



UMStor 分布式存储 技术白皮书

优刻得科技股份有限公司
上海 · 北京 · 深圳

目 录

1	概述	3
2	产品价值	4
2.1	灵活 FLEXIBILITY	4
2.2	高可靠 STABILITY	5
2.3	高扩展 SCALE UP	5
2.4	高性能 SPEED	6
3	UMSTOR 系统架构	7
3.1	UMSTOR 架构概述	7
3.2	数据可靠性技术	8
3.3	系统扩展性设计	9
3.4	多活部署方案设计	9
3.5	容灾备份方案设计	9
3.5.1	跨数据中心容灾备份	10
3.5.2	存储桶级别容灾备份	13
3.5.3	块存储容灾备份	14
4	技术优势	16
4.1	完全对称设计	16
4.2	多种数据服务	17
4.3	智能故障自愈和并行数据恢复	18
4.4	机柜级故障域隔离	19
4.5	多种数据保护模式	20
4.6	灵活的数据调度	20
4.7	混合云存储	21
4.8	线性横向扩展 SCALE-OUT	21
4.9	高聚合性能	23
4.10	数据湖存储平台	24
4.11	数据生命周期管理	24

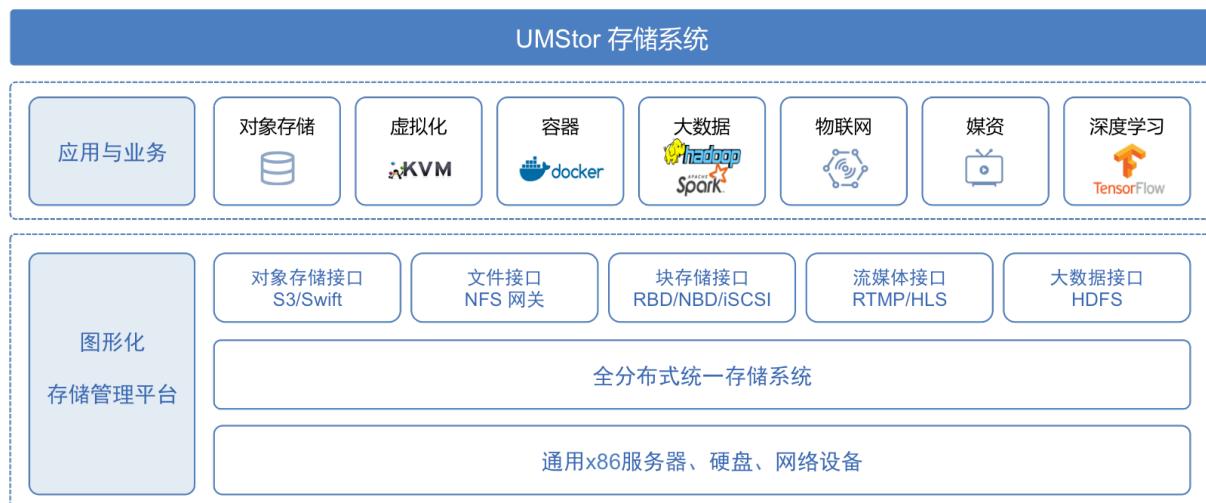
4.12	高效易用的管理平台.....	25
5	技术指标.....	26
5.1	技术规格.....	26
5.2	性能指标.....	27

1 概述

云计算与大数据正在推动数据中心存储基础架构发生深刻的变革，传统的异构存储设备难以解决统一管理和数据共享的难题，而且也不能适配虚拟化、云计算、大数据、物联网、混合云等场景，导致成本居高不下。

在海量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？如何建立统一开放的大数据体系？如何进行数据资源共享？

UCloud 统一存储产品 UMStor 为云而生，适用于虚拟化、云计算、大数据、物联网、混合云等使用场景，能够适配多种应用接口，基于通用的 x86 服务器构造统一存储资源池，提供多种数据服务，并且互联互通，打破数据调度壁垒。



UMStor 是真正的软件定义存储系统 (SDS)，采用领先的全分布式全冗余架构，没有单点故障，具有高弹性和高可靠性，性能和容量可以横向扩展。我们可以通过灵活的软件配置和硬件选型，自定义存储系统的性能、容量、数据保护能力等，满足当前和未来战略的需求。

2 产品价值

UMStor 提供全分布式高扩展云存储，为企业存储和管理不断增长的数据，主要价值如下：

2.1 灵活 Flexibility

统一存储，多种接口，数据调度，存储分级

- ▶ 多种数据服务接口，UMStor 在一套存储系统中提供多种应用接口，提供块、对象、大数据存储服务，把应用与数据连接起来。块存储服务支持虚拟化、容器、物理服务器，支持 OpenStack 云平台和 VMware；对象存储服务支持 Amazon 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；大数据存储提供 HDFS 接口，支持 Hadoop、Spark、Hive、HBase 等大数据应用，也支持深度学习框架 Tensorflow。数据的收集、处理、分发都可以放在 UMStor 上。
- ▶ 数据调度，UMStor 支持存储分级功能，可以根据可靠性、性能、访问频度定义存储池级别，例如分成高性能存储池、大容量存储池、归档存储池等。创建卷或者对象的时候，可以指定存储池。而且可以实时热迁移卷或对象到其他存储池上，实现数据的全生命周期管理。UMStor 的对象存储服务还支持存储桶跨域复制功能，可以把存储桶异步复制到另外一个 UMStor 系统上。对象存储服务还支持 Cloud Sync 功能，可以把本地数据复制到公有云或外部存储系统上，实现数据就近访问。
- ▶ 数据共享，UMStor 的对象存储支持完善的 ACL 访问控制，可以实现多个用户和多个应用对于数据的同时访问。UMStor 促进企业深度利用数据，将数据变成“生产资料”，产生更大的价值。

- ▶ 软件定义存储 SDS , UMStor 软硬件分离 , 使用通用 x86 服务器和网络交换机 , 没有硬件厂商锁定。

2.2 高可靠 Stability

全分布式全冗余架构 , 多站点数据保护

- ▶ 全分布式架构 , UMStor 所有软硬件全冗余 , 无单点故障 , 具有超高可用性。
- ▶ 故障自愈 , UMStor 具有故障检测和自动恢复功能 , 数据恢复不需要人工接入 , 在数据恢复期间 , 数据访问正常 , 服务不中断。 UMStor 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。
- ▶ 数据多副本 , UMStor 支持存储池粒度的不同副本数量存储策略 , 副本数支持 2~10 。支持任意 N-1 个硬盘损坏或任意 N-1 个服务器损坏。
- ▶ 数据 EC 纠删码 , UMStor 也支持 N+2 、 N+3 、 N+4 等多种纠删码策略 , 以便获得更多有效存储容量。
- ▶ 数据保护 , UMStor 支持多站点容灾备份 , 可以对块存储的卷增量备份到远端 UMStor 上 , 可以把对象存储的数据异步复制到远端 UMStor 上。

2.3 高扩展 Scale up

按需在线扩容 , 最大支持 50PB 容量

- ▶ 快速落地 , UMStor 配置灵活 , 可以最少从 3 台 x86 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。

- ▶ 超大规模，UMStor 可以支持 1024 个存储池，每个存储池可以支持 10PB 的存储容量，所有的存储池组成全局统一命名空间。

2.4 高性能 Speed

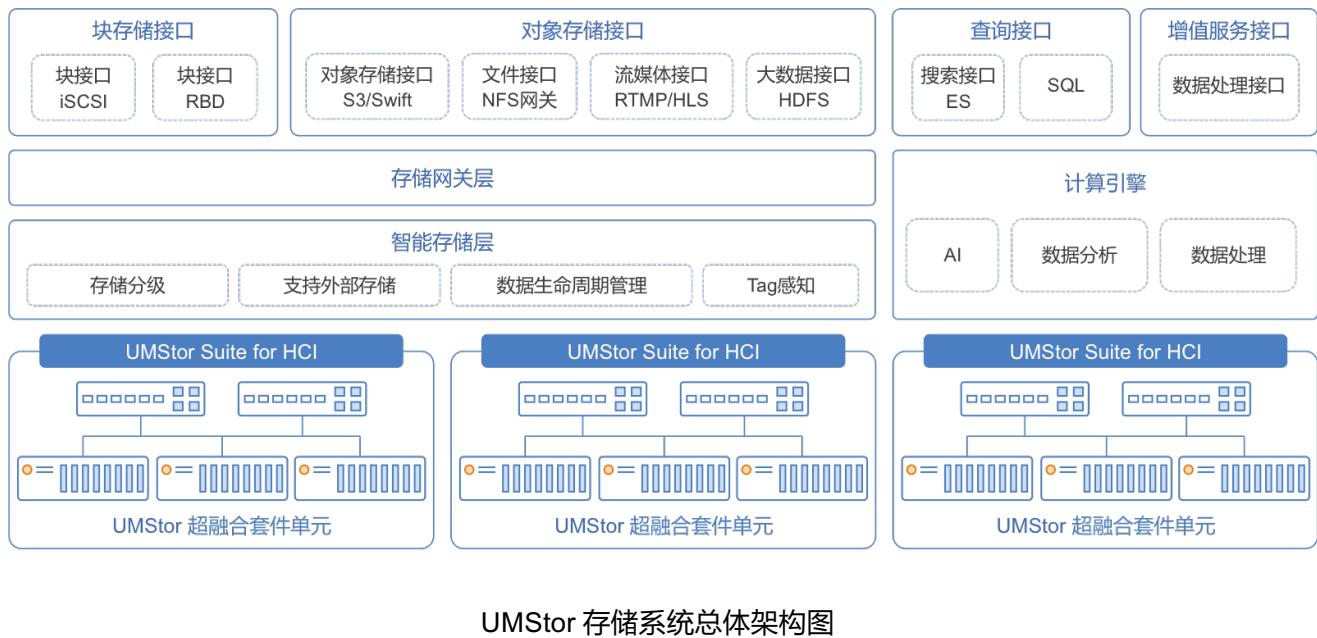
百万级别 IOPS，自动化运维

- ▶ 超高性能，UMStor 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。
- ▶ 秒级创建快照和克隆，UMStor 的块存储服务支持秒级操作，加快业务创新。
- ▶ 精简配置，UMStor 的块存储服务创建新卷时不占用实际存储空间，只有当实际写入数据时，才会占用存储空间。
- ▶ 图形化操作，UMStor 支持图形化界面进行部署、配置、监控、管理整个存储系统。

3 UMStor 系统架构

3.1 UMStor 架构概述

UMStor 基于通用的 x86 架构硬件平台构建统一的存储资源池，提供多种数据服务，并且互联互通，打破数据调度壁垒。系统总体架构如下图所示，



UMStor 存储系统由 UDP 统一存储平台、存储网关模块、存储管理平台、增值服务模块组成。

底层 **统一存储平台** 是整个 UMStor 存储系统的基础支撑平台，具体负责数据的实际存放与管理。由存储硬件层、存储 OS 内核、智能存储层、外部存储层组成，包括硬盘故障检测、故障恢复、数据动态重分布等功能，可提供多副本、纠删码两种数据冗余方式，具备存储分级、数据生命周期管理、Tag 感知能力。

存储网关模块 提供功能强大的存储网关和丰富的协议接口。UMStor 具备强大的协议能力，提供包括块存储接口（iSCSI / RBD）、对象存储接口（S3 / Swift）、文件存储接口

(NFS)、流媒体接口 (RTMP / HLS)、大数据接口 (HDFS)，充分满足用户对各类数据的存储和使用需求。

存储管理平台 通过图形化界面实现对存储系统的自动化部署、运维、监控、告警、运营等工作。

增值服务模块 为用户提供大数据分析、AI等数据增值服服务，包括AI、大数据分析和处理、ES弹性搜索、SQL查询等。

3.2 数据可靠性技术

UMStor存储系统采用全分布式架构，所有软硬件全冗余，无单点故障，具有超高可用性。

硬件故障是常态，而不是异常。整个UMStor存储系统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，UMStor允许系统中的部分部件失效。

UMStor存储系统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接入，在数据恢复期间，数据访问正常，服务不中断。并且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，保证数据的可靠性。

同时，UMStor支持存储池粒度的不同副本数量存储策略，副本数支持2~10（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），允许任意N-1个硬盘损坏或任意N-1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、多个数据中心甚至不同的国家和地区之中，最大限度提高数据容灾能力。

UMStor 存储系统支持 N+2、N+3、N+4 等多种纠删码策略，以便获得更多有效存储容量。

UMStor 存储系统支持多站点容灾备份，可以对块存储的卷增量备份到远端 UMStor 上，可以把对象存储的数据异步复制到远端 UMStor 存储系统上。

3.3 系统扩展性设计

UMStor 存储系统采用分布式架构设计，节点由多个独立的 x86 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 (x86 物理服务器) 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。支持理论上无限水平扩展，支持 PB 级别的大规模存储。

UMStor 存储系统支持最小规模从 3 台 x86 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。同时，扩容过程中业务无需中断。

UMStor 存储系统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。

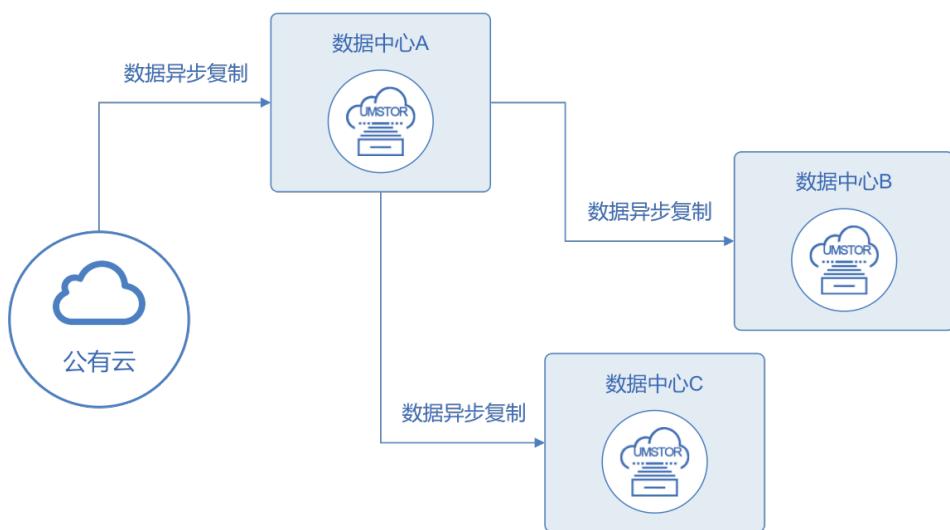
3.4 多活部署方案设计

3.5 容灾备份方案设计

根据业务需要，UMStor 存储系统可实现多种不同级别的数据容灾备份。

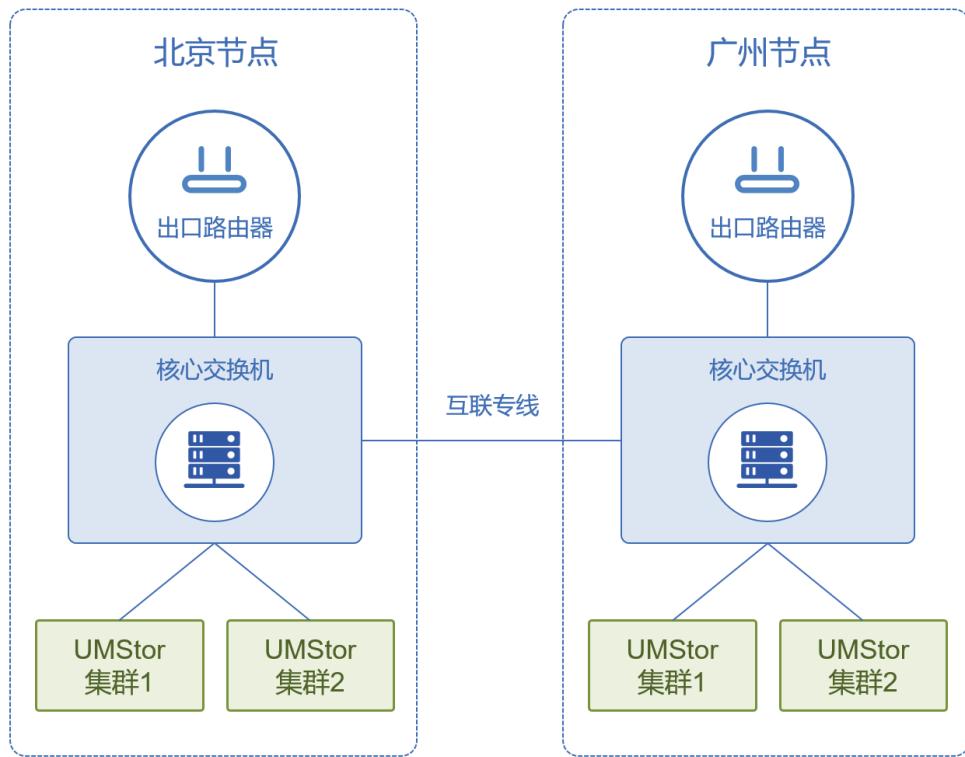
3.5.1 跨数据中心容灾备份

UMStor 存储系统支持多站点间的数据复制功能，可以实现跨数据中心的数据容灾备份，数据复制粒度是整个存储集群。



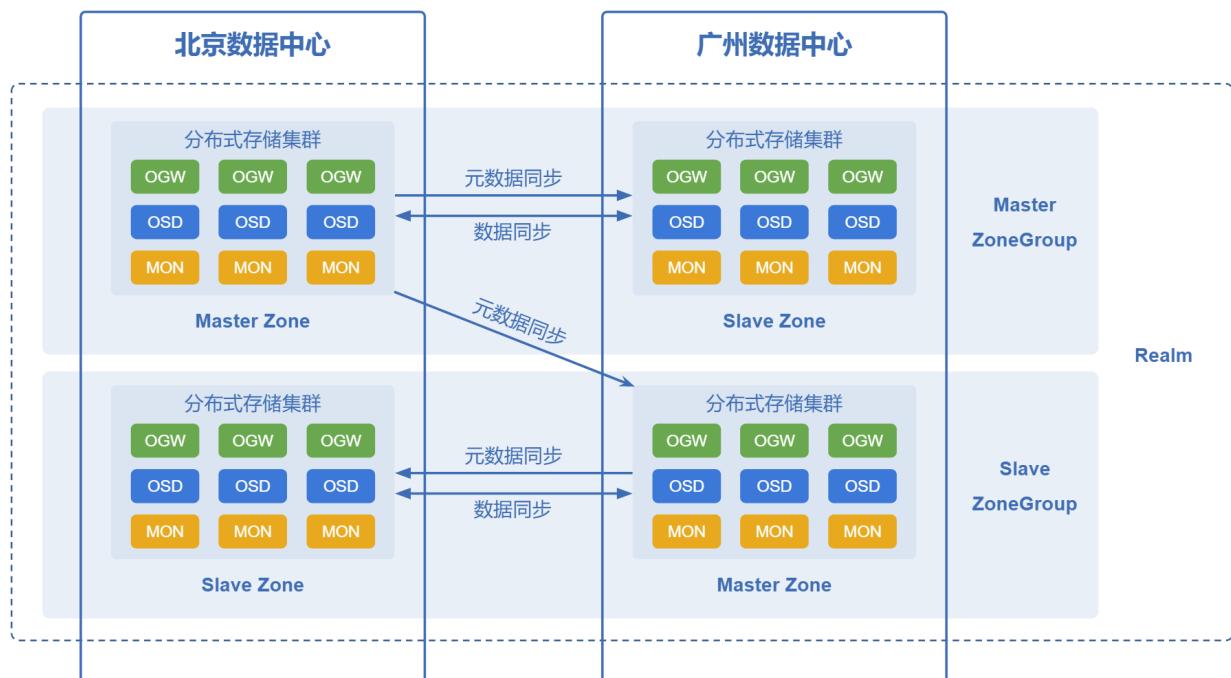
UMStor 实现多数据中心间容灾备份示意图

下面是两个站点间的网络架构，因为站点间的通信是使用 HTTP，所以也可以使用 WAN。



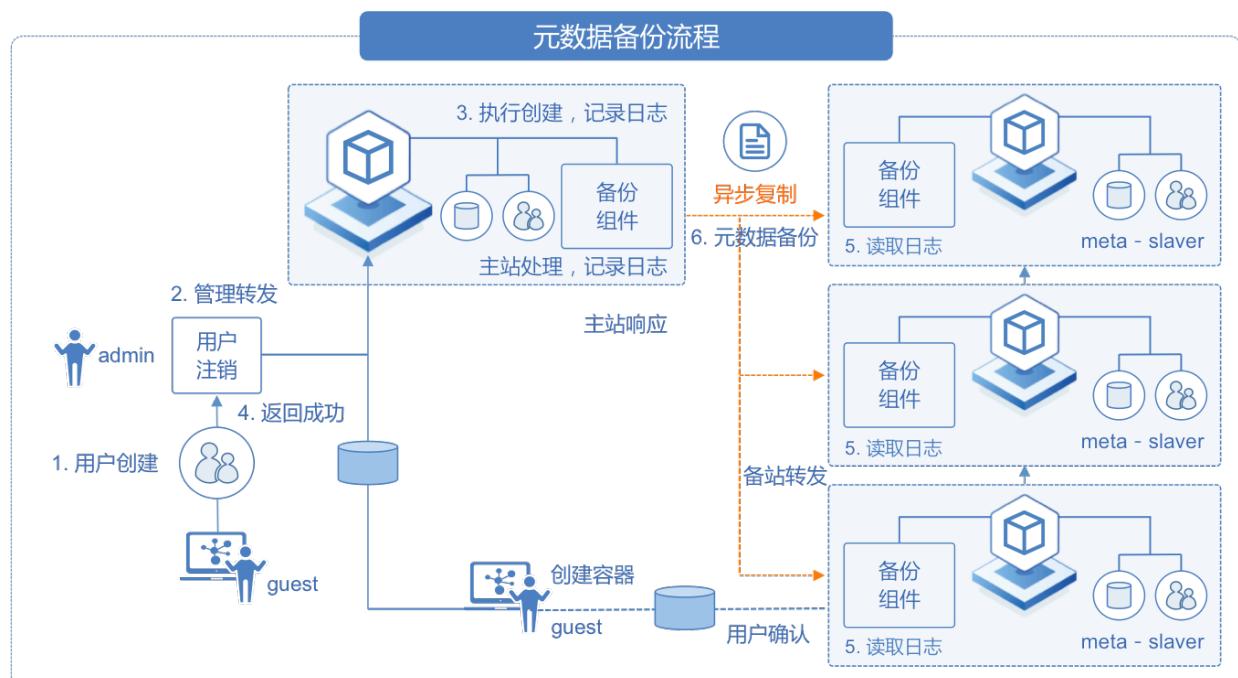
UMStor 跨数据中心容灾备份网络架构示意图

下面是逻辑架构图：

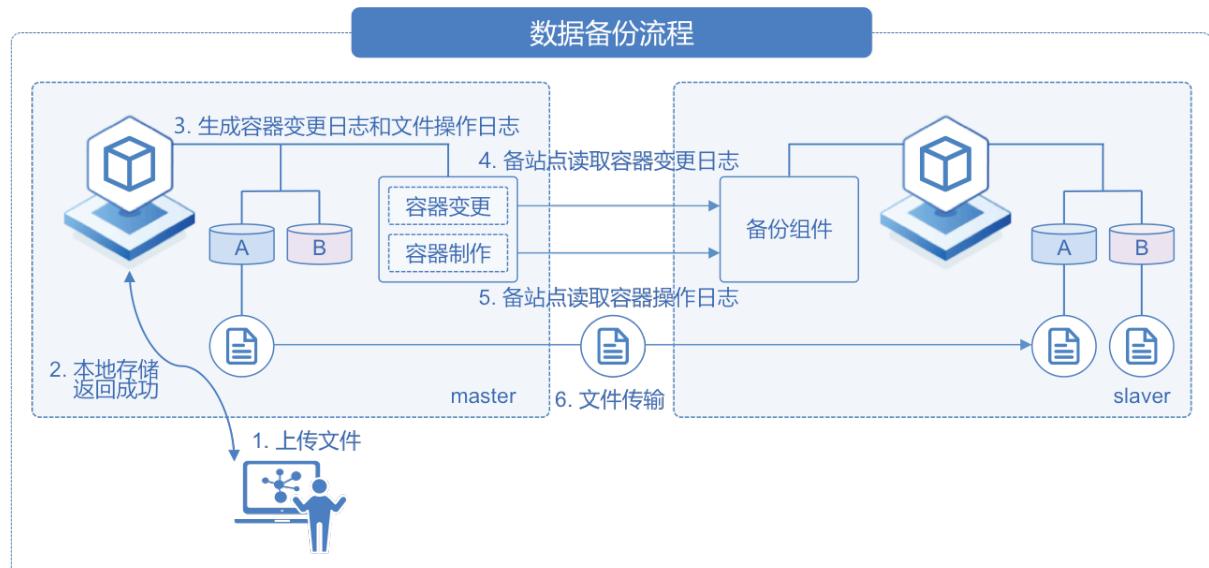


UMStor 跨数据中心容灾备份逻辑架构图

多站点间数据复制的原理如下图所示：



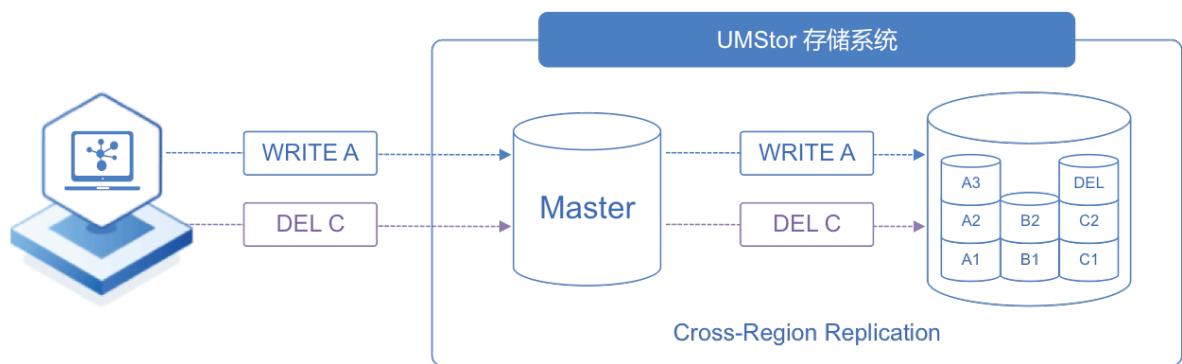
UMStor 多站点间元数据备份流程图



UMStor 多站点间数据备份流程图

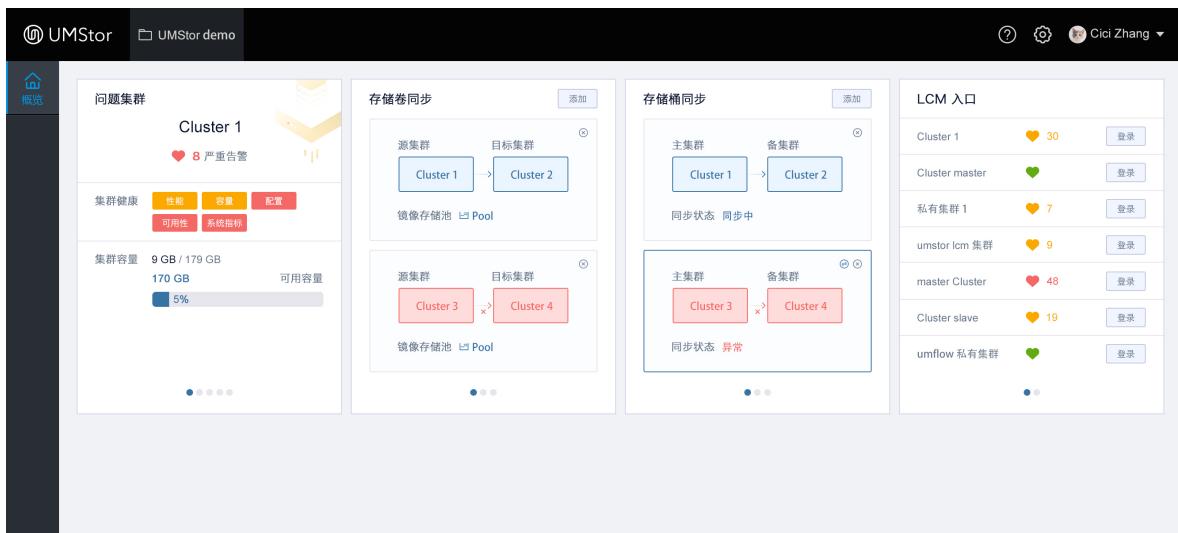
3.5.2 存储桶级别容灾备份

UMStor 支持存储桶跨域复制，数据复制粒度是单个存储桶，与集群粒度容灾备份相比，其数据复制粒度要更小，更灵活方便，成本更低。



UMStor 存储桶级别容灾备份示意图

下面是如何设置存储桶跨区域复制：

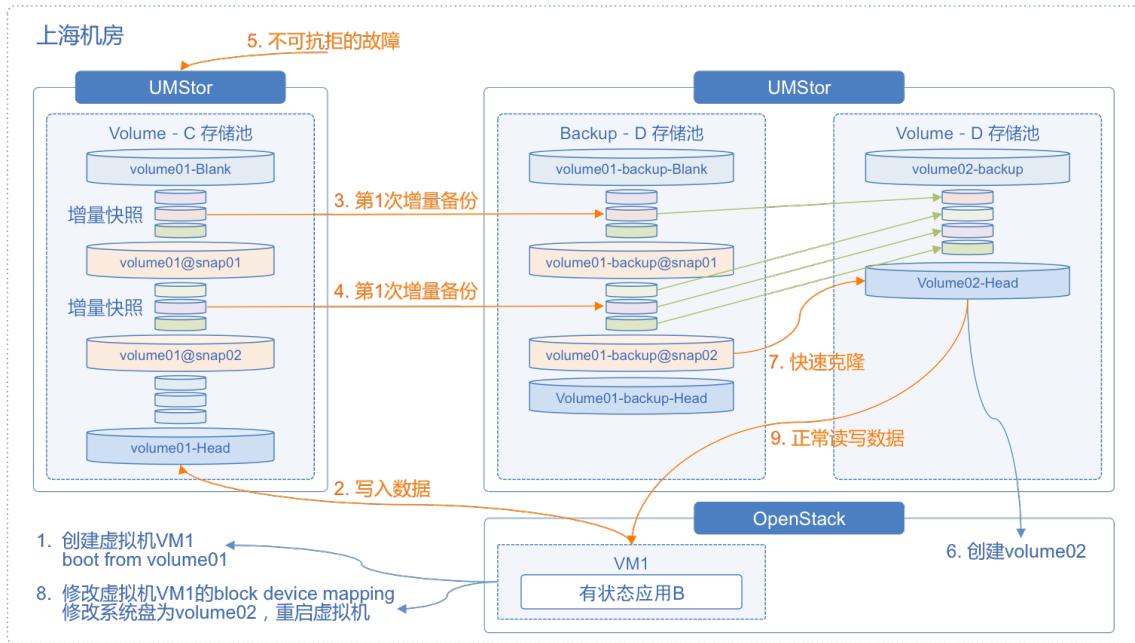


UMStor 存储桶跨区域设置图

3.5.3 块存储容灾备份

UMStor 可以实现块存储级别的全量备份和增量备份，可以将数据备份到 UMStor 分布式存储集群或公有云和外部存储。

下图展示两个 UMStor 块存储间的备份原理，可以实现存储增量备份。并且，当一个 UMStor 块存储集群出现故障时，可以快速进行故障切换到另外一个块存储集群。（切换速度）

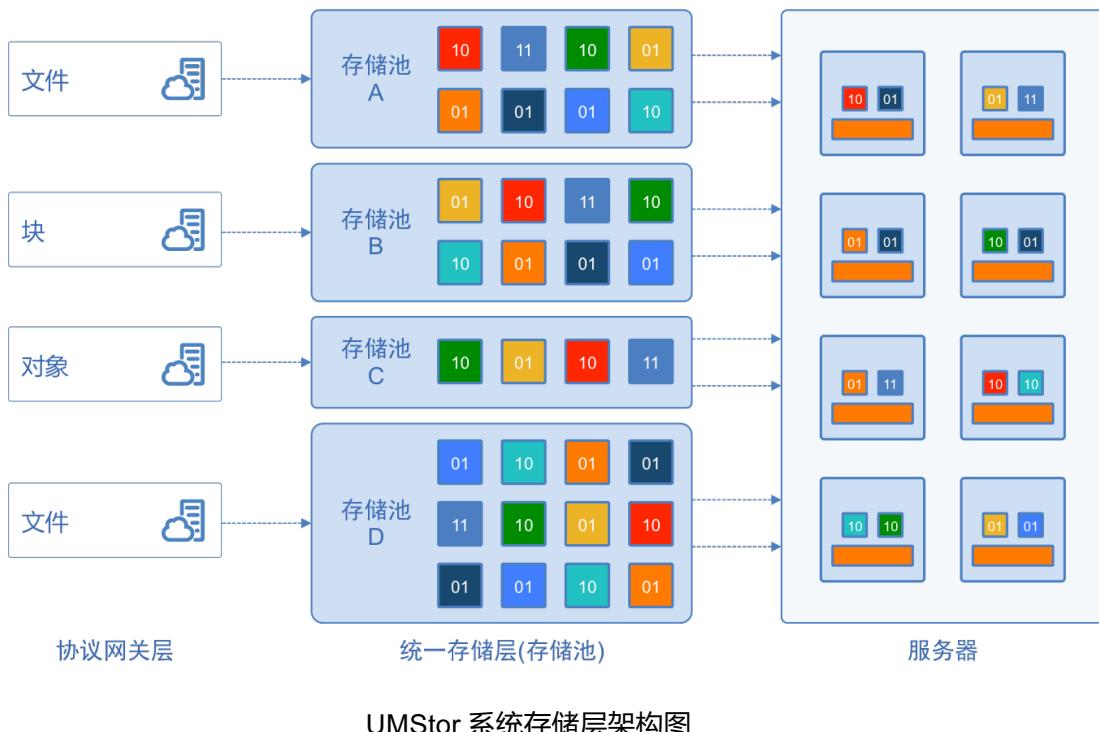


UMStor 块存储级别容灾备份原理示意图

当集群故障恢复时，我们可以通过使用存储热迁移，进行 fail-back。UMStor 分布式集群可以有效跟 OpenStack 云平台进行容灾备份恢复上的整合，保证云平台的稳定运行。

4 技术优势

4.1 完全对称设计



UMStor 存储系统采用全分布式架构，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)，负责提供提供存储空间，并保证故障自愈和数据恢复。

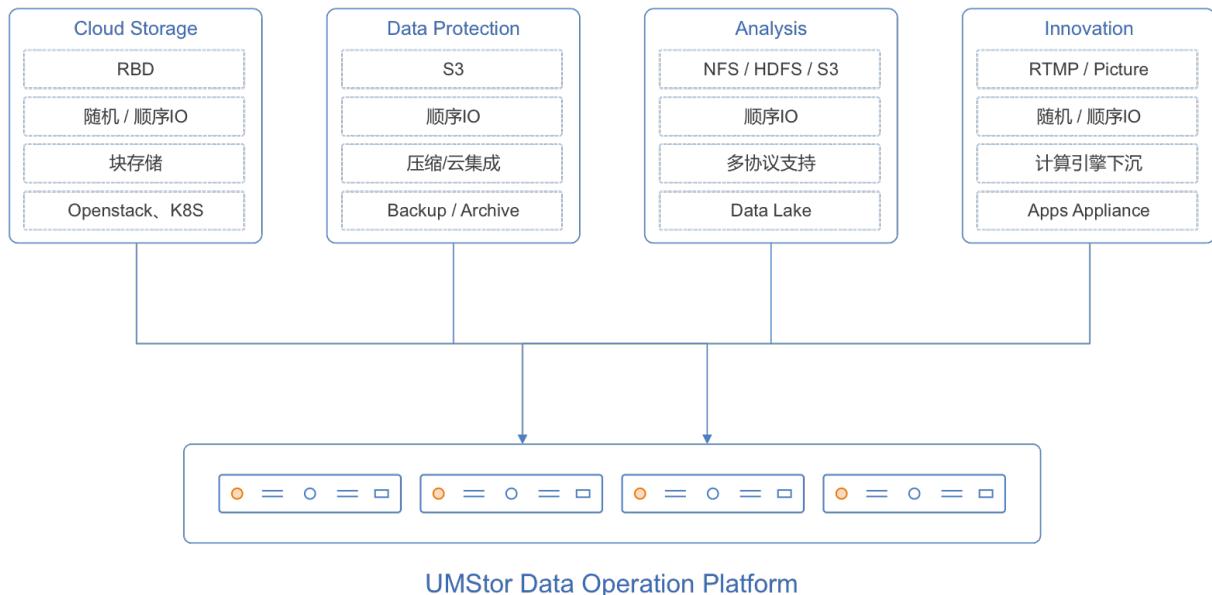
OpenStack 块存储和容器持久化存储直并行接访问 OSD，不需要代理和转发，具有非常高的性能。

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。

4.2 多种数据服务



UMStor 存储系统具备强大的协议能力，支持包括块存储接口（ iSCSI / RBD ）、对象存储接口（ S3 / Swift ）、文件存储接口（ NFS ），可对外提供多种数据服务。

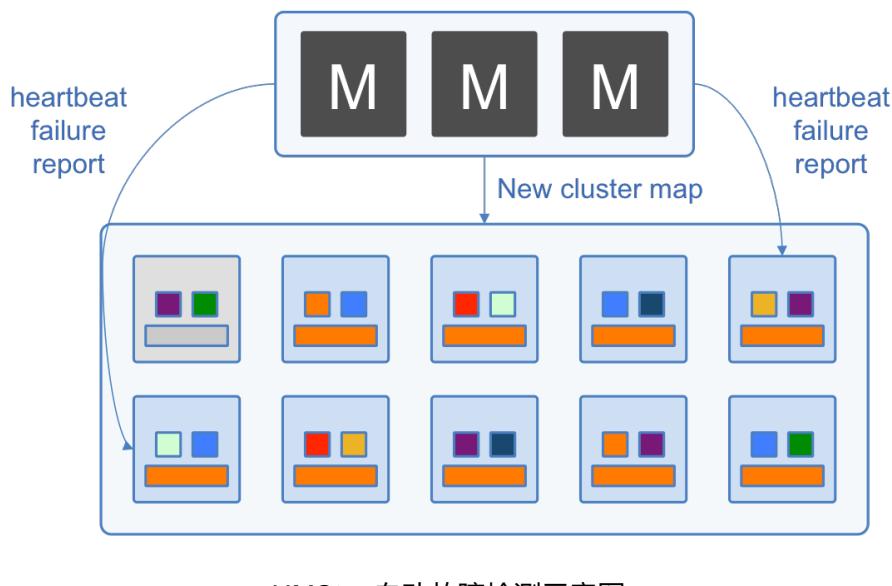
UMStor 可提供块存储服务，支持虚拟化应用、容器应用、物理服务器等，为 OpenStack 云平台和 VMware 提供存储服务。

UMStor 可提供对象存储服务，可以用于通用云存储服务，如存放文件、图片、视频等非结构化数据的存储；也可以用于备份和归档，支持主流备份软件；支持 Amazon S3 接口，无缝对接 S3 完善的生态系统。

UMStor 提供多种应用接口，如流媒体接口（ RTMP / HLS ） 大数据接口（ HDFS ）等。支持 Hadoop、Spark、Hive、HBase 等大数据应用，也支持深度学习框架 Tensorflow。对于数据的收集、处理、分发都可以放在 UMStor 上实现。

4.3 智能故障自愈和并行数据恢复

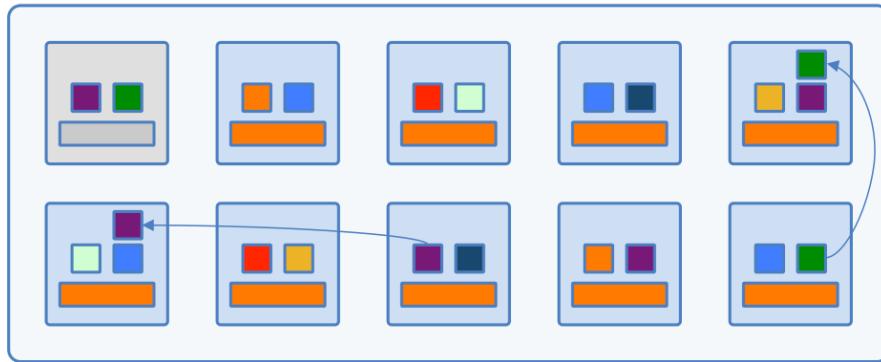
整个 UMStor 存储系统由数百或数千个存储着文件数据片段的服务器组成，每一个组成部分都很可能出现故障，UMStor 允许系统中的部分部件失效。UMStor 存储系统具有强大的故障检测和自动恢复能力，数据恢复不需要人工接入，在数据恢复期间，数据访问正常，服务不中断。



UMStor 自动故障检测示意图

UMStor 存储系统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。

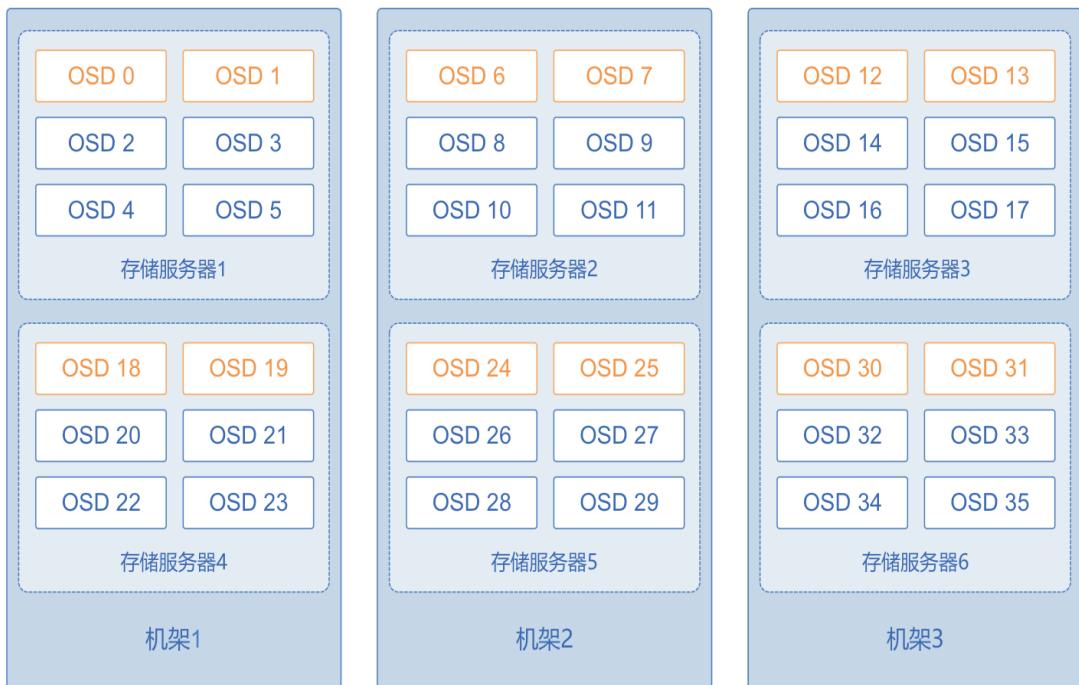
Distributed Recovery



UMStor 数据并行恢复示意图

4.4 机柜级故障域隔离

UMStor 存储系统可以定义数据分布策略，例如定义多副本分布在不同的机架上，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。



数据跨机架分布示意图

4.5 多种数据保护模式

目前 UMStor 存储系统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。

UMStor 支持存储池粒度的不同副本数量存储策略，副本数支持 2~10（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），允许任意 N-1 个硬盘损坏或任意 N-1 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、多个数据中心甚至不同的国家和地区之中，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，

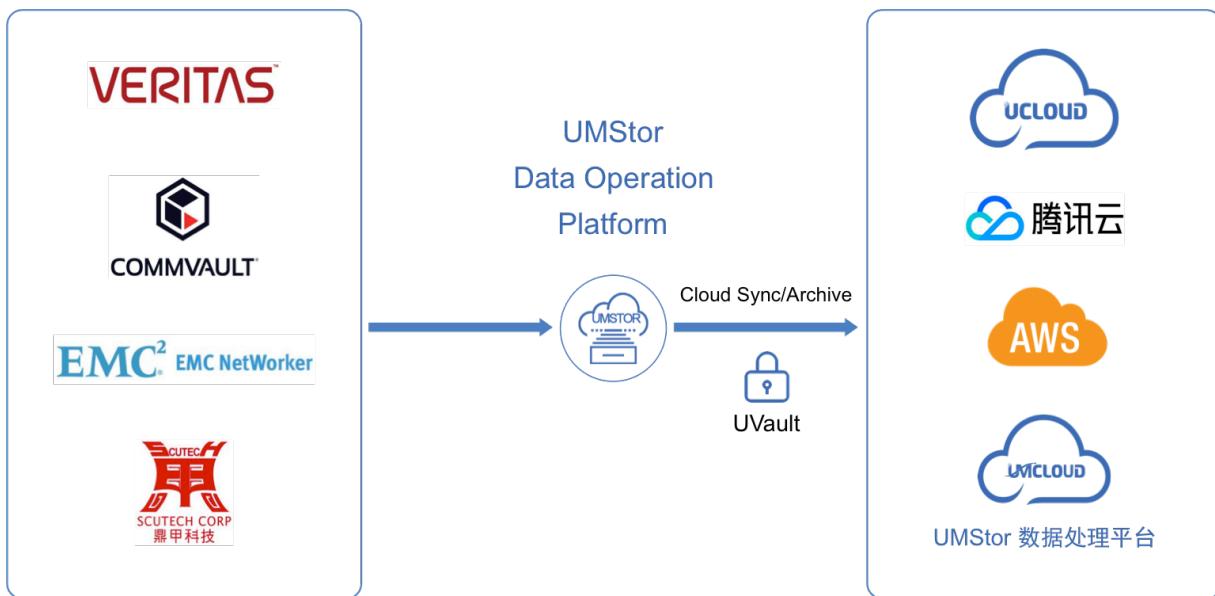
- ▶ 可跨地理位置存储数据和完全冗余副本
- ▶ 读性能延迟低
- ▶ 写性能延迟高
- ▶ 数据恢复速度快

同时，UMStor 支持纠删码策略，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。

4.6 灵活的数据调度

- ▶ 可以把存储池从一个Domain热迁移到另外一个Domain上。
- ▶ 可以指定块存储的云硬盘创建在指定的存储池上。
- ▶ 可以指定对象存储的bucket创建在指定的存储池上。
- ▶ 可以热迁移云硬盘到另外一个存储池上。

4.7 混合云存储



多云存储示意图

UMStor 外部存储层可以接入外部存储设备，包括其它第三方公有云存储、私有云存储、NAS 存储等。对于企业在今后很少会访问到的冷数据，我们建议可以存放到公有云厂商提供的云存储以有效降低存储成本。UMStor 内部已经深度集成了 AWS 和 UCloud 公有云存储接口，结合策略，用户可以通过图形界面实现数据一键转移。

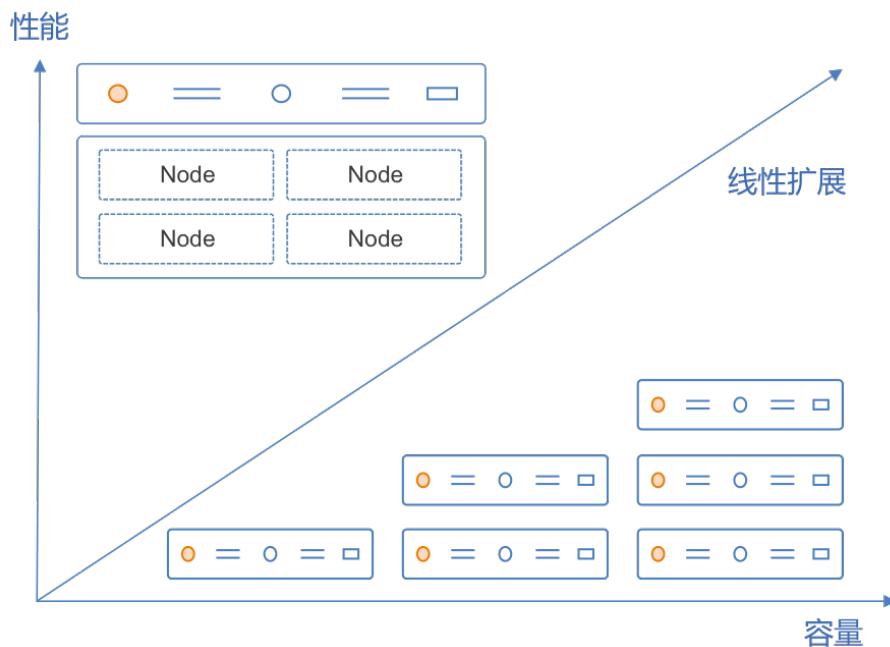
4.8 线性横向扩展 Scale-out

UMStor 存储系统采用分布式架构设计，节点由多个独立的 x86 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 (x86 物理服务器) 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。

UMStor 存储系统支持最小规模从 3 台 x86 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

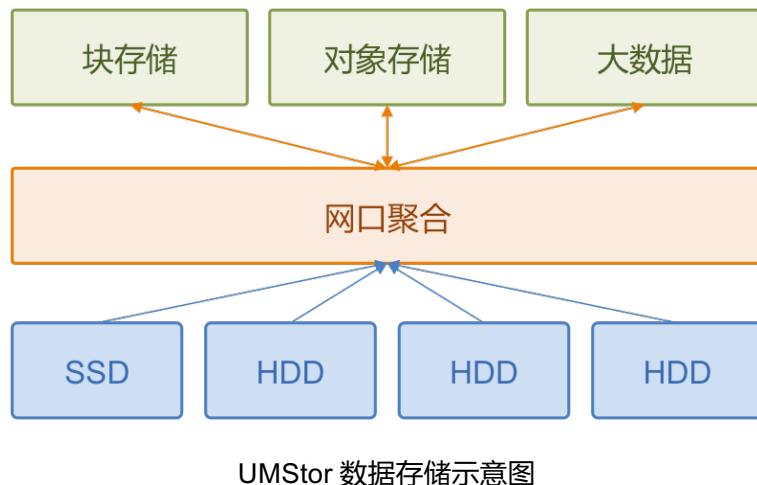
展，可以一次增加单台或者多台存储服务器，理论上没有容量限制。支持理论上无限水平扩展，支持 PB 级别的大规模存储。单集群可支持 200 个节点 5000 块硬盘，裸容量超过 50PB。

UMStor 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。UMStor 对象存储服务可支持 10 个集群，多个集群形成统一命名空间。



UMStor 容量系统线性扩展示意图

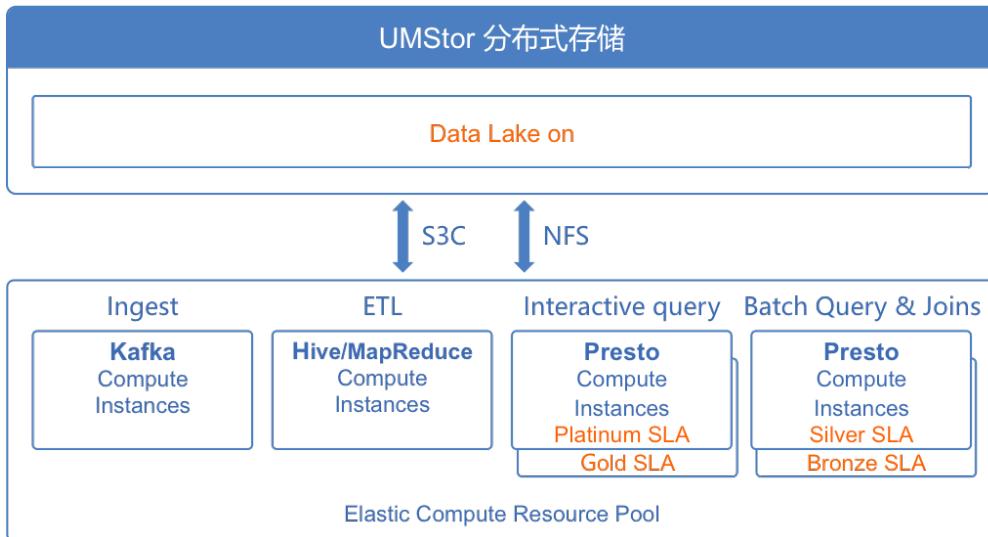
4.9 高聚合性能



UMStor 数据定位与访问不经过元数据服务器，没有元数据瓶颈。Client 和 Server 直接通信，不需要代理和转发。所有存储服务器和硬盘并行工作，数据分布在所有存储单元上，每个存储单元按权重承载工作负载，有效解决负载均衡问题，提高系统并行度。

- ▶ 单节点 4MB 文件读取 : 2000MB/s
- ▶ 单节点 4MB 文件写入 : 1000MB/s
- ▶ N 存储节点 4MB 文件读取 : N*2000MB/s
- ▶ N 存储节点 4MB 文件写入 : N*1000MB/s

4.10 数据湖存储平台



数据湖存储平台工作示意图

基于 UMStor 良好的在线扩展能力和低廉的存储成本，企业无需担心数据超出预期寿命后进行销毁以节省存储空间，可以方便的存储想要存储的所有时间跨度的数据，包括全生命周期的数据。

UMStor 可以对企业中的所有数据进行统一存储，从原始数据（源系统数据的精确副本）转换为用于报告、可视化、分析和机器学习等各种任务的转换数据，包括结构化数据从关系数据库（行和列），半结构化数据（CSV、XML、JSON 的日志），非结构化数据（电子邮件，文档，PDF）和二进制数据（图像、音频、视频）从而形成一个容纳所有形式的统一数据湖存储平台，便于企业在任何需要的时候对数据进行方便的管理、再处理和分析。

4.11 数据生命周期管理

根据行业规定要求，有些数据需要至少保存 N 天，有些数据在不同时期的访问频度不同，需要定期归档。UMStor 存储系统支持生命周期管理，可以设置被上传的文件对象在一

段时间后自动归档，迁移到其他介质存储池上以降低成本，也可以设置文件对象在一段时间后，被自动删除，既遵守了行业规定，又保证了存储空间的有效使用。

4.12 高效易用的管理平台

UMStor 提供友好的图形化操作界面，实现最大程度的部署运维自动化，配置管理和运营管理简单直观易用，有效降低学习成本和管理维护成本，提高效率。

The screenshot displays the UMStor graphical user interface. At the top, there's a header bar with the title 'UMStor' and a sub-title 'umstor181'. It also shows network activity: '6.25 KB/s' up and '0 B/s' down. On the right of the header are search, help, and user ('admin') icons.

The main area is divided into several sections:

- 集群概览 (Cluster Overview):** Shows cluster information (version v2.8.0, created 2019/1/7, 1 day running, work mode), cluster capacity (169.90 GB total, 9.21 GB used / 179.11 GB available), S3C and MCV interfaces, and a fault summary (0 alerts, 3 warnings, 0 notifications, 0 pending alerts).
- 存储池 (Storage Pool):** Displays 2 storage pools, both marked as '正常' (Normal) with 100% health.
- 对象存储 (Object Storage):** Shows 2 buckets and 1 user.
- 文件存储 (File Storage):** Shows 0 shared directories.
- 块存储 (Block Storage):** Shows 0 volumes and 0 client groups.
- 服务状态 (Service Status):** Monitors management services (3 healthy, 0 warning), storage services (9 healthy, 0 warning), object gateway (3 healthy, 0 warning), block gateway (3 healthy, 1 warning), and file gateway (3 healthy, 1 warning).
- 硬件资源 (Hardware Resources):** Includes tabs for 性能 (Performance), 容量 (Capacity), 配置 (Configuration), 可用性 (Availability), and 系统指标 (System Metrics).

UMStor 图形化操作界面

5 技术指标

5.1 技术规格

规格项	规格描述
系统支持的卷个数	131072
系统支持的快照个数	131072
系统支持的服务器个数	2048
系统支持的存储池个数	512
系统支持的客户端个数	4096
系统支持的存储桶个数	100 万
系统支持的对象个数	10000 亿
单存储池支持的最大硬盘数量	4096
单卷的最大容量	256 TB
单卷的最大快照个数	128
单存储桶的最大对象数	100 万

单服务器支持的最大硬盘数	48
数据冗余	1~10 副本 ; EC 纠删码
故障域隔离	支持多级故障隔离，包括服务器级别、机柜级别、机房级别

5.2 性能指标

存储资源池	指标	吞吐率
常规块存储 (三副本)	每节点 8KB 随机读	10,000 IOPS
	每节点 8KB 随机写	4,000 IOPS
	每节点 1MB 顺序读	2,000 MB/s
	每节点 1MB 顺序写	1,000 MB/s
高性能块存储 (三副本)	每节点 8KB 随机读	160,000 IOPS
	每节点 8KB 随机写	40,000 IOPS
	每节点 1MB 顺序读	2,000 MB/s
	每节点 1MB 顺序写	1,000 MB/s
大容量对象存储 (三副本)	每节点 64KB 文件读取	10,000 TPS
	每节点 64KB 文件写入	4,000 TPS
	每节点 4M 文件读取	2,000 MB/s

	每节点 4M 文件写入	1,000 MB/s
	每节点 64KB 文件读取	30,000 TPS
全闪存对象存储 (三副本)	每节点 64KB 文件写入	15,000 TPS
	每节点 4M 文件读取	2,000 MB/s
	每节点 4M 文件写入	1,000 MB/s