云容器实例

产品介绍

文档版本 01

发布日期 2023-06-30

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。

目 录

1 什么是云容器实例 1
2 产品优势 4
3 应用场景 6
4 基本概念 10
5 安全 14
5.1 责任共担 14
5.2 身份认证与访问控制 15
5.3 数据保护技术 17
5.4 审计与日志 18
5.5 监控安全风险 18
6 约束与限制 20
7 计费说明 24
8 权限管理 25
9 区域和可用区 32
10 与其他服务的关系 34

1 什么是云容器实例

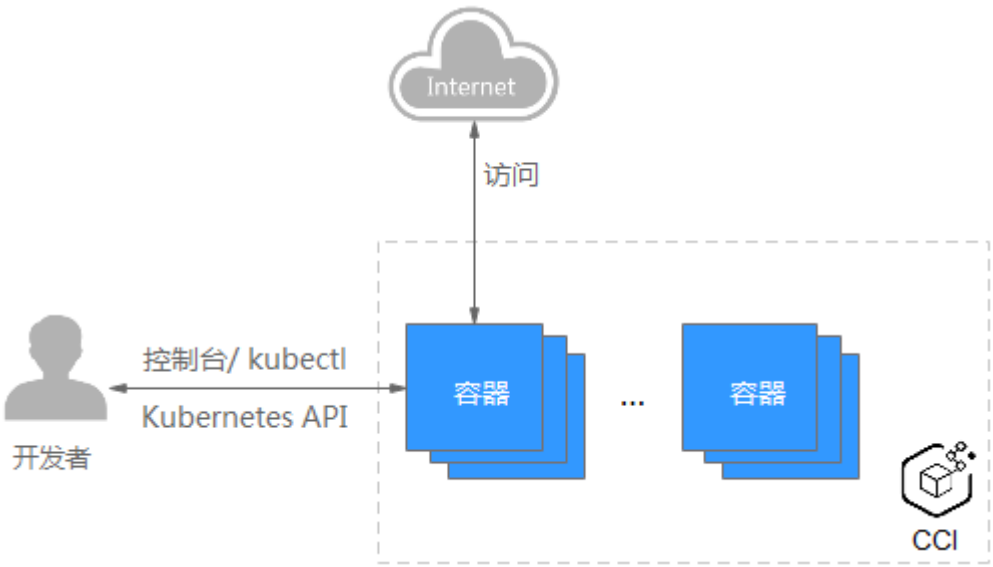
什么是云容器实例

云容器实例（Cloud Container Instance，CCI）服务提供Serverless Container（无服务器容器）引擎，让您无需创建和管理服务器集群即可直接运行容器。

Serverless是一种架构理念，是指不用创建和管理服务器、不用担心服务器的运行状态（服务器是否在工作等），只需动态申请应用需要的资源，把服务器留给专门的维护人员管理和维护，进而专注于应用开发，提升应用开发效率、节约企业IT成本。传统上使用Kubernetes运行容器，首先需要创建运行容器的Kubernetes服务器集群，然后再创建容器负载。

云容器实例的Serverless Container就是从使用角度，无需创建、管理Kubernetes集群，也就是从使用的角度看不见服务器（Serverless），直接通过控制台、kubectl、Kubernetes API创建和使用容器负载，且只需为容器所使用的资源付费。

图 1-1 使用云容器实例



产品功能

一站式容器生命周期管理

使用云容器实例，您无需创建和管理服务器集群即可直接运行容器。您可以通过控制台、kubectl、Kubernetes API创建和使用容器负载，且只需为容器所使用的资源付费。

支持多种类型计算资源

云容器实例提供了多种类型计算资源运行容器，包括CPU，GPU（提供NVIDIA Tesla V100、NVIDIA Tesla T4显卡）。

支持多种网络访问方式

云容器实例提供了丰富的网络访问方式，支持四层、七层负载均衡，满足不同场景下的访问诉求。

支持多种持久化存储卷

云容器实例支持将数据存储在云服务的云存储上，当前支持的云存储包括：云硬盘存储卷（EVS）、文件存储卷（SFS）、对象存储卷（OBS）和极速文件存储卷（SFS Turbo）。

支持极速弹性扩缩容

云容器实例支持用户自定义弹性伸缩策略，且能在1秒内实现弹性扩缩容，并可以自由组合多种弹性策略以应对业务高峰期的突发流量浪涌。

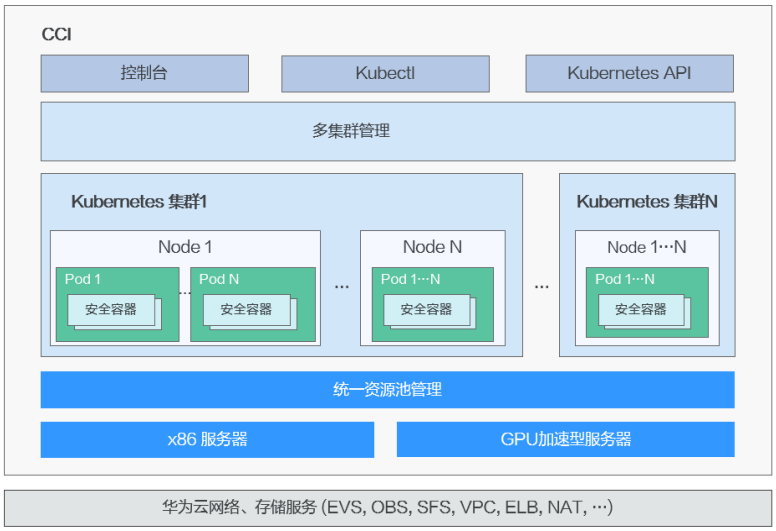
全方位容器状态监控

云容器实例支持监控容器运行的资源使用率，包括CPU、内存、GPU和显存的使用率，方便您实时掌控容器运行的状态。

产品架构

云容器实例提供Serverless Container服务，拥有多个异构的Kubernetes集群，并集成网络、存储服务，让您方便的通过控制台、kubectl、Kubernetes API创建和使用容器负载。

图 1-2 产品架构



- 基于云平台底层网络和存储服务（VPC、ELB、NAT、EVS、OBS、SFS等），提供丰富的网络和存储功能。

- 提供高性能、异构的基础设施（x86服务器、GPU加速型服务器、Ascend加速型服务器），容器直接运行在物理服务器上。
- 使用安全容器提供虚拟机级别的安全隔离，结合自有硬件虚拟化加速技术，提供高性能安全容器。
- 多集群统一管理，容器负载统一调度，使用上无需感知集群存在。
- 基于**Kubernetes**的负载模型提供负载快速部署、弹性负载均衡、弹性扩缩容、蓝绿发布等重要能力。

云容器实例学习路径

您可以借助云容器实例**成长地图**，快速了解产品，由浅入深学习使用和运维CCI。

2 产品优势

随启随用

业界领先的Serverless Container架构，用户无需创建Kubernetes服务器集群，直接使用控制台、kubectl、Kubernetes API创建容器。

极速弹性

云容器实例的Kubernetes集群是提前创建好的，且从单一用户角度看资源“无限大”，所以能够提供容器秒级弹性伸缩能力，让您能够轻松应对业务快速变化，稳健保障业务SLA。

按需秒级计费

根据实际使用的资源数量，按需按秒计费，避免业务不活跃时段的费用开销，降低用户成本。

完全开放的原生平台

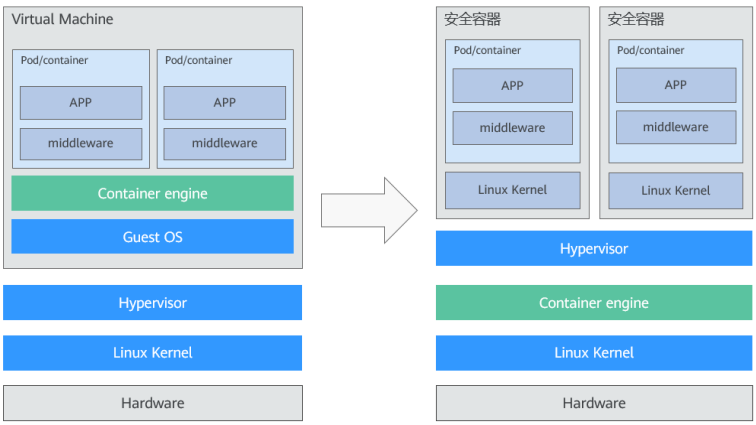
- 紧跟Kubernetes社区，迅速同步最新版本
- 原生支持Kubernetes API

高安全

云容器实例同时具备容器级别的启动速度和虚拟机级别的安全隔离能力，提供更好的容器体验。

- 原生支持安全容器。
- 基于安全容器的内核虚拟化技术，为您提供全面的安全隔离与防护。
- 自有硬件虚拟化加速技术，让您获得更高性能的安全容器。

图 2-1 通过安全容器实现多租户容器强隔离



3 应用场景

大数据、AI 计算

当前主流的大数据、AI训练和推理等应用（如Tensorflow、Caffe）均采用容器化方式运行，并需要大量GPU、高性能网络和存储等硬件加速能力，并且都是任务型计算，需要快速申请大量资源，计算任务完成后快速释放。

云容器实例提供如下特性，能够很好的支持这类场景。

- **计算加速**：提供GPU/Ascend等异构芯片加速能力
- **大规模网络容器实例调度**：支持大规模、高并发的容器创建和管理
- **随启随用、按需付费**：容器按需启动，按资源规格和使用时长付费

图 3-1 大数据 AI 计算场景

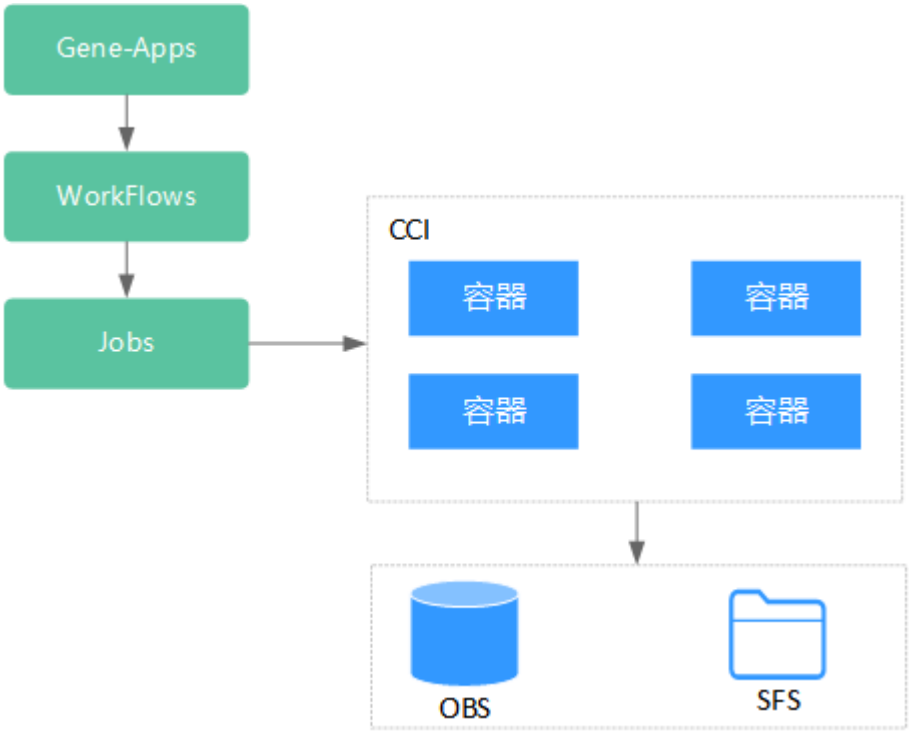
生物基因、药物研发等科学计算

生物基因、药品研发等领域需要高性能、密集型计算，同时对成本较敏感，需要低成本、免运维的计算平台。科学计算一般都是任务型计算，快速申请大量资源，完成后快速释放。

云容器实例提供如下特性，能够很好的支持这类场景。

- **高性能计算**：提供高性能计算、网络和高I/O存储，满足密集计算的诉求
- **极速弹性**：秒级资源准备与弹性，减少计算过程中的资源处理环节消耗
- **免运维**：无需感知集群和服务器，大幅简化运维工作、降低运维成本
- **随启随用、按需付费**：容器按需启动，按资源规格和使用时长付费

图 3-2 科学计算



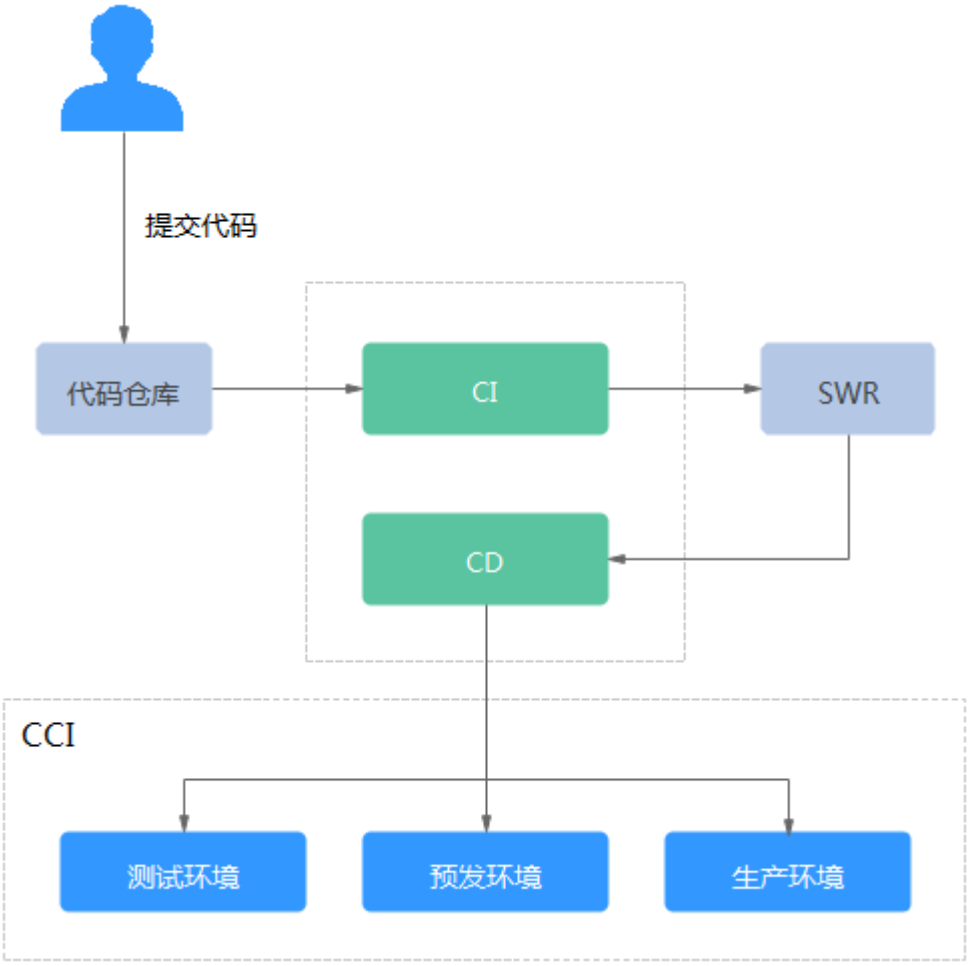
DevOps 持续交付

软件开发型企业，希望构建从代码提交到应用部署的DevOps完整流程，提高企业应用迭代效率。DevOps流程一般都是任务型计算，如企业CI/CD（持续集成/持续发布）流程自动化，需要快速申请资源，完成后快速释放。

云容器实例提供如下特性，能够很好的支持这类场景。

- **流程自动化**：无需创建和维护集群，实现从CI/CD的全流程自动化
- **环境一致性**：以容器镜像交付，可以无差别地从开发环境迁移到生产环境
- **随启随用、按需付费**：容器按需启动，按资源规格和使用时长付费

图 3-3 DevOps 持续交付



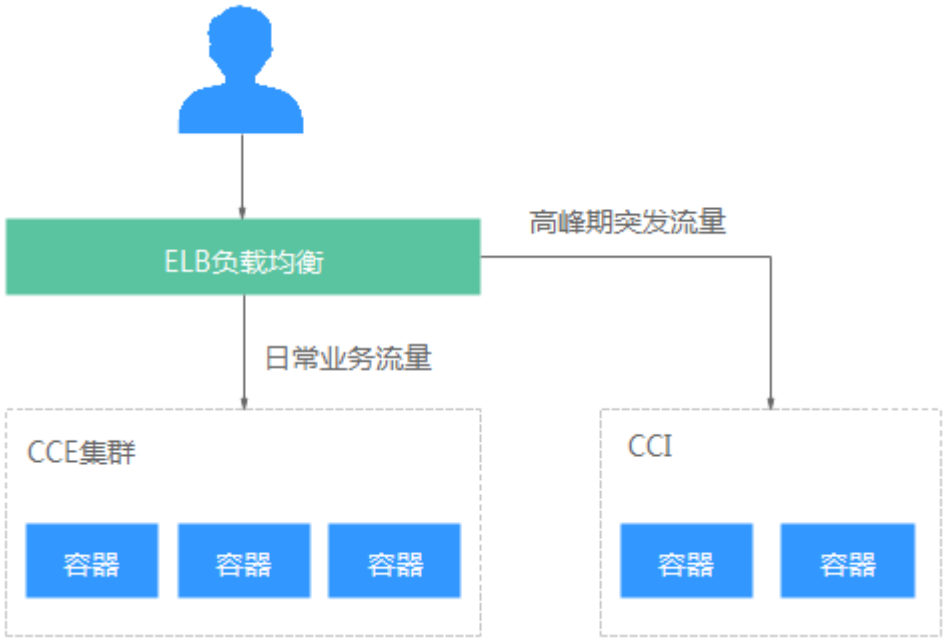
高弹性业务

业务波峰波谷较明显的业务，日常流量稳定，高峰期又需要快速扩展资源，并对成本有一定诉求，如视频直播、媒体资讯、电商、在线教育等应用。

云容器实例提供如下特性，能够很好的支持这类场景。

- **快速弹性伸缩**：业务高峰时，业务能够快速从CCE弹性扩展到CCI，保障业务稳定运行
- **低成本灵活计费**：业务平稳期在CCE上包周期计费，节省成本；高峰期弹性扩容到CCI上，按需计费，高峰期结束后又可以快速释放资源，降低成本

图 3-4 弹性扩展



4 基本概念

云容器实例基于Kubernetes的负载模型增强了容器安全隔离、负载快速部署、弹性负载均衡、弹性扩缩容、蓝绿发布等重要能力。

云容器实例提供Kubernetes原生API，支持使用kubectl，且提供图形化控制台，让您能够拥有完整的端到端使用体验，使用云容器实例前，建议您先了解相关的基本概念。

镜像（Image）

容器镜像是一个特殊的文件系统，除了提供容器运行时所需的程序、库、资源、配置等文件外，还包含了一些为运行时准备的配置参数（如匿名卷、环境变量、用户等）。镜像不包含任何动态数据，其内容在构建之后也不会被改变。

容器（Container）

镜像和容器的关系，就像是面向对象程序设计中的类和实例一样，镜像是静态的定义，容器是镜像运行时的实体。容器可以被创建、启动、停止、删除、暂停等。

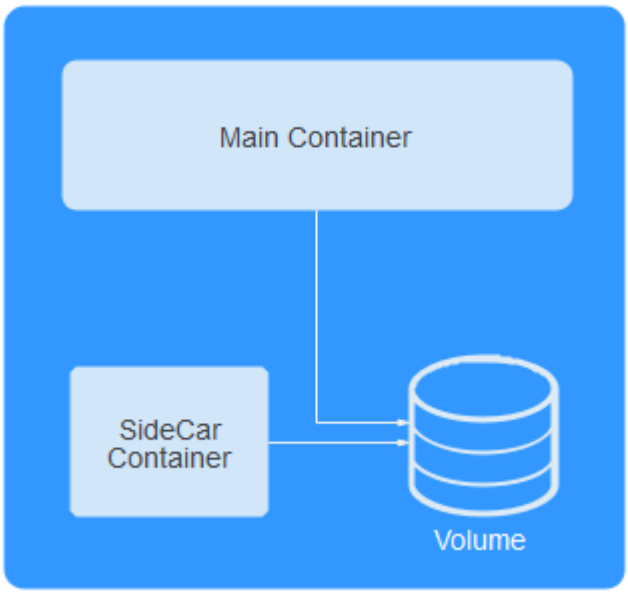
命名空间（Namespace）

命名空间是一种在多个用户之间划分资源的方法。当你的项目和人员众多的时候可以考虑根据项目属性，例如生产、测试、开发划分不同的namespace。

Pod

Pod是Kubernetes创建或部署的最小单位。一个Pod封装一个或多个容器、存储资源、一个独立的网络IP以及管理控制容器运行方式的策略选项。

图 4-1 Pod



Pod使用主要分为两种方式：

- Pod中运行一个容器。这是Kubernetes最常见的用法，你可以将Pod视为单个封装的容器，但是Kubernetes是直接管理Pod而不是容器。
- Pod中运行多个需要耦合在一起工作、需要共享资源的容器。

实际使用中很少直接创建Pod，而是使用Kubernetes中称为Controller的抽象层来管理Pod实例，例如Deployment。Controller可以创建和管理多个Pod，提供副本管理、滚动升级和自愈能力。通常，Controller会使用Pod Template来创建相应的Pod。

详细信息请参见**Pod**。

Init 容器（Init-Containers）

Init-Containers，即初始化容器，顾名思义容器启动的时候，会先启动一个或多个容器，如果有多个，那么这几个Init Container按照定义的顺序依次执行，只有所有的Init Container执行完后，主容器才会启动。由于一个Pod里的存储卷是共享的，所以Init Container里产生的数据可以被主容器使用到。

Init Container可以在多种K8S资源里被使用到如Deployment、Job等，但归根结底都是在Pod启动时，在主容器启动前执行，做初始化工作。

详细信息请参见**Init容器**。

标签

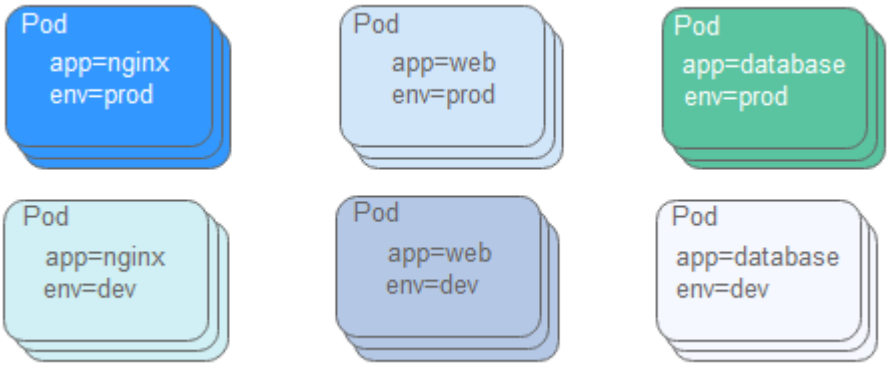
Label（标签）是一组附加在对象上的键值对，用来传递用户定义的属性。

标签常用来从一组对象中选取符合条件的对象，这也是Kubernates中目前为止最重要的节点分组方法。

比如，你可能创建了一个“tier”和“app”标签，通过Label（tier=frontend，app=myapp）来标记前端Pod容器，使用Label（tier=backend，app=myapp）标记后台Pod。然后可以使用Selectors选择带有特定Label的Pod，并且将Service或者Deployment应用到上面。

详细信息请参见**Label**。

图 4-2 使用 Label 组织的 Pod



无状态负载（Deployment）

Deployment是Pod Controller的一种。

一个Deployment可以包含一个或多个Pod，每个Pod的角色相同，所以系统会自动为Deployment的多个Pod分发请求。Deployment中的所有Pod共享存储卷。

使用Deployment时，您只需要在Deployment中描述您想要的目标状态是什么，Deployment就会帮您将Pod的状态改变到目标状态。

详细信息请参见**Deployment**。

短时任务（Job）

Job是用来控制批处理型任务的资源对象。批处理业务与长期伺服业务（Deployment）的主要区别是批处理业务的运行有头有尾，而长期伺服业务在用户不停止的情况下永远运行。Job管理的Pod根据用户的设置把任务成功完成就自动退出了。

Job的这种用完即停止的特性特别适合一次性任务，比如持续集成，配合云容器实例按秒计费，真正意义上做到按需使用、按需付费。

详细信息请参见**Job**。

定时任务（CronJob）

定时任务是基于时间控制的短时任务（Job），类似于Linux系统的crontab文件中的一行，在指定的时间周期运行指定的短时任务。

详细信息请参见**CronJob**。

服务（Service）

Pod是有生命周期的，它们可以被创建，也可以被销毁，然而一旦被销毁生命就永远结束。通过Pod Controller能够动态地创建和销毁Pod（例如，需要进行扩缩容，或者执行滚动升级）。每个Pod都会获取它自己的IP地址，但这些IP地址不总是稳定可依赖的。 这会导致一个问题：如果一组Pod（称为backend）为其它Pod（称为frontend）提供服务，那么那些frontend该如何发现，并连接到这组Pod中的哪些backend呢?

Service定义了这样一种抽象：一个Pod的逻辑分组，一种可以访问它们的策略（通常称为微服务）。 这一组Pod能够被Service访问到，通常是通过Label Selector实现的。

举个例子，考虑一个图片处理backend，它运行了3个Pod副本。这些副本是可互换的（frontend不需要关心它们调用了哪个backend副本）。 然而组成这一组backend的Pod实际上可能会发生变化，frontend不应该也没必要知道，而且也不需要跟踪这一组backend的状态。Service定义的抽象就是用来解耦这种关联。

详细信息请参见**Service**。

Ingress

Service和Pod仅可在内部网络中通过IP地址访问，外部的请求需要通过负载均衡转发到Service在Node上暴露的NodePort上，然后再由kube-proxy将其转发给相关的Pod。

Ingress是授权入站连接到达集群服务的规则集合。您可以给Ingress配置外部可访问的URL、负载均衡、SSL、基于名称的虚拟主机等。

详细介绍请参见**Ingress**。

PVC

PersistentVolumeClaim（PVC）是用户存储的请求。 它类似于Pod，Pod申请CPU和内存，PVC申请存储资源。在云容器实例中，你可以通过PVC申请EVS、SFS等存储资源。

详细信息请参见**PVC**。

ConfigMap

ConfigMap用于保存配置数据的键值对，可以用来保存单个属性，也可以用来保存配置文件。ConfigMap跟Secret很类似，但它可以更方便地处理不包含敏感信息的字符串。

详细信息请参见**ConfigMap**。

Secret

Secret是Kubernetes中一种加密存储的资源对象，用户可以将认证信息、证书、私钥等保存在密钥中，在容器启动时以环境变量等方式加载到容器中。

详细信息请参见**Secret**。

5 安全

5.1 责任共担

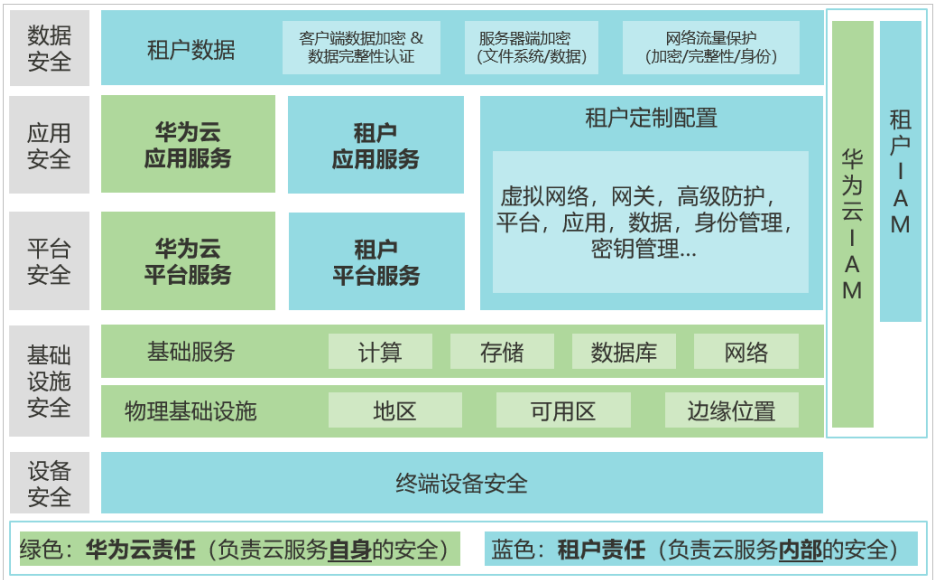
华为云秉承“将公司对网络和业务安全性保障的责任置于公司的商业利益之上”。针对层出不穷的云安全挑战和无孔不入的云安全威胁与攻击，华为云在遵从法律法规业界标准的基础上，以安全生态圈为护城河，依托华为独有的软硬件优势，构建面向不同区域和行业的完善云服务安全保障体系。

安全性是华为云与您的共同责任，如图5-1所示。

- **华为云**：负责云服务**自身**的安全，提供安全的云。华为云的安全责任在于保障其所提供的 IaaS、PaaS 和 SaaS 类云服务自身的安全，涵盖华为云数据中心的物理环境设施和运行其上的基础服务、平台服务、应用服务等。这不仅包括华为云基础设施和各项云服务技术的安全功能和性能本身，也包括运维运营安全，以及更广义的安全合规遵从。
- **租户**：负责云服务**内部**的安全，安全地使用云。 华为云租户的安全责任在于对使用的 IaaS、PaaS 和 SaaS 类云服务内部的安全以及对租户定制配置进行安全有效的管理，包括但不限于虚拟网络、虚拟主机和访客虚拟机的操作系统，虚拟防火墙、API 网关和高级安全服务，各项云服务，租户数据，以及身份帐号和密钥管理等方面的安全配置。

《华为云安全白皮书》详细介绍华为云安全性的构建思路与措施，包括云安全战略、责任共担模型、合规与隐私、安全组织与人员、基础设施安全、租户服务与租户安全、工程安全、运维运营安全、生态安全。

图 5-1 华为云安全责任共担模型



5.2 身份认证与访问控制

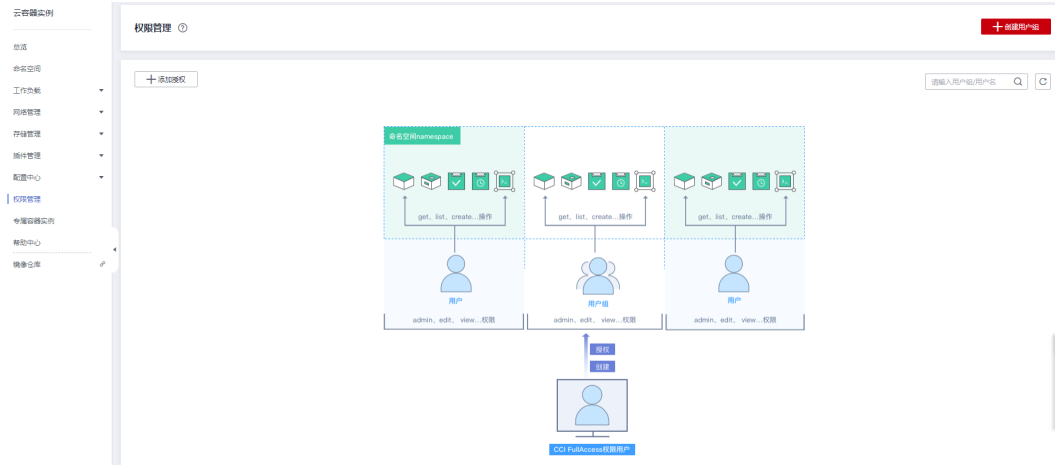
CCI当前认证鉴权是在Kubernetes的角色访问控制（RBAC）与统一身份认证服务（IAM）的能力基础上，提供的基于IAM的细粒度权限控制和IAM Token认证，同时支持命名空间级别及命名空间以下资源的权限控制，帮助用户便捷灵活的对租户下的IAM用户、用户组设定不同的操作权限。

- **命名空间权限：**是基于Kubernetes RBAC能力的授权。通过权限设置可以让不同的用户或用户组拥有操作指定Namespace下Kubernetes资源的权限。
- **CCI权限：**是基于IAM的细粒度授权。通过命名空间级别权限设置可以控制用户操作Namespace（如创建、删除Namespace等）。更多细粒度权限说明请参见**CCI细粒度鉴权系统策略关联Actions**。

说明

- 创建Namespace时，打开RBAC鉴权开关，则此Namespace下资源访问受RBAC鉴权控制；如果未打开RBAC鉴权开关，则RBAC鉴权不生效。
- 创建开启RBAC鉴权的Namespace后，需要先对用户授权后，用户才能使用这个Namespace。
- network、clusterRole和roleBinding资源不受RBAC权限影响，只受IAM细粒度鉴权控制。network受控于network相关action，clusterRole与roleBinding受控于rbac相关action。
- 支持对当前用户下的所有命名空间进行授权，以提供更好的前端显示体验。

图 5-2 CCI 权限管理



命名空间权限

Kubernetes RBAC API定义了四种类型：Role、ClusterRole、RoleBinding与ClusterRoleBinding。当前CCI仅支持ClusterRole、RoleBinding，这两种类型之间的关系和简要说明如下：

- **ClusterRole：**描述角色和权限的关系。在Kubernetes的RBAC API中，一个角色定义了一组特定权限的规则。整个Kubernetes集群范围内有效的角色则通过ClusterRole对象实现。
- **RoleBinding：**描述subjects（包含users，groups）和角色的关系。角色绑定将一个角色中定义的各种权限授予一个或者一组用户，该用户或用户组则具有对应绑定ClusterRole定义的权限。

表 5-1 RBAC API 所定义的两种类型

类型名称	说明
ClusterRole	ClusterRole对象可以授予整个集群范围内资源访问权限。
RoleBinding	RoleBinding可以将同一Namespace中的subject（用户）绑定到某个具有特定权限的ClusterRole下，则此subject即具有该ClusterRole定义的权限。

⚠ 注意

当前仅支持用户使用ClusterRole在Namespace下创建RoleBinding。

CCI中的命名空间权限是基于Kubernetes RBAC能力的授权，通过权限设置可以让不同的用户或用户组拥有操作不同Kubernetes资源的不同权限。

CCI的kubernetes资源通过命名空间进行权限设置，目前包含**cluster-admin**、**admin**、**edit**、**view**四种角色，详见表5-2。

表 5-2 用户/用户组角色说明

默认的ClusterRole	描述
cluster-admin	具有Kubernetes所有资源对象操作权限。
admin	允许admin访问，可以限制在一个Namespace中使用RoleBinding。如果在RoleBinding中使用，则允许对Namespace中大多数资源进行读写访问。这一角色不允许操作Namespace本身，也不能写入资源限额。
edit	允许对命名空间内的大多数资源进行读写操作。
view	允许对多数对象进行只读操作，但是对secret是不可访问的。

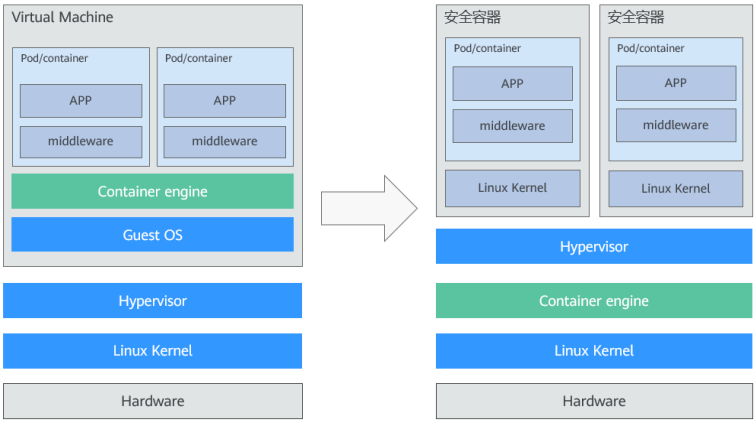
更多Kubernetes RBAC授权的内容可以参考**Kubernetes RBAC官方文档**。

5.3 数据保护技术

云容器实例同时具备容器级别的启动速度和虚拟机级别的安全隔离能力，提供更好的容器体验。

- 原生支持安全容器。
- 基于安全容器的内核虚拟化技术，为您提供全面的安全隔离与防护。
- 自有硬件虚拟化加速技术，让您获得更高性能的安全容器。

图 5-3 通过安全容器实现多租户容器强隔离



SSL

SSL（安全套接层，Secure Sockets Layer）是一种安全协议，目的是为互联网通信，提供安全及数据完整性保障。

云容器实例支持上传SSL证书，在使用HTTPS访问时，云容器实例将SSL证书自动安装到七层负载均衡器上，实现数据传输加密。详细信息请参见**SSL证书**。

Secret

Secret是Kubernetes中一种加密存储的资源对象，用户可以将认证信息、证书、私钥等保存在密钥中，在容器启动时以环境变量等方式加载到容器中。

详细信息请参见**Secret**。

5.4 审计与日志

审计

云审计服务（Cloud Trace Service，CTS），是华为云安全解决方案中专业的日志审计服务，提供对各种云资源操作记录的收集、存储和查询功能，可用于支撑安全分析、合规审计、资源跟踪和问题定位等常见应用场景。

用户开通云审计服务并创建和配置追踪器后，CTS可记录您从云管理控制台或者开放API发起的云服务资源操作请求以及每次请求的结果。

CTS的详细介绍和开通配置方法，请参见**CTS快速入门**。

CTS支持追踪的CCI操作列表，请参见**云审计服务支持的CCI操作列表**。

CCI记录的审计日志会上报到CTS，供用户查询和分析，详细介绍和配置方法，请参见**查看云审计日志**。

日志

CCI为用户提供日志管理功能，用户可配置容器的日志路径和日志上报地址，Pod中集成的fluentbit插件会从日志路径采集日志，并上报到LTS，详细介绍和配置方法，请参见**日志管理**。

整体上CCI的安全日志能力已对接CTS、CLS、LTS、AOM服务，其后续相关安全性、完整性等能力由相关承载服务跟踪。

5.5 监控安全风险

通过 AOM 查看 Pod 监控数据

为使用户更好的掌握工作负载的运行状态，CCI配合AOM对其进行全方位的监控。

通过AOM界面您可监控CCI的基础资源和运行在CCI上的应用，同时在AOM界面还可查看相关的日志和告警。

更多内容，请参见**监控管理**。

Pod 资源监控指标

CCI支持Pod资源基础监控能力，提供CPU、内存、磁盘、网络等多种监控指标，满足对Pod资源的基本监控需求。

Pod内置系统agent，默认会以http服务的形式提供Pod和容器的监控指标。

- CCI支持的资源监控指标，请参见**资源监控指标**。

- Pod资源基础监控能力，请参见**Pod资源监控指标**。

6 约束与限制

本章介绍CCI相关的使用限制，以便于您更好地使用CCI。

CCI 实例限制

下表为CCI实例相关的使用限制。

限制项	限制描述
创建CCI实例的用户帐号限制	已通过实名认证。
单个用户的资源数量和容量配额限制	云容器实例对单个用户的资源数量和容量限定了配额，您可以登录华为云控制台，在“资源 > 我的配额>服务配额”页面，查看各项资源的总配额及使用情况。 **说明** 如果当前配额不能满足业务要求，可申请扩大配额。配额的详细信息请参见**关于配额**。
单个CCI实例的vCPU数量	0.25核-32核，或者自定义选择48核、64核。
支持的容器操作系统类型	仅支持Linux容器。
CCI实例的网络类型	仅支持VPC网络。

Kubernetes 应用限制

基于华为云的安全性带来的限制，CCI目前还不支持Kubernetes中HostPath、DaemonSet等功能，具体如下表所示。

不支持的功能	说明	推荐替代方案
HostPath	挂载本地宿主机文件到容器中	使用云盘或者SFS文件系统
HostNetwork	将宿主机端口映射到容器上	使用type=LoadBalancer的负载均衡

不支持的功能	说明	推荐替代方案
DaemonSet	DaemonSet（守护进程集）在集群的每个节点上运行一个Pod，且保证只有一个Pod	通过sidecar形式在Pod中部署多个容器
Privileged权限	容器拥有privileged权限	使用Security Context为Pod添加Capability
type=NodePort的Service	将宿主机端口映射到容器上	使用type=LoadBalancer的负载均衡

Pod 规格限制

云容器实例当前支持使用GPU，您可以根据需要选择，实例收费详情请参见**产品价格详情**。

当不使用GPU时，Pod规格需满足如下要求：

表 6-1 Pod 规格限制要求

Pod规格限制项	限制取值范围
Pod的CPU	• 0.25核-32核，或者自定义选择48核、64核。 • 单个容器的CPU必须为0.25核的整数倍。
Pod的内存	• 1GiB-512GiB。 • 内存必须为1GiB的整数倍。
Pod的CPU/内存配比值	在1:2至1:8之间。
Pod的容器	一个Pod内最多支持5个容器。 单个容器最小配置是0.25核、0.2GiB，最大同容器实例的最大配置。
Pod中所有容器和InitContainer（启动容器）	两者规格中的request和limit相等。

说明

- Pod规格计算详情请参见**Pod规格计算方式**。
- InitContainer是一种特殊容器，在 Pod 内的应用容器启动之前运行。有关InitContainer更多解释请参见**对容器进行初始化操作**。

GPU加速型Pod提供3种显卡，具体的规格如下所示：

表 6-2 GPU 加速型 Pod 规格

显卡类型	具体规格	可用区域
NVIDIA Tesla T4显卡	• NVIDIA Tesla T4 x 1，CPU 8核，内存32GiB • NVIDIA Tesla T4 x 2，CPU 16核，内存64GiB • NVIDIA Tesla T4 x 4，CPU 32核，内存128GiB • NVIDIA Tesla T4 x 8，CPU 64核，内存256GiB	华北-北京四
NVIDIA Tesla V100 16G显卡	• NVIDIA Tesla V100 16G x 1，CPU 4核，内存32GiB • NVIDIA Tesla V100 16G x 2，CPU 8核，内存64GiB • NVIDIA Tesla V100 16G x 4，CPU 16核，内存128GiB • NVIDIA Tesla V100 16G x 8，CPU 32核，内存256GiB	华东-上海一
NVIDIA Tesla V100 32G显卡	• NVIDIA Tesla V100 32G x 1，CPU 4核，内存32GiB • NVIDIA Tesla V100 32G x 2，CPU 8核，内存64GiB • NVIDIA Tesla V100 32G x 4，CPU 16核，内存128GiB • NVIDIA Tesla V100 32G x 8，CPU 32核，内存256GiB	华北-北京四

云容器实例支持使用NVIDIA GPU的驱动版本为**460.106**和**418.126**，您应用程序中使用的CUDA需满足如**表6-3**所示的配套关系。CUDA与驱动的配套关系来源于NVIDIA官网，详细信息请参见**CUDA Compatibility**。

表 6-3 NVIDIA GPU 驱动与 CUDA 配套关系

NVIDIA GPU驱动版本	CUDA Toolkit版本
460.106	CUDA 11.2.2 Update 2 及以下
418.126	CUDA 10.1 (10.1.105)及以下

GPU镜像

CUDA和cuDNN都是与GPU相关的技术，用于加速各种计算任务，特别是深度学习任务。在使用NVIDIA GPU进行深度学习时，通常需要安装CUDA和cuDNN。请使用配套关系的基础镜像。

Pod 存储空间限制

如果没有挂载EVS等磁盘，应用数据存储在容器的rootfs，每个Pod存储空间限制如下所示：

表 6-4 每个 Pod 存储空间限制

Pod类型	存储空间限制
CPU型Pod	20G
GPU型Pod	20G

7 计费说明

计费项

实例资源包含CPU、内存、GPU等，根据您申请的实际实例资源规格，以及每个实例实际运行时长按秒计费。计费时长从开始下载容器镜像 (docker pull) 到CCI实例停止使用进行计算。

计费模式

云容器实例分为按需收费和购买套餐包两种计费模式。具体计费详情请参见**产品价格详情**。

说明

费用计算核时为核数和时间相乘，例如：730核时，表示您可以730核用1小时，也可以730小时用1核。

- 1 核*时= 1 * 3600（核*秒）
- 1 核*时：1核的CPU连续跑1个小时所用的资源量
- 1 核*秒：1核的CPU连续跑1秒所用的资源量

按需计费模式

以实例为单位，采用按量付费的计费模式，按秒计费，以小时为出账周期。

套餐包模式

- 购买的资源包在有效期内，扣费方式是先扣除已购买的资源包内的额度后，超出部分以按需付费的方式进行结算。
- 用户可重复购买资源包，当存在多个资源包时，优先从过期时间早的资源包中扣除。

  说明

  额度重置：选择额度重置的套餐包，资源套餐包的额度在当前周期结束后，下个周期会自动恢复额度。

8 权限管理

如果您需要对云平台上购买的CCI资源，为企业中的员工设置不同的访问权限，以达到不同员工之间的权限隔离，您可以使用统一身份认证服务（Identity and Access Management，简称IAM）进行精细的权限管理。该服务提供用户身份认证、权限分配、访问控制等功能，可以帮助您安全的控制云资源的访问。

通过IAM，您可以在您的云账号中给员工创建IAM用户，并授权控制他们对云资源的访问范围。例如您的员工中有负责软件开发的人员，您希望他们拥有CCI的使用权限，但是不希望他们拥有删除CCI等高危操作的权限，那么您可以使用IAM为开发人员创建用户，通过授予仅能使用CCI，但是不允许删除CCI的权限，控制他们对CCI资源的使用范围。

如果您的云账号已经能满足您的要求，不需要创建独立的IAM用户进行权限管理，您可以跳过本章节，不影响您使用CCI服务的其它功能。

IAM是云平台提供权限管理的基础服务，无需付费即可使用，您只需要为您账号中的资源进行付费。关于IAM的详细介绍，请参见**IAM产品介绍**。

CCI 权限

默认情况下，管理员创建的IAM用户没有任何权限，需要将其加入用户组，并给用户组授予策略或角色，才能使得用户组中的用户获得对应的权限，这一过程称为授权。授权后，用户就可以基于被授予的权限对云服务进行操作。

CCI部署时通过物理区域划分，为项目级服务。授权时，“作用范围”需要选择“区域级项目”，然后在指定区域（如华北-北京四）对应的项目（cn-north-4）中设置相关权限，并且该权限仅对此项目生效；如果在“所有项目”中设置权限，则该权限在所有区域项目中都生效。访问CCI时，需要先切换至授权区域。

根据授权精细程度分为角色和策略。

- 角色：IAM最初提供的一种根据用户的工作职能定义权限的粗粒度授权机制。该机制以服务为粒度，提供有限的服务相关角色用于授权。由于云平台各服务之间存在业务依赖关系，因此给用户授予角色时，可能需要一并授予依赖的其他角色，才能正确完成业务。角色并不能满足用户对精细化授权的要求，无法完全达到企业对权限最小化的安全管控要求。
- 策略：IAM最新提供的一种细粒度授权的能力，可以精确到具体服务的操作、资源以及请求条件等。基于策略的授权是一种更加灵活的授权方式，能够满足企业对权限最小化的安全管控要求。例如：针对CCI服务，管理员能够控制IAM用户仅能对某一类云容器实例资源进行指定的管理操作。多数细粒度策略以API接口为粒度进行权限拆分，CCI支持的API授权项请参见**权限策略和授权项**。

如表8-1所示，包括了CCI的所有系统策略。

表 8-1 CCI 系统策略

策略名称	描述	策略类别
CCI FullAccess	云容器实例所有权限，拥有该权限的用户可以执行云容器实例所有资源的创建、删除、查询、更新操作。 **说明** 对象存储服务OBS为全局级服务，若需要使用对象存储服务请为其单独授予权限，授权操作请参见**对象存储服务权限控制**。	系统策略
CCI ReadOnlyAccess	云容器实例只读权限，拥有该权限的用户仅能查看云容器实例资源。	系统策略
CCI CommonOperations	云容器实例普通用户，拥有该权限的用户可以执行除RBAC、network和namespace子资源创建、删除、修改之外的所有操作。	系统策略
CCI Administrator	云容器实例管理员权限，拥有该权限的用户可以执行云容器实例所有资源的创建、删除、查询、更新操作。	系统角色

CCI FullAccess策略权限如下：

表 8-2 CCI FullAccess 策略主要权限

操作（Action）	说明
cci:*:*	CCI（云容器实例）服务的所有权限
vpc:*:*	VPC（虚拟私有云）服务的所有权限
elb:*:*	ELB（弹性负载均衡）服务的所有权限
sfs:*:*	SFS（弹性文件服务）服务的所有权限
evs:*:*	EVS（云硬盘）服务的所有权限
aom:*:*	AOM（应用运维管理）服务的所有权限
apm:*:*	APM（应用性能管理）服务的所有权限
swr:*:*	SWR（容器镜像服务）服务的所有权限
nat:*:*	NAT（NAT网关）服务的所有权限
kms:cmk:*	DEW（数据加密服务）服务的所有权限

CCI ReadOnlyAccess策略权限如下：

表 8-3 CCI ReadOnlyAccess 策略主要权限

操作（Action）	说明
cci:*:get	CCI（云容器实例）所有资源详情的查看权限
cci:*:list	CCI（云容器实例）所有资源列表的查看权限
vpc:*:get	VPC（虚拟私有云）所有资源详情的查看权限
vpc:*:list	VPC（虚拟私有云）所有资源列表的查看权限
ecs:*:get	ECS（弹性云服务器）所有资源详情的查看权限
ecs:*:list	ECS（弹性云服务器）所有资源列表的查看权限
elb:*:get	ELB（弹性负载均衡）所有资源详情的查看权限
elb:*:list	ELB（弹性负载均衡）所有资源列表的查看权限
sfs:*:get*	SFS（弹性文件系统）所有资源详情的查看权限
sfs:*:list	SFS（弹性文件系统）所有资源列表的查看权限
obs:*:get*	OBS（对象存储服务）服务所有资源详情的查看权限
obs:*:list	OBS（对象存储服务）服务所有资源列表的查看权限
evs:*:get*	EVS（云硬盘）服务所有资源详情的查看权限
evs:*:list	EVS（云硬盘）服务所有资源列表的查看权限
aom:*:get	AOM（应用运维管理）服务所有资源详情的查看权限
aom:*:list	AOM（应用运维管理）服务所有资源列表的查看权限
amp:*:get	APM（应用性能管理）服务所有资源详情的查看权限
apm:*:list	APM（应用性能管理）服务所有资源列表的查看权限
swr:*:get	SWR（容器镜像服务）服务所有资源详情的查看权限
swr:*:list	SWR（容器镜像服务）服务所有资源列表的查看权限
nat:*:get	NAT（NAT网关）服务所有资源详情的查看权限
nat:*:list	NAT（NAT网关）服务所有资源列表的查看权限
kms:cmk:get	查询密钥信息
kms:cmk:list	查询密钥列表

CCI CommonOperations策略权限如下：

表 8-4 CCI CommonOperations 策略主要权限

操作（Action）	说明
cci:rbac:get	查询rbac信息
cci:rbac:list	查询rbac列表
cci:namespace:get	查询所有namespaces
cci:namespace:list	列出所有namespaces
cci:network:get	查询network详情
cci:network:list	查询network列表
cci:namespaceSubResource:*	namespace子资源的所有权限
cci:addonTemplate:*	插件模板的所有权限
cci:addonInstance:*	插件实例的所有权限
vpc:*:*	VPC（虚拟私有云）服务的所有权限
elb:*:*	ELB（弹性负载均衡）服务的所有权限
sfs:*:*	SFS（弹性文件服务）服务的所有权限
obs:*:*	OBS（对象存储服务）服务的所有权限
evs:*:*	EVS（云硬盘）服务的所有权限
aom:*:*	AOM（应用运维管理）服务的所有权限
apm:*:*	APM（应用性能管理）服务的所有权限
swr:*:*	SWR（容器镜像服务）服务的所有权限
nat:*:*	NAT（NAT网关）服务的所有权限
kms:cmk:*	DEW（数据加密服务）服务的所有权限

CCI细粒度鉴权系统策略关联Actions如下：

表 8-5 CCI 细粒度鉴权系统策略关联 Actions

操作（Action）	说明
CCI:rbac:get	查询rbac详情
CCI:rbac:list	查询rbac列表
CCI:rbac:update	更新rbac
CCI:rbac:delete	删除rbac
CCI:rbac:create	创建rbac

操作（Action）	说明
CCI:namespaceSubResource:Create	创建namespace下子资源
CCI:namespaceSubResource:List	查询kubernetes资源列表
CCI:namespaceSubResource:Get	查询kubernetes资源
CCI:namespaceSubResource:Delete	删除kubernetes资源
CCI:namespaceSubResource:Update	更新kubernetes资源
CCI:network:update	更新network
CCI:network:create	创建network
CCI:network:delete	删除network
CCI:network:list	查询network列表
CCI:network:get	查询network详情
CCI:addonInstance:create	创建插件实例
CCI:addonInstance:update	更新升级插件实例
CCI:addonInstance:delete	删除插件实例
CCI:addonInstance:get	获取插件实例
CCI:addonInstance:list	列出所有插件实例
CCI:addonTemplate:list	列出所有插件模板
CCI:addonTemplate:get	获取插件模板
CCI:namespace:get	获取指定namespace
CCI:namespace:update	更新namespace
CCI:namespace:create	创建namespace
CCI:namespace:list	列出所有namespaces
CCI:namespace:delete	删除namespace

表8-6列出了CCI常用操作与系统策略的授权关系，您可以参照该表选择合适的系统策略。

表 8-6 常用操作与系统策略的关系

操作	CCI FullAccess	CCI ReadOnlyAccess	CCI CommonOperations
创建无状态负载	√	x	√
删除无状态负载	√	x	√
查看无状态负载	√	√	√
升级负载	√	x	√
伸缩负载	√	x	√
删除Pod	√	x	√
查看Pod	√	√	√
创建任务	√	x	√
删除任务	√	x	√
查看任务	√	√	√
创建定时任务	√	x	√
删除定时任务	√	x	√
查看定时任务	√	√	√
查看资源使用率	√	√	√
添加云硬盘卷	√	x	√
删除云硬盘卷	√	x	√
查看云硬盘卷	√	√	√
创建文件存储卷	√	x	√
删除文件存储卷	√	x	√
查看文件存储卷	√	√	√
创建ConfigMap	√	x	√
删除ConfigMap	√	x	√
查看ConfigMap	√	√	√
创建Secret	√	x	√
删除Secret	√	x	√
查看Secret	√	√	√

操作	CCI FullAccess	CCI ReadOnlyAccess	CCI CommonOperations
添加SSL证书	√	x	√
删除SSL证书	√	x	√
查看SSL证书	√	√	√
添加日志存储	√	x	√
查看日志	√	√	√
安装插件	√	x	√
删除插件	√	x	√
查看插件	√	√	√
查看授权	√	√	√
新增授权	√	x	x
删除授权	√	x	x
获取指定namespace	√	x	√
创建namespace	√	x	x
删除namespace	√	x	x
创建network	√	x	x
删除network	√	x	x
查询network列表	√	√	√
查询network详情	√	√	√

相关链接

- IAM产品介绍
- 创建用户组、用户并授予CCI权限
- 权限策略和授权项

9 区域和可用区

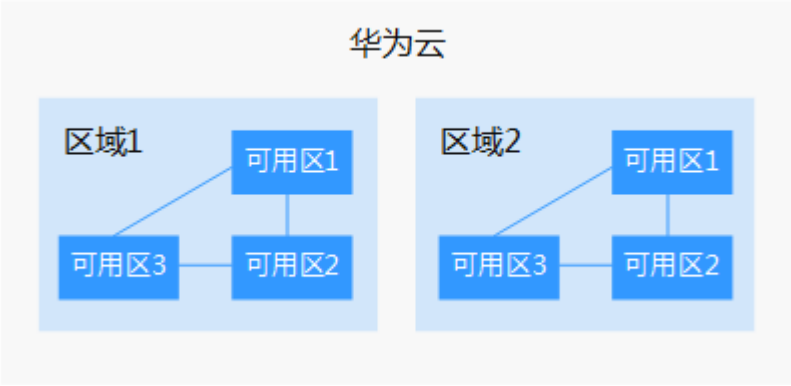
什么是区域、可用区?

区域和可用区用来描述数据中心的位置，您可以在特定的区域、可用区创建资源。

- 区域（Region）：从地理位置和网络时延维度划分，同一个Region内共享弹性计算、块存储、对象存储、VPC网络、弹性公网IP、镜像等公共服务。Region分为通用Region和专属Region，通用Region指面向公共租户提供通用云服务的Region；专属Region指只承载同一类业务或只面向特定租户提供业务服务的专用Region。
- 可用区（AZ，Availability Zone）：一个AZ是一个或多个物理数据中心的集合，有独立的风火水电，AZ内逻辑上再将计算、网络、存储等资源划分成多个集群。一个Region中的多个AZ间通过高速光纤相连，以满足用户跨AZ构建高可用性系统的需求。

图9-1阐明了区域和可用区之间的关系。

图 9-1 区域和可用区



目前，华为云已在全球多个地域开放云服务，您可以根据需求选择适合自己的区域和可用区。更多信息请参见华为云全球站点。

如何选择区域?

选择区域时，您需要考虑以下几个因素：

- 地理位置

  一般情况下，建议就近选择靠近您或者您的目标用户的区域，这样可以减少网络时延，提高访问速度。
- 资源的价格

  不同区域的资源价格可能有差异，请参见华为云服务价格详情。

如何选择可用区?

是否将资源放在同一可用区内，主要取决于您对容灾能力和网络时延的要求。

- 如果您的应用需要较高的容灾能力，建议您将资源部署在同一区域的不同可用区内。
- 如果您的应用要求实例之间的网络延时较低，则建议您将资源创建在同一可用区内。

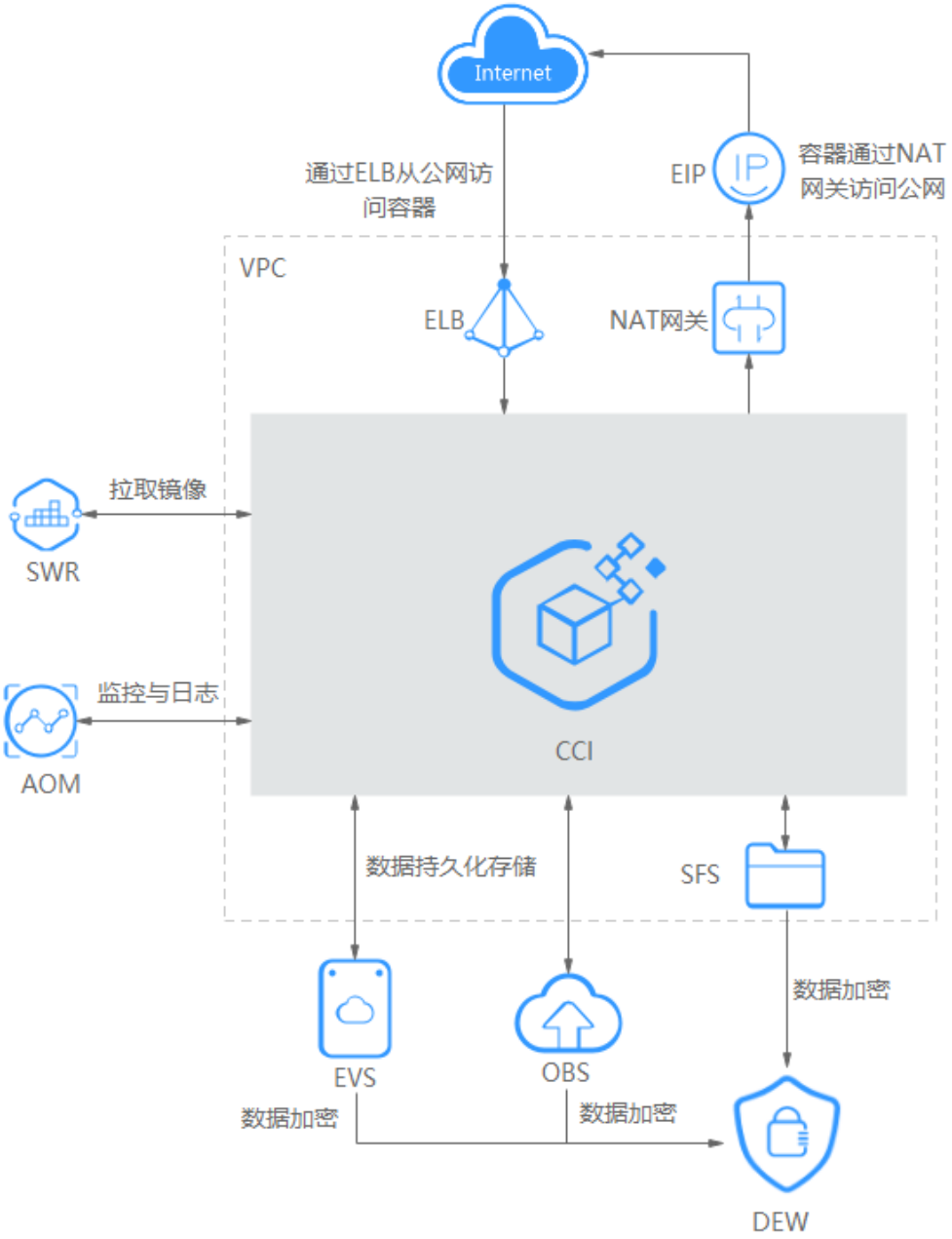
区域和终端节点

当您通过API使用资源时，您必须指定其区域终端节点。有关华为云的区域和终端节点的更多信息，请参阅**地区和终端节点**。

10 与其他服务的关系

云容器实例需要与其他云服务协同工作，云容器实例需要获取如下云服务资源的权限。

图 10-1 云容器实例与其他服务的关系



- **容器镜像服务**

  容器镜像服务（Software Repository for Container，SWR）是一种支持容器镜像全生命周期管理的服务， 提供简单易用、安全可靠的镜像管理功能，帮助用户快速部署容器化服务。

  您可以使用容器镜像服务中的镜像创建负载。

- **虚拟私有云**

  虚拟私有云（Virtual Private Cloud，VPC）是用户在云平台上申请的隔离的、私密的虚拟网络环境。用户可以自由配置VPC内的IP地址段、子网、安全组等子服务，也可以申请弹性带宽和弹性IP搭建业务系统。

  您创建命名空间时，需要创建或关联VPC，创建在命名空间的容器都运行在VPC之内。

- **弹性负载均衡**

  弹性负载均衡（ Elastic Load Balance，ELB）将访问流量自动分发到多台云服务器，扩展应用系统对外的服务能力，实现更高水平的应用容错。

  您可以通过弹性负载均衡，从外部网络访问容器负载。

- **应用运维管理**

  应用运维管理（Application Operations Management，AOM）为运维人员提供一站式立体运维平台，实时监控应用、资源运行状态，通过数十种指标、告警与日志关联分析，快速锁定问题根源，保障业务顺畅运行。

  云容器实例对接了AOM，AOM会采集容器日志存储中的“.log”等格式日志文件，转储到AOM中，方便您查看和检索；并且云容器实例基于AOM进行资源监控，为您提供弹性伸缩能力。

- **云硬盘服务**

  云硬盘（Elastic Volume Service，EVS）提供持久性块存储的服务，通过数据冗余和缓存加速等多项技术，提供高可用性和持久性，以及稳定的低时延性能。您可以对云硬盘做格式化、创建文件系统等操作，并对数据做持久化存储。

  您可以使用云硬盘作为容器的持久化存储，在创建负载的时候挂载到容器上。

- **对象存储服务**

  对象存储服务（Object Storage Service，OBS）是一个基于对象的海量存储服务，为客户提供海量、安全、高可靠、低成本的数据存储能力，包括：创建、修改、删除桶，上传、下载、删除对象等。

  您可以使用对象存储服务作为容器的持久化存储，在创建任务的时候挂载到容器上。

- **弹性文件服务**

  弹性文件服务（Scalable File Service）为用户提供托管的共享文件存储，符合标准文件协议（NFS），能够弹性伸缩至PB规模，具备可扩展的性能，为海量数据、高带宽型应用提供有力支持。

  您可以使用弹性文件服务作为容器的持久化存储，在创建任务负载的时候挂载到容器上。

- **弹性云服务器**

  弹性云服务器（Elastic Cloud Server）是一种可随时自助获取、可弹性伸缩的云服务器，帮助用户打造可靠、安全、灵活、高效的应用环境。

  云容器实例通过ECS将数据导入到SFS，进而供容器业务使用。

- **NAT网关**

  NAT网关能够为VPC内的容器实例提供网络地址转换（Network Address Translation）服务，SNAT功能通过绑定弹性公网IP，实现私有IP向公有IP的转换，可实现VPC内的容器实例共享弹性公网IP访问Internet。

  您可以通过NAT网关设置SNAT规则，使得容器能够访问Internet。

- **数据加密服务**

  数据加密服务（Data Encryption Workshop）是一个综合的云上数据加密服务。它可以提供专属加密、密钥管理、密钥对管理等功能。其密钥由硬件安全模块（HSM） 保护，并与许多华为云服务集成。用户也可以借此服务开发自己的加密应用。

