中数国科分布式存储
技术白皮书

2023 年 8 月

目 录

1. 概述

该系统是一款全对称、去中心化的分布式架构存储系统，提供融合且统一的企业级块、文件、对象、大数据的存储服务。

系统内每个节点都能提供相同分布式存储软件，包括四个层：

高可用集群管理层、存储协议层、统一资源层以及系统平台层，系统内无独立元数据服务节点，消除性能瓶颈，不存在单点故障，在节点扩容、故障场景下都能无缝平滑切换，业务无感知。

作为存储系统的基础，每个节点采用全活的不共享方式，每个节点都有独立的 CPU/内存/硬盘资源，分布式存储系统的数据平均分布在各个节点上，避免了系统资源争用，消除了系统瓶颈；如果出现单个节点故障，系统也能够自动识别故障的节点，自动重构故障节点涉及的数据和元数据，使故障对业务透明，完全不影响业务连续性。整系统采用全互联全冗余的组网机制，全对称分布式集群设计，实现存储系统节点的全局统一命名空间，从而允许系统中任何节点并发访问整系统的任何存储对象；内部从多个节点并发访问不同的对象或相同对象的不同区域，实现无锁高并发高性能无阻塞读写。

2. 软硬件架构

2.1. 架构逻辑

分布式融合存储集群提供了块、文件、对象、大数据等通用存储协议功能，实现了高可靠性、横向扩展等企业级存储系统的优点，并具备海量数据管理的能力。整个集群有很好伸缩性，线性扩展可支持上百 PB 乃至 EB 级的存储数据池。集群中不共享资源的存储节点以通用服务器硬件为基础，使用软件定义存储的设计思想，使用以太网络将多活的存储节点互联组成超大规模局域网络硬盘池，每个网络硬盘池都可以继续抽象为提供存储能力的统一存储池，NX-SPool 组织起了大量能够提供并发读写存储能力的统一存储设备，节点之间靠 TCP/RDMA 网络来实现具有分布式一致性的冗余数据事务、并能够基于全局逻辑空间动态地平衡数据分布，具备统一存储能力的 NX-Spool 将块、文件、对象、大数据等通用存储协议产生的数据进行融合统一管理，为计算场景提供高性能存储服务。

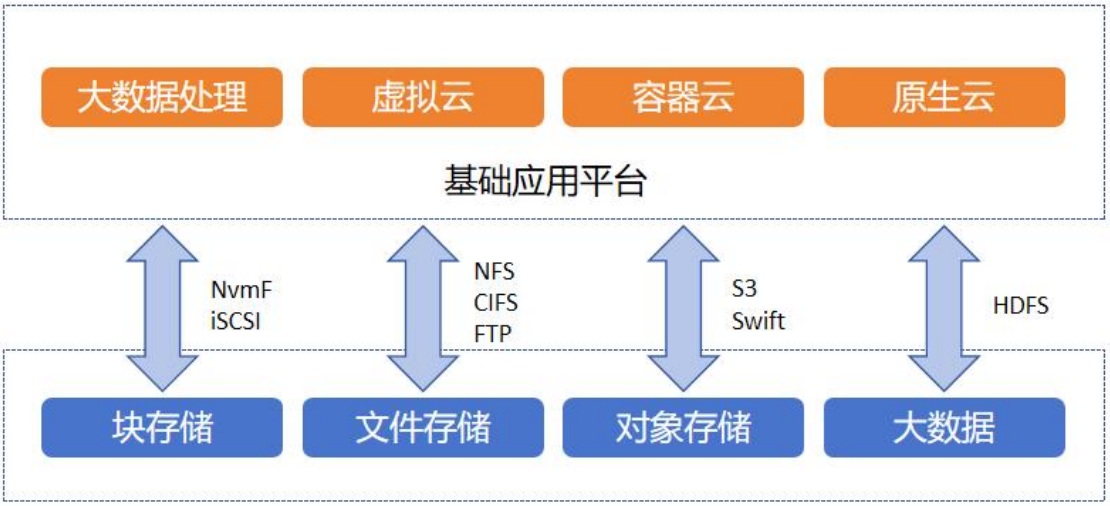
产品逻辑图

2.2. 硬件组成架构

分布式存储系统采用全对称架构，对相同类型硬件的节点而言，节点内部的软硬件配置完全对等，这样的设计使得用户首次购买或者扩容时，不需要考虑独立元数据服务器或独立网管服务器等问题，只需根据实际需求计算出需要的节点数即可，最小 3 台起配。

2.3. 运行环境及服务

通过 NFS 共享、CIFS 共享和 FTP 等方式为用户提供文件存储服务；通过 iSCSI 方式为用户提供块存储服务；通过 S3 和 swfit 为用户提供对象存储服务；通过 HDFS 为用户提供大数据存储服务。



2.4. 存储先进性

2.4.1. 未来就绪，统一数据平台

使用集群存储系统统一数据平台，提供块、文件、对象等所有存储类型接口，支撑上层数据整个生命周期，从数据采集、存储、管理、利用；提供统一的数据平台，实现存储节点采用分布式集群存储架构，支持全局单一文件系统和统一命

名空间。

对于数据管理则可以通过各种企业级特性保障数据安全可靠，同时结合数据生命周期和数据分层可以对数据的存储要求和性能要求按需分层存储和数据流动， 以及数据生命周期结束后自动删除。数据分层可以根据分层策略自动分层到蓝光 系统，降低整体投入成本。

可以实现通过访问存储管理可视化统一管理界面，即可完成对存储系统存储节点、软件系统和业务的统一管理，无需专业技术人员，减低管理成本。所有的业务操作都可以通过统一管理界面完成。

支持对物理存储节点、存储介质、存储池数据冗余状态、负载监控及管理；支持对象、块和文件的网关和链路健康检测；支持存储介质根据 SMART 信息预测设备寿命，提醒坏盘。支持物理服务器 CPU 、内存、网络、负载监控；支持存储介质读写 IOPS 、带宽和延迟监控；支持存储池读写 IOPS 、带宽和延迟监控；支持卷、桶、文件目录读写 IOPS 、带宽和延迟监控；支持上述指标统计保留天数 自定义设置。支持集群内所有资源的告警，在存储系统的各级软硬件产生故障时，由管理控制台向管理员提示告警，有助于及时了解资源使用情况和处理突发事件。支持自定义告警通知，同时支持邮件告警。告警会显示告警原因和建议解决意见，当故障恢复之后，告警自动恢复。

2.4.2. 数据永活，释放数据价值

Make Data Alive 让数据流动起来，数据战略长期专注在存储子系统如何更好 的保存、管理、再次基于数据进行价值创造，无论这些数据是对象，文件，还是块。面向未来的 5 年，2022-2027 年，除了基于数据如何存，还关心数据如何流动，进入数据湖架构，关联数据生命周期。

2.4.3. 异构平台，海量数据统一管理

分布式系统能够统一管理异构芯片基础设施，包含基于国产鲲鹏、飞腾、海光、兆芯等 CPU 的硬件设备，提高数据中心管理效率，降低分布式系统建设和运维管理成本。分布式系统可以支持混合芯片与异构厂商产品。

跨生态的数据统一管理，实现业务无感知的平滑生态融合主流信创生态体系可以在一个分布式系统环境下构建不同生态的存储池，统一的数据管理界面，实现跨生态的管理融合。跨生态的数据内部流动，实现业务无感知的平滑生态过渡，

以生态存储池为单位定义热、温、冷数据池，有效利用各个生态资源；按自定义策略自动对数据进行分层管理，实现数据的生命周期管理；跨生态的数据在线迁移，实现业务无感知的平滑生态切换；跨生态的数据迁移，实现跨生态的存储切换；业务无感知的数据迁移。

2.4.4. 一池多芯、多池多芯

分布式系统的任意存储资源池底层硬件设备可通过三种或以上服务器架构，不同操作系统搭建，实现软件灵活适配，提供分布式系统实施落地的能力。

2.4.5. 全存储介质支持

多存储介质支持。兼容 SAS 、NL-SAS 、SATA HDD 、SATA SSD 、m.2 SSD、u.2 NVMe 等接口。

2.4.6. 全协议支持

分布式系统支持 S3 、POSIX 、NFS 、HDFS 、CIFS 、FTP 、RBD 、iSCSI、FC 、Local SCSI 、CSI 和 RBD 原生访问等标准协议和访问接口。

2.4.7. 全业务场景适配

分布式系统端到端的场景化解决方案对接，可扩展接入各类应用服务器，适配大数据、容器、两地三中心、私有云、多套应用软件（包括数据管理软件、运维软件和显控软件等）、数据备份、数据归档等各种应用场景。满足不同的应用场景需求，提供文件、块、对象存储服务的软件定义存储系统，自由调配块存储、文件存储、对象存储的存储池划分。共享访问全系统的存储空间；实现基础架构的融合统一。

2.4.8. 无缝接入，管理简单高效

可以满足并实现 100% CLI 或者 GUI 管理方式，可无缝接入网络环境，无需 额外设备就可以实现对存储系统的管理。

2.4.9. 优化性能，海量数据加速

通过构建分布式的系统设计，完整的企业级功能和面向未来的架构简化，数据存储子系统平台可以支持全协议，并通过不同类型存储节点利用，能同时在高性能主存储和次级存储领域进行使用。

打不爆的缓存：利用海量数据加速技术，实现了高性能 SSD 和大容量 HDD 的深度融合。所有业务随机小块 IO 会写入多副本 SSD 层，然后通过创新的 IO 冷 热感知聚合技术，将 SSD 层的随机写合并成顺序大块回写到使用 EC 冗余策略的 HDD 层，同时避免了 EC 冗余策略下的写放大问题。另外，SSD 层和 HDD 层的 解耦设计，允许各自独立扩容，并且 SSD 在故障情况下，换盘只在 SSD 层内部进 行重构，完全不影响 HDD 层的数据健康。

强大的非结构化数据平台，在架构上分别支持高性能文件系统和千亿级海量对象存储系统，为非结构化数据场景提供从核心业务、在线分析到离线备份归档的端到端场景覆盖。同时内置强一致能力，为业务提供数据中心高可用。

2.4.10. 性能容量线性增长

分布式系统应具有良好的可扩展性，应支持超大容量的存储空间；进行扩容存储节点，后不需要做大量的数据搬迁，系统可以快速达到负载均衡状态。支持灵活的扩容方式，可以独立扩容、硬盘、存储节点，或者同时进行扩容操作。分布式存储平台的控制接口、存储带宽和缓存都均匀分布到各个分布式存储节点上，系统 IOPS 、吞吐量和缓存随着节点的扩容而线性增加。性能没有瓶颈。

分布式系统可以通过存储节点的增加，网络接口的扩容来实现千万 IOPS 的能力 。支持 1GE 、10GE 、25GE 、40GE 、100GE 、4GBFC 接口、8GB FC 接口、 16GB FC 接口、32GB FC 接口、100GB InfiniBand 接口、200GB InfiniBand 接口等 丰富的主机接口类型，满足高性能 IO 的需求。

2.4.11. 极致性能底座

分布式系统基于软件定义技术实现企业级存储能力，极致性能紧耦合硬件能力，甚至深度感知并整合。信创领域的 ARM 架构，众核利用都是用好 CPU 性能的核心武器。针对 CPU Socket 、Memory Node 、PCI-E 接口分布，都跟 IO 流的所有组件进行了整合，保证端到端的 NUMANode 亲和度。

通过 NUMA Node 的划分，不同的业务进程隔离在不同核上运行，避免了不 同业务分组对 CPU 资源的争抢和冲突 ，在多核系统上可以有效减少 Atomic
operation、Spinlock 等互斥机制的开销，提高整系统的线性扩展能力。

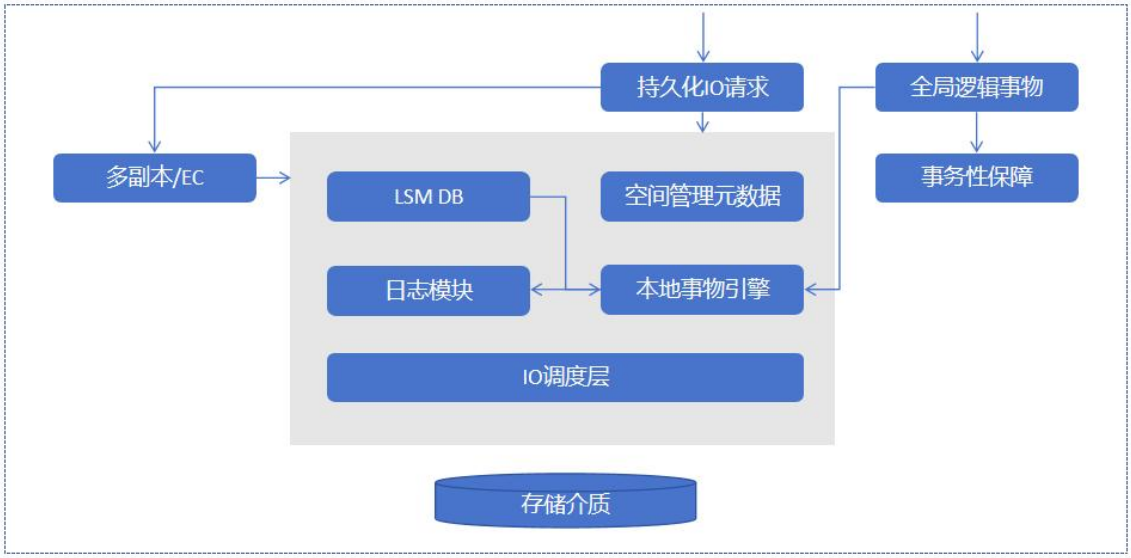
2.4.12. IO 全局智能调度

为保证稳定时延，分布式系统对所有 IO 都进行优先级标识，根据标识的优先 级，系统在优先级队列处理、线程并发度方面进行控制，实现高优先级 IO 优先处 理。

所有客户端业务 IO 为高优先级，系统后台 IO 根据任务的重要性分别对应优 先级，同时赋予恢复 IO ，数据 GC IO 进行了动态调控，结合数据安全的要求，进行恢复 IO 抑制或支持。如在高可靠性业务系统中，当存储节点出现多点故障，大量业务 IO 仍然占据主要 IO 资源。这个时候，按照 PG 粒度，根据 IO 命中情况以及数据冗余度危险情况，会自动将部分恢复 IO 标记为最高优先级处理，对业务高压力进行抑制，从而在数据安全和业务高负载的情况下做动态平衡。这种情况也适用于数据 GC IO 场景。

2.4.13. 高效存储介质管理

分布式系统对存储介质的空间管理设计了基于 KV 数据库的元数据管理和 IO 事务系统，区别于传统文件系统 BTree 或者 BITMAP 的持久化方式，基于 KV 数据库采用了 LSM 方式进行写入，同时由于元数据的修改通常都在几十字节到几百字节内，单独的元数据 WAL Log 会造成一定的写放大和 IO 资源浪费。分布式系统创新性的实现了元数据和数据 IO 合并的能力，对于随机 IO 的写放大控制在 10%-15%的损耗，相对于传统方式降低了 50%以上的写放大。



高效介质管理图

同时，针对当前越来越大的 HDD 和 SSD 容量，同时为了支持 4KB 粒度

最小写入单元，介质空间管理元数据随着介质容量从数 TB 到十几 TB 的增长而膨胀，空间管理元数据越来越难以全量放入内存中。分布式系统采用了两级元数据缓存管理，将所有在线数据（非快照数据）元数据并放入内存，将所有“冷 ”元数据聚合压缩写入 SSD 中，保证了存储引擎对于快照场景的性能支撑，又避免了元数据需要频繁 SWAP 的性能损耗。

2.4.14. 池级扩容及休眠技术

分布式系统应对海量数据的存储，超大规模数据的重平衡会产生规模效应，传统扩容模式不能有效的应对，分布式系统应提供了按存储池扩容技术，以及冷池休眠技术。存储资源池存储空间不足时，创建一个新的存储资源池，将新的资源池激活成活动资源池，按资源池粒度扩容。非活动资源池可进入休眠状态，不参与重平衡计算。进入休眠状态的资源池，在有数据需要访问时能够快速响应数据通过压缩处理，大幅降低存储空间消耗。配置存储策略，选择 EC 存储资源池存放数据，大幅提高存储空间利用率。

2.4.15. 业务连续性保障

副本、EC 、具有弹性高效的数据存取能力，采用去中心化的全分布式架构，通过横向扩展能线性增加整系统的容量和性能，系统可以轻松扩展至 PB ，甚至 EB。

通过拓扑规划功能对存储集群支持多种安全级别的数据安全保护，如支持硬盘、服务器级别、机架级别、数据中心级别的故障域。使存储系统可靠性及持续在线特性得到有效保证。

支持灵活的存储策略来保证存储稳定可靠性，包括可配置的多副本支持。副本数的增加可以提高数据的可靠性与并发访问的性能，支持 1-6 副本选择。

主副本存储策略，使数据就近访问，易扩展；副本数据强一致性，故障时无须切换。

2.4.16. 全局负载均衡，安全可靠

存储子系统架构分两层路由，一层在数据服务，每个网关服务节点承载前端服务请求，根据服务协议特点，进行数据请求均匀下发到不同的存储对象（分布式存储池的读写单元），确保不存在热点对象的设计，另一层在分布式存储池，尽可能的均衡存储空间，并加快故障下的数据重构速度。任何协议请求到数据服

务网关时，都需要转换为相应的存储对象请求，这个转换过程即存储服务负载均衡。而存储对象在存储池里如何分布来实现空间和性能的均衡的过程，即空间分布负载均衡。

2.4.17. 存储服务负载均衡

根据存储服务的协议平面不同，具体服务负载均衡方式会不尽相同。

存储协议的路由核心是针对卷空间的读写请求映射。为了让每个卷的数据均衡分布以及简化路由算法，会直接把每个卷按照固定的粒度(如 4MB)划分成若干个对象，每个对象按照“卷 ID+起始 LBA ”来计算哈希值，将每个对象落到某一个具体 PG 上。

对象存储和文件存储针对桶、 目录、对象和文件的元数据部分，都会存储到索引服务提供的键值存储中。而数据部分类似块存储的卷，将数据按照一定规则进行切片成对象，落到具体 PG 上。

2.4.18. 空间分布负载均衡

当存储对象本身的哈希映射到 PG 的过程实现均衡后，实现从 PG 到具体存储节点和介质的均衡是空间分布负载均衡的目标。

存储子系统采用类 DHT 算法，实现了 PG 到 OSD 服务的映射关系，分配算法保证了主副本和备用副本在不同服务器和不同硬盘上的均匀分布，即每个 OSD 上的主副本和备副本数量是均匀的。

在扩容节点或者故障缩容节点时，分配算法会考虑尽可能减少数据迁移的情况下，保证了系统中各节点负载的均衡性。数据能够均匀地分布到所有的节点中。当有新节点加入系统中，系统会重新做数据分配，数据迁移仅涉及新增节点，现有节点上的数据不需要做很大调整。在做数据分布计算时，算法是可以动态调整的，当系统中出现性能、负载不一致的节点时，算法可以根据调整输入参数优化算法，重新平衡负载。

2.4.19. 强一致性写入与掉电保护

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

如果系统中的某个硬盘出现短暂故障，存储系统会暂时不写这个硬盘上的数

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

如果系统中的某个节点在写入过程出现掉电，同样不影响任何时刻的数据一致性，因为所有写入都是经过持久化后才会与客户端进行确认。

2.4.20. 故障增量恢复和并行重建

当 HDD/SSD 或存储节点短时间离线（如盘误拔、存储节点重启）后又重接入系统时，存储系统不会进行全盘数据重构。所有 OSD 会按照 PG 粒度，逐步启动状态协商，主要是根据每次数据写入时附带的日志信息进行分析，对比新接入 OSD 的日志信息匹配，协商完成后使用该部分日志进行增量恢复。

当节点离线较久后重新接入时，增量恢复日志已经超过阈值，这时候会按照 PG 粒度进行全量恢复。由于数据通常都是分布到不同 PG 中，即副本或者 EC 条带数据会按照策略打散到不同的存储节点的不同硬盘，数据修复会在不同的节点上同时启动，每个节点上只需增量或全量修复一小部分数据，多个节点并行工作，有效避免单个节点重建大量数据所产生的性能瓶颈。

2.4.21. 自动化部署，运维敏捷，成本可控

自动化部署、一键安装、资源创建引导图标；图形化快速完成资源的基础部署。

一套存储资源池，提供块，文件，对象的数据平台支持能力，规模，高可用，业务连续性，弹性扩容，按需部署。全自动运维模式，容易上手，服务器级别运维，备件通用性强。

全系统选取高效能的双路多线程处理器、高效冗余芯片组、冗余热插拔风扇和硅脂散热片、冗余热插拔电源、冗余网卡等部件，提高系统的能效利用率，降低系统能耗。

2.4.22. 最大化存储介质，享受架构设计福利

海量数据，按照热、温、冷，按策略调整，使用最合适的存储介质保存相应的数据。降低数据管理成本，享受硬件福利。

2.4.23. 数据生命周期和业务需求，进行数据流动，数据分层

一份数据从写入存储，到频繁访问、偶尔访问、备份、归档，再到删除，行程整个生命周期过程，存储平台提供数据生命周期管理功能，是数据生命周期过程自动化，数据根据设置的策略，决定处于生命周期的阶段，从“热 ”到“冷 ”再到删除，无需人工干预，提升数据管理效率，也提升存储使用效能。

2.4.24. 动态接入网络，支持网络接口丰富

可以支持 1GE 、10GE 、25GE 、40GE 、100GE 、4GBFC 接口、8GB FC 接口、16GB FC 接口、32GB FC 接口、100GB InfiniBand 接口、200GB InfiniBand 接口等丰富的主机接口类型，满足高性能 IO 的需求。

2.5. 关键技术突破

2.5.1. 统一业务接口，统一命名空间

真正实现了统一存储功能，同时提供块、对象以及文件接口。基于块存储、文件存储和对象存储统一数据平台可以对接接入各类应用服务器，以及同时支持多套应用软件，共享访问全系统的存储空间。满足不同的应用场景需求，自由调配块存储、文件存储、对象存储的存储池划分。实现基础架构的融合统一。

分布式系统中的数据存储满足全局统一命名，海量的数据存储全局共享统一的存储信息，实现分布式系统提供全局数据存储访问服务。

分布式网关层提供两种网关类型：

分布式块网关：可以提供 SCSI 、iSCSI 、FC 以及 librbd 的块接口。对于 SCSI 协议支持完整的 SCSI-3 协议栈，兼容 HANA 、MSCS 等集群。

分布式对象网关：提供 S3 以及 NFS 接口。

企业级文件存储网关：提供 NFS 、CIFS 、FTP 、POSIX 等标准文件存储接口。

分布式系统块存储网关和对象网关都是无状态网关，横向扩展不受限制，可以随着集群规模的扩展近似线性的提升 IOPS 性能。

对象接口：提供 S3 对象接口。

文件接口：提供 NFS 、CIFS 、FTP 文件访问接口。

分布式系统提供统一平台为任何应用场景，任何部署架构提供完整的存储解决方案。其通过基于三层的分布式存储架构，支持文件、对象、大数据、块基于

同一个存储池提供服务，并进行统一管理。

分布式系统提供灵活的部署方式，允许每个节点根据存储使用规划进行角色定义，既可以作为存储节点，也可以作为监控、管理节点。

分布式系统通过构建分布式的系统设计，完整的企业级功能的架构简化。一个数据中心一套存储分布式系统提供了全面的标准协议存储，覆盖块、对象、文件、大数据场景，并提供服务编排和自动化能力。

分布式系统通过领先的软件架构分层设计，统一的组件通信协议和硬件自适应能力，加上核心组件以用户态形式运行，使得分布式系统可以运行在不同 CPU 架构、不同 OS 内核以及复杂网络环境下，解决了信创生态平台混搭、过渡和迁移等一系列问题和需求。分布式系统支持单个存储池跨 CPU 架构并提供高可靠服务的存储平台。

分布式系统实现了高性能 SSD 和大容量 HDD 的深度融合。所有业务随机小 块 IO 会写入多副本 SSD 层，然后通过创新的 IO 冷热感知聚合技术，将 SSD 层的随机写合并成顺序大块回写到使用 EC 冗余策略的 HDD 层，同时避免了 EC 冗余策略下的写放大问题。另外，SSD 层和 HDD 层的解耦设计，允许各自独立扩容， 并且 SSD 在故障情况下，换盘只在 SSD 层内部进行重构，完全不影响 HDD 层的数据健康。

分布式系统在架构上支持高性能文件系统和千亿级海量对象存储系统，为非结构化数据场景提供从核心业务、在线分析到离线备份归档的端到端场景覆盖。同时内置数据安全能力，为业务提供数据中心高可用，大大提升了业务可用性。

分布式系统在全局服务引导，关联信息检索， 自定义排序和搜索、系统亚健康管理等，都通过直观的 Web GUI 服务来提供，支持运维可视化，无需第三方软件或插件，即可支持同一 Web 界面管理实时存储系统和离线存储系统 2 套存储集 群用以高度自动化的存储资源配置的工作流。通过 Web GUI 服务可以管理不大于 20 套统一数据平台集群，满足长期多集群的统一管理；管理员可以监控到所有存储设备及其全方面性能和可用性指标，并调配相关资源分布，可以大大减少运维操作的开销，消除由于人为错误造成的停机时间。

2.5.2. 精简配置

提供了精简配置功能，为应用提供比实际物理存储更多的虚拟存储资源。相

比直接分配物理存储资源，可以显著提高存储空间利用率。

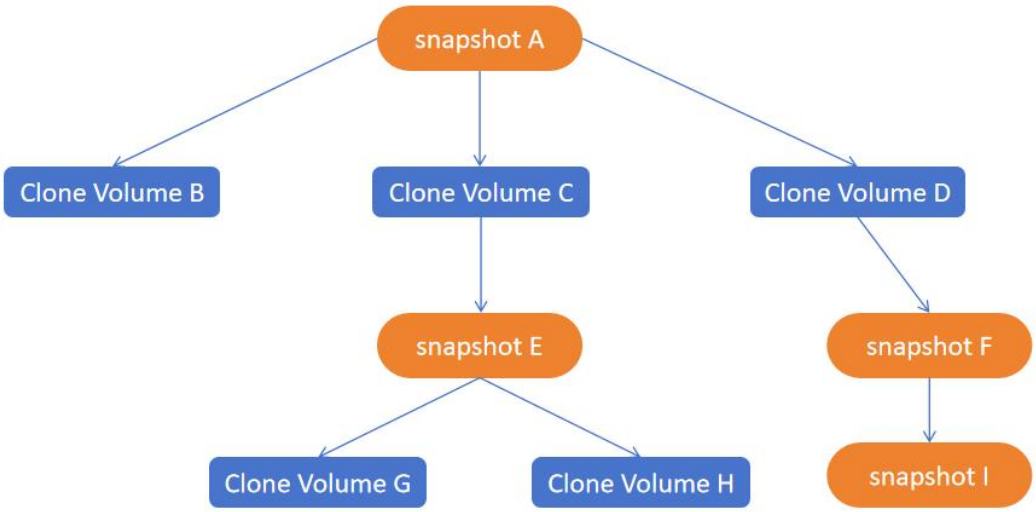
如果应用程序所使用的存储空间已满，就会崩溃。因此，存储管理员通常分配比应用程序实际需要的存储空间更大的存储容量，以避免任何潜在的应用程序故障。这种做法为未来的增长提供了“headroom ”（净空），并减少了应用程序出故障的风险。但却需要比实际更多的物理磁盘容量，造成浪费。

自动精简配置以“从一个存储池中按需提供存储给应用程序 ”作为基本原则。 存储管理员就可以像往常一样分配逻辑存储给应用程序，但仅在需要时才真正占用物理容量。当该存储的利用率接近预定阈值时，会自动从存储池中分配空间来扩展该卷，而不需要存储管理员的人工干预。因此应用程序认为它有充足的存储空间，但实际上并没有浪费存储空间。自动精简配置是一种按需存储技术，基本上消除了已分配但未使用的空间的浪费。自动精简配置可以提高存储空间的整体利用率。

2.5.3. 快照

快照是对指定数据集合的一个完全可用的拷贝，该拷贝包含源数据拷贝时间点的静态映像，用于生产测试、数据备份与恢复。快照技术是众多数据备份技术中的一种，其原理与日常生活中的拍照类似，通过拍照可以快速记录下在拍照时间点被拍照对象的状态。由于可以瞬间生成快照，通过快照技术，系统管理员能够实现数据的零窗口备份，从而满足客户对业务连续性和数据可靠性的要求。

提供了快照机制，将用户的卷数据在某个时间点的状态保存下来，后续可以作为导出数据、恢复数据之用。快照数据在存储时采用 ROW（Redirect-On- Write）写时重定向机制，更新源数据卷中的原始数据时，将源数据卷数据指针表中的被更新原始数据指针重定向到新的存储空间。



快照示意图

●文件定时快照

文件定时快照是基于底层块的快照来实现的，文件定时快照策略同卷定时快照，支持每天，每周，每月的定时快照策略，配置亦同卷定时快照，需要在数据保护中配置文件定时快照策略，再在文件系统中启用定时快照策略；当文件系统定时快照数达到最大规格数，会删除最老的快照，如果快照有链接克隆，为了保护克隆数据，该快照不会删除；当打定时快照的时候，系统中有其他冲突任务（如文件系统回滚），该快照会被跳过文件定时快照是一种数据保护方案，文件定时快照可以在文件系统遇到异常，故障或病毒等情况下，帮助文件系统恢复到没有出问题之前的某个状态。

●桶快照

桶快照管理；存储系统桶快照管理功能，支持桶快照创建、修改、删除，以及读取指定快照数据、获取快照差异信息等。

桶快照回滚；存储系统桶快照回滚功能，可以将桶中数据回滚到任一快照的历史状态。

●桶复制

基于数据集的复制策略：

存储系统可以根据前缀、后缀、元数据、标签、拥有者等过滤条件定义不同的数据集，每个数据集可以使用不同的复制策略，包括：

●可以设定一个或多个复制目标，复制目标可以是公有云存储、私有云对象

存储。

●删除对象时，可选择是否同步删除复制目标上的数据。

●选择数据复制范围，可以对包括历史存量数据在内的全部数据复制，也可以选择仅对新写增量数据复制。

可以指定在目标存储上的对象命名规则，比如叠加前缀。

当启动 ROW 快照后，系统会创建一个和源卷对应的快照空间和索引日志，在创建 ROW 快照后，源卷会保持初始状态不变，如源卷中的数据需要修改，新数据会被写入对应的快照空间中，同时该数据位的指针会重定向到新的写入存储空间地址位，这样当读这个数据时会按指针位置找到新更新的数据。

对比 ROW 快照和 COW 快照，可见 ROW 快照写入数据的效率更高，不会导致快照后源卷写性能下降。

如图：a 、b 、c 、d 、e 、f、g 、h 、i 分别是地址位 0-8 的数据块中的数据，在未启动快照前把 p 写入地址位块 1 中，这时系统会把块 1 中原有的数据 b 删除，写入 p 。启动快照功能后，ROW 都会建立独立的快照空间存放索引和日志，这时在地址位 6 写入 z。

●ROW 快照：

直接将地址位 6 和数据 z 写入快照空间，原数据 g 保持不变；

记录地址位 6 数据发生改变，改变目录索引。

分布式系统的 ROW 快照技术，极大地改善了快照后的存储性能。

2.5.4. 一致性组快照

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

2.5.5. 链接克隆

提供链接克隆机制，支持基于一个卷快照创建出多个克隆卷，各个克隆卷刚创建出来时的数据内容与卷快照中的数据内容一致，后续对于克隆卷的修改不会影响到原始的快照和其他克隆卷。

克隆卷继承普通卷所有功能：克隆卷可支持创建快照、从快照恢复以及再次作为母卷进行克隆操作。

2.5.6. 多资源池

分布式系统使用不同性能存储介质以及故障隔离，支持多资源池特性。不同的资源池可以提供不同的性能以及不同的副本策略。

分布式系统使用相同的一组OSD 创建多个资源池，资源池的类型、副本数量可以各自不同，这些共享OSD 的资源池形成一组关联资源池，它们共享分配使用磁盘的裸存储空间。

分布式系统SSD 盘建索引池和HDD 盘建数据的部署方式。分布式系统元数据存放在SSD 盘，小对象数据先存SSD，后经文件归并形成大文件存入HDD。

●混合盘场景

在海量规模、小文件性能较好的场景，使用混合盘部署分布式系统。通常SSD 盘和 HDD 盘的数量比是 1:5 ，将 SSD 盘划分 5 个分区，每个分区和一块 HDD 盘组成一个混合盘，基于混合盘创建的 OSD ，omap 分区和缓存分区使用 SSD ，数据存储使用 HDD。在混合盘 OSD 上创建一个副本类型的索引池和一个 EC 类型的数据池，对象元数据都存储在 SSD，大文件数据利用缓存大块 bypass 特性直接写到HDD 盘，小文件数据先写入 SSD 缓存，而后归并成大文件后写入数据池。

●全 HDD 场景

在海量对象规模存储、低成本、小文件性能无要求的场景，使用全 HDD 部署 对象存储，基于同一组 HDD 盘创建一个副本类型的索引池和 EC 类型的数据池， 二者成为一组关联资源池。小文件数据先写入副本类型的索引池，而后通过文件归并形成大文件存入 EC 类型的数据池。

●全 HDD 小文件写性能优化场景

全 HDD 场景部署模式下，小文件直接写HDD 盘，而 HDD 盘随机小IO性能差。在全 HDD 部署基础上，为了提升小文件写性能，可以采用少量 SSD 盘来创建副本类型的资源池，用做缓存小文件数据，小文件会先写入 SSD 资源池，而后经文件归并形成大文件后存入 HDD 数据池。

●全 SSD 场景

在海量小文件存储、对小文件读写性能都有比较高的需求的场景，使用全SSD 部署对象存储。基于 SSD 创建一个副本类型的索引池和一个 EC 类型的

数据池，提升 SSD 存储空间利用率，兼顾读写性能和存储成本。小文件数据先写入副本类型的索引池，而后通过文件归并形成大文件存入 EC 类型的数据池。

2.6. 业务 QoS

业务 Qos 功能，支持在线设置最大 IOPS 、突发 IOPS 、最大带宽、突发带宽， 保证核心业务的优先策略。技术上采用漏桶与令牌桶相结合的 IO 流管理策略。

对于分布式系统，数据按时间顺序排队（IO Queue）进行写入。这些 IO 可以划分为两大类，一类是客户端过来的业务 IO；另一类是系统内部活动产生的 IO ，包括副本复制、Recovery 和 SnapTrim 等。QoS 可实现系统资源的合理分配，平衡业务 I/O 和系统内部 I/O 的性能需求。

业务 QoS 支持在线设置存储卷的最大/突发 IOPS 、最大/突发带宽，可以通过 设置业务 QoS ，实现不同存储卷性能的量化管控，进而应对多样化业务对性能的需求。采用漏桶(Leaky bucket)和令牌桶(Token Bucket)相结合的 IO 流管理策略：漏桶算法能够强行限制数据的传输速率；令牌桶算法能够在限制平均传输速率的同时允许一定的突发传输。

●漏桶算法

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

●令牌桶算法

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

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

2.6.1. Recovery QoS

分布式系统硬件异常时，或者进行硬件更换维护时，分布式系统会在超过硬

盘离线时间（系统默认 180 分钟，可以在界面配置）后进入 Recovery 状态，将失效硬件上的数据重新分布在其他节点，此时将产生 Recovery IO ，当 Server3 异常离线时，Server1 和 Server2 的 PG 除了处理业务 IO 之外，还需要处理 RecoveryIO ，此时，如果不进行控制，应用层发起的业务 IO 将会受到很大的冲击。 Recovery Qos 能对 Recovery IO 进行控制，并制定策略，根据用户需求保证业务 IO 或 Recovery IO 正常进行。

在存储池的范围内设置 Recovery QoS ，根据不同的业务类型设置不同的策略：

●静态设置 QoS

低速恢复：数据恢复带宽为单 OSD 10MB/s；低速恢复优先保证业务带宽，恢复时间相对较长，长时间恢复过程中再有硬件故障可能会降低数据安全级别。

中速恢复：数据恢复带宽为单 OSD 20MB/s；中速恢复保证业务和恢复带宽同等优先级，恢复时间中等，在性能饱和情况下可能会增加 I/O 延时。

高速恢复：数据恢复带宽为单 OSD 50MB/s；高速恢复优先保证恢复带宽，恢复时间相对较短，在性能饱和情况下可能会影响客户端 IO 性能。

当集群硬件异常需进行数据重平衡操作时，会产生 Recovery I/O ，分布式系统数据重平衡进程拥有独立的工作队列和线程池，数据重平衡 QoS 会对两个参数进行控制:

●控制不同 PG 重平衡的优先级 控制重平衡的 PG 并发数目

控制不同 PG 重平衡的优先级:瓣先将 PG 加入优先级队列，再根据 PG 的优先级从高到低的顺序出队列，优先级高的 PG 先进行重平衡。

控制重平衡的 PG 并发数目:瓣数据重平衡过程一般会有客户端业务数据写入，大量 PG 参与数据重平衡及业务写入操作，往往会导致业务 I/O 需要等待、延时增大，最终影响业务性能。用户设置静态 QoS 后，数据恢复 QoS 会根据已有数据量、异常类型、冗余策略等信息，根据 QoS 策略智能控制参与数据重平衡的 PG 的数目。

2.7. 业务在线迁移

提供在同一集群内部做业务的在线迁移工作，为了保证存储资源的利用率，在不影响系统上业务的情况下，可以将 SSD 性能池中的数据迁移至 HDD 容量

池中，同时支持副本池数据迁移至 EC 池中。

2.8. 按需定义的异步复制

远程异步复制是容灾备份的核心技术，同时也是保持远程数据同步和实现灾难恢复的基础。它可以将数据备份到异地容灾站点，以防主站点由于各种突如其来、不可预知的原因导致的数据不可访问。对于配置了远程异步复制任务的复制主从集群，当主集群系统故障或者用户进行主集群系统维护升级、临时下线等操作时，可以通过远程异步复制提供的基本操作，使用远程异步复制任务的从集群为上层业务提供服务，减少用户业务应用的中断时间。

支持目录级别的异步远程复制，支持 1 对 1、1 对多、多对 1、双向复制，保证多套存储集群间容灾功能；数据复制：支持非结构化数据复制、Qos 、回收站、灾难演练、10 分钟级 RPO 、丰富复制策略、目标端快照保护、NFS alc 拷贝等功能。

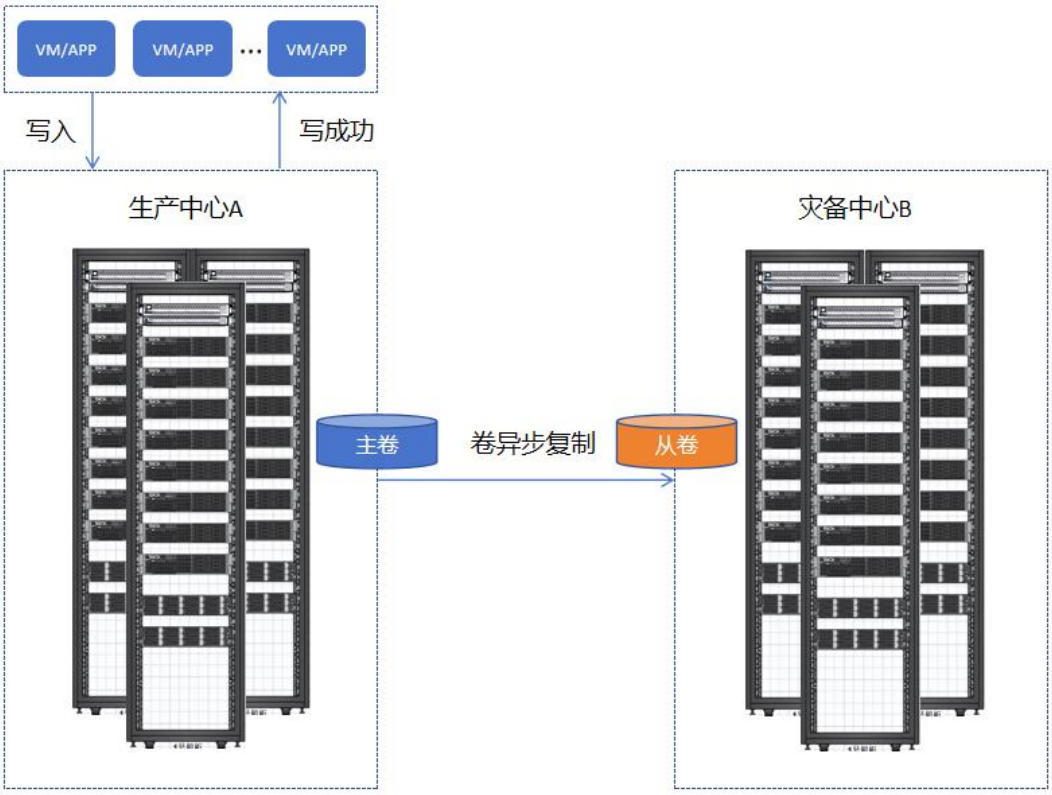
●数据异步复制

在远程异步复制任务中将主集群卷和从集群构成复制策略，主从复制目标分别位于不同的存储集群。

创建远程复制任务后，系统对主集群卷执行快照，作为初始同步的基准快照。根据指定远程异步复制策略的定时同步方式，系统基于主集群卷的快照，由主集群的保护网关将快照时刻的数据导出复制到从集群，当从集群目标卷数据与主集群卷的基准快照一致时，则复制完成。同时支持自定义主集群和从集群中的保留快照数，以减少容量占用。

如果在某一时间段，用户不希望数据由主端目录同步到从端目录，例如，链路带宽不足或对关键业务有影响，可暂停链路上远程异步复制数据的同步。

以容灾场景为例，生产站点的数据周期性地从生产站点复制到灾备站点，当主站点发生火灾、水灾或者地震等区域性灾难的时候，业务从生产站点切换到灾备站点。如图：



异步远程复制架构图

●以存储桶为粒度，可以按需将数据同步到指定的数据中心，在本地数据中心就近读、写数据。

●站点跨区域部署，实现跨区域数据共享，一个桶可以跨十个站点做数据同步。

●桶在站点间的数据互为备份，实现数据异地容灾功能。

●Zone 故障迁移：

当非 Master Zone 发生不可用时，不会影响整个 Region 的可用性。

当 Master Zone 发生不可用时，切换 Master Zone ，需要重启 Master Zone 的 RadosGW 来实现 Master Zone 接管。

●数据流：

多个存储站点通过异步复制的方式复制数据，文件写入时会将数据和元数据先写入本地站点的存储，写入成功后即返回客户端写入成功，本地站点会通知其它站点拉取新增数据，最终完成站点间数据同步。

2.9. 海量小文件合并技术

提供小文件合并功能，将小文件合并成大文件再存储到系统中，能从容面对

海量小文件（每日千万个文件出入）存储需求：

●配置存储策略，小对象存入高性能副本存储池，大对象存入 EC 存储池。

●配置存储策略，开启合并归档功能。

●后台自动启动小对象合并归档任务，将存放在高性能副本池中的小对象合并成大对象后存入 EC 存储池。

存储系统支持归并空间空洞回收功能，小文件归并后，因部分小对象被删除而产生的归并空间空洞，可以主动回收，避免空间浪费。

●支持设置空洞率阈值，空洞率超过阈值后，系统自动回收空洞空间。

●支持按指定空洞率做可回收空洞空间预估。

在达到自动回收空洞空间阈值前，支持人工触发空洞空间回收。

2.10. 存储 WORM 写保护

WORM 是指一次写入，多次读出的技术。数据写入提交后，在设置的文件保护期之内，不能够对文件再次进行修改、删除、移动，但是可以多次读取。

存储系统桶 WORM 写保护功能和存储系统文件 WORM 写保护功能，支持配置和修改 WORM 锁定期和缺省写保护模式，包括监管模式、法规遵从模式。

存储系统的 WORM 特权用户功能，特权用户仅对处于监管模式下的对象具备相关特权，包括：修改 WORM 保护期和保护模式；对受监管模式保护的对象做修改、删除操作。

2.11. 池级扩容及休眠技术

提供了按存储池扩容技术，以及冷池休眠技术：

数据通过压缩处理，大幅降低存储空间消耗；

配置存储策略，选择 EC 存储资源池存放数据，大幅提高存储空间利用率；

存储资源池存储空间不足时，创建一个新的存储资源池，将新的资源池激活成活动资源池，按资源池粒度扩容；

非活动资源池可进入休眠状态，大幅降低电力、制冷成本；

进入休眠状态的资源池，在有数据需要访问时能够快速响应。

2.12. 数据生命周期管理

存储桶数据生命周期管理支持，可以对存储桶内的数据通过数据前缀或整桶进行归档和删除，支持实时归档及延时删除。通过数据生命周期管理功能，用户可以更加高效的定义数据生命周期，使存储的利用更加高效。

●生命周期数据流动

对象在集群内数据存储冷热分层；存储系统可以根据前缀、后缀、元数据、标签、拥有者等过滤条件定义不同的数据集，将数据集中的对象数据根据冷热程度在集群内不同数据存储类别间流动。

从内部存储同时分层到不同的次级存储；存储系统支持根据前缀、后缀、标签、元数据、拥有者等过滤条件定义数据集，将数据集中的对象同时分层到不同的次级存储平台。

从外部次级存储分层流动到集群内部数据存储；存储系统支持根据前缀、后缀、标签、元数据、拥有者等过滤条件定义数据集，将存储在外部次级存储平台的数据，分层流动到集群内部数据存储平台。

分层后数据解冻还原；存储系统中对象分层到次级存储平台后支持数据解冻还原到本地缓存，在缓存有效期内高效访问。

同时归档到不同的次级存储；存储系统支持根据前缀、标签、元数据、拥有者等过滤条件定义数据集，将数据集中的对象同时归档到不同的次级存储平台。

●生命周期过期删除

生命周期过期规则；存储系统可以根据前缀、后缀、元数据、标签、拥有者等过滤条件定义不同的数据集，对数据集设置过期时间，支持按设定天数过期，也支持按指定时间点过期，到期后自动删除对象；多版本对象支持当前版本和历史版本分别过期删除；支持对未完成分段上传的对象过期删除；支持对象可保留历史版本的最大数量设置，超过限定数量，系统会自动删除最靠前的历史版本。

●对象数据重删

对象数据重删；存储系统的对象数据重删功能，支持相同数据只存储一份，节省存储空间。

●对象数据更新

对象追加写；存储系统的对象追加写功能，支持在对象尾部追加写入数据。

对象局部更新写；存储系统的对象局部更新写功能，支持对指定位置的对象

数据做局部更新。

对象秒合；存储系统的对象秒合功能，可以按指定顺序将多个对象快速合成为一个大对象。

对象软链接；存储系统的对象软链接功能，支持基于已经存在的对象创建软链接，通过软连接对象访问链接目标对象数据。

●回收站

对象回收站；存储系统对象回收站功能，对象删除后会先进入对象回收站，可以从回收站中恢复被删除的对象；可配置对象在回收站中停留的时间，到期后从回收站中删除对象。

桶回收站；存储系统桶回收站功能，桶删除后会进入桶回收站，可以从桶回收站恢复被删除的桶。

●元数据查询

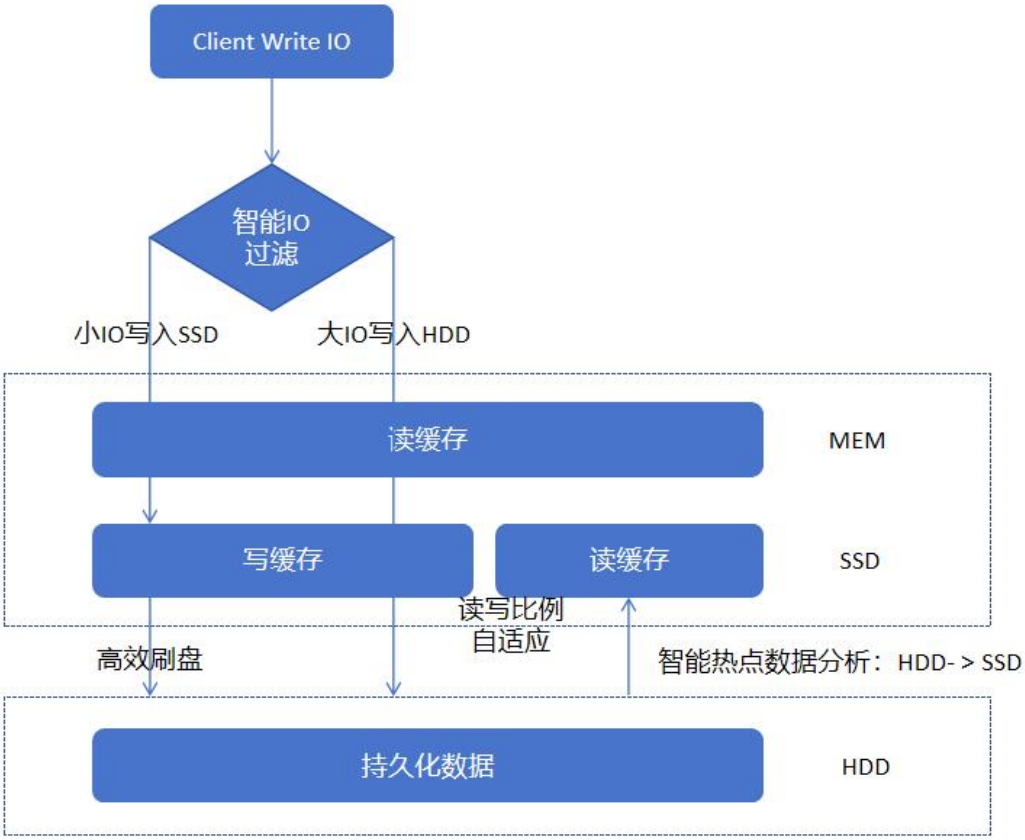
存储系统的对象元数据查询功能，支持查询字段定义，动态定义对象元数据、标签的数值类型；基于查询字段执行多查询字段、多组合条件的查询检索。

2.13. Cache 缓存技术

基于 Ram ，SSD 多级 Cache ，提升存储在多负载情况下的业务识别和靶向加速能力。为了能准确识别业务模型，Cache 以策略驱动的方式设计，将“机制 ” 和“策略 ”分开实现。Cache 并非在单一模块实现，而是联动了客户端和持久化层，实现 IO 路径端到端的加速，包括 RAM ，SSD 读写缓存机制，提供面向各类业务的加速策略。配合高效的合并刷盘算法，保证产品在各类业务下的性能提升。

通过创新性的 IO 聚合，将 SSD 缓存层的随机写转换为顺序写到 HDD 数据层，解决了高负载情况下的刷盘场景下业务性能问题。在 SSD 缓存被打爆的情况下，依旧能满足客户需求。

整体框架采用 Tier 分层模式，缓存层由高速介质组成，以三节点起，每节点 最少一块盘，数据层由低速介质组成，以三节点起，如果数据层采取副本方式存储数据，则每节点最少一块盘，如果数据层采取 EC4+2:1 方式存储数据，则每节点最少两块盘，存储两份数据。



多级缓存

●IO 写流程如下：

写请求通过数据路由算法寻址主 OSD；

主 OSD 收到请求并进行合法性校验后，立刻分发给副 OSD；

主 OSD 会将元数据写入元数据区(KVDB)，数据写入 Cache 当本地数据和元数据都写完成后，同时副 OSD 也反馈成功，这时主 OSD 返回前端写成功；

●IO 读流程如下

读请求通过数据路由算法寻址主 OSD；

如果数据读 IO 命中内存缓存，则从 MEM 中读取数据后返回；

如果数据读 IO 命中 SSD Cache，则从 Cache 中读取数据后返回；

都不命中，则从 HDD 中读取数据后返回；

主 OSD 返回前端读成功。

●读缓存热点识别

支持 SSD 读缓存功能，提升随机 IOPS：为了提升随机小 IO 的并发访问

能力，缓存算法会进行热点数据识别，在数据读取过程中判断热点数据，将数据长久保存在 SSD 中，随着时间的推移，算法会配合 SSD 使用率，动态的将“冷 ”数据回刷至 HDD 中，将“热点 ”数据持续更新在 SSD 中，提升数据库持续读取性能。分布式读写缓存算法，支持 SSD 读写缓存功能，在 SSD Cache 的支持下可以使用户获得更高的性能支撑。支持添加介质时配置读缓存大小。

●IO 流预读

在面向业务端并发读取多流数据（多个并发大块顺序读），存储端收到的 IO 在某个时刻呈现随机分布状态，缓存算法的流预测技术，可以准确的对访问进行识别和预测，针对每一路的访问提前预取。流预测策略和 RAM 读缓存机制配合使用，提升业务场景并发顺序读取性能。

2.14. S3/NFS 互操作

支持将对象存储桶通过 NFS 网关以目录的形式挂载给 NFS 客户端，NFS 客户端可以访问桶中已经上传的对象文件，可以在桶中创建子目录、写新的文件、删除文件或目录、文件追加写、文件内容更新、列表查看各级目录等，兼容各种常用的 POSIX 接口；通过 NFS 客户端写到桶中的文件，可以使用 S3 API 直接下载。

主要应用于符合 S3 的一次性写、多次读的 IO 访问模型场景，以及低频度文件内容更新的场景。

●应用从文件访问方式向对象访问方式平滑演进

应用投入 S3 开发过渡阶段需要文件、对象两种访问方式共存；或者先用文件接口对接现存应用，支持 S3 的应用完成开发后可以直接切换到对象访问已经写入的文件，不需要做文件到对象的数据迁移。

●文件接口的备份归档软件对接

传统备份归档软件主要是支持文件接口，采用文件方式备份数据，备份数据可以享受对象存储的低成本、弹性伸缩、生命周期管理等优势。

●对象内容局部更新

标准的 S3 API 并不支持追加写、局部内容更新，使用 NFS 访问可以弥补标 准 S3 API 的这些不足。

●多协议访问

NFS 文件访问；存储系统的 NFS 接入访问功能，支持 NFS 追加写、局部更新写、Truncate；支持创建软链接；支持文件和目录重命名。

HDFS 访问；存储系统的 HDFS 接入访问功能，支持 Hadoop 应用对 XEOS 对象存储的访问，包括读取对象、写入对象、删除对象等。

●多协议互操作；存储系统 S3/NFS/HDFS/CIFS 协议互操作功能，任一协议写入的数据，都可以通过其它协议访问。

2.15. 配额管理

由于是多人多任务的环境，所以会有多人共同使用一个硬盘空间的情况发生，如果其中有少数几个用户大量的占掉了硬盘空间的话，那势必压缩其他用户的使用权力。因此应该适当的限制硬盘的容量给用户，以妥善的分配系统资源。

●用户配额
存储系统支持用户配额管理功能，可控制的配额包括：

- 存储数量配额
- 数据容量配额
- 总数据容量

●数据存储类别的数据容量

- 次级存储类别的数据容量
- 文件数量配额
- 总文件数量
- 数据存储类别的对象数量
- 次级存储类别的对象数量
- 缺省配额
- 数据容量配额
- 总数据容量

●数据存储类别的数据容量

●次级存储类别的数据容量

- 对象数量配额

●总对象数量

●数据存储类别的对象数量

- 次级存储类别的文件数量

●文件配额

测试存储系统支持文件配额管理功能，可控制的配额包括：

- 数据容量配额
- 总数据容量
- 数据存储类别的数据容量
- 次级存储类别的数据容量
- 数量配额
- 总对象数量
- 数据存储类别的文件数量
- 次级存储类别的文件数量
- 配额超标处理模式

●存储系统在配额使用达到上限时的差异化处理功能，可以选择配置不同的处理模式，包括：拒绝新的数据写入

●按存入时间先后顺序，删除最早存入的数据

●支持对用户、用户组、 目录设置配额；配额类型支持统计配额、限制配额；

统计配额仅监控存储的使用情况，不限制使用；限制配额监控存储使用情况的同 时并限制使用，超出阈值告警。

●灵活定义用户权限和配额限制，不同的用户拥有不同的操作权限。同时支持自定义容量配额。可以设置软配额和硬配额两种方式。硬配额是空间使用容量一旦超出硬配额设定范围，立刻停止写入。软配额是空间使用容量超出软配额设定范围，系统会产生告警，但不会限制数据写入，只有在超过硬配额之后，系统才会立即停止写入。

2.16. 分级存储

可以结合生命周期管理功能，设置数据冷热分级流动策略，实现数据自动分级；可以延长热点资源的存续时间；可以提高热点资源的使用等。

●多存储类别

●存储系统支持定义 4 种以上不同级别的存储类别，用于存放不同活动程度 的数据。可以实现分级到硬盘、蓝光库、磁带库等存储介质。

●客户端选择使用不同存储类别

●存储系统支持客户端上传对象时按需选择使用不同的存储类别，将数据存放到对应的资源池内。

●存储类别按规则自动匹配

存储系统支持按存储服务端配置的规则使用存储类别，匹配某条规则的对象，将数据存放到对应存储类别的资源池内；没有匹配到任何规则的对象，可以拒绝写入，也可以选择写入缺省存储类别。

●存储类别使用模式优先级

存储系统可以同时支持两种存储类别使用模式（客户端指定和存储服务端规则匹配），并可配置二者的使用优先级。

●基于既定策略将文件迁移到特定存储介质上。分级策略包括文件名、文件大小、修改时间、访问时间、元数据修改时间进行数据迁移。

●可以根据数据性质选择数据持久化存储级别，如：重要的数据用高安全级别的存储类型，非重要数据使用低安全级别的存储类型；性能要求高的数据存高性能存储池；需要提供存储空间利用率时选择 EC 存储。

2.17. 用户权限

配置权限管理，权限管理根据共享协议的不同，有不同的权限管理。

●用户 ACL

存储系统的用户 ACL 访问权限控制功能，支持授权用户访问资源的相关权限。

●用户权限控制策略

存储系统的用户权限控制策略(user policy)功能，通过不同的权限策略，授权或禁止用户对满足特定条件的资源所具备的访问操作权限。

●文件权限

文件 ACL；存储系统的文件 ACL 访问权限控制功能，支持授权不同用户、用 户组访问文件的相关权限。

●文件继承桶 ACL 权限

存储系统的文件继承桶 ACL 权限功能，开启该功能后，未设置 ACL 的对象， 可默认继承文件 ACL 权限。

●权限控制策略

存储系统的权限控制策略(bucket policy)功能，通过不同的权限策略，授权或禁止一个或多个用户对满足特定条件的资源所具备的访问操作权限。具备数据的安全权限管理、用户权限管理功能，对访问存储系统的用户实现权限管理与控制，可支持不同用户对存储数据的读、写、删、查等不同权限的灵活组合。

2.18. 存储加密压缩

支持在线压缩、加密功能，通过 GUI 界面开启关闭压缩加密功能。在 IO 实时读写的过程中通过网关层进行处理队列，压缩加密的粒度为：副本是 4M 粒度,EC 是根据数据块数量动态决定（4*k）M 。并且每次加入处理队列后会检查处理队列中处理情况，并做相应操作。开启压缩加密后只有数据压缩加密完，才会将数据写下去，否则继续接受数据处理。

●队列处理并发为 4 个

压缩加密的处理顺序为先压缩再加密，最终落盘

● 自适应压缩

对于可压缩的数据，在 IO 聚合模块 LogAppend 中，会对数据进行一次压缩，压缩算法、压缩标志和原始数据长度和压缩后长度会记入元数据中。压缩功能减小了后端的容量占用和减少对后端的 IO 写入次数，为后端节省空间和减轻磁盘压力。由于压缩功能是后台行为，而且基本不涉及额外的 IO 操作，因此在结构化数据场景时，大部分情况下可以提升性能。

压缩算法默认采用的是 snappy ，可以支持的算法:snappy ，zlib ，zstd ，lz4。

存储池压缩算法的修改不会影响已经压缩的数据，已经压缩的数据在读取时会根据压缩时的算法进行解压。

2.19. 访问日志

分布式系统可以开启访问日志功能，可以详细记录每次资源访问请求及处理结果的详细信息，包括请求时间、桶名、对象名、发起请求的 IP 地址、请求操

作类型、请求 URI 、响应码、对象大小、访问的字节数、响应时长等。这些字段信 息按照 S3 协议的标准日志格式形成一条纯文本日志记录，一组日志记录形成一个日志对象持久化存储到存储桶中。经过授权的用户，可以从桶中下载日志对象，通过日志解析工具分析所有记录下来的日志信息。

配合多租户管理，可以支持三权分立：业务管理员 A 创建了存储桶 bucket-1，并授权 B 用户使用，同时给该桶配置桶访问日志功能，授权审计人员 C 对产生的 日志具有下载权限。

支持 S3 的 PUT Bucket logging API 和 S3 的 GET Bucket logging API

●应用场景

标准的日志格式，可以直接对接开源生态中的日志分析平台。比如可以采用 ELK 专门提供实时日志分析平台，通过 Logstash 提供的数据采集、转换、优化和输出能力，Kibana 提供的强大可视化管理界面，以及 Elasticsearch 提供的实时分布式搜索和分析引擎，对日志信息进行实时分析、了解业务情况以及用户行为。

●审计

通过桶访问日志分析特定的访问者对某些资源的访问和操作历史，了解业务请求以及用户行为，起到事后审计和追踪的作用。

●热点分析

通过桶访问日志分析某些资源被访问的频度，识别热点资源，为策略驱动模型提供决策依据。例如，可以结合生命周期管理功能，设置数据冷热分层流动策略，实现数据自动分层；可以延长热点资源的存续时间；可以提高热点资源的使用费用等。

2.20. 并行处理能力

海量存储节点分布式文件系统，将多个存储节点上的磁盘组织成为一个逻辑上统一的磁盘供业务应用使用。MPI I/O 提供两种类型的接口:Independent I/O 和 Collective I/O 。Independent I/O 中，每个进程独立地处理 I/O ，Independent I/O 基本的操作是，MPI_File_read()和 MPI_File_write()。

分布式系统通过标准的 NFS 、SMB 、POSIX 接口协议，支持 MPI-IO ，兼容 OpenMPI 、MPICH2 等。

整个 IO 栈中的最底层的接口是 POSIX 接口，涉及最基本的文件操作，如 open 、close 、read 、write 、stat 等 。POSIX HPC IO 扩展的设计目的 ，是提高 POSIX 在大规模 HPC 环境中的性能，大多数普通应用程序的 API 都是在 POSIX 语义之上构建的。

当进程数量少且每个进程操作的数据量较大时，使用 FPP 策略；

当数据量较小时，使用一次写单个共享文件策略；

当每个进程操作的数据量较大且进程数与 OST 的比值较小时，使用同时写单个共享文件策略；

当进程数较多且数据量较大时，使用划分子集写共享文件的策略，并选择好子集的数目。

将一次 I/O 操作分成两个阶段来完成。第一个阶段是所有计算进程交换彼此 I/O 信息，确定每个进程的文件域，从磁盘上读取相应 I/O 数据；第二个阶段是按照应用需求将缓冲区的数据交换分配到各进程的用户缓冲区，写操作与读操作类似。在这两个阶段中，进程之间需要进行相互通信，同步同一通信域中的相关进程的数据信息，从而得到程序总体的 I/O 行为信息，以便整合各个进程独立的 I/O 请求。

第一阶段：

(1) 确定参加读操作的通信域中所有进程，并进行同步；

(2) 确定每个进程各自需要读取的一个包含所请求数据的连续的文件区域；

(3) 每个进程根据自己的文件区域进行 I/O 访问，并将这些数据存放在一个临 时缓冲中，同步所有进程到每个进程的 I/O 操作结束。

第二阶段：

(1) 进行 I/O 数据的置换，将缓存中数据发送给目标进程；

(2) 同步所有进程到数据置换结束。

通过多级缓存加速、IO 聚合顺序写入，改进集中式操作和直接 I/O 技术，能够有效解决 I/O 瓶颈，提升并行计算速度。

2.21. 开放数据处理框架

●任务管理和调度

开放数据处理框架支持数据处理任务的创建、管理、调度、执行功能。

- 自定义数据处理组件

开放数据处理框架扩展能力，支持按需定义或引入第三方数据处理组件，灵活拓展数据处理服务功能。

●打包上传

开放数据处理框架支持打包上传功能，客户端将多个文件打包压缩成单一大文件，压缩文件上传到存储桶后，存储系统通知开放数据处理框架执行解压缩和文件存储任务，将解压出来的文件存入到指定位置。

●图片处理

存储系统支持图片处理服务，支持图片旋转、放大、缩小、裁剪、圆角矩形、模糊、锐化、明暗对比度调节、添加文字/图片水印、质量转换、格式转化等图片处理功能。

3. 软件系统设计

3.1. 统一存储架构

分布式存储系统内置分布式存储为虚拟化业务提供弹性精简配置的强一致性块 存储服务，内部采用独特的并行架构、创新的缓存算法、自适应的数据分布算法，既消除了热点也提高了性能，并且能够以超快的重建时间实现自动化自修复，提供卓越的可用性和可靠性。

线性扩展和弹性：分布式存储系统的分布式存储采用全分布式数据路由架构，将所有元数据按规则分布在各节点，避免了元数据瓶颈，支持线性扩展。分布式存储采用了分配、传输等长的数据分块切片技术，配以基于一致性 HASH 的数据路由算法 Crush，可以将虚拟机镜像的所在卷块数据均匀的分散到较大的资源池故障域范围内，使得每个卷可以获得更大的吞吐量(IOPS)和带宽(MBPS)的性能，也使得每个硬件资源的负载相对均衡。

高性能：分布式存储系统分布式存储免锁化调度的 I/O 软件子系统，彻底解决了分布式锁冲突，使得 I/O 路径上无需进行任何锁操作和元数据查询，IO 路径短、时延低；同时分布式存储系统分布式存储采用分布式 NVMe Cache 技术，配合大容量的 NVMe 盘做主存，使得系统的性能可以具备 NVMe 的性能和 SAS/SATA 的容量。

高可靠性：分布式存储系统分布式存储支持多种数据冗余保护机制，如 2 副

本、3 副本、纠删码等，在此基础上，分布式存储支持设置灵活的数据可靠性策略，允许将不同的副本放在不同的服务器上，保证在服务器故障的情况下，数据仍然不丢失、仍然可访问。同时采用对有效数据分片进行数据的冗余保护，在硬盘、服务器故障的时候，能够对有效数据进行并行重建，1TB 硬盘的重建时间小于 30 分钟，大大增强系统的可靠性。

分布式存储采用分布式集群控制技术和 Crush 路由技术，提供分布式存储功能特性。

名称	统一说明
协议层	通过 RBD/iSCSI 驱动接口向 QEMU 提供逻辑卷设备作为虚拟机镜像。
服务层	提供各种存储高级特性，如卷快照、快照克隆、卷精简配置、分布式缓存等存储业务功能。
引擎层	提供分布式存储基本功能，包括集群管理状态控制、分布式数据路由、强一致性复制技术、集群故障自愈与并行数据重建等技术。
持久层	实现智能 OSD 设备，持久化 DB 存储功能，保持全局事务一致性。
OS 层	为分布式存储系统提供 LIBAIO、epoll 等关键操作系统组件。
驱动层	为分布式存储系统提供多种硬件能力支持。

3.2. I/O 流程

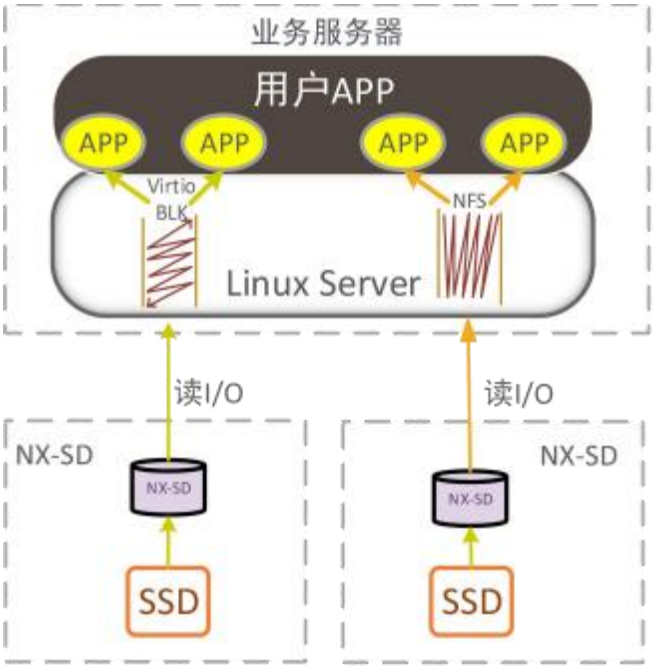
(1) 数据路由映射方法

应用及前端负载均衡->后端容量均衡？

(2) 数据 I/O 路径方法

①、读 I/O 路径

分布式存储系统分布式存储系统中的读 I/O 流程如图所示。

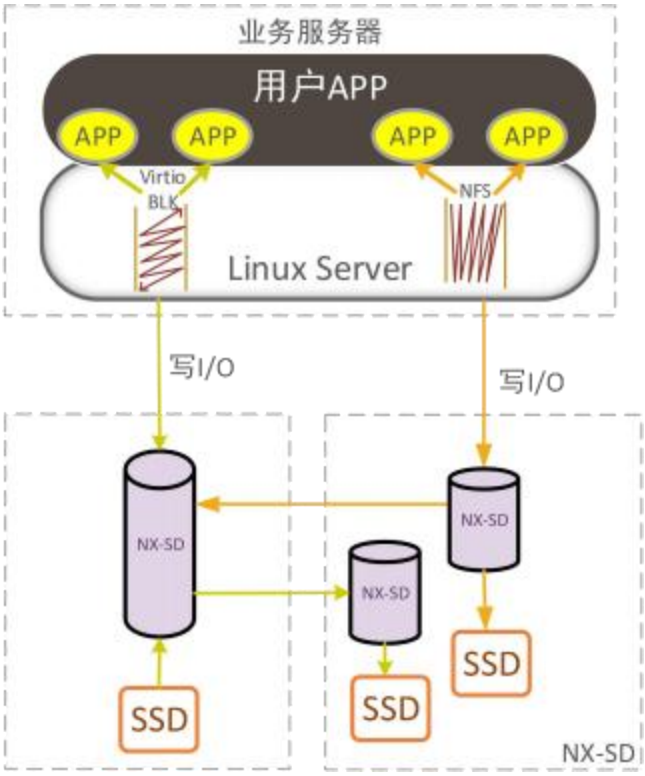


分布式存储系统 分布式存储读 IO 流程

APP 下发读 I/O 请求到虚拟机 OS ，虚拟机 OS 通过 Virtio 转发该 I/O 请求到本服务器的 librbd 模块。librbd 根据读 I/O 信息中的卷和 LBA 信息，通过数据映射机制确定数据所在的 主 OSD 。如果此时主 OSD 故障，librbd 会选择备 OSD 读取所需数据。

②、写 I/O 路径

分布式存储系统分布式存储系统中的写 I/O（2 副本）流程如图所示。



分布式存储系统虚拟机写 I/O（2 副本）流程

APP 下发写 I/O 请求到客户端 OS ，客户端 OS 通过 Virtio 转发该 I/O 请求到本服务器的 librbd 模块。librbd 根据写 IO 信息中的卷和 LBA 信息，通过数据映射机制确定数据所在的主 OSD 。主 OSD 接收到写 I/O 请求后，同时以同步方式写入到本节点 SSD 以及数据副本所 在其他节点的备 OSD ，备 OSD 也会同步写入本节点 SSD 。主 OSD 接收到两个都写成功后，返回写 I/O 成功给 librbd 。librbd 返回写 I/O 成功给虚拟机 OS 完成写流程。

3.3. 多协议网关

丰富的协议支持：支持 NFS、CIFS、FTP、POSIX、S3、HTTP、iSCSI、FC、Local SCSI 、CSI 和 RBD 原生访问等协议；标准协议和访问接口，简化业务 IT 架构的同时解除对业务的锁定。

新兴业务场景支持：支持 HPC 、IPFS 、大数据和容器等新兴负载。

1.灵活扩展

（1）软件定义，可自定义节点属性，并支持各种品牌的通用服务器；

（2）灵活部署，可以扩展到上千节点，满足不同业务需求；

（3）按需扩展，性能和容量随节点数增加而增长，满足不断增长的业务对

性能和容量的需求；

（4）整池扩容，业务无感知；

（5）全存储介质支持。兼容 SAS 、NL-SAS 、SATA HDD 、SATA SSD 、m.2 SSD 、u.2 NVMe 等接口；

（6）支持专用热备、全局热备、空闲硬盘热备、资源池内热备；支持自主规划集群物理设备拓扑，使允许多级故障隔离，包含硬盘、主机、机柜、机房故障隔离能力。

2.丰富的企业级功能

（1）支持目录级别快照，支持手动快照和定时快照；

（2）支持 WORM 功能，保护数据免遭意外、恶意更改，删除；

（3）支持基于 DNS 的负载均衡，支持多种负载均衡策略；

（4）支持网关负载均衡和 HA 保护，支持 AD 域、LDAP 域对接，本地认证等多种认证方式；

（5）支持配额管理，支持目录配额和用户/用户组配额，配额包括容量配额和文件数配额；

（6）通过数据管理系统可以实现文件的复制、迁移、备份、归档等丰富的数据管理功能；

作为企业数据湖底座，支持丰富的大容量非结构化数据保存和分析场。

3.文件共享、办公存储

单一全局命名空间，使用简单。支持文件共享、网盘、FTP 等办公场景。

4.海量数据存储

横向扩展，滚动升级，数据永久保存。

5.大数据、HPC 后端存储

兼容 HDFS,高效文件元数据处理机制，灵活应对 AI/ML 数据分析要求。

6.集中灾备资源池

利用数据管理系统和 ODPF（开放数据保护框架），可以作为大容量的共享灾备资源池。

7.企业数据湖底座

支持 Hadoop 存算分离部署，接口协议丰富，可以扩展到上千节点。

3.4. 产品技术手册:

产品技术手册

特性分类	特性	描述
多管理接口	RESTful API	支持 RESTful 接口，以利于上层应用无缝整合存储系统。
	CLI 命令行工具	支持 CLI 命令行工具，最大化管理效率。
	丰富的 SDK 开发包	支持主流语言的 SDK 开发包，与上层应用无缝集成。
	GUI 管理接口	100%可视化 GUI 管理接口。
运行管理与分析	全局即刻搜索	支持对系统内所有资源的全部信息搜索，列表信息排序及相应字段的过滤，允许快速访问关键资源。
	全方位性能指标监控	支持物理服务器 CPU、内存、网络、负载 监控。支持存储介质读写 IOPS、带宽、延迟 和盘 IO 利用率监控监控。支持存储池读写 IOPS、带宽、IO 大小和延迟监控。支持文件 系统数据和元数据的读写 IOPS、延迟、数据 带宽和 IO 大小监控。支持上述指标统计保留 天数自定义设置。
	实时健康管理	支持对象、块和文件网关和链路健康检 测。
		支持对物理服务器、存储介质、存储池
		数据冗余状态监控及管理。支持硬盘和链路健康检测。支持存储介质根据 S.M.A.R.T.信息预测设备寿命，提醒坏盘可能。
	自动重平衡	自定义闲时，对不均衡的 OSD 自动重平衡
	容量预警	根据智能算法预测未来容量使用增长，可以预测剩余容量将在几天后被写满，并在容量使

		用天数剩余30天内给与提示和告警
	自助巡检工具	支持自助健康巡检，生成巡检报告。
	事件中心	支持系统和用户触发产生的及关键事件日志，包括记录重要的系统触发、操作员行为触发及系统关键事件等（系统触发如服务器、硬盘离线上线、存储池重平衡等，用户 行为触发如创建、修改、删除资源等），便于排错、审计和跟踪，方便用户全方面掌控 存储运行情况。同时支持事件日志导出。
	可视化拓扑	支持可视化的硬盘及网络拓扑，展示硬盘的基本信息及从属关系，基本信息包括：硬盘名称、状态、容量、已使用容量、数据恢复情况、硬盘介质、IO利用率等。从属关系包括从属服务器视图和从属存储池视图。同时支持鼠标hover时显示硬盘详细信息。可视化硬件网络拓扑，直观展现集群网络情况、数据中心、机架、服务器、网卡信息。同时可以可视化展现集群网络中各个模块的异常情况。支持不同网卡的监控信息及监控 历史。
	磁盘定位	支持通过可视化界面点灯进行硬盘定位的功能。
	自定义容量阈值	自定义集群中硬盘的容量阈值，该阈值是硬盘被安全写满的阈值。达到该阈值后该硬盘将不可再写入数据，但该阈值可以根据业务需求调整。
	告警中心	支持集群内所有资源的告警，在存储系统的各级软硬件产生故障时，由管理控制台向管理员提示告警，显示告警得原因和系统状态，有助于及时了解资源使用情况和处理 突发事件。支持自定义告警通知，同时支持 邮件告警。

	标准统计接口	支持 Prometheus 等第三方标准统计接口，与已有统计系统无缝集成。
	SNMP 支持	支持 SNMP V2/V3，支持 TRAP，GET，SET 操作
	接口支持	支持 SMB/CIFS、NFS、FTP、POSIX、 FUSE、CSI、HDFS 等多种协议支持
文件存储	用户管理	灵活定义用户权限，不同的用户拥有不同的操作权限。
	配额	支持基于目录和用户/用户组的容量、文件配额，支持硬配额，软配额，建议配额
	快照	支持基于文件目录级的快照，支持手动快照和定时快照
	WORM	支持对文件系统中任意目录配置 WORM 保护功能，可设定自动锁定时间和 WORM 保护期
	负载均衡	支持文件客户端基于 CPU 负载，吞吐量、空闲内存，轮询等方式的负载均衡
	SMB/CIFS 共享回收站	支持 SMB/CIFS 共享目录删除的文件放到回收站中防止误操作
	Session 统计	可实时查看访问共享目录的 Session 信息，并显示共享路径总的连接数，Session 信息至少包括：登录的客户端 IP、登录用户、协议版本、登录时间
	文件目录树管理	支持文件系统下目录树浏览，该目录下所有的子目录，文件信息查看，基于目录可快速灵活的设定并管理基于指定 目录的共 享、快照、配额、WORM 等功能
	用户权限	支持本地用户以及对接 AD 域和 LDAP 域。可同时支持本地用户和 AD 域/LDAP 域混合认证方式
	多协议共享	同一文件系统多种访问协议
	支持ACL权限	支持 Windows ACL 权限和 Linux POSIX ACL 权限

	多活文件网关	支持最多 128 个多活文件网关，多网关并发读写统一 IO 空间，单个 VIP 高可用，自动切换，性能随网关横向扩展而线性增长
	元数据集群	使用高性能KV数据库实现文件元数据，独有专利技术的键值设计，支持完整的权限、ACL等文件属性
	文件权限管理	SMB/CIFS:完全控制/读写/只读三种权限选其一；FTP：可设置查看文件列表、创建文件夹、上传文件、下载文件、删除文件、重命名等权限；NFS：挂载控制为读写、只读两种权限。
	在线整池扩容	用户可不改变任何访问路径的情况下实现扩容；可避免大规模池内扩容导致海量数 据重平衡，解决海量数据存储扩容问题，有 效避免数据重平衡对前端业务的冲击；
	支持活动池和非活动池	支持活动池和非活动池，非活动池将不再接受新文件的数据分配写入，只支持已有文件的数据读、写、删除操作
	专用客户端FUSE Client	专用客户端直接部署在应用的 Linux 主机上，采用私有协议和存储集群通讯，不需要分布式网关层处理 IO；所有客户端均支持并发访问所有存储节点，具备极致的性能
	卷和快照管理	支持卷管理操作，自动精简配置，在线扩容。支持ROW秒级快照，在连续快照/克隆负载下，性能变化幅度小于 10%。支持按定时策略创建快照，达到定期对块存储卷进行数据备份。支持将克隆卷与快照关系断链。
	卷级实时 QoS	实时调整附加在卷上的 IOPS 和带宽限制属性，即时生效。
		支持 iSCSI，FC*，CSI，Local SCSI 和 RBD 原生

块存储	多协议支持	访等标准协议和访问接口，简化业务 IT 架构的同时解除对业务的锁定
	自动精简配置	支持自动精简配置，按写入有效数据容量分配实际空间，使存储空间能够根据需要自动扩展，而不必将存储空间全部分配出 去，因此只需要配置少量硬盘即可开展业 务，后续再根据存储空间使用情况新增硬 盘，从而降低初次购买成本和TCO。
	KVM 主机侧访问聚合	支持在 KVM 计算主机上聚合所有存储访问操作，大幅度降低 CPU 及内存相关资源消 耗，提升单卷性能
	链路冗余	支持 iSCSI 和 FC 链路冗余，最大支持 4 路径 MPIO，业务链路更安全。
	iSCSI VIP	在无法启用 MPIO 的场景下，使用 VIP 功 能自动切换故障端口业务，业务无感知，提高 iSCSI 路径的可靠性
	在线卷迁移	支持卷在线跨池迁移，比如，从 SSD 性 能池迁移至 HDD 容量池，从副本池迁移至 EC 池等。
	卷回收站	支持卷回收站，防止数据误删除
	跨池克隆	满足 OpenStack 场景下多虚拟机/多桌面 /多应用的快速复制需求。
	一致性组	支持一致性组快照和一致性组异步复制
	协议支持	支持 Amazon S3 标准接口，兼容 S3 生态体系，支持专用 Hadoop HDFS 高性能客户端，支持 NFS、CIFS、FTP、POSIX、S3、HTTP 等协议；
	用户配额及权限管理	灵活定义用户权限和配额限制，不同的用户拥有不同的操作权限，同时会限制不同用户的总容量、总存储桶、总对象数配额。
		支持 Bucket ACL 和 Bucket Policy 两种 权限控制，ACL 是比较粗的用户粒度权限。 Bucket

对象存储	桶权限控制	policy用于控制存储系统中的桶、对象等底层资源的访问权限，是比Bucket ACL 更细粒度的资源权限控制，包括桶的访问权限控制、桶内对象访问权限控制，从访问来源、访问目标、操作类型、过滤条件几个方面提供丰富的控制策略。
	桶配额管理	支持自定义存储桶配额，包括容量、对象数
	桶策略管理	对不同类型的对象数据可以自定义存储 策略，设定元数据、数据存放的资源池，以 及不同数据存放到不同的资源池，如大文件 存EC池、小文件存副本池。
	对象压缩	支持对象数据网关层进行压缩，实现端到端的传输减少。
	对象加密	支持对象数据网关加密，避免数据通过其他非法途径获取，保证数据安全。
	对象软链接	支持设置软链接，用于快速访问对象存储空间内的常用文件。
	对象秒合	支持将多个对象合并成一个对象
	对象桶回收站	存储桶支持设置回收站桶，删除后的数据在回收站内继续保留一段时间，便于后续 对删除数据的找回。
	数据分级存储	可以根据数据性质选择数据持久化存储级别，如：重要的数据用高安全级别的存储类型，非重要数据使用低安全级别的存储类型；性能要求高的数据存高性能存储池；需要提供存储空间利用率时选择EC存储。
	对象数据流动	数据按策略进行“热”、“温”、“冷”双向智能流动，实现高效访问和成本最优。同时支持复制、分层、归档等流动策 略。

	对象回源重定向	当客户端访问本地立思辰对象存储时，如果本地存储中没有被访问的数据，可以通过回源规则从源站获取对应数据。支持重定向、代理、镜像、CDN 四种回源模式。源站类型支持对象、文件、Web服务。
	数据生命周期管理	存储桶数据生命周期管理支持，可以对存储桶内的数据通过数据前缀或整桶进行归 档和删除，目前支持实时归档及延时归档， 同时支持延时删除。
	整池扩容	支持按池扩容，避免存储池大规模扩容导致数据重平衡从而影响业务可用性。
	多协议互通	NFS、HDFS、S3协议互通实现文件协议和对象协议的互操作。通过对象存储协议写入的数据可以通过NFS或HDFS协议访问，反之亦然。
	对象存储负载均衡	内置对象存储负载均衡，在保证对象存储高可用的同时，实现负载均衡作用。
	对象存储 SSL加密	对象存储支持 SSL访问加密。在客户端和服务器端之间建立加密通道，保证数据在传输过程中不被窃取或篡改。
	WORM写保护	存储桶支持WORM,一次写入，多次读取模式，对关键数据实行写保护，杜绝病毒破坏，非法篡改。
	桶访问日志	支持存储桶访问日志，记录桶资源访问详情，便于审计、分析、计费，可通过 S3 API获取日志信息。
	对象多版本	对象存储桶支持多版本，开启多版本 后，桶中的对象都以多版本形式存储，版本 数量无限制。
		内置高效数据查询引擎，数据查询引擎通过

	对象属性查询	并行计算的方式构建，以对象的元数据 （名称、大小、日期、自定义属性）、标签 等信息为基础构造查询检索条件，通过查询 引擎并行遍历系统内数据信息，并查找出符 合指定条件的对象集合。查询引擎计算能力 可根据需求按需扩展，实现每秒百万文件甚 至千万文件的查询效率。
	对象存储多站点	通过跨多集群建立统一的用户和桶管理 视图，桶中对象数据在多个站点间异步复制，实现数据异地容灾、就近访问、负载跨 站点分布。支持多站点同步进度查询。
	对象海量小文件	自动识别文件大小，将小文件存储在归并池中，大文件存储在数据池中，以不同的存储介质来存储不同大小的文件，以此提升小文件初期写入效率。当小文件达到一定大小后，存储系统可自动将小文件合并成大文 件后归并到数据池中，以此提升存储系统整 体空间利用率。
	对象迁移和复制	支持通过 X3DS 对对象存储进行数据迁移和异步复制。数据迁移支持对象到对象，对象到文件及文件到对象间的数据迁移服务，同时支持断点续传及增量迁移。
	CDP数据保护	支持将桶内数据本地持续保护，可恢复到一定范围内的任意时间点的数据，减轻应用灾难恢复的工作量。
	图像处理	在线图片处理，支持图片缩放、裁剪、格式转换、水印、旋转、质量调节等十几种图片处理功能
	S3Console	面向 S3 用户的便捷管理对象存储服务的图形化 Web 应用程序，主要提供了丰富的 桶、

		文件夹及文件等级别的管理功能。
	卷云端备份	支持将块数据备份到AWS、阿里云、腾讯云、电信天翼云等公有云平台，同时支持将块数据备份到异地对象存储集群或S3协议的其他存储平台。同时支持备份数据的去重、加密和压缩功能。
数据保护	策略驱动备份	通过丰富时间频率，间隔策略，支持将保护资源备份到多个保护平台,通过跨数据中心或云端的存储提供更高数据安全性。
	资源恢复	任意存储集群只要可以认证连接到任一 保护平台获取灾备资源,都轻松恢复到本地。
	传输加密	所有数据在不同平台间传输全部加密进 行，保护数据在不可信环境的安全性。
	异步复制	支持将块数据基于快照异步复制到远端集群，支持块数据容灾。
数据可靠	延展集群	可同时实现块、文件、对象的同城双活
数据安全	双向 CHAP支持	支持 iSCSI双向CHAP认证
	数据副本	支持存储池粒度的 1-6 副本存储策略， 以获得不同等级的数据可靠性。同时支持在线修改副本数
	数据EC	支持多种纠删保护机制。支持 2+4； 4+4；8+4；10+4;12+4;16+4 等模式。支持业 务需求，自定义调整 EC 保护模式。也支持 2+6；4+6；8+6；10+6；12+6；16+6等保护模 式，可以达到单一存储系统支持最多容忍任意 6 个数据节点同时失效，支持最大容忍任意 6 块硬盘同时失效。对块存储、对象存储及文件存储均支持。
	端到端校验	针对在线实时读写数据时做 CRC（Cyclic Redundancy Check)校验,防止静默数据错 误。

数据可靠	故障域	支持自主规划集群物理设备拓扑，使允许多级故障隔离，包含硬盘、主机、机架、数据中心故障隔离能力。
	热备盘	支持专用热备、全局热备、空闲硬盘热备、资源池内热备。
	不中断维护	在任何规定硬件维护、软件升级、容量扩充、介质热更换及故障窗口时间内，避免集群存取性能不受影响。
	硬盘维护模式	硬盘维护模式是辅助对硬盘进行维护操 作时数据不进行重平衡，利用该模式可用于 主动的硬件设备下线维护场景。
	数据恢复 QoS 控制	在数据较长时间处于降级状态时，例如节点丢失或副本丢失，系统会自动触发数据重建恢复。用户可设定数据恢复的带宽规 则，最小化数据恢复对业务的影响。
	磁盘漫游	支持将故障中的存储节点硬盘和 SSD 更换到新设备上，重新恢复至原存储池，避免数据多次重平衡
	磁盘重建	硬盘或者 SSD 出现故障时一键完成新盘在线替换，不需要重新创建OSD，原OSD数据会自动重建到新盘上
	磁盘亚健康处理	自动检测集群中的坏盘和慢盘，告警并自动隔离，保护存储集群稳定。
	网络亚健康处理	自动检测集群网络故障，告警并自动隔离，保护存储集群稳定。
	延展集群	通过延展集群的方式支持同城双活，自定义集群主副本位置，实现本地读优先。
		系统全冗余，控制器之间采用高可用架构；即支持元数据和数据分离部署（需要元数据

	全冗余架构	节点，二者独立扩展）也支持元数据和数据的融合部署（不需要独立的元数节点），任何一个控制器出现故障，不影响数据的正常访问。
	基于角色访问制	通过权限角色划分,保护系统访问安全。
数据安全	全面审计记录	所有系统操作、维护、IP信息等记录。
	存储系统控制台SSL加密	存储控制台支持SSL访问加密。在客户端和服务器端之间建立加密通道，保证数据在传输过程中不被窃取或篡改。
	身份认证	支持第三方平台无需使用本存储管理平台的认证信息，而是直接使用第三方平台已 有的认证信息，即可登录使用本平台的功能。
	密码安全策略设置	通过密码安全策略更加安全的保证控制台中的密码规则及账户登录。
数据缓存	内存读缓存	通过流预测算法在内存中提前获取目标 数据，大大提高读性能。支持添加介质时配 置读缓存大小。
	SSD读写缓存	通过热点预测，写入合并等技术将高速设备与低速设备结合，大幅提升读写性能。 支持给低速介质配备50GB到几百GB范围持久高速缓存。
软硬件兼容	多存储介质支持	兼容 SAS、NL-SAS、SATA HDD，SATA SSD、m.2 SSD、U.2 NVMe等接口。
	服务器兼容性	支持工业标准的x86 通用硬件,硬件不限定品牌，不限定硬件部件的型号及技术参数。升级、扩容的过程中用户可以选择自行增加硬件部件。
	操作系统升级	在不剔除节点配置的前提下完成操作系统的在线升级。
	存储系统兼容性	分布式软件可以安装在通用发行版Linux 操作

		系统上，无需定制操作系统支持。
	硬盘及服务器异构	支持同一存储资源池混插磁盘。支持统一存储资源池异构服务器。
基础架构易用 性	操作界面	界面显示 license、服务更新、产品功能、版本号等各种更新。
	自动化配置	自动化部署、一键安装；图形化快速完成资源的基础部署，新增节点上线时间缩短。
	操作系统在线升级	在不剔除节点的前提下完成操作系统的在线升级。
	可视化操作引导	通过可视化操作引导，一键配置所需资 源，快速完成相应资源的使用，降低产品复 杂度，减少产品学习成本及配置时间。

4. 分布式存储产品规格

4.1. 节点

采用分布式存储架构，存储节点采用标准 X86 架构服务器，根据内置硬盘密度不同细分为 3 种节点。

节点配置从 3 节点起配，最大可横向扩展至 512 个节点，存储容量和性能线性增长。因此可用于构建、维护 PB 级数据集群。

4.1.2 内存

元数据服务器和 Monitor 必须可以尽快地提供它们的数据，所以节点标配 32GB 内存，确保每进程可分配足够的内存空间 。

OSD 的日常运行不需要那么多内存（如每进程 500MB ）；然而在恢复期间它们占用内存比较大（如每进程每 TB 数据需要约 1GB 内存）。节点最大可支持 512GB 内存。

4.1.1. CPU

元数据服务器对 CPU 敏感，它会动态地重分布它们的负载，元数据服务器

支持配置双路龙芯处理器，具有强悍的处理能力。

OSD 运行着 RADOS 服务、用 CRUSH 计算数据存放位置、复制数据、维护其自己的集群运行图副本， OSD 同样可配置双路龙芯处理器，具备强大的处理能力。

Monitor 用于维护集群运行图的副本，配置单 CPU 即可。

4.1.2. 数据存储

可实现海量存储，考虑到成本和性能的平衡，为用户规划数据存储配置。来自操作系统的并行操作和到单个硬盘的多个守护进程并发读、写请求操作会影响性能。不同的文件系统也有局限性：btrfs 尚未稳定到可以用于生产环境的程度，但它可以同时记日志并写入数据，而 xfs 和 ext4 却不能。

因为发送 ACK 前必须把所有数据写入日志（至少对 xfs 和 ext4 来说是），因此均衡日志和 OSD 性能相当重要。

4.1.3. 硬盘驱动器

OSD 应该有足够的空间用于存储对象数据。考虑到大硬盘的每 GB 成本，我们建议用容量更大的硬盘。单个驱动器容量越大，其对应的 OSD 所需内存就越大，特别是在重均衡、回填、恢复期间。根据经验，1TB 的存储空间大约需要 1GB 内存。

存储驱动器受限于寻道时间、访问时间、读写时间，还有总吞吐量，这些物理局限性影响着整体系统性能，尤其在系统恢复期间。因此我们推荐独立的驱动器用于安装操作系统和软件，另外每个 OSD 守护进程占用一个驱动器。大多数“slow OSD ”问题的起因都是在相同的硬盘上运行了操作系统、多个 OSD 、和/或多个日志文件。鉴于解决性能问题的成本差不多会超过另外增加磁盘驱动器，你应该在设计时就避免增加 OSD 存储驱动器的负担来提升性能。

允许在每块硬盘驱动器上运行多个 OSD ，但这会导致资源竞争并降低总体吞吐量；也允许把日志和对象数据存储在相同驱动器上，但这会增加记录写日志并回应客户端的延时，因为必须先写入日志才会回应确认了写动作。btrfs 文件系统能同时写入日志数据和对象数据， xfs 和 ext4 却不能。因此建议应该分别

在单独的硬盘运行操作系统、OSD 数据和 OSD 日志。

4.1.4. SSD

使用 SSD 来降低随机访问时间和读延时，同时增加吞吐量。SSD 和硬盘相比每 GB 成本通常要高 10 倍以上，但访问时间至少比硬盘快 100 倍。SSD 没有 可移动机械部件，所以不存在和硬盘一样的局限性。但 SSD 的顺序读写性能很重要，在为多个 OSD 存储日志时，具有高顺序读写吞吐量的 SSD 对系统整体性能提升至关重要。

对于日志和 SSD 时还有几个重要考量：

●写密集语义：记日志涉及写密集语义，SSD 写入性能好于硬盘。

●顺序写入：在一个 SSD 上为多个 OSD 存储多个日志时也必须考虑 SSD 的顺序写入极限，因为它们要同时处理多个 OSD 日志的写入请求。

●分区对齐： 采用了 SSD 的一个常见问题是常常忽略了分区对齐，这会导致 SSD 的数据传输速率慢很多，所以请确保分区对齐了。把 OSD 的日志存到 SSD、把对象数据存储到独立的硬盘可以明显提升性能。

提升 FS 文件系统性能的一种方法是从 FS 文件内容里分离出元数据。系统提供了默认的 metadata 存储池来存储 FS 元数据，所以不需要给 FS 元数据创建存储池，但是可以给它创建一个仅指向某主机 SSD 的 CRUSH 运行图。

4.1.5. 硬盘类型

支持 SSD、SAS、NL-SAS 和 SATA 等多种硬盘接口和容量点选择，用户可根据存储容量和存储性能有多种选择。

4.1.6. 节点网卡

存储集群每个节点部署两个双端口 10Gbps 网卡分别用于公网（前端）和集群网络（后端）。集群网络用于处理由数据复制产生的额外负载。使用 10Gbps 复制 1TB 数据的时间为 20 分钟。在一个 PB 级集群中， OSD 磁盘失败是常态，而非异常；在性价比合理的的前提下，系统管理员想让 PG 尽快从 degraded（降级）状态恢复到 active + clean 状态。可使用 VLAN 来提高网络和硬件可

管理性使用 VLAN 来处理集，群和计算栈（如 OpenStack、CloudStack 等等）之间的 VM 流量时，采用 10G 网卡性能可以得到保障。每个网络的机架路由器到核心路由器应该有更大的带宽，如 40Gbps 到 100Gbps。

节点还配置 2 个 GE 网口用于管理和部署，包括 SSH 访问、VM 映像上传、操作系统安装、端口管理等等。