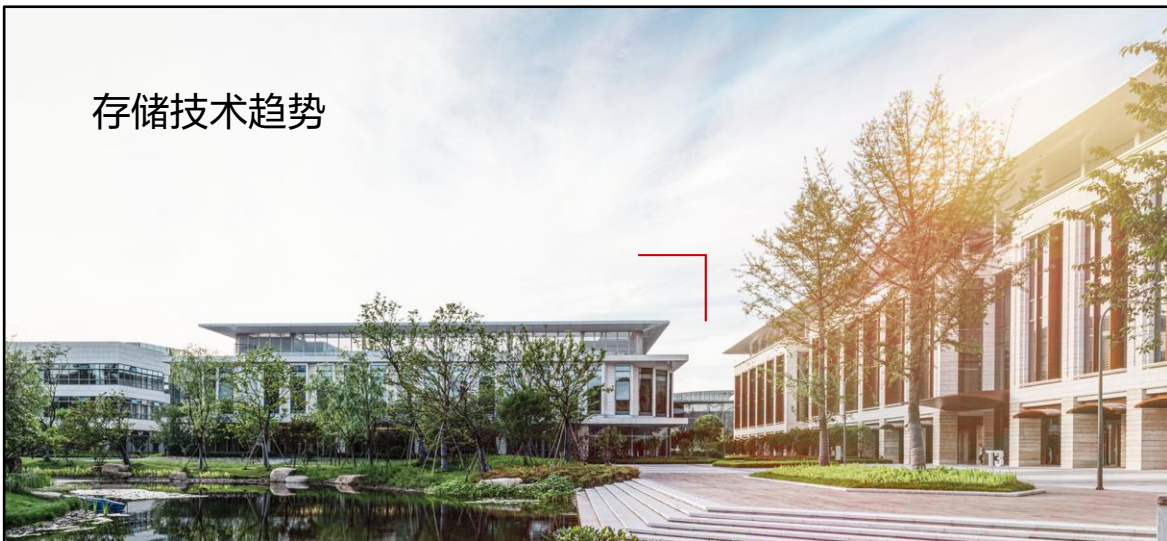


存储技术趋势



前言

- 信息在网络中的传输，需要借助数据作为载体，那么信息和数据的关系是什么，数据存储的作用是什么？本章主要介绍在计算机领域信息和数据的定义，以及它们之间的关系，数据存储的概念、发展历史和发展趋势。

目标

- 学完本课程后，您将能够：
 - 了解信息与数据的定义
 - 了解数据存储的概念
 - 了解数据存储的发展历史
 - 了解数据存储产品的发展趋势

目录

1. 数据与信息
2. 数据存储
3. 存储技术的发展
4. 存储产品的发展趋势

什么是数据

- SNIA (Storage Networking Industry Association, 全球网络存储工业协会) 对数据的定义:

▫ 数据是对所有事物的数字表示。



电子邮件



数字音乐



数字视频



电子书

数据在存储中的表现形式

```
0101001010100010000011
1100011100010001110001
1100000111101010100101
0101001010100101001010
1001010101001010100010
1010010101001010101010
0101010101010010100010
0101001010101010100101
0101001010101010100101
0101001010101010100101
0101010010100101001000
1010101001001010010010
```

- 数据是指所有能输入到计算机并被计算机程序处理的符号，是用于输入计算机进行处理，具有一定意义的数字、字母和符号等的通称。
- DAMA对数据的定义：数据是以文本、数字、图形、图像、声音和视频等格式对事实进行表现。
- DAMA是指The Global Data Management Community，即国际数据管理协会。

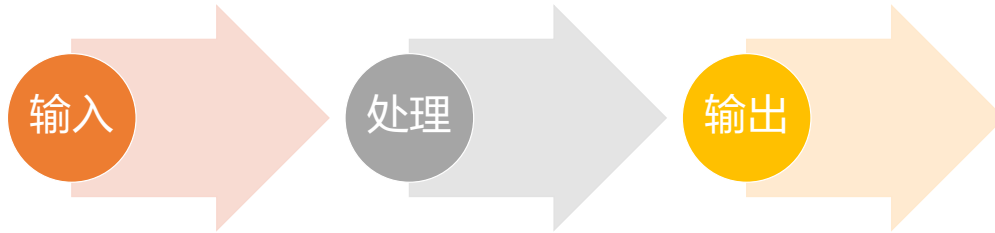
数据的种类



- 根据数据的保存和管理方式，数据分为结构化数据、半结构化数据和非结构化数据。
- 结构化数据
 - 可以使用关系型数据库表示和存储，常表现为二维表。
 - SQL server, MySQL, Oracle。
- 半结构化数据
 - 不符合关系型数据库或其他数据表的形式，但使用相关标记分隔语义元素、或对记录和字段分层。
 - XML, HTML, JSON。
- 非结构化数据
 - 数据结构不规则或不完整，没有预定义的数据模型。
 - 文本、图片、各类报表、图像和音频/视频信息。

数据的处理周期

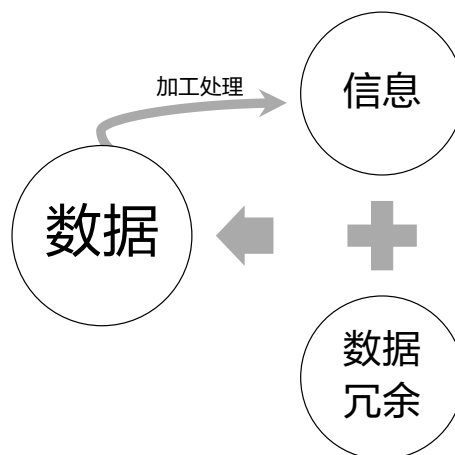
- 数据处理是人或机器对数据进行的重组或重新排序，以增加其特定的价值。数据处理包括以下基本步骤：输入、处理和输出。这三个步骤构成了数据处理周期。



- 输入：将数据以特定的形式输入进行处理，形式将取决于处理机制。例如，当使用电子计算机时，输入数据可以记录在几种类型的输入介质上，例如磁盘、磁带等。
- 处理：将输入数据更改生成更有用的数据形式。例如，可以从时间卡中计算薪水，或者可以从销售订单中计算当月的销售摘要。
- 输出：将收集处理步骤的结果。输出数据的特定形式取决于数据的用途。例如，输出数据可以是雇员的薪水。

什么是信息

- 信息是已经被处理、具有逻辑关系的数据，是对数据的解释。
- 信息是经过处理的数据，其中包括具有上下文，相关性和目的的数据。



数据与信息的关系

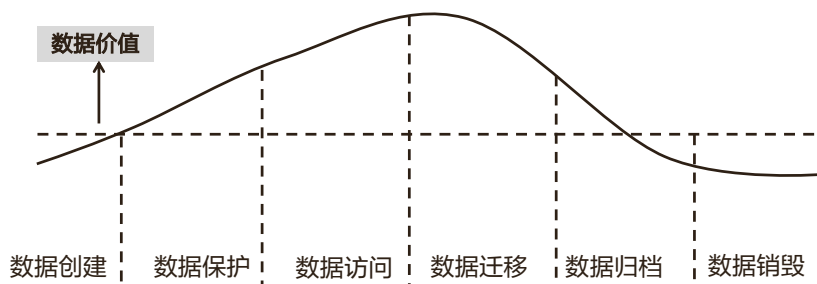
- 数据经过加工处理之后，能够成为信息；
- 而信息需要经过数字化转变成数据才能存储和传输。

对比项	数据	信息
特征	没有特定目的	有价值的、有逻辑的
本质	原始素材	加工后的数据
依赖关系	不依赖于信息	取决于数据
举例	各区域演唱会的门票销售量 气象数据和卫星影像资料	销售报告 天气预报

- 数据是未经处理的原始事实，需要对其进行处理以使其变得有意义，而信息是根据给定要求以有意义的方式处理的一组数据。
- 数据没有任何特定目的，而“信息”具有通过解释数据指定的含义。
- 仅数据本身没有意义，而信息本身就具有重要性。
- 数据从不依赖于信息，而信息则依赖于数据。
- 另一方面，以位和字节为单位的数据以有意义的单位（例如时间，数量等）来衡量信息。
- 数据可以是结构化的，表格数据、图形、数据树，而信息是基于给定数据的语言、想法和思想。

信息生命周期管理

- ILM (Information life cycle management , 信息生命周期管理) 是指从信息产生和初始储存阶段到最后过时被删除时的一套综合管理方法。



- 数据创建阶段：数据从终端产生并保存到存储设备。
- 数据保护阶段：信息生命周期管理将按照数据和应用系统的等级，采用不同的数据保护技术，以保证各类数据和信息得到及时有效的保护。存储系统会提供数据保护功能，如RAID、HA、容灾和权限管理。
- 数据访问阶段：信息必须便于访问，可以在企业的多个业务环节和业务应用之间共享，以提供最大限度的业务价值。
- 数据迁移阶段：IT设备使用过程中，经常需要设备升级和替换，设备中的数据需要从旧设备迁移到新的设备中。
- 数据归档阶段：数据归档系统可以从多个方面支持企业的业务运作，提供交易和决策记录查询。数据的重删压缩功能常会应用在此场景中。
- 数据销毁阶段：数据在一段时期后，没有再继续保存的价值。对没有必要保留或保存的数据进行销毁或回收，从存储，以及数据仓库中清除。

目录

1. 数据与信息
- 2. 数据存储**
3. 存储技术的发展
4. 存储产品的发展趋势

什么是数据存储



- 狭义的存储指的是根据不同的业务，将信息数据存放在具有冗余、保护、迁移等功能的物理媒介。比如软盘、CD、以及DVD和硬盘，甚至磁带。
- 广义存储是一组为企业提供信息存取、保护、优化和利用的整体解决方案，是数据为中心信息架构的支柱。
- 本课程中的数据存储指的是广义的存储。

数据存储系统



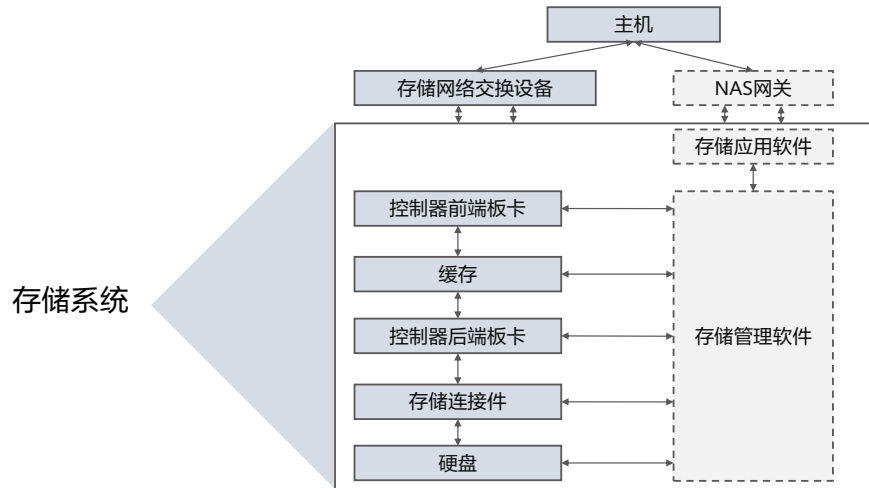
- 容灾解决方案
- 备份解决方案

- 存储管理软件
- 快照、镜像软件
- 备份软件
- 多路径软件

- | | |
|---------------|---------------|
| • 存储设备 | • 连接设备 |
| □ 磁盘阵列 | □ HBA卡 |
| □ 磁带库 | □ 交换机 |
| □ 虚拟磁带库 | □ 连接线缆 |
| □ | □ |

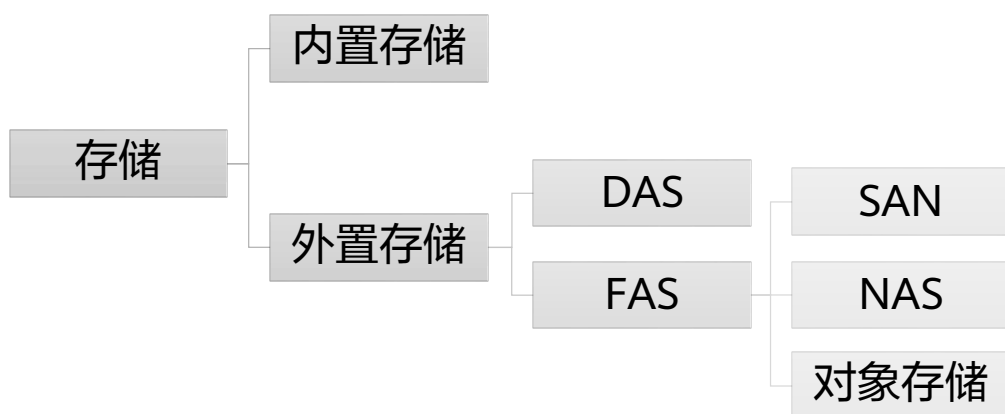
- 存储技术不是一个单独而孤立的技术，实际上，完整的存储系统应该是由一系列组件构成。
- 存储系统分为存储硬件、存储软件以及存储解决方案。硬件部分又分为存储设备和连接设备，主要是指存储设备和存储网络连接设备，比如磁盘阵列、磁带库和FC交换机等。存储软件使存储设备的可用性得到了大大的提高，数据的镜像、复制、自动的数据备份等数据操作都可以通过对存储软件的控制来完成。

存储的物理结构



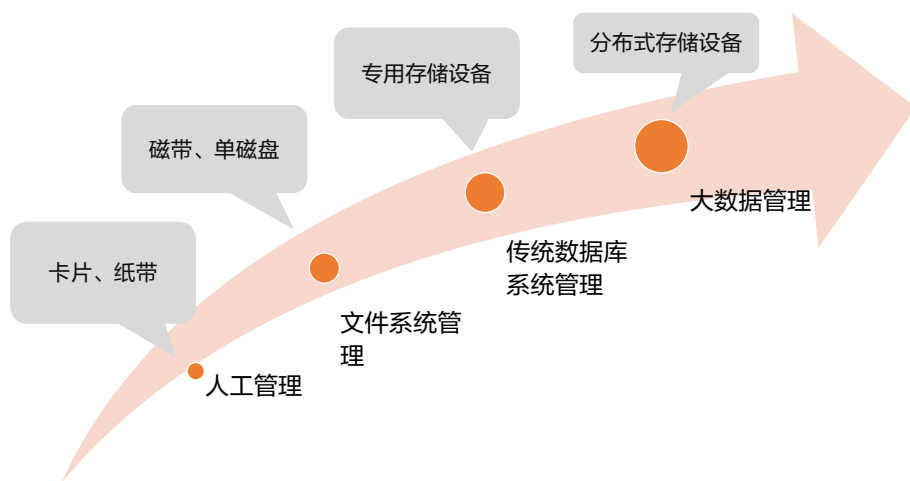
- 一个单一的存储设备包括硬盘子系统、控制子系统、连接子系统和存储管理软件子系统四大部分。
- 存储从物理结构上来看，底层主要是磁盘，其通过相关的连接件如光纤线、串口线等，与存储的内部后端卡 and 控制器相连。
- 存储通过前端板卡与存储网络交换设备连接为主机提供数据访问服务。
- 存储管理的软件是用于配置、监控和优化存储内部的众多子系统和连接件。

数据存储分类



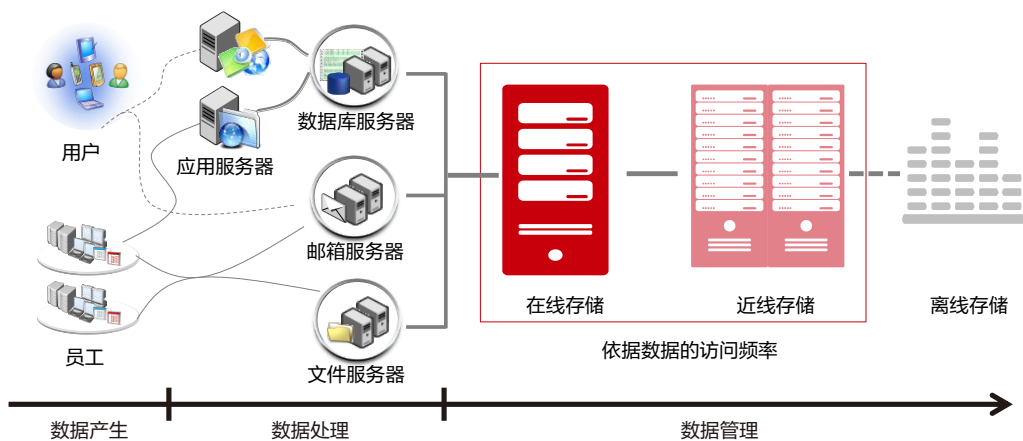
- 存储系统可以根据存储器和主机的位置关系，大致分为内置存储和外置存储。
- 内置存储系统直接与主机总线相连，主要包括CPU运算所需要的高速缓存、内存，与计算机主板直接相连的硬盘、光驱等。容量往往不大，不方便扩展。
- 外置存储根据连接的方式分为：直连式存储（ Direct-Attached Storage，简称DAS ）和网络化存储（ Fabric-Attached Storage，简称FAS ）。
- 网络化存储根据传输协议又分为：网络接入存储（ Network-Attached Storage，简称NAS ）、存储区域网络（ Storage Area Network，简称SAN ）和对象存储（ Object Storage ）。

数据管理技术的演变



- 数据管理是利用计算机硬件和软件技术对数据进行有效的收集、存储、处理和应用的過程。其目的在于充分有效地发挥数据的作用。实现数据有效管理的关键是数据组织。
- 数据管理技术是对数据进行分类、组织、编码、输入、存储、检索、维护和输出的技术。计算机数据存储设备及计算机应用系统的发展，推动了数据库和数据管理技术的发展。按照时间顺序计算机系统中的数据管理经历了人工管理、文件系统管理、传统数据库系统管理、大数据管理4个阶段。

数据存储的应用

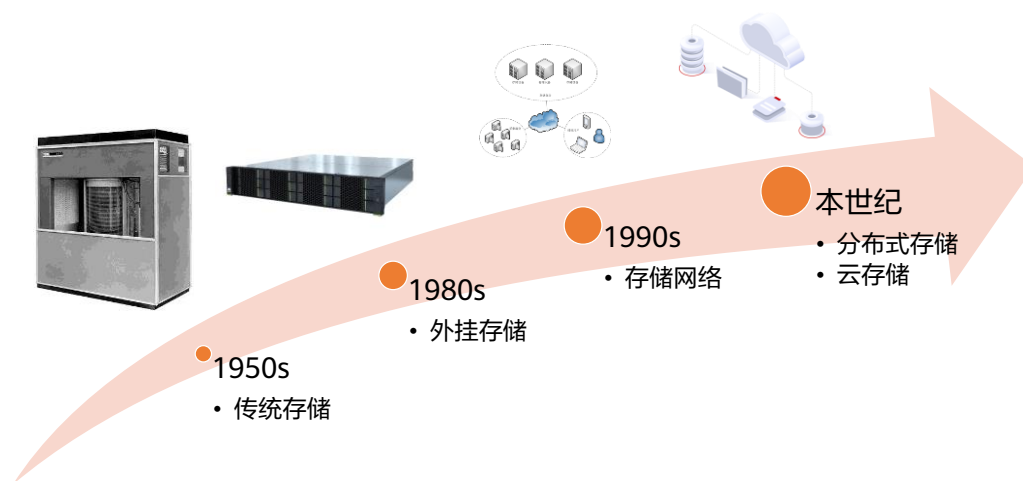


- 由个人和组织产生的数据，通过计算系统的处理保存在数据存储中。
- ICT背景下，存储的主要作用：
 - 数据的存取
 - 数据的保护-安全性
 - 数据的管理

目录

1. 数据与信息
2. 数据存储
- 3. 存储技术的发展**
 - 存储架构的发展
 - 存储介质的发展
 - 接口协议的发展
4. 存储产品的发展趋势

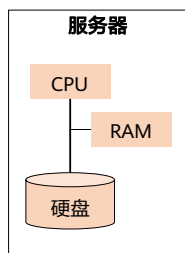
存储架构历程



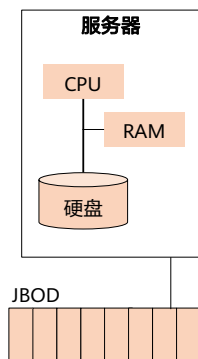
- 存储架构的发展历程经历了传统存储、外挂存储、存储网络、分布式存储和云存储几个阶段。
- 传统存储是由单硬盘组成，1956年IBM发明了世界上第一块机械硬盘，这个硬盘有50个24英寸盘片，容量只有5MB。体积与两台冰箱差不多大，重量超过一吨。被用于当时的工业领域，且独立于主机以外。
- 外挂存储即直连存储，最早的形态是JBOD，仅仅是将一些磁盘串联在一起，被称为JBOD（Just a Bundle Of Disks，磁盘簇），主机看到的就是一堆独立的硬盘，只增加了容量，无法提供安全保障。
- SAN是典型的存储网络，主要使用FC网络传输数据，随后出现了IP存储区域网络。
- 分布式存储采用通用服务器硬件构建存储资源池，更适合云计算的场景。

从硬盘到磁盘阵列

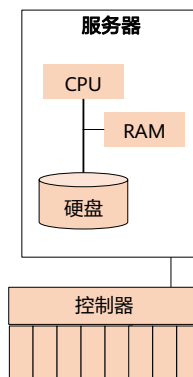
硬盘在服务器内部



早期外挂存储

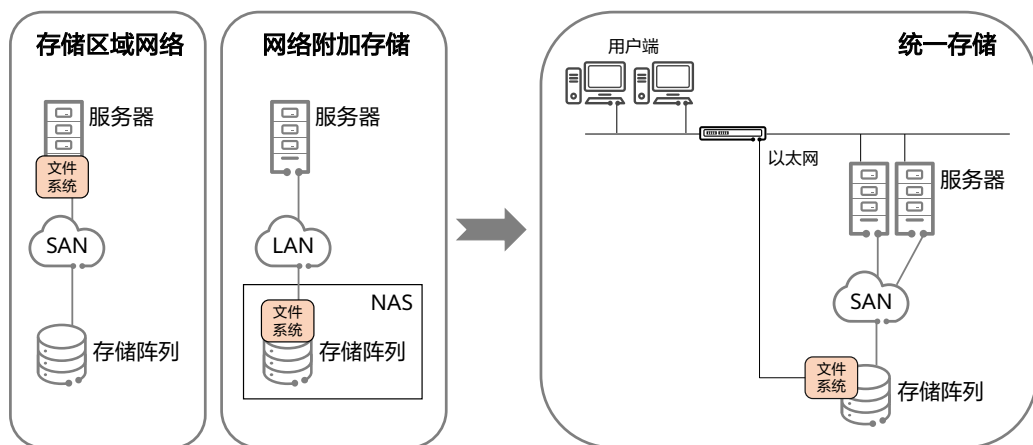


存储阵列



- 硬盘在服务器内部时存在的问题：
 - 槽位有限，容量不足；
 - 单盘存放数据，可靠性差；
 - 硬盘成为系统性能瓶颈；
 - 存储空间利用率低；
 - 数据分散，存储在不同服务器本地。
- JBOD的出现一定程度上解决了槽位有限的问题，后续RAID技术的出现则帮助解决了可靠性和性能的问题。随着时间的发展，外挂存储渐渐发展为存储阵列带有控制器，控制器中包含Cache，支持RAID（Redundant Arrays of Independent Disks，冗余独立磁盘阵列）功能，可以配置专用管理软件，主机看到的是一块大的、高性能的、有冗余特性的硬盘。

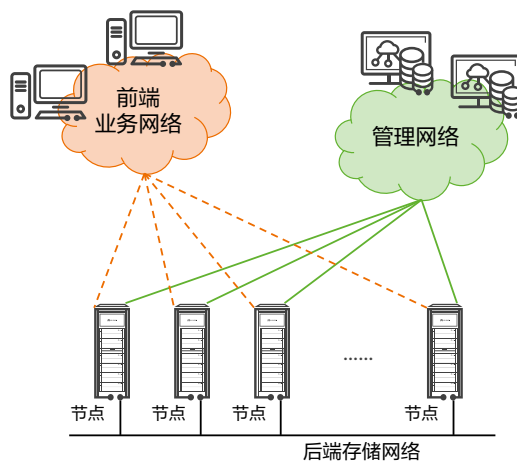
从分离到融合



- DAS (Direct-Attached Storage，直连式存储) 的特点：
 - 数据分散
 - 存储空间利用率低
- 存储发展需求：
 - 数据共享
 - 提升资源利用率
 - 距离延长
- 网络的出现给存储带来了新的生机
 - SAN (Storage Area Network，存储区域网络)：在存储设备与服务器之间组成网络，主要提供块存储服务。
 - NAS (Network Attached Storage，网络附加存储)：使用网络连接服务器与带有文件系统的存储，主要提供文件存储服务。
- 2011年左右，支持SAN和NAS多协议的统一存储迅速成为热点，融合的趋势再次出现：通过整合降低成本，将NAS和SAN结合起来，同时提供数据库与文件共享服务，简化存储管理，提高存储利用率。

分布式存储

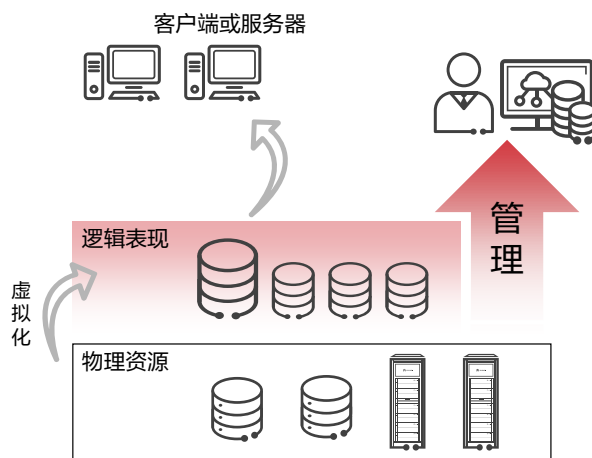
- 通过软件将物理资源组织起来构成高性能逻辑存储池，在保证可靠性的同时提供多种存储服务。
- 一般而言，分布式存储是将数据分散存储在多台独立的设备上，采用可扩展的系统结构、利用多台存储服务器分担存储负荷、利用位置服务器定位存储信息。



- 传统架构扩展能力和灵活性有限。
- 分布式存储架构：
 - 通用硬件、架构统一、存储和网络不被绑定；
 - 性能、容量线性扩展，目前可达数千个节点；
 - 资源弹性伸缩，利用率高。

存储虚拟化

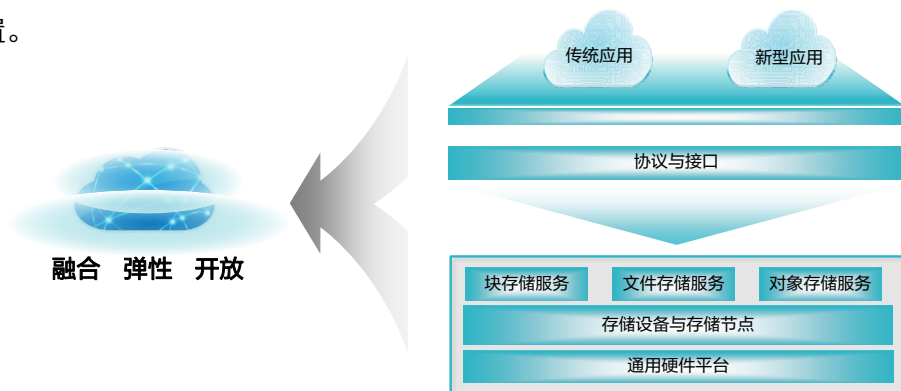
- 存储虚拟化可以将存储设备进行抽象，以逻辑资源的方式呈现，统一提供全面的存储服务。
- 可以在不同的存储形态或设备类型之间提供统一的功能。



- 在物理存储系统和服务器之间增加一个虚拟层，它管理和控制所有存储并对服务器提供存储服务。
- 存储硬件的增减、调换、分拆、合并对服务器完全透明。
- 作用：
 - 隐藏物理设备复杂度；
 - 集成使用现有的功能；
 - 摆脱物理容量的局限。

云存储

- 云存储系统是一个多存储设备、多应用、多服务协同工作的集合体，它使用高度虚拟化的多租户基础设施为企业提供可扩展的存储资源，可以根据组织的要求动态配置。



- 通过集群应用、网格技术或分布式文件系统等功能，将网络中不同类型的存储设备通过应用软件集合起来协同工作，共同对外提供数据存储和业务访问功能。

目录

1. 数据与信息
2. 数据存储
- 3. 存储技术的发展**
 - 存储架构的发展
 - 存储介质的发展
 - 接口协议的发展
4. 存储产品的发展趋势

机械硬盘发展简史

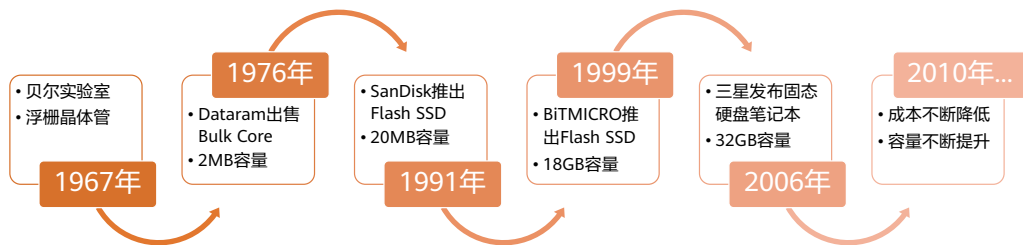
- 容量越来越大，体积越来越小。



- 机械硬盘的发展：
 - 1970年到1991年间，硬盘盘片的存储密度以每年25%到30%的速度增长；
 - 1991年开始，以每年60%到80%的速度增长；
 - 从1997年开始以100%甚至200%的惊人速度提升，这得益于IBM的GMR（Giant Magneto Resisitive，巨磁阻）技术，它使磁头灵敏度进一步提升，进而提高了储存密度。
- IBM 1301：使用空气轴承消除了摩擦力，容量达到了28MB。
- IBM 3340：被装进了盒子，容量达到了30MB，恰好是当时的“温彻斯特来福枪”的口径和填弹量，因此也被称为“温彻斯特”。
- 1992年，1.8寸HDD诞生。

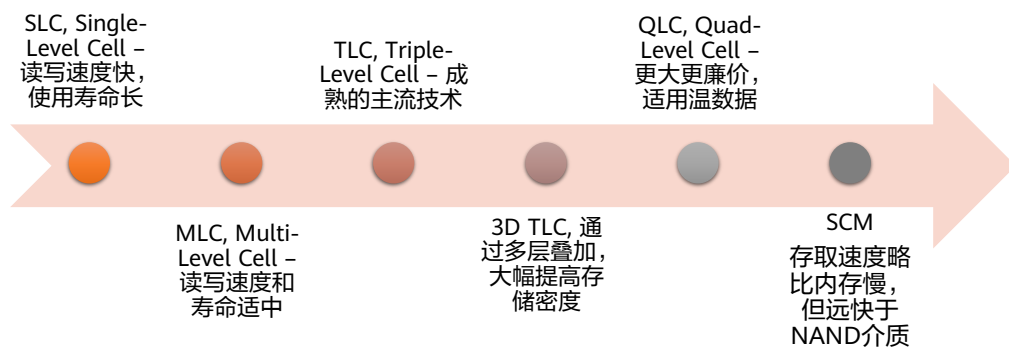
固态硬盘发展简史

- 固态硬盘出现也很早，但是因为20世纪末机械硬盘快速发展而被遗忘。
- 但随着对存取速度要求的提高，固态硬盘又开始繁荣发展。



- 浮栅晶体管，Floating Gate Transistor：由韩裔科学家姜大元和华裔科学家施敏一起发明，是NAND Flash技术的基础。
- Bulk Core：介质使用的是RAM，2MB，8块大电路板，每个板子有18个256K的RAM。

闪存介质发展



- 介质发展趋势
 - 容量更大，速度更快。
 - 介质空间持续扩大。

存储级内存 - SCM

- Storage Class Memory (SCM) 是非易失性内存，该类介质的存取速度略比内存慢，但是远快于NAND类介质。
- 在研的SCM介质种类繁多，但是比较主流的有PCRAM、ReRAM、MRAM和NRAM四大类产品。

	当前存储技术		SCM			
	DRAM	NAND Flash	PCRAM	ReRAM	MRAM	NRAM
非易失性	否	是	是	是	是	是
读延迟	10-60 ns	25 μ s	48 ns	<10 ns	<10 ns	<30 ns
写延迟	10-60 ns	200 μ s	40-150 ns	~10 ns	12.5 ns	50 ns
擦写次数	>10 ¹⁵	10 ⁴	10 ⁸	10 ⁵	>10 ¹⁵	>10 ¹⁴
寻址单位	字节	页	字节	字节	字节	字节

- 2016年问世的存储级内存，结合了DRAM、NAND的性能优势，具有大容量、时延低，相对价格低廉以及非易失存储等特性，被认为颠覆存储系统设计的新一代介质。
- PCRAM, Phase-Change RAM**, 利用特殊合金材料在晶态和非晶态下的导电性差异来表示0或者1数据。
- ReRAM, Resistive RAM**, 通过在上下电极间施加不同的电压，控制Cell内部导电丝的形成和熔断的状态对外呈现不同的阻抗(忆阻器)值来表示数据。
- MRAM, Magnetic RAM**, 磁性随机存储器通过电流磁场改变电子自旋方向来表示不同数据状态。
- NRAM, Nantero's CNT RAM**, 碳纳米管随机存储器采用碳纳米管作为开关，控制电路通断表示不同的数据状态。

目录

1. 数据与信息
2. 数据存储
- 3. 存储技术的发展**
 - 存储架构的发展
 - 存储介质的发展
 - 接口协议的发展
4. 存储产品的发展趋势

接口协议

- 硬盘接口是硬盘与主机系统间的连接部件。
- 接口协议（Interface Protocol）指的是需要进行信息交换的接口间需要遵从的通信方式和要求。



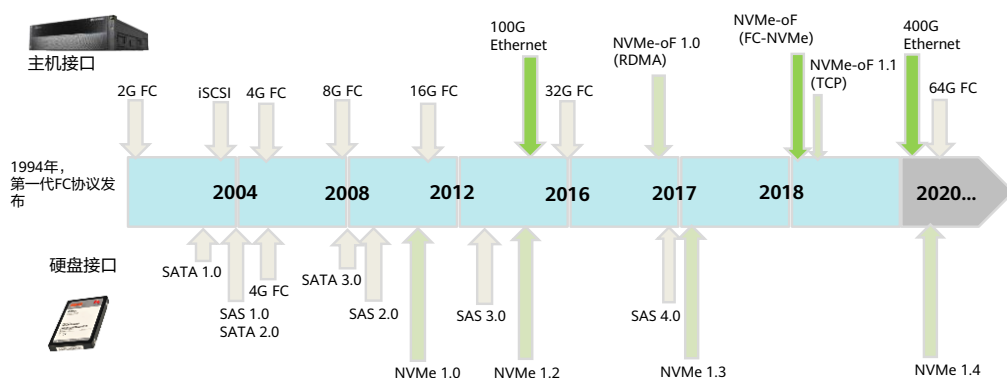
并行接口



串行接口

- 接口的作用是在硬盘缓存和主机内存之间传输数据；不同的硬盘接口决定着硬盘与控制器之间的连接速度。
- 在整个系统中，硬盘接口的性能高低对磁盘阵列整体性能有直接的影响。

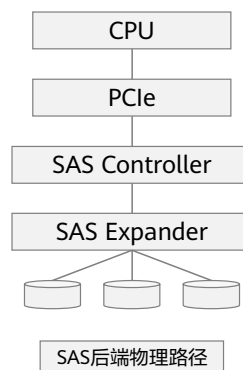
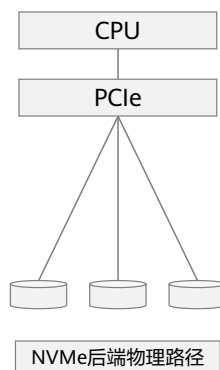
接口协议发展简史



- 存储协议一方面是传输速率的提升，另一方面随着介质从机械硬盘发展到固态存储，协议从SCSI演进到NVMe，包括基于PCIe的NVMe协议和前端接主机网络的NVMe over Farbic（NVMe-oF）协议。
- NVMe Over Fabric（简称NVMe-oF）借助RDMA等超低延时的传输协议，可远程访问SSD，解决了下一代数据中心在横向扩展 (Scale-Out) 时所遭遇的性能、功能、容量三者之间难以取舍的权衡问题。
- 最初NVMe-oF规范发布于2016年，它支持通过光纤通道（Fibre Channel）和远程直接内存访问（RDMA），在RDMA协议框架下，InfiniBand支持融合以太网和广域Internet RDMA协议。
- 2018年11月发布的NVMe-oF1.1规范中，添加了TCP作为架构选项，即RoCE（RDMA over Converged Ethernet），不需要缓存，CPU可直接访问硬盘。

NVMe与NVMe-oF

- NVMe, Non-Volatile Memory Express
 - 提升性能
 - 降低时延
- NVMe-oF, NVMe over Fabrics
 - 潜力：低延迟，高带宽
 - 目的：加速服务器主机和目标存储设备之间的数据传输

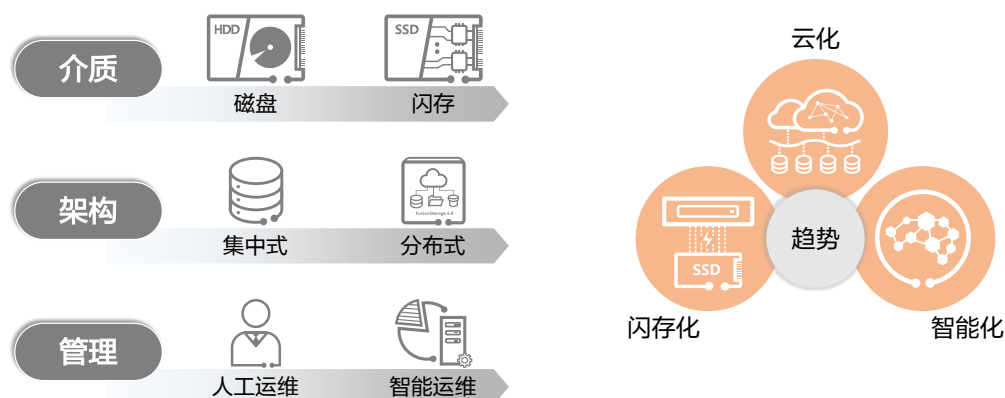


- NVMe (Non-Volatile Memory Express) 是一个SSD控制器接口标准，为基于PCIe接口的SSD设计，它的目标是最大限度释放闪存性能。在数据密集型行业，如生命科学、金融服务、多媒体和娱乐领域为企业级工作负载提供密集计算能力。
- NVMe SSD常用于数据库应用场景，如果利用NVMe速度和低延迟来设计文件系统，则通过NVMe的全闪存阵列可实现卓越的读写性能。可以通过NVMe+全闪存性能实现存储、网络交换、元数据高效通信。

目录

1. 数据与信息
2. 数据存储
3. 存储技术的发展
- 4. 存储产品的发展趋势**

存储产品发展历程



- 介质趋势：从传统磁盘到混合闪存存储、全闪存存储，存储闪存化趋势明显。
- 架构趋势：从集中式存储架构到分布式存储架构，将计算、存储、网络分离，组成共享池，并提供软件来进行统一管理，优化了资源配置，降低了运维成本。
- 管理趋势：从传统人工运维发展到存储管理自动化；从存储管理向数据管理发展。
- 专业存储：存储产品与特定场景结合，带来新的客户价值，如配套容器使用的存储。
- 整体趋势：
 - 闪存化-对应存储介质：闪存颗粒价格逐年下降，机械硬盘逐渐退化成类似磁带地位的介质。
 - 云化-对应存储架构：除了分布式架构带来的弹性可扩展之外，云化还意味着可以减少持有成本。
 - 智能化-对应智能运维：除了软件带来的智能运维功能外，智能化还包括一些带有智能化功能的硬件，比如智能硬盘框。

智能时代来临



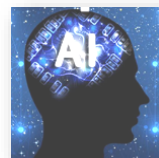
蒸汽时代



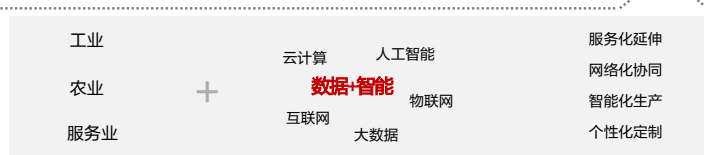
电气时代



信息时代

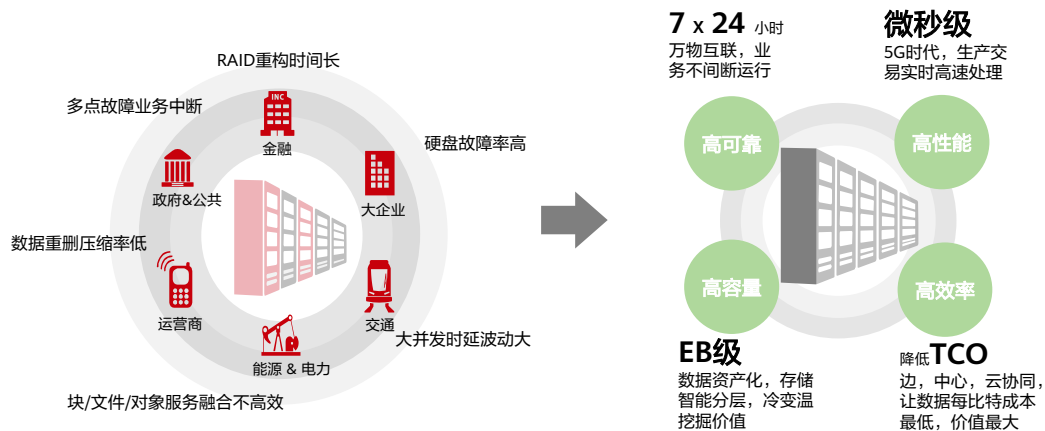


智能时代



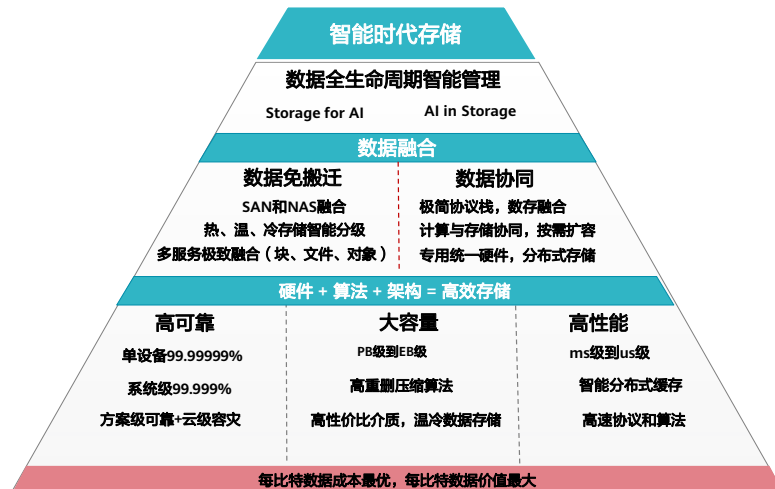
- 从历史上看，人类社会经历了蒸汽技术、电力技术、信息技术三次影响深远的技术革命，近来又在发生以智能化为代表的第四次革命。在这次技术革命中，人工智能、云计算、物联网、大数据等新技术推动包括工业、农业、服务业等许多行业、产业进行大规模的数字化变革，逐渐形成以数据+智能为中心的新型业务，推动服务化延伸、网络化协同、智能化生产和个性化定制等新的变化。人类正快速迈入智能时代。

智能时代对数据存储的新要求



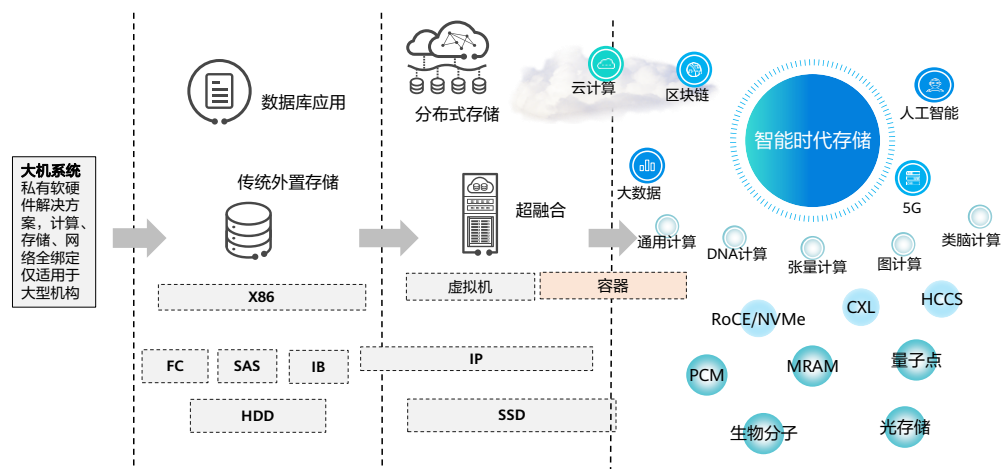
- 数据存储传统架构越来越难以满足严苛的可靠性、性能和效能要求。

智能时代数据存储的特征



- 智能时代数据存储要满足三个主要特征，即智能、融合、高效。
- 智能体现在实现数据全生命周期的智能管理，我们可以称之为Storage for AI和AI in Storage。
- 数据融合主要体现在多协议、多服务的融合，存储智能分级管理，实现数据免搬迁，以及对数据处理和分析的有效协同，采用专有硬件实现极简协议栈、让数据库与存储高效融合，对海量大数据分析进行计算与存储协同，按需扩容。
- 面向业务，要基于硬件+算法+架构，突破冯·诺依曼的瓶颈，利用最新的技术实现数据存储的极致可靠、极致性能以及极致的容量，满足智能时代全新业务的数据存储新需求。

数据存储发展趋势



- 在智能时代，需要重新看待硬件、协议、技术的变化。
- 从IBM大机时代，到x86时代，再到虚拟化时代，存储介质组件走向全闪存化，网络协议走向ALL IP化。
- 智能时代下，基于超高速互联新标准、新协议HCCS（Huawei Cache Coherence System，华为缓存一致性系统）/CXL（Compute Express Link）可以实现CPU、NPU（Neural Processing Unit，神经网络处理器）异构处理器高速互联，RoCE/NVMe支持高速数据传输，容器化技术，加上新的硬件，新的技术，为数据存储提供了丰富的选择。

存储介质发展趋势

蓝光存储



寿命长 可靠性高 单盘100 G+

金纳米玻璃



低功耗 长期稳定存储 单盘10 TB

50年 >>> 600年

DNA存储

- 少量人造DNA就能保存大量数据，并且可以冷冻干燥、运输和储存数据长达几千年。
- DNA作为存储介质的优势：
 - 体积小；
 - 密度大；
 - 稳定性强。
- 现阶段的瓶颈与局限：
 - DNA分子合成成本高；
 - 读取和搜索数据的效率不高。



- 随着科技的发展，硬盘容量越来越大，体积也已经比最初缩小了许多。
- 但相对于DNA而言，单块硬盘的体积仍旧很大，但所存储的信息量又远不及DNA。
- 最初，有团队尝试将数据写入到活细胞的基因组中。但活细胞中的基因会复制，会复制就意味着存在突变的可能；活细胞也会死亡，死亡则意味着数据丢失。因此，后来的团队开始尝试使用脱离细胞的人工合成DNA来存储数据，就现在取得的成果而言，虽然基因存储密度足够大，但读写效率并不高。
- 虽然现阶段DNA分子合成的成本很高，但随着基因测序技术的发展，成本将会有所降低。

原子存储

- 1959年，物理学家理查德·费曼曾提出，如果原子能够被有序排列的话，那么用原子存储信息就是可能的。
- 因原子足够小，原子存储器的单位体积容量也会比现有介质更大。
- 随着科技的发展，有序排列原子成为了现实。
- 现阶段的瓶颈与局限：
 - 原子存储器对运行环境要求严苛。



量子存储

- 现在，电子设备中信息的存储和移动是通过流动电子来实现的。
- 如果电子被光量子取代，计算机内部信息的移动可能以光速发生。
- 尽管存储器的存储效率和存储寿命有所提升，但现阶段量子存储仍然难以广泛适用。
- 已实现的量子存储器难以同时满足下列条件：
 - 高存储效率；
 - 低噪音；
 - 长寿命；
 - 室温下使用。



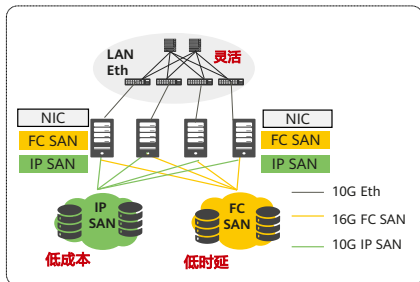
存储网络发展趋势

现状

FC SAN/IP SAN独立两张网

网络成本高：FC专网时延低，价格昂贵；IP SAN成本低，但时延和性能略差

运维代价高：IP/FC SAN短期无法融合，需要专人运维，且无法云网协同

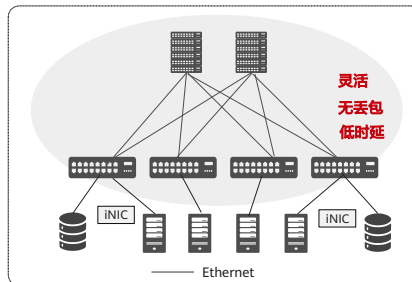


未来

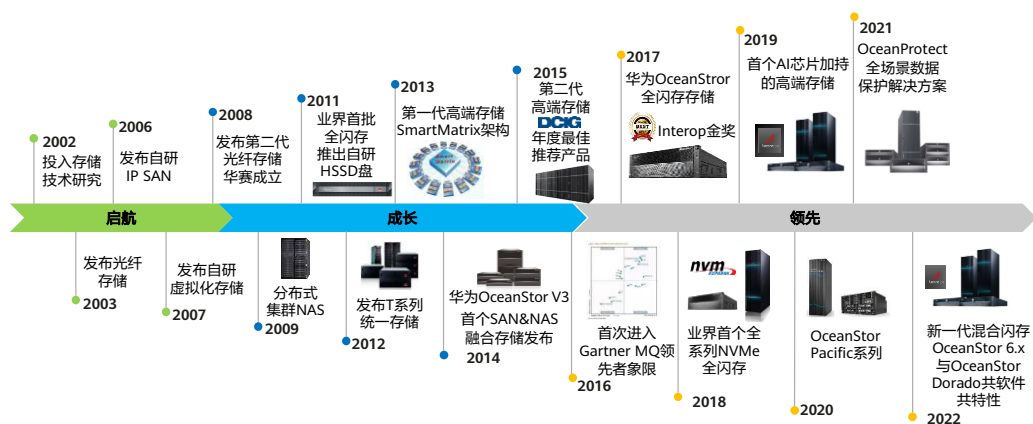
AI Fabric融合网络

降低网络成本：开放以太网承载高性能低时延低成本存储网络

降低运维成本：无需专用技能运维，支持数据中心统一网络管理



华为存储产品发展历程



- 2003 发布第一代光纤存储
- 2006 发布第一代自研IP SAN
- 2007 发布自研FC SAN-S5000
- 2008 发布云存储节点T3000
- 2009 发布第一代分布式NAS
- 2011 首款全固态存储Dorado2100
- 2012 发布T系列统一存储
-

思考题

1. （多选）以下哪些是数据的种类()
 - A. 结构化数据
 - B. 半结构化数据
 - C. 非结构化数据
 - D. 海量数据
2. （多选）下列关于存储的说法哪些是正确的()
 - A. 存储指的就是硬盘
 - B. 存储体系包括存储硬件、存储软件、存储解决方案
 - C. 存储类型包括块存储、文件存储、对象存储
 - D. 文件存储一般用来保存数据库等数据应用的数据

• 参考答案：

- ▣ ABC
- ▣ BC

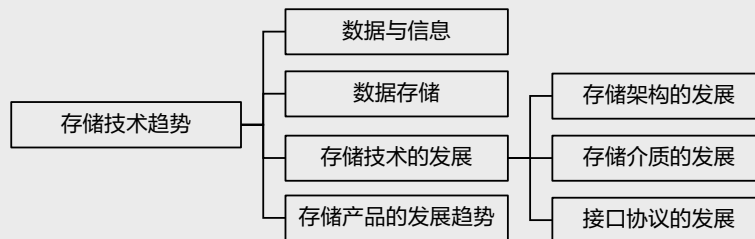
思考题

3. （多选）以下哪些是云存储的特点()
- A. 融合
 - B. 开放
 - C. 弹性
 - D. 纵向扩展
4. （多选）以下哪些是人工智能融入存储的目标()
- A. 简单
 - B. 高效
 - C. 高耗能
 - D. 易用

- 参考答案:

- ABC
- ABD

本章总结



学习推荐

- 华为官方网站
 - 企业业务: <http://enterprise.huawei.com/cn/>
 - 技术支持: <http://support.huawei.com/enterprise/>
 - 在线学习: <http://learning.huawei.com/cn/>
- 热门工具
 - HedEx Lite
 - 网络资料工具中心
 - 信息查询助手

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©2022 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.



智能数据存储系统



前言

- 本章主要对存储系统的组件进行介绍，包括控制框、硬盘框、硬盘、接口模块等组件，以及存储系统的扩展方式。介绍了存储介质、控制框、硬盘框、硬盘及接口模块的结构、工作原理及特点。

目标

- 学习完本章课程后，您将能够：
 - 了解存储产品的形态
 - 了解控制框和硬盘框的作用及组件
 - 了解机械硬盘和固态硬盘的工作原理
 - 了解Scale-up和Scale-out的概念，以及相关的线缆及接口模块

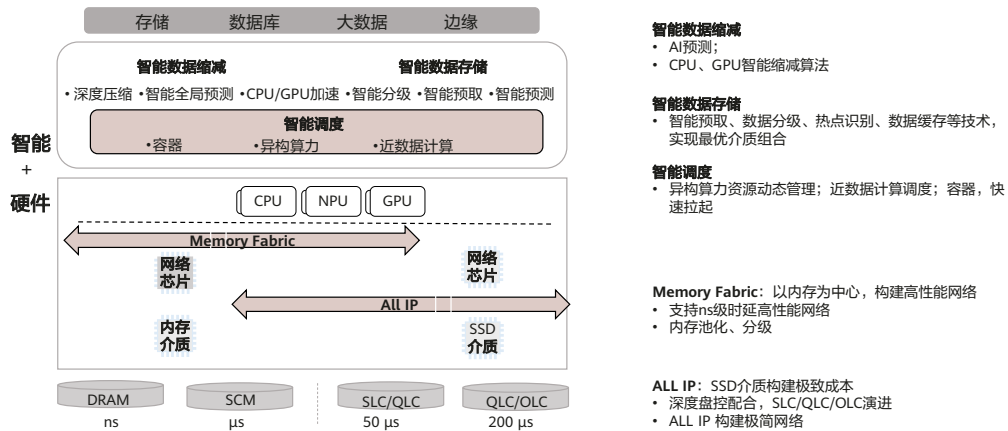
目录

1. 智能数据存储系统介绍
2. 智能数据存储组件
3. 存储系统扩展方式

智能时代存储

- 整个人类社会都在快速迈进智能社会，在这个过程中，数据量迎来爆炸式增长。
- 越来越多的企业已经意识到数据基础设施是智能化成功的关键，而存储是数据基础设施的核心底座。
- 智能时代需要存储智能化
 - Storage for AI
 - 存储要更好支撑企业利用智能技术，满足类似AI智能训练、应用的需要。
 - AI in Storage
 - 存储本身也要应用AI技术，将AI融入存储全生命周期管理，让存储管理、性能、效率、稳定性更加出色。

智能数据存储架构



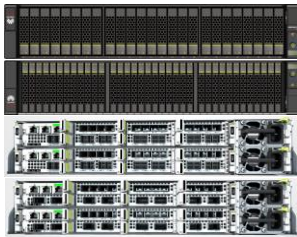
- 华为面向智能时代数据存储打造了基于智能+硬件+算法的创新架构。
 - 基于Memory Fabric打造以内存/SCM为中心的高性能层;
 - 基于ALL IP全闪存高性能的容量层, 提供数据存储的智能分层管理;
 - 面向CPU、NPU和GPU基于创新算法和高速互联协议提供万核级计算资源池化, 智能调度;
 - 面向业务提供基于容器的异构微服务化;
 - 基于数据全生命周期的智能管理系统, 为万物互联的智能时代持续提供符合时代要求的创新存储产品。
- OLC, Octa-Level Cell,每Cell要存储8 bit。

目录

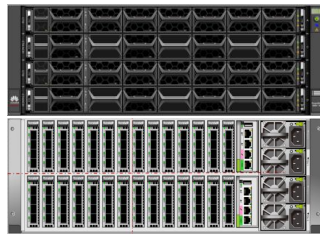
2. 智能数据存储组件

- 控制框
 - 硬盘框
 - 级联模块
 - 硬盘
 - 接口模块

存储产品形态



2U盘控一体



4U盘控分离

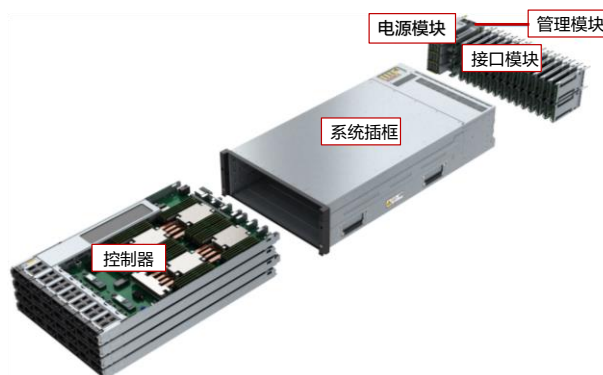


整机柜

说明：以华为OceanStor Dorado V6系列产品为例

控制框形态

- 控制框采用部件模块化设计，主要由系统插框、控制器（内含风扇模块）、BBU模块、电源模块、管理模块和接口模块等组成。



- 控制框是承载了控制器的机框，对外提供存储服务，它是存储系统中的核心组件。
 - 系统插框：通过背板为各种接口模块提供可靠的连接，实现各个模块之间的信号互连与电源互连。
 - BBU模块：能够在系统外部供电失效的情况下，提供后备电源支持，以保证存储阵列中业务数据的安全性。
 - 管理模块：主要提供管理接口，包括管理网口、串口和维护网口。
 - 电源模块：PSU（Power Supply Unit）电源模块为交流电源模块，可以支持控制框在最大功耗模式下正常运行。
 - 接口模块：IM，承载业务或管理接口的现场可更换模块。

控制框正视图详解



2U 盘控一体 控制框



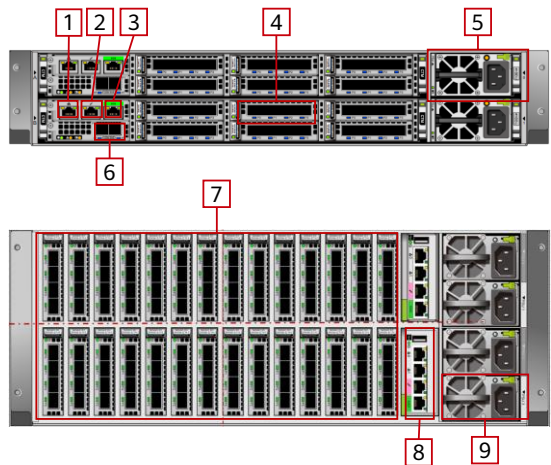
4U 盘控分离 控制框

说明：以华为OceanStor Dorado V6产品为例

图标	右挂耳图标说明
	框ID指示灯
	框定位指示灯 1. 蓝色，闪烁：控制框正在定位。 2. 灭：控制框未定位。
	框告警指示灯 1. 琥珀色，亮：控制框出现告警。 2. 灭：控制框正常运行。
	电源指示灯/电源按钮

- Huawei OceanStor Dorado V6控制框：
 - 2U盘控一体内置25*2.5寸硬盘；
 - 2U盘控一体内置36*PALM硬盘（NVMe）；
 - 4U盘控分离控制框。

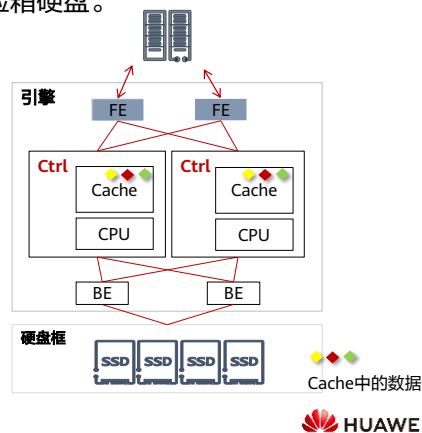
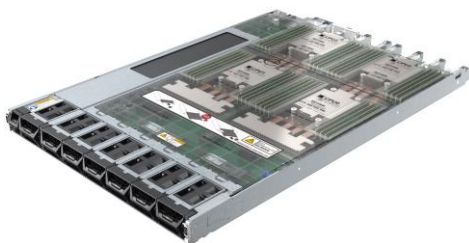
控制框后视图详解



编号	说明
1	管理口
2	维护口
3	串口
4	接口模块
5	电源、BBU一体模块
6	SAS级联口
7	接口模块
8	管理板
9	电源

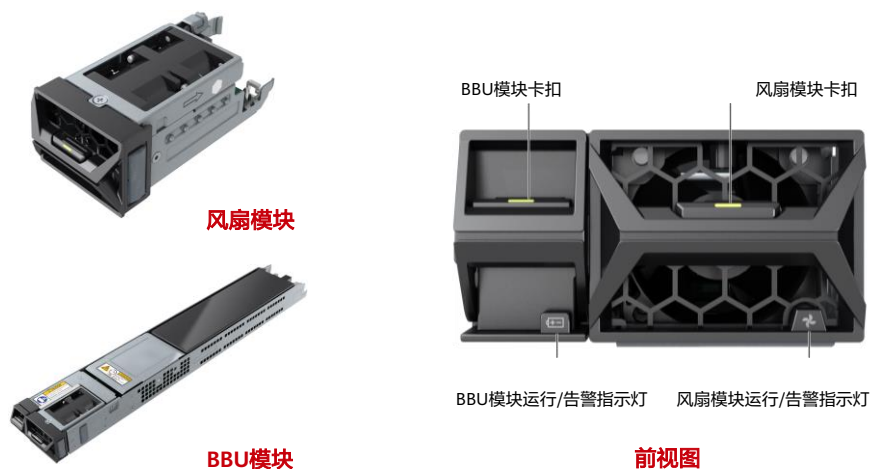
控制器

- 控制器是设备中的核心部件，主要负责处理存储业务、接收用户的配置管理命令、保存配置信息、接入硬盘和保存关键信息到保险箱硬盘。



- 控制器，Ctrl：存储系统中承载业务处理的核心模块。
 - 控制器上的CPU和Cache共同实现对来自主机系统I/O请求的操作和对磁盘阵列的RAID管理。
 - 每个控制器可内置多个硬盘，用于保存存储系统的数据和系统掉电后Cache中的数据。控制器间的硬盘互相冗余。
- 前端端口，FE（Front End）：用于应用服务器和存储设备进行业务通信，即对主机I/O进行处理。
- 后端端口，BE（Back End）：负责连接控制框和硬盘框，是数据从硬盘读写的通道。
- 缓存（Cache）：是硬盘控制器上的一块内存芯片，具有极快的存取速度，它是硬盘内部存储和外界接口之间的缓冲器。
- 引擎（Engine）：引擎（Engine）是电子平台上开发程序或系统的核心组件，通常是一个程序或一套系统的支持部分。

BBU模块和风扇模块



- 风扇模块为系统提供散热功能，可以支持控制框在最大功耗下的正常运行。
- BBU能够在系统外部供电失效的情况下，提供后备电源支持，以保证存储阵列中业务数据的安全性。
- 在系统电源输出正常时处于备份状态，当外部电源断开时，BBU能够继续给系统供电。BBU支持失效隔离，当BBU出故障时不会影响系统的正常运行。当系统掉电时BBU能确保系统将Cache数据写入控制器内置硬盘，保证数据不丢失。在系统外部电源故障恢复时，驱动软件将保存在控制器内置硬盘上的数据读出恢复至Cache中。使用锂电池备电的系统，都需要通过对电池进行充放电来刷新和检测电池容量，用于提前发现电池部件在网长时间不使用导致容量衰减，无法满足系统备电应用，导致数据备份失败的问题，提升系统掉电数据保护的可靠性。

保险箱盘



说明：以华为OceanStor Dorado V6产品为例

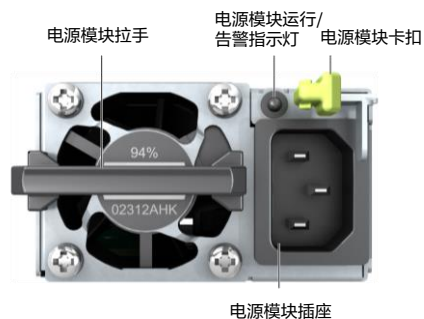
14 Huawei Confidential



- 保险箱盘用于保存存储系统的数据、系统配置信息和日志、系统掉电后Cache中的数据。
- 存储系统中存在两种保险箱盘：内置的保险箱盘和外置的保险箱盘。
- 内置的保险箱盘：
 - Huawei OceanStor Dorado V6每个控制器内置1块或2块SSD硬盘作为保险箱盘，详细配置请参考产品文档。
- 外置的保险箱盘：
 - 存储系统会自动选择系统中的4块硬盘规划为保险箱盘，每块硬盘各取2GB空间组成RAID1，保险箱盘上的其他容量可以用来存储业务数据。当保险箱盘故障时，为保障冗余性，系统会自动从其余硬盘中选择状态正常的盘来替代故障保险箱盘。

电源模块

- 电源模块为交流电源模块，可以支持控制框在最大功耗模式下正常运行。



说明：以华为OceanStor Dorado V6产品为例

- 4控制器存储产品每个4U控制框有4个电源模块（PSU，Power Supply Unit），分别为PSU 0、PSU 1、PSU 2、PSU 3。分为两个电源平面，PSU 0与PSU 1为一个电源平面且互为冗余，为A、C控制器供电。PSU 2和PSU 3为另外一个电源平面且互为冗余，为B、D控制器供电。为保障电源模块供电的可靠性，建议PSU 0和PSU 2插入一个PDU，PSU 1和PSU 3插入另外一个PDU。
- 2控制器存储产品每个控制框有2个电源模块（PSU，Power Supply Unit），分别为PSU 0、PSU 1，组成一个电源平面且互为冗余，为A、B控制器供电。为保障电源模块供电的可靠性，建议PSU 0和PSU 1分别插入不同的PDU。

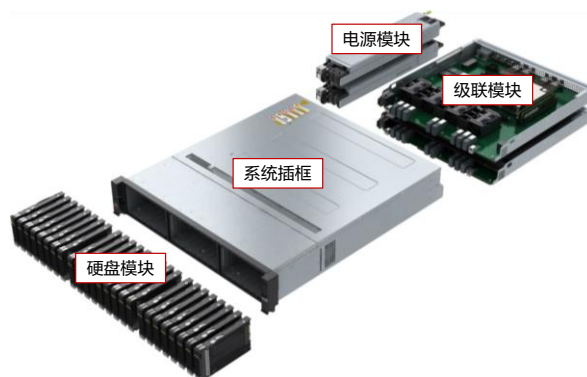
目录

2. 智能数据存储组件

- 控制框
- 硬盘框
- 级联模块
- 硬盘
- 接口模块

硬盘框形态

- 硬盘框采用部件模块化设计，主要由系统插框、级联模块、电源模块和硬盘模块等组成。



- 系统插框：通过背板为各种接口模块提供可靠的连接，实现各个模块之间的信号互连与电源互连。
- 级联模块：通过级联端口来级联控制框和其他硬盘框，实现与控制框和其他硬盘框的通信，是控制框与硬盘框之间、硬盘框与硬盘框之间进行数据传输的连接点。
- 电源模块：可以支持硬盘框在最大功耗模式下正常运行。
- 硬盘模块：为存储系统提供存储容量，可作为系统保险箱，并且可以实现业务数据、系统数据和缓存数据的存储作用。

硬盘框正视图



2U 25盘位 智能SAS硬盘框



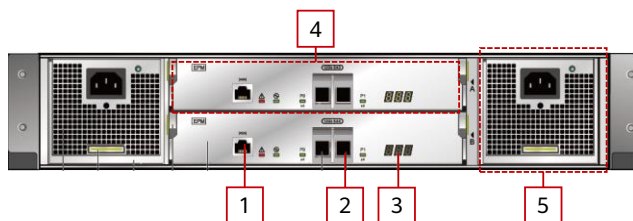
2U 36盘位 智能NVMe硬盘框

说明：以华为OceanStor Dorado V6系列产品为例，
如图为2U SAS硬盘框和2U智能NVMe硬盘框的正视图。

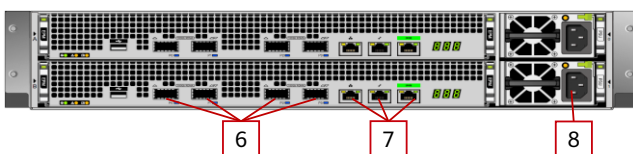
图标	图标说明
	硬盘框ID指示灯
	硬盘框定位指示灯 1. 蓝色，闪烁：硬盘框正在定位。 2. 灭：硬盘框未定位。
	硬盘框告警指示灯 1. 黄色，亮：硬盘框出现告警。 2. 灭：硬盘框正常运行。
	硬盘框电源指示灯 1. 绿色，亮：硬盘框已上电。 2. 灭：硬盘框未上电。
	硬盘框电源指示灯/电源按钮 1. 硬盘框电源按钮无效，即不支持上下电功能。

- 华为OceanStor Dorado V6 系列全闪存存储系统提供了三种硬盘框：SAS硬盘框、智能SAS硬盘框和智能NVMe硬盘框。
- 硬盘框上电后，可以通过观察指示灯初步检查硬盘框当前的工作状态。

硬盘框后视图



2U 智能SAS硬盘框



2U 智能NVMe硬盘框

编号	说明
1	串口
2	Mini SAS HD级联端口
3	硬盘框ID显示器
4	级联模块
5	电源模块
6	板载级联口
7	板载管理口
8	电源模块

说明：以Huawei OceanStor Dorado V6系统为例，如图为2U智能SAS硬盘框和2U智能NVMe硬盘框的后视图。

- 智能硬盘框对外提供RoCE（RDMA over Converged Ethernet）接口用于和控制框的互联，内部通过配置鲲鹏CPU和DDR内存获得了计算能力。
- 智能硬盘框提供智能SAS硬盘框（2U25盘位）和智能NVMe硬盘框（2U36盘位）。

目录

2. 智能数据存储组件

- 控制框
- 硬盘框
- 级联模块
- 硬盘
- 接口模块

级联模块



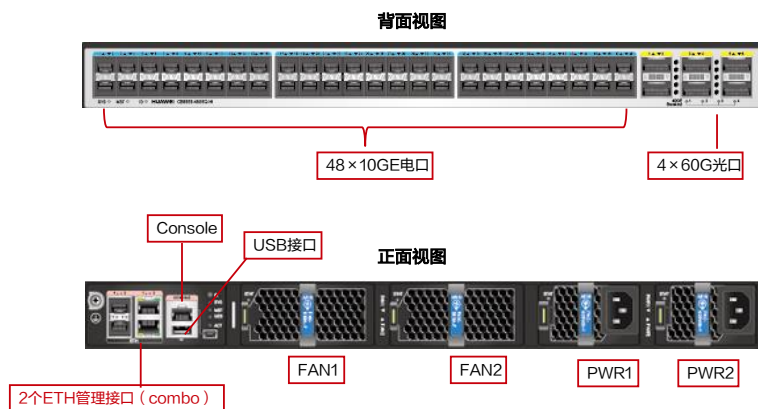
2U SAS硬盘框中级联模块
通过级联端口级联控制框



2U 智能NVMe硬盘框中级联模块
通过级联端口级联控制框与其他硬盘框

- 每个级联模块提供1个级联端口“P0”和1个级联端口“P1”。级联模块通过级联端口来级联控制框或硬盘框，实现与控制框或硬盘框的通信，是控制框与硬盘框之间进行数据传输的连接点。

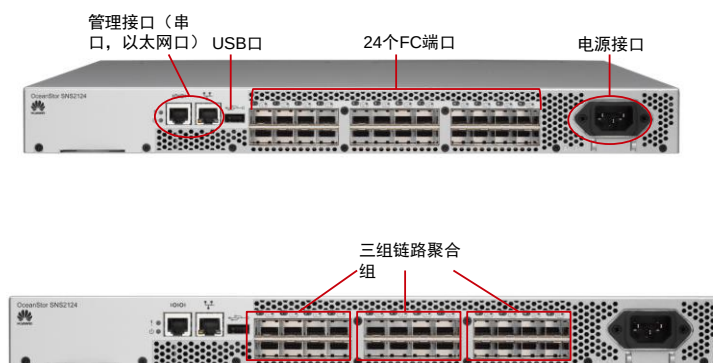
CE交换机



说明：CE交换机以华为CE6800系列交换机为例

- 华为CloudEngine 系列盒式交换机是面向数据中心推出的新一代高性能、高密度、低时延以太网交换机，采用灵活的前后/后前风道设计。
- 可以应用于IP SAN组网，分布式存储组网场景。

FC交换机



说明：FC交换机以华为SNS2124产品为例

- 光纤交换机是一种高速的网络传输中继设备，它较普通交换机而言采用了光纤电缆作为传输介质。光纤传输的优点是速度快、抗干扰能力强。
- 可应用于FC SAN组网场景。

设备线缆



1. 串口线缆



2. mini SAS HD电缆



3. mini SAS HD光缆



4. 100G QSFP28线缆



5. 25G SFP28线缆



6. MPO-4*DLC光纤



7. 光纤

- 串口线缆用来连接设备的系统串口到维护终端。
- mini SAS HD线缆：存储系统中mini SAS HD线缆用于连接级联端口，连接控制框到硬盘框。mini SAS HD线缆分为mini SAS HD电缆和mini SAS HD光缆。
- 100G QSFP28线缆：100G QSFP28线缆主要用于多控间直连组网或连接智能硬盘框。
- 25G SFP28线缆：25G SFP28线缆主要用于前端主机组网。
- 光纤：设备通过光纤与光纤交换机进行通信。光纤的一端可连接FC（Fibre Channel）HBA（Host Bus Adapter）卡，另一端连接光纤交换机或存储。光纤两端都为LC（Lucent Connector）连接器。存储到光纤交换机可以使用MPO-4*DLC光纤线连接，MPO-4*DLC光纤是8Gb FC接口模块（8端口）和16Gb FC接口模块（8端口）专用线缆。

目录

2. 智能数据存储组件

- 控制框
- 硬盘框
- 级联模块
- 硬盘
 - 机械硬盘
 - 固态硬盘
- 接口模块

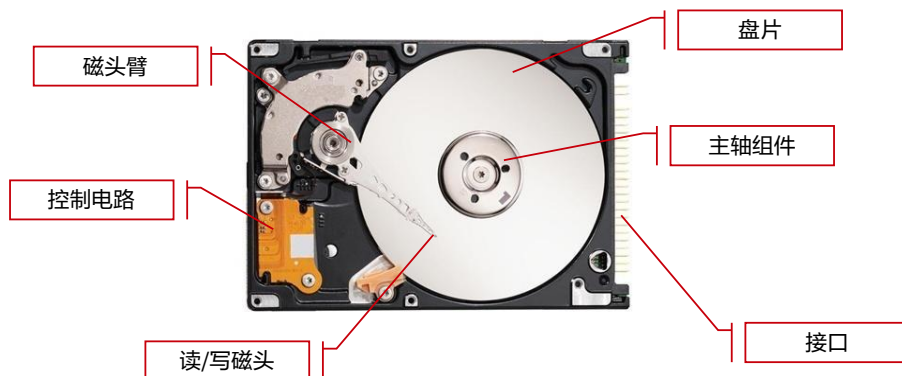
硬盘类型



- 硬盘种类有多种，可以按照接口类型、尺寸大小、应用场景、硬盘实现等多个维度分类。

机械硬盘结构

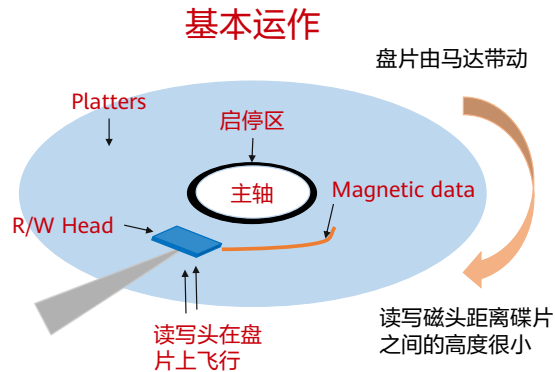
- 机械硬盘的结构大致包括：盘片、磁头臂、读/写磁头、主轴、硬盘接口和控制电路等。



- 硬盘结构组件：
 - 盘片：盘片被磁性材料覆盖（两个表面都涂覆磁性材料），盘片上的磁性粒子被极化以表示一个二进制信息单元（或比特）。
 - 读/写磁头：负责向盘片读取和写入数据。通过磁头改变盘片表面上极细微的磁性粒子簇的N、S极性加以储存。
 - 磁头臂：带动读/写磁头，将磁头移动到指定的位置。
 - 主轴：马达电机和轴承在其下方，转动盘片，将盘片上的指定位置移动到读/写磁头下。
 - 控制电路：控制硬盘的速度、磁头臂的移动，向磁头下发命令等。

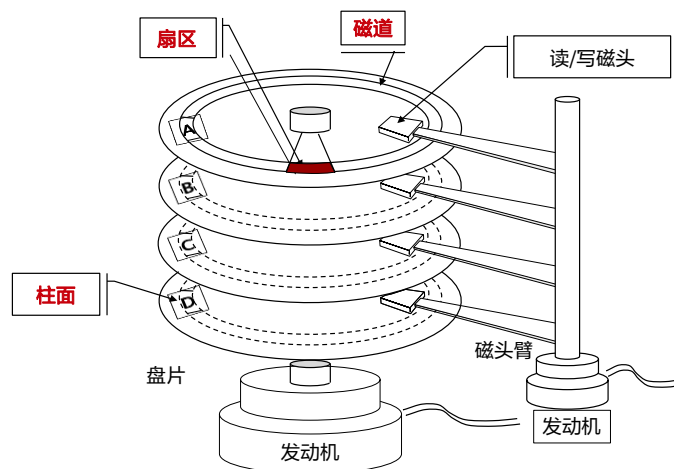
机械硬盘工作原理

- 盘片 Platters
- 磁道 Track
- 柱面 Cylinder
- 扇区 Sector
- 磁头 Head



- 硬盘的每个盘片都有两个读写磁头，分别用于盘片的两个表面的数据的读写。
- 磁头是利用气流飘浮在盘片上，并没有接触到盘片，因而可以在各轨间高速来回移动。若磁头距离盘片太高，则读取的信号会太弱，若距离太低则会磨到盘片表面，因此盘片表面必须相当光滑平整，任何异物和尘埃均会使得磁头摩擦到磁面而造成数据永久性损坏。
- 工作原理：
 - 起初，读/写磁头停靠在盘片在主轴附件的一个特殊区域，启停区。
 - 主轴连接所有盘片，并连接到一个马达上。主轴电机以恒定的速度旋转，带动盘片旋转。
 - 主轴旋转时，读/写磁头和盘片间有个很微小的空气间隙，称磁头飞行高度。
 - 读/写磁头被安装在磁头臂顶端，磁头臂带动磁头移动到需要被写入或取出数据的盘片位置上方。
 - 磁头在硬盘表面以二进制的形式读写数据，读取的数据储存在硬盘的flash芯片中，最后传到程序中运行。

硬盘上的数据组织



- 盘面：硬盘的每一个盘片都有两个盘面，每个盘面都能存储数据，成为有效盘片。每一个有效盘面都有一个盘面号，按从上到小的顺序从0开始依次编号。在硬盘系统中，盘面号又叫磁头号，因为每一个有效盘面都有一个对应的读写磁头。
- 磁道（Track）：磁道是在盘片上围绕在主轴周围的同心环，数据被记录在磁道上。磁道从最外圈向内圈从0开始顺序编号。硬盘的每一个盘面有300~1024个磁道，新式大容量硬盘每面的磁道数更多，通常用盘片上每英寸的磁道数（TPI，也称磁道密度）来衡量盘片上磁道排列的紧密程度。磁道是肉眼看不见的，只是盘面上以特殊形式磁化了一些磁化区。
- 柱面（Cylinder）：在同一个硬盘中所有盘片（包含上下两个盘面）具有相同编号的磁道形成一个圆柱，称之为硬盘的柱面。每个柱面上的磁头由上而下从0开始编号，数据的读写按柱面进行，即磁头读写数据时先在同一柱面内从0磁头开始进行操作，依次往下在同一柱面的不同盘面（即磁头）上进行操作。只有同一柱面所有的磁头全部读写完成后磁头才转移到下一个柱面，因为选取磁头只需通过电子切换即可，而选取柱面侧必须通过机械切换，即寻道。通常硬盘中磁头的位置由柱面号来说明，而不是用磁道号来说明。
- 扇区（Sector）：每个磁道被分为更小的单位，称为扇区，划分扇区的目的是为了使数据存储更加条理化。扇区是硬盘中可以单独寻址的最小存储单元。不同硬盘磁道的扇区数可以不同。通常情况下，一个扇区可以保存512字节的用户数据，但也有些硬盘可以被格式化为更大的扇区大小，如4KB扇区。

硬盘容量和缓存

- 硬盘容量
 - 硬盘容量=柱面数*磁头数*扇区数*扇区大小，单位为MB或GB，影响硬盘容量的因素有**单碟容量**和**碟片数量**。
- 缓存（Cache）
 - 由于CPU和硬盘之间存在巨大的速度差异，为解决硬盘在读写数据时CPU的等待问题，在硬盘上设置适当的高速缓存，以解决二者之间速度不匹配的问题，硬盘缓存与CPU上的高速缓存作用一样，是为了提高硬盘的读写速度。

- 硬盘可以分为单碟盘和多碟盘，前者在盘体内只有一张盘片，后者则有多张。但是硬盘每个时刻只允许一个磁头来读写数据，因此盘片和磁头再多，也不能提高硬盘的吞吐量和IO性能，只能提高容量。

影响硬盘性能的因素



转速

是影响硬盘连续IO时吞吐量性能的首要因素。

寻道速度

是影响硬盘随机IO性能的首要因素。

单碟容量

是影响硬盘性能的一个间接因素。

接口速度

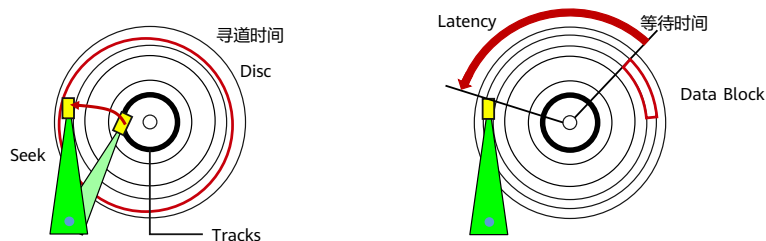
是影响硬盘性能的一个最不重要的因素。

- 影响硬盘性能的因素：

- 转速：指硬盘盘片转动圈数/分钟，单位rpm。读写数据时，磁头不会动，靠盘片的转动来感应读取数据，因此盘片转得越快，数据传输时间就越短。在连续IO情况下，磁头臂寻道次数少，因此要提高吞吐量或IOPS值，转速就是首要影响因素。
- 寻道速度：随机IO情况下，磁头臂需要频繁更换磁道，用于数据传输的时间相比较于换道消耗的时间来说是很少的，因此若磁头臂能够以很高的速度更换磁道，那么就会提升随机IOPS值。
- 单碟容量：单碟容量越高，证明相同空间内的数据量越大，即数据密度越大。在相同转速和寻道速度条件下，具有高数据密度的硬盘会显示出更高的性能。
- 接口速度：目前的接口速度在理论上都已满足硬盘所能达到的最高外部传输带宽。在随机IO环境下，接口速度显得更加不重要，因此瓶颈几乎都在寻道速度上。

平均访问时间

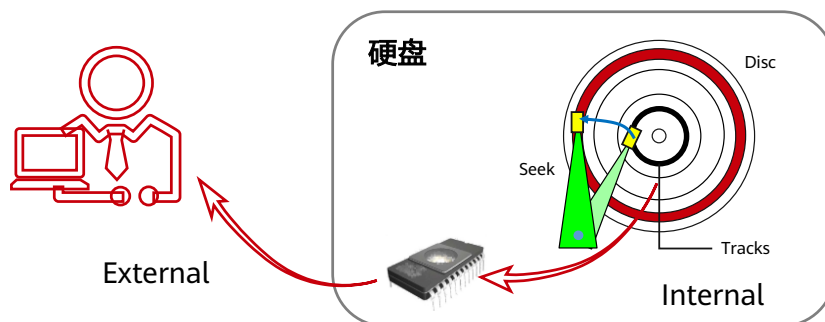
- 平均访问时间由以下两项构成
 - 平均寻道时间（Average Seek Time）
 - 平均等待时间（Average Latency Time）



- 平均寻道时间：指硬盘的磁头从初始位置移动到盘面指定磁道所需的时间，是影响硬盘内部数据传输率的重要参数。这个时间越小越好。
- 平均等待时间：指磁头已处于要访问的磁道，等待所要访问的扇区旋转至磁头下方的时间。平均等待时间通常为盘片旋转一周所需时间的一半，因此硬盘转速越快，等待时间就越短。

数据传输速率

- 数据传输速率（Data Transfer Rate）
 - 内部传输速率（Internal Transfer Rate）
 - 外部传输速率（External Transfer Rate）/接口传输率



- 数据传输率：硬盘的数据传输率是指硬盘读写数据的速度。数据传输率包括外部数据传输率和内部数据传输率两个指标，单位MB/s。
 - 内部传输率：也称为持续传输率（Sustained Transfer Rate），指磁头读写硬盘时的最高速率。不包括寻道以及等待扇区旋转磁头下所耗费的时间的影响。它是一种理想情况，即假设磁头读写的时候不需要换道，也不专门读取某个扇区，只在一个磁道上连续的循环读写这个磁道的所有扇区，此时的速率就叫做硬盘的内部传输速率。
 - 外部传输率：也称为突发数据传输率（Burst Data Transfer Rate）或接口传输率。它指的是系统总线与硬盘缓冲区之间的数据传输率，与硬盘接口类型和硬盘缓存的大小有关。

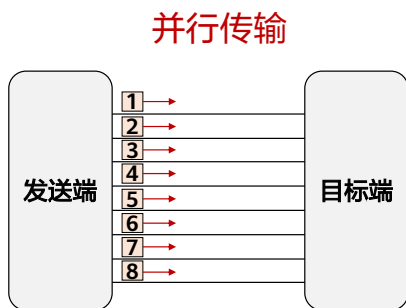
硬盘的IOPS和传输带宽

- IOPS
 - IOPS (Input/Output Per Second)指每秒的输入输出量(或读写次数)，是衡量硬盘性能的主要指标之一。
 - IOPS的计算：由寻道时间、旋转延迟和数据传输时间三部分组成。
- 传输带宽（Throughput吞吐量）
 - 指单位时间内成功传输的数据数量，即传输数据流的速度。例如写入10000个1KB的文件需要10s，则此时传输带宽只达到1MB/s；若写入一个10MB的文件用了0.1s，则此时的传输带宽为100MB/s。

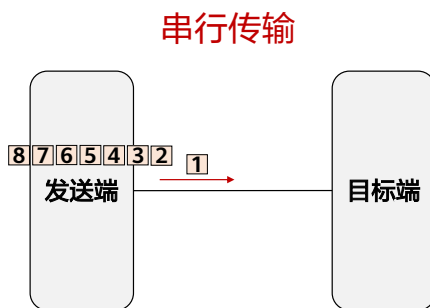
- IOPS计算方法：由寻道时间、旋转延迟和数据传输时间三部分构成。
 - 寻道时间：寻道时间越短，I/O操作越快，目前硬盘的平均寻道时间一般在3-15 ms。
 - 旋转延迟：指盘片旋转将请求数据所在扇区移至读写磁头下方所需要的时间。旋转延迟取决于硬盘转速，通常使用硬盘旋转一周所需时间的1/2表示。如7200 rpm的硬盘平均旋转延迟大约是 $60 \times 1000 / 7200 / 2 = 4.17$ ms，而转速为15000 rpm的硬盘其平均旋转延迟约2 ms。
 - 数据传输时间：指完成传输所请求的数据所需要的时间，它取决于数据传输率，其值等于数据大小除以数据传输率，比如IDE/ATA能达到133 MB/s，SATA II可达到300 MB/s的接口数据传输率。
 - 随机IO情况下，磁头臂需要频繁更换磁道，用于数据传输的时间相对于换道消耗的时间来说是很少的，不在同一个数量级上。因此可忽略数据传输时间。
- 理论上可以计算出硬盘的最大IOPS，即 $IOPS = 1000 \text{ ms} / (\text{寻道时间} + \text{旋转延迟})$ ，忽略数据传输时间。如假设硬盘平均物理寻道时间为3 ms，硬盘转速为7200, 10K, 15K rpm，则硬盘的IOPS理论最大值分别为140, 167和200。

并行传输和串行传输

- 现在有8个数字从1到8，需要传送给对方。传输方式如下：



与对方连接多条线，每条线上传输一个字符。

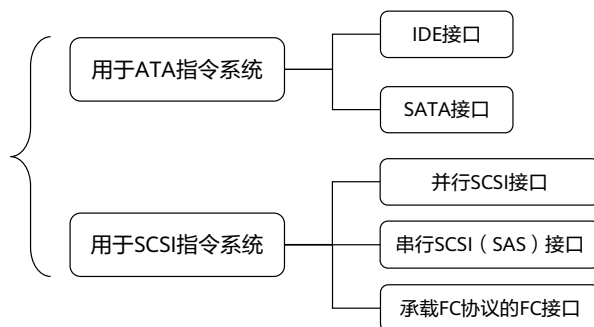


只用一条连线来连接对方，依次在这条线上发送8个数字，需要发送8次才能将数字全部传送到对方。

- 并行传输：
 - 并行传输的传输效率高，但是传输距离不长，传输频率不高；
 - 在长距离上，使用多条线路要比使用一条单独线路昂贵；
 - 长距离传输要求较粗的导线，以便降低信号的衰减，但把它们捆到一条单独电缆中相当困难；
 - 长距离传输时，受导线电阻或其他因素影响，每条线上的数据到达对方的速度不同，从而导致接收方必须等8条线上的所有数字都到达后，才能发起下一轮传送；
 - 传输频率太高时，电路高速震荡的时候，数据线之间会产生很大的干扰，因此并行传输的频率不能太高。
- 串行传输：
 - 串行传输在效率上显然比并行传输低得多，但是可以通过提高传输频率来搞传输速度，总体来说串行传输反而比并行传输的总体速率更快；
 - 串行传输用于远距离传输，目前如PCI接口都转向串行传输方式。PCIe接口就是典型的串行传输方式，其单条线路传输速率高达2.5 Gb/s。

机械硬盘接口技术

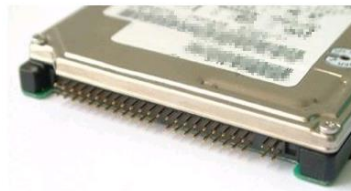
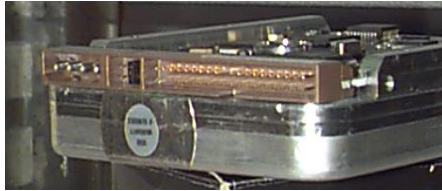
- 不管硬盘内部多么复杂，它必定要给使用者一个简单的接口，用来对其访问读取数据。硬盘向用户提供的物理接口包括如下几种：



- 按接口将硬盘分为IDE、SCSI、SATA、SAS、FC硬盘等。该方式最常见，但是这种分类方式下硬盘的差别不只是接口，还有机械底盘。
- IDE和SATA硬盘都采用ATA系的机械底盘，适用于单任务处理。
- SCSI、SAS、FC硬盘都采用SCSI系的机械底盘，能更好地满足服务器多任务的需要。
- 两者对比：
 - ATA系硬盘面对大数据吞吐量的服务时，速度不如SCSI系硬盘；
 - 面对多任务数据读取，硬盘磁头频繁来回摆动，导致硬盘过热是ATA系硬盘比较严重的问题。
 - SCSI系机械底盘可靠性比ATA系的机械底盘高。

IDE硬盘接口

- IDE接口，也称为PATA接口（Parallel ATA，并行传输ATA）
 - ATA，Advanced Technology Attachment高级技术附加装置
 - ATA硬盘也经常称为IDE Integrated Drive Electronics硬盘
 - ATA接口为并行ATA技术



- ATA接口发展到今，推出多个不同版本，分别是：ATA-1（IDE）、ATA-2（EIDE Enhanced IDE/Fast ATA）、ATA-3（FastATA-2）、ATA-4（ATA33）、ATA-5（ATA66）、ATA-6（ATA100）、ATA-7（ATA133）。
- ATA接口具有以下优缺点：
 - 优点：价格低廉、兼容性非常好。
 - 缺点：速度慢、只能内置使用、对接口电缆的长度有很严格的限制。
- 在当今的许多大型企业中，PATA现有的传输速率已经逐渐不能满足用户的需求。

SATA硬盘接口

- SATA: Serial ATA , 串行ATA
 - SATA采用串行方式进行数据传输, 接口速率比IDE接口高。
 - SATA硬盘采用点对点连接方式, 支持热插拔, 即插即用。



SATA Version	Line Code	Transfer Rate	Throughput
1.0	8b/10b	1.5 Gbit/s	150 MB/s
2.0	8b/10b	3 Gbit/s	300 MB/s
3.0	8b/10b	6 Gbit/s	600 MB/s

- 在数据传输的过程中, 数据线和信号线独立使用, 并且传输的时钟频率保持独立, 因此与PATA相比, SATA的传输速率可以达到并行的30倍。
- 优点:
 - SATA接通常为7+15针, 采用单通道, 传输速率要比ATA更快。
 - SATA对指令及数据封包采用CRC循环冗余校验技术, 保障数据传输可靠性。
 - SATA具有比ATA更好的抗干扰能力。

SCSI硬盘接口

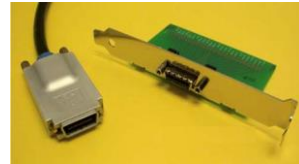
- SCSI (Small Computer System Interface) 小型计算机系统接口。



- SCSI英文全称：Small Computer System Interface，它出现的原因主要是因为原来的IDE接口的硬盘转速太慢，传输速率太低，因此高速的SCSI硬盘出现。其实SCSI并不是专为硬盘设计的，实际上它是一种总线型接口，独立于系统总线工作。
- 优点：
 - 适应面广，在一块SCSI控制卡上就可以同时挂接15个设备；
 - 高性能（具有多任务、CPU占用率低、转速快及传输率高等特点）；
 - 具有外置和内置两种，支持热插拔。
- 缺点：
 - 价格昂贵，安装设置复杂。

SAS硬盘接口

- SAS(Serial Attached SCSI) 串行连接SCSI
 - SAS是一种点对点、全双工、双端口的接口
 - SAS向下兼容SATA
 - 速率每路600M
 - SAS具有高性能、高可靠性、强大的扩展性能



- SAS和Serial ATA(SATA)硬盘相同，都是采用串行技术以获得更高的传输速度，并通过缩短联结线改善内部空间等。
- SAS是并行SCSI接口之后开发出的全新接口，此接口的设计是为了改善存储系统的效能、可用性和扩充性，并且SAS可以向下兼容SATA，二者的兼容性主要体现在物理层和协议层的兼容。
- 优点：
 - SAS在传输速率、抗干扰性方面强于SCSI，实现更长的连接距离。
- 缺点：
 - SAS接口硬盘价格相对更高。

FC硬盘接口

- FC硬盘采用FC-AL(Fiber Channel Arbitrated Loop) 光纤通道仲裁环
 - FC-AL是一种基于SCSI协议设计的双端口的串行存储接口
 - FC-AL支持全双工工作方式
 - FC为上层协议 (SCSI,IP等) 提供一个通用硬件传输平台, FC是一种高速度、高可靠、低延迟、高吞吐量的串行数据传输接口。



40-pin Male FC-SCA
II Connector

- FC和SCSI接口一样,最初也不是为硬盘设计开发的接口技术,而是专门为网络系统设计的,但随着存储系统对速度的需求,才逐渐应用到硬盘系统中。
- 优点:
 - 具有很好的升级性,可以用非常长的光纤电缆,可超过10公里;
 - 具有非常高的带宽;
 - 具有很强的通用性。
- 缺点:
 - 价格非常昂贵;
 - 组建复杂。

目录

2. 智能数据存储组件

- 控制框
- 硬盘框
- 级联模块
- 硬盘
 - 机械硬盘
 - 固态硬盘
- 接口模块

固态硬盘概述

- 相对于HDD而言，固态硬盘SSD在性能、可靠性、能耗、轻便性方面有着绝对的优势，目前已经在各领域得到广泛应用。
- SSD特点：
 - 使用flash技术存储信息，数据传输速度比HDD快
 - 内部没有机械结构因此耗电量更小、散热小、噪音小
 - SSD盘使用寿命受擦写次数影响

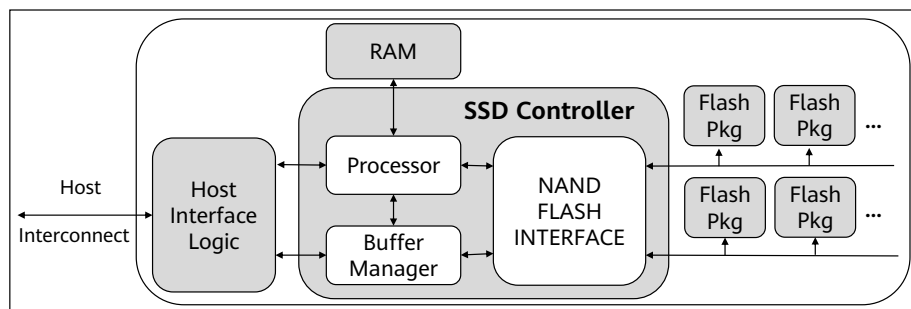


- SSD特点：
 - SSD盘并不像传统硬盘那样采用磁性材料来存储数据，而是采用基础单位被称为cell（存储单元）的NAND flash来存储数据。NAND Flash是一种非易失性随机访问存储介质，其特点是断电后数据不消失，这种技术可以很快很紧凑存储数字信息。
 - SSD无高速旋转部件，性能高，功耗低，无噪音。
 - SSD没有内部机械部件，但是，这并不意味着他们的生命周期是无限的。由于NAND Flash是非易失性介质，在写入新数据之前必须擦除原有数据，再写入数据。但是每个cell的擦写次数是有限的，当被擦写次数达到极限，则该cell就不能继续读或写数据了。

SSD架构

- SSD主要由控制单元和存储单元（当前主要是Flash闪存颗粒）组成。

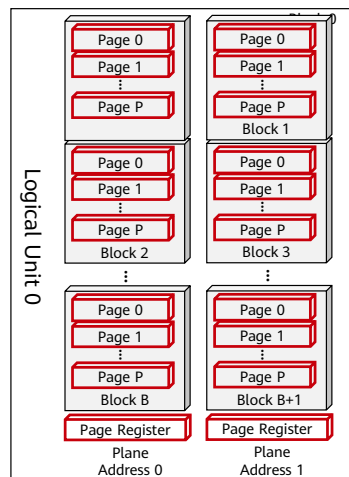
- 控制单元：SSD控制器、主机接口、DRAM等；
- 存储单元：NAND FLASH颗粒。



- 主机接口：主机访问SSD的协议和物理接口，如SATA、SAS和PCIe。
- SSD控制器：负责主机到后端介质的读写访问和协议转换，表项管理、数据缓存及校验等，是SSD的核心部件。
- DRAM：FTL（Flash translation layer，闪存转换层）表项和数据的缓存，以提供数据访问性能。
- NAND FLASH：数据存储的物理器件，是一种非易失性随机访问存储介质。
- 多通道并发，通道内Flash颗粒复用时序。支持TCQ/NCQ，一次响应多个IO请求。

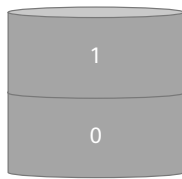
NAND Flash

- NAND Flash内部存储单元组成包括：
 - LUN、Plane、Block、Page、Cell
- 对NAND Flash的读写数据的操作主要涉及擦除（Erase）、编程（Program）和读（Read）。
- NAND Flash是非易失性介质，在写入新数据之前必须保证Block被擦除过。对Block擦除一次后再写入一次称为一次P/E Cycle。



- NAND Flash存储原理：
 - NAND Flash采用浮栅晶体管存储数据，其工作原理是利用浮栅上是否储存有电荷或储存电荷的多少来改变晶体管的阈值电压，通过读取到的晶体管阈值电压来实现数据信息的表征。
- LUN：能够独立封装的最小物理单元，通常包含多个plane。
- Plane：拥有独立的Page寄存器，通常LUN包含1K或2K奇数Block或偶数Block。
- Block：能够执行擦除操作的最小单元，通常由多个Page组成。
- Page：能够执行编程和读操作的最小单元，通常大小为16KB等。
- Cell：Page中的最小操作擦读写单元，对应一个浮栅晶体管，可以存储1 bit或多bit。
- 编程和读的基本操作单位是Page，擦除的基本单位是Block。
- 擦写过程会对浮栅晶体管的绝缘层造成一定程度的损坏，在擦除失败或编程失败后会置操作的Block为坏块，厂商定义当坏块数量达到一定程度（4%以内）的时候，则认为NAND Flash达到使用寿命上限。

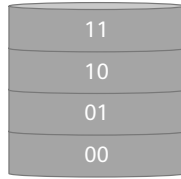
SLC、MLC、TLC和QLC



SLC-1bit

SLC

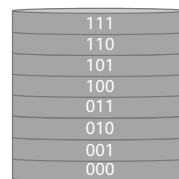
- 1、可擦写次数和可靠性最好的，一般5w-10w间；
- 2、但是存储容量相对较少；
- 3、成本也最高。



MLC-2bit

MLC

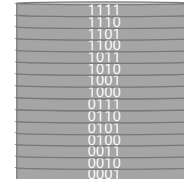
- 1、可擦写次数较少，一般3K左右；
- 2、2bit数据读写使得MLC速度比SLC慢；
- 3、存储容量较大；
- 4、价格比较便宜。



TLC-3bit

TLC

- 1、数据密度更大，可擦写次数只有几百次到上千次；
- 2、可靠性和性能很低；
- 3、具有成本优势，一般用在个人消费产品中，不能满足企业产品要求。



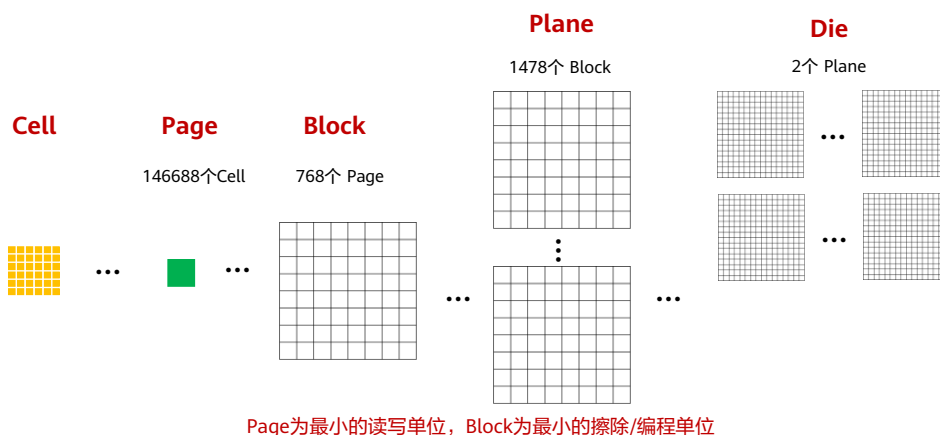
QLC-4bit

QLC

- 1、容量增加了33%；
- 2、与TLC相比较，写入性能、P/E寿命再次减少。

- NAND Flash的颗粒可根据Cell储存bit数据位的不同分为四类：
 - SLC (Single Level Cell)，单层式存储单元，在SLC中，1个cell能够存储1位数据：0或1。
 - MLC (Multi Level Cell)，多层式存储单元，在MLC中，1个cell能够存储2位数据：00，01，10和11。
 - TLC (Triple Level Cell)，三层式存储单元，在TLC中，1个cell可以存储3位数据：000,001,010,011,100,101,110和111。
 - QLC(Quad Level Cell)，四层式存储单元，在QLC中，1个cell可以存储4位数据：0000,0001,0010,0011,0100,0101,0110,0111,1000,1001,1010,1011,1100,1101,1110,1111。
- 四种不同的类型的cell，虽然存储的数据量不同，但是成本却是大致相同的，这也是SSD的容量不断增大的一个原因。最初的SSD只有64GB或更小，而现在最大TLC型的SSD盘可以存储多达2TB的数据。不过，不同类型的SSD盘的抗磨损的能力不同，导致硬盘的可靠性不同。SSD盘的抗磨损能力也是选择SSD盘的一个重要参数。

闪存颗粒数据关系



Page为最小的读写单位, Block为最小的擦除/编程单位

- 如图，以一片Flash芯片（Toshiba 3D-TLC）逻辑图为例：
 - 每 $18336 \times 8 = 146688$ 个Cell逻辑上形成一个Page，每个Page中可以存放16KB的内容和1952B的ECC校验数据，Page也是Flash芯片IO的最小单位。
 - 每768个Page组成一个Block。每1478个Block组成一个区域Plane。
 - 一整片Flash芯片由两个区域组成，一个区域存储奇数序号的Block，另一个存储偶数序号的Block，两个Plane可以并行操作。
- 考虑到要对存入NAND FLASH中的数据做ECC，NAND FLASH中的Page大小并非16KB整数，而是多出一部分Bytes。如16K Page实际大小为 $16384 + 1952$ Bytes，16384Bytes是用于存储数据，而1952Bytes则用于存放ECC数据校验码。

地址映射管理

LBA, Logical Block Address ——→ 中华人民共和国浙江省杭州市滨江区XX路26号

PBA, Physics Block Address ——→ 东经120° 12' ,北纬30° 16'

HDD: LBA和PBA一对一, 不变更

- 覆盖写

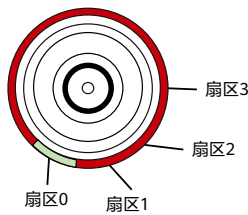
SSD: LBA和PBA的关系不固定

- 不覆盖写: 先擦除才能再写, 异处更新

FTL (Flash Translation Layer, 闪存转换层): 负责LBA和PBA的转换工作

- LBA: Logical Block Address, 逻辑区块地址, 可以指某个数据区块的地址或者某个地址上所指向的数据区块。
- PBA: Physics Block Address, 物理区块地址。
- Host是通过LBA访问SSD的, 每个LBA代表着一个Sector (一般为512 B大小), 操作系统一般以4K为单位访问SSD, 把Host访问SSD的基本单元叫做用户页 (Host Page)。
- 在SSD内部, SSD主控与Flash之间是Flash Page为基本单元访问Flash的, 称Flash Page为物理页 (Physical Page)。Host每写入一个Host Page, SSD主控会找一个Physical Page把Host数据写入, SSD内部同时记录了这样一条映射 (Map)。
- 有了这样一个映射关系后, 下次Host需要读取某个Host Page时, SSD就知道从Flash的哪个位置把数据读取上来。

FTL闪存转换层



主控制器
FTL映射表：存放在内部
SRAM/DRAM或者外部DRAM
还有NAND颗粒里。

扇区0			
			扇区2
	扇区1		
		扇区4	
扇区3			扇区5

操作系统扇区（512B），
文件系统用512B为单位
读/写操作。

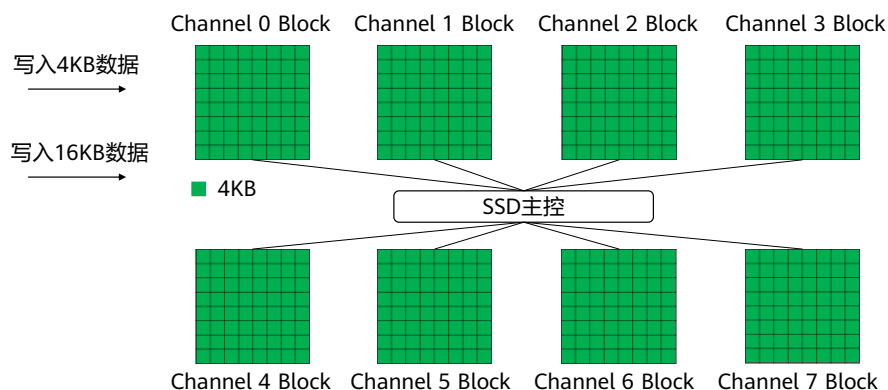
FTL映射操作，主控制
按照映射表地址来映射
地址。

NAND物理地址上的地
址数据，是根据映射表
的安排来存放。

- FTL（Flash translation layer，闪存转换层）

SSD数据写入过程 (1)

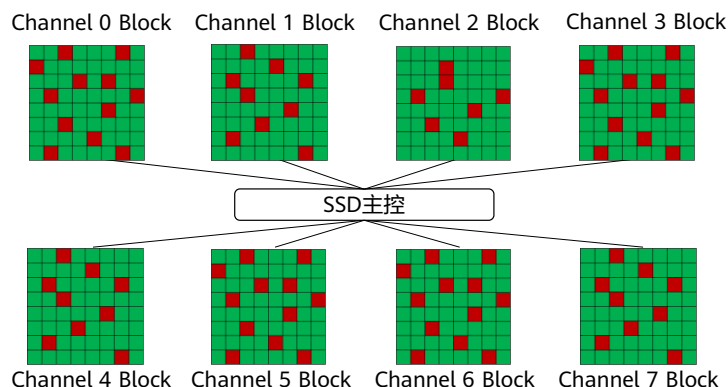
- 以8通道为例，分析Host怎么写SSD。



- 以8通道为例，分析Host怎么读写SSD。
- 主控通过8通道连接8个Flash DIE，为方便解释，这里只画了每个DIE里的一个Block，其中每个小方块表示一个Page（假设大小为4 KB）。
 - Host写入4KB数据：往channel 0的Block里写入了一个4 KB（占1Page）；
 - Host继续写入16KB数据：往channel 1-4的Block里各写入4 KB；
 - Host继续写入，最后整个Block都写满。
- 当所有Channel上的Block都写满的时候，SSD主控会挑选下一个Block以同样的方式继续写入。

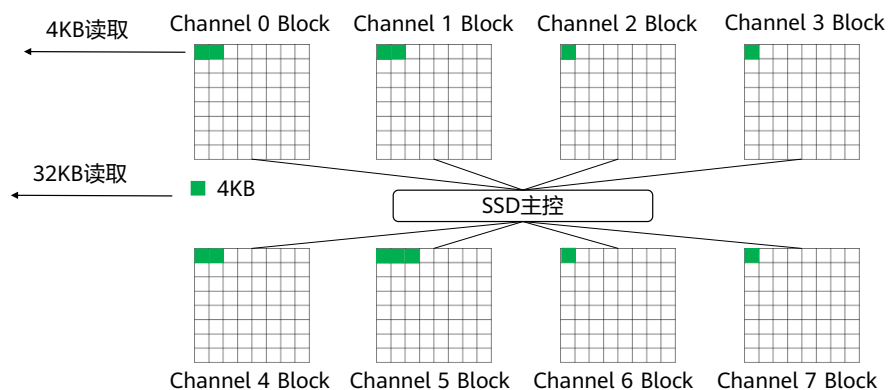
SSD数据写入过程 (2)

- 当整个SSD写满后，从用户角度来看，如果想写入新的数据，必须删除一些数据，腾出空间再写。用户在删除和写入数据的过程中，会导致一些Block里面的数据变无效或者变老。



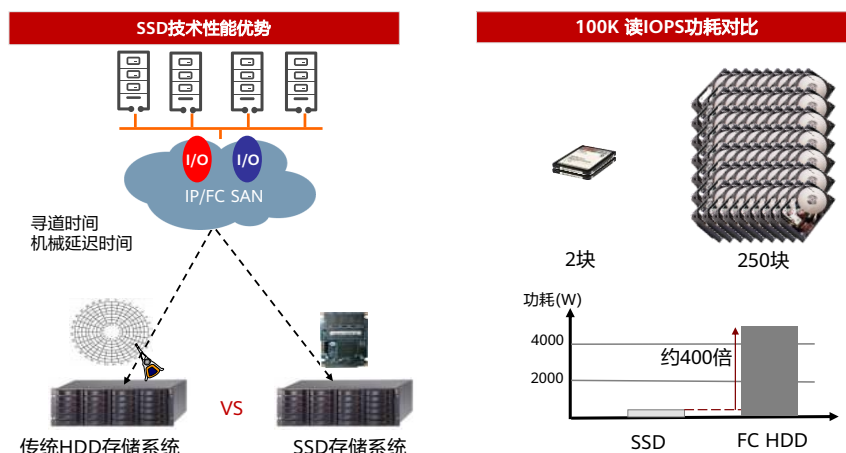
- 绿色方块代表有效数据，红色方块代表无效数据。Block中的数据变老或者无效，是指没有任何映射关系指向它们，用户不会访问到这些FLASH空间，它们被新的映射关系所取代。
- 比如有一个Host Page A，开始它存储在FLASH空间的X,映射关系为A->X。后来，HOST重写了该Host Page，由于FLASH不能覆盖写，SSD内部必须寻找一个没有写过的位置写入新的数据，假设为Y，这个时候新的映射关系建立：A->Y，之前的映射关系解除，位置X上的数据变老失效，我们把这些数据叫垃圾数据。
- 随着HOST的持续写入，FLASH存储空间慢慢变小，直到耗尽。如果不及时清除这些垃圾数据，HOST就无法写入。

SSD数据读取过程



- 速度是不是8倍提升取决于要读的数据是不是平均分布在每个channel的Block里，如果32KB数据是存储在图中channel1-4的Block的，那么读取就最多只有4倍的提升了，这也就是为什么文件越小传输率越低的。

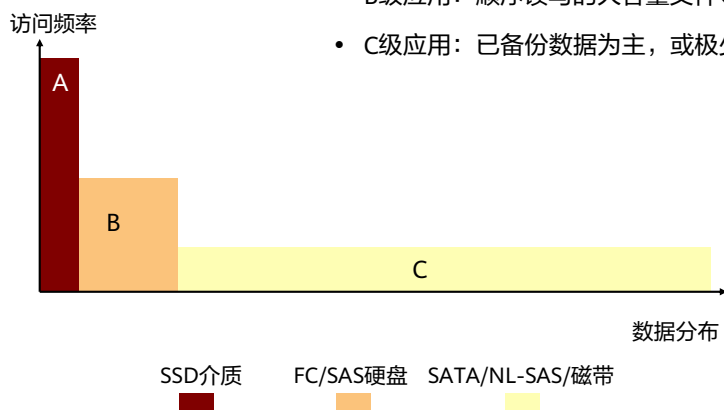
SSD性能优势



- 响应时间短
 - 传统硬盘的机械特性导致大部分时间浪费在寻道和机械延迟上，数据传输效率受到严重制约。而SSD硬盘内部没有机械运动部件，省去了寻道时间和机械延迟，可更快捷的响应读写请求。
- 读写效率高
 - 机械硬盘在进行随机读写操作时，磁头不停地移动，导致读写效率低下。而SSD通过内部控制器计算出数据的存放位置，并进行读写操作，省去了机械操作时间，大大提高了读写效率。
- 较少的硬盘功耗优势不是很明显，但是当使用大量的硬盘数量的时候，功耗所产生的成本差别就比较大，也成为企业在选择方案时候的一个考虑因素。

SSD在存储中的应用

- A级应用：以高并发随机读写为主，如数据库应用；
- B级应用：顺序读写的大容量文件、图片、流媒体等；
- C级应用：已备份数据为主，或极少使用的数据。



- 二八原则

- 用户需要频繁改动或读写的数据，一般占存储总量的20%，称为热数据，对应于A级应用。

- 分级存储

- 将热数据存放在SSD上，B级和C级应用的数据存放在高速HDD和一般HDD上，提升性能并减少投资。

SCM卡

- SCM (Storage Class Memory) 是新近问世的存储级内存，是非易失性内存，该类介质的存取速度略比内存慢，但是远快于NAND类介质。
- SCM卡是SCM介质类型的缓存加速卡，OceanStor Dorado V6全闪存存储（6.1.0及后续版本）使用SmartCache特性时，需要在控制框上安装SCM卡。



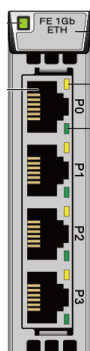
- SCM (Storage Class Memory) 即储存级内存，新一代存储介质，具备持久化和快速访问特点，能够提供比快闪存储器更快速的读写速度。
- SCM介质种类繁多：
 - PCRAM/PCM：相变存储器
 - MRAM：磁性随机存储器
 - STT-RAM：自旋转移力矩随机存取
 - FeRAM：铁电存储器
 - RRAM：阻变式存储器
 - PFRAM：聚合物存储器等

目录

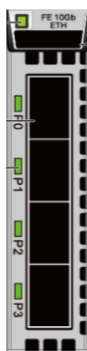
2. 智能数据存储组件

- 控制框
- 硬盘框
- 级联模块
- 硬盘
- 接口模块

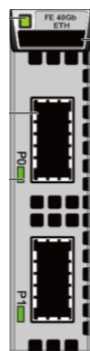
前端接口-GE接口模块



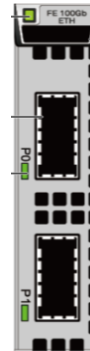
GE电接口模块



10GE电接口模块



40GE接口模块



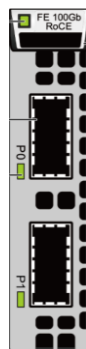
100GE接口模块

- GE电接口模块：GE电接口模块提供4个传输速率为1Gbit/s的电端口，可用作双活仲裁。
- 10GE电接口模块：10GE电接口模块提供4个传输速率为10Gbit/s的电端口，需要插入电模块后才能使用，主要用于存储设备与应用服务器之间的连接。
- 40GE接口模块：40GE接口模块提供2个传输速率为40Gbit/s的光口，主要用于存储设备与应用服务器之间的连接。
- 100GE接口模块：100GE接口模块提供2个传输速率为100Gbit/s的光口，主要用于存储设备与应用服务器之间的连接。

前端接口-RoCE接口模块



25Gb RoCE接口模块



100Gb RoCE接口模块

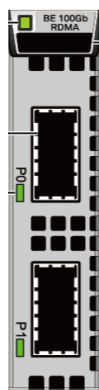
- 25Gb RoCE接口模块：主要用于存储设备与应用服务器之间的连接。
- 100Gb RoCE接口模块：主要用于存储设备与应用服务器之间的连接。

前端接口-SmartIO接口模块



- SmartIO接口模块：支持8Gbit/s、10Gbit/s、16Gbit/s、25Gbit/s、32Gbit/s五种光模块，支持8Gb FC、16Gb FC、32Gb FC、10Gb ETH、25Gb ETH端口，主要用于存储设备与应用服务器之间的连接。
- 光模块速率需要与接口模块上的标签速率一致，否则存储系统会触发告警且该端口将无法使用。

后端接口-100Gb RDMA接口模块和12Gb SAS级联模块



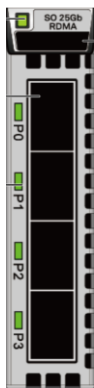
100Gb RDMA接口模块



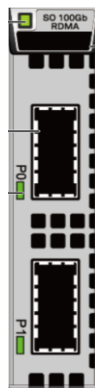
12Gb SAS级联模块

- 100Gb RDMA接口模块：提供2个传输速率为100Gbit/s的光口，用于控制框与Scale-out交换机相连，或控制框与智能硬盘框之间的连接。其中标签中的SO表示Scale-Out，BE表示Back-end。
- 12Gb SAS级联模块：提供4个传输速率为4*12Gbit/s的mini SAS HD级联端口，通过级联端口连接控制框和2U SAS硬盘框，是控制框和2U SAS硬盘框之间进行数据传输的连接点。

Scale-out接口-100Gb RDMA接口模块和25Gb RDMA接口模块



25Gb RDMA接口模块



100Gb RDMA接口模块

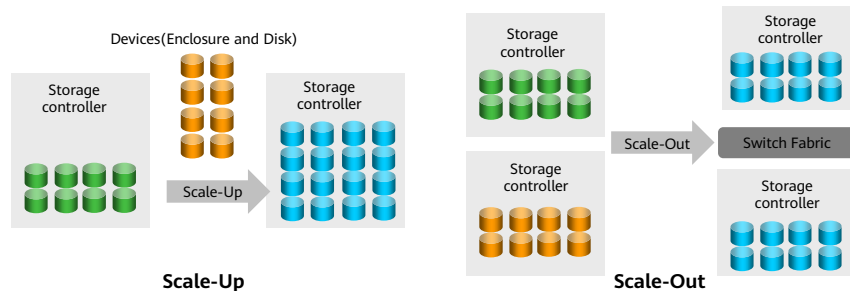
- 25Gb RDMA接口模块：提供4个传输速率为25Gbit/s的光口，用于直连组网中控制框之间的连接。
- 100Gb RDMA接口模块：提供2个传输速率为100Gbit/s的光口，用于控制框与Scale-out交换机相连，或控制框与智能硬盘框之间的连接。其中标签中的SO表示Scale-Out，BE表示Back-end。

目录

1. 智能数据存储系统介绍
2. 智能数据存储组件
- 3. 存储系统扩展方式**

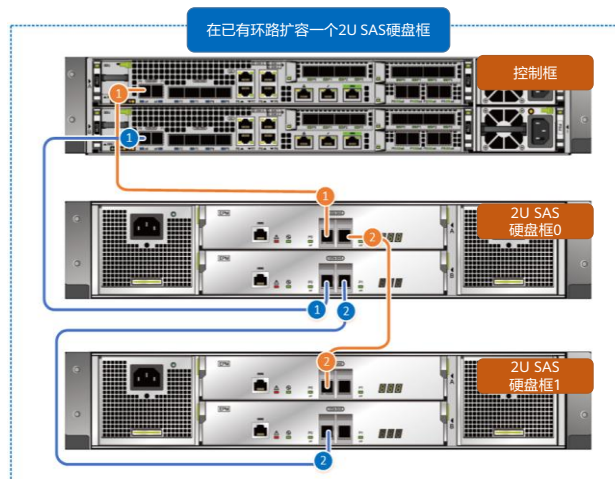
Scale-Up与Scale-Out

- 随着企业信息化进程的加快和业务规模不断扩展，业务数据不断增加，存储系统初期配置的容量已不能满足现有业务需求，存储系统扩容已经成为系统管理员重点关注的问题。主要采用两种方法进行扩容，分别是Scale-Up和Scale-Out。下面以华为存储产品为例描述两种扩展方式。



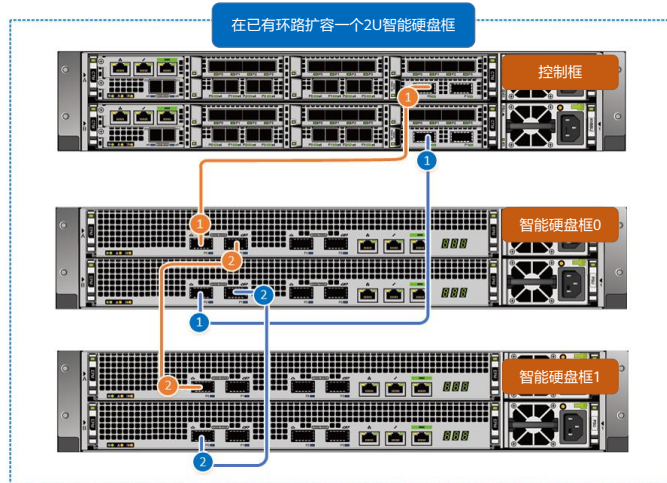
- Scale-Up:
 - 传统的纵向扩展架构，即利用现有的存储系统，通过不断增加存储器件来满足需求；
 - 优点：初期操作简单；
 - 缺点：随着存储系统规模增加，资源的增加已经到达瓶颈。
- Scale-Out:
 - 横向扩展架构的扩容通常是以节点为单位，通过扩展更多的控制器来满足需求；
 - 优点：随着规模增加，单位价格降低，效率提升；
 - 缺点：软件复杂度和管理复杂度增加。

SAS硬盘框Scale-Up组网原则



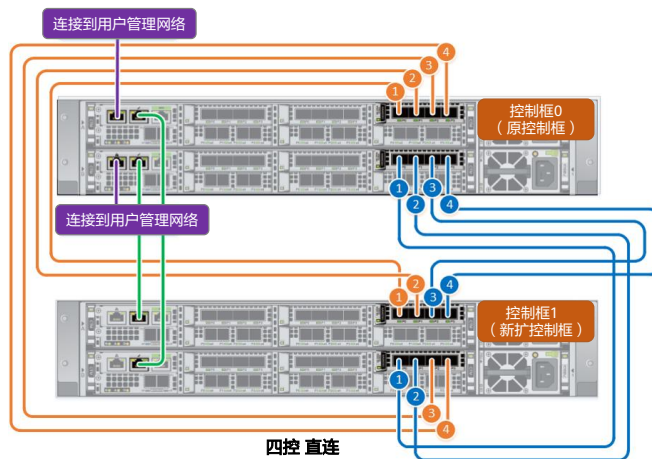
- 以华为SAS硬盘框为例说明。
 - 端口对应：同环路内上一级硬盘框的下行口（EXP）接下一级硬盘框的上行口（PRI）。
 - 双平面组网：A级联板接A控制器，B级联板接B控制器。
 - 对称组网：A、B控同槽位同端口接同一个框。
 - 正反接组网：A级联板正接，B级联板反接。（Dorado V6和融合存储X10 全正接）
 - 级联深度：级联硬盘框数目不能超过系统最大级联深度。

智能硬盘框Scale-Up组网原则



- 以华为智能硬盘框为例说明。
 - 端口对应：同环路内上一级硬盘框的下行口(P1)接下一级硬盘框的上行口(P0)。
 - 双平面组网：A级联板接A控制器，B级联板接B控制器。
 - 对称组网：A、B控同槽位同端口接同一个框。
 - 正接组网：A、B级联板正接。
 - 级联深度：级联硬盘框数目不能超过系统最大级联深度。

Scale-Out直连组网



图为Huawei OceanStor Dorado 5500 V6/5600 V6的Scale-Out组网。

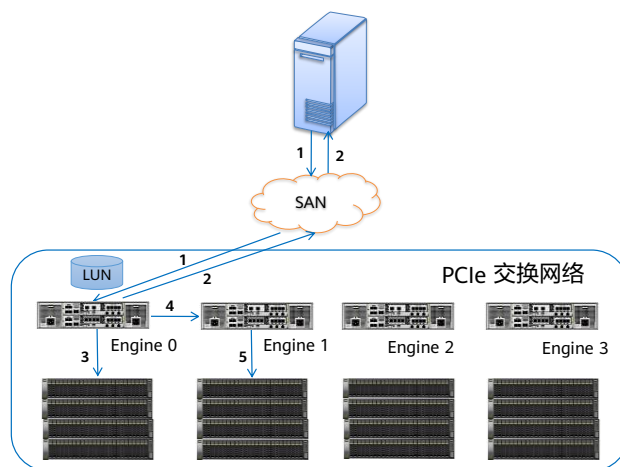
Scale-Out交换组网



图为Huawei OceanStor Dorado 6800 V6/18500 V6/18800 V6 的Scale-Out组网。

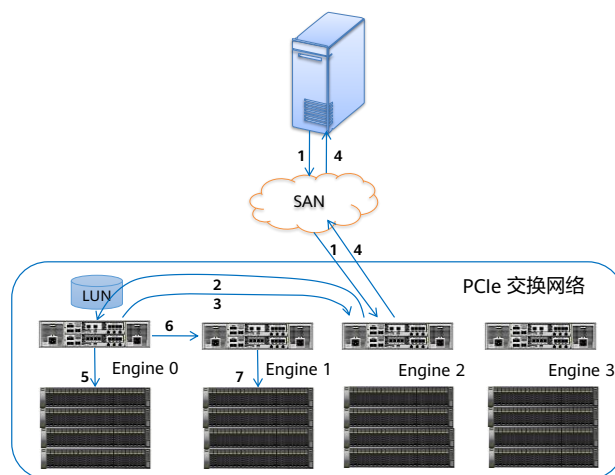


本地写流程



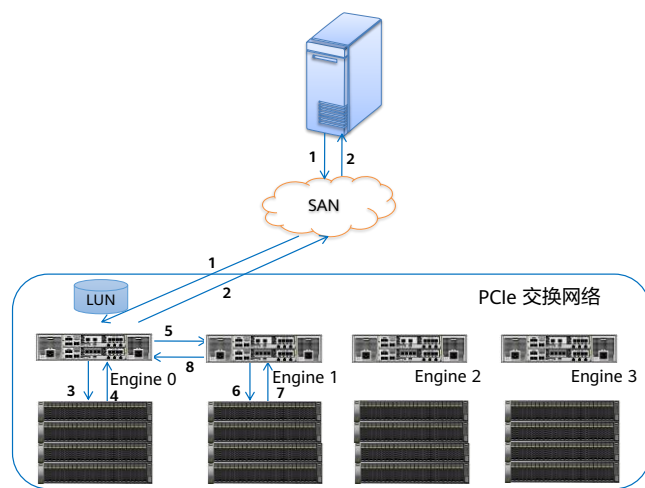
- 主机写的LUN归属于其下发写IO的引擎，其流程如下：
 1. 主机下发写IO到引擎0；
 2. 引擎0将写数据写入本地Cache（并做镜像保护）后，返回主机IO成功；
 3. 引擎0将脏数据刷盘，若需要写的盘在本地，则直接下发；
 4. 若所写盘不在本地，则转发给盘所在的引擎（引擎1）；
 5. 引擎1将脏数据写盘。

非本地写流程



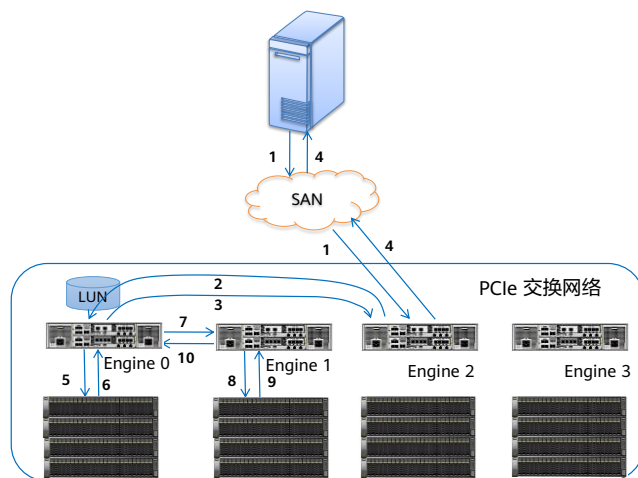
- 主机写的LUN不归属于其下发写IO的引擎，其流程如下：
 - LUN归属引擎0，主机下发写IO给引擎2；
 - 引擎2发现写的LUN归属于引擎0，则将写IO转发给引擎0；
 - 引擎0将数据写入本地Cache（并做镜像保护）后，返回引擎2 IO成功；
 - 引擎2返回主机IO成功；
 - 引擎0将脏数据刷盘，若需要写的盘在本地，则直接下发；
 - 若所写盘不在本地，则转发给盘所在的引擎（引擎1）；
 - 引擎1将脏数据写盘。

本地读流程



- 主机读的LUN归属于其下发写IO的引擎，其流程如下：
 1. 主机下发读IO到引擎0；
 2. 如果读IO在引擎0 Cache中命中，则返回主机数据；
 3. 若不命中，则引擎0将需要读盘，若需要写的盘在本地，则直接下盘读；
 4. 本地读盘数据返回，引擎0返回主机数据；
 5. 若所读盘不在本地，则转发给盘所在的引擎（引擎1）；
 6. 引擎1读盘；
 7. 引擎1读盘完成；
 8. 引擎1将数据返回引擎0，引擎0将数据返回给主机。

非本地读流程



- 主机读的LUN不归属于其下发写IO的引擎，其流程如下：
 1. LUN归属引擎0，主机下发读IO到引擎2；
 2. 引擎2发现读的LUN归属于引擎0，则将IO转发给引擎0；
 3. 如果读IO在引擎0Cache中命中，则返回引擎2数据；
 4. 引擎2返回主机数据；
 5. 若不命中，则引擎0将需要读盘，若需要写的盘在本地，则直接下盘读；
 6. 本地读盘数据返回，引擎0返回引擎2数据，引擎2将数据返回主机；
 7. 若所盘不在本地，则转发给盘所在的引擎1；
 8. 引擎1读盘；
 9. 引擎1读盘完成；
 10. 引擎1将数据返回引擎0，引擎0将数据返回给引擎2，引擎2将数据返回主机。

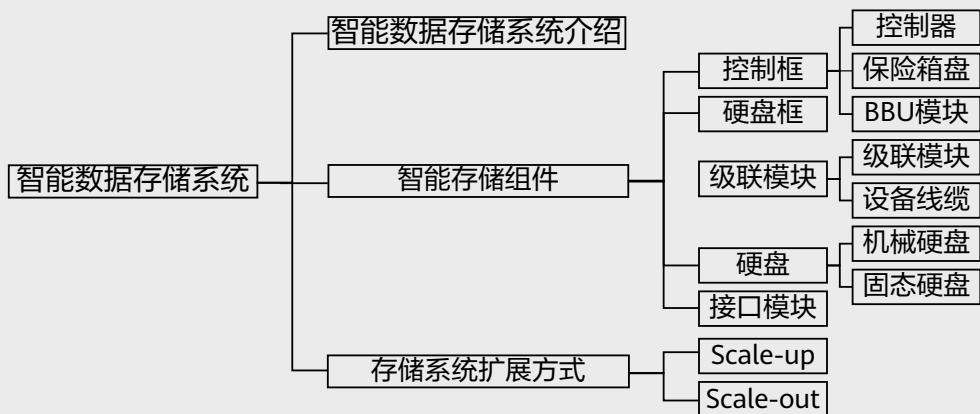
思考题

1. （多选）SSD硬盘的闪存颗粒类型有以下哪些？
 - A. SLC
 - B. MLC
 - C. TLC
 - D. QLC
2. （多选）衡量机械硬盘性能的参数有哪些？
 - A. 硬盘容量
 - B. 硬盘转速
 - C. 数据传输率
 - D. 平均访问时间

- 参考答案：

- ABCD
- BCD

本章总结



学习推荐

- 华为官方网站
 - 企业业务：<http://enterprise.huawei.com/cn/>
 - 技术支持：<http://support.huawei.com/enterprise/>
 - 在线学习：<http://learning.huawei.com/cn/>
- 热门工具
 - HedEx Lite
 - 网络资料工具中心
 - 信息查询助手

- 热门工具：
 - HedEx Lite：华为产品文档管理工具，支持浏览、搜索、更新和管理产品资料。
 - eStor：图形化存储仿真平台，该平台通过对真实华为OceanStor全闪存仿真模拟，帮助广大ICT从业者和用户快速熟悉华为存储系列产品，了解并掌握相关产品的操作和配置。
 - 网络资料工具中心：汇集网络产品的资料工具，是您进行投标支撑、网络规划、项目交付和升级维护的好助手。
 - 信息查询助手：提供华为产品命令和告警信息查询。

Thank you.

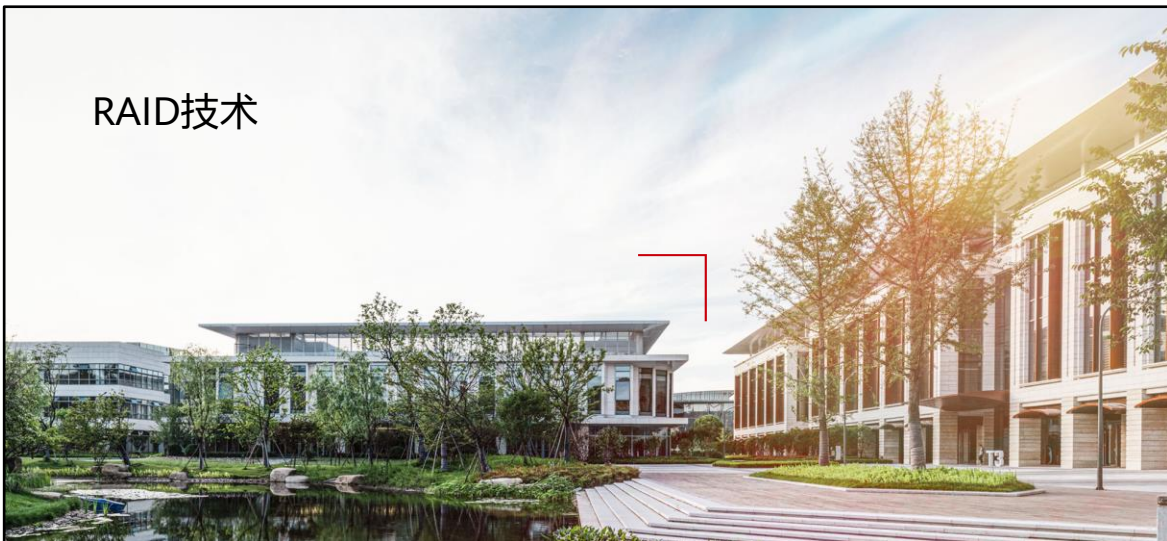
把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。
Bring digital to every person, home, and
organization for a fully connected,
intelligent world.

Copyright©2022 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.



RAID技术



前言

- 本章课程我们主要学习传统RAID和RAID2.0+技术相关知识。RAID技术发展的重要的目的是数据保护和性能提升。

目标

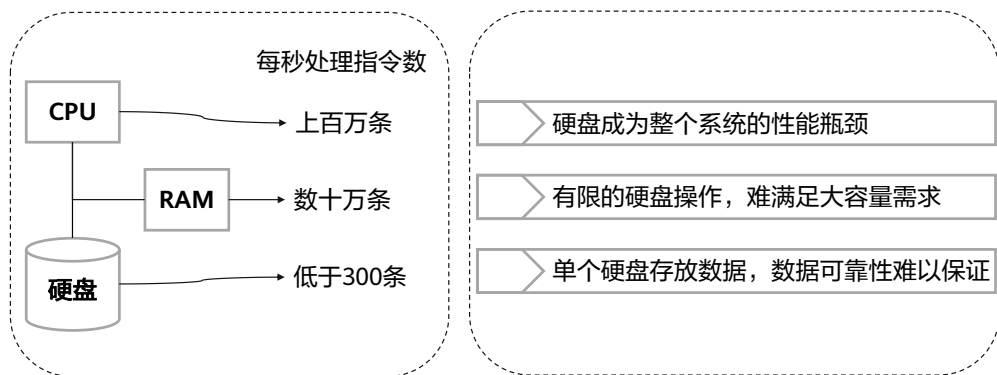
- 完成这门课程后，您将能够：
 - 列出常见的RAID类型
 - 列出不同RAID类型提供的不同级别的数据保护类型
 - 了解RAID2.0+技术原理
 - 了解动态RAID算法和RAID-TP

目录

1. 传统RAID技术
2. RAID2.0+技术
3. 其他RAID技术

RAID技术背景

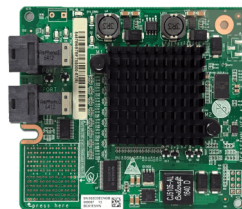
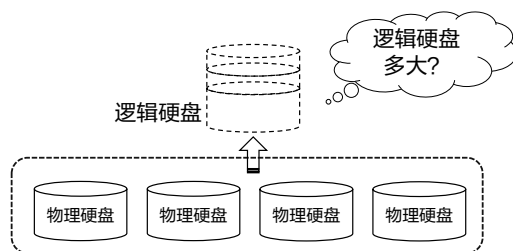
- 传统计算机系统的问题



- 单硬盘容量太小，没有冗余，硬盘故障会导致数据丢失，且单硬盘性能有限，不易管理。

RAID基本概念

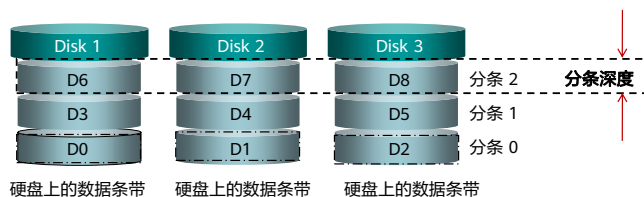
- RAID: Redundant Array of Independent Disks, 独立硬盘冗余阵列, RAID技术将多个单独的物理硬盘以不同的方式组合成一个逻辑硬盘, 从而提高了硬盘的读写性能和数据安全性。
- 实现方法: 硬件RAID和软件RAID。



- RAID技术的主要功能:
 - 把多个硬盘组合成一个逻辑盘组, 以提供更大容量的存储。
 - 将数据分割成数据块, 对多个硬盘并行进行写入/读出, 提高硬盘访问速度。
 - 通过提供镜像或奇偶校验来提供容错。
- 在存储设备中, 可以通过2种方式实现RAID功能: 硬件RAID和软件RAID。
 - 硬件RAID使用专用的RAID适配器、硬盘控制器或存储处理器。RAID控制器有自己的处理器, I/O处理芯片, 和内存, 用来提高资源利用率和数据传输速度。RAID控制器管理路由、缓冲区, 控制主机与RAID间数据流。硬件RAID通常在服务器中使用。
 - 软件实现的RAID没有它自己的处理器或I/O处理芯片, 而是完全依赖于主机处理器。因此, 低速CPU不能满足RAID实施的要求。软件RAID通常在企业级存储设备上使用。

RAID的数据组织形式

- 硬盘的条带化：将硬盘空间按照设定的大小分为多个条带，数据写入时也按照条带的大小来划分数据模块。
- 条带 (strip)：硬盘中单个或者多个连续的扇区构成一个条带，是组成分条的元素。
- 分条 (stripe)：同一硬盘阵列中的多个硬盘驱动器上的相同“位置”（或者说是相同编号）的条带。



- 分条宽度
 - 指一个磁盘组在进行分条时所使用的磁盘的个数，例如一个由3个成员盘组成的磁盘组，其条带宽度为“3”。
- 分条深度
 - 指一个条带的容量大小。

RAID的数据保护方式

- 方法一：镜像(Mirroring)，在另一块冗余的硬盘上保存数据的副本。
- 方法二：奇偶校验算法 (XOR)。
 - XOR运算广泛地使用在数字电子和计算机科学中。
 - XOR校验的算法——相同为假，相异为真：
 - $0 \oplus 0 = 0$ ； $0 \oplus 1 = 1$ ； $1 \oplus 0 = 1$ ； $1 \oplus 1 = 0$ ；

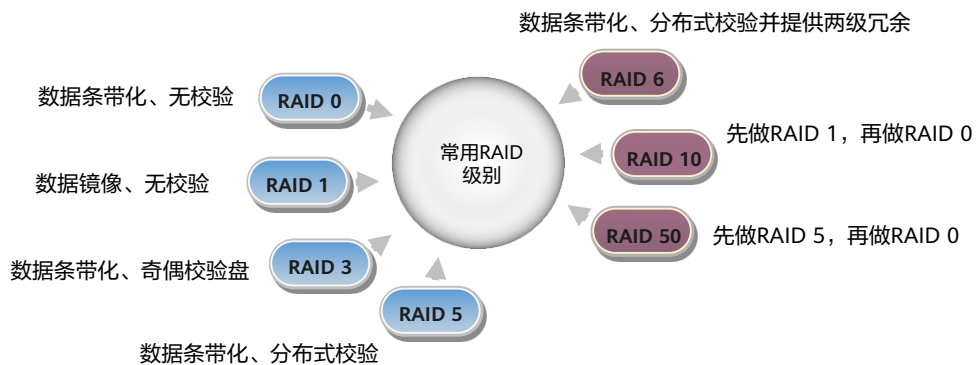


异或校验冗余备份

- RAID技术通常有2种不同的方式进行数据保护。
 - 一种方法是在另一块冗余的硬盘上保存数据的副本，提高可靠性，并可以提高读性能。
 - 二是使用奇偶校验算法。奇偶校验码是使用用户数据计算出的额外信息。对于使用奇偶校验的RAID类型，它意味着需要额外的校验硬盘。奇偶校验采用的是异或（XOR的计算符号 $\oplus$ ）算法。

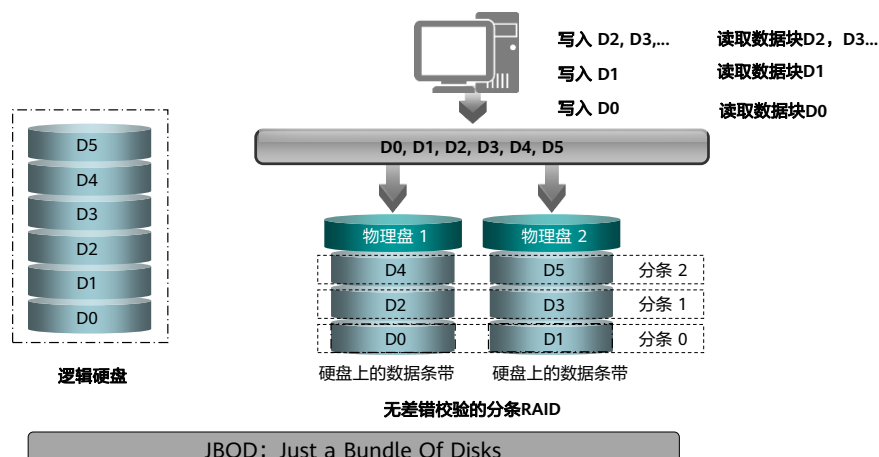
常用RAID级别与分类标准

- 根据不同的组合方式可以分为不同的RAID级别。



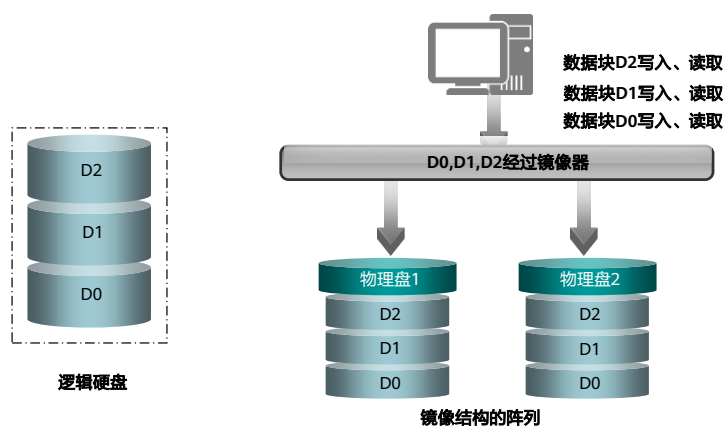
- 随着阵列技术的发展，已经产生了很多不同类型的RAID，但现在只有少数几种RAID仍在使用的。

RAID 0的工作原理



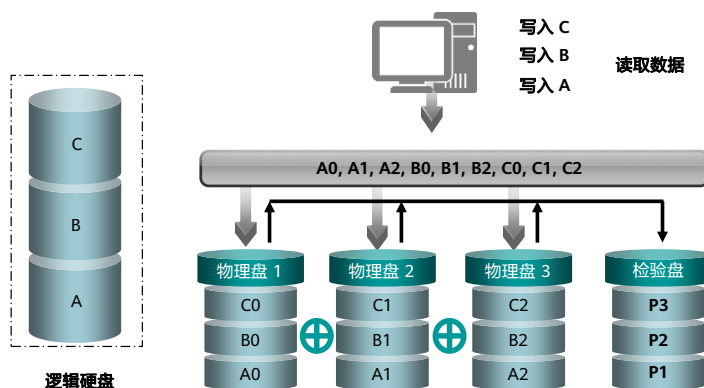
- 在所有RAID级别中，RAID 0（也被称为条带化RAID）具有最高的存储性能。RAID 0使用条带化技术将数据分布存储在RAID组的所有硬盘中。
- 一个RAID 0包含至少2个成员盘。RAID 0组将数据分为大小不等的从512个字节至兆字节的数据块（通常是512字节的倍数），并行将其写入到不同的硬盘中。如图所示的两个硬盘（驱动器）构成的RAID中：前两块数据被写入到分条0上，其中，第一个数据块被写在硬盘1的条带0上，第二个数据块并行存放在硬盘2的条带0上；这时，再下一个数据块被写到硬盘1上的下一个条带（条带1）上，以此类推。以这种方式，I/O的负载平衡分布在RAID中的所有硬盘上，由于数据传输总线上的速度远大于硬盘读写速度，因此，RAID组上的硬盘可以认为在同时进行读写。
- RAID 0像是提供了一个单一的大容量的硬盘，还同时具有非常快速I/O的特点。在RAID 0技术使用之前，类似RAID 0的一种技术被称为JBOD。一个JBOD（Just a Bundle Of Disks，简称一堆硬盘）是一组硬盘组合成一个虚拟的大硬盘。与RAID 0最大的区别是，一个JBOD的数据块不是同时并行写入不同硬盘的。在JBOD中，只有将第一块硬盘的存储空间使用完，才会使用第二块硬盘。所以JBOD总的可用容量是所有个硬盘容量的总和，但性能是单个硬盘的性能！
- 当RAID 0接收数据读取请求时，它会在所有硬盘上搜索目标数据块并读取数据。在图中，我们可以看到整个读取过程。RAID 0组的读写性能与硬盘的数量成正比。

RAID 1的工作原理



- RAID 1（也被称为镜像结构的硬盘阵列）旨在建立一个高安全性的RAID级别。RAID 1使用2个相同的硬盘系统，并设置了镜像。当数据写入到一个硬盘上时，数据的副本会同时存储在镜像硬盘上。当源硬盘（物理）失败时，镜像硬盘从源硬盘接管服务，保证服务的连续性。镜像盘作为备份，提供高数据可靠性。
- 一个RAID 1组存储的数据量只是单个硬盘的容量，另一硬盘保存的是数据的副本，相当于每一G字节的数据存储占用了2G字节的硬盘空间，所以说两个硬盘组成的RAID 1的空间利用率是50%。
- RAID 0采用条带化技术将不同数据并行写入到硬盘中，而RAID 1则是同时写入相同的数据到每个硬盘，数据在所有成员硬盘中都是相同的。在上图所示，数据块D0，D1和D2，等待写入到硬盘。D0和D0的副本同时写入到两个硬盘中（硬盘1和硬盘2），其他数据块也以相同的方式（镜像）写入到RAID 1硬盘组中。通常来说，一个RAID 1的写性能是单个硬盘的写性能。
- RAID 1读取数据时，会同时读取数据盘和镜像盘，以提高读取性能。如果其中一个硬盘失败，可以从另一个硬盘读取数据。
- RAID 1系统的读取性能等于两个硬盘的性能之和。在RAID组降级的情况下，性能下降一半。

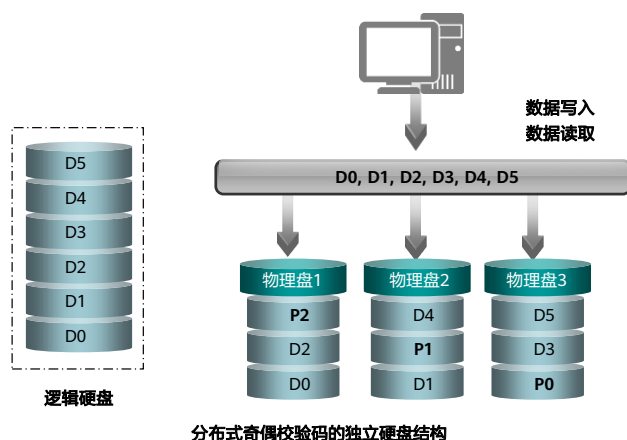
RAID 3工作原理



注：当新写入的数据较少，只需写入一个或两个硬盘时，会产生RAID 3的“写惩罚”。

- RAID 3与RAID 0类似，不同之处在于RAID 3带有专用的奇偶校验的分条。在RAID 3中，一块专用硬盘（校验盘）用来保存同一条带上其他硬盘上的相应的条带中的数据的奇偶校验值。如果检测到不正确的数据或硬盘出现故障，我们可以利用奇偶校验信息来恢复故障硬盘上的数据。RAID 3适用于数据密集型或单用户环境，需要长期、连续访问的数据块。RAID 3将数据写入操作分配给RAID组内的数据成员盘。但是，当有新数据需要写入时，无论写入哪个硬盘，RAID 3都需要重新计算并重写校验信息。因此，当某个应用程序需要大量写入时，RAID 3的奇偶校验盘将有很大的工作量。因为需要等待奇偶校验，所以会对RAID 3组的读写性能有一定影响。此外，因为校验盘有较高的工作负载，它往往是RAID 3里最容易失效的硬盘。而少量写入数据的场景，因为需要额外的读和写操作，相对于对单个硬盘进行数据写入，并没有提升硬盘的性能，这种情况被称为RAID 3的“写惩罚”。
- RAID 3采用单硬盘容错和并行数据传输。RAID 3采用分条技术将数据分块，这些块进行异或算法，并将奇偶校验数据写到最后一个盘——RAID 3组的奇偶校验硬盘。
- RAID 3的写入性能取决于更改数据的数量、硬盘的数目、以及计算和存储奇偶校验信息所需的时间。假定一个RAID 3的硬盘数为N，当所有成员盘的转速相同时，在不考虑写惩罚，满分条写的情况下，RAID 3的顺序IO写性能理论上略小于 N-1倍单个硬盘的性能（计算冗余校验需要额外的计算时间）。
- 在RAID 3中，数据以分条的方式进行读取。RAID中的每个硬盘的硬盘驱动器被控制，所以RAID 3里同一条带上的数据块可以并行读取。
- RAID 3使用并行数据读（写）模式。RAID 3的读取性能取决于读取的数据量和RAID 3阵列的硬盘数量。

RAID 5的工作原理



- RAID 5是改进版的RAID 3，使用条带化并计算奇偶校验信息，在RAID 5中，数据以分条的形式写入硬盘组中。RAID 5的不同分条的奇偶校验数据不是单独存在一个固定的校验盘里的，而是按一定规律分散存放的。
- RAID 3在少量的数据被写入时有写惩罚，RAID 5类似。
- RAID 5的写入性能取决于所写的的数据量和RAID 5组中硬盘的数量。假定一个RAID 5的硬盘数为N，当所有成员盘的转速相同时，在不考虑写惩罚，满分条写的情况下，RAID 5的顺序IO写性能理论上略小于 N-1倍单个硬盘的性能（计算冗余校验需要额外的计算时间）。
- 在RAID 3级别和RAID 5级别的硬盘阵列中，如果一个硬盘失效，该硬盘组将从在线（正常）状态转变为降级状态，直到完成重构失效硬盘。如果RAID中的另一个硬盘也出现故障，则硬盘组的数据将丢失。

RAID 6概述

- RAID 6
 - 需要至少 $N+2$ ($N>2$)个硬盘来构成阵列，一般用在数据可靠性、可用性要求极高的应用场合。
- 常用的RAID 6技术有：
 - RAID6 P+Q
 - RAID6 DP

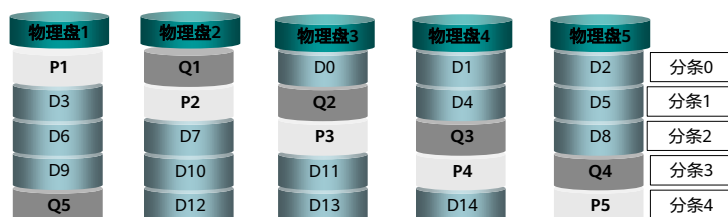
- 前面讨论到的RAID组数据保护都是考虑单一硬盘失效的场景（RAID 0排除在外）。现在，硬盘的容量已经增加了很多，同时重构时间也增加了。很多大容量的硬盘组合起来形成的一个RAID 5组重建失效硬盘可能需要几天，而不是几个小时。在重建过程中，系统处于降级状态，这种情况下，任何额外的硬盘故障都会导致硬盘组失效和数据丢失。这就是为什么一些组织或单位需要一个双冗余系统。换句话说：一个RAID组应该允许2个硬盘故障时，同时所有的数据应该是可访问的。这种双重冗余数据保护类型的实现有一些不同的方式：
 - 第一种是多重镜像。多重镜像是指数据块存储在主盘时同步存储多个副本到多余硬盘的方法。这种方式意味着大量的开销。
 - 第二种方式是RAID 6级别硬盘阵列。RAID 6组对2个硬盘失效提供保护。这些硬盘甚至可以在同一时间失效。
- RAID 6的正式名称是分布式双校验RAID。本质上它是一种改进的RAID 5，也具有条带化和分布式奇偶校验。现在在RAID 6有双校验，这意味着两点：
 - 写入用户数据时，附加的双校验计算需要进行。所以，在所有RAID 类型中，RAID 6是“最慢”的。
 - 额外的校验信息需要占用两个盘的存储空间。这就是为什么我们把RAID 6看作是一个 $N+2$ 类型的RAID。
- 目前，RAID 6没有一个统一的标准。不同公司以不同的方式实施RAID 6。不同种模式获得校验数据的方法不同。

RAID 6 P+Q工作原理

- RAID6 P+Q需要计算出两个校验数据P和Q，当有两个数据丢失时，根据P和Q恢复出丢失的数据。校验数据P和Q是由以下公式计算得来的：

$$P = D0 \oplus D1 \oplus D2...$$

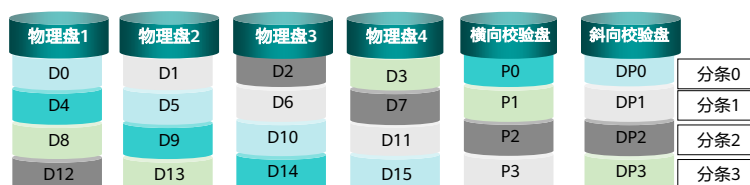
$$Q = (\alpha * D0) \oplus (\beta * D1) \oplus (\gamma * D2)...$$



- RAID 6采用P+Q校验时，P和Q是2个彼此独立的校验值。它们使用不同的算法，用户数据和校验数据分布在同一分条的所有硬盘上。
- 如图所示，P1是通过D0，D1，D2所在的分条0进行异或操作获得的，P2是对D3，D4，D5所在的分条1异或操作实现的，P3则是对D6，D7，D8所在的分条2条进行异或操作。
- Q1是对D0，D1，D2所在的分条0条进行GF变换再异或操作实现的，Q2是对D3，D4，D5所在的分条1进行GF变换再异或运算，Q3实现D6，D7，D8分条2进行GF变换再异或。
- 如果一个硬盘中的一个分条失效，只需有P校验值即可恢复失效硬盘上的数据，异或运算在P校验值和其它数据硬盘间执行。如果同一个分条有2个硬盘同时故障，不同的场景有不同的处理方法。如果Q校验值不在失效的一个硬盘上，数据可以被恢复到数据盘上，然后重新计算校验信息。如果Q在其中一个失效的硬盘上，两个的公式都需要使用才能恢复两个失效硬盘上的数据。

RAID 6 DP工作原理

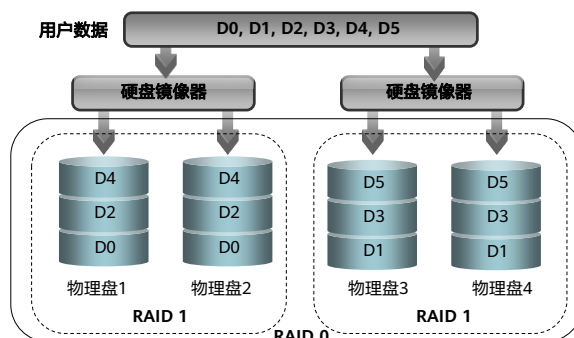
- DP - Double Parity，就是在RAID4所使用的的一个行XOR校验硬盘的基础上又增加了一个硬盘用于存放斜向的XOR校验信息
- 横向校验盘中P0 - P3为各个数据盘中横向数据的校验信息
- 例：P0=D0 XOR D1 XOR D2 XOR D3
- 斜向校验盘中DP0 - DP3为各个数据盘及横向校验盘的斜向数据校验信息
- 例：DP0=D0 XOR D5 XOR D10 XOR D15



- RAID 6 DP也有两个独立的校验数据块。第一个校验信息与RAID 6 P+Q的第一个校验值是相同的，第二个不同于RAID 6 P+Q，采用的是斜向异或运算得到行对角奇偶校验数据块。行奇偶校验值是同一分条的用户数据异或运算获得到，所图所示：P0是由分条0上的D0，D1，D2和D3异或运算得到，P1由分条1上的D4，D5，D6，D7异或运算，等等。所以， $P0 = D0 \oplus D1 \oplus D2 \oplus D3$ ， $P1 = D4 \oplus D5 \oplus D6 \oplus D7$ ，如此类推。
- 第二个校验数据块是由阵列的对角线数据块进行异或运算。数据块的选择过程比较复杂。DP0是由硬盘1 的分条0上的D0，硬盘2的分条1上的D5，硬盘3上的分条2的D10，和硬盘上4 分条3上的D15异或操作得到。DP1是对硬盘2 的分条0上的D1，硬盘3的分条1上的D6，硬盘4上分条2的 D11，和的第一块校验硬盘上分条3 上的P3进行异或运算得到。DP2是硬盘3 分条0上的D2，硬盘4上的分条1的 D7，奇偶硬盘分条2的P2，和硬盘1 分条3上的D12进行异或运算得到。所以， $DP0 = D0 \oplus D5 \oplus D10 \oplus D15$ ， $DP1 = D1 \oplus D6 \oplus D11 \oplus P3$ ，如此类推。
- 一个RAID 6阵列能够容忍双硬盘失效。
- 一个RAID 6组的性能，无论算法是DP还是P+Q，相对都比较慢。因此，RAID 6适用两种场景：
 - 数据非常重要，需要尽可能长的时间处于在线和可使用的状态。
 - 使用的硬盘容量非常大（通常超过2T）。大容量硬盘的重建时间较长，两个硬盘都失效是会造成数据较长时间不能访问。在RAID 6中，可以实现一个硬盘重构时另一个硬盘失效。一些企业希望在使用大容量硬盘后，存储阵列的供应商使用一个双重保护的RAID组。

RAID 10

- RAID 10是将镜像和条带进行组合的RAID级别，先进行RAID 1镜像然后再做RAID 0。RAID 10也是一种应用比较广泛的RAID级别。

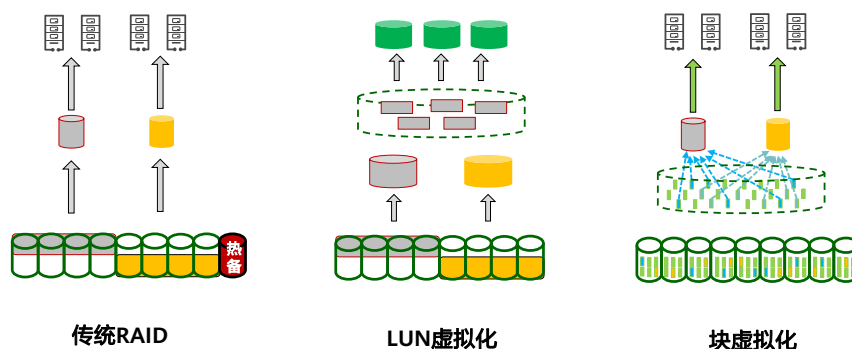


- 对于大多数的企业客户而言，RAID 0并不是一个真正可以操作的选择，而RAID 1受限于硬盘容量利用率。RAID 10组合了RAID 1和RAID 0，提供了最好的解决方案，特别是在随机写入时，由于不存在写惩罚，性能优势比较明显。
- RAID 10组的硬盘数量总是偶数。一半硬盘进行用户数据写入，另一半保存用户数据的镜像副本。镜像基于分条执行。
- 在图中，物理硬盘1和2构成一个RAID 1，物理硬盘3和物理硬盘4形成另一个RAID 1。这2个RAID 1子组再形成RAID 0。
- RAID 10组写入数据时，子组间采用并行的方式写入数据块，子组内数据采用镜像的方式写入。如图所示，D0将写入物理硬盘1，副本将被写入物理硬盘2。
- 当硬盘在不同的RAID 1组故障（例如硬盘2和4），RAID 10组的数据访问不受影响。这是因为其他2个硬盘（3和1）上有故障盘2和4上数据的完整副本。但是，如果同一RAID 1子组的硬盘（例如，硬盘1和2）在同一时间失败，数据将不能访问。
- 从理论上讲，RAID 10可以忍受总数一半的物理硬盘失效，然而，从最坏的情况来看，在同一个子组的两个硬盘故障时，RAID 10也可能出现数据丢失。通常RAID 10用来保护单一的硬盘失效。

目录

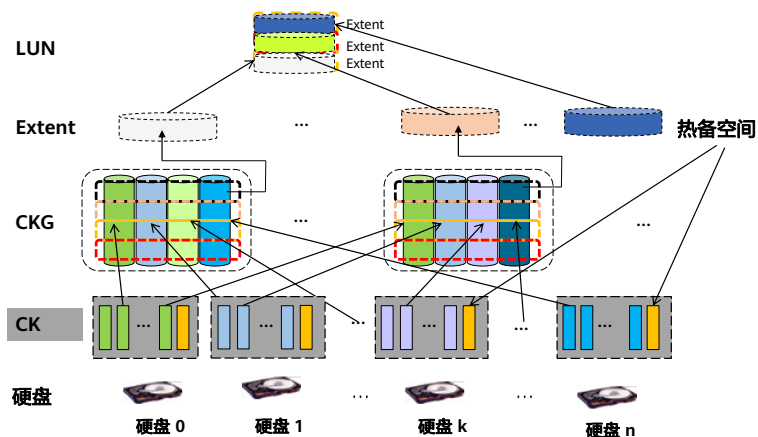
1. 传统RAID技术
- 2. RAID2.0+技术**
3. 其他RAID技术

RAID技术演变



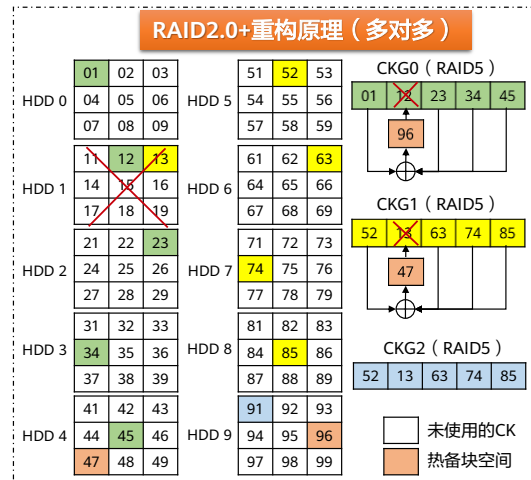
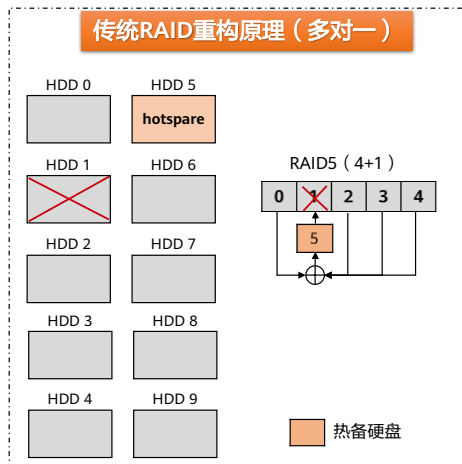
- 作为一种成熟、可靠的硬盘系统数据保护标准，RAID技术自诞生以来一直作为存储系统的基础技术而存在，但是近年来随着数据存储需求的快速增长，单盘容量不断增大，高性能应用的不断涌现，传统RAID逐渐暴露出越来越多的问题，特别是大容量硬盘出现故障而需要进行数据重构的场景。
- 传统RAID技术存在着以下不足：
 - 数据丢失风险大：硬盘容量越来越大，重构时间越来越长，数据丢失风险越来越高。重构过程中数据处于单冗余状态，再有硬盘故障或者出现硬盘坏块，将导致数据丢失。重构时间的延长增加了数据丢失的风险。
 - 对业务影响大：重构过程中，RAID组成员盘忙于重构，服务能力明显下降，影响上层业务的运行。
- 为了解决传统RAID的上述问题，同时顺应虚拟化技术的发展趋势，如下替代方案应运而生。
 - LUN虚拟化：在传统RAID基础上将RAID组进行更细粒度地切分，再将切分的单元进行组合，构建主机可访问的空间。
 - 块虚拟化：将存储池中的硬盘划分成一个个小粒度的数据块，基于数据块来构建RAID组，使得数据均匀地分布到存储池的所有硬盘上，然后以数据块为单元来进行资源管理。

RAID 2.0+ 基本原理

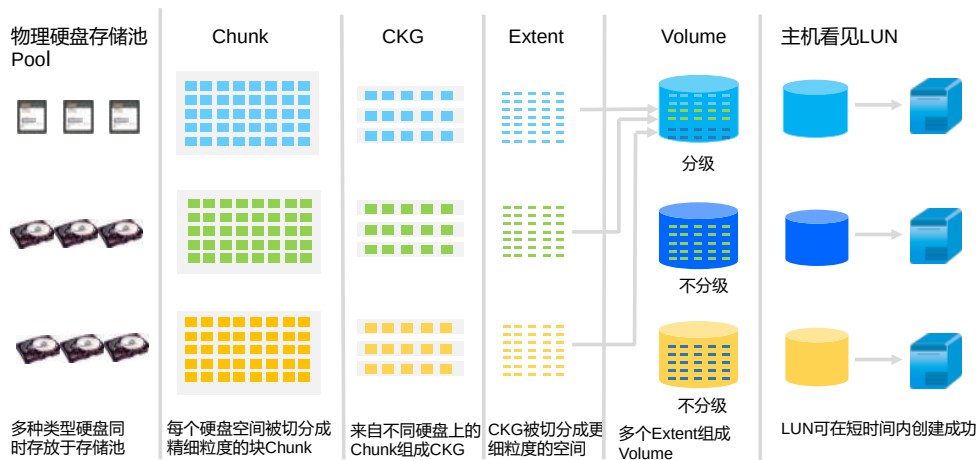


- RAID2.0+技术，把物理硬盘划分成很多数据块（chunk，CK），不同盘上的CK组成具备RAID关系的CKG（Chunk Group），多个CKG组成庞大的存储资源池，主机所用的资源从资源池中获取。
- 相对于传统RAID机制，RAID2.0+具备如下优势：
 - 业务负载均衡，避免热点：数据打散到资源池内所有硬盘上，没有热点，硬盘负荷平均，避免个别盘因为承担更多的写操作而提前达到寿命的上限。
 - 快速重构，缩小风险窗口：当硬盘故障时，故障盘上的有效数据会被重构到资源池内除故障盘外的所有盘上，实现了多对多的重构，速度快，大幅缩短数据处于非冗余状态的时间。
 - 全盘参与重构：资源池内所有硬盘都会参与重构，每个盘的重构负载很低，重构过程对上层应用无影响。

RAID 2.0+重构原理



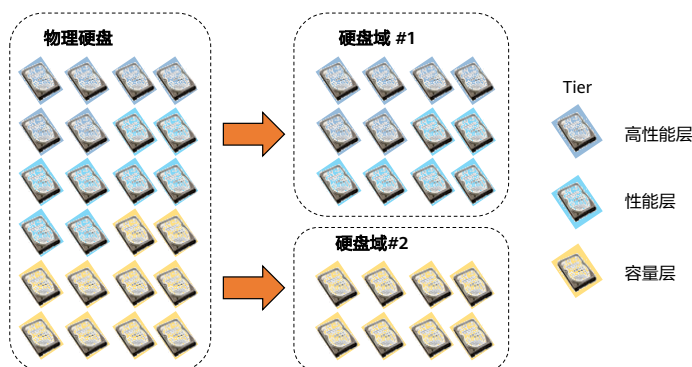
RAID 2.0+ 软件逻辑对象



- 华为RAID 2.0+：底层介质虚拟化 + 上层资源虚拟化，同时解决数据快速重构问题和资源的智能分配问题。
- 数据快速重构：缩短数据重构时间，从小时级到分钟级，使重构速度提升近20倍，大幅降低重构过程对业务的影响和多盘失效的概率。
- 存储池内所有硬盘参与重构，仅重构业务数据，由传统RAID多对一的重构模式转变为多对多快速重构模式。

硬盘域

- 硬盘域是一堆硬盘的组合（可以是整个系统所有硬盘）这些硬盘整合并预留热备容量后统一向存储池提供存储资源。



- 在传统RAID中，要将硬盘空间提供给业务主机的第一步操作是 创建RAID组。但是对创建RAID组这一操作有一些限制和要求：RAID组中的硬盘的类型是相同的，硬盘的大小和转速需要保持一致，而且一个RAID组中的硬盘个数不建议超过12个。
- 华为RAID 2.0+技术中采用了另一种实现方法。在该方法中，第一步操作不是 创建RAID组，而是创建硬盘域。一个硬盘域就是一组硬盘。一个硬盘只能属于一个硬盘域。OceanStor存储系统可以创建一个或多个硬盘域。看起来，硬盘域的概念似乎跟RAID组的概念类似，都是一组硬盘，但实际上还是有很大的区别。在创建的RAID组中，这组硬盘已经按照设定某个RAID组级别进行了绑定，且这组硬盘的类型，大小，转速都必须是一样的。但是在一个硬盘域中，硬盘的个数可以大很多，比如上百个，而且可以最多包含3种类型的硬盘，每种类型的硬盘对应分配到一个存储层级，SSD盘对应分配到高性能层，SAS盘分配到性能层，NL-SAS盘分配到容量层。如果硬盘域中没有相应的硬盘类型，则该硬盘域就没有相应的存储层。更为关键的区别是，硬盘域的主要作用是将这一组盘和哪一组盘分开，目的是实现故障、性能和存储资源等的完全隔离。在创建硬盘域时，RAID级别尚未指定，即数据的冗余保护方式还未指定。事实上，RAID 2.0+ 对数据冗余保护方式的设置更加灵活，更加精细化。硬盘域中硬盘组成的存储空间，将被划分成更小粒度的存储池，以及为各个存储层级内共享的热备空间。热备空间大小是系统根据管理员为该硬盘域设置的热备策略（高、低、无）以及该硬盘域各存储层的硬盘个数自动设置的，而不再是传统RAID组中那样，由管理员指定某个特定的盘做热备盘。

Storage Pool & Tier

- Storage Pool即存储池，是存放存储空间资源的容器，所有应用服务器使用的存储空间都来自于存储池。
- Tier即存储层级，存储池中性能类似的存储介质集合，用于管理不同性能的存储介质，以便为不同性能要求的应用提供不同存储空间。

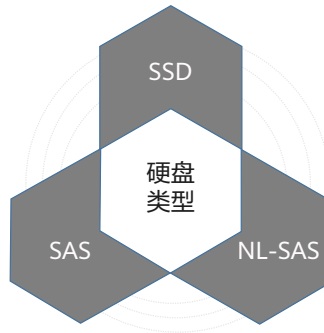
存储层级	层级名称	支持硬盘类型	应用
Tier0	高性能层	SSD	性能和价格较高，适合存放访问频率很高的数据
Tier1	性能层	SAS	性能较高，价格适中，适合存放访问频率中等的的数据
Tier2	容量层	NL-SAS	性能较低，价格最低且单盘容量大，适合存放大容量的数据以及访问频率较低的数据

RAID级别	RAID策略
RAID1	1D+1D, 1D+1D+1D+1D
RAID10	系统自动选择2D +2D或4D+4D
RAID3	2D+1P, 4D+1P, 8D+1P
RAID5	2D+1P, 4D+1P, 8D+1P
RAID50	(2D+1P)*2, (4D+1P)*2, (8D+1P)*2
RAID6	2D+2P, 4D+2P, 8D+2P, 16D+2P

- 一个存储池基于指定的一个硬盘域创建，可以从该硬盘域上动态的分配Chunk（CK）资源，并按照每个存储层级（Tier）的“RAID策略”组成Chunk Group（CKG）向应用提供具有RAID保护的存储资源。
- 存储池根据硬盘类型可划分为多个Tier，OceanStor存储系统支持的存储层级和硬盘类型上图左表所示。
- 创建存储池可以指定该存储池从硬盘域上划分的存储层级（Tier）类型以及该类型的“RAID策略”和“容量”。
- OceanStor存储系统支持RAID1、RAID10、RAID3、RAID5、RAID50和RAID6，支持的RAID策略和配置上图右表所示。
- 容量层由大容量的SATA、NL-SAS盘组成，RAID策略建议使用双重校验方式的RAID6。（SATA盘较少使用，部分产品规格上已经不支持）

Disk Group (DG)

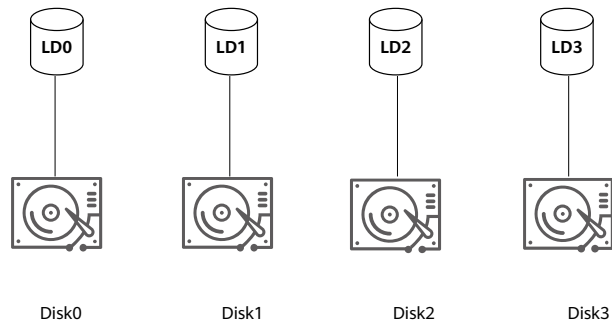
- Disk Group (DG)即硬盘组，由硬盘域内相同类型的多个硬盘组成的集合，硬盘类型包括SSD、SAS和NL-SAS三种。



- OceanStor存储系统会在每个硬盘域内根据每种类型的硬盘数量自动划分为一个或多个Disk Group (DG)。
- 一个Disk Group (DG) 只包含一种硬盘类型。
- 任意一个CKG的多个CK来自于同一个Disk Group (DG) 的不同硬盘。
- Disk Group (DG) 属于系统内部对象，主要作用为故障隔离，由OceanStor存储系统自动完成配置，对外不体现。

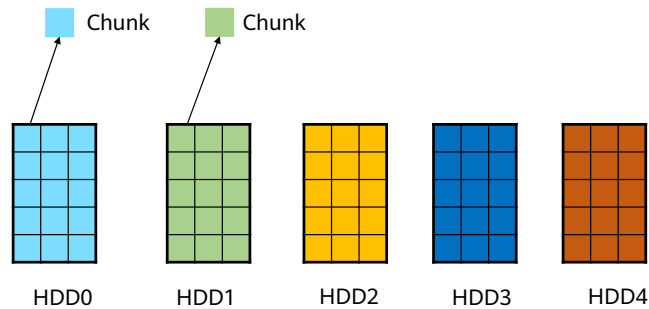
LD

- Logical Drive (LD)即逻辑磁盘，是被存储系统所管理的硬盘，和物理硬盘一一对应。



CK

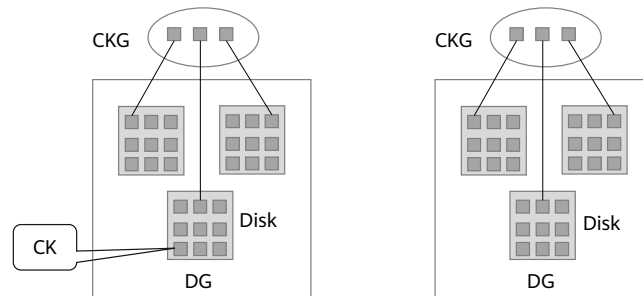
- Chunk简称CK，是存储池内的硬盘空间切分成若干固定大小的物理空间，是组成RAID的基本单位。



- 块大小是系统在切分物理空间的时候固定的大小，不能进行更改。

CKG

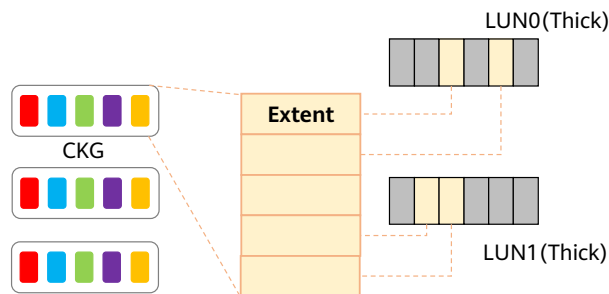
- Chunk Group简称CKG，是由来自于同一个DG内不同硬盘的CK按照RAID算法组成的逻辑存储单元，是存储池从硬盘域上分配资源的最小单位。



- 一个CKG中的CK均来自于同一个DG中的硬盘，CKG具有RAID属性（RAID属性实际配置在Tier上），CK和CKG均属于系统内部对象，由存储系统自动完成配置，对外不体现。

Extent

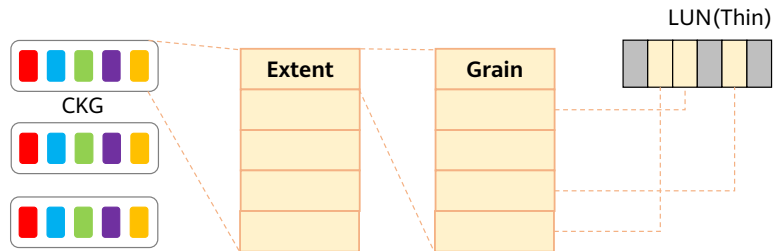
- Extent是在CKG基础上划分的固定大小的逻辑存储空间，大小可调，是热点数据统计和迁移的最小单元（数据迁移粒度），也是存储池中申请空间、释放空间的最小单位。



- 一个Extent归属于一个Volume或一个LUN，Extent大小在创建存储池时可以进行设置，创建之后不可更改，不同存储池的Extent大小可以不同，但同一存储池中的Extent大小是相同的。

Grain

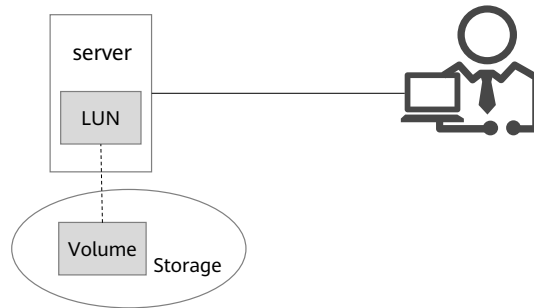
- 在Thin LUN模式下，Extent按照固定大小被进一步划分为更细粒度的块，这些块称之为Grain。Thin LUN以Grain为粒度进行空间分配，Grain内的LBA是连续的。



- Thin LUN以Grain为单位映射到LUN，对于Thick LUN，没有该对象。

Volume & LUN

- Volume即卷，是存储系统内部管理对象。
- LUN是可以直接映射给主机读写的存储单元，是Volume对象的对外体现。



- 一个Volume对象用于组织同一个LUN的所有Extent、Grain逻辑存储单元，可动态申请释放Extent来增加或者减少Volume实际占用的空间。

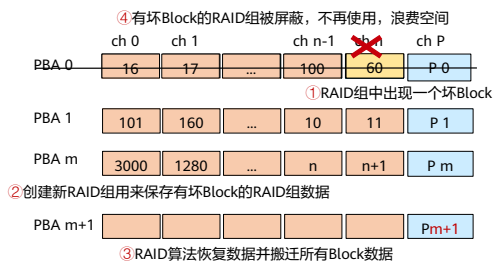
目录

1. 传统RAID技术
2. RAID2.0+技术
- 3. 其他RAID技术**

华为动态RAID算法

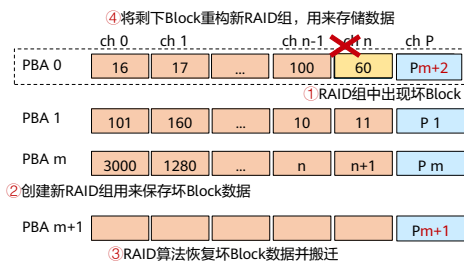
普通RAID算法

- RAID组中坏一个Block后，恢复坏Block数据，并搬迁整个RAID组数据，再屏蔽该RAID组空间；
- 效果：大量可用的Flash空间被浪费。



华为动态RAID算法

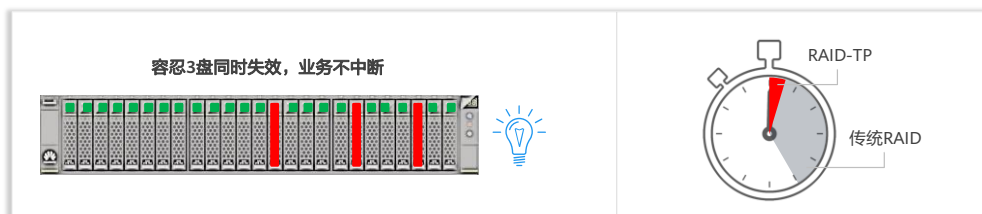
- RAID组中坏一个Block后，恢复坏Block数据并搬迁，同时屏蔽坏Block，用剩下的Block重构新RAID组；
- 效果：充分有效利用所有Flash空间。



- 能在发生Flash器件失效后，将主动恢复故障Flash中的数据并继续对数据进行RAID保护。
- 该算法能够动态调整RAID中数据块的数目，从而满足系统可靠性和容量要求。当出现某个CHUNK损坏，并且无法从RAID的成员盘以外选择新的CHUNK时，系统使用动态方式将原来N+M分条中的数据重构至新的(N-1)+M分条中。当有新的SSD盘插入时，系统重构RAID，动态的将原来(N-1)+M分条中的数据搬移至新的N+M分条中，提升硬盘利用率。
- 动态RAID采用EC（Erasure Code）算法，在全SSD盘配置下，能够动态调整CKG中CK的数目，从而满足系统可靠性和容量要求。

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%。

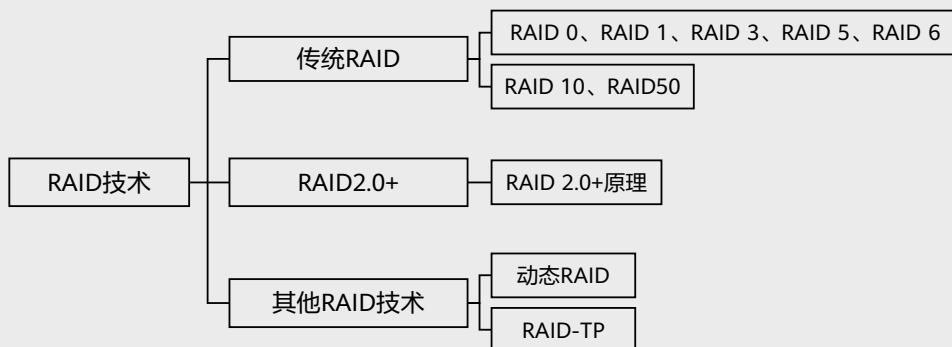
思考题

1. （简答题）条带和分条的区别是什么？
2. （简答题）如果一个用户关心的是可靠性和随机写性能，你会推荐什么样的RAID级别？
3. （判断题）RAID 10级别的RAID中的任何1个硬盘的故障不会影响数据访问。

- 参考答案：

1. 条带（strip）：硬盘中单个或者多个连续的扇区构成一个条带，他是一块硬盘上进一次数据读写的最小单元。它是组成分条的元素。分条（stipe）：同一硬盘阵列中的多个硬盘驱动器上的相同“位置”（或者说是相同编号）的条带的组合。
2. RAID 10
3. T

本章总结



学习推荐

- 华为官方网站
 - 企业业务: <http://enterprise.huawei.com/cn/>
 - 技术支持: <http://support.huawei.com/enterprise/>
 - 在线学习: <http://learning.huawei.com/cn/>
- 热门工具
 - HedEx Lite
 - 网络资料工具中心
 - 信息查询助手

- 热门工具:
 - HedEx Lite: 华为产品文档管理工具, 支持浏览、搜索、更新和管理产品资料
 - eStor: 图形化存储仿真平台, 该平台通过对真实华为OceanStor全闪存仿真模拟, 帮助广大ICT从业者和用户快速熟悉华为存储系列产品, 了解并掌握相关产品的操作和配置。
 - 网络资料工具中心: 汇集网络产品的资料工具, 是您进行投标支撑、网络规划、项目交付和升级维护的好助手。
 - 信息查询助手: 提供华为产品命令和告警信息查询。

Thank you.

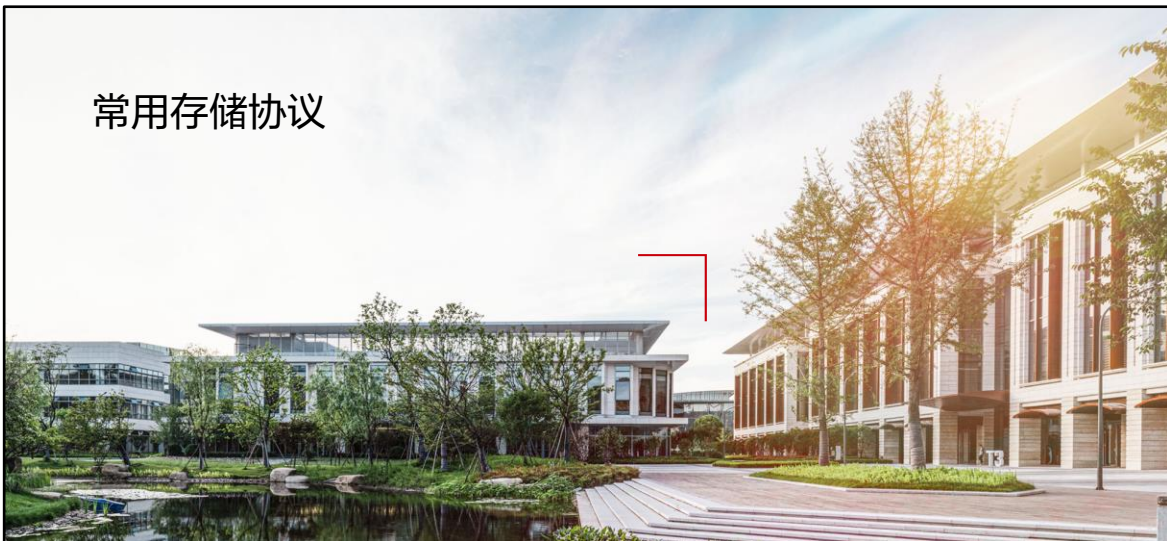
把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。
Bring digital to every person, home, and
organization for a fully connected,
intelligent world.

Copyright©2022 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.



常用存储协议



前言

- 协议是通信计算机双方必须共同遵从的一组约定。如怎么样建立连接、怎么样互相识别等。只有遵守这个约定，计算机之间才能相互通信交流。
- 协议不仅要定义双方互相通信所使用的语言，还要规定所使用的硬件、传输介质、传输协议以及各自的接口技术等。本章节主要介绍存储不同协议的定义及原理。

目标

- 学完本课程后，您将能够：
 - 了解存储系统中常见的协议；
 - 了解各种协议工作原理及其特点。

目录

1. SAN协议

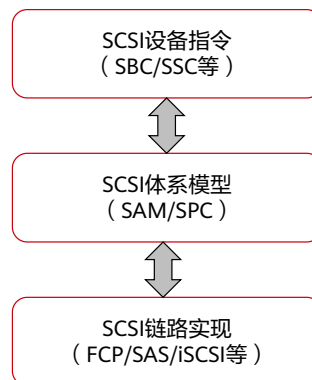
- SCSI、SAS
 - iSCSI、FC
 - PCIe、NVMe
 - RDMA、RoCE

2. NAS协议

3. 对象存储和大数据存储协议

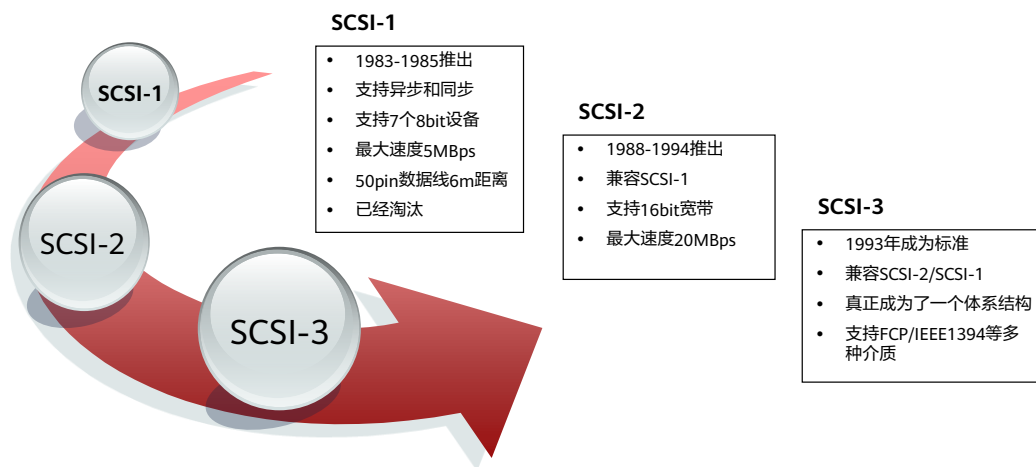
SCSI协议简介

- SCSI 即 小 型 计 算 机 接 口 (Small Computer System Interface)，指的是一个庞大协议体系，到目前为止经历了SCSI-1/SCSI-2/SCSI-3变迁。
- SCSI协议定义了一套不同设备利用该框架进行信息交互的模型和必要指令集。
- SCSI协议本质上同传输介质无关，SCSI可以在多种介质上实现，甚至是虚拟介质。

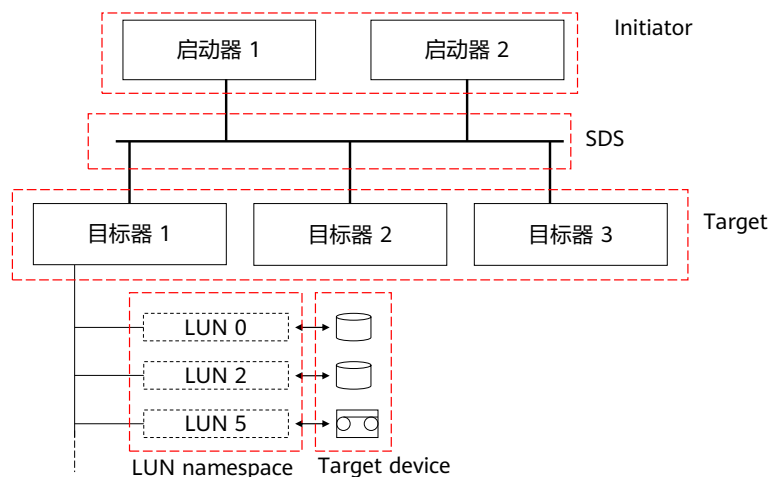


- SCSI的参考文档是一个大的集合，大致分为三类：设备/模型/链路。
 - SCSI体系结构文档：包括了基本架构模型SAM和SPC文档，描述了SCSI整体架构细节，如任务队列模型，基本共有指令模型等。
 - SCSI设备实现文档：包括了具体设备实现，比如块设备（磁盘）SBC，流设备（磁带）SSC的指令系统。
 - SCSI传输链路实现文档：包括FCP/SAS/iSCSI/FCoE等，描述了SCSI协议在介质上的实现细节。

SCSI的历史



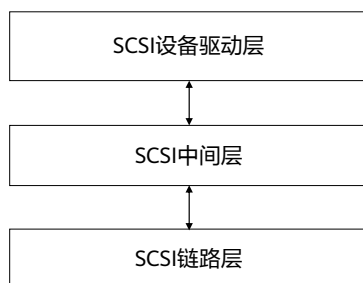
SCSI逻辑拓扑



- 启动器（Initiator），本质上，SCSI是一个C/S架构，其中客户端成为启动器，负责向SCSI目标器发送请求指令，一般主机系统都充当了启动器的角色。
- 目标器（Target），处理SCSI指令的服务端称为目标器，它接收来自主机的指令并解析处理，比如磁盘阵列的角色就是目标器。
- 逻辑单元（LUN），LUN是SCSI目标器中所描述的名字空间资源，一个目标器可以包括多个LUN，而且每个LUN的属性可以有所区别，比如LUN#0可以是磁盘，LUN#1可以是其他设备。
- SCSI的Initiator与Target共同构成了一个典型的C/S模型，每个指令都是“请求/应答”这样的模型来实现。
 - Initiator主要任务：发出SCSI请求。
 - Target主要任务：回答SCSI请求，通过LUN提供业务，并通过任务管理器提供任务管理功能。

SCSI启动器模型

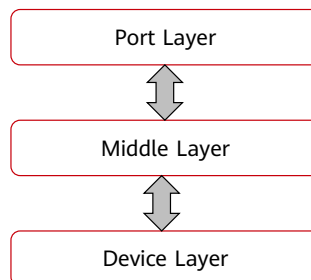
- 主机的SCSI系统一般工作中“启动器”模式下面，由于SCSI的体系结构来说，分为了“架构层（中间层）”，“设备层”，“传输层”，故此一般的操作系统，都将SCSI分为三个层次，包括Windows，Linux，AIX，Solaris，BSD。



- Windows将SCSI启动器分为三个逻辑层次，分别是Storage/Tape等驱动、ScsiPort和MiniPort。
 - 其中ScsiPort负责实现SCSI的基本框架性处理流程，比如设备发现，名字空间扫描等。
- Linux将SCSI启动器分为三个逻辑层次，分别是SCSI设备驱动、scsi_mod中间层和SCSI适配卡驱动（HBA）。
 - 其中scsi_mod中间层复杂处理SCSI设备无关和适配器无关的流程处理，比如一些异常，名字空间维护等。
 - HBA驱动提供SCSI指令的打包解包传输等链路实现细节，设备驱动实现特定的SCSI设备驱动，比如著名的sd（SCSI磁盘）驱动,st（SCSI磁带）驱动,sr(SCSI光盘设备)驱动等。
- Solaris的结构和Linux/Windows很类似，分别是SCSI设备驱动、SSA中间层和SCSI适配卡驱动。
- AIX的结构分为三层，分别是SCSI设备驱动、SCSI中间层和SCSI适配驱动。

SCSI目标器模型

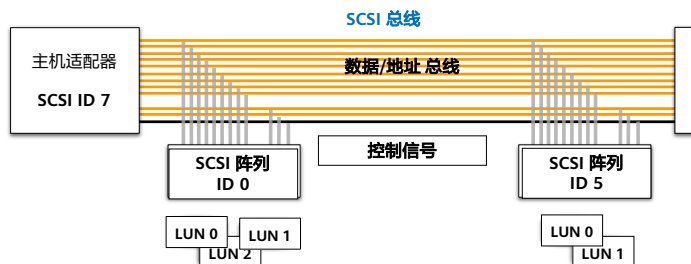
- Target一般依据SCSI体系结构，分为三个层次，分别是链路端口层（Port Layer），中间层（Middle Layer），目标设备层（Device Layer）。
- 其中最重要的是中间层，在中间层中将以SAM/SPC为规范，对LUN命名空间，链路端口，目标设备，任务，任务集，会话等进行管理维护。
- 端口层的驱动都以注册的形式动态载入，设备层的驱动也是动态载入。



- 目标器中的PORT模型：PORT主要完成的任务是对链路上携带的SCSI指令的解包和封包，如将指令打包为FPC或者iSCSI或者SAS等，或者将指令从FCP/iSCSI/SAS中解出。
- 目标器中的Device模型：Device是一个SCSI指令“分析器”，通过处理INQUIRY高速Initiator当前LUN是什么设备，通过READ/WRITE处理IO。
- 目标器中间层：中间层维护了LUN空间”，“任务集”，“任务（命令）”等模型。对LUN空间维护又两种方法：一种是所有PORT都看见的是一个全局的LUN，另一种是为每个PORT维护一个LUN空间。

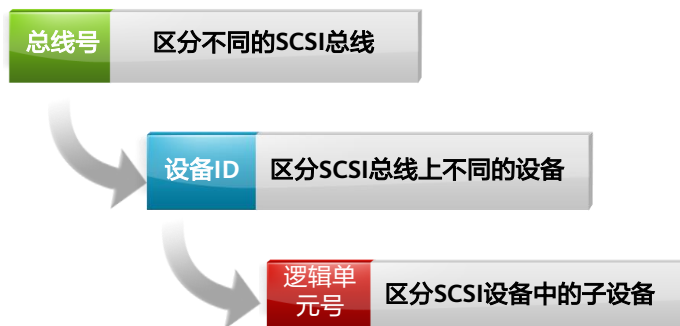
SCSI协议和存储系统

- SCSI协议是主机与存储设备通信的基本协议。
- DAS使用SCSI协议实现主机服务器与存储设备的互联。



- 控制器首先向总线处理器发出请求使用总线的信号。该请求被接受之后，控制器高速缓存就开始执行发送操作。在这个过程中，控制器占用了总线，总线上所连接的其它设备都不能使用总线。当然，由于总线具备中断功能，所以总线处理器可以随时中断这一传输过程并将总线控制权交给其它设备，以便执行更高优先级的操作。
- SCSI 控制器相当于一个小型CPU，有自己的命令集和缓存。SCSI是一种特殊的总线结构，可以对计算机中的多个设备进行动态分工操作，对于系统同时要求的多个任务可以灵活机动的适当分配，动态完成。

SCSI协议寻址



- 传统的SCSI控制器连接单条总线，相应的只具有一个总线号。企业级的一个服务器则可能配置了多个SCSI控制器，从而就可能有多条SCSI总线。在引入存储网络之后，每个FC HBA（Host Bus Adapter）或iSCSI（Internet SCSI）网卡也都各连接着一条总线，因此必须对每一条总线分配一个总线号，在他们之间依靠不同的总线号加以区分。
- 为了对连接在SCSI总线上的设备寻址，SCSI协议引入了SCSI设备ID和逻辑单元号LUN（Logical Unit Number）。在SCSI总线上的每个设备都必须有一个唯一的设备ID，当然服务器中的主机总线适配器也拥有自己的设备ID，固定为7。每条总线，包括总线适配器，最多允许有8个或者16个设备ID。设备ID一方面用以寻址，另一个作用是标识该设备在总线使用上的优先级。
- 每一个存储设备可能包括若干个子设备，如虚拟磁盘、磁带驱动器等。因此SCSI协议引入了逻辑单元号LUN ID，以便于对存储设备中的子设备进行寻址。
- 使用一个三元描述标识一个SCSI目标：总线号/目标设备ID/逻辑单元号LUN ID。

SAS的背景

- 并行总线已经发展到顶峰，到达带宽的极限；
- Fibre Channel、InfiniBand、Ethernet等串行总线对于存储应用来说都存在缺点：
 - a) FC：价格昂贵，适用于复杂组网、远距离的场景
 - b) InfiniBand：组网复杂，价格昂贵
 - c) iSCSI：时延高、传输速率慢



串行连接SCSI: SAS

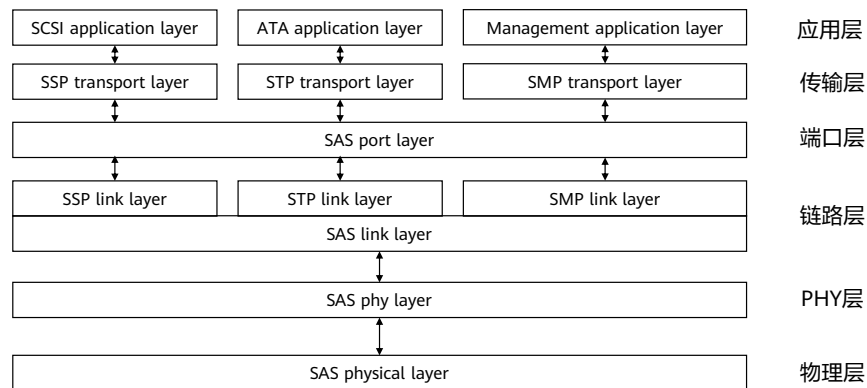
- SAS是SCSI总线协议的串行标准。
- 串行接口结构简单，支持热插拔，传输速度快，执行效率高。一般情况下，较大的并行电缆会带来电子干扰，SAS的电缆结构可以解决这个问题。SAS的电缆结构节省了空间，从而提高了使用SAS硬盘服务器的散热、通风能力。

什么是SAS

- SAS: Serial Attached SCSI, 是SCSI总线协议的串行标准, 即串行连接SCSI。
- SAS采用串行技术以获得更高的传输速率和更好的扩充性, 并兼容SATA硬盘。
- SAS采用点对点架构, 目前的传输速率高达3Gbps、6Gbps、12Gbps 或更高, 支持全双工模式。

- 更低的成本:
 - SAS的背板可以兼容SAS和SATA两类硬盘, 使用不同类型的硬盘时不需要再重新进行投资。
 - 不需要为SCSI和SATA标准分别设计不同的产品, 同时也减少了设计时布线的复杂度和PCB的层数, 因此降低了成本。
 - 系统集成商也不需要为给客户配置不同硬盘而采购不同的背板和线缆。
 - 可连接更多的设备。
 - SAS技术引入了SAS扩展器, 使SAS系统可以连接更多的设备, 其中每个扩展器允许连接多个端口, 每个端口可以连接SAS设备、主机或其他SAS扩展器。
- 高可靠性
 - 器件可靠性同SCSI硬盘、FC盘, 比SATA好。
 - 保留了经过验证的SCSI命令集。
- 高性能
 - 端口单向速率高。
- 与SATA兼容
 - SATA硬盘可以直接装入SAS环境里。
 - 可以在同一个系统里使用SATA或SAS盘, 符合现在日益流行的分级存储战略。

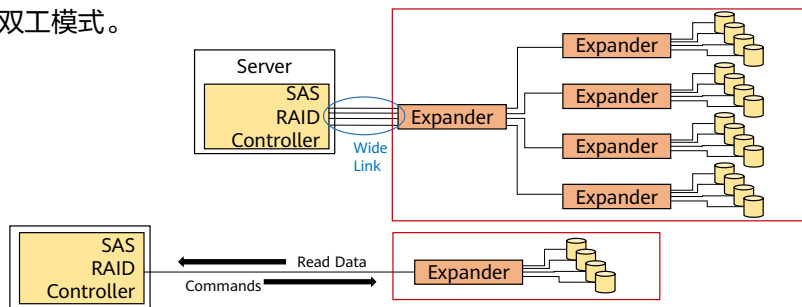
SAS协议层次结构



- SAS标准将SAS架构分为6层，从低到高依次是物理层、phy层、链路层、端口层、传输层和应用层，每层负责一定的功能。
 - 物理层：对于线缆、接头、收发器等硬件进行了定义；
 - PHY层：包括了最低级的协议，比如编码方案和供电/复位序列等等；
 - 链路层：描述的是如何控制phy层连接管理，以及原语、CRC校验和加解扰、速率匹配处理；
 - 端口层：描述的是链路层和传输层的接口，包括如何请求、中断、如何选择建立连接；
 - 传输层：定义了如何将所传输的命令、状态、数据封装在SAS帧中，以及如何分解SAS帧；
 - 应用层：描述了如何在不同类型的应用下使用SAS的细节。

SAS特点

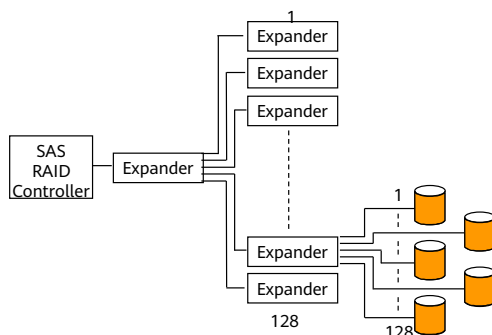
- SAS的串行通信方式允许多个数据通路全速与各个设备通信。
- SAS支持多个窄端口捆绑形成宽端口。
- SAS结构采用扩展器（expander）进行接口扩展，具有非常好的扩展能力。
- SAS采用全双工模式。



- SAS相对SCSI而言，最重要的提升在于：
 - 因为采用串行通信方式，可以提供更高的吞吐量，并为将来更高的性能成为可能。
 - 四个窄端口可以绑定成一个宽端口（Wide Link），提供更高的吞吐量。
- SAS采用全双工（双向）通信模式，而不是单向通信。传统的并行SCSI只可以在一个方向上进行通信，当设备接收到并行SCSI的一个数据包后，如果该设备要响应该数据包，就需要在上一个链路断开后，再重新建立一个新的SCSI通信链路。而SAS，则可以进行双向通信。每个SAS电缆有4根电缆，2根输入2根输出。SAS可以同时进行数据的读写，全双工的数据操作提高数据的吞吐效率。

SAS的可扩展性

- SAS结构采用扩展器（expander）进行接口扩展，具有非常好的扩展能力，1个SAS域最多可以连接16384个磁盘设备。



- SAS结构采用扩展器（expander）进行接口扩展，具有非常好的扩展能力，最多可以连接16384个磁盘设备。
- SAS Expander：SAS域中的互联设备，类似于以太网交换机，通过 Expander的级联可以大大增加终端设备的连接数，从而节约HBA花费。每个Expander最多可以连接128个终端设备或者128个Expander。1个SAS域由以下几个部分组成：SAS Expander、终端设备、连接设备（即SAS连接线缆）。
 - SAS Expander配备了一个地址的路由表跟踪，记录了所有SAS驱动器的地址。
 - 终端设备包括启动器（通常为SAS HBA卡）和目标器（SAS/SATA硬盘，也可以是处于目标模式的HBA卡）。
- SAS 域中不能形成环路，以保证其发现终端设备流程的正常进行。
- 实际使用中，因为带宽的原因，扩展器连接的终端设备比128个少很多。

SAS连线的原则

- SAS线缆一般有四个通道，每个通道现在常用的是12 Gb/s。
- SAS设备以环路（也称为链）的形式连接在一起。
- 线缆的带宽为4X12 Gb/s，限制了环路中的硬盘数。
- 当前一个环路中的最大硬盘数的最佳实践是168块盘，也就是说最多7个24个硬盘槽位的硬盘框组成一个环路。



- 大多数存储设备供应商现使用SAS线缆连接硬盘框到控制框，或者硬盘框之间的互连。SAS线缆通常将4个独立的通道（窄端口）捆绑成一个宽端口来提供更多的带宽。4个独立的通道都可以在12 Gb/s上进行，因此整个宽端口可以提供48 Gb/s 的带宽。为了确保SAS线缆上的数据量不超过整个SAS线缆的最大带宽，因此我们需要限制被连接在一个SAS环路上的硬盘的总个数。
- 对于华为的设备来说，这个最大硬盘数是168块盘，也就是说最多7个24个硬盘槽位的硬盘框组成一个环路。但是，这个前提条件是在该环路中的所有硬盘都是传统的SAS硬盘。现在SSD盘用得越来越多了，我们必须意识到SSD盘传输速度远远高于SAS盘。因此对于SSD盘，最大环路硬盘数的最佳实践则为96块盘，也就是4个24个硬盘槽位的硬盘框组成的环路。
- SAS线缆接口在SAS线缆单通道为6 Gb/s时称为 Mini SAS 线缆，现在单通道速度提升到12 Gb/s了，其相应的SAS线缆称为高密度Mini SAS线缆。

目录

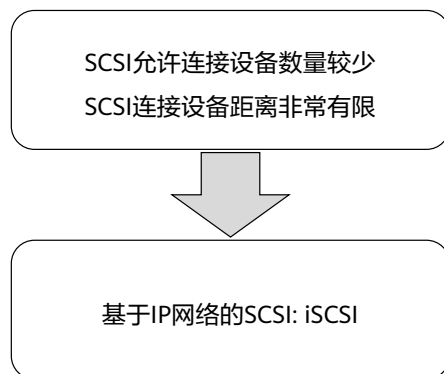
1. SAN协议

- SCSI、SAS
- iSCSI、FC
- PCIe、NVMe
- RDMA、RoCE

2. NAS协议

3. 对象存储和大数据存储协议

iSCSI的诞生



- iSCSI协议最早由IBM、CISCO、HP发起，2004年起作为正式的IETF标准，现有的iSCSI协议依据AM2(SCSI Architecture Model-2)。
- iSCSI (Internet SCSI) 把SCSI命令和块状数据封装在TCP中在IP网络中传输。
- iSCSI作为SCSI的传输层协议，基本出发点是利用成熟的IP网络技术来实现和延伸SAN。

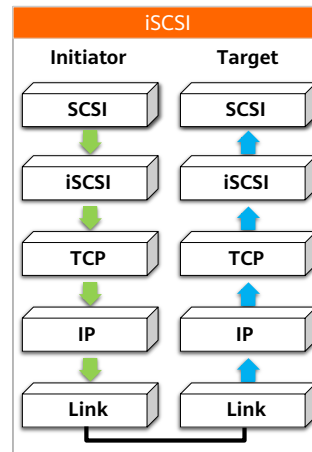
iSCSI简介



- iSCSI是互联网小型计算机系统接口（Internet Small Computer System Interface）的简称，是一种在TCP/IP上进行数据块传输的标准，可以理解为SCSI over IP。
- iSCSI（Internet SCSI）把SCSI命令和块状数据封装在TCP中在IP网络中传输。iSCSI作为SCSI的传输层协议，基本出发点是利用成熟的IP网络技术来实现和延伸SAN。
- SCSI协议层负责生成CDB，并将其送到iSCSI协议层，然后由iSCSI协议层进一步封装成PDU，经IP网络进行传送。

iSCSI启动器/目标器

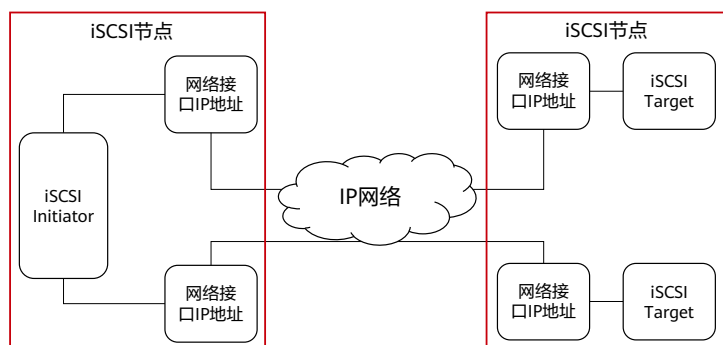
- 发起端（Initiator）
 - SCSI层负责生成CDB（命令描述符块），将CDB传给iSCSI。
 - iSCSI层负责生成iSCSI PDU（协议数据单元），并通过IP网络将PDU发给target。
- 目标器（Target）
 - iSCSI层收到PDU，将CDB传给SCSI层。
 - SCSI层负责解释CDB的意义，必要时发送响应。



- iSCSI的通信体系仍然继承了SCSI的部分特性，在iSCSI通信中，具有一个发起I/O请求的启动器设备（Initiator）和响应请求并执行实际I/O操作的目标器设备（Target）。在Initiator和Target建立连接后，Target在操作中作为主设备控制整个工作过程。
 - iSCSI Initiator：iSCSI启动器，可分为三种，即软件Initiator驱动程序、硬件的TOE（TCP Offload Engine，TCP卸载引擎）卡以及iSCSI HBA卡。就性能而言，软件Initiator驱动程序最差、TOE卡居中、iSCSI HBA卡最佳。
 - iSCSI Target：iSCSI目标器iSCSI Target，通常为iSCSI 磁盘阵列、iSCSI 磁带库等。
- iSCSI协议为Initiator和Target定义了一套命名和寻址方法。所有的iSCSI节点都是通过其iSCSI名称被标识的。这种命名方式使得iSCSI名称不会与主机名混淆。

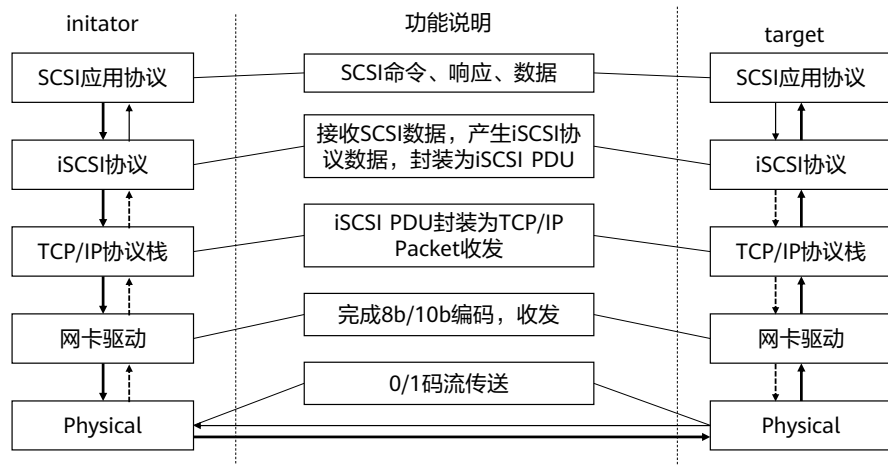
iSCSI体系结构

- iSCSI节点将SCSI指令和数据封装成iSCSI包，然后该数据封装被传送给TCP/IP层，再由TCP/IP协议将iSCSI包封装成IP协议数据以适合在网络中传输。



- 在支持iSCSI的系统中，用户在一台SCSI存储设备上发出存数据或取数据的命令，操作系统对该请求进行处理，并将该请求转换成一条或者多条SCSI指令，然后再传给目标SCSI控制卡。iSCSI节点将指令和数据封装（Encapsulation）起来，形成一个iSCSI包，然后该数据封装被传送给TCP/IP层，再由TCP/IP协议将iSCSI包封装成IP协议数据以适合在网络中传输。也可以对封装的SCSI命令进行加密处理，然后在不安全的网络上传送。
- 数据包可以在局域网或Internet上传送。在接收存储控制器上，数据包重新被组合，然后存储控制器读取iSCSI包中的SCSI控制命令和数据发送到相应的磁盘驱动器上，磁盘驱动器再执行初始计算机或应用所需求的功能。如果发送的是数据请求，那么将数据从磁盘驱动器中取出进行封装后发送给发出请求的计算机，而这整个过程对于用户来说都是透明的。尽管SCSI命令的执行和数据准备可以通过使用标准TCP/IP和现成的网络控制卡的软件来完成，但是在利用软件完成封装和解封装的情况下，在主机处理器上实现这些功能需要很多的CPU运算周期来处理数据和SCSI命令。如果将这些事务交给专门的设备处理，则可以将对系统性能的影响减少到最小程度，因此，发展在iSCSI标准下并执行SCSI命令和完成数据准备的专用iSCSI适配器是有必要的。iSCSI适配器结合了NIC和HBA的功能。这种适配器以块方式取得数据，利用TCP/IP处理引擎在适配卡上完成数据分化和处理，然后通过IP网络送出IP数据包。这些功能的完成使用户可以在不降低服务器性能的基础上创建一个基于IP的SAN。

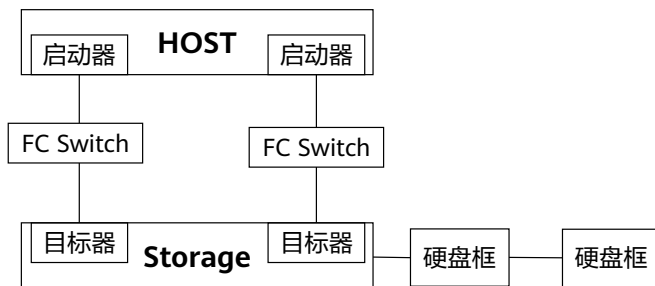
iSCSI和SCSI、TCP和IP的关系



- PDU：协议数据单元。
- TCP/IP：传输控制协议/因特网互联协议。
- 8b/10b编码是将一组连续的8位数据分解成两组数据，一组3位，一组5位，经过编码后分别成为一组4位的代码和一组6位的代码，从而组成一组10位的数据发送出去。

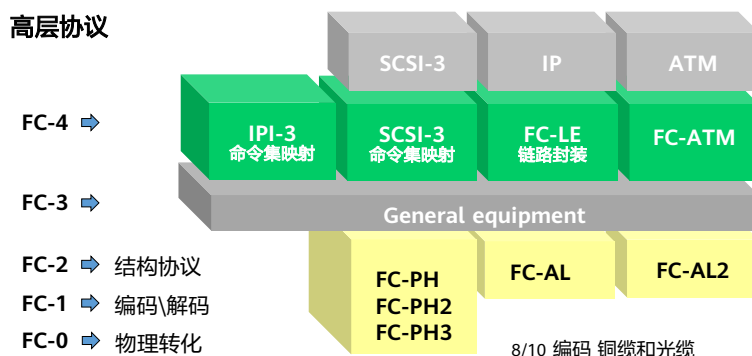
存储中的FC

- FC（Fibre Channel）可称为FC协议，或者FC网络、FC互联。由于其性能较高，逐渐发展到前端作为主机接口，并逐渐发展出点对点、交换机等组网。
- 存储网络引入FC后，有如下优势：
 - 提高扩展性
 - 增加传输距离
 - 解决安全性问题



- 像TCP/IP一样，FC协议集同样具备TCP/IP协议集以及以太网中的很多概念，如FC交换、FC交换机、FC路由、FC路由器、SPF路由算法等。

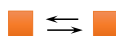
FC协议结构



- FC-0: 物理界面 这是最下一层, FC-0定义物理连接。它的主要用途是选择协议操作的不同物理介质和数据速率。这种方式确保系统具有最大限度的灵活性, 允许采用现有连接线和不同的技术以满足不同系统的需求。常用连接线为铜缆和光缆。
- FC-1: 字节编码 本层记录8b/10b传输编码用于平衡传输位流。此外, 编码还可以作为数据传输和接受错误检测机制。选择8b/10b编码是因为它的优秀的传输特性, 这个极为稳定的编码降低了部件设计成本同时保证了良好的传输密度易于时钟恢复。注: 8b/10b设计也用于 IBM ESCON。
- FC-2: 数据分发 本层包含在网络内发送数据的基本规则。包括:
 - 1) 如何将数据切分为小帧;
 - 2) 一次可以发送多少数据(流控制);
 - 3) 帧应当发送到什么地方。它同时包括根据应用定义服务等级。
- FC-3: 通用服务 本层定义包括条带(通过多条通道传输数据), 多点传送(将一个信息发送到多个目标)和查寻组(为一个节点映射多个端口)等先进功能。于是当FC-2为单一端口定义功能时, FC-3层为可以跨端口定义功能。
- FC-4: 高层协议映射 本层将光纤通道性能映射为IP, SCSI或ATM等协议。SCSI是FC协议的子集。

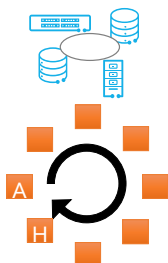
FC的拓扑结构

点对点



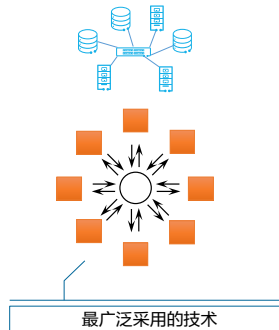
只能连接 两个设备
(直接连接)

FC-AL仲裁环



最多支持127个设备

FC交换网络

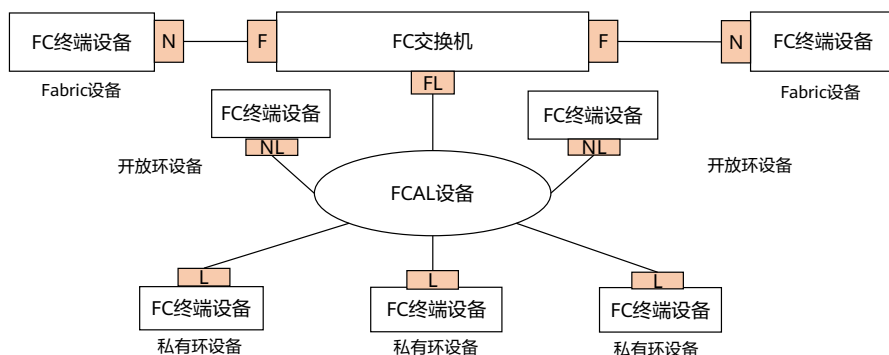


最多支持1千6百万个设备

- 与以太网类似，FC提供了几种网络拓扑结构：
- 点对点连接：
 - 最简单的拓扑结构，允许两节点之间直接通信，一般是一个存储设备和一台服务器。
- FC-AL拓扑：
 - 类似于以太网共享总线拓扑，但是连接方式不是总线，而是一条仲裁环路（Arbitral Loop）。每个FC-AL设备首尾相连构成了一个环路。
 - 数据帧在仲裁环内是一跳一跳被传输的，并且任何时刻数据帧只能按照一个方向向下游传输。如图A节点想与H节点通信，在A节点赢得仲裁后，向H节点发送数据帧，但是此环的数据是顺时针方向传递，所以A发出的数据帧，按照B-C-D-E-F-G-H的顺序传送，需要绕一圈，效率低。
- Fabric拓扑：
 - 与以太网交换拓扑类似，Fabric意为“网状结构”，表明这种拓扑是一个网状交换矩阵。
 - 相对于仲裁环路来说，转发效率大大提升。
 - FC设备通过光纤或者铜线等各种标注的线缆连接到Fabric交换机，便可以实现各个节点基于FC Fabric拓扑方式的点对点通信。

FC协议中的7种端口类型

- 在FC网络中，存在7种类型的端口。



- 光纤通道解除了工作站对每一端口的管理工作，每一端口只需管理自己到Fabric简单的点对点连接，其他的Fabric功能由光纤通道交换机来完成。
- 设备（节点）端口：
 - N_Port：Fabric直连接设备。
 - NL_Port：Loop连接设备。
- 交换机端口：
 - E_Port：扩展端口（交换机到交换机）。
 - F_Port：Fabric端口，是Fabric面向用户N_Port的连接点。
 - FL_Port：Fabric Loop端口。
 - G_Port：普通（Generic）端口，可以转换为E或F端口。
 - U_Port：通用（Universal）端口，用于描述自动端口检测的术语。

FC适配器

- FC主机总线适配器，即FC HBA（Host Bus Adapter）。支持连接FC网络的应用，实现高带宽高性能存储组网方案。



- FC HBA卡：光纤通道卡，Fibre Channel HBA，用于服务器和光纤阵列的连接。

目录

1. SAN协议

- SCSI、SAS
- iSCSI、FC
- PCIe、NVMe
- RDMA、RoCE

2. NAS协议

3. 对象存储和大数据存储协议

PCIe

- PCI Express（又称PCIe）是一种高性能、高带宽串行通讯互连标准，最早由英特尔提出，后由外设组件互连特别兴趣组（PCI-SIG）制定，以取代基于总线的通信架构，如：PCI、PCI Extended（PCI-X）以及加速图形端口（AGP）。



- PCI特点：
 - 总线结构简单、成本低、设计简单；
 - 并行总线无法连接太多设备，总线扩展性比较差；
 - 当连接多个设备时，总线有效带宽将大幅降低，传输速率变慢。
- 1991年，Intel公司首先提出了PCI的概念，随着现代化处理器技术的发展，在互连领域，使用高速差分总线替代并行总线是大势所趋。与单端并行信号相比，高速差分信号用于更高的时钟频率，PCIe总线应运而生。
- 相对于传统的PCI总线，PCI Express有如下优点：
 - 双通道，高带宽，传输速率快：实现类似全双工传输模式（RX和TX分开）；传输速率高，第一代PCIe X1的带宽是2.5千兆位/秒（Gbps），第二代则达到5.0Gbps，而PCIe 3.0标准则能支持8.0Gbps的速率，PCIe 4.0支持16.0Gbps，PCIe 5.0高达32.0Gbps；对更大的带宽，则可以通过扩展链路数目来实现，而带宽则是其N条通路的N倍。
 - 兼容性：在软件层保持与PCI兼容，版本升级，向下兼容PCI软件。
 - 易用性：支持热插拔，PCIe总线接口插槽中包括“热插拔检测信号”，可以像USB进行热拔插和热交换。
 - 具有错误处理和先进的错误报告功能：得益于PCI Express总线是分层结构，其软件层具有错误处理和提供错误报告能力，详细后续介绍。

为什么用PCIe

转向PCIe的主要动机是以更低的生产成本获得明显提高的系统吞吐量、可扩展性和灵活性，而上述这些是传统基于总线的互连几乎根本无法达到的。



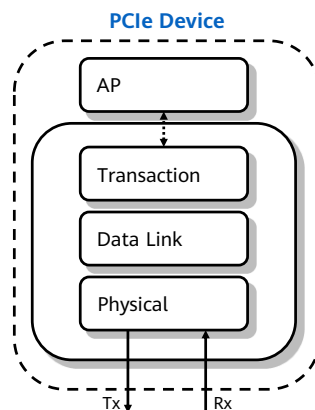
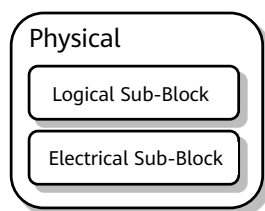
高性能、高带宽串行通讯互连标准：PCI Express (PCIe) 。

- PCI Express标准的制定是着眼未来的，它还在继续发展为系统提供更高的吞吐量。在PCIe标准继续利用最新的技术以提供不断增加的吞吐量的同时，利用分层协议、通过使驱动程序保持与现有PCI应用的软件兼容性将简化从PCI到PCIe的过渡。
- PCIe协议的特性包括：
 - 点对点连接
 - 高可靠
 - 树形组网
 - 全双工
 - 基于帧结构的传输

PCIe协议结构

- PCIe设备的层次：主要分为物理层、数据链路层、事务层、应用层：

- 物理层（Physical Layer）
- 数据链路层（Data Link Layer）
- 事务层（Transaction Layer）
- 应用层（Application Layer）



- PCI Express总线架构中的物理层主要决定总线的物理特征。在未来的发展中，可以通过速度提升或更改编解码方式来进一步提升PCI Express总线性能，而这些变更都只会作用于物理层，不会影响其它结构，为升级带来了便利。
- 数据链路层（Data Link Layer）
 - 数据链路层的重要作用就是要保证PCI Express总线传输的数据包的正确性和可靠性。它会检查数据包封装是否完整正确，并将序列号和冗余校验代码（CRC）添加到数据中进行检测和纠错，并采用ack/nack握手协议进行检测纠错。
- 事务层（Transaction Layer）
 - 处理层的主要作用有两个，一是接受来自软件层送来的读、写请求，或者本身创建一个请求封装包传到数据链路层，这种数据包称为“处理层数据包”（Transaction Layer Packet, TLP另一个作用是接受从链路层传来的响应数据包（Data Link Layer Packet, DLL P），并将其与相关的软件请求相关联，传送至软件层处理。

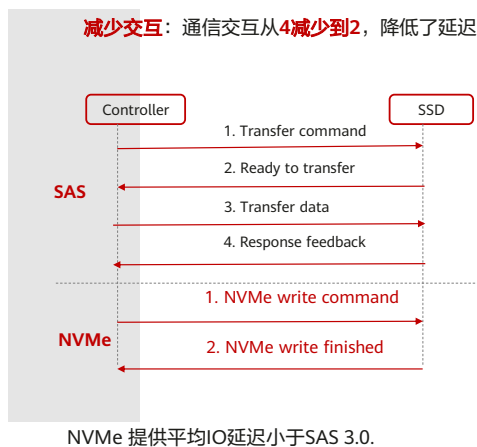
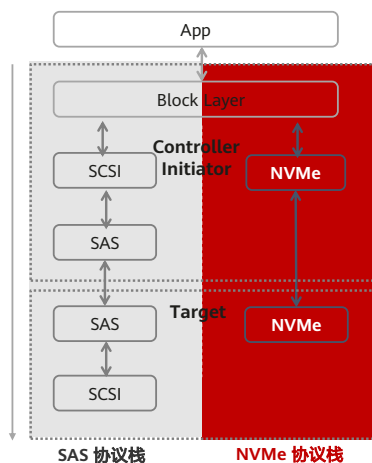
NVMe

- NVMe，全称Non-Volatile Memory Express，非易失性存储器标准。
- NVMe标准是面向PCI-E SSD的，使用原生PCI-E通道与CPU直连可以免去SATA与SAS接口的外置控制器（PCH，Platform Controller Hub）与CPU通信所带来的延时。
- PCI-E是接口形态和总线标准，NVMe是为PCI-E SSD定制的标准接口协议。



- 就存储整个流程来说，NVMe不仅仅是逻辑上的协议接口，还有一种指令标准、一种指定协议。
- 此规范目的在于充分利用PCI-E通道的低延时以及并行性，还有当代处理器、平台与应用的并行性，在可控制的存储成本下，极大的提升固态硬盘的读写性能，降低由于AHCI接口（Advanced Host Controller Interface）带来的高延时，彻底解放SATA时代固态硬盘的极致性能。

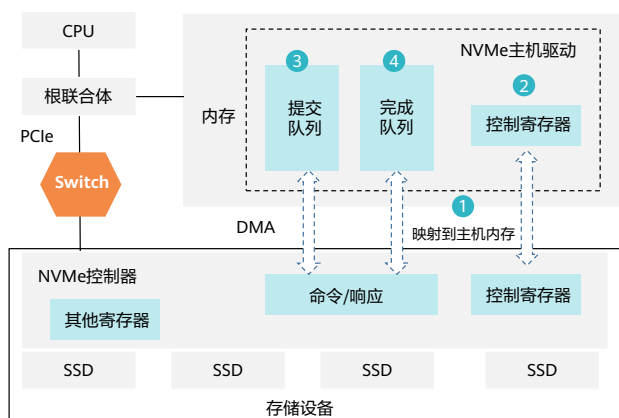
NVMe协议栈



- 在传输路径层面，SAS 全闪存阵列的 IO 从前端服务器通过存储设备的 FC/IP 前端接口协议传输到 CPU-> CPU 通过 PCIe 链路和 PCIe 交换机 --> SAS 芯片 -->SAS Expander-->SAS SSD。
- 而华为 NVMe 全闪存支持端到端 NVMe，数据 I/O 从前端服务器通过存储设备的 FC-NVMe/NVMe Over RDMA 前端接口协议到达 CPU，后端采用 100G RDMA 直接到达 NVMe SSD。显然，NVMe 全闪存阵列的 CPU 和 NVMe SSD 直接通信，传输路径更短，传输效率更高，传输延迟更少。
- 在软件协议解析层面，SAS 和 NVMe 全闪存 阵列在数据写入的协议交互技术差异较大。一次完整的主机数据写入请求，如果采用SAS后端的SCSI协议，一共需要通过 4 次协议交互，才能完成一次完整的数据写入 操作；而华为NVMe全闪存只需要2次协议交互就能完成一次写操作，处理写请求的效率比 SAS 全闪存高 1 倍。

NVMe工作原理（DMA）

- NVMe使用DMA（Direct Memory Access，直接内存存取）传输NVMe指令（Command）。

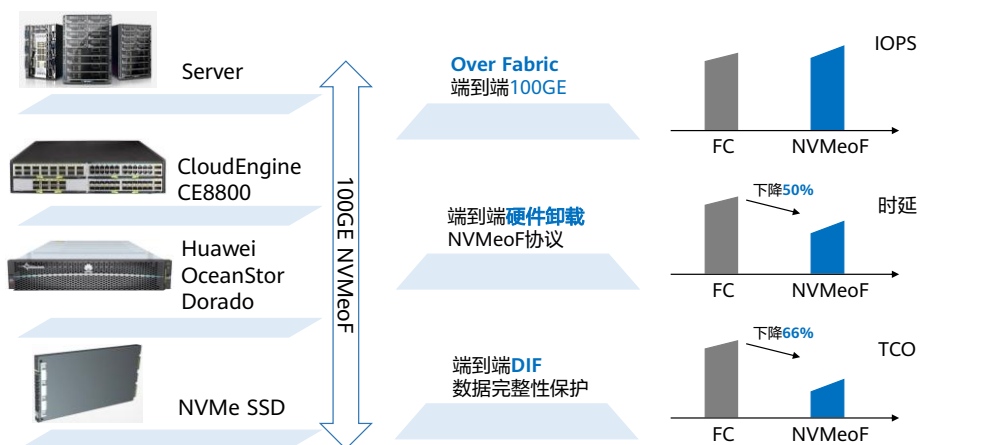


- 1、NVMe主机驱动将存储设备的控制寄存器映射到主机内存空间。
- 2、通过寄存器实现主机对存储设备的控制和状态读取。
- 3、主机将I/O请求放入请求队列，主机发送NVMe指令。
- 4、通过DMA和设备传输数据，完成后NVMe设备向主机反馈结束。

NVMe支持多队列传输数据，大并发、低时延。

- NVMe的队列分为2种，其中一种是用于管理的队列，称为Admin Queue（管理队列），仅有一个。另外一种是指令队列（Command Queue），最多可以有65535个。其中命令队列的数量和模式都是通过管理队列来设置的。其中每一个队列实际上是一个队列对，也就是包括两个队列，分别是提交队列（Submission Queue）和完成队列（Completion Queue）。提交队列用于主机端向NVMe设备发送NVMe命令，而完成队列则用于NVMe设备向主机反馈命令执行情况。实际上NVMe还有另外一种模式，就是多个提交队列共享同一个完成队列的情况。
- 更多NVMe工作原理，请参考NVMe系列白皮书，<https://nvmexpress.org/white-papers/>。

NVMe的优势和应用



- NVMe的优势：
 - 较低时延：NVMe执行命令不需要读取寄存器，从而单个IO延迟低；
 - 带宽高：例如x4 PCIe可为单个驱动器提供高达4GB/s的吞吐量；
 - 较高IOPS：NVMe可以把最大队列深度从32提升到64000，SSD的IOPS能力也会得到大幅提升。
 - 较低能耗：自动功耗状态切换和动态能耗管理功能大大降低了功耗；
 - 驱动适用性广：NVMe标准的出现解决了不同PCIe SSD之间的驱动适用性问题。
- RDMA（Remote Direct Memory Access，远程直接内存访问功能）利用相关的硬件和网络技术，使服务器的网卡之间可以直接读内存，最终达到高带宽、低时延和低资源消耗率的效果。但是RDMA专用的InfiniBand网络架构封闭，无法兼容现网，使用成本较高。RoCE（RDMA over Converged Ethernet）技术的出现有效解决了这些难题。RoCE即使用以太网承载RDMA的网络协议，有两个版本：RoCEv1是一种链路层协议，不同广播域下无法使用；RoCEv2是一种网络层协议，可以实现路由功能。
- 华为 OceanStor Dorado 全闪存凭借 NVMe over Fabrics实现SSD资源全共享，同时提供32G FC-NVMe和NVMe Over 100G RDMA组网设计，实现前端网络连接、后端硬盘框连接、scale-out的控制器互联均采用同一种网络协议。

目录

1. SAN协议

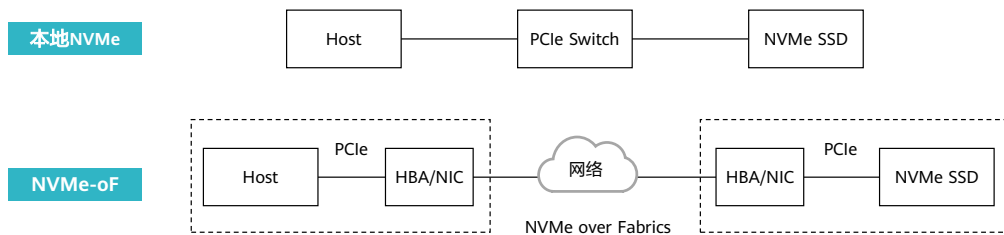
- SCSI、SAS
- iSCSI、FC
- PCIe、NVMe
- RDMA、RoCE

2. NAS协议

3. 对象存储和大数据存储协议

NVMe over Fabrics简介

- NVMe over Fabrics (简称NVMe-oF)，是NVMe协议在以太网和FC网络上的拓展。其目的是：想像访问本地NVMe设备一样访问远端NVMe设备。
- 基于PCIE的扩展性问题，2016年NVMe发布标准定义了两种类型的Fabric传输，实现了端到端的全路径NVMe协议：
 - **NVMe over Fabrics using RDMA**
 - NVMe over Fabrics using Fibre Channel (FC-NVMe)

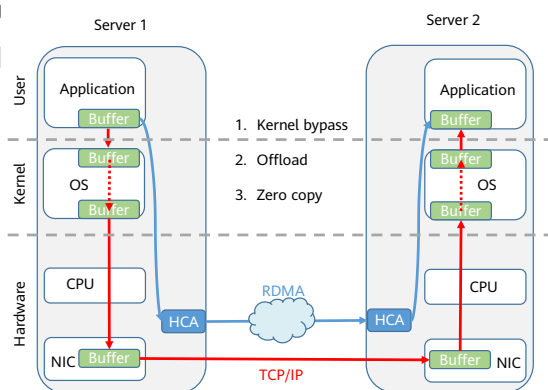


- 基于PCIE存在几个问题：
 - PCIE扩展性受限，最大支持255节点
 - PCIE不方便多主机共享
 - PCIE无法拉远

RDMA简介

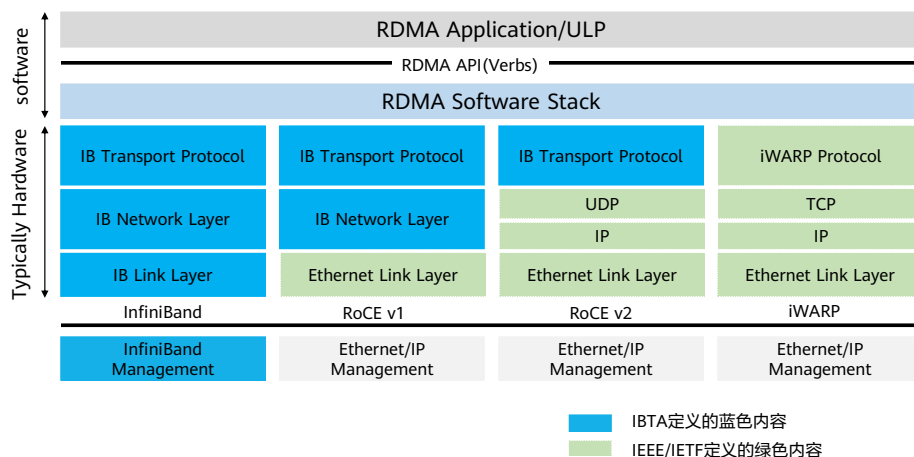
- RDMA (Remote Direct Memory Access) 远程直接内存访问，这是一种通过网络在两个服务器的应用软件之间搬运缓冲区里的数据的方法。

- 低时延
- 高吞吐量
- CPU、OS资源占用很少



- 传统模式与RDMA模式对比：
 - 相比传统DMA的内部总线IO，RDMA通过网络在两个端点的应用软件之间实现Buffer的直接传递。
 - 相比传统的网络传输，RDMA又无需操作系统和协议栈的介入。
 - RDMA可以轻易实现端点间的超低延时、超高吞吐量传输，且基本不需CPU、OS资源介入，不必为网络数据的处理和搬移耗费过多其他资源。
- HCA, host channel adapter, 主机通道适配器。

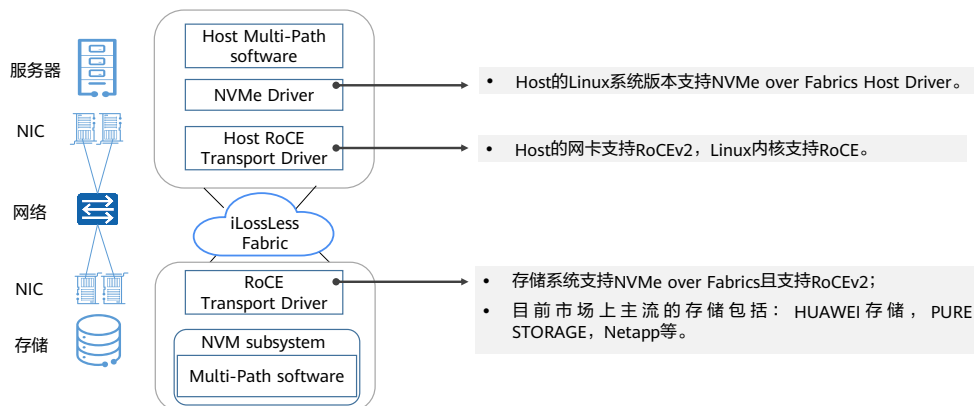
RDMA承载网络



- 目前大致有三类RDMA网络，分别是Infiniband、RoCE、iWARP。其中Infiniband是一种专为RDMA设计的网络，从硬件级别保证可靠传输，而RoCE和iWARP都是基于以太网的RDMA技术，支持相应的verbs接口，如图所示。
- RoCE协议中，RoCE v1是基于以太网链路层实现的RDMA协议，交换机需要支持PFC等流控技术，在物理层保证可靠传输；而RoCE v2是以太网TCP/IP协议中UDP层实现。
- InfiniBand (IB)：从一开始就支持RDMA的新一代网络协议。需要支持该技术的网卡NIC和交换机。
- RDMA over Ethernet (RoCE)：一种允许通过以太网进行RDMA的网络协议。其较低的网络标头是以太网标头，其较高的网络标头（包括数据）是IB标头，这支持在标准以太网基础设施（交换机）上使用RDMA，只有网卡是特殊的，支持RoCE。
- 互联网广域RDMA协议（iWARP）：允许通过TCP执行RDMA的网络协议。IB和RoCE中存在的功能在iWARP中不受支持。这支持在标准以太网基础设施（交换机）上使用RDMA。只有网卡应该是特殊的，并且支持iWARP（如果使用CPU卸载），否则所有iWARP堆栈都可以在SW中实现，并且丧失了大部分RDMA性能优势。

NVMe over RoCE介绍

- NVMe over RoCE是NVMe-oF(基于RDMA)的一种，在性能、成本、网络管理、技术发展等方面有显著优化，正逐步成为NVMe over fabric的最佳应用。



- Multi-Path是存储多路径软件，实现多路负载均衡，以及冗余备份。

目录

1. SAN协议

2. **NAS协议**

- 文件系统介绍

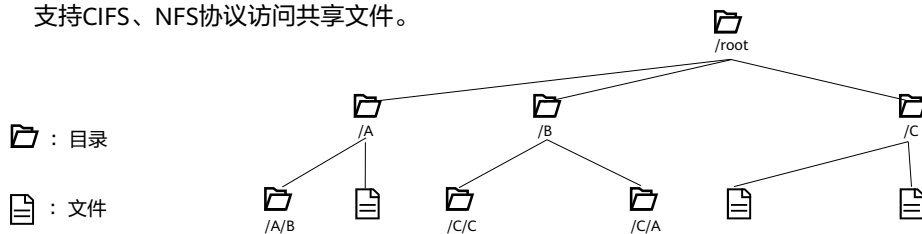
- CIFS、NFS、跨协议

- HTTP、FTP、NDMP

3. 对象存储和大数据存储协议

文件系统介绍

- 文件系统是计算机用来管理数据的一种方式，它以文件和目录的形式将数据组织起来。文件系统组成一个多叉树形结构，叶子结点是文件，中间节点是各级目录，最顶级是根目录。
- 文件服务用于提供稳定可靠的文件共享功能，也是NAS存储系统的基本特性。文件服务同时支持CIFS、NFS协议访问共享文件。



目录

1. SAN协议

2. **NAS协议**

- 文件系统介绍

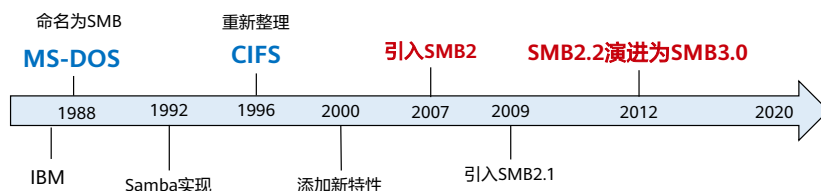
- CIFS、NFS、跨协议

- HTTP、FTP、NDMP

3. 对象存储和大数据存储协议

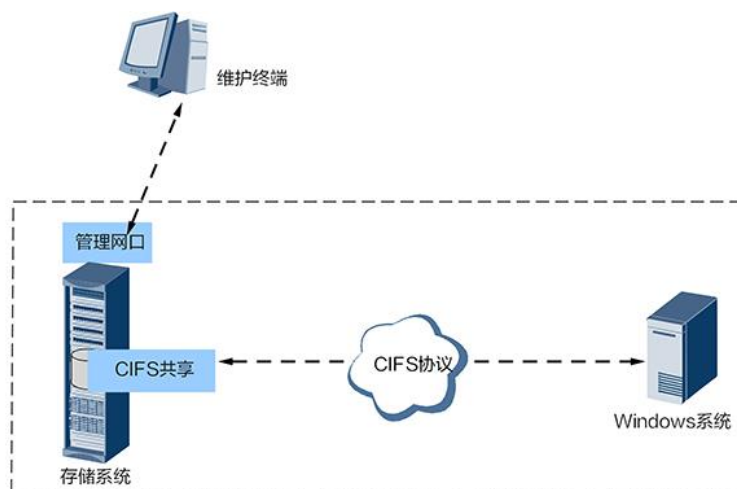
CIFS协议

- 1996年，微软提出将SMB改称为CIFS（Common Internet File System protocol），并加入了许多新的功能。现在CIFS指SMB的统称，目前各个版本分别为SMB1，SMB2，SMB3.0。
- CIFS采用C/S模式，基本网络协议：TCP/IP和IPX/SPX。CIFS共享特性主要应用于Windows客户端进行文件共享的场景，常用于无域环境和AD域环境中。



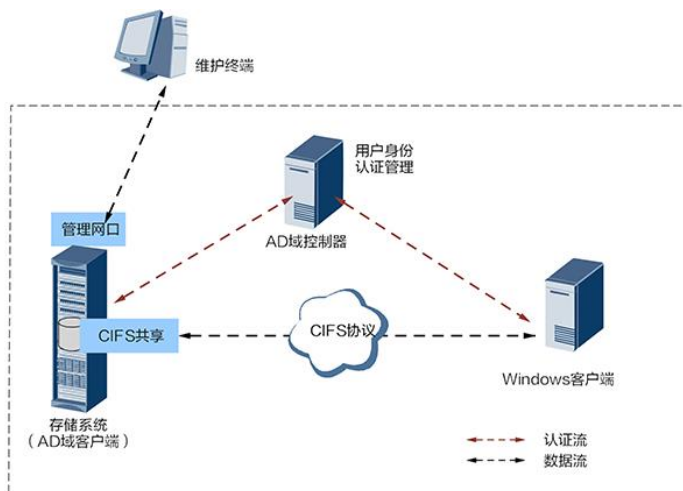
- SMB：Server Message Block，服务器消息块，由微软开发的一种软件程序级的网络传输协议，主要用来使得一个网络上的机器共享计算机文件、打印机、串行端口和通讯等资源。
- CIFS：Common Internet File System protocol，传统的微软环境下的文件共享协议，基于服务器消息块协议。
 - CIFS是一个基于网络的共享协议，其对网络传输的可靠性要求很高，所以它通常是使用TCP/IP。
- CIFS（Common Internet File System）共享主要应用于Internet和Internet中的Windows主机访问网络中的文件或其他资源。通过CIFS共享，Windows客户端能够识别并访问网络中存储系统提供的共享资源，用户能够像使用本机一样对保存在存储系统中的文件进行读、写、创建等操作。可以解决大量用户访问共享文件速度下降和系统响应慢的问题。

无域环境中的CIFS共享



- 存储系统使用CIFS共享的方式，把文件系统以目录的形式共享给某个用户。该用户可以查看或访问给予其权限的共享目录。
- 存储系统作为CIFS服务器，通过CIFS协议，向客户端提供对文件系统的共享访问。客户端将共享目录映射到本地后，用户可以像访问本地文件一样远程访问服务器中的文件系统。在存储系统中通过设置本地认证的用户名和密码，确定允许访问该文件系统的本地认证信息。

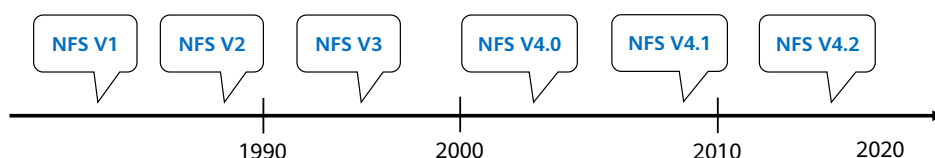
AD域环境中的CIFS共享



- 随着局域网、广域网规模越来越大，很多企业使用AD域进行Windows网络管理，使得管理更便捷、更具有扩展性。
- 存储系统能够加入AD域，作为AD域的客户端，实现和AD域环境的无缝对接。AD域控制器中保存了域环境中所有的用户信息、群组信息等。AD域客户端访问存储系统提供的CIFS共享时，认证均通过AD域控制器完成。通过设置AD域用户的权限，对不同域用户访问的共享目录进行不同的权限控制。同时，存储系统支持AD域客户端只能访问与其名称相同的共享目录，无法查看并访问其他域客户端的共享目录。

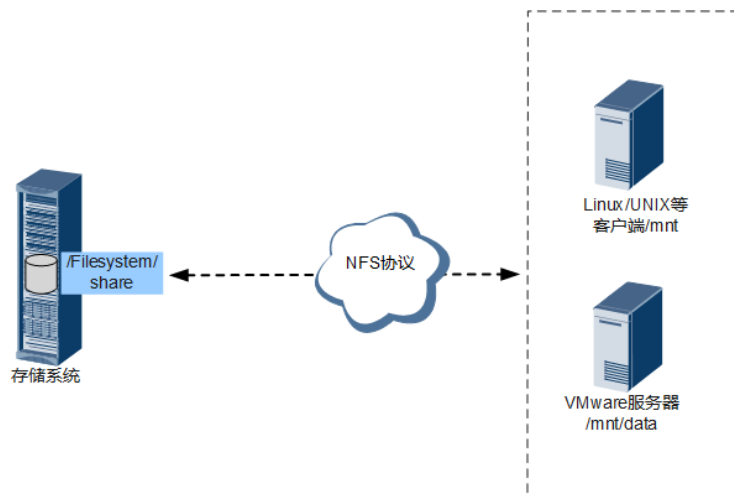
NFS协议

- 网络文件系统（ Network File System ），在Linux/Unix环境里普遍使用的网络文件共享协议，是IETF标准规范的网络文件共享协议。
- NFS使用客户端/服务器架构。服务器程序向其他计算机提供对文件系统的访问，客户端程序对共享文件系统进行访问。NFS通过网络让不同类型操作系统的客户端彼此共享文件，适用于多种网络环境，例如无域环境、LDAP域环境和NIS域环境等。



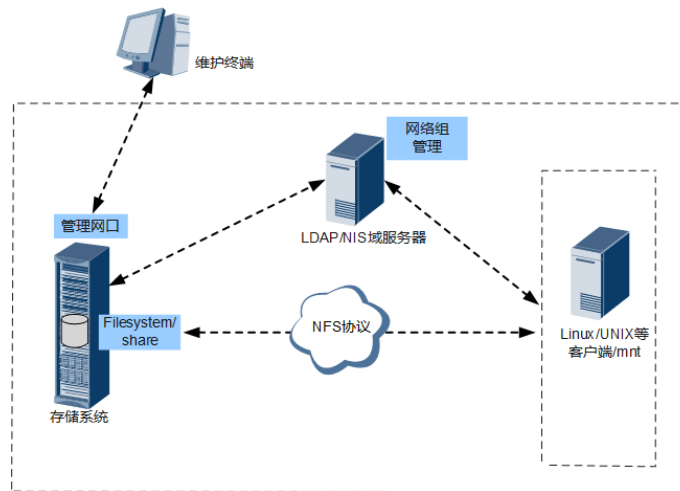
- NFS： Network File System protocol,传统的UNIX环境文件共享协议。
 - NFS是一个客户机/服务器应用程序，使用远程过程调用（RPC）在计算机之间进行通信。用户就像使用自己的计算机一样存储、更新远程NAS上的文件。
 - 用户的系统需要一个NFS客户端连接到NFS服务器。NFS用于独立的传输，所以它使用TCP或UDP。
 - 用户或系统管理员可以使用NFS挂载所有文件系统或文件系统的一部分（任意目录或子目录分级树中的一部分）。被挂载的文件系统可使用权限来控制访问（例如，只读或读-写权限）。
- NFS v3和NFS v4的区别:
 - NFS v4设计成了一种有状态的协议，自身实现了文件锁功能和获取文件系统根节点功能，不需要NLM和MOUNT协议协助了。NFS v3是一种无状态的协议，NFS v3需要NLM协议（ Network Lock Manager ），协助才能实现文件锁功能。
 - NFS v4增加了安全性，支持RPCSEC-GSS身份认证。
 - NFS v4只提供了两个请求NULL和COMPOUND，所有的操作都整合进了COMPOUND中，客户端可以根据实际请求将多个操作封装到一个COMPOUND请求中，增加了灵活性。
 - NFS v4文件系统的命令空间发生了变化，服务器端必须设置一个根文件系统（fsid=0），其他文件系统挂载在根文件系统上导出。
 - NFS v4相较于NFS v3增强了跨平台特性。

无域环境下NFS共享



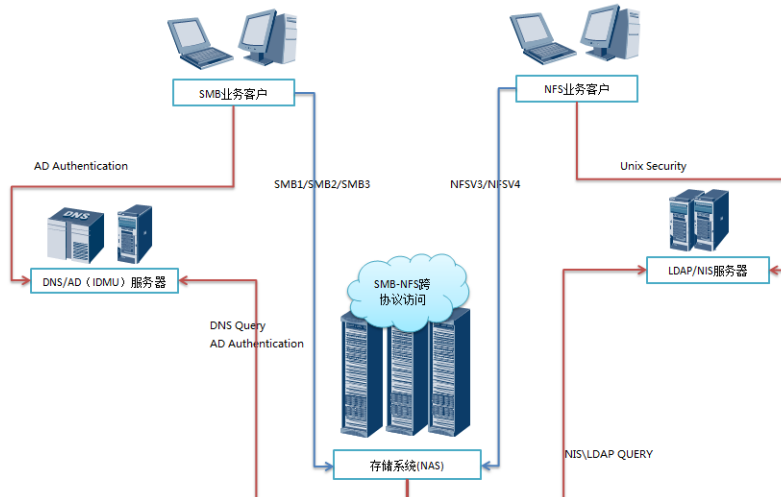
- 无域环境下的NFS共享常应用于中小企业中。存储系统作为NFS服务器，通过NFS协议，向客户端提供对文件系统的共享访问。客户端将共享文件挂载到本地后，用户像访问本地文件一样远程访问服务器中的文件。在存储系统中通过设置客户端IP地址的过滤条件，限定访问NFS共享的客户端信息。

域环境下NFS共享



- 域实现了例如帐号管理，应用管理，网络管理等集中式的管理。Linux环境中可以将域分为LDAP（Lightweight Directory Access Protocol）域和NIS（Network Information Service）域两种。
 - LDAP是一个开放的、可扩展的网络协议。基于LDAP的认证应用主要是实现一个以目录为核心的用户认证系统，即LDAP域环境。在LDAP域环境中，当客户端的用户需要访问应用程序时，客户端将用户名和密码提供给LDAP服务器，LDAP服务器将其与目录数据库中的认证信息进行比对来确定用户身份的合法性。
 - NIS是一种可以集中管理系统数据库的目录服务技术，提供了一个网络黄页的功能。NIS使用客户端/服务器架构。如果某个用户的用户名以及密码被保存在NIS服务器中的数据库中，NIS允许此用户在NIS客户端上登录，并且通过维护NIS服务器中的数据库，可以统一管理整个局域网系统中的网络信息。
- 在域环境中，当客户端需要访问存储系统提供的NFS共享时，存储系统通过域服务器网络组完成可访问IP地址的验证，保证文件系统数据可靠。

CIFS-NFS跨协议访问服务



- 存储系统支持对同一个文件系统或Dtree同时设置NFS共享和CIFS共享。不同协议的客户端可以同时访问同一个文件系统或Dtree。由于Windows环境与Linux/Unix环境对于用户认证及访问控制等方面采取的机制不同，存储系统通过统一管理实现不同环境下用户的映射关系和权限控制，以保障CIFS-NFS跨协议访问的安全性。
- 用户使用CIFS协议的客户端访问存储系统时，存储系统首先对本地或AD域用户进行接入身份认证。若访问的文件或目录具有UNIX权限（UNIX Mode bits）时，认证时会根据已设置的用户映射规则，将CIFS用户映射为NFS用户，再对用户访问权限进行Unix权限鉴权。
- 用户使用NFS协议的客户端访问存储系统时，若访问的文件或目录具有NT ACL，存储系统根据已设置的用户映射规则将NFS用户映射为CIFS用户，再根据映射出的CIFS用户，进行NT ACL权限鉴权。

目录

1. SAN协议

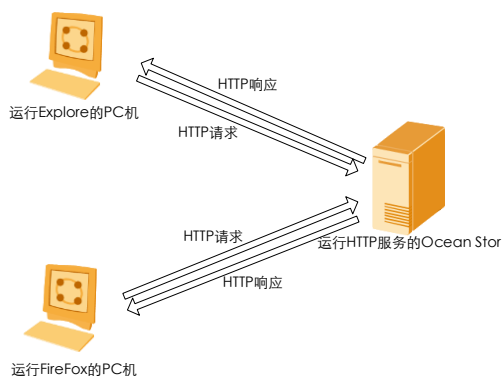
2. **NAS协议**

- 文件系统介绍
- CIFS、NFS、跨协议
- HTTP、FTP、NDMP

3. 对象存储和大数据存储协议

HTTP共享文件系统

- 存储系统支持HTTPS共享文件系统。通过开启HTTPS服务，可以将一个文件系统以HTTPS的方式共享。
- 共享资源管理主要通过WebDAV协议实现，它是一个HTTP的扩展协议，它允许客户端对服务器上共享目录中的资源进行复制、移动、修改、锁定、解锁、搜索等处理。



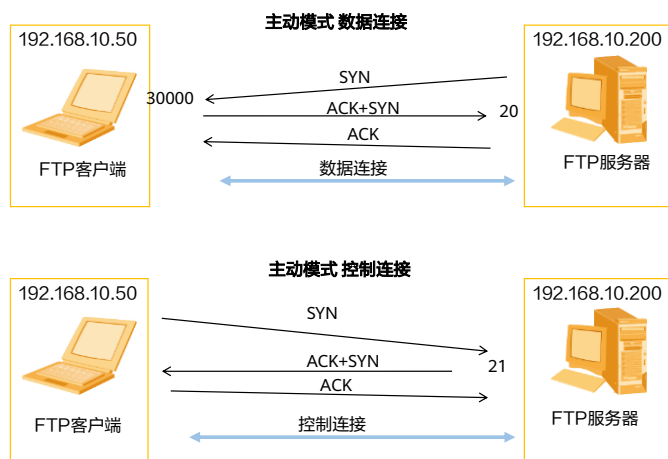
- 超文本传输协议HTTP(Hypertext transfer protocol)是一种详细规定了浏览器和万维网服务器之间互相通信的规则，通过Internet传送万维网文档的数据传送协议。
- HTTP协议定义Web客户端如何从Web服务器请求Web页面，以及服务器如何把Web页面传送给客户端的协议。
- HTTP采用短连接方式进行报文传输，每次传输结束后连接就中断。HTTP协议使用80端口，HTTPS协议使用443端口。

FTP概述

- FTP（File Transfer Protocol，文件传输协议）用于在远端服务器和本地主机之间传输文件，是IP网络上传输文件的通用协议。
- FTP协议在TCP/IP协议族中属于应用层协议，用于在远端服务器和本地客户端之间传输文件，使用TCP端口20和21进行传输。端口20用于传输数据，端口21用于传输控制消息。FTP协议基本操作在RFC959中进行了描述。
- 存储系统支持FTP服务，在设备上配置FTPS服务器功能后，用户可以利用FTP客户端安全访问远程的设备。
- OceanStor存储系统FTP有两种工作方式：
 - 主动方式（PORT）：建立数据连接时由FTP服务器发起连接请求，当FTP客户端处于防火墙内时不适用（如FTP客户端处于私网内）。
 - 被动方式（PASV）：建立数据连接时由FTP客户端发起连接请求，当FTP服务器限制客户端连接其高位端口（一般情况下大于1024）时不适用。

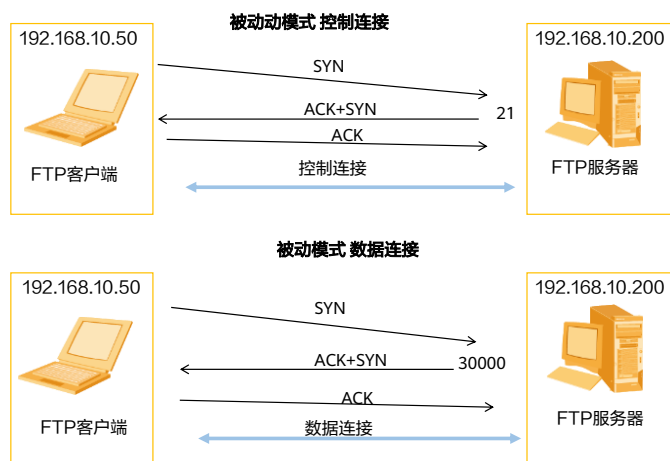
- FTP 是File Transfer Protocol(文件传输协议)的英文简称，而中文简称为“文传协议”。用于Internet上的控制文件的双向传输。同时，它也是一个应用程序(Application)。基于不同的操作系统有不同的FTP应用程序，而所有这些应用程序都遵守同一种协议以传输文件。
- FTP允许用户以文件操作的方式(如文件的增、删、改、查、传送等)与另一主机相互通信。

FTP服务器主动模式



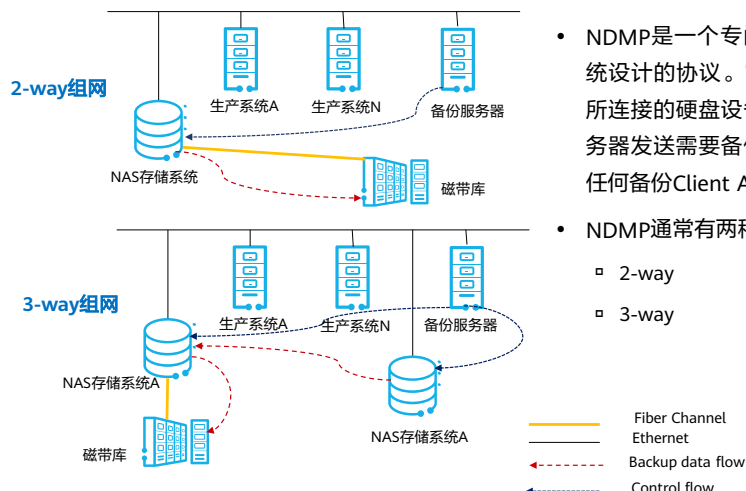
- FTP客户端会采用port命令告诉服务器，自己会在哪一个IP地址，哪一个临时端口等待FTP服务器使用20端口为源端口主动发起的数据连接请求，因为是FTP服务器主动发起的数据连接请求，所以叫主动模式。假设本例中FTP客户端用port命令告诉FTP服务器的临时端口为30000，IP地址为192.168.10.50。
- 控制连接建立成功后会接着建立数据连接，如图。数据连接建立成功的标志是在FTP客户端里看到了FTP服务器上的文件列表，大家经常碰到的FTP客户端连接FTP服务器时候的列目录超时，就是数据连接无法建立的标志。

FTP服务器被动模式



- FTP客户端会采用pasv命令告诉FTP服务器，自己会主动发起数据连接请求，然后FTP服务器会使用port命令告诉FTP客户端，自己会在哪一个IP，哪一个临时端口上等待数据连接的建立。假设本例中FTP服务器在192.168.10.200的30000端口等待FTP客户端数据连接请求。然后FTP客户端会向192.168.10.200的30000端口发起数据连接请求，因此是FTP服务器被动接受数据请求，所以叫被动模式。
- 数据连接建立成功的标志是在FTP客户端里看到了FTP服务器上的文件列表，大家经常碰到的FTP客户端连接FTP服务器时候的列目录超时，就是数据连接无法建立的标志。

NDMP协议



- NDMP是一个专门为NAS设备的数据备份系统设计的协议。它可以让NAS设备直接向其所连接的硬盘设备或者位于网络上的备份服务器发送需要备份的数据，这个过程不需要任何备份Client Agent的参与。

- NDMP通常有两种组网方式：

- 2-way
- 3-way

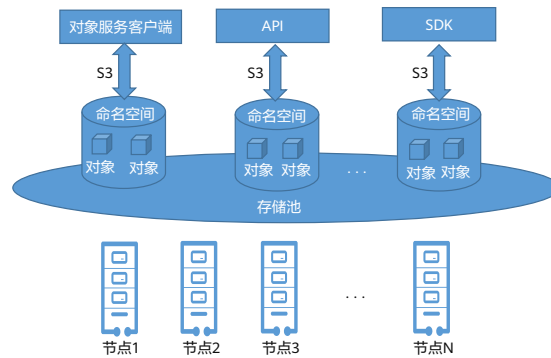
- 传统的NAS存储备份方式实现过程：
 - NAS设备是个封闭的存储系统，一般很难在NAS设备上安装备份软件的Client Agent(只能安装在生产系统上)，所以传统的网络备份都是先通过CIFS或者NFS共享协议从NAS设备中将需要备份的数据读出，然后再通过网络将这些数据传送到备份服务器上备份。
 - 传统的这种备份方式缺点就是会占用网络资源及生产系统、备份服务器的资源，性能偏低，很难满足大数据量的备份。
- 2-way组网方式中，备份介质不再与备份服务器连接，而是直接与NAS存储系统连接。备份服务器执行备份操作的时候，通过以太网将备份命令传递给NAS存储，NAS存储直接将数据备份到与其连接的磁带库上。
 - NDMP 2-way这种备份方式下，数据流直接传递到备份介质上，大大提升了传输性能，同时减少了服务器资源的占用。但是由于磁带库是连接给NAS存储的，所以磁带库只能给这一台NAS存储做备份。
 - 磁带库的价格是非常昂贵的，为了可以让不同的NAS存储共享磁带设备，NDMP同时还支持3-way这种备份方式。
- 3-way备份方式下，NAS存储系统可以通过专用备份网络传递备份数据到连接了磁带库的NAS存储上，由该NAS存储将数据备份到磁带库中去。

目录

1. SAN协议
2. NAS协议
- 3. 对象存储和大数据存储协议**
 - 对象存储协议
 - 大数据存储协议

对象服务

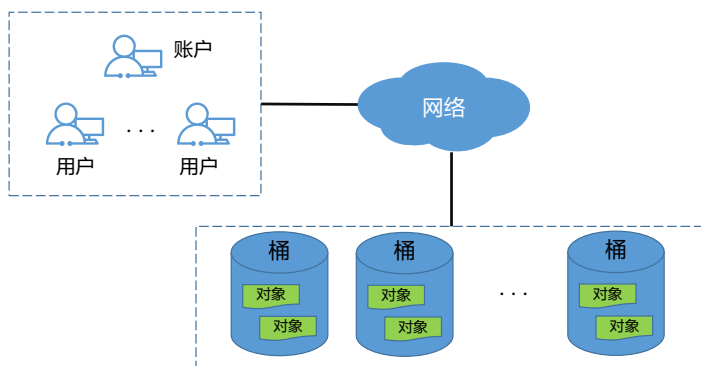
- 对象服务是基于对象的海量数据存储服务，提供易扩展、高安全、高可靠、低成本的数据存储能力。
- 对象服务对外提供标准S3协议接口，是一种基于HTTP/HTTPS协议的REST风格协议，用户可通过对象服务客户端、API接口以及SDK方式方便地管理和使用对象服务数据，以及开发多种类型的上层应用。



- 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。

S3相关概念

- 桶（Bucket）、对象（Object）、帐户、用户之间的关系如图所示。



- Simple Storage Service（S3），简单存储服务，是一个公开的云存储服务，Web 应用程序开发人员可以使用它存储数字资产，包括图片、视频、音乐和文档。
- S3 提供一个 RESTful API 以编程方式实现与该服务的交互。
- S3 存储和检索的资产被称为对象。对象存储在桶（bucket）中。您可以用硬盘进行类比：对象类似于文件，桶类似于文件夹（或目录）。

RESTful

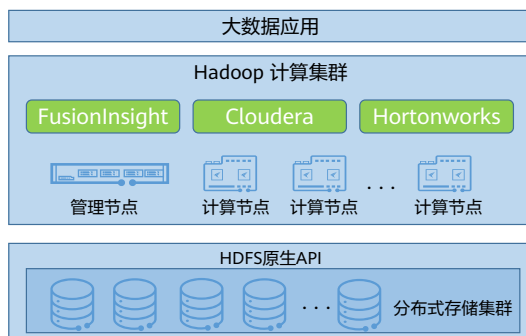
- REST
 - REST是REpresentational State Transfer的缩写，全称是Resource Representational State Transfer，是指资源在网络中以某种表现形式进行状态转移。
 - Resource：资源，即数据。比如 newfeed，friends等；
 - Representational：某种表现形式，比如用图像、视频等；
 - State Transfer：状态变化。通过HTTP动词实现。
 - REST就是选择通过使用http协议和URI，利用client/server模型对资源进行增删改查操作。
 - REST不是规范，而是一种架构，一种网络应用的架构。可以把REST理解成一种设计模式，只不过REST这种设计模式是应用在网络应用架构上的。
- RESTful
 - 如果一个架构符合REST原则，就称它为RESTful架构；
 - 它主要用于客户端和服务交互类的软件。基于这个风格设计的软件可以更简洁，更有层次，更易于实现缓存等机制。

目录

1. SAN协议
2. NAS协议
- 3. 对象存储和大数据存储协议**
 - 对象存储协议
 - 大数据存储协议

大数据服务

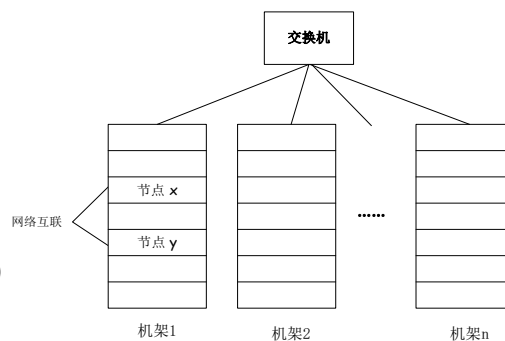
- 大数据服务提供基于原生HDFS的大数据存算分离方案，实现存储与计算资源按需配置，提供相同用户体验的同时降低总拥有成本；支持与原有计算存储一体化架构共存。
- 大数据服务提供原生HDFS接口，对接FusionInsight、Cloudera CDH、Hortonworks HDP等大数据平台，完成大数据的存储与计算，提供大数据分析服务供上层大数据应用使用。



- 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。
- HDFS: Hadoop Distributed File System，是开源Hadoop的主要组件之一，适用于大数据存储。HDFS由NameNode和DataNode组成。

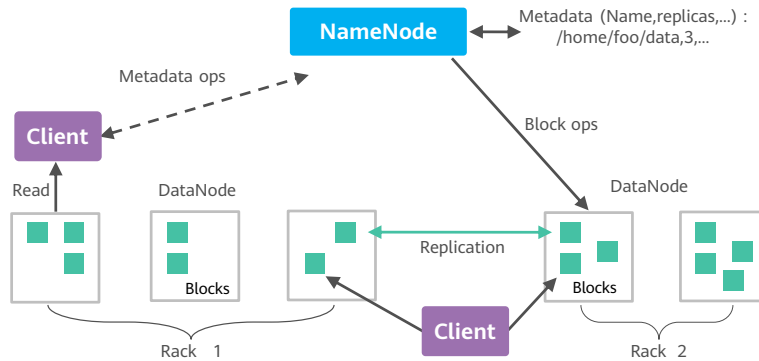
分布式文件系统

- 分布式文件系统把文件分布存储到多个计算机节点上，成千上万的计算机节点构成计算机集群。
- 目前的分布式文件系统所采用的计算机集群都是由普通硬件构成的，这就大大降低了硬件上的开销。
- HDFS(Hadoop Distributed File System) 基于Google发布的GFS论文设计开发，运行在通用硬件上的分布式文件系统。



HDFS系统架构

HDFS Architecture



- HDFS架构包含三个部分：NameNode，DataNode，Client。
 - NameNode：NameNode用于存储、生成文件系统的元数据。运行一个实例。
 - DataNode：DataNode用于存储实际的数据，将自己管理的数据块上报给NameNode，运行多个实例。
 - Client：支持业务访问HDFS，从NameNode,DataNode获取数据返回给业务。多个实例，和业务一起运行。

HDFS通信协议

- HDFS是一个部署在集群上的分布式文件系统，因此，很多数据需要通过网络进行传输
 - 所有的HDFS通信协议都是构建在TCP/IP协议基础之上的
 - 客户端通过一个可配置的端口向名称节点主动发起TCP连接，并使用客户端协议与名称节点进行交互
 - 名称节点和数据节点之间则使用数据节点协议进行交互
 - 客户端与数据节点的交互是通过RPC（ Remote Procedure Call ）来实现的。在设计上，名称节点不会主动发起RPC，而是响应来自客户端和数据节点的RPC请求

思考题

1. （多选题）FC拓扑结构有以下哪些？
 - A. 仲裁环网络
 - B. 点对点网络
 - C. 交换网络
 - D. 双交换网络
2. （多选题）NFS协议有哪些版本？
 - A. NFS V1
 - B. NFS V2
 - C. NFS V3
 - D. NFS V4

- 参考答案：

- ABC
- ABCD

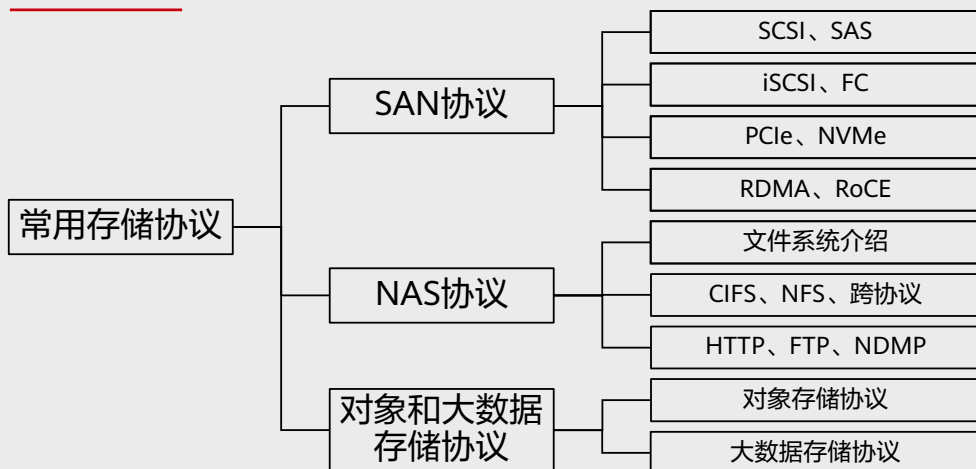
思考题

3. （多选题）文件共享协议有以下哪些？
- A. HTTP
 - B. iSCSI
 - C. NFS
 - D. CIFS
4. （多选题）CIFS协议工作需要经过以下哪几个步骤？
- A. 协议握手
 - B. 安全认证
 - C. 共享连接
 - D. 文件操作
 - E. 断开连接

- 参考答案：

- ACD
 - ABCDE

本章总结



学习推荐

- 华为官方网站
 - 企业业务: <http://enterprise.huawei.com/cn/>
 - 技术支持: <http://support.huawei.com/enterprise/>
 - 在线学习: <http://learning.huawei.com/cn/>
- 热门工具
 - HedEx Lite
 - 网络资料工具中心
 - 信息查询助手

- 热门工具:
 - HedEx Lite: 华为产品文档管理工具, 支持浏览、搜索、更新和管理产品资料
 - eStor: 图形化存储仿真平台, 该平台通过对真实华为OceanStor全闪存仿真模拟, 帮助广大ICT从业者和用户快速熟悉华为存储系列产品, 了解并掌握相关产品的操作和配置。
 - 网络资料工具中心: 汇集网络产品的资料工具, 是您进行投标支撑、网络规划、项目交付和升级维护的好助手。
 - 信息查询助手: 提供华为产品命令和告警信息查询。

Thank you.

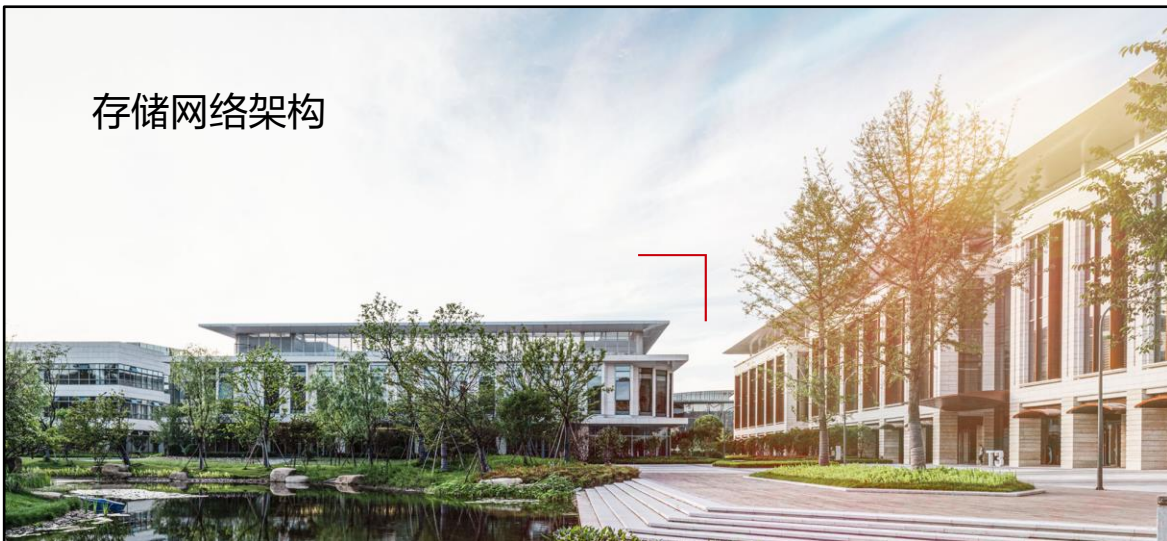
把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。
Bring digital to every person, home, and
organization for a fully connected,
intelligent world.

Copyright©2022 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.



存储网络架构



前言

- 随着主机、磁盘、网络等技术的发展，存储系统架构演变的同时，存储网络架构也在不断的发展，以适应业务需要。本章主要介绍了存储网络架构。

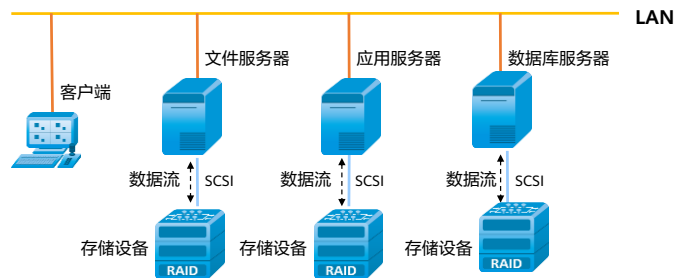
目标

- 学完本课程后，您将能够：
 - 了解存储网络架构演变
 - 了解存储组网技术演变

目录

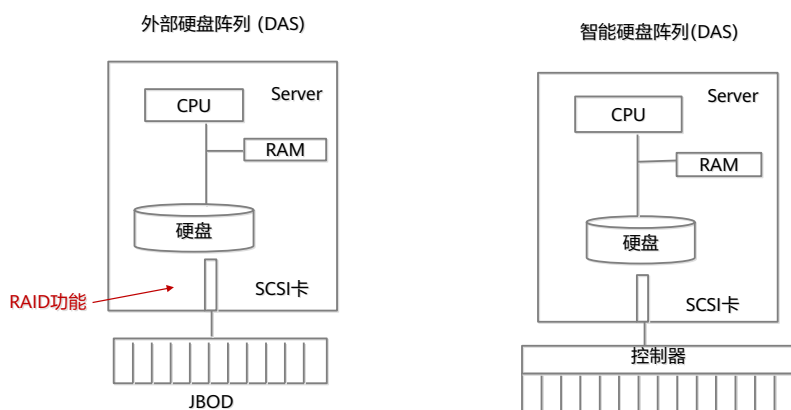
1. DAS
2. NAS
3. SAN
4. 分布式

DAS (Direct Attached Storage)



- 存储设备直接连接到主机
- 数据分散管理
- 存储容量利用率低
- 扩展性差
- DAS特别适合于对存储容量要求不高、服务器的数量很少的中小型局域网，其主要的优点在于存储容量扩展的实施非常简单，投入的成本少而见效快。

DAS直连式存储



- DAS的定义：DAS是1个或多个直接连接到使用它们的服务器上的指定存储设备，这些存储设备为服务器提供块级数据访问服务。
- DAS分类：基于存储设备与服务器间的位置关系，DAS分为内部DAS和外部DAS。
- 外部硬盘阵列：
 - JBOD 即Just a Bunch Of Disks，在逻辑上把几个物理磁盘串联在一起，其目的纯粹是为了增加磁盘的容量，并不提供数据安全保障。能够解决内置存储有限硬盘槽位，容量扩展不足的问题。但仍然是基于单硬盘存放数据，可靠性差。
- 智能硬盘阵列：
 - 控制器中包含RAID功能、大容量Cache，同时使得磁盘阵列具有多种实用的功能，配置专用管理软件进行配置管理。
- 主要使用SCSI线缆连接主机和存储。

DAS遇到的挑战

可扩展性低	• 可连接到主机的端口数目有限 • 可寻址的磁盘数有限 • 距离限制
维护不方便	• 维护时系统需要下电
资源共享性较差	• 阵列前端端口、存储空间 难以共享 • 导致资源孤岛：如存储空间紧张的DAS不能共享存储资源过剩的DAS的剩余空间

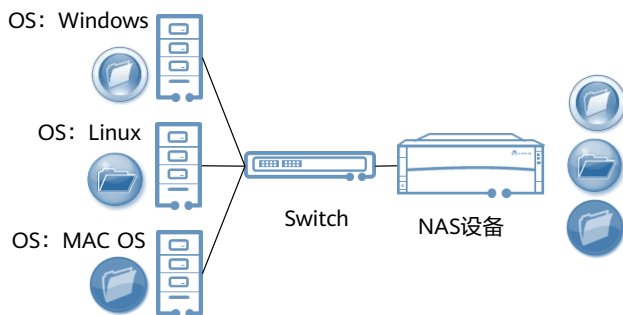
- 直连式存储依赖服务器主机操作系统进行数据的 IO 读写和存储维护管理，数据备份和恢复要求占用服务器主机资源（包括 CPU、系统 IO 等），数据流需要回流主机再到服务器连接着的磁带机（库），数据备份通常占用服务器主机资源 20-30%，因此许多企业用户的日常数据备份常常在深夜或业务系统不繁忙时进行，以免影响正常业务系统的运行。直连式存储的数据量越大，备份和恢复的时间就越长，对服务器硬件的依赖性和影响就越大。
- 直连式存储与服务器主机之间的连接通道通常采用SCSI连接，带宽为10MB/s、20MB/s、40MB/s、80MB/s 等，随着服务器 CPU 的处理能力越来越强，存储硬盘空间越来越大，阵列的硬盘数量越来越多，SCSI 通道将会成为 IO 瓶颈；服务器主机 SCSI ID 资源有限，能够建立的 SCSI 通道连接有限。
- 无论直连式存储还是服务器主机的扩展，从一台服务器扩展为多台服务器组成的群集 (Cluster)，或存储阵列容量的扩展，都会造成业务系统的停机，从而给企业带来经济损失，对于银行、电信、传媒等行业 7×24 小时服务的关键业务系统，这是不可接受的。并且直连式存储或服务器主机的升级扩展，只能由原设备厂商提供，往往受原设备厂商限制。

目录

1. DAS
- 2. NAS**
3. SAN
4. 分布式

NAS

- NAS (Network Attached Storage) 网络附加存储，即将存储设备连接到现有的网络上，提供数据和文件服务。
- NAS最常用的网络共享协议：CIFS和NFS。

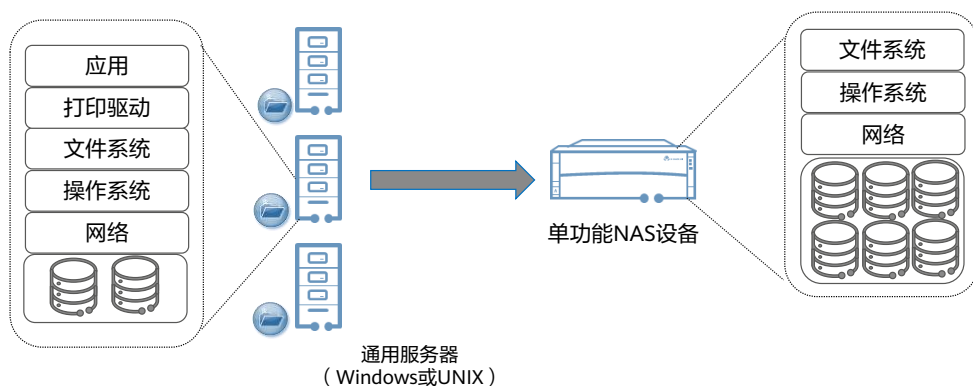


特点:

- 提高效率
- 提高灵活性
- 集中存储
- 简化管理
- 可扩展性
- 高可用性
- 安全 (用户认证授权)

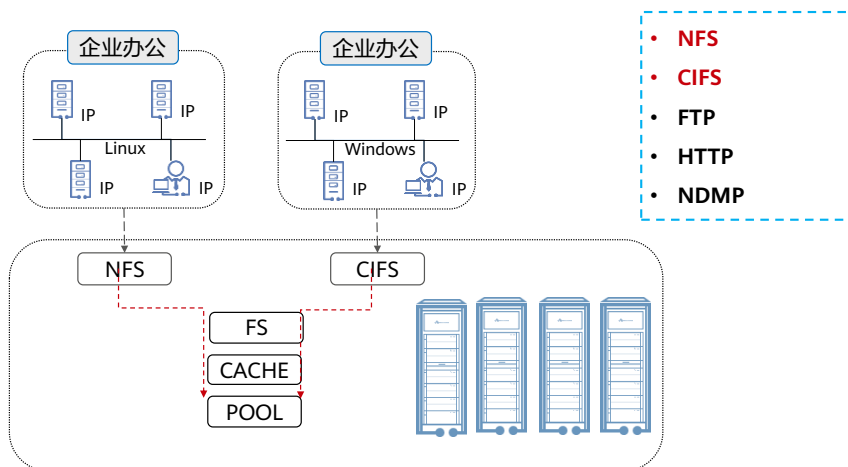
- 企业不仅需要存储庞大的数据，而且需要通过网络把存储的数据共享出去，NAS无疑是一个不错的选择。对于服务器/主机而言，NAS是一个外部设备，通过网络部署具有非常强的灵活性。且NAS是基于文件级的共享，而不是块级，这使得客户端更易于通过网络访问。
 - NAS通过文件级的数据访问和共享提供存储资源，使客户能够以最小的存储管理开销快速直接共享文件；
 - 采用NAS可以不用建立多个文件服务器，是首选的文件共享存储解决方案；
 - NAS还有助于消除用户访问通用服务器时的瓶颈；
 - NAS使用网络 and 文件共享协议进行归档和存储，这些协议包括进行数据传输的TCP/IP和提供远程文件服务的CIFS和NFS。
- UNIX和Microsoft Windows用户能够通过NAS无缝共享相同的数据，通常有NAS和FTP两种数据共享方式。采用NAS共享的时候，UNIX通常使用NFS，Windows使用CIFS。
- NAS实际上就是一个专门优化了的文件服务器

通用服务器与NAS设备



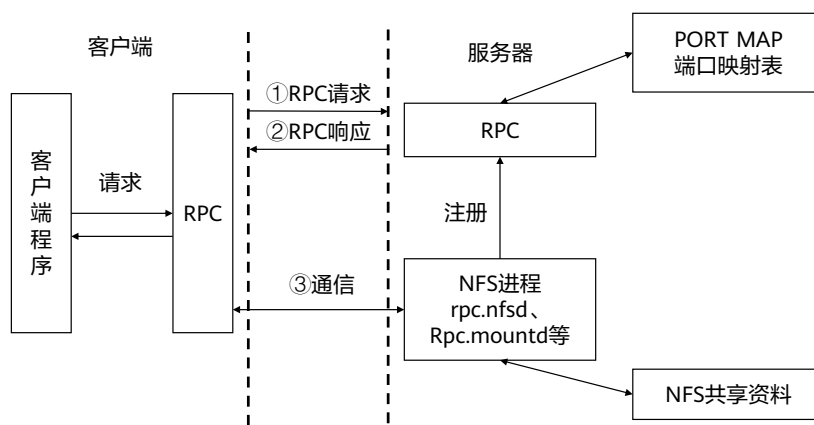
- NAS设备在通用服务器的基础上进行了优化，比如文件服务功能、存储、检索、访问应用程序管客户端文件。
- 如图，一个通用服务器可以用来承载任何应用程序，它运行一个通用的操作系统。NAS不像通用服务器，它专用于文件服务，通过使用开放标准协议对其它操作系统提供文件共享服务。为了提升NAS设备的高可用性，有些NAS供应商还支持NAS集群功能。
- NAS设备包含如下组件：
 - NAS引擎(CPU和内存等)；
 - 一个或多个网络接口卡（NIC），提供网络连接。如千兆以太网卡、万兆以太网卡；
 - 一个优化的操作系统用于NAS功能管理；
 - 协议NFS和CIFS；
 - 采用工业标准存储协议的磁盘资源，如ATA、SCSI、FC。
- NAS环境包括通过使用标准协议的IP网络访问NAS设备的客户端。

NAS常见协议



- NFS: Network File System protocol,传统的UNIX环境文件共享协议。
- CIFS: Common Internet File System protocol,传统的微软环境下的文件共享协议，基于服务器消息块协议。
- CIFS是一个基于网络的共享协议，其对网络传输的可靠性要求很高，所以它通常是使用TCP/IP。NFS用于独立的传输，所以它使用TCP或UDP。
- NFS的一个缺点是客户端必须配备专用软件。CIFS被集成到操作系统中，不需要额外的软件。
- NFS是一个无状态的协议，而CIFS是一个有状态的协议。NFS连接故障后可自动恢复连接，而CIFS不能。CIFS仅发送少许的冗余信息，因此它具有比NFS更高的传输效率。
- FTP: File Transfer Protocol，文件传输协议是TCP/IP 协议组中的协议之一。FTP协议包括两个组成部分，其一为FTP服务器，其二为FTP客户端。其中FTP服务器用来存储文件，用户可以使用FTP客户端通过FTP协议访问位于FTP服务器上的资源。
- HTTP: HyperText Transfer Protocol，是一个用于传输超媒体文档（例如 HTML）的应用层协议。它是为 Web 浏览器与 Web 服务器之间的通信而设计的，但也可以用于其他目的。
- NDMP: Network Data Management Protocol，网络数据管理协议为网络连接存储 (NAS) 的网络备份提供了一个开放式标准。NDMP 可以使数据绕过备份服务器而直接写入磁带存储，提高了 NAS 数据保护的速度和效率。

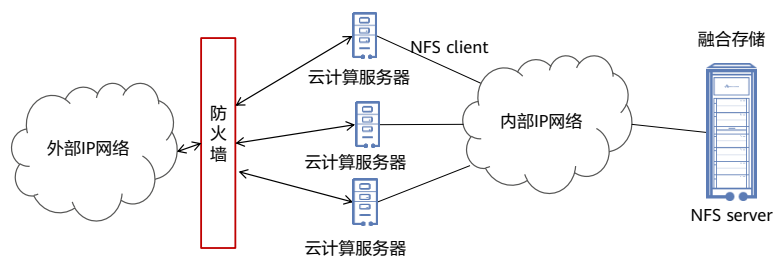
NFS工作原理



- RPC (Remote Procedure Call) —远程过程调用，它是一种通过网络从远程计算机程序上请求服务，而不需要了解底层网络技术的协议。RPC协议假定某些传输协议的存在，如TCP或UDP，为通信程序之间携带信息数据。在OSI网络通信模型中，RPC跨越了传输层和应用层。RPC使得开发包括网络分布式多程序在内的应用程序更加容易。
- RPC采用客户机/服务器模式。请求程序就是一个客户机，而服务提供程序就是一个服务器。首先，客户机调用进程发送一个有进程参数的调用信息到服务进程，然后等待应答信息。在服务器端，进程保持睡眠状态直到调用信息到达为止。当一个调用信息到达，服务器获得进程参数，计算结果，发送答复信息，然后等待下一个调用信息，最后，客户端调用进程接收答复信息，获得进程结果，然后调用执行继续进行。

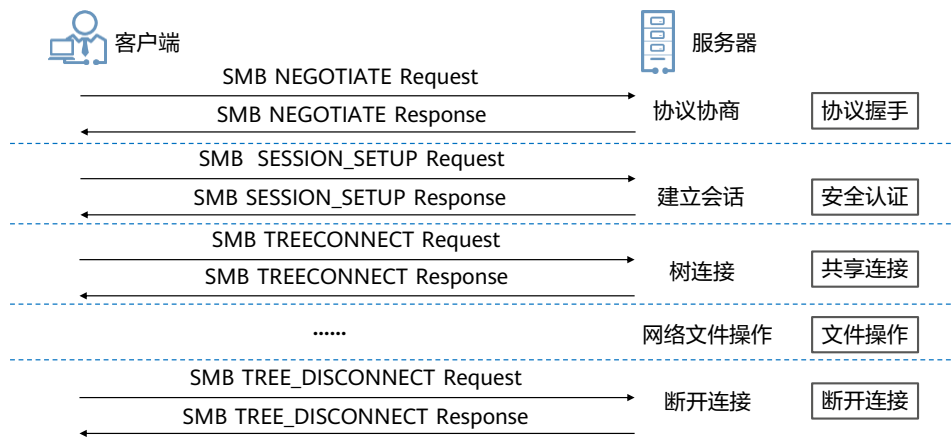
NFS典型应用案例：云计算共享存储

- 云计算使用NFS服务器作为内部共享存储:



- 云虚拟化软件（如：VMware）专门对NFS客户端进行了优化，用来把虚拟机存储空间建立在NFS服务器的共享空间上。
- 这个根据云计算优化过的NFS客户端提供更好的性能和可靠性。

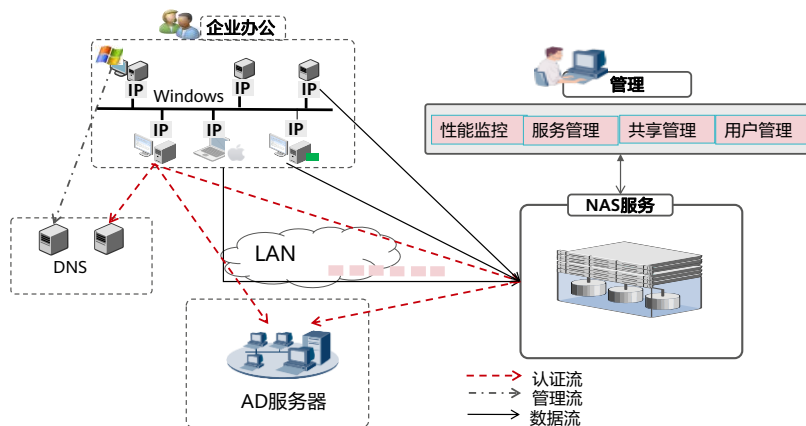
CIFS工作原理



- 认证方式:NTLM，Kerberos 等。
- 网络通信可以over TCP 或其他方式如：RDMA等。
- 网络文件操作经处理后转交给实际文件系统。
- NTLM安全性支持提供者服务。NTLM的意思即NT LanManger，是NT下提供的认证方法之一，使用了64位的加密手段。
- Kerberos 网络认证。Kerberos 是一种网络认证协议，其设计目标是通过密钥系统为客户机/服务器应用程序提供强大的认证服务。

CIFS典型应用案例：文件共享服务

- 文件共享服务场景适用于企业文件服务器、媒资等应用场景。



- DNS: Domain Name Service, 用于实现域名和IP的映射。
- AD: Active Directory, AD域, 提供目录服务。
- CIFS协议主要用于文件共享, 这里介绍两个典型的应用场景:
 - 文件共享服务场景
 - 文件共享服务是最典型的应用场景, 其主要应用于企业文件服务器, 媒资等等, 为用户提供文件共享服务。
 - Hyper-v虚拟机应用场景
 - 微软主推Hyper-v虚拟机, 使用SMB共享虚拟机镜像。该场景下需要依赖SMB3.0的Failover特性(故障透明切换), 该特性可以保证在结点故障情况下, 进行结点切换, 业务不中断, 从而保证了虚拟机运行的可靠性。

目录

1. DAS

2. NAS

3. SAN

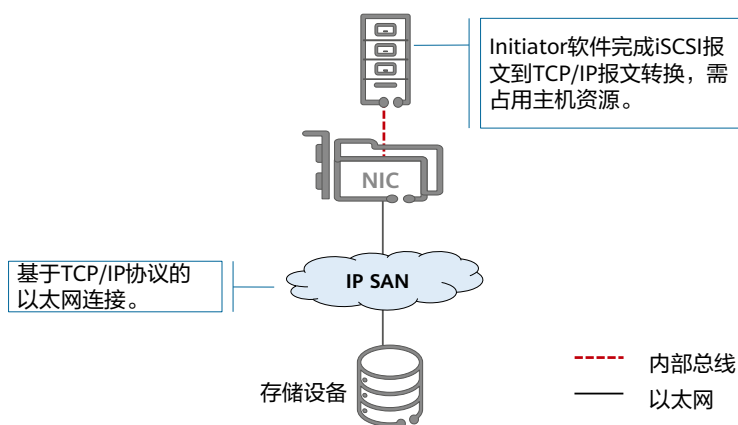
- IP SAN 相关技术

- FC SAN 相关技术

- IP SAN和FC SAN区别

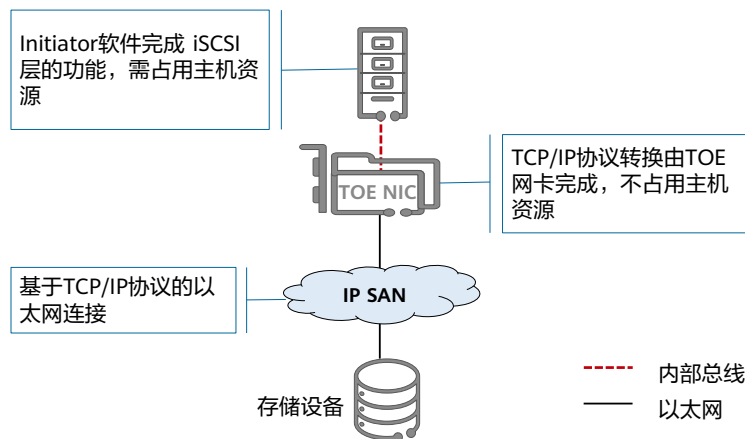
4. 分布式

NIC + initiator软件



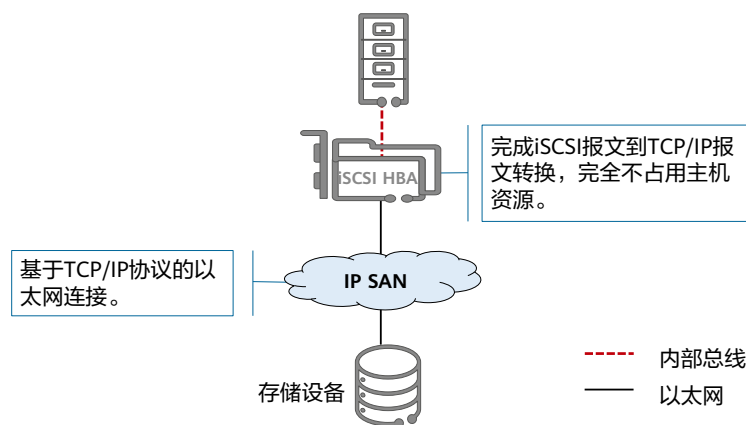
- 服务器、工作站等主机设备使用标准的NIC卡连接到以太网交换机，iSCSI存储也同样连接到以太网交换机，或直接连接到主机的NIC卡。在主机上安装Initiator软件以便将以太网卡虚拟为iSCSI卡，用以接收和发送iSCSI数据包，从而实现主机和iSCSI设备之间的iSCSI协议和TCP/IP协议传输功能。由于采用普通的标准以太网卡和以太网交换机，无需额外配置适配器，因此此种方式硬件成本最低。然而，这种模式在进行iSCSI和TCP/IP数据包转换的时候，需要占用主机资源，增加主机操作开销，降低系统性能。不过在对于I/O和带宽性能要求较低的应用环境中基本能够满足数据访问要求。

TOE NIC + initiator软件



- TOE NIC卡处理TCP/IP协议层的功能，而主机则处理iSCSI协议层的功能。因此，采用TOE网卡可以大幅度提高数据的传输速率。与纯软件的方式相比较，这种方式极大的降低了主机的操作开销，而又不会使网络构建成本过多增加，是一种比较折衷的配置方案。

iSCSI HBA卡

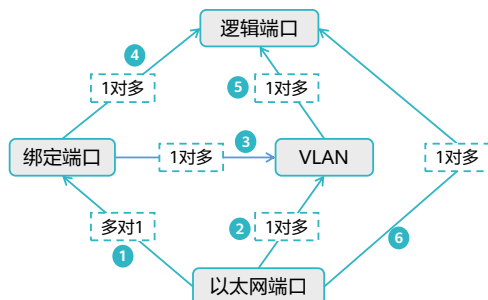


- 在主机上安装iSCSI HBA适配卡，从而实现主机与交换机之间、主机与存储设备之间的高效数据交换。iSCSI协议层和TCP/IP协议栈的功能均由主机总线适配器（HBA）来完成，对主机CPU的资源消耗最少。这种方式提供了最好的数据传输性能，但是系统构建成本也最高。
- iSCSI的通信体系继承了SCSI的部分特性，在iSCSI通信中，具有一个发起I/O请求的启动器设备（Initiator）和响应请求并执行实际I/O操作的目标器设备（Target）。在Initiator和Target之间建立连接后，Target在操作中作为主设备控制整个工作过程。目标包含iSCSI磁盘阵列和iSCSI磁带库。
- iSCSI协议为Initiator和Target定义了一套命名和寻址方法。所有的iSCSI节点都是通过其iSCSI名称被标识的。这种命名方式使得iSCSI名称不会与主机名混淆。
- iSCSI使用iSCSI唯一名称（IQN）来鉴别启动设备和目标设备。地址会随着启动设备和目标设备的移动而改变，但是名字始终是不变的。建立连接时，启动设备发出一个请求，目标设备接收到请求后，确认启动设备发起的请求中所携带的iSCSI Name是否与目标设备绑定的iSCSI Name一致，如果一致，便建立通信连接。每个iSCSI节点只允许有一个iSCSI Name，一个iSCSI Name可以被用来建立一个启动设备到多个目标设备的连接，多个iSCSI Name可以被用来建立一个目标设备到多个启动设备的连接。

逻辑端口

- 逻辑端口是用户基于绑定端口、VLAN的端口或者以太网端口创建的，是承载主机业务的虚拟端口。
- 每个逻辑端口都会被分配一个唯一的IP地址来承载相应的业务。

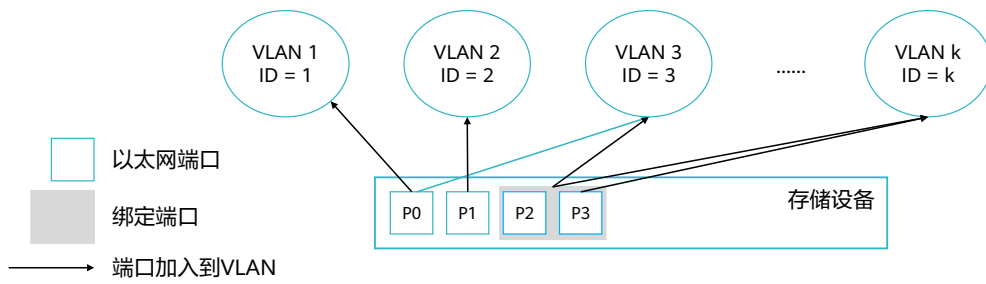
标号	说明
1	表示可以由多个以太网端口绑定在一起形成绑定端口。
2	表示可以将一个以太网端口加入到多个VLAN。
3	表示将一个绑定端口加入到多个VLAN。
4	表示一个绑定端口可以创建多个逻辑端口。
5	表示VLAN的一个端口可以创建多个逻辑端口。
6	表示一个以太网端口可以创建多个逻辑端口。



- 以太网端口：**存储系统接口卡上物理可见的以太网端口。绑定端口、VLAN和逻辑端口的创建都是基于可用的以太网端口的。
- 绑定端口：**为了提高文件系统访问路径的可靠性和增加带宽，可以将同一个控制器下的多个以太网端口绑在一起形成绑定端口。
- VLAN：**存储系统上的以太网端口或者绑定端口资源在逻辑上划分成多个广播域，在VLAN内发送或接收业务数据时，给数据加上VLAN ID，实现VLAN之间的网络隔离和业务隔离，从而提高业务数据的安全性和可靠性。

VLAN划分

- VLAN是将一个物理的LAN在逻辑上划分成多个广播域的通信技术。
- 在存储中，可以将系统中的以太网端口或绑定端口加入到多个独立的VLAN。将不同的业务配置在不同的VLAN中，可确保业务数据的安全性和可靠性。



- VLAN, Virtual Local Area Network, 虚拟局域网。
- LAN, Local Area Network, 局域网。
- 每个VLAN的ID, 可以是1~4094之间的任意数字。

IP漂移

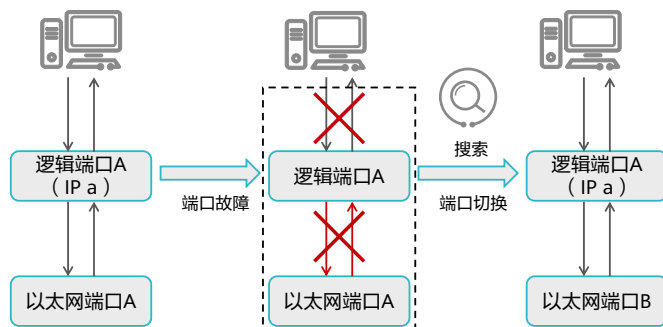
- IP漂移是指逻辑IP地址从故障端口漂移到可用端口上，从而在不中断主机业务的情况下，将故障端口的业务快速切换到其它可用端口上，当故障端口恢复工作后，又可重新接管之前的业务。
- IP漂移的主要功能是在用户无感知的情况下，将故障端口上的业务，快速自动地切换到其它可用端口上，保障业务不中断，提高访问文件系统路径的可靠性。
- IP漂移的实质是当前承载业务端口的切换，这些端口可以是以太网端口、绑定端口或者VLAN的端口。

- 漂移组：存储系统上可用于IP漂移的端口组合。
- 回切模式：简称回切。当故障端口恢复工作后，故障端口重新接管之前的业务模式，有手动切换和自动切换两种。
- IP漂移不只适用于IP SAN，也适用于NAS。

基于以太网端口的IP漂移

- 为了提高访问文件路径的可靠性，可以基于以太网端口创建IP漂移。
- 当创建逻辑端口对应的以太网端口发生故障时：

- 系统搜索到相同类型的其他可用以太网端口；
- 系统删除故障以太网端口上的逻辑端口；
- 系统在可用端口上创建IP相同的逻辑端口来承载业务；
- 业务继续运行。

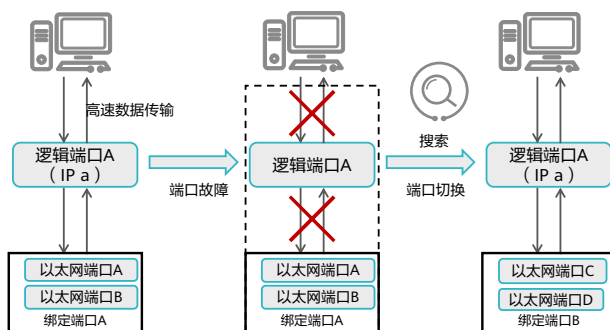


- 主机业务运行在以太网端口A对应的逻辑端口A上，对应的IP地址为a。以太网端口A故障，不能对外提供业务。启用IP漂移后，存储系统会自动搜索可用的以太网端口B，同时删除以太网端口A对应的逻辑端口A的相关配置，在以太网端口B上创建并配置逻辑端口A。这样，主机的业务就在以太网端口B对应的逻辑端口上运行，快速完成故障端口业务切换。应用IP漂移技术完成故障端口的业务切换过程是非常短暂的，主机业务不中断，用户无感知。

基于绑定端口的IP漂移

- 为了提高访问文件路径的可靠性，可以将多个以太网端口绑在一起，形成绑定端口。
- 当创建逻辑端口对应的以太网端口发生故障时：

- 系统搜索到可用端口；
- 系统删除故障端口上创建的逻辑端口；
- 在可用端口上创建相同IP的逻辑端口；
- 将业务切换到可用端口上；
- 当故障端口恢复后，可以重新接管之前的业务。

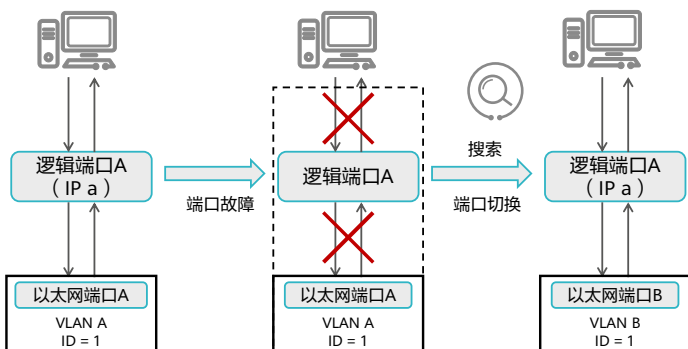


- 当创建绑定端口的某一个以太网端口故障，业务运行端口还是继续在该绑定端口上，只有当所有创建绑定端口的以太网端口都发生故障，才会应用IP漂移技术。
- 将多个以太网端口绑定在一起形成绑定端口A，基于绑定端口A创建的逻辑端口A可以提供高速数据传输业务。由于各种原因造成以太网端口A和以太网端口B都发生故障时，系统会自动搜索其它可用的绑定端口，同时删除逻辑端口A，并将在搜索到的绑定端口B上创建对应的逻辑端口A，业务运行的当前端口就会切换到绑定端口B上。当故障端口恢复工作后，如果启用了IP回切功能，业务运行的当前端口会重新切换到绑定端口A上。由于切换过程非常短暂，这就实现了用户无感知的业务迁移。

基于VLAN的IP漂移

- 当用户需要将不同的业务隔离开来，可以通过创建VLAN来实现。
- 当创建VLAN的以太网端口发生端口故障时：

- 系统搜索到相同类型的可用端口；
- 系统删除故障端口上创建的逻辑端口；
- 系统在可用端口上创建相同的逻辑端口；
- 将故障端口的业务切换到其他可用端口上。



- 使用基于VLAN的IP漂移需要先创建VLAN，每个VLAN都分配唯一的ID，将不同的业务隔离开来。当某个VLAN上的以太网端口故障时，系统会自动搜索VLAN ID相同的可用以太网端口，将业务切换到该可用端口上。当故障端口恢复正常后，又可以重新接管之前的业务。
- VLAN A和VLAN B为创建VLAN时自动生成的VLAN名称，实际名称会与存储系统版本有关。
- 将以太网端口和对应的交换机端口划分为多个VLAN，并分配不同的VLAN ID，将不同的业务隔离开来。在以太网端口A上创建VLAN A，VLAN ID为1，基于VLAN A创建的逻辑端口A可以起到多业务间隔离作用。由于各种原因造成以太网端口A发生故障时，系统会自动搜索到其他可用的VLAN B，搜索到VLAN ID为1的端口，同时删除逻辑端口A，并创建对应逻辑端口A。业务运行的当前端口会切换到VLAN B上。当故障端口恢复工作后，如果启用了IP回切功能，业务运行的当前端口会重新切换到VLAN A上。
- 一个以太网端口也可以属于多个VLAN，当该以太网端口故障时，基于该端口创建的所有VLAN都会故障，需要将业务端口切换到其它可用的VLAN的端口上。由于端口IP地址切换过程非常短暂，这就实现了用户无感知的业务迁移。

目录

1. DAS

2. NAS

3. SAN

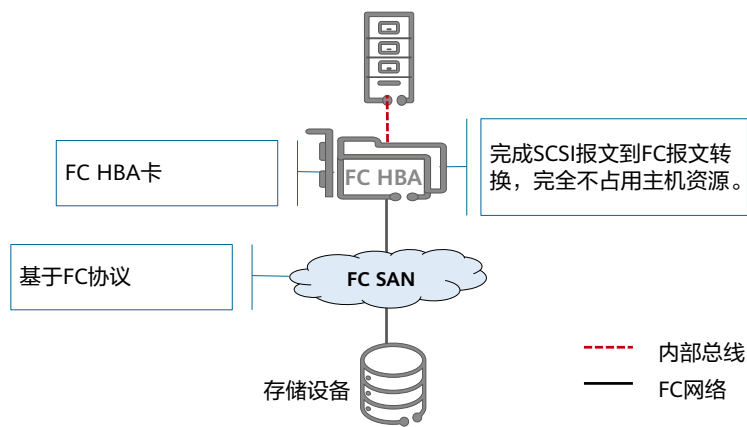
- IP SAN 相关技术

- FC SAN 相关技术

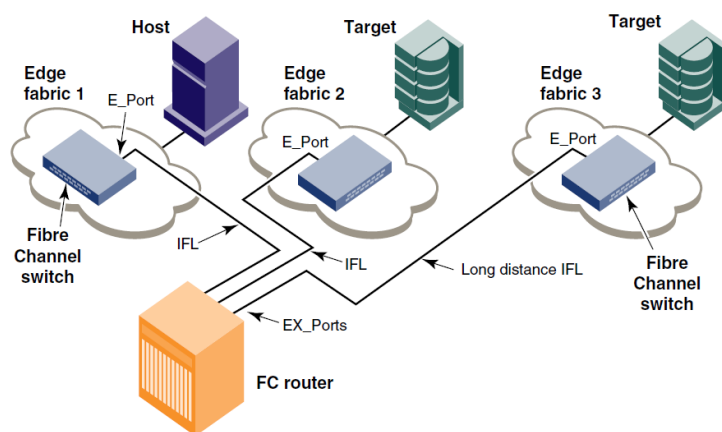
- IP SAN和FC SAN区别

4. 分布式

FC HBA卡

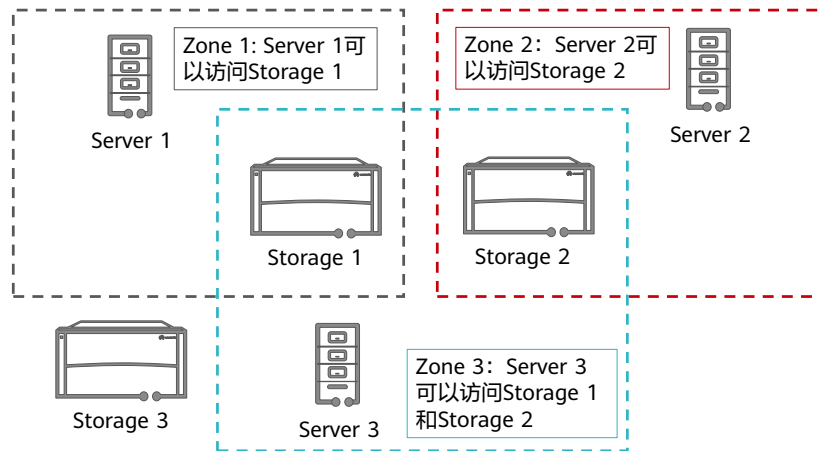


FC组网



- FCR (Fibre Channel Routing) 用于链接不同Fabric网络中的设备。与普通交换机 E_Port级联不同，交换机通过FCR交换机链接后，两Fabric网络并没有融合，还是之前的两个独立的Fabric。两Fabric之间的链接交换机功能类似于路由器。
- FC Router:启用了FC-FC Routing服务的交换机。
- EX_Port: 类似于E_Port类型的端口类型，但不传送Fabric服务或拓扑结构信息。
- Backbone Fabric: 具有启用FC Router服务的交换机的Fabric可看作为Backbone Fabric。
- Edge Fabric: 连接FC Router的Fabric可看做是Edge Fabric。
- IFL (Inter Fabric Link) ， 在E_Port和EX_Port， 或VE_Port和VEX_Port之间的链路，即FC路由链路

Zoning



- Zone是可进行互通的端口或设备的名称构成的集合。
- 在一个zone里的设备只能与同一个zone中的其他设备通信。
- 一个设备可以同时多个zone里。
- 基本分区（ Basic Zones ）：设置基本分区，控制各设备或端口之间的访问权限。
- 流量隔离分区（ Traffic Isolation Zones ）：当存在多条ISL级联链路（多个E_Port）时，用于指定某条ISL链路只负责传送与该链路E_Port在同一TI Zone内的端口的流量。

目录

1. DAS

2. NAS

3. SAN

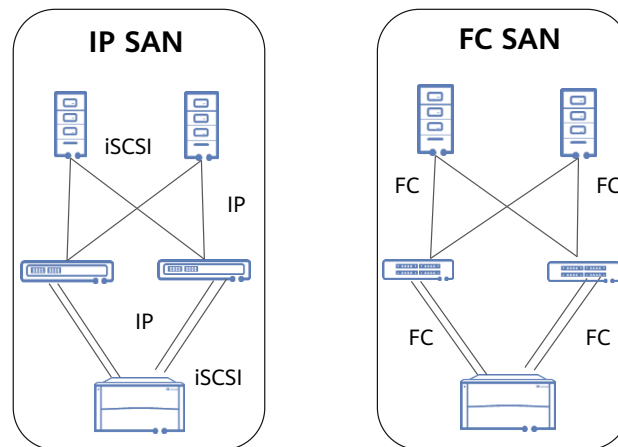
- IP SAN 相关技术

- FC SAN 相关技术

- IP SAN和FC SAN区别

4. 分布式

IP SAN和FC SAN



- FC SAN为解决DAS扩展性差的问题，将存储设备网络化，可以同时连接上百台服务器。
- IP SAN为解决FC-SAN在价格及管理上的诸多问题而产生。
- SAN (Storage Area Networks)
 - 协议: FC/iSCSI，使用这两种协议的SAN架构，分别为FC SAN和IP SAN。
 - 裸设备访问，适宜传统数据库访问。
 - 依赖应用主机提供文件访问。共享访问需要集群软件支持，处理冲突访问开销大，性能较差，难以支持异构环境共享。
 - 性能好，高带宽，低延迟，但是代价高，可扩展性差。

IP SAN与FC SAN比较

描述	IP SAN	FC SAN
网络架构	使用现有IP网络	单独建设光纤网络和HBA卡
传输距离	理论上没有距离限制	受到光纤传输距离的限制
管理、维护	与IP设备一样操作简单	技术和管理较复杂
兼容性	与所有IP网络设备都兼容	兼容性差
成本	与FC SAN相比，购买与维护成本都较低，有更高的投资收益比例	购买（光纤交换机、HBA卡、光纤磁盘阵列等）、维护（培训人员、系统设置与监测等）成本高
容灾	本身可以实现本地和异地容灾，且成本低	容灾的硬件、软件成本高
安全性	较低	较高

- IP SAN只需要很少的硬件配置，并且这些硬件都是广泛使用的，所以，这使得IP SAN的成本比FC SAN低廉得多。大多数的主机都已经配置了合适的网卡和交换机，这些网卡和交换机同样也适合（虽然不完美）于iSCSI协议的传输。高性能的IP SAN，通常需要需要配备专门的iSCSI HBA卡和高端交换机。

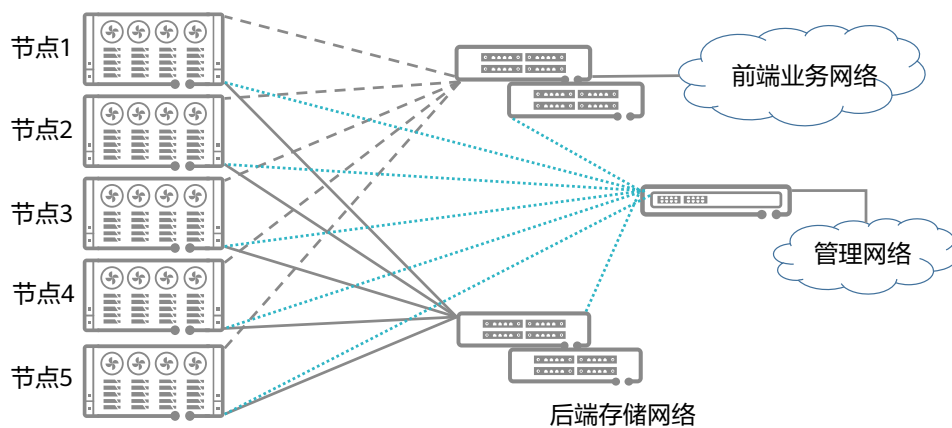
DAS、NAS 、SAN对比

存储系统架构	DAS	NAS	SAN
数据传输协议	SCSI/FC/ATA	TCP/IP	FC
传输对象	数据块	文件	数据块
使用标准文件共享协议	否	是（NFS/CIFS.....）	否
集中式管理	不一定	是	需要管理工具
提高服务器效率	否	是	是
灾难忍受度	低	高	高，专有方案
适合对象	中小企业服务器，捆绑磁盘（JBOD）	中小企业，监控，光电编播	大型企业，数据中心
应用环境	局域网。 文档共享程度低，独立操作平台，服务器梳理少。	局域网。 文档共享程度高，异质格式存储需求高。	光纤通道存储域网。 网络环境复杂，文档共享程度高，异质操作系统平台，服务器数量多。
容量扩充能力	低	中	高

目录

1. DAS
2. NAS
3. SAN
- 4. 分布式**

分布式存储组网

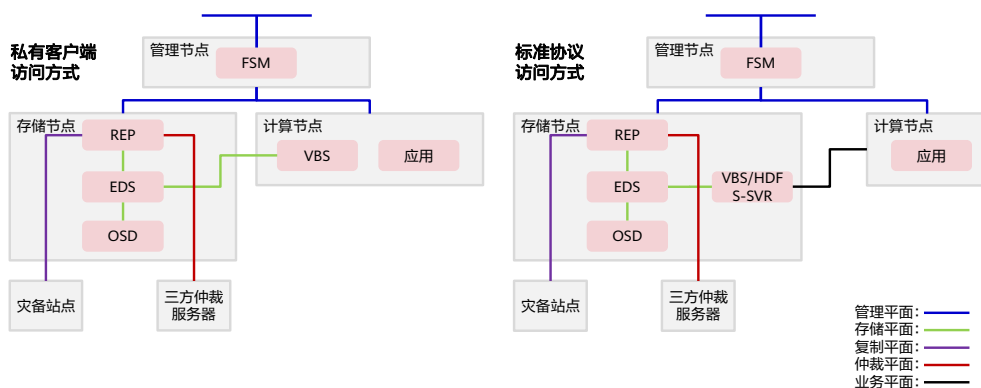


- 分布式存储简单的来说，将通用服务器的本地HDD、SSD等存储介质组织成一个大规模的存储资源池，然后将数据分散存储到多个数据存储服务器上。
- 分布式存储后端网络一般采用10GE，25GE和IB网络；前端网络一般采用GE，10GE，25GE网络。

组网概述

- 前端业务/租户网络：
 - 前端业务/租户网络用于分布式存储与用户网络对接，为租户用户提供租户UI，完成资源申请、使用情况查询等操作，并处理租户客户端或API发送的业务请求。
- 后端存储网络：
 - 后端存储/内部管理网络用于节点间内部互联，提供HA（High Availability）组件如DSS（Data Service SubSystem）的心跳通信，以及各组件之间的内部通信和数据交互。
- 管理网络：
 - 管理网络用于与用户维护网络对接，为系统管理员提供管理UI，完成系统配置、租户管理、资源管理、服务发放等业务操作，以及告警/性能/拓扑等维护操作。同时可以汇聚所有物理节点的Mgmt接口，提供远程设备维护能力，如远程登录设备虚拟KVM、查看温度、电压等硬件运行数据等。

网络平面划分



- 管理平面：与用户管理网络对接的平面，用于进行系统管理和维护；
- BMC平面：用于接入管理节点/存储节点Mgmt接口，提供远程硬件设备管理功能；
- 存储平面：内部平面，用于系统内部所有节点之间的业务数据通信；
- 业务平面：与客户应用对接，通过iSCSI/HDFS等标准协议访问存储；
- 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，只有块服务规划双活功能时才会规划该平面。
- FSM：FusionStorage Manager，华为分布式存储的管理进程，提供告警、监控、日志、配置等操作维护功能，推荐主备部署。
- VBS：Virtual Block Service，VBS通过SCSI或iSCSI接口提供分布式存储接入点服务，使应用服务器能够通过VBS访问分布式存储资源。
- OSD：Object Storage Device，华为分布式存储的组件，承载分布式集群中用户数据的存储设备。
- REP：Replication，数据复制网络。
- EDS：Enterprise Data Service，企业数据服务组件，接收到来自VBS的I/O业务之后，执行具体的I/O操作。

组网规则

- 节点在机柜中优先从底向上部署。
- 在实际项目应用中，需要计算机柜中存储节点、交换机、KVM的总功耗和总重量，并根据客户机房条件综合判断一个机柜中可容纳的节点数量，设计适合具体项目实际情况的部署方案。
- 在典型配置中，基础柜内部的节点通过网线和SFP+线缆连接到本机柜中的交换机，扩展柜内部的节点需要通过网线和光纤连接到基础柜中的交换机。

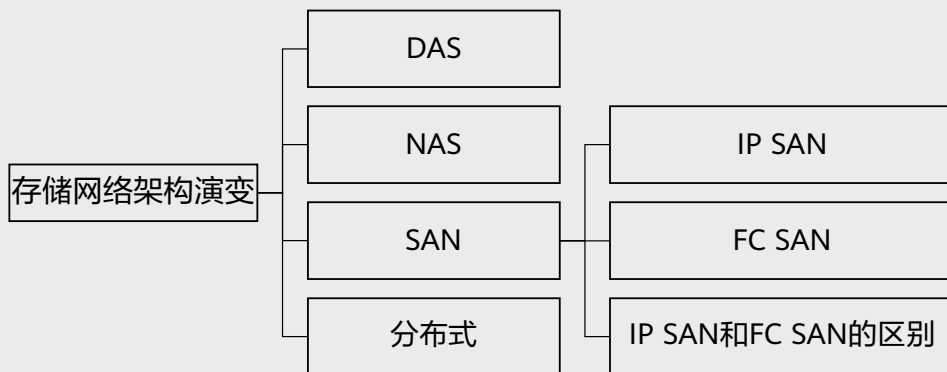
思考题

1. （多选题）分布式存储组网包括以下哪些选项？
 - A. 管理网络
 - B. 前端业务网络
 - C. 前端存储网络
 - D. 后端存储网络
2. （多选题）SAN组网中的常见协议有哪些？
 - A. FC
 - B. iSCSI
 - C. CIFS
 - D. NFS

- 参考答案：

- ABD
- AB

本章总结



学习推荐

- 华为官方网站
 - 企业业务: <http://enterprise.huawei.com/cn/>
 - 技术支持: <http://support.huawei.com/enterprise/>
 - 在线学习: <http://learning.huawei.com/cn/>
- 热门工具
 - HedEx Lite
 - 网络资料工具中心
 - 信息查询助手

- 热门工具:
 - HedEx Lite: 华为产品文档管理工具, 支持浏览、搜索、更新和管理产品资料
 - eStor: 图形化存储仿真平台, 该平台通过对真实华为OceanStor全闪存仿真模拟, 帮助广大ICT从业者和用户快速熟悉华为存储系列产品, 了解并掌握相关产品的操作和配置。
 - 网络资料工具中心: 汇集网络产品的资料工具, 是您进行投标支撑、网络规划、项目交付和升级维护的好助手。
 - 信息查询助手: 提供华为产品命令和告警信息查询。

缩略语表

DAS: Direct Attached Storage, 直接连接存储, 是指将外置存储设备通过连接电缆, 直接连接到一台计算机上。

RAID: Redundant Array of Independent Disks, 独立磁盘冗余数组, 一种把多块独立的硬盘 (物理硬盘) 按不同方式组合起来形成一个硬盘组 (逻辑硬盘), 从而提供比单个硬盘更高的存储性能和提供数据冗余的技术。

ROW: Redirect on write, 是指写时重定向技术, 是实现文件系统快照的核心技术。当源文件系统收到写数据请求需要修改源文件系统中的数据时, 存储系统在存储池中为新写入的数据指定新的存储位置, 并将被修改数据块的指针指向新的存储位置。

VLAN: Virtual Local Area Network, 虚拟局域网 一种在交换局域网的基础上, 采用网络管理软件构建的可跨越不同网段、不同网络的端到端的逻辑网络。

HBA: Host Bus Adapter, 主机总线适配器。

KVM: Keyboard, Video, and Mouse, 原义指键盘、显示器和鼠标。通过KVM远程查看终端主机的屏幕, 或者通过本地的鼠标和键盘远程控制终端主机, 方便管理员通过远程的方式帮终端用户解决终端主机出现的问题。

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©2022 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.



华为智能存储产品介绍



前言

- 本章主要介绍华为智能存储系列产品，包括华为全闪存存储、混合闪存存储、分布式存储、超融合、以及专用备份存储产品的特点、产品定位以及典型应用场景。

目标

- 学完本课程后，您将能够：
 - 熟悉存储系列的产品特点
 - 熟悉存储系列的产品定位
 - 熟悉产品典型应用场景

目录

1. 全景介绍
2. 全闪存存储
3. 混合闪存存储
4. 分布式存储
5. 超融合
6. 专用备份存储

华为提供全面丰富的数据存储产品系列

解决方案



智能数据湖



视频云



HPDA



企业应用加速



双活



备份

数据管理

智能运维



eService

数据管理引擎



DME Storage

设备管理



DeviceManager

存储产品

全闪存存储



混合闪存存储



分布式存储



超融合



专用备份存储



目录

1. 全景介绍
- 2. 全闪存存储**
3. 混合闪存存储
4. 分布式存储
5. 超融合
6. 专用备份存储

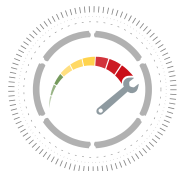
全闪存产品系列



- 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。
- OceanStor Pacific全系列产品分为性能型和容量型两类，性能型和容量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；
- 容量型整机以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前端接口，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。

产品特点

永快一步



业界最高性能
业界最低时延

2100万 IOPS
0.05ms 时延
NAS性能领先业界30%

永远在线



SmartMatrix全互联架构
关键业务持续在线

容忍控制器8坏7
SAN&NAS一体化AA双活

永智高效



SAN&NAS极致融合
全生命周期智能管理

智能运维
端云智能协同

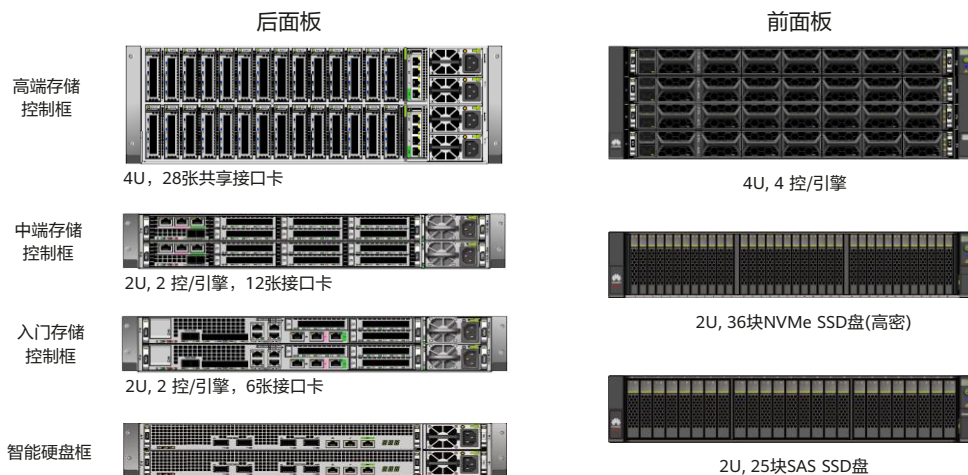


华为OceanStor Dorado
全闪存存储

- 华为全新一代全闪存存储主要有三个方面的优势：
 - 一是性能好，能够提供基于SPC-1模型下2100万IOPS，0.05ms的极致时延，同时NAS性能领先业界30%
 - 二是可靠性好，可以容忍可靠性最高的控制器8坏7，且做到了业界唯一的SAN&NAS一体化真AA双活，保障业务持续在线。
 - 三是高效，智能运维，端云智能协同，让您使用存储资源更为智能、简洁。
- 采用SmartMatrix3.0全互联均衡架构、该架构采用了高速、矩阵式全互联无源背板。可以连接多个控制器节点，同时接口模块采用全共享方式接入背板，允许主机从任意端口接入，直达任意控制器进行处理。SmartMatrix全互联均衡架构允许控制节点之间紧密协调、简化软件模型，实现Active-Active细粒度均衡、高效低时延和协同作业。
- FlashLink®技术，具备高IOPS（Input/Output Operations Per Second）并发能力，同时保持稳定的低时延。FlashLink®技术的核心是通过一系列针对闪存介质的算法优化技术，实现了控制器板载CPU和SSD板载专用CPU的联动，保障了SSD算法在不同CPU之间的协同，实现系统的高性能和高可靠。
- SSD(固态硬盘)是一种主要以闪存（NAND Flash）作为永久性存储器的存储设备，相较于传统HDD硬盘具有：速度快、能耗低、低延迟、体积小、重量轻、抗震等特点。

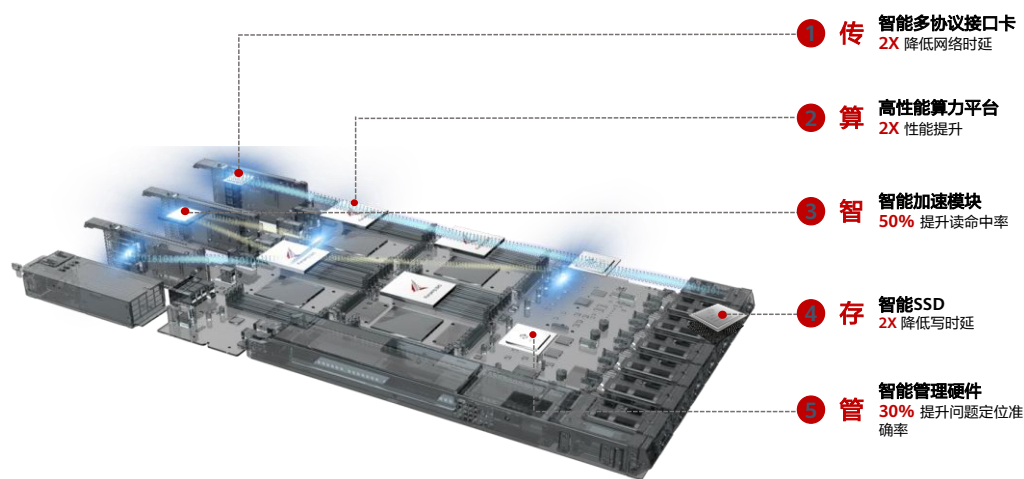
- 高性能
 - 全SSD硬盘，具备高IOPS、低时延；
 - 支持支持FlashLink，如：智能众核、高效RAID、冷热数据分流、低时延保证等技术。
- 高可靠性
 - 部件失效保护，双冗余设计和A/A工作模式；支持SmartMatrix3.0存储控制器引擎采用前后端全互联架构，高效低时延和协同作业；
 - 部件采用双冗余设计、支持断电保护、保险箱盘等技术；
 - 高级数据保护技术：如HyperSnap、HyperReplication、HyperClone、HyperMetro等；
 - RAID 2.0+底层虚拟化。
- 高可用性
 - 支持部件在线更换，如：控制器、电源、接口模块、硬盘等；
 - 支持硬盘漫游，可自动识别更改槽位后的硬盘，并自动恢复其原有业务；
 - 支持对第三方存储系统中的存储资源进行集中管理。

产品形态示例



- OceanStor Dorado V6采用新一代创新硬件平台，在控制器和硬盘框上都采用了高密设计：
 - 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，最大支持12张接口卡，通过100Gb RDMA网络实现引擎内镜像，通过25Gb RDMA网络实现跨引擎互联；
 - 低端存储2U盘控一体方式部署，最大支持6张接口卡，通过40Gb RDMA网络实现引擎内镜像，通过25Gb RDMA网络实现跨引擎互联；
- 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。

创新硬件，智能加持，关键路径加速永快一步

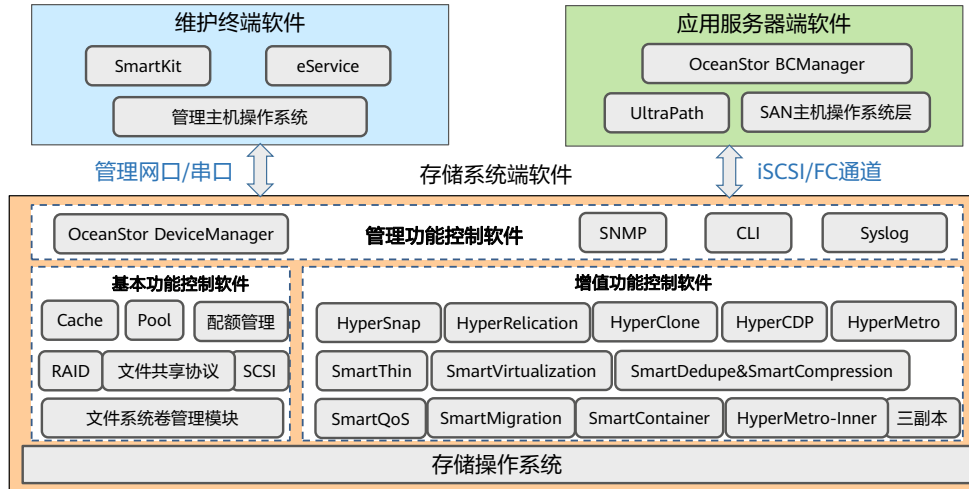


11 Huawei Confidential

HUAWEI

- 为了能给客户带来极致的性能体验，华为在影响性能的所有关键IO路径上，均实现了部件的自研：
- 第一个是智能多协议接口卡，这个卡的作用是协议卸载和解析的工作，通过自研，使其实现了多协议卸载和解析，从而释放了控制器CPU的负载，提升了整体的性能。
- 第二，在算力方面，以往都是采用INTEL CPU来实现，这一代采用了华为鲲鹏高性能算力平台，该处理器的能效比同档位的INTEL要高25%，加上在CPU数量和核数上的绝对优势，大大提高了存储系统的整体性能表现，比如在高端上一个控制器就部署了4颗芯片，4U4控就部署了16颗芯片，192核。
- 第三，在智能上，自研了一颗智能加速模块，它是以接口卡的形态插在控制器槽位上，配合着内部的智能缓存算法来进行使用。当上层的IO数据流过来时，智能加速模块会自动捕捉业务流，学习业务流的规律，分析规律，从而推测出后续可能的行为，提前预取缓存资源，提升了读缓存的命中率，特别是在批量读的场景非常有效。
- 第四，智能SSD控制器，它是SSD的大脑，通过自研，能够加速盘内读和写以及降低时延，单盘数据读和写要比友商快15%。
- 第五，智能管理硬件，它内置了华为存储十年来的故障基因库，可以在芯片内定位出原因，并给出相应的解决方案，自愈的措施，问题定位准确率从60%定位到93%
- 通过5部分自研硬件，从传、算、智、存、管等在关键路径上的加速，华为全闪存可以做到比友商永快一步。

软件架构

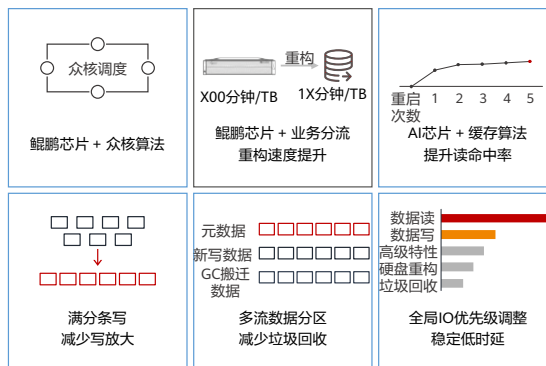


- 存储系统端软件采用OceanStor OS专用操作系统，实现硬件管理和支撑存储业务软件件的运行。
 - 存储系统通过基本功能控制软件实现基础的数据存储和读写功能。
 - 存储系统通过增值功能控制软件实现各种备份、容灾和性能调优等高级功能。
 - 存储系统通过管理功能控制软件实现对存储系统的管理功能。
- 维护终端软件用于系统配置与维护。用户可以通过维护终端的SmartKit、eService等软件对存储系统进行配置和维护。
- 应用服务器端软件包括OceanStor BCManager和UltraPath。

“芯”智能



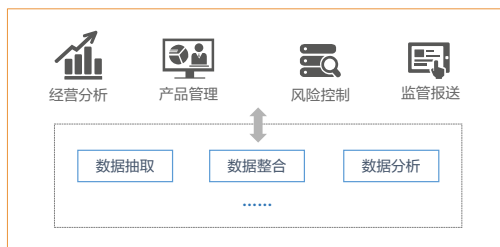
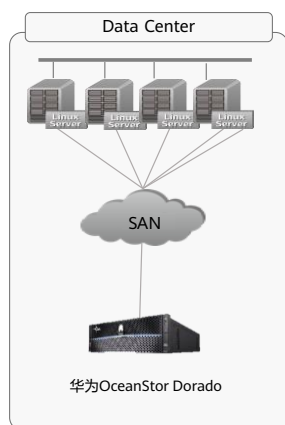
FLASHLINK® 智能算法



- FlashLink®技术主要包含以下关键技术：
- 智能众核技术
 - 存储系统采用华为自研CPU，在同一个控制器内实现了业界存储产品最多的CPU数和CPU核数。通过智能众核技术，实现了存储设备性能随着CPU数量和核数线性增加。
- 高效RAID技术
 - 存储系统采用ROW满分条写入的设计，ROW满分条写对所有数据都采用新写模式，不需要传统RAID写流程所需的数据读和校验修改写而产生RAID写惩罚，有效降低了写入过程阵列控制器的CPU开销与对SSD盘的读写压力。相比传统的RAID覆盖写（Write In Place）的方式，ROW满分条写方式使得各种RAID级别都能实现高性能。
- 冷热数据分流
 - 存储系统识别系统中的冷热数据，通过SSD盘和控制器的配合，提升垃圾回收的性能，减少盘内SSD颗粒擦写次数，延长SSD盘的使用寿命。
- 低时延保证技术
 - 存储系统采用更快的协议，采用最新一代自研SSD盘，并对系统的I/O处理流程进行优化，使得I/O时延保持在低位水准。

- 智能盘框
 - 存储系统采用最新一代自研智能硬盘框，硬盘框内有CPU和内存资源，能够对盘故障重构等任务进行卸载，减轻控制器的负载，实现在盘故障场景下，重构对业务性能无影响。
- 高效时间点技术
 - 存储系统使用高效的分布式时间点技术实现数据保护特性，用户的读写I/O均携带了对应的时间点信息，可以快速定位到元数据信息，因此对性能开销较小。
- 全局磨损均衡与全局反磨损均衡
 - 全局磨损均衡：SSD盘在使用过程中，盘与盘之间磨损不均匀，部分SSD盘会提前损坏，会造成存储系统过早的出现SSD盘失效，维保成本就会上升。存储系统通过全局磨损均衡，使所有SSD盘的磨损率保持均衡，避免SSD盘提前失效。
 - 全局反磨损均衡：存储系统通过全局反磨损均衡，在SSD盘磨损度达到一定程度后，通过优先磨损少量的盘，让SSD盘一块一块地单独损坏，避免多盘同时失效，从而大幅提高存储系统的可靠性。

典型应用场景 - 关键业务加速



- 常见问题：

- 随着移动互联高速发展，海量的客户数据、交易数据爆炸式增长，如何挖掘产生数据真正价值，则必须要进行高效的数据采集、分析、整合提取，推动以数据为中心的战略执行。原有IT系统压力越来越大，如银行、大型企业的账单、库存等系统，数据处理、数据仓库整合，时间长达数小时，导致经营分析、业务查询等业务无法及时获取或查询速度缓慢，严重影响效率。

- 解决方案：

- 采用华为高性能全闪存方案很好的解决此类难题，采用高端全闪存承载多种核心应用（如：交易系统数据库等业务），处理时间可降低一半以上，响应时延也有极大的提升，业务效率成倍提升。

目录

1. 全景介绍
2. 全闪存存储
- 3. 混合闪存存储**
4. 分布式存储
5. 超融合
6. 专用备份存储

混合闪存存储系列



- 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。
- OceanStor Pacific全系列产品分为性能型和容量型两类，性能型和容量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；
- 容量型整机以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前端接口，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。

新一代OceanStor混合闪存



全面升级，极致高效

- SmartMatrix AA全负载均衡架构，最大化性能释放
- 热数据自适应布局，性能提升100%



多样性功能负载，面向未来

- 适应多种应用生态，多样云化路线支持
- 多种安全保护手段，守护企业价值数据



更简单，更经济，更智能

- 无需SAS HDD节省建设成本，跨代设备互通利旧
- 智能、自动化运维管理，大幅降低运维成本

- 融合存储：
 - SAN和NAS存储技术的融合；
 - 协议支持iSCSI、FC、NFS、CIFS、FTP等存储网络协议；
- 高性能：
 - 高性能处理器、高速大容量缓存和多种高速接口模块，提供更加卓越的存储性能；
 - 支持SSD硬盘加速，大幅提升存储性能；
- 高扩展性：
 - 支持多种硬盘类型；
 - 支持多种接口模块；
 - 支持Scale-out等技术；
- 高可靠性：
 - SmartMatrix全互联架构及部件全冗余设计，以及A/A工作模式，提供更高的可靠性，RAID 2.0+等技术；
 - 提供多种数据保护技术，如：掉电数据保护、数据预拷贝、保险箱盘、硬盘坏道修复等；

- 高可用性：
 - 提供多种高级数据保护技术，如：快照、LUN拷贝、远程复制、克隆、卷镜像和双活等，支持NDMP协议。
- 智能、高效：
 - SmartTier、SmartQoS、SmartThin等多种控制管理功能，提供精细化控制管理等操作；
 - DeviceManager支持GUI操作管理界面；
 - eService提供自助式智能运维；

产品形态示例

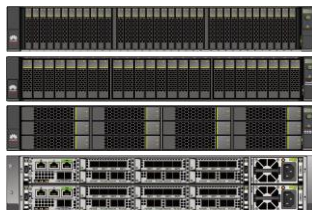
2U存储 如：华为新一代OceanStor混合闪存 5310



4U存储 如：华为新一代OceanStor混合闪存 6810



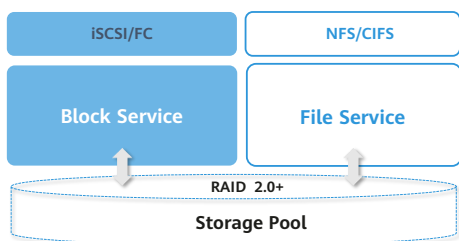
2U存储 如：华为新一代OceanStor混合闪存 5510/5610



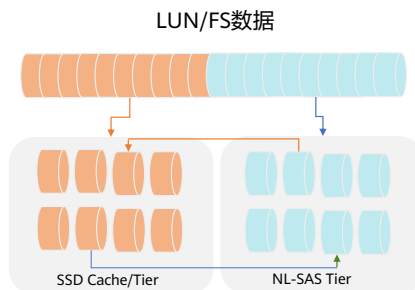
4U存储 如：华为新一代OceanStor混合闪存 18510/18810



SAN&NAS融合



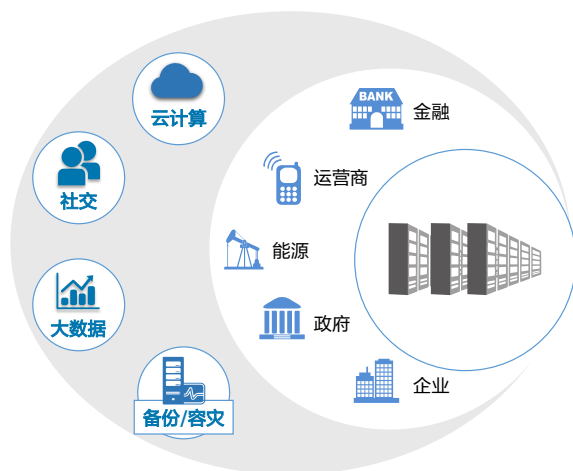
华为新一代OceanStor混合闪存系列SAN&NAS资源融合



SAN/NAS冷热数据智能分级流动

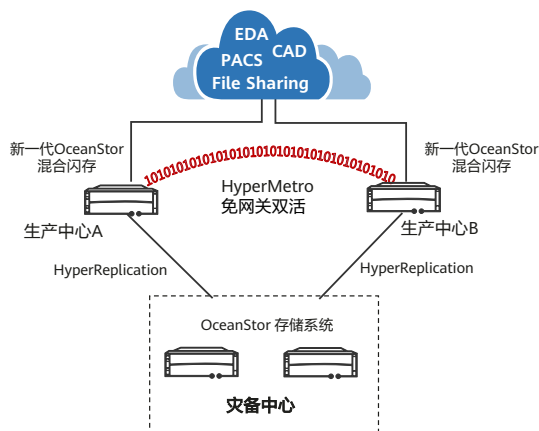
- 存储SAN和NAS存储技术的融合：一套存储设备可同时支持SAN服务和NAS服务，共享存储设备资源，主机可以通过任一控制器的前端端口访问任意的LUN或文件系统，在整个数据的生命周期中，会逐渐从热数据变为冷数据，如果变冷的数据长期占用Cache或者SSD这类快速的设备，不仅会浪费宝贵的资源，还会影响存储系统长期的性能表现。存储可以通过智能分级技术来做到后台灵活调配数据存储介质。
- 智能分级技术需要部署在有不同介质类型的设备上，数据被实时监控，长期不被访问的数据会被标记为冷数据，逐渐从高性能介质转移到慢速介质，保证设备不会越用越慢；当冷数据被激活后也可以快速移动到快速介质上，让整个系统性能保持平稳，不会因为数据量的增加而变得越来越慢。
- 迁移策略可以手工可以自动。

多业务场景支持



- 华为OceanStor混合闪存存储系统集SAN和NAS于一体，支持多种存储协议；广泛适用于多种通用场景，包括但不限于政府、金融、运营商、制造、备份/容灾等行业。

应用场景：数据中心双活

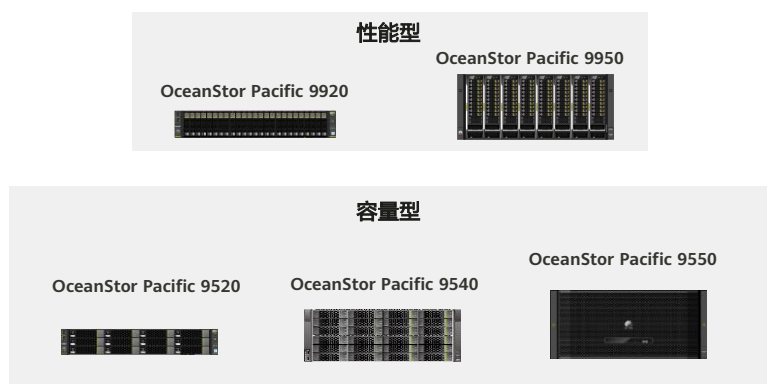


- 实现控制器负载均衡；
- 关键业务RPO=0，RTO≈0；
- 融合SAN与NAS，一台设备同时支持部署SAN和NAS双活服务，单控制器故障支持本地切换等；
- 为客户建设业务不间断运行的解决方案；
- 双活解决方案可广泛用于医疗、金融、社保等行业。

目录

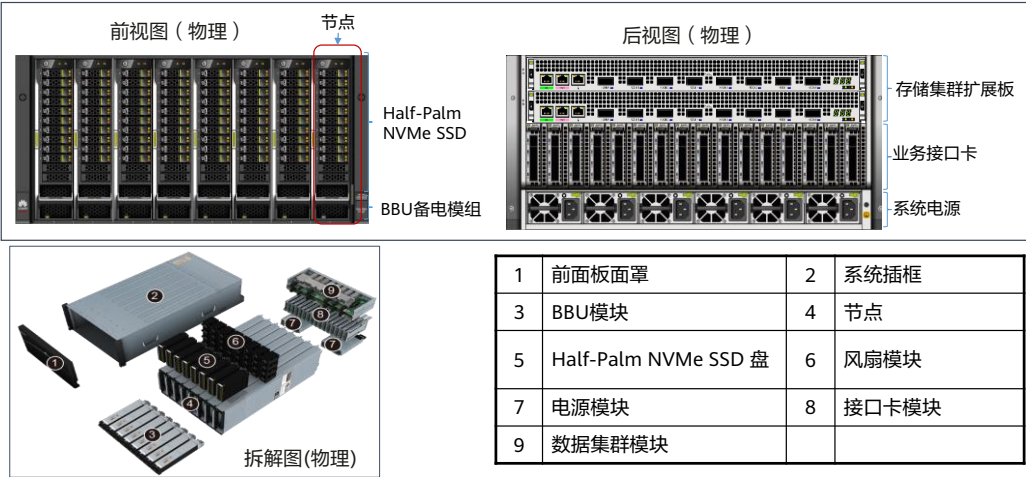
1. 全景介绍
2. 全闪存存储
3. 混合闪存存储
- 4. 分布式存储**
5. 超融合
6. 专用备份存储

分布式存储系列



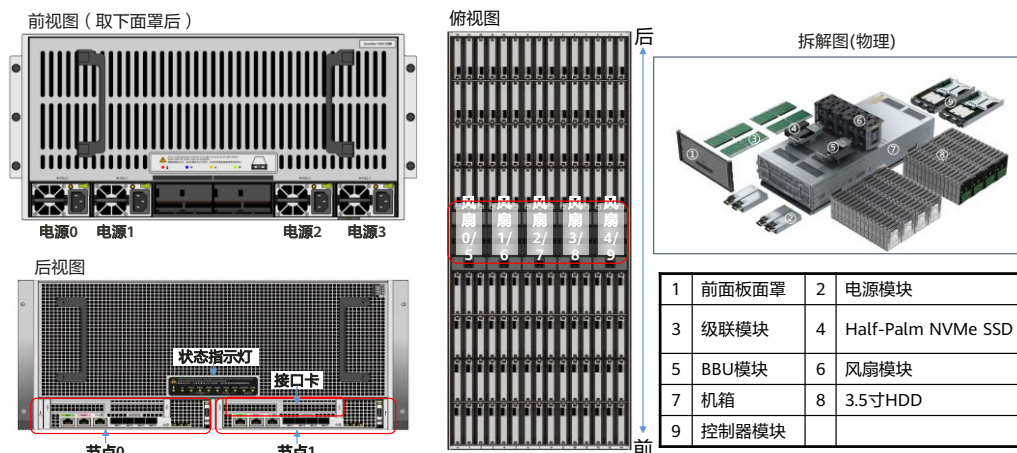
- 在二十世纪第一个十年，随着互联网数据爆发，基于分布式架构的数据存储成为了应对海量数据增长的一种有效手段。
- 在第二个十年，随着企业拉开了数字化转型序幕，数据快速增长，海量数据存储在企业市场迎来了飞速发展。
- OceanStor Pacific全系列产品分为性能型和容量型两类，性能型和容量型可采用分级存储的方式满足客户在不同场景的应用诉求。
- 性能型整机以OceanStor Pacific 9950为典型代表，提供客户极致的海量存储全闪存性能，单机达到160GB/s带宽性能；
- 容量型整机以OceanStor Pacific 9550为典型代表，单整机提供120盘位，配套当前主流的14T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前端接口，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



- 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。

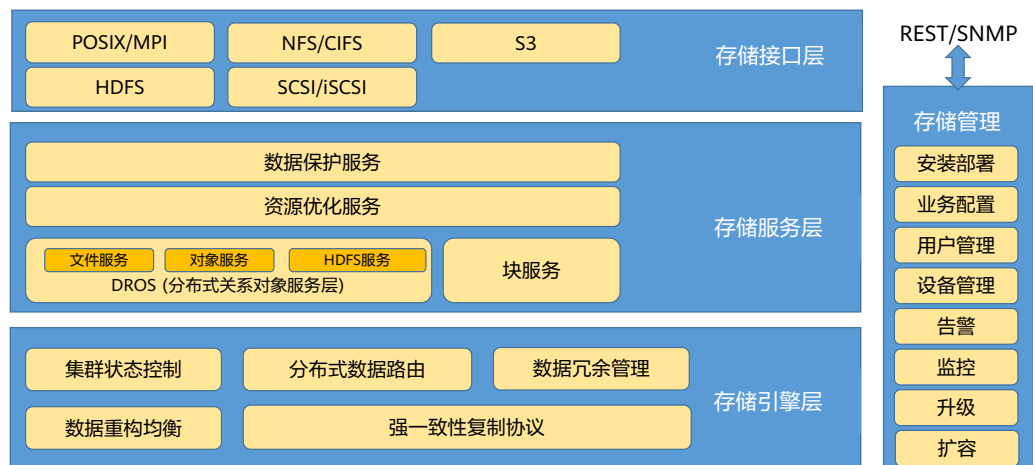
OceanStor Pacific 9550产品物理形态



- 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。
- 最右力是物理硬盘区的更细致划分，整个硬盘区划分为8个锯齿状的子硬盘区域，每个子硬盘区域以7+8的两行排布分布，并通过级联板连接。
- 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

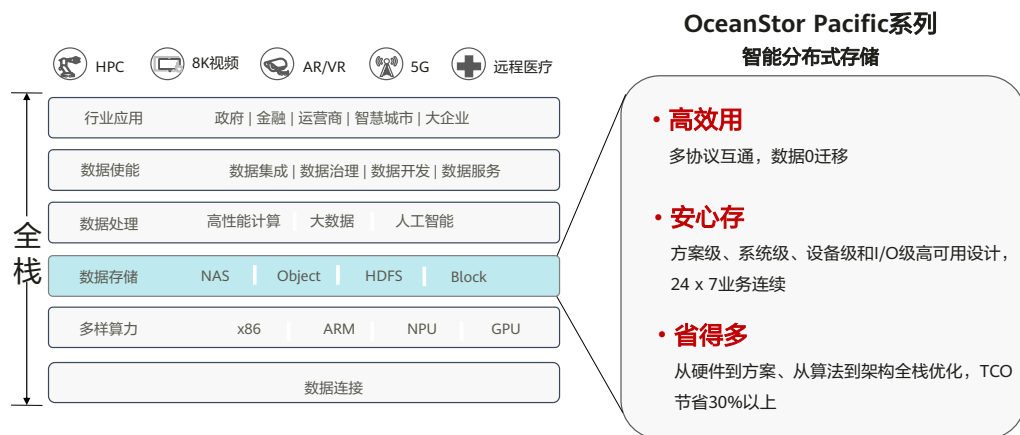
整个风扇坏了，就得立即更换。

软件总体架构



- NFS/CIFS: 文件标准协议场景的共享协议，支持NFS/CIFS共享协议，提供文件系统共享功能。在网络环境中提供各种类型操作系统下的文件共享。
- POSIX/MPI: 在文件DPC（Distributed Parallel Client）场景，通过兼容标准MPI语义和POSIX语义，提供并行接口技术和智能数据缓存算法，使上层应用更智能的访问存储空间。
- S3: 负责处理Amazon S3和NFS协议消息，处理对象存储服务业务逻辑等。
- HDFS: 对外提供HDFS标准接口。
- SCSI/iSCSI: 通过SCSI标准驱动在本地映射卷，或与多路径软件配合通过iSCSI协议接口在应用服务器映射卷，向操作系统、数据库提供卷设备。
- 数据冗余管理: 负责Erasure Coding的计算，保证数据的高可靠性。
- 分布式数据路由: 负责将数据和元数据按规则均匀分布在各存储节点。
- 集群状态控制: 实现对分布式集群的状态控制。
- 强一致性复制协议: 块服务中，负责双活Pair间数据的一致性。
- 数据重构均衡: 负责数据的重建、均衡等。

产品特点



- 华为OceanStor Pacific系列打造面向不同行业应用场景的专属硬件，匹配用户的多样化需求，实现硬件领先；提供块、对象、大数据及文件四类存储协议，通过一套架构满足多样性数据的存储要求；面向不同行业不同业务场景，实现产品与解决方案领先，打造更高效、更可靠和更经济的存储产品，帮助客户实现：海量数据存得下、用得好。

FlashLink众核技术

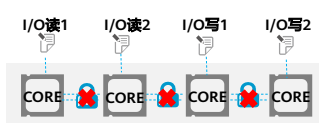
CPU分组算法



I/O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



I/O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。

I/O智能调度



I/O优先级保障，降低业务影响

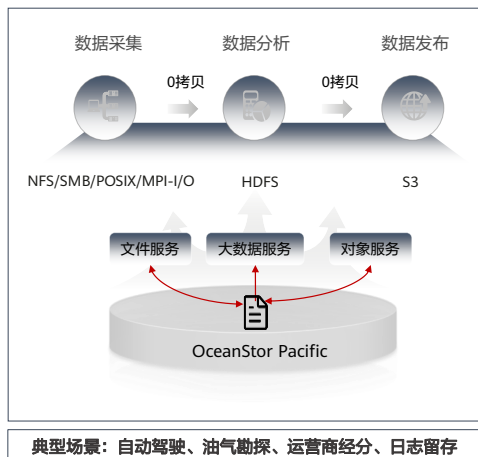
数据读写I/O永远第一优先级响应，其他I/O等数据读写I/O完成后重新启动，保障数据读写I/O时延最低。

稳定低时延，关键业务极速响应

- 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。我们知道，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。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。

- 数据从CPU进入缓存。缓存对存储而言是非常重要的，缓存算法的优劣是决定存储性能的关键要素之一，缓存算法具备三大亮点：
 - 1、摒弃传统存储的二叉树算法，采用闪存专用的哈希表算法。众所周知，二叉树算法的空间占用少，但是查找速度慢；哈希表算法查找速度快，空间占用高，采用多级哈希即满足空间占用同时满足高效查找速度
 - 2、元数据和数据缓存资源分区提升元数据命中率，读数据和写数据缓存分区避免读写缓存资源互抢；
 - 3、哈希索引支持冷热分区机制，提升查找速度

SmartInterworking多协议互通



提升分析效率

多种协议共享一份数据，0迁移高效访问

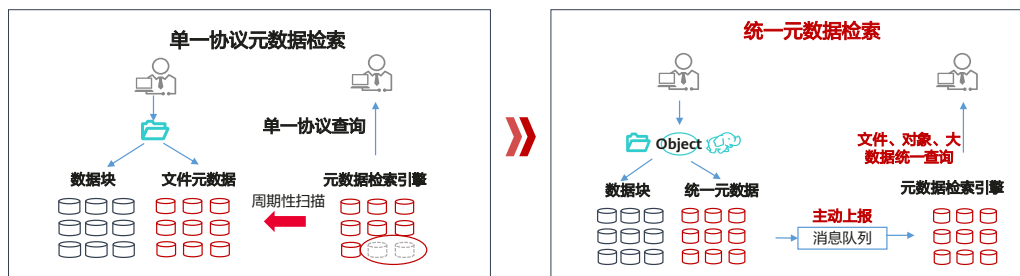
节省存储空间

避免数据重复存储，节省空间开销

规格	D厂商P产品	D厂商E产品	H厂商H产品	OceanStor Pacific
原生语义	√	×	×	√
免网关	√	×	√	√
语义完整度	中	低	低	高
性能影响	中	大	中	小
跨协议实时访问	√	√	×	√
多协议共享特性	×	×	×	√

- 针对自动驾驶等HPC、大数据分析场景数据分析链路长，数据量大，拷贝周期长的问题，OceanStor Pacific系列推出了多协议互通（SmartInterworking）特性，实现了文件、对象、大数据多协议融合互通、语义和性能无损的能力，一份数据不需要格式转换，即可实现多种语义的访问，避免了数据拷贝的过程，大幅度提升数据分析效率，同时消除了传统方式生产系统和数据分析、发布系统分别存储的数据冗余，进一步节省了空间；基于文件、对象、大数据协议互通能力，一份数据可同时被多个环节的业务访问，无需数据迁移，效率更高，真正发挥数据湖的独特价值。在同类型产品功能中，DELL EMC的产品能做到非结构化协议互通，但是有语义损失，存在部分原生功能缺失；而Inspur和H3C均是基于开源Ceph二次开发的产品，这类产品常是在某种存储协议上叠加网关的方案支持多种协议的互通，由于多了一层逻辑转换，所以会带来较大的性能损失。

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



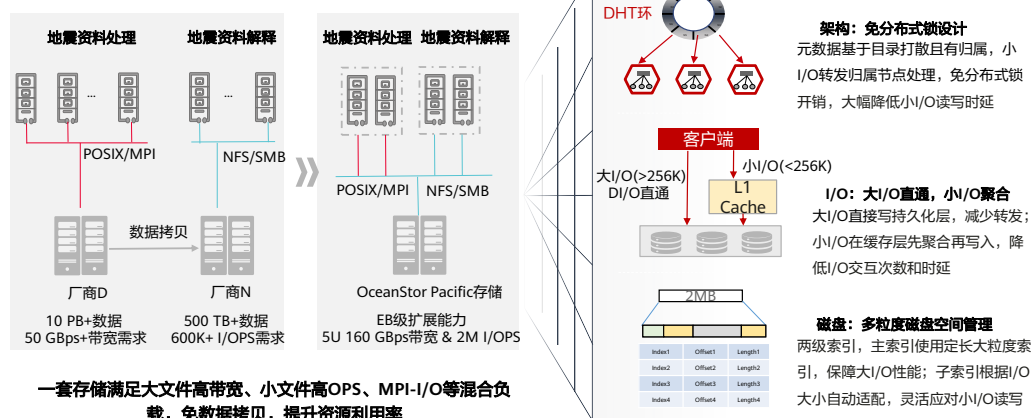
秒级返回结果

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

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I/OPS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

- 每PB最大性能

DPC分布式并行客户端



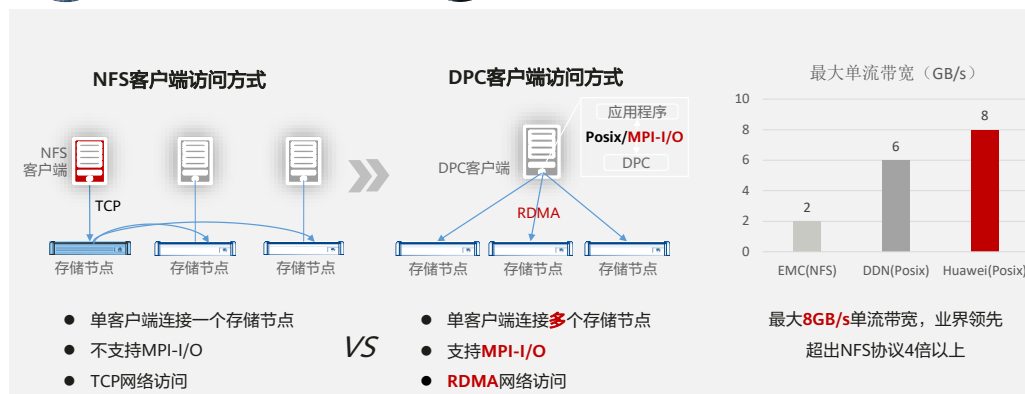
MPI-I/O实现气象
数据分析应用提速



单流高性能
保障卫星数据接收

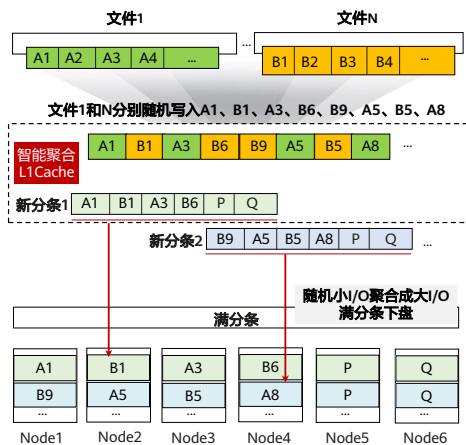


单客户端高性能释放
超算胖客户端潜能



- 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最佳的访问性能;
- 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。

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

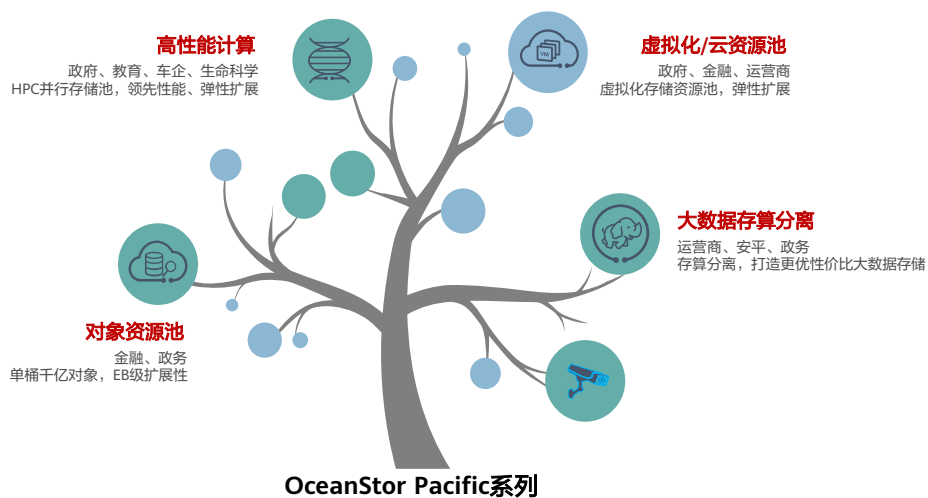
小I/O智能聚合

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，避免小写，发挥HDD极限性能

四大典型场景化解决方案



- OceanStor Pacific系列当前提供四大水平场景方案：高性能计算存储、虚拟化/云资源池存储、对象资源池及大数据分析场景的大数据存算分离存储方案。

目录

1. 全景介绍
2. 全闪存存储
3. 混合闪存存储
4. 分布式存储
- 5. 超融合**
6. 专用备份存储

超融合系列

虚拟化/VDI



FusionCube 1000H

数据库



FusionCube 1000D

企业分支



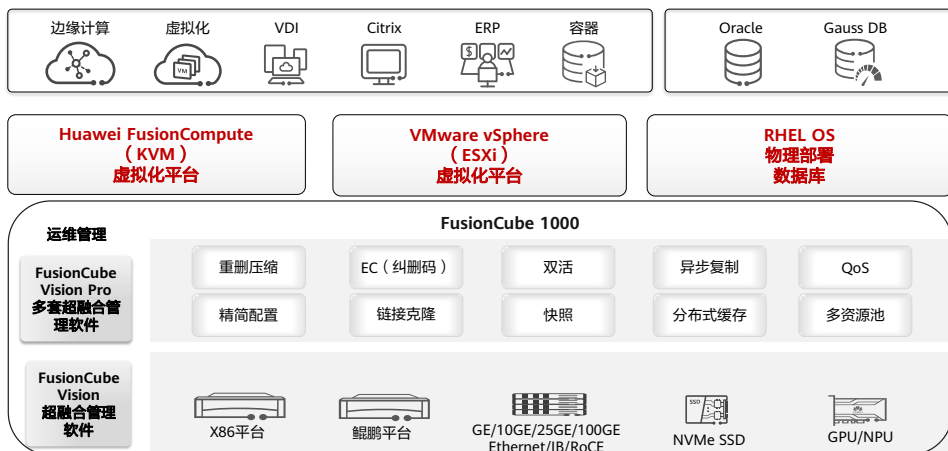
FusionCube 1000C

- 华为FusionCube是基于超融合架构的IT基础设施平台。FusionCube不再是单独的计算、网络、存储设备，而是一体化设备，实现了计算、存储和网络资源的融合，无需额外配置存储、网络等资源。融合计算、存储为一体，并预集成分布式存储引擎、虚拟化和云管理软件，资源可按需调配、线性扩展。主要应用于数据库、桌面云、容器、虚拟化等多种混合负载的数据中心场景场景。
- 华为FusionCube 1000是一体化设计整柜交付的边缘IT基础设施解决方案。主要用于边缘分支机构以及垂直行业边缘应用场景，比如加油站、园区、煤矿、电网等。可部署在办公室，并通过FusionCube Center实现远程集中管理。该方案结合华为产品与FCV(FusionCube Center Vision)平台，实现一站式整柜交付、一站式业务部署、一站式运维管理、一站式故障处理，大幅缩短分支站点部署周期、降低运维成本。
- 超融合基础架构（Hyper-Converged Infrastructure，或简称“HCI”）是指在同一套单元设备中不仅仅具备计算、网络、存储和服务器虚拟化等资源和技术，而且还包括备份软件、快照技术、重复数据删除、在线数据压缩等元素，而多套单元设备可以通过网络聚合起来，实现模块化的无缝横向扩展，形成统一的资源池。

FusionCube主要场景及解决方案概览



超融合数据中心解决方案逻辑架构



关键优势

融合高效



虚拟化-容器-数据库-桌面云与存储 4类融合

2节点起配，降低起配门槛
EC+重删压缩，容量利用率提升3倍

4类服务，融合高效

极致可靠



设备级-数据中心内-数据中心间 10级保护

亚健康检测、端到端DIF，数据极致可靠
真AA双活，99.9999%可靠性，业务持续在线

10级保护，极致可靠

云化易用



超融合-多站点-资源池-服务化 4级管理

开箱即用超融合，一键交付
统一管理平台，极简运维

4级管理，云化易用

企业分支：简单存、统一管、云协同

简单存

快速部署 | 即插即用 | 一键配置



存储 | 计算 | AI | 网络 | 安全

50% 交付周期缩短

- 全栈设计，一站式交付
- 即插即用，一键配置

统一管

统一管理 | 集中监控



设备 | 虚拟机 | 容器 | 应用

1套 软件统管

- 应用、虚拟机、容器、服务器、防火墙、交换机、网关、存储、UPS和传感器
- 20000+站点可视化远程管理

云协同

算法协同 | 数据协同



推理 | 训练

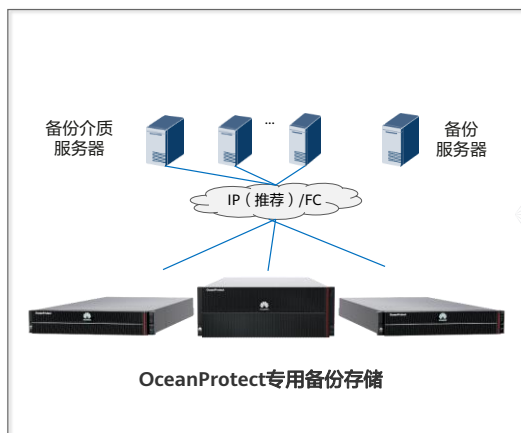
智能自治

- 中心训练，近端推理
- 云端训练，近端推理
- 数据集中备份，快速恢复

目录

1. 全景介绍
2. 全闪存存储
3. 混合闪存存储
4. 分布式存储
5. 超融合
- 6. 专用备份存储**

专用备份存储系列



备份快，恢复快

端到端备份加速

即时恢复



用量省

备份数据预处理和变长重删

压缩/压紧



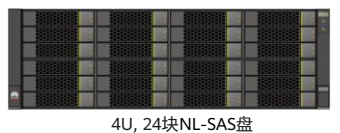
高可靠

数据一致性校验

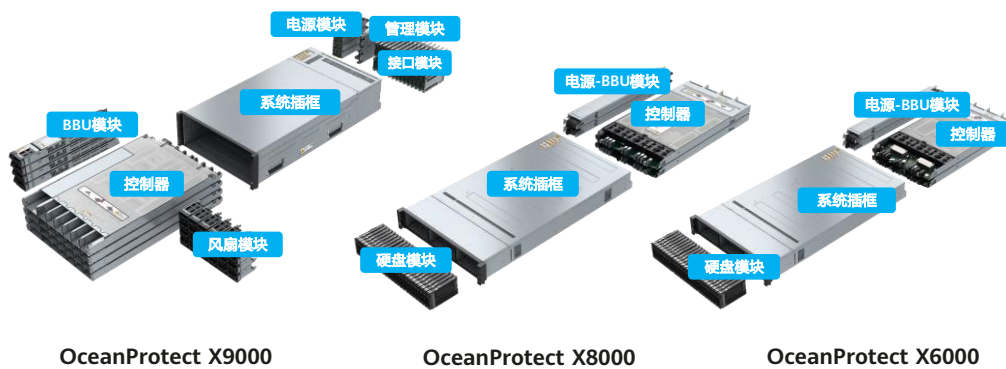
系统级可靠

主动预见式运维

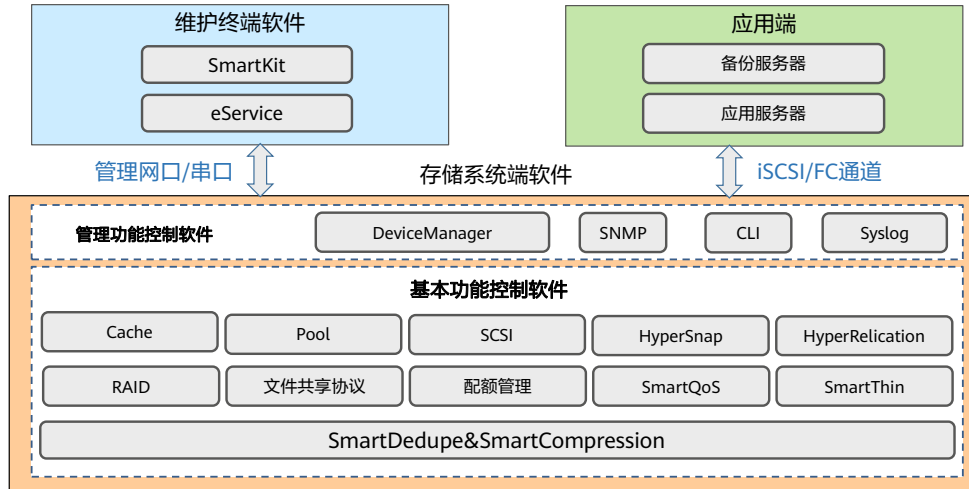
产品形态示例



硬件架构

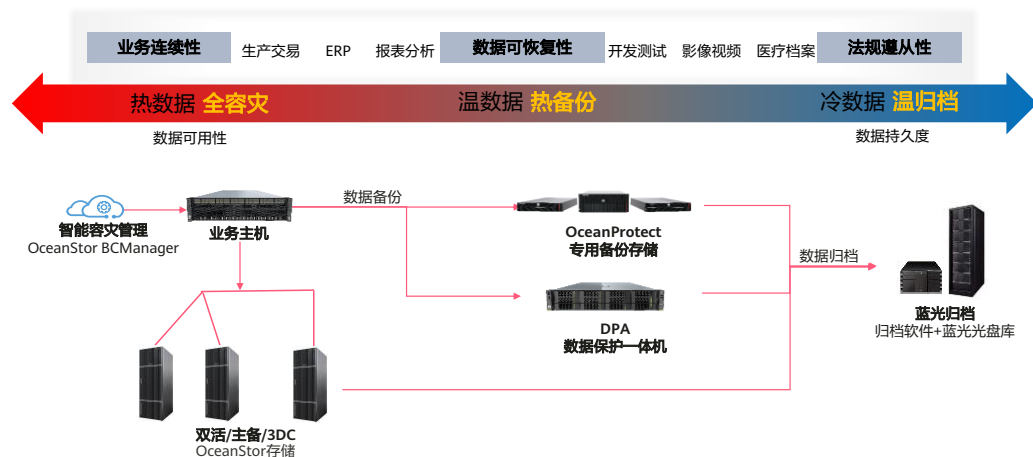


软件架构



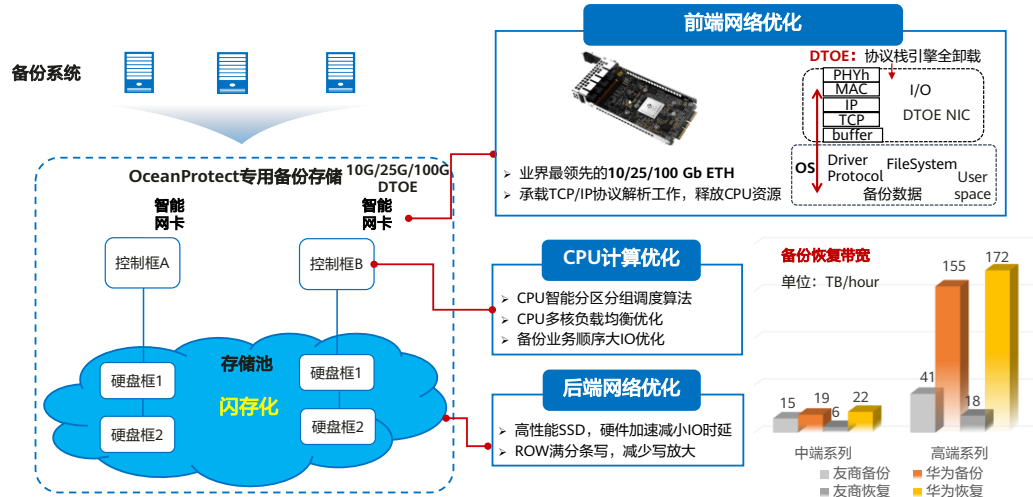
- OceanProtect专用备份存储专用备份存储提供的软件包括存储系统端软件、维护终端软件。软件相互配合，从而智能、高效、经济地实现各种备份业务。
- 存储系统端软件采用专用操作系统，实现硬件管理和支撑存储业务软件的运行。
- 维护终端软件用于系统配置与维护。用户可以通过维护终端的SmartKit、eService等软件对系统进行配置和维护。
- 在SAN网络下，应用服务器端软件可与存储系统通信，从而使双方能够配合执行某些操作。应用服务器端包括备份服务器（通过备份软件备份数据到OceanProtect专用备份存储）、应用服务器（应用服务器通过自带的备份工具备份数据到OceanProtect专用备份存储）。

全场景数据保护，为智能世界保驾护航

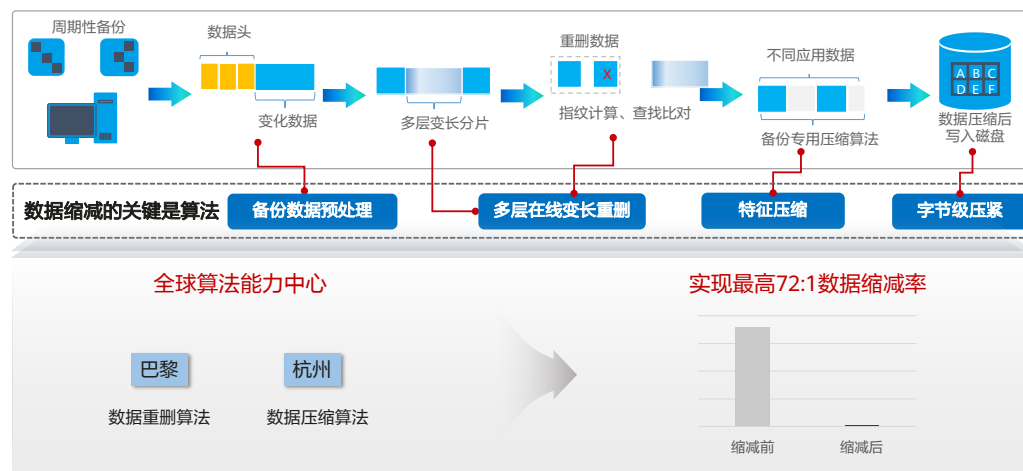


- 华为对存储产品的投入，不仅聚焦于盒子，而是致力于为客户打造端到端完善的产品和解决方案。
- 华为OceanProtect针对快速增长的多样性业务数据，围绕数据全生命周期实现热数据全容灾、温数据热备份、冷数据温归档，和全场景智能融合，支撑客户业务不中断、数据不丢失、信息长期留存。
- 在容灾场景下，华为存储提供完善的容灾特性，包括主备，双活，3DC等方案，并且通过OceanStor BCManager实现统一、智能、简化的容灾调度和运维管理。
- 在备份场景下，华为提供集中备份和一体化备份两种方式，满足客户不同场景需求。
- 在归档场景下，华为提供数据归档至本地蓝光光盘库的方案。

端到端备份加速，全面提升系统性能

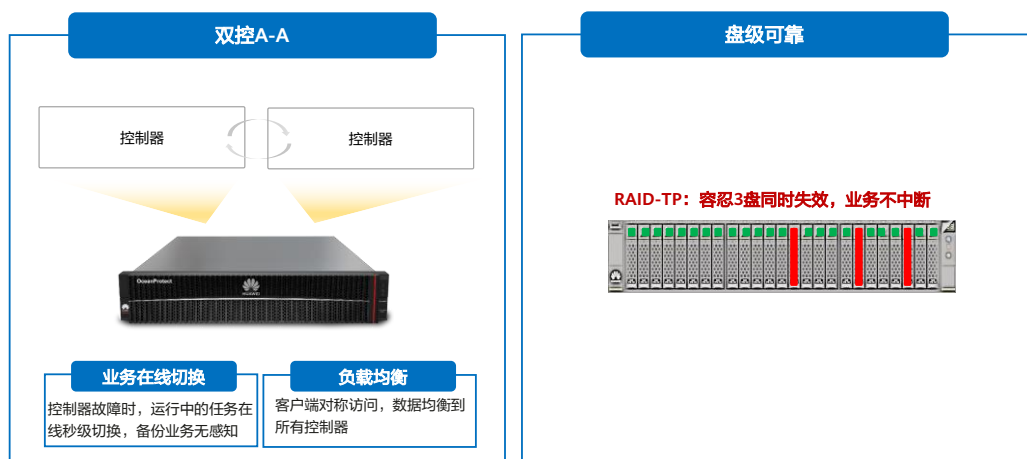


全数据流重删和压缩技术，极致数据缩减



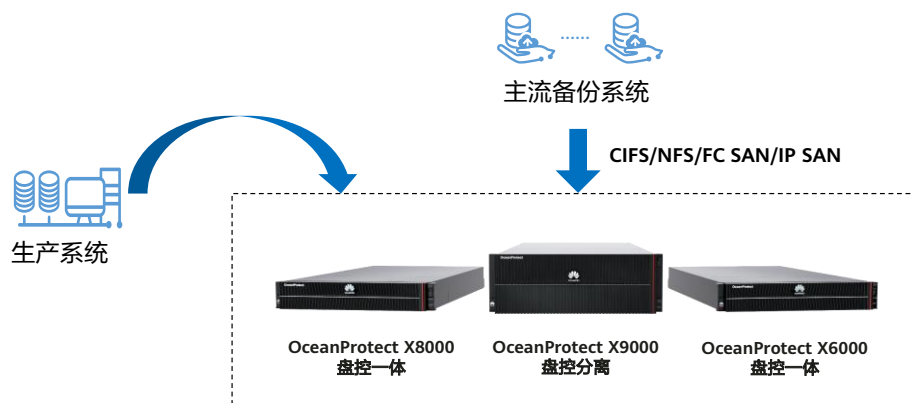
- 重删率72:1的计算模型：
 - 虚拟化场景--200VM
 - 28天全量备份
 - 每天4%修改量
 - 每天0.125%新数据

A-A架构和RAID-TP 技术，保障业务和数据的系统级可靠



- 擦写次数是SSD盘片寿命的主要因素，华为在硬盘上采用了SSD全局磨损均衡&全局反磨损均衡的专利技术，来提升SSD寿命：
 - 1、SSD全局磨损均衡指的是通过华为RAID 2.0+使数据按照指纹全局打散，数据均衡分布在所有SSD，尽可能均衡所有SSD磨损程度，增强SSD可靠性
 - 2、全局反磨损均衡的技术则是多块盘接近磨损极限时，选定某块SSD，加大单块SSD业务量，避免多盘集体失效
- 通过SSD全局磨损均衡&全局反磨损均衡的技术，华为存储能够降低多块硬盘同事故障的风险，有效延长SSD寿命，提升系统可靠性

应用场景：标准NAS/SAN协议兼容备份软件生态



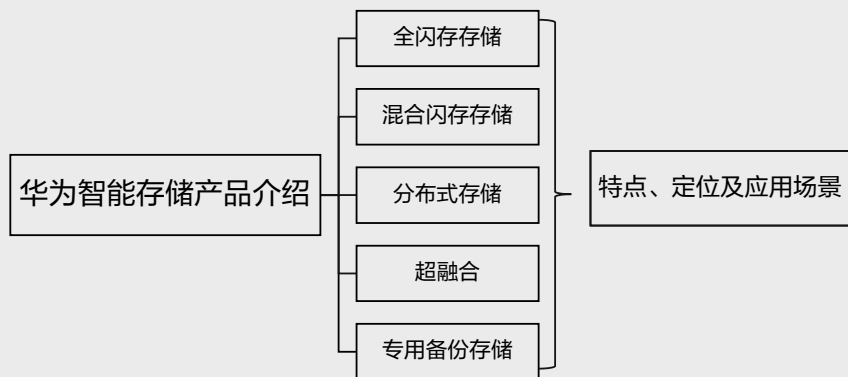
思考题

1. （单选题）以下关于华为OceanStor全闪存存储描述错误的选项是哪个？
 - A. 支持SSD硬盘和NVME硬盘
 - B. 支持SAS接口硬盘
 - C. 支持HDD硬盘
 - D. 支持Palm SSD硬盘
2. （多选题）华为分布式存储支持以下哪些存储服务？
 - A. 块存储
 - B. 文件存储
 - C. 对象存储
 - D. HDFS存储
 - E. 链式存储

- 参考答案：

- C
- ABCD

本章总结



学习推荐

- 华为官方网站
 - 企业业务: <http://enterprise.huawei.com/cn/>
 - 技术支持: <http://support.huawei.com/enterprise/>
 - 在线学习: <http://learning.huawei.com/cn/>
- 热门工具
 - HedEx Lite
 - 网络资料工具中心
 - 信息查询助手

- 热门工具:
 - HedEx Lite: 华为产品文档管理工具, 支持浏览、搜索、更新和管理产品资料
 - eStor: 图形化存储仿真平台, 该平台通过对真实华为OceanStor全闪存仿真模拟, 帮助广大ICT从业者和用户快速熟悉华为存储系列产品, 了解并掌握相关产品的操作和配置。
 - 网络资料工具中心: 汇集网络产品的资料工具, 是您进行投标支撑、网络规划、项目交付和升级维护的好助手。
 - 信息查询助手: 提供华为产品命令和告警信息查询。

缩略语表

DME: Data Management Engine, 数据管理引擎。

HPDA: High Performance Data, 高性能数据分析。

RAID: Redundant Array of Independent Disks, 独立磁盘冗余数组, 一种把多块独立的硬盘(物理硬盘)按不同方式组合起来形成一个硬盘组(逻辑硬盘), 从而提供比单个硬盘更高的存储性能和提供数据冗余的技术。

ROW: Redirect on write, 是指写时重定向技术, 是实现文件系统快照的核心技术。当源文件系统收到写数据请求需要修改源文件系统中的数据时, 存储系统在存储池中为新写入的数据指定新的存储位置, 并将被修改数据块的指针指向新的存储位置。

OLTP: On-Line Transaction Processing, 联机事务处理。

OLAP: On-Line Analytical Processing, 为联机分析处理。

NDMP: Network Data Management Protocol, 网络数据管理协议, 关于文件服务系统基于网络的备份的一种开放协议, 允许独立于平台的数据存储。

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©2022 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.



存储系统操作管理



前言

- 本章重点介绍管理存储系统的三种工具：OceanStor DeviceManager，CLI和UltraPath，以及存储系统管理内容和相关操作。

目标

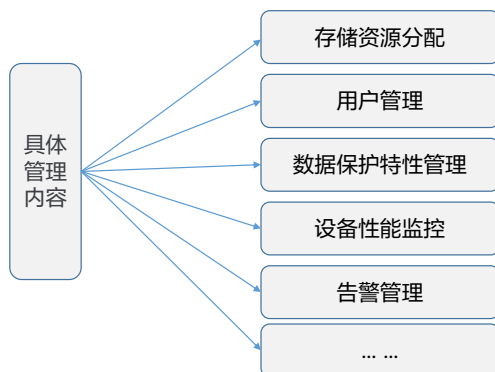
- 学完本课程后，您将能够：
 - 了解DeviceManager，CLI和UltraPath存储系统管理工具
 - 了解存储系统的基础管理操作

目录

1. 存储管理概述
2. 存储管理工具介绍
3. 基础管理操作介绍

什么是存储管理

- 存储管理就是用户可通过管理工具对存储系统进行查询、设置、管理及维护。



常用的存储系统访问方式



通过哪些方式能登录存储系统呢？

登录Web UI: DeviceManager界面

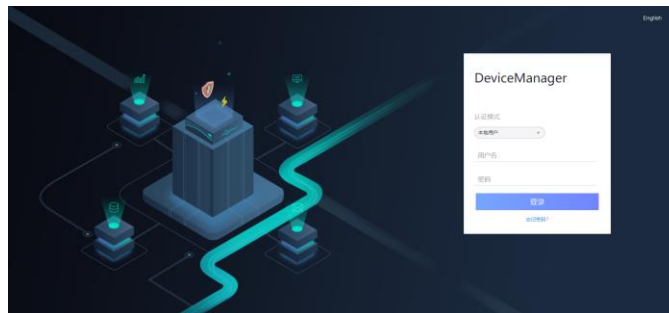
登录CLI管理界面



- 存储系统提供了DeviceManager和CLI两种操作与维护方式，以适应不同环境下、不同使用习惯的要求。
- 登录存储系统CLI控制界面，可以对存储系统进行查询、设置、管理及维护。

DeviceManager介绍

- DeviceManager是一款华为存储单设备管理软件，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是华为技术有限公司开发的集成存储管理软件。存储系统出厂时，DeviceManager管理软件已经加载在存储系统中。

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。

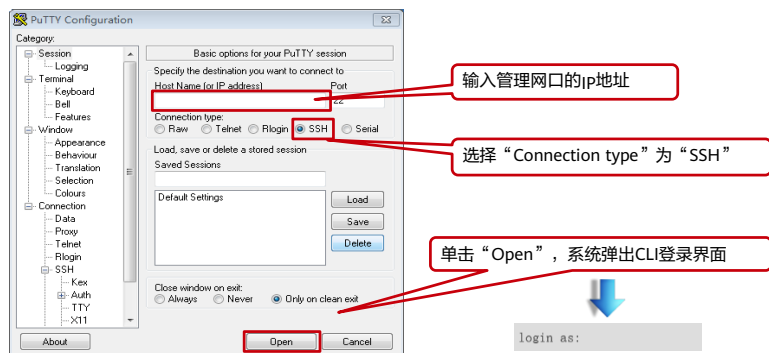
具体输入格式：`https://xxx.xxx.xxx.xxx:8088`



- 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。

CLI简介

- CLI (Command Line Interface) 提供命令行的方式对存储系统进行管理, 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。



- 有两种方式可以登录CLI:
 - 通过存储系统的串口进行登录。由于必须在设备旁边才能连接到串口, 因此该方式通常在不知道存储系统管理IP地址或者是存储系统异常时使用。
 - 通过存储系统的管理网口进行登录。在路由可达的情况下, 在任意一台与存储系统连接的维护终端上, 利用PuTTY等工具通过SSH协议访问存储系统控制器的管理网口IP地址, 即可登录CLI理界面。由于IP网络四通八达, 即使距离存储系统很远也能够登录, 因此该方式是最常用的登录方式。

CLI登录界面介绍

- 在CLI界面提示下，输入用户名和密码。初次登录时，系统会提示您修改密码，请立即修改密码，以保证系统的安全。登录成功后界面显示如下：

1、输入用户名

2、输入密码

3、修改密码

```
Storage login: 
Authorized users only. All activity may be monitored and reported.
Using keyboard-interactive authentication.
password: 
WARNING: You have accessed the system operated by Huawei.
You are required to have a personal authorisation from the system administrator before you use this computer. Unauthorized access to or misuse of this system is prohibited.

For security purposes, please change the initial password and log in to the system using the new password.
Old password:*****
New password:*****
Reenter password:*****

System Name      : Huawei.Storage
Health Status    : Normal
Running Status   : Normal
Total Capacity   : 6.240TB
SN               : XXXXXXXXXXXXXXXXXXXX
Location         :
Product Model    : XXXXX
Product Version   : XXXXX
Time             : XXXX-XX-XX/16:38:22 +08:00
admin:/>
```

UltraPath介绍

- OceanStor UltraPath是华为技术有限公司开发的多路径软件，能够实现屏蔽冗余LUN、选择最优路径、I/O负载均衡、故障切换和故障恢复等功能，为存储网络提供智能、稳定、快速的使用体验。



目录

1. 存储管理概述
- 2. 存储管理工具介绍**
3. 基础管理操作介绍

DeviceManager界面介绍 (1)



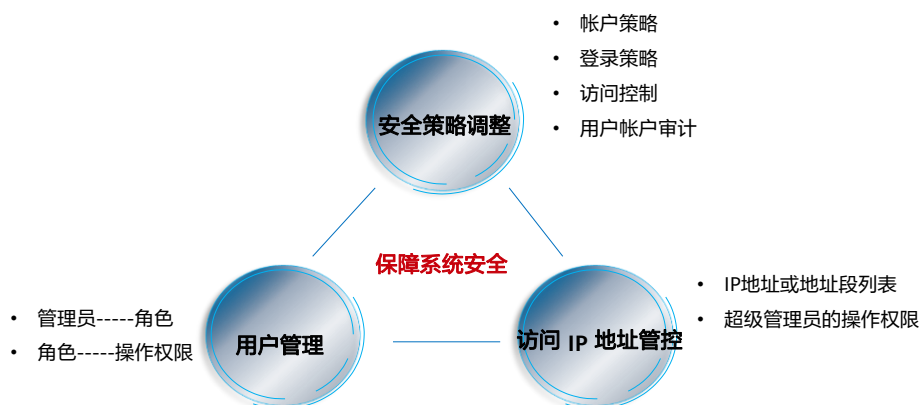
说明：不同产品版本和型号的界面可能会有少许差异，界面展示的信息请以实际界面为准。

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

DeviceManager界面介绍 (2)

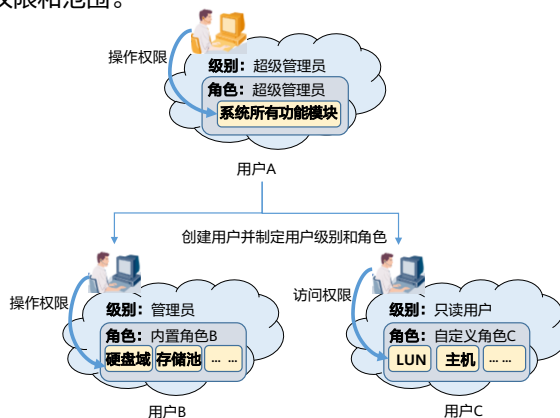
序号	名称	说明
1	信息展示区	信息展示区显示了当前操作的相关界面，并可供用户进行操作。
2	导航栏	导航栏上列出了整个存储系统的各个功能模块，从逻辑上将存储系统划分为几个模块。
3	告警和任务统计区	告警统计区上显示了系统各级别告警的数量，用户可以了解当前系统的运行状态。 任务统计区上显示了用户执行的所有任务，可以了解任务是否执行成功。
4	设备管理区域	设备管理区域支持查看并修改设备信息，下电或重启设备等功能。
5	注销和语言切换区	注销语言切换区列出了注销按钮和语言切换按钮。 DeviceManager 支持简体中文和英文两种语言类型。

管理存储系统访问权限



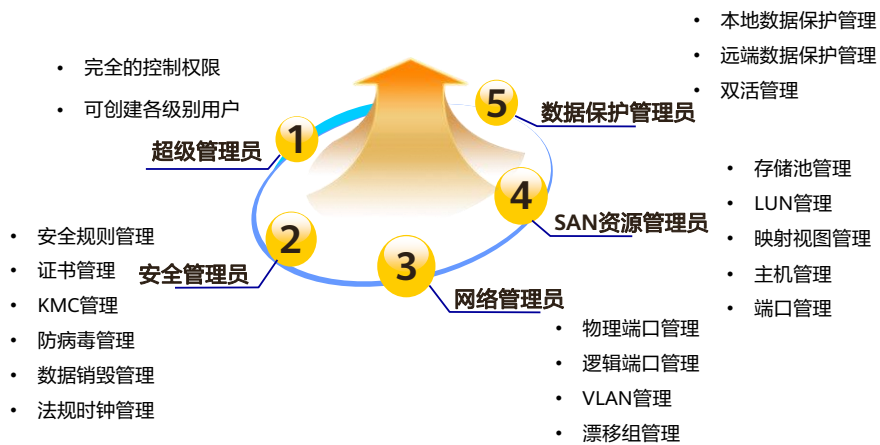
存储系统用户管理

- 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



DeviceManager Demo下载路径

登录华为技术支持网站：
support.huawei.com



选择企业业务网站，
选择“工具专区”



点击“Storage
Simulator Demo”，
选择相应版本的模拟
器进行安装



搜索“Storage Simulator”



- 可以在华为support网站上找到对应的产品，可以下载对应各种不同版本的DeviceManager的Demo。

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=?
```

1

2

3

- 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，命令 `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。

① 当需要查询所有命令的第一个字段时，按一次“Tab”

```
admin:/> //按“Tab”  
^  
add      change    create  
delete   exit      export  
help     import    poweroff  
poweron  reboot    remove  
scan     show      swap
```

② 当输入了命令的前一个字段，需要查询所有可用的下一个字段时，按一次“Tab”

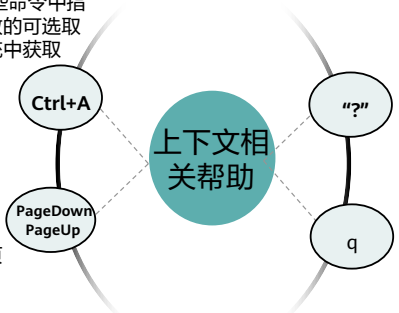
```
admin:/>add//按“Tab”  
cache_partition  clone      consistency_group  
disk_domain      host       host_group  
lun_copy         lun_group  mapping_view  
notification      port       port_group  
remote_device    security_rule smartqos_policy  
snmp             storage_pool
```

③ 当输入了命令所有所需的字段，符合命令可执行的条件时，按“Tab”后系统会提示命令可以被执行。此时可以直接按“Enter”执行命令

```
admin:/>add port ipv4_route eth_port_id=0 type=net target_ip=192.168.3.0 mask=255.255.255.0 gateway=10.0.0.1//按“Tab”  
Command is executable now.
```

- 当输入了命令某个字段的一部分，并且已输入的字母能够确定是某个字段时，接着按“Tab”就可以补全此字段，并提示所有可用的下一字段。

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
 - 通过输入“？”，可以查询CLI基本操作使用说明，以及命令参数的详细说明
 - 当输入命令的前一个字段并空格后，输入“？”可以查询所有可用的下一个字段及每个字段的详细说明
 - 退出上下文帮助界面
- 
- ```
graph TD; A((Ctrl+A)) --- C((上下文相关帮助)); B((PageDown
PageUp)) --- C; D((?)) --- C; E((q)) --- C;
```

- 在CLI中，上下文相关帮助可以在命令输入时提供实时的帮助和指引。

# CLI命令过滤功能



## CLI列过滤命令 - filterColumn

```
show xxx|filterColumn { exclude | include } columnList=?
```

二者选其一

可供筛选的列项

exclude: 过滤掉不需展示的信息

Include: 仅保留需要展示的列项

当需要查询多个列项时, 使用","隔开

```
admin:/>show bbu general|filterColumn exclude//按 "Tab"
<columnList=?> column list separated by comma, select one or more
 seperated by comma, the spaces are replaced with \s in the
 parameter list.
columnList=Inter\sID columnList=ID
columnList=Health\sStatus columnList=Running\sStatus
columnList=Current\sVoltage(V) columnList=Number\sOf\sDischarges
columnList=Firmware\sVersion columnList=Delivered\sOn
columnList=Owning\sController columnList=Electronic\sLabel
```

```
admin:/>show bbu general |filterColumn include
columnList=Inter\sID,ID
```

```
Inter ID ID

0.0A.0 CTE0.0
0.0A.1 CTE0.1
```

## CLI行过滤命令 - filterRow

```
show xxx|filterRow column=? predict=? [predict2=?] value=? [logicOp=?]
```

需要筛选的列项

筛选条件

筛选条件2

取值

列项间的逻辑关系

### predict=?

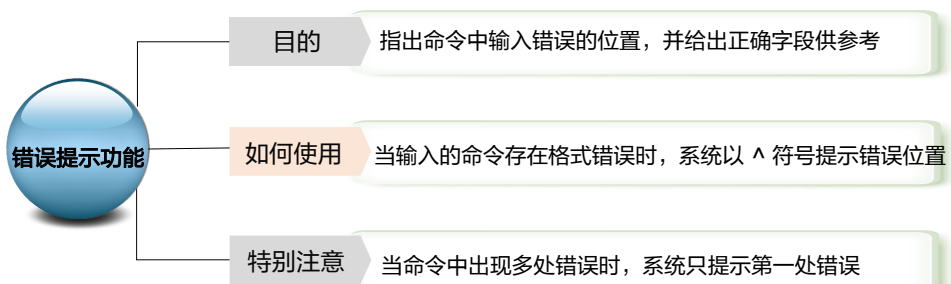
- not: 逻辑“非”
- equal\_to: 等于value=?的值。
- greater\_than: 大于value=?的值。
- greater\_equal: 大于等于value=?的值。
- less\_than: 小于value=?的值。
- less\_equal: 小于等于value=?的值。
- match: 正则匹配value=?的值。

### logicOp=?

- and: 显示同时满足多个column=?参数取值的信息。
- or: 显示满足任意一个column=?取参数值的信息。

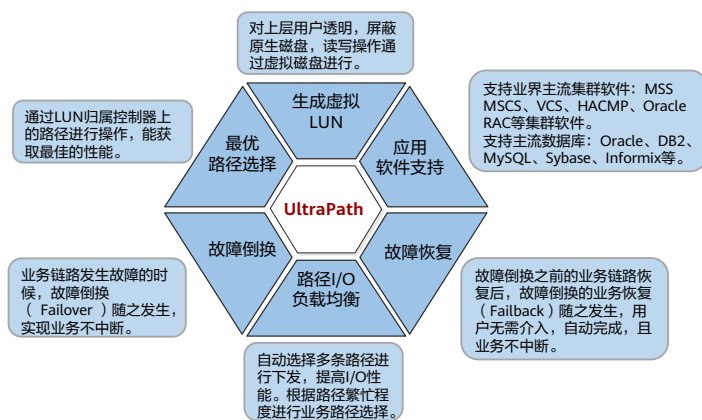
- xxx表示查询命令的后几个字段。
- 可同时选择多个列项，并使用logicOp=? 建立多个列项之间的逻辑关系。
- 当选择的列项名中存在空格时，中间使用“\s”代替。例如查询LUN的“Write Policy”列项值，命令格式为admin:/>show lun general|filterRow column=Write\sPolicy。

# 错误提示功能



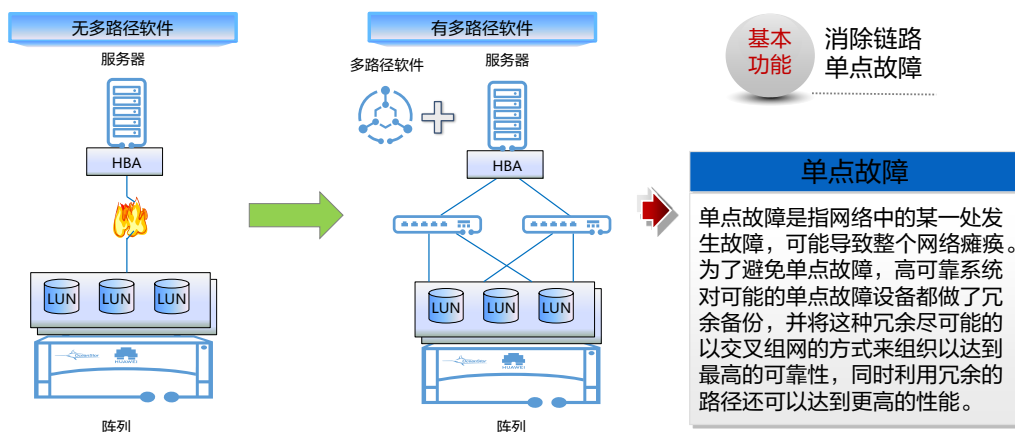
```
admin:/>add part
 ^
port port_group
admin:/>add part
```

## 华为多路径软件主要功能

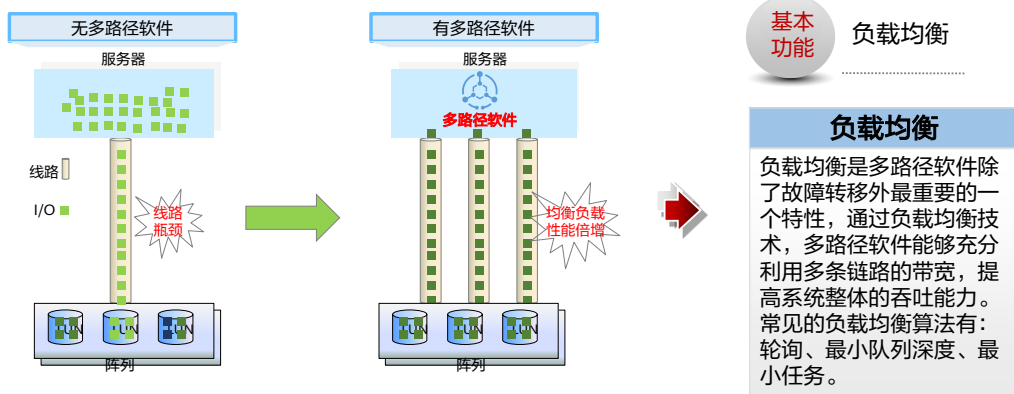




## 多路径软件基本定位



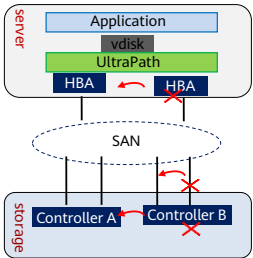
# 多路径软件基本定位



# 多路径软件基本定位

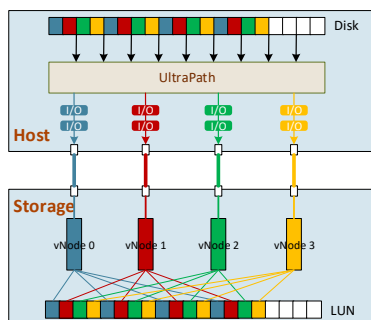
## 定位

UltraPath是运行在主机内核中的一种过滤驱动软件，可以拦截OS原生的磁盘创建、删除以及I/O下发等事件并进行处理。利用多路径软件可以实现冗余路径的可靠利用。如果一条路径不能使用，或不能满足规定的性能要求，多路径软件会自动而透明地将I/O流转移到其他可用的路径上，确保I/O流有效、可靠地继续传输。如图所示，多路径软件可以处理包括HBA故障、链路故障、存储控制器故障在内的多种路径故障。



| 基本功能           | 重要度 | 概述                        |
|----------------|-----|---------------------------|
| Failover(故障转移) | 高   | 路径故障时可以自动将I/O转移到其他可用路径    |
| Failback(故障恢复) | 高   | 检测到路径故障恢复后会自动重新将业务I/O转移回去 |
| 负载均衡           | 高   | 利用多条链路的带宽，提高系统整体的吞吐能力     |

## 支撑OceanStor V6 全均衡（A-A）架构



### OceanStor LUN无归属，均衡打散

Lun的数据，按照64MB的粒度，划分成Slice。每个Slice按照Hash（LunID+LBA）的结果，分发到不同的VNode上。

### 负载均衡

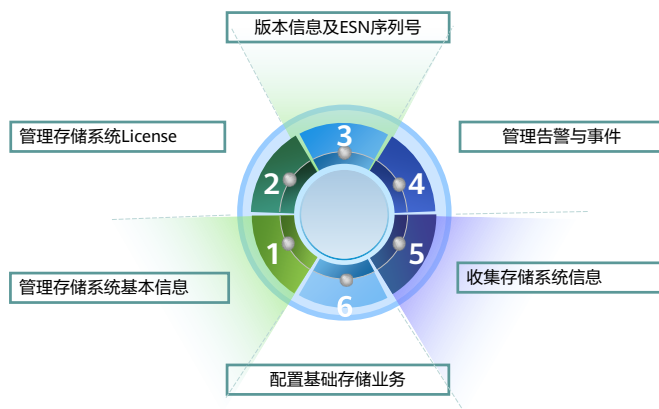
UltraPath通过与存储系统进行协商交互，对每个下发的I/O计算Hash（LunID+LBA）得到对应处理的VNode，选择VNode对应的物理路径下发I/O，减少存储节点内跨CPU分发，提升端到端性能。

# 目录

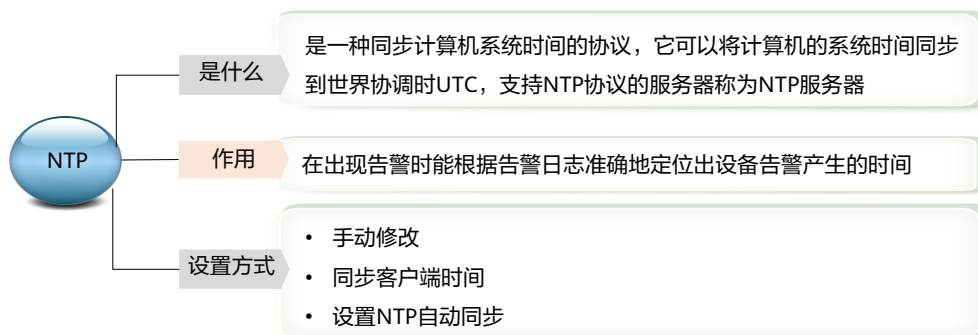
---

1. 存储管理概述
2. 存储管理工具介绍
- 3. 基础管理操作介绍**

# 基础管理操作



## 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：导入License、激活License和更新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

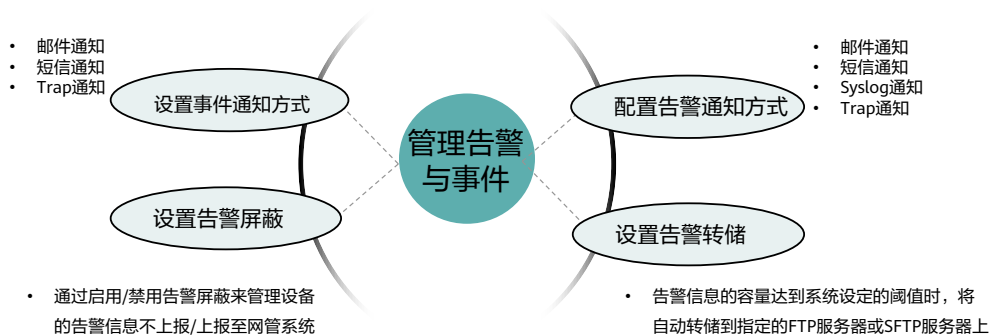


- 使用CLI获取ESN序列号

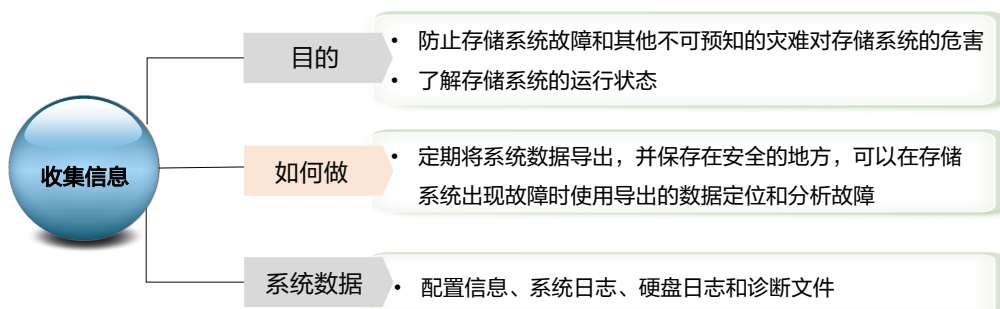
>> 执行 show system general 命令

```
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 :
```

# 管理告警与事件



## 收集存储系统信息 (1)



## 收集存储系统信息 (2)

### 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。

## 配置存储基础业务



作用

- 将存储系统提供的存储空间划分为多个LUN
- 将LUN映射给应用服务器
- 应用服务器能够使用存储系统提供的存储空间

### 使用 DeviceManager配置存储基础业务

- 创建存储池：DeviceManager支持通过“推荐”或“自定义”的两种方式创建存储池。
- 分配存储资源，继续通过创建LUN组或者创建文件系统的形式将存储资源进行分配。



# 使用CLI配置基础存储业务



# 华为多路径基本配置操作指导

UltraPath命令行配置方式：

以Windows系统为例，通过点击“开始->所有程序->UltraPath->upadm”或直接在Windows的CMD命令行中输入upadm，进入upadm命令行模式，并通过相应CLI命令配置UltraPath。

注：文档中主要以Windows系统为例介绍华为多路径的基本配置命令，其他操作系统配置命令与Windows系统类似，具体命令请参考相应系统的《用户指南》。

- 华为多路径安装、卸载操作具体步骤及注意事项请参考对应多路径版本各个操作系统的《 OceanStor UltraPath for xxxxxx 用户指南 》文档。升级操作具体步骤及注意事项请参考多路径对应版本的《 OceanStor UltraPath xxxxxx 升级指导书 》。
- 用户指南和升级指导书的获取地址：
  - 企业业务：<http://support.huawei.com/enterprise/zh/storage/ultrapath-pid-8576127>

# 华为多路经常用参数在典型场景下的配置

大多数场景下，推荐使用华为多路径的默认配置。部分典型场景，可参考以下说明对华为多路径进行配置：

upadm set workingmode={0/1}

- 设置存储控制器层面的负载均衡模式，0为控制器间负载均衡，1为控制器内负载均衡。
- 默认为控制器内负载均衡，多路径会根据每个存储LUN的归属控制器去选择路径下发I/O。
- 而控制器间负载均衡时，多路径会在所有路径上下发I/O，但这会增加I/O在存储控制器间转发的处理时延。

| 典型场景                 | 推荐配置             |
|----------------------|------------------|
| 主机和存储系统之间的传输路径成为性能瓶颈 | 0（控制器间负载均衡）      |
| 其他场景                 | 1（默认配置，控制器内负载均衡） |

## 华为多路经常用参数在典型场景下的配置

upadm set loadbalancemode={round-robin|min-queue-depth|min-task}

- 设置链路层面的负载均衡算法。round-robin即为轮询算法，min-queue-depth即为最小队列深度算法，min-task即为最小任务算法。
- 默认为最小队列深度算法，多路径会根据当前每一条可用路径上的I/O个数，选择所有路径中I/O数目最少的路径下发I/O。
- 使用轮询算法时，多路径会根据应用服务器和存储系统当前的可用路径，依次选择每一条路径下发I/O数据。

| 典型场景                              | 推荐配置                           |
|-----------------------------------|--------------------------------|
| 主机下发的业务I/O模型差异较小，且需要业务在每一条路径上均衡下发 | round-robin（轮询算法）              |
| 主机下发的业务I/O模型数据块较大                 | min-task（最小任务算法）               |
| 其他场景                              | min-queue-depth（默认配置，最小队列深度算法） |

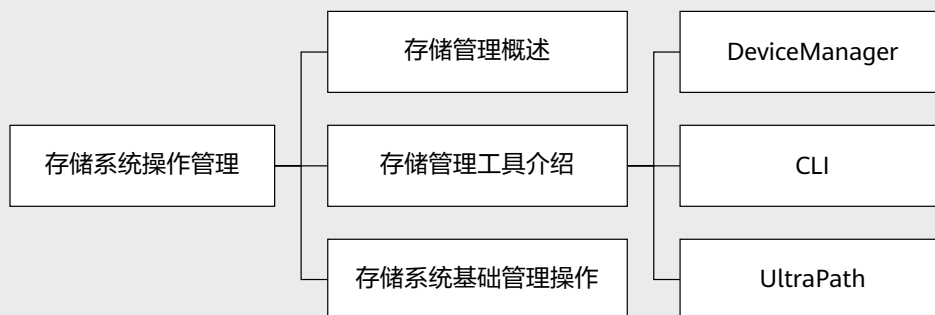
## 思考题

1. （单选题）某存储设备的管理IP地址为192.168.5.12，工程师A需要在浏览器输入以下哪些内容才可以正确登录？
  - A. 192.168.5.12
  - B. http://192.168.5.12
  - C. https://192.168.5.12
  - D. https://192.168.5.12:8088
2. （判断题）通过DeviceManager可以监控控制器、前端、后端端口的性能。

- 参考答案：

- D
- T

## 本章总结



## 学习推荐

- 华为官方网站
  - 企业业务: <http://enterprise.huawei.com/cn/>
  - 技术支持: <http://support.huawei.com/enterprise/>
  - 在线学习: <http://learning.huawei.com/cn/>
- 热门工具
  - HedEx Lite
  - 网络资料工具中心
  - 信息查询助手

- 热门工具:
  - HedEx Lite: 华为产品文档管理工具, 支持浏览、搜索、更新和管理产品资料
  - eStor: 图形化存储仿真平台, 该平台通过对真实华为OceanStor全闪存仿真模拟, 帮助广大ICT从业者和用户快速熟悉华为存储系列产品, 了解并掌握相关产品的操作和配置。
  - 网络资料工具中心: 汇集网络产品的资料工具, 是您进行投标支撑、网络规划、项目交付和升级维护的好助手。
  - 信息查询助手: 提供华为产品命令和告警信息查询。

## 缩略语表

LUN: Logical Unit Number 逻辑单元号，是用来标识一个逻辑单元的数字，这个逻辑单元是通过SCSI寻址的设备。

SAN: Storage Area Network，存储区域网络

VLAN: Virtual Local Area Network，虚拟局域网 一种在交换局域网的基础上，采用网络管理软件构建的可跨越不同网段、不同网络的端到端的逻辑网络。

NTP: Network Time Protocol，网络时间协议，是应用层协议，用于在分布式时间服务器和客户端之间进行时间同步。



# 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©2022 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.



## 存储资源调优技术与应用



## 前言

- 本章节主要讲述存储资源调优的相关技术，其中包含了SmartThin、SmartTire、SmartCache 、 SmartAcceleration 、 SmartQoS 、 SmartDedupe 、 SmartCompression、SmartVirtualization和SmartMigration等功能特性，各功能特性的业务特点、实现原理和应用场景。

# 目标

- 学完本课程之后，你将能够了解以下特性的业务特点、实现原理和应用场景：
  - SmartThin
  - SmartTire&SmartCache
  - SmartAcceleration
  - SmartQoS
  - SmartDedupe&SmartCompression
  - SmartVirtualization
  - SmartMigration

# 目录

---

1. **SmartThin**
2. SmartTier&SmartCache
3. SmartAcceleration
4. SmartQoS
5. SmartDedupe&SmartCompression
6. SmartVirtualization
7. SmartMigration

## 概述

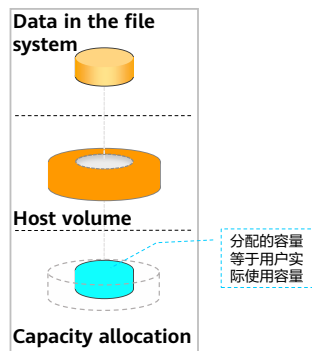
- 传统存储系统部署方式存在下面一些问题：
  - 存储空间扩容时影响，甚至中断业务系统
  - 存储空间利用率不均衡
  - 存储效能比较低
- SmartThin
  - 使用过程中采用“按需分配”的存储空间分配策略，提高存储资源使用效率，更大限度满足业务的实际要求。

- 当出现数据容量超过预期的情况时，可以动态调整该LUN的空间。未使用的空间作为公共的空间可以分配给任何需要空间的LUN。这样，不存在私有的一直不能被使用到的空间，提高了利用率和效能比。同时，动态空间调整提供了在线调整LUN空间大小的能力，可以做到扩容的同时不影响业务。
- SmartThin技术做了如下改进：
  - 改进一：采用SmartThin技术的LUN空间在创建时不真正分配具体空间，而在需要真正使用空间时，再分配具体空间。
  - 改进二：有了改进一的基础后，采用SmartThin技术的LUN空间可以在创建后动态调整大小。

## 精简LUN

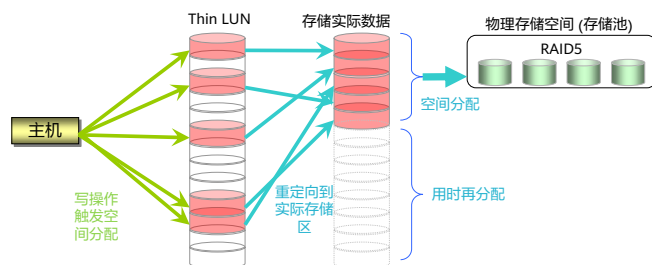
- 定义：主机可访问的一个逻辑硬盘，按用户实际使用容量从存储池中动态分配存储资源，英文名称为thin LUN。

- ✓ 数据集合：对存储阵列来说是可映射给主机的LUN。
- ✓ 完全可用：可以正常读写。
- ✓ 动态分配：写时分配资源。



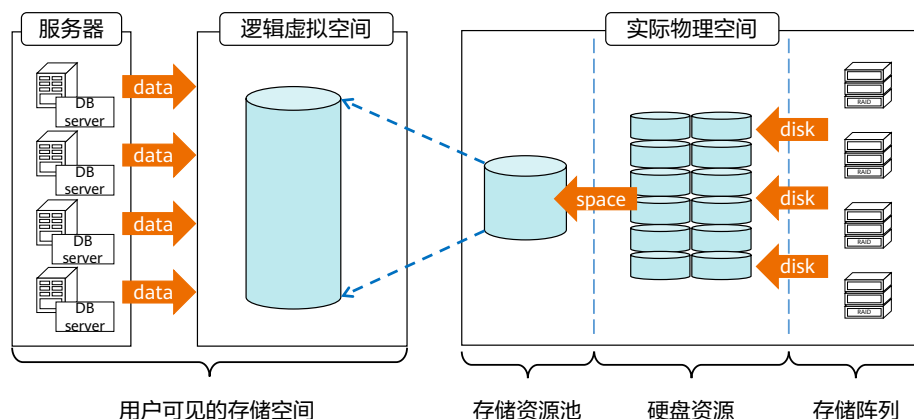
## 存储虚拟化

- Capacity-on-write: 实现写时空分配, 空间按照均衡原则从内部各引擎分配。
- Direct-on-time: 实现读写重定向。



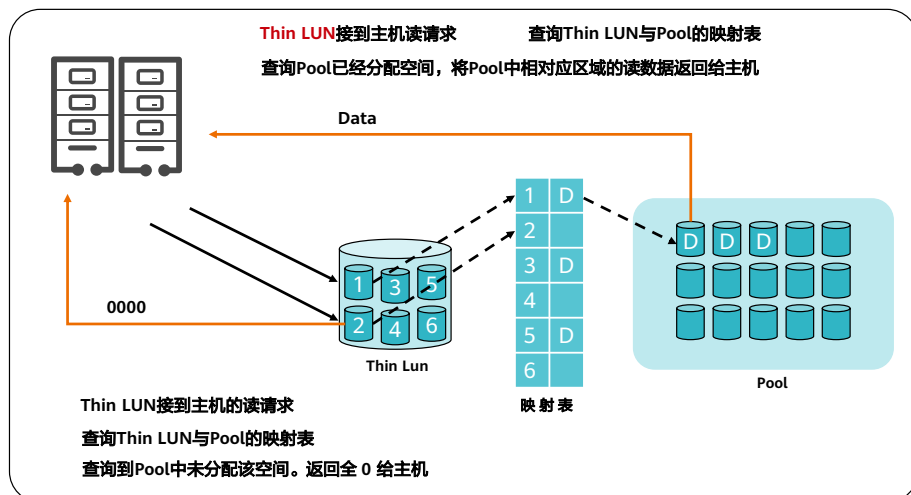


## SmartThin工作原理



- SmartThin 的工作原理为：存储资源的虚拟化。
- SmartThin以一种按需分配的方式来管理存储设备。
- SmartThin不会预先分配所有空间，而是将大于物理存储空间的容量形态呈现给用户，使用户看到的存储空间远远大于系统实际分配的空间。用户对这部分还未分配空间的使用实现“按需分配”的原则，即：用多少提供多少。如果用户的物理存储空间不足，可通过扩充后端存储单元的方式来进行系统扩容，整个扩容过程无需系统停机，对用户完全透明。
- SmartThin基于RAID 2.0+存储虚拟资源池创建Thin LUN，即Thin LUN和传统的Thick LUN共存于同一个存储资源池中。精简LUN（Thin LUN）是在存储池中创建的并可以映射为主机直接访问的逻辑单元。Thin LUN的容量大小并不是实际的物理空间，而是一个虚拟值，只有在对Thin LUN进行真正I/O读写时，才通过写时分配的策略从存储资源池中申请物理空间。
- SmartThin 允许主机可感知容量大于Thin LUN实际存储空间。主机可感知容量指的是用户能够创建的Thin LUN大小，也即是Thin LUN创建成功后映射给主机，在主机侧显示的卷容量（逻辑虚拟空间）；Thin LUN实际存储空间指的是Thin LUN真实占用的存储池物理空间的大小。SmartThin会向主机隐藏Thin LUN实际存储空间大小，而向主机提供Thin LUN的名义存储空间。
- 除此之外，SmartThin 支持创建大于存储池最大物理可用空间的Thin LUN。例如存储池提供的最大物理空间是 2TB，但SmartThin支持创建大于10 TB的Thin LUN。

## SmartThin读操作



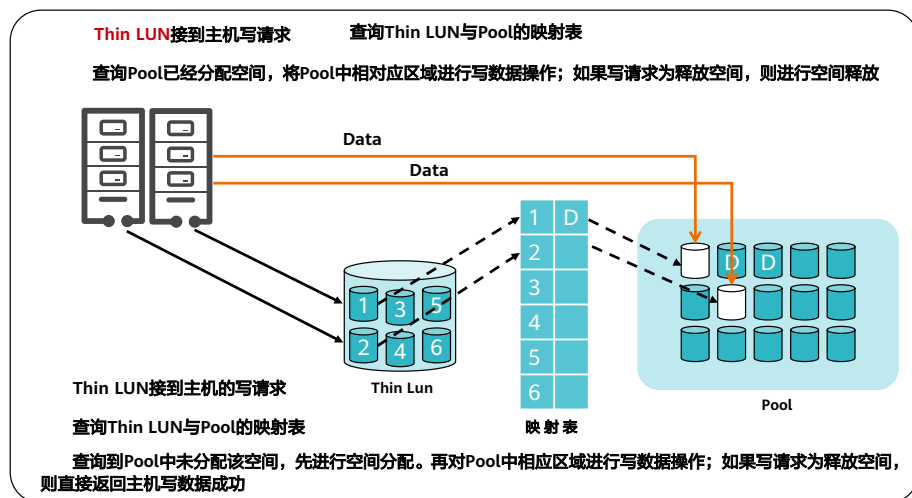
- SmartThin 的读操作步骤如下：

- 当Thin LUN收到主机读请求，首先查询Thin LUN与Pool的映射表，确认Pool是否已经分配实际存储空间给Thin LUN。
- 如果Pool已经分配实际存储空间，SmartThin 通过 Direct-on-time 技术从Pool中对应的实际存储空间中读取出数据并返回给主机。
- 如果Pool尚未分配实际存储空间，则表明该数据未曾写入。SmartThin 返回全 0 给主机。具体过程如图所示。

- Direct-on-time

- 由于采用了 Capacity-on-write 技术，数据的实际存储区域和逻辑存储区域的关系不再是按照确定的公式可以固定不变计算出来的，而是按照写时分配的原则随机映射确定的。所以在对Thin LUN进行读写时需要重定向实际存储区域和逻辑存储区域的关系，重定向依赖于映射表。映射表的主要作用是用来记录实际存储区域和逻辑存储区域的映射关系。在写过程中动态更新映射表，在读过程中查询映射表。因此，Direct-on-time 重定向操作也就分为读重定向和写重定向。

## SmartThin写操作



- SmartThin 的写操作步骤如下：

- 当Thin LUN收到主机写请求，首先查询Thin LUN与Pool的映射表，确认Pool是否已经分配实际存储空间给Thin LUN。
- 如果Pool已经分配实际存储空间，对Pool中相应区域进行写数据操作（Direct-on-time 技术实现）。如果写请求为释放空间，则进行空间释放后返回主机写成功。
- 如果Pool尚未分配实际存储空间，SmartThin 先通过 Capacity-on-write 技术从Pool中给写请求分配实际存储空间，然后再通过 Direct-on-time 技术建立逻辑存储空间与实际存储空间的联系，对分配的实际存储空间写入数据。如果写请求为释放空间，则直接返回主机写成功。

- Capacity-on-write

- 当Thin LUN接收到主机写数据请求，首先会通过Direct-on-time 技术判断该写数据请求的逻辑存储区域是否已经分配了实际存储区域，如果尚未分配就会触发空间分配，分配的最小粒度称为Grain，Grain大小为64 KB（说明：对于开启重删或压缩功能的Thin LUN，该粒度创建时可配置），然后将数据写入到新分配的实际存储区域中。

## 应用场景

- 对业务连续性要求较高的系统核心业务，使用智能精简配置，可以在线对系统进行扩容，不会中断业务。例如：银行票据交易系统。
- 应用系统数据增长速度无法准确评估的业务，使用智能精简配置，可以按需分配物理存储空间，避免浪费，例如：E-mail 邮箱服务、网盘服务等。
- 多种业务系统混杂并且对存储需求不一的业务，使用智能精简配置，可以让不同业务去竞争物理存储空间，实现物理存储空间的优化配置，例如：运营商服务等。

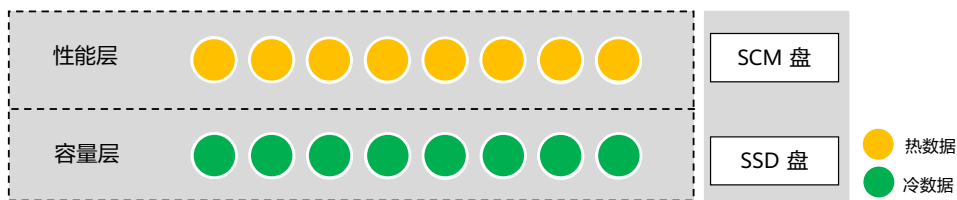
# 目录

---

1. SmartThin
- 2. SmartTier&SmartCache**
3. SmartAcceleration
4. SmartQoS
5. SmartDedupe&SmartCompression
6. SmartVirtualization
7. SmartMigration

## SmartTier

- 分级存储是一种利用数据迁移，实现冷热数据或者不同使用价值的数据在不同介质上进行存储迁移，从而有效的平衡性能。
- SmartTier特性又叫智能数据分级特性。SmartTier提供智能化数据存储管理的功能，通过统计和分析数据的活跃度，将不同活跃度的数据和不同特点的存储介质动态匹配。



- 由于硬盘技术的进步，存储系统支持的存储介质类型和数量已显著增加。SCM介质作为一种比SSD性能更好的介质，在存储系统中主要用作Cache和主存，用来加速关键应用。在OceanStor Dorado 6.1.0及后续版本中，通过增加支持SCM介质实现数据的分级存储，进而提升业务性能。
  - 热数据：存储在或迁移至性能层硬盘的较活跃数据；
  - 冷数据：存储在或迁移至容量层硬盘的“空闲”数据，且数据迁移后，其现有性能不会受到影响。

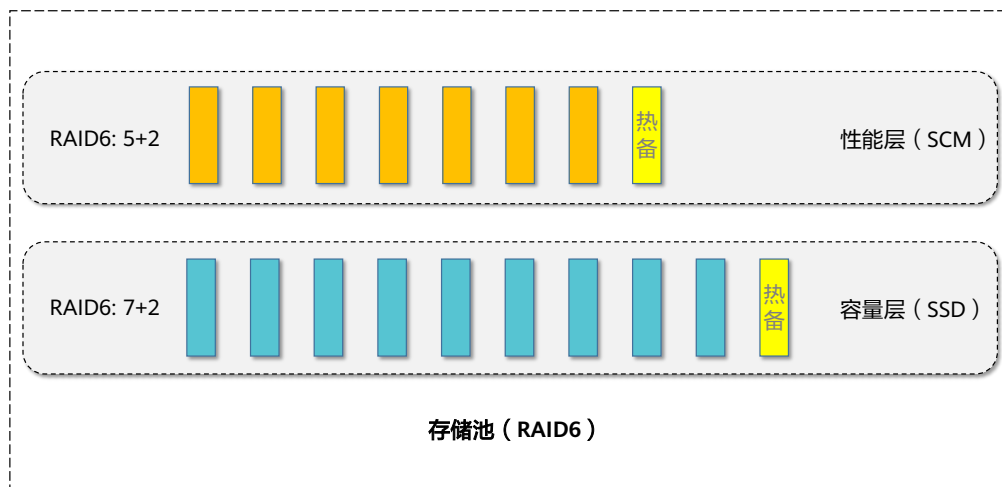
## 存储层层级划分

- 在同一存储池中，存储层是相同性能的存储介质集合。SmartTier将不同存储介质按性能高低划分为两个存储层，由高到低分别是性能层和容量层（SCM做性能层、SSD做容量层）。每个存储层分别使用相同的硬盘类型和RAID策略。

| 存储层级 | 硬盘类型 | 硬盘特点                     | 数据特点                 |
|------|------|--------------------------|----------------------|
| 性能层  | SCM盘 | SCM盘的响应时间很短，每单位存储容量成本很高。 | 性能层适宜存储访问频率非常高的数据。   |
| 容量层  | SSD盘 | SSD盘的响应时间较短，每单位存储容量成本适中。 | 容量层适宜存储访问频率适中或较低的数据。 |

- 存储池是一个或多个存储层的逻辑组合，最多可支持三个级别的存储层。存储池包含的硬盘类型决定其可以创建的存储层。由单一硬盘类型组成的存储池无法创建不同存储层，因此不能应用 SmartTier进行智能化数据存储管理。
- 由于硬盘技术的进步，存储系统支持的存储介质类型和数量已显著增加。SCM介质作为一种比SSD性能更好的介质，在存储系统中主要用作Cache和主存，用来加速关键应用。在OceanStor Dorado 6.1.0及后续版本中通过增加支持SCM介质实现数据的分级存储。
- OceanStor Dorado支持SCM和SSD混插，SCM是比SSD性能更优的介质。当存储池中存在这两种硬盘类型时，主机新写入的数据优先存入性能层，以提高主机写入存储的性能；当数据长时间不被访问时，存储系统会将其迁移到容量层。当优选层级空间不足时，从其他层级分配。

## 成员盘管理



- 创建了SmartTier的存储池，SCM和SSD盘都作为成员盘一起管理。
- 热备盘数：每层tier单独预留自己的热备盘，比如存储池配置热备策略为低时，性能层和容量层分别预留1盘做热备。
- 大小盘：每层tier独立计算自己的大小盘，每层tier允许配置最多2种容量的盘。
- RAID：不同介质分别组RAID，不跨介质组CKG。
- 容量：存储池可用容量为各层tier可用逻辑容量之和。



## 数据迁移

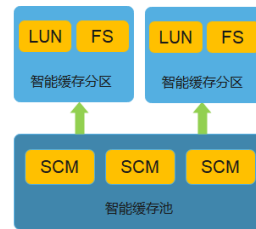
- SCM是比SSD性能更优的介质，配置SmartTier后，主机新写入数据优先存入性能层，以提高主机写入存储的性能，在后台将长时间不访问的冷数据迁移到容量层中。
- 数据管理和迁移策略如下：
  - **元数据：**元数据业务写和回收写（GC）都优先保存在性能层（SCM）中，保证大容量场景高性能访问。当性能层空间不足时，才从容量层分配空间。
  - **用户数据：**用户数据优先保存在性能层（SCM）中，当性能层达到一定水位之后，后台就会启动把性能层的冷数据迁移到容量层（SSD）中，保证性能层可以保存更多的热数据和元数据，提高用户的访问速度。

- 当存储池中存在SCM和SSD两种硬盘类型混插时，主机新写入的数据优先存入性能层，以提高主机写入存储的性能。当数据长时间不被访问时，存储系统会将其迁移到容量层。当优选层级空间不足时，从其他层级分配。

# SmartCache

## 新介质SCM的优势:

- 比SSD更低的时延和更高的IOPS
- 比DRAM更大的容量和价格优势
- 适合做缓存



## • SmartCache特性:

基于SCM盘对数据读取响应时间短的特点，利用SCM盘组成智能缓存池，将访问频率高的读热点数据缓存到智能缓存池中，缩短热点数据的读取响应时间，有效提高热点数据的访问效率，从而提升系统的性能。

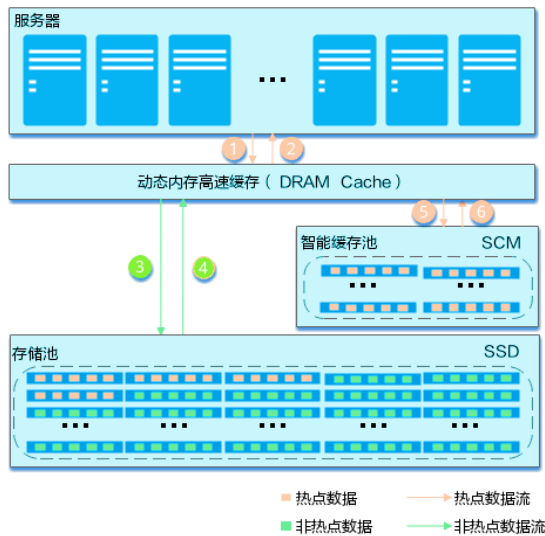
## • 智能缓存池（SmartCache Pool）:

由SCM盘组成，作为DRAM Cache的扩充，用于存储热点数据。

## • 智能缓存分区（SmartCache Partition）:

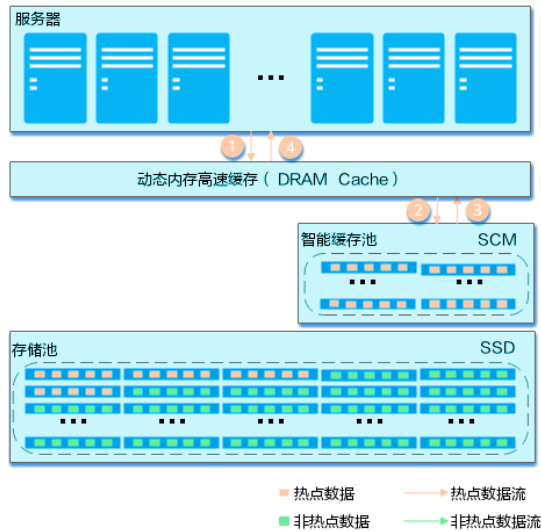
建立在智能缓存池之上的逻辑概念，将需要LUN和文件系统业务放入其中。

# SmartCache写流程



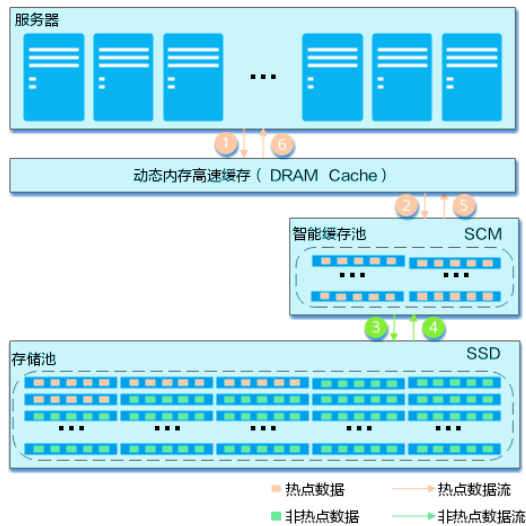
1. 当存储系统收到服务器发起的LUN或文件系统写I/O请求时，先将数据发送到DRAM Cache。
2. IO写入DRAM Cache完成，返回服务器。
3. DRAM Cache将数据发送到存储池管理模块。
4. 数据持久化到SSD盘，IO返回。
5. DRAM Cache把数据副本发送到SmartCache中，数据经过冷热识别算法过滤后将识别出的热点数据写入SCM，并在内存中建立数据和SCM映射关系的元数据。
6. 数据缓存至智能缓存池中，IO返回。

# SmartCache读命中



1. 当存储系统收到服务器发起的LUN或文件系统读I/O请求时，将首先访问DRAM Cache。
2. DRAM Cache中没有找到对应数据，DRAM Cache将读I/O请求转发到SmartCache。
3. 数据在智能缓存池中，直接从SCM中读取相应数据并将数据返回至DRAM Cache。
4. DRAM Cache将数据返回至服务器。

# SmartCache读未命中



1. 当存储系统收到服务器发起的LUN或文件系统读I/O请求时，将首先访问DRAM Cache。
2. 在DRAM Cache未命中，DRAM Cache将读I/O请求转发到SmartCache。
3. 在SmartCache也未命中，SmartCache模块将读I/O请求转发至存储池管理模块，从SSD盘读取对应数据。
4. 在SSD盘中读取到对应的数据后，将数据返回至SmartCache模块。
5. SmartCache模块将数据返回至DRAM Cache。
6. DRAM Cache将数据返回至服务器。

## SmartCache的特点

- **支持SmartCache池动态扩容**

- 用户可以在需要的时候可在线向智能缓存池添加新的SCM来实现动态扩容。
- 建议对同一引擎上的各控制器都配置相同个数和容量的SCM资源，这样可保证工作在同一引擎中的多个控制器上的LUN或文件系统的加速效果相当。

- **灵活的开启策略**

- 支持对指定LUN或者文件系统开启或关闭SmartCache功能，不影响业务正常运行。
- 支持将指定的单个或多个LUN或者文件系统添加到指定的智能缓存分区来缩短数据的读取响应时间。

- **自适应开关**

- 当检测到SmartCache策略低效（例如SmartCache命中率低的业务、CPU工作繁忙的情况）时，会主动关闭SmartCache IO开关，减少对业务的影响，此时IO将不会进入SmartCache中缓存，读IO也不会SmartCache命中。为保证数据的一致性，同时会清空SmartCache中的数据。
- 当检测到适合SmartCache的场景，会主动打开IO开关，恢复到之前的状态。

# 目录

---

1. SmartThin
2. SmartTier&SmartCache
- 3. SmartAcceleration**
4. SmartQoS
5. SmartDedupe&SmartCompression
6. SmartVirtualization
7. SmartMigration

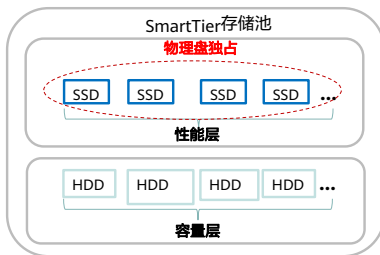
# SmartAcceleration

- 性能层需SSD物理盘独占，无法多DD共享
- 性能层一旦配置难以调整，灵活性不够
- Tier/Cache相互独立，都需单独规划物理盘

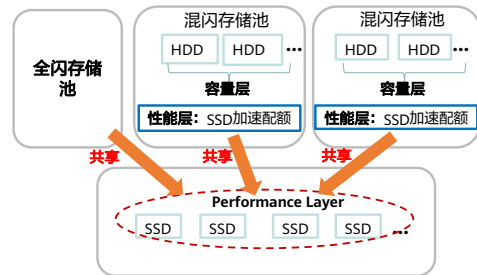


- 性能层无需独占，多DD共享，利用率更高
- 性能层系统默认推荐配额，也可灵活伸缩（增/减）
- Tier/Cache统一融合，系统内部自动配置，效率最佳

传统SmartTier



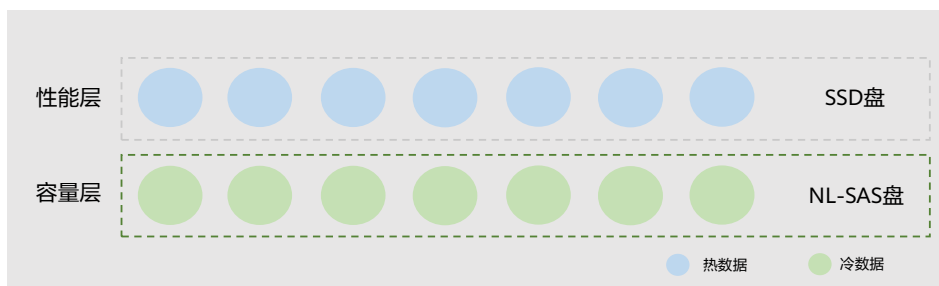
SmartAcceleration



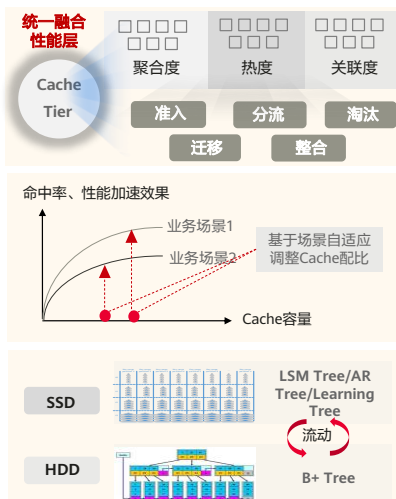
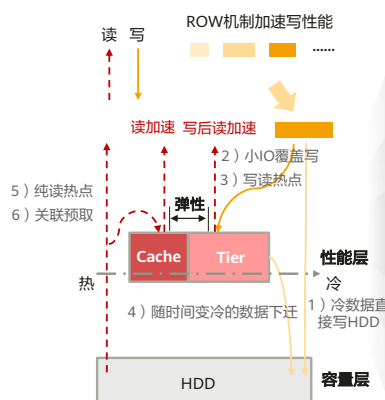


## 存储层

- 冷热数据的随机分布无法使各种介质的硬盘发挥各自最大特点，SmartAcceleration 技术采用多维自适应冷热数据感知算法实现统一性能层，自动识别数据冷热特征，对冷热数据自动布局，实现冷热数据最优分布。

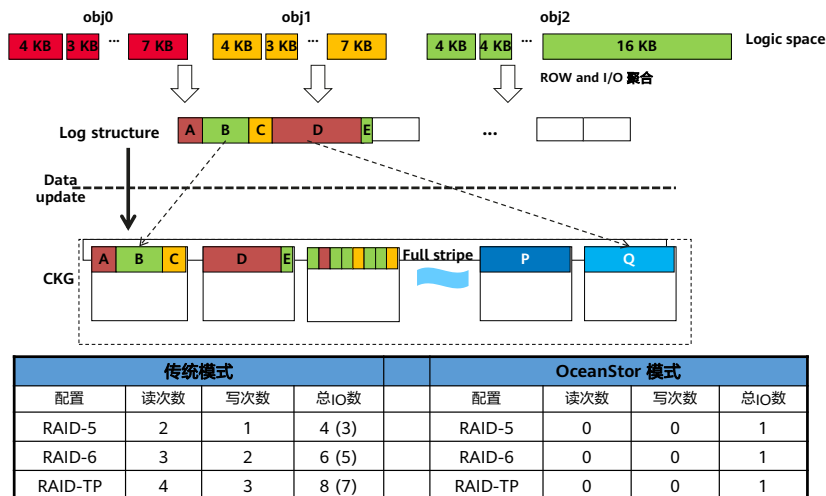


## 弹性融合性能层



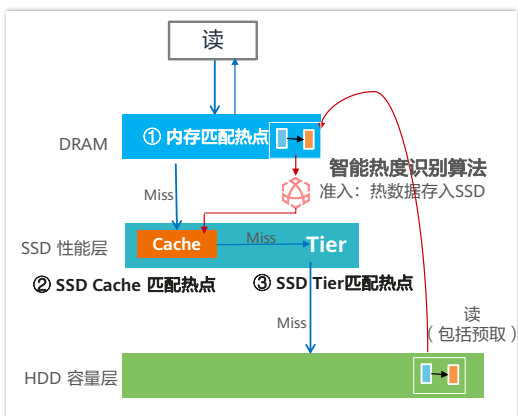
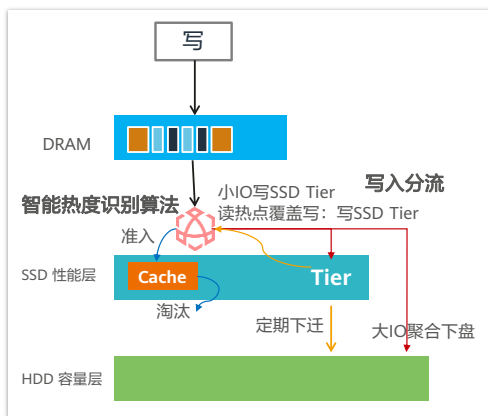
- 基于全局统一的冷热感知和数据协同算法，打破Cache和Tier的边界，实现最优数据布局 and 极简配置：
  - Cache和Tier不再相互割裂，避免数据反复流动效率低
  - 全局热度，统一Cache准入、淘汰和Tier分流、下迁
  - 感知不同介质的性能、容量、寿命，综合判断数据放置
- 基于场景自适应的弹性Cache调整，实现性能和容量的最佳平衡：
  - Cache弹性调整，和Tier可动态转化，追求最佳性价比
  - 客户无需独立配置Cache和Tier（物理），配置极简
- 性能层和容量层采用多模索引技术适应不同数据布局，实现多变混合模型下的效率最佳：
  - 性能层PelagoDB，中小容量规模，数据快速更新
  - 容量层KVDB，中大容量规模，读和写均衡效率

## ROW大块顺序写



- 盘控配合，也是华为OceanStor存储的一大特点，FlashLink-ROW大块顺序写机制，是指基于自研SSD，控制器感知硬盘内数据布局，从而把多段小块离散写，通过聚合成一个连续的大数据块写操作，顺序写入SSD。使得RAID5，RAID6和RAID-TP都只用做一次IO操作，不用产生小块离散写带来的多次读写操作。在写性能上RAID5，RAID6和RAID-TP保持一致，不再有明显差异。

## SmartAcceleration工作原理



# 目录

---

1. SmartThin
2. SmartTier&SmartCache
3. SmartAcceleration
- 4. SmartQoS**
5. SmartDedupe&SmartCompression
6. SmartVirtualization
7. SmartMigration

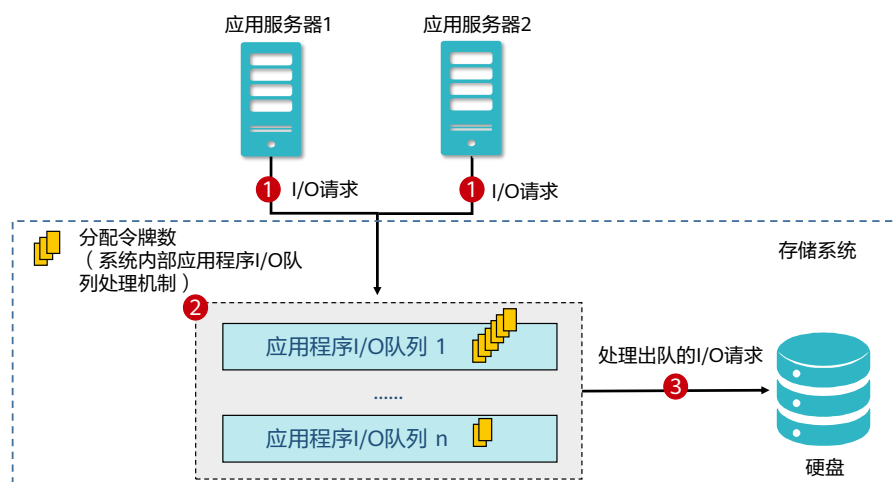
## 概述



- **SmartQoS:**  
又称智能服务质量控制，通过动态地分配存储系统的资源来满足某些应用程序的特定性能目标。
- SmartQoS的被控对象包括：**LUN、LUN组、主机以及文件系统。**
- SmartQoS依功能点可分为：流控管理、突发流控管理和下限保障。
- 依据配置简化，其它特点包括：支持多维度流控、策略分层和策略分配。

- SmartQoS是对存储系统的一个极为重要的增强特性，尤其是在某些应用程序需要高服务级别的环境中。
- 在同一台存储设备上部署多个应用程序时，通过合理配置SmartQoS特性，用户可以得到最大化整合的利益：
  - 通过性能调控，减少不同应用程序之间的相互影响，保证关键业务的性能。
  - 通过限制提供给非关键应用程序的资源，从而提供给关键应用程序更多的资源，保证关键应用程序的高性能。

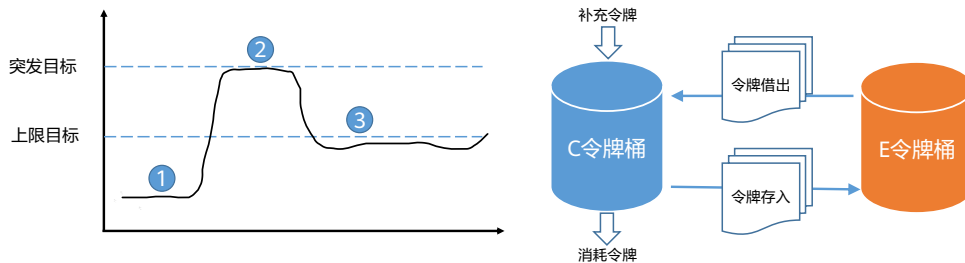
## 上限流控管理



- SmartQoS流控管理通过对被控对象的I/O分类队列管理、令牌分发和出队控制三部分实现。
- 基于被控对象的I/O队列管理通过令牌机制实现存储资源的分配，某个被控对象的I/O队列所拥有的令牌数越多，系统分配给这个被控对象的I/O资源也越多，存储系统会优先处理该被控对象的I/O队列中的I/O请求。
- SmartQoS为关联了策略的LUN维护一个令牌桶和一条IO队列，将QoS策略的上限配置值转换为令牌桶速度。
- SmartQoS策略配置类型为IOPS，一个IO消耗一个令牌；SmartQoS策略配置类型为带宽，一个sector对应一个令牌。
- 策略配置的上限值越大，对应令牌桶拥有的令牌数就越多，对应LUN就能拥有更高的性能。
- 系统内部应用程序I/O队列处理机制：
  1. 应用服务器下发I/O请求后，存储系统将I/O下发到相应的被控对象的I/O队列中。
  2. 存储系统根据被控对象的优先级调整I/O队列所持有的令牌数，通过降低优先级低的被控对象的令牌数，保证优先级高的被控对象可占有的系统资源，优先保证高优先级的被控对象的运行（系统内部为I/O出队操作）。
  3. 处理出队的被控对象的I/O请求。

## 突发流控管理

- 配置突发时必须同时配置上限，其含义为：系统性能瞬时值可达到突发值，而性能平均值不超过上限值。
- ①阶段的业务压力小于上限目标，系统累积了突发能力。
- ②阶段业务压力增大，性能可以突破上限目标达到突发目标，总的突发时长为①阶段累积突发时长和QoS策略配置突发时长的较小值。
- 一旦结束突发到了③阶段，业务性能就会下跌至上限目标以下。



31 Huawei Confidential

HUAWEI

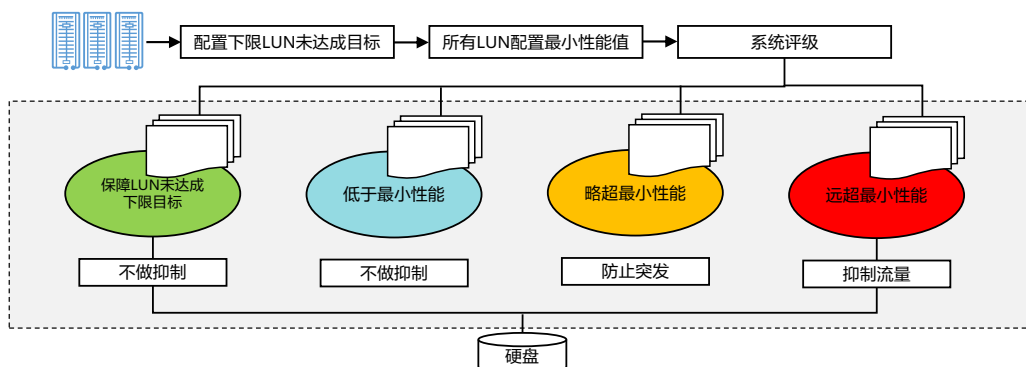
- 详细原理：

- LUN的burst能力来自于性能累积的burst信用，所谓burst信用是指LUN的实际性能小于上限流控值产生的性能累积。
- QoS框架使用双令牌桶模式实现burst信用的增加和消耗。C令牌桶用于达成LUN的上限目标，E令牌桶用于达成LUN的burst能力。
- 如果LUN的业务压力小于上限配置值，即C令牌桶的令牌消耗速度小于令牌补充速度，C令牌桶中令牌逐渐累积，超过上限配置值则溢出。溢出的令牌被当做burst信用存入E令牌桶，E令牌桶中令牌超过E桶大小令牌被丢弃，E令牌桶中有剩余令牌则具有burst能力。
- 如果LUN的业务压力大于上限配置值，C令牌桶中令牌逐渐会被消耗光。之后会从E令牌桶中获取令牌，实现burst，直到E令牌桶中令牌消耗光。



## 下限保障

- 当系统多种业务抢占资源，少量关键业务在资源抢占中不占优势时，下限保障功能可自动抑制非关键业务，为关键业务预留资源。



- 详细原理：

- 对所有LUN配置下限目标：用户对少量关键业务的LUN配置下限目标，系统对其余LUN依据系统标称性能的10%（可更改）配置共享下限目标，时延目标换算为IOPS目标。
- 对系统所有LUN进行负载评级，有下限保障LUN未达成目标，则触发系统评级、调整流程。
- 对于性能远超其下限目标的LUN进行流量抑制，释放资源。

## 支持多维度流控

- SmartQoS策略提供了多维度的流程，最大程序简化用户操作：
  - 支持对LUN进行流控
  - 支持对LUN组进行流控
  - 支持对Host进行流控
  - 支持对文件系统进行流控
- 说明：
  - 各策略仅可添加一种对象。
  - 当策略内添加LUN组时，仅可添加一个LUN组。
  - 一个LUN、LUN组、Host、文件系统仅可添加进一个策略。
  - 一个LUN可以添加进不同的LUN组，不同的LUN组可以添加进不同策略。QoS有单独处理这种情况，LUN最终的流控值会全局取小，确保所有策略达成。
  - Host策略不支持下限保障功能。

## 分层管理

- 分层策略：依用户/租户/子公司维度将系统能力进行划分，防止不同维度间扰邻。  
分层策略下可添加多个普通策略。各子策略性能总和将满足分层策略要求。
- 普通策略：单个用户/租户/子公司内的不同业务配置流控策略。
- 说明：
  - 分层策略具有普通策略的所有特性，且可以添加普通策略作为子策略。
  - 添加有LUN、LUN组的子策略可添加进同一分层策略。
  - LUN/LUN组的策略配置、Host的策略配置、文件系统的策略配置均两两互斥，不可添加进同一策略（包括分层、普通）中。
  - 下限保障策略不可添加进分层策略。

## 策略分配

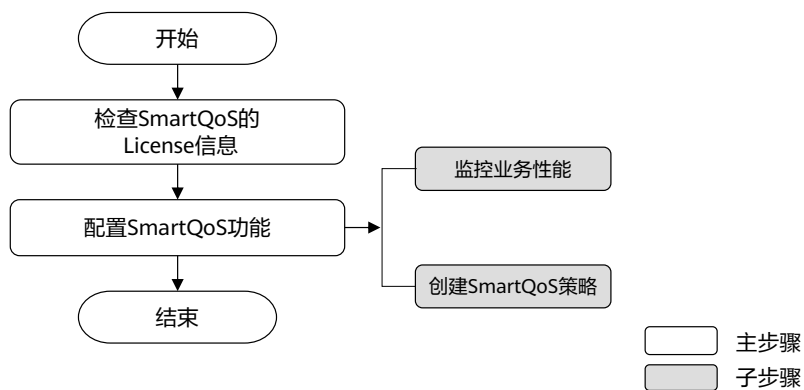
- SmartQoS策略配置的上限流控目标为共享型，即添加到策略中的所有对象共享此策略的流控目标。SmartQoS模块会定期采集流量控制策略中所有子对象的性能数据及需求数据，然后通过分配算法，将流量控制策略总体的控制目标分配给策略中具体的对象。
- 目前采用的分配算法为max-min权重分配算法，分配算法总体流程如右图。



## SmartQoS应用场景

- 上限流控策略：
  - 适用于多种业务混合的应用场景，避免多业务之间的扰邻现象。对于系统中的一些影响较大，但重要性较低的业务进行最大性能限制，避免其对系统中其它业务造成冲击。
  - 存在明确性能限制目标的场景，比如在线业务和备份业务混合的场景，通过限制备份业务的最大流量，从而在确保备份时间窗的同时，减小对于在线业务的性能影响。
- 突发流控：
  - 对于时延极度敏感的业务，如数据库业务、虚拟机部署、虚拟机启动。
- 下限保障策略：
  - 针对最核心的少量关键业务，确保服务质量（ IOPS、BPS、时延 ）。

## 配置流程



- 当性能作为一个重要因素需要被一个应用程序或多个共用一个存储系统的应用程序考虑时，需要了解应用程序在存储系统中的真实性能。存储系统提供的业务性能监控功能可以帮助用户确定应用程序的性能瓶颈。

# 目录

---

1. SmartThin
2. SmartTier&SmartCache
3. SmartAcceleration
4. SmartQoS
- 5. SmartDedupe&SmartCompression**
6. SmartVirtualization
7. SmartMigration

## 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。

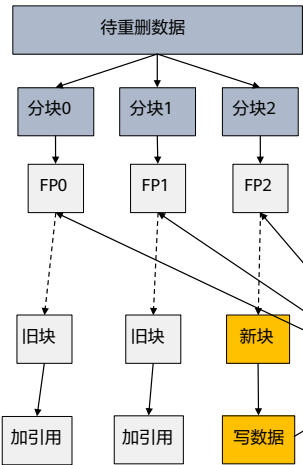


# 压缩

- **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。

# 在线重删

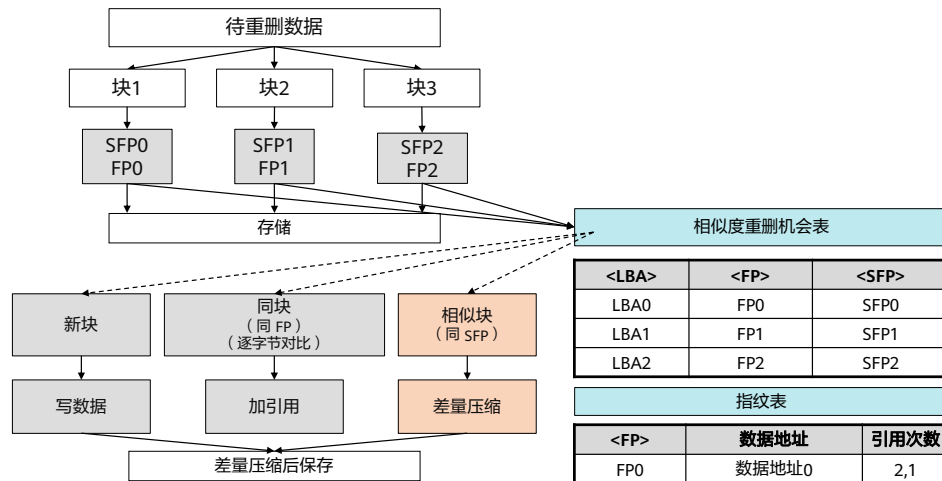
- 1、对数据进行分块。
- 2、对分块数据计算指纹。
- 3、到指纹库查找指纹是否存在。
- 4、存在即为旧块，不存在即为新块。



- 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。

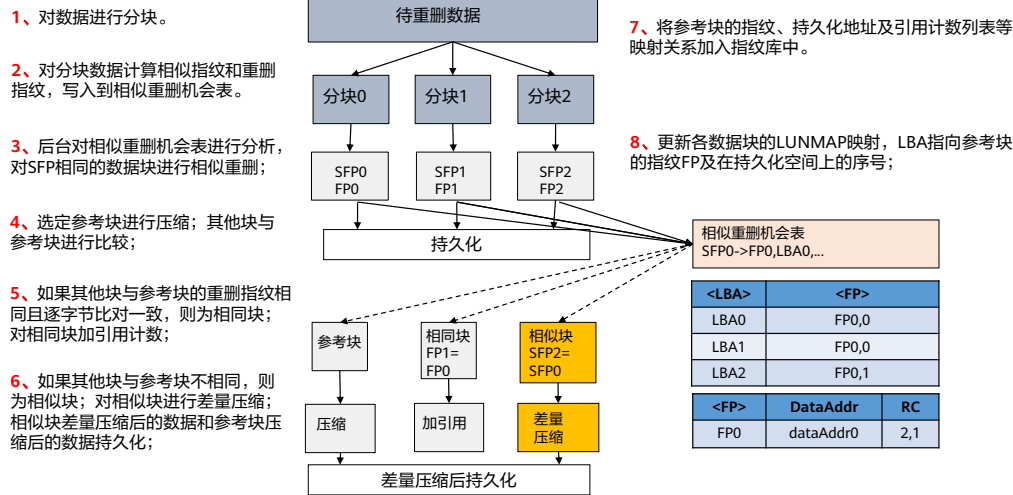
| 指纹表     |           |      |
|---------|-----------|------|
| FP      | DataAddr  | rc   |
| FP0=FPx | dataAddr0 | +1   |
| FP1=FPy | dataAddr1 | +1   |
| FP2     | dataAddr2 | 1    |
| FPx     | dataAddrx | 1->2 |
| Fpy     | dataAddry | 1->2 |

## 后处理相似重删原理



- 存储系统对新写入的数据进行分块，默认数据块大小为8KB。
- 存储系统利用相似指纹算法计算新写入数据块的相似指纹信息。
- 写入数据块到硬盘，并将数据块的指纹信息和位置信息写入到机会表。
- 存储系统周期性的检查机会表中是否存在相似的指纹信息
  - 是 => 执行2
  - 否 => 继续周期性检查
- 通过逐字节比较结果确认相似的数据块是否相同。
  - 相同 => 删除该数据块，并在指纹库中将该数据块的指纹信息及存储位置映射关系指向原有数据块的记录。
  - 相似 => 存储系将数据块进行增量压缩，记录指纹信息到指纹表，更新指纹信息到数据块元数据中，并回收这些数据块的空间。

# 后台相似重删



# 压缩原理

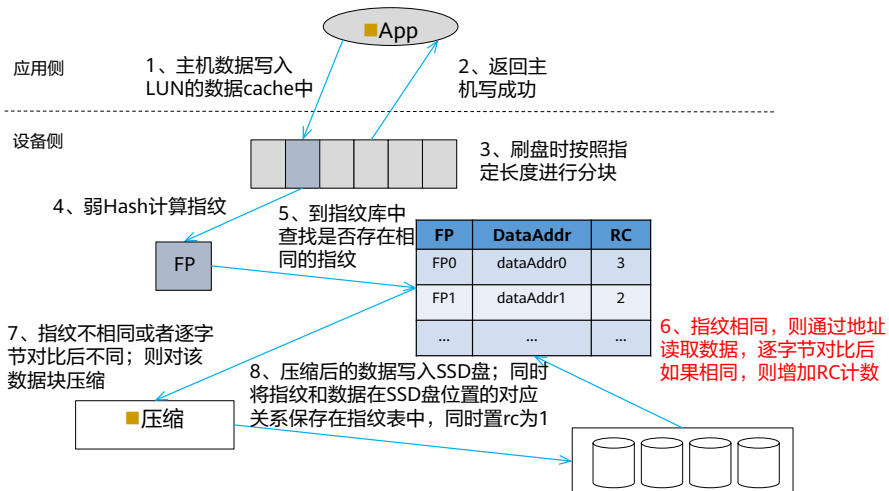
LZ77示例    abcdbbccaaabaeaaabae

- 假设有一个要压缩的串ab~~cd~~bbccaaabaeaaabaee，前面10个字符已经被压缩，并且设置滑动窗口（字典）为10个字符。

| 窗口                      | 未压缩         | 相同字符串   | 相同字符串起始位置 | 相同字符串长度 | 相同字符串后一字符 | 压缩码     |
|-------------------------|-------------|---------|-----------|---------|-----------|---------|
| ab <del>cd</del> bbccaa | abaeaaabaee | ab      | 0         | 2       | a         | (0,2,a) |
| dbbccaaba               | eaabaee     | null    | 0         | 0       | e         | (0,0,e) |
| bbccaabaee              | aaabaee     | aaabaee | 4         | 6       | e         | (4,6,e) |

- 最后输出的压缩结果为(0,2,a)(0,0,e)(4,6,e)

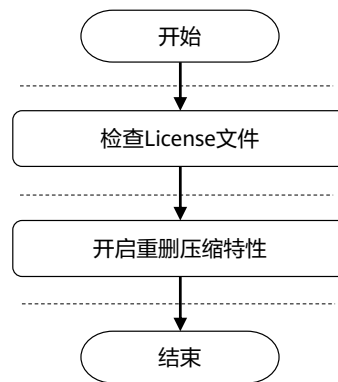
## 重删压缩流程



## 重删压缩特性应用场景

- 重复数据删除和数据压缩两种功能结合使用，高效地实现存储空间节省。
- 场景包括：
  - VDI/VSI场景。
  - 数据测试或开发系统。
  - 文件服务系统。
  - 工程数据系统。

## 配置流程





# 目录

---

1. SmartThin
2. SmartTier&SmartCache
3. SmartAcceleration
4. SmartQoS
5. SmartDedupe&SmartCompression
- 6. SmartVirtualization**
7. SmartMigration

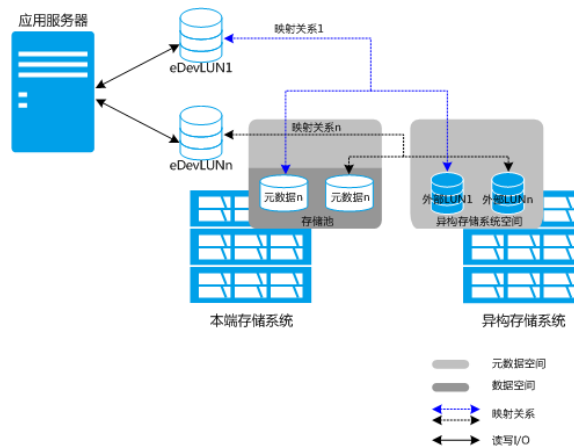
## SmartVirtualization相关概念

- 思考：SmartVirtualization特性允许外部LUN为OceanStor Dorado 系列存储系统提供物理存储空间，如何实现呢？
- 实现过程中提到的“本端存储系统”、“异构存储系统”、“外部LUN”、“eDevLUN”、“在线接管”、“离线接管”、“托管”分别是什么意思呢？

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

## eDevLUN与外部LUN的关系

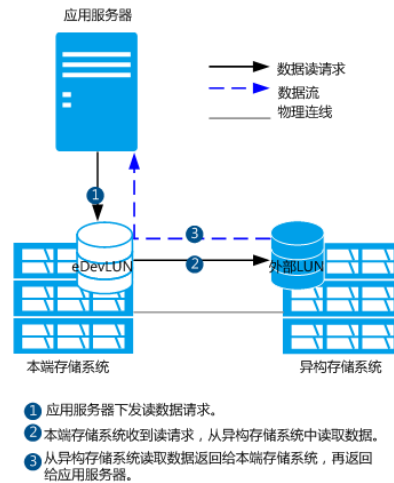
- eDevLUN由实际数据和元数据两部分组成，通过映射关系建立元数据和实际数据的对应关系。



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

## 数据读流程

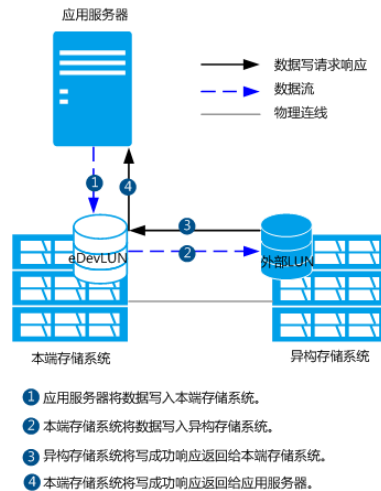
- 使用SmartVirtualization特性对异构存储系统中的外部LUN进行托管后，当应用服务器下发读取外部LUN数据请求时，本端存储系统的eDevLUN收到数据读取请求后从异构存储系统的外部LUN中读取。



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

## 数据写流程

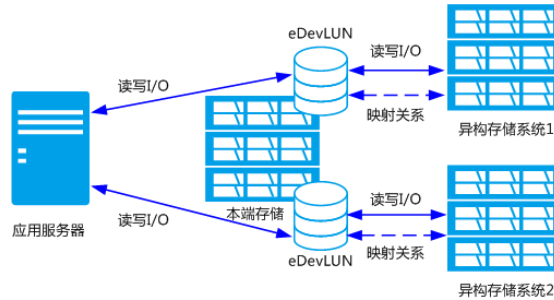
- 使用SmartVirtualization特性对异构存储系统进行托管后，数据写流程如下：



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

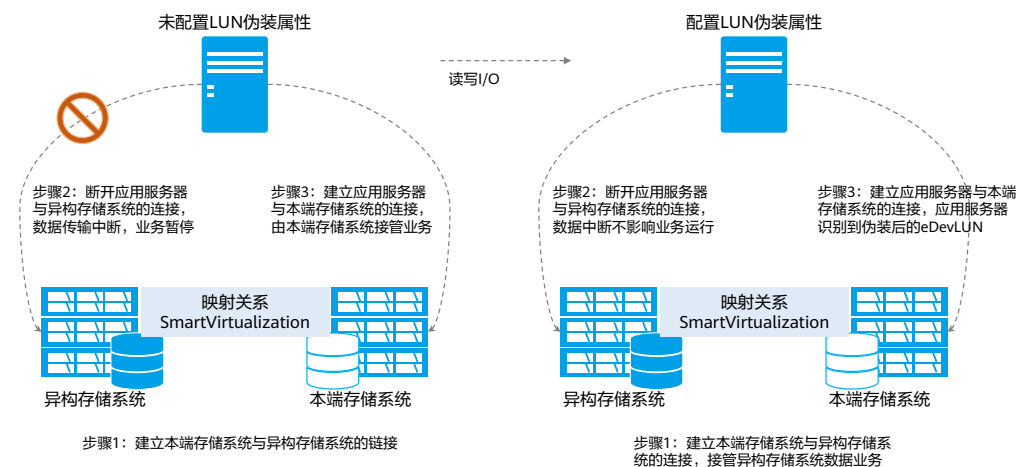
## 存储资源集中管理

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



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

## 离线接管和在线接管



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

## 业务数据迁移至新购存储系统

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

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



## 冷数据迁移至原有存储系统

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

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

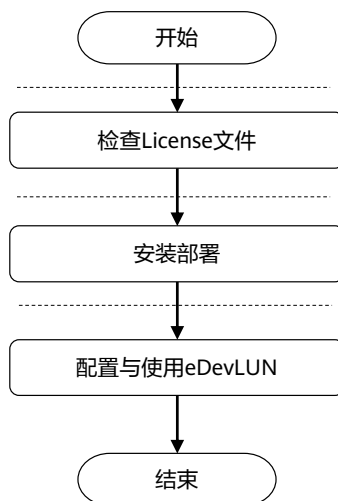
## 接管方式的选择

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

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

- 离线接管方式适用于所有兼容的其他华为异构存储系统及第三方厂商的异构存储系统。离线接管需要暂停应用服务器业务，eDevLUN的伪装属性为“不伪装”。
- 在线接管华为异构存储系统时，eDevLUN的伪装属性为“基本伪装”或“扩展伪装”。基本伪装和扩展伪装的选择与选择的多路径厂商及版本、其他华为异构存储系统的版本有关，具体请根据版本查看相关产品文档。
- 有关版本、兼容性的问题需要在实际部署前查看相关文档或兼容性列表，本章主要是讲解SmartVirtualization的理论。

## 配置流程



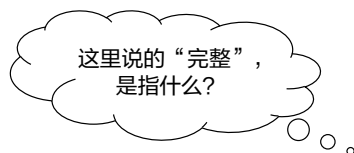
# 目录

---

1. SmartThin
2. SmartTier&SmartCache
3. SmartAcceleration
4. SmartQoS
5. SmartDedupe&SmartCompression
6. SmartVirtualization
- 7. SmartMigration**

## 概述

- SmartMigration是实现业务迁移的关键技术，可以在不中断主机业务的情况下实现源LUN上的业务完整地迁移到目标LUN上，并在复制结束后使目标LUN完全替代源LUN来承载业务。



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

## SmartMigration原理

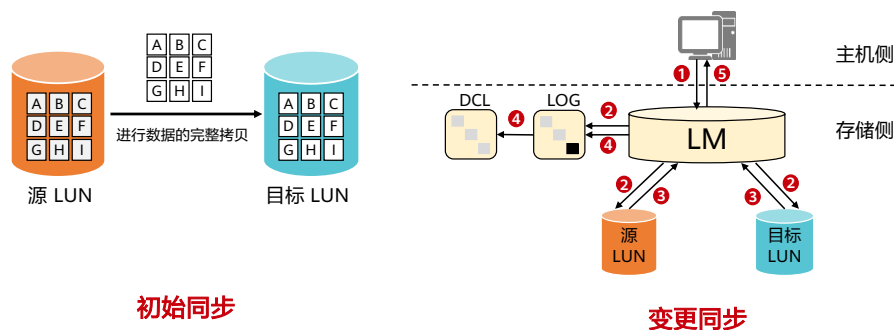
- SmartMigration特性的主要用途是通过对LUN上的数据进行迁移，从而调整业务性能或实现设备升级换代以满足用户需求。
- SmartMigration 的实现过程分为两个阶段：



- 存储系统采用虚拟化存储技术。存储池中的虚拟化数据由元数据卷和数据卷两部分组成。
  - 元数据卷：记录数据的存储位置，其中包含LUN ID和数据卷 ID。LUN ID用以识别 LUN，数据卷ID用以识别数据卷的物理空间。
  - 数据卷：存放实际的用户数据。

## SmartMigration业务数据同步

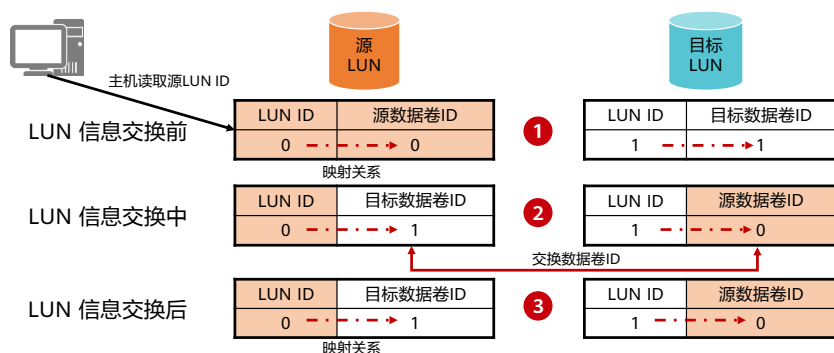
- 创建 SmartMigration，源LUN和目标LUN之间建立Pair关系。
- 源LUN和目标LUN之间的业务数据同步包括：初始同步和变更同步。



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

## SmartMigrationLUN信息交换

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

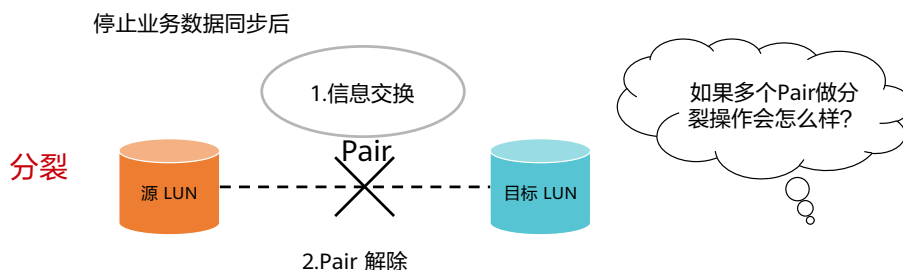


- LUN信息交换是业务信息同步后目标LUN能够顺利地代替源LUN来承载业务的前提。
- 存储系统中，每个LUN和对应的数据卷都有属于自己的唯一标识，分别为LUN ID和数据卷ID。其中，LUN是逻辑上的概念，数据卷是物理上的概念，两者之间形成一一对应的映射关系。
- LUN信息交换前：主机通过源LUN ID识别源 LUN。LUN ID和数据卷ID形成映射关系；
- LUN信息交换中：交换源数据卷ID和目标数据卷ID。源LUN指向的物理空间变为目标数据卷；
- LUN信息交换后：源LUN ID未发生变化，主机读写不受影响（用户无感知）。源LUN ID和目标数据卷ID形成新的映射关系。主机实际读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



## SmartMigration分裂

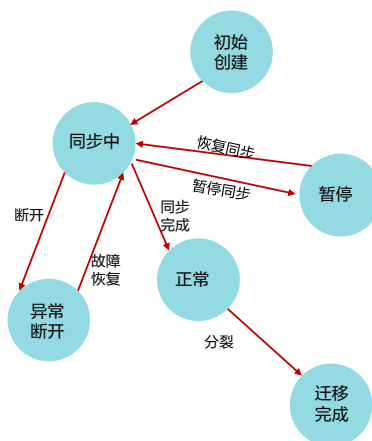
- 分裂是针对单个Pair的操作，它的过程包括构成Pair的单个源LUN和目标LUN停止业务数据同步开始进行LUN信息交换，并在LUN信息交换后解除数据迁移关系。



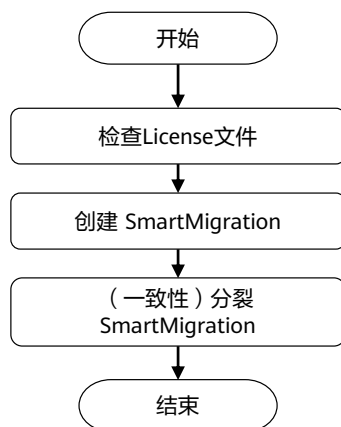
- 分裂时，主机业务被悬挂，信息交换后，业务下发到迁移后的LUN，从而实现用户无感知情况下的业务迁移。
- Pair解除：指源LUN和目标 LUN，在信息交换后解除数据迁移关系。
  - Pair解除后，如果主机侧下发I/O请求至存储侧，数据只写入源LUN（源LUN ID指向的物理空间为目标数据卷），不再进行双写。
  - 目标LUN将封存Pair解除时刻源LUN的所有数据。
  - Pair 解除后不能再重新建立联系。
- SmartMigration的一致性分裂是指多个Pair同时进行LUN信息交换，并在信息交换完成后同时解除Pair关系，保证分裂前后各时刻点数据的一致性。
- 在某些有多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

## SmartMigration状态转换

- **迁移中**：源LUN正在向目标LUN同步数据。
- **正常**：源LUN与目标LUN的数据成功同步完成。
- **暂停**：该pair对处于等待队列。
- **异常断开**：SmartMigration遇到IO错误导致源LUN和目标LUN之间的复制关系断开。
- **迁移完成**：源LUN与目标LUN的数据成功同步完成，并且完成分裂。

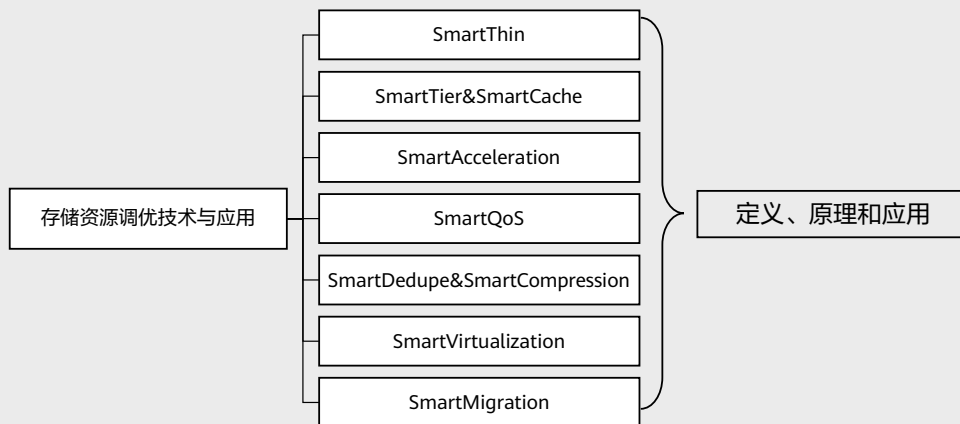


## 配置流程



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

## 本章总结



## 思考题

1. （判断题）单一硬盘类型组成的存储池不可以启用 SmartTier。
2. （选择题）可对LUN设置哪些迁移策略？
  - A. 自动迁移
  - B. 向高性能层迁移
  - C. 向低性能层迁移
  - D. 不迁移

- 参考答案：

- T
- ABCD

## 思考题

3. （单选题）LUN迁移执行一致性分裂前必须是什么状态？

- A. 迁移中
- B. 暂停
- C. 正常
- D. 迁移完成

• 参考答案：

- D

## 学习推荐

- 华为官方网站
  - 企业业务: <http://enterprise.huawei.com/cn/>
  - 技术支持: <http://support.huawei.com/enterprise/>
  - 在线学习: <http://learning.huawei.com/cn/>
- 热门工具
  - HedEx Lite
  - 网络资料工具中心
  - 信息查询助手

- 热门工具:
  - HedEx Lite: 华为产品文档管理工具, 支持浏览、搜索、更新和管理产品资料
  - eStor: 图形化存储仿真平台, 该平台通过对真实华为OceanStor全闪存仿真模拟, 帮助广大ICT从业者和用户快速熟悉华为存储系列产品, 了解并掌握相关产品的操作和配置。
  - 网络资料工具中心: 汇集网络产品的资料工具, 是您进行投标支撑、网络规划、项目交付和升级维护的好助手。
  - 信息查询助手: 提供华为产品命令和告警信息查询。

# 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©2022 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.





## 存储数据保护技术与应用



## 前言

- 传统的数据保护解决方案专注在对数据的周期性备份上，因此一直伴随有备份窗口、数据一致性以及对生产系统的影响等问题。
- 本章节主要介绍一些如HyperSnap、HyperClone、HyperCDP和HyperLock等存储数据保护技术。

# 目标

- 学完本课程后，您将能够了解以下特性的原理、配置方法与应用场景：
  - HyperSnap
  - HyperClone
  - HyperCDP
  - HyperLock

# 目录

---

1. **HyperSnap**
2. HyperClone
3. HyperCDP
4. HyperLock

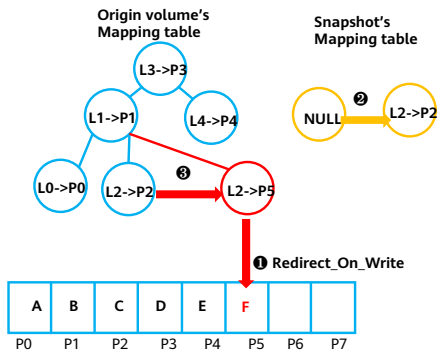
## 概述



- HyperSnap：  
又叫快照特性，是指源数据在某个时间点的一致性数据副本，是对指定数据集的一个完全可用的拷贝，该拷贝包含源数据在拷贝时间点的静态映像。
- 快照的实现机制包括**写前拷贝**（COW，Copy-on-write）技术和**写时重定向**（ROW，Redirect-on-write）技术：
  - COW机制在初次写过程中，需要完成数据的拷贝，数据拷贝过程会影响主机写性能。
  - ROW机制虽然不会导致数据拷贝，但是频繁覆盖写后，会破坏源LUN的数据分布，源LUN数据分布离散影响主机顺序读性能。

- 快照的目的：
  - 备份归档：快照可以作为备份和归档的数据源；
  - 快速恢复：快照可以对存储设备上的数据灵活且频繁地生成多个恢复时间点，在需要时可以快速地恢复数据；
  - 瞬间生成：可以瞬间生成，不影响主机业务，得到源LUN在某个时间点的数据副本。

# ROW原理



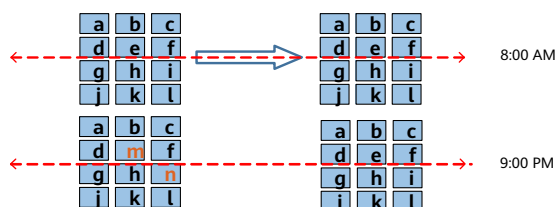
## LUN快照功能原理

- **定义：**

- 快照是指源数据在某个时间点的一致性数据副本。快照生成后可以被主机读取，也可以作为某个时间点的数据备份。

- **主要特点：**

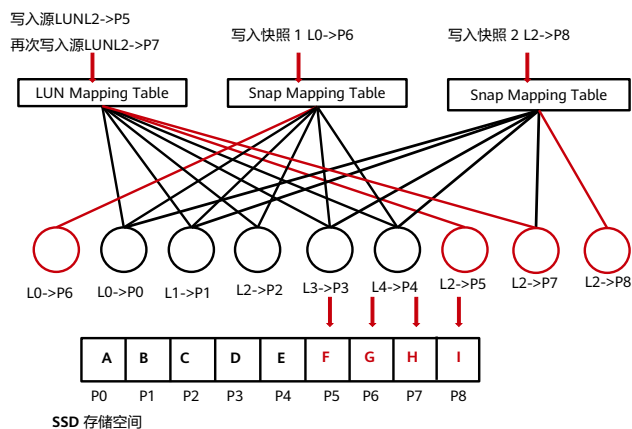
- 瞬间生成：存储系统可以在几秒钟内生成一个快照，获取源数据的一致性副本。
- 占用存储空间少：生成的快照数据并非完整的物理数据拷贝，不会占用大量存储空间。所以即使源数据量很大，也只会占用很少的存储空间。



- **快照常用术语：**

- 源卷（source volume）：需要进行快照操作的源数据所在的卷，对用户而言表现形式为 LUN。
- 快照卷（snapshot volume）：对源卷创建快照后，逻辑上生成的数据副本。对用户而言表现形式为快照 LUN。
- 重定向写（redirect on write）：重定向写，覆盖写的场景也需要重新分配空间，写成功后释放原有的空间。
- 快照回滚（snapshot rollback）：将快照LUN的数据复制给源 LUN，从而使源 LUN的数据恢复成激活快照LUN时刻的数据。
- 未激活状态（inactive）：快照的一种状态，在这种状态下快照不可用，需要进行激活操作后可用。

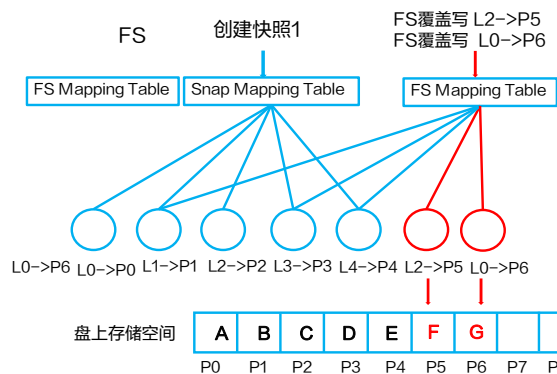
## LUN快照功能原理 - 性能无损



- 写入源LUN L2 的请求直接申请新空间 P5 写入；原来的空间 P2 被快照引用。
- 写入快照 1 L0 的请求直接申请空间 P6 写入，没有额外的读写开销。
- 再次写入源LUN L2 的请求直接申请新空间 P7 写入；原来的空间 P5 由于没有快照引用，直接释放。
- 再创建并激活一个快照2。



## FS快照功能原理 - 性能无损

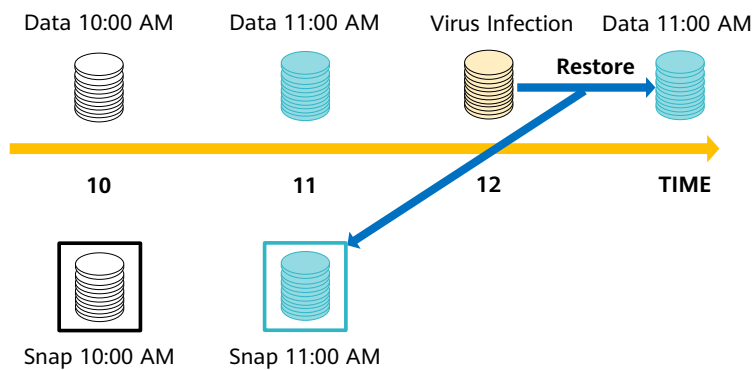


快照不影响源FS读写性能

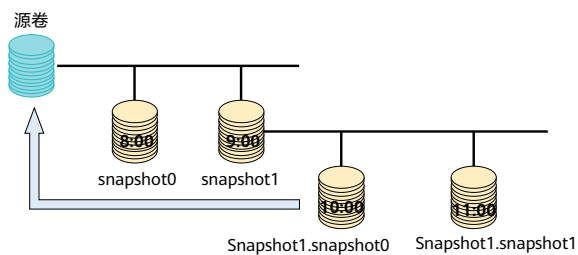
快照读性能与源FS读性能相当  
(文件系统快照只读)

①写入源FS L0/L2的请求直接申请新空间  
P5/ P6写入；原来的P0 P2被快照引用；

## 快照功能原理 - 回滚

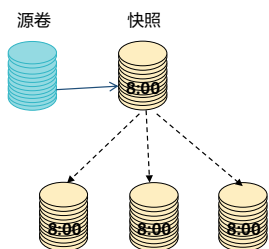


## 快照功能原理 - 级联快照与跨级回滚



- **级联快照：**即快照的快照，与快照副本不同的地方在于，级联快照包含快照写过的数据，其他功能与普通快照没有差别。
- **跨级回滚：**是指相同源卷的快照之间包括源卷，可以进行回滚没有层级约束。

## 快照功能关键技术 - 副本



如何得到**多份**相同快照的数据？



因为快照是**虚拟**的，因此可以很快的对快照进行复制。



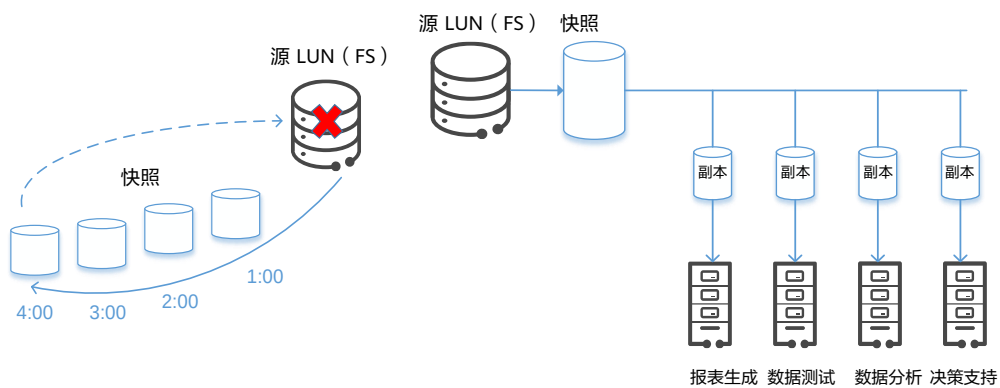
## 快照功能关键技术 - 写前回滚



如何能够**瞬时**完成数据恢复?

回滚过程中, 主机对源LUN**进行写时**, 快照会先将快照中的数据块拷贝到源LUN后, 主机IO再继续写。  
**没有主机读写时**, 快照中的数据是依次向源卷回滚。

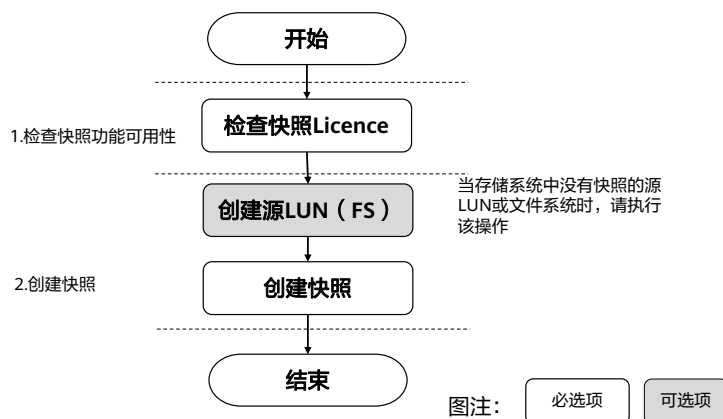
## 应用场景



- 数据持续保护场景

- 数据备份与恢复场景

## 配置流程



- 快照配置术语：
  - 激活：快照的一种状态，在这种状态下快照可用。
  - 未激活：快照的一种状态，在这种状态下快照不可用。
  - 重建快照：重建快照命令是取消激活快照和激活快照命令的组合。
  - 快照一致性组：当同一类业务需要同时使用多个LUN创建快照，以保持所有快照在时间上的一致性，您需要对业务归属的保护组创建快照一致性组，以便日后统一维护。
- 更多信息请查找华为数据存储信息中心：<http://support-it.huawei.com/storage/#/home>

# 目录

---

1. HyperSnap
- 2. HyperClone**
3. HyperCDP
4. HyperLock

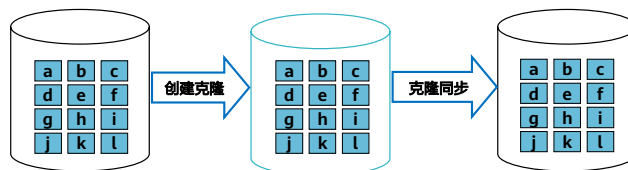


## HyperClone概述

- 定义
  - HyperClone可以创建特定时间点(启同步时刻)的源LUN的全量的数据副本 (目标LUN)
  - 文件系统克隆特性是基于文件系统或文件系统某一个时刻的快照进行创建，克隆文件系统创建完成后，该文件系统中的数据和父文件系统对应时刻点保持一致（包括父文件系统对应时刻点的Dtree配置及Dtree数据）。
- 特点
  - 目标LUN可立即读写，不需要等待后台拷贝完成；
  - 支持全量同步和差异同步；
  - 支持正向同步和反向同步；
  - 支持一致性组。

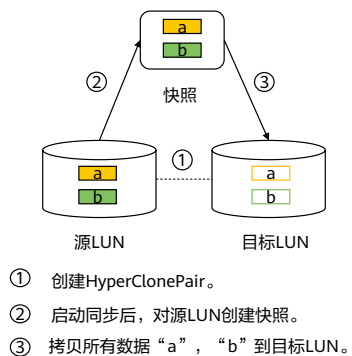
## 克隆功能原理

- **定义：**克隆是指源数据在某个时间点的一致性数据副本，数据同步完成后成为完整的数据副本。克隆生成后可以被主机读取，也可以作为某个时间点的数据备份。
- **主要特点：**
  - **瞬间生成：**存储系统可以在几秒钟内生成一个克隆，获取源数据的一致性副本，创建后的克隆支持立即读写，创建的克隆支持选择不同的重删压缩属性。
  - **在线分裂：**支持在不中断业务的场景下进行克隆分裂，分裂读写不影响源LUN的IO处理。

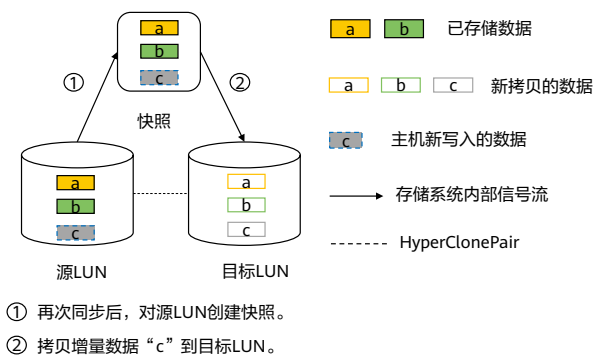


## 功能原理 - 同步

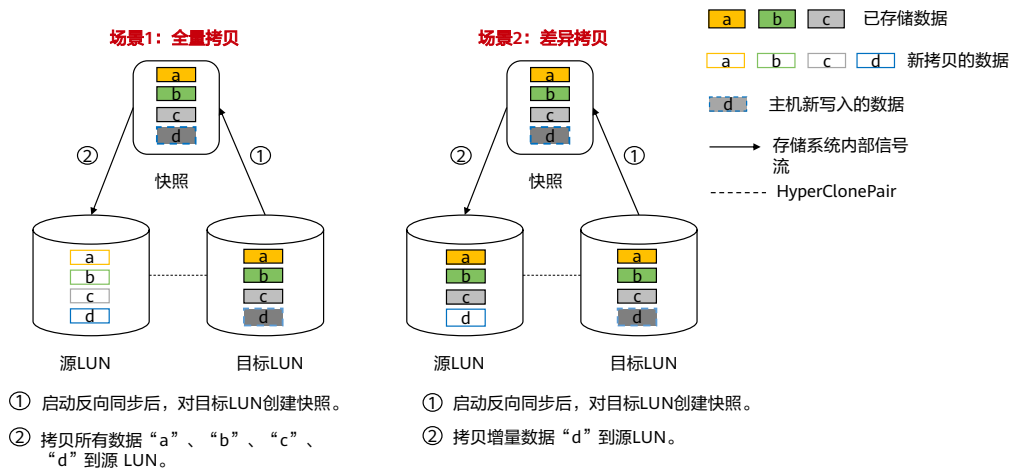
场景1：首次同步、全量拷贝



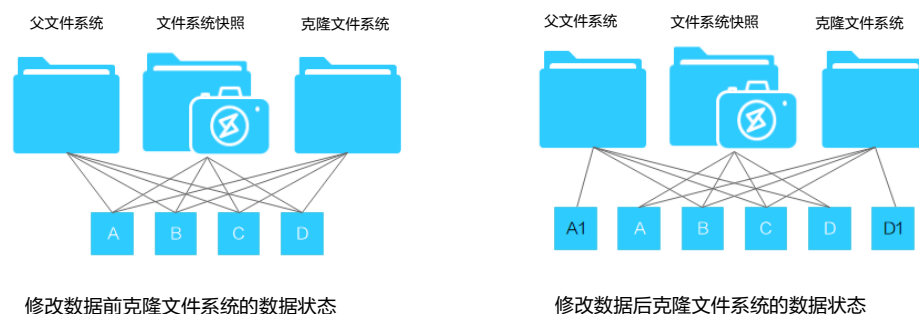
场景2：首次同步后再次同步、差异拷贝



## 功能原理 - 反向同步

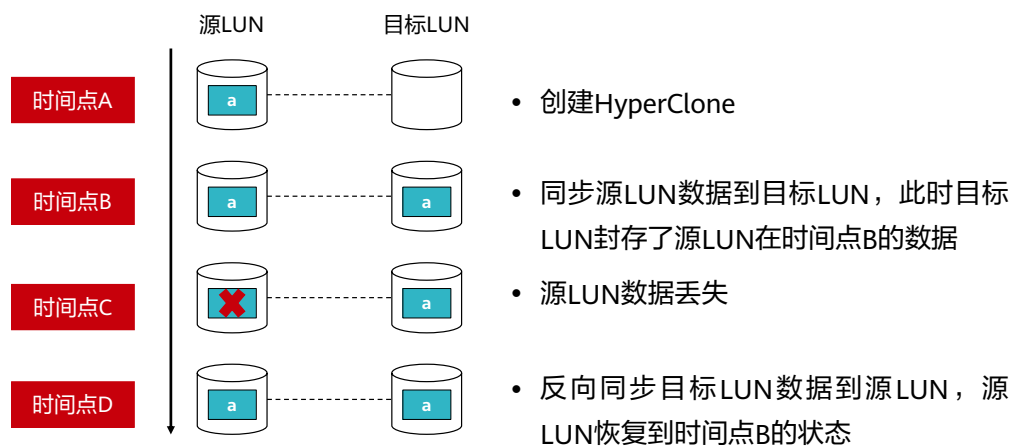


## 克隆文件系统读写原理



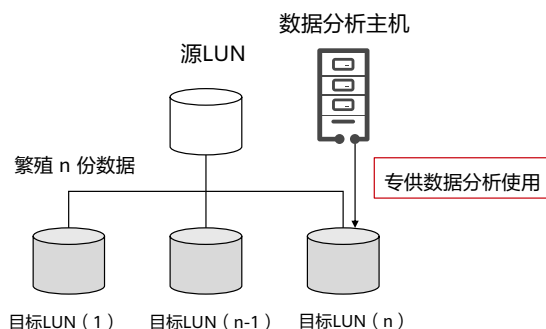
- 初始创建时，数据未被修改前，克隆文件系统与父文件系统共享源数据，通过文件系统快照保证在克隆时间点的数据一致性。克隆文件系统共享给应用服务器进行读写，此时读出的数据是父文件系统的源数据。
- 当应用服务器对父文件系统或者克隆文件系统已有数据的数据块写入新数据时，由于快照ROW技术的保护，会将新数据写入新分配的存储空间，不会覆盖原有数据。修改父文件系统A数据块时，存储池会新分配一个A1数据块用于存储新数据，A数据块不释放；同样，修改克隆文件系统D数据块时，存储池也会新分配一个D1数据块用于存储新数据，D数据块不释放。此时，文件系统快照数据保持不变。

## 应用场景 - 数据备份与恢复



- 将源数据内容复制到一个或多个副本中，可实现对源数据的时间点备份，当源LUN数据破坏时，可以从目标LUN恢复。使用目标LUN可以实现在线数据备份和方便快捷的数据恢复。

## 应用场景 - 数据分析与繁殖



### • 数据分析

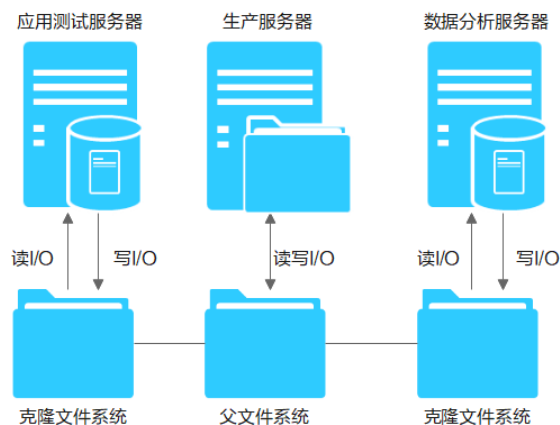
- 数据分析业务使用目标LUN中的数据，避免了数据分析业务和生产业务抢占源LUN资源，影响性能。

### • 数据繁殖

- 即HyperClone可以从同一源LUN添加多个目标LUN拷贝多份数据。

- 数据分析为了提取有用信息、形成结论、提供决策支持而对大量数据进行详细研究和概括总结的过程。在数据分析场景下，数据分析业务使用目标LUN中的数据，避免了数据分析业务和生产业务抢占源LUN资源，影响性能。
- 数据克隆 即HyperClone可以从同一源LUN添加多个目标LUN拷贝多份数据。

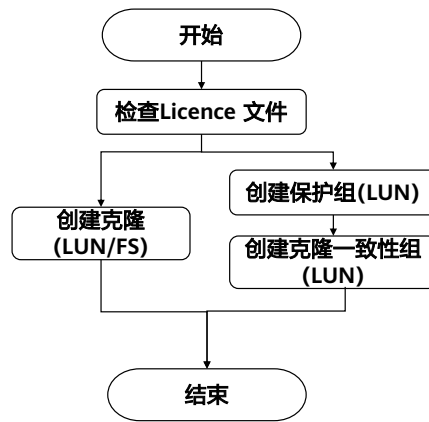
## 应用场景 – 应用测试与数据分析



- 应用测试通常是指对软件系统进行功能、性能、安全性和可用性等方面的测试。数据分析是对大量数据进行研究，提取有用信息，为决策提供支撑。应用测试和数据分析都涉及对大量数据进行读写，从克隆文件系统中读取数据，既能保证数据在克隆时间点与父文件系统的一致性，又避免了测试和分析业务对生产系统造成影响，可以并行开展各项业务活动。



## 配置流程



# 目录

---

1. HyperSnap
2. HyperClone
- 3. HyperCDP**
  - **LUN HyperCDP**
  - FS HyperCDP
4. HyperLock

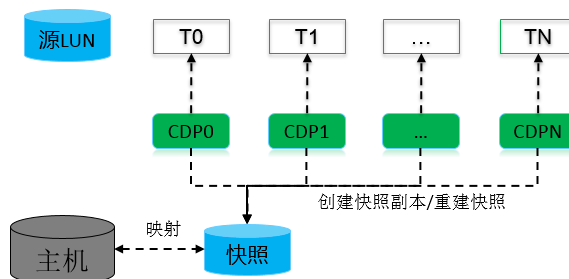
## HyperCDP（块业务）

- HyperCDP是一种高密快照技术，可为用户提供持续的数据保护功能。HyperCDP生成的快照与普通快照类似，是指源数据在某个时间点的一致性数据副本，是对指定数据集合的一个完全可用的拷贝，该拷贝包含源数据在拷贝时间点的静态映像。
- 主要特点包括：
  - 数据保护密集、持久。单LUN支持6万个HyperCDP，最短定时间隔支持3秒。
  - 秒级保护，性能无损，占用存储空间少。
  - 内置定时计划功能，支持间隔、天、周、月4种定时策略。
  - 支持HyperCDP一致性组。
  - 不能直接映射主机读写，可通过创建副本转换为可写快照映射给主机。

- HyperCDP基于无损快照技术，创建HyperCDP对源LUN性能几乎没有影响，与可写快照相比，HyperCDP无需构建LUN对象，极大减少了系统内存开销，从而实现比可写快照更密集、更持久的保护。
- HyperCDP是一项增值特性，需要购买License才能使用。
- 在License文件中，HyperCDP的“特性名称”显示为“HyperCDP”。
- 当有HyperCDP的license时，无需快照license也可以进行快照的所有操作，包括创建、激活、取消激活、删除、重建、创建副本、回滚、停止回滚等。
- HyperCDP具有以下优势：
  - 秒级保护，性能无损，占用存储空间少。
  - 支持定时计划功能。存储系统内置定时计划功能，支持间隔、天、周、月4种定时策略，满足用户远疏近密的备份诉求。
  - 数据保护密集、持久。单个LUN支持的HyperCDP规格更大，能为存储设备上的数据生成更多更密集的恢复点，可以提供更小的保护间隔，更长的保护周期，从而为用户提供持续的数据保护。
- 目的与收益
  - 节省存储空间，保护用户投资
  - 生成HyperCDP副本，可供多种应用使用
  - 持续数据保护

## HyperCDP功能原理

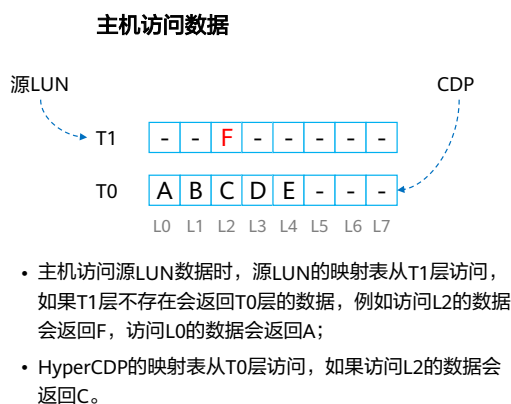
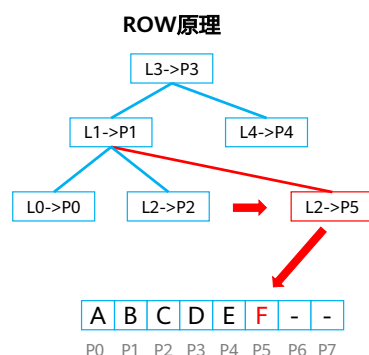
- HyperCDP基于无损快照技术，创建HyperCDP对源LUN性能几乎没有影响，HyperCDP无需构建LUN对象，极大减少了系统内存开销，从而实现比可写快照更密集、更持久的保护。
- HyperCDP不支持直接映射主机，当需要读取某个HyperCDP的数据时，可以对HyperCDP创建快照副本，再将快照副本映射给主机。



- HyperCDP是一种高密快照特性，可为用户提供持续的数据保护功能。HyperCDP基于无损快照技术，创建HyperCDP对源LUN性能几乎没有影响，与可写快照相比，HyperCDP无需构建LUN对象，极大减少了系统内存开销，从而实现比可写快照更密集、更持久的保护。
- 系统内置HyperCDP定时计划，通过配置不同的策略满足客户不同的备份诉求。HyperCDP不支持直接映射主机，当需要读取某个HyperCDP的数据时，可以对HyperCDP创建快照副本，再将快照副本映射给主机。

## 创建HyperCDP

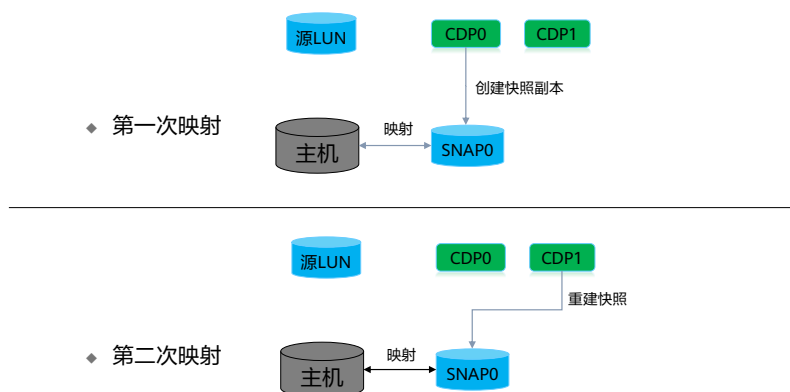
- 创建HyperCDP的过程就是保存源LUN在激活时间点的数据状态的过程。



- 创建HyperCDP的过程就是保存源LUN在激活时间点的数据状态的过程，原理如下：OceanStor Dorado全闪存存储系统的普通LUN都采用了ROW（redirect on write）的读写方式：每次主机写入新数据，系统会重新分配空间存储新数据，同时更新LUN的映射表指向新的数据空间（图中的L0~L4为逻辑地址，P0~P7为物理地址，A~F为具体的数据）；
- ROW
  - ROW是指写时重定向技术，是实现HyperCDP的核心技术。当系统收到写数据请求需要修改源数据时，存储系统在存储池中为新写入的数据指定新的存储位置，并将被修改数据块的指针指向新的存储位置。
- 基于ROW的原理，在T0时刻在源LUN上创建HyperCDP后的主机读写过程如下图所示（图中T0、T1为具体的时刻，A~F为具体的数据，L0~L7为逻辑地址），其中F为T1时刻主机写入L2的数据：主机访问源LUN数据时，源LUN的映射表从T1层访问，如果T1层不存在会返回T0层的数据，例如访问L2的数据会返回F，访问L0的数据会返回A；HyperCDP的映射表从T0层访问，如果访问L2的数据会返回C；

## HyperCDP读写

- HyperCDP不能直接映射主机读写，需提升为可写快照后映射至主机。

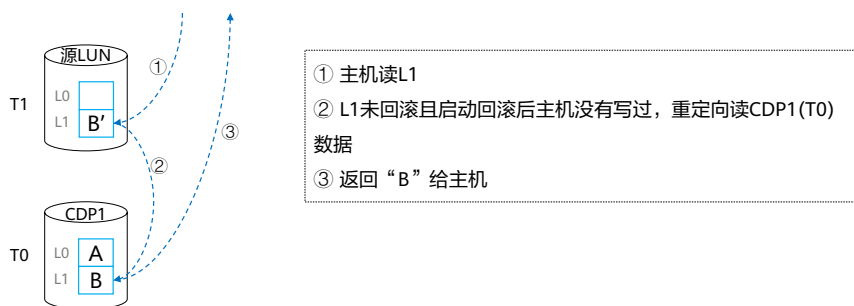


- HyperCDP不能直接映射主机读写，当用户需要读取某个HyperCDP的数据时，可以对HyperCDP创建快照副本（创建出来的快照副本与HyperCDP数据相同），再将快照副本映射给主机。如下图所示，对CDP0创建快照副本SNAP0并映射给主机，快照SNAP0的数据与CDP0的数据完全相同。
- 后续如果想读取该LUN的另一个时刻的HyperCDP的数据，可基于另一时刻的HyperCDP重建快照，重建后的快照立即获得另一时刻的HyperCDP的数据。如下图所示，用户想读取CDP1的数据，则可以基于CDP1重建快照SNAP0（由于之前SNAP0已经映射，因此此处无需再映射），重建后SNAP0的数据与CDP1的数据完全相同。

## 回滚HyperCDP (1)

- 回滚HyperCDP的过程是将HyperCDP的数据拷贝至源LUN的过程。启动回滚后，源LUN支持立即使用（源LUN的数据即为HyperCDP的数据）。

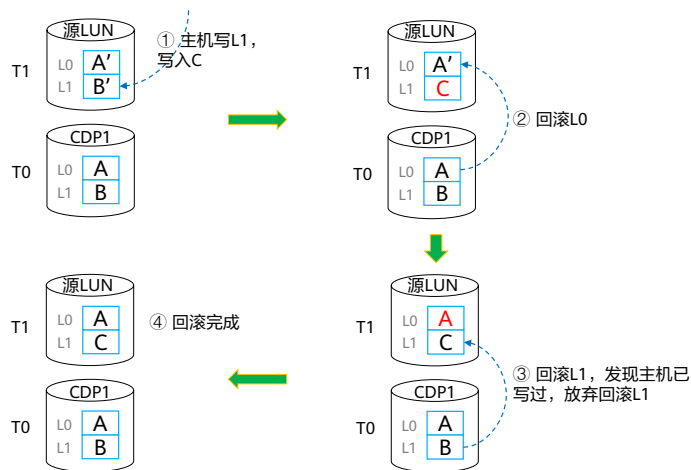
CDP回滚时读源LUN，执行读重定向



- 读源LUN，需要执行读重定向；读重定向的流程如下图所示，CDP1回滚数据至源LUN，主机来读取源LUN地址为L1的数据，如果此地址还未被回滚且启动回滚后该位置未被主机写过，需要返回CDP1对应L1地址的数据；

## 回滚HyperCDP (2)

### CDP回滚时写源LUN

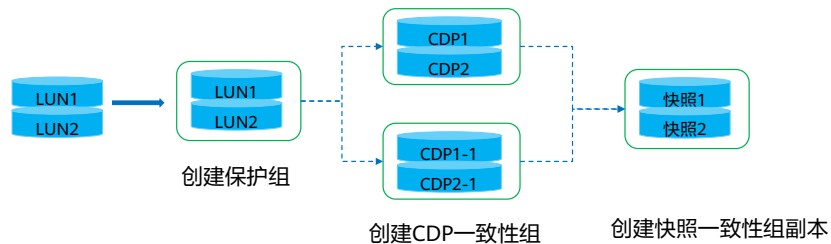


- 写源LUN，直接写入主机数据，回滚后台拷贝时若发现该位置主机已写过，则不再回滚。
- 停止回滚操作目前不支持将源LUN的数据恢复至回滚前的状态，所以建议在启动回滚前对源LUN创建HyperCDP做数据保护。
- 一旦执行了停止回滚命令，源LUN上已经回滚的数据不会自动回退，停止回滚命令只会停止数据拷贝任务。



## HyperCDP一致性组

- 在大中型数据库应用中，数据、日志、修改信息等存储在磁盘阵列的不同LUN中，缺少其中一个LUN的数据，都将导致其他LUN中的数据失效，无法继续使用。通过HyperCDP一致性组功能，可实现在数据恢复时应用数据的一致性。
- 与单个HyperCDP一样，HyperCDP一致性组支持创建、删除、回滚、停止回滚操作，支持对HyperCDP一致性组创建快照一致性组副本，支持基于HyperCDP一致性组重建快照一致性组。



## 定时HyperCDP计划

- 定时HyperCDP计划的内容包括间隔、天、周、月4种策略，每个策略可以设置不同的保留个数，间隔、天、周、月策略可以复选。
- 同一个定时HyperCDP计划支持添加多个LUN和LUN一致性组，同一个LUN或LUN一致性组只能添加至一个定时HyperCDP计划。
- 定时HyperCDP计划最小的间隔周期为3秒，最大保留个数为60000。

- 同一个计划如果选择了多个策略，策略的执行时间可能会重叠，在重叠的时间点会给每个策略分别创建一个HyperCDP。例如计划配置了间隔（间隔周期为10s）、天（周日凌晨00:00）2种策略，则在周日凌晨00:00时定时计划会创建2个HyperCDP。

## 存储池容量阈值

- 由于定时HyperCDP计划创建HyperCDP的最小间隔为3 s，持续写源LUN可能会产生大量的数据保护容量导致存储池容量耗尽，进而影响源LUN的主机业务。
- 系统支持对存储池的已使用容量和保护容量分别设置容量阈值，通过定时数据保护自动删除开关控制容量阈值是否生效。

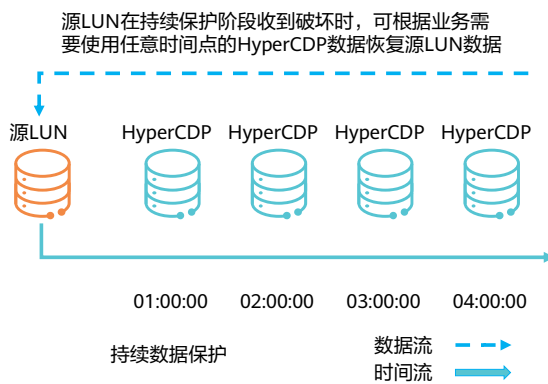


- 容量阈值开关默认关闭，建议用户配置HyperCDP时将容量阈值开关打开。
- 容量阈值功能只删除定时HyperCDP计划创建的HyperCDP，不会删除用户手动创建的HyperCDP。

## 应用场景 - 快速数据备份和恢复

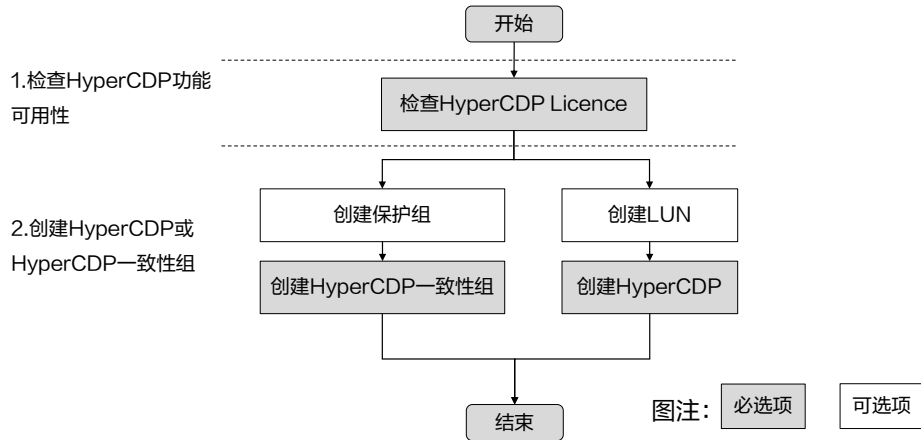
- 使用HyperCDP备份可以在以下场景中迅速备份和恢复数据：

- 病毒感染
- 人为误操作
- 恶意篡改
- 系统宕机造成的数据损坏
- 应用程序BUG造成的数据损坏
- 存储系统BUG造成的数据损坏



- 利用HyperCDP进行数据备份和恢复有以下几个优势：
  - 显著降低数据恢复目标时间（RTO），即便是在回滚数据较多的情况下，也可以在数秒内让应用感知到数据已经恢复完毕。
  - 能够实现频繁和不停机的数据备份操作，同时应用程序运行不受影响。
  - 显著减少或消除备份窗口的时间要求。

## 配置流程



# 目录

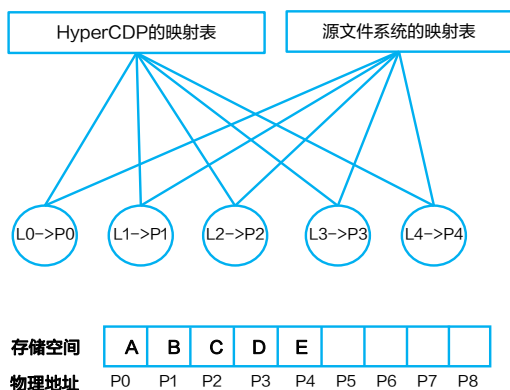
---

1. HyperSnap
2. HyperClone
- 3. HyperCDP**
  - LUN HyperCDP
  - **FS HyperCDP**
4. HyperLock

## HyperCDP（文件业务）

- HyperCDP是一种高密快照特性，可为用户提供持续的数据保护功能。HyperCDP基于无损快照技术，创建HyperCDP对源文件系统性能几乎没有影响，与可写快照相比，HyperCDP无需构建文件系统对象，极大减少了系统内存开销，从而实现比可写快照更密集、更持久的保护。
- OceanStor系统内置HyperCDP定时计划，通过配置不同的策略满足客户不同的备份诉求。HyperCDP不支持直接映射主机，当需要读取某个HyperCDP的数据时，需要把文件系统共享给主机，从而读取HyperCDP创建时间点的源文件系统数据。

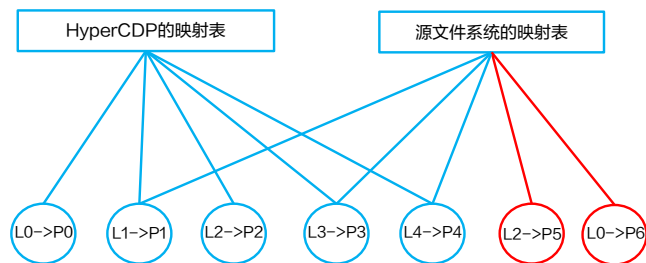
## 创建HyperCDP副本



- 创建HyperCDP时，系统会为文件系统生成一个CDP副本，CDP副本创建成功后共享源文件系统的映射表，且CDP副本创建之后不影响主机对源文件系统的数据读写性能。且CDP副本是只读的。
- 源文件系统的初始数据为“ABCDE”，CDP副本创建成功后共享源文件系统的映射表。其中，L0~L4为逻辑地址，P0~P8为物理地址，A~E为具体的数据。



## 主机写源文件系统



存储空间  
物理地址

|    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|
| A  | B  | C  | D  | E  | F  | G  |    |    |
| P0 | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 |

- HyperCDP创建成功后，主机写源文件系统 L0和L2（写入数据"F"和"G"），由于采用ROW（写时重定向）机制，直接申请新的物理空间 P5和P6写入，原来的地址P0和P2被CDP引用。
- 整个源文件系统的写流程与创建HyperCDP前的源文件系统写流程完全一致没有额外的开销，保证了性能无损。

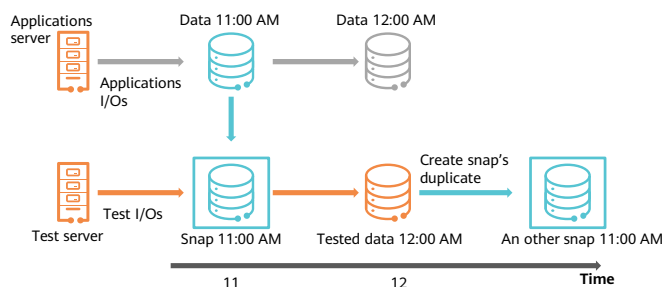
## HyperCDP的读原理

- 存储系统可以通过NFS和CIFS共享服务将文件系统共享给客户端。文件系统HyperCDP是文件系统定时只读快照，文件系统快照创建后，在源文件系统使用期间，客户端应用程序可以通过访问文件系统指定的共享路径（.snapshot目录）读取HyperCDP创建时间点的源文件系统数据。
- 定时快照访问支持下面3种方式：
  - 特定目录下（文件系统共享根目录和dtree目录），.snapshot目录可见，可通过进入.snapshot目录访问。
  - 可从任意目录进入.snapshot目录，显示保护了该目录的所有定时快照。
  - 任意目录可通过输入路径访问.snapshot，再进入到对应的快照中访问。访问定时快照时，若定时快照中不包含该目录信息，则访问失败。

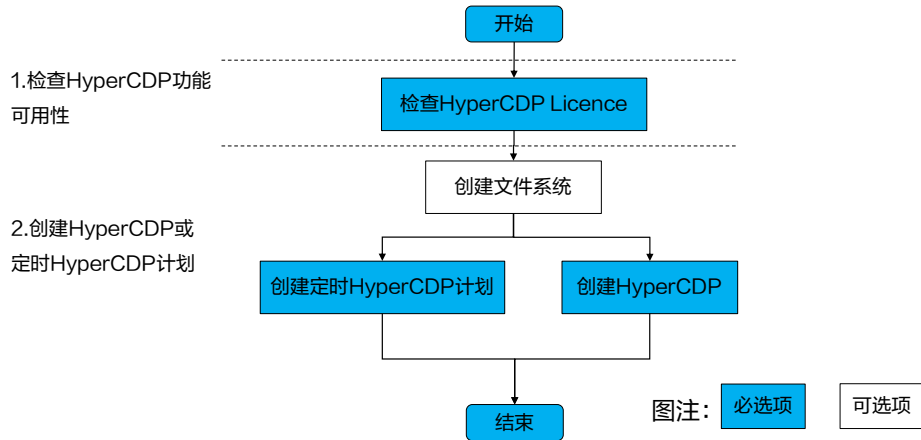
- 下面举例说明上面3个快照访问方式：
  - 假设文件系统根目录下有目录d1、d2，文件系统有定时快照s1、s2及s3，其中，s1和s2保护了d1，s3保护了d2，使用文件系统根目录创建共享。
  - 在共享的根路径可以看到.snapshot目录，进入.snapshot目录，可以看到s1、s2及s3三个定时快照，进入任意一个快照即可访问快照的数据。
  - 当前在如下路径 /d1，执行ls，看不到.snapshot目录，但是cd .snapshot目录会执行成功。进入.snapshot目录ls，只能查看到s1和s2(看不到s3)。
  - 执行cd /d2/.snapshot/s1/，会报错，因为s1未保护d2。执行cd /d2/.snapshot/s3/ 访问成功，因为s3保护了d2。
  - 通过读取定时快照数据，使用cp命令，可以从快照精准恢复某个文件和目录数据。

## 应用场景 - 数据挖掘和测试场景

- 对HyperCDP创建快照副本，对快照副本进行读写，可以达到数据挖掘和测试的目的，同时，不会对业务数据产生影响：
  - 对需要测试的数据生成某一个时间点（上午11点）的HyperCDP；
  - 使用测试服务器直接对HyperCDP创建快照副本，对快照副本进行测试（读写快照副本），此测试过程并不会对源数据和源数据上运行的业务造成任何影响；
  - 一个小时后，源数据和快照副本数据各自基于11点的数据演变为不同的两份数据；
  - 测试完成后，用户还可以通过对HyperCDP创建快照副本重新获取11点的数据，并使用新的副本重新进行测试。



## 配置流程



# 目录

---

1. HyperSnap
2. HyperClone
3. HyperCDP
- 4. HyperLock**

# 概述

- WORM特性又叫HyperLock特性，通过一次写入多次读取技术，保障原始数据在存储介质中的完整性、保密性及可存取性，满足日益增长的数据安全存储需求。
- WORM特性是指文件被写入完成后即可通过去掉文件的写权限，使其进入只读状态。在该状态下文件只能被读取，无法被删除、修改或重命名。通过配置WORM特性对存储数据进行保护后，可以防止其被意外篡改，满足企业或组织对重要业务数据安全存储的需求。
- 具有WORM特性的文件系统（以下简称WORM文件系统）只能由管理员进行设置。根据管理员权限不同，WORM文件系统可分为法规遵从模式（Regulatory Compliance WORM，简称 WORM-C）和企业遵从模式（Enterprise WORM，简称 WORM-E）。

| 模式   | 主要应用场景                     |
|------|----------------------------|
| 法规遵从 | 该模式主要应用于遵从法规施行数据保护机制的归档场景。 |
| 企业遵从 | 该模式主要应用于企业内部管理。            |

# WORM法规时钟

- 存储系统为防止由于系统时间被篡改而导致文件的保护期被修改提供了WORM法规时钟。WORM法规时钟包括全局安全法规时钟和WORM文件系统法规时钟。

| 时钟分类         | 作用范围                                         | 详细说明                                                   |
|--------------|----------------------------------------------|--------------------------------------------------------|
| 全局安全法规时钟     | 整个存储系统内部维护一个全局安全法规时钟，该时钟为所有WORM类型的文件系统提供时钟源。 | 在首次创建WORM文件系统时，系统管理员必须初始化全局安全法规时钟，初始化成功后不能再修改。         |
| WORM文件系统法规时钟 | 每个WORM文件系统内部维护一个法规时钟，文件的保护期以此时钟为基准。          | 在创建WORM文件系统时，系统会自动使用全局安全法规时钟初始化WORM文件系统法规时钟，无需用户手动初始化。 |

## 文件状态

- WORM文件系统中的文件被定义了四种状态，各状态的含义如表所示：

| 状态   | 详细说明                                        |
|------|---------------------------------------------|
| 初始状态 | 所有新建文件均处于该状态，此时文件是可读、可写状态，所有用户可以对文件进行任意改动。  |
| 锁定状态 | 处于该状态的文件，所有用户都不能修改、删除或重命名，只能读取和查看属性。        |
| 过期状态 | 处于该状态的文件，所有用户可以删除、读取及查看属性，但不可以修改或重命名。       |
| 追加状态 | 处于该状态的文件，所有用户只能以追加的方式向文件末尾添加内容，不能删除、截断或重命名。 |



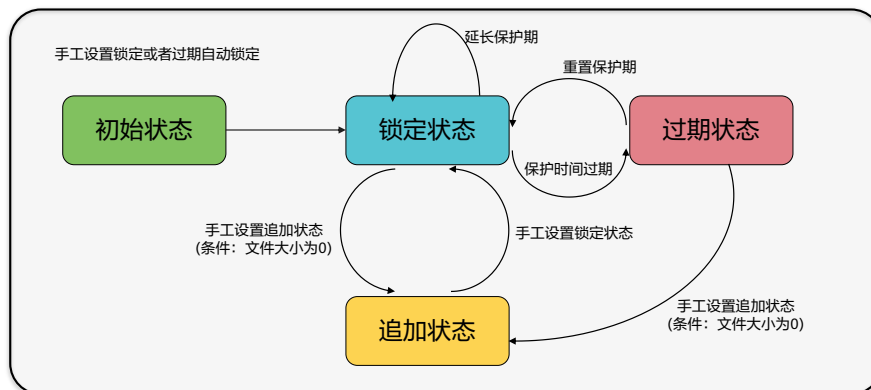
## WORM属性

- 当为文件系统配置了WORM特性后，文件系统具有WORM属性，这些属性参数作用于放入WORM文件系统上的文件。系统通过文件的WORM属性，判断该文件何时进入锁定状态，何时过期。文件系统的WORM属性如表所示。

| 属性     | 说明                                                                                     |
|--------|----------------------------------------------------------------------------------------|
| 模式     | 当前系统仅支持法规遵从模式。                                                                         |
| 最小保护期  | WORM文件系统支持的最短保护时间。为WORM文件系统中的各文件设置保护期时，文件的保护期最小不能小于WORM文件系统的最小保护期。                     |
| 最大保护期  | WORM文件系统支持的最长保护时间。为WORM文件系统中的各文件设置保护期时，文件的保护期最大不能大于WORM文件系统的最大保护期。                     |
| 默认保护期  | WORM文件系统支持的默认保护时间。当放入WORM文件系统中的文件未设置保护期时，文件的保护期为WORM文件系统的默认保护期。                        |
| 自动锁定   | 开启自动锁定功能后，放入WORM文件系统中的文件在结束修改超过一段特定时间（锁定等待时长）后自动进入锁定状态，修改包括修改文件数据和元数据两种类型的操作。          |
| 锁定等待时长 | 修改完成的文件自动进入保护状态的默认等待时间，“自动锁定”被启用的情况下该属性才有效。                                            |
| 自动删除   | 开启自动删除功能后，存储系统会自动删除已经过了保护期的文件，即自动删除过期文件。<br><b>说明：</b> 开启此操作前，请确认文件过期后不需要再保护，需要系统自动删除。 |

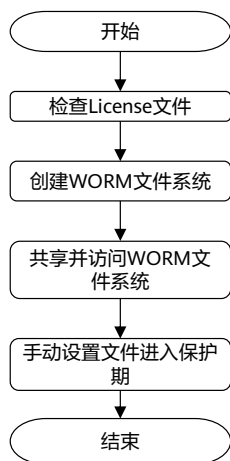
## 实现原理

- WORM特性是在普通文件系统的基础上增加了WORM属性，使WORM文件系统内的文件在保护期内只能被读取。创建WORM文件系统后，通过NFS或者CIFS协议映射给应用服务器。



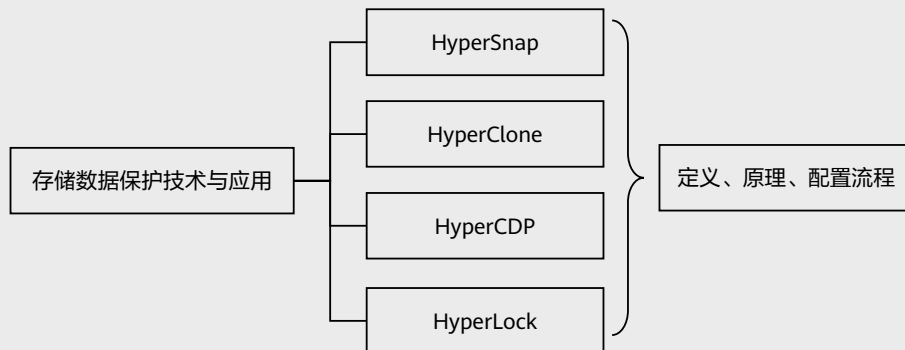
- 通过使用WORM特性，存在于WORM文件系统中的文件可以在初始状态、锁定状态、追加状态以及过期状态之间进行转换，从而防止重要数据在指定周期内被意外或恶意篡改。

## 配置流程



- License文件是使用增值特性的权限凭证。在配置增值特性前，请确认License文件中包含相关信息。
- 首次创建WORM文件系统时，需要初始化存储系统的全局WORM法规时钟。
- 通过创建WORM文件系统可以使文件系统以文件目录的形式共享存储资源。
- 创建完成WORM文件系统后，需要将文件系统共享给客户端，用户可以将需要保存的文件放入WORM文件系统中，实施防篡改保护。
- 创建WORM文件系统成功后，如果没有启用自动锁定，需要在客户端手动设置WORM文件系统中的文件进入锁定状态。同时可以设置文件由锁定状态进入追加状态，对文件进行内容追加。

## 本章总结



## 思考题

1. （判断题）同一个源LUN可以与不同的目标LUN组成多个不同的HyperClone Pair。一个目标LUN只能加入一个到HyperClone Pair。
2. （简答题）执行快照回滚命令后，什么时候主机可以对快照源LUN进行读写？

- 参考答案：

- T
- 启动回滚成功后就可以对快照源LUN进行读写

## 思考题

3. （单选题）以下哪种状态不是WORM文件系统中文件的状态（ ）
- A. 初始状态
  - B. 锁定状态
  - C. 删除状态
  - D. 过期状态

- 参考答案：

- C

## 学习推荐

- 华为官方网站
  - 企业业务: <http://enterprise.huawei.com/cn/>
  - 技术支持: <http://support.huawei.com/enterprise/>
  - 在线学习: <http://learning.huawei.com/cn/>
- 热门工具
  - HedEx Lite
  - 网络资料工具中心
  - 信息查询助手

- 热门工具:
  - HedEx Lite: 华为产品文档管理工具, 支持浏览、搜索、更新和管理产品资料
  - eStor: 图形化存储仿真平台, 该平台通过对真实华为OceanStor全闪存仿真模拟, 帮助广大ICT从业者和用户快速熟悉华为存储系列产品, 了解并掌握相关产品的操作和配置。
  - 网络资料工具中心: 汇集网络产品的资料工具, 是您进行投标支撑、网络规划、项目交付和升级维护的好助手。
  - 信息查询助手: 提供华为产品命令和告警信息查询。

# 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©2022 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

- 随着存储设备成本的下降，大容量存储设备已经逐渐被更多企业所应用，用于存储企业业务应用系统和IT系统产生的数据，比如电子邮件、文档、业务数据、数据备份等。因此对存储设备的有效管理关系到企业业务的持续性和稳定性。

# 目标

- 学完本课程后，您将能够：
  - 了解通用运维管理流程
  - 了解常用存储系统运维管理工具
  - 了解典型存储系统运维场景流程方法

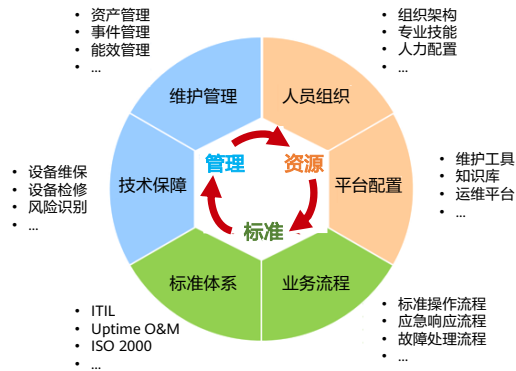
# 目录

---

1. 运维概述
2. 运维工具
3. 运维场景

## 什么是运维

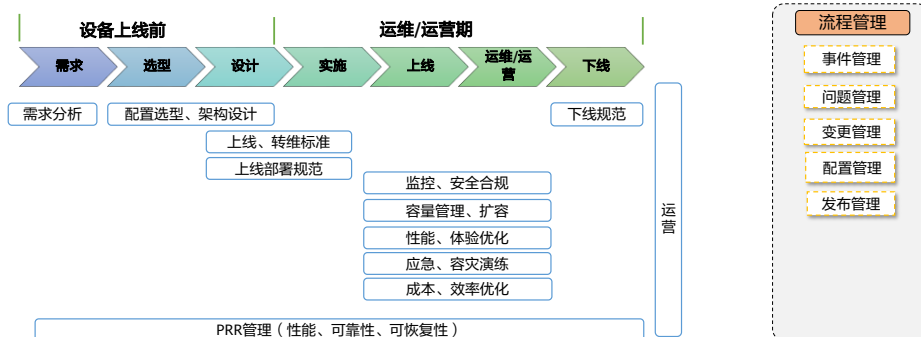
- 运维，本质上是对网络、服务器、服务的生命周期各个阶段的运营与维护，在成本、稳定性、效率上达成一致可接受的状态。



- 运维以技术为基础，通过技术保障产品提供更高质量的服务。运维是一个非常广泛的定义，在不同的公司、不同的阶段有着不同的职责与定位。

## 如何做好运维

- 技术层：梳理各产品运维生命周期的工作，识别各项工作的关键举措。
- 流程层（ITIL流程管理框架）：变更、事件、问题管理等。



# 事件管理



## 目标

- 尽快将服务恢复到正常状态
- 最小化突发事件对业务运作的负面影响
- 确保服务质量和可用性满足SLA指标



## 定义

- 突发事件
  - 会引起或可能引起服务中断、服务质量下降的任何事件
  - 硬件故障、软件故障、服务请求中断

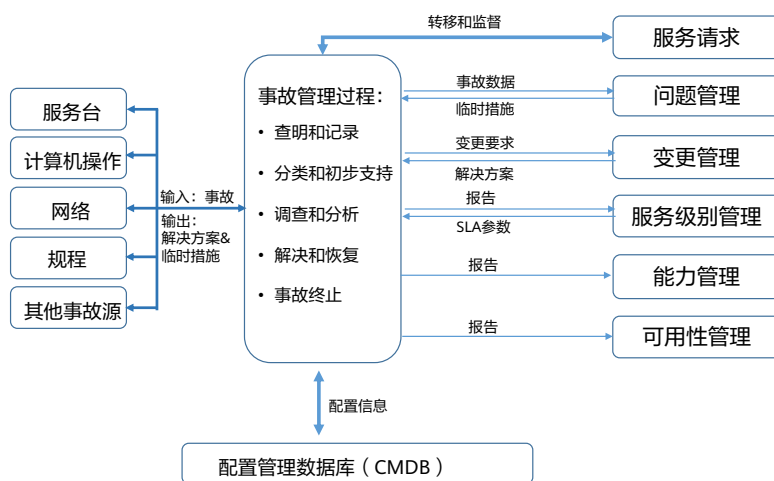


## 任务

- 探测和记录
- 分类和在线支持
- 确认优先级 (根据影响和紧急度)
- 调查和诊断
- 解决和恢复
- 结束
- 职责、监控、跟踪和交流

- SLA ( Service Level Agreement )：服务等级协议。

## 事件管理流程



- CMDB ( configuration management database ) : 配置管理数据库。



# 问题管理



## 目标

- 负责查找问题根源，并采取措施消除已知错误
- 最小化减少由于IT基础架构错误引起的突发事件数量，并减小问题的负面影响；防止与错误相关的突发事件的再次发生



## 定义

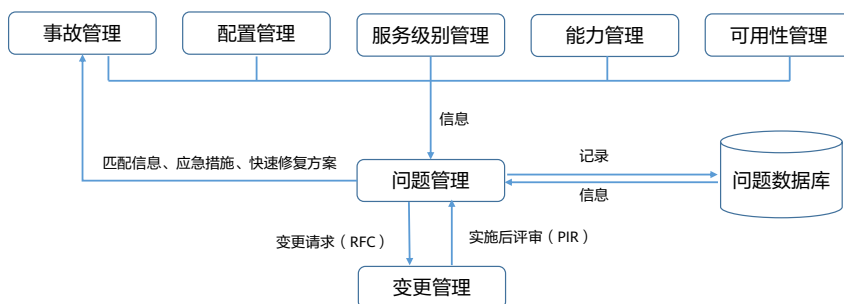
- 问题：从多个具有相同现象的突发事件或一个重大的突发事件获取的、表明存在某个未知原因的错误的情况
- 已知错误：已成功诊断问题的根源、找到解决方案的情况



## 任务

- 问题控制
- 已知错误控制
- 积极的问题管理
  - 趋势分析
  - 重大问题回顾

## 问题管理流程



- RFC ( Request for Change )：变更请求。
- PIR ( post implement review )：实施后评审。

# 变更管理



## 目标

- 确保通过标准化的手段和流程有效的控制和处理所有变更，以最小风险、高效的、高费效比的实施被批准的变更



## 定义

- 变更：导致一项或多项IT基础架构CI状态变化的一个行动
- 标准变更（事先经过审批）
- 变更请求（RFC）
- 变更下一步日程安排（FSC）
- 变更顾问委员会（CAB）



## 任务

- 受理、记录、批准、计划、测试、实施并回顾审视变更请求
- 提供IT基础设施的变更报告
- 驱动CMDB的修改

- CI（configuration item）：配置项。
- FSC（Forward Schedule of Changes）：变更进度计划表。
- CAB（Change Advisory Board）：变更顾问委员会。

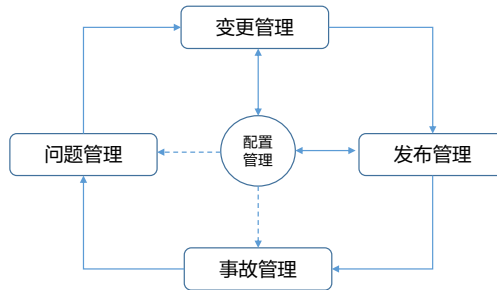
# 变更管理流程

输入信息包括：

- ❑ 变更请求
- ❑ 配置管理数据库提供的信息，特别是有关变更影响的信息
- ❑ 变更实施进度表
- ❑ 能力管理提供的数据库以及财务管理流程提供的预算信息等

输出的信息包括：

- ❑ 更新的变更实施进度表
- ❑ 触发配置管理和发布管理开始运作的信号
- ❑ 变更咨询委员会的议程、会议记录和行动项目
- ❑ 变更管理报告



# 配置管理



## 目标

- 计量组织和服务中所使用的所有IT资产和配置项的价值
- 为其它服务管理流程提供有关IT基础架构配置的准确信息
- 为事故管理、问题管理、变更管理和发布管理的运作提供支持
- 核实有关IT基础架构的配置记录的正确性并纠正发现的错误



## 定义

- 识别和定义配置项
- 规划、定义与管理配置管理数据库
- 定期验证CMDB的准确性和完整性
- IT资产的详细报告

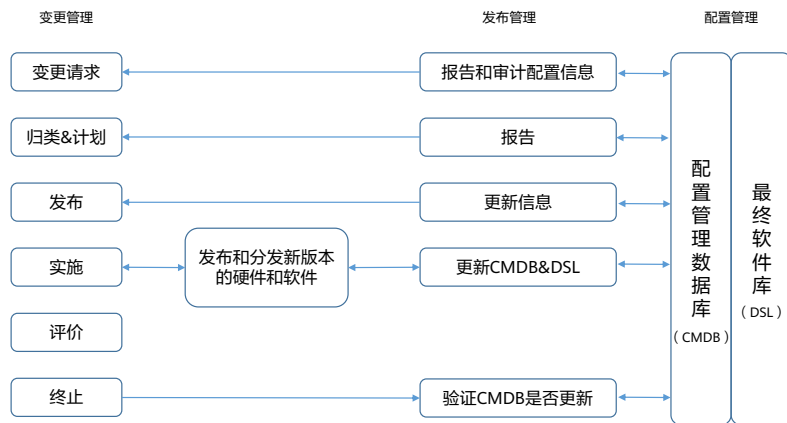


## 任务

- 受理、记录、批准、计划、测试、实施并回顾审视变更请求
- 提供IT基础设施的变更报告
- 驱动CMDB的修改

- 在配置管理中，最基本的信息单元是配置项（ Configuration Items， CIs ）。所有软件、硬件和各种文档，比如变更请求、服务、服务器、环境、设备、网络设施、台式机、移动设备、应用系统、协议、电信服务等都可以被称为配置项。
- 配置管理数据库不仅保存IT基础架构中特定组件的配置信息，而且还包括各配置项相互关系的信息。
- 配置管理不同于IT资产管理。后者是一个计量过程，用于控制和管理超过一定价值的资产的折旧过程；它记录了资产的购置时间、购买价格、折旧年限、折旧方法以及资产所处状态和位置等方面的情况。而配置管理除了记录配置项本身的信息外，还记录了各配置项之间的关系以及有关配置项的标准和授权方面的信息，同时它还记录了配置项的当前状态和变更情况。

# 配置管理流程



# 发布管理



## 目标

- 发布管理全面考核对IT服务的变更，确保综合考虑一项发布的各个方面，包括技术和非技术因素



## 定义

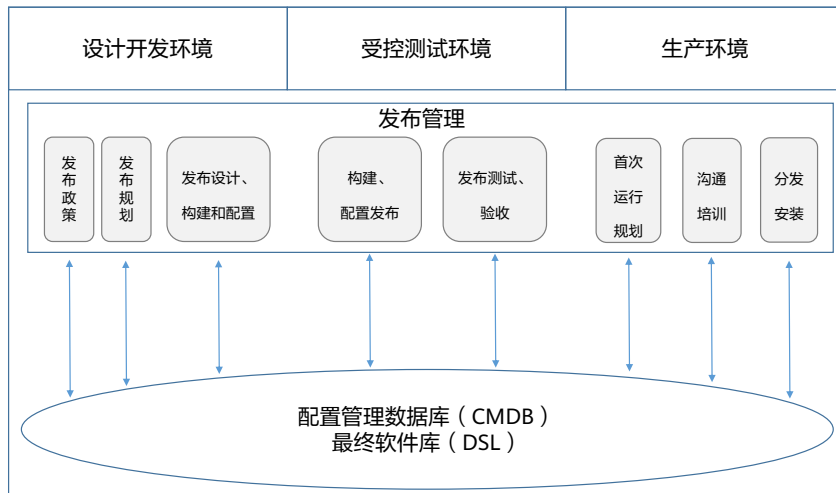
- 发布
  - 增量发布 ( Delta Release )
  - 完整发布 ( Full Release )
  - 包发布 ( Package Release )
- 紧急发布 ( Emergency Release )
- 发布策略 ( Release Policy )



## 任务

- 发布规划
- 设计、开发和配置一项发布
- 发布审核
- 上线计划
- 交流、准备和培训
- 分发和安装

# 发布管理流程



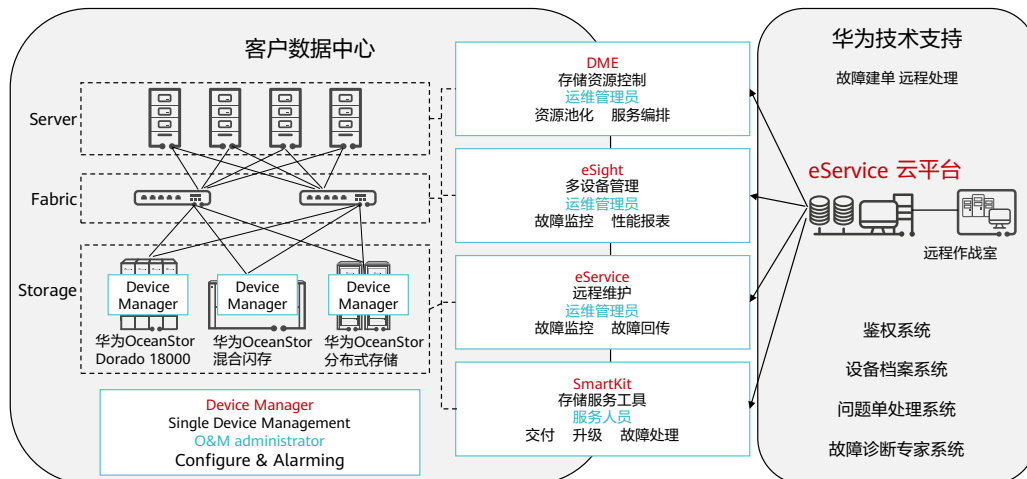


# 目录

---

1. 运维概述
- 2. 运维工具**
3. 运维场景

## 华为企业存储运维系统构成



- DeviceManager：单设备级的运维软件。
- SmartKit：面向华为技术支持工程师的专业工具，包括兼容性评估、规划设计、一键故障信息采集、巡检、升级、FRU更换等。
- eSight：面向客户的多设备维护套件，具有故障监控、可视化运维的特点。
- DME：面向客户的统一管理存储资源软件，编排服务目录，提供按需供给的存储服务和数据应用服务。
- eService客户端：部署在客户机房，实时发现存储设备异常，通知到华为维护中心。
- eService云平台：部署在华为维护中心，实时监控全网的设备，变被动维护为主动维护，甚至可以做到代维。

## 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



## SmartKit简介



### 1、统一平台

桌面工具管理平台，集成云计算、  
服务器、存储三大领域的运维工具



### 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

- SmartKit 是所有 IT 工具统一的桌面端管理平台，包含了对 IT 设备进行部署、维护、升级等操作所需的各类工具，能够帮助产品用户、服务工程师和维护工程师在上述过程中对设备进行精准操作，降低操作难度、提升工作效率。

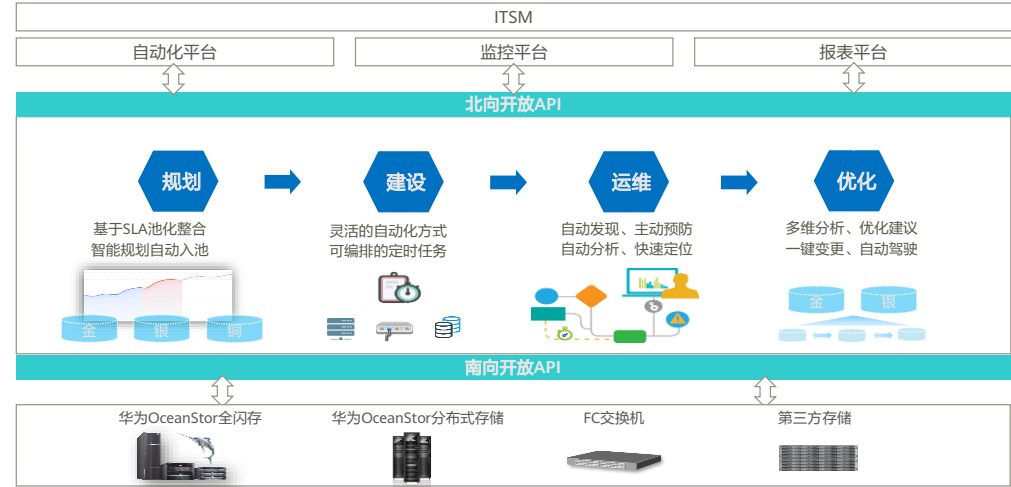
## SmartKit界面

- 各服务场景作业标准化、流程化，提升作业效率。



- 设备：所需要维护的IT设备管理入口。
  - 一次添加，多个工具及场景共享。
  - 批量导入/导出设备，无需每次手动输入。
- SmartKit 的典型应用场景包括以下：
  - 开局交付（安装部署）：标准化及规范化开局检查，确保开局后设备正常运行。
  - 例行维护：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：可以根据流程指引对设备进行升级和打补丁等操作，提升升级效率。
  - 扩容：支持提升设备性能或增加系统容量。
  - 部件更换：支持对故障FRU进行向导式更换。
  - 故障处理：针对华为存储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。

# DME Storage简介



## DME Storage功能特性概述





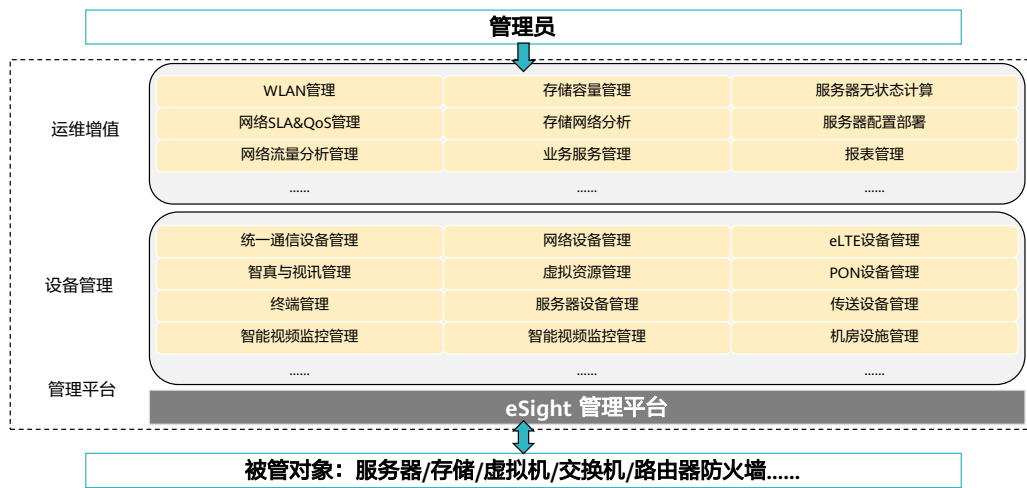
## eSight简介

- eSight支持多厂商设备适配，轻松实现全网设备统一管理；组件化的架构，按需构建企业运维平台；轻量级、Web化，减少系统维护与升级成本。



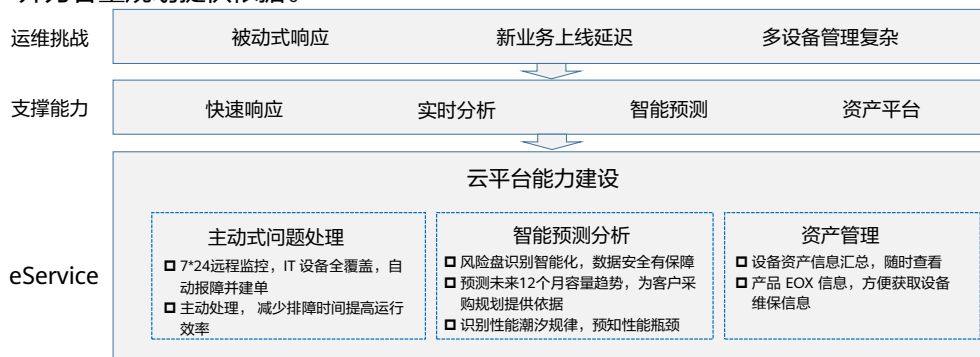
- 多厂商设备适配，轻松实现全网设备统一管理
  - eSight支持华为服务器、存储、虚拟化、交换机、路由器、WLAN、防火墙、PON、eLTE、统一通信、智真、视频监控等设备以及应用系统和机房设施的统一管理，并预集成对HP、Cisco、H3C等第三方主流设备的管理能力。
- 组件化的架构，按需构建企业运维平台
  - eSight采用组件化架构，在统一的eSight管理平台之上，提供丰富多样的组件，客户可以根据自己的情况选择所需要的组件。
- 轻量级、Web化，减少系统维护与升级成本
  - eSight采用B/S架构，客户端无需安装任何插件，可随时随地访问。当系统升级或维护时只需更新服务器端软件，减少系统维护与升级成本。

## eSight逻辑架构示意图



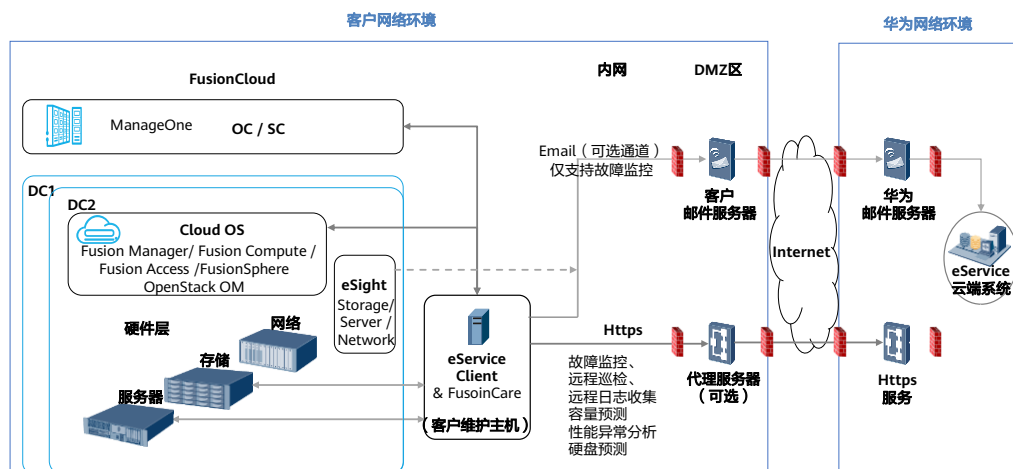
## eService简介

- 华为eService云智能管理平台，基于Cloud-Native（云原生），利用大数据分析和AI技术，提供自动报障、容量性能预测以及硬盘风险预测等服务，提前预防潜在风险，并为容量规划提供依据。



- 随着华为云计算、存储、服务器等IT产品的大规模交付使用，客户对提升故障处理效率的需求越来越强烈。传统的服务支持方式为全人工本地服务，在故障发现环节，技术人员面临着问题发现不及时、信息传递不到位的挑战。eService云智能管理系统（简称 eService），系统集云上管理、远程维护、自动建单、预测预防等功能，全方位提升客户的运维能力并将潜在风险转为可计划的维护动作。
  - eService提供自助式智能运维系统，为客户提供精准化的贴身信息服务。
  - eService云端系统使用华为云，通过大数据分析和AI人工智能技术驱动IT运维活动，提前识别故障，降低运维难度，全面提升运维效率。
  - 数据传输加密，保证信息安全传输且客户授权后eService才能接入客户系统。
  - 自动建单，提供 7\*24 小时不间断服务，主动式运维，安全可靠。
  - 客户可以在任何时间、地点使用任意PC访问eService，查看设备信息。

## eService系统架构示意图



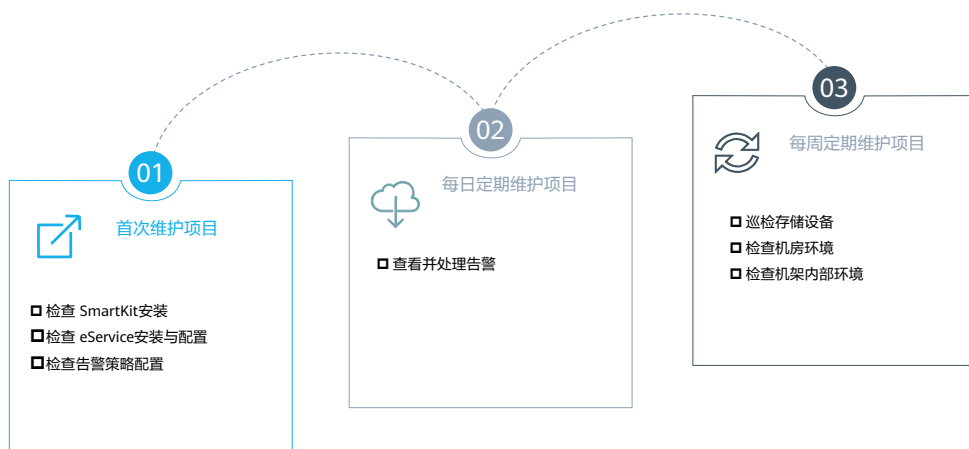
- eService系统由前端系统和云端系统协同工作。
  - 前端系统：部署在客户侧。
    - 能够采集客户的设备告警等信息，并及时将其发送到华为侧的云端系统，实现远程维护功能，如远程巡检、远程日志收集等。
  - 云端系统：部署在华为技术支持中心。
    - 能够全天候地接收客户侧前端系统发回的设备告警等信息，并自动通知华为技术服务人员进行及时处理，同时可发起对客户侧设备的自动巡检和日志收集。

# 目录

---

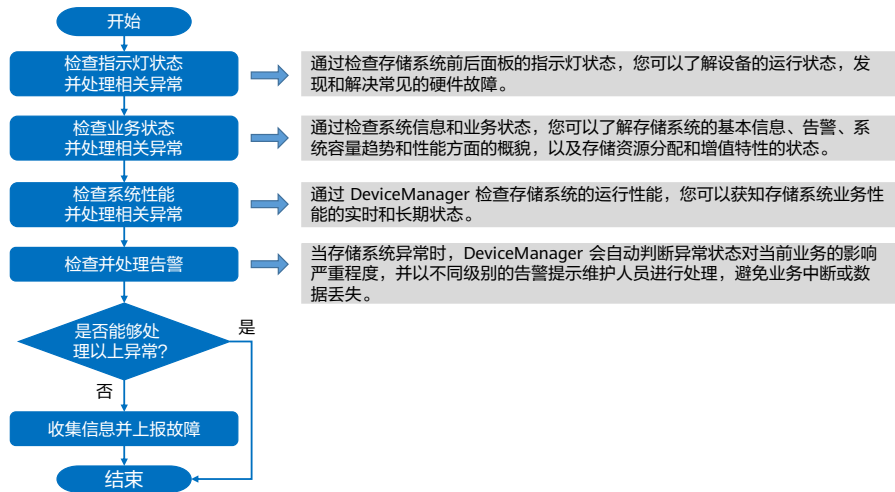
1. 运维概述
2. 运维工具
- 3. 运维场景**

## 维护项目总览



- 通过维护项目及周期列表，系统管理员能够确认设备环境、设备状态的情况，如出现异常状态可及时进行处理和维护，确保存储系统持续健康运行。
- 根据维护的频率和维护项目分为：首次维护、每天定期维护、每周定期维护。

## 快速维护流程



# 运维场景一 - 巡检

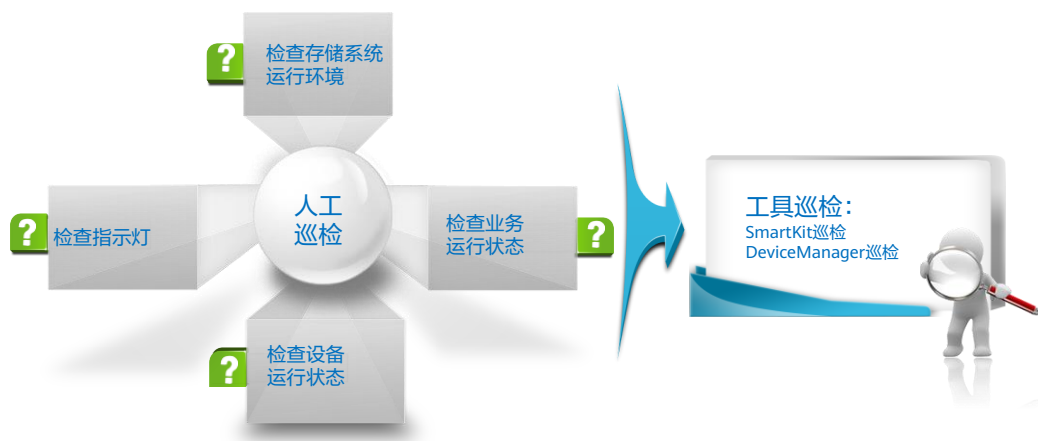
01

## 背景介绍

- 公司E采购的存储完成部署实施后，现已部署业务，正常开展各项业务活动。为了确保业务系统核心设备的存储的安全，公司IT部门的工程师A负责存储设备的巡检工作，请帮助工程师A制定巡检计划。



## 巡检方法



- 工具巡检：
  - 可以使用SmartKit或者DeviceManager对存储设备进行巡检；
  - 当未安装巡检工具时，可以通过CLI检查交换机的健康状态和运行状态。
- 人工手动巡检：
  - 包括例行检查机房和机柜内部环境，以及设备运行指示灯状态等，能够帮助维护工程师了解设备运行情况，及时发现设备出现的故障。

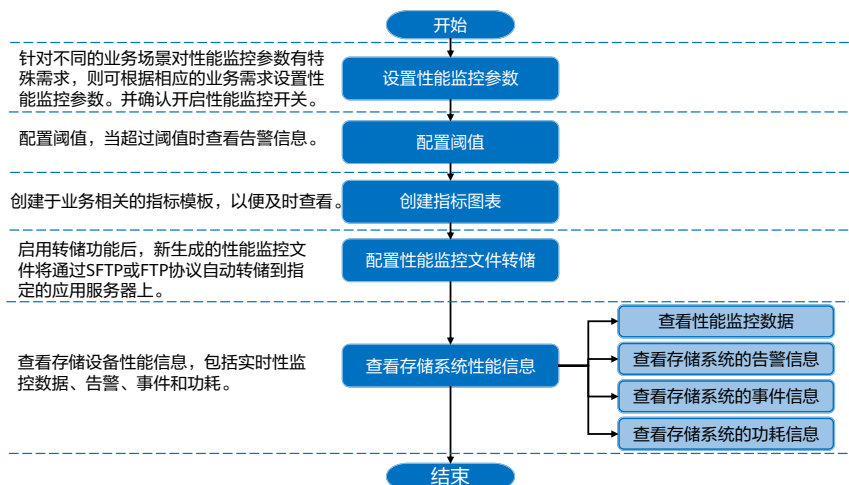
## 运维场景二 - 性能监控



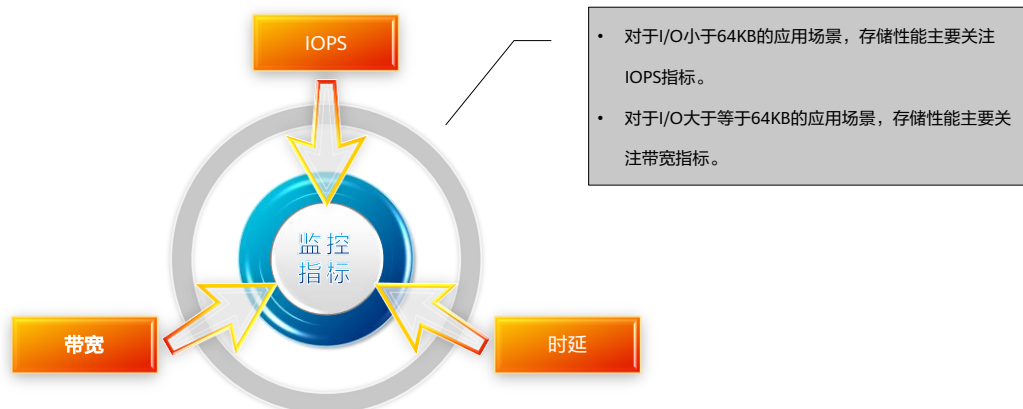
### 背景介绍

- 公司E业务已经上线。为了了解存储设备性能使用情况，公司IT部门的工程师B负责存储设备的性能监控工作，请帮助工程师B监控存储设备性能。

# 性能监控流程



## 性能监控指标



- 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。在数据库OLTP（Online Transaction Processing）业务等应用场景，通常以IOPS衡量系统的性能。
  - 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。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。
  - 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。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。

## 性能指标项

|   |         |   |            |   |           |
|---|---------|---|------------|---|-----------|
| 1 | 快照      | 1 | 逻辑端口       | 1 | 异构iSCSI链路 |
| 2 | 前端以太网端口 | 2 | 主机         | 2 | 异构FC链路    |
| 3 | LUN优先级  | 3 | 控制器        | 3 | 远程复制一致性组  |
| 4 | 后端SAS端口 | 4 | LUN        | 4 | FC复制链路    |
| 5 | 前端FC端口  | 5 | 存储池        | 5 | 远程复制      |
| 6 | 前端绑定端口  | 6 | SmartQoS策略 | 6 | 系统        |
| 7 | 硬盘      | 7 | LUN组       | 7 | 主机组       |

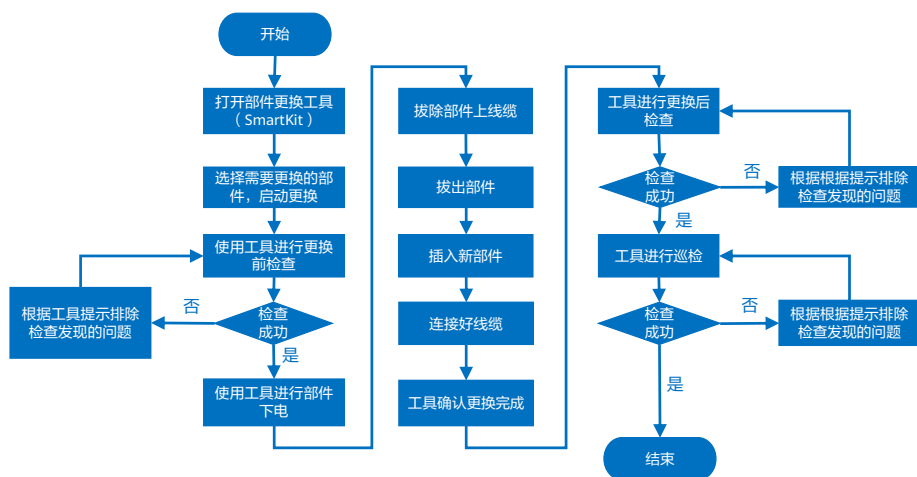
## 运维场景三 - 部件更换

03

### 背景介绍

- 企业E采购的存储设备使用一段时间后，存储系统上报硬盘故障，需要更换硬盘。华为技术支持工程师C收到该需求后，准备前往客户现场更换硬盘。

## 部件更换流程



## 可更换部件

| FRU    | CRU    |
|--------|--------|
| 控制器    | 电源模块   |
| 接口模块   | BBU 模块 |
| 系统插框   | 风扇模块   |
| 管理模块   | 硬盘模块   |
| 线缆     | 级联模块   |
| 辅助散热模块 | 光模块    |
| 仲裁服务器  | -      |
| 数据交换机  | -      |

- FRU(Filed Replaceable Unit) ， 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。
- CRU(Customer Replaceable Unit) 用户可更换单元，是指用户根据指导操作在设备现场进行更换的部件。



## 备件查询助手

- 华为备件查询助手提供基于存储产品型号、部件编码及SN号的方式快速获取备件信息，并支持备件手册的导出。
- 您可以直接登录华为网站<http://support.huawei.com/enterprise/>，在搜索栏中，输入“存储备件查询助手”，使用该工具。

华为数据存储信息中心 > 备件查询

按编码或SN查询    按产品查询

产品系列 OceanStor Dorado V6    产品型号 OceanStor Dorado 5300 V6    导出手册

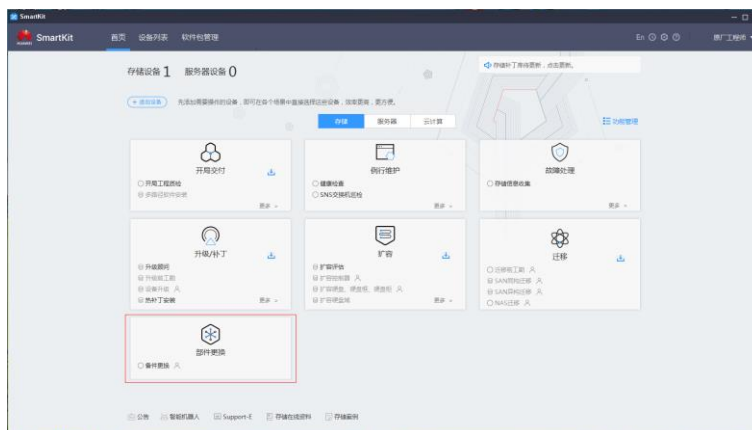
|         |                                        |                                             |                                          |                                         |                                          |                                          |                                          |
|---------|----------------------------------------|---------------------------------------------|------------------------------------------|-----------------------------------------|------------------------------------------|------------------------------------------|------------------------------------------|
| 机框      | <input checked="" type="checkbox"/> 全部 | <input checked="" type="checkbox"/> 控制器模块   | <input checked="" type="checkbox"/> 电源模块 | <input checked="" type="checkbox"/> BBU | <input checked="" type="checkbox"/> 风扇模块 | <input checked="" type="checkbox"/> 级联模块 | <input checked="" type="checkbox"/> 系统插槽 |
| 接口模块    | <input checked="" type="checkbox"/> 全部 | <input checked="" type="checkbox"/> 接口模块    | <input checked="" type="checkbox"/> 光模块  |                                         |                                          |                                          |                                          |
| 硬盘组件    | <input checked="" type="checkbox"/> 全部 | <input checked="" type="checkbox"/> SSD硬盘单元 |                                          |                                         |                                          |                                          |                                          |
| 服务器/交换机 | <input checked="" type="checkbox"/> 全部 | <input checked="" type="checkbox"/> 仲裁服务器   |                                          |                                         |                                          |                                          |                                          |
| 辅料      | <input checked="" type="checkbox"/> 全部 | <input checked="" type="checkbox"/> 线缆      | <input checked="" type="checkbox"/> PDU  |                                         |                                          |                                          |                                          |

## 硬盘更换要点

- 为防止硬盘模块损坏，在对硬盘模块进行操作时，只能接触硬盘模块的边缘。
- 拔插硬盘模块时用力要均匀，避免用力过大或强行拔插等操作，以免损坏部件的物理外观或导致接插件故障。
- 当对硬盘模块进行插拔时，插拔硬盘模块的时间至少间隔1分钟，即在拔出硬盘模块1分钟后，再插入硬盘模块，或在插入硬盘模块1分钟后再拔出硬盘模块，避免损坏硬盘模块。
- 为防止数据丢失，只更换硬盘告警/定位指示灯亮黄色的硬盘模块。
- 建议在硬盘拔出后的5分钟内完成硬盘模块的更换，否则系统散热性会受到影响。
- 更换风险盘（暂未故障），请务必使用SmartKit工具。
- 请确保更换后的硬盘插入到和更换前硬盘相同的槽位。插入错误的槽位可能会导致系统异常。

## 硬盘更换工具入口

- 打开SmartKit工具，选择部件更换工具。



## 硬盘更换向导



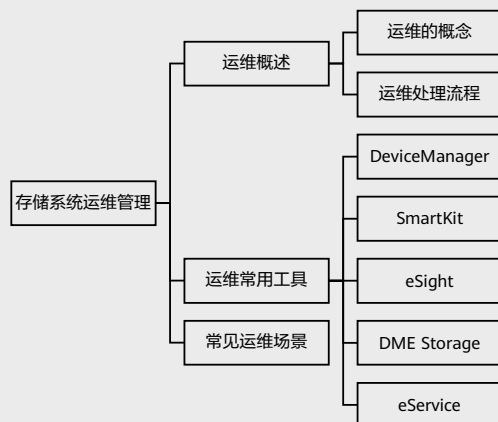
## 思考题

1. （多选题）常用管理软件有以下哪些（ ）
  - A. DeviceManager
  - B. eSight
  - C. SmartKit
  - D. eService
2. （判断题）SmartKit中包含了对IT设备进行部署、维护、升级等操作所需的各类工具，能够帮助产品用户、服务工程师和维护工程师在上述过程中对设备进行精准操作，降低操作难度、提升工作效率。（ ）

- 参考答案：

- ABCD
- T

## 本章总结



## 学习推荐

- 华为官方网站
  - 企业业务: <http://enterprise.huawei.com/cn/>
  - 技术支持: <http://support.huawei.com/enterprise/>
  - 在线学习: <http://learning.huawei.com/cn/>
- 热门工具
  - HedEx Lite
  - 网络资料工具中心
  - 信息查询助手

- 热门工具:
  - HedEx Lite: 华为产品文档管理工具, 支持浏览、搜索、更新和管理产品资料
  - eStor: 图形化存储仿真平台, 该平台通过对真实华为OceanStor全闪存仿真模拟, 帮助广大ICT从业者和用户快速熟悉华为存储系列产品, 了解并掌握相关产品的操作和配置。
  - 网络资料工具中心: 汇集网络产品的资料工具, 是您进行投标支撑、网络规划、项目交付和升级维护的好助手。
  - 信息查询助手: 提供华为产品命令和告警信息查询。

# 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©2022 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.

