

闪存存储产品技术与应用



前言

- 华为闪存存储产品包含OceanStor Dorado全闪存存储和OceanStor融合存储。
- 本章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等内容。

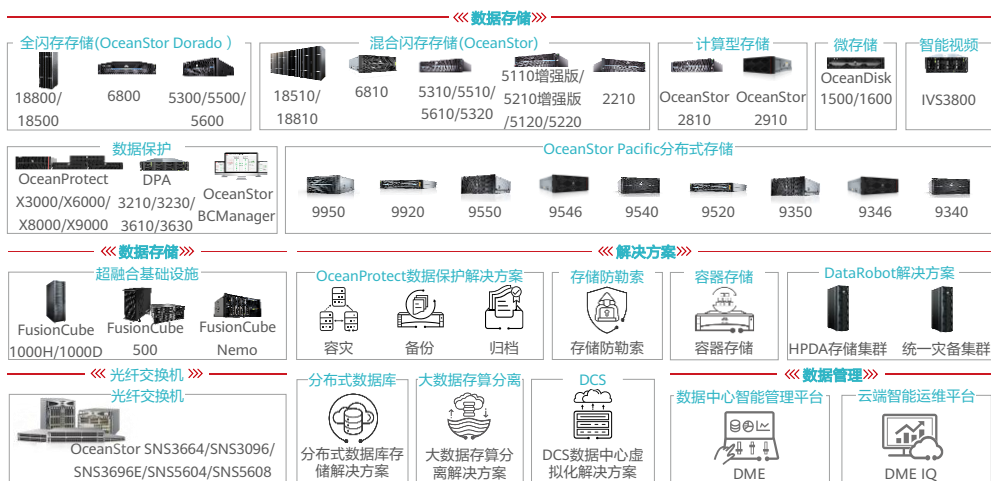
目标

- 学完本课程后，您将能够了解华为闪存存储系列产品的：
 - 产品概览
 - 软硬件架构
 - 关键技术
 - 增值特性
 - 典型应用场景

目录

1. 产品概览
2. 产品架构
3. 关键技术
4. 增值特性
5. 应用场景

数据存储全景图



全闪存存储产品概览



	中端				高端			
	OceanStor Dorado 5300	OceanStor Dorado 5500	OceanStor Dorado 5600	OceanStor Dorado 5600K (信创)	OceanStor Dorado 6800	OceanStor Dorado 18500	OceanStor Dorado 18800	OceanStor Dorado 18800K
控制框	2U/2C	2U/2C	2U/2C	2U/2C	4U/4C (2P新硬件)	4U/4C (2P新硬件)	4U/4C	4U/4C
控制器规格	2-16	2-32	2-32	2-32	2-32	2-32	2-32	2-32
最大盘位	1000	1200	1500	1500	3200	6400	9600	9600
最大有效容量 (TiB)/双控	500	1024/2048	2048/4096	2048/4096	4096/8192	4096/8192	4096/8192	4096/8192
缓存/双控	256G	384G/512G	768G/1024G	1024G	512G/1024G /2048G	512G/1024G /2048G	1024G/2048G	1024G/2048G
前端端口类型	8/16/32G FC/FC-NVMe; 10/25/40/100G Ethernet, 25/100G NVMe over RoCE (仅Dorado 55/56支持100G), 10GE电口							
后端端口类型	SAS 3.0/100G Ethernet (K系列后端只有100G)							

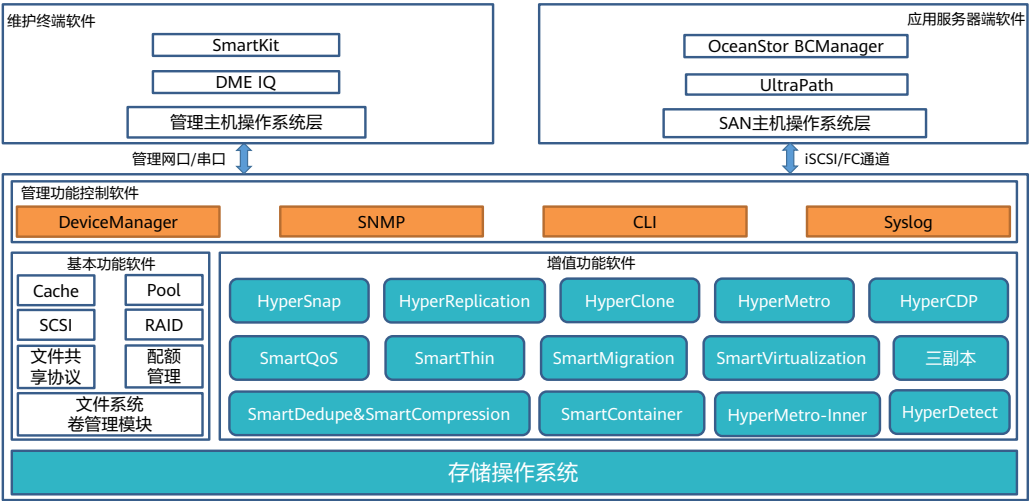
- 以2022年Q2产品信息为例。

新一代OceanStor融合存储产品概览



	中端			高端		
类型	OceanStor 5310	OceanStor 5510	OceanStor 5610	OceanStor 6810	OceanStor 18510	OceanStor 18810
控制框高度	2U/2C	2U/2C	2U/2C	4U/4C	4U/4C	4U/4C
控制框扩展	2-16	2-16	2-16	2-16	2-32	2-32
最大盘位	1200	1600	2000	3200	6400	9600
缓存/双控	128G/256G	384G/512G	768G/1024G	512G/1024G/2048G	512G/1024G/2048G	1024G/2048G
前端端口类型	8/16/32G FC/NoF(FC NVMe); 1/10/25/40/100G Ethernet; 25G NoF (NVMe over RoCE)					
后端端口类型	SAS 3.0/100Gb RDMA					

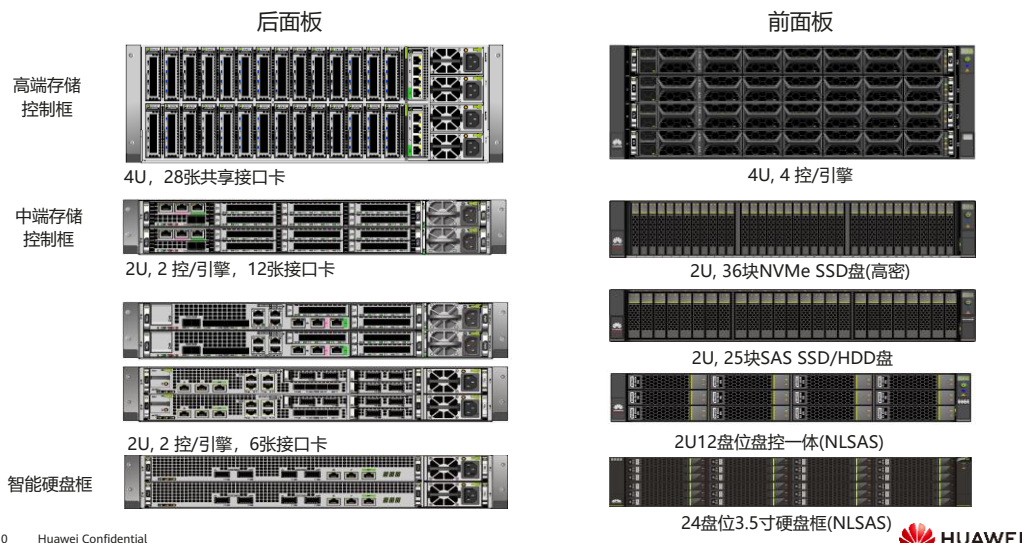
闪存存储功能架构



目录

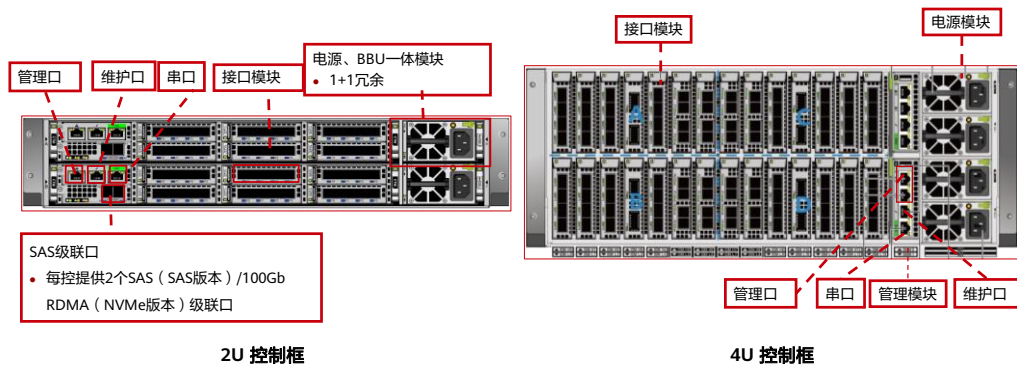
1. 产品概览
2. **产品架构**
 - **硬件架构**
 - 软件架构
3. 关键技术
4. 增值特性
5. 应用场景

灵活高密高性能硬件设计



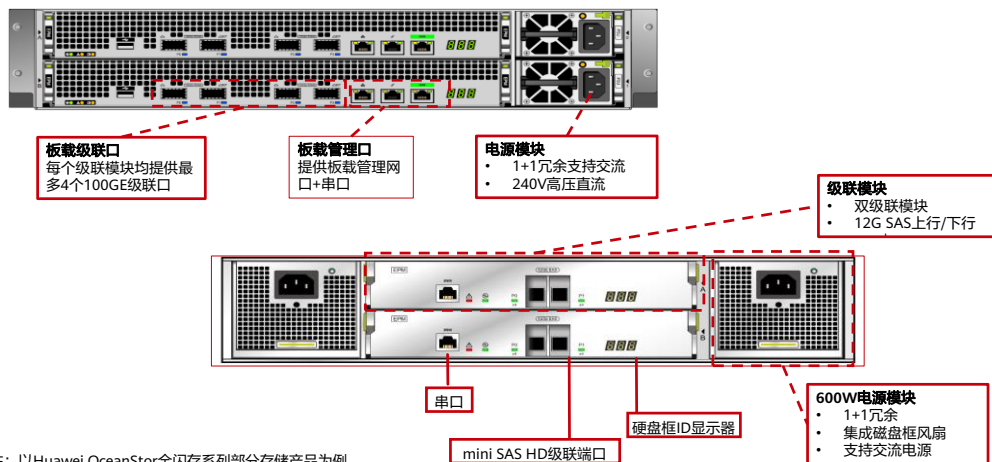
- 增加2U12盘位3.5寸盘控一体、4U24盘位NL_SAS硬盘框、高密框；
- 高端存储4U独立机头方式部署，最大支持28张接口卡，通过100Gb RDMA网络实现引擎内镜像和跨引擎互联；
- 中端存储2U盘控一体方式部署，不同型号的可插拔接口板数量有差异，最大支持12张接口卡，通过100Gb RDMA网络实现引擎内镜像，25Gb RDMA跨引擎直连，100Gb RDMA的跨引擎交换机互联；
- 智能框支持标准的25盘位SAS SSD以及高密的36盘位NVMe SSD，普通硬盘框支持标准的25盘位SAS SSD。

控制框后视图



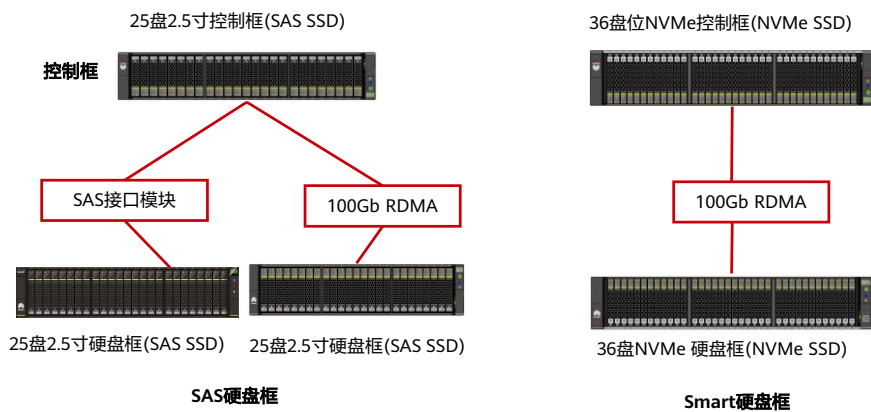
注：以Huawei OceanStor全闪存系列部分存储产品为例。

硬盘框



注：以Huawei OceanStor全闪存系列部分存储产品为例。

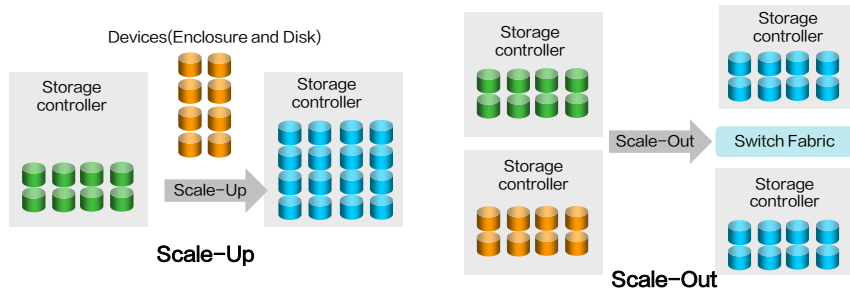
级联架构概览



注：以Huawei OceanStor全闪存系列部分存储产品为例。

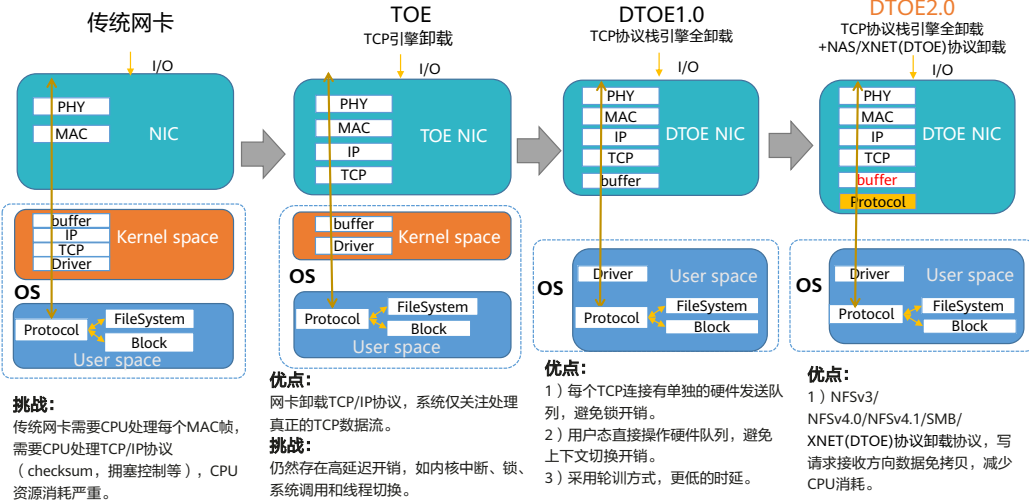
Scale-Up&Scale-Out

- 在存储系统中，Scale-Up即传统的纵向扩展架构，即利用现有的存储系统，通过不断增加存储器件来满足用户日益增长的需求；而Scale-Out横向扩展架构的升级通常是以节点为单位，通过扩展更多的控制器来满足用户日益增长的需求。

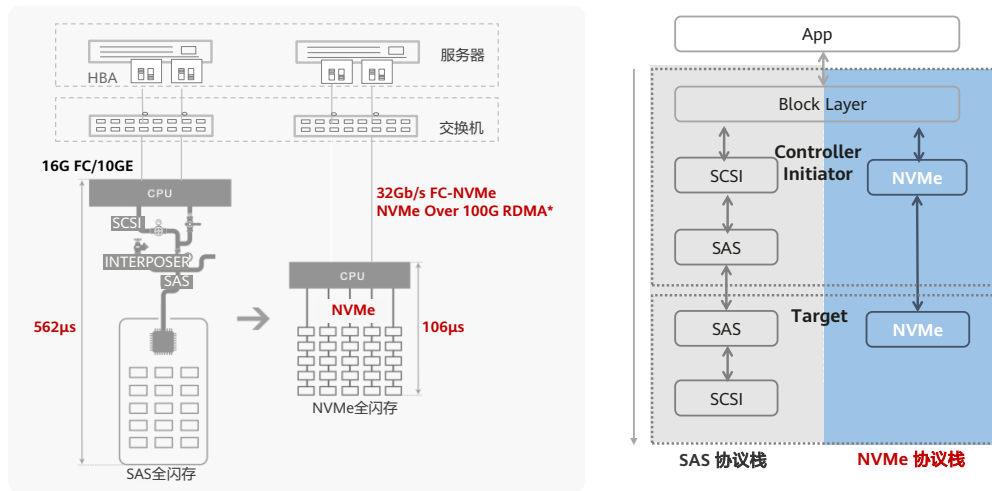


- PCIE Scale-Out是构建在PCIE协议上的Scale-Out技术，IP Scale-Out则是构建在IP协议上。
- 融合存储V3、V5系列入门级和中端产品、融合存储x10 V5 Kunpeng系列、Dorado V6系列和新融合系列使用IP Scale-Out，融合存储18000 V3、18000 V5、Dorado V3系列使用PCIE Scale-Out。

DTOE协议卸载技术



端到端NVMe架构



16 Huawei Confidential

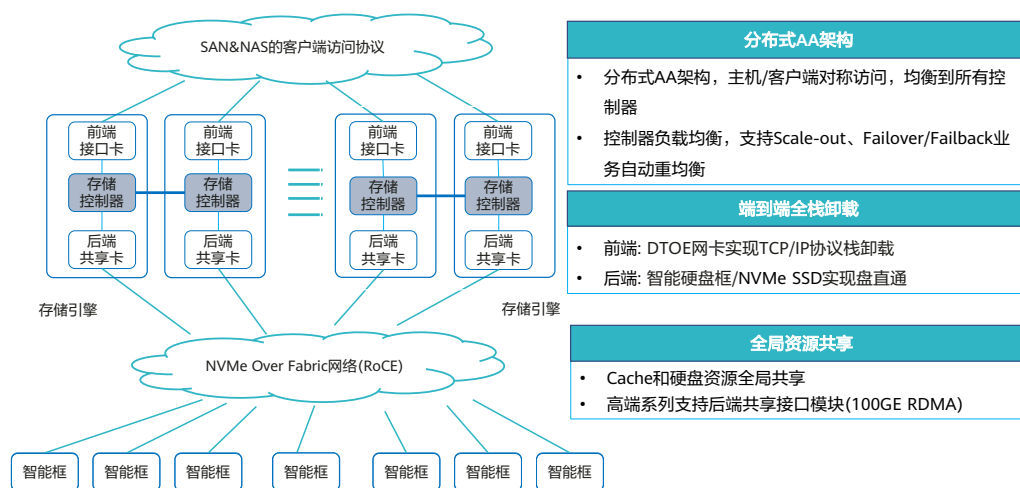


- NVMe是一种抽象协议层，提供可靠的NVMe命令和数据传输。为了支持数据中心的网络存储，通过NVMe over Fabric可以实现NVMe标准在多种网络上的扩展，达到降低存储网络协议栈处理开销并提供高并发低延时的目的，适应SSD介质驱动的存储架构演进。NVMe over Fabric支持把NVMe命令和数据映射到多个Fabric传输链路，主要包括FC、InfiniBand、RoCE v2、iWARP和TCP。
- 华为全闪存存储系统支持端到端的NVMe（部分型号不支持NVMe进行后端组网），包括：
 - 存储与主机网络支持FC-NVMe（NVMe over FC），后续将扩展到NVMe over 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。

目录

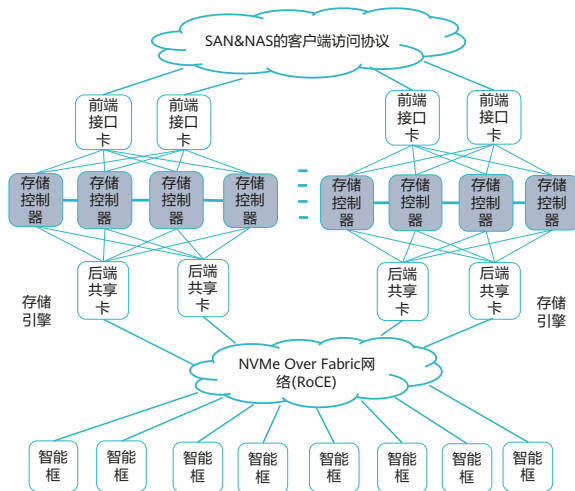
1. 产品概览
2. 产品架构
 - 硬件架构
 - **软件架构**
3. 关键技术
4. 增值特性
5. 应用场景

AA基础架构（中端&入门级）



- OceanStor 存储采用华为自研的分布式架构，单个LUN/FS可以部署运行在多控制器上，全局负载均衡的高性能和高可靠的LUN/FS服务；在整体软件架构上进行技术创新和改进，实现了真正的分布式AA架构，端到端NVMe，同时全局资源共享。

AA基础架构（高端）



分布式AA架构

- 分布式AA架构，主机/客户端对称访问，均衡到所有控制器
- 控制器负载均衡，支持Scale-out、Failover/Failback业务自动重均衡

端到端全栈卸载

- 前端: DTOE网卡实现TCP/IP协议栈卸载
- 后端: 智能硬盘框/NVMe SSD实现盘直通

全局资源共享

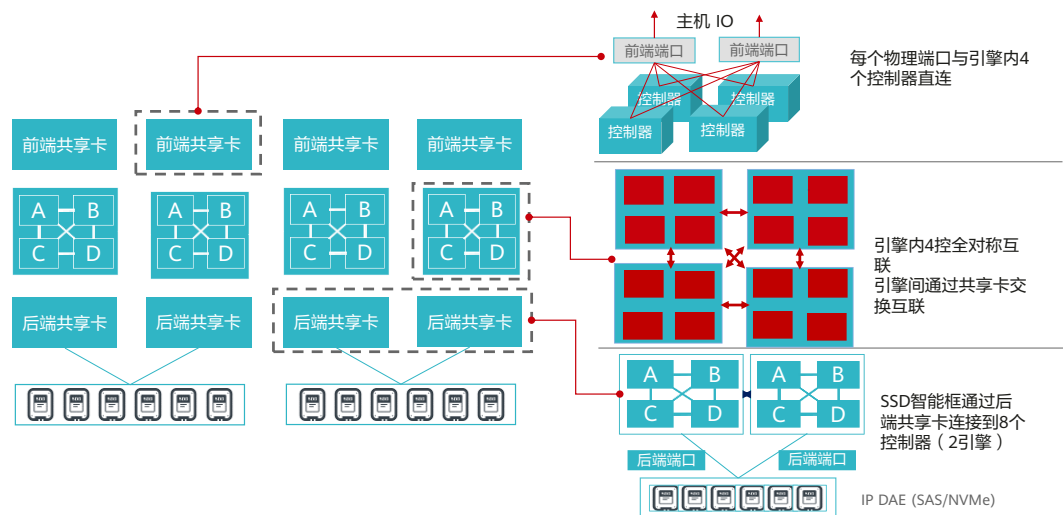
- Cache和硬盘资源全局共享
- 高端系列支持后端共享接口模块(100GE RDMA)

• SmartMatrix架构（真正的分布式AA架构）：

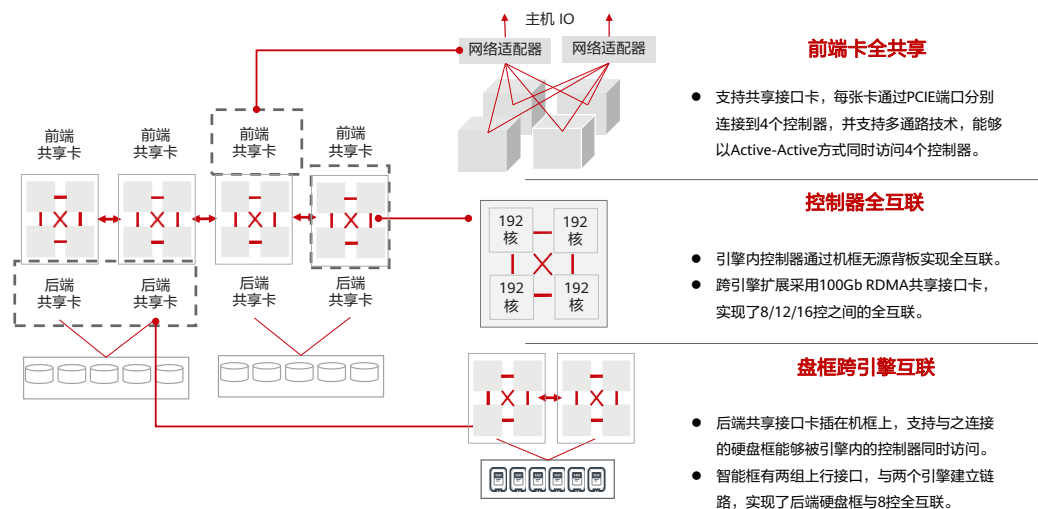
- 全系产品都采用分布式、对称AA架构，高端存储在硬件上做到全互联，软件上支持跨引擎镜像；
- LUN无归属，每个LUN的主机请求按一定粒度切片打散到所有控制器上并行处理；FS无归属，按目录打散到所有的控制器上并行处理。
- 控制框扩容时，原有LUN/FS的业务可以自动均衡到所有新老控制器上使用全局cache和全局pool，存储池将启动后台任务在所有控制框的所有盘间进行数据重排布；

- Cache、POOL全局资源共享：
 - 全局Cache：通过切片+DHT算法将每个LUN均衡打散到所有控制器上并行处理，写请求写入全局cache，读请求在全局cache进行预取和命中，cache刷盘和预取请求也在各个控制器上并行处理；
 - 全局POOL：通过RAID2.0+技术引擎创建池，LUN空间可以均衡分布到所有SSD盘上，实现后端盘能力的共享；
 - 高端存储支持前端和后端共享卡，实现引擎实现跨引擎内全互联，通过100GE RoCE实现引擎间共享后端智能框；
- 端到端NVMe：
 - 支持FC-NVMe（NVMe over FC）和NVMe over RoCE；后端通过100GE RoCE技术支持NVMe SSD；
 - 前端NOF技术上，FC-NVMe生态相对成熟度好一些（一个是为大多数AFA厂商采纳，二是更好利用旧现有企业基础设施），NVMe over RoCE实现客户all IP的网络部署，简化客户网络；

SmartMatrix全互联架构



端到端全互联架构，业务持续在线

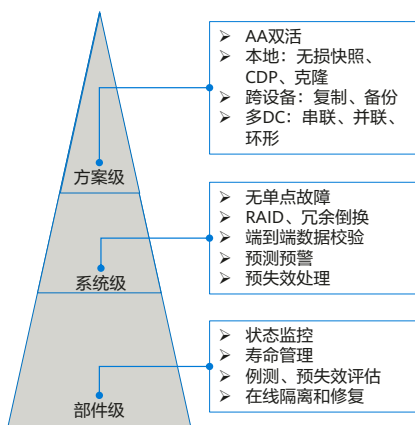


• 端到端全互联主要包括：

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。

AA高可靠性架构

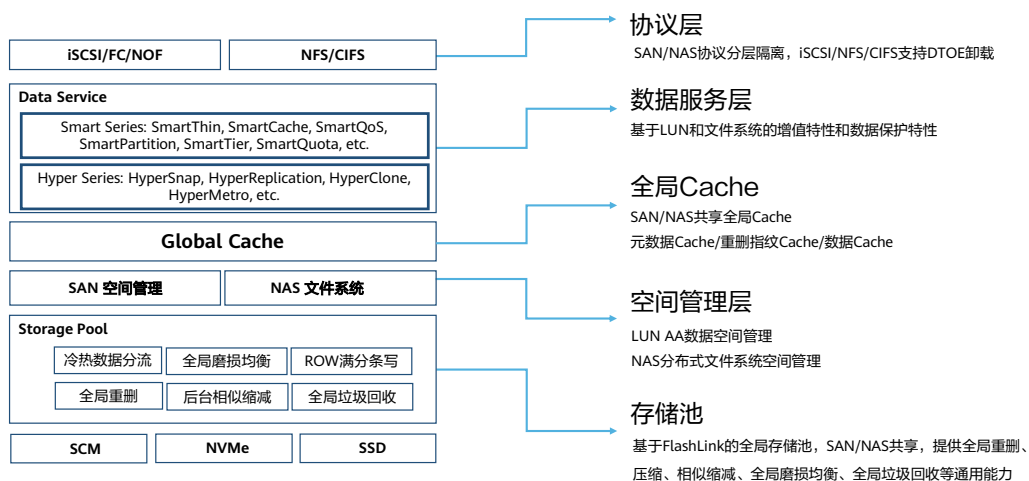
多层级可靠防护



全生命周期可靠运营



SAN/NAS一体化架构

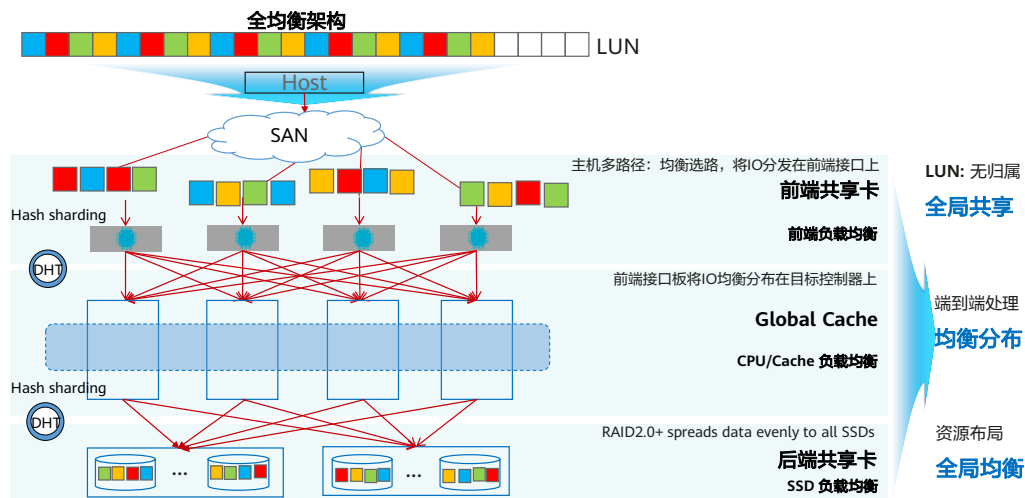


- OceanStor Dorado采用SAN/NAS一体化的AA架构设计，基于共同的存储资源池，特性共同的设计实现框架，叠加SAN和NAS不同的逻辑语义的差异构建一体化的特性服务能力。
- OceanStor Dorado的软件架构从上往下分为：协议层、数据服务层、Cache层、文件和块空间管理层以及存储池层。
- 协议层负责协议的解析和权限管理；SAN和NAS是两套的语义逻辑 因此采用不同的服务入口组件分别交付实现。
- 数据服务层实现了各种容灾备份增值特性，也是特性的框架再叠加SAN和NAS的差异化语义进行插件化的实现。
- 全局Cache为文件服务和块服务提供元数据和数据的Cache读写加速和智能预取。
- 空间管理层：SAN部分，空间管理逻辑相对简单一些；而NAS部分涉及到目录/文件的组织结构和数据布局，因此在空间管理框架上支持服务的差异化插件实现。
- 存储池提供底层的空间分配，FlashLink智能算法底座功能。

目录

1. 产品概览
2. 产品架构
- 3. 关键技术**
4. 增值特性
5. 应用场景

LUN AA均衡

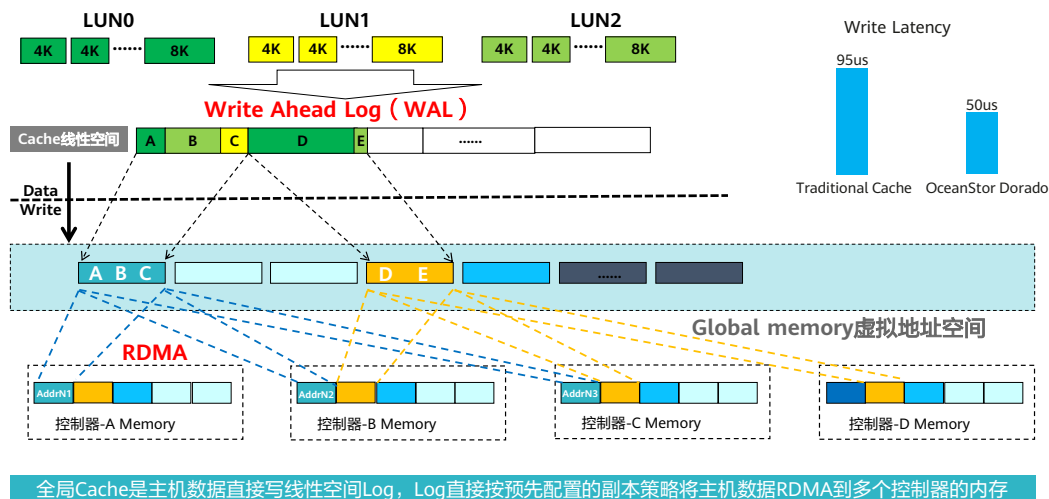


26 Huawei Confidential



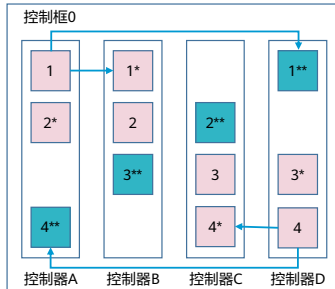
- 全系列产品LUN支持Active-Active访问，业务均衡分布在前端链路上，LUN无归属控制器，易于使用，负载均衡（采用DHT算法将LUN按固定大小均匀打散到所有控制器）。
- 高端产品提供智能共享前端接口卡，可直接将LUN的固定大小分片发送给目标控制器，降低时延。
- 主机请求可在任意控制器写入Cache缓存，任意控制器智能读缓存可预取所有LUN的数据和元数据，提升Cache命中率。
- 支持跨引擎创建存储池，采用RAID2.0+细粒度管理存储池内所有SSD盘。

全局Cache均衡

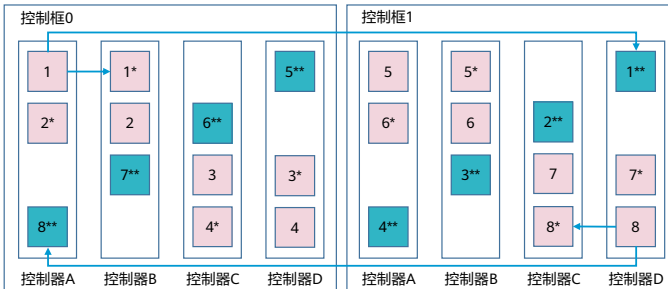


- 闪存存储把IO从前端链路均衡分发到控制器上，采用全局Cache接收数据。全局Cache是基于RDMA & WAL设计，能做到更短时延下把数据均衡到各控制器。
- 采用WAL的日志Cache机制的时延为20~50us，采用传统流程交互Cache机制时延为95us，其处理过程对比如下：
 - 日志Cache是主机数据直接写组成线性空间，直接按配置副本策略把主机数据RDMA到多个节点上；
 - 流程交互Cache机制是数据通过消息发送到远端节点，远端节点将数据写入本地Cache缓存，应答消息返回本节点，本节点再将数据写入本地Cache。
- 全局Cache采用类似操作系统虚拟内存地址空间管理方法，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。

缓存多副本



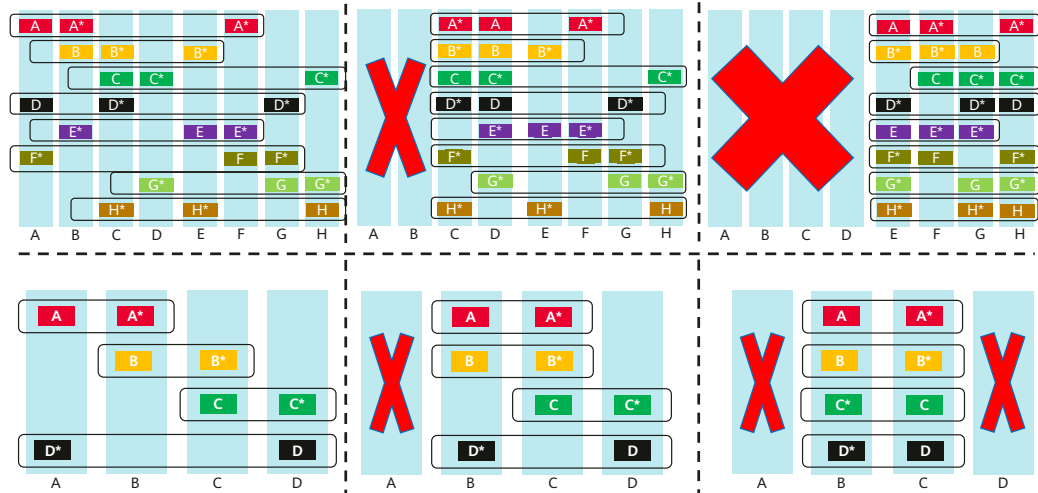
当只有一个控制框时，第三个副本在本控制框内，可以保证任意两个控制器同时故障数据不丢失；



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。

- 多副本指的是cache的数据实时镜像为3分，1主2从；主有读写权限，当主故障的时候其中给一个从会升级为主。仅限高端全NVMe形态支持3副本；中端为2副本。

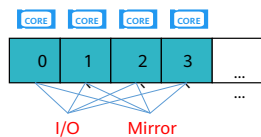
持续镜像



- 持续镜像指当主或从副本故障的情况下会从其他的控制器中选择新的来填补，快速 cacheup 恢复为副本一致性。
- 当同一个控制框内正常工作的控制器为2个或两个以上时，在控制框内持续镜像。
- 当一个控制框内故障1个控制器后只剩下最后1个控制器时，会跨控制框在其它控制框上找控制器作为副本对象，直至整个阵列只剩下一个控制器为止。

CPU多核负载均衡优化

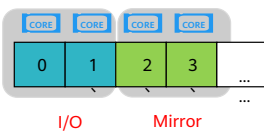
传统方式



挑战:

不同任务竞争CPU不同核的时间片，
导致数据I/O在不同核间频繁切换，导
致时延高

CPU Core分组



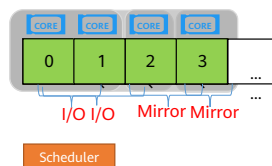
优点:

避免干扰和资源频繁切换

挑战:

不同业务，部分核过载，时延高

OceanProtect
CPU Core分组+智能调度

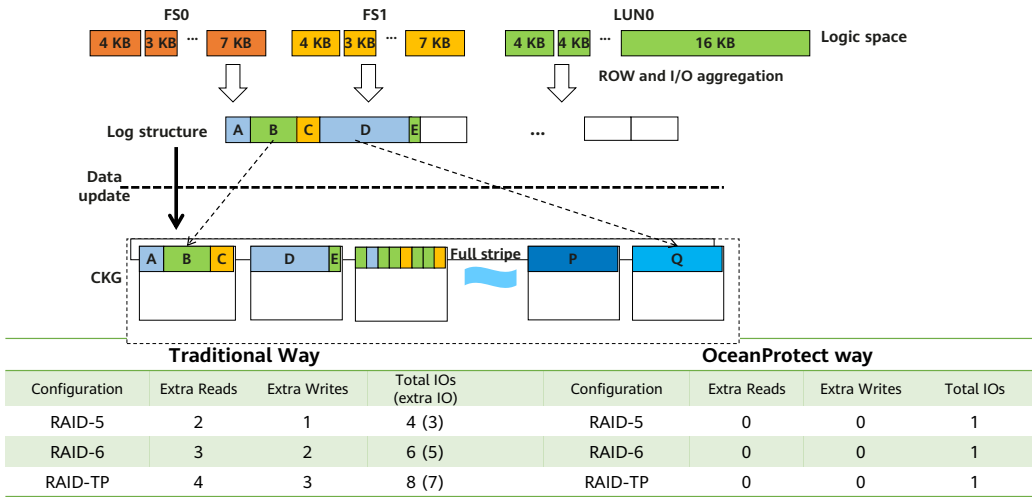


优点:

智能调度器根据业务负载状态，将任务
分发到其它核，实现负载均衡

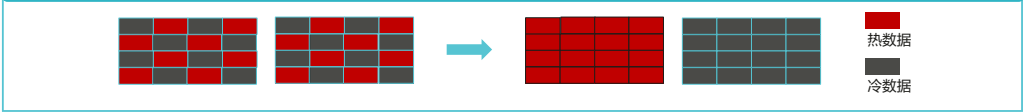
- 支持内核的自适应调度，满足高耗能群组的临时需求。

基于ROW的满分条写

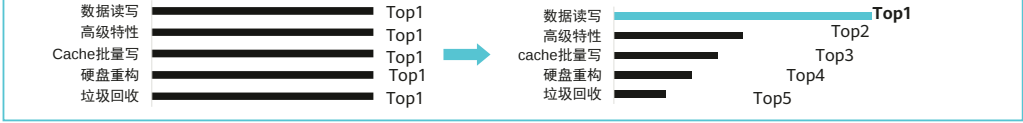


FlashLink

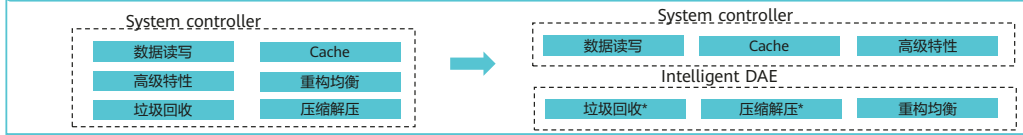
深度联动1：冷热数据分区处理，减少垃圾回收，减小性能抖动



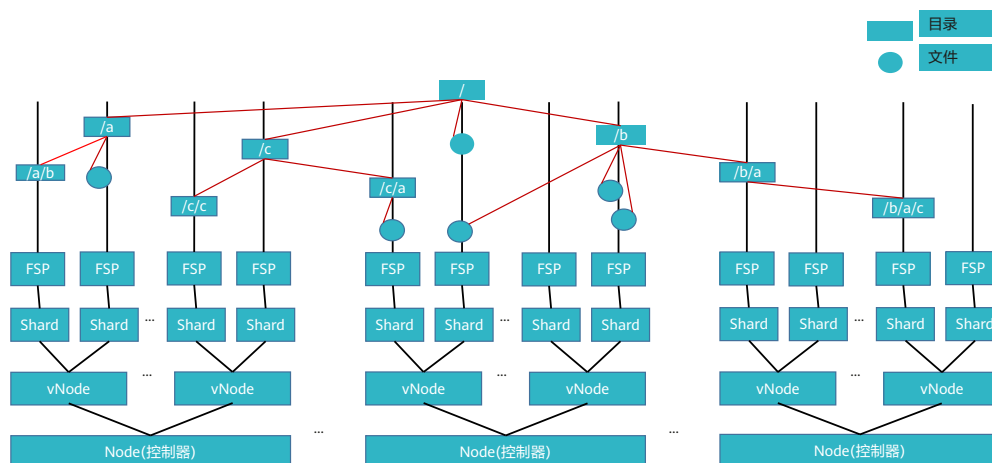
深度联动2：全局IO优先级，保障稳定低时延



深度联动3：系统后台业务卸载到智能框，提升端到端性能

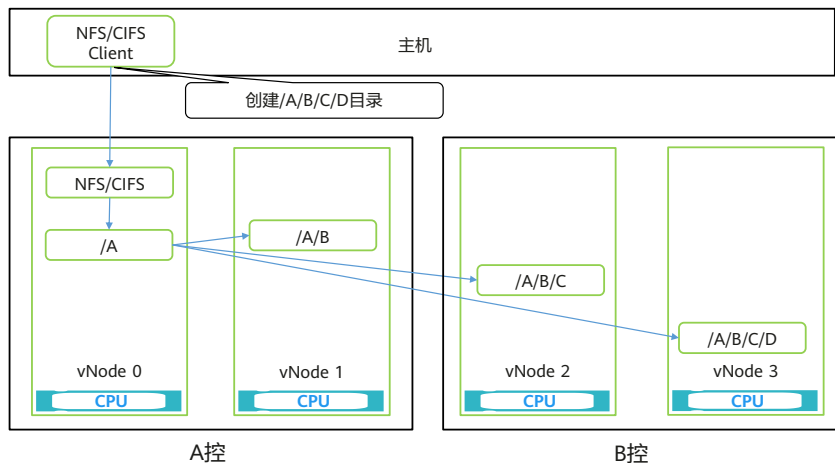


分布式文件系统



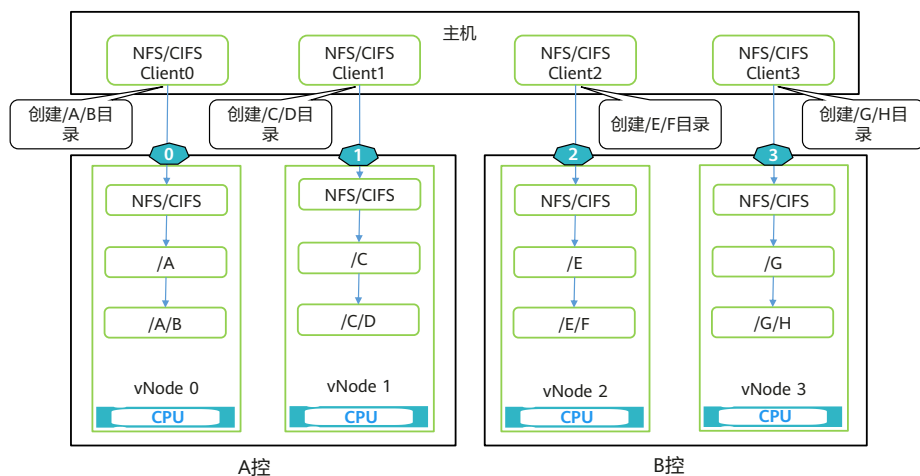
- NAS客户端创建的目录和文件均衡的写入整个存储集群。
- 数据分布式分发粒度：
 - 文件系统从架构选择上，采用了基于目录为粒度的分布式分发策略，将目录分发到多个CPU去处理，而目录下的子文件分发到相同的控制器节点，保持目录和子文件在同一个CPU处理，可以获得最佳的文件读写性能和目录处理性能。
- 分布式分发过程(创建过程)：
 - 图上的Node代表我们的一个控制器，而vNode代表控制器内的CPU，对于高端存储每个控制器我们有4颗CPU，也就是有4个vNode，而中端存储就是每控2个vNode。
 - 在vNode之上，我们设计了一个叫做FSP（Filesystem Partition）的逻辑分区，用于承载文件系统的目录和文件，FSP是均衡分布在所有控制器的vNode上面的，当存储系统从双控扩展到更多控制器，FSP会自动均衡排布到新的控制器，从而使不管多少vNode，上面的FSP数量都是一样的，这样多控多CPU的文件和目录处理就均衡了。
 - 将目录和子文件设置到不同的FSP中去运行的过程，只要目录和文件能均衡到所有FSP中，FSP会自动保证能均衡到多控的多个vNode上去处理。
- 目录创建分发模式：
 - OceanStor闪存存储支持两种分布式分发模式，均衡模式和性能模式，其中均衡模式适合绝大多数场景。

文件系统访问均衡模式



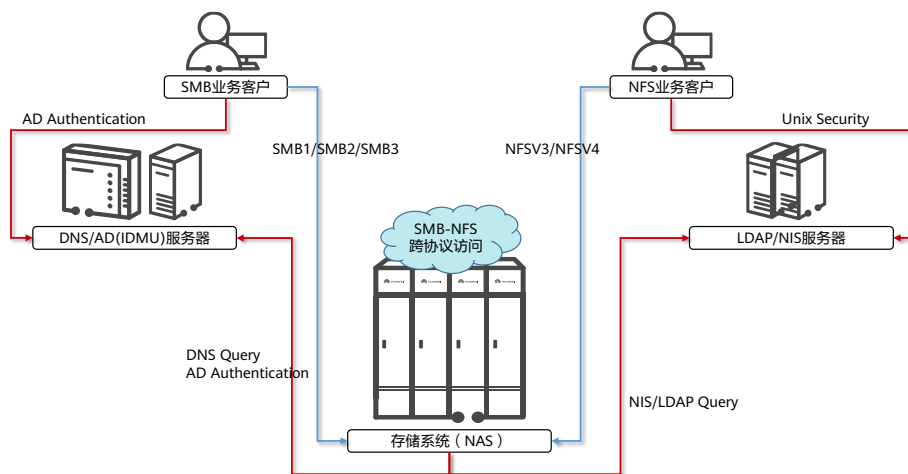
- 主机Client通过IP访问控制器，以目录为粒度在多个控制器的vNode间均衡，100%发挥存储性能。
- 主机Client通过IP地址与存储控制器连接，Client创建的目录均衡分布在多个vNode上；文件跟着目录也均衡到对应的vNode上。

文件系统访问性能模式

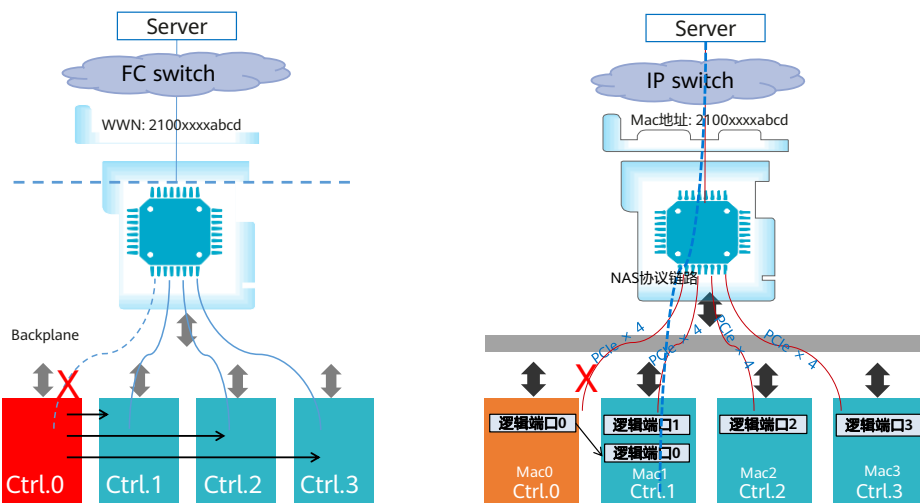


- 主机Client通过IP访问控制器，其目录/文件在本控制器亲和处理。
- 主机ClientX通过IP地址X与存储控制器连接，Client创建的目录/文件在接入控制器的vNode上均衡分布，同时相同IP创建的目录/文件保持在本vNode的亲和性，减少跨vNode或跨控制访问。
- 不同vNode的容量相差5%以上时，会将新创建目录均衡到其它vNode上，达成容量均衡。

跨协议访问



接口模块/链路冗余保护



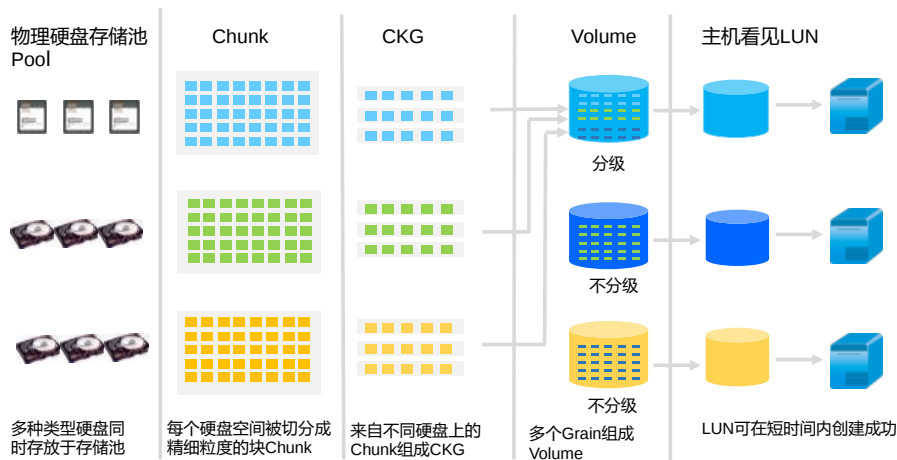
• FC场景

- 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。

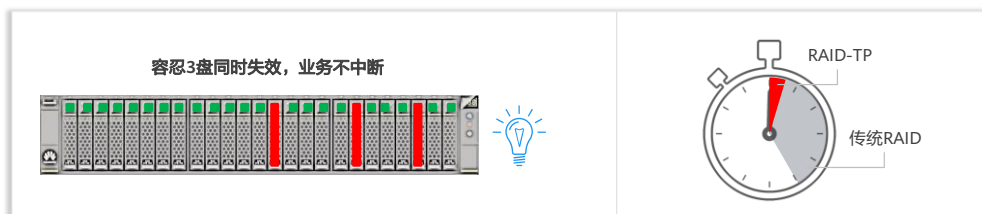
RAID2.0+



- 多块硬盘组成一个存储池。
- 存储系统将存储池中的每个硬盘划分为固定大小的CK（通常为4MB）进行逻辑空间管理。
- 来自不同硬盘的CHUNK按照用户设置的“RAID策略”组成CKG。
- CKG再被划分为更细粒度的Grain（通常为8KB），为Voulume使用的最小粒度。
- 用户创建LUN时，存储系统以Grain为单位映射到LUN，实现对存储资源的精细化管理。

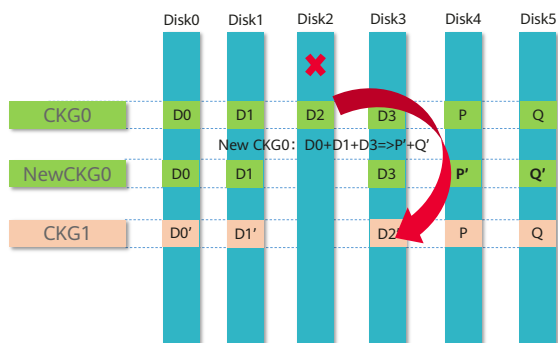
RAID-TP

- 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%。

缩列重构



- **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

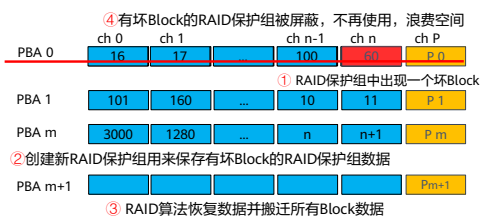
- 1、Disk2故障后, 硬盘域内可用盘数少于RAID成员盘数时, 触发缩列重构;
- 2、数据D0+D1+D3缩列计算出校验列P'Q' 构成NewCKG0;
- 3、故障盘数据D2新写入成D2'与D0',D1',Q组成CKG1 **缩列重构后, RAID冗余级别不变**

- 动态RAID重构技术支持两种重构方式: 普通重构和缩列重构。系统自动选择重构方式, 当硬盘域可用盘数小于RAID成员盘数时, 智能自动缩小数据列, 保证各种场景下RAID冗余度不降低, 数据高可靠。

盘内动态RAID

盘内普通RAID算法

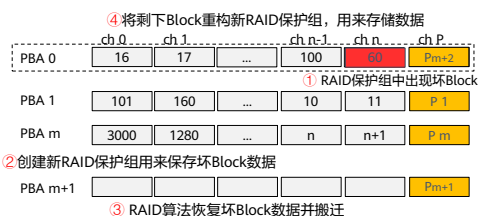
- RAID保护组中坏一个Block后，恢复坏Block数据，并搬迁整个RAID保护组数据，再屏蔽该RAID保护组空间



问题：大量可用的Flash空间被浪费

华为盘内动态RAID算法

- RAID保护组中坏一个Block后，恢复坏Block数据并搬迁，同时屏蔽坏Block，用剩下的Block重构新RAID保护组



效果：充分有效利用所有Flash空间

- 即针对盘内的数据，以物理Die为单位建立RAID，降低静默错误导致数据丢失的概率。
- HSSD内的每个颗粒由多个物理Die构成，针对下盘的数据，通过RAID4形成冗余，保证单个Die失效，盘内数据不丢失。业界SSD的不可恢复错误率(uber)的为10-17左右，而通过支持盘内RAID，HSSD将出现不可恢复错误率(UBER)降低到10-18(下降一个数量级)。例如：不带盘内RAID，当对盘读写11PB的数据就会出现1个坏块，而如果带盘内RAID，往盘上写入110PB的数据才能出现1次坏块，将出现坏块的概率降低到以前的十分之一。

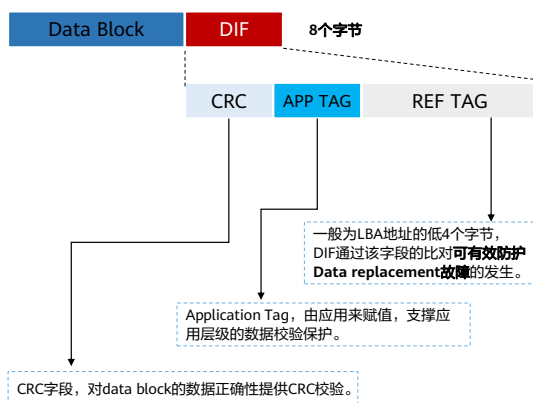
盘内DIF

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



框级RAID

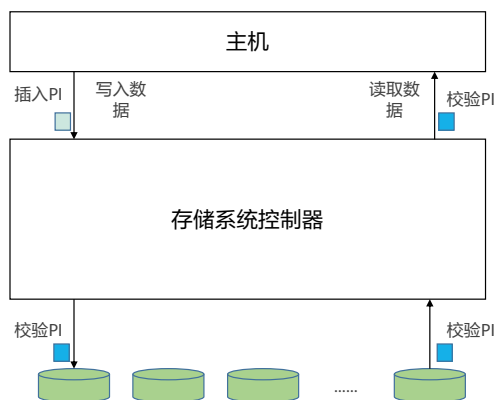
- 框级冗余在满足盘级冗余的前提下，考虑硬盘和硬盘框的归属关系。每次组建 Chunk组时从每个硬盘框内最多选2块盘的Chunk组成Chunk组，Chunk组的Chunk数量由硬盘框的数量决定。
- 以系统有4个框，每个框3盘的配置，采用RAID6为例，相同颜色的Chunk组成1个Chunk组。冗余盘数为2，因此框故障后，每个Chunk组只丢失2份数据，未超出Chunk组的冗余数量，业务不中断。



- 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。

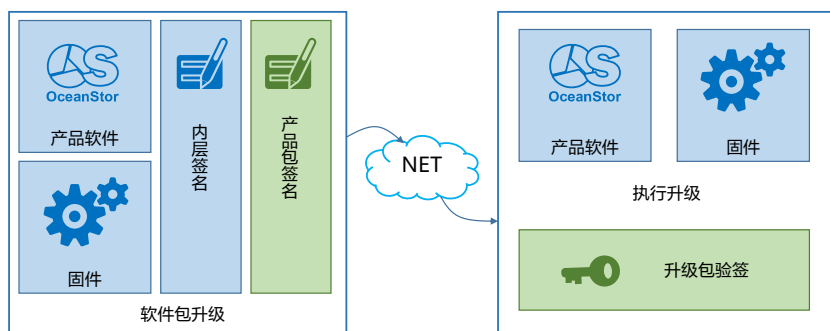
端到端PI (Protection Information)

- 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。

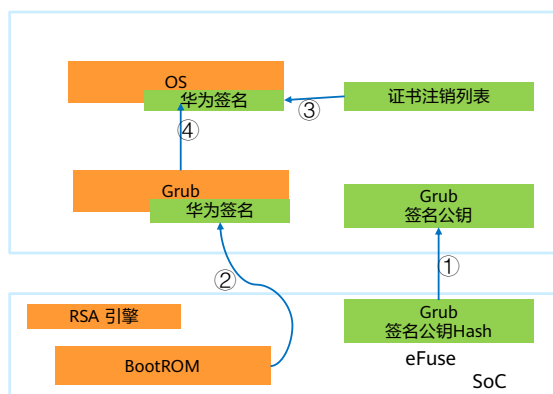


软件完整性保护

- 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。



硬件可信安全启动

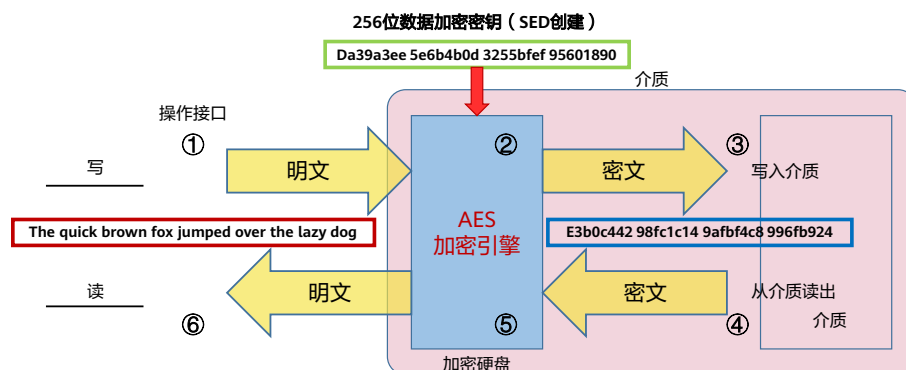


安全启动的软件验证、加载过程:

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

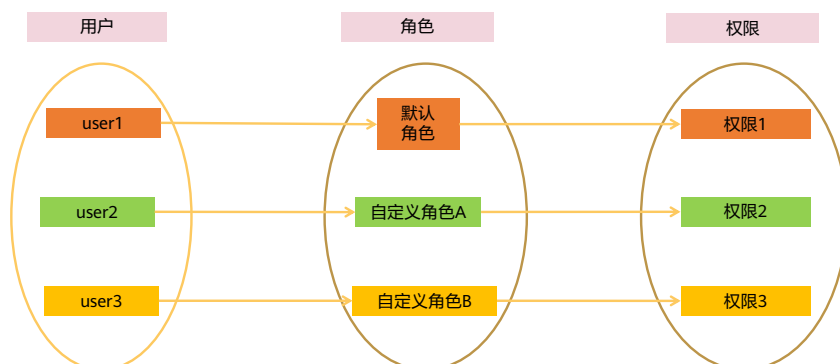
- 设备上电启动后，从初始启动模块开始启动，并逐级先验证，验证通过再启动。通过数字签名校验固件的完整性，防止固件和操作系统被篡改。
- 相关技术如下：
 - 可信根RoT(Root of Trust)，集成到CPU内部，可防止软件攻击，以及物理攻击，安全等级业界最高。
 - 通过两级数字签名（根密钥+二级密钥）保证软件完整性、通过数字证书保证软件唯一性。
 - 采用业界安全等级最高的算法之一：RSA 2048/4096。
 - 支持二级密钥（代码签名密钥）吊销。
 - 处理器内置RoT，可防止ARM片外Flash固件篡改、系统盘替换等恶意启动。

数据加密



- 通过配置SED和内置密管或者外置密管，和存储系统配合完成数据静态加密，从而保证数据的安全性。
- 内置密管：存储系统自带的密钥管理应用，采用NIST SP 800-57最佳实践设计，对密钥进行生命周期管理，包括密钥分层保护、密钥备份、密钥恢复、密钥产生、密钥更新、密钥销毁。内置密管具有易部署、易配置、易管理的特点；如果数据中心的密钥管理仅用于存储系统，且无需通过密码模块安全标准FIPS 140-2的认证，建议选择内置密管。
- 外置密管：采用KMIP+TLS的标准协议，满足FIPS 140-2 SafeNet的Key Secure密管。如果数据中心多场景需要密钥集中管理，或者需要通过密码模块安全标准FIPS 140-2的认证，建议选择外置密管。

基于角色权限管理



权限=功能1: 对象1,对象2; 功能2: 对象3, 对象4;
(一组或者多组功能和对象的关系)

- 预置默认角色：系统预置系统管理用户和租户管理用户的默认值角色。
- 系统管理用户默认角色：
 - 超级管理员 (super_administrator)
 - 管理员(administrator)
 - 安全管理员 (security_administrator)
 - SAN管理员 (san_administrator)
 - NAS管理员(nas_administrator)
 - 数据保护管理员 (dataProtection_administrator)
 - 网络管理员 (network_administrator)
- 租户管理用户角色：
 - 租户管理员(vStore_administrator)
 - 租户数据保护管理员(vStore_dataProtection)
 - 租户协议管理员 (vStore_protocol)
- 自定义角色：支持用户依据自己场景自定义角色。用户自定义角色时可以选择多个功能，每个功能可以选择多个对象。自定义角色支持删除和修改。

目录

1. 产品概览
2. 产品架构
3. 关键技术
- 4. 增值特性**
5. 应用场景

目录

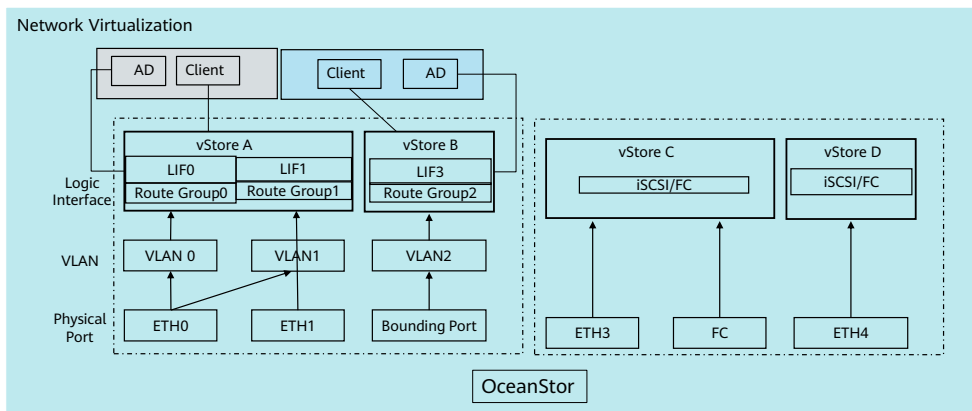
4. 增值特性

- 效率提升特性
 - **SmartMulti-Tenant**
 - SmartDedupe&SmartCompression
- 数据保护特性

特性简介

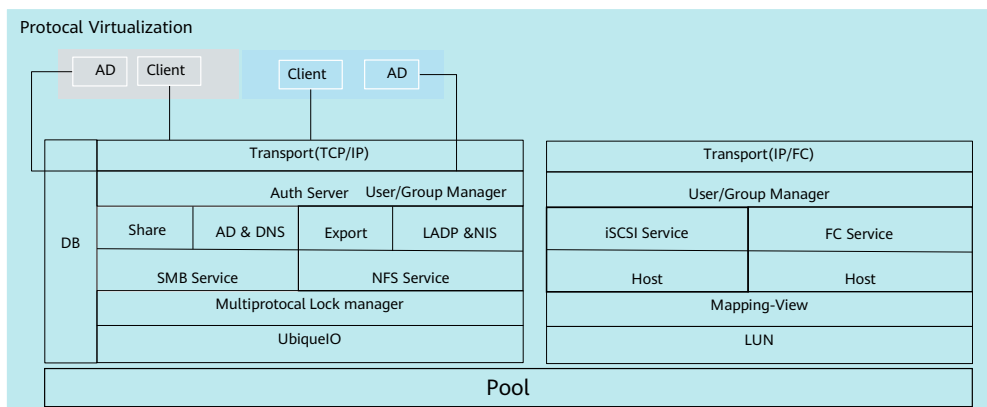
- 多租户特性又称SmartMulti-Tenant，将物理存储系统分为多个虚拟存储系统的功能，以便为不同客户端提供服务并提高存储利用率。
- 在支持多租户特性的存储设备中，存储池、LUN、LUN组、主机、主机组、端口组、NAS的文件系统（包含Dtree和配额配置），认证用户（包含UNIX、Windows本地认证用户和本地认证用户组），共享，域配置，LIF（逻辑接口）、FC端口、QoS策略、快照、复制、双活等逻辑资源均与租户建立归属关联，以实现资源隔离，提高安全性。

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、快照等资源。
- 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UNIX认证用户、Window认证用户、用户映射规则、NFS共享、CIFS共享、AD域、NIS域、LDAP域、DNS服务等。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，从而确保数据均是按租户隔离的。

目录

4. 增值特性

- 效率提升特性
 - SmartMulti-Tenant
 - **SmartDedupe&SmartCompression**
- 数据保护特性

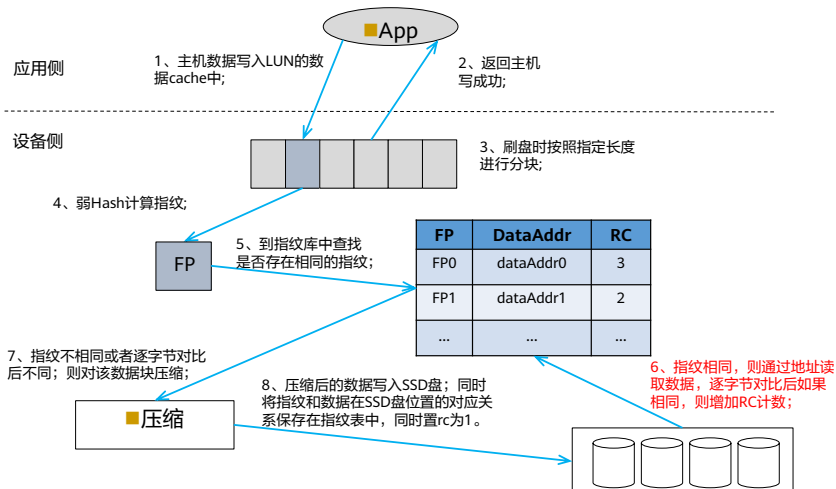
重删

- 重复数据删除（重删）：是一种节约数据存储空间的技术。在计算机中存储了很多重复数据，这些数据占用了大量硬盘空间，利用重复数据删除技术，可以只存储一份数据。重复数据删除是在比较大的范围内查找大块的重复数据，一般重复数据块尺寸在4KB以上。重复数据删除技术被广泛应用于网络硬盘、电子邮件、磁盘备份介质设备等。
- 重删的分类：
 - 在线重删：数据在写入存储介质的过程中进行重删。
 - 后台重删：数据在写入存储介质持久化后，再读取进行重删。
 - 定长重删：数据按照固定粒度分块之后再进行重删。
 - 变长重删：数据按照自身内容划分成大小不固定的块进行重删，主要应用在备份领域。
 - 相似重删：数据按照固定粒度分块，进行相似度分析，将相同的数据块进行重删，不相同但相似度高的数据块进行合并压缩或差量压缩；

压缩

- 数据压缩（压缩）：在计算机科学和信息论中，数据压缩或者源编码是按照特定的编码机制用比未经编码少的数据比特（或者其它信息相关的单位）表示信息的过程。
- 压缩的分类：
 - 在线压缩：数据写入存储介质的过程中进行压缩。
 - 后台压缩：数据流写入存储介质后，再读取出来进行压缩。
 - 软件压缩：利用系统的CPU执行压缩算法的过程。
 - 硬件压缩：将压缩算法逻辑固化到硬件设备，如ARM CPU的加速引擎/FPGA/ASIC等，在使用时调用硬件设备提供的压缩接口的过程。
 - 有损压缩：压缩后不能完全恢复成原始数据的压缩方式，常见用于处理声音、视频、图像。
 - 无损压缩：压缩后能通过解压缩完全恢复成原始数据的压缩方式。

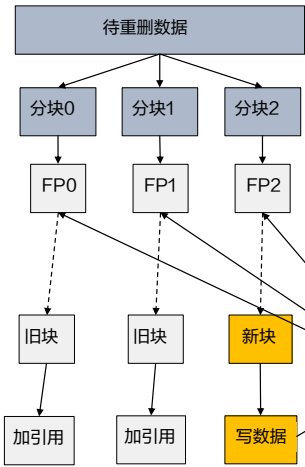
自适应重删压缩处理流程



- 闪存存储系统支撑LUN和文件系统的数据缩减, 提供在线重删和相似重删结合的自适应重删策略。
 - 用户写入数据时, 重删自适应预判算法根据业务数据特征, 识别出适合在线重删的数据, 则直接进行在线重删处理。
 - 重删自适应预判算法根据业务数据特征, 识别出适合相似重删的数据, 则计算数据的相似指纹信息 (Similar Fingerprint, SFP), 并把SFP插入相似重删机会表。然后将用户数据压缩处理后再写到存储池中, 并返回成功。
 - 后台运行重删任务将从机会表中找出有相似度的数据, 从盘上读取这些SFP对应的数据进行相似重删处理, 成功处理完后更新指纹表。

在线重删

- 1、对数据进行分块。
- 2、对分块数据计算指纹。
- 3、到指纹库查找指纹是否存在。
- 4、存在即为旧块，不存在即为新块。



- 5、旧块加引用计数，返回已经存在的数据地址；新块则写入存储空间。
- 6、将新块的指纹和保存的地址，这一映射关系加入指纹库中。

FP	DataAddr	rc
FP0=FPx	dataAddr0	+1
FP1=FPy	dataAddr1	+1
FP2	dataAddr2	1
FPx	dataAddrx	1->2
Fpy	dataAddry	1->2

后台相似重删

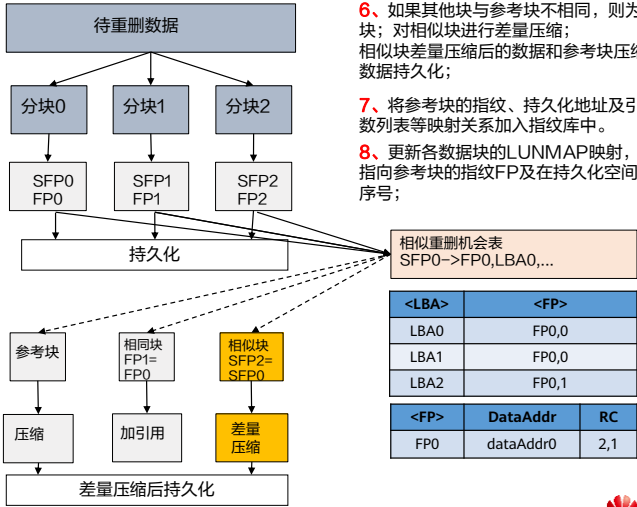
1、对数据进行分块。

2、对分块数据计算相似指纹和重删指纹，写入到相似重删机会表。

3、后台对相似重删机会表进行分析，对SFP相同的数据块进行相似重删；

4、选定参考块进行压缩；其他块与参考块进行比较；

5、如果其他块与参考块的重删指纹相同且逐字节比对一致，则为相同块；对相同块加引用计数；

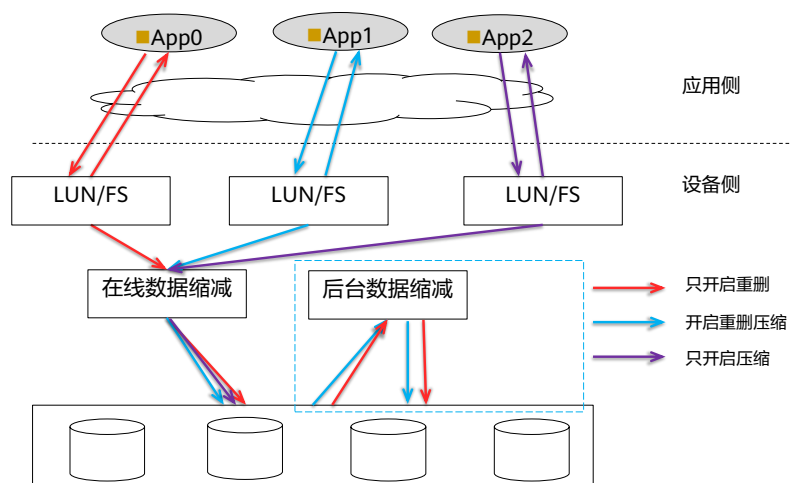


6、如果其他块与参考块不相同，则为相似块；对相似块进行差量压缩；相似块差量压缩后的数据和参考块压缩后的数据持久化；

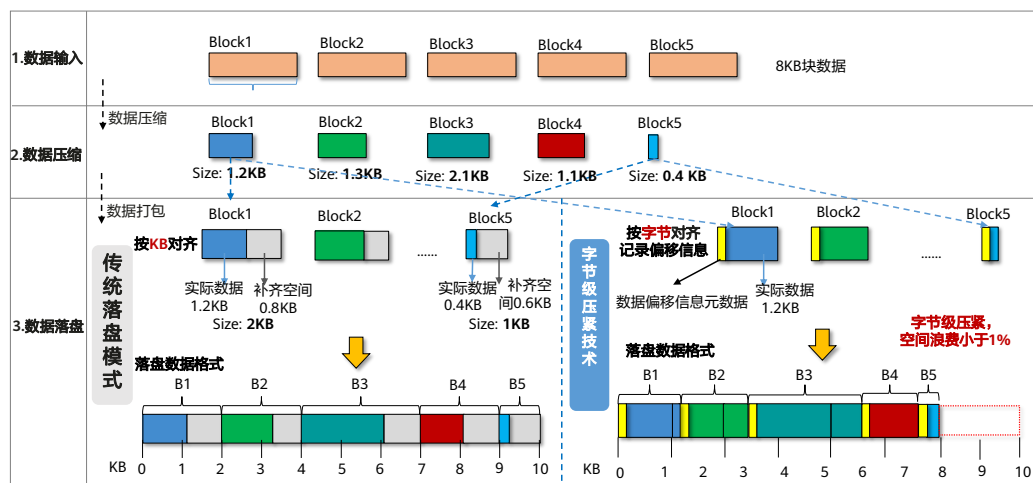
7、将参考块的指纹、持久化地址及引用计数列表等映射关系加入指纹库中。

8、更新各数据块的LUNMAP映射，LBA指向参考块的指纹FP及在持久化空间上的序号；

数据缩减



字节级压紧技术



- 闪存存储和专用备份存储都采用字节级排布技术，在不增加元数据开销的情况下，使得空间浪费小于1%。
- 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。

DIF重排



- 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。

目录

4. 增值特性

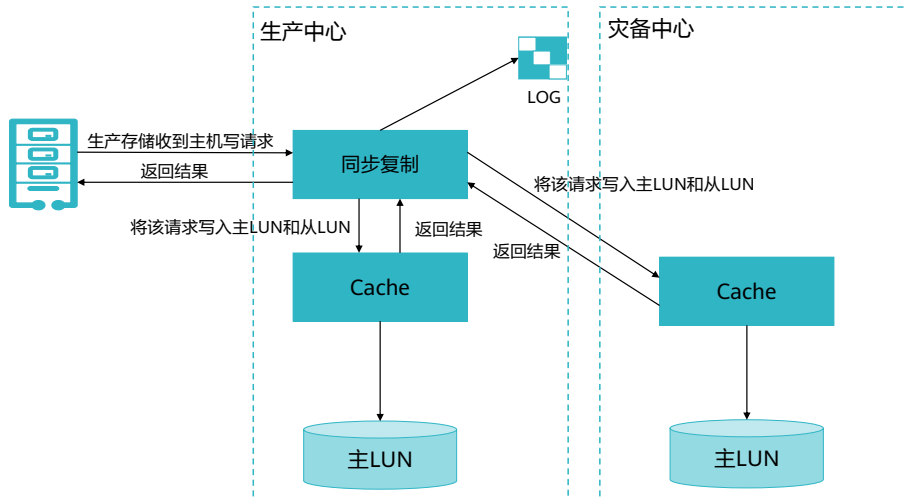
- 效率提升特性
- 数据保护特性
 - **HyperReplication**
 - HyperMetro
 - HyperDetect

简介

- 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。

LUN同步远程复制原理

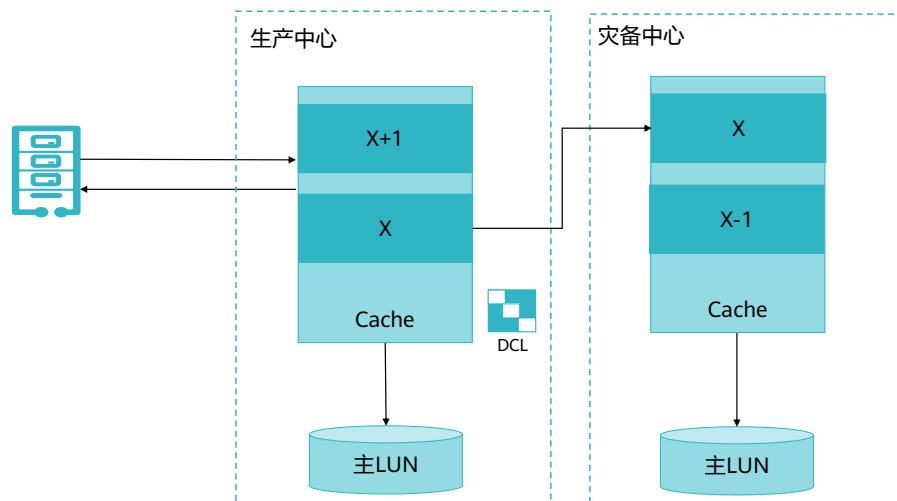


65 Huawei Confidential



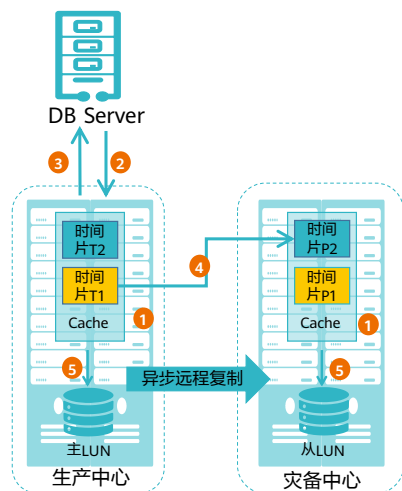
- 初始同步：生产中心的主LUN和灾备中心的从LUN建立同步远程复制关系，启动初始同步。
 - 将主LUN数据全量拷贝到从LUN。
 - 初始同步中主LUN收到主机写请求也会同样写到从LUN。
- 双写状态：初始同步完成以后，进入正常状态，此时主、从LUN数据相同。
 - 生产存储收到主机写请求。HyperReplication将该请求记录日志。日志中只记录地址信息，不记录数据内容。
 - 将该请求写入主LUN和从LUN。通常情况下LUN是回写状态，数据会写入Cache。
 - HyperReplication等待主LUN和从LUN的写处理结果都返回，如果写从LUN超时或失败时远程复制关系断开。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。
 - 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



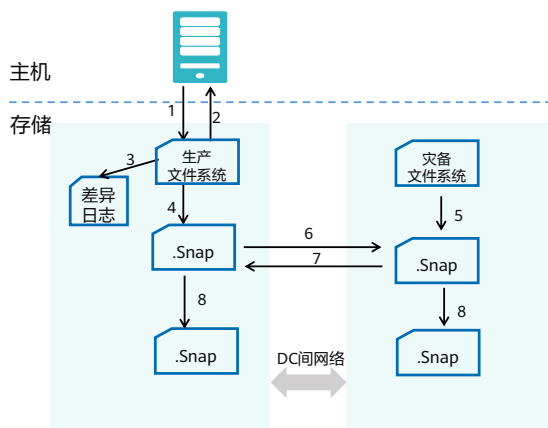
最小3秒一个一致性点

- 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**。

NAS异步复制



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:

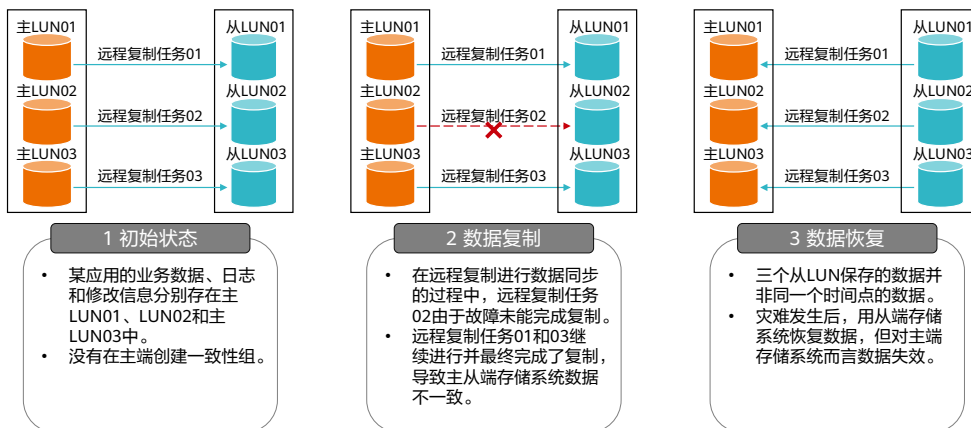
- 1、客户端写生产文件系统。
- 2、写本地完成,回主机IO。
- 3、cache刷盘时后台记录写差异日志。

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。

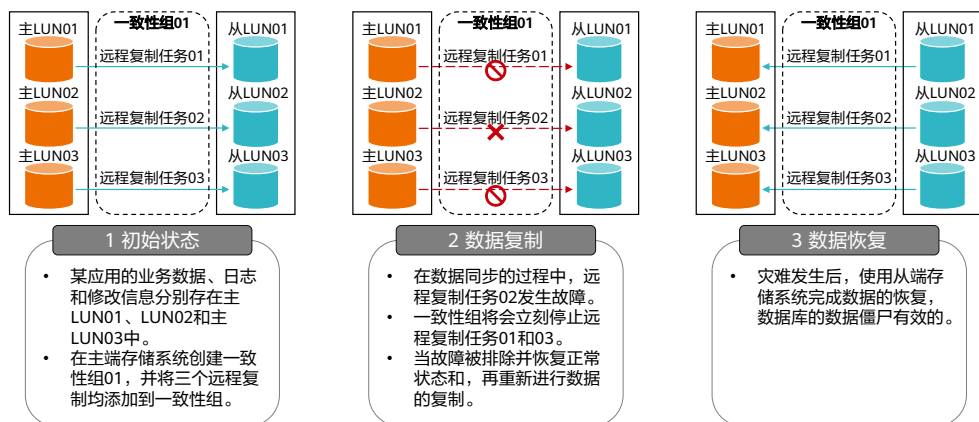
- HyperReplication文件系统异步远程复制采用基于对象层的方式进行数据复制,最短RPO支持15秒。
- 文件系统的所有内容,比如文件、文件属性、目录,都是由对象构成。
- 基于对象层的复制直接将对象从主文件系统复制到从文件系统,不需要关心复杂的文件层的信息,比如文件与目录间的依赖关系、各种文件操作,从而使复制变得更加简单高效。
- 支持高中低设备型号互通,在线升级时不用中断复制关系。

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Pair运行状态，可以及时地根据状态取值对远程复制进行同步、分裂、主从切换等操作，并在操作结束后判断操作是否成功。本页以LUN为例说明，文件系统同理。

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。

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

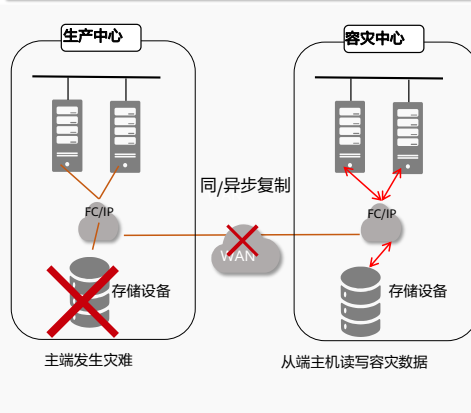
- 用户需要使用从LUN进行数据分析、挖掘，但不愿意影响主LUN业务。
- 当生产阵列故障，容灾阵列需要接管业务，但无法进行正常的主从切换。

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

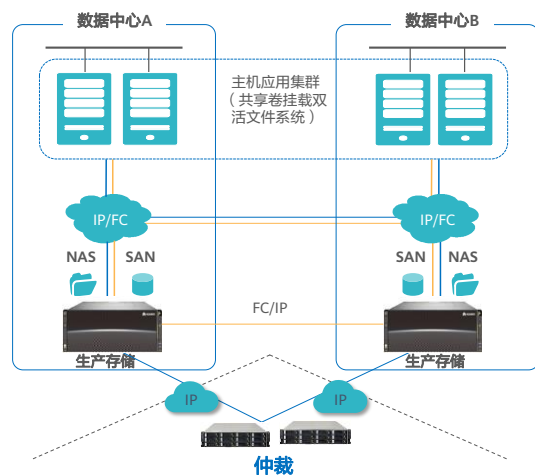


目录

4. 增值特性

- 效率提升特性
- 数据保护特性
 - HyperReplication
 - **HyperMetro**
 - HyperDetect

SAN&NAS一体化双活



基本特征

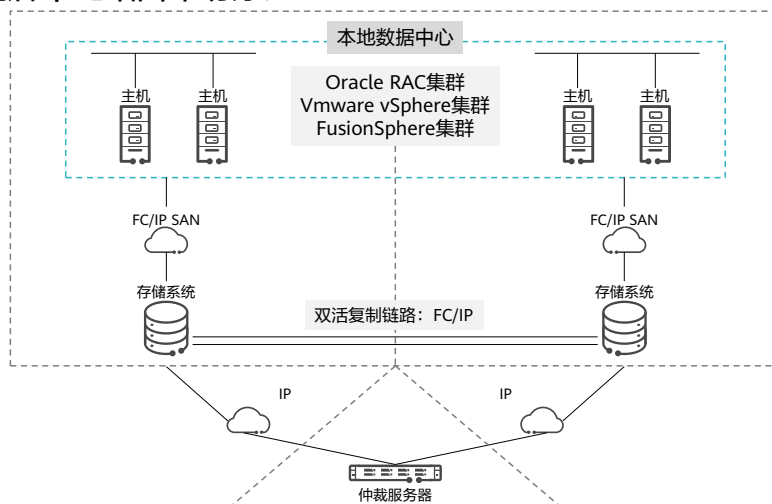
- AA双活，RPO=0，RTO≈0。
- 设备双活，免网关；阵列间单网络，易部署。
- SAN&NAS一体化双活，共享仲裁。

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%

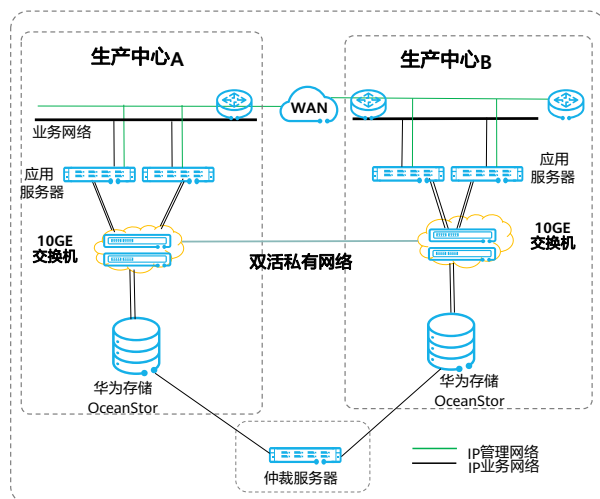
- 华为HyperMetro特性又称双活特性（以下统称双活特性），双活特性中的两个数据中心互为备份，且都处于运行状态。当一个数据中心发生设备故障，甚至数据中心整体故障时，业务自动切换到另一个数据中心，解决了传统灾备业务无法自动切换的问题。提供给用户高级别的数据可靠性以及业务连续性的同时，提高存储系统的资源利用率。
- 华为提出以OceanStor为基础的存储双活架构，为客户建设业务不间断运行的解决方案，包括本地数据中心部署逻辑架构和跨数据中心部署逻辑架构。

本地数据中心部署场景



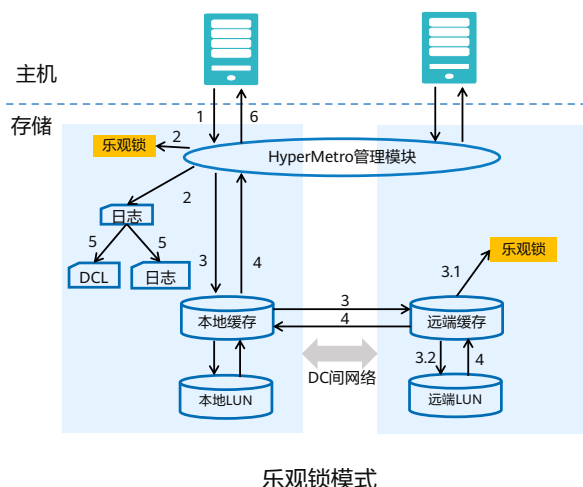
- 本地数据中心部署
 - 通常部署在同园区的不同机房中。
 - 主机与存储间通常通过交换机方式通信，对于NAS文件系统，需要采用IP交换机。同时，存储两端通过双活复制网络连接，保证双活业务的运行。

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。

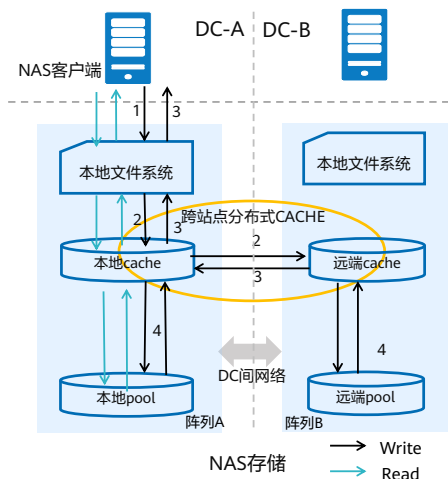
SAN: 高性能乐观锁



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。

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。

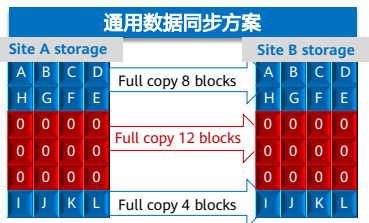
双活读:

- 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, 有分布式cache实现一致性保证。
- 2、跨阵列之间数据同步, 实时, 且一次用户数据传输。
- 3、基于租户的双活配置, 网络和鉴权配置自动同步。

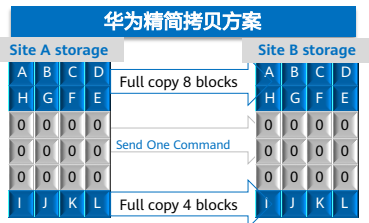
快速初始/增量数据同步



■ 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

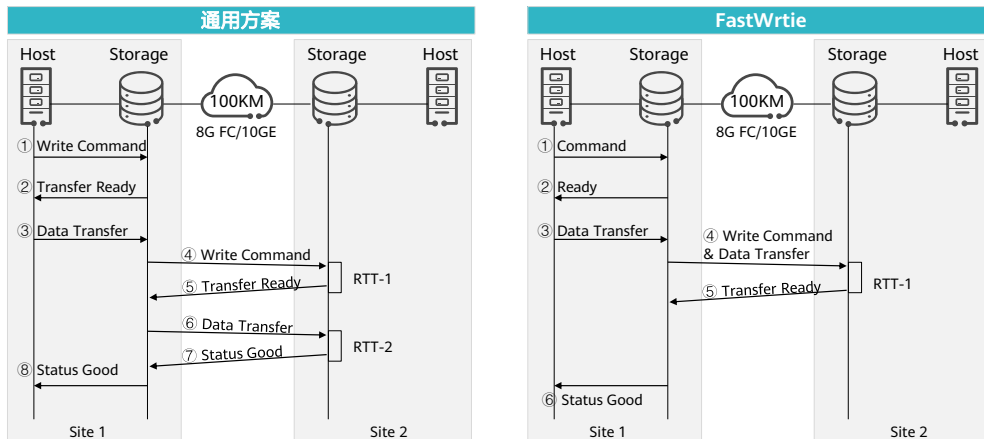
- **精简拷贝方案：**数据同步时，智能识别“全零数据”，只传输一个特殊标识，不传输数据
- 数据初始同步时间**减小90%**，数据初始同步链路**带宽占用减小90%**

NAS双活支持三层组网

- 概述：
 - 支持BGP协议，并新增一种逻辑端口VIP LIF，通过BGP通告VIP LIF的路由信息，从而实现VIP LIF跨子网漂移。
 - 新一代数据中心使用三层网络机制，双活两个站点可能部署在不同的子网中，需要LIF具备跨子网漂移的能力。
- 原理：
 - BGP LIF不提供业务，不在站点间同步，不支持漂移：BGP LIF与外部路由器建立BGP对等体，通过BGP LIF向外部路由器通告/撤销VIP LIF的路由信息
 - VIP LIF提供业务，在站点间同步，支持漂移，与二层网络的NAS LIF的功能一致：控制器或者站点故障时，VIP LIF漂移到另一个控制器或站点，配置新的静态路由，并通过BGP协议的network命令发布静态路由信息。主机到存储设备的路由切换，保障主机到存储设备间的IO不中断。

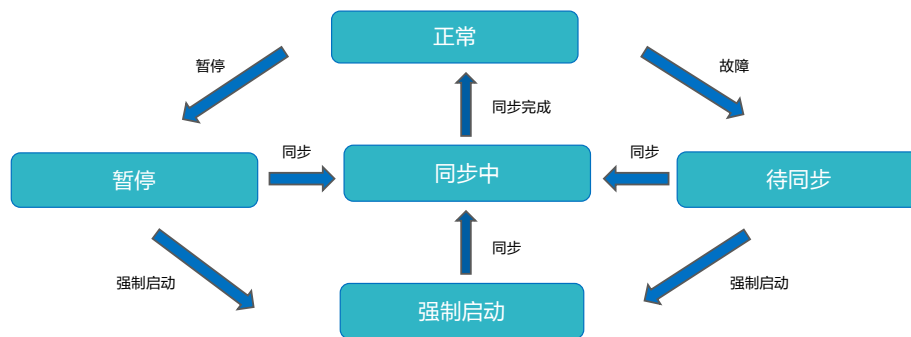
- 以控制器为单位配置BGP。
- 基于前端业务端口（ETH、VLAN、BOND）创建BGP LIF。
- 建立BGP LIF与路由器间的BGP对等体关系。
- 创建自定义漂移组，并将物理口加入自定义漂移组。（例如，对于VLAN LIF，漂移组要求VLAN TAG不一样，所以肯定来自不同的物理口）
- 在创建自定义漂移组时，系统自动创建一个虚拟端(VPORT)，VPORT的状态根据漂移组成员的状态进行刷新。
- 基于内部的VPORT虚拟端口创建VIP LIF。
- 创建VIP LIF时须指定漂移组，且漂移组中成员必须配BGP LIF。
- 系统自动配置VIP LIF到漂移组成员对应的BGP LIF的静态路由。
- 系统自动通过BGP将VIP LIF的路由通告给用户网络。

FastWrite

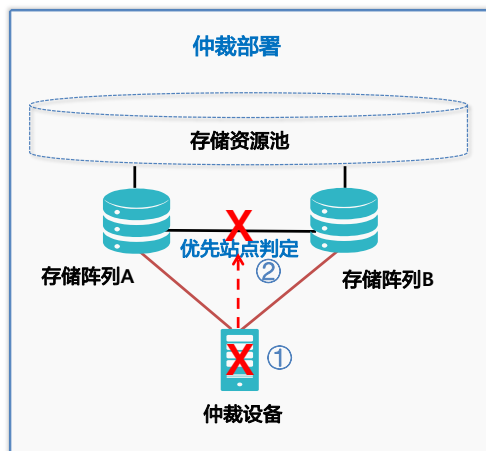


- 通用方案：写IO在传输的双端要经历“Write Command”和“Data Transfer”两次交互。
- FastWrite：通过对传输协议进行优化，将“写命令”和“写数据”合并为一次发送，跨站点写IO交互次数减少一半。

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。

目录

4. 增值特性

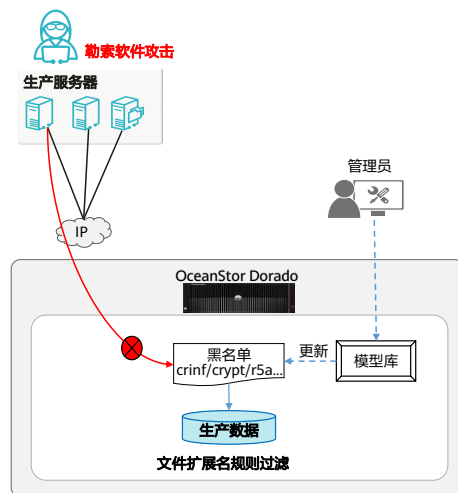
- 效率提升特性
- **数据保护特性**
 - HyperReplication
 - HyperMetro
 - **HyperDetect**

简介

- 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，通过分析文件I/O行为，快速筛选出异常文件，并对异常文件进行内容深度分析，检测出被勒索的文件。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。
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文件系统。对未识别出勒索感染的文件系统，分析对比文件系统快照变化特征，使用机器学习算法，进一步检测文件系统是否被勒索软件感染。

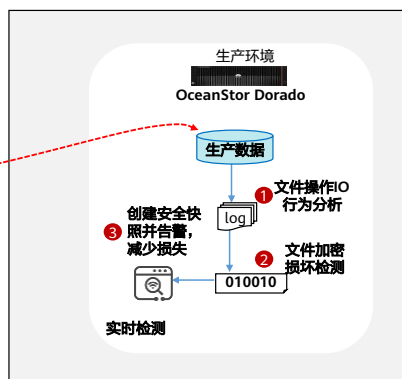
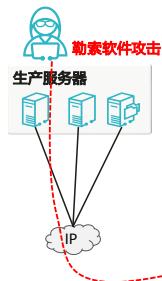
勒索文件拦截



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。

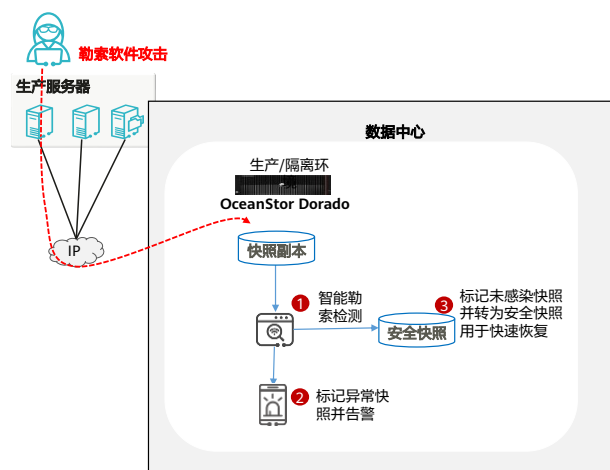
实时勒索检测



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。

- 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



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文件系统快照。对未识别出勒索感染的文件系统快照，分析对比文件系统快照变化特征，使用机器学习算法，进一步检测文件系统是否被勒索软件感染。未感染的快照转为安全快照并可用于快速恢复数据。

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。

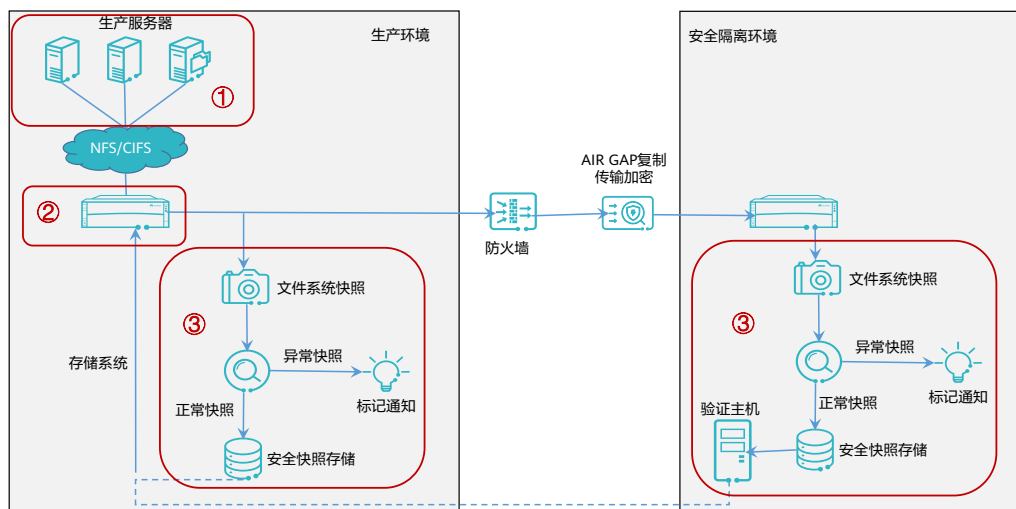
- 已知勒索软件攻击特征检测

- 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。

- 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。

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。

目录

1. 产品概览
2. 产品架构
3. 关键技术
4. 增值特性
- 5. 应用场景**

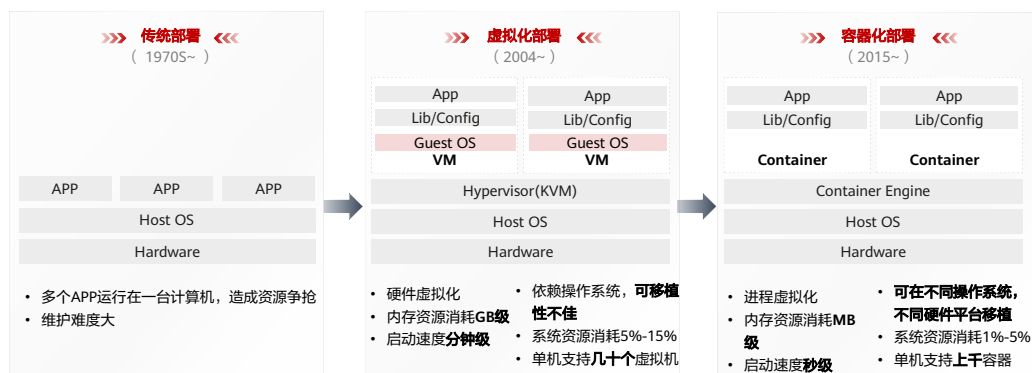
目录

5. 应用场景

- 华为容器存储解决方案
- 分布式数据库解决方案

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。
- K8s是一个比较典型的二层架构和 server-client 架构，Master 作为管理节点，运行K8S管理相关组件，其他节点用来给各服务使用，称之为Node/数据节点。

✓ 资源调度

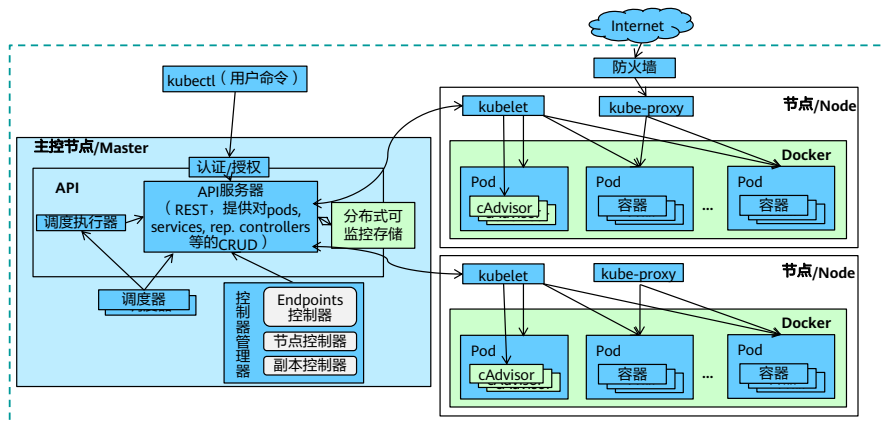
✓ 部署运行

✓ 服务发现

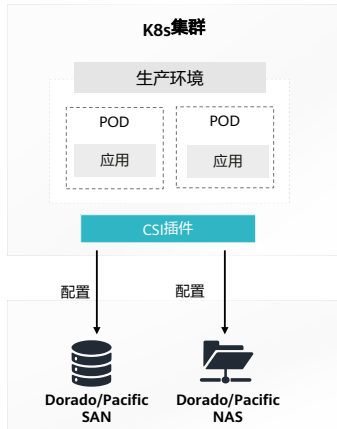
✓ 扩容缩容

✓ 健康监控

....



华为CSI简介



• 什么是CSI？

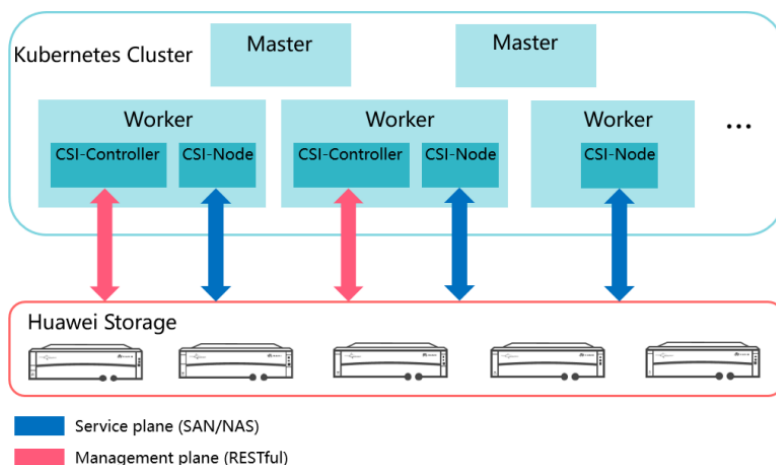
CSI全称为Container Storage Interface，译为容器存储接口，CSI是将任意块和文件存储系统暴露于Kubernetes等容器编排系统（CO）上的容器化工作负载的标准。

• 华为CSI

华为CSI是华为公司实现的满足，社区官方标准的CSI插件，用于给kubernetes/openshift等CO供应华为存储的LUN/文件系统将华为存储资源挂载到容器后，供容器使用。

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

CSI工作原理



- 如上图所示，华为存储为Kubernetes的工作节点通过多种协议提供SAN/NAS存储资源服务，华为存储和华为CSI驱动之间则使用Restful进行通信。
- 图中，绿色部分是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，并在需要的时候发起对CSI Driver调用。
- 黄色部分是华为开发的CSI Driver，用于将sidecar发起的调用在华为存储上实施。如在
- Kubernetes上创建一个持久化卷的操作，落实在华为存储上便是创建一个LUN/文件系统。
- 华为CSI Driver分为两个组件：
 - CSI Controller：以Deployment方式的Pod单独运行在一个工作节点中，用于和存储交互，对存储侧资源进行创建、删除等操作，如创建LUN/文件系统，扩容等。
 - CSI Node：以DaemonSet方式的Pod运行在Kubernetes的每个工作节点中，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、卸载、格式化文件系统等操作。

华为容器存储解决方案



华为提供

产品交付:

- 存储盒子
- 华为CSI插件
- 华为CDR插件

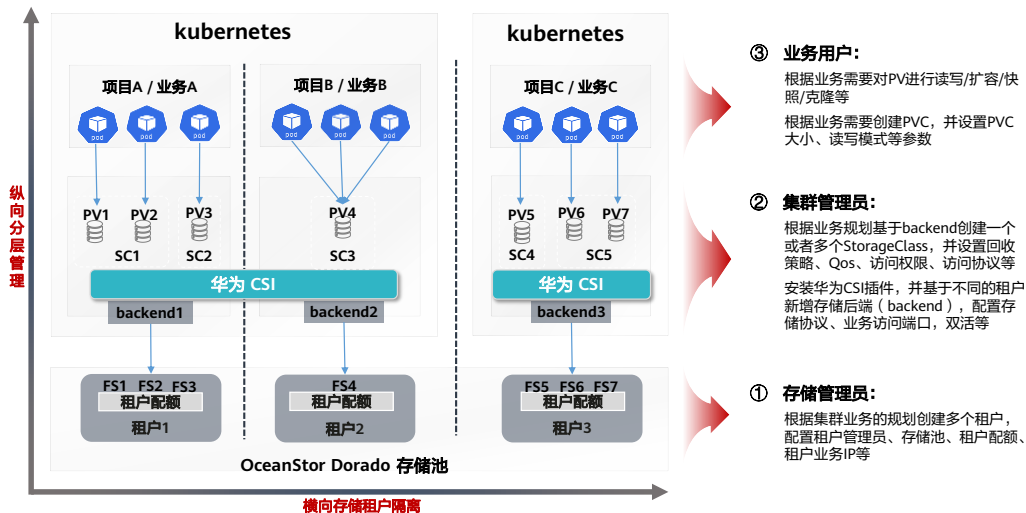
解决方案交付:

- MySQL容器化对接Dorado最佳实践
- Kafka容器化对接Dorado最佳实践
- 华为CSI插件使用最佳实践

验收标准:

- 以产品验收标准为准

CSI最佳实践



目录

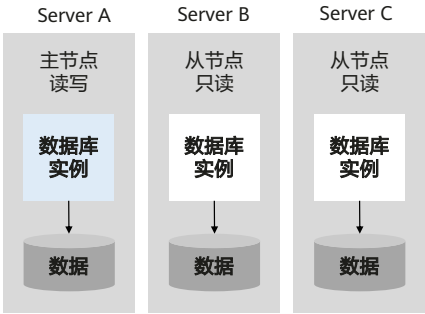
5. 应用场景

- 华为容器存储解决方案
- 分布式数据库解决方案

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

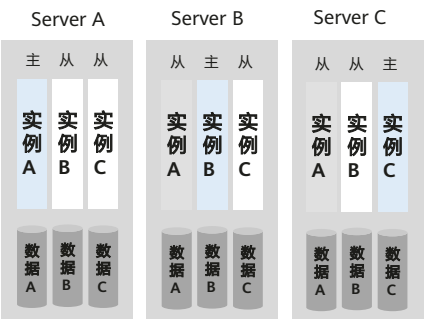


优点：数据一致性易保障，效率高

缺点：大库性能差、扩展性差

分布式部署

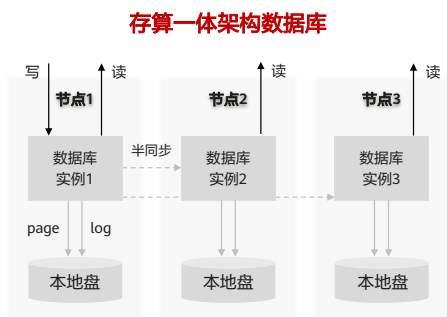
数据切片后分散存放，代表：OceanBase



优点：并发访问效率高，易扩展

缺点：难以保障数据一致性

存算一体的数据库架构存在的缺陷



局部拥塞导致业务停滞

单写+分库分表不均导致性能瓶颈
高峰期数据库性能下滑至无法服务

数据不强一致

主从间只保证日志同步
高并发写时从库读存在不一致风险

性能与可靠性难以兼容

容灾站点间同步复制性能损失大
异步复制备集群数据可能不一致

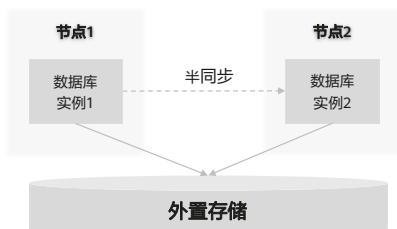
海量服务器硬盘管理难

服务器缺乏硬盘健康管理能力
海量主机下硬盘故障难以预测

国产数据库选择存算一体，一方面是满足超大规模部署诉求，一方面是鲸吞数据库全栈市场的商业动机

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，补从快
- 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

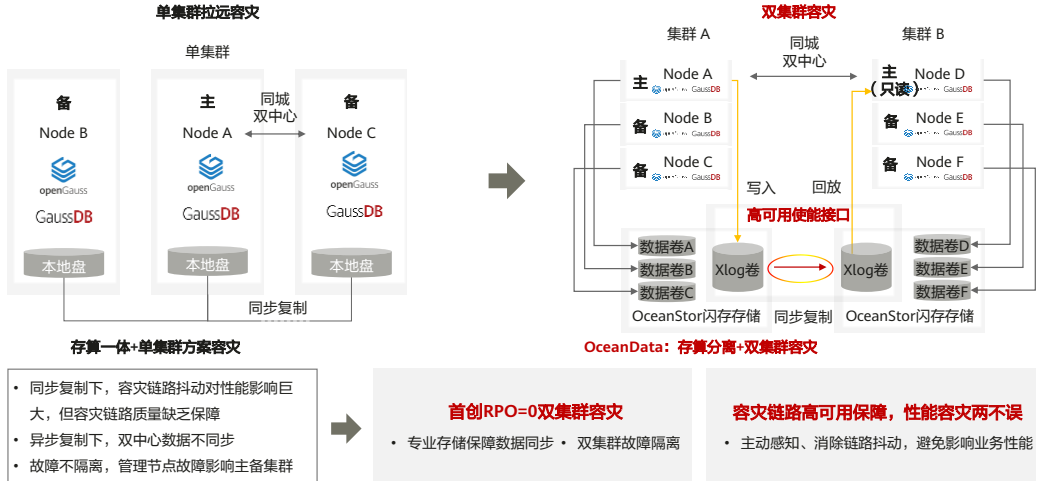
性能提升

- 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

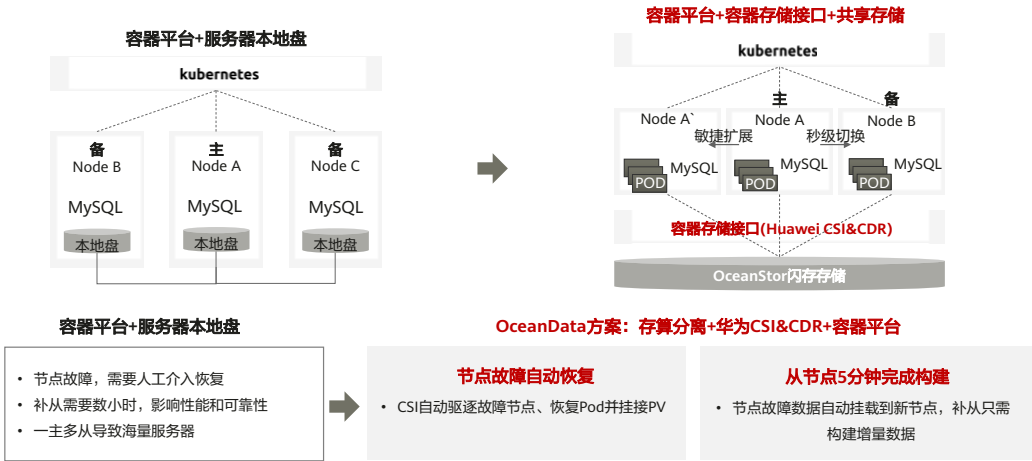
TCO降低

- 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

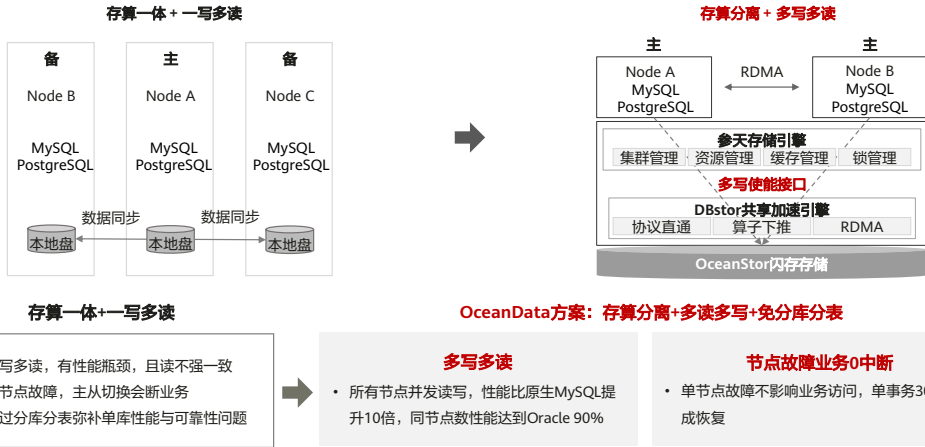
分布式新核心改造解决方案



数据库微服务改造解决方案



传统核心免分库分表改造解决方案



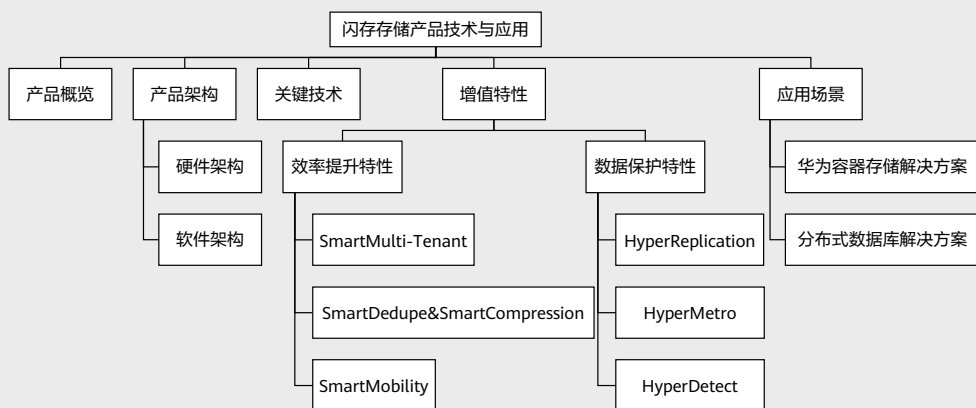
思考题

1. （多选题）华为 OceanStor Dorado全闪存系列存储的典型应用场景有哪些？
 - A. 关键业务存储
 - B. 虚拟环境存储
 - C. 虚拟桌面
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：

- ABC
- C
- T

本章总结



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://enterprise.huawei.com/cn/>
- 技术支持: <https://support.huawei.com/enterprise/>
- 在线学习: <https://support.huawei.com/learning/Index!toTrainIndex>
- 华为Support案例库: <https://support.huawei.com/enterprise/servicecenter?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.

Copyright©2024 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.



数据保护产品技术与应用



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OceanProtect备份产品
 - 熟悉OceanStor BCManager管理软件
 - 了解数据保护产品应用场景

目录

1. OceanProtect备份产品

- 产品介绍
- 数据缩减技术
- 高性能设计
- 高可靠设计
- 高安全设计

2. OceanStor BCManager

3. 应用场景

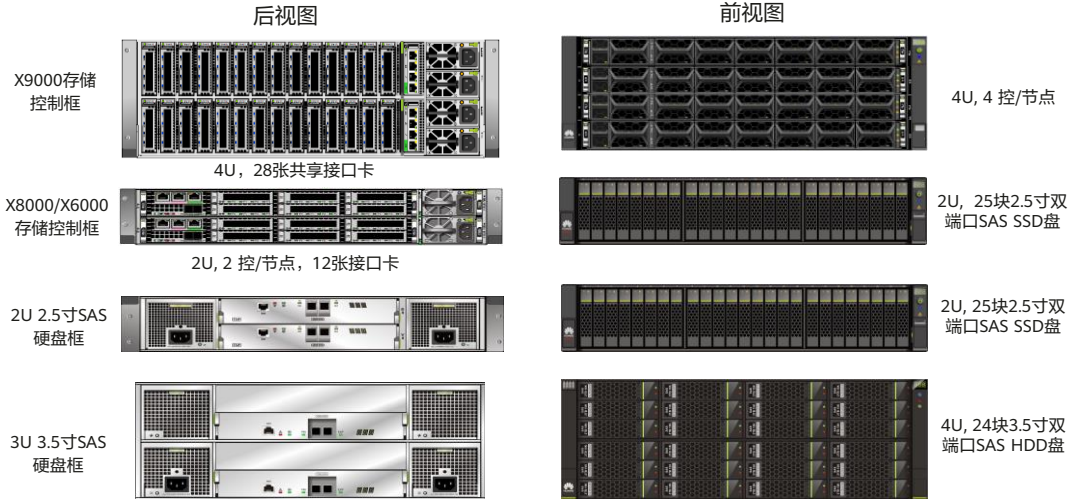
产品规格



产品型号	OceanProtect X3000	OceanProtect X6000		OceanProtect X8000		OceanProtect X9000	
形态	HDD形态	全闪形态	HDD形态	全闪形态	HDD形态	全闪形态	HDD形态
数据缩减率*	Up to 72:1						
系统物理备份带宽	Up to 6 TB/hour	Up to 19 TB/hour		Up to 55 TB/hour		Up to 155 TB/hour	
系统逻辑备份带宽	Up to 10 TB/hour	Up to 45 TB/hour		Up to 117 TB/hour		Up to 310 TB/hour	
恢复带宽	Up to 1 TB/hour	Up to 22 TB/hour	Up to 8 TB/hour	Up to 57 TB/hour	Up to 24 TB/hour	Up to 172 TB/hour	Up to 48 TB/hour
控制框高度	2U	2U		2U		4U	
控制器扩展	2	2		2-4		4-8	
系统可用容量	16 TB-60 TB (Three capacity levels: 16TB 36TB 60TB)	16 TB-300 TB		150 TB-2.0 PB		480 TB-3.6 PB	
逻辑容量*	Up to 4.3 PB	Up to 21.6 PB		Up to 144 PB		Up to 259.2 PB	
数据盘类型	4 TB/8 TB/14 TB NL-SAS	3.84 TB/7.68 TB SAS SSD	4 TB/8 TB/14 TB NL-SAS	3.84 TB/7.68 TB SAS SSD	4 TB/8 TB/14 TB NL-SAS	7.68 TB SAS SSD	8 TB/14 TB NL-SAS

- 实际数据缩减率取决于应用类型和备份策略。
- 逻辑容量根据以下条件计算：最多72:1数据缩减率；实际逻辑容量取决于系统可用容量和实际数据缩减率。
- OceanProtect专用备份存储包含两种形态：全闪形态和HDD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。

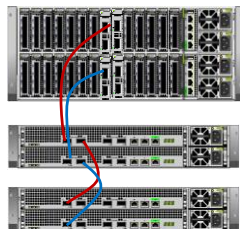
硬件视图



- OceanProtect专用备份存储主要由控制框和硬盘框组成。OceanProtect专用备份存储的HDD形态支持2U SAS硬盘框和4U SAS硬盘框，OceanProtect专用备份存储的全闪存形态只支持2U SAS硬盘框。
- X9000:
 - A/A硬件架构，全部件冗余，无单点故障；
 - 前面板上下并排插入四个控制器，每个控制器有独立的BBU和风扇，每控192Core，每控有2块2.5寸系统盘；
 - 后面板最大支持插28张接口卡，在4个控制器间全局共享；有4个电源，分两个供电平面；
 - 通过前后对插设计，风扇、电源、BBU、接口卡支持在线维护。
- X8000/X6000
 - A/A硬件架构，全部件冗余，无单点故障；
 - 电源、BBU、接口卡支持在线热拔插维护，风扇更换需要拔插控制器；
 - 控制框0.82米长需要1米机柜；
 - IOB边上有内存插槽，但没有内存条。

Scale-up和Scale-out

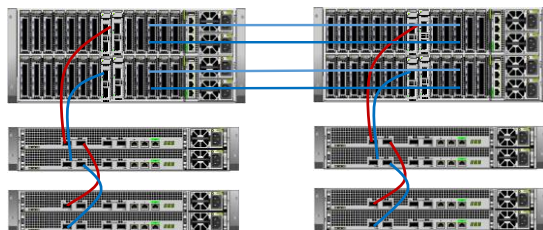
Scale-up



单系统容量最大支持4.6PB物理容量

单系统性能最大支持91TB/H

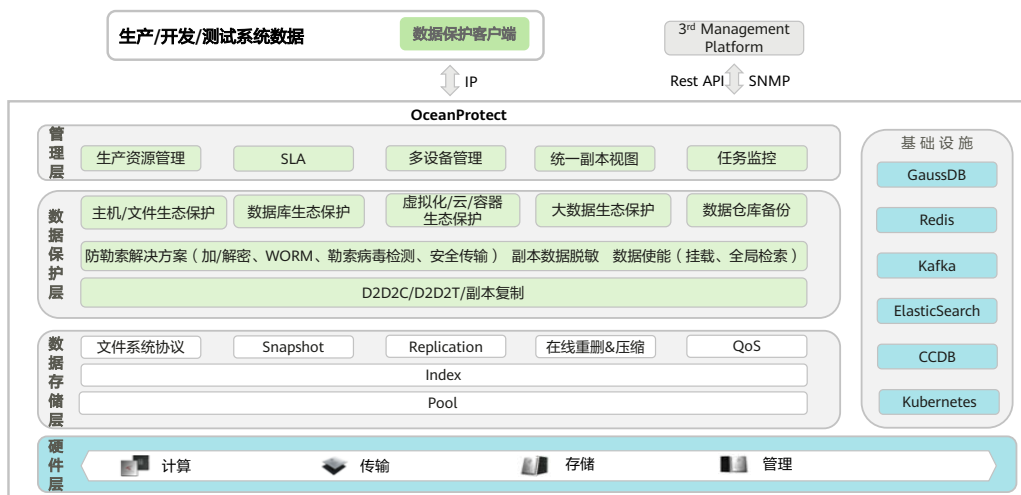
Scale-out



单系统容量最大支持9.2PB物理容量

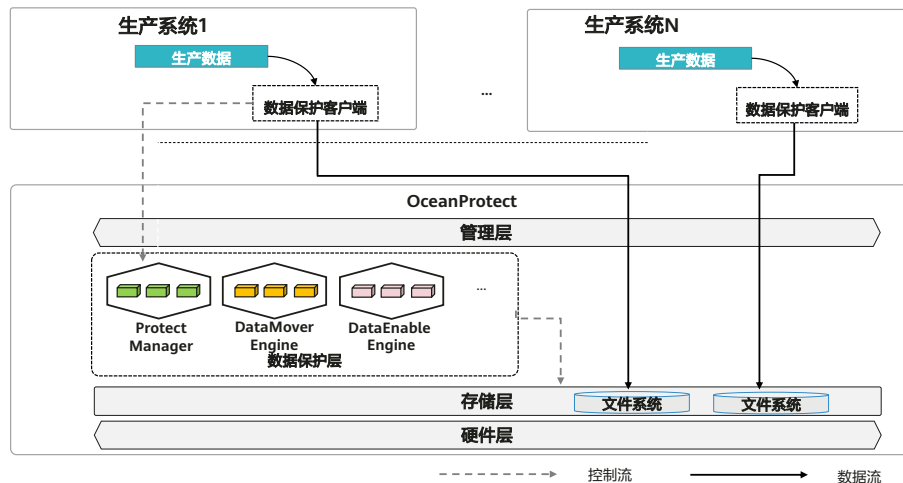
单系统性能最大支持155TB/H

专用存储软件架构



- **数据存储层 (Data Store Layer)**：主要包含存储基础平台和硬件系统，为数据保护软件提供统一的数据存储能力，后端采用全NAS协议访问，支持备份数据直接在主机使用NAS协议访问存储，同时支持备份数据通过回传到一体化备份存储介质内通过后端存储容器卡卡间交换高效写入后端存储。
- **数据保护层 (Data Protect Layer)**：数据保护层负责处理数据保护核心业务，提供数据/日志备份、数据/日志恢复/即时恢复、副本归档/复制、副本/日志删除、副本挂载、数据脱敏、全局检索等功能。
- **数据管理层 (Manager Layer)**：数据管理层提供全图形化界面客户视图，负责备份、恢复、归档、复制等相关任务调度，副本数据管理等功能。
- **数据保护基础设施 (Common Components)**：包括Gauss DB数据库、ElasticSearch搜索引擎、kafka、zookeeper等基础组件。
- **数据保护客户端 (Protect Client)**：备份客户端，根据需保护的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。

备份一体机架构



- 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。
- 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。
 - Protect Manager：数据管理模块，负责备份、恢复、归档、复制等任务调度。
 - DataMover Engine：数据流动模块，负责备份、恢复、复制、归档、副本挂载的数据处理。
 - DataEnable Engine：数据使能模块，负责全局检索、数据脱敏。
- 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。

数据备份特性(1)

分类		特性描述
即时可用	副本即时可用	高效备份 - 多节点并行流备份，实现大数据PB级超大规模集群备份恢复 - 基于ROW的永久增量备份：秒级合成全量副本，提升备份效率 - 源端重删+短IO路径，更高传输效率
		高效恢复 - 数据即时恢复，后台热迁移，业务恢复时间从4小时到5分钟 - 支持重删副本挂载恢复 - 百亿级文件秒级检索，单文件细粒度恢复 - 副本挂载可用，激活冷数据价值
	多场景快速备份恢复	数据库场景生态兼容： Oracle MySQL SQL Server GaussDB T OpenGauss PostgreSQL Dameng MariaDB
		文件场景NAS海量小文件备份，主机文件备份， 极致性能，挂载恢复，永久增量备份
		大数据场景生态兼容： MRS FusionInsight Cloudera CDH 开源Hadoop 组件支持： HDFS HBase Hive ElasticSearch Redis ClickHouse
		DWS数仓场景DWS GaussDB OLAP 全功能界面，分布式并行流备份
		虚拟化/云/容器场景生态兼容： VMware FusionCompute Huawei Cloud Stack Kubernetes (FlexVolume)

数据备份特性(2)

分类			特性描述
经济可靠	多样数据缩减模式		数据保护服务器与容灾备份存储融合，部署简单，节省空间
			在线变长重删和专利字节级压缩算法，业界先进数据缩减率
	数据安全高可用	数据脱敏	副本数据脱敏，安全使用副本数据
		端到端副本防勒索	副本防勒索病毒，保障备份数据安全 副本防篡改，WORM只读属性设置
		副本数据流动	副本归档，副本长期留存 副本数据存储层异步复制
	全冗余硬件架构		双控A-A高可靠架构 最大容忍3盘失效，备份链路冗余，保障业务高可靠
极简易用	保护自动化		数据保护全流程自动化 Dashboard SLA策略
	恢复极简化		副本3步恢复 自定义脱敏规则 全局检索
	生态全面化		

目录

1. OceanProtect专用备份存储

- 产品介绍
- **数据缩减技术**
- 高性能设计
- 高可靠设计
- 高安全设计

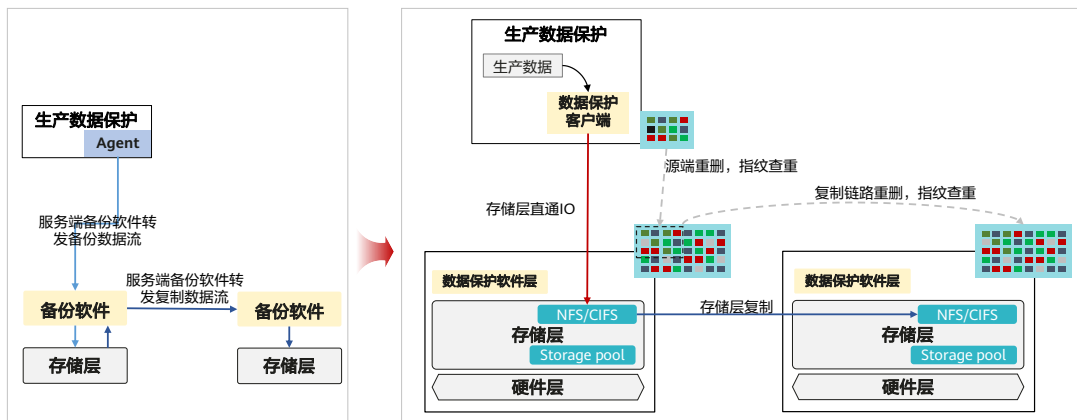
2. OceanStor BCManager

3. 应用场景

源端重删+短IO路径

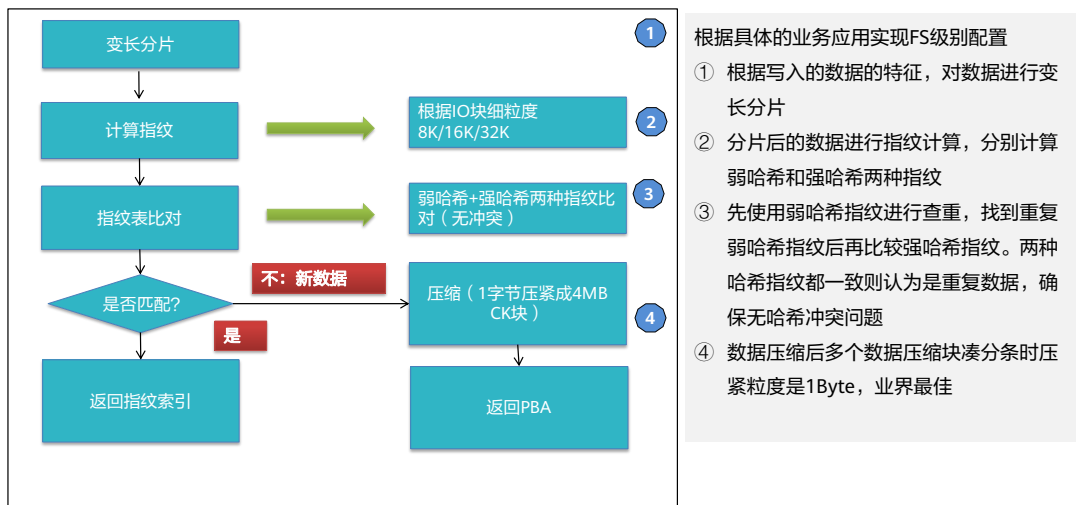
传统备份软件：数据流经过备份软件转发

数据流IO直通存储层



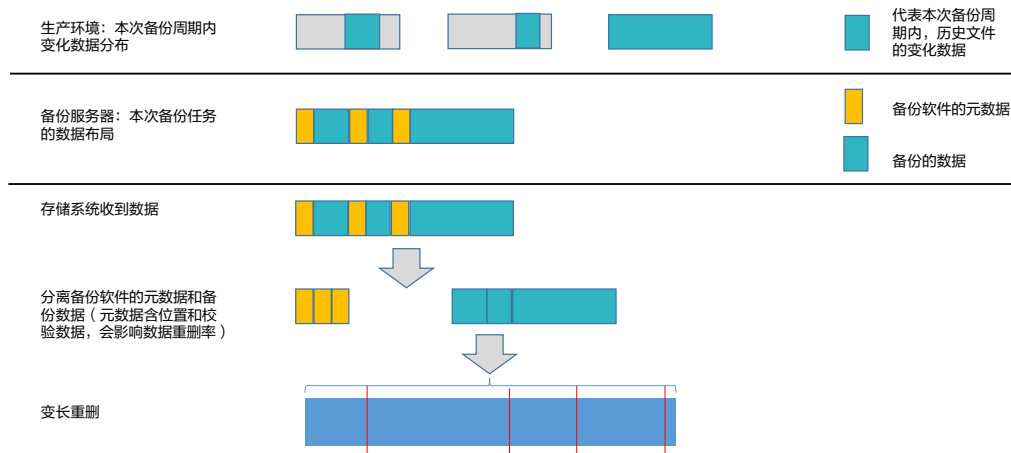
- 由于OceanProtect后端为多控制器设备，为避免数据的跨控转发，源端重删客户端与存储之间会建立多条路径，并且多条路径链接存储的每个控制器节点，DataTurbo将IO请求直接下发到所归属的控制器节点，避免跨控制器节点，IO路径短，提升设备写性能。

在线重删压缩



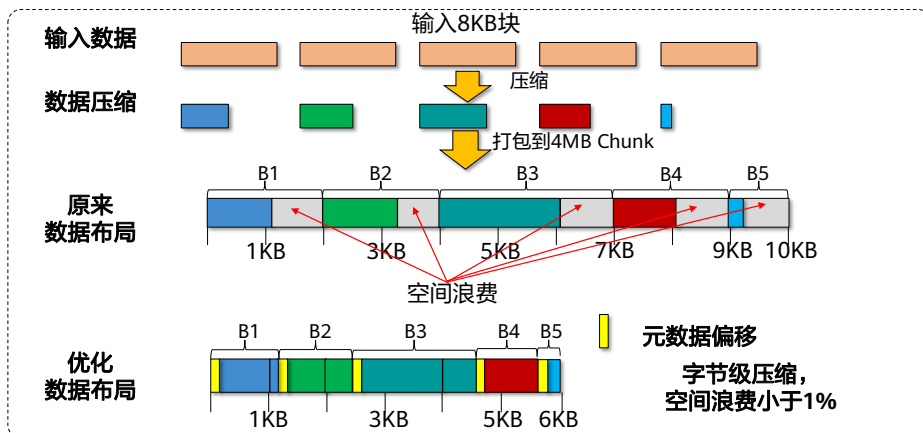
- 注：如果使用LUN，不支持重删支持压缩，需要配合备份软件的重删功能使用。

写入数据预处理+变长重删



- OceanProtect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。
- OceanProtect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），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。

按字节压紧技术



- 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。

DIF重排



- 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。

目录

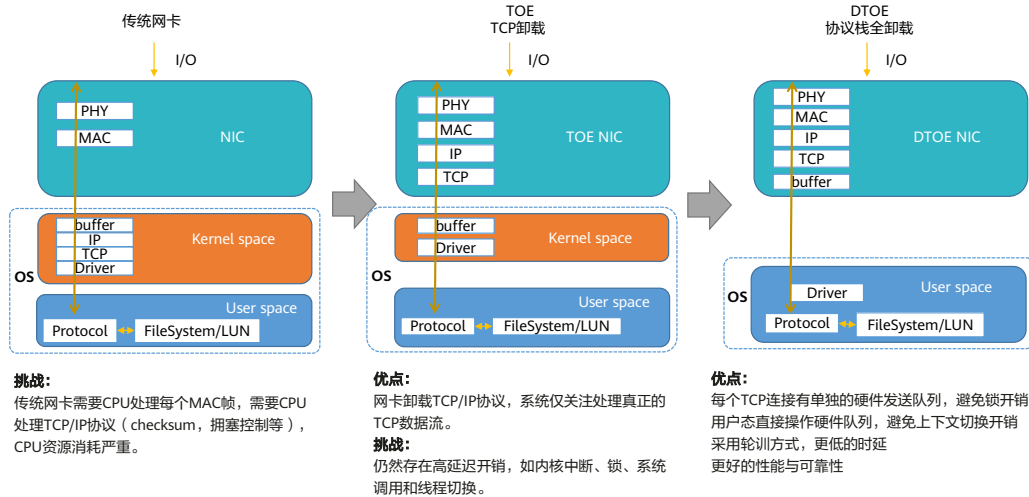
1. OceanProtect专用备份存储

- 产品介绍
- 数据缩减技术
- **高性能设计**
- 高可靠设计
- 高安全设计

2. OceanStor BCManager

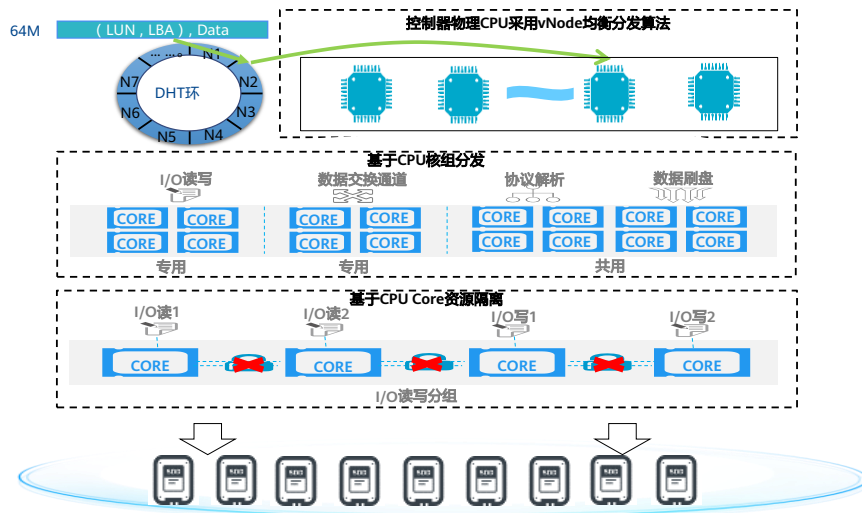
3. 应用场景

智能网卡优化



- DTOE实现协议卸载，释放CPU资源，提供更大性能。
- DTOE: direct TCP offload engine，华为深度优化的TOE技术，用于TCP/IP协议卸载。

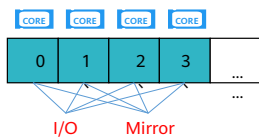
业务负载均衡



- Step1: LUN以64M粒度打散到不同的vnode上, vnode与CPU做亲和绑定, 减少跨物理CPU调度和CPU Cache Miss率, 提升CPU使用效率;
- Step2: vNode内CPU分组算法: 读写I/O与其他I/O部署在不同分组上避免相互干扰;
- Step3: vNode内CPU分核算法: 一次请求在同一个核上连续执行并实现免锁设计, 避免频繁多核切换。

CPU多核负载均衡优化

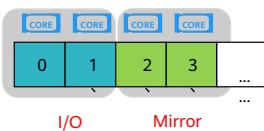
传统方式



挑战:

不同任务竞争CPU不同核的时间片，
导致数据I/O在不同核间频繁切换，导
致时延高

CPU Core分组



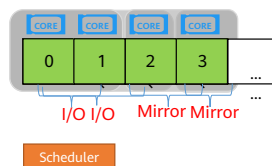
优点:

避免干扰和资源频繁切换

挑战:

不同业务，部分核过载，时延高

OceanProtect
CPU Core分组+智能调度

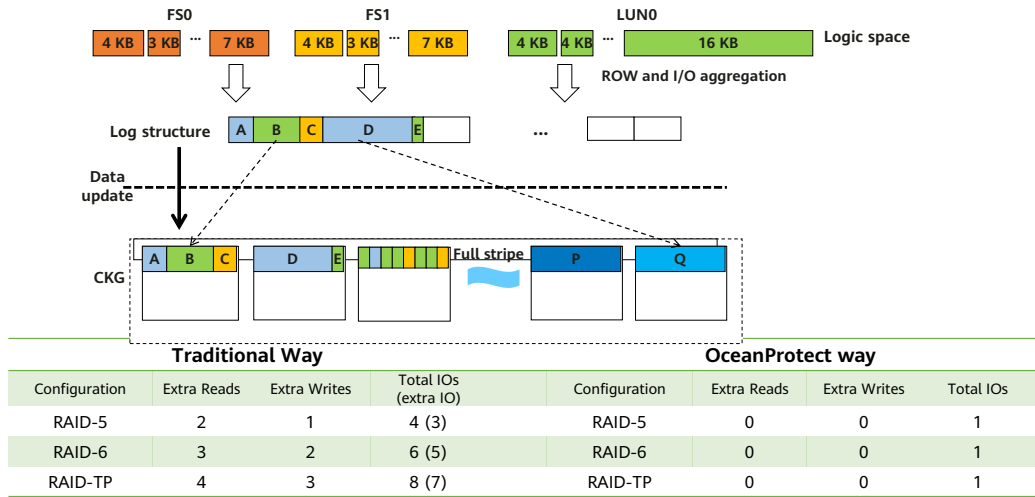


优点:

智能调度器根据业务负载状态，将任务
分发到其它核，实现负载均衡

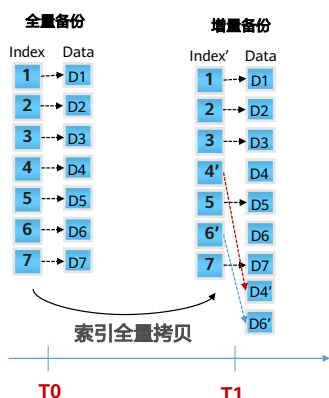
- 支持内核的自适应调度，满足高耗能群组的临时需求。

基于ROW的满分条写

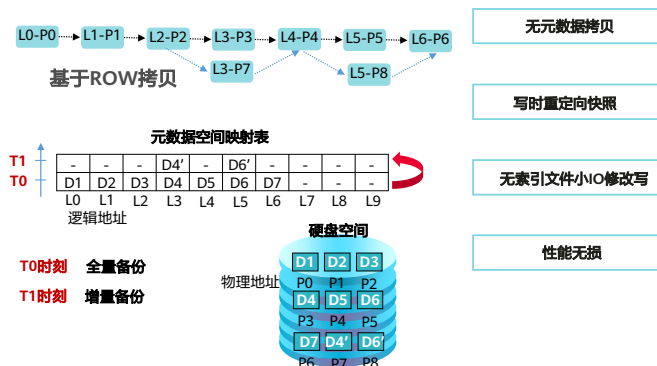


基于ROW的永久增量备份

传统方案：基于索引拷贝的永久增量备份

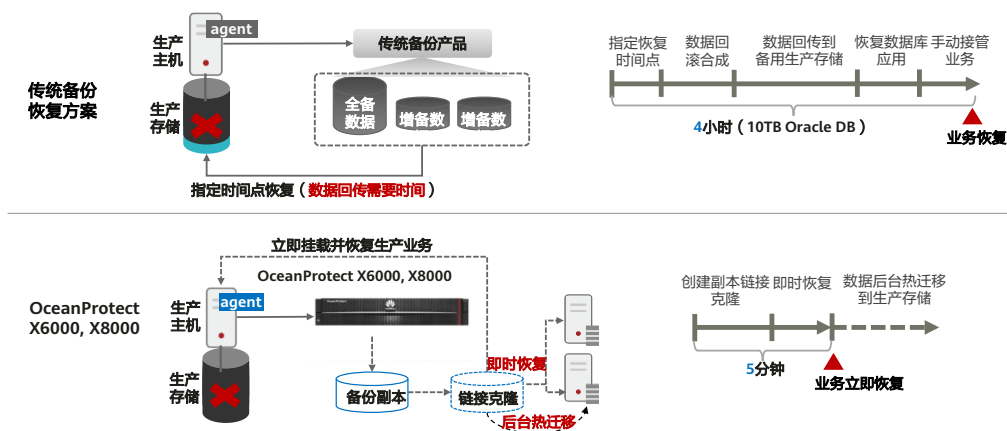


基于ROW的永久增量备份



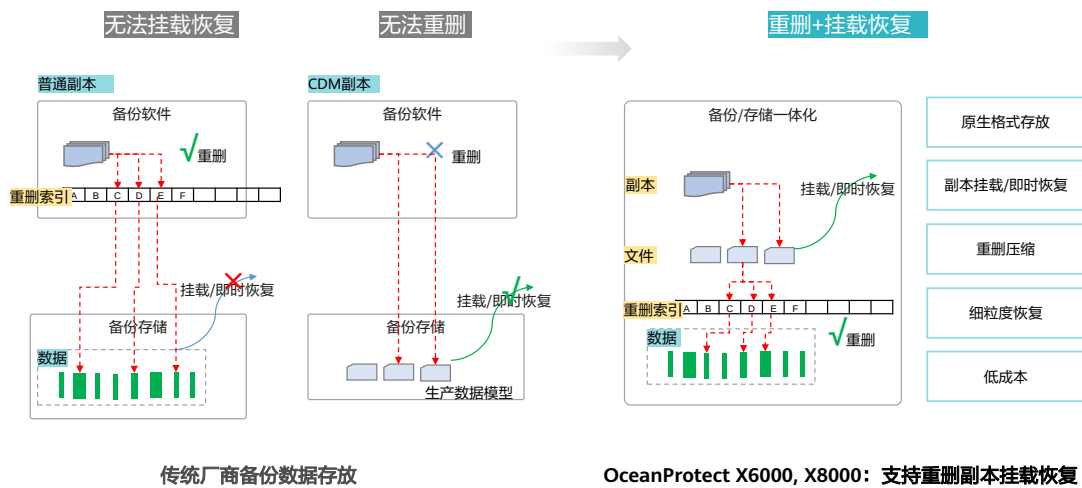
- 传统厂商在实现永久增量备份时，需要先进行一次上次备份副本的索引文件全量拷贝，再根据增量信息修改对应的偏移索引。得到永久增量副本的全量索引文件。或者拷贝真实数据后再合并写入成新文件写入，这种实现方式需要拷贝一次索引文件，并多次对索引文件进行修改写，或者拷贝整文件内容合并后写入，对存储层带来极大的IO负担。
- OceanProtect支持的大部分应用，均采用永久增量备份方式进行增量数据存储，和传统后期合成永久增量不同，备份软件采用ROW（Redirect-On-Write）写时重定向方式进行合成。备份数据以原始文件格式或固定文件格式写入，当增量备份时，分析增量数据与上一次备份数据的差异数据块和偏移位置，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，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。

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

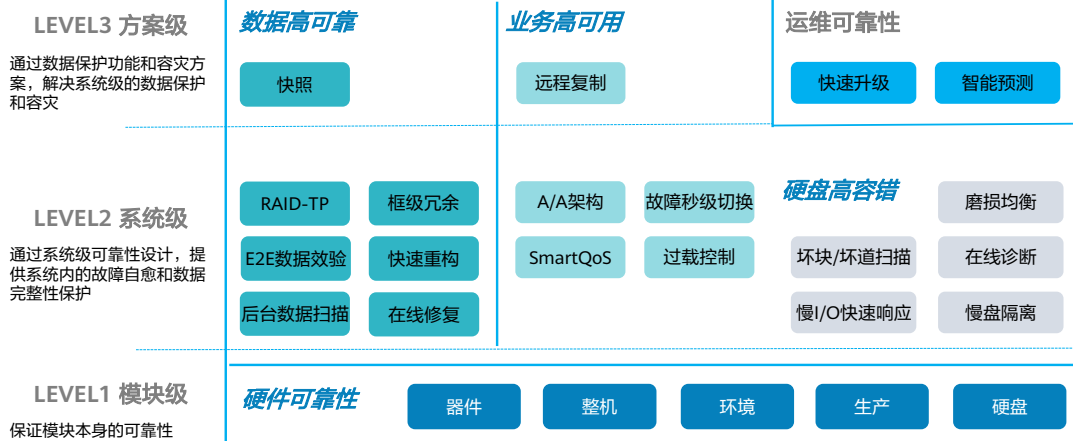
1. OceanProtect专用备份存储

- 产品介绍
- 数据缩减技术
- 高性能设计
- **高可靠设计**
- 高安全设计

2. OceanStor BCManager

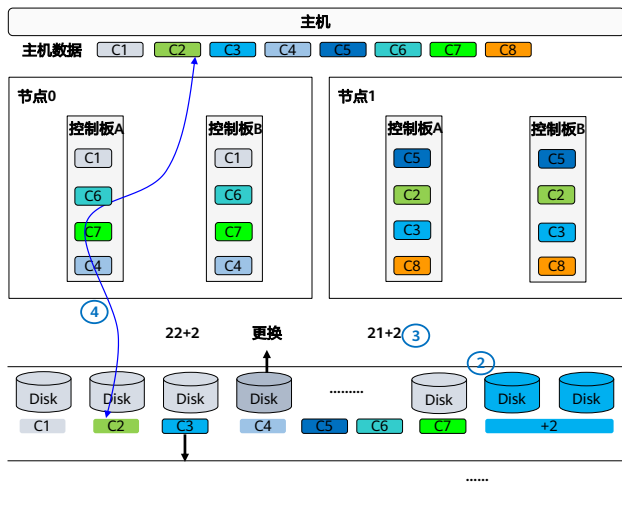
3. 应用场景

主存级可靠



- OceanProtect存储系统通过从模块级，系统级和方案级三层角度构建系统可靠性。
 - 模块级可靠性主要指通过工程，工艺等措施保证良品率，尽量保证引入的情况故障率低，持久耐用。
 - 系统级可靠性主要指在特定失效率的硬件部件下，通过冗余和软件逻辑保证整个阵列的可靠性。
 - 方案级可靠性指通过阵列内的数据备份，克隆等，阵列间通过远程复制，双活等保证存储阵列故障后业务不中断，数据不丢失能力。
- 除了上述维度外，我们将整个可靠性分为数据可靠，业务可用和运维可靠。通过这些维度可以确保各专项活动，方案技术有一定的倾向性，并从这些维度去解决可能出现的数据丢失，业务中断以及维护上的问题。

数据高可靠

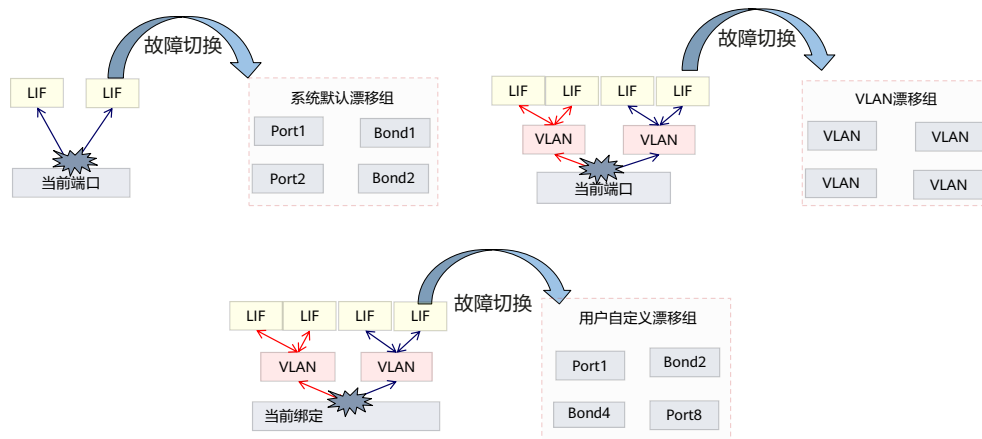


端到端数据冗余保护

1. **缓存数据冗余**：缓存于控制器的主机写入数据支持两副本，控制器故障缓存不丢失。BBU掉电保护，系统异常掉电缓存不丢失。
2. **存盘数据冗余**：存储于盘上的数据通过RAID2.0+冗余存放，多盘相继或同时故障用户数据不丢失；通过快速重构，进一步保证数据可靠性。
3. **成员盘不足数据冗余度不下降**：在成员盘数量因为更换不及时不满足RAID要求时，通过减少参与数据盘数，维持冗余，利用缩列重构保证数据冗余不下降。
4. **端到端数据一致性保证**：用通过端到端DIF，父子检验保证IO传输路径数据不损坏。

- 数据高可靠主要包括如下几个方面，如图所示：
 - 首先，为了保证向主机IO响应时延，针对写入阵列的数据，会先写到缓存，然后向主机返回成功。支持缓存数据的双副本功能，支持多重故障数据不丢失。
 - 其次，OceanProtect存储系统支持RAID2.0+，通过EC-1/EC-2/EC-3等技术保证即使单盘，双盘，以及三盘故障，落盘的数据也不丢失。成员盘不足，数据冗余度也不下降。
 - 再次，针对已经落到盘上的数据，HSSD支持盘内RAID，即使单个DIE失效，以及出现盘目坏块，盘也能利用盘内的冗余恢复数据，对外不体现数据丢失。
 - 最后，针对主机读写IO，从阵列接收到IO到IO落盘，从盘上读取IO到向主机返回，通过PI和父子校验保证传输路径的数据完整性和一致性。

漂移组



• 系统默认漂移组

- 当LIF故障时，存储在对端控制器上找到有相同位置、类型（物理/绑定）、速率（GE/10GE）、MTU的端口，并漂移过去。如果这个端口又故障了，存储会按此原则在另外一个控制器上找到合适的端口。在对称组网中，只需在创建LIF的时候选择该漂移组。

• VLAN漂移组

- 创建VLAN端口时，系统自动创建VLAN漂移组。当LIF故障时，存储会漂移到漂移组中有着相同tag和MTU且状态正常的VLAN端口。当使用VLAN组网时，更容易部署LIF，使用该漂移组。

• 用户自定义漂移组

- 当用户想要自定义故障切换策略时，可以创建该组并指定使用的LIF。当LIF故障时，系统会根据之前指定的组成员找到合适的故障切换目标端口。

• 故障回切

- 支持手工和自动回切，主用端口开始自动回切到恢复正常大概5秒钟。

目录

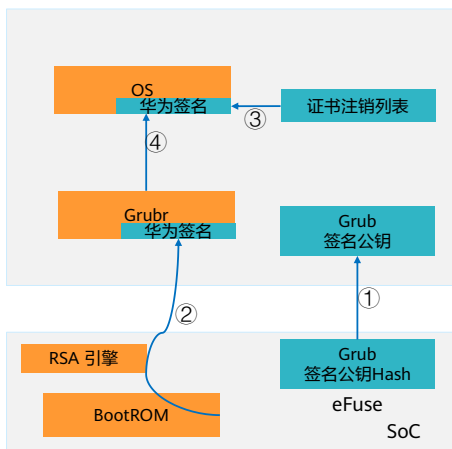
1. OceanProtect专用备份存储

- 产品介绍
- 数据缩减技术
- 高性能设计
- 高可靠设计
- **高安全设计**

2. OceanStor BCManager

3. 应用场景

硬件可信安全启动



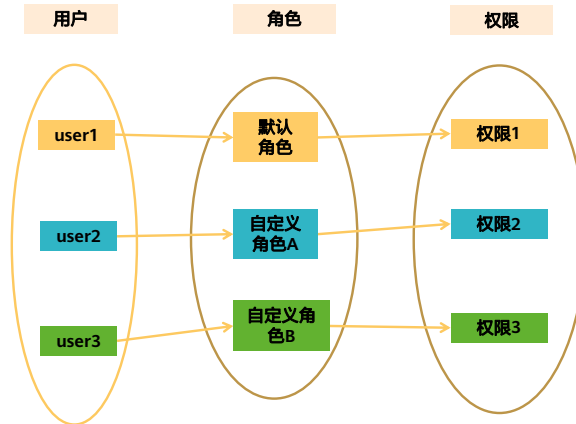
安全启动思想：建立硬件信任根（不可篡改），从信任根开始，一级认证一级，建立信任链，把信任扩展到整个系统，实现系统行为可预期。

华为 OceanProtect 基于硬件可信根构建安全启动，保证启动过程加载的软件没有遭到黑客或者恶意软件篡改。

安全启动的软件验证、加载过程：

1. **验证Gurb的签名公钥**：BootROM验证Grub签名公钥完整性；如果验证不通过，启动过程中止；
2. **验证并加载Grub**：BootROM验证Grub签名；验证通过后加载Grub；如果验证不通过，启动过程中止；
3. **验证软件签名证书状态**：Grub根据证书注销列表验证软件签名证书的状态。如果证书无效，启动过程中止；
4. **验证并加载OS**：Grub验证OS的签名。验证通过后加载OS；如果验证不通过，启动过程中止。

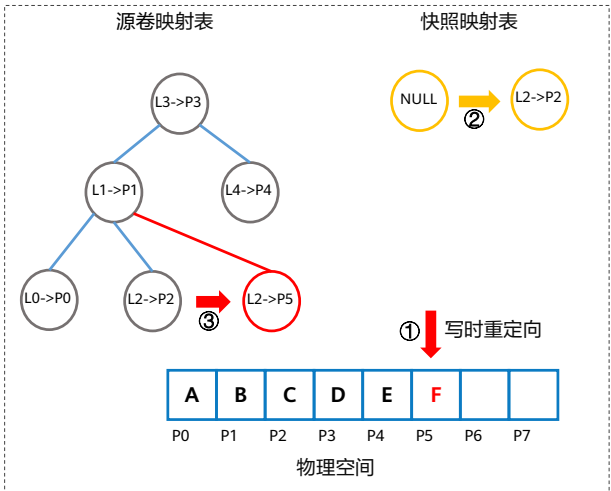
基于角色权限管理



权限=功能1: 对象1,对象2; 功能2: 对象3, 对象4;(一组或者多组功能和对象的关系)

- 预置默认角色：系统预置系统管理用户和租户管理用户的默认值角色。
- 系统管理用户默认角色：
 - 超级管理员 (super_administrator)
 - 管理员(administrator)
 - 安全管理员 (security_administrator)
 - SAN管理员 (san_administrator)
 - NAS管理员(nas_administrator)
 - 数据保护管理员 (dataProtection_administrator)
 - 网络管理员 (network_administrator)
- 租户管理用户角色：
 - 租户管理员(vStore_administrator)
 - 租户数据保护管理员(vStore_dataProtection)
 - 租户协议管理员 (vStore_protocol)
- 自定义角色：支持用户依据自己场景自定义角色。用户自定义角色时可以选择多个功能，每个功能可以选择多个对象。自定义角色支持删除和修改。

性能无损快照



目录

1. OceanProtect专用备份存储

2. OceanStor BCManager

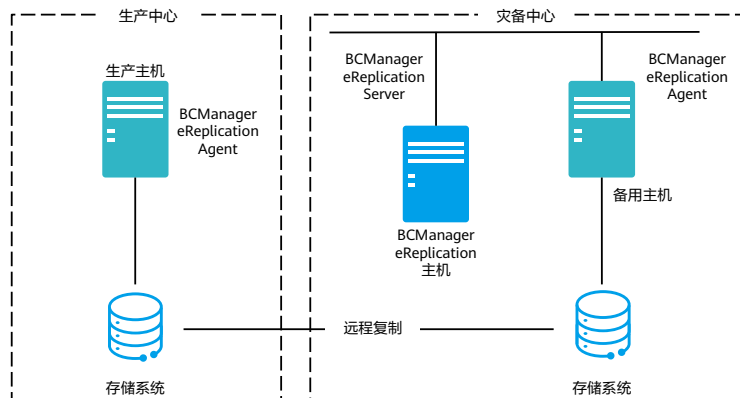
- eReplication

- eBackup

3. 应用场景

BCManager eReplication介绍

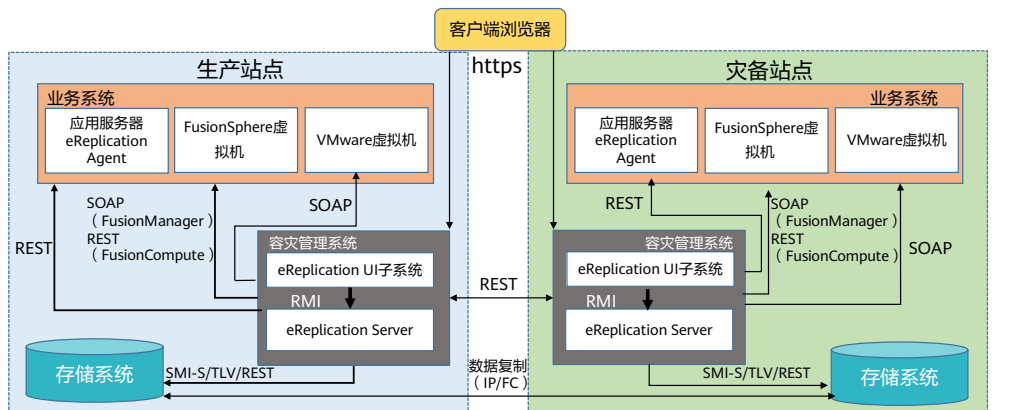
- BCManager eReplication是一款基于应用数据一致性、存储快照、远程复制技术，针对华为典型容灾解决方案，提供可视化、流程化、简单、快捷的操作与监控平台的灾备管理软件。



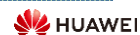
- 灾备中心部署BCManager eReplication主机，安装BCManager eReplication Server软件；生产中心业务主机和备用主机上安装BCManager eReplication Agent软件。

BCManger eReplication软件架构

- BCManager eReplication基于B/S架构，通过浏览器进行容灾管理，包括BCManager eReplication Agent、BCManager eReplication Server两个子系统。

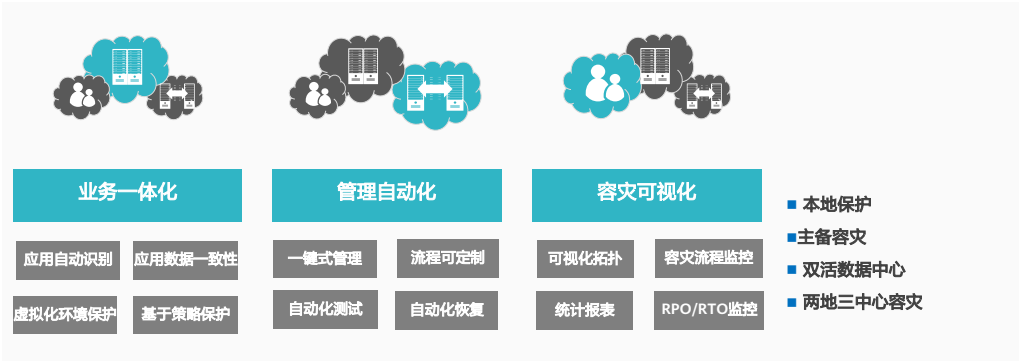


36 Huawei Confidential

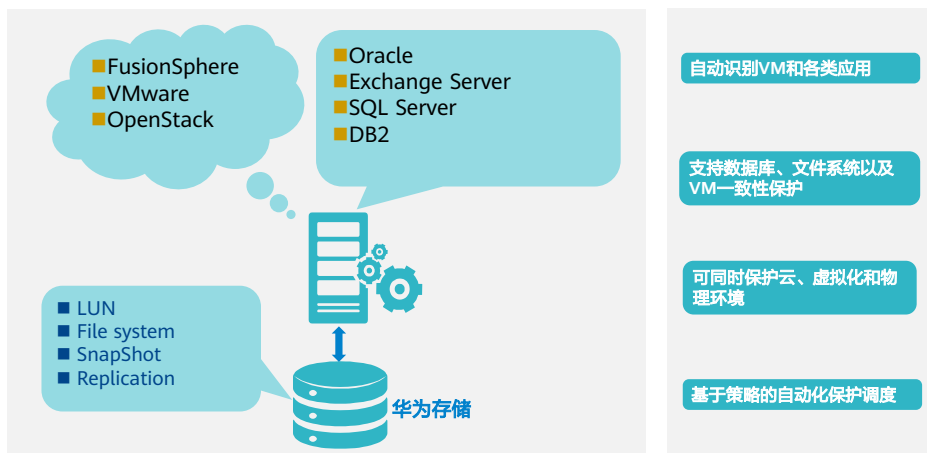


- eReplication包含2个子系统：
 - eReplication Agent安装在业务主机上，提供主机应用的发现、应用数据一致性的保证和应用的恢复。此外，eReplication Agent还负责将对应的应用系统信息传递给eReplication Server，从而让操作人员更加方便的管理和配置数据保护与容灾任务。
 - eReplication Server安装在容灾管理服务器上，提供容灾业务流程控制、容灾策略调度、容灾资源、容灾业务配置数据管理。其中eReplication Server的UI子系统提供容灾业务管理图形化操作界面。

产品特性

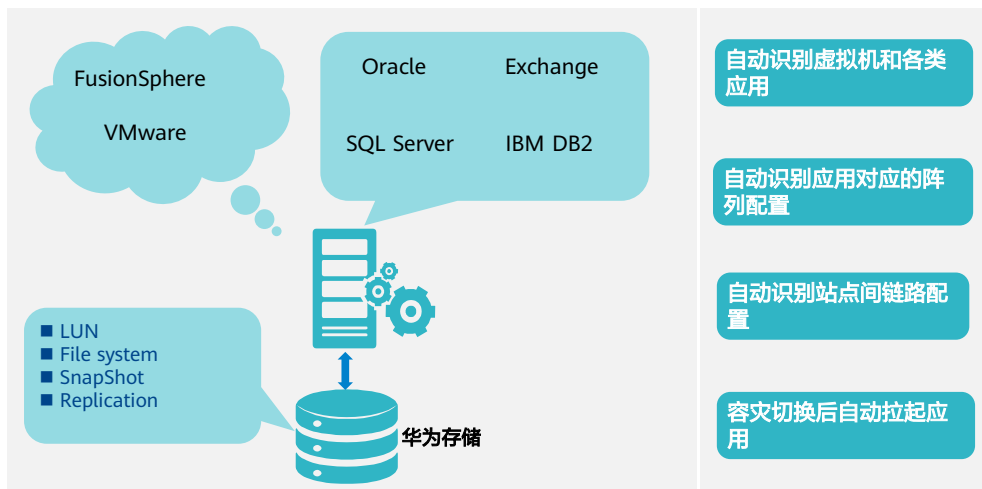


业务一体化



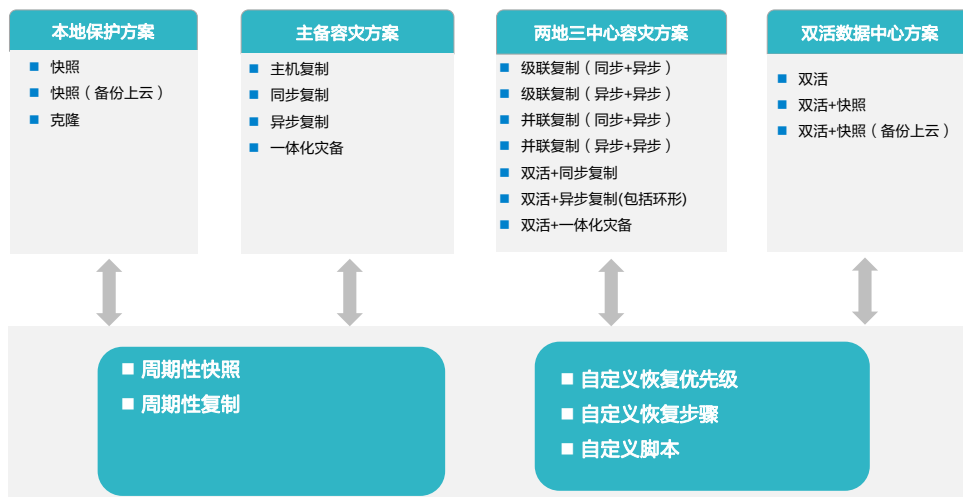
- 自动识别出应用、待保护对象（数据库实例）以后，会显示相应的详细信息，便于客户简单灵活的选择保护对象，降低管理难度，同时减少对业务应用的影响。
- 能自动识别出跟应用相关联的阵列以及阵列内的LUN、文件系统。

应用自动识别与拉起



- 自动识别出应用、待保护对象（数据库实例）以后，会显示相应的详细信息，便于客户简单灵活的选择保护对象，降低管理难度，同时减少对业务应用的影响。
- 能自动识别出跟应用相关联的阵列以及阵列内的LUN、文件系统。

模板化策略配置



- BCManager提供了4大类丰富的预置模板，满足各种组网环境下的需求。
- 配合华为存储阵列，提供周期性快照、周期性复制、备份上云功能。

目录

1. OceanProtect专用备份存储

2. OceanStor BCManager

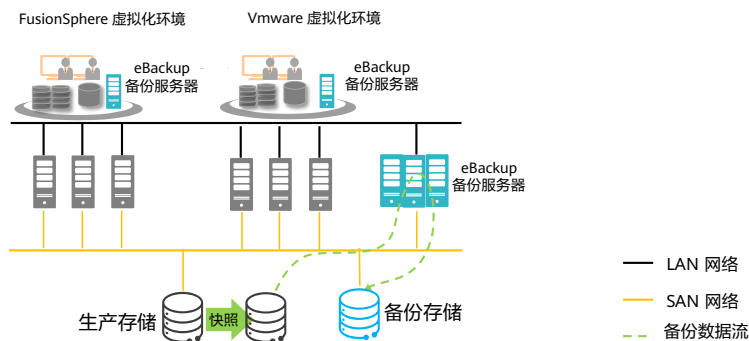
- eReplication

- **eBackup**

3. 应用场景

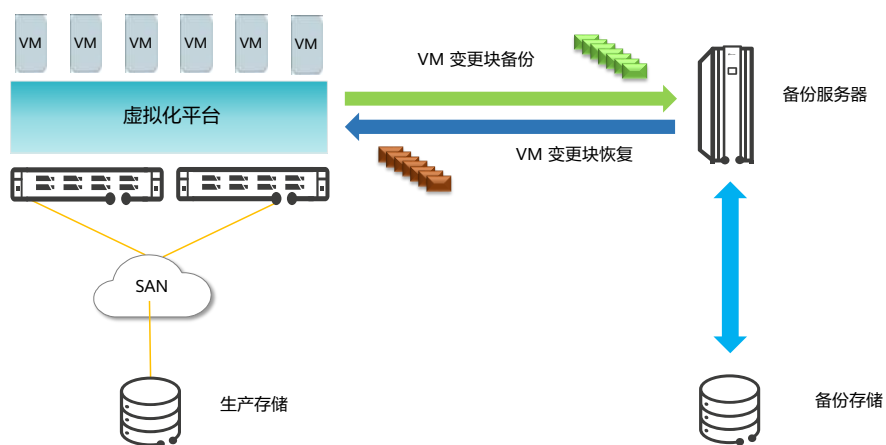
OceanStor eBackup产品定位

- eBackup是一款针对华为FusionSphere和VMware vSphere虚拟化平台的备份软件，基于虚拟机快照和CBT（change block tracking）技术，对虚拟机数据提供全面的保护。它是一个满足海量虚拟机备份场景、简单易用、性价比极高的数据保护方案。



- 满足海量虚拟机备份常见需求
 - 备份服务器与备份代理各司其职，分布式可扩展架构。
 - 虚拟机智能选择，自动保护新增虚拟机，减轻维护工作量。
 - 任务负载均衡、故障切换，保证可靠性。
- 备份管理简单易用
 - 提供快速入门说明，简化管理复杂度，轻松上手。
 - 提供配置向导，四步完成虚拟机保护配置。
 - 备份系统运行状态都可直观的在首页监控。
- 性价比极高的数据保护方案
 - 通过备份数据重复删除、压缩，永久增量备份等技术，降低35%用户备份存储购置成本。
 - 支持块级增量备份/恢复，缩减95%备份恢复窗口。

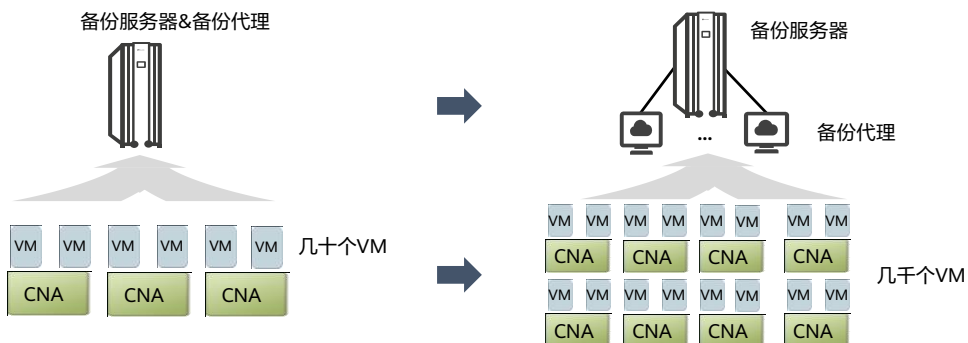
无代理VM快照备份



- 支持在线备份，不需要虚拟机停止业务
- 通过虚拟机快照实现备份，无需在虚拟机上安装代理，减少或者消除对虚拟机影响
- 支持VM块级增量备份/增量恢复，极大的提高备份/恢复效率，缩短备份/恢复窗口

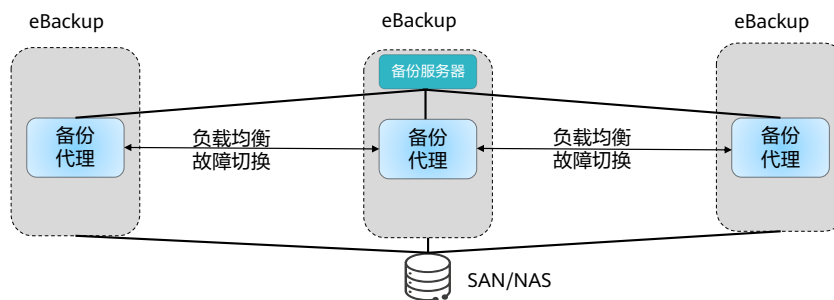
分布式部署、弹性扩展

- eBackup备份系统，采用分布式架构，可弹性扩展
- 备份服务器作为备份系统中的控制中心，对备份系统进行集中管理
- 备份代理负责具体事务处理，随着虚拟机数量增加，可弹性扩展，满足未来业务发展需求



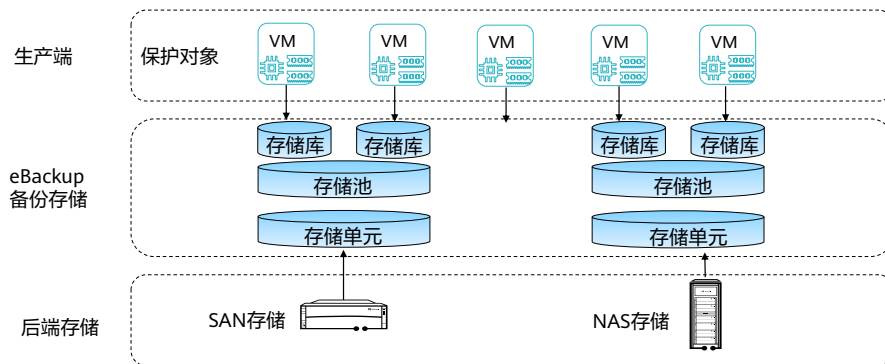
高可靠的备份系统

- 支持备份任务在不同备份代理之间负载均衡和故障切换，提高备份任务可靠性。
- 提供备份任务失败重试机制。
- 提供备份服务器自身数据库备份机制，备份服务器故障后可迅速恢复。



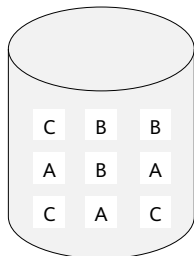
灵活的存储管理

- 支持SAN、NAS作为备份介质。
- 备份存储按照存储单元、存储池、存储库分层管理。业务增长后，可以通过修改存储库配置容量实现扩容（存储池有富余空间），无需调整备份计划，减轻维护工作量。

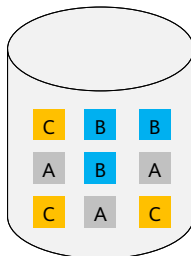


重复删除

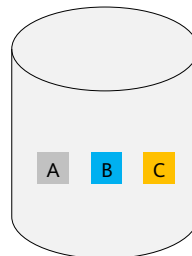
- 作为一种数据压缩技术，重复数据删除通过自动查找重复数据，对于相同的数据只保留一个副本，并使用指向单一副本的指针替换其他重复副本，从而消除冗余数据。
- eBackup备份管理系统支持重复数据删除技术，以减少对后端存储空间的消耗，满足日益增长的存储容量需求。



数据分块



相似性检查



重复数据删除

备份方式

- 第一次备份（全量）
 - eBackup备份管理系统对磁盘启动第一次备份任务，默认执行全量备份，将磁盘中全部数据进行备份。
- 后续备份（增量）
 - FusionSphere场景当存储类型为虚拟化存储时，增量数据由CBT记录。通过生产端提供的CBT方式获取前后两次备份之间的数据变更信息，从磁盘中取出有变更的数据进行备份。
 - FusionSphere场景当存储类型为FusionStorage时，通过对比生产端磁盘的前后两次快照获取前后两次备份之前的数据变更信息，从磁盘中取出有变更的数据进行备份。

目录

1. OceanProtect专用备份存储
2. OceanStor BCManager
- 3. 应用场景**

目录

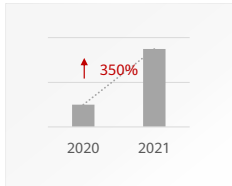
3. 应用场景

- 存储防勒索解决方案
- 统一备份解决方案

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%
- 在过去的一年中，新型勒索软件变种的数量增加了 46%。

网络威胁最热问题



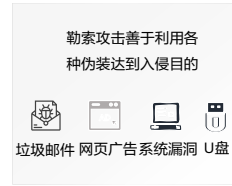
- 《2022 年全球网络安全展望》勒索软件攻击是全球网络领导者在网络威胁关心问题中排第一；

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。

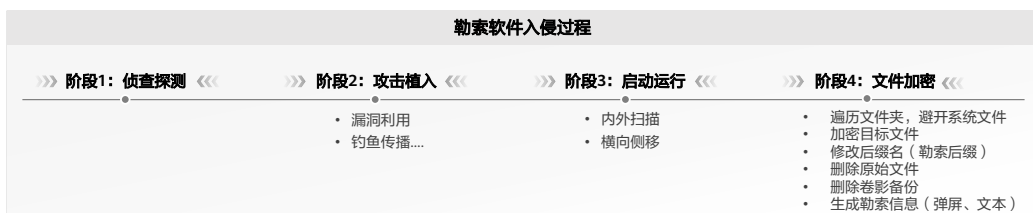
隐蔽性极强



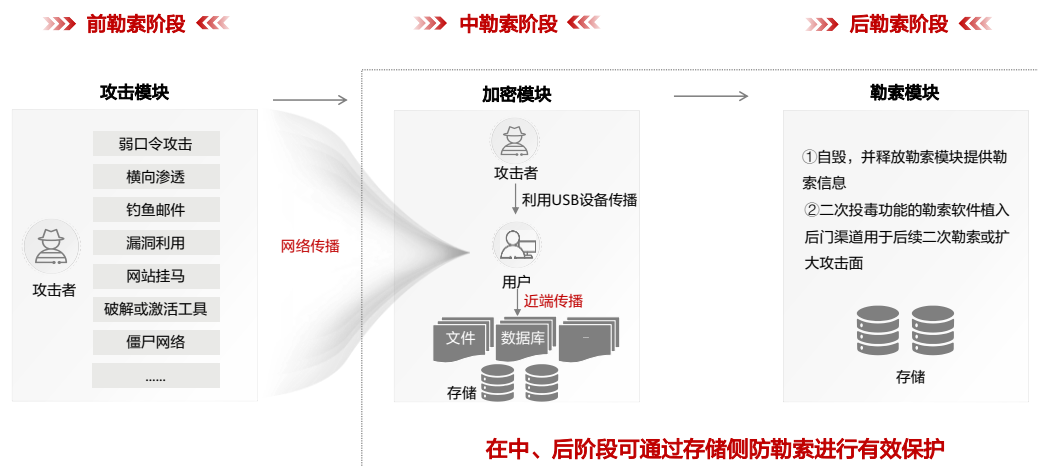
- 钓鱼式攻击：透过伪装获得敏感信息攻击
- 暴力破解：穷举常见弱口令密码破解
- 零时差漏洞攻击：系统发现漏洞前发起攻击

- 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

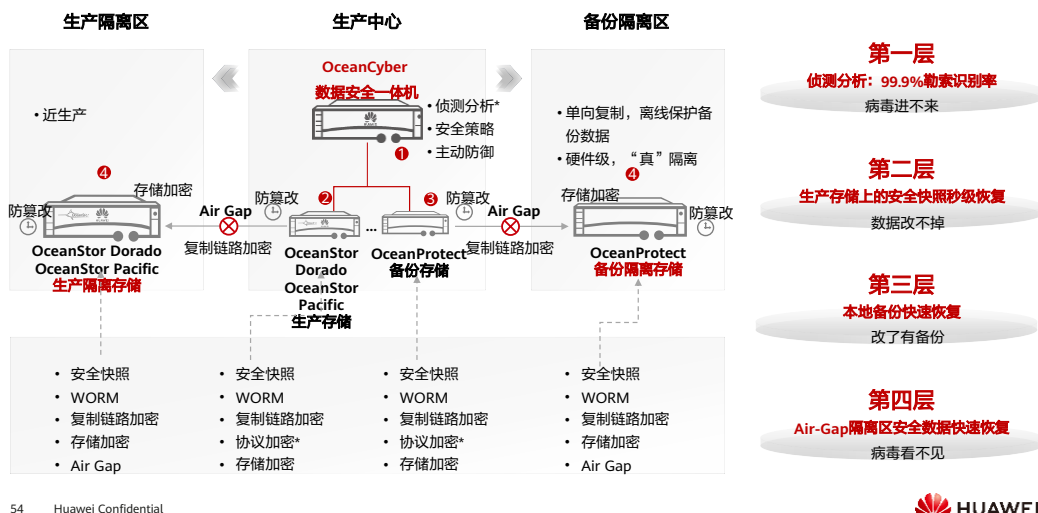


勒索软件分三个阶段进行攻击，存储是最后一道防线



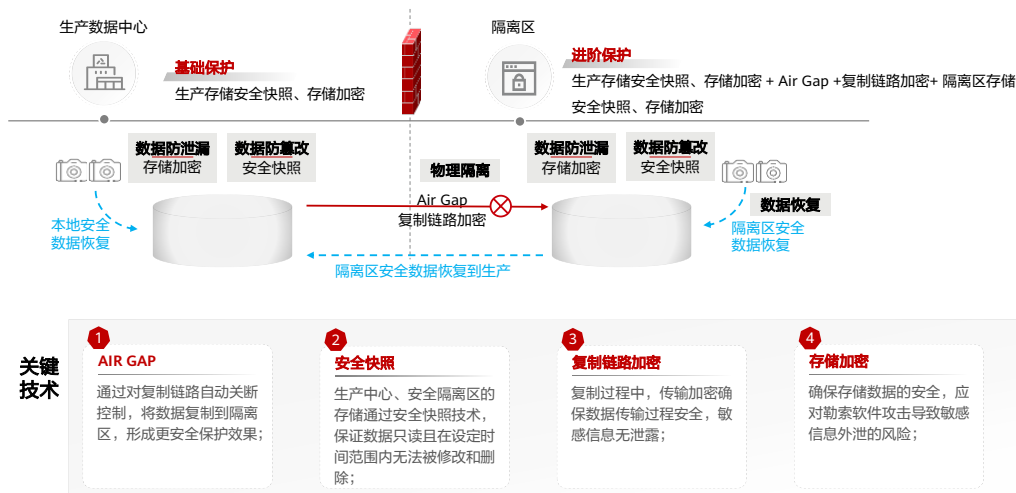
- 勒索病毒攻击模块主要通过携带各类漏洞exploit、密码字典等形式，对攻击目标进行漏洞检测以及服务口令爆破等。2020年起勒索病毒的攻击手段全线“开花”，弱口令攻击、横向渗透、钓鱼邮件、漏洞利用、网站挂马、破解或激活工具、僵尸网络、供应链攻击等手段悉数登场。
- 当攻击模块成功获取本地权限后，加密模块将开始运行，通过公钥加密或对称密钥加密等形式进行本地文件加密操作。具有窃密行为的勒索病毒还会在该环节进行特定格式文件窃取上传。在该阶段勒索病毒将对文件造成真实损害，因故称之为中勒索阶段，勒索病毒执行加密模块后会加密特定类型的文件，不同的勒索病毒会加密几十到几百种类型的文件，基本都会包括常见的文档、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4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



- 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。
- 下面我介绍一下我们是如何做到的。
- SAN存储有4大技术可有效防御勒索攻击：

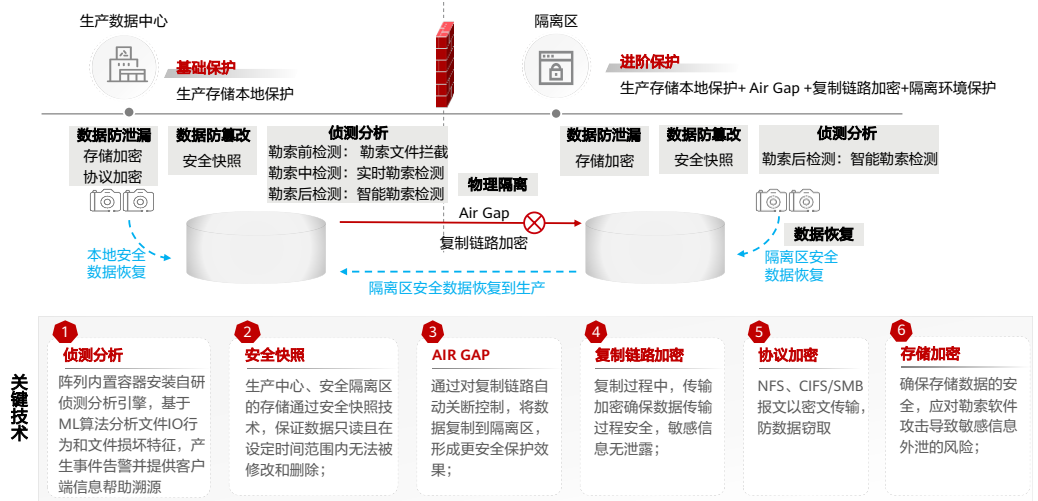
1. Air-Gap技术构建一个物理隔离的安全保护区，我们称之为隔离区。在隔离区内，部署一套生产存储，用于生产数据的备份和保护。部署隔离区需要配置华为BCManager软件，该软件部署在隔离区内，控制生产存储到生产隔离区的链路的定时通断，在需要备份数据时，链路up，在备份完后，相应的端口down掉，实现物理上的隔离。Air-Gap技术可有效保护隔离区不被黑客攻破。

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。

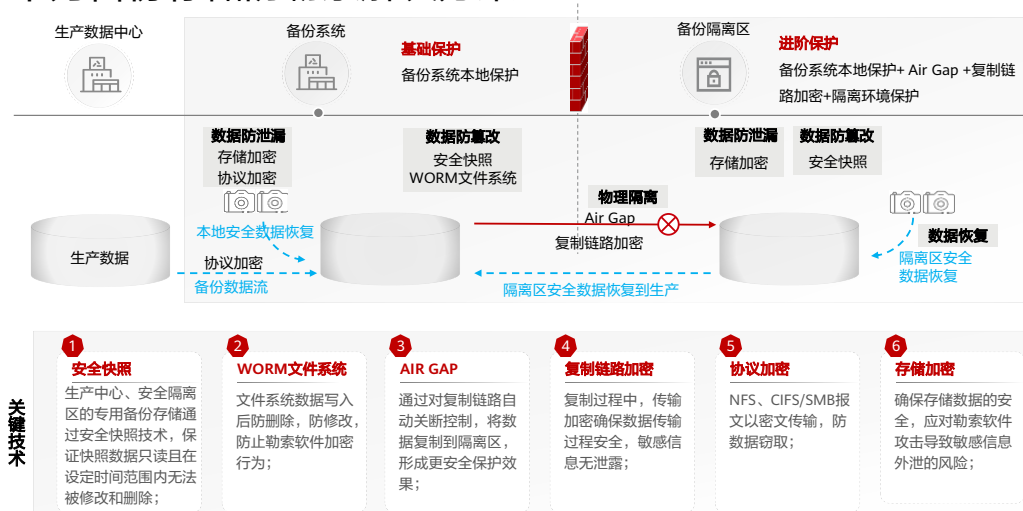
4. 存储加密，在生产存储、生产隔离区的存储进行加密，保障即使硬盘被盗也无法从存储中窃取数据。

华为NAS存储防勒索解决方案



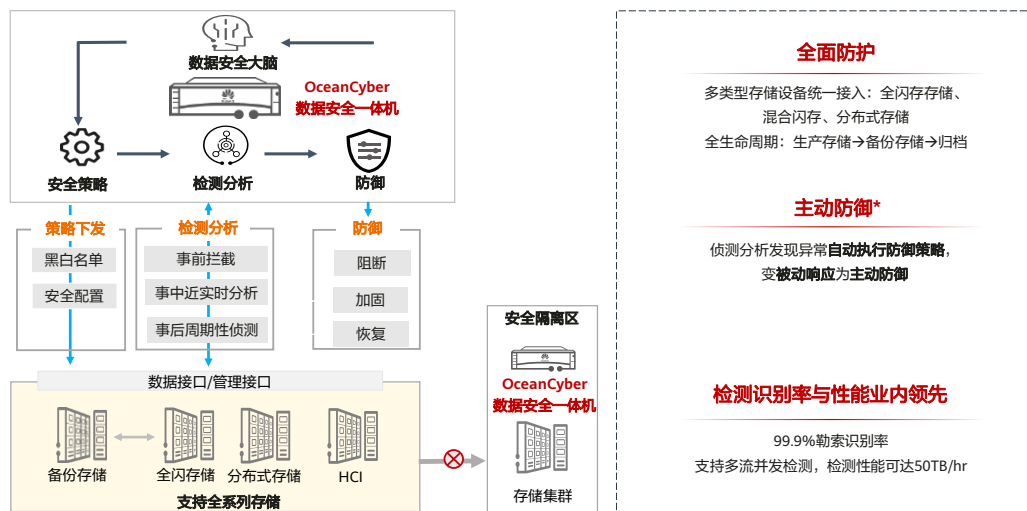
- NAS存储防勒索同样分为基础保护和进阶保护。NAS存储相对SNA存储多了侦测分析技术，即勒索检测技术，可以及时有效的检测到勒索病毒的攻击行为，并主动触发对数据的保护行为。
- 侦测分析是华为开发的，业界准确率第一的勒索检测技术。侦测分析可以部署在生产存储和生产隔离区的存储上，它的准确度高达99.9%。侦测分析总共分为三个阶段：
 - 事前（勒索病毒发动攻击前）—勒索文件拦截，可以识别勒索软件特征，阻止勒索软件对数据的破坏行为，达到拦截勒索攻击的作用；
 - 事中（勒索病毒加密数据过程中）—实时勒索检测，可以在线监测勒索病毒攻击的IO行为，发现勒索攻击后，会立即触发安全快照功能对检测到勒索攻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勒索解决方案



目录

5. 应用场景

- 存储防勒索解决方案
- 统一备份解决方案

灾备建设走向集约化、归一化

政务先行

建设现状

建设目标

《全国一体化政务大数据体系建设指南》

等保三级还有差距，需全栈方案、统一规划

安全、低碳的政务灾备体系，实现统一监管

（四）主要任务

合理利用全国一体化大数据中心协同创新体系，完善政务大数据算力管理措施，**整合建设全国一体化政务大数据体系主节点与灾备设施**，优化全国政务云建设布局，提升政务云资源管理运营水平，提高各地区各部门政务大数据算力支撑能力。

十四五规划



9个国家级灾备中心
14个省级灾备中心

X局备份系统



数据无专业备份：

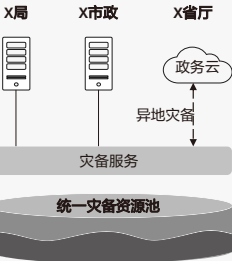
- 脚本不专业
- 本机不可靠
- 不安全

X市政备份系统



数据无异地备份：

- 不满足等保要求
- 多架构烟囱，不低碳，运营复杂



核心诉求：

数据安全可靠、统一服务化运维运营、绿色低碳

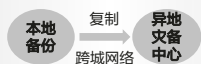
统一备份关键能力诉求

高效备份与恢复

① **本地高效备份**：确保在业务闲时完成数据备份，而不占用业务高峰期资源，尽可能降低本地备份窗口，留出更多时间用于异地远程复制，以减少跨城网络带宽需求。



② **远程高效复制**：生产数据本地备份后，还需跨城远程复制到异地灾备中心，需要在有限的跨域带宽下尽可能提升复制效率，满足复制窗口要求。



③ **即时挂载恢复**：为了满足数据应急恢复需求，降低RTO，需要支持直接将备份副本挂载到生产云主机或其它云主机，副本挂载后数据立即可用。

全面备份能力

① **极致重删压缩**：为节省备份存储空间，节省本地备份网络带宽占用，降低跨城远程复制带宽成本，需要端到端支持源端重删、复制链路重删、备份介质重删压缩能力。



② **数据安全防护**：勒索病毒日益猖獗，备份系统是数据安全的最后一道防线，需要支持端到端防勒索能力，确保遭勒索攻击后数据可快速恢复。



③ **备份服务化**：备份系统需要支持租户服务化，包括租户隔离、租户配额计量等，支持按照重删压缩前备份数据配额计量等灵活的商业变现模式。

全场景备份生态

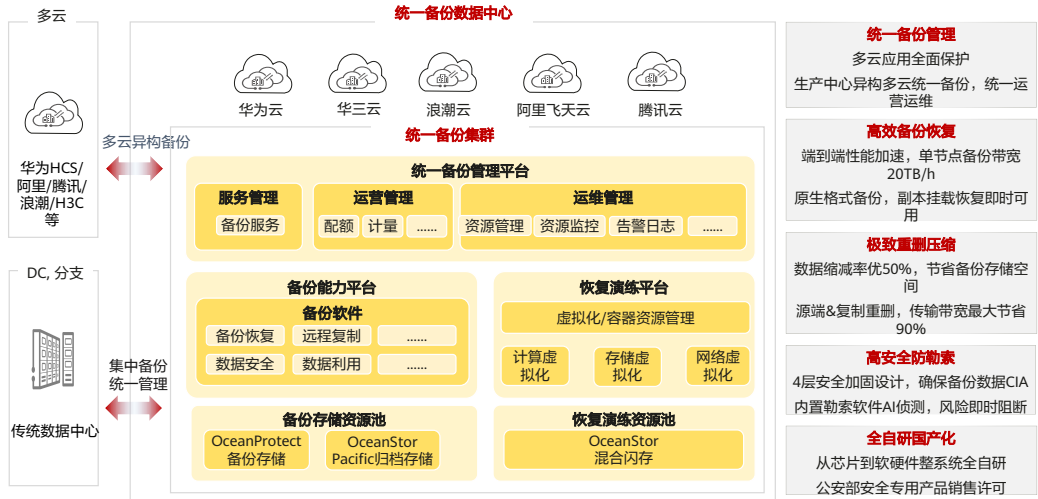
① **多云统一备份**：需要支持全省政务云数据的统一备份，屏蔽生产云技术架构差异。



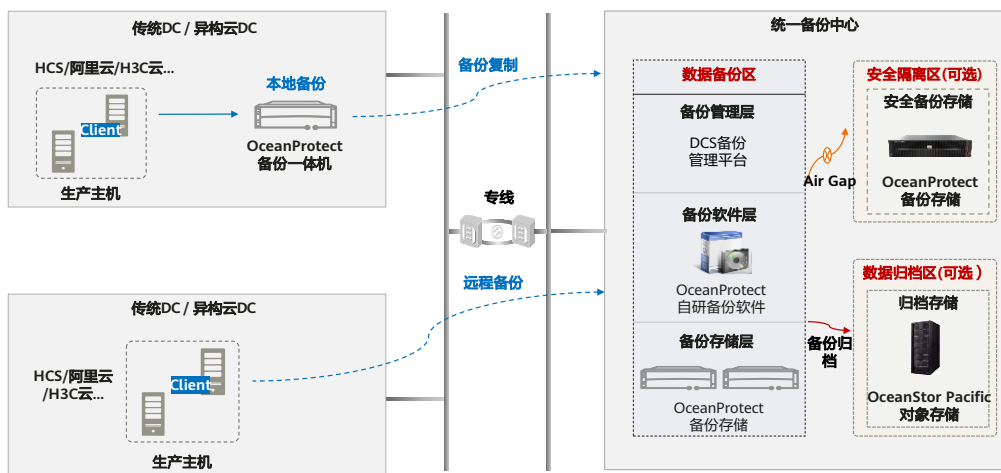
② **全面备份生态**：需要支持全面的应用生态备份，例如主流数据库、文件、应用、大数据组件、云主机等的备份。



多云统一备份中心实现备份中心集约化



统一备份解决方案



- 备份网络：增量备份带宽=增量备份数据量/增量备份窗口
- 复制网络：复制带宽=每日复制总容量*链路缩减率/复制窗口/带宽利用率（一般为80%），复制网络要求单向时延<50ms

思考题

1. (多选) OceanProtect X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

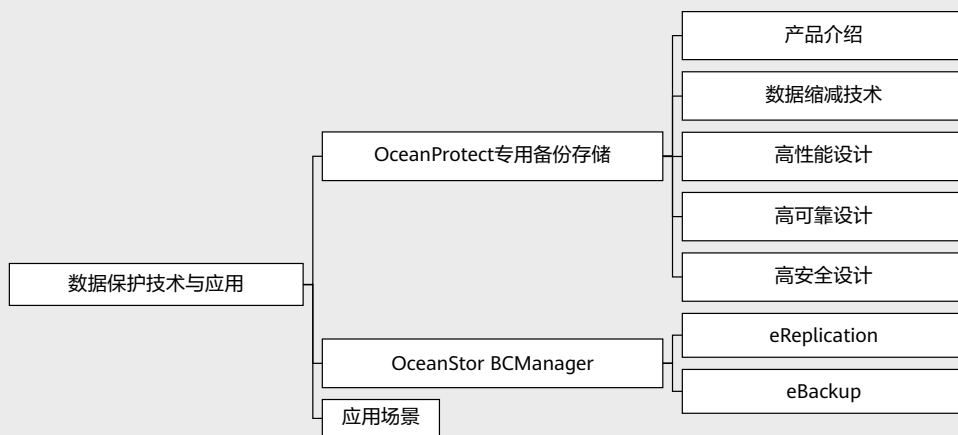
思考题

2. (多选) 关于数据重删的说法，以下哪些选项是正确的？（ ）
- 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



更多信息

- 华为存储信息自助服务平台：
 - https://enterprise.huawei.com/topic/Self_service_2013/knowledge.html
- 企业业务：<https://enterprise.huawei.com/cn/>
- 技术支持：<https://support.huawei.com/enterprise/>
- 在线学习：<https://support.huawei.com/learning/Index!toTrainIndex>
- 华为Support案例库：<https://support.huawei.com/enterprise/servicecenter?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.

Copyright©2023 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.



分布式存储产品技术与应用



前言

- 华为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，是既具备云基础架构的弹性按需服务能力、又满足企业级关键业务和新兴业务需求的存储系统。
- OceanStor Pacific系列存储通过存储系统软件将每个硬件节点的本地存储资源组织起来，按需为上层应用提供文件存储、对象存储、大数据存储和块存储服务。实现海量数据储存更经济、多样性数据使用更高效、在线业务承载更可靠。

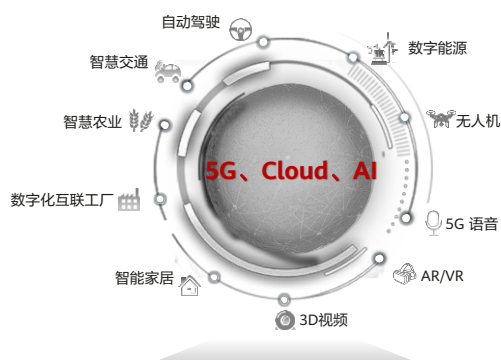
目标

- 学完本课程后，您将能够：
 - 了解分布式存储概念和发展趋势
 - 掌握OceanStor Pacific产品软硬件架构
 - 掌握OceanStor Pacific产品的关键技术
 - 掌握OceanStor Pacific产品的关键特性
 - 熟悉OceanStor Pacific产品的应用场景

目录

1. 产品介绍
2. 关键技术
3. 关键特性
4. 应用场景

YB数据时代到来



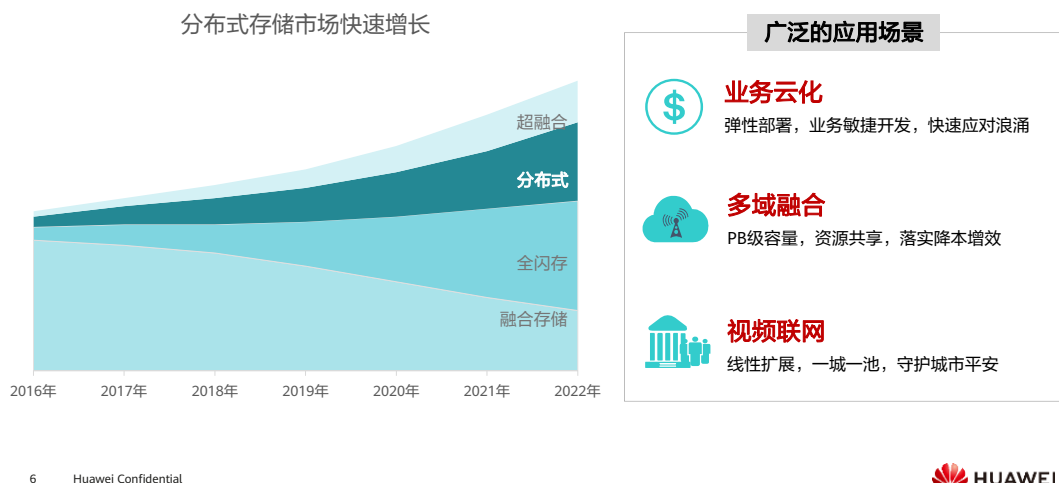
千亿	30x	TB	➡	PB
物联网	交通	传统数据		视频影像数据
5G, 8K, VR		企业		
64 TB		1 PB		
每日训练数据量		每日生产数据量		
自动驾驶		数字化互联工厂		

到2030年，全球每年产生的数据总量将突破 **1YB量级**

$1\text{YB} = 1000\text{ZB} = 1000,000\text{EB} = 1000,000,000\text{PB} = 1000,000,000,000\text{TB}$

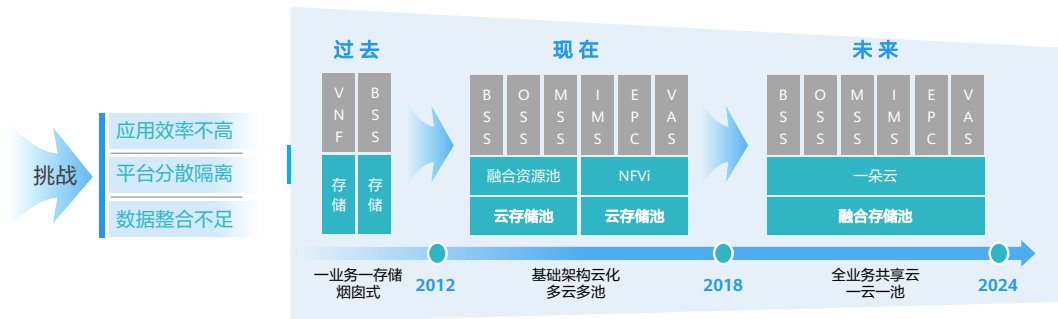
- 自动驾驶、数字能源、AR/VR等全球智能化应用加速数据的产生和流动。例如自动驾驶每日训练数据量达到64TB，数字化互联工厂每日生产数据量达到1PB。预估到2030年，全球每年新增数据量将突破1YB量级。
- 数据是数字经济时代的新生产要素，存取海量数据的地方，一定要用专业的海量数据存储。

数据整合与业务云化



- 在二十世纪第一个十年，随着互联网数据爆发，基于分布式架构的数据存储成为了应对海量数据增长的一种有效手段。
- 在第二个十年，随着企业拉开了数字化转型序幕，数据快速增长，海量数据存储在企业市场迎来了飞速发展。
- 几个典型的应用场景：
 - 银行、ISP基于业务创新和敏捷的目标，全面云化转型。
 - 运营商提出降本增效目标，稳步推进BOM域的设备 and 数据整合。
 - 公共安全成为人类共同的挑战，驱动的视频联网应用快速发展。
- 这些应用带来的数据快速增长需求，也成为了过去十年海量存储在企业市场普及的主要推动力。

存储池化

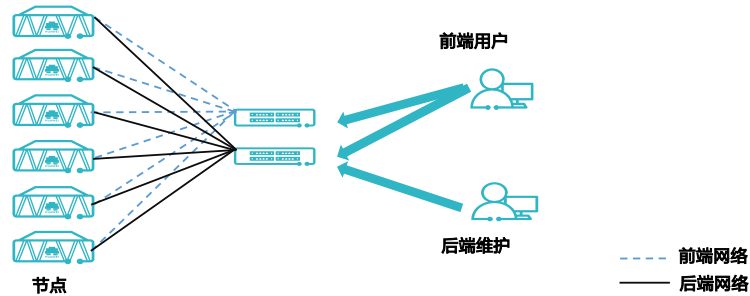


运营商业演进催生IT基础架构从离散走向融合：存储资源池化

- 存储池化是未来云化建设的必经阶段。

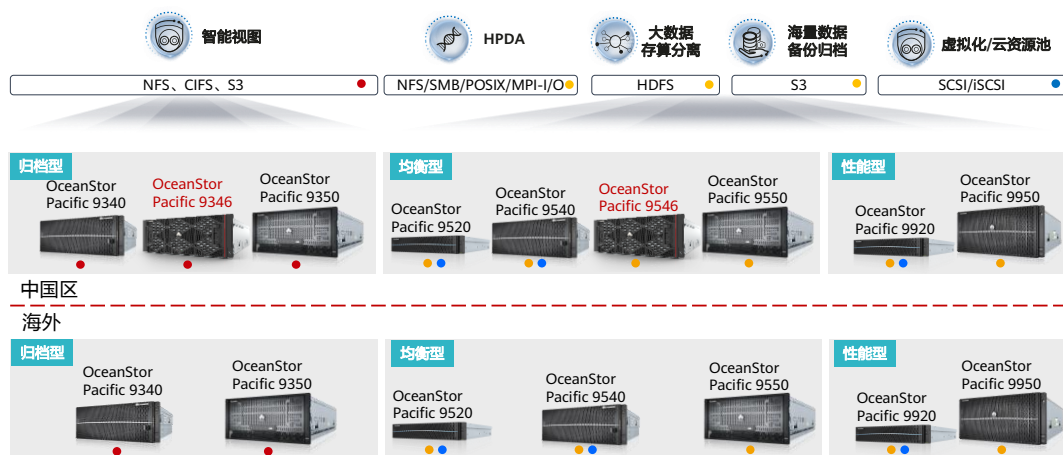
什么是分布式存储

- 分布式存储简单来说，将通用服务器的本地HDD、SSD等存储介质组织成一个大规模的存储资源池，然后将数据分散存储到多个数据存储服务器上。
- 分布式存储目前多借鉴Google的经验，在众多的服务器搭建一个分布式文件系统，再在这个分布式文件系统上实现相关的数据存储业务。



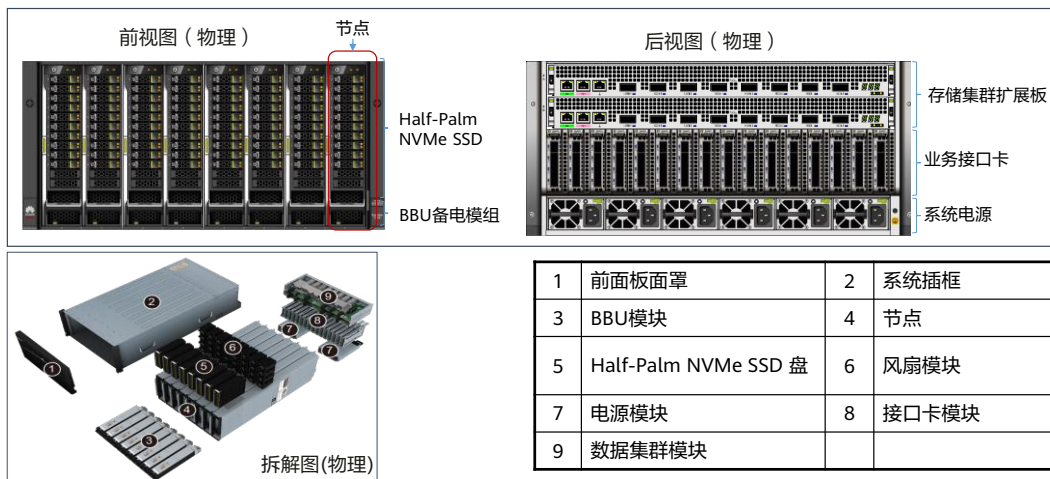
- 底层采用了分布式存储技术为上层应用提供服务。

OceanStor Pacific 产品概览



- 产品信息引用自2022Q3产品规格表，在实际的项目选型中，应以华为官网的产品信息为准。
- OceanStor Pacific全系列产品的规格：
 - OceanStor Pacific全系列产品分为性能型、均衡型和归档型，可采用分级存储的方式满足客户在不同场景的应用诉求。
 - 性能型整机以OceanStor Pacific 9950为典型代表，提供客户极致的海量存储全闪存性能，单机达到160GB/s带宽性能。
 - 均衡型整机以OceanStor Pacific 9550为典型代表，单整机提供120盘位，配套当前主流的14T硬盘，单机可提供1.68PB的存储容量。
 - OceanStor Pacific系列产品支持主流的海量存储网络接口，支持10/25GE、40/100GE，HDR-100/EDR IB接入能力，支持TCP/IP标准协议接入，也支持RDMA、RoCE协议接入。

OceanStor Pacific 9950产品物理形态

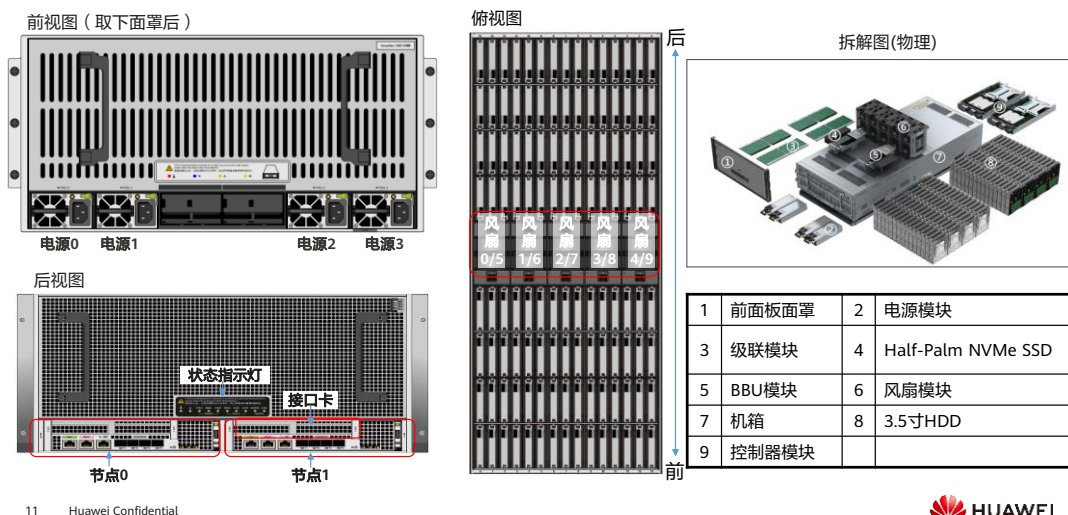


10 Huawei Confidential



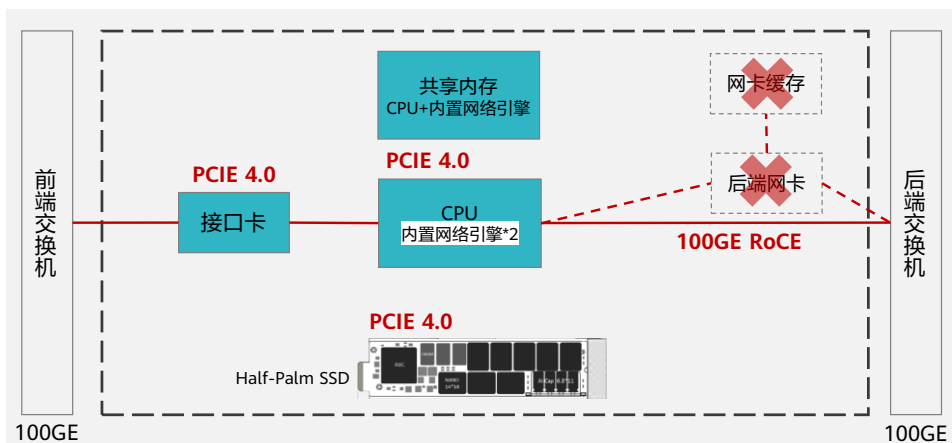
- OceanStor Pacific 9950是基于华为鲲鹏920处理器开发的5U 8节点全闪存高密整机，为客户提供极致的分布式存储性能。具有高性能、高密度、极致TCO、高可靠等特点。
- 整机的前面板由8个独立可插拔的节点组成，每个节点支持10 Half Palm NVMe SSD，整机支持80个Half Palm NVMe SSD盘。每个节点支持32GB保电内存空间，在系统异常掉电场景，提供数据保护功能。
- 整机的后面板由三部份组上，上部的两块数据集群模块，为每节点提供2个100GE RoCE双平面的存储后端网络接口，中间的16张接口卡模块用于前端网络接入，每两张接口卡归属一个节点，下部的6个电源槽位组成，提供2+2电源冗余可靠性。
- OceanStor Pacific 9950采用全FRU设计，整机的控制器节点、BBU模块、Half Palm NVMe SSD盘、风扇模组、数据集群模块、接口卡模块、电源模块均支持独立插拔，方便维护更换。

OceanStor Pacific 9550产品物理形态



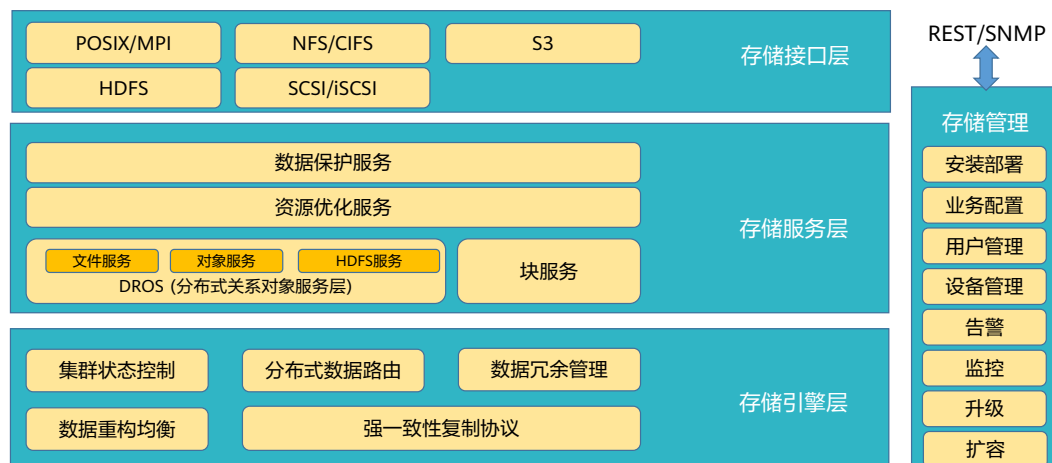
- 本页介绍OceanStor Pacific 9550存储机箱的物理视图：
- 前视图：包含网络状面防尘面罩及设备状态显示面板（最右是操作灯-指示是推进还是拉出，定位灯-更换哪类部件就琥珀色灯长亮（主存硬盘、风扇、BBU、Half-Palm NVMe SSD缓存盘），告警灯-前-盘、风扇、电源、BBU，硬盘，后面-节点控制板、后半框盘；过温灯；可见左右各两个电源模块。
- 后视图：可见两个存储节点集成板，节点0和节点1，每个节点上集成了1颗Kunpeng 920处理器，及各种制式的接口卡插槽；华为打破了业界前盘后控、垂直背板设计的惯性思维，采用了上盘下控的设计，每个集成控制板管理其上方从前至后的60个硬盘，通过柔性线缆——坦克链连接硬盘盒与CPU集成板。
- 俯视图：可见前后两个物理硬盘区域及中间的风扇集成区域；前后两排硬盘区域包含共计120个硬盘插槽、中间的风扇区域采用5*2上、下两层堆叠放置，形成双层风道。针对120盘位的高风阻，华为专门定制的10个80mm航空对旋增压风扇，形成了一道风扇墙，上下两层风道，上层风道为主风道，满足前排硬盘散热，下层风道满足CPU和后排硬盘散热。增压风扇前抽后吹，风道前后贯穿，完美解决了高密设备散热难题。相比采用业界主流的散热技术，OceanStor Pacific 9550的器件可靠性提升了1倍以上。
- OceanStor Pacific 9550采用全FRU设计，节点控制器、电源、BBU、halfpalm NVMe SSD缓存盘、级联模块、风扇模块、硬盘模块均支持独立插拔，方便维护更换。
- 风扇N+1冗余：5组风扇，每个风扇2个电机（转子），坏一个电机后，系统正常运行，若整个风扇坏了，就得立即更换。

前后端PCIe 4.0



- 全系统采用PCIe 4.0设计，业界最先推出的全PCIe 4.0全闪存系统，PCIe 4.0相比现有的PCIe 3.0带宽提升一倍，具备更大带宽，更低时延的优势。
- CPU的前端业务接口带宽分配的2个PCIe4.0X8的接口，提供的总带宽能力为20GB/s。采用免后端网卡设计，最大化地发挥了CPU的接入能力，节省了成本，同时后端数据集群接口采用CPU直出的2个100GE接口端口，该接口支持RDMA技术，极大地降低了网络时延，可以有效的将集群的数据，高效的传输到其他节点上，同时只消耗少量的CPU。本地闪存接口采用10个PCIe4.0X2的接口，使用的SSD接口协议是NVMe，使得CPU可以直接通过PCIe直接访问介质，相比传统的SAS协议，每次I/O时延可以降低50%以上。

功能架构

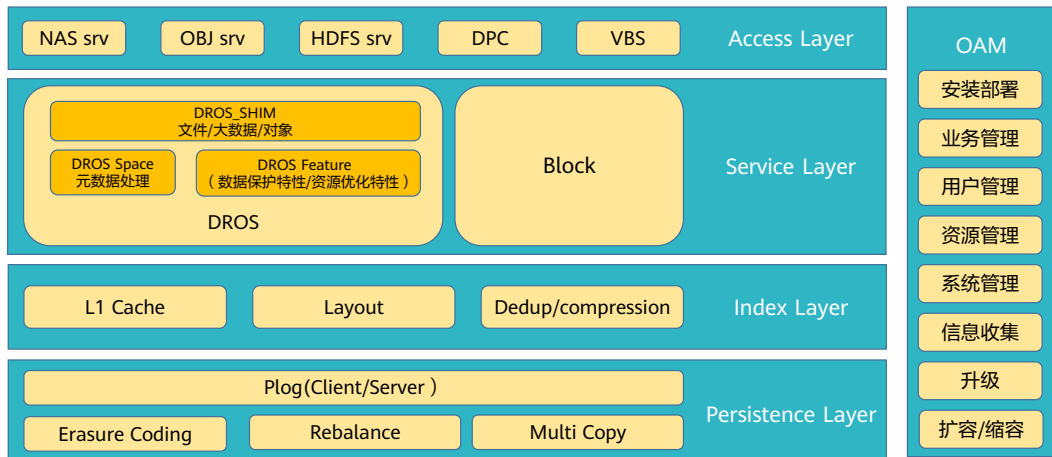


13 Huawei Confidential



- 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、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。

软件逻辑架构



- 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。

目录

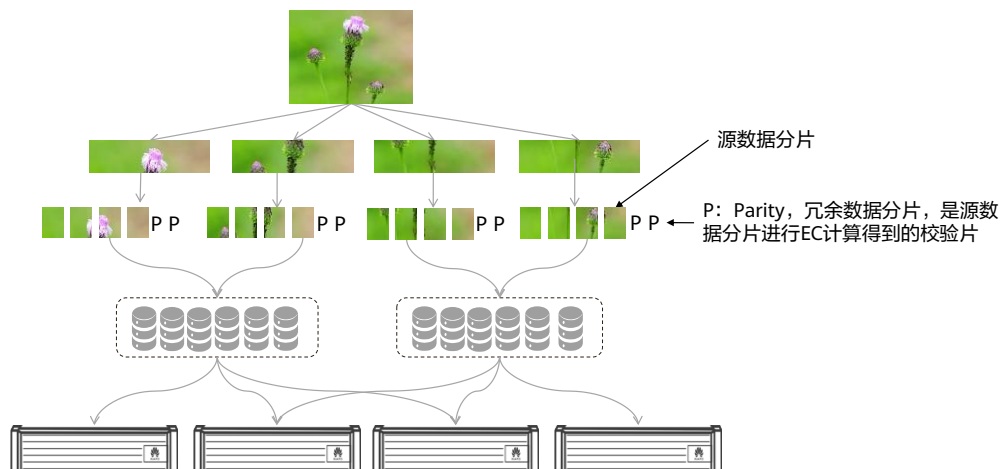
1. 产品介绍
- 2. 关键技术**
3. 关键特性
4. 应用场景

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。

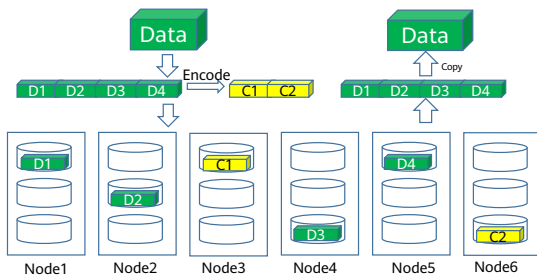
Erasure Coding特性原理



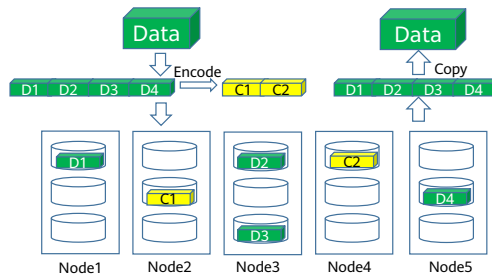
- 数据分片写到不同的节点上，不仅可以提高读写性能，又能保证数据的高可靠性和服务的高可用性，还能保持高的磁盘利用率，节约投资。
- 一个存储池内，任意不超过M（冗余份数）个硬盘故障后，系统都可以通过节点间数据重构的方式快速恢复出所损失的数据，进而恢复整个系统的数据可靠性。
- 系统中任意可用空间都可以作为“热备”空间使用，解决了传统RAID的热备盘问题，进一步提高了存储利用率。

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。



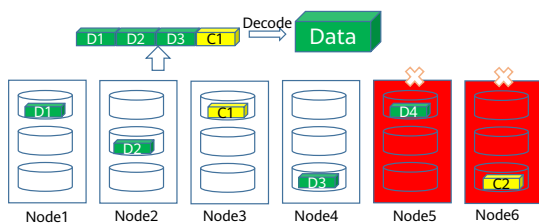
$N=4$ 、 $M=2$ ，6个节点 采用 $N+M$ 冗余配比



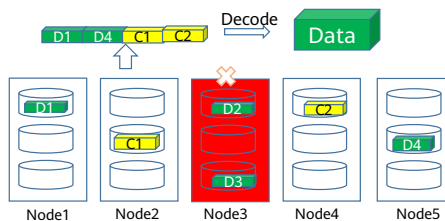
$N=4$ 、 $M=2$ ，5个节点 采用 $N+M:1$ 冗余配比

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。

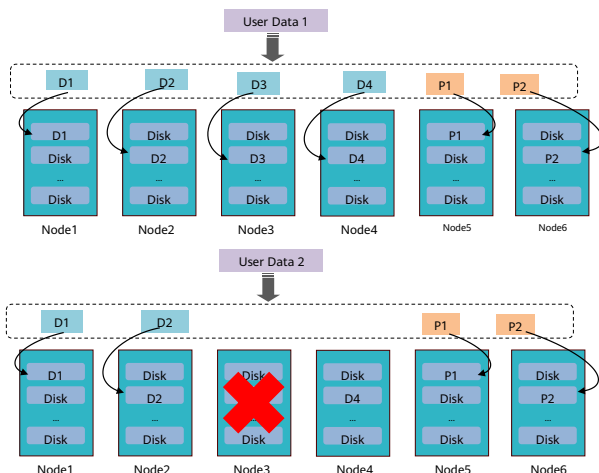


$N=4$ 、 $M=2$ ，6个节点 采用 $N+M$ 冗余配比



$N=4$ 、 $M=2$ ，5个节点 采用 $N+M:1$ 冗余配比

Erasure Coding Turbo功能



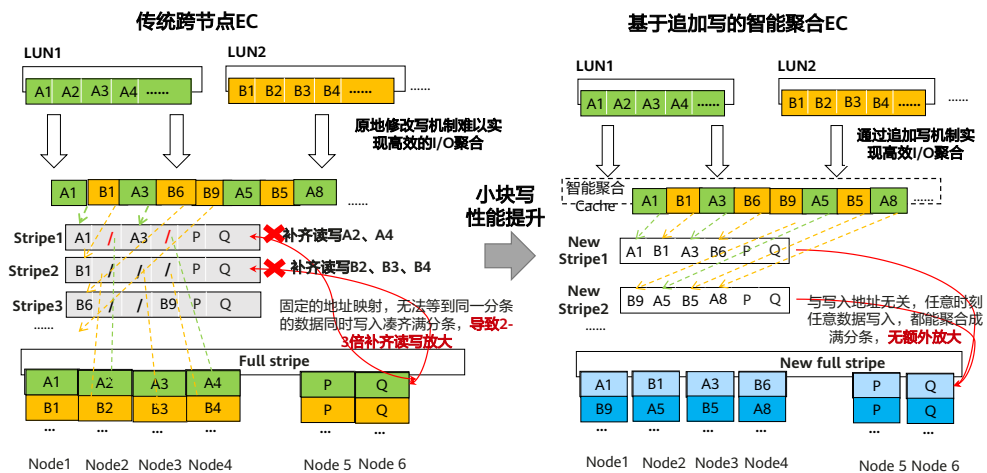
用户数据1被切成N（例如4）个数据分片另外再计算出M(例如2)个冗余分片 4+2。

当发生节点3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。

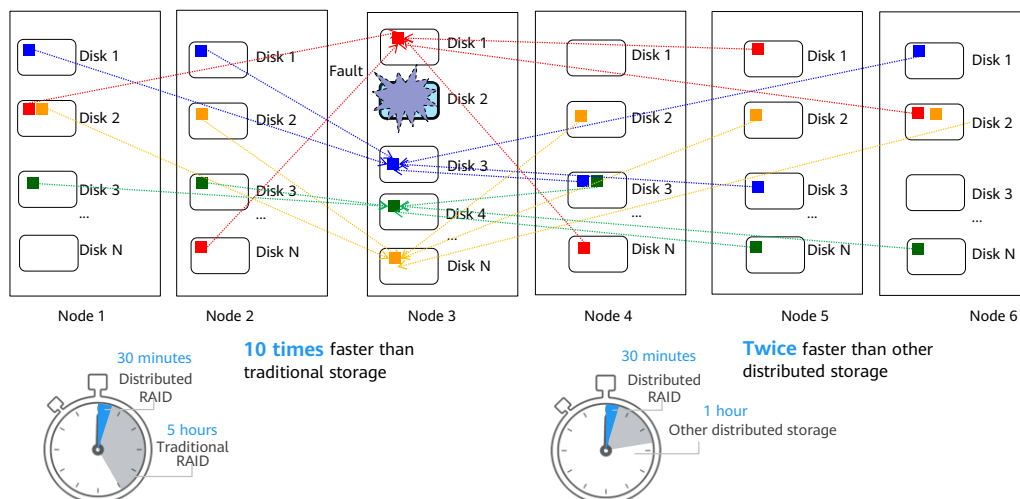
- OceanStor Pacific 支持动态 EC:

- 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



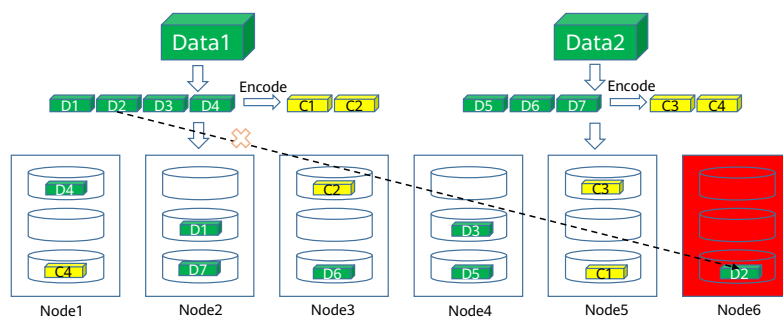
数据重构原理



- 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。

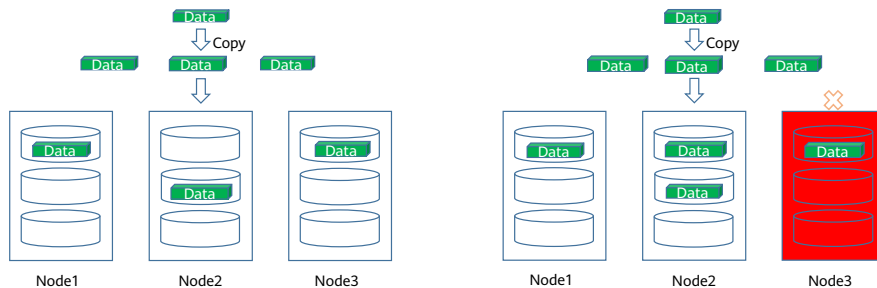
弹性EC的技术优势

- 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。

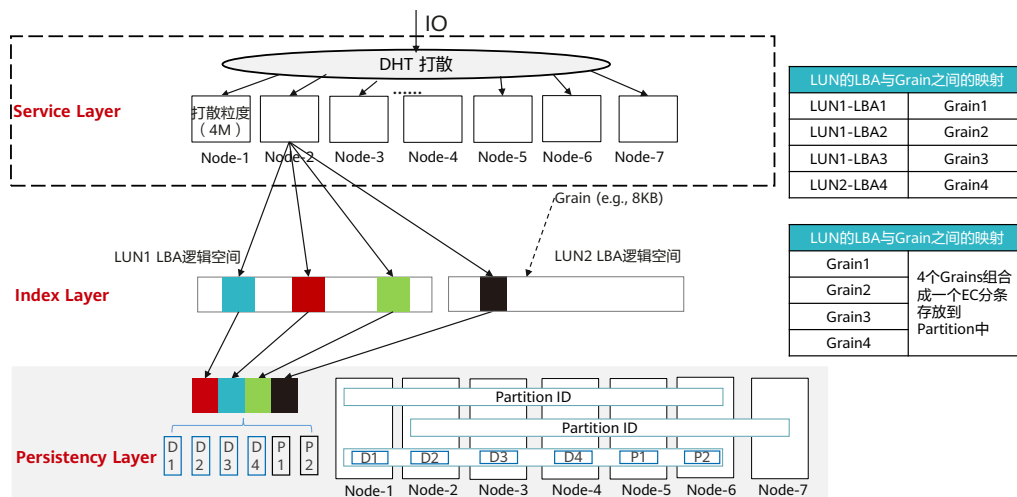


- 传统的硬盘级RAID模式将数据存放于单节点内的不同硬盘，当整节点发生故障时，无法有效恢复数据。存储系统将数据在节点间进行多副本构建，有效避免数据丢失。多副本是通过将相同的数据在不同的节点上存储多份来实现数据保护的一种技术，支持三副本和两副本，推荐三副本。
- 三副本的空间利用率为33.3%，两副本的空间利用率50%。
- 节点级安全：将多副本存储于不同的节点中，故障2个（三副本）或1个（两副本）节点或硬盘，系统仍可正常读写数据，业务不中断，数据不丢失。
- 机柜级安全：将多副本存储于不同的机柜中，故障2个（三副本）或1个（两副本）机柜、节点或硬盘，系统仍可正常读写数据，业务不中断，数据不丢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

DHT路由技术

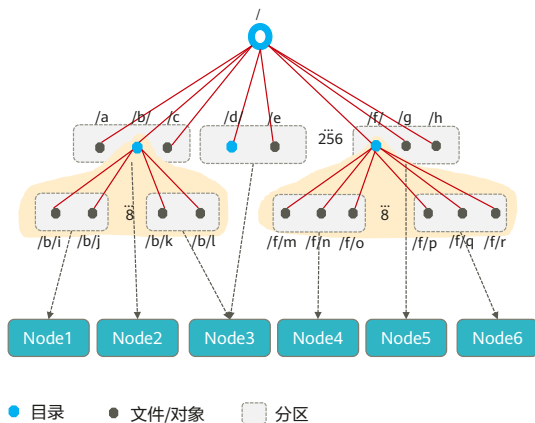


端到端负载均衡



- 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。

语义层负载均衡



- 将目录划分为多个逻辑分区，用于数据处理的负载均衡。
- 目录下的所有成员(子目录/文件/对象)按Hash算法映射到不同分区。
- 每个分区映射到不同节点，由相应节点处理该分区负载。
- 目录默认划分为8个分区(可调整到最大256)；作为对象服务的桶，根目录默认划分为256个分区，解决单桶海量对象的数据均衡问题。

- 目录分区打散，实现语义层负载均衡，消除热点。

DPC分布式并行客户端



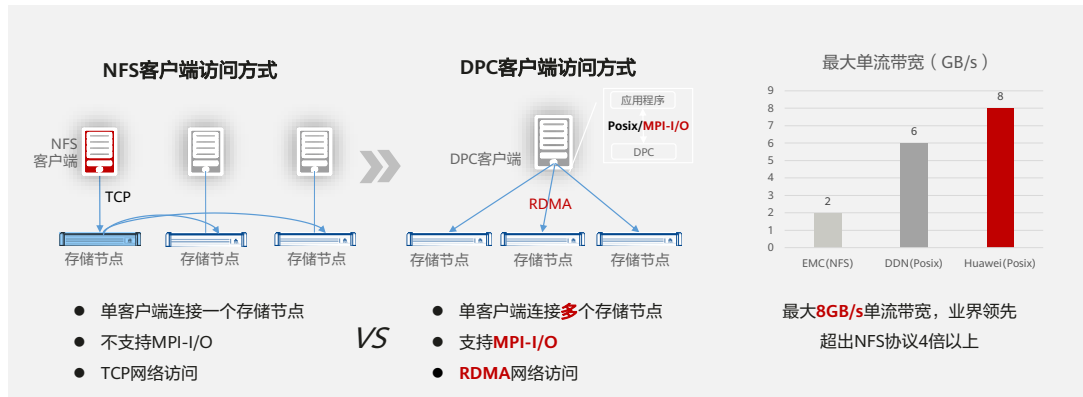
MPI-I/O实现气象
数据分析应用提速



单流高性能
保障卫星数据接收



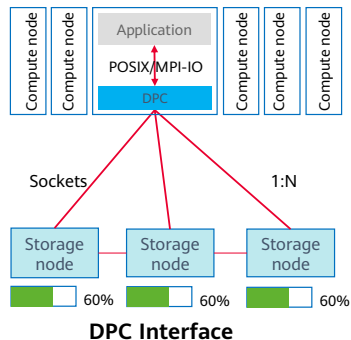
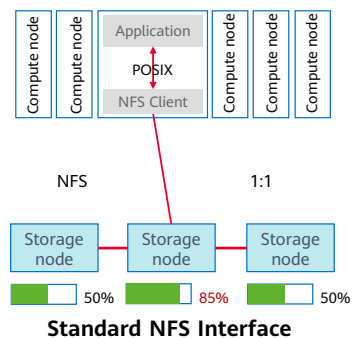
单客户端高性能释放
超算客户端潜能



- 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。

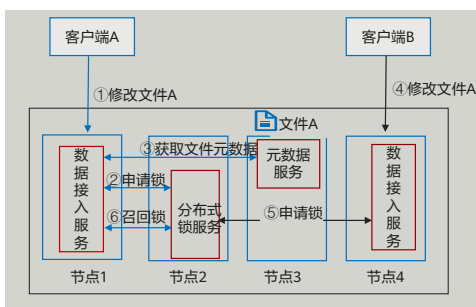
I/O级的负载均衡能力（文件服务）

- 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。



小I/O处理（非结构化服务支持）

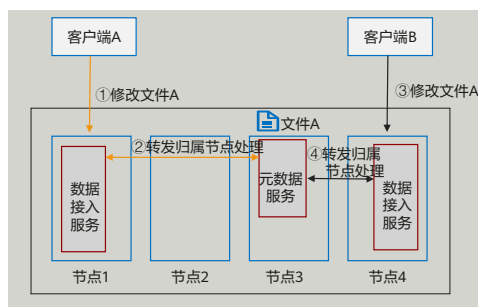
基于分布式锁的小I/O处理流程



- 元数据服务和分布式锁服务独立，多次交互
- 多客户端访问同一文件，涉及锁召回，进一步增加交互次数



基于归属转发的小I/O处理流程

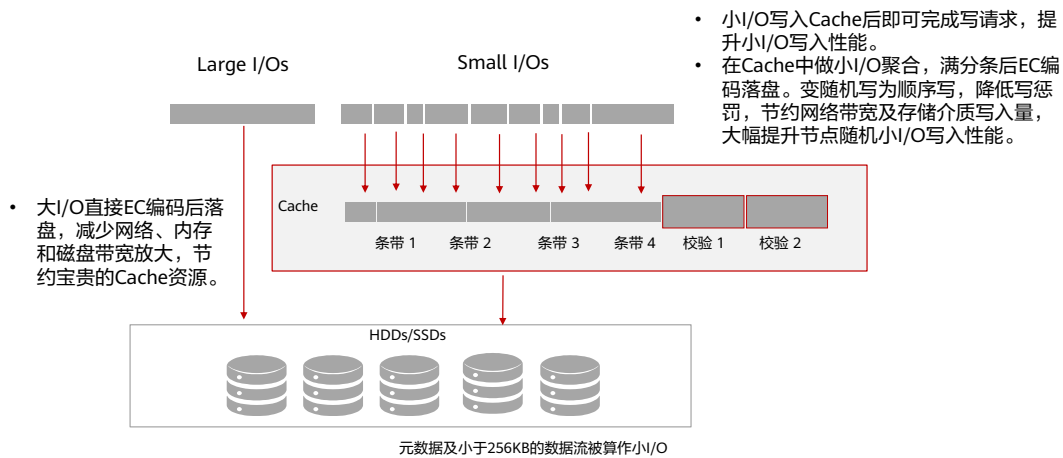


- 元数据有归属，无需分布式锁服务
- 多客户端访问同一文件，归属节点内部排队处理，无额外开销

VS

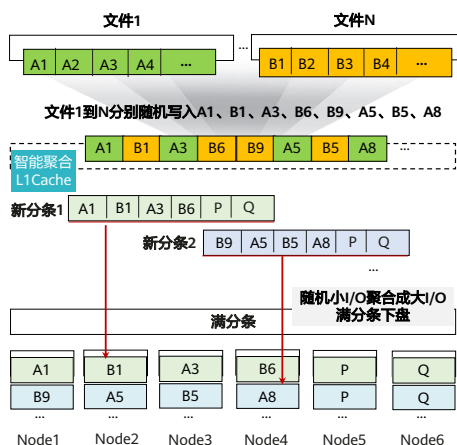
- 小I/O归属转发，避免分布式锁带来的多次交互开销。

智能I/O自适应



- 智能I/O自适应技术能同时兼顾大、小IO的不同性能诉求。

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

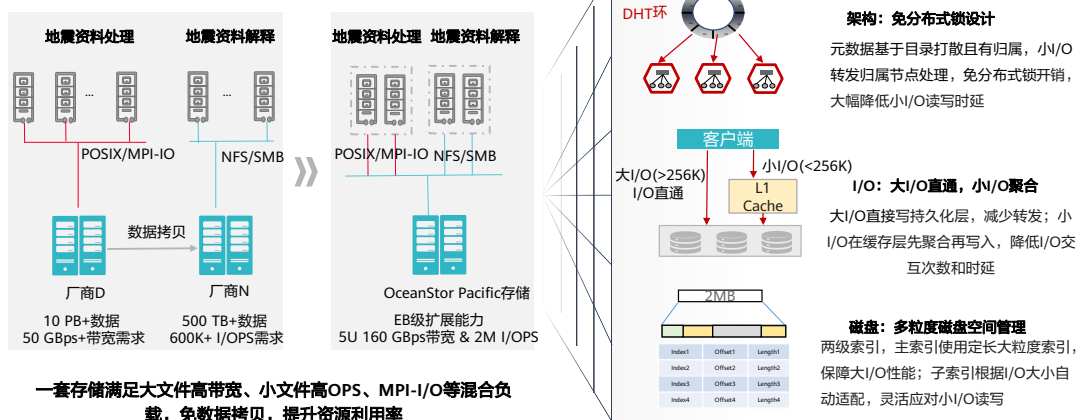
智能追加写

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。

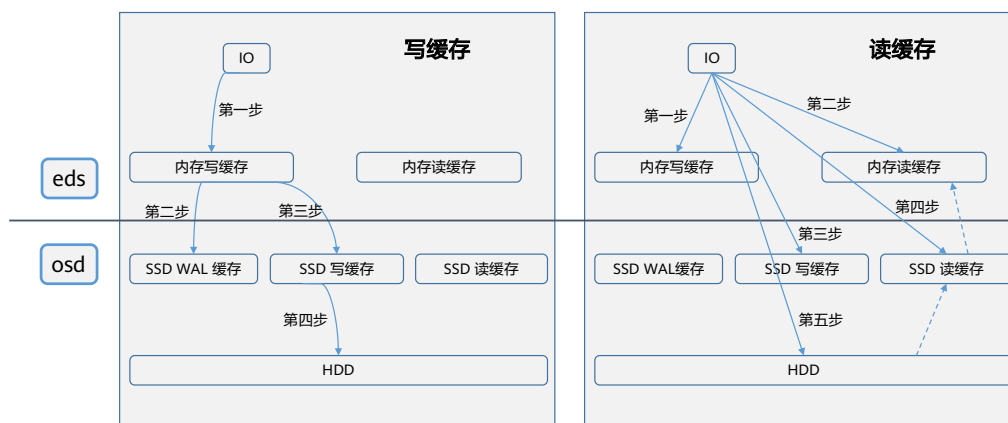
并行文件系统

• 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。

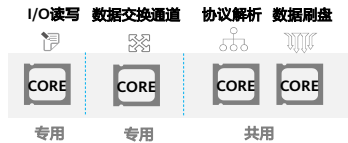
多级缓存技术



- OceanStor Pacific系列产品采用多级缓存技术逐层对IO进行聚合加速，提升系统整体IO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IO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。

FlashLink

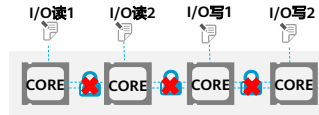
CPU分组算法



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

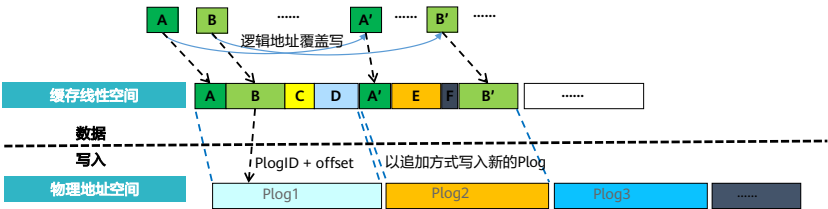
- 数据从I/O卡流入CPU。华为OceanStor Pacific系列的CPU利用算法具备三大显著优势：
 - CPU采用最新的智能分区算法，包括CPU分组算法和CPU分核算法。
 - CPU分组算法是把比较重要的I/O单独一个分组（比如：读写，数据交换），从而读写I/O与其他I/O部署在不同分组，避免相互干扰，保障读写I/O的性能；优先级相对低的I/O在一个分组内，共享CPU资源。这样能够保障资源利用的最大化。
 - CPU分核算法是华为独有的优势。I/O工作在CPU的时间片是轮询的，一段时间在CPU的Core1上工作，过一段时间轮询到Core2上工作。其一，从Core1跳转到Core2是要花时间的。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。

Append Only Plog技术

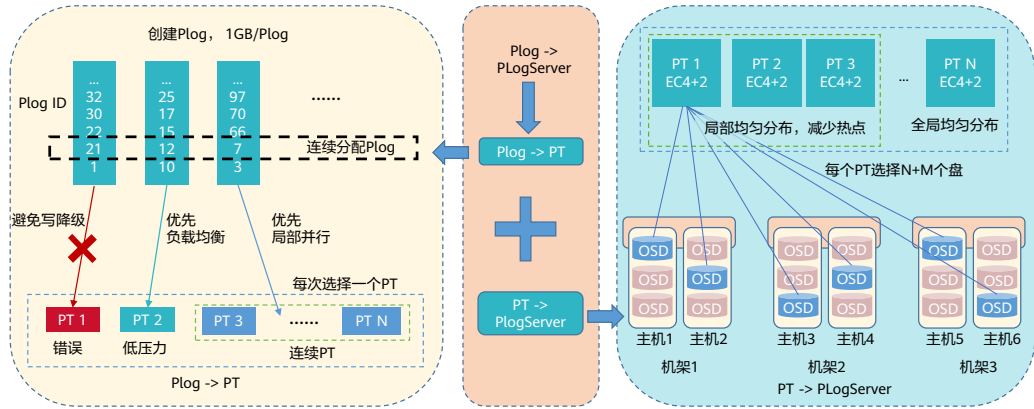
磁盘和新介质在不同IO Pattern下均体现出巨大的性能差异：

随机写（8K）			顺序写（8K aggregate->Large size）		
硬盘类型	性能	GC写入	硬盘类型	性能	GC写入
HDD	150 IOPS / 1.2 MB/s	/	HDD	5120 IOPS / 40 MB/s	/
SSD	40K IOPS / 312MB/s	Bad	SSD	153K IOPS / 1200 MB/s	Good

Append Only Plog技术可面向介质提供最佳的下盘性能模型



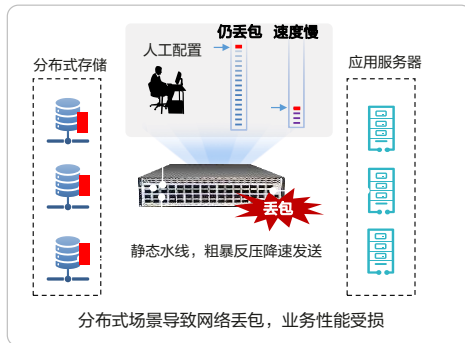
动态智能分区和静态选盘算法



网络加速

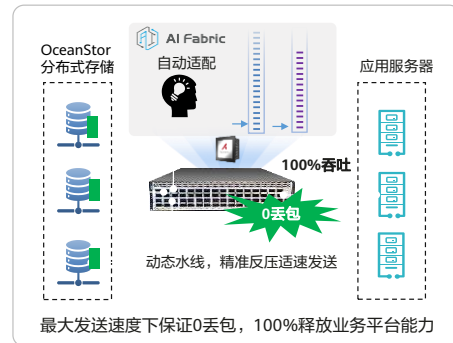
传统RoCE网络

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

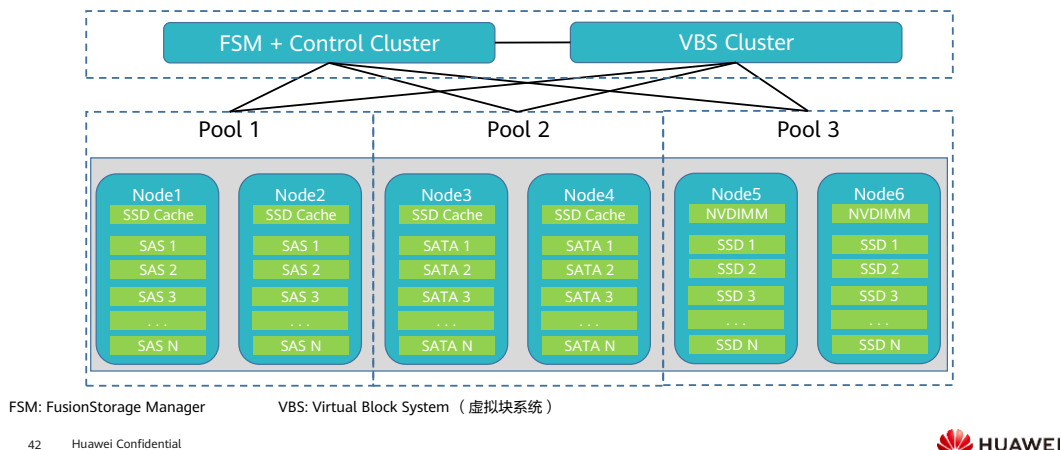
交换机内置智能芯片和算法，动态调整水线，确保100%吞吐下0丢包。



- 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。

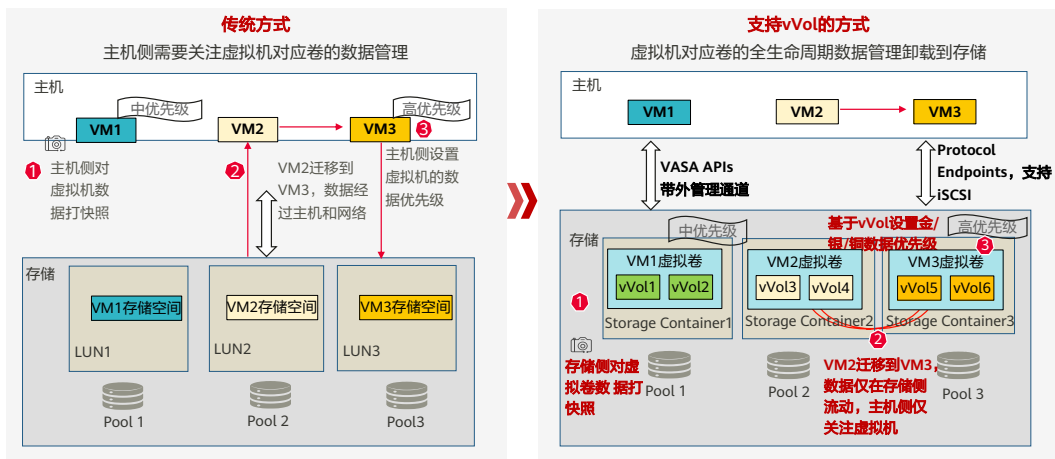
多资源池

- 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。



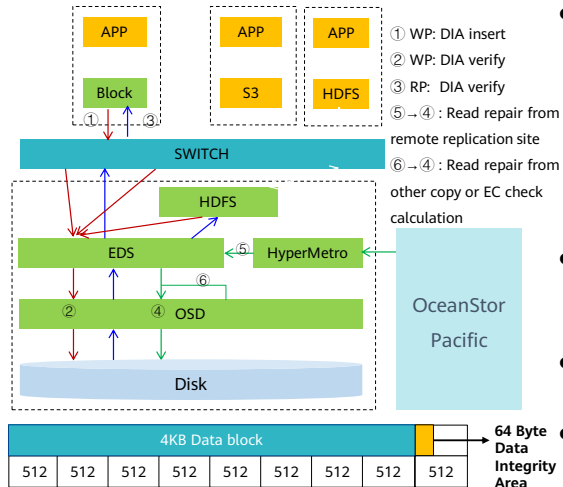
- 软件定义存储策略：支持以资源池为单位配置主存类型、缓存类型、副本冗余类型、安全等级策略（机柜级/服务器级）。
- 高可靠性：通过故障隔离，多MDC等机制，各资源池相对独立，故障互不影响。

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。

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。

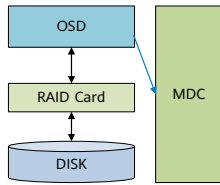
● 三种校验机制：CRC 32保护用户4K数据，除此之外，还支持主机LBA校验，盘LBA校验等逻辑校验，完美解决跳变，读偏，写偏等所有静默数据场景。

● 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。

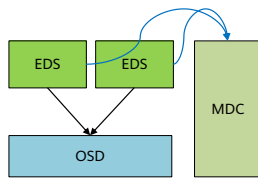
● Block/NAS/HDFS均支持端到端校验，Object暂不支持。

亚健康管理技术

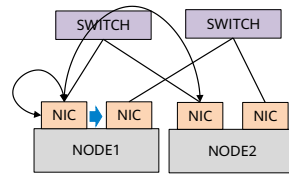
1、盘亚健康管理



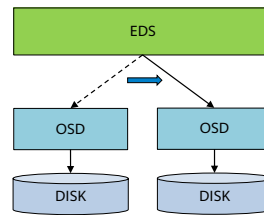
3、进程/服务亚健康管理



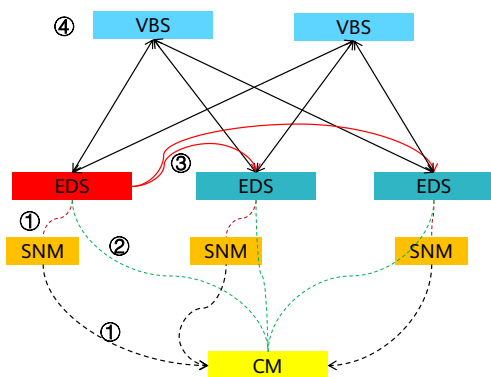
2、网络亚健康管理



4、Fast-Fail (快速换路重试)



故障快速切换技术

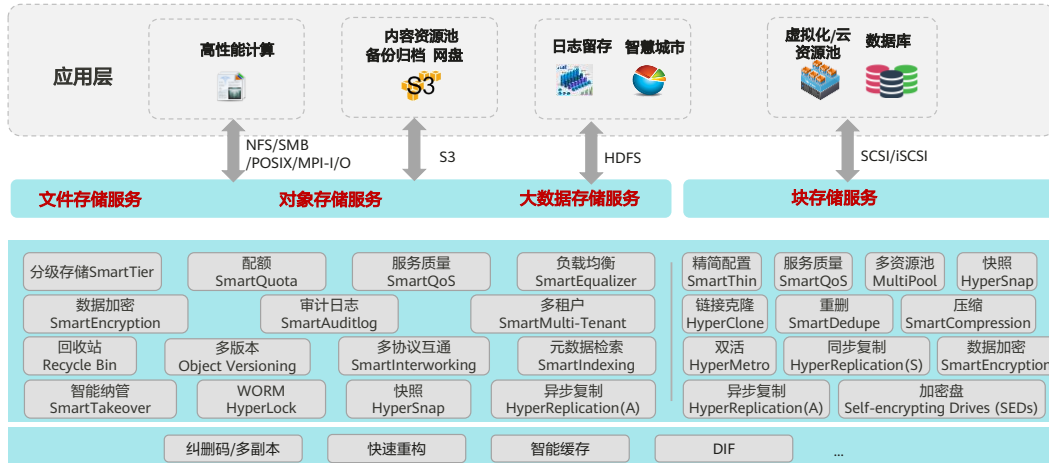


- **进程故障检测通知**：部署监控进程，可达到对进程故障的毫秒级监控。检测故障后立即拉起进程并通知集群主，由集群主触发业务本地的恢复。
- **心跳检测机制**：针对网络/节点等故障场景导致集群主和业务进程同时失联时，采用“心跳群发+关联诊断”算法在控制新主上快速判定业务进程失联/故障。
- **快速倒换接管**：将需要接管的业务并发倒换到剩余的多节点上，一方面缩短接管时长，一方面使接管后性能更均衡。同时倒换流程使用延迟加载、并行加载、加载数据量缩减、索引快速重建等技术来加速接管过程。
- **主机IO重定向**：VBS层毫秒级检测发往某节点/进程的IO是否返回，如无返回则主动拉取该节点/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

目录

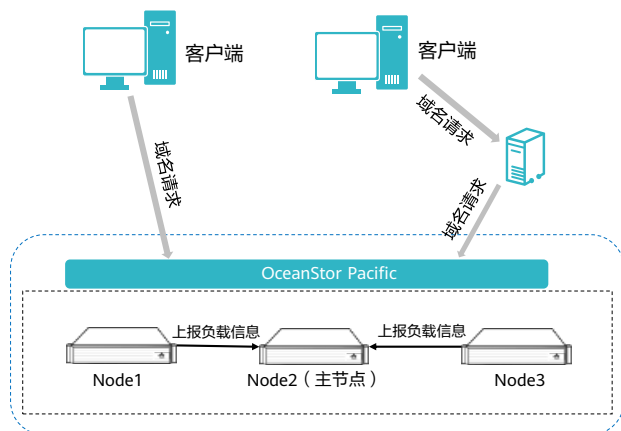
1. 产品介绍
2. 关键技术
- 3. 关键特性**
 - **效率提升**
 - 数据保护
4. 应用场景

关键特性



- OceanStor Pacific系列提供块、对象、大数据及文件四种存储协议，通过一套架构满足多样性数据的存储要求；同时提供了丰富企业级特性，对于文件、对象和大数据存储服务，我们开创性地发布了多协议互通特性，一份数据支持多存储存储协议访问，同时，所有企业级特性均同时适用于文件、对象和大数据存储。

负载均衡特性



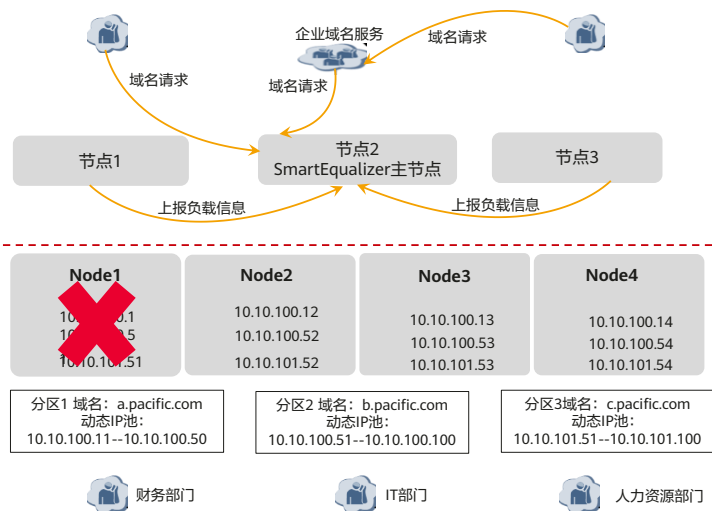
- SmartEqualizer为华为OceanStor Pacific系列分布式存储服务提供了负载均衡特性，SmartEqualizer周期性的采集存储系统各个节点的CPU利用率、网络吞吐量等信息，提供一种可配置负载策略的负载均衡服务。

- 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（非结构化服务支持）



• 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。

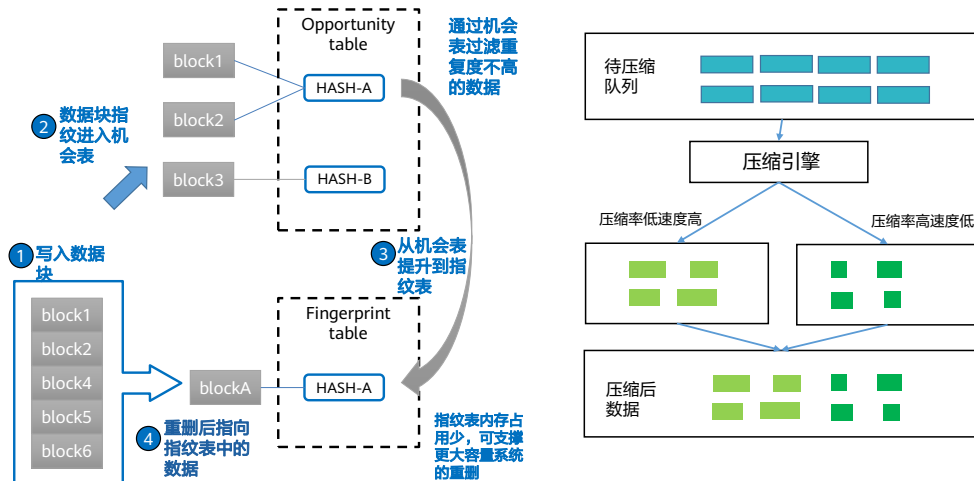
• IP资源池化

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

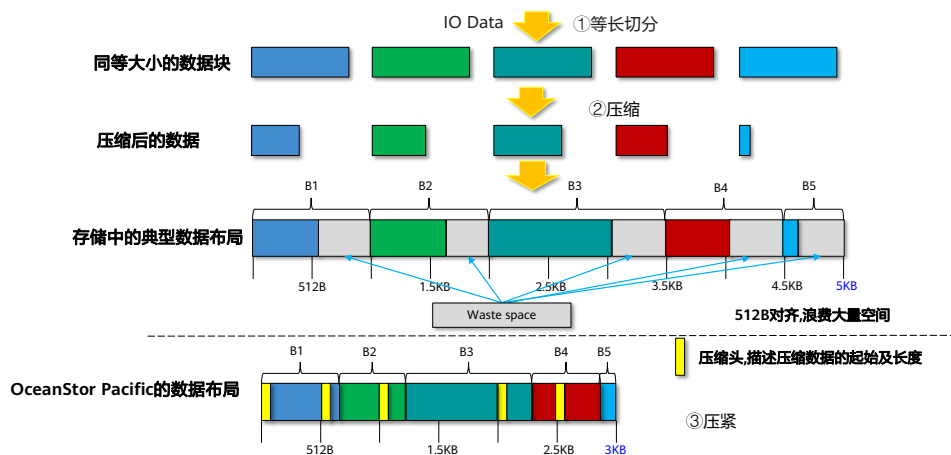
分区内IP池动态分配到每个节点, 节点故障后, IP可以自动被多个节点分别接管, 保障可靠性及负载均衡; 也可以配置为IP地址不自动漂移的静态IP资源池模式。

自适应全局重删压缩



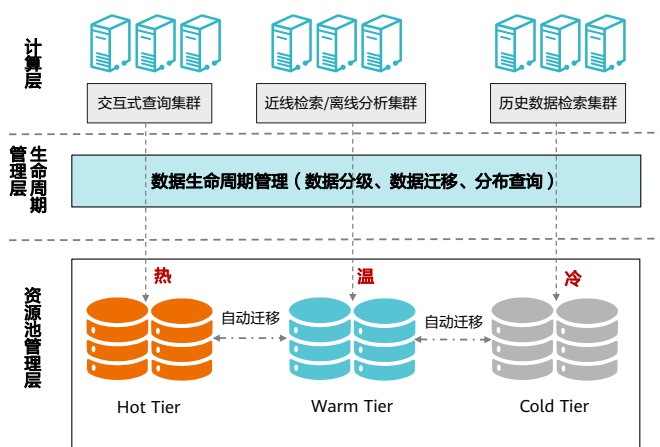
- OceanStor Pacific系列结构化服务采用了智能的自适应重删技术，以用户需求为导向，在用户数据处理负载较高的情况下，前重删会自动关闭，优先确保性能，由后处理完成数据缩减。在负载较低的情况下自动开启前重删，避免了后处理的读写放大。自适应技术以用户为导向，在用户不感知的情况下根据负载自动切换重删方式，避免了在线重删和后重删两种方案的缺点。
- 为了获取较好的重删压缩效果，OceanStor Pacific系列结构化服务采用了全局重删。同时分布式存储空间巨大，为了减少指纹表的内存空间消耗，引入指纹机会表的机制（如图）。数据块的指纹先进入机会表计数，只有相同数据块被多次写入达到阈值（默认3，可配置），方可进入指纹表执行重删。
- 如果压缩未开启，则直接申请该数据块长度存储空间保存数据。开启了压缩的情况下，则会进行压缩。数据块将被压缩引擎压缩，然后以最小512字节的粒度进行保存。
- OceanStor Pacific系列结构化服务压缩引擎采用两种不同压缩算法组合运行，一种是压缩率低速度高的算法，一种是压缩率高但速度低的算法。通过配置两种压缩算法不同的执行比例，可以得到不同的性能和数据缩减率。在同一个存储池只能选择一个压缩算法。存储池压缩算法的修改不会影响已经压缩的数据，已经压缩的数据在读取时会根据压缩时的算法进行解压。

压缩流程



- 压缩支持以LUN/FS为粒度按需打开关闭。
- 支持两种压缩算法：LZ4 与 HZ9可选. HZ9是深度压缩算法，压缩率比LZ4高20%.
- 压缩后的数据通常长短不一，可以将多个小于512字节的数据块合并压紧，按512对齐存储到存储空间，压紧后仍然非512对齐则补0后存储，进一步节省空间。

分级存储



按应用分级

不同目录数据可写入不同的资源池，一套存储池支持多种大数据应用

数据自动迁移

用户自定义迁移策略，实现冷热数据的自动迁移

统一命名空间

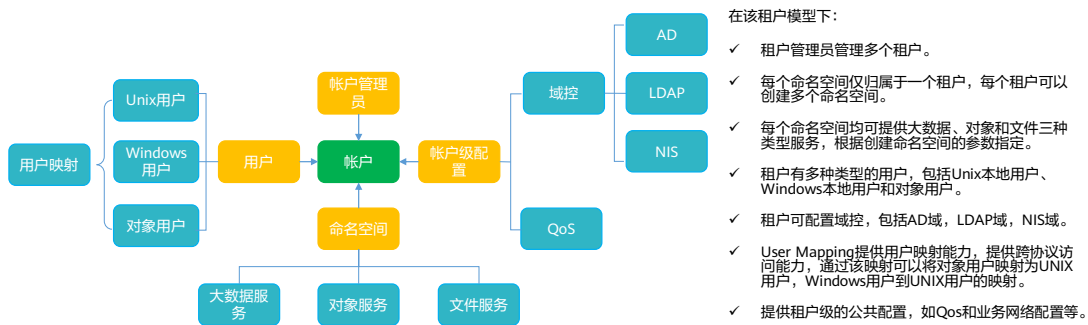
数据迁移计算侧无感知，可根据数据实际位置直接访问

- OceanStor Pacific系列非结构化服务提供的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 workflows 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。

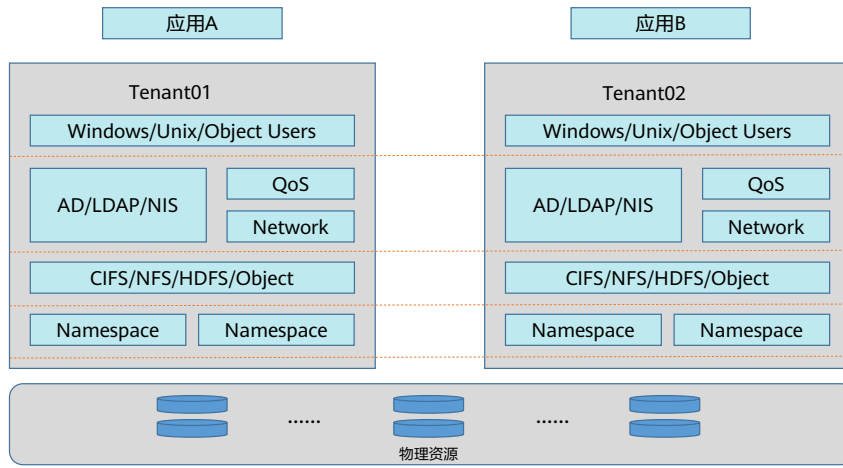
- SmartTier的分级策略分为：放置策略、迁移策略和删除策略。
- 放置策略：
 - 用于确定文件数据初始放置位置，该类策略条件包括：FS/DTree、文件名、UID/GID。如果未匹配到任何策略，根据系统默认的放置策略决定放置位置。
- 迁移策略：
 - 周期性迁移策略：用于确定文件数据搬迁的目标位置。该类策略条件包括：FS/DTree、文件名、文件大小、创建时间、修改时间、访问时间、状态修改时间、UID/GID。系统自动按照策略将数据搬迁到指定等级的硬盘池中。数据搬迁的时间可配置，避开业务的高峰期，默认在每日凌晨零时执行。
 - 一次性迁移策略：用于确定文件数据搬迁的目标位置。该类策略条件包括：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。

多租户

- 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命名空间管理、用户管理、QoS管理等。



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。

SmartIndexing自定义元数据检索



多协议统一检索

对象、文件、大数据存储统一元数据



更新主动上报

元数据更新主动上报，更智能

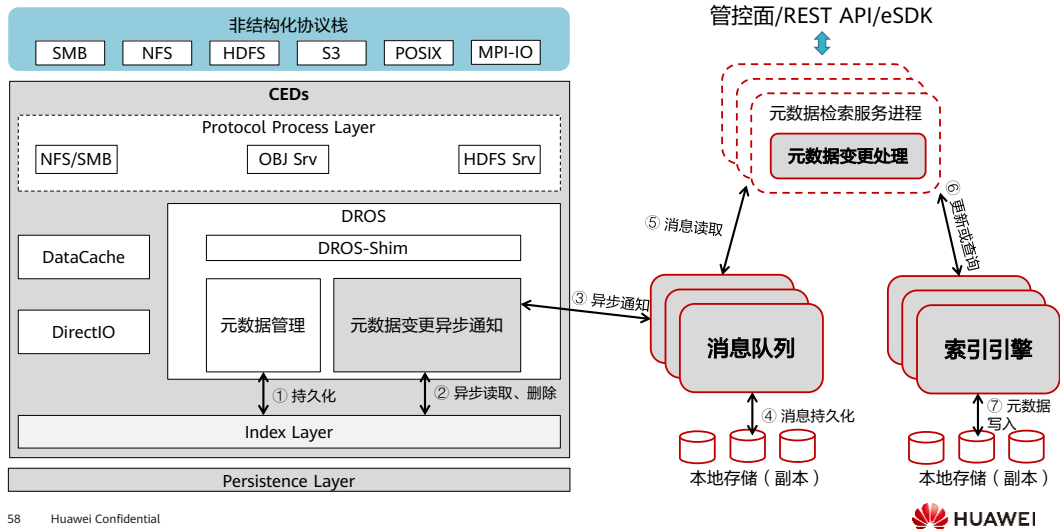


秒级返回结果

自定义检索千亿元数据时

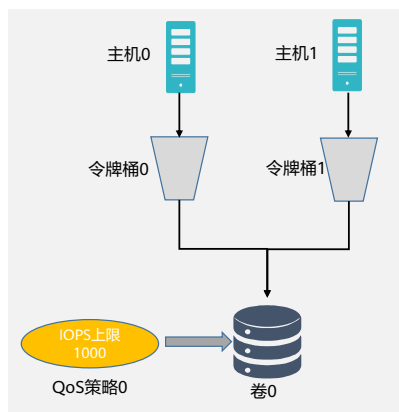
- 海量数据场景，用户需要基于元数据才能对海量数据进行管理，例如获取后缀名为 jpeg 的所有图片文件列表、获取文件大小大于 10GBytes 的文件列表、获取给定日期之前创建的文件列表，在用户能够快速获取符合条件的文件或是对象列表之后才能高效管理对应的数据。
- OceanStor Pacific 系列提供了基于非结构化（文件、对象和大数据）存储池的元数据检索功能 SmartIndexing，支持以命名空间(或文件系统)为粒度开启或是关闭元数据检索功能。对于开启了元数据检索的存储系统，在前台业务 I/O 变更文件、对象或是目录的元数据之后，异步、主动推送变更的元数据到检索系统。有别于传统存储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。

SmartIndexing

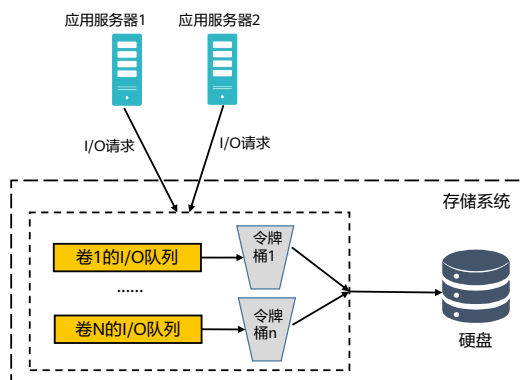


- OceanStor Pacific 采用统一的存储底座进行用户数据及元数据的管理，在该底座上支持各种数据访问协议，从而实现对NFS/CIFS/HDFS/S3/POSIX协议写入数据的元数据检索，不同协议下发的操作都由同样的模块进行处理，统一了非结构化数据的元数据，从而轻松支持多协议（NFS/CIFS/HDFS/S3/POSIX）的元数据检索。
- 持久化（persistence）：一种数据存储过程，该过程将数据写入磁盘文件等存储介质，以供内存读取。
- OceanStor Pacific 海量存储引入快速事务，将元数据变更日志和变更之后的元数据持久化（上图步骤①）。事务的引入确保了元数据的一致性，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。

SmartQoS（结构化场景）



基于负反馈的自适应算法



基于卷的IO流控管理算法

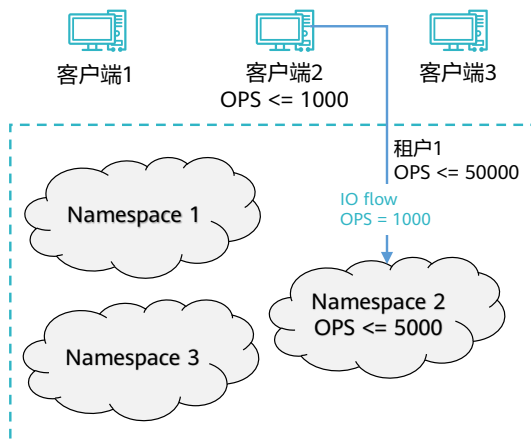
- 基于负反馈的自适应调整算法

- OceanStor Pacific系列产品中存在一个卷挂载到多个VBS节点的场景，在此场景下为了限制该卷上的业务对系统资源的消耗。需要限制该卷在系统中的总性能，需要一个分布式流控参数协调的过程。
- 以上图举例，系统为初始状态，卷0的IOPS上限为1000，QoS框架会采集主机0和主机1对卷0的压力情况，自适应调整卷0在各个节点令牌桶的令牌速度，以达成整体IOPS被限制到1000的目标。

- 基于卷的IO流控管理算法

- QoS流控管理通过对卷的IO队列管理、令牌分发和出队控制三部分实现。当用户为某个QoS策略设置了性能上限目标，那么通过上文讲述的分布式流控参数协调，会确定每个VBS节点上的流控参数目标，进而这个性能上限会被转化成对应的令牌。在存储系统中，如果用户要限制的流量类型是IOPS，那么一个IO会消耗一个令牌；如果设定的性能目标是带宽，那么一个字节对应一个令牌。基于卷的IO队列管理通过令牌机制实现存储资源的分配，某个卷的IO队列所拥有的令牌数越多，系统分配给这个卷的IO资源也越多。
- 如图所示，应用程序的IO进入存储系统后，首先进入各自的IO队列。QoS框架周期性地处理队列中的IO等待，将队列首元素从队列中摘除，尝试到令牌桶中获取令牌。如果令牌桶内剩余令牌满足队首元素的令牌请求，则将该元素交由存储系统其它模块继续处理，并且QoS框架继续处理下一个队首元素；如果桶内剩余令牌不能满足队首元素的令牌请求，将其重新挂入队列首部，QoS框架结束本轮IO出队动作。

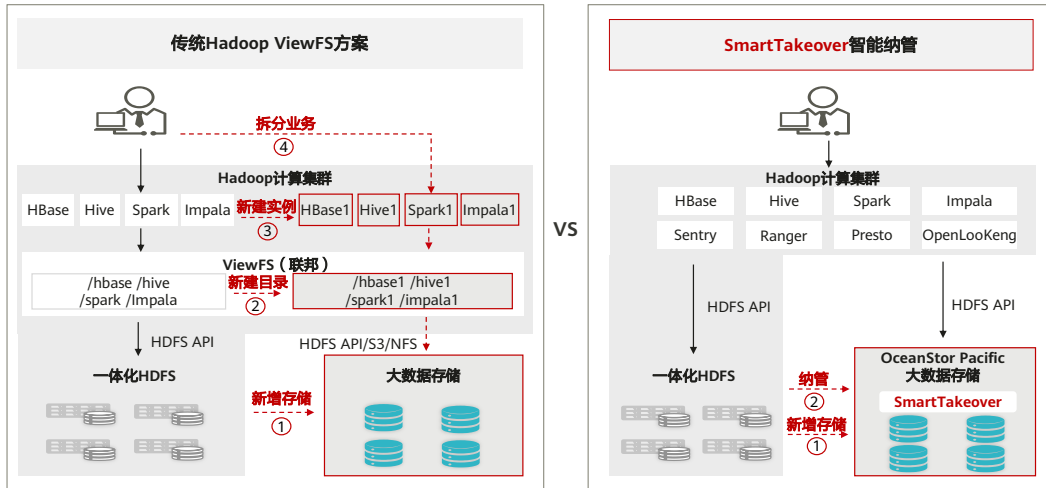
SmartQoS（非结构化场景）



- 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，最严的约束决定了业务的实际流量。
- 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。

SmartTakeover

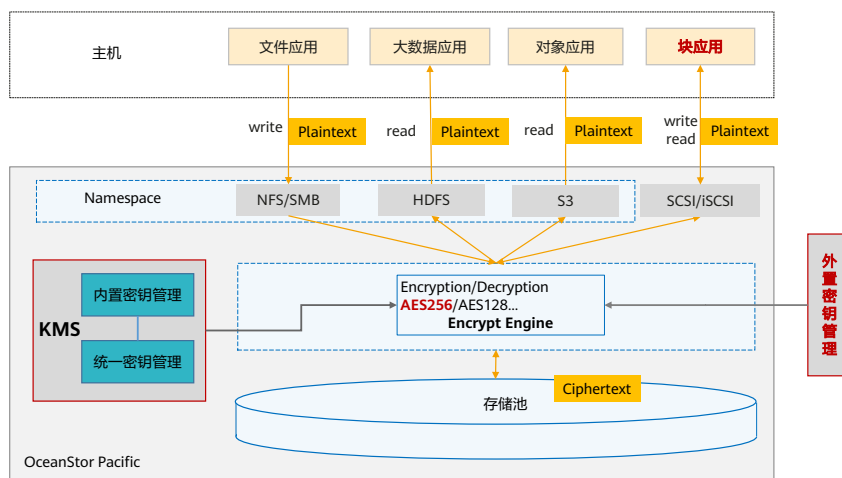


- OceanStor Pacific系列大数据服务提供了SmartTakeover元数据网关方案，对外提供全局Namespace，实现在不迁移数据的前提下对其它HDFS Namespace进行纳管。在用户现有的大数据平台下可以实现无缝接入，对上层业务无影响。
- 修改配置：
 - Hadoop集群中，将HBase/Hive/Spark/Presto/openLooKeng等组件的数据访问路径更改到OceanStor Pacific系列大数据服务的NameNode。
 - 给元数据网关增加原HDFS的NameNode路径。

SmartEncryption (数据加密)

- 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，通过两层秘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秘钥管理安全性。
- 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。

软件加密(Data at rest)



- 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。

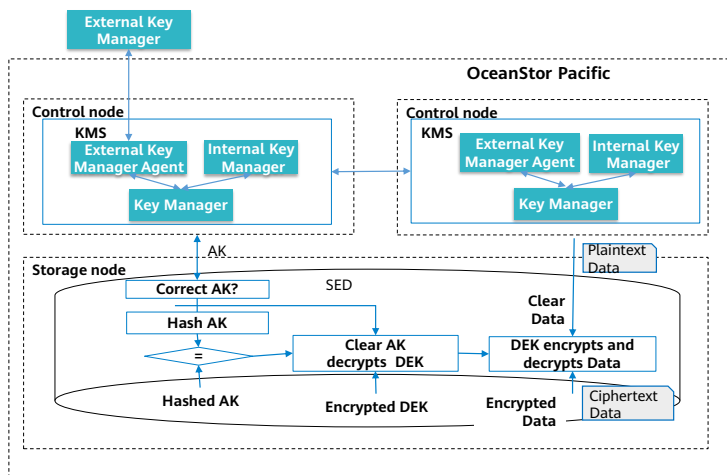
- 密钥管理服务（KMS：Key Manager Service）

- 支持内置密钥管理和外置密钥管理。
- 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，返回给应用。

加密盘(SED Data at rest)

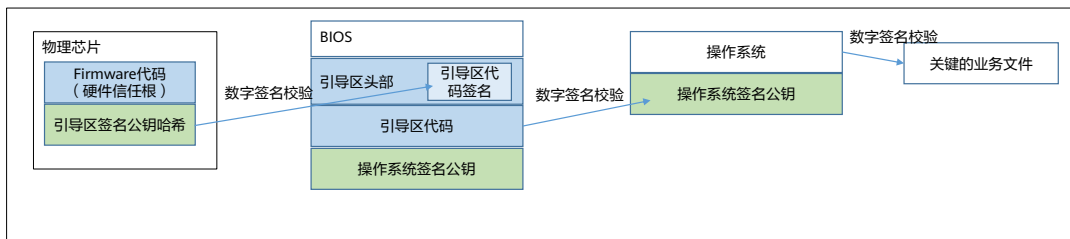


- 加密盘 (SED: Self Encryption Disk) :
 - SED具备两层安全保护, 分别使用 AK(Authentication Key) 和DEK(Data Encryption Key)两个安全密钥。AK: 身份认证密钥, 初始化硬盘时用于身份认证的密钥, AK由密钥管理服务 (Key Manager Service) 生成。DEK: 数据加密密钥, SED读写时用户加解密数据的密钥, 由加密盘生成。
 - 加密能力以存储池为力度按需启用。

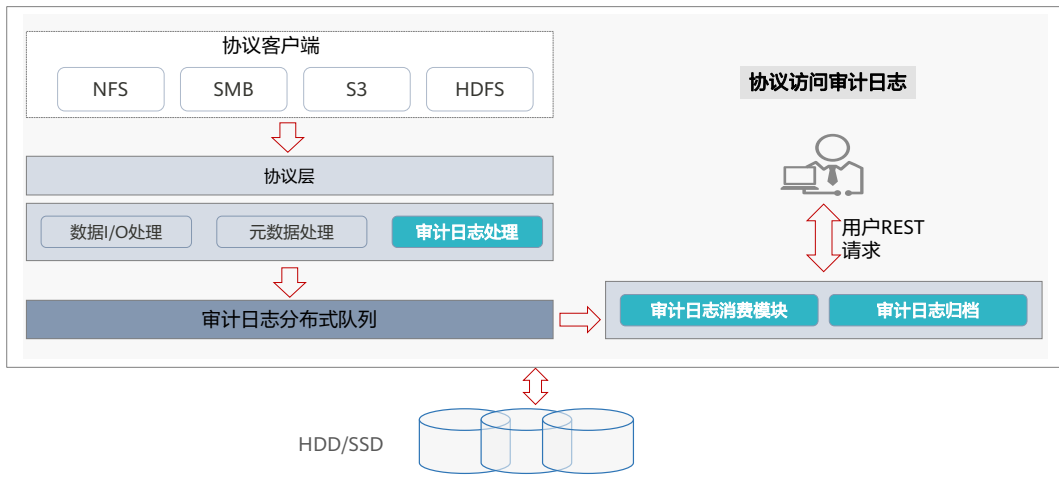
- 密钥管理服务 (KMS: Key Manager Service) :
 - AK密钥的获取, 支持内置密钥密管和外置密钥管理。
 - 外置密钥管理: AK由外置密钥管理服务器产生, 遵循标准的KMIP协议, 当前支持KMIP1.0、KMIP1.1、KMIP1.2三个协议版本。
 - 内置密钥管理: AK密钥由安全随机数产生, 采用三层体系结构保护密钥安全, 支持密钥的备份、恢复、更新、销毁。
- AK认证原理:
 - 当在存储系统上首次启用数据加密特性时, 存储会用KMS生成AK并计算HASH值, 并替换硬盘内的缺省AK的HASH, 且保存到硬盘中, 并对盘内置的DEK进行加密且保存在盘内。
 - 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。

安全启动

- 安全启动（Secure Boot）也叫验证启动（Verified Boot），是利用公私钥对启动模块进行签名和验证。在启动过程中，前一个模块验证后一个模块的数字签名。如果验证通过，则运行后一个模块；如果验证不通过，则暂停启动。安全启动检测设备启动阶段固件（Firmware）以及软件是否被篡改，以防止恶意软件侵入和修改。通过安全启动可以保证存储系统启动过程中各模块的完整性，防止未经认证的模块被加载运行，防止对存储系统及用户数据产生安全威胁。



SmartAuditlog



- 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。

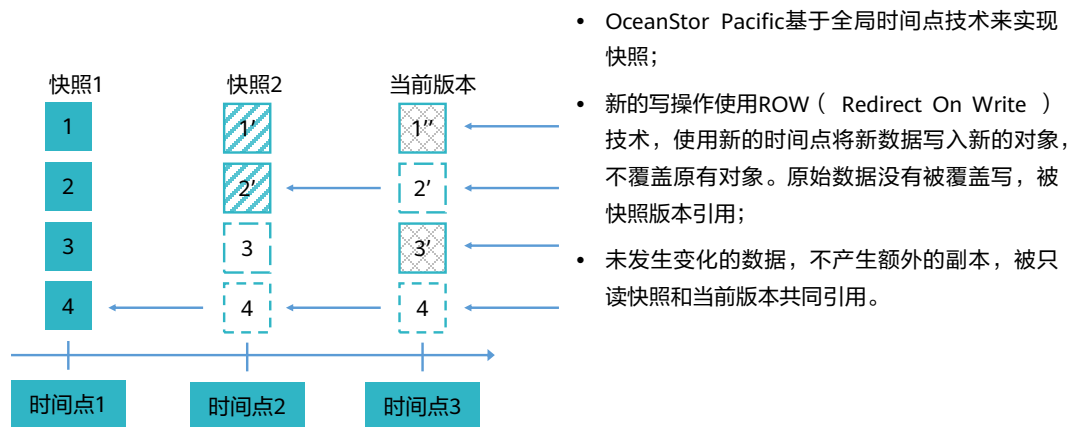
目录

1. 产品介绍
2. 关键技术
- 3. 关键特性**
 - 效率提升
 - **数据保护**
4. 应用场景

快照

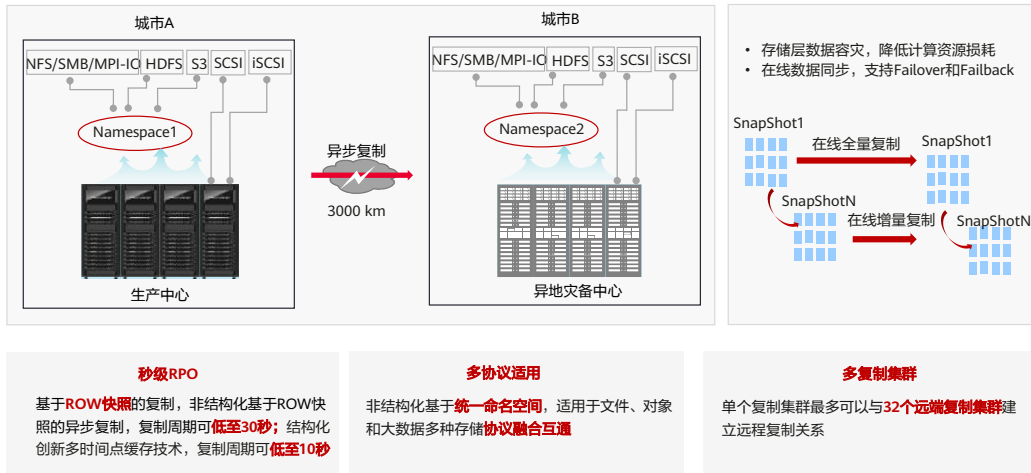
- 概述
 - OceanStor Pacific的快照是只读快照，数据不可更改。
 - 快照是指源命名空间或Dtree在某个时间点的镜像。
 - 快照生成后可以立即被读取。
- 主要特点
 - 瞬间生成：存储系统可以在秒级生成一个快照，获取源命名空间或Dtree数据的一致性副本。
 - 占用存储空间少：生成的快照数据并非完整的物理数据拷贝，不会占用大量存储空间。
 - 性能无损：基于ROW机制，快照性能与源命名空间性能相当。

ROW



- 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。

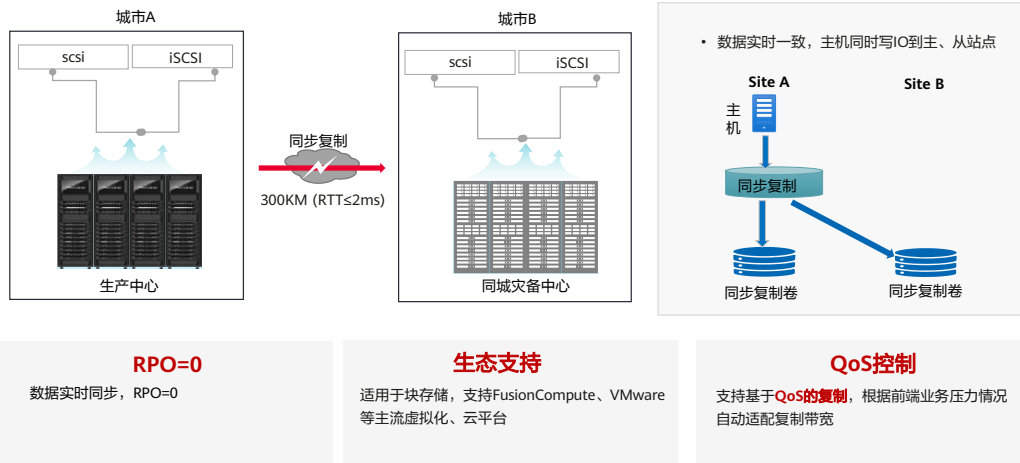
HyperReplication(A)



- 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。

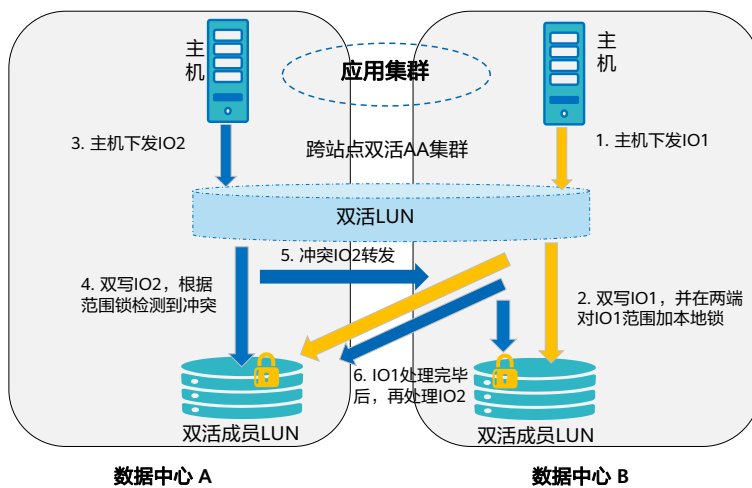
- 当灾难发生时，主站点失效，主从命名空间之间的远程复制链路异常断开。此时，需要管理员取消从资源保护，使从命名空间变为“读写”状态，以供备用主机访问。
 - 业务恢复：远程复制中的主站点失效后，从站点临时接管业务，待主站点故障恢复后，再将业务重新切换回主站点。当主站点故障恢复后，需要将主站点的数据同步到主站点，从而让主站点重新对外提供服务。

HyperReplication(S)



- 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”）。

HyperMetro

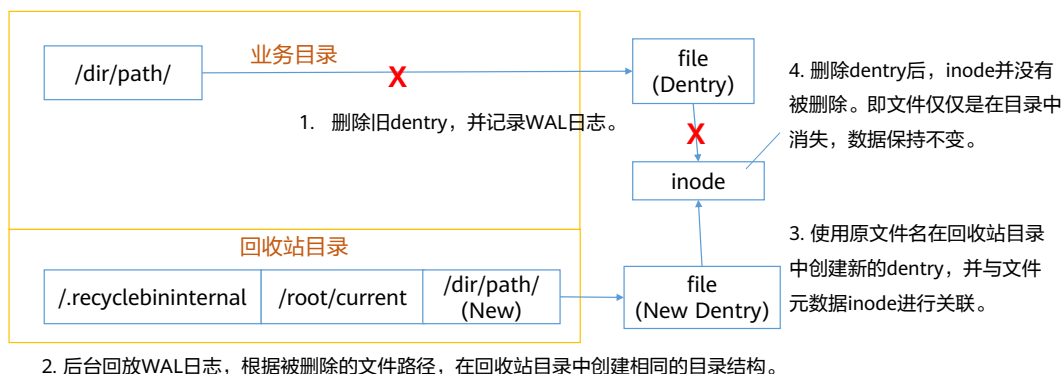


- 本端存储池和远端存储池向双活模块返回写I/O结果。对LOG的操作有两种情况：

- 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。

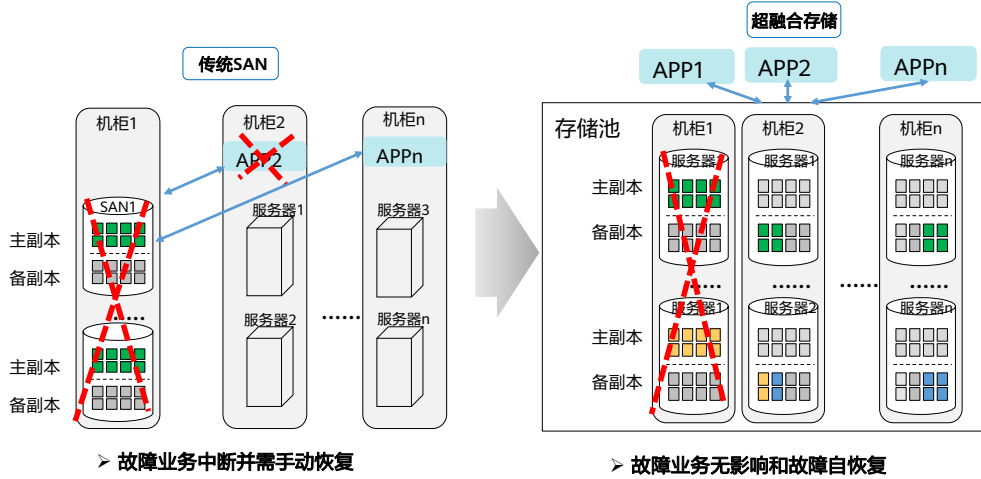
回收站



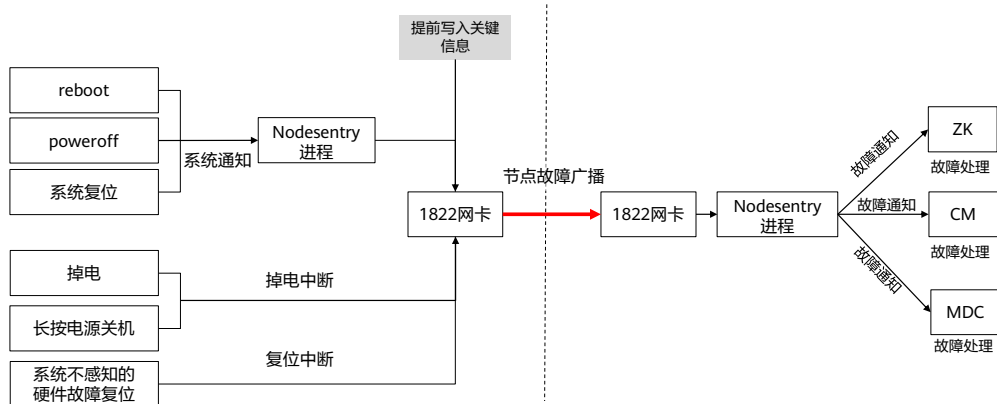
- 回收站的概念

- OceanStor Pacific回收站是命名空间/dtree根目录下面的一个特殊目录, 名为: `/.recyclebininternal`, 用于临时/长期存储用户删除的文件。
- 与windows的回收站不同, OceanStor Pacific回收站下面的文件不是平铺的, 而是按照文件删除时的目录结构存储的。在开启回收站后, 文件在删除时不删除数据, 而是修改元数据将文件隐藏起来, 并在后台创建回收站中的目录结构, 并再次修改文件元数据将其移到对应的目录中。

机柜级可靠性



系统不感知硬件故障复位



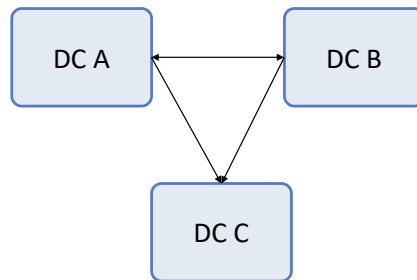
对象跨站点容灾技术

- 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/Passvie)两种模式

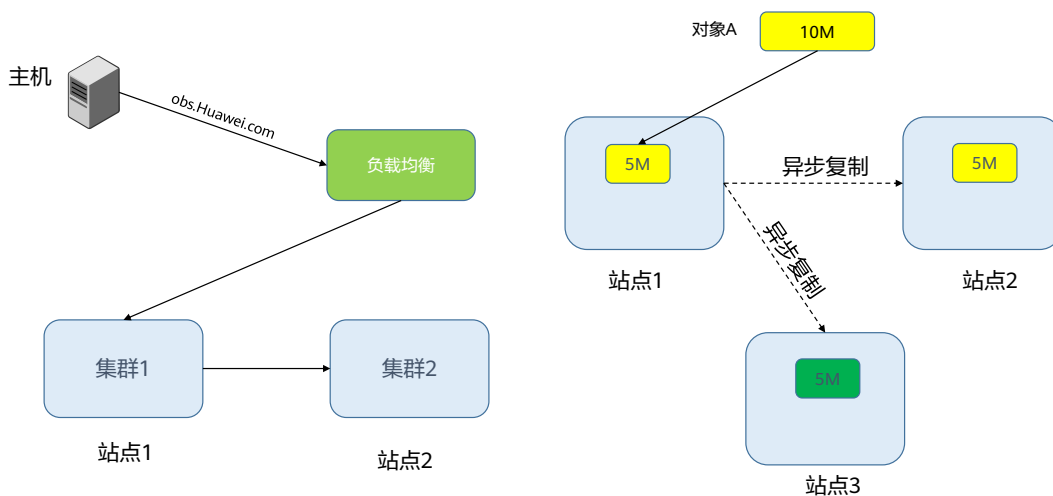
- 常见容灾模式：

- 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。

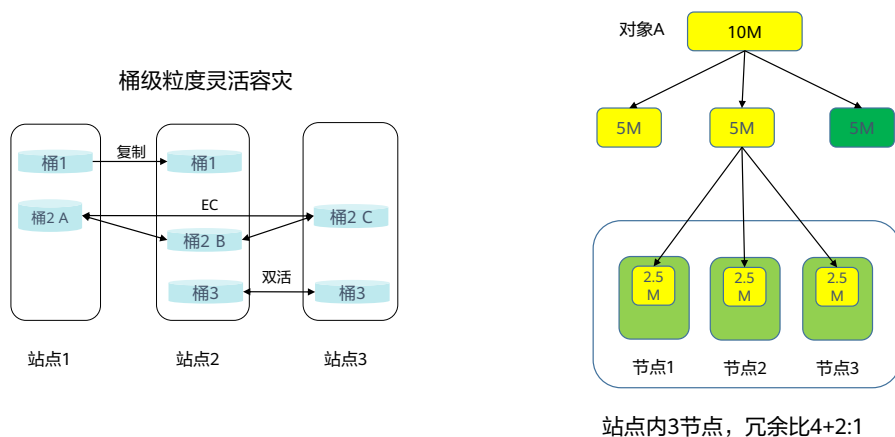
- 站点间副本冗余

- 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）



- 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。
 - 利用率: $0.67 * 0.67 = 0.45$ 。
 - 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。

目录

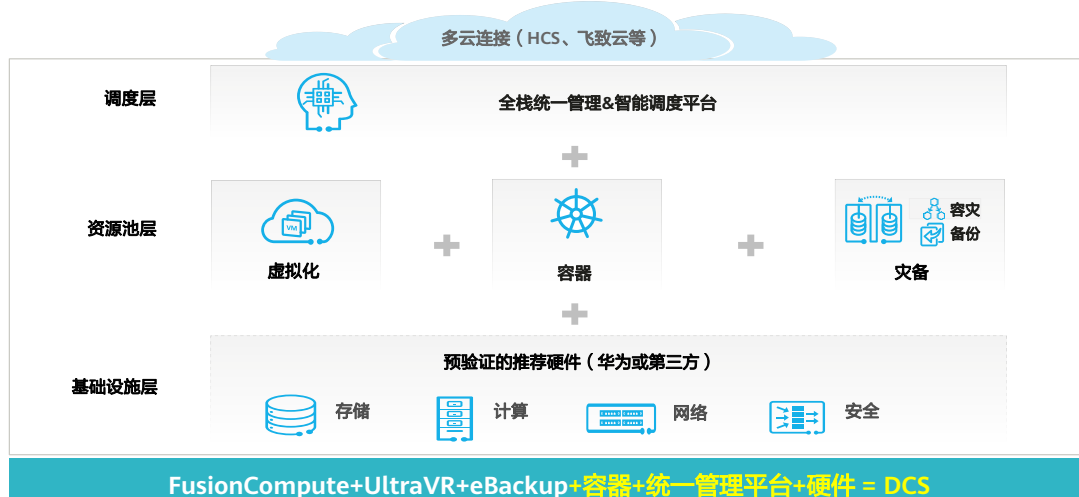
1. 产品介绍
2. 关键技术
3. 关键特性
- 4. 应用场景**

企业IT发展趋势



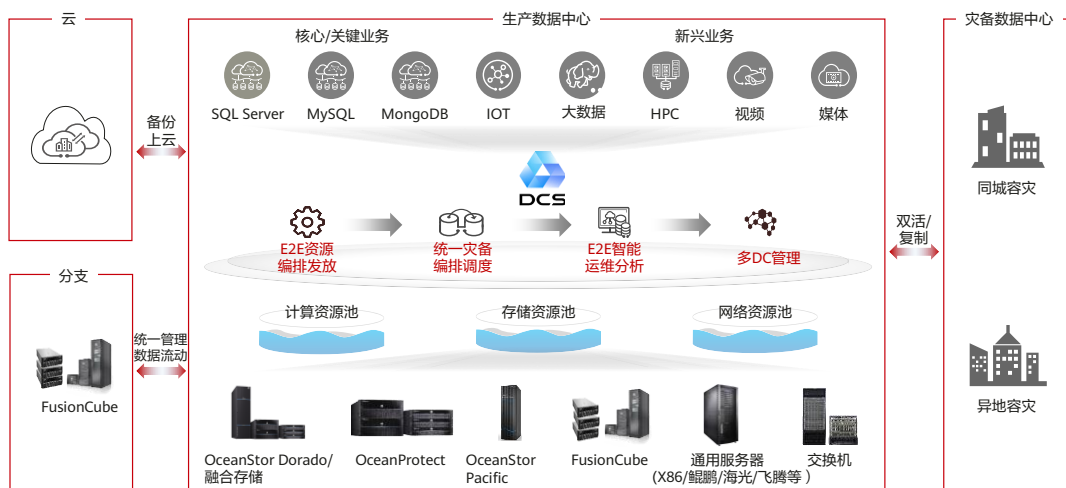
- 多样诉求推动数据中心基础设施迭代进化，助力企业IT数字化业务敏捷。

华为DCS数据中心虚拟化解决方案



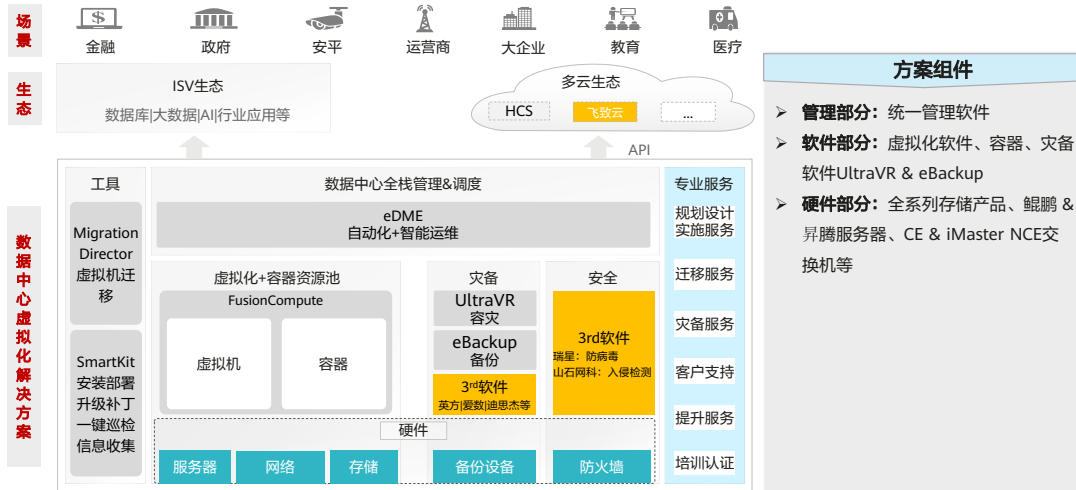
- 采用eDME作为数据中心全栈管理软件，通过统一的管理界面、开放的API、云上联动的AI使能、多维度智能风险预测与智能调优，实现“规划、建设、运维、优化”资源全生命周期自动化管理与智能运维，对多个数据中心提供统一管理，帮助客户简化管理，提升数据中心运维效率；
- 采用FusionCompute作为云操作系统软件，对各个物理数据中心资源进行整合，主要负责硬件资源的虚拟化，以及对虚拟资源、业务资源、用户资源的集中管理，从而降低业务的运行成本，保证系统的安全性和可靠性，协助运营商和企业构筑安全、绿色、节能的云数据中心能力；
- 采用eBackup&UltraVR虚拟化备份与容灾软件，实现虚拟机数据的备份和容灾，为数据中心提供全地域、全场景统一的灾备保护方案。

DCS定位



- 构建企业数据中心最佳基础设施，助力企业数字化转型。
- 通过虚拟化系列软件与ICT硬件组合，形成统一解决方案，为客户提供轻量弹性、多元开放的现代化基础设施。

DCS组件介绍



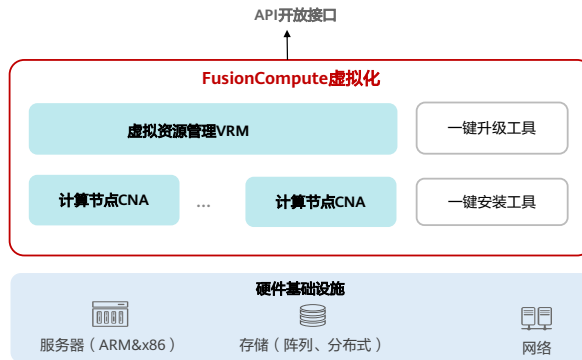
85 Huawei Confidential



• 关键组件介绍

- eDME：数据中心全栈管理软件，实现对数据中心各种资源池的接入、管理、E2E资源编排发放和E2E智能运维等功能。
- FusionCompute：虚拟化软件，实现对服务器、存储和网络的虚拟化，形成VM和容器弹性资源池，实现资源自动调度和管理。
- UltraVR：容灾软件，基于存储远程复制和存储双活，支持主备容灾，城域双活容灾，3DC容灾等场景，保障极端情况下的业务连续性。
- eBackup：备份软件，提供从虚拟机到磁盘阵列的全面备份能力，实现指定对象按指定策略备份，保障关键业务数据不丢失。
- MigrationDirector：虚拟机迁移软件，实现异构平台向DCS迁移能力。
- SmartKit：服务工具软件，实现软硬件安装部署、升级补丁、信息收集、巡检等。

FusionCompute



FusionCompute: 由VRM (Virtual Resource Management)和CNA (Computing Node Agent)组成, 实现对服务器、存储和网络的虚拟化, 形成弹性资源池, 实现资源自动调度和管理。

VRM: 虚拟资源管理节点, 提供统一的虚拟计算资源、虚拟存储资源、虚拟网络资源管理, 提供统一资源调度和业务发放, 提供统一的运维管理接口等。

CNA: 计算节点代理, 供虚拟计算能力, 管理计算节点虚拟机, 管理计算节点上计算、存储、网络资源。

- 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。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, 协助运营商和企业构筑安全、绿色、节能的云数据中心能力。

FusionCompute特点

极简

- **超轻量部署**: 2节点起步, 2小时交付
- **管理至简**: 分支站点可视化, VDC分权分域统一管理, 向导式不停机升级
- **敏捷开放**: 一键安装易集成, 标准Rest接口开放, 低门槛快速适配

轻量且易被集成
极低的成本提升数据中心资源利用率

极速

- **计算加速**: NUMA拓扑优化, 大页内存访问, 虚拟机性能提升30%
- **高速网络**: OVS+DPDK提升6倍网络转发效率, 支持40GE高带宽网卡
- **昇腾AI加速**: Atlas算力精准分配, 10倍性价比提升

多项性能调优技术
保障虚拟化高性能

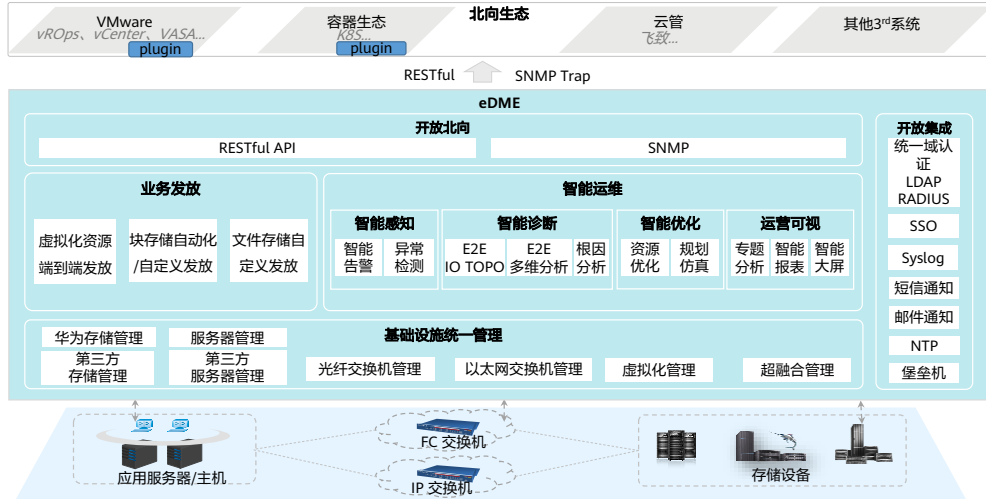
多元算力

- **技术创新**: 7年ARM和虚拟化深度开发, I/O性能提升30%
- **多元算力**: 支持x86、鲲鹏、飞腾、海光集群混合部署, 实现业务平滑过渡
- **一步上云**: 华为云全系统引擎, “0”成本接入混合云计算资源池

为政企数字化转型提供差异化、高性能、自主创新的算力支撑

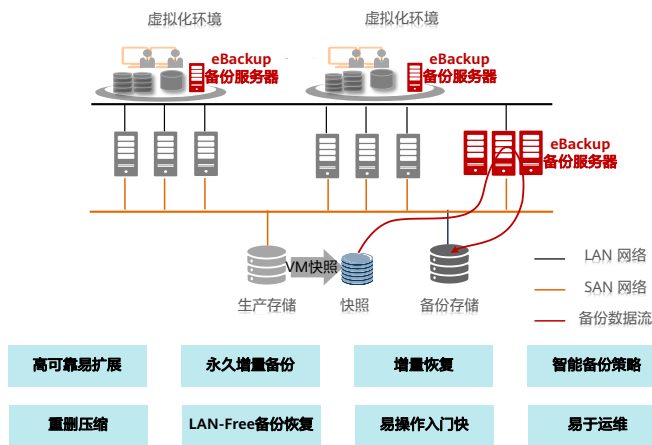
- 极简: 方案轻量, 快速部署, 业务快速上线
- 极速: 计算存储网络性能好, 满足不同业务诉求
- 多元算力: 支持ARM+x86双栈架构部署, 多元算力

eDME



- eDME定义为虚拟化场景下的软硬件统一管理软件，协同管理了数据库、数据工具、服务器、交换机、存储的控制、管理和分析。其核心是基于华为统一的智能运维管理平台，提供了基于告警的实时运维和问题闭环能力，基于统一数据集的丰富报表分析能力、基于策略和AI的自动化运维能力等。在南北向生态上，eDME通过Restful、SNMP、Telnet、Redfish等协议，可北向对接客户现有运维平台和各类云管平台，同时通过业界已开放的标准接口南向支持典型的第三方设备，这里包含服务器、交换机和存储设备。

eBackup



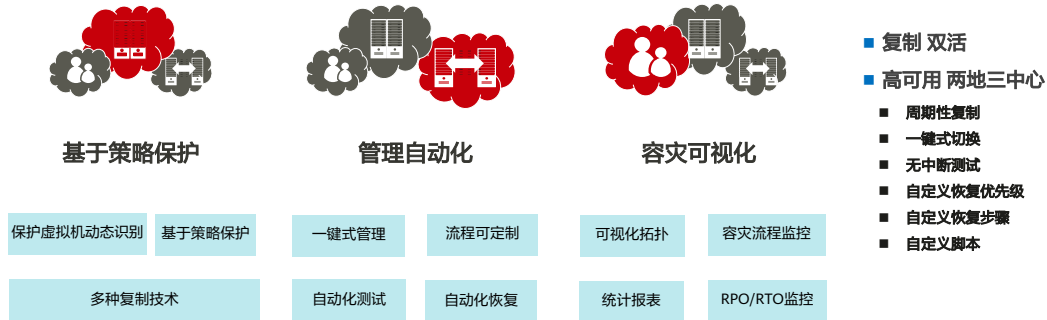
产品概述

eBackup是一款针对云化和虚拟化环境的备份软件，基于虚拟机快照和CBT（change block tracing）技术，对虚拟机数据提供全面的保护。它是一个满足海量虚拟机备份场景、简单易用、性价比极高的数据保护方案。

- eBackup虚拟化备份方案针对华为FusionCompute虚拟化平台，基于虚拟机/磁盘快照和CBT（Change Block Tracing）技术，对海量虚拟机数据提供全面的保护。eBackup虚拟化备份方案包括以下4个组成部分：
 - **备份对象：**即受保护对象，eBackup虚拟化备份方案可以对华为FusionCompute环境中的虚拟机提供数据保护。
 - **eBackup备份软件：**对虚拟化环境中的数据进行备份和恢复管理，eBackup备份软件支持模板部署和软件包部署两种方式。在一套eBackup备份系统中不能同时使用模板和软件包两种部署方式，可通过增加备份代理实现性能平滑扩展。
 - **备份存储：**支持NAS（Network Attached Storage）存储，S3（Simple Storage Service）存储和SAN（Storage Area Network）存储。
 - **备份策略：**支持为虚拟机、单个或多个磁盘创建保护策略。支持永远增量备份，减少备份数据容量。

UltraVR

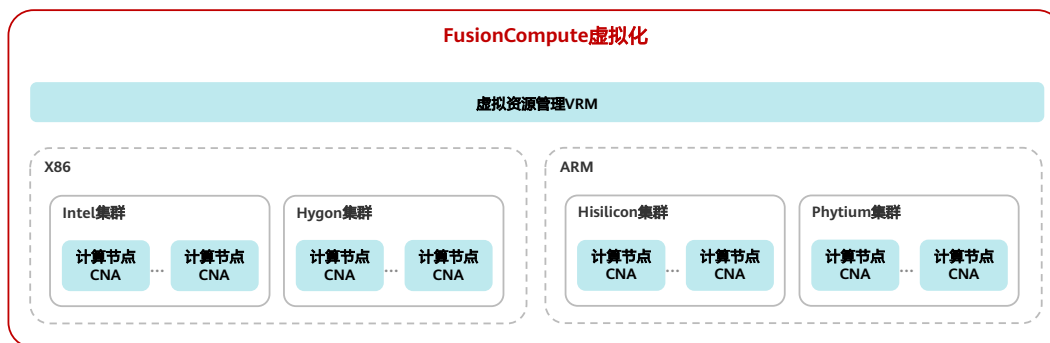
- **UltraVR**定位为企业数据中心容灾业务管理软件，简单高效地完成容灾业务配置，清晰可视地掌控系统容灾业务的运行情况，快速方便地完成数据恢复和测试。



- 产品特点:

- **简单高效:** 流程化引导容灾业务配置。一键式容灾测试、计划性迁移、故障恢复和重保护。
- **可视化:** 通过查看全局容灾物理拓扑、业务保护逻辑拓扑，掌控系统全局。保护组和恢复计划的执行状态一目了然。
- **集成化:** 与存储资源管理集成，满足多种应用场景（主备、两地三中心、双活数据中心）容灾运维需求，减少运维系统的重复投入，提升运维效率。
- **高可靠:** 多站点部署保障容灾业务管理调度系统的可靠性。管理数据自动备份保证快速恢复管理系统。

虚拟化资源池



- 虚拟化资源池是基于物理基础设施构建的虚拟计算、虚拟存储和虚拟网络资源池。
- FusionCompute提供对虚拟计算、虚拟存储和虚拟网络的资源池化和管理能力。

虚拟化灾备



城域双活容灾

- 在相距较远的两地建立两个站点，站点间使用华为存储的HyperMetro特性，配合FusionCompute的虚拟机HA功能和DRS功能实现的容灾方案。



阵列复制容灾

- 在两地建立两个站点，分别为生产站点和灾备站点，利用存储设备的远程复制功能，将虚拟机数据从生产站点复制到灾备站点，再使用UltraVR容灾管理软件，在灾备站点将灾备存储上的虚拟机注册到虚拟化平台，并自动启动。



两地三中心容灾

- 通过建立同城灾备中心和异地灾备中心，实现对生产中心数据及业务的多重保护。



华为分布式块存储复制容灾

- 在两地建立两个站点，分别为生产站点和灾备站点，利用分布式存储设备的远程复制功能，将虚拟机数据从生产站点复制到灾备站点，再使用UltraVR容灾管理软件，在灾备站点将灾备存储上的虚拟机注册到虚拟化平台，并自动启动。



虚拟机备份

- 使用华为eBackup备份软件，配合FusionCompute快照功能和CBT (Changed Block Tracking) 备份功能实现的虚拟机数据备份方案。

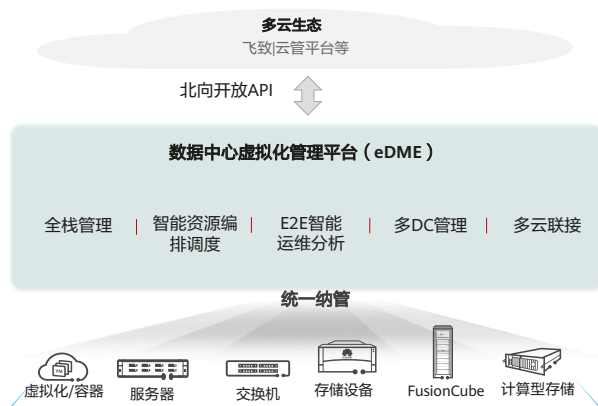
- 基于存储远程复制和双活能力提供主备、双活、3DC容灾管理保障业务连续性，为数据中心提供全地域、全场景统一的灾备保护方案。

虚拟化安全



- 提供数据存储、网络传输、管理运维、主机系统、虚拟机全栈的安全防护，保障业务安全访问。

虚拟化管理平台



融合

- 数据中心全栈管理
- 数据全生命周期管理
- 多DC统一管理

智能

- E2E智能分析
- 业务最佳安置
- 智能风险预测

开放

- 北向API开放，对接多云
- 多厂商设备管理
- 异构资源池纳管

- 提供从硬件基础设施到虚拟化资源的站点资源全栈运维管理，支持不同地域下的多个数据中心资源站点的统一管理，为系统管理员（包含硬件设备管理员）提供一个简单易用、软硬件协同的统一智能运维管理平台。

DCS关键特性全景图



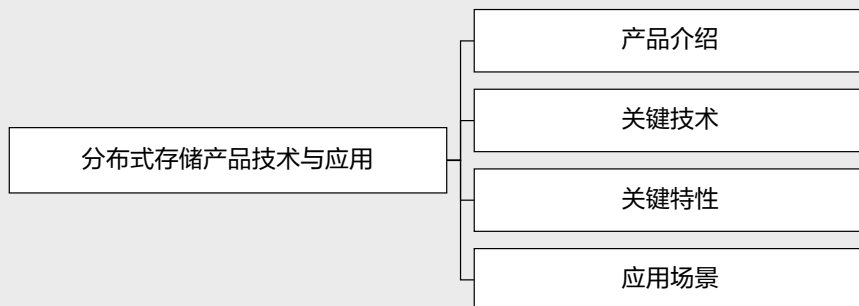
思考题

1. （多选题）华为分布式存储支持的存储服务四合一是哪4个？
 - A. 块服务
 - B. 文件服务
 - C. 对象服务
 - D. HDFS服务
2. （多选题）以下哪些选项是描述华为DCS数据中心虚拟化解决方案的特点？
 - A. 轻量弹性
 - B. 敏捷应用
 - C. 智能管理
 - D. 多元生态

- 参考答案：

1. ABCD
2. ABCD

本章总结



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://enterprise.huawei.com/cn/>
- 技术支持: <https://support.huawei.com/enterprise/>
- 在线学习: <https://support.huawei.com/learning/Index!toTrainIndex>
- 华为Support案例库: <https://support.huawei.com/enterprise/servicecenter?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.

Copyright©2023 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.



闪存存储产品部署



前言

- 本章节主要讲述华为闪存存储产品的部署流程，如存储系统安装、系统初始化、业务部署等。

目标

- 学完本课程后，您将能够：
 - 掌握闪存存储产品安装流程
 - 掌握闪存存储业务配置流程

目录

1. 闪存存储产品安装

- 设备安装
- 存储系统初始化
- OceanProtect产品初始化

2. 闪存存储业务配置

安装流程



安装前准备

准备安装工具

安装辅助工具

安装仪表

软件工具

.....

检查安装环境

检查温度湿度

检查电源

检查机柜散热

.....

开箱验货

设备验收

拆卸设备包装

安装辅料验收

.....

安装工具



记号笔



十字螺丝刀
(M3 ~ M6)



一字螺丝刀
(M3 ~ M6)



斜口钳



防静电腕带



压线钳



裁纸刀



浮动螺母
安装条



防静电工
作服



防静电手
套

安装配件



设备安装模板



浮动螺母和M6螺钉



miniSAS HD线



RDMA线缆



设备前面板



可调节滑道



滚珠滑轨



拔纤器



多模光纤



网线



接地线



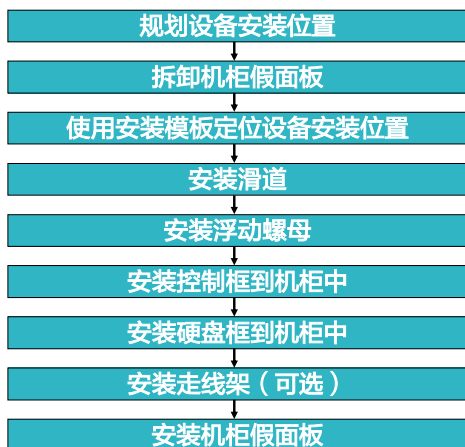
U型固定支架

软件工具

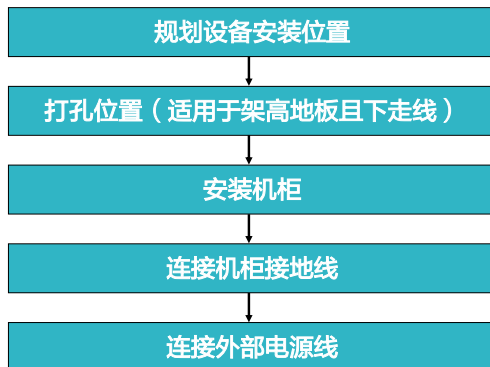
软件名称	用途	获取途径
SSH工具	登录存储系统、维护终端与存储系统之间的文件传输。	本文档以PuTTY软件为例进行说明。您可以访问chiark主页下载PuTTY软件。低版本的PuTTY软件可能导致登录存储系统失败，建议使用最新版本的PuTTY软件。
OceanStor SmartKit	信息收集、巡检、开局质量检查、扩容、升级等	工具下载请登录华为support网站搜索“OceanStor SmartKit”，下载对应版本的软件及文档。

设备安装流程

中低端设备安装流程



高端设备安装流程

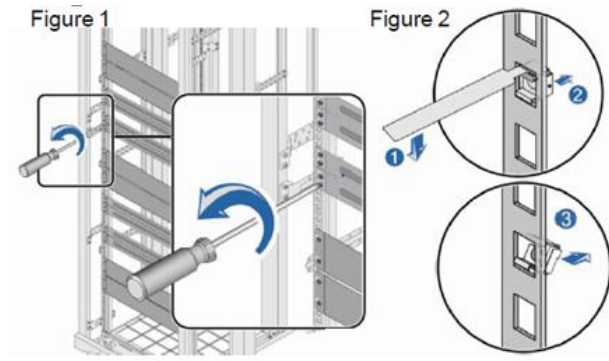


机柜空间规划

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。

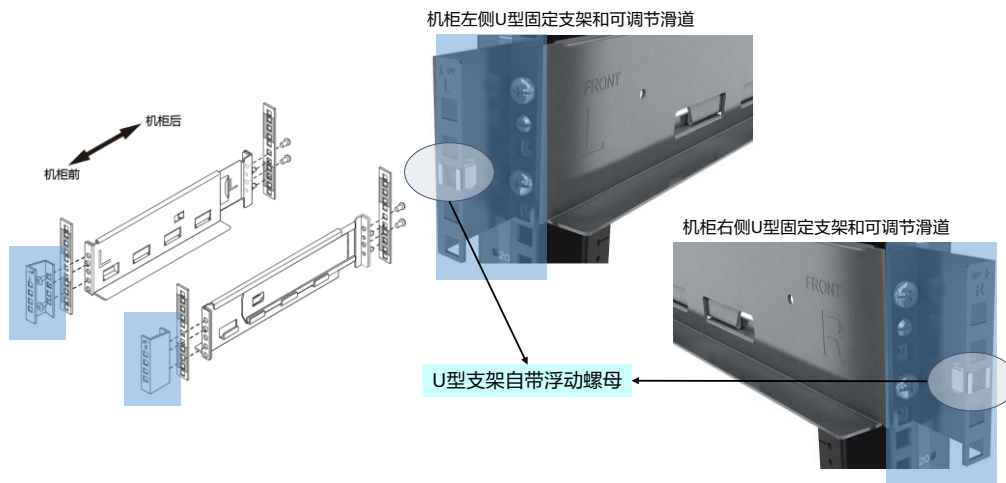
拆卸机柜假面板



- 所需工具:
 - 十字螺丝刀
 - 浮动螺母安装条（或一字螺丝刀）

- 注意:
 - 如果没有浮动螺母安装条，您也可以使用一字螺丝刀来拆卸浮动螺母。
 - 拆卸浮动螺母时，请勿用力过猛，以免浮动螺母安装条或一字螺丝刀伤及手部或损坏设备。
- 说明:
 - 用十字螺丝刀拆卸假面板上的M6螺钉，取下假面板。
 - 用浮动螺母安装条将浮动螺母的一只脚向下拉，然后将浮动螺母取出。

安装U型固定支架和可调节滑道



- 说明:

- 浮动螺母为用于固定设备的M6螺钉提供螺丝孔。
- 将浮动螺母的一只脚插入上方定位标识位置的方孔内，使其卡住方孔框
- 用浮动螺母安装条拉浮动螺母的另外一只脚，使其卡住方孔框。
- 检查浮动螺母的安装情况。

2U控制框线缆连接

低端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。

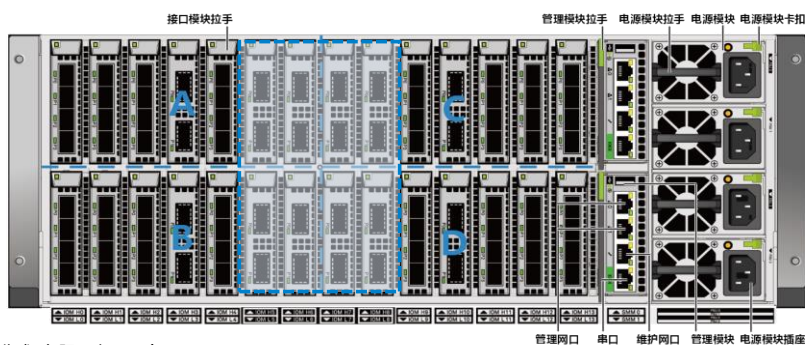
中端2U控制框



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。

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；

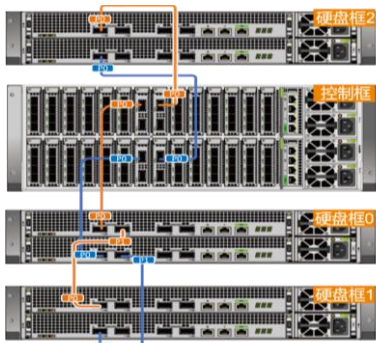
原则5： Scaleout卡配置在H3， L3， H10， L10；

硬盘框组网原则

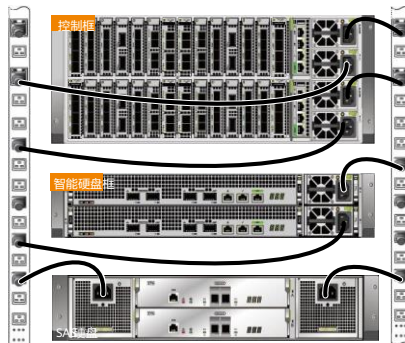
- 硬盘框级联组网采用全正接。
- 2U控制框：控制器A上的级联接口模块连接到硬盘框的级联模块A，控制器B上的级联接口模块连接到硬盘框的级联模块B。
- 4U控制框：框上部的级联接口模块连接到硬盘框的级联模块A，框下部的级联接口模块连接到硬盘框的级联模块B。
- 控制框上连接同一个硬盘框的两个端口所属接口模块要对称，端口号要一致。
- 控制框的级联接口模块端口或者板载级联端口只能连接到硬盘框的P0端口。
- 硬盘框的级联模块P1端口连接到下一级硬盘框级联模块的P0端口。
- 取消保险箱盘（配置框）必须是第1框前4盘的约束，对于硬盘框的安装顺序不再有约束。

- 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



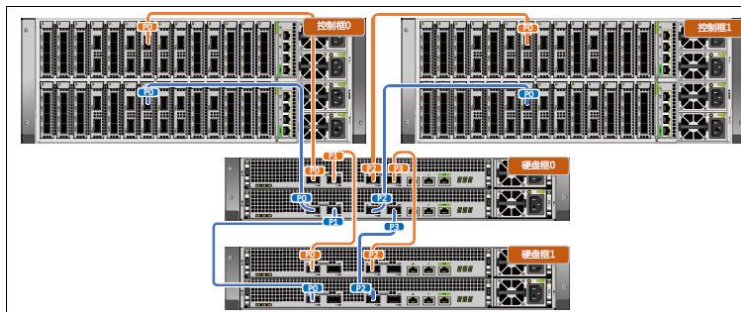
控制框和硬盘框连接电源线

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。

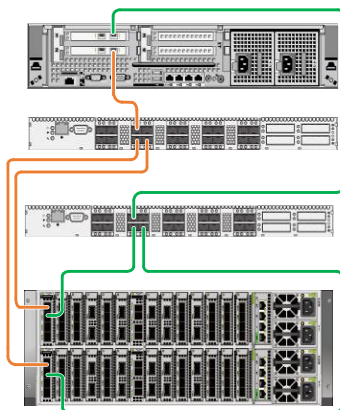
内双活组网

- **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。

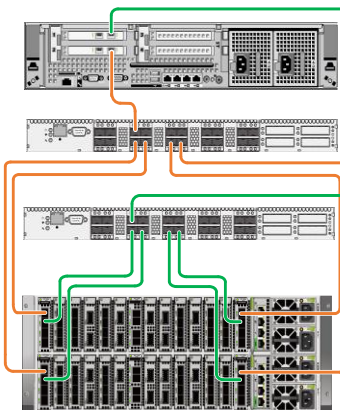


连接控制框至应用服务器

4U2控：FC/iSCSI组网



4U4控：FC/iSCSI组网



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



目录

1. 闪存存储产品安装

- 设备安装
- **闪存存储系统初始化**
- OceanProtect产品初始化

2. 闪存存储业务配置

初始化存储系统流程



修改管理网口IP地址

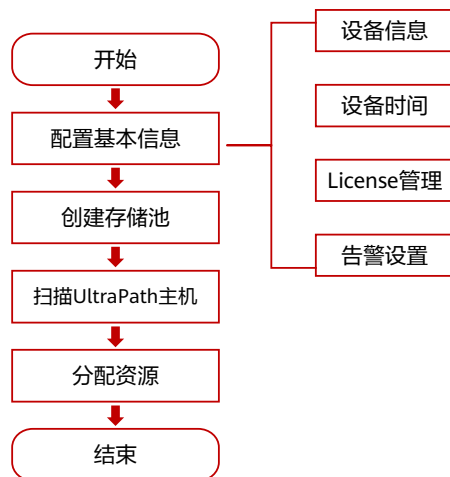
- 使用CLI修改管理网口IP地址：
 - 管理网口默认IP登录前提条件
网线连接维护终端和存储设备管理网口，维护终端的IP地址与存储管理网口的默认IP地址在同一网段。
 - 串口登录前提条件
串口线缆连接维护终端和存储设备串口，终端程序（例如PuTTY软件）的波特率需设置为115200。
- 登录后运行change system management_ip命令修改管理网口IP地址。如：设置管理模块0的管理网口IPv4地址为“172.16.190.2”，子网掩码为“255.255.255.0”，网关地址为“172.16.0.1”。

```
admin:/>change system management_ip eth_port_id=CTE0.A.MGMT ip_type=ipv4 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.
```

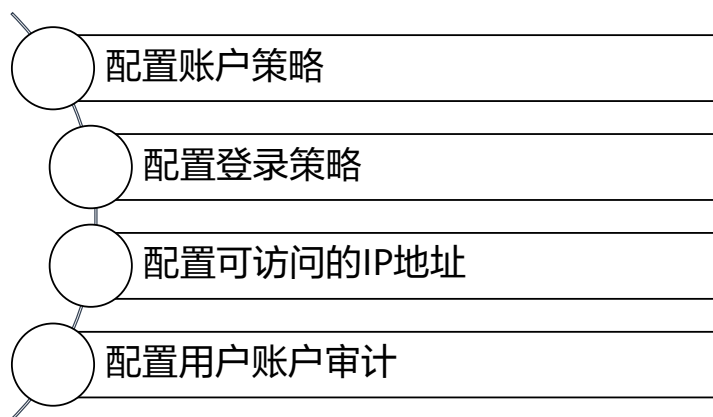
- 注意
 - 管理网口IP地址的默认值为192.168.128.101/192.168.128.102，子网掩码的默认值为255.255.0.0。
 - 管理网口IP地址与维护网口IP地址不能在同一个网段，否则会造成路由冲突。默认的维护网口IP地址为172.31.128.101或172.31.128.102，子网掩码为255.255.0.0，所以管理网口不能设置为172.31.XXX.XXX网段的IP地址。
 - 首次登陆，需要初始化默认密码。
 - 您也可以通过设备默认IP登录DeviceManager，通过DM进行IP的修改。

初始化配置向导



说明：以华为全闪存存储产品为例，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

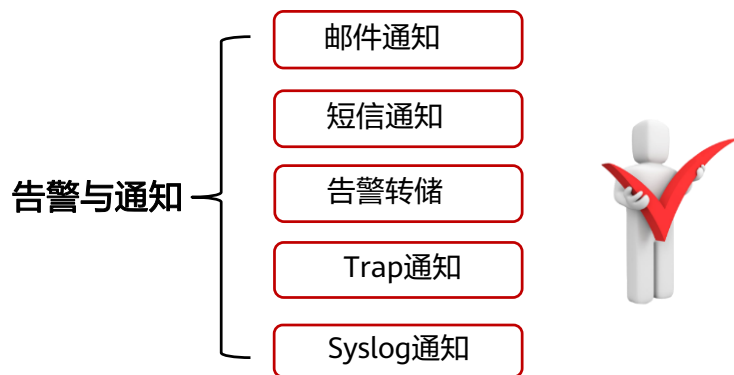
用户/角色管理

- **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



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“Info TV”中可以查找存储安装的视频；也可以选择OceanStor Dorado V6信息中心查找产品文档。



目录

1. 闪存存储产品安装

- 设备安装
- 闪存存储系统初始化
- **OceanProtect**产品初始化

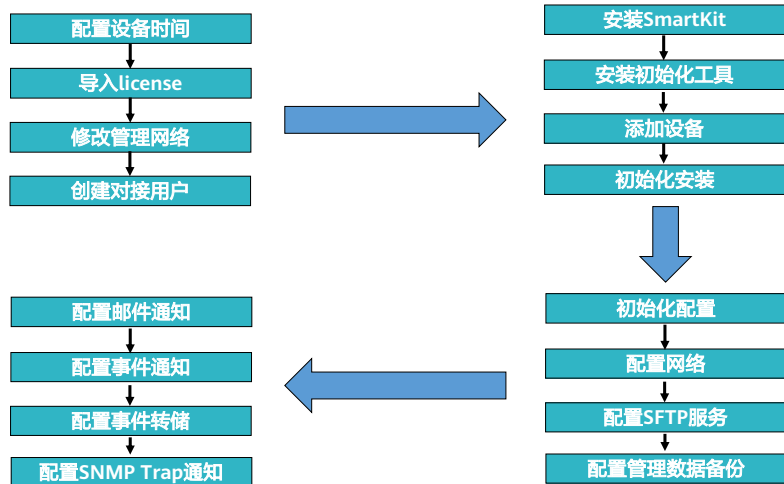
2. 闪存存储业务配置

安装前规划与准备

分类	子类	说明
局点信息	局点的应用信息、操作系统等信息	重点了解客户的应用信息、BOQ、操作系统，LLD等相关收集。
规划工具	能耗计算	提供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://support-it.huawei.com/powercalc/#/home
	OceanProtect兼容性查询	根据机房现场环境在线快速查询产品与业界主流厂商、主流设备的兼容性关系。 访问地址： https://support-it.huawei.com/comp/#/oceanprotect
	LLDesigner	提供硬件配置、设备组网等功能，能快速生成LLD文档和设备的配置文件。主要功能：支持BOQ一键式导入，自动识别设备，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。
安装辅助工具	螺丝刀、斜口钳、防静电的工作服、手套、腕带等	
软件工具	SSH工具	用于登录OceanProtect 专用备份存储和维护终端与存储系统之间的文件传输。
	SmartKit	用于扩容控制器、开局巡检等
安装资料	《产品描述》、《管理员指南》、《通用线缆操作手册》	访问地址： https://support.huawei.com/enterprise/zh/centralized-storage/oceanprotect-x8000-pid-251807257
备份软件	License	http://app.huawei.com/isdp/
	软件包	https://support.huawei.com/enterprise/zh/centralized-storage/oceanprotect-x8000-pid-251807257/software

- 在进行安装规划的时候，需要提前了解客户的应用信息及部署相关工具准备。

华为备份一体机初始化流程



配置设备时间

- 通过该操作，可以同步客户端时间到OceanProtect专用备份存储、设置NTP自动同步或手动修改设备时间，确保OceanProtect专用备份存储时间与客户端时间保持一致，以便在出现告警时能根据告警日志准确地定位出设备告警产生的时间。
- 前提条件
 - 设置NTP自动同步前，需要完成NTP服务器的配置。NTP服务器的配置方法，请参考NTP服务器的相关配置文档。
 - 在有防火墙的环境中使用NTP自动同步时，需要开启123端口。
 - 系统管理员已登录设备管理平台DeviceManager。
- 操作步骤：
 1. 在DeviceManager的菜单栏上，选择“设置 > 基本信息 > 设备时间”。
 2. 单击当前设备时间右侧的“修改”。
- 共三种方式设置设备时间：
 - 同步客户端时间
 - 手动修改
 - 设置NTP自动同步

导入License

- License文件是使用增值特性的权限凭证。在配置数据备份特性前，请先导入并激活数据备份特性的License。如果您需要使用数据加密功能，请同时导入并激活数据加密特性的License。
- 导入License后，你可以选择“设置 > License管理”，在中间信息展示区，确认“特性名称”列中已包含了“数据备份容量”。如果您导入了数据加密特性的License，“特性名称”列中包含了“数据加密”。

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 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

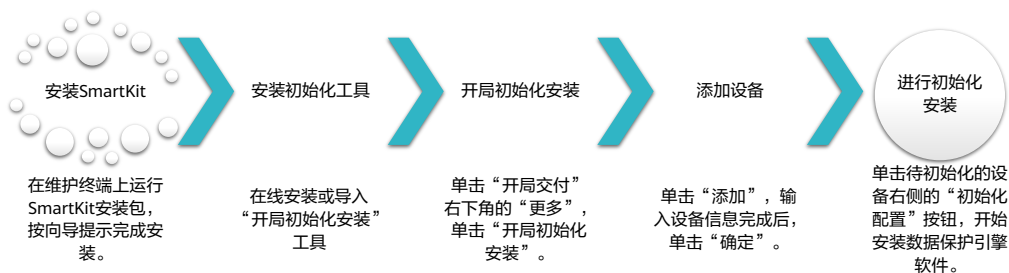
Are you sure you really want to perform the operation?(y/n)y
Command executed successfully.
```

- 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。
 - 您也可以通过设备默认IP登录DeviceManager，通过DM进行IP的修改。

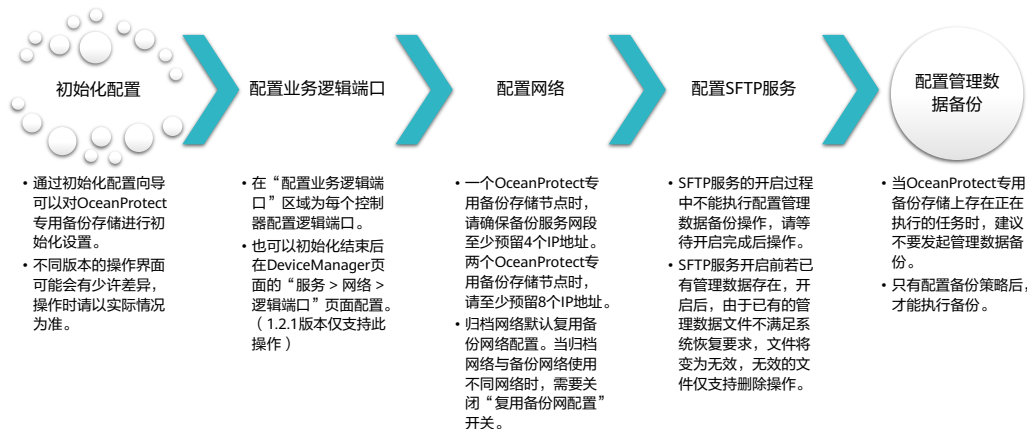
创建对接用户

- 在DeviceManager上，使用超级管理员用户创建一个对接用户，用于DeviceManager 与 ProtectManager对接。
- 操作步骤：
 1. 登录DeviceManager，浏览地址：<https://XXX.XXX.XXX.XXX:8088>，XXX.XXX.XXX.XXX表示控制器的管理IP地址。
 2. 选择“设置 > 用户与安全 > 用户和角色 > 用户管理”，系统进入“用户管理”界面。
 3. 单击“创建”，系统弹出“创建用户”对话框。
 4. 设置用户信息。
 - 类型：本地用户
 - 用户名/密码/确认密码：对接用户的用户名和密码
 - 角色：超级管理员
 - 密码永不过期：开启
 - 登录认证：登录密码
 5. 单击“确定”，创建用户完成。

数据保护引擎初始化安装向导



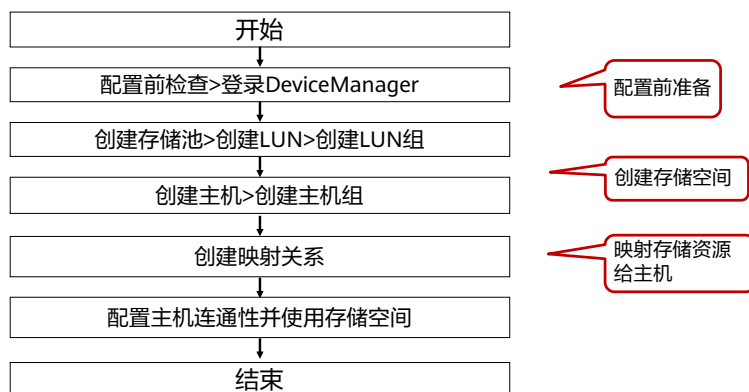
OceanProtect初始化配置



目录

1. 闪存存储产品安装
- 2. 闪存存储业务配置**

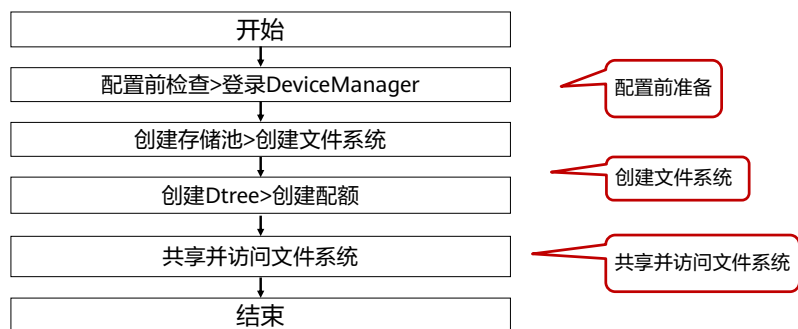
块服务业务配置流程



说明：以华为混合闪存存储产品为例，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：以华为混合闪存存储产品为例，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。

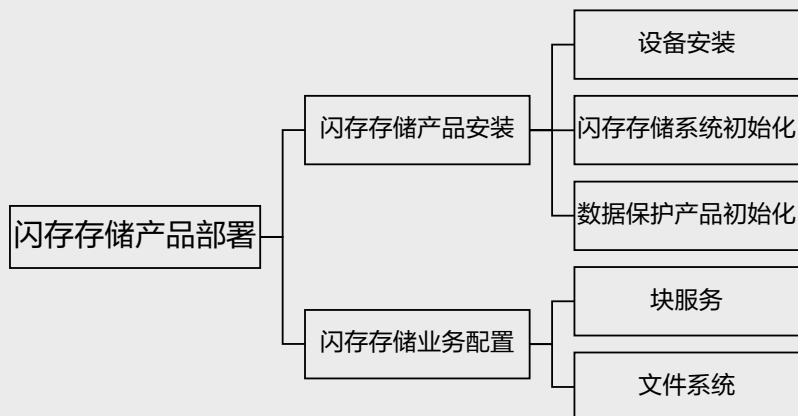
思考题

1. （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。
2. （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。
3. （多选）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：

1. F
2. T
3. ABCD

本章总结



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://enterprise.huawei.com/cn/>
- 技术支持：<https://support.huawei.com/enterprise/>
- 在线学习：<https://support.huawei.com/learning/Index!toTrainIndex>
- 华为Support案例库：<https://support.huawei.com/enterprise/servicecenter?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.

Copyright©2023 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.



分布式存储产品部署



前言

- 本章主要讲述华为分布式存储的安装部署流程，如存储系统安装、系统初始化、业务部署等。

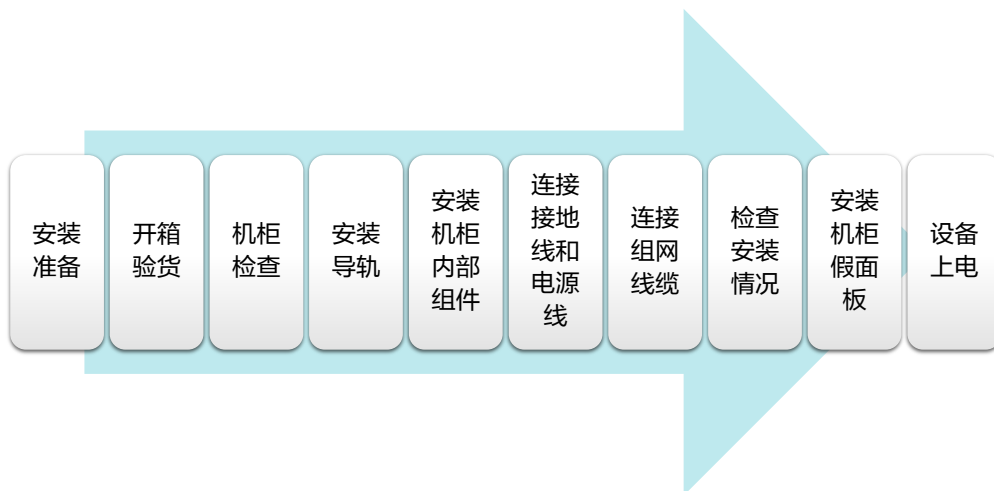
目标

- 学完本课程后，您将能够：
 - 掌握硬件安装流程
 - 掌握组网规划
 - 掌握软件的安装部署
 - 掌握分布式存储业务配置

目录

- 1. 硬件安装流程**
2. 组网规划
3. 软件安装流程
4. 业务配置

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。

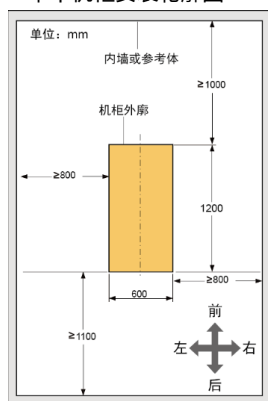
能耗计算工具：

华为数据存储信息中心>信息速查>规划设计>能耗计算

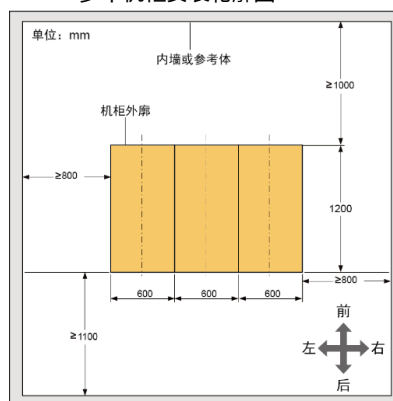
<https://info.support.huawei.com/storage/powercalc/#/home>

机柜空间规划

单个机柜安装轮廓图



多个机柜安装轮廓图



- 若设备安装在第三方机柜，请确保满足如下基本空间尺寸要求：
 - 机柜符合IEC 60297 482.6mm（19英寸）系列机械结构尺寸要求；
 - 2U节点和4U节点机柜深度 ≥ 1000 mm；
 - 2U节点和4U节点机柜滑道（托盘）承重能力 ≥ 50 kg；
 - 5U节点机柜深度 ≥ 1100 mm；
 - 5U节点机柜滑道（托盘）承重能力 ≥ 170 kg；
 - $600\text{mm} \leq$ 前后方孔条的距离 $\leq 900\text{mm}$ ；
 - 前方孔条离前门距离 $\geq 50\text{mm}$ 。

设备摆放规划

- **2U/4U节点**

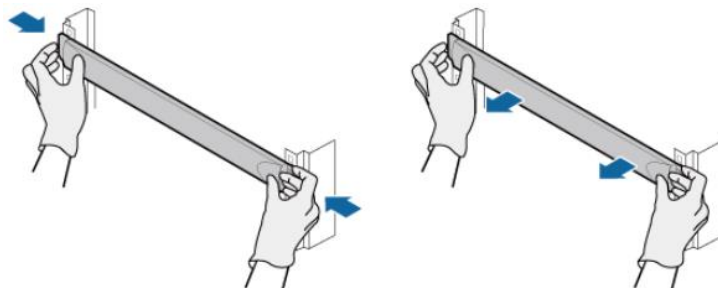
- 每个机柜最多放置12个2U节点，8个4U节点；
- 建议管理节点摆放在在机柜最下方，多机柜场景均分在不同机柜；
- 集群控制节点从机柜最上方依次向下摆放，多机柜场景尽量均匀分布在不同机柜中。

- **5U节点**

- 基于5U设备的重量、电源和功耗参数，匹配客户机房的承重能力、供电能力和散热能力，单机柜建议为摆放2~5台节点；
- 设备建议安装在机柜的6U~31U之间，机柜下方至少预留3U空间，上方预留4U空间。且始终按照从下到上的顺序安装，以防机柜因上重下轻而倾倒。

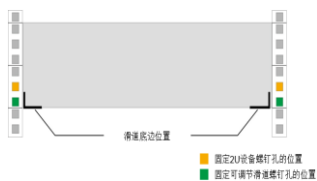
拆卸机柜假面板

- 根据设备在机柜中的位置信息确定待拆卸假面板的位置；
- 双手握住机柜中待拆卸假面板两侧的卡扣，用力向内侧按下并用力向外拔出。

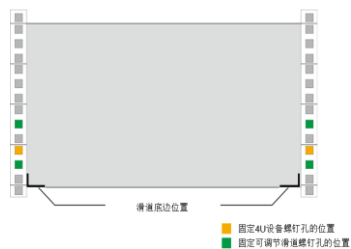


定位安装位置

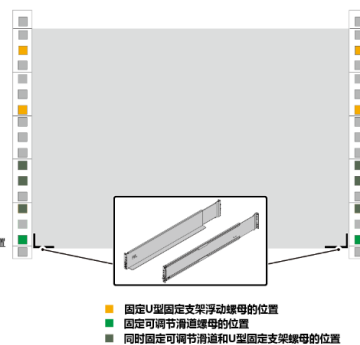
定位2U设备安装位置



定位4U设备安装位置

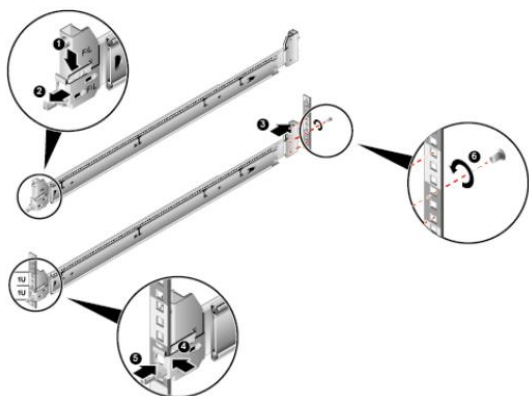


定位5U设备安装位置



- 每三个方孔对应1U高度。每1U高度都用刻度线标出。
- 如果机柜前后立柱上没有1U刻度线，请自行标记1U刻度线，以方便定位设备安装位置。
- 总体规划原则为交换机等重量较轻的设备安装在机柜上层，较重的设备安装在机柜下层。

安装抱轨

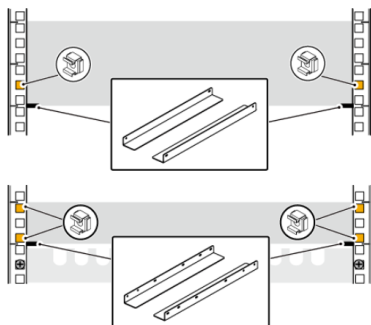


1. 按下抱轨前端挡片，同时向前拉升挂钩，如图中的步骤①、②所示；
2. 将抱轨后端定位销，插入机柜后侧的方孔，如图中的步骤③所示；
3. 将抱轨前端对准立柱孔位，向前推动抱轨卡入立柱孔位，如图中的步骤④所示；
4. 向后推动挂钩，使挂钩的金属片贴住立柱，如图中的步骤⑤所示；

可选：在抱轨后端立柱上的第三个方形孔上，安装一颗M6螺钉，以便固定抱轨，如图中的步骤⑥所示。

使用同样方法安装另一个抱轨。

安装L型滑道

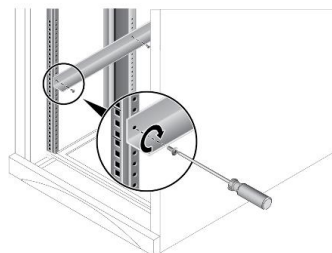


1. 确定L型滑道及浮动螺母的安装位置，如上图；
(1U设备（包括KVM、交换机）使用的L型滑道
为非标准滑道，区别在于非标准滑道中间有孔洞。)

2. 安装浮动螺母，如右图。

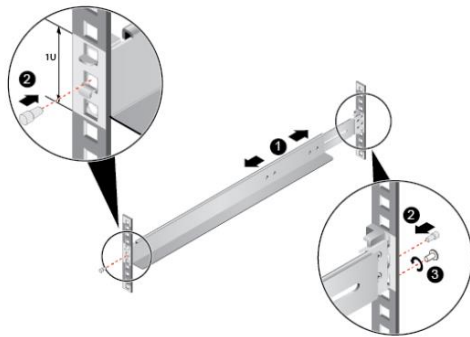


3. 安装L型滑道，如下图。

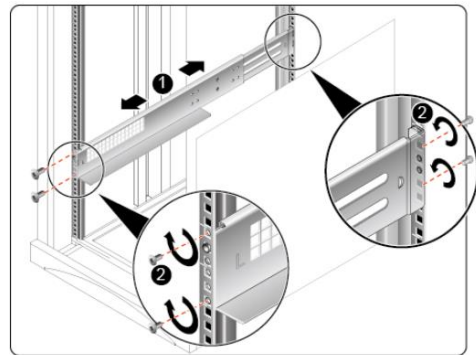


- KVM: Keyboard Video Mouse

安装可调节滑道



2U节点的可调整滑道安装

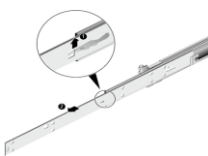


4U节点的可调整滑道安装

安装节点-2U设备

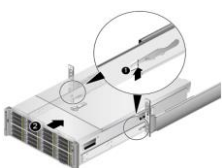
1. 将抱轨的内轨拉出轨道直至无法移动；
2. 将节点上的固定钉对准内轨的固定孔位，然后向前推进，直到听见“咔”的一声，确保卡扣弹起完全挡住挂钉，使节点固定到内轨上；
3. 按住节点两侧的解锁按钮，将节点推入抱轨，如上图①、②所示；
4. 按顺时针方向拧紧挂耳上的松不脱螺钉，固定节点；
5. 安装理线架：
 - a. 将支撑杆卡入外轨两侧的固定钉中，如图中①所示；
 - b. 将理线架支架的外壁卡入左侧外轨的固定钉中，并沿箭头方向拉固定理线架，如图中②所示；
 - c. 将理线架支架的内壁卡入左侧内轨的固定钉中，并沿箭头方向拉固定理线架，如图中③所示。
6. 安装前置硬盘模块。

安装节点-4U设备

- 

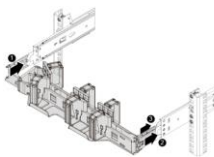
1. 将抱轨的内轨拉出轨道直至无法移动，向上拨解锁按钮，同时往外将内轨完全拉出抱轨并取下，如图中①、②所示；
- 

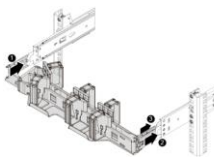
2. 将节点上的固定钉对准内轨的固定孔，沿箭头方向推动内轨，直到听见“咔”的一声，确保卡扣弹起完全挡住挂钉，使节点固定到内轨上；
- 

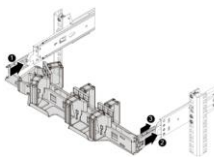
3. 至少四人从两侧将节点抬起，使节点上的内轨对准抱轨轨道，同时将节点推入机柜；
- 

4. 按住节点两侧的解锁按钮，将节点推入抱轨，如图中①、②所示；
- 

5. 按顺时针方向拧紧挂耳上的松不脱螺钉，固定节点；
6. 安装理线架：

 - 

a. 将支撑杆的左外侧卡槽卡入左侧外轨中，如图中①所示；
 - 

b. 将理线架支架的右侧卡槽卡入右侧外轨中，如图中②所示；
 - 

c. 将理线架支架的右内侧卡槽卡入右侧内轨中，如图中③所示。
- 

7. 安装前置硬盘模块。

安装节点-5U设备 (9550)



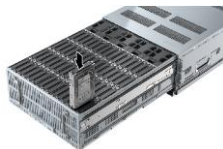
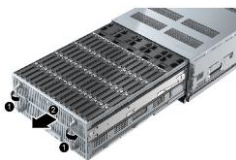
1、连接电源线：

将电源线的一端插入电源模块的线缆接口，如上图 1 所示。使用魔术贴将电源线固定在电源模块的拉手上。将电源线分别沿系统插框两侧的走线槽进行布线，如上图 2 所示。



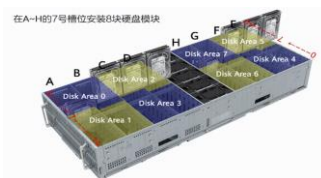
2、安装系统插框：

至少3人使用物料升降平台车将系统插框上架，并沿滑道将系统插框平稳推入机柜中，如上图中的步骤 1 所示。使用十字螺丝刀固定系统插框前面板左右两侧的6颗螺钉，如上图中的步骤 2 所示。



3、安装前置硬盘抽屉中的硬盘模块：

i. 将前置硬盘抽屉的拉手拉到垂直状态，握住前置硬盘抽屉的拉手缓慢拉出前置硬盘抽屉直至出现风扇模块，如上图 1、2 所示；
ii. 将硬盘模块沿滑道推入硬盘抽屉，直至听见“咔”的一声；
然后按上述方式完成后置硬盘抽屉中硬盘模块的安装。



注：前8个硬盘模块分别插入硬盘抽屉A～H的7号槽位。剩余硬盘模块平均分配的方式从左到右、从外到内依次插入到每个硬盘区域（Disk Area），以Disk Area 0为例，插入硬盘模块的顺序依次为A0～A6、B0～B6。硬盘模块的插法请参见上图。



4、安装前面板面罩：

将前面板面罩左侧卡扣与系统插框左侧卡槽完全对齐，旋转推动前面板面罩右侧，使前面板面罩右侧锁扣锁定到系统插框，如上图 1、2 所示。

安装节点-5U设备 (9950)



1、安装系统插框：

A、至少3人使用物料升降平台车将系统插框上架，并沿滑道将系统插框平稳推入机柜中；

B、使用十字螺丝刀固定系统插框前面板左右两侧的4颗螺钉，如上图所示。

2、安装节点：

A、按压节点解锁按钮，节点扳手自动弹开，向上下两侧完全打开节点扳手，将节点沿滑道完全推入系统插框，如右图中①、②所示；

B、当节点推入系统插框后，同时复位节点扳手，直至听到“咔”的一声，此时节点完全锁定到系统插框；

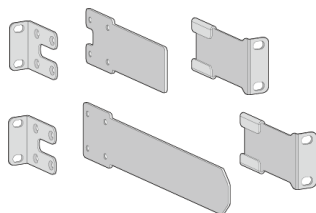
注：节点按照从0~7的顺序依次安装，当系统插框中的节点未满载时，请在空闲的节点槽位安装辅助散热模块，节点编号如下图所示。



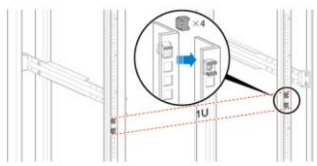
3、安装前面板面罩：

将前面板面罩左侧卡扣与系统插框左侧卡槽完全对齐，旋转推动前面板面罩右侧，直到听见“咔”的一声，确保前面板面罩固定到系统插框，如上图中①、②所示。

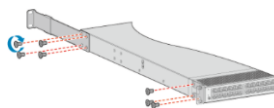
安装交换机



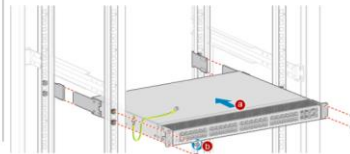
1. 对于不同款型的交换机，标配的前挂耳型号及安装方式不同，如图所示。安装前挂耳到交换机时请使用与设备配套发货的挂耳；



2. 根据规划好的安装位置，确定浮动螺母在方孔条上的安装位置。用一字螺丝刀在机柜前方孔条上安装4个浮动螺母，左右各2个，挂耳上的固定孔对应着方孔条上间隔1个孔位的2个安装孔。保证左右对应的浮动螺母在一个水平面上；

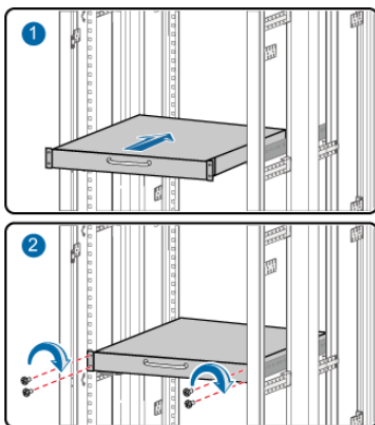


3. 参考上图场景完成前挂耳、后挂耳和后挂耳滑道安装；



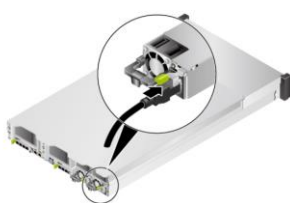
4. 安装交换机到机柜
 - a. 双手托住交换机将交换机搬进机柜中，使两边的后挂耳都对准后挂耳滑道后再慢慢插入；
 - b. 用十字螺丝刀和M6螺钉将交换机前挂耳固定到方孔条的浮动螺母上。

安装KVM



1. 用M4螺钉把KVM两侧的平衡杆安装在滑道上；
2. 将KVM轻轻放在机柜的相应滑道上，并平稳推入机柜内；
3. 使用M6螺钉将KVM两侧的挂耳固定到机柜前面侧柱的方孔条上；
4. 检查KVM的安装情况，确认KVM水平安装，且固定螺钉已经拧紧。

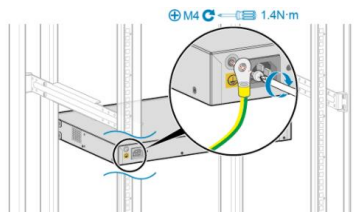
连接保护地线和电源线



2U节点电源线连接



4U节点电源线连接



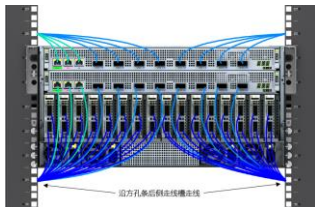
交换机连接接地线缆



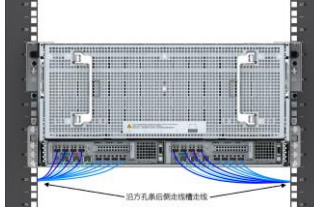
KVM连接接地线缆和电源线

连接节点信号线

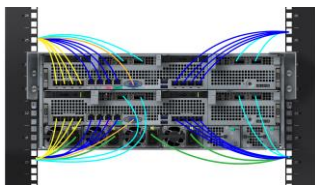
OceanStor Pacific 9950线缆布放规则



OceanStor Pacific 9550线缆布放规则



OceanStor Pacific 9546(1机箱2节点, 每节点2 CPU)线缆布放规则



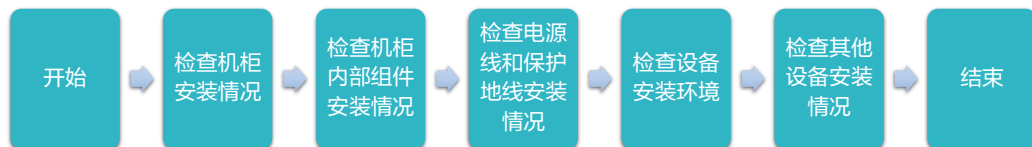
OceanStor Pacific 9546(1机箱1节点, 每节点2 CPU)线缆布放规则



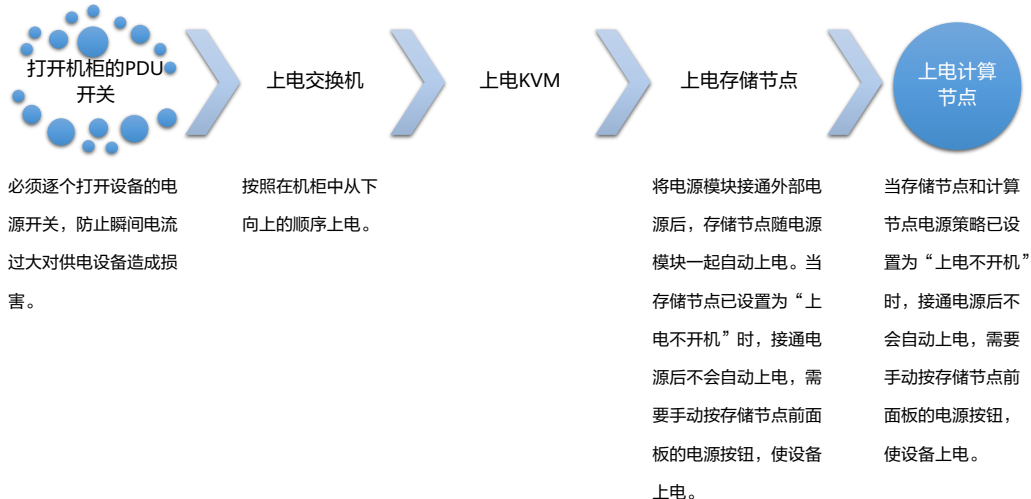
- 操作步骤：
 - 佩戴防静电腕带或防静电手套。
 - 识别存储设备的端口，确定信号线缆的连接关系。
 - 获取所需类型的线缆，并粘贴标签。
- 存储设备中所有信号线缆的布放不能影响该设备和相邻设备中硬盘抽屉的抽拉及部件的拆卸。

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。



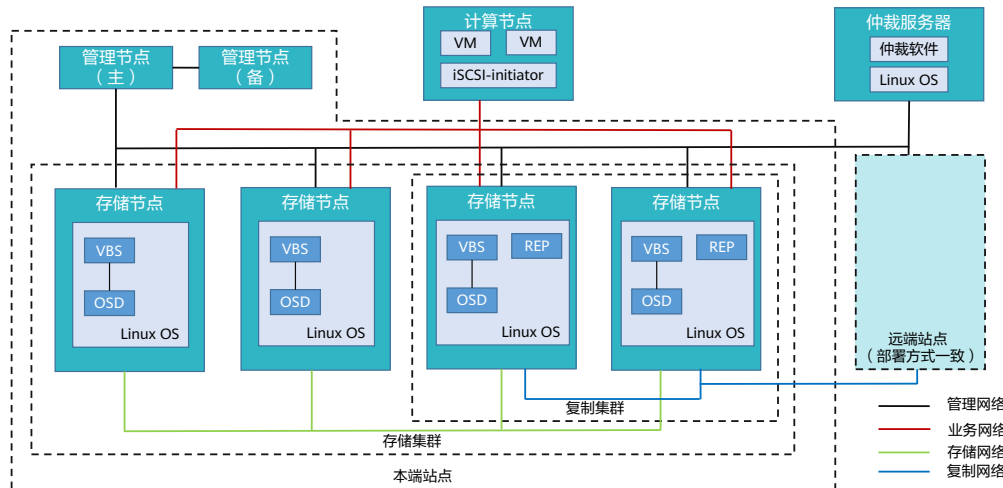
分布式存储设备上电流程



目录

1. 硬件安装流程
- 2. 组网规划**
3. 软件安装流程
4. 业务配置

网络平面划分（结构化服务场景）

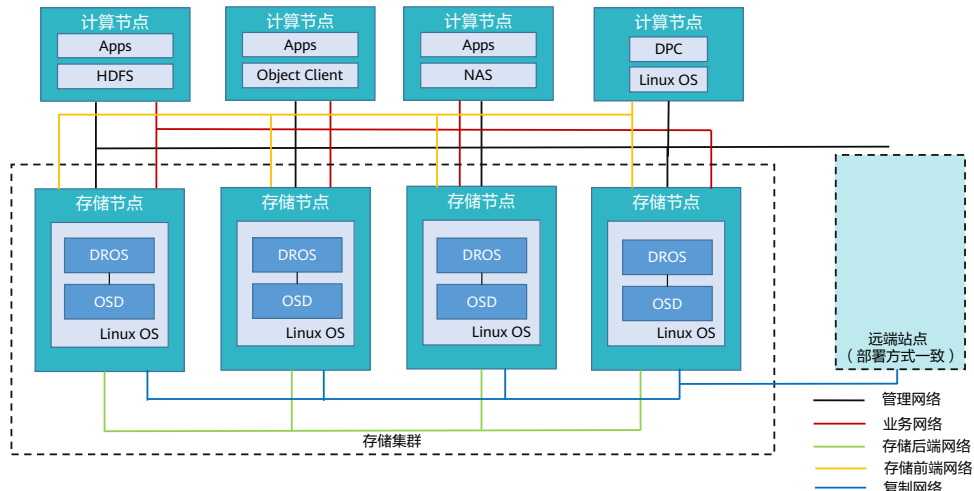


25 Huawei Confidential



- 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。

网络平面划分（非结构化服务场景）

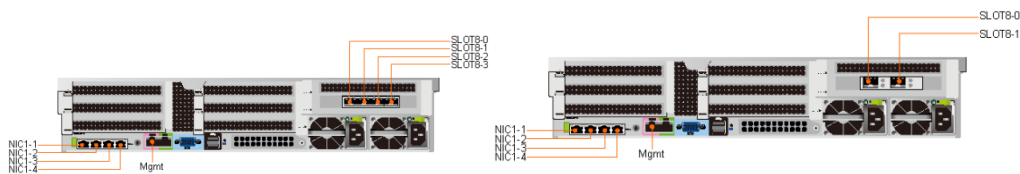


26 Huawei Confidential



- 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。

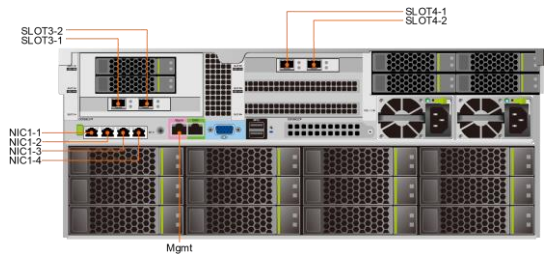
OceanStor Pacific 9920 Kunpeng(NVMe SSD)节点接口规划



接口名称	接口类型	接入网络	说明
SLOT8-0和SLOT8-1	10GE/25GE/100GE/IB接口	存储网络	连接到存储交换机，2个接口组成逻辑上的bond。
NIC1-1	GE接口	管理网络	连接到管理交换机。 说明： 当需要将管理网口组成bond时，管理网口选择NIC1-1和NIC1-2。
Mgmt	GE接口	BMC网络	连接到BMC交换机。

- Database/原生OpenStack/文件共享/华为云生态（VBS部署在计算节点，前后端共享网络）为例，其他场景请参考产品文档。

OceanStor Pacific 9540 Kunpeng 节点接口规划

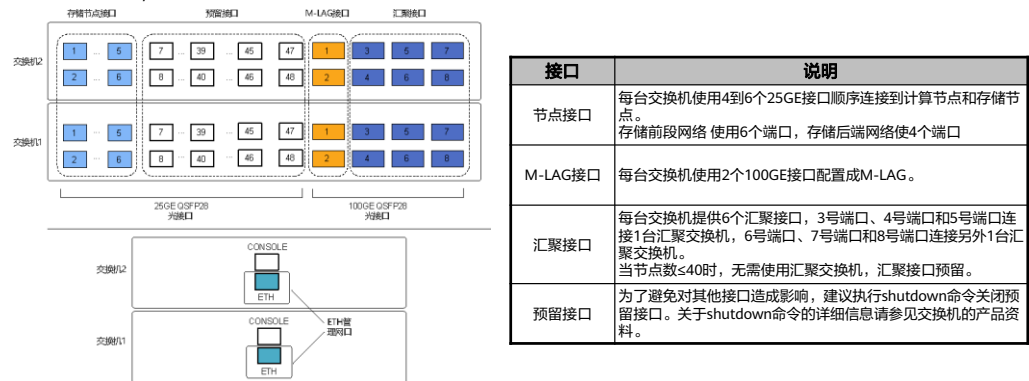


接口名称	接口类型	接入网络	说明
SLOT4-1和SLOT4-2	25GE/100GE/100Gbps IB接口	存储网络/控制网络	连接到存储交换机。
NIC1-1 说明： 当需要将管理网口组成bond时，管理网口选择NIC1-1和NIC1-2。	GE接口	管理网络	连接到管理交换机。
Mgmt	GE接口	BMC网络	连接到BMC交换机。

- 文件服务（DPC场景）,当存储网络采用25GE/100GE/100Gbps IB网络时为例，其他场景请参考产品文档。

25GE交换机接口规划

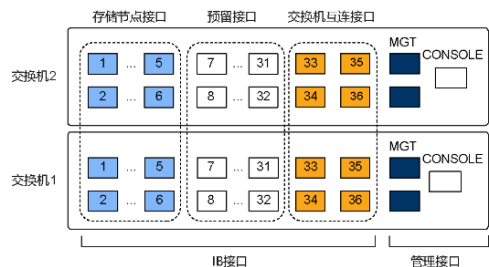
- 以部署6个节点（2个计算节点，4个存储节点），业务交换机、存储交换机和管理交换机分别独立部署为例，介绍25GE组网的交换机接口规划。
- 选用CE6860/CE6865交换机时，业务交换机和存储交换机接口规划&接口说明。



IB交换机接口规划

- 存储交换机根据节点数不同，分为以下几种情况：
 - 当节点数≤32时，存储交换机选用2台SB7800，无需汇聚。
 - 当32 < 节点数≤112时，存储交换机选用4台SB7800，汇聚交换机选用SB7800。
 - 当112 < 节点数≤192时，存储交换机选用2台CS7520，无需汇聚交换机。
 - 当192 < 节点数≤256时，存储交换机选用4台CS7510，无需汇聚交换机。
- 以部署2个计算节点和4个存储节点，业务交换机、存储交换机和管理交换机分别独立部署为例，介绍100Gb IB组网的交换机接口规划。

以组网为100Gb IB时为例：



接口	说明
节点接口	每台交换机使用6个IB接口顺序连接到各节点。
交换机互连接口	每台交换机分别使用4个IB接口进行互连。
MGT接口	每台交换机的1个MGT接口连接到BMC/管理交换机。
预留接口	为了避免对其他接口造成影响，建议执行shutdown命令关闭预留接口。关于shutdown命令的详细信息请参见交换机的产品资料。

管理交换机和BMC交换机接口规划

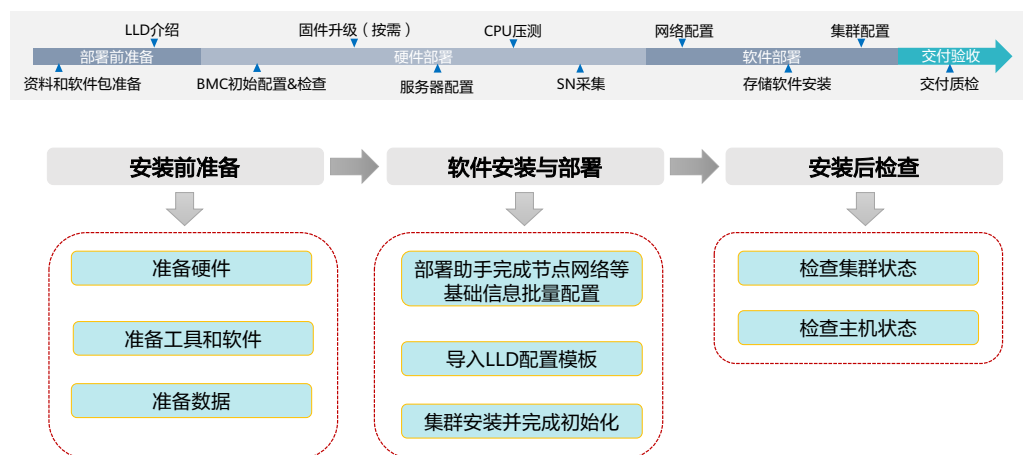
交换机	接口	说明
管理交换机	NIC接口	管理交换机使用6个GE接口分别连接到各存储节点的NIC接口。 如果存储节点的NIC口采用bond模式，此时管理交换机需要规划12个GE接口。
	交换机ETH管理接口	管理交换机使用5个GE接口连接到业务交换机、存储交换机和BMC交换机的ETH管理网口。
	汇聚接口	管理交换机使用2个10GE接口连接到用户的管理汇聚交换机。 当节点数 ≤ 40 时，无需使用汇聚交换机，汇聚接口预留。
	预留接口	为了避免对其他接口造成影响，建议执行shutdown命令关闭预留接口。关于shutdown命令的详细信息请参见交换机的产品资料。
BMC交换机	Mgmt接口	BMC交换机使用6个GE接口分别连接到各存储节点的Mgmt接口。
	汇聚接口	BMC交换机使用2个10GE接口连接到用户的BMC汇聚交换机。 当存储节点数 ≤ 40 时，无需使用汇聚交换机，汇聚接口预留。
	预留接口	为了避免对其他接口造成影响，建议执行shutdown命令关闭预留接口。关于shutdown命令的详细信息请参见交换机的产品资料。

- 当管理网络独立使用交换机时，BMC交换机和管理交换机使用情况分为两种：
 - 管理交换机和BMC交换机独立使用
 - 管理交换机和BMC交换机共同使用

目录

1. 硬件安装流程
2. 组网规划
- 3. 软件安装流程**
4. 业务配置

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

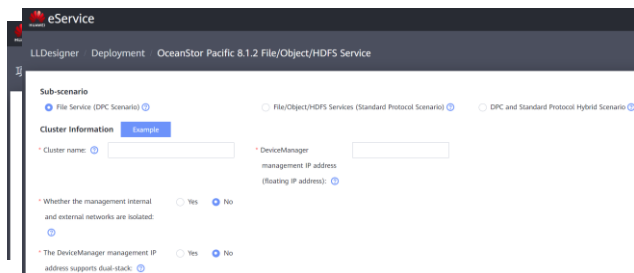
交付阶段	任务类别	实施内容	使用工具
部署前准备	部署前准备	准备部署需要用到资料、工具和软件包	NA
硬件部署	设备上架及上电	设备上架及上电	NA
	组网连线及交换机配置	组网连线及交换机配置	网线、Console通信线缆
	BMC初始配置&检查	配置BMC IP、子网掩码/前缀、网关和BMC主机名 固件版本、OS版本和驱动版本配套检查	SmartKit分布式部署助手
	固件升级（按需）	BMC、BIOS、CPLD、RAID卡CPLD等固件升级	SmartKit分布式存储固件升级
	OS安装（按需）	批量安装EulerOS	SmartKit分布式存储开局交付
	服务器配置（按需）	BMC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介绍



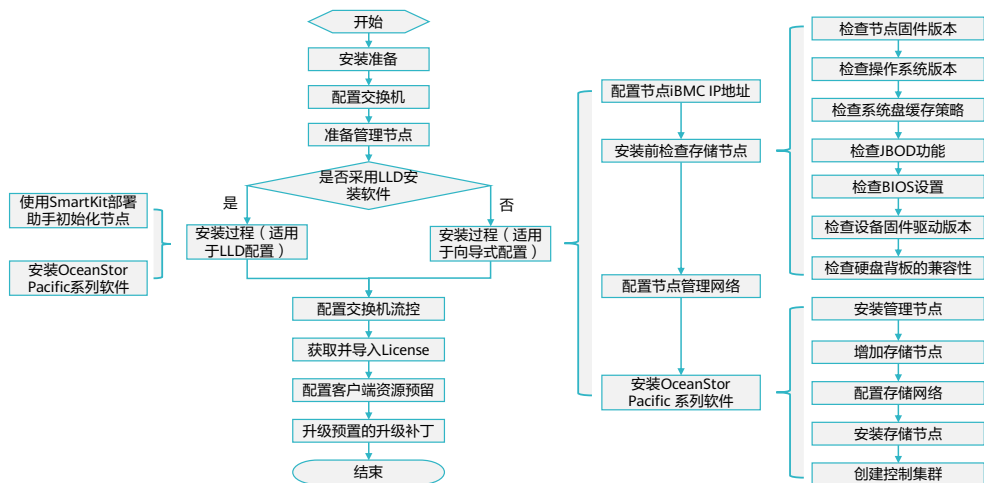
支持的子场景

- 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



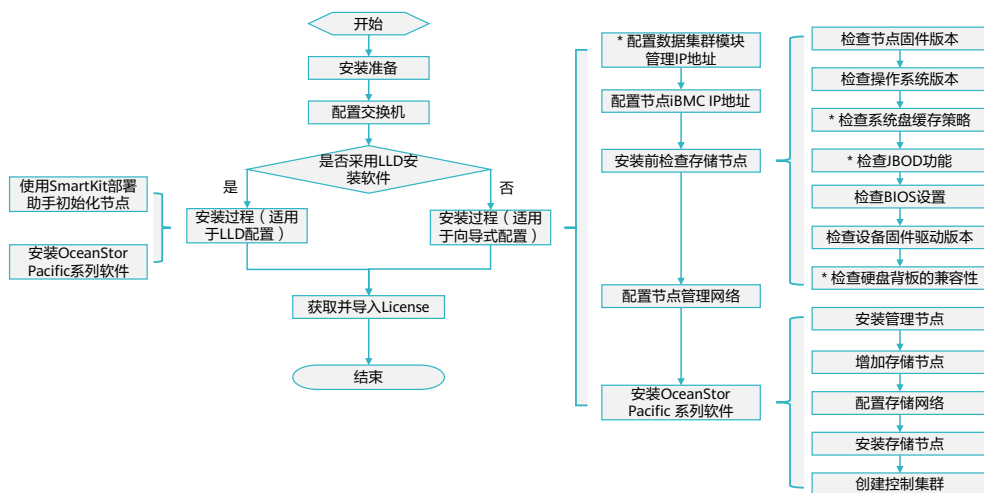
- 软件安装支持LLD配置和向导式配置两种方式，推荐使用LLD配置方式：
 - LLD配置模式：
 1. 需要先使用LLDesigner工具完成节点规划、网络规划和集群规划，并导出部署配置文件；
 2. 根据部署配置文件使用SmartKit部署助手完成节点初始化（节点初始化包含：配置节点iBMC IP地址、安装前检查存储节点、配置节点管理和存储网络）；
 3. 通过DeviceManager Client采用LLD配置模式完成OceanStor Pacific系列软件的安装。
 - 向导式配置模式：
 1. 手动在DeviceManager界面上完成节点添加、网络配置和控制集群的创建；
 2.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

块服务场景软件安装流程



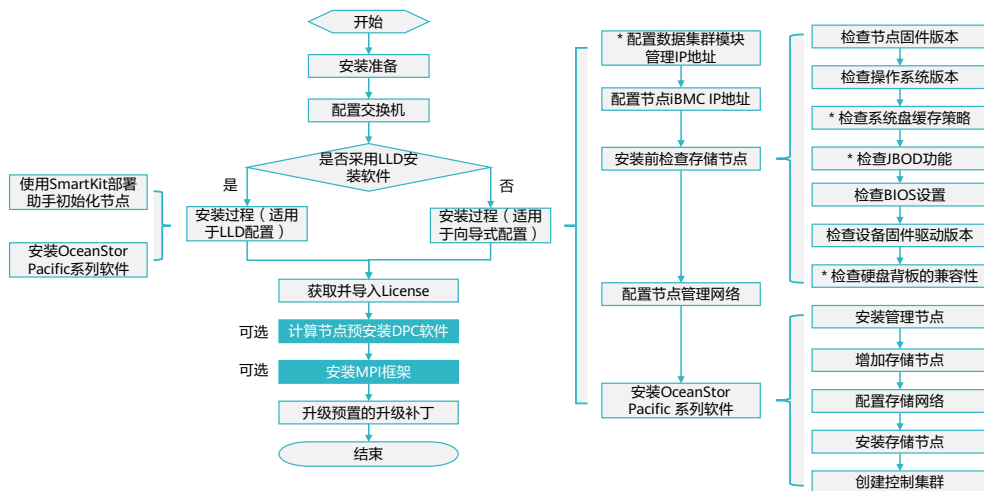
- OceanStor Pacific系列配套FusionCompute的典型分离部署方案，完成软件安装部署后，需要创建存储池并启用VBS，以便应用服务器能够通过VBS访问分布式存储资源。

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

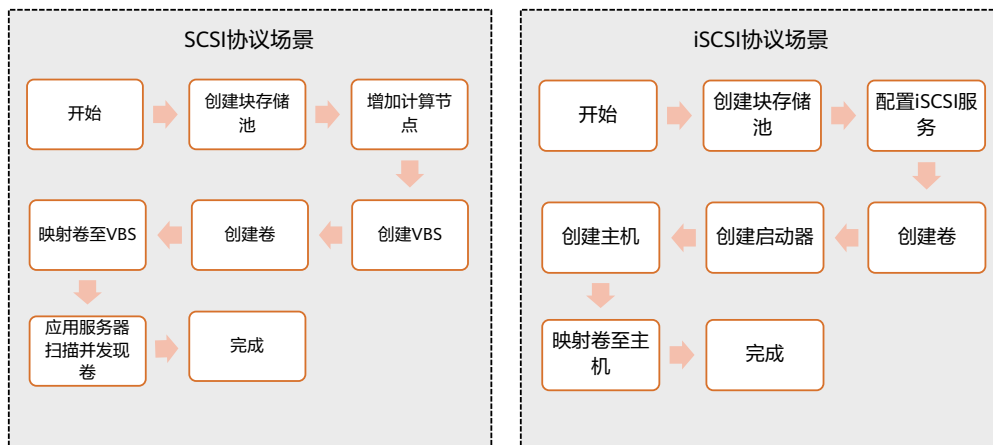


- DPC（Distributed Parallel Client）分布式并行客户端场景软件安装包含管理节点和存储节点的安装部署，采用管理节点和存储节点合并部署的方式。
- 同时也包含了在计算节点上预安装DPC软件的相关内容。
- 带 * 号内容表示适用于部分机型。

目录

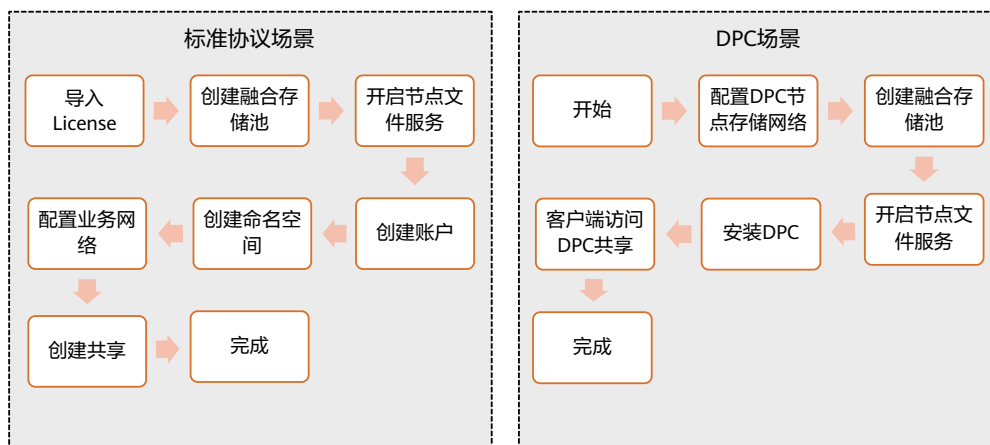
1. 硬件安装流程
2. 组网规划
3. 软件安装流程
- 4. 业务配置**

块服务业务配置流程



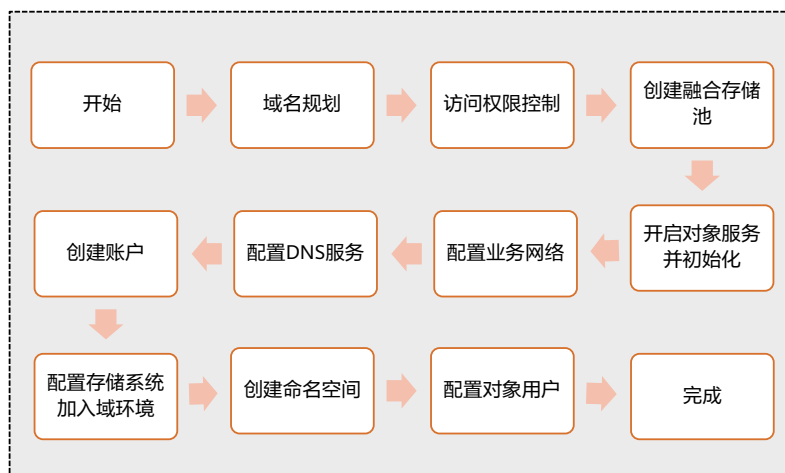
- 存储系统通过SCSI和iSCSI接口向上层应用提供高性能与高扩展能力的块服务，满足虚拟化、云资源池及数据库等场景的存储需求。
- 块服务当前的主要应用场景分为两大类：
 - 资源池：将存储节点池化，建立大规模块存储资源池。可以和各种虚拟化平台集成，如华为FusionSphere、VMware vSphere、开源OpenStack等，按需分配存储资源，支持各种业务应用（如SQL、Web、行业应用等）。
 - 数据库：以SSD做Cache或主存，极大地提升了存储系统的性能，又保留了分布式存储的高扩展性基因，从而解决了企业关键数据库、关键ERP（Enterprise Resource Planning）/CRM（Customer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

文件服务业务配置流程



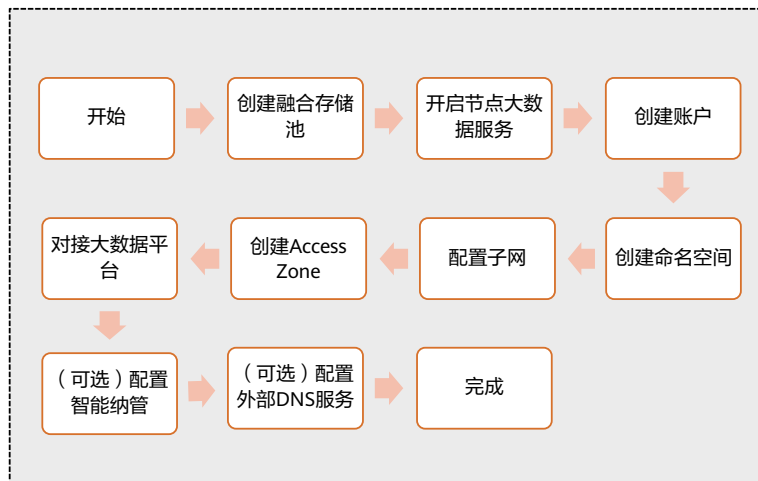
- 文件服务用于提供稳定可靠的文件共享功能。
- 文件共享是NAS存储系统的基本特性。OceanStor Pacific系列文件服务同时支持CIFS、NFS协议访问共享文件和DPC客户端访问共享文件，兼容多种操作系统，可以实现异构网络环境下的文件资源共享，支持对命名空间同时设置CIFS、NFS和DPC共享。

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



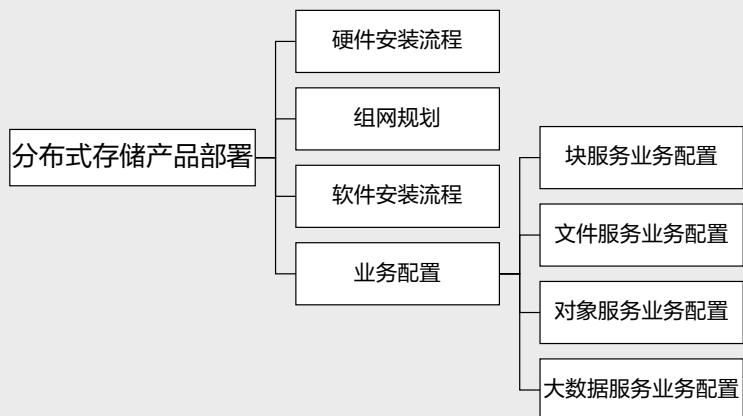
- 大数据服务提供基于原生HDFS的大数据存算分离方案，实现存储与计算资源按需配置，提供相同用户体验的同时，助您降低总拥有成本；支持与原有计算存储一体化架构共存。广泛应用于金融大数据、互联网日志留存大数据、政务大数据和智慧城市大数据等场景。
- 大数据服务使用OceanStor Pacific系列存储，提供原生HDFS接口，对接FusionInsight、Cloudera CDH、Hortonworks HDP等大数据平台，完成大数据的存储与计算，提供大数据分析服务供上层大数据应用使用。

思考题

1. 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。
2. （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

- 答案
- 1、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。
- 2、A。

本章总结



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://enterprise.huawei.com/cn/>
- 技术支持: <https://support.huawei.com/enterprise/>
- 在线学习: <https://support.huawei.com/learning/Index!toTrainIndex>
- 华为Support案例库: <https://support.huawei.com/enterprise/servicecenter?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.

Copyright©2023 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.



存储系统性能调优



前言

- 存储系统是企业IT基础架构重要的组成部分，为企业内部众多的IT系统提供数据存储服务。随着数字化转型的深入，企业的IT系统建设也进一步加快，这一方面带来了数据量的急剧增长，另一方面也提高了数据的访问频率，存储的性能瓶颈的影响也会被进一步放大。
- 本章主要讲述存储系统性能相关概念，存储性能评估方法，存储系统性能问题定界定位思路和方法，存储系统性能调优以及性能分析测试工具使用方法。

目标

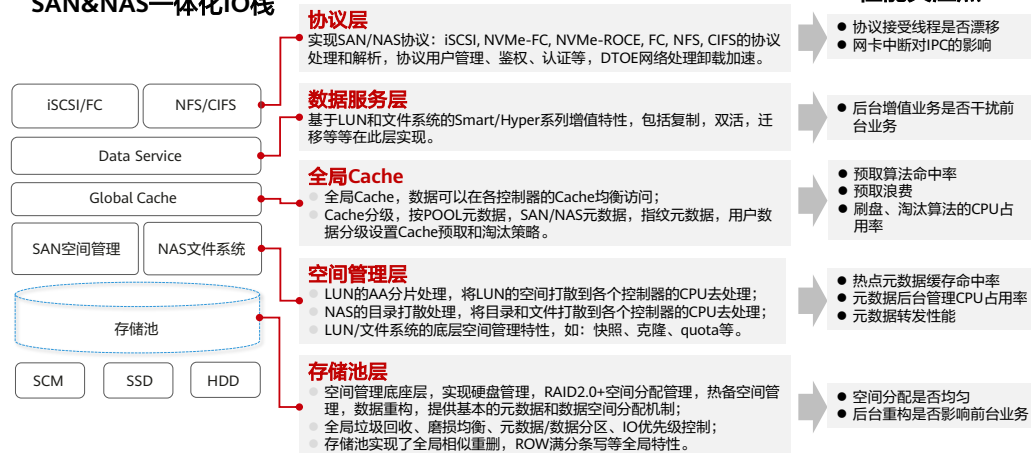
- 学完本课程后，您将能够：
 - 理解存储性能的相关概念；
 - 了解存储性能评估方法；
 - 理解存储系统性能问题定位方法；
 - 掌握存储系统性能调优方法；
 - 掌握存储系统性能测试及工具使用方法。

目录

1. 存储性能概述
2. 存储性能评估
3. 存储性能问题
4. 存储性能调优
5. 存储性能测试

闪存存储IO栈简介

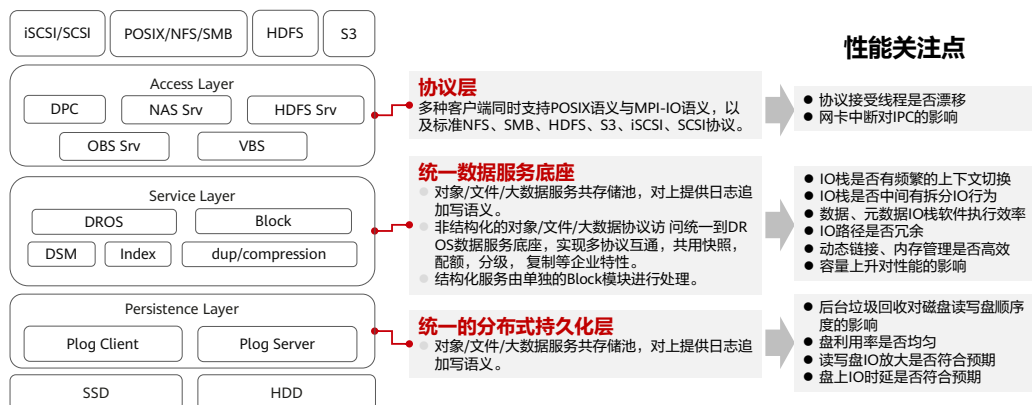
SAN&NAS一体化IO栈



- 以常用的FC-SAN存储为例, 服务器操作系统一般会使用文件系统管理文件, 文件系统是建立在存储LUN之上, 文件的读写会对应着存储的IO操作; 文件会被分为多个Block, Block大小固定, 一般是4KB-16KB; 数据Block块会被发送到服务器的HBA卡, HBA卡再将其转换为FC协议的数据帧 (Data frame), 并通过SAN网络传输到存储系统的前端口; 存储的前端口继续将这些数据帧重新封装成数据Block, Block大小一般为4KB, 并将这些数据Block块传输到存储控制器中; 存储控制器中会有存储缓存 (Cache), 分为读缓存与写缓存, 根据缓存的算法规则, 部分缓存命中的IO数据流会立刻返回IO确认, 缓存未命中的IO数据流则会需要继续访问磁盘阵列; 由于使用RAID 2.0+技术, 一个数据IO流实际上对应着多个磁盘的并发读写。
- 以NFS为例, NAS存储是有着明显异于SAN存储的特点, 比如客户端缓存、Server的无状态性等。首先客户端并不是直接访问NAS文件系统, 而是通过客户端缓存, 是服务端的文件系统目录树映射到了客户端, 实际在文件读写时, 需要循环读写固定大小的页面, 比如64KB; 而Server端的无状态性体现在不需要维护客户端的协议状态信息, 客户端通过RPC调用操作Server端的文件系统数据, 但也不能获取Server端的状态, 当连接中断时, 可以不停地连接重试。

分布式存储IO栈简介

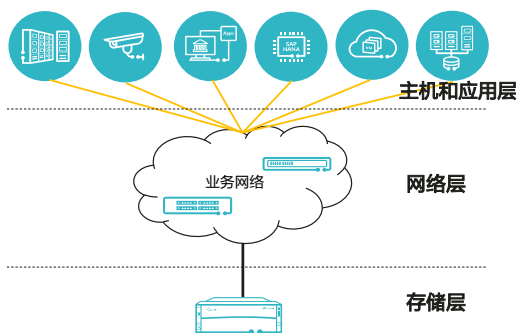
结构化/非结构化IO栈



- 分布式存储系统是采用可扩展的集群架构，通过数据副本算法将数据分散存储在多台独立的设备上，分布式集群之间一般通过通用 TCP/IP 网络连接。相比于其传统的集中式存储阵列，分布式存储系统可以通过多台存储服务器来分担存储负荷，可以满足大规模存储应用的需要。
- 以OceanStor Pacific的块存储为例，上层应用下发写IO请求到存储服务的VBS（Virtual Block Service），VBS收到该IO请求，根据第一层的DHT hash算法将数据转到指定存储节点；由这个存储节点上的EDS(Block+Index)模块处理该数据；EDS接收到写IO请求后，如果是小IO，以小比例EC形式写入L1Cache，同时该EDS所在存储节点的内存中仍然保持一份该数据，EDS返回写IO成功给VBS，再由VBS返回给上层应用。待L1 Cache内存中的数据聚合到更大的块，聚合成满分条计算EC冗余后，通过PLOG接口写到OSD。如果是大IO，就不需要进入L1 Cache进行聚合，直接按用户配置的EC比例计算冗余，写入OSD。

存储典型组网

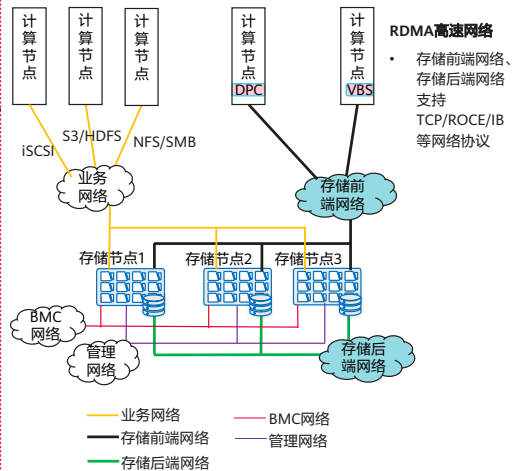
闪存存储



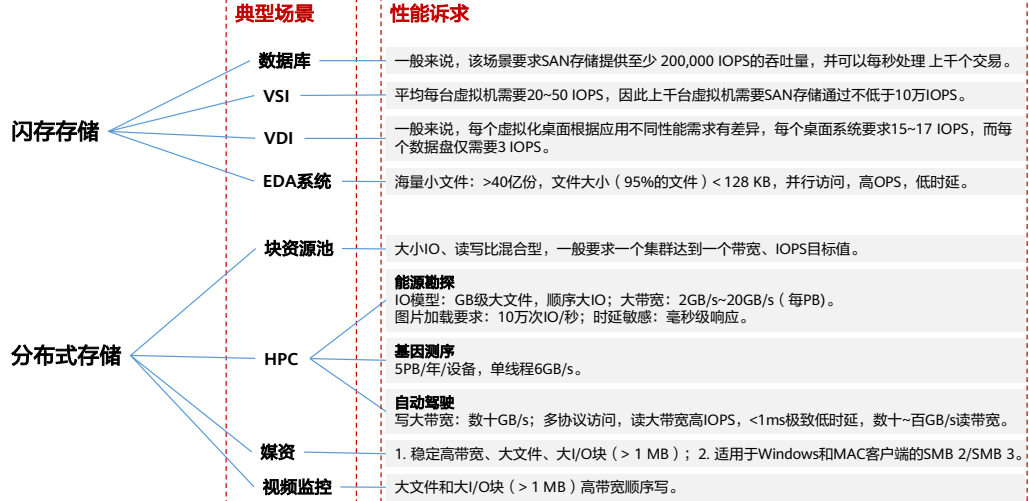
网络分层简单

- 主机层和存储层可统计时延，性能问题易定界

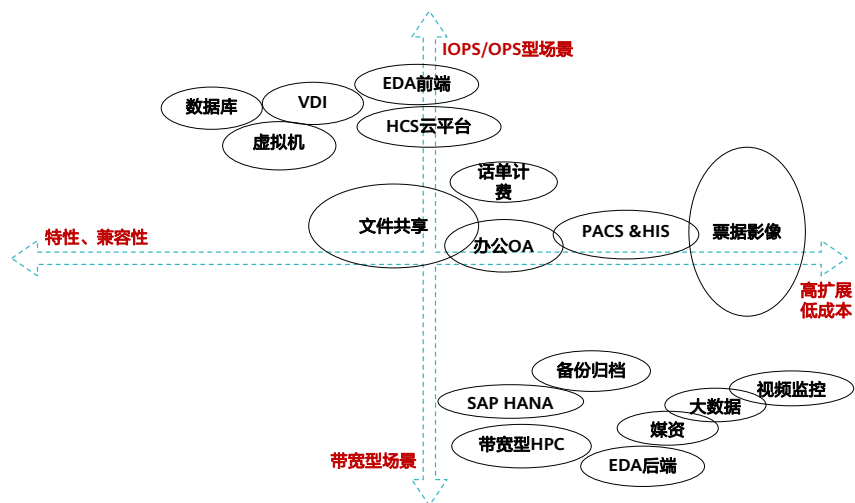
分布式存储



存储典型场景及性能诉求



业务场景



- 存储行业服务泛政府、医疗、金融、制造、能源、教育、运营商等诸多领域，数据类型和业务场景复杂。

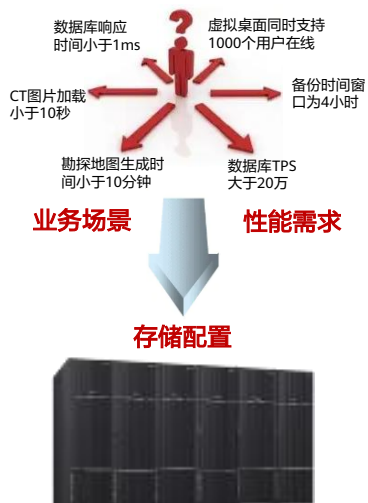
常用场景的典型业务特征

业务场景	业务特征	场景画像
数据库OLTP	以随机小IO为主，读多写少，热点区域明显，有少量顺序写IO	8KB随机读写，比例为7:3，20%热点
数据库OLAP	以顺序读大IO为主，一次批处理数据量决定热点区域	256KB顺序读写，比例为9:1
虚拟机VDI	启动风暴：随机小IO读密集型操作 登录风暴：随机小IO写密集型操作 平稳状态：多路顺序小IO写多读少	启动风暴：8KB随机读写，比例为9:1 登录风暴：8KB随机读写，比例为2:8 平稳状态：8KB随机读写，比例为2:8，80%热点
EDA前端	海量小文件随机小IO业务，60%元数据操作	64KB海量小文件随机读写，读写比例为4:6
EDA后端	90%的GB级大文件写操作	GB级文件，读写比例为1:9，顺序IO大小为512KB-1MB
票据影像	KB级随机小文件，有明显冷热特性	256-512KB小文件随机创建写+随机读，读写比例1:1
备份归档	以顺序大IO为主，备份和归档为写业务，恢复为读业务	备份归档：1MB顺序写 恢复：1MB顺序读

目录

1. 存储性能概述
- 2. 存储性能评估**
3. 存储性能问题
4. 存储性能调优
5. 存储性能测试

存储性能评估



什么是存储性能评估？

- 把客户的业务场景和性能诉求转化为存储产品的业务模型和性能需求，评估出可以满足的最小存储配置。

存储性能评估的目的是什么？

- 用最少的硬件资源和最合理的软件配置，提供某一种或多种业务的最佳性能，达到性能和成本的平衡。

为什么需要做性能评估？

- 存储的应用场景繁多，数据类型多变，业务模型复杂。
- 存储的商用领域广泛，不同客户对存储的性能诉求不同，同一客户对不同业务场景的性能诉求也不同。
- 存储软硬件配置灵活多变，网络、控制器、磁盘、内存都影响性能。

存储性能评估体系 – 存储性能指标

4+1 个评价维度，4个**基础性能指标**（IOPS、带宽、时延、OPS）+ **第三方评测**（SPC、SPEC、IO500）

IOPS	带宽
▪ 定义：存储每秒钟可以处理I/O的个数，用于衡量存储的I/O处理能力。 ▪ 适用场景：数据库。	▪ 定义：存储每秒钟可以处理的数据总量，用于衡量存储的吞吐量。 ▪ 适用场景：媒资和视频监控。
时延	OPS
▪ 定义：存储处理一个I/O所需花费的时间，单位为ms，用于衡量存储的快速响应能力。 ▪ 适用场景：数据库，VDI，视频点播和视频监控。	▪ 定义：存储每秒钟可以处理的文件操作数，用于衡量存储在NAS场景下的文件处理能力。 ▪ 适用场景：文件服务器。

- 在通常情况下，如果数据的读写是在连续的磁盘空间上，可以认为是顺序IO；如果应用读取的数据分布在不连续的磁盘空间，且无固定的顺序，则视为随机IO；如果一部分数据是顺序读写，一部分数据是随机读写，则视为混合类型IO。
- 性能指标是系统性能的直观体现，用户可以通过性能指标了解系统的运行状态。衡量存储系统性能的主要指标包括：
 - IOPS (Input/Output Operations Per Second)，即存储每秒能处理的I/O个数，用于衡量存储的快速响应能力。用来衡量存储系统的I/O处理能力。在数据库OLTP（Online Transaction Processing）业务、SPC-1认证等应用场景，通常以IOPS衡量系统的性能。
 - OPS(Operation Per Second)，针对NAS业务的性能指标，表示存储每秒能处理的操作个数，这些操作包括了read、write、open、close、readdir、delete、mkdir等文件操作。
 - 带宽也被叫做吞吐量(Throughput)，表示存储每秒能处理的数据的总量，用于衡量存储的吞吐能力，常以MB/s或GB/s为单位。在数据库OLAP（Online Analytical Processing）业务、媒资业务、视频监控业务等应用场景，通常以带宽衡量系统性能。如果每个IO的大小都相同， $\text{带宽} = \text{IOPS} \times \text{IO Size}$ 。
 - 时延(Latency)指的是发起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。例如，数据库OLTP业务一般时延要求10ms以下；VDI（Virtual Desktop Infrastructure）场景一般时延要求30ms以下；视频点播和视频监控的时延要求随码率的不同而不同。对于并发数稳定的系统来说， $\text{IOPS} = \text{并发数} / \text{平均时延}$ 。

存储性能评估体系 – 第三方评测之SAN性能认证

SPC组织介绍

SPC (Storage Performance Council) 是存储性能委员会的简称。它的建立就是为了定义、标准化以及促进存储子系统的基准，同时向计算机厂商和用户发布客观、权威、公正的性能数据，为客户提供参照，为存储厂商提供交流平台携手提升存储性能。它公布的性能数据得到了各大存储厂商和客户的广泛认可。SPC会员包括Huawei, IBM, HDS, NetApp, HP等主流存储厂商。

SPC认证基准介绍

SPC基准组织制定了SPC-1和SPC-2两大SAN性能测试基准。**SPC-1主要针对随机小IO快速响应的应用场景**，用于度量SAN阵列在OLTP、Database和Email三种应用下的性能表现。**SPC-2主要针对大IO顺序访问的应用场景**，考察存储系统大文件处理，大型数据库查询及VOD视频点播三种应用的性能表现。

- SPC-1对存储的IOPS性能进行认证，要求时延小于30ms；SPC-2对存储的带宽性能进行认证。

存储性能评估体系 – 第三方评测之NAS性能认证

SPEC组织介绍

SPEC(Standard Performance Evaluation Corporation标准性能评估机构)是一个全球性的、权威的第三方应用性能测试组织，它旨在确立、修改一系列服务器以及存储性能评估标准。该组织成立于1988年，是由斯坦福大学、清华大学、微软、等全球几十所知名大学、研究机构、IT企业组成的第三方测试组织。SPEC经过多年的发展积累，该测试标准得到了全球众多用户的广泛认可。

SPEC认证NAS基准介绍

SPEC SFS 是SPEC基准组织制定的、全球唯一、业界认可的NAS性能测试基准。SPEC组织包含Huawei, EMC, NetApp, IBM等主流存储厂商。**SPEC SFS的认证结果对众多海量小文件随机访问应用领域有重要的借鉴意义，包括HPC、广电行业渲染、芯片仿真设计等。**目前SPEC SFS2014基准已发布，分为了四种场景测试：SWBUILD, VDA, VDI, DATABASE, 更为全面的考察存储系统在各应用场景下的性能表现。

存储性能评估体系 – 第三方评测之IO500性能认证

IO500组织介绍

IO500是高性能计算领域针对存储性能评测的全球排行榜，是高性能计算领域最权威的榜单之一。IO500测试包括数据带宽BW(GiB/S)和元数据性能MD(KIOP/S)两大部分，各项分数取几何平均后得到总分。在HPC计算领域，不仅CPU算力非常重要，IO系统的数据传输更是瓶颈。自2017年11月开始，每年IO500榜单会在高性能计算领域的顶级会议——美国SC大会和德国ISC大会上发布。

IO500认证HPC基准介绍

IO500定义了一个全面的基准性能测试套件，可以对高性能存储系统进行标准的测试和比较。从测试方法上，IO500进行两组测试，分别完成以下场景的测试：①**理想测试流程**，存储系统的最优性能（例如大文件读写），这组测试可以充分发挥存储厂商在性能调优上的主动性，测试包括：数据读写IO（IOR easy: write/read）、元数据操作（MDTest Easy: create/stat/delete）；②**严苛测试流程**（如随机读写3901 Byte数据），以及海量小文件的读写，以此来评估存储系统在极端场景下的性能底线。测试包括：数据读写IO（IOR Hard: write/read）、元数据操作（MDTest Hard: create/stat/read/delete）、Find索引。最终，通过IOPS和带宽的测试数据，计算出整个存储系统的性能评分。

存储性能评估标准

从三个方面来衡量性能评估的好坏：**成本、竞争力、可达性**，三者缺一不可。

以下案例中哪一个性能评估最好？

- OLTP业务场景，OceanStor Dorado 5000性能是300K IOPS，OceanStor Dorado 6000性能是480K IOPS；
- SSD硬盘时延<1ms，单盘性能为30K IOPS；
- 3.84TB硬盘价格比7.68TB硬盘价格便宜20%。

数据库OLTP业务，容量为50TB，客户要求性能达到300K IOPS@1ms，请配置一个存储方案。

方案一：
• 控制器：OceanStor Dorado 6000
• 硬盘：20块3.84TB SSD

方案二：
• 控制器：OceanStor Dorado 5000
• 硬盘：17块1.92TB SSD

方案三：
• 控制器：OceanStor Dorado 5000
• 硬盘：17块3.84TB SSD

方案四：
• 控制器：OceanStor Dorado 5000
• 硬盘：8块7.68TB SSD

方案一：性能和容量满足，成本偏高——高成本、高竞争力、高可达性

方案二：性能满足，容量不满足——低成本、高竞争力、低可达性

方案三：性能和容量满足，成本最优——低成本、高竞争力、高可达性👍

方案四：性能不满足，容量满足——低成本、低竞争力、低可达性

性能的影响因素

控制器的处理能力

包括内存、CPU、系统内部IO路径上的各种算法、前后端口等因素而决定的能力，即存储系统的处理能力。

盘的IO承载能力

后端硬盘能够提供的IO能力，随着盘的数量和类型的不同而不同。

性能的配置

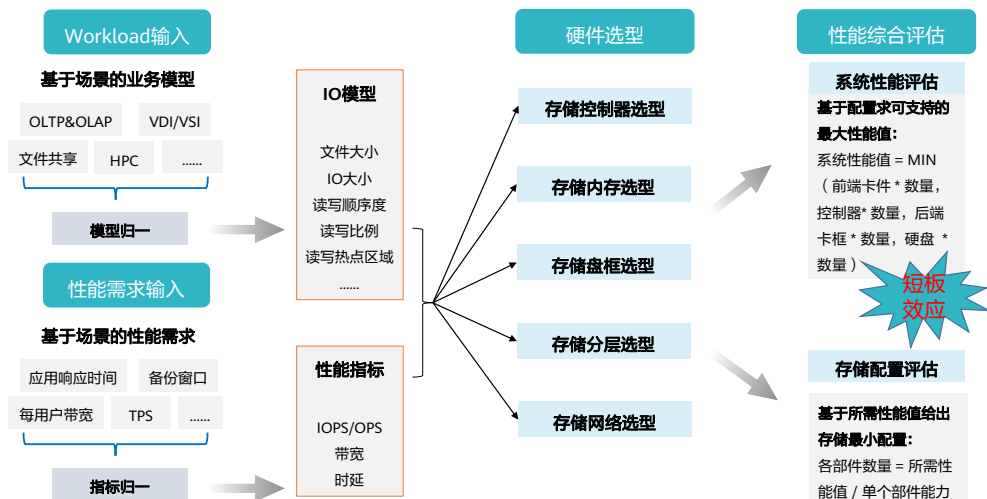
存储设备的相关软件配置和业务模型也会影响到性能。

增值特性配置

例如快照、远程复制、压缩、双活、SmartCache等特性，由于对IO路径上的处理流程产生了影响，导致了配置了上述特性之后，整体性能受到了影响。

特性名称	影响因素	说明
快照	COW（V3的快照实现方式）	COW模式对于写IO有影响，每个写IO在初次覆盖写需要Copy On Write。
远程复制	Syn	IO需要等待从端返回，时延会增大，IOPS降低。 IO同步同样会增加内部的IO流程。
	Async	基于快照技术实现，所以存在COW操作，原因同快照。
	synchronizing cycle	同步周期涉及到做快照的频率。
双活	本地优先模式	类似于同步远程复制，在阵列cache层面进行双写。
	负载均衡模式	在主机层面根据IO地址进行分别轮询下发到阵列，还是有阵列层面进行镜像。

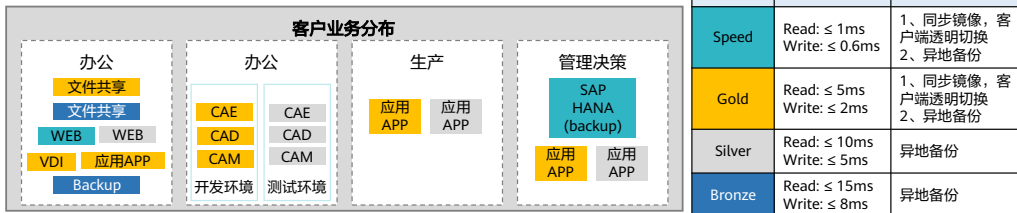
存储性能评估方法



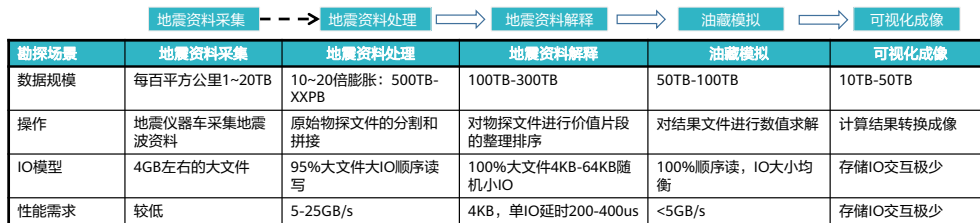
- 业务场景、性能需求和存储选型影响性能评估结果。

性能需求

同一客户在不同业务场景的性能需求不同：

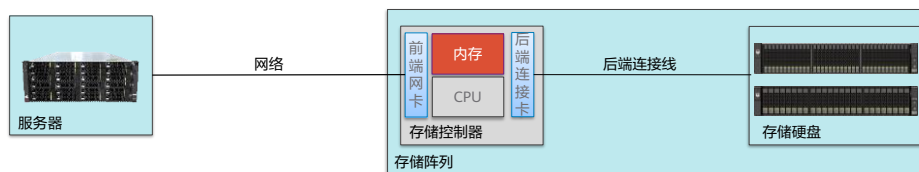


在同一个业务场景的不同业务阶段，性能需求不同：



- 由于性能属性是客户的主观感受，同时存储业务场景多样性，导致每个局点的性能需求不同，甚至同一局点不同业务的性能也不同。
- 客户层面的性能需求有很多，归一化处理后转换为带宽类、IOPS/OPS类、时延类三类性能指标。

存储选型



服务器和网络

服务器的性能决定了客户发出业务的压力大小，网络的吞吐量决定了存储阵列接收到的业务压力大小。

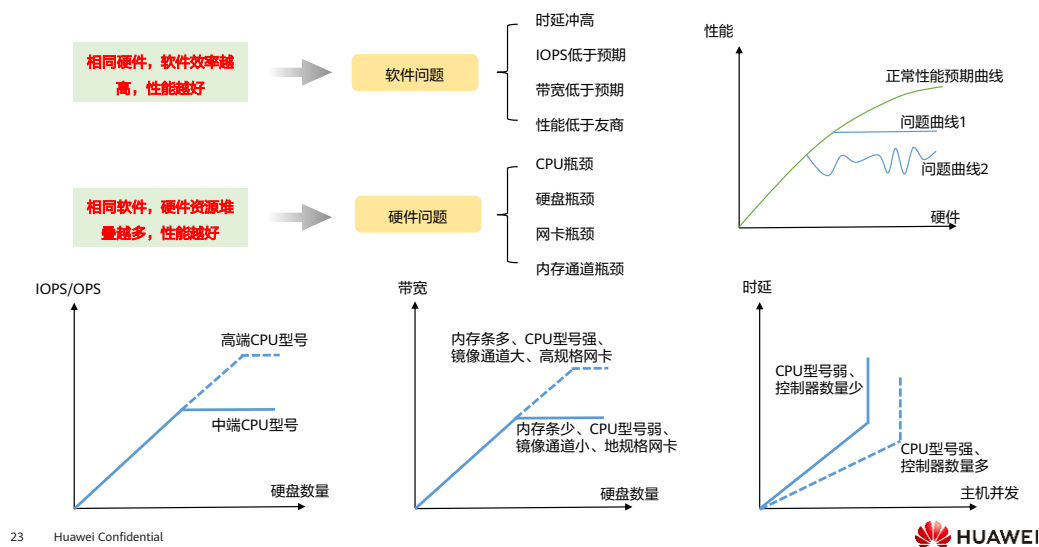
存储各组件对性能关系	性能影响因子
CPU：存储侧业务处理核心，性能三类指标均相关	CPU型号、CPU个数
内存：存储侧业务处理核心，性能三类指标均相关，同时决定了以下组件的数量上限	内存大小、内存条数
前端网卡：存储侧业务接收组件，带宽类和IOPS/OPS类指标相关	网卡型号、网卡数量
后端连接卡：存储控制器与硬盘连接组件，带宽类指标相关	卡件型号、卡件数量
后端连接线：存储控制器与硬盘连接通道，带宽类指标相关	线缆类型、线缆数量
硬盘框：存储硬盘承载组件，带宽类指标相关	硬盘框类型、硬盘框数量
硬盘：业务数据存储单元，性能三类指标均与硬盘相关	硬盘类型、硬盘数量

- 存储能承载多少业务性能，最终取决于存储软硬件选型。软件是存储性能的基础，硬件决定了存储性能的上限。
- 广义的存储是由服务器、网络（交换机）、存储阵列组成的复杂系统；狭义的存储即存储阵列，由前端网卡、存储控制器、后端硬盘连接卡、后端连接线、硬盘框和硬盘组成，性能高低是由这些组合因子共同决定。

目录

1. 存储性能概述
2. 存储性能评估
- 3. 存储性能问题**
4. 存储性能调优
5. 存储性能测试

性能问题概述

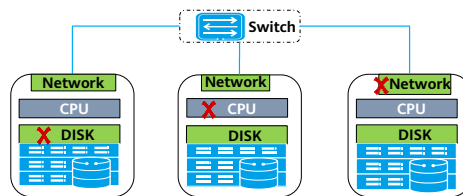


- IOPS/OPS随磁盘数量增加而增加，最终受限于CPU瓶颈，在一定拐点后不再增长。CPU型号越高，计算能力越强，最终达到拐点后的最大IOPS能力最强。
- 带宽随磁盘数量增加而增加，受限于内存条、控制器、CPU、网卡的内存通道、镜像通道、网卡通道等瓶颈，在一定拐点后不再增长。通道能力越强，最终达到拐点后的最大带宽能力越强。
- 时延随主机并发增加而增加，受限于CPU能力和控制器数量，在一定拐点后性能不再增长，但是时延会随业务压力增加继续增大。

性能问题定界

常见的四种系统性能瓶颈：

- **磁盘性能瓶颈：**例如只要不是在SSD全命中的场景，机械盘配置下瓶颈通常在于机械盘的随机能力。
- **CPU性能瓶颈：**CPU物理核瓶颈，CPU线程瓶颈等。
- **网络性能瓶颈：**网络通道的带宽、网络的时延、网络协议栈的IOPS处理能力都可能成为系统瓶颈。
- **内存性能瓶颈：**内存通道瓶颈。



主存盘的性能基线

硬盘类型	IOPS	带宽 (MB/s)
NL-SAS HDD	180	40
SAS HDD	280	70
SAS SSD	32000	500
NVMe SSD	100000	1500
Intel NVMe SSD	50000	700

- 整个IO流程包含三大步，即计算、网络、存储。在定位性能问题的时候我们不能一开始就直接怀疑存储侧软件的问题，而应该先将问题定界，然后再具体问题具体分析。
- 性能问题定界可以通过对比计算和存储的平均时延差异来大致判断是计算侧问题还是存储侧问题：
 - 1、如果计算侧时延与存储侧时延都很大且相差较小，说明问题可能出现在存储上（常见如达到磁盘性能上限、CPU瓶颈等）；
 - 2、如果计算侧观察到的时延比存储侧大很多，说明问题可能出现在计算上（计算节点CPU跑满、块设备内核参数限制）或者网络链路出现问题（如带宽达到瓶颈、交换机配置不当等）；
 - 3、确认是计算、存储、网络问题后再进行具体分析定位。

性能问题快速定界

收集信息

- 定位问题前，首先要了解清楚存储的配置，至少需要先收集到主存盘是SSD，还是HDD。

描述问题

- 了解清楚真正的问题点，如虚拟机卡顿，带宽低，时延高，是一直有这个问题，还是突发的，问题发生前是否有特别的操作等。

定界是否为存储问题

- 判断存储是否有问题，最直接的就是判断时延是否正常。存储时延在没有硬件瓶颈的情况下，主要取决于盘的类型、网卡类型和IO大小，而盘的类型对时延又有着较大的影响。

排除故障和硬件瓶颈

- 首先，排查有没有故障，有故障优先解决故障。其次，排查CPU、网络、盘和内存是否有瓶颈。

存储内部定界

- 在确认没有任何硬件瓶颈，且是存储时延大以后，可以通过FTDS日志进一步分析时延大的模块。

性能问题定位

 客户主观感觉业务变慢

 IOPS/OPS/带宽低

 时延明显升高

 性能波动

 硬件过载告警

资源问题

时延问题

压力问题



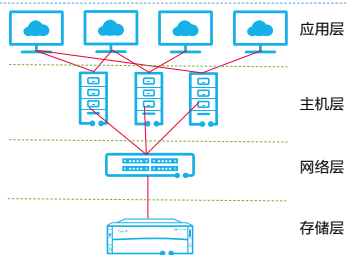
看配置: 服务器、网络、存储

看数据: 时延分布、资源利用率

系统拓扑

性能压力

异常提取



先定界

再定位

看时延

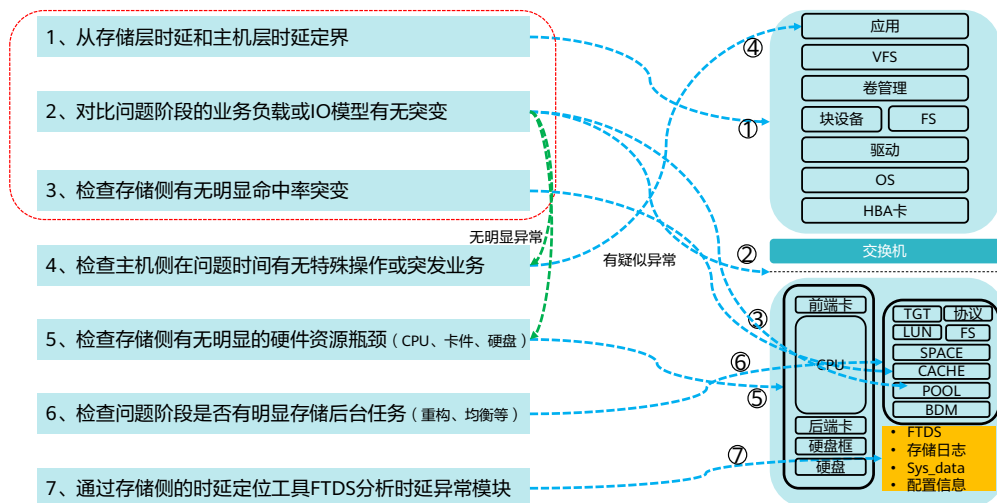
算并发

看资源

看告警

- 对于存储而言, 当客户反馈有性能问题时, 首先要判断是否为存储的问题。主机侧感知到应用卡顿或者时延大时, 可能存储本身并没有性能问题。问题有可能发生在主机侧, 如主机侧的应用、操作系统或者网络资源等。因此, 应该从上往下逐层定位。非存储的问题, 需要联系相关领域的专业人员来定位才更高效。

时延类问题



- 对于混合闪存存储，读IO优先尝试读缓存盘，如该IO未在缓存盘命中，则继续从主存盘读取。当读时延大时，写时延不高，优先怀疑时延来自读HDD盘。可以通过iostat观察HDD盘时延是否也较大，如果较大，说明业务前端的读时延来自读HDD盘上。如果所有节点盘上的时延都不高，可以进一步分析网络和CPU是否有瓶颈点。如果没有，则需要进一步通过FTDS分析时延来自哪个模块。
- 对于全闪存存储，由于SSD盘本身的时延都不高，大概率不会来自盘上。所以优先检查CPU和网络是否有瓶颈。为稳妥起见，还是要通过iostat观察盘的使用情况。
- FTDS：故障跟踪诊断系统。

性能低问题

1、检查存储侧时延有无明显问题

无问题

2、尝试增加或优化上层并发，或主机侧时延有无异常

有问题

3、着手分析存储侧时延异常的原因，尝试优化时延

4、检查存储侧有无明显的硬件资源瓶颈（CPU、卡件、硬盘）

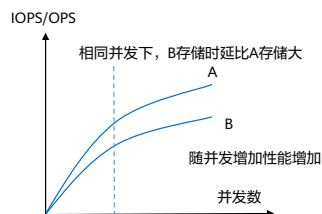
5、检查问题阶段是否有明显存储后台任务（重构、均衡等）

6、通过存储侧的时延定位工具FTDS分析时延异常模块

并发=IOPS*时延

增加并发

减少时延



性能波动问题

1、检查波动时间点是否有明显IO模型或业务负载变化

性能波动通常和业务变化有关

2、检查波动时间点是否有明显的命中率突变

3、检查波动时间点是否出现极小IO场景 (IOPS<100)

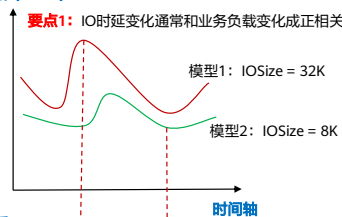
4、检查波动时间点存储侧有无明显的硬件资源瓶颈

5、检查波动时间点是否有明显存储后台任务 (重构、均衡等)

6、通过存储侧的时延定位工具FTDS分析时延波动模块

负载 (IOPS)

要点1: IO时延变化通常和业务负载变化成正相关



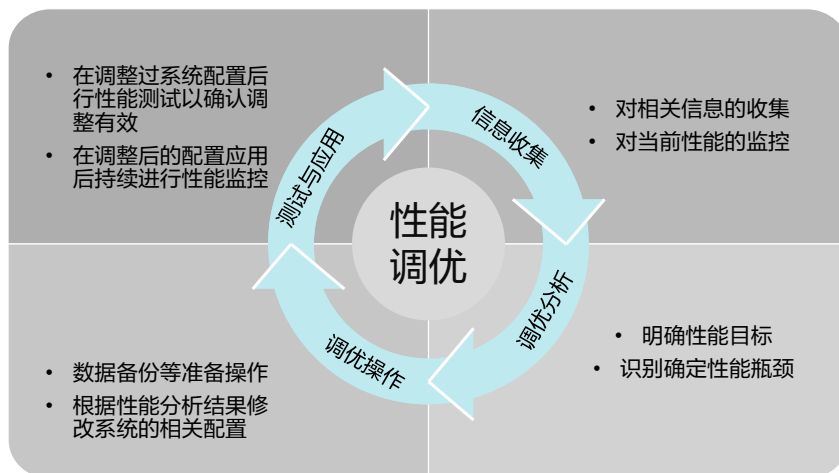
时延



目录

1. 存储性能概述
2. 存储性能评估
3. 存储性能问题
- 4. 存储性能调优**
5. 存储性能测试

性能调优流程



性能调优注意事项

迭代渐进

- 调优过程是迭代渐进的过程
- 每一次调优的结果都要反馈到后续的系统运行中
- 运行中的性能指标也应该始终被监控

设计为主

- 系统的性能更大程度上取决于良好的设计
- 在系统的设计开发过程中，应始终把性能放在考虑的范围内
- 调优技巧只是辅助手段

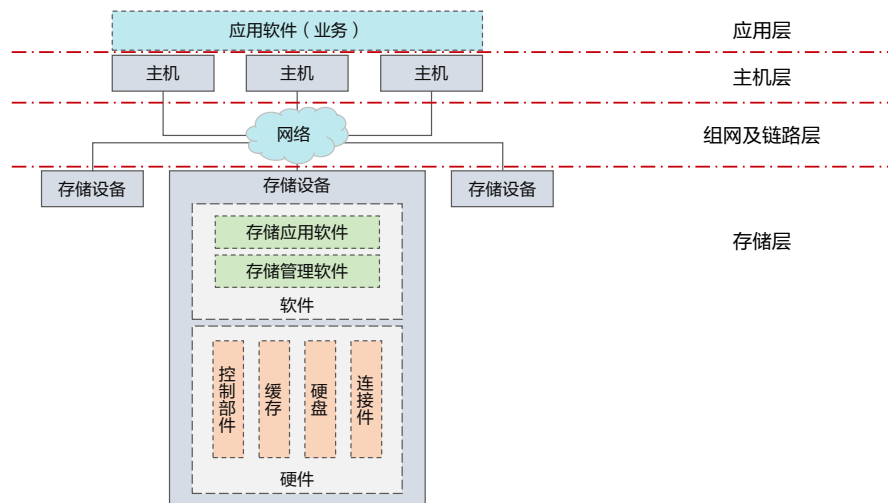
多方排查

- 性能是一个全流程相关的工作
- 问题可能出现在整个I/O路径的各个环节
- 系统的性能是由系统中出现短板的环节决定的
- 因此应该从多个方面逐一排查，找出整个系统的性能瓶颈所在并针对性地解决

关键与底线

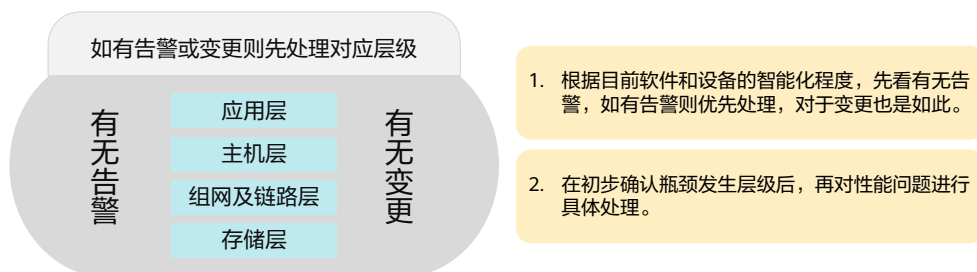
- 关键：确定清晰明确的性能目标是性能调优的关键
- 底线：必须保证调优后的系统正常运行

影响业务层级

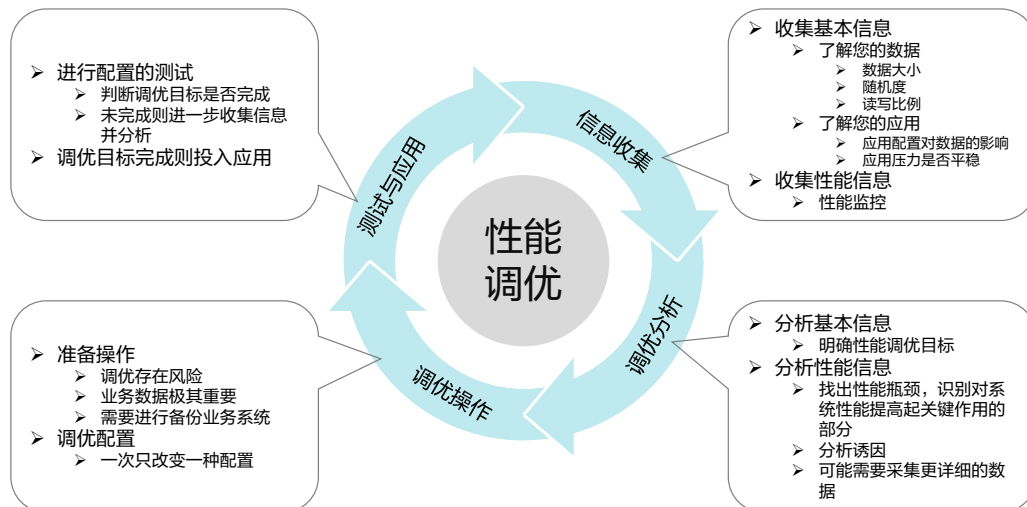


存储系统性能调优思路

- 目前无论应用、主机、网络、存储层上的相关软件和设备都有比较智能化的告警系统，如发现告警则优先处理（参照故障处理章节的胶片），处理后大概率的性能影响会消失。



存储性能调优流程

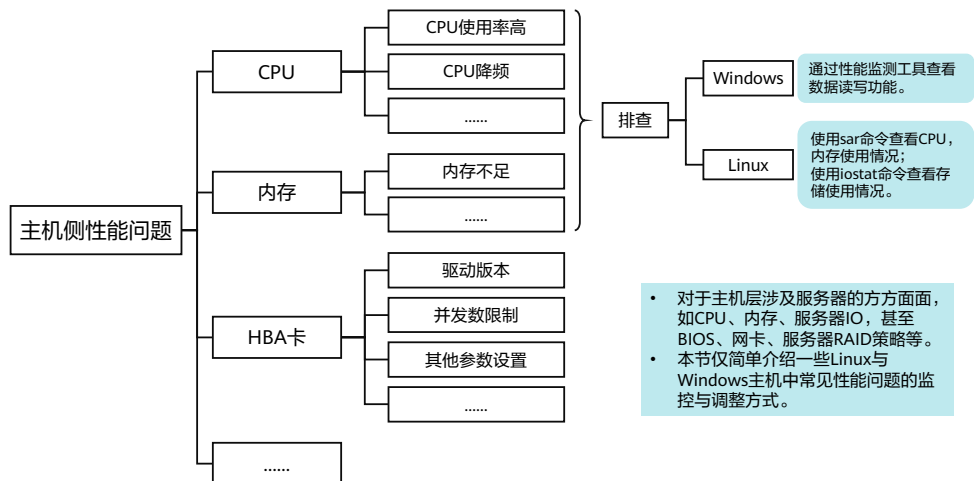


目录

4. 存储性能调优

- 主机层调优
- 网络层调优
- 存储层调优

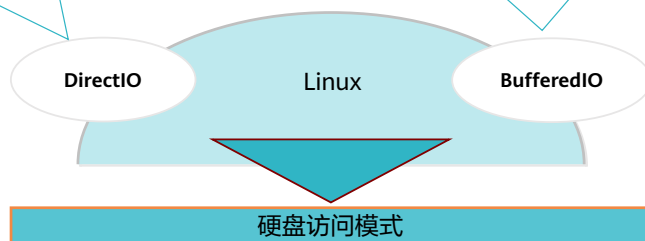
主机层影响因素



Linux硬盘访问模式

- 当进行常见的性能测试时，在Linux操作系统下，建议指定硬盘的访问模式为DirectIO。
- 通过运行dd命令将IO方式设置为直接IO（一般是在打开块设备的时候设置一个O_DIRECT标记）：
`dd if=/dev/sds of=/dev/null bs=256K iflag=direct。`

- 当使用Vdbench工具或dd命令等进行性能测试时的默认值。



Linux块设备参数调优 - 队列深度

Linux系统默认
值为128

队列深度

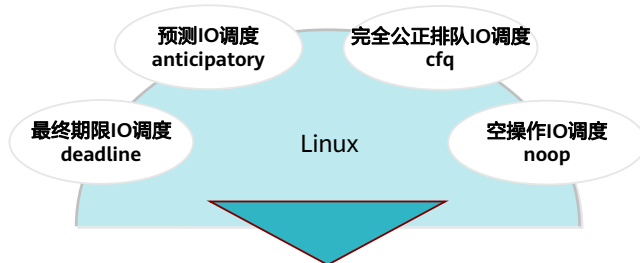
决定

块设备写I/O的最大并发数

一般情况下不建议用户修改此参数。

当对系统进行极限性能测试时，为了增大主机写I/O的压力及I/O在队列中被合并的概率，可以适当增大此参数。

Linux块设备参数调优 - 调度算法

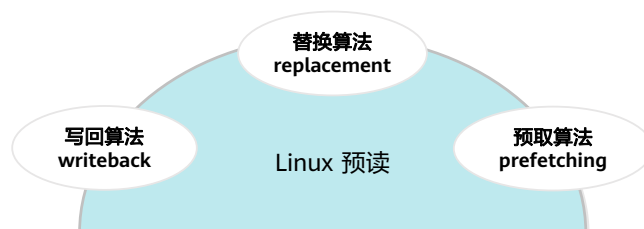


调度算法配置不当会对系统性能产生影响，如I/O并发等。临时修改调度算法可以用于性能调优，当应用服务器重启后，调度算法将恢复为默认值。

- IO调度器算法的选择，既取决于硬件特征，也取决于应用场景。
- 在传统的SAS盘上，CFQ、DEADLINE、ANTICIPATORY都是不错的选择；对于专属的数据库服务器，DEADLINE的吞吐量和响应时间都表现良好。然而在新兴的固态硬盘比如SSD、Fusion IO上，最简单的NOOP反而可能是最好的算法，因为其他三个算法的优化是基于缩短寻道时间的，而固态硬盘没有所谓的寻道时间且IO响应时间非常短。
- 假设我们要对sda进行操作，如下所示：
 - `cat /sys/block/sda/queue/scheduler`
 - `echo "cfq" > /sys/block/sda/queue/scheduler`

- IO调度器（IO Scheduler）是操作系统用来决定块设备上IO操作提交顺序的方法。存在的目的有两个，一是提高IO吞吐量，二是降低IO响应时间。然而IO吞吐量和IO响应时间往往是矛盾的，为了尽量平衡这两者，IO调度器提供了多种调度算法来适应不同的IO请求场景。
- NOOP算法的全写为No Operation。该算法实现了最简单的FIFO队列，所有IO请求大致按照先来后到的顺序进行操作。之所以说“大致”，原因是NOOP在FIFO的基础上还做了相邻IO请求的合并，并不是完全按照先进先出的规则满足IO请求。
- CFQ算法的全写为Completely Fair Queuing。该算法的特点是按照IO请求的地址进行排序，而不是按照先来后到的顺序来进行响应。
- DEADLINE在CFQ的基础上，解决了IO请求饿死的极端情况。除了CFQ本身具有的IO排序队列之外，DEADLINE额外分别为读IO和写IO提供了FIFO队列。读FIFO队列的最大等待时间为500ms，写FIFO队列的最大等待时间为5s。FIFO队列内的IO请求优先级要比CFQ队列中的高，而读FIFO队列的优先级又比写FIFO队列的优先级高。优先级可以表示如下：FIFO(Read) > FIFO(Write) > CFQ。
- CFQ和DEADLINE考虑的焦点在于满足零散IO请求上。对于连续的IO请求，比如顺序读，并没有做优化。为了满足随机IO和顺序IO混合的场景，Linux还支持ANTICIPATORY调度算法。ANTICIPATORY是在DEADLINE的基础上，为每个读IO都设置了6ms的等待时间窗口。如果在这6ms内OS收到了相邻位置的读IO请求，就可以立即满足。

Linux块设备参数调优 - 预读功能



当业务需要对大量的大文件进行读取时，可以通过适当调大预读量提升性能。

CPU性能问题

性能问题

- 系统访问卡顿，QPS、TPS降低，响应时间延长，网络吞吐量降低；
- 应用服务器内存和IO正常，CPU利用率增高。

性能分析

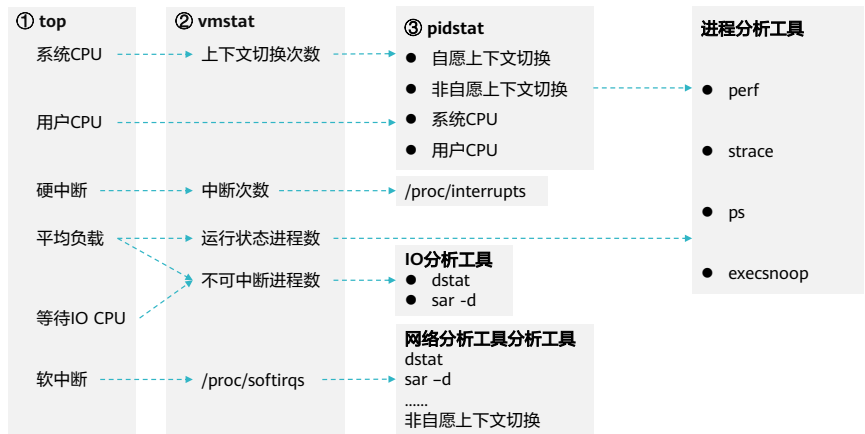
- 线程太多，上下文切换太频繁；
- GC回收使用了过高的CPU资源；
- 某段代码陷入了死循环；
- 锁争用激烈。

性能调优

- 考虑使用更高级的CPU代替当前的CPU；
- 对于多CPU，考虑CPU之间的负载分配；
- 考虑在其他体系上设计系统，例如增加前置机，设置并行服务器。

CPU性能工具

在实际环境中，通常都希望尽可能快地定位系统的瓶颈，然后尽可能快地优化性能。为了缩小排查范围，通常会先运行几个支持指标较多的工具，如top、vmstat、pidstat和perf，并用虚线表示关系，对应出性能分析的下一步方向。



内存性能问题

性能问题

- 项目内存持续增加；响应时间成规律性的先增加后回落；
- 查看应用日志，会出现OutOfMemoryError错误；GC日志发出FULL GC警告；
- 系统长时间运行后出现访问错误或宕机。

性能分析

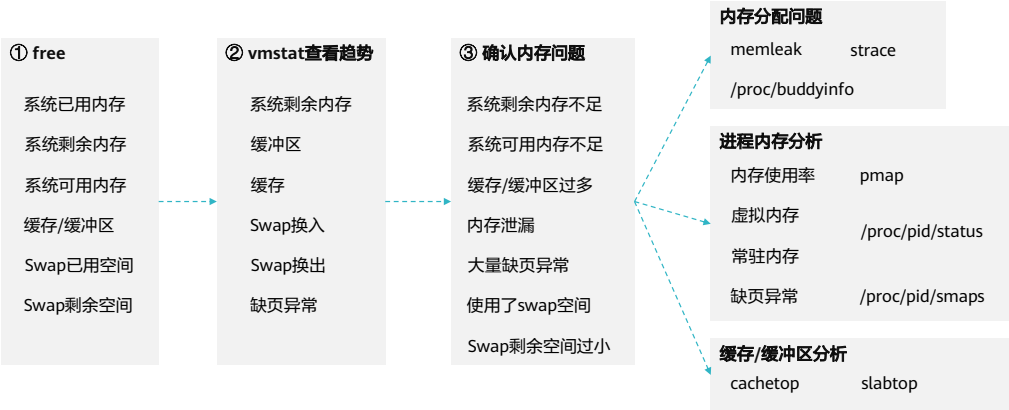
- 启动参数内存值设定得过小；
- 代码中存在死循环或者循环产生过多重复的对象实体；
- 集合类中有对对象的引用，使用完后未清空，使JVM不能回收。

性能调优

- 保证保留列表具有较大的邻接内存块；
- 调整数据块缓冲区大小（用数据块的个数表示）；
- 将最频繁使用的数据保存在存储区中。

内存性能工具

为了迅速定位内存问题，通常会先运行几个覆盖面比较大的性能工具，比如free、top、vmstat、pidstat等，具体的分析思路如下：



磁盘IO性能问题

性能问题

- 磁盘读写速率、IOPS过高，系统出现卡顿。

性能分析

- SQL写法、参数配置不合理；
- 交换机故障，网线老化；
- 存储阵列带宽不足，缓存不足，Qos限制，RAID级别设置不当。

性能调优

- 通过把日志和数据库对象分布在独立的设备上；
- 把不同的数据库放在不同的硬盘上。

磁盘性能工具

在性能分析过程中，通常会先运行那几个支持指标较多的工具，如iostat、vmstat、pidstat等。然后再根据观察到的现象，结合系统和应用程序的原理，寻找下一步的分析方向。

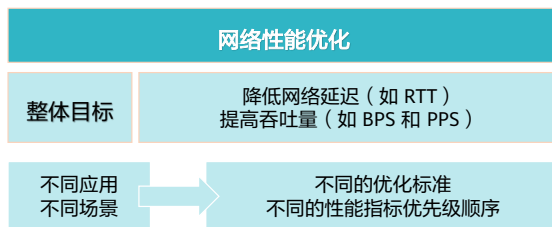
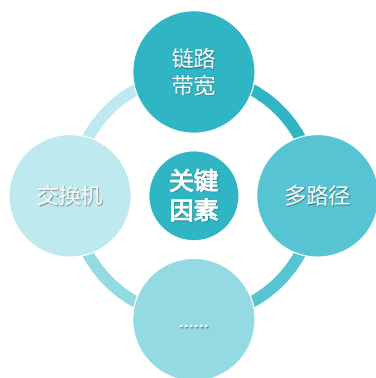


目录

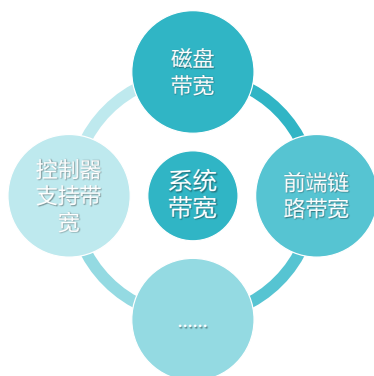
4. 存储性能调优

- 主机层调优
- **网络层调优**
- 存储层调优

网络层调优



链路带宽瓶颈



带宽性能主要针对的是大IO

影响带宽性能的因素较多

带宽受制于组网中的短板

前端带宽

$$= \text{单端口带宽} \times \text{端口数量}$$

硬盘提供的有效读带宽

$$= \text{单盘读带宽} \times \text{硬盘数量}$$

硬盘提供的有效写带宽

$$= \text{单盘写带宽} \times \text{硬盘数量} \times (\text{数据盘数量} / \text{总盘数})$$

可提供的带宽

$$= \text{MIN}(\text{硬盘提供的有效带宽}, \text{前端带宽})$$

交换机端口

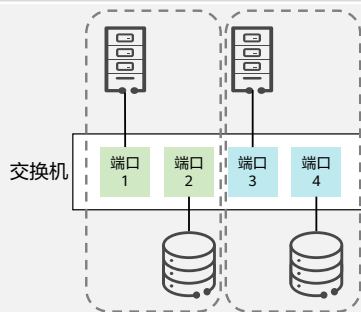
FC交换机 - Zone

IP交换机 - VLAN

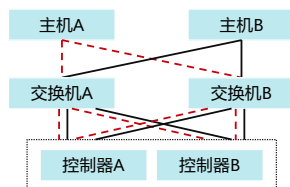
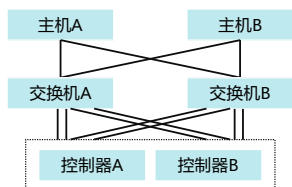
避免各业务系统相互影响

改善IT基础架构的灵活性

- 在双交换组网的前提下，若交换机上有多种业务，最小化原则可以帮助保障性能：每个Zone仅包含一个主机端口和一个存储端口。
- 对于IP交换机的VLAN划分也遵循一个VLAN包含一个存储端口、一个主机侧端口的最小化原则。
- 虽然使用VLAN或Zone能够隔离不同种类的业务，但VLAN的数量不宜过多。当业务项目过多时，推荐按照宏观的业务类别进行划分。



案例 - FC交换机Zone划分



问题描述

- 某局点有两台主机，采用交叉组网分别连接到两台交换机，每台交换机到控制器A和控制器B各有两根FC光纤连接，未划分Zone。
- 最大读带宽远小于当前组网环境下的理论带宽。
- 原因：两台交换机到存储的8条物理路径被两台主机共享，竞争链路带宽。

调优

- 划分Zone对共享的链路进行分组隔离，每台主机各固定使用交换机到控制器的四条链路，保证链路上没有多台主机的I/O竞争。
- 除了划分Zone之外，还可以通过多路径软件对路径进行禁用、绑定LUN组和端口组的方式实现共享链路的分组隔离。

HBA卡

HBA卡性能

并发数限制

当HBA卡并发数不够会导致主机压力不能完全传递到存储上，出现主机侧时延和存储侧时延相差较大时。

用户可以通过HBA卡厂商提供的管理软件查询当前HBA卡的并发数，并对其进行适当的修改。

驱动问题

若主机HBA卡驱动版本过低，可能会造成大I/O被HBA卡拆分、默认并发不足、I/O时延增大等现象，特别是对于低负载、全命中和带宽型的场景。

建议将主机HBA卡驱动更新到最新驱动版本。

网络性能工具

从网络性能指标出发，更容易把性能工具同系统工作原理关联起来，对性能问题有宏观的认识和把握。这样，想查看某个性能指标时，就能清楚知道，可以用哪些工具。

性能指标	工具	说明
吞吐量（BPS）	sar/nethogs/iftop	分别可以查看网络接口、进程以及IP地址的网络吞吐量
PPS	sar	查看网络接口的PPS
连接数	netstat	查看网络连接数
延迟	ping	通过ICMP、TCP等测试网络延迟
连接跟踪数	conntrack	查看和管理连接跟踪状况
路由	route/traceroute	查看路由并测试链路信息
防火墙和NAT	Iptables	配置和管理防火墙及NAT规则
网卡功能	ethtool	查看和配置网络接口功能
抓包	tcpdump/wireshark	抓包分析网络流量
内核协议栈跟踪	bcc/systemtap	动态跟踪内核协议栈的行为
吞吐量压测	iperf/ib_write_bw/ib_read_bw	压力测试网络流量、时延等

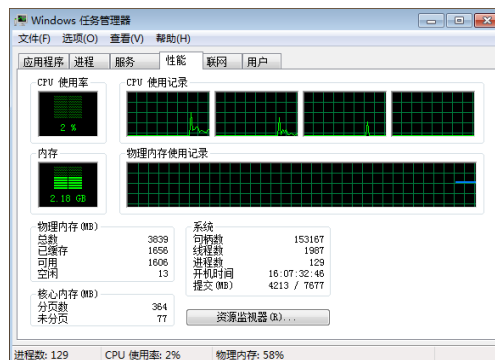
网络性能参数调整

- 所有的 TCP/IP 调优参数都位于 /proc/sys/net/ 目录，以下列举了一些调优参数及其含义。

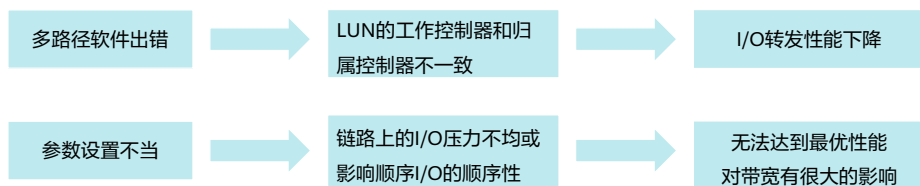
参数	含义
/proc/sys/net/core/rmem_max	最大的 TCP 数据接收缓冲区大小。
/proc/sys/net/core/wmem_max	最大的 TCP 数据发送缓冲区大小。
/proc/sys/net/ipv4/inet_peer_gc_mintime	每一次碎片收集（garbage collection）之间的最小时间间隔。内存压力较大时可调整该参数。
/proc/sys/net/ipv4/tcp_timestamps	时间戳（开启与否）。一个1Gbps的链路肯定会遇到以前用过的序列号，时间戳能够让内核接受这种“异常”的数据包。
/proc/sys/net/ipv4/tcp_window_scaling	支持更大的 TCP 窗口，在网速要求较高的网络中可调整此参数。

Windows性能监控方法

- 查看主机内存和CPU使用情况
 - 自带的性能监控器（Windows Performance Monitor）
 - 任务管理器查看
- 查看主机是否降频
 - DirectX诊断工具（开始>运行>DXDIAG）
- 查看HBA卡驱动
 - Server Manager

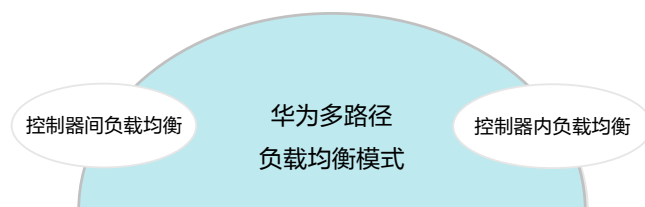


多路径软件



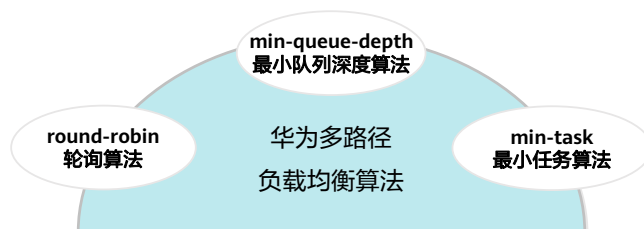
多路径软件保障了主机和存储间链路的冗余可靠和高性能，建议配套使用华为的UltraPath多路径软件。
大多数场景下，推荐使用华为多路径的默认配置。

多路径软件 - 负载均衡模式



典型场景	推荐配置
当主机和存储系统之间的传输路径成为性能瓶颈时	控制器间负载均衡
其他场景	控制器内负载均衡（默认配置）

多路径软件 - 负载均衡算法



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
主机下发的业务I/O模型数据块较大	min-task
其他场景	min-queue-depth (默认配置)

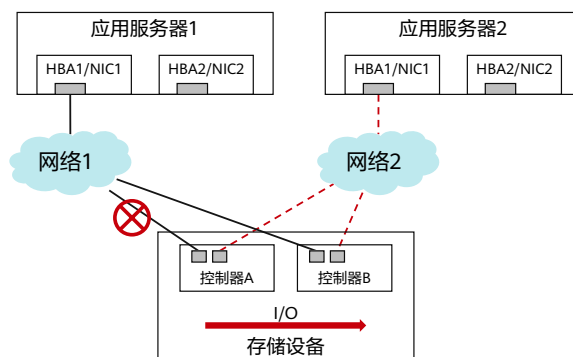
乒乓效应

LUN的工作控制器来回切换
(乒乓效应)

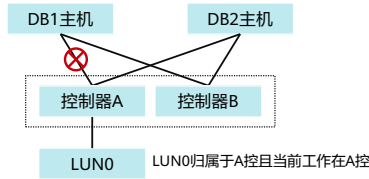
LUN的访问性能下降

应用服务器上容易出现
I/O超时等问题

当存储出现乒乓效应时，应将多路径软件的LUN自动切换功能关闭，并请尽快恢复断开的链路，保证每个节点到存储系统的每个控制器链路连接正常。



案例 - 多路径乒乓效应导致数据库无法访问



- DB: Database, 数据库。
- 案例延伸：在应用中，应尽量避免将一个LUN同时映射给两台及以上应用服务器。如果在特定应用场景要将LUN同时映射给两台应用服务器，则要在应用服务器上安装集群软件并配制集群预留。

问题描述	调优
• 某局点数据库业务中，双机通过多路径使用存储资源，多路径版本为UltraPath for Windows V100R002C01。 • 存储设备上有一个LUN0映射给双机使用，但数据库双机中的一台应用服务器无法访问。 • 原因：多路径乒乓效应导致存储设备反复来回切换LUN的工作控制器，导致数据库无法访问。	• 在命令行下运行upadm show option命令查看Failover功能是否关闭；若未关闭，则运行upadm set failover=off命令将Failover功能关闭。 • 运行upadm start updateimage命令更新多路径配置。 • 运行upadm show option命令查看多路径配置确认多路径功能已经关闭。

目录

4. 存储性能调优

- 主机层调优
- 网络层调优
- **存储层调优**

收集基本信息

- 基本信息的收集能够存储侧性能问题优化定位提供依据，帮助明确调优性能目标，是规划监控参数提供基础。需要收集的信息主要包括当前存储系统所承载的业务类型及I/O特征以及存储资源规划信息。

业务类型



I/O特征

产品型号规格与软件版本

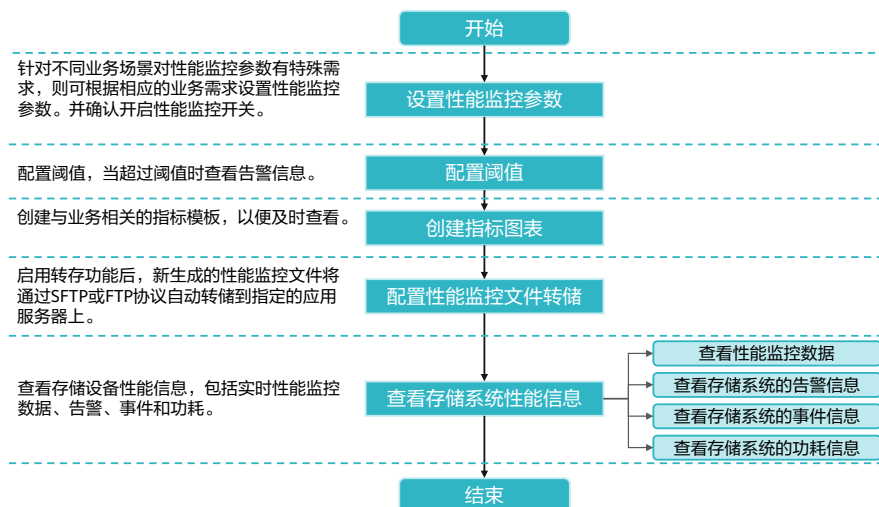
前后端接口模块数量与连接的数量

存储资源规划

LUN属性

增值功能配置

性能监控流程



- 在运用DeviceManager管理软件对存储系统进行性能监控之前，首先应该了解性能监控的流程，以便能够准确的对存储侧进行日常的性能监控。

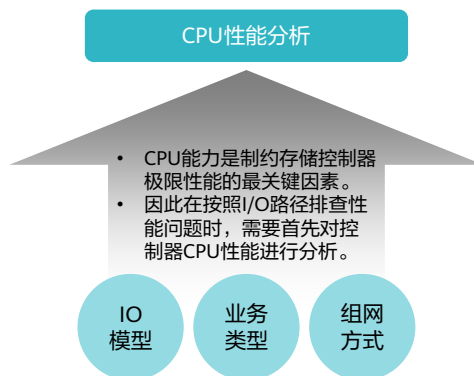
内部事务及增值特性

- 正如前文提过，在优化性能时首先需要排查的就是存储本身所产生的告警和变更。如硬盘、BBU、电源、端口模块、特性等均及可能对性能造成影响，这些内部事务或特性的告警与变更都需要首先进行排查。
- 如果性能有问题，需首先确认是否为内部事务或增值特性影响导致。



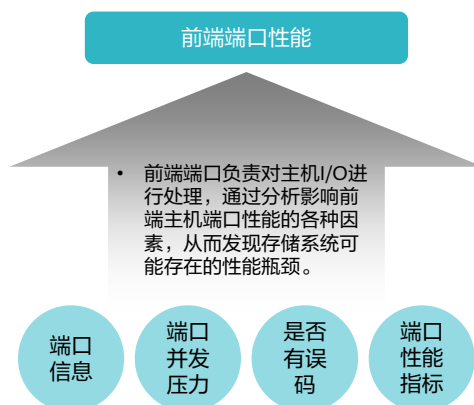
CPU性能分析

- 存储系统CPU占用率高会导致系统调度时延增大，造成I/O延时增大。
- 可以通过DME性能分析工具以及CLI命令查看当前控制器的CPU利用率。
- 存储系统CPU消耗与I/O模型、业务类型和组网方式有紧密的联系：
 - 写I/O比读I/O消耗CPU更多
 - 随机I/O比顺序I/O消耗CPU更多
 - IOPS型业务比带宽型业务消耗CPU更多
 - iSCSI组网比FC组网消耗CPU更多



前端端口性能分析

- 在分析前端端口性能之前，需要确认接口模块的位置、端口的连接数量、工作状态、速率等信息。
- 前端端口主要关注的性能指标包括：平均读I/O响应时间、平均写I/O响应时间、平均I/O大小、IOPS、带宽。



Cache性能分析

- Cache主要用来缓存主机数据，使主机在进行数据读写时不用每次都对硬盘进行操作，从而提高系统的读写性能。



- Cache是提升性能、改善用户体验的关键模块，有关Cache的问题多数都跟存储故障告警有关，回顾之前所讲，优先处理告警，多数存储侧性能问题将迎刃而解。

Cache预取策略

- 在有主机读请求下发时，Cache会先将磁盘中数据读到内存中，并启动预取，根据数据的连续性将磁盘中后续可能会被主机读到的数据预取到Cache中，增加后续主机请求的读命中概率。



预取浪费 预取的数据量远大于主机实际读取到的数据量。

预取不足 对顺序I/O流没有提前在盘上做足量的预取，导致所有I/O都必须直接写入硬盘读取数据，而无法提前在Cache中命中返回。

Cache写策略

- Cache的写策略默认设置为回写，在业务运行过程中，当出现电池故障、系统单控、高温转透写、LUN Fault页面超标等原因时可能会触发LUN转透写。

回写

透写

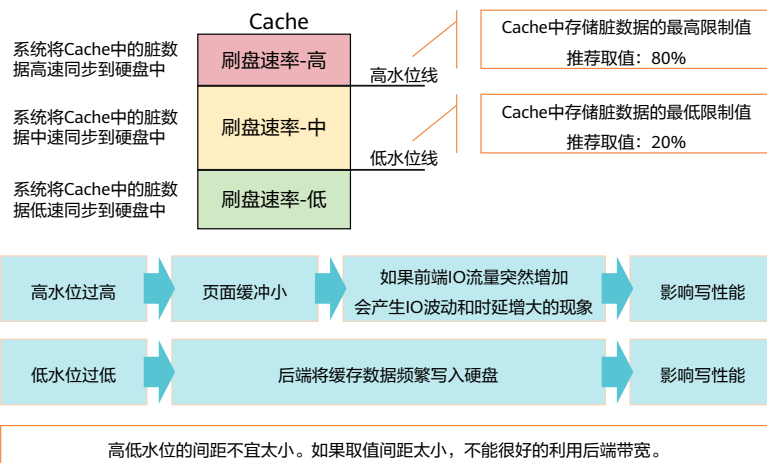
机械盘：回写 > 透写

由于写入Cache的速度比机械盘读写快很多

SSD盘：回写 ? 透写

回写对比透写不一定有性能优势
取决于算法的具体实现

高低水位



LUN性能分析

- 通过对LUN的性能分析，了解LUN的类型及LUN的本端访问对性能的影响，从而发现存储系统可能存在的性能瓶颈。
- 在使用SmartThin特性时，Thin LUN与Thick LUN在读写性能上存在一定的差异。

LUN类型	写性能		读性能	
	首次写	覆盖写	顺序读	随机读
Thick LUN	Thick LUN在LUN创建完成后就进行格式化。首次写入只涉及主机I/O的写操作，性能优于Thin LUN。	性能相当	Thick LUN在创建时，系统通过自动资源配置技术从存储资源中一次性为其分配存储资源，最大限度上保证了其映射到磁盘上的空间连续性。对于机械磁盘顺序读取效率更高。	性能相当
Thin LUN	Thin LUN在新数据写入时，同时会进行空间的分配，从而会带来较多元数据I/O的读写操作，写入流程更长且会对后端硬盘造成额外的压力。		Thin LUN是按照用户的实际需求动态分配存储资源，空间分配在时间上存在不连续性，因此会影响到其映射到磁盘上的空间连续性。	

RAID性能分析

- 对于不同的I/O模型，不同的RAID级别在读写性能上有不同的表现。
- RAID对系统性能的影响主要体现在RAID级别和分条深度。
- 当然，对于传统RAID，RAID组中成员盘的个数也会影响到系统性能。

随机读

- 各种RAID级别性能相当

随机写

- 由于RAID算法计算校验和镜像的处理，对于随机写，都会带来额外的I/O开销，被称为写惩罚（Write Penalty）。
- RAID算法的写惩罚越大，则随机写性能越差。
- 因此RAID10随机写性能优于RAID5，而RAID5优于RAID6。

顺序读

- 在RAID2.0+的架构下，没有单独的校验盘，硬盘域中的所有硬盘都可以提供读性能。
- 因此RAID5、RAID6的顺序读性能较好，而RAID10由于镜像的原因性能比RAID5、RAID6稍低。

顺序写

- 顺序写的场景下，写入硬盘时基本能够做到满分条下盘，则此时写额外的开销在于写校验位（RAID5、6）和镜像双写（RAID10）。
- 因此顺序写情况下，校验位占总位数比例越大，性能越差。

分条深度

- 分条深度决定了主机I/O按照RAID算法处理后写入硬盘最大I/O的大小。
- 对于不同特征的I/O，分条深度对性能会造成一定的影响。

属性	小I/O		大I/O
	随机	顺序	
大小	一般指的是小于16KB大小的I/O		一般指大于256KB大小的I/O
描述	由于I/O的随机性，在Cache中得不到聚合，写入硬盘时基本保持原始大小。因此当分条深度较大达到I/O大小的若干倍时（当前默认为128KB）时，小I/O跨分条单元的几率较低，即I/O被拆分的几率较低。	由于I/O的顺序性，多个小I/O在Cache中会聚合成大I/O，在理想情况下会聚成分条深度大小的I/O写入硬盘。因此，较大的分条深度有利于提升写入硬盘的效率、减少写入硬盘的I/O数量。	此时如果选择小于I/O大小的分条深度，会导致在RAID层I/O被拆分，从而影响效率。
分条深度建议	受分条深度的影响较小 建议保持默认分条深度128KB		建议选择最大分条深度512KB

后端端口性能分析

- 后端端口指的是后端SAS端口，负责连接控制框和硬盘框，是数据从硬盘读写的通道。后端SAS端口会对性能造成一定程度的影响，主要体现在硬盘框环路连接上。

由于单个端口的带宽有限，因此应当确保环路所连接的端口支持的带宽大于环路上连接的硬盘框上所有硬盘的总带宽。

随着环路上硬盘框数增多，级联链路上所消耗的时延也越大，影响后端I/O的时延，从而影响IOPS。

在具有足够的SAS端口的情况下，建议：

硬盘框均匀分布在多个环路上。

单个环路级联的硬盘框数小于2。

硬盘框间要尽量保证对称插盘。

单控上有多个后端接口模块的情况下，建议：

环路尽量分布在多个后端接口模块上。

硬盘性能分析

- 海量的数据读写对磁盘访问造成巨大压力，单盘的性能是也影响整个系统性能的重要因素。

硬盘类型

- 不同存储介质在存储成本和性能方面存在较大差异。
- 因此，用户在进行存储规划时，必须根据业务压力和I/O特性来选择硬盘类型。

优先排查

- 当出现性能问题排查存储前端无异常时，应优先排查硬盘是否达到性能上限。
- 当硬盘达到性能极限时（硬盘利用率接近100%），存储性能受制于后端硬盘性能，IOPS和带宽性能无法提升。

硬盘利用率

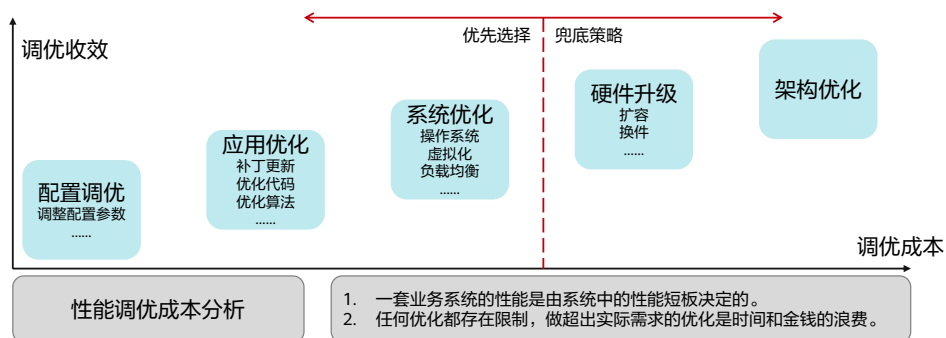
- 为保证硬盘可靠性和使用寿命，建议保持硬盘的利用率低于70%为最佳。
- 当出现硬盘域中大部分硬盘的利用率均大于90%时，建议对硬盘域加盘扩容或将业务迁移到性能更优的硬盘上。

工具推荐

- 可以通过SystemReporter工具查看硬盘利用率。

调优操作

- 调优操作的第一步是进行数据备份。
- 调优的具体配置涉及各类应用场景，需具体情况具体分析，根据分析结果，找出当前系统的性能瓶颈，做出相对应的调整。



目录

1. 存储性能概述
2. 存储性能评估
3. 存储性能问题
4. 存储性能调优
- 5. 存储性能测试**

性能测试概述

- 性能测试开展之前需要做许多准备工作，其中性能需求分析是整个性能测试工作的基础。收集客户需求，通过历史业务数据和性能指标调查，明确具体的性能指标，然后选择合适的工具进行性能测试，帮忙快速定位性能问题。
- 性能测试在不同情况下是有着不同的目的，清楚每次性能测试的目标，能够让我们在开展性能测试的时候更有针对性。

性能指标

这是最常见的性能测试目的，常规性能测试通常都有这个需求。在正常压力（无压力）流量下对系统进行性能测试，建立一个可度量的参考标准，为其他测试场景或者调优过程提供对比参考。

性能瓶颈

系统容量评估因为比较难自动化，所以通常这类需求频率会低一些，可能会周期性评估。通过“拐点法”进行性能测试，逐步增大系统的压力，直到性能指标不可接受或者出现了明显的拐点，从而找到系统的瓶颈点。

测试系统稳定性

系统稳定性压测是要知道在各种情况下系统能否稳定运行。

系统参数调整

需要配合研发，测试在不同系统参数或者部署方案下，系统性能情况的对比。

系统性能回归

在版本升级时，通常需要对比新老版本的性能情况。

性能测试需求分析

性能测试需求收集的途径/方法

- **客户需求**: 客户如果对自身系统非常了解, 并且提出明确的性能需求, 或者希望系统能负载多大的压力的时候, 我们可以采用客户的性能需求。
- **历史业务数据**: 通过产品经理或者运营收集以往公司的业务情况, 如网站PV、UV, 每星期/每月的峰值业务量, 每天业务交易量最多的时间段, 用户的增长速度, 用户对系统的哪些模块使用最频繁等。
- **性能指标的调查**: 对系统性能进行调研, 收集系统各方面数据, 如TPS、平均响应时间、90%/95%响应时间、CPU、内存、磁盘、网卡等。
- **参考同行项目经验**: 如果公司没有类似的项目经验, 我们可以参考同行业的相关项目需求, 借鉴他们成熟的需求考虑自己项目会哪些情况。在借鉴的同时需要结合自身产品情况, 不能完全照搬。

性能测试点的选取

- **核心业务**: 以公司主要的业务为第一核心, 无论如何首先要优先满足这些业务的性能测试。如何确定哪些业务是关键业务, 看产品的定性与定调, 如果电商行业、金融交易行业, 主要的业务就是与交易挂钩, 如下单、付款等接口。
- **高访问量的功能点**: 根据项目各个功能点的访问量, 可以以小时、天、周、月为统计单位, 如果访问量高, 表示该功能点承受压力也就大, 同时它又是关键业务, 那么基本可以确定该功能点为测试点。
- **业务逻辑复杂度高**: 业务复杂度高的功能点通常是核心业务且请求量高。如果一个业务的访问量不高, 业务逻辑却很复杂, 也是需要性能测试的。因为往往出错的可能性会很高, 严重影响系统的正常运行。

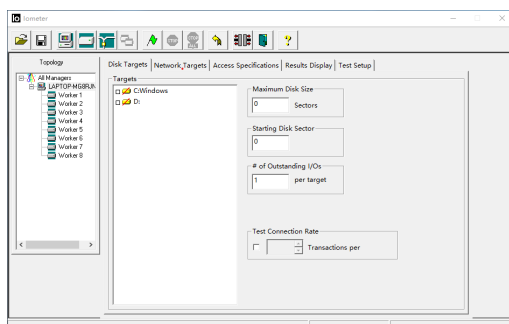
性能测试工具

性能测试工具

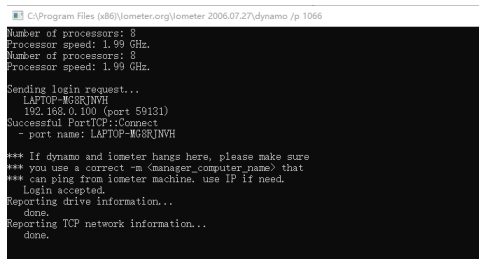
- **Iometer**: 一个免费的开源的测试磁盘性能的工具。和其他磁盘工具相比，可以测试I/O的传输率和平均的I/O响应时间。
- **Vdbench**: 业界认可和应用广泛的SAN和NAS存储性能测试工具。
- **Fio**: 开源压力测试工具，主要是用来测试磁盘io性能。
- **iozone**: 一个文件系统的benchmark工具，可以测试不同的操作系统中文件系统的读写性能。
- **Postmark**: 是NetApp开发的针对文件系统在邮件系统或电子商务系统中性能，这类应用的特点是：需要频繁、大量地存取小文件。
- **IOR**: Interleaved or Random是一种常用的文件系统基准测试程序，特别适用于评估并行文件系统的性能。

Iometer

- GUI界面

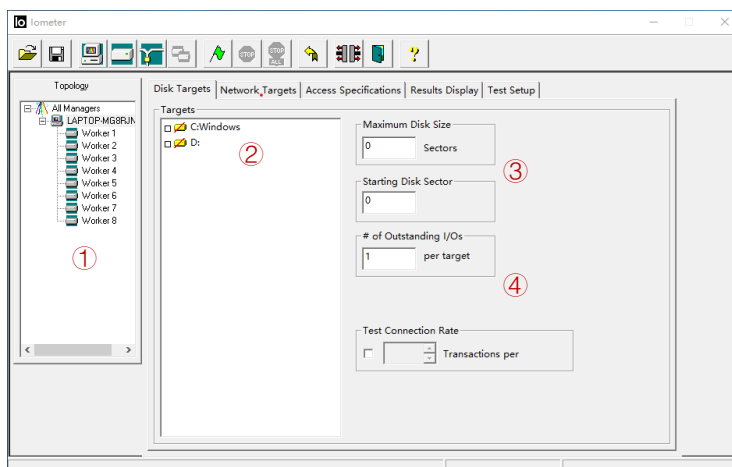


- Dynamo



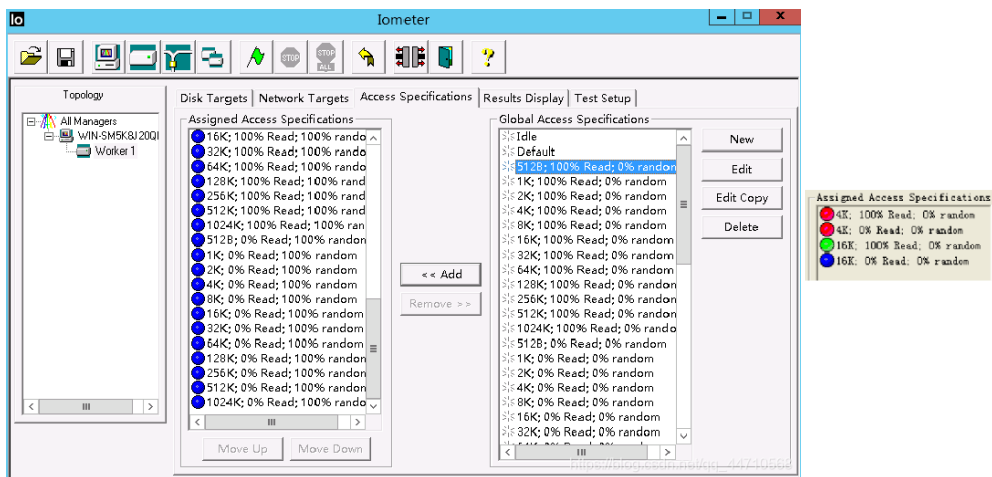
- Iometer是一个免费的开源的测试磁盘性能的工具。和其他磁盘工具相比，可以测试I/O的传输率和平均的I/O响应时间。
- Iometer 包含了两个程序，Iometer 和 Dynamo。
- Iometer 是控制程序。使用图形用户接口（GUI），你可以配置负载，设置操作参数，启动和停止测试。Iometer 告诉Dynamo 去做什么，搜集分析数据，将分析数据输出到文件中。在某一时刻，只能有一个Iometer副本运行；典型的情况是运行在服务器上。
- Dynamo 是负载生成器。它没有用户界面。当接收到Iometer 发送过来的命令，Dynamo执行相应的I/O 操作并且记录性能信息，然后将数据返回给Iometer。它可以有多个副本同时运行；典型的情况是服务器上运行一个副本，每个客户端运行一个副本。Dynamo 是多线程的；每一个副本都可以模拟多客户程序的工作负载。Dynamo 中的每一个运行的副本称为一个管理者；副本中的一个线程称为工作者。

lometer GUI界面



1. 界面中展示Manager数量，每个树根表示一台主机，每台主机下的worker数表示主机的CPU核数，CPU越强劲，worker数越多。
2. 选择树根后展现出来的Targets表示在这台主机上有的文件系统盘符或裸设备，黄色表示文件系统，蓝绿色表示裸设备。复选框打钩表示选择该设备进行测试。点击树根按ctrl依次多选，默认一个worker对应一个设备。建议一个worker对应一个磁盘设备。
3. 表示针对磁盘设备的测试扇区，Maximum Disk Sizes 是指lometer对多大的扇区进行访问，默认是0，意味着对整个磁盘扇区进行访问。Starting Disk Sector 是指从哪个扇区开始对磁盘进行访问。默认是0，意思是从第一个扇区开始访问磁盘。
4. # of Outstanding I/Os 是并发执行的指令数，表示每个worker工作线程大小，数值越大，线程越多，产生压力越大。

Access Specifications界面介绍



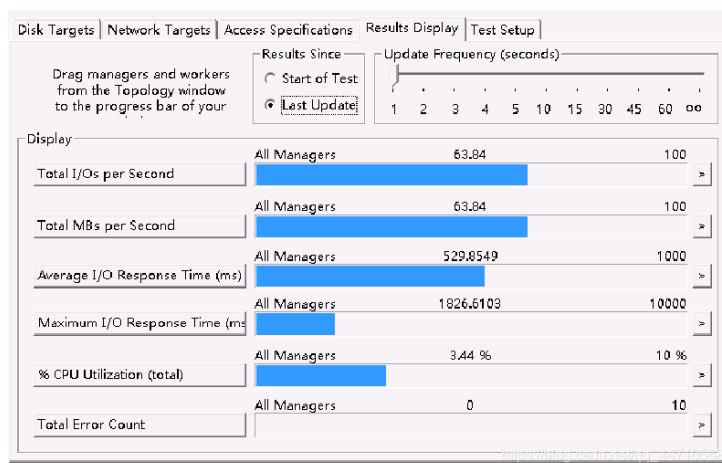
- Assigned access specifications下可以查看执行状态，红色表示已执行完，绿色表示正在执行，蓝色表示还未执行。当全部执行完后，全部恢复成蓝色。
- Global access specifications列出常用的IO参数，用户可以直接选中添加到左侧使用，也可以点击edit自定义参数。

Edit Access Specification界面介绍

Size	OKB	OB	% Access	% Read	% Random	Delay	Burst	Alignment	Reply
1MB	OKB	OB	18	100	100	0	1	sector	none
1MB	256KB	OB	26	0	48	0	1	sector	none
2MB	OKB	OB	26	43	48	0	1	sector	none
0MB	256KB	OB	28	0	48	0	1	sector	none

- Name: 测试任务的名字，无限制。
- Access Specification Line: 测试数据块列表，定义访问的数据块有哪些，显示各占比重有多大，各自读写比例是多少，顺序或随机访问的比例又是多少。
- Transfer Request Size: 传输数据块大小。
- Percent of Access Specification: 该大小的数据块在访问中出现的概率。
- Percent Read/Write Distribution: 读写比例。
- Percent Random/Sequential Distribution: 随机访问和顺序访问的比例。
- Default Assignment: 测试任务的类型，分为磁盘线程（只用于测试磁盘）、网络线程（只用于测试网络）和全局线程（可用于测试磁盘或网络），还有None（非测试用途）。
- Burstiness: 爆发设置。Transfer Delay是完成传输后的休息时间，Burst Length是指一次传输要完成多少个I/O操作。
- Align I/Os on: 设置每次读写数据在硬盘的排列方式。此项对网络测试没有作用，通常我们选按扇区排列。
- Reply Size: 设置应答块的大小。若设置了应答块的大小，则每次完成读操作，会有一个写操作。

Results Display界面介绍



- Start of Test 是指从开始测试以来的平均值。
- Last Update 显示最后更新的值，与更新频率有关，如：更新频率是5s，那么显示的值就是最后5s的平均值。
- Update Frequency 就是刷新显示数据的频率。
- Total I/Os per Second:IOPS，与transfer request size有关。
- Total MBs per Second:每秒传输的数据量，即MBPS。
- Average I/O Response Time:平均IO响应时间。
- Maxium I/O Response Time:最大IO响应时间。

Vdbench

- Vdbench是一款由SUN公司的Henk Vandenbergh开发的存储产品基准测试工具。它既可以测试裸盘的性能和数据一致性，也可以测试文件系统的性能和数据一致性，除此以外还提供了一些非常实用的数据后处理功能。
- 对裸盘的性能测试是Vdbench的基本功能，而且启动测试也非常简单，执行`./vdbench -f parmfile`，其中parmfile是后面一节会详细介绍的裸盘负载参数文件，Vdbench会按照parmfile中的设置对裸盘进行性能测试。parmfile包含5大类参数，分别是：General，HD，SD，WD和RD。在负载参数文件中，必须要按照General->HD->SD->WD->RD的顺序进行参数设置，每一类参数占据一行或者多行，但是不同类别的参数不能排列在同一行中。General在99.9%的情况下都不会用到，HD只在联机测试时使用。
- 文件系统的性能测试是Vdbench的基本功能，而且启动测试也非常简单，与裸盘测试相同，如下所示：`./vdbench -f parmfile`，与裸盘相比，文件系统多了目录和文件的概念，那么在Vdbench进行文件系统测试时，目录名会按照一定的方式产生，如`vdb_x.y.dir`。文件名也按照一定的方式产生，如`vdb_fnnnn.file`，其中nnnn表示文件的序号。包含3个大类参数，分别是：FSD，FWD和RD。在负载参数文件中，必须要按照FSD->FWD->RD的顺序进行参数设置。

Vdbench参数

- **裸盘测试参数：**

- **General：**负载参数文件的入口参数，包含各种全局设置。一般的性能、功能测试几乎不会用到这些参数。
- **HD：**Host Definition的简称，即主机定义。只有在多台主机联机测试的场合，或者需要对单台主机重新设置JVM数目的时候才有设置的必要。
- **SD：**Storage Definition的简称，即存储定义。这类参数是必须设置的，主要包含磁盘设备的位置、名称、大小、偏移和并发访问数等信息。
- **WD：**Workload Definition的简称，即工作负载定义。这类参数是可选的，主要是描述需要对哪些SD执行哪种I/O负载类型的下发，比如数据块大小，读写比例，随机和顺序访问的比例等等。
- **RD：**Run Definition的简称，即运行定义。这类参数是必须设置的，主要是描述需要执行哪些WD，需要产生多大的I/O负载压力，需要运行多少时间等等。

- **文件系统测试参数：**

- **HD：**Host Definition的简称，即主机定义。只有在多台主机联机测试的场合，或者需要对单台主机重新设置JVM数目的时候才有设置的必要。
- **FSD：**Filesystem Storage Definition的简称，即文件系统存储定义。这类参数是必须设置的，主要包含文件系统的挂载目录，目录结构，文件个数，文件大小等等。
- **FWD：**Filesystem Workload Definition的简称，即文件系统工作负载定义。这类参数是必须设置的，主要是描述需要对哪些FSD执行哪种文件访问方式和哪种I/O负载类型的下发。
- **RD：**Run Definition的简称，即运行定义。这类参数是必须设置的，主要是描述需要执行哪些FWD，需要产生多少操作数等等。

Vdbench裸盘测试参数

SD	size 设置测试区域的大小	threads 设置I/O并发数	openflags =o_direct表示下发的IO绕过内核缓存	offset 设置偏移量(512B的整数倍), 对于裸设备, 偏移量达到32K即可 对于文件系统, windows偏移32M, linux偏移1024	align 设置LBA地址的对齐边界, 必须是512B的整数倍 确认分条与IO大小, 偏移量是整除分条且大于IO
WD	seekpct 设置随机I/O的比例	rdpct 设置读I/O的百分比	xfersize 设置I/O数据块大小		
RD	iorate 限制IOPS最大值, =max不限制	warmup 设置测试热身时间	elapsed 设置任务运行时间	interval 设置测试数据的打印时间间隔	
可以用Vdbench工具, 对存储进行性能遍历测试, 对不同并发下, 不同读写模型 进行遍历测试。					
用例	工具配置			### Linux 1 RawDisks 遍历测试4k, 8k I/O, 分别在100%Random 100%Read和100%write的性能	
ID	I/O速率	随机比例 (%)	读比例 (%)	I/O Size	并发
1	max	100%	100%	4K	16
2	max	100%	100%	8K	16
3	max	100%	0%	4K	16
4	max	100%	0%	8K	16
root/vdbench50403 # cat examplefile7 sd=sd1,lun=/dev/sdb,size=200g,threads=16,openflags=o_direct,offset=4k,align=4k wd=wd1,sd=sd1,seekpct=100 rd=rd1,wd=wd1,iorate=max,warmup=10,elapsed=120,interval=1,forrdpct=(100,0),forxfersize=(4k,8k)					



Vdbench文件系统测试参数

FSD	size 设置生成的文件大小	depth 设置目录深度	width 设置目录宽度	files 设置每个目录的文件数量	anchor 指定创建新的目录结构生成文件并进行测试的路径
FWD	operation 设置文件系统操作类型	threads 设置并发数	fileio 设置每个文件中I/O操作的类型，顺序或者是随机	fileselect 选择目录和文件的方式是顺序的还是随机	
RD	fwdrate 限制负载每秒操作数ops，=max不限制	warmup 设置测试热身时间	elapsed 设置任务运行时间	interval 设置测试数据的打印时间间隔	format 运行之前对目录和文件进行预处理yes/no
fsd=fsd1,anchor=/testdir,depth=1,width=1,files=512,size=20m,openflags=o_direct fwd=fwd1,fsd=fsd1,operation=write,xfsize=10m,fileio=random,fileselect=random,threads=8 rd=rd1,fwd=fwd1,fwdrate=max,format=yes,elapsed=30,interval=1				生成固定同样大小的文件： 文件数量=depth*width*files 文件固定大小20MB format=yes 每次执行 都将目录以及文件删除重新创建	
fsd=fsd1,anchor=/testdir,depth=1,width=1,files=512,size=(4k,25,8k,25,1g,10,32k,40),openflags=o_direct fwd=fwd1,fsd=fsd1,operation=write,xfsize=10m,fileio=random,fileselect=random,threads=8 rd=rd1,fwd=fwd1,fwdrate=max,format=yes,elapsed=30,interval=1				生成不同比例不同大小的文件： 4K文件占25%；8K文件占25%；1G文件占10%；32K文件占40% 即，每个目录都会有files=512个文件，4K；8K；1G；32K文件各自占25%，25%，10%，40%	
fsd=fsd1,anchor=/testdir,depth=1,width=1,files=512,size=(4k,25,8k,25,1g,10,32k,40),openflags=o_direct fwd=format,threads=32,xfsize=512k fwd=fwd1,fsd=fsd1,operation=write,xfsize=10m,fileio=random,fileselect=random,threads=8 rd=rd1,fwd=fwd1,fwdrate=max,format=yes,elapsed=30,interval=1				指定生成文件时的压力： 默认fwd=default，测试文件系统生成文件时默认配置 threads=8,xfsize=128k fwd=format，需要按照指定的参数生成文件，需要在该参数后面指定threads和xfsize参数	

Fio

- Fio是一个开源压力测试工具，主要用来测试硬盘io性能。Fio是由Jens Axboe编写的，目的是实现对Linux I/O子系统和调度程序的灵活测试。
- 测试者可以使用fio进行各种混合IO测试，它支持13种不同类型IO引擎（libaio、sync、mmap、posixaio、network等等）。它可以测试块设备或文件，可以通过多线程或进程模拟各种io操作，可以测试统计IOPS、带宽和时延等性能。我们主要使用fio工具进行存储性能测试。
- 目前 fio 支持的操作系统有 Debian，Ubuntu，Redhat，Fedora，CentOS，Mandriva，Arch Linux，Solaris，Windows和BSDs。

Fio常用参数

参数	参数值	解释
filename	设备名或文件名，如：裸设备- /dev/sdb，文件-/home/test.img	定义测试对象，一般是设备或者文件。如果想测试裸盘设备/dev/sdb，可以设置filename=/dev/sdb；如果想要测试文件系统性能，则可以设置filename=/home/test.img
name	测试名称	定义测试名称。必填项，本地fio测试的名称，对性能没有影响。
rw	测试类型，可选值有：read， write，rw，randread， randwrite，randrw	定义测试的读写类型：read-顺序读，write-顺序写，rw(readwrite)-混合顺序读写，randread-随机读，randwrite-随机写，randrw-混合随机读写。对于读写混合类型来说，默认读写比例为1:1
rwmixwrite， rwmixread	混合读写中读写占用比例，可选 值为[0,100]	定义在混合读写模式中，写或读所占的比例。举例来说，rwmixread值为10，则表示读写比为10:90。
ioengine	io引擎选择，可选值有sync、 libaio、psync、vsync、mmap等 等	定义fio如何下发io请求。sync：基本的read、write io。libaio：linux原生的异步io，linux只支持non-buffer情况下的排队操作。默认值为psync，该模式为同步io模型，异步io模型libaio，一般结合direct=1使用。
direct	0或1	定义是否使用direct io：值为0，表示使用buffered io；值为1，表示使用direct io。
bs	带单位数字	定义io的块大小，单位是k，K，m，M。默认值为4k。
numjobs	正整数	定义测试的进程/线程数，默认值为1，如果指定了thread参数，则使用单进程多线程模式，否则使用多进程模式。
iodepth	正整数	定义每个进程/线程可以同时下发的io任务数，默认为1，适用于异步io场景，同步io场景要等前一个io任务完成才能下发下一个任务，所以iodepth并不起作用。

- fio是可以根据测试要求配置不同参数进行io性能测试的工具，可配置的参数有上百个，几乎可以覆盖所有的io模型，详细参数可查询fio的官网获取：
https://fio.readthedocs.io/en/latest/fio_doc.html。

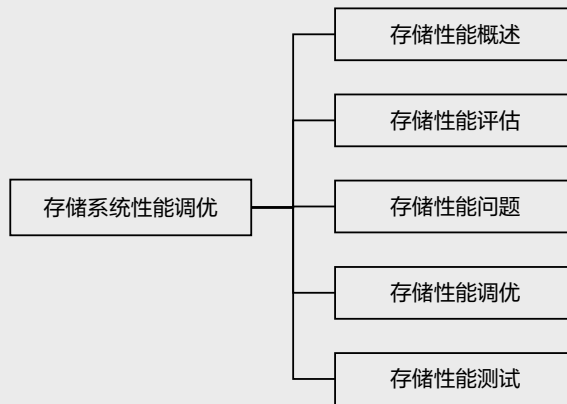
思考题

1. (单选题)在配置存储池时，如果存储的I/O类型主要是随机大I/O，则适合的分条深度是下列哪项？（ ）
 - A. 64KB
 - B. 128KB
 - C. 256KB
 - D. 512KB
2. (多选题)下列哪些原因可能会触发Cache写策略转变？（ ）
 - A. 出现电池故障
 - B. 系统单控
 - C. 高温转透写
 - D. LUN Fault页面超标

- 参考答案：

- D
- ABCD

本章总结



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://enterprise.huawei.com/cn/>
- 技术支持：<https://support.huawei.com/enterprise/>
- 在线学习：<https://support.huawei.com/learning/Index!toTrainIndex>
- 华为Support案例库：<https://support.huawei.com/enterprise/servicecenter?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.

Copyright©2023 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.



存储系统运维管理



前言

- 本章主要讲述存储系统运维管理，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：
 - 掌握使用存储运维管理工具
 - 掌握存储日常管理操作

目录

1. 存储系统运维管理工具

- CLI
- DeviceManager
- SmartKit
- DME IQ

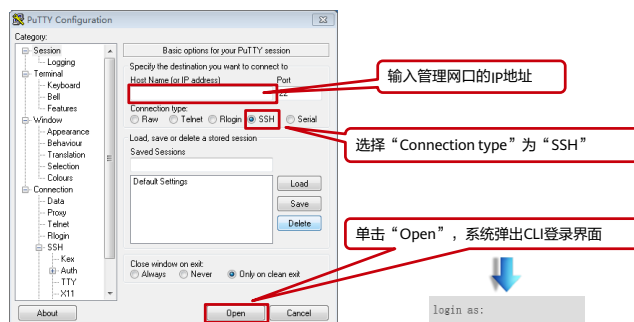
2. 日常管理

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。

目录

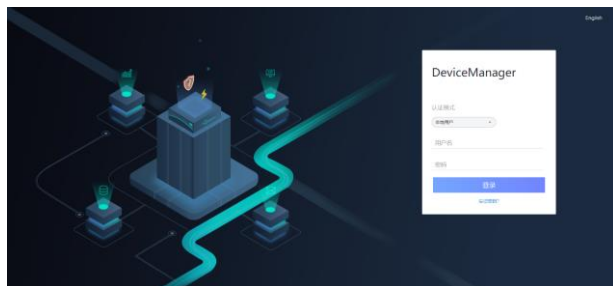
1. 存储系统运维管理工具

- CLI
- **DeviceManager**
- SmartKit
- DME IQ

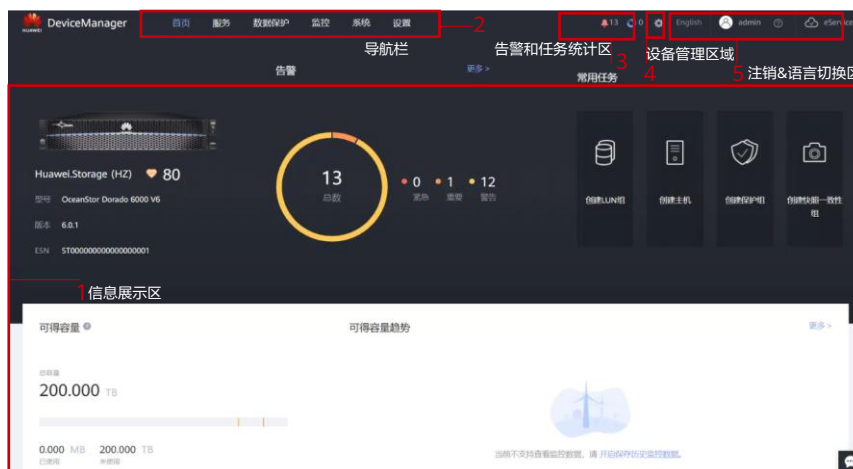
2. 日常管理

DeviceManager介绍

- DeviceManager是一款华为存储设备管理软件，可以轻松便捷地配置、管理和维护存储设备。
- 主要软件功能包括：存储资源分配、用户管理、数据保护特性管理、设备性能监控、告警管理等。



DeviceManager界面介绍 (1)



说明：不同产品版本和型号的界面可能会有少许差异，界面展示的信息请以实际界面为准。

- DeviceManager是一款华为存储设备管理软件，可以轻松便捷的配置、管理和维护存储设备。

DeviceManager界面介绍 (2)

序号	名称	说明
1	信息展示区	信息展示区显示了当前操作的相关界面，并可供用户进行操作。
2	导航栏	导航栏上列出了整个存储系统的各个功能模块，从逻辑上将存储系统划分为几个模块。
3	告警和任务统计区	告警统计区上显示了系统各级别告警的数量，用户可以了解当前系统的运行状态。 任务统计区上显示了用户执行的所有任务，可以了解任务是否执行成功。
4	设备管理区域	设备管理区域支持查看并修改设备信息，下电或重启设备等功能。
5	注销和语言切换区	注销语言切换区列出了注销按钮和语言切换按钮。 DeviceManager 支持简体中文和英文两种语言类型。

DeviceManager登录方式介绍



登录时一定要在管理网口IP地址后面加入“8088”端口号，否则会出现下列无法登录的情况。

具体输入格式：https://xxx.xxx.xxx.xxx:8088



- 任意一台与存储系统连接的维护终端，通过浏览器访问存储系统的管理网口IP地址，使用本地用户名或域用户名，就可以登录DeviceManager管理界面。

DeviceManager管理 (1)

存储空间检查



- ☐ 存储池
- ☐ LUNs
- ☐ 文件系统
- ☐ 配额
- ☐ ...

存储资源状态检查



- ☐ 存储池
- ☐ LUNs
- ☐ 文件系统
- ☐ ...

增值特性状态检查



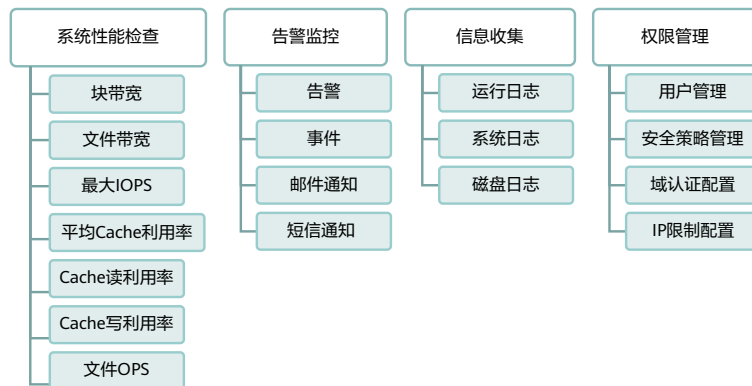
- ☐ 双活
- ☐ 远程复制
- ☐ 一致性组
- ☐ 克隆任务
- ☐ 快照任务
- ☐ LUN拷贝任务
- ☐ ...

容量趋势检查

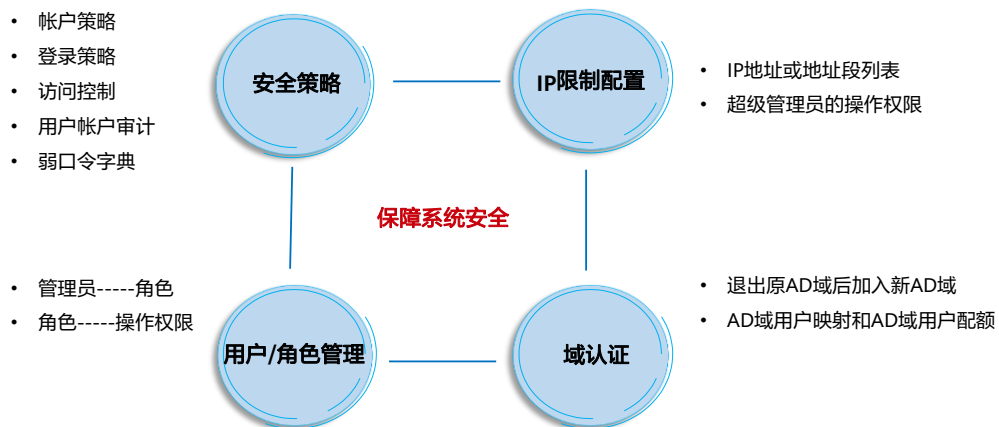


- ☐ 总容量
- ☐ 已使用容量
- ☐ 可用容量
- ☐ 已映射LUN容量
- ☐ 未映射LUN容量
- ☐ Thin LUN总容量
- ☐ 已分配Thin LUN容量
- ☐ ...

DeviceManager管理 (2)

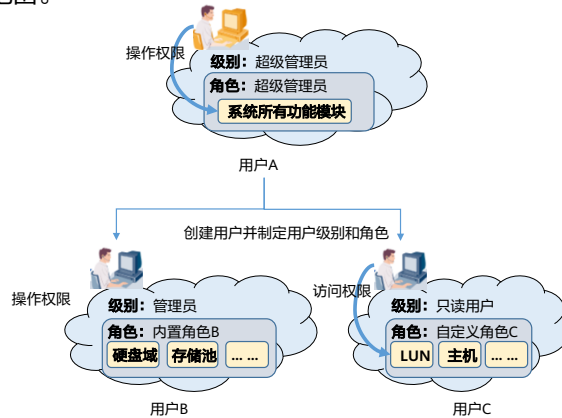


权限管理



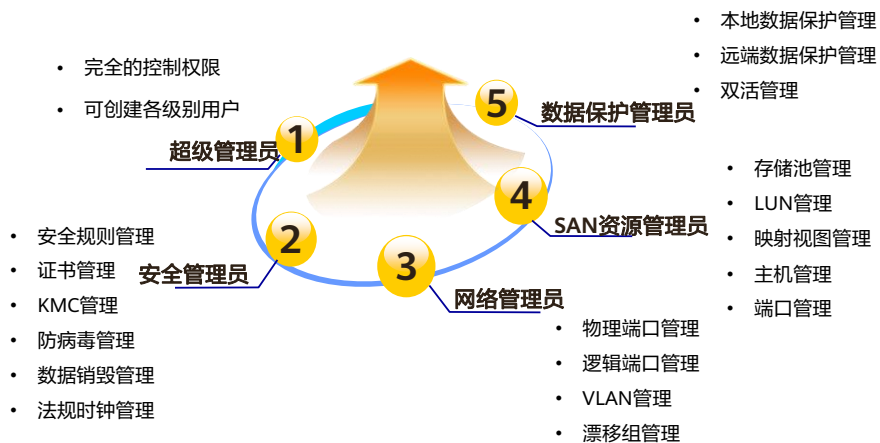
存储系统用户管理

- 为了防止错误的操作影响到业务系统的稳定性及业务数据的安全性，系统通过级别和角色来控制用户的操作权限和范围。



- 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；
 - 不能进行用户管理、升级系统、修改系统时间、设备下电和重启设备等操作。
- 只读用户
 - 对存储设备只有访问权限；
 - 登录设备后，只能进行查询操作。

用户角色及权限



目录

1. 存储系统运维管理工具

- CLI
- DeviceManager
- **SmartKit**
- DME IQ

2. 日常管理

SmartKit简介



1、统一平台

桌面工具管理平台，集成存储、服务器、虚拟化、云计算及机器视觉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
- 服务器
- 云计算
- 虚拟化
- 机器视觉

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、服务器、云计算、虚拟化、机器视觉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。

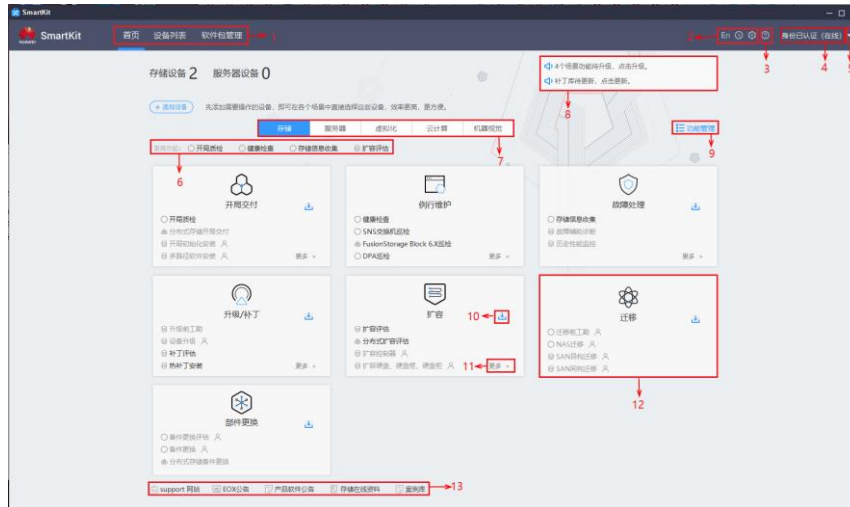
闪存存储服务场景（2）

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。

首页界面功能描述

区域	功能描述
1	功能导航栏：通过此导航栏可以选择功能类别。
2	定时任务与系统设置按钮。
3	帮助、关于与语言切换栏。
4	显示当前是否完成身份认证。
5	立即认证：单击立即认证可以认证身份以获得更多的工具功能和支持。 退出认证：单击退出认证可以退出当前使用的账号。 可查看获取的是企业网还是运营商服务。
6	常用功能：列出您常用的最多四个场景功能。
7	领域切换栏：点击可切换存储、服务器、虚拟化、云计算及机器视觉领域。
8	信息公告：提示用户升级工具箱、场景功能以及更新存储补丁库。 说明：如果离线状态下的工具箱180天未升级，给出相应升级提示。
9	功能管理：包含了工具箱所集成的功能管理列表，在功能管理界面可对单个或多个功能进行下安装、升级、卸载等操作。
10	📌：点击此图标可安装当前场景任务下的所有待安装功能。 🔄：点击此图标可升级当前场景任务下的所有待升级功能。
11	更多：点击该按钮，系统弹出该场景任务下的所有功能对话框。
12	场景化入口：单击此入口，可以使用相应的场景化功能。
13	链接（企业用户）：公告、智能机器人和Support-E等网站的超链接，存储、服务器或云计算领域的在线资料列表和案例。 链接（运营商）：Support网站、EOX公告和软件版本推荐的超链接，存储、服务器或云计算领域的在线资料列表和案例库。 说明：当用户类型为“客户”或者“访客”时，首页不显示此链接。

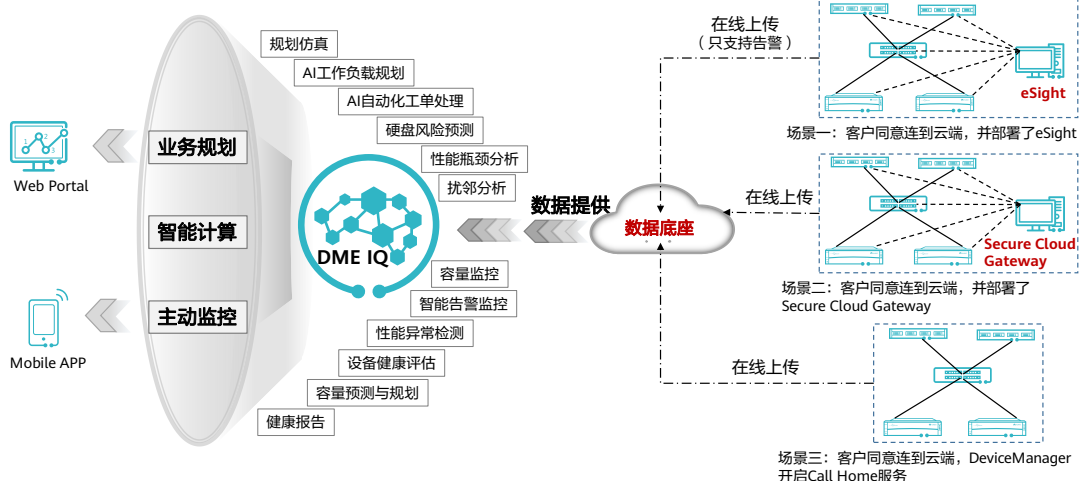
目录

1. 存储系统运维管理工具

- CLI
- DeviceManager
- SmartKit
- **DME IQ**

2. 日常管理

DME IQ



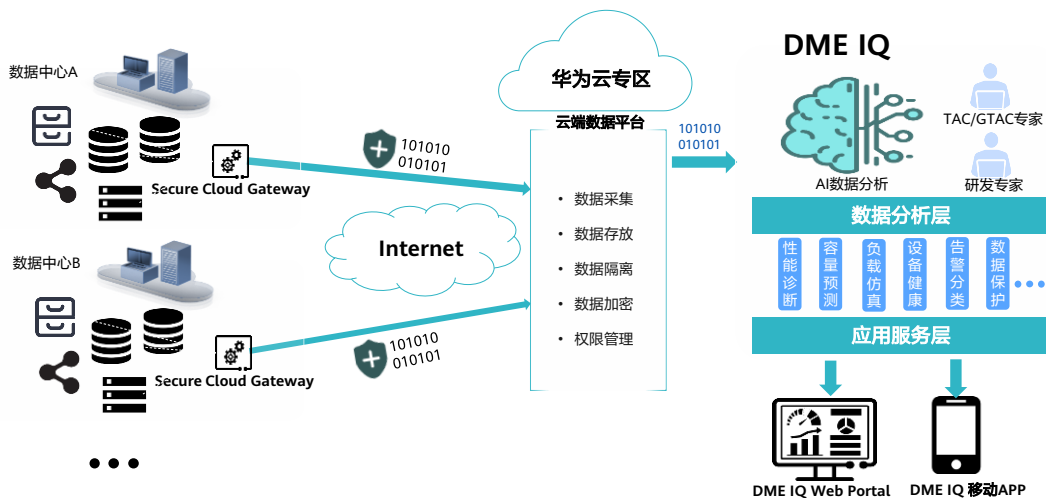
- DME IQ定位云端智能运维平台，基于云端基础底座提供的大量数据，通过大数据分析和人工智能AI技术，驱动IT运维活动的开展，降低运维难度、全面提升运维效率。同时，DME IQ还支持手机APP端运维方式，可以随时随地查看设备运行状态，方便快捷。
- DME IQ云端：IT存储集中运维的能力扩展，云端管理，提供智能运维、规划仿真、预测预防等特性。作为云端运维的基础数据底座，能够收集IT存储设备的告警、性能、配置、系统日志、硬盘信息等数据，并提供远程维护和存量经营能力。同时，DME IQ云端提供的大量运维数据，是DME IQ业务孵化和演进的核心基础。
- Secure Cloud Gateway：Secure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

应用场景



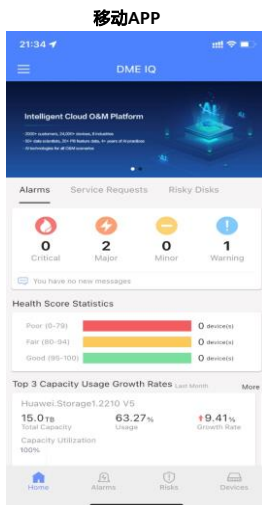
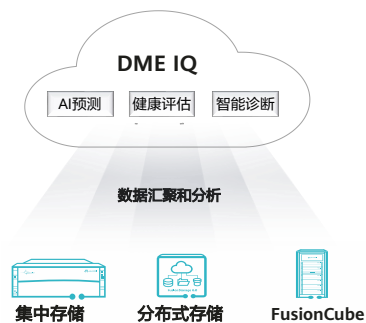
- DME IQ运行在华为云，利用机器学习（Machine Learning）和深度学习（Deep learning）等AI技术进行大数据分析，主动帮助客户设备监控，分析和规划华为存储、服务器等IT基础设施。实现随时随地主动式、预见式运维，降低运维成本，提升运维效率。为客户提供一站式运维解决方案。

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。

主动监控



① 告警/风险通知

故障/风险主动通知，实时掌控设备状态

② 智能诊断

拓扑展示故障详情，影响范围可视

③ 进展查询

随时查看问题单进展，跟踪解决进度

- 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。

智能分析

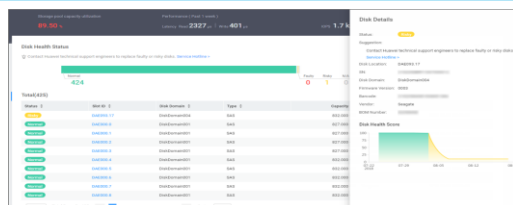


提前排除风险，提升可靠性



提前**14天**识别故障硬盘

采用融合模型+专家经验算法，持续
在线学习，识别率达到80%，误报率
小于0.1%



海量数据分析，构建成熟的风险盘预测模型



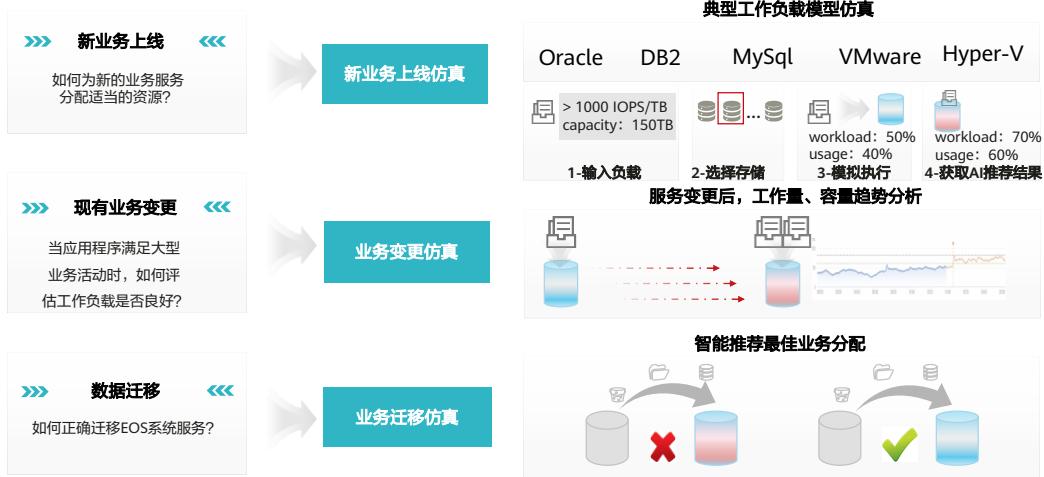
累计分析60万+硬盘，200亿+条
特征数据

企业数据中心 模型验证1000天+



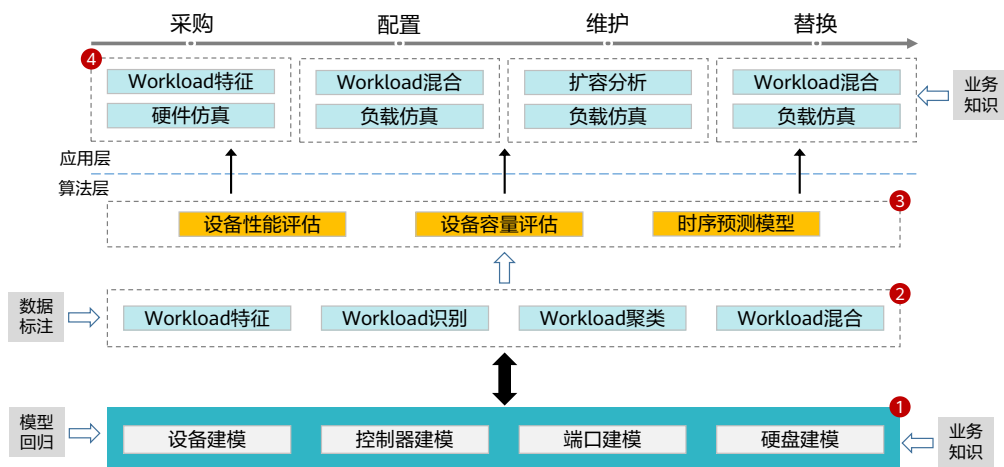
- DME IQ采用AI来动态分析硬盘日志的变化、硬盘性能指标波动规律，同时结合华为内置的硬盘定位日志关联分析，确保预测更加准确，而且可以给出寿命的预测。
- 智能识别设备性能异常并提供分析建议，将性能问题扼杀在萌芽阶段，保障业务稳定运行。
- 从VM到存储的全栈可视化，快速发现存储网络性能问题。
- 快速诊断Oracle数据库性能异常，定位问题根本原因。

规划仿真



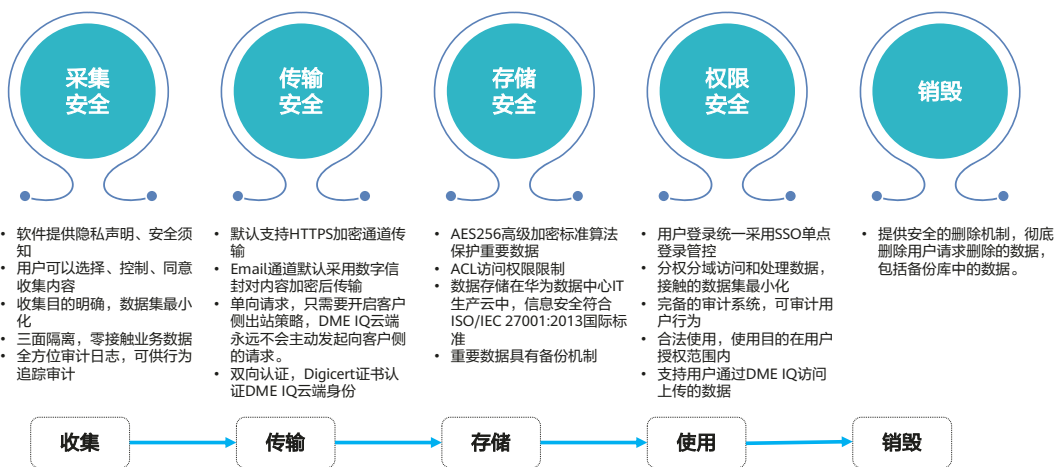
- 规划仿真是面向存储设备全生命周期的仿真评估, 支持业务上线、业务SCALE和业务迁移3大负载变更场景, 支持Oracle、DB2等八大业务类型自动化评估与最佳业务安置配置智能推荐, 实现负载和容量趋势分析和预测, 准确率大于80%。另外, 能够为运维工程师提供设备性能和容量变化趋势评估能力, 辅助其方案变更决策。

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



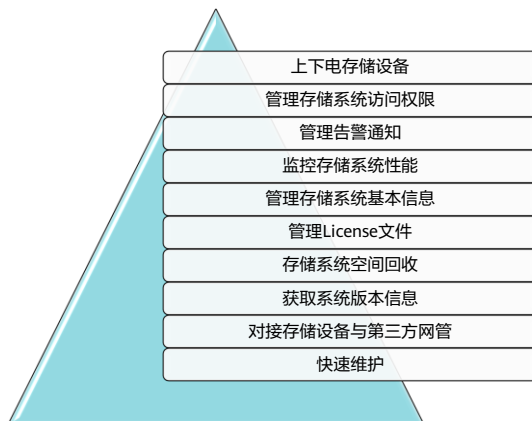
- DME IQ从业务流程、整体架构设计和技术实现方案上都进行了仔细的安全考量，做到对客户安全进行严密防护。
- 采集安全
 - 经过客户允许才会部署并符合当地法律，不碰设备的业务数据，只获取相关运维信息，详细收集的信息条目均列在《授权协议书》中。
 - Secure Cloud Gateway上进行的所有操作均有记录，上传记录可追溯，并采用完整性校验，确保记录不被篡改。
 - 在架构设计上，Secure Cloud Gateway的所有连接都从客户侧向外发起，任何第三方包括华为公司都不能主动从外部网络向客户侧创建网络连接。
- 传输安全
 - 双向认证：Secure Cloud Gateway通过证书认证华为云端，华为云端通过Support帐号认证Secure Cloud Gateway。
 - 单向访问：Secure Cloud Gateway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。
 - 数据保存时均经过AES256加密处理，确保信息不泄露。
- 权限安全
 - Secure Cloud Gateway和DME IQe云端，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. 存储系统运维管理工具
2. 日常管理

存储日常管理项目总览



上电存储系统（电源按钮方式）



2U控制框的电源按钮



4U控制框的电源按钮



- 系统正常上电顺序：
 1. 打开各设备对应的外部电源开关。
 2. 依次按每个控制框上的电源按钮。
 3. 打开交换机的电源开关（如果配置了交换机，但尚未上电）。
 4. 打开应用服务器的电源开关。
- 上电时只需按一下电源按钮。控制框电源指示灯绿色闪烁，表示存储系统开始上电。请勿长按电源按钮，长按超过5秒钟将导致存储系统下电。
- 存储系统上电需要15~30分钟。
- 当控制框上电后，与其级联的硬盘框会自动上电。
- 上电完成后，硬盘会自动初始化，初始化时间与初始化硬盘数量相关，请以具体实施情况为准。

上电存储系统（命令行远程上电）

- 集群内所有控制框均未上电时：

- 使用BMC管理员用户通过管理网口IP地址登录BMC系统。

```
$ ssh admin@XXX.XXX.XXX.XXX
The authenticity of host 'XXX.XXX.XXX.XXX (XXX.XXX.XXX.XXX)' can't be established.
ECDSA key fingerprint is SHA256:RV2Qwdr0wL2XXXXXXXXXXXXXXXXXXXXXX.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'XXX.XXX.XXX.XXX' (ECDSA) to the list of known hosts.

Authorized users only. All activities may be monitored and reported.
admin@xx.xx.xx.xx's password:
```

- 进入BMC系统后，执行cliset -d clusterpoweron 1命令，上电存储系统。

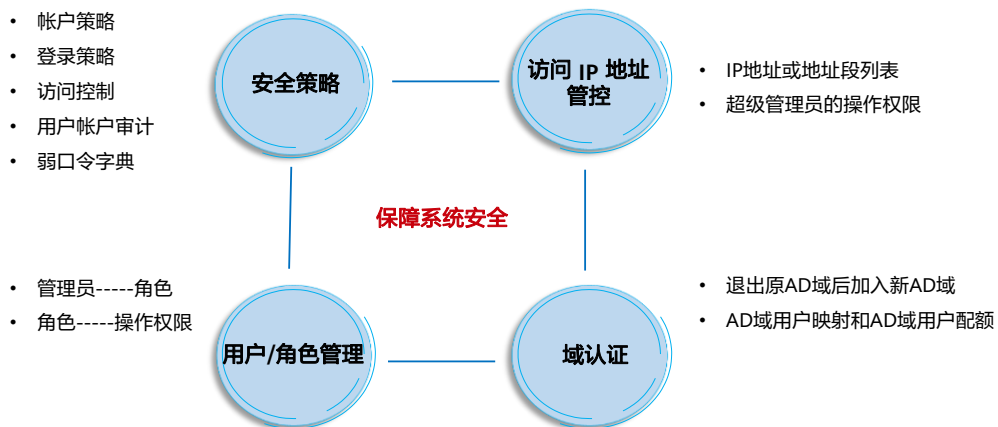
```
iBMC:/->cliset -d clusterpoweron 1
Cluster power on OK.
```

- 对于单控制框存储系统
 - 控制框电源线缆已正确连接。
 - 维护终端已连接至控制框的任意一个管理网口。
- 对于多控制框存储系统
 - 所有控制框电源线缆已正确连接。
 - 维护终端已连接至任意控制框的任意一个管理网口。
 - 所有控制框的Scale-out线缆已正确连接，所有控制框的维护网口已按照标准组网正确连接。

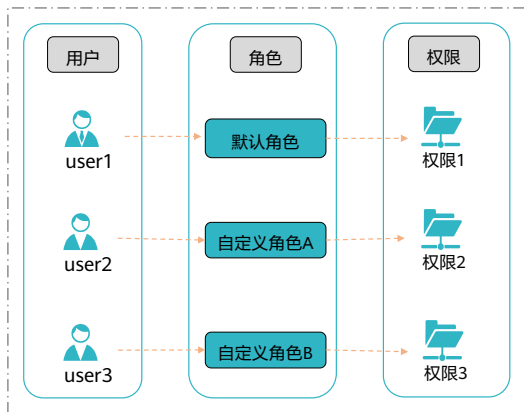
下电存储系统

- 如果您需要停止使用存储系统、更换系统插框或者因设备供电链路故障需要下电存储系统时，请通过正常的下电操作对存储系统进行下电。
- 存储系统正常下电顺序：停止主机业务 > 下电控制框 > 断开外部电源。
- 控制框正常下电的操作方式有两种：
 - 长按控制框电源按钮5秒；
 - 在控制框的DeviceManager上执行下电操作。

管理存储系统访问权限



基于角色权限管理-RBAC



权限=功能1: 对象1,对象2; 功能2: 对象3, 对象4;(一组或者多组功能和对象的关系)

预置默认角色: 系统预置系统管理用户和租户管理用户的默认值角色。

系统管理用户默认角色:

超级管理员 (super_administrator)
管理员(administrator)
安全管理员 (security_administrator)
SAN管理员 (san_administrator)
NAS管理员 (nas_administrator)
数据保护管理员 (dataProtection_administrator)
网络管理员 (network_administrator)

租户管理用户角色:

租户管理员 (vStore_administrator)
租户数据保护管理员 (vStore_dataProtection)
租户协议管理员 (vStore_protocol)

自定义角色: 支持用户依据自己场景自定义角色。

用户自定义角色时可以选择多个功能, 每个功能可以选择多个对象。
自定义角色支持删除和修改。

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

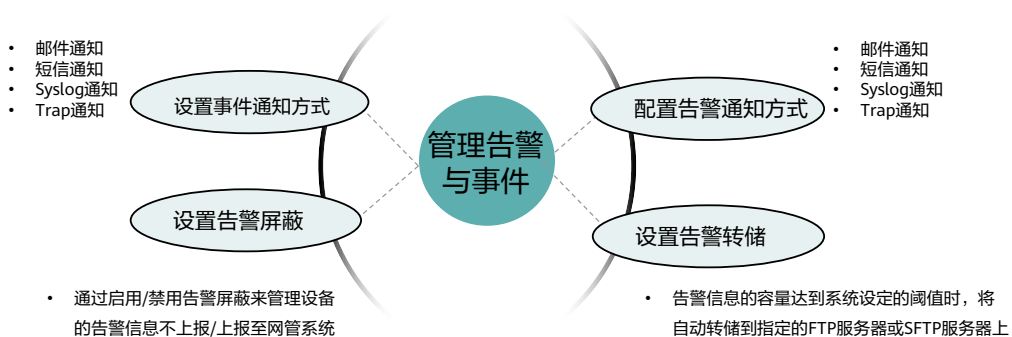
配置CAS服务器

- 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

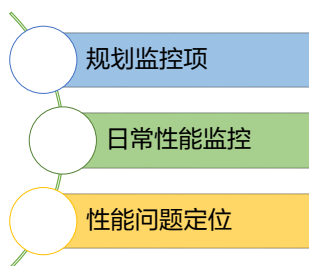
第三方网管平台 免密登陆

- 获取令牌
- 访问方法: PUT

管理告警与事件

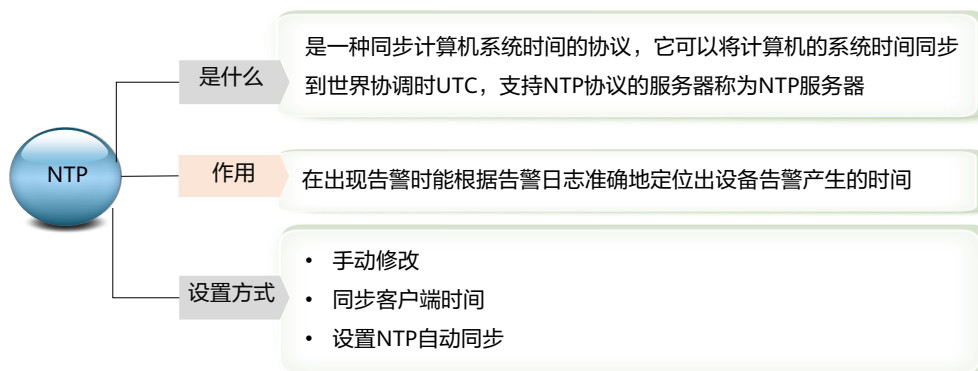


监控存储系统性能



- 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

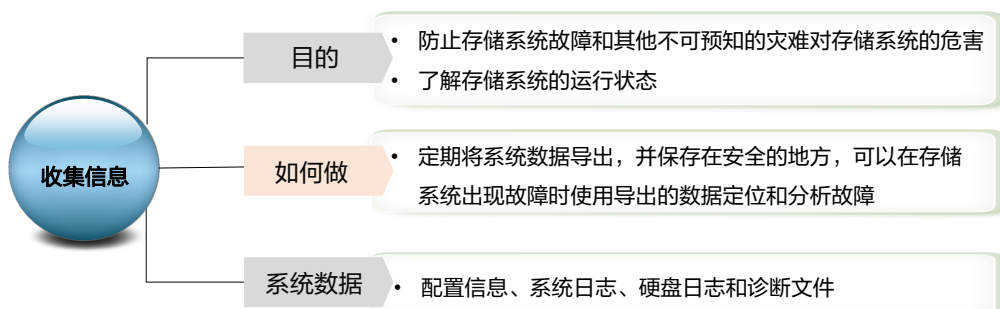


- 使用CLI获取ESN序列号

>> 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

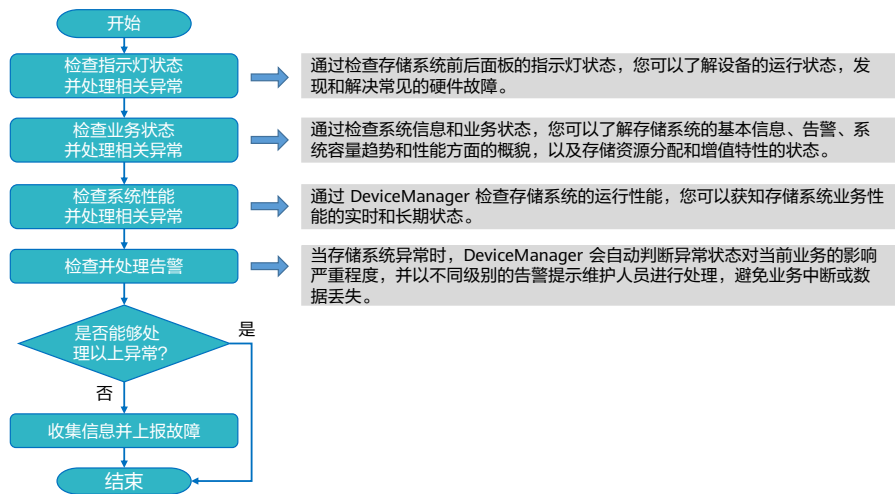
信息类型	名称	说明
基本信息	设备序列号和版本	提供存储设备的序列号与版本信息。
	客户信息	提供具体的联系人及其联系方式。
故障信息	故障发生时间	记录故障发生的准确时间。
	故障现象	详细记录故障发生时的现象，例如弹出的错误提示对话框、收到的事件通知内容等。
	故障前执行的操作	回忆并记录故障发生前执行的操作。
	故障后执行的操作	回忆并记录故障发生后到向维护人员上报故障前执行的操作。
存储设备信息	硬件模块配置	详细记录存储设备硬件模块的配置信息。
	指示灯状态	记录存储设备指示灯状态，尤其需要记录橙色和红色状态的指示灯信息。存储设备的各部件指示灯状态，请参见对应产品型号的《产品描述》。
	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



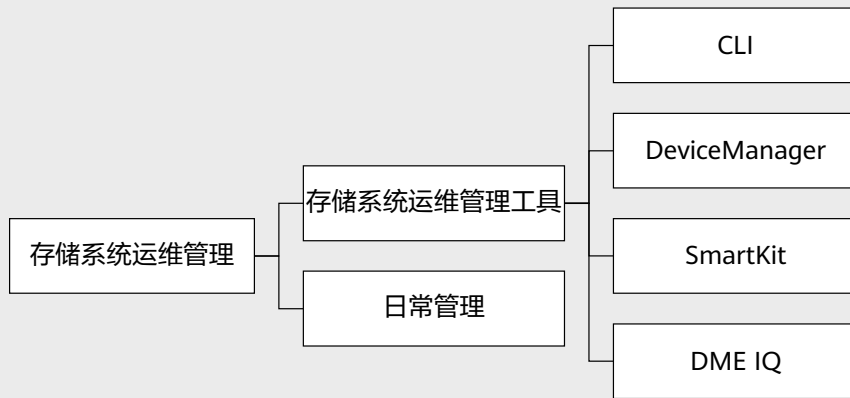
思考题

1. (多选)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



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/>
- 华为人才在线学习: <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.

Copyright©2023 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.



存储系统故障处理



前言

- 本章主要讲述存储系统故障类型、故障处理流程、故障处理基本原则、常用的故障定位方法和部件更换操作。

目标

- 学完本课程后，您将能够：
 - 掌握存储系统故障处理基础知识
 - 描述存储系统故障处理流程
 - 理解存储系统故障处理方法
 - 掌握部件更换的方法

目录

1. 故障处理基础
2. 故障处理流程
3. 收集信息与上报故障
4. 部件更换
5. 案例分析

故障定义

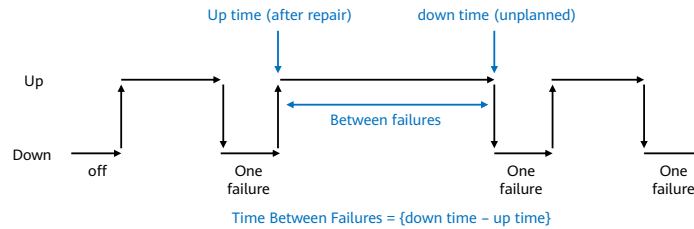
- 故障 fault
 - 产品不能执行规定功能的状态。预防性维修或其他计划性活动或缺乏外部资源的情况除外。故障通常是产品本身失效后的状态，但也可能在失效前就存在。(GB/T3187-94)
 - A fault is defined as an abnormal condition or defect at the component, equipment, or sub-system level which may lead to a failure. (ISO/CD 10303-226)

故障分析

- 故障分析，又称为故障诊断，是指为了确定故障原因以及如何防止其再次发生而收集和分析数据的过程。
- 故障分析的重要性：
 - 故障分析乃是制造行业众多分支之中的一门重要学科。例如，在电子行业，新产品开发和已有产品改进时的重要手段就是故障分析。
 - 在故障分析过程中，需要采用各种各样的方法和手段，收集故障部分的数据和信息，以便于故障原因（一种或多种）的后续分析。（摘自维基百科）

MTBF和MTTR

- MTBF (Mean Time Between Failures , 平均故障间隔)
 - 是描述产品在操作使用或测试期间的平均连续无故障时间。
- MTTR (Mean time to recovery, 平均修复时间)
 - 是描述产品由故障状态转为工作状态时修理时间的平均值。
- 可用性 (也称有效性)
 - 是指可维修产品在规定的条件下使用时, 在某时刻具有或维持其功能的能力。其值可用MTBF和MTTR来计算: 可用性 = $MTBF / (MTBF + MTTR)$ 。



可用性与DT的关系

- DT (Downtime, 停机时间)
 - 是描述产品处于不能工作状态的累积时间。一般采用“分钟/年”为单位。

DT与可用性A的换算关系:

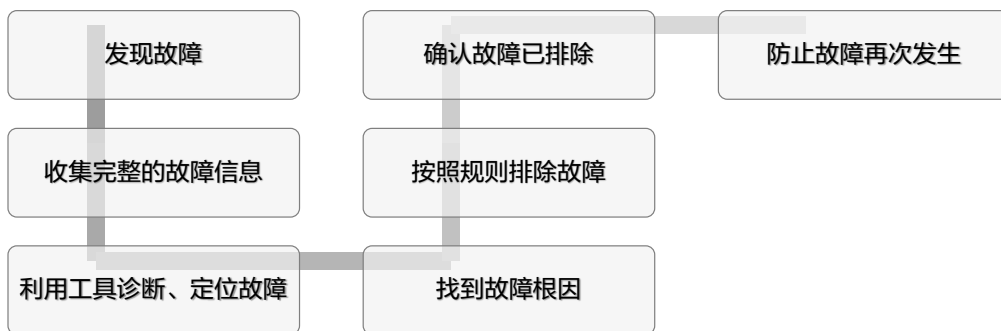
$$DT = (365 \times 24 \times 60) \text{ 分钟} \times (1 - A) = 525600 \text{ 分钟} \times (1 - A)$$

对应5个9, (99.999%) 的可用性, 停机时间为:

$$525600 \text{ 分钟} \times (1 - A) = 525600 \times 0.00001 = 5.256 \text{ 分钟/年}$$

约为: 5分钟/年

故障处理关键动作



目录

1. 故障处理基础
- 2. 故障处理流程**
3. 收集信息与上报故障
4. 部件更换
5. 案例分析

对维护人员的要求

- 专业素质和技能
 - 熟悉RAID、DAS、NAS、SAN等存储技术。
 - 熟悉以太网技术。
 - 熟悉TCP/IP的基础知识。
- 掌握基本操作
 - 能够操作存储设备。
 - 能够操作各种类型的应用服务器。
 - 能够操作与存储业务相关的数据传输设备，例如：以太网交换机、FC（Fibre Channel）交换机和路由器等。
- 掌握常用测试仪表的使用方法
 - 光功率计
 - 线路测试仪
 - 万用表
- 熟悉存储组网信息
 - 熟悉常见存储组网类型。
 - 熟悉维护局点组网情况。
 - 熟悉维护局点设备的运行情况。
- 采集和保存现场数据
 - 现场数据的采集和保存包括设备正常运行时的定期采集和设备发生故障时的采集。一般在进行故障处理前，要求维护人员首先要采集、保存现场数据。

故障处理工具和仪表清单

工具名称	用途
光功率计	用于测量光功率。
光纤跳线	用于替换故障尾纤。
光纤绑扎带	用于捆扎替换后的故障尾纤。
网线	用于替换出故障的网线。
网线测试仪	用于测试以太网网线的连通性。
万用表	用于测量各种电气参数。
防静电腕带	防止人体静电损坏敏感元器件。
防静电包装袋	保护需要防静电的部件。
防静电服	防止人体静电损坏敏感元器件。
十字螺丝刀	用于拆装设备螺丝。
标签纸	用于标识设备或者线缆。

故障分级

- 根据故障的严重程度和影响范围，存储系统故障问题等级可分为3级：普通故障、中级故障、严重故障。

故障等级	特点	处理措施
普通故障	存储系统性能或主机业务没有任何影响的故障。如无法登陆 DeviceManager、串口无法激活登陆等	普通故障处理
中级故障	存储系统性能下降导致主机I/O延迟或主机业务中断。如接口模块故障、应用服务器重新启动后无法加载多路径软件等	应急处理
严重故障	存储发生多个部件故障导致部分生产系统无法使用、部分数据丢失的风险很高或已经丢失。如控制器故障等	应急处理

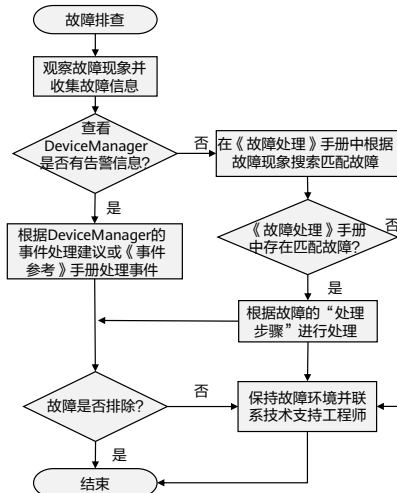
故障分类

- 根据故障发生的部件不同可以将故障分为存储侧故障及外部环境故障两大类。

- 存储侧故障：**主要指存储内部硬件或软件引起的存储系统本身故障，通常可以通过存储的告警平台获取相关故障信息。
- 外部环境故障：**数据从主机端经过网络到存储的过程中出现的软硬件故障，此类故障大部分是由于网络链路导致的。通常可以通过操作系统的系统日志、应用程序日志、或交换机的日志查看故障信息。

故障分类	故障部件	故障现象
存储系统内部故障	硬盘故障	单块或多块硬盘故障
	电源模块故障	单个电源模块故障
	BBU故障	单个或多个BBU模块故障
	控制器故障	控制器故障
	接口模块故障	接口模块故障
外部环境故障	风扇故障	风扇故障
	HBA卡故障	HBA卡硬件故障
	光纤线故障	HBA的链路故障
	交换机SFP故障	存储前端的链路故障 交换机SFP故障
	交换机板卡故障	交换机端口卡故障
	交换机故障	交换机宕机
	链路闪断	由于HBA、SFP、存储前端、光纤线问题引起链路闪断

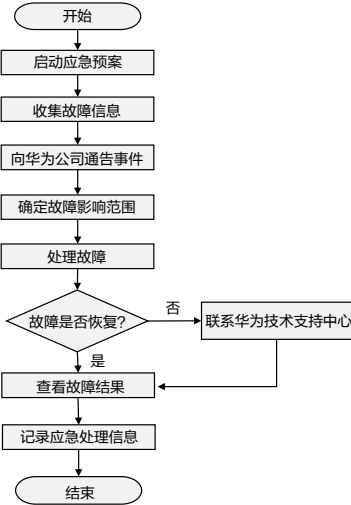
普通故障处理



- 普通故障处理是指存储系统性能或主机业务没有任何影响的故障，故障等级为普通故障。

操作步骤	操作说明
收集故障信息	当故障发生时，请注意收集故障处理所需的相关信息，以便更好、更快地定位故障发生的原因，从而排除故障。收集信息的相关内容，请参见收集故障信息。
登陆DeviceManager管理软件	登陆DeviceManager管理软件，可以查询存储设备的运行状态，以及是否有告警等事件信息产生。
根据事件处理建议处理事件	如果在DeviceManager中查询到事件信息，则根据DeviceManager提供的事件处理建议或参见《事件参考》处理事件。
定位故障产生的原因	定位故障是指从众多可能原因中找出故障发生的具体原因的过程，可通过分析、比较各种可能的故障原因，不断排除不可能因素，最终确定故障发生的具体原因。常用的故障定位方法请参见故障定位的方法中的描述。
联系技术支持中心	当您无法排除故障并收集好故障的相关信息后，请您联系华为技术有限公司客户服务中心以获取技术支持。

应急处理



- 应急处理是指当突发情况下，为迅速排除故障，恢复系统正常运行、尽量减少事故损失而进行的处理措施，故障等级为中级故障或严重故障。

操作步骤	操作说明
启动应急预案	应急预案指针对可能发生的事故，为迅速、有序地开展应急处理而预先制定的行动方案，当中级故障或严重故障发生时需要启动预先制定的应急预案。
收集故障信息	当故障发生时，请注意收集故障处理所需的相关信息，以便更好、更快地定位故障发生的原因，从而排除故障。收集信息的相关内容，请参见收集故障信息。
向华为公司通告事故	当紧急故障发生时，应第一时间通知华为公司，并反馈收集到的故障信息。
确定故障影响范围	当故障发生时，根据故障现象确定故障的影响范围，以降低故障带来的损失。
处理故障	处理故障以恢复主机业务。
联系华为技术	当您无法排除故障并收集好故障的相关信息后，请您联系华为技术有限公司客户服务中心以获取技术支持。
记录应急处理信息	记录在应急过程中的故障信息和处理步骤等信息。

故障处理基本原则

先定位外部 后定位内部

- 外部设备问题包括光纤、光缆、客户设备和断电等问题。
- 内部问题包括硬盘、控制器、接口模块等问题。

先高级别告警 后低级别告警

- 在分析告警时，应该按照级别的优先级依次进行分析，如先分析紧急级别的告警，再分析重要和警告级别的告警。
 - 紧急：已经造成客户业务中断或者可能引起设备不可用的故障，需要立即采取措施修复。
 - 重要：对设备产生部分影响或者对系统性能造成影响，需要采取纠正措施，以防止更严重的故障的发生。所有硬件类型的故障均属于该级别。
 - 警告：对设备没有影响，系统检测到潜在的或者即将发生的影响业务的故障，且当前还没有影响业务。

先共性告警 后个别告警

- 在分析告警时，应先分析是个别问题还是共性问题，确定问题的影响范围。
- 需要确定是一个模块出问题，还是多个模块出现类似问题。

故障定位的方法

一、告警信息分析法

- 分析原理：
 - 当系统发生故障时，一般会伴随有大量的告警信息产生，通过查看告警信息并配合对性能数据的分析，可大概判断出所发生故障的类型和位置。
- 应用场景：
 - 在能够正常收集到告警信息的情况下，告警分析法适用于定位任何故障。
- 应用举例：
 - 某视频业务运行过程中，突然出现业务性能下降，视频画面断断续续的情况。登陆管理界面后发现当前有紧急告警。进入告警管理查看发现有告警信息，并直接定位到是某硬盘出现故障。通过更换该硬盘，故障排除。
- 小结：
 - 通过分析告警，可以定位故障的具体部位或原因，也可以配合其他方法定位故障原因。

二、替换法

- 分析原理：
 - 替换法就是使用一个工作正常的部件去替换一个疑似工作不正常的部件，从而达到定位故障、排除故障的目的。这里的部件可以是一根网线、一个控制器或者一个级联模块。
- 应用场景：
 - 替换法适用于硬件故障的处理，往往可以快速准确地定位出发生故障的部件，并且对维护人员没有特别的要求。使用替换法的局限在于事先必须准备相同的备件，因此要求进行较充分的前期准备工作。
- 应用举例：
 - 某图书馆设备巡检过程中，发现无法通过ISM登陆存储设备，主机端也无法ping通存储设备。通过拔插或更换网线发现，网线松动或损坏导致该问题。
- 小结：
 - 替换法的优势是可以将故障定位到较细的位置，且对维护人员的要求不高，因此是一种比较实用的方法。

目录

1. 故障处理基础
2. 故障处理流程
- 3. 收集信息与上报故障**
4. 部件更换
5. 案例分析

故障信息收集

- 在故障发生后，通过及时收集和上报基本信息、故障信息、存储设备信息、组网以及应用服务器信息，可以帮助维护人员更快速的定位故障原因并排除故障。
- 信息收集项目
 - 需要收集的信息项目主要包括基本信息、故障信息、存储设备信息、组网以及应用服务器信息。
- 导出系统数据
 - 为了准确了解存储系统的运行状态，可以定期将系统数据导出并进行查看和分析。导出的系统数据包含运行数据、系统日志和硬盘日志。
- 导出告警和日志信息
 - 告警和日志信息用于保存存储设备在运行过程中自动记录的故障和事件。当存储设备出现故障时，维护人员可以通过查看告警和日志信息来定位和解决故障。

信息收集项目

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。 说明： 可以通过登录DeviceManager，在“基本信息”区域框中查询到存储设备序列号和版本信息。
	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	记录应用服务器上安装的操作系统类型和版本号。
	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》。
	操作系统日志	查看并导出操作系统日志信息。

- 需要收集的信息项目主要包括基本信息、故障信息、存储设备信息、组网以及应用服务器信息。

收集文件系统故障信息

- 收集Windows文件系统故障信息
 - 收集主机操作系统版本信息。
 - 右键单击“计算机”选择“属性”。
 - 在弹出的“系统”界面中收集Windows系统版本信息。
 - 收集Windows系统日志。
 - 右键单击“计算机”选择“管理”。
 - 在“事件查看器”中选择“Windows日志”。
 - 选择“系统”，在左侧收集Windows系统故障的相关日志信息。
- 收集AIX文件系统故障信息
 - 收集“/var/adm/ras/errlog”文件。
 - “errlog”记录了系统的软硬件故障信息。
- 收集ExtX文件系统故障信息
 - 执行**fdisk -l**命令查看系统内所有磁盘及分区情况。
 - 执行**mount**命令查看当前操作系统挂载信息。
 - 将设备挂载到某个挂载点收集报错信息。
 - 执行**dd**命令对分区的前10Mi进行诊断，收集文件系统结构信息。
 - 收集主机日志“/var/log/messages”文件。

收集卷管理故障信息

- 收集SUSE操作系统的LVM故障信息。
 - 以root用户登录服务器。
 - 收集“/etc/lvm/backup”目录中文件以收集LVM配置信息。
 - 收集PV状态信息。
 - 执行pvscan > /home/pvscan命令重新扫描PV，并将结果重定向至文件（pvscan）中，收集“pvscan”文件。
 - 执行pvdisplay > /home/pvdisplay命令并将结果重定向至文件（pvdisplay），收集“pvdisplay”文件。
 - 收集VG状态信息。
 - 执行vgscan命令查看所有VG情况。
 - 执行vgscan > /home/vgscan命令重新扫描VG，并将结果重定向至文件（vgscan）中，收集“vgscan”文件。
 - 对故障VG执行vgdisplay -v vname > /home/vgdisplay_vname命令，并将结果重定向至文件（vgdisplay_vname）中。收集“vgdisplay_vname”文件。其中vname是故障的VG名称。
 - 收集LV状态信息。
 - 执行lvdisplay > /home/lvdisplay命令，并将结果重定向至文件（lvdisplay）中。收集“lvdisplay”文件。

收集数据库故障信息

- 收集Oracle故障信息
 - 收集Oracle的alert日志文件。将Oracle启动到nomount或者mount状态，执行**show parameter background_dump_dest**命令；在查询得到的路径下获取alert日志文件。
 - Oracle 10g：在“\$ORACLE_BASE/admin/dbname/bdump”路径下获取alert日志文件。
 - Oracle 11g：在“\$ORACLE_BASE/diag/rdbms/dbname/inst_name/trace”路径下获取alert日志文件。
 - **可选**：收集Oracle ASM的alert日志文件。
 - Oracle 10g：Oracle用户下，在“\$ORACLE_BASE/admin/+asm/bdump”路径下获取alert日志文件。
 - Oracle 11gR1：Oracle用户下，在“\$ORACLE_BASE/diag/asm/+asm/trace”路径下获取alert日志文件。
 - Oracle 11gR2：grid用户下，在“\$ORACLE_BASE/diag/asm/+asm/trace”路径下获取alert日志文件。
 - **可选**：在ASM或者数据库实例下执行**select group_number,type from v\$asm_diskgroup**命令收集ASM diskgroup信息。

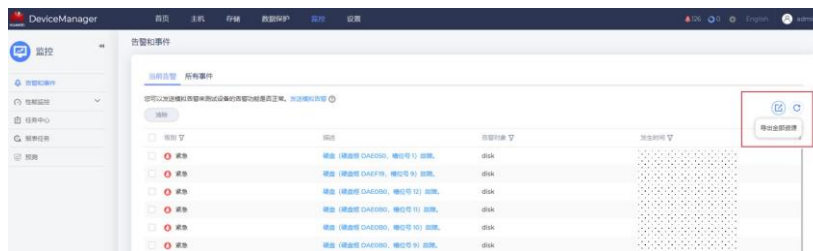
收集存储系统信息

- 通过DeviceManager导出的系统数据，包括：
 - 运行数据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运行数据和状态信息，例如用户信息或LUN的配置信息等。导出的运行数据文件类型是*.txt。
 - 系统日志包括存储设备上的运行数据、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。
 - 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和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。



收集告警和事件信息

- 告警和事件信息用于保存存储设备在运行过程中自动记录的故障和事件。当存储设备出现故障时，维护人员可以通过查看告警和事件信息来定位和解决故障。
- 通过DeviceManager管理界面可以指定级别和发生时间，选择导出的事件信息。
 - “当前告警”中包括“紧急”、“重要”和“警告”三种级别的故障信息。
 - “所有事件”中包括“紧急”、“重要”、“警告”和“提示”四种级别的事件信息，在“当前告警”中的所有告警信息都会导入到“所有事件”中。



利用SmartKit收集信息

导出系统数据

- 存储系统支持的信息收集项包括系统日志、诊断信息、告警信息、历史性能数据、配置信息、电子标签、硬盘、CLI回显、硬盘数据擦除报告、历史故障信息表等。

导出历史性能数据

- 当存储系统已安装SmartKit历史性能监控工具时，您可以通过历史性能监控工具对设备进行历史性能数据收集，以便实时地了解设备的相关信息。

收集设备档案信息

- 当存储系统已安装SmartKit档案信息收集工具时，您可以通过档案信息收集工具收集设备配置信息及部署相关信息，并生成设备档案文件以便归档。

收集主机信息

- 当存储系统已安装主机信息收集工具时，您可以通过主机信息收集工具一键式收集主机或数据库系统的相关信息。

- 利用SmartKit可以收集设备系统数据、档案信息、主机信息等。

目录

1. 故障处理基础
2. 故障处理流程
3. 收集信息与上报故障
- 4. 部件更换**
5. 案例分析

安全操作指引

- 操作设备时，需要遵循警告和安全标识提示的注意事项，避免造成人身伤害或设备损坏。
- 在进行设备的安装、维护等操作时，需要遵循防止静电的安全注意事项，避免造成人身伤害或设备损坏。
- 在进行设备的安装、维护等操作时，需要遵循使用激光的安全注意事项，避免造成人身伤害或设备损坏。
- 安全正确地使用光纤，确保设备正常运行，避免造成人身伤害或设备损坏。
- 在进行设备的安装、维护等操作时，操作工具的使用和放置应遵守工具操作规范，避免操作工具等金属物体造成设备短路。
- 在进行设备的安装、维护等操作时，需要遵循安全带电注意事项，避免造成人身伤害或设备损坏。
- 在进行设备的安装前，需要确保没有凝露，避免设备上电异常。当室外温度与机房温度相差超过15℃时，设备从室外被搬运至机房后，需静置8小时才能开始安装。

部件可更换/拆卸说明

- 存储系统中的可更换单元，如硬盘、控制器、风扇模块、接口模块、电源模块、BBU，均是冗余的，避免单点故障。所有这些可更换单元均可热插拔，并支持在线更换。
 - 存储系统支持对硬盘（非系统盘）、控制器、接口模块、电源模块、系统插框、BBU等部件的独立更换。
 - 存储系统不支持对内存、系统盘和纽扣电池的独立更换，仅支持能对内存、系统盘和纽扣电池的独立拆卸；存储系统不支持对CPU的独立更换和独立拆卸。对于CPU、内存、系统盘和纽扣电池，只能通过更换CPU、内存、系统盘和纽扣电池所在控制器的方式完成相关部件的更换。

CRU和FRU

- CRU
 - CRU（Customer Replaceable Unit）即用户可更换单元，是指用户根据操作指导在设备现场进行更换的部件。
 - 操作对象：用户。
- FRU
 - FRU（Field Replaceable Unit）即现场可更换单元，是指工程师根据操作指导在设备现场进行更换的部件。
 - 操作对象：华为公司或通过华为公司认证的技术支持工程师和维护工程师。

- 客户只可以更换CRU，禁止更换非CRU，避免造成人身伤害或设备损坏。
- 为有效进行FRU更换前后的软、硬件检查，确保部件更换成功，请使用SmartKit中的FRU更换工具，根据操作向导进行部件更换。

识别CRU和FRU

- OceanStor 18510/18810的所有部件均为FRU。对于非OceanStor 18510/18810型号的设备，不同部件的CRU/FRU属性也不同，FRU和CRU列表如下：

CRU	FRU
风扇模块（ OceanStor 6810的4U控制框、4U SAS硬盘框 ） ^a	风扇模块
BBU	系统插框
电源模块	控制器
硬盘模块	线缆
级联模块	接口模块
光模块	辅助散热模块
电模块	管理模块
-	仲裁服务器
a： OceanStor 6810的4U控制框和4U SAS硬盘框的风扇模块为CRU， 其它系统框的风扇模块均为FRU。	

部件更换操作指导



- 部件更换多媒体可针对部件更换的相关步骤进行动态的展示，您可以通过Support网站进行访问：

1. 登录<https://support.huawei.com/enterprise/>。
2. 在“按产品找支持 > 数据存储”版块选择“闪存存储”。
3. 单击对应的产品型号。
4. 单击“视频”页签，在“场景”中选择“故障处理”。
5. 选择部件更换多媒体，在“目录”选择所需观看的操作视频。

- 也可以通过数据存储信息中心进行访问：

1. 登录数据存储信息中心<https://info.support.huawei.com/storage/#/home>。
2. 单击“Info TV”进入视频专区。
3. 在“视频检索”区域，选择“产品类别”。
4. 选择“产品系列”。
5. 选择“标签类别”为“部件更换”，查看所需视频。

目录

1. 故障处理基础
2. 故障处理流程
3. 收集信息与上报故障
4. 部件更换
- 5. 案例分析**

案例分析

硬件模块类

管理软件类

基础存储业务类

增值业务类

多路径软件类

硬件模块类故障概述

硬件发生故障时，一般会伴随硬件模块指示灯异常的现象和产生相关告警，通过查看告警信息，可判断出故障硬件的位置。

- **常见故障**

可在线更换的部件（FRU或CRU部件），如：硬盘、BBU、控制器、接口模块、电源风扇等外置部件。

不可更换的部件，如：控制框背板、级联框背板等内置器件。

- **常用处理方法**

一般CRU/FRU坏件后，需确认备件编码，确认方法：

步骤一：通过日志的配置文件中电子标签，确认FRU备件编码

步骤二：通过命令查询FRU部件，获取SN信息。从SN信息中得到备件编码

步骤三：从备件手册中获取和坏件匹配的备件编码。

- **参考资料**

登陆华为support官网 > 技术支持 > 闪存存储 > 具体产品型号 > 产品文档 > 参考**故障处理或部件更换指南手册**

接口模块故障

- 现象描述: 登陆DeviceManager, 选择“系统 > 硬件 > 设备”。单击图标 , 切换到存储设备的后视图。单击被红色框标识的接口模块, 可以看到该接口模块的“健康状态”为“故障”。在存储设备侧接口模块电源指示灯亮黄色。
- 告警信息: 在DeviceManager的“告警和事件”界面上, 选择“当前告警”页签, 可以看到“接口模块故障”的告警信息。
- 可能原因: 接口模块故障。
- 影响范围及程度: 接口模块故障将导致该模块无法继续承载业务, 业务单链路运行, 增加业务中断的风险。
- 处理步骤:
 - 1.在存储设备侧查看接口模块电源指示灯是否亮黄色。
是 => 2。否 => 保持故障环境联系技术支持工程师进行处理。
 - 2.重新拔插该接口模块。拔插该接口模块后, 查看接口模块电源指示灯状态是否亮绿色
且在DeviceManager中接口模块的“健康状态”显示为“正常”。是 => 处理完毕。否 => 3。
 - 3.请更换该接口模块。更换接口模块的方法请参见对应产品型号的《部件更换》。
 - 4.查看接口模块电源指示灯状态是否亮绿色且在DeviceManager中接口模块的“健康状态”显示为“正常”。是 => 处理完毕。否 => 保持故障环境联系技术支持工程师进行处理。

BBU模块故障

- 现象描述：在存储设备侧BBU运行/告警指示灯亮黄色。登陆DeviceManager，选择“系统 > 硬件 > 设备”。
 - 对于2U控制框，单击图标 ，切换到存储设备的后视图。在存储设备的后视图上，单击被红色框标识的BBU模块，可以看到该BBU模块的“健康状态”为“故障”。
 - 对于4U控制框，在存储设备的前视图上，单击被红色框标识的BBU模块，可以看到该BBU模块的“健康状态”为“故障”。
- 告警信息：在DeviceManager的“告警和事件”界面上，选择“当前告警”页签，可以看到“BBU模块故障”的告警信息。
- 可能原因：BBU故障。
- 影响范围及程度：BBU故障会导致存储系统意外断电时Cache中数据无法完整的保存到保险箱盘中，系统数据丢失。
- 处理步骤：
 - 1.在存储设备侧查看BBU运行/告警指示灯是否亮黄色。
是 => 2。否 => 保持故障环境联系技术支持工程师进行处理。
 - 2.请更换该BBU模块。更换BBU模块的方法请参见对应产品型号的《部件更换》。
 - 3.操作结束后，查看BBU运行/告警指示灯状态是否亮绿色且在DeviceManager中BBU模块的“健康状态”显示为“正常”。是 => 处理完毕。否 => 保持故障环境联系技术支持工程师进行处理。

管理软件类故障概述

管理软件类故障是指用户无法对存储系统进行管理和维护。

- **常见故障**

- 1、通过串口无法激活登陆。
- 2、登陆DeviceManager异常。
- 3、Device manager界面操作异常类故障。

- **常用处理方法**

- 1、串口线缆连接异常或串口连接参数错误造成的，通过拔插串口线缆或者重新设置串口连接参数，通常可以解决此类问题。
- 2、浏览器兼容性问题，请选择指定版本浏览器或者重新设置浏览器配置。

DeviceManager登陆异常故障案例

- 故障现象1描述：使用IE10登陆DeviceManager后，出现以下现象：
 - 页面显示空白或无法单击。
 - 按“F12”打开IE10“控制台”，提示：script7002: XMLHttpRequest: 0x2f7d网络错误。
- 告警信息：无
- 可能原因：浏览器通过HTTPS（HTTP + SSL）协议访问DeviceManager，IE10在证书认证链中可能发生错误，属于IE10固有问题。
- 处理步骤：
 - 单击“F5”刷新该页面。操作完成后，查看故障是否解决。
 - 是 => 处理完毕。
 - 否 => 保持故障环境并联系技术支持工程师进行处理。
- 故障现象2描述：当用户导入或回滚数字证书后，使用Firefox浏览器无法正常进入DeviceManager界面，出现“此连接不受信任”页面，且“添加例外...”按钮无法单击。
- 告警信息：无
- 可能原因：Firefox浏览器缓存未更新。
- 处理步骤：
 - 在Firefox浏览器中，按“Ctrl+Shift+Del”快捷键，在弹出的窗口中按提示清除全部历史记录。操作完成后，重启Firefox，查看故障是否解决。
 - 是 => 处理完毕。
 - 否 => 保持故障环境并联系技术支持工程师进行处理。

DeviceManager界面操作异常故障案例

- 故障现象1描述：
 - DeviceManager页面一直处于加载中或显示异常。
 - 登陆DeviceManager后，进入设备图，在设备图未完全加载完成时，单击其他功能入口切换出去，连续切换多次，浏览器偶尔出现当前TAB页崩溃的现象。
- 告警信息：无
- 可能原因：
 - 由于网络问题，偶然出现页面无法加载。
 - 由于浏览器兼容性问题，偶然出现页面无法加载。
 - 由于浏览器缓存数据异常。
- 处理步骤：
 - 按“F5”重新加载页面或单击当前TAB页面的“重新加载”。操作结束后，查看页面是否加载成功。是 => 处理完毕。否 => 2。
 - 在浏览器中，按“Ctrl+Shift+Delete”快捷键，在弹出的窗口中按提示清除全部。操作结束后，重新登陆并查看页面是否加载成功。是 => 处理完毕。否 => 保持故障环境并联系技术支持工程师进行处理。
- 故障现象2描述：
 - 导入License文件时，提示“License文件导入失败”。
- 可能原因：
 - 网络环境断开，导致License上传文件失败。
 - 上传的License文件被防火墙或系统监控软件拦截。
- 处理步骤：
 - 检查网络环境是否正常。是 => 3。否 => 2。
 - 修复网络环境，重新尝试上传License文件。查看操作是否成功。是 => 处理完毕。否 => 3。
 - 检查系统是否安装有网络防火墙或监控软件，是否拦截了上传的文件以及上传文件同级目录下的文件。是 => 4。否 => 保持故障环境并联系技术支持工程师进行处理。
 - 请关闭防火墙或监控软件，重新尝试上传License文件。查看操作是否成功。是 => 处理完毕。否 => 保持故障环境并联系技术支持工程师进行处理。

基础存储业务类故障概述

基础存储业务类故障时，一般会伴随有告警产生，通过查看告警信息，依据告警修复建议进行处理。

- 常见故障

通过串口无法激活登陆

远端设备增加iSCSI链路失败

应用服务器无法扫描到LUN

删除LUN提示超时

AIX特殊机制导致存储设备第一次对接AIX应用服务器不正常

存储系统无法发现HP-UX服务器提供的启动器

应用服务器无法扫描到LUN

- 现象描述：当用户为应用服务器映射了LUN后，在应用服务器侧无法发现映射的LUN。
- 告警信息：无
- 可能原因：
 - 存储池故障。
 - 链路通信异常。
 - 设备节点文件丢失（适用于UNIX或Linux）。
 - 应用服务器动态检测机制影响扫描LUN（适用于Mac OS X）。
 - 应用服务器没有“主机LUN ID”为0的LUN映射（适用于HP-UX）。
 - 应用服务器自动停止扫描LUN操作（适用于Solaris 9）。
- 影响范围及程度：应用服务器无法扫描到LUN，将导致应用服务器无法使用存储系统提供的资源。

应用服务器无法扫描到LUN处理步骤 (1)

- 原因1：存储池故障。
 1. 请检查存储系统中是否存在“存储池故障”的告警。是 => 2。
否 => 原因2：链路通信异常。
 2. 根据“存储池故障”告警处理存储池故障。
 3. 在应用服务器上重新扫描LUN，查看故障是否解决。是 => 处理完毕。否 => 原因2：链路通信异常。
- 原因2：链路通信异常。
 1. 在DeviceManager界面导航栏单击“系统 > 硬件 > 设备”。查看与应用服务器连接的FC或iSCSI主机端口的“运行状态”是否显示为“未连接”。是 => 2。否 => 原因3：设备节点文件丢失。
 2. 确认是iSCSI组网还是FC组网。FC组网 => 3。iSCSI组网 => 4。
 3. 处理FC组网的链路问题
 4. 处理iSCSI组网的链路问题
 5. 在应用服务器上重新扫描LUN，查看故障是否解决。是 => 处理完毕。否 => 原因3：设备节点文件丢失。
- 原因3：设备节点文件丢失（适用于UNIX或Linux）
 1. 确认应用服务器的操作系统是否为Linux或UNIX。是 => 2。否 => 保持故障环境并联系技术支持工程师进行处理。
 2. 在应用服务器侧查看“dev”目录下，是否有对应的设备节点文件，例如“dev/sdb”。是 => 原因4：应用服务器动态检测机制影响扫描LUN。否 => 3。
 3. 在应用服务器的“Terminal”中执行命令mknod，创建设备节点。
 4. 在应用服务器上重新扫描LUN，查看故障是否解决。是 => 处理完毕。否 => 原因4：应用服务器动态检测机制影响扫描LUN。
- 原因4：应用服务器动态检测机制影响扫描LUN（适用于Mac OS X）。
 1. 确认应用服务器的操作系统是否为Mac OS X。是 => 2。否 => 原因5：应用服务器没有“主机LUN ID”为0的LUN映射（适用于HP-UX）。
 2. 请重新启动Mac OS X应用服务器。
 3. 在应用服务器上重新扫描LUN，查看故障是否解决。是 => 处理完毕。否 => 保持故障环境并联系技术支持工程师进行处理。

应用服务器无法扫描到LUN处理步骤 (2)

- 原因5：应用服务器没有“主机LUN ID”为0的LUN映射（适用于HP-UX）。
 1. 确认应用服务器的操作系统是否为HP-UX。是 => 2。否 => 原因6：应用服务器自动停止扫描LUN操作
 2. 在DeviceManager上查看HP-UX应用服务器的LUN映射，是否存在“主机LUN ID”为0的LUN映射。是 => 保持故障环境并联系技术支持工程师进行处理。否 => 3。
 3. 为HP-UX应用服务器添加“主机LUN ID”为0的LUN映射。
 4. 在应用服务器上重新扫描LUN，查看故障是否解决。是 => 处理完毕。否 => 保持故障环境并联系技术支持工程师处理。
- 原因6：应用服务器自动停止扫描LUN操作（适用于Solaris 9）。
 1. 确认应用服务器的操作系统是否为Solaris 9。是 => 2。否 => 保持故障环境并联系技术支持工程师进行处理。
 2. 否 => 保持故障环境并联系技术支持工程师进行处理。是 => 3。否 => 保持故障环境并联系技术支持工程师进行处理。
 3. 运行cfgadm al命令查询端口的WWPN（World Wide Port Name）号。
 4. 运行cfgadm al -o show_FCP_dev c2::WWPN命令重新启动端口扫描LUN操作。其中show_FCP_dev表示显示LUN信息，c2表示3中查询到的端口号，WWPN表示3中查询到的WWPN号。查看故障是否解决。是 => 处理完毕。否 => 保持故障环境并联系技术支持工程师进行处理。

远端设备增加iSCSI链路失败

- 故障现象描述：
 - 修改iSCSI启动器名称后立即向远端设备增加iSCSI链路，系统提示“通信异常或系统繁忙”。
 - 登陆DeviceManager，单击“数据保护 > 远端设备”，在“iSCSI链路”中没有新增的iSCSI链路详细信息。
- 可能原因：修改iSCSI启动器名称后，30秒内向远端设备增加iSCSI链路导致。
- 处理步骤：
 1. 确认是否30秒内修改过iSCSI启动器名称。是 =>2。否 =>4。
 2. 检查远端设备是否已存在iSCSI链路。是 =>3。否 =>4。
 3. 删除远端设备已经存在的iSCSI链路。
 4. 等待30秒以上，重新增加iSCSI链路。
 5. 操作完成后，查看增加iSCSI链路是否成功。
 - 是 => 处理完毕。
 - 否 => 保持故障环境，并联系技术支持工程师进行处理。

增值业务类概述

- 增值业务类故障多发生在人为配置错误或者配置不合理，导致系统发生运行错误或者引发性能问题。
- **常见故障：**
 - 基础业务配置，比如：RAID级别规划、热备策略规划、LUN参数规划、硬盘域规划等。
 - 高级功能特性配置，比如：远程复制配置、SmartPartition配置等。
 - 业务网络环境配置，比如：主机配置、交换机配置、业务接口卡配置等。
 - 性能调优，比如：Oracle数据库性能调优、VMware VDI性能调优等。
 - 维护类，比如：阵列日常巡检、阵列故障诊断维护等。
 - 主机类，比如：主机文件系统配置管理、主机操作系统兼容性、主机多路径配置等。
- 常用处理方法
 - 增值业务类故障时，一般会伴随有告警产生，通过查看告警信息，依据告警修复建议进行处理。存储主从设备间异常及业务配置错误。

存储系统间复制链路不一致

- 现象描述：在两台互相连接的存储系统间配置链路时：

1. 分别在两台存储系统上登陆DeviceManager，选择“数据保护 > 远端设备”，选中远端设备查看所包含的复制链路。
2. 查看到两台存储系统上的复制链路条数不一致。比如一端可以看到两条复制链路，另一端只能看到一条复制链路。

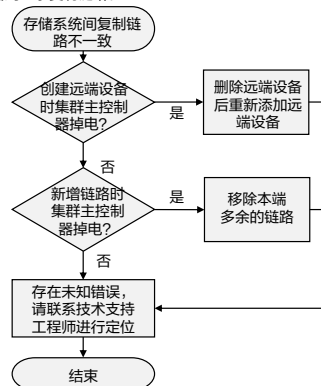
- 可能原因：

- 在创建远端设备的过程中，集群主控制器断电。
- 在远端设备中新增链路时，集群主控制器断电。

- 定位思路：如右图所示。

- 处理步骤：

1. 通过DeviceManager登陆在“远端设备”中查看到的、复制链路数更多的存储系统。
检查其是否仅包含一条复制链路。是 => 2。否 => 3。
2. 删除远端设备后重新添加远端设备。查看两端复制链路配置是否一致，复制链路条数是否相等。
是 => 处理完毕。否 => 保持故障环境并联系技术支持工程师进行处理。
3. 移除本端多余的链路。查看两端复制链路配置是否一致，复制链路条数是否相等。
是 => 处理完毕。否 => 保持故障环境并联系技术支持工程师进行处理。



远程复制一致性组状态失效

- 现象描述：登陆CLI，运行show consistency_group general命令，查看远程复制一致组信息。发现一致性组健康状态为“fault”。
- 告警信息：在DeviceManager的“告警和事件”界面上，选择“当前告警”页签，可能存在“远程复制一致性组不可用”告警。
- 可能原因
 - 人为误操作，在链路断开条件下做了本地删除一致性组。
 - 添加、移除一致性组成员操作失败或一致性组主从切换操作失败。
- 处理步骤
- 原因1：远程复制链路断开。
 - 查看是否存在“复制链路断开”的告警。是 => 断开告警处理；否 => 查看原因2。
- 原因2：添加、移除一致性组成员操作失败或一致性组主从切换操作失败。
 - 主端和从端分别删除一致性组后再重新创建一致性组。
 - 添加一致性组成员。查看故障是否解决。是 => 处理完毕；否 => 保持故障环境并联系技术支持工程师进行处理。

多路径软件类故障概述

多路径软件无法正常使用，将导致存储性能降低。

- **常见故障**

- 1、应用服务器重新启动后无法加载多路径软件
- 2、SUSE应用服务器无法发现多条路径
- 3、Windows操作系统下安装多路径软件时蓝屏

- **常用处理方法**

出现此类问题，一般是由于多路径软件被屏蔽导致的，可能的屏蔽原因：服务器启动项中没有多路径信息、HBA卡驱动自带Failover功能屏蔽了冗余路径，解除屏蔽即可解决此类问题。

同时排查：是否为链路或者交换机故障；并排查是否为控制器本身故障。

Windows操作系统下安装多路径软件时蓝屏

- 现象描述：在安装了Windows操作系统（Windows Server 2003或Windows Server 2008）的应用服务器上安装多路径软件时，意外出错，导致蓝屏。
- 告警信息：无。
- 可能原因：Windows操作系统中没有安装最新的SP（Service Pack）补丁程序。
- 影响范围及程度：Windows操作系统下安装多路径软件时蓝屏将导致多路径软件安装失败。存储系统单链路运行，降低系统性能，增加了业务运行风险。
- 处理步骤：
 1. 重新启动操作系统，进入“Windows高级选项菜单”并选择“最后一次正确的配置”。
 2. 重新登陆系统后，登陆微软官方网站下载最新的Windows SP补丁。
 3. 安装下载的补丁包。
 4. 安装完成后，请重新启动应用服务器，查看故障是否解决。
 - 是 => 处理完毕；
 - 否 => 保持故障环境并联系技术支持工程师进行处理。

未安装多路径软件，控制器B故障导致业务中断

- 现象描述：
 - 控制器B的指示灯全部熄灭；控制器A的指示灯显示无异常。
 - 与控制器B连接的应用服务器读写请求无法下发给存储设备，导致业务中断。登陆DeviceManager，在“性能监控”页面中，控制器B上的主机端口写I/O流量或读I/O流量变为0。
- 可能原因：控制器故障。
- 影响范围及程度：未安装多路径的情况下，控制器故障导致主机业务中断，可以通过手动切换到正常控制器的方式恢复主机业务。
- 处理步骤：
 1. 将控制器B的业务切换到控制器A上。
 2. 对控制器B下电并重新启动。观察控制器B的指示灯是否呈绿色常亮状态。
 - 是 => 3; 否 => 保持故障环境并联系技术支持工程师。
 3. 将原来属于控制器B的业务重新从控制器A切换至控制器B。
 4. 观察业务读写情况。

获取帮助

- 各产品的最新资料都在华为技术有限公司技术支持网站（<https://support.huawei.com/enterprise/>）上发布，该网站提供产品文档的在线检索在线浏览和下载服务。
 - 多媒体：
 - 您可以直接在该网站上申请帐号和密码。使用申请的帐号和密码登陆后，在搜索栏中，输入产品型号，选择搜索栏联想出的路径，进入对应产品型号的文档页面，检索和下载对应版本的多媒体。
 - 产品软件：
 - 您可以直接在该网站上申请帐号和密码。使用申请的帐号和密码登陆后，在搜索栏中，输入产品型号，选择搜索栏联想出的路径，进入对应产品型号的文档页面，然后单击“软件”页签进入软件下载页面，检索和下载对应版本的产品软件。
 - 在线知识库：
 - 您可以直接在该网站上申请帐号和密码。使用申请的帐号和密码登陆后，在搜索栏中，输入产品型号，选择搜索栏联想出的路径，进入对应产品型号的文档页面，然后单击“案例”页签进入案例下载页面，检索、浏览和下载对应版本的案例。

思考题

1. (判断题)根据故障的严重程度和影响范围，存储系统故障问题等级可分为3级。
2. (判断题)当存储系统产生故障告警时，使用替换法处理是最好的方法。

- 参考答案：
 - 正确
 - 错误

本章总结

存储系统故障处理

故障处理基础

故障处理流程

收集信息与上报故障

部件更换

案例分析

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/>
- 华为人才在线学习: <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.

Copyright©2023 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

