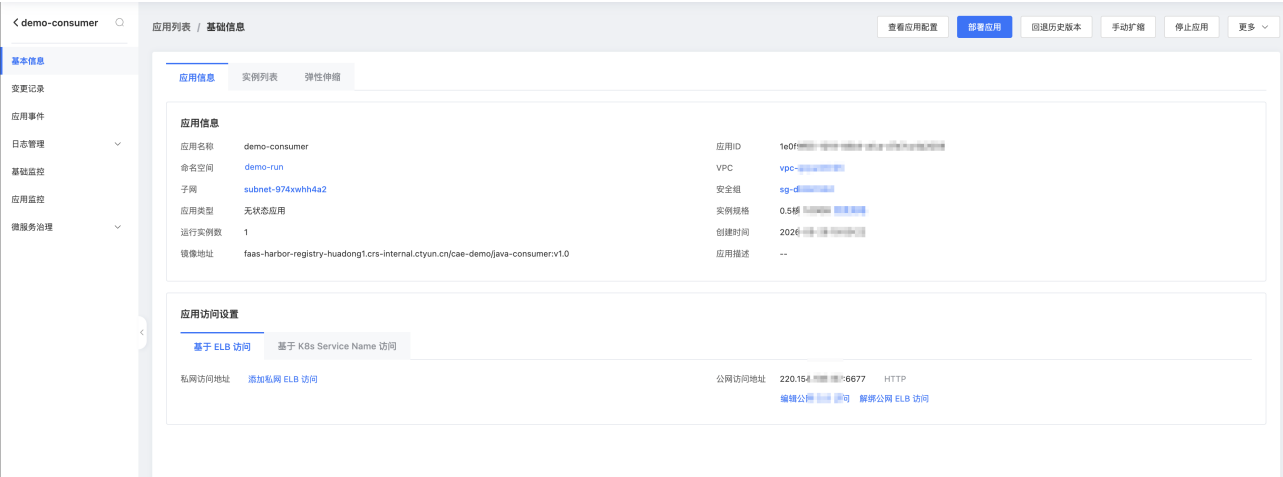
弹性 IP 80、443、8080 等端口默认为关闭状态，只有在备案完成后才会被放通。

网络协议	HTTP端口	容器端口	操作
HTTP	6677	26160	删除

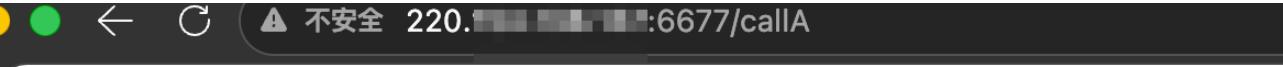
+添加

验证结果

当应用绑定公网 ELB 成功后，控制台将显示应用的公网访问地址。



此时可以通过浏览器访问 `http://<公网访问地址:端口>/callA`（其中<公网访问地址:端口>需要替换为实际值），验证能够通过公网来访问应用。



a_cae-consumer_192.168.0.25 -> B_cae-provider_192.168.0.23

资源清理

在完成本教程后，建议清理相关资源（删除 provider 和 consumer 应用、解绑ELB访问等），避免持续产生费用。

通过命名空间实现开发环境、测试环境和生产环境的隔离

如果您在部署应用时需要区分开发环境、测试环境和生产环境，您可以为每套环境创建一个命名空间。同一命名空间的应用之间可以通过内网实现方法调用，不同命名空间的应用可以实现相互隔离。

方案概览

本文将指引您创建dev、test、prod三个命名空间，分别用于开发、测试和生产环境。将应用部署到一个命名空间后，您可以将应用复制到其他命名空间并灵活更改应用的配置。您可以批量启停开发、测试或生产环境中的应用，以提升运维效率。

操作指引

1. 创建命名空间

1. 登录 CAE 控制台，在顶部菜单栏选择应用部署的地域，本文以华东1为例。
2. 同一地域中可以创建多个命名空间。我们首先创建dev命名空间。在左侧导航栏选择命名空间，单击创建命名空间。命名空间名称设置为dev。
3. 我们将为dev命名空间创建并关联一个VPC。单击创建 VPC，跳转到专有网络管理控制台。确认专有网络的地域为应用部署的地域，名称设置为dev-vpc。您可以根据需求配置VPC的网段，详见 [创建和管理专有网络](#)。

VPC 是应用所在的专有网络，一个命名空间只能关联一个 VPC，这意味着同一命名空间中的应用将处于同一 VPC 内部。

4. 应用必须部署在子网内。我们将在dev-vpc中创建2个子网，以提高应用的可用性。您可以自定义子网的名称、可用区、IPv4网段，详见 [创建和管理子网](#)。然后单击+添加来创建另一个子网。单击确定。

同一 VPC 内的不同子网之间内网互通。

5. 返回 CAE 控制台的创建命名空间面板，选择上一步中创建的 VPC 作为专有网络VPC。单击确定。

至此，您已成功创建dev命名空间。您可以重复上述步骤创建test、prod命名空间。

如果要实现不同命名空间中的应用相互隔离，您需要为每个命名空间分别创建并关联不同的 VPC。否则，不同命名空间中的应用可以通过 VPC 内网互通，可能造成相互干扰。

2. 部署应用

您可以先将应用部署到dev命名空间，再将应用复制到test、prod命名空间。

1. 在左侧导航栏选择应用管理 > 应用列表，命名空间选择dev，单击创建应用。将应用逐个部署到dev命名空间，为应用指定上一步中创建的 VPC 和子网。详细步骤请参见 [应用部署](#)。
2. 在应用列表中选择已创建的应用，在操作列单击复制。命名空间选择test或prod，即可将应用复制到相应的命名空间中。您还可以更改应用的配置以满足不同环境中的需求。

3. 批量启停应用

您可以批量启停开发、测试或生产环境中的应用。您可以手动批量启停应用。