UCLOUD优刻得

中国第一家公有云科创版上市公司

股票代码：688158

UCloudStack HCI 超融合 产品白皮书

优刻得私有云

构建下一代可持续云基础设施

赋能企业未来

文档更新：2024 年 02 月 06 日

版权信息

关于文档

优得刻科技股份有限公司在编写本文档时已尽最大努力保证其内容准确可靠，但优得刻科技股份有限公司不对本文本中的遗漏、不准确或错误导致的损失和损害承担责任。

由于产品版本升级或其它原因，本文档内容会不定期更新，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

前言

UCloud（优刻得科技股份有限公司）是中立、安全的云计算服务平台，坚持中立，不涉足客户业务领域。公司自主研发 IaaS、PaaS、大数据流通平台、AI 服务平台等一系列云计算产品，并深入了解互联网、传统企业在不同场景下的业务需求，提供公有云、私有云、混合云、专有云在内的综合性行业解决方案。

依托公司在莫斯科、圣保罗、拉各斯、伦敦等全球部署的 32大高效节能绿色数据中心，以及国内北、上、广、深、杭等 11 地线下服务站，UCloud 已为全球上万家企业级客户提供云服务支持，间接服务终端用户数量达到数亿人。UCloud 深耕用户需求，秉持产品快速定制、贴身应需服务的理念，推出适合行业特性的产品与服务，业务已覆盖包含互联网、金融、新零售、制造、教育、政府等在内的诸多行业。

公司核心团队来自腾讯、阿里、百度、华为、VMware 等国内外知名互联网和 IT 企业，同时引进传统金融、医疗、零售、制造业等行业精英人才，目前员工总数超过 1000人。

在云计算之前，企业业务应用上线，需要经历组网规划、容量规划、设备选型、下单、付款、发货、运输、安装、部署、调试的整个完整过程。随着云计算、大数据、人工智能等新技术对各行各业的赋能，以虚拟化和软件定义技术方向构建新一代数据中心为基础，实现业务的集中管理、资源动态调配、业务快速部署及统一运营运维，满足企业关键应用向 x86 及国产化系统迁移时，对资源高性能、高可靠、安全性、高可用及易用性上的要求，同时提高基础架构的自动化管理水平及业务快速交付能力，继而推动企业的数字化转型与业务创新。

UCloud 提供计算、存储、网络、数据库、中间件、云分发、多媒体音视频、大数据、人工智能及云安全防护等产品服务，为多种业务场景提供全方位的公有云计算产品和服务。为帮助企业级客户实现数字化转型，快速构建新一代基于虚拟化的私有数据中心，优刻得私有化解决方案，构建下一代可持续云基础设施，赋能企业未来。通过构建基于公有云且自主可控的下一代云基础设施，提供超融合、私有云、分布式存储及智能大数据平台等“底座”，凭借多年公有云运营经验

和解决方案能力，助力企业数字化转型。同时依托客户实践，提供超融合和信创架构满足敏捷交付、降低成本和信创转型等需求。

在私有化层面，UCloud 基于多年运营的公有云架构，复用服务器操作系统内核及虚拟化核心组件，同时优化核心调度系统及管理平台，对公有云平台进行定制重构，缩减部署规模及业务开销，提高超融合平台的可靠性及可用性，提供一套可私有化交付且与公有云一致稳定性的超融合平台——UCloudStack HCI。

通过 UCloudStack HCI 超融合平台，企业可在现有数据中心及设备上快速构建一套成熟且完整的超融合解决方案。提升组织管理和业务管理效率的同时，降低业务转型及信息系统的总体拥有成本，助力企业数字化转型。

1 产品简介

1.1 产品概述

UCloudStack HCI 超融合平台，提供计算虚拟化、分布式存储、智能资源调度、监控日志、账户权限等统一管理能力，助力企业数字化转型。

平台基于 UCloud 公有云基础架构，复用内核及核心虚拟化组件，具有自主可控、稳定可靠、持续进化及开放兼容等特点，企业可通过控制台或 API 快速构建资源及业务，帮助企业快速构建安全可靠的业务架构。

UCloudStack HCI 定位为轻量级交付，3 节点即可构建生产环境且可平滑扩容，兼容 X86 和 ARM 架构，并提供统一资源调度和管理，支持纯软件、超融合一体机交付模式，有效降低用户管理维护成本，为用户提供一套安全可靠且自主可控的超融合平台。

1.2 核心优势

- **超十年云计算技术沉淀**

超 10 年公有云运营成熟经验和技术积累，从海量项目实践中积累了产品及解决方案能力。核心组件历经上万家企业级客户大规模的磨炼和验证，确保产品稳如磐石。

- **中立安全自主可控**

中立、安全的超融合平台，复用核心虚拟化组件，自主研发，工信部权威机构代码溯源检测，核心代码自研率达 96%以上，为业界领先水平，可控性高且可靠性经上万家企业验证。

- **技术架构轻量灵活**

轻量灵活且先进的技术架构，最小 3 个节点即可构建生产环境，并可平滑扩容至数千、数万节点。资源规划尽在掌控，轻松应对业务增长挑战。

- **稳定可靠开放兼容**

平台服务高可用，虚拟资源智能调度，数据存储多副本，自愈型分布式网络，为业务保驾护航。兼容 X86、ARM、龙芯、申威等架构及生态适配，设备异构搭建统一管理。

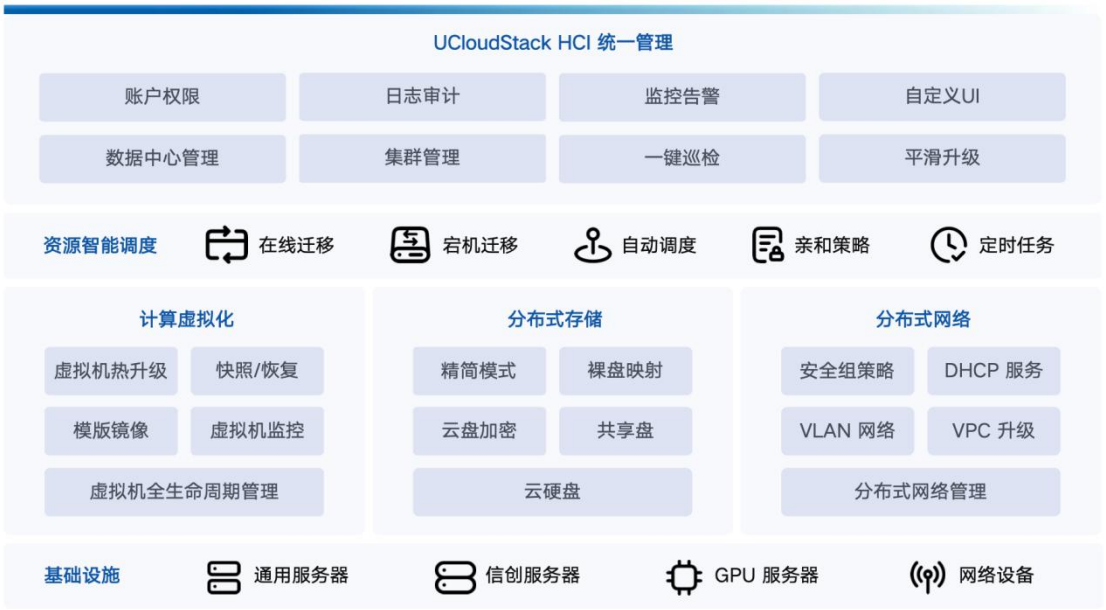
- **全方位的服务保障**

1 对 1 金牌服务，精准匹配客户需求，量身打造数字转型方案，以最佳实践协助企业轻松转型，为项目提供全生命周期服务保障，助力企业数字化建设。

- **众多用户可信赖选择**

适配兼容 100+ 信创生态软硬件，全面满足企业数字化改造等多场景业务需求。累计服务政府、金融、教育、制造等 14+ 行业、400+上市企业，成为全球 50000+用户及伙伴可信赖的数字转型服务商。

1.3 产品架构



UCloudStack HCI 超融合平台整体架构由基础硬件设施、虚拟核心引擎、智能调度系统、统一管理平台组成，为平台管理员及运营人员提供统一管理和服务。

(1) 基础设施

用于承载 UCloudStack HCI 平台的服务器、交换机及存储设备等。

- 平台支持并兼容通用 X86、ARM 架构硬件服务器，不限制服务器和硬件品牌；
- 支持 SSD、SATA、SAS 等磁盘存储，同时支持对接磁盘阵列设备，无厂商锁定；
- 支持华为、思科、H3C 等通用交换机、路由器网络设备接入，仅需物理交换机支持 Vlan、Trunk、IPv6、端口聚合、堆叠等特性；

(2) 虚拟核心引擎

承载平台核心的操作系统内核、虚拟化计算、存储、网络的实现和逻辑。

- **内核模块**：承载平台运行的服务器操作系统及内核模块，复用公有云深度优化的 Linux 内核；同时兼容 UOS、银河麒麟等国产服务器操作系统；
- **虚拟化计算**：通过 KVM、Libvirt 及 Qemu 实现计算虚拟化，支持标准虚拟化架构，提供虚拟机全生命周期管理，兼容 X86 和 ARM 架构体系，支持热升级、重装系统、CPU 超分、GPU 透传、在线迁移、宕机迁移、反亲和部署等特性，并支持导入导出虚拟机镜像满足业务迁移需求；
- **分布式网络**：无集中网络转发节点，所有生产网络仅在计算节点上传输，无需通过管理服务或网络节点进行转发，避免集中网络瓶颈。通过 OVS 虚拟交换机实现扁平网络，通过 VLAN 实现多个扁平网络之间的隔离及外部物理网络的打通。通过对集群内所有虚拟交换机的统一配置管理，为虚拟机提供简单高效、集群内配置一致的网络连接。
- **分布式存储 SDS**：基于 Ceph 实现分布式高性能存储，为平台提供块存储服务，支持虚拟磁盘在线扩容、克隆、快照及回滚功能；同时底层数据多副本存储并支持数据重均衡和故障重建能力，保证性能和数据安全性。

(3) 智能调度系统

- 支持亲和反亲和性调度部署策略，保证业务的高可用性和高可靠性；
- 支持在线迁移技术，实时感知物理机状态和负载信息；

- 物理主机故障或超过负载时，自动迁移虚拟机至低负载物理主机；
- 创建虚拟机时，自动调度至低负载、健康的物理主机；

(4) 核心产品资源

- **地域（数据中心）：**数据中心指资源部署的物理位置分类，数据中心之间相互独立，如无锡数据中心、上海数据中心等。平台支持多数据中心管理，使用一套平台管理遍布各地数据中心的超融合平台；
- **集群：**用于区分不同资源在一个数据中心下的分布情况，如 x86 计算集群、ARM 计算集群、SSD 存储集群及 SATA 存储集群，一个数据中心可以部署多个集群；
- **弹性计算：**运行在物理主机上的虚拟机，支持从镜像创建、重启/关机/启动、删除、VNC 登陆、重装系统、重置密码、热升级及安全组、挂载数据盘及反亲和策略部署等虚拟机全生命周期功能，同时支持将虚拟机制作为镜像及磁盘快照能力，提供快捷的业务部署及备份能力；
- **GPU 虚拟机：**平台提供 GPU 设备透传能力，支持用户在平台上创建并运行 GPU 虚拟机，让虚拟机拥有高性能计算和图形处理能力；
- **镜像：**虚拟机运行时所需的操作系统，提供 CentOS、Windows、Ubuntu 等常用基础操作系统镜像；支持将虚拟机导出为镜像，通过自制镜像重建虚拟机；支持镜像的导入导出，便于用户自定义镜像；支持导入 ISO 镜像文件，方便用户创建自定义操作系统镜像。
- **虚拟硬盘：**一种基于分布式存储系统为虚拟机提供持久化存储空间的块设备。具有独立的生命周期，支持随意绑定至不同虚拟机使用及解绑，基于网络分布式访问，并支持容量扩容、克隆、快照等特性，为虚拟资源提供高安全、高可靠、高性能及可扩展的磁盘；
- **快照：**提供磁盘快照及快照回滚能力，可应用于容灾备份及版本回退等业务场景，降低因误操作、版本升级等导致的数据丢失风险；
- **扁平网络：**为虚拟机提供的多个独立、安全的二层网络环境。虚拟机在

同一扁平网络内可以直接通信，不同扁平网络的虚拟机则通过 VLAN 隔离，保障网络的安全性；通过采用一致性网络配置和分布式虚拟交换机，为虚拟机提供可靠、简单、易用、一致的网络环境；

- **安全组：**虚拟防火墙，提供出入双方向流量访问控制规则，定义哪些网络或协议能访问资源，用于限制虚拟资源的网络访问流量，支持 TCP、UDP、ICMP 及多种应用协议，为平台提供必要的安全保障；
- **监控告警：**支持虚拟机、磁盘等资源各维度监控数据收集及展示，同时可通过告警模板快速配置资源监控指标的告警策略和通知规则；
- **操作日志：**平台所有资源及平台自身的操作和审计日志，支持多时间跨度的日志收集和展示，提供操作失败原因；
- **回收站：**资源删除后暂存的位置，支持回收资源、恢复资源及彻底删除资源等操作；
- **隔离组：**隔离组是一种针对虚拟机的简单编排策略，支持组内或组之间的实例分散到不同物理机上，用以保障业务的高可用。

(5) 统一管理平台

- 平台提供 Web 控制台和 API 接口两种方式接入和管理平台；
- 通过 WEB 控制台用户可快捷的使用并管理平台资源，如虚拟机、虚拟硬盘等；
- 开发者可通过 API 自定义构建平台资源，支持无缝迁移。

(6) 运维管理平台

为平台管理员提供的运维运营管理平台，包括平台管理账号、多地域、全局资源视图、物理资源管理、虚拟资源管理、QoS 配置管理、资源模版、标签管理、监控告警、通知组、操作日志、资源事件、回收站、报表统计、大屏监控等功能模块。

- **平台管理账号：**支持创建多个管理员账号，根据地域的授权范围分为系

统管理员和地域管理员。

- **多地域：**多地域管理支持在多个数据中心部署超融合平台，多套平台通过统一管理界面进行运维运营管理，用户可在统一控制台使用多个地域下的资源。
- **全局资源视图：**支持全局资源管理视图，针对多地域资源使用情况提供总体视图展示，支持查看单地域和全部地域资源的使用情况。
- **物理资源管理：**支持对平台物理资源进行管理和调度，包括物理数据中心、集群、节点资源、存储资源、网络 IP 网段资源池等。
- **虚拟资源管理：**平台为管理员提供虚拟资源全生命周期运营和管理能力，包括虚拟机、虚拟硬盘、监控告警、安全组、回收站等，使平台管理员可通过控制台统一管控平台的虚拟资源。
- **QoS 配置管理：**支持全局虚拟硬盘 QoS 配置，新建虚拟硬盘会根据平台算法赋予默认 QoS 值，同时支持管理员对平台所有虚拟硬盘自定义设置 QoS 值。
- **资源模版：**支持用户预定义创建资源参数配置，保存到模版中，便于后续快速创建，可以通过指定机型、规格、镜像、虚拟硬盘、网络、安全组等相关基础信息一键创建虚拟机模板，用于从模板创建虚拟机实例。
- **标签管理：**支持通过标签标记各项虚拟资源，从不同维度对具有相同特征的资源进行分类、搜索和聚合，让资源管理变得更加方便。
- **监控告警：**提供平台物理设备、组件及所有虚拟资源的监控数据，并支持自定义监控报警和通知。
- **通知组：**通知组包含一个或多个联系人，在资源发生告警时通过所设置的通知方式至所有通知人，支持邮件及 Webhook 通知方式。
- **操作日志：**提供操作日志用于记录用户通过控制台或 API 对资源进行的操作行为及登录登出平台的审计信息。支持操作记录查询及筛选，通过操作日志可实现安全分析、资源变更追踪以及合规性审计。

- **资源事件：**资源事件记录用户在资源类型的部分核心操作事件，提供事件详细记录查询及筛选，并可配合通知规则及时通知用户、定位问题。
- **回收站：**支持资源回收站，为平台删除资源提供暂时保留区，用户删除的虚拟机、磁盘、自制镜像等资源，会在删除后自动进入回收站中。
- **巡检服务：**支持一键巡检扫描检查平台健康情况，将对平台管理节点、计算节点的资源使用情况进行扫描并展示当前设备参数现状及针对性的建议，便于用户进一步处理。
- **报表统计：**支持资源用量统计及资源统计表，将各种数据整理成易于理解和分析的形式，提高平台整体运营和管理效率。
- **大屏监控：**支持资源可视化大屏，展示平台宏观维度的监控数据，帮助管理员快速了解平台的整体运行情况。

1.4 技术架构特性

1.4.1 API 幂等性

幂等性是指一次和多次请求某一个资源应该具有同样的作用，保证资源请求无论调用多少次得到的结果始终一致。如多次调用更新虚拟机的 API 请求，返回的结果都是一致的。

平台通过分布式锁、业务字段唯一约束及 Token 唯一约束等技术手段保证平台资源 API 幂等性。对虚拟机、虚拟硬盘等资源的操作请求（除创建请求）均支持重复提交，并保证多次调用同一个 API 请求返回结果的一致性，同时避免网络中断导致 API 未能获取确切结果，从而导致重复操作的问题。

1.4.2 全异步架构

平台使用消息总线进行服务通信连接，在调用服务 API 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，即触发回调函数回复任务结果。

平台服务之间和服务内部均采用异步调用方法，通过异步消息进行通信，并结合异步 HTTP 调用机制，保证平台所有组件均实现异步操作。

基于异步架构机制，平台可同时管理数十万以上的虚拟机及虚拟资源，后端系统每秒可并发处理上万条 API 请求。

平台采用的插件机制，每个插件设置相应的代理程序，同时在 HTTP 包头为每个请求设置回调 URL,插件任务结束后，代理程序发送应答给调用者的 URL。

1.4.3 分布式

(1) 分布式底层系统

平台核心模块提供计算、存储及调度等分布式底层支持，用于智能调度、资源管理、安全管理、集群部署及集群监控等功能模块。

- **智能调度**

基于分布式和远程服务调用为用户提供智能调度模块。智能调度模块实时监测集群和所有服务节点的状态和负载，当某集群扩容、服务器故障、网络故障及配置发生变更时，智能调度模块将自动迁移被变更集群的虚拟资源到健康的服务器节点，保证平台的高可靠性和高可用性；

- **资源管理**

通过分布式资源管理模块，负责集群计算、存储、网络等资源的分配及管理，为平台用户提供资源申请、资源调度、资源占用及访问控制，提升整个集群的资源利用率；

- **安全管理**

分布式底层系统提供安全管理模块，为用户提供身份认证、授权机制、访问控制等功能。通过 API 密钥对和用户名密码等多种方式进行服务间调用及用户身份认证；通过角色权限机制进行用户对资源访问的控制；通过安全组对资源网络进行访问控制，保证平台的安全性；

- **集群部署**

分布式底层系统为平台提供自动化部署集群节点的模块，为运维人员提供集群部署、配置管理、集群管理、集群扩容、在线迁移及服务节点下线等功能，为平台管理者提供自动化部署通道；

- **集群监控**

监控模块主要负责平台物理资源和虚拟资源信息收集、监控及告警。监控模块在物理机及虚拟资源上部署 Agent，获取资源的运行状态信息，并将信息指标化展示给用户；同时监控模块提供监控告警规则，通过配置告警规则，对集群的状态事件进行监控及报警，并有效存储监控报警历史记录；

(2) 分布式存储系统

平台采用高可靠、高安全、高扩展、高性能的分布式存储系统，提供块存储服务，保证本地数据的安全性和可靠性。

- 软件定义分布式存储，将大量通用机器的磁盘存储资源聚合在一起，采用通用的存储系统标准，对数据中心的所有存储进行统一管理；
- 分布式存储系统采用多副本数据备份机制，写入数据时先向主副本写入，由主副本负责向其他副本同步数据，并将每一份数据的副本跨磁盘、跨服务器、跨机柜、跨数据中心分别存储于不同磁盘上，多维度保证数据安全；
- 多副本机制存储数据，将自动屏蔽软硬件故障，磁盘损坏和软件故障，系统自动检测到并自动进行副本数据备份和迁移，保证数据安全性，不会影响业务数据存储和使用；
- 分布式存储服务支持水平扩展、增量扩容及数据自动平衡性，保证存储系统的高扩展性；
- 支持 PB 级存储容量，总文件数量可支持亿量级；
- 支持不间断数据存储和访问服务，保证存储系统的高可用性；
- 支持高性能虚拟硬盘，IOPS 和吞吐量随存储容量规模线性增长，保证响应时延；

在部署上，可将 SSD 磁盘构建为高性能的存储池，SATA/SAS 磁盘构建为普通性能存储池。分布式存储系统将块设备内建为弹性块存储，可供虚拟机直接挂载使用，在数据写入时通过三副本、写入确认机制及副本分布策略等措施，最大限度保障数据安全性和可用性。在本地可通过快照技术，将本地数据定时备份，在数据丢失或损坏时，可通过快照快速恢复本地业务的数据。

1.4.4 高可用

平台架构从硬件设施、网络设备、服务器节点、虚拟化组件、分布式存储均提供高可用技术方案，保证整个平台业务不间断运行：

- 数据中心机柜级别冗余性设计，所有设备均对称部署于机柜，单机柜掉电或故障不影响业务；
- 网络设备扩展性设计，所有网络设备分为核心和接入两层架构，一套核心可水平扩展几十套接入设备；
- 网络设备冗余性设计，所有网络设备均为一组两台堆叠，避免交换机单点故障；
- 交换机下联接入冗余性设计，所有服务器双上联交换机的接口均做 LACP 端口聚合，避免单点故障；
- 服务器网络接入冗余性设计，所有服务器节点均做双网卡绑定，避免单点故障；
- 管理节点冗余性和扩展性设计，多台管理节点均为 HA 部署，并支持横向扩展，避免管理节点单点故障；
- 通过智能调度系统将虚拟机均衡部署于集群节点，可水平扩展集群节点数量；
- 分布式存储冗余性设计，将数据均衡存储于所有磁盘，并多副本、写确认机制及副本分布策略保证数据安全；
- 进行服务器节点及存储扩展时，只需增加相应数量的硬件设备，并相应

的配置资源调度管理系统；

- 平台内各组件均采用高可用架构设计，如管理服务、调度服务等，保证平台高可用；

1.4.5 组件化

平台将所有虚拟资源组件化，支持热插拔、编排组合及横向扩展。

- 组件化包括虚拟机、磁盘、安全组等；
- 组件支持热插拔，如将一个安全组绑定至一个在运行中的虚拟机；
- 组件支持横向扩展，如增加虚拟机的磁盘，提升整体平台的健壮性。

1.5 应用场景

1.5.1 虚拟化

通过将业务系统和内部应用部署至 UCloudStack HCI 平台，可为用户提供一套集虚拟化、分布式存储为一体的超融合平台。平台支持多数据中心管理，可将业务部署至多个数据中心构建 IT 基础设施灾备机制，帮助政企快速构建安全可靠的业务架构。

1.5.2 业务快速交付

平台服务所见即所得，可通过统一管理平台一键部署并管理业务交付所需的基础设施；同时平台支持镜像导入导出，可方便快捷将业务系统迁移至平台，并可对所有业务系统的资源进行统一管理。

1.5.3 超融合一体机

平台提供一体机交付模式，多款机型应用于不同业务场景，集成 UCloudStack HCI 超融合平台，出厂预装开箱即用，服务模块热插拔可按需部署，提供虚拟化、网络、存储的统一管理能力。

1.6 交付和服务

UCloudStack HCI 定位为轻量级交付，3 节点即可构建生产环境且可平滑扩容，并提供统一资源调度和管理，支持纯软件、超融合一体机及超融合机柜多种交付模式，有效降低用户管理维护成本，为用户提供一套安全可靠且自主可控的服务平台。

- **纯软件交付**

客户提供符合兼容性要求的硬件服务器、网络设备及相关基础设施，UCloud 优刻得提供 UCloudStack HCI 超融合软件；通常在基础网络设施环境完备的情况下，软件可在 2 小时内完成部署并交付。

- **超融合一体机**

客户仅需提供数据中心基础设施，UCloud 优刻得提供超融合一体机（出厂预装 UCloudStack HCI），通常在基础网络设施环境完备的情况下，可在小时内完成初始化并交付。

- **超融合机柜**

客户仅需提供数据中心即可，UCloud 优刻得提供超融合一体机柜（包含网络设备、服务器节点&一体机、PDU、线缆及 UCloudStack HCI 软件），通常以一个机柜的形式进行交付。

在服务方面，提供全面服务保障体系，用户可根据业务场景和需求，自主选择适合的服务，如基础质保服务、高级质保服务、金牌质保服务、续保服务、培训服务及驻场服务；并可提供一对一专家的方案咨询、架构设计、业务迁移、巡检调优等。

2 平台物理架构

UCloudStack HCI 超融合平台服务主要包含管理服务、计算服务、存储服务。

- **管理服务**

平台核心管理服务，负责虚拟资源全生命周期的管理；同时承载平台的北向接口服务，包括帐户认证、资源管理、网关及监控等服务，提供标准 API 和 WEB 控制台两种接入和管理方式。

- **计算服务**

通过 KVM 和 Qemu 等 Hypervisor 组件及技术，将物理服务器计算资源进行虚拟和池化，将 CPU、内存等物理资源转化为一组可统一管理、调度和分配的逻辑资源，并基于虚拟机在物理机上构建多个同时运行、相互隔离的虚拟机执行环境。

- **存储服务**

通过将通用服务器的磁盘存储资源融合池化，构建可伸缩的统一分布式存储集群，实现对存储资源的统一管理及调度，向计算服务提供块存储接口，供平台虚拟机自由分配并使用存储资源池中的存储空间。

面向不同的需求场景，平台支持灵活的部署架构，企业可以根据其特定的业务需求选择超融合模式或存算分离模式，中小型企业和大型企业都能够根据其具体规模选择适当的部署架构，以实现更灵活的部署。

- **超融合架构**

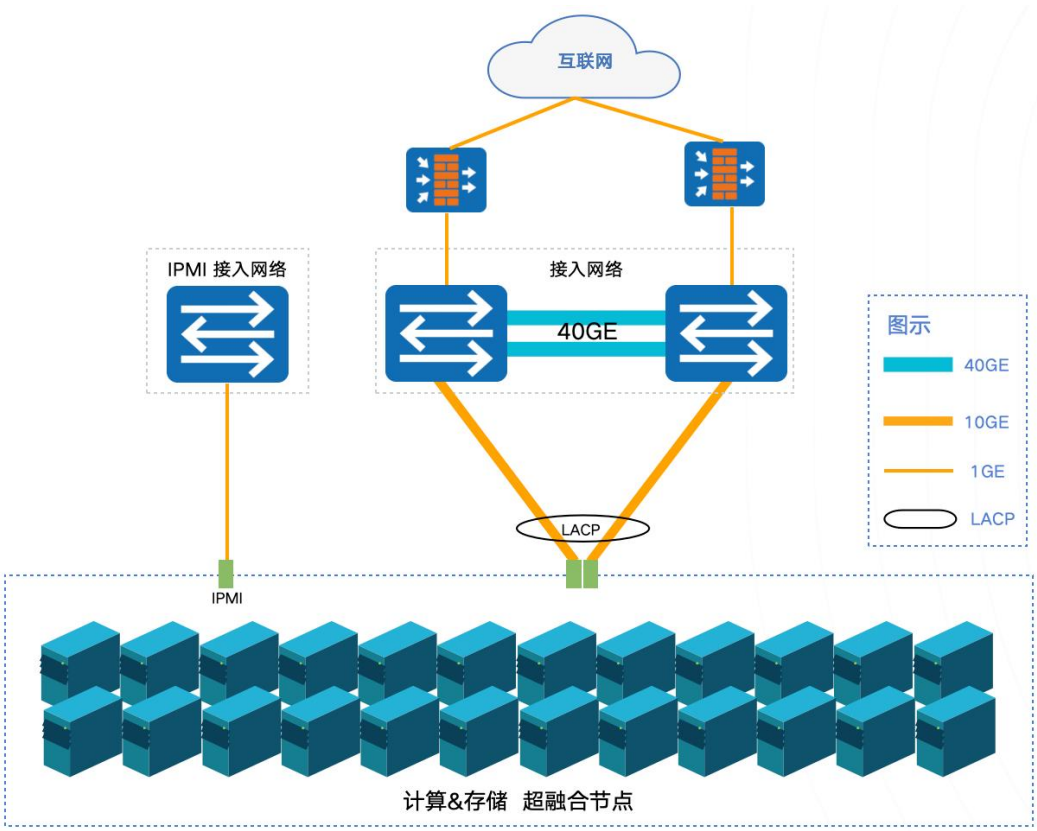
计算存储融合节点将计算、存储和网络等基础设施组件整合到同一硬件上，通过软件定义实现整个基础设施的自动化管理，降低了复杂性，使得企业可以更轻松地部署和管理整个基础设施，从而简化管理、提高灵活性。

- **存算分离架构**

面向大规模场景，平台支持存算分离部署，允许计算和存储分别部署在不同的硬件节点上。这种架构可以更好地满足大规模数据处理和存储的需求，支

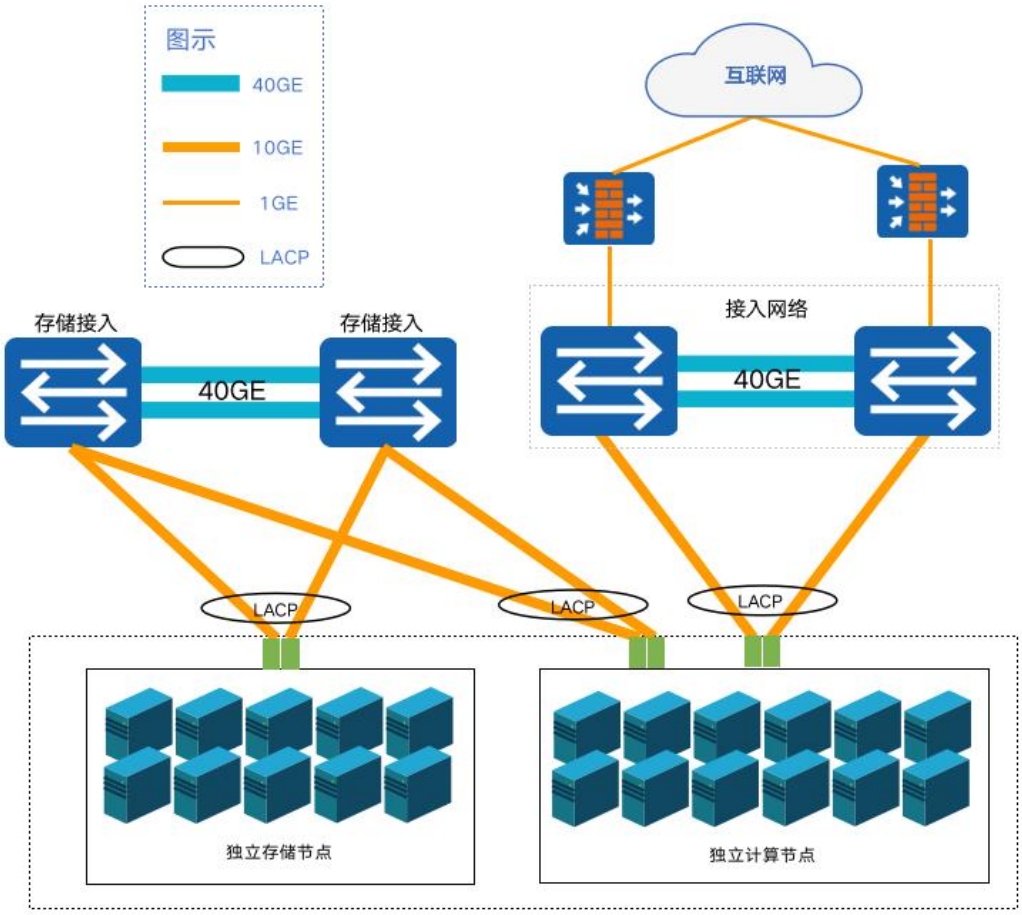
持更灵活的资源扩展和优化。存算分离的部署方式使得企业可以根据需要独立地扩展计算和存储资源，从而更好地适应快速增长或变化的业务需求。

2.1 超融合架构



- 计算存储融合节点，同时包含计算资源和存储资源，用于运行虚拟机、虚拟网络、分布式存储等资源，同时承载智能调度控制和监控服务。
- 采用 SATA+SSD 缓存的方案构建超融合节点，须保证 SSD 缓存盘和 HDD 数据盘的容量比不高于 1:20，数量比不高于 1:5。
- 生产环境至少部署 3 台以上，保证分布式系统的正常部署和运行。

2.2 存算分离架构



- 计算节点用于独立运行计算和网络资源，通过挂载独立存储节点的磁盘作为平台的存储资源。生产环境至少部署 2 台以上，保证虚拟机的调度及稳定迁移；通常建议将相同配置的计算节点服务器放置在一个集群内进行虚拟资源的调度。
- 采用 SATA+SSD 缓存的方案构建独立存储节点，须保证 SSD 缓存盘和 HDD 数据盘的容量比不高于 1:20；盘数量比不高于 1:5。

2.3 硬件选型

2.3.1 最低硬件配置

用户在部署平台时可选择超融合节点或存算分离节点进行部署，超融合节点

和独立节点最低硬件配置要求如下：

(1) 超融合节点

用于生产环境的最低主机和网络硬件配置，一般生产环境至少需要 3 台超融合节点和 2 台万兆接入交换机。针对测试环境和生产环境最低配置要求如下：

配置项	测试环境	生产环境	备注
CPU	CPU 不低于 10 核	CPU 不低于 16 核	每增加一块数据盘，需增加 2 核
内存	内存不低于 32GB	内存不低于 32GB	每增加一块数据盘，需增加 4GB
网卡	1 个 10GB 网口	2 个 10GB 网口	若考虑网卡冗余，推荐 2 张 10GB 网卡
系统盘	单块硬盘不小于 240GB	2 个 SSD 480GB RAID1	
数据盘	SSD 盘：最少 1 块。	SSD 盘：最少 1 块	1、整个分布式存储集群，需要 2TB 可用容量，作为平台自身的容量预留。 2、如节点只配置 1 块数据盘，单块 HDD 不低于 2TB，单块 SSD 不低于 1.92TB。
	HDD 盘：最少 1 块，且必须配置 SSD/NVME 缓存盘，存储盘的容量比不高于 1:20；数量比不高于 1:5	HDD 盘：最少 1 块，且必须配置 SSD/NVME 缓存盘，存储盘的容量比不高于 1:20；数量比不高于 1:5	
节点数量	3 台	3 台	
接入交换机	1 台万兆以太网交换机	2 台万兆以太网交换机	

注意 最低配置建议，只保证平台能正常部署，稳定运行，不考虑用户的资源预留。**生产环境，服务器硬件和架构层面必须保证冗余机制。**

(2) 存算分离节点

用于生产环境的最低主机和网络硬件配置，一般生产环境至少需要 2 台计算节点、3 台存储节点及 2 台万兆接入交换机。针对测试环境和生产环境最低配置要求如下：

<table>
<tr><th>硬件类型</th><th>配置项</th><th>测试环境</th><th>生产环境</th></tr>
<tr><td rowspan="4">计算节点</td><td>CPU</td><td>CPU 不低于 10 核</td><td>CPU 不低于 16 核</td></tr>
<tr><td>内存</td><td>内存不低于 32GB</td><td>内存不低于 32GB</td></tr>
<tr><td>网卡</td><td>1 个 10GB 网口</td><td>2 个 10GB 网口做网卡冗余</td></tr>
<tr><td>系统盘</td><td>1 个 SSD 480GB</td><td>2 个 SSD 480GB RAID1</td></tr>
<tr><td rowspan="8">存储节点</td><td>节点数量</td><td>1</td><td rowspan="2">2
CPU 不低于 16 核</td></tr>
<tr><td>CPU</td><td>CPU 不低于 10 核</td></tr>
<tr><td>内存</td><td>内存不低于 32GB</td><td>内存不低于 32GB</td></tr>
<tr><td>网卡</td><td>1 个 10GB 网口</td><td>2 个 10GB 网口做网卡冗余</td></tr>
<tr><td>系统盘</td><td>1 个 SSD 480GB</td><td>2 个 SSD 480GB RAID1</td></tr>
<tr><td rowspan="3">数据盘</td><td>SSD 盘：最少 1 块。</td><td>SSD 盘：最少 1 块</td></tr>
<tr><td>HDD 盘：最少 1 块，且必须配置 SSD/NVME 缓存盘，存储盘的容量比不高于 1:20；数量比不高于 1:5</td><td>HDD 盘：最少 1 块，且必须配置 SSD/NVME 缓存盘，存储盘的容量比不高于 1:20；数量比不高于 1:5</td></tr>
<tr><td colspan="2">1、整个分布式存储集群，需要 2TB 可用容量，作为平台自身的容量预留。
如节点只配置 1 块数据盘，单块 HDD 不低于 2TB，单块 SSD 不低于 1.92TB。</td></tr>
<tr><td>节点数量</td><td>3</td><td>3</td></tr>
<tr><td>网络设备</td><td>接入交换</td><td>1 台万兆以太网交换机</td><td>2 台万兆以太网交换机</td></tr>
</table>

	机		

> **注意** 最低配置建议，只保证平台能正常部署，稳定运行，不考虑用户的资源预留。**生产环境，服务器硬件和架构层面必须保证冗余机制。**

2.3.2 推荐硬件配置

(1) 网络设备推荐配置

业务	配置描述	构建方案
内网接入交换机	48*10GE + 6*100GE	必选
内网核心交换机	CE88-40G 板卡(16 口)*4，64*40GE	可选（48 节点以上扩容）
存储接入交换机	48*10GE + 6*100GE	可选（独立存储区域）
管理汇聚交换机	48*GE + 4*10GE + 2*40GE	可选（构建运维管理网）
IPMI 接入交换机	48*GE + 4*10GE + 2*40GE	可选（构建 IPMI 管理网）
网络设备 MGT 接入	48*GE + 4*10GE + 2*40GE	可选（构建 MGT 管理网）

(2) 服务器推荐配置

机型	配置描述
融合型——低配	Factor Form 2U CPU Intel Xeon Silver 4310 Processor(12CORES_2.1GHz_120W_X86) *2 DDR4_32GB_RDIMM_3200MHz *4 OS HDD 480G_SSD_SATA3_512E_2.5"_6Gb/s *2 Cache HDD 960G_SSD_U.2_N/A_512E_2.5"_32Gb/s*2 Data HDD SATA3_HDD_8TB *4 LSI-9311-8I*1

	双口万兆光口网卡(不含光模块)*2 PSU=800W*2/导轨
融合型——中配	Factor Form 2U CPU Intel Xeon Silver 4310 Processor(12CORES_2.1GHz_120W_X86) *2 DDR4_32GB_RDIMM_3200MHz *8 OS HDD 480G_SSD_SATA3_512E_2.5"_6Gb/s *2 Cache HDD 1.92T_SSD_U.2_N/A_512E_2.5"_32Gb/s*2 Data HDD SATA3_HDD_8TB *10 LSI-9311-8I 双口万兆光口网卡(不含光模块)*2 PSU=800W*2/导轨
融合型——高配	Factor Form 2U CPU Intel Xeon Silver 4310 Processor(12CORES_2.1GHz_120W_X86) *2 DDR4_32GB_RDIMM_3200MHz *16 OS HDD 480G_SSD_SATA3_512E_2.5"_6Gb/s *2 Cache HDD 3.84T_NVME_U.2_N/A_512E_2.5"_32Gb/s*2 Data HDD SATA3_HDD_16TB *10 LSI-9311-8I 双口万兆光口网卡(不含光模块)*2 PSU=800W*2/导轨
存储型	Factor Form 2U CPU Intel Xeon Silver 4310 Processor(12CORES_2.1GHz_120W_X86) *2 DDR4_32GB_RDIMM_3200MHz *4 OS HDD 480G_SSD_SATA3_512E_2.5""_6Gb/s *2 Cache HDD 3.84T_NVME_U.2_N/A_512E_2.5""_32Gb/s*2 Data HDD SATA3_HDD_16TB *10

	LSI-9311-8I 双口万兆光口网卡(不含光模块)*2 PSU=800W*2/导轨
计算型	Factor Form 2U CPU Intel Xeon Silver 4310 Processor(12CORES_2.1GHz_120W_X86) *2 DDR4_32GB_RDIMM_3200MHz *8 OS HDD 480G_SSD_SATA3_512E_2.5""_6Gb/s *2 LSI-9311-8I 双口万兆光口网卡(不含光模块)*2 PSU=800W*2/导轨

2.4 平台资源占用

平台运行本身需要占用服务器的 CPU、存储及存储资源，具体如下：

模块	角色	数量	CPU	内存	存储	说明
调度服务	调度管理服务	3	4C	8GB	400GB	CPU 内存占用管理集群物理资源，存储占用管理节点本地存储资源
	每计算节点——计算服务	N	4C	4GB	400GB	CPU 内存占用计算集群物理资源，存储占用计算节点本地存储资源
	存储服务	3	4C	4GB	400GB	均占用存储节点本地物理资源
存储服务	每存储节点——N 块硬盘存储服务	N	2C	4GB		均占用存储节点本地物理资源
	缓存加速模式	每 TB		4GB		每 TB 缓存容量需要消耗 4GB 内存
管理	管理服务	1	4C	8GB	240GB	CPU 内存占用计算集群资源，存储数据占用分布式存

服务						储资源
	公共服务	1	4C	8GB	640GB	CPU 内存占用计算集群资源，存储数据占用分布式存储资源

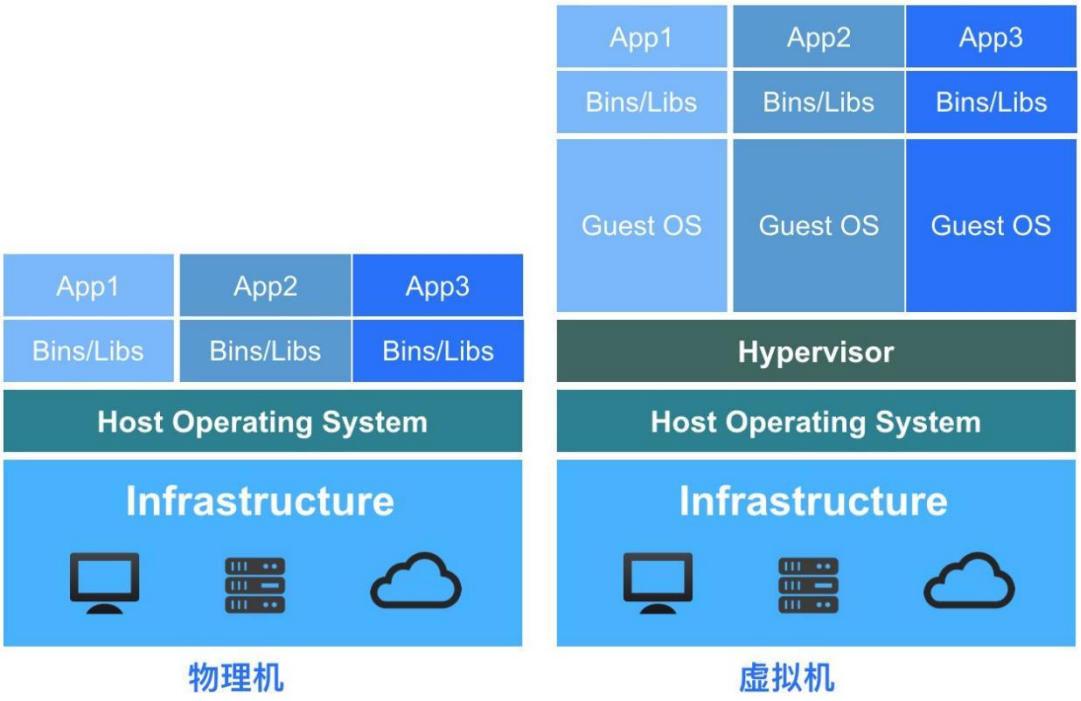
注意 存储服务的 CPU 和内存为预估值，实际生产环境中，根据使用负载等情况，可能会有变动。

3 平台技术架构

UCloudStack HCI 平台复用公有云核心虚拟化组件，具备计算虚拟化、智能调度、存储虚拟化、网络虚拟化的基础能力，为用户提供软件定义的计算、存储、网络及资源管理等服务，在保证资源服务性能、可用性及安全性的同时，提供统一资源调度及管理服务，适应企业基础设施服务的多种应用场景。

3.1 计算虚拟化

计算虚拟化是在硬件之上增加一个 Hypervisor，通过它虚拟出多个完全隔离的主机并可安装不同的操作系统，承载不同的应用程序运行，解决了一台物理机被一个系统或一个应用占用的问题，有效的提高资源使用率。



如上图所示，物理机和虚拟机在应用部署及资源占用上有本质区别：

(1) 物理机环境

- 操作系统是直接安装在物理机上，通常一台物理机只支持安装一个操作系统；
- 所有的应用程序和服务均需部署在物理机操作系统上，共享底层硬件资

源；

- 多个应用程序对底层操作系统及组件要求不一致时，可能会导致应用无法正常运行，需要将两个应用程序分别部署至不同物理机上，在非业务高峰时资源利用率较低。

(2) 虚拟机环境

- 在硬件底层及操作系统之上增加 Hypervisor 层，作为计算虚拟化的引擎；
- 虚拟化引擎支持将底层硬件虚拟为多个主机，即虚拟机；
- 每个虚拟机都拥有独立的硬件设施，如 CPU、内存、磁盘、网卡等；
- 每个虚拟机可以独立安装并运行不同的操作系统（GuestOS），相互完全隔离，彼此不受影响；
- 每个虚拟机操作系统与物理机操作系统能力一致，拥有独立的组件及库文件，可运行专属应用服务；
- 多个应用程序的虚拟机在完全隔离且彼此不影响的情况下运行在一台物理机上，并共享物理机的资源，提高物理机的资源使用率及管理效率。

优刻得超融合平台计算虚拟化采用 KVM 和 Qemu 等 Hypervisor 组件及技术，将通用架构的 x86/ARM 服务器资源进行抽象，以虚拟机的方式呈现给用户。虚拟机将 CPU、内存、I/O、磁盘等服务器物理资源转化为一组可统一管理、调度和分配的逻辑资源，并基于虚拟机在物理机上构建多个同时运行、相互隔离的虚拟机执行环境，可充分利用硬件辅助虚拟化技术，实现高资源利用率的同时满足应用更加灵活的资源动态分配需求，如快速部署、资源均衡部署、重置系统、在线变更配置及热迁移等特性，降低应用业务的运营成本，提升部署运维的灵活性及业务响应的速度。

平台计算虚拟化通过 KVM 硬件辅助的全虚拟化技术实现，因此需要 CPU 虚拟化特性的支持，即要求计算节点 CPU 支持虚拟化技术，如 Intel VT 和 AMD V 技术。KVM 属于 Linux Kernel 的一个模块，虚拟化平台可通过加载内核模块

的方式启动 KVM，管理虚拟硬件的设备驱动，用于模拟 CPU 和内存资源，同时需要加载 QEMU 模块模拟 I/O 设备。KVM 虚机包括虚拟内存、虚拟 CPU 和虚机 I/O 设备，其中 KVM 用于 CPU 和内存的虚拟化，QEMU 用于 I/O 设备的虚拟化。

虚拟机不直接感知物理 CPU，它的计算单元会通过 Hypervisor 抽象的 vCPU 和内存进行呈现，通过与 GuestOS 的结合共同构建虚拟机系统。I/O 设备的虚拟化是 Hypervisor 复用外设资源，通过软件模拟真实硬件进行呈现，为虚拟机提供诸如网卡、磁盘、USB 设备等外设。

计算虚拟化是 UCloudStack HCI 超融合平台的服务器虚拟化组件，是整个平台架构的核心组件。在提供基础计算资源的同时，支持 CPU 超分、QCOW2 镜像文件、GPU 透传、USB 透传、物理机纳管及集群平滑扩容等特性。

3.1.1 CPU 超分

平台支持平台物理 CPU 超分，即平台可虚拟化的 vCPU 数量可大于 pCPU 数量，在分配给虚拟机的 CPU 资源未全部使用时，共享未使用的部分给其它虚拟机使用，进一步提高平台 CPU 资源使用率。

支持用户按比例进行 CPU 超分，调整 CPU 超分比例延时生效，超分获得的可分配核数，以超分结果为准。平台自服务支持超分比例 100%（流畅）、150%，200%（低风险）、250%、300%、350%、400%（高风险）、450%、500%，550%、600%。

具体超分可分配核数以 1 台双路 CPU 的计算节点服务器为例：

- 双路 CPU 即为 2 颗物理 CPU，每颗物理 CPU 为 12 核，开启双线程；
- 每颗 CPU 为 24 核，两颗 CPU 为 48 核，即可分配 48 vCPU；
- 正常情况下，能提供的虚拟机 vCPU 为 48C；

若平台管理员开启 CPU 超分，并设置超分比例为 1:2，即代表可使用的 vCPU 数量是实际 CPU 数量的 2 倍。服务器（48C）在开启 2 倍超分后，可实际创建使用的 vCPU 为 96，即可创建 96C 的虚拟机。但不支持向下修改，即如

果已经设置了超分比为 1:2，则不再允许将超分比调为 1:1。

仅支持平台专业的运维人员设置并管理 CPU 超分比，平台管理员可查看平台 CPU 的实际使用量及 vCPU 的使用量。由于开启超分后，可能存在多台虚拟机共用 vCPU 的情况，为不大幅影响虚拟机的性能及可用性，通常建议尽量降低 CPU 超分比例，甚至不建议开启 CPU 超分。

如平台实际共 48 vCPU，经过超分后可创建 96 vCPU 的虚拟机，在虚拟机业务峰值时可能会真正占满 48 vCPU 的性能，通过超分资源运行的虚拟机性能会极速下降，甚至会影响虚拟机的正常运行。

CPU 超分比例需通过长期运行运营的数据进行调整，与平台虚拟机上所运行的业务应用程序有强关联性，需要长期考察平台在峰值业务时需要的 CPU 资源量进行灵活调整。

3.1.2 镜像文件

为方便镜像文件的上传、下载及分发，平台使用占用空间更小的 QCOW2 格式的镜像文件。

当创建虚拟机时，平台会将第一次使用到的镜像自动转换为 RAW 格式并导入到分布式存储中，通过分布式存储快照能力实现虚拟机磁盘的快速创建。

平台支持导入 ISO 镜像文件，当需要使用一个全新的操作系统时，可以选择使用包含操作系统介质的 ISO 镜像，直接使用 ISO 介质引导并安装到虚拟机中，方便用户快速创建自定义操作系统的虚拟机。

转换格式后的镜像及运行的虚拟机块设备均会存储于统一分布式存储系统中，方便虚拟机的迁移和故障恢复。

3.1.3 GPU 透传

平台支持 GPU 设备透传能力，为平台用户提供 GPU 虚拟机服务，让虚拟机拥有高性能计算和图形处理能力。GPU 虚拟机在科学计算表现中比传统架构性能提高数十倍，可同时搭配 SSD 虚拟硬盘，IO 性能亦在普通磁盘的数十倍以

上，可有效提升图形处理、科学计算等领域的计算处理效率，降低 IT 成本投入。

GPU 虚拟机与标准虚拟机采用一致管理方式，包括内网 IP 分配及安全组管理，并可对 GPU 虚拟机进行全生命周期管理，包括重置密码，变更配置及监控等，使用方式与普通的虚拟机一致，支持多种操作系统，如 CentOS、Ubuntu、Windows 等，在不增加额外管理的基础上，为用户提供快捷的 GPU 计算服务。

为让 GPU 发挥最佳性能，平台对 GPU、CPU 及内存的组合定义如下：

GPU	CPU	内存
1 颗	4 核	8G，16G
	8 核	16G，32G
2 颗	8 核	16G，32G
	16 核	32G，64G
4 颗	16 核	32G，64G
	32 核	64G，128G

平台本身不限制 GPU 品牌及型号，即支持任意 GPU 设备透传，已测试并兼容 GPU 型号为 NVIDIA 的 K80、P40、V100、2080、2080Ti、T4 及华为 Atlas300。

3.1.4 USB 透传

平台支持 USB 设备透传能力，使虚拟机可以使用物理机上的 USB 设备，从而使虚拟机可与宿主机 USB 设备进行数据交互。平台提供 USB Passthrough（设备直通）及 USB Redirection（设备转发）两种模式。

3.1.4.1 USB 直通

- **实现：**直通模式是将主机上的 USB 控制器或具体的 USB 设备直接分配给虚拟机的技术，通过 Qemu hostdev 设备方式挂载到虚拟机实例中，使得虚拟机能够直接访问和控制 USB 设备。

- **优势：**USB 设备直通能够提供更高的性能和更低的延迟，适用于对性能要求较高的 USB 设备，如外部存储设备等。
- **限制：**直通模式下虚拟机只能使用所在宿主机上的 USB 设备，迁移虚拟机时需要卸载 USB 设备。

3.1.4.2 USB 转发

- **实现：**转发模式通过 Qemu redir 设备方式挂载到虚拟机中，底层协议是 TCP 协议，主要原理是将“USB I/O 消息”封装成“TCP/IP 消息”，虚拟化层通过网络访问远端 USB 设备；转发模式下的 USB 设备不可再直通给虚拟机使用。
- **优势：**转发模式相对灵活，虚拟机可以跨节点使用其他宿主机上的 USB 设备，迁移时不需要卸载此 USB 设备。
- **限制：**延迟相对较高，不适用于对性能要求较高的设备。

在实际应用中，选择直通还是转发模式通常取决于具体的需求和应用场景。性能敏感的应用可能更倾向于使用 USB 设备直通，而对性能要求不苛刻的场景可能会选择 USB 设备转发，以保持更大的灵活性。平台提供了相应的功能界面和接口，使用户在创建虚拟机时能够根据需要配置选择对应的模式。

3.1.5 集群平滑扩容

平台支持平滑扩容集群节点，新增的节点不会影响已有节点及虚拟资源的运行。通过平滑扩容，平台管理员可轻松解决平台因业务增长而带来的资源扩展，包括硬件资源不足、高负载主机维护、新业务上线资源扩容等场景。

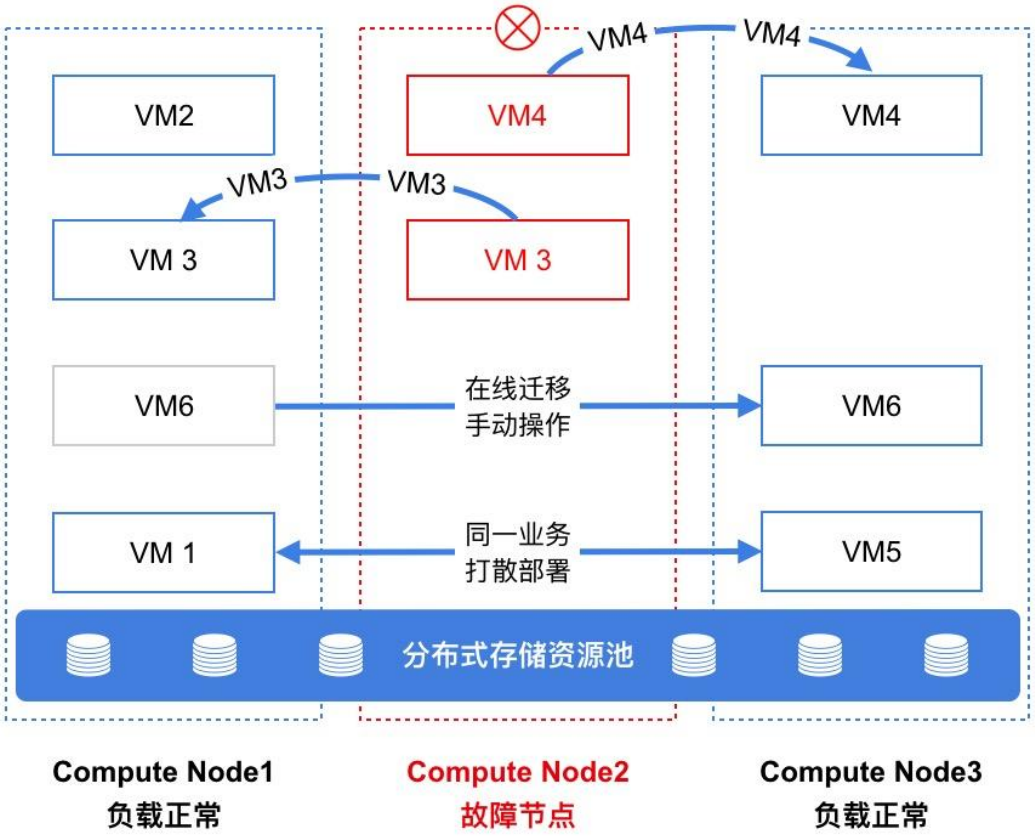
集群扩容可保证节点扩展过程中业务不中断，虚拟资源均正常运行，并提供简单快速的部署操作，支持自动化脚本一键部署上线。在扩容后平台支持在线为节点添加磁盘功能，使管理员可在不影响平台稳定运行的情况下，为平台横向及纵向的扩展资源。

平滑扩容成功后，平台原有的虚拟资源会保持原始状态，待平台有新的虚拟

资源需要运行和部署时，智能调度平台会将新的虚拟资源（如虚拟机）调度至平滑扩容的节点；若平台有物理机发生故障，原物理机上的虚拟机会根据调度策略迁移至新扩容的节点。支持平台管理员手动将一台虚拟机迁移至新扩容的节点，用于平衡平台整体资源使用率。

3.2 智能调度

智能调度是平台虚拟资源调度管理的核心，由调度模块负责调度任务的控制和管理，用于决策虚拟机运行在哪一台物理服务器上，同时管理虚拟机状态及迁移计划，保证虚拟机可用性和可靠性。



3.2.1 均衡调度

智能调度系统实时监测集群所有节点计算、存储、网络等负载信息，作为虚拟机调度和管理的数据依据。当有新的虚拟资源需要部署时，**调度系统会优先选择低负荷节点进行部署**，确保整个集群节点的负载相对均衡。如上图所示，新创建的虚拟资源将会通过调度检测，自动部署至负载较低的 Node3 节点上。

调度系统在优先选择低负荷节点进行虚拟资源部署的同时，分别提供打散部署、在线迁移、宕机迁移等能力，整体保证平台的可靠性。

平台使用分布式存储提供存储服务，如上图所示，虚拟机均运行于分布式存储池之上，且分布式存储池可跨多台物理机构建统一分布式存储资源池。

虚拟机的系统盘、镜像文件及挂载的硬盘均存储于统一分布式存储池中，每台节点均可通过分布式存储池中的虚拟机磁盘文件及配置信息注册一个相同的虚拟机进程，可作用于在线迁移或宕机迁移任务。

3.2.2 亲和策略

亲和反亲和策略，允许用户自定义虚拟机与其它虚拟机或宿主机之间的调度关系，使用户可以根据业务需求、性能要求或其它因素定义虚拟机调度逻辑，以满足特定的架构和运行要求。平台通过亲和反亲和策略提供了灵活的资源调度机制，用户可以根据实际需求动态调整调度策略，以适应不同业务负载和调度需求。

- 亲和策略用于将策略相关的虚拟机实例调度部署在同一物理主机上，可以最大程度地提高它们之间的数据传输和通信效率。
- 反亲和策略，是将虚拟机实例分散在不同物理主机上的调度策略，旨在降低单点故障的风险，提高系统的可靠性和容错性。

3.2.3 在线迁移

在线迁移（虚拟机热迁移）是计划内的迁移操作，即虚拟机不停机的情况下，在不同的物理机之间进行在线跨机迁移。首先是在目标物理机注册一个相同配置的虚拟机进程，然后进行虚拟机内存数据同步，最终快速切换业务到目标新虚拟机。整个迁移切换过程非常短暂，几乎不影响或中断用户运行在虚拟机中的业务，适用于平台资源动态调整、物理机停机维护、优化服务器能源消耗等场景，进一步增强平台可靠性。

由于采用分布式统一存储，虚拟机在线迁移时只迁移【计算】的运行位置，不涉及【存储】（系统盘、镜像、虚拟硬盘）位置迁移。迁移时仅需通过统一存储内的源虚拟机配置文件在目的主机上注册一个相同配置且状态置为暂停的虚

拟机进程，然后反复迁移源虚拟机的内存至目的虚拟机，待虚拟机内存同步一致后，关闭源虚拟机并激活目标虚拟机进程，最后进行网络切换并成功接管源虚拟机业务。

整个迁移任务仅在激活目标虚拟机及网络切换时业务处于短暂中断，由于激活和切换所用时间很短，少于 TCP 超时重传时间，因此源虚拟机业务几乎无感知。同时由于无需迁移虚拟机磁盘及镜像位置，虚机挂载的虚拟硬盘迁移后不受影响，可为用户提供无感知且携带存储数据的迁移服务。具体迁移过程如下：

(1) 注册目标虚拟机

- 调度系统使用统一分布式存储内的源虚拟机配置文件在目标主机上注册一个相同配置的虚拟机进程；
- 注册的虚拟机进程为不可提供服务的暂停【paused】状态，并通过监听一个 TCP 端口接收迁移数据；
- 注册目标虚拟机的阶段为瞬间完成，通常耗时为几毫秒，此时源虚拟机处于正常提供业务的状态。

(2) 迁移源虚拟机内存

- 在目标虚拟机注册完成的同时，调度系统会立即将源虚拟机的全量内存数据迁移至目标虚拟机；
- 为保证数据迁移的一致性，迁移过程中源虚拟机的内存更新也需要进行同步，因此调度系统通过多次迭代将源虚拟机产生的新内存数据迁移至目标端，耗时与物理机的网络带宽、性能及虚拟机的内存大小有关；
- 内存迁移时源虚拟机正常提供业务，待内存数据反复迭代迁移完成时立即暂停源虚拟机进程，避免产生新的内存数据；
- 源虚拟机进程暂停后，会再进行一次内存数据的同步，保证源端和目标端的数据一致性。

(3) 接管源虚拟机服务

- 完成内存同步的收尾工作，调度系统会关闭源虚拟机并激活目标虚拟机的进程，实现虚拟机平滑运行；
- 虚拟机从源主机迁移至目标主机，系统会将虚拟机的网络切换至目标主机，通过目标主机的虚拟交换机进行通信，成功接管源虚拟机服务。

整个迁移过程中，从源虚拟机暂停至目标虚拟机激活并完成网络切换为停机时间，由于激活虚拟机及网络切换时间非常短暂，通常小于几百毫秒，少于 TCP 超时重传时间，对大多数应用服务来说可忽略不计，因此虚拟机业务几乎不会感知到迁移停机。如智能调度图中的 VM6 默认运行在 Node1 上，管理员通过在线迁移功能手动将 VM6 迁移至 Node3 的流程如下：

- 调度系统收到迁移指令后，会立即使用 VM6 的配置文件在 Node3 节点上注册一个暂停状态的虚拟机进程；
- 立即迁移 VM6 的全量进程数据至 Node3 节点的 VM6’，并反复多次迁移更新内存数据；
- 调度系统暂停 Node1 上的 VM6 虚拟机，再次进行内存数据的迁移并关闭 VM6 虚拟机；
- 激活 Node3 节点上的 VM6 虚拟机进程，完成网络切换并接管 VM6 的业务服务及通信；
- 若 VM6 有挂载的虚拟硬盘，迁移成功后，不影响虚拟硬盘的挂载信息及配置，可正常读写虚拟硬盘。

3.2.4 离线迁移

离线迁移，是虚拟机在关机状态下，在不同的集群之间进行的离线跨集群迁移。离线迁移不涉及存储及数据迁移，更改虚拟机的集群配置，在虚拟机下次启动的时候，重新进入启动流程，重新分配物理机，在目标物理机注册一个相同配置的虚拟机进程，然后进行虚拟机内存数据同步，最终重新启动虚拟机到目标物理机。

离线迁移适用于平台资源动态调整、物理机停机维护、优化服务器能源消耗

等场景，进一步增强平台可靠性。

由于采用分布式统一存储，虚拟机的系统盘及写进系统盘的数据均存储在底层分布式存储中，虚拟机离线迁移只迁移【计算】的运行位置，不涉及【存储】（系统盘、镜像、虚拟硬盘）位置迁移，仅需选择可迁移的集群，在下次虚拟机启动时保证网络通信即可。迁移过程如下：

- 更改虚拟机的 xml 配置，修改集群信息，清空物理机信息
- 虚拟机在关机状态下，不进行任何操作，下一次启动重新走虚拟机启动流程，智能调度到空闲物理主机
- 调度系统使用统一分布式存储内的源虚拟机配置文件在目标主机上注册一个相同配置的虚拟机进程

3.2.5 宕机迁移

宕机迁移又称虚拟机高可用（High Availability），指平台底层物理机出现异常或故障而导致宕机时，调度系统会自动将其所承载的虚拟资源快速迁移到健康且负载正常的物理机，尽量保证业务的可用性。

整体宕机迁移不涉及存储及数据迁移，新虚拟机可快速在新物理机上运行，平均迁移时间为 90 秒左右，可能会影响或中断运行在虚拟机中的业务。

由于采用分布式统一存储，虚拟机的系统盘及写进系统盘的数据均存储在底层分布式存储中，虚拟机宕机迁移只迁移【计算】的运行位置，不涉及【存储】（系统盘、镜像、虚拟硬盘）位置迁移，仅需在新物理机上重新启动虚拟机并保证网络通信即可。迁移机制说明如下：

- 平台调度管理系统会周期性检测除本物理机之外的所有物理机，间隔时间为 10 秒；
- 当检测到某物理机出现网络中断，则会重试 3 次；
- 如果重试 3 次之后都不成功，就会将此物理机标记为不可达；
- 在所有物理机中，有超过半数的物理机都标记某台物理机为不可达，就

会判定此物理机为宕机，此物理机所有的虚拟机会在该集群（Set）内进行宕机迁移操作；

- 调度系统使用分布式存储内故障虚拟机的系统盘及数据重新在新物理机上启动虚拟机，启动过程及状态流转与新建虚拟机一致，平均启动时间为 30 秒左右；
- 虚拟机在新物理机上启动后，会将虚拟机网络切换至新物理机；

整个迁移过程，从检测到故障至迁移成功平均为 90 秒左右。虚拟机启动时间与源虚拟机的组件及配置有关，如绑定虚拟硬盘；同时由于虚拟机规格过大、底层物理资源不足、底层硬件故障等原因可能会导致宕机迁移失败，通常建议尽量保证底层物理资源充足。

如智能调度图中的 Node2 节点故障，智能调度系统自动将 VM3 和 VM4 分别迁移至 Node1 和 Node3 节点，具体流程如下：

- 调度系统经过周期性监测及二层检测，判断 Node2 节点故障，VM3/VM4 两台虚拟机不可用，需要进行宕机迁移操作；
- 调度系统根据收集的集群节点信息，使用分布式存储系统中 VM3 的系统盘及数据在 Node1 节点启动 VM3 虚拟机，并在启动后将 VM3 的网络信息切换至 Node1；
- 使用分布式存储系统中 VM4 的系统盘及数据盘在 Node3 节点启动 VM4 虚拟机，并在启动后将 VM4 的网络信息切换至 Node3；

宕机迁移的前提是集群中至少有 2 台以上的物理服务器，且在迁移过程中需保证健康节点的资源充足及网络连通性。通过宕机迁移技术，为业务系统提供高可用性，极大缩短由于各种主机物理故障或链路故障引起的中断时间。

3.3 存储虚拟化

平台通过硬件辅助的虚拟化技术最大程度上提高资源利用率和业务运维管理的效率，整体降低 IT 基础设施的总拥有成本，并有效提高业务服务的可用性、

可靠性及稳定性。在解决计算资源的同时，企业还需考虑虚拟化计算平台的数据存储，包括存储的安全性、可靠性、可扩展性、易用性、性能及成本等。

虚拟化计算KVM平台可对接多种类型的存储系统，如本地磁盘、商业化SAN存储设备、NFS 及分布式存储系统，分别解决虚拟化计算在不同应用场景下的数据存储需求。

- **本地磁盘：**服务器上的本地磁盘，通常采用 RAID 条带化保证磁盘数据安全。性能高，扩展性差，虚拟化环境下迁移较为困难，适用于高性能且基本不考虑数据安全业务场景。
- **商业化存储：**即磁盘阵列，通常为软硬一体的单一存储。性能高，成本高，需配合共享文件系统进行虚拟化迁移，适用于 Oracle 数据库等大型应用数据存储场景。
- **NFS 系统：**共享文件系统，性能较低，易用性较好，无法保证数据安全性，适用于多台虚拟机共享读写的场景
- **分布式存储系统：**软件定义存储，采用通用分布式存储系统的标准，将大量通用 x86 服务器的磁盘资源聚合在一起，提供统一存储服务。通过多副本的方式保证数据安全，高可靠、高性能、高安全、易于扩展、易于迁移且成本较低，适用于虚拟化、云计算、大数据、企业办公及非结构化数据存储等存储场景。

每一种类型的存储系统，在不同的存储场景下均有优劣势，虚拟化计算平台需根据业务特证选择适当的存储系统，用于提供存储虚拟化功能，在某些特定的业务模式下，可能需要同时提供多种存储系统，用于不同的应用服务。

在传统的存储结构中，客户端与单一入口点的集中式存储组件进行通信，可能会限制存储系统的性能和可伸缩性，同时可能带来单点故障。

UCloudStack HCI 平台采用分布式存储系统作为虚拟化存储，用于对接 KVM 虚拟化计算及通用数据存储服务，消除集中式网关，使客户端直接与存储系统进行交互，并以多副本/纠删码、多级故障域、数据重均衡、故障数据重建等数据保护机制，确保数据安全性和可用性。

3.3.1 分布式存储

平台基于 Ceph 分布式存储系统适配优化，为虚拟化计算平台提供一套纯软件定义、可部署于通用服务器的高性能、高可靠、高扩展、高安全、易管理且较低成本的虚拟化存储解决方案，同时具有极大可伸缩性。作为平台的核心组成部分，为用户提供多种存储服务及 PB 级数据存储能力，适用于虚拟机、数据库等应用场景，满足关键业务的存储需求，保证业务高效稳定且可靠的运行。

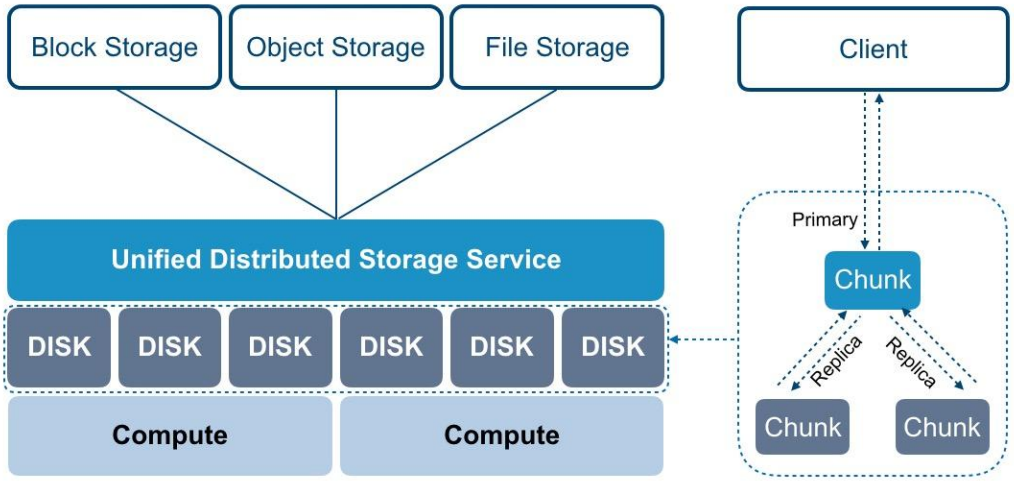
分布式存储服务通过将大量通用服务器的磁盘存储资源融合在一起进行【池化】，构建一个无限可伸缩的统一分布式存储集群，实现对数据中心所有存储资源的统一管理及调度，向虚拟化计算层提供【块】存储接口，供平台虚拟机或虚拟资源根据自身需求自由分配并使用存储资源池中的存储空间。

存储功能所见即所得，用户无需关注存储设备的类型和能力，即可在平台快捷使用虚拟化存储服务，如虚拟磁盘挂载、扩容、增量快照、监控等，平台用户像使用通用服务器的本地硬盘一样的方式使用虚拟磁盘，如格式化、安装操作系统、读写数据等。平台管理和维护者可以全局统一配置并管理平台整体虚拟化存储资源，如 QoS 限制、存储池扩容、存储规格及存储策略配置。

分布式存储系统提供块存储服务，同时可保证数据的安全性及集群服务的可靠性。在块存储的部署上，通常推荐使用同一类型的磁盘构建存储集群：

- SSD 磁盘构建为高性能的存储集群；
- SATA/SAS 磁盘构建为普通性能存储集群；
- SATA/SAS 磁盘构建的普通性能存储集群，在平台需通过 SSD 缓存加速的方式提升存储性能，缓存盘可采用 SSD 或 NVME 磁盘介质，平台要求缓存盘容量配比不高于 1:20，数量比不高于 1:5。

分布式存储系统提供的块存储服务可供虚拟机直接挂载使用，在数据写入时通过三副本、写入确认机制及副本分布策略等措施，最大限度保障数据安全性和可用性。逻辑架构如下：



分布式存储系统是整个平台架构不可或缺的核心组件，通过分布式存储集群体系结构提供基础存储资源，并支持在线水平扩容，同时融合智能存储集群、超大规模扩展、多副本与纠删码冗余策略、数据重均衡、故障数据重建、数据清洗、自动精简配置及快照等技术，为虚拟化存储提供高性能、高可靠、高扩展、易管理及数据安全性保障，全方面提升存储虚拟化及平台的服务质量。

3.3.2 高可用和高可靠

为构建全平台高可用的分布式存储服务，保证虚拟化计算及应用服务数据存储的可靠性，分布式存储系统从多方面保证存储服务的稳健运行。

- **基础设施高可用**

物理网络设备支持 10GE/25GE 底层存储堆叠网络架构，同时服务器层面均采用双链路，保证数据读写的 IO 性能及可用性。

- **存储监视器高可用**

集群监视器维护存储集群中数据映射信息，包括集群成员、状态、变更、以及存储集群的整体健康状况等。客户端会通过监视器获取最新集群映射图，为保证平台服务的可用性，支持监视器高可用，当一个监视器因为延时或错误导致状态不一致时，存储系统会通过算法将集群内监视器状态达成一致。

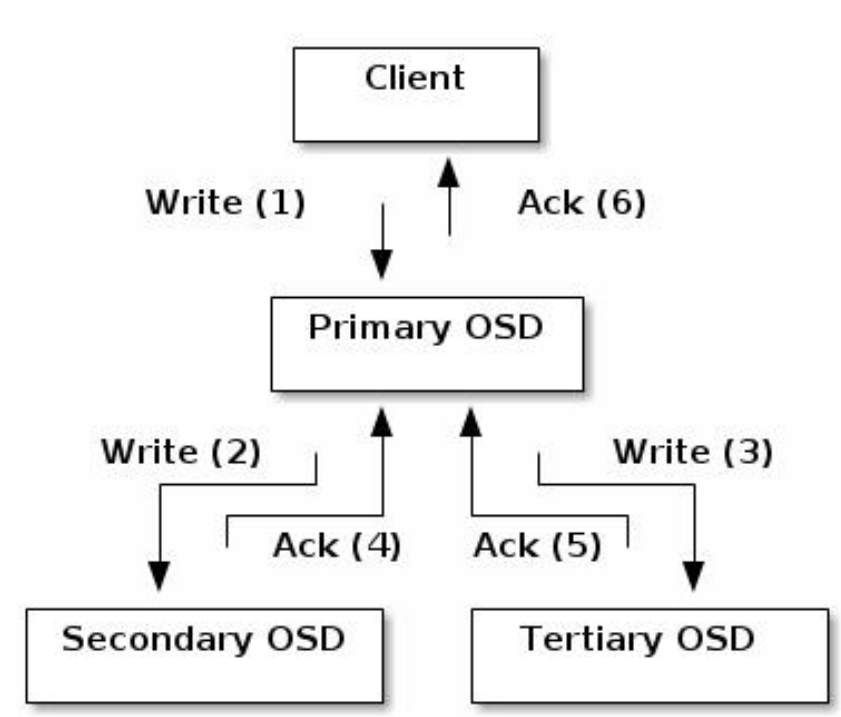
3.3.3 多副本冗余机制

多副本机制是指将写入的数据保存多份的数据冗余技术，并由存储系统保证多副本数据的一致性。平台分布式块存储系统默认采用多副本数据备份机制，**写入数据时先向主副本写入数据，由主副本负责向其他副本同步数据**，并将每一份数据的副本跨节点、跨机柜、跨数据中心分别存储于不同磁盘上，多维度保证数据安全。**存储客户端在读取数据会优先读取主副本的数据，仅当主副本数据故障时，由其它副本提供数据的读取操作。**

分布式存储系统通过多副本、写入确认机制及副本分布策略等措施，最大限度保障数据安全性和可用性。多副本机制存储数据，将自动屏蔽软硬件故障，当磁盘损坏和软件故障导致副本数据丢失，系统自动检测到并自动进行副本数据备份和同步，不会影响业务数据的存储和读写，保证数据安全性和可用性。本章节以三副本为例，具体描述多副本的工作机制：

(1) 三副本

用户通过客户端写入分布式存储的数据，会根据 Pool 设置的副本数量 3 写入三份，并按照副本分布策略，分别存储于不同物理主机的磁盘上。分布式存储保证数据安全的副本数量至少为 2 份，以便存储集群可以在降级状态下运行，保证数据安全。



(2) 写入确认机制

如上图所示，三副本在写入过程中，只有三个写入过程全部被确认，才返回写入完成，确保数据写入的强一致性。

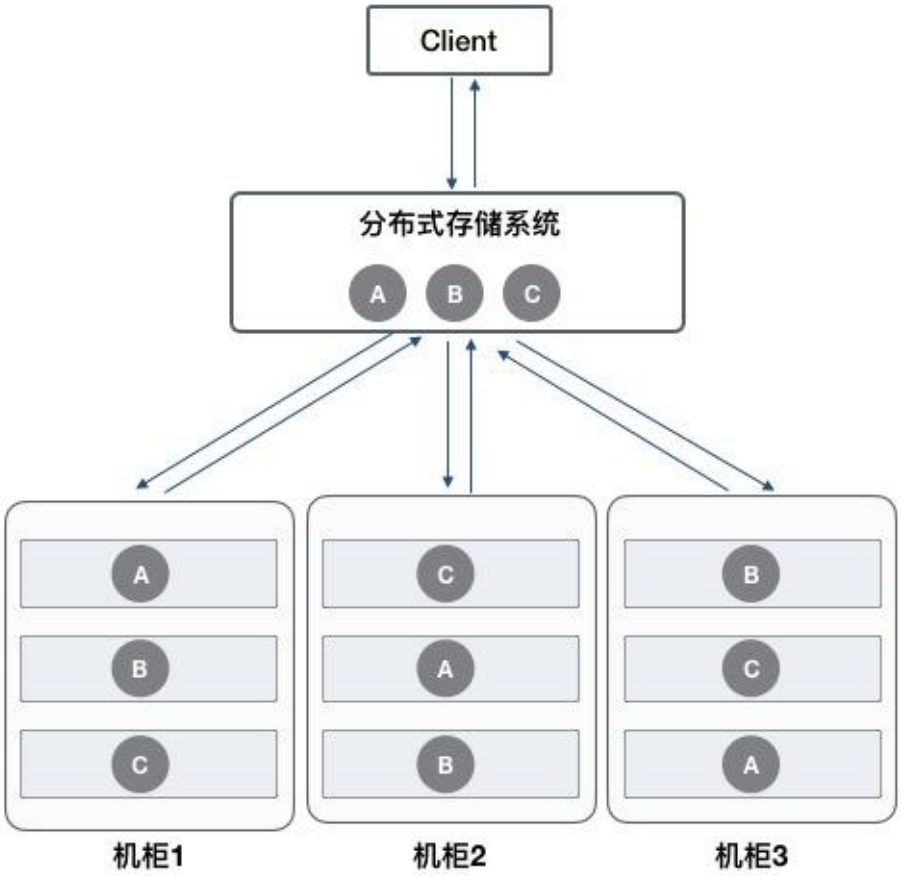
客户端将对象写入到目标 PG 的主 OSD 中，然后主 OSD 通过 GRUSH 映射关系图定位用于存储对象副本的第二个和第三个 OSD，并将对象数据复到 PG 所对应的两个从 OSD，当三个对象副本数据均写入完成，最后响应客户端确认对象写入成功。

(3) 副本分布策略

分布式存储支持副本数据落盘分布策略（多级故障域），使用 CRUSH 算法根据存储设备的权重值分配数据对象，尽量确保对象数据的均匀分布。平台支持节点级、机柜级、数据中心级故障域，可将副本数据分布在不同主机、不同机柜及不同数据中心，避免因单主机、单机柜及单数据中心整体故障造成数据丢失或不可用的故障，保证数据的可用性和安全性。

为保证存储数据的访问时延，通常建议最多将数据副本保存至不同的机柜，若将数据三副本保存至不同的机房，由于网络延时等原因，可能会影响虚拟硬盘

的 IO 性能。



如上图所示，客户端通过分布式存储系统写入 ABC 三个对象数据，根据 CRUSH 规则定义的故障域，需要将三个对象的副本分别存储于不同的机柜。以 A 对象为例，存储系统提前设置副本分布策略，尽量保证对象副本分布在不同柜柜的服务器 OSD 中。当分布式存储系统计算出写入对象的 PG 及对应的 OSD 位置时，会优先将 A 写入到机柜 1 的服务器 OSD 中，同时通过主 OSD 复制副本 A'至机柜 2 的服务器 OSD 中，复制 A"至机柜 3 的服务器 OSD 中，数据全部复制写入成功，即返回客户端对象 A 写入成功。

在存储节点无网络中断或磁盘故障等异常情况时，对象副本数据始终保持为3副本。仅当节点发生异常时，副本数量少于3时，存储系统会自动进行数据副本重建，以保证数据副本永久为三份，为虚拟化存储数据安全保驾护航。如上图第三个节点发生故障，导致数据D1-D5丢失并故障，存储系统会将对象数据的PG自动映射一个新的OSD，并通过其它两个副本自动同步并重建出D1‘-D5’，以保证数据始终为三副本，保证数据安全。

3.3.4 数据重均衡

平台分布式存储集群在写入数据时，会通过数据分片、CRUSH映射关系、多副本分布策略尽量保证数据对象在存储池中的均衡。随着存储集群的长期运行及对平台的运维管理，可能会导致存储池内的数据失衡，如节点和磁盘扩容、存储部分数据被删除、磁盘和主机故障等。

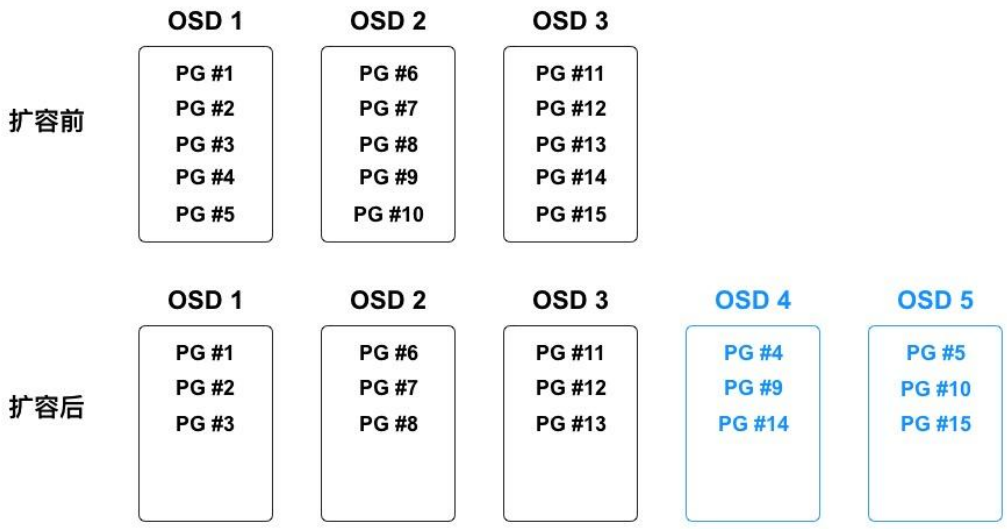
- 存储节点及磁盘扩容后，平台总存储容量增加，新增容量未承载数据存储，导致集群数据失衡；

- 用户删除虚拟机或虚拟硬盘数据，导致集群内出现大量空闲空间；
- 磁盘和主机故障下线后，部分数据对象副本会重建至其它磁盘或主机，故障恢复后处于空闲状态。

为避免扩容及故障导致存储集群数据分布失衡，平台分布式存储系统提供数据重均衡能力，在存储集群及磁盘数据发生变更后，通过 CRUSH 规则及时对数据的部分对象进行重新分发和均衡，使存储池中的对象数据尽量均衡，避免产生数据热点及资源浪费，提升存储系统的稳定性及资源利用率。

(1) 集群扩容重均衡

平台支持水平扩展存储节点或在线向存储节点中增加磁盘的方式扩容存储集群的容量，即分布式存储集群支持在运行时增加 OSD 进行存储池扩容。当集群容量达到阈值需要扩容时，可将新磁盘添加为集群的 OSD 并加入到集群的 CRUSH 运行图，平台会按照新 CRUSH 运行图重新均衡集群数据分布，将一些 PG 移入/移出多个 OSD 设备，使集群数据回到均衡状态。如下图所示：



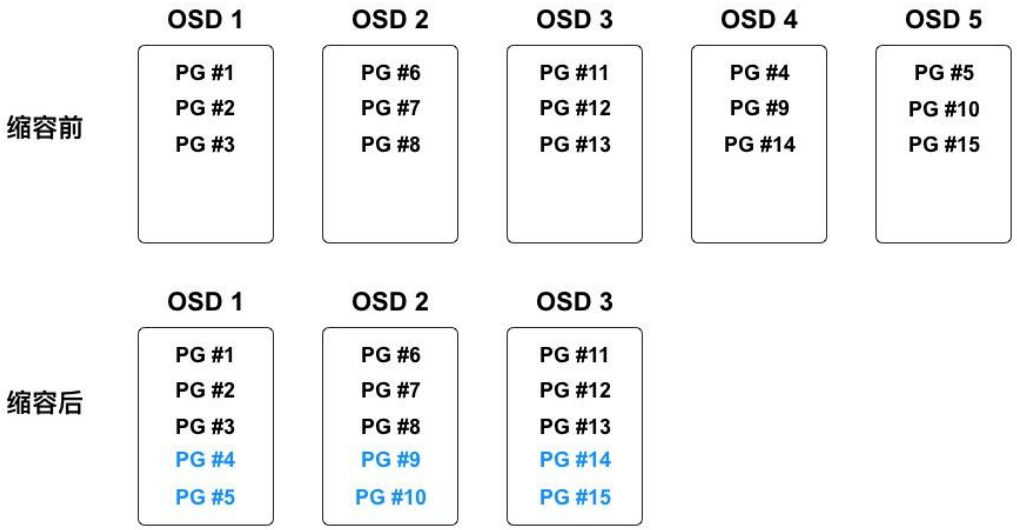
在数据均衡过程中，仅会将现有 OSD 中的部分 PG 到迁移到新的 OSD 设备，不会迁移所有 PG，尽量让所有 OSD 均腾出部分容量空间，保证所有 OSD 的对象数据分布相对均衡。如上图中新增 OSD 4 和 OSD 5 后，有三个 PG（PG#4、PG#9、PG#14）迁移到 OSD 4，三个 PG（PG#5、PG#10、PG#15）迁移到 OSD 5，使五个 OSD 中映射的 PG 均为 3 个。为避免 PG 迁移导致集群性能整体降低，**存储系统会提高用户读写请求的优先级，在系统空闲时间进行 PG 迁移**

操作。

> **说明** PG 在迁移过程中，原 OSD 会继续提供服务，直到 PG 迁移完成才将数对象写入新 OSD 设备。

(2) 集群容量缩减重均衡

存储集群在运行过程中可能需要缩减集群容量或替换硬件，平台支持在线删除 OSD 及节点下线，用于缩减集群容量或进入运维模式。当 OSD 被在集群中被删除时，存储系统会根据 CRUSH 运行图重新均衡集群数据分布，将被删除的 OSD 上的 PG 迁移至其它相对空闲的 OSD 设备上，使集群回到均衡状态。如下图所示：



在数据均衡过程中，仅会将被删除 OSD 上的 PG 迁移至相对空闲的 OSD 设备，尽量保证所有 OSD 的对象数据分布相对均衡。如上图中即将被删除的 OSD 4 和 OSD 5 上共映射 6 个 PG，删除后分别分有 2 个 PG 会被迁移至剩余 3 个 OSD 中，使 3 个 OSD 中映射的 PG 均为 5 个。

(3) 故障数据重均衡

分布式存储在长期运行中会存在磁盘、节点的物理损坏、系统崩溃及网络中断等故障，均会中断节点的存储服务。存储集群提供容错方法来管理软硬件，PG 作为对象与 OSD 的中间逻辑层，可保证数据对象不会直接绑死到一个 OSD 设备，意味着集群可在“降级”模式下继续提供服务。详见数据故障重建。

> **说明** 通过数据重均衡机制，可支持分布式存储集群平滑扩容，包括横向扩容和纵向扩容，即可以在线添加存储节点及存储磁盘。

3.3.5 数据故障重建

根据多副本保护机制，存储集群在把数据对象通过CRUSH写入到指定OSD后，OSD 会通过运行图计算副本或数据块的存储位置，并将数据副本或数据块写入到指定 OSD 设备中，通常数据对象会被分配至不同故障域中，保证数据安全性和可用性。

当磁盘损坏或节点故障时，即代表节点部分/全部 OSD 设备下线或无法为 PG 内对象提供存储服务，同时也表示有部分对象数据的副本数量不完整，如 3 副本可能因为磁盘损坏变为 2 副本。故障时对象数据的 PG 被置为“降级”模式继续提供存储服务，并开始进行数据副本重建操作，按照最新 CRUSH 运行图将故障节点或磁盘上的对象数据重映射到其它 OSD 设备上，即重新复制对象数据的副本至其它 OSD 设备，保证副本数量与存储池设置一致。

故障数据重建时会遵循存储集群中配置的故障域（主机级、机柜级及数据中心级），选择符合故障域定义的 OSD 作为故障数据重建的位置，让同一对象数据的多副本数据间位置互斥，避免数据块均位于同一个故障域，保证数据安全性和可靠性。同时为提高故障数据的重建速度，多个故障数据重建任务的 I/O 会并发进行，实现故障数据的快速重建。

故障节点或磁盘恢复后，OSD 被重新加入至集群的 CRUSH 运行图，平台会按照新CRUSH运行图重新均衡集群数据分布，将一些PG移入/移出多个OSD设备，使集群数据回到均衡状态。**为保证存储集群的运营性能，副本数据恢复及迁移时，会限制恢复请求数、线程数、对象块尺寸，并提高用户读写请求的优先级，保证集群可用性和运行性能。**

3.3.6 数据清洗

分布式存储集群在长期运行及数据重平衡的过程中，可能会产生一些脏数据、缺陷文件及系统错误数据。如一块 OSD 磁盘损坏，集群在重均衡后重建数

据至其它 OSD 设备，当故障 OSD 设备恢复后可能还存储着之前数据的副本，这些副本数据在集群重新平衡时需及时进行清洗。

分布式存储的 OSD 守护进程可进行 PG 内对象的清洗，即 OSD 会比较 PG 内不同 OSD 的各对象副本元数据，如果发现有脏数据、文件系统错误及磁盘坏扇区，会对其进行深度清洗，以确保数据的完整性。

3.3.7 自动精简配置

自动精简配置（Thin Provisioning）是一种利用虚拟化技术减少物理存储部署的技术。通过自动精简配置，可以用较小的物理容量提供较大容量的虚拟存储空间，且真实的物理容量会随着数据量的增长及时扩展，可最大限度提升存储空间的利用率，并带来更大的投资回报。

平台分布式存储系统支持自动精简配置，在创建块存储服务时，分配逻辑虚拟容量呈现给用户，当用户向逻辑存储容量中写入数据时，按照存储容量分配策略从物理空间分配实际容量。如一个用户创建的虚拟硬盘为 1TB 容量，存储系统会为用户分配并呈现 1TB 的逻辑卷，仅当用户在虚拟硬盘中写入数据时，才会真正的分配物理磁盘容量。若用户在虚拟硬盘上存储的数据为 100GB，则虚拟硬盘仅使用存储池的 100GB 容量，剩余的 900GB 容量可以供其它用户使用。

平台分布式存储系统支持对真实物理容量的监控，可提供真实物理已使用容量和逻辑的已分配容量。通常建议真实已使用容量超过总容量的 70%时对存储集群进行扩容。自动精简配置类似 CPU 超分的概念，即可供用户创建使用的存储容量可大于物理总容量，自动按需分配物理存储空间给块存储设备，消除已分配但未使用的存储空间浪费。

通过自动精简配置，平台管理员无需对业务存储规模进行细化且准确预判，更不需提前为每个业务做精细的空间资源规划和准备，配合逻辑存储卷的容量分配策略，有效提升运维效率及存储空间的整体利用率。

3.3.8 块存储服务

平台通过软件定义的分布式存储重新定义数据存储服务，基于通用服务器构

建统一存储层，为应用提供块存储服务，用户无需关注底层存储设备及架构，即可在平台构建并使用存储服务。

平台基于分布式存储系统为用户提供块设备，即虚拟硬盘服务，为计算虚拟化的虚拟机提供持久化存储空间的块设备。具有独立的生命周期，支持随意绑定/解绑至多个虚拟机使用，并能够在存储空间不足时对虚拟硬盘进行扩容，基于网络分布式访问，为虚拟机提供高安全、高可靠、高性能及可扩展的数据磁盘。

根据底层物理存储设备的磁盘介质不同，平台可为用户提供普通和高性能多种架构类型的虚拟硬盘：

- 普通虚拟硬盘使用 SATA/SAS 磁盘作为存储介质，并采用 SSD/NVME 作为缓存加速盘，保证普通虚拟硬盘性能。
- 性能型虚拟硬盘使用 SSD/NVME 磁盘作为存储介质。

虚拟硬盘数据均通过三副本机制进行存储，并在分布式存储系统的基础之上通过块存储系统接口为用户提供虚拟硬盘资源及全生命周期管理。

支持组建多个存储集群，如 SATA 存储集群和 SSD 存储集群，并支持虚拟机跨集群挂载集群上的块存储服务。

- 分布式块存储服务直接通过物理网络进行挂载。
- 通过 libvirt 融合分布式存储 rbd 和 qemu，qemu 通过 librbd 操作分布式存储。
- 虚拟化进程与分布式存储进程通过本机&跨物理机内网进行通信。

不同存储集群间，数据存储完全隔离。一个存储集群中不同块存储设备的存储策略完全隔离，互不干扰。分布式存储系统为虚拟机系统盘、镜像及虚拟硬盘提供统一存储及管理，提高虚拟机与系统盘、虚拟硬盘的数据传输效率，实现虚拟机快速创建及恢复，并支持系统盘和虚拟硬盘的在线快速扩容和迁移。

在业务数据安全方面，平台分布式存储支持磁盘快照能力，可降低因误操作、版本升级等导致的数据丢失风险，是平台保证数据安全的一个重要措施。支持对虚拟机的系统盘和数据盘进行手动或定时快照，在数据丢失或损坏时，可通过快

照快速恢复本地业务的数据，实现业务分钟级恢复，包括数据库数据、应用数据及文件目录数据等。

3.4 网络虚拟化

网络是虚拟化计算和分布式存储为平台提供服务时不可或缺的核心部分，平台通过 OVS 实现分布式虚拟交换机，为虚拟机提供基于 VLAN 隔离的二层网络，不同物理节点上的虚拟交换机能够协同工作，实现虚拟机之间的跨节点通信及不同节点间的无缝迁移。

3.4.1 分布式网络

分布式网络架构，无集中网络转发节点，所有生产网络仅在计算节点上传输，无需通过管理服务或网络节点进行转发，避免集中网络转发节点成为性能瓶颈。每个扁平网络通过独立的虚拟交换机实现，多个扁平网络之间不复用虚拟交换机，保证多个网络之间的完全隔离。对于每个创建的扁平网络，平台会在集群内的每个节点上均创建一个独立的虚拟交换机并设定相同的网络配置，虚拟机在集群内多个节点进行迁移后依然可以拥有配置一致的网络连接。

- **多网络隔离**

平台允许管理员创建多个扁平网络，多网络的支持使得平台能够满足虚拟化环境中网络的多样化需求。每个扁平网络通过 VLAN 实现网络隔离，为虚拟机提供多个独立、安全的二层网络环境。虚拟机在同一扁平网络内可以直接通信，而不同扁平网络的虚拟机则通过 VLAN 隔离，保障网络的安全性。

- **独立虚拟交换机**

每个扁平网络都通过独立的虚拟交换机实现，确保网络之间的彻底隔离。这有助于避免网络中的潜在干扰，提高网络的稳定性和安全性。

- **一致性网络配置**

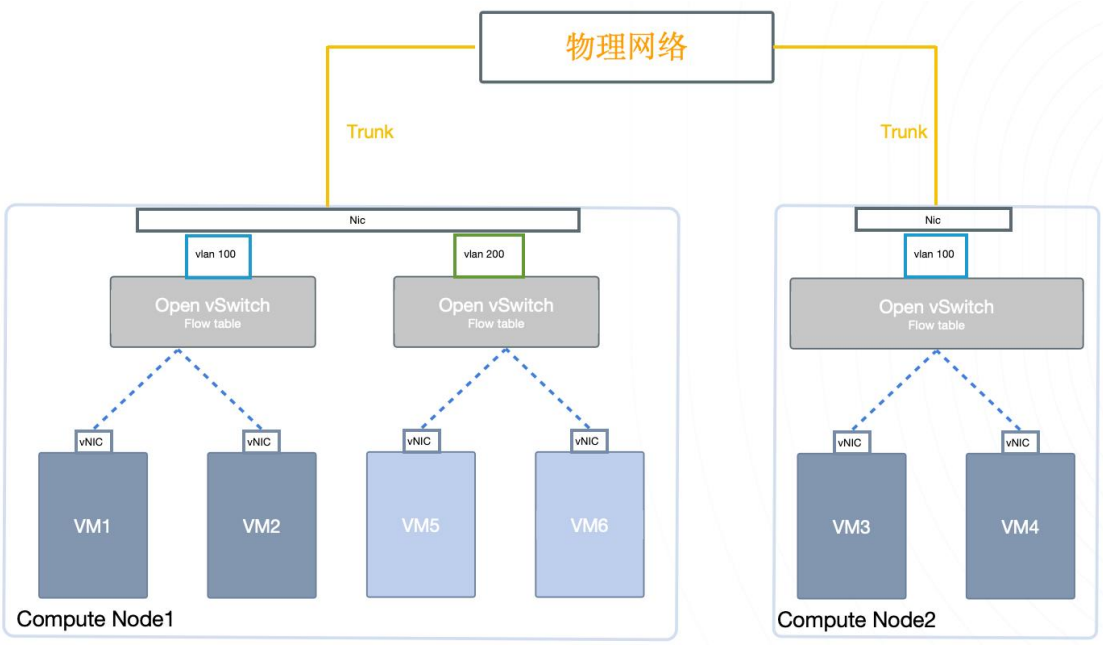
每个节点上对应的虚拟交换机都被设定为相同的网络配置，通过配置相同的 VLAN 子接口保证所有加入该网络的虚拟机位于相同的 VLAN 隔离网络。这

确保在整个集群内，不论虚拟机位于哪个节点，其所连接的扁平网络都拥有相同的网络环境。当虚拟机在集群内的不同节点之间进行迁移时，由于每个节点都有相同配置的虚拟交换机，虚拟机依然能够拥有配置一致的网络连接。

通过采用一致性网络配置和分布式虚拟交换机，平台为虚拟机提供可靠、简单、易用、一致的网络环境。

3.4.2 分布式架构

平台采用 OVS 作为虚拟交换机，网络数据面组件以分布式的方式部署于每个计算节点服务器，结合自研的网络控制器实现虚拟交换机的协同工作及自动化配置，为虚拟机提供高效易用的网络服务。



如上图所示，由 OVS 实现的虚拟交换机分布式运行在所有计算节点，即每个计算节点均部署 OVS 组件，由平台对所有虚拟交换机统一配置管理，通过分布式交换机提供集群内配置一致的网络连接：

- 分布式架构，无集中网络转发节点，所有生产网络仅在计算节点上传输，无需通过管理服务或网络节点进行转发，避免集中网络转发节点成为性

能瓶颈。

- 每个计算节点仅承载运行在本机的虚拟机网络转发和传输，单节点故障，不影响其它节点的虚拟网络通信。
- 管理服务故障，不影响已部署好的虚拟资源运行及通信。

分布式网络架构将业务数据传输分散至各个计算节点，所有虚拟化资源的业务流量均分布在集群节点上，即平台业务扩展并不受管理节点数量限制。

3.4.3 通信机制

平台通过扁平网络为虚拟机提供基于 VLAN 隔离的二层网络，具体通信原理如下：

- 相同扁平网络中，同一个物理主机上的虚拟资源可直接通过 OVS 进行网络数据通信；
- 相同扁平网络中，跨物理主机虚拟资源间的数据经由 OVS 后通过物理 VLAN 网络通信；
- 不同扁平网络间通过 VLAN 实现二层隔离，不可直接通信，可通过三层路由实现多个网络互通；

如分布式网络架构图所示，假设 VM1、VM2、VM3 属于同一个扁平网络，VM5、VM6 同属另一个不同的扁平网络，则虚拟机通信机制如下：

- VM1 和 VM2 属于同一个扁平网络且在同一台物理主机，可以通过 Open vSwitch 直接进行通信。
- VM1 和 VM3 属于同一个扁平网络但位于不同的物理主机，数据通过 Open vSwitch 后流经物理 VLAN 网络进行互相通信。
- VM1 和 VM5 属于不同的扁平网络，由于不同 VLAN 之间二层隔离，二者不可直接通信，需要通过三层路由实现网络互访。

3.4.4 网络控制器

网络控制器负责虚拟交换机及虚拟防火墙的自动化管理。针对每个扁平网络，集群所有节点均有对应的独立虚拟交换机，网络控制器根据扁平网络的网卡及 VLAN 配置进行所有节点的自动化配置，实现扁平网络在集群中所有节点的网络配置一致性。根据虚拟机设定的安全组，网络控制器自动下发对应的网络规则，实现虚拟网卡级别的网络安全防护。

网络控制器和智能调度系统一样，由【核心调度及管理模块】提供虚拟网络控制及管理，支持集群架构，结合物理网络及链路的冗余架构，整体提升虚拟网络的可用性。

- 每一个地域仅需部署一套高可用（主备模式）的网络控制模块，可在两台或多台节点上进行部署。
- 当部署网络控制模块所在的主计算节点服务器物理故障时，部署调度模块的备计算节点将自动接替调度服务，保证核心调度及网络安全规则控制服务的可用性。
- 网络控制器仅承载控制服务，不透传生产网络传输。

网络控制器高可用架构全部故障，仅影响新建虚拟资源的管理，不影响已部署的虚拟资源运行及通信。由于网络控制器及平台管理服务不承载业务数据转发和传输，均由分布在所有计算节点的 OVS 组件进行所属虚拟机的网络传输，所以计算节点在水平扩展的同时，承载生产流量的虚拟网络也同步进行了扩展，整体提升平台的可用性和可靠性。

3.4.5 网络功能简介

平台通过虚拟交换机实现简单高效的扁平网络架构，为虚拟机提供多个全局统一配置的分布式二层网络。通过安全组实现的虚拟防火墙，可灵活控制每一个虚拟网卡的网络安全规则。

- **扁平网络**

通过虚拟交换机实现简单、高效的二层网络。支持对扁平网络预设 IP 地址段，实现虚拟机 IP 地址的自动化分配和配置。支持自定义路由及内置 DHCP 服务，以简化平台网络的使用和管理。

- **网络隔离能力**

每个扁平网络通过不同的 VLAN 标识进行二层隔离，每个扁平网络内的设备可直接进行通信，不受其他扁平网络的影响。

- **安全组**

虚拟防火墙，提供出入双方向流量访问控制规则，定义哪些网络或协议能访问资源，用于限制虚拟资源的网络访问流量，支持 TCP、UDP、ICMP 及多种应用协议，为平台提供必要的安全保障。

- **网络资源管理**

支持平台管理员查看并管理平台所有的网络资源，包括物理网络资源和虚拟网络资源。

3.5 复用公有云

UCloudStack HCI 基于 UCloud 公有云核心技术，复用内核及核心虚拟化组件，历经 10 年的大规模磨炼和验证，保证平台底层的稳定性。具有自主可控、稳定可靠、持续进化及开放兼容等特点，助力企业轻松实现数字化转型。

专业的内核及测试团队及时跟进兼容性问题、安全问题、性能问题，确保平台时刻拥有一个健壮的底座。

根据权威机构评测，平台代码自研率达 96%以上，为业界领先水平。在国产化替代的大背景下，自主可控显得尤为重要。产品演进不受开源项目干扰和限制，紧跟客户需求，为用户解决实际问题。

3.6 异构管理

UCloudStack HCI 平台提供通过信创互认证的 IaaS 功能，兼容硬件、操作

系统到上层应用的全信创生态，助力企业快速构建自主可控的数字化底座。

- **异构资源统一管理**：多种架构资源统一管理，建设信创资源池的同时利旧和纳管传统架构资产，方便不同业务可按需选用合适的资源类型。
- **软硬件生态的快速适配**：专业的适配团队，快速完成硬件、操作系统、中间件等的适配，充分的测试保证兼容稳定性，满足用户不同场景需求。
- **多数据中心统一管理**：通过多地域统一管理能力实现多数据中心、多集群的统一调度和管理，提供统一运维和运营的一致体验，简化整体管理。

UCloudStack HCI 超融合平台从芯片到应用进行了全面的适配，CPU 已完成鲲鹏、飞腾、海光、龙芯、兆芯、申威主流国产芯片的适配，宿主机层面也已实现银河麒麟和统信等国产操作系统的全面兼容。

4 核心产品服务

4.1 基本概念

4.1.1 地域

地域（Region）指物理数据中心的地理区域，是平台中的一个逻辑概念，指资源部署的物理位置分类，可对应机柜、机房或数据中心，如上海、北京、杭州、主数据中心、备数据中心等。

数据中心之间资源和网络完全物理隔离，可通过一套管理平台管理遍布各地数据中心的超融合平台。

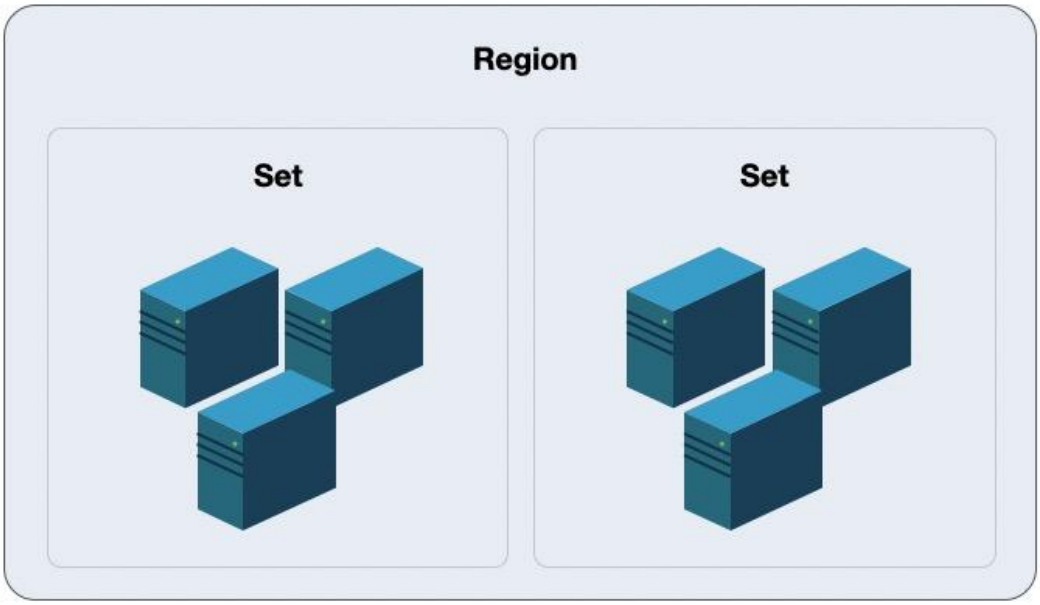
平台默认内置一个地域，管理服务通过本地数据中心平台提供的 API 端点管理地域内计算、存储及网络资源。支持对数据中心内资源的生命周期管理，包括计算集群、存储集群、基础镜像及自制镜像等资源的查看和维护。

- 不同地域间完全物理隔离，平台资源创建后不能更换地域；
- 不同地域间网络完全隔离，资源内网不能互通，可通过公网或专线进行网络通信；

4.1.2 集群

集群（Set）是平台物理资源的逻辑划分，用于区分不同配置规格及不同存储类型的服务器节点，如 X86 计算集群、ARM 计算集群或 SSD 存储集群。

一个数据中心可支持部署多个计算和存储集群，一个集群通常由一组配置、用途相同的物理节点组成，且服务节点一般具有相同的 CPU/内存、磁盘类型及操作系统。



- 一个地域可包含多个集群，使用统一管理平台进行集群管理和运营，资源仅支持在单集群调度；
- 一个集群至少由3台服务器节点组成，集群内服务器须具有相同的CPU/内存、磁盘类型及操作系统；不同磁盘类型的节点划分为一个集群，如SATA 存储节点集群；
- 通常一个集群的服务器建议接入同一组接入交换机，业务数据网络仅在集群内进行传输；
- 若采用独立存储节点，可将其与计算节点划分为一个集群进行磁盘挂载；
- 虚拟机支持跨集群挂载分布式块存储设备，用于数据存储。

平台支持将X86、ARM、GPU等异构计算集群统一管理，并可统一管理SSD、STAT、NVME 多种架构存储集群。

用户可将虚拟资源部署于不同的计算集群，并分别对虚拟资源挂载不同存储集群的块存储设备。

4.1.2.1 计算集群

计算集群是一组配置、用途相同的计算节点（物理机）组成，用于部署并承

载平台上运行的虚拟计算资源。一个数据中心可部署多个不同类型的计算集群，如 X86 集群、ARM 集群、GPU 集群等，不同的集群可运行不同类型的虚拟机资源，如 GPU 集群可为用户提供 GPU 虚拟机，ARM 集群可为用户提供基于 ARM 或国产化 OS 的虚拟机。

为保证虚拟机高可用，平台基于集群维度提供虚拟化智能调度策略，包括打散部署、在线迁移、离线迁移及宕机迁移，即虚拟资源可在集群内的所有计算节点中进行调度、部署及迁移，提升业务的可用性。

- **打散部署**

  平台用户创建虚拟机时默认会将创建的虚拟机尽量打散部署于集群内的所有节点上，保障硬件或软件故障等异常情况下用户业务服务的可用性。

- **在线迁移**

  手动将一台虚拟机从集群的一个物理机迁移到另一台物理机，释放源物理机的资源，支持随机分配和指定物理节点两种模式。

- **离线迁移**

  手动将一台关机的虚拟机从一个集群迁移到另一个集群，调整集群的机器数量，支持指定集群迁移。

- **宕机迁移**

  运行虚拟机的物理机出现异常或故障导致宕机时，调度系统会自动将其所承载的虚拟资源快速迁移至集群内健康且负载正常的物理机，尽量保证业务的可用性。

基于在线迁移、离线迁移和宕机迁移的逻辑，通常在部署上推荐将相同 CPU 和内存配置的物理机节点规划为一个计算集群，避免因 CPU 架构或配置不一致，导致虚拟机迁移后异常或无法启动。

默认情况下平台会根据 CPU 平台架构设定集群名称，管理员可根据平台自身使用情况修改集群名称；同时支持管理员管理计算集群内的物理机和计算实例。

4.1.2.2 存储集群

存储集群为平台分布式块存储集群，通常由一组配置相同的存储节点（物理机）组成，用于部署并承载分布式存储资源。一个数据中心可部署多个不同类型的存储集群，如 SSD 集群、SATA 集群、容量型集群、性能型集群等，不同的集群可提供不同类型的虚拟硬盘源，如 SSD 存储集群可为用户提供 SSD 类型的虚拟硬盘。

平台通过分布式存储集群体系结构提供基础存储资源，并支持在线水平扩容，同时融合智能存储集群、多副本机制、数据重均衡、故障数据重建、数据清洗、自动精简配置、QOS 及快照等技术，为虚拟化存储提供高性能、高可靠、高扩展、易管理及数据安全性保障，全方面提升存储虚拟化及平台的服务质量。

分布式存储集群默认支持 3 副本策略，写入数据时先向主副本写入数据，由主副本负责向其他副本同步数据，并将每一份数据的副本跨磁盘、跨服务器、跨机柜分别存储于不同磁盘上，多维度保证数据安全。在存储集群中存储服务器节点无网络中断或磁盘故障等异常情况时，副本数据始终保持为 3 副本，不区分主副本和备副本；当存储节点发生异常副本数量少于 3 时，存储系统会自动进行数据副本重建，以保证数据副本永久为三份，为虚拟化存储数据安全保驾护航。

默认情况下平台会根据存储架构设定集群名称，管理员可根据平台自身使用情况修改集群名称；同时支持管理员管理存储集群。

4.2 虚拟机

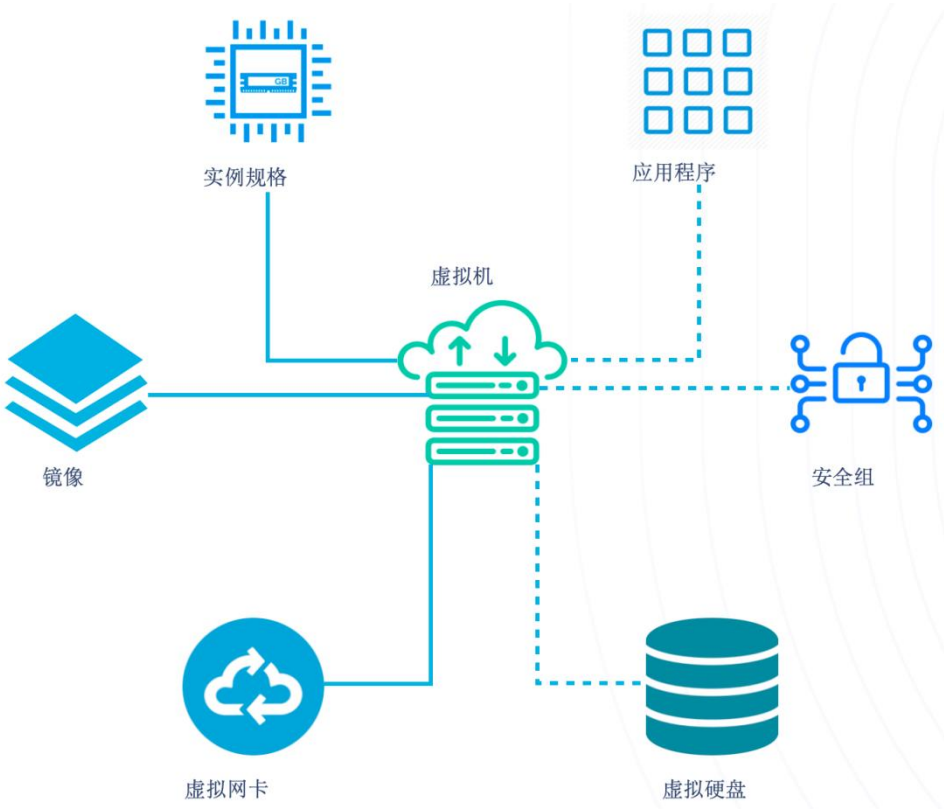
4.2.1 概述

虚拟机是平台的核心服务，提供可随时扩展的计算能力服务，包括 CPU、内存、操作系统等基础的计算组件，并与网络、磁盘等服务结合提供完整的计算环境。

- 平台通过 KVM(Kernel-based Virtual Machine)将物理服务器计算资源虚拟化，为虚拟机提供计算资源；

- 一台虚拟机的计算资源只能位于一台物理服务器上，当物理服务器负载较高或故障时，自动迁移至其它健康的物理服务器；
- 虚拟机计算能力通过虚拟 CPU(vCPU)和内存表示，存储能力通过存储容量和性能体现；
- 虚拟机管理程序通过控制 vCPU、内存及磁盘的 QoS，用于支持虚拟机资源隔离，保证多台虚拟机在同一台物理服务器上互不影响。

虚拟机是平台用户部署并运行应用服务的基础环境，与物理计算机的使用方式相同，提供创建、关机、断电、开机、暂存、重置密码、重装系统、升降级等完全生命周期功能；支持 Linux、Windows 等不同的操作系统，并可通过 VNC、Spice 等方式进行访问和管理，拥有虚拟机的完全控制权限。虚拟机运行涉及资源及关联关系如下：



如图所示，实例规格、镜像、网络是运行虚拟机必须指定的基础资源，即指定虚拟机的 CPU 内存、操作系统、网络信息。在虚拟机基础之上，可绑定虚拟硬盘及安全组，为虚拟机提供数据盘及网络防火墙，保证虚拟机应用程序的数据

存储和网络安全。

在虚拟化计算能力方面，平台提供 GPU 设备透传能力，支持用户在平台上创建并运行 GPU 虚拟机，让虚拟机拥有高性能计算和图形处理能力。**支持透传的设备包括 NVIDIA 的 K80、P40、V100、2080、2080Ti、T4 及华为 Atlas300 等。**

4.2.2 实例规格

实例规格是对虚拟机 CPU 内存的配置定义，为虚拟机提供计算能力。CPU 和内存是虚拟机的基础属性，需配合镜像、虚拟硬盘、安全组，提供一台完整能力的虚拟机。

- 默认提供 1C2G、2C4G、4C8G、8C16G、16C32G、32C64G 等实例规格；
- 支持自定义实例规格，提供多种 CPU 内存组合，以满足不同应用规模和场景的负载要求；
- 支持升降级虚拟机 CPU 和内存配置，可通过更改实例规格进行调整；
- 实例规格通过关机后变更，需重新启动虚拟机生效；

平台支持自定义实例规格，支持为不同集群创建不同的虚拟机规格，即可为不同的机型创建不同的规格，用户创建虚拟机选择不同机型时，即可创建不同规格的虚拟机，适用于不同集群硬件配置不一致的应用场景。可分别定义 CPU 和内存：

- CPU 规格支持（C）：除 1 以外，以 2 为步长进行增加，如 1C、2C、4C、6C、8C，最大值为 240C。
- 内存规格支持（G）：除 1 以外，以 2 为步长进行增加，如 1G、2G、4G、6G、8G ，最大值为 1024G。

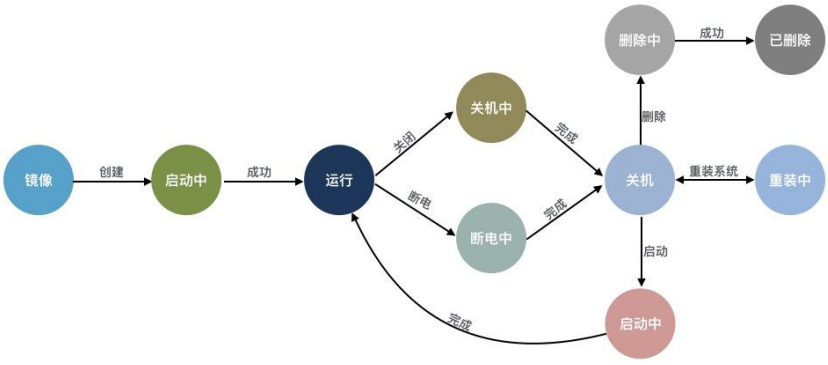
4.2.3 生命周期管理

平台为虚拟机提供完整生命周期管理，用户可自助创建虚拟机，并对虚拟机

进行关机、断电、开机、暂存、重置密码、重装系统、升降级配置、热升级、制作镜像、修改名称/备注、修改告警模板及删除等基本操作；同时支持与虚拟机相关联资源的绑定和解绑管理，包括虚拟硬盘及安全组等。

- 关机是对虚拟机操作系统的正常关机，断电是将虚拟机强制关机；
- 重装系统即更换虚拟机镜像，Linux 仅支持更换 Linux 类型镜像，Windows 仅支持更换 Windows 类型镜像；
- 升降级配置是对虚拟机的规格配置进行升级或降级的变更操作；
- 热升级指在虚拟机开机（running）状态下，支持升级虚拟机的 CPU、内存，不支持在线降级操作；
- 暂存是将运行中的虚拟机的内存信息以及状态保存到文件中，恢复虚拟机时可通过文件加载内存状态数据，实现虚拟机的快速状态还原。
- 销毁虚拟机会自动删除系统盘及默认虚拟网卡，同时会自动解绑相关联的虚拟资源；
- 一个虚拟机支持绑定多个虚拟硬盘。

虚拟机完整生命周期包括启动中、运行、关机中、断电中、关机、启动中、重装中、删除中及已删除等资源状态，各状态流转如下图所示：



4.2.4 镜像服务

镜像（Image）是虚拟机实例运行环境的模板，通常包括操作系统、预装应用程序及相关配置等。虚拟机管理程序通过指定的镜像模板作为启动实例的系统

盘，生命周期与虚拟机一致，虚拟机被销毁时，系统盘即被销毁。平台虚拟机镜像分为基础镜像和自制镜像。

4.2.4.1 基础镜像

基础镜像是由官方提供，包括多发行版 Centos、Ubuntu 及 Windows 等原生操作系统。

默认提供的基础镜像包括 CentOS 6.5 64、CentOS 7.4 64、Windows 2008r2 64、Windows 2012r2 64、Ubuntu 14.04 64、Ubuntu 16.04 64。

- 基础镜像均经过系统化测试，并定期更新维护，确保镜像安全稳定的运行和使用；
- 基础镜像为系统默认提供的镜像，仅支持查看及通过镜像运行虚拟机，不支持修改；

Linux 镜像默认系统盘为 40GB，Windows 镜像默认系统盘为 40GB，支持创建时进行系统盘容量扩容，也可以在虚拟机创建后做系统盘扩容操作（需要用户手动进入虚拟机内部进行文件系统扩容操作）。

支持重装系统，即更换虚拟机镜像，Linux 虚拟机仅支持更换 Centos 和 Ubuntu 操作系统，Windows 虚拟机仅支持更换 Windows 其它版本的操作系统。

> **注意** Windows 操作系统镜像为微软官方提供，需自行购买 License 激活。

4.2.4.2 自制镜像

自制镜像由用户通过虚拟机自行制作或自定义导入已有的自有镜像，可用于创建虚拟机，平台用户有权限查看和管理。

- 支持用户制作、导入和导出自定义镜像
- 支持通过自制镜像创建虚拟机、删除自制镜像、修改自制镜像名称
- 支持用户自有 ISO 镜像导入

ISO 镜像是一种将光盘或 DVD 中的数据以文件的形式保存在计算机硬盘上的方法。当需要使用一个全新的操作系统时，可以选择使用包含操作系统介质的 ISO 镜像，直接使用 ISO 介质引导并安装到虚拟机中。通过这种方式，可以快速的部署一个全新的虚拟机。在部署完成后通常会进行以下操作，用于进一步提升将来的部署效率。

- 将安装完操作系统的虚拟机，修改符合企业标准的系统参数，并转化成一个标准的虚拟机镜像，方便下一次更加高效的创建一个符合企业信息化标准的全新虚拟机实例。
- 进一步部署相关的业务系统，转化成一个包含业务系统的虚拟机镜像，以提升下一次部署相同业务时的效率，同时也保障业务系统的符合企业信息化标准规范。

自制镜像和 ISO 镜像可用于创建虚拟机，并支持用户下载虚拟机镜像到本地，同时镜像管理支持查看镜像、修改名称和备注、从镜像创建虚拟机、导入镜像、下载镜像及删除镜像等生命周期管理。

4.2.4.3 镜像存储

基础镜像和用户自制镜像默认均存储于分布式存储系统，保证性能的同时通过三副本保证数据安全。

- 镜像支持 QCOW2 格式，可将 RAW、VMDK 等格式镜像转换为 QCOW2 格式文件，用于 V2V 迁移场景；
- 所有镜像均存储于分布式存储系统，即镜像文件会分布在底层计算存储超融合节点磁盘上；
- 若为独立存储节点，则分布存储于独立存储节点的所有磁盘上；
- 地域的镜像只能创建本地域的虚拟机，不支持跨地域镜像创建虚拟机。

4.2.5 虚拟机存储

虚拟机的系统盘和数据盘支持块存储作为后端存储，并统一池化为虚拟硬

盘，用户可以像使用物理机硬盘一样格式化并建立文件系统来使用虚拟硬盘。

针对虚拟机的虚拟硬盘和数据安全，平台支持虚拟硬盘加密特性，使用 LUKS 加密规范来对磁盘全盘加密，保护用户的数据不被未经授权的访问者获取，甚至在磁盘丢失或被盗的情况下也可以保证数据的机密性。

(1) 虚拟硬盘

一种基于分布式存储系统为虚拟机提供持久化存储空间的块设备。虚拟硬盘基于网络分布式访问，为虚拟机提供高安全、高可靠、高性能及可扩展的数据磁盘。可用于虚拟机的系统盘和数据盘。

- 支持对虚拟硬盘类型的系统盘进行扩容、快照及加密。
- 支持对虚拟硬盘类型的数据盘进行绑定、解绑、扩容、快照及加密。
- 支持通过虚拟硬盘创建并启动虚拟机。
- 支持将虚拟硬盘设置为共享盘，多个虚拟机同时进行挂载使用。

> **说明** 平台支持对虚拟硬盘本身进行全生命周期管理，包括虚拟硬盘创建、查看、绑定、解绑、扩容、克隆、删除、快照、设为共享盘等，详见【虚拟硬盘】章节描述。

(2) 共享盘

共享盘是一种数据块级存储设备，能够同时支持多个虚拟机并发读写访问。这种存储设备具有多挂载点、高可靠性等特点，适用于需要支持集群和高可用性（HA）能力的关键企业应用场景，多个虚拟机可以同时访问一个共享盘。

支持将虚拟硬盘设置为共享盘，并作为虚拟机的数据盘，使多个虚拟机同时对共享盘进行数据读写操作。同时支持对共享盘进行创建、绑定、解绑、扩容、克隆及删除等操作。

4.2.6 存储热迁移

存储热迁移是平台提供的在不中断虚拟机运行的情况下可以动态更换虚拟机存储的能力，支持对 Intel/AMD x86 架构虚拟机的任意磁盘进行动态更换。该

能力允许在虚拟机运行的同时进行存储迁移变更，可以在不影响业务连续性的情况下执行存储维护、变更或优化操作，以提高系统的可用性。

4.2.6.1 应用场景

存储热迁移的能力不仅限于系统盘，同时支持对虚拟机中任意数据盘的动态更换，通常应用于如下场景：

- **存储空间均衡**

通过创建新的存储集群对平台进行存储扩容后，通常面临着新旧存储集群空间使用不均衡的问题。存储热迁移的能力使得在不中断虚拟机业务的前提下，可以灵活地调整存储资源，实现存储空间的均衡分配，从而有效缓解原有存储集群的存储压力，提升整体系统的性能和稳定性。

- **更换高性能存储**

随着业务对虚拟机磁盘性能需求的提升，可能需要将虚拟机磁盘更换为性能更高的存储集群。通过存储热迁移的能力，可以在不中断虚拟机服务的情况下，将虚拟机磁盘迁移到高性能存储上，满足业务对性能的更高要求，提升应用的响应速度和整体处理能力。

- **存储设备下线**

在存储设备需要进行维护或下线的情况下，存储热迁移能够实现虚拟机磁盘的平稳迁移，将数据从即将下线的存储设备迁移到长期稳定可用的存储集群中。这种无缝迁移的过程保证了虚拟机业务的连续性，同时让管理员可以灵活地进行存储设备的维护工作，确保整个环境的可维护性和稳定性。

4.2.6.2 迁移模式

平台支持通过内置分布式存储为虚拟机提供虚拟硬盘，存储热迁移能力支持在不同存储集群之间进行热迁移。迁移时选择目标存储集群即可，平台会自动在目标集群创建新存储并在迁移完毕后自动删除旧存储。

注意 共享盘允许多个虚拟机同时挂载使用，不支持针对共享盘进行存储热迁移。

4.2.6.3 迁移过程

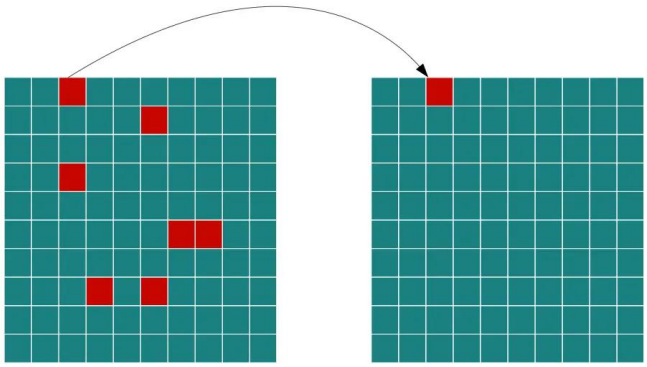
运行中的虚拟机在不断变更存储中的数据，为确保在迁移完成时目标存储包含迁移过程中的所有变更数据，平台通过迁移准备、数据迁移、迁移收尾三个阶段实现完整的存储热迁移。

(1) 迁移准备

将磁盘按 block 为单位组织成一个数组，块级别的组织结构使得系统能够更加细致地管理和操作数据，便于系统实现对数据的分批迁移。同时开启脏块记录机制用于数据变更追踪，系统在进行数据迁移时可以有选择性地迁移仅包含脏块的部分，降低了迁移的成本和复杂性。

(2) 数据迁移

首先进行磁盘的全量数据迁移，依次将每个 block 迁移到目标存储。通过将整个磁盘的数据迁移到目标存储，建立一个基准状态，用于确保后续增量迁移的准确性和完整性。



然后通过多次迭代，将迁移过程中虚拟机产生的新数据迁移到目标存储。迭代过程逐渐收敛的特性使得迁移操作更具渐进性和可控性。Qemu 将边迁移边记录剩下的脏数据大小，随着迭代的进行，剩余的脏数据不断减小，最终收敛到一个较小的值。迭代收敛的过程确保了最终一致性，当脏数据逐渐趋近零时，系统达到了一种稳定状态，Qemu 进程就会暂停，从而避免产生新的脏数据，以便

进行迁移收尾工作。

(3) 迁移收尾

在虚拟机暂停之后，整个迁移过程进入第三阶段的收尾工作，这一阶段的主要任务是确保迁移数据的完整性和一致性。Qemu 进程会将剩余的磁盘脏数据一次性同步到目标端，完成时虚拟机新旧存储的数据将会一致。

4.2.7 虚拟网卡

虚拟网卡（Virtual NIC）是虚拟机与外部通信的虚拟网络设备，创建虚拟机时会自动创建虚拟网卡。虚拟网卡与虚拟机的生命周期一致，无法进行分离，虚拟机被销毁时，虚拟网卡即被销毁。

虚拟网卡基于 Virtio 实现，QEMU 通过 API 对外提供一组 Tun/Tap 模拟设备，将虚拟机的网络桥接至宿主机网卡，通过 OVS 与其它虚拟网络进行通信。

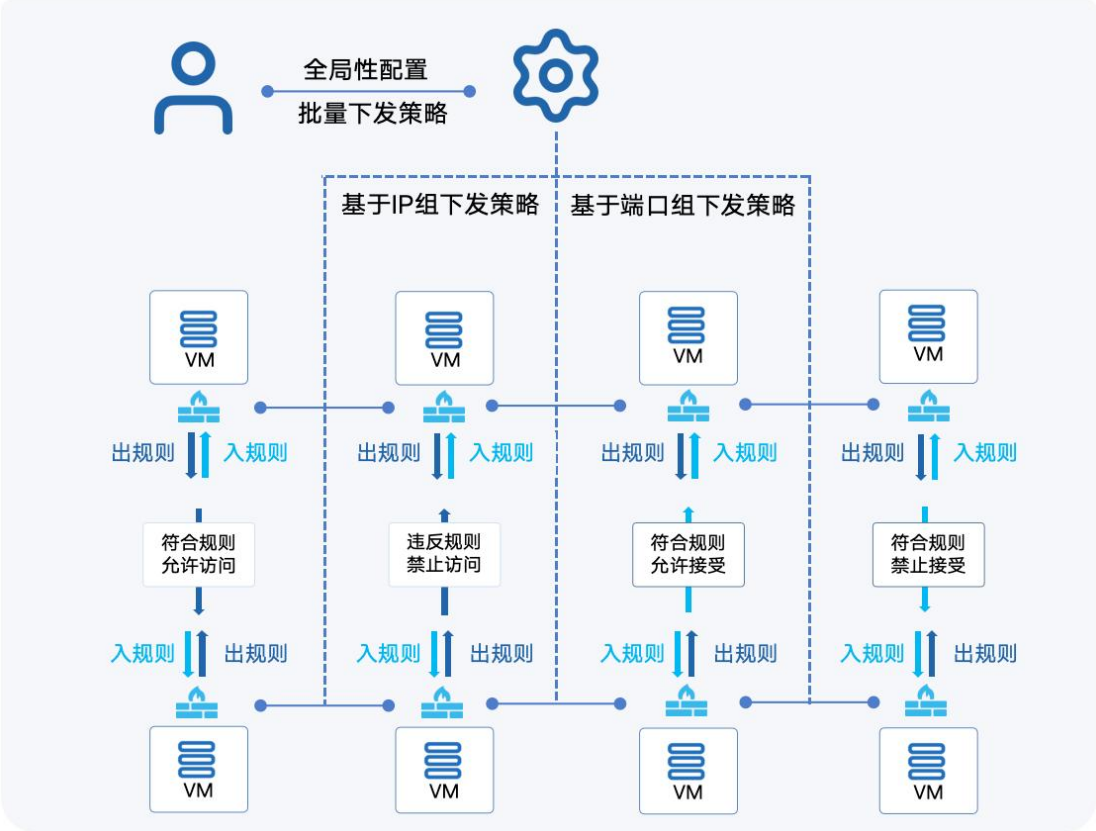
- 每个虚拟机会生成一块虚拟网卡，用于虚拟机与扁平网络的通信。
- 加入开启 IP 地址管理的网络时，虚拟机会自动分配和配置 IP 地址。
- 虚拟网卡支持绑定安全组，提供网卡级别安全控制。

4.2.8 安全组

安全组（Security Group）是一种类似 IPTABLES 的**虚拟防火墙**，提供出入双方向流量访问控制规则，定义哪些网络或协议能访问资源，用于限制虚拟资源的网络访问流量，支持 IPv4 和 IPv6 双栈限制，为平台提供必要的安全保障。

4.2.8.1 实现机制

平台安全组基于 Linux Netfilter 子系统，通过在 OVS 流表中添加流表规则实现，需开启宿主机 IPv4 和 IPv6 包转发功能。每增加一条访问控制规则会根据网卡作为匹配条件，生成一条流表规则，用于控制进入 OVS 的流量，保证虚拟资源的网络安全。



安全组具有独立的生命周期，可以将安全组与虚拟机绑定在一起，提供安全访问控制，与之绑定的虚拟资源销毁后，安全组将自动解绑。

- 安全组对虚拟机的安全防护针对的是网卡，即安全组是与虚拟机的虚拟网卡绑定在一起，通过访问控制规则限制网卡的出入网络流量；
- 一个安全组支持同时绑定至多个虚拟机实例；

创建安全组规则时，支持通过引用端口组及 IP 组一次性定义多个相关的协议、端口和 IP 地址，减少规则数量有助于降低配置的复杂性。

支持创建虚拟机时不指定安全组，支持虚拟机启动后再进行调整，支持随时修改安全组的出入站规则，新规则生成时立即生效，可根据需求调整安全组出/入方向的规则。支持安全组全生命周期管理，包括安全组创建、修改、删除及安全组规则的创建、修改、删除等生命周期管理。

4.2.8.2 安全组规则

安全组规则可控制允许到达安全组关联资源的入站流量及出站流量，提供双栈控制能力，支持对 IPv4/IPv6 地址的 TCP、UPD、ICMP、GRE 等协议数据包进行有效过滤和控制。

每个安全组支持配置多条规则，根据优先级对资源访问依次生效。**规则为空时，安全组将默认拒绝所有流量；规则不为空时，除已生成的规则外，默认拒绝其它访问流量。**

支持有状态的安全组规则，可以分别设置出入站规则，对被绑定资源的出入流量进行管控和限制。每条安全组规则由协议、端口、地址、动作、优先级及方向六个元素组成：

- 协议：支持 TCP、UDP、ICMPv4、ICMPv6 四种协议数据包过滤。
  - ALL 代表所有协议和端口，ALL TCP 代表所有 TCP 端口，ALL UDP 代表所有 UDP 端口；
  - 支持快捷协议指定，如 FTP、HTTP、HTTPS、PING、OpenVPN、PPTP、RDP、SSH 等；
  - ICMPv4 指 IPv4 版本网络的通信流量；ICMPv6 指 IPv6 版本网络的通信流量。
- 端口：源地址访问的本地虚拟资源或本地虚拟资源访问目标地址的 TCP/IP 端口。
  - TCP 和 UDP 协议的端口范围为 1~65535 ；
  - ICMPv4 和 ICMPv6 不支持配置端口。
- 地址：访问安全组绑定资源的网络数据包来源地址或被安全组绑定虚拟资源访问的目标地址。
  - 当规则的方向为入站规则时，地址代表访问被绑定虚拟资源的源 IP 地址段，支持 IPv4 和 IPv6 地址段；

- 当规则的方向为出站规则时，地址代表被绑定虚拟资源访问目标 IP 地址段，支持 IPv4 和 IPv6 地址段；
- 支持 CIDR 表示法的 IP 地址及网段，如 120.132.69.216 、0.0.0.0/0 或 ::/0 。

- 动作：安全组生效时，对数据包的处理策略，包括 “接受” 和 “拒绝” 两种动作。
- 优先级：安全组内规则的生效顺序，包括高、中、低三档规则。
  - 安全组按照优先级高低依次生效，优先生效优先级高的规则；
  - 同优先级的规则，优先生效精确规则。
- 方向：安全组规则所对应的流量方向，包括出站流量和入站流量。
- 描述：每一条安全组规则的描述，用于标识规则的作用。

安全组支持数据流表状态，规则允许某个请求通信的同时，返回数据流会被自动允许，不受任何规则影响。即安全组规则仅对新建连接生效，对已经建立的链接默认允许双向通信。如一条入方向规则允许任意地址通过互联网访问虚拟机外网 IP 的 80 端口，则访问虚拟机 80 端口的返回数据流（出站流量）会被自动允许，无需为该请求添加出方向允许规则。

说明 通常建议设置简洁的安全组规则，可有效减少网络故障。

4.2.8.3 端口组和 IP 组

创建安全组规则时，除了通过自定义端口方式指定端口信息外，平台提供的端口组功能允许用户更灵活地定义安全组规则。用户可以将一组端口和协议组织成一个逻辑单元，并在安全组规则中引用该端口组，从而简化规则的管理和维护。

平台提供的 IP 组功能提供了类似的灵活性，允许用户指定包含多个 IP 地址、网段或连续地址段的规则。这使得在规则定义中，可以更加便捷地表示一组相关的网络地址，从而提高规则的可读性和管理效率。

端口组和 IP 组功能允许用户在单个规则中定义多个协议、端口及 IP 信息，从而简化规则的配置。用户无需为每个协议或端口创建单独的规则，而是可以通过选择端口组及 IP 组一次性定义相关的协议、端口和 IP 地址，这样的简化有助于降低配置的复杂性。通过集中管理相关信息和减少规则数量有助于降低配置错误的风险。

(1) 端口组

创建端口支持以下格式:

- 单个端口，如：80
- 多个单端口，如：80,443
- 连续端口段，如：3306-20000

单个端口组可包含多条“协议:端口”信息，如：

```
TCP:80,443,3306-10000
UDP:53,1000
```

(2) IP 组

创建 IP 支持以下格式:

- 单个 IP，如：10.0.0.1 或 FF05::B5
- 网段，如：10.0.1.0/24 或 FF05:B5::/60
- 连续地址段，如：10.0.0.1-10.0.0.100

单个 IP 组可包含多个 IP 信息，如：

```
10.0.1.10
172.16.1.0/24
10.0.1.100-10.0.1.200
```

4.2.9 隔离组

隔离组，又称亲和反亲和策略，允许用户自定义虚拟机与其它虚拟机或宿主机之间的调度关系，意味着用户可以根据业务需求、性能要求或其它因素定义虚

拟机调度逻辑，以满足特定的架构和运行要求。

隔离组提供了灵活的资源调度机制，用户可以根据实际需求动态地调整隔离组的配置，以适应不同业务负载和调度需求。隔离组根据策略对象类型分为虚拟机组和节点组。

4.2.9.1 虚拟机组

策略对象类型为虚拟机组的隔离组，用于控制一组虚拟机之间或多组虚拟机之间的调度关系，支持亲和、反亲和两种策略。

(1) 亲和策略

亲和策略用于将策略相关的虚拟机实例调度部署在同一物理主机上，可以最大程度地提高它们之间的数据传输和通信效率。物理主机内部的虚拟机通信会经过高速内部网络，从而实现更高的性能和更低的通信延时。在同一物理主机上运行的虚拟机之间的通信不需要经过物理网络，减少了对网络带宽的占用，有助于减轻网络拥塞和提高整体系统的可扩展性。

亲和策略通常适用于有业务关联性的虚拟机实例，例如构成同一业务服务的多个虚拟机，虚拟机之间通常需要频繁地进行通信和数据共享，因此将它们调度到同一物理主机上有助于提高整体服务的性能。

- 非强制亲和：当物理节点资源小于虚拟机资源需求时，系统将会忽略亲和策略，选择其它合适的物理主机创建虚拟机。
- 强制性亲和：当物理节点资源无法满足虚拟机资源需求时，则虚拟机会因不满足强制性策略而阻塞调度。

(2) 反亲和策略

反亲和策略，也被称为非亲和性策略，是将虚拟机实例分散在不同物理主机上的调度策略，旨在降低单点故障的风险，提高系统的可靠性和容错性。通过分散调度降低了某一物理主机发生故障对整个系统造成影响的风险，当某个节点出现问题时，其它节点上的虚拟机可以继续运行，确保整个系统的持续性，以提高

系统的可靠性、容错性和弹性。

- 非强制反亲和：当集群节点资源不能满足相关虚拟机完全分散调度时，系统将会忽略反亲和策略，从而出现多个虚拟机调度到相同物理主机的情况。
- 强制性反亲和：当集群节点资源不能满足相关虚拟机完全分散调度时，则虚拟机会因不满足强制性策略而阻塞调度。

4.2.9.2 节点组

策略对象类型为节点组的隔离组，用于控制虚拟机和物理主机之间的调度关系。创建后支持加入目标节点及虚拟机实例，支持亲和、反亲和两种策略。

(1) 亲和策略

将组内的虚拟机实例调度部署在已加入组内的物理主机上，以实现虚拟机启动时的定向调度。

- 非强制亲和：当物理节点资源小于虚拟机资源需求时，系统将会忽略亲和策略，选择其它合适的物理主机创建虚拟机。
- 强制性亲和：当物理节点资源无法满足虚拟机资源需求时，则虚拟机会因不满足强制性策略而阻塞调度。

(2) 反亲和策略

将组内的虚拟机实例避免调度在已加入组内的物理主机上。

- 非强制反亲和：当集群节点资源不能满足虚拟机资源需求时，系统将会忽略反亲和策略，从而出现组内虚拟机调度到组内物理主机的情况。
- 强制性反亲和：当集群节点资源不能满足虚拟机资源需求时，则虚拟机会因不满足强制性策略而阻塞调度。

4.2.10 USB 透传

平台支持 USB 透传功能，物理机 USB 设备可直接透传至该物理机上所运

行的虚拟机，USB 设备包含以下两种模式：

- **直通**

将 USB 设备加载到此物理机上的虚拟机，迁移虚拟机时需要卸载此 USB 设备，通常用于对 USB 设备性能有要求的场景。

- **转发**

将 USB 设备加载到此物理机所在计算集群，通过网络转发 USB 设备内的数据，迁移虚拟机时不需要卸载此 USB 设备。

4.2.11 VNC 登录

VNC（Virtual Network Console）是平台为用户提供的一种通过 WEB 浏览器连接虚拟机的登录方式，适应于无法通过远程登录客户端（如 SecureCRT、PuTTY 等）连接虚拟机的场景。通过 VNC 登录连到虚拟机，可以查看虚拟机完整启动流程，并可以像 SSH 及远程桌面一样管理虚拟机操作系统及界面，支持发送各种操作系统管理指令，如 CTRL+ALT+DELETE。

支持用户获取虚拟机的 VNC 登录信息，包括 VNC 登录地址及登录密码，适用于使用 VNC 客户端连接虚拟机的场景，如桌面云场景。为确保 VNC 连接的安全性，每一次调用 API 或通过界面所获取的 VNC 登录信息有效期为 300 秒，如果 300 秒内用户未使用 IP 和端口进行连接，则信息直接失效，需要重新获取新的登录信息；同时用户使用 VNC 客户端登录虚拟机后，300 秒内无任何操作将会自动断开连接。

支持用户获取虚拟机的 Spice 登录信息，包括 Spice 登录地址及登录密码，同样适用于使用 Spice 客户端连接虚拟机的场景，如桌面云场景，与 VNC 连接一致，限制有效期 300 秒，保证连接的安全性。

4.2.12 自定义启动源

平台支持自定义虚拟机启动源，不仅可以采用常规方式选择镜像进行虚拟机创建，还可以充分发挥灵活性，选择已有的虚拟硬盘作为系统盘进行虚拟机的创

建和启动。

在选择已有盘作为启动源时，要求所选盘中包含完整的操作系统，以确保虚拟机能够在启动时正常运行。该功能赋予用户极大的灵活性，使虚拟机的创建方式更为快捷、个性化，为用户提供了定制化的虚拟机创建方式。

4.2.13 自定义主机名称

支持自定义虚拟机主机名称，用于自动设置虚拟机操作系统内部的计算机名。批量创建时会在当前填写主机名添加有序后缀。

- Windows 系统，长度为 2~15 个字符。允许使用大小写字母、数字或连接符（-）。不能以连字符（-）开头或结尾，不能连续使用连字符（-），也不能仅使用数字。
- Linux 系统，长度为 2~63 个字符。允许使用大小写字母、数字、点号（.）或连接符（-）。不能以点号（.）或连字符（-）开头或结尾，不能连续使用点号（.）或连字符（-），也不能仅使用数字。

4.2.14 自定义 DNS

虚拟机默认 DNS 指定为 114.114.114.114，用于提供通用域名解析服务。同时系统支持用户自定义配置最多两个自定义 DNS 地址。

企业内部网络通常会使用自己的 DNS 服务器来处理内部域名解析，用户可以选择自定义虚拟机的 DNS 配置，以满足特定网络环境或业务需求。通过自定义 DNS，虚拟机能够适应特定的企业网络环境，确保内部域名的正确解析。

4.2.15 自定义 MAC

虚拟机的 MAC 地址为平台随机分配，以避免冲突和确保网络的唯一性。为满足特定场景下如授权与指定 MAC 地址绑定的需求，用户可以选择自定义虚拟机的 MAC 地址，在控制台上即可对关机状态下的虚拟机进行 MAC 地址设定。

在自定义 MAC 地址时，需要确保所选择的地址在整个网络中是唯一的，平台会对输入的 MAC 地址进行重复性校验以避免冲突。

4.2.16 自定义引导方式

虚拟机默认通过 BIOS 固件引导，对于启动盘磁盘格式是 GPT 的镜像需要通过 UEFI 方式引导。创建虚拟机时，引导方式默认和所选镜像保持一致，平台支持自定义选择引导方式，支持 BIOS 和 UEFI。

- BIOS 使用 Master Boot Record（MBR）进行引导，使用文本模式的启动界面，限制在 2TB 以下的硬盘容量，启动速度相对较慢，标准成熟，对于一些老旧的硬件和操作系统具有较好的兼容性。
- UEFI 使用 GUID Partition Table（GPT）进行引导，支持图形化的用户界面，提供更直观的操作和信息显示，可以支持大于 2TB 的硬盘容量，启动速度更快，支持并行加载驱动和应用程序。

4.2.17 自定义 CPU 启动模式

支持虚拟机自定义选择 CPU 启动模式，分为默认（Custom）和直通（Host-passthrough），方便用户根据使用场景灵活选择。

- Custom 模式通过提炼 CPU 通用指令集，最大限度保障了虚拟机在不同宿主机之间热迁移时的兼容性。
- Host-passthrough 模式将宿主机的 CPU 指令集全部透传给虚拟机，可以最大限度的使用宿主机 CPU 指令集，但是在热迁移时要求目的节点的 CPU 和源节点的完全一致。

4.2.18 自定义高可用模式

支持设定虚拟机的高可用模式，提供【永不停止】和【无】两种选项，方便用户根据业务类型选择合适的高可用性模式。

- 选择【永不停止】模式的虚拟机将会开启高可用模式，确保虚拟机关闭后能够自动重启，无论是因为宿主机故障、虚拟机异常关闭或则其它原因，虚拟机将会尽快重新启动，使其处于长期运行状态，以维持业务的连续性和可用性。

- 选择【无】模式的虚拟机不启用高可用模式，虚拟机关闭后将不会自动重启，对于不需要长期运行的虚拟机，便于用户更灵活的控制运行状态。

4.3 GPU 虚拟机

4.3.1 概述

平台提供 GPU 设备透传能力，支持用户在平台上创建并运行 GPU 虚拟机，让虚拟机拥有高性能计算和图形处理能力。

GPU 虚拟机可以提供更好的成本效益。通过共享和灵活分配 GPU 资源，可以更有效地利用硬件资源，降低硬件投资和运营成本。同时虚拟机的动态调整和弹性扩展功能，可以根据实际需求进行资源分配，避免资源浪费。

- 支持用户选择 GPU 颗数，选择 GPU 规格创建 GPU 虚拟机，GPU 虚拟机与虚拟机的管理功能和生命周期一致。
- 支持用户关机状态下修改虚拟机配置、解绑 GPU。

支持透传的设备包括 NVIDIA 的 K80、P40、V100、2080、2080Ti、T4 及华为 Atlas300 等。

针对 GPU 虚拟机，平台支持最高配置 4 颗 GPU 芯片，为使 GPU 虚拟机发挥最佳性能，平台限制最小 CPU 内存规格为 GPU 颗数的 4 倍以上：

- 1 颗 GPU 芯片最小需要 4 核 8G 规格
- 2 颗 GPU 芯片最小需要 8 核 16G 规格
- 4 颗 GPU 芯片最小需要 16 核 32G 规格

4.3.2 应用场景

- **GPU 资源共享**

GPU 虚拟机允许多个用户共享同一台物理服务器上的 GPU 资源。每个虚拟机实例可以分配物理服务器上的一个或多个 GPU 资源，以满足不同用户的

需求。

- **高性能图形处理**

GPU 虚拟机提供了强大的图形处理能力，可以加速图形密集型任务，如游戏渲染、图像处理和视频编码等。通过虚拟化技术，多个用户可以同时享受到高性能的图形处理能力。

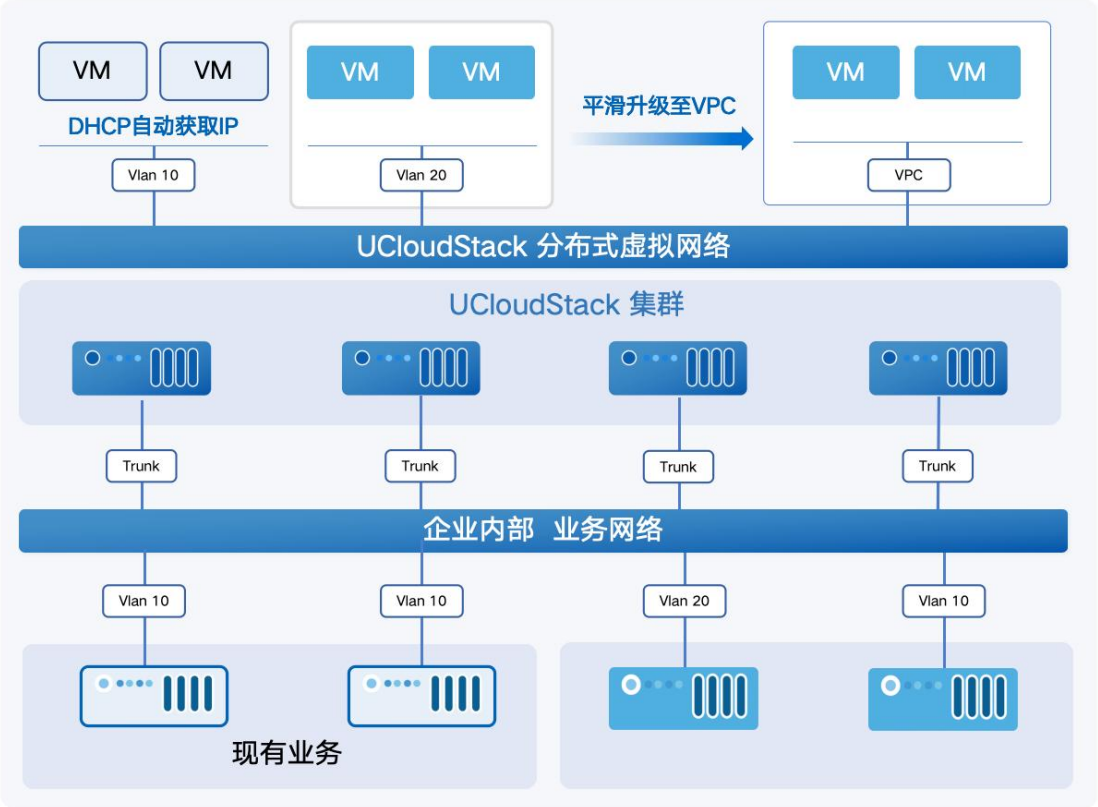
- **GPU 加速计算**

GPU 虚拟机不仅可以用于图形处理，还可以用于加速通用计算任务。虚拟机实例可以利用 GPU 的并行计算能力，加速科学计算、机器学习、数据分析、AI 训练、AI 推理等工作负载。

4.4 扁平网络

4.4.1 概述

扁平网络是一种易用高效的网络模型，通过简化网络拓扑、降低延迟、提高性能，提升网络的灵活性和易用性，为虚拟化环境提供简单高效的网络解决方案。在扁平网络中，虚拟机直接连接到二层网络，无需经过隧道网络，从而避免数据封包和解包的额外开销。这种连接方式使得集群内工作负载的通信更为直接，减少网络流量的路径，从而提高网络性能。



扁平网络使得网络结构更为简单灵活，当需要调整网络拓扑时，管理员可以迅速进行配置更改，无需深入考虑复杂的网络层次结构，可以根据业务需求和变化快速扩展或调整网络配置。这种灵活性使得网络可以更好地满足业务需求，而不必受制于复杂的网络结构。

4.4.2 功能特性

扁平网络通过虚拟交换机实现，通过 VLAN 实现多个扁平网络之间的隔离及外部物理网络的打通。通过提供 IP 地址管理、自定义路由及内置 DHCP 服务，以简化平台网络的使用和管理。平台通过对集群内所有虚拟交换机的统一配置管理，为虚拟机提供简单高效、集群内配置一致的网络连接。

- **多网络支持**

  平台支持根据网络规划创建多个扁平网络，创建时需指定网卡设备，通常该设备由两个物理网卡组合为 bond 以提供网络设备的高可用，且配置为 Trunk 模式用于承载多个 VLAN 的网络数据。每个扁平网络通过独立的虚拟交换机实现，为平台虚拟机提供多个不同 VLAN 隔离的二层网络。

- **IP 地址管理**

除基础的二层网络连接能力外，平台针对扁平网络提供 IP 地址自动管理功能，允许为扁平网络预设一个指定的 IP 地址段，加入到该网络的虚拟机可以分配和自动配置该网段的一个 IP 地址，以简化虚拟机 IP 地址的管理和分配过程。

- **自定义路由**

已开启 IP 地址管理的扁平网络，平台支持为该网络设置自定义路由，用户可以根据网络拓扑和需求，定义适用于扁平网络的路由规则，包括目的地网络、下一跳信息。平台自动为加入该网络的虚拟机下发路由信息，且路由信息变更时可以实时更新虚拟机内部的自定义路由信息。

- **DHCP 服务**

通过开启 DHCP 功能平台为该网络提供内置的 DHCP 服务，对虚拟机的 DHCP 网络请求进行响应，以满足使用传统 IP 地址请求分配的网络方式。

4.4.3 IP 地址管理

平台针对扁平网络同时提供 IP 地址的自动管理功能。允许管理员为扁平网络预设一个指定的 IP 地址段，加入到该网络的虚拟机可以自动分配和配置该网段的一个 IP 地址，从而简化虚拟机 IP 地址的管理和分配过程，提高网络配置的效率和便捷性。

- **预设 IP 地址段：**管理员可以在创建扁平网络时预设一个指定的 IP 地址段。这有助于网络管理员更好地控制和规划 IP 地址的分配，确保网络中的 IP 地址符合特定的规范和需求。
- **IP 地址自动管理：**虚拟机加入扁平网络时，系统自动分配和配置该网络预设的 IP 地址段中的一个 IP 地址，避免手动规划和分配。
- **简化管理和分配：**用户无需手动为每个虚拟机配置 IP 地址，简化虚拟机 IP 地址的管理和分配过程，提高整体网络配置的效率。

- **降低配置错误风险：**自动管理 IP 地址减少手动配置 IP 地址时可能发生的错误风险。系统根据预设的规则进行自动分配，减少配置错误的可能性，提高网络的稳定性和可靠性。

通过 IP 地址自动管理功能，平台为扁平网络提供更加简单易用的网络配置选项，使得管理员能够更轻松地进行网络规划和管理。对于大规模虚拟化环境尤为重要，可以降低网络配置的复杂性，提高整个系统的可维护性。

4.4.4 自定义路由

对于已开启 IP 地址管理的扁平网络，平台提供更进一步的网络功能，即通过设置自定义路由规则，自动为加入该网络的虚拟机下发自定义路由信息。

- **自动路由下发：**扁平网络支持管理员设置自定义路由规则，以适应特定的网络拓扑和需求。当虚拟机加入该网络时，平台会根据预设的路由规则自动下发路由信息，确保虚拟机能够正确地进行网络通信。
- **简化网络配置：**自动下发路由信息可以简化网络配置的复杂性。管理员无需手动配置每个虚拟机的路由信息，系统能够自动化地为虚拟机提供正确的路由，减轻管理负担。
- **提高网络灵活性：**自定义路由规则的设置可以增加网络的灵活性。管理员可以根据实际需求定制路由策略，以适应不同的应用场景和业务需求，提高网络配置的适应性。

自定义路由及自动化下发能力，进一步增强网络的可管理性，使得网络管理更为灵活和高效。

4.4.5 DHCP 服务

平台为已开启 IP 地址管理的扁平网络提供内置的 DHCP 网络服务，用户可根据实际需求选择是否开启。

对于加入扁平网络的虚拟机，平台优先通过 QEMU Guest Agent （QGA）对虚拟机进行 IP 地址的自动化配置，没有安装 QGA 的虚拟机除手动配置方式

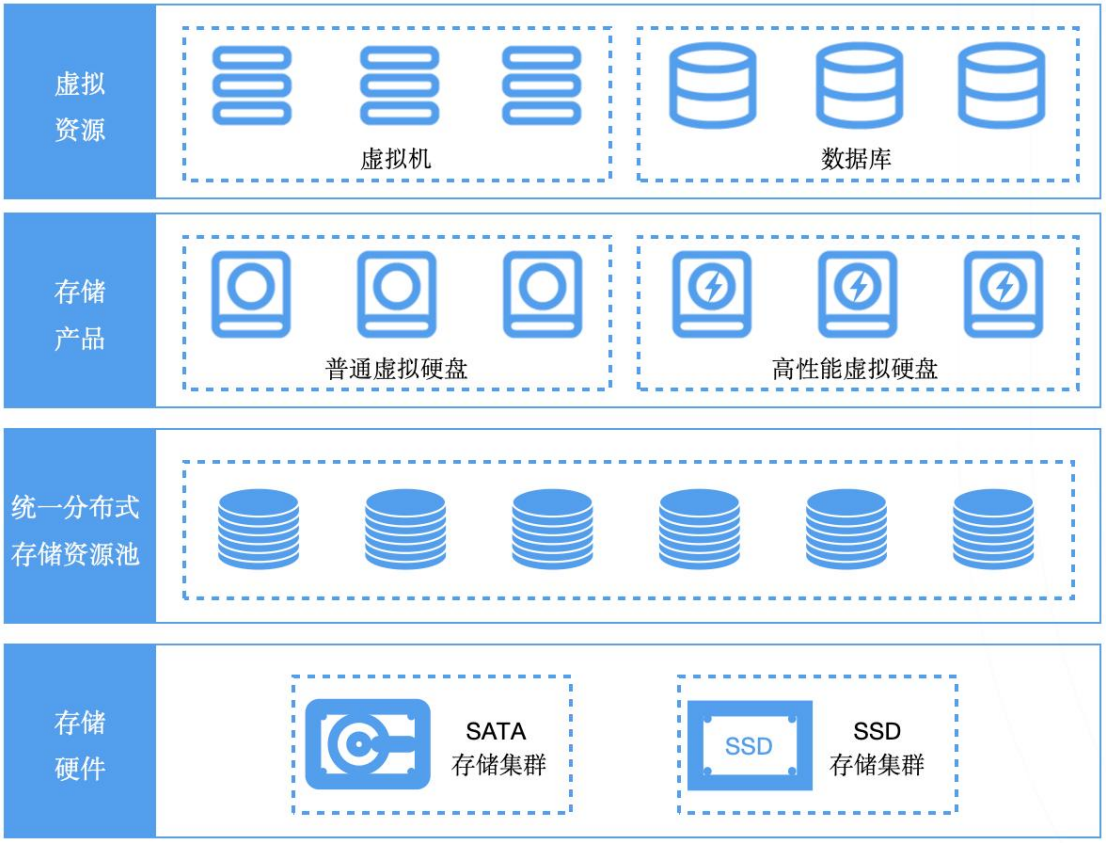
外，则可以通过 DHCP 客户端进行 IP 地址请求及自动化配置。

通过平台提供的 DHCP 功能，用户无需再额外搭建 DHCP 服务，简化整个网络配置的部署过程，减少用户的工作量和网络复杂性。

4.5 虚拟硬盘

4.5.1 虚拟硬盘概述

虚拟硬盘是一种基于分布式存储系统为虚拟机提供持久化存储空间的块设备。具有独立的生命周期，支持绑定/解绑至虚拟机使用，并能够在存储空间不足时对虚拟硬盘进行扩容，基于网络分布式访问，为虚拟机提供高安全、高可靠、高性能及可扩展的数据磁盘。



存储系统兼容并支持多种底层存储硬件，如通用服务器（计算存储超融合或独立通用存储服务器），并将底层存储硬件分别抽像不同类型集群的存储资源池，由分布式存储系统统一调度和管理。在实际应用场景中，可以将普通 SATA 接口

的机械盘统一抽像为【SATA 存储集群】，将 SSD 全闪磁盘统一抽象为【SSD 存储集群】，分别由统一存储封装后提供平台用户使用。

如示意图所示，将 SATA 存储集群的资源封装为普通虚拟硬盘，将 SSD 全闪存储集群的资源封装为高性能虚拟硬盘。平台的虚拟机可根据需求挂载不同存储集群类型的磁盘，支持同时挂载多种集群类型的虚拟硬盘。平台管理员可通过控制台自定义存储集群类型的别名，用于标识不同磁盘介质、不同品牌、不同性能或不同底层硬件的存储集群，如 EMC 存储集群、SSD 存储集群等。

通常 SSD 磁盘介质的虚拟硬盘的性能与容量的大小成线性关系，容量越大提供的 IO 性能越高，如对 IO 性能有强烈需求，可考虑扩容 SSD 磁盘介质的虚拟硬盘。

分布式存储底层数据通过 PG 映射的方式进行数据存储，同时以多副本存储的方式保证数据安全，即写入至平台存储集群的数据块会同时保存多份至不同服务器节点的磁盘。

多副本存储的数据提供一致性保证，可能导致写入的多份数据因误操作或原始数据异常导致数据不准确；为保证数据的准确性，平台提供硬盘快照能力，将虚拟硬盘数据在某一时间点的数据文件及状态进行备份，在数据丢失或损坏时，可通过快照快速恢复数据，包括数据库数据、应用数据及文件目录数据等，可实现分钟级恢复。

4.5.2 功能与特性

虚拟硬盘由统一存储从存储集群容量中分配，为平台虚拟资源提供块存储设备并共享整个分布式存储集群的容量及性能；同时通过块存储系统为用户提供虚拟硬盘资源及全生命周期管理，包括虚拟硬盘的创建、绑定、解绑、扩容、克隆、快照及删除等管理。

虚拟硬盘容量是由统一存储的从存储集群容量中分配的，所有虚拟硬盘共享整个分布式存储池的容量及性能。

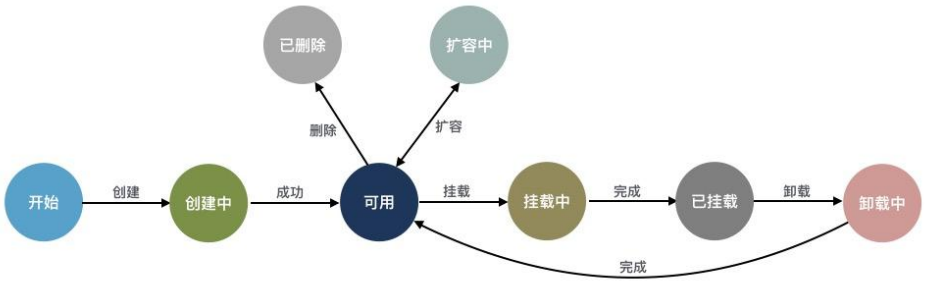
- 支持虚拟硬盘创建、挂载、卸载、磁盘扩容、删除等生命周期管理，单

块虚拟硬盘同时仅能挂载一台虚拟机。

- 支持在线和离线的方式扩容磁盘容量，磁盘扩容后需要在虚拟机的操作系统进行磁盘容量的扩容操作。
- 为保证数据安全性及准确性，虚拟硬盘仅支持磁盘扩容，不支持磁盘缩容。
- 虚拟硬盘最小支持 10G 的容量，步长为 1GB，可自定义控制单块虚拟硬盘的最大容量。
- 虚拟硬盘具有独立的生命周期，可自由绑定至任意虚拟机，解绑后可重新挂载至其它虚拟机；
- X86 架构的虚拟机最多支持绑定 25 块虚拟硬盘，ARM 架构虚拟机最多支持绑定 3 块虚拟硬盘；
- 支持虚拟硬盘克隆，即将虚拟硬盘内的数据复制成为一个新的虚拟硬盘；
- 支持对虚拟硬盘进行快照备份，包括虚拟机的系统盘快照及数据盘快照，并可从快照回滚数据至虚拟硬盘，用于数据恢复和还原场景；
- 支持对全局及每一块虚拟硬盘的 QoS 进行配置，可根据不同业务模式调整磁盘的性能，以平衡平台整体性能；
- 支持从虚拟硬盘创建虚拟机，虚拟硬盘需要有能正常启动的镜像系统。

支持自动精简配置，在创建虚拟硬盘时，仅呈现分配的逻辑虚拟容量。当用户向逻辑存储容量中写入数据时，按照存储容量分配策略从物理空间分配实际容量。如一个用户创建的虚拟硬盘为 1TB 容量，存储系统会为用户分配并呈现 1TB 的逻辑卷，仅当用户在虚拟硬盘中写入数据时，才会真正的分配物理磁盘容量。

高性能型虚拟硬盘的性能与容量的大小成线性关系，容量越大，提供的 IO 性能越高，如果对 IO 性能有强烈需求，可考虑扩容性能型虚拟硬盘。虚拟硬盘完整生命周期包括创建中、可用、挂载中、已挂载、卸载中、扩容中、已删除等资源状态，各状态流转如下图所示：



4.5.3 应用场景

(1) 普通虚拟硬盘（SATA+SSD 缓存）

- 适用于对容量要求较高且数据不被经常访问或 I/O 负载低的应用场景；
- 需要低成本并且有随机读写 I/O 的应用环境，如大型视频、音乐、离线文档存储等；

(2) 高性能虚拟硬盘（SSD/NVME）

- 适用于 I/O 负载高且数据经常被读写的应用场景；
- 中大型关系数据库；
- 中大型开发测试环境；
- 中大型实时响应服务类环境；

4.6 共享盘

共享盘是一种支持多个虚拟机并发读写访问的数据块级存储设备，具备多挂载点、高并发性、高性能、高可靠性等特点。主要应用于需要支持集群、HA（High Available，指高可用集群）能力的关键企业应用场景，多个虚拟机可同时访问一个共享盘。

用户可通过指定共享硬盘的类型、容量及名称即可快速创建一块虚拟硬盘，作为虚拟机的共享数据盘。

支持将虚拟硬盘设置为共享盘，并作为虚拟机的数据盘，使多个虚拟机同时对共享盘进行数据读写操作。同时支持对共享盘进行创建、绑定、解绑、扩容、

克隆及删除等操作。

4.7 快照服务

平台分布式存储支持磁盘快照能力，可降低因误操作、版本升级等导致的数据丢失风险，是平台保证数据安全的一个重要措施。

快照是某一时间点一块虚拟硬盘的数据状态文件，可以理解虚拟硬盘某个时刻的数据备份，虚拟硬盘的数据写入和修改不会对已创建的快照造成影响。

支持定时快照策略，即一个可周期性执行的自动创建快照的策略，快照策略与快照分离，拥有独立的生命周期。在实际应用中，磁盘快照可降低因误操作、版本升级等导致的数据丢失风险，可大致应用于以下业务场景：

- **容灾备份：**定时为虚拟硬盘制作快照，当系统出现问题时，可快速回退，避免数据丢失。
- **版本回退：**在业务做重大升级时，建议预先做好快照，当升级版本出现系统问题无法修复时，可通过快照恢复到历史版本。

用户可为某块虚拟硬盘创建快照，同时支持对虚拟机系统盘进行快照备份。为保证数据及磁盘的安全：

- 仅支持对未绑定及已绑定的硬盘进行快照操作，若硬盘在扩容或快照中，无法进行快照备份；
- 创建快照时，不可进行磁盘挂载/卸载及修改虚拟机状态（如开机或关机），否则可能会导致快照创建异常；
- 快照仅捕获已写入硬盘的数据，不包含应用程序或操作系统缓存在内存中的数据，建议在暂停对硬盘的 I/O 操作后进行快照制作，如关机或卸载硬盘。

平台支持对已绑定虚拟机的系统盘及数据盘进行快照操作，并支持快照回滚操作，即将虚拟硬盘回滚到快照时刻的数据状态，以满足数据恢复的应用场景。同时支持通过快照创建虚拟硬盘。

(1) 回滚快照

将虚拟硬盘回滚到快照时刻的数据状态，应对快照数据恢复的应用场景。回滚时虚拟硬盘必须处于未绑定或绑定的虚拟机必须处于关机状态，仅支持正常状态的快照进行回滚操作。

(2) 从快照创建虚拟硬盘

创建的硬盘大小与快照的原始硬盘大小相等，继承加密属性；从快照创建虚拟硬盘，该虚拟硬盘只能与快照所对应的原始虚拟硬盘归属同一存储集群，可以用系统盘快照创建的虚拟硬盘创建虚拟机。

(3) 快照删除

平台采用 Copy-on-Write（COW，写时复制）快照技术。系统通过元数据记录每个快照引用的数据块，元数据包含了关于每个数据块在哪个快照中被引用的信息。多个快照可以引用相同的数据块，系统不需要在存储上复制数据块的多个副本，从而有效减少了存储空间占用。当用户删除某个快照时，系统会检查该快照引用的数据块，并且只会删除仅有该快照引用的数据，多个快照引用的数据块则不会被删除，以保障其它任意快照的完整性。

5 运维运营管理

5.1 统一管理服务

平台为底层异构计算、存储、网络等基础设施提供统一管理服务，支持将x86、ARM、龙芯、申威、GPU、SSD 存储、HDD 存储等多种架构物理资源统一抽象为计算和存储集群资源，为上层客户提供资源服务能力。

平台以不同的集群划分不同配置、不同架构、不同用途的服务节点，可在一个数据中心下部署多个不同类型的计算集群或存储集群，并通过一套平台统一进行资源分发配置和管理，打平不同架构资源的管理面。

用户在平台上进行不同架构资源的部署时，只需选择适合的集群即可创建出相应架构的资源，如选择 GPU 集群，则可以创建 GPU 虚拟机，并可对所有资源进行统一的生命周期管理。

同时针对不同架构的资源，平台通过资源的统一抽象，统一提供资源接入、账号认证、资源管理、监控告警、日志事件、统计分析及权限控制等，从下而上形成一个统一的整体，满足数据中心统一管理与运维的需求。

为方便整个平台资源的统一运维和运营，平台在账号认证体系上提供多账号管理、多地域管理、全局资源视图、物理资源管理、虚拟资源管理、QoS 配置、资源模板、标签管理、监控告警、通知组、操作日志、资源事件、回收站、报表统计、大屏监控等服务，全面覆盖平台运营运维的使用场景。

5.2 平台管理账号

5.2.1 管理员概述

平台除内置的系统管理员账号外，支持创建多个管理员账号，根据地域的授权范围分为系统管理员和地域管理员。

角色等级	默认	数据范围	功能范围

系统级	系统管理员	系统+地域	地域管理、集群管理、物理资源管理、虚拟资源管理、网络服务管理、账号管理、运维与管理、监控大屏、自定义 UI 管理、全局配置管理
	系统只读管理员	系统+地域	地域查看、集群查看、物理资源查看、虚拟资源查看、网络服务查看、账号查看、运维与查看、监控大屏、自定义 UI 查看、全局配置查看
地域级	地域管理员	地域	地域管理、集群管理、物理资源管理、虚拟资源管理、网络服务管理、账号管理-我的账号、运维与管理、监控大屏、全局配置-产品策略管理
	地域只读管理员	地域	地域查看、集群查看、物理资源查看、虚拟资源查看、网络服务查看、账号管理-我的账号查看、运维查看、监控大屏、全局配置-产品策略查看

系统管理员拥有平台所有管理权限，用于全局管理和运营整个平台。可通过系统管理员账号管理平台的地域、集群、管理员、资源、安全及平台全局配置。

地域管理员拥有平台特定地域下的管理权限。可通过地域管理员账号管理平台特定地域下的集群、资源及平台配置。

平台账号体系使用邮箱地址作为平台的登录账号，在使用前需确保提供的账号邮箱地址可用，方便接收告警邮件或找回密码等。

5.2.2 管理员账号安全

针对管理员账号自身的安全，平台提供修改登录密码、找回密码、双因子验证登录、登录访问限制、修改登录邮箱等安全防护功能。

- **修改登录密码：**支持在使用管理员账号时，更改管理员密码。
- **找回密码：**平台支持管理员账号在忘记密码时通过控制台自主找回密码，找回密码时需通过邮箱进行验证，需确保管理员添加的账号为真实可用的邮箱。

- **双因子验证**：平台为管理员账号提供免费的基于 TOTP（Time-Based One-Time Password Algorithm）登录二次认证服务，开通本服务后，管理员每次登录控制台均需通过授权认证。支持国密硬件版和普通软件版，用户可根据需要通过部署进行配置。

- **登录访问限制**：为保证账号登录的安全及对特定安全场景的需求，平台提供账号登录访问限制能力，为管理员账号设置登录控制台和访问 API 的客户端 IP 地址，配置后管理员账号只能从指定的 IP 登录或发起 API 访问，保证管理员登录及资源管理的安全性。

  - 支持配置多个 IP 地址或 IP 地址段，多个 IP 地址/段间使用英文逗号进行分隔。

  - 配置的 IP 地址或 IP 地址段为白名单模式，即配置的 IP 地址/段客户端才可正常登录控制台或访问 API。

  - 默认不指定任何 IP，代表不限制登录控制台和访问 API 的客户端 IP 地址，即默认全网可访问登录控制台。

- **修改登录邮箱**：平台支持管理员通过控制台右上角管理员账号头像中的【修改登录邮箱】来进行修改管理员的账号邮箱地址，用于将管理员账号修改为真实可用且实际需要接收告警邮件的邮箱地址。

同时平台支持获取系统管理员和地域管理员的 API 密钥信息，用于通过 API 接口管理平台全局资源和获取运行信息。

5.2.3 管理员账号管理

(1) 自定义创建管理员

平台管理员可创建系统管理员和地域管理员，地域管理员不可创建管理员，创建管理员时需要添加一个账号邮箱作为管理员的账户。

可同步设置账号首次登陆强制修改密码，同时支持对账户设置管理类别、是否只读及开通地域。

(2) 冻结/解冻管理员

冻结管理员是指将一个管理员进行锁定，被成功冻结的管理员将无法登录平台行。仅支持状态为【使用中】的管理员进行冻结操作。

当管理员被冻结后，管理员的状态为冻结中，支持具有管理权限的系统管理员解冻管理员，管理员解冻后，可正常登录控制台。

(3) 地域授权管理

地域授权管理可设定一个地域管理员在地域下的授权情况。只有在授权地域下，地域管理员才可在该地域正常使用服务。企业可以根据平台实际运营情况配置管理员在对应地域的开通情况。

(4) 删除管理员

支持具有管理权限的系统管理员删除管理员账号，管理员账号被删除后，无法再登录平台。

5.2.4 账号权限管理

平台支持通过角色对账号进行权限管理，以实现平台细粒度的权限控制。

角色是一组权限的集合，通过对资源权限进行集合配置管理，可支持的资源包括虚拟机、镜像、虚拟硬盘、快照、安全组、资源模板、监控告警、扁平网络等资源，并支持对资源的增、删、改、查及相关管理 API 的操作权限进行开启和关闭，以满足多场景用户权限管理及控制场景。

平台提供内置角色，根据地域性或全局性以及只读或管理维度，默认提供如下四种角色：

- 系统管理员：包含了平台所有资源操作和管理授权，属于平台最高权限角色。
- 系统只读用户：对平台所有资源有查看权限，通常为平台运维管理人员赋予该角色。
- 地域管理员：相比系统管理员角色，无账号管理等全局性权限。为账号

赋予角色的同时可选定部分地域，以实现账号对指定地域的所有资源有操作和管理授权。

- 地域只读用户：对指定地域的所有资源有查看授权，通常为地域运维管理人员赋予该角色。

内置角色是平台根据应用场景提供的默认角色，可进行快速权限配置，若内置角色无法满足场景需求，支持用户自定义创建角色，并对角色进行修改、删除管理，同时支持查看一个角色已绑定的账号。

创建管理员账号时，支持设定授权范围，【系统管理员】表示授权所有地域，【地域管理员】只对选定的地域进行授权。同时通过所选角色对账号进行功能授权，即账号只拥有角色中定义的权限集合。平台通过授权范围及角色控制账号的权限以及资源范围，以实现平台细粒度的权限控制。

> **说明** 角色授权时，支持多个角色一起授权，最终账号获得的权限为多个角色的权限的并集，即所有角色权限的集合。

5.3 多地域

5.3.1 多地域概述

地域（Region）是平台中的一个逻辑概念，指资源部署的物理位置分类，可对应机柜、机房或数据中心。一个数据中心内平台支持部署多个计算和存储集群；数据中心之间资源和网络完全物理隔离，可通过一套管理平台管理遍布各地数据中心的超融合平台。

- 地域在平台也称为数据中心，通常数据中心之间完全隔离以保证最大程度的稳定性和容错性。
- 平台默认内置一个地域，管理服务通过本地数据中心平台提供的 API 端点管理地域内计算、存储及网络资源。
- 支持对数据中心内资源的生命周期管理，包括计算集群、存储集群、基础镜像及自制镜像等资源的查看和维护。

多地域管理是指在一个组织或企业内部建立和管理多个数据中心，并将多个数据中心部署在不同的地理位置，提供更高的可用性、容错性和灵活性，使组织能够更好地满足地理分布和业务需求。

多地域管理支持在多个数据中心部署超融合平台，多套平台通过统一管理界面进行运维运营管理，用户可在统一控制台使用多个地域下的资源。

在部署结构上可以将一个机柜、模块或机房作为一个地域或数据中心，部署一套超融合平台，可应用于边缘节点构建、双活数据中心或两地三中心等场景。

多套超融合平台均通过上层统一平台进行统一管理，多套平台共享一套账号认证、监控告警、审计日志、API 网关、操作界面及相关通用套件，每个数据中心的平台资源均部署于自己的地域中，通过地域标识保证资源的隔离和安全性。

5.3.2 多地域特性

多地域管理通过一套控制台对多个地域的多套超融合平台进行统一管理和调度，具有多地域部署、统一管理、安全隔离及平滑扩展等特性。

- **多地域部署**

多地域管理允许将数据中心分布在不同的地理位置上，可以是不同的城市、国家或洲际，用于减少地理位置带来的单点故障风险，提高系统的可靠性和可用性。

- **统一管理**

为方便管理和监控多个地域的数据中心，平台提供统一管理服务，支持对各个地域集中管理，包括资源配置、监控和故障排除等能力。用户可在平台一键切换地域，并进行资源创建和管理；可通过管理控制台统一管理整个平台的多地域架构。

- **安全与隔离**

多地域管理，提供安全和隔离机制，以确保各个地域之间的数据和资源不被未授权的访问。具备网络隔离、访问控制手段来实现，确保地域之间的独立

性和安全性。

- **平滑扩展**

多地域管理具备平滑的扩展能力，以应对业务的变化和需求的增长。当需要增加新的地域或调整已有地域的规模时，多地域管理机制不影响现有业务，扩展和调整不会对整个架构产生影响。

5.3.3 多地域管理

平台支持资源多地域统一管理，通过地域切换可进行不同地域的管理，在不同的产品服务页面，切换地域后将会仅展示和操作当前地域的资源。

平台支持管理当前已部署的所有地域，包括地域信息的查看、编辑地域信息、查看地域资源用量统计及地域资源管理等。

- **地域信息获取**：支持查看平台管理的所有地域列表，以及查看地域下 CPU 核数用量、内存用量、存储用量、物理 GPU 用量。
- **地域资源用量统计**：支持查看地域的资源使用情况，按照已分配数量和总量生成扇形图，支持查看多地域中资源使用率最高的前五个地域排行展示。
- **地域资源管理**：提供全平台所有地域物理资源的生命周期管理和运维能力，使平台管理员可通过控制台统一管控地域中整体物理资源，包括物理机资源（节点）、计算集群、存储集群及网络资源。

5.4 全局资源视图

平台提供全局资源管理视图，针对多地域资源使用情况提供总体视图展示，支持查看单地域和全部地域资源的使用情况，方便平台管理者统计所有物理资源及虚拟资源的使用状况，如使用率、使用量及资源分配数量等。

- 支持平台所有地域或单个地域的资源用量统计，如 vCPU、内存及存储资源的资源总量、已分配量及未分配量。
- 支持平台所有地域或单个地域的 CPU、内存、存储资源的使用率排行前五统计。

5.5 节点管理

节点管理是指对地域内的所有物理节点的管理，包括查看节点信息、锁定、解锁、进入维护模式、退出维护模式及修改告警模板等，同时可支持对每个计算节点中已存在的计算实例、USB 设备、网络设备、磁盘设备及集群信息进行查看和管理。

(1) 节点信息

支持查看节点的名称、IP 地址、CPU 信息（如 CPU 型号、CPU 核心、CPU 槽数、CPU 线程）、CPU 总量、内存总量、序列号、状态、架构、节点类型、地域及 NTP 信息。

同时支持查看节点所属的集群信息，如集群 ID、集群类型、CPU 使用率、内存用量、GPU 用量、地域等。

(2) 计算实例

支持管理员通过节点详情页面，查看节点中的计算实例列表及信息，包括名称、计算实例 ID、资源 ID、状态、节点 IP、镜像 ID、GPU、CPU、内存、创

建时间及更新时间。

(3) 监控告警

支持查看节点机器的监控信息，包括：网卡入带宽、网卡出带宽、硬盘读吞吐、硬盘写吞吐、平均负载、内存使用率、空间使用率、硬盘读次数、硬盘写次数、网卡入包量、网卡出包量、CPU 使用率、TCP 连接数、阻塞进程数。

支持管理员根据节点监控指标修改相应的告警模板，用于监控节点的健康状态，并支持多种方式的告警通知。

(4) 锁定节点

节点被锁定后，新建计算实例不会被调度至该节点，不影响节点内已有计算实例，可配合节点进入维护模式功能，以实现节点维护、升级等操作。

(5) 解锁节点

管理员将锁定的节点进行解锁，可对外提供计算服务，计算实例可被调度并部署至节点。

(6) 进入维护模式

当需要维护节点时，比如扩展内存、升级、修复硬件等维护场景下，平台支持将节点进入维护模式，使节点上的虚拟资源自动迁移至同计算集群中其他物理节点上，使节点处于空闲状态，确保对物理节点维护时不影响平台的虚拟资源运行，保证业务的可用性。节点进入维护模式前必须保证节点状态为已锁定。

(7) 退出维护模式

退出维护模式是指将节点重新加入至调度系统，为平台提供计算能力。仅支持状态为【维护模式】的节点退出维护模式。退出维护模式，将会自动恢复并进入至锁定状态，需进行解锁才可加入智能调度系统以提供计算能力。

(8) USB 设备管理

支持管理员通过节点页面，查看节点中 USB 设备列表及信息，包括设备名、设备 ID、状态、厂商、类型、序列号、虚拟机、USB 版本。

5.6 虚拟资源管理

平台为管理员提供虚拟资源全生命周期运营和管理能力，使平台管理员可通过控制台统一管控平台的虚拟资源。

虚拟资源管理包括所有产品服务，如虚拟机、资源模板、隔离组、安全组、虚拟硬盘、快照、监控告警、资源事件等资源。

5.7 QoS 配置管理

平台默认提供全局虚拟硬盘 QoS 配置，即新创建的虚拟硬盘会根据平台公式赋予 QoS 值，限制平台用户对磁盘性能强行占用。同时支持管理员对平台所有虚拟硬盘自定义设置 QoS 值，仅当全局 QoS 配置开启时，管理员为每个虚拟硬盘自定义的 QoS 才可生效。

虚拟硬盘创建出来后，管理员可在虚拟硬盘列表上针对每一块虚拟硬盘进行“QoS 配置”，同时可对虚拟机详情磁盘中的系统盘进行 QoS 配置。

- **读/写 IOPS**

  当磁盘的 Arch 架构为 HDD 时，可设置的读/写 IOPS 范围为 0~50000，默认值为 1000，配置为 0 不限速。

  当磁盘的 Arch 架构为 SDD 时，可设置的读/写 IOPS 范围为 0~50000，默认值为计算公式根据当前硬盘容量计算的值，配置为 0 不限速。

- **读写带宽（MBps）**

  当磁盘的 Arch 架构为 HDD 时，可设置的读/写带宽范围为 0~1000Mbps，默认为 100，配置为 0 则不限速。

  当磁盘的 Arch 架构为 SSD 时，可设置的读/写带宽范围为 0~1000Mbps，默认为计算公式根据前当前硬盘容量计算的值，配置为 0 则不限速。

硬盘扩容容量后，会根据计算公式重新计算新容量的 QoS 值，根据计算的新值重新设置硬盘的 QoS。

- 若硬盘扩容前设置的 QoS 值 < 新容量 QoS 值，则以新容量 QoS 值为准。
- 若硬盘扩容前设置的 QoS 值 > 新容量 QoS 值，则以扩容前设置的值为准。

硬件介质和容量会影响硬盘的读写 IOPS 和宽带速率，若配置的 QoS 超过硬件本身性能，以硬件性能为准。系统会默认分配 QoS 值，如需取消一块硬盘的限速功能，可将 IOPS 和宽带均配置为 0。

5.8 资源模板

资源模版支持用户预定义创建资源的参数配置，保存到模版中，便于后续快速创建。

平台用户可以通过指定机型、规格、镜像、虚拟硬盘、网络、安全组等虚拟机相关基础信息一键创建虚拟机模板，用于从模板创建虚拟机实例。

- 虚拟机模板仅作为资源创建的模板配置，不占用实际资源。
- 支持通过资源模板一键创建资源，创建时可进行配置变更。
- 支持用户更新资源模板的配置项和标签，并支持对资源模板进行克隆。
- 支持用户删除资源模板，删除后对通过资源模板创建的资源无影响。

5.9 标签管理

5.9.1 概述

标签用于标记各项虚拟资源，从不同维度对具有相同特征的资源进行分类、搜索和聚合，让资源管理变得更加方便。标签由一对键值对（key:value）构成，用户可根据需求自定义键值对内容，绑定不同资源。

- 支持标签批量创建，单次创建，删除标签功能。
- 支持查看资源，展示该条标签下所有绑定的资源。

- 支持绑定资源，可选择不同地域下不同资源类型进行绑定。
- 支持解绑资源，可批量解绑。

同时，支持资源创建时选择需要的标签进行添加，支持在资源界面对标签进行添加与删除操作。资源界面将会展示当前资源所绑定的标签键值对。

支持统一的搜索入口，可根据 key/value，资源 ID，资源类型，三个维度进行绑定资源的查询，灵活操作资源，可在资源界面以及标签管理界面进行搜索，方便查询管理较大数量的标签，以及快速的匹配资源。

5.9.2 资源类型

支持的资源类型包括：虚拟机、虚拟机模版、镜像、扁平网络、虚拟硬盘、快照、安全组、隔离组。

5.9.3 使用限制

- **标签命名限制**

标签键以及 value 值支持最大 127 位字符，不能为空，区分大小写-标签 key 以及 value 内容支持 utf-8 格式表示的大小写数字、汉字、数字、空格以及特殊字符

- **数量规范**

1 个资源最多可以绑定 50 个标签-1 个标签包含 1 个标签键和 1 个标签值（tagKey:tagValue）-1 个资源上的同一个标签键只能对应 1 个标签值-单次批量创建标签数量最多不超过 5 个

- **资源状态限制**

虚拟机除过删除，删除中和失败的资源不能更新标签，其他状态下可修改资源绑定的标签内容。

5.10 监控告警

5.10.1 概述

监控告警是平台全线产品的运维监控及告警服务，提供全线资源实时监控数据及图表信息，可批量为资源设置告警策略，并在资源故障或监控指标超过告警阈值时，以邮件的方式给予通知及预警；同时监控告警服务实时为用户提供资源告警状态，让用户精准掌控业务和平台产品的健康状况，全方位保障业务的可靠性和安全性。

监控告警服务提供监控图表、告警模板、告警记录及通知组四大架构功能，整体架构功能均以监控数据为基准：

- 平台通过智能化数据采集系统，对虚拟机、虚拟硬盘等资源指定的监控指标数据进行完整挖掘；同时可对节点、计算集群、存储集群指定的监控指标数据进行完整挖掘。
- 平台将采集来的监控数据存储至数据库中，并根据指定规则对数据进行检索及统计，通过指定的时间维度及数据粒度以图形化的方式显示监控图表。
- 基于已有的监控数据，用户可通过配置告警模板，为指定的监控指标指定告警阈值、持续时间、重要程度、通知组及对比方式，可通过设置告警持续时间，判定区分不同等级的告警及通知。
- 可为告警模板配置通知组，指定在发生告警时通知事件的通知人及通知方式。
- 在告警期间，可通过告警记录查询实时告警信息，以判断故障的发生时间和重要程度。

5.10.2 监控图表

监控图表指平台将智能化采集的资源运行数据，根据指定的资源及指标等筛选规则进行检索并统计，通过指定的数据粒度及时间维度以图形化的方式显示监

控图表。通过监控图表，用户可以直观的查看并了解平台上已运行虚拟资源的性能、容量及网络状态等状态，及时了解资源的健康状况。

平台为虚拟机提供多种监控指标的实时和历史监控图表，并可根据监控指标项配置相关告警模板，用于阈值超标时给予告警及通知。同时平台为节点、计算集群、存储集群提供多种监控指标的实时监控图表，并可按照用户的需求对告警模版进行配置，用于阈值超标的时给予告警和通知。

- **虚拟机监控图表**：通过虚拟机详情页面的监控信息栏可查看单台虚拟机的监控信息，包括 CPU 使用率、CPU 平均负载、内存使用率、磁盘空间使用率、磁盘读/写吞吐、网卡出/入带宽、网卡出/入包量、磁盘读/写次数、TCP 连接数、阻塞进程数；
- **节点监控图表**：通过节点详情页面的监控信息可查看，包括 CPU 使用率、CPU 平均负载、内存使用率、磁盘空间使用率、磁盘读/写吞吐、网卡出/入带宽、网卡出/入包量、磁盘读/写次数、TCP 连接数、阻塞进程数。
- **计算集群监控图表**：通过计算集群列表页面的监控信息可查看，包括 CPU 使用率、内存用量、GPU 用量。
- **存储集群监控图表**：通过存储集群列表页面的监控信息可查看，包括存储用量。

监控图表可根据时间维度展示实时监控数据，同时支持查看 1 小时及自定义时间的监控数据及图表信息。

5.10.3 告警模板

告警模板是平台监控告警服务为用户提供的一种批量设置资源告警的功能，通过预先定义模板中的告警规则及通知规则，将模板中定义的规则应用到虚拟资源；若虚拟资源的监控指标数据达到或超过告警规则中设定的阈值及条件，则根据通知规则中定义的通知方式发送告警通知到指定的联系人。

根据不同的资源类型，可定制不同监控指标及阈值的告警规则，满足多种应

用场景下的监控报警需求。

- 告警模板是由多条告警规则及关联资源构成的；
- 一个告警模板仅支持绑定一种类型资源，包括虚拟机、节点、计算集群及存储集群等。
- 每个告警模板可包含多条告警规则，每条告警规则包含监控指标、对比方式、告警阈值、持续时间、重要程度及通知组；

平台支持管理员为资源创建告警模板，并提供告警模板的全生命周期管理，用户可通过告警模板自定义所对应资源类型的告警规则，通过告警规则对资源进行监控指标的告警触发和处理。

告警规则是告警模板的核心，每个告警模板均由 1 条或多条告警规则组成。被绑定至告警模板的资源监控指标数据会根据告警规则中定义的阈值触发相关告警策略，并通过告警规则中的通知方式进行告警信息的通知，以便快速入处理告警或故障。

- **监控指标：**仅可选择告警模板资源类型所包含的监控指标，一条告警规则仅支持一个监控指标：
- **对比方式：**指监控指标的实际数据与告警阈值的比较方式，代表当前告警规则的告警逻辑，包括>=、<=、>、<：
  - 选择>=时，即代表监控数据大于或等于阈值时触发一次告警周期；
  - 选择<=时，即代表监控数据小于或等于阈值时触发一次告警周期；
  - 选择>时，即代表监控数据大于阈值时触发一次告警周期；
  - 选择<时，即代表监控数据小于阈值时触发一次告警周期；
- **告警阈值：**指监控指标数据的临界值，与监控指标数据进行对比，符合对比方式即触发一次告警周期，如 CPU 使用率的告警阈值为 80，对比方式为大于等于，即 CPU 使用率大于等于 80%即触发一次告警周期；
- **持续时间：**监控指标数据触发阈值持续的时间，持续时间内均达到告警

阈值才会触发告警；

- **重要程度**：用户可根据业务需要在创建告警规则时选择合适的等级，分为一般、重要、危险三种，在告警记录中可根据重要程度进行记录的筛选；
- **通知组**：即触发告警周期且需要发送通知时，发送告警通知的方式及联系人。

5.10.4 告警记录

通过告警记录可查阅实时及历史告警信息，包括告警的指标说明、模版类型、当前值、状态、重要程度及告警时间。

- **指标说明**：触发当前告警记录的资源监控指标项，即数据来源；
- **模版类型**：触发当前告警记录的资源类型及资源；
- **当前值**：即触发告警或恢复告警时当前告警记录监控指标的数据值；
- **状态**：告警记录的当前状态，分为触发中、待触发、未触发，可根据需求进行状态的筛选；
- **重要程度**：根据监控规则显示当前告警记录的重要程度，包含危险、重要、一般，可根据需求进行告警记录的筛选；
- **告警时间**：触发告警规则的具体时间。

5.11 通知组

通知组是监控报警发送告警通知的方式及联系人信息，通过对用户邮箱、Webhook 地址的记录，将不同资源告警通过邮件或 Webhook 方式通知给通知人，以便划分全责，精细化处理告警通知。

在使用监控告警模板时，需要先创建一个通知组，添加相关联系人信息，并设置通知组的通知方式，以便关联告警模板。

通知组是一组通知人的组合，可以包含一个或多个联系人，在资源发生告警

时会通过所设置的通知方式至所有通知人。

通知人是指告警规则发送通知的具体联系人，同一个联系人，可以加入多个通知组，支持邮件通知和 Webhook 通知两种方式。

- **邮件通知：**支持配置联系人名称及邮箱地址信息，用于发送告警通知至配置的联系人邮箱。
- **Webhook 通知：**支持配置 Webhook 地址及发送警告信息的请求方式，请求方法支持 GET 和 POST 两种方式进行信息传输。

5.12 操作日志

5.12.1 操作日志

操作日志是指用户在控制台或 API 对资源进行的操作行为及登录登出平台的审计信息。操作日志会记录用户在平台中的所有资源操作，提供操作记录查询及筛选，通过操作日志可实现安全分析、资源变更追踪以及合规性审计。

通过操作日志用户可查看平台全部地域所有的资源操作及平台审计日志。

操作日志支持的资源模块包括虚拟机、USB、资源模板、镜像、虚拟硬盘、安全组、监控告警模板、定时器、账户等。

平台支持获取日志的操作名称、所属模块、地域、关联资源、操作者、操作结果、备注及操作时间等信息；并支持导出用户的操作审计日志为本地 Excel 表格，方便账户管理和运营。

- **操作名称：**指操作日志的操作名称，包括调用 API 的接口名称及操作的界面展示名称，如调整带宽。
- **所属模块：**指操作日志操作的资源类型，如虚拟机类型。
- **地域：**操作资源所属的地域。
- **关联资源：**操作日志对应的资源标识符，并可查看一个操作中所有关联的资源标识。

- **操作者：**操作日志对应的操作者。
- **操作结果：**操作日志的结果，如操作成功、操作失败、参数异常、存储集群物理资源不足等。
- **备注：**操作日志的备注信息。
- **操作时间：**操作日志的操作时间。

为方便用户便捷的查看操作审计日志，支持日志的筛选和搜索检索，包括所属模块、操作状态及查询时间范围等纬度。

所属模块支持所有产品模块的筛选，同时支持查看全部资源的日志及审计信息，即不对所属模块进行筛选。操作状态支持状态为成功、失败的日志筛选，同时支持查看全部状态的日志和审计信息。查询时间范围支持 1 小时及自定义时间的日志筛选，最长可查询半年的操作日志。

5.12.2 通知规则

操作日志通知规则对资源操作日志进行监控并通过通知组进行事件信息的告知。支持用户为操作日志配置通知规则，当资源操作日志符合通知规则要求时，即发送监控邮件到通知组内的成员。

支持用户配置操作日志通知规则的监控地域、通知组、监控模块及监控级别等信息，并支持对操作日志的通知规则进行修改和删除。

- **监控地域：**通知规则的地域信息。
- **通知组：**邮件通知的通知组信息，仅支持选择一个通知组。
- **监控模块：**监控的资源模块内容，如虚拟机、虚拟硬盘等，支持批量配置多个模块。
- **监控级别：**操作日志的操作结果，包括操作成功、操作失败。

5.13 资源事件

5.13.1 资源事件

资源事件是用于对平台核心资源的部分操作及状态进行记录及通知，如资源生命周期状态的变化、操作运维执行情况等。

资源事件记录用户在资源类型的部分核心操作事件，提供事件详细记录查询及筛选，并可配合通知规则及时通知用户、定位问题。

资源事件支持的的资源类型包括虚拟机。支持获取资源事件的资源 ID、资源类型、事件类型、事件等级、事件内容、事件发生次数、开始时间及更新时间。

- **资源 ID**：指资源事件监控的资源 ID。
- **资源类型**：当前资源事件记录所指定的资源类型。
- **事件类型**：分为生命周期变化事件和操作运维事件，如虚拟机调度，虚拟机开关机，挂载磁盘等。
- **事件等级**：事件等级的类型，包括正常、警告、错误。
- **事件内容**：详细记录触发事件的具体信息。
- **发生次数**：记录该事件累计触发次数。
- **开始时间**：第一次资源事件发现的时间。
- **更新时间**：第二次及以后触发资源事件的时间。

资源事件对整个平台及所有用户资源事件进行记录，为方便用户便捷的查看资源事件日志，支持事件日志的筛选和搜索检索，包括所属地域、资源类型、资源及事件周期等纬度；同时平台资源事件日志支持用户导出资源事件为本地 Excel 表格，方便用户查看和定位。

5.13.2 通知规则

资源事件通知规则对事件日志进行监控并通过通知组进行事件信息的告知。

支持用户为资源事件配置通知规则，当事件日志符合通知规则要求时，即发送监控邮件到通知组内的成员。

支持用户配置资源事件通知规则的监控地域、通知组、监控模块及监控级别等信息，并支持对资源事件的通知规则进行修改和删除。

- **监控地域：**通知规则的地域信息。
- **通知组：**邮件通知的通知组信息，仅支持选择一个通知组。
- **监控模块：**监控的资源模块内容，如虚拟机、集群等。
- **监控级别：**对实例正常运行的影响程度进行划分，含正常、警告、错误。

5.14 回收站

5.14.1 概述

平台提供资源回收站，是平台资源删除的暂时保留区，用户删除的资源包括虚拟机、磁盘、自制镜像等资源，会在删除后自动进入回收站中。

平台资源被用户手动删除时，会自动进入回收站暂时留存。保留期间用户可在回收站中查看资源的信息，如资源 ID、资源名称、资源状态、资源类型、删除时间等。

- **资源状态：**当前资源的状态，包括已删除、销毁中；
- **资源类型：**当前留存资源的资源类型，包括虚拟机、硬盘、自制镜像等；
- **删除时间：**指当前留存资源被手动删除进入回收站的时间。

在回收站中的资源支持恢复及销毁操作，销毁会将资源彻底删除，不可恢复。

为方便用户对回收站资源的维护，支持对进入回收站的资源进行批量操作，包括批量恢复资源及批量销毁资源。

5.14.2 恢复资源

恢复资源是指手动恢复被误删而进入回收站的资源，若资源被用户手动删

除，可直接通过恢复资源操作进行恢复。

5.14.3 销毁资源

销毁资源是指用户手动销毁留存在回收站的资源，资源被销毁后无法恢复。支持用户批量对回收站中的资源进行批量销毁，以方便对账户的清理和维护。

5.15 巡检服务

一键巡检，是平台提供的用来检查平台健康情况的特性能力。通过对平台管理节点、计算节点的巡检项扫描，检查平台节点 CPU、内存、磁盘等资源的使用情况，使管理员更方便地对问题进行评估。

巡检主要是对平台进行全面扫描，包括管理节点的时间源同步检查、CPU 使用率、内存使用率检查、磁盘使用率检查；计算节点的物理机 CPU 平均使用率检查、物理机内存使用率检查、物理机系统盘已用容量检查等，一键巡检内容如下：

巡检类型	巡检项	巡检项含义	结果展示	巡检建议
管理节点	时间源一致性检查	检查是否设置时间源同步	提供节点当前时间源，和推荐时间源	若检测到时间源与集群内其他节点时间源不一致或物理机系统时钟未与时间源同步，请 SSH 登录对应系统，检查时间源配置
	CPU 使用率检查	检查平台管理节点 CPU 的使用占比	提供当前占比，若超过 80%，提供最高使用率的五个进程	若检测到平台 CPU 使用率在 10 分钟内持续的超过 80%的使用率，请尽快联系平台相关人员进行热升级或问题评估，以继续正常使用本平台功能
	内存使用率检查	检查平台管理节点内存的使用占比	提供当前占比，若超过 80%，提供最高使用率的五个进程	若检测到平台内存使用率在 10 分钟内持续的超过 80%的使用率，请尽快联系平台相关人员进行热升级或问题评估，以继续正常使用本平台功能
	磁盘容量检查	检查平台管理节点磁盘的使用占比	提供当前占比，若超过 70%，提供占比最高的十个文件路径和文件大小	若检测到平台磁盘数据容量已占用管理节点所在磁盘超过 70%的容量，请尽快联系平台相关人员进行磁盘检查和扩容，以继续正常使用本平台功能
	管理服务检查	检查平台管理服务的运行情况	提供当前服务名称及状态，若服务异常，提供	若检测到服务状态异常，请根据提供的节点信息，SSH 登录对应系统，检查服务的状态

			异常的节点 IP	
计算节点	物理机 CPU 平均使用率检查	检查平台上物理机 CPU 平均使用率	提供当前占比，若超过 80%，提供最高使用率的五个进程	若检测到物理机 CPU 平均使用率超过 70%，请登录物理机系统，确认物理机上是否存在异常进程。若未存在异常进程，建议考虑对集群进行扩容
	物理机内存使用率检查	检查平台上物理机内存平均使用率	提供当前占比，若超过 80%，提供最高使用率的五个进程	若检测到物理机内存使用率超过 80%甚至 90%，请立即登录物理机系统，检查物理机上是否存在业务异常，并按需优化运行业务。必要时，建议对集群进行扩容
	物理机系统盘已用容量检查	检查平台上物理机系统盘使用率和使用量	提供当前占比，若超过 70%，提供占比最高的十个文件路径和文件大小	若检测到物理机系统盘容量使用率超过 70%甚至 90%，请立即登录至物理机系统，检查并清理对业务无影响的数据

支持用户下载巡检报告，通过浏览器将巡检报告下载到本地。通过巡检报告可查看详细的报告内容，如地域、节点、名称、巡检项、巡检结果、现状、分数、建议以及最高使用率。

当巡检结果异常时，展示当前设备参数现状及针对性的建议，并展示导致结果异常的主要文件名称及大小，使管理人员及时了解物理机状态并介入处理。同时针对已完成巡检任务的巡检报告，平台支持管理员删除巡检报告。

5.16 报表统计

报表统计是平台用于汇总和分析平台内各种资源数据的机制，包括资源用量统计及资源统计表，将各种数据整理成易于理解和分析的形式，提高平台整体运营和管理的效率。

5.16.1 资源用量统计

资源用量是平台聚合全平台资源监控，根据多维度查询和指标分析，展示资源使用情况，支持管理员创建报告并导出 Excel 表格。

支持创建资源用量报告，从资源用量周期、地域、资源类型多个维度对控制台资源进行统计。

支持虚拟机、计算集群和存储集群三种资源类型的自定义时间周期资源用量报告，支持范围为 1 小时 ~ 12 个月的用量，可选择 1 天/3 天/7 天/14 天/30 天/自定义，自定义可将开始时间和结束时间精确到小时。

- **计算集群资源用量报表信息**

  地域、资源类型、资源 ID、资源名称、架构、CPU 超分比例、CPU 总量、CPU 分配量、CPU 分配率、内存总量、内存分配量、内存分配率及统计时间。

- **存储集群资源用量报表信息：**

  地域、资源类型、资源 ID、资源名称、集群架构、总量、已分配量、已分配率、已使用量、已使用率及统计时间。

- **虚拟机资源用量报表信息：**

  地域、资源 ID、资源名称、资源类型、状态、CPU 规格、内存规格、GPU 规格、CPU 平均使用率、CPU 最大使用率、内存平均使用率、内存最大使用率、磁盘分区使用情况及统计时间。

当资源类型为虚拟机时，可查看虚拟机 CPU 使用率分布和虚拟机内存使用率分布，以使用率为横轴，虚拟机数量为纵轴进行统计。

支持资源用量统计的自动创建策略，通过制定保留数量、重复周期、执行时间、资源用量周期。

- **保留数量：**需要保留的数量，超出此数量的最旧的将被删除。
- **重复周期：**支持单次、每天、每周、每月、间隔多种模式，单次执行默认当天执行，若执行时间已过则为次日执行；间隔支持按分种或和小时进行间隔执行。
- **资源用量周期：**生成时间范围内的资源用量报告。

支持查看平台所有创建的资源用量报表，并支持管理员删除资源用量报告。方便运营数据的统计，平台支持导出资源用量报告为本地 Excel 表格文件。

5.16.2 资源统计表

资源统计表是呈现平台各类资源清单的离线 Excel 表格，用户可以统一收集各资源的基本信息，用于报表分析和自定义的数据处理。

平台支持虚拟机、虚拟硬盘、共享硬盘、快照、扁平网络、安全组、节点计算实例、节点硬件设备、管理员账号、操作日志、资源事件、资源模板等信息报表统计。

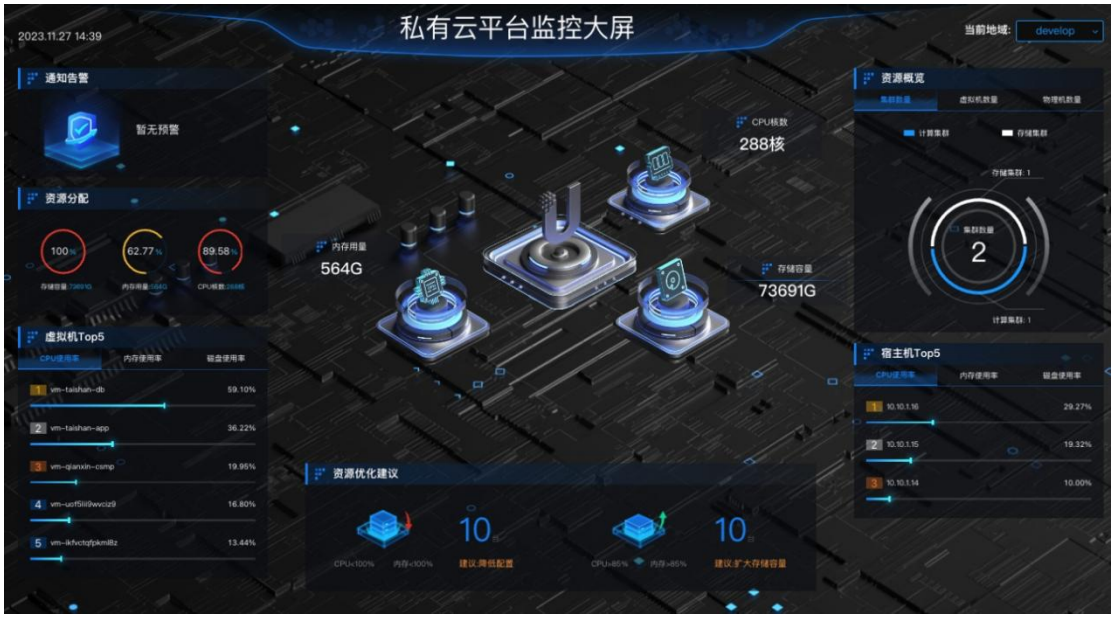
- **虚拟机：**名称、资源描述、资源 ID、状态、宿主机 IP、扁平网络 ID、扁平网络名称、集群与特性、镜像、GPU、CPU、内存、系统盘总容量、数据盘总容量、CPU 使用率、内存使用率、磁盘使用率、IP、标签、高可用级别、创建时间。
- **虚拟硬盘：**名称、资源描述、资源 ID、状态、是否加密、集群架构、集群、硬盘类型、硬盘容量、绑定资源类型、绑定资源、标签、创建时间。
- **共享硬盘：**名称、资源描述、资源 ID、状态、是否加密、集群架构、集群、硬盘类型、硬盘容量、绑定资源类型、绑定资源、标签、创建时间。
- **快照：**名称、资源描述、资源 ID、状态、是否加密、硬盘 ID、硬盘类型、标签、快照来源、创建时间。
- **扁平网络：**名称、资源 ID、状态、IP 版本、网段、IP 范围、 网卡、VLAN、标签、创建时间、更新时间。
- **安全组：**名称、资源描述、资源 ID、状态、规则数量、绑定资源数量、标签、创建时间。
- **节点计算实例：**名称、计算实例 ID、资源 ID、状态、节点 IP、镜像 ID、GPU、CPU、内存(MB)、创建时间、更新时间。
- **节点硬件设备：**设备名、设备 ID、设备类型、状态、厂商、产品、序列号、BUS、Device、虚拟机、创建时间、更新时间。
- **管理员账号：**账号名称、账号 ID、账号邮箱、状态、管理类别、创建时

间。

- **操作日志：**操作(API)名称、所属模块、地域、关联资源、操作者、操作结果、操作时间。
- **资源事件：**资源 ID、资源类型、地域、事件类型、事件等级、事件内容、事件发生次数、开始时间、更新时间。
- **资源模板：**名称、资源描述、资源 ID、状态、资源类型、标签、创建时间、更新时间。

5.17 大屏监控

监控大屏是平台为企业提供的平台资源可视化大屏，主要展示平台宏观维度的监控数据，帮助企业平台运营者快速了解平台的整体运行情况，支持自定义拖拽模块及全屏展示，方便管理员进行分屏管理。



- **通知和告警：**展示最近 5 条平台告警信息。
- **物理机 TOP5：**展示 CPU 使用率、硬盘读吞吐、硬盘写吞吐、内存使用率在前 5 名的物理节点 IP 。
- **虚拟机 TOP5：**展示 CPU 使用率、硬盘读吞吐、硬盘写吞吐、内存使用率在前 5 名的虚拟机 ID 。

- **资源分配：**展示平台 CPU、内存、存储的总容量以及已分配容量的百分比。
- **资源概览：**展示物理机总量以及状态分布（可用、锁定）、计算集群和存储集群的数量分布、虚拟机总量以及状态分布（运营、关机、其他）。

6 平台管理

6.1 客制化能力

平台为客户提供平台客制化能力，支持管理员自定义平台 UI 展示样式，包括网站基本设置、登录页设置，如修改平台 logo 图片、登录页图片及标题等。

6.1.1 定义网站展示

网站设置是平台为企业和管理员提供的客制化能力，包括网站 Favicon 图片、网站 Title、平台 Logo 图片，即自定义平台的 Logo 及浏览器标志等。同时支持设置是否展示帮助文档、收藏夹及默认语言等配置。

- **网站 Favicon 图片**：浏览器标签页上展示的 Favicon 图片，必须为 ico 格式，最大不超过 100KB，推荐尺寸 48px*48px。
- **网站 Title**：浏览器标签页上展示的网站说明，如中立安全可信赖的云计算服务商，支持中英文及特殊字符。
- **平台 Logo 图片**：管理员控制台导航栏上方的 logo，允许管理员自定义，图片支持 png、jpeg、jpg 格式，最大不超过 200KB，推荐尺寸 352px*72px。
- **帮助文档**：允许管理员开启或关闭控制台上的帮助文档的展示。
- **收藏夹**：允许管理员开启或关闭控制上的收藏夹功能。
- **默认语言**：允许管理员开启或半闭英文控制台。
- **资源失败状态原因**：允许管理员控制资源操作失败时是否在状态旁边展示失败原因提示。
- **告警单位默认 MBytes**：允许管理员设置监控告警单位默认使用 MBytes，而不是 Bytes。

6.1.2 监控大屏标题

支持自定义设定监控大屏标题，以满足个性化和特定需求。

6.1.3 定义登录页

登录页设置是平台为企业和管理员提供的登录客制化能力，包括登录页标题、登录页标题颜色、背景图、登录页面 Logo、登录页联系电话、登录页输入框位置、登录页面版权、登录页描述信息、登录面输入框背景透明等。

- **登录页标题**：代表登录框上的标题描述，如超融合，可支持中英文及特殊字符。
- **登录页标题颜色**：代表登录框上标题的颜色，支持白色、黑色、红色、黄色、绿色、青色、蓝色、棕色、紫色、橙色、灰色及金色，以适应不同背景图片上文字的可读性。
- **背景图**：代表登录框后的背景图片，图片支持 png、jpeg、jpg 格式，最大不超过 500KB。
- **登录页面 Logo**：图片支持 png、jpeg、jpg 格式，最大不超过 200KB，推荐尺寸 352px*72px
- **登录页联系电话**：登录页的联系电话。
- **登录页输入框位置**：支持居中、居左和居右。
- **登录页面版权**：支持自定义登录页面的版本说明信息。
- **登录页描述信息**：登录面的描述信息，登录页账户输入区域位置设置为居中时不展示该信息。
- **登录面输入框背景透明**：可设置输入框为透明色，如不透明，背景色为白色。

6.2 平台系统配置

6.2.1 邮箱配置

邮箱设置是指平台邮件服务的配置，主要功能是找回密码、监控告警邮件的接收和发送。平台支持管理员定义邮箱的是否支持 SSL、发件人邮箱地址、发件人邮箱密码、邮件服务器 IP、邮箱服务器 Port 及邮件主题前缀。

- **邮箱主题前缀：**配置平台发送的邮件主题前缀，如超融合平台等。
- **发件人邮箱地址：**配置发件人的邮箱地址。
- **发件人邮箱密码：**配置发件人邮箱密码。
- **邮箱服务器地址：**设置发件邮箱服务器的 IP 地址或域名。
- **邮箱服务器端口：**设置发件邮箱服务器的端口，默认值为 994，范围支持 0-65535。
- **邮箱支持 SSL：**配置邮箱是否支持 SSL。

平台部署时默认必须提供邮箱设置，避免无法接收找回密码及监控告警邮件。邮箱配置完成后，支持用户对邮箱是否配置正确进行测试。

6.2.2 磁盘设置

6.2.2.1 全局磁盘 QoS

硬盘管理支持管理员对平台全局虚拟硬盘开启或关闭 QoS 控制，以保证平台所有虚拟硬盘资源的性能可靠性。

(1) 平台默认全局开启磁盘 QoS，即代表平台全局硬盘 QoS 生效，包括新建硬盘和已有硬盘的 QoS。

- 硬盘默认创建出来会根据平台计算公式赋予 QoS 值。
- 已有硬盘的 QoS 根据已赋予的默认值或管理员修改的值生效。

- 配置为开启时，管理员为每个硬盘自定义的 QoS 才可生效。

(2) 配置为关闭时，平台全局硬盘 QoS 失效，包括新建硬盘和已有硬盘的 QoS。

- 新创建的硬盘 QoS 不受限制。
- 已有硬盘的 QoS 不受限制。
- 配置为关闭时，管理员为每个硬盘自定义的 QoS 不会生效。

(3) 硬盘扩容容量后，会根据计算公式重新计算新容量的 QoS 值，根据计算的 QoS 值重新设置硬盘的 QoS。

- 若硬盘扩容前设置的 QoS 值<新容量 QoS 值，则以新容量 QoS 值为准。
- 若硬盘扩容前设置的 QoS 值>新容量 QoS 值，则以扩容前设置值为准。

6.2.2.2 全局磁盘设置

平台针对磁盘支持配置共享盘绑定虚拟机数量，并支持自定义设置虚拟硬盘可手动创建的快照数量。

- **共享盘绑定虚拟机数量：**支持限制共享盘可同时绑定的虚拟机数量，范围为 2~30 个。
- **单个硬盘快照数量：**设置单个硬盘手动快照的数量上限，默认为 10 个，支持的范围为 0~200 个，0 代表单个硬盘创建快照的数量不受限制，管理员可根据平台实际使用情况调整上限数值。

6.2.3 回收策略

平台支持管理员对全局资源开启或关闭删除能力，若开启则允许资源被手动删除。若关闭资源删除能力，平台所有资源将不允许手动删除。

6.3 平台数据备份

平台数据备份服务是针对平台自身的数据库及配置文件进行备份，保证平台

本身的数据安全性。支持平台数据库和配置文件备份。

- **平台数据库备份：**平台部署后会自动生成一条自动备份策略，默认每天备份一次。
- **平台配置文件备份：**平台部署后会自动生成一条自动备份策略，默认每天备份一次。

6.4 自定义规格

规格配置是平台为企业和管理员提供的自定义规格能力，管理人员可通过自定义规格配置调整平台上架产品服务的规格类型，包括虚拟机、硬盘等。

- 虚拟机支持定义 CPU 和内存规格；
- 硬盘支持定义可创建虚拟硬盘的容量范围；

平台针对虚拟机、硬盘会默认提供建议型的规格，管理员可根据企业需求对规格进行变更，包括查看、删除、修改等。

(1) 创建规格

平台支持创建虚拟机 CPU/内存的规格，虚拟硬盘规格由平台默认生成，仅支持修改。创建虚拟机规格支持根据不同的集群创建不同的规格，即可为不同的集群创建不同的规格，适应不同集群硬件配置不一致的应用场景。

(2) 修改规格

可根据业务需求在不同的集群中创建不同的规格，同时支持管理员对已创建和默认生成的规格值进行修改。

(3) 删除规格

支持管理员删除指定自定义虚拟机规格，不支持删除硬盘的规格。规格删除后平台用户即不可在所属集群中创建当前规格的虚拟机，但不影响通过该规格创建资源的正常运行。

虚拟机规格在每个集群内会生成一条无法删除的默认规格，以避免平台上所

有规格均被删除，导致无法创建虚拟机。

6.5 统一授权

统一授权支持客户按需对基础服务模块和增值服务模块分开授权，支持用户按照架构区分授权节点，根据业务需求选择各模块授权的生效时间和失效时间，平台通过授权证书激活保证了密钥不可克隆验证的唯一性。

平台为用户提供完整的授权管理能力，包括授权管理和节点管理两大模块。

6.5.1 授权管理

管理员可通过【信息采集】下载当前平台硬件及服务授权需求信息，并将信息发送给平台运营平台，运营平台确定用户需求生成授权证书，并由平台管理员将证书上传至平台。

通过授权管理用户可查看平台基础许可、拓展许可和服务许可，并可详细了解产品授权的状态、生效时间、失效时间和数量限制等信息。

- **基础许可**：支持基础设施管理系统套件-标准版、基础设施管理系统套件-信创版和分布式块存储套件的授权；
- **拓展许可**：支持 USB 透传服务、GPU 服务和异构平台迁移软件服务的授权。
- **服务许可**：支持基础实施管理系统+增值服务 7*24 维保服务和金牌 VIP 维保服务。

6.5.2 节点管理

用户可通过节点管理查看平台所有节点授权状态及节点信息情况等。如节点名称、序列号、状态、CPU 型号、CPU 总量、内存总量和架构等。

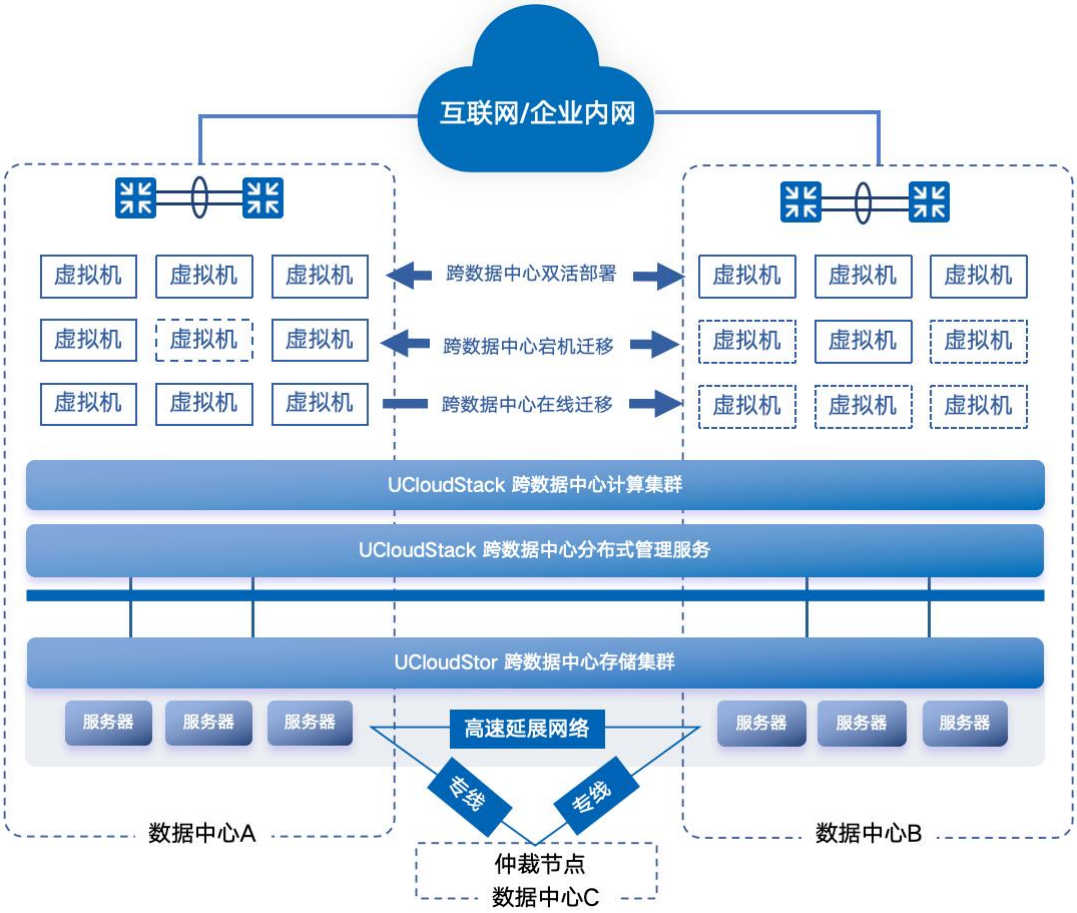
同时用户可通过节点管理查看授权节点的基本信息、CPU 信息和 Memory 信息。

7 双活数据中心

7.1 概述

出于数据隐私和安全性考量，可通过“同城双活”及“两地三中心”的高可用架构保障生产环境稳定性和业务过程连续性。同时超融合在企业数字化转型中可提供更加快速灵活的 IT 资源交付和管控，支撑业务创新和变革。

平台具备多数据中心部署和统一管理能力，帮助用户降低双活数据中心的建设门槛，快速实现业务跨数据中心的故障转移保障机制，从而进一步提升并保障业务连续性。



双活数据中心方案采用两个相互独立、互为备份的数据中心，将两个数据中心资源通过一套平台的相同集群进行统一管理。

数据会在两个数据中心之间实时同步，当一个数据中心出现故障或宕机，则会自动实现故障切换，避免业务系统因单点故障而中断，确保业务任何情况都能

保持稳定运行。同时双活数据中心具备高度的可扩展性，可根据客户的需求进行自定义配置和扩容，满足不同业务场景下的需求。

通过平台提供的双活数据中心能力，对数据安全和业务连续性保障进行全面梳理和支撑，提升系统可靠性和连续性，助力企业在数字化转型中创造优质价值。

7.2 部署结构

两个数据中心的节点间网络可为二层或三层网络通信，需将两个数据中心间通过专线打通内网，保证两个数据中心间网络性能，避免影响分布式存储或虚拟机存储服务的性能。

双活数据中心的存储节点共同构建为一套存储集群，共享一套分布式存储系统，将存储管理服务分别部署于两个数据中心和一个仲裁数据中心，避免数据中心脑裂，保证数据中心存储服务的可用性。

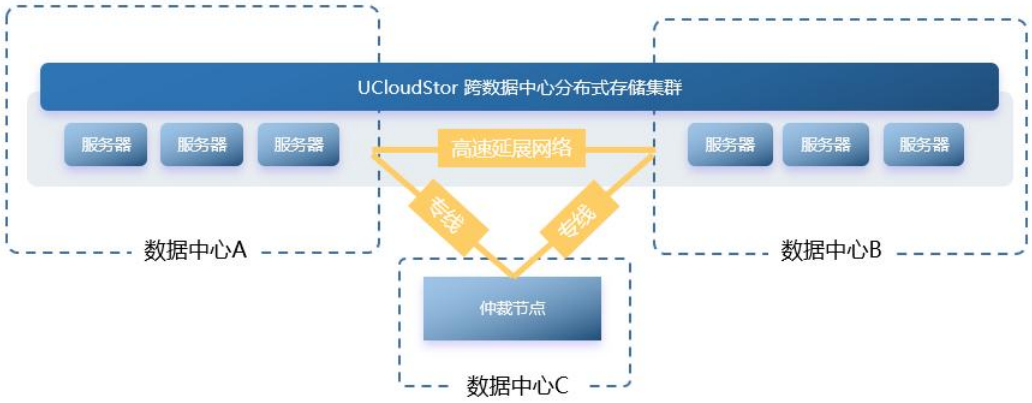
分布式存储数据冗余机制通过 4 副本的方式分别将 2 副本存放于 2 个数据中心，采用数据同步、数据复制、多级故障域及故障自恢复等技术，实现数据在不同数据中心间的强一致同步，提高数据的可用性和可靠性。

平台管理服务分别部署于两个数据中心和一个仲裁数据中心，采用分布式系统架构，保障平台调度和管理服务的健壮性和可用性，使平台在多个数据中心健康运行。

7.3 双活机制

(1) 跨数据中心强一致性数据保障机制，防止单点故障和数据丢失

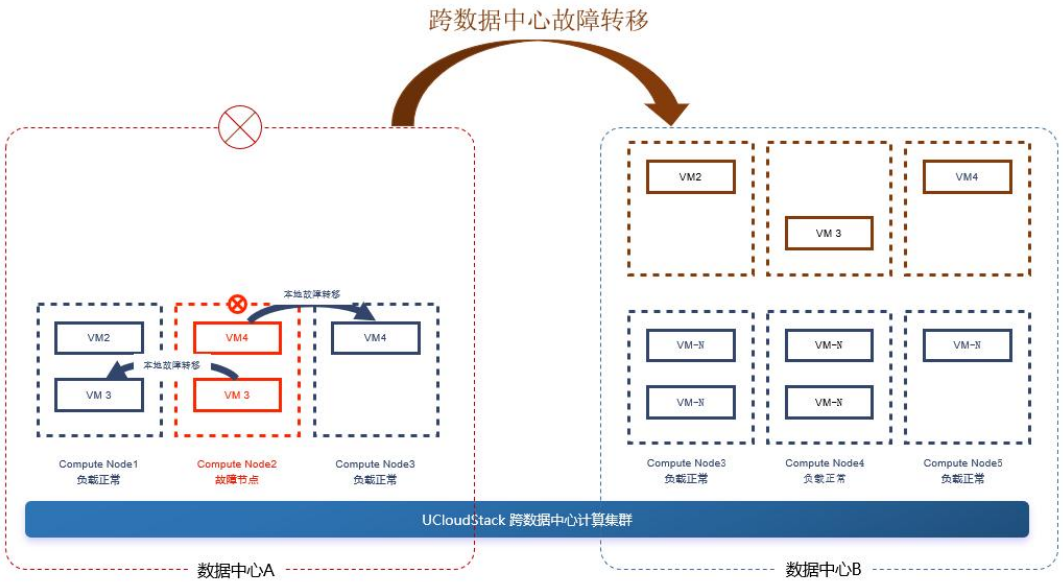
平台内置分布式存储，作为双活数据中心解决方案的核心存储系统。依托跨数据中心的存储双活能力，可实现跨数据中心强一致性数据保障，是双活数据中心方案的核心技术之一。



通过将数据存储在多个数据中心中，并采用数据同步和数据复制等技术，实现数据在不同数据中心之间的双向同步和备份，提高数据的可用性和可靠性，防止数据中心单点故障和数据丢失。在数据中心之间的网络质量符合方案要求的前提下，可以实现 RPO=0，RTO≈0，保证数据零丢失。

(2) 跨数据中心故障转移机制，有效降低故障恢复时间

平台默认提供本地故障转移调度策略和机制，当物理服务器发生宕机或故障时，实现在同数据中心内进行故障转移，即：将故障物理机的虚拟机，向另一台有空闲资源的物理机上迁移并启动，从而大幅降低系统的故障恢复时间，RPO=0，RTO<5min。

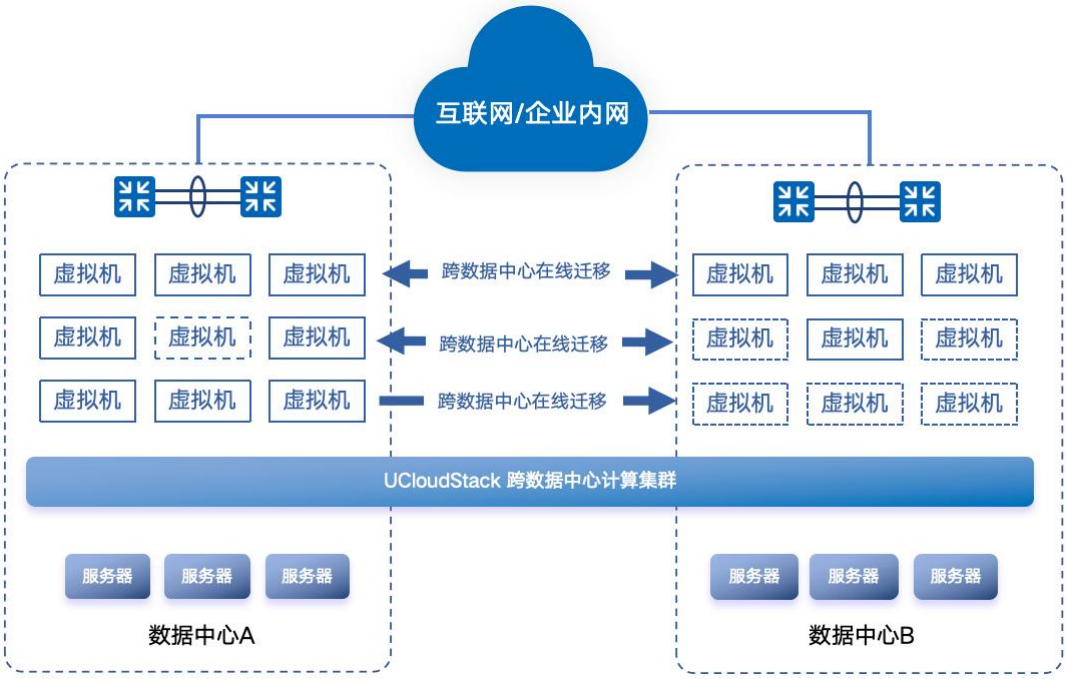


跨数据中心的故障转移机制，在优先采用本地调度策略的基础上，增加多数

据中心调度属性，当数据中心出现极端的故障时，在对应用不做任何改造的条件下，将实例迁移至健康数据中心的物理机，实现业务系统跨数据中心的容灾恢复，RPO=0，RTO<5min。

(3) 跨数据中心在线迁移机制，多数据中心资源平衡

在线迁移是计划内的虚拟机热迁移操作，即：虚拟机内的业务应用保持着持续对外服务的同时，虚拟机在不同的物理机之间进行在线跨物理机迁移，业务应用近似无感知。



跨数据中心在线迁移机制，即提供多数据中心迁移能力，使在线迁移不受限于同一数据中心。跨数据中心在线迁移机制，可以有效的进行多数据中心之间的资源平衡，以及计划内的跨数据中心热迁移。

(4) 跨数据中心分布式管理服务机制，保障系统健康运行

跨数据中心分布式管理是一种基于分布式系统架构的管理服务，用于支撑平台自身健壮性的一组管理服务，同时可保障平台在多个数据中心健康运行。该服务机制支持跨数据中心运行，通过将管理功能和资源分布到不同的数据中心，以实现跨数据中心的分布式管理和协作。

- **管理服务自愈能力**

基于分布式系统的建设原理，通过智能化和自动化的管理策略，可以自动监控和维护多个数据中心内的健康状况，减少人工干预和管理成本。在面临数据中心级别故障或异常情况时，管理服务可自动检测、定位、诊断和修复，从而保证平台的稳定性和可靠性。

- **可视化监控**

提供全面的可视化监控和报告功能，帮助管理员了解平台的状态和性能，及时发现和解决问题。

- **统一管理接口**

提供统一的管理接口和管理策略，方便管理员对整个系统进行集中化的管理和协作。

7.4 双活收益

(1) 降低双活数据中心建设门槛

传统的双活数据中心建设是一项较为复杂的集成类项目，项目周期长、涉及的软硬件产品多、运维成本高、建设效果参差不齐。

平台将双活数据中心建设所需的基础能力标准化和产品化，客户在建设过程中无需集成第三方产品，采用平台标准的建设步骤即可快速完成双活数据中心的建设。同时配合平台轻量化特性，有效降低双活数据中心的建设成本。

(2) 进一步提升应用的业务连续性

通过双活数据中心建设，可在不改造业务系统的情况下，实现业务系统的跨数据中心故障转移机制，即跨数据中心宕机迁移和跨数据中心在线迁移，从而再次提升系统的业务连续性。

(3) 为跨数据中心双活应用的建设夯实基础

通过双活数据中心建设，即完成必要的双活数据中心基础建设，并内置一定的故障转移机制，为双活应用的改造提供充分准备。

用户仅需为应用增加跨数据中心的访问流量调度机制和业务系统本身的跨数据中心高可用，即可完成跨数据中心的双活应用改造。

7.5 方案场景

双活数据中心解决方案，可帮助客户进一步提升业务系统的可靠性，同时保障数据安全和隐私性，通过降低建设门槛和复杂性，提升建设效率，赋能客户加快数字化转型的进程。适用场景如下：

- 高可用性要求较高的应用场景，如金融、电商、物流等行业，在业务高峰期和数据中心故障时，能够保证系统的稳定运行。
- 考虑建设灾备数据中心的客户，可在数据中心出现故障时，快速将应用切换到另一个数据中心，保证业务的连续性和数据安全性。
- 建设双活数据中心需要付出高额成本的客户，平台双活数据中心可实现异地备份和故障转移，降低运营成本；同时支持在多个数据中心之间实现负载均衡和资源共享，有效提高资源利用率和运营效率。
- 对于数据安全和合规性的有较高要求的客户，平台双活数据中心，提供数据备份和异地容灾能力，保证客户业务的数据的安全性和完整性，同时基于平台的管理机制及安全等保，全面满足监管和合规要求。

平台全面的双活数据中心能力，已在多个行业客户案例中得到验证。如某专注于智能营销云的大型集团，需针对数据爬虫、网站及大数据分析业务资源进行多数据中心部署和有效运营。

为保障客户业务的连续性和安全性，平台提供双活数据中心能力，通过跨数据中心的数据强一致性保障、故障转移及跨数据中心统一管理等机制，为客户提供跨城双活灾备方案，使客户通过一套平台统一管理并调度多数据中心资源，在提升业务连续性和数据可靠性的同时，大幅降低运营成本、提高资源利用效率。

8 平台安全性

平台提供多维度且全面的安全保障体系，包括控制台安全、账号认证授权、网络安全控制、数据存储安全及日志审计等体系，并结合信息安全等级保护三级保证平台和业务的安全性。

8.1 控制台安全性

平台的管理控制台通过多次登录失败冻结、密码登录有效期、禁止多点登录、无操作自动退出、密码过期强制修改、自定义密码复杂度及密码不符合规则时强制修改等多方面保证平台的安全性和排它性。

- **多次登录失败冻结：**支持管理员为平台开启或关闭全局账号登录多次失败冻结功能，开启后账号登录输入密码错误超过次数将被冻结。同时支持设置多次登录失败锁定时长，支持设置 15~1440 分钟。
- **密码登录有效期：**支持管理员为平台全局账号默认开启密码到期强制修改策略，保证账号安全。默认设置为 90 天，可修改天数，每 90 天账号登录控制台时会强制要求修改账户密码。
- **密码过期强制修改：**支持管理员为平台全局设置密码过期强制修改，开启后平台所有账号密码过期后必须强制修改；若关闭则密码过期后不强制要求修改。
- **禁止多点登录：**支持管理员为平台全局账号开启多点登录，以适应对平台账号不同的登录需求。关闭时平台账号支持多个客户端同时登录，即一个账号在不同的客户端均可同时登录并管理平台资源。开启时代表平台账号仅支持单点登录，即一个账号同一时间仅支持在一个客户端进行登录，在其它客户端进行登录时将会自动退出已登录的客户端连接。
- **无操作自动退出时长：**支持管理员为平台全局设置空闲时长，保证控制台资源和数据的安全。默认值为 30 分钟，即代表控制台在 30 分钟内无任何操作即会自动退出，支持设置 15~1440 分钟。

- **自定义密码复杂度：**支持管理员对平台账号和虚拟机密码的长度进行配置，支持自定义配置为 6-30 位；同时支持自定义密码复杂度，如密码须包含有大写字母、小写字母、数字、特殊符号(除空格)中的两种或以上,不能包含[A-Z],[a-z],[0-9]和[()`~!@#$%^&*-+=_|{}[]:;’ <>,.?/]之外的非法字符。

- **密码不符合规则强制修改：**支持管理员配置用户账号密码不符合长度及复杂度规则时强制修改。

8.2 账号认证授权

平台提供管理员账号用于整体平台资源管理及配置。基于账号认证安全，平台提供账号认证安全、角色权限授权、账号冻结、API 签名等保障体系。

- **账号认证安全：**平台账号支持登录密码修改、登录邮箱修改、找回密码、双因子验证、数字证书、国密认证、API 密钥鉴权及登录访问限制等多种方式安全防护保障，保证支平台 API 访问入口及控制台的认证安全。

- **角色权限授权：**支持通过角色授权为不同账号授权不同的产品功能，达到不同人员以不同权限访问平台资源的效果。

- **账号冻结：**支持平台对管理员账号进行冻结，当账号被冻结时，无法进行登录和管理操作，保证平台账号的安全性。

- **API 签名：**平台向用户开放产品服务资源操作 API 接口，并提供 API 接口访问密钥对（公钥和私钥），在用户调用 API 接口时，支持将公钥作为参数包含在每一个请求中发送，私钥负责生成请求串的签名，保证平台 API 访问入口的安全性。

8.3 网络安全控制

平台通过安全组实现网络流量安全和控制。安全组作为虚拟防火墙，提供东西向出入双方向流量的访问控制规则，支持 TCP、UPD、ICMP、GRE 等协议数据包的过滤和控制，用于限制虚拟资源的东西向网络访问流量。

8.4 数据存储安全

平台数据存储通过数据保护机制、存储安全、存储加密、存储快照、磁盘 QoS 及平台数据备份等安全体系，全面保证底层存储和数据存储的安全性。

- **数据保护机制**：提供多副本数据冗余机制，通过多副本、写入确认机制及副本分布策略等措施，自动屏蔽软硬件故障并自动进行副本数据备份和同步，保证数据安全性和可用性。
- **存储安全**：虚拟机存储在分布式存储系统中的数据，完全打散写入至整个存储集群的所有磁盘中，读取数据经过元数据和其它盘上块文件进行数据整合，保证数据安全。
- **存储加密**：针对虚拟机的虚拟硬盘和数据安全，平台提供虚拟硬盘加密特性，使用 LUKS 加密规范来对磁盘全盘加密，保护用户的数据不被未经授权的访问者获取，甚至在磁盘丢失或被盗的情况下也可以保证数据的机密性。
- **存储快照**：平台分布式存储提供磁盘快照能力，支持对虚拟机系统盘、虚拟硬盘、共享盘进行手动和自动快照，降低因误操作、版本升级等导致的数据丢失风险，是平台保证数据安全的一个重要措施。
- **磁盘 QoS**：平台全局默认提供全局虚拟硬盘 QoS 配置，即新创建的虚拟硬盘会根据平台公式赋予 QoS 值，限制平台用户对磁盘性能强行占用，保证平台所有虚拟硬盘资源的性能可靠性。
- **平台数据备份**：针对平台自身的数据库及配置文件支持备份特性，保证平台本身的数据安全性。支持平台数据库和配置文件备份。

8.5 日志审计体系

平台提供全面的日志审计能力，包括操作日志及事件管理，实现安全分析、资源变更追踪以及合规性审计。

- **操作日志审计**：平台提供全面操作日志，包括控制台或 API 资源的操作

行为及登录登出审计信息。操作日志会记录用户在平台中的所有资源操作，提供操作记录查询及筛选，通过操作日志可实现安全分析、资源变更追踪以及合规性审计。

- **事件日志审计**：平台提供资源事件审计能力，对平台核心资源的部分操作及状态进行记录及通知，如资源生命周期状态的变化、操作运维执行情况等。资源事件记录用户在资源类型的部分核心操作事件，提供事件详细记录查询及筛选，并可配合通知规则及时通知用户定位问题。

9 平台可靠性

平台可靠性通过数据中心、硬件设施、平台软件、平台服务及平台升级等多维度高可用设计保证平台整体可靠性，进一步保证用户业务连续性。

9.1 数据中心

(1) 双活数据中心

平台具备多数据中心部署和统一管理能力，采用两个相互独立、互为备份的数据中心，将两个数据中心资源通过一套平台的相同集群进行统一管理。

数据会在两个数据中心之间实时同步，当一个数据中心出现故障或宕机，则会自动实现故障切换，避免业务系统因单点故障而中断，确保业务任何情况都能保持稳定运行。同时双活数据中心具备高度的可扩展性，可根据客户的需求进行自定义配置和扩容，满足不同业务场景下的需求

通过平台提供的双活数据中心能力，快速实现业务跨数据中心的故障转移保障机制，提升系统可靠性和连续性，助力企业在数字化转型中创造优质价值。

(2) 异地容灾

在数据中心维度支持容灾方案，根据不同容灾等级的需求，分别通过专线、SD-WAN、VPN、互联网等方式互联多个异地数据中心。将业务和数据按需求部署或异步复制到异地数据中心，并通过智能 DNS 进行异地数据中心的业务切换。

(3) 机柜级冗余

网络设备和服务器硬件设备均对称部署于机柜，如 3 节点服务器分别对称部署于 3 个机柜，一个机柜一台服务器，单机柜掉电或故障不影响业务正常运行和使用。

9.2 硬件设施

(1) 网络设备高可用

- **网络设备扩展**：网络设备扩展性设计，所有网络设备分为核心和接入两层架构，一套核心可水平扩展几十套接入设备。
- **网络设备冗余**：网络设备冗余性设计 ，所有网络设备均为一组两台堆叠，避免交换机单点故障，实现交换机级别高可用。
- **网络接入冗余**：交换机下联接入冗余性设计，所有服务器双上联交换机的接口均做 LACP 端口聚合，避免单点故障，实现交换机互联高可用。

(2) 服务器高可用

- **接入层冗余**：服务器网络接入冗余性设计，所有服务器节点均做双网卡绑定，避免单点故障，实现服务器网络接入高可用。
- **管理节点冗余**：管理节点冗余性和扩展性设计，多台管理节点分布式部署，并支持横向扩展，避免管理节点单点故障，实现管理服务高可用。
- **计算节点高可用**：平台通过智能调度系统将虚拟机均衡部署于计算节点，可水平扩展计算节点数量。支持虚拟机在线迁移、故障转移、智能均衡部署，实现计算节点高可用及虚拟机高可用。
- **分布式存储节点**：分布式存储冗余性设计，将数据均衡存储于所有磁盘，并通过三副本保证数据安全。数据及副本可放置于不同机柜、不同节点及不同磁盘上，尽可能保证数据安全性；同时分布式存储服务器本身提供冗余性负载设计，节点损坏不影响分布式存储使用且不会丢失数据。

9.3 平台软件

- **分布式调度**：基于分布式服务和远程服务调用为用户提供智能调度模块。智能调度模块实时监测集群和所有服务节点的状态和负载，当某集群扩容、服务器故障、网络故障及配置发生变更时，智能调度模块可根据隔离组自动迁移虚拟资源到健康的服务器节点，保证平台的高可靠性和高可用性。
- **分布式网络**：通过独立虚拟交换机实现的扁平网络，保证多个网络之间

的完全隔离。对于每个创建的扁平网络，平台会在集群内的每个节点上均创建一个独立的虚拟交换机并设定相同的网络配置，虚拟机在集群内多个节点进行迁移后依然可以拥有配置一致的网络连接。

- **资源管理**：通过分布式资源管理模块，负责集群计算、存储、网络等资源的分配及管理，为平台用户提供资源创建、资源调度、资源占用及访问控制，提升整个集群的资源利用率。
- **安全管理**：为用户提供身份认证、授权机制、访问控制等功能。通过API 密钥对和用户名密码等多种方式进行服务间调用及用户身份认证；通过角色权限机制控制用户对资源的访问；通过安全组对资源网络进行访问控制，保证平台的安全性。
- **集群部署**：提供自动化部署集群节点的模块，为运维人员提供集群部署、配置管理、集群管理、集群扩容、在线迁移及服务节点下线等功能，为平台管理者提供自动化部署通道。
- **集群监控**：提供平台集群物理资源和虚拟资源信息收集、监控及告警。通过自动化获取资源的运行状态信息，并将信息指标化展示给用户；同时提供监控告警规则，通过配置告警规则，对集群的状态事件进行监控及报警，并有效存储监控报警历史记录。
- **平台升级**：平台支持在线扩展节点，并支持平滑升级所有服务，保证平台扩容时的业务可用性。

9.4 平台服务

(1) 弹性计算

- **智能调度**：智能调度优先选择低负荷节点进行虚拟资源部署，并提供打散部署、在线迁移、离线迁移及宕机迁移等能力，整体保证平台的可靠性。
- **部署策略**：将多个虚拟机加入隔离组，用于控制虚拟机的分布以保证业务高可用性。可自定义虚拟机与其他虚拟机或宿主机之间的亲和关系。

支持亲和性和反亲和性两种策略类型，可有效提高业务服务的可用性。

- **自制镜像**：自制镜像是由平台用户通过虚拟机自行导出的自有镜像，可用于创建虚拟机，提高平台虚拟机资源的可用性。

(2) 网络服务

- **二层隔离**：平台通过虚拟交换机实现简单高效的扁平网络架构，为虚拟机提供多个全局统一配置的分布式二层网络。
- **安全组**：提供出入双方向流量访问控制规则，定义可访问资源的网络或协议，用于限制虚拟资源的网络访问流量，为平台提供必要的安全保障。

(3) 弹性存储

- **分布式存储**：虚拟硬盘采用大规模分布式存储系统，将整个集群中的存储资源虚拟化后，整合在一起对外提供统一的存储服务。分布式存储系统通过三副本、写入确认机制及副本分布策略等措施，最大限度保障数据安全性和可用性。
- **三副本冗余**：用户通过虚拟机应用程序写入虚拟硬盘的数据，会根据分布式存储系统三副本机制存储三份，并按照副本分布算法，分别存储于不同物理主机的磁盘上。三副本机制存储数据，将自动屏蔽软硬件故障，磁盘损坏和软件故障，导致副本数据丢失，系统自动检测到并自动进行副本数据备份和同步，不会影响业务数据的存储和读写，保证数据安全性和可用性。
- **写入确认机制**：三副本在写入过程中，只有三个写入过程全部被确认，才返回写入完成，确保数据写入的强一致性。
- **数据分布策略**：支持副本数据落盘分布策略，可将三副本数据分布在不同磁盘、不同主机、不同机柜甚至不同机房，避免因单主机及单机柜整体故障造成数据丢失或不可用的故障，保证数据的可用性和安全性。为保证虚拟硬盘数据访问时延，通常建议最多将数据副本保存至不同的机柜，若将数据三副本保存至不同的机房，由于网络延时等原因，可能会

影响虚拟硬盘的 IO 性能。

(4) 日志和监控

- **操作日志**：平台提供资源级别操作日志及平台组件化系统操作日志，记录平台所有操作信息，方便定位故障及相关平台运营信息。
- **审计日志**：提供平台所有登录登出操作审计信息。
- **资源事件**：平台提供核心资源的部分操作及状态进行记录及通知，如资源生命周期状态的变化、操作运维执行情况等。
- **监控告警**：平台全线产品的运维监控及告警服务，提供全线资源实时监控数据及图表信息，可根据监控数据批量为资源设置告警策略，并在资源故障或监控指标超过告警阈值时，以邮件的方式给予通知及预警，全方位保障业务的可靠性和安全性。