天翼云

iStack 云原生一体机

技术白皮书

天翼云科技有限公司

版权声明

前言

ChinaTelecom Cloud（天翼云科技股份有限公司）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。

ChinaTelecom Cloud（天翼云科技股份有限公司）建设的天翼云 4.0，依托中国电信集团公司，采用 2+4+31+X 布局，力图构建符合网络、用户、数据、能源多因素分布特征的分布式云。其中 2 指的是内蒙，贵州两大核心数据中心，4 指的是京津翼、长三角、珠三角、川陕渝四大区域中心呢，31 指的是中国 31 省一省一池，X 指的是县区级“千城百池”。

ECX（智能边缘云，Edge Computing X）是 ChinaTelecom Cloud 旗下边缘计算品牌，在天翼云 4.0 中的 X 位置聚焦于低时延、大流量、本地化业务场景，提供虚拟机、云存储、函数计算、裸金属等云服务和解决方案。

iStack 是源于 ECX 的更为轻量级的私有云解决方案，两者都源自天翼云科技的边缘技术栈 ECStack（Edge Computing Stack）。iStack 更关注于以客户为中心的现场私有化云服务，可提供超融合一体机和纯软件部署的两种交付方式。

客户处于自身的管理和安全行考虑出发，往往会倾向于自建私有云平台，但客户自建私有云平台往往存在下列问题：

- **可控性差**

业界基于开源架构封装的私有云核心组件和服务源自社区，可控性差且可靠性未经验证，平台特性升级受限于社区且需专精运维人员，同时开源框架构繁杂，部署实施环节复杂，实施难度大。

- **投入成本大**

OEM 公有云直接部署的专有云平台，所有服务均需独立的服务器集群，起始部署规模较大且通常限制硬件架构及品牌，部署实施需要投入大量基础设施和人力资源；在运维方面通常需要托管运维，建设成本较高。

- **运维复杂**

自建数据中心及通用虚拟化系统，对于业务构建所需的数据库、缓存、负载均衡等一系列应用，需自己通过虚拟机进行搭建并维护，同时还需考虑服务的集群部署、监控、日志、备份、容灾及可靠性和可用性等。

- **兼容性差**

通用数据中心及虚拟化系统，对国产化硬件、操作系统、中间件的适配及兼容性较差。
通过 iStack 私有云平台，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私有

云及混合云解决方案，为企业简化业务上云过程，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。

目录

产品简介

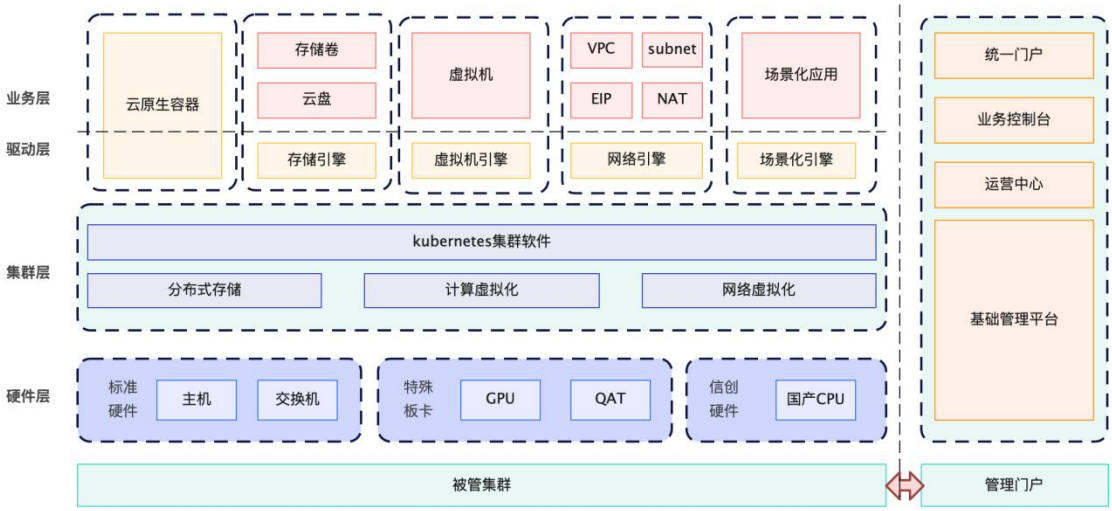
iStack 私有云平台，提供虚拟化、SDN 网络、分布式存储等核心服务的统一管理、资源调度、监控日志及运营运维等一整套云资源管理能力，助力政企数字化转型。

iStack 基于 ECX 公有云基础架构，复用内核及核心虚拟化组件，将公有云架构私有化部署，具有自主可控、稳定可靠、持续进化及开放兼容等特点，可通过控制台或 APIs 快速构建资源及业务，支持与公有云无缝打通，灵活调用公有云能力，帮助政企快速构建安全可靠的业务架构。

iStack 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，不强行绑定硬件及品牌，兼容 X86 和 ARM 架构，并提供统一资源调度和管理，支持纯软件、一体机柜等多种交付模式，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云服务平台。

产品架构

iStack 平台采用被管集群和管理平台分离的架构，这样的架构可实现多集群管理。架构分为两个部分：被管集群和管理门户。



被管集群

被管集群指的是 iStack 所管理的云资源池，包括计算资源，存储资源和网络资源。被管集群的架构分为硬件层、集群层、驱动层、业务层等层次。

硬件层

用于承载 iStack 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，同时支持计算存储超融合节点及对接磁盘阵列设备，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，所有网络功能均通过 SDN 软件定义，仅需物理交换机支持 VIan、Trunk、IPv6、端口聚合、堆叠等特性；
- 支持用于特殊用途的 PCIe 板卡，如 GPU 显卡，QAT 加速设备等；
- 支持主流信创设备；

集群层

用于支撑存储虚拟化、计算虚拟化、网络虚拟化的集群软件。

- Kubernetes 集群软件：提供云原生环境的容器集群调度软件，为 iStack 提供强大的调度和容错能力；
- 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，支持云盘在线扩容、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；
- 计算虚拟化：通过 KVM 、 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上云需求；
- 分布式网络 SDN ：通过 OVS + Geneve 实现虚拟网络，纯软件定义分布式网络，提升网络转发性能的同时对传统数据中心物理网络进行虚拟化，为云平台资源提供 VPC 隔离网络环境、弹性 IP、NAT 网关、负载均衡、安全组及网络拓扑等网络功能，并支持 IPv4&IPv6 双栈；

驱动层

用于支撑特定云服务的 kubernetes 控制器插件。

- 存储引擎：用于在集群中提供分布式存储服务；
- 计算引擎：用于在集群中提供虚拟机管理服务；
- 网络引擎：用于在集群中提供虚拟化网络服务；
- 场景化引擎：用于在集群中提供特定业务场景的服务。

业务层

iStack 对用户提供的核心业务产品（包括但不限于以下产品）。

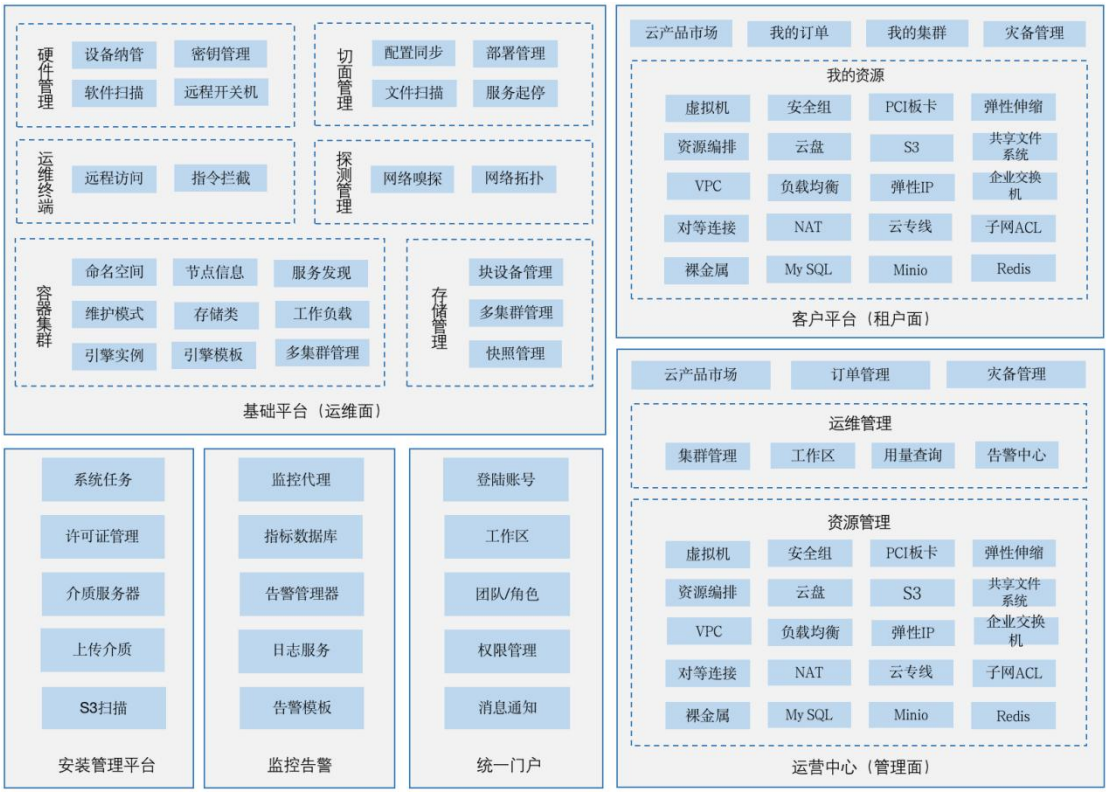
- **边缘虚拟机（EVM）**：运行在物理主机上的虚拟机，支持从镜像创建、重启/关机/启动、删除、console 登陆、VNC 登陆、重装系统、重置密码、绑定弹性 IP 及安全组、挂载数据盘等虚拟机全生命周期功能，同时支持将虚拟机制作为镜像及磁盘快照能力，提供快捷的业务部署及备份能力；
- **GPU 虚拟机（EVM-G）**：平台提供 GPU 设备透传能力，支持用户在平台上创建并运行 GPU 虚拟机，让虚拟机拥有高性能计算和图形处理能力；
- **云硬盘（EVD）**：一种基于分布式存储系统为虚拟机提供持久化存储空间的块设备。具

有独立的生命周期，支持将单独创建的云硬盘绑定给虚拟机使用，或将已绑定的云硬盘解绑，再绑定给其他虚拟机使用；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；

- **VPC 网络**：软件定义虚拟专有网络，用于租户间数据隔离，提供自定义 VPC 网络、子网规划及网络拓扑；
- **弹性 IP**：用于虚拟机、负载均衡、NAT 网关等资源的外网 IP 接入，用于与平台外网络进行连接，如虚拟机访问互联网或访问 IDC 数据中心的物理机网络；
- **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，为云平台提供必要的安全保障；
- **NAT 网关**：企业级 VPC 网关，为云平台资源提供 SNAT 和 DNAT 代理，支持自动和白名单两种网络出口模式，并为 VPC 网络提供端口映射代理服务，使外部网络通过 NAT 网关访问虚拟机；
- **负载均衡**：基于 TCP/UDP/HTTP/HTTPS 协议将网络访问流量在多台虚拟机间自动分配的控制服务，类似于传统物理网络的硬件负载均衡器，用于多台虚拟机间实现流量负载及高可用，提供内外网 4 层和 7 层监听及健康检查服务；
- **裸金属服务器**：裸金属云服务器兼具虚拟机的灵活弹性和物理机高稳定、强劲的计算性能，能与 istack 全产品（如网络、数据库等）无缝融合，在大数据、高性能计算、云游戏等领域都有广泛应用；
- **弹性伸缩组**：弹性伸缩从用户的业务需求和策略出发，自动调整其弹性计算资源的管理服务。可预先配置相关的伸缩策略来保证计算能力，使业务需求上升时自动增加虚拟机实例，业务需求下降时自动减少虚拟机实例，降低人为反复调整资源以应对业务变化和高峰压力的工作量，保障业务平稳健康运行，帮助用户节约资源和人力成本；
- **云备份**：提供一个专注于备份相关业务的服务给用户使用。用户可以在此定义备份空间、备份策略、备份库等。备份中心用于专门处理平台云产品实例的备份，以标准化的备份配置以及备份流程去拓展不同实例的备份。当发生病毒入侵、人为误删除、软硬件故障等事件时，可将数据恢复到任意备份点。

管理门户

iStack 管理门户用于对被管集群进行从底层到业务层的管理，为平台租户、管理员及运营人员提供云平台管理和服务，包括统一门户、客户平台、运营平台、基础管理平台、安装管理平台和监控告警等 6 部分。



统一门户

统一门户为用户提供安全认证和团队管理功能，包括：

- **个人信息管理**：为平台使用 者提供个人信息管理功能，包括：个人信息维护、个人密码维护、个人登录凭据维护，支持密码登录，天翼云官网登录，客户私有 LDAP 登录等；
- **工作区管理**：以工作区为维度划分团队和资源归属，包括：团队管理，角色管理，授权管理等。

客户平台

客户平台提供工作区（工作区即租户）的云资源管理功能，工作区成员根据工作区所有者分配的权限进行云资源的管理，包括：

- **云产品市场**：集中展现了平台中可以开通的云产品，包括但不限于：虚拟机、云硬盘、VPC 网络、NAT 网关、负载均衡、EIP 等；可展现云产品的说明文档，已开通的实例和订单等；
- **我的资源**：集中展现了工作区当前所开通的业务实例，包括但不限于：虚拟机、云硬盘、VPC 网络、NAT 网关、负载均衡、EIP、安全组、私有镜像、特殊板卡等；为每个实例提供个性化详情页，便于使用产品实例；
- **我的订单**：集中展现了工作区的订单以及订单详情，包括订单的处理过程和生成的实例信息；
- **我的集群**：集中展现了管理员为工作区所开通的集群及相应的配额信息；
- **灾备管理**：集中展现了工作区创建的备份数据和数据迁移信息，包括备份空间、备份策略、备份库以及同平台迁移、跨平台迁移、存储复制等信息。

运营平台

运营平台为平台的管理团队提供云资源管理功能，管理团队成员根据超级管理员所分配的权限进行云资源的管理，包括：

- **云产品市场**：集中展现了平台中可以开通的云产品，包括但不限于：虚拟机、云硬盘、VPC 网络、NAT 网关、负载均衡、EIP 等；可展现云产品的说明文档，已开通的实例和订单等；可控制产品的上下线以及控制用户对相关产品的可见范围；
- **资源管理**：集中展现了平台中当前所开通的业务实例，包括但不限于：虚拟机、云硬盘、VPC 网络、NAT 网关、负载均衡、EIP、安全组、私有镜像、特殊板卡等；为每个实例提供个性化详情页，便于使用产品实例；
- **订单管理**：集中展现了平台中的订单以及订单详情，包括订单的处理过程和生成的实例信息；可对订单进行审核；
- **运维管理**：集中展现了关于运营人员可执行的对租户层面以及集群层面的查询及管理权限，包括但不限于：工作区管理、集群管理、用量查询、运维查询、告警查询等相关功能。
- **灾备管理**：集中展现了工作区创建的备份数据和数据迁移信息，包括备份空间、备份策略、备份库以及同平台迁移、跨平台迁移、存储复制等信息。

基础管理平台

基础管理平台为平台的管理团队提供基础资源的管理功能，管理团队成员根据超级管理员所分配的权限进行基础资源的管理，包括：

- **硬件管理**：按照站点/机架/资产的分类架构对硬件资源进行管理，包括主机和交换机的纳管，主机的登录密钥管理和自动维护，对主机上安装软件进行扫描，对主机进行远程开关机（需 ipmi 管理网络），对主机进行自动发现（获取主机型号，制造商，CPU，内存，磁盘，文件系统等信息）等功能。
- **切面管理**：以横向维度对集群软件进行跨主机管理，提供软件的文件扫描，配置同步，一键部署，一键起停等功能；
- **切面管理**：提供对主机进行远程登录功能，并对指令进行拦截和审计；
- **存储管理**：提供对 ceph 分布式存储进行管理，包括块设备池管理，块设备快照及恢复管理，块设备的备份管理等；可进行多存储集群管理，可管理第三方的 ceph 集群；
- **容器集群**：提供对 kubernetes 集群的原生管理功能，包括节点管理，命名空间管理，服务发现及工作负载管理；可进行多容器集群管理，可管理第三方的 kubernetes 容器集群；
- **驱动引擎管理**：提供驱动引擎的安装和卸载功能，可通过驱动引擎管理将第三方的 kubernetes 容器集群改造为 iStack 被管集群；
- **探测管理**：提供一个立体化监控平台，可通过对不同探测点进行网络探测，实现网络质量分析、性能分析等目的。

监控告警

监控告警为 iStack 提供基础的数据采集和告警处理功能，包括：

- **监控代理**：负责采集 iStack 平台所管理的硬件资源（主机/交换机）和集群软件资源的

监控指标；

- **指标数据库**：负责存储 iStack 平台所采集的时序数据，用于告警和前端展现；
- **告警服务器**：根据指标计算，对告警进行智能处理，包括：指标分组，抑制等操作；
- **日志服务器**：负责存储 iStack 平台的日志数据，包括软件的安装日志，系统部署日志，业务开通日志等。

安装管理平台

安装管理平台用于平台初始化部署，作为单独部署的工具为 iStack 提供相关的界面化安装引导服务，帮助用户更快速地安装平台。

- **系统任务**：负责平台安装部署的任务，包括任务进度、任务向导和操作日志；
- **许可证管理**：负责平台许可证的生成和管理；
- **介质管理**：提供界面化的介质管理功能，实现介质服务器的管理以及介质的上传、下载、删除等基本操作。

产品特性

- **自主可控**

基于公有云架构，复用核心虚拟化组件自主研发，可控性高且可靠性经上万家企业验证。

- **稳定可靠**

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。

- **简单易用**

3 节点构建生产环境，规模轻量可水平扩展，支持业务平滑迁移，助力政企轻松上云。

- **开放兼容**

不绑定硬件品牌，兼容 X86 和 ARM 架构及生态适配，设备异构搭建统一管理。

- **云原生**

基于 kubernetes 生态链，在提供传统虚机服务的同时，提供云原生的容器服务。

技术架构特性

分布式

分布式底层系统

iStack 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**：

基于分布式服务调用和远程服务调用为租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，保证云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，为云平台租户提供资源配额、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，为租户提供身份认证、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 VPC 隔离机制和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**：

分布式底层系统为云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；

分布式存储系统

iStack 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、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，SLA 为 99.95% ，保证存储系统的高可用性；
- 支持高性能云硬盘，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，计算节点自带 SSD 磁盘构建为高性能的存储池，计算节点自带的 HDD 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。

分布式网络

采用分布式 Overlay 网络，提供 VPC 、NAT 网关、负载均衡、安全组、弹性 IP 等网络功能。
分布式网络架构将业务数据传输分散至各个计算节点，除业务逻辑等北向流量需要管理服务外，所有虚拟化资源的业务实现等南向流量均分布在计算节点或存储节点上，即平台业务扩展并不受管理节点数量限制。

高可用

iStack 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，保证整个云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；
- 网络服务区域隔离设计，内网业务和外网业务在物理设备上完全隔离，避免内外网业务相互影响；
- 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，分别接入内网和外网，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计算节点，可水平扩展计算节点数量；
- 分布式存储冗余性设计，将数据均衡存储于所有磁盘，并三副本、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；
- 云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、网络流表分发服务等，保证平台高可用；
- 云平台提供的产品服务，如负载均衡、NAT 网关均采用高可用架构构建，保证云平台提供服务的可靠性。

业务实现分离

iStack 架构从业务逻辑上分为北向接口和南向接口，将云平台的业务逻辑和业务实现进行分离，业务管理逻辑不可用时，不影响虚拟资源的正常运行，整体提升云平台业务可用性和可靠性。
业务实现分离后，当云平台业务端（如 WEB 控制台）发生故障时，并不影响已运行在云平台上的虚拟机及运行在虚拟机中的业务，一定程度上保证业务高可用。

组件化

iStack 将云平台的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、网卡、IP、路由器、交换机、安全组等；
- 每种组件均支持热插拔，如将一个 IP 绑定至一个在运行中的虚拟机；
- 每种组件均支持横向扩展，如横向增加虚拟机的磁盘，提升整体云平台的健壮性 。

应用场景

虚拟化&云化

通过将业务系统和内部应用部署至 iStack 平台，可为用户提供一套集虚拟化、分布式存储、SDN 网络为一体的私有云平台。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灾备云或边缘计算，同时支持与公有云无缝打通，灵活调用公有云能力，帮助政企快速构建安全可靠的业务架构。

业务快速交付

平台服务所见即所得，可通过自服务云管理平台一键部署并管理业务交付所需的基础设施和中件间，包括在线扩容、负载分发、数据库缓存及监控日志等应用基础环境服务能力；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云平台，并可对所有业务系统的资源进行统一管理。

超融合一体机

平台提供一体机交付模式，多款机型应用不同业务场景，集成 iStack 私有云平台，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、云管等一系列云服务能力。

政企专有云

iStack 提供租户控制台和管理员控制台，支持多租户、账户注册等功能特性，同时为云平台管理者提供运营运维管理功能，包括资源管理、租户管理等服务，为政企提供行业专有云解决方案。

交付模式

iStack 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云服务平台。

- **纯软件交付-iStack**

客户提供承载云平台运行的硬件服务器、网络设备及相关基础设施，天翼云提供 iStack 轻量级私有云软件；通常在基础网络设施环境完备的情况下，iStack 软件可在 2 小时内完成部署并交付。

- **超融合机柜-iStack**

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

平台物理架构

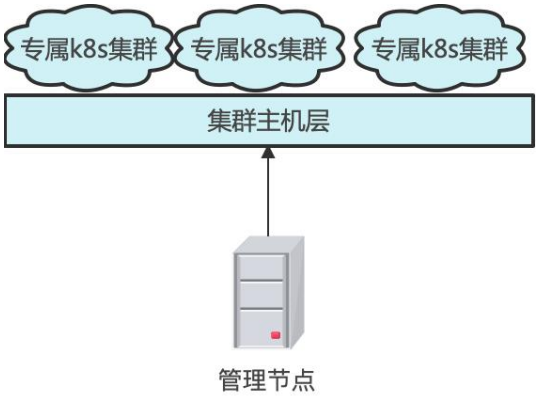
iStack 云平台在架构上分为管理节点和被管集群，一个管理节点可以管理多个被管集群。

管理节点

管理节点部署了核心的管理服务，以 WEB 和 API 的方式对外提供服务，管理节点和被管集群的部署方式可分为三种：

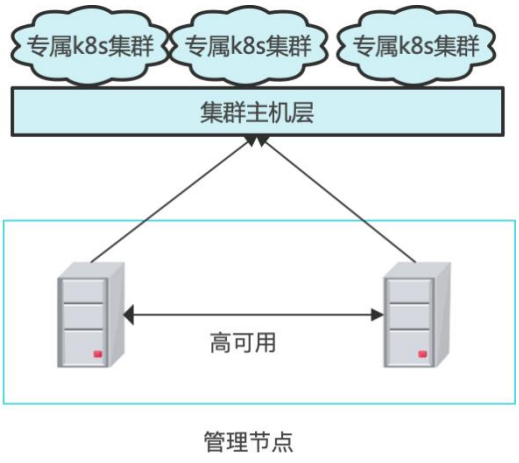
单机外置部署

指的是将管理节点以单机的模式部署在被管集群的外部，其优势是对资源消耗少，但由于是单机部署，不能保证高可用，仅仅适用于开发、测试等对高可用性不高的场合。



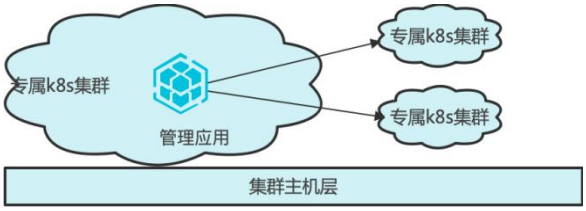
双机外置部署

指的是将管理节点以双机高可用的方式部署在被管集群的的外部，支持主备级高可用，但需要专门的管理节点资源，适合于中大规模的集群。



集群内置部署

指的是将管理节点以应用的方式部署在其中一个被管集群之中，支持高可用，同时适合超 mini 小集群和中大型集群。



被管集群

被管集群主要有计算存储融合节点、独立计算节点和独立存储节点等三种形式。

计算存储融合节点

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、数据库服务、缓存服务等资源，同时承载智能调度控制和监控服务。云平台分布式存储使用所有计算节点的数据磁盘，每个节点仅支持部署一种类型的数据磁盘，如 SATA、SSD 等（使用 SSD 作为缓存的场景除外）。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。

在部署上，每台计算节点均会部署用于运行计算存储网络的 KVM、Qemu、Libvirt、OVS、Ceph 等核心组件，同时在每个节点中至少有 3 台计算节点会部署核心调度及管理模块。

独立计算节点

集群内宿主机节点：

- 用于独立运行所有计算和网络资源，通过挂载独立存储节点的磁盘作为云平台的存储资源；
- 一般由几台到几千台服务器组成，生产环境至少部署 3 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。

独立存储节点

独立存储节点，用于独立承载分布式存储的节点，构建独立存储区域。适合将计算和存储分离，搭建独立存储网络的场景。独立存储节点，使用独立的存储网络接入设备，与计算业务物理或逻辑隔离。

- 部署独立存储节点，可节省计算节点的 CPU、内存等资源；
- 一般由几台和几千台服务器组成，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；
- 独立存储节点为【可选】节点，如果采用融合节点，可使用计算存储超融合节点上的数

据磁盘作为分布式存储的存储池。

部署存储节点时，每个节点需配置相同介质类型的数据磁盘，如全 SSD 存储节点、HDD 存储节点等（使用 SSD 作为缓存的场景除外），将相同磁盘类型的节点组成一个存储集群，分别作为普通存储和高性能存储资源池。

推荐节点方案

iStack 物理节点方案会根据企业业务需求及应用场景进行调整，可部署计算存储融合节点（例如 3 节点方案，即 3 台超融合计算节点，管理服务部署于计算集群中，后续可根据业务规模水平扩展，如将超融合计算节点扩容为 12 台），也可部署独立计算节点+独立存储节点（后续可根据业务规模水平扩展计算节点或存储节点）。

SSD 和 HDD 节点的配比取决于业务需求，如高存储容量需求较大，则需配置较多的 HDD 节点；若高性能业务需求较多，则需配置较多的 SSD 全闪节点。

最佳实践中，生产环境至少需要 3 台 超融合节点部署搭建 iStack 平台，即 iStack 最小生产规模为 3 台服务器。

平台技术架构

iStack 平台基于 ECX 公有云平台，复用公有云核心组件，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。

计算虚拟化

云计算技术是虚拟化技术的延伸，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，最大程度上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。

物理机和虚拟机在应用部署及资源占用上有本质区别：

- 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的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虽部署至一台物理机上，在非业务高峰时资源利用率较低。

- 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的操作系统一致，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。

iStack 计算虚拟化采用 kubernetes 生态圈的 kubevirt 组件，以容器的方式运行虚机，底层基于 kvm，Qemu 等 Hypervisor 组件及技术，将通用裸金属架构的 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，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的完全虚拟化技术，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。

iStack 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。

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 GuestOS 提供诸如网卡、磁盘、USB 设备等外设。

计算虚拟化是 iStack 企业专有云平台的服务器虚拟化组件，是整个云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、虚拟机异常重启及集群平滑扩容等特性。

CPU 超分

iStack 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。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 1:3 ，即代表可使用的 vCPU 数量是实际 CPU 数量的 3 倍。服务器（48C ）在开启 3 倍超分后，可实际创建使用的 vCPU 为 144 ，即可创建 144C 的虚拟机。支持向下修改，如果已经设置了超分比为 **1:3**，后续支持将超分比调为 **1:1**。

当前仅支持平台专业的运营人员设置并管理 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 144 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。

QCOW2 镜像文件

iStack 平台使用 QCOW2 格式的镜像作为虚拟机的虚拟磁盘文件，即原始镜像。

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。

热插拔数据盘

iStack 平台支持对数据盘进行热插拔，即在虚机不关机的情况下，在线添加数据盘或删除数据盘，并且添加或删除后，虚机内部立刻可以相应地识别并正常使用。

快照

iStack 平台支持对虚机执行快照，以及快照的恢复。虚机执行快照时，会保留当前虚机的数据情况以及运行状态，后续可对虚机进行回滚，回滚到某一个快照点。

快照功能实际上是对虚机的系统盘以及数据盘进行快照，因此虚机的快照能力，依赖于底层分布式存储的快照能力，以及 kubernetes 的快照能力。

自定义镜像

iStack 平台支持用户提交自定义的镜像。用户使用平台公共镜像创建虚机后，可对虚机做一些应用配置，将虚机提交为自定义的镜像，后续可以从自定义的镜像去创建新虚机。此功能方便用户创建批量相同应用功能的虚机，节省了应用部署的时间。

GPU 透传

iStack 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 云硬盘，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，包括内弹性 IP 分配、弹性网卡、子网及安全组管理，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，为租户提供快捷的 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	8GiB，16GiB
	8 核	16GiB，32GiB
2 颗	8 核	16GiB，32GiB
	16 核	32GiB，64GiB
4 颗	16 核	32GiB，64GiB
	32 核	64GiB，128GiB

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。
平台默认不支持 GPU 虚拟化，如需 GPU 虚拟化能力，需购买 GPU 虚拟化授权。

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，适用于云平台资源动态调整、物理机停机维护、优化服务器能源消耗等场景，进一步增强云平台可靠性。

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、云硬盘）位置迁移。迁移时仅需通过 kubernetes 集群内保存的源虚拟机配置，在目的主机上注册一个相同配置的虚拟机进程，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，**虚机挂载的云盘迁移后不受影响**，可为用户提供无感知且携带存储数据的迁移服务。

宕机迁移

宕机迁移又称离线迁移（Offline Migration）或虚拟机高可用（ 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、云硬盘）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。

存储虚拟化

云计算平台通过硬件辅助的虚拟化计算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，企业还需考虑适用于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 SAN 存储：即磁盘阵列，通常为软硬一体的单一存储，采用 RAID 保证数据安全。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 廉价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。

商业化 SAN 存储设备

iStack 对 FCSAN 以及 IPSAN 均支持，其中包含的基础概念如下：

- SCIS：小型计算机系统接口，是计算机主机内部设备之间（硬盘、软驱、光驱、打印机、扫描仪等）系统级接口的独立处理器标准。

- **iSCIS**：Internet 小型计算机系统接口。iSCSI 就是用广域网仿真了一个常用的高性能本地存储总线。
- **SAN**：存储区域网络，一种专门为存储建立的独立于 TCP/IP 网络之外的专用网络。
- **FCSAN**：基于 FC 的 SAN 存储，采用光纤通道(Fibre Channel，FC)技术，通过光纤通道交换机连接存储阵列和服务器主机，建立专用于数据存储的区域网络。客户端只能通过 HBA 来连接存储。
- **IPSAN**：基于 IP 的 SAN 存储，以 iSCSI 来对外提供服务，客户端可以通过 IP 协议来连接存储。
- **LUN**：逻辑单元（Logical Unit Number），集中存储上的逻辑管理单元，对于主机来讲，就是一块裸设备；对于存储来讲是一个可被主机识别的独立存储单元。

以 IPSAN 为例，iStack 支持 SAN 的方案分为五个层次：

- **存储层**：在 IPSAN 设备上创建 LUN 并通过 iSCSI 协议开放给上层。
- **设备层**：在每台主机上，通过 iSCSI 将存储层映射为主机上的存储裸设备。
- **文件系统层**：由集群软件将多台主机上的设备构建成共享文件系统。
- **存储引擎**：在共享文件系统的基础上，为虚拟化提供磁盘文件。
- **虚拟化层**：存储引擎所提供的磁盘文件以裸设备的形式映射到虚拟内部，作为磁盘使用。

如果是 FCSAN 存储，可以在主机上映射为裸设备，对于上层来说完全透明。出于性价比的考虑，建议采用 IPSAN。

分布式存储

iStack 云平台基于 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 x86 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。作为云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。

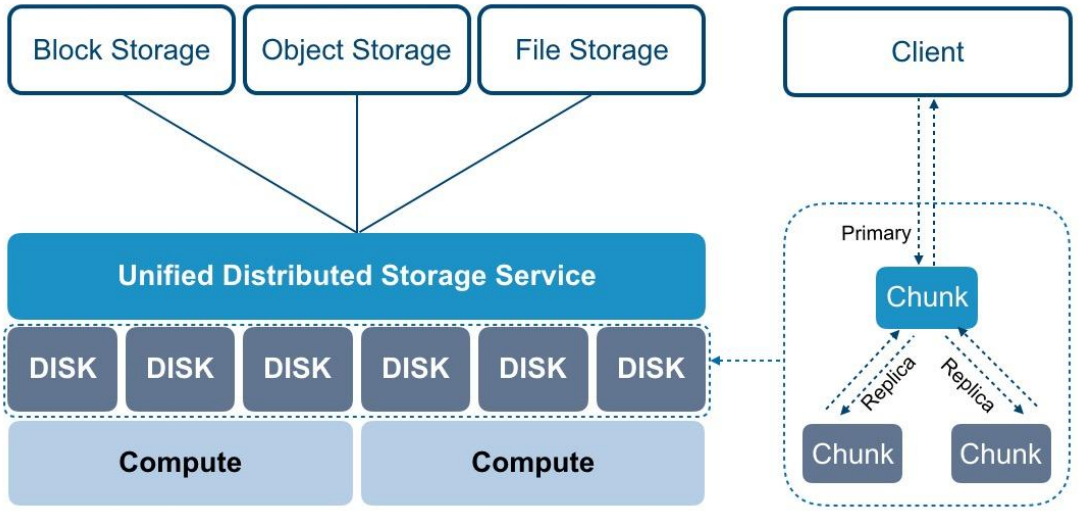
分布式存储服务通过将大量 x86 通用服务器的磁盘存储资源融合在一起进行【池化】，构建一个无限可伸缩的统一分布式存储集群，实现对数据中心所有存储资源的统一管理及调度，向虚拟化计算层提供【块】存储接口，供云平台虚拟机或虚拟资源根据自身需求自由分配并使用存储资源池中的存储空间。同时云平台虚拟化通过 iSCSI 协议对接 IPSAN 商业存储设备，将商业存储作为虚拟化后端存储池，提供存储池管理及逻辑卷分配，可直接作为虚拟机的系统盘及数据盘进行使用，即只要支持 iSCSI 协议的存储设备均可作为平台虚拟化的后端存储，适应多种应用场景；可利旧企业用户的集中存储设备，整体节省信息化转型的总拥有成本。

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，即可在云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，云平台用户像使用 x86 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。云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。

分布式存储系统可提供块存储、文件存储及对象存储服务，适用于多种数据存储的应用场景，同时可保证数据的安全性及集群服务的可靠性。文件存储和对象存储在一个数据中心部署一

套集群，支持机械盘和高性能盘混合部署且可逻辑划分多个存储池，如高性能存储池和容量型存储池。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，如超融合计算节点和独立存储节点自带 SSD 磁盘构建为高性能的存储集群；超融计算节点和独立存储节点自带的 SATA/SAS 磁盘构建为普通性能存储集群。

分布式存储存储系统将集群内的磁盘设备通过 OSD 内建不同的存储资源池，分别提供弹性块存储服务、对象存储及文件存储服务，其中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。文件存储和对象存储可提供诸如 NFS、CIFS、S3 等多种协议接口为应用服务提供非结构化数据存储服务，同时结合多副本及纠删码数据冗余策略满足多种场景下的数据存储和处理，逻辑架构如下：



iStack 分布式存储系统是整个云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云平台的服务质量。

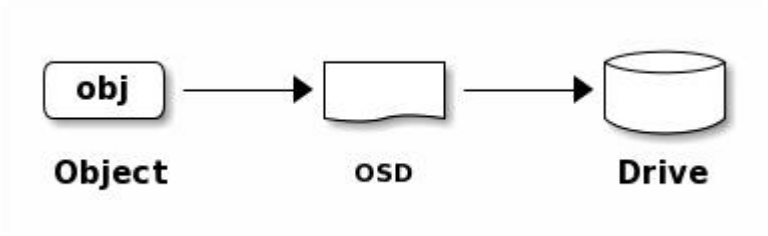
智能存储集群

分布式存储集群可包含数千个存储节点，通常至少需要一个监视器和多个 OSD 守护进程才可正常运行及数据复制。分布式智能存储集群消除集中控制网关，使客户端直接和存储单元 OSD 守护进程交互，自动在各存储节点上创建数据副本确保数据安全性和可用性。其中包括的基础概念如下：

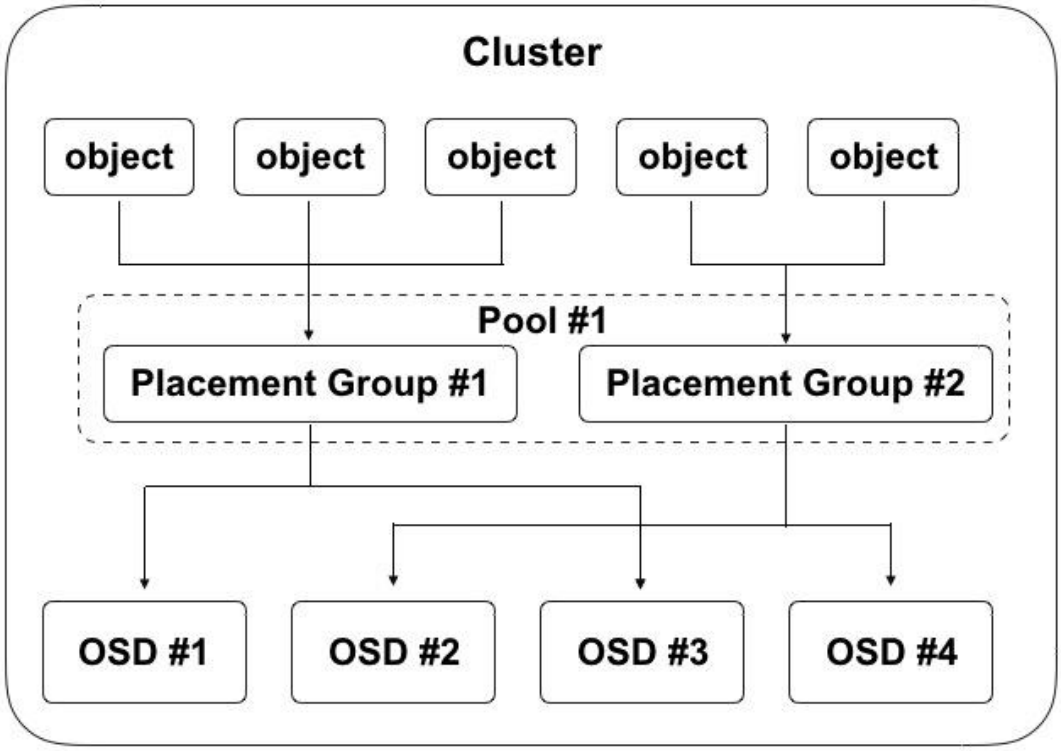
- **OSD**：通常一个 OSD 对应物理机一块磁盘、一个 RAID Group 或者一个物理存储设备，主要负责数据存储、处理数据复制、恢复、回填及数据重均衡，并负责向监视器报告检测信息。单集群至少需要两个 OSD，并在物理架构可划分为多个故障域（机房、机架、服务器），通过策略配置使多副本位于不同的故障域中。
- **监视器 Monitor**：实现存储集群的状态监控，负责维护存储集群的 Object、PG 及 OSD 间的映射关系图，为数据存储提供强一致性决策，同时为客户端提供数据存储的映射关系。

- **元数据服务 MDS**：实现文件存储服务时，元数据服务（MDS）管理文件元数据。
- **客户端**：部署在服务器上，实现数据切片，通过 CRUSH 算法定位对象位置，并进行对象数据的读写。通常包括块设备、对象存储及文件系统客户端，读/写操作由 OSD 守护进程处理。
- **CRUSH 算法**：用于保证数据均匀分布的伪随机算法，OSD 和客户端均使用 CRUSH 算法来按需计算对象的位置信息，为存储集群动态伸缩、重均衡和自修复功能提供支撑。

存储数据时，存储集群从客户端（块设备、对象存储、文件系统）接收数据，并将数据分片为存储池内的对象 Object，每个对象直接存储至 OSD 的裸存储设备上，由 OSD 进程处理裸设备上的读写操作。如下图所示：



客户端程序通过与 OSD 或监视器交互获取映射关系数据，在本地通过 CRUSH 算法计算得出对象存储位置后，直接与对应的 OSD 进行通信，完成数据读写操作。为实现分布式存储集群可自主、智能且自我修复的存取数据，智能存储集群通过 CURSH 算法、存储池 Pool、放置组 PG 及 OSD 等多种逻辑概念相互关联承载数据存储流程，逻辑架构图如下：



- 一个集群可逻辑上划分为多个 Pool ，Pool 是一个命名空间，客户端存储数据时需指定一个 Pool；
- 一个 Pool 包含若干个逻辑 PG（Placement Group），可定义 Pool 内的 PG 数量和对象副本数量；
- PG 是对象和 OSD 的中间逻辑分层，写对象数据时，会根据 CRUSH 算法计算每个对

象要存储的 PG ；

- 一个物理文件会被切分为多个 Object ，每个 Object 会被映射到一个 PG ，一个 PG 包含多个 Object ；
- 一个 PG 可映射到一组 OSD ，其中第一个 OSD 为主 ，其它 OSD 为从，Object 会被均匀分发至一组 OSD 上进行存储；
- 承载相同 PG 的 OSD 间相互监控存活状态，支持多个 PG 同时映射到一个 OSD 。

在存储集群的机制中，承载相同 PG 的主从 OSD 间需要彼此交换信息，确保彼此的存活状态。客户端首次访问会首先从监视器获取映射关系的数据，存储数据时会与 OSD 对比映射关系数据的版本。由上图示意图得知，一个 OSD 可同时承载多个 PG ，在三副本机制下每个 PG 通常为 3 个 OSD 。如上图所示，数据寻址流程分为三个映射阶段：

1、将用户要操作的文件映射为存储集群可处理的 Object ，即将文件按照对象大小进行分片处理；
2、通过 CRUSH 算法将所有文件分片的 Object 映射到 PG ；
3、将 PG 映射到数据实际存储的 OSD 中 ，最后客户端直接联系主 OSD 进行对象数据存储操作。
分布式存储客户端从监视器获取集群映射关系图，并将对象写入到存储池。集群存储数据的逻辑主要取决于存储池的大小、副本数量、CRUSH 算法规则及 PG 数量等。

超大规模扩展

在传统集中式架构中，中心集群组件作为客户端访问集群的单一入口，这将严重影响集群的性能和可扩展性，同时引入单点故障。在分存储存储集群的设计中，存储单元 OSD 和存储客户端能直接感知集群中的其它 OSD 及监视器信息，允许存储客户端直接与存储单元 OSD 交互进行数据读写，同时允许每个 OSD 与监视器及其它节点上的 OSD 直接交互进行数据读写，这种机制使得 OSD 能够充分利用每个节点的 CPU/RAM ，将中心化的任务分摊到各个节点去完成，支持超大规模集群扩展能力，提供 EB 级存储容量。

- OSD 直接服务于客户端，存储客户端直接与 OSD 进行通信，消除中心控制器及单点故障，提升整体集群的性能及可扩展性。

- OSD 之间相互监测彼此的健康状态，并主动更新状态给监视器，使监视器可以轻量化部署和运行。

- OSD 使用 CRUSH 算法，用于计算数据副本的位置，包括数据重平衡。在多副本机制中，客户端将对象写入主 OSD 中后， 主 OSD 通过自身的 CRUSH 映射图识别副本 OSD 并将对象复制到副本 OSD 中；凭借执行数据副本复制的能力，OSD 进程可减轻存储客户端的负担，同时确保高数据可用性和数据安全性。

为消除中心节点，分布式存储客户端和 OSD 均使用 CRUSH 算法按需计算对象的位置信息，避免对监视器上集群映射图的中心依赖，让大部分数据管理任务可以在集群内的客户端和

OSD 上进行分布式处理，提高平台的可伸缩性。
为消除中心节点，分布式存储客户端和 OSD 均使用 CRUSH 算法按需计算对象的位置信息，避免对监视器上集群映射图的中心依赖，让大部分数据管理任务可以在集群内的客户端和 OSD 上进行分布式处理，提高平台的可伸缩性。

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**

存储集群不强行绑定硬件及品牌，可采用通用服务器及网络设备，支持存储集群异构。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 Object、PG 及 OSD 间的主映射图，包括集群成员、状态、变更、以及存储集群的整体健康状况等。OSD 和客户端均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。

- **存储接入负载均衡**

对象存储和文件存储接入网关支持负载均衡服务，保证对象存储和文件存储网关高可用，同时为存储网关提供流量负载分发，提升存储的整体性能。在负载均衡的接入机制下，读写 I/O 会均衡到集群中所有网关服务上，当其中一台网关服务器出现异常时，会自动剔除异常网关节点，屏蔽底层硬件故障，提升业务的可用性。

多冗余策略

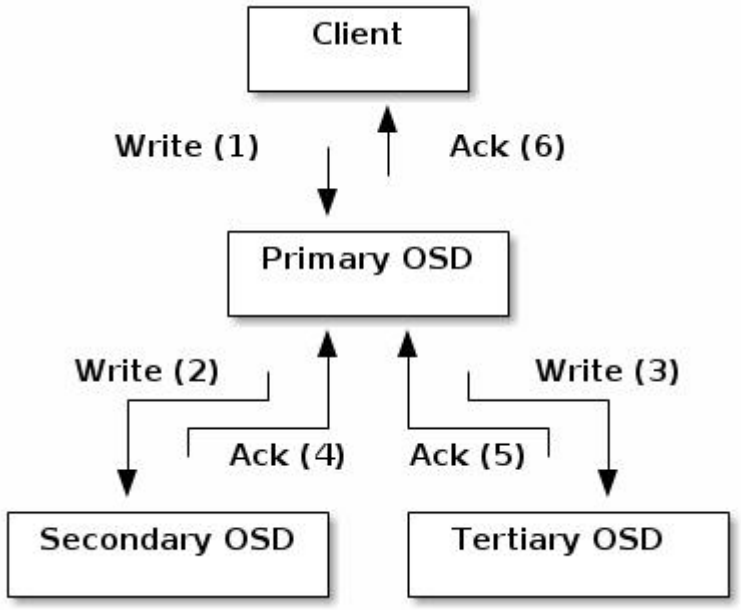
多副本机制

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。iStack 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**。

iStack 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

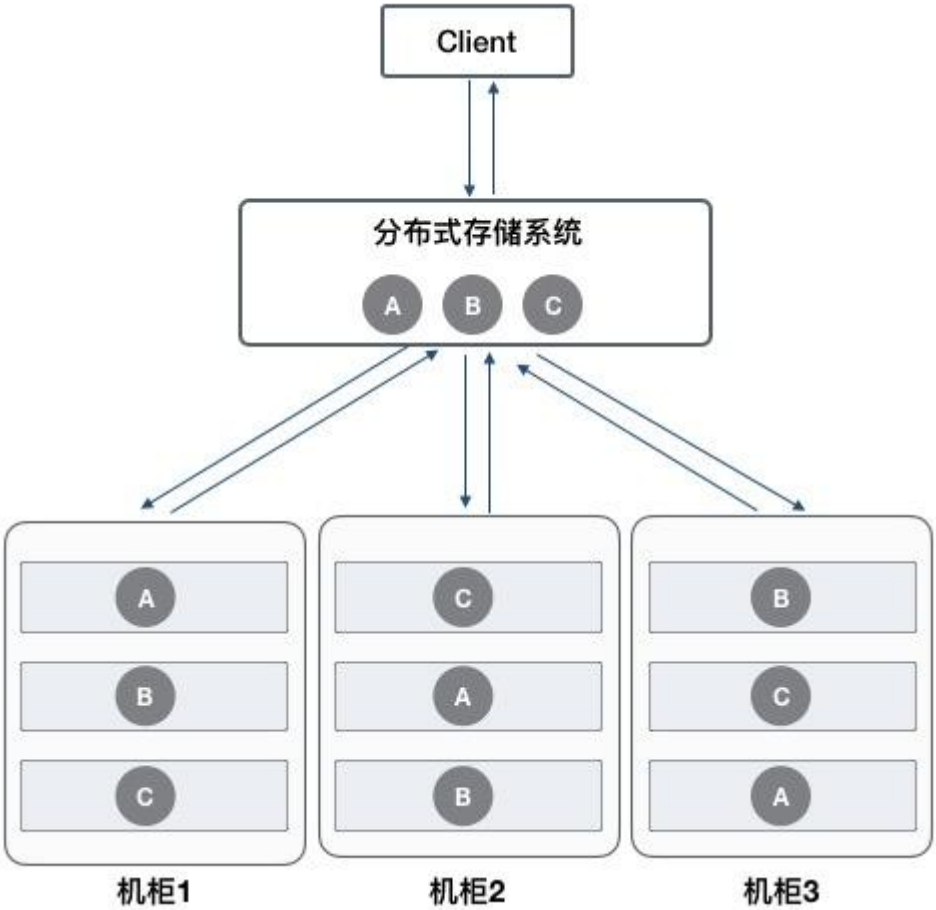
如上图所示，三副本在写入过程中，只有三个写入过程全部被确认，才返回写入完成，确保数据写入的强一致性。

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。平台通过定义存储桶类型，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，可能会影响云硬盘的 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，即定义机柜和主机存储桶。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。

纠删码策略

纠删码（Erasure Coding，EC）是一种数据保护方法，类似商业存储中的 RAID5 技术，它将数据分割成片段，把冗余数据块扩展、编码，并将其存储在不同的位置，比如磁盘、存储节点或者其它地理位置。iStack 分布式对象存储和文件存储可采用纠删码策略进行数据冗余保护。

纠删码策略可兼备数据安全性和磁盘利用率，在分布式存储系统中，纠删码策略将写入的数据进行分片（称为数据块），基于分片编码生成备份冗余数据（称为较验块），最后将原始分片数据和备份数据分别写入不同的存储介质，以保证数据的安全性。同时数据块和校验块可通过故障域分别存储于跨节点、跨机柜、跨数据中心的不同 OSD 磁盘上，多维度保证数据安全。

通过纠删码策略将数据分段的块数称为 K，编码较验块称为 M，所有数据块个数称为 N ，即 N= K+M；基于此磁盘利用率可通过 K/N 获得，如 K=9，M=3 ，N=12 ，则磁盘总空

简利用率为 9/12=75% ，即磁盘利用率为 75%。对象存储和文件存储的存储集群根据冗余策略不同，磁盘利用率不同。在三副本机制下，磁盘利用率为集群总容量的三分之一；而在纠删码策略中，利用率与 K+M 的比例值相关，即不同的 K+M 值，会有不同的磁盘利用率，可根据实际使用场景自定义纠删码策略的 K+M 值，平台默认推荐 4+2。

以 4+2 为例，在写入数据时存储系统会先将对象数据映射到一个 PG ，再由 PG 映射到一组 OSD 中（OSD 的数量取决于 K+M 的值，即 OSD 数量与 N 的值相等）；同时在 OSD 中选举出主 OSD ，由主 OSD 将对象文件分片为 4 个数据块，在通过 4 个数据块编码出 2 个校验块，最后将 4 个数据块和 2 个校验块分别写入 6 个 OSD 中。在读取数据时，由主 OSD 分别从相同 PG 的其它 OSD 中读取所有需要的分片数据，最后由主 OSD 统一汇总拼拼凑出指定的对象文件，向客户端应答。

由于纠删码是将数据切片并发写入至多个 OSD 磁盘，并无多副本机制中多倍写放大的问题，因此写性能较有优势。而读数据时需要先计算数据分片，再将多个 OSD 中的数据读出来进行汇总，因此读性能相对偏低。

纠删码的原理证明，存储集群中允许损坏的数据块数量小于等于 M（较验块），在存储节点或磁盘无故障或异常时，对象数据块和校验块始终不变。仅当节点发生异常时，数据块和较验块小于 N+M 值时，需要通过剩余的数据块和较验块一起进行解码计算出损坏数据，并将期恢复在正常的 OSD 设备中。如上图 5+3 的 EC 策略中，允许失败的数据为 3 ，即在实际生产环境中可允许失败 3 块磁盘；当分片数据 D1 损坏时，主 OSD 计算获取到对象文

件剩余的数据块（D2—D5 ）及较验块信息（ P1—P3），通过 EC 对数据块和较验块的信息解码，计算出损坏的 D1 数据 D1'，最后将 D1' 数据恢复至正常的 OSD 设备中，完成损坏数据的恢复。

若在读取数据时正好有一个分片数据损坏，则会同步进行数据解码恢复操作，则读取该数据的时延较大，会影响整体数据的读取性能。

由于纠删码在存取数据时需要消耗更多的计算资源，因此纠删码对节点的计算要求相较多副本高，然纠删码以其灵活多变的数据备份策略、较高的存储空间利用率非常适合存储大量对时延不敏感的数据，如备份数据、办公应用数据、日志数据等。基于此，iStack 的文件存储和对象存储可提供纠删码和多副本两种冗余保护策略，而块存储仅采用多副本机制进行数据安全保护。

数据重均衡

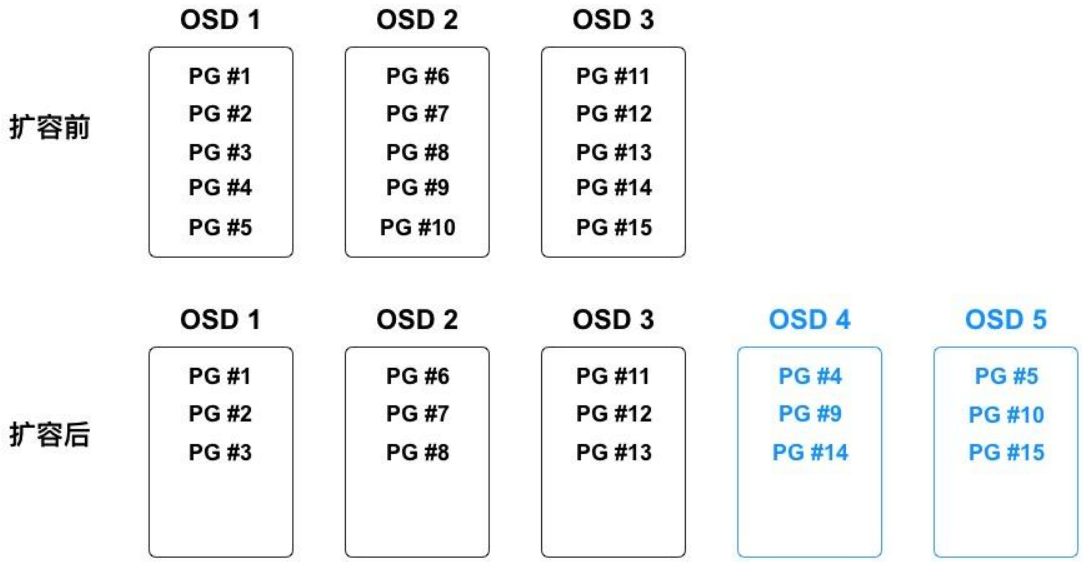
iStack 云平台分布式存储集群在写入数据时，会通过数据分片、CRUSH 映射关系、多副本或纠删码分布策略尽量保证数据对象在存储池中的均衡。随着存储集群的长期运行及对平台的运维管理，可能会导致存储池内的数据失衡，如存储节点和磁盘扩容、存储部分数据被删除、磁盘和主机故障等。

- 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；
- 用户删除虚拟机或云硬盘数据，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，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。

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：

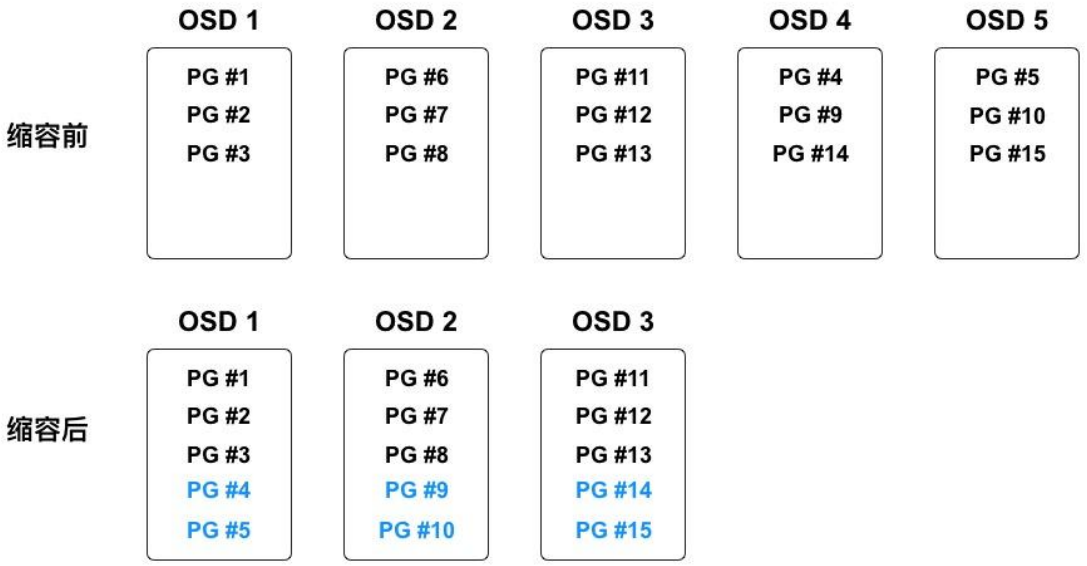


在数据均衡过程中，仅会将现有 OSD 中的部分 PG 到迁移到新的 OSD 设备，不会迁移所有 PG ，尽量让所有 OSD 均腾出部分容量空间，保证所有 OSD 的对象数据分布相对均衡。如上图中新增 OSD 4 和 OSD 5 后，有三个 PG（PG #4、PG #9、PG #14） 迁移到 OSD 4 ，三个 PG （PG #5、PG #10、PG #15）迁移到 OSD 5 ，使五个 OSD 中映射的 PG 均为 3 个。为避免 PG 迁移导致集群性能整体降低，**存储系统会提高用户读写请求的优先级**，在系统空闲时间进行 PG 迁移操作。

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。

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。

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。
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。

数据故障重建

根据多副本和 EC 纠删码的保护机制，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。

在 EC 纠删码策略下，节点或磁盘设备故障时会导致部分数据块或校验块丢失，如 4+2 的

纠删码数据会丢失一个数据块或校验块，此时对象数据的 PG 被置为“降级”模式继续提供存储服务，并开始进行纠删数据的解码和恢复操作，按照最新 CRUSH 运行图将故障数据块或校验块数据重新恢复至其它健康的 OSD 设备上，保证对象数据的完整性和可用性。

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或 EC 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，副本或纠删码 EC 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。

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。

自动精简配置

自动精简配置（Thin Provisioning），又称【超额申请】或【运行时空间】，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。

iStack 云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。如一个用户创建的云硬盘为 1TB 容量，存储系统会为用户分配并呈现 1TB 的逻辑卷，仅当用户在云硬盘中写入数据时，才会真正的分配物理磁盘容量。若用户在云硬盘上存储的数据为 100GiB ，则云硬盘仅使用存储池的 100GiB 容量，剩余的 900GiB 容量可以供其它用户使用。

云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，即可供租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。

块存储服务

iStack 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、对象及文件存储服务，同时提供多种数据接口，用户无需关注底层存储设备及架构，即可在云平台构建并使用存储服务，适用于虚拟化、云计算、大数据、物联网及企业应用等使用场景。

iStack 基于分布式存储系统为云平台租户提供块设备，即云硬盘服务，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云硬盘进行扩容，基于网络分布式访问，为云主机提供高安全、高可靠、高性能及可扩展的数据磁盘。

云平台为租户提供普通和高性能两种架构类型的云硬盘，普通云硬盘使用 SATA/SAS 磁盘作为存储介质，性能型云硬盘使用 SSD/NVME 磁盘作为存储介质。云硬盘数据均通过 PG 映射及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云硬盘资源及全生命周期管理。

在业务数据安全方面，云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。

网络虚拟化

网络是虚拟化计算和分布式存储为云平台提供服务时不可或缺的核心部分，通常可采用硬件定义的 UnderLay 网络或软件定义的 OverLay 网络与虚拟化计算对接，为云平台提供多应用场景的网络及信息传输服务。

- **UnderLay 网络**

传统 IT 架构中硬件方式定义的单层物理网络，由物理设备和物理链路组成，即当前数据中心物理基础转发架构层——物理底层承载网，包括一切现有的传统网络技术，负责互联互通。常见的物理设备有交换机、路由器、负载均衡、防火墙、IDS/IPS 等。

- **OverLay 网络**

虚拟网络，基于底层 UnderLay 网络架构上叠加隧道技术构建的逻辑网络，实现网络资源虚拟化，以软件的方式在虚拟化平台上完整再现物理网络的功能。

OverLay 网络的核心是隧道技术，只负责虚拟化计算资源的网络通信，具有独立的控制面和转发面（SDN 的核心理念）。对于连接到 OverLay 的终端设备（例如服务器）来说，物理网络是透明的，从而可以实现承载网络和业务网络的分离。

作为云计算核心技术之一的虚拟化计算已被数据中心普遍应用，UnderLay 网络和 OverLay 网络均可为虚拟化计算提供网络服务。随着业务规模的增长，虚拟机数量的快速增长和迁移已成为一个常态性业务，如果采用传统 IT 架构中硬件方式定义的 UnderLay 网络，可能会给云平台带来一些问题：

- **网络隔离能力限制**

UnderLay 主流的网络隔离技术是 VLAN ，由于 IEEE 802.1Q 中定义的 VLAN ID 为 12 比特，仅能实现 4096 个 VLAN ，无法满足大二层网络中标识大量租户或租户群的需求。同时由于 VIan 技术会导致未知目的广播数据在整网泛滥，无节制消耗网络交换能力与带宽，仅适合小规模的云计算虚拟化环境。

- **虚拟机迁移范围受网络架构限制**

为保证虚拟机热迁移，需保持虚拟机的 IP 地址和 MAC 地址保持不变，即要求业务网络为二层网络且需具备多路径的冗余备份和可靠性。传统物理网络 STP、设备虚拟化等技术部署反锁且不适合大规模网络，限制虚拟机的网络扩展性，通常仅适用于数据中心内部网络。为大规模网络扩展的 TRILL/SPB/FabricPath/VPLS 等技术，虽可解决规模问题，但均需网络中的软硬件进行升级而支持此类新技术，增加云计算平台的部署成本。

- **虚拟机规模受网络规格限制**

在传统二层网络环境下，数据报文是通过查询 MAC 地址表进行二层转发，而网络设备 MAC 地址表的容量限制了虚拟机的数量。若选择适配较大容量 MAC 地址表的网络设备，则会提升网络建设成本。

- **部署缓慢且僵化**

虚拟化计算快速部署及灵活扩展特性上，均需网络提供强有力的支撑。传统网络中虚拟机部署业务及上线，均需对系统及网络设备进行繁琐的配置，甚至需要改变物理设备部署位置，降低业务发布效率的同时，难以快速响应新业务灵活部署及发布。

基于上述的问题和场景，可在 UnderLay 网络基础架构上采用 OverLay 网络解决方案，构建大二层虚拟网络，实现业务系统间网络隔离，并通过 NFV 实现网络中所需的各类网络功能和资源，按需灵活的调度资源，功能所见即所得，从而实现云计算平台中的网络虚拟化，满足虚拟化计算对网络的能力需求。

- **网络隔离能力**

OverLay 网络虚拟化提供多种隧道隔离技术，如 VXLAN 、GRE、NVGRE、STT 等，均引入类似 VIan 的用户隔离标识，并对隔离标识进行极大扩展，如 VXLAN 支持 24 比特，可支持千万级以上的网络隔离标识。

- **隧道路由网络**

OverLay 通过隧道技术，将二层以太报文封装在三层 IP 报文之上，通过路由的方式在网络中分发传输。路由网络本身无特殊网络结构限制，具备大规模扩展能力和高性能转发能力，

同时路由三层网络会缩小二层广播域，大幅降低网络广播风暴的风险，具备很强的故障自愈能力和负载均衡能力。通过 OverLay 技术的路由网络，虚拟机迁移不受网络架构限制，企业部署的现有网络便可用于支撑新的云计算业务。

- **大规模虚拟机规模**

虚拟机发出的数据包封装在 IP 数据包中，对网络只表现为封装后的网络参数，即隧道端点的地址。因此极大的降低大二层网络（UnderLay）对 MAC 地址表容量的需求，可支撑大规模虚拟机场景。

- **快速灵活部署**

基础网络不感知虚拟网络业务变化，OverLay 网络中应用部署的位置将不受限制，网络功能所见即所得，支持即插即用、自动配置下发及自动运行，可快速并灵活的部署业务，并支持业务在虚拟网络中进行迁移和变更。

网络虚拟化通过结合软件定义网络（SDN Software Defined Network）和网络功能虚拟化（NFV Network Function Virtualization）提供服务。SDN 是一种全新的网络架构，核心思想是通过标准化技术（如 openflow ）将网络控制面和数据转发面进行分离，由控制器统一计算并下发流表，进而实现对网络流量集中化、灵活化、细粒度的控制。NFV 是指具体网络设备的虚拟化，使用通用服务器和软件实现并运行网络功能，如虚拟网卡、虚拟交换机、虚拟防火墙等，实现网络功能灵活配置、快速部署及定制编程能力。

SDN 和 NFV 是高度互补关系，各有侧重，分别从不同角度提供解决方案满足不同虚拟化场景的网络需求。SDN 通过将控制平台和数据转发面分离实现集中的网络控制，而 NFV 技术是通过软硬件分离，实现网络功能虚拟化。二者的关系如下：

- SDN 技术在流量路由方面所提供的灵活性，结合 NFV 的虚拟化架构，可更好地提升网络的效率，提高网络整体的敏捷性。
- NFV 不依赖 SDN ，可在无 SDN 的情况下进行虚拟化部署，但 SDN 中控制和数据转发分离可改善 NFV 网络性能、易用性及可维护性，可实现 NFV 的快速部署及网络构建。

iStack 通过 **OVS+Geneve** 的 OverLay 网络及软件定义的 SDN 控制器，构建大二层虚拟网络，实现业务系统间网络隔离；并通过 NFV 实现网络中所需的各类网络功能和资源，用于对接 KVM 虚拟化计算服务，结合分布式网络架构为平台提供高可用、高性能且功能丰富的网络虚拟化能力及管理。

iStack 通过软件定义网络 （ SDN ）对传统数据中心物理网络进行虚拟化，虚拟化网络功能所见即所得，用户无需关注底层设备类型及网络架构，即可在云平台构建使用虚拟网络服务，包括私有网络 VPC 、网络隔离、弹性 IP、NAT 网关、负载均衡、安全组等网络服务，承载云平台上虚拟资源的网络通信及安全。

VPC 网络

软件定义虚拟专有网络，用于租户间数据隔离。提供自定义 VPC 网络、子网规划及网络拓

朴，可将虚拟机加入私有网络和子网，为虚拟机提供二层网络服务。

网络隔离能能力

由 VPC 提供的逻辑隔离的二层网络广播域环境，为云平台用户或子帐号提供网络隔离能力，不同 VPC 网络间网络完全隔离，不可进行通信。

弹性 IP

用于 VM、负载均衡及 NAT 网关等资源的互联网接入。支持多运营商线路接入并可调整弹性 IP 的带宽上限。

NAT 网关

企业级 VPC 网关，为云平台资源提供 SNAT 和 DNAT 代理，支持外网和物理网两种网络地址转换能力，并支持 VPC 级、子网级及实例级 SNAT 规则。

负载均衡

基于 TCP/UDP/HTTP/HTTPS 协议将网络访问流量在多台虚拟机间自动分配的控制服务，类似于传统物理网络的硬件负载均衡器.用于多台虚拟机间实现流量负载及高可用，提供内外网 4 层和 7 层监听及健康检查服务。

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，为云平台提供必要的安全保障。

产品功能架构

核心概念

集群

集群（ Cluster）是 iStack 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。

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 集群可为租户提供 GPU 虚拟机，ARM 集群可为租户提供基于 ARM 或国产化 OS 的虚拟机。

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、HDD 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云盘源，如 SSD 存储集群可为租户提供 SSD 类型的云硬盘。

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云平台的服务质量。

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。集群默认对所有租户开放权限，平台支持对存储集群进行权限控制，用于将部分物理存储资源独享给一个或部分租户使用，适用于专属私有云场景。修改集群权限后，集群仅可对指定的租户开放并使用，无权限的租户无法查看并使用受限的集群创建云盘资源。

虚拟机

虚拟机是 iStack 云平台的核心服务，提供可随时扩展的计算能力服务，包括 CPU 、内存、操作系统等最基础的计算组件，并与网络、磁盘、安全等服务结合提供完整的计算环境。通过与负载均衡、数据库、缓存、对象存储等服务结合共同构建 IT 架构。

- iStack 云平台通过 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）和虚拟内存表示，存储能力通过云存储容量和性能体现；
- 虚拟机管理程序通过控制 vCPU、内存，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。

虚拟机是云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、重启、开机、重置密码、快照、快照恢复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 、SSH 等方式进行访问和管理，拥有虚拟机的完全控制权限。

实例规格、镜像、VPC 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、虚拟网卡及 IP 信息。在虚拟机基础之上，可绑定云硬盘、弹性 IP 及安全组，为虚拟机提供数据盘、公网 IP 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。

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、VPC 网络、云硬盘、安全组及密钥，提供一台完整能力的虚拟机。

- 默认提供 1C2GiB 、2C4GiB 、4C8GiB 、8C16GiB 、16C32GiB 等实例规格；
- 虚拟机规格支持 vCPU 设置 3 倍超分；
- 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；

创建虚拟机规格支持根据不同的集群创建不同的规格，即可为不同的机型创建不同的规格，租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，适应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 的倍数进行增加，如 1C、2C、4C、6C ，最大值为 240C。
- 内存规格支持（GiB）：除 1 以外，以 2 的倍数进行增加，如 1GiB、2GiB、4GiB、6GiB ，最大值为 1024GiB。

创建出的规格即可被所有租户看到并使用，可根据业务需求在不同的集群中创建不同的规格。

镜像

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公共镜像和私有镜像，支持将私有镜像共享给平台所有租户使用。

公共镜像

公共镜像是由 iStack 提供，包括多发行版 Centos 、Ubuntu 及 Windows 等原生操作系统。

- 公共镜像默认所有租户均可使用，默认提供的镜像包括 Centos 7.5、Centos 7.9 64 、Centos 8.2、Windows 2016、Windows 2019、Ubuntu 16.04、debian 9.13；
- 公共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；
- 公共镜像为系统默认提供的镜像，支持查看、更新、导出镜像以及通过镜像运行虚拟机，不支持修改、创建、删除；
- Linux 镜像默认系统盘为 50GiB ，Windows 镜像默认系统盘为 50GiB ，支持系统盘容量扩容；

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

私有镜像

私有镜像由租户或管理员通过虚拟机自行创建的自有镜像，或自行上传 QCOW2 文件格式的镜像，后续可用于创建虚拟机，除平台管理员外仅账号自身有权限查看和管理。私有镜像支持导出、删除以及共享镜像给平台其他租户共同使用。

镜像存储

公共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；

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，创建虚拟机时随 VPC 网络默认创建的虚拟网卡。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上 kubernetes 集群网络，通过 kubeovn 组件与其它虚拟网络进行通信。

- 每个虚拟机默认会生成 1 块虚拟网卡，可根据需要添加直通网卡和附加网卡，并且支持对网卡端口限速；
- 在虚拟机启动时，根据选择的 VPC 子网自动发起 DHCP 请求以获取内网 IP 地址，并将网络信息配置在虚拟网卡上，为虚拟机提供内网访问；
- 虚拟机启动后，可申请 EIP（浮动 ip，也即弹性 IP）绑定至虚拟机，提供互联网访问服务；
- 支持预留部分内网 IP 作为虚拟 IP，通过绑定虚拟机，可以提高虚拟机实例的内网访问可靠性；
- 支持修改虚拟网卡的 IP 地址，手动修改的 IP 地址需在虚拟机关联的子网网段中。

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，为云平台提供必要的安全保障。

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 200 条安全组规则，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、方向及描述六个元素组成：

- 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，全部 TCP 代表所有 TCP 端口，全部 UDP 代表所有 UDP 端口；
  - 支持快捷协议指定，如 FTP、HTTP、HTTPS、ICMP、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 120.132.69.216/32、0.0.0.0/0。

- 优先级：安全组内规则的生效顺序，按数值作为优先级规则配置。
  - 安全组按照数值（数值越小优先级越高）依次生效，优先生效优先级高的规则；
- 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。

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弹性 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。

注：通常建议设置简洁的安全组规则，可有效减少网络故障。

VNC 登录

VNC（Virtual Network Console）是 iStack 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。

云服务器组

云服务器组（Server Group）是对虚拟机的一种逻辑划分，云服务器组中的虚拟机遵从同一策略。当前云服务器组支持反亲和性策略。

将业务涉及到的虚拟机分散部署在不同的物理服务器上，以此保证业务的高可用性和底层容灾能力。

- 反亲和性策略：同一云服务器组中的虚拟机分散地创建在不同的主机上，提高业务的可靠性。

弹性伸缩组

弹性伸缩 ASG（Auto Scaling Group）从用户的业务需求和策略出发，自动调整其弹性计算资源的管理服务。可预先配置相关的伸缩策略来保证计算能力，使业务需求上升时自动增加虚拟机实例，业务需求下降时自动减少虚拟机实例，降低人为反复调整资源以应对业务变化和高峰压力的工作量，保障业务平稳健康运行，帮助用户节约资源和人力成本。

根据用户预设的伸缩策略，自动调整其弹性计算资源，无需人工干预，避免因人为的手动操作而可能引入的低错。

- 弹性扩张时：
  - 自动创建指定数量、指定规格配置的虚拟机实例，确保伸缩内所有实例的计算能力能满足业务上升需求；
  - 如果伸缩组关联了负载均衡，自动为创建的虚拟机实例关联负载均衡，访问流量将通过负载均衡监听器自动分发到伸缩组内所有的虚拟机实例。
- 弹性收缩时：
  - 自动移出指定数量、指定规格配置的虚拟机实例，确保冗余的资源及时得到释放；
  - 如果伸缩组关联了负载均衡，自动为移出的虚拟机实例消关联负载均衡。负载均衡不再给该虚拟机实例分发访问请求。

云网络

私有网络

产品概述

私有网络（VPC —Virtual Private Cloud）是一个属于用户的、逻辑隔离的二层网络广播域环境。在一个私有网络内，用户可以构建并管理多个三层网络，即子网（Subnet），包括网络拓扑、IP 网段、IP 地址、虚拟网关等虚拟资源作为租户虚拟机业务的网络通信载体。

iStack 通过软件定义网络（SDN）对传统数据中心物理网络进行虚拟化，采用 OVS 作为虚拟交换机，Geneve 隧道作为 OverLay 网络隔离手段，通过三层协议封装二层协议，用于定义虚拟私有网络 VPC 及不同虚拟机 IP 地址之间数据包的封装和转发。

私有网络 VPC 是虚拟化网络的核心，为云平台虚拟机提供内网服务，包括网络广播域、子网（IP 网段）、IP 地址等，是所有 NVF 虚拟网络功能的基础。

功能特性

私有网络是子网的容器，不同私有网络之间是绝对隔离的，保证网络的隔离性和安全性。可将虚拟机、负载均衡、NAT 网关等虚拟资源加入至私有网络的子网中，提供类似传统数据中心交换机的功能，支持自定义规划网络，并通过安全组对虚拟资源 VPC 间的流量进行安全防护。

VPC 网络具有数据中心属性，每个 VPC 私有网络仅属于一个数据中心，数据中心间资源和网络完全隔离，资源默认内网不通。租户内和租户间 VPC 网络默认不通，从不同维度保证租户网络和资源的隔离性。

弹性 IP

产品概述

弹性 IP（Elastic IP Address，简称 EIP），是平台为用户的虚拟机、NAT 网关、负载均衡等虚拟资源提供的弹性 IP 地址，为虚拟资源提供平台 VPC 网络外的网络访问能力，如互联网或 IDC 数据中心物理网络，同时外部网络也可通过 EIP 地址直接访问平台 VPC 网络内的虚拟资源。

EIP 资源支持独立申请和拥有，用户可通过控制台或 API 申请 IP 网段资源池中的 IP 地址，并将 EIP 绑定至虚拟机、 NAT 网关、负载均衡上，为业务提供外网服务通道。

功能特性

EIP 为浮动 IP ，可随故障虚拟机恢复漂移至健康节点，继续为虚拟机或其它虚拟资源提供外网访问服务。
当一台虚拟机所在的物理主机发生故障时，调度系统会自动对故障主机上的虚拟机进行宕机迁移操作，即故障虚拟机会在其它健康的主
机上重新拉起并提供正常业务服务。若虚拟机已绑定外网 IP，调度系统会同时将外网 IP 地址及相关流表信息一起漂移至虚拟迁移后所在
的物理主机，并保证网络通信可达。

支持平台管理员自定义外网 IP 资源池，即自定义外网 IP 网段，并支持配置网段的路由策略。租户申请网段的外网 IP 绑定至虚拟资源后，下发目的路由地址的流量自动以绑定的外网 IP 为网络出口。
EIP 具有弹性绑定的特性，支持随时绑定至虚拟机、NAT 网关、负载均衡等虚拟机资源，并可随时解绑绑定至其它资源。
提供外网 IP 网段获取服务，支持租户手动指定 IP 地址申请 EIP，并提供 IP 地址冲突检测，方便用户业务网络地址规划。

弹性 IP 具有集群属性，仅支持绑定相同集群的虚拟资源。用户可通过平台自定义申请 EIP，

并对 EIP 进行绑定、解绑、调整带宽等相关操作。

NAT 网关

产品概述

NAT 网关（NAT Gateway）是一种类似 NAT 网络地址转换协议的 VPC 网关，为云平台资源提供 SNAT 和 DNAT 代理，支持互联网或物理网地址转换能力。平台 NAT 网关服务通过的 SNAT 和 DNAT 规则分别实现 VPC 内虚拟资源的 SNAT 转发和 DNAT 端口映射功能。

- SNAT 规则：通过 SNAT 规则实现 VPC 级、子网级及虚拟资源实例级的 SNAT 能力，使不同维度的资源通过 NAT 网关访问外网；
- DNAT 规则：通过 DNAT 规则，可配置基于 TCP 和 UDP 两种协议的端口转发，将 VPC 内的云资源内网端口映射到 NAT 网关所绑定的弹性 IP，对互联网或 IDC 数据中心网络提供服务。

作为一个虚拟网关设备，需要绑定弹性 IP 作为 NAT 网关的 SNAT 规则出口及 DNAT 规则的入口。NAT 网关具有集群（数据中心）属性，仅支持相同数据中心下同 VPC 虚拟资源的 SNAT 和 DNAT 转发服务，

虚拟机通过 NAT 网关可访问的网络取决于绑定的弹性 IP 所属网段在物理网络上的配置，若所绑定的弹性 IP 可通向互联网，则虚拟机可通过 NAT 网关访问互联网；若所绑定的弹性 IP 可通向 IDC 数据中心的物理网络，则虚拟机通过 NAT 网关访问 IDC 数据中心的物理网络。

功能特性

NAT 网关支持绑定多个弹性 IP 地址，使 SNAT 规则中的资源可通过多个弹性 IP 地址访问外网，DNAT 端口转发规则中的虚拟资源，可通过指定的弹性 IP 地址访问 VPC 内网服务。

支持用户查看已绑定至 NAT 网关的所有弹性 IP 地址，同时支持对弹性 IP 地址的解绑，解绑后相关联的 SNAT 规则和 DNAT 规则网络通信都将失效。用户可通过修改 SNAT 和 DNAT 规则，分别设置新的出口 IP 及入口源 IP 地址。

SNAT 规则

NAT 网关通过 SNAT 规则支持 SNAT（Source Network Address Translation 源地址转换）能力，每条规则由源地址和目标地址组成，即将源地址转换为目标地址进行网络访问。平台 SNAT 规则支持子网出外网场景，即 NAT 网关所属 VPC 下被指定子网中的所有虚拟机可通过 NAT 网关访问外网。

规则的目标地址为 NAT 网关绑定的弹性 IP 地址，将源地址子网的 IP 地址转换为网关绑定的弹性 IP 进行网络通信，即通过 SNAT 规则虚拟机可在不绑定弹性 IP 的情况下与平台外网进行通信，如访问 IDC 数据中心网络或互联网。

用户配置 SNAT 规则后，NAT 网关会自动下发默认路由至源地址匹配的虚拟机，使虚拟机通过 SNAT 规则的弹性 IP 访问外网。

DNAT 规则

NAT 网关支持 DNAT（Destination Network Address Translation 目的地址转换），也称为端口转发或端口映射，即将弹性 IP 地址转换为 VPC 子网的 IP 地址提供网络服务。

- 支持 TCP 和 UDP 两种协议的端口转发，支持对端口转发规则进行生命周期管理；
- 支持批量进行多端口转发规则配置，即支持映射端口段，如 1024—1030 。

NAT 网关绑定弹性 IP 时，端口转发规则为 VPC 子网内的虚拟机提供互联网外网服务，可通过外网访问子网内的虚拟机服务。

负载均衡

产品概述

负载均衡（ Load Balance ）是由多台服务器以对称的方式组成一个服务器集合，每台服务器都具有等价的地位，均可单独对外提供服务而无须其它服务器的辅助。平台负载均衡服务是基于 TCP/UDP/HTTP/HTTPS 协议将网络访问流量在多台虚拟机间自动分配的控制服务，类似于传统物理网络的硬件负载均衡器。

通过平台负载均衡服务提供的虚拟服务地址，将相同数据中心、相同 VPC 网络的虚拟机添加至负载均衡转发后端，并将加入的虚拟机构建为一个高性能、高可用、高可靠的应用服务器池，根据负载均衡的转发规则，将来自客户端的请求均衡分发给服务器池中最优的虚拟机进行处理。

支持内外网两种访问入口类型，分别提供 VPC 内网和 EIP 外网的负载访问分发，适应多种网络架构及高并发的负载应用场景。提供四层和七层协议的转发能力及多种负载均衡算法，支持会话保及健康检查等特性，可自动隔离异常状态虚拟机，同时提供 SSL Offloading 及 SSL 证书管理能力，有效提高整体业务的可用性及服务能力。

当前 LB 为接入的虚拟机服务池提供基于 NAT 代理的请求分发方式，在 NAT 代理模式下，所有业务的请求和返回数据都必须经过 LB ，类似 LVS 的 NAT 工作模式。

用户也可将负载均衡服务分配的 IP 地址与自有域名绑定在一起，通过域名访问后端应用服务。

功能特性

平台负载均衡服务提供四层和七层转发能力，支持内网和外网两种网络入口，在多种负载调度算法基础之上支持健康检查、会话保持、连接空闲超时、内容转发及 SSL Offloading 和 SSL 证书管理等功能，保证后端应用服务的可用性和可靠性。

- 支持内网和外网两种类型负载均衡器，满足 VPC 内网、IDC 数据中心及互联网服务负载均衡应用场景。

- 提供四层和七层业务负载分发能力，支持基于 TCP、UDP、HTTP 及 HTTPS 协议的监听及请求转发。

- 支持加权轮询、最小连接数和基于源地址的的负载调度算法，满足不同场景的流量负载业务。
  - 加权轮询：基于权重的轮询调度，负载均衡器接收到新的访问请求后，根据用户指定的权重，按照权重概率分发流量至各后端虚拟机，进行业务处理；
  - 最小连接数：基于后端服务器最小连接数进行调度，负载均衡器接收到新的访问请求后，会实时统计后端服务器池的连接数，选择连接数最低的虚拟机建立新的连接并进行业务处理；
  - 源地址：基于客户端源 IP 地址的调度策略，采用哈希算法将来源于相同 IP 地址的访问请求均转发至一台后端虚拟机进行处理。

- 提供会话保持功能，在会话生命周期内，保证同一个客户端的请求转发至同一台后端服务节点上。四层和七层分别采用不同的方式进行会话保持 。
  - 针对 UDP 协议，基于 IP 地址保证会话保持，将来自同一 IP 地址的访问请求转发到同一台后端虚拟机进行处理，支持关闭会话 UDP 协议的会话保持；
  - 针对 HTTP 和 HTTPS 协议，提供 Cookie 植入的方式进行会话保持，支持自动生成 KEY 和自定义 KEY。自动生成 key 是由平台自动生成 Key 进行植入，自定义 Key 是由用户自定义 Key 进行植入。

- 健康检查：支持端口检查和 HTTP 检查，根据规则对后端业务服务器进行业务健康检查，可自动检测并隔离服务不可用的虚拟机，待虚拟机业务恢复正常后，会将虚拟机重新加入至后端组并分发流量至虚拟机。
  - 端口检查：针对四层和七层负载均衡，支持按 IP 地址 + 端口的的方式探测后端服务节点的健康状况，及时剔除不健康的节点；
  - HTTP 检查：针对七层负载均衡，支持按 URL 路径和请求 HOST 头中携带的域名进行健康检查，筛选健康节点。

- 内容转发：针对七层 HTTP 和 HTTPS 协议的负载均衡，支持基于域名和 URL 路径的流量分发及健康检查能力，可将请求按照域名及路径转发至不同的后端服务节点，提供更加精准的业务负载均衡功能。

- SSL 证书：针对 HTTPS 协议，提供统一的证书管理服务和 SSL Offloading 能力，并支持 HTTPS 证书的单向和双向认证。SSL 证书部署至负载均衡，仅在负载均衡上进行

解密认证处理，无需上传证书到后端业务服务器，降低后端服务器的性能开销。

- TCP 获取客户端真实 IP：TCP 监听器采用 Nginx 官方的 Proxy—Protocol 方案。
  - 使 LB TCP 监听收到客户端的请求后，在转发请求至后端服务节点时，将客户端的源 IP 地址封装在 TCP 请求数据包中，发送给后端服务节点，使服务端通过解析 TCP 数据包后即可获取客户端 IP 地址。
  - Proxy Protocol 是一种 Internet 协议，用于将连接信息从请求连接的源传送到请求连接的目的地，通过为 TCP 报文添加 Proxy Protocol 报头来获取客户端源 IP，因此需要后端服务节点做相应的适配工作，解析 Proxy Protocol 报头以获取客户端源 IP 地址。

负载均衡为用户提供业务级别的高可用方案，可以将业务应用同时部署至多个虚拟机中，通过负载均衡和 DNS 域名的方案设置流量均衡转发，实现多业务级别的流量负载均衡。当大并发流量通过负载均衡访问虚拟机业务时，可通过最小连接数、加权轮询等算法，将请求转发给后端最健壮的虚拟机进行处理，请通过负载均衡将请求结果返回给客户端，保证业务可用性和可靠性。

云专线

产品概述

云专线用于搭建用户本地数据中心与 VPC 之间高速、低时延、稳定安全的专属连接通道，充分利用云服务优势的同时，继续使用现有的 IT 设施，实现灵活一体，支撑企业业务快速上云。

虚拟网关是云专线服务的重要组成部分，需要基于和云专线直连的 VPC 下创建，实现通过虚拟网关关联用户访问的 VPC。

- 虚拟网关：物理连接的接入路由器，是实现物理连接访问 VPC 的逻辑接入网关。

功能特性

用户通过云专线可以在办公网络打通租户化网络(VPC)，使得用户办公网络和 VPC 内的资源的网络互通，一个 vpc 支持多个专线网关接入，形成灵活可伸缩的混合云部署。

对等连接

产品概述

对等连接是指两个 VPC 之间的网络连接。通过对等连接用户可以使用私有 IP 地址在两个 VPC 之间进行通信，就像两个 VPC 在同一个网络中一样。用户可以在自己的 VPC 之间创建对等连接，也可以在自己的 VPC 与其他帐户的 VPC 之间创建连接。VPC 可位于不同的工作区内。

功能特性

用户根据业务需求可以在同一工作区内的两个 VPC 之间建立对等连接，也可以与其他工作区相同的 VPC 创建对等连接。同一工作内的 VPC 之间创建对等连接，无需手动接受，默认为自动接受。跨工作区之间创建对等连接，支持对端工作区同意接受或拒绝接受，同意后将会建立对等连接。

配置两端路由是两个 VPC 建立对等连接后互相通信的前提条件。同一工作区内对等连接需在本工作区下添加本端路由规则和对端路由规则。不同工作区对等连接需分别在本工作区下添加本端路由规则，在对端工作区下添加对端路由规则。

企业交换机

产品概述

企业交换机（Enterprise Switch，简称 ES）可以基于云专线、VPN 等实现虚拟私有云（Virtual Private Cloud，VPC）大二层互联等增强网络转发能力，实现企业私有 IP 不变业务无感知迁移上云，业务可选择部署在云上或云下，具备高可靠容灾能力。

功能特性

企业交换机基于云专线、VPN 建立云上与云下之间的二层网络，解决云上和云下网络二层互通问题。 允许在不改变子网、IP 规划的前提下将数据中心或私有云主机业务部分迁移上云。

云下 IDC子网和云上 VPC子网网段可以重叠，灵活的二层三层业务互访，支持访问云上高级服务，简化云上网络规划。网络迁移粒度由“子网”变为“虚拟机”，同时支持同一个子网跨云上和云下互通，云上及云下之间的大二层隧道，私有 IP 不变业务无感知迁移上云，实现无缝上云。

企业交换机创建完成后，建立本端二层连接子网和远端 VXLAN 交换机之间的二层网络通信。二层连接子网是云上 VPC 与云下 IDC 准备建立二层互通的子网，包括本端二层连接子网和远端二层连接子网。

- 本端二层连接子网：VPC 的子网，该子网需要和 IDC子网建立二层网络通信。
- 远端二层连接子网：IDC 的子网，该子网需要和 VPC子网建立二层网络通信。

隧道子网基于云专线或者 VPN 实现三层网络通信，包括本端隧道子网和远端隧道子网。企业交换机需要基于隧道子网之间的三层网络，为需要互通的云上和云下子网提供二层连接通道。

- 本端隧道子网：VPC 的子网，该子网需要与 IDC子网建立三层网络通信。
- 远端隧道子网：IDC 的子网，该子网需要与 VPC子网建立三层网络通信。

- 隧道ID：即云下IDC连接企业交换机所需要的VXLAN隧道号，即VXLAN网络标识号（VNI），是VXLAN隧道的标识，用于区分不同的VXLAN隧道。对于同一个VXLAN隧道，云下IDC和云上隧道号一致，即本端和远端隧道号一致。
- 隧道端口：云下IDC连接企业交换机所需要的VXLAN隧道端口号。
- 隧道IP：企业交换机需要和云下IDC建立VXLAN隧道实现二层网络通信，VXLAN隧道两端各需要一个IP，隧道IP 属于远端隧道子网。
- 本端IP：企业交换机需要和云下IDC建立VXLAN隧道实现二层网络通信，VXLAN隧道两端各需要一个IP，本端IP 属于本端隧道子网，限制在 198.19.0.0/16子网中。

- 使用限制：
  - 已被企业交换机二层连接绑定的VPC子网，不能再被其他二层连接或者企业交换机使用；
  - 企业交换机建立二层通信网络时，依赖隧道子网之间的三层网络，因此使用企业交换机前，请确保 已通过VPN或者云专线打通本端和远端隧道子网的三层网络；
  - 企业交换机建立二层网络通信时，需要和IDC侧建立VXLAN隧道，IDC侧交换机必须支持VXLAN功能；
  - 一个二层连接可以连通一对本端和远端二层连接子网，一个企业交换机支持建立1个二层连接；
  - 通过二层连接连通本端二层连接子网和企业交换机时，需要占用本端二层连接子网中的1个IP地 址，用作接口IP。这个IP地址不能被本端资源占用，也不能与远端二层连接子网内的其他IP地址冲突。

云存储

云盘

产品概述

云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云硬盘进行扩容，基于网络分布式访问，为云主机提供高安全、高可靠、高性能及可扩展的数据磁盘。

功能特性

云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云硬盘资源及全生命周期管理，包括云硬盘的创建、绑定、解绑、快照及删除等管理。

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

- 支持云硬盘创建、挂载、卸载、删除、克隆等生命周期管理，单块云硬盘同时仅能挂载一台虚拟机；

- 云硬盘最小支持 10GiB 的容量，步长为 1GiB，可自定义控制单块云硬盘的最大容量；
- 云硬盘具有独立的生命周期，可自由绑定至任意虚拟机或数据库服务，解绑后可重新挂载至其它虚拟机；
- 支持对云硬盘进行快照备份，包括虚拟机的系统盘快照及弹性云盘快照，并可从快照回滚数据至云硬盘，用于数据恢复和还原场景。

共享云盘

产品概述

共享云硬盘是一种支持多个云服务器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，多个云服务器可同时访问一个共享云硬盘。

根据是否支持高级的 SCSI 命令来划分磁盘模式，分为 RBD(虚拟块存储设备 ，Virtual Block Device)类型和 SCSI（小型计算机系统接口，Small Computer System Interface）类型。

- RBD 类型：RBD 类型的磁盘只支持简单的 SCSI 读写命令。
- SCSI 类型：SCSI 类型的磁盘支持 SCSI 指令透传，允许云服务器操作系统直接访问底层存储介质。除了简单的 SCSI 读写命令，SCSI 类型的磁盘还可以支持更高级的 SCSI 命令。

功能特性

共享云硬盘本质是将同一块云硬盘挂载给多个云服务器使用，类似于将一块物理硬盘挂载给多台物理服务器，每一台服务器均可以对该硬盘任意区域的数据进行读取和写入。如果这些服务器之间没有相互约定读写数据的规则，比如读写次序和读写意义，将会导致这些服务器读写数据时相互干扰或者出现其他不可预知的错误。

共享云硬盘为云服务器提供共享访问的块存储设备，但其本身并不具备集群管理能力，因此需要自行部署集群系统来管理共享云硬盘，如企业应用中常见的 Windows MSCS 集群、Linux RHCS 集群、Veritas VCS 集群和 CFS 集群应用等。

如果用户将共享云硬盘挂载到多个云服务器，首先请根据不同的应用选择不同的磁盘模式，包括 RBD 和 SCSI。SCSI 类型的共享云硬盘支持 SCSI 锁，但是需要在云服务器系统中安装驱动并保证镜像在兼容性列表中。

裸金属

裸金属

产品概述

裸金属云服务器兼具虚拟机的灵活弹性和物理机高稳定、强劲的计算性能，能与 istack 全产品（如网络、数据库等）无缝融合，在大数据、高性能计算、云游戏等领域都有广泛应用。裸金属云服务器可以在极短时间为您构建云端独享的高性能、安全隔离的物理服务器集群，是极致性能追求者的最佳选择。

专享裸金属产品提供能够接入特定 vpc，以及挂载好共享存储的裸金属商品给客户使用。

功能特性

创建裸金属云服务器时，用户指定的实例类型决定了实例的主机硬件配置。每个实例类型提供不同的计算、内存和存储功能。用户可基于需要部署运行的应用规模，选择一种适当的实例类型。这些实例由 CPU、内存、存储、异构硬件和网络带宽组成不同的组合，可灵活地为应用程序选择适当的资源。

纳管裸金属时，支持丰富多样类型的规格和实例，适用于中大型数据库系统、缓存、搜索集群，高网络包收发场景，如视频弹幕、直播、游戏等，视频编解码、视频渲染等对单核性能敏感的应用。

中间件

Etcd 元数据存储

产品概述

etcd 是一个分布式键值对存储，设计用来可靠而快速的保存关键数据并提供访问。基于 Raft 协议，通过复制日志文件的方式来保证数据的强一致性。

客户端应用写一个 key 时，首先会存储到 etcd Leader 上，然后再通过 Raft 协议复制到 etcd 集群的所有成员中，以此维护各成员（节点）状态的一致性与实现可靠性。

虽然 etcd 是一个强一致性的系统，但也支持从非 Leader 节点读取数据以提高性能，而且写操作仍然需要 Leader 支持，所以当发生网络分时，写操作仍可能失败。

etcd 具有一定的容错能力，假设集群中共有 N 个节点，即便集群中(n−1) /2 个节点发生了故障，只要剩下的(n+1) /2 个节点达成一致， 也能操作成功，因此，它能够有效地应对网络分区和机器故障带来的数据丢失风险。

etcd 默认数据一更新就落盘持久化，数据持久化存储使用 WAL（write ahead log） ，预写式日志。

格式 WAL 记录了数据变化的全过程，在 etcd 中所有数据在提交之前都要先写入 WAL 中；etcd Snapshot （快照）文件则存储了某一时刻 etcd 的所有数据，默认设置为每 10 000 条记录做一次快照，经过快照后 WAL 文件即可删除。

功能特性

etcd 的核心优势是使用简洁的方式实现 Raft 协议。

- 简单：支持 RESTful 风格的 HTTP+JSON 的 API，v3 版本增加了对 gRPC 的支持，同时也提供 rest gateway 进行转化。Go 语言编写，跨平台，部署和维护简单 ，使用 Raft 算法保证强一致性，Raft 算法可理解性好。
- 安全：支持 TLS 客户端安全认证。
- 可靠：使用 Raft 算法充分保证了分布式系统数据的强一致性 etcd 集群是一个分布式系统，由多个节点相互通信构成整体的对外服务，每个节点都存储了完整的数据，并且通过 Raft 协议保证了每个节点维护的数据都是一致的。

Minio 对象存储

产品概述

MinIO 是在 GNU Affero 通用公共许可证 v3.0 下发布的高性能对象存储。 它是与 Amazon S3 云存储服务兼容的 API。 使用 MinIO 为机器学习、分析和应用程序数据工作负载构建高性能基础架构。

其设计的主要目标是作为私有云对象存储的标准方案。主要用于存储海量的图片，视频，文档等。非常适合于存储大容量非结构化的数据，例如图片、视频、日志文件、备份数据和容器/虚拟机镜像等，而一个对象文件可以是任意大小，从几 kb 到最大 5T 不等。

Minlo 所有读写操作都严格遵守 read-after-write 一致性模型。

功能特性

在对象存储里，元数据包括 account（用户）， bucket， bucket index 等信息。Minlo 没有独立的元数据服务器，这个和 GlusterFs 的架构设计很类似，在 Minlo 里都保存在底层的本地文件系统里。

在本地文件系统里，一个 bucket 对应本地文件系统中的一个目录。一个对象对应 bucket 目录下的一个目录(在 EC 的情况下对应多个 part 文件)。目录下保存者对象相关的数据和元数据。

My SQL 数据库

产品概述

MySQL 数据库是一款安全可靠、可弹性扩展和便于管理的关系型数据库服务。支持 MySQL 数据库主流版本引擎，提供自动备份、监控、容灾、快速扩容等高级特性，满足您的数据库需求。

功能特性

平台提供单机版本、集群版本、高可用版本三种规格类型，能满足高并发场景下的读写请求，以及快速数据访问与弹性扩容的业务场景。My SQL 数据库实例创建成功后，支持用户绑定弹性 IP，在公共网络访问数据库实例。对于已绑定弹性 IP 的实例，需解绑后，才可重新绑定其他弹性公网 IP。用户还可根据自身需要通过绑定 4 层负载均衡、NAT 网关访问数据库。

单机版只包含一个节点，没有备用节点，因此当发生故障时，恢复时间较长。单机部署，适用于对可用性要求不高的场景，例如：适用于测试、个人学习等场景。

高可用版采用一主一从的双机架构，适用于连续性、安全性要求非常高等多种场景。主节点故障时，主从节点秒级完成切换，整个切换过程对应用透明；从节点故障时，自动新建从节点以保障高可用性。

企业版采用一主两从架构，是企业生产环境数据库的最优选择。可实现集群节点自动故障恢复，手动主从切换。所有备选节点均可参与容灾切换，保证了系统的高可用稳定性。

除系统默认创建的 root 帐号外，您可根据业务的需要创建其他的业务帐号并且支持为账号配置访问权限。

My SQL 数据库提供库表级操作、实时监控、SQL 窗口、数据管理为一体的数据库管理服务。支持数据库的库表增删、以及查看表数据、表结构等数据库操作。提供数据库状态信息、数据库连接以及流量相关的多维度监控。

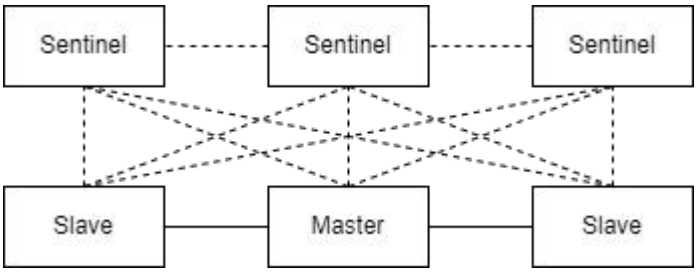
My SQL 数据库为用户提供 SQL 编译器，支持常用 SQL 模板以及自定义 SQL 保存，支持 SQL 语句执行结果的展示。

Redis 数据库

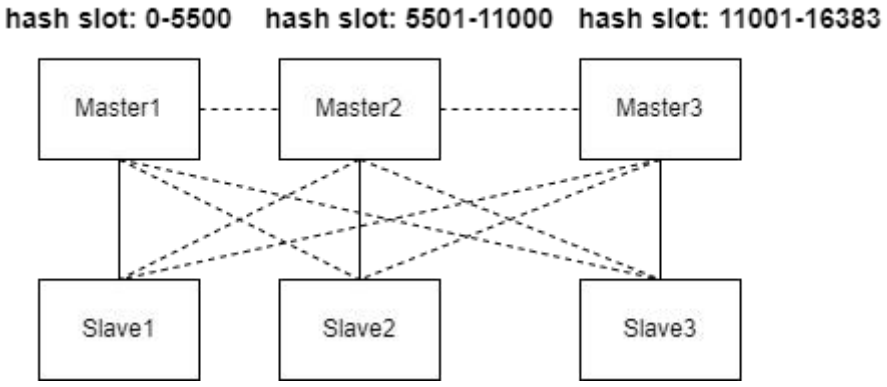
产品概述

Redis 数据库是兼容开源 Redis 协议标准的缓存数据库服务，基于 sentinel（哨兵）主从架构和集群架构，支持单机、主从、集群等多种规格类型，具备高可用、高可靠、高弹性等特征，可满足业务在缓存、存储、计算等多种不同场景中的需求。

Redis 架构分为 sentinel（哨兵）主从架构和集群架构，如下图：



上图是 sentinel 主从架构，其中实线代表数据同步，虚线代表心跳连接。在这种架构下，每个 sentinel 都会监控所有节点，当超过一定数量的 sentinel 发现 master 挂了后，会将其中一个 slave 选作新的 master，实现主备切换。数据的读写都需要在 master 上执行，因此 redis 支持 sentinel 模式，从 sentinel 获取 master 的地址和端口。



sentinel 架构中，写数据请求都需要落在 master 节点上，并且集群中只有一个 master，当 tps 量级很大时，master 就成为了瓶颈，因此产生了集群架构，如上图所示。这种架构的思想就是在集群中生成 16384 个 slot，并且将 slot 分配到 master 上，这种情况下 master 也称作分片。当需要写数据时，会将 key 作 CRC16 运算，再余 16384，决定将数据写到 master 上面，这样就可以解决 sentinel 中的问题。同时，需要给每个 master 配置 slave 或者说副本，避免 master 挂了后无法提供服务。

功能特性

平台提供单机版本、集群版本、高可用版本三种规格类型，能满足高并发场景下的读写请求，以及快速数据访问与弹性扩容的业务场景。

支持将当前时间点的实例缓存数据进行备份，以便在缓存实例发生异常后能够使用备份数据进行恢复，保障业务正常运行。同样支持将实例数据在同一个工作区内迁移至另一实例。

Redis 数据库还提供可视化 Web 管理界面，支持在线完成实例重启、重置密码、参数修改等操作。用户也可通过监控，观察实例资源运行状态。

资源编排

资源编排是一种云资源的自动化编排服务，通过使用 Yaml 格式的模板或者可视化工具描述多个复杂的云计算资源的配置以及依赖关系，自动完成云资源的创建和部署。

使用资源编排服务的流程大概如下：用户需要创建一个资源模板，模版内容用于描述需要创建的云资源配置以及资源间的依赖关系、引用关系等，用户可以根据需求自行定义资源模版。资源编排服务将根据模板来创建和配置这些云资源。例如创建虚拟机，用户只需要创建编写资源模板定义虚拟机、虚拟私有云和子网，并定义虚拟机与虚拟私有云、子网之间的依赖关系，子网与虚拟私有云的依赖关系，然后通过资源编排使用该模板创建资源栈，虚拟私有云、子网和虚拟机就创建成功了。

- 资源模板：体现为 YAML 格式的文本文件，模板用于创建资源栈，是描述资源基础设施和架构的前提。
- 资源栈：资源栈是针对云资源的集合。资源栈将一组云资源作为一个整体来进行创建等操作。

资源编排支持编排平台所有云资源。通过创建资源栈，用户可大批量创建不同规格的云资源，快速完成资源的统一编排，提高工作效率。编排语言支持 YAML 语法来定义需要的元素。资源编排支持用户上传已有的资源模板，还可通过 web 界面进行更改参数，控制要部署资源对象的规格等，并且支持将私有模板共享为公共模板，平台所有用户均可使用，从而实现模板的重复利用。初次之外，图形化编辑模板界面让新手用户更容易快速上手定义所需的资源模板。

资源中心

资源中心通过将云资源进行分组的形式为用户提供一系列管理云资源的服务，主要包括资源组、资源实例、权限管理等操作功能。用户可以使用资源组将云资源划分，分别管理云上资源，使用权限管理决定工作区成员是否拥有管理、查看云上资源的权限。

资源组（Resource Group）是用户基于工作区维度进行资源分组管理的一种机制，资源组能够帮助用户解决单个工作区的资源分组和授权管理的复杂性问题。

- 对单个工作区下多种云资源进行集中的分组管理，组内所有操作将写入操作日志中。
- 可为每个资源组设置一个或多个管理员，资源组管理员可以独立管理资源组内的所有资源。
- 访客角色角色仅支持查看资源组内的资源，不具备操作管理权限。
- 未得到任何授权的用户将无法查看、管理资源组内的资源。
- 资源组内支持有操作管理权限的用户将其资源进行组别变更。
- 资源组内支持用户申请调整管理组内实例个数，管理员可选择接受审批或驳回审批，并且可查看审批记录，以便后续进行回溯。