ALIBABA CLOUD

Alibaba Cloud HCI
AliyunHCI-Z版 超融合产品

技术白皮书

产品版本：V1.0.0

文档版本：20230406

阿里云

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。
2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。
3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。
4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。
5. 阿里云文档中所有内容，包括但不限于图片、架构设计、页面布局、文字描述，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。
6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

目录

1 概述

行业背景

随着数据不断增长以及互联网业务的兴起，新兴业务的激增、业务数据呈现几何倍数增加，传统服务器+存储的架构已经无法很好满足业务发展需求，分布式、云化技术应运而生。越来越多的企业采用虚拟化与云计算技术来构建IT系统，提升IT系统的资源利用率以及缩短业务上线周期。但在应用过程中，企业面临如下挑战：

- 虚拟平台部署和管理复杂，运维费用仍然维持增长趋势。
- 安装部署复杂，硬件来自多厂商，规划、部署、调优需要丰富的经验支撑。
- 多厂商设备，售后支持界面多，解决问题慢。
- 系统庞大（不同厂商硬件设备维护、虚拟平台管理），维护难度大。
- 企业越来越关注成本控制、业务敏捷、风险管控，希望能拥有总成本低、新业务的上线时间快、资源可弹性伸缩、安全可靠、高性能的IT系统。

AliyunHCI-Z应运而生

AliyunHCI-Z版 超融合产品（简称AliyunHCI-Z）是基于超融合架构研发的新一代IT基础设施平台。在超融合内集成了计算虚拟化、存储虚拟化、网络虚拟化等业界前沿技术，通过开箱即用实现快速交付，并通过统一管理平台实现IT资源的可视化管理和弹性扩展，帮助用户快速构建简单、稳定、安全、高效的新型IT基础架构。

2 技术优势

极简

- 极简安装部署：开箱即用，一键初始化系统，实现业务和测试系统敏捷应用交付，缩短部署实施周期。
- 极简上云：自带V2V迁移、VMware纳管功能，一键纳管VMware平台或将其他虚拟化平台的云主机及数据完整迁移至当前超融合内，异构平台业务极简上云。

高效

- 高效性能：NUMA绑定、大页内存、智能Cache、IO优化、增强LibRBD显著提升超融合计算存储性能。
- 统一运维管理界面：资源页面可视化监管，告警页面多维度分级展示，管理更加精细化。

可靠

- 全冗余架构：支持硬件系统冗余、多链路冗余和计算/存储高可用，无单点故障。
- 灵活数据策略：支持多副本、EC纠删码和自定义故障域等方式来对数据进行保护，可容忍一定量的数据丢失，让数据更可靠。

安全

- 丰富的网络安全能力：支持网络隔离、安全组、防火墙、Netflow、端口镜像等多种网络防护方式，全方位保证网络安全。
- 精细化运维安全能力：涵盖权限、计费、操作审计、双因子认证、无缝升级等特色安全功能，高效运维。

3 产品架构

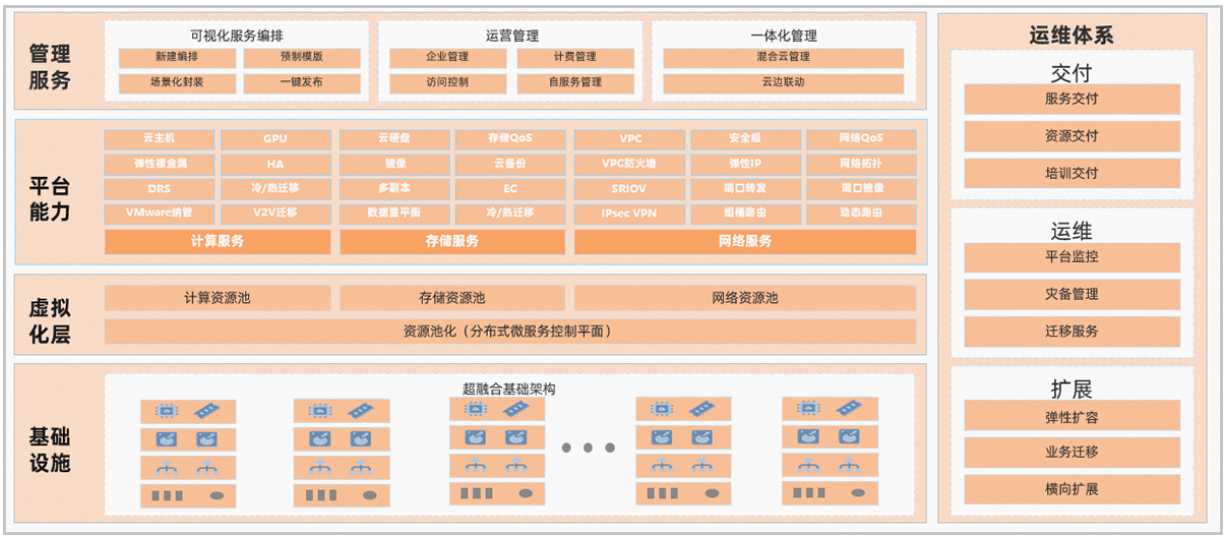
3.1 系统架构

系统总体架构

AliyunHCI-Z架构主要由两部分组成：超融合基础设施、云计算管理平台。在企业级云架构之上，用户可以基于自身需求构建业务系统。

如图 *3-1:* 系统总体架构所示：

图 3-1: 系统总体架构



管理服务层架构

AliyunHCI-Z管理服务层架构特点：

1. 全异步架构：异步消息、异步方法、异步HTTP调用。
   - 使用消息总线进行各服务的通信连接，在调用服务时，源服务发消息给目的服务，并注册一个回调函数，然后立即返回；一旦目的服务完成任务，就会触发回调函数回复任务结果。异步消息可以并行处理。
   - 服务之间采用异步消息进行通信，对于服务内部，一系列相关组件或插件，也是通过异步方法来调用，调用方法与异步消息一致。
   - 采用的插件机制，给每个插件设置相应的代理程序。为每个请求设置了回调URL在HTTP的包头，任务结束后，代理程序会发送应答给调用者的URL。

- 基于异步消息、异步方法、异步HTTP调用这三种方式，构建了一个分层架构，保证了所有组件均能实现异步操作。
- 基于全异步架构机制，单管理节点的每秒可并发处理上万条API请求，还可同时管理上万台服务器和数十万台云主机。

2. 无状态服务：单次请求不依赖其他请求。

- 计算节点代理、存储代理、网络服务、控制台代理服务、配置服务等，均不依赖其他请求，一次请求可包含所有信息，相关节点无须维护存储任何信息。
- 使用一致性哈希环对管理节点、计算节点或者其他资源以UUID为唯一ID进行认证的哈希环处理，消息发送者无需知道待处理消息的服务实例，服务也无须维护、交换相关的资源信息，服务只需单纯的处理消息即可。
- 管理节点间共享的信息非常少，两个管理节点即可满足高可用性和可扩展性需求。
- 无状态服务机制让系统更为健壮，重启服务器不会丢失任何状态信息，数据中心的弹性扩展和伸缩性维护更为简单。

3. 无锁架构：一致性哈希算法。

- 一致性哈希算法保证了同一资源的所有消息均被同一个服务实例来处理。这种聚合消息到特定节点的方法，降低了同步与并行的复杂度。
- 使用工作队列来避免竞争锁的问题，串行任务以工作队列的方式保存在内存中，工作队列可对任意资源的任意操作进行并行处理来提高系统并行度。
- 基于队列的无锁架构，使得任务可以简单地控制并行度，从而提升系统性能。

4. 进程内微服务：微服务解耦。

- 使用消息总线对各服务进行隔离控制，例如，云主机服务、身份认证服务、快照服务、云盘服务、网络服务、存储服务等。所有的微服务都集合在管理节点的进程内，各服务之间利用消息总线进行交互，所有消息发送到消息总线后，再通过一致性哈希环选择目的服务进行转发处理。
- 进程内微服务，以星状架构实现各服务独立运行，将高度集中的控制业务进行解耦，实现了系统的高度自治和高度隔离，任何服务出现故障并不影响其他组件。可靠性与稳定性得到有效保障。

5. 全插件结构：插件支持横向扩展。

- 使用中任何新加入的插件对目前其他的插件没有任何影响， 均是独立自主提供服务。

- 支持策略模式和观察者模式进行插件设计。策略插件会继承父类的接口然后执行具体实现；观察者插件，会注册listener进行监控内部的业务逻辑的事件变化，当应用内部发现事件时，插件会对此事件做出自响应，在插件自身的代码里执行相应的业务流。
- 支持插件的横向扩展，云平台可以快速更迭，而整体系统架构依然健壮。

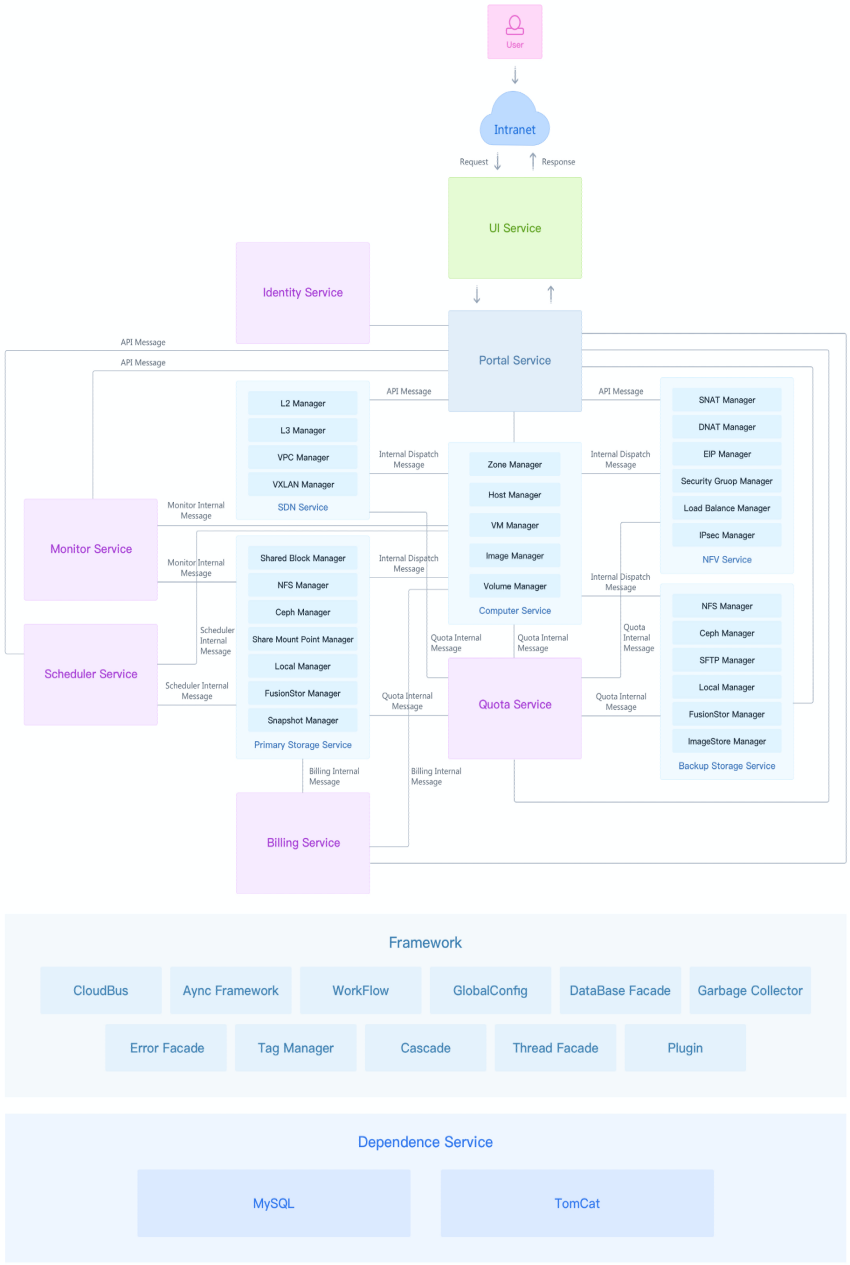
6. 工作流引擎：顺序管理，出错回滚。
   - 工作流基于XML对每个工作流程进行清晰定义，在任何步骤出现错误均可按照原本执行路径进行回滚，清理掉执行过程的垃圾资源。
   - 每个工作流还可以包含子工作流用于扩展业务逻辑。
7. 标签系统：支持业务逻辑变更，增加资源属性。
   - 支持利用系统标签和插件机制对原本的业务逻辑进行扩展变更。
   - 使用标签机制，可对资源进行分组划分，支持对指定标签进行资源搜索。
8. 瀑布流架构：支持资源的级联操作。
   - 使用Cascade Framework对资源管理进行瀑布状的级联操作，对资源进行卸载或者删除时，会对相关的资源进行级联操作。
   - 资源也可以通过插件形式加入到瀑布框架中，加入或者退出瀑布框架，并不影响其他资源。
   - 级联机制使得的配置灵活轻便，快速满足客户资源配置的变更。
9. 全自动化Ansible部署：Ansible无代理自动部署。

   使用Ansible进行无代理的全自动化安装依赖、配置物理资源，部署代理程序，全过程对用户透明，无须额外干预，可透过重连代理程序对代理进行升级。
10. 全API查询：

    任意资源的任意属性均可查询。支持数百万个条件的资源查询，支持全API查询，支持任意组合。

如图 *3-2:* 管理服务层架构所示：

图 3-2: 管理服务层架构



存储层架构

AliyunHCI-Z存储层架构特点：

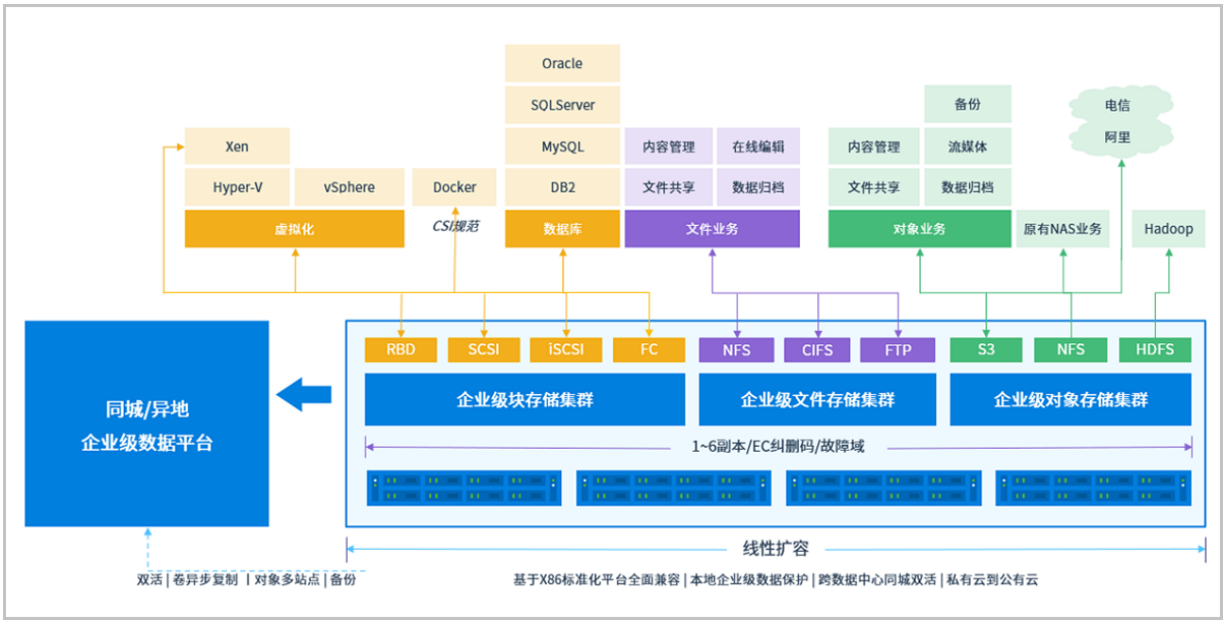
1. 云原生，无缝对接云平台。
   - 可全协议、100%无缝替代社区版Ceph。
   - 基于主流的开源分布式存储系统，在保留云原生RBD协议的基础上，对协议层进行了优化重构，为IO的持续性提供了更高的技术支撑。

- 丰富的RESTful API接口，可以和多种云平台无缝对接。
- 支持跨集群云主机和存储在线迁移。

2. IO路径多维度优化，发挥硬件潜能，持续提升系统性能。
   - 针对系统IO栈进行了深度代码优化，包括网络和磁盘处理效率优化、数据分层和缓存机制优化等，在应对高并发、高输入输出效率的需求时更加流畅，延迟更低。
   - 支持高性能多级Cache加速引擎，包括RAM和SSD两级Cache，其中Ram Read Cache采用自主研发的增强型预读算法，提升系统读性能；SSD R/W Cache提供智能IO合并算法和热点数据分析算法，大大提升业务系统读写性能。
   - 自研技术的librbd汇聚代理，使单客户端性能提升20%，每节点CPU利用率降低40%。
3. 企业级存储功能，完善的数据保护功能，满足容灾备份需求。
   - 支持IO级别数据校验，防止磁盘静默错误。
   - 在数据存储过程中，节点故障、硬盘损坏等会自动触发系统数据重平衡操作，用户可在数据快速恢复和业务性能优先之间，根据自己的需求进行自定义，避免给前端业务带来性能冲击。
   - 即时无损ROW快照、克隆独立卷、延展集群等高级功能，适应企业存储管理的各种容灾场景，分布式存储提供了从磁盘级别到方案级别的数据保护及容灾恢复处理途径，用户可按需自由选择。
   - 对关键IO路径及管理控制模块进行了冗余设计，管理模块采用一主两备的方式，可实现信息实时同步、故障自动切换，保障管理功能的高可用。
   - 通过拓扑规划功能对存储集群进行多种安全级别的数据安全保护，如支持服务器级别、机架级别、数据中心级别的故障域，有效保证存储系统可靠性及持续在线特性。
   - 支持硬盘和网络亚健康处理：自动识别坏盘和慢盘并自动隔离，自动识别集群网络故障并告警通知，协助管理员快速定位和解决集群故障。
   - S3对象存储功能选件，可靠的镜像仓库管理。
4. 敏捷交付，简化运维。
   - 向导式安装部署，快速扩容升级，AI辅助运维，满足客户统一管理的需求。
   - 100%可视化管理界面操作，产品管理界面简洁明了。客户的任何操作都可通过GUI界面完成，一键安装部署，快速扩容，无需关注硬件的复杂性，提升运维效率。
   - CLI、日志及分级告警功能能够协助用户迅速发现并定位存储集群各类突发状况，降低了日常运维的复杂度。
   - 支持自助巡检工具，对集群健康状况进行智能分析，提前识别系统潜在隐患。

如图 *3-3: 存储层架构*所示：

图 3-3: 存储层架构



3.2 核心设计

3.2.1 资源虚拟化

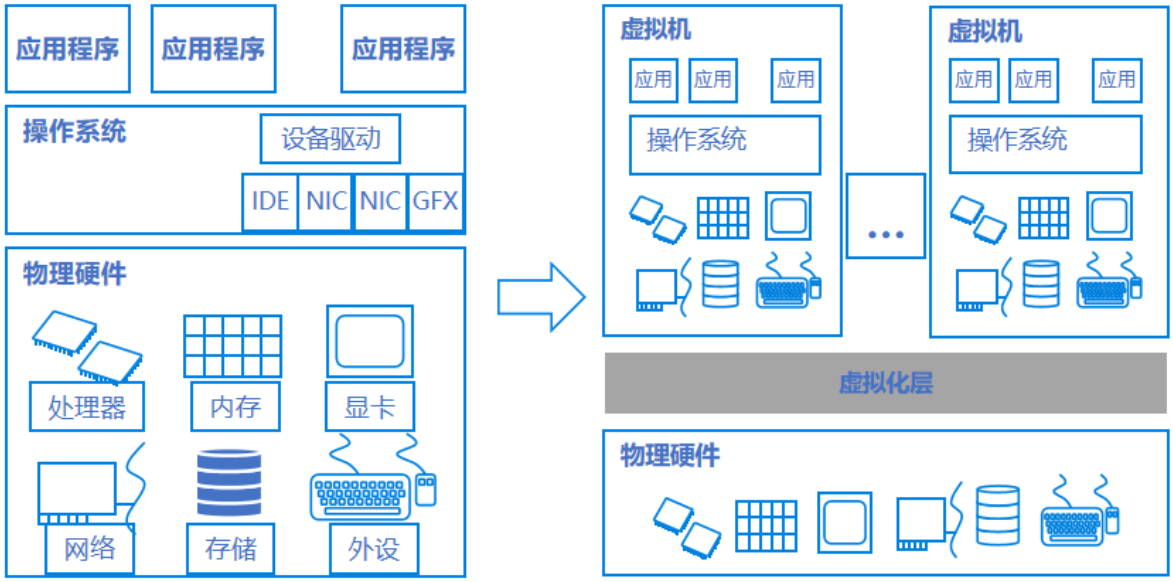
3.2.1.1 计算虚拟化

3.2.1.1.1 概述

计算虚拟化是将物理服务器资源通过虚拟化技术抽象成逻辑资源，让一台物理服务器变成多台相互隔离的虚拟服务器，CPU、内存、磁盘、I/O设备等硬件资源变成虚拟化资源池进行统一动态管理，从而提高资源利用率，降低系统管理成本，让IT对业务变化更具适应力。

如图 *3-4: 服务器虚拟化*所示：

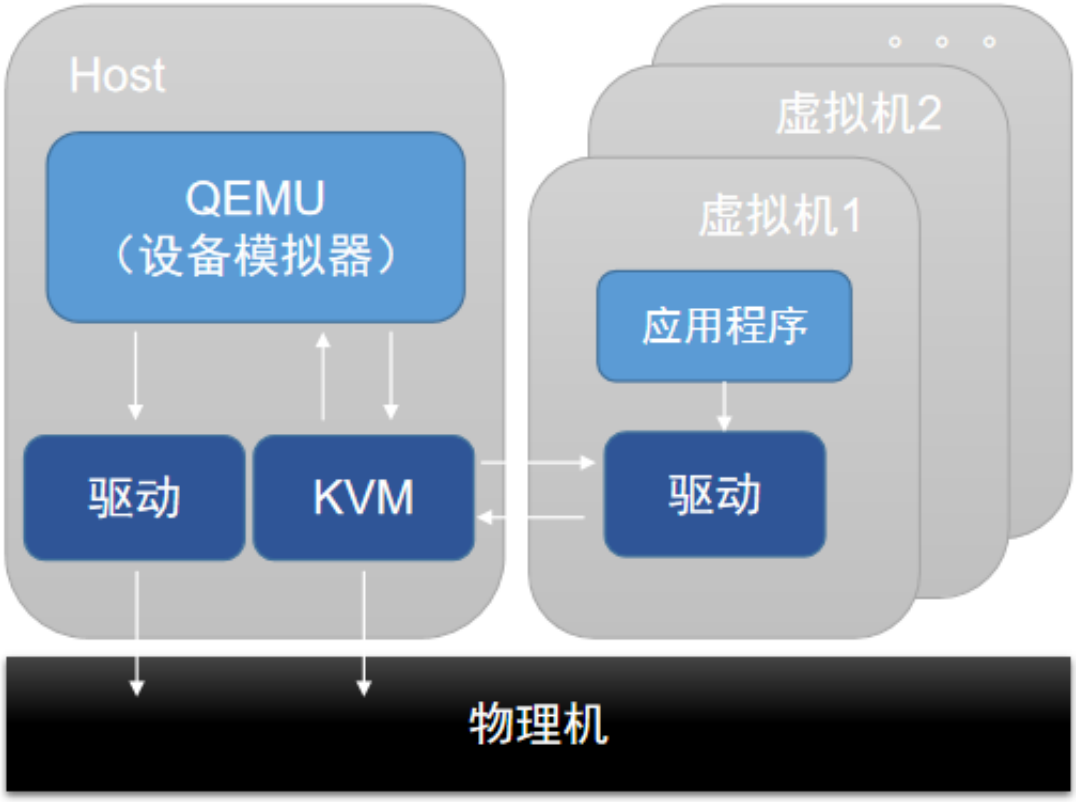
图 3-4: 服务器虚拟化



AliyunHCI-Z采用基于KVM的硬件虚拟化技术。KVM是一个Linux内核模块，将Linux内核变成一个Hypervisor。KVM在Linux系统内以进程形式出现，由标准Linux调度程序进行调度，因此KVM能够使用Linux内核已有功能，例如：内存管理、CPU调度等。但是，KVM本身仅提供CPU与内存虚拟化，I/O设备虚拟化需要结合Qemu才能完成。Qemu是一个用户态的设备模拟器，为云主机提供虚拟设备模型，负责各种虚拟设备的创建、调用及管理。

如图 *3-5: KVM虚拟化技术*所示：

图 3-5: KVM虚拟化技术



3.2.1.1.2 技术特性

3.2.1.1.2.1 CPU虚拟化

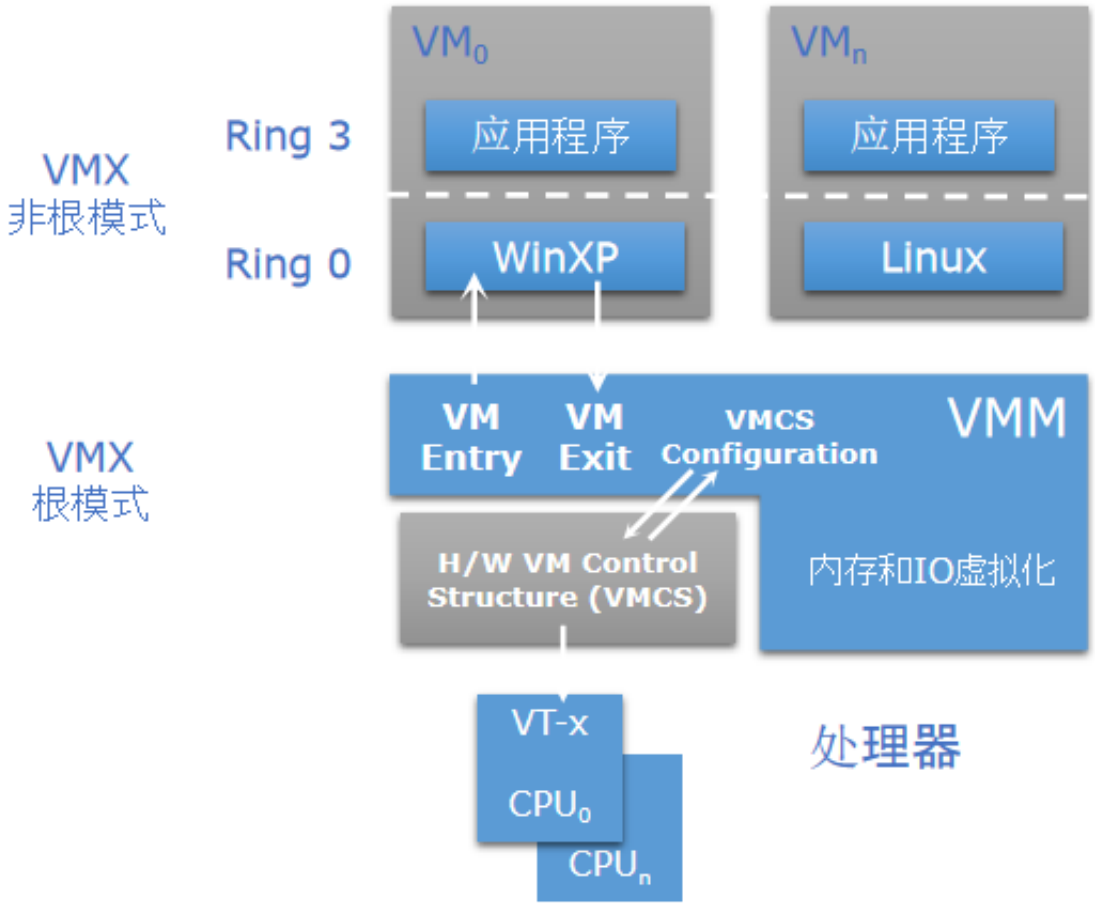
在x86体系架构上，CPU一般有4个特权级别：ring0~ring3，用于给操作系统和应用程序访问硬件。在Linux中，仅使用其中2个特权级别：ring0（内核态）、ring3（用户态）。

VMX根模式与VMX非根模式。对于硬件辅助虚拟化而言，为能在不对操作系统做任何修改的前提下使用云主机，CPU引入2种运行模式：VMX根模式、VMX非根模式。宿主机运行在根模式下，宿主机内核处于ring0，用户态程序处于ring3。云主机运行在非根模式下，云主机内核处于ring0，用户态程序处于ring3。

VM Exit与VM Entry。处于非根模式的云主机，当外部中断或缺页异常，或主动调用VMCALL指令来调用VMM服务时，CPU会从非根模式切换至根模式，整个过程称为VM Exit。相反，当VMM通过显式调用VMLAUNCH或VMRESUME指令切换至非根模式时，硬件自动加载云主机上下文，运行云主机指令，这一转换称为VM Entry。

如图 *3-6:* 非根模式与根模式所示：

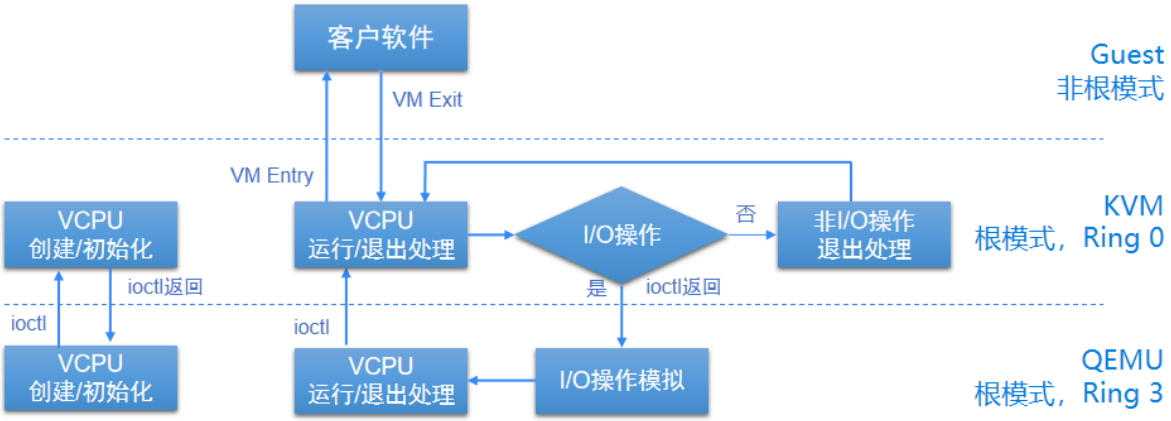
图 3-6: 非根模式与根模式



当云主机通过VM Exit从非根模式退出至根模式后，KVM会根据退出原因执行进一步操作，若是I/O操作则交由Qemu处理，若是非I/O操作则由KVM自行处理，处理完成后会通过VM Entry再次切回至云主机非根模式下运行。

如图 *3-7:* 模式切换所示：

图 3-7: 模式切换



3.2.1.1.2.2 内存虚拟化

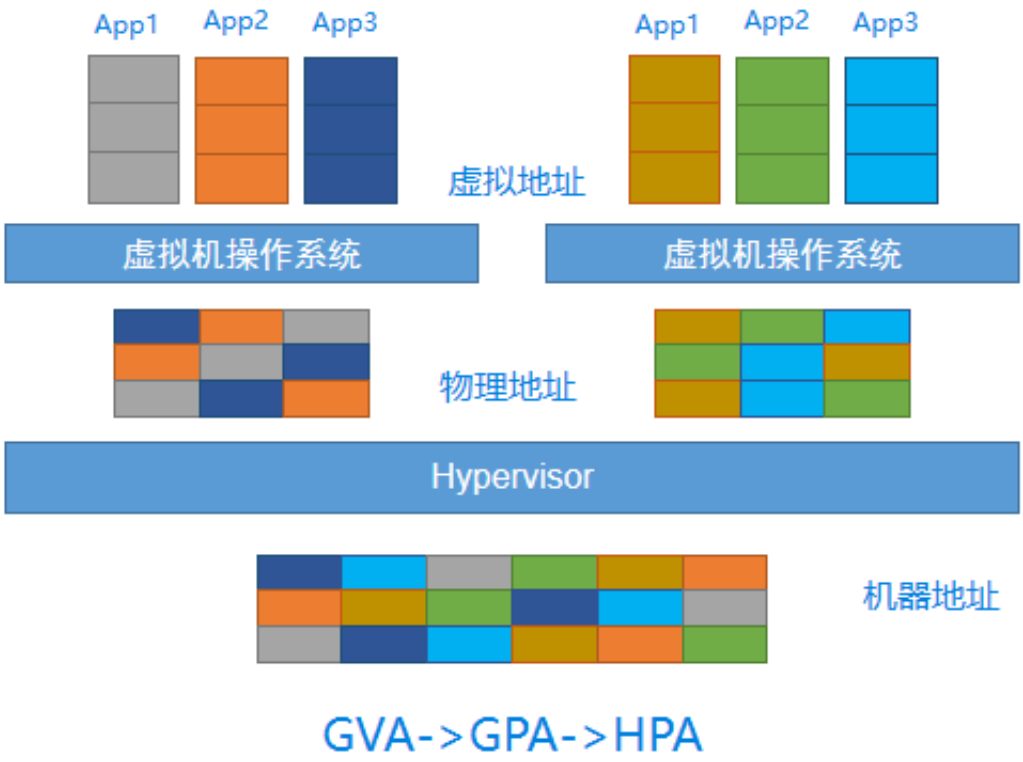
VMM负责管理和分配每个云主机的物理内存。云主机操作系统看到的是一个虚构的云主机物理地址空间，操作系统的内存管理模块负责将云主机虚拟地址（GVA）映射到云主机物理地址（GPA），其指令目标地址也是一个云主机物理地址。这样的地址在无虚拟化情况下，其实是实际物理地址。但在有虚拟化情况下，这样的地址不能被直接处理使用，需VMM先将云主机物理地址转换成一个物理机物理地址（HPA），再交由物理处理器执行。

由于引入云主机物理地址空间，内存虚拟化主要处理以下两方面问题：

- 维护云主机物理地址到宿主机物理地址之间的映射关系。
- 当云主机访问云主机物理地址时，根据映射关系，将其转换成宿主机物理地址。

如图 *3-8: GVA->GPA->HPA*所示：

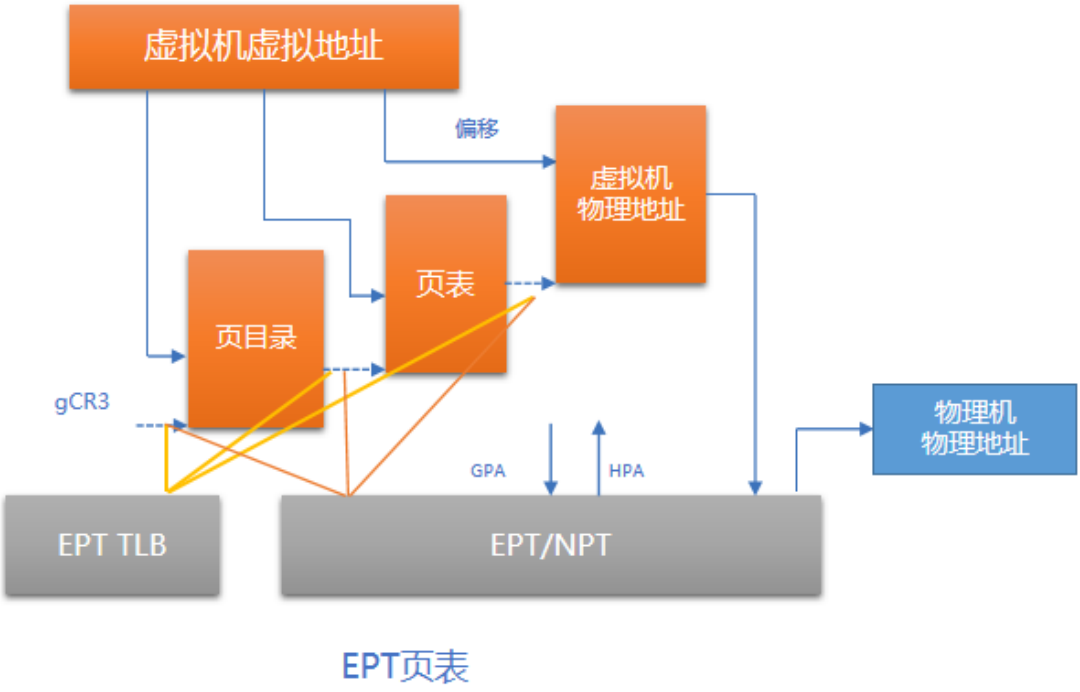
图 3-8: GVA->GPA->HPA



内存的硬件辅助虚拟化使用扩展页表技术，通过硬件完成云主机虚拟地址到物理机物理地址的转换。

如图 *3-9: EPT*页表所示：

图 3-9: EPT页表



3.2.1.1.2.3 设备虚拟化

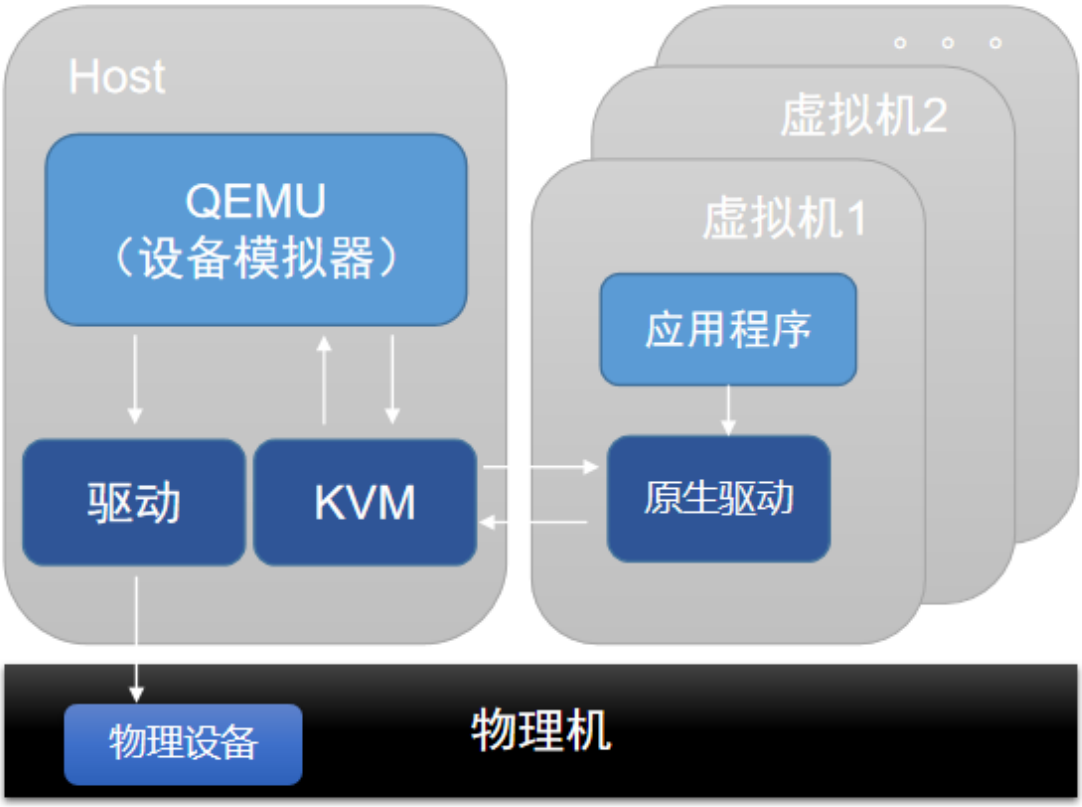
设备虚拟化方式主要有三种：设备模拟、半虚拟化设备、设备直通。

设备模拟

设备模拟是通过Qemu提供的设备模型，可完全模拟出与物理设备一样的接口。因此，在云主机操作系统中，使用原生驱动即可使用设备。设备模拟只能模拟出具有基本功能的设备，不支持复杂功能和模型的设备。完全模拟的设备兼容性好，但由于是纯软件模拟，性能相对较低。

如图 *3-10:* 设备模拟所示：

图 3-10: 设备模拟



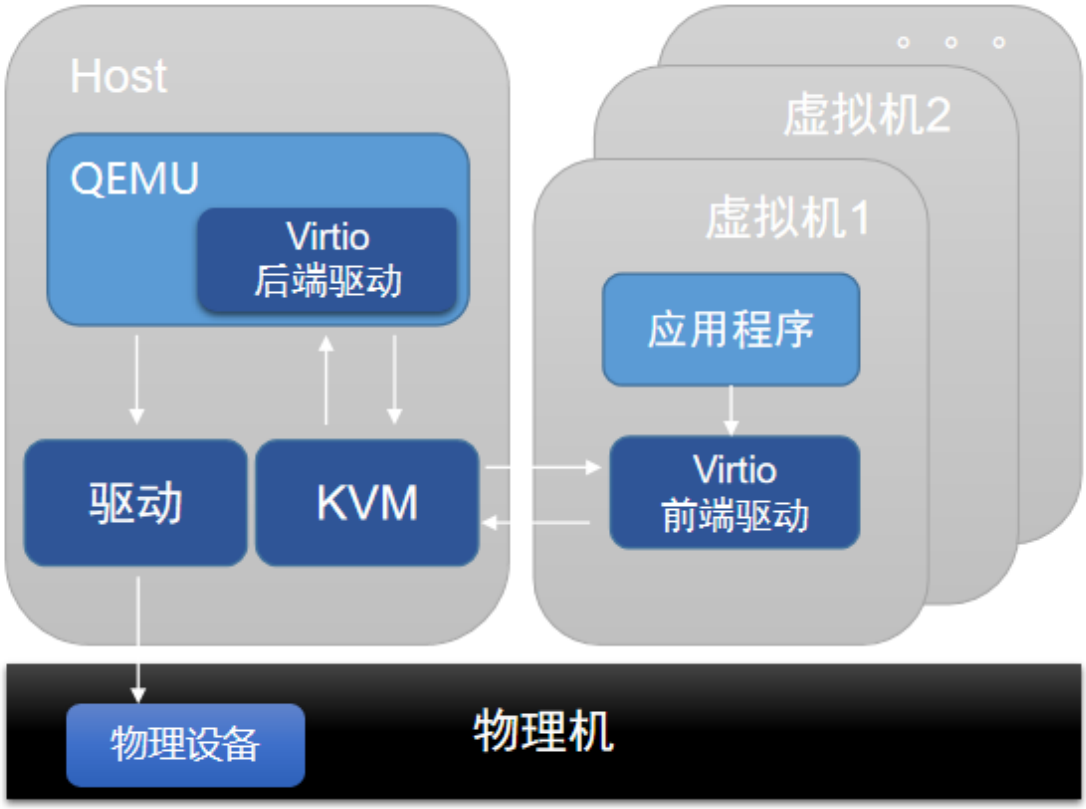
半虚拟化设备

半虚拟化设备实现前后端驱动。利用云主机中的前端驱动，通过基于事务的通信机制，将请求直接发给宿主机端的后端驱动，从而很大程度上减少上下文切换的开销，性能相比完全设备模拟有较大提升。然而，Virtio后端驱动仍在Qemu中实现，在IO处理过程中会经过用户态与内核态之间的多次切换。为进一步提升性能，可将Virtio后端驱动的功能放至内核态实现，称为vhost-kernel后端，于是数据仅需经过从用户态到内核态的一次切换，就可完成数据传输，实现性能提升。

随着技术发展，将数据放入用户态处理可得到更灵活的形式。因此，在原有vhost架构中进行改动，增加vhost-user后端，搭配DPDK、SPDK中相关用户态函数库，性能进一步提升。

如图 *3-11:* 半虚拟化设备所示：

图 3-11: 半虚拟化设备

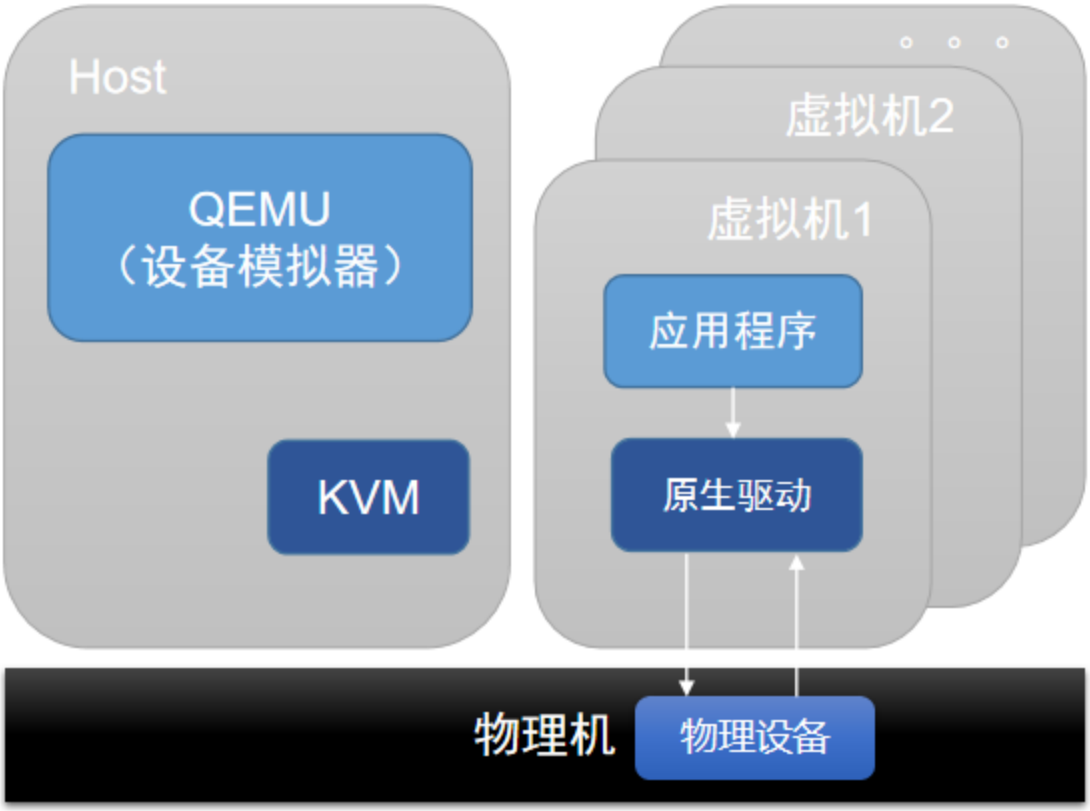


设备直通

设备透传基于硬件的设备虚拟化技术，支持将PCI/PCIe物理设备直接映射到云主机的地址空间，在云主机中，使用原生设备驱动就可直接使用设备，达到近乎物理设备的性能。物理设备被透传后被云主机独享，其它云主机无法共享使用该设备。

如图 *3-12:* 设备透传所示：

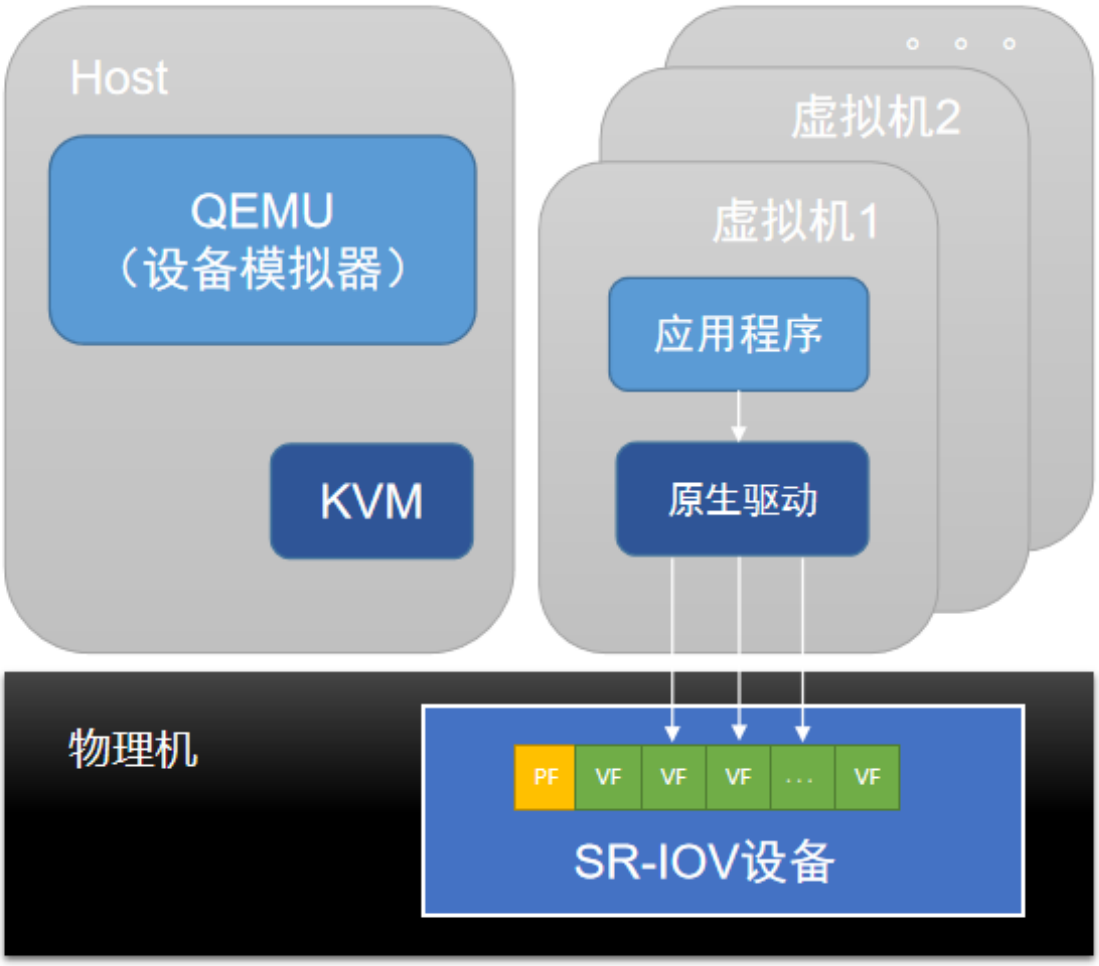
图 3-12: 设备透传



SR-IOV是由PCI-SIG组织定义的PCIe规范的扩展规范，目的是通过提供一种标准规范，为云主机提供独立的内存空间、中断、DMA数据流。 SR-IOV支持单个物理PCIe设备（PF）虚拟出多个虚拟PCIe设备（VF），然后通过设备透传技术将虚拟PCIe设备直通到各云主机，以实现单个物理PCIe设备支撑多云主机的应用场景。

如图 *3-13: SR-IOV*所示：

图 3-13: SR-IOV



3.2.1.2 存储虚拟化

3.2.1.2.1 概述

业界典型的分布式存储技术主要有：分布式文件存储、分布式对象存储和分布式块存储。分布式存储具备高性能、高可靠、高扩展、易管理等特点。

AliyunHCI-Z通过存储虚拟化技术将服务器存储资源池化，实现服务器存储资源的统一整合、管理及调度，并向上层提供多种存储接口，供云主机根据自身的存储需求灵活分配使用资源池中的存储空间。整体存储逻辑主要包含以下6层：

- 应用层：一套存储软件通过块存储接口、对象存储接口、文件存储接口，支撑了行业云、私有云、桌面云、数据库资源池、海量媒体数据、影像数据、智能制造数据等不同类型的应用场景。

- 分布式网关层：块存储通过RBD/SCSI/iSCSI/FC 驱动接口向操作系统、数据库提供卷设备；对象存储通过标准的S3接口对接海量的非结构化数据；文件存储通过标准的NFS、CIFS/SMB、FTP协议可以提供文件共享服务。
- 存储服务层：提供各种存储高级特性，如ROW快照、克隆、精简配置、数据链路HA、目录快照、小文件归并、生命周期管理、整池扩容、分级存储；各种资源管理功能，如多副本、EC纠删码、多资源池管理、恢复QoS、卷级QoS、主副本位置、多级故障域、亚健康等；系统加速功能，如智能SSD缓存、大IO透传、热卷锁定等。
- 数据服务层：提供各种数据保护及数据生命周期功能，包括延展集群、卷备份、云备份、卷异步复制、对象多站点、桶云端归档、资源恢复、小文件归并及对象延时删除等功能。
- 存储引擎层：存储基本功能，包括集群状态控制、数据智能路由、强一致性协议、集群自检自愈、数据并行恢复及Block持久化等。
- 硬件设备层：以通用x86服务器为基础硬件平台，支持异构服务器。
  - 硬盘方面：兼容各类SSD、SAS以及SATA硬盘，包括PCIe SSD。
  - 网络访问：兼容主流千兆、万兆网卡、交换机，以及IB设备。
  - 接口方面：兼容主流网卡和光纤卡。
  - 异构方面：支持将不同类型、配置的服务器放置在同一存储系统中，共同使用，解决企业利旧问题，提升资源使用率。

3.2.1.2.2 技术特性

3.2.1.2.2.1 自动精简配置

支持为应用提供比实际物理存储更多的虚拟存储资源，相比直接分配物理存储资源，可以显著提高存储空间利用率。采用CRUSH数据路由技术，系统无需使用专门的集中元数据来记录卷的精简分配情况，和传统SAN相比，不会带来性能下降。

3.2.1.2.2.2 ROW无损快照

采用ROW快照技术，极大改善快照后的存储性能，支持手动快照和定时快照：

- 手动快照：应用于用户在当前时间点对某一指定卷手动创建快照。
- 定时快照：针对不同卷设置不同的快照时间策略，周期性自动创建快照。可按照**时间段**或**时间点**方式定时自动创建快照策略。

快照适用场景：

- 可利用快照技术，进行快照克隆、快照复制等一系列操作。

- 使用快照技术可以在以下场景中迅速恢复数据：
  - 病毒感染
  - 人为误操作
  - 恶意篡改
  - 系统宕机造成的数据损坏
  - 应用程序BUG造成的数据损坏

3.2.1.2.2.3 一致性快照

块存储服务支持一致性组，将同一业务场景中的多个块存储卷在同一个时间点创建的快照集合。首先需要将块存储卷分配到一致性组中，以确保在同一个时间点为该组中的所有卷创建快照，从而实现数据的一致性。

3.2.1.2.2.4 链接克隆

支持基于一个卷快照或文件系统快照创建出多个克隆卷或克隆文件系统，各个克隆卷或文件系统刚创建出来时的数据内容与快照中的数据内容一致，后续对于克隆卷或文件系统的修改不会影响到原始的快照和其他克隆资源。

克隆卷或文件系统继承普通卷和文件系统的所有功能：克隆出的资源也可支持创建快照、从快照恢复以及再次作为母资源进行克隆操作。

3.2.1.2.2.5 多资源池

为了满足使用不同性能存储介质以及故障隔离，支持多资源池特性。不同资源池可提供不同性能以及不同副本策略。

3.2.1.2.2.6 企业级QoS

支持在线设置存储卷的最大/突发IOPS、最大/突发带宽，用户通过设置业务QoS，实现不同存储卷性能的量化管控，进而应对多样化业务对性能的需求。

支持采用漏桶（Leaky bucket）和令牌桶（Token Bucket）相结合的IO流管理策略。漏桶算法能够强行限制数据的传输速率，令牌桶算法能够在限制平均传输速率的同时允许一定的突发传输。

- 漏桶算法：IO进入存储队列如同水进入到漏桶里，桶里的水通过下面的孔以固定的速率流出。漏桶算法能强行限制数据的传输速率。

- 令牌桶算法：系统会以一个恒定的速度往桶里放入令牌，如果请求需要被处理，则需要先从桶里获取一个令牌，当桶里没有令牌可取时，则拒绝服务。一旦需要提高速率，则按需提高放入桶中的令牌的速率。

漏桶算法对于存在突发特性的流量来说缺乏效率，不能够有效地使用网络资源，而令牌桶算法则能够有效支持具有突发特性的流量。采用漏桶与令牌桶相结合的I/O流管理策略，并对两种算法进行优化：漏桶无上限，避免了漏桶算法的/IO溢出、漏桶丢包的现象；令牌桶具有令牌租借、归还策略，提升了整体性能。

3.2.1.2.2.7 纳管卷

在本地创建虚拟卷，元数据保存在本地，只将IO下发到被纳管卷上。应用可以直接使用纳管卷读写IO，以达到异构存储的目的。纳管卷允许读写 IO，可以在线迁移，支持基本的管理面功能，不支持扩容、缩容和快照操作。

3.2.1.2.2.8 在线卷迁移

在线将卷从同一个集群的一个池迁移到另一个池。在迁移过程中，保证业务IO不中断；迁移完成后，源卷自动删除，应用端无需做任何修改，真正做到应用无感知。

3.2.1.2.2.9 卷回收站

默认支持卷回收站功能，防止由于误操作导致数据丢失，增强对数据的保护能力。默认情况下，删除的块存储卷将进入回收站进行管理，等待一段时间（默认30天）后将自动删除。

进入回收站的卷可以进行手动还原或者永久删除。

同时支持托管卷和非托管卷的回收站功能。非托管卷默认启用回收站功能，而托管卷则默认不启用。

3.2.1.3 网络虚拟化

3.2.1.3.1 概述

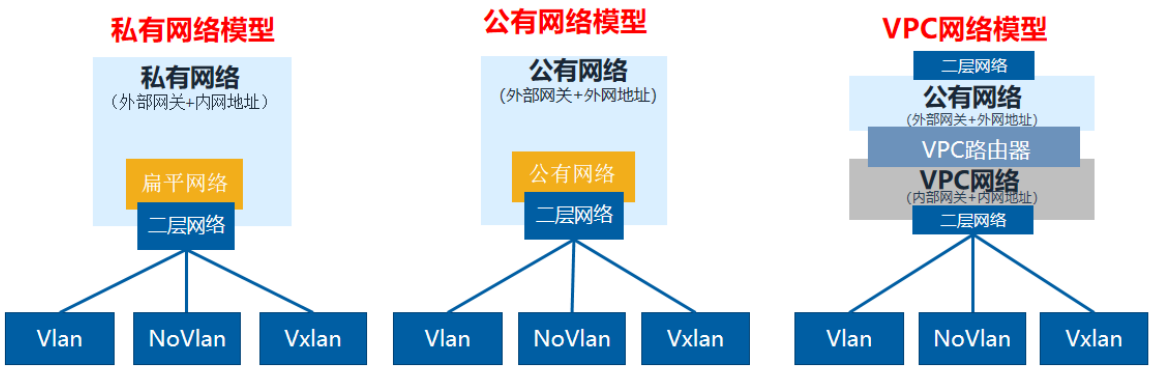
网络虚拟化是一种脱离专用网络硬件抽象出的虚拟网络技术。底层硬件仅需提供基础的数据包转发服务，网络虚拟化可提供多种网络服务，包括交换、路由、安全组和防火墙等，使网络体验同物理网络一样。

AliyunHCI-Z将网络模型抽象为二层网络和三层网络。二层网络对应于二层广播域，提供一种二层网络隔离的方式。三层网络主要与OSI七层模型中第4层~第7层网络服务相对应。可提供二层隔离技术的NoVLAN、VLAN、VXLAN、SDN等均可作为二层网络。创建二层网络，相当于在二层网络所挂

载的集群内所有物理机上，创建对应的虚拟交换机来提供广播域。在二层网络之上，创建的三层网络类型包括：扁平网络、公有网络、VPC网络。基于三层网络可提供各种网络服务，包括：DHCP、DNS、弹性IP、端口转发、负载均衡等。

如图 *3-14:* 二层网络与三层网络所示：

图 3-14: 二层网络与三层网络



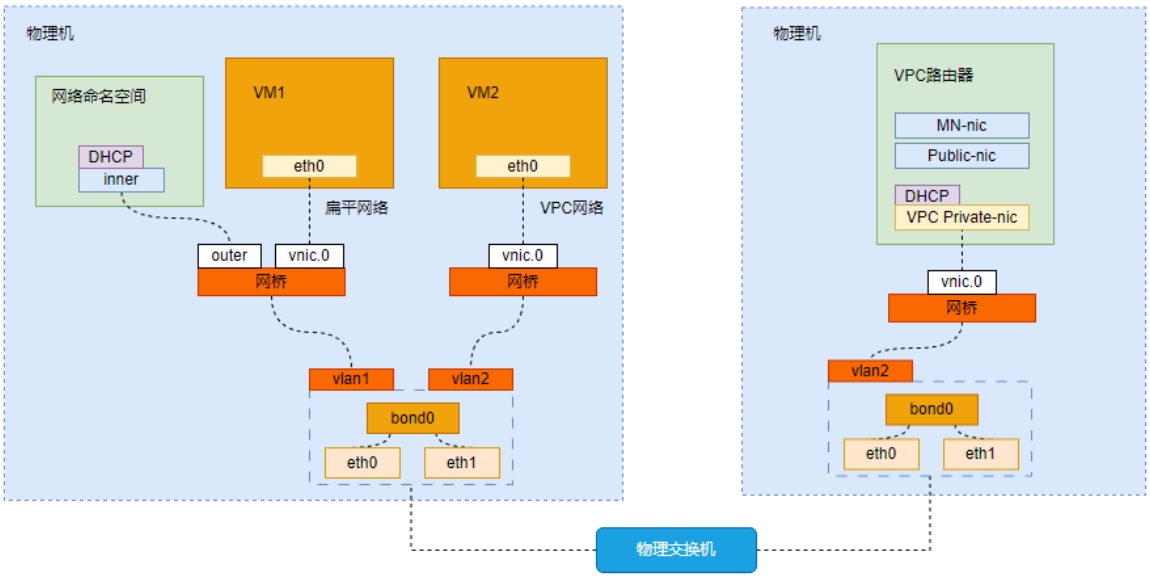
3.2.1.3.2 技术特性

3.2.1.3.2.1 三层网络

三层网络包括：扁平网络、公有网络、VPC网络。扁平网络可给云主机分配私有网络地址，同时云主机可通过分布式EIP访问公有网络。扁平网络支持DHCP、User Data、弹性IP、安全组、端口镜像等网络服务。VPC网络是一块可由租户自定义的网络空间，其目的是让租户在云平台上构建出一个隔离的、可自行管理配置及策略的虚拟网络环境，从而进一步提升租户在云环境中的资源安全性。VPC网络和服务由VPC路由器提供，一个VPC路由器下可提供多个相互隔离的VPC网络，给云主机提供DHCP、DNS、SNAT、路由表、弹性IP、端口转发、负载均衡、IPsec隧道、安全组、动态路由、组播路由、VPC防火墙、Netflow等网络服务。

如图 *3-15:* 三层网络所示：

图 3-15: 三层网络

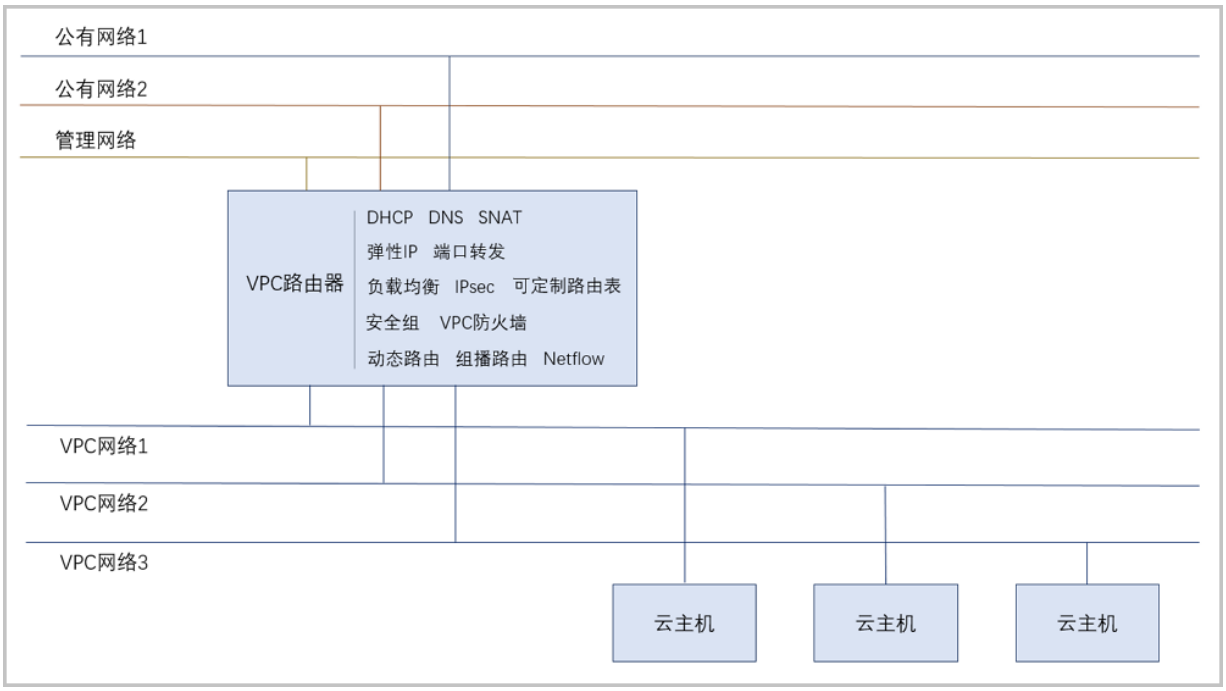


3.2.1.3.2.2 VPC路由器

VPC路由器是一种专用的云主机，运行着定制的Linux操作系统，以及管理服务代理程序。一个VPC路由器下可提供多个相互隔离的VPC网络，每个VPC路由器中包含一个管理服务代理程序，通过HTTP协议接收来自管理节点的命令来配置网络服务，可为云主机提供DHCP、DNS、SNAT、路由表、弹性IP、端口转发、负载均衡、IPsec隧道、安全组、动态路由、组播路由、VPC防火墙、Netflow等网络服务。

如图 *3-16: VPC路由器*所示：

图 3-16: VPC路由器



3.2.1.3.2.3 负载均衡

负载均衡提供各种灵活分配算法将全部网络请求均衡分布至后端服务器组上，通过合理管理流量分发以减轻单个服务器的负担，从而应对大流量、高并发的访问，满足客户业务场景需求。

负载均衡同时支持四层负载均衡协议（TCP/UDP）和七层负载均衡协议（HTTP/HTTPS）。

负载均衡转发前端流量至后端服务器时，支持以下算法：

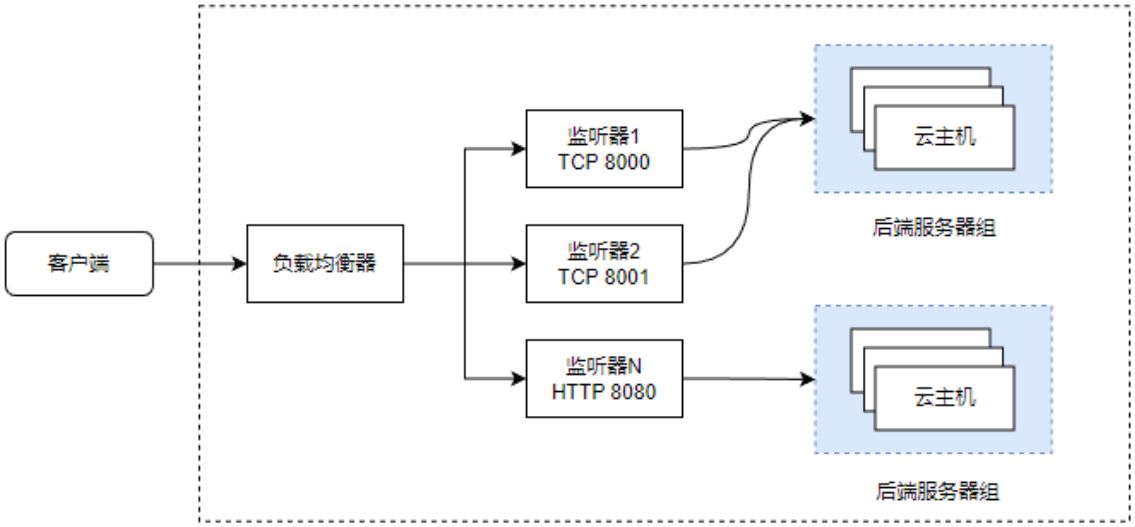
- 轮询算法：按照顺序轮流分配访问请求至后端服务器。轮询是最简单的一个算法，无须关注后端服务器本身的连接数和系统负载等状态，主要应用于各个后端服务器性能差异不大的场景。
- 加权轮询算法：根据后端服务器权重转发访问请求。一般情况下，权重基于硬件配置进行设置，为静态值。权重值越高，被轮询的次数（概率）越高。加权轮询是轮询的一种特殊形式，主要应用于各个后端服务器性能差异较大的场景。
- 源地址哈希：使用客户端请求的源 IP 地址与目标 IP 地址生成唯一的哈希密钥，将请求分配给特定的后端服务器，适合后端服务器需处理客户端请求差异较大的场景。
- 最小连接算法：将新的连接请求分配到当前连接数最小的后端服务器，适用于请求占用后端服务器时间相差较大的场景，常用于长连接服务。

AliyunHCI-Z支持基于TCP/UDP协议的四层会话保持机制，以及基于HTTP/HTTPS协议的七层会话保持机制。可识别客户端与后端服务器之间的交互关联性，将客户端访问请求定向转发至特定的后端服务器，从而保证业务会话连续性。

- 四层会话保持机制：负载均衡将同一个源IP地址的访问请求都转发至一台后端服务器上。
- 七层会话保持机制：不同负载均衡算法下，七层会话保持机制不同。轮询算法或加权轮询算法使用基于Cookie的会话保持机制，负载均衡可通过Cookie将访问请求定向转发至之前记录的后端服务器。源地址哈希算法通过哈希函数计算客户端源IP地址，同一个源IP地址的访问请求都将转发至一台后端服务器上。

如图 *3-17:* 负载均衡所示：

图 3-17: 负载均衡



3.2.1.4 虚拟资源管理

3.2.1.4.1 云主机调度策略

云主机调度策略可为云主机分配物理机资源编排策略，用于保障业务高性能和高可用。支持将云主机加入一个云主机调度组，通过为该云主机调度组绑定调度策略实现云主机调度。

功能原理

支持将云主机加入一个云主机调度组，通过为该云主机调度组绑定调度策略实现云主机调度。

- 若绑定互斥云主机或聚集云主机策略，无需指定物理机调度组，云主机按照策略及其执行机制分配物理机。
- 若绑定云主机亲和物理机或云主机互斥物理机调度策略，需指定对应的物理机调度组，云主机按照策略及其执行机制分配物理机。

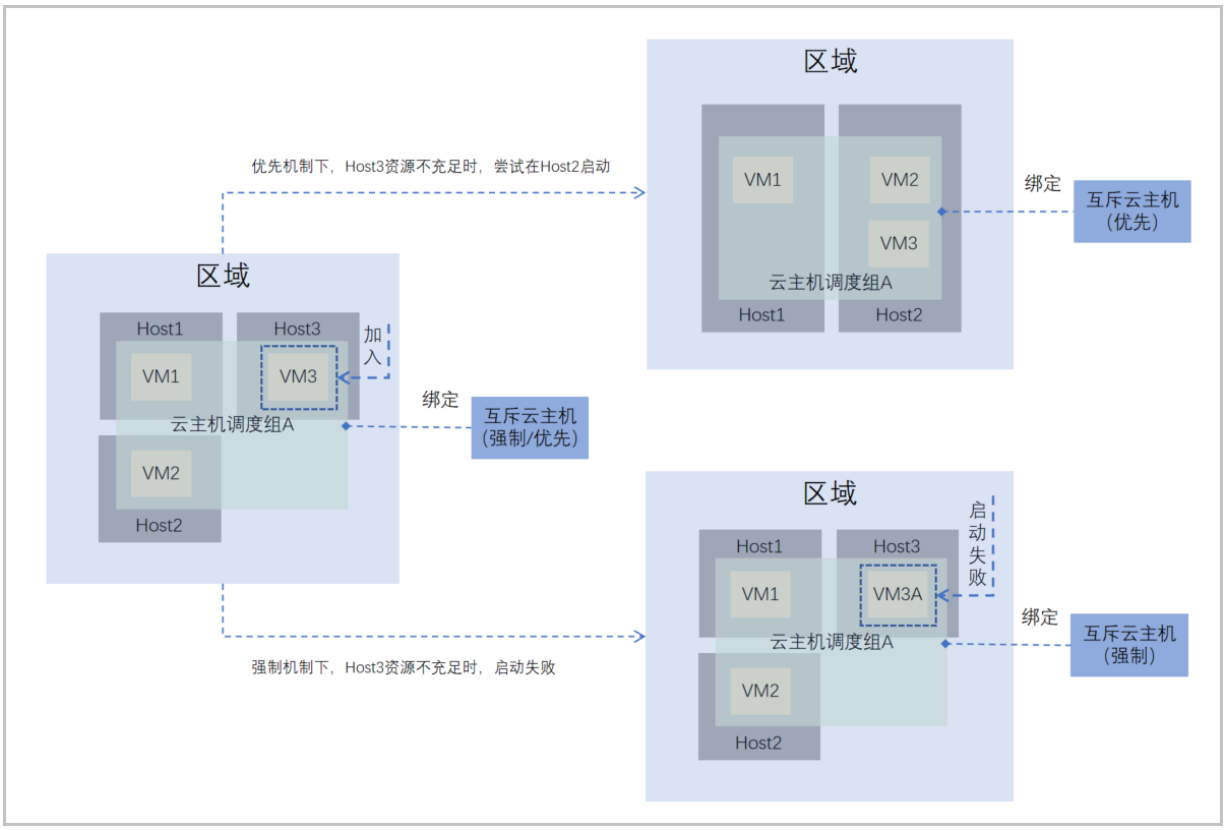
下文通过四个场景说明四类调度策略的工作原理：

场景1：假定区域内有三台物理机Host1、Host2、和Host3。云主机调度组A已绑定**互斥云主机**调度策略，且云主机VM1和VM2已加入该调度组，并分别运行在物理机Host1和Host2上。此时将云主机VM3加入该调度组，不同执行机制下云主机VM3行为如下：

- 强制机制下，云主机VM3遵循与组内其他云主机强制互斥原则：
  - 若物理机Host3资源充足，可正常在Host3上启动并运行。
  - 若物理机Host3资源不足，无法在Host3上启动。
- 优先机制下，云主机VM3遵循与组内其他云主机尽量互斥原则，优先在Host3上启动：
  - 若物理机Host3资源充足，可正常在Host3上启动并运行。
  - 若物理机Host3资源不足，VM3可尝试在其他资源充足的物理机上启动。在该场景下，VM3在Host2上启动并运行。

如图 *3-18: 互斥云主机#强制/优先#*所示：

图 3-18: 互斥云主机（强制/优先）

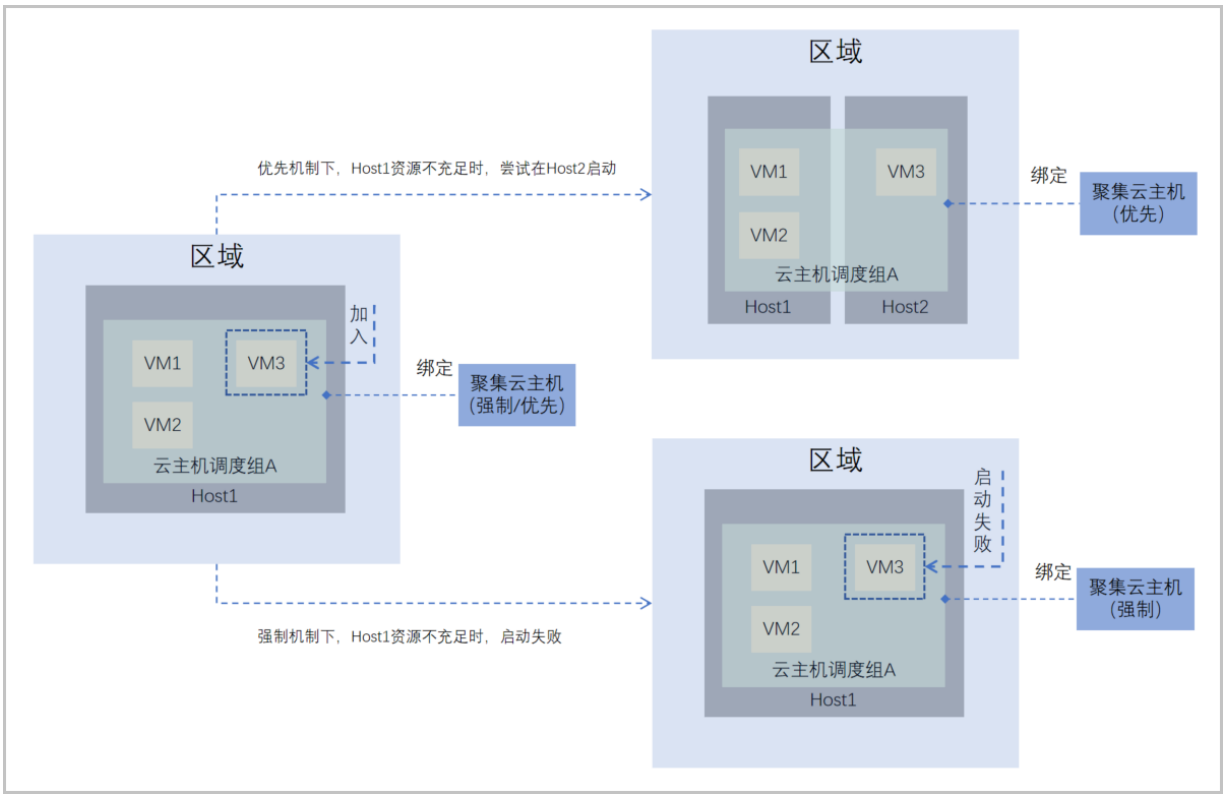


场景2：假定区域内有两台物理机Host1和Host2。云主机调度组A已绑定**聚集云主机**调度策略，且云主机VM1和VM2已加入该调度组，并运行在物理机Host1上。此时将云主机VM3加入该调度组，不同执行机制下云主机VM3行为如下：

- 强制机制下，云主机VM3遵循与组内其他云主机强制聚集原则：
  - 若物理机Host1资源充足，可正常在Host1上启动并运行。
  - 若物理机Host1资源不足，无法在Host1上启动。
- 优先机制下，云主机VM3遵循与组内其他云主机尽量聚集原则，优先在Host1上启动：
  - 若物理机Host1资源充足，可正常在Host1上启动并运行。
  - 若物理机Host1资源不足，VM3可尝试在其他资源充足的物理机上启动。在该场景下，VM3在Host2上启动并运行。

如图 *3-19:* 聚集云主机#强制/优先#所示：

图 3-19: 聚集云主机（强制/优先）

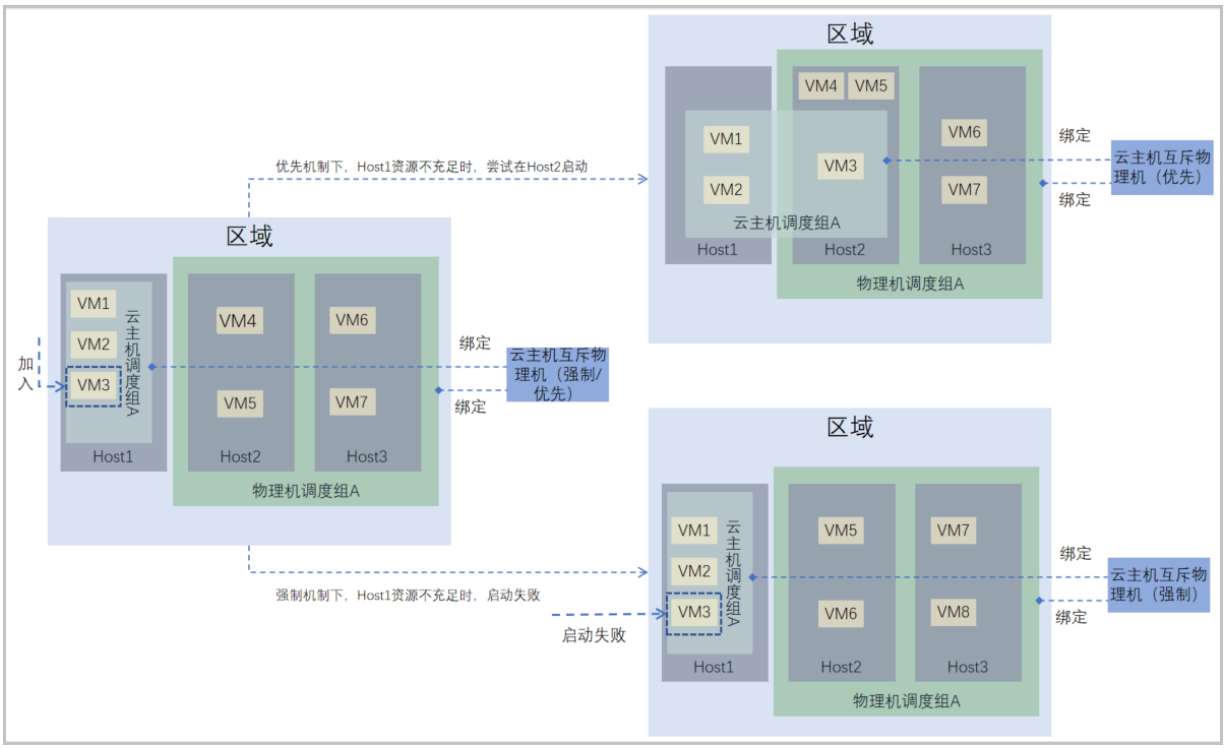


场景3：假定区域内有三台物理机Host1、Host2、和Host3。云主机调度组A已绑定**云主机互斥物理机**调度策略，且云主机VM1和VM2已加入该调度组，并运行在物理机Host1上。物理机调度组A也已绑定**云主机互斥物理机**调度策略，且物理机Host2和Host3上已加入该调度组，分别运行两台云主机。此时将云主机VM3加入云主机调度组A，不同执行机制下云主机VM3行为如下：

- 强制机制下，云主机VM3遵循与物理机调度组A内的物理机强制互斥原则：
  - 若物理机Host1资源充足，可正常在Host1上启动并运行。
  - 若物理机Host1资源不足，无法在Host1上启动。
- 优先机制下，云主机VM3遵循与物理机调度组A内的物理机尽量互斥原则，优先在Host1上启动：
  - 若物理机Host1资源充足，可正常在Host1上启动并运行。
  - 若物理机Host1资源不足，VM3可尝试在其他资源充足的物理机上启动。在该场景下，VM3在Host2上启动并运行。

如图 *3-20: 云主机互斥物理机#强制/优先#*所示：

图 3-20: 云主机互斥物理机（强制/优先）

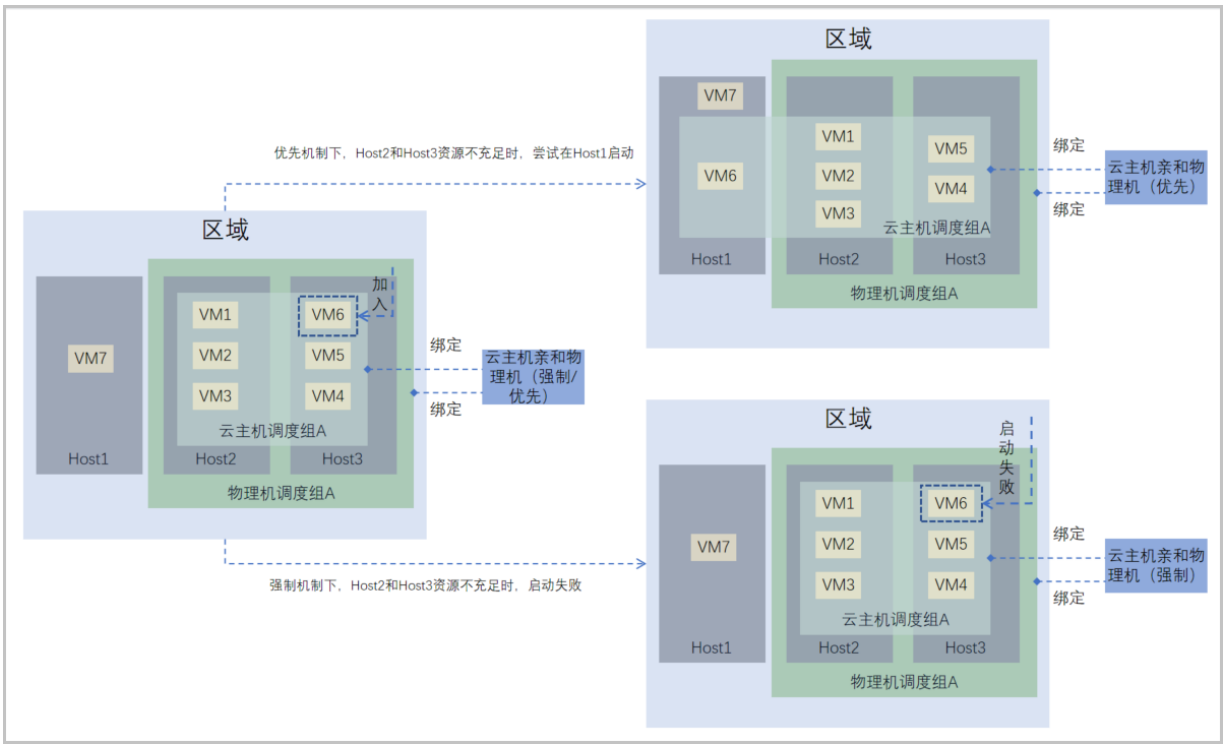


场景4：假定区域内有三台物理机Host1、Host2、和Host3。云主机调度组A已绑定**云主机亲和物理机**调度策略，且云主机VM1~VM5已加入该调度组，并分别运行在物理机Host2和Host3上。物理机调度组A也已绑定**云主机亲和物理机**调度策略，且物理机Host2和Host3上已加入该调度组。此时将云主机VM6加入云主机调度组A，不同执行机制下云主机VM6行为如下：

- 强制机制下，云主机VM3遵循与物理机调度组A内的物理机强制聚集原则：
  - 若物理机Host2或Host3资源充足，可正常在Host2或Host3上启动并运行。
  - 若物理机Host2和Host3资源不足，无法在Host2或Host3上启动。
- 优先机制下，云主机VM3遵循与物理机调度组A内的物理机尽量聚集原则，优先在Host2或Host3上启动：
  - 若物理机Host2或Host3资源充足，可正常在Host2或Host3上启动并运行。
  - 若物理机Host2和Host3资源不足，VM6可尝试在其他资源充足的物理机上启动。在该场景下，VM3在Host1上启动并运行。

如图 *3-21: 云主机亲和物理机#强制/优先#*所示：

图 3-21: 云主机亲和物理机（强制/优先）

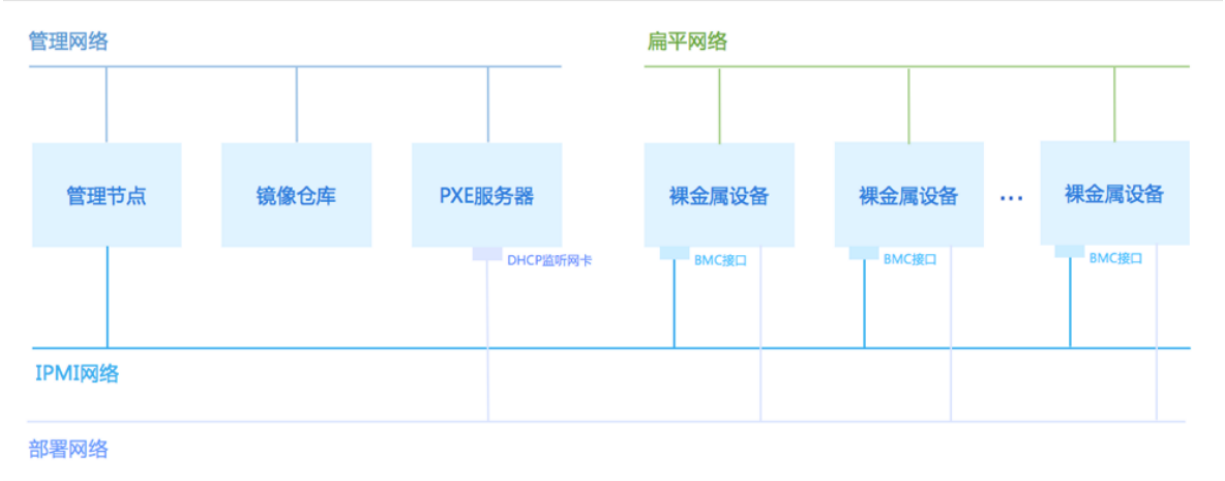


3.2.1.4.2 裸金属管理

裸金属管理的网络流量模型包括管理网络、扁平网络、IPMI网络和部署网络。管理网络主要用于管理云平台相关的硬件资源。扁平网络主要用于裸金属主机的业务网络，对外提供应用服务。IPMI网络用于管理节点对裸金属设备与裸金属主机的开关机、重启、获取硬件信息等操作。部署网络用于PXE服务器通过DHCP服务下发IP地址以及通过TFTP服务传输镜像。

如图 *3-22:* 裸金属管理网络拓扑所示：

图 3-22: 裸金属管理网络拓扑



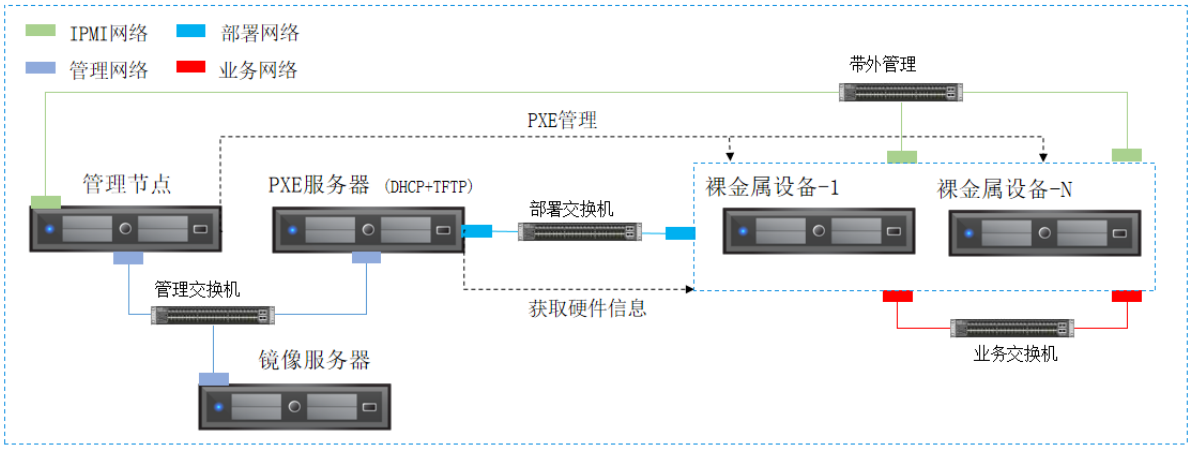
裸金属管理核心包括硬件信息获取和裸金属设备无人值守部署。

管理节点通过PXE服务器指定裸金属设备通过PXE启动。PXE服务器封装DHCP、TFTP和镜像存储服务，在裸金属设备PXE启动后通过DHCP获得IP地址，随后到TFTP服务器上，下载pxelinux.0和引导文件，并把内核加载至内存中执行，引导进LiveCD系统。在LiveCD系统中执行检测脚本，将裸金属设备的硬件信息回传至管理节点。根据回传的硬件信息，对裸金属设备设置预配置模板，包括：分区信息、设置网卡Bond、IP地址等。选择一个待安装的操作系统ISO，即可部署裸金属主机。重启裸金属设备至PXE启动，PXE部署服务器会预先下载待安装操作系统ISO，裸金属设备根据预设好的配置模板实行无人值守部署，部署完毕后会自动根据预配置模板配置网卡等信息，至此裸金属主机配置完毕。

为更好地运维裸金属主机，PXE服务器可下发裸金属监控服务，实现裸金属主机内部数据的实时监控，包括：CPU、内存、磁盘容量、磁盘IO、网卡等信息。

如图 *3-23:* 裸金属管理硬件拓扑所示：

图 3-23: 裸金属管理硬件拓扑



3.2.2 运维管理

3.2.2.1 管理节点监控

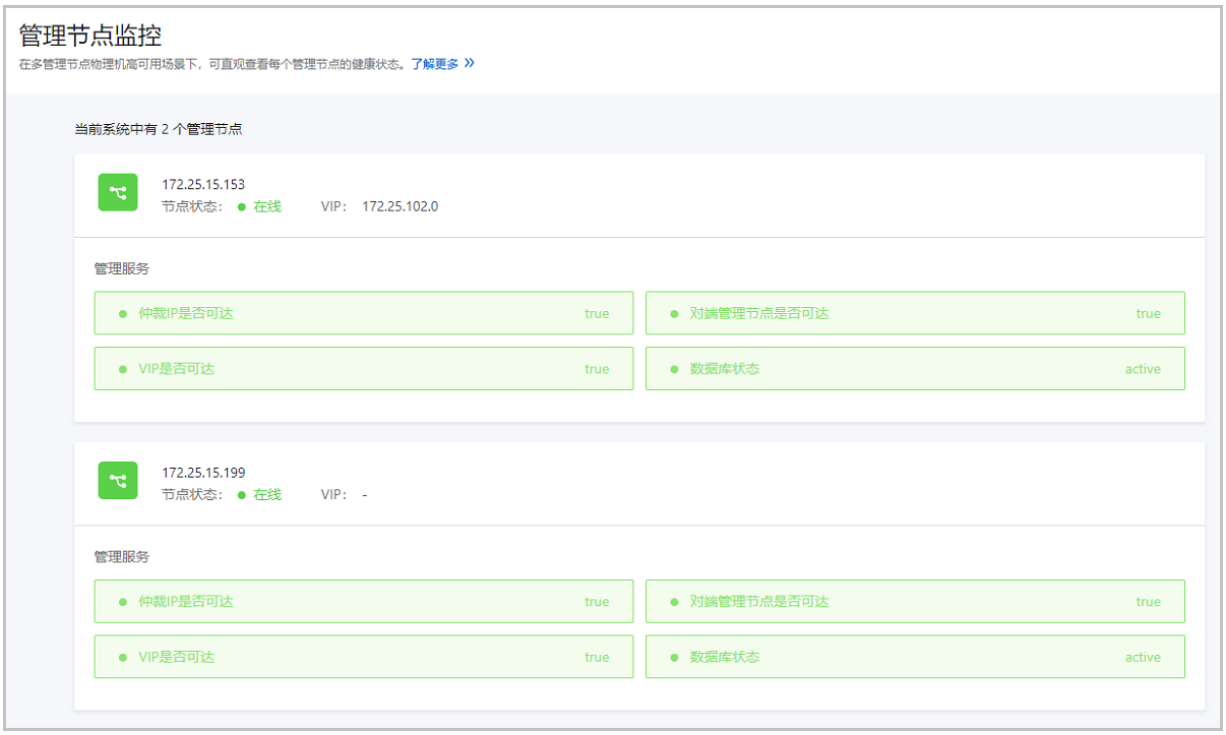
在多管理节点物理机高可用场景下，可直观查看每个管理节点的健康状态。

管理节点监控支持显示多个管理节点的管理节点IP、节点状态、VIP和管理服务状态，主要包括以下管理服务：

- 仲裁IP是否可达：监控用于判断主备管理节点的仲裁IP是否可达，若不可达可能导致管理节点高可用功能失效。
- 对端管理节点是否可达：监控备管理节点是否可达，若备管理节点不可达，无法与备管理节点通信。
- VIP是否可达：监控VIP是否可达，若VIP不可达，主管理节点不能通过VIP访问UI界面。
- 数据库状态：监控数据库状态，若数据库异常或多管理节点数据库不同步，可能存在数据丢失风险，请及时恢复故障。

如图 *3-24:* 管理节点监控所示：

图 3-24: 管理节点监控

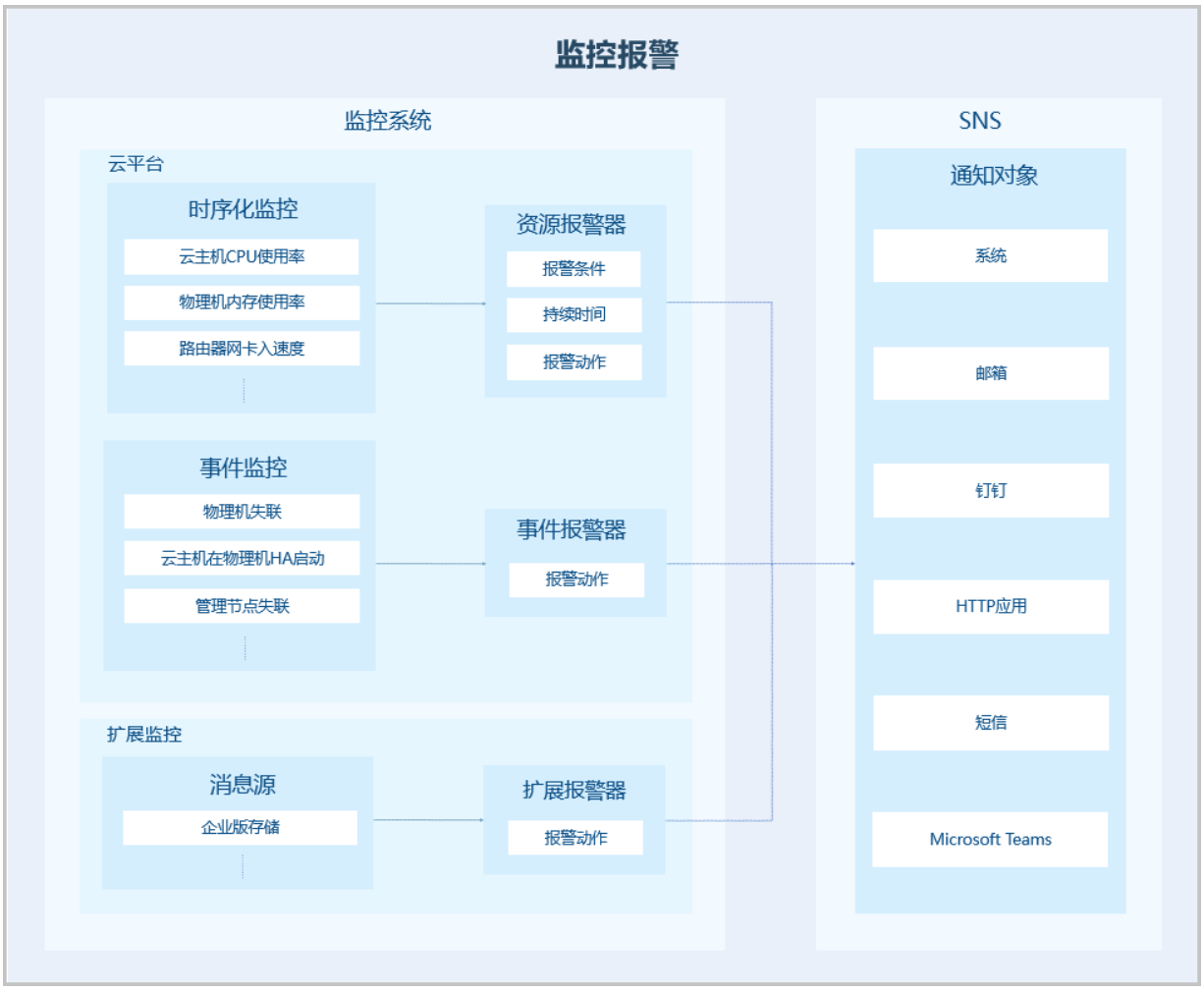


3.2.2.2 监控报警

监控报警支持对时序化数据（如资源负载数据和资源容量数据）以及系统中发生的预定义事件进行监控，并通过通知服务（SNS）推送报警消息至指定的通知对象。支持资源报警器、事件报警器和扩展报警器三种报警器类型，支持系统/邮箱/钉钉/HTTP应用/短信/Microsoft Teams通知对象类型，部分资源报警器需安装agent才能使用。

如图 *3-25:* 监控报警功能框架所示：

图 3-25: 监控报警功能框架

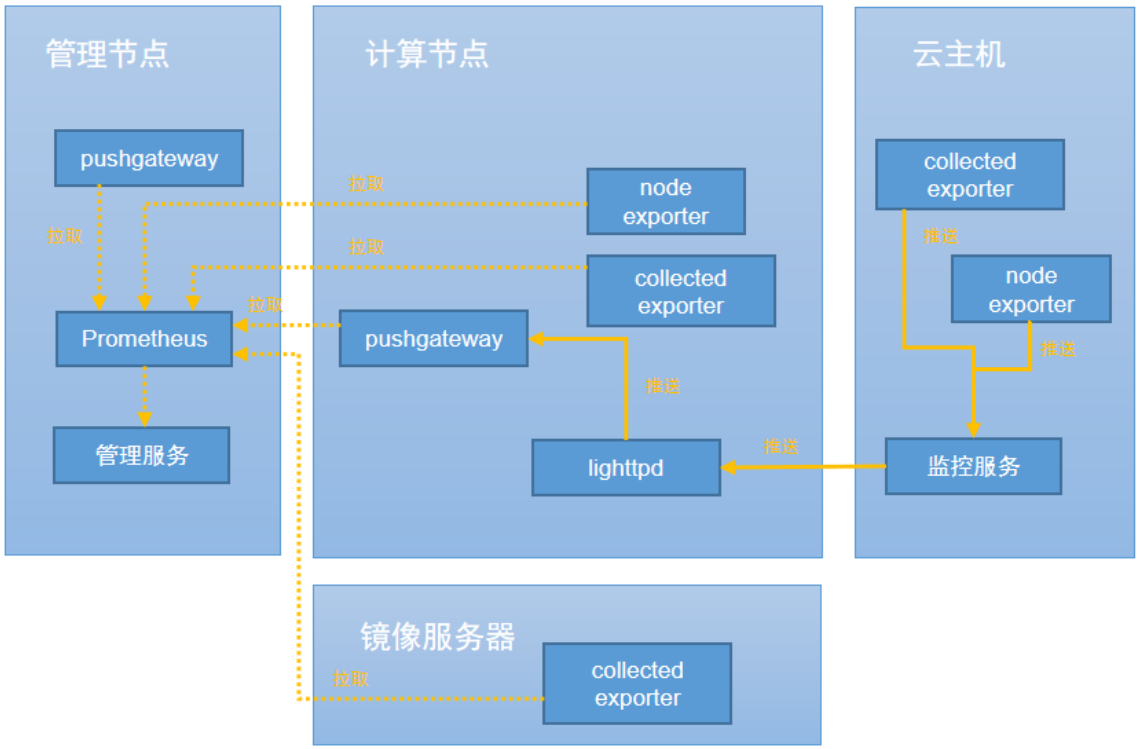


时序监控数据由Prometheus提供，在监控业务数据时，需将不同数据汇总，由Prometheus统一收集。

在Prometheus架构设计中，Prometheus服务器并不直接服务监控特定目标，其主要负责数据的收集、存储，并对外提供数据查询支持。因此，为监控到样本数据，如：物理机CPU使用率，需通过Exporter周期性采集监控样本。云平台针对不同监控目标，分别使用拉取模式和推送模式来采集监控数据。当物理机或云主机外部监控作为监控目标时，Prometheus服务会周期性使用拉取模式采集物理机上Exporter收集到的数据。另外，由于网络问题或安全问题，Prometheus无法直接访问到云主机内部或裸金属服务器内部。此时需一个pushgateway作为中间者完成中转工作。采集端仍通过Exporter采集监控数据， 并采用推送方式周期性将数据推送给pushgateway，随后Prometheus采用拉取方式采集pushgateway数据，从而完成数据的统一收集。

如图 *3-26:* 监控数据采集原理所示：

图 3-26: 监控数据采集原理



3.2.2.3 一键巡检

AliyunHCI-Z支持对关键资源和服务进行全方位一键式健康检查，并根据巡检结果为巡检资源和服务进行健康评分，同时提供巡检建议和巡检报告，助力高效运维，确保云平台资源和服务处于最佳状态。一键巡检适用于需要对云平台进行集中高效运维场景。

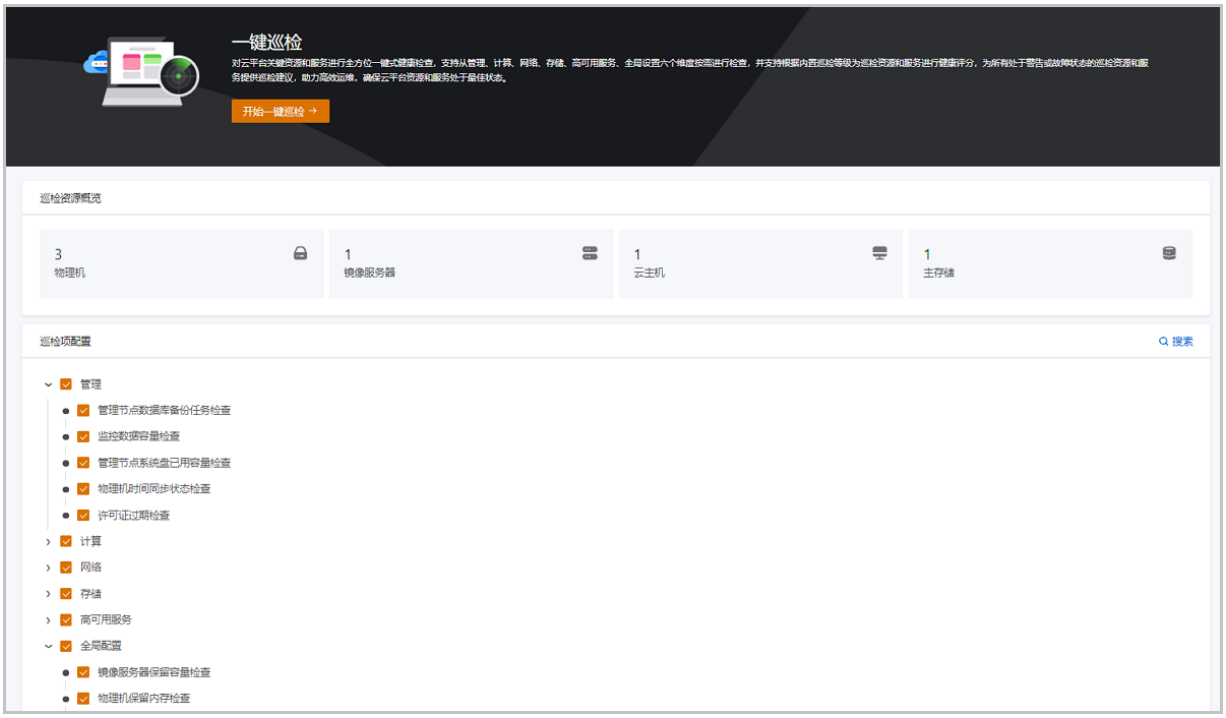
一键巡检提供平台、计算、网络、存储、全局设置五大类别巡检项，支持对管理节点、物理机和云主机、镜像服务器和主存储、物理/虚拟网络和网卡、许可证等云平台关键资源和服务进行巡检：

- 平台：检测云平台基础服务和运行状态。
- 计算：检测云平台物理计算资源和虚拟化计算资源使用状况和运行状态。
- 网络：检测云平台物理网络和虚拟化网络配置和状态。
- 存储：检测云平台物理存储资源使用状况和运行状态。
- 全局设置：检测云平台全局性重要资源的配置情况。

用户可自定义根据类别选择巡检项进行一键巡检，启动巡检后，云平台将对所选择的巡检项涉及的资源或服务进行健康检查。一键巡检内置健康评分机制，支持对所巡检的资源或服务的健康状态进行量化评分，帮助用户直观准确把握云平台整体运行状态。

如图 *3-27: 一键巡检*所示：

图 3-27: 一键巡检



3.2.3 数据保护

3.2.3.1 灾备管理

3.2.3.1.1 数据备份

支持基于Qemu块设备层的数据备份，各类型主存储上的云主机均支持备份。备份类型可分为：全量备份、增量备份。全量备份包含完整的数据集合，增量备份仅包含自上一次备份后所有更新的数据集合。全量备份和增量备份均仅备份真实数据。

默认情况下，备份策略是在首次全量备份后，每63个增量备份的下一次备份就会自动执行一次全量备份。这是因为增量备份之间有依赖关系，在做新的一次全量备份后，才能对之前的增量备份进行删除。这里增量备份的数量可以通过全局配置修改。实际上，系统内部有更智能灵活的应对策略来决定使用哪种合适的备份方式，以确保备份数据的安全可靠。

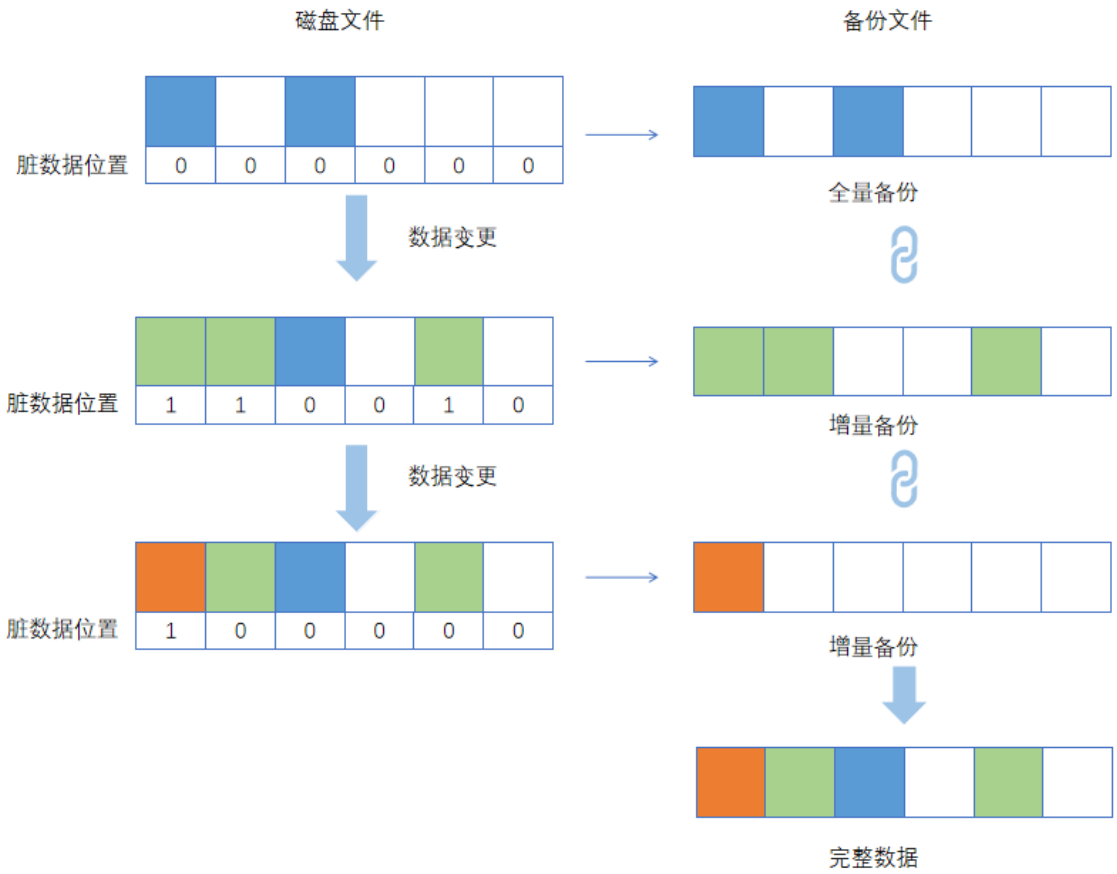
数据备份可分为三部分：数据复制、数据传输、数据保存。

3.2.3.1.1.1 数据复制

备份模块利用Qemu块设备层的脏数据跟踪功能（dirty bitmap）实现的备份数据导出。云主机磁盘文件数据发生变化的位置被称为脏数据位置，dirty bitmap记录自上次备份后，虚拟磁盘文件上产生脏数据的所有位置记录，根据位置记录，就可导出自上次备份后所有被修改过的数据，即增量的备份数据。最终全量备份文件和各个增量备份文件会产生一个完整的备份链，保存完整的数据。dirty bitmap存在于Qemu进程的内存中，云主机重启后就会丢失这部分信息，因此当云主机重启后的下一次备份，系统会自动选择全量备份。

如图 *3-28:* 数据复制所示：

图 3-28: 数据复制



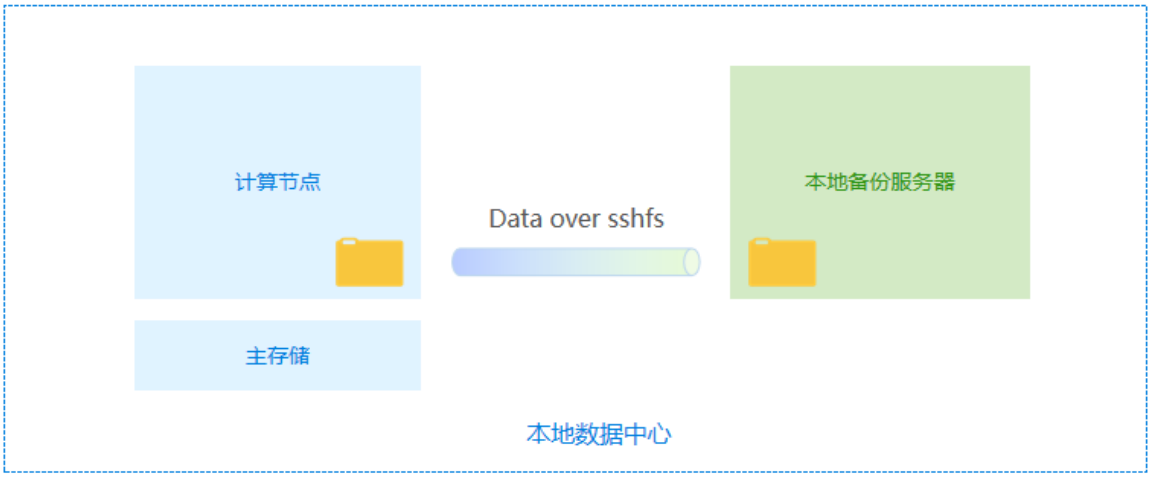
3.2.3.1.1.2 数据传输

针对不同的虚拟化组件版本，支持两套不同的实现方案，主要区别在备份数据的传输上。

第一种方案。使用sshfs在计算节点上挂载远程备份服务器的备份目录，然后把备份数据导入备份服务器。sshfs是一个简单的fuse over ssh方案，数据链路由ssh会话加密，每个备份任务有单独的sshfs链路。

如图 *3-29: Data over sshfs*所示：

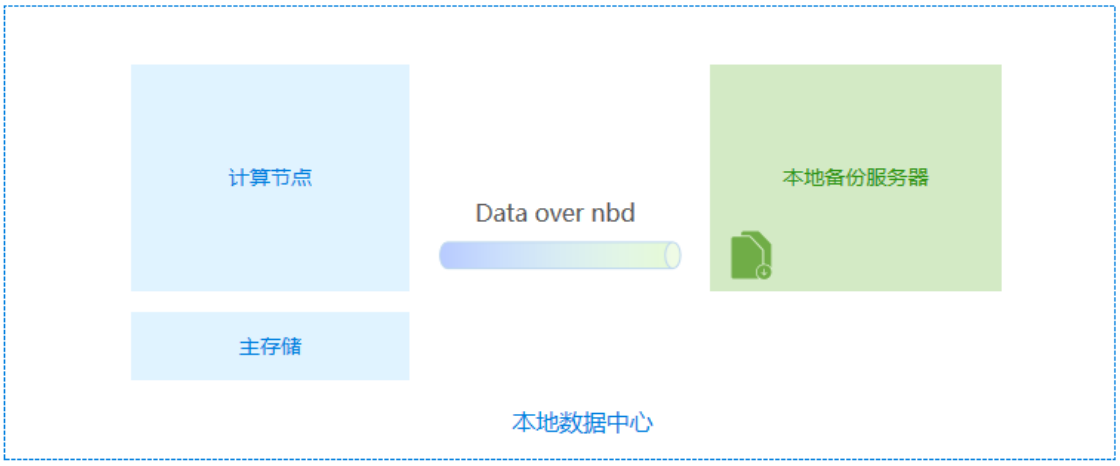
图 3-29: Data over sshfs



第二种方案。在备份服务器上使用nbd模块导出一个备份磁盘，然后在计算节点通过qemu的块设备任务（block-job），直接把备份数据导入到备份磁盘中。

如图 *3-30: Data over nbd*所示：

图 3-30: Data over nbd



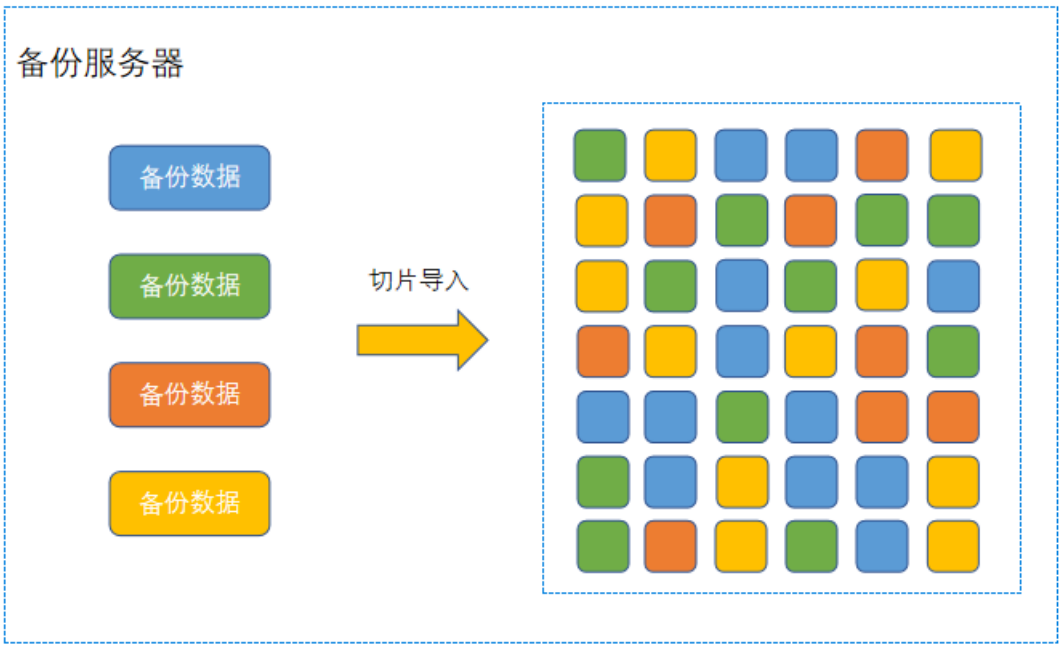
3.2.3.1.1.3 数据保存

备份服务器支持多种存储介质，包括：SAN、NAS、磁盘阵列以及带库等。

备份数据在备份服务器中是切片去重存放的，备份数据会被切分成64MB大小的数据块，然后计算hash，建立索引。拥有相同hash的数据块不会被存储多份。

如图 *3-31:* 数据切片保存所示：

图 3-31: 数据切片保存

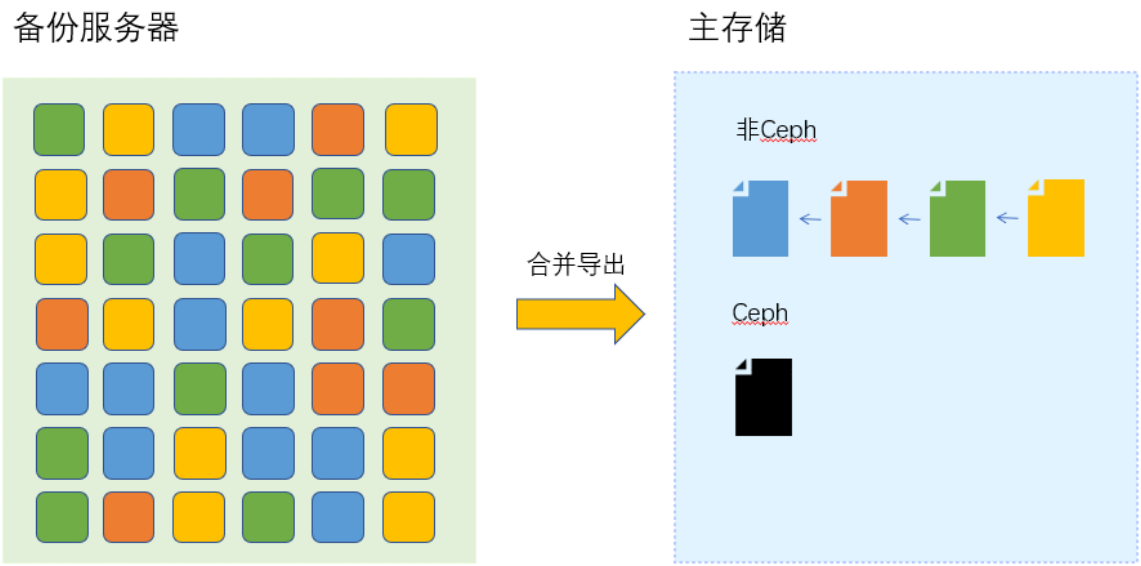


3.2.3.1.2 数据恢复

从本地备份数据恢复云主机/云盘，会把备份服务器上的切片数据合并导入主存储中，合并后的备份恢复数据会在非Ceph主存储上以磁盘链的形式存放，Ceph主存储上会把磁盘链合并成单个磁盘文件存放。 如果是新建恢复，则恢复到主存储的磁盘数据会被当作镜像缓存来创建新云主机/云盘；如果是覆盖恢复，则会把恢复到主存储的磁盘路径更新到当前云主机/云盘的数据库记录中，随后删除旧的云主机/云盘文件。

如图 *3-32:* 数据恢复所示：

图 3-32: 数据恢复



3.3 关键流程

业务云主机部署在KVM虚拟化平台上，存储采用分布式存储资源，云主机IO通过SCSI协议，直接与运行在Hypervisor内核中的分布式存储软件进行数据交互。

4 高性能

4.1 通用SSD读写缓存

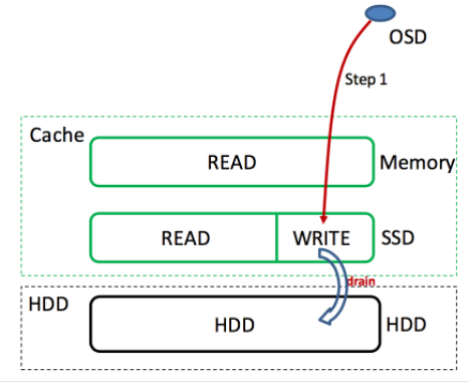
分布式Cache算法支持SSD读写Cache，在SSD Cache的支持下使用户获得更高的性能支撑。

实现原理

OSD 在收到XDC发送的写IO操作时，会将写IO缓存在SSD Cache上完成本节点写操作。当写Cache达到刷盘水位时，OSD会将SSD Cache中的写IO数据批量写入HDD硬盘；写Cache水位值75%（读写cache比40%：60%，所以写cache水位为60% * 75%），在没有IO的情况下，5分钟后也会将Cache中数据写入到HDD硬盘中。

如图 *4-1: 实现原理*所示：

图 4-1: 实现原理



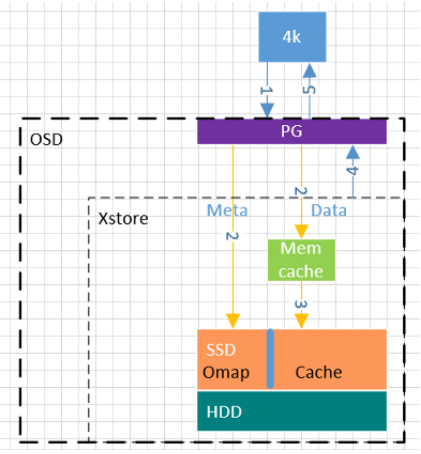
SSD Cache读写流程

- 写流程

1. 写请求发送至PG，元数据写入元数据区，数据写入数据区。
2. 如果数据写IO命中MEM，则写入MEM命中部分，继续；如没有命中，继续。
3. 数据写入Cache，当数据和元数据都写完成后Xstore返回写成功。
4. OSD返回前端写成功。

如图 *4-2: SSD Cache写流程*所示：

图 4-2: SSD Cache写流程

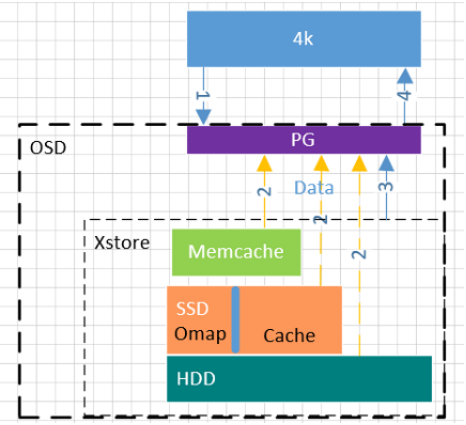


- 读流程

1. 读请求发送至PG，从数据区读取数据。
2. 如果数据读IO命中MEM，则从MEM中读取数据后返回。
3. 如果数据读IO命中Cache，则从Cache中读取数据后返回。
4. 如果数据读IO命中Disk，则从Disk中读取数据后返回。
5. Xstore返回读成功。
6. OSD返回前端读成功。

如图 *4-3: SSD Cache*读流程所示：

图 4-3: SSD Cache读流程



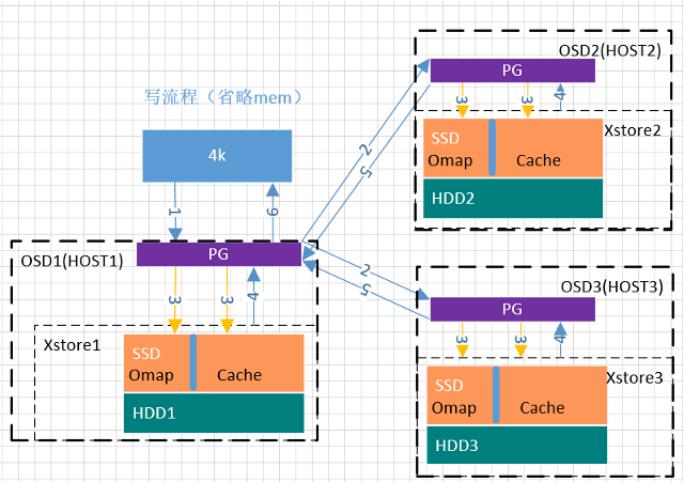
SSD Cache OSD多副本读写流程

- 写流程

1. 对于每个OSD写流程同SSD Cache。
2. 三个OSD的PG写是异步写；Xstore写返回也是异步。
3. OSD2和OSD3的返回是异步。
4. 主PG在收到所有的写正确返回后，所在OSD向前端返回写成功。

如图 *4-4: SSD Cache OSD多副本写流程*所示：

图 4-4: SSD Cache OSD多副本写流程

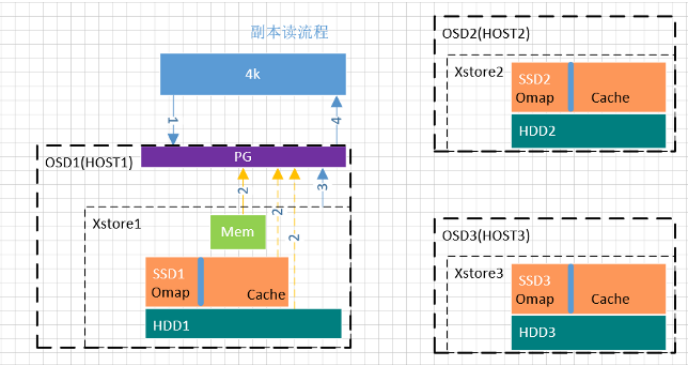


- 读流程

1. 多副本只读主PG。
2. 流程同单OSD读流程。

如图 *4-5: SSD Cache OSD多副本读流程*所示：

图 4-5: SSD Cache OSD多副本读流程

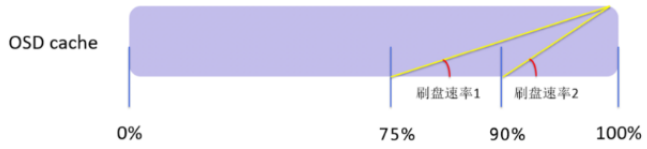


SSD Cache刷盘流程

1. 每个OSD都有自己的SSD Cache，刷盘操作各自独立。
2. 默认刷盘水位为Cache容量75%。
3. 水位越高，刷盘速率越快。由系统算法决定，不能进行修改设置。
4. 当写Cache水位下降至75%以下后不再进行刷盘；若5分钟后无IO会全速刷盘；5分钟内有稀疏IO（心跳或查询命令）会有按照稀疏IO的大小阶梯刷盘。
5. 刷盘IO与前端读写IO无关，但刷盘操作会影响前端IO性能，当系统检测到前端IO时，刷盘会调整策略尽量降低对前端IO操作的影响。
6. 当刷盘IO与前端写IO命中同一Cache 32K块时，待前端IO写完后再一次刷盘，因此不会产生一致性错误等问题。

如图 *4-6: SSD Cache刷盘流程水位*所示：

图 4-6: SSD Cache刷盘流程水位



4.2 大块IO优化写入策略

对于大小块混合写的业务场景，在SSD Cache中可以识别出大于等于128K的IO，将其Bypass直接下发到HDD。用户可以自定义大块IO Bypass的SSD Cache阈值，默认大块IO优先写入SSD Cache中，当达到阈值后，超出的大块IO会选择下发到HDD中。但是当HDD盘利用率也较高时，大块IO仍然回到SSD Cache中。

通过该方式，大块IO可以同时占满HDD和SSD的带宽，从而解决以下两方面问题：

- 大块IO访问SSD Cache会受限于SSD的带宽，影响整个集群的带宽。
- 大块IO访问SSD Cache会增大SSD的压力，大幅影响SSD上的其他小块IO时延。

4.3 通用RAM读缓存

在软件定义分布式存储的Cache体系中，内存只用于读Cache用，避免由于服务器掉电带来的数据丢失问题。RAM读缓存将符合策略的数据提前缓存在内存中，在业务下一次读IO请求时直接从内存读取数据返回，可以显著提升读性能。

4.4 SSD Cache热点识别技术

面对高性能数据库需求，为了提升随机小IO的并发访问能力，支持在数据读取过程中判断热点数据，将数据长久保存在SSD中，随着时间推移，算法会配合SSD 使用率，动态地将冷数据回刷至HDD中，将热点数据持续更新在SSD中，提升数据库持续读取性能。

4.5 流预测技术

在面向视频读取场景时，业务端并发读取多路数据，存储端收到的IO在某个时刻呈现随机分布状态。例如：在业务同时获取200路视频流时，同一时刻存储端接收到200个随机读取请求，从磁盘读取这些随机IO性能非常低下。流预测技术支持准确对数百路流式访问进行识别和预测，针对每一路访问提前预取。流预测策略和RAM读缓存机制配合使用，提升流媒体业务场景并发读取性能。

4.6 合并刷盘技术

在IO栈中，离散随机小块IO会先写到SSD中，通过合并刷盘技术将离散随机小块IO合并成连续大块IO写入到HDD中，提升系统整体IOPS性能，同时减少HDD硬盘寻道次数，降低硬盘损耗，延长硬盘使用寿命。

4.7 热卷缓存锁定

对指定卷实现专用缓存加速，避免性能瓶颈。热卷缓存锁定功能将内存与数据盘之间进行关联，锁定内存空间，形成虚拟缓存盘，存放数据盘内指定的云主机镜像母卷，云主机启动及运行都是基于内存的读，可防止大量VDI在短时间内同时启动产生大量存储 I/O，造成云主机启动慢访问延迟高等问题，提升VDI场景用户读写体验。

实现原理

创建性能优先卷之后，系统会自动启动热卷缓存锁定功能，性能优先卷所在的存储池中的每个硬盘都会与本服务器中的128MB内存空间建立关联，并将关联的内存空间锁定，形成性能优先卷的专属Cache。如存储池中有5块硬盘，则该卷的专属Cache为：128MB * 5 = 640MB。

数据写入性能优先卷后都会被打上标记，该数据在写入数据盘之后，默认会被一直缓存在专属Cache中，即只要存储系统启动，被标记的数据块就会被放置在性能优先卷的专属Cache中，如性能优先卷写入的数据量大于专属Cache的容量，则采用性能优先卷热点数据算法机制，优先将访问频率高的数据缓存在专属Cache中。

性能优先卷的专属Cache不会与其它内存混用，也不会被其它业务影响。

5 扩展性

5.1 性能容量线性增长

- 扩容存储节点后无需做大量数据搬迁，系统快速达到负载均衡状态。
- 支持灵活扩容，可独立扩容计算节点、硬盘、存储节点，或同时进行扩容。在扩容计算节点时同步扩容存储空间，扩容后的系统仍可为计算和存储融合。
- 控制器、存储带宽和Cache均匀分布到各个节点上，系统IOPS、吞吐量和Cache随着节点的扩容而线性增加。

5.2 数据自动负载均衡

CRUSH算法可保证上层应用对数据IO操作均匀分布在不同服务器的不同硬盘上，不会出现局部热点，实现全局负载均衡。

- 系统自动将每个卷的数据块打散存储在不同服务器的不同硬盘上，冷热不均的数据会均匀分布在不同的服务器上，不会出现集中的热点。
- 数据分片分配算法保证了主用副本和备用副本在不同服务器和不同硬盘上的均匀分布，即，每块硬盘上的主用副本和备副本数量是均匀的。
- 扩容节点或者故障减容节点时，数据恢复重建算法保证了重建后系统中各节点负载的均衡性。

CRUSH算法具备以下特点：

- 均衡性：数据能够能均匀的分布到所有的节点中。
- 单调性：当有新节点加入系统中，系统会重新做数据分配，数据迁移仅涉及新增节点，现有节点上的数据不需要做很大调整。
- 适应性：与DHT不同的是，CRUSH在做数据分布计算时，算法是可以动态调整的，当系统中出现性能、负载不一致的节点时，CRUSH算法可以根据调整输入参数优化算法，重新平衡负载。

6 可靠性

6.1 数据存储冗余

多副本

支持用户数据按照设定的1-6副本进行冗余存储。以3节点2副本为例，任意1个节点上的主副本数据，其备副本数据都会均匀分布在其他节点上，单点故障不会丢失数据。

存储系统通过强一致性复制协议来保证数据多个副本的一致性。只有当数据的所有副本都写入成功后，才会返回前端数据写入完成。正常情况下存储系统可以保证每个副本上的数据都是完全一致的，从任意副本读到的数据都是相同的。

如果系统中的某个硬盘出现短暂故障，存储系统会暂时不写这个硬盘上的数据，通过日志记录的方式，记录此硬盘上数据的变化，等硬盘恢复后通过日志信息恢复该硬盘上的数据，如果硬盘长时间或者永久故障，存储系统会将硬盘从存储系统中移除掉，并统计出此硬盘上所有数据的副本位置，将这些丢失数据恢复到其它服务器的硬盘中。

以下为副本与服务器最小数量对应关系：

副本数量	空间利用率	服务器最小数量	备注
2副本	50%	3台	3台MON与存储服务器复用
3副本	33%	5台	5台MON与存储服务器复用
4副本	25%	4台存储服务器+3/5台MON	4副本以上建议采用独立MON服务器，需额外增加3台独立MON服务器
5副本	20%	5台存储服务器+3/5台MON	
6副本	16%	6台存储服务器+3/5台MON	

说明：

- 考虑MON与存储服务器复用的情况，副本数与服务器最小数量对应关系为：服务器最小数量=2n-1，n为副本数量。
- 如无需考虑MON，即MON在系统内处于独立的状态，则服务器最小数量=副本数量。

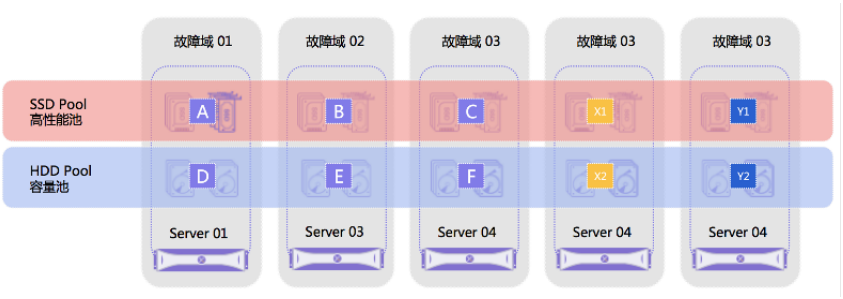
EC纠删码

支持EC纠删码机制，流程如下：

1. 按照4MB*N（N为N+M中的N）的粒度对数据进行切片，并将切片后的数据根据负载均衡算法写入到主OSD所在的存储服务器的内存中。
2. 在主OSD的内存中再次将数据片段切分成N块数据并通过计算编码得到M块冗余校验数据。
3. 将第一份数据存储在主OSD中，并将其它数据分布式地通过后端专用网络并行地存储在其它存储服务器中。
4. 当所有N+M块数据写入完成后，返回主OSD数据写入完成，主OSD返回客户端写入完成。
5. 客户端通过流式写入方法将其它切片的数据块写入主OSD。

如图 *6-1: 纠删码*所示：

图 6-1: 纠删码



6.2 故障域隔离

故障域由具有共同单点故障的服务器（或机架）组成，数据副本分布到不同故障域，保障数据安全。可为机架、服务器、硬盘提供故障恢复能力。无论磁盘、服务器发生硬件故障，还是机架、机房出现故障，都不会造成停机或数据丢失。

6.3 数据强一致性

采用强一致性复制协议保证多副本数据的一致性，即只有当所有副本都写成功，才返回确认信号。正常情况下存储系统可以保证每个副本上的数据都是完全一致的，从任一副本读到的数据都是相同的。如果某个副本中的磁盘短暂故障，存储系统会暂时不写该副本，等磁盘恢复后再恢复该副本数据。如果磁盘长时间或者永久故障，存储系统会把该磁盘从群集中移除，并为副本寻找新的磁盘，通过重建机制使得数据在磁盘上均匀分布。

6.4 令牌桶I/O流控

支持基于令牌桶的IO流控技术，实现IO资源过载流控。当IO过载时，根据流控策略有选择地减少低优先级业务，优先保证高优先级业务成功。根据系统的过载程度，流控模块相应的调整系统处理的IO业务量，尽快将系统恢复到正常负载。

引入令牌桶IO流控技术，可应对各种异常对IO的冲击，增强系统鲁棒性。

6.5 磁盘数据可靠性

支持硬盘S.M.A.R.T检测、快慢盘检测、磁盘SCSI错误处理、硬盘热插拔和识别处理、磁盘扫描等，上层业务根据Smart Data返回的相关IO错误和磁盘状态信息， 完成读修复、有效数据磁盘扫描及纠错、Smart超阈值告警和处理。

读修复功能（Read Repair）

在读操作时如果发现有读失败，通过错误类型判断，若是磁盘扇区读取错误，可从其它副本读取数据然后重新写入进行数据恢复。这是磁盘的一种特性，大部分读扇区错误都可以修复。如果此方法还不能修复，那么就通过隔离流程为副本选择其它硬盘并把故障的硬盘踢出集群。

有效数据磁盘扫描及纠错

通过对数据进行读扫描，防止静默数据错误（Silent Data Corruption），如果因为出现坏道导致扫描失败（返回扩展的EIO），则进行更细粒度的扫描，找出具体是哪些扇区故障，并针对故障扇区进行读修复。

Smart超阈值告警和处理

当检测到超阈值或者慢盘时，发出告警，系统优先将该盘上的主分区迁移，同时预先重建另一份拷贝（如果原有为2份拷贝，新增1份变为3份拷贝），待这份拷贝重建完成后，再将超阈值或慢盘进行移除磁盘处理。

6.6 故障检测

支持免人工干预的磁盘、主机自动故障检测和报警。

6.7 故障自愈

支持在更换磁盘、主机后自动进行数据重建和均衡，且数据重建速度可设置。

6.8 数据一致性校验

支持后台数据一致性校验，所有存取请求携带校验并连同数据存储到介质，可以有效防止数据的静默错误。

6.9 磁盘维护模式

支持硬盘维护模式，当设置硬盘维护模式后，硬盘处于不会out状态，避免在维护硬盘和服务器时，对硬盘上的数据进行重平衡。

6.10 磁盘重建

提供基于硬盘或者SSD的磁盘重建方案，可以帮助用户解除因更换硬盘/SSD而产生的数据多次重建带来的系统风险。

正常情况下，普通硬盘（OSD）故障后需要将该硬盘从存储池中移除，再将该OSD删除，插入新的物理盘，创建为新的OSD再加入存储池。同样如果一块SSD缓存盘故障后，会导致相关联的OSD都故障。重新添加缓存盘，需要将每个OSD执行删除并重新添加进集群。上述两种情况下都需要删除故障OSD和重新添加新OSD，从而导致两次数据平衡，对业务的压力会比较大，而且整个流程比较冗长和复杂。

磁盘重建功能可以有效应用在计划内的磁盘替换，配合SSD寿命预测功能，有效在SSD寿命即将到期时做预警，并进行SSD替换。

6.11 磁盘漫游

支持将故障存储节点中的硬盘和SSD更换到新的存储节点设备上，同时重新恢复OSD至原存储池，有效地避免因更换存储节点导致的数据多次重平衡，方便运维团队快速更换整机，减少维护问题。

当存储节点发生CPU、内存、主板等硬件故障，或整机系统故障无法启动时，需要进行整机硬件更换或者重新安装操作系统，这时旧设备下线、新设备上线，会出现两次数据重平衡，但是其实原存储节点中的磁盘并没损坏，数据依旧在，只是需要重新识别。该功能是将故障存储设备的硬盘和SSD转移到新更换的存储节点上，通过管理操作界面将新存储节点重新加入到原存储池，而无需做多次数据重平衡。

6.12 亚健康

DBSCAN聚类算法

支持通过DBSCAN聚类算法实现对磁盘的分析，以OSD为单位，将亚健康的指标数据（包括时延、iostat等）进行聚类。为消除突发时延对于亚健康判断的影响，指标数据将从一个滑动窗口获取，和一个更长时间内的样本做比较，OSD向MON上报指标数据，MON通过对指标数据的分析，判断磁盘是否处于亚健康的状态。在当前的数据冗余度，满足数据安全的前提下进行亚健康磁盘的隔离，从而去除亚健康设备对于整个存储系统的影响。

LOF局部离群因子算法

支持通过LOF局部离群因子算法，即：基于距离的异常点检测算法，寻找观测值和参照值之间有意义的偏差，主要目的是为了检测出那些与正常数据行为或特征属性差别较大的异常数据或行为，将OSD间链路上的相关数据（OSD间网络请求的传输往返延迟与重传请求的数量等）作为网络亚健康的指标数，这些数据将通过OSD上报到MON服务，进行局部离群因子算法分析，MON通过对指标数据的分析，判断OSD间链路是否处于的亚健康的状态。

7 安全性

7.1 计算安全

7.1.1 HTTPS加密登录UI

支持HTTPS方式登录UI管理界面，进一步提升系统安全性。

- HTTPS方式默认不启用。
- 启用HTTPS后，系统默认支持5443端口，且支持自定义指定其它端口登录。
- 启用HTTPS后，如使用HTTP方式的5000端口登录，将会自动重定向到HTTPS方式。目前仅支持HTTP的5000端口自动重定向到HTTPS。
- 系统默认支持**PKCS12**格式证书。目前仅提供PKCS12/JKS格式证书支持，如使用其它格式证书，请自行格式转换。

7.1.2 云主机控制台

云主机控制台为用户提供了快捷监控管理云主机的入口，用户必须具有相应权限才可登录云主机控制台。提供两种认证方式登录云主机控制台：SSH密钥方式、用户名/密码方式。

- SSH密钥方式
  - 支持SSH密钥方式登录云主机，目前仅适用于Linux云主机。
  - SSH密钥是通过一种加密算法生成的一对密钥：一个为**公钥**，对外公开，一个为**私钥**，由用户自己保留。
  - 云主机注入公钥后，用户可在另一台云主机中，通过私钥SSH登录已注入公钥的云主机，而无需输入密码。
  - 前提需确保云主机镜像已预先安装cloud-init，且cloud-init推荐版本为：0.7.9、17.1、19.4、19.4以后版本。
- 用户名/密码方式
  - 支持用户名/密码方式登录云主机。
  - Linux云主机固定用户名**root**，Windows云主机固定用户名**administrator**。
  - 云主机注入密码后，用户可在另一台云主机中，通过用户名/密码SSH登录已注入密码的云主机。
  - 前提需确保云主机镜像已预先安装cloud-init，且cloud-init推荐版本为：0.7.9、17.1、19.4、19.4以后版本。

7.1.3 高可用性

云主机高可用

云主机支持设置高可用模式，当云主机因日常维护（计划）或突发异常（非计划）导致停机时，该策略可触发云主机自动重启，提高云主机可用性。

NeverStop云主机高可用机制：

- 通过轮询、触发等机制检测云主机状态，如果确定云主机已停止，设置高可用的云主机将直接自动重启。
- 通过轮询、触发等机制检测云主机状态，如果不能确定云主机状态，将根据以下步骤进行检测：

  1. 根据已有网络配置，选择最精准的方式探测云主机所在的物理机状态。
  2. 如果物理机状态异常，设置高可用的云主机将尝试自动重启。

负载均衡

多台云主机可使用负载均衡服务组成集群，消除单点故障，提升应用的可用性。

7.1.4 防IP/MAC/ARP欺诈

在传统网络里，IP/MAC/ARP欺骗一直是网络面临的严峻考验。通过IP/MAC/ARP欺骗，黑客可以扰乱网络环境，窃听网络机密。

在物理机数据链路层隔离由云主机向外发起的异常协议访问，并阻断云主机MAC/ARP欺骗，在物理机网络层防止云主机IP欺骗。

7.1.5 镜像与快照

镜像

支持对云主机/云盘创建镜像。云主机/云盘的数据信息完整包含在镜像中，通过镜像可快捷复制相应资源。

云平台支持对镜像的完整性和安全性保护：

- 镜像采用加密算法保护完整性。当镜像从镜像仓库下载至主存储时，需通过加密算法校验，只有校验成功才可下载镜像。
- 镜像文件以切片方式存储于镜像仓库，切片的镜像文件需通过云平台拼接完整后，才能读取具体内容，从而实现镜像数据的安全性保护。

快照

支持对云主机/云盘创建快照。快照实质为某一时间点某一磁盘的数据状态文件。做重要操作前，对云主机/云盘创建快照，可保留特定时间点的数据状态（包括内存状态），方便出现故障后迅速回滚。如需长期备份，建议使用灾备服务。

快照包括手动快照和自动快照两种类型：

- 手动快照：用户随时手动对云主机根云盘或数据云盘创建快照。
- 自动快照：通过定时任务创建快照，或系统在特定场景触发一次性自动快照。

快照功能适用于以下应用场景：

- 故障迅速还原：当生产环境出现异常故障，可使用快照回滚功能迅速还原至正常状态。该手段为临时方案，考虑到数据的长期完善保护，建议使用灾备服务。
- 数据开发：通过对生产数据创建快照，从而为数据挖掘、报表查询和开发测试等应用提供近实时的真实生产数据。
- 提高操作容错率：在系统升级或业务数据迁移等重大操作前，建议创建一份或多份快照。一旦升级或者迁移过程中出现任何问题，可以通过快照及时恢复到正常的系统数据状态。

7.1.6 密码加密存放

支持对云平台所有明文密码加密存放，从而保护用户数据的隐私性和自主性。

支持的密码加密存放场景包括但不限于：

- 物理机密码：非明文展示。
- 主存储密码：非明文展示。
- 数据库密码：通过密钥加密存放，直接对用户隐藏。
- 日志密码：云平台所有日志密码非明文展示或直接对用户隐藏。

7.1.7 资源删除保护

删除策略控制

支持对重要资源进行删除策略控制，降低误操作风险。

目前删除策略包括：立刻删除、延时删除、永不删除。

- 立刻删除：资源直接被物理删除，并在数据库中删除，无法恢复。

- 延时删除：资源首先在数据库中被标记为删除，但不会物理删除。在一定时间内，用户可通过UI界面的回收站功能或使用云平台API恢复资源。在此期间，资源仍物理存在，仍会占用物理空间（例如磁盘空间）。超过一定时间后，资源会被物理删除，无法再恢复。
- 永不删除：资源在数据库中标记为删除，永不会被物理删除，会一直占用物理空间。

目前支持删除策略控制的资源包括：云主机、云盘、镜像、裸金属主机、弹性裸金属实例。

- 云主机删除策略：立刻删除、延时删除、永不删除。默认为延时删除。
- 云盘删除策略：立刻删除、延时删除、永不删除。默认为延时删除。
- 镜像删除策略：立刻删除、延时删除、永不删除。默认为延时删除。
- 裸金属主机删除策略：立刻删除、延时删除。默认为延时删除。
- 弹性裸金属实例删除策略：立刻删除、延时删除、永不删除。默认为延时删除。

> **说明**：
>
> 弹性裸金属实例是一个定制的云主机，弹性裸金属实例与云主机受同一套删除策略控制。若云主机删除策略发生变更，则弹性裸金属实例删除策略同步变更。

UI删除提醒

在UI界面对重要资源删除提供保护机制，系统会提醒删除此资源的后果，并展示与此资源直接关联的云主机、云盘数量。用户需确认后才能进行删除，降低误操作风险。

7.1.8 国密数据保护

提供基于国密算法（SM3、HMAC-SM3和SM4）的数据保护功能，开启该功能后，可对日志、口令、镜像等重要数据进行加密保护，保护数据的机密性和完整性。

如需使用该功能，需确保已安装密评合规模块许可证并开启平台密评合规。

7.1.9 监控报警

主要通过监控系统以及通知系统提供监控报警功能，监控系统对时序化数据和事件进行监控，通知系统推送报警消息至指定的接收端。

通过监控系统提供包括系统性能、资源用量在内的监控数据指标，以大屏监控/仪表盘/可视化图表/横幅提示等形式，让用户全面了解云平台资源使用情况、系统运行状态以及健康度。用户还可自定义报警器以及接收端，实现细粒度灵活监控，及时发现并诊断相关问题。

监控系统功能特点：

- 时序化监控：目前支持监控两种时序化数据类型。
  - 资源负载数据：例如云主机CPU使用率、物理机内存使用率等。
  - 资源容量数据：例如可用IP数量、运行中云主机的总数量等。
- 事件收集：收集云平台中发生的预定义事件，例如物理机失联，云主机高可用功能启动等。
- 报警功能：对时序化数据或事件进行报警，并针对重要资源进行全局提示，例如主存储可用物理容量等。
- 审计功能：记录所有操作并提供搜索。
- 自定义功能：用户可自定义设置报警器和报警消息模板。

通知系统功能特点：

- 推送报警消息至指定的接收端。
- 系统默认提供一个系统类型接收端，用户可自行设置邮箱/钉钉/HTTP应用/短信/Microsoft Teams类型接收端。

7.1.10 安全场景封装

为安全场景提供一键全局设置封装，方便快速将云平台设置为所需状态，满足用户实际生产环境的安全需求。

表 7-1: 全局设置-安全场景封装

名称	描述
云平台登录IP黑白名单	默认为false，用于设置是否开启IP黑白名单功能，开启后，云平台将对登录IP进行防护。
物理机密码加密存储开关	默认为None，用于设置物理机密码在数据库中的加密存储策略。可选策略为：None、LocalEncryption。 • None：不进行加密存储。 • LocalEncryption：使用云平台自带的加密功能进行加密存储。 **说明**： 如已开启平台密评合规功能，“当前生效值”会变为SecurityResourceEncryption，表示使用密码机进行加密。此时不支持更换加密存储策略。

名称	描述
禁止同一用户多会话连接开关	默认为false，用于设置是否禁止同一用户多会话连接。若为true，则同一用户只能存在一个登录会话，历史会话将强制退出。
会话超时时间	默认为7200，单位为s/m/h/d（即：秒/分/小时/天）。 **说明**： 当前会话登录超过该会话时间后，系统将不可用，需重新登录。
SSL证书检查开关	默认为false，用于设置是否开启跳过LDAP SSL证书的所有检查的开关。若为true，表示跳过所有LDAP SSL证书的检查。
云平台登录验证码策略	默认为false，用于设置是否启用登录控制中的验证码功能。开启后连续登录失败次数超过上限将触发验证码保护机制，要求输入正确的账户名、密码以及验证码才能成功登录云平台。
云平台登录密码更新周期	默认为false，用于设置是否开启按周期修改密码功能。若设置为true，密码使用时间达到所设置的密码更新周期后，重新登录将提示修改密码。
云平台登录密码不重复次数	默认为false，若设置为true，则在重新设置密码时，新密码不能与之前已使用过的历史密码重复，不重复次数可配置。
云平台连续登录失败锁定用户	默认为false，用于设置是否启用连续登录失败锁定用户。若设置为true，则用户连续登录失败数次，账户会被锁定一段时间。
云平台登录密码强度	默认为false，若设置为true，则可以手动设置密码的长度和选择是否启用数字、大小写和特殊字符组合的策略。
云平台登录双因子认证开关	默认为false，登录云平台时，是否开启双因子认证。
VNC控制台密码强度	默认为false，用于设置是否启用密码登录VNC控制台。 **说明**：

名称	描述
	VNC密码长度范围格式为m~n，取值范围[6，8]的整数，默认为6~8，并支持选择是否启用数字、大小写和特殊字符组合的策略。
云主机密码强度	默认为false，用于设置是否启用密码登录云主机。 **说明**： • 云主机密码长度范围格式为m~n，取值范围[8，18]的整数，默认为8~18，并支持选择是否启用数字、大小写和特殊字符组合的策略。 • 设置云主机密码需确保云主机镜像中已安装cloud-init，且cloud-init推荐版本为：0.7.9、17.1、19.4、19.4以后版本。

7.1.11 持续数据保护（CDP）服务

以单独的功能模块形式提供持续数据保护（CDP）服务。CDP服务为云主机中的重要业务系统提供秒级细粒度的持续备份，即可以将云主机数据恢复到指定时间状态，又可以在不恢复系统的情况下找回文件。CDP恢复支持新建云主机和恢复到原云主机两种策略，用户可根据自身业务需求，灵活选择合适的恢复方式。

典型CDP场景：

- 本地CDP恢复 | 恢复到原云主机
  - 支持将本地部署的镜像仓库作为**本地备份服务器**，用于存放本地云主机数据。
  - 支持为多台云主机创建CDP任务，对批量云主机提供统一CDP保护。创建CDP任务时，支持秒/分钟级别的RPO设置。进行重要业务调整时，用户可对恢复点进行标记和锁定，长期保存重要恢复点数据。
  - 当发生本地数据误删，或突发故障导致数据损坏，由于用户的业务应用有硬件授权，为快速验证业务的可用性，可找到锁定恢复点，将数据恢复到原云主机查看应用是否正常。支持通过新建云盘方式恢复到原云主机，恢复前的云盘支持全部保留并重新加载回云主机，最大限度保证数据安全。
  - CDP恢复时，云主机会快速拉起，RTO最低可达到秒级，有效保障业务连续性。
- 本地CDP恢复 | 新建云主机

- 支持将本地部署的镜像仓库作为**本地备份服务器**，用于存放本地云主机数据。
- 支持为多台云主机创建CDP任务，对批量云主机提供统一CDP保护。创建CDP任务时，支持秒/分钟级别的RPO设置。
- 进行重要恢复演练时，可在不影响当前云主机正常运行的情况下，基于所选恢复点新建云主机，确认数据无误后再恢复到原环境中。
- CDP恢复时，云主机会快速拉起，RTO最低可达到秒级，有效保障业务连续性。

7.2 存储安全

7.2.1 基于角色访问控制

通过权限角色划分，保护系统访问安全。分布式存储系统支持创建多个帐户，分别赋予管理员和观察者两种身份权限，限制不同用户角色的访问权限，可通过访问日志查看每个用户的登录行为。基于登录密码可设置如下的安全管理策略：密码的登录错误次数、登录后锁定时间、密码的有效期设置。

7.2.2 传输安全

存储控制台支持SSL访问加密，在客户端和服务器端之间建立加密通道，保证数据在传输过程中不被窃取或篡改。

用户需将网站服务由HTTP转变成 HTTPS，可通过购买受信任CA认证中心颁发的数字证书，然后应用在存储平台，将HTTP访问转换成HTTPS，提供认证加密功能。

客户端和服务器端之间建立安全通信的流程：

1. 客户端访问使用HTTPS连接的资源。
2. 客户端产生一个唯一的会话密钥，并且使用服务器端证书生成的公钥加密传输。
3. 服务器端收到会话密钥后，用自己的私钥进行解密。
4. 连接建立后，客户端和服务器端之间就可安全通信。

7.2.3 访问安全

访问令牌（Access Token）是访问分布式存储系统API所需的认证密钥，请妥善保存。如果怀疑访问令牌泄露，请及时更换。

7.3 网络安全

7.3.1 安全组

为云主机提供三层网络安全组控制，控制TCP/UDP/ICMP等数据包进行有效过滤，对指定网络的指定云主机按照指定的安全规则进行有效控制。

7.3.2 防火墙

支持对VPC路由器配置防火墙。VPC防火墙创建后，系统为VPC路由器自动配置入方向规则集，用户可灵活配置出方向规则集。VPC路由器的每个接口方向允许应用一个规则集，通过对VPC路由器接口处的南北向流量进行过滤，可有效保护整个VPC的通信安全以及VPC路由器安全。与作用于云主机虚拟网卡、侧重于保护VPC内部东西向通信安全的安全组相辅相成。

7.3.3 VPC路由器高可用组

支持VPC路由器高可用组功能。可在一个VPC路由器高可用组内部署一对互为主备的VPC路由器，当主VPC路由器状态异常，会秒级触发高可用切换，自动切换至备VPC路由器工作，从而保障业务持续稳定运行。

7.3.4 Netflow

支持VPC路由器定向导出Netflow网络流分析监控。通过Netflow对VPC路由器网卡的进出流量进行分析监控，从而快速定位整个网络的流量瓶颈，优化网络拓扑以及网络带宽，防止恶意攻击，增强网络安全。目前支持Netflow V5、V9两种数据流输出格式。

7.3.5 端口镜像

支持端口镜像功能。将云主机网卡的出入流量转发至另一台云主机上，在不影响源端口正常业务吞吐的情况下，可获取云主机端口上的业务报文进行分析，方便企业对内部网络数据进行监控管理，快速定位网络故障。端口镜像需使用单独的流量网络，不与其它网络复用，确保传输效率。

7.4 权限管理安全

7.4.1 三员分立

支持三员分立权限管理，将超级管理员（admin）权限分解并赋予系统管理员、安全管理员和安全审计员。其中，系统管理员负责云平台资源管理、安全管理员负责云平台权限管理、安全审计员负责云平台审计管理，三者之间相互独立，相互制约。

三员分立将超级管理员的超级权限分解，由三员分而治之，可有效降低因超级管理员权限过大带来的安全风险，进一步加强云平台安全。

7.4.2 企业管理权限

以单独的功能模块形式提供企业管理功能。企业管理主要为企业用户提供组织架构管理，以及基于项目的资源访问控制、工单管理、独立区域管理等功能。

企业管理权限功能特点：

- 在企业管理中，用户与角色分离，角色作为一组权限的集合，可灵活绑定到企业管理用户或从企业管理用户解绑。
- 角色分为系统角色和自定义角色，系统角色是云平台默认提供的预定义权限范围的角色，自定义角色是用户按需自行创建的角色。
- 企业管理用户可在UI界面进行API级别的权限控制，灵活适配各种场景的权限配置需求。

7.4.3 国密证书登录

提供基于国密算法（SM2）的证书登录功能。开启该功能后，需使用UKey进行登录认证，确保身份的真实性。

如需使用该功能，需确保已安装密评合规模块许可证并开启平台密评合规。

支持为admin或租户开启证书登录。若需为租户开启证书登录，需确保云平台已安装企业管理模块许可证。

7.4.4 双因子认证

在静态密码认证基础上支持第二层防护：双因子认证。当云平台开启双因子认证后，每次登录均需正确输入身份验证器APP提供的6位动态安全码才能成功登录。

当开启双因子认证并首次成功登录后，不再展示登录二维码，有效防止恶意登录，进一步提升系统安全。

7.4.5 AccessKey认证

支持AccessKey认证功能。

AccessKey包括：

- 本地AccessKey：包括AccessKey ID和AccessKey Secret，是云平台授权第三方用户调用云平台API来访问云平台云资源的安全凭证，需严格保密。

- 第三方AccessKey：包括AccessKey ID和AccessKey Secret，是第三方用户授权用户调用第三方API来访问第三方云资源的安全凭证，需严格保密。

7.4.6 统一认证

支持标准单点登录（SSO）协议。通过企业管理/子账户对接第三方统一身份认证系统，对接完成后，第三方用户可通过认证系统单点登录云平台使用相关功能。

目前企业管理支持AD/LDAP/OIDC/OAuth2/CAS第三方认证，子账户支持OIDC第三方认证。在云平台上添加完成第三方认证服务器、并配置映射规则后，第三方用户信息将按照映射规则同步至云平台，并生成免密登录URL，第三方用户即可免密登录云平台。

7.4.7 操作审计

为用户提供统一的操作日志管理，记录云平台各类账号下的用户登录及资源操作，包括：操作描述、任务结果、操作员、登录IP、任务创建/完成时间，以及操作返回详情。通过操作日志审计，可满足用户进行安全分析、入侵检测、资源变更追踪以及合规性审计等需求。

8 开放兼容性

在服务器虚拟化技术中，云主机操作系统（Guest OS）运行在虚拟化内核层之上，虚拟化内核模拟出完整的主板芯片组、BIOS 和硬件I/O，Guest OS就承载在这些模拟出来的硬件上，因此，虚拟化内核必须满足Guest OS所需的所有硬件驱动程序。对于通用的服务器虚拟化内核软件，同时兼容Windows、Linux、UNIX等不同操作系统是一件非常困难的事情，尤其是随着开源NFV（例如：MikroTik、Quagga、VyOS、BIRD 等）的迅速涌现，这些开源NFV软件在虚拟化内核软件上的安装与部署首先就必须解决虚拟网络的兼容性问题与性能问题，而且这种兼容性问题的解决与处理一定是一个逐步优化改进的过程。

云计算的最终目的不是资源池化，而是提供敏捷的应用与服务，而应用与服务的核心就是安装在Guest OS内的应用软件。与Guest OS一样，很多应用软件也依赖于与硬件的配合，这类软件在启动阶段会对运行环境进行检测，如果发现是在虚拟化环境中，考虑到云主机提供的外设（USB 加密设备、显卡或网卡）不能很好地支持应用程序的表现，影响客户体验，软件开发商在程序中做了运行环境限制。针对应用软件与虚拟化软件的兼容性问题，需要虚拟化厂商与应用软件厂商从技术层面协同解决。

AliyunHCI-Z不断积累与主流软件厂商的兼容性开发与测试经验，在操作系统和应用软件的兼容性上不断投入适配力度，并以开放合作的姿态，与众多操作系统和应用软件厂商建立官方的互认证机制，具备良好软件兼容性。

术语表

云主机（VM Instance）

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

云盘（Volume）

为云主机提供存储，包括两种类型：根云盘、数据云盘。

根云盘（Root Volume）

云主机的系统云盘，用于支撑云主机的系统运行。

数据云盘（Data Volume）

云主机的数据云盘，用于云主机扩展的存储使用。

镜像（Image）

云主机或云盘使用的镜像模板文件，包括两种类型：系统镜像、云盘镜像。

计算规格（Instance Offering）

云主机涉及的CPU数量、内存、网络设置等规格定义。

云盘规格（Disk Offering）

云盘涉及的容量大小的规格定义。

GPU规格（GPU Specification）

物理GPU设备或vGPU设备涉及的GPU帧数、显存、分辨率等规格定义，包括两种类型：物理GPU规格、vGPU规格。

vNUMA配置（vNUMA Configuration）

通过CPU绑定透传关联的物理机NUMA节点（pNUMA Node）拓扑，为云主机生成vNUMA节点（vNUMA Node）拓扑，实现云主机CPU优先访问所在vNUMA节点本地内存，提高云主机性能。

NUMA（Non-Uniform Memory Access）

非一致性内存访问，是一种计算机内存设计架构。该架构下，CPU访问内存的时间取决于CPU与内存的相对位置。通过优先访问相对位置较近的内存可缩短延迟，从而可提升主机系统性能。

pNUMA节点（pNUMA Node）

基于物理机NUMA架构预定义的NUMA节点，用于物理机CPU和内存管理。

pNUMA拓扑（pNUMA Topology）

CPU厂商基于NUMA架构预定义的物理机NUMA节点拓扑。

vNUMA节点（vNUMA Node）

基于CPU绑定透传关联的物理机NUMA节点而生成的云主机NUMA节点，用于云主机CPU和内存管理。

vNUMA拓扑（vNUMA Topology）

基于CPU绑定生成的云主机NUMA节点（vNUMA Node）拓扑。

本地内存（Local Memory）

CPU（pCPU或vCPU）通过所在NUMA节点（pNUMA节点或vNUMA节点）非CPU核部件中内存控制器可直接访问的内存。相比非本地内存，CPU访问本地内存的延迟更低。

CPU绑定配置（CPU Pinning）

将云主机的虚拟CPU（vCPU）与物理机的物理CPU（pCPU）严格关联，可为云主机分配特定的pCPU，提高云主机性能。

EmulatorPin配置（EmulatorPin Configuration）

将云主机中除vCPU和IO线程外的其他线程与物理机pCPU进行绑定，使云主机相关线程只运行在对应的pCPU上。

弹性伸缩组（Auto Scaling Group）

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

快照（Snapshot）

某一时间点某一磁盘的数据状态文件。

云主机调度策略（VM Scheduling Policy）

为云主机分配物理机的资源编排策略，可用于保障业务的高性能和高可用。

区域（Zone）

云平台内最大的一个资源定义，包括：集群、二层网络、主存储等资源。

集群（Cluster）

一组物理机（计算节点）的逻辑集合。

物理机（Host）

为云主机实例提供计算、网络、存储等资源的物理主机。

主存储（Primary Storage）

用于存储云主机磁盘文件（包括：根云盘、数据云盘、根云盘快照、数据云盘快照、镜像缓存等）的存储服务器。

镜像服务器（Backup Storage）

用于存储云主机镜像模板（含ISO）的存储服务器。

iSCSI存储（iSCSI Storage）

基于iSCSI协议构建的SAN存储。用户可将iSCSI存储上划分的块设备添加为Shared Block主存储，或直接透传给云主机使用。

FC存储（FC Storage）

基于FC协议构建的SAN存储。用户可将FC存储上划分的块设备添加为Shared Block主存储，或直接透传给云主机使用。

二层网络（L2 Network）

对应于一个二层广播域，进行二层相关的隔离。一般用物理网络的设备名称标识。

VXLAN网络池（VXLAN Pool）

在一组VXLAN隧道端点（VTEP）之上创建的VXLAN网络的集合，同一个VXLAN网络池内VXLAN网络标识符（VNI）不能重复。

三层网络（L3 Network）

云主机使用的网络配置，包括IP地址范围、网关、DNS等。

公有网络（Public Network）

一般表示可直接访问互联网的网络，由于公有网络是一个逻辑概念，在无法连接互联网的环境中，用户也可以自定义该网络。

扁平网络（Flat Network）

可与物理机网络直通，也可直接访问互联网的网络。云主机可使用扁平网络提供的分布式EIP访问公有网络。

VPC网络（VPC Network）

云主机使用的私有网络，可通过VPC路由器访问互联网。

管理网络（Management Network）

管理控制云平台相关物理资源的网络，例如：配置访问物理机、主存储、镜像服务器、VPC路由器时使用的网络。

流量网络（Flow Network）

端口镜像的专用网络，用于将网卡的网络流量镜像到远端。

VPC路由器（VPC vRouter）

一个定制的云主机，用于提供多种网络服务。

VPC路由器高可用组（VPC vRouter HA Group）

一对互为主备的VPC路由器，当主VPC路由器状态异常，自动切换至备VPC路由器，为业务高可用提供保障。

路由器镜像（vRouter Image）

封装网络服务，用于创建VPC路由器/负载均衡实例，包括两种类型：VPC路由器镜像、高性能实例型负载均衡镜像。路由器镜像不能直接用于创建业务云主机。

VPC路由器镜像（VPC vRouter Image）

封装了多种网络服务，只能用于创建VPC路由器，不能直接用于创建业务云主机。

高性能实例型负载均衡镜像（LB Image）

封装了高性能实例型负载均衡服务，只能用于创建负载均衡实例，不能直接用于创建业务云主机。

路由器规格（vRouter Offering）

定义VPC路由器使用的CPU、内存、镜像、管理网络、公有网络配置，用于创建VPC路由器，为公有网络/VPC网络提供多种网络服务。

负载均衡实例规格（LB Offering）

定义负载均衡实例使用的CPU、内存、镜像、管理网络配置，用于创建负载均衡实例，为公有网络/扁平网路/VPC网络提供负载均衡服务。

SDN控制器（SDN Controller）

云平台支持添加外部SDN控制器来控制外部交换机等网络设备。

安全组（Security Group）

为云主机提供三层网络安全控制，控制TCP/UDP/ICMP等数据包进行有效过滤，对指定网络的指定云主机按照指定的安全规则进行有效控制。

虚拟IP（VIP）

在桥接网络环境中，使用虚拟IP地址提供弹性IP、端口转发、负载均衡、IPsec隧道等网络服务，数据包会被发送到虚拟IP，再路由至云主机网络。

弹性IP（EIP）

基于网络地址转换（NAT）功能，将一个网络的IP地址转换成另一个网络的IP地址，从而可通过其他网络访问内部私有网络。

端口转发（Port Forwarding）

基于VPC路由器提供的三层转发服务，可将指定公有网络的IP地址端口流量转发到云主机对应协议的端口。在公网IP地址紧缺的情况下，通过端口转发可提供多个云主机对外服务，节省公网IP地址资源。

负载均衡（Load Balancer）

将虚拟IP的访问流量分发到后端服务器上，自动检测并隔离不可用的后端服务器，从而提高业务的服务能力和可用性。

监听器（Listener）

负责监听负载均衡的前端请求，按照指定策略分发给后端服务器，且监听器会对后端服务器进行健康检查。

转发规则（Forwarding Rule）

将来自不同域名或者不同URL的请求转发到不同的后端服务器组处理。

后端服务器组（Backend Server Group）

一组负责处理负载均衡分发的前端请求的后端服务器。负载均衡实例进行流量分发时，流量分配策略以后端服务器组为单位生效。

后端服务器（Backend Server）

负责处理负载均衡分发的前端请求的服务器。支持添加云主机或云平台之外的服务器作为后端服务器。

前端网络（Frontend Network）

负载均衡的前端网络，负载均衡将来自该网络的客户端请求按照指定策略分发给后端服务器。

后端网络（Backend Network）

负载均衡的后端网络，负责处理负载均衡分发前端请求的后端服务器所在的网络。

负载均衡实例（Load Balancer Instance）

一个定制的云主机，专用于提供负载均衡服务。

证书（Certificate）

当负载均衡监听器使用HTTPS协议，需绑定证书使用。支持上传证书和证书链。

防火墙（Firewall）

在VPC网络场景下，负责管控经由VPC路由器的流量，通过配置规则集和规则管控网络的访问控制策略。

防火墙规则集（Firewall RuleSet）

防火墙规则的集合，包含了一组规则，需要绑定到VPC路由器网卡的某个方向上才能生效。

防火墙规则（Firewall Rule）

配置至防火墙用于控制VPC网络流量的访问策略，由规则优先级、匹配条件、以及行为三部分组成。

规则模板（Rule Template）

将一组规则保存为模板，向防火墙或规则集添加规则时可直接选用该模板。

IP/端口集合（IP/Port Set）

将一组IP或端口进行保存，在向防火墙或规则集添加规则时可直接选用已建好的IP/端口集合。

IPsec隧道（IPsec Tunnel）

通过对IP协议的分组加密和认证来保护IP协议的网络传输数据，实现站点到站点（Site-to-Site）的虚拟私有网络（VPN）连接。

OSPF区域（OSPF Area）

OSPF协议按照一定标准将一个自治系统划分为不同区域，用于分层管理路由器。

NetFlow

通过Netflow对VPC路由器网卡的进出流量进行分析监控，支持两种数据流输出格式：V5、V9。

端口镜像（Port Mirroring）

将云主机网卡的网络流量复制一份到远端，对端口上的业务报文进行分析，方便对网络数据进行监控管理，在网络故障时可以快速定位故障。

路由表（Route Table）

用户自定义配置路由信息，包括目标网段、下一跳地址、路由优先级。

资源编排（CloudFormation）

一款帮助云计算用户简化云资源管理和自动化部署运维的服务。通过资源栈模板，定义所需的云资源、资源间的依赖关系、资源配置等，可实现自动化批量部署和配置资源，轻松管理云资源生命周期，通过API和SDK集成自动化运维能力。

资源栈（Resource Stack）

资源编排通过资源栈模板快速创建和配置一组资源（以及资源间的依赖关系），这组资源定义为一个资源栈，通过管理资源栈，维护这组资源。

资源栈模板（Stack Template）

一个UTF8编码格式的文件，基于模板可快速创建资源栈，用户在模板中定义所需的云资源、资源间的依赖关系、资源配置等，资源编排将解析模板，自动完成所有资源的创建和配置。

资源栈示例模板（Sample Template）

云平台提供了常用的示例模板，用户可基于已有示例模板创建资源栈。

可视化编排（Designer）

一个可视化的资源编排工具，通过在画布上拖曳连线建立资源间的依赖关系，直观高效编排云资源。

裸金属集群（Bare Metal Cluster）

为裸金属设备提供单独的集群管理。

部署服务器（Deployment Server）

一台单独的服务器，为裸金属设备提供PXE服务和控制台代理服务。

裸金属设备（Bare Metal Chassis）

用于创建裸金属主机，通过BMC接口以及IPMI配置进行唯一识别。

预配置模板（Preconfigured Template）

通过预配置模板，可快速生成预配置文件，实现无人值守批量安装裸金属主机操作系统。

裸金属主机（Bare Metal Instance）

安装操作系统的裸金属设备。

弹性裸金属管理

不仅可为应用提供专属物理服务器，保障核心应用的高性能和稳定性，而且结合云平台中资源的弹性优势，可实现灵活申请，按需使用。

部署网络（Deployment Network）

创建弹性裸金属实例时，用于PXE流程及下载镜像的专属网络。

弹性裸金属集群（Bare Metal Cluster）

为裸金属节点提供单独的集群管理。

网关节点（Gateway Node）

云平台和弹性裸金属实例的流量转发节点。

裸金属节点（Bare Metal Node）

用于创建弹性裸金属实例，通过BMC接口以及IPMI配置进行唯一识别。

裸金属规格（Bare Metal Offering）

弹性裸金属实例涉及的CPU、内存、CPU架构、CPU型号等规格定义。

弹性裸金属实例（Elastic Bare Metal Instance）

性能媲美物理服务器的云实例，结合云平台中资源的弹性优势，可实现灵活申请，按需使用。

网络拓扑（Network Topology）

通过可视化方式展示云平台网络规划，帮助用户更高效地进行网络规划、管理和性能改进，支持两种类型：全局拓扑、自定义拓扑。

性能分析（Performance Analysis）

通过列表方式展示云平台核心资源的性能监控指标，提供外部和内部两种监控方式，支持按资源查看性能分析结果和自定义导出分析报表，方便用户掌控云平台性能状态，提高运维效率。

容量管理（Capacity Management）

通过可视化方式展示云平台核心资源的容量信息，方便用户掌控云平台容量使用情况，提高运维效率。

管理节点监控（MN Monitoring）

在多管理节点物理机高可用场景下，可直观查看每个管理节点的健康状态。

报警器（Alarm）

用于监控并响应时序性数据和事件的状态变化，支持资源报警器、事件报警器和扩展报警器。

一键报警（One-Click Alarm）

将种类繁多的资源监控项进行归纳整合，用于快速建立各种资源的监控报警服务。

报警模板（Rule Template）

一组报警器规则的通用模板，关联资源分组后，将对组内资源创建相应的报警器进行监控。

资源分组（Resource Group）

按照业务对资源进行分组，关联报警模板后，报警规则将直接作用于组内全部资源。

消息模版（Message Template）

报警器或事件向SNS系统的主题发送消息时使用的文本模板。

消息源（Message Source）

用于连接扩展消息源，接管扩展报警消息并结合报警器统一推送至各类通知对象。

通知对象（Endpoint）

用户获取订阅主题信息的方式，通知对象类型包括：系统、邮箱、钉钉、HTTP应用、短信、Microsoft Teams。

报警消息（Alarm Message）

报警器触发时发送的即时提示消息。

当前任务（Current Task）

展示当前正在进行中的操作，提供集中查看和管理。

操作日志（Operation Log）

云平台运行过程中变化的一种抽样，内容为指定对象的某些操作及其操作结果按时间的有序集合。

审计（Auditing）

实时监控并记录云平台的所有活动，可用于操作追踪、等保合规、安全分析、问题排查、自动运维等场景。

一键巡检（One-Click Inspection）

对云平台关键资源和服务进行全方位一键式健康检查，并根据巡检结果为巡检资源和服务进行健康评分，同时提供巡检建议和巡检报告，助力高效运维，确保云平台资源和服务处于最佳状态。

灾备管理

灾备管理以业务为中心，融合定时增量备份、定时全量备份等多种灾备技术到云平台中，支持本地灾备、异地灾备等多种灾备方案，用户可根据自身业务特点，灵活选择合适的灾备方式。

灾备服务

云主机数据在线备份到备份服务器，支持本地，跨地域和混合云多种备份场景，数据更可靠 。

备份任务（Backup Job）

通过备份任务，可将本地云主机/云盘/数据库定时备份到指定的存储服务器。

本地备份数据（Local Backup Data）

本地云主机/云盘/数据库的备份数据，存放在本地备份服务器中。

持续数据保护（CDP）

为云主机中的重要业务系统提供秒级细粒度的持续备份，既可以将云主机数据恢复到指定时间状态，又可以在不恢复系统的情况下找回文件。

CDP任务（CDP Task）

通过CDP任务，可将云主机数据持续备份到指定的备份服务器，实现持续数据保护。

CDP数据（CDP Data）

对云主机进行持续数据保护产生的备份数据，存放在指定的备份服务器中。

恢复点

进行CDP持续备份时产生的数据点，一个恢复点对应用户指定时间间隔内的云主机数据变化。

锁定恢复点

支持将恢复点进行锁定或解锁，锁定恢复点后，恢复点对应的数据将不会被自动清理或删除。

恢复任务

提供向导式的恢复流程，帮助用户通过指定CDP任务和恢复点快速进行数据恢复。同时提供恢复操作记录，方便后续审计和追溯。

密评合规

通过建立以商用密码为核心的云安全保障体系，满足企业在密评场景的云平台合规要求。

密码资源

提供密码服务的资源，包括密码机资源（密码机资源池/密码机）和密码服务资源（第三方密码服务）。

第三方密码服务

支持添加第三方密码服务平台，用于对外提供验签、加密等密码服务。

密码机资源池

一组密码机的逻辑集合，对外提供统一的验签、加密等密码服务。

密码机

运用密码技术对信息实施加解密处理和认证的专用设备。

平台密评合规

通过密码机资源池提供的密码能力，满足云平台密评合规要求。

证书登录

可通过UKey设备对用户进行身份标识和鉴别，保证用户身份的真实性。

数据保护

可对云平台重要数据进行保护，保证数据的机密性和完整性。

定时任务（Scheduled Job）

一种预设任务，在指定时间执行指定行为。与定时器配合使用。

定时器（Scheduler）

承载定时任务的容器，尤其适用于长时间运行的操作。

标签（Tag）

一种资源标记，便于快速搜索和资源聚合。

迁移服务（Migration Service）

提供V2V迁移服务，可将其它虚拟化平台的云主机系统及数据完整迁移至当前云平台。

V2V迁移（V2V Migration）

将VMware或KVM源云平台的云主机迁移至当前云平台。

迁移服务器（V2V Conversion Host）

V2V迁移需指定目标集群内的一台物理机作为迁移服务器，云主机系统和数据先缓存在迁移服务器中，再导入目标主存储。

用户（User）

表示自然人，是企业管理中的最基本单位。

成员组（Member Group）

有双重含义，既表示一组自然人的集合，也表示一组项目成员的集合，支持以成员组为单位进行权限控制。

角色（Role）

权限的集合，为用户和成员组赋予权限可获得调用相关API进行资源操作的能力。包括平台角色和项目角色两类。

第三方认证（3rd Party Authentication）

云平台提供的第三方登录认证服务，支持无缝接入第三方登录认证系统，相应第三方用户将直接登录云平台，便捷使用云资源。

项目（Project）

项目是租户的一种，指在特定时间、资源、预算下指定相关人员完成特定目标的任务。企业管理以项目为导向进行资源规划，可为一个具体项目建立独立的资源池。

项目成员（Project Member）

项目的基本组成人员，可在权限范围内使用项目资源完成任务目标，包括项目负责人、项目管理员和普通项目成员。

流程管理（Process Management）

为了更高效对项目提供基础资源支持，工单审批引入流程管理，包括默认流程和自定义流程两种类型。

我的审批（My Approvals）

仅admin/项目负责人/自定义审批人员拥有审批权限，可通过或驳回申请。若审批通过，资源会自动部署并下发至项目生效。

账单（Bills）

按计费价目在指定时间段统计的资源费用，计费精确至秒级。

计费价目（Pricing List）

一张包含不同资源计费单价的集合表，资源单价基于资源规格、资源使用时间而设置。

控制台代理（Console Proxy）

可通过代理地址登录云主机控制台。

AccessKey管理（AccessKey Management）

访问云平台API的身份凭证，具有该云平台完全的权限，包括：AccessKey ID（访问密钥 ID）和AccessKey Secret（秘密访问密钥）。

IP黑白名单（IP Blocklist/Allowlist）

云平台登录IP的黑白名单，通过对访客身份的识别和过滤，进一步提升云平台访问控制安全。

应用中心（Application Center）

提供云平台增强功能以及第三方应用的快速访问。

子账户管理（Sub-Account Management）

子账户是租户的一种，由admin创建或第三方认证系统同步，且受admin管理，子账户对自己创建的虚拟资源拥有管理权限。

主题外观（Theme）

用户可自定义设置云平台主题外观。

邮箱服务器（Email Server）

云平台报警选择邮箱类型的通知对象，需设置邮箱服务器，用来发送报警邮件。

日志服务器（Log Server）

添加日志服务器至云平台，可用于收集管理节点日志，快速定位问题，提高云平台运维效率。

全局设置（Global Setting）

提供平台层面的功能特性设置，一经设置，云平台全局范围内生效。

场景封装（Scenario Template）

基于用户实际生产场景需求，提供场景化的一键全局设置，方便快速将云平台设置为所需状态，提高运维效率。

API Inspector

云平台API调用的监视工具，用于记录用户在云平台上操作时所关联调用的API。