UCLOUD优刻得

中国第一家公有云科创版上市公司

股票代码：688158

UCloudStor 5.1

统一分布式存储产品白皮书

文档更新：2023 年 7 月 17 日

下一代云基础设施

筑基数字化底座

版权信息

注意

您购买的产品、服务或特性等应受优刻得科技股份有限公司商业合同和条款约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用权利范围之内。除非合同另有约定，优刻得科技股份有限公司对本文档内容不做任何明示或暗示的声明或保证。

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，不涉足客戶业务领域。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戶提供云服务支持，间接服务终端用戶数量达到数亿人。UCloud 深耕用戶需求，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。

随着云计算、大数据、人工智能、5G 网络等新技术快速发展，数字化时代已然到来，各种新业务、新应用层出不穷，而数据量也呈指数级增长。传统的集中式存储架构难以处理大规模的数据存储和处理需求。

UCloudStor 统一分布式存储以“数据”为导向，从数据存储、数据管理的维度，提供跨站点数据同步、多集群管理、数据冗余、故障自愈、卷备份等企业级存储服务，为企业客户提供可靠、稳定、统一的数字化底座基石，助力企业数字化转型。

1 产品简介

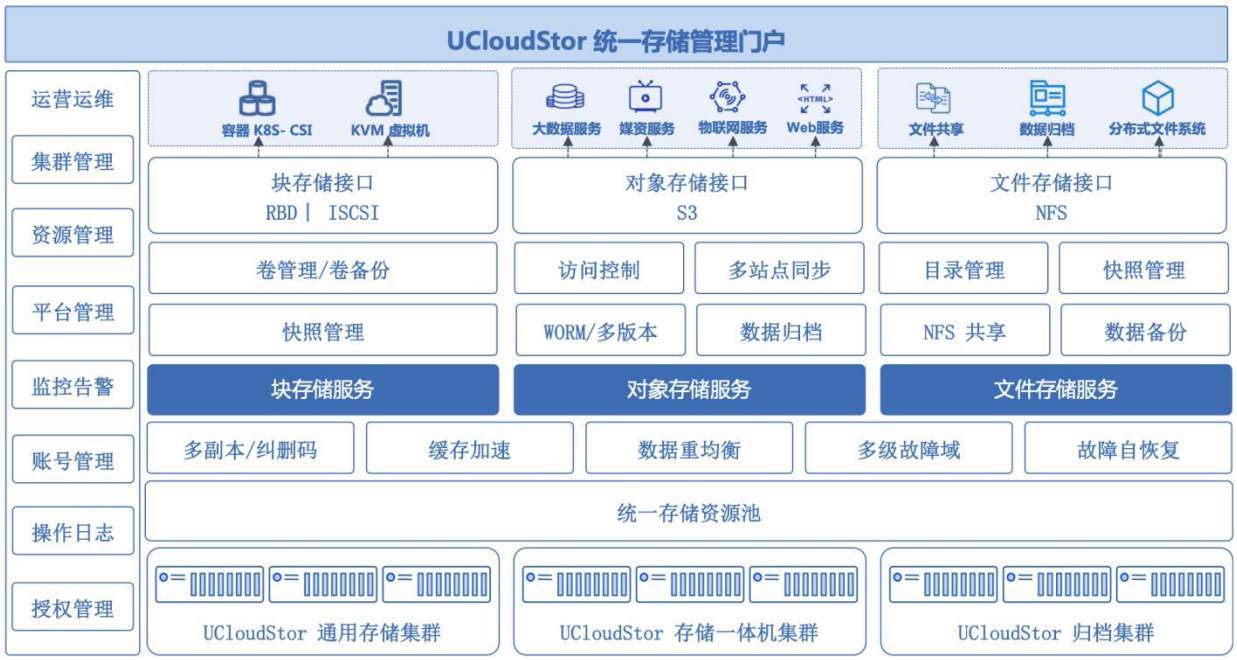
1.1 产品概述

在海量非结构化数据不断增长和业务场景不断变化的情况下，哪种存储系统可以满足当前与未来的需求？如何打破信息壁垒和“孤岛”，构建统一高效、互联互通、安全可靠的数据资源体系？如何为大数据业务提供统一的存储架构？如何进行数据资源共享?

UCloud 推出自主研发的 UCloudStor 统一分布式存储产品，它是企业级、软件定义、融合统一的数据存储平台，为虚拟化/云平台、政务归档、企业办公、广电媒资库、医疗/金融影像、数据中心灾备等场景，提供高性能块存储、文件存储和大容量对象存储一站式存储方案和服务。

1.2 产品架构

UCloudStor 统一分布式存储基于多年公有云存储的技术积累，在 Ceph 开源基础上自研扩展，向上层提供标准的协议接口及管理门户，提供多种数据存储服务及解决方案。



产品由基础硬件设施、统一资源池，智能存储层、核心存储服务、运营运维管理及统一管理门户组成，为上层应用服务及运营运维人员提供全面的存储服务及管理服务。

基础硬件设施，支持通用 X86 架构硬件服务器，软硬解耦，不限制品牌和型号，并支持利旧现有通用 x86 服务器硬件资源。针对信创场景，支持鲲鹏、飞腾、海光、申威等国产芯片服务器。

统一资源池，具体负责数据的实际存放与管理，承载平台 OS 内核模块及软件定义存储引擎 UStore 的实现逻辑，同时基于智能存储层提供三种核心存储服务。

智能存储层，为平台数据存储服务提供智能的可靠性和安全性保证，包括多副本及纠删码冗余策略、缓存加速、数据动态重均衡、多级故障域、故障自恢复及数据校验增强等能力。

核心存储服务，平台为上层应用提供三种类型的核心存储服务，包括块存储服务、对象存储服务和文件存储服务，包括卷管理、快照备份、对象多站点同步、NFS 共享文件等，充分满足用户对各类数据的存储和使用需求。

运维运营管理，提供图像化的资源管理、集群管理、平台管理、账号管理、监控告警等运维运营能力，满足客户对存储系统的运维运营述求。

统一管理门户，UCloudStor 提供 WEB 控制台、API 接口两种方式接入和管理存储系统，可针对块、文件及对象存储进行统一管理、监控和运维。

1.3 核心优势

- **统一存储，多种接口**

多种数据服务接口，UCloudStor 在一套存储系统中提供多种应用接口，提供块、对象、文件存储服务，把应用与数据连接起来。块存储服务支持 RBD、iSCSI 协议，支持 K8S CSI 接入；对象存储服务支持 S3 接口，无缝对接 S3 完善的生态系统，可以用于通用云存储服务，存放文件、图片、视频等非结构化数据，也可以用于备份和归档，支持主流备份软件；文件存储提供 NFS 协议，支持文件目录、文件快照等能力，适应网盘、AI 训练等存储文件的场景。

- **全分布式全冗余架构，多站点数据保护**

全分布式架构，UCloudStor 所有软硬件全冗余，无单点故障，具有超高可用性。

数据多副本，UCloudStor 支持存储池粒度的不同副本数量存储策略，副本数支持 2~6，生产环境建议采用三副本模式。2N+1 节点数量下可允许 N 个服务器的不可用。

数据 EC 纠删码，UCloudStor 也支持 4+2、6+3、8+4 等多种纠删码策略，以便获得更多有效存储容量。

UCloudStor 支持多站点容灾备份，可以把对象存储的数据异步复制到远端 UCloudStor。

- **按需在线扩容，最大支持 EB 级容量**

UCloudStor 配置灵活，可以最少从 3 台服务器开始部署，快速构建统一存储系统，承载业务运行。

按需扩容，性能和容量都可以横向扩展，而且扩容时业务不中断，可以按阶段购买软件和硬件进行扩容，降低成本。

- **高性能方案，智能化运维**

UCloudStor 可以充分利用硬件优势，比如 Intel 至强多核 CPU、SSD、10Gb/40Gb 网络等，可以提供百万级别的 IOPS 和极低的响应延迟。

UCloudStor 具有故障检测和自动恢复功能，数据恢复不需要人工介入，在数据恢复期间，数据访问正常，服务不中断。UCloudStor 可以实现数据并行恢复，多块硬盘同时进行数据恢复，极大的降低了数据恢复时间，提高了数据的可靠性。

1.4 技术架构特性

- **高可靠、高可用**

全分布式架构，所有的节点和磁盘均可独立提供存储服务，所有软硬件全冗余设计，无单点故障，具有超高可用性。并可在不影响业务系统的情况下无缝升级平台系统。

- **高扩展性**

支持灵活的扩容方式，提供按需扩容能力，性能和容量均可在线横向扩展，可以独立扩容存储节点、硬盘、网络节点，或者同时进行扩容。

- **高度兼容开放**

提供标准的 API 接口和 SDK，客户可为上层应用或业务灵活定制开发，深度集成。

不绑定硬件及品牌，兼容 X86 和主流国产 CPU 架构，可兼容通用服务器、网络设备及现有利旧硬件资源。

- **自主可控**

基于开源，经过 5 年多的技术积累和深度开发，产品代码自有率高达 97%以上（国家测评中心认证），降低被控和盗窃风险，保障其客户的自主权和利益。

1.5 客户痛点

数字经济时代，随着新场景、新应用不断涌现，数据量呈爆发性增长，数据类型也更加丰富，这对数据存储带来新的挑战和机会，而传统的存储架构，已经不能满足用户数字化转型升级的需求：

- 传统 SAN 存储方案，需要专业存储和网络设备，硬件和后期维护的成本高；
- 传统（双控）控制器架构，一旦软硬件增加扩容，存储性能将随之急剧降低；
- 传统存储厂商软硬件封闭，兼容性差，不利于客户长期投资，损害客户利益；
- NAS 文件存储在亿级文件、目录层次深场景下，受架构性能限制，无法满足业务应用；
- NAS 文件存储受限于控制器和设备容量，扩展性比较差；

面对海量非结构化数据存储场景（PB 级容量、亿级文件数及以上），SAN 和 NAS 更是显得“力不从心”。

1.6 典型应用场景

- **替换传统 NAS 设备系统**

针对客户已有 NAS 数据存储，业务市场后期快速扩张，数据量爆发增长，传统 NAS 文件受限于自身架构，无法满足 PB 级规模的文件存储需求，UStor 提供替换数据迁移方案，一是传统 NAS 文件数据迁移至对象存储，业务中断时间分钟级（老数据平滑迁移）；另外传统 NAS 文件迁移至分布式文件存储。通过灵活专业的数据迁移方案和服务，满足客户日益增长业务数据存储的同时，降低替换风险。

- **海量非结构化存储**

在政府政务、视频监控、企业办公、电子签章等场景下，非结构化数据量巨大，保存周期长，且存在备份归档等诉求，传统存储难以满足该类海量数据的存储需求。针对这类场景，UCloudStor 统一分布式存储提供对象存储方案，可实现弹性平滑扩容。

- **灾备中心场景**

金融、能源、医疗等行业的业务数据保护有着严格标准要求，如双录系统、电力系统、PACS 系统等，UCloudStor 统一分布式存储实现了跨地域、跨集群的对象数据实时同步，保障业务数据安全和连续性。

- **数据归档**

以非结构化数据为主，一次写入，长期保存，数据量只增不减的业务场景，如金融行业，受监管要求其保存周期通常数十年；互联网类业务，如电子签章（签名）、电子发票等非结构化数据，需要保存 3 年以上；政府政务，绝大部分电子档案需保存 20 年到 70 年，重要档案需要永久保存；医疗行业，住院病历数据保存不得少于 30 年；保管保存医疗影像数据时间要求不少于 15 年，UCloudStor 统一分布式存储提供数据全生命周期管理，可按照“时间”维度，对数据进行自动归档，提升数据处理效率。

1.7 交付模式

- **纯软件交付**

客户自身提供通用的 x86 或国产芯片服务器，可以利旧已有服务器硬件资源，只需采购软件平台，授权相应软件功能模块即可。

- **软硬一体交付**

软硬件都有 UCloud 提供，交付到客户侧直接开箱即用，无需单独采购硬件设备，整体缩短业务上线周期，降低总体 TCO 成本。

2 平台物理架构

2.1 物理集群节点

2.1.1 融合节点

融合节点指的是一种将多个不同功能、不同角色能力的节点集成在一起的架构，以实现更高效的资源利用和更简化的管理。在分布式存储系统中，融合节点通常指的是具有管理、网关、存储功能的节点，它们可以同时提供存储、网关等服务。

融合节点是一种非常重要的架构形态。通过将不同功能、角色的节点合并为一个节点，可以减少节点数量，从而简化系统管理和维护工作，提高资源利用率和性能。此外，融合节点可以提供更高效的数据传输和处理能力，从而提高整个 UStor 系统的性能和可靠性。

融合节点是一种架构，一种产品形态，它可以提高资源利用率和性能，简化系统管理和维护工作，并提高整个分布式存储系统的可靠性和可伸缩性。

2.1.2 独立管理节点

管理节点承载 UCloudStor 存储系统核心管理服务及文件存储元数据服务，负责存储平台的集群管理、集群状态管理、文件存储元数据服务、管理控制台及 API 网关等服务模块的运行，提供标准 API 和 WEB 控制台两种接入管理方式。

管理服务仅负责数据调度、集群管理及上层业务的接入流量，存储业务扩展不受管理节点数量限制。生产环境管理节点 3 台起（节点数量需为奇数），保证管理服务各功能模块的高可用。

集群状态管理使用 Paxos 算法保证集群状态一致，为保证服务的个数为奇数，需在 3 个管理节点均进行部署，保证管理服务的高可用并防止存储集群脑裂。同时由于 Paxos 算法选举机制要求必须有半数以上成员存活才可对外提供服务，因此同一时间仅允许 1 个管理节点宕机。

管理服务比较轻量且对资源消耗较少，通常小规模存储集群可直接融合部署至 3 个存储节点，并可在后续进行横向扩展，提高存储系统交付效率，降低总体建设成本。

2.1.3 独立网关节点

网关节点涉及块 ISCSI 协议、文件 NFS 协议和对象 S3 协议服务，即可单独部署，也可部署在相同节点。网关节点至少 2 台，采用 keepalived+vip 互为主备模式，保证高可用。

块存储网关，标准的 iSCSI 协议和接口，通过映射和挂载卷，为上层应用提供块存储。

文件存储网关，NFS 文件网关，为上层应用提供标准的 NFS 文件共享协议和接口，并提供访问共享目录的访问入口。

对象存储网关，为上层应用提供对象存储 S3 接口的 HTTP 服务，接收和解析上层应用 HTTP 客户端请求；同时作为存储集群的对象存储客户端，为对象应用提供数据存储通道，将 HTTP 服务传入的对象数据写入集群，并负责对象数据的认证和访问控制。

2.1.4 独立存储节点

存储节点是集群内实际存储数据的节点，承载存储服务 OSD 的运行，负责数据的实际存放与管理，并处理多副本和纠删码数据复制、数据恢复、数据回滚、故障检测恢复、重均衡等功能，同时负责向管理服务报告检测信息。每个 OSD 可作为独立组件提供存储服务，存储客户端可直接与 OSD 磁盘进行数据读写，将中心化任务分摊到所有磁盘完成，使得管理服务和客户端更加轻量，支持大规模扩展。

通常一个 OSD 存储服务可对应一块磁盘，一个集群至少需要 2 个 OSD 存储服务；同时存储服务支持节点、机柜及数据中心级故障域，可通过策略配置使数据均衡分布于不同的节点、机柜或数据中心，默认采用节点级故障域，即将冗余策略的数据平均分布至集群的不同节点磁盘上。

生产环境至少需要 3 个存储节点，结合多副本和纠删码策略保证数据的安全性，单集群支持横向扩展至 200 个存储节点。每个存储节点支持混合 SSD、SATA、SAS、NVME 等多类型磁盘，存储节点采用 SSD+HDD 缓存方案，配置标准：SSD 与 HDD 盘数比 1:5，容量比 1:20。

将所有存储节点上相同类型的磁盘划分为一个存储池，并定义存储池的故障域和冗余策略，用于不同的存储场景。相同介质的磁盘构建存储池，建议使用容量、类型一致的磁盘，避免因不一致引发集群数据不均衡等问题，而导致无法正常存储数据。

在大规模场景下（如 EB 级），可将对象和文件存储的元数据、索引数据单独部署，存放至 SSD 介质，以提高读写性能。

2.1.5 推荐方案

UCloudStor 统一存储物理节点方案会根据业务需求及应用场景进行调整，可采用 3 个管理节点+N（N≧3）个存储节点进行集群规划和部署，后续可根据业务规模水平扩展存储节点，或独立扩展对象存储和文件存储网关节点。

最佳实践中，推荐将管理服务和存储服务融合部署至 N 台（N≧3）存储节点服务器，搭建 UCloudStor 统一存储平台，即最小生产规模为 3 台服务器。

•每个节点均部署存储服务 OSD，若有 SSD 加速需求，可将所有或部分 SSD 磁盘构建为对象文件元数据存储池，提升对象和文件存储性能。

•核心管理服务中的 Mon 均衡部署于 3 台节点，Mgr 集群管理均衡部署于 2 台节点，控制台管理服务均衡部署于 2 台节点。

•对象网关服务可均衡部署于 2 台节点，做主备高可用

•文件网关服务可均衡部署于 2 台节点，做主备高可用。

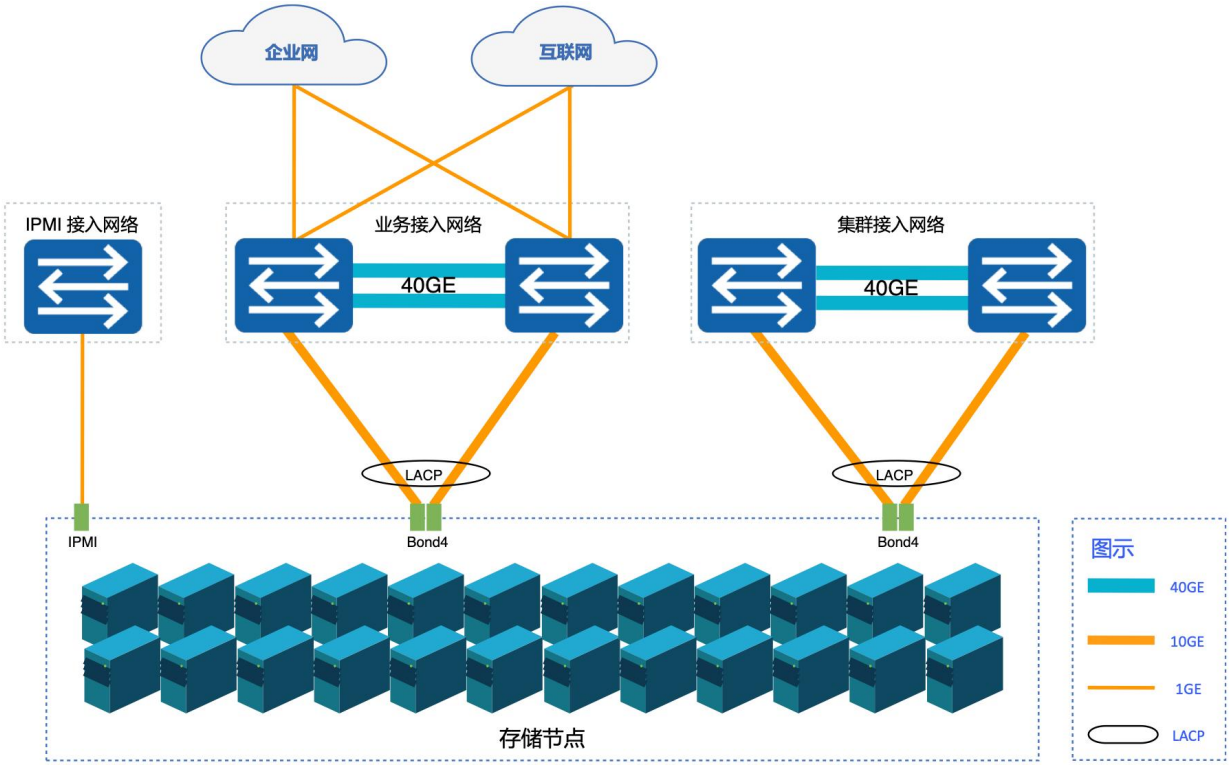
•文件元数据服务可均衡部署于 3 台节点，做主备高可用。

在大规模存储集群案例中，可根据业务需求及场景将存储节点拆分为多个集群，通常每个存储集群裸容量建议不超过 5PB，也可根据存储存储场景进行集群拆分，如拆分为文件存储集群和对象存储集群等。

2.2 物理网络架构

2.2.1 标准网络架构

为构建高可用、高可靠、高安全的统一分布式存储平台，UCloudStor 标准网络架构均采用高可用冗余性设计。本文以标准网络拓扑图为基础进行描述。在标准网络架构中至少需要 4 台万兆交换机和 N（N≧3）台存储节点服务器。若有服务器 IPMI 及网络设备带外管理需求，可根据需求增加 IPMI 接入交换机并接入网络。



整个存储系统网络设计为集群接入（ClusterNetwork）和业务接入（PublishNetwork）两张网络区域，分别承载集群内通信和对外业务通信，两张网络在网络设备层面物理隔离。

- 集群接入网络是集群内服务器间业务网络通信的网络，用于处理冗余数据复制、恢复、回填、重均衡等数据传输产生的网络流量。
- 业务接入网络是企业网或互联网应用与 UCloudStor 存储集群通信的网络，用于处理与块、对象及文件存储客户端对集群数据读写所产生的网络流量。

标准网络架构为单集群网络架构，为保证存储保群的高可用，所有交换机设备和服务器节点均采用冗余网络设计，业务接入交换机和集群接入交换机均采用两台交换机堆叠，一组堆叠交换机共可提供 96 个网络接口，堆叠检测及备用占用 3*2 个接口，可用业务端口为 90 个。每台服务器均采用两个万兆口绑定在一起，双上联至一组堆叠交换机的 2 个接口，即标准网络方案中的一组交换机可接入 45 台存储服务器。

2.2.2 网络区域

网络区域的设备通常包括业务接入交换机和集群接入交换机，若有 IPMI 和网络带外管理需求，可增加 IPMI 网络接入交换机设备。

- 业务接入交换机：采用 2 台万兆交换机堆叠作为一组业务接入交换机，用于承载 45 台存储节点业务接入。

•集群接入交换机：采用 2 台万兆交换机堆叠作为一组集群接入交换机，用于承载 45 台存储节点集群接入。

业务接入和集群接入均为大二层环境，所有网络接入均为端口聚合，保证高可用，同时通过控制接口广播报文流量，抑制网络广播风暴。

所有服务器接入交换机的接口均配置为 Trunk 模式，便于在服务器节点上透传 VLan 网络信息，以方便划分集群网络和业务网络，其中业务网络需要提供网关功能。

业务接入与互联网、数据应用、办公环境及虚拟化云平台网络的连接可支持二层聚合、三层聚合等互连模式，同时支持串联或旁挂防火墙、IDS、IPS 及防 DDOS 等安全设备。

提供文件存储和对象存储服务时，文件存储和对象存储网关通过主备方式保证高可用，暂不支持负载均衡。

提供块存储服务时，块存储客户端 RBD 需要和业务网络中每一台存储节点进行通信，以满足 RBD 分布式写数据的需求。

分布式存储系统属于网络存储，数据的复制和写入均需通过网络进行传输，在生产环境中必须采用 10GE 或 25GE 以上的网络，以保证存储的读写性能。除特殊要求外，接入交换机设备只要支持通用能力即可，如堆叠、VLan、Trunk、LACP 及 IPV6 等。

2.2.3 服务器区域

服务器区域设备通常包括存储节点和管理节点，若直接使用存储节点部署所有管理服务，可省去物理管理节点。

•存储节点【必选】：采用通用服务器作为存储节点，用于承载存储服务 OSD 的运行，负责数据的实际存放与管理，同时负责向管理服务报告检测信息。

•采用 2 个 10GE 网口分别上联到两台集群接入交换机,并做双网卡 bond,作为存储节点集群接入。

•采用 2 个 10GE 网口分别上联到两台业务接入交换机,并做双网卡 bond,作为存储节点业务接入。

•采用 1 个 1GE 网口上联至 IPMI 接入交换机，作为存储节点的 IPMI 带外管理接入。

若将管理服务同时部署于存储节点，则业务接入 Bond 网口通过划分子接口并配置 Vlan 用于区分业务接入网络、管理网络、对象文件网关网络等。

为保证分布式存储的性能，存储节点必须采用万兆以上速率的网卡作为集群接入和业务接入网口。

管理节点【可选】：UCloudStor 默认推荐使用 3 台存储节点融合部署管理服务，如需独立服务器承载并运行管理服务，可采用通用服务器作为存储平台管理节点。

•采用 2 个 10GE 网口分别上联到两台业务接入交换机，并做双网卡 bond，作为管理节点业务接入。

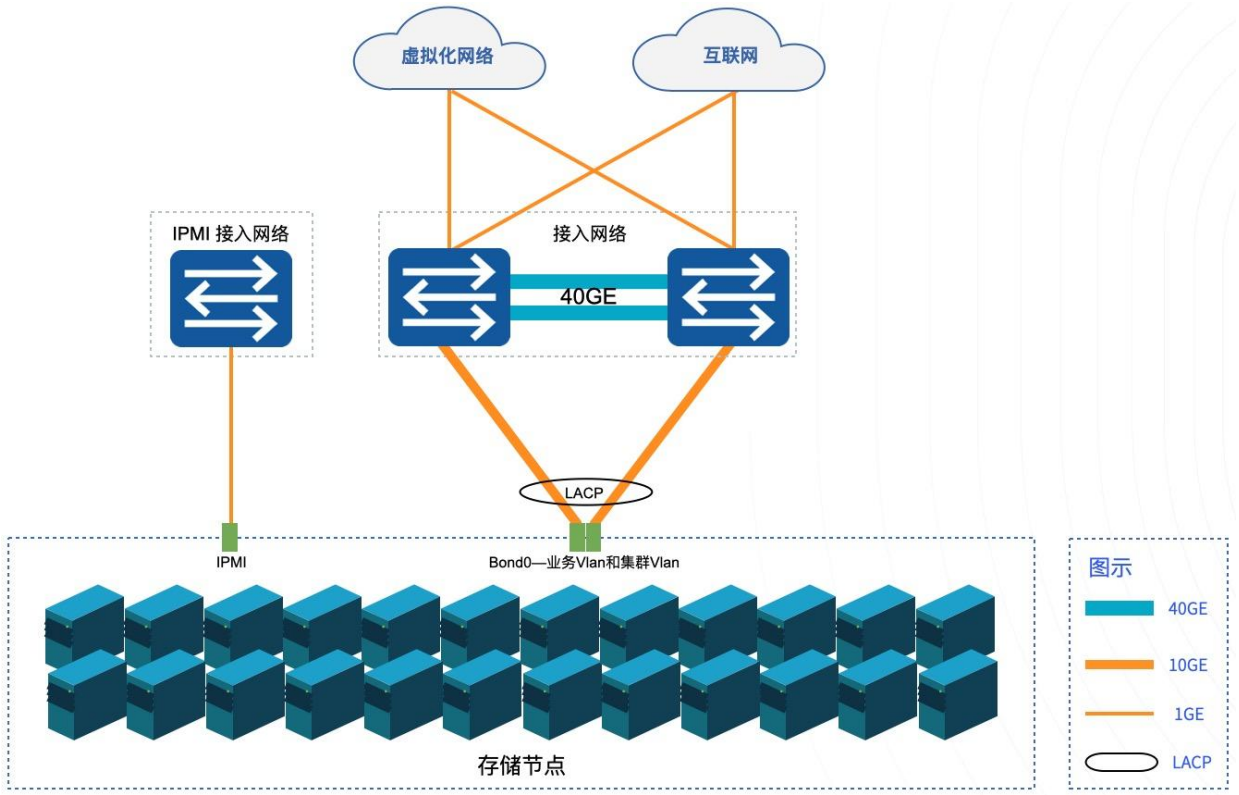
- 采用 2 个 10GE 网口分别上联到两台集群接入交换机,并做双网卡 bond,作为管理节点集群接入。
- 采用 1 个 1GE 网口上联至 IPMI 接入交换机，作为管理节点的 IPMI 带外管理接入。

以上网卡 bond 均采用“mode=4”模式，即 IEEE802.3ad 动态链路聚合。

2.2.4 扩展架构

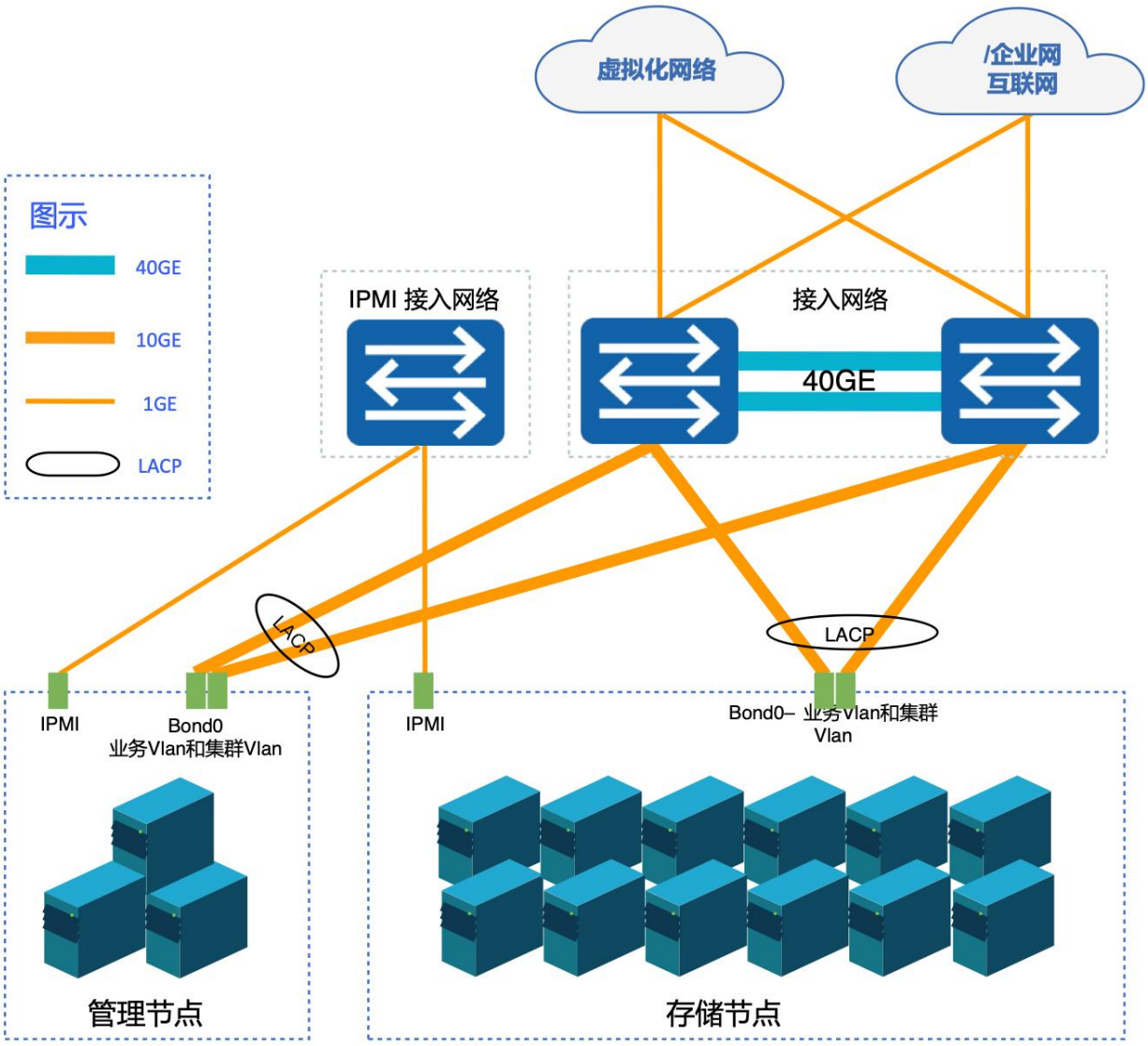
在实际项目中，根据用户需求和所提供的环境，可对标准网络架构进行调整，以适配不同规模及场景的存储需求。

2.2.4.1 业务网和集群网逻辑隔离



可采用 2 台万兆交换机堆叠，承载存储集群的所有网络流量，服务器使用 2 个接口绑定接入交换机 Trunk 接口，通过在服务器操作系统内划分 Vlan（即子接口）隔离业务网络和集群网络。如集群规模扩展至 45 台以上，需要在上层增加核心层交换设备，同时需要再增加一组接入交换机连接更多存储节点服务。

2.2.4.2 独立管理节点网络架构



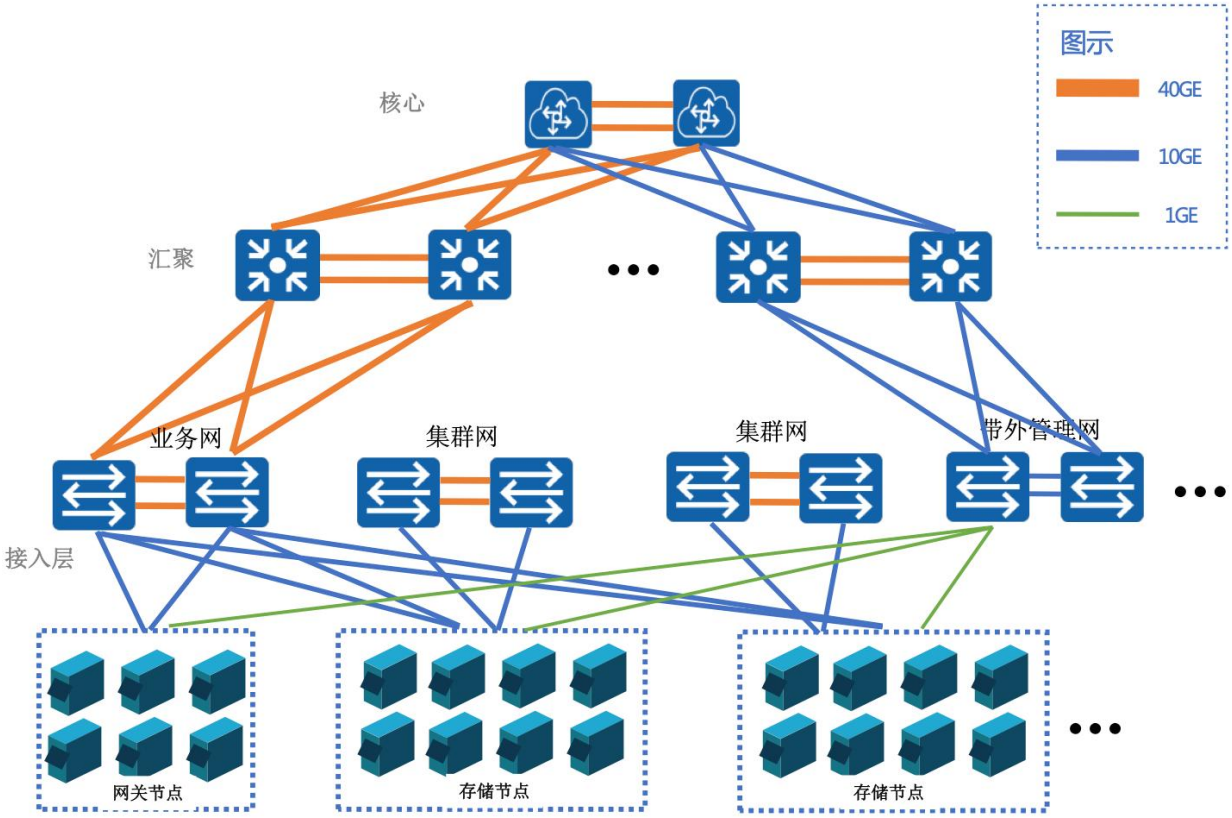
可采用 2 台万兆交换机堆叠，承载存储集群的所有网络流量，存储节点和管理节点服务器均使用 2 个接口绑定接入交换机 Trunk 接口，通过在服务器操作系统内划分 Vlan（即子接口）隔离业务网络和集群网络。管理节点同时接入业务网络和集群网络，为存储管理、业务接入、元数据读写等提供网络通信能力。

如需横向扩展对象存储网关和文件存储网关，可在管理节点处直接增加对象存储网关和文件存储网关服务器节点即可。对象存储网关和文件存储网关服务仅作网络转发，通常对服务器配置要求较低，保证网络带宽即可。

如需提供测试环境，可使用 1 台交换机通过 Vlan 进行网络划分，规划并部署测试环境。在实际环境中通常建议网络全冗余设计，同时在高安全要求的环境中，也可规划三张网络，分别承载管理网络、集群网络和业务网络，所有存储节点均增加网口直接上联至管理网络，直接与管理节点进行通信。

2.2.4.3 独立管理、网关节点网络架构

针对大容量（EB 级）规模场景，需要将管理节点、网关节点和存储节点单独分离部署。



采用传统三层架构（接入-汇聚-核心）；接入交换机和汇聚交换机采用堆叠技术，保证带宽和冗余；

服务器节点与接入交换机间做链路聚合（LACP）；

- 集群网:万兆网络，负责业务数据的落盘，主从 OSD 的数据复制，以及故障数据恢复、OSD 间数据自均衡方面。每个集群独立的内网交换机；
- 业务网:万兆网络，客户端入口；mon 与 osd 间的数据通信传输（如 map 数据等）；
- 带外管理网:对服务器资源通过网络进行集中管理（BMC、IPMP 口）；

2.3 硬件选型

2.3.1 最低配置要求

针对测试环境和生产环境最低配置要求如下：

	测试环境	生产环境	备注
CPU	不低于 8 核	不低于 10 核	
内存	单台主机不低于 32GB	单台主机不低于 64GB	
网卡	1 个 10GB 网口 (无性能要求)	2 个 10GB 网口 （做 bond4）	如需网卡级别冗余，2 张 10GB 网卡
系统盘	1 块 SSD480G	2 块 SSD480G	生产系统盘做 RAID1
数据盘	**全闪方案：** SSD 盘：至少 1 块，如果只配置 1 块，单块硬盘不低于 480G	**全闪方案：** SSD 盘：至少 1 块，如果只配置 1 块，单块硬盘不低于 960G	1、推荐三副本； 2、缓存方案，推荐 SSD 和 HDD 的盘数比为 1:5，容量比为 1:10
	缓存方案： SSD 和 HDD 的盘数比不高于 1:5，容量比不高于 1:20	**缓存方案：** SSD 和 HDD 的盘数比不高于 1:5，容量比不高于 1:20	
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	
服务器数量	三台	三台	

注：

- 最低配置只保证平台正常部署，稳定运行，不包括实际业务数据容量、副本数、性能要求；
- 生产环境，服务器硬件和架构层面必须保证冗余机制；
- 全 HDD 方案不推荐，只仅限用于冷存、归档业务；

2.3.2 推荐硬件配置

2.3.2.1 服务器推荐配置

机型	详细配置描述	备注

存储一体机	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	
存储一体机（冷存/归档）	Factor Form 4U CPU Intel® Xeon® Silver 4314 Processor(16CORES_2.4GHz_135W_X86) *2 DDR4_32GB_RDIMM_3200MH*8 OS HDD 480G_SSD_SATA3_512E_2.5"_6Gb/s *2 Data HDD SATA3_HDD_16TB *36 LSI-9311-8I 双口万兆光口网卡(不含光模块)*2 PSU=1300W*2/导轨	不作为生产存储，仅限于归档、备份冷存

2.3.2.2 网络设备推荐配置

业务类型	详细配置描述
万兆接入交换机(10G)	锐捷 RG-S6250-48XS8CQ 48*10G+2*40G+2 个扩展插槽（可再扩展 4*40G），1U，满负荷 234W
	华三 S6520X-54QC-EI 48*10G+2*40G+2 个扩展插槽（可再扩展 4*40G），1U，满负荷 234W
万兆汇聚交换机(25G)	锐捷 RG-S6510-48VS8CQ(V2.0) 48*25G+8*100G，1U，满负荷 413W
	华三 S6850-56HF 48*25G+8*100G，1U，满负荷 413W
千兆交换机（IPMI 带外管理）	锐捷 RG-S6000C-48GT4XS-E 48*1G+4*10G+1 个扩展插槽（可再扩展 8*10G 或者 2*40G），1U，满负荷 93W
	华三 5560X-54C-EI 48*1G+4*10G+1 个扩展插槽（可再扩展 8*10G 或者 2*40G），1U，满负荷 93W
5M10GAOC 线缆	5M10GAOC 线缆
10M10GAOC 线缆	10M10GAOC 线缆
40G 交换机堆叠线	40G 交换机堆叠线
5M 千兆管理网线	5M 千兆管理网线（CAT-6 网线）
10M 千兆管理网线	10M 千兆管理网线（CAT-6）

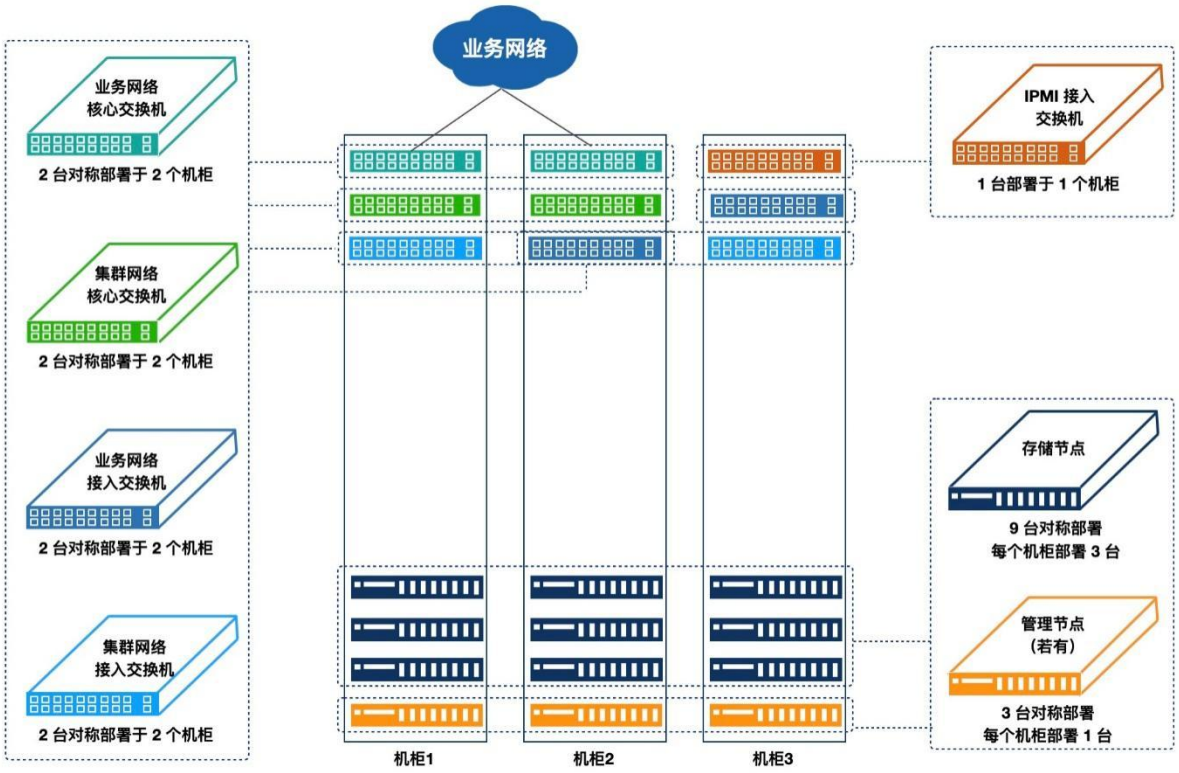
2.4 平台资源占用

平台运行本身需要占用服务器的 CPU、内存及存储资源，具体如下：

模块	角色	数量	CPU	内存 GB	存储 GB	说明
OSD 服务	存储节点	N	1C	4		一块硬盘对应一个 OSD N 块硬盘即为 N
Monitor 监视器	管理节点	固定 3 台	2C	4	60GB	
MDS 元数据	管理节点	固定 3 台	2C	2		主要保存文件存储的元数据，文件存储时使用，其他无需
网关 /ganesha	网关节点	2	1C	2		主备模式，文件存储时使用，其他无需
网关 /RGW	网关节点	2	1C	2		主备模式（vip），对象存储时使用，其他无需
网关 iSCSI	网关节点	2	1C	2		主备模式，iscsi 块存储时使用，其他无需
控制台 console	管理节点	2	4C	8	200G	包括监控及管理服务

注意：以上服务的 CPU 和内存为预估值，实际生产环境中，根据使用负载等情况会有变动。

2.5 机柜空间规划



所有设备在机柜中对称部署，结合管理节点高可用及存储节点故障域实现机柜级冗余，单机柜掉电或故障不影响云平台业务。通常一个机柜可支撑 15 个 2U 节点，根据网络架构设计一组接入交换机支撑 45 个节点，即一组接入交换机支撑 3 个机柜。

3 个机柜为 1 组，平均 1 组机柜支撑 45 个节点、1 组业务网络接入交换机、1 组集群网络接入交换机、1 台 IPMI 接入交换机。

另外两组 4 台核心交换机可根据机柜位置对称部署于多个机柜中，即每组核心分布在 2 个机柜中。

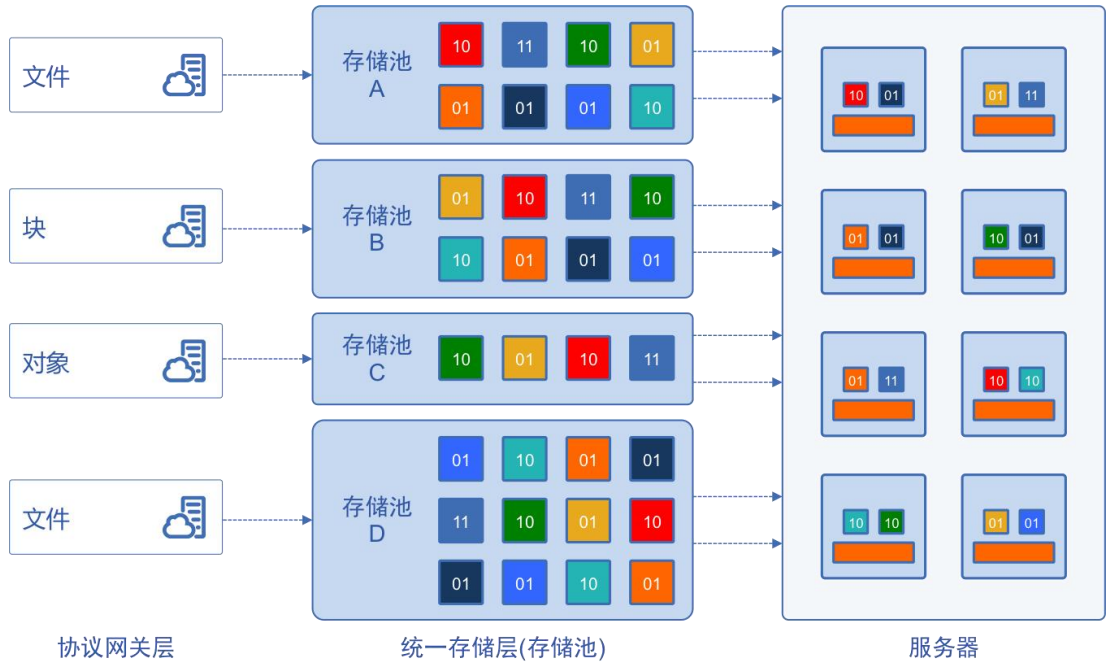
以上图为例，包括 8 台业务交换机、1 台运维管理交换机、12 台服务器设备及 3 个机柜：

- 一组业务网络核心交换机分别部署于 2 个机柜，即其中两个机柜各部署 1 台。
- 一组业务网络接入交换机分别部署于 2 个机柜，即其中两个机柜各部署 1 台。
- 一组集群网络核心交换机分别部署于 2 个机柜，即其中两个机柜各部署 1 台。
- 一组集群网络接入交换机分别部署于 2 个机柜，即其中两个机柜各部署 1 台。
- 1 台 IPMI 接入交换机部署于 1 个机柜。
- 9 台存储节点分别部署于 3 个机柜，即每个机柜各部署 3 台。
- 3 台管理节点分别部署于 3 个机柜，即每个机柜各部署 1 台。

若服务器分集群部署存储平台，建议不同集群的服务器分别部署于多个机柜中。

3 平台技术特性

3.1 分布式、全对称设计



平台采用全分布式架构，所有组件都是对等的，可以分布部署在多个存储服务器上，没有单点故障和性能瓶颈。

每块硬盘上运行着一个智能存储组件(OSD)，负责提供存储空间，并保证故障自愈和数据恢复。

块存储为直接并行访问 OSD，不需要代理和转发，具有非常高的性能。

对象存储服务可以使用硬件负载均衡，也可以使用软件负载均衡组件。对象存储网关组件(OGW)属于无状态服务，可以横向扩展部署多个 OGW，提高并发访问性能和吞吐率。

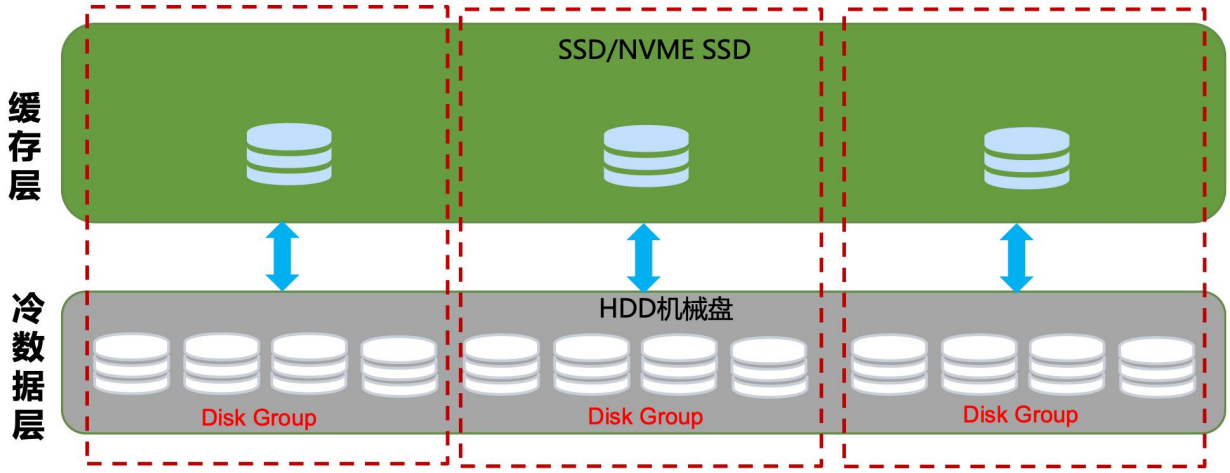
分布式存储集群管理组件(MON)管理整个集群的状态，MON 使用 Paxos 算法保证集群状态一致，因此需要保证 MON 的个数为奇数，这样才能防止集群脑裂。MON 和 OSD 之间有心跳检测，OSD 之间也会有心跳检测。当某个 OSD 发生故障时，MON 或者其他 OSD 可以检测到这个 OSD 故障，并更新集群的状态。所以存储系统使用多副本或者 EC 数据保护模式，其他 OSD 会马上接替这个 OSD 的工作，保证存储服务的高可用性。

3.2 缓存存储机制

不同客户的不同业务场景，对存储性能要求差异很大，HDD 的随机访问受到磁头寻道时间的限制，与 SSD 相比，导致随机访问性能大幅下降，如何以最优的成本提高随机读写的性能？UStor 基于 open cas，结合 ceph 架构，提出了自己的一套缓存机制方案。

把 UStor 存储的缓存模块（UCache）安装到 Linux 操作系统内核，将高性能缓存盘（SATA SSD 或 NVME SSD）和大容量存储盘（如 HDD）绑定生成新的设备分区，当对新的设备分区读写时，可利用高性能缓存盘提高读写性能，降低读写延时。

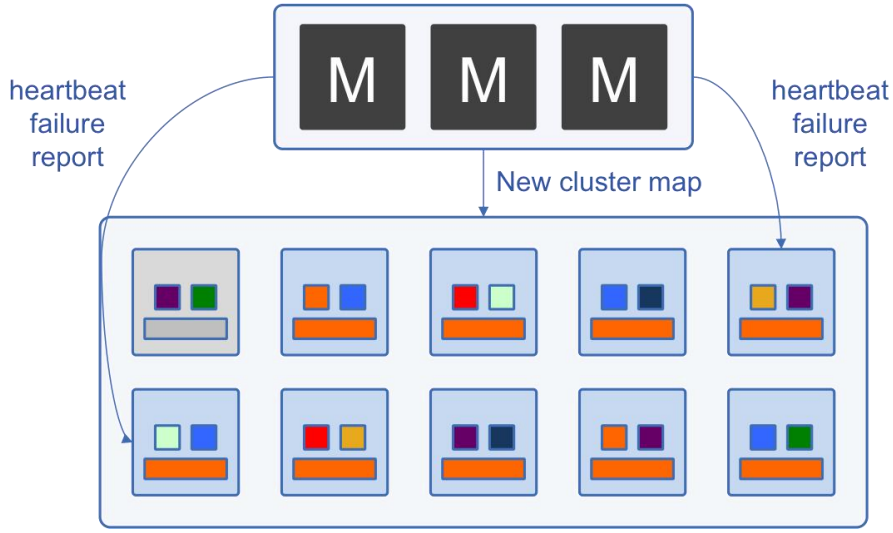
UStor 支持 SSD 智能分层技术，每块 SSD 缓存分区可对应多块 HDD 磁盘，组合为缓存磁盘组；冷热数据根据算法自动流转分别存储至固体盘和机械盘，缓存命中率越高性能越高。



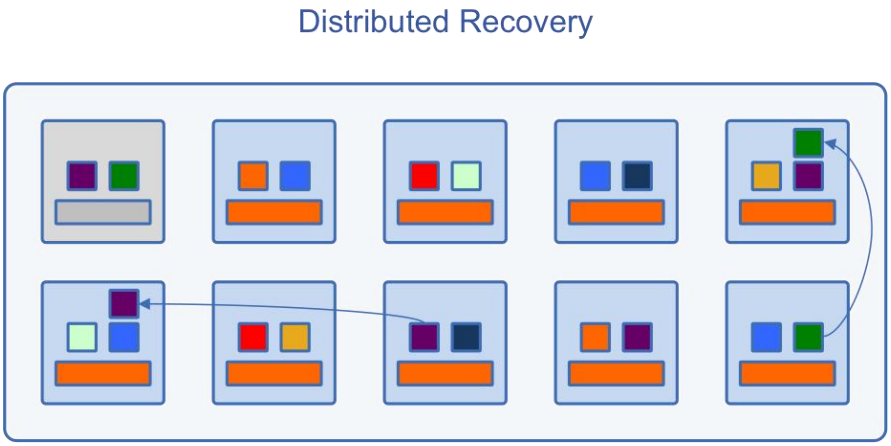
- 缓存盘支持普通 SATA SSD 和 NVME SSD 固态盘
- 标准配置：SSD 和 HDD 盘数配比 1:5;容量配比 1:20
- 各节点的缓存盘和数据盘，类型、数量和容量上需保持一致

3.3 智能故障检测与恢复

整个 UCloudStor 存储平台由数百或数千个存储着文件数据片段的服务器组成，每一个组成部分都很可能出现故障，UCloudStor 允许平台中的部分部件失效。UCloudStor 存储平台具有强大的故障检测和自动恢复能力，数据恢复不需要人工介入，在数据恢复期间，数据访问正常，服务不中断。

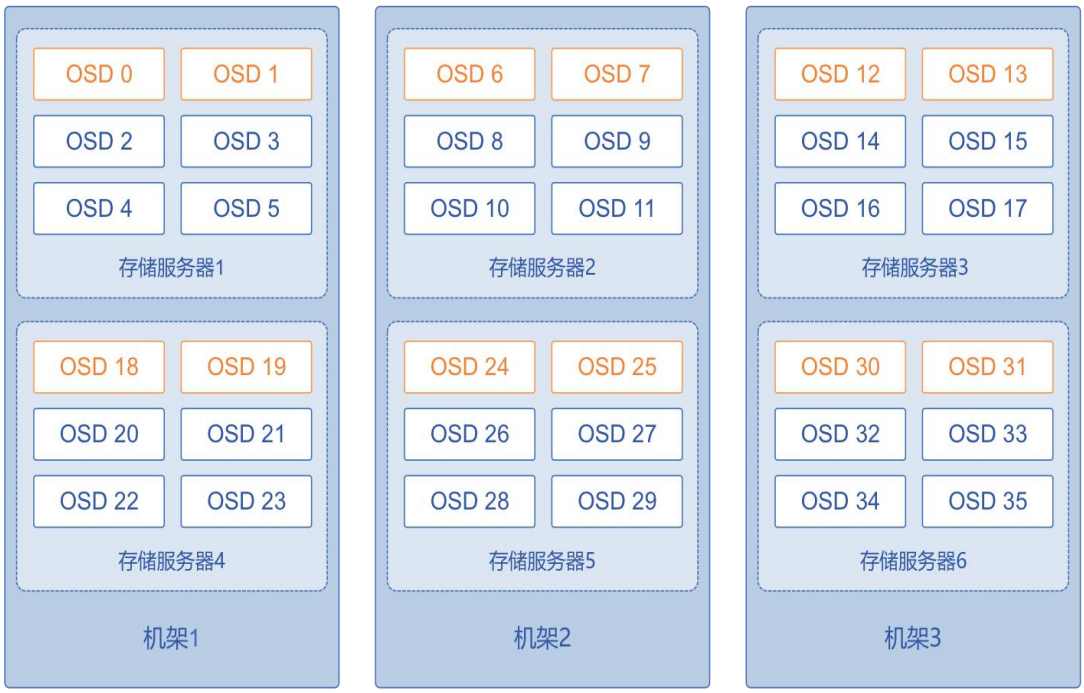


UCloudStor 存储平台可以实现数据并行恢复，多块硬盘同时进行数据恢复，极大的降低了数据恢复时间，提高了数据的可靠性。



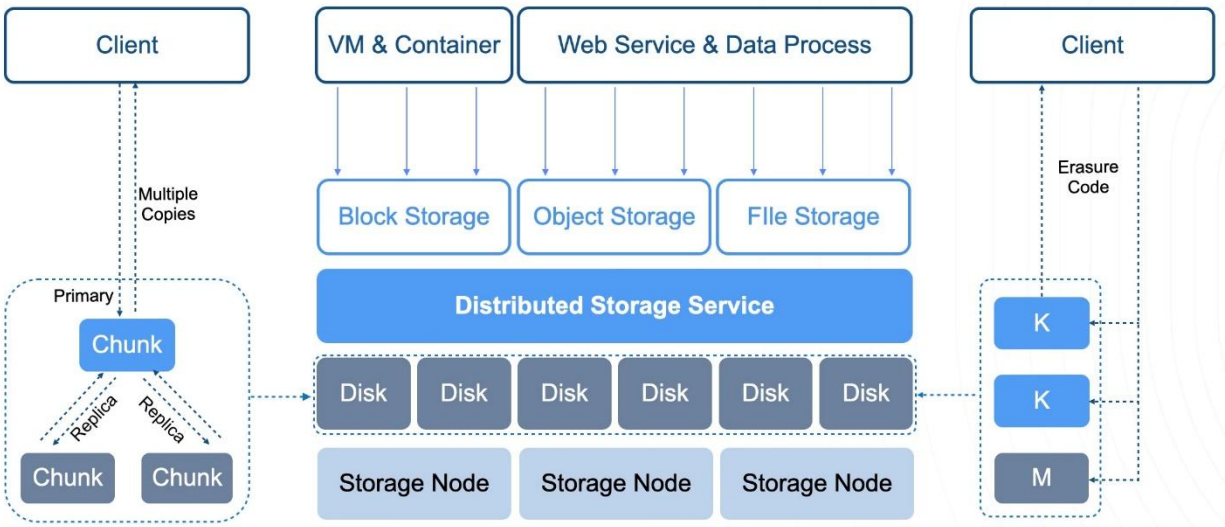
3.4 多级故障域隔离

UCloudStor 存储系统可以定义数据分布策略，自定义故障域隔离级别，让多副本数据分布在不同的节点、机架或机房上。如数据分布在多个机架服务器中，当任一机架掉电，不影响数据服务的提供，有效提高数据的可靠性。



3.5 多种数据保护模式

UCloudStor 存储平台数据保护支持多副本和纠删码，可以灵活设置存储池的数据保护模式。



UCloudStor 支持存储池粒度的不同副本数量存储策略，副本数支持 2~6（一般推荐采用三副本并且进行强一致性验证，保障每一份数据的 3 份副本），2N+1 节点情况下允许 N 个服务器损坏，系统仍能正常运行。通过系统的实时多副本技术，保证数据高可靠，可以根据用户需要设置数据副本数量和复制策略，把数据同时存在于多台服务器、多个机架之中，最大限度提高数据容灾能力。多副本存储策略具有以下优点：

- 读性能延迟低

- 写性能延迟高
- 数据恢复速度快

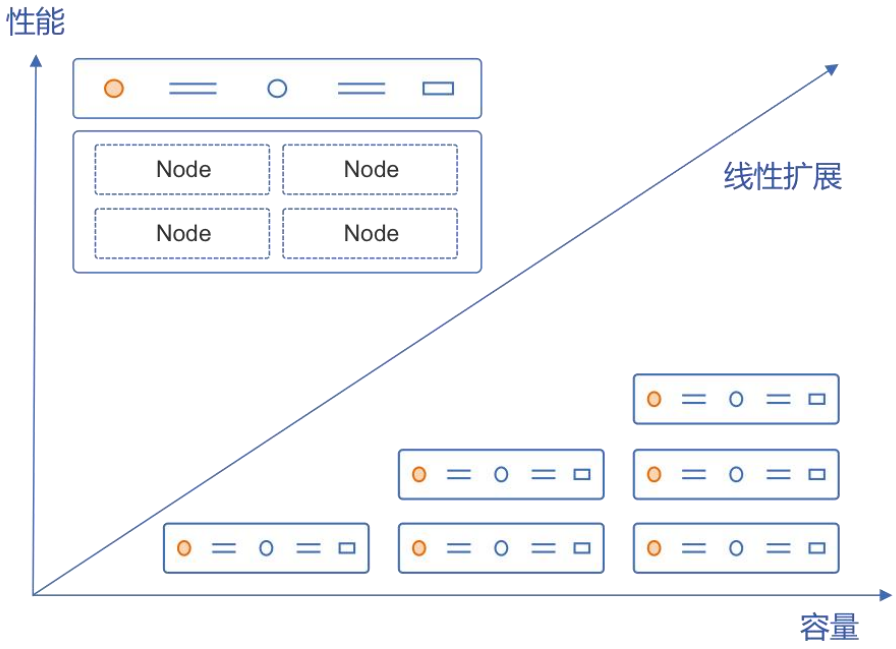
同时，UCloudStor 支持纠删码策略 4+2、8+3 及 16+4 等，对延迟不敏感的场景，包括备份、对象存储等采用纠删码存储策略可以获得更多有效存储容量。

3.6 弹性、线性横向扩展

UCloudStor 存储平台采用分布式架构设计，节点由多个独立的服务器实现(利用本地硬盘)，所有节点是完全对称架构，无主次之分，物理节点可以在不停机的情况下动态增加/删除，实现存储容量和性能的动态扩展，“对称”意味着各节点可以完全对等，能极大地降低系统维护成本，且无单点故障。

UCloudStor 存储平台支持最小规模从 3 台服务器开始部署，在业务初期可以有效控制建设成本。随着业务需求的增长，可以通过增加系统中存储服务器数量便捷地进行弹性扩展，可以一次增加单台或者多台存储服务器，理论上没有容量限制。支持 PB 级别的大规模存储，单集群可支持 200 个节点，裸容量超过 30PB。

UCloudStor 支持可预测的水平扩展，实现自动负载均衡，扩展节点后，可根据集群中各个服务器节点的负载和容量使用情况做负载均衡，以达到整个系统的负载均衡，避免单点过热的情况出现，整个扩容过程无需中断业务。



3.7 平台服务可靠性

平台模块、组件和服务做了主备或集群模式，保证平台自身的稳定、可靠运行。

- 存储监视器（Monitor）采用 3 节点集群模式；
- RGW、ISCSI 等网关服务 2 节点，采用 keepalived+vip 实现主备高可用；
- 文件存储 MDS 元数据 3 节点，采用主备模式；
- 监控告警服务（Prometheus），2 节点部署，保证高可用；
- 平台管理服务，如 api-server,controller-manager,mysql 等，2 节点部署，采用 keepalived+vip 实现高可用；

4 核心能力

4.1 块存储

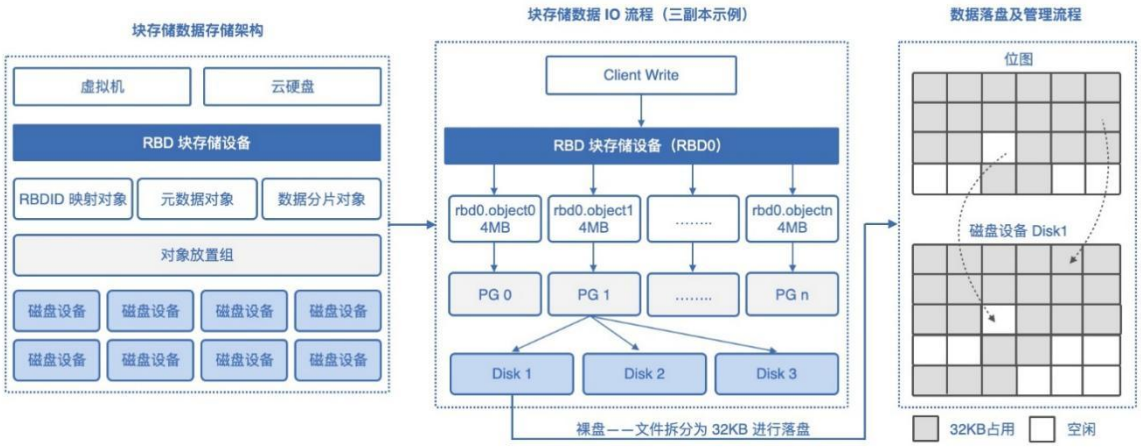
4.1.1 概述

块存储是一种数据存储方式，将数据划分为固定大小的块并以块为单位进行读写。每个块都有一个唯一的标识符，通常使用逻辑块地址(LBA)来标识。块存储通常用于存储设备，如硬盘驱动器(HDD)、固态硬盘（SSD）和网络存储设备。

块存储提供了对数据的随机访问能力，允许以任意顺序读取或写入块。这使得块存储适用于许多应用场景，包括文件系统、数据库和虚拟化环境。

UStor 块存储支持 RBD、ISCSI 协议接口，提供在线扩容、快照管理、定时备份策略等能力，适用于虚拟化、私有云、超融合、容器云等应用场景。

4.1.2 块存储机制



- **块存储数据存储架构**

虚拟机和云硬盘创建后，会在分布式存储系统中分别生成一个 RBD 块存储设备，即 KVM 引擎客户端读写数据的载体，同时针对一个块存储设备会生成 RBDID 映射对象、元数据对象及数据分片对象。

•RBDID 映射对象：指每个 RBD 块存储设备在存储系统中映射的 ID，作为全局唯一标识符，如 RBD0 对应的标识符为 RBD00001。

•元数据对象：指 RBD 块存储设备的元数据描述信息，包括块设备的创建时间、更新时间、属性、容量等。

•数据分片对象：RBD 块存储设备的数据分片对象文件，每个分片默认为 4MB，分片数量取决于 RBD 设备的大小，如 400MB 的云硬盘，分片数量即为 100 个对象文件。

所有的对象文件分别会通过算法计算对象的 PG 存储放置组，三副本模式下，一个放置组通常对应三个磁盘设备，即 RBDID 对象、元数据对象及数据分片的所有对象均会对应一个放置组，同时会将对象数据写入放置组对应的三个磁盘设备中。

- **块存储数据 IO 流程**

虚拟机和云硬盘的虚拟化客户端在写数据至 RBD 块设备时，会自动对数据进行切片操作。如上图中 RBD 块设备为 RBD0，每个分片大小为 4MB，即会自动将写入的数据切分为 4M 大小的对象文件，同时包括元数据对象及 RBDID 映射对象。每个对象文件都有一个名字，即 rdb 设备+object+序号，如 rbd0.object0。

每一个 rbd.objectn 的对象文件通过放置组进行副本位置的分配，放置组通过 Cursh 算法定位出三个磁盘设备，作为对象文件的存储位置，即数据及元数据会首先进行对象文件的拆分，并根据放置组及磁盘设备的对应关系，分别存储至存储系统中的所有磁盘中。

- **数据落盘及管理流程**

分布式存储系统使用裸盘进行磁盘管理及数据落盘操作，在进行对象文件 rbd.objectn 文件的存储和落盘时，会通过存储管理系统将每一个对象文件再次进行拆分进行存储，即通过位图的方式计算拆分后文件的在物理磁盘上的存储位置，将每个 4MB 对象文件拆分后存储至磁盘设备中，默认拆分大小为 32KB。

在写数据时根据位图计算出 32KB 文件在磁盘介质上的存储位置，同时在位置上将占用的位置标示为 1（占用），未被占用的磁盘位置标示为 0（空闲）。

整体存储数据的过程，会将文件拆成 4MB 大小的对象文件，对应至不同的磁盘设备；同时在落盘时再次将 4MB 文件拆分成 32KB 大小的块存储至磁盘设备中。

- **删除数据机制**

根据上面存储数据和落盘状况，存储在分布式存储系统中的文件被两次拆分成 32KB 的块文件，完全打散写入至整个存储集群的所有磁盘中，包括存储文件的元数据文件；在读取数据或找回数据时，需通过元数据计算数据是由哪些对象文件组成，同时需要结合磁盘位图计算对象数据中由哪些 32KB 的块数据组成，即其中一个 32KB 的数据是无法读取或恢复一个文件，必须将文件打散存储在存储集

群中所有磁盘的 32KB 数据组合为一个对象文件，再通过元数据拼接对象文件，才可读取和恢复一个文件。

在平台上删除虚拟机、云硬盘或删除虚拟机中的文件时，存储系统会将文件的元数据进行删除，同时到磁盘管理的位图中将相关的 32KB 块置为 0（仅将块置为空闲，不真正清除数据），即用于恢复数据和读取数据的元数据被清除，同时 32KB 的块被置为空闲，可以被其它数据占用和写入。

- 若 32KB 块空间被其它数据占用后，则之前的数据会被新的数据覆盖；
- 若 32KB 块空间未被其它数据占用，则可通过恢复软件找回的 32KB 数据，但 32KB 数据由于无元数据及位图，无法找出其它关联的 32KB 数据及相对应的对象文件，保证数据的安全性。

4.1.3 卷管理

卷管理是指对块存储系统中的卷（Volume）进行创建、配置、管理和监控的过程。卷是一种逻辑存储单元，可以包含一个或多个物理存储设备，并提供给操作系统和应用程序使用。

- 创建卷，支持 UStor 控制台图像化操作创建逻辑卷，设置卷名和卷大小，卷名可以重名，单个卷值设置范围 1~8000G。创建成功的卷，不能进行更新修改，可以进行在线扩容操作，没有映射启动器组的卷才可以被删除，删除后卷被存放在在平台回收站中，可以进行恢复；
- 配置卷，如上所述，可以对其卷名、卷大小进行配置；
- 扩展卷，对已有创建成功并在使用的卷，如需要更多存储空间，可以在控制台进行在线扩容操作，从而扩容卷容量，扩容过程中，对用户业务无感知，无任何影响；
- 自动精简配置，系统默认配置，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；
- 卷监控，主要是对资源使用情况进行监控，包括已用容量、读写 IOPS、读写带宽、读写延迟，支持按照时间维度（1 小时、自定义）图形化直观展示数据信息，可设置每 15s 自动刷新数据；
- 删除卷，没有映射启动器组的卷才允许删除。删除后的卷被放在回收站中，可以进行恢复或销毁。

4.1.4 卷复制备份

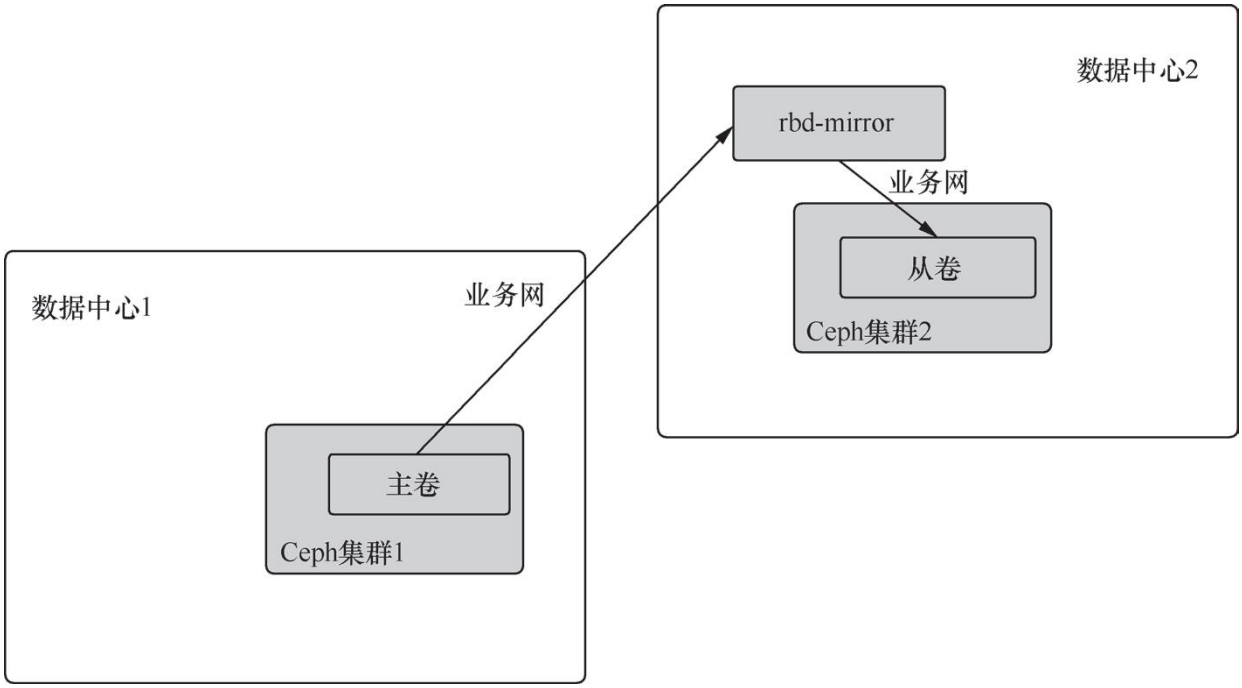
目前市场上大部分存储产品、方案主要通过快照、备份技术来构建数据的副本，在故障时通过数据的副本恢复用户数据，且 ceph 本身多副本冗余等方式，也是提升数据的安全性。但这 3 种技术都存在不足：

- 快照与源卷存放在同一个集群，在源卷误删除场景下可以恢复，但如果遇到断电、火灾、地震等整机房级别的故障，该方案就显得无能为力；

•备份虽然可以将数据复制到对象存储或者其他备份存储系统中，但是备份和恢复的时间可能要数分钟到十几小时，无法保证业务的连续性；

•而多副本方式采用的是强一致性同步模型，所有副本都必须完成写入操作才算一次用户数据 I/O 写入成功，这导致了 Ceph 存储集群在跨域部署时性能表现欠佳，因为如果副本在异地，网络时延就会增大，拖垮整个集群的 I/O 写入性能。

对数据安全性和业务连续性要求高的场景，UStor 存储提供支持异地容灾功能，基于 snapshot 模式的卷的异步复制，使用定期计划或手动创建 RBD 快照，然后基于快照将主集群的数据异步复制到备份集群的策略。借助 RBD 的 fast-diff 功能，无须扫描整个 RBD 卷即可快速确定更新的数据块。备份集群根据两边快照之间的数据或元数据差异，将增量数据快速同步到本地。



•只支持单向同步，当数据从主集群备份到备用集群的时候，rbd-mirror 进程仅在备份群集上运行。

•支持开启、设置自动数据同步策略，最小粒度可达分钟级（RPO），减少对卷性能的影响。

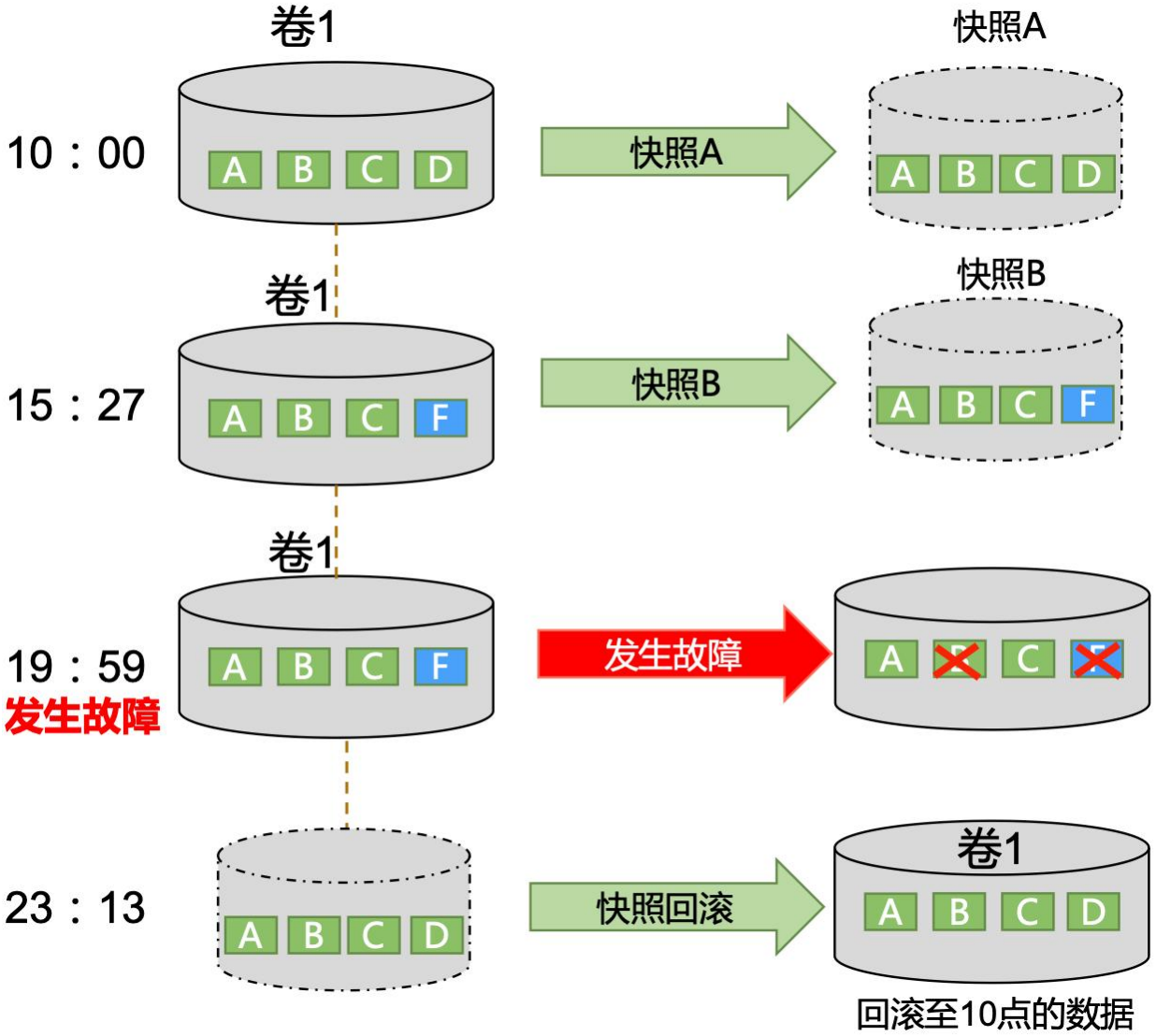
•目前 UStor 支持两种模式的异步复制，分别为 Pool 模式和 Image 模式，如果配置为 Pool 模式，那么存储池中所有 RBD 卷都会备份到远端集群；如果配置为 Image 模式，表示只会对特定的某些卷开启异步复制功能。

•当设置为 pool 模式，备集群需要创建和主集群名称一样的池。

•当主集群出现故障，需要人工手动进行主备切换（将主池或卷降级 non-primary，备集群卷升级为 primary），并将升级为主的卷挂载给客户端，方可正常使用。

4.1.5 快照管理

快照管理是指对块存储中的快照进行创建、管理和使用的过程。快照是存储系统中某个特定时间点上的数据副本，可以用于数据保护、恢复、测试和其他用途。



UStor 提供秒级快照机制，将用户的逻辑卷数据在某个时间点的状态保存下来，作为备份恢复数据之用。单卷最大快照个数达到 256 个。支持手动一次性快照和自动快照策略两种方式。

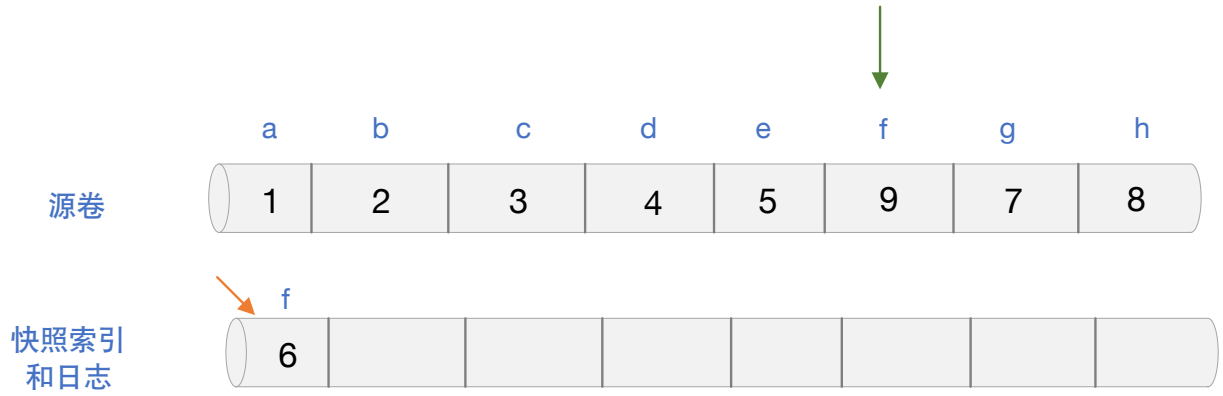
快照前写数据

9 写入块 f，直接覆盖源数据

快照后写数据

地址块 f 数据 6 写入日志，并记录原地址位 f

9 写入源卷位置块 f 的地方



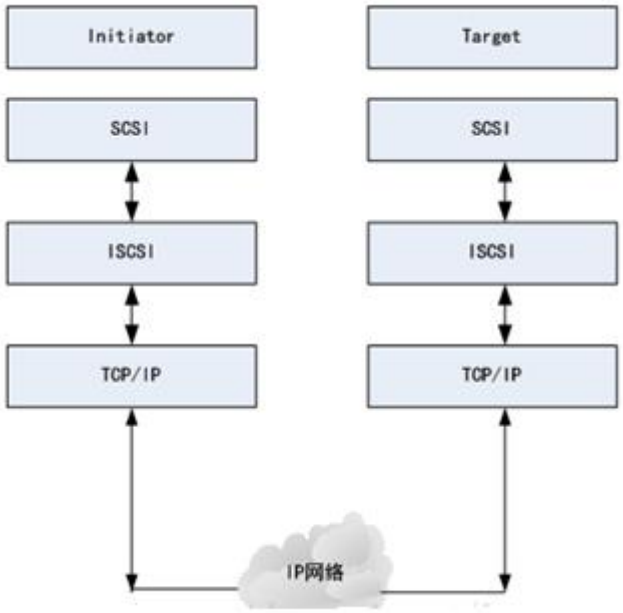
基于写时复制机制（COW），在创建快照时，存储系统会记录当前卷的元数据信息，并创建一个指向原始数据块的元数据映射表，当发生数据写入时，存储系统会先将要修改的数据块复制到新的位置（快照卷），然后在新的位置上进行写入操作，这样，原始数据卷上的数据保持不变，而快照卷上的数据随着写入操作而改变。快照只存储修改的数据和元数据信息，因此在存储空间方面具有一定的节省效果。

- 自定义自动快照策略，设置定时快照时间、保留时间、保留数量等信息，支持查看自动快照执行任务信息，包括快照源、快照 ID、执行任务状态、执行/结束时间等信息；
- 提供秒级手动快照，仅占用少量元数据空间，快照间数据相互独立，对某个快照的操作不影响其他快照；
- 可进行回滚操作，将数据恢复至某个（时间点）快照；
- 支持更新和删除操作。

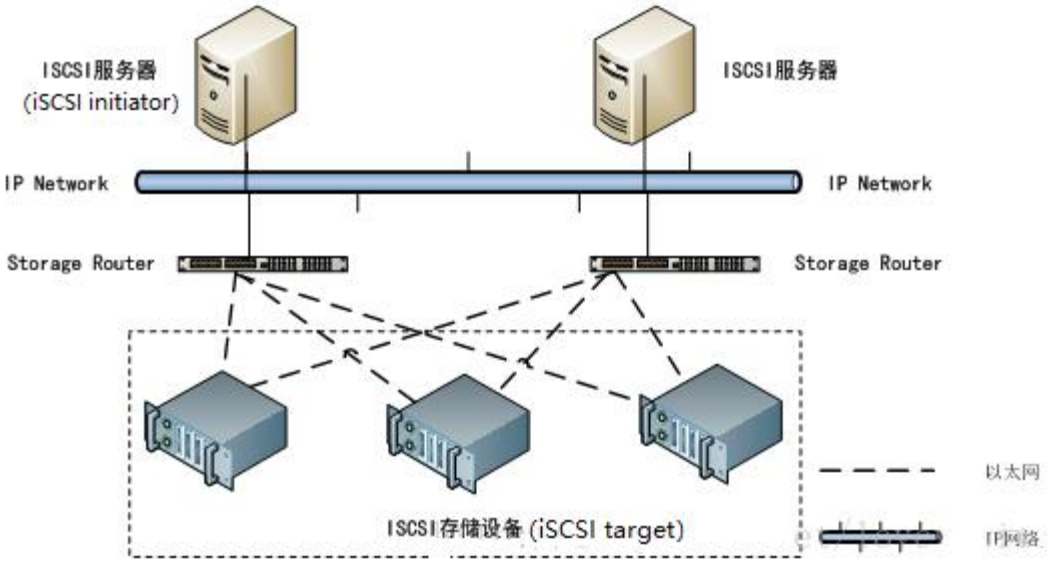
快照机制能够有效地保护用户的重要数据，在误删、误覆盖、被篡改的场景下，提供完备的数据恢复机制，大大提高了重要数据的安全性。也能够在业务出现逻辑错误，从而导致大量数据错误的场景下，快速、自动地对数据进行回滚恢复，极大的降低了逻辑错误场景下的数据恢复成本。

4.1.6 卷映射（ISCSI）

iSCSI（Internet Small Computer System Interface）映射是一种将远程存储设备通过网络连接到主机系统的技术。它允许主机系统通过网络访问远程存储设备，就像直接连接在本地主机上一样。



针对客户端通过 iSCSI 协议访问存储设备的场景，UStor 统一存储提供启动器组功能，可以将卷与 iSCSI 客户端进行映射，启动器组内的所有启动器（iSCSI 客户端 IQN）均可访问映射的卷，即通过不同的 IQN 创建多个启动器，多个启动器归属于某个启动组。



1）iSCSI Target 设备配置：在远程存储设备上配置一个 iSCSI Target，这是提供存储服务的设备。iSCSI Target 将通过网络提供存储卷或逻辑单元给主机系统。

2）主机系统配置：在主机系统上安装 iSCSI Initiator 软件。iSCSI Initiator 是主机系统上的软件或驱动程序，用于与远程存储设备进行通信和管理连接。

3）iSCSI Target 发现：主机系统通过网络发现可用的 iSCSI Target 设备。这可以通过在主机上配置 iSCSI Initiator 来实现，指定 iSCSI Target 的 IP 地址或主机名。

4) 建立连接：一旦 iSCSI Target 被发现，主机系统会与其建立连接。它会通过 iSCSI 协议与 iSCSI Target 进行握手和身份验证，建立起可靠的通信通道。

5) 逻辑单元映射：连接建立后，主机系统可以请求 iSCSI Target 上的存储卷或逻辑单元进行映射。iSCSI Target 将响应并将存储卷映射到主机系统上，使其在操作系统中可见。

存储卷映射完成，主机系统就可以像访问本地存储设备一样访问远程存储设备。它可以将其用作磁盘驱动器、卷或文件系统，执行读取、写入和其他操作。

UStor 统一存储通过启动组管理，实现卷映射、访问等操作。启动组管理包括以下功能：

- 创建/删除启动组
- 开启 CHAP 认证（设置、修改用户名和密码），默认关闭认证
- 支持启动组内创建多个启动器，卷同时映射给多个客户端
- 支持在映射中移除启动器或卷

4.2 文件存储

4.2.1 概述

文件存储是一种将数据以文件形式保存在存储介质中的方法，文件通常以层次化的目录结构进行组织，使用户能够方便地组织和管理文件。每个文件都有一个唯一的文件名，用于标识和访问该文件。

UStor 文件存储，是一个强大的分布式文件系统架构，将文件数据分散存储在多个节点或服务器上，以实现高可靠性、可扩展性和高性能的文件存储和访问。这种架构在大规模数据存储和高并发访问的场景下非常有用，并且可以为用户提供可靠和高效的文件存储解决方案。

UStor 文件存储支持开启/关闭 NFS 共享协议，支持创建共享协议策略，如白名单 IP、写入权限控制、root 权限限制。

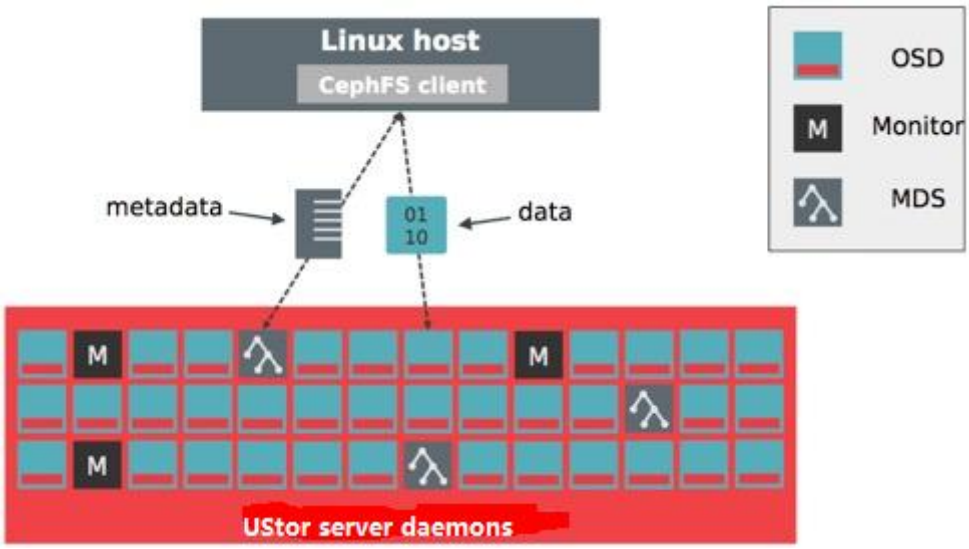
4.2.2 目录管理

UStor 文件存储基于 cephFS 文件系统，是一个高性能、分布式的文件系统，提供了一个统一的命名空间，使得用户可以像使用传统的文件系统一样访问存储在 UStor 存储集群中的数据。

CephFS 使用元数据服务器来管理文件系统的元数据。元数据包括文件和目录的信息，如文件名、权限、层次结构等。MDS 服务器负责处理客户端的元数据操作请求，并将元数据存储在 RADOS 中。CephFS 使用了一种称为 Ceph Metadata Server Cluster (MDS Cluster) 的机制来实现高可用性和负载均衡。

文件的分布，采用 CRUSH（Controlled Replication Under Scalable Hashing）的分布式算法来确定文件在存储集群中的位置。CRUSH 算法将文件的元数据和数据位置映射到存储集群中的物理位置，以便客户端可以直接与存储节点进行通信，而无需中央调度。

客户端通过与 MDS 服务器通信来进行文件系统操作，如文件的创建、读取和写入。客户端可以使用标准的文件系统接口（如 POSIX 接口）与 CephFS 进行交互，从而使应用程序能够透明地访问 UStor 存储集群中的数据。

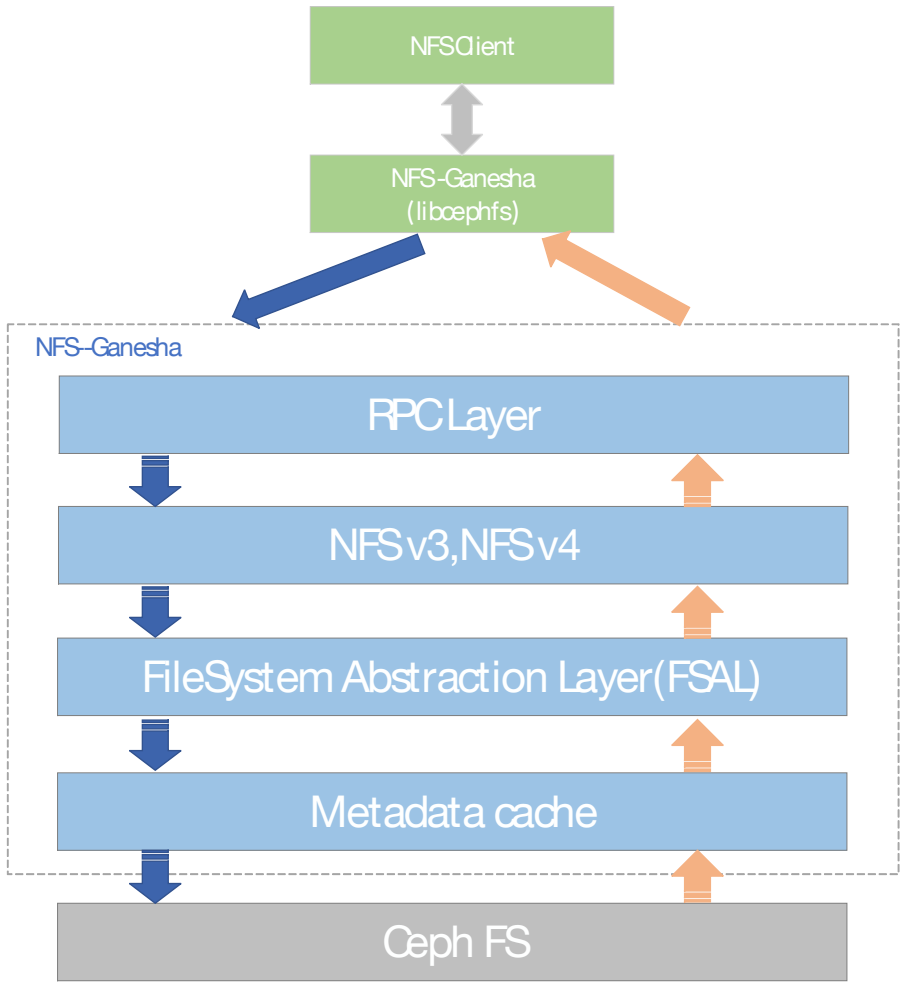


- 支持创建文件目录，以树型方式管理目录、子目录和文件，只支持删除、强制删除目录
- 支持查看目录、文件大小、路径、更新时间等信息
- 支持查看目录的详情，包括当前目录的父目录、大小、共享状态、路径等信息，也可在详情中创建快照、开启 NFS 共享等功能

4.2.3 NFS 协议共享

NFS（Network File System）是一种用于在网络上共享文件的协议。它允许计算机通过网络透明地访问远程文件和目录，就像它们是本地文件一样。NFS 最初由 Sun Microsystems 开发，并成为 UNIX 和类 UNIX 系统中最常用的网络文件共享协议之一。

UStor 文件存储支持 NFS 接口（NFSv3 和 NFSv4），客户端可通过 NFS 协议，mount 挂载远端文件目录，对其进行数据读写、权限设置等操作。



基于开源软件 nfs-ganesha，它是一个 NFS 服务器，支持将不同的文件系统通过 FSAL（File System Abstraction Layer）导出为 nfs，支持 CephFS、GPFS 等文件系统导出为 NFS。

- Metadata cache：后端文件系统的元数据缓存
- FSAL：文件系统抽象层，统一后端不同文件系统的 API
- NFS：nfs 服务，支持 v3 和 v4 版本

UStor 文件存储对相应目录可以开启/关闭 NFS 共享协议，设置协议策略，包括白名单 IP 段/IP 地址、写入权限、root 权限限制，支持查看共享地址和挂载命令显示。

4.2.4 快照管理

文件存储快照是一种用于存储系统的快照技术，用于创建存储设备上文件或文件系统的时间点备份或副本。它记录存储设备上文件或文件系统的状态，以便在需要时能够回滚到先前的状态。

UStor 文件存储快照能力由 cephfs 提供，通常通过使用 mkdir 命令创建快照目录，快照目录是一个隐藏的特殊目录，在目录列表中不可见。

•数据保护和备份：文件快照可以用于保护和备份文件系统中的数据。通过创建快照，可以在意外删除、文件损坏或恶意操作发生时，恢复到之前的文件系统状态。这样可以有效防止数据丢失或损坏。

•版本控制和历史记录：文件快照可以提供文件系统的版本控制和历史记录功能。每个快照都代表了文件系统在某个特定时间点的状态，因此用户可以随时查看、恢复或比较不同时间点的文件版本，方便进行文件修复、回滚或版本回退。

•快速恢复和回滚：文件快照可以在需要时快速恢复文件系统到先前的状态。当发生数据丢失、错误或意外更改时，可以使用快照来回滚文件系统，并恢复文件和目录到之前的状态，以避免重要数据的丢失或损坏。

•存储空间优化：文件快照通常采用写时复制（Copy-on-Write）技术，即在快照创建时，只复制已更改的数据块，而不会复制整个文件系统。这种方式可以节省存储空间，并提高快照创建和删除的效率。

UStor 文件存储支持手动创建一次性快照和自动快照策略两种方式，客户根据实际场景情况，灵活选择快照创建方式进行快照操作。

•手动快照，支持目录选择和回滚操作；

•快照策略，可以设置重复周期（天、周、月、间隔）、执行时间（1 小时单位）；

•支持天数和时间的多选；

•保留快照数（默认最多 20 个）。

4.3 对象存储

4.3.1 概述

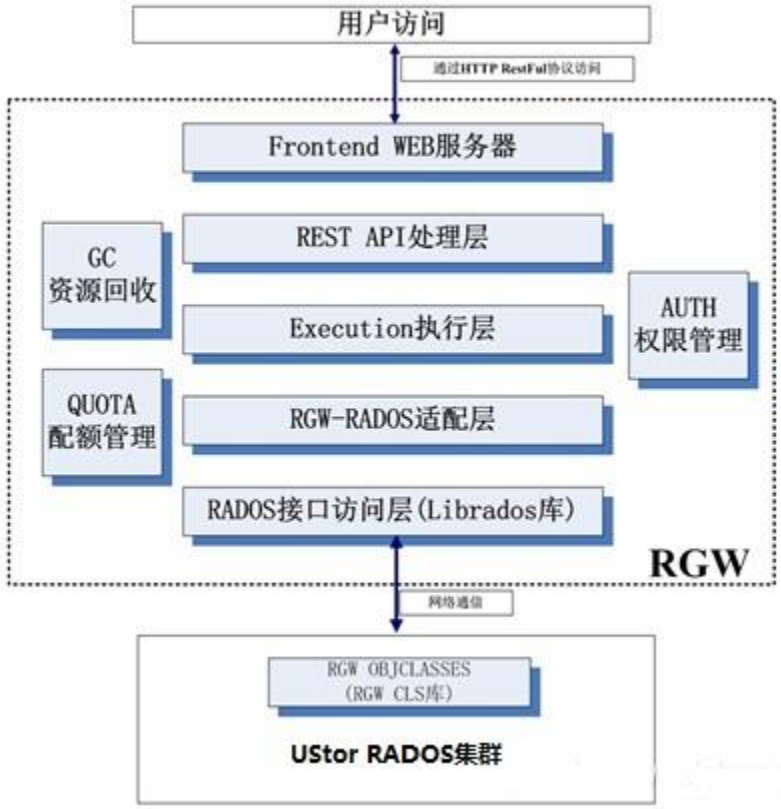
对象存储（Object Storage）是一种数据存储模型，用于存储和检索大规模、非结构化数据。与传统的文件系统和块存储相比，对象存储提供了更高的可伸缩性、耐久性和灵活性。

在对象存储中，数据以对象（Object）的形式存储。每个对象由数据本身、元数据和一个唯一的标识符（通常是一个 URL 或键）组成。对象存储不像传统文件系统那样有层次结构，而是使用一个扁平的命名空间，可以根据需要创建任意数量的存储桶（Bucket）来组织对象。存储桶是一种逻辑容器，用于存放相关对象。

UStor 对象存储基于 Ceph RGW 组件，提供对象存储接口，允许用户通过 HTTP 和 S3 协议与 UStor 对象存储集群进行交互。Ceph RGW 充当了 UStor 存储集群与外部应用程序或客户端之间的门户，使得应用程序可以通过标准的对象存储接口(S3)与 UStor 进行交互。

4.3.2 逻辑架构

RGW 是运行于 UStor RADOS 集群之上的一个客户端实例，是 UStor 存储集群对外提供对象存储服务的一个网关，它允许用户通过 restful api（S3 协议）的方式访问 ustor 集群。向下 RGW 通过 librados 接口库访问 UStor rados 集群。



•Frontend web 服务器用于监听并接受 http 服务请求，默认情况下 RGW 已内嵌该进程，相关请求在 RGW 进程内直接处理；

•Rest API 负责处理 S3 协议的 API 特性逻辑。用户操作请求交由该层处理，该层依据具体的 S3 协议，去除协议特性，按照 RGW 对象标准操作格式提交给下一层；

•Execution 层按照 RGW 对象的标准格式处理各类操作请求，这一层会根据操作请求处理 RGW 对象的元数据和内容数据，并操纵 RGW-RADOS 适配层进行 IO 操作；

•RGW-RADOS 适配层会将 RGW 对象的操作转化为对 UStor RADOS 对象的操作。操作 UStor RADOS 对象主要通过 LIBrados 接口实现；

•LIBRADOS 是 UStor RADOS 系统的对外窗口，在 RGW 应用环境中它运行在 RGW 进程内，并通过网络协议访问 UStor 存储系统内的 OSD、Monitor。

•GC 资源回收，辅助模块，对临时资源、磁盘空间进行资源回收。

•QUOTA 配额管理模块，用于存储桶和用户的配额管理，包括桶内对象大小、对象数，用户配额可限制用户所有对象的大小和对象数量。

•AUTH 权限管理，涉及 S3 协议用户的身份认证、用户对资源的操作权限（读、写、删除等）。

4.3.3 桶管理

在对象存储中，"桶"（Bucket）是一种逻辑容器，用于组织和管理对象。UStor 对象存储桶管理涉及创建、删除、配置和监视存储桶，以及管理与桶相关的访问权限和属性。桶必须赋予用户，即桶用户。

•创建桶，是指在 UStor 对象存储系统中创建一个新的存储桶。需要指定桶的名称和所属用户。创建桶时，还可以设置桶的访问类型、是否开启多版本等属性。

•删除桶，从 UStor 对象存储系统中删除一个现有的存储桶及其所有对象。删除桶需要确保桶为空，即没有任何对象存在于其中,不然无法删除该桶。

•桶用户，支持单独创建桶用户，赋予给某个桶，用户可访问、读取该桶，如用户没有赋予该桶，用户无法访问该桶及桶内数据。桶与用户是多对一的关系，即一个桶只能赋予一个用户，一个用户可以赋予给多个桶。

•配置桶属性，包括访问类型、是否开启多版本、是否开启对象锁定、生命周期管理等。通过配置桶属性，可以更好的使用桶，满足客户多样性场景的存储需求。

•监视桶，指跟踪和记录桶的活动和使用情况。监视桶可以提供有关桶中对象数量、使用量、请求次数、上传下载带宽等指标的信息。这些指标对于容量规划、性能优化和安全审计非常有用。

•访问控制，通过访问控制机制，可以控制桶的公开访问或私有访问，指定允许的用户，令牌管理，控制访问 IP（白名单、黑名单）、令牌权限以及授权文件等安全控制项。

•跨区域复制，UStor 对象存储系统支持跨区域复制功能，允许将一个桶中的对象复制到另一个区域的桶中。通过配置跨区域复制策略，可以实现数据备份、灾难恢复或数据分发等需求。详见“多站点数据复制”章节。

•生命周期管理,生命周期管理功能允许自动管理桶中对象的生命周期。可以定义对象的保留时间、转换存储类别、自动删除等规则，以便根据对象的时间和使用情况自动执行操作。详见“生命周期管理”章节。

4.3.4 令牌管理

令牌管理是指在 UStor 对象存储系统中管理访问令牌（Token）的创建、更新等操作。令牌是一种安全凭证，用于验证和授权访问对象存储系统中的资源。

对某些特殊业务场景，应用程序需要通过 token 密钥的方式访问 ustor 存储系统，UStor 对象存储提供访问令牌，以及相关能力。

- 创建生成令牌，在 ustor 对象存储系统中，支持创建令牌，上层应用通过令牌来进行身份验证和授权，成功后才能访问对象存储资源。
- 删除令牌，支持删除已经创建的令牌，注意删除后，即便该令牌没失效，但已经删除了，则该令牌无法继续使用。
- 临时令牌，usor 对象存储系统支持临时令牌。临时令牌是一种有时效性的令牌，可以在一定时间后自动过期。
- 令牌权限，通过令牌管理，可以定义不同令牌的授权对象、权限级别和访问范围。授权对象为桶级别；权限级别涉及上传、下载、删除和获取列表；访问范围可以设置为“所有文件”和“设置前缀”。为不同的用户或应用程序生成不同权限的令牌，以控制对对象存储资源的访问和操作。
- 令牌过期和续期，对于即将过期的令牌，需要进行续期，以保持持续的访问权限。令牌更新续期有用户触发操作。如过期不进行更新，则令牌失效，无法通过该令牌访问 ustor 对象存储。
- 令牌黑白名单设置，支持按照 ip 段方式设置令牌的黑白名单，防止非法访问。

令牌管理在 UStor 对象存储系统中是非常重要的，它提供了对资源的安全访问和控制。

4.3.5 文件管理

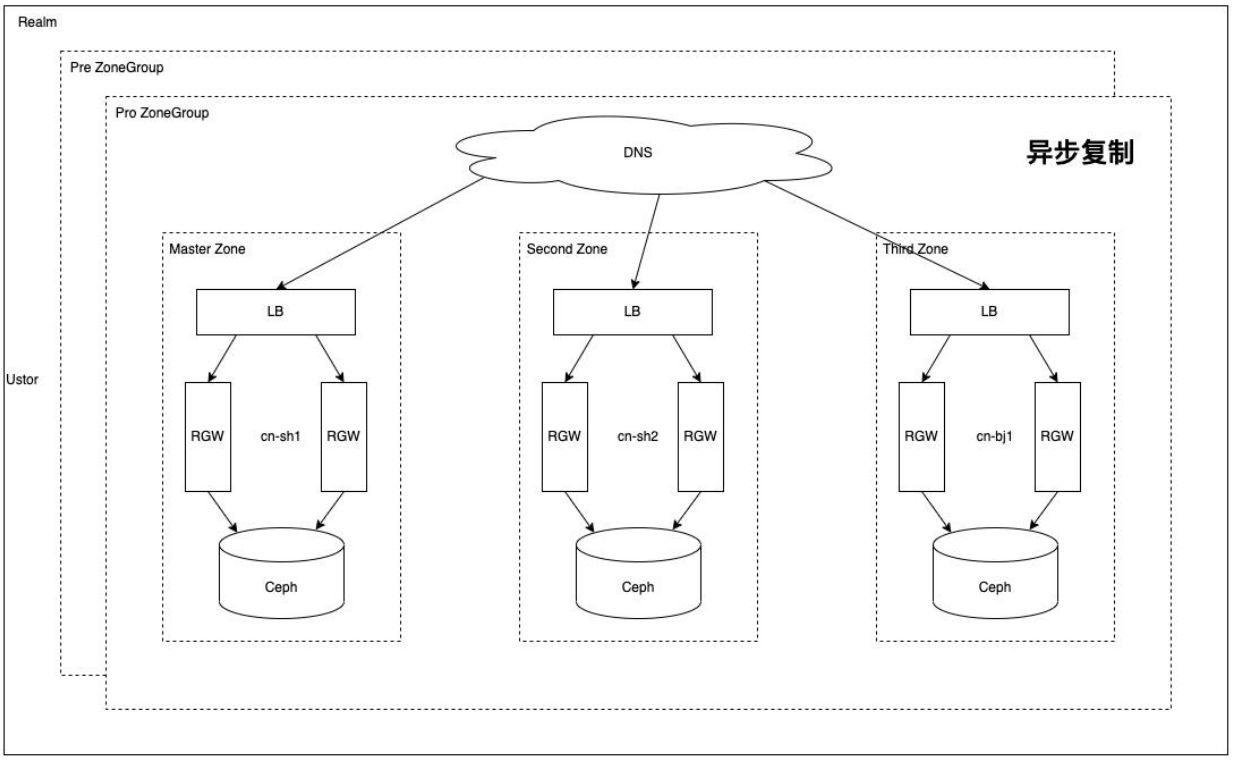
文件管理即 S3 client，通过标准 S3 sdk 封装，用户可在控制台实现图形化的对桶文件进行操作，包括下载、上传等文件管理操作。

- 支持上传文件至存储桶，下载桶内文件至本地。
- 可查看上传文件进度。
- 支持直接删除桶内文件，可一次删除多个文件。
- 支持在桶下直接创建、删除目录，目录必须为空方可删除。
- 支持一键清除桶内所有文件、目录和文件碎片。
- 支持查看没有上传成功的文件列表。
- 支持查看被删除对象的历史版本（删除标记）。

4.3.6 多站点数据复制

对象存储的多站点同步是指将对象存储系统中的数据在多个站点之间进行同步，以实现数据的冗余和高可用性。这种同步是异步进行的，即数据在主站点进行写入后，异步地复制到其他站点。

UStor 对象存储多站点数据同步，它是双向、实时的，即多个站点互相同步，同步策略设置在桶级别，可根据前缀过滤同步的文件数据。



•区域（Zone):定义了由一个或多个（可以对客户端的请求做负载均衡）ceph-radosgw 实例组成的逻辑组。

•区域组（Zone Group):包含一个或多个 zone。在一个 zone group 中，一个 zone 将会被配置成 master zone。master zone 处理所有 bucket 和 user 的变更。Secondary zone 可以接受 bucket 和 user 操作请求，然后将操作请求重定向到 master zone。如果 master zone 出现故障，secondary zone 将会被提升为 master zone。

•领域（Realm):它代表一个全局唯一的命名空间，包含一个或者多个 zone group。但必须要有一个 master zone group。Realm 使用 period 的概念来管理 zone group 和 zone 的配置状态。每次对 Zone group 或 zone 进行变更，都会对 period 做 update 和 commit 操作。

在实际部署生产的时候，推荐采用单 Realm、单 ZoneGroup、多 Zone 来部署。也可以在相同的 ceph 集群上，一同部署，使用多 ZoneGroup 将生产和开发环境隔开。

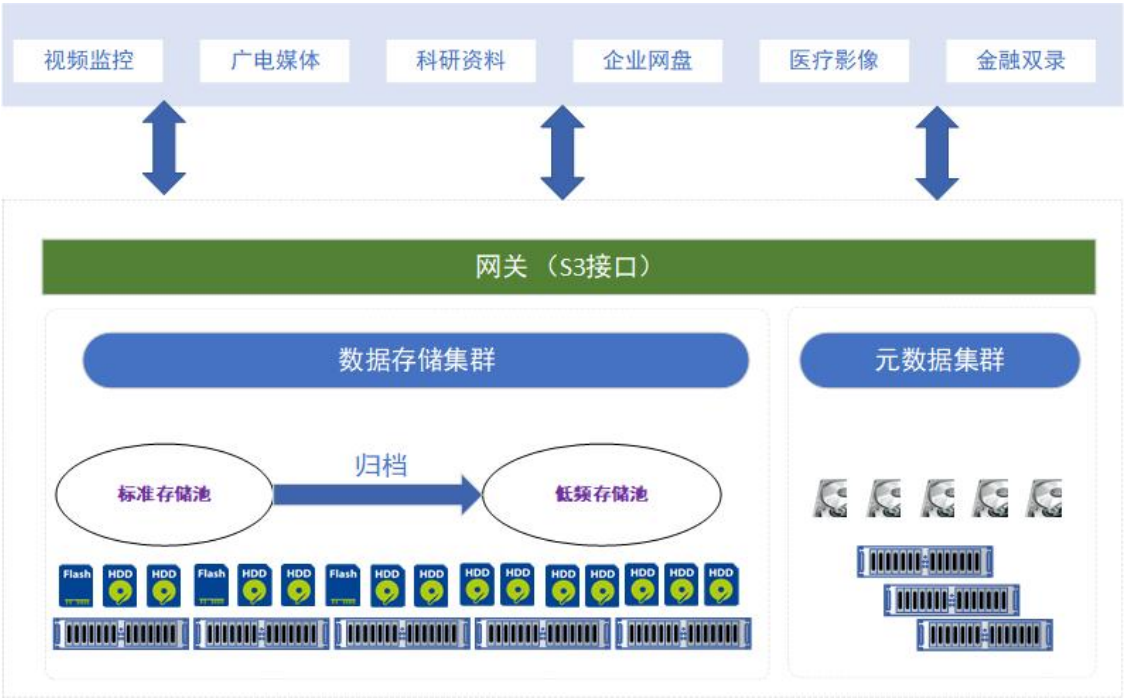
需要注意的是，在备站点可以对桶内对象文件，以及告警进行读写操作外，不允许对其他任何对象进行写更新操作，如用户、令牌等，只能读取。

金融、医疗等行业的业务数据保护述求，对数据合规性和安全性有严格要求，如双录系统、PACS系统等；又如总部集团—分支机构、政府上下级单位等，存在大规模数据分发或汇聚的场景，UStor对象存储实现跨地域、跨集群的对象数据实时双向、异步同步，保障业务数据安全和连续性的同时，为客户大数据分析等业务提供数据汇聚或分发能力。

4.3.7 生命周期管理

对象存储生命周期管理是一种管理和自动化对象存储中数据生命周期的策略和过程。通过定义生命周期规则，可以根据数据的特性和需求，自动化执行数据的迁移、转换、存储级别变更和删除等操作，以优化存储资源的利用和数据管理的效率。

UStor 存储支持数据的全生命周期管理，可以自定义管理策略，按照时间维度、整桶或指定前缀范围，将数据自动归档至低介质存储池，实现数据分级存储，长期保存数据的同时降低硬件投入成本。针对失效或过期的数据，也支持设置自动删除策略，防止无效历史数据占用大量空间容量。提升数据处理效率。



- 支持按照时间维度设置规则策略。
- 支持在同一集群存储中，创建多个存储类（不同介质不同存储池）。
- 支持对整桶或指定前缀文件进行删除或归档操作。

- 支持对象数据归档至其他存储池，应用访问路径不变，无需任何修改，为上层用户提供透明统一的访问路径。

- 配置策略包括当前版本的删除或迁移归档、历史版本的删除或迁移归档以及清理未完成的分片上传，即配置分片上传多少天后将其删除。

生命周期管理的应用场景针对非结构化对象数据，一次写入，长期保存，数据量只增不减的应用业务，如金融行业，受监管要求其保存周期通常数十年；互联网类业务，如电子签章（签名）、电子发票等非结构化数据，需要保存 3 年以上；政府政务，绝大部分电子档案需保存 20 年到 70 年，重要档案需要永久保存；医疗行业，住院病历数据保存不得少于 30 年；保管保存医疗影像数据时间要求不少于 15 年。

5 资源管理

5.1 集群资源

对 UStor 存储集群进行管理和设置，从集群视角，让客户直观了解集群层面信息，包括集群 ID、架构、存储用量、副本冗余策略、角色。

支持对集群层面进行管理设置，包括：

- NTP 服务，设置 NTP 地址。
- 集群恢复，默认开启，集群将运行故障数据恢复和数据重均衡操作，并且分为 3 档位，分别对应不同级别的数据重建占用带宽，级别越大占用带宽越高，数据恢复和重建的速度最高。
- 硬盘离线超时时间，可设置离线超时时间，当硬盘故障导致存储服务不可用时间达到硬盘离线超时时间之后，故障硬盘将被下线，其上承载数据将自动均衡到其他健康硬盘上。

5.2 节点资源

节点管理是指在一个计算机集群中对节点进行配置、监控和维护，通过 ustor 控制台可快速查看、了解当前集群节点状态和各资源使用情况，便于管理人员对集群进行运营运维计划和操作。

- 节点信息列表展示，展示信息包节点 IP、状态、CPU、内存、类型等信息
- 监控数据，节点当前 CPU 使用率、内存使用率、CPU 负载、磁盘总吞吐、TCP 连接数、网卡带宽等信息；
- 基本信息，包括 cpu 信息、内存信息和 NTP 信息；
- 网络信息，当前节点的网卡数显示，每个网卡的 bond 模式、MAC 地址、IP 地址和掩码信息；
- 磁盘信息，节点分区和挂载点，磁盘介质、分区大小和使用量；
- 存储 OSD，节点 OSD 信息，包括 osd id、健康状态、大小、使用量，所属节点等信息；

5.3 硬盘资源

针对存储集群的运营运维方面，除了常见的监控告警、自动数据均衡等能力外，ustor 还提供了细粒度的硬盘操作。通过图形化的硬盘维护管理操作，大大缩减维护复杂性，提高效率，降低维护风险和成本。

- 支持节点—硬盘间的关系拓扑展示。

- 支持硬盘监控状态、介质类型、OSD id、硬盘空间使用率（硬盘总容量、已有容量）信息展示。
- 支持选中单个硬盘，开启/关闭维护模式（如计划性对硬盘进行下线等维护操作，开启维护模式后，数据不会写入该硬盘）。
- 支持按照 OSD 级别，监控展示相关数据，包括 OSD iops、OSD 数据恢复、OSD 延迟、osd io 等信息，多维度、全方位了解当前硬盘性能。

6 运维运营管理

6.1 账号管理

账号管理在 UStor 存储中起着重要的作用。它确保了访问控制、安全性、资源管理、合规性等方面需求，为存储系统提供了有效的身份验证和授权机制，保护 ustor 存储中数据和资源的安全性。

UStor 系统提供多种灵活的身份验证和访问控制手段。

- 登录密码，支持对当前登录账号密码进行修改。
- 支持 FortiToken 身份认证，动态码方式登录。
- 支持登录访问限制设置，登录控制台的客户端 IP 限制，账号只能从指定的 IP 登录或访问 API。
- 支持 API 密码方式，通过密钥信息调用访问 ustor 存储 API。
- 数字证书设置，通过数据证书进行验证。
- 支持角色管理，系统默认角色和自定义角色，系统内置角色包括系统管理员和系统只读用户。

6.2 多集群管理

在集群管理方面，针对多地域、多站点、多集群场景，支持 web 管理控制台进行统一管控，系统全局设置。web 管理控制台通过多集群切换，对远端集群进行配置、监控和运维。

- 支持控制台在多个集群间无缝切换。
- 支持系统全局设置，如系统设置、账号管理等。

满足同城、异地灾备（多站点、多集群）场景下的统一管理诉求。适用于集团总部——分支机构、政府上下级业务单位存储建设场景。

6.3 回收站

回收站指，它允许用户删除文件或文件夹时，将这些文件暂时保存在一个特定的位置，以便在需要时能够还原或永久删除。

当你删除文件或文件夹时，它们通常会被移动到回收站而不是立即永久删除。这样做的目的是为了防止意外删除或错误删除重要文件。

UStor 存储系统的回收站针对块存储的卷而言，是一个用于临时存放已删除卷文件的地方，它提供了一种安全的方式来管理和恢复误删除的卷文件，当对卷出现误删等操作时，被删除的卷会存放在回

收站中，可执行恢复操作，被删除的卷及数据会被恢复。

在回收站中，你可以查看已删除的文件列表，恢复还原它们到原来的位置，或者永久删除它们。恢复操作会将文件恢复到它们最初的位置，就好像它们从未被删除过一样。销毁操作将彻底删除文件，使其无法被恢复。

另外，回收站中的文件也可能会占用一定的存储空间，如果回收站中的文件太多或者占用过多的空间，你需要定期清空回收站来释放存储空间，如果不对回收站中的文件做清理销毁操作，文件会一直在回收站中占用存储空间。

6.4 操作日志

操作日志（操作记录）是记录 UStor 系统中的操作活动的记录。它可以用来追踪和审计系统的使用情况，包括用户的操作、事件信息等。

操作日志对于系统管理和故障排除非常有用。通过查看 ustor 操作日志，管理员可以了解存储系统的状态、用户的行为。还可以用于故障排查和安全审计，帮助发现和解决问题，追踪存储系统异常和监测潜在的安全威胁。

- 支持按照时间和操作结果（成功、失败）条件过滤查询。
- 支持精确查询，如通过操作名称、关联资源 id 等精确查询相关操作日志。
- 支持创建通知规则设置，针对监控的模块，以邮件的方式通知其操作事件。

6.5 通知组

UStor 通知组是一组相关信息的通知，可以根据特定标准或条件进行组织和管理，用于监控告警、操作日志等功能中，以便将相关信息发送给特定的接收者。

- 支持创建或删除通知组
- 支持邮件方式创建通知

另外通过 webhook 方式，可以支持钉钉、微信等第三方通讯应用通知相关用户人员。

Webhook 是一种应用程序间通信的方式，它允许一个应用程序通过 HTTP 请求将数据发送到另一个应用程序。简单来说，Webhook 允许一个应用程序将实时数据推送到另一个应用程序，而不是等待另一个应用程序轮询或定期拉取数据。

6.6 监控告警

检测 ustor 软硬件、组件等异常或重要事件时发送通知，告知相关人员。

帮助管理员或运维人员及时了解系统的状态和资源使用情况，以确保存储系统的可靠性和稳定性。

- 支持告警模板创建/删除设置，指定资源类型，包括块存储、集群、文件存储、节点、对象存储、硬盘。
- 支持创建/更新/删除告警规则，涉及监控指标、对比方式、告警阈值、持续时间、重要程度、通知组。
- 关联资源信息列表展示，包括资源 ID、模板名称、模板 ID。
- 支持查看告警记录，包括指标说明、目标 ID、当前值、模板类型、状态、重要程度、告警时间信息。

监控告警的目的是及时发现和解决存储系统的问题，以减少潜在的停机时间、数据丢失或用户体验问题。通过有效的监控告警，管理员可以在问题发生前获得及时通知，并采取适当的措施以避免或最小化潜在的影响。

6.7 开放 API

API（Application Programming Interface）是一组定义和规范了软件组件之间交互的接口。它定义了一系列规则和协议，用于不同软件系统之间相互通信和互操作。

UStor API 是一种通过 HTTP 协议提供服务的 API（RESTful API）。它允许开发人员使用 HTTP 请求和响应来访问和操作远程服务器上的资源，如获取虚拟机、执行操作等。

在 ustor 管理平台开放 API 控制台，允许不同软件系统之间的通信和数据交换，以实现特定的功能。可直接通过控制台“开放 API”图形化操作调试、测试 API 接口，便于与第三次软硬件厂商产品进行对接和集成。

- 存储产品，包括卷、文件和对象存储相关 API 信息，以及每个请求参数的详细说明。
- 管理控制平面，包括集群、节点、全局配置、操作日志、回收站账号等功能相关的 API 信息，以及请求参数的详细说明。

通过使用“开放 API”，进行测试和调试，促进了不同系统之间的集成和互操作，使得不同应用程序能够共享数据和功能，实现更强大的功能和增强用户体验。

7 平台管理

7.1 自定义 UI

针对传统企业客户，某些特殊业务场景，如内部专有平台、第三方厂商等，建设云存储平台，需要替换存储平台 logo、标题等信息。UStor 支持自定义 UI 界面相关内容，可以根据企业或特定领域的需求进行自定义，以提供更符合用户喜好、提高易用性和品牌识别度的用户界面。。

- 支持登录页设置，包括 logo、标题、背景、描述信息等
- 支持网站设置，包括网站 logo、网站 title 等信息

通过 UStor 平台自定义 UI 能力，帮助客户或者合作伙伴构建自己的软件产品品牌形象和企业标识，加强品牌识别度和用户体验。

7.2 邮箱管理

平台支持全局邮箱设置，通过 UStor 邮件管理，保护邮箱账户安全的同时，便于告警监控、操作日志等事件通知。

- 支持设置发件人邮箱地址、密码、邮箱服务器以及邮箱主题信息
- 支持是否开启 SSL 加密
- 支持对邮件进行收发测试

7.3 统一授权

为防止非正规途径或渠道获取产品，侵害客户和用户的利益，同时也为产品的市场策略和销售需要，在新版本中，产品授权进行了更严谨、更充分的设计，让客户直观、清晰的了解当前产品授权详细信息，保证客户产品的合法性。

- 一键采集硬件信息、上传授权。
- 基础许可、高级功能许可和服务许可的授权信息详细展示。

软件授权通常使用许可证来明确授权条款和条件，基本包括以下内容：

授权范围：明确规定用户可以使用软件的方式、用途、限制和维保服务方面。

使用限制：规定用户不得以未授权的方式复制、修改、分发或销售软件。

有效期限：规定授权的时间范围，包括永久授权、临时授权（3 个月）。

权益保护：规定违反授权条款的后果和可能采取的法律措施。

对于 UStor 存储用户来说，遵守软件授权是合法使用软件的前提，同时也是维护软件开发者权益和保护用户利益的重要措施。

8 性能数据

8.1 硬件配置

3 节点缓存盘模式

节点*3	服务器型号	浪潮 Rack Mount Chassis
	CPU	Intel(R) Xeon(R) Silver 4314 CPU @ 2.40GHz*2
	内存	32GB*8
	磁盘	NVME SSD 8T *2 16T HDD * 10 480G SSD *1
	网口	双万兆网卡

8.2 集群信息

- 副本数： 三副本
- Ceph 版本： 16.2.11-1
- 系统版本： CentOS Linux release 8.3.2011
- 内核版本： 4.18.0-348.7.1.el8_5.x86_64

8.3 测试项

8.3.1 对象存储

测试工具：cosbench

Op-Type	IOPS	带宽	时延
小文件(4k)读	9342.84 op/s	37.37 MB/S	3.41 ms
小文件(4k)写	3065.61 op/s	12.26 MB/S	10.36 ms
大文件(4M)读	236.07 op/s	966.93 MB/S	40.81 ms

大文件(4M)写	189.95 op/s	778.02 MB/S	125.4 ms

8.3.2 文件存储

测试工具：vdbench

Type	IOPS	带宽	时延
4k 随机写	2282.3	8.92 mb/sec	0.013ms
4k 随机读	6413.05	25.05 mb/sec	0.004ms
4M 顺序写	164.45	657.75 mb/sec	76.708ms
4M 顺序读	233.9	935.4 mb/sec	68.033ms

8.3.3 块存储

RBD

- 测试工具：fio
- 测试引擎：rbd

Op-Type	IOPS	带宽	时延
4k 512 随机读	150.67k	588.67MiB/s	3.4ms
4k 512 随机写	80.57k	314.67MiB/s	6.36ms
4M 512 顺序读	629.34	2520.67MiB/s	820ms

4M 512 顺序写	**359**	**1437MiB/s**	**1430ms**
4k 1 随机读	**4073.67**	**15.94MiB/s**	**0.25ms**
4k 1 随机写	**904**	**3.53MiB/s**	**1.11ms**

ISCSI

- 测试工具：fio
- 压测引擎：libaio

Op-Type	IOPS	带宽	时延
4k 512 随机读	**62.7k**	**245 MB/S**	**8.17 ms**
4k 512 随机写	**11.8k**	**46.03 MB/S**	**43.45 ms**
4M 512 顺序读	**222**	**889.5 MB/S**	**2279.54 ms**
4M 512 顺序写	**153.5**	**616.5 MB/S**	**3338.67 ms**
4k 1 随机读	**2186.5**	**8.54MiB/s**	**0.91ms**
4k 1 随机写	**657**	**2.57MiB/s**	**1.51ms**

