

闪存存储技术



前言

- 随着企业数字化转型的加速推进，数据存储已成为确保业务连续性的关键基础设施。华为OceanStor全闪存与混合闪存存储系列，凭借其端到端NVMe架构、SmartMatrix全互联设计以及AI智能引擎，为企业构建了一个具备“零延迟、零故障、零运维”特性的智能数据底座。
- 本章节将详细介绍华为闪存存储系列产品的软硬件架构、关键技术、增值特性以及典型应用场景部署等内容。

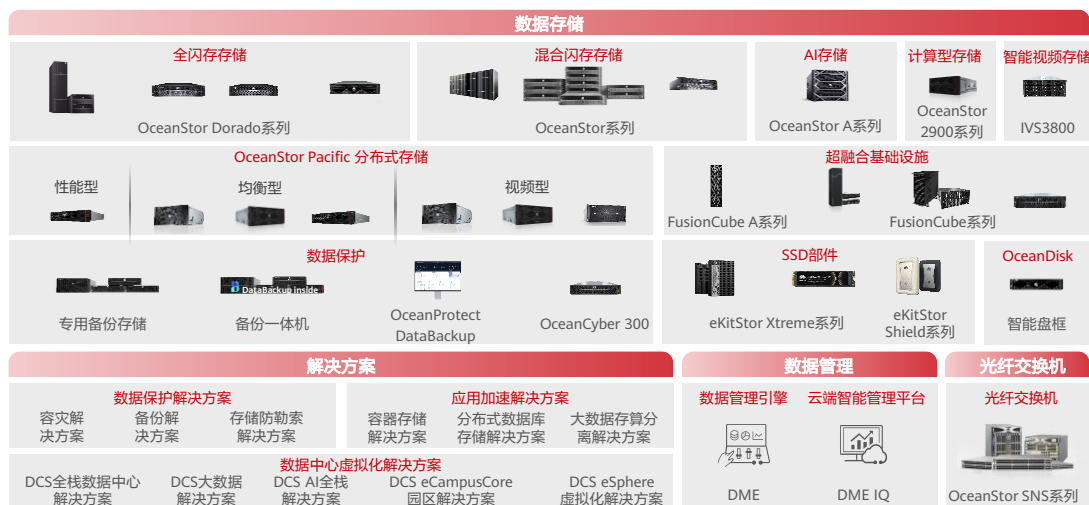
目标

- 学完本课程后，您将能够：
 - 了解华为闪存存储产品体系架构及应用场景
 - 熟悉华为闪存存储产品部署流程
 - 熟悉华为闪存存储增值特性

目录

1. **数据存储全景**
2. 闪存存储产品介绍
3. 闪存存储产品部署
4. 闪存存储增值特性

数据存储全景图

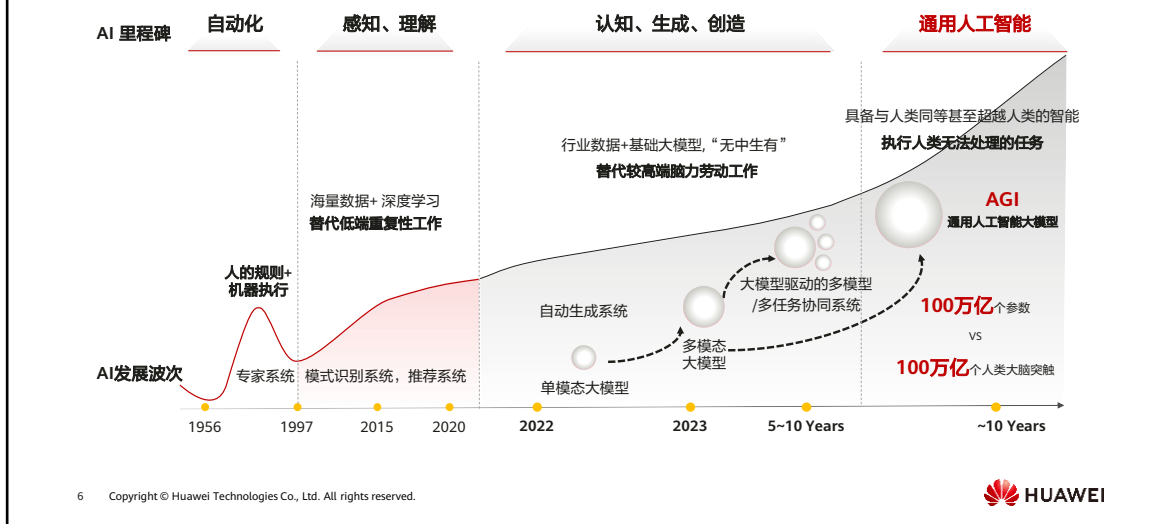


5 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



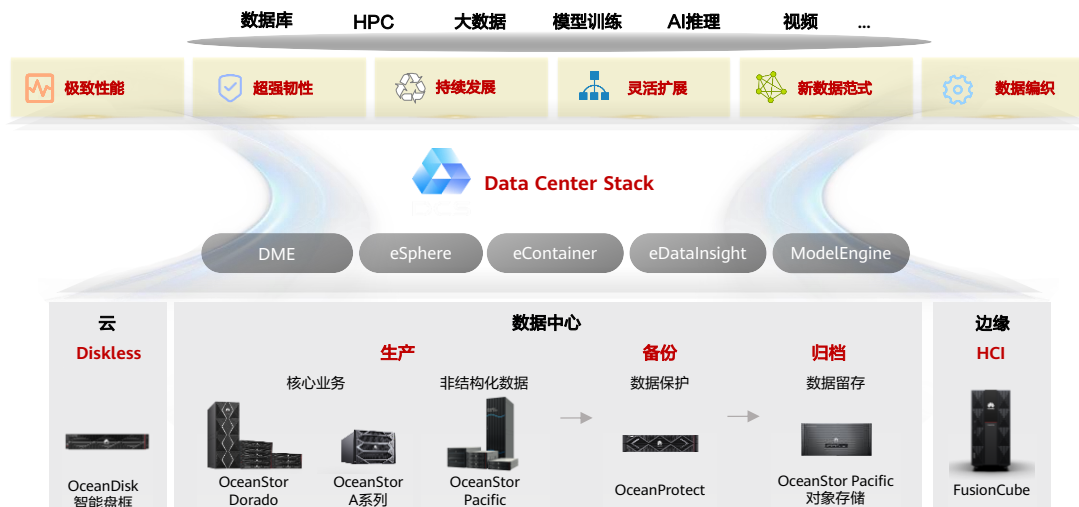
- 数字经济时代，数据作为新型的生产要素，已经成为基础性资源和战略性资源，我们正在迎来YB数据时代。数据应用蓬勃发展需要有强大、安全、可靠的数据设施保驾护航。
- 在数字化时代，存储发挥着至关重要的作用。存储容量决定了智能世界中数据的总量；企业拥有的数据越多，其数字化转型的基础就越坚实。
- 存储性能决定了数据存取的效率。只有当存储、网络 and 计算三者的性能完全匹配时，数据才能真正发挥其实时价值。
- 存储的可靠性确保了数据的持久性和可用性，一旦数据丢失或无法访问，将对数字社会和数字经济的运行造成不可挽回的损失。
- 存储安全作为数据安全的最后一道防线，包括识别和保护关键数据，检测和响应不可预见的攻击，并恢复关键数据，从而构成了完整的数据安全体系架构。

人工智能跨越拐点，从专用走向通用



- 随着AI技术的不断进步，大模型所使用的数据量和参数规模呈现“指数级”增长，带来智能算力和存力需求爆炸式增加。超大规模预训练模型技术不断成熟，10万亿多模态、超大规模预训练大模型是人工智能产业的标配，巨大的模型规模为其提供了强大的表达能力和学习能力。大模型需要大量数据进行训练，只有高质量大规模数据才能发挥大模型的参数规模优势。强大的计算资源让计算新范式加速实现，训练大模型通常需要数十万卡并行工作，数PB级的数据参与训练，单次训练也大量的时间，通常需要几周到几个月，因此对AI的基础设施，尤其是存储系统提出了高性能、大规模共享、高可靠的诉求。不仅如此，华为AI存储在高效支持训练的同时，也支持大模型推理加速能力，通过将推理过程中token生成的KV Cache持久化至存储，在后续的推理任务中以查代算，减少推理重复计算，加速推理过程，可降低首token响应时延（TTFT），提升推理吞吐性能，支持主流昇腾、英伟达推理生态，支持业界主流MindIE和vLLM等推理框架。

先进数据存力构筑AI时代根基



7 Copyright © Huawei Technologies Co., Ltd. All rights reserved.

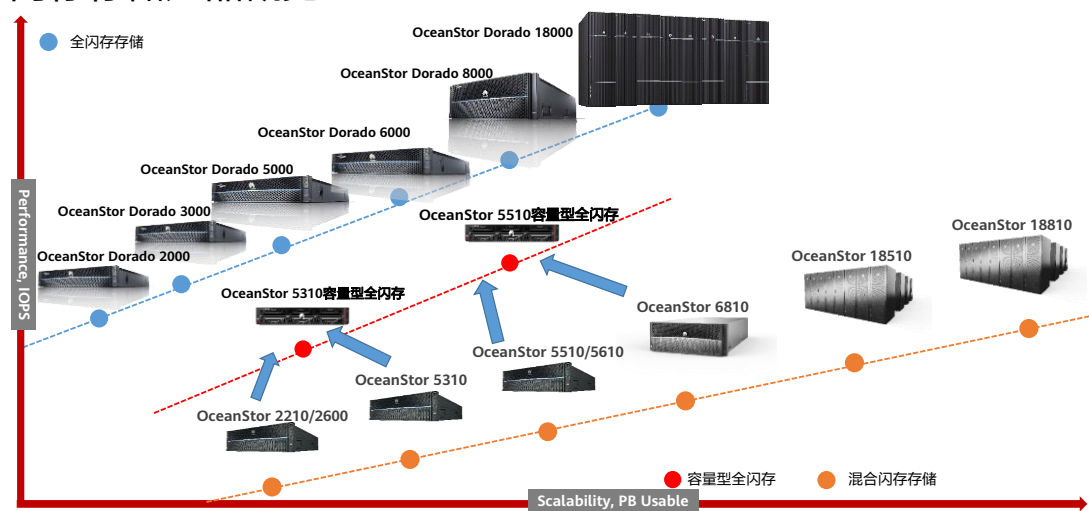


- 华为沿着“架构、介质、算法”三个方向不断突破创新，打造面向全域全场景的先进数据存力，帮助客户全面提升数据使用效率，使能上层数据应用，实现向智能化的跃迁。
- 全域全场景存储指的是华为存储打造全系列高性能高可靠的产品与解决方案，满足客户“生产-备份-归档”全业务场景的数据存储需求。此外华为存储还提供企业分支的超融合产品以及可用于云数据中心Diskless架构的OceanDisk智能盘框。
- 在提升数据使用效率方面，华为通过Omni-Dataverse全局文件系统实现跨域跨系统的数据可视可管和按需流动，数据调度效率提升3倍；通过eDataMate数据治理平台大幅提升数据处理效率，30+算子按需调用，将数据归集、清洗等预处理时间从30天缩短到1天，处理后的数据检索准确率可提升50%；通过DME数据管理引擎实现全生命周期数据管理，TCO节省20%。
- 基于这样的先进数据存力底座，华为在支持企业关键业务存储持续领先的同时，全面使能AI、大数据、云原生等新兴数据应用，助力客户持续成功。

目录

1. 数据存储全景
2. 闪存存储产品介绍
 - 产品体系
 - 产品架构
 - 应用场景
3. 闪存存储产品部署
4. 闪存存储增值特性

闪存存储产品概览



- OceanStor闪存存储采用全新一代的硬件平台和SmartMatrix全互联架构，基于全闪存深度优化的软件架构，结合多种高级数据应用和数据保护技术，使存储系统具有高性能、高稳定性、高可扩展性、高可靠性和高可用性等特点，满足大中型企业对存储产品的各种要求。并且OceanStor Dorado采用SAN/NAS一体化设计，不需要单独的NAS网关设备，一套软硬件同时支持SAN和NAS，支持NFS、CIFS、S3、HTTP、FTP、FC（支持基于SCSI接口的FC-SCSI协议和基于NVMe接口的FC-NVMe协议）、iSCSI、NVMe-oF等协议。SAN和NAS均支持多控的Scale-out技术，主机可以从任意1个控制器上的前端主机端口访问任意1个LUN或文件系统。
- 以2026H1公布信息为例进行展示。

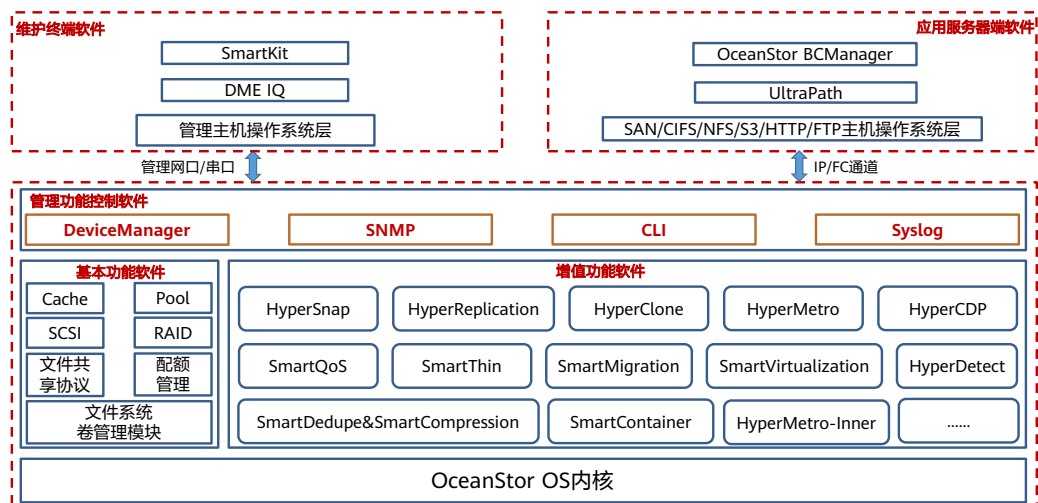
目录

1. 数据存储全景
2. 闪存存储产品介绍
 - 产品体系
 - **产品架构**
 - 应用场景
3. 闪存存储产品部署
4. 闪存存储增值特性

硬件架构

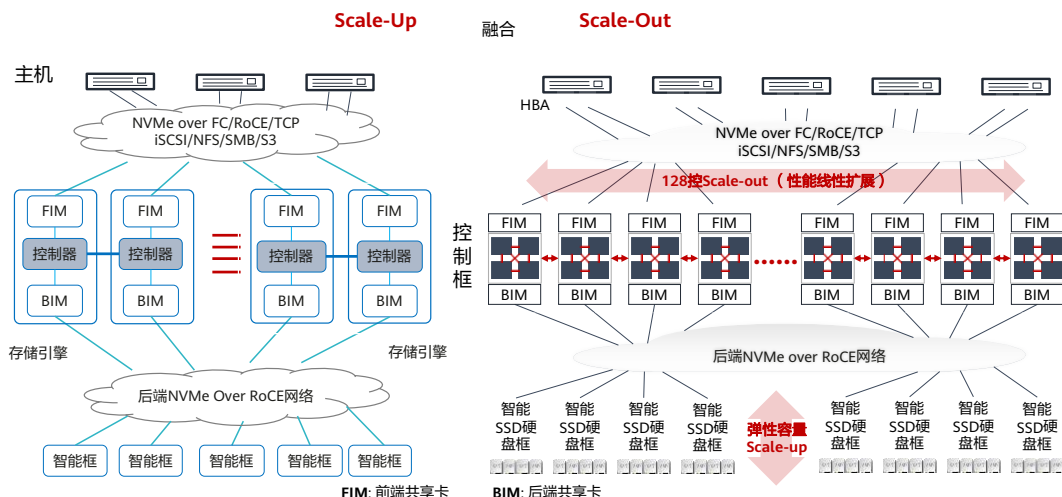
- 3D体验中心提供全新的硬件3D仿真交互，支持全方位浏览硬件模型并手动拆卸部件，了解各模块功能，并提供产品关联的文档、工具、多媒体、软件和成功案例等资源。您可以通过如下方式进行访问：
- 1. 访问华为数据存储信息中心（ <https://info.support.huawei.com/storageinfo/#/home> ）。
- 2. 在首页菜单“信息速查”页签，单击“3D体验中心”。
- 3. 选择“全闪存存储”或“混合闪存存储”。
- 4. 单击对应的产品型号。
- 5. 选择所需查看的物理设备，进入展示页面。

闪存存储功能架构



- 维护终端软件用于系统配置与维护。用户可以通过维护终端的SmartKit、DME IQ等软件对存储系统进行配置和维护。
- 应用服务器端软件可与存储系统通信，从而使双方能够配合执行某些操作。应用服务器端软件包括OceanStor BCManager和UltraPath。
- 存储系统端软件是华为自研的存储操作系统(OceanStor OS)，主要包括以下组件：
 - 基本功能控制软件：实现基础的数据存储和读写功能。
 - 增值功能控制软件：实现各种备份、容灾和性能调优等高级功能。
 - 管理功能控制软件：实现对存储系统的管理功能。
 - OceanStor OS内核：实现管理硬件和支撑存储业务软件运行的功能。

分布式全AA架构

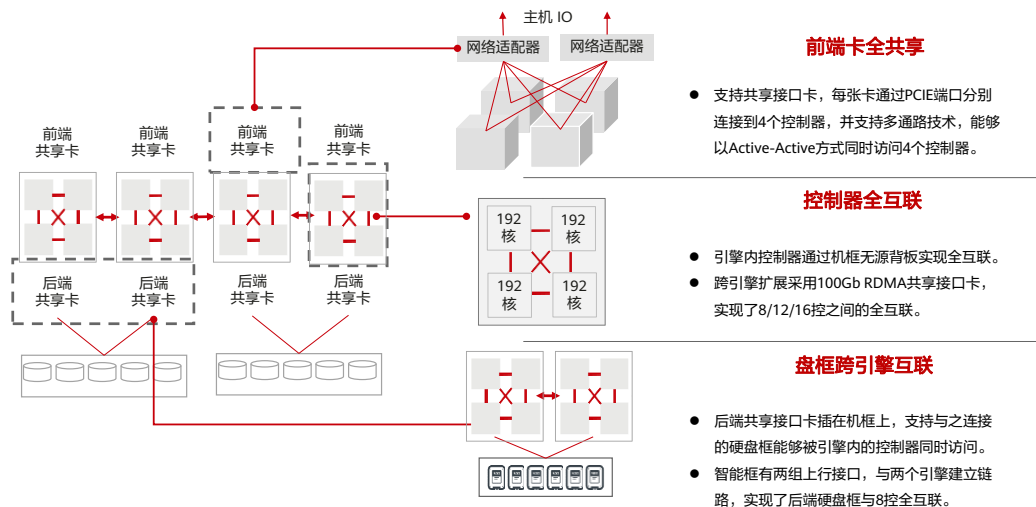


13 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



- OceanStor闪存存储采用SmartMatrix全互联均衡架构，该架构采用了具有高速、矩阵式全互联无源背板，可以连接多个控制器节点，同时接口模块（FC和后端级联）采用全共享方式接入背板，允许主机从任意端口接入，直达任意控制器进行处理。同时具备Scale-out和Scale-up能力，融合了企业存储和分布式存储双重的优势。
- 华为OceanStor系列中端产品的控制框采用盘控一体专用高可靠硬件设计，每个控制框内置2个控制器，框内控制器之间通过无源背板互连，采用RDMA协议高效协议通信。前后端接口模块和控制板采用PCIe3.0互联，提供前后端的存储访问协议能力。
- Scale-out：中高端支持128控多控制器扩展能力，新一代软件上具备Symmetric A-A能力，业务负载均衡到多个控制器上，做到随着控制器的增加性能线性扩展。
- Scale-up：新一代高端全闪存打破硬盘框归属与某一控制框的局限，控制框和硬盘框之间提供100G RoCE交换网络，实现硬盘框实现与控制框的全局互联。硬盘框和控制框的扩展解耦，客户按需选择，单独扩控制框或者单独扩硬盘框。

端到端全互联架构，业务持续在线



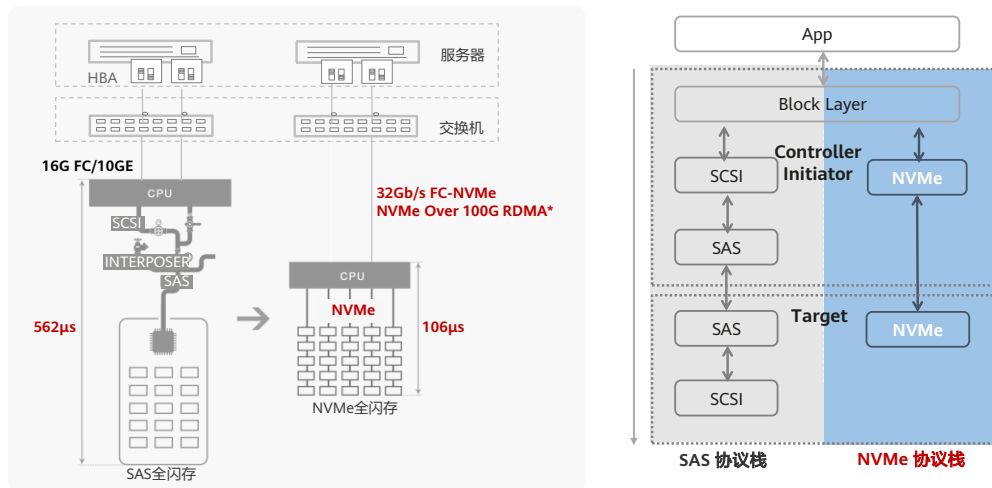
14 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



• 端到端全互联主要包括：

1. 前端4控共享互联：通过共享接口卡，实现接口卡与引擎内所有控制器的全互联。共享接口卡插在控制框中，每张卡通过4个PCIE 3.0 X4端口分别连接到4个控制器，共享卡支持Multi-Host（多通路）技术，能够以Active-Active方式同时访问4个控制器。
2. 控制器全互联：控制框内的4个控制器通过控制框无源背板上的12条高速100Gb RDMA通道进行连接，采用全对称交叉的连接方式，任何控制器之间有2条直连路径。扩展到8控至16控，可以采用交换机组网和免交换直连组网两种方式，实现4/8/12/16控之间的全互联。
3. 后端8控共享互联：通过后端共享接口卡进行后端扩展。共享接口卡插在控制框内，每张卡连接到引擎内的四个控制器，支持被四个控制器同时访问。同时，智能框能够同时与两个引擎建立链路，就实现了两个引擎（8个控制器）能够同时访问相同的盘框，实现了后端硬盘框与8控全互联。

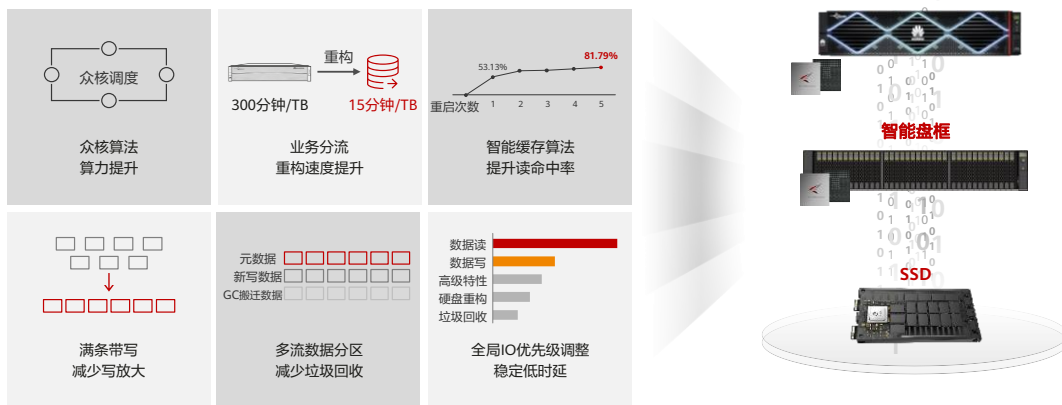
端到端NVMe架构



- NVMe是一种抽象协议层，提供可靠的NVMe命令和数据传输。为了支持数据中心的网络存储，通过NVMe over Fabric可以实现NVMe标准在多种网络上的扩展，达到降低存储网络协议栈处理开销并提供高并发低延时的目的，适应SSD介质驱动的存储架构演进。NVMe over Fabric支持把NVMe命令和数据映射到多个Fabric传输链路，主要包括FC、InfiniBand、RoCE v2、iWARP和TCP。
- 华为全闪存存储系统支持端到端的NVMe（部分型号不支持NVMe进行后端组网），包括：
 - 存储与主机网络支持FC-NVMe（NVMe over FC），后续将扩展到NVMe over RoCE v2，NVMe over TCP/IP。
 - 更高的接口速率，支持32Gbps FC-NVMe和100G NOF。
 - iSCSI连接TCP/IP协议栈卸载技术，减少CPU消耗以及降低网络协议栈时延。
 - 为鲲鹏920多核定制设计的NVMe多队列轮询，免锁处理多并发IO，处理器算力发挥更彻底。
 - 针对NVMe SSD访问而设计，对下盘请求提高读优先级调度，减少在写SSD盘的时候，对主机读请求的访问阻塞时长，使得更快响应。
 - 采用端到端NVMe设计的华为全闪存存储系统最低访问时延降低至100us以内。

FLASHLINK®

FLASHLINK® 智能算法



16 Copyright © Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

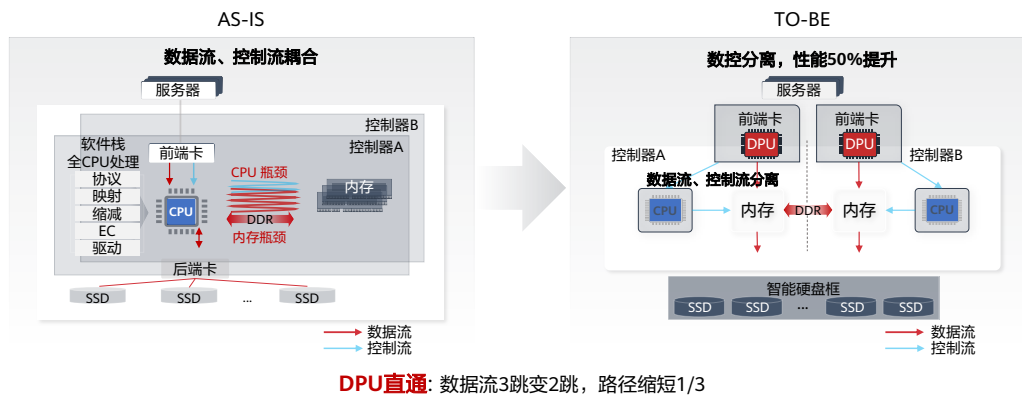
- FlashLink®技术的核心是通过一系列针对闪存介质的算法优化技术，实现了控制器板载CPU和SSD板载专用CPU的联动，保障了SSD算法在不同CPU之间的协同，实现系统的高性能和高可靠。FlashLink®技术主要包含以下关键技术：
 - 智能众核技术：采用高性能CPU，每个控制器内最多4颗CPU，多达192核；在同一个控制器内实现了业界最多的CPU个数和CPU核数，提供了强劲的运算能力。通过智能众核技术，充分发挥了CPU的强劲的运算能力，实现了性能随着CPU数量线性增加。智能众核技术，由CPU分组技术、业务分组技术、核间免锁技术组成。
 - 大块顺序写技术：SSD盘相对HDD，SSD盘片上的Flash颗粒具有擦写次数的限制。在传统的RAID覆盖写（Write In Place）的方式下，如果某块盘上的数据成为了热点，那么对这些数据的不断改写，就会导致对应的Flash颗粒的擦写次数很快的耗尽。OceanStor闪存存储通过对所有的写（包括新写数据和对老数据的改写等）都采用ROW（Redirect On Write）大块顺序写的方式，每次写通过分配新的Flash颗粒来写，实现每个Flash颗粒擦写次数均衡。同时，避免了传统RAID写流程所需的数据读和校验修改写而产生RAID写惩罚，有效降低了写入过程阵列控制器的CPU开销与对SSD盘的读写压力。相比传统的RAID覆盖写的方式，ROW大块顺序写方式使得各种RAID级别都能实现高性能。
 - 多流分区技术：SSD内部使用NAND Flash来作为存储介质，其逻辑结构如下图所示。每个SSD盘由多个NAND Flash颗粒组成，每个NAND Flash颗粒包含多个Block，每个Block包含多个Page(4KB 或 8KB)。由于NAND的特性，Flash颗粒中的Block需要先擦除再写。擦除的单位是Block，在擦除前需要把该

Block中的有效数据搬移走，对有效数据的搬移就产生了盘内写放大。通过多流写技术，对存储系统内的数据进行分类，不同类型的数据存放到SSD盘上不同的Block中，相同类型的数据同时为有效数据或者同时为垃圾数据的概率较大，减少了擦除Block时搬移的数据量，减少了盘内写放大，提升了SSD盘的性能和寿命。

- 智能盘框：智能硬盘框内有CPU和内存资源，能够对控制器进行卸载（比如盘故障重构），减轻控制器的负载。盘重构卸载，能实现在盘故障场景下，智能硬盘框的任务卸载可以大幅度减轻阵列控制器的压力。
- 智能技术：基于智能芯片 提供的强大的运算能力，构建了智能学习框架。通过对业务负载、设备健康状态等持续学习，系统越用越智能。基于智能学习强化的智能读缓存，能够对业务模型预测更准确，实现更高的读命中率，保障复杂业务模型下系统高性能；基于智能学习强化的智能服务质量控制（QoS），能够对系统内各种不同负载进行识别、分类，保障关键业务，抑制非关键业务，保证业务平稳运行；基于智能学习强化的数据缩减，能够对数据模式进行识别，合理统筹在线重删与后重删，以及针对不同数据模式采用不同的缩减算法，实现缩减比和性能最优。

数控分离技术

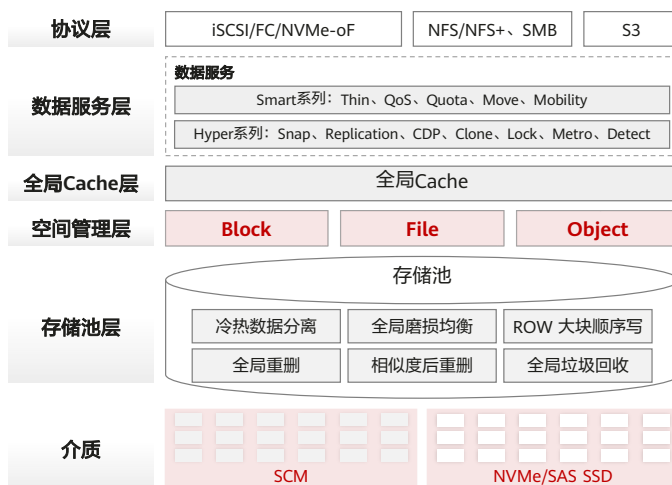
CPU为中心 → CPU+DPU 分布式算力架构



DPU直通: 数据流3跳变2跳, 路径缩短1/3

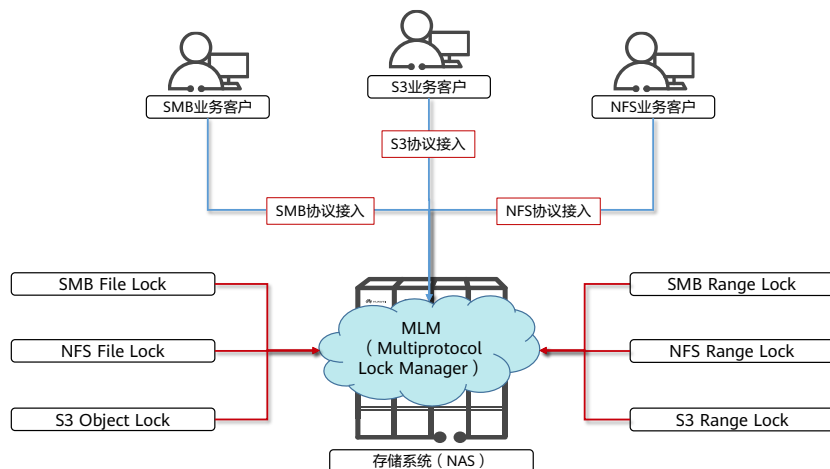
- 数控分离技术指的是对存储访问的请求区分请知中的命令控制部分和数据部分分离处理的技术。前端DPU卡、控制器CPU、盘控制器的多元算力协调配合，采用业务请求的控制流和数据流分离布局设计；控制流由控制器的CPU进行逻辑处理确保逻辑一致性，数据流则卡到内存、盘的直通共享访问。
- DPU（Data Processing Unit）是以数据为中心构造的专用处理器，一个具有加速能力并可卸载服务器的CPU功能的网络适配器，采用软件定义技术路线支撑基础设施层资源虚拟化，支持存储、安全、服务质量管理等基础设施层服务。
- 相比传统的数控一体架构中以CPU算力为中心的模式，数控分离架构采用的CPU算力+DPU算力+盘算力的混合算力模型，实现接口模块、控制器和SSD盘协同处理，数据不随业务逻辑反复拷贝，有效规避了CPU的算力和内存总线的放大瓶颈，实现I/O的随路处理能力，达成总线级的系统端到端带宽能力。

协议融合



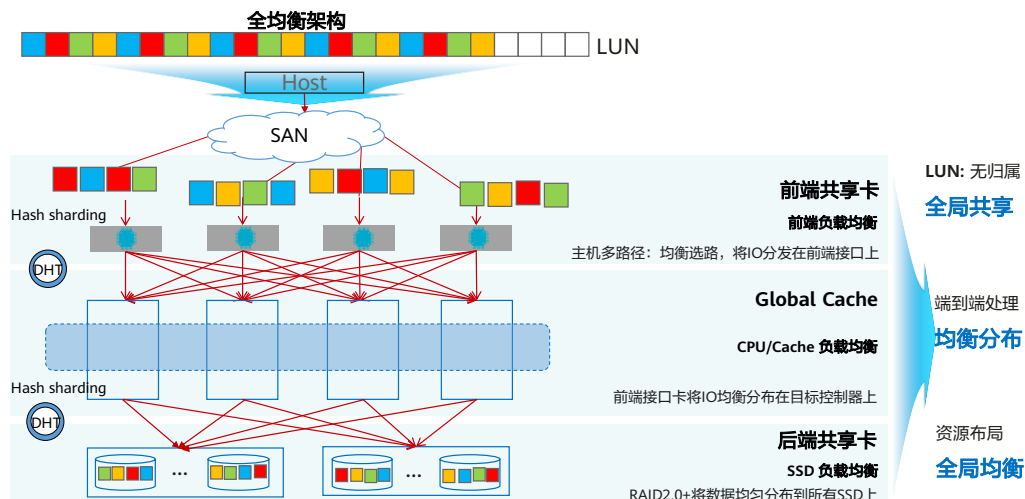
- OceanStor闪存存储采用SAN/NAS/Object一体化设计，无需单独的NAS/S3网关设备。一套软硬件同时支持SAN和NAS、Object，支持NFS、CIFS文件访问协议、S3对象协议以及FC、iSCSI、NOF等块访问协议。SAN和NAS、Object均支持多控的Scale-out技术，主机可以从任意一个控制器的前端主机端口访问任意一个LUN、文件系统或对象。文件系统、LUN和对象均直接从存储池分配空间并提供服务，形成平行架构。这意味着LUN、文件系统和对象可以直接与底层存储池进行交互。相较于传统架构中在LUN基础上创建文件系统或对象，或在文件系统上创建LUN和对象的方式，这种软件架构能够为不同类型的服务提供最短的I/O路径，从而提升存储的读写效率。同时，不同服务之间的空间管理保持独立，互不影响，进一步增强了系统的可靠性。
- 存储池 (Storage Pool): 基于Flashlink技术提供全局统一的存储池服务，采用ROW技术进行空间分配，为LUN、文件系统、对象提供存储空间，具备全局重删、压缩功能，识别元数据与数据并分流写入SSD,支持RAID2.0+全局快速重构和后台垃圾回收。
- 空间管理层 (Space Management): 基于Thin Provision机制，为LUN、文件系统、对象提供空间分配与回收管理。
- 全局 Cache (Global Cache): 全局统一的Cache服务，为LUN、文件系统、对象提供读写缓存及元数据缓存。
- 数据服务层 (Data Service Layer): 统一的数据复制、复制管理、配置管理和复制网络管理服务，为LUN、文件系统、对象提供远程复制、双活等容灾能力。
- 协议层 (Protocol Layer): 为LUN、文件系统和对象提供协议服务，负责协议解析、I/O收发及错误处理。

跨协议访问



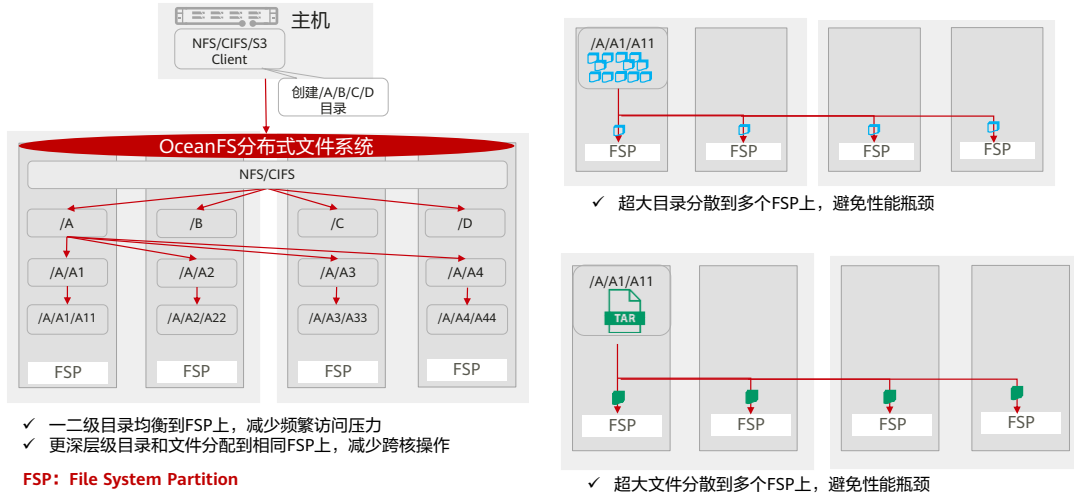
- NFS/SMB/S3跨协议访问，能为一个文件系统既配置NFS共享服务又配置SMB共享服务，同时还能通过S3协议共享访问。系统通过Multiprotocol Lock manager进行分布式锁管理，保证 NFS/SMB/S3 能互斥的访问相同的文件，不会造成文件损坏或数据不一致。
- 跨协议访问场景下支持4种安全模式：
 - NT 模式：仅允许在Windows端通过SMB设置文件的属性和ACL权限，系统会自动根据配置好的用户映射关系，将SMB用户的文件权限映射到NFS端，从而NFS/S3端使用对应的Linux 用户来访问该文件时可以成功鉴权。在NFS/S3端该模式会禁止用户设置文件的Mode和ACL权限。
 - UNIX模式：仅允许在Linux/UNIX端通过NFS/S3设置文件的权限，系统会自动根据配置好的用户映射关系，将Linux用户的文件权限映射到SMB端，从而SMB端使用对应的用户来访问该文件时可以成功鉴权。在SMB端该模式会禁止用户设置文件的ACL权限。
 - Mixed模式：UNIX客户端和Windows客户端均可进行权限设置，文件权限会相互覆盖，以最后设置的权限为准。
 - Native模式：UNIX客户端和Windows客户端均可进行权限设置， NFS/SMB/S3端权限独立存储，各自鉴权。

Active-Active的结构化数据（SAN）逻辑架构



- 在ALUA（Asymmetric Logic Unit Access）架构中，LUN有归属控制器，客户在创建LUN的时候，需要对LUN的归属进行规划，以尽可能的实现系统中每个控制器的负载均衡。但是由于不同LUN的业务压力各不相同，同一个LUN不同时段的压力也不相同，在实际情况下ALUA架构难以实现系统的负载均衡。
- OceanStor闪存存储采用Symmetric Active-Active架构。通过均衡算法，实现每个控制器接收到的主机读写请求是均衡的；通过全局缓存技术实现LUN无归属，每个控制器收到的读写请求，就在本控制器处理（而不像ALUA存储需要转发到LUN归属控制器处理），实现了控制器压力均衡；通过RAID2.0+技术，数据均匀的分布到存储池内的所有硬盘上，实现盘的压力均衡。
- 每个主机读写请求需确定存储系统中哪个控制器进行处理，这是通过计算读写请求LBA的哈希值来实现的。华为的主机多路径（UltraPath）、前端共享接口模块以及控制器采用了相同的哈希计算方法和参数，从而实现智能的请求分发。在没有使用华为多路径及前端共享接口模块的情况下，控制器在接收到主机请求后，会根据主机请求LBA的哈希计算结果，将请求转发至最终处理的控制器，确保每个控制器处理的主机读写请求均衡分布。
- 华为主机多路径将主机I/O请求发送至最优控制框的共享接口模块，共享接口模块再将I/O请求发送至最优控制器进行处理。在单控制框场景下，无需依赖多路径能力，共享接口模块即可实现I/O请求的最优分发，有效避免了框内控制器间的转发。而在多控制框场景下，推荐使用华为多路径软件，以有效避免跨框之间的前端I/O转发。

Active-Active的非结构化数据（NAS&对象）逻辑架构

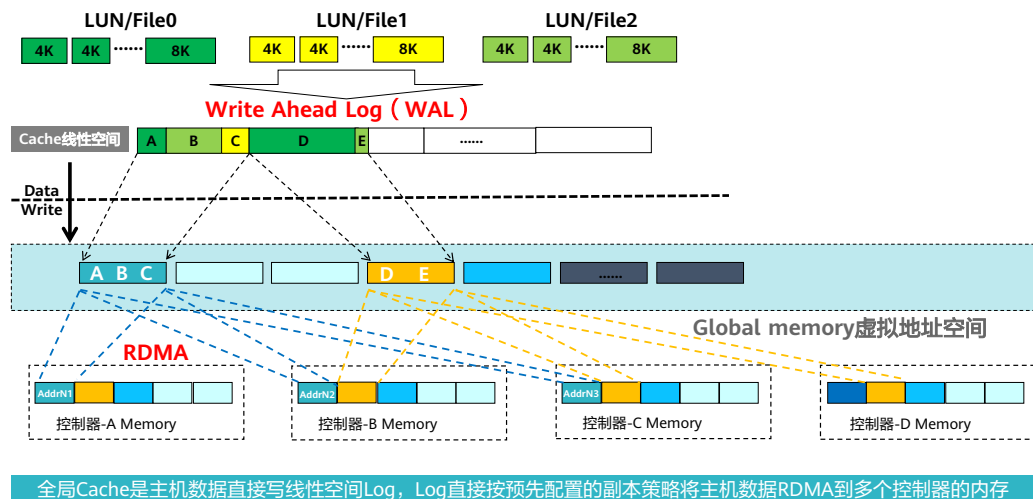


- 在传统的NAS文件系统架构中，业界通常采用Active-Passive架构。在这种架构下，文件系统归属某个特定的控制器。在创建文件系统时，需要对文件系统的归属进行规划，通过在不同的控制器上运行多个文件系统来实现系统中每个控制器的负载均衡。然而，这种NAS架构存在一定的局限性。如果仅有一个文件系统，那么只能发挥一个控制器的硬件性能，无法充分利用多控制器的硬件资源以获得更高的性能，因此无法支持单一命名空间（Single Name Space）。而如果采用创建多个文件系统的方式，由于不同文件系统的业务压力各不相同，也很难实现系统的负载均衡。
- 相比之下，OceanStor闪存存储的NAS采用了分布式文件系统架构。在这种架构下，文件系统没有归属特定的控制器。通过均衡算法将文件系统的目录和文件分配到每个控制器上，实现了每个控制器接收到的主机读写请求的均衡分布。这使得即使只有一个文件系统，也能充分利用整个存储集群的资源。因此，客户可以根据自己的业务需求灵活地选择使用单一命名空间的文件系统或多个文件系统。
- OceanStor闪存存储的分布式文件系统架构兼具海量小文件和大文件并存的文件共享场景，采用基于目录的分布式打散策略，以目录为粒度将数据均衡打散写入到各个控制器，达到负载均衡的目的。目录与目录下的子文件归属相同的控制器进行I/O处理，避免跨控制器转发，以提升目录遍历查询、属性遍历查询、批量属性设置等场景的性能。针对大文件，当大文件写入存储池时，通过RAID2.0+的全局数据块打散技术，可以将大文件的数据块打散到存储池的所有硬盘上，提升大文件的写入带宽。
- 业务数据下发时，数据根据FSP（File System Partition）视图进行处理。每个控制器在全局FSP视图中都有对应的FSP。当客户端发出请求时，全局FSP视图对该请求进行分发。客户端的请求通常归属在请求接入节点的FSP，以便对本地接入的客户端提供

就近读服务。

- OceanStor闪存存储文件系统支持两种目录均衡打散策略：
 - Local Allocation：主机写入的目录优先从所挂载IP的控制器直接写入和读取，以获得最佳性能和时延。系统后台根据各个控制器管理的数据容量、文件数量和负载，动态调整远端分配比例，即当某个控制器的负载、管理的数据容量或者管理的文件数量高于其他控制器一定阈值后，系统会自动将远端控制器分配比例调高，以让每个控制器的负载和管理容量达到均衡。推荐与OceanStor混合闪存的内置DNS服务功能配合使用，内置DNS服务将自动为主机挂载进行均衡的分配。
 - Balance Allocation：主机写入的目录直接进行按控制器负载均衡的打散写入，以在一开始就获得最佳的负载均衡能力，系统后台仍然会根据各控制器的管理容量等信息进行均衡调整。此模式为OceanStor闪存存储的默认分配模式，适合主机挂载存储的IP地址不均匀、压力不均匀，应用与对时延不高的文件共享场景。

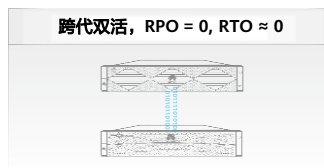
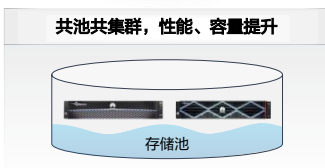
全局Cache均衡



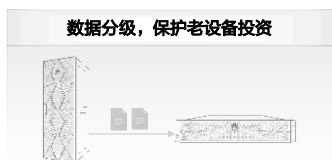
- 闪存存储把IO从前端链路均衡分发到控制器上，采用全局Cache接收数据。全局Cache是基于RDMA & WAL设计，能做到更短时延下把数据均衡到各控制器。
- 采用WAL的日志Cache机制的时延为20~50us，采用传统流程交互Cache机制时延为95us，其处理过程对比如下：
 1. 日志Cache是主机数据直接写组成线性空间，直接按配置副本策略把主机数据RDMA到多个节点上；
 2. 流程交互Cache机制是数据通过消息发送到远端节点，远端节点将数据写入本地Cache缓存，应答消息返回本节点，本节点再将数据写入本地Cache。
- 全局Cache采用类似操作系统虚拟内存地址空间管理方法，把各个控制器的内存统一编址，组成全局的DSM内存空间进行访问。对于写操作，Cache将数据写入线性空间，Cache线性空间由Plog组成，每个Plog是三副本模型，上层业务对于Cache全局地址空间（AddrL，Len）写操作，将会转换为对控制器A（AddrN1，Len）、控制器B（AddrN2，Len）和控制器C（AddrN3，Len）的并行写操作。

跨代融合

两代产品共框、共集群、共双活，提升性能、容量及可靠性

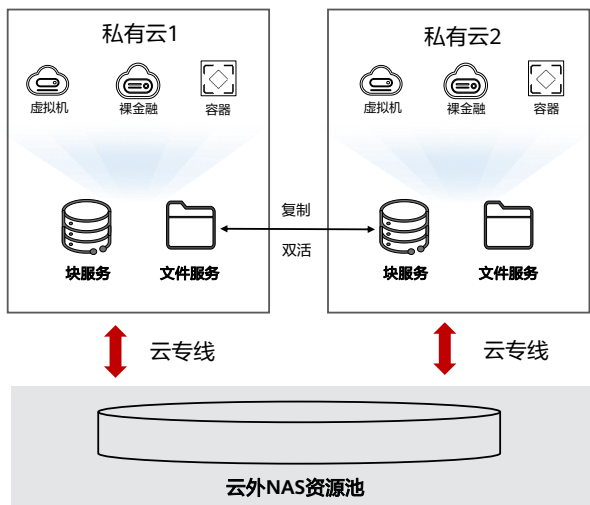


数据迁移、流动以及容灾，投资充分保护



- FlashEver设备永新功能，在业务不中断情况下实现跨代次存储设备在线升级替换。
FlashEver设备永新支持控制器永新和控制框永新两种实现方式：
 - 控制器永新：通过使用新一代兼容的控制器在线更换V6代次控制器，在不中断业务、不迁移数据的情况下实现控制器跨代次升级。
 - 控制框永新：通过在线扩容新一代控制框的方式，实现跨代次设备共集群。

云融合



多云对接

- 支持与HCS，腾讯，阿里云、青云及多种云平台对接，具备兼容性认证
- 提供Cinder和Manila接口为OpenStack云提供块及文件服务

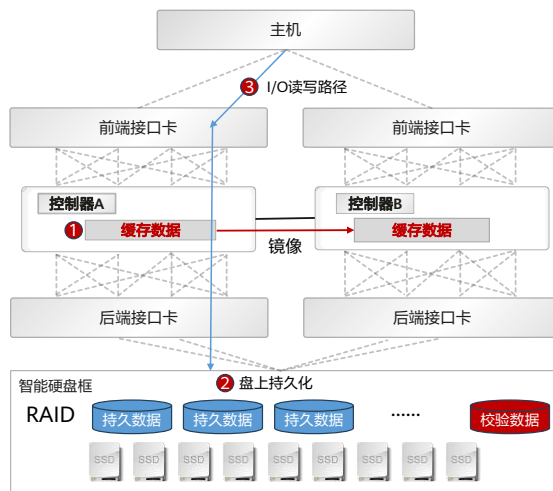
云内高性能高可靠

- 端到端NVMe为混合云提供云内数据库极致时延保障
- OceanFS文件系统为云内小文件场景提供10倍性能提升
- 支持跨AZ，跨Region的数据复制及双活解决方案，保障云内高可用
- 联合HCS组成云内防勒索解决方案，保障关键数据安全

跨云数据共享

- 云外专线对接，提供NAS文件服务，可以跨云进行数据流动及共享

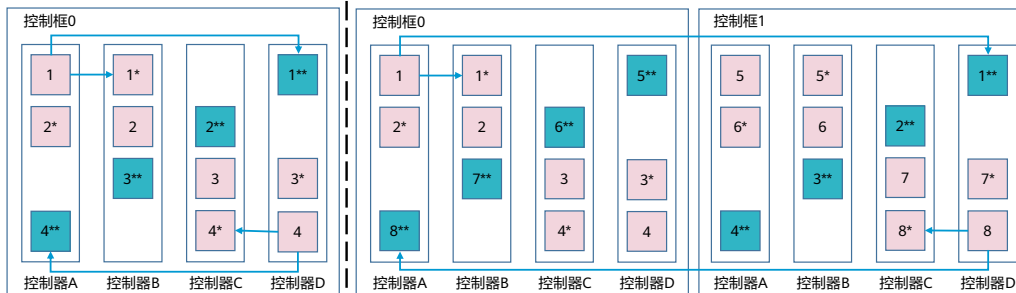
数据可靠性



- 主机写入到存储的数据，会经历数据缓存、盘上持久化和数据路径传输三个过程。下面我们分别对这三个过程的数据可靠性措施进行说明。
- 缓存数据可靠性
 - 为了提升写入数据的速度，OceanStor闪存存储提供了写缓存机制。即数据写入到控制器的内存缓存及其跨控的副本就向主机返回成功，然后后台将缓存数据下盘。
 - 储存于控制器内存中的用户数据，一旦遇到系统掉电或是控制器故障，将存在内存数据丢失进而导致内存数据丢失的风险。为了应对上述故障模式，系统提供了缓存跨控多副本和掉电保护功能保证数据可靠性。
- 持久数据可靠性
 - 通过盘内RAID技术保证单盘数据可靠性，保证盘内数据不丢失；通过系统RAID2.0+技术、缩列重构保证系统级的数据可靠性，即冗余范围内的单盘故障，多盘故障也不会导致数据丢失或是冗余度下降。
- I/O路径数据可靠性
 - 数据在存储系统内部传输时，会经过了多个部件、多种传输通道和复杂的软件处理，其中任意一个错误都可能会导致数据错误。如果这种错误无法被立即检测出来，就可能将这些错误的写入到持久化的盘、内部计算或是将错误的返回给主机，造成业务异常。
 - 为了解决上述问题，OceanStor闪存存储通过端到端PI功能保证传输路径数据错

误（数据块内部跳变）能被检测出来并纠错。通过矩阵校验功能保证数据整体跳变（整块数据被老数据或其它数据覆盖）也能被检测出来。通过上述措施，保证 I/O 路径数据可靠性。

缓存多副本

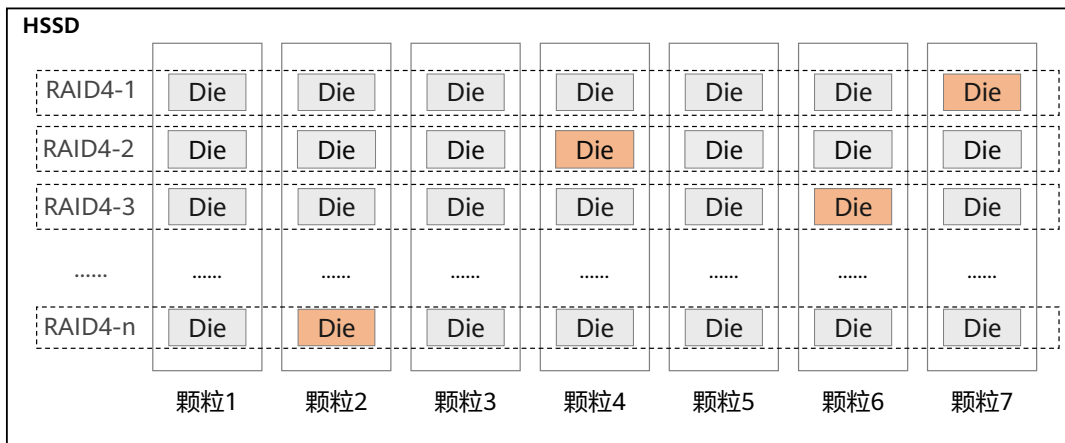


当只有一个控制框时，第三个副本在本控制框内，可以保证任意两个控制器同时故障数据不丢失；

在大于等于两个控制框时，可以将其它控制框内的控制器作为第三个副本的存放的控制器，除可以保证任意双控制器同时故障数据不丢失外，还可以保证单个控制框（包含4个控制器）故障，数据不丢失。

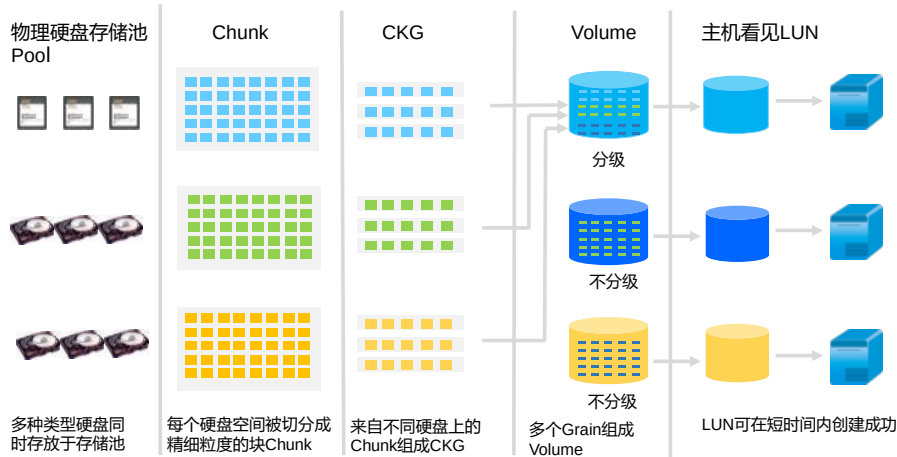
- OceanStor闪存存储支持两副本和三副本两种模式（中端系列只支持两副本模式）。通过设置三副本模式，实现了存储系统两个控制器同时故障，写缓存数据不丢失，业务不中断。对包含两个控制框的8控系统，三副本会分布在两个控制框内，保证一个控制框故障，写缓存数据不丢失。
- 在只有一个控制框的场景下，在控制框内选择多副本；在有两个控制框（并且有内双活License的场景下）场景下，除在当前框内选择副本外，还会跨控制框选择另外一个副本。
- OceanStor闪存存储内置了BBU模块（Backup Battery Unit，备份电池单元），当存储阵列在供电故障后，能利用BBU模块将各控制器内存中的缓存数据刷入到保险箱中。待阵列电力恢复后，系统在启动时，再将保险箱中的缓存数据恢复到内存，保证数据不丢失。

盘内RAID



- 硬盘除了会出现整体故障外，用于存储数据的颗粒也会出现区域性损坏，称之为静默失效（坏块）。这些坏块不会体现为单盘故障，但会导致盘上相应的数据无法访问。
- 通过常规的坏块扫描可以提前发现静默失效数据并进行修复，但因为扫描对盘访问会占用较多资源，为了避免影响前台业务，扫描速率必须有所控制，在盘容量且数量较大时，扫描一次花费时间少则数周，多则数月才能完成所有盘的扫描工作。而在扫描间隔期间，出现坏块并叠加盘故障，则可能导致数据损坏无法恢复。
- 在坏块扫描基础之上，为了应对扫描隔离内的静默失效问题，OceanStor闪存存储结合HSSD(Huawei SSD)，提供了盘内RAID特性。即针对盘内的数据，以Die为单位建立RAID，通过RAID4形成冗余，容忍单个Die失效，盘内数据不丢失。

RAID2.0+



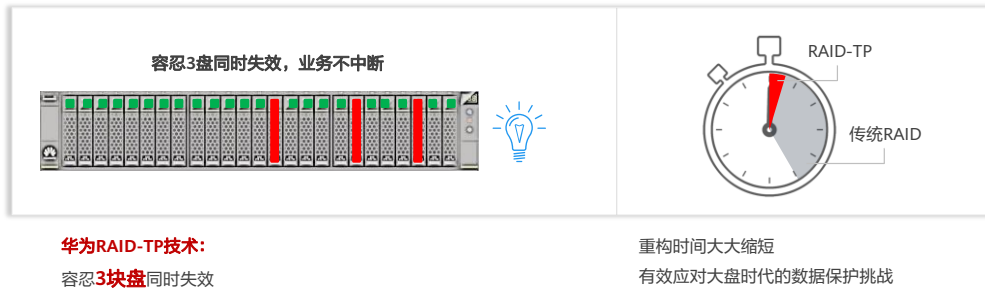
28 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



- 传统RAID的存储系统中RAID组的成员盘是固定的几个物理盘，用户使用的 LUN/FS 从某一个RAID组中划分。由于系统中每个LUN/FS的访问频度不同，这就导致系统中一部分 RAID 组中的硬盘异常繁忙，形成“热点”，而其它RAID组的硬盘即使空闲，也无能为力。另外，硬盘如果长期工作，它的故障率就会明显升高，结果快速故障。因此，传统RAID存储系统中的“热点”硬盘，时刻面临负载不均和部分盘着“过劳死”的风险。
- OceanStor闪存存储将每个SSD盘切分成固定大小的Chunk（简称CK，大小为4MB），将所有盘上Chunk按RAID冗余组成Chunk组，形成RAID2.0+，相对于传统RAID机制，RAID2.0+具备如下优势：
 - 业务负载均衡，避免热点。数据打散到资源池内所有硬盘上，没有热点，硬盘负荷平均，避免个别盘因为承担更多的写操作而提前达到寿命的上限。
 - 快速重构，缩小风险窗口。当硬盘故障时，故障盘上的有效数据会被重构到资源池内除故障盘外的所有盘上，实现了多对多的重构，速度快，大幅缩短数据处于非冗余状态的时间。
 - 全盘参与重构。资源池内所有硬盘都会参与重构，每个盘的重构负载很低，重构过程对上层应用无影响。

RAID-TP

- RAID保护是存储系统必备的基本功能，提供高可靠性和性能保证，但随着容量剧增，RAID重建的时间越来越难保证，RAID保护的可靠性降低。
- RAID-TP功能实现了最佳性能、可靠性和容量利用率。



- 现有系统的持续升级意味着客户需要不断购买更高容量的磁盘来逐渐取代现有磁盘，不同容量的磁盘将用于同一系统中。问题是如何在不同容量磁盘混用的环境中保证最佳利用率。
- RAID-TP采用了华为优化的FlexEC算法，这个算法允许系统容忍三盘同时故障，这样不仅可以提高可靠性，还允许更长的重建时间。
- 与传统的RAID系统相比，使用FlexEC算法重建时实际从单个磁盘读取的数据量减少了70%，降低了对系统性能的影响。
- 在典型的4:2 RAID 6系统中，利用率通常约为67%，而配备25个磁盘的华为OceanStor全闪存存储的容量利用率则大幅度提高了20%。

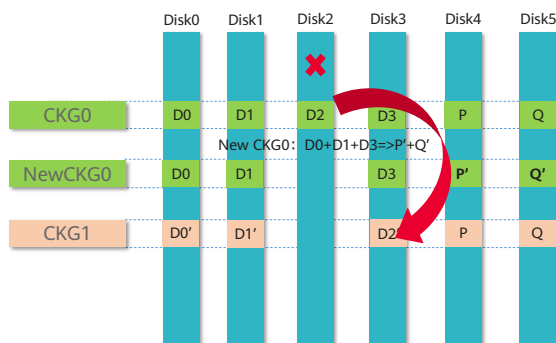
盘级冗余与框级冗余

- 框级冗余在满足盘级冗余的前提下，考虑硬盘和硬盘框的归属关系。每次组建Chunk组时从每个硬盘框内最多选2块盘的Chunk组成Chunk组，Chunk组的Chunk数量由硬盘框的数量决定。
- 以系统有4个框，每个框3盘的配置，采用RAID6为例，相同颜色的Chunk组成1个Chunk组。冗余盘数为2，因此框故障后，每个Chunk组只丢失2份数据，未超出Chunk组的冗余数量，业务不中断。



- 盘级冗余策略（普通RAID模式）是把CHUNK打散到不同的硬盘上，支持不超过RAID冗余能力的硬盘故障，不能直接支持硬盘框故障。金融核心交易系统等核心业务场景对存储系统的可靠性提出了更为苛刻的诉求，要求存储设备能容忍单个硬盘框故障。基于此，华为OceanStor存储系统的框级冗余策略（跨框RAID特性）应运而生。当存储池具备框级冗余能力时，可以实现：
 - RAID 6支持单硬盘框故障，业务不中断、数据不丢失。
 - RAID-TP支持单硬盘框故障后，再故障另外一块硬盘，业务不中断、数据不丢失。
 - 故障的单盘在热备重构完成后，支持再故障单个硬盘框。
 - 故障硬盘框恢复后，自动恢复框级冗余的能力。

动态RAID重构技术



- **冗余度不降**: 缩列重构根据硬盘配置和运行状态, **智能判定RAID列数, 各个场景下RAID冗余度不降**, 保证数据高可靠
- **智能恢复**: 盘故障恢复, 硬盘域总可用硬盘数满足RAID成员盘数时, RAID数据列自动恢复正常值

- 1、Disk2故障后, 硬盘域内可用盘数少于RAID成员盘数时, 触发缩列重构;
- 2、数据D0+D1+D3缩列计算出校验列P'Q' 构成NewCKG0;
- 3、故障盘数据D2新写入成D2'与D0',D1',Q组成CKG1 **缩列重构后, RAID冗余级别不变**

- 传统硬盘域因为连续盘故障、盘更换等原因导致可用成员盘数小于RAID成员盘数时, 将导致重构无法进行, 用户数据冗余无法保证。
- 为了应对该问题, 存储系统采用动态RAID重构技术 (也称缩列重构技术), 重构粒度为CK。即存储池总的可用硬盘小于RAID成员盘数, 缩列重构时保持M (校验列) 不变, 减少N (数据列) 的方式进行重构, 重构前后RAID校验列数不变, 数据列数变少。缩列重构完成后, RAID组成员盘数减少, 但是RAID冗余级别不变。当故障盘更换完成后, 系统会根据存储池内的可用硬盘数, 增加N (数据列), 新写数据就会采用新的RAID方式, 故障期间写的数据也会逐渐转换为新的RAID方式。

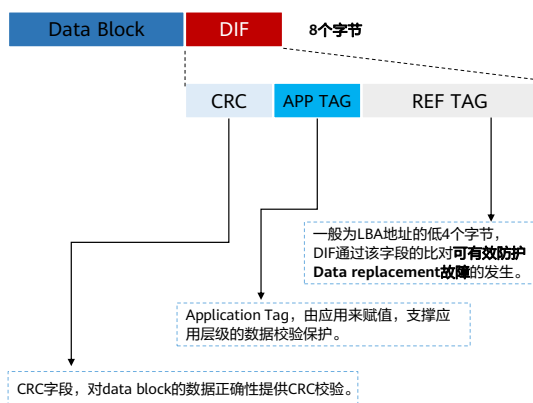
盘内DIF

静默数据破坏

数据在传输过程中，经过了多个部件、多种传输通道和复杂的软件处理，其中任意一个错误都可能会导致数据错误。但是这种错误**无法被立即检测出来**，而是在后续的数据检查和访问中才发现数据已经出错。

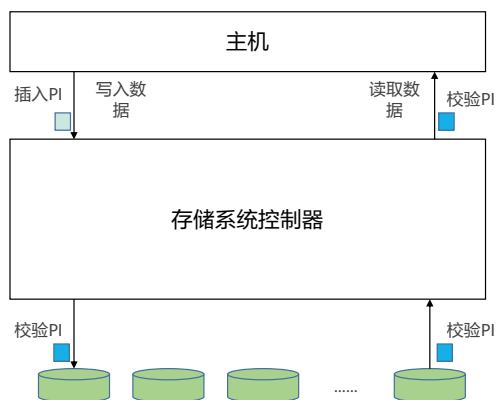
产生静默数据破坏的原因：

- 硬件错误：内存、CPU、硬盘、数据传输链路等
- Firmware错误：HBA、硬盘等
- 软件bug：系统软件、操作系统、应用程序等

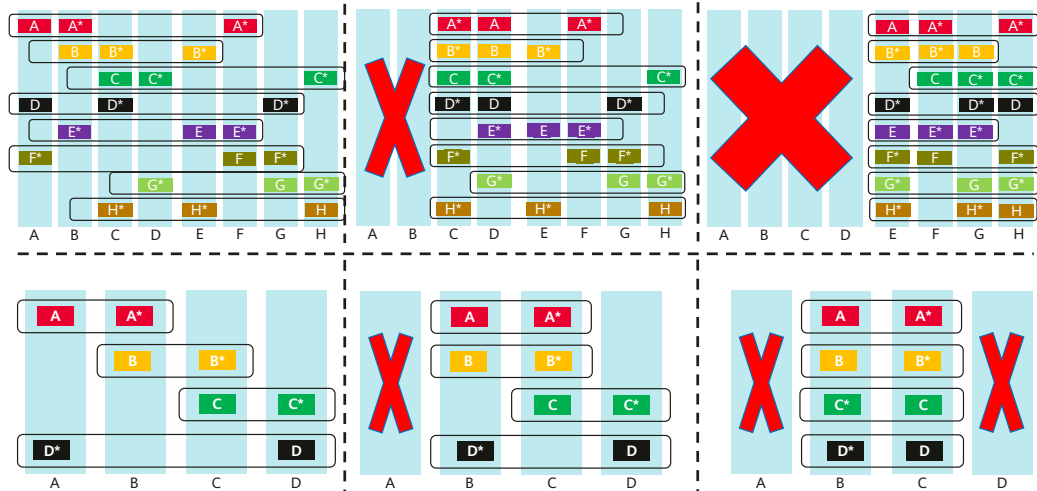


端到端PI (Protection Information)

- 当数据从主机下发到达阵列后，阵列会在收到数据后，首先对每512个字节插入8字节的PI，然后再做内部处理。
- 数据在达到盘以后，盘本身会对数据进行PI校验，保证数据在从前端接收（插入PI）到下盘这段时间出现跳变或改写能被校验出来。

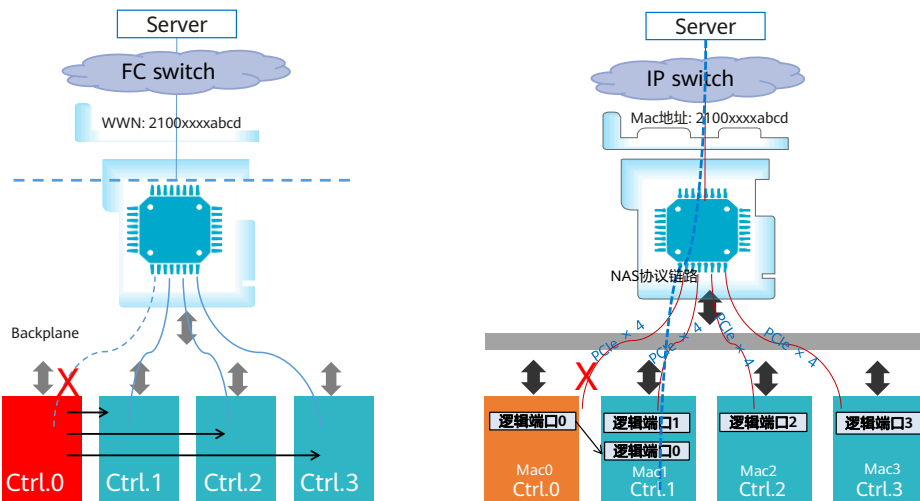


持续镜像



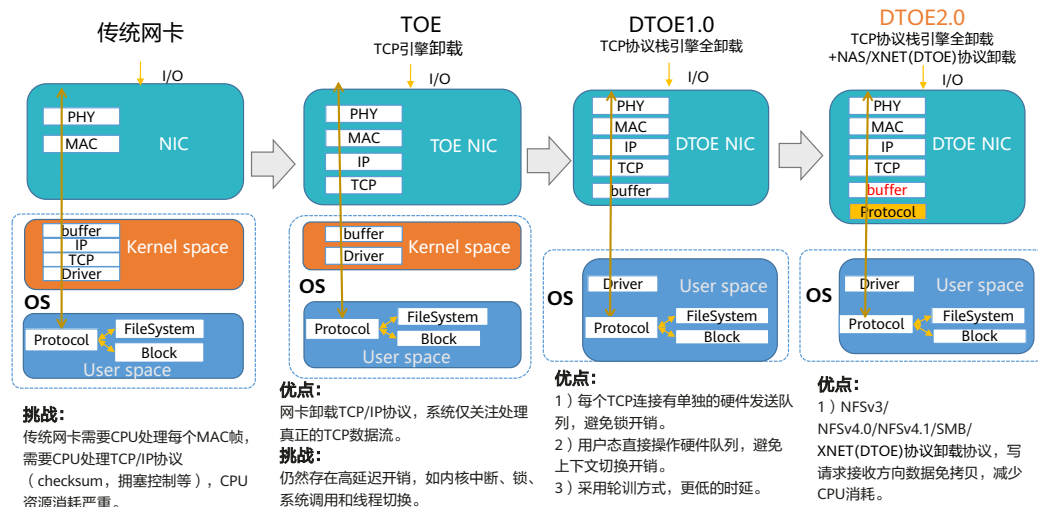
- OceanStor闪存存储高端系列提供了持续镜像功能。举例：A控故障，B控的缓存数据选择C控或者D控做缓存镜像；当控制器D再故障，选择B控和C控做缓存镜像，提供了4坏3的业务高可用能力。在配合双控制框，最大可支持8坏7的业务可用能力。这样的设计保证了即使单控制器故障，客户未及时更换，再次出现控制器故障，也不会导致数据丢失和业务中断，最大限度的保证了业务连续性。
- 当单控故障，其承载的业务快速切换到镜像控制器，剩余的控制器重新建立数据的镜像关系。业务请求可以保持回写，保障了系统性能，同时数据镜像冗余保障了系统的数据可靠性和业务连续性。
- 缓存数据的三副本功能，在两个控制器同时故障时能保证缓存数据不丢失。当阵列只存在单个控制框（4控）时，每个节点会选择另外两个控制器作为数据副本和接管对象。当双控同时故障，剩下的副本接管控制器会接管故障节点的业务，保证了控制框的高可用。
- 当阵列存在两个及以上控制框（单框4控）时，每个节点除在当前控制框选择一份数据副本和接管关系外，还会选择除当前控制框外的其它控制器作为第三副本接管对象，因此即使单控制框掉电或故障，剩下控制框的控制器也会接管故障控制器框的业务，保证业务连续性（内双活）。

接口模块/链路冗余保护



- OceanStor闪存存储是一个全冗余的存储系统。对接主机的前端，连接硬盘的后端，以及控制器之间的通讯，均有链路接口模块冗余保护。
- FC场景
 - 每个前端端口统一为主机提供1个FC会话，主机从每个存储端口只看到1个FC会话和WWN。
 - 4个内部链路连接到引擎内每个控制器，一个引擎内每个控制器都与主机建立本控的FC会话。
 - 一个引擎任一控制器故障，故障控制器的I/O被前端接口芯片重定向到其它控制器，主机无感知，FC链路一直保持up，业务正常运行，无任何告警/事件。
- IP场景
 - 主机与普通以太网端口一样，单物理端口单物理链路。
 - 4个内部链路，连接到引擎内每个控制器，每个控制器产生本地端口，独立Mac地址。
 - IP地址配置在控制器的逻辑端口，主机与本地逻辑端口建立IP连接。
 - 控制器故障对应逻辑端口Down，IP地址自动漂移到其它控制器的逻辑端口，免费ARP发送，更新IP和MAC映射表，NAS客户端重连，业务恢复。

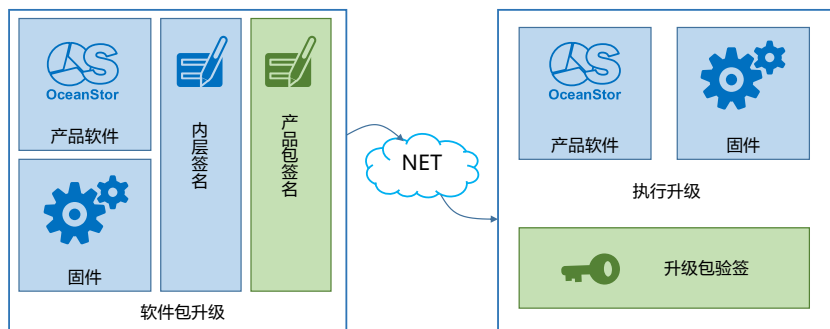
DTOE协议卸载技术



- 传统网卡，控制器处理的网络协议栈层次深，处理每个数据包都要触发一次中断，CPU资源开销严重。
- TOE技术，网卡对TCP/IP协议层进行卸载，每个应用程序完成一次完整的数据处理才触发一次中断，显著减轻中断开销；但是在这种优化情况下还有部分驱动等运行在内核态，存在用户态和内核态的系统调用、线程切换等开销带来额外的时延。
- DTOE技术，将iSCSI/NAS协议TCP数据路径卸载到网卡，传输层的处理包括DIF校验功能卸载到网卡微码里，彻底消除了CPU开销；同时，系统中的工作线程，采用的轮询的方式，周期性去检查接口模块的接收队列，有请求就立即处理。减少了请求来了以后，再去唤醒工作线程带来的时延开销。

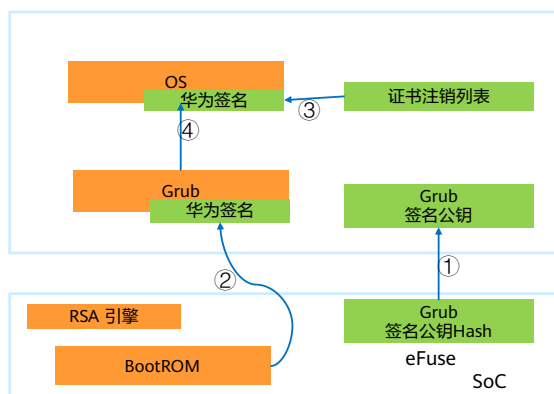
软件完整性保护

- 软件包采用内层文件数字签名和产品包数字签名，通过阵列上升级模块经过数字验签，可以保证升级包以及内部软件模块的完整性与唯一性。



- 产品软件包、升级包、固件包等公司研发到产品现场过程中，存在被篡改等风险。通过对产品软件包采用数据签名等技术，保护从产品研发，到客户现场使用的各类升级包的完整性，没有被篡改。软件包采用内层文件数字签名和产品包数字签名，通过网络发送到客户后，通过阵列上升级模块经过数字验签，验签通过后才开始升级，可以保证升级包以及内部软件模块的完整性与唯一性。

硬件可信安全启动



安全启动的软件验证、加载过程:

1. 验证Grub的签名公钥
2. 验证并加载Grub
3. 验证软件签名证书状态
4. 验证并加载OS

- 设备上电启动后，从初始启动模块开始启动，并逐级先验证，验证通过再启动。通过数字签名校验固件的完整性，防止固件和操作系统被篡改。
- 相关技术如下：
 - 可信根RoT(Root of Trust)，集成到CPU内部，可防止软件攻击，以及物理攻击，安全等级业界最高。
 - 通过两级数字签名（根密钥+二级密钥）保证软件完整性、通过数字证书保证软件唯一性。
 - 采用业界安全等级最高的算法之一：RSA 2048/4096。
 - 支持二级密钥（代码签名密钥）吊销。
 - 处理器内置RoT，可防止ARM片外Flash固件篡改、系统盘替换等恶意启动。

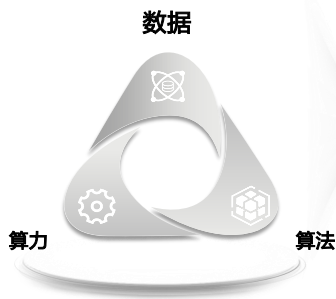
目录

1. 数据存储全景
2. 闪存存储产品介绍
 - 产品体系
 - 产品架构
 - 应用场景
3. 闪存存储产品部署
4. 闪存存储增值特性

AI智能体应用落地面临的挑战

AI赋能变革，重塑生产力

数据是AI核心要素



来源：华为MI, 2025

幻觉频发



大模型与现实场景存在脱节，若未结合私有数据，极易产生幻觉

推理卡顿



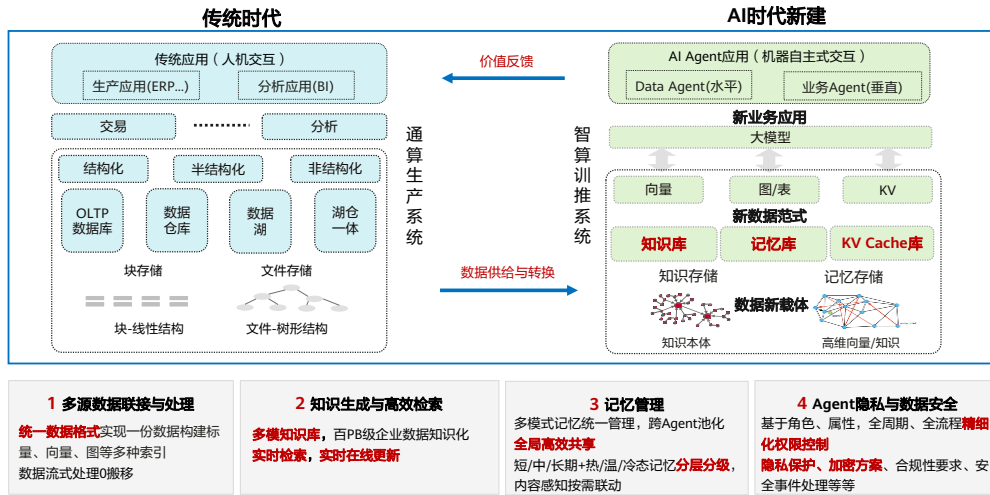
长序列推理与多轮对话场景下，推理吞吐不足、首token时延过长

智识浅薄



智能体缺乏上下文记忆，总结与反思能力欠缺，难以实现迭代进化

面向AI应用的数据平台



- 面向AI演进的数据平台需构建4大能力

华为AI数据平台



- 华为AI数据平台，整合知识库、KV Cache存储、记忆库，协同UCM（Unified Cache Manager）推理记忆数据管理技术，面向AI智能体提供知识生成与检索、KV Cache推理加速、记忆萃取与召回三大关键能力，助力智能体推理实现更精准、更高效、更智能，加速企业AI应用落地。
- 知识生成与检索，为智能体提供实时的、高精度多模态知识检索
 - 基于知识库，持续感知源数据变化，近实时完成原始数据与知识转换；通过多模无损解析、Token级编码等技术，将多模态数据转化为高精度知识，实现超95%的检索准确率。
- KV Cache推理加速，利用历史记忆数据提升智能体推理效率
 - 通过KV智能分级缓存管理，大幅减少推理过程中的重复计算，显著降低推理时延，提升推理吞吐与用户体验，为长序列、复杂逻辑的智能体推理提供关键性能支撑。
- 记忆萃取与召回，为智能体提供个性化、持续总结的记忆
 - 依托记忆库，沉淀AI智能体交互过程中的任务记忆与经验记忆，支持记忆回溯与多智能体协同学习，持续优化推理精度与效率，实现“越用越聪明”的迭代进化。

华为AI数据平台解决方案



- 随着AI技术的发展，AI推理已经在医疗、金融、制造等各行业智能体应用，然而在实际落地过程中仍面临诸多挑战与问题，往往会因为多模态知识生成难与检索不准导致的检索精度低的问题。并且随着上下文序列长度变长，出现推理时延剧增，或长序列无法一次完成推理等问题，大大降低推理体验。由于缺少个性化记忆，智能体无法真正像人一样积累经验，每次都是全新的对话，或者说着说着突然断片。这些都是应用落地中出现的 key 问题。
- 相对于传统IT系统，AI对存储提出的新的要求，需要处理的不仅仅是数据库、湖仓等结构化、非结构化数据，出现了向量、图/表、KV的等新的语义。需要将企业的数据转换为知识与记忆直接供AI使用。因此华为推理出AI数据平台方案，该方案基于 OceanStor A800一体化方案和AIDE（AI数据引擎，可由A800提供）与OceanStor Dorado的组合方案提供2库3服务的能力，构建多模态知识库与PB级共享记忆的记忆库。通过高效高精度的知识生成与检索，让数据实时转换为知识，实现多模态只是检索精度>90%，帮助智能体推理更精准；通过UCM推理加速实现对推理过程中产生的KV Cache持久化保存与动态管理，实现以查代算，首token时延降低90%+，推理吞吐提升2倍+，推理体验与效率大幅提升；构建PB级共享记忆，并且对记忆数据凝练保存，让智能体记得住历史工作内容，具备工作学习经验，越用越聪明，任务效率10倍提升。

华为AI数据平台部署模式

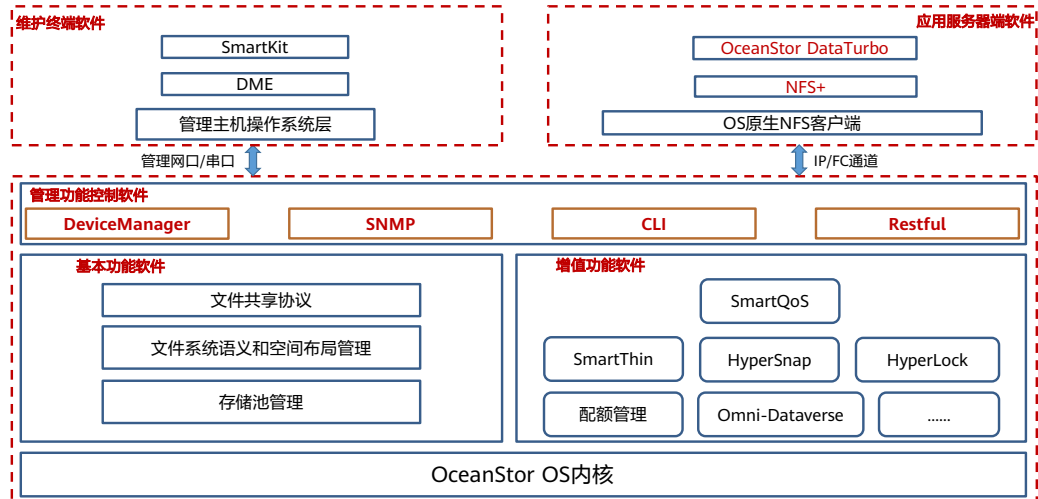
一体化部署



分离式部署

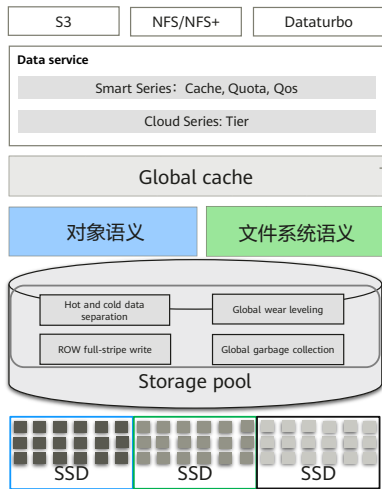


AI存储功能架构



- AI存储的系统功能架构与闪存存储不同，更聚焦AI应用场景。
- 安装部署在存储访问的应用主机上的存储客户端软件，该客户端软件是存储服务在主机侧的一个驱动，应用程序可以通过调用该软件提供的API接口、或调用该软件注册到服务器OS中的接口实现来访问存储的数据，实现高性能、高可靠的存储服务使能。OceanStor A800配套的存储软件包括DataTurbo、NFS+和OS自带的NFS驱动软件。
- 华为基于开源NFS软件基础上做了功能增强，主要是在访问I/O栈的优化、客户端和服务端多路径链接的可靠性、多路径并发上优化，实现高可靠高性能的访问。其中NFS+多路径的通用能力已经纳入openEuler的开源社区于2023年发布上市，存储配合定制优化部分则配套存储产品在华为support网站发布使用。
- 华为面向不同的应用系统提供适配应用业务特征的性能、可靠性和定制功能的软件组件。OceanStor DataTurbo DTFS（DataTurbo FileSystem）是访问文件存储的客户端组件，基于华为技术有限公司开发的高性能的DataTurbo文件系统共享协议与存储系统进行数据通信。它兼容标准POSIX语义和MPI-IO语义，具备高性能、高可靠性、强一致性、MPI-IO等能力。通过在客户端上部署OceanStor DataTurbo DTFS软件并在存储系统上配置DataTurbo共享，可以提升单流和单客户端性能，降低访问时延。

AI存储系统架构



协议层

- 实现跨协议共享能力：统一命名空间支持标准S3、NFS协议、增强NFS+协议、私有DataTurbo共享访问，用户管理，用户鉴权，客户端认证，ACL权限控制等
- 协议分层卸载加速：RDMA、TCP/IP，NFS、S3

数据服务层

- 基于Smart/Cloud系列增值特性，包括Qos、Quota、Tier等在此层实现。

全局Cache层

- 共享使用的全局Cache，数据可以在各控制器的Cache均衡访问
- Cache分级策略：按元数据，数据分层分级设置Cache缓存策略

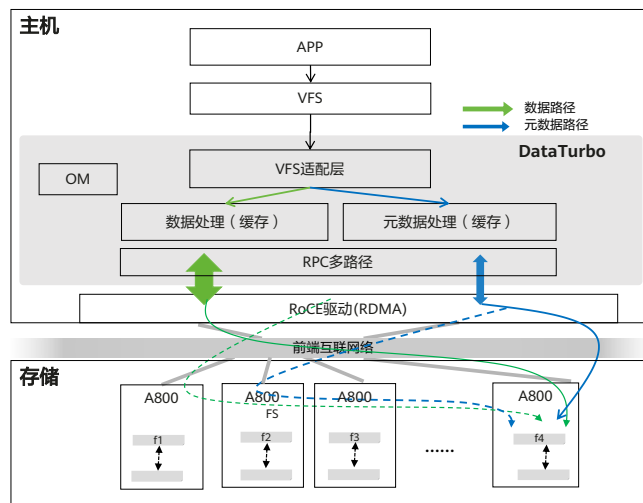
空间管理层

- 实现文件系统的目录/文件打散，均衡到各个控制器的CPU去处理
- 实现桶/对象的打散、均衡到各控制器的CPU去处理

存储池层

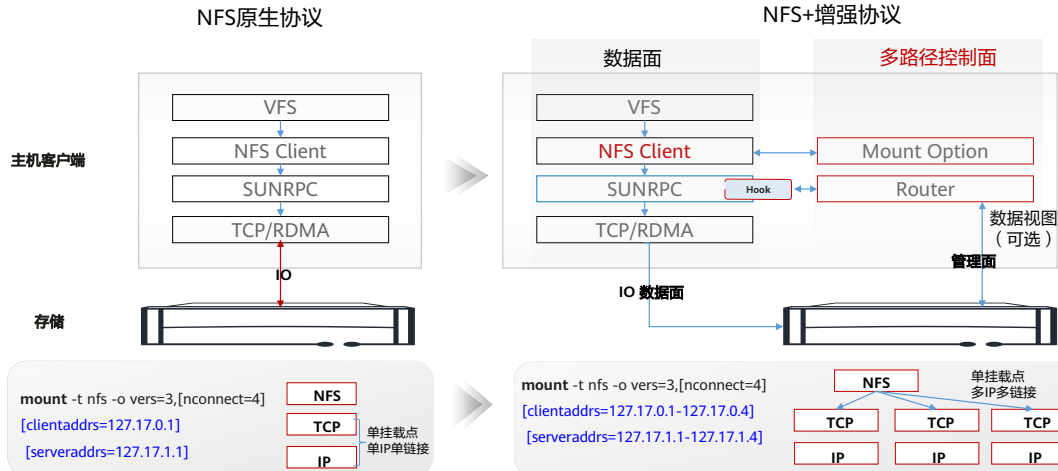
- 存储池对硬盘的管理，RAID2.0+空间分配管理，热备空间管理，数据重构，提供基本的元数据和数据空间分配机制
- 盘框配合的FlashLink系列特性：如全局垃圾回收，全局磨损均衡和反磨损均衡，元数据/数据分区，IO优先级控制
- 存储池I/O处理和空间管理：ROW/满条写等全局特性

DataTurbo高性能客户端



- DataTurbo协议客户端是华为OceanFS定制的私有文件访问客户端，安装部署在主机侧，对接OS的VFS层，提供标准的POSIX语义，支持MPI-IO等访问协议对接能力，应用程序不需要做任何修改就可以直接使用并挂载文件系统，安全易用。DataTurbo本身提供应用加速能力，高性能缓存提升客户端访问性能；链路管理层提供多路径管理能力，实现高可靠和高性能的存储链接访问能力。
- DataTurbo客户端关键技术能力：
 - 存储连接层：支持主机多网卡多IP和存储集群的多服务IP的多对的链接能力，实现高可靠高性能的多路径能力。单客户端可同时连接多个存储节点，提升链路可靠性、单流和单客户端访问并发能力度，提升总带宽。
 - 存储连接层：支持RDMA协议，可以实现更低的时延和更低的CPU开销。
 - 存储连接层：解决标准协议只能在Mount时进行负载均衡的问题，实现IO级负载均衡。
 - 应用加速层：支持客户端高性能缓存技术，降低访问I/O栈长度，降低端到端的访问延时，支持随机小IO聚合为大IO提升系统带宽。
 - 应用适配层：支持VFS和fuse的对接，标准的POSIX语义，支持MPI-IO对接，应用无需修改直接使用。

NFS+高性能客户端



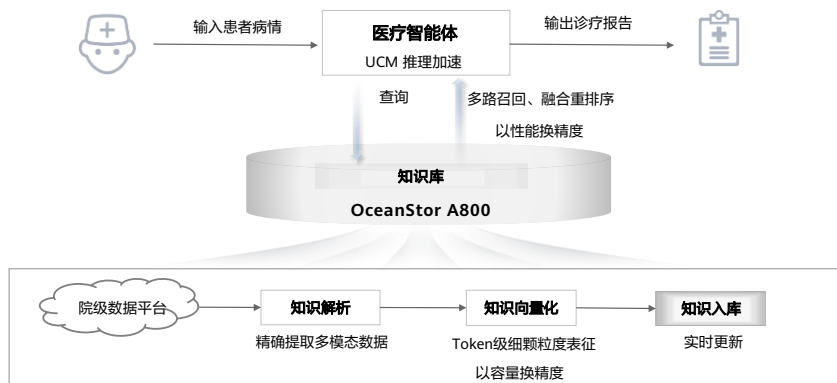
- 基础NFS协议中，单客户端只能基于单服务IP进行挂载访问，受限于单IP的性能和可靠性的局限性，华为NFS+扩展标准NFS协议的实现能力，单客户端可用多服务IP挂载同一个文件系统，客户端的多IP和存储服务端的多IP之间建立多个IP链接并提供多链路多路径管理技术能力，提升可靠性和访问并发性能。
 - NFS+链路可靠性：当单IP因为网络、网卡或驱动等因素出现故障，系统会根据链接状态检测和I/O的超时时间识别异常，并自动切换到正常链接进行请求直接重试处理；无需等待IP漂移后重新更新链接信息后重试。大幅降低的链路故障的处理效率，从s级别缩减到ms级。
 - NFS+链路性能：单IP链路下，请求只能在该链路进行发送和排队处理，对带宽型访问严重受限该IP下的端口速率、传输并发和延时，导致带宽受限；NFS+的多路径有多IP对链接，可以并发处理数据报文，能够充分利用网络端口带宽能力实现单客户端总带宽能力倍数级提升。
 - NFS+主机生态兼容性：继承NFS的主机应用生态能力和协议功能能力，方案组网设计和业务对接访问接口不变，应用无需适配修改，简单易用。
 - NFS+存储性能增强：多路径选路策略上支持和存储系统的协商，能够感知存储数据布局进行链路监控状态管理和最佳路径信息维护，实现I/O分发均衡选路，端到端的极致时延和带宽。

AI数据平台一体化部署场景

某Top医院深度融合临床医学专家实践经验、医学指南以及医疗文献等数据，提供权威可信的辅助诊疗智能体系统

业务痛点：医学文献解析精度差、知识召回精度<70%（召回低、排序不准确）

采用华为AIDP知识库辅助诊疗业务流



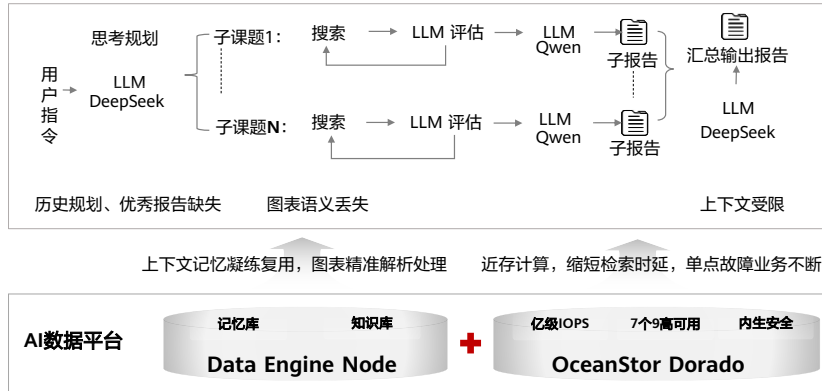
- 某医院使用辅助诊疗，知识检索更精度、更安全。

AI数据平台分离式部署场景

DeepResearch企业信贷分析报告场景，在信贷审批与贷后管理场景，生成动态风险报告，识别潜在经营与信用风险。

业务痛点：① 研报质量不足 ② 研报生成效率低 ③ 推理业务中断恢复慢

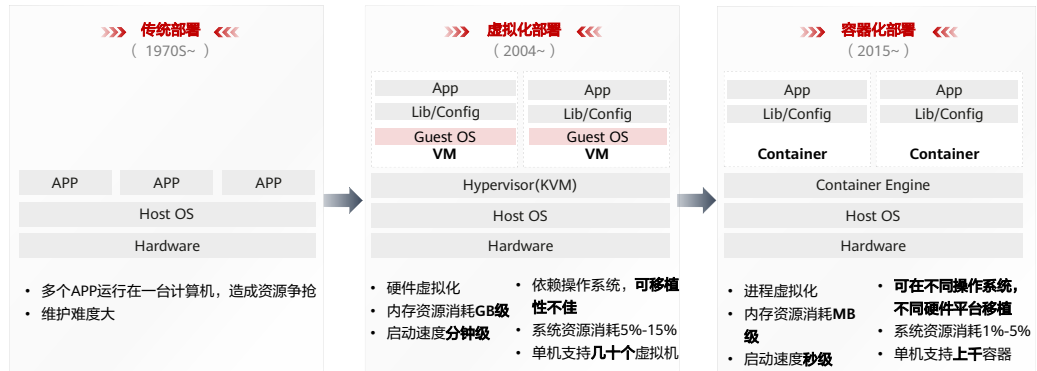
采用华为AIDP方案提升报告质量



- 某金融客户DeepResearch场景应用，推理效果提升。

应用的容器化部署趋势

容器是轻量的虚拟化技术



为了应对业务创新的加速和解决互联网规模的挑战, 灵活、快速迭代的云原生应用架构应运而生。容器的敏捷、高效、弹性等特点实现了软件交付流程的标准化, 一次构建, 随处部署, 是云原生应用的基石。

- 企业业务要想真正的云化, 不仅要在基础设施和平台层面实现, 而且应用本身也应该基于云的特点进行开发, 从架构设计、开发方式、部署维护等各个阶段和方面重新设计, 构建真正应“云”而生的“云原生应用”。

存储成为容器化进程的主要挑战之一

容器应用对存储的主要需求

有状态应用

注1

数据库
消息中间件
大数据



存储诉求

高性能
数据高可用
数据安全
易运维

无状态应用

注2

搜索引擎
WEB
CI/CD



存储诉求

数据共享访问
数据高可用
易运维

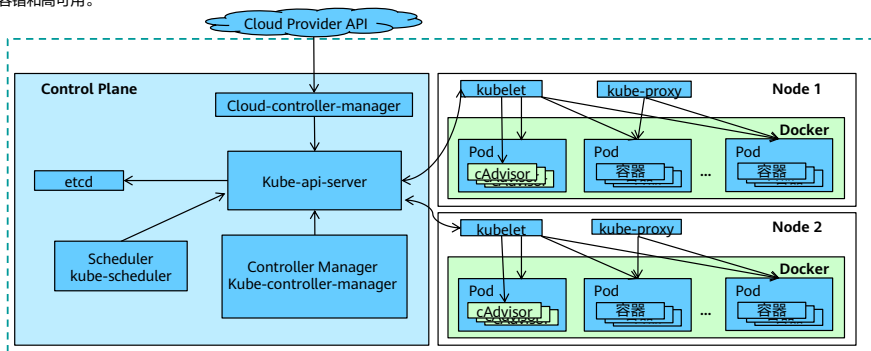
注1：有状态应用指需要持久化保存中间数据或会话数据的应用

注2：无状态应用指不需要保存任何中间会话数据的应用，相当于一个外界访问数据的通道

Kubernetes简介

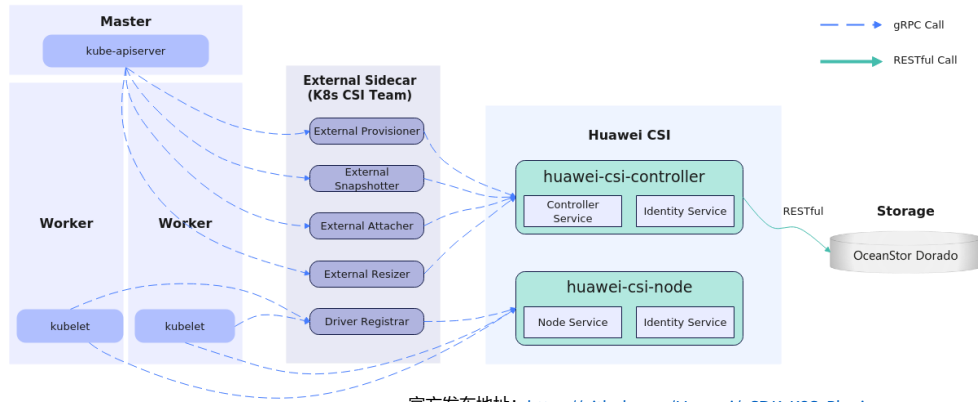
- 真正的生产环境中会涉及部署在多个服务器上的大量容器。如何跨集群的构建、管理、调度、持续监控这些容器？因此有了Kubernetes。
- Kubernetes 集群由一个控制平面和一组用于运行容器化应用的工作机器组成，这些工作机器称作节点（Node）。每个集群至少需要一个工作节点来运行 Pod。
- 工作节点托管着组成应用负载的 Pod。控制平面管理集群中的工作节点和Pod。在生产环境中，控制平面通常跨多台计算机运行，而一个集群通常运行多个节点，以提供容错和高可用。

- ✓ 资源调度
- ✓ 部署运行
- ✓ 服务发现
- ✓ 扩容缩容
- ✓ 健康监控
-



华为CSI简介

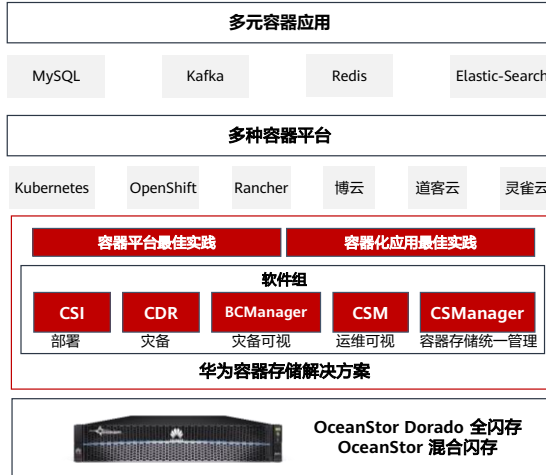
容器存储接口（Container Storage Interface），简称 CSI，是一种行业标准协议，用于对接Kubernetes等容器平台与底层存储系统。华为CSI插件作为必备组件，实现华为企业存储和分布式存储产品与Kubernetes集群的对接，为容器工作负载提供持久化存储服务。



官方发布地址: https://github.com/Huawei/eSDK_K8S_Plugin

- Kubernetes通过其官方维护的一系列sidecar组件负责注册监听Kubernetes对象资源，并调用CSI驱动程序，触发华为存储的物理操作。如创建一个持久卷（Persistent Volume，PV）时，华为CSI驱动会在存储设备中创建对应的LUN（块存储）或文件系统。

华为容器存储解决方案



华为提供

55 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



敏捷易用

华为CSI支持按需、快速发放存储资源，支持多租户、QoS、卷纳管、卷双活等丰富的企业级特性。

业务可靠

华为CDR为容器有状态应用提供主机级、站点级和集群级可靠性保护，及跨阵列一体化备份能力。

运维监控

华为CSM支持对接Prometheus和DM界面，帮助K8s集群管理员、存储管理员实现存储资源的可视化监控与管理。

极简管理

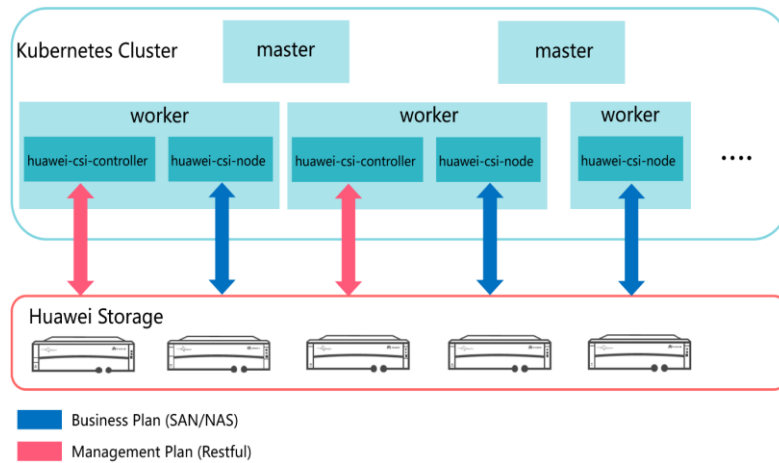
华为CSManager作为容器场景可视化统一管理软件，支持可视化集群管理、存储资源管理、容灾管理等。

数据安全

存储多租户特性，实现管理面、网络面、业务面三层隔离。

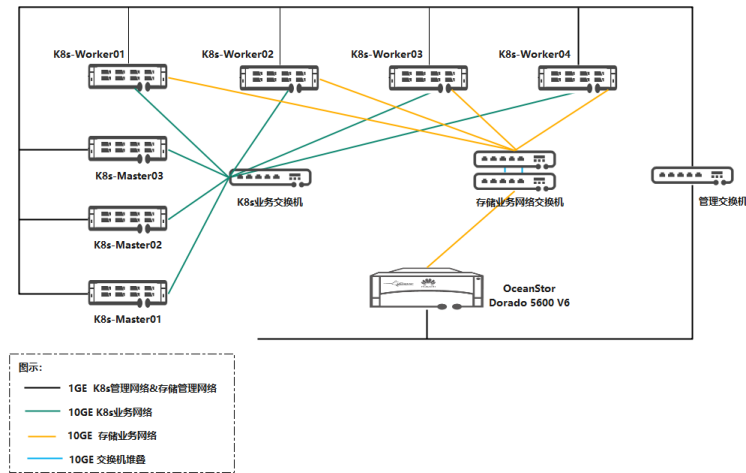
- 华为容器存储解决方案提供面向容器场景的多个软件，针对MySQL、Kafka、Redis和Elasticsearch等分布式中间件，一方面通过提供华为CSI（Container Storage Interface）软件对接Kubernetes，使客户可以在Kubernetes及以此为基础的容器平台上，便捷地利用华为存储向Kubernetes提供持久化的高性能、高可靠性、按需配置的SAN和NAS存储。再通过华为CSM(Container Storage Monitor)插件，对Kubernetes容器场景下华为存储的资源和Kubernetes中关联的资源进行可视化展示，提升运维体验。
- 另一方面，华为提供华为CDR（Container Disaster Recovery）容器容灾恢复软件，结合华为存储对容器资源进行备份、容灾及恢复。CDR将华为存储设备的卷管理特性和备份容灾特性结合，从而提升容器场景下业务应用的数据安全性，业务应用可靠性。同时结合OceanStor BCManager，实现容器容灾可视化管理，清晰可视地掌控容灾业务全局。

华为CSI部署模型



- 华为CSI主要有两大组件，分别为huawei-csi-controller和huawei-csi-node：
 - huawei-csi-controller：包含Controller Service和Identity Service，以Deployment方式运行的一个或多个Pod，主要负责与华为存储交互，使用RESTful方式进行通信，因此运行huawei-csi-controller组件的节点需要连通存储的管理面网络。
 - huawei-csi-node：包含Node Service和Identity Service，以DaemonSet方式运行在Kubernetes工作节点上的Pod，用于在工作节点上对华为存储提供的LUN/文件系统资源进行挂载和卸载等操作，因此运行huawei-csi-node组件的节点需要连通存储的业务面网络。

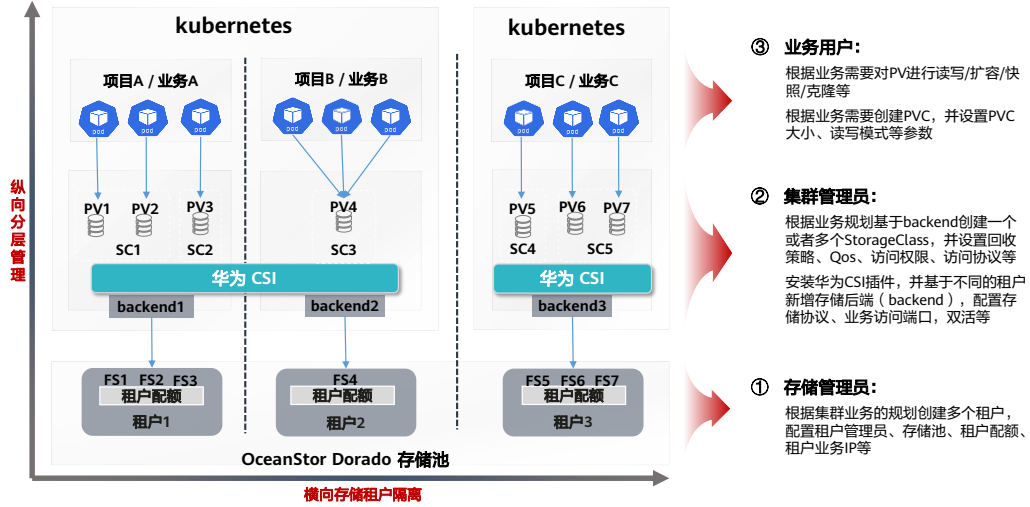
华为CSI部署组网要求



• 存储对接容器平台建议：

- 一套存储对接一个K8S集群使用系统租户；对接多个K8S集群时可使用系统租户，若希望存储管理员对各个集群做资源隔离则使用自建租户，一个租户对接一个K8S集群。
- 系统租户对接容器平台，和存储相关的网络共两张：1GE的存储管理网络（黑色）；10GE的存储业务网络（黄色）。
- 自建租户对接容器平台，建议将租户管理网、业务网合并，统一使用黄色网络。

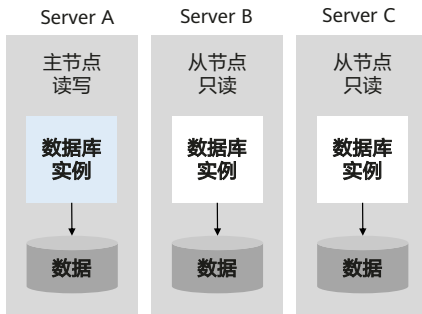
CSI最佳实践



新生态数据库存在两种部署模式

集中式部署

数据不做切片，代表：GaussDB

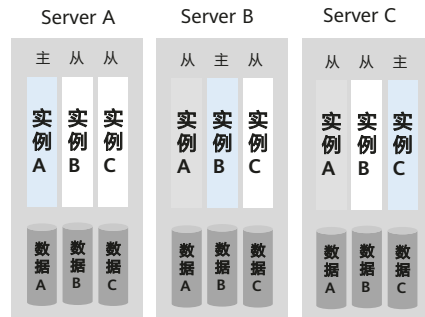


优点：数据一致性易保障，效率高

缺点：大库性能差、扩展性差

分布式部署

数据切片后分散存放，代表：OceanBase

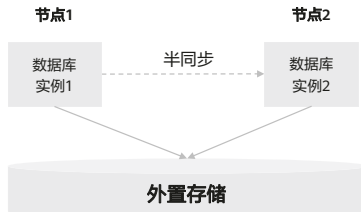


优点：并发访问效率高，易扩展

缺点：难以保障数据一致性

存算分离可以全面提升数据库的各项能力

存算分离架构数据库



整体方案优势

华为方案独家优势

可靠性提升

- 专业外置存储的可靠性大幅提升
- 节点故障数据仍可访问
- 专业存储保障容灾效率和可靠性
- 数据共享保障强一致性
- 多节点并发写一份数据，故障无缝切换

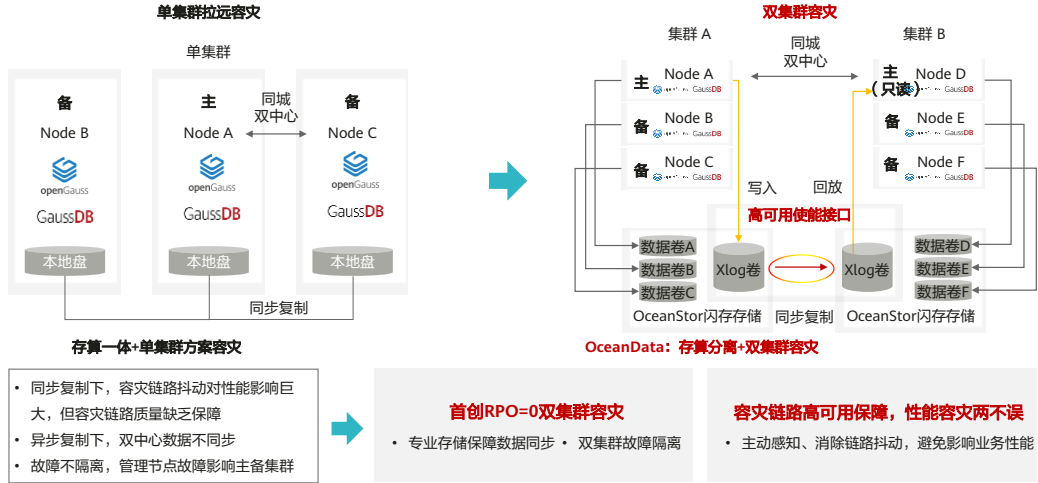
性能提升

- 减少副本间同步，提升性能
- 容灾效率提升，性能影响小
- 减少分片带来的服务器开销
- 多节点并发写，大幅提升性能
- 免数据同步/分布式事务，性能提升

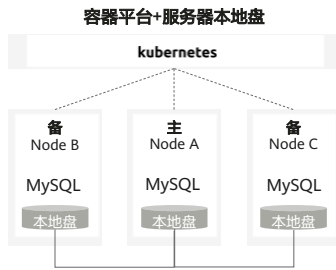
TCO降低

- 可靠性提升，可大幅减少副本数量
- 硬盘健康度统一智能管理
- 存算分别按需扩展
- 数据共享，全局可一份副本
- 无需分库分表，降低改造开销

分布式新核心改造解决方案

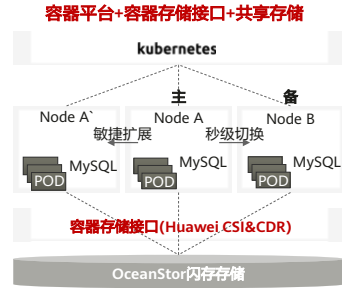


数据库微服务改造解决方案



容器平台+服务器本地盘

- 节点故障，需要人工介入恢复
- 补从需要数小时，影响性能和可靠性
- 一主多从导致海量服务器



OceanData方案：存算分离+华为CSI&CDR+容器平台

节点故障自动恢复

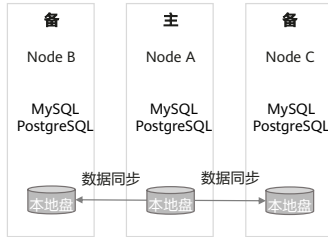
- CSI自动驱逐故障节点、恢复Pod并挂接PV

从节点5分钟完成构建

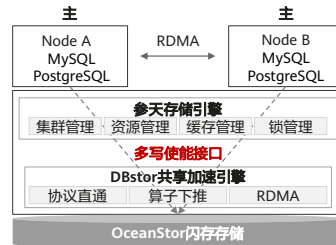
- 节点故障数据自动挂载到新节点，补从只需构建增量数据

传统核心免分库分表改造解决方案

存算一体 + 一写多读



存算分离 + 多写多读



存算一体 + 一写多读

- 一写多读，有性能瓶颈，且读不强一致
- 主节点故障，主从切换会断业务
- 通过分库分表弥补单库性能与可靠性问题



OceanData方案：存算分离 + 多读多写 + 免分库分表

多写多读

- 所有节点并发读写，性能比原生MySQL提升10倍，同节点数性能达到Oracle 90%

节点故障业务0中断

- 单节点故障不影响业务访问，单事务30s内完成恢复

目录

1. 数据存储全景
2. 闪存存储产品介绍
- 3. 闪存存储产品部署**
4. 闪存存储增值特性

闪存存储安装流程

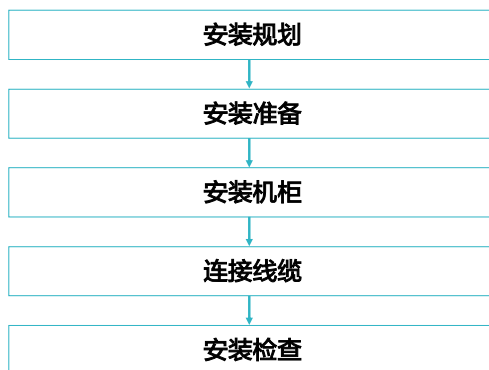


设备安装流程

非整柜交付场景



整柜交付场景



规划工具介绍

工具名称	简介
能耗计算	提供存储产品能耗快速计算功能，用于规划阶段评估机房供电情况。 访问地址： https://info.support.huawei.com/storage/powercalc/#/home
兼容性查询	根据机房现场环境在线快速查询存储产品与业界主流厂商、主流设备的兼容性关系。 访问地址： https://info.support.huawei.com/storageinfo/comp/#/oceanstor
LLDesigner	提供硬件配置、设备组网、资源分配等功能，能快速生成LLD文档和存储设备的配置文件。主要功能如下所示： •支持BOQ一键式导入，自动识别存储设备，控制器，接口卡、硬盘框、硬盘类型以及数量，快速完成硬件信息规划。 •支持通过模板创建LLD，覆盖多控等典型组网场景。 •支持自动导出离线配置文件，通过DeviceManager导入到存储设备，完成设备基础业务配置。

机柜空间规划

设备	0.8米机柜	1.0米机柜	1.1米机柜	1.2米机柜
4U独立机头高端控制器（865mm）	/	/	需U型固定支架	需U型固定支架
2U 2.5寸中端盘控（820mm）	/	需U型固定支架	无需U型固定支架	无需U型固定支架
2U PALM中端盘控（920mm）	/	/	需U型固定支架	需U型固定支架
2U 2.5寸低端盘控（520mm）	无需U型固定支架	无需U型固定支架	无需U型固定支架	无需U型固定支架
2U 智能NVMe硬盘框（620mm）	需U型固定支架	无需U型固定支架	无需U型固定支架	无需U型固定支架
2U 智能SAS硬盘框（520mm）	无需U型固定支架	无需U型固定支架	无需U型固定支架	无需U型固定支架
2U 2.5寸SAS硬盘框（410mm）	无需U型固定支架	无需U型固定支架	无需U型固定支架	无需U型固定支架

- 以上要求以华为存储机柜为例，其他第三方机柜由于机柜立柱位置不同，对U型固定支架的需求也不同。因此会随控制框发货3种类型U型固定支架（25mm、50mm、75mm），以满足不同机柜的要求。

2U控制框线缆连接

低端2U控制框



中端2U控制框



原则1： 后端SAS接口模块配置在2号接口卡槽位；
原则2： **Scale out**接口模块配置在1号接口卡槽位。

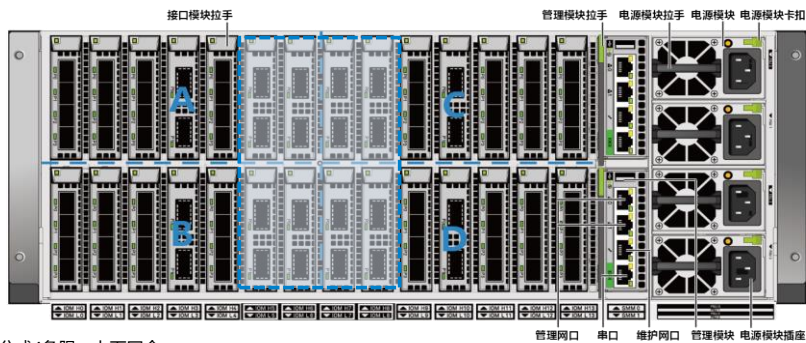
建议1： 前端接口模块安装先后顺序：0、1、2；
建议2： 后端RDMA接口模块配置先后顺序：2、1、0。

原则1： 后端SAS接口模块配置在3至5号接口卡槽位；
原则2： **Scale out**接口模块配置在2号接口卡槽位。

建议1： 前端接口模块配置先后顺序：0、1、2、3、4、5；
建议2： 后端RDMA接口模块配置先后顺序：5、4、3、2、1、0；
建议3： 后端SAS接口模块配置先后顺序：5、4、3。

- 此部分内容举例说明线缆连接方式，产品安装部署过程中应以产品文档内容为准。

4U控制框线缆连接



原则1: 每控分成4象限，上下冗余；

原则2: 同一功能，配两张卡时，插入象限A和B；配四张卡时，新增的2张卡插入象限C和D；

原则3: SAS卡只能配置在H5~H8，L5~L8，最大支持8张SAS共享卡；

原则4: 100G RDMA后端卡由H6/L6向两边插，前端卡由两边向中间插；

原则5: Scaleout卡配置在H3，L3，H10，L10；

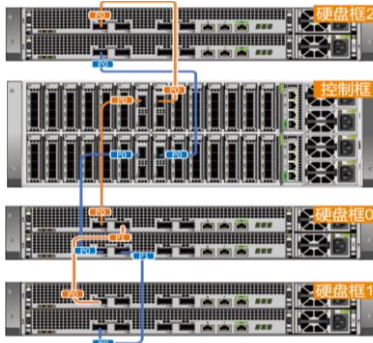
- 此部分内容举例说明线缆连接方式，产品安装部署过程中应以产品文档内容为准。

硬盘框组网原则

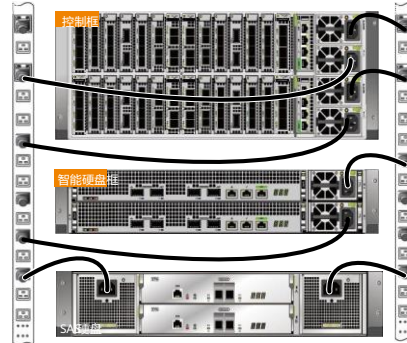
- 硬盘框级联组网采用全正接。
 - 2U控制框：控制器A上的级联接口模块连接到硬盘框的级联模块A，控制器B上的级联接口模块连接到硬盘框的级联模块B。
 - 4U控制框：框上部的级联接口模块连接到硬盘框的级联模块A，框下部的级联接口模块连接到硬盘框的级联模块B。
- 控制框上连接同一个硬盘框的两个端口所属接口模块要对称，端口号要一致。
- 控制框的级联接口模块端口或者板载级联端口只能连接到硬盘框的P0端口。
- 硬盘框的级联模块P1端口连接到下一级硬盘框级联模块的P0端口。
- 取消保险箱盘（配置框）必须是第1框前4盘的约束，对于硬盘框的安装顺序不再有约束。

- 硬盘框号生成规则
 - CTE：代表控制框。CTE0代表0号控制框；CTE1代表1号控制框。
 - DAE：代表硬盘框。DAEXXX表示硬盘框编号，第一个X为所属控制框号，第二个X为级联环路号，第三个X为该级联域中的硬盘框号。比如DAE010表示控制框0的1号级联环路中0号框。

连接线缆



控制框和硬盘框线缆连接



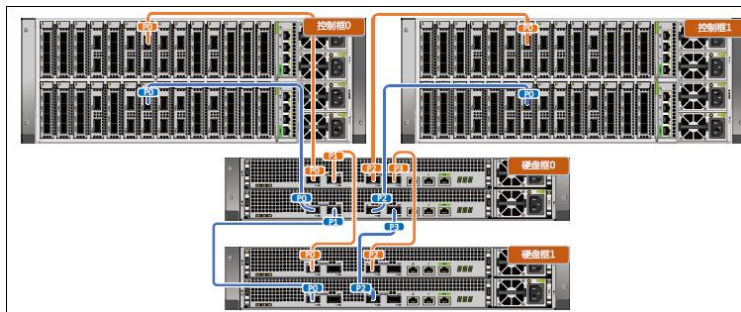
控制框和硬盘框连接电源线

说明：以华为OceanStor Dorado全闪存存储为例。

- 硬盘框P0端口是上行口，连接来自控制框或上一级硬盘框的P1端口。
- 硬盘框P1端口是下行口，只能连接下一级硬盘框的P0。

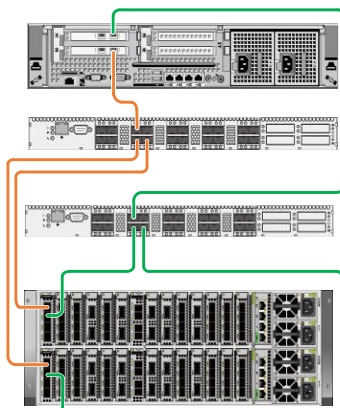
内双活组网

- **控制框0和硬盘框内双活组网：**硬盘框P0端口是上行口，连接来自控制框0或上一级硬盘框的P1端口。硬盘框P1端口是下行口，只能连接下一级硬盘框的P0。
- **控制框1和硬盘框内双活组网：**硬盘框P2端口是上行口，连接来自控制框1或上一级硬盘框的P3端口。硬盘框P3端口是下行口，只能连接下一级硬盘框的P2。

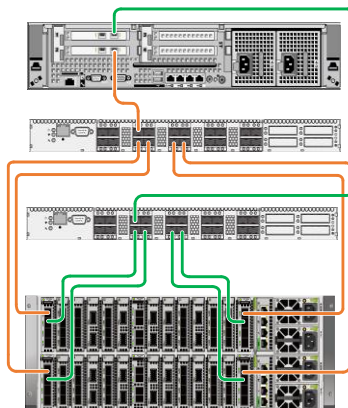


连接控制框至应用服务器

4U2控：FC/iSCSI组网



4U4控：FC/iSCSI组网



- 此部分内容举例说明线缆连接方式，产品安装部署过程中应以产品文档内容为准。

安装与连线检查

【安装检查】

检查项	正常	异常
控制框	- 平稳地安装在滑道上，无位移。 - 螺钉已拧紧。	- 设备倾斜或不在位。 - 螺钉松动或不在位。
硬盘框		

【连线检查】

检查项	正常	异常
光纤、mini SAS HD 线、RDMA线缆、网线、串口线等	线缆正确固定并且完全插入。	线缆松动或脱落。
交流电源线	- 交流电源线冗余连接至不同的独立电源。 - 交流电源线卡扣扣好。	- 交流电源线连接到同一个独立电源。 - 交流电源线卡扣未扣好。

设备上电顺序



初始化存储系统流程



修改管理网口IP地址

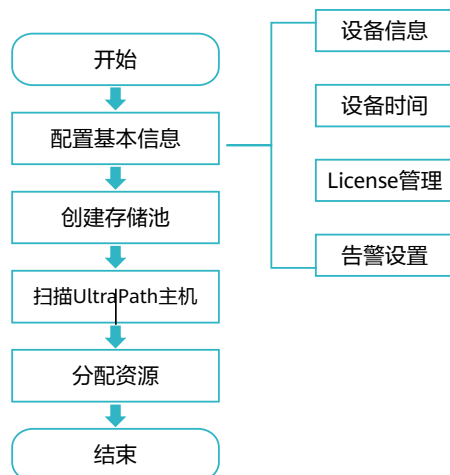
- 使用CLI修改管理网口IP地址：
 - 管理网口默认IP登录前提条件
网线连接维护终端和存储设备管理网口，维护终端的IP地址与存储管理网口的默认IP地址在同一网段。
 - 串口登录前提条件
串口线缆连接维护终端和存储设备串口，终端程序（例如PuTTY软件）的波特率需设置为115200。
- 登录后运行change system management_ip命令修改管理网口IP地址。如：设置管理模块0的管理网口IPv4地址为“172.16.190.2”，子网掩码为“255.255.255.0”，网关地址为“172.16.0.1”。

```
admin:/>change system management_ip eth_port_id=CTE0.A.MGMT ip_type=ipv4_address ipv4_address= 172.16.190.2 mask=255.255.255.0 gateway_ip=172.16.0.1
WARNING: You are about to change the IP address of management network port. If you enter an unavailable IP address, the DeviceManager will become inaccessible. This operation will clear the configured routes on the IP address that is modified on this management port.
Suggestion: Before performing this operation, ensure that the entered IP address is available.
Have you read warning message carefully?(y/n)y

Are you sure you really want to perform the operation?(y/n)y
Command executed successfully.
```

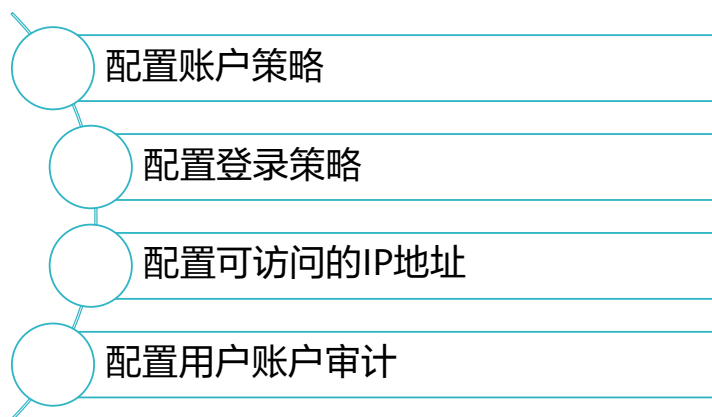
- 注意
 - 管理网口IP地址的默认值为192.168.128.101/192.168.128.102，子网掩码的默认值为255.255.0.0。
 - 管理网口IP地址与维护网口IP地址不能在同一个网段，否则会造成路由冲突。默认的维护网口IP地址为172.31.128.101或172.31.128.102，子网掩码为255.255.0.0，所以管理网口不能设置为172.31.XXX.XXX网段的IP地址。
 - 首次登陆，需要初始化默认密码。
 - 也可以通过设备默认IP登录DeviceManager，通过DM进行IP的修改。
- 此部分内容举例说明存储默认IP修改，产品安装部署过程中应以产品文档内容为准。

初始化配置向导



说明：以华为闪存存储产品为例，不同产品版本和型号的操作界面可能会有差异，操作时请以实际情况为准。

配置安全策略



安全配置管理

用户/角色管理

- **超级管理员**：对存储设备有完全的控制权限，可以创建各级别的用户。
- **管理员**：对存储设备具有一定的控制权限，但不能进行用户管理、升级系统、修改系统时间、激活License、设备下电、重启设备。
- **只读用户**：对存储设备只有访问权限。登录存储设备后，只读用户只能进行查询操作。
- **USM用户**：当使用SNMPv3协议进行数据传输时，需要设置USM用户，为SNMPv3协议进行鉴权和数据加密。

证书管理

- 在存储系统和外部设备进行通信时，建议采用证书校验的方式来提高业务的安全性。

配置告警和事件处理策略

告警与通知

邮件通知

短信通知

告警转储

Trap通知

Syslog通知



说明：具体配置方式可参照存储产品文档。

更多工具获取



教授，存储安装步骤比较复杂，有没有安装指导资料呢？

我们教材里只列举了部分主要的安装步骤，更多的资料可以登录华为数据存储信息中心（<https://info.support.huawei.com/storage/#/home>），在“数存学院”中可以查找存储安装的视频；也可以选择OceanStor Dorado信息中心查找产品文档。



规划网络

规划以太网端口

- 以太网端口是指在设备上物理可见的端口。以太网端口是创建VLAN、绑定端口和逻辑端口的基础。

规划绑定端口

- 绑定端口可以增加链路的带宽和冗余。端口绑定后，每个会话连接依然通过单个端口进行数据传输，因此只有在有多个会话连接的情况下才可能提高总的带宽。请根据实际环境情况确定是否需要绑定端口。

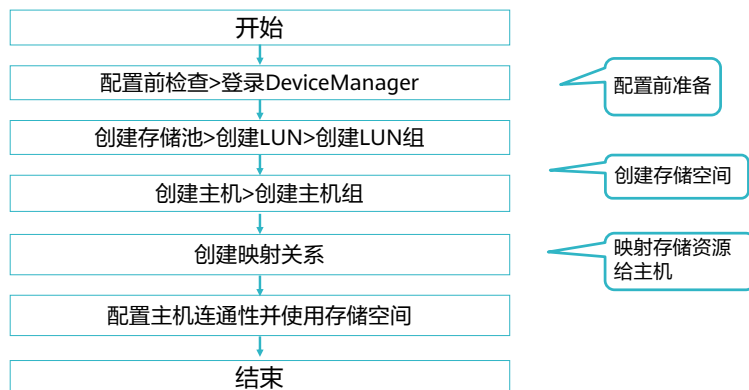
规划VLAN

- 存储系统上的以太网端口资源在逻辑上划分成多个广播域，在VLAN内发送或接收业务数据时，给数据加上VLAN ID，实现VLAN之间的网络隔离和业务隔离，从而确保业务数据的安全性和可靠性。
- VLAN的创建基于以太网端口或绑定端口，一个物理端口可以从属于多个VLAN。已经创建为绑定端口的以太网端口不能单独用来创建VLAN，只能通过其组成的绑定端口创建。

规划逻辑端口

- 逻辑端口的创建基于以太网端口、绑定端口或VLAN，用于业务的运行。

块服务业务配置流程



说明：以华为闪存存储产品为例，不同产品版本和型号的操作界面可能会有差异，操作时请以实际情况为准。

- 此处所示的配置流程为较常用的配置流程。DeviceManager采用全新的界面交互设计，资源分配操作极简，支持多种灵活的方式配置映射关系，用户可以结合实际业务需要选择合适的方式进行配置。

文件系统业务配置流程



说明：以华为闪存存储产品为例，不同产品版本和型号的操作界面可能会有差异，操作时请以实际情况为准。

- 存储系统支持应用服务器通过NFS（Network File System）、CIFS（Common Internet File System）协议访问共享的文件。
- 存储系统支持GUI（Graphical User Interface）和CLI（Command-Line Interface）两种方式管理文件访问协议。

目录

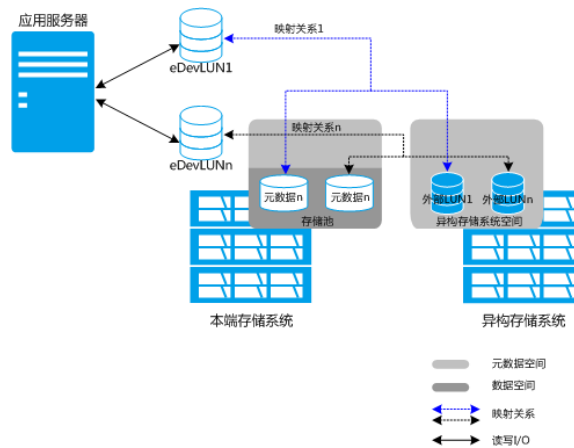
1. 数据存储全景
2. 闪存存储产品介绍
3. 闪存存储产品部署
- 4. 闪存存储增值特性**
 - **效率提升特性**
 - 数据保护特性

SmartVirtualization概述

- 随着用户数据量的增长，管理和扩容现有存储系统的重要性也逐渐凸显，而在对存储系统的管理和扩容过程中存在如下问题：
 - 如果用户使用新购存储系统替换现有存储系统，需要将现有存储系统的业务数据迁移至新购存储系统中，但由于不同厂商之间的存储系统兼容性存在差异，数据迁移时间长，严重时会导致数据迁移失败。
 - 如果用户新购存储系统，且对多台存储系统采用分散管理方式时，随着存储系统数量增加，投入的维护成本也随之增加，且原有存储系统和新购存储系统之间的存储资源也无法进行有效整合和管理。
- 使用异构虚拟化技术可以有效地规避上述风险，由于屏蔽了不同存储系统的物理特性，可以简化异构存储系统间的配置和管理，有效利用存储资源。
- 华为技术有限公司开发的SmartVirtualization特性又叫异构虚拟化特性，是指本端存储系统（OceanStor存储系统）与异构存储系统（包括其他华为存储系统和第三方厂商的存储系统）相互连接后，本端存储系统能够将异构存储系统提供的存储资源当作本地存储资源进行使用并对其进行集中管理，无需关注存储系统间软件架构和硬件架构的差异。
- 由于解决了不同存储系统之间的兼容性问题，用户可以集中管理OceanStor存储系统及异构存储系统中的存储资源，同时用户仍然可以使用旧存储系统的存储资源，降低对原有投资的浪费。

eDevLUN与外部LUN的关系

- eDevLUN由实际数据和元数据两部分组成，通过映射关系建立元数据和实际数据的对应关系。

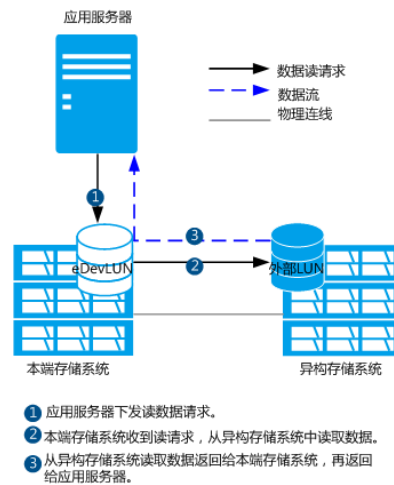


- 本端存储系统：即指OceanStor系列闪存存储系统。
- 异构存储系统：不仅包括其他主流厂商的存储系统，还包括某些型号的华为存储系统。
- 外部LUN：位于异构存储系统中的LUN。在DeviceManager上显示为“远端LUN”。
- eDevLUN：在本端存储系统的存储池中，映射的外部LUN会被按照虚拟化数据组织形式创建为裸存储设备，通过这种方式创建的裸存储设备称为eDevLUN。eDevLUN在本端存储系统中占用的实际物理空间仅为元数据所需的存储空间，业务数据的存储空间仍位于异构存储系统上。应用服务器可以通过eDevLUN访问外部LUN上的数据，并在eDevLUN上配置SmartMigration增值业务。
- 在线接管：在线接管的整个过程业务不会中断，可以保证业务的连续性和数据的完整性。在线接管技术通过伪装异构LUN的关键识别信息，让多路径自动识别新存储系统，并将I/O切换到新存储系统，实现全程在线的异构接管。通过在线接管技术可以大幅降低数据迁移的技术复杂度以及缩短实施时间。
- 离线接管：离线接管过程中会断开异构存储系统和应用服务器之间的连线，业务会有短暂的中断。该方式适用于所有兼容的其他华为异构存储系统及第三方厂商的异构存储系统。
- 托管：将异构存储系统中的LUN映射给本端存储系统，供本端存储系统使用并进行管理的过程。
- 实际数据所需的物理空间为外部LUN的存储空间，由异构存储系统提供，不占用本端存储系统的空间。

- 元数据用于对eDevLUN的数据存储位置进行管理。元数据所需的存储空间由本端存储系统的存储池中的元数据空间提供。由于元数据占用的存储空间较小，所以创建eDevLUN后，本端存储系统占用的存储空间较小（未对eDevLUN配置任何增值特性时，每个eDevLUN在本端存储系统的存储池中大约占几十K空间）。
- 当配置了增值特性后，与本地LUN一样，eDevLUN也会占用本端存储系统的空间用于存放增值特性的元数据。请在创建eDevLUN时做好存储空间规划，保证增值特性能正常运行。

数据读流程

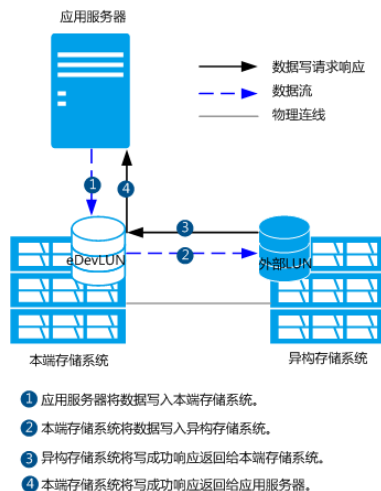
- 使用SmartVirtualization特性对异构存储系统中的外部LUN进行托管后，当应用服务器下发读取外部LUN数据请求时，本端存储系统的eDevLUN收到数据读取请求后从异构存储系统的外部LUN中读取。



- 使用SmartVirtualization特性后，应用服务器可以通过本端存储系统对位于异构存储系统外部LUN中的数据进行读写操作，其流程与应用服务器对本端存储系统中的LUN进行读写操作类似。

数据写流程

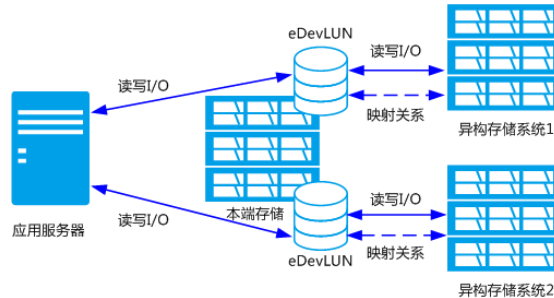
- 使用SmartVirtualization特性对异构存储系统进行托管后，数据写流程如下：



- 使用SmartVirtualization特性后，应用服务器可以通过本端存储系统对位于异构存储系统外部LUN中的数据进行读写操作，其流程与应用服务器对本端存储系统中的LUN进行读写操作类似。

存储资源集中管理

- 用户现场存在多套异构存储系统时，可能面临如下两个问题：
 - 由于存储系统和主机多路径软件的兼容性问题，一台应用服务器的多路径软件不能同时兼容多套存储系统。
 - 受到组网环境限制，一台应用服务器可能只能与一套存储系统连接（例如，光纤直连组网环境），但实际应用中应用服务器需要将业务分发至多套存储系统中。



- SmartVirtualization特性提供类似虚拟网关的功能。通过使用SmartVirtualization特性，用户能够通过本端存储系统发现多台异构存储系统中的存储资源，对其下发读写指令并且进行集中管理。

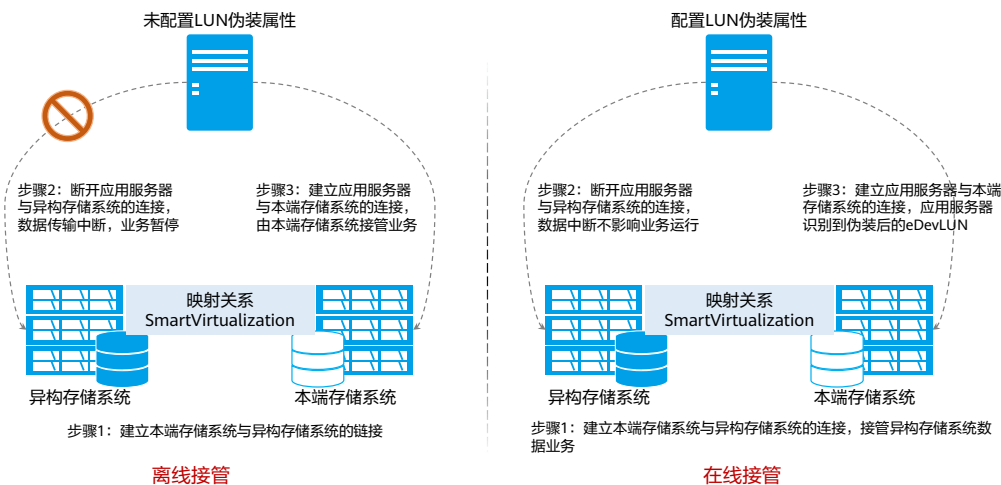
接管方式的选择

- SmartVirtualization特性提供离线接管和在线接管两种方式对异构存储系统进行接管，而在线接管又分为配置基本伪装接管、配置扩展伪装接管和配置第三方接管。选择哪种接管方式由异构存储系统的厂商及版本、多路径软件的厂商及版本等决定。

接管方式	是否伪装	说明
离线接管	不伪装	适用于所有兼容的其他华为异构存储系统及第三方厂商的异构存储系统，离线接管需要暂停应用服务器业务。
在线接管	基本伪装或扩展伪装	基本伪装和扩展伪装的选择与选择的多路径厂商及版本、其他华为异构存储系统的版本有关，具体请根据版本查看相关产品文档。

- 有关版本、兼容性的问题需要在实际部署前查看相关文档或兼容性列表，本章主要是讲解SmartVirtualization的理论。

离线接管和在线接管



- 离线接管

- 离线接管方式适用于所有兼容的其他华为异构存储系统及第三方厂商的异构存储系统。离线接管需要暂停应用服务器业务，eDevLUN的伪装属性为“不伪装”。

- 在线接管

- 在线接管华为异构存储系统时，eDevLUN的伪装属性为“基本伪装”或“扩展伪装”。基本伪装和扩展伪装的选择与选择的多路径厂商及版本、其他华为异构存储系统的版本有关。

业务数据迁移至新购存储系统

- 由于业务不断增长，需要存储的数据量越来越多，原有存储系统的容量已经无法满足数据存储容量和性能的需求。此时，用户可以通过购买容量更大、性能更高的存储系统，将原有存储系统进行升级换代。由于两套存储系统的软件和硬件都会存在差异，数据迁移过程可能会出现业务中断甚至数据丢失的风险。通过使用SmartVirtualization特性可以屏蔽两套存储系统的差异，将原有存储系统中的外部LUN映射成新购存储系统中的eDevLUN，然后配合使用SmartMigration特性，业务数据就能从原有存储系统中完整可靠的迁移至新购存储系统中，且保持业务的连续运行。

新购存储系统



原有存储系统



- 此应用场景中，新购存储系统被视为本端存储系统，原有存储系统被视为异构存储系统。
- 请播放动画

冷数据迁移至原有存储系统

- 存储系统升级换代后，经过一段时间的运行，用户发现存储系统中存在一些使用率偏低的冷数据。这些数据如果大量存储在新购的存储系统中，会使得存储资源的利用率减低，导致存储空间的浪费。为了降低运营成本，通过SmartVirtualization特性配合使用SmartMigration特性，可以将冷数据迁移至原有的异构存储系统中进行存储，实现资源利旧。



- 此应用场景中，新购存储系统被视为本端存储系统，原有存储系统被视为异构存储系统。
- 请播放动画

SmartMigration概述（块业务）

- 随着存储技术的不断发展和各行业的数字化推进，产生了因存储系统升级换代、资源优化利用以及业务性能调整而进行的存储系统内部和存储系统间业务数据迁移的需求。尤其在某些核心领域，中断主机业务可能会给用户带来巨大的损失和影响，如何实现能够支持在线情况下进行的业务数据迁移便成了需要解决的问题。SmartMigration技术应运而生，它不仅实现了在不中断主机业务的情况下完成存储系统内部的业务数据迁移，还实现了华为存储系统和与其兼容的其他存储系统之间的业务数据迁移。
- 华为技术有限公司研发的SmartMigration特性又称智能数据迁移特性，是业务迁移的关键技术，可以实现存储系统内和存储系统间的业务数据迁移。

- SmartMigration的特点：
 - 可靠的业务连续性：支持在线的业务数据迁移，避免了迁移过程中由于业务中断给客户造成的损失。
 - 稳定的数据一致性：进行业务数据迁移的过程中，主机产生的数据变更将及时同步至建立数据迁移关系的两个LUN中，保证了迁移完成后数据的一致性，避免遗漏。
 - 便捷的性能可调性：根据业务需求，实现不同存储介质间、不同 RAID 级别间的数据迁移，完成业务性能的自由调节。
 - 兼容的异构存储系统间数据迁移：SmartMigration不仅支持存储系统内部的业务数据迁移，还支持华为存储系统和与其兼容的异构存储系统之间的业务数据迁移。

SmartMigration原理（块业务）

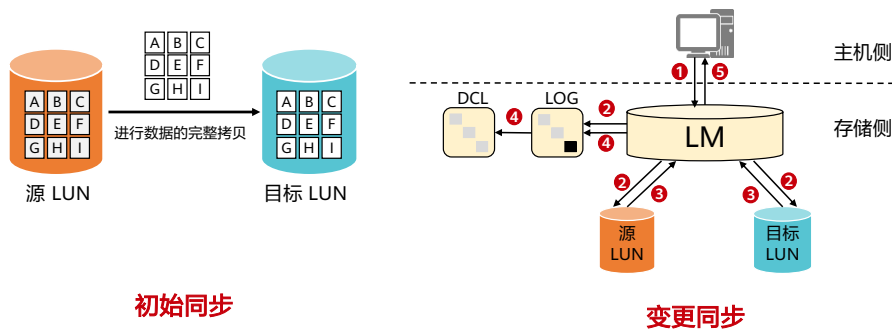
- SmartMigration特性的主要用途是通过对LUN上的数据进行迁移，从而调整业务性能或实现设备升级换代以满足用户需求。
- SmartMigration的实现过程分为两个阶段：



- 数据组织形式：存储系统采用虚拟化存储技术。存储池中的虚拟化数据由元数据卷和数据卷两部分组成。
 - 元数据卷：记录数据的存储位置，其中包含LUN ID和数据卷ID。LUN ID用以识别LUN，数据卷ID用以识别数据卷的物理空间。
 - 数据卷：存放实际的用户数据。
- 源LUN：业务数据迁出的LUN。
- 目标LUN：业务数据迁入的LUN。
- LM模块：存储系统中用于管理SmartMigration的模块。
- Pair：在SmartMigration特性中，Pair用于表示源LUN和目标LUN之间的数据迁移关系。一个源LUN仅和一个目标LUN构成一个Pair。
- 双写：业务数据迁移中，将数据变更同时写入源LUN和目标LUN的过程。
- LOG：记录数据变更写入源LUN的操作日志，用以判断目标LUN是否双写成功。
- DCL：数据变更日志（Data change log），在数据变更同步中记录未成功写入目标LUN的差异数据。
- 分裂：源LUN和目标LUN停止业务数据同步，开始LUN信息交换，并在信息交换后解除数据迁移关系的过程。

SmartMigration业务数据同步（块业务）

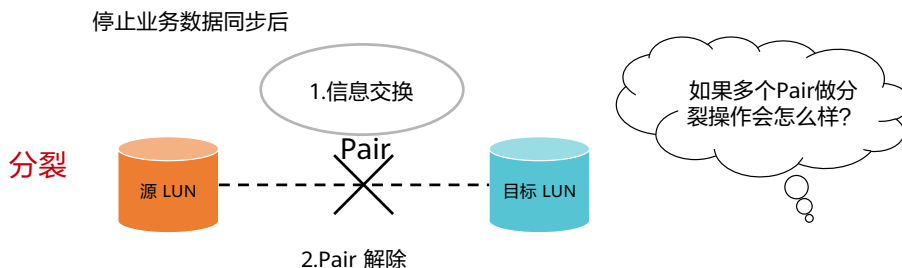
- 通过业务数据同步将源LUN数据同步至目标LUN，数据同步完成之后，目标LUN和源LUN上的业务数据保持完全一致。
- 业务数据同步包括初始同步和变更同步，两种同步方式相互独立，可同时进行。保证了源LUN业务数据同步至目标LUN的过程中，主机的业务数据变更能完整地同步至源LUN和目标LUN两端。



- Pair：在SmartMigration特性中，Pair用于表示源LUN和目标LUN之间的数据迁移关系。一个源LUN仅和一个目标LUN构成一个Pair。
- 两种业务数据同步方式是独立的，可同时进行。保证了源LUN业务数据同步至目标LUN的过程中，主机的业务数据变更能完整地同步至源LUN和目标LUN两端。
- 初始同步
 - 创建SmartMigration，源LUN和目标LUN之间建立Pair关系，源LUN启动业务数据同步，将初始业务数据完整拷贝至目标LUN。
- 变更同步：
 - 1. 主机下发数据写I/O请求至存储侧存储系统的LM模块。
 - 2. LM 模块同时将数据写入源LUN和目标LUN，并将写I/O操作记录至LOG。
 - 3. 源LUN和目标LUN返回数据写I/O结果至LM模块。
 - 4. LM 模块根据写I/O结果，判断是否清除LOG。
 - 5. 返回主机写I/O成功。

SmartMigration分裂（块业务）

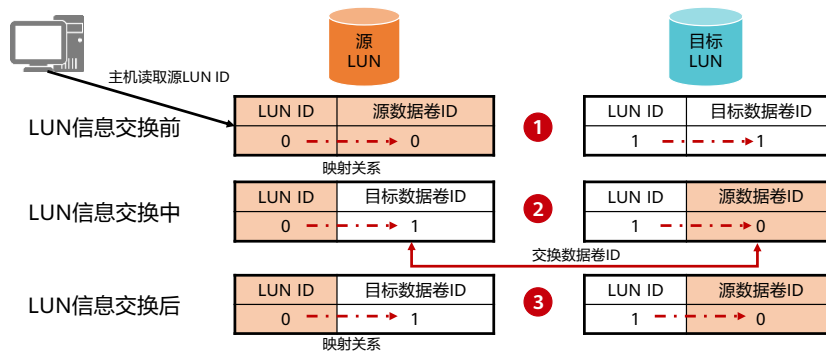
- 分裂是针对单个Pair的操作，它的过程包括构成Pair的单个源LUN和目标LUN停止业务数据同步开始进行LUN信息交换，并在LUN信息交换后解除数据迁移关系。



- 分裂时，主机业务被悬挂，信息交换后，业务下发到迁移后的LUN，从而实现用户无感知情况下的业务迁移。
- Pair解除：指源LUN和目标 LUN，在信息交换后解除数据迁移关系。
 - Pair解除后，如果主机侧下发I/O请求至存储侧，数据只写入源LUN（源LUN ID指向的物理空间为目标数据卷），不再进行双写。
 - 目标LUN将封存Pair解除时刻源LUN的所有数据。
 - Pair 解除后不能再重新建立联系。
- SmartMigration的一致性分裂是指多个Pair同时进行LUN信息交换，并在信息交换完成后同时解除Pair关系，保证分裂前后各时刻点数据的一致性。
- 在某些有多个Pair的场景下，例如大中型数据库，数据、日志、修改记录等文件存放在存储系统的有关联的不同LUN中，使用分裂无法保证每个时刻点下各个LUN中的信息都相互关联，缺失了其中某些LUN的数据，都将导致其他LUN中数据失效。针对这样的情况，一致性分裂便提供了保障。

SmartMigration LUN信息交换（块业务）

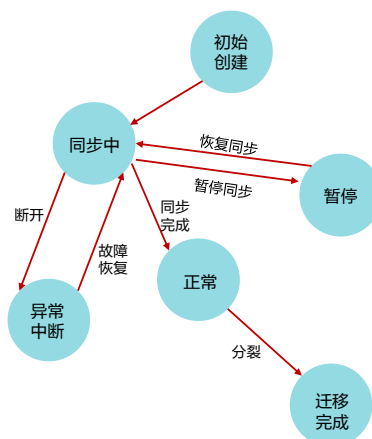
- LUN信息交换主要是针对LUN和数据卷之间的映射关系，即源LUN和目标LUN的LUN ID保持不变的情况下，将两者的数据卷ID相互交换。



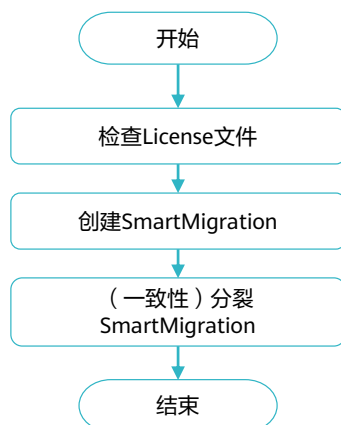
- LUN信息交换是业务信息同步后目标LUN能够顺利地代替源LUN来承载业务的前提。
- 存储系统中，每个LUN和对应的数据卷都有属于自己的唯一标识，分别为LUN ID和数据卷ID。其中，LUN是逻辑上的概念，数据卷是物理上的概念，两者之间形成一一对应的映射关系。
- LUN信息交换前：主机通过源LUN ID识别源 LUN。LUN ID和数据卷ID形成映射关系；
- LUN信息交换中：交换源数据卷ID和目标数据卷ID。源LUN指向的物理空间变为目标数据卷；
- LUN信息交换后：源LUN ID未发生变化，主机读写不受影响（用户无感知）。源LUN ID和目标数据卷ID形成新的映射关系。主机实际读写到的是目标LUN的物理空间。

SmartMigration迁移状态转换（块业务）

- **正在迁移中**：源LUN的数据正在后台拷贝到目标LUN。
- **正常**：源LUN的数据和目标LUN的数据一致，主机下业务时进行实时双写。
- **暂停**：迁移中手动停止迁移。
- **异常中断**：在迁移过程中，由硬盘域故障、源LUN故障、目标LUN故障或存储池满等迁移I/O错误导致迁移中断，可自愈恢复。
- **迁移故障**：删除失败、创建失败或分裂失败时会显示故障。
- **迁移完成**：源LUN与目标LUN的数据成功同步完成，并且完成分裂。



块业务迁移配置流程



- 上图所示为存储系统内SmartMigration任务的配置流程，如果要进行存储系统间的SmartMigration任务，需要借助 SmartVirtualization 特性的LUN接管功能，因此检查License文件后，需进行“接管异构存储系统的 LUN”操作，然后再执行“创建 SmartMigration”、“（一致性）分裂 SmartMigration”等配置。

SmartMigration概述（文件业务）

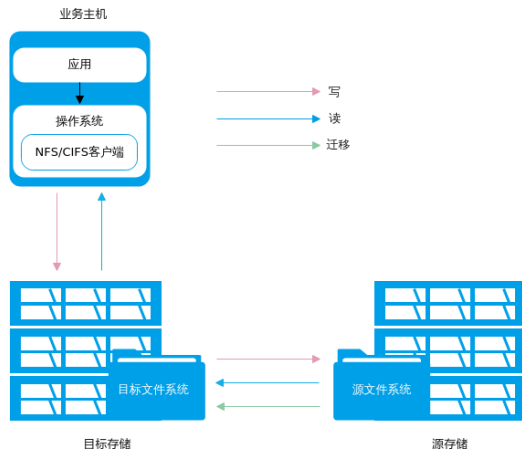
- 随着存储技术的不断发展和各行业的数字化推进，产生了因存储系统升级换代、资源优化利用以及业务性能调整而进行的存储系统间业务数据迁移的需求。NAS数据迁移中存在着数据一致性、业务连续性和迁移速度等多个问题，SmartMigration通过Takeover-first、Copy-first和Scan模式三种关键技术，实现了华为存储系统之间、华为存储系统和与其兼容的其他存储系统之间的NAS数据迁移。

目的和受益	详细说明
割接停机时间短	与传统的主机层迁移方案相比，可以将任意大小的文件系统割接停机时间控制到分钟级。
兼容的异构存储系统间数据迁移	SmartMigration不仅支持华为存储系统之间的业务数据迁移，还支持华为存储系统和与其兼容的异构存储系统之间的业务数据迁移。
免迁移服务器迁移	无需部署迁移服务器，即可进行存储系统之间的数据迁移，节省配置成本。

SmartMigration原理（文件业务）

- SmartMigration分为Scan模式、Takeover-first模式和Copy-first模式。其中，Scan模式用于进行正式迁移前的预迁移检查，为正式迁移提供质量保证；Takeover-first模式优先接管业务，先将源存储的业务割接到目标存储上，再进行数据迁移；Copy-first模式优先进行数据迁移，再将源存储的业务割接到目标存储上。
- 系统支持将源端的任意共享迁移至目标端文件系统或Dtree级别的共享。
- Scan模式
 - 在源存储不停止业务的情况下进行一次数据迁移。仅迁移文件的元数据，不迁移数据。用于Takeover-first或Copy-first模式正式迁移前的数据评估和演练，提前判断数据能否正常迁移。
- Takeover-first模式
 - 在Takeover-first模式下，目标文件系统可以在数据没有迁移到目标存储的情况下对外提供CIFS、NFS或跨协议共享。迁移完成后的Takeover-first模式的目标文件系统可以在客户端IO不中断的情况下断开与源文件系统的关系，切换为普通文件系统。
- Copy-first模式
 - Copy-first模式包括普通拷贝和最终拷贝两种后台拷贝方式，没有完成最终拷贝任务的文件系统或Dtree为只读状态。最终拷贝任务完成后，手动删除迁移任务，转为客户端可读写状态。

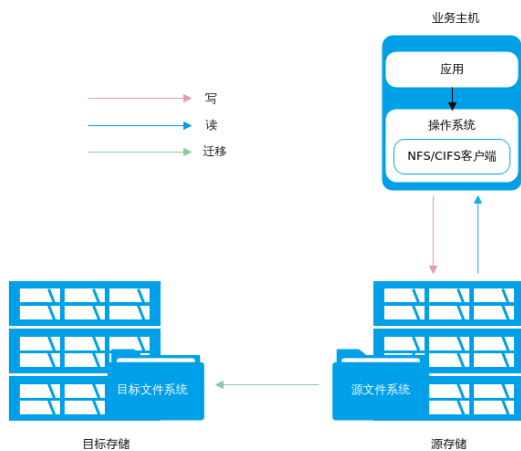
Takeover-first模式



- Takeover-first的迁移过程:

- 在目标存储使用目标文件系统配置共享，并配置迁移任务。
- 通过修改权限或删除与客户端交互的IP等方式停止源文件系统处理客户端业务，使得客户端挂起。在客户端重新挂载共享目标文件系统，完成目标文件系统接管业务。对客户端的I/O数据读取请求，通过I/O重定向对未迁移的源文件系统的数据进行读取。对客户端的I/O数据写入请求，通过I/O重定向对源文件系统和目标文件系统进行双写。
- 启动数据迁移，存储系统在客户端进行访问期间将所有文件元数据和数据从源文件系统后台拷贝到目标文件系统。
- 所有数据迁移完成后，删除迁移任务。

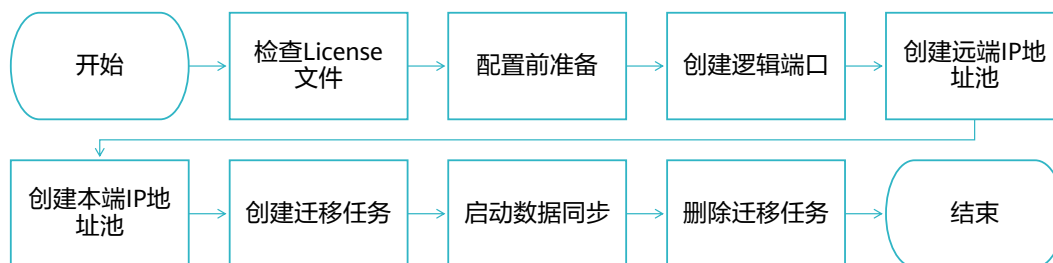
Copy-first模式



Copy-first模式的迁移过程：

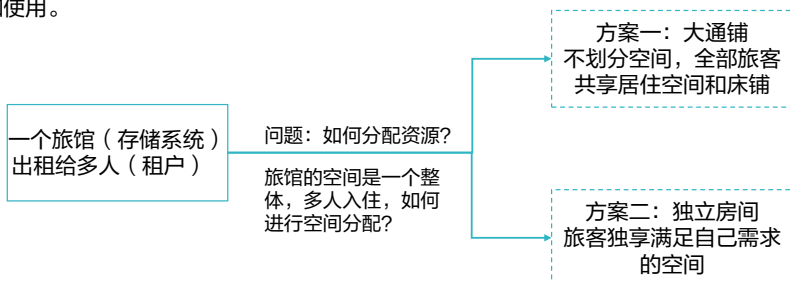
- 在目标存储使用目标文件系统配置共享，并配置迁移任务。
- 启动数据迁移，期间源存储的源文件系统继续处理客户端业务。
- 在一次全量迁移和一轮或多轮从源存储到目标存储的增量迁移后，停止源文件系统的客户端I/O，进行最后一次增量最终拷贝。
- 所有数据迁移完成后，删除迁移任务。
- 通过修改权限或删除与客户端交互的IP等方式停止源文件系统处理客户端业务，使得客户端挂起。在客户端重新挂载共享目标文件系统，完成目标文件系统接管业务。

文件业务迁移配置流程



SmartMulti-Tenant概述

- 多租户特性又称SmartMulti-Tenant，实现了在一套物理存储系统中创建多个虚拟存储系统，让租户在多协议统一存储架构中既能共享相同的存储硬件资源，又不影响相互的数据安全性和隐私。
- 多租户技术可以帮助在共用的数据中心内以单一系统架构与服务提供多数客户端相同甚至可定制化的服务，并且仍然可以通过逻辑隔离保障租户的业务及网络安全。每个租户下的资源都是相对独立的，其他租户无法看到和使用。



- 从上面的例子可以看出：很多旅客（租户）可以共用同一个旅馆（存储系统），不同的旅客有不同的需求。根据对旅馆房间不同的划分，旅客可使用满足自己需求的房间。例如有50名旅客要租房，根据旅客的使用需求和预算情况，有的需要面积（存储容量）大的房间，有的只需要面积小的房间。那么旅馆就可以将整栋楼划分为多个面积不同的房间供旅客选择，而不是给每个旅客去建一栋适应旅客需求的旅馆。

相关概念

- vStore (Virtual Store)
 - 存储系统可被划分成多个虚拟存储，这些虚拟存储被称为vStore。一套物理存储系统可以划分出多个vStore，vStore间共享存储池。系统管理员和租户用户为每个vStore中分配存储资源。系统自带默认租户System_vStore，不指定租户创建的资源(LUN/FS/协议等等)都归属于默认租户。
- 系统视图
 - 系统管理员的默认视图，系统管理员可以在该视图下创建租户并进行全局和租户级资源管理。未激活多租户License时，该视图下不能创建租户，作为单台设备创建存储资源并管理，主要给没有多租户需求的客户使用，以免改变以往设备管理的模式。激活多租户License后，可以创建租户并对租户的资源进行配置和管理操作。
- 租户视图
 - 配置和管理租户业务。激活多租户License后，系统管理员可以在系统视图下创建租户和租户用户，然后使用租户用户登录到租户视图，配置和管理对应租户的存储资源。
- 预置默认角色
 - 系统预置给系统管理用户和租户用户的默认角色。
- 自定义角色
 - 系统管理用户依据场景自定义系统或租户下的用户角色。

租户隔离（块业务）

- 多租户技术可以让当前的存储系统为多个租户提供资源分配和管理，提高资源利用率的同时提升租户安全性。
- 资源分配
 - 存储系统可被划分成多个虚拟存储分区，这些分区被称为vStore。一套物理存储系统可以划分出多个vStore，vStore间共享存储池。系统管理员可以为每个Vstore分配块存储资源，租户用户也可以直接登录对应的租户视图配置和管理块存储资源。每个租户下创建的LUN资源可被映射给上层应用服务器使用。
- 用户管理
 - 用户的管理通过权限和角色共同定义，权限和角色的所属组分为系统组和租户组，其中租户组的角色仅用于租户。不同角色的用户可拥有不同的访问控制权限。
 - 系统角色：可以为存储系统划分租户并分配存储资源的系统默认角色或系统自定义角色。
 - 租户角色：可以在租户视图下完成租户的相关配置的默认租户角色或自定义租户角色。

租户隔离（文件）

- 多租户特性主要解决租户之间的隔离问题，包括管理隔离、业务隔离、网络隔离。租户之间不能相互访问数据，以此来达到安全隔离的效果。
 - **管理隔离：**华为闪存存储系统支持租户管理用户隔离，不同租户下可以配置不同的管理用户。不同租户下能够操作和管理的资源统称为视图，租户管理员仅能在本租户视图下进行资源管理，不能访问其他租户视图和系统视图。租户管理用户支持基于角色的权限控制，创建租户管理用户时，需要为用户配置所需权限对应的角色。
 - **网络隔离：**租户使用逻辑端口LIF配置对象业务，一个LIF只能归属一个租户，做到逻辑上的端口隔离。可以选择从绑定口、VLAN口及以太网端口创建LIF。
 - **业务隔离：**多租户的业务隔离，主要体现在用户的业务数据、业务访问和业务配置（NAS协议配置等）隔离。
 - 业务数据隔离：每个租户有自己独立的文件系统，以此达到租户文件系统的隔离。同理基于文件系统的配额等业务也是隔离的。
 - 业务访问隔离：每个租户都具有独立的NAS协议实例，包括NFS服务、NDMP服务。
 - 业务配置隔离：每个租户可以有自己独立的用户、用户组、用户映射规则、安全策略、NFS共享、AD域、DNS服务、LDAP服务、NIS服务等。

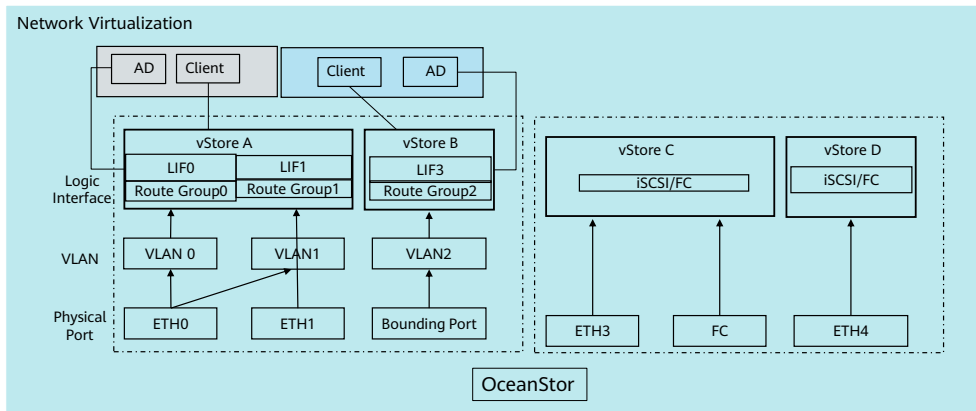
- 多租户的业务隔离，主要体现在用户的业务数据、业务访问和业务配置隔离。
 - 业务数据隔离：每个租户有自己独立的对象业务资源，以此达到租户级的隔离。
 - 业务访问隔离：每个租户都具有独立的S3协议实例。
 - 业务配置隔离：每个租户可以有自己独立的桶，桶策略，对象用户，对象用户组，权限策略等。

租户隔离（对象）

- 多租户特性主要解决租户之间的隔离问题，包括管理隔离、业务隔离、网络隔离。租户之间不能相互访问数据，以此来达到安全隔离的效果。
 - **管理隔离：**华为闪存存储系统支持租户管理用户隔离，不同租户下可以配置不同的管理用户。不同租户下能够操作和管理的资源统称为视图，租户管理员仅能在本租户视图下进行资源管理，不能访问其他租户视图和系统视图。租户管理用户支持基于角色的权限控制，创建租户管理用户时，需要为用户配置所需权限对应的角色。
 - **网络隔离：**租户使用逻辑端口LIF配置对象业务，一个LIF只能归属一个租户，做到逻辑上的端口隔离。可以选择从绑定口、VLAN口及以太网端口创建LIF。
 - **业务隔离：**多租户的业务隔离，主要体现在用户的业务数据、业务访问和业务配置隔离。
 - 业务数据隔离：每个租户有自己独立的文件系统，以此达到租户文件系统的隔离。同理基于文件系统的配额等业务也是隔离的。
 - 业务访问隔离：每个租户都具有独立的S3协议实例。
 - 业务配置隔离：每个租户可以有自己独立的桶，桶策略，对象用户，对象用户组，权限策略等。

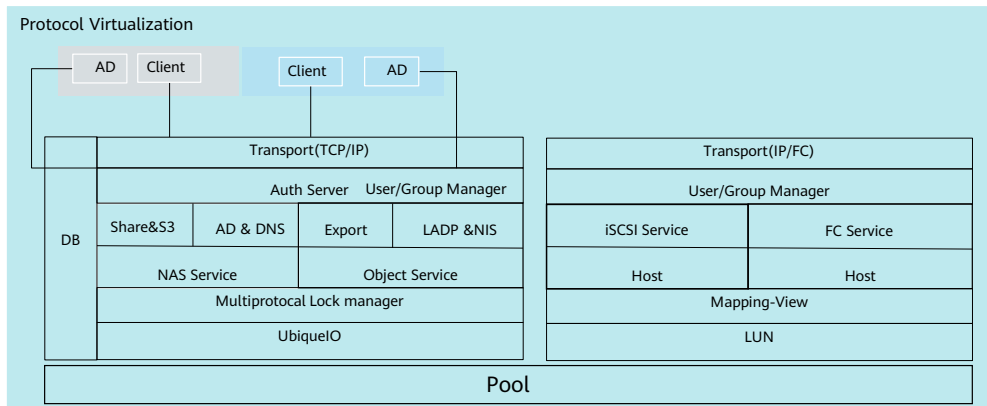
- 多租户的业务隔离，主要体现在用户的业务数据、业务访问和业务配置隔离。
 - 业务数据隔离：每个租户有自己独立的对象业务资源，以此达到租户级的隔离。
 - 业务访问隔离：每个租户都具有独立的S3协议实例。
 - 业务配置隔离：每个租户可以有自己独立的桶，桶策略，对象用户，对象用户组，权限策略等。

网络隔离



- 一个LIF只能归属于一个租户，从而确保LIF资源的租户间隔离。
- 每个租户只能使用分配好的LIF口提供业务访问，客户端通过指定的LIF只能访问LIF归属的租户下的LUN、文件系统、对象等资源。
- 每个租户只能使用自己的路由表与外部通信，包括与客户端、AD、LDAP、NIS、DNS、主机等。
- SAN网络隔离：ISCSI/FC均隔离了启动器，相当于在逻辑链路上做了隔离，FC口支持物理隔离。

业务隔离



- 业务配置隔离
 - 每个租户可以有自己独立的用户、用户组、用户映射规则、NFS共享、AD域、NIS域、LDAP域、DNS服务、桶、桶策略等。其中，租户内的共享只能基于本租户的文件系统创建，租户内的映射视图只能关联本租户的主机和LUN。
- 业务访问隔离
 - 每个租户都具有独立的协议实例，包括SMB、NFS、iSCSI、FC、S3等服务。
- 业务数据隔离
 - 不同租户间无法访问对方的LUN、文件系统 and 对象，从而确保数据均是按租户隔离的。

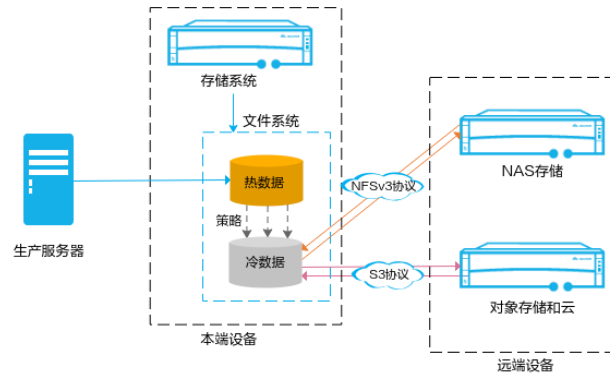
SmartMobility概述

- 随着日益迅速的数据增长，越来越多的企业希望将不常访问的数据迁移到成本更低的云或其他NAS设备中去。但是企业同时也希望不要影响到现有应用的业务，希望应用不感知文件系统的数据迁移。SmartMobility特性应运而生，它在上层应用不感知的情况下，将不常访问的冷数据迁移到云和其他NAS设备。这样的最大好处是通过节省本地系统空间，从而达到了降低成本的目的。
- SmartMobility特性是将文件系统的数据按照冷热数据进行分级存储，热数据存放在本地高性能的存储设备，冷数据存放在远端的存储设备（对象存储设备、其他NAS设备），并且支持对已经迁移到远端设备的文件进行实时读写和后台召回，对用户不仅做到了透明访问，而且做到了空间和性能的平衡。

相关概念

- 迁移：将文件系统的文件数据按照策略自动迁移到远端设备（对象存储、NAS设备）。
- 召回：当用户读写或修改已迁移到远端设备的对象时，会在后台自动将远端设备的对象数据拷贝回本端设备，然后将远端的对象进行删除。
- 迁移策略：一个文件系统能够支持10条迁移策略，每个文件系统有不同的迁移策略。迁移策略包括迁移模式（周期/一次性），按照文件名称（支持模糊匹配，有完全匹配，前缀匹配、后缀匹配三种方式）、文件大小、文件操作时间、文件目录进行迁移。多个迁移策略之间是“或”的关系，只要文件满足任意迁移策略就进行迁移。
- 迁移模式：
 - 1）周期性：用户通过按照月和周进行设置，周期性的进行迁移
 - 2）一次性：只执行一次，用于手工触发迁移
- 开始迁移阈值：文件系统剩余容量小于此阈值时，就开始迁移
- 停止召回阈值：文件系统已用容量超过此阈值时，就停止召回

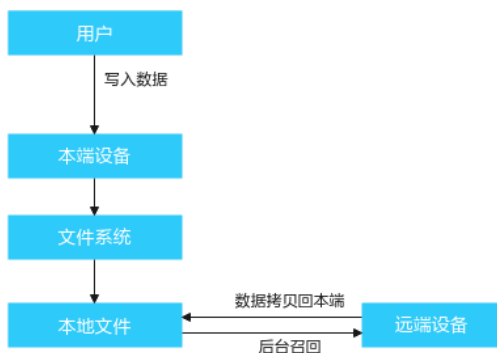
迁移原理



- SmartMobility特性通过制定的策略将本端的文件系统的数据分成热数据和冷数据，热数据存放在本端设备，冷数据通过不同的协议迁移到远端设备。
 - 文件系统通过NFSv3协议与远端NAS设备进行通信；通过S3协议与远端对象存储设备进行通信。
 - 一个文件系统只支持与一种远端设备连接，不支持同时连接对象存储设备和其他NAS设备。
 - 一个文件系统的冷数据存放于远端NAS设备的共享路径或远端对象存储设备的桶。
 - 与本端对接的远端设备的文件系统/命名空间，只允许本端设备写入，不允许远端设备其他业务写入。

写召回原理

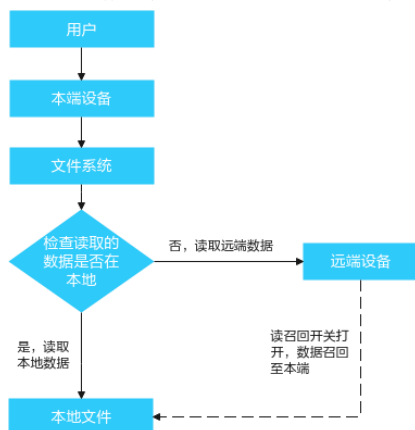
对于那些已经被迁移到远端设备的文件，在用户进行写入时，会先将写入数据存放本地，然后再通过后台将文件从远端拷贝回本地。这样的好处在于不用等待文件被召回后再写入，所以速度非常快。



1. 如果一个文件已经迁移到远端设备，在用户写入新数据到此文件时，会将写入数据存放本地文件系统。
2. 用户在写入数据时，文件系统会记录写召回日志。后台会处理此召回日志，将此文件从远端设备进行召回。
3. 召回完成后，会将远端设备的对象删除，用户访问此文件时，直接从本地设备进行访问。

读召回原理

通常场景下，远端设备的性能都比本端设备较差，在读取已经迁移到远端设备的文件时，会先从本地读取数据，没有读到再从远端设备读取要访问的数据返回给用户，最后在后台将其召回。召回后，用户将不再读取远端的数据，用户的读取性能就能提升。



1. 如果一个文件已经迁移到远端设备，在用户读取文件时，可能本地存放有最新的数据，所以优先从本地读取
2. 如果本地读取未命中，说明数据存放在远端，就会直接从远端设备读取（不用召回后再从本地读取）。
3. 用户读取文件后，会触发读召回（读召回有开关，默认开启，用户可以关闭）。这时文件系统会记录读召回日志。后台会处理此召回日志，将此文件从远端设备进行召回。

SmartMove概述

- 在实际的生产中经常遇到文件系统所属存储池容量不足、文件系统所在存储池的性能较低或控制器负载不均衡等问题，需要将文件系统迁移到容量更大，性能更好的存储池或低负载控制器的存储池上。或者因存储系统升级换代，需要将文件系统迁移到新一代存储系统上。因此如何实现文件系统迁移变成了需要解决的问题。
- SmartMove特性可以实现在不中断业务的情况下完成阵列内和同构阵列间的文件系统数据迁移。
- SmartMove包括阵列内SmartMove和阵列间SmartMove。
 - 阵列内SmartMove能够在不中断主机业务的前提下实现存储系统内不同存储池文件系统之间的业务数据迁移。
 - 阵列间SmartMove能够在不中断主机业务的前提下实现同构存储系统（即OceanStor存储系统）之间文件系统的业务数据迁移，并在数据同步完成后将业务切换到目标文件系统上。

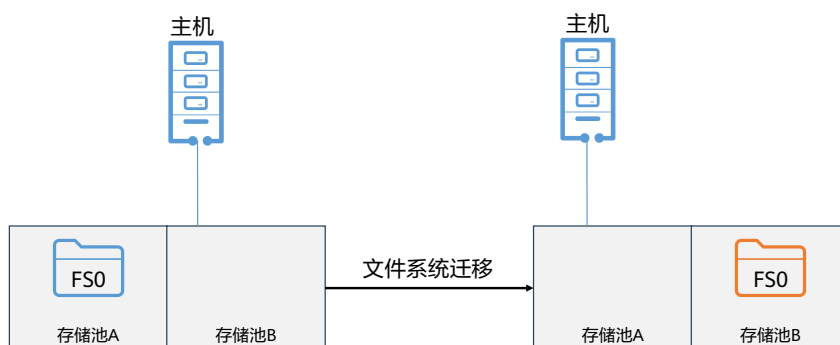
相关概念（1）

- 源文件系统
 - 业务数据迁出的文件系统。
- 目标文件系统
 - 业务数据迁入的文件系统。
- 逻辑端口
 - 逻辑端口的创建基于物理以太网端口、绑定端口或VLAN，用于主机访问存储系统以及SmartMove两端存储系统的对接。
- SmartMove域
 - 是指本端设备与远端设备的集合。配置SmartMove租户Pair前，需要配置SmartMove域，每个SmartMove租户Pair必须都是基于SmartMove域创建的。
- 租户
 - 租户是指在一套物理存储系统中创建多个虚拟存储系统，让多个租户既能共享相同的存储硬件资源，又不影响相互的数据安全性和隐私。借助租户特性，可以在存储架构中实现更加灵活、更易于管理、部署成本更低的共享存储。

相关概念（2）

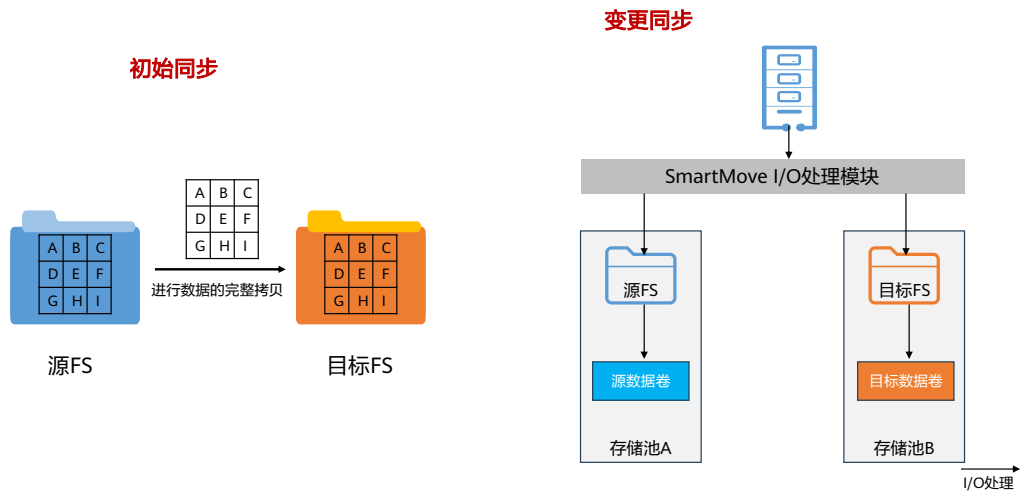
- SmartMove租户Pair
 - SmartMove租户Pair是指源端存储系统的租户和目标存储系统的租户之间的SmartMove关系。
- SmartMove文件系统Pair
 - SmartMove文件系统Pair是指源端文件系统 and 目标文件系统之间的SmartMove关系。
- 本端资源、远端资源
 - 在SmartMove文件系统Pair中，本端存储系统中的文件系统称为本端资源，远端存储系统中的文件系统称为远端资源，两者是相互的。
- 本端资源角色
 - 配置SmartMove时执行操作的一端为源端，另一端为目标端。本端资源角色为本端资源在SmartMove租户Pair中的角色，文件系统数据由源端同步到目标端。

阵列内SmartMove



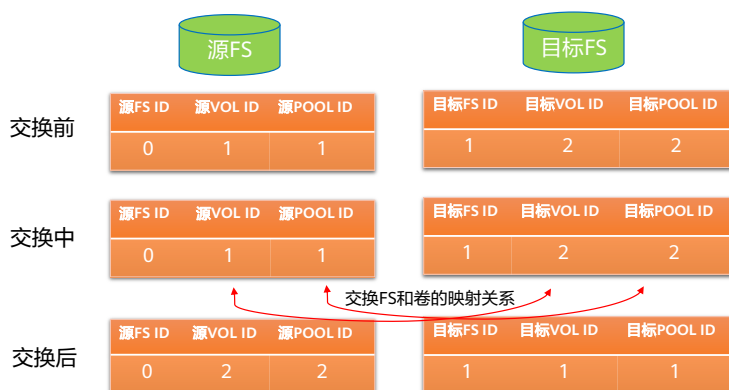
- 基于文件系统数据同步技术，实现阵列内不同文件系统数据和配置同步，同步完成后启动源文件系统 and 目标文件系统交换映射关系，交换完成后对主机展现的文件系统的属性不变。文件系统迁移后改变了文件系统的物理位置，但是逻辑位置保持不变。

业务数据同步



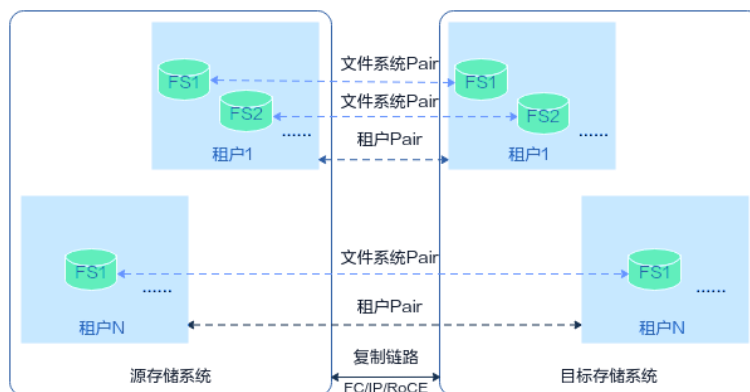
- 创建SmartMove，源文件系统和目标文件系统之间建立Pair关系。源文件系统和目标文件系统之间的业务数据同步包括初始同步和变更同步。两种同步方式是独立的，可同时进行。保证了源文件系统业务数据同步至目标文件系统的过程中，主机的业务数据变更能完整地同步至源文件系统和目标文件系统两端。
 - 初始同步：源文件系统启动业务数据同步，将初始业务数据完整拷贝至目标文件系统。
 - 变更同步：同步过程中，主机无需中断业务。当主机出现业务数据变更，下发I/O数据写入请求至存储系统，存储系统对源文件系统的数据卷和目标文件系统的数据卷进行双写。

交换映射对象



- 交换映射对象是业务信息同步后目标文件系统能够顺利地代替源文件系统来承载业务的前提。
 - 交换映射对象前，主机通过源文件系统ID识别到源文件系统。由于源文件系统ID和用以识别物理空间的源数据卷ID间存在映射关系，主机读取到源文件系统的物理空间信息。目标文件系统ID和目标数据卷ID此时也存在映射关系。
 - 交换映射对象中，文件系统ID均保持不变，交换源文件系统和目标文件系统的数据卷ID，使得源文件系统ID和目标数据卷ID形成新的映射关系。
 - 交换映射对象后，主机仍然通过源文件系统ID识别到源文件系统，但由于源文件系统ID和目标数据卷ID形成了新的映射关系，实际读取的物理空间信息由源数据卷变为目标数据卷，从而实现了用户无感知情况下的业务迁移。

阵列间SmartMove



- 阵列间的SmartMove可实现同构存储系统之间的文件系统数据迁移，并在迁移结束后将业务切换至目标文件系统。
- 阵列间的SmartMove为租户粒度迁移。源文件系统和目标文件系统归属于不同阵列的不同租户，通过建立SmartMove租户Pair建立起两边租户之间的SmartMove关系，并将源端租户的配置信息同步至目标端租户，包括域控（AD域、NIS域、LDAP域、DNS服务）、LIF（LIF、DNS Zone、路由）、Unix用户（本地用户、本地用户组、用户映射规则）、Windows用户（本地用户、本地用户组、用户映射规则）、全局配置（SMB2/SMB3开关、CIFS签名配置开关等）、CIFS（共享及共享权限）、NFS（共享及共享权限）等。基于SmartMove租户Pair为租户中所有文件系统建立SmartMove文件系统Pair并进行数据同步。在SmartMove文件系统Pair的数据同步完成后，将源文件系统的配置信息（包括文件系统基本属性、Dtree、配额、快照、审计日志配置项和WORM配置）同步至目标文件系统，然后将业务切换至目标存储系统。

目录

1. 数据存储全景
2. 闪存存储产品介绍
3. 闪存存储产品部署
- 4. 闪存存储增值特性**
 - 效率提升特性
 - **数据保护特性**

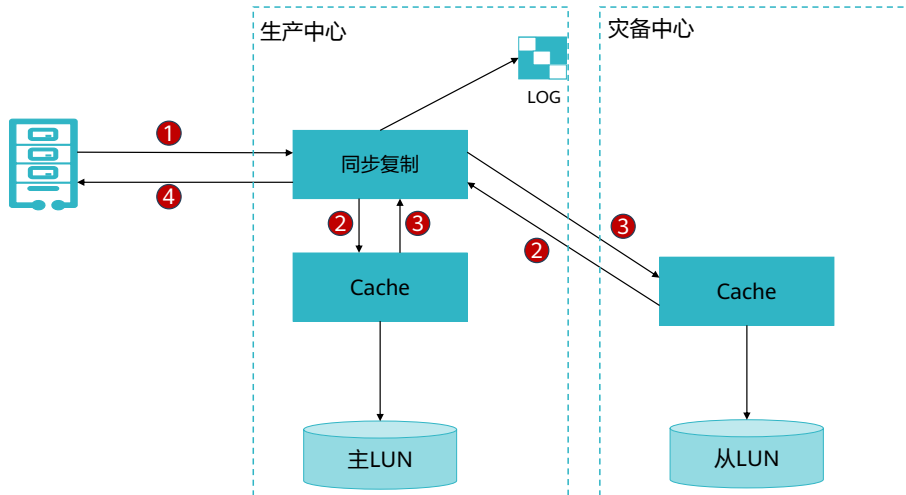
HyperReplication概述

- 远程复制是容灾备份的核心技术，可以实现远程数据备份和灾难恢复。远程复制有以下两种复制模式：
 - 同步远程复制
 - 实时地同步数据，最大限度保证数据的一致性，以减少灾难发生时的数据丢失量。
 - 异步远程复制
 - 周期性地同步数据，最大限度减少由于数据远程传输的时延而造成的业务性能下降。

远程复制的作用	目的	受益
远端备份和恢复	业务数据失效后，可以通过远端的备份数据进行恢复。	避免主站点数据失效后，由于数据丢失给客户造成的损失。
持续的业务支撑	灾难发生时可以通过从站点快速接管主站点的业务数据，以保证业务的可靠性。	避免主站点故障后，由于业务中断给客户造成的损失。
容灾恢复	灾难发生后可以通过从站点的数据实现对主站点数据的恢复。	避免灾难发生后，由于业务数据丢失或恢复时间过长给客户造成的损失。

- 主站点：指主端存储系统、应用服务器、链路等构成的生产中心。
- 从站点：指从端存储系统、应用服务器、链路等构成的备份中心。
- 主机：如果没有特殊说明，本课程中的主机特指应用服务器。

LUN同步远程复制原理

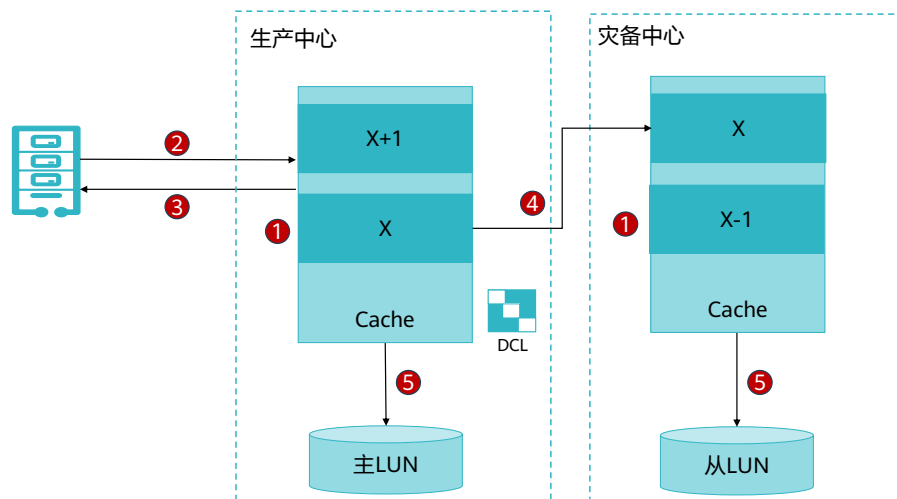


130 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



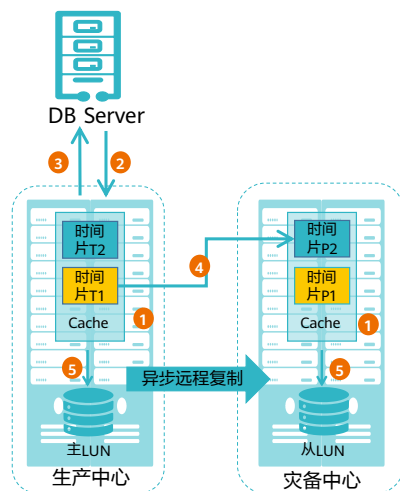
- 初始同步：生产中心的主LUN和灾备中心的从LUN建立同步远程复制关系，启动初始同步。
 - 将主LUN数据全量拷贝到从LUN。
 - 初始同步中主LUN收到主机写请求也会同样写到从LUN。
- 双写状态：初始同步完成以后，进入正常状态，此时主、从LUN数据相同。
 1. 生产存储收到主机写请求。HyperReplication将该请求记录日志。日志中只记录地址信息，不记录数据内容。
 2. 将该请求写入主LUN和从LUN。通常情况下LUN是回写状态，数据会写入Cache。
 3. HyperReplication等待主LUN和从LUN的写处理结果都返回，如果写从LUN超时或失败时远程复制关系断开。如果都写成功，清除日志；否则保留日志，写入盘上DCL（Data Change Log 数据变更日志）中，进入异常断开状态，后续启动同步时重新复制该日志地址对应的数据块。
 4. 返回主机写请求处理结果，以写主LUN的处理结果为准。
- LOG：记录数据写入操作的日志。
- DCL：Data change log，数据变化位图。

LUN异步远程复制原理



- 当生产站点的主LUN和远端复制站点的从LUN建立异步远程复制关系以后，默认情况下会启动一个初始同步，将主LUN数据全量拷贝到从LUN，以保证主LUN和从LUN的数据完全一致。
- 初始同步完成后，从LUN数据状态变为完整（即从LUN为主LUN的过去某个时刻的一致性拷贝），然后开始按照下面的流程进行I/O处理：
 - 异步远程复制同步任务启动时，在主LUN和从LUN分别生成快照并更新时间点（主LUN快照X，从LUN快照X-1）；
 - 主机新写入的数据缓存在主LUN缓存的时间点X+1中；
 - 响应主机写完成；
 - 将主LUN时间点X的数据基于DCL（Data Change Log）增量差异记录直接复制到从LUN，当数据同步完成后，删除主LUN与从LUN的快照；
 - 主从LUN各自将收到的数据下盘，数据下盘的同时记录DCL，数据下盘完成后从LUN最新数据即为主LUN时间点X的完整数据。

异步远程复制多时间点技术



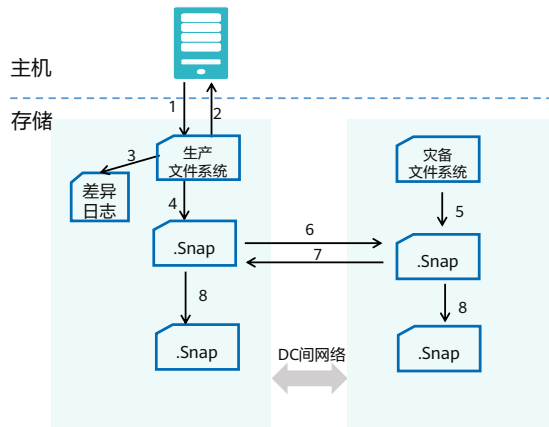
最小3秒一个一致性点

- 1 每个复制周期启动时在主LUN和从LUN的缓存中产生新的时间片 (T2, P2);
- 2 主机新写入的数据缓存在主LUN Cache的时间片T2中;
- 3 响应主机写完成;
- 4 将时间片T1的数据直接复制到从LUN, 写入从LUN的时间片P2中;
- 5 主从LUN各自将收到的数据下盘。

➢ 拷贝时**直接从Cache读取数据**，时延小。

➢ 快照不需实时更新数据，同步对性能的影响小，复制周期可缩短到**3秒**。

基于文件系统的远程复制



初始同步:

当主文件系统和从文件系统建立文件系统远程复制关系之后,通常会启动初始同步,初始同步前创建主文件系统的快照,将快照时间点的主文件系统数据全部复制到从文件系统,以保证从文件系统数据和主文件系统在某一时刻数据完全一致。初始同步完成后,从文件系统数据状态变为完整(即从文件系统为主文件系统的过去某个时刻的一致性拷贝)。

初始同步完成后,后续数据复制过程如下:

主机写:

- 1、客户端写生产文件系统。
- 2、写本地完成,回主机IO。
- 3、cache刷盘时后台记录写差异日志。

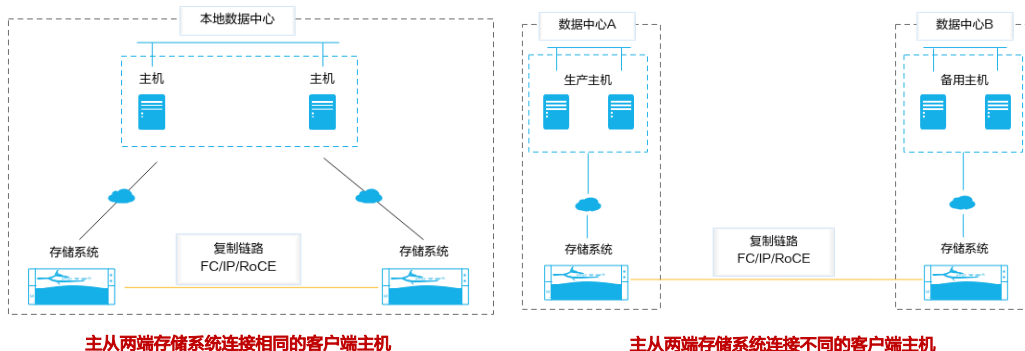
后台异步复制:

- 4、生产文件系统打快照,更新快照一致性状态。
- 5、灾备文件系统打快照,更新快照一致性状态。
- 6、生产快照增量同步到灾备文件系统。
- 7、数据同步完成。
- 8、根据快照保护策略,删除历史快照。

- HyperReplication文件系统异步远程复制采用基于对象层的方式进行数据复制,最短RPO支持15秒。
- 文件系统的所有内容,比如文件、文件属性、目录,都是由对象构成。
- 基于对象层的复制直接将对象从主文件系统复制到从文件系统,不需要关心复杂的文件层的信息,比如文件与目录间的依赖关系、各种文件操作,从而使复制变得更加简单高效。
- 支持高中低设备型号互通,在线升级时不用中断复制关系。

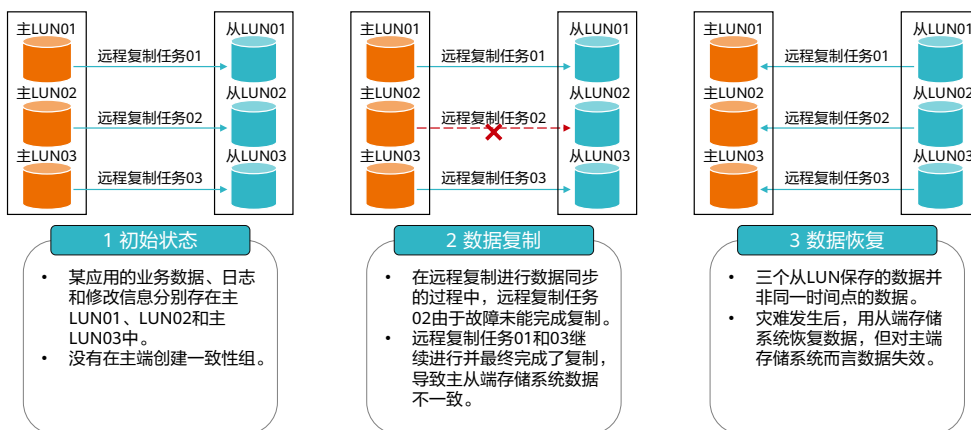
基于租户的远程复制

- 在使用远程复制功能时，用户期望能够将主端的配置信息同步到从端，当灾难发生时用户可以在从端直接使用该配置进行业务接管，进而缩短RTO。华为闪存存储提供基于租户粒度的远程复制，相比基于文件系统粒度的远程复制，租户粒度的远程复制将同步主端文件系统数据至从端，也会将租户及文件系统的相关配置同步至从端。其中，同步文件系统数据的原理同数据复制。



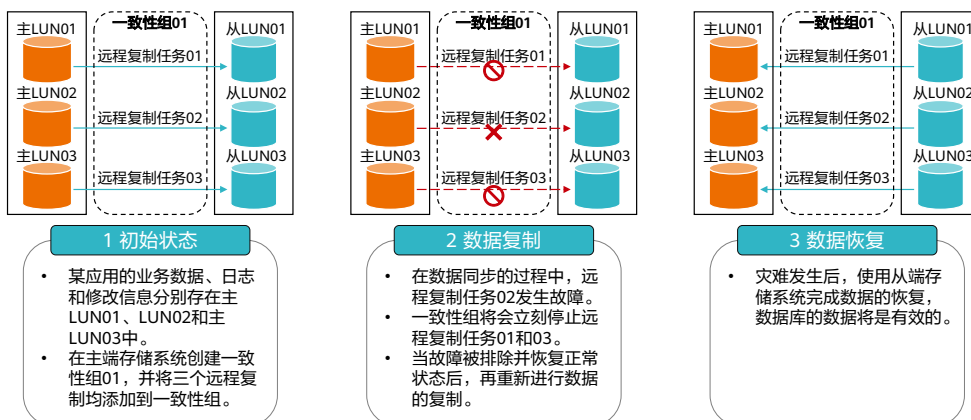
- 用户选择租户创建远程复制租户Pair，再选择租户Pair创建远程复制Pair，进而执行基于租户的远程复制。在创建远程复制租户Pair时，用户可根据实际部署场景选择是否同步共享、鉴权以及网络配置。按照主从端存储系统所服务的主机对象是否相同而言，我们可以将部署场景分为以下两种：
- 主从两端存储系统连接相同的客户端主机
 - 当主从两端存储系统部署在同一数据中心同一网段内，且主从存储系统可同时连接同一主机时，如果主端的配置信息可全部同步至从端，那么当生产中心发生故障时管理员可快速将业务切换至从端。对客户端主机来说共享路径等信息不变。
 - 针对该场景，用户在创建远程复制租户Pair时，需要同时开启“同步共享和鉴权”和“同步网络配置”。开启后，主端将向从端同步文件系统的共享权限、域控配置，以及租户的逻辑端口配置以及相关网络配置（如DNS Zone、路由等）。当生产中心故障时，通过启用从端存储系统的逻辑端口等简单操作，便可实现业务快速切换。
- 主从两端存储系统连接不同的客户端主机
 - 主从两端存储系统部署在不同的数据中心不同的网段，且主从存储系统连接至不同的主机（主端存储系统连接生产主机，从端存储系统连接备用主机）。
 - 针对该部署场景，根据主从存储系统是否加入域环境来判断远程复制过程中需要同步的配置项。

未使用一致性组的远程复制



说明：本页以LUN为例说明，文件系统同理。

使用一致性组的远程复制



说明: 本页以LUN为例说明, 文件系统同理。

Pair的运行状态

- 通过查看已有远程复制Pair的运行状态，对远程复制状态进行有效的监控。

运行状态	状态说明
正常	当Pair运行状态为“正常”时，主LUN和从LUN的数据成功同步完成。
分裂	当Pair运行状态为“分裂”时，主LUN和从LUN之间的数据复制暂停。由于业务需要，通过手动断开主LUN和从LUN之间的Pair关系，则Pair运行状态为“分裂”。
异常断开	远程复制所用的链路断开或远程复制主LUN或从LUN故障，导致主LUN和从LUN之间的Pair关系断开，则Pair运行状态为“异常断开”。
待恢复	Pair异常断开，在故障恢复以后，如果需要以“手动”策略恢复远程复制，则Pair运行状态被标识为“待恢复”，提示用户需手动进行同步操作，恢复原有主LUN和从LUN的Pair关系。
失效	Pair异常断开后，如果在主端或从端改变Pair属性（例如远程复制链路故障时，在主端或从端删除Pair），导致主从站点配置不一致，则Pair运行状态为“失效”。
正在同步	主LUN正在向从LUN同步数据，此时，从LUN不能被读写，如果此时发生灾难，此时的从LUN数据不能用于业务恢复。当从LUN数据状态为“完整”时，从LUN的数据可用于业务恢复。

说明：本页以LUN为例说明，文件系统同理。

取消远程复制Pair从资源保护

定义

指从LUN可以接收来自主机的数据。远程复制配置完成后对主机来说从LUN默认为只读，如果主LUN出现故障，需要管理员执行“取消远程复制Pair从资源保护”操作，手动设置从LUN为可写，以便从端存储系统能够接管主机业务，保证业务持续运行。

应用场景

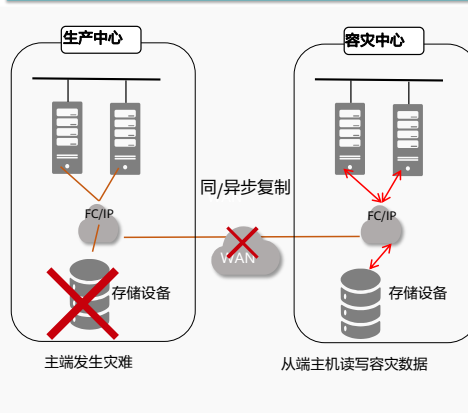
- > 用户需要使用从LUN进行数据分析、挖掘，但不愿意影响主LUN业务。
- > 当生产阵列故障，容灾阵列需要接管业务，但无法进行正常的主从切换。

优势

通过从LUN可写功能，可**加快业务恢复时间**，从LUN读写后可进行增量同步，使得灾难恢复后业务可迅速回切。

说明：本页以LUN为例，文件系统同理。

从LUN可写功能

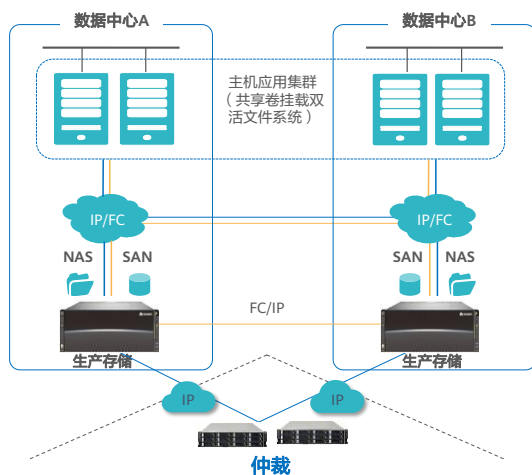


HyperMetro概述

- 华为HyperMetro特性又称双活特性（以下统称双活特性），双活特性中的两个数据中心互为备份，且都处于运行状态。当一个数据中心发生设备故障，甚至数据中心整体故障时，业务自动切换到另一个数据中心，解决了传统灾备业务无法自动切换的问题。提供给用户高级别的数据可靠性以及业务连续性的同时，提高存储系统的资源利用率。
- 华为提出以OceanStor为基础的存储双活架构，包括本地数据中心部署逻辑架构和跨数据中心部署逻辑架构，为客户建设业务不间断运行的解决方案。

- SAN双活解决方案采用数据双写技术，确保存储冗余，单存储或生产中心故障时，数据零丢失，业务快速切换，最大化保护客户业务的连续运行。满足业务RTO=0，RPO=0的业务需求。
- NAS双活AA（Active-Active）模式两个数据中心同时提供读写业务，业务负载均衡，充分利用灾备资源。任一站点故障时，数据零丢失，业务零中断。
- NAS双活AP（Active-Passive）模式任一站点故障时，数据零丢失，业务零中断，保障客户业务的连续运行。

SAN&NAS一体化双活



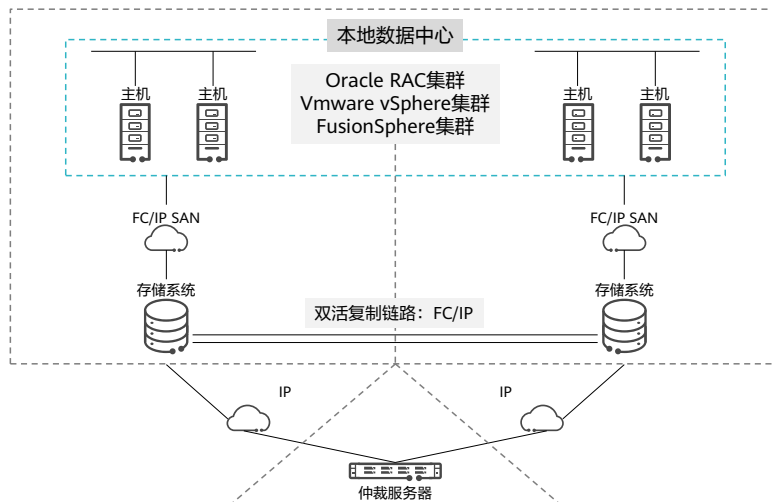
基本特征

- AA双活，RPO=0，RTO≈0。
- 设备双活，免网关；阵列间单网络，易部署。
- SAN&NAS一体化双活，共享仲裁。

HyperMetro SAN&NAS双活设计

- **架构：**跨阵列对称式AA双活架构，支持单LUN/FS 双端阵列都可以读写服务，数据实时同步至对端；
- **高可靠性设计：**双重仲裁机制，提高系统可靠性；
- **高性能设计：**
 - 跨站点通信协议级优化，
 - FC：FASTWRITE，跨站点数据交互减半
 - IP：RDMA通信能力，提升带宽降低时延
 - 跨站点数据同步：一次同步
 - 单FS访问性能
 - AA模式：相比本地单阵列性能提升30%
 - AP模式：相比本地单阵列性能下降10%

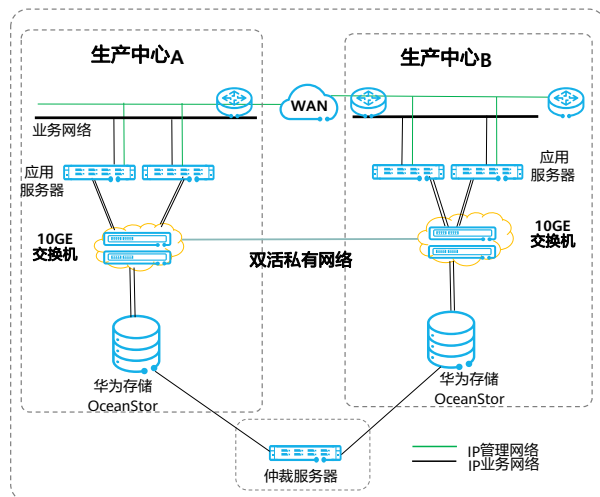
本地数据中心部署场景



- 本地数据中心部署

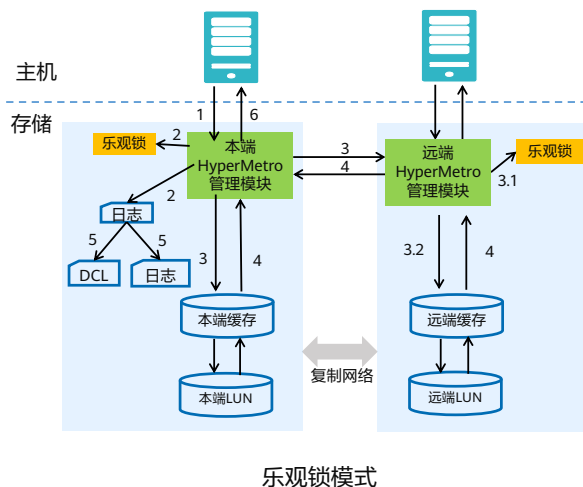
- 通常部署在同园区的不同机房中。
- 主机与存储间通常通过交换机方式通信，对于NAS文件系统，需要采用IP交换机。同时，存储两端通过双活复制网络连接，保证双活业务的运行。

跨数据中心部署模式



- 通常部署在同城或相近城市存在的两个数据中心，其物理距离在300km以内，两个数据中心均处于运行状态，可同时承担相同业务，提高数据中心的整体服务能力和系统资源利用率，当单数据中心故障时，业务自动切换到另一个数据中心。
- 对于跨数据中心场景，在长距离传输时，方案需要配套波分设备来保证网络传输时延。同时存储侧两端通过双活复制网络连接，保证双活业务的运行。

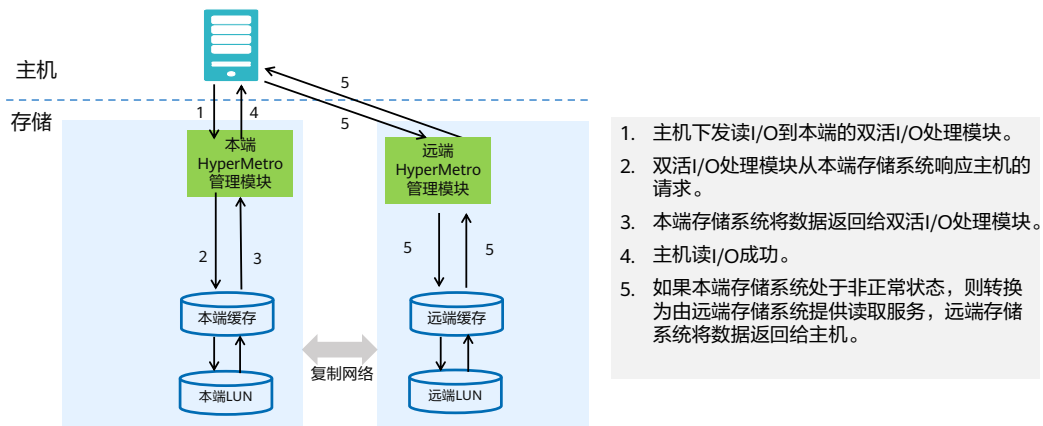
SAN: 写I/O流程



1. 主机下发写I/O到双活I/O处理模块。
2. 向本端乐观锁申请写权限。获得权限后，记录LOG。LOG中只记录地址信息，不记录具体的写数据内容。
3. 执行双写：双活I/O处理模块同时将数据写入本端LUN的Cache和远端LUN的Cache。远端收到写请求时，同样先向乐观锁申请写权限，获得写权限后将数据写入Cache。
4. 本端Cache和远端Cache向双活I/O处理模块返回写I/O结果。
5. 根据4的结果进行处理：
 - 如果两端存储系统都返回写成功，则清除LOG。
 - 如果任意一端返回写失败，则进行以下处理：
 - 将LOG转换成DCL，转换成功后清除LOG，记录本端LUN和远端LUN的差异数据。
 - 双活Pair关系断开，双活Pair的运行状态变为待同步。I/O变成单写，写成功的一端继续提供主机业务，写失败的一端停止主机业务。
6. 返回主机I/O响应成功。

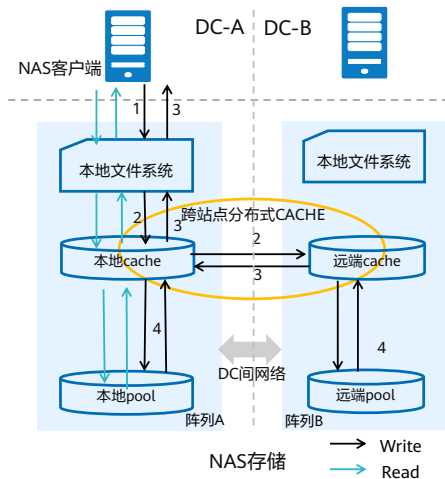
- 实现两台存储设备数据一致的核心就是高效的锁分配机制和高效的数据双写机制。
 - 在业务运行中，数据变更可通过双写和DCL来完成同步。双写可以使得应用服务器的I/O请求同时下发到本端Cache和远端Cache，从而保证本端Cache与远端Cache的变更数据一致性。在一个数据中心故障时，DCL可以记录业务运行中数据中心的数据变更。待故障恢复后，同时跟踪变更数据同步到该数据中心的存储设备中，以防止变更的数据在同步过程中出现遗漏。
 - 具备双活特性的两台存储系统，能够同时处理主机的I/O请求。为了防止两台存储系统同一时间收到主机写请求修改同一个数据块，而发生数据冲突，需要设计锁分配机制来避免冲突，只有获取锁分配机制允许的存储系统才能写入数据。没有获取该数据块锁分配机制允许的存储系统，需要等待锁分配机制释放之后才能重新获取锁分配机制允许写。

SAN: 读I/O流程



- 两端的LUN数据实时同步，且都能提供主机读写访问。当任何一端存储系统故障的情况下主机将切换访问路径到正常的一端继续业务访问。

NAS：跨站点分布式Cache高性能AA响应



双活写：

前端流程：

- 1、客户端写文件系统。
- 2、写本地cache，同时镜像到本地cache和远端cache。
- 3、所有写cache副本成功完成，回主机写IO成功。

后台流程：

- 4、本地cache控制发起cache刷盘；本地cache数据写入本地pool，远端cache写入远端pool。

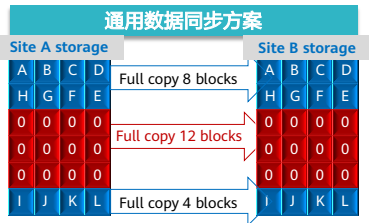
双活读：

- 1、客户端读文件系统。
- 2、本地cache读，如果命中则返回；不命中则下盘读数据返回。

优势：

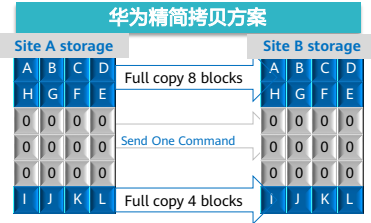
- 1、AB阵列支持同时访问同一个文件系统，由分布式cache实现一致性保证。
- 2、跨阵列之间数据同步，实时用户数据传输。
- 3、基于租户的双活配置，网络和鉴权配置自动同步。

快速初始/增量数据同步



全量拷贝的“非零”数据块
 全量拷贝的“零页面”数据块

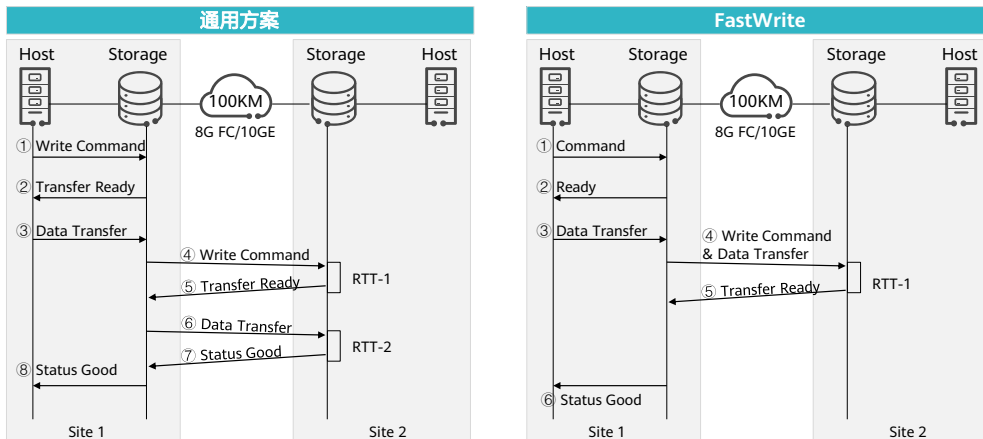
- **通用方案：**数据同步时，不识别“全零数据”，逐块拷贝所有数据
- 数据初始同步占用带宽高、传输时间长



“零拷贝”的“零页面”数据块

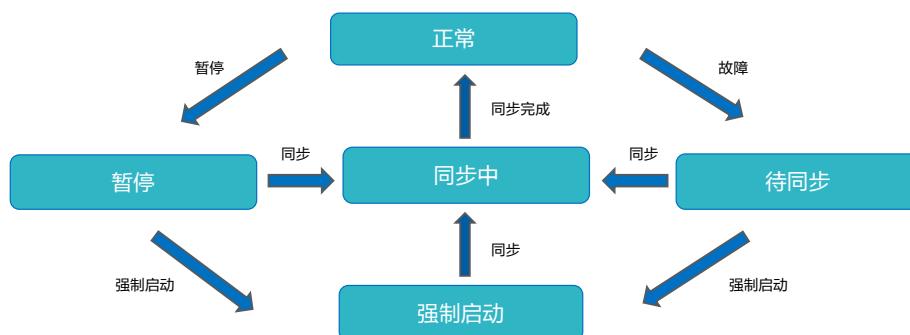
- **精简拷贝方案：**数据同步时，智能识别“全零数据”，只传输一个特殊标识，不传输数据
- 数据初始同步时间**减小90%**，数据初始同步链路**带宽占用减小90%**

FastWrite

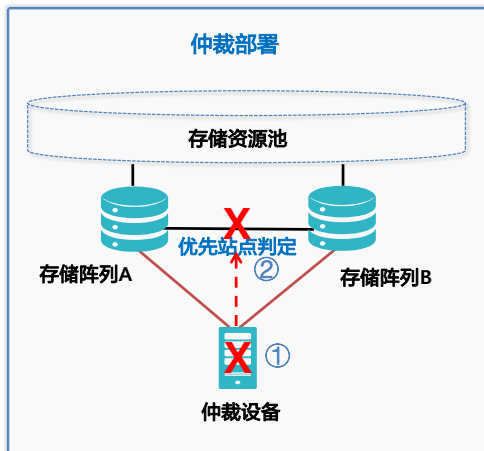


- 通用方案：写IO在传输的双端要经历“Write Command”和“Data Transfer”两次交互。
- FastWrite：通过对传输协议进行优化，将“写命令”和“写数据”合并为一次发送，跨站点写IO交互次数减少一半。

双活业务状态转换



双活仲裁机制



第一重：仲裁服务器模式

- **原理：**心跳不通时，仲裁服务器发起仲裁，仲裁获胜的一方继续提供服务，仲裁失败方停止服务，优先站点具有仲裁优先权

第二重：静态优先级模式

- **原理：**在未配置仲裁服务器或仲裁服务器无法访问时，双活仲裁模式为静态优先级模式。心跳不通时，预先设置的优先方存活。

- 当两个数据中心之间的链路故障或其中一个数据中心故障时，两个数据中心之间无法实时同步，此时只能由一端继续提供服务。为了保证数据一致性，双活通过仲裁机制决定数据中心的的服务优先级。
- 静态优先级模式
 - 用户在未配置仲裁服务器或仲裁服务器无法访问时，双活仲裁模式为静态优先级模式，双活Pair发生仲裁时优先站点端获得仲裁并提供业务。
- 单仲裁服务器
 - 单仲裁服务器模式下，一旦出现任一数据中心整体故障或阵列间链路故障时，阵列向仲裁服务器发起仲裁请求，由仲裁服务器综合判断哪端获胜。仲裁获胜的一方继续提供服务，另一方停止服务。单仲裁服务器模式下优先站点会优先获得仲裁胜利。
- 双仲裁服务器
 - 仲裁站点配置两台仲裁服务器时，如果主仲裁服务器到两端阵列之间所有链路断开，且备仲裁服务器与两端阵列连接正常、双活站点间链路也正常时，系统在主仲裁故障后1分钟内自动切换到备仲裁服务器，由备仲裁服务器来执行仲裁功能。

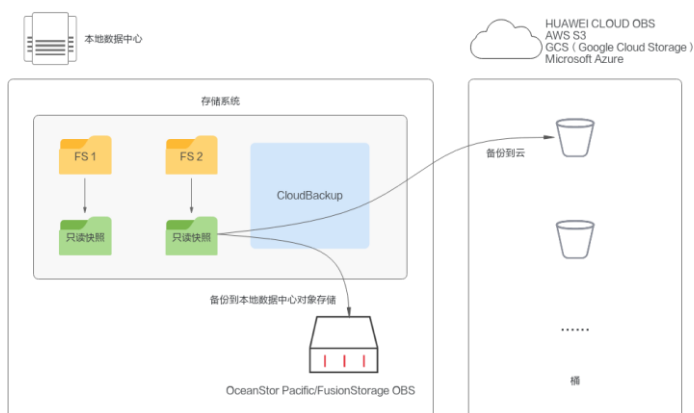
CloudBackup概述

- 为了避免数据中心存储设备因人为原因或物理故障而导致数据丢失或损坏，需要备份存储设备中的数据，并在需要时通过恢复找回数据。
- 云的普及使得越来越多的企业选择将数据直接备份到云端对象存储中，以减少备份存储的购置和维护成本，达到较高的性价比和可靠性。
- CloudBackup（云备份）特性是一种基于存储系统容器部署的数据保护技术，可以将存储侧文件系统中的数据备份到本地或者云端的目标存储，无需购买额外的备份服务器。
- 当存储侧文件系统中的数据损坏或丢失时，CloudBackup特性可以利用用户保存的指定时间点的备份副本恢复数据。

基本概念

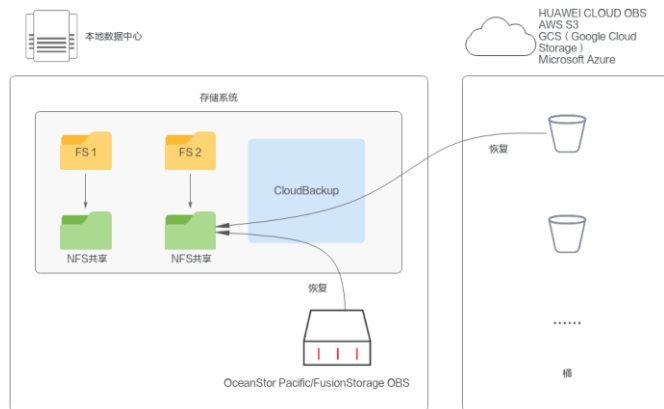
概念	说明
只读快照	源数据在某个时间点的一致性数据副本。
快照对比	两个只读快照之间的差异对比。在增量备份的时候，仅备份差异数据。
全量备份	将文件系统上的全部数据进行备份。第一次执行备份任务时，默认全量备份。 特点：备份的时间较长，占用空间较大。
增量备份	将本次备份与上次备份之间有变更的数据进行备份。 特点：每次备份的数据较少，备份时间较短。
备份副本	对文件系统进行一次备份所产生的数据集合称为备份副本。 一个文件系统会有多个不同时间点的备份副本。 用户可以根据时间戳选择指定的备份副本作为数据恢复的源头，恢复生产端丢失或损坏的数据。

备份原理



1. 为受保护对象创建只读快照。
2. 获取待备份的数据信息。
 - 扫描快照差异，计算出本次需要备份的变化数据。如果是第一次备份，则备份整个文件系统（全量备份）。
3. 读取数据到备份存储。
 - 根据2获取待备份的数据信息，为快照创建共享并挂载，读取待备份的数据和元数据，传输到数据中心的对象存储或者远端公有云的对象存储中，实现数据备份保护。

恢复原理



1. 在本地文件系统中选择需要恢复的文件系统，查看备份副本数据。
2. 选择需要恢复的备份副本。根据部署场景不同，该备份副本可以是远端公有云上的备份副本，也可以是本地数据中心对象存储中的备份副本。
3. 选择恢复方式：整副本恢复或者文件级恢复。
4. 选择恢复目标文件系统：原位置恢复或者新位置恢复。
5. 执行恢复。

恢复的过程中，CloudBackup会从对象存储中，将指定备份副本数据读出，并写到恢复目标文件系统的共享中。

HyperDetect概述

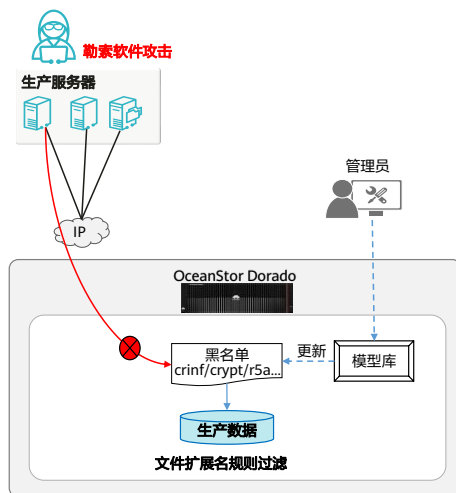
- 勒索软件（又称勒索病毒）已成为全球互联网的高级持续威胁之一，每年造成数千亿美元损失。勒索软件本质是一种特殊的恶意软件，通过对数据恶意加密/窃取，导致计算机或特定LUN/文件不可用或不可读，只有受害人支付了赎金，才有可能获取用于恢复数据的密钥。因此，在存储系统构建防勒索能力，贴身保护数据，具有重要的意义。
- HyperDetect特性是一种基于存储系统容器部署的数据保护技术，可以提供勒索文件拦截、实时勒索检测、智能勒索检测的勒索软件检测能力。

目的和受益	详细说明
勒索文件拦截	勒索软件发起攻击时，通常会对正常文件进行加密，产生带有特殊扩展名的加密文件。通过禁止该类型文件的写入，拦截已知勒索软件的勒索行为，保护存储系统中的文件系统。
实时勒索检测	勒索软件的攻击行为具有相似的I/O行为特征，通过分析LUN或文件的I/O行为，快速筛选出异常文件，并对异常LUN或文件进行内容深度分析，检测出被勒索的LUN或文件。然后对相应的LUN或文件系统创建安全快照并发送告警，提醒数据保护管理员及时介入，避免勒索软件影响进一步扩散，减少损失。
智能勒索检测	通过对已知勒索软件攻击特征的检测，识别出被勒索软件感染的LUN或文件系统。对未识别出勒索感染的LUN或文件系统，通过分析对比LUN或文件系统快照变化特征，使用机器学习算法，进一步检测LUN或文件系统是否被勒索软件感染。

基本概念

概念	说明
文件扩展名过滤规则	文件扩展名即文件后缀。每条文件扩展名过滤规则对应一种文件扩展名，用于阻止用户对该类型的文件的操作。
检测模型	勒索软件检测的模型。包含已知勒索软件特征和机器学习模型。
I/O行为	用户对文件的创建、读、写、删除和重命名等行为；对LUN的读、写行为。
块	SAN存储中的数据单元。

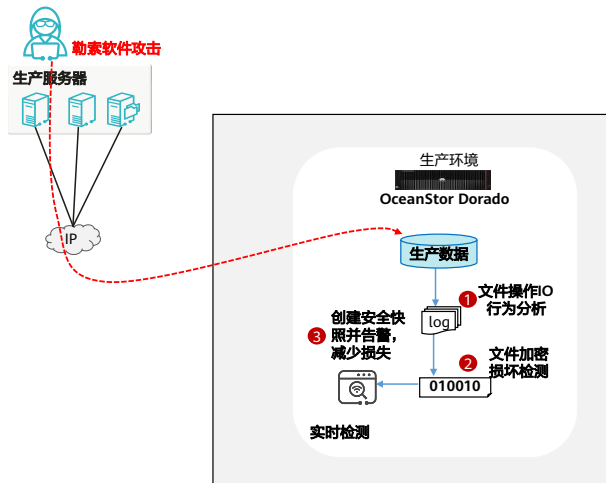
勒索文件拦截



勒索软件发起攻击时，通常会对正常文件进行加密，产生带有特殊扩展名的加密文件，通过禁止该类型文件的写入，拦截已知勒索软件的勒索行为，保护存储系统中的文件系统。

- 系统预置了常用的检测模型，检测模型中预置了已知的勒索文件扩展名。
- 可以通过导入检测模型来更新勒索文件扩展名，还可以通过创建文件扩展名过滤规则来新增要拦截的勒索文件扩展名。
- 用户可以根据业务需求选择租户并更新文件扩展名过滤规则，租户下的所有文件系统根据设置的文件扩展名过滤规则，检测待操作的文件扩展名是否和过滤规则相同，若相同，则对勒索文件进行拦截。

实时勒索检测

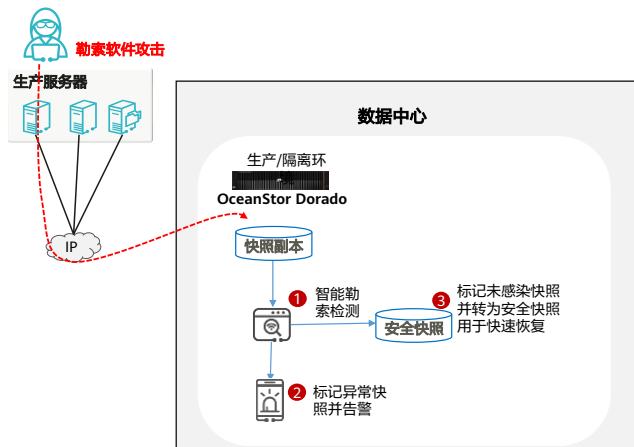


勒索软件的攻击行为具有相似的I/O行为特征，通过①分析文件I/O行为，快速筛选出异常文件，②并对异常文件进行加密损坏检测，检测出被勒索的文件。③然后对相应的文件系统创建安全快照并发送告警，提醒数据保护管理员及时介入，避免勒索软件影响进一步扩散，减少损失。

说明：以文件业务场景为例，块业务场景同理

- 异常I/O行为指勒索软件的I/O行为。实时勒索检测主要通过异常I/O行为快速筛查和文件损坏深度检测结合，检测数据是否被勒索软件勒索。
- 异常I/O行为快速筛查：通过分析存储系统的I/O行为，识别出异常I/O行为。将异常I/O行为操作对应的文件进行文件损坏深度检测。
- 文件损坏深度检测：通过文件内容基本特征检测，识别出已损坏的文件，对于未识别出基本特征异常的文件，基于机器学习算法对文件内容进行深度检测，进一步检测文件是否被损坏。
- 创建安全快照并发送告警：对已损坏的文件所在的文件系统创建安全快照并发送告警，提醒数据保护管理员及时介入，避免勒索软件影响进一步扩散，减少损失。

智能勒索检测（文件业务）



- 智能勒索检测包括已知勒索软件攻击特征检测和前后快照数据变化行为的检测。

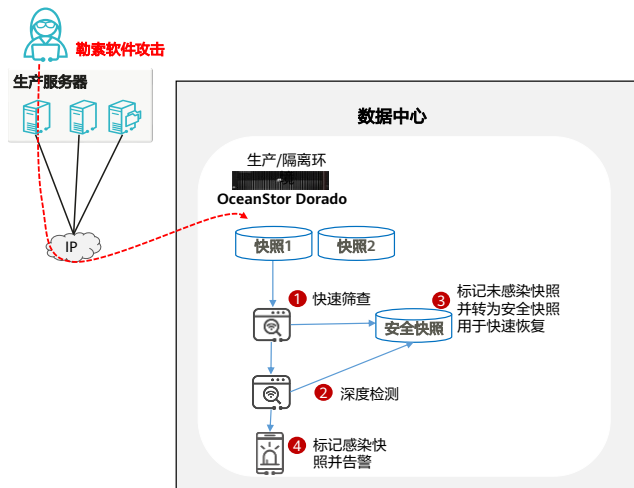
1. 已知勒索软件攻击特征检测

- 勒索软件通常有两种常见行为
 - 篡改原文件的扩展名并加密。
 - 留下勒索提示信息文件（如：勒索邮件等）。
- 对于已知的勒索软件，可以通过扩展名和勒索邮件两种类型的特征来发现快照是否被勒索软件感染。

2. 前后快照数据变化行为的检测

- 基于快照的元数据特征，快速提取若干基础变化特征，判断待检测快照的变化是否可疑。若被判断为可疑快照，再进一步提取可疑快照的全量变化特征，使用机器学习算法检测快照是否被勒索软件感染。

智能勒索检测（块业务）



- 智能勒索检测首先基于快照的元数据特征和分区信息，使用快照比对获取快照间的差异数据位置，对差异数据进行采样，并快速提取若干数据特征，根据基线模型判断当前快照的变化是否可疑。若被判断为疑似感染，再进一步分析可疑快照内采样数据块周围的连续变化数据，基于数据内容提取多种的特征，利用特征融合技术，结合机器学习模型进行检测。智能勒索检测主要通过分区采样变化数据快速筛查和分区连续变化数据深度检测结合，检测快照中的数据是否被勒索病毒加密

1. 快速筛查分区采样变化数据

- 通过勒索加密分布模型筛选加密数据块，初步判断快照LUN中分区数据是否被勒索病毒加密。
- 将检出的异常分区对应的连续加密数据块进行块数据深度检测。

2. 深度检测分区连续变化数据

- 通过对分区中加密块所对应的连续变化数据内容提取多种特征，并利用特征融合技术将数据特征融合，基于深度学习算法对分区连续数据进行深度检测，检测分区数据是否被勒索病毒加密。

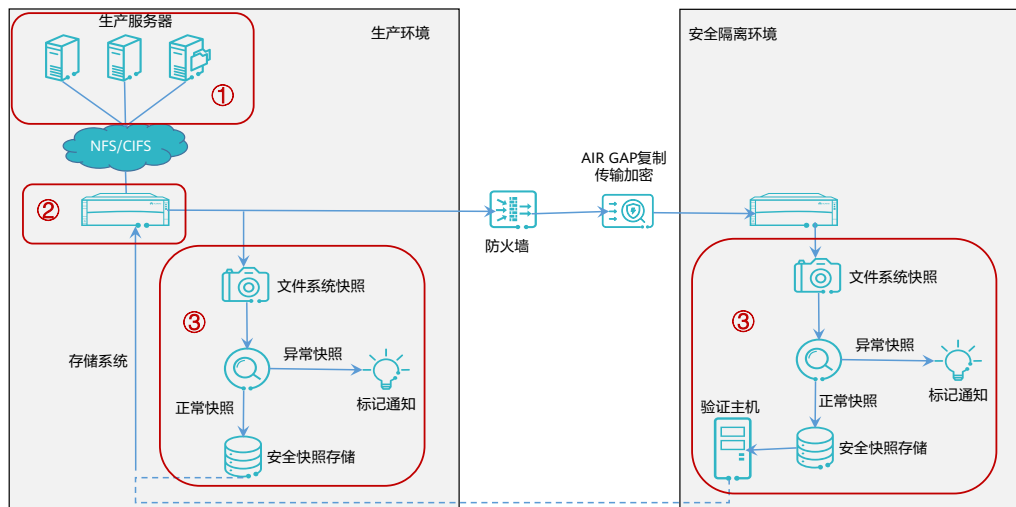
3. 将未感染的快照转换成安全快照

- 若快照检测为未感染，则将该快照转换成安全快照，用于后续恢复。

4. 发送告警

- 若快照检测为感染则发送告警，提醒数据保护管理员及时介入，避免勒索病毒影响进一步扩散，减少损失。

应用场景



- 在生产环境中：
 - ① 通过对特定扩展名文件的写入阻止，拦截已知的勒索软件对数据的篡改。
 - ② 通过异常I/O行为快速筛查和文件损坏深度检测结合，及时发现勒索软件勒索行为，标记数据感染的时间点，从而确保数据的丢失量最小。
 - ③ 通过已知勒索特征检测和前后快照数据变化行为的检测，标记异常快照。并可以将正常快照转换为安全快照，防止被篡改，最后生成检测报告。
- 在隔离环境中：通过已知勒索特征检测和前后快照数据变化行为的检测，标记异常快照。并可以将正常快照转换为安全快照，防止被篡改，最后生成检测报告。
- AIR GAP通过对复制链路自动关断控制，将数据复制到隔离区，形成更安全保护效果。

思考题

1. （多选题）华为 OceanStor Dorado全闪存系列存储的数据保护特性包括以下哪些？
 - A. HyperReplication
 - B. HyperMetro
 - C. CloudBackup
 - D. HyperDetect
2. （单选题）RAID-TP可容忍几块盘同时失效？
 - A. 1
 - B. 2
 - C. 3
3. （判断题）ROW满分条写对所有数据都采用新写模式，这样不需要因为传统RAID写流程所需的数据读和校验修改写而产生RAID写惩罚，有效降低了写入过程阵列控制器的CPU开销与对SSD盘的读写压力。

- 参考答案：

- ABCD
- C
- T

思考题

4. （判断题）存储系统提供初始化向导，可以完成设备管理IP配置。
5. （判断题）设备上电遵循先给机柜上电，再给机柜内设备上电的上电顺序。
6. （多选题）存储系统支持哪些告警处理策略？
 - A. 邮件通知
 - B. syslog通知
 - C. 短消息通知
 - D. Trap方式通知

- 参考答案：

4. F
5. T
6. ABCD

本章总结

- 本章主要介绍华为闪存存储产品体系、系统架构、典型应用场景、安装部署以及增值特性。

更多信息

- 华为数据存储信息中心: <https://info.support.huawei.com/storage/#/home>
- 企业业务: <https://e.huawei.com/cn/>
- 技术支持: <https://support.huawei.com/enterprise/>
- 在线学习: <https://e.huawei.com/cn/talent/portal/#/>
- 华为Support案例库: <https://support.huawei.com/enterprise/zh/knowledge?lang=zh>

Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home, and
organization for a fully connected,
intelligent world.

版权所有 © 华为技术有限公司 2026。

保留一切权利。

未经华为事先书面许可，任何人不得对本资料进行复制、修改、
整合、改编、翻译，也不得将本资料或其任何部分或基于本资
料的衍生作品提供、出售、转让、再许可给任何人用于华为
许可以外的任何其他用途。

Copyright © Huawei Technologies Co., Ltd. 2026.

All Rights Reserved.

Without prior written permission of Huawei, no one is
allowed to reproduce, modify, integrate, adapt, translate
this Document, or provide, sell, transfer, sublicense this
Document or any part of this Document or its derivative to
any third party for any other purpose otherwise permitted
by Huawei.



数据保护产品技术



前言

- 随着数据的数量、种类和增长速度呈现指数级变化，同时，由于人为错误、病毒、自然灾害以及其他网络安全威胁等原因，企业面临越来越多的数据丢失的风险，数据保护的重要性日益明显。华为数据保护产品为各种复杂的操作系统、数据库、虚拟化平台等提供更高效的数据保护，帮助用户实现更高的SLA要求。

目标

- 学完本课程后，您将能够：
 - 熟悉数据保护产品及应用场景
 - 熟悉数据保护产品部署

目录

1. 数据保护产品介绍
2. 数据保护产品应用场景
3. 数据保护产品部署

当我们说数据保护时，到底在说什么...

业务连续性



容灾

数据可靠性



备份

法规遵从性



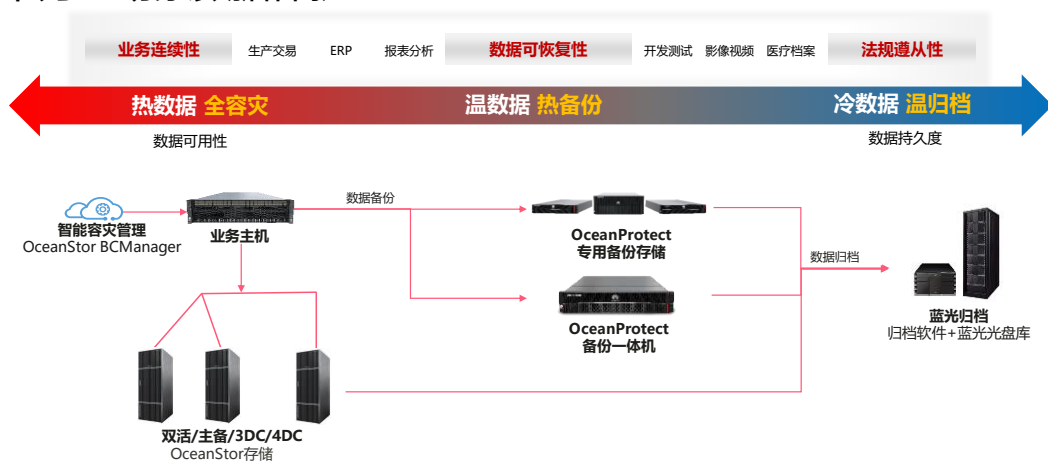
归档

在灾难未发生前利用IT技术对信息系统的数据和应用程序进行保护

应用场景不断演进，数据保护要求越来越高

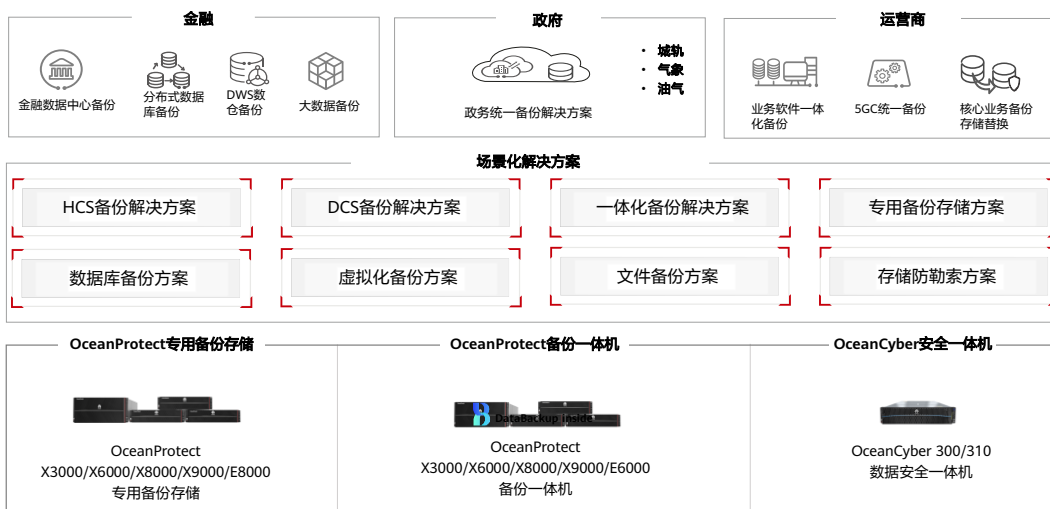


华为全场景数据保护



- 随着业务应用场景不断演进，数据保护的等级、保护对象、保护性能和保护规模都出现新的要求，围绕数据生命周期提供全面保护变得越来越重要。华为 OceanProtect 数据保护，围绕数据的全生命周期，实现热数据全容灾、温数据热备份、冷数据温归档和全场景智能融合，保证客户业务不中断、数据不丢失、信息长期留存，为智能世界保驾护航。

华为数据保护产品解决方案



- OceanProtect专用备份存储基于端到端加速和 Active-Active 高可靠架构设计，具有备份快、恢复快、用量省、更安全的特点，可帮助用户实现高效备份恢复，大幅节省 TCO，广泛适用于政府、金融、运营商、医疗、制造等行业。
- OceanProtect备份一体机，基于专用备份存储底座AA高可靠存储架构，内置备份软件特性，将备份软件、硬件底座和防勒索三合一，实现备份恢复全流程端到端加速。这款备份设备具有备份快、高安全、生态优的特点，帮助用户实现数据高效备份，快速利用，节省数据保护投资，广泛适用于政府、金融、运营商、医疗和制造等行业。
- OceanCyber数据安全一体机作为数据存储的安全引擎，为多类型存储设备提供安全策略配置管理、侦测分析及防御能力，助力客户基于存储这一数据的最终载体，部署存储防勒索能力，构筑数据安全最后一道防线。
- OceanProtect DataBackup是面向新一代企业数据中心和多云环境的数据保护软件，秉承“极致保护 + 高效利用”的设计理念，帮助客户构建边缘、中心、多云的线上线下高效协同的全场景数据保护方案。系统提供全面的生态保护能力、面向多云的备份和服务能力、强韧性的数据安全防护能力，通过统一管理平台，实现高效、易用的数据备份恢复管理、副本利用和数据安全检测。帮助用户有效降低数据安全风险，提高企业对于备份系统的投资回报。

目录

1. 数据保护产品介绍

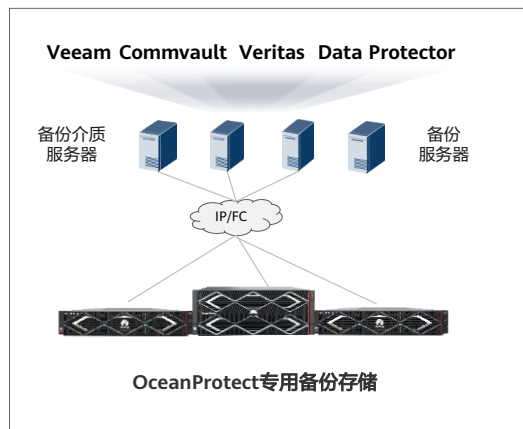
▪ OceanProtect专用备份存储

- OceanProtect备份一体机
- OceanCyber数据安全一体机

2. 数据保护产品应用场景

3. 数据保护产品部署

OceanProtect专用备份存储产品简介



备份快、恢复快

最高**172TB/h**恢复速度
保障数据即时可用



用量省

以**先进算法**和**数据缩减卡**实现数据缩减，
大幅提升有效容量；以**先进介质**，缩减空间和
降低能耗，节省系统投资



更安全

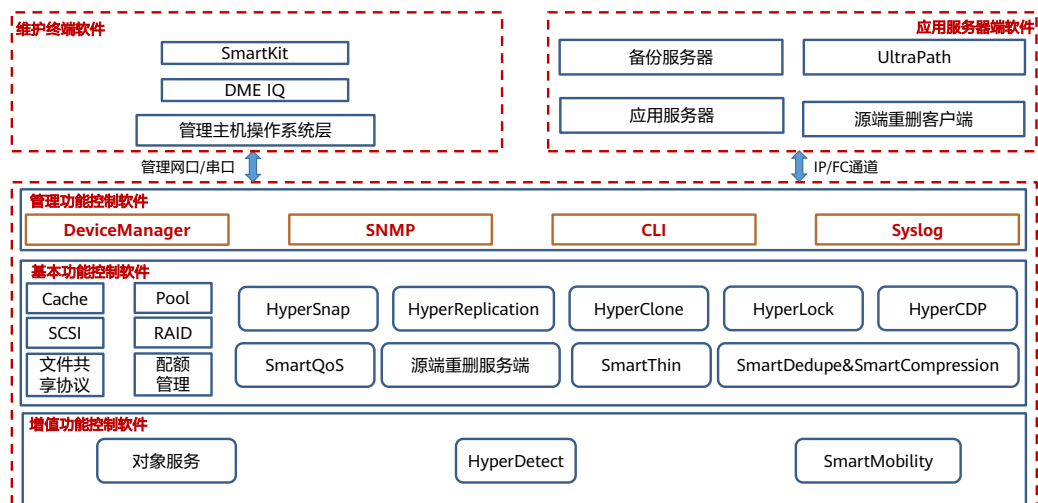
业界领先**防勒索方案**，业界独家**A-A架构**
保障副本数据安全可靠

- 面向集中备份场景，OceanProtect专用备份存储与业界主流备份软件形成数据保护解决方案。具有备份快、恢复快、用量省、更安全的特点，可帮助用户实现高效备份恢复，大幅节省TCO（Total Cost of Ownership），广泛适用于政府、金融、运营商、医疗、制造等行业。除此之外，还提供易于使用的管理方式和方便快捷的本地/远程维护方式，大大降低了设备管理和维护的成本。
- OceanProtect专用备份存储包含两种形态：
 - 全闪形态：系统只支持SSD（Solid-State Drive），业务数据和元数据都存放在SSD中。
 - HDD形态：系统支持HDD（Hard Disk Drive）+SSD，业务数据存放在HDD盘组成的存储池中，SSD仅用于存放系统元数据，通过SSD性能优势加速系统的元数据访问，进而提升读写性能。

硬件架构

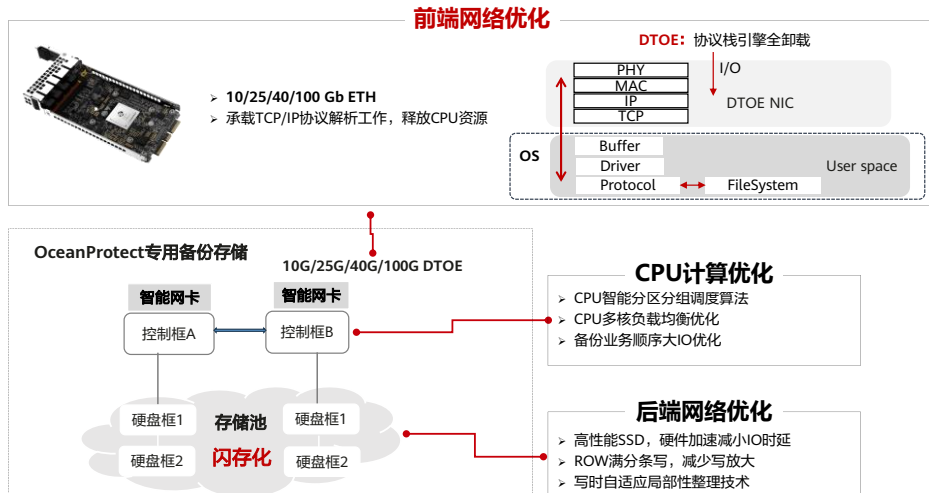
- 3D体验中心提供全新的硬件3D仿真交互，支持全方位浏览硬件模型并手动拆卸部件，了解各模块功能，并提供产品关联的文档、工具、多媒体、软件和成功案例等资源。您可以通过如下方式进行访问：
- 1. 访问华为数据存储信息中心（<https://info.support.huawei.com/storageinfo/#/home>）。
- 2. 在首页菜单“信息速查”页签，单击“3D体验中心”。
- 3. 选择“数据保护”下的产品类型。
- 4. 单击对应的产品型号。
- 5. 选择所需查看的物理设备，进入展示页面。

OceanProtect专用备份存储功能架构



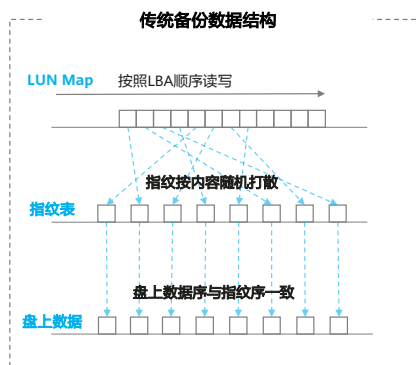
- 存储系统端软件采用专用操作系统，实现硬件管理和支撑存储业务软件的运行。
- 维护终端软件用于系统配置与维护。用户可以通过维护终端的SmartKit、DME IQ等软件对系统进行配置和维护。
- 在SAN网络下，应用端软件可与存储系统通信，从而使双方能够配合执行某些操作。
- 应用端包括备份服务器（通过备份软件备份数据到OceanProtect专用备份存储）、应用服务器（应用服务器通过自带的备份工具备份数据到OceanProtect专用备份存储）、源端重删客户端（与源端重删服务端软件模块配合，在生产端实现重复数据删除）和UltraPath（是一款与存储系统配套的基于应用服务器的存储设备驱动程序）。

基于闪存端到端加速

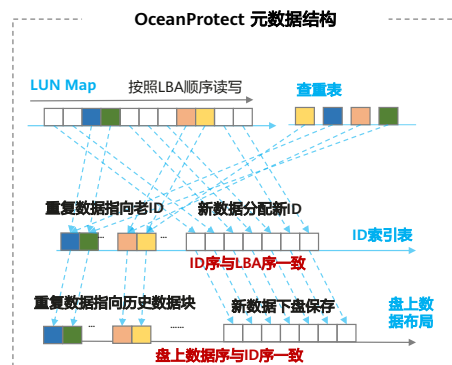


- OceanProtect专用备份存储通过从前端网络、控制器、后端网络等全I/O路径的软件优化, 充分利用强大的硬件能力为客户提供极致的高带宽。
- 前端网络优化: OceanProtect专用备份存储的DTOE技术, 将IP数据路径卸载到网卡, 传输层的处理包括DIF校验功能下移到华为定制的网卡微码里, 彻底消除了CPU开销; 同时, 系统中的工作线程, 采用的轮询的方式, 周期性去检查接口模块的接收队列, 有请求就立即处理。减少了请求来了以后, 再去唤醒工作线程带来的时延开销。
- 控制器内优化: OceanProtect专用备份存储采用了业界最多的CPU个数和CPU核数, 提供了强劲的运算能力。通过智能众核技术, 降低了跨CPU访问等一系列多核、多CPU架构带来的性能下降等负面影响, 充分发挥了CPU的强劲的运算能力, 实现了系统性能随着CPU数量线性增加。智能众核技术, 由CPU分组技术、业务分组技术、核间免锁技术组成。
- 后端网络优化: OceanProtect专用备份存储实现了三类冷热数据分流: 元数据、新写的数据、垃圾回收搬移的有效数据。元数据最热, 新写的数据中, 如果长时间没有改写的, 那么就会被垃圾回收搬移, 这些被搬移的数据就是改动概率最小, 最冷的数据。通过三类分流, 减少了SSD盘写放大, 极大提升GC回收效率, 减少对SSD颗粒的写数据量, 提升系统的写性能。延长SSD盘的寿命。OceanProtect专用备份存储通过对所有的写 (包括新写数据和对老数据的改写等) 都采用ROW (Redirect On Write) 大块顺序写的方式, 避免了传统RAID写流程所需的数据读和校验修改写而产生RAID写惩罚, 有效降低了写入过程阵列控制器的CPU开销与对SSD盘的读写压力。同时通过ROW技术每次写通过分配新的Flash颗粒来写, 实现每个Flash颗粒擦写次数均衡, 避免对固定地址的Flash颗粒反复擦写而耗尽Flash颗粒的寿命导致SSD快速故障。

三层元数据结构



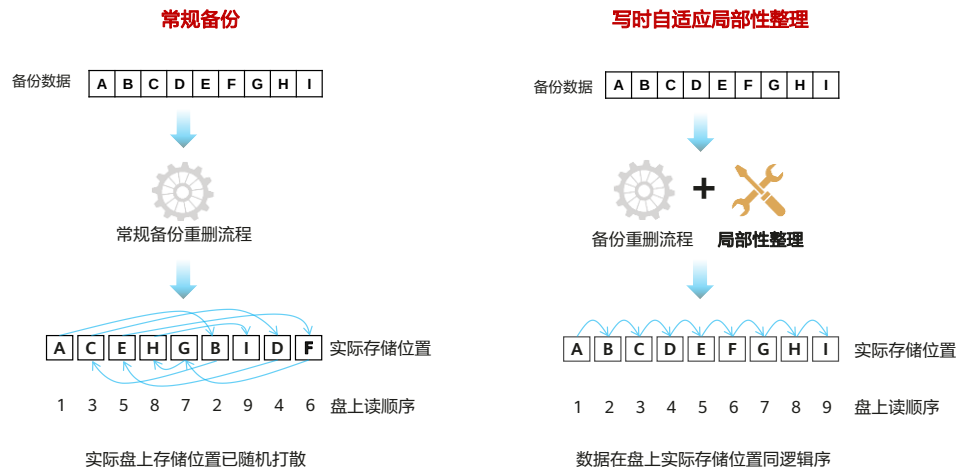
以指纹表为中心进行数据布局，数据经过重删后，会打散IO的顺序性，在HDD环境下随机访问性能低，影响恢复性能。



首创以ID索引表为中心的三层元数据结构，减少HDD的随机访问，提升HDD环境下的恢复性能。

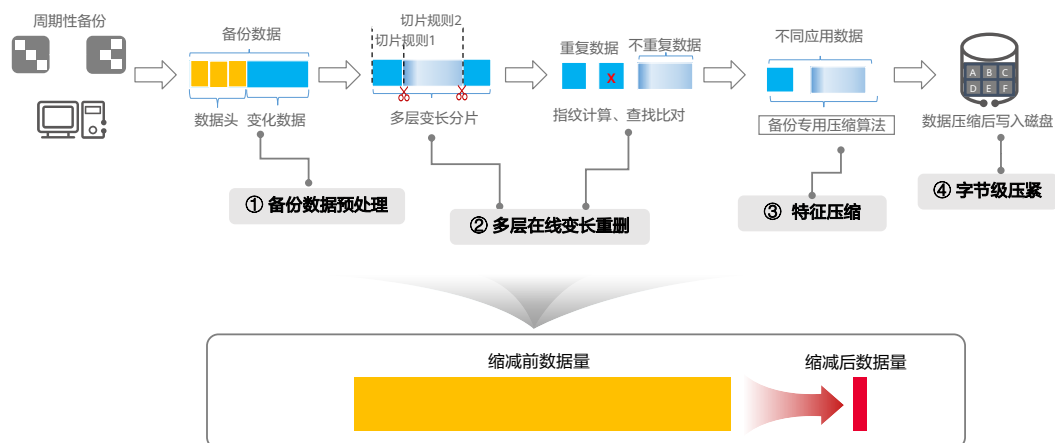
- 备份任务一般情况下是按照备份任务生成的文件的逻辑地址顺序写入，备份恢复时大部分的IO都是按照文件逻辑地址顺序读取，读数据的顺序与写数据的顺序基本一致。OceanProtect专用备份存储针对备份软件的这个特性专门设计了一套数据布局架构。
- 传统的重删架构以数据块的指纹表为中心进行数据布局，顺序的数据经过重删后，会按照指纹进行打散存放，盘上数据按照指纹归属存放，与LBA顺序无关。扰乱了业务数据的顺序性。这种方式适合全SSD存储，但是在HDD为存储介质的存储上，HDD盘随机访问的性能差，将会影响写数据的重删查找和恢复性能。
- OceanProtect专用备份存储针对备份任务执行时一般是按照生成备份文件的逻辑地址顺序写，备份任务恢复时一般按照备份文件的逻辑地址顺序读的特点，结合HDD盘顺序读写性能高，随机读写性能低的特点。专门设计了以ID为中心的三层元数据结构，解决重删扰乱数据布局的顺序性导致HDD盘乱序读盘性能差的问题。

写时自适应局部整理



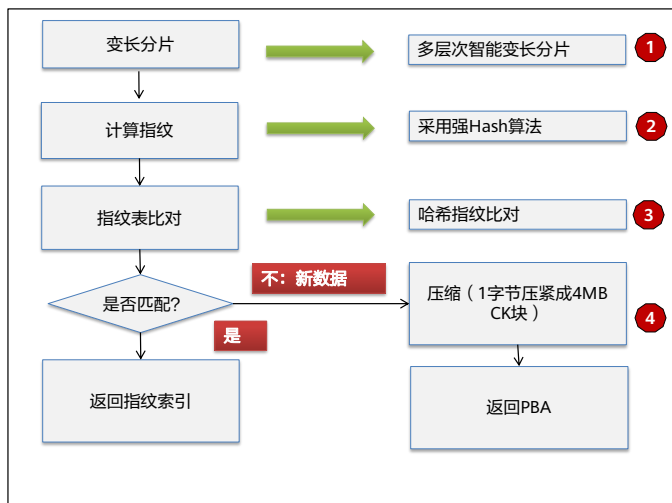
- 备份系统随着备份次数的增加以及增量备份与全量备份时的重删关系，会使后写入的备份数据的重删数据的物理排布越来越离散，导致恢复读性能会逐渐变差；其次备份恢复的使用场景决定了最新写入的备份数据用于恢复的概率最高，但是其恢复性能因为在HDD盘上布局的离散性大往往最差，导致用户在遇到数据丢失或突发状况时的恢复时间往往无法接受。为构筑OceanProtect的性能优势，打造用户数据的即时恢复能力，华为创新研发出针对备份数据恢复性能的局部性整理算法。
- 由于最新的备份数据往往会重删在早期多次的备份数据上，所以在恢复最新一次备份时需要多次随机读取老备份上的数据，引入大量的随机读操作，从而导致HDD的读性能较低。局部性整理算法的核心在于减少备份恢复时的随机读操作。

数据缩减技术全景



- SmartDedupe是重复数据删除技术，完全相同的数据块在存储时只保留一份数据块，其他数据直接引用同一份相同的数据块，以此减少数据的存储空间。
- SmartCompression是数据压缩技术，数据在写入存储介质前先经过无损压缩算法进行压缩后把压缩后的数据块写入存储介质，以此减少数据的存储空间。

数据缩减 - 在线重删压缩

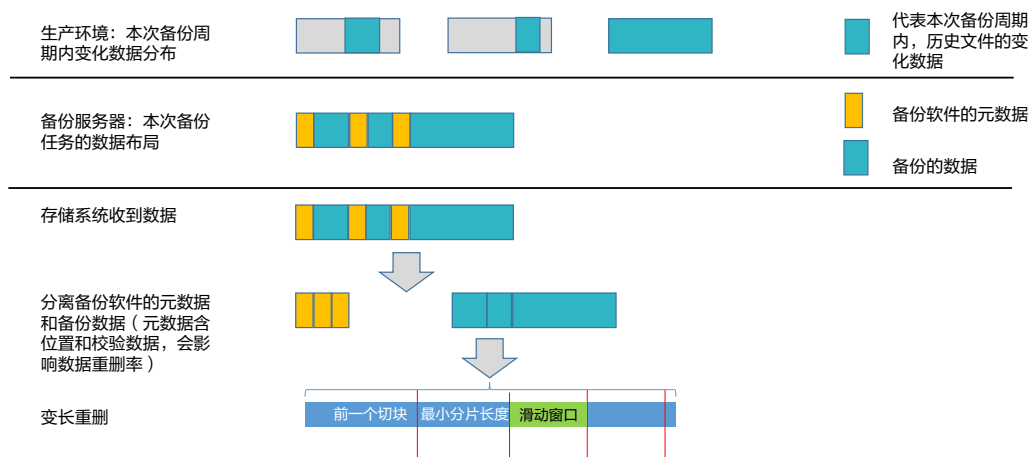


实现原理

- ① 用户写入数据时，先进行变长分片处理：系统根据多层次智能变长分片的算法，对写入的数据进行指纹特征计算，根据计算结果对数据块进行变长分片，把连续的数据块切分成不等长的数据。
- ② 数据重删流程，经过变长分片切分后的数据块经过SHA1算法的计算生成指纹。
- ③ 用哈希指纹进行指纹比对，如果在系统的指纹库中找到完全一样的指纹，说明本次写入的数据块是重复数据，更新数据应用计数后返回已经存在的指纹索引。
- ④ 如果哈希指纹在指纹数据库中找不到完全一致的，证明本次写入的数据块是新数据。进入后续的压缩流程+写盘流程。

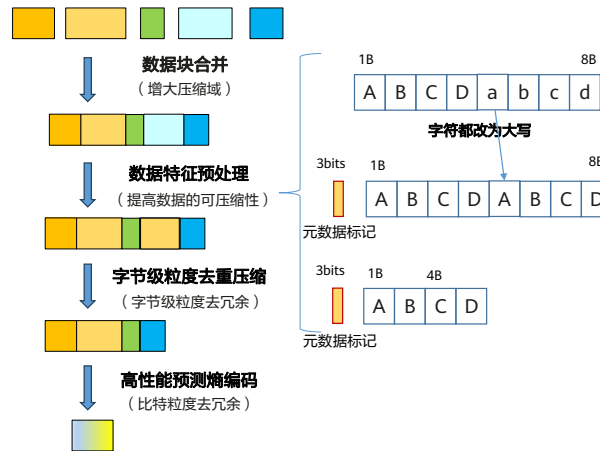
- OceanProtect专用备份存储的SmartDedupe&SmartCompression特性采用先重删再压缩的设计。重删采用在线变长重删技术，变长重删分片大小为4 KB~32 KB，压缩采用华为专有的高效压缩算法+压缩后数据字节级压紧存储技术。

数据缩减 - 写入数据预处理+变长重删



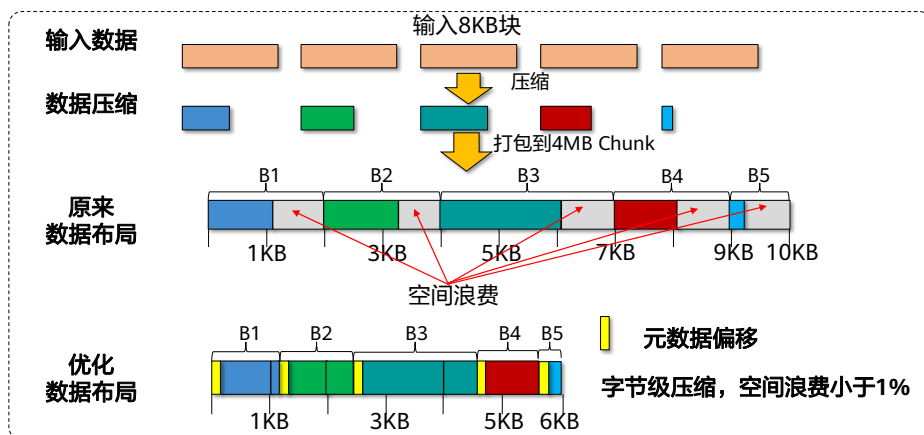
- 备份软件写入的数据一般包含元数据和客户备份数据。部分备份软件会把元数据和备份数据合并在一个文件中，元数据散布到整个文件中。元数据包含时间戳、校验等随机数据，因此元数据本身的重删率很低。如果把元数据和客户备份数据放在一起进行重删会降低整体数据的重删效果。
- OceanProtect专用备份存储针对备份软件的备份数据和元数据布局做了特征识别处理，把备份软件插入的元数据和备份数据分离，用户数据做聚合后进行变长分片重删，元数据单独做压缩处理。
- 传统的存储系统将用户文件按照固定长度分片后存储，在备份场景下这会导致严重的问题，因为一旦用户文件中插入或者删除一部分数据之后，再次定长分片会得到大段和修改前有极小差异的定长分片，导致修改前和修改后的数据有大量的比例无法重删。OceanProtect专用备份存储采用变长分片技术来解决这类问题，变长分片技术通过计算数据流的特征，根据特征来决定切点的位置（切分后得到的数据块长度不一致，范围在4KB~32KB之间），当用户插入或者删除一部分数据后，仅仅在插入或者删除位置附近极小范围的位置，数据流特征有变化，其他位置数据特征会保持不变，这样修改前后的数据的特征差异程度仅仅和插入或者删除的数据量有关，没有发生变化的地方将会生成完全一样的变长分片，最终被存储系统重删掉，从而实现较高的重删率。
- OceanProtect专用备份存储采用的是基于内容定义的分块算法，主要用于解决数据偏移导致重删率低的问题，如图展示基于内容定义分块（CDC）的基本原理，通过使用滑窗内容判断是否应该切块，从而达到按内容定义切点的目的。

数据缩减 - 备份数据特征压缩



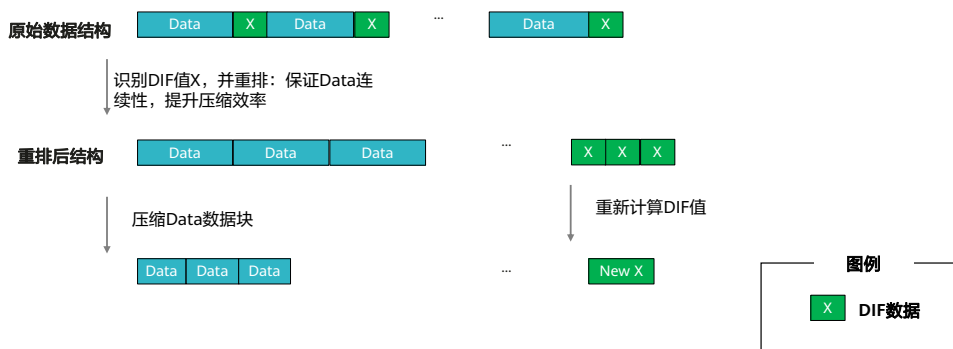
- 数据压缩主要包括两个过程，首先是采用压缩算法对输入的数据块进行压缩，形成一个个的小数据块，然后再把这些压缩后的数据块拼接在一起（数据压紧，Data Compaction）下盘。
- OceanProtect专用备份存储采用合并压缩技术，系统将连续的数据块合并成大的数据块（128KB）后一次压缩。利用备份软件顺序写，恢复时同样顺序读的特点，ID连续的数据块合并一起压缩，在执行备份恢复时，一次读取后解压得到的ID连续的数据块会在读cache被后续备份软件发起的连续的读请求命中，减少读盘时间从而提升整体性能。
- 在数据压缩前，采用华为自主研发的多种预处理算法，针对主要备份场景下的数据特点进行特征处理，达到备份全场景覆盖，预处理的特点是通过数据流特征识别数据场景，并按照数据特征对数据进行清洗（去除软件在数据中插入的Tag等特殊标识）、重排（按照数据类型特征进行可逆的数据重排，提升压缩率）、去冗余（按照数据特征去除冗余数据），提升压缩率。预处理只对数据做特征处理，处理完成的数据和识别的数据特征传输到自适应LZ模块进行去重压缩。区别于普通的LZ压缩算法，自适应LZ压缩算法会按照预处理模块识别的数据特征，匹配多种预置的压缩策略对预处理后数据的进行压缩处理。

数据缩减 - 数据压紧 (Data Compaction)



- OceanProtect压缩后的用户数据按照字节级对齐，进行数据压紧，减少浪费物理空间，相对于1kB（相对业界主流对齐粒度）粒度对齐，提供更高缩减率。
- 字节级压缩，比如1.1K 装进2K的Chunk，会有0.9K的浪费，系统会给有空的地方做空余标记，新数据接着在原来的地方写，保证装满，专用备份存储控制粒度是1K对齐的；
- 技术难点在于数据凑整和数据回收的配合；定长压成变长，变长凑整。
- 通过字节级对齐，减少物理空间的浪费，相对于1kB粒度对齐节省了2kB物理空间，从而提升压缩率。

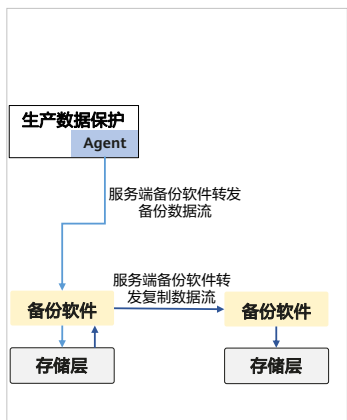
数据重排



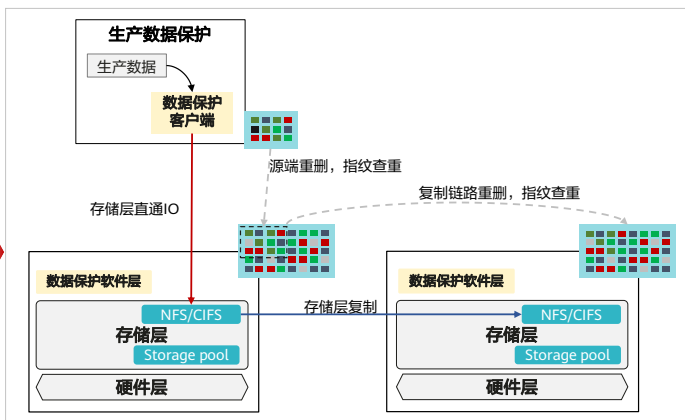
- OceanProtect中采用了数据重排技术，对于一个Grain粒度的数据块，OceanProtect系统会在每个512B后面都会插入数据的校验信息。由于校验信息中存在随机数据，对于压缩算法，随机数据会打乱压缩算法的计算，降低压缩率。因此，数据重排技术将用户数据和校验数据分开，对用户数据进行压缩，同时对校验数据进行定制化缩减。这样能够得到更好的压缩率和压缩解压速度。
- DIF数据包含校验数据和其他数据，随机性较大，DIF数据的压缩效果较低。DIF数据和DATA放在一起压缩会降低数据整体的压缩率。
- 采用剔除DIF数据后对大数据块单独进行压缩，对大数据块重新计算DIF数据可以提升整体的数据缩减率。
- DIF: data integrity field, 数据完整性字段，数据完整性字段包括了用户数据的CRC校验、LBA信息及其他自定义信息。

源端重删+短IO路径

传统备份软件：数据流经过备份软件转发

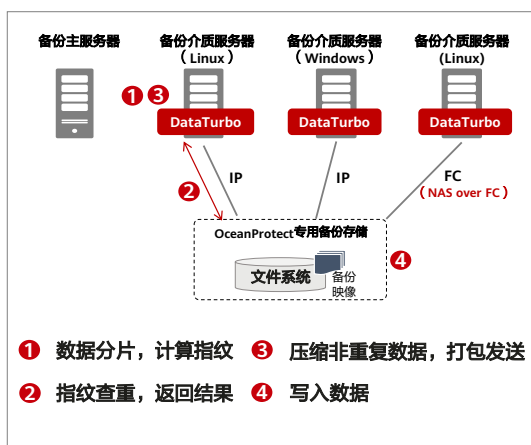


数据流IO直通存储层



- 由于OceanProtect专用备份存储后端为多控制器设备，为避免数据的跨控转发，源端重删客户端与存储之间会建立多条路径，并且多条路径链接存储的每个控制器节点，DataTurbo将IO请求直接下发到所归属的控制器节点，避免跨控制器节点，IO路径短，提升设备写性能。

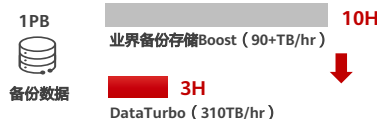
基于DataTurbo客户端的源端重删技术



降低TCO

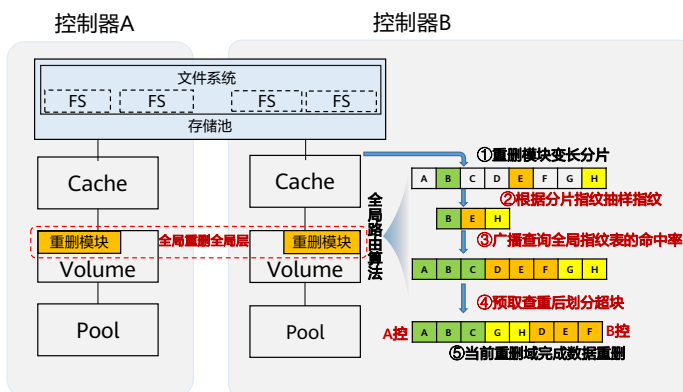
- 在备份介质服务器部署DataTurbo源端重删客户端, 减轻生产环境的工作负载;
- 通过源端重删技术, 仅传输非重复数据, 节省网络带宽, 提高逻辑备份带宽;
- Linux/Windows环境支持NAS over FC特性, 从而实现FC网络利用, 节约建设成本。

缩短备份时间



- DataTurbo客户端, 作为存储客户端运行在计算节点上, 通过网络协议与后端存储节点进行数据交换。客户端通过兼容标准POSIX语义, 提供数据缓存算法和源端重删压缩算法, 使上层应用能更高效的访问存储。
- DataTurbo客户端兼容标准POSIX语义, 通过MetaDataCache和DataCache实现元数据与数据缓存的聚合, 并通过源端重删 (SourceDedupe) 和源端压缩 (SourceCompression) 技术, 在数据写入存储层前先进行重删和压缩处理, 从而节省网络带宽并提升性能。DataTurbo作为部署在备份服务器上的用户态软件, 后端访问OceanProtect专用备份存储的NAS文件系统, 前端则通过Linux操作系统Fuse框架注册用户态文件系统, 使备份软件能够像使用NAS文件系统一样使用DataTurbo提供的文件系统进行备份操作。
- DataTurbo支持通过FC网络访问OceanProtect专用备份存储, 为客户提供更多组网选择, 以便部署备份解决方案。DataTurbo客户端和备份存储支持建立多条FC路径, 每条路径均采用标准SCSI协议与存储进行通信。DataTurbo客户端将文件系统IO封装为SCSI协议的读写命令, 其中写命令用于将文件系统IO传输至存储, 读命令用于获取IO执行结果。存储前端协议层解析SCSI命令中的文件系统IO, 并将其发送至存储后端处理。

跨控全局指纹管理+存储池级重删



* 全局重删基于同一存储池实现，跨池文件系统不支持全局重删

全局去重

- 跨控制器存储池级全局指纹去重，多文件系统间数据可重删

高效并发

- 多控制器AA并行处理，全面发挥多控架构优势，提升去重处理效率

交付维护更简单

- 单存储池只需规划1个文件系统（最高32变1）
- 部署实施、扩容维护更简单

* 规划文件系统数量根据实际交付局点决定

- OceanProtect专用备份存储可以对接不同厂商的备份软件以及上层业务应用，备份数据写入存储的打散逻辑存在不确定性。同时理论上，查找重复数据的范围越大，找到相同或相似数据的概率越高，从而获得更高的重删率。OceanProtect实现了在存储池（Storage Pool）内的全局重删功能，即重删执行时，所有属于同一存储池中的文件系统，相同或相似数据均可通过重删消除冗余。
- 全局重删算法继承现有的变长分片能力，同时会在存储池范围内查询所有控制器上的指纹，达到最好的重删效果。

目录

1. 数据保护产品介绍

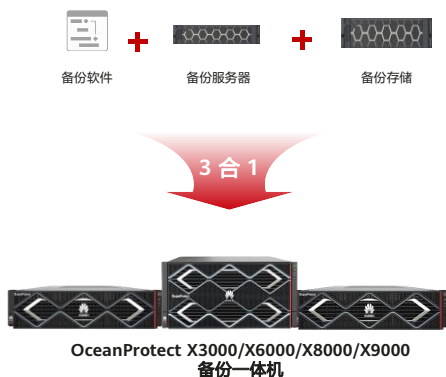
- OceanProtect专用备份存储
- **OceanProtect备份一体机**
- OceanCyber数据安全一体机

2. 数据保护产品应用场景

3. 数据保护产品部署

OceanProtect备份一体机简介

传统方案



敏捷高效

1. **硬件加速备份快**，增量备份秒级合成，小文件10倍加速
2. **用量省**，多层次数据缩减，重删率业界领先40%
3. **高效恢复**，副本即时可用，百亿文件秒级检索
4. **绿色低碳**，全闪高密，容量密度提升5倍，能耗降37%

安全可靠

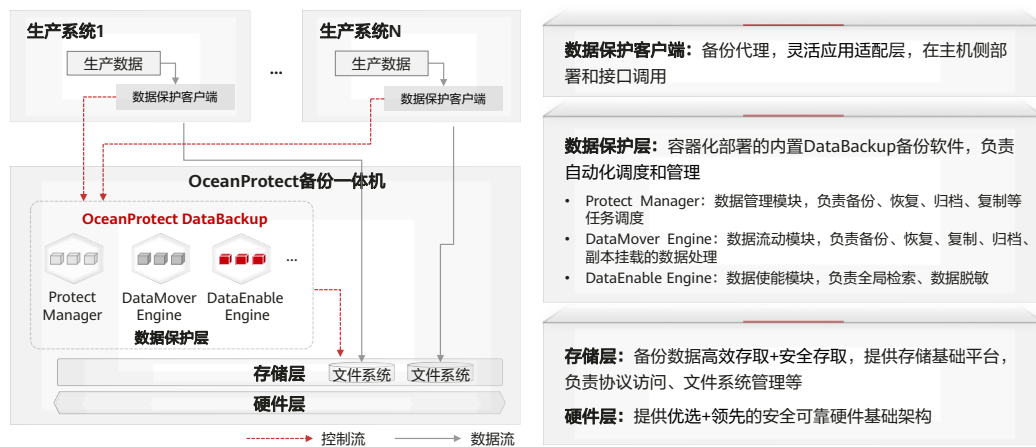
1. **集群AA架构**，任务多节点均衡，故障业务无影响
2. **全面防勒索方案**，勒索病毒检测，副本安全无忧
3. **安全恢复**，自动脱敏和演练，数据安全恢复有保障

生态丰富

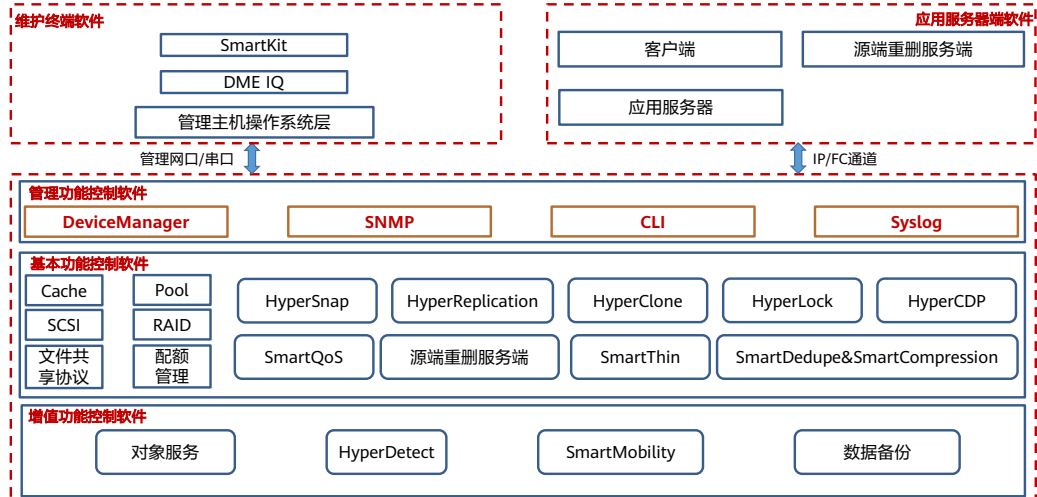
1. **广泛覆盖**传统+新兴应用生态，本地数据中心和云服务
2. **存储层备份加速**，数据库、虚拟化和文件细粒度恢复
3. **大数据组件支持最全**，6大组件全支持

- OceanProtect备份一体机通过3合1架构融合了备份存储、备份软件、备份服务器功能；其中数据备份软件（OceanProtect DataBackup）将以容器的方式部署到控制器节点，控制器节点将由存储系统软件和数据备份软件共享。OceanProtect备份一体机为主流数据库、虚拟化、操作系统等应用提供E2E（End to End）包含数据备份、数据恢复和副本数据再利用的数据全生命周期管理，帮助用户实现关键数据高效备份，价值数据快速利用，节省数据保护投资。

OceanProtect备份一体机产品架构

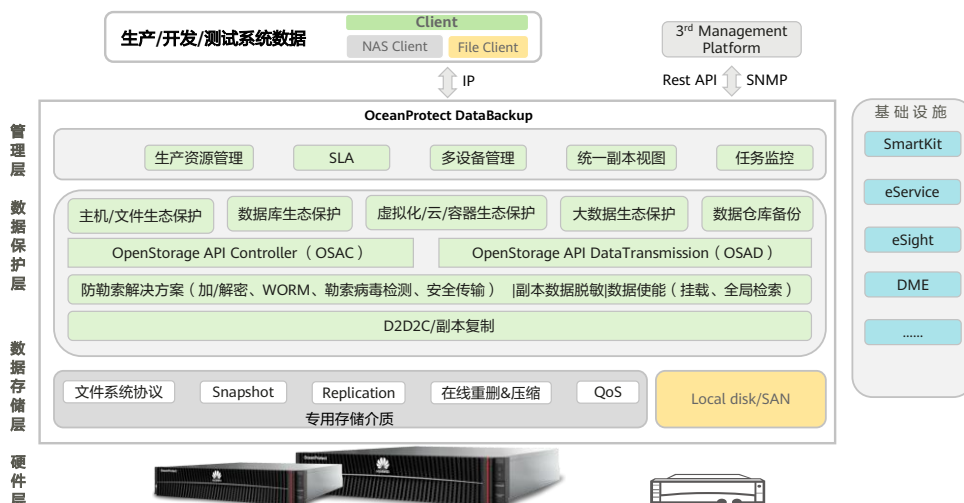


功能架构



- 存储系统端软件采用专用操作系统，实现硬件管理和支撑存储业务软件的运行。
- 维护终端软件用于系统配置与维护。用户可以通过维护终端的SmartKit、DME IQ等软件对系统进行配置和维护。
- 应用端包括：
 - 客户端：作为OceanProtect与生产环境的数据连接组件，用于备份恢复过程中的流程控制和数据传输。
 - 应用服务器：部署了待保护应用的服务器。
 - OceanStor DataTurbo：即源端重删客户端，与源端重删服务端软件模块配合，在生产端实现重复数据删除。

OceanProtect DataBackup备份软件架构

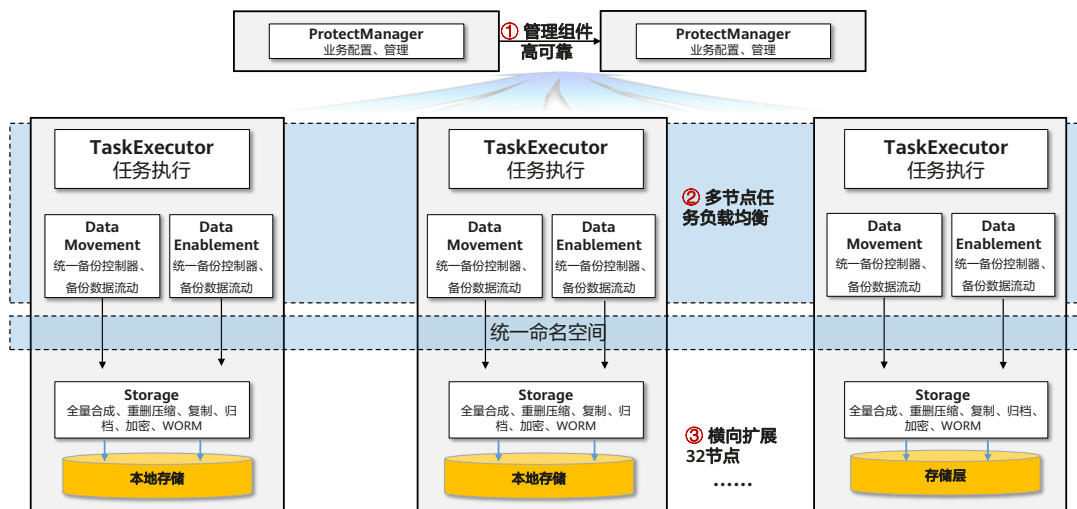


29 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



- 数据保护客户端：备份代理，根据需保护的应⽤类型识别需要安装的代理软件，在主机侧完成应⽤数据保护，I/O模块包含NAS Client（数据通过NAS协议直接写入后端X系列专用备份存储介质）和FileClient（数据通过私有协议传回一体机OSAD模块后，又OSAD模块组织数据格式写入本地盘或LUN存储）。
- 数据保护层：负责处理数据保护业务，容器化部署，分布式架构；主要支持主机文件、数据库生态、虚拟化/云/容器生态、大数据生态保护，提供端到端副本数据防勒索解决方案。以及副本数据流动能力。
- 数据存储层：存储基础平台，负责协议访问、文件系统管理等
- 硬件层：提供硬件，以及相关的驱动
- 基础设施：⼯具化安装部署、运维，对接存储产品线多种网管平台，同时支持标准协议对接三方网关平台。

集群软件架构



30 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



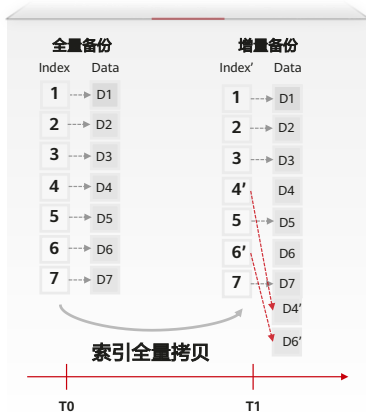
- OceanProtect备份一体机在应对大规模场景，需要将多套独立的一体机组成一个统一管控且支持Scale-out横向扩展的备份集群，突破单套设备容量、性能、控制器数量的限制，以满足大容量、高性能的需求。
 - X3000一体机仅支持和X3000一体机组成集群；
 - X8000一体机和X6000一体机可以组成混合集群，均可以扮演主节点、备节点、成员节点角色，但主备节点必须是用相同型号。
- 多套OceanProtect 备份一体机（每一套一体机为一个独立的K8s集群）组成一个大的备份集群，集群间共享GaussDB数据库，成员集群通过Service映射外部IP的方式访问主集群上的GaussDB。
- 为避免主节点故障导致整集群故障，支持配置主节点HA，主从集群GaussDB数据进行同步，当主集群故障后，从集群升主并使用同步过来的数据拉起GaussDB为整个备份集群提供数据服务。
- 被保护的生产资源关联特定的备份一体机节点，以确保该保护对象的副本数据都保存在对应的节点上，避免跨节点引发增量转全量的问题。PM生成任务时，根据被保护对象关联的节点将任务分发给对应节点上的ProtectManager，如果生产资源尚未与节点关联，则按照一定的负载均衡算法分配节点并进行关联。节点收到任务后便将任务终结在本集群内，整个任务过程与原来单K8s集群一致。
- 由于多套备份一体机之间存储池彼此独立，为了防止单套设备整体故障引发副本数据不可用的问题，可以通过在SLA中配置存储策略，将不同时刻的全量副本链分散到不同的备份域存储中，这样当其中某个节点存储故障时，在其它正常的节点中还有其它

时刻的数据副本可用。

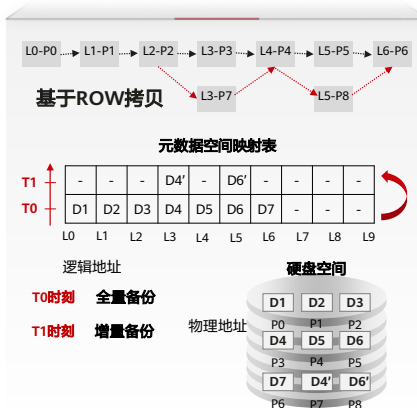
- ProtectMananger限定在主从集群上通过管理口访问，部分操作只能在主集群上进行，如ProtectAgent的安装包管理、推送安装等（导入的Agent安装包需在主从集群同步）。
- 跨集群的业务数据（包括PM上任务分发）使用备份网络平面，管理口故障不影响备份业务。

合成加速

传统方案：基于索引拷贝的永久增量备份



OceanProtect：基于ROW的永久增量备份



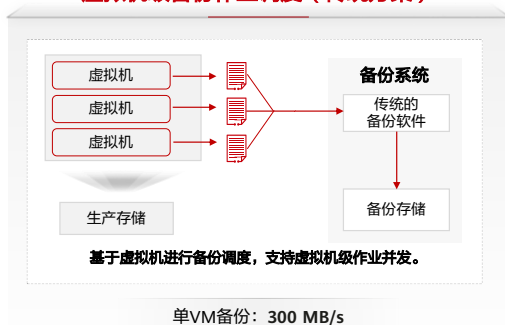
秒级合成



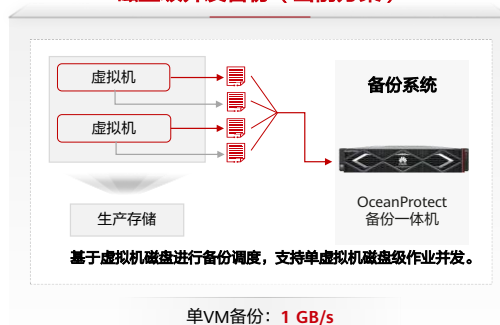
- OceanProtect备份一体机支持的大部分应用，均支持永久增量备份方式进行增量数据存储，和传统后期合成永久增量不同，备份软件采用ROW（Redirect-On-Write）写时重定向方式进行合成。备份数据以原始文件格式或固定文件格式写入，当增量备份时，分析增量数与上一次备份数据的差异数据块和偏移位置，并将差异数据覆盖写入文件，得到最新的全量合成副本，副本合成后生成快照，快照记录以来关系，当下次增量备份时，基于上次文件系统快照，新数据分片重定向写入新位置。省略中间合成过程。提升备份和恢复效率。
- 传统厂商在实现永久增量备份时，需要先进行一次上次备份副本的索引文件全量拷贝，再根据增量信息修改对应的偏移索引。得到永久增量副本的全量索引文件。或者拷贝真实数据后再合并写入成新文件写入，这种实现方式需要拷贝一次索引文件，并多次对索引文件进行修改写，或者拷贝整文件内容合并后写入，对存储层带来极大的IO负担。
- OceanProtect 备份一体机永久增量备份技术，针对索引文件采用ROW快照技术，变化数据块仅新增变化数据索引信息，及数据块前后偏移的索引依赖关系，不拷贝全量索引。极大缩减存储IO开销，提升备份性能。

多任务加速

虚拟机级备份作业调度（传统方案）



磁盘级并发备份（当前方案）



虚拟化、云、容器并发备份：

将虚拟机、容器或云主机作业拆分成多个磁盘子作业，分配给多个备份代理节点。

大数据生态应用并发备份：

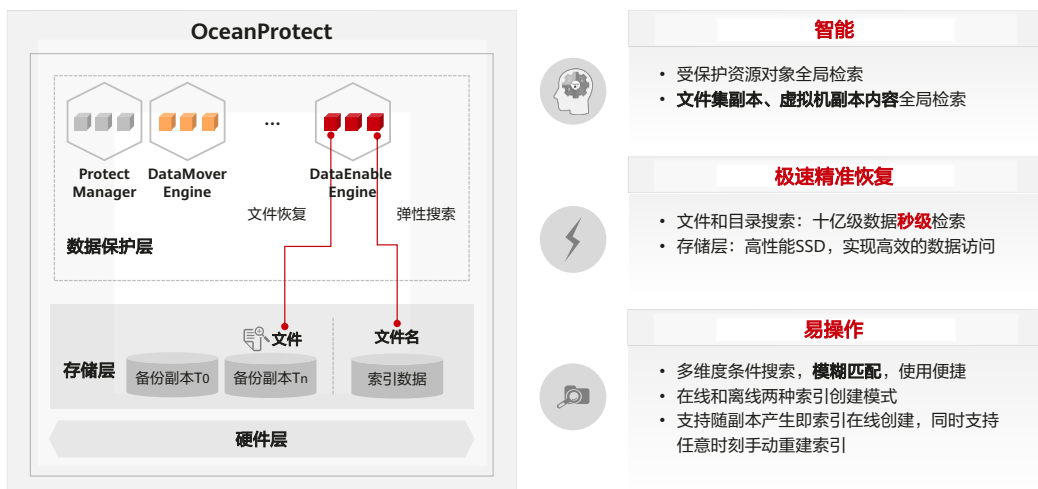
将一个作业拆分成多个文件级（HDFS）或表级（HBase/Hive）的子作业，分配给多个备份代理节点。

DWS备份：

根据节点生成的备份对象，将作业分配给多台OceanProtect设备。

- OceanProtect 备份一体机针对大数据量应用，为提升备份带宽，将任务拆分为多个子任务并分摊到多个节点内，最大程度利用多节点的计算和传输能力，以满足PB级数据量的备份性能诉求。
 - 虚拟化/云/容器类应用并发：将虚拟机/容器/云主机整机任务拆分为多个磁盘任务，以磁盘子任务粒度分摊到多个备份代理节点上。
 - 大数据生态类应用并发：以文件粒度（HDFS）或表级粒度（HBASE/Hive）拆分为多个子任务，并分摊到多个备份代理节点上。
 - DWS数仓备份：以节点生成的备份对象粒度分摊到多个OceanProtect设备上。

全局检索

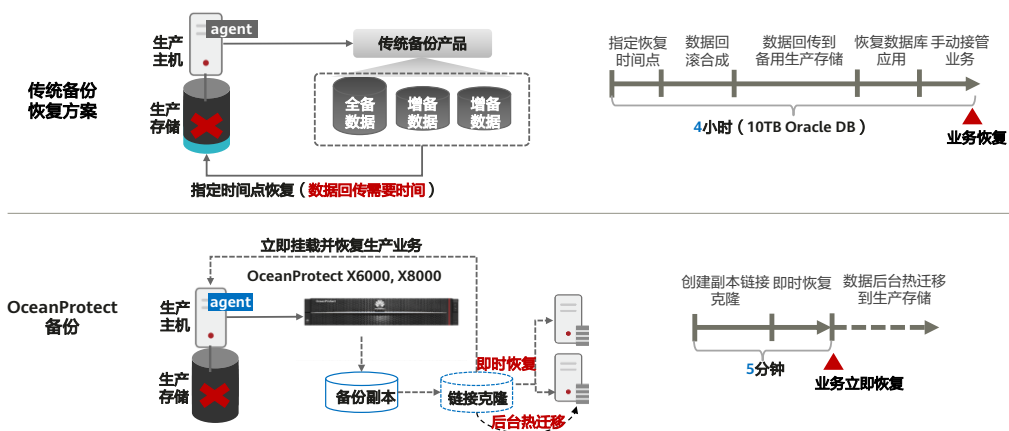


33 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



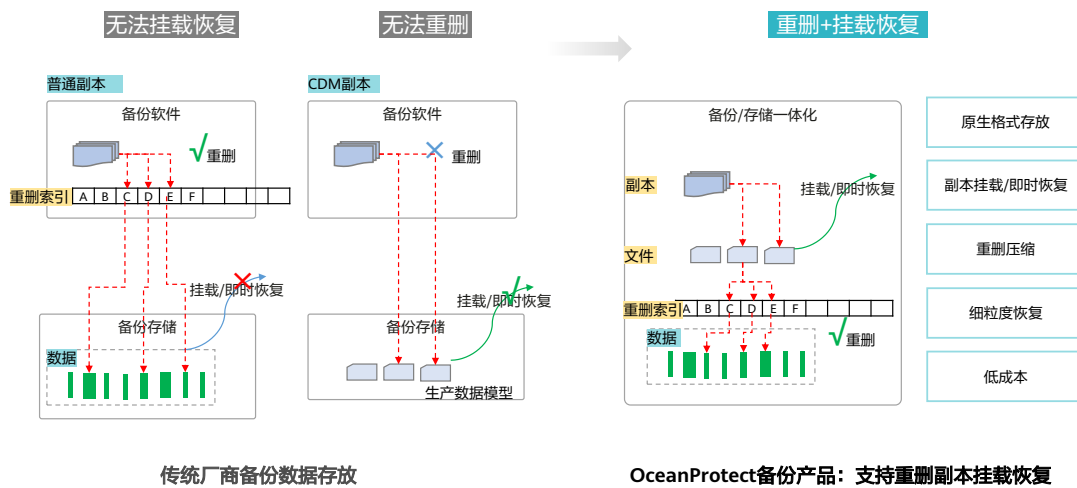
- OceanProtect 备份一体机内置全局检索能力，提供整系统受保护资源的搜索能力和备份副本数据搜索能力，支持文件、数据库、虚拟化/私有云/容器的生产资源类型按照资源名称进行搜索，并快速创建备份恢复任务，提升管理效率。支持对文件、虚拟机、HDFS等的备份副本数据全局检索。支持模糊匹配文件名、目录名搜索。能快速查找受保护环境中的所有对象（如添加的VMware vCenter中的虚拟机），同时能秒级检索VMware虚拟机备份副本，主机文件级备份副本中的目录、文件，并针对查找到的目标文件进行细粒度的恢复和下载。
- 同时提供在线索引和离线索引两种索引创建方式，在线索引在备份副本生成后立即进行索引创建，离线索引在备份完成后，手动选择副本创建索引。备份完成后自动或用户在界面手动触发副本数据索引创建任务，主机文件级的索引数据直接在备份过程中的元数据导出，VMware整机备份副本的索引在索引创建任务中，首先将原生格式副本挂载到节点上，识别并挂载副本中的文件系统，文件系统挂载到节点后，遍历目录和文件生成副本内容的索引信息。
- 全局检索能力可以快速直观的展示副本内容，帮助用户快速发现多个备份副本中的任意文件的变化过程，并对细粒度恢复提供基础能力。

数据即时恢复



- 为便于用户即时查看副本内容，或利用备份副本进行开发测试、数据分析等操作，OceanProtect支持Oracle、VMware、NAS备份、MySQL备份、文件备份等备份副本和复制副本的Live-Mount挂载能力以及Oracle、VMware的后端数据迁移能力。

重删副本挂载恢复



- 由于原生格式副本需要按照通用协议可识别的方式组织文件存储，这样生产应用才可识别这些备份文件，所以需要支持如SCSI、NFS、CIFS、HDFS、S3等标准协议，传统备份软件由于重删层处于前述协议层之上，因此一旦由备份软件重删后，数据就不可被标准协议识别，或者以标准协议格式存放，就无法进行重删压缩、加密操作，因此，传统备份软件通常是挂载能力与重删压缩等能力互斥，给客户使用带来极大不便。
- OceanProtect与数据备份特性架构垂直整合，将重删能力构筑在文件系统协议之下，并支持标准NFS、CIFS协议，将挂载、即时恢复等快速恢复能力与重删压缩、加密等能力兼而顾之，完整的满足客户诉求。

目录

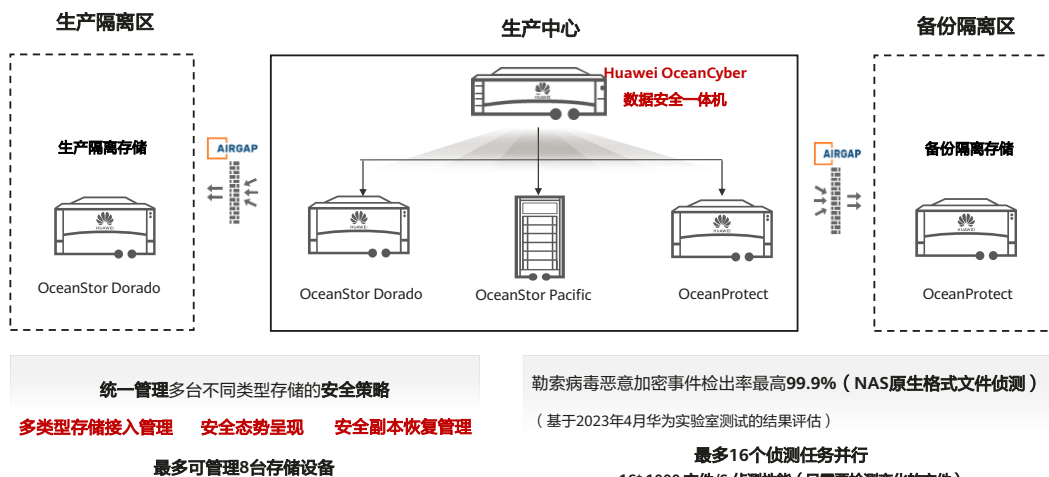
1. 数据保护产品介绍

- OceanProtect专用备份存储
- OceanProtect备份一体机
- **OceanCyber数据安全一体机**

2. 数据保护产品应用场景

3. 数据保护产品部署

数据安全一体机产品简介

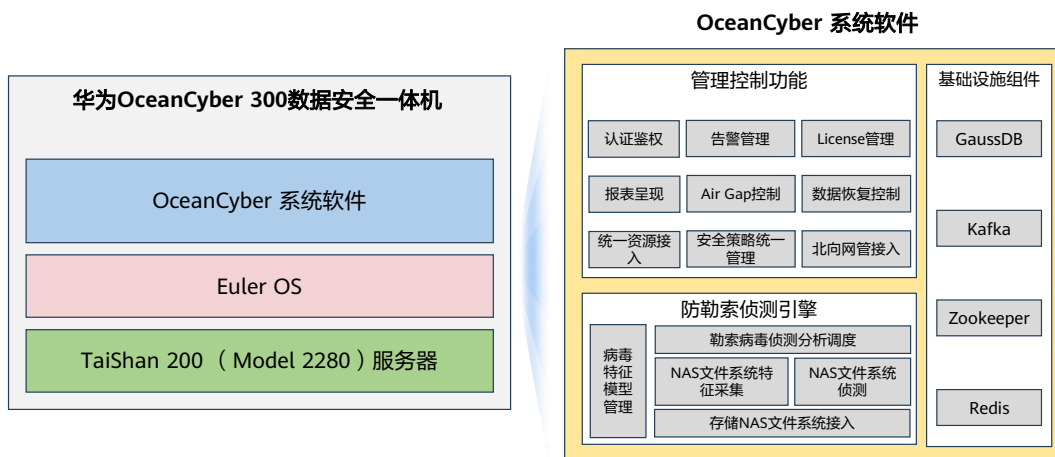


37 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



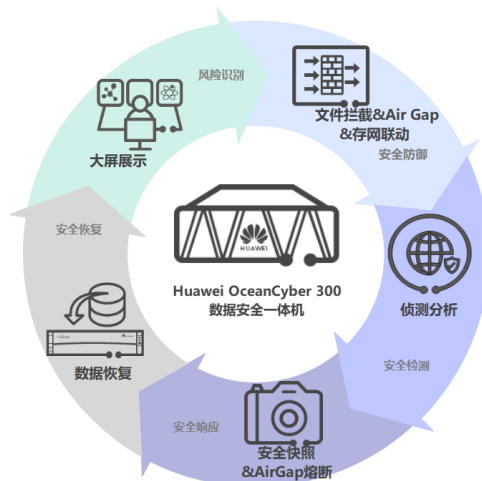
- 华为OceanCyber 300数据安全一体机是一款集华为OceanCyber系统软件、操作系统（Euler OS）、通用服务器于一体的即插即用、稳定可靠的一体化产品，基于OceanCyber系统端软件提供针对勒索病毒对于存储数据恶意加密的侦测、保护、响应以及恢复能力。
- 华为OceanCyber 300基于华为自研勒索侦测算法，支持多台、多类型设备的统一防勒索管理，针对勒索病毒攻击行为，具有识别准、侦测性能高、防护全、数据恢复快的特点。

产品架构



- 华为 OceanCyber 300以稳定可靠的华为通用服务器与Euler操作系统为基石，软硬全
- 栈自研，为OceanCyber系统软件提供安全、稳定、高效的运行环境。
- 关键的可靠性保障：
 - 主要器件均采用冗余结构，网卡端口Bond、数据盘系统盘隔离、RAID保护，预防单点器件故障，保障系统稳定运行。
- OceanCyber 系统软件运行于专用操作系统之上，使用全容器化、微服务化部署模式，按应用资源隔离，极大减少故障影响域。从架构视角，OceanCyber系统软件分为如下层级：
 - 华为OceanCyber 300通过管理控制功能软件，实现对系统的设备管理以及防勒索业务的统一管理功能。
 - 华为OceanCyber 300通过防勒索侦测引擎软件，实现对于NAS文件系统的勒索病毒恶意加密事件的侦测分析能力。
 - 华为OceanCyber 300数据安全一体机通过基础设施组件，提供系统业务运行的软件框架能力。

OceanCyber 防勒索特性全景



- 华为 OceanCyber 300 数据安全一体机作为华为存储防勒索解决方案的重要组成部分，针对勒索病毒等网络威胁攻击行为，对NAS文件系统提供了勒索病毒恶意加密行为的侦测分析能力以及对应的快速响应恢复能力；
- 基于上述的能力，在存储侧应对勒索病毒攻击时，实现可预防、可感知、可恢复，助力提升数据安全韧性。
- 侦测分析（实时侦测）：通过分析存储设备NAS文件系统的审计日志，OceanCyber 300获取被保护的文件系统的异常I/O行为，并且利用在文件系统中部署的少量诱饵文件进行攻击捕获，可以近实时的识别文件系统是否正在被勒索病毒感染（被勒索病毒恶意加密）。
- 侦测分析（智能侦测-NAS文件系统事后侦测）：基于勒索病毒对数据攻击的特征进行侦测，识别出被勒索软件感染（恶意加密）的文件系统，同时也能够评估可用于数据安全恢复的未被感染的“干净”快照副本。
- 侦测分析（智能侦测-备份副本深度侦测）：针对华为OceanProtect专用备份存储上的备份副本文件，数据安全一体机提供了备份副本文件深度侦测能力，可以识别备份的原始数据是否被感染（被恶意加密）。
- 文件拦截（拦截已知勒索病毒后缀文件）：安全一体机通过向存储设备下发已知勒索病毒对应的文件扩展名过滤规则，实现勒索攻击前（勒索病毒传播阶段）拦截已知勒索病毒感染后缀文件的写入。
- Air Gap管理：安全一体机管理Air Gap策略，策略应用到存储设备，支撑针对勒索病毒等网络威胁攻击的安全隔离能力；安全一体机勒索侦测发现异常事件，可以立即断

开Air Gap复制链路（Air Gap熔断），保证隔离域完全“离线”，提升安全防护级别。

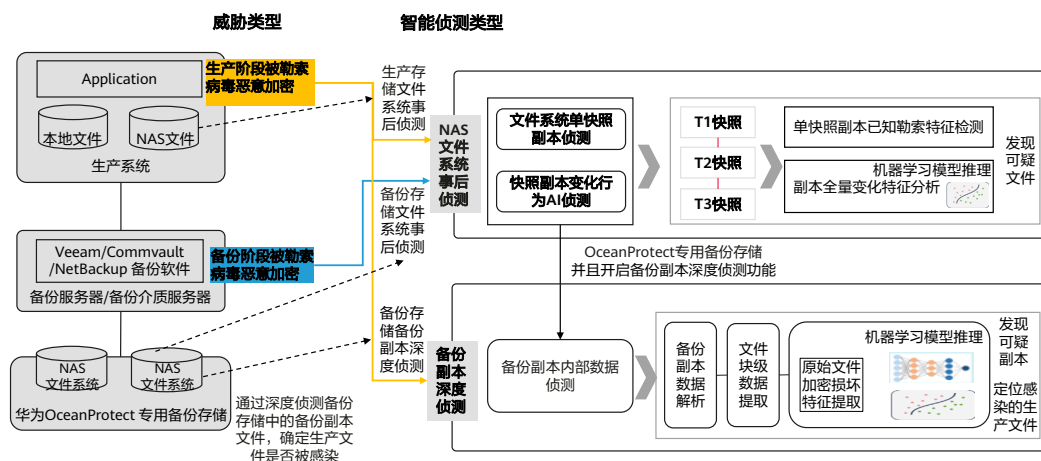
- 存网联动：基于华为的数据中心全栈安全能力构建针对勒索病毒的多层防线，通过联动防御机制，快速侦测勒索病毒、快速响应处置（触发Air Gap熔断），快速阻断非法数据转移。
- 安全快照：针对没有被感染的文件系统快照，OceanCyber 设置“干净”副本的标签，快照副本可以被设置为安全快照，支撑用户后续的文件系统恢复。
- 数据恢复：数据安全一体机侦测到勒索事件之后，支撑用户选择“干净”的副本用于恢复。
- 大屏展示：展示存储设备状态、异常文件系统列表、告警概览信息、任务概览信息、Air Gap状态信息等安全态势。

智能侦测-NAS文件系统事后侦测



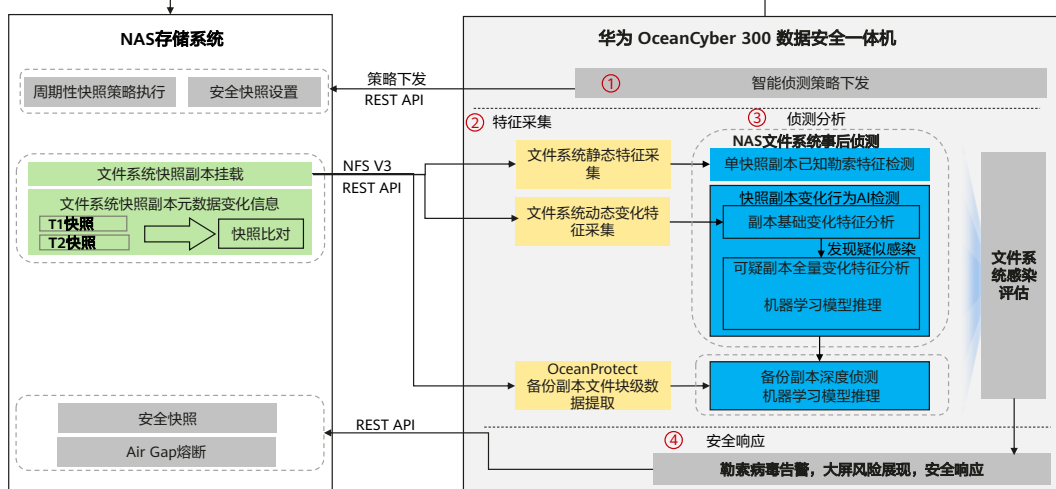
- 华为OceanCyber 300数据安全一体机的智能侦测特性，针对文件系统的快照副本以及快照副本变化的特征，采用机器学习算法对于文件系统进行勒索软件恶意加密的周期性侦测，可以有效的感知勒索病毒针对文件系统的加密行为，向用户生成告警信息，同时可提供标记为“干净”的未感染副本用于数据的恢复。
- 智能侦测针对勒索软件对客户数据的恶意加密行为，对于NAS文件系统进行周期性快照副本侦测，在获取有效的侦测结果知会用户数据风险的同时，提供“干净”的副本用于客户的数据恢复。
- 针对现网各种勒索病毒家族对于存储数据的恶意操作行为，华为内部实验室进行统一的汇总、分析、抽象，提取对应的勒索病毒攻击特征，结合机器学习技术，形成了自研的文件系统智能侦测算法。

智能侦测-备份副本深度侦测



- 对于OceanProtect专用备份存储之中的备份副本文件，上述的NAS文件系统事后侦测可以检测备份软件生成的副本文件是否被勒索病毒恶意加密（对应备份服务器被勒索病毒感染的场景）；但是如果需要分析备份副本之中的原始数据（生产存储设备上的文件，虚拟机或者数据库）是否被感染，就要通过开启智能侦测的“备份副本深度侦测”功能进行侦测分析。
- 对于备份软件在OceanProtect专用备份存储中生成的备份副本文件，OceanCyber数据安全一体机的备份副本深度侦测功能通过解析副本文件的内容，对副本文件的内部结构进行归一化处理；基于归一化后的副本文件内部数据，采用文件块级全量数据提取的方式，对于副本文件进行全量数据采集；采集后的数据，经过计算得出文件的损坏程度特征信息；基于特征信息，使用华为自研的机器学习算法进行侦测，分析评估对应的备份副本文件之中的原始文数据、虚拟机数据或者数据库数据是否被感染；
- 备份副本深度侦测是通过对于备份存储上的备份副本文件进行深度解析，判断备份副本之中生产系统的原始文件/虚拟机/数据库是否被勒索病毒感染；由于生产系统之中文件I/O负载模式多样，部分场景的I/O负载与勒索病毒有可能存在相似性（如大批量写入加密文件，大批量删除文件等），为了降低这种相似性造成的勒索事件误报，推荐在开启了OceanProtect文件系统的备份副本深度侦测功能之后，同步开启“学习模式”；
- “学习模式”用于在指定的学习期内，OceanCyber数据安全一体机对于当前的备份副本文件内容以及文件变化趋势进行学习，进而获取备份副本之中的原始文件在正常I/O场景下的变化趋势特征；这些特征作为备份副本深度侦测机器学习算法的一部分，参与对备份副本文件进行侦测，用于消减上述特殊I/O场景引起的误报。

智能侦测业务交互

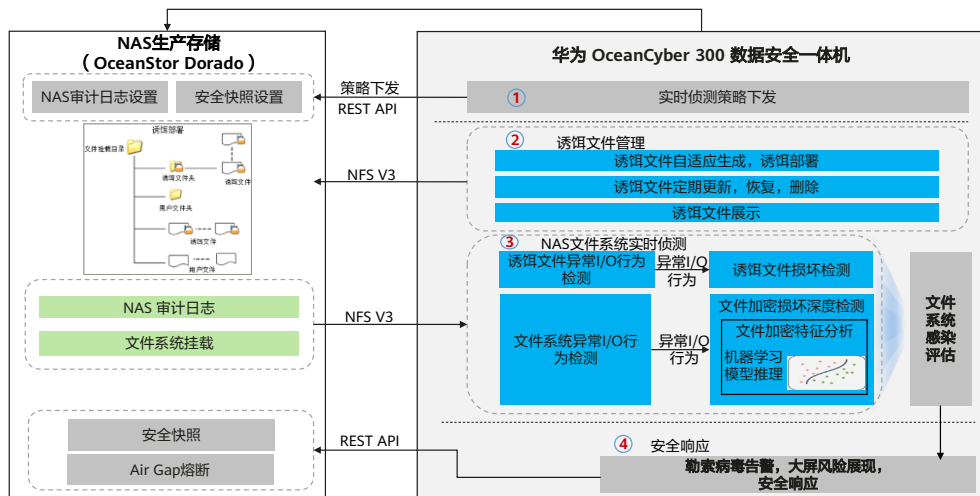


1. 侦测策略下发：安全一体机调用存储设备的REST接口，下发勒索侦测快照策略，对于文件系统创建周期性快照，基于创建的快照副本数据，进行勒索病毒的周期性侦测。
2. 特征采集：安全一体机通过NFS 共享的方式（NFS V3），挂载被侦测的文件系统，用于采集文件系统特征信息；安全一体机同时通过调用存储设备的文件系统快照比对API接口，快速获取多个快照副本之间的文件元数据变化信息（文件路径、文件大小、mtime等信息），支撑元数据变化采集，提升系统的性能；
3. 侦测分析（智能侦测）：通过机器学习算法对文件系统采样特征进行分析，评估与勒索软件行为是否匹配，识别文件系统是否被勒索病毒感染。
 - 单快照副本的已知勒索特征侦测：对于文件系统指定快照副本的疑似勒索病毒感染痕迹进行检测。
 - 快照副本变化行为AI侦测：采用机器学习算法，针对文件系统前后快照副本数据变化的行为侦测。
 - 备份副本深度侦测：对OceanProtect专用备份存储的备份副本文件，采用机器学习算法，对备份副本的原始数据进行加密损坏深度检测。
 - 多种检测技术相互补充，对于已知勒索特征检测以及具有相似恶意行为的未知勒索病毒的侦测都具有较好的侦测效果。
4. 安全响应：侦测到快照副本数据被感染之后，OceanCyber告警提示用户，用户可以进行相应的响应以及恢复操作，尽快恢复数据；同时可以根据用户的配置，断开对

应存储设备的Air Gap复制链路，防止勒索病毒造成的感染文件扩散（Air Gap熔断）。

- 针对没有被感染的文件系统快照，OceanCyber设置“干净”副本的标签，也可以设置为安全快照，支撑用户后续的文件系统恢复。

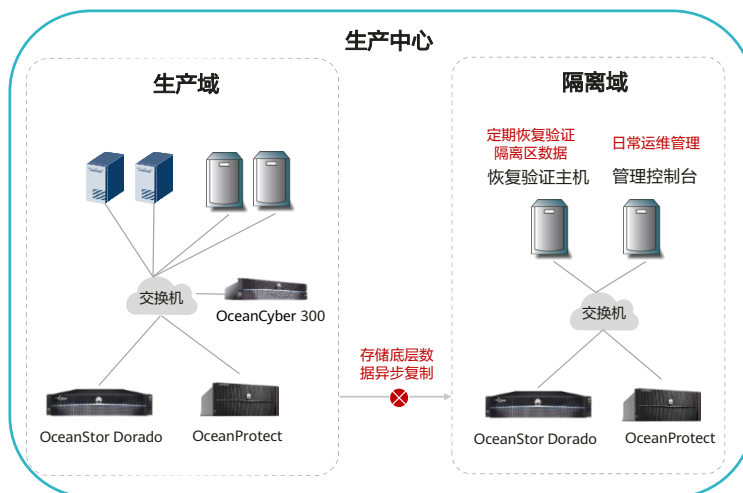
实时侦测业务交互



1. 实时侦测策略下发：安全一体机调用存储设备的REST接口，开启需要被保护的文件系统的NAS审计日志，基于审计日志，对文件系统进行勒索病毒实时侦测。
2. 诱饵文件管理：安全一体机提供在生产存储的文件系统中部署诱饵文件的能力，通过欺骗勒索病毒对诱饵文件进行异常操作，能够快速识别勒索恶意加密的攻击行为。诱饵文件采用自适应生成方式，安全一体机通过对文件系统的文件特征（如文件名、文件大小、修改时间等）提取并建模，拟合用户文件系统环境进而生成针对性的诱饵文件，确保生成的诱饵文件有较高的引诱性（优先被勒索病毒攻击）同时难以被勒索病毒识别并绕过。
3. NAS文件系统实时侦测：安全一体机基于文件系统NAS审计日志，分析文件系统的I/O行为（如文件读、写操作，加密文件、删除文件等），使用算法评估是否存在异常I/O操作。通过机器学习算法对异常I/O涉及的文件进行深度检测，评估与勒索软件行为是否匹配，近实时的识别文件系统是否被勒索病毒感染。
 - 诱饵文件侦测：对于存在异常I/O行为的诱饵文件，进行诱饵文件内容损坏检测，能够近实时检测出勒索病毒的攻击。
 - 文件系统I/O在线侦测：对于存在异常I/O行为的用户文件，采用机器学习算法，进行文件深度检测。
 - 上述两种检测技术相互补充，对已知勒索病毒以及具有相似恶意行为的未知勒索病毒都具有较好的侦测效果。
4. 安全响应：侦测到文件系统被感染之后，安全一体机告警提示用户，并且立即对文件系统设置安全快照。用户可以进行相应的响应处置操作，同时结合智能侦测获取

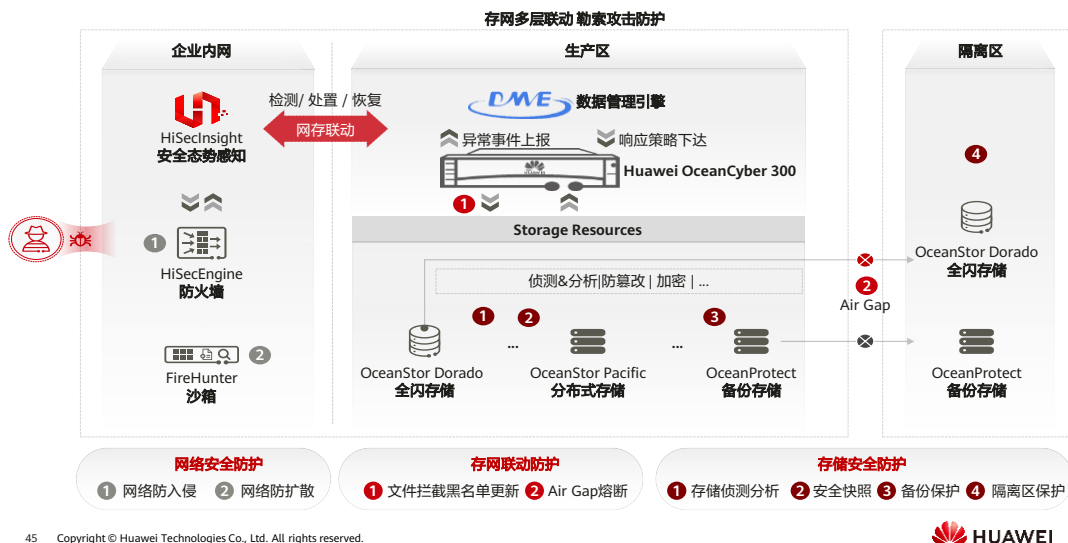
的文件系统“干净”副本，尽快恢复数据。安全一体机也可以自动断开对应存储设备的Air Gap复制链路，防止勒索病毒造成的感染文件扩散（Air Gap熔断）。

Air Gap管理



- 基于零信任安全模型，假设在应对勒索病毒攻击时，网络侧防护失效，勒索病毒攻击到生产中心，造成整个生产中心被“污染”后，存在一个安全信任的环境，该环境中有一份可以用于恢复的黄金副本，企业有信心恢复业务数据。
- 因此业内网络安全框架大多都要求建立一个隔离区，为减少暴露面，该隔离区和生产环境仅做必要的复制连接，且仅当复制时候连接才会建立，复制完成后，该连接自动断开，这个隔离区处于“离线”状态。这个复制过程称之为Air Gap复制。
- 华为OceanCyber 300数据安全一体机提供管理存储设备Air Gap端口能力。通过配置的Air Gap策略，OceanCyber通过调用存储设备管理面的REST API，控制存储设备用于Air Gap复制的逻辑端口的通断状态，限制业务数据在指定的时间段内才能够允许复制，从而实现安全隔离。

存网联动



- 华为OceanCyber 300 数据安全一体机通过与华为网络安全设备（HiSecInsight）的威胁情报共享，构建网络安全防护以及存储安全防护的两道安全防线和多层防御体系，实现E2E防勒索。
- 业务流程：网络安全设备提前感知到异常风险，通知华为DME存储管理平台，DME根据预定义策略，触发OceanCyber主动执行Air Gap熔断以及文件拦截黑名单更新，快速防止勒索病毒的感染扩散。

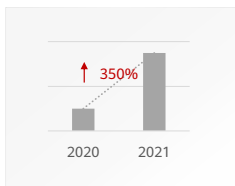
目录

1. 数据保护产品介绍
- 2. 数据保护产品应用场景**
3. 数据保护产品部署

勒索软件已成为全球主要网络威胁

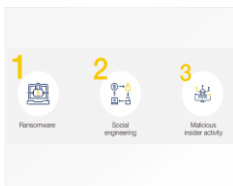
勒索软件本质是恶意软件攻击，对重要数据加密/窃取，导致计算机或特定文件不可用或不可读；只有受害人支付了赎金，才有可能获取用于恢复电脑或解密被加密文件的密钥。勒索软件变种多、变化快，普通的杀毒软件一般难以快速识别并查杀。

攻击量倍增



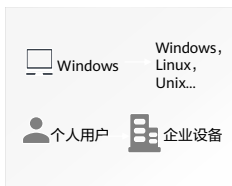
- 勒索攻击攻击率年增长率超 350%
- 2020-2021年，新型勒索软件变种的数量增加了 46%。

网络威胁最热问题



- 《2022 年全球网络安全展望》勒索软件攻击是全球网络领导者在网络威胁关心问题中排第一。

攻击目标更多元



- 2021年全球制造业、服务业、建筑、金融、能源、医疗、工控和政府组织机构等频遭勒索软件攻击，给全球产业产值造成严重损失。

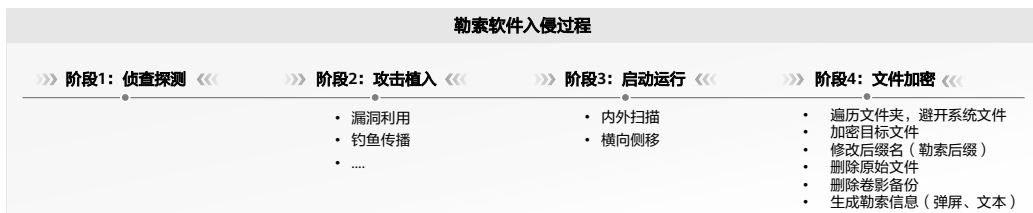
隐蔽性极强



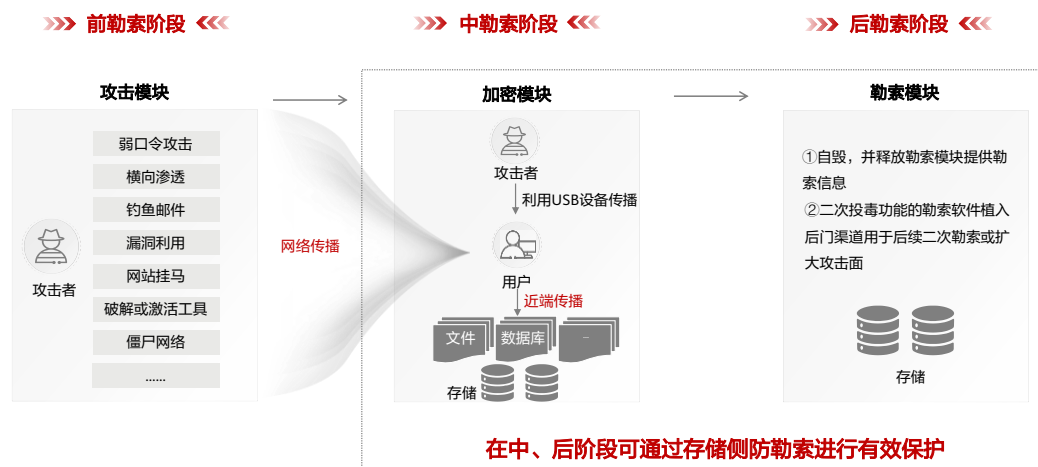
- 钓鱼式攻击：通过伪装，获得敏感信息的攻击。
- 暴力破解：穷举常见弱口令。
- 零日漏洞攻击（Zero-day）：在系统漏洞被发现前，发起攻击。

- 勒索软件已成为全球的主要网络威胁，攻击量增长了350%，是全球网络领导者最关心的威胁，对全球的政府、金融、医疗等各行业都造成了严重损失，其攻击隐蔽性极强，善于伪装，防不胜防。

传统网络安全、容灾、备份难以应对勒索软件攻击

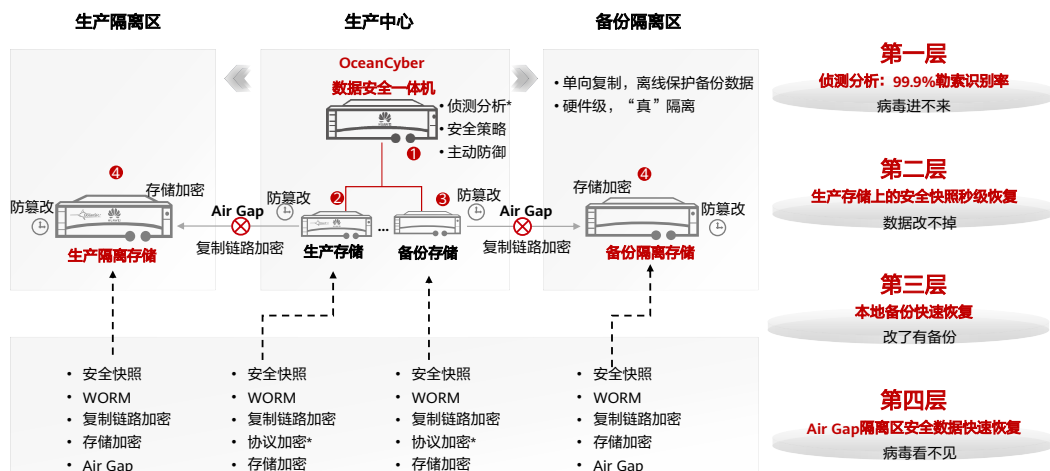


勒索软件分三个阶段进行攻击，存储是最后一道防线



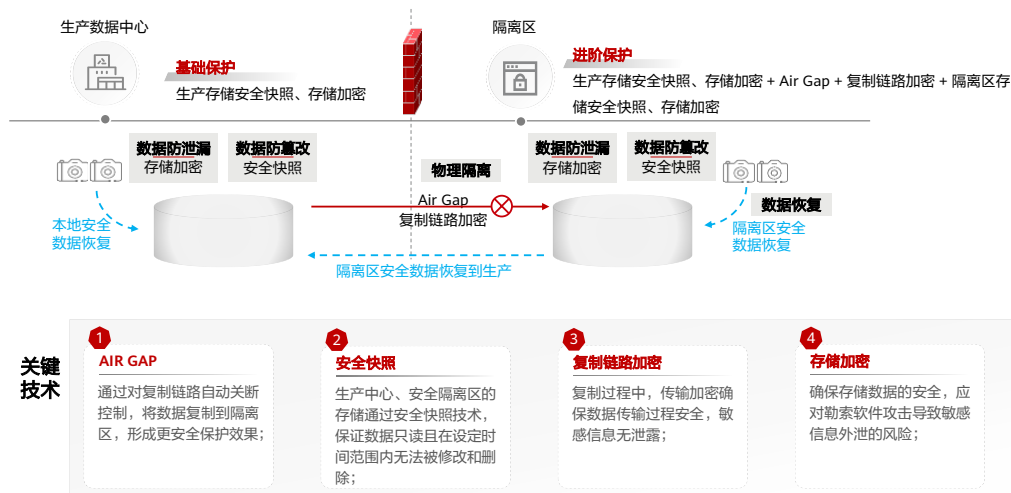
- 勒索病毒攻击模块主要通过携带各类漏洞exploit、密码字典等形式，对攻击目标进行漏洞检测以及服务口令爆破等。2020年起勒索病毒的攻击手段全线“开花”，弱口令攻击、横向渗透、钓鱼邮件、漏洞利用、网站挂马、破解或激活工具、僵尸网络、供应链攻击等手段悉数登场。
- 当攻击模块成功获取本地权限后，加密模块将开始运行，通过公钥加密或对称密钥加密等形式进行本地文件加密操作。具有窃密行为的勒索病毒还会在该环节进行特定格式文件窃取上传。在该阶段勒索病毒将对文件造成真实损害，因故称之为中勒索阶段，勒索病毒执行加密模块后会加密特定类型的文件，不同的勒索病毒会加密几十到几百种类型的文件，基本都会包括常见的文档、图片、数据库文件。
- 一旦勒索病毒加密完成，其往往将进行部分模块自毁，以防止自身加密逻辑或内存残留信息被利用于解密，并释放勒索模块向受害者用户提供勒索信息，要求受害者用户通过特定渠道支付赎金。在该阶段，主要勒索过程均已完成，只剩余提示勒索信息，因故称之为后勒索阶段。同时具有二次投毒功能的勒索病毒在该环节还会根据当前操作系统植入不同类型的持久化后门渠道用于后续人工操作二次勒索或扩大攻击面。

华为存储防勒索解决方案四层防护



- 华为存储防勒索解决方案拥有四层防护。
- 第一层，常见勒索病毒发起攻击时，侦测分析的勒索文件拦截功能进行事前拦截，识别病毒特征，阻断病毒的加密行为，成功拦截已知勒索病毒攻击。
- 第二层，未知勒索病毒加密数据时，侦测分析的实时勒索检测功能的机器学习算法自动检测到文件的异常IO行为，判断为勒索攻击，自动对被攻击的文件系统打安全快照，避免损失扩大，并发送告警，通知管理员介入处理。安全快照和WORM防篡改技术，可以保护数据副本被勒索病毒加密，在生产存储的数据被加密攻击后，可使用生产存储上保护的安全数据副本进行快速恢复。
- 第三层，生产存储被黑客攻破后，勒索病毒已经将用户的生产数据全部被加密或破坏。（备份存储侦测分析的智能勒索检测功能，已经在后台识别出安全的备份副本数据）用户可以使用备份存储上保存的干净的备份副本数据进行数据的快速恢复。
- 第四层，生产存储和备份存储都被黑客攻破，生产数据和备份数据都全部损失。此时，部署在生产隔离区或者备份隔离区的存储由于Air Gap的物理隔离保护，并未被攻击。可使用隔离区的备份副本进行数据的快速恢复。
- 四层防护可以及时检测并拦截网络和主机层遗漏的勒索攻击，主动对数据进行保护，减少数据损失，始终保证系统中存在一份干净的数据副本用于数据的快速恢复。构建业界最全的数据存储防勒索防线，确保数据安全的最后一道防线万无一失。

华为SAN存储防御勒索解决方案



51 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



- 华为的SAN存储防勒索分为基础保护和进阶保护两个版本，基础保护在生产存储上使用数据防篡改功能对数据副本进行保护，在遭遇勒索攻击，数据被损坏后，可使用被保护的数据副本进行恢复。本地安全快照的恢复速度可以实现秒级恢复。
- 而进阶保护可以保障当生产存储被黑客攻击，所有数据都被破坏的情况下，隔离区还有安全的数据副本用于恢复。
- SAN存储有4大技术可有效防御勒索攻击：

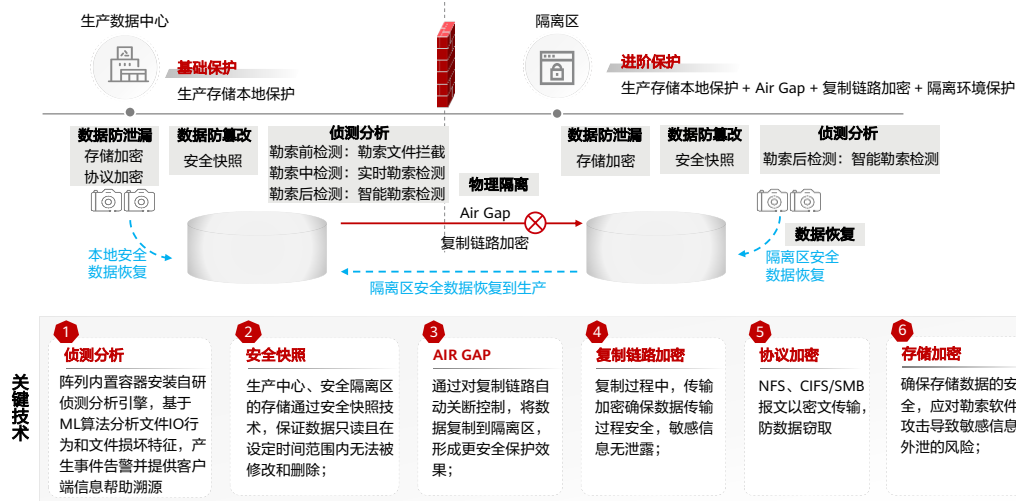
1. Air Gap技术构建一个物理隔离的安全保护区，我们称之为隔离区。在隔离区内，部署一套生产存储，用于生产数据的备份和保护。部署隔离区需要配置华为BCManager软件，该软件部署在隔离区内，控制生产存储到生产隔离区的链路的定时通断，在需要备份数据时，链路up，在备份完后，相应的端口down掉，实现物理上的隔离。Air Gap技术可有效保护隔离区不被黑客攻破。

2. 在生产存储和隔离区存储上部署安全快照，防止勒索攻击篡改价值数据。安全快照是一种带保护期的快照，在设置的保护期内，快照不能被修改或者删除。有了安全快照，客户不必担心副本数据被修改或删除。

3. 复制链路加密，在生产存储到生产隔离区的存储进行ipsec加密，保障数据不会在复制过程中被窃取。

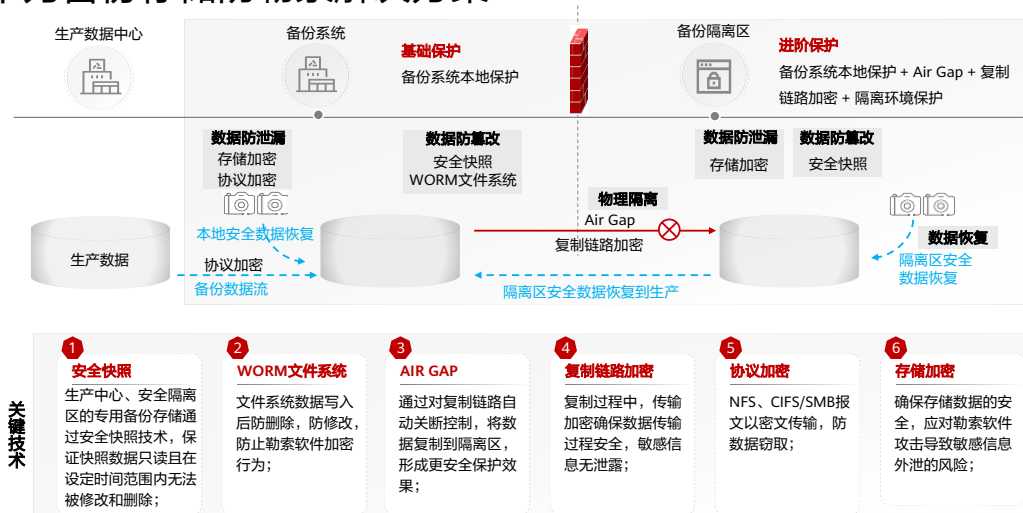
4. 存储加密，在生产存储、生产隔离区的存储进行加密，保障即使硬盘被盗也无法从存储中窃取数据。

华为NAS存储防勒索解决方案



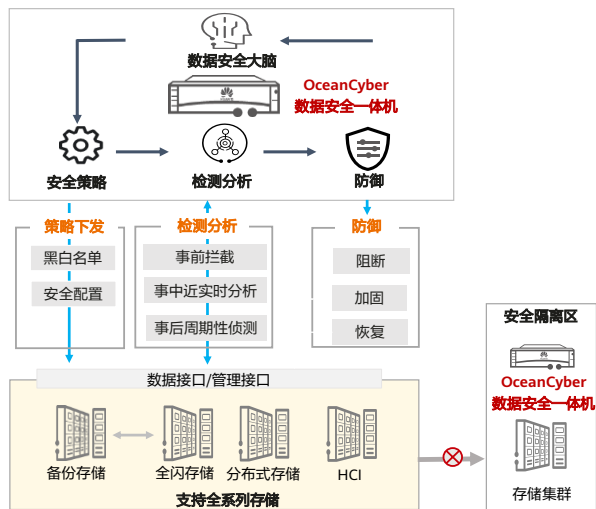
- NAS存储防勒索同样分为基础保护和进阶保护。NAS存储相对SNA存储多了侦测分析技术, 即勒索检测技术, 可以及时有效的检测到勒索病毒的攻击行为, 并主动触发对数据的保护行为。
- 侦测分析是华为开发的, 业界准确率第一的勒索检测技术。侦测分析可以部署在生产存储和生产隔离区的存储上。侦测分析总共分为三个阶段:
 - 事前 (勒索病毒发动攻击前) —勒索文件拦截, 可以识别勒索软件特征, 阻止勒索软件对数据的破坏行为, 达到拦截勒索攻击的作用;
 - 事中 (勒索病毒加密数据过程中) —实时勒索检测, 可以在线监测勒索病毒攻击的IO行为, 发现勒索攻击后, 会立即触发安全快照功能对检测到勒索攻击的文件系统进行主动保护, 减少数据损失;
 - 事后 (勒索病毒加密数据后) —智能勒索检测, 可以检测出数据副本是否被勒索病毒感染, 确保用于恢复的数据副本的绝对安全。
- 安全快照, Air Gap, 复制链路加密, 存储加密原理同SAN安全快照, 这里不再展开。
- 协议加密, 即NAS的访问协议, 包含NFS协议, CIFS协议或者SMB协议, 在服务器和存储前端传输数据时, 对数据进行加密传输, 保障数据完整性和安全性。

华为备份存储防勒索解决方案



- 备份存储需配合专用备份软件使用，专用备份软件读取应用数据并以文件形式写入备份存储。华为存储生态广泛，支持和业界主流的备份软件对接。
- 备份存储的防勒索技术大部分和NAS防勒索技术一致。
- 备份存储防勒索支持WORM文件系统防篡改技术，WORM文件系统是一个保护期内不可修改不可删除的文件系统，在专用备份软件支持和备份存储WORM联动的情况下，推荐使用WORM对文件进行保护。WORM的保护级别为文件级别。

华为数据安全一体机防勒索解决方案



全面防护

多类型存储设备统一接入：全闪存存储、混合闪存、分布式存储
全生命周期：生产存储→备份存储→归档

主动防御*

检测分析发现异常**自动执行防御策略**，
变**被动响应**为**主动防御**

检测识别率与性能业内领先

99.9%勒索识别率
支持多流并发检测，检测性能可达50TB/hr

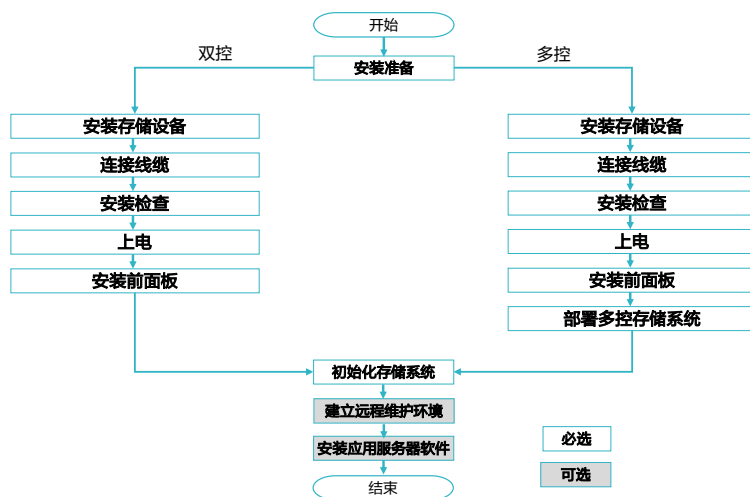
目录

1. 数据保护产品介绍
2. 数据保护产品应用场景
- 3. 数据保护产品部署**

规划工具介绍

工具名称	简介
能耗计算	提供存储产品能耗快速计算功能，用于规划阶段评估机房供电情况。 访问地址： https://info.support.huawei.com/storage/powercalc/#/home
OceanProtect兼容性查询	根据机房现场环境在线快速查询存储产品与业界主流厂商、主流设备的兼容性关系。 访问地址： https://info.support.huawei.com/storage/comp/#/oceanprotect
LLDesigner	提供硬件配置、设备组网、资源分配等功能，能快速生成LLD文档和存储设备的配置文件。主要功能如下所示： •支持BOQ一键式导入，自动识别存储设备，控制器，接口卡、硬盘框、硬盘类型以及数量，快速完成硬件信息规划。 •支持通过模板创建LLD，覆盖多控等典型组网场景。 •支持自动导出离线配置文件，通过DeviceManager导入到存储设备，完成设备基础业务配置。

专用备份存储开局安装



初始化存储系统流程



规划网络

规划以太网端口

- 以太网端口是指在设备上物理可见的端口。以太网端口是创建VLAN、绑定端口和逻辑端口的基础。

规划绑定端口

- 绑定端口可以增加链路的带宽和冗余。端口绑定后，每个会话连接依然通过单个端口进行数据传输，因此只有在有多个会话连接的情况下才可能提高总的带宽。请根据实际环境情况确定是否需要绑定端口。

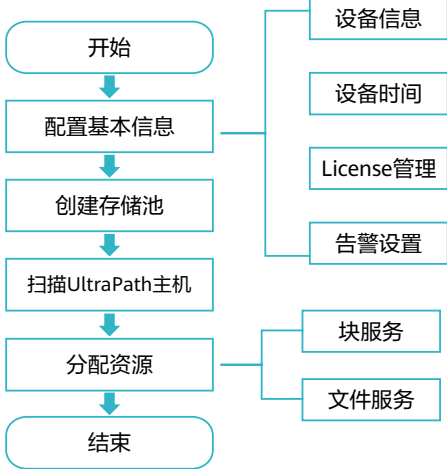
规划VLAN

- 存储系统上的以太网端口资源在逻辑上划分成多个广播域，在VLAN内发送或接收业务数据时，给数据加上VLAN ID，实现VLAN之间的网络隔离和业务隔离，从而确保业务数据的安全性和可靠性。
- VLAN的创建基于以太网端口或绑定端口，一个物理端口可以从属于多个VLAN。已经创建为绑定端口的以太网端口不能单独用来创建VLAN，只能通过其组成的绑定端口创建。

规划逻辑端口

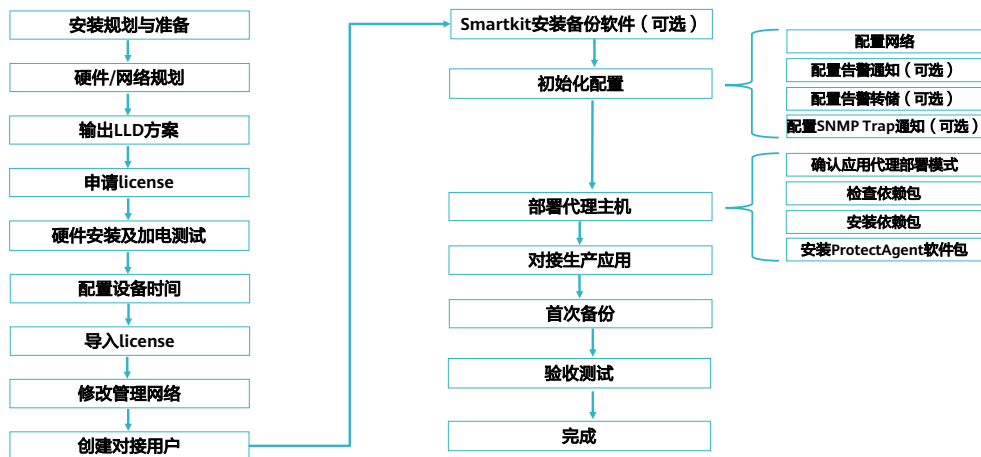
- 逻辑端口的创建基于以太网端口、绑定端口或VLAN，用于业务的运行。

初始化配置向导

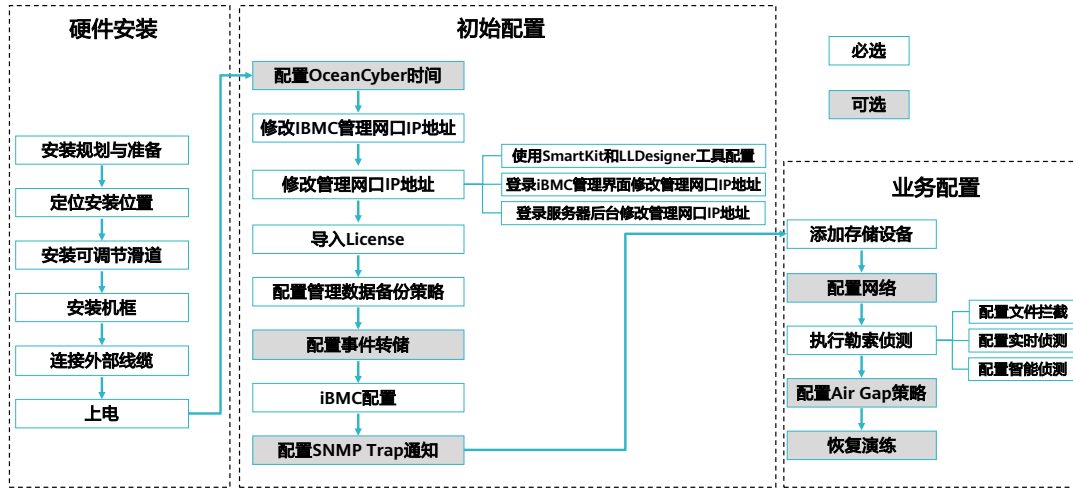


说明：不同产品版本和型号的操作界面可能会有差异，操作时请以产品文档为准。

备份一体机安装流程



OceanCyber安全一体机安装流程



思考题

1. (多选题) OceanProtect专用存储支持以下哪些数据重删技术? ()
- A. 在线变长重删
 - B. 数据无冲突重删
 - C. DIF重排
 - D. 漂移组

- 参考答案:

1. AB

思考题

2. (多选题)以下关于数据重删的描述，正确的是哪些项？（ ）
- A. 全局重删是指全局范围内共用一个指纹库，重删率高
 - B. 源端重删是指在备份数据传输到存储介质之前进行重删操作
 - C. 源端重删会增大备份窗口
 - D. 并行重删在多个不同节点上构建重删指纹库，有效解决存储空间压力问题

- 参考答案：

2. ABD

本章总结

- 本章节主要讲解OceanProtect专用备份存储，备份一体机和OceanCyber安全一体机的产品技术以及部署流程。

更多信息

- 华为数据存储信息中心: <https://info.support.huawei.com/storage/#/home>
- 企业业务: <https://e.huawei.com/cn/>
- 技术支持: <https://support.huawei.com/enterprise/>
- 在线学习: <https://e.huawei.com/cn/talent/portal/#/>
- 华为Support案例库: <https://support.huawei.com/enterprise/zh/knowledge?lang=zh>

Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。
Bring digital to every person, home, and
organization for a fully connected,
intelligent world.

版权所有 © 华为技术有限公司 2026。

保留一切权利。

未经华为事先书面许可，任何人不得对本资料进行复制、修改、
整合、改编、翻译，也不得将本资料或其任何部分或基于本资
料的衍生作品提供、出售、转让、再许可给任何人用于华为
许可以外的任何其他用途。

Copyright © Huawei Technologies Co., Ltd. 2026.

All Rights Reserved.

Without prior written permission of Huawei, no one is
allowed to reproduce, modify, integrate, adapt, translate
this Document, or provide, sell, transfer, sublicense this
Document or any part of this Document or its derivative to
any third party for any other purpose otherwise permitted
by Huawei.



分布式存储技术



前言

- 在当今数字化时代，数据量呈现爆炸式增长，传统的集中式存储架构在面对海量数据存储、高并发访问以及高可用性需求时，逐渐暴露出性能瓶颈和可扩展性问题。分布式存储作为一种数据存储技术，通过将数据分散存储在多个独立的物理节点上，利用网络通信实现数据的共享和透明访问，为解决这些问题提供了有效的解决方案。
- 本章主要讲述分布式存储技术，包括Ceph相关知识，华为分布式存储，分布式存储部署以及产品特性。

目标

- 学完本课程后，您将能够：
 - 理解Ceph相关技术
 - 熟悉华为分布式存储产品
 - 熟悉分布式存储部署
 - 理解分布式存储特性

目录

1. Ceph

▪ Ceph架构

- Ceph块设备
- Ceph文件存储
- Ceph对象网关
- Ceph部署

2. 分布式存储产品介绍

3. 分布式存储产品部署

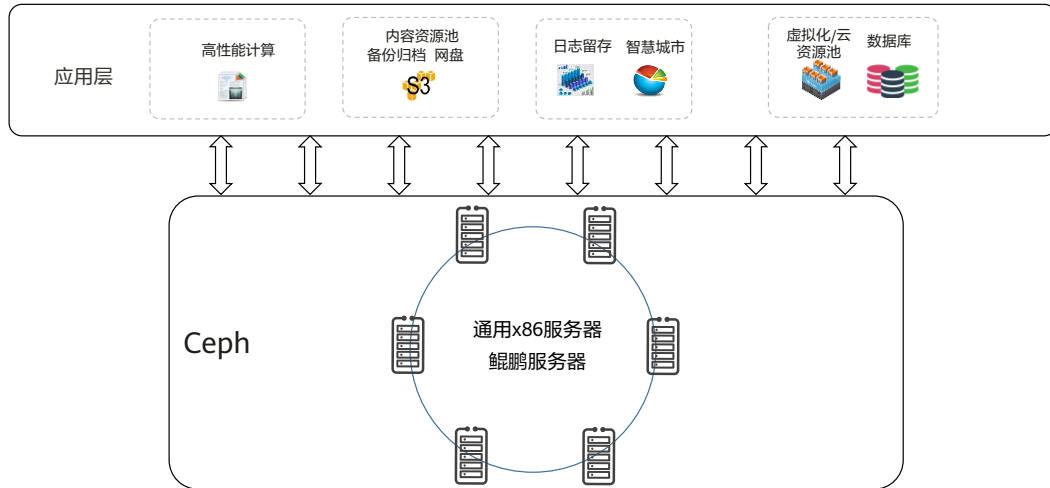
4. 分布式存储特性

Ceph简介

- Ceph是一款开源的SDS分布式存储，它具备极高的可用性、扩展性和易用性，可用于存储海量数据。
- Ceph的存储节点可部署在通用服务器上，这些服务器的CPU可以是x86架构的，也可以是ARM架构的。Ceph存储节点之间相互通信，以便对存储的数据进行动态复制和重分配。
- 采用RADOS系统将所有数据作为对象，存储在存储池中。
- 使用CRUSH算法计算出对象所对应的存储位置。
- 去中心化，客户端可根据CRUSH算法自行计算出对象存储位置，然后进行数据读写。
- 集群可自动进行扩展、数据再平衡、数据恢复等。

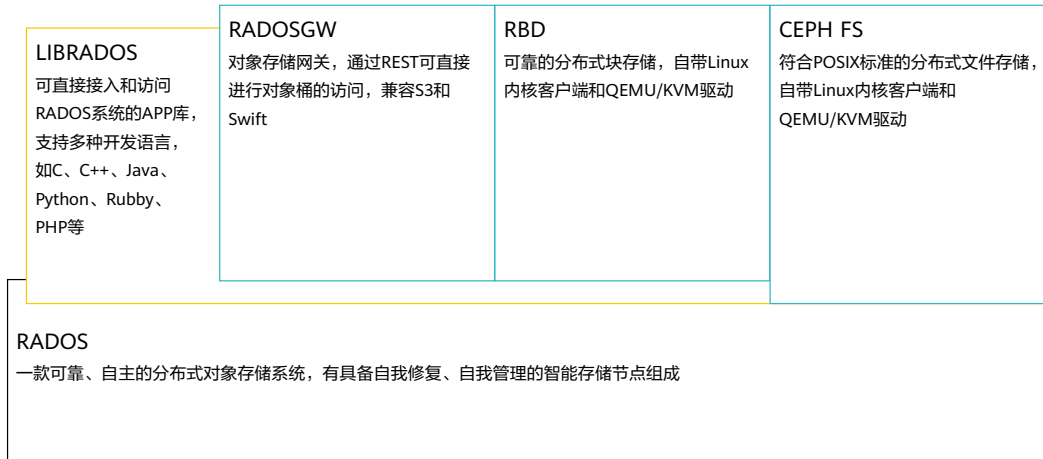
- SDS: software-defined storage, 软件定义存储
- 华为的鲲鹏芯片是典型的ARM架构CPU，Ceph可部署在基于鲲鹏芯片的泰山服务器中

Ceph的使用场景



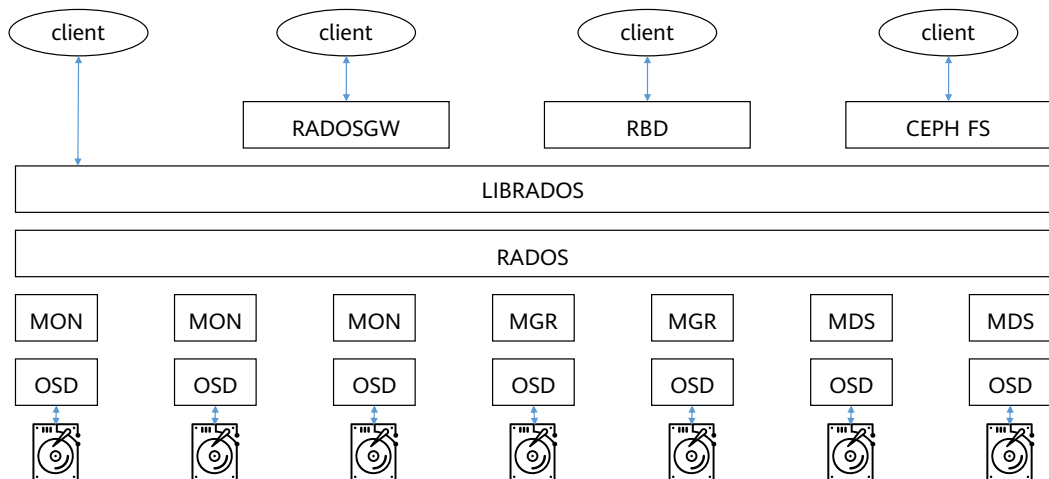
- Ceph在统一系统中独特地提供了对象、块和文件存储。其具有高度可靠性、易于管理且免费。Ceph具备卓越的可扩展性，支持数千个客户端访问从PB到EB级别的数据。Ceph节点采用商用硬件和智能守护程序，而Ceph存储集群则容纳大量节点。这些节点通过互相通信实现动态数据复制和重新分配。

Ceph的架构



- RADOS的全称为**Reliable Autonomic Distributed Object Store**，是Ceph的基础，Ceph提供了一个基于RADOS的无限可扩展Ceph存储集群，RADOS是一种可靠的分布式存储服务，它利用每个节点中的智能来保护其存储的数据并向 客户端 提供这些数据。
- Ceph支持多种接入方式：
 - LIBRADOS：RADOS库
 - RADOSGW：RADOS gateway，对象存储网关，对外提供对象存储
 - RBD：RADOS block device，RADOS块设备，对外提供块存储
 - CEPH FS：可对外提供文件存储

Ceph的组件总览



- MON、MGR、MDS和OSD属于Ceph的后端存储组件
- RADOSGW、RBD、CEPHFS及LIBRADOS属于Ceph的访问方式
- Ceph Monitors维护集群映射的主副本，并将其提供给Ceph客户端。Ceph集群中存在多个Monitor可确保在其中一个Monitor守护程序或其主机发生故障时仍可继续使用。
- Ceph OSD Daemon检查自身的以及其他OSD的状态，并向Monitor报告。
- Ceph Manager用作监视、编排和插件模块的端点。
- Ceph Metadata Server (MDS)在使用CephFS提供文件服务时管理文件元数据。

Ceph组件简介 - Monitor

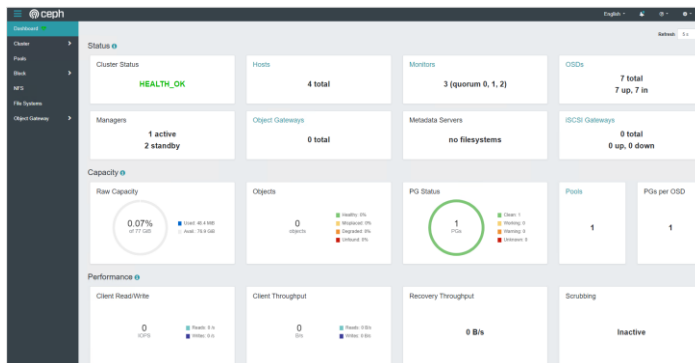
- Monitor用于维护Ceph集群的“Cluster Map”，Ceph各组件之间的交互配合需要根据“Cluster Map”进行。
- “Cluster Map”可使Ceph客户端找到所有的ceph Monitor、OSD、Ceph元数据服务器，Cluster Map包括：
 - Monitor Map
 - OSD Map
 - PG Map
 - CRUSH Map
 - MDS Map
- 为了保证Monitor的高可用性，Monitor一般会以集群的方式部署，并通过一定的算法和多数服从少数的方式来保证Monitor之间的数据一致性。

- Monitor采用多数服从少数的形式来保证数据的一致性，例如3台中的2台、5台中的3台、6台中的4台等等
- Monitor对应的进程一般会被称为mon
- 在生产中，MON的节点数一般为单数，来保证投票仲裁机制
- Monitor map包含以下信息：
 - Cluster 的fsid。fsid是一个自动产生的、唯一的用于区别Ceph集群的uuid
 - 每一个mon的位置、名称、地址及端口号
 - Map的时间戳
 - 使用命令ceph mon dump可以查看monitor map的信息
- OSD map包含以下信息：
 - Cluster 的fsid
 - 存储池（pool）清单
 - 副本数量
 - PG数量
 - OSD清单及其状态
 - Map的时间戳
 - 使用命令ceph osd dump可以查看osd map的信息

- PG map包括：
 - PG的版本
 - PG的空间占用率
 - PG的详细信息，如PG ID、PG的状态、资源池的数据利用率统计等
 - Map的时间戳
 - 使用命令`ceph pg dump`可以查看pg map的信息
- CRUSH map包括：
 - 存储设备清单
 - 故障域级别（设备级、主机级、机柜级、机柜列级、机房级）
 - 数据恢复规则
 - 使用命令`ceph osd crush`可以查看CRUSH map的信息
- MDS map包括：
 - 元数据存储池信息
 - Metadata server的清单
 - Metadata server的状态
 - Map的时间戳
 - 使用命令`ceph fs dump`可以查看fs map的信息

Ceph组件简介 - Manager

- Ceph的Manager进程（ceph-mgr）用于查看Ceph集群的运行状态，可以提供图形化配置页面，并可以对接第三方的管理插件，如Prometheus、Zabbix等。
- 数据的读写不依赖Manager，但无法对集群的相关数据进行查询，建议每个集群至少配置两个Manager。



Ceph组件简介 - MDS

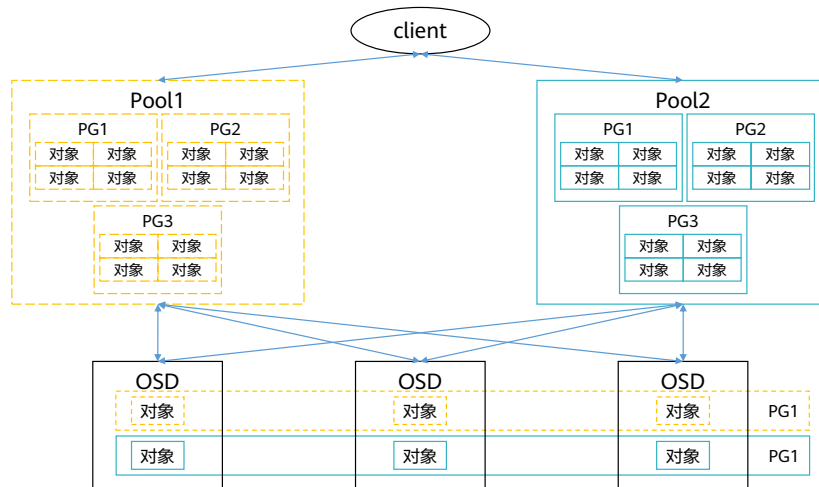
- MDS全称为“MetaData Server”，也被称为ceph-mds，当运行了CephFS时，MDS必须运行，它用来保存文件系统的元数据。
- MDS使用RADOS来保存元数据，因此，在MDS配置时，系统会要求创建两个存储池，分别是元数据存储池和数据存储池。
- MDS会将元数据加载到内存中，以加速文件系统的访问。

Ceph组件简介 - OSD

- OSD的全称为Ceph Object Storage Daemon，它是Ceph存储集群的构建单元、数据的存储位置。Ceph存储节点运行多个OSD，OSD将数据保存在存储节点的裸硬盘中，通常情况下，每个OSD对应一个裸硬盘。
- 为了使Ceph的性能尽可能的接近物理设备，Ceph使用CRUSH算法将对象存储到OSD中。
- CRUSH的全称为Controlled Replication Under Scalable Hashing，用于计算对象的存储位置。CRUSH将对象和PG进行关联，通过伪随机的算法将对象分散在不同的PG中，并计算出OSD和PG之间的映射。一旦出现故障，Ceph将PG重新映射到不同的OSD，同时将根据配置的数据保护策略进行数据同步。
- Ceph采用副本机制，将数据保存在不同的OSD中，用于保证数据的安全性。仅有一个OSD会被选为主OSD，其他OSD为从OSD，主OSD和从OSD在Ceph中有着不同的功能，其中客户端进行数据读写时只会和主OSD进行通信。
- OSD有四种状态：
 - Up且in：表示OSD运行正常且已承载了至少一个PG；
 - Up且out：表示OSD运行正常，但未承载任何一个PG，新加入集群的OSD为此状态；
 - Down且in：表示OSD运行异常，但承载了至少一个PG，该状态下的OSD采取一定措施，可能会恢复到up且in状态；
 - Down且out：表示OSD运行异常，且不再承载任何PG。

- 早期的OSD既可以通过BlueStore将数据保存在裸盘中，也可以通过FileStore保存在非裸盘的文件系统中，但是从Ceph的Luminous版本开始，BlueStore作为默认的存储方式，并从Reef版本后，不再支持FileStore，因此目前数据仅能被保存在裸盘中，这样可带来接近裸设备的性能
- 主OSD的功能包括：
 - 响应所有的IO请求
 - 数据的复制和保护
 - 检查数据的一致性
 - 数据重平衡和恢复
- 从OSD的功能包括：
 - 接收主OSD的指令
 - 主OSD故障后，升级为主OSD

Ceph的数据存储机制



- Ceph存储集群从Ceph客户端 接收数据——无论是通过Ceph块设备、Ceph对象存储、Ceph文件系统，还是通过使用librados创建的自定义实现。Ceph存储集群接收到的数据以RADOS对象的格式存储。每个对象都存储在Object Storage Device（也称为“OSD”）上。Ceph OSD控制存储驱动器上的读取、写入和复制操作。默认的BlueStore后端以单片、类似数据库的方式存储对象。

目录

1. Ceph

- Ceph架构
- **Ceph块设备**
- Ceph文件系统
- Ceph对象网关
- Ceph部署

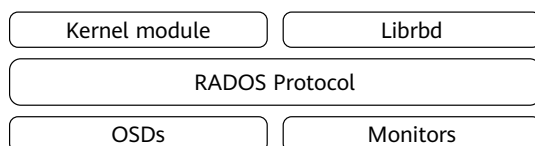
2. 分布式存储产品介绍

3. 分布式存储产品部署

4. 分布式存储特性

Ceph块存储简介

- Ceph块存储功能基于RADOS构建，经过精简配置、大小调整后，在多个OSD上实现存储数据条带化。Ceph块存储可实现快照、复制和存储的强一致性。
- 应用访问块存储通过两种方式实现：
 - Librbd用户态接口：librados.so库作为客户端连接RADOS集群。
 - Krbd内核态接口：通过rbd命令将块设备映射为主机设备，可挂载至指定目录。



- Krbd用于linux操作系统直接挂载Ceph块设备。
- Librbd大多会用到虚拟化、云平台等场景。

RBD设备特性

- 在创建RBD设备时，使用参数 “--image-feature” 可开启对应特性，RBD设备支持的特性有：
 - layering：克隆的分层
 - Striping：可提高性能的条带 v2版本
 - exclusive-lock：独占锁
 - object-map：对象映射，依赖exclusive-lock特性
 - fast-diff：快速差异计算，依赖object-map特性
 - deep-flatten：快照的扁平化
 - Journaling：日志功能，依赖exclusive-lock特性
 - data-pool：纠删码池支持

RBD设备特性配置

- 使用命令“`rbid info`”可查看到RBD设备开启的特性。

```
[ceph: root@ceph01 /]# rbd info test-rbd/rbd-6G
rbd image 'rbd-6G':
.....
features: layering, exclusive-lock, object-map, fast-diff, deep-flatten
```

- RBD设备创建会后开启或关闭某个特性的命令为：`rbd feature enable/disable [pool_name/] image-name feature-name`。

```
[ceph: root@ceph01 /]# rbd feature enable test-rbd/rbd-6G journaling
[ceph: root@ceph01 /]# rbd info test-rbd/rbd-6G
rbd image 'rbd-6G':
  size 6 GiB in 1536 objects
  order 22 (4 MiB objects)
  snapshot_count: 0
  id: 172505c22d28b
  block_name_prefix: rbd_data.172505c22d28b
  format: 2
  features: layering, exclusive-lock, object-map, fast-diff, deep-flatten, journaling
  op_features:
.....
```

RBD客户端缓存机制

- 当客户端使用librbd调用Ceph块存储时，应用的数据需先写入到挂载到所在操作系统的块设备（ rbd设备 ）上，在通过librbd与RADOS交互，最终写到OSD中。如果OSD有延时，则应用等待写入完成响应的时间就变长，从而造成应用的运行卡顿，因此需借助RBD客户端的缓存机制避免此类问题。
- RBD客户端的缓存机制有两种模式：
 - 直写模式
 - 回写模式
- RBD客户端缓存机制所使用的缓存空间由客户端所在主机的内存提供。

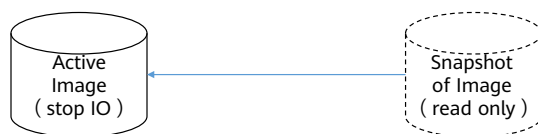
- RBD客户端缓存机制主要用于使用librbd调用Ceph块存储的场景中
- 直写模式也叫透写模式，是数据直接被写入到OSD磁盘中
- 回写模式是由librbd将数据先写入到服务器的本地缓存中，再周期性刷盘到OSD磁盘中
- 回写模式可提高Ceph块设备的读写性能，但是有数据丢失的风险

RBD客户端缓存配置参数

参数	说明	默认值
rbd_cache	开启或关闭缓存功能，true表示开启，false表示关闭	True
rbd_cache_policy	选择缓存机制，可配置writeback（绕写）和writethrough（直写）和writeback（回写）	writearound
rbd_cache_size	为每个RBD设备分配的缓存空间大小，用于直写和回写模式	30M
rbd_cache_max_dirty	每个RBD设备的最大脏数据大小，该值需小于rbd_cache_size，用于直写模式	24M
rbd_cache_target_dirty	每个RBD设备开始清除脏数据时的缓存大小，该值需小于rbd_cache_max_dirty	16M
rbd_cache_max_dirty_age	清空缓存脏数据的期限，单位为秒	1

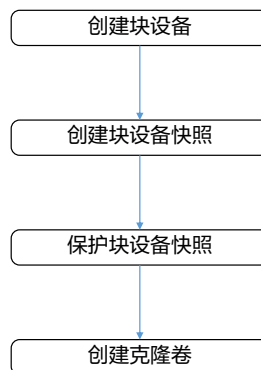
RBD设备的快照

- RBD设备的快照是RBD设备生命周期中某个时间点被创建的只读副本。
- RBD设备快照支持分层特性，可基于快照快速克隆出一个新的RBD设备。
- RBD设备的快照功能可通过rbd命令实现，也可以由上层应用（QEMU、libvirt、OpenStack等）的接口调用实现。
- 在创建快照时，建议暂停或停止RBD设备的IO读写。



快照克隆卷的创建流程

- 创建快照卷的流程如下：
 - 1、创建RBD块设备
 - 2、为RBD块设备创建快照
 - 3、为了保证克隆卷的可用性，需对块设备快照进行保护
 - 4、基于RBD设备的快照创建克隆卷



目录

1. Ceph

- Ceph架构
- Ceph块存储
- **Ceph文件系统**
- Ceph对象网关
- Ceph部署

2. 分布式存储产品介绍

3. 分布式存储产品部署

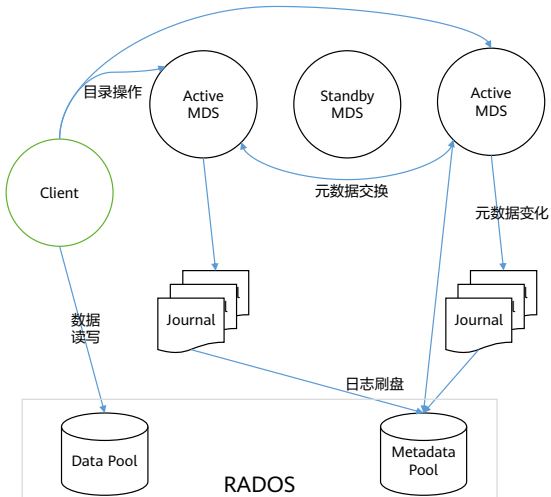
4. 分布式存储特性

Ceph文件系统简介

- Ceph文件系统（CephFS）是基于Ceph分布式存储RADOS构建的兼容POSIX的文件系统。
- CephFS致力于提供多用户、高可用、高性能的先进文件存储。
- CephFS支持的场景包括共享主目录、高性能计算暂存空间及工作流共享存储等。
- CephFS需要在集群中运行MDS服务，其守护进程为ceph-mds，用于管理CephFS中所存储文件的元数据，及协调对Ceph集群的访问。
- CephFS通过ceph-mds将元数据与数据分开，可降低复杂性并提高可靠性。

Ceph文件系统工作原理

- MDS管理着CephFS客户端的元数据，ceph-mds为客户端接入提供相关信息，如文件在文件系统结构中的具体位置等。
- MDS还管理着目录结构及文件的元数据，如所属用户、时间戳、权限等。
- MDS同样负责响应缓存的接入和管理，来维护缓存一致性。
- MDS有两个状态：
 - 活动：负责CephFS的元数据管理
 - 备用：当活动MDS无法响应时转变为活动状态
- CephFS在数据访问时，会首先通过MON的验证，并获取到集群映射信息，然后从MDS请求到相关的元数据，并通过请求到的元数据直接与OSD进行通信。



- Ceph支持MDS多活部署，但该功能处于beta状态，不建议应用于生产系统

目录

1. Ceph

- Ceph架构
- Ceph块存储
- Ceph文件系统
- **Ceph对象网关**
- Ceph部署

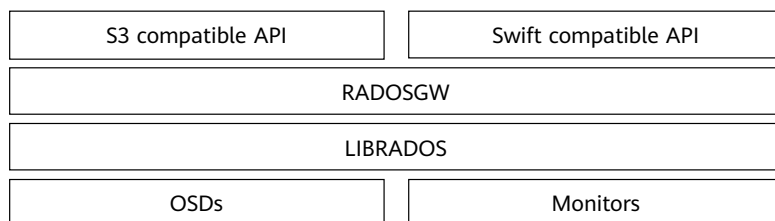
2. 分布式存储产品介绍

3. 分布式存储产品部署

4. 分布式存储特性

对象存储网关简介

- Ceph的对象存储服务可直接开发应用调用Rados实现，也可以通过RGW（Rados gateway）实现
- RGW是建立在librados上一个对象存储接口，它为应用访问Ceph存储集群提供了RESTful网关，当前支持两种接口：
 - S3：提供对象存储功能，其接口兼容Amazon S3 RESTful API的大部分子集。
 - Swift：提供对象存储功能，其接口兼容OpenStack Swift API的大部分子集。



- Ceph对象存储使用Ceph对象网关守护程序（radosgw），这是一个旨在与Ceph存储集群交互的HTTP服务器。Ceph对象网关提供与Amazon S3和OpenStack Swift兼容的接口，并拥有自己的用户管理。Ceph对象网关可以使用单个Ceph存储集群来存储来自Ceph文件系统和Ceph块设备客户端的数据。S3 API和Swift API共享一个共同的命名空间，这意味着可以使用一个API将数据写入Ceph存储集群，然后使用另一个API检索该数据。

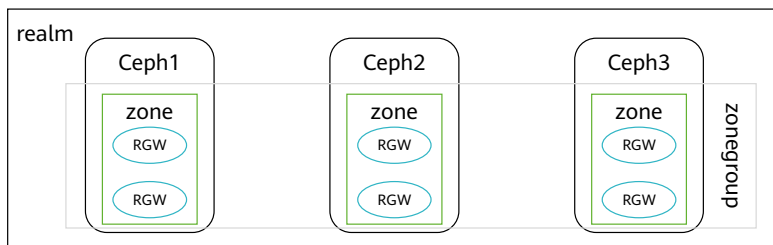
Ceph对象存储网关所涉概念

- Realm

- 在RGW中，realm是一个全局的、独一无二的、由zonegroup组成的命名空间，每个realm都有对应的period，在进行配置修改时，需更新并提交period，使配置生效。

- Zone

- 在RGW中，zone是一个由一个或者多个RGW实例组成的逻辑概念，zone的配置保存在period中。多个zone可组成zonegroup，用于多站点服务。



对象存储网关多站点部署简介

- 通过对象存储网关多站点部署可以实现Ceph容灾功能。
- 在对象存储网关多站点部署的方案中：
 - 一个zone可部署多个RGW，一个Ceph集群中可以部署多个zone，但zone不可以跨Ceph集群。
 - Zonegroup由一个或多个zone组成，且其中一个为master zone，剩余为secondary zone，zonegroup内的zone之间进行数据同步。
 - Realm由一个或多个zonegroup组成，其中一个为master zonegroup，其余为secondary zonegroup，所有的zonegroup共享realm的全局配置，realm中的所有RGW都为与master zonegroup和master zone的RGW拉取配置。

realm

zonegroup3

zonegroup2

zonegroup1

Cluster1

master zone

RGW RGW

mon mon mon

osd osd osd

Cluster2

secondary zone

RGW RGW

mon mon mon

osd osd osd

对象存储网关多站点架构

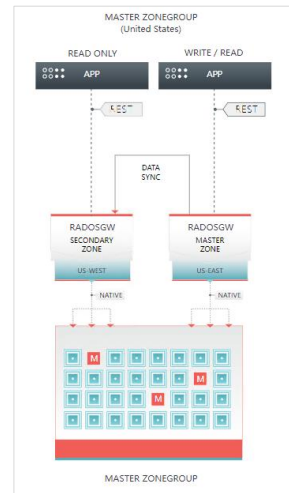
- 对象存储网关多站点部署有多种实现架构：
 - Single-zone: 一个realm中包含一个zonegroup和zone，一个zone中部署多个RGW。
 - Multi-zone: 一个realm中包含一个zonegroup，该zonegroup中包含多个zone。
 - Multi-zonegroup: 一个realm中包含多个zonegroup，每个zonegroup中包含一个或者多个zone。
 - Multi-realms: 部署多个realm。

Periods和Epochs的版本管理

- 在Ceph中，period用于记录和跟踪zone、zonegroup及realm的配置状态，而epoch用于记录和跟踪zone period配置变更的版本号。
- 每个zone都有与之关联的period，每个period都会关联对应的epoch。
- 每个period对有唯一ID，并包含着zone的当前配置，及上一版本period的ID。
- 在更新zone的配置时，需更新zone当前的period。

对象存储多区域同步流程

- RGW在所有的zonegroup之间同步元数据和数据。
- 多区域配置完成后，RGW会在master zone和非master zone之间执行一次完全同步，后续执行增量更新同步。
- 当用户将数据写入到zonegroup的某个zone时，该zone会将数据同步给同zonegroup的其他zone。
- 当RGW之间同步数据时，所有处于active状态的网关会更新数据日志并通知其他网关。
- 当RGW因用户操作而进行元数据同步时，主网关会更新元数据日志并通知其他RGW。



- 图片来源：<https://docs.ceph.com/en/quincy/radosgw/multisite/#delete-default-zonegroup-and-zone>

目录

1. Ceph

- Ceph架构
- Ceph块存储
- Ceph文件系统
- Ceph对象网关
- **Ceph部署**

2. 分布式存储产品介绍

3. 分布式存储产品部署

4. 分布式存储特性

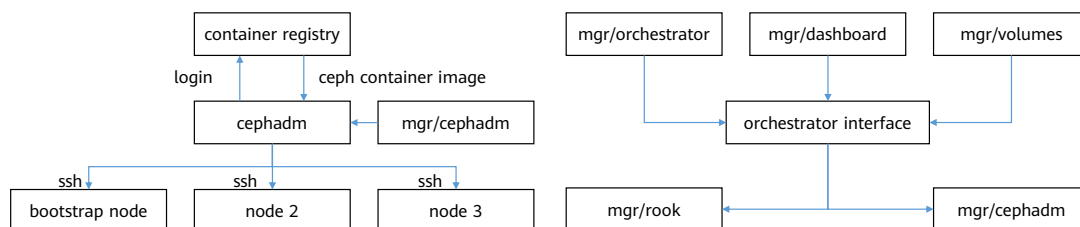
Ceph的安装方式简介

- 官方推荐部署方式
 - Cephadm：基于容器部署，支持Octopus及以后版本，部署完成后可通过命令行和图形界面进行Ceph集群的管理。
 - Rook：基于kubernetes部署，支持Nautilus及以后版本，部署完成后可通过命令行和图形界面进行Ceph集群的管理，同时可通过kubernetes对组件所在pod进行管理。
- 其他部署方式
 - ceph-ansible
 - ceph-deploy
 - DeepSea
 - 手动部署

- ceph-ansible通过ansible进行Ceph部署
- Ceph-deploy已经不再维护，不建议使用该方式部署
- DeepSea通过salt进行Ceph部署

Cephadm简介

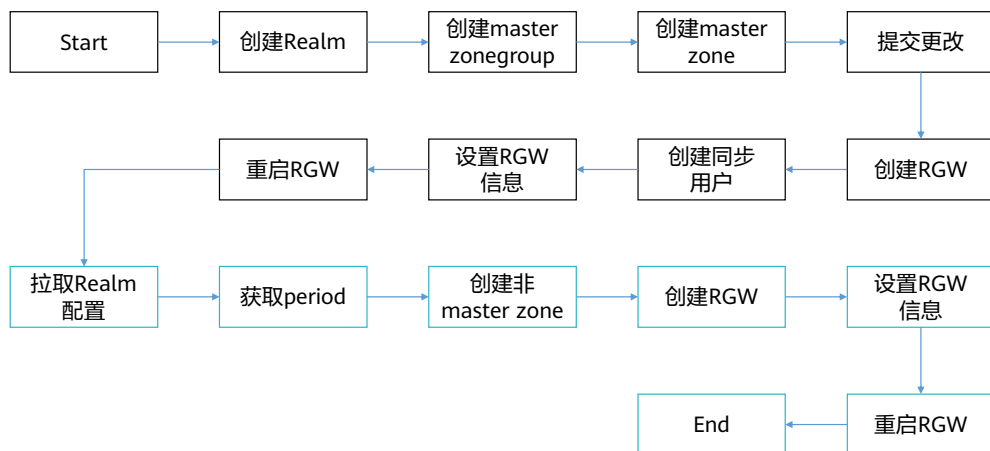
- Cephadm是一个Ceph全生命周期管理工具，通过“引导（ bootstrapping ）”可创建一个包含一个MON和一个MGR的单节点集群，后续可通过自带的编排接口进行集群的扩容、主机添加、服务部署。
- Cephadm的主要功能包括：
 - 添加基于容器的Ceph集群
 - 从Ceph集群中移出Ceph容器
 - 更新Ceph容器



使用cephadm部署Ceph集群流程

- 安装前准备工作
- 安装cephadm
- 下载ceph yum源文件
- 新创Ceph集群
- 添加node节点
- 添加组件

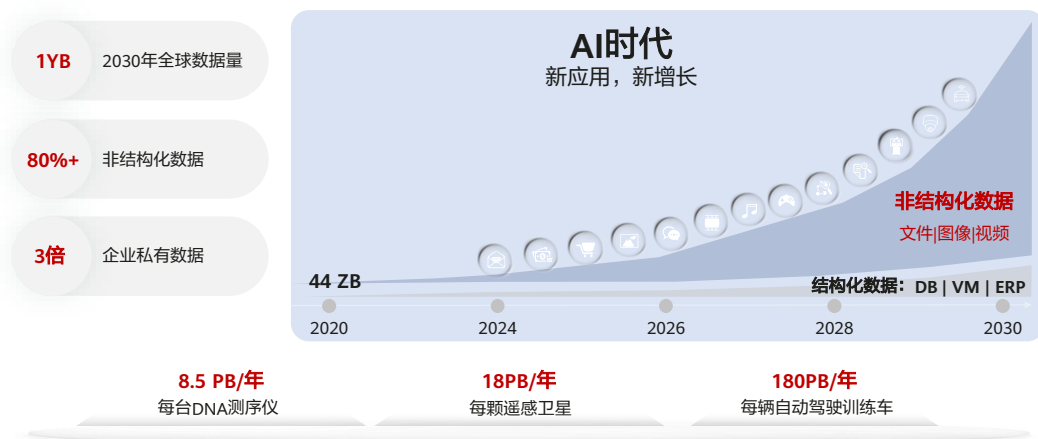
多站点架构配置流程



目录

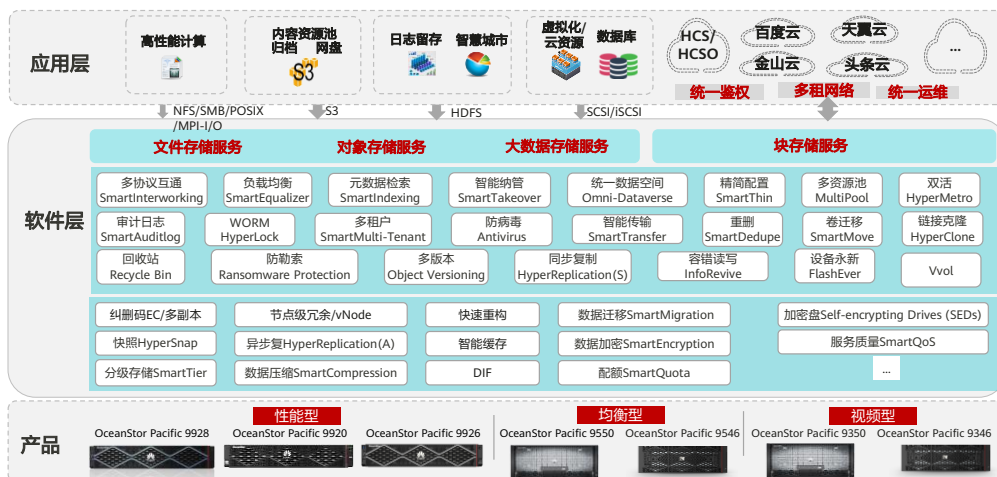
1. Ceph
2. 分布式存储产品介绍
 - 产品架构
 - 关键技术
 - 应用场景
3. 分布式存储产品部署
4. 分布式存储特性

新应用驱动数据爆炸式增长



- 随着云、互联网及智能技术的兴起与普及，企业新兴应用层出不穷，数据快速增长，业务加速创新对于存储能力的快速迭代提出了更高要求。例如，在金融行业，银行要抓住互联网、特别是移动互联网金融 崛起带来的机遇，同时也不得不迎接由此带来的挑战：新业务小时级快速上线，以及更精准的用户需求分析等；在智慧城市、智能制造、电信运营商等领域中，随着新兴业务的激增，业务数据呈现几何倍数的快速增长。企业业务的发展，使得企业数据中心存储系统开始面临新的挑战。

OceanStor Pacific系列分布式存储

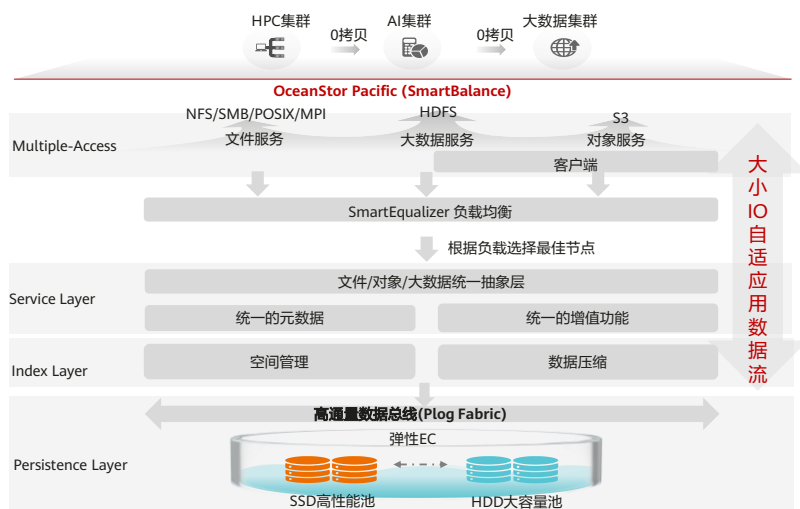


- 华为OceanStor Pacific系列正是一款可大规模横向扩展、弹性伸缩的数据中心级智能分布式存储产品。OceanStor Pacific系列通过系统软件将存储节点的本地存储资源组织起来构建分布式存储池，可向上层应用提供分布式文件存储服务、分布式对象存储服务、分布式大数据存储服务与分布式块存储服务，以及丰富的业务功能和增值特性；其独有的弹性Erasure Coding（以下简称EC）技术，能够在保证业务性能前提下为客户提供更多可用空间。产品支持用户根据业务需要灵活购买和部署存储服务，帮助企业轻松应对业务快速变化，实现数据灵活、高效、按需存取能力。

硬件架构

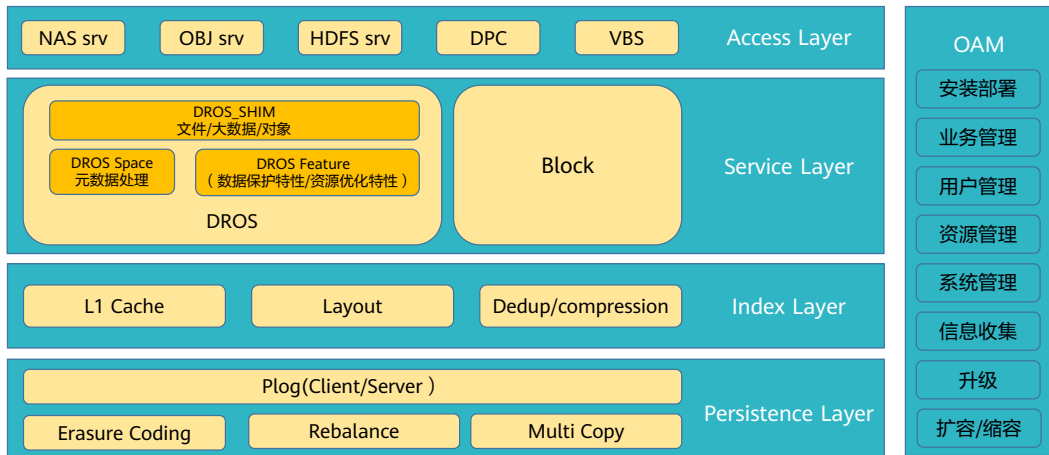
- 3D体验中心提供全新的硬件3D仿真交互，支持全方位浏览硬件模型并手动拆卸部件，了解各模块功能，并提供产品关联的文档、工具、多媒体、软件和成功案例等资源。您可以通过如下方式进行访问：
- 1. 访问华为数据存储信息中心（<https://info.support.huawei.com/storageinfo/#/home>）。
- 2. 在首页菜单“信息速查”页签，单击“3D体验中心”。
- 3. 在分布式存储栏目下，选择“OceanStor Pacific”。
- 4. 单击对应的产品型号。
- 5. 选择所需查看的物理设备，进入展示页面。

基于SmartBalance的全对称分布式架构

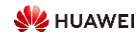


- OceanStor Pacific系列采用全对称分布式架构，支持通过横向扩展硬件节点线性增加整系统容量与性能，无需复杂的资源需求规划；系统可轻松扩展至数千节点及EB级容量，满足用户的业务规模增长需求。系统提供自动负载均衡策略，数据与元数据均匀分布于各节点，消除元数据访问瓶颈，保障规模扩展场景下的系统性能。系统通过智能分条聚合、IO优先级智能调度等 FlashLink技术，配合多级 Cache，大数据直通等系列关键技术，为用户提供高带宽，低时延的极致性能。无论您的数据中心在未来需要扩展I/O密集型、时延敏感型、大带宽或大容量需求业务，OceanStor Pacific系列提供的分布式存储都可以应需承载。

软件逻辑架构



42 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



- OceanStor Pacific系列是一款可大规模横向扩展的智能分布式存储产品，可提供块服务、文件服务、对象服务和大数据服务，用户可以按需部署相应的存储服务。如上图所示，OAM(Operation Administrator and Maintenance)提供OceanStor Pacific系列产品的管控面服务，包括资源管理、业务管理、系统管理、用户管理、安装部署、升级、扩容/缩容、巡检/信息收集等。IO面在逻辑架构上从上往下分为：Access Layer，Service Layer，Index Layer和Persistence Layer。
- Access Layer负责接收并处理各种服务的访问，对于非结构化服务，主要有以下功能模块：
 - NAS srv模块负责提供标准的NAS协议(NFS/SMB)接入处理。
 - OBJ srv模块负责提供标准的S3对象访问接入处理。
 - HDFS srv模块负责提供标准的HDFS协议接入访问处理。
 - DPC模块是针对HPC场景设计的兼容标准POSIX/MPI-IO语义访问。
 - 对于结构化服务，VBS(Virtual Block Service)提供标准的iSCSI协议访问，以及为高性能场景提供的本地SCSI语义访问。
- Service Layer负责提供非结构化服务与结构化服务功能，所有非结构化服务处理统一到DROS(Distributed Relational Object Service)平台，共享一份元数据，一份数据，共享增值特性。结构化服务由单独的Block模块进行处理。
- 在DROS模块内，DROS_SHIM是语义抽象层，针对文件/大数据/对象三种非结构化服务语义进行统一抽象，上层不同的非结构化语义接入后会把相应的语义操作映射到

DROS_SHIM提供的功能接口上。DROS_Space提供基础的非结构化服务的元数据处理，确保所有非结构化服务共一份元数据。DROS_Feature为非结构化服务提供统一的增值服务，如配额，分级，快照等。

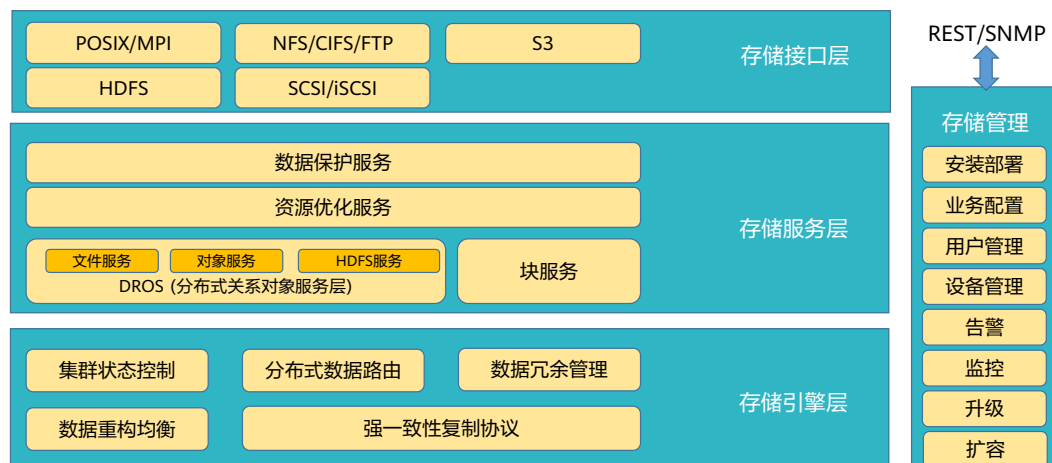
- Index Layer 在Persistence Layer之上提供细粒度的空间布局（Layout）管理能力。Index模块负责将DROS或block中的数据对象映射到下层的Plog空间上。对于非结构化服务提供了基于保电内存的缓存能力。Dedup/compression模块为结构化服务提供重删压缩能力节省成本。
- Persistence Layer负责HDD、SSD等存储介质的空间分配与故障管理，通过跨节点的EC或副本提供可靠的、分布式化的数据管理能力。Persistence Layer聚焦于可靠性、性能、扩展性等存储底层相关的能力。向上提供基于PLOG（persistence log）语义的空间访问能力。
- PLOG（Persistence LOG）是Persistence Layer提供的一组按照固定大小管理的物理地址的集合，提供EC或多副本的冗余保护能力，只能追加写，不能覆盖写，通过PlogID+Offset的方式来访问具体的物理磁盘上的LBA地址。

微服务形态

微服务包名称	功能描述	对应的逻辑功能架构元素
DPC(Distributed parallel client)	并行客户端，部署在计算节点，提供标准posix与MPI-IO访问。	DPC、plog client
VBS(Virtual Block Service)	块客户端，可以提供标准的iscsi/scsi访问，可以部署在计算节点，也可以部署在存储节点。	VBS
EDS(Enterprise Data Service)	企业数据服务，提供结构化服务及相应的增值特性。	Block、index layer、plog client
CEDS(Converged Enterprise Data Service)	融合企业数据服务，提供融合的非结构化服务。	NAS srv、OBJ srv、HDFS srv、DROS、index layer、plog client
OSD(Object Storage Device)	提供数据持久化能力。	persistent layer

- 前面介绍了OceanStor Pacific产品的软件逻辑功能架构，整个产品软件是微服务形态发布，以上的软件架构元素，会构建在对应的微服务包内。

功能架构



44 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



- 华为OceanStor Pacific系列提供基于三层的分布式存储架构，融合分布式文件、对象、大数据和块多个服务形态，支持文件、对象、大数据服务部署在一个集群，并统一管理。
- 华为OceanStor Pacific系列整体功能架构由存储接口层、存储服务层、存储引擎层和存储管理组成，如图所示。
- 存储接口层：提供应用访问存储系统的标准访问接口，支持NFS/CIFS/FTP、POSIX（Portable Operating System Interface）/MPI（Message Passing Interface）、S3、HDFS，以及SCSI/iSCSI接口协议。其中NFS、CIFS、POSIX、MPI、S3和HDFS协议可互通访问。
- 存储服务层：提供基于文件、对象、大数据和块的存储服务，以及HyperReplication、HyperMetro、HyperSnap、SmartQuota、SmartQoS、SmartTier、SmartEqualizer、SmartCache、SmartIndexing等高级特性。
- 存储引擎层：提供持久化存储能力，采用Plog接口（一种Append Only的ROW写机制），提供Erasure Coding、数据均衡与重构等能力，并对盘进行管理和数据读写。
- 存储管理：负责系统的操作、管理和维护，提供安装部署、业务配置、用户管理、设备管理、监控、告警、升级、扩容等功能。

全局负载均衡

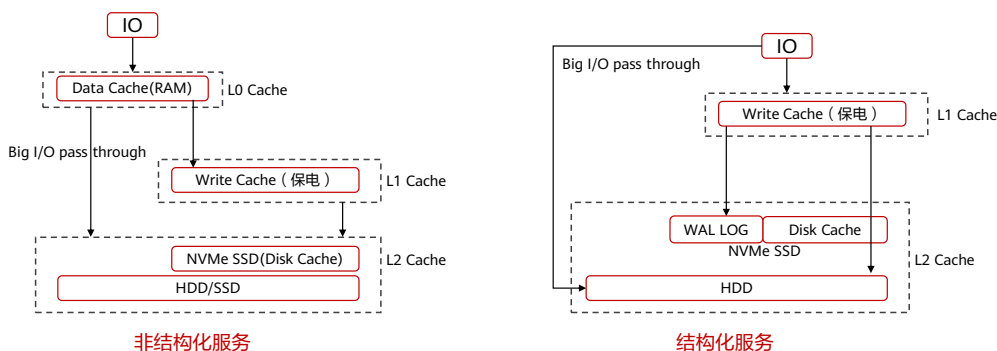


- OceanStor Pacific系列全对称分布式软件架构分两层路由，一层在Service layer，确保每个节点都能分担服务处理能力，另一层在Persistence layer，尽可能的均衡存储空间同时加快故障重构速度。
- Service layer非结构化服务与结构化服务由于语义的差异有不同的路由打散策略，但是都是基于同一个基础路由模型进行映射的。Persistence layer的路由打散策略是基于STORE DHT。
- OceanStor Pacific系列存储采用 Shard DHT算法作为 Service layer的基础路由策略，每一个资源池拥有一个 Shard DHT哈希环，每一个哈希环有 4096个 Shard，对于每个存储节点 Node能够处理的 Shard数量，为资源池成员存储节点数量均分 4096，比如资源池由 16个节点组成，那么就由 16个节点均分 4096，即每个存储节点处理 $4096/16=256$ 个 Shard。
- OceanStor Pacific系列对象会被映射成文件，对象的桶会被映射成根目录，非结构化数据服务 共一份元数据管理，对象、大数据与文件服务使用相同的打散策略。具体来说，为了解决大目录文件过多的性能问题，采用了基于目录分片的分布式打散策略，达到负载均衡的目的。目录分片与目录分片下的子文件归属相同的节点进行处理。
- 请求经过前端负载均衡打散后，在Service层归属节点处理完成处理，对于读请求，需要快速定位到数据持久化节点，采用了 STORE DHT算法进行快速路由。对于写请求，数据最终以 plog语义写入 Persistence层，在选择 plog时，需要考虑到后端节点及硬盘容量的均衡性，采用了动态智能分区和静态选盘算法。STORE DHT中的 key由 PlogID和 Offset构成，经过 hash计算后，确定数据存放在哪一个分区 (PT)，进而映

射到硬盘的具体位置。分区 (PT)支持机柜级、节点级、硬盘级，默认是跨节点组织分区 (PT), 系统所有容量按分区 (PT)个数均匀分配容量空间。

多级缓存技术

- OceanStor Pacific系列产品采用多级缓存技术逐层对I/O进行聚合加速，提升系统整体I/O性能。在SSD与HDD混配场景下，构建内存、SSD、HDD的多层缓存机制，减少下 HDD盘的I/O提升整体性能。在全闪存配置场景下，构建内存、SSD的多层缓存机制，以获取极致性能体现。
- 全闪存配置下，非结构化数据服务没有L2 Cache；结构化数据服务没有L2 Cache中的Disk Cache。



• 非结构化数据服务多级缓存:

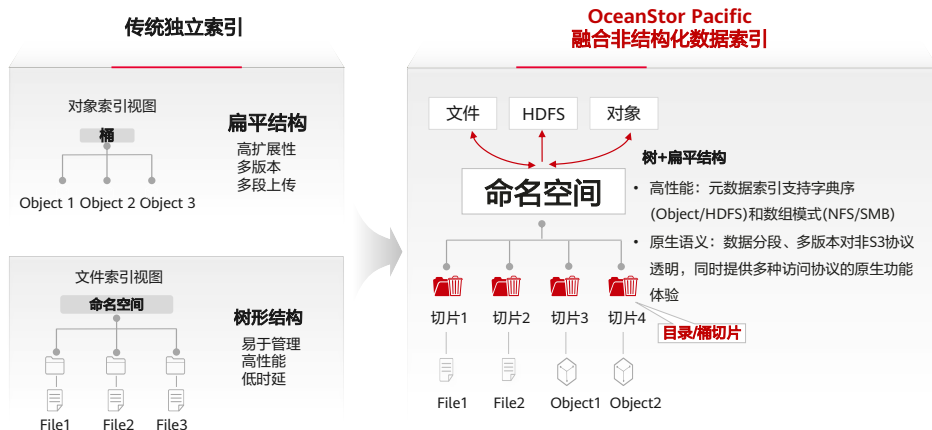
- OceanStor Pacific系列非结构化数据服务的节点角色通常分为：接入节点，文件归属节点，EC数据持久化节点。我们会在这三个角色分别构建缓存进行加速。L0位于系统的接入节点，L1位于文件的归属节点，L2位于数据持久化节点。
- 如果是标准协议访问，请求接收的存储节点就是接入节点，如果是DPC（Distributed Parallel Client，分布式并行客户端）访问DPC就是接入节点。
- L0 Cache是不持久化的，计算节点应用的异步写 I/O直接写入 L0 Cache就返回成功，当应用下发 flush命令或 L0 Cache缓存智能聚合后会写到后端存储。
- L1 Cache基于保电内存构建，同时会在节点间镜像，确保节点掉电或故障 cache数据不丢失。

• 结构化数据服务多级缓存:

- OceanStor Pacific系列结构化数据服务不支持保电内存，为实现缓存加速，把SSD从功能上分为两个部分：
- 第一个部分给L1 Cache做持久化使用，通常为主机业务服务，即主机的 Write I/O通过此部分用来确保数据临时持久化，不会因掉电导致数据丢失，对于这部分的Cache，我们通常称为WAL Cache，通常使用小比例的跨节点EC，保证数据的可靠性；小I/O在 L1Cache进行智能聚合，聚合成大块I/O，会以大比例EC写入后端HDD介质。
- 第二部分通常是作为主存机械盘的缓存区域L2 Cache，确保数据在进入HDD机

械盘之前，先通过这个部分的 Cache将数据缓存住，这样 I/O不需要到达机械盘，从而降低了时延，对于这部分的Cache，通常称之为Disk Cache。如果是全闪存配置，不会构建Disk Cache。

多协议无损互通



- OceanStor Pacific系列非结构化数据服务 共享存储资源池 (storage pool)，这是多服务语义互通的前提（一份数据写入，多种协议可访问）。为了实现语义的互通，统一了多种非结构化数据服务的概念模型。OceanStor Pacific系列非结构化数据服务 提供统一的多Namespace，文件服务称之为文件系统，对象服务称之为Bucket。各个Namespace的用户数据相互隔离，系统资源争抢通过SmartQos机制控制，一个Namespace可以同时支持大数据 /文件 /对象服务。
- 基于DROS底座，为了实现非结构化数据服务 数据互通， Namespace映射为对象的 bucket，把文件的目录结构与对象的 URL进行了映射，映射关系示意如上图所示，Namespace被映射成桶bucket1，目录/dir1下的文件object1，object2，被映射成桶 bucket1下的对象 dir1/object1，dir1/object2，根目录下的文件object3，object4，object5被映射成桶bucekt1下的对象object3，object4，object5。
- 反过来用户从S3对象侧创建的对象也会相应的映射成对应的目录结构，如果一个空的 Namespace(映射成桶 bucket1)，用户从S3对象侧在桶bucket1下创建dir1/object3对象，对应的DROS会为此对象创建相应的目录dir1，及目录dir1下的文件object3.如果用户从 S3对象侧在桶bucket1下创建 object3对象，对应的DROS会在Namespace的根目录下创建文件object3。
- 如此可以保证数据一份数据，从文件服务写入，对象服务可以访问，对象服务写入的文件，文件服务也可以访问。
- 基础的模型的统一，只能解决最基础的多服务互通访问，NAS/HDFS/对象各自的服务语义是有差异的，如锁、权限等。解决了上述的基础模型映射后，IO语义互通的

还有三个主要关键点：

- 语义无损互通：如S3协议的ListObjects接口、多段、多版本接口；如HDFS协议的Concat接口。
- 权限互通：同一个文件被多个协议同时访问时，如何做到权限互通，保证数据安全。
- 锁互通：同一个文件被多个协议访问时，不同协议的锁如何做到互通，如何保证数据并发访问的一致性。

目录

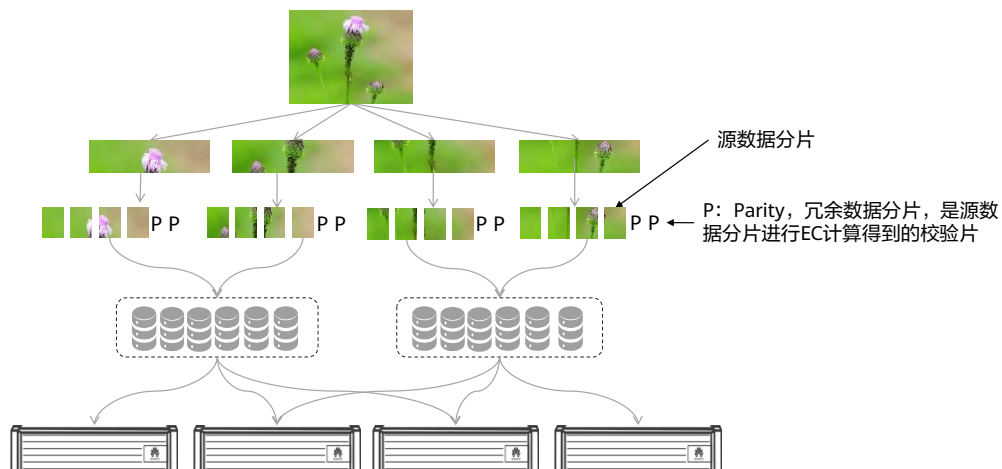
1. Ceph
2. 分布式存储产品介绍
 - 产品架构
 - **关键技术**
 - 应用场景
3. 分布式存储产品部署
4. 分布式存储特性

Elastic Erasure Coding（弹性EC）

- Erasure Coding（简称EC，即纠删码）是一种冗余保护机制，分布式存储系统在写入数据时，将数据切分为N个数据块（N为偶数），通过EC编码算法计算得到M个校验块（M取值2、3或4）的方式实现数据冗余保护。
- Elastic Erasure Coding（弹性EC）指的是在故障场景下，当故障后剩余的节点数量 $<N+M$ 时，在故障恢复前系统会将新写入的数据缩列为 $(N-1)+M$ 的配比（最小可缩列为 $N/2+M$ ），保证IO不中断的同时可靠性级别不降低，故障恢复后，系统冗余配比恢复为 $N+M$ 。
- 分布式存储系统支持节点级安全，机柜级安全；以节点级安全为例：
 - 当存储节点的个数大于等于 $N+M$ 时采用 $N+M$ 冗余配比，允许故障M块盘或M个节点；
 - 当存储节点的个数大于等于 $(N+M)/M$ 并且小于 $N+M$ 时采用 $N+M:1$ （折叠）冗余配比，允许故障M块盘或1个节点。

- EC冗余方式的空间利用率约为 $N/(N+M)$ ，N越大，空间利用率越高，数据的可靠性由M值的大小决定，M越大可靠性越高。基于性能和可靠性综合考虑，推荐配置4+2。
- 节点级安全：将 $N+M$ 个数据块和校验块存储于不同的节点中，故障M个节点或M块硬盘，系统仍可正常读写数据，业务不中断，数据不丢失。
- 机柜级安全：将 $N+M$ 个数据块和校验块存储于不同的机柜中，故障M个机柜、M个节点或M块硬盘，系统仍可正常读写数据，业务不中断，数据不丢失。

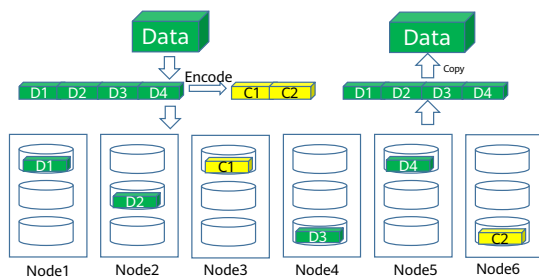
Erasure Coding特性原理



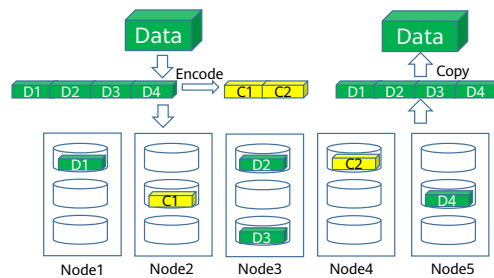
- 数据分片写到不同的节点上，不仅可以提高读写性能，又能保证数据的高可靠性和服务的高可用性，还能保持高的磁盘利用率，节约投资。
- 一个存储池内，任意不超过M（冗余份数）个硬盘故障后，系统都可以通过节点间数据重构的方式快速恢复出所损失的数据，进而恢复整个系统的数据可靠性。
- 系统中任意可用空间都可以作为“热备”空间使用，解决了传统RAID的热备盘问题，进一步提高了存储利用率。

弹性EC的读写原理

- 以 $N=4$ 、 $M=2$ 为例：
- 写原理：将用户写入数据Data切分成4个数据块（D1~D4），同时通过EC编码得到2个校验码（C1~C2）。
- 读原理：系统从多个节点的不同硬盘读取数据块（D1~D4），并通过Copy的方式将4个数据块拼装成Data返回给用户。



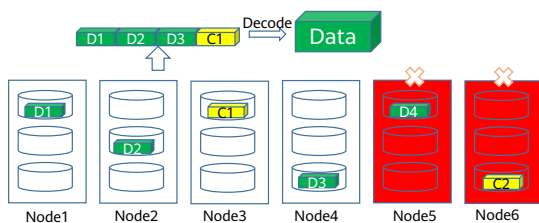
$N=4$ 、 $M=2$ ，6个节点 采用 $N+M$ 冗余配比



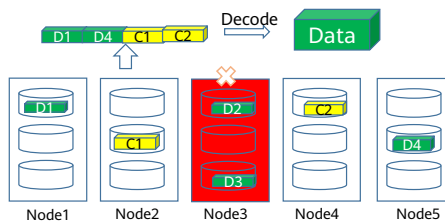
$N=4$ 、 $M=2$ ，5个节点 采用 $N+M:1$ 冗余配比

弹性EC故障场景读数据

- 以 $N=4$ 、 $M=2$ 为例：
- 故障场景读数据：当EC成员节点（成员盘）故障时，系统会任意读取 N 个正常节点（盘）的数据（不管是数据分片还是校验分片），通过EC解码恢复数据返回给用户。

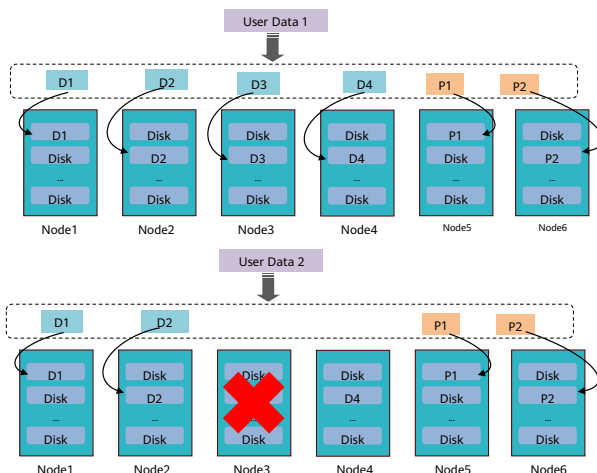


$N=4$ 、 $M=2$ ，6个节点 采用 $N+M$ 冗余配比



$N=4$ 、 $M=2$ ，5个节点 采用 $N+M:1$ 冗余配比

Erasure Coding 缩列



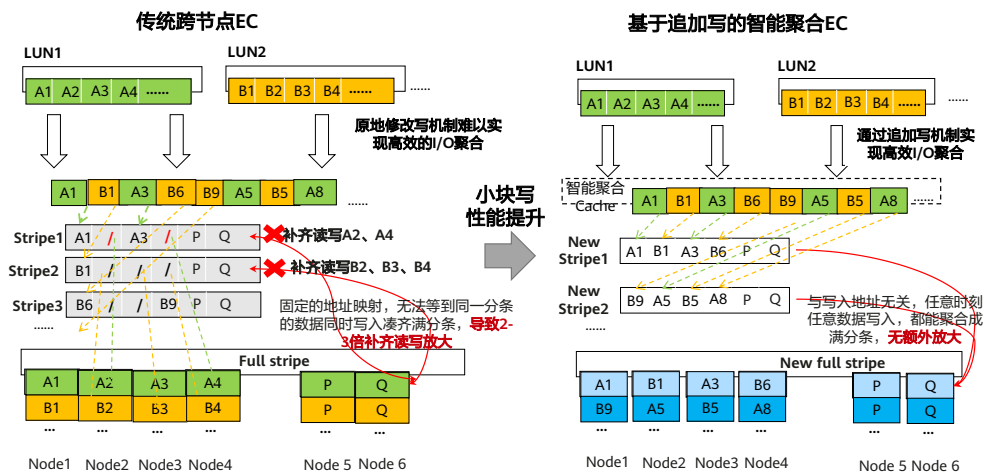
用户数据1被切成N（例如4）个数据分片另外再计算出M（例如2）个冗余分片 4+2。

当节点3发生故障以后，用户数据2被切成N（例如2）个数据分片另外再计算出M（例如2）个冗余分片 2+2，如图所示写入到盘。

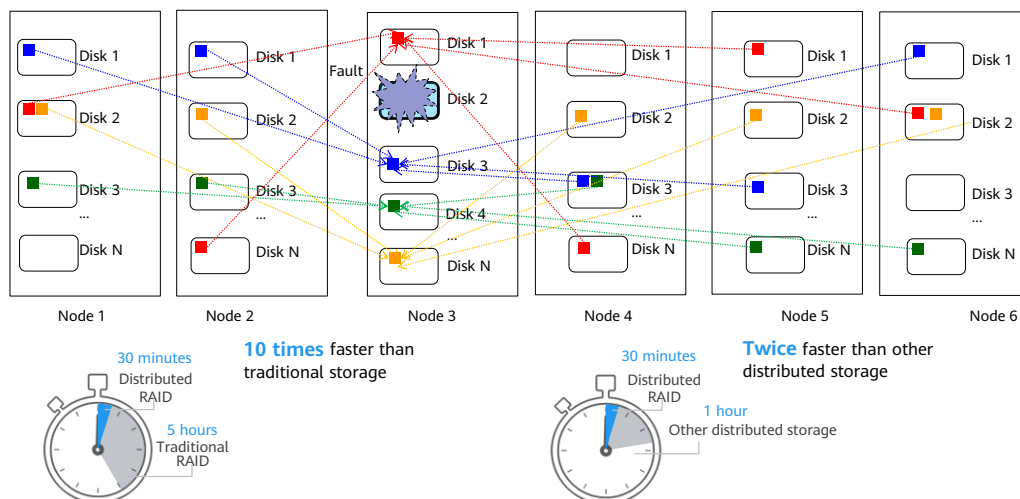
- OceanStor Pacific 支持动态 EC:

- 在系统发生节点故障时，既能保证数据冗余度，又能保证性能。如图所示，如果数据最开始4+2写入到系统，当系统发生一个节点故障以后，用户数据可以按照2+2的缩列方式，不发生降级写的方式写入各个节点。
- OceanStor Pacific 支持EC折叠：8.0版本最小3节点起配置，3节点也支持4+2的EC，提升空间利用率。

EC智能聚合技术



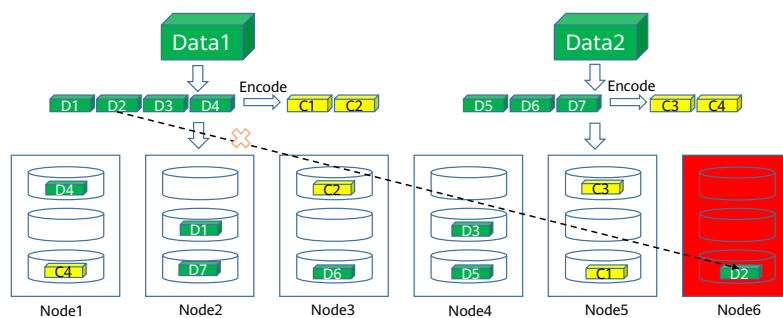
数据重构原理



- 以 $N=4$ 、 $M=2$ ，6个存储节点为例。
- 当EC成员节点Node3（成员盘）故障时，系统会任意读取 N 个正常节点（盘）的数据（不管是数据分片还是校验分片），通过EC解码恢复故障成员节点（成员盘）的数据块，并将其存入到其他的正常盘中（图中包含2个数据块的盘）。
- 重构策略：
 - 当硬盘发生故障时，硬盘池等待15分钟后启动数据重构。
 - 当节点发生故障时，硬盘池还在冗余保护时，延迟7天启动数据重构；当硬盘池不存在冗余保护时，延迟24小时启动数据重构。

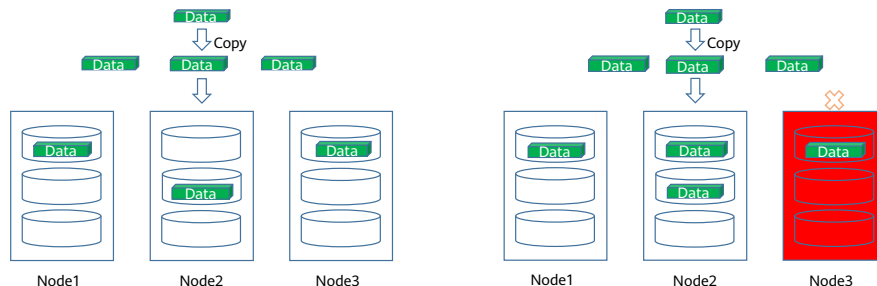
弹性EC的技术优势

- 支持最大的EC冗余配比为（22+2），空间利用率可以达到91.66%；
- 支持数据写入不降级，以N=4、M=2、存储节点个数=6为例，数据Data1在以4+2的冗余配比写入过程中，EC成员节点突然故障，此时为了保证可靠性不变，系统将新写入的数据Data2缩列为3+2冗余配比。



多副本

- 写原理：以3节点3副本池为例，数据copy3份，分别写入3个节点中；故障单节点后，支持将3副本数据写入到剩余的2个节点中。
- 读原理：系统从任意节点读取Data并发送给用户；当Node3节点故障后，系统在剩余2个节点中读取1个副本返回给用户。

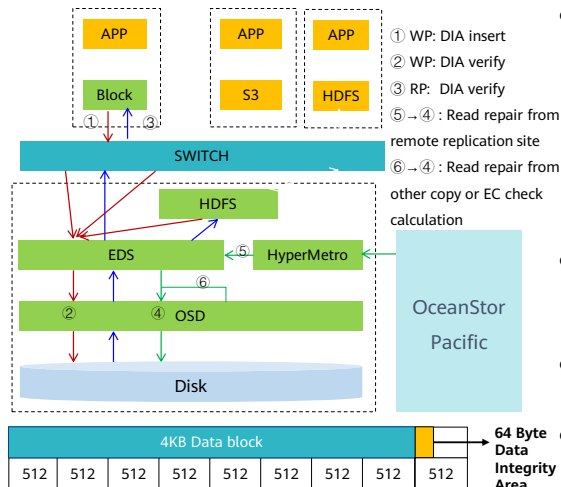


- 传统的硬盘级RAID模式将数据存放于单节点内的不同硬盘，当整节点发生故障时，无法有效恢复数据。存储系统将数据在节点间进行多副本构建，有效避免数据丢失。多副本是通过将相同的数据在不同的节点上存储多份来实现数据保护的一种技术，支持三副本和两副本，推荐三副本。
- 三副本的空间利用率为33.3%，两副本的空间利用率50%。
- 节点级安全：将多副本存储于不同的节点中，故障2个（三副本）或1个（两副本）节点或硬盘，系统仍可正常读写数据，业务不中断，数据不丢失。
- 机柜级安全：将多副本存储于不同的机柜中，故障2个（三副本）或1个（两副本）机柜、节点或硬盘，系统仍可正常读写数据，业务不中断，数据不丢失。每机柜需部署3个及以上的存储节点。

EC&副本对比差异

对比项	EC (N+M) 以4+2为例	多副本以三副本为例
空间利用率	66% 随着N值增加，利用率也随之增加	33%
可靠性	• 机柜级安全：允许故障任意2个机柜或2个节点 • 节点级安全：允许故障任意2个节点或2个硬盘	允许任意2机柜（机柜级可靠性）/2节点/2盘故障 • 机柜级安全：允许故障任意2个机柜 • 节点级安全：允许故障任意2个节点
重构	数据需要通过EC解码后得到目标分片	直接从其他盘拷贝后重建
SSD缓存寿命	业务IO相同情况下，无效IO少，总写次数是三副本的2/3，SSD缓存消耗少	业务IO相同情况下，无效IO多，总写次数是EC的1.5倍，SSD缓存消耗多

端到端数据校验技术



两种校验方式:

- **实时校验:** 写请求在系统入口处 (VBS) 插入校验, 用户数据写入磁盘前在OSD再次进行校验, 校验成功才写入磁盘; 主机读时在VBS处进行校验。
- **周期性后台校验:** 当系统业务压力较小时, 系统会自动启动后台周期校验和自愈。

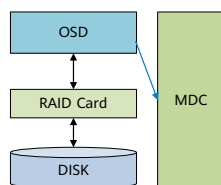
三种校验机制: CRC 32保护用户4K数据, 除此之外, 还支持主机LBA校验, 盘LBA校验等逻辑校验, 完美解决跳变, 读偏, 写偏等所有静默数据场景。

两种自愈机制: 本地冗余机制修复, 还可利用双活冗余数据进行修复。

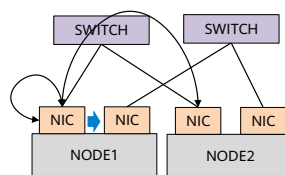
Block/NAS/HDFS均支持端到端校验, Object暂不支持。

亚健康管理技术

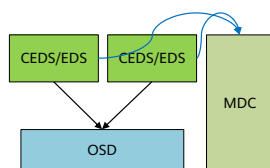
1、硬盘亚健康管理



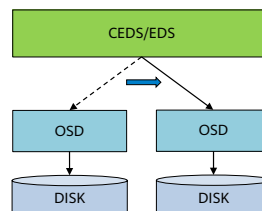
2、网络亚健康管理



3、进程/服务亚健康管理



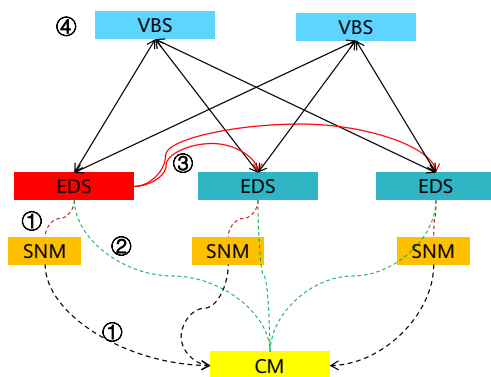
4、Fast-Fail（快速换路重试）



- 亚健康也称Fail slow，是指对应硬件可以正常运行但性能低于预期的一种状态。导致亚健康的原因非常多，包含但不限于 Firmware Bug、硬件自身设计缺陷、温度、环境（如震动）、配置错误等。
- 多种硬件均有可能进入到亚健康状态，包含但不限于盘（SSD/HDD）、网络、CPU、内存等，一旦某硬件进入到亚健康状态，如果存储系统未采取有效监控和容错，则极有可能会造成存储系统响应主机的时延增大、IOPS/BPS降低，甚至导致无法响应主机，进而导致主机业务中断。OceanStor Pacific系列存储系统对盘、网络、服务（涵盖CPU\内存等）等进行全面的亚健康状态监控，并进行智能诊断后实现自动隔离，实现单部件亚健康存储系统性能无感知，让主机享受一致性性能体验。
- 硬盘亚健康管理
 - 华为OceanStor Pacific系列存储系统对其使用的主存盘、缓存盘、系统盘、元数据盘均实现了全方位的亚健康状态监控，主要监控内容包含但不限于：SMART信息（如硬盘扇区重映射数超过门限、读错误率统计超标等），IO时延（含单IO级时延、IO平均时延等），IO错误（含静默数据错误数等）。
- 网络亚健康管理
 - 网卡降速，丢包/错包率增加，协商速率不匹配等都会导致集群网络性能降级，进入亚健康状态。系统通过检测网络资源状态的变化，定位受到网络亚健康影响的节点，进行bond主备切换或者节点隔离。
- 服务亚健康管理

- 分布式集群节点在运行过程中出现软硬件问题是普遍现象。由于节点软硬件问题导致节点进入亚健康状态，比如CPU降速，内存反复纠错导致访问降速等；在这种场景下，整系统服务时延受到单个节点影响而降级。针对这一类问题场景，系统通过收集时延信息检测出处于亚健康状态的节点，对问题节点或节点的问题资源进行隔离。
- 快速换路径重试机制
 - 华为OceanStor Pacific系列存储系统设计实现了专有的快速换路重试特性，即Fast-fail特性，确保单点亚健康的I/O时延可控。

故障快速切换技术



- **进程故障检测通知**: 部署监控进程, 可达到对进程故障的毫秒级监控。检测故障后立即拉起进程并通知集群主, 由集群主触发业务本地的恢复。
- **心跳检测机制**: 针对网络/节点等故障场景导致集群主和业务进程同时失联时, 采用“心跳群发+关联诊断”算法在控制新主上快速判定业务进程失联/故障。
- **快速倒换接管**: 将需要接管的业务并发倒换到剩余的多节点上, 一方面缩短接管时长, 一方面使接管后性能更均衡。同时倒换流程使用延迟加载、并行加载、加载数据量缩减、索引快速重建等技术来加速接管过程。
- **主机IO重定向**: VBS层毫秒级检测发往某节点/进程的IO是否返回, 如无返回则主动拉取该节点/进程状态视图, 如状态视图标示该节点/进程故障则主动断开与该节点网络连接并将IO重定向到新节点/进程。

DPC分布式并行客户端



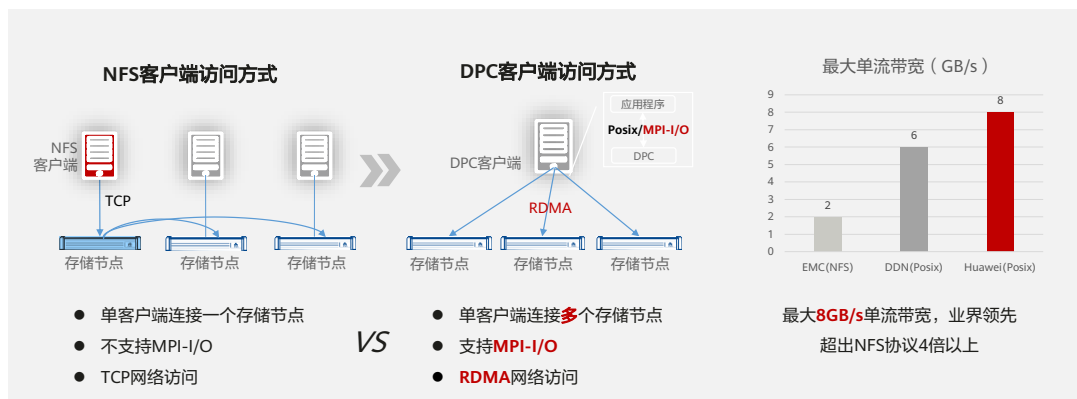
MPI-I/O实现气象
数据分析应用提速



单流高性能
保障卫星数据接收



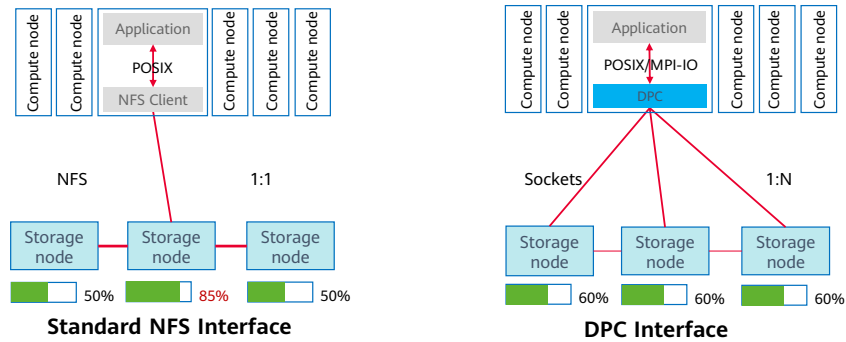
单客户端高性能释放
超算客户端潜能



- DPC: Distributed Parallel Client, 华为分布式存储客户端的缩写。
- HPC的一些场景, 会对单流带宽、单客户端带宽有较高的要求; 另外一些场景则可能用到MPI-I/O (多个客户端并行访问同一文件), 由于NFS协议下单个客户端只能连接一个存储节点, 网络访问以TCP为主, 也不支持MPI-I/O, 这些场景需求都无法很好的应对。为此, OceanStor Pacific推出了新一代DPC (Distributed Parallel Client) 分布式并行客户端, DPC部署在计算节点中, 可以实现单个计算客户端连接多个存储节点, 从而消除因存储节点配置造成的性能瓶颈, 发挥出计算节点的最大能力; 同时, DPC访问方式下的I/O级负载均衡, 相比NFS仅能在挂载时进行负载均衡判断可以获得更佳的均衡型, 从而把存储集群的性能更好的发挥出来; 再者, DPC可同时支持POSIX和MPI-I/O方式的访问, MPI类应用也无需修改即可获得更佳的访问性能; 最后, DPC支持IB和RoCE两种方式的RDMA网络, 可以在客户端直接将数据进行EC打散, 从而获得更低的数据读写时延。

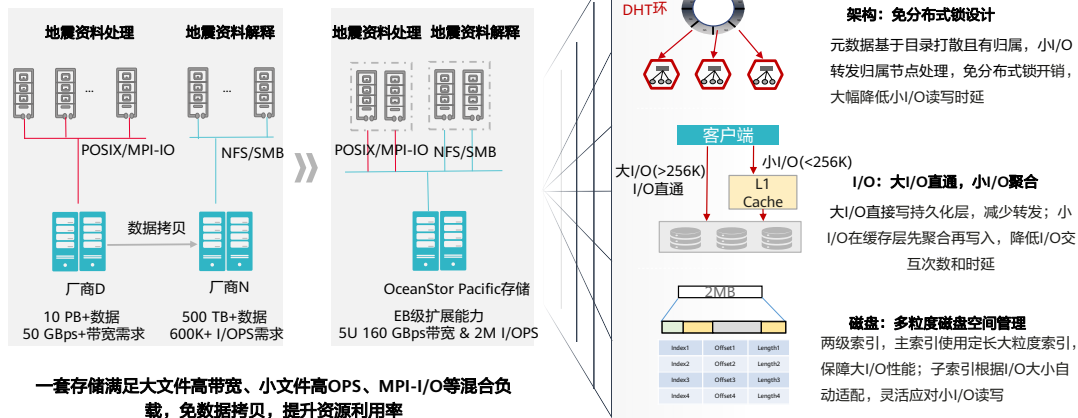
I/O级的负载均衡能力

- DPC与存储节点之间**使用RDMA通信**，有效降低计算及存储节点的CPU资源占用。
- DPC实现**IO级负载均衡**，单客户端同时访问多个节点，避免存储节点之间进行业务负载转发。
- 客户端支持MPI-IO，支持单文件在多个存储节点之间并发处理，实现跨节点负载均衡能力。



并行文件系统

• 面向混合负载设计



- HPC/HPDA、大数据类应用流程有多个环节，有的环节需要大I/O高带宽，有的环节又需要小I/O高I/OPS，传统HPC存储大都仅擅长一种性能模型，这就导致不同环节需要使用不同的存储产品，不同环节间交换数据需要进行拷贝，很大程度降低了流程效率。
- OceanStor Pacific搭载的新一代并行文件系统，通过系列创新实现了高带宽和高IOPS兼得，一套存储满足混合负载要求：
 - 架构上，OceanStor Pacific将元数据按目录打散到多个节点，从而实现了类似集中式存储的元数据有归属，小I/O可以转发到归属节点直接处理，免分布式锁开销，从而大幅降低小I/O的读写时延。
 - I/O流方面，大I/O采用直通方式下盘，提升带宽能力；小I/O通过Cache聚合后再下盘，既获得了理想的时延，又提升了小文件利用率。
 - 数据布局上，OceanStor Pacific采用了两级索引方式，主索引使用定长的大粒度索引，保障大I/O顺序读写的性能；子索引可根据I/O大小自动调整，避免小I/O写入时的写惩罚。

小I/O处理归属转发

《》 基于分布式锁的小I/O处理流程 《》



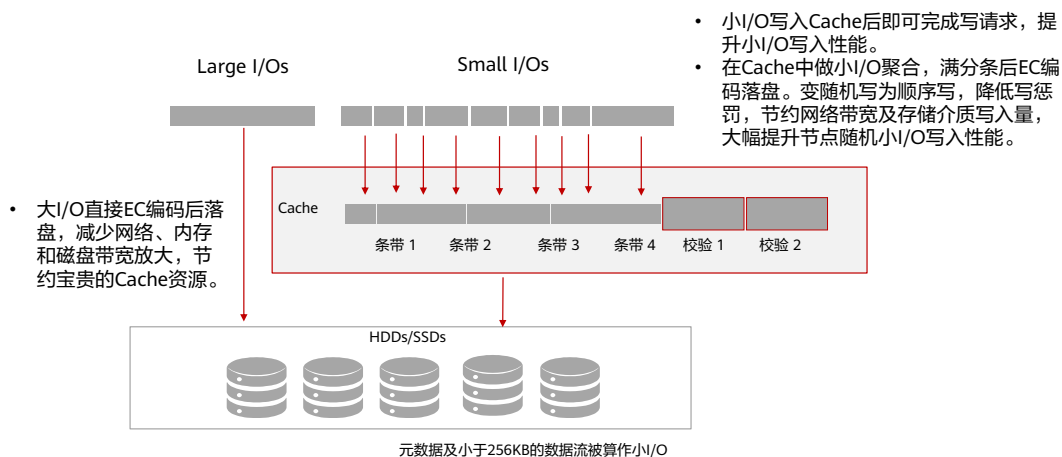
VS

《》 基于归属转发的小I/O处理流程 《》



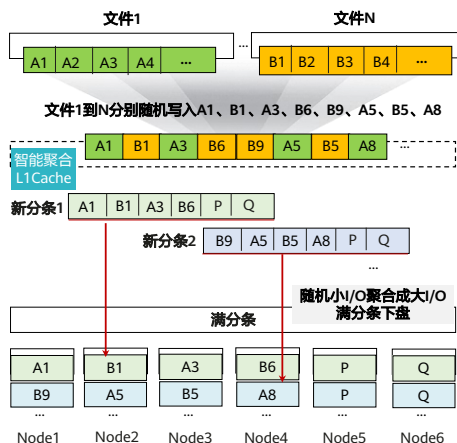
- 小I/O归属转发，避免分布式锁带来的多次交互开销。

智能I/O自适应



- 传统阵列数据写入，所有IO都先写入到内存，完成节点镜像后返回给业务，对于小IO来说路径最短，时延最好；但对大IO带宽型业务，网络和内存带宽均放大一倍，业务带宽只有网络和内存带宽的一半。面向混合负载，大IO走直通到盘，减少网络和内存带宽放大，实现最高的带宽性能；小IO写缓存聚合，完成镜像后响应，实现最低时延，但是放大了CPU调度的次数，进一步加剧了CPU的性能瓶颈。OceanStor Pacific系列文件系统在此基础上又实现了单边RDMA方式的数据拷贝，仅占用主存储节点的CPU资源，将总CPU开销降低了20-30%，混合负载下带宽和IOPS均能做到最佳。

智能分条聚合技术



大I/O直通

大I/O直接组EC分条下盘，不经过Cache，节省Cache资源并提高缓存寿命

小I/O智能聚合

I/O聚合成满分条EC下盘，降低写时延

智能追加写

追加写代替传统覆盖写，随机写转变为100%顺序写，避免小块写，发挥HDD极限写性能

- OceanStor Pacific系列产品支持EC冗余保护，为提供较高的空间利用率，支持高效的大比例EC，如20+2的大比例EC配比。在大比例的EC条带配比下，随机小I/O往往很难凑成满分条下盘，传统的处理方式会产生额外的写惩罚IO，导致性能的下降。为解决这种问题，OceanStor Pacific系列产品基于PLOG append only的语义设计实现了针对小I/O智能聚合的技术。Plog是Persistence Layer提供的一组按照固定大小管理的物理地址的集合，提供EC的冗余保护能力，通过PlogID+Offset的方式来访问具体的物理磁盘上的LBA地址。PlogID+Offset的逻辑地址到物理LBA地址的映射关系，是在写入数据时，才真正完成分配映射的。以文件服务为例，多个文件的随机小I/O写入L1Cache，在L1Cache进行聚合，聚合成大IO，满分条写入plog，减少写惩罚，提升系统性能。

FlashLink

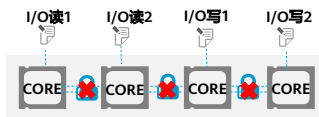
CPU分组算法



IO分组避免相互干扰

专核专用，保障关键业务资源投入，降低时延
多核共用，自适应调配多个业务的负载均衡

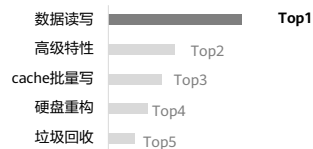
CPU分核算法



IO连续执行避免切换消耗

一次请求在同一个核上连续执行，利用同核不发生线程切换从而保证操作原子性来实现免锁设计，避免频繁多核切换，提升CPU Cache命中率

IO智能调度

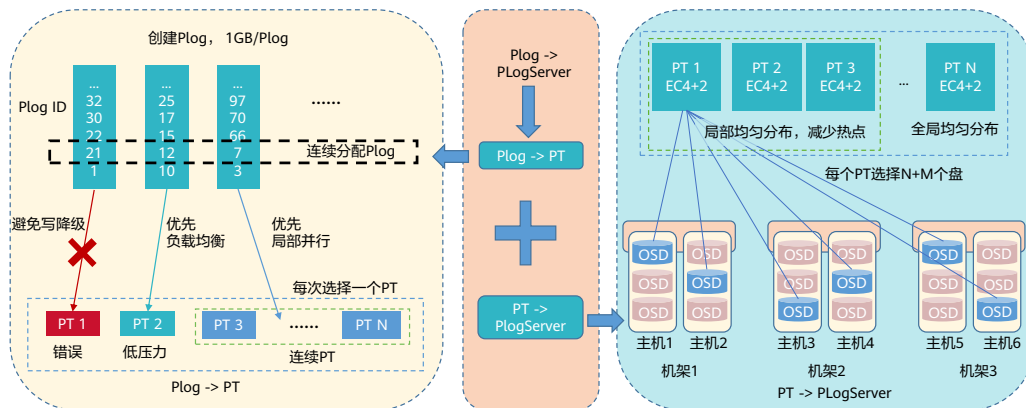


I/O优先级保障，降低业务影响

数据读写I/O永远第一优先级响应，其他I/O等数据读写I/O完成后重新启动，保障数据读写I/O时延最低

- 数据从I/O卡流入CPU。华为OceanStor Pacific系列的CPU利用算法具备三大显著优势：
 - CPU采用最新的智能分区算法，包括CPU分组算法和CPU分核算法。
 - CPU分组算法是把比较重要的I/O单独一个分组（比如：读写，数据交换），从而读写I/O与其他I/O部署在不同分组，避免相互干扰，保障读写I/O的性能；优先级相对低的I/O在一个分组内，共享CPU资源。这样能够保障资源利用的最大化。
 - CPU分核算法是华为独有的优势。I/O工作在CPU的时间片是轮询的，一段时间在CPU的Core1上工作，过一段时间轮询到Core2上工作。其一，从Core1跳转到Core2是要花时间的。其二，为了保障数据的一致性，I/O从core1跳转到core2时，CPU core1会加锁，保证I/O不会在Core1上运行，只在core2上运行，给I/O加锁也是需要时间的。当I/O在core2上工作一段时间，如果他要跳转回Core1，他需要为Core2加锁，同时为Core1解锁，此时加锁和解锁都是需要时间的。
 - CPU分核算法，能很好解决CPU轮询和加/解锁的问题，它让1次读请求或者写请求就在CPU的1个core上连续执行，更为重要的是，它采用了免锁设计，这样一个I/O就在一个core上运行，从开始到结束，有效避免了多个Core之间切换，以及加锁、解锁之间的时间消耗。

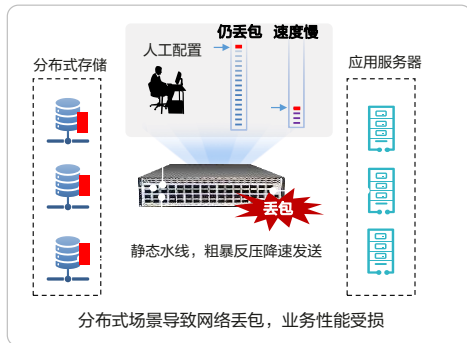
动态智能分区和静态选盘算法



网络加速

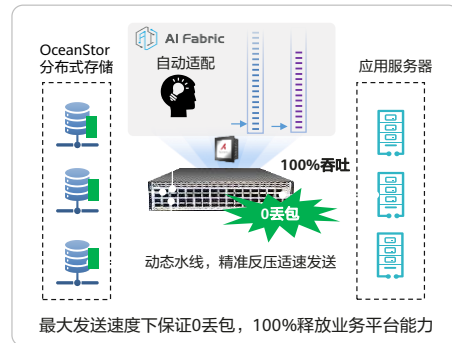
传统RoCE网络

静态配置水线，无法根据流量状态变化，要么丢包，要么降速。



AI Fabric使能的RoCE网络

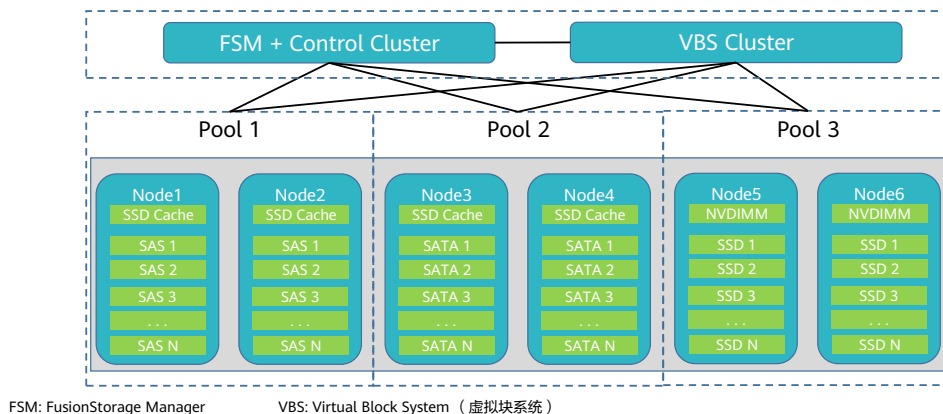
交换机内置智能芯片和算法，动态调整水线，确保100%吞吐下0丢包。



- 分布式存储单套集群往往会有几十甚至上千个存储节点，数据通过存储网络打散到不同节点中，这就对网络的性能和可靠性提出了更高的要求。OceanStor Pacific系列文件服务所有节点之间通过支持RDMA的RoCE网络互联起来，可以使用RDMA进行低时延通信，这是OceanStor Pacific系列能够达成高I/OPS、高带宽、低时延的基础之一。
- 传统RoCE网络静态配置流量水线、无法根据流量状态变化的情况，遇到I/O流程过大时，要么丢包，要么降速；华为OceanStor Pacific系列配合华为智能交换机构建存储网络，该类型交换机内置智能芯片(我们称为AI Fabric功能)，动态调整流程流水线，交换机网络与存储节点网卡主动协同加速，彻底避免拥塞和调度震荡，实现分布式存储I/O吞吐量提升，确保100%吞吐下实现0丢包。

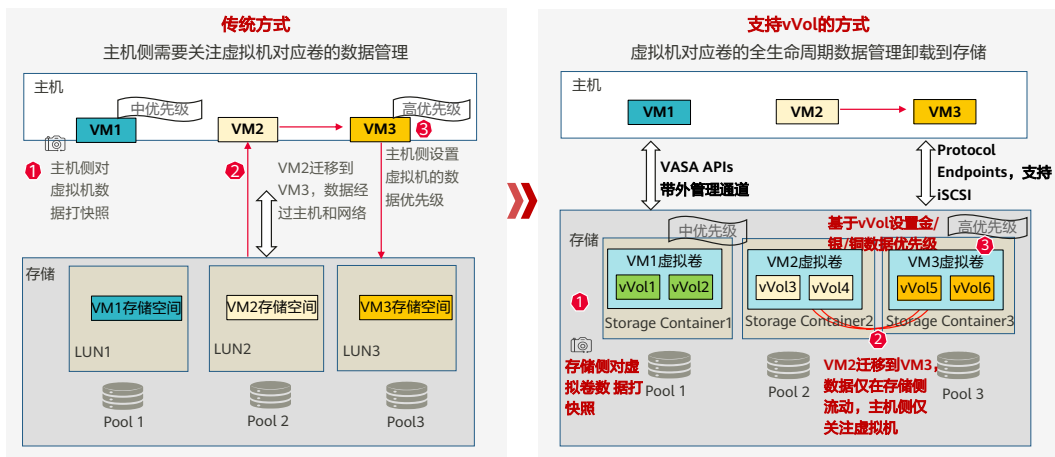
多资源池

- OceanStor Pacific Block同时支持不同性能的资源池满足性能需求，并且多资源池可以实现故障隔离，提升可靠性。



- 软件定义存储策略：支持以资源池为单位配置主存类型、缓存类型、副本冗余类型、安全等级策略（机柜级/服务器级）。
- 高可靠性：通过故障隔离，多MDC等机制，各资源池相对独立，故障互不影响。

vVol

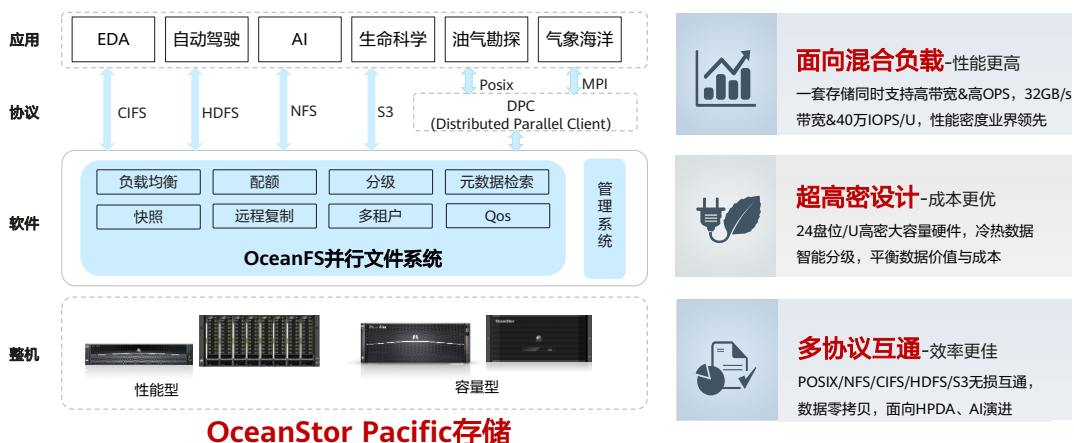


- vVol技术主要把对虚拟卷的全生命周期数据管理操作卸载到存储侧完成，通过VASA API带外管理通道可基于vVol创建、查询、快照、克隆等操作，也可以以vVol粒度自动进行智能数据迁移、空间回收。
- 图中第1点，在存储侧可完成虚拟卷的快照，主机侧不需要关注虚拟卷，简化了主机侧的管理。图中第2点，如果数据要从虚拟机2迁移到虚拟机3，传统方式要经过主机和网络，而采用vVol技术，数据仅在存储侧流动，既不占用主机CPU资源，也不用占用网络带宽，本地处理还可降低数据处理的时延，有利于整体性能提升。
- 一个或多个pool组成了Storage Container（存储容器）的存储空间，vVol则从存储容器中分配空间，不同虚拟机的业务天然进行了I/O隔离，可基于vVol为粒度配置金、银、铜等不同优先级的存储策略，保证关键业务性能优先级最高。也可以针对性配置Qos策略，对不同的vVol业务进行单独的性能优化。
- OceanStor Pacific 8.1.0版本的Protocol Endpoint支持iSCSI协议，1个主机能支持256个PE，1个PE可管理成千上万个vVol，主机可管理的卷规模也得到了大大的提升。

目录

1. Ceph
- 2. 分布式存储产品介绍**
 - 产品架构
 - 关键技术
 - **应用场景**
3. 分布式存储产品部署
4. 分布式存储特性

HPDA场景

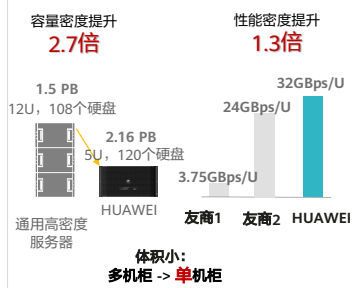


- 超算是构建下一代数据产业和科学计算的基石，更是开启未来数据文明的钥匙。传统超算从单一计算中心走向多元算力中心（HPC&AI& BigData），通过海量数据统一存储加速多元算力融合，实现高性能数据分析（HPDA），使超算从算力服务时代走向数据价值时代，数据存储举足轻重。华为HPDA解决方案，以OceanStor Pacific等并行文件存储为核心，通过与应用、软件平台、计算、网络等联合优化，形成油气勘探、基因测序、卫星遥感、自动驾驶等业务场景解决方案，构建数据密集型超算，助力科学创新和产业升级。
- HPDA解决方案以OceanStor Pacific存储为核心，通过与计算、网络、调度软件等周边资源联合优化，提供更高效、更经济、更可靠的存储服务，满足HPDA技术发展对存储系统提出的新挑战。
- HPDA解决方案使用华为OceanStor Pacific存储，提供基于文件、对象、大数据的存储服务。

面向下一代HPDA的存储解决方案



超高密度设计

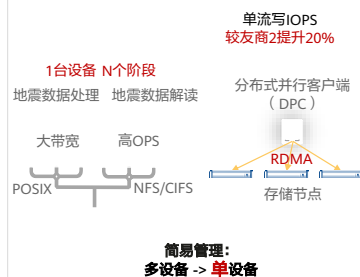


多协议互通

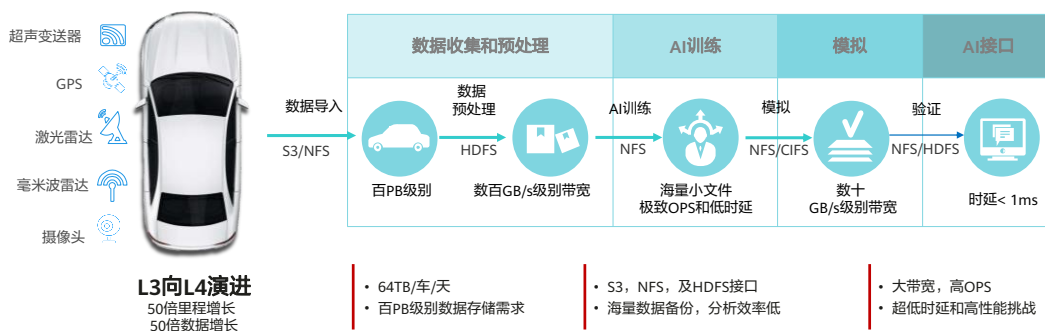
类型	友商1	友商2	OceanStor Pacific
原生语义	√	×	√
无网关	√	×	√
语义完整性	中	低	高
性能影响	中	大	小

经济高效:
多数据副本 -> **单数据EC打散**

面向混合负载



HPDA场景1：自动驾驶



关键优势

- 一套存储系统覆盖整个流程, 提供多协议互通, 数据0拷贝, 分析效率提升25%。
- 一框5U高性能存储可提供: (NFS)48GB/s带宽和440K IOPS; (DPC)160GB/s带宽和640万IOPS
- 冷热数据智能分级, 每U 24盘高密硬件, 全生命周期数据管理, TCO降低20%。

HPDA场景2：生命科学-基因测序

基因测序数据流



关键要求

- 二代测序仪：6TB/40小时
- 测序仪内存：大带宽
- 分析性能：几分钟即可处理100GB的样本（大文件和顺序小I/Os），需要大存储带宽。
- 目前归档成本较高。
- 归档海量数据耗时巨大。

关键优势

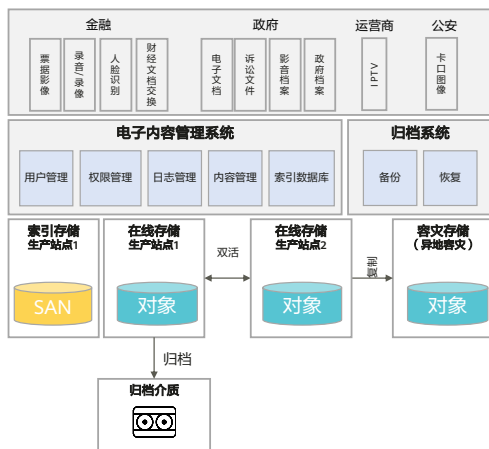
- DPC特性提供高单流带宽（8 GB/s）。
- 容量密度：2.4 TB/U
- 无损压缩算法，较友商产品效率提高1.5倍。

内容资源池及备份归档场景的对象存储解决方案



- OceanStor Pacific的另一个大场景是对象资源池。它包括海量内容资源池和备份和归档。
- 针对对象资源池场景，OceanStor Pacific有三个亮点：
- 可靠，最大支持12个双活站点，可容忍2个站点故障，提供6个9的高可用。
- 高效，支持千亿级存储桶，稳定搜索时延小于10 ms。
- 高性价比：120盘/5U高密硬件，22+2高比EC，节省空间和功耗。

对象存储场景1：内容资源池



现状和挑战

- 服务体验差：电子文档数据的CAGR（包括政府、公安机关、检察院、法院、医疗PACS，和金融文档影像）**超过20%**。文件量从数十增长到**百万甚至十亿级别**。目前的企业/分布式NAS系统检索效率低。时延超过200ms会影响服务体验。
- 备份困难：数据备份和恢复耗费**数日**，无法满足服务要求。

为何选择对象存储

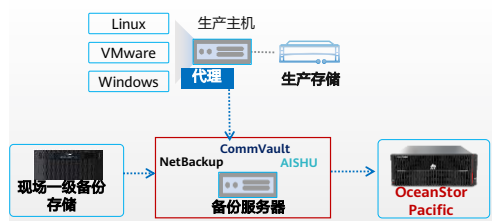
- **检索高效**：NAS采用树形元数据结构，在线索**数亿**个文件时，需要**多次I/O交互**。对象存储采用哈希算法和扁平结构来确保性能稳定，仅需**一次I/O交互**。
- **免数据备份**：NAS需要数据备份。对象存储采用版本控制和多站点容灾来恢复损坏的数据。

为何选择华为

- **性能稳**：数百亿对象，数据读写时延小于50ms（新加坡Sea）。
- **99.9999%可用性**：跨区域2到3个主站点，确保业务连续性。
- **免备份**：版本控制、桶级快照、企业级防误删除、批量数据恢复。
- **多协议互通**，允许新旧业务共存。

- 针对内容资源池场景，OceanStor Pacific提供以下关键特性：
- 性能稳定：百亿对象，数据读写时延< 50 ms（新加坡Sea）
- 99.9999%高可用：跨区域2~3个活跃站点，保障业务连续性
- 免备份：多版本控制、桶级快照、企业级防误删除、批量数据恢复
- 多协议互通，新老业务共存

对象存储场景2：备份



场景：OceanStor Pacific作为二级备份存储
(备份软件将数据从一级备份存储迁移到OceanStor Pacific)。

现状和挑战

- **烟囱式建设**：每个应用系统都有自己的备份系统。资源无法高效利用。
- **海量数据**：生产数据不断增加，备份数据长期保留。
- **快速恢复需求**：生产业务必须全天候运行，需要快速故障恢复。
- **高可靠需求**：备份是生产的最后一道保护，保留期内的任何数据都必须可以恢复。

为何选择对象存储

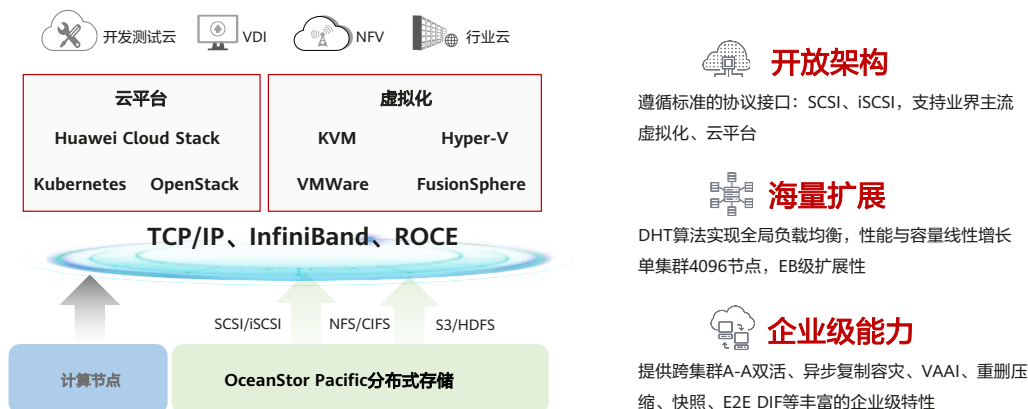
- 分布式架构比集中式设备具有更好的可扩展性。对象存储支持跨站点EC，可实现高可靠低成本。
- 对象数据的扁平存储，避免文件系统碎片化，在处理海量数据时提供了稳定的性能。

为何选择华为

- **高性能**：每节点2 GB/s读写带宽，随节点增加线性增长。
- **性能稳**：数百亿对象，数据读写时延小于50ms（新加坡Sea）。
- **99.9999%可用性**：跨区域2到3个主站点，确保业务连续性。
- **高可靠**：版本控制、桶级快照、企业级防误删、批量数据恢复、多站点容灾、WORM。

- 针对二级备份场景，OceanStor Pacific提供以下关键特性：
- 高性能：单节点2 GB/s读写带宽，随节点增加线性增长
- 性能稳定：百亿对象，数据读写时延< 50 ms
- 99.9999%高可用：跨区域2~3个活跃站点，保障业务连续性
- 高可靠：多版本控制、桶级快照、企业级防误删除、批量数据恢复、多站点容灾、WORM

云/虚拟化场景



- OceanStor Pacific系列在虚拟化场景中应用，具有如下特点：
 - 块服务可支撑当前核心基础数据库和虚拟机，文件集中共享减少拷贝并可弹性扩展。
 - 基于通用服务器硬件平台，一套存储同时提供分布式对象、大数据和文件服务，可根据业务类型灵活上线和回收业务。
 - 基于统一网管秒级分发对象和块存储业务，并可基于业务等级灵活分配存储资源，满足不同业务的性能和容量需求。
- OceanStor Pacific系列在云场景中应用，具有如下特点：
 - 提供对象服务可存储文档、图片、音视频等非结构化数据。
 - 支持按需线性扩容，提供海量、安全、高可靠、低成本的数据存储能力。
 - 支持1000亿级对象下的数据实时透视功能，实时展示系统中的热点对象和对象数量等信息。

人工智能技术革命



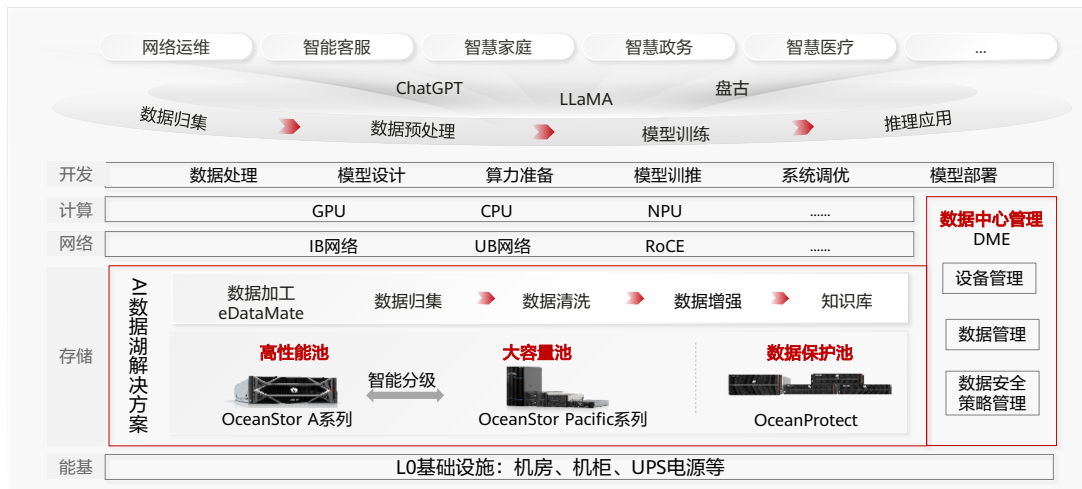
- AI大模型的商业正循环，需要依赖在各行各业中的商业变现。
- AI大模型可以分为：L0基础大模型，L1行业大模型，L2场景大模型
 - L0基础大模型，面向2B和2C市场，提供智能对话，图像生成的基础能力；
 - L1行业大模型，利用行业知识和通用大模型的能力进行训练，能够满足特定行业的基础应用；
 - L2场景大模型，在AI大模型的应用上更加深入，更加聚焦。通常用于解决特定行业、特定场景的具体问题，可以预见，未来这类应用会像今天的手机APP一样丰富多彩。

AI大模型时代的存储



- 在过去传统的三个维度上，AI存储的要求更高，比如：
 - 性能：AI存储需要10倍以上的性能，需要PB级带宽，亿级IOPS能力；
 - 数据范式：AI存储还需要支持Vector，Tensor，RAG等新的存储范式；
 - 可靠性：除了过去我们定义的6个9的传统可靠性，AI存储还需要支持勒索病毒检测，Checkpoint快速恢复等能力。
- 此外，还有三个新的维度，比如：
 - 数据编织：AI存储需要支持数据全局可视、可管，高效的数据移动和数据处理等能力；
 - 扩展性：AI存储需要支持EB级的容量扩展，还要支持GPU、DPU、NPU等横向扩展能力，用于IO处理的加速；
 - 可持续性：AI存储需要满足省电、省空间，我们定义了两个标准供参考，1 Watt/TB即每TB的功耗小于1Watt，1PB/U即每U空间支持的容量大于1PB。

AI数据湖

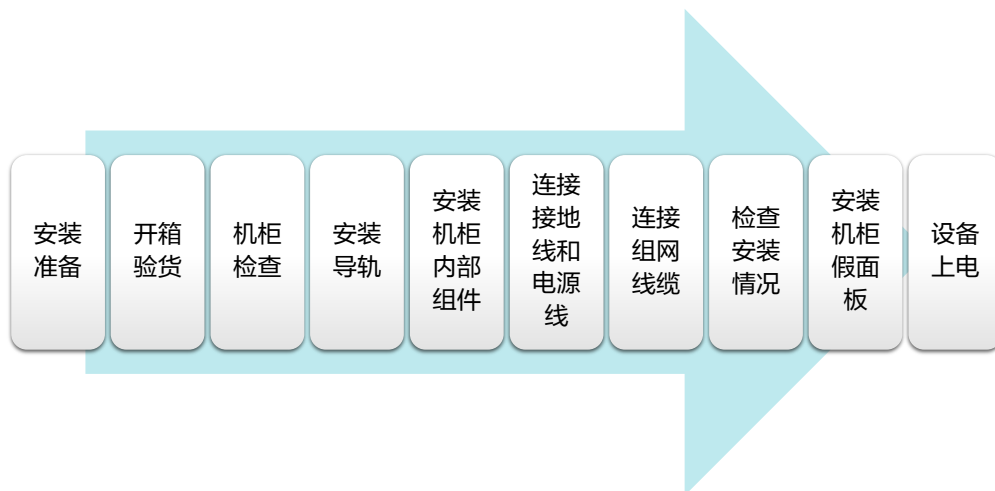


- AI数据湖解决方案是面向新型数据中心场景的综合性解决方案，满足客户对海量数据高效存储、实时分析、跨域管理、绿色节能的要求，打造“全栈协同、融合高效”的全栈解决方案；引领应用加速引擎和全局数据管理新产业方向；引领面向数据应用加速、全栈协同的新型基础设施建设。
- 在以OceanStor A系列为代表的高性能资源池基础上，增加以OceanStor Pacific系列为代表的大容量存储资源池，以OceanProtect系列为代表的的海数据保护资源池，组成AI数据湖解决方案，满足客户AI基础设施中对不同存储类型的需求。
- 此外，通过DME数据管理软件（集成Omni-Dataverse全局文件系统），可以让客户的数据全局可视、可管、可用。
- 为了提升客户AI场景下的数据预处理效率，华为也提供了eDataMate数据预处理工具链，加速从数据到知识的加工过程。
- 采用华为的AI数据湖解决方案后，可以帮助可以提升30%的算力可用度，EB级容量扩展，以存储内生的安全防护。

目录

1. Ceph
2. 分布式存储产品介绍
- 3. 分布式存储产品部署**
4. 分布式存储特性

设备安装流程



配电规划

设备在安装之前应该合理规划配电方案，保证设备的安全性、稳定性和可靠性。

场景	注意事项
机房	机柜的配电箱应支持独立的两路电源输入，且输入电压相同。
	电源线的额定负载流量应该满足机柜和设备的最大运行功率要求。
机柜内部配电	机柜中需要冗余电源输入的设备，每一路输入应接输入电源的不同支路。
	应给机柜中设备提供有效的过载、短路保护。
	机柜中电源线和信号线宜分开走线，两者间距离宜保持在3cm以上。

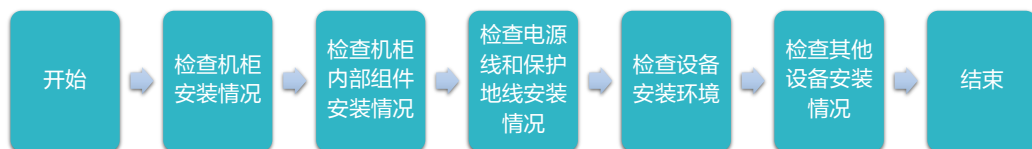
能耗计算工具：

华为数据存储信息中心>信息速查>规划设计>能耗计算

<https://info.support.huawei.com/storageinfo/powercalc/#/home>

检查安装情况

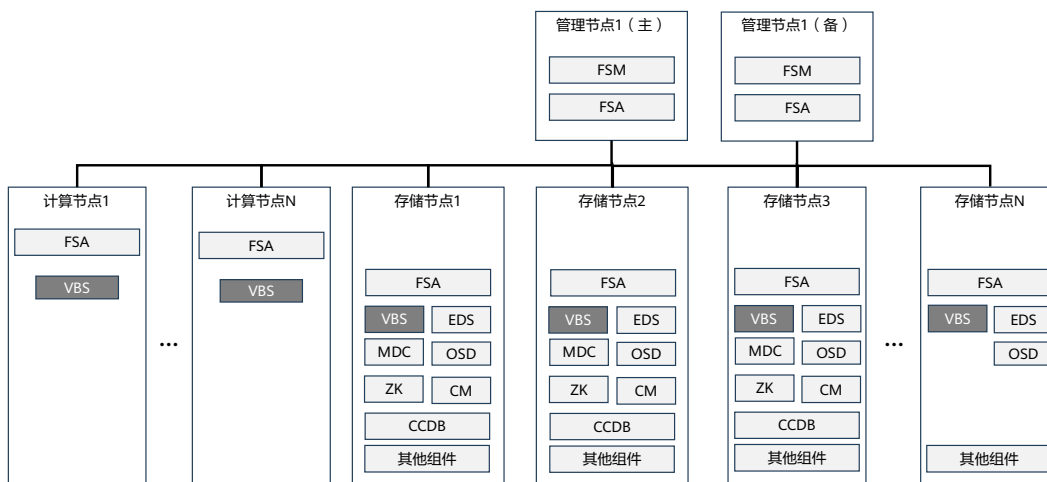
- 在完成所有硬件的安装后，要根据华为技术有限公司硬件质量标准，对所有的硬件安装进行检查，确保安装工作规范正确，为后续的软件安装及系统调试工作做好准备。



分布式存储设备上电流程

打开配电柜的电源开关	打开配电盒或PDU的电源开关	上电交换机	上电存储节点	确认设备的状态
• 检查设备机柜中配电盒的所有空气开关或PDU的电源开关均处于关闭状态。 • 由客户方具有专业资质的人员依次打开配电柜给所有机柜的供电电源开关。	• 逐个打开机柜中配电盒的空气开关或PDU的电源开关。 • 每闭合一个开关，在配电盒输出端测量并确认输出电压在合适的范围之内。	• 按照在机柜中从下向上的顺序上电。	• 将电源模块接通外部电源后，存储节点随电源模块一起自动上电。当存储节点已设置为“上电不开机”时，接通电源后不会自动上电，需要手动按存储节点前面板的电源按钮，使设备上电。	• 观察设备的风扇，确认风扇运行由全速运转降到匀速运行，风扇旋转的声音无异常，机箱后面的通风孔有强气流排出。 • 观察设备面板上的指示灯状态，确认设备运行正常。

块服务部署视图（VBS计算存储分离）



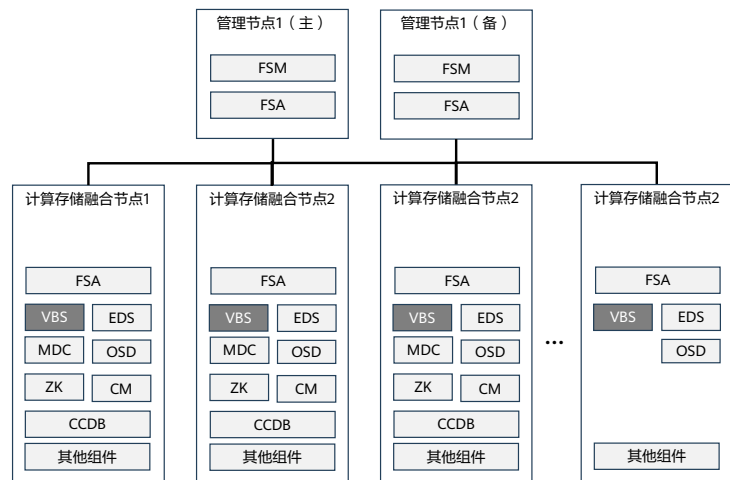
90 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



- FSM
 - OceanStor Pacific系列的管理模块，提供告警、监控、日志、配置等操作维护功能。
 - 安装管理节点时，启动FSM进程。
- FSA
 - OceanStor Pacific系列的管理代理进程，实现各节点与FSM的通信，可收集各节点的监控与告警信息，也可在升级本节点软件组件时，接收升级包并执行升级。
 - 安装存储节点和计算节点时，启动FSA进程。
- MDC
 - 元数据控制软件，实现对分布式集群的状态控制，以及控制数据分布规则和数据重建规则等。
 - 创建控制集群时，控制集群节点启动MDC进程。
- ZK
 - 为MDC集群提供选主仲裁。由ZK集群在多个MDC中选举主MDC，主MDC对其它MDC进行监控，主MDC故障时产生新的主MDC。
 - 创建控制集群时，控制集群节点启动ZK进程。

- VBS
 - 虚拟块存储管理组件（块服务独有），执行卷元数据管理。通过SCSI或iSCSI接口提供分布式存储接入服务，使计算资源能通过VBS访问存储系统，从而把卷映射给客户的应用服务器使用。
 - 选择节点启用VBS后，该节点上启动VBS进程，部署VBS进程的节点形成VBS集群。
- EDS
 - 接收来自VBS的I/O业务，并执行具体的I/O操作。在EDS中，可执行各服务相关的特性操作（如快照等），同时还可管理存储空间，建立数据与存储空间之间的索引关系，确保每块数据通过索引都能找到对应的存储位置。同时在数据存储到物理空间前，进行重删压缩处理。
 - 加入存储池的节点开启服务时，启动EDS进程。
- OSD
 - 管理硬盘的进程，处理VBS下发的I/O消息，执行具体的I/O数据存储空间操作，进行数据冗余保护并持久化到存储介质中。
 - 节点加入存储池时，启动OSD进程。
- CM
 - 集群管理软件，用于管理整个存储集群的状态信息，包括各组件的状态信息，实时监控各组件状态。
 - 当创建控制集群时，控制集群节点会启动CM进程。
- CCDB
 - 集群配置数据库，用于保存用户配置信息的数据库，在EDS组件中采用CCDB存放配置信息。
 - 当创建控制集群时，控制集群节点会启动CCDB进程。

块服务部署视图（VBS计算存储融合）



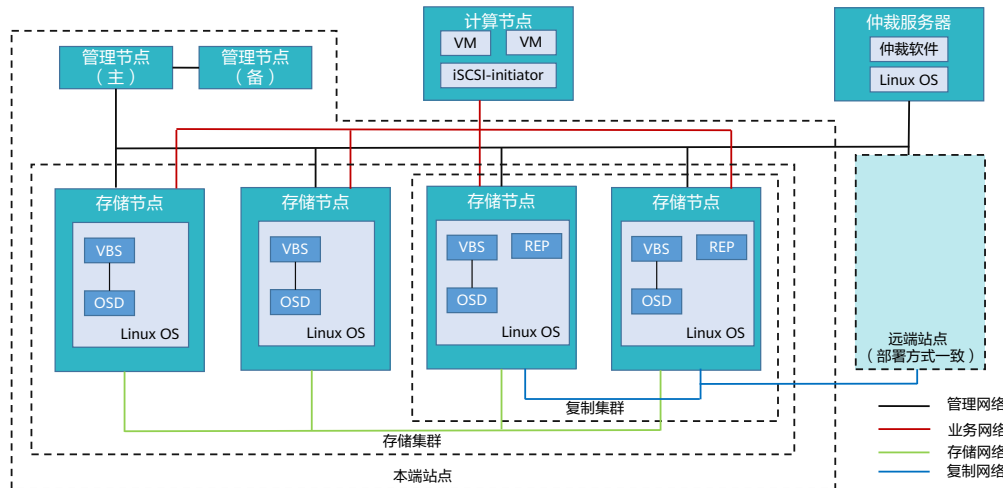
块服务节点角色及关键组件

节点角色	描述	可部署的组件
管理节点	部署FSM（FusionStorage Manager）管理进程的节点，负责系统的告警、监控、日志、配置等操作维护功能。必须部署2个节点，可以与计算节点或存储节点合并部署。	•FSM •FSA
计算节点	客户实际运行业务的节点。	•VBS •FSA
存储节点	提供存储资源的节点。分布式存储中主存盘和缓存盘所在的服务器，同时也是运行OceanStor Pacific系列软件的服务器。	•FSA •ZK •EDS（Enterprise Data Service） •OSD •CM •CCDB •MDC •VBS
计算存储融合节点	计算节点和存储节点使用同一个存储节点融合部署时，该节点既可作为计算节点，也可作为存储节点。仅适用于块服务的VMware场景。	•FSA •VBS •ZK •EDS •OSD •CM •CCDB •MDC

块服务组件部署原则

组件名称	部署要求
FSM	部署2个，主备模式。
FSA	每个节点上部署1个。
MDC	一个集群至少部署3个，形成MDC集群。
ZK	一个集群至少部署3个，可选择3、5、7、9个。
VBS	•SCSI场景：VBS个数与计算节点、存储节点之和一致，每1个计算节点和存储节点上默认都创建1个VBS。 •iSCSI场景：在需要VBS的每个存储节点上创建1个。
EDS	每个存储节点部署1个。
OSD	每个存储节点部署1个。
CM	一个集群至少部署3个。
CCDB	一个集群至少部署3个。

网络平面划分（块服务场景）

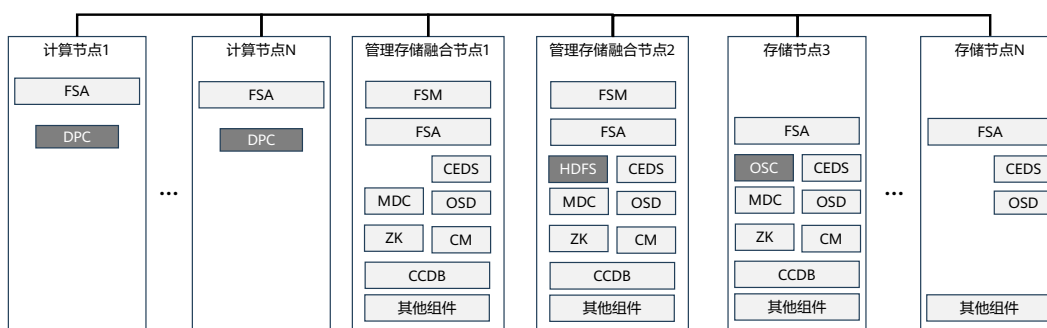


94 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



- 管理网络：与客户管理网络对接，用于系统管理和维护，管理平面支持双网络平面（管理内部网络、管理外部网络）。
 - 管理内部网络：用于存储节点和管理节点之间的数据通信，包括部署升级，资源配置，性能告警上报，存储节点NTP时间同步，证书密钥更新等。
 - 管理外部网络：用于客户端访问管理界面，REST接口，CLI，告警转发，与外部NTP服务时间同步等。
- 存储网络：内部网络平面，用于系统内部所有节点之间的业务数据通信，存储平面支持双网络平面（存储前端网络、存储后端网络）。
 - 存储前端网络：用于处理存储节点之间的前台数据。可用于VBS与计算节点通信，处理VBS与OSD间的通信数据。
 - 存储后端网络：用于处理存储节点OSD和OSD间的后台通信数据。
- 业务平面：与客户应用系统对接，通过iSCSI协议访问存储；
- 复制平面：用于复制节点间进行数据同步复制；
- 仲裁平面：用于与双活仲裁服务器进行通信，上图是复用管理网络。

文件/对象/大数据服务部署视图



- FSM

- OceanStor Pacific系列的管理模块，提供告警、监控、日志、配置等操作维护功能。
- 安装管理节点时，启动FSM进程。

- FSA

- OceanStor Pacific系列的管理代理进程，实现各节点与FSM的通信，可收集各节点的监控与告警信息，也可在升级本节点软件组件时，接收升级包并执行升级。
- 安装存储节点和计算节点时，启动FSA进程。

- MDC

- 元数据控制软件，实现对分布式集群的状态控制，以及控制数据分布规则和数据重建规则等。

- 创建控制集群时，控制集群节点启动MDC进程。
- ZK
 - 为MDC集群提供选主仲裁。由ZK集群在多个MDC中选举主MDC，主MDC对其它MDC进行监控，主MDC故障时产生新的主MDC。
 - 创建控制集群时，控制集群节点启动ZK进程。
- CEDS
 - 接收来自DPC、NFS客户端、HDFS的I/O业务，并执行具体的I/O操作。在CEDS中，可执行各服务相关的特性操作（如快照等），同时还可管理存储空间，建立数据与存储空间之间的索引关系，确保每块数据通过索引都能找到对应的存储位置。
 - 加入存储池的节点开启服务时，启动CEDS进程。
- OSD
 - 管理硬盘的进程，处理DPC、OSC或HDFS下发的I/O消息，执行具体的I/O数据存储空间操作，进行数据冗余保护并持久化到存储介质中。
 - 节点加入存储池时，启动OSD进程。
- CM
 - 集群管理软件，用于管理整个存储集群的状态信息，包括各组件的状态信息，实时监控各组件状态。
 - 当创建控制集群时，控制集群节点会启动CM进程。

- CCDB
 - 集群配置数据库，用于保存用户配置信息的数据库，在CEDs组件中采用CCDB存放配置信息。
 - 当创建控制集群时，控制集群节点会启动CCDB进程。
- OSC
 - 对象服务控制组件（对象服务独有），处理Amazon S3协议消息，完成对象存储服务业务逻辑处理。
 - 加入对象存储池的节点启动OSC进程，一个节点启动一个OSC进程。
- DPC
 - 分布式并行客户端（文件服务独有），作为存储客户端运行在计算节点上，与后端存储节点进行数据交换。DPC支持MPI协议和POSIX协议的文件语义访问。
 - 文件服务DPC场景下，安装计算节点时，启动DPC进程。
- HDFS
 - Hadoop文件系统协议组件（大数据服务独有），提供原生HDFS标准接口及HDFS语义的处理能力。
 - 加入大数据存储池的节点启动HDFS进程，一个节

点启动一个HDFS进程。

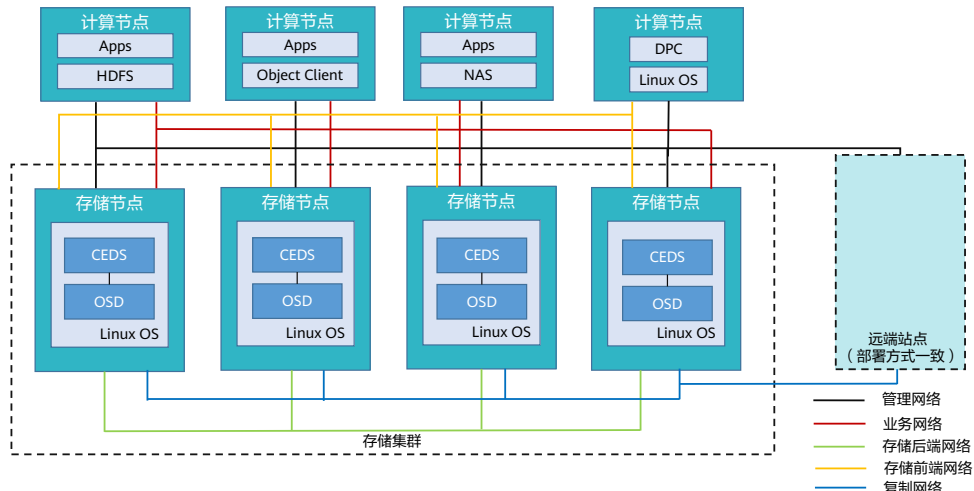
文件/对象/大数据服务节点角色及关键组件

节点角色	描述	可部署的组件
管理存储融合节点	在文件/对象/大数据服务中，管理节点和存储节点需要融合部署。该节点既可以做管理节点，也可以做存储节点。 •管理节点：部署FSM（FusionStorage Manager）管理进程的节点，负责系统的告警、监控、日志、配置等操作维护功能。必须部署2个节点，与存储节点合并部署。 •存储节点：提供存储资源的节点。分布式存储中主存盘和缓存盘所在的服务器，同时也是运行OceanStor Pacific系列软件的服务器。	•FSM •FSA（FusionStorage Agent） •ZK（ZooKeeper） •CEDS（Converged Enterprise Data Service） •OSD（Object Storage Device） •OSC（Object Service Controller） •HDFS •CM（Cluster Manager） •CCDB（Cluster Configuration Database） •MDC（MetaData Controller）
计算节点	客户实际运行业务的节点。	•FSA •DPC
存储节点	提供存储资源的节点。分布式存储中主存盘和缓存盘所在的服务器，同时也是运行OceanStor Pacific系列软件的服务器。	•FSA •ZK •CEDS •OSD •OSC •HDFS •CM •CCDB •MDC

文件/对象/大数据服务组件部署原则

组件名称	部署要求
FSM	部署2个，主备模式。
FSA	每个节点上部署1个。
MDC	一个集群至少部署3个，形成MDC集群。
ZK	一个集群至少部署3个，可选择3、5、7、9个。
CEDS	每个存储节点部署1个。
OSD	每个存储节点部署1个。
CM	一个集群至少部署3个。
CCDB	一个集群至少部署3个。
OSC	每个存储节点部署1个。
DPC	•DPC场景：每个计算节点创建1个。 •DPC和标准协议混合场景：在需要DPC的每个计算节点上创建1个。
HDFS	每个存储节点部署1个。

网络平面划分（文件/对象/大数据服务场景）

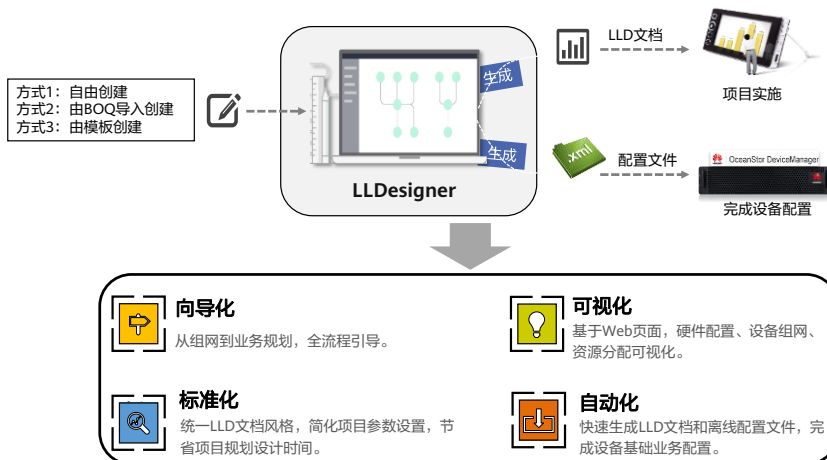


98 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



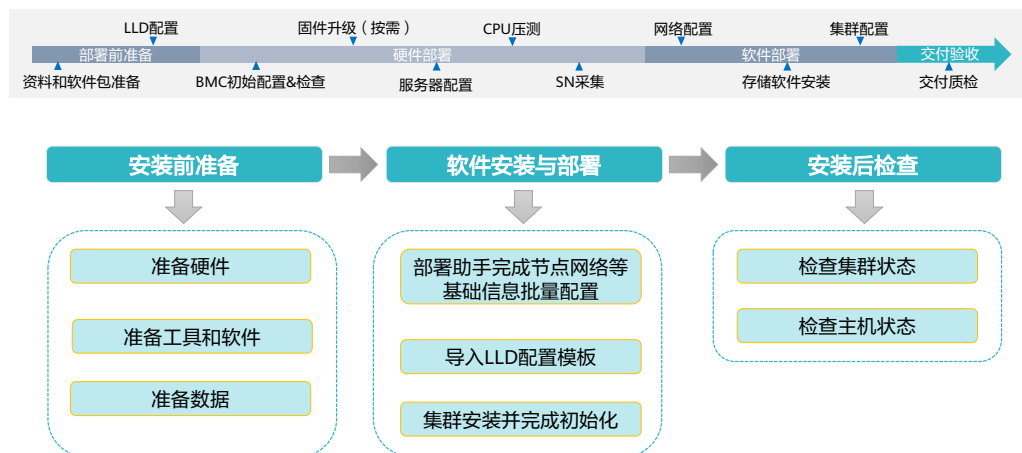
- 管理网络：与客户管理网络对接，用于系统管理和维护，管理平面支持双网络平面（管理内部网络、管理外部网络）。
 - 管理内部网络：用于存储节点和管理节点之间的数据通信，包括部署升级，资源配置，性能告警上报，存储节点NTP时间同步，证书密钥更新等。
 - 管理外部网络：用于客户端访问管理界面，REST接口，CLI，告警转发，与外部NTP服务时间同步等。
- 存储网络：内部网络平面，用于系统内部所有节点之间的业务数据通信，存储平面支持双网络平面（存储前端网络、存储后端网络）。
 - 存储前端网络：用于处理存储节点之间的前台数据。例如，在文件服务DPC场景下，可用于与计算节点通信，处理DPC和OSD间的通信数据。
 - 存储后端网络：用于处理存储节点OSD和OSD间的后台通信数据。
- 业务平面：与客户应用系统对接，通过S3、HDFS、NFS、CIFS、POSIX等接口访问存储；
- 控制网络：控制网络为控制集群提供IP地址，用于处理系统内部组件与控制集群之间，以及控制集群内部的通信，通过控制网络提供的IP地址访问控制集群。存储前端网络存在多个IP地址的场景下，因系统可靠性要求，控制网络无法与存储前端网络合一部署，故产生此网络。
- 复制平面：用于复制节点间进行数据同步复制。

eService LLDesigner介绍



- eService LLDesigner可以实现项目规划可视化，自动生成规划文档。功能包括：自由创建、以及通过双活模板创建LLD项目、存储组网规划、存储资源规划、LLD文档和离线配置导出。
- LLDesigner工具平台：
https://support.eservice.huawei.com/#/Delivery/LLD_Storage;domain=storage

分布式存储部署流程



- 请根据要安装的服务类型及场景（例如：文件/对象/大数据服务的标准协议场景）选择对应的安装流程。执行软件安装时，需要先配置存储交换机、管理交换机和各节点的网络，然后通过DeviceManager Client工具完成所有管理节点的安装，再通过管理节点完成存储节点的安装。
- OceanStor Pacific系列存储出厂时已经预装了操作系统和相关驱动，当操作系统和相关驱动的版本和版本配套表的版本一致时不需要单独安装，可直接通过DeviceManager Client工具安装OceanStor Pacific系列软件。如果操作系统和相关驱动的版本与版本配套表不一致，则需要重新安装各节点的操作系统和相关驱动。

部署前准备

交付阶段	任务类别	实施内容	使用工具
部署前准备	部署前准备	准备部署需要用到资料、工具和软件包	NA
硬件部署	设备上架及上电	设备上架及上电	NA
	组网连线及交换机配置	组网连线及交换机配置	网线、Console通信线缆
	BMC初始配置&检查	配置BMC IP、子网掩码/前缀、网关和BMC主机名 固件版本、OS版本和驱动版本配套检查	SmartKit分布式部署助手
	固件升级（按需）	BMC、BIOS、CPLD、RAID卡CPLD等固件升级	SmartKit分布式存储固件升级
	OS安装（按需）	批量安装EulerOS	SmartKit分布式存储开局交付
	服务器配置（按需）	BMC时区、用户设置、http开关设置、LLDP开关设置、BIOS密码修改、 资产信息和位置信息等	SmartKit分布式存储开局交付
	SN采集（按需）	采集设备SN	SmartKit存储健康检查
软件部署	OS配置	配置Bond、VLAN、存储前后端网络	SmartKit分布式部署助手
	驱动安装（按需）	安装网卡驱动、ES3000驱动等	SmartKit分布式存储开局交付
	存储软件安装	安装OceanStor Pacific系列软件	DeviceManager Client
	集群配置	关闭OS的LLDP，打开BMC移动定制开关	SmartKit分布式存储开局交付
	云平台对接	云平台对接	NA
交付验收	交付质检	对集群进行开局质检，配置是否满足要求	SmartKit开局工程质检
	客户验收测试	客户验收测试	NA

LLD配置

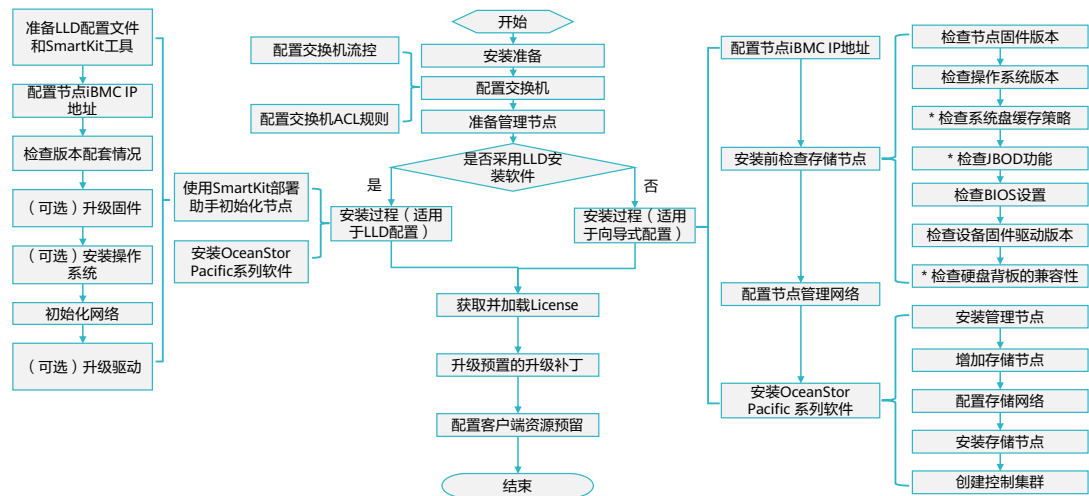


支持的子场景

- 块服务
- 文件服务（DPC场景）
- 文件/对象/大数据服务（标准协议场景）
- DPC和标准协议混合场景

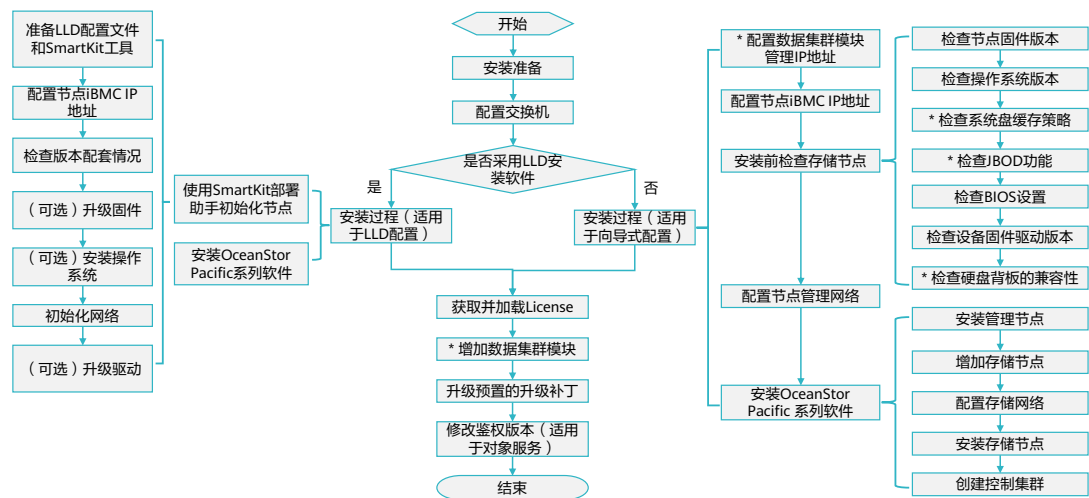
- 软件安装支持LLD配置和向导式配置两种方式，推荐使用LLD配置方式：
 - LLD配置模式：
 1. 需要先使用LLDesigner工具完成节点规划、网络规划和集群规划，并导出部署配置文件；
 2. 根据部署配置文件使用SmartKit部署助手完成节点初始化（节点初始化包含：配置节点iBMC IP地址、安装前检查存储节点、配置节点管理和存储网络）；
 3. 通过DeviceManager Client采用LLD配置模式完成OceanStor Pacific系列软件的安装。
 - 向导式配置模式：
 1. 手动在DeviceManager界面上完成节点添加、网络配置和控制集群的创建；
 2. 安装OceanStor Pacific系列软件。

块服务场景软件安装流程



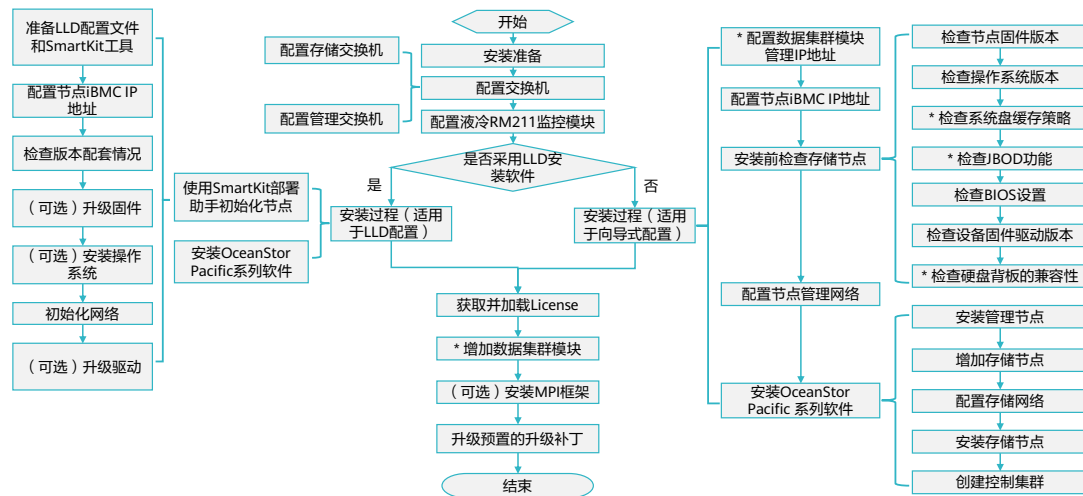
- OceanStor Pacific系列配套FusionCompute的典型分离部署方案，完成软件安装部署后，需要创建存储池并启用VBS，以便应用服务器能够通过VBS访问分布式存储资源。
- 其他块服务场景请参考产品文档。

文件/对象/大数据服务（标准协议场景）软件安装流程



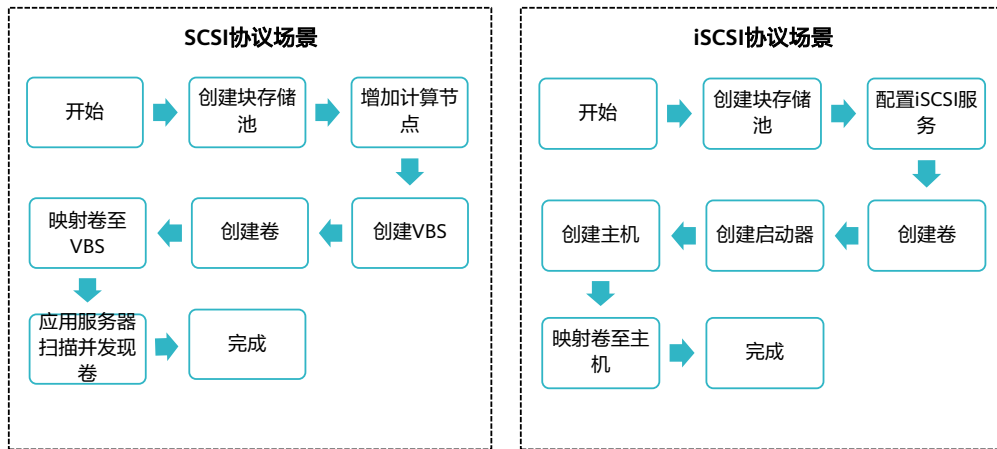
- 标准协议场景的软件安装包含管理节点和存储节点的安装部署，采用管理节点和存储节点合并部署的方式，存储网络采用“前后端共享网络”模式。
- 带 * 号内容表示适用于部分机型。

文件服务（DPC场景）软件安装流程



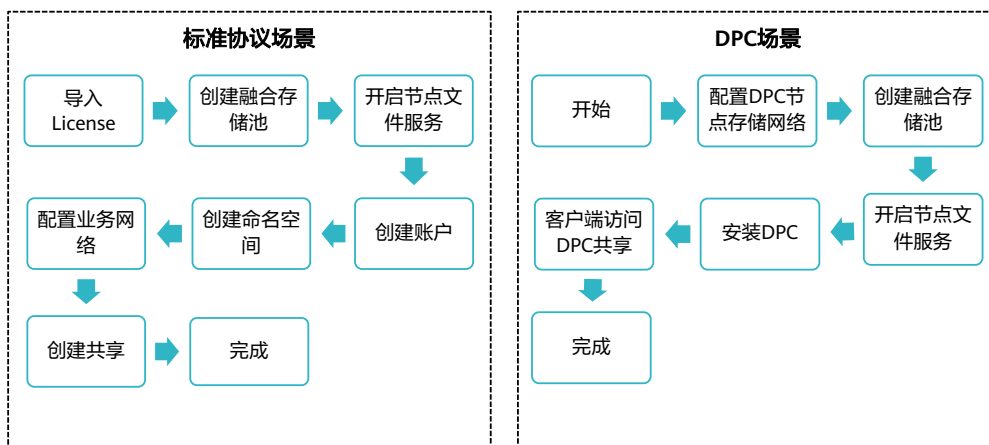
- DPC（Distributed Parallel Client）分布式并行客户端场景软件安装包含管理节点和存储节点的安装部署，采用管理节点和存储节点合并部署的方式。
- 同时也包含了在计算节点上预安装DPC软件的相关内容。
- 带 * 号内容表示适用于部分机型。

块服务业务配置流程



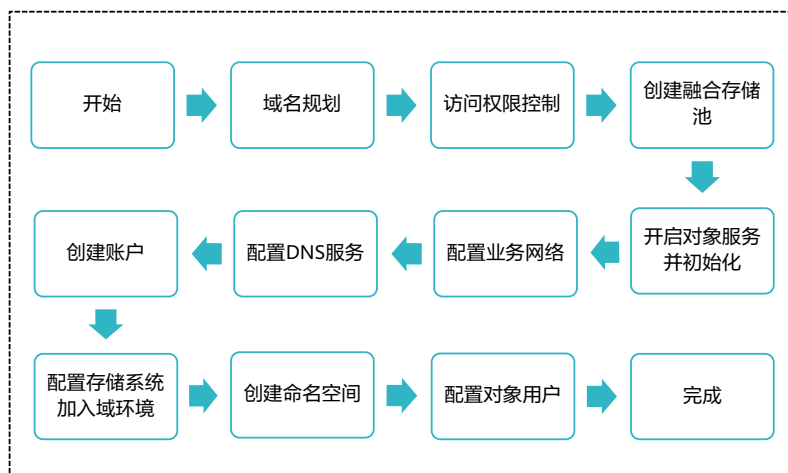
- 存储系统通过SCSI和iSCSI接口向上层应用提供高性能与高扩展能力的块服务，满足虚拟化、云资源池及数据库等场景的存储需求。
- 块服务当前的主要应用场景分为两大类：
 - 资源池：将存储节点池化，建立大规模块存储资源池。可以和各种虚拟化平台集成，如华为FusionSphere、VMware vSphere、开源OpenStack等，按需分配存储资源，支持各种业务应用（如SQL、Web、行业应用等）。
 - 数据库：以SSD做Cache或主存，极大地提升了存储系统的性能，又保留了分布式存储的高扩展性基因，从而解决了企业关键数据库、关键ERP（Enterprise Resource Planning）/CRM（Customer Relationship Management）等应用的高性能需求。

文件服务业务配置流程



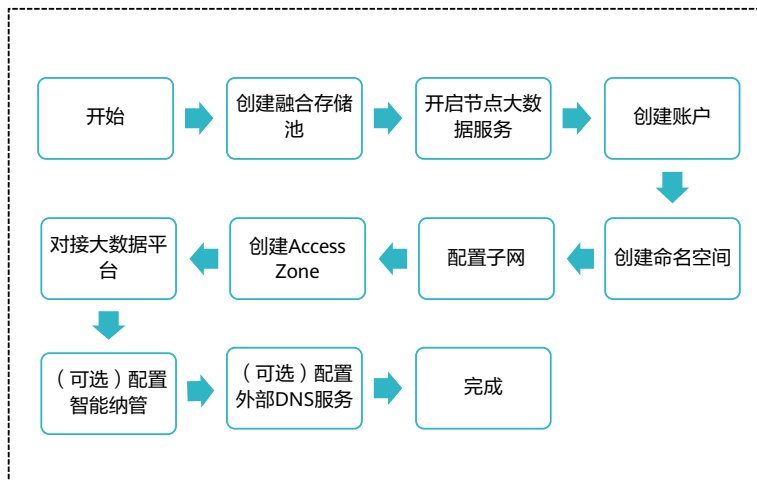
- 文件服务用于提供稳定可靠的文件共享功能。
- 文件共享是NAS存储系统的基本特性。OceanStor Pacific系列文件服务同时支持CIFS、NFS协议访问共享文件和DPC客户端访问共享文件，兼容多种操作系统，可以实现异构网络环境下的文件资源共享，支持对命名空间同时设置CIFS、NFS和DPC共享。

对象服务业务配置流程



- 对象服务是基于对象的海量数据存储服务，提供易扩展、高安全、高可靠、低成本的数据存储能力。适合存储任意类型的文件，一般应用于大规模数据存储场景，如互联网海量内容（视频、图片、照片、图书、音像、杂志等）、网盘、数字媒体、备份、归档等。
- 对象服务对外提供标准S3协议接口，是一种基于HTTP/HTTPS协议的REST风格协议，用户可通过对象服务客户端、API接口以及SDK方式方便地管理和使用对象服务数据，以及开发多种类型的上层应用。

大数据服务业务配置流程



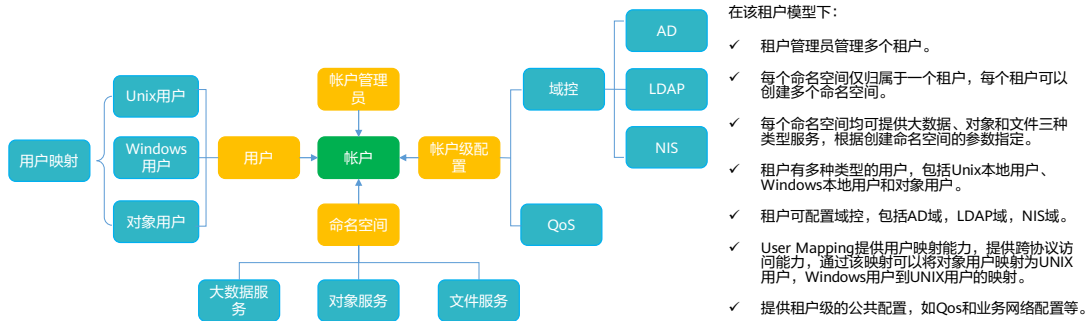
- 大数据服务提供基于原生HDFS的大数据存算分离方案，实现存储与计算资源按需配置，提供相同用户体验的同时，助您降低总拥有成本；支持与原有计算存储一体化架构共存。广泛应用于金融大数据、互联网日志留存大数据、政务大数据和智慧城市大数据等场景。
- 大数据服务使用OceanStor Pacific系列存储，提供原生HDFS接口，对接FusionInsight、Cloudera CDH、Hortonworks HDP等大数据平台，完成大数据的存储与计算，提供大数据分析服务供上层大数据应用使用。

目录

1. Ceph
2. 分布式存储产品介绍
3. 分布式存储产品部署
- 4. 分布式存储特性**
 - **效率提升**
 - 数据保护

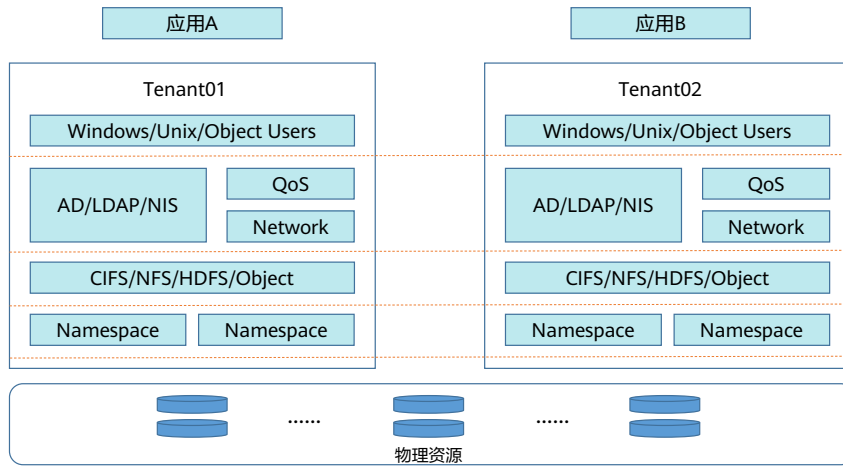
SmartMulti-Tenant介绍（文件/对象/大数据服务）

- 多租户特性提供基于租户粒度的统一资源管理，以租户为单位分配和管理资源。
- 多个租户共享同一套物理存储系统，租户间资源隔离，确保安全性和隐私。
- 多租户管理提供了一个通用的可扩展的多租户管理模型，通过多租户管理可实现租户级的命名空间管理、配额管理、QoS管理、域控配置等。



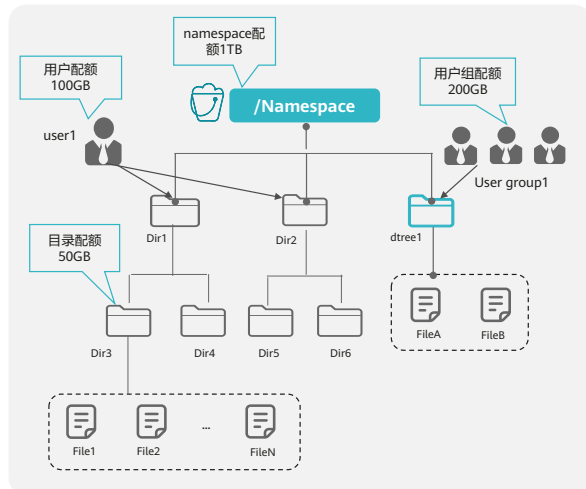
- 一个租户可以有多个Namespace，每个Namespace可以同时提供大数据、对象和文件服务，一个Namespace只能属于一个租户。多租户特性提供基于租户粒度的统一资源管理，以租户为单位分配和管理资源。多个租户共享同一套物理存储系统，租户间资源隔离，确保安全性和隐私。多租户管理提供了一个通用的可扩展的多租户管理模型，通过多租户管理可实现租户级的配置。

业务隔离模型



- 多租户的业务隔离，主要体现为租户间的业务数据、业务访问和业务配置隔离。业务数据隔离主要指租户管理员分配不同的命名空间给不同的租户，以此达到租户命名空间的隔离。业务访问隔离主要指每个租户都具有独立的访问协议。业务配置隔离主要指每个租户可以有自己独立的用户、用户组、LDAP域、NIS域等。

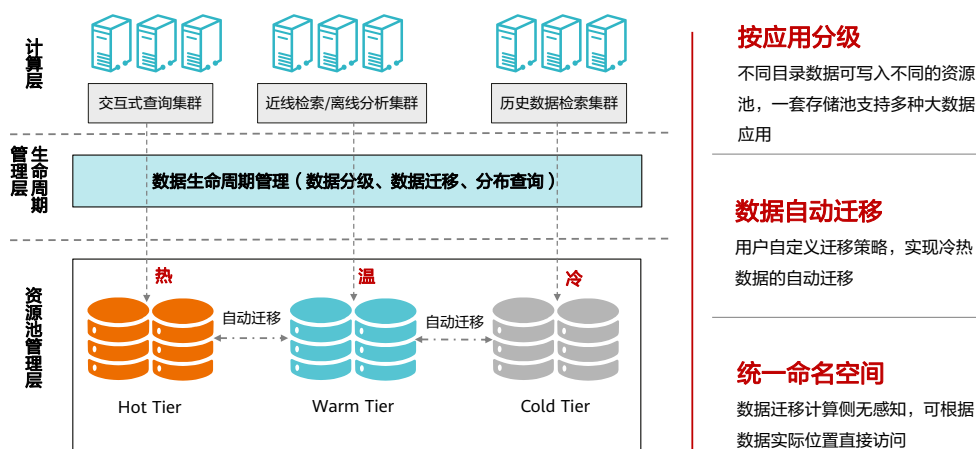
SmartQuota介绍（文件/对象/大数据服务）



- **多级配额**
 - ✓ 命名空间 & Dtree配额
 - ✓ **目录配额**（支持任意目录配额，非空目录配额，支持8层嵌套）
 - ✓ **租户配额**
 - ✓ 用户&用户组配额
- **多种类型配额**
 - ✓ 容量配额、文件数配额
- **多种配额阈值**
 - ✓ 建议配额/软配额/硬配额
 - ✓ 支持软配额宽限时间配置
- **多协议互通**
 - ✓ 支持DPC、NFS、CIFS、S3、HDFS、FTP写入文件

- 配额是一种限制或者统计所使用的存储空间和文件数量的技术。
- OceanStor Pacific系列文件/对象/大数据服务支持SmartQuota配额特性，可以应用于帐户、Namespace或者Dtree（Directory Tree），管理员通过配置合适的配额策略，限制帐户、Namespace或Dtree中资源的使用量，并在资源使用量达到配置的阈值级别时发送告警通知。
- Dtree可以在包括文件系统根目录在内的任意目录下创建，但Dtree间不能嵌套。
- SmartQuota提供简单易用的配置方式，使管理员能合理地为用户、用户组、
- Dtree分配存储资源，避免因某个帐户、用户、用户组、Dtree的过度使用造成其他用户无法正常工作，甚至影响系统运行，避免由于存储资源使用的失控，可能造成的系统崩溃，提高了系统的安全性。

SmartTier介绍（文件/对象/大数据服务）



114 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



- OceanStor Pacific系列文件/对象/大数据服务提供的SmartTier分级特性，允许将同一个存储池内的不同类型物理节点划分成不同的硬盘池。硬盘池是指具有相同特征（物理类型/访问性能）的节点的集合。SmartTier允许用户基于文件池策略定义 workflow 中数据的价值，将高价值的文件放置在高可用性、高性能的存储设备上，低价值的文件放置在成本较低的、性能和可用性规格较低的设备上。
- 如上图所示的Tier，是指具有相同特点/能力的硬盘池的集合。存储空间被分为热、温、冷三个等级，每个硬盘池都有对应一种分级等级，如SSD节点硬盘池为“热”，SAS节点硬盘池为“温”，SATA或NL-SAS节点硬盘池为“冷”。每个等级可包括若干个硬盘池。同一存储等级内部多个硬盘池间系统自动实现负载均衡，其中包括压力均衡和容量均衡。
- SmartTier允许管理员设置基于文件属性的文件池策略，使得数据可以根据存储策略在不同的硬盘池(硬盘池是EC数据分布的逻辑资源池，还是EC数据故障隔离域)之间自动迁移，从而在保证高性能的同时，更合理的使用存储空间。也允许用户手动规划自己的业务，根据自己的业务特点进行分级，把某些重要的目录，文件存放在性能较高的分级上。这些策略包括文件名、文件大小、创建时间、修改时间、访问时间、状态修改时间、UID/GID。针对对象还有一些独有属性，如前缀匹配，对象tag匹配，当前版本过期时间匹配，历史版本过期时间匹配。
- SmartTier的分级策略分为：放置策略、迁移策略和删除策略。
- 放置策略：
 - 用于确定文件数据初始放置位置，该类策略条件包括：FS/Dtree、文件名、

UID/GID。如果未匹配到任何策略，根据系统默认的放置策略决定放置位置。

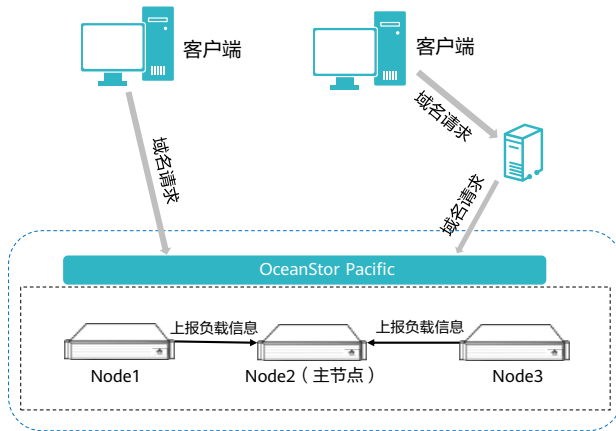
- 迁移策略：

- 周期性迁移策略：用于确定文件数据搬迁的目标位置。该类策略条件包括：FS/DTree、文件名、文件大小、创建时间、修改时间、访问时间、状态修改时间、UID/GID。系统自动按照策略将数据搬迁到指定等级的硬盘池中。数据搬迁的时间可配置，避开业务的高峰期，默认在每日凌晨零时执行。
- 一次性迁移策略：用于确定文件数据搬迁的目标位置。该类策略条件包括：文件名、文件大小、创建时间、修改时间、访问时间、状态修改时间、UID/GID。系统自动按照策略将数据搬迁到指定等级的硬盘池中。顾名思义，该搬迁策略只会在配置后执行一次，只有执行当时符合策略的文件数据才会搬迁。

- 删除策略：

- 周期性删除策略：周期性删除指定的文件。该类策略条件包括：文件名、文件大小、创建时间、修改时间、访问时间、状态修改时间、UID/GID。对于开启S3服务的命名空间，还可以指定前缀、标签、过期策略、非当前版本过期天数等条件。系统自动按照删除策略，直接删除文件，无需用户干预。
- 一次性删除策略：用于用户删除指定的文件。与周期删除策略不同，该任务只会执行一次。

SmartEqualizer介绍（文件/对象/大数据服务）



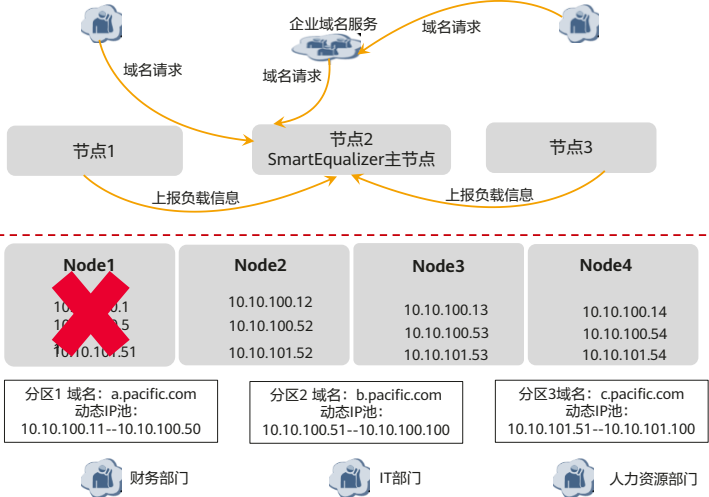
- SmartEqualizer是华为OceanStor Pacific系列分布式存储服务提供的负载均衡特性，SmartEqualizer周期性的采集存储系统各个节点的CPU利用率、网络吞吐量等信息，提供一种可配置负载策略的负载均衡服务。

- 各节点通过Paxos算法选出SmartEqualizer主节点(同一时间点只有一个主节点)。
- 从节点周期采集各自节点负载信息上传至主节点作为负载均衡的依据。
- SmartEqualizer基于当前配置负载策略进行负载计算并返回合适的业务IP给客户端。
- 若客户端无DNS服务器，可以配置本机DNS服务器地址为存储系统的DNS IP，直接访问存储系统的DNS服务。

- SmartEqualizer负载策略：

- 轮询方式(默认策略)：按顺序依次选择节点处理客户端连接请求。
- CPU使用率：选择CPU使用率最低的节点处理客户端连接请求。
- 节点连接数：选择文件服务连接数量最低的节点处理客户端连接请求。
- 节点吞吐量：选择网络吞吐量最低的节点处理客户端连接请求。

接入层负载均衡及故障换路



• 域名负载均衡

SmartEqualizer基于当前所配置的负载均衡策略进行负载计算，返回合适的节点IP给客户端，供客户端接入访问存储系统。

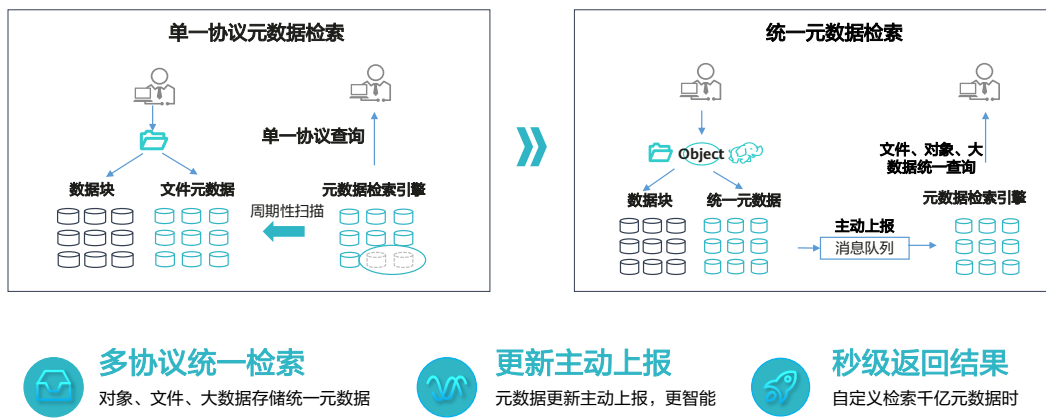
• IP资源池化

单个网络接口支持16个网络分区，每个分区可以归属不同的租户、子网，实现安全隔离及分区内负载均衡。

• 负载均衡 & 故障自动接管

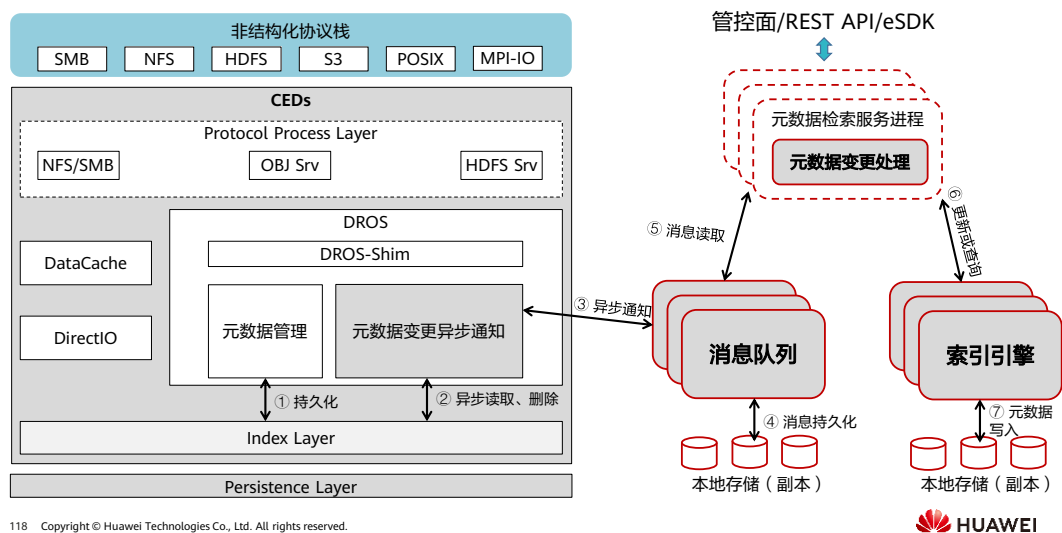
分区内IP池动态分配到每个节点，节点故障后，IP可以自动被多个节点分别接管，保障可靠性及负载均衡；也可以配置为IP地址不自动漂移的静态IP资源池模式。

SmartIndexing介绍（文件/对象/大数据服务）



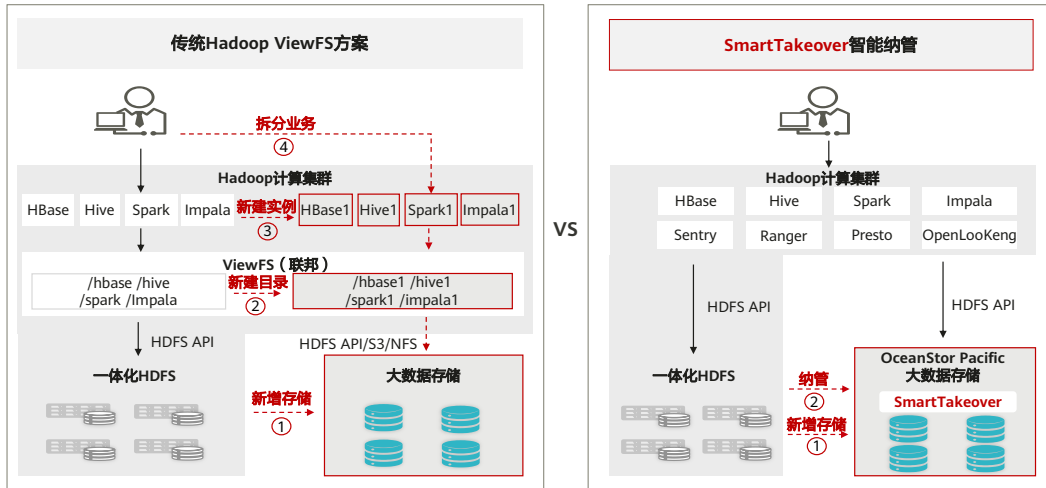
- 海量数据场景，用户需要基于元数据才能对海量数据进行管理，例如获取后缀名为 jpeg 的所有图片文件列表、获取文件大小大于 10GBytes 的文件列表、获取给定日期之前创建的文件列表，在用户能够快速获取符合条件的文件或是对象列表之后才能高效管理对应的数据。
- OceanStor Pacific 系列提供了基于非结构化（文件、对象和大数据）存储池的元数据检索功能 SmartIndexing，支持以命名空间(或文件系统)为粒度开启或是关闭元数据检索功能。对于开启了元数据检索的存储系统，在前台业务 I/O 变更文件、对象或是目录的元数据之后，异步、主动推送变更的元数据到检索系统。有别于传统存储产品采用的周期性元数据扫描，在避免对生产系统带来性能影响的同时，满足用户对海量数据的快速检索，例如，在千亿级元数据规模情况下，检索结果返回时间可低至秒级。该特性广泛适用于电子票据影像、自动驾驶训练等业务。
- 应用场景：电子票据影像、自动驾驶数据预处理等。
- 特性价值：目录层次多、海量文件场景快速检索；支持多维度的元数据检索，检索维度是 find 的 2 倍以上。

元数据检索原理



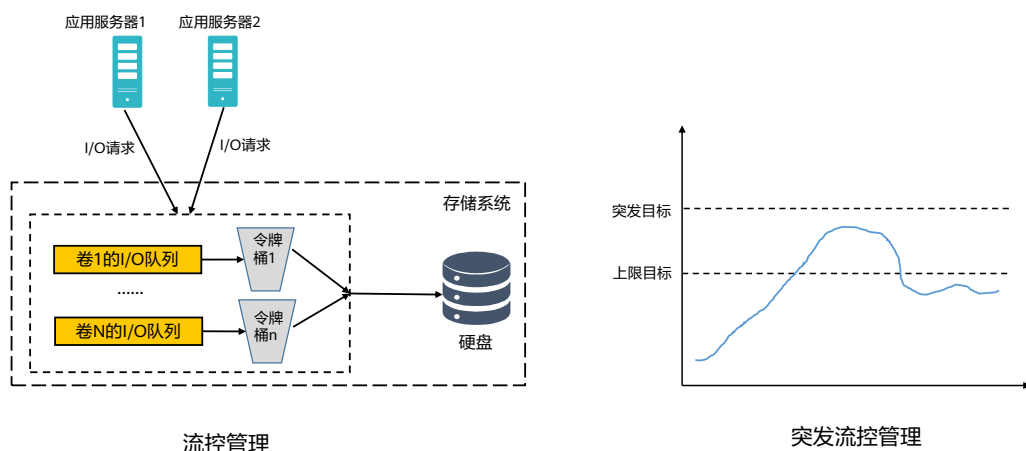
- OceanStor Pacific采用统一的存储底座进行用户数据及元数据的管理，在该底座上支持各种数据访问协议，从而实现对NFS/CIFS/HDFS/S3/POSIX协议写入数据的元数据检索，不同协议下发的操作都由同样的模块进行处理，统一了非结构化数据的元数据，从而轻松支持多协议（NFS/CIFS/HDFS/S3/POSIX）的元数据检索。
- 持久化（persistence）：一种数据存储过程，该过程将数据写入磁盘文件等存储介质，以供内存读取。
- OceanStor Pacific海量存储引入快速事务，将元数据变更日志和变更之后的元数据持久化（上图步骤①）。事务的引入确保了元数据的一致性，要么变更日志和元数据变更都成功，要么都失败。业务进程中的“变更处理”将异步批量读取元数据变更日志（上图步骤②）并推送到元数据检索的“消息队列”（上图步骤③）。“消息队列”在接收到消息请求之后，采用副本的方式（推荐3副本）持久化到本地存储（上图步骤④），用于缓冲来自存储节点的消息推送压力。“消息队列”在将信息持久化之后将通知“变更处理”模块删除本次通知的消息，释放其占用的存储池空间。
“元数据检索服务”将从消息队列读取消息（上图步骤⑤），在处理之后将写入“索引引擎”（上图步骤⑥）并将该消息从“消息队列”的本地存储空间中清除。至此，一条元数据的变更消息处理完毕。

SmartTakeover介绍（大数据服务）



- 智能纳管是指将OceanStor Pacific系列存储转换为网关集群从而具备统一管理多个HDFS存储服务的能力。
- 智能纳管特性仅适用于大数据HDFS服务。
- OceanStor Pacific系列提供了一种管理多个HDFS命名空间的方法。假设现在有两个集群，两个命名空间，且两个命名空间之间相互独立，物理存储也独立，无法做到在不迁移数据的情况下两个命名空间同时共存。SmartTakeover对外提供全局命名空间，对内通过一定规则进行元数据管理，实现在不迁移数据的前提下对HDFS命名空间进行纳管。
- 智能纳管主要分为全转发模式、本地优先模式、本地联邦模式、网关联邦模式。全转发模式下，元数据网关部署在存储集群中，唯一的被纳管HDFS存储服务称为“远端”。本地优先模式下，元数据网关部署在存储集群中，存储集群称为“本地”，唯一的被纳管HDFS存储服务称为“远端”。本地联邦模式下，元数据网关也部署在存储集群中，但不同于本地优先模式，本地联邦模式可以纳管多个HDFS存储服务。而网关联邦模式为元数据网关和存储集群分别独立部署。

SmartQoS介绍（块服务）



- 分布式应用场景，存储系统承载业务的多样性和复杂性显著增大，保障关键业务服务质量，保证系统提供稳定服务的能力显得尤为重要。SmartQoS（Quality of Service）特性就是为了解决融合虚拟化架构下混合业务的资源使用问题，避免业务对资源的自然争抢，保障单个或某些业务的存储性能。
- SmartQoS通过动态地分配存储系统的资源来满足某些应用程序的特定性能目标。
- OceanStor Pacific产品的QoS策略分为共享型和独享型。
 - 共享型QoS策略以存储池为粒度提供限流功能，配置的QoS策略只支持关联存储池，限制存储池的总性能不超过配置上限值。
 - 独享型QoS策略可以关联卷、快照和存储池，最终生效目标对象为单个卷或快照。限制单个对象（卷、快照）的性能不超过配置上限值。
- 流控管理
 - SmartQoS策略的流控功能通过I/O队列管理、令牌桶管理和出队控制三部分实现。
 - SmartQoS流量上限控制采用经典的令牌桶算法。存储系统将SmartQoS策略配置的实际性能上限转化为令牌桶的令牌速度，存储系统以该速度每秒向令牌桶中补充对应数量的令牌。每一个关联了SmartQoS策略的被控对象各自维护着一条I/O队列和一个令牌桶，一个I/O请求对应消耗一个令牌，当令牌桶中令牌消耗完后，新来的I/O请求将进入队尾等待，待补充令牌后，队首的I/O将优先出队。

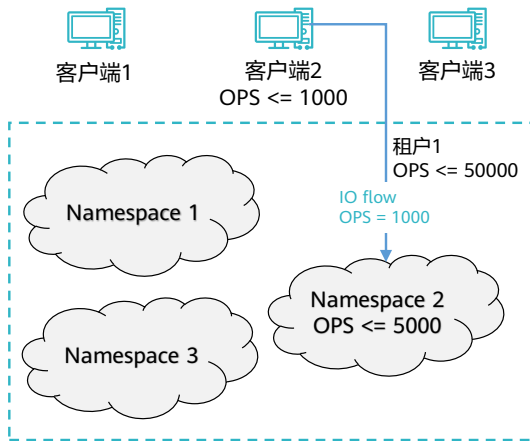
- 突发流控管理

- 为了应对某些应用的突发压力，存储系统支持Burst功能，当卷在需要高性能的情况下，能够在一定的时间内实现每秒可进行读写操作次数的突增。
- 其工作原理为积攒下前期末消耗的性能，当业务压力突然增大的情况下，可利用此前积攒下的性能，短时间获得突破上限目标的性能。长时间来看，其平均流量应低于上限目标。

- Burst功能包括：

- IOPS型的Burst功能：影响的性能参数包括IOPS和时长。Burst IOPS表示单个卷使用Burst功能时，每秒能够进行读写操作次数的最大值，时长表示当被控对象实际的IOPS达到Burst IOPS时，所能持续的最长时间。
- 带宽型的Burst功能
 - 独享型：影响的性能参数为Burst MBPS/TB，表示单个卷使用Burst功能时，每TB容量每秒可传输的数据量的最大值，其中Burst MBPS表示单个卷使用Burst功能时，每秒可传输数据量的最大值。
 - 共享型：影响的性能参数为Burst MBPS/时长，表示存储池使用Burst功能时，在所持续的时间内每秒可传输的数据量的最大值。时长表示当被控对象实际的MBPS达到Burst MBPS时，所能持续的最长时间。

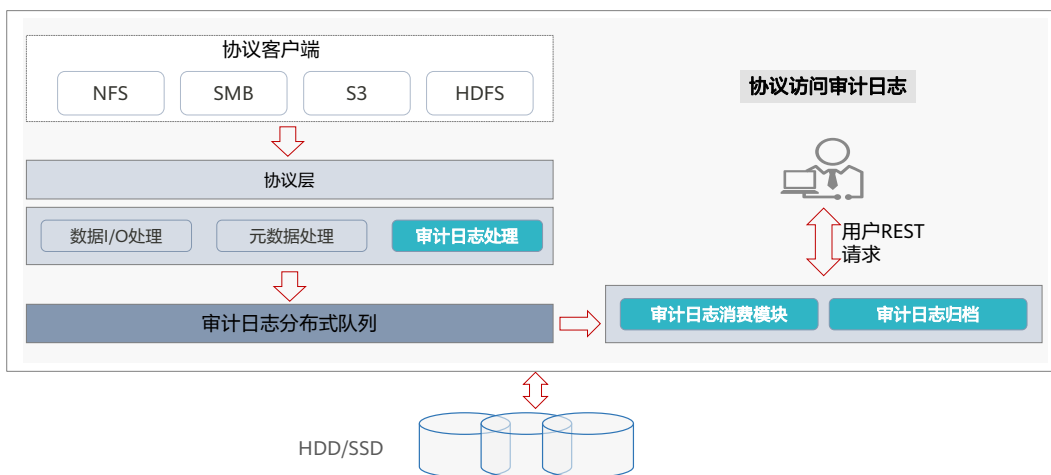
SmartQoS介绍（文件/对象/大数据服务）



- 华为OceanStor Pacific系列QoS技术支持以租户、命名空间与客户端IP为粒度进行控制。租户名下可以拥有多个命名空间，命名空间由文件服务和大数据服务的文件系统以及对象服务的桶所共享，客户端包括DPC（分布式并行客户端）和通用协议客户端（NAS协议、HDFS协议和S3协议）。其中，DPC客户端直接通过存储网络下发业务，不存在多个业务IP，因此控制粒度为单个DPC。对多个控制粒度同时施加QoS控制时，采用叠加式控制原则，即所有约束都生效，业务的实际流量由最严的约束决定。
- 如图所示，租户1的OPS上限被设置为50000其名下的命名空间2的OPS上限被设置为5000客户端2的OPS上限被设置为1000因此从客户端2下发到命名空间2的业务流量被限制在1000。

- 华为OceanStor Pacific系列QoS技术支持对总OPS（每秒操作数和总BPS（带宽）进行上限控制，针对文件服务和大数据服务，OPS的意义是存储服务端每秒处理的操作字总数，包括读写操作字和各类由客户端拆分出来的元数据操作字；针对对象服务，OPS的意义是服务端每秒处理的事务数，包括读（GET）、写（PUT）、LIST和DELETE。对象服务下，QoS在总OPS和总BPS控制之外，额外提供了读写操作字的OPS和BPS控制。
- 华为OceanStor Pacific系列存储服务器支持文件、大数据和对象服务的融合，采用命名空间统一涵盖文件系统和桶的概念，三大服务共享存储服务端的存储与访问能力，因此租户级与命名空间级 QoS 控制不针对某一种服务进行控制，而是对三大服务所共享的租户和命名空间的访问流量进行OPS和带宽的上限控制。
- 各个服务向同一个租户或命名空间下发业务时，相互之间可能争抢OPS和带宽，若有防扰邻需求，可搭配使用客户端IP粒度的QoS进行性能隔离。
- QoS技术通过创建和管理QoS对象实现不同粒度的QoS控制。任一租户、命名空间或客户端IP在配置了QoS控制后都将生成一个对应的QoS对象，该对象接收处理IO的令牌申请，判断是否需要排队，以此控制着与之对应的租户、命名空间或客户端IP的业务流量。
- QoS技术通过管理不同的令牌实现不同维度的QoS控制，例如IO通过申请总OPS令牌实现总OPS控制，通过申请写BPS令牌实现写BPS控制。

SmartAuditlog介绍



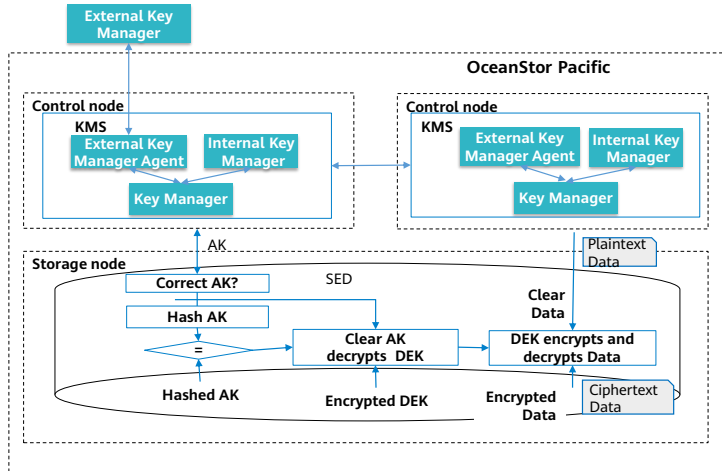
- OceanStor Pacific系列非结构化服务支持多协议互通访问的Namespace，在多协议支持下，NFS，SMB，S3和HDFS用户可以访问给定的文件。OceanStor Pacific系列审计特性，通过在Namespace上配置审计开关和操作字，记录发生在该文件系统上的文件或目录操作行为，支持审计日志的导出，导出文件格式支持evtx格式，审计日志的保留空间和时间可以由客户决定。
- 分布式存储节点协议层接收来自协议客户端（NFS，SMB，S3和HDFS）的文件访问操作时，协议层完成审计日志信息收集，传递审计日志信息给审计日志管理模块。审计日志管理模块完成审计日志记录序列化，推送审计日志记录到分布式队列。审计日志消费模块从分布式队列消费日志记录，并推送给审计日志归档模块，完成审计日志的归档存储。

SmartEncryption介绍

- 数据加密特性，支持通过内置密管和内置加密引擎完成静态数据加密，通过两层密钥管理，提供租户级安全性，从而保证用户的数据加密持久化和密钥管理安全性。
- OceanStor Pacific系列的SmartEncryption特性，支持硬盘加密，业务下发的所有数据都通过加密盘进行存储。支持NFS、CIFS、HDFS、S3或DPC任何一种协议加密写入，然后通过NFS、CIFS、HDFS、S3或DPC任意协议解密读取。文件服务客户端通过NFS/CIFS/DPC协议写入明文数据，加密盘从密管服务器获取身份认证密钥并将明文数据进行加密，然后持久化到加密盘中。当通过HDFS或S3协议读取时，加密盘从密管服务器获取身份认证密钥，然后在加密盘中读取密文数据并解密，再把明文返回给协议层，由协议层将明文数据返回给客户端应用。
- OceanStor Pacific系列的SmartEncryption特性，支持NFS、CIFS、HDFS或S3任何一种协议加密写入，然后通过NFS、CIFS、HDFS或S3任意协议解密读取。文件服务客户端通过NFS/CIFS协议写入明文数据，存储加密服务组件通过密管服务器获取身份认证密钥并将明文数据进行加密，再持久化到存储池中。当通过HDFS或S3协议读取时，存储加密服务组件首先从存储池中读取密文数据，然后通过密管服务器获取身份认证密钥并进行数据解密，再把明文返回给协议层，由协议层将明文数据返回给客户端应用。

- 本节以文件/对象/大数据服务场景为例，块服务场景原理相似。

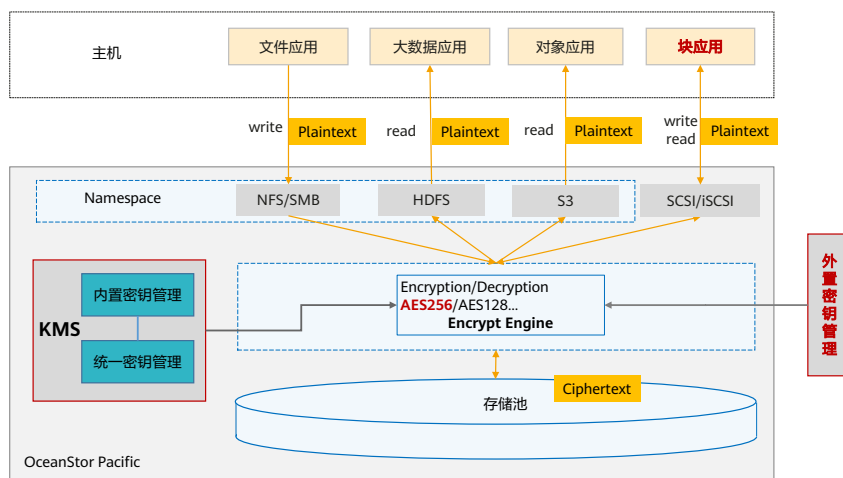
盘加密(SED Data at rest)



- 加密盘 (SED: Self Encryption Disk) :
 - SED具备两层安全保护, 分别使用 AK(Authentication Key) 和DEK(Data Encryption Key)两个安全密钥。AK: 身份认证密钥, 初始化硬盘时用于身份认证的密钥, AK由密钥管理服务 (Key Manager Service) 生成。DEK: 数据加密密钥, SED读写时用户加解密数据的密钥, 由加密盘生成。
 - 加密能力以存储池为粒度按需启用。

- 密钥管理服务 (KMS: Key Manager Service) :
 - AK密钥的获取, 支持内置密钥密管和外置密钥管理。
 - 外置密钥管理: AK由外置密钥管理服务器产生, 遵循标准的KMIP协议, 当前支持KMIP1.0、KMIP1.1、KMIP1.2三个协议版本。
 - 内置密钥管理: AK密钥由安全随机数产生, 采用三层体系结构保护密钥安全, 支持密钥的备份、恢复、更新、销毁。
- AK认证原理:
 - 当在存储系统上首次启用数据加密特性时, 存储会用KMS生成AK并计算HASH值, 并替换硬盘内的缺省AK的HASH, 且保存到硬盘中, 并对盘内置的DEK进行加密且保存在盘内。
 - 硬盘每次接入时, 需要存储系统从密管服务器获取硬盘的AK并计算HASH, 如果与硬盘上的AK HASH匹配, 硬盘就通过AK将盘内加密后的DEK解密, 解密后的DEK再用于数据加解密。
- AK修改原理:
 - 先用老的AK认证成功获取盘的访问权限, 然后采用KMS生成的新AK计算HASH并存放到盘上, 重新采用新AK来加解密DEK。

软件加密(Data at rest)



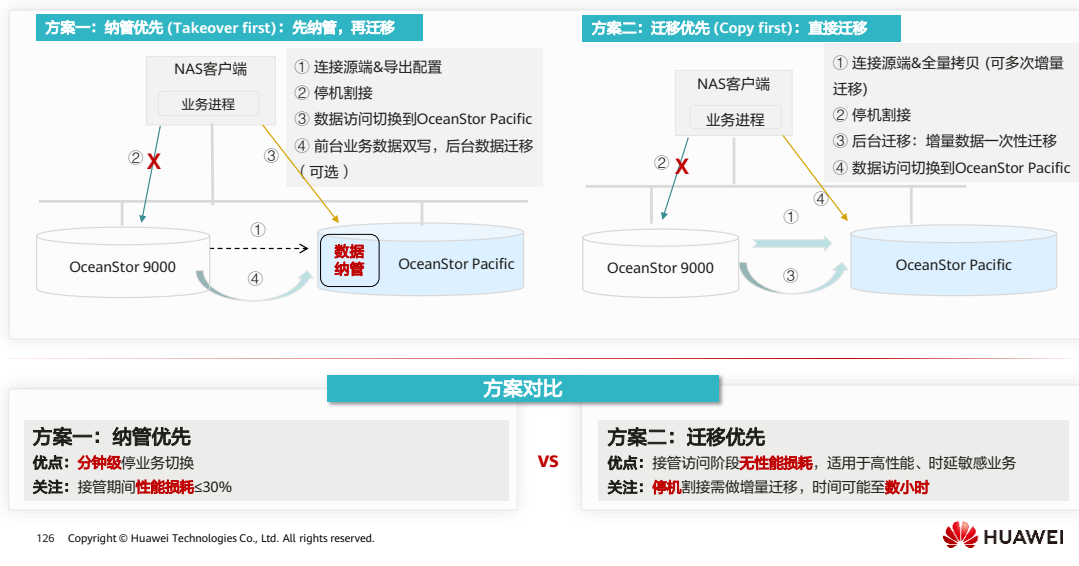
125 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



- OceanStor Pacific系列支持基于软件加密和加密盘两种静态数据加密能力，保护客户存储到系统中的数据不被泄露。
- 软件加密
 - 支持结构化和非结构化数据的软件加密，支持XTS-AES-128、XTS-AES-256和国密SM4-XTS三种加密算法。理论上，XTS-AES-256算法相较于XTS-AES-128算法对性能的影响不超过10%。软件加密通过内置密管和内置加密引擎完成静态数据加密，通过两层密钥管理提供帐户级安全。数据写入存储系统后，先实现软件层面的加密，再将密文写入到盘中保存；读数据时，先从盘中读取密文，再在软件层面解密成明文，然后再提供给用户。
- 功能介绍
 - 软件实现静态数据加密（Data-at-Rest）。
 - 支持结构化/非结构化所有标准协议（SCSI/iSCSI/NFS/SMB/S3/HDFS）。
 - 多层密钥设计，认证密钥（AK）和数据密钥（DEK），密钥支持更新。
 - 支持国际标准加密算法AES128-XTS/AES256-XTS和AES-NI指令加速。
 - 基于TLS证书的安全通道确保密钥传输安全，证书支持更新。
 - 加密能力以租户为粒度按需启用；基于最小块粒度(4KB)进行加解密。
 - 支持内置密钥管理和外置密钥管理，以及内置加密引擎。
- 密钥管理服务（KMS：Key Manager Service）

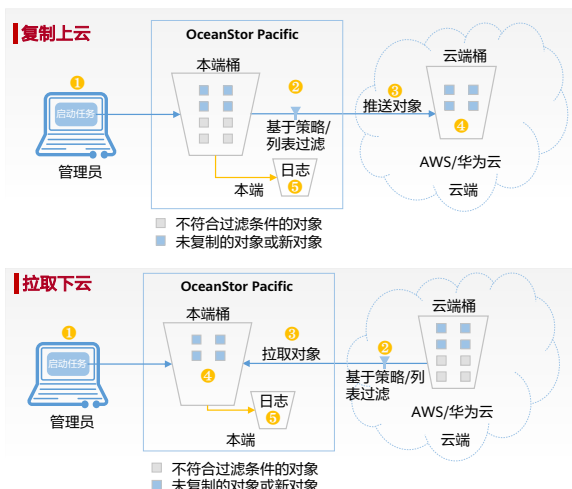
- 支持内置密钥密管和外置密钥管理。
- 外置密钥管理：遵循KMIP协议，支持KMIP1.0、KMIP1.1、KMIP1.2三个协议版本。
- 内置密钥管理：AK密钥由安全随机数产生，采用三层体系结构保护密钥安全，支持密钥的备份、恢复、更新、销毁。
- 加解密过程
 - 从KMS获取明文AK，将密文DEK解析成明文DEK。
 - 写操作，加密引擎使用明文DEK将写入的明文数据加密成密文数据，写入存储池。
 - 读操作，加密引擎使用明文DEK将从存储池读取到的密文数据解密成明文数据，返给应用。

SmartMigration介绍（文件服务）



- SmartMigration分为优先接管模式、优先拷贝模式和试运行模式。其中，优先接管模式优先接管业务，先将远端设备的业务割接到本端设备上，再进行数据迁移；优先拷贝模式优先进行数据迁移，再将远端设备的业务割接到本端设备上；试运行模式用于进行正式迁移前的预迁移检查，为正式迁移提供质量保证。
- 系统支持将源命名空间的任意共享迁移至目标命名空间或Dtree级别的共享。
- 优先接管模式
 - 在优先接管模式下，目标命名空间可以在数据没有迁移到本端存储系统的情况下对外提供SMB、NFS或跨协议共享。
 - 迁移完成后的优先接管模式的目标命名空间可以在客户端IO不中断的情况下断开与源命名空间的关系，切换为普通命名空间。
- 优先拷贝模式
 - 优先拷贝模式包括普通拷贝和最终拷贝两种后台拷贝方式，没有完成最终拷贝任务的命名空间或Dtree为只读状态。最终拷贝任务完成后，手动删除迁移任务，转为客户端可读写状态。

SmartSync介绍（对象服务）



关键技术点

双向的数据传输策略、自由的任务配置
灵活的带宽QoS、多样的过滤条件

方案价值

兼容标准的AWS S3协议
云上云下数据自由流动，数据共享无忧

- 随着机器学习等新兴技术的发展，以计算密集型负载为主要特征的HPC需求逐渐增长。由于借助大数据挖掘数据价值能够提升业务产出质量，因此基于海量数据存储利用HPC资源开展高性能数据分析（HPDA, High Performance Data Analytics）逐渐成为应用趋势。
- SmartSync特性为用户提供了将大容量存储资源和强算力资源联合使用的机会，利用S3协议建立了二者的数据流动通道。通过SmartSync特性，用户能够将海量业务数据从OceanStor Pacific系列存储推送至云端存储进行大数据分析、模型计算，快速开展高性能数据分析，满足基于数据密集型的高价值业务场景需求。
- SmartSync特性支持将海量业务数据复制上云、拉取下云，从而帮助用户开展高性能数据分析、进行数据备份。
- OceanStor Pacific系列的SmartSync特性可以实现以下功能：
 - 推送和拉取海量数据
 - 过滤数据
 - 控制带宽
 - 提供复制日志
- SmartSync特性支持对象复制上云、对象拉取下云两种模式。其中，对象复制上云模式（PUSH）将本端桶中的对象推送至云端桶，对象拉取下云模式（PULL）将云端桶中的对象拉取至本端桶。

SmartTransfer介绍（文件/对象服务）



☑ 支持文件和对象协议接入远端设备

- 文件协议：支持OceanStor Dorado/Lustre的远端设备接入
- 对象协议：支持OceanStor Pacific/华为云的远端设备接入

☑ 支持两种拉取模式

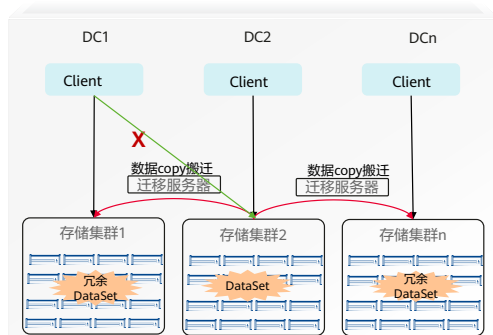
- Pre-fetch模式：提前预加载外接设备上的数据和元数据信息。
- On demand模式：支持用户获取外接设备上的元数据，数据按需获取。

☑ 支持灵活配置一次性或周期性任务策略

- 随着存储技术的不断发展和各行业的数字化推进，产生了因存储系统资源优化、利用以及业务性能调整而进行的存储系统间业务数据互相流动的需求。通过SmartTransfer特性，用户能够使用OceanStor Pacific系列存储设备与其他外接设备的数据流动。
- SmartTransfer特性为用户提供了将大容量存储资源和强算力资源联合使用的机会，利用文件协议或对象协议建立了二者的数据流动通道。用户能够将海量业务数据从OceanStor Pacific系列存储进行大数据分析、模型计算，快速开展高性能数据分析，满足基于数据密集型的高价值业务场景需求。

全域数据服务Omni-Dataverse（文件服务）

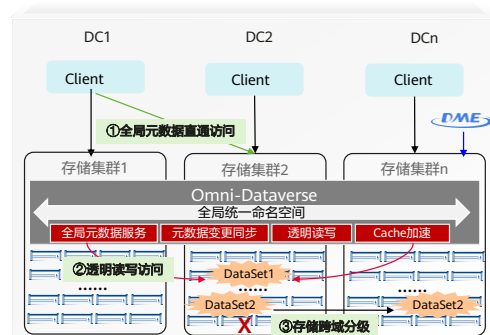
《》 集群烟囱式构建，集群间无联接，数据成孤岛 《》



关键问题：多集群为数据孤岛，跨集群的数据访问需要数据拷贝，无数据协同

- 拷贝需要迁移服务器，copy速度慢
- Copy数据到不同的数据集群，形成新的数据孤岛，产生新的冗余数据
- 不同集群间无统一元数据视图，无数据协同

《》 基于Omni-Dataverse，多个孤岛统一访问 《》



关键技术：构建元数据全局视图，提供统一数据访问与加速，数据免搬迁

- Client基于全局元数据支持数据直通存储访问
- 任意Client、任意站点无感透明访问，数据预取、缓存加速
- 支持数据跨域分级，数据协同、无冗余数据

- Omni-Dataverse，也被称为全域数据服务，是一种全局文件系统的技术。其核心特点是全局可视，即支持将多个集群的文件系统加入到一个抽象的全局文件系统中，一个集群的文件系统可以将自己的数据共享给其他集群，做到统一的数据视图，即全局可视。这种特性是通过在DME中构建的Omni-Dataverse全域文件系统实现的，使得企业数据资产变得可见和可管理。Omni-Dataverse特性主要包含全局元数据管理、元数据存储空间、透明读写和跨域数据分级这几个主要功能。

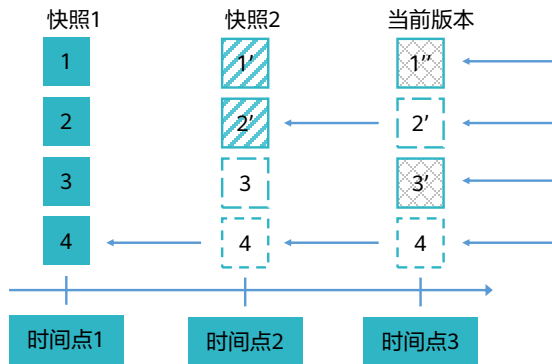
目录

1. Ceph
2. 分布式存储产品介绍
3. 分布式存储产品部署
- 4. 分布式存储特性**
 - 效率提升
 - **数据保护**

HyperSnap介绍

- 概述
 - OceanStor Pacific的快照是只读快照，数据不可更改。
 - 快照是指源命名空间或Dtree在某个时间点的镜像。
 - 快照生成后可以立即被读取。
- 主要特点
 - 瞬间生成：存储系统可以在秒级生成一个快照，获取源命名空间或Dtree数据的一致性副本。
 - 占用存储空间少：生成的快照数据并非完整的物理数据拷贝，不会占用大量存储空间。
 - 性能无损：基于ROW机制，快照性能与源命名空间性能相当。

ROW



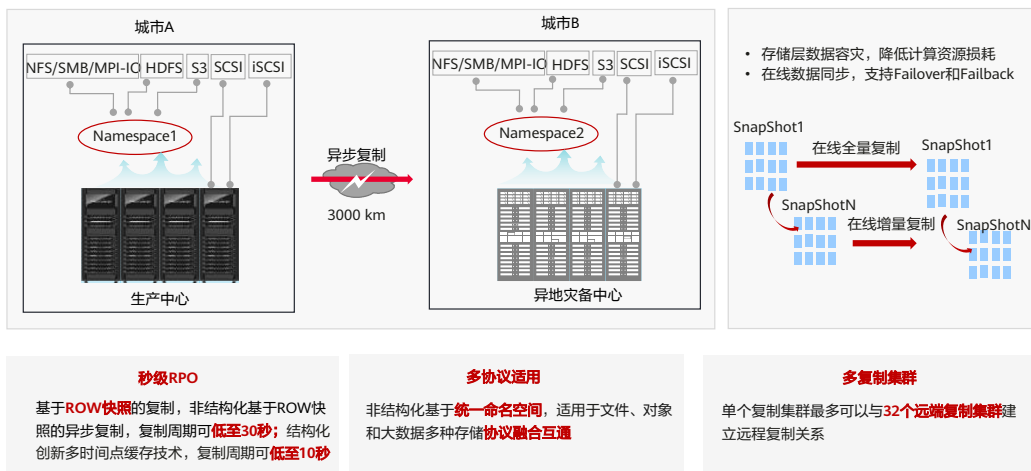
- OceanStor Pacific基于全局时间点技术来实现快照；
- 新的写操作使用ROW（Redirect On Write）技术，使用新的时间点将新数据写入新的对象，不覆盖原有对象。原始数据没有被覆盖写，被快照版本引用；
- 未发生变化的数据，不产生额外的副本，被只读快照和当前版本共同引用。

- 如图所示，创建快照Snapshot2之后，访问数据时，对于数据块1，由于当前版本产生了新的数据，则直接访问数据块1''；对于数据块2，由于当前版本并没有产生新的数据，则数据前向依赖于上一个快照版本的数据，访问快照Snapshot2的数据；对于数据块3，由于当前版本产生了新的数据，则直接访问数据块3'；对于数据块4，由于当前版本并没有产生新的数据，则数据前向依赖于上一个快照版本的数据，由于Snapshot2同样没有产生ROW数据，则最终访问快照Snapshot1的数据4。

HyperReplication介绍

- 随着各行业数字化进程的推进，数据逐渐成为企业的运营核心，用户对承载数据的存储系统的稳定性要求也越来越高。虽然企业可以拥有稳定性极高的存储设备，但还是无法防止各种自然灾害对生产系统造成不可恢复的毁坏。
- 为了保证业务数据的持续性、可恢复性和高可用性，远程容灾备份解决方案应运而生，而远程复制技术则是远程容灾备份解决方案中的关键技术之一。
- 远程复制有以下两种复制模式：
 - 同步远程复制：对主从两端存储系统的数据进行实时地同步，做到RPO=0，保证数据一致性，以减少灾难发生时的数据丢失量。
 - 异步远程复制：对主从两端存储系统的数据进行周期性地同步，做到RPO≥10秒，最大限度减少由于数据远程传输的时延而造成的业务性能下降。

HyperReplication(文件/对象/大数据服务)



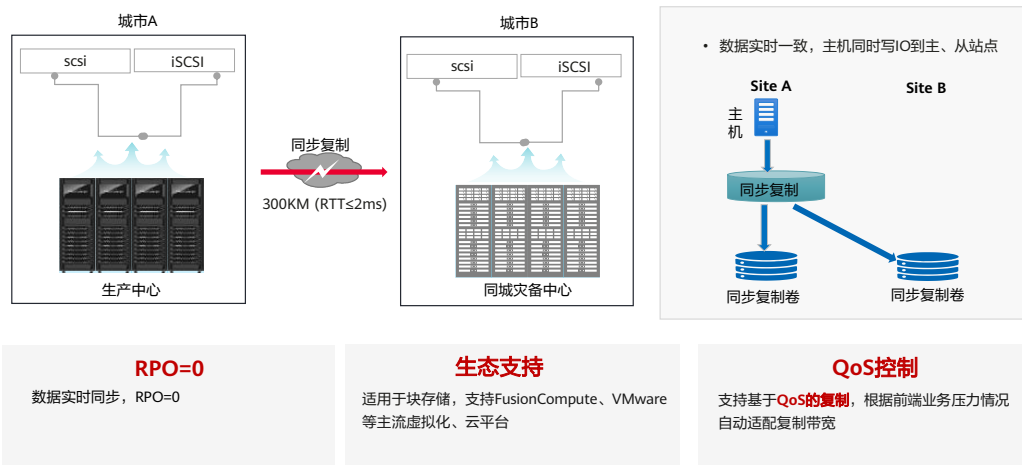
134 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



- 文件/对象/大数据服务只支持异步远程复制。
- 异步复制是对主从两端存储系统的数据进行基于只读快照的周期性增量同步，实现系统容灾，从而最大限度减少由于数据远程传输的时延而造成的业务性能下降。
- 异步复制特性主要功能包括：
 - 数据复制：远程复制是基于命名空间级的快照技术，实现主端存储系统和从端存储系统间的数据异步复制。命名空间的快照是命名空间在某个时间点的完整数据副本。远程复制基于快照进行数据同步，不会影响业务运行。远程复制包括两个阶段：初始同步和增量同步。
 - 初始同步：当主命名空间与从命名空间建立远程复制关系后，首先会启动初始同步，初始同步前创建主命名空间的快照，将该快照时间点的主命名空间数据全量复制到从命名空间，以保证从命名空间数据和主命名空间（某一快照时间点）数据一致。
 - 增量同步：初始同步完成后，后续复制通过快照差异日志同步增量数据。
 - 业务切换：当主站点遭受灾难，远程复制可以实现从站点迅速接管业务，保证业务持续运行。
 - 从端存储系统能够接管业务需要满足以下条件：
 - 从端存储系统中存在历史业务数据。从资源的数据状态必须是“完整”。
 - 备用主机上已配置与生产主机相同的业务应用。

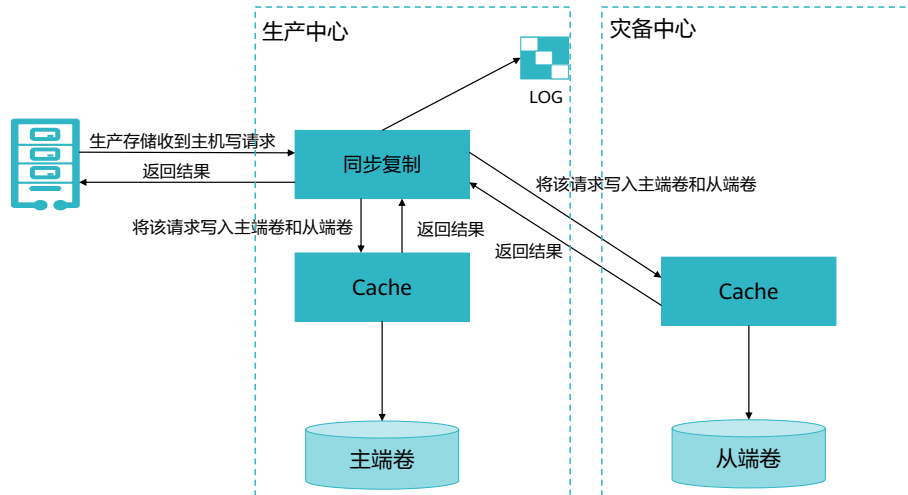
- 备用主机可正常读写从端存储系统的命名空间。
- 当灾难发生时，主站点失效，主从命名空间之间的远程复制链路异常断开。此时，需要管理员取消从资源保护，使从命名空间变为“读写”状态，以供备用主机访问。
 - 业务恢复：远程复制中的主站点失效后，从站点临时接管业务，待主站点故障恢复后，再将业务重新切换回主站点。当主站点故障恢复后，需要将主站点的数据同步到主站点，从而让主站点重新对外提供服务。

HyperReplication（块服务）



- 同步复制是主机写入生产的数据，会在生产存储和灾备存储都写入成功后，才返回主机写成功，能够保证RPO为0。基于同步复制，能够实现灾难恢复级别较高的数据级容灾（“第6级：数据零丢失和远程集群支持”）。

块服务同步远程复制原理

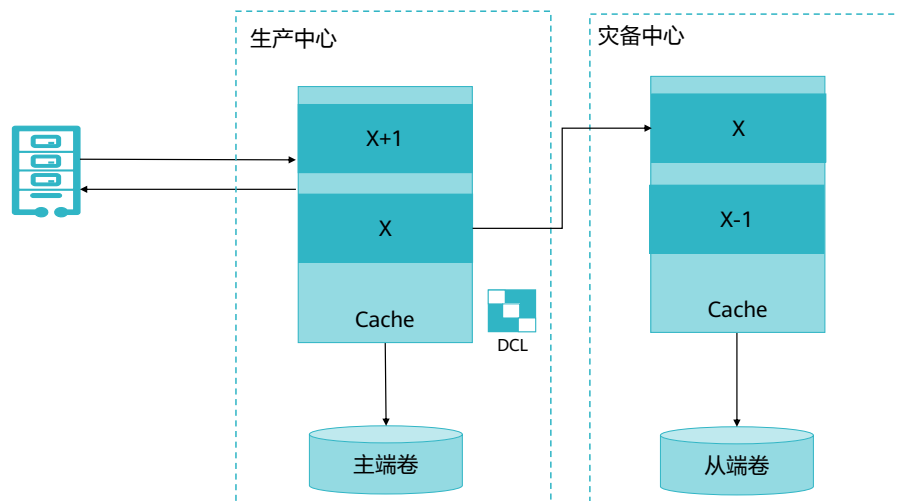


136 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



- 同步远程复制基于数据双写机制，将主端存储系统上的数据实时地同步到从端存储系统上，保证主从端存储系统的数据一致性。
- 同步远程复制包括两个阶段：初始同步和数据双写状态。
- 初始同步：生产中心的主端卷和灾备中心的从端卷建立同步远程复制关系，启动初始同步。
 - 将主端卷数据全量拷贝到从端卷。
 - 初始同步中主端卷收到主机写请求也会同样写到从端卷。
- 双写状态：初始同步完成以后，进入正常状态，此时主、从端卷数据相同。
 - 生产存储收到主机写请求。HyperReplication将该请求记录日志。日志中只记录地址信息，不记录数据内容。
 - 将该请求写入主端卷和从卷。
 - HyperReplication等待主端卷和从端卷的写处理结果都返回，如果写从端卷超时或失败时远程复制关系断开。如果都写成功，清除日志；否则保留日志，写入盘上DCL（Data Change Log 数据变更日志）中，进入异常断开状态，后续启动同步时重新复制该日志地址对应的数据块。
 - 返回主机写请求处理结果，以写主端卷的处理结果为准。
- LOG：记录数据写入操作的日志
- DCL：Data change log，数据变化位图

块服务异步远程复制原理



137 Copyright © Huawei Technologies Co., Ltd. All rights reserved.

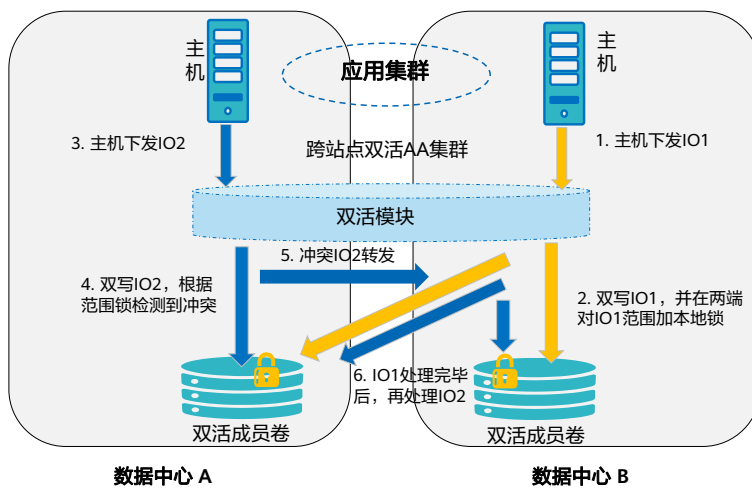


- 当生产站点的主端卷和远端复制站点的从端卷建立异步远程复制关系以后，默认情况下会启动一个初始同步，将主端卷数据全量拷贝到从端卷，以保证主端卷和从端卷的数据完全一致。
- 初始同步完成后，从端卷数据状态变为完整（即从端卷为主端卷的过去某个时刻的一致性拷贝），然后开始按照下面的流程进行I/O处理：
 - 异步远程复制同步任务启动时，在主端卷和从端卷分别生成快照并更新时间点（主端卷快照X，从端卷快照X-1）；
 - 主机新写入的数据缓存在主端卷缓存的时间点X+1中；
 - 响应主机写完成；
 - 将主端卷时间点X的数据基于DCL（Data Change Log）增量差异记录直接复制到从端卷，当数据同步完成后，删除主端卷与从端卷的快照；
 - 主从端卷各自将收到的数据下盘，数据下盘的同时记录DCL，数据下盘完成后从端卷最新数据即为主端卷时间点X的完整数据。

HyperMetro介绍

- HyperMetro特性中的两个数据中心互为备份，且都处于运行状态。当一个数据中心发生设备故障，甚至数据中心整体故障时，业务自动切换到另一个数据中心，解决了传统灾备中心不能承载业务和业务无法自动切换的问题。提供给用户高级别的数据可靠性以及业务连续性的同时，提高存储系统的资源利用率。
- HyperMetro特性具有以下特点及优势：
 - 基于两套集群构建Active-Active读写访问能力，任意一个数据中心故障，数据零丢失，系统自动切换到另外一个集群运行，做到RPO=0，RTO≈0，保证上层业务连续性。
 - 基于两套存储集群虚拟出一个跨站点的虚拟卷，该虚拟卷的数据在两个存储集群之间实时同步，且两套存储集群能够同时处理计算节点的I/O读写请求。
 - 具有良好的弹性扩展能力，对于大规模应用客户，每个存储集群可以配置多个节点，每个节点都可分担数据同步的负载，满足后续的业务增长。
 - 支持多种仲裁模式：第三方仲裁和静态优先仲裁。在配置的第三方仲裁服务器故障的情况下，系统自动切换到静态优先仲裁模式，提高业务连续性。

I/O处理机制

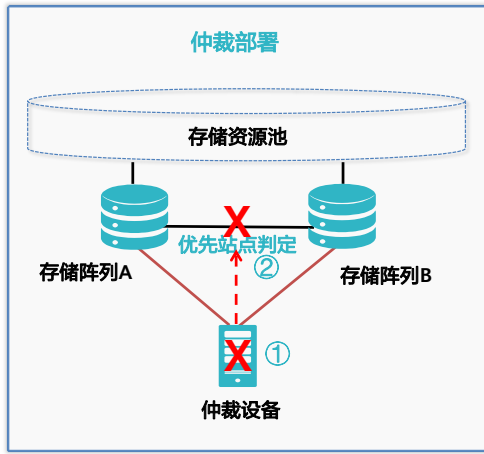


- 本端存储池和远端存储池向双活模块返回写I/O结果。对LOG的操作有两种情况：

- 如果都写成功：清除LOG。
- 如果有一端失败：双活Pair关系断开，写成功的一端继续提供业务，写失败的一端停止提供业务。待双活间通信恢复时，系统会根据LOG记录的信息将差异数据同步到写失败的一端。

- OceanStor Pacific系列结构化服务的双活支持增量同步，当一个站点发生故障时，按照系统的仲裁机制，会有仲裁获胜的站点提供业务，IO请求由双写转为单写，当故障站点恢复系统，支持增加将故障过程的数据同步到恢复的站点，快速恢复系统。
- 同时, OceanStor Pacific系列结构化服务增加了对逻辑写错误的处理，当系统正常运行，但是一端IO返回写失败的场景，支持将IO重定向写到正常的站点，写失败的站点故障修复以后，支持将增加的数据同步回来，减少逻辑写错误导致上层应用切换的问题。
- 通过数据双写和同步差异数据机制实现存储层数据的双活，本端和远端两个站点同时对主机提供数据读写能力。
- 实现两台存储系统数据一致的核心就是高效的乐观锁机制和高效的数据双写机制。
 - 在业务运行中，数据变更可通过双写和读取差异数据来完成同步。双写可以使得主机的I/O请求同时下发到本端存储池和远端存储池，从而保证本端存储池与远端存储池的变更数据一致性。在一个数据中心故障时，LOG可以记录业务运行中数据中心的数据变更。待故障恢复后，存储设备根据LOG记录数据变更进行数据同步，防止数据丢失。
 - 具备双活特性的两台存储系统，能够同时处理主机的I/O请求。为了防止不同的主机同时访问同一存储地址，而发生访问冲突，需要乐观锁机制来避免访问冲突，所有的双写I/O都会先尝试进行两个站点的双写操作，并在两端加本地锁，当检测到位置冲突I/O，则将I/O转发到一端进行等待处理。

双活仲裁机制



第一重：仲裁服务器模式

- **原理：**心跳不通时，仲裁服务器发起仲裁，仲裁获胜的一方继续提供服务，仲裁失败方停止服务，优先站点具有仲裁优先权

第二重：静态优先级模式

- **原理：**在未配置仲裁服务器或仲裁服务器无法访问时，双活仲裁模式为静态优先级模式。心跳不通时，预先设置的优先方存活。

- 当两个数据中心之间的链路故障或其中一个数据中心故障时，两个数据中心之间无法实时同步，此时只能由其中一个数据中心继续提供服务。为了保证数据一致性，双活系统通过仲裁机制决定哪端数据中心继续提供业务。
- 双活提供了两种仲裁模式：
 - 静态优先级模式：应用在无第三方站点仲裁服务器的场景。
 - 仲裁服务器模式：应用在有第三方站点仲裁服务器的场景。使用独立的物理服务器或者虚拟机作为仲裁设备，仲裁服务器建议部署在第三方站点。这样可以避免单数据中心整体发生灾难时，仲裁设备也同时故障。

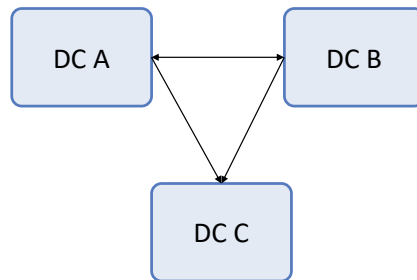
对象跨站点容灾技术

- 对象主要应用在备份归档场景，对可靠性和可用性要求高，通常要求具备跨站点容灾能力：

- 从可靠性上，分为副本冗余和EC冗余两种模式
- 从可用性上，分为AA (Active/Active)，AP (Active/Passive)两种模式

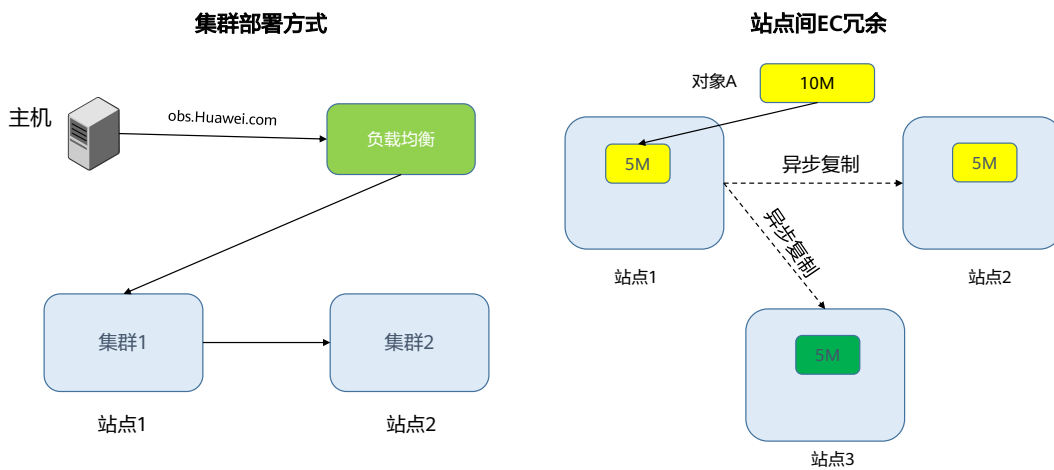
- 常见容灾模式：

- 两站点异步复制（AP，2副本）
- 两地三中心（AAP，3副本）
- 3AZ（AAA，EC）



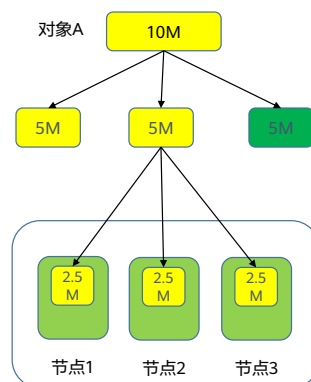
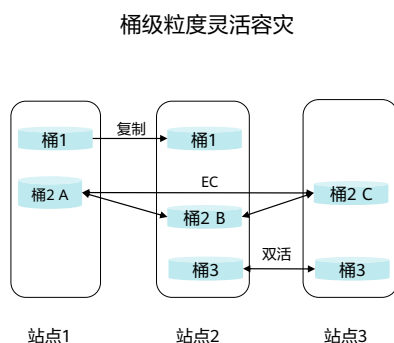
两地三中心

跨站点容灾可靠性（1）



- 集群部署方式
 - 全局域名：使用统一的域名访问。
 - 每个站点都是独立的，有独立的DM界面，独立的EC配置，独立部署，独立升级，独立运维，站点间松耦合。
- 站点间副本冗余
 - 数据按照副本方式存储，每个站点都拥有完整的对象。
 - 特点：读写效率高，但存储利用率低，适用于对时延要求较高的生产型业务。
- 站点间EC冗余
 - 数据按照EC方式存储，每个站点只拥有对象的一部分数据。
 - 特点：存储利用率高，读写效率低，适用于对时延要求不高的备份归档类场景。

跨站点容灾可靠性（2）



站点内3节点，冗余比4+2:1

- 桶级粒度灵活容灾

- 桶级粒度灵活容灾，同时支持副本和EC，兼容不同业务容灾诉求。
- 创桶时，可选择副本 or EC，以及部署在哪些站点。

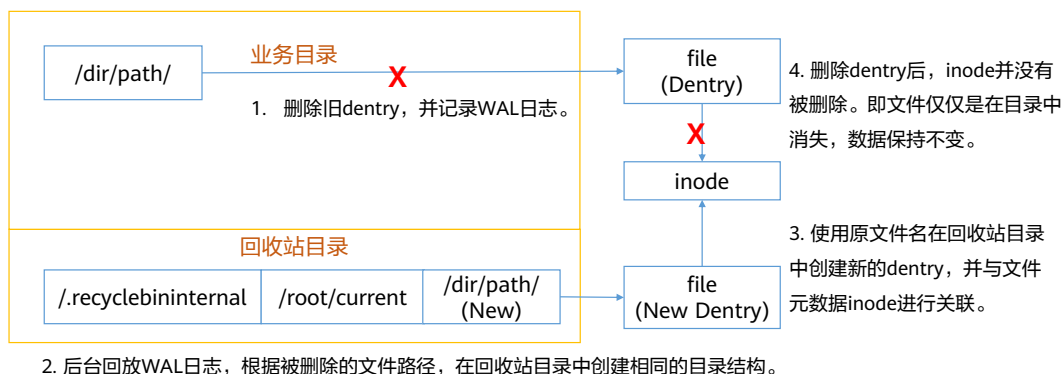
- 站点内的数据保护方式

- 每个站点都是独立的，站点内有冗余，站点内硬盘/节点故障在站点内重构，避免消耗站点间流量。
- 站点内按照EC方式存储，EC最小4+2，最大22+2。
- 整体利用率：站点间利用率*站点内利用率，如：3DC EC，站点间2+1，站点内4+2:1。

1. 利用率: $0.67 * 0.67 = 0.45$ 。

2. 可靠性：支持故障一个站点+另外两个站点各故障1个节点/2块盘。

回收站



- OceanStor Pacific系列非结构化数据服务 提供了基于Namespace/Dtree的回收站, 在用户删除文件时并不真正删除数据, 而是文件在后台移入到回收站目录, 以使用户找回数据。文件从原来的位置移入回收站, 并不会真正删除也不会释放存储空间。为了及时释放空间管理员可以配置文件在回收站内的保留时长, 系统会自动删除回收站内超过保留时长的文件。
- 每个Namespace/Dtree创建一个回收站目录 /.recyclebininternal, 文件在删除时不删除数据, 而是在后台移到回收站中, 在回收站中保持删除前的相对路径。如归属于 root用户的文件在Dtree中的相对路径为/a/b/c.txt, 在回收站中的路径为 /.recyclebininternal/root/current/a/b/c.txt。用户可访问回收站内的目录与文件由回收站操作权限控制, 普通用户权限下用户可访问归属于自己的目录与文件, 回收站作为文件系统内一个特殊目录, root用户也可以添加权限以便其他用户能访问。
- 系统会定期检查回收站内的目录以及数据。文件进入回收站时会首先放入 /.recyclebininternal/用户目录/current目录下, 系统会首先把current目录rename到以当前时间为名称的目录下, 然后再检测当前时间与创建该目录的时间差, 判断是否超过保留时长, 如果超过保留时长, 则删除该目录。

思考题

1. （多选题）在使用Ceph块存储服务时，以下哪些场景会用到RBD设备的快照功能？
 - A. 数据备份
 - B. 应用的破坏性测试
 - C. Ceph版本升级
 - D. Ceph资源池扩容
2. （简答题）请简单描述分布式存储产品部署时，需要划分几个网络平面。

- 答案
- 1、AB
- 2、结构化场景：管理网络，存储网络，业务网络，复制网络和仲裁网络；非结构化场景：管理网络，存储网络，业务网络，复制网络和控制网络。

思考题

3. （判断题）部署OceanStor Pacific存储软件时，FSM和FSA能够部署在同一节点上。
- A. True
- B. False
4. （多选题）华为分布式存储支持提供以下哪些服务？
- A. 块服务
- B. 文件服务
- C. 对象服务
- D. HDFS服务

- 答案
- 3、A
- 4、ABCD

本章总结

- 本章主要讲述分布式存储技术，包括Ceph相关知识，华为分布式存储，分布式存储部署以及产品增值特性，包括效率提升特性和数据保护特性。

更多信息

- 华为数据存储信息中心: <https://info.support.huawei.com/storage/#/home>
- 企业业务: <https://e.huawei.com/cn/>
- 技术支持: <https://support.huawei.com/enterprise/>
- 在线学习: <https://e.huawei.com/cn/talent/portal/#/>
- 华为Support案例库: <https://support.huawei.com/enterprise/zh/knowledge?lang=zh>

Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。
Bring digital to every person, home, and
organization for a fully connected,
intelligent world.

版权所有 © 华为技术有限公司 2026。

保留一切权利。

未经华为事先书面许可，任何人不得对本资料进行复制、修改、
整合、改编、翻译，也不得将本资料或其任何部分或基于本资
料的衍生作品提供、出售、转让、再许可给任何人用于华为
许可以外的任何其他用途。

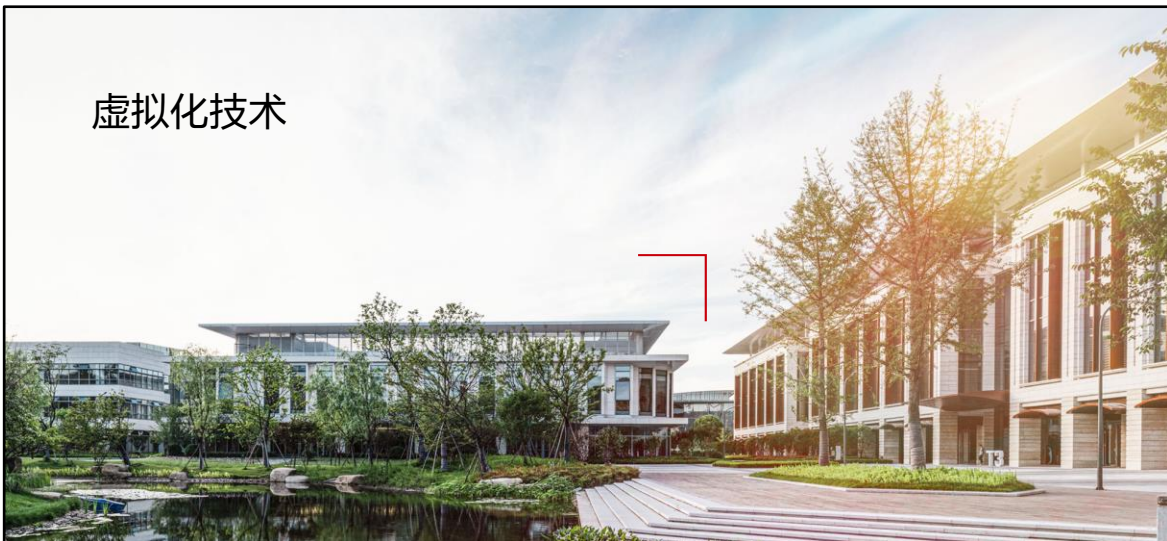
Copyright © Huawei Technologies Co., Ltd. 2026.

All Rights Reserved.

Without prior written permission of Huawei, no one is
allowed to reproduce, modify, integrate, adapt, translate
this Document, or provide, sell, transfer, sublicense this
Document or any part of this Document or its derivative to
any third party for any other purpose otherwise permitted
by Huawei.



虚拟化技术



前言

- 云计算已成为IT技术领域最热门的词汇之一，现在很多互联网服务都声称自己是“云服务”。在云计算发展的背景下，支撑云计算服务的最底层、最基础的虚拟化技术也得到了快速发展，成为技术研究和应用的热点之一。
- 虚拟化技术的本质是什么？主流的虚拟化产品和方案有哪些？本章将一一为您揭开。

目标

- 学完本课程后，您将能够：
 - 掌握虚拟化技术相关概念
 - 掌握虚拟化技术基本原理
 - 了解常见的虚拟化软件
 - 了解数据中心虚拟化应用

目录

1. 虚拟化技术概览

▪ 虚拟化技术简介

- 虚拟化技术基本原理

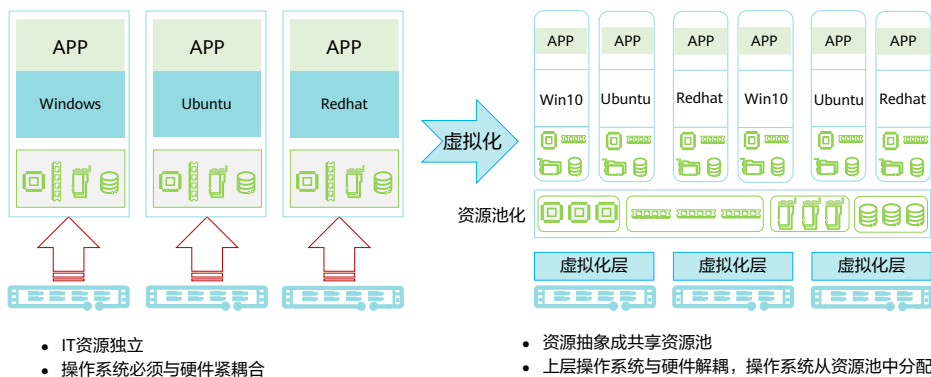
2. 典型虚拟化软件介绍

3. 华为虚拟化软件FusionCompute

4. 数据中心虚拟化DCS

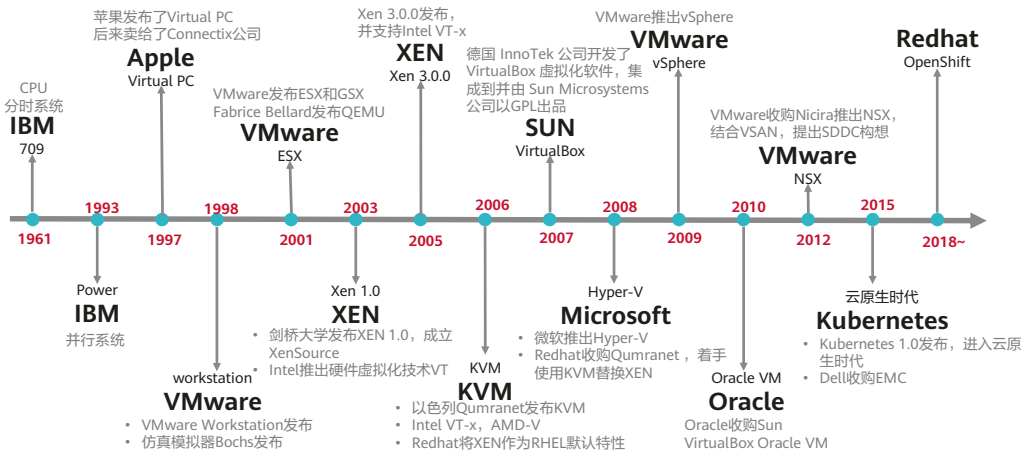
什么是虚拟化

- 虚拟化（Virtualization）的含义很广泛。将任何一种形式的资源抽象成另一种形式的技术都是虚拟化，是资源的一种逻辑表示，解除了物理硬件和操作系统之间的紧耦合关系。



- “虚拟化是以某种用户和应用程序都可以很容易从中获益的方式来表示计算机资源的过程，而不是根据这些资源的实现、地理位置或物理包装的专有方式来表示它们。换句话说，它为数据、计算能力、存储资源以及其他资源提供了一个逻辑视图，而不是物理视图。” —— Jonathan Eunice, Illuminata Inc.
- “虚拟化是表示计算机资源的逻辑组（或子集）的过程，这样就可以用从原始配置中获益的方式访问它们。这种资源的新虚拟视图并不受实现、地理位置或底层资源的物理配置的限制。” —— Wikipedia
- “虚拟化：对一组类似资源提供一个通用的抽象接口集，从而隐藏属性和操作之间的差异，并允许通过一种通用的方式来查看并维护资源。” —— Open Grid Services Architecture Glossary of Terms
- “虚拟化是资源的逻辑表示，它不受物理限制的约束。” —— IBM 公司

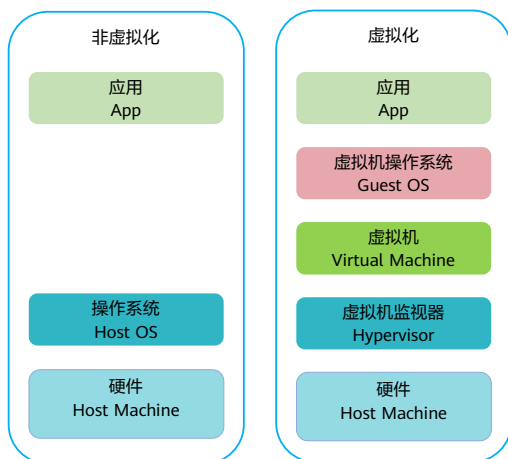
虚拟化发展历程



- 1964年，“蓝色巨人”IBM就开始尝试在大型机上实现虚拟化。
- 1972年，IBM正式将system370机的分时系统命名为虚拟机。
- 1999年，VMware推出了最早的能在x86架构上运行的虚拟化产品。
- 2002年，Xen正式开源，并先后推出了1.0和2.0版本。在这之后，Xen作为虚拟化解决方案，开始被诸如Redhat、Novell和Sun的Linux发行版集成。2004年，Intel的工程师开始为Xen添加硬件虚拟化的支持，从而为即将上市的新款处理器做必需的软件准备。在他们的努力下，2005年发布的Xen 3.0，开始正式支持Intel的VT技术和IA64架构，因此Xen虚拟机就可以运行完全没有修改的操作系统了。
- 2006年10月，在完成基本功能、动态迁移以及主要的功能和性能的优化后，Qumranet正式对外宣布KVM的诞生。同年10月，KVM模块的源代码被正式纳入Linux Kernel，成为内核源代码的一部分。
- 在2006年至2010年期间，各大传统IT厂商在虚拟化方面都推出了自己的产品。2007年，HP推出了Integrity虚拟机，微软在Windows Server 2008 R2中加入了Hyper-V。2010年11月，Redhat公司推出了新的企业版Linux——RHEL 6，该发行版集成了最新的KVM虚拟机，替换了在RHEL 5.x系列中集成的Xen。
- 2011年IBM、红帽、惠普和英特尔成立开放虚拟化联盟，加速KVM推广。
- 2013年，推出了Docker容器项目。
- 2014年，Rocket推出，它最初是由CoreOS开发的，专门用于解决Docker中存在的部分缺陷。CoreOS目标是在满足安全性与生产要求能力上超越Docker。
- 2018年，Redhat收购Kubernetes 与容器原生解决方案创新者与领导厂商 CoreOS。
- 2018年，IBM收购Redhat。
- 2021年，VMware独立。

- 2023年，博通收购VMware。

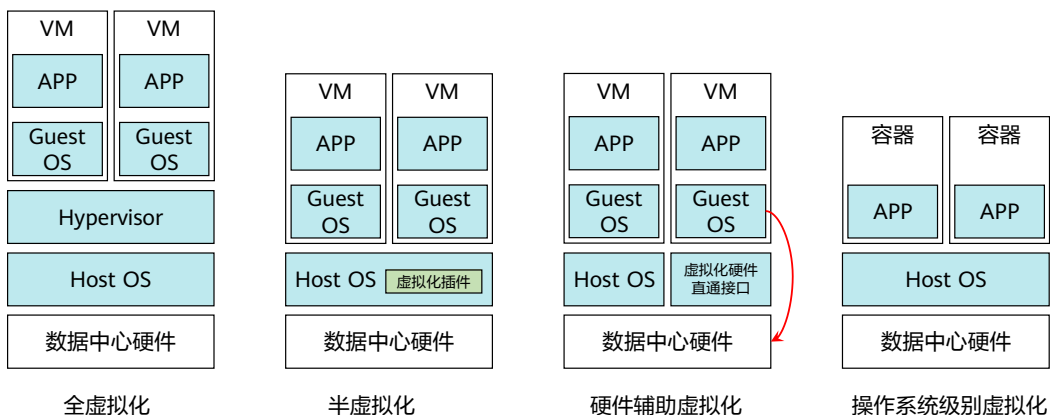
虚拟化中的相关概念



- **Hypervisor:** VMM，也可称为VMM（Virtual Machine Monitor，虚拟机监视器）。计算虚拟化的核心组件，它是运行于物理硬件之上的软件层，负责在物理硬件和虚拟机之间建立一层抽象层，管理其上的虚拟机，使得虚拟机能够像物理机一样运行。Hypervisor对虚拟机进行监控和统计，帮助用户更好地了解虚拟机的运行状态和性能表现。
- **Host OS:** 宿主操作系统，运行在物理机上的操作系统。
- **Virtual Machine:** 虚拟机（Virtual Machine, VM）是一种计算机程序或系统，它通过Hypervisor模拟物理计算机的硬件运行环境，使得多个操作系统能够在同一台物理机器上同时运行。
- **Guest OS:** 客户操作系统，运行在虚拟机内的操作系统。

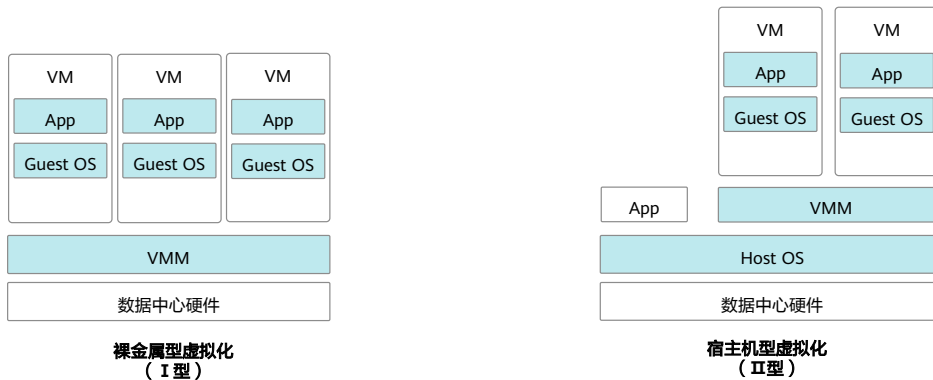
- 如图所示，虚拟化环境由三个部分组成：硬件、VMM和虚拟机。在没有虚拟化的情况下，操作系统直接运行在硬件之上，管理着底层物理硬件，这就构成了一个完整的计算机系统，也就是下文的“物理机”。
- 在虚拟化环境里，虚拟机监视器VMM抢占了操作系统的位置，变成了真实物理硬件的管理者，同时向上层的软件呈现出虚拟的硬件平台，“欺骗”着上层的操作系统。而此时操作系统运行在虚拟平台之上，仍然管理着它认为是“物理硬件”的虚拟硬件，这就是图中的“虚拟机”。
- 虚拟机可以看作是物理机的一种高效隔离的复制，上面的定义里蕴含了三层含义：同质、高效和资源受控，这也是一个虚拟机所具有的三个典型特征。

虚拟化的类型



- 全虚拟化：使用VMM实现CPU、内存、设备I/O的虚拟化，而Guest OS和计算机系统硬件都不需要进行修改。该方式兼容性好，但会给处理器带来额外开销。
- 半虚拟化：使用VMM实现CPU和内存虚拟化，设备I/O虚拟化由Guest OS实现。需要修改 Guest OS，使其能够与VMM协同工作。该方式兼容性差，但性能较好。
- 硬件辅助虚拟化：借助硬件（主要是处理器）的支持来实现高效的全虚拟化。该方式不需要修改Guest OS，兼容性好。该技术将逐渐消除软件虚拟化技术之间的差别，成为未来的发展趋势。
- 操作系统级别虚拟化：是一种服务器虚拟化的方法，其要求操作系统内核允许多个隔离的user-space实例。这样的实例（经常被称为容器、Virtualization engines, virtual private servers）看起来像一个真实的服务器。

主流虚拟化层方案



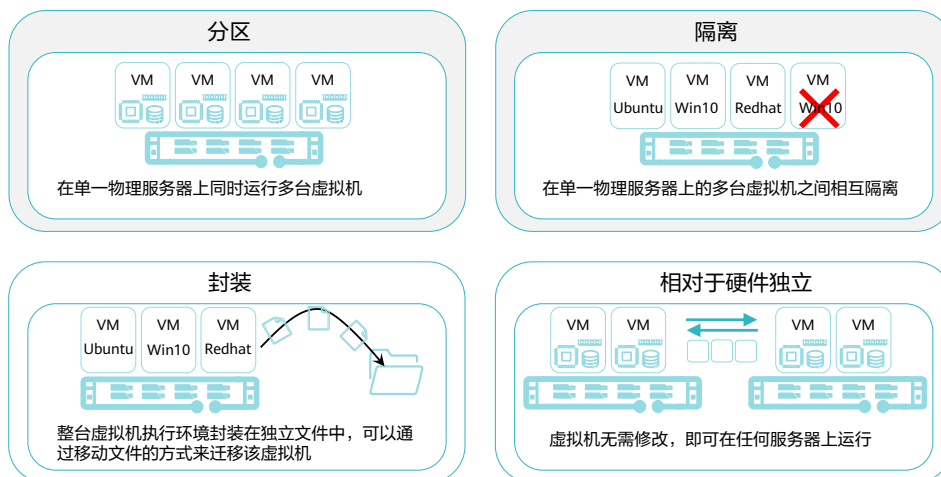
- 裸金属型虚拟化 (I 型)

- 裸金属虚拟化是指在没有操作系统的情况下直接在物理服务器上运行虚拟机实例的技术。VMM直接位于硬件之上，不经过操作系统，或者它本身就是这个操作系统。VMM作为中间层，直接在服务器硬件上运行虚拟机实例，实现真正的硬件资源共享，提供了更高的性能、更低的延迟、更好的资源隔离。
- 裸金属虚拟化通常通过CPU的硬件虚拟化技术，如Intel VT-x、AMD-V、Kunpeng-V来实现，一般需要在BIOS中开启CPU虚拟化能力。
- 典型的虚拟化软件：VMware vSphere、Huawei FusionCompute等。

- 宿主机型虚拟化 (II 型)

- 宿主机虚拟化是在一台物理服务器上运行多个虚拟机实例的技术。它利用虚拟化软件来创建和管理这些虚拟机实例，每个虚拟机实例都可以运行独立的操作系统和应用程序。
- 典型虚拟化软件：VMware workstation、VirtualBox等。

虚拟化的特点



- **分区:** 分区意味着虚拟化层为多台虚拟机划分服务器资源的能力; 每台虚拟机可以同时运行一个单独的操作系统 (相同或不同的操作系统), 使您能够在一台服务器上运行多个应用程序; 每个操作系统只能看到虚拟化层为其提供的“虚拟硬件” (虚拟网卡、CPU、内存等), 以使它认为运行在自己的专用服务器上。
- **隔离:** 虚拟机是互相隔离的。
 - 一台虚拟机的崩溃或故障 (例如, 操作系统故障、应用程序崩溃、驱动程序故障, 等等) 不会影响同一服务器上的其他虚拟机。
 - 一台虚拟机中的病毒、蠕虫等与其它虚拟机相隔离, 就像每台虚拟机都位于单独的物理机器上一样。
 - 可以进行资源控制以提供性能隔离: 您可以为每台虚拟机指定最小和最大资源使用量, 以确保某台虚拟机不会占用所有的资源而使得同一系统中的其他虚拟机无资源可用。
 - 可以在单一机器上同时运行多个负载/应用程序/操作系统, 而不会出现我们刚才讨论传统x86服务器体系结构的局限性时所提到的那些问题 (应用程序冲突、DLL冲突等)。
- **封装:** 封装意味着将整套虚拟机 (硬件配置、BIOS配置、内存状态、磁盘状态、CPU状态) 储存在独立于物理硬件的一小组文件中。这样, 您只需复制几个文件就可以随时随地根据需要复制、保存和移动虚拟机。
- **相对于硬件独立:** 因为虚拟机运行于虚拟化层之上, 所以只能看到虚拟化层提供的虚拟硬件; 此虚拟硬件也同样不必考虑物理服务器的情况; 这样, 虚拟机就可以在任何x86服务器 (IBM、Dell、HP等) 上运行而无需进行任何修改。这打破了操作系统和硬件以及应用程序和操作系统/硬件之间的约束。

目录

1. 虚拟化技术概览

- 虚拟化技术简介

- **虚拟化技术基本原理**

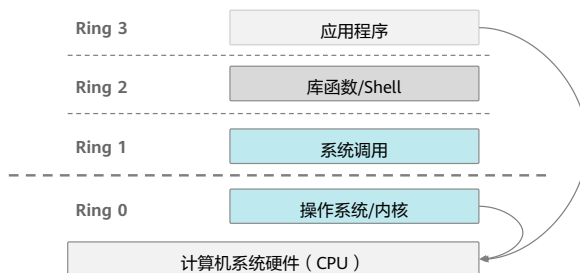
2. 典型虚拟化软件介绍

3. 华为虚拟化软件FusionCompute

4. 数据中心虚拟化DCS

内核态和用户态

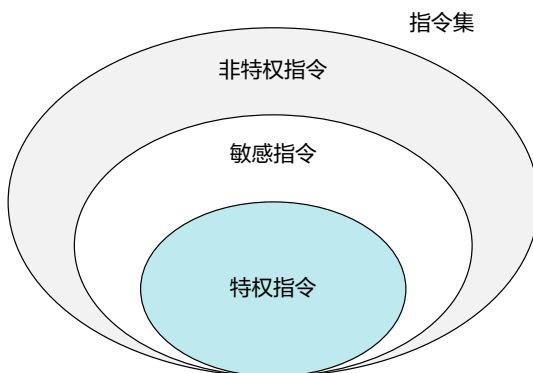
- 内核态（Kernel mode）：如果 CPU 处于内核态，执行的程序可以执行任何 CPU 指令，并且访问内存中的所有地址，包括外围设备，比如硬盘/网卡等等；
- 用户态（User mode）：如果处于用户态，只能访问受限的资源，而且不能引用内存或者直接访问外围设备。



- 在CPU的设计中，用户态指非特权状态。在此状态下，执行的代码被硬件限定，不能进行某些操作，比如写入其他进程的存储空间。
- 在操作系统的设计中，用户态也类似，指非特权的执行状态。内核禁止此状态下的代码进行潜在危险的操作，比如写入系统配置文件、杀掉其他用户的进程、重启系统等。
- x86架构提供四个特权级别给操作系统和应用程序来访问硬件。Ring 是指 CPU 的运行级别，Ring 0是最高级别，Ring1次之，Ring2更次之。就 Linux+x86 来说：
 - 操作系统（内核）需要直接访问硬件和内存，因此它的代码需要运行在最高运行级别 Ring0上，这样它可以使用特权指令，控制中断、修改页表、访问设备等等。
 - 应用程序的代码运行在最低运行级别上Ring3上，不能做受控操作。如果要做，比如要访问磁盘，写文件，那就要通过执行系统调用（函数），执行系统调用时，CPU的运行级别会发生从Ring3到Ring0的切换，并跳转到系统调用对应的内核代码位置执行，这样内核就为你完成了设备访问，完成之后再从Ring0返回Ring3。这个过程也称作用户态和内核态的切换。

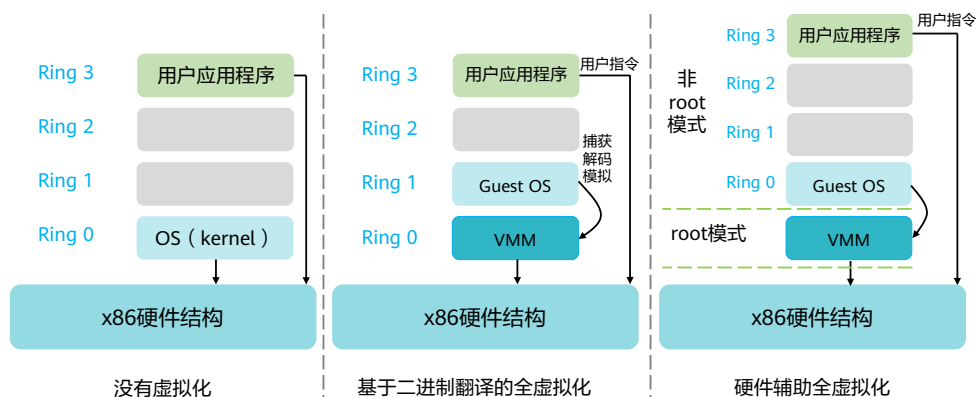
指令

- 在 CPU 虚拟化中，指令的区分对于理解虚拟化机制至关重要。根据指令对系统资源和权限的要求，我们可以将指令分为特权指令、非特权指令和敏感指令三类。



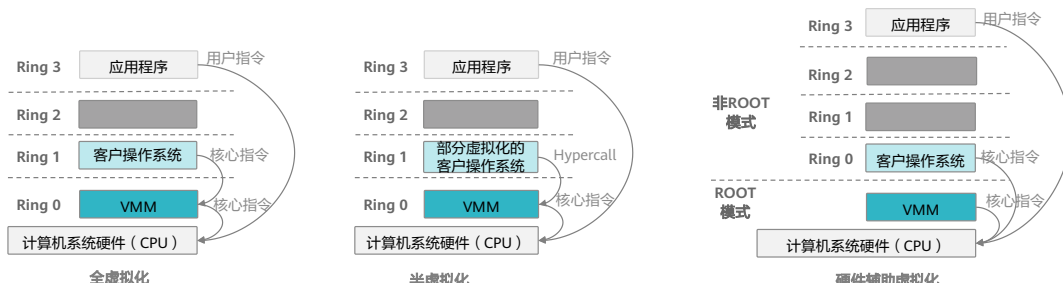
- 特权指令：Ring 0才能运行，具有最高权限级别。这类指令可以直接访问和控制系统硬件和核心资源，如：进行硬件I/O通信、内存管理、中断管理等。
- 敏感指令：Ring 1上运行，部分指令在Ring 0上运行。这类指令可以直接访问和控制系统硬件和核心资源，但其执行权限级别较低，通常由需要执行敏感操作的应用程序或服务使用，如：访问用户隐私数据、修改系统设置等。
- 非特权指令：在Ring 2、Ring 3级别上运行，部分指令在Ring 1上运行，在CPU普通权限级别上运行。这类指令主要用于数据处理、逻辑运算、内存访问等常规操作。

指令模拟



- 指令模拟的核心思想是，用户在虚拟机执行指令时，如果该指令需要模拟（如特权指令），则由虚拟化软件（如 Hypervisor）捕获该指令的执行，并将其转化为可在物理 CPU 上执行的指令序列。这一过程涉及到指令的捕获、解码、模拟执行以及结果的返回。
- 虚拟化在这里就遇到了一个难题，因为宿主操作系统是工作在 Ring 0 的，客户操作系统就不能也在 Ring 0 了，但是它不知道这一点，以前执行什么指令，现在还是执行什么指令，但是没有执行权限是会出错的。所以这时候虚拟机管理程序（VMM）需要避免这件事情发生。VM 通过 VMM 实现 Guest CPU 对硬件的访问，根据其原理不同有三种实现技术：
 - 全虚拟化
 - 半虚拟化
 - 硬件辅助的虚拟化

CPU虚拟化技术



维度	全虚拟化	半虚拟化	硬件辅助虚拟化
实现技术	利用二进制翻译和直接执行。	Hypervisor实现CPU和内存虚拟化，设备IO虚拟化由管理OS实现。	Intel-VT技术和AMD-V技术。CPU 采用VT-x技术，内存采用EPT技术，设备IO采用VT-d/VT-c技术，遇到特权指令转到root模式执行。
客户操作系统兼容性	无需修改客户操作系统，最佳兼容性。	客户操作系统需要修改来支持hypercall，因此它不能运行在物理硬件本身或其他的hypervisor上，兼容性差，不支持Windows	无需修改客户操作系统，最佳兼容性。
性能	差	好。半虚拟化下CPU性能开销几乎为0，虚机的性能接近于物理机。	全虚拟化下，CPU需要在两种模式之间切换，带来性能开销；但是，其性能在逐渐逼近半虚拟化。
典型厂商	VMware workstation	XEN	VMware vSphere、Microsoft Hyper-V、KVM

• 全虚拟化：

- 客户操作系统运行在Ring 1，它在执行特权指令时，会触发异常（CPU的机制，没权限的指令会触发异常），然后VMM捕获这个异常，在异常里面做翻译、模拟，最后返回到客户操作系统内，客户操作系统认为自己的特权指令工作正常，继续运行。但是这个性能损耗非常的大，简单的一条指令执行要通过复杂的异常处理过程。

• 半虚拟化：

- 半虚拟化的思想就是，修改操作系统内核，替换掉不能虚拟化的指令，通过超级调用（Hypercall）直接和底层的虚拟化层Hypervisor来通讯，Hypervisor同时也提供了超级调用接口来满足其他关键内核操作，比如内存管理、中断和时间保持。这种做法省去了全虚拟化中的捕获和模拟，大大提高了效率。所以像XEN这种半虚拟化技术，客户机操作系统都是有一个专门的定制内核版本，和X86、mips、arm这些内核版本等价。这样一来，就不会有捕获异常、翻译、模拟的过程了，性能损耗非常低。这就是XEN这种半虚拟化架构的优势。

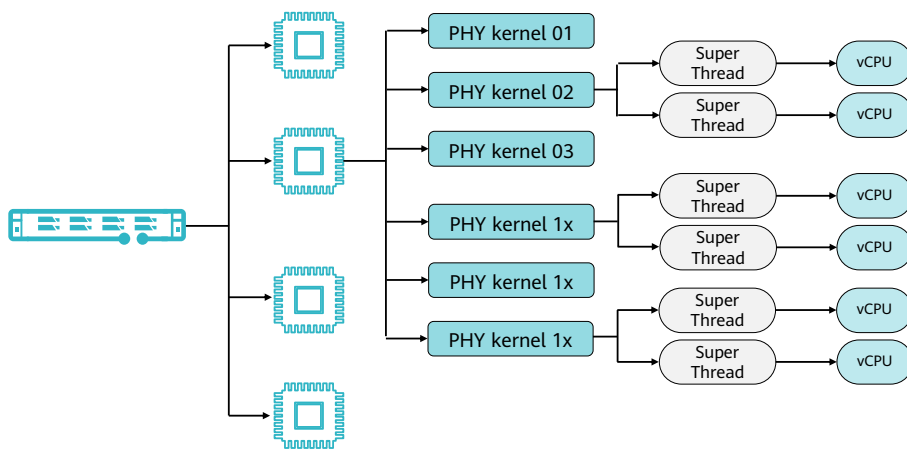
• 硬件辅助虚拟化：

- 在 X86架构中，Intel和AMD两大处理器厂商都提供了专门的硬件辅助虚拟化技术，以支持高效的指令模拟。Intel的VT-x（Virtualization Technology for X86）技术通过提供一组专门的指令和处理器模式，使得Hypervisor能够更加高效地捕获和控制虚拟机的指令执行。
- ARM架构的硬件辅助虚拟化技术中，ARMv8架构特别引入了专门的虚拟化扩

展VHE（Virtualization Host Extension），此举显著增强了ARM处理器对虚拟机指令模拟的高效支持。VHE作为ARM架构专为提升虚拟化性能而设计的硬件特性，其核心思想在于精简虚拟化流程中的冗余上下文切换与陷阱（trap）操作，进而优化虚拟机的运行效率。

- 2005年后，CPU厂商Intel和AMD开始支持虚拟化了。Intel引入了Intel-VT（Virtualization Technology）技术。这种CPU，有VMX root operation和VMX non-root operation两种模式，两种模式都支持Ring 0 ~ Ring 3共4个运行级别。这样一来，VMM可以运行在VMX root operation模式下，客户OS运行在VMX non-root operation模式下。
- 以Intel-VT技术为例，它增加了两种运行模式：VMX root模式和VMX non-root模式。通常来讲，主机操作系统和VMM运行在VMX root模式中，客户机操作系统及其应用运行在VMX non-root模式中。因为两个模式都支持所有的Ring，因此，客户机可以运行在它所需要的Ring中（OS运行在Ring 0中，应用运行在Ring 3中），VMM也运行在其需要的Ring中（对KVM来说，QEMU运行在Ring 3，KVM运行在Ring 0）。CPU在两种模式之间的切换称为VMX切换。从root mode进入non-root mode，称为VM-Entry；从non-root mode进入root mode，称为VM-Exit。可见，CPU受控制地在两种模式之间切换，轮流执行VMM代码和Guest OS代码。
- 对KVM虚拟机来说，运行在VMX Root Mode下的VMM在需要执行Guest OS指令时执行VMLAUNCH指令将CPU转换到VMX non-root mode，开始执行客户机代码，即VM-Entry过程；在Guest OS需要退出该mode时，CPU自动切换到VMX Root mode，即VM-Exit过程。可见，KVM用户机代码是受VMM控制直接运行在物理CPU上的。QEMU只是通过KVM控制虚拟机的代码被CPU执行，但是它们本身并不执行其代码。也就是说，CPU并没有真正地被虚拟化成虚拟CPU给客户机使用。

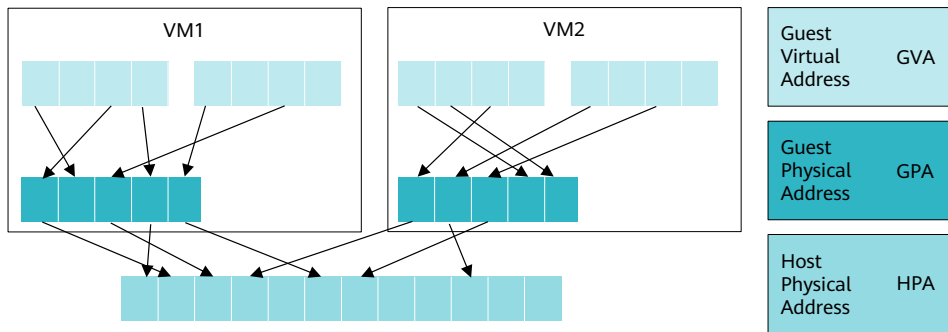
CPU和vCPU的对应关系



- vCPU数量和物理CPU对应关系如图所示。
- 以x86服务器使用2.6 GHz主频CPU为例，单台服务器有2颗物理CPU，每颗CPU有8核，又因为超线程技术可以提供每个物理内核两个处理线程，因此每颗CPU有16线程，总vCPU数量为 $2 \times 8 \times 2 = 32$ 个vCPU。总资源为 $32 \times 2.6 \text{ GHz} = 83.2 \text{ GHz}$ 。
- 虚拟机vCPU数量不能超过单台服务器节点可用vCPU数量。多台虚拟机间可以复用同一颗物理CPU，因此单服务器节点上运行的虚拟机vCPU数量总和可以超过实际vCPU数量。

内存虚拟化

- 把物理机的真实物理内存统一管理，包装成多份虚拟的内存动态分配给若干虚拟机使用。
- KVM通过内存虚拟化共享物理系统内存，动态分配给虚拟机。



- KVM中，虚拟机的物理内存即为qemu-kvm进程所占用的内存空间。KVM使用CPU辅助的内存虚拟化方式。
- 内存虚拟化 - 影子页表：
 - 由于宿主机MMU不能直接装载客户机的页表来进行内存访问，所以当客户机访问宿主机物理内存时，需要经过多次地址转换。即首先根据客户机页表把客户机虚拟地址（GVA）转换成客户机物理地址（GPA），然后再通过客户机物理地址（GPA）到宿主机虚拟地址（HVA）之间的映射转换成宿主机虚拟地址，最后再根据宿主机页表把宿主机虚拟地址（HVA）转换成宿主机物理地址（HPA）。而通过影子页表，则可以实现客户机虚拟地址到宿主机物理地址的直接转换。
 - Intel的CPU提供了EPT（Extended Page Tables，扩展页表）技术，直接在硬件上支持GVA->GPA->HPA的地址转换，从而降低内存虚拟化实现的复杂度，也进一步提升内存虚拟化性能。
- KVM为了在一台机器上运行多台虚拟机，需要增加一个新的内存虚拟化层（客户机物理地址空间），这个地址空间不是真正意义上的物理地址空间，它们之间还有一层转换。客户机虚拟地址（GVA）到客户机物理地址（GPA）的转换。
- 但是客户操作系统不能直接访问实际机器内存，因此VMM需要负责映射客户物理内存到实际机器内存（GPA->HPA）。

内存虚拟化问题

- Native操作系统对内存的认识与管理达成以下两点认识：
 - 内存都是从物理地址0开始的；
 - 内存都是连续的。
- 内存虚拟化需要解决两个问题：
 - 从物理地址0开始的：物理地址0只有一个，无法同时满足所有客户机从0开始的要求。
 - 地址连续：虽然可以分配连续的物理地址，但是内存使用效率不高，缺乏灵活性。

I/O虚拟化

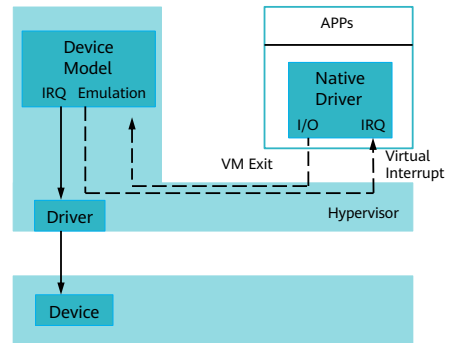
- I/O虚拟化可以被看作是位于服务器组件的系统和各种可用I/O处理单元之间的硬件中间件层，使得多个guest可以复用有限的外设资源。
- 设备虚拟化（I/O虚拟化）的过程，就是模拟设备的这些寄存器和内存，截获Guest OS对IO端口和寄存器的访问，通过软件的方式来模拟设备行为。
- 在QEMU/KVM中，客户机可以使用的设备大致可分为三类：
 - 模拟设备：完全由QEMU纯软件模拟的设备。
 - Virtio设备：实现VIRTIO API的半虚拟化设备。
 - PCI设备直接分配（PCI device assignment）。

I/O虚拟化问题

- I/O虚拟化需要解决两个问题：
 - 设备发现：
 - 需要控制各虚拟机能够访问的设备。
 - 访问截获：
 - 通过I/O端口或者MMIO对设备的访问；
 - 设备通过DMA与内存进行数据交换。

I/O虚拟化 - 全模拟

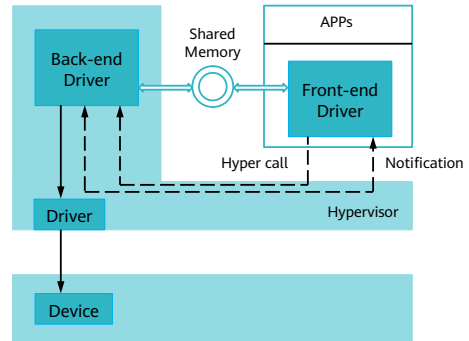
- 用软件完全模拟一个特定的设备
 - 保持一样的软件接口，如：PIO、MMIO、DMA、中断等。
 - 可以模拟出跟系统中的物理设备不一样的虚拟设备。
- 每次I/O操作需要多次上下文切换
 - VM <-> Hypervisor
 - QEMU <-> Hypervisor
- 软件模拟的设备不影响虚拟机中的软件栈
 - 原生驱动



- 模拟I/O设备方式的优点是对硬件平台依赖性较低、可以方便模拟一些流行的和较老旧的设备、不需要宿主机和客户机的额外支持，因此兼容性高；而其缺点是I/O路径较长、VM-Exit次数很多，因此性能较差。一般适用于对I/O性能要求不高的场景，或者模拟一些老旧遗留（legacy）设备（如RTL8139的网卡）。

I/O虚拟化 - Virtio

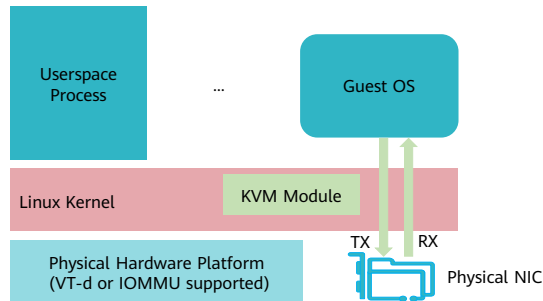
- 虚拟出特殊的设备
 - 特殊的设备驱动，包括VM中的Front-end驱动和主机上的Back-end驱动；
 - Front-end和Back-end驱动之间的高效通信。
- 减少VM和主机的数据传输开销
 - 共享内存；
 - Batched I/O；
 - 异步事件通知Eventfd轻量级进程间“等待/通知”机制。



- Virtio半虚拟化设备方式的优点是实现了VIRTIO API，减少了VM-Exit次数，提高了客户机I/O执行效率，比普通模拟I/O的效率要高很多；而其缺点是需要客户机中与Virtio相关驱动的支持（较老的系统默认没有自带这些驱动，Windows系统中需要额外安装Virtio驱动），因此兼容性较差，而且I/O频繁时的CPU使用率较高。

PCI设备直接分配

- KVM虚拟机支持将宿主机中的PCI、PCI-E设备附加到虚拟化的客户机中，从而让客户机以独占方式访问这个PCI或PCI-E设备。通过硬件支持的VT-d技术将设备分配给客户机后，在客户机看来，设备是物理上连接在其PCI或PCI-E总线上的，客户机对该设备的I/O交互操作和实际的物理设备操作完全一样，不需要或很少需要Hypervisor的参与。



- 设备直接分配让客户机完全占有PCI设备，这样在执行I/O操作时大量地减少甚至避免了VM-Exit陷入到Hypervisor中，极大地提高了I/O性能，可以达到几乎和Native系统中一样的性能。尽管virtio的性能也不错，但VT-d克服了其兼容性不够好和CPU使用率较高的问题。不过，VT-d也有自己的缺点，一台服务器主板上的空间比较有限，允许添加的PCI和PCI-E设备是有限的，如果一台宿主机上有较多数量的客户机，则很难向每台客户机都独立分配VT-d的设备。另外，大量使用VT-d独立分配设备给客户机，让硬件设备数量增加，这会增加硬件投资成本。

存储虚拟化介绍

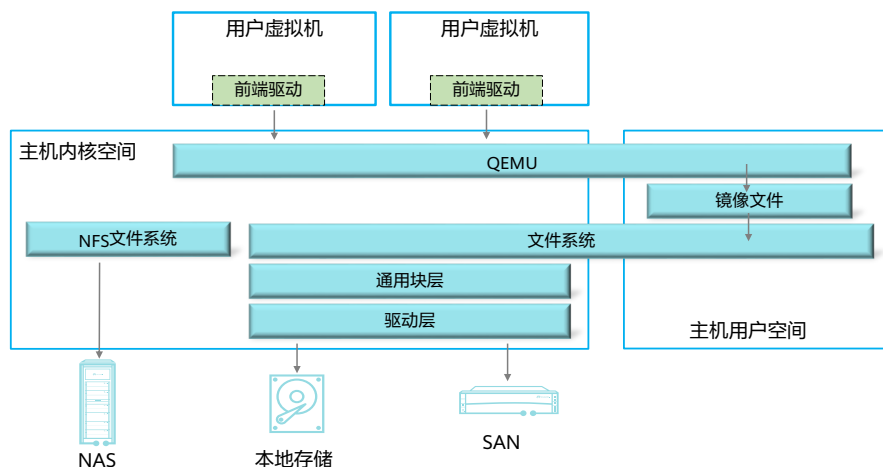
- 存储虚拟化是将存储设备抽象为数据存储，虚拟机在数据存储中作为一组文件存储在自己的目录中。数据存储是逻辑容器，类似于文件系统，它将各个存储设备的特性隐藏起来，并提供一个统一的模型来存储虚拟机文件。存储虚拟化技术可以更好地管理虚拟基础架构的存储资源，使系统大幅提升存储资源利用率和灵活性，提高应用的正常运行时间。
- 本节以FusionCompute为例介绍存储虚拟化相关内容。

FusionCompute中的存储概念

- 存储资源
 - 主机本地磁盘
 - 专用存储物理设备，例如SAN存储、NAS存储、OceanStor Pacific、OceanStor系列和OceanStor Dorado系列等。
- 存储设备
 - 存储设备表示存储资源中的管理单元，类似LUN、华为分布式块存储上的存储池、NAS共享目录等。
- 数据存储
 - FusionCompute对存储资源上的存储单元进行的统一封装。
- 能够封装为数据存储的存储单元包括：
 - SAN存储（包括iSCSI或光纤通道的SAN存储）上划分的LUN
 - NAS存储上划分的文件系统
 - 华为分布式块存储上的存储池
 - 主机的本地硬盘（虚拟化）
 - eVol存储上的存储池

- FusionCompute可使用的存储资源来自主机本地磁盘或专用的存储设备。专用的存储设备与主机之间应通过网线或光纤连通。
- 存储资源封装成数据存储并与主机关联后，就能够进一步创建出若干虚拟磁盘，供虚拟机使用。
- 存储池是存放存储空间资源的容器，所有应用服务器使用的存储空间都来自于存储池。
- 从虚拟机操作系统使用的角度观察，在不同存储资源上创建的虚拟磁盘之间不存在差异，使用方式均与物理PC的磁盘相同。
- FusionCompute存储虚拟化技术定位于构建针对电信业务环境的高竞争力的虚拟化平台，通过对开源KVM进行安全加固、功能扩展、性能优化和可靠性保障，具备如下特点：
 - 存储设备兼容性：对于不同的存储设备，包括IP SAN、FC SAN、NAS、本地磁盘。可以做到以文件系统进行屏蔽，统一提供文件级别的业务操作。
 - 丰富的功能：提供了包括精简磁盘、增量快照、存储冷热迁移、链接克隆虚拟机、虚拟机磁盘扩容等众多功能。
 - 业务能力同质化：业务是在虚拟化层进行，基于各种存储设备可以提供相同的能力，对存储设备无特殊要求。

存储虚拟化架构



- FusionCompute存储虚拟化平台主要由文件系统、磁盘驱动和磁盘工具组成。SAN设备和本地磁盘等块设备连接到服务器上后，经过设备驱动层和通用块层后，对主机呈现一个磁盘设备。
- 文件系统建立在主机能够访问的存储设备上，创建文件系统的过程就是主机对存储设备执行格式化，在存储设备上写入文件系统的元数据和iNode信息，建立文件到裸设备块的映射，并负责裸设备块的管理，包括空间分配和释放。对于虚拟化来说，文件系统屏蔽了块层的操作，提供了文件级的抽象操作。虚拟机磁盘就是放在文件系统文件中的文件。
- 虚拟机磁盘在使用时，需要由磁盘驱动挂载给虚拟机，并通过QEMU管理起来，用户虚拟机的所有读写IO都会由前端驱动截获，转发给QEMU进程，并转换为用户态驱动的读写操作，最终写入磁盘文件中。
- 虚拟机磁盘包含了属性信息和数据区域，磁盘工具提供了一系列对虚拟机磁盘的操作，可以解析磁盘文件的文件头，读取或修改磁盘的属性信息，创建磁盘的数据区等操作。

数据存储使用方式

01 虚拟化存储

- 虚拟化的数据存储创建普通磁盘速度较慢，但可支持部分高级特性，如果创建精简磁盘，还能支持更多的高级特性，可提高存储的资源利用率和系统的安全性、可靠性。
- 格式化：存储设备首次添加为数据存储时，请确认该存储设备上的数据已经备份或不再使用，并将“是否格式化”设置为“是”，此时系统会将待使用的存储设备格式化为华为虚拟化文件系统。

02 非虚拟化存储

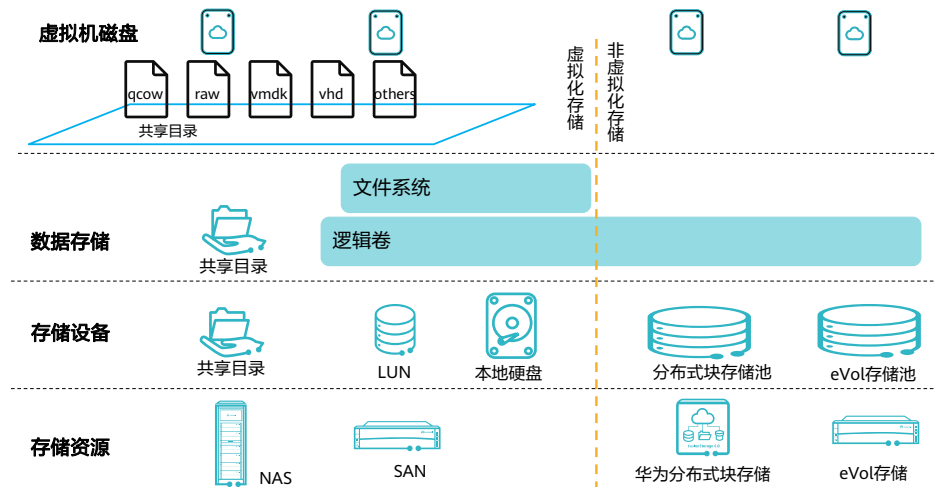
- 当存储资源为华为分布式块存储或者eVol存储时，创建数据存储时默认为非虚拟化。

03 裸设备映射

- 将SAN存储的物理LUN直接作为磁盘绑定给业务虚拟机，使SAN存储具有更高的性能。

- 裸设备映射的数据存储只能整块当做裸设备映射的磁盘使用，不可分割，因此只能创建与数据存储同等容量的磁盘，且不支持虚拟化存储的高级功能。

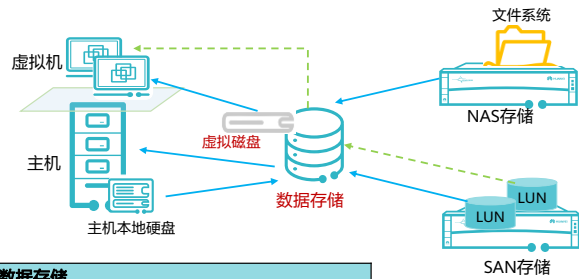
存储模型



- 本地硬盘添加数据存储时可以选择虚拟化的使用方式。
- SAN添加数据存储时可以选择虚拟化和裸设备映射的使用方式。
- 华为分布式块存储添加数据存储时默认非虚拟化的使用方式。

数据存储关联模型

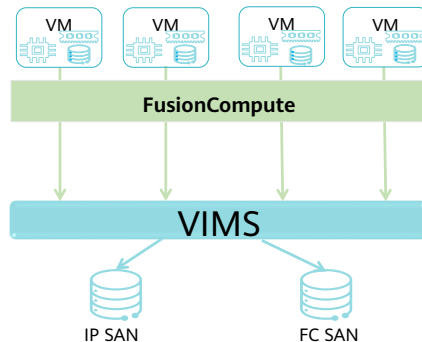
- 数据存储一般关联关系
 - - - - - 裸设备映射



存储设备	存储资源	数据存储
本地硬盘	-	虚拟化本地硬盘（虚拟化）
SAN存储（IP SAN、FC SAN）	SAN存储	虚拟化SAN存储（虚拟化） 裸设备共享存储（裸设备映射）
NAS存储	NAS存储	NAS存储（虚拟化）
华为分布式块存储	华为分布式块存储	华为分布式块存储（非虚拟化）
eVol存储	eVol存储	eVol存储（非虚拟化）

VIMS虚拟集群存储文件系统

- VIMS（虚拟镜像管理系统）是一种高性能的集群文件系统，使虚拟化技术的应用超出了单个存储系统的限制，可让多台虚拟机共同访问一个整合的集群式存储池，从而显著提高了资源利用率。VIMS是跨越多个存储服务器实现虚拟化的基础，它可启用存储热迁移、DRS（Storage Dynamic Resource Scheduler）和High Availability等各种服务。



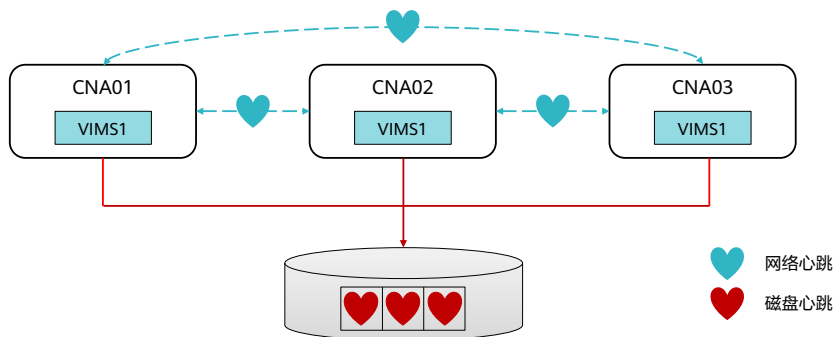
- 虚拟镜像管理系统VIMS 是一种高性能的集群文件系统，使用时先将数据存储格式化成VIMS格式，统一提供文件级别的业务操作。
- 数据存储可支持如下文件系统格式：
 - 虚拟镜像管理系统（VIMS）：为存储虚拟机而优化的高性能文件系统。主机可以将虚拟镜像管理系统数据存储部署在任何基于SCSI的本地或联网存储设备上，包括光纤通道、以太网光纤通道和iSCSI SAN设备。
 - 网络文件系统（NFS）：NAS设备上的文件系统。FusionCompute支持NFS V3协议，可以访问位于NFS服务器上指定的NFS磁盘，挂载该磁盘并满足任何存储需求。
 - EXT4：FusionCompute支持服务器的本地磁盘虚拟化。
- VIMS的关键能力：
 - 为存储虚拟机而优化的高性能文件系统；
 - 屏蔽存储设备差异，异构/利旧存储设备；
 - 物理存储设备池化管理调度，显著提高资源利用率。

VIMS分布式锁

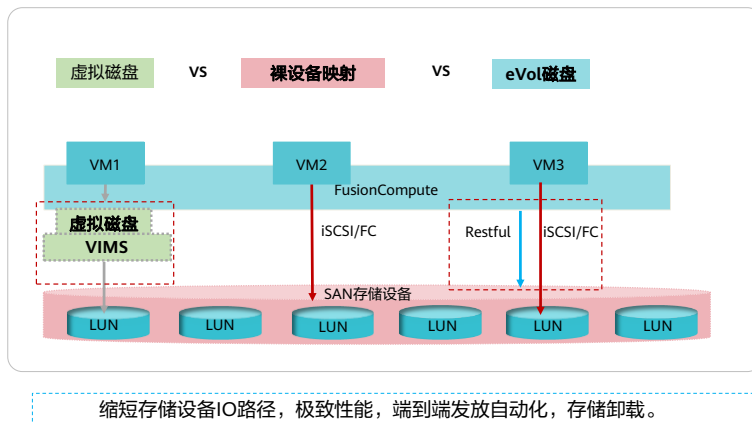
- 一个VIMS卷同时被多个CNA节点挂载，因此每个CNA节点都可以访问VIMS卷上的文件。为保证多节点读写同一文件的数据一致性，VIMS需要实现分布式文件锁。VIMS的DLM（Distributed lock manager）模块负责实现分布式文件锁，它提供集群概念上的锁服务，调用者通过DLM保证集群间的同步要求。
- VIMS采用分布式全对称锁机制。在VIMS中有多个资源管理者（master），每个master只对应一个锁资源。不同的master并不会集中在同一个节点上，无管理中心节点。
- 正常情况下成为某个锁资源的master方式有两种：
 - 第一个申请访问某资源的节点；
 - 如果多节点同时访问某资源，以VIMS节点号较小的节点作为master。
- 当节点发生故障时，此节点负责管理的资源会重新选举出master。

VIMS心跳

- VIMS存在两种心跳，磁盘心跳用于检测主机是否可以正常读写共享存储，网络心跳用于检测主机间网络通信是否正常。作为集群文件系统，挂载了VIMS卷的CNA节点从来都不是单独的个体，作为集群节点成员之一，通过网络心跳确保与其他节点进行正常的网络通信。



eVol存储



适用场景：
集群文件系统，视频图片处理等存储高IO、低延迟业务，业务发放自动化

- eVol存储（原Advanced San存储）把存储阵列的一个存储池作为FusionCompute的数据存储，在FusionCompute创建磁盘时会自动在存储阵列创建对应的LUN，将存储阵列的一个LUN直接虚拟化为虚拟机的一个虚拟磁盘。
- eVol存储不通过VIMS进行存储虚拟化，不需要规划LUN资源。
- eVol存储将各类存储管理操作卸载到存储阵列：如虚拟机磁盘的发放、回收、扩容、快照、克隆、迁移。
- eVol支持的存储型号为OceanStor Dorado系列和OceanStor 2910。

虚拟磁盘参数

01 磁盘类型

- 普通：普通磁盘只能单台虚拟机使用。
- 共享：共享磁盘可以绑定给多台虚拟机使用。

02 配置模式

- 普通：根据磁盘容量为磁盘分配空间，在创建过程中会将物理设备上保留的数据置零。
- 精简：该模式下，系统首次仅分配磁盘容量配置值的部分容量，后续根据使用情况，逐步进行分配，直到分配总量达到磁盘容量配置值为止。
- 普通延迟置零：根据磁盘容量为磁盘分配空间，创建时占用极少空间做标记，初次使用时分配所有空间。

03 磁盘模式

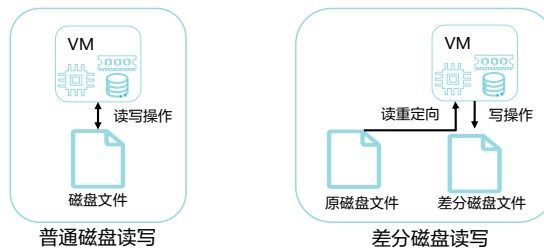
- 从属：快照中包含该从属磁盘，更改将立即并永久写入磁盘。
- 独立-持久：更改将立即并永久写入磁盘，持久磁盘不受快照影响。
- 独立-非持久：关闭电源或恢复快照后，丢弃对该磁盘的更改。

• 配置模式：

- 普通：根据磁盘容量为磁盘分配空间，在创建过程中会将物理设备上保留的数据置零。这种格式的磁盘性能要优于其他两种磁盘格式，但创建这种格式的磁盘所需的时间可能会比创建其他类型的磁盘长。建议系统盘使用该模式。适用于对IOPS要求较高的场景。
- 普通延迟置零：根据磁盘容量为磁盘分配空间，创建时不会擦除物理设备上保留的任何数据，但后续从虚拟机首次执行写操作时会按需要将其置零。创建速度比“普通”模式快；IO性能介于“普通”和“精简”两种模式之间。只有数据存储类型为“虚拟化本地硬盘”或“虚拟化SAN存储”时，支持该模式。适用于对发放速度要求高，但对IOPS要求不高的场景。
- 精简：该模式磁盘在创建时不进行空间分配，而是在用户IO写入磁盘文件时才进行空间动态分配，性能较普通磁盘有所下降。适用于用户对存储需求不明确，或是规划的容量比实际使用的容量多的场景。使用精简模式可能导致数据存储超分配，建议超分配比例不超过50%。超分配率可通过数据存储的详细信息页签“已分配容量”和“总容量”的比率关系来确定。

差分磁盘

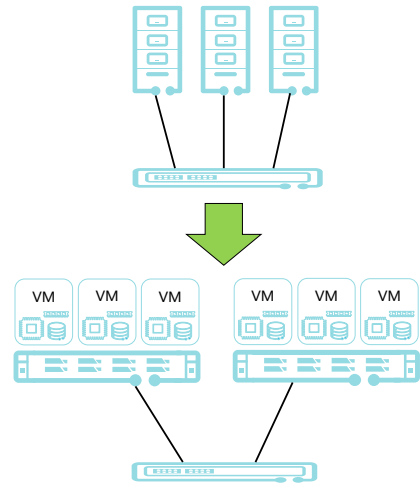
- 差分磁盘必须基于一个已有的父磁盘来创建，它只记录相对于父磁盘的差异数据，包括数据的增减，差分磁盘不能脱离父磁盘而存在，如果父磁盘进行了修改，则差分磁盘的数据将不再可用。
- 该磁盘用于FusionCompute系统中的快照、非持久化磁盘、链接克隆等功能，起到保护源盘不再被修改，并可以跟踪虚拟机磁盘差异数据的作用。



- 对差分磁盘进行读操作，会首先读取差分磁盘中的数据区域，如果不存在，则表示该数据区域未被修改，则通过文件头的索引找到父磁盘，再对该区域进行读取。对差分磁盘进行写操作时，则直接写在差分磁盘上。
- 差分磁盘的结构和动态磁盘完全一致，文件头中会记录父磁盘的路径，数据区只记录和父磁盘差异数据，大小也会随着数据的写入而自动增长。
- 差分磁盘不支持扩容。

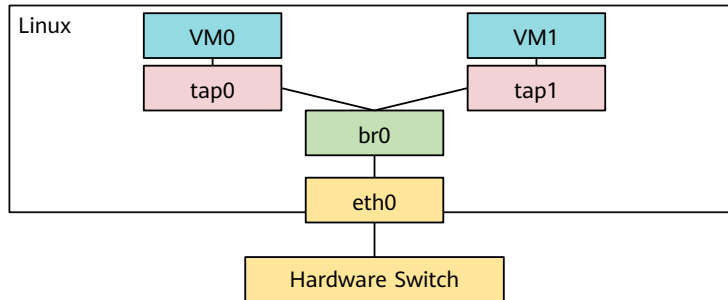
计算虚拟化驱动网络虚拟化的发展

- 计算虚拟化驱动网络虚拟化的发展。传统数据中心的，一台服务器运行一个操作系统，通过物理网线与交换机相连，由交换机实现不同主机的交换、流量控制、安全控制等功能。在计算虚拟化后，一台服务器虚拟化成多台虚拟的主机，每个虚拟主机有自己的虚拟CPU、内存和网卡。同一服务器上的不同主机之间既需要维持原有的通信，同时由于共享物理设备，引出了新的安全隔离、以及对流控更高的需求，对虚拟交换技术的诉求由此产生。
- 为统一和简化对各台主机的虚拟交换机的配置管理，业界引入分布式虚拟交换机。分布式虚拟交换机一方面可以对多台服务器的虚拟交换机统一配置、管理和监控，另一方面也可以保证虚拟机在服务器之间迁移时网络配置的一致性。



Linux Bridge简介

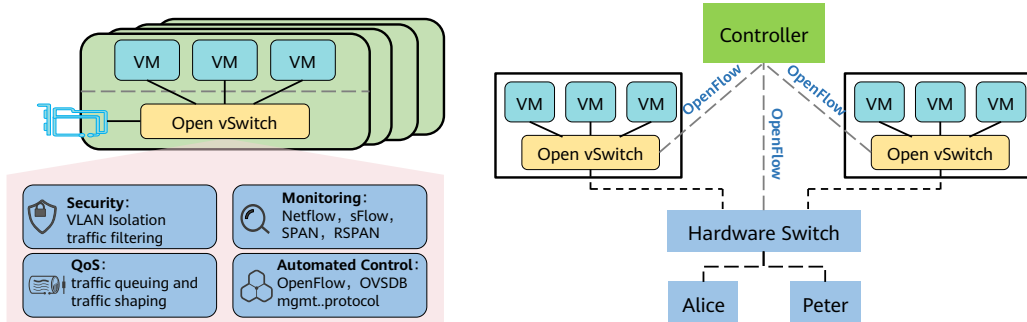
- Linux Bridge（网桥）是工作于二层的虚拟网络设备，功能类似于物理的交换机。
- Bridge可以绑定其他Linux网络设备作为从设备，并将这些设备虚拟化为端口，当一个从设备被绑定到Bridge上时，就相当于真实网络中的交换机端口插入了一个连接有终端的网线。



- Bridge设备br0绑定了实际设备eth0与虚拟设备tap0/tap1，此时，对于Hypervisor的网络协议栈上层来说，只看得到br0，并不会关心网桥的细节。当这些从设备接收到数据包时，会将其提交给br0决定数据包的去向，br0会根据MAC地址与端口的映射关系进行转发。

OVS简介

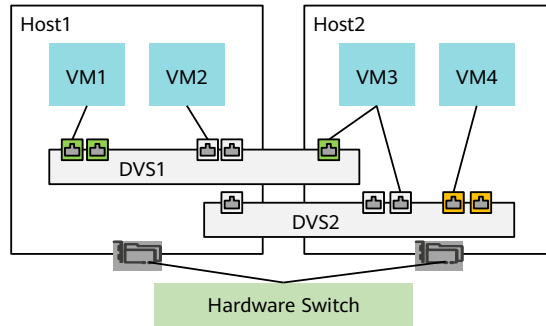
- Open vSwitch (OVS) 是一款基于软件实现的开源虚拟以太网交换机，能够支持多种标准的管理接口和协议，还可以支持跨多个物理服务器的分布式环境。提供了对OpenFlow协议的支持，并且能够与众多开源的虚拟化平台相整合。
- 主要有两个作用：传递虚拟机VM之间的流量，以及实现VM和外界网络的通信。



- Open vSwitch (OVS) 就是一款基于软件实现的开源虚拟交换机。它遵循Apache 2.0许可证。能够支持多种标准的管理接口和协议【例如NetFlow、sFlow、SPAN、RSPAN (Remote Switched Port Analyzer, 远程交换端口分析器)、CLI (Command Line Interface, 命令行接口)、LACP、802.1ag等】，还可以支持跨多个物理服务器的分布式环境（类似于VMware的vSwitch或Cisco的Nexus 1000V）。OVS提供了对OpenFlow协议的支持，并且能够与众多开源的虚拟化平台相整合。
- OpenFlow是Software Defined Network的一种，由斯坦福大学的Nick McKeown教授在2008年4月ACM Communications Review上发表的一篇论文《OpenFlow: enabling innovation in campus networks》里首先提出来。它最初的出发点是用于网络研究人员实验其创新网络架构、协议，考虑到实际的网络创新思想需要在实际网络上才能更好地验证，而研究人员又无法修改在网的网络设备，故而提出了OpenFlow的控制转发分离架构，将控制逻辑从网络设备盒子中引出来，研究者可以通过一组定义明确的接口对网络设备进行任意的编程从而实现新型的网络协议、拓扑架构而无需改动网络设备本身。
- OpenFlow交换机将原来完全由交换机/路由器控制的报文转发过程转化为由OpenFlow交换机 (OpenFlow Switch) 和控制服务器 (Controller) 来共同完成，从而实现了数据转发和路由控制的分离。控制器可以通过事先规定好的接口操作来控制OpenFlow交换机中的流表，从而达到控制数据转发的目的。
- OpenFlow网络由OpenFlow交换机、FlowVisor和Controller三部分组成。OpenFlow交换机进行数据层的转发；FlowVisor对网络进行虚拟化；Controller对网络进行集中控制，实现控制层的功能。
- OpenFlow交换机由流表、安全通道和OpenFlow协议三部分组成。

DVS简介

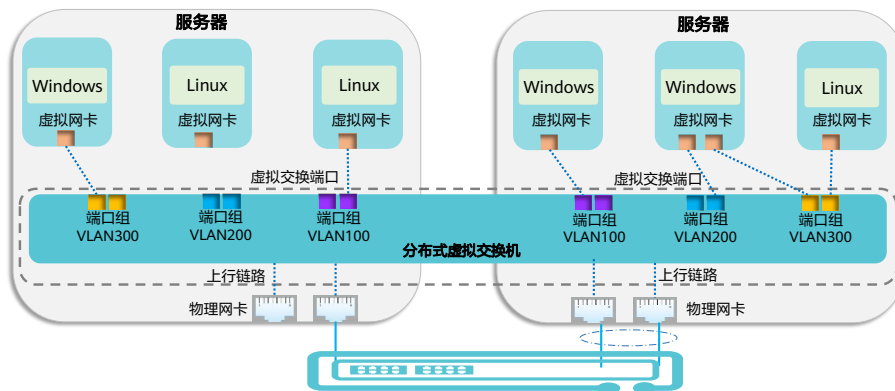
- DVS（分布式虚拟交换机）的功能类似于普通的物理交换机，每台主机都连接到分布式交换机中。分布式虚拟交换机的一端是与虚拟机相连的虚拟端口，另一端是与虚拟机所在主机上的物理以太网适配器相连的上行链路。通过它可以连接主机和虚拟机，实现系统网络互通。另外，分布式虚拟交换机在所有关联主机之间作为单个虚拟交换机使用。此功能可使虚拟机在跨主机进行迁移时确保其网络配置保持一致。



- DVS: Distributed Virtual Switch, 分布式虚拟交换机

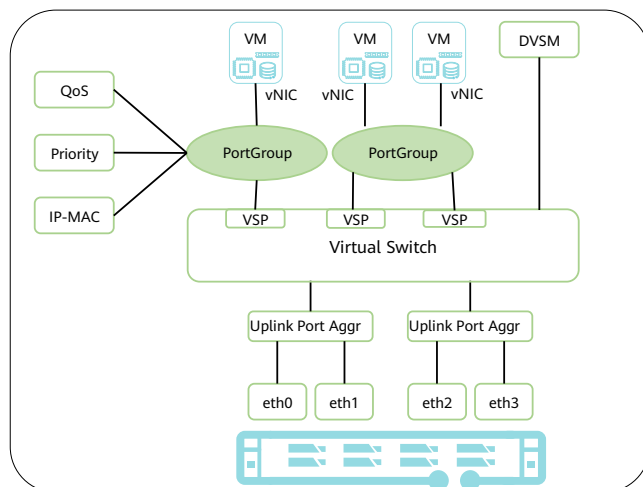
分布式交换机

- 分布式虚拟交换机可以向虚拟机提供独立的网络平面。像物理交换机一样，不同的网络平面间通过VLAN进行隔离。



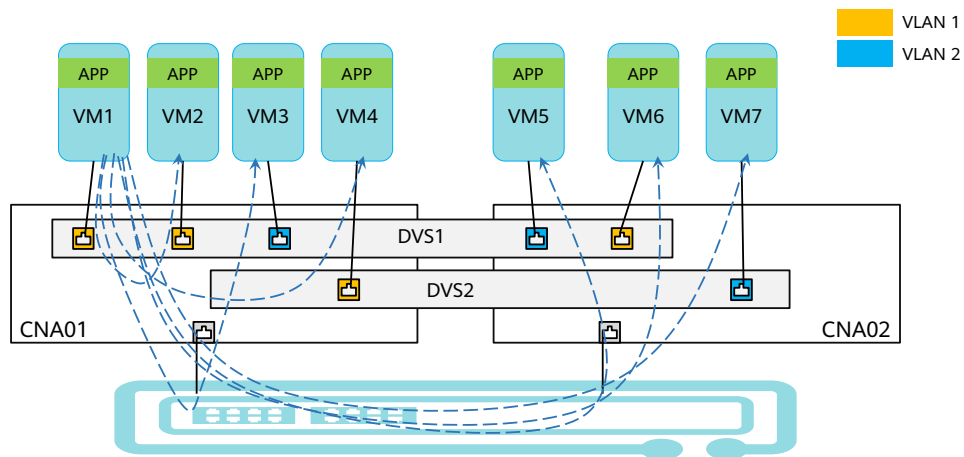
- 通过分布在各物理服务器的虚拟交换机，提供虚拟机的二层通信、隔离、QoS的能力。
- 分布式交换机模型基本特征：
 - 虚拟化管理员可以配置多个分布式交换机，每个分布式交换机可以覆盖集群中的多个CNA节点；
 - 每个分布式交换机具有多个分布式的虚拟端口VSP，每个VSP具有各自的属性（速率），为了管理方便采用Port Group组管理相同属性的一组端口，相同端口组的VLAN相同；
 - 虚拟化管理员或业务系统（例如VDI/IDC），可选择管理/存储/业务使用的不同物理接口；每个分布式交换机可以配置一个Uplink端口或者一个Uplink端口聚合组，用于VM对外的通信。Uplink端口聚合组可以包含多个物理端口，端口聚合组可以配置负载均衡策略；
 - 每个VM可以具有多个vNIC接口，vNIC可以和交换机的VSP一一对接；
 - 虚拟化管理员或业务系统可根据业务需求，选择在一个集群中允许进行2层迁移的服务器创建虚拟二层网络，设置该网络使用的VLAN信息。
- 华为分布式交换机相关概念：上行链路、端口组、VLAN池。

虚拟交换模型



- 管理员可通过定义端口组 属性（安全/QoS）简化对虚拟机端口属性的设置；设置端口组属性，不影响虚拟机正常工作。
- 端口组：是网络属性相同的一组端口的属性集合。
- 上行链路：是分布式交换机连接主机物理网卡的链路，用于虚拟机数据上行。
- 上行链路聚合：分布式交换机关联的服务器绑定网口，绑定网口可以包含多个物理网口，这些物理网口可以配置主备或负载均衡策略。

虚拟机通信

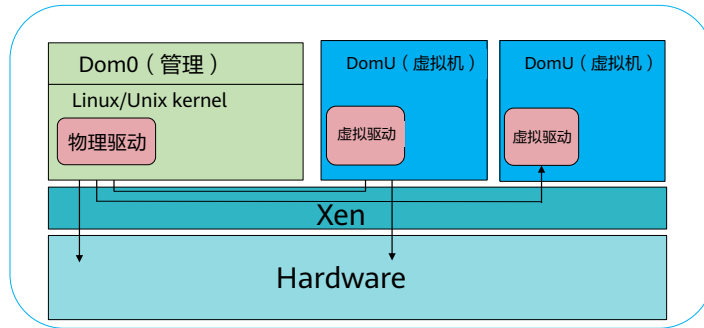


目录

1. 虚拟化技术概览
2. 典型虚拟化软件介绍
 - **Xen和KVM**
 - VMware ESXi
 - Hyper-V
3. 华为虚拟化软件FusionCompute
4. 数据中心虚拟化DCS

Xen虚拟化简介

- Xen的Hypervisor是服务器经过BIOS启动之后载入的首个程序，然后启动一个具有特定权限的虚拟机，称之为Domain 0（简称Dom0）。Dom0的操作系统可以是Linux或Unix，Domain 0实现对Hypervisor控制和管理功能。在所承载的虚拟机中，Dom0是唯一可以直接访问物理硬件（如存储和网卡）的虚拟机，它通过本身加载的物理驱动，为其他虚拟机（Domain U，简称DomU）提供访问存储和网卡的桥梁。



- Xen最初是剑桥大学Xensource的一个开源研究项目，2003年9月发布了首个版本Xen 1.0，2007年Xensource被Citrix公司收购，开源Xen转由www.xen.org继续推进，该组织成员包括个人和公司（如 Citrix、Oracle等）。该组织在2011年3月发布了版本Xen 4.1。
- 相对于ESX和Hyper-V来说，Xen支持更广泛的CPU架构，前两者只支持CISC的x86/x86_64 CPU架构，Xen除此之外还支持RISC CPU架构，如IA64、ARM等。
- Xen支持两种类型的虚拟机，一类是半虚拟化（PV，Paravirtualization），另一类是全虚拟化（Xen称其为HVM，Hardware Virtual Machine）。半虚拟化需要特定内核的操作系统，如基于Linux paravirt_ops（Linux内核的一套编译选项）框架的Linux内核，而Windows操作系统由于其封闭性则不能被Xen的半虚拟化所支持，Xen的半虚拟化有个特别之处就是不要求CPU具备硬件辅助虚拟化，这非常适用于2007年之前的旧服务器虚拟化改造。全虚拟化支持原生的操作系统，特别是针对Windows这类操作系统，Xen的全虚拟化要求CPU具备硬件辅助虚拟化，它修改的QEMU仿真所有硬件，包括：BIOS、IDE控制器、VGA显示卡、USB控制器和网卡等。为了提升I/O性能，全虚拟化特别针对磁盘和网卡采用半虚拟化设备来代替仿真设备，这些设备驱动称之为PV on HVM，为了使PV on HVM有最佳性能，CPU应具备MMU硬件辅助虚拟化。
- Xen的Hypervisor层非常薄，少于15万行的代码量，不包含任何物理设备驱动，这一点与Hyper-V是非常类似的，物理设备的驱动均是驻留在Dom0中，可以重用现有的Linux设备驱动程序。因此，Xen对硬件兼容性也是非常广泛的，Linux支持的，它就支持。

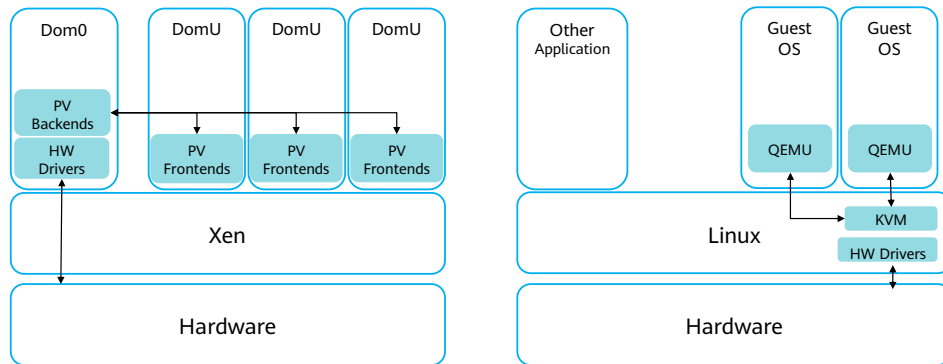
KVM虚拟化简介

- KVM (Kernel-based Virtual Machine) 是基于内核的虚拟机。
- KVM本质是Linux内核中的虚拟化功能模块kvm.ko，利用Linux做大量的事，如任务调度、内存管理与硬件设备交互等。
- KVM是开源软件，于2007年2月被集成到Linux 2.6.20内核中。
- KVM中，虚拟机其实就是一个Linux进程，由CPU进行调度运行。
- KVM运行在内核空间，提供CPU、内存的虚拟化，它本身不执行任何模拟。运行在用户空间的QEMU提供硬件I/O的虚拟化模拟。

- KVM的全称是Kernel-based Virtual Machine，字面意思是基于内核的虚拟机。其最初是由Qumranet公司开发的一个开源项目。2008年，Qumranet被RedHat所收购，但KVM本身仍是一个开源项目，由RedHat、IBM等厂商支持。
- KVM之所以叫做基于内核的虚拟机，是因为KVM本身是一个Linux内核模块，当安装有Linux系统的物理机安装了这个模块后，就变成了Hypervisor，而且还不会影响原先在该Linux上运行的其它应用程序。
- 与Xen类似，KVM支持广泛的CPU架构，除了x86/x86_64 CPU架构之外，还将会支持大型机（S/390）、小型机（PowerPC、IA64）及ARM等。
- KVM充分利用了CPU的硬件辅助虚拟化能力，并重用了Linux内核的诸多功能，使得KVM本身是非常瘦小的，KVM的创始人Avi Kivity声称KVM模块仅有约10000行代码，但我们不能认为KVM的Hypervisor就是这个代码量，因为从严格意义来说，KVM本身并不是Hypervisor，它仅是Linux内核中的一个可装载模块，其功能是将Linux内核转换成一个裸金属的Hypervisor。
- 通过KVM模块的加载将Linux内核转变成Hypervisor，Linux本身运行于内核模式，主机进程运行于用户模式，虚拟机则运行于客户模式，使得转变后的Linux内核可以将主机进程和虚拟机进行统一的管理和调度，这也是KVM名称的由来。
- KVM历史：
 - 2006年10月，以色列公司Qumranet发布KVM；
 - 2006年12月，KVM合入内核（Linux 2.6.20rc）；
 - 2007年2月，Linux 2.6.20正式版发布；
 - 2008年9月，Redhat以1.07亿美元收购Qumranet；
 - 2009年9月，RHEL 5.4开始支持KVM（同时支持Xen）；

- 2010年11月，RHEL 6.0之后仅支持KVM。

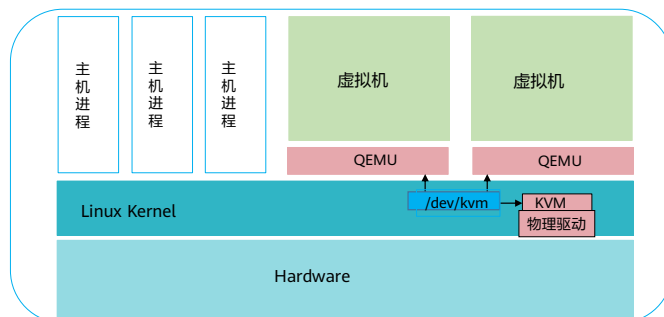
Xen vs KVM



- Xen平台架构侧重安全性：为保证安全性，各Domain对共享区域的访问和映射必须通过Hypervisor授权。
- KVM平台架构侧重性能：VM之间以及与Host Kernel之间对共享区域的访问和映射无需Hypervisor进行授权，故整个访问路径较短。使用Linux Baremetal内核，无PVOPS性能损耗。

KVM与QEMU

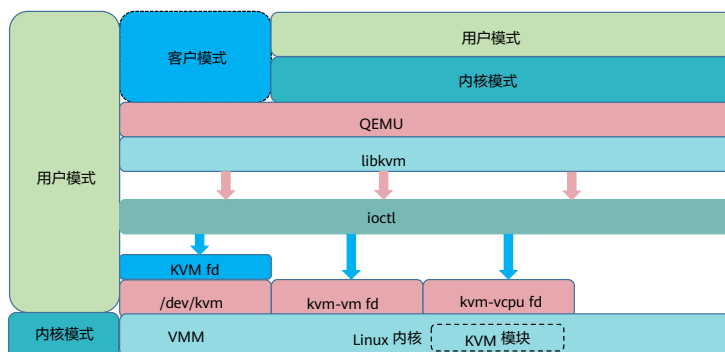
- 在KVM虚拟化方案中，KVM主要用于管理CPU和内存的虚拟化，IO设备的虚拟化则由QEMU来完成。
- QEMU是一个纯软件实现的开源（模拟）软件，它能够模拟整套虚拟机的实现，包括CPU、内存、IO设备、USB、网卡等。



- KVM用来模拟CPU的运行，但缺少了对Network和I/O的支持。QEMU-KVM是一个完整的模拟器，它基于KVM，提供了完整的I/O模拟支持。其中OpenStack为了跨VM性，所以不会直接控制QEMU-KVM，而是通过Libvirt的库去间接控制QEMU-KVM，后续我们会对Libvirt进行介绍。
- KVM离不开QEMU。KVM实现初期，为了简化开发和代码重用，在QEMU基础上进行了修改，主要是将比较消耗CPU性能的CPU虚拟化和内存虚拟化部分移交到了内核中实现，保留IO虚拟化模块在用户空间实现。避免了用户态和内核态的频繁切换，优化使用性能。
- QEMU离不开KVM。之前我们提到，QEMU是一个纯软件的实现，运行在用户空间，性能非常低下，所以从QEMU的角度可以说是QEMU使用了KVM的虚拟化功能，为自身虚拟机提供资源与加速。
- /dev/kvm接口是QEMU和KVM交互的“桥梁”。/dev/kvm本身是一个设备文件，我们可以通过ioctl函数来对该文件进行控制和管理，从而完成用户空间与内核空间的数据交互。KVM与QEMU的通信过程主要就是一系列针对该设备文件的ioctl系统调用。

KVM工作原理

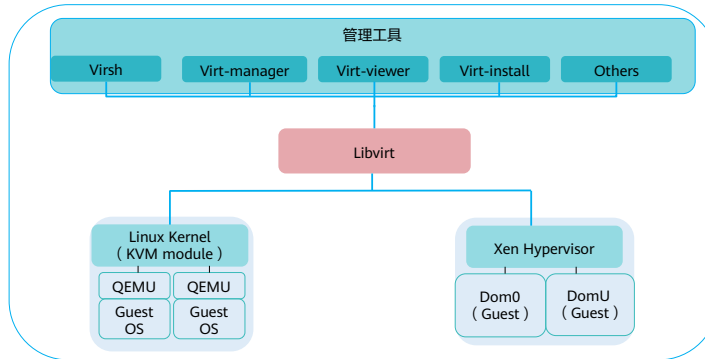
- KVM是Linux Kernel的一个模块，运行在内核空间。
- QEMU运行在用户空间，提供硬件I/O设备的虚拟化模拟。
- Linux系统安装KVM模块后，会有如下三种运行模式：客户模式、用户模式、内核模式。



- KVM基本结构如上图。KVM已经是内核模块，被看作是一个标准Linux字符集设备（/dev/kvm）。QEMU通过Libkvm应用程序接口，用fd（文件描述符）通过ioctl向设备驱动来发送创建、运行虚拟机命令。设备驱动KVM会解析命令。
- KVM模块让Linux主机成为了一个虚拟机监视器（VMM），在原有执行模式基础上，增加了客户模式。在虚拟机运行时，三种模式的工作为：
 - 客户模式：执行非I/O的客户代码，虚拟机运行在这个模式下。
 - 用户模式：代表用户执行I/O指令，QEMU运行在这个模式下，它用来为虚拟机模拟执行I/O类的操作请求。
 - 内核模式：实现客户模式切换，处理因I/O或者其他指令引起的从客户模式退出动作（VM-Exit）。KVM模块工作在这个模式下。此模式下可以真正操作硬件，当Guest OS执行I/O类操作或特权指令操作时，需要向用户模式提交请求，然后由用户模式再次发起硬件操作请求给内核模式，从而真正操作硬件。

虚拟化平台管理工具 - Libvirt

- Libvirt是一套由C语言开发的API，主要目标是提供一种通用并且稳定的软件层，来管理物理主机上多种不同的虚拟化方式和虚拟主机，并支持远程管理。
- Libvirt是Linux上的虚拟化库，Libvirt也是一个开源项目，它是一个非常强大的虚拟化平台管理工具，被管理的虚拟化平台可以是KVM，也可以是Xen、VMware以及Hyper-V等。



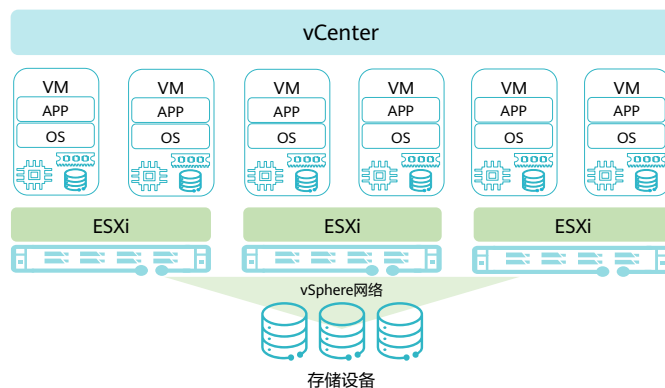
- 虚拟化领域针对不同的场景提出了很多虚拟化解方案（KVM、Xen等），为了支持更多厂商以及更多领域，很多IaaS解决方案需要融合多种虚拟化技术，在这个需求背景下，Libvirt就为用户提供了一个平台类的管理工具，同时支持多种虚拟化方案。
- Libvirt是为了更方便地管理平台虚拟化技术而设计的开放源代码的应用程序接口、守护进程和管理工具，它不仅提供了对虚拟化客户机的管理，也提供了对虚拟化网络和存储的管理。
- Libvirt对多种不同的Hypervisor的支持是通过一种基于驱动程序的架构来实现的。Libvirt对不同的Hypervisor提供了不同的驱动：对Xen有Xen的驱动，对QEMU/KVM有QEMU驱动。
- Libvirt作为中间适配层，让底层Hypervisor对上层用户空间的管理工具做到完全透明，因为Libvirt屏蔽了底层各种Hypervisor的细节，为上层管理工具提供了一个统一的、较稳定的接口（API）。

目录

1. 虚拟化技术概览
2. 典型虚拟化软件介绍
 - Xen和KVM
 - **VMware ESXi**
 - Hyper-V
3. 华为虚拟化软件FusionCompute
4. 数据中心虚拟化DCS

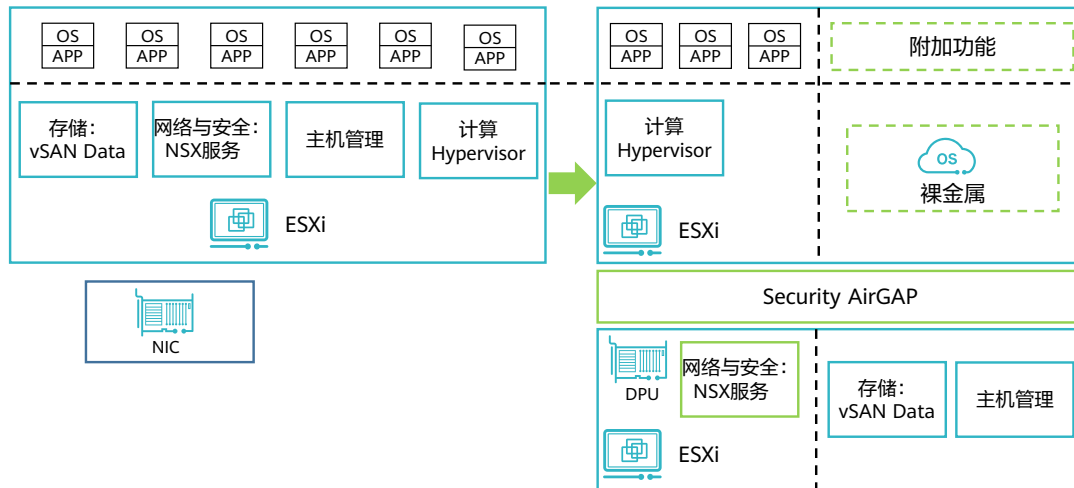
VMware vSphere介绍

- VMware vSphere是VMware的虚拟化平台，可将数据中心转换为包括CPU、存储和网络资源的聚合计算基础架构。



- vSphere的两个核心组件是ESXi和vCenter Server。ESXi是用于创建并运行虚拟机和虚拟设备的虚拟化平台。vCenter Server是一项服务，用于管理网络中连接的多个主机，并将主机资源池化。

vSphere架构演进



- vSphere Distributed Services Engine® 是一项核心 vSphere 功能，使客户能够将DPU（data processing units，数据处理单元）与vSphere和VMware Cloud Foundation 结合使用。
- vSphere Distributed Services Engine通过在DPU上引入VMware vSphere Distributed Switch和VMware NSX网络和可观察性来卸载和加速DPU上的基础架构功能，从而使您能够主动监控、识别和缓解网络基础架构瓶颈，而无需使用复杂的网络。DPU成为扩展基础架构功能的新控制点，并支持与工作负载域分离的无代理安全控制。

vSphere组件介绍

- VMware vSphere是用于虚拟化的软件组件套件。这些软件组件包括ESXi、vCenter Server和其他软件组件（如vCenter Single Sign-On）以及vCenter Server 插件（可在vSphere环境中实现多种不同的功能）。

ESXi

- 通过ESXi，可以运行虚拟机，安装操作系统，运行应用程序以及配置虚拟机。配置包括识别虚拟机的资源，如存储设备。

vCenter Server

- 一种充当连接网络的VMware ESXi主机的中心管理员的服务。vCenter Server指导虚拟机和ESXi主机上的操作。

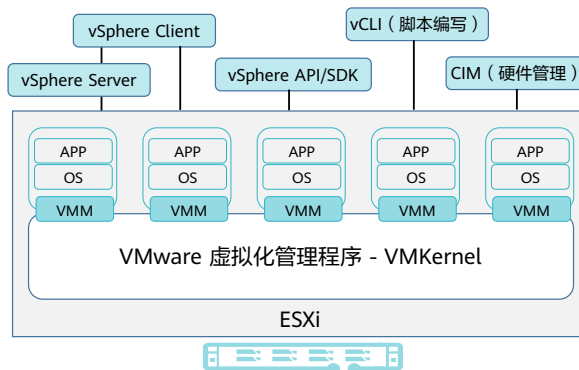
vCenter Single Sign-On

- vCenter Single Sign-On身份验证服务支持各种vSphere软件组件彼此进行相互通信，从而提高了VMware Cloud基础架构平台的安全性。vCenter Single Sign-On身份验证服务使用安全令牌交换机制，而无需使用每个组件单独对目录服务（如Active Directory）进行用户身份验证。

- 2009年，VMware推出ESX，可以直接安装在物理服务器上，这种安装方式称为裸机安装。但是ESX并不能完全的摒弃宿主操作系统，他的解决方法是将虚拟化程序和操作系统整合到一起。也就是说，他将虚拟化的程序写入到Linux的操作系统中。此时，ESX通过Linux系统的Servers Console来运行，而资源和虚拟机的管理工作则通过合作代理伙伴和编写脚本来执行。ESX有效的解决了对宿主操作系统过于依赖的问题。但是这种架构依然有他自己的缺陷，首先，由于虚拟化程序中包含Linux的操作系统，Linux的操作系统中非虚拟化部分的进程会占用主机上的部分资源，造成资源的浪费。其次，在进行资源和虚拟机的管理时，只能通过脚本和代理，非常的不方便。
- 2011年，VMware推出了ESXi，即为当前应用最广泛的虚拟化产品。同样，ESXi也是裸机安装在物理计算机上的，它做的改进是将虚拟化层中繁杂的Linux层剔除，只保留了VMkernel虚拟化内核对资源进行管理。这样，便大大降低了虚拟化层的大小。同时，也减小了虚拟化层对物理化层的开销。ESXi做的第二大改进便是将控制台从虚拟化程序中移除，变成一个独立的组件，即vSphere Client，使得管理工作更加的轻松便捷。

ESXi架构

- 从体系结构来说ESXi包含虚拟化层和虚拟机，而虚拟化层有两个重要组成部分：虚拟化管理程序VMkernel和虚拟机监视器VMM。ESXi主机可以通过vSphere Client、vCLI、API/SDK和CIM接口接入管理。



- ESXi是组成vSphere基础架构核心的虚拟化管理器，可直接安装在物理服务器之上，并允许多个虚拟机运行于虚拟化层之上。每个虚拟机与其他虚拟机共享相同的物理资源，并且它们可以同时运行。与其他虚拟化管理程序不同，ESXi的所有管理功能都可以通过远程管理工具提供。
- ESXi体系结构独立于任何通用操作系统运行，可提高安全性、增强可靠性并简化管理。紧凑型体系结构设计旨在直接集成到针对虚拟化进行优化的服务器硬件中，从而实现快速安装、配置和部署。

ESXi组件

VMkernel

- VMkernel 是虚拟化的核心和推动力，由 VMware 开发并提供与其他操作系统提供的功能类似的某些功能，如进程创建和控制、信令、文件系统和进程线程。

虚拟机监视器VMM

- 每个 ESXi 主机的关键组件是一个称为 VMM 的进程。对于每个已开启的虚拟机，将在 VMkernel 中运行一个 VMM。虚拟机开始运行时，控制权将转交给 VMM，然后由 VMM 依次执行虚拟机发出的指令。

虚拟机

- 虚拟机是一个由VMkernel控制的软件构造体。所有虚拟机配置信息、状态信息和数据都封装在存储在数据存储中的一组离散文件中。这使虚拟机具有可移动性，并且易于备份或克隆。

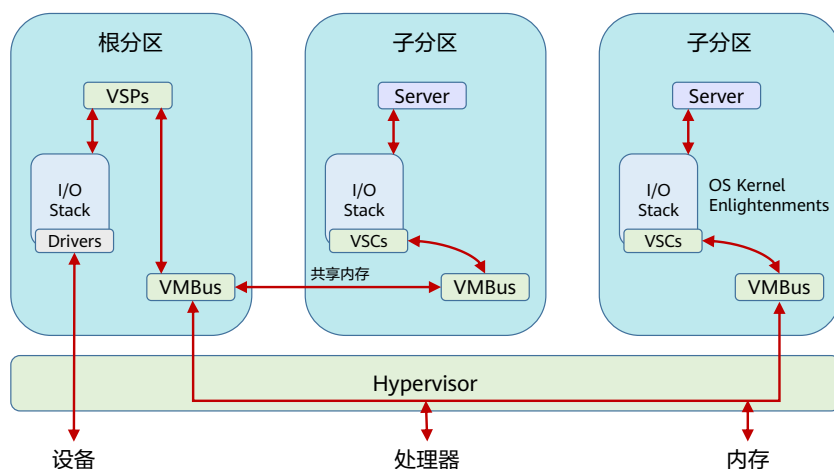
目录

1. 虚拟化技术概览
2. 典型虚拟化软件介绍
 - Xen和KVM
 - VMware ESXi
 - **Hyper-V**
3. 华为虚拟化软件FusionCompute
4. 数据中心虚拟化DCS

Hyper-V介绍

- Hyper-V 是 Windows Server 和 Azure Local 中的虚拟化服务器角色。虚拟化服务器可以托管多个彼此隔离的虚拟机，但通过虚拟化处理器、内存和 I/O 设备来共享基础硬件资源。通过将服务器合并到单台计算机上，虚拟化可以提高资源使用率和节能性，并降低服务器的运营和维护成本。此外，虚拟机和管理 API 为管理资源、均衡负载和预配系统提供了更大的灵活性。
- Hyper-V 是基于虚拟机监控程序的虚拟化技术，适用于某些 x64 版本的 Windows。虚拟机监控程序是虚拟化的核心。它是特定于处理器的虚拟化平台，允许多个独立的操作系统共享单个硬件平台。

Hyper-V体系结构



- Hyper-V具有基于类型1虚拟机监控程序的体系结构。虚拟机监控程序可虚拟化处理器和内存。它为根分区中的虚拟化堆栈提供机制来管理子分区、虚拟机（VM），并向VM公开 I/O（输入/输出）设备等服务。
- 根分区拥有并可以直接访问物理I/O设备。根分区中的虚拟化堆栈为VM、管理API和虚拟化I/O设备提供内存管理器。它还实现模拟设备，例如集成设备电子（IDE）磁盘控制器和PS/2输入设备端口。它支持特定于Hyper-V的合成设备，以提高性能和降低开销。
- Hyper-V的I/O体系结构由根分区中的虚拟化服务提供商（VSP，virtualization service providers）和子分区中的虚拟化服务客户端（VSC，virtualization service clients）组成。每个服务作为设备通过VM总线（VMBus）暴露，该总线充当I/O总线，并通过共享内存（Shared Memory）等机制在使用这些机制的VM之间实现高性能通信。来宾操作系统的即插即用管理器枚举这些设备，包括VM总线，并加载相应的设备驱动程序、虚拟服务客户端。除I/O以外的服务也通过此体系结构公开。

Hyper-V的主要性能特征

改进的硬件共享体系结构

- Hyper-V在运行具有虚拟机监控程序感知内核的来宾操作系统时，提供对核心资源（例如磁盘、网络和视频）的改进访问和利用率，这些操作系统配备了必要的虚拟服务器客户端（VSC）代码（称为Hyper-V启发式I/O）。启蒙是操作系统的增强功能，可帮助降低某些操作系统功能（如内存管理）的成本。集成组件（包括VSC驱动程序）也可用于其他客户端操作系统。

处理器硬件辅助虚拟化支持

- Hyper-V充分利用最新处理器技术提供的处理器硬件辅助虚拟化支持。

多核（SMP）来宾操作系统支持

- Hyper-V提供在虚拟机环境中最多支持四个处理器的功能，使应用程序能够充分利用虚拟机中的多线程功能。

32位和64位来宾操作系统支持

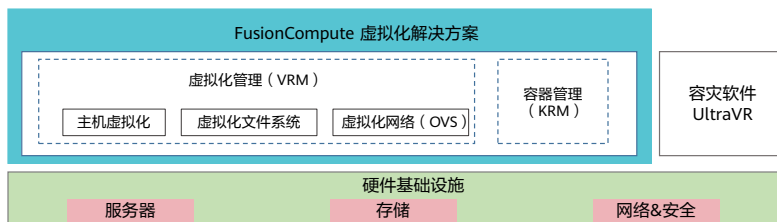
- Hyper-V为同时运行不同类型的操作系统提供广泛支持，包括跨不同服务器平台（如Windows、Linux等）的32位和64位系统。

目录

1. 虚拟化技术概览
2. 虚拟化软件介绍
3. **华为虚拟化软件FusionCompute**
 - **FusionCompute介绍**
 - FusionCompute特性
4. 数据中心虚拟化DCS

FusionCompute虚拟化解决方案

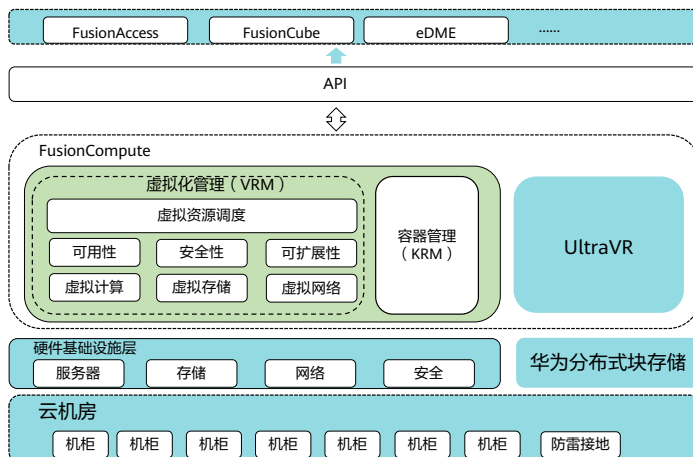
- 华为FusionCompute虚拟化是业界领先的虚拟化解决方案，能够帮助客户带来如下的价值，从而大幅提升数据中心基础设施的效率：
 - 帮助客户提升数据中心基础设施的资源利用率；
 - 帮助客户成倍缩短业务上线周期；
 - 帮助客户成倍降低数据中心能耗；
 - 利用虚拟化基础设施的高可用和强恢复能力，实现业务快速自动化故障恢复，降低数据中心成本和增加系统应用的正常运行时间。
- FusionCompute虚拟化解决方案通过在服务器上部署虚拟化软件，使一台物理服务器可以承担多台服务器的工作。通过整合现有的工作负载并利用剩余的服务器资源以部署新的应用程序和解决方案，实现较高的整合率。



- 应用场景介绍：
 - 单虚拟化场景：单虚拟化场景适用于企业只采用FusionCompute作为统一的操作维护管理平台对整个系统进行操作与维护的应用场景。包含资源监控、资源管理、系统管理等。FusionCompute主要负责硬件资源的虚拟化，以及对虚拟资源、业务资源、用户资源的集中管理。它采用虚拟计算、虚拟存储、虚拟网络等技术，完成计算资源、存储资源、网络资源的虚拟化。同时通过统一的接口，对这些虚拟资源进行集中调度和管理，从而降低业务的运行成本，保证系统的安全性和可靠性。
- UltraVR是可选部件。UltraVR定位为企业数据中心容灾业务管理软件，简单高效地完成容灾业务配置，清晰可视地掌控系统容灾业务的运行情况，快速方便地完成数据恢复和测试。

FusionCompute的定位

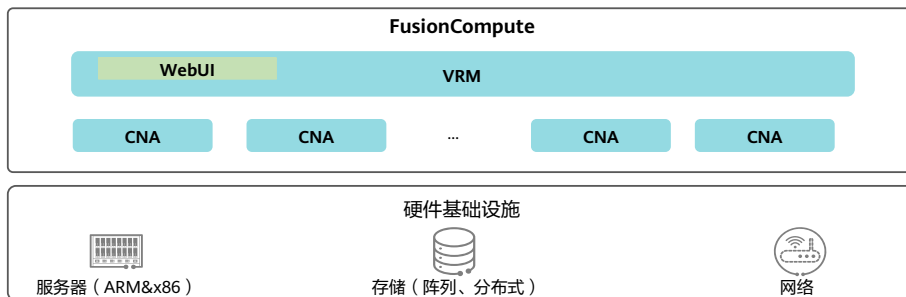
- FusionCompute是**云操作系统软件**，主要负责硬件资源的虚拟化，以及对虚拟资源、业务资源、用户资源的集中管理。它采用虚拟计算、虚拟存储、虚拟网络等技术，完成计算资源、存储资源、网络资源的虚拟化。
- 通过统一的接口，对虚拟资源进行集中调度和管理，从而降低业务的运行成本，保证系统的安全性和可靠性，协助运营商和企业构筑安全、绿色、节能的云数据中心能力。



- 虚拟化技术除了商用解决方案FusionCompute外，还支持了华为众多业务，比如研发计算云，超融合FusionCube，桌面云FusionAccess等，为这些产品提供基座能力。
- FusionCompute支持容器管理特性提供了企业级K8S平台，并提供租户项目管理、集群生命周期管理、容器镜像和应用管理、容器监控运维的能力，是用户应用现代化的最佳底座平台。本课程不展开介绍。
- 云机房：为云数据中心运营所提供的一系列基本配套设施和空间，主要包括供电系统、消防系统、布线系统、制冷系统等。
- 硬件基础设施层：硬件资源包括服务器、存储、网络、安全等全面的云计算基础物理设备，支持用户从中小规模到大规模的新建或扩容，可运行从入门级到企业级的各种企业应用。设备类型丰富，可为客户提供灵活的部署选择。
- 华为分布式块存储：华为分布式块存储是一种存储与计算高度融合的分布式存储软件，在通用服务器上部署该软件后，可以把所有服务器的本地硬盘组织成一个虚拟存储资源池，提供块存储功能。
- UltraVR：UltraVR是容灾业务管理软件，利用底层SAN存储系统提供的异步远程复制特性，提供虚拟机关键数据的数据保护和容灾恢复。

FusionCompute产品架构

- FusionCompute包含两个模块：
 - CNA（Compute Node Agent），部署在需要虚拟化的服务器上。
 - VRM（Virtual Resource Management），部署在VM上或者部署在物理服务器上；VRM对外提供网页操作界面给管理维护人员。

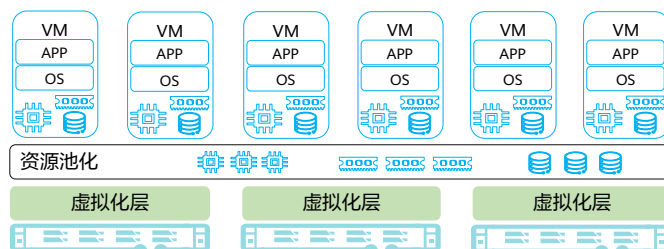


FusionCompute功能模块

模块	介绍	功能
CNA	CNA (Compute Node Agent) ，部署在需要虚拟化的服务器上。	• 提供虚拟计算功能； • 管理计算节点上的虚拟机； • 管理计算节点上的计算、存储、网络资源。
VRM	VRM (Virtual Resource Management) ，可以作为一台虚拟机进行部署，也可以直接部署在物理服务器上； VRM对外提供网页操作界面给管理维护人员。	• 管理集群内的块存储资源； • 管理集群内的网络资源 (IP/VLAN) ，为虚拟机分配IP地址； • 管理集群内虚拟机的生命周期以及虚拟机在计算节点上的分布和迁移； • 管理集群内资源的动态调整； • 通过对虚拟资源、用户数据的统一管理，对外提供弹性计算、存储、IP等服务； • 通过提供统一的操作维护管理接口，操作维护人员通过WebUI远程访问FusionCompute对整个系统进行操作维护，包含资源管理、资源监控、资源报表等。

FusionCompute产品功能 - 虚拟化计算

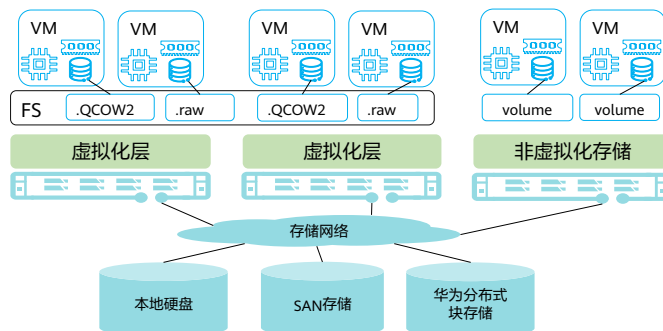
- FusionCompute支持将服务器物理资源抽象成逻辑资源，让一台服务器变成几台甚至上百台相互隔离的虚拟服务器，不再受限于物理上的界限，让CPU、内存、磁盘、I/O等硬件变成可以动态管理的“资源池”，从而提高资源的利用率，简化系统管理。
- 对于最终用户，虚拟机比物理机的优势在于它可以很快速地发放，很方便地调整配置和组网。对于维护人员来讲，虚拟机复用了硬件，这样硬件更少，加上云平台的自动维护能力，维护成本显著降低。对于系统管理员，可以很直观地知道资源使用的总量及变化趋势，以便决策是否扩容。



- 将服务器物理资源抽象成逻辑资源，让一台服务器变成几台甚至上百台相互隔离的虚拟服务器，不再受限于物理上的界限，而是让CPU、内存、磁盘、I/O等硬件变成可以动态管理的“资源池”，从而提高资源的利用率，简化系统管理。同时硬件辅助虚拟化技术提升虚拟化效率，增加虚拟机的安全性。

FusionCompute产品功能 - 虚拟化存储

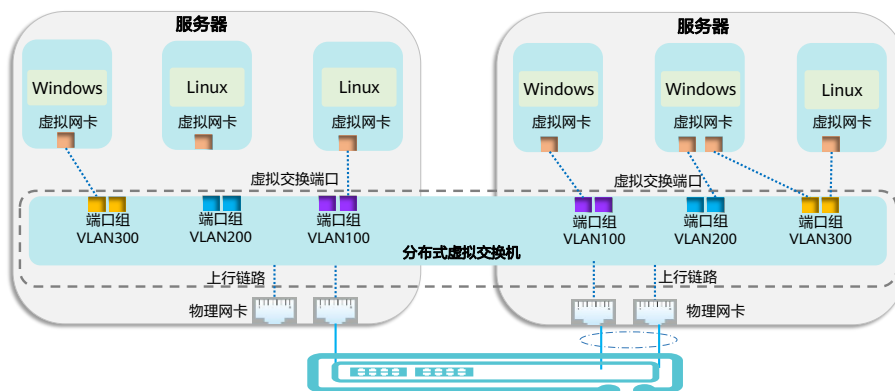
- FusionCompute存储虚拟化技术可以更好地管理虚拟基础架构的存储资源，使系统大幅提升存储资源利用率和灵活性，提高应用的正常运行时间。FusionCompute支持将SAN设备、计算节点本地硬盘以及华为分布式块存储提供的虚拟存储空间统一管理，以虚拟卷的形式分配给虚拟机使用。



- 存储设备的能力、接口协议等差异性很大，存储虚拟化技术可以将不同存储设备进行格式化，将各种存储资源转化为统一管理的数据存储资源，可以用来存储虚拟机磁盘、虚拟机配置信息、快照等信息。用户对存储的管理更加同质化。
- 虚拟机磁盘、快照等内存均以文件的形式存放在数据存储上，所有业务操作均可以转化成对文件的操作，操作更加直观、便捷。
- 基于存储虚拟化平台提供的众多存储业务，可以提高存储利用率，更好的可靠性、可维护性、可以带来更好的业务体验和用户价值。
- 华为提供基于主机的存储虚拟化功能，用户不需要再关注存储设备的类型和能力。存储虚拟化可以将存储设备进行抽象，以逻辑资源的方式呈现，统一提供全面的存储服务。可以在不同的存储形态，设备类型之间提供统一的功能。
- 对于最终用户，就像x86服务器使用本地硬盘一样的方式使用，可以格式化，安装文件系统，安装操作系统，读写。同时，虚拟化的存储还具备快照能力，可以调整大小，这是物理硬盘不能实现的。
- 对于管理员来说，虚拟存储卷并不一对一映射到某块具体的磁盘，而是收敛到几台SAN设备。由于SAN设备已经有了可靠性保障，所以更换硬盘的工作量大规模下降。同时，虚拟存储具备瘦分配，灵活调整，QoS可限制，可迁移等等比物理盘强的特性，在整体成本方面优势很明显。

FusionCompute产品功能 - 虚拟化网络

- FusionCompute支持分布式虚拟交换，可以向虚拟机提供独立的网络平面。像物理交换机一样，不同的网络平面间通过VLAN进行隔离

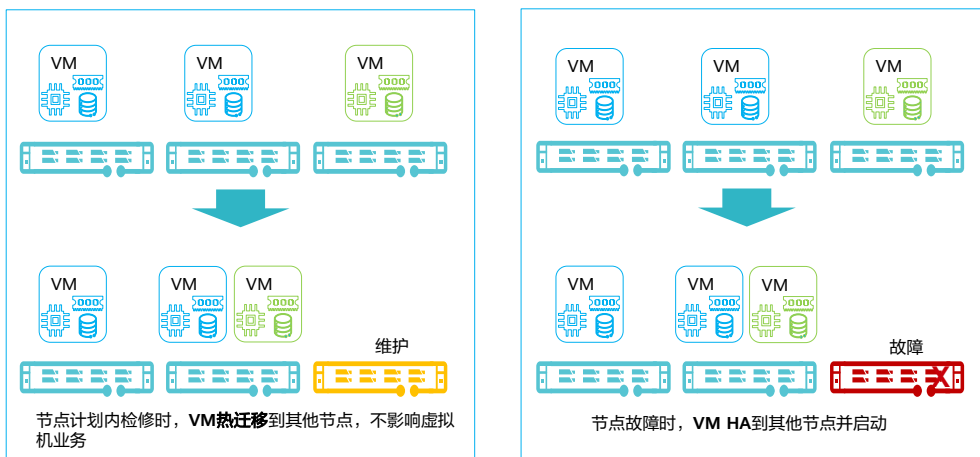


67 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



- 虚拟网卡均有自己的IP地址、MAC地址，从网络角度来看，虚拟网卡与物理网卡一致。
- 网络QoS策略提供带宽配置控制能力，包括基于端口组成员接口发送方向与接收方向的带宽控制，基于端口组的每个成员接口提供流量整形、带宽优先级的控制能力。
- 分布式交换机的功能类似于普通的物理交换机，每台主机都连接到分布式交换机中。分布式交换机的一端是与虚拟机相连的虚拟端口，另一端是与虚拟机所在主机上的物理以太网适配器相连的上行链路。通过它可以连接主机和虚拟机，实现系统网络互通。另外，分布式交换机在所有关联主机之间作为单个虚拟交换机使用。此功能可使虚拟机在跨主机进行迁移时确保其网络配置保持一致。

FusionCompute产品功能 - 可用性



- 热迁移：FusionCompute支持跨计算集群内自由迁移虚拟机。在虚拟机迁移期间，用户业务不会有任何中断。如果迁移失败，目的端的虚拟机将销毁，而用户仍可以使用源端虚拟机。该功能可避免因服务器维护造成的业务中断，降低数据中心的电能消耗。
- 故障迁移（HA，High Availability）：支持虚拟机故障后自动重启。配置集群时，可以选择是否开启HA功能，即是否支持HA功能。系统周期检测虚拟机状态，当物理服务器故障引起虚拟机故障时，系统会根据主机故障处理策略在原物理服务器或其他物理服务器上重新启动虚拟机，保证虚拟机能够快速恢复。重新启动的虚拟机，会像物理机一样重新开始引导，加载操作系统，所以发生故障时未保存的内容将丢失。
- 目前系统能够检测到的引起虚拟机故障的原因包括物理硬件故障、系统软件故障。

FusionCompute产品功能 - 安全

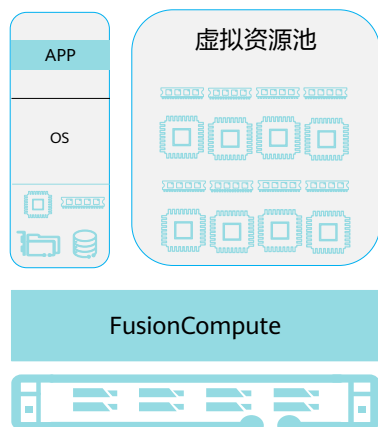
- 虚拟网络访问控制
 - 通过设置虚拟网卡的VLAN ID划分虚拟机使用的网络范围；
 - 通过动态修改虚拟网卡所属PortGroup，实现虚拟网卡VLAN ID的动态修改；
 - 动态修改虚拟网卡的VLAN ID，不再通过添加新网卡，重新绑定VLAN实现VLAN的修改。

- 除了FC中的虚拟网络访问控制，根据云计算领域面临的威胁与挑战，华为还针对不同的安全方向提供对应的安全解决方案：
 - 云平台安全：数据存储安全、虚拟机隔离和网络传输安全。
 - 运维管理安全：权限管理、账号密码管理和日志管理。
 - 主机安全：包括Web安全、操作系统加固、数据加固和安全补丁。

目录

1. 虚拟化技术概览
2. 虚拟化软件介绍
3. **华为虚拟化软件FusionCompute**
 - FusionCompute介绍
 - **FusionCompute特性**
4. 数据中心虚拟化DCS

虚拟机资源管理 - CPU/内存在线调整



技术原理

- vRAM, vCPU不仅支持离线添加/删除, 还支持在线添加。

技术特点

- 平台支持在虚拟机运行情况下调整CPU和内存规格, 虚拟机不需要重启即可生效。

适用场景

- 根据业务需要, 灵活调整虚拟机的CPU、内存数量配置。

应用价值

- 根据虚拟机的需求, 灵活调整其配置;
- 纵向扩展有效保证单个虚拟机QoS;
- 与横向扩展有机结合, 保证集群QoS。

- 支持主流Linux系统虚拟机在线调整CPU和内存; 支持主流Windows系统虚拟机在线调整内存, 在线调整CPU后重启生效。
- 虚拟机处于离线状态, 可以减少虚拟机vCPU数目/内存容量。
- FusionCompute支持离线/在线添加/删除网卡、离线/在线挂载磁盘。虚拟机处于离线或在线状态, 且虚拟机使用的磁盘是虚拟化存储时, 用户可通过增加已有磁盘容量的方式进行虚拟机存储容量的扩容。

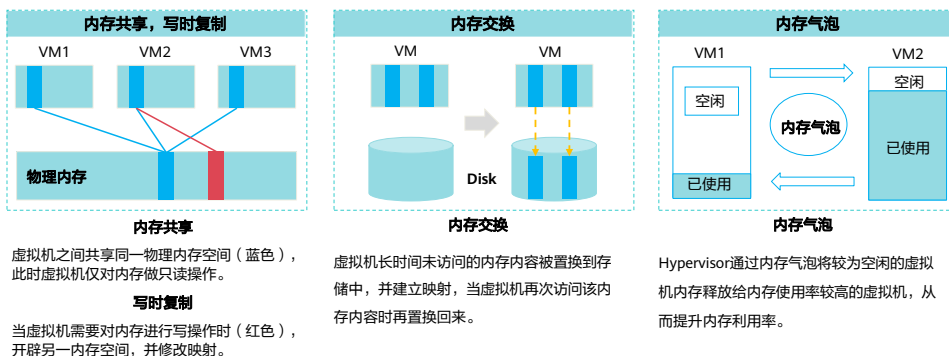
虚拟机资源管理 - CPU QoS

- IT管理员可以对虚拟机设定资源的使用上限，防止非关键应用和恶意用户争抢共享资源。
- 虚拟机的CPU QoS用于保证虚拟机的计算资源分配，隔离虚拟机间由于业务不同而导致的计算能力相互影响，满足不同业务对虚拟机计算性能的要求，最大程度复用资源，降低成本。
- 创建虚拟机时，可根据虚拟机预期部署业务对CPU的性能要求而指定相应的CPU QoS。不同的CPU QoS代表了虚拟机不同的计算能力。指定CPU QoS的虚拟机，系统对其CPU的QoS保障，主要体现在计算能力的最低保障和资源分配的优先级。

- QoS: Quality of Service，服务质量，虚拟机的CPU QoS用于保证虚拟机的计算资源分配，最大程度复用资源，降低成本。
- 一句话总结：计算资源按需分配，优化竞争优先级，资源预留，计算性能最低保障，确保关键业务性能。
- CPU QoS包含如下三个参数：
 - CPU资源份额：CPU份额定义多个虚拟机在竞争物理CPU资源的时候按比例分配计算资源。
 - CPU资源预留：CPU预留定义了多个虚拟机竞争物理CPU资源的时候分配的最低计算资源。
 - CPU资源限额：控制虚拟机占用物理CPU资源的上限。

虚拟机资源管理 - 主机内存超分配

- FusionCompute支持虚拟机总配置内存超过物理主机实际运行内存，通过虚拟机之间使用内存气泡、内存交换以及内存共享等技术实现物理内存超规格使用。

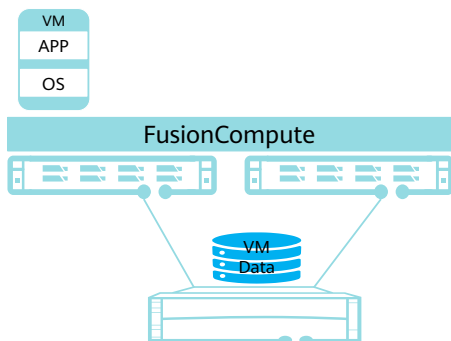


同等内存资源条件下，提升虚拟机密度，降低50%的硬件（内存）采购成本

- 内存复用技术是否开启，可配置。一旦开启，则三种皆开启。
- Swap交换卷——内存置换：将外部存储虚拟成内存给虚拟机用，将虚拟机上应用暂时不用的数据存放到外部存储上，系统需要这些数据时再和预留内存上的数据进行交换。
- 内存气泡：多个相同OS的虚拟机可以共享物理内存。VM系统主动回收虚拟机暂时不用的物理内存，分配给需要复用内存的虚拟机。内存的回收和分配均为系统动态执行，虚拟机上的应用无感知。整个物理服务器上的所有虚拟机使用的分配内存总量不能超过该服务器的物理内存总量。
- 内存交换：将外部存储虚拟成内存给虚拟机使用，将虚拟机上暂时不用的数据存放到外部存储上。系统需要使用这些数据时，再与预留在内存上的数据进行交换。
- 内存共享：多台虚拟机共享数据内容相同的内存页。

虚拟机热迁移

- 支持在不中断业务的情况下，将同一个集群中虚拟机从一台物理服务器移动至另一台物理服务器。



技术特点

- 基于内存压缩传输技术，虚拟机热迁移效率提升1倍；
- 虚拟机磁盘数据位置不变，只更改映射关系。

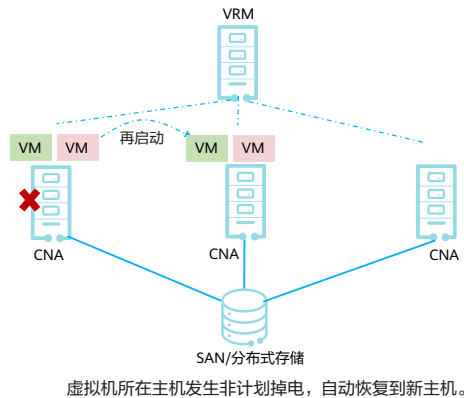
适用场景

- 可容忍短时间中断，但必须要快速恢复业务，比如轻量级数据库业务，桌面云业务。

- 支持在不影响用户使用或中断服务的情况下在服务器之间实时迁移虚拟机，从而无需为进行计划内服务器维护而安排应用中断。
- 虚拟机管理器提供内存数据快速复制和共享存储技术，确保虚拟机迁移前后数据不变。
- 虚拟机热迁移是后续动态资源调度和分布式电源管理技术的基础。
- 在进行服务器操作维护前，系统维护人员将该服务器上的虚拟机迁移到其他服务器，降低操作维护过程中业务中断的风险。
- 在进行服务器升级操作前，系统维护人员将该服务器上的虚拟机迁移到其他服务器，升级完成后将所有虚拟机迁回，降低服务器升级过程中业务中断的风险。
- 将空闲服务器上的虚拟机迁移到其他服务器，将没有负载的服务器关闭，降低业务运行成本。

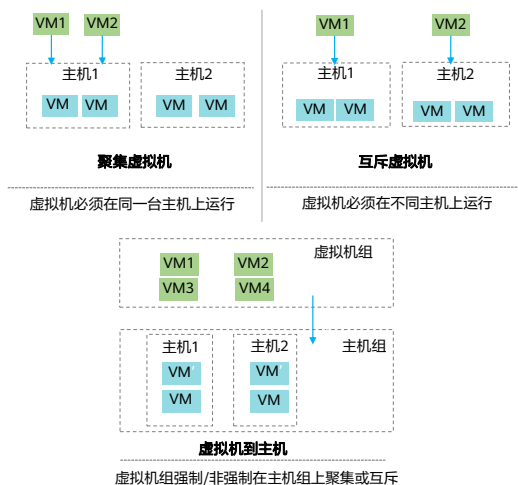
虚拟机HA

- 当计算节点上的虚拟机出现故障时，系统自动将故障的虚拟机在正常的计算节点上重新创建，使故障虚拟机快速恢复。



- 支持主机、虚拟平台、虚拟机内部多种故障场景的检测和虚拟机恢复。
- 支持集中控制HA和集群自治HA两种机制，可自选。
- 可设置HA心跳消息的网络平面，降低网络压力。
- 多种故障判断机制，避免漏判、误判故障。
- 支持共享存储与本地存储虚拟机HA。
- 当计算节点掉电恢复或重启时，系统将计算节点上具有高可用属性的虚拟机重新创建至其他计算节点。当系统检测到虚拟机故障时，系统将选择正常的计算节点，将故障虚拟机在正常的计算节点上重新创建。
- 虚拟机蓝屏（x86）当系统检测到虚拟机蓝屏故障且该虚拟机蓝屏处理策略配置为高可用性时，系统选择其他正常的计算节点重新创建虚拟机。

规则组



聚集虚拟机

- 规则组内虚拟机必须在同一主机上运行；
- 多台业务虚拟机，需求规划聚合，以及特殊业务激活需求绑定机器码等场景。

互斥虚拟机

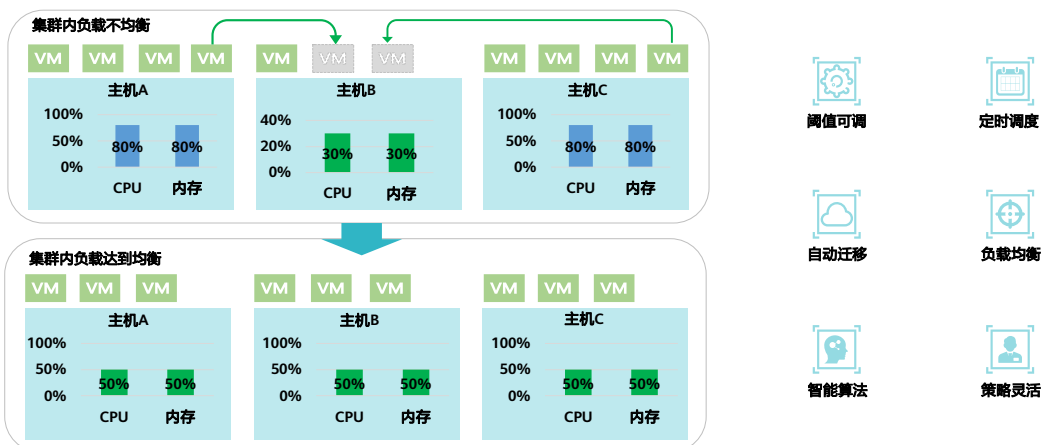
- 规则组内虚拟机必须在不同主机上运行；
- 多台关键业务虚拟机或前后台业务等为保证冗余可靠性，必须分布在不同主机，例如Oracle RAC或其他集群类业务。

虚拟机到主机

- 虚拟机组必须在主机组上运行；
- 虚拟机组应该在主机组上运行；
- 虚拟机组禁止在主机组上运行；
- 虚拟机组不应该在主机组上运行；
- 更多灵活虚拟机组对主机组的规则，包含强制，非强制关系。

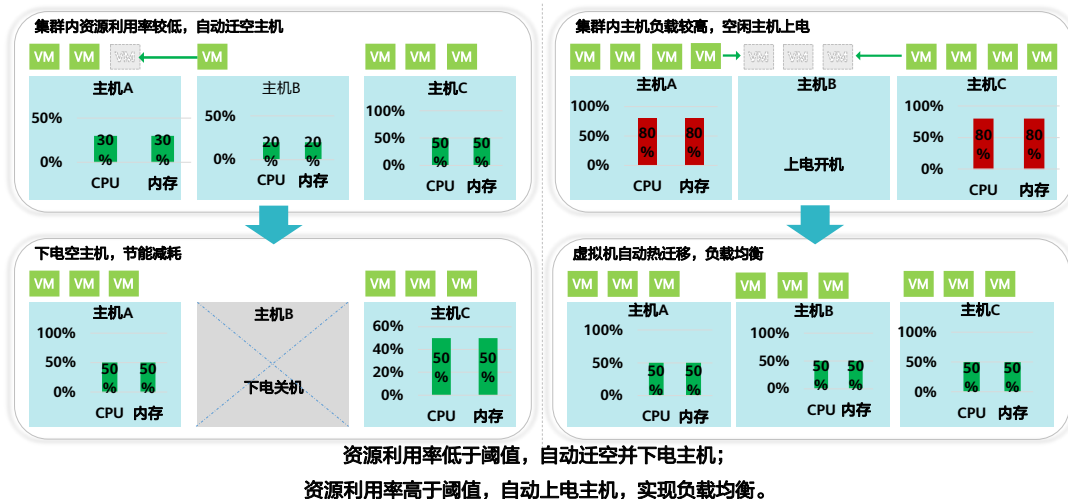
- 设定规则组可以控制虚拟机与虚拟机、虚拟机与主机之间的位置关系，以满足不同应用场景的需求。

动态资源调度



- DRS: Dynamic Resource Schedule，动态资源调度又称为计算资源调度自动化。指采用智能负载均衡调度算法，并结合动态电源管理功能，通过周期性检查同一集群资源内各个主机的负载情况，在不同的主机间迁移虚拟机，从而实现同一集群内不同主机间的负载均衡，并最大程度降低系统的功耗。
- FusionCompute可以在虚拟机创建和运行的过程中感知全局物理资源的使用情况，并通过智能调度算法，确定适合虚拟机运行的最佳主机，通过热迁移等手段将虚拟机运行在最佳主机之上，从而提升全局业务体验。
- FusionCompute的计算集群，配合基于VIMS文件系统的共享存储；DRS算法实时监控集群中每个计算节点的资源使用情况，引用VMOTION功能，智能迁移负荷高节点上虚拟机到资源充足的节点上；均衡各节点资源使用并保障业务有充足资源可用。因此，DRS是实现自动负载均衡的基础。
- 系统轻载时，将迁移部分虚拟机，并将其集中在部分物理主机，随即将空闲主机下电。
- 系统重载时，将启动部分物理主机，并将虚拟机均衡分布在各主机中，以保证资源的供应和用户的体验。
- 通过计划任务，可根据系统运行的情况，分时段采取不同的资源调度策略，以满足不同场景的用户需求。
- 虚拟机存在以下绑定关系时，动态资源调度对该虚拟机失效。
 - 绑定主机
 - 绑定USB设备
 - 绑定直通设备
 - 绑定GPU资源组

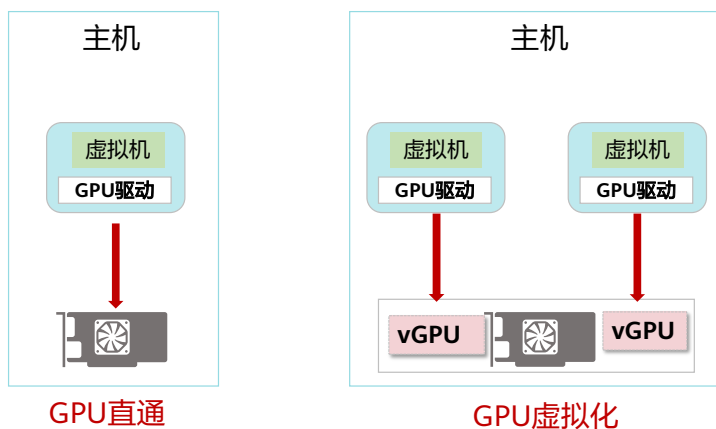
动态电源管理



- 动态电源管理，又称为电源管理自动化功能会周期性地检查集群中服务器的资源使用情况，如果集群中资源利用率不足，则会将多余的主机下电节能，下电前会将虚拟机迁移至其他主机；如果集群资源过度利用，则会将离线的主机上电，以增加集群资源，减轻主机的负荷。
- 开启电源管理自动化时，必须同时开启计算资源调度自动化。待主机上电后，便自动对虚拟机进行负载均衡。
- 电源管理的分时阈值设置，可以满足不同时段的调度需求。在业务平稳运行的时段，建议将电源管理自动化的级别调低，以减小对业务的影响。
- 电源管理自动化会检测集群的资源利用率，当资源利用率低于轻载阈值的持续时间超过下电主机评估历史时间（默认为40分钟）时，才会对主机进行下电操作。同样，当资源利用率高于重载阈值的持续时间超过上电主机评估历史时间（默认为5分钟）时，才会对主机进行上电操作。下电主机评估历史时间和上电主机评估历史时间支持在高级选项中进行自定义配置。
- 技术特点：
 - 系统自动选择合适的物理机上下电，减小迁移VM数量。
 - 保证小部分物理机处理休眠态，以快速满足新增业务所需资源。
- 使用场景：
 - 夜间低负载，自动迁移虚拟机，下电空闲主机。
 - 日间业务需求上升，自动上电主机，迁移虚拟机到新上电主机。

GPU直通&GPU虚拟化

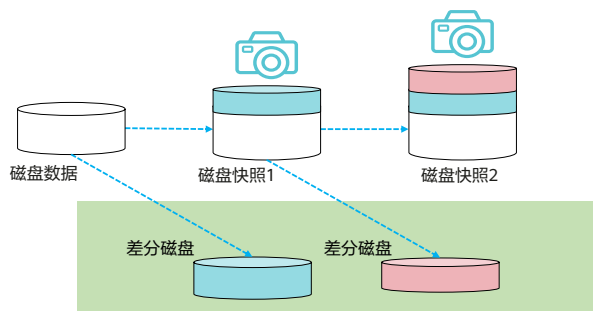
- FusionCompute支持GPU直通与GPU虚拟化，提升高性能图形图像应用体验。



- GPU直通是将物理服务器的GPU直接关联给虚拟机，虚拟机直接访问物理GPU设备，使虚拟机的图形视频处理能力接近物理GPU的性能。
- GPU虚拟化是指一个物理GPU在硬件的支持下可以创建出多个vGPU设备，从而实现被多个虚拟机共享，能同时向多个虚拟机提供2D图形处理和3D图形渲染的加速服务。GPU虚拟化主要满足GPU场景的中高端用户需求，既可以支持多用户同时共享一个GPU的能力又有着良好的性能，同时降低了用户的使用成本。

快照

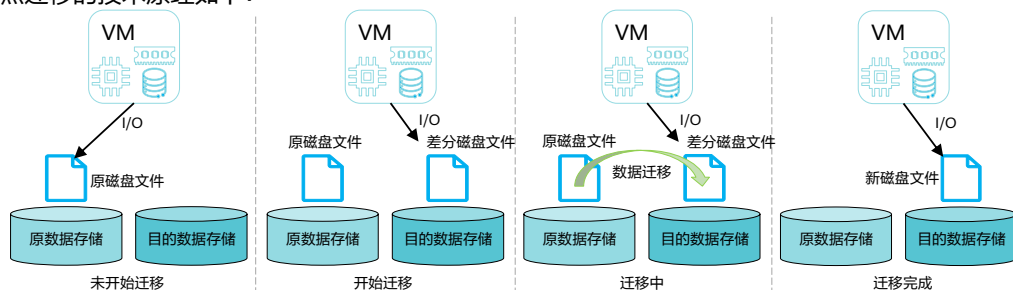
- 虚拟机可以将当前状态保存在快照文件中，包括磁盘内容、内存和寄存器数据。用户可以通过恢复快照多次回到这一状态，虚拟机用户在执行一些重大、高危操作前，例如系统补丁，升级，破坏性测试前执行快照，可以用于故障时的快速还原。
- FusionCompute支持普通快照、一致性快照以及内存快照。



- 创建快照会创建一个差分磁盘文件，该磁盘文件与源磁盘文件存放同一目录下。源磁盘文件会转为只读，新写入的数据存放在差分磁盘文件中。
- 回滚快照时会删除差分磁盘文件中的数据。
- 删除快照时，系统会整合原磁盘与差分磁盘文件里面的数据，形成新的磁盘文件。

存储热迁移

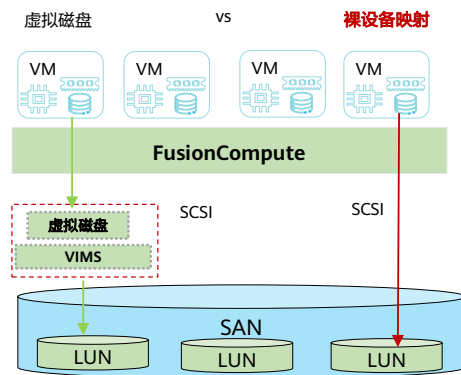
- 虚拟机正常运行时，管理员可通过手动操作，将虚拟机的磁盘迁移至其他存储单元。存储热迁移可在存储虚拟化管理下的同一个存储设备内、不同存储设备之间进行迁移。热迁移使客户在业务无损的情况下动态调整虚拟机存储资源，以实现设备维护等操作。
- 热迁移的技术原理如下：



- FusionCompute提供了虚拟机磁盘的冷迁移和热迁移，冷迁移是在虚拟机关机时候，将其磁盘文件从一个存储移动到另一个存储，热迁移可以在不中断业务的前提下，将虚拟机磁盘从一个存储迁移至另一个存储。
- 热迁移首先使用写时重定向，将虚拟机数据写入目的存储的一个差异磁盘，这样，原磁盘文件就变成只读的。
- 将源卷的所有数据块依次读取出来并合并到目标端的差异磁盘中，等数据合并完成后，目的端的差分磁盘就拥有虚拟磁盘的所有最新数据。
- 去除目的端快照对源卷的依赖，将差分磁盘修改为动态磁盘，这样，目的端磁盘文件可以独立运行。
- 依赖：
 - 不支持迁移已挂载的“共享”类型的磁盘。
 - 不支持非持久化磁盘的迁移。在条件允许的情况下，建议将虚拟机关闭后迁移。
 - 不支持跨华为分布式块存储资源之间进行迁移。

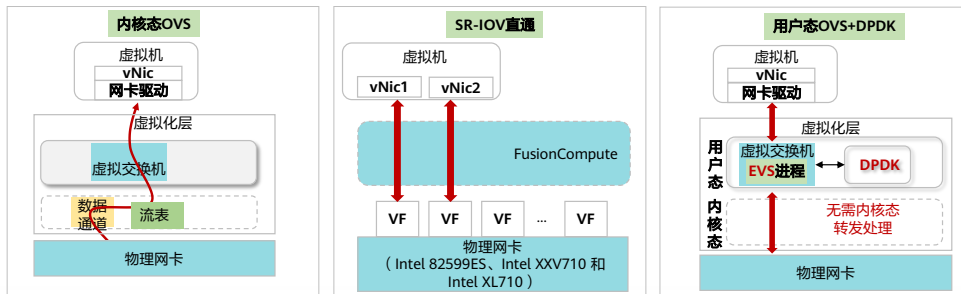
存储资源裸设备映射（RDM）

- RDM为虚拟机提供了一种机制来直接访问物理存储子系统（仅限光纤通道或iSCSI）上的LUN，通过使用物理设备映射，可以让虚拟机识别SCSI磁盘。



- Raw Device Mapping绕过Hypervisor，物理裸设备映射（RDM）能够将虚拟机下发的SCSI命令直接透传，使得虚拟机中的业务能够直接操作物理SCSI设备，避免由于虚拟化层的模拟导致功能丢失。
- 裸设备映射优势：缩短存储设备IO路径，提供逼近物理存储的IO性能，支持SCSI指令透传。
- 不支持存储瘦分配、磁盘在线/离线扩容、存储增量快照、iCache、存储热迁移、存储QoS、磁盘备份、虚拟机转为模板等。
- 技术特点：
 - 虚拟机直接通过SCSI命令操作裸存储设备。
 - 兼容FC光纤存储和IP SAN存储。
- 适用于需要高性能存储的应用，比如Oracle RAC，集群文件系统，视频图片处理等存储高IO，低延迟业务。

OVS+DPDK



结合OVS技术提供**IPMAC绑定、安全组**等高级网络特性，**网络QoS**控制更加精细；

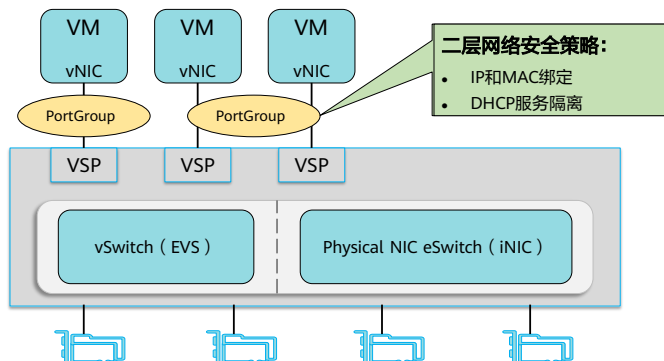
利用物理网卡提供的硬件虚拟化能力，将硬件虚拟设备挂载给VM，VM网络I/O性能、时延**逼近物理网卡**；

利用DPDK技术加速网络转发，并通过大页、用户态减少报文拷贝，**最大提升6倍网络**转发效率，PPS大幅提升。

- 通过使用DPDK（Data Plane Development Kit，数据平面开发套件，DPDK是一系列库和驱动的集合）技术，用来在x86平台进行快速的数据包处理。它通过环境抽象层旁路内核协议栈、轮询模式的报文无中断收发、优化内存/缓冲区/队列管理、基于网卡多队列和流识别的负载均衡等多项技术，实现了在x86处理器架构下的高性能报文转发能力，提高虚拟机网络性能。
- 用户态交换模式支持的网卡型号为Intel 82599ES、Intel XL710、Intel X710和Mellanox MT27712A0。
- 使用用户态交换模式时，需要先配置主机大页内存并配置主机的用户态交换机规格。
- SR-IOV实现了将PCI功能分配到多个虚拟接口以在虚拟化环境中共享一个PCI设备的资源。SR-IOV能够让网络传输绕过软件模拟层，直接分配到虚拟机。这样就降低了软件模拟层中的I/O开销。如果虚拟机使用了SRIOV直通虚拟网卡，则不能使用休眠、唤醒、在线迁移、完整迁移、内存快照、内存热添加、网卡在线添加与删除、IP MAC绑定、安全组功能。
- SRIOV直通模式支持型号为Intel 82599ES、Intel XXV710、Intel X710和Intel XL710的网卡，每个网口支持的最大虚拟网口个数为63。

二层网络安全策略

- 二层网络安全策略主要包括：防止用户虚拟机IP和MAC地址仿冒，防止用户虚拟机DHCP Server仿冒。



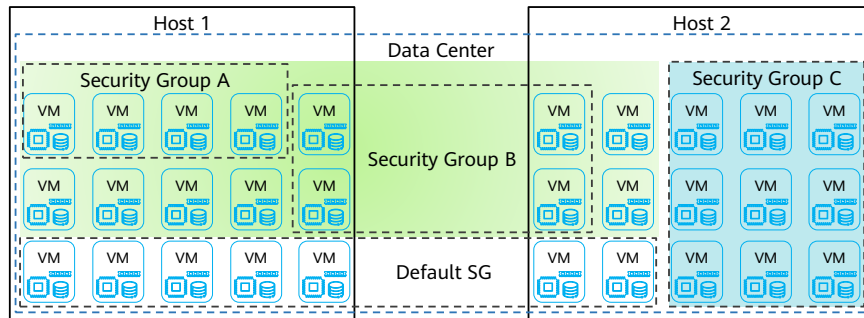
- 防止IP地址和MAC仿冒（IP和MAC绑定）：防止虚拟机用户通过修改虚拟网卡的IP、MAC地址发起IP、MAC仿冒攻击，增强用户虚拟机的网络安全。通过生成IP-MAC的绑定关系，基于IP源侧防护（IP Source Guard）与动态ARP检测（DAI）对非绑定关系的报文进行过滤。
- 防止DHCP Server仿冒（DHCP Server隔离）：禁止用户虚拟机启动DHCP Server服务，防止用户无意识或恶意启动DHCP Server服务，影响正常的虚拟机IP地址分配过程。

广播报文抑制

- 在服务器整合、桌面云等企业应用场景，如果发生网络攻击或病毒发作等引起的广播报文攻击，可能造成网络通信异常，此时可以开启虚拟交换机的广播报文抑制功能。
- 虚拟交换机提供虚拟机虚端口发送方向的广播报文抑制开关，以及抑制阈值设置功能。可以通过开启虚拟机网卡所在端口组的广播报文抑制开关设置阈值，减少过量广播报文对二层网络带宽的消耗。
- 管理员可以通过系统Portal，基于虚拟交换机端口组对象，配置报文抑制开关和报文抑制带宽阈值。

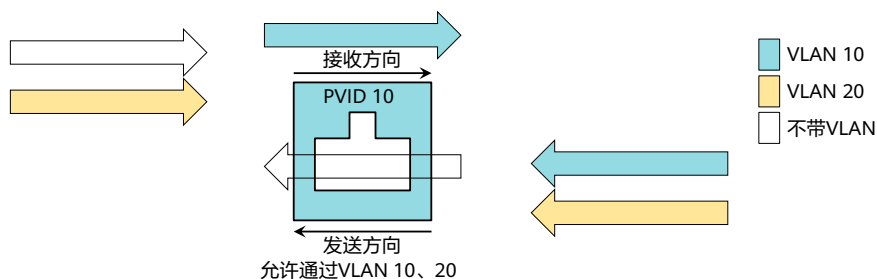
安全组

- 用户根据虚拟机安全需求创建安全组，每个安全组可以设定一组访问规则。当虚拟机加入安全组后，即受到该访问规则组的保护。用户通过在创建虚拟机时选定要加入的安全组来对自身的虚拟机进行安全隔离和访问控制。



- 位于同一个安全组的所有虚拟机网卡都将使用该安全组规则进行网络通信。虚拟机一块网卡只能加入一个安全组中。

Trunk口

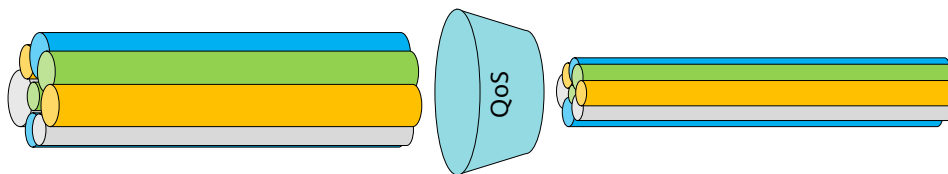


Trunk口介绍

- 虚拟机网卡通过虚端口接入虚拟交换机进行网络数据包的收发；
- 虚拟交换机虚端口支持配置为Trunk类型，并允许设置Trunk的VLAN ID范围，之后虚端口便具备了同时收发携带不同VLAN标签的网络数据包的功能，从而满足了虚拟网卡支持Trunk类型端口的需求。

- 普通类型的虚端口只能属于一个VLAN，中继类型的虚端口可以允许多个VLAN接收和发送报文。普通虚拟机选择普通类型的端口，虚拟机的网卡启用VLAN设备的情况下选择中继类型的端口，否则虚拟机的网络可能不通。
- 端口组配置为中继的方式后，可以在Linux虚拟机内创建多个VLAN设备，这些VLAN设备通过1个虚拟网卡即可以收发携带不同VLAN标签的网络数据包。使虚拟机不用创建多个虚拟网卡，即可收发携带不同VLAN标签的网络数据包。

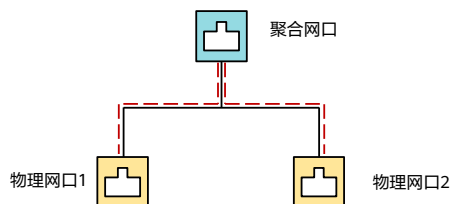
网络QoS



网络QoS策略提供带宽配置控制能力，包含如下方面：

- 基于端口组成员接口发送方向与接收方向的带宽控制；
- 基于端口组的每个成员接口提供流量整形、带宽优先级的控制能力。

网口绑定



主机网口绑定

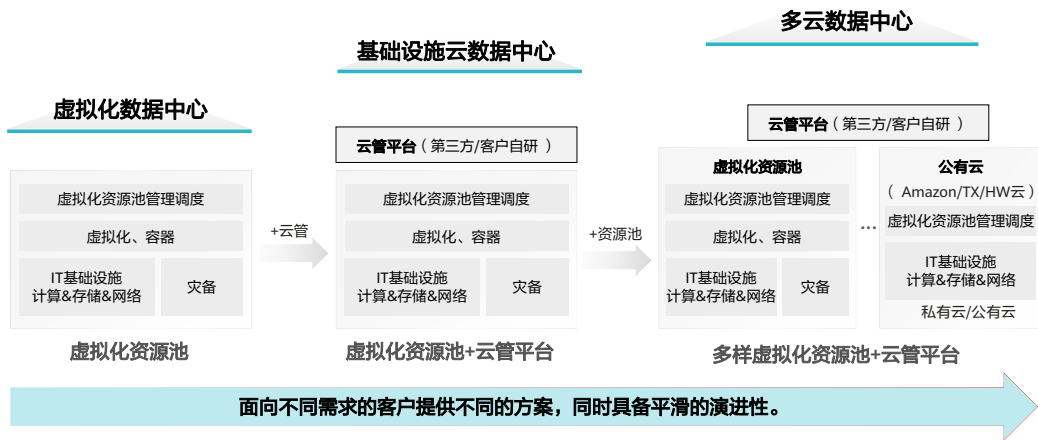
- 管理员可以通过FusionCompute绑定CNA主机的网口，以提高网络的可靠性；
- 针对普通网卡和DPDK驱动的物理网卡均可以设置端口绑定。

- 对于普通网卡，绑定模式可以选择以下几种：
 - 主备
 - 轮询
 - 基于源目的IP和端口负荷分担
 - 基于源目的MAC的负荷分担
 - 基于源目的MAC的LACP
 - 基于源目的IP的LACP
- 对于支持DPDK驱动的物理网卡，绑定模式可以选择以下几种：
 - DPDK驱动的主备模式
 - DPDK驱动的基于源目的MAC的LACP模式
 - DPDK驱动的基于源目的IP和端口的LACP模式

目录

1. 虚拟化技术概览
2. 典型虚拟化软件介绍
3. 华为虚拟化软件FusionCompute
- 4. 数据中心虚拟化DCS**

数据中心虚拟化和云将长期共存



DCS全栈数据中心解决方案组件介绍



92 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



• 关键组件介绍

- DME：智能资源管理软件，实现对资源池的接入、管理、E2E资源编排发放和E2E智能运维等功能
- eSphere：虚拟化平台，提供计算、存储、备份、容灾等能力，形成弹性资源池，实现资源自动编排调度和管理
- eDataInsight：大数据平台，围绕数据“采、存、算、管、用”，构建轻量灵活、开放的大数据平台
- eCampusCore：统一南向集成接入和数据处理，通过全面服务能力开放，支撑业务应用快速构建
- ModelEngine：AI全流程工具链，提供行业大模型的开发、训练、微调、推理的全流程服务能力
- MigrationDirector：虚拟机迁移软件，实现异构平台向DCS迁移能力

• DCS框架特点

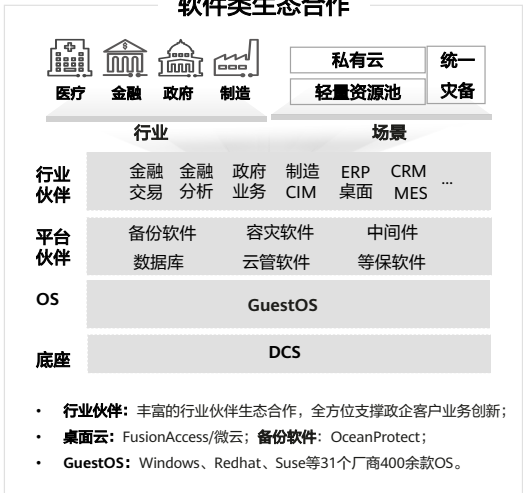
- 轻量：最低2节点资源池建设
- 积木式：灵活叠加容器、备份、管理等组件
- 开放联接：南向兼容，北向开放

软硬件生态

硬件类兼容合作



软件类生态合作



虚拟机和容器双栈共池



eSphere双栈资源池（虚拟化&容器）



虚拟机



容器

轻量高效

- 敏捷弹性：2节点起步，一键集群发放和管理、支持单集群128节点
- 自研容器引擎iSulad，性能佳

企业级增强

- 安全增强：开源漏洞管理、证书轮转、密钥加密、镜像仓加密
- 升级增强：跨版本升级、批量节点升级、不影响业务
- 应用一致性备份：PV的快照和克隆，应用一致性备份（含应用配置，PV卷数据）和恢复

开放灵活

- 开放：支持标准Chart和Yaml应用模板，支持部署自定义组件
- 灵活：应用部署方式多样，命令行部署、上传部署，Git自动同步

思考题

1. （判断题）全虚拟化使用VMM实现CPU和内存虚拟化，设备I/O虚拟化由Guest OS实现。需要修改Guest OS，使其能够与VMM协同工作。该方式兼容性差，但性能较好。
2. （判断题）Libvirt是Linux上的虚拟化库，主要目标是提供一种通用并且稳定的软件层，来管理物理主机上多种不同的虚拟化方式和虚拟主机，并支持远程管理。

- 参考答案：

- 错误
- 正确

本章总结

- 本章我们一起学习了虚拟化技术的本质与价值，了解了主流的虚拟化技术有哪些，掌握了主流虚拟化技术基本原理。

更多信息

- 华为数据存储信息中心: <https://info.support.huawei.com/storage/#/home>
- 企业业务: <https://e.huawei.com/cn/>
- 技术支持: <https://support.huawei.com/enterprise/>
- 在线学习: <https://e.huawei.com/cn/talent/portal/#/>
- 华为Support案例库: <https://support.huawei.com/enterprise/zh/knowledge?lang=zh>

Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。
Bring digital to every person, home, and
organization for a fully connected,
intelligent world.

版权所有 © 华为技术有限公司 2026。

保留一切权利。

未经华为事先书面许可，任何人不得对本资料进行复制、修改、
整合、改编、翻译，也不得将本资料或其任何部分或基于本资
料的衍生作品提供、出售、转让、再许可给任何人用于华为
许可以外的任何其他用途。

Copyright © Huawei Technologies Co., Ltd. 2026.

All Rights Reserved.

Without prior written permission of Huawei, no one is
allowed to reproduce, modify, integrate, adapt, translate
this Document, or provide, sell, transfer, sublicense this
Document or any part of this Document or its derivative to
any third party for any other purpose otherwise permitted
by Huawei.



存储系统日常运维



前言

- 本章主要讲述IT运维的相关概念，存储系统运维工具，日常管理和系统升级，通过本章的学习，可以了解到存储运维管理工具的使用和存储系统的日常管理。

目标

- 学完本课程后，您将能够：
 - 掌握存储运维管理
 - 掌握存储日常管理操作

目录

1. 存储智能运维
2. 存储运维工具
3. 存储日常管理
4. 存储系统升级

IT运维的定义

第三方机构对IT运维的定义

与IT服务管理相关的人员及管理流程，其目的是将具有成本与质量要求的服务交付给用户

Gartner

由组织的IT部门提供，同时面向组织内外部用户的一系列流程及服务，其中包括了对硬软件的管理和运行维护

TechTarget

IT服务全生命周期中的一个阶段，通过对IT服务与IT基础设施进行监控，实现备份恢复与作业调度等活动

ITIL

《企业IT运维发展白皮书》的定义

**理论
角度**

IT运维是指以组织的内、外部用户需求为导向，通过一系列流程、技术、方法，确保为用户提供的IT服务或产品符合一定要求。

**操作
角度**

通过运用互联网、云计算、大数据、人工智能等新型信息通信技术，通过监控、作业调度、备份与恢复等手段，结合管理流程手段，维护生产环境以及和生产环境相关的基础设施，包括硬件设备、基础软件、网络等，以保障生产环境稳定、高效、安全、低成本运行。

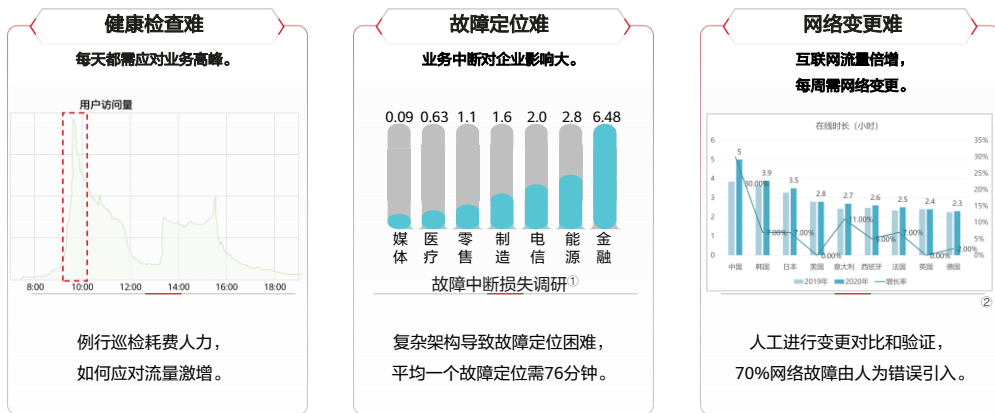
- IT运维（IT Operations）的定义非常广泛，因其在不同组织中的不同阶段代表着不同的内涵。

IT运维发展历程



- 从企业信息系统规模、复杂程度变化及对信息通信技术应用的深入等方面考虑，IT运维的发展可以分为以下五个阶段：
 - 第一阶段为手工运维阶段，依赖个人知识、技术、经验解决信息系统问题。
 - 第二阶段为流程化、标准化运维阶段，看重流程说明、标准规范等文档的建立与管理。
 - 第三阶段为平台化、自动化运维阶段，聚焦企业运维平台建设，具备自动化运维能力。
 - 第四阶段为DevOps（研发运营一体化）阶段，助力企业实现软件生命周期的全链路打通，持续运营与优化。
 - 第五阶段为AIOps（智能运维）阶段，尝试将AI技术及海量数据应用于运维场景。

传统运维面临的挑战



- 应用动态随迁，流量激增，数据中心运维需要智能化。
 - 数据中心规模不断扩大，故障域也随之扩大。
 - 虚拟网络延伸到服务器内部（如vSwitch），网络 and IT 的运维边界越来越模糊。
 - 网络动态感知 VM 动态迁移及应用的弹性扩缩，配置变化频繁，流量激增，数据中心内应用策略及互访关系日益复杂，传统运维手段实施起来难度越来越大。
 - 为了提高用户体验，保证关键业务的高可靠性，需要对故障进行实时定位和消除。
- 数据来源^①： *Network Computing, the Meta Group and Contingency Planning Research*
- 数据来源^②： *APP ANNIE*

数字化转型面临的挑战



- 数据中心作为企业数字化的新基建底座以及数据智能的萃取中心，其在数字时代的重要性不言而喻。某种程度上来说，数据中心对创新业务的支撑力度、高效运营决定了企业数字化的程度。

企业IT设施运维面临的挑战

- 数字化全面兴起，对IT设施稳定运行要求越来越高，设备数量和种类呈指数级增加，IT运维团队工作压力和挑战
- 当前IT行业还是以被动式运维为主，待问题发生后依赖告警等才能感知，业务易受影响



被动式运维

针对硬盘故障、容量或性能瓶颈等，如不能提前识别，只能被动处理、**业务易受影响**



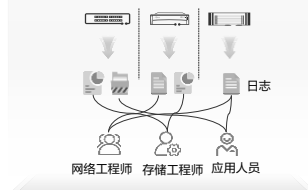
问题响应慢

传统模式客户发现问题建单、技术伙伴处理，复杂问题需要厂家参与，**响应处理时间长**



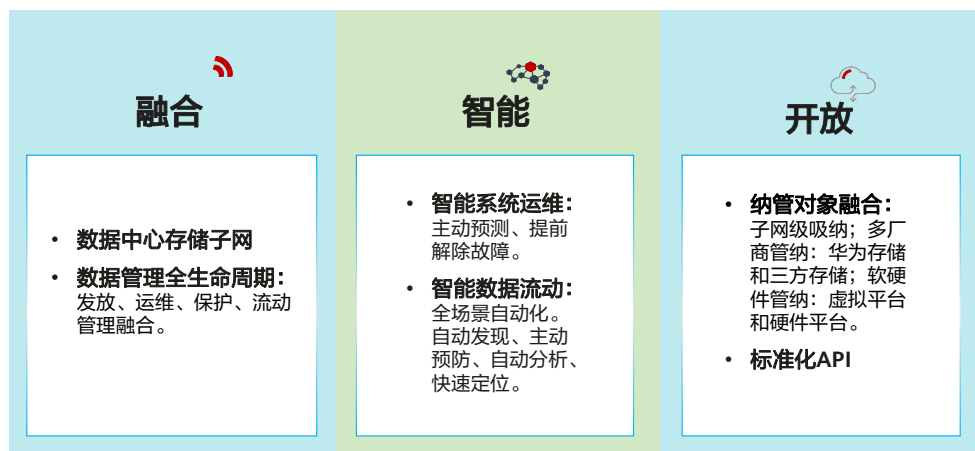
问题定界难

性能定位，需计算、网络、存储**多领域分析**日志，平均闭环时间4h，**周期长，定位难**

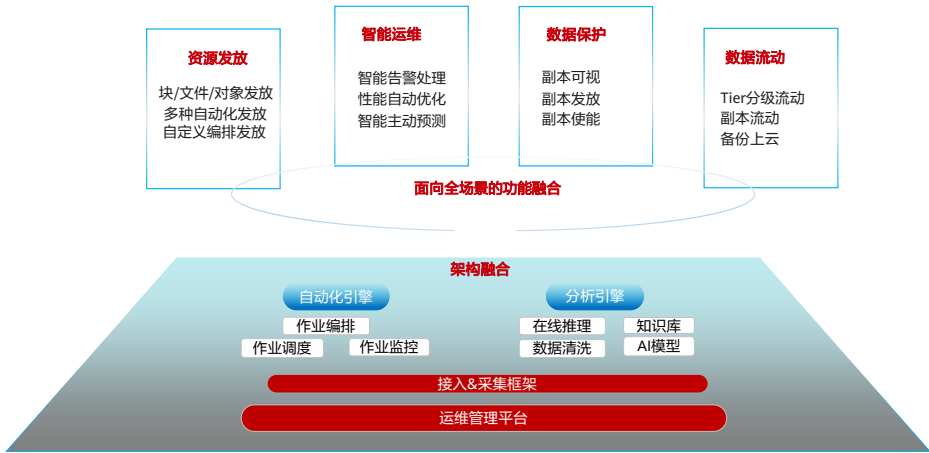


- 近年来，IT技术革命和软硬件创新，为企业数据中心的集约化和高效率运营提供了有力的支持。IT数据是企业信息化的重要核心，但设备规模越来越大，设备种类繁多，运维难度越来越高，一旦出现问题，将对企业带来巨大影响和损失。

数据存储管理与运维趋势

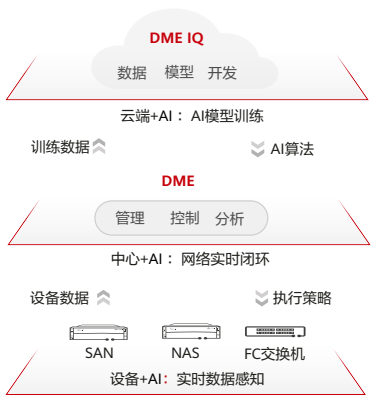


数据存储管理与运维趋势 - 融合

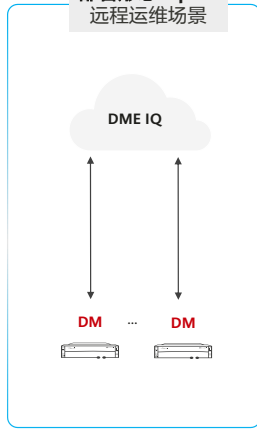


数据存储管理与运维趋势 – 智能 (1)

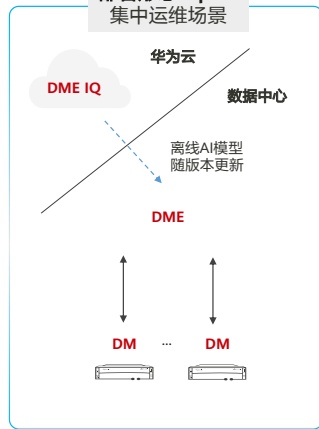
三层 AI 实现架构



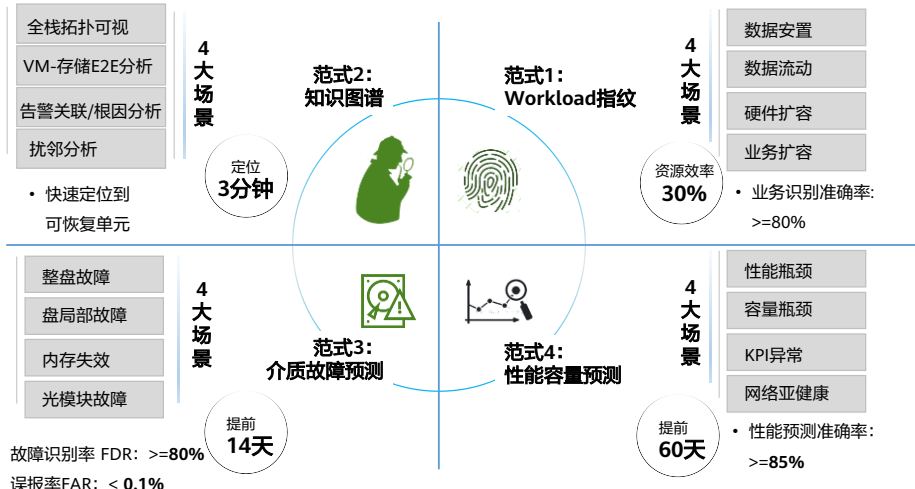
部署形态 Op. 1 远程运维场景



部署形态 Op. 2 集中运维场景



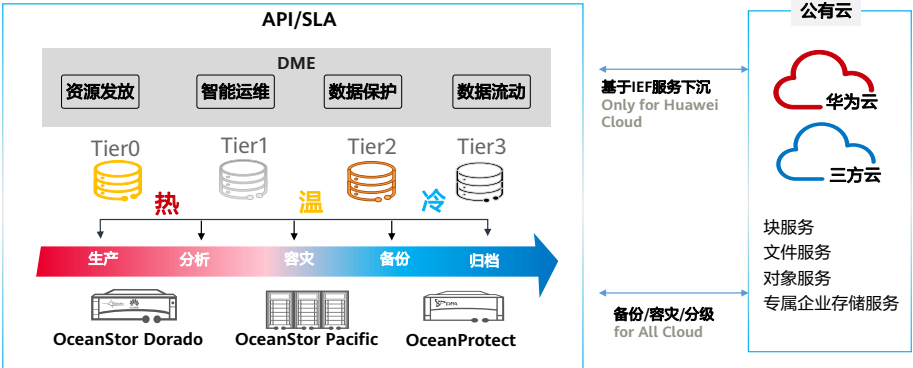
数据存储管理与运维趋势 – 智能 (2)



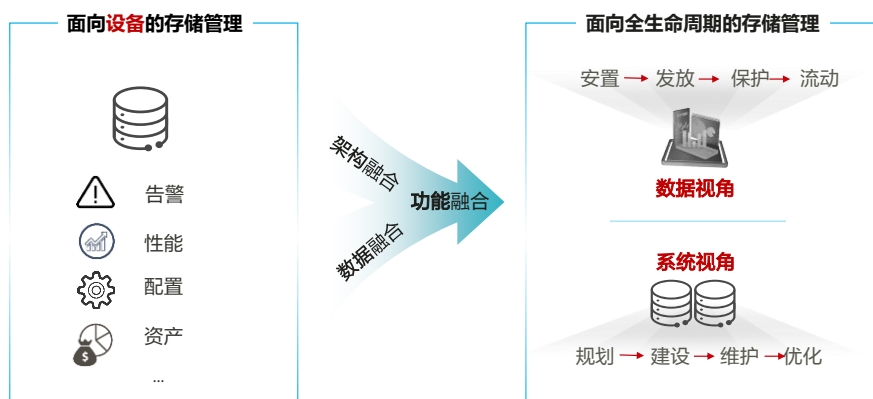
数据存储管理与运维趋势 – 开放 (1)



数据存储管理与运维趋势 – 开放 (2)

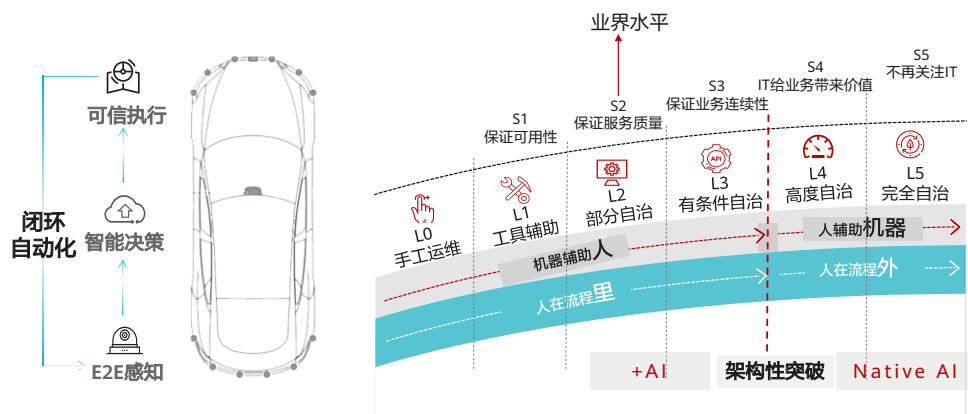


从面向设备到面向全场景



- 数据是智能时代最重要的资产已经成为共识，而数据中心存储作为数据的载体，同样也是数据基础设施中最关键的设备，存储管理和运维的重要性不言而喻。
- 从存储的管理员视角来看，要实现运维效率的质的飞跃，需要还原到存储的全生命周期场景：
 - 从基础设施的视角来看，包括规划，建设，维护和优化几个环节
 - 从数据的视角来看，包括数据的安置，发放，保护和流动等环节。
- 因此，存储的管理和运维走向数据、架构以及功能的融合，是实现全生命周期场景自动化的一个基本前提。

从自动到自治

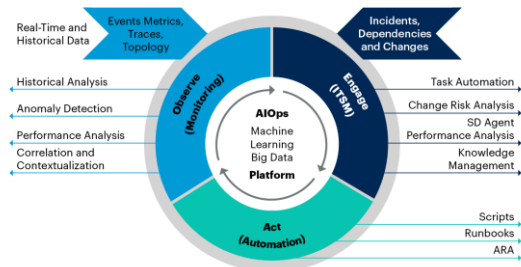


- 自动化的下一步是什么？是自治。英文是Autonomous；自动化和自治的核心区别在于在场景中是否具备自我决策的能力。
- 也就是在一个特定或者更大范围的场景中，要实现大家说的Zero touch，也就是无人工干预。打个比方，高级别的自动驾驶（L3/L4）要实现机器对人的替代，基于机器要实E2E的感知，智能的决策和可信的执行，三个能力（眼 脑 手）有机结合，才有可能实现。
- 华为致力于实现数据存储运维的自动驾驶，Gartner 把IT运维分为S1-S5五个阶段，这与华为倡导的自动驾驶的level基本是呼应的。

智能运维

Gartner从市场角度定义了AIOps平台

AIOps Platform Enabling Continuous Insights Across IT Operations Management (ITOM)



Source: Gartner
756577_0

Gartner

AIOps (Artificial Intelligence for IT Operations)

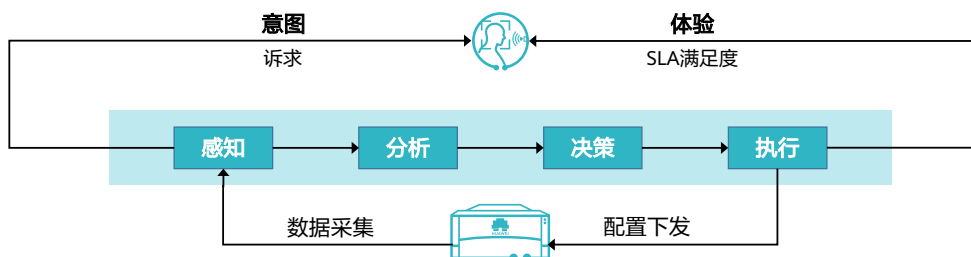
《企业级AIOps实施建议》梳理了AIOps主要场景



高效运维社区，《企业级AIOps实施建议》，2018

- Gartner在2021年修订的《AIOps平台市场指南》中给出了AIOps平台的定义：AIOps平台结合数据存储与数据分析能力，满足系统运维支持需求，并基于系统数据为相关团队提供洞见。
- AIOps的主要任务包括：异常检测、诊断信息、根因分析用于提升监测、服务管理和自动化任务
- 2018年高效运维社区、AIOps标准工作组牵头发布了《企业级AIOps实施建议》白皮书，参与方包括华为、清华、阿里、腾讯、百度、360、平安等
- 白皮书给出了AIOps定义：将人工智能应用于运维领域，基于已有的运维数据（日志、监控信息、应用信息等），通过机器学习的方式来进一步解决自动化运维没办法解决的问题。
- 白皮书梳理了AIOps常见应用场景，主要包括：效率提升、质量保障、成本管理等3项主要目标，每个主要目标又分为4个小场景，每个小场景又包含若干具体问题。

存储智能运维演进模型

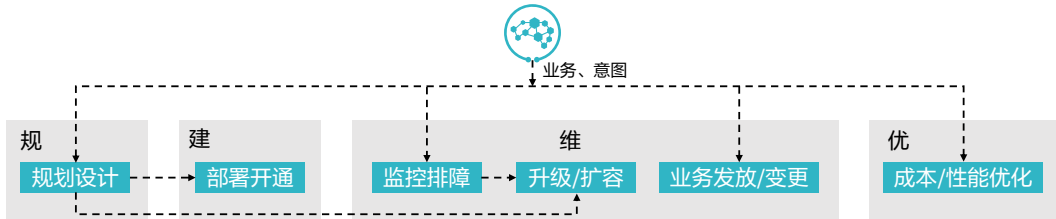


L1~L5演进模型

L1（工具辅助操作）		L2（自动化操作+辅助分析）		L3（特定场景自治）		L4（全场景自治）		L5（无人值守完全自治）	
执行	人+系统	执行	系统	执行	系统	执行	系统	执行	系统
监控	人	监控	人+系统	监控	系统	监控	系统	监控	系统
分析	人	分析	人+系统	分析	人+系统	分析	系统	分析	系统
决策	人	决策	人	决策	人+系统	决策	系统	决策	系统
意图/体验	人	意图/体验	人	意图/体验	人	意图/体验	人+系统	意图/体验	系统

- 参照自动驾驶网络2.0的分级体系，同时考虑到系统分析能力演进所需的时间较长，因此在“分析”环节从L2阶段开始加入系统辅助的能力。
- 意图/体验：人对存储的诉求，包括商业诉求（客户）、业务意图（存储资源管理员）、维护意图（维护人员）；系统对意图的接收和翻译；体验指SLA满足度。
- 感知：在意图翻译的基础上，监测存储运行状态，收集、处理监测对象及其外部数据，输出分析所需的数据。
- 分析：在感知阶段输出的数据基础上，通过运用模型推理和分析等技术，输出备选方案。
- 决策：在分析阶段输出的备选方案基础上，通过仿真验证等评估活动，确定优选方案。
- 执行：根据决策阶段确定的优选方案，进行方案实施，并返回结果信息给意图方。
- 表中的“系统”泛指工具和系统：
 - 工具：离线、单次使用。
 - 系统：在线、常驻。如产品自身、集中管理系统和云端系统。系统可以积攒历史数据、向智能化演进，系统间可以联动，使流程拉通。

存储智能运维场景定义



运维场景		定义
规	规划设计	根据客户商业意图、业务发展目标，并结合现网勘察和分析洞察情况，输出存储规划方案(HLD)和硬件/网络/业务配置等详细设计方案(LLD)。包括针对新建/升级/扩容的规划设计。
建	部署开通	新建场景，根据详细设计方案完成硬件安装、软件部署、网络配置和业务调测，并通过验收、达到转维条件。
维	监控排障	根据客户运维策略(如重保)对存储设备健康状态进行监测，及时发现故障或隐患、进行故障界定定位、分析问题根因并修复故障或消除隐患。
	升级/扩容	根据监控排障或规划设计(客户业务发展需要)产生的升级/扩容请求，分析对客户业务的影响，输出变更方案并实施变更，满足监控排障或客户业务发展诉求。
	业务发放/变更	根据客户对存储资源的需求(包括容量/性能/保护等方面)，下发业务配置，完成业务开通/变更。
优	性能/成本优化	根据对存储性能情况、资源利用情况的监控和分析，制定优化方案并实施，满足客户对性能/成本优化的意图。

从AIOps走向AgenticOps



- 传统的人工智能驱动的IT运维（AIOps, AI-driven IT operations）显著提升了大型网络的可视性和分析能力。通过利用遥测数据和机器学习技术，AIOps平台能够检测异常、关联事件，并向操作人员提供建议。然而，尽管取得了这些进步，AIOps本质上仍然是一种“人在回路”（human-in-the-loop）的范式：AI系统仅停留在告警和诊断阶段，而决策和执行仍由人类操作负责。
- AgenticOps代表了超越这一模式的质的飞跃。它不仅辅助人类操作，还引入了能够自主推理操作目标、规划行动并在明确约束下自主执行决策的AI agent。在这一范式下，AI系统不仅能够解读网络状态，还能以闭环方式对其采取行动，从而实现更快的问题预防、减少操作工作量，并在跨领域中实现持续的系统级优化。这一转变并非突然发生，而是操作自动化长期演进的结果，大致可以分为五个阶段：
 - 基于规则和脚本驱动的自动化（1980年代至1990年代）：早期的自动化依赖于人工编写的规则和脚本，通常以预定义的if-else逻辑形式表达。这些系统在处理重复性任务时效果显著，但缺乏适应性和上下文感知能力。
 - 基于工作流和策略的编排（1990年代至2000年代）：自动化逐渐向工作流引擎和策略驱动的编排框架发展，实现了更复杂任务序列和流程的自动化。虽然可扩展性有所提升，但决策逻辑仍由人类静态地显式编码。
 - 机器学习赋能的AIOps（2000年代至2010年代）。机器学习的引入带来了异常检测、模式识别和根因关联等能力。AIOps系统现在能够从统计角度“理解”网络行为，但其作用仍主要局限于检测和建议，而非自主行动。
 - 生成式AI（Generative AI）辅助的运维（2010年代末至2020年代初）：大型语

言模型实现了自然语言交互、解释和运维洞察的总结。然而，大多数生成式AI应用仍处于被动和单代理状态，仅能对查询做出响应，而没有持续的目标或执行权限。

- 代理式AI与AgenticOps（2020年代至今）：代理式AI的出现标志着向目标导向、自主系统的根本性转变。AI代理能够进行多步推理、任务优先级排序、调用工具和API，并与其它代理或人类协调以实现操作目标。在AgenticOps中，这些代理在治理和安全约束下运行，实现了从感知到行动的闭环。
- 从这个角度来看，AgenticOps不仅仅是AIOps的渐进式增强，而是一种全新的操作范式——它将AI从辅助组件转变为可问责、目标驱动的操作实体。

目录

1. 存储智能运维
2. **存储运维工具**
 - **SmartKit**
 - iMasterCloud DME IQ
 - iMaster DME
3. 存储日常管理
4. 存储系统升级

SmartKit简介



1、统一平台

桌面工具管理平台，集成存储、存储服务器、服务器、虚拟化以及解决方案四大领域的运维工具



2、场景化引导

按运维场景，一站式下载所需工具



3、标准化作业

向导化引导，工程场景简单、智能化

① 5大领域运维

- 存储
- 存储服务器
- 服务器
- 虚拟化
- 解决方案

② 7大运维场景

- 开局交付
- 例行维护
- 故障处理
- 升级/补丁
- 扩容
- 迁移
- 部件更换

③ 设备共享

- 工具箱一次添加，多场景共享
- 模板批量添加设备

- SmartKit为存储、存储服务器、服务器、虚拟化、解决方案五大领域的产品提供了统一的服务工具平台，同时也支持原OceanStor Toolkit的所有功能。

场景任务说明

- SmartKit工具箱集成了存储、服务器、虚拟化、云计算和机器视觉领域的场景任务，五大领域下的部分场景功能名称及描述参见下表：

场景	任务描述
开局交付	开局交付场景中，您可以根据流程指引对设备进行开局巡检操作。
例行维护	在例行维护场景中，您可以根据流程指引对设备进行日常检查、维护等操作。
升级/补丁	在升级/补丁场景中，您可以根据流程指引对设备进行升级和打补丁等操作。
故障处理	故障处理场景针对产品常见问题和故障，梳理对应的现象和排查逻辑，帮助您快速甄别和处理问题。
扩容	扩容场景可以帮助您对设备进行扩容操作。
迁移	迁移场景可以帮助您完成对数据的迁移操作。
部件更换	在部件更换场景中，您可以根据流程指引对设备进行更换部件的操作。

闪存存储服务场景（1）

场景	关键功能	功能描述
开局交付	开局工程质检	对设备硬件、软件状态和告警等开局质检标准检查。
	开局兼容性评估	收集存储和主机兼容性信息，支持在兼容性网站上完成兼容性评估。
例行维护	健康检查	日常巡检、信息收集和硬盘健康分析。
故障处理	存储信息收集	收集存储设备日志、配置信息、硬盘Smart信息、硬盘日志等。
	历史性能监控	收集历史性能数据。
	主机信息采集	收集主机信息，用于升级、迁移兼容性评估。
	故障辅助诊断	深度分析设备故障，并提供修复手段。
升级/补丁	升级前工勘	通过巡检、升级前评估或兼容性评估，提前识别升级风险。
	设备升级	通过传包、升级前检查、升级和升级后检查等流程，快速完成存储软件升级。
	补丁评估	补丁安装前风险检查。
	热补丁安装	一键存储补丁安装。
	硬盘固件升级	批量升级硬盘固件。

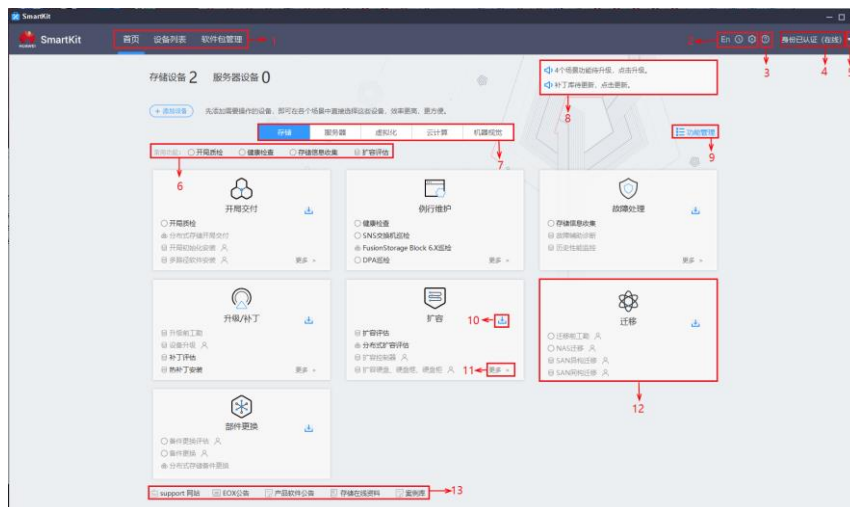
闪存存储服务场景（2）

场景	关键功能	功能描述
扩容	扩容评估	扩容前风险评估。
	扩容控制器	向导式扩容控制器，简化和标准化操作。
	扩容硬盘、硬盘框、硬盘柜	向导式扩容扩容硬盘、硬盘框、硬盘柜，简化和标准化操作。
备件更换	备件更换评估	备件更换前风险评估（控制器，PCIE接口卡，IP Scale Out接口卡）。
	备件更换	向导式备件更换。
迁移	NAS迁移	支持CIFS，NFS文件系统间的数据迁移。
	SAN异构迁移	第三方存储（EMC，HDS）到换完存储的一键式Lun迁移。
	SAN同构迁移	华为存储间的在线一键式迁移。

分布式存储服务场景

场景	关键功能	功能描述
开局交付	开局工程质检	检查进程、配置和资源状态，在开局阶段识别风险。
	分布式部署助手	导入LLD，配置iBMC IP，配套检查和存储前后端网络配置。
例行维护	健康检查	检查设备进程、配置和资源状态。
故障处理	存储信息收集	收集日志、配置信息等，支持按节点、时段收集。
	故障辅助诊断	深度分析设备故障，并提供修复手段。
升级/补丁	升级前巡检	提前识别升级风险。
	分布式存储计算节点升级	独立批量升级计算节点。
	分布式存储固件升级	支持硬盘固件、iBMC、BIOS、RAID卡等固件批量升级。
扩容	扩容评估	扩容前风险评估。
备件更换	备件更换	向导式支持部件更换。（主存盘、缓存盘）

SmartKit首页界面



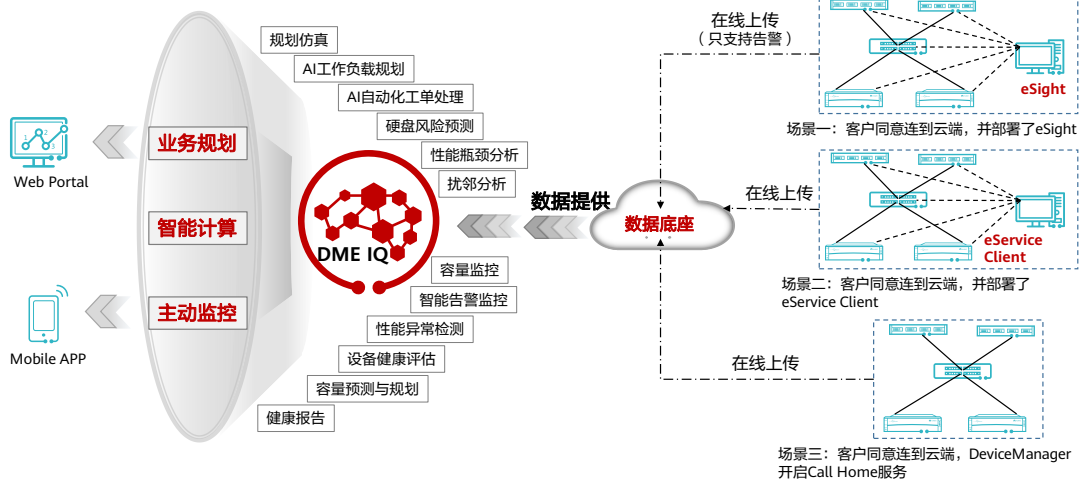
- 由于SmartKit中包含的工具数量多且较零散，导致技术服务人员任务执行成本高、效率低、难度大，所以将服务日常的活动场景化、流程化，在“首页”中使用统一的入口引导服务人员按照流程一步一步完成日常工作，提高工作效率。

#

目录

1. 存储智能运维
2. **存储运维工具**
 - SmartKit
 - **iMasterCloud DME IQ**
 - iMaster DME
3. 存储日常管理
4. 存储系统升级

iMasterCloud DME IQ



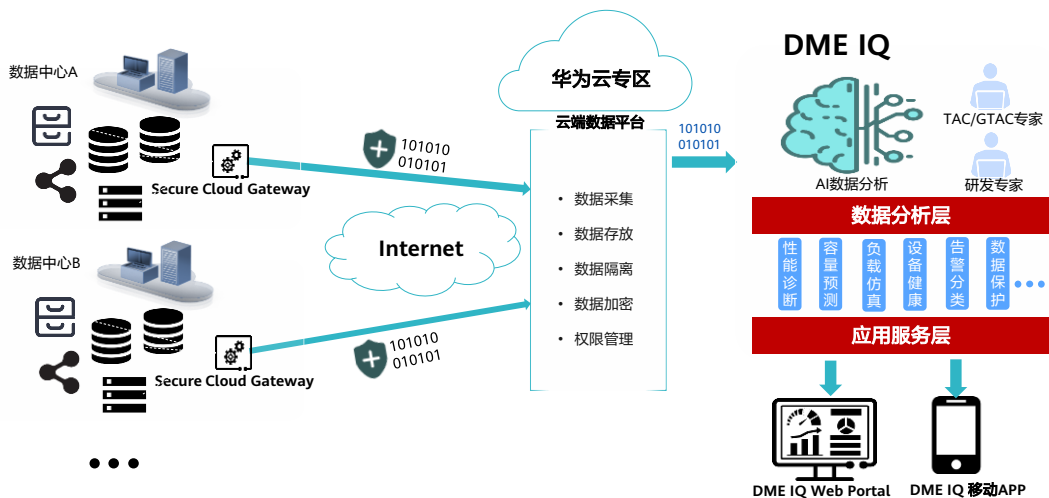
- DME IQ定位云端智能运维平台，基于云端基础底座提供的大量数据，通过大数据分析和人工智能AI技术，驱动IT运维活动的开展，降低运维难度、全面提升运维效率。同时，DME IQ还支持手机APP端运维方式，可以随时随地查看设备运行状态，方便快捷。
- DME IQ云端：IT存储集中运维的能力扩展，云端管理，提供智能运维、规划仿真、预测预防等特性。
- eService安全接入服务：作为云端运维的基础数据底座，能够收集IT存储设备的告警、性能、配置、系统日志、硬盘信息等数据，并提供远程维护和存量经营能力。同时，eService安全接入服务提供的大量运维数据，是DME IQ业务孵化和演进的核心基础。
- eService Client：eService Client支持上传告警、性能、配置、系统日志、硬盘信息至DME IQ云端，需提供物理机、虚拟机或SVP安装eService Client软件。
- DeviceManager：存储DeviceManager开启Call Home服务后，支持上传告警、性能、配置、系统日志、硬盘信息至eService安全接入服务，无需额外提供物理机或虚拟机安装eService Client软件。
- eSight：eSight目前只支持邮件上传告警至eService安全接入服务，需额外提供物理机或虚拟机安装eSight软件。

应用场景



- DME IQ运行在华为云，利用机器学习（Machine Learning）和深度学习（Deep learning）等AI技术进行大数据分析，主动帮助客户设备监控，分析和规划华为存储、服务器等IT基础设施。实现随时随地主动式、预见式运维，降低运维成本，提升运维效率。为客户提供一站式运维解决方案。

工作流程



- 通过DME IQ Client的监控设备和Support账号关联，提供自助式智能运维系统，为用户提供精准化的贴身信息服务。

实时统一管理

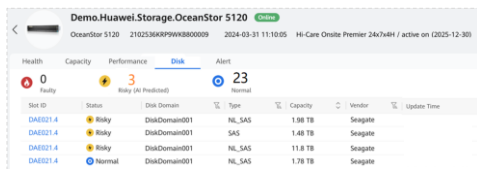


- 移动APP打造无所不在的智能运维体验。
- 问题单自动创建并处理。
- 设备全面健康度AI评分，确保系统高可用。

风险预测

风险盘识别

提前30天识别硬盘风险，主动预防
(HDD支持提前30天, SSD提前14天)



容量预测

预测12个月业务未来对容量的诉求，提前适配业务对存储系统的需求



性能预测

预测未来2个月的性能趋势，提前预知性能瓶颈



硬盘故障预测

- 硬盘是存储阵列的基础组成，虽然当今存储阵列广泛使用各种冗余技术，但都只能保证有限块盘故障的场景，如RAID5只允许一块异常失效，当出现两块硬盘失效，存储系统为保障数据可靠，将停止对外提供服务。
- 为了减少性能波动和多盘故障带来影响，存储系统每天定时将硬盘S.M.A.R.信息、性能指标以及硬盘LOG发送到eService安全接入服务，DME IQ提取硬盘特征指标后，通过华为自研的硬盘故障预测模型，每天进行学习和训练，综合分析并提前2周准确预测即将故障硬盘，从而实现主动式故障预防，显著提升系统可靠性。
- DME IQ系统，提供硬盘健康状态的查看，硬盘健康趋势统计，方便全面了解所有系统硬盘状况。

容量趋势预测

- 容量如何达到低成本与高性能的平衡，企业一直在积极寻找容量管理的解决方案，但由于无法精准预判当下业务增长对容量的具体诉求，导致临时紧急扩容；同时又受多种因素影响（节假日、临时事件等），人工预测容量不准确，无法精确给出中长期的业务需求规划，长期困扰IT运维人员。
- DME IQ基于eService基础数据底座采集的设备配置信息、性能数据、告警事件等数据，使用机器学习来学习和训练，选取最佳预测模型，预测未来12个月的容量消耗情况。

- 性能异常检测

- DME IQ提供实时、精准的性能异常检测与诊断能力。
- 业务能否平稳运行，是企业最关心的问题，但由于性能问题的复杂性与隐蔽性，用户难以提前识别并解决性能问题。等到问题恶化并影响业务时，已经给用户造成损失。
- DME IQ性能异常检测功能，依托于华为强大的专家经验，实时检测性能异常，并予以分析，保障异常检测有效、可靠、及时。
- DME IQ异常检测具备以下特点：
 - 1. 实时
 - 分钟级别异常检测，全天24小时实时防护。
 - 2. 精准
 - 针对业务出现的时延问题，从I/O变化、业务配置、硬件瓶颈、硬件故障等方面进行精准定位并给出修复建议。

- 性能瓶颈分析

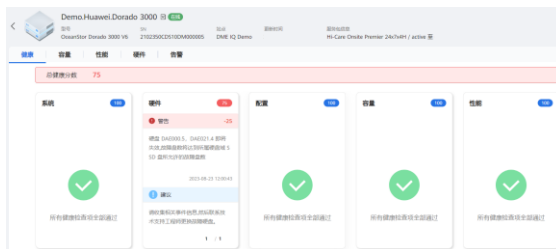
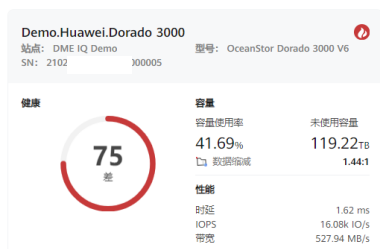
- 随着业务上线，运维人员普遍关心设备性能压力是否过大、业务能否稳定运行。
- 由于影响设备性能的因素复杂（设备硬件配置、设备软件配置、业务类型、性能数据等），设备性能压力评估以及性能调优成为难题。
- DME IQ性能瓶颈分析对设备硬件配置，软件配置，性能数据等进行分析，评估设备性能压力并结合华为强大的专家经验，准确识别性能瓶颈并给出相应的优化建议。

健康评估

设备健康度实时监测

五个维度全面评估，智能分析设备硬件、配置、容量、性能、系统等健康情况，及时风险同步，确保存储高可用&高性能。

类型		监测要素举例
	硬件	硬盘、线缆、风扇等
	配置	完整性、配置合规、数据保护等
	容量	Pool容量、数据保护预留空间等
	性能	CPU利用率、时延、IOPS等
	系统	License、EOS、补丁版本等



36 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



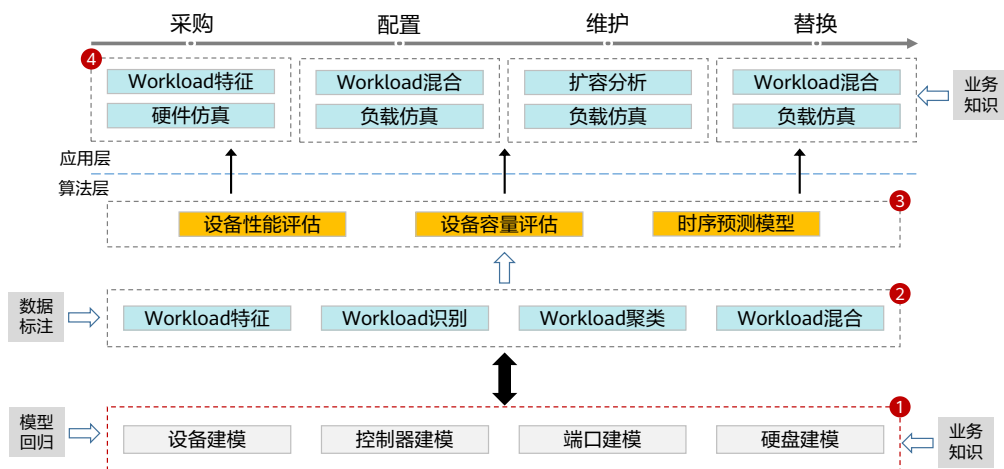
- 通常情况下，设备上线后，客户通过巡检来预防设备风险，这种方式面临两个问题。
 - 巡检频次问题：巡检一般固定周期（月度或季度）进行，导致客户不能及时发现问题。
 - 巡检深度问题：只能检查当前设备是否存在问题，缺少对设备系统、硬件、配置、容量、性能风险的分析。
- DME IQ从系统、硬件、配置、容量、性能维度实时评估设备健康，发现设备潜在风险，通过健康分数直观反映设备运行状况，同时为客户提供解决方案，实现风险的主动预防。并根据风险级别策略建单，跟踪闭环。
- 此外，可以对关心的设备进行报告订阅，订阅后每月自动生成设备健康报告并通过Email发送到您的邮箱。同时还可以DIY报告内容，包括自定义报告的模板、内容、时间段等，创建一份独有的设备健康报告。
- 基于华为全球维护经验，并综合运用多种先进技术（包括不限于大数据分析、机器学习等），智能分析设备硬件、配置、容量、性能、系统等健康情况，识别设备运行中存在的问题和潜在的风险，进行总结报告并提出相应的措施，主动预防。

规划仿真



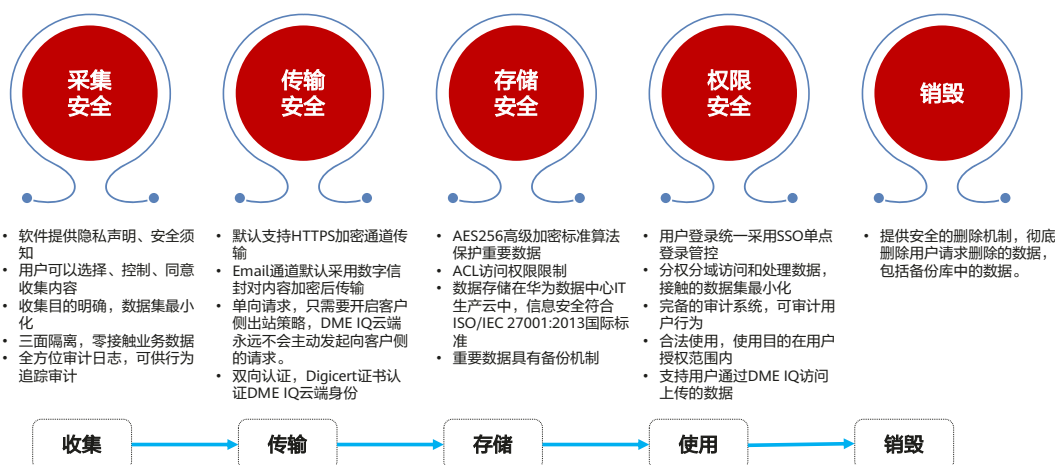
- 规划仿真是面向存储设备全生命周期的仿真评估, 支持业务上线、业务SCALE和业务迁移3大负载变更场景, 支持Oracle、DB2等八大业务类型自动化评估与最佳业务安置配置智能推荐, 实现负载和容量趋势分析和预测, 准确率大于80%。另外, 能够为运维工程师提供设备性能和容量变化趋势评估能力, 辅助其方案变更决策。

规划仿真特性原理



1. 硬件建模：通过数据标注、模型回归和业务知识，对控制器、端口和硬盘关键存储模块建模，实现整个存储设备建模；
2. 负载建模：通过数据标注和模型回归的方式，能够对业务Workload关键特征进行提取，进一步识别业务类型和业务聚类，研究Workload混合机制；
3. 性能/容量评估：通过硬件建模和负载建模，构建设备性能和容量评估算法；
4. 场景能力：结合业务知识和性能、容量评估算法，构建设备采购、业务上线、业务扩展和业务迁移场景仿真能力。

安全无忧



39 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



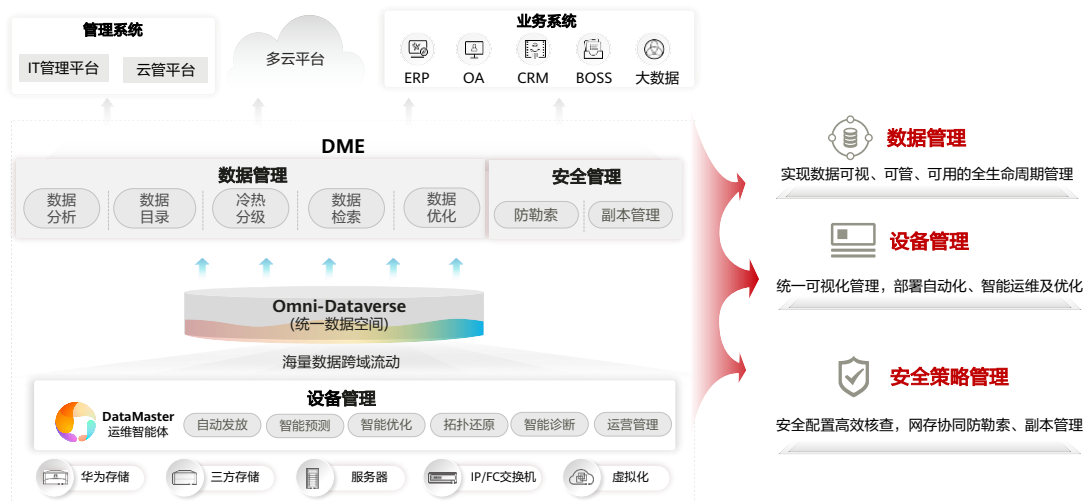
- DME IQ从业务流程、整体架构设计和技术实现方案上都进行了仔细的安全考量，做到对客户安全进行严密防护。
- 采集安全
 - 经过客户允许才会部署并符合当地法律，不碰设备的业务数据，只获取相关运维信息，详细收集的信息条目均列在《授权协议书》中。
 - eService Client上进行的所有操作均有记录，上传记录可追溯，并采用完整性校验，确保记录不被篡改。
 - 数据保存时均经过AES256加密处理，确保信息不泄露。
 - 在架构设计上，eService Client的所有连接都从客户侧向外发起，任何第三方包括华为公司都不能主动从外部网络向客户侧创建网络连接。
- 传输安全
 - 双向认证：eService Client通过证书认证华为云端，华为云端通过Support帐号认证eService Client。
 - 单向访问：eService Client使用Https连接华为云端，通过Https加密通道进行传输，确保安全连接。
 - 通过国家电子计算机质量监督检验中心认证，安全项测试覆盖100%，并出具《安全认证报告》。
- 存储安全

- DME IQ云端所在部署环境，信息安全管理体系符合ISO27001标准要求，确保客户数据的安全性、完整性和可用性。
- DME IQ云端管理系统部署为多套，分区域覆盖，确保数据不会传出规定区域。
- 权限安全
 - eService Client 和eService安全接入服务，以及DME IQ云端管理系统具备完善的认证、授权、审计机制（3A）和有效的数据加密保护措施。

目录

1. 存储智能运维
2. **存储运维工具**
 - SmartKit
 - iMasterCloud DME IQ
 - **iMaster DME**
3. 存储日常管理
4. 存储系统升级

iMaster DME



41 Copyright © Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

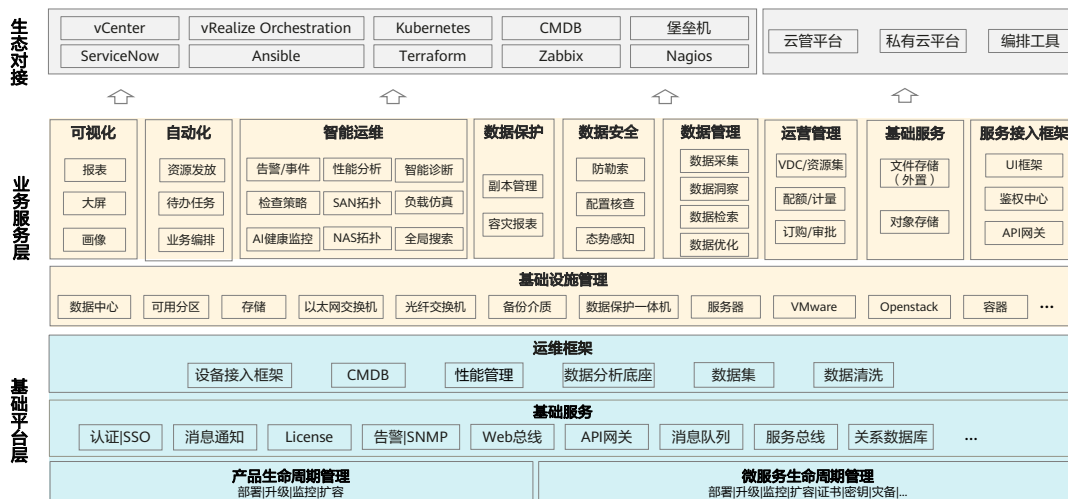
- iMaster DME（本课程简称为DME）是华为公司推出面向数据中心存储设备的集中管理运维软件，覆盖设备管理生命周期的四个阶段：规划、建设、维护、优化，并提供面向非结构化数据的管理能力：检索、分析、优化。通过提升管理的自动化程度和问题闭环率，实现资源管理降本增效。
- 华为DME致力于为用户提供针对华为存储的全生命周期自动化管理，涵盖规、建、维、优四个维度。业界率先实现管理、控制、分析在一个软件上融合，用户只需登录一个GUI（Graphical User Interface），在统一的界面体验下完成数据的存储、分析、备份和容灾。

数据全生命周期管理



- DME定位为存储多设备管理软件，提供IT基础设施生命周期管理、数据生命周期管理两大功能。IT基础设施管理围绕存储资源提供控制、管理和分析能力。数据生命周期管理，围绕非结构化数据提供检索、分析和优化能力。
- 针对IT基础设施生命周期管理，DME的价值如下：
 - 可视化：提供多维度的资源报表和大屏监控，可根据用户需求灵活定制数据中心的资源报表和监控，全面掌握资源使用情况。提供数据中心在线洞察能力，全面分析整个数据中心的风险，以可视化视图方式直观展示。
 - 自动化：覆盖存储生命周期的全流程自动化能力，让管理员从重复繁琐的工作中抽离出来，同时提供负载仿真模拟、自动化编排、预检审批能力，有效避免了人为错误。
 - 智能化：基于AIOps提前发现问题、一键定位问题、快速闭环问题，提升业务服务质量，避免业务异常，降低业务宕机时间。

业务架构



43 Copyright © Huawei Technologies Co., Ltd. All rights reserved.



- DME采用了ServiceMesh服务化架构设计思想，基础平台层提供产品组件和微服务生命周期管理、基础服务、运维框架，上层提供基础设施管理、可视化、业务编排、智能运维、数据保护、数据安全、数据管理几大业务模块，涵盖存储和数据的全生命周期管理，实现对存储管控融合。让您通过一个界面入口、一套API实现更简单、更高效的管理。
- 该架构的关键特点：
 - 分布式微服务架构，支持高可靠部署，在多节点部署场景下，单节点故障时系统仍可以继续提供服务，可靠性可达99.9%。
 - 支持横向扩展，满足大规模存储设备的管理需求。
 - 南北向全面开放系统，北向提供RESTful、SNMP、生态插件（如Kubernetes CSI、Ansible等）对接北向生态，南向通过SSH，SNMP、RESTful等协议实现对存储资源、交换机资源、主机资源的管理。

DataMaster

运维门槛较高

DME典型运维场景平均需要>3次
页面点击



DME有超过200个功能点，需
要专业培训



运维经验全靠人工积累



LUI基于自然语言的人机交互

人为操作 -> AI操作



华为+企业运维经验知识沉淀



20+典型场景, 100+API

场景	关键功能
监控 分析	查询资源列表
	查询资源关联对象
	查询性能数据
	查询健康分析
诊断	...
	执行智能诊断

从意图理解到任务完成，全链路运维自动化闭环

智能体自动化任务编排，自动执行

- 意图理解 -> 知识检索 -> 方案生成 -> 智能体编排
- 基于知识图谱+RAG的Agent任务决策引擎，实现任务动态分配



开放API对接，支持第三方模型*

- 南北辅助开发，快速融入客户运维系统
- 支持对接第三方AI模型



- 对话式运维（DataMaster）基于生成式AI技术，针对具体的运维角色，辅助问题定位和问题闭环，是人机交互的数字助理。其关键能力包括自然语言对话（LUI）、任务规划（LLM）和资源查询和诊断。

基础设施集中管理



- 基础设施管理通过标准对接协议，发现存储、网络、计算、备份系统等物理资产以及应用等软件资产，并对这些资产的主要数据资产进行发现，按类型进行统一呈现，实现多资源的一站式管理。
- 主要分为存储、网络、计算、备份系统、应用资产管理。
 - 存储资产管理：管理华为存储、第三方存储，提供存储设备的访问、配置、台账能力。
 - 网络资产管理：管理华为、博科、思科光纤交换机，提供对交换机的管理能力，以及对交换机Zone、别名等资源的配置能力。
 - 计算资产管理：管理存储主机、物理主机、vCenter服务器，提供对主机的接入、查看、修改能力。
 - 备份系统管理：管理备份存储、备份一体机、备份集群，提供备份存储的访问、配置能力，对备份一体机进行统一呈现。
 - 应用管理：管理应用信息，整合应用关联资源，统一呈现应用以及应用与计算、存储资源的关联关系。
- 资产管理主要实现逻辑为：
 - 用户提供资产接入信息在界面接入对应的资产。
 - DME对资产进行发现，并采集资产的资源，进行呈现。
 - 用户通过DME对应的资产界面进行资产的查看、管理操作。

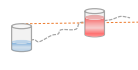
智能分析

风险被动响应

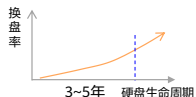
业务扩张致性能不足，但扩容周期长



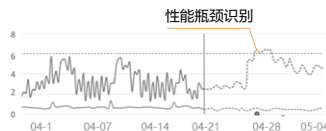
规划不合理，致容量超预期达到预警值



硬盘周期过3年后，坏盘率提升、被动更换



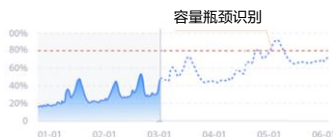
**提前14天
性能预测**



准确率：80%以上

采用Lasso回归、岭回归、随机森林的融合预测算法

**提前365天
容量预测**



准确率：95%以上

采用向量机（多种核函数）、Arima模型、对数回归、线性回归的融合预测算法

**提前30天
硬盘预测故障**



检测准确率：80%

支持范围：HDD、SSD
HDD支持提前30天，SSD提前14天

- 智能分析通过分析计算、网络、存储各阶段时延消耗，结合I/O全路径资源性能瓶颈分析，完成性能问题定界，提升性能问题分析效率。主要提供以下关键能力，通过不同关键能力组合，可解决客户SAN性能问题定界、流量分析、NAS扰邻分析等需求。
- 性能问题定界：识别分析对象关联的虚拟机/存储主机/VBS客户端时延是否存在突变、上升，并分析计算、网络、存储分段时延消耗，锁定故障源范围。
- 资源瓶颈分析：分析主机、交换机端口、存储前端端口、控制器、后端端口、硬盘、VBS客户端等资源是否已达到性能瓶颈。
- 业务变化分析：
 - SAN场景：分析与瓶颈对象或入口对象相关的LUN负载变化，根据I/O大小变化、I/O命中率变化与时延变化的相关性评估负载变化对时延的影响，并分析使用瓶颈对象或入口对象的虚拟机/存储主机的带宽、IOPS变化，展示TOPN对象。
 - NAS场景：
 - 闪存存储：分析使用瓶颈对象或入口对象的IP客户端或逻辑端口带宽、OPS变化，展示TOPN对象。
 - OceanStor A系列存储：分析zone内性能瓶颈资源，展示访问该资源的TOPN并行客户端性能统计。

主动健康检查



47 Copyright © Huawei Technologies Co., Ltd. All rights reserved.

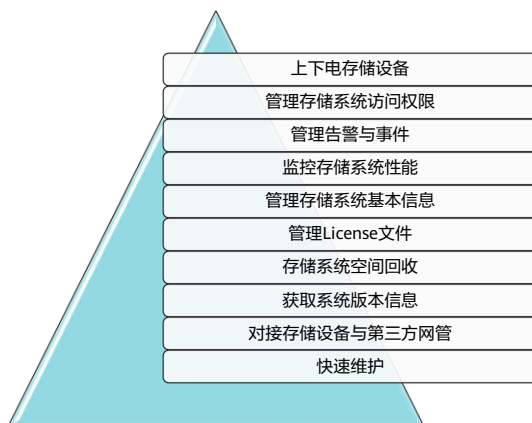


- 通过AI算法推断或预测资源风险，根据风险的严重程度对资源的健康状态进行综合评估，给出直观的健康分数，让用户主动提前发现并解决问题，避免对业务造成严重影响。
- 健康检查的维度包括告警、性能异常、性能预警、容量预警几个方面，针对每类对象的各个维度的风险，DME根据风险程度内置了扣分规则，用户可以进行自定义。用户可直观的查看每类对象的健康概览，掌握处于不同健康级别的对象个数，并按照健康度进行排序，查看影响健康度的维度，和各维度的风险详情，并执行常见的运维或修复动作。
- 健康检查特性主要包含如下功能：
 - 管理健康检查策略：查看和管理健康检查策略，包括性能异常检查策略、性能预警策略、容量预警策略，用户可设置检查对象、指标、阈值，其中性能异常检查策略还支持设置检查时段、检查算法、检查周期等高级选项。
 - 设置健康打分规则：用户可定义健康打分规则，如每个维度不同风险级别的扣分值、每类对象各维度的最大扣分值、各指标的最大扣分值。
 - 查看健康打分结果：查看每类对象处于不同健康级别的对象个数，按健康分数排序或按属性筛选对象，查看每个对象的健康分数和各维度的扣分概况，查看对象的详细属性和风险详情。

目录

1. 存储智能运维
2. 存储运维工具
- 3. 存储日常管理**
4. 存储系统升级

存储日常管理项目总览



- 以闪存存储产品为例讲解，其他产品可参考产品文档内容。

管理存储系统访问权限



管理存储系统安全配置管理

安全策略

- 帐户策略
- 登录策略
- 访问控制
- 用户帐户审计
- 弱口令字典

管理用户

- 本地用户
- 域用户
- 登陆方式
- 用户角色
- 用户密码

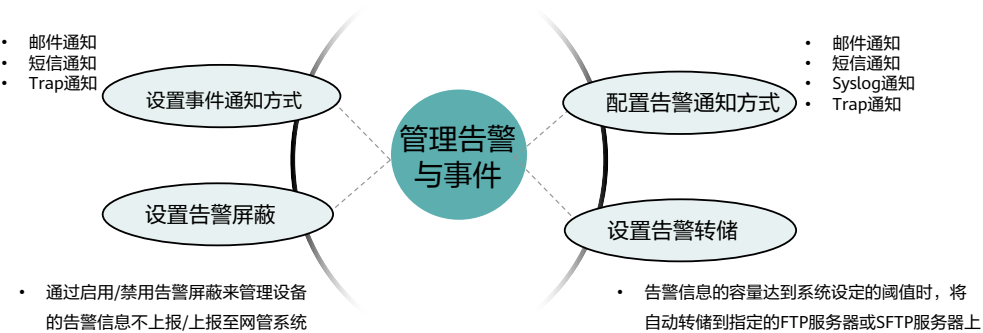
配置SSO服务器

- SSO (Single Sign-On, 单点登录) 是身份管理中的一部分
- CAS (Central Authentication Service) 是一种网页单点登录协议
- 存储系统通过SAML协议对接AD FS服务器或CyberArk服务器, 实现存储系统的统一身份认证和多因子认证。

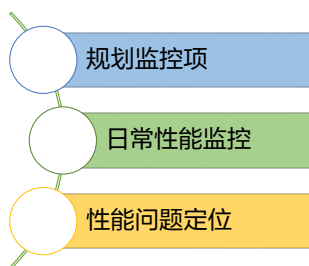
第三方网管平台 免密登陆

- 获取令牌
- 访问方法:
PUT

管理告警与事件

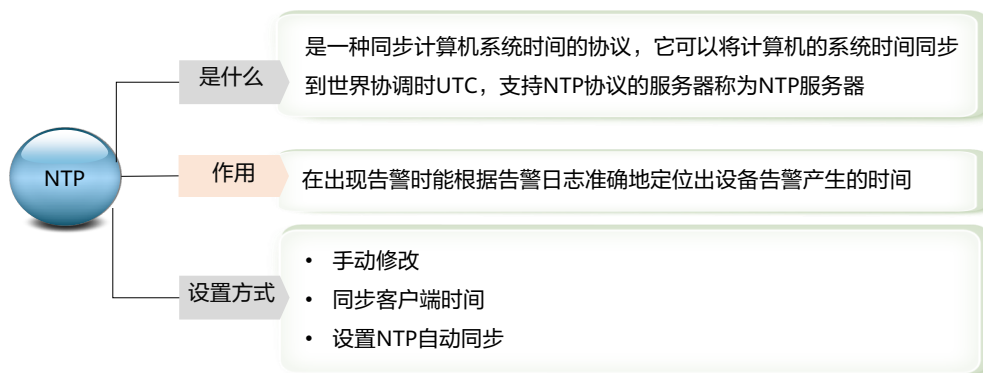


监控存储系统性能



- 衡量存储系统性能的主要指标包括：
 - **IOPS**: I/O per second, 即每秒钟可以处理的I/O个数, 用来衡量存储系统的I/O处理能力。
 - **带宽**: 每秒钟可以处理的数据量, 常以MB/s或GB/s为单位, 用于衡量存储系统的吞吐量。
 - **时延**: 发起I/O请求到I/O处理完成的时间间隔, 常以毫秒 (ms) 为单位。常用指标包括平均响应时间和最大响应时间。

管理存储系统基本信息 - 设置设备时间 (1)



- 设置NTP自动同步前，需要完成NTP服务器的配置。
- 在有防火墙的环境中使用NTP自动同步时，需要开启123端口。
- 通过同步客户端时间，可以将存储系统时间调整为和客户端时间一致。
- 通过设置NTP自动同步，可以将NTP服务器作为外部时间源，定期、自动地同步到设备。

管理存储系统基本信息 - 设置设备时间 (2)

- 登录DeviceManager管理设备时间



设备时间 ⓘ

当前设备时间

修改方式

- ☐ 同步客户端时间 通过同步客户端时间，可以将设备时间调整为和本地客户端时间一致。
- ☐ 手动修改 手动修改设备的时间或时区。
- ☒ 设置NTP自动同步 将NTP服务器上的时间自动同步到设备。

* NTP自动同步 [配置](#)

- 登录CLI管理设备时间
 - **change ntp_server config** 用于将存储系统的时间自动与NTP服务器的时间同步
 - **show system ntp** 用于查询NTP的设置信息
 - **show ntp status** 用于查询NTP状态
 - **show ntp_server general** 用于查询时间同步功能的设置信息

- 说明：请设置正确的时区和时间，否则可能导致告警或日志中记录的时间与实际时间不一致，影响后续问题定位。

管理设备License (1)



是什么

使用各种增值特性（例如快照、远程复制、克隆、SmartQoS等）的**权限凭证**

注意事项

在设备的日常管理中，需要关注 License 文件是否可用

使用DeviceManager管理 License

根据存储系统当前是否已导入或激活License的状态，在“License管理”界面会自动显示不同的License操作：更新License和备份License。

对于已激活的License文件，DeviceManager有两种控制方式：

- 基于运行时间控制：显示该License的“运行截止时间”
- 基于容量控制：显示该License的“已使用/总容量”

管理设备License (2)

使用CLI管理设备License

- export license 命令用于导出 License 文件

例： export license ip=? user=? password=? license_path=? [port=?] [protocol=?]

- import license 命令用于导入 License 文件。

例： import license ip=? user=? password=? license_path=? [port=?] [protocol=?]

- show license 命令用于查询系统中已导入的 License 文件的功能配置信息。

例： show license

- show license_active 命令用于查询已激活的 License 信息。

例： show license_active

获取设备当前版本信息

版本信息

作用

可根据系统版本准确判断配套软件版本

- 登录DeviceManager获取系统当前版本信息



- 以超级管理员身份登录CLI

- 执行 show system general 命令
- “Product Version” 即为当前存储系统的版本信息

```
admin:/>show system general
System Name       : XXX.Storage
Health Status    : Normal
Running Status   : Normal
Total Capacity   : 3.186TB
SN               : 210235G6EHZ0CX0000XX
Location         :
Product Model    : XXXX
Product Version   : VX00R00XCXX
High Water Level(%) : 80
Low Water Level(%) : 20
WWN              : XXXX
Time             :
Patch Version    : SPCXXX SPHXXX
Description      :
```

获取设备ESN序列号



是什么 用来唯一标识设备的字符串

使用场景 申请License、设备报修和配置eService 服务等场景

- 使用DeviceManager获取ESN序列号

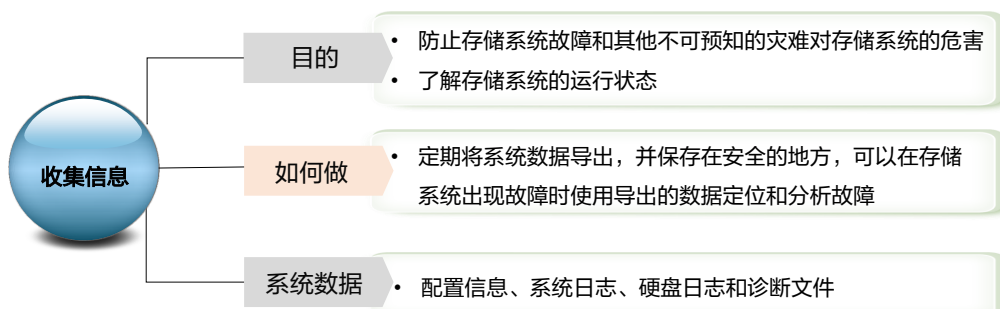


- 使用CLI获取ESN序列号

>> 执行 show system general 命令

```
admin:/>show system general
System Name       : XXX.Storage
Health Status    : Normal
Running Status   : Normal
Total Capacity   : 3.186TB
SN                : 210235G6EHZ0CX0000XX
Location         :
Product Model    : XXXX
Product Version  : VX00R00XCXX
High Water Level(%) : 80
Low Water Level(%) : 20
WWN              : XXXX
Time             :
Patch Version    : SPCXXX SPHXXX
Description      :
```

收集存储系统信息



收集存储系统信息方式

使用 DeviceManager 收集存储系统数据

在DeviceManager中：

- 可直接导出“配置信息”，以此可搜集系统当前运行状态信息。
- 可选择下载“近期日志”或“全部日志”，以此可搜集存储设备上的配置信息、事件信息和调试日志等信息。
- 可选择下载“DHA运行日志”或“HSSD日志”，以此可搜集硬盘运行日志、I/O统计和寿命、S.M.A.R.T等日志信息。
- 可直接导出“诊断文件”，以此可搜集存储设备的故障信息。

使用CLI收集存储系统配置数据

通过超级管理员帐户登录存储系统的CLI，执行以下命令后将配置文件导出到FTP服务器或SFTP服务器上。

```
export configuration_data ip=? user=? password=? db_file=? [ port=? ] [ protocol=? ] [ clean_device_file=? ]
```

- 系统日志包括存储设备上的配置信息、事件信息和调试日志等信息，可用于分析存储设备的运行状态。导出的系统日志文件类型是*.tgz。
- 配置信息是指系统当前的运行状态数据，反映的是实时的和需要关注的系统配置信息和状态信息，例如用户信息或LUN的配置信息等。导出的配置信息文件类型是*.txt。
- DHA运行日志是指硬盘的日常运行日志，主要包含关于硬盘健康统计的日常采集信息、I/O统计信息和硬盘寿命信息。导出的DHA运行日志文件类型是*.tgz。
- HSSD日志是HSSD盘的工作日志，主要包含硬盘的S.M.A.R.T信息和硬盘运行日志。导出的HSSD日志文件类型是*.tgz。
- 硬盘数据擦除报告主要包括执行数据擦除的硬盘信息、擦除方式、执行时间和校验结果等。导出的硬盘数据擦除报告文件类型是*.csv。
- 诊断文件主要包含设备的故障信息。导出的诊断文件类型是*.tgz。

收集存储系统信息内容

信息类型	名称	说明
基本信息	设备序列号和版本	提供存储设备的序列号与版本信息。
	客户信息	提供具体的联系人及其联系方式。
故障信息	故障发生时间	记录故障发生的准确时间。
	故障现象	详细记录故障发生时的现象，例如弹出的错误提示对话框、收到的事件通知内容等。
	故障前执行的操作	回忆并记录故障发生前执行的操作。
	故障后执行的操作	回忆并记录故障发生后到向维护人员上报故障前执行的操作。
存储设备信息	硬件模块配置	详细记录存储设备硬件模块的配置信息。
	指示灯状态	记录存储设备指示灯状态，尤其需要记录橙色和红色状态的指示灯信息。存储设备的各部件指示灯状态，请参见对应产品型号的《产品描述》。
	存储系统数据	手工导出存储设备的运行数据和系统日志。
	告警和日志	手工导出存储设备的告警和日志信息。
组网信息	连接方式	描述应用服务器与存储设备间的网络连接方式，如FC组网、iSCSI组网等。
	交换机型号	如果网络中存在交换机，请记录交换机的型号。
	交换机诊断信息	手动导出交换机目前运行的诊断信息，包括启动配置、当前配置、接口信息、时间、系统版本等。
	网络拓扑结构	描述应用服务器与存储设备间的网络拓扑结构或提供组网图。
应用服务器信息	IP地址信息	如果应用服务器与存储设备间采用iSCSI组网，需要描述IP地址划分原则或提供IP地址分配列表。
	操作系统版本与日志	记录应用服务器中安装的操作系统类型、版本号与系统日志。
	端口速率	记录与存储设备相连的应用服务器端口速率。如何查看端口速率请参见《联机帮助》。

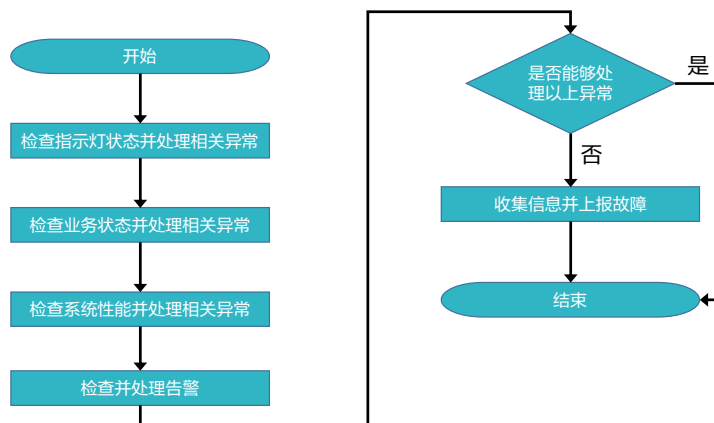
对接存储设备与第三方网管

对接方式	说明	操作指导参考文档
SMI-S	通过将SMI-S provider安装到第三方的Windows/Linux服务器上，用户可以通过SMI-S provider对华为存储系统进行管理。	《eSDK Enterprise Storage Plugins X.X 用户指南（SMI-S Provider）》
SNMP	第三方网管可以通过SNMP协议查看存储设备的信息，如LUN、端口和存储池等信息。	•对应产品型号的《初始化指南》的“配置Trap通知”章节。
VAAI NAS	华为VAAI NAS插件基于VMware提供的vSphere APIs for Array Integration (VAAI)框架开发，以支持硬件加速。通过硬件加速，ESXi主机可以与NAS设备相集成并使用NAS存储提供的多个硬件操作。 VAAI NAS框架支持这两种版本的NFS数据存储：NFS 3和NFS 4.1。	《eSDK Enterprise Storage Plugins X.X 用户指南（VAAI NAS）》
VASA	VASA Provider是华为技术有限公司开发的符合VASA标准的软件，用以实现通过vCenter Server对华为存储设备进行管理。VASA Provider支持VMware的vVol（Virtual Volume）功能。	《eSDK Enterprise Storage Plugins X.X 用户指南（VASA）》
REST	RESTful API是OceanStor DeviceManager提供的基于REST(Representational State Transfer)标准的开放接口。第三方开发者能使用RESTful API授权访问OceanStor DeviceManager开放的资源，例如：告警查询、性能监控、资源分配等。	《OceanStor 6.1.X REST 接口参考》

- REST：Representational State Transfer，表现层状态转换
- SMI-S：Storage Management Initiative Specification，存储管理计划规范
- SNMP：Simple Network Management Protocol，简单网络管理协议

快速维护流程

存储系统的快速维护流程如下图所示：



- 通过检查存储系统前面板的指示灯状态，您可以了解设备的运行状态，发现和解决常见的硬件故障。
- 通过DeviceManager存储系统的业务状态，可以了解存储系统的基本信息、告警、系统容量趋势和性能方面的概貌。
- 通过DeviceManager检查存储系统的运行性能，您可以获知存储系统业务性能的实时和长期状态。如果存储系统出现性能问题，您可以及时采取措施，避免业务中断或者数据丢失。
- 当存储系统异常时，DeviceManager会自动判断异常状态对当前业务的影响严重程度，并以不同级别的告警提示维护人员进行处理，避免业务中断或数据丢失。

配置存储基础业务



作用

- 将存储系统提供的存储空间划分为多个LUN
- 将LUN映射给应用服务器
- 应用服务器能够使用存储系统提供的存储空间

使用 DeviceManager配置存储基础业务

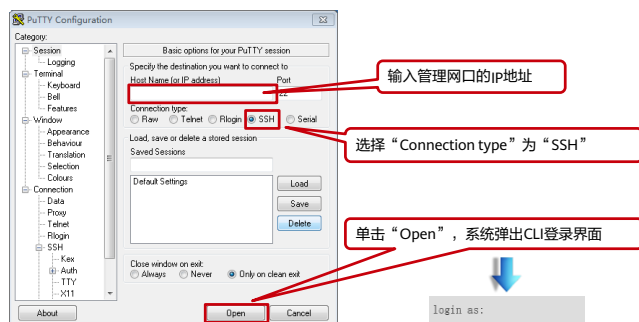
- 创建存储池：DeviceManager支持通过“推荐”或“自定义”的两种方式创建存储池。
- 分配存储资源，继续通过创建LUN组或者创建文件系统的形式将存储资源进行分配。

使用CLI配置基础存储业务



CLI登录

- CLI (Command Line Interface) 提供命令行的方式对存储系统进行管理和维护，配置命令通过键盘输入，由程序编译执行。执行结果以文本或图形的方式回显在CLI操作界面。
- 登录CLI时，需要使用终端软件，此处以PuTTY软件登录CLI为例。



CLI格式约定 (1)

- 在CLI命令的使用过程中，需要遵循CLI格式约定。典型的命令格式如下所示：

```
change storage pool relocation speed relocation_speed=?  
show host
```

- 第1个字段：进行的操作。例如：change（修改）、show（查询）。
- 第2个字段：执行此操作的对象。例如：storage_pool（存储池）、host（主机）。
- 第3个字段（仅部分命令含有）：执行操作的对象属性。例如：relocation_speed（迁移速率）。
- 其他字段：执行命令所需的参数。

CLI格式约定 (2)

格式	意义
粗体	命令的关键字（命令中保持不变）采用 加粗 字体表示。
<i>斜体</i>	命令的参数（命令中必须由实际值进行替代的部分）采用 <i>斜体</i> 表示。
[]	表示用“[]”括起来的部分在命令配置时是可选的。
{ x y ... }	表示从两个或多个选项选取一个。
[x y ...]	表示从两个或多个选项选取一个或者不选。
{ x y ... } *	表示从两个或多个选项选取多个，最少选取一个，最多选取所有选项。
[x y ...] *	表示从两个或多个选项选取多个或者不选。

- 例如，`change user user_name=? { level=? | action=? }`
 - 命令 **change user** 保持不变。
 - 参数 **user_name=?** 必选，**level=?** 和 **action=?** 只能二选一。
 - 参数 **level=?** 中，**level=** 保持不变，**?** 必须用可选取值替代，如 **level=admin**。
- 正确命令范例：`admin:/>change user user_name=newuser level=admin`

CLI命令快捷功能

■ CLI命令补全功能

在CLI中，通过按“Tab”或空格键，可以使用命令补全功能。

二者的区别：空格键只补全当前的字段，而“Tab”键在补全的同时提示用户所有可能的取值。

■ CLI命令过滤功能

该功能可去掉冗余信息，按需显示有效内容
相关命令有：

- filterColumn 列过滤命令
- filterRow 行过滤命令

■ 上下文相关帮助功能

通过按“Ctrl+A”，可以在某些命令中指定参数时，自动提示某些参数的可选取值。这些取值通常需要从系统中获取。

■ 错误提示功能

指出命令中输入错误的位置，并给出正确字段供参考。

当输入的命令存在格式错误时，系统以^符号提示错误位置。

华为多路径基本配置操作指导

UltraPath命令行配置方式：

以Windows系统为例，通过点击“开始->所有程序->UltraPath->upadm”或直接在Windows的CMD命令行中输入upadm，进入upadm命令行模式，并通过相应CLI命令配置UltraPath。

注：文档中主要以Windows系统为例介绍华为多路径的基本配置命令，其他操作系统配置命令与Windows系统类似，具体命令请参考相应系统的《用户指南》。

- 华为多路径安装、卸载操作具体步骤及注意事项请参考对应多路径版本各个操作系统的《 OceanStor UltraPath for xxxxxx 用户指南 》文档。升级操作具体步骤及注意事项请参考多路径对应版本的《 OceanStor UltraPath xxxxxx 升级指导书 》。
- 用户指南和升级指导书的获取地址：
 - 企业业务：<http://support.huawei.com/enterprise/zh/storage/ultrapath-pid-8576127>

华为多路经常用参数在典型场景下的配置

大多数场景下，推荐使用华为多路径的默认配置。部分典型场景，可参考以下说明对华为多路径进行配置：

upadm set workingmode={0/1}

- 设置存储控制器层面的负载均衡模式，0为控制器间负载均衡，1为控制器内负载均衡。
- 默认为控制器内负载均衡，多路径会根据每个存储LUN的归属控制器去选择路径下发I/O。
- 而控制器间负载均衡时，多路径会在所有路径上下发I/O，但这会增加I/O在存储控制器间转发的处理时延。

典型场景	推荐配置
主机和存储系统之间的传输路径成为性能瓶颈	0（控制器间负载均衡）
其他场景	1（默认配置，控制器内负载均衡）

华为多路径常用参数在典型场景下的配置

upadm set loadbalancemode={round-robin|min-queue-depth|min-task}

- 设置链路层面的负载均衡算法：round-robin即为轮询算法，min-queue-depth即为最小队列深度算法，min-task即为最小任务算法。
- 使用轮询算法时，多路径会根据应用服务器和存储系统当前的可用路径，依次选择每一条路径下发I/O数据。
- 默认为最小队列深度算法，多路径会根据当前每一条可用路径上的I/O个数，选择所有路径中I/O数目最少的路径下发I/O。
- 使用最小任务算法时，多路径会根据当前每一条可用路径上I/O数据的大小，选择所有路径中总I/O数据量最小的路径下发I/O。

典型场景	推荐配置
主机下发的业务I/O模型差异较小，且需要业务在每一条路径上均衡下发	round-robin（轮询算法）
主机下发的业务I/O模型数据块较大	min-task（最小任务算法）
其他场景	min-queue-depth（默认配置，最小队列深度算法）

目录

1. 存储智能运维
2. 存储运维工具
3. 存储日常管理
- 4. 存储系统升级**

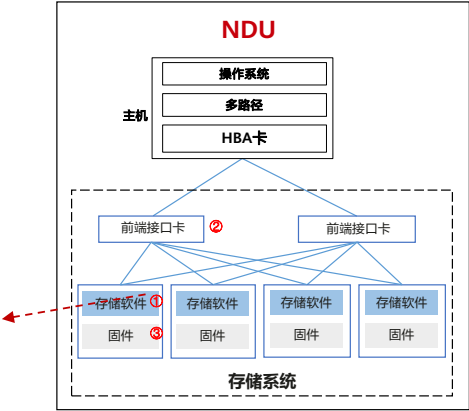
华为闪存存储系统支持的升级方式

- 快速升级（推荐升级方式）
 - 快速升级是一种通过进程替换，不复位控制器的快速升级方式，整个升级过程中主机链路不中断，主机对整个升级过程无感知，是华为全闪存存储推荐使用的升级方式。
- 滚动升级（备选升级方式）
 - 滚动升级方式是一种控制器分批升级并复位生效的升级方式，升级过程中业务将在控制器间进行倒换，配合主机多路径软件，可保证升级过程中主机业务不中断。
- 并行升级方式（仅推荐在无任何业务配置下使用）
 - 并行升级方式是在存储设备未关联任何主机情况下推荐使用的一种升级方式，所有控制器一起升级，升级时长是滚动升级方式的一半。

无中断升级

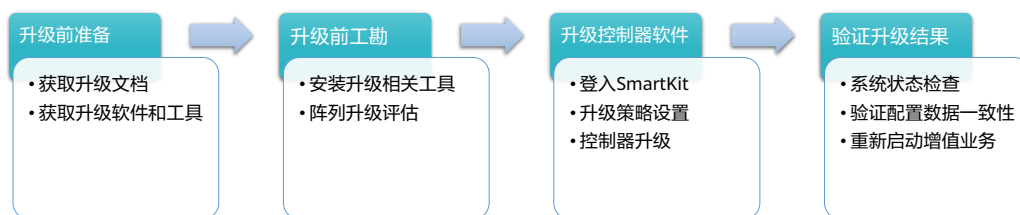
升级业务无感知：

- ① 软件组件化快速升级，且I/O路径上的组件1秒内完成升级
- ② 前端接口卡为自研芯片，支持升级前端连接不中断
- ③ 固件支持弱配套和补丁式升级；极端场景依靠前端接口卡支持连接不中断



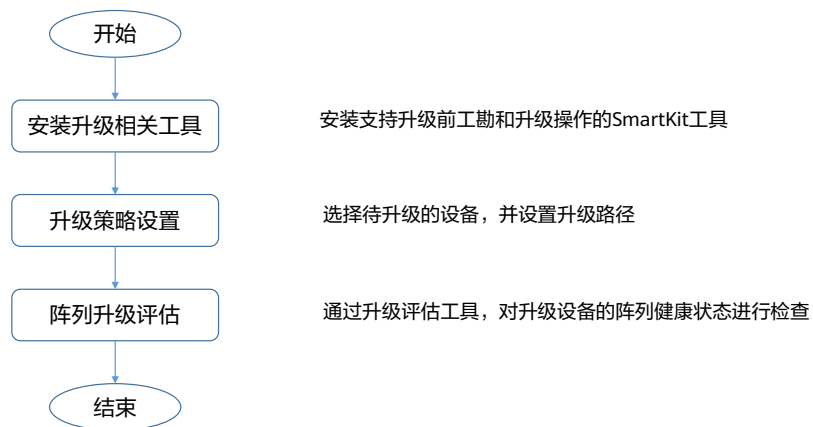
NDU: non-disruptive upgrade, 无中断升级

升级流程图



- 安装支持升级前工勘和升级操作的SmartKit工具
- 选择待升级的设备，并设置升级路径
- 通过升级评估工具，对升级设备的阵列健康状态进行检查
- 选择升级设备、导入升级包，升级控制器软件

升级前工勘



升级控制器软件



升级结果验证

验证项目	验证标准
系统状态检查	通过巡检工具，检查系统状态是否正常，确保升级过程未影响系统状态
验证配置数据一致性	• 用户配置与升级之前是否一致 • 升级后License是否可用
重新启动增值业务	如果升级前对增值业务（包含：快照和远程复制）进行了暂停、停止或分裂处理，升级后需要恢复到升级前状态

版本回退

- **前提：**升级过程中发生异常，系统会尝试自动修复系统。版本回退只能在离线的环境下进行，请在版本回退前确认主机业务已停止。版本回退必须在两端控制器软件版本均已升级成功的情况下进行。需要进行版本回退的可能因素有：
 - 升级完成后业务无法恢复。
 - 升级完成后系统性能下降。
 - 升级完成后配置数据丢失并无法恢复。
- **方式：**通过Smartkit中的升级工具进行版本回退
 - 通过升级工具集进行版本回退的操作时，请导入指定的升级包后，其余操作与通过升级工具升级控制软件版本的操作完全一致。

思考题

1. （多选题）为保证系统安全性，可以通过以下哪些方式保证登录存储系统安全性？
 - A. 域认证
 - B. 配置可访问IP地址
 - C. 管理用户权限
 - D. 配置用户账户审计
2. （判断题）重新启动存储系统不会对主机业务造成任何影响。
 - A. True
 - B. False

- 参考答案：

1. ABCD
2. B

本章总结

- 本章主要讲解IT运维的相关概念和发展趋势，介绍了存储运维工具SmartKit、iMasterCloud DME IQ和iMaster DME，存储系统日常管理任务和系统升级。

更多信息

- 华为数据存储信息中心: <https://info.support.huawei.com/storage/#/home>
- 企业业务: <https://e.huawei.com/cn/>
- 技术支持: <https://support.huawei.com/enterprise/>
- 在线学习: <https://e.huawei.com/cn/talent/portal/#/>
- 华为Support案例库: <https://support.huawei.com/enterprise/zh/knowledge?lang=zh>

Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。
Bring digital to every person, home, and
organization for a fully connected,
intelligent world.

版权所有 © 华为技术有限公司 2026。

保留一切权利。

未经华为事先书面许可，任何人不得对本资料进行复制、修改、
整合、改编、翻译，也不得将本资料或其任何部分或基于本资
料的衍生作品提供、出售、转让、再许可给任何人用于华为
许可以外的任何其他用途。

Copyright © Huawei Technologies Co., Ltd. 2026.

All Rights Reserved.

Without prior written permission of Huawei, no one is
allowed to reproduce, modify, integrate, adapt, translate
this Document, or provide, sell, transfer, sublicense this
Document or any part of this Document or its derivative to
any third party for any other purpose otherwise permitted
by Huawei.

