# 网络存储笔记

# BUAA 王宇轩

2019.9

课程目标

•描述在当今复杂的信息管理环境中发现的挑战。

•描述存储技术解决方案(如DAS、NAS、SAN)。

•描述存储的关键业务驱动因素:信息可用性和业务连续性。

•描述常见的存储管理角色和职责。

•解释从经典数据中心到虚拟数据中心再到云计算的过渡阶段

•描述IT基础设施的计算、存储、网络、桌面和应用程序层的虚拟化技术

•描述VDC环境中的业务连续性解决方案

•解释云的关键特征、服务和部署模型

•描述云基础设施组件和服务管理流程

•描述云安全问题和解决方案

•列出迁移到云计算的主要考虑事项

平时作业（20%）+ 考勤成绩（20%）+ 实验（10%）+ 综合案例分析报告（30%）+在线测试（20%），额外课堂积极表现分。

参考书：

英文版 《Information Storage and Management》

中文版 《信息存储与管理》人民邮电出版社

# 目 录

[模块一 信息存储简介](#_Toc20331_WPSOffice_Level1) [1](#_Toc20331_WPSOffice_Level1)

[学习成果](#_Toc1650_WPSOffice_Level2) [1](#_Toc1650_WPSOffice_Level2)

[1.1 基础术语介绍](#_Toc15757_WPSOffice_Level2) [1](#_Toc15757_WPSOffice_Level2)

[模块一总结](#_Toc16230_WPSOffice_Level2) [3](#_Toc16230_WPSOffice_Level2)

[测验题目](#_Toc16336_WPSOffice_Level2) [4](#_Toc16336_WPSOffice_Level2)

[模块二 数据中心环境](#_Toc1650_WPSOffice_Level1) [5](#_Toc1650_WPSOffice_Level1)

[学习成果](#_Toc31323_WPSOffice_Level2) [5](#_Toc31323_WPSOffice_Level2)

[2.1 应用程序、DBMS和主机](#_Toc15757_WPSOffice_Level1) [5](#_Toc15757_WPSOffice_Level1)

[2.2 连接有关](#_Toc16230_WPSOffice_Level1) [10](#_Toc16230_WPSOffice_Level1)

[2.3 存储有关](#_Toc16336_WPSOffice_Level1) [11](#_Toc16336_WPSOffice_Level1)

[2.3.1 存储选项](#_Toc28240_WPSOffice_Level2) [11](#_Toc28240_WPSOffice_Level2)

[2.3.2 磁盘驱动器性能及有关计算](#_Toc25316_WPSOffice_Level2) [13](#_Toc25316_WPSOffice_Level2)

[2.3.3 基于应用程序要求和磁盘驱动器性能的存储设计](#_Toc8773_WPSOffice_Level2) [15](#_Toc8773_WPSOffice_Level2)

[2.3.4 主机对存储的访问](#_Toc8568_WPSOffice_Level2) [15](#_Toc8568_WPSOffice_Level2)

[模块二总结](#_Toc28104_WPSOffice_Level2) [17](#_Toc28104_WPSOffice_Level2)

[测验题目](#_Toc25934_WPSOffice_Level2) [17](#_Toc25934_WPSOffice_Level2)

[模块三：数据保护 – RAID](#_Toc31323_WPSOffice_Level1) [19](#_Toc31323_WPSOffice_Level1)

[学习成果](#_Toc31436_WPSOffice_Level2) [19](#_Toc31436_WPSOffice_Level2)

[3.1 RAID基础介绍](#_Toc20947_WPSOffice_Level2) [19](#_Toc20947_WPSOffice_Level2)

[2.2 RAID所用技术介绍](#_Toc9000_WPSOffice_Level2) [20](#_Toc9000_WPSOffice_Level2)

[2.3 RAID各级别介绍](#_Toc5407_WPSOffice_Level2) [22](#_Toc5407_WPSOffice_Level2)

[2.4 RAID对性能的影响及有关计算](#_Toc4550_WPSOffice_Level2) [27](#_Toc4550_WPSOffice_Level2)

[2.4.1 影响描述](#_Toc16336_WPSOffice_Level3) [27](#_Toc16336_WPSOffice_Level3)

[2.4.2 RAID性能损失计算示例](#_Toc31323_WPSOffice_Level3) [27](#_Toc31323_WPSOffice_Level3)

[2.4.3 计算示例](#_Toc28240_WPSOffice_Level3) [29](#_Toc28240_WPSOffice_Level3)

[模块三总结](#_Toc22236_WPSOffice_Level2) [29](#_Toc22236_WPSOffice_Level2)

[测验题目](#_Toc32366_WPSOffice_Level2) [30](#_Toc32366_WPSOffice_Level2)

[模块四 智能存储系统 ISS](#_Toc28240_WPSOffice_Level1) [32](#_Toc28240_WPSOffice_Level1)

[学习成果](#_Toc27499_WPSOffice_Level2) [32](#_Toc27499_WPSOffice_Level2)

[4.1 智能存储系统的关键组件](#_Toc24776_WPSOffice_Level2) [32](#_Toc24776_WPSOffice_Level2)

[4.1.1 ISS概述](#_Toc25316_WPSOffice_Level3) [32](#_Toc25316_WPSOffice_Level3)

[4.1.2 ISS关键组件介绍](#_Toc8773_WPSOffice_Level3) [32](#_Toc8773_WPSOffice_Level3)

[4.1.3 缓存有关](#_Toc8568_WPSOffice_Level3) [35](#_Toc8568_WPSOffice_Level3)

[4.2 存储资源调配和ISS实施](#_Toc21854_WPSOffice_Level2) [37](#_Toc21854_WPSOffice_Level2)

[4.2.1 存储资源调配方案](#_Toc9000_WPSOffice_Level3) [37](#_Toc9000_WPSOffice_Level3)

[4.2.2 LUN（逻辑盘）变形](#_Toc5407_WPSOffice_Level3) [39](#_Toc5407_WPSOffice_Level3)

[4.2.3 ISS的类型](#_Toc4550_WPSOffice_Level3) [40](#_Toc4550_WPSOffice_Level3)

[模块四总结](#_Toc7306_WPSOffice_Level2) [42](#_Toc7306_WPSOffice_Level2)

[测验题目](#_Toc12145_WPSOffice_Level2) [42](#_Toc12145_WPSOffice_Level2)

[模块五 光纤通道存储区域网络 (FC SAN)](#_Toc25316_WPSOffice_Level1) [44](#_Toc25316_WPSOffice_Level1)

[学习成果](#_Toc30695_WPSOffice_Level2) [44](#_Toc30695_WPSOffice_Level2)

# 模块一 信息存储简介

## 学习成果

• 定义数据和信息

• 描述数据类型

• 描述存储体系结构的发展历史

• 描述数据中心的核心元素

• 列出数据中心的关键特征

• 概述虚拟化和云计算

## 1.1 基础术语介绍

### 数据

可从中得出结论的未经处理的事实的集合

数据类型：结构化数据，非结构化数据（创建的大部分数据都是非结构化数据）

大数据：指大小超出常用软件工具在可接收时间限制内的捕获、存储、管理和处理能力大的数据集。

包括各种源生成的结构化和非结构化数据

大数据实时分析需要提供以下功能的新技术和工具：高性能，大规模并行处理（MPP）数据平台，高级分析

通过大数据分析，可以将大量数据转换为正确的决策

结构化数据：数据库

### 存储

存储个人和组织创建的数据，提供对数据的访问以备进一步处理

发展历史：

以服务器为中心的存储体系结构：



转为了以信息为中心的存储体系结构：存储将得到集中管理且独立于服务器

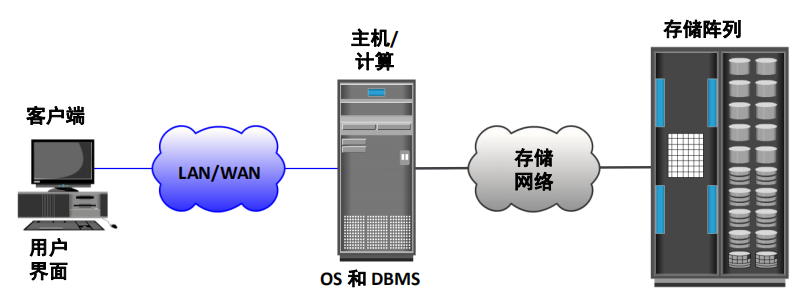


### 数据中心

包含存储，计算，网络和其他IT资源以提供集中式数据处理功能的设备

核心元素：应用程序，数据库管理系统（DBMS），主机或计算，网络，存储

一个完整数据中心的例子：



关键特征：可管理性，数据完整性，可用性，安全性，性能，可扩展性，容量

可扩展性：是指存储解决方案能够随业务一同增长

管理活动：

监视：收集数据中心运行的各种元素和服务的信息的持续过程

报告：有关资源性能，容量和利用率的详细信息

资源调配：配置和分配资源以可满足容量，可用性，性能和安全要求

虚拟化和云计算改变了数据中心基础架构资源的调配和管理方式

### 虚拟化

指抽象化物理资源并让其显示为逻辑资源的技术，例如原始磁盘的分区

共用物理资源并提供物理资源功能的聚合视图

可根据共用物理资源创建虚拟资源（提高物理IT资源的利用率）

### 云计算

支持个人和组织通过网络将IT资源作为服务使用

支持自助请求且自动化请求完成过程

支持用户快速纵向扩展计算资源的使用

支持基于消耗量的计量：用户只为他们使用的资源付费，例如使用的CPU小时数，数据传输量和数据存储量

## 模块一总结

本模块涵盖以下要点：

• 数据和信息

• 数据类型

• 大数据

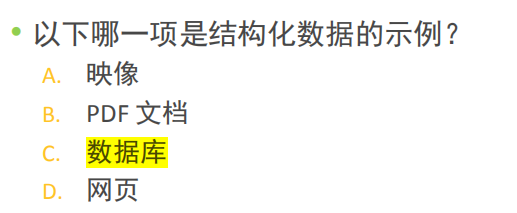
• 存储体系结构的发展历史

• 数据中心的核心元素

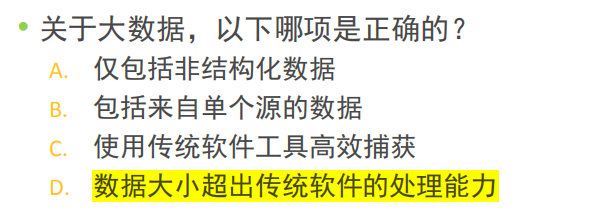
• 数据中心的关键特征

• 虚拟化和云计算

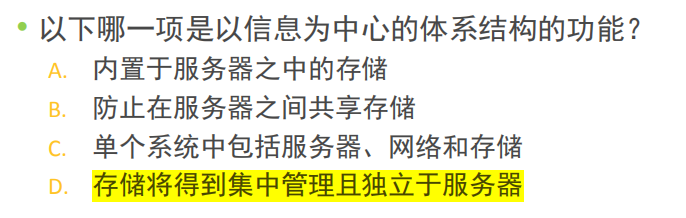
## 测验题目



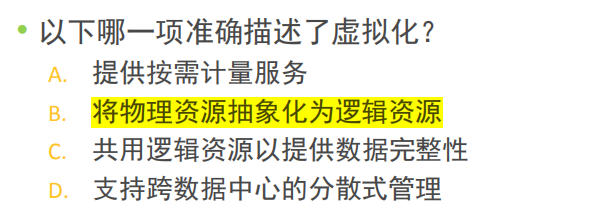
数据库中的数据均为结构化的，因此成为结构化数据



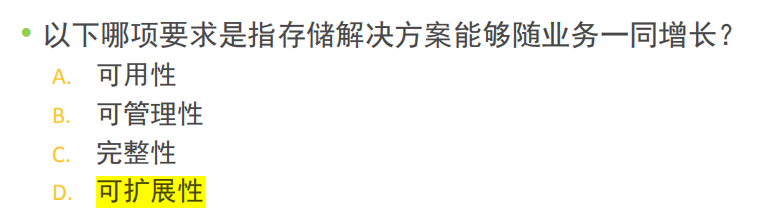
见大数据定义



见存储体系结构中以信息为中心的存储体系结构的示意图，先集中再独立管理。服务器为中心的是直接在服务器上进行存储。



见虚拟化定义



增长即扩展

# 模块二 数据中心环境

## 学习成果

• 描述数据中心的核心元素

• 描述应用程序和主机层的虚拟化

• 描述磁盘驱动器的组件和性能

• 描述主机通过 DAS 访问存储

• 描述闪存驱动器的工作和优势

## 2.1 应用程序、DBMS和主机

### 应用程序

为计算操作提供逻辑的软件程序

数据中心中通常部署的应用程序

业务应用程序——电子邮件，ERP，DSS

管理应用程序——资源管理，性能调整，虚拟化

数据保护应用程序——备份，复制

安全应用程序——身份验证，反病毒

关键IO特征

读取密集型与写入密集型

按序与随机

IO大小

### 应用程序虚拟化

一种向最终用户提供应用程序而无需任何安装，集成或底层计算平台上的依赖项的技术

允许在单独环境中提供应用程序

将操作系统资源和应用程序聚合到虚拟化容器中

确保操纵系统和应用程序的完整性

避免不同应用程序或同意应用程序的不同版本之间发生冲突

### 数据库管理系统DBMS

数据库是一种结构化存储方式，可将数据存储在相互关联并按逻辑组织的多个表中：有助于优化数据的存储和检索

DBMS便可控制数据库的创建，维护和使用

处理应用程序的数据请求

指示OS从存储中检索相应数据

SQL，SQL Server都是DBMS

### 主机（计算）

借助底层计算组件运行应用程序的资源

例如服务器，大型机，笔记本电脑

包含硬件和软件组件

硬件：CPU，内存和IO设备

软件：OS，设备驱动程序，文件系统，卷管理器

传统环境中，OS位于应用程序和硬件之间：负责控制环境

虚拟化环境中，虚拟化层在OS和硬件之间工作

虚拟化层可控制环境

OS仅操作一部分且仅控制应用程序环境

在某些实现中，会修改OS与虚拟曾进行通信

设备驱动程序是2使OS能够识别特定设备的软件

### 内存虚拟化

使呈现给应用程序的内存量大于实际可用内存的一项OS功能

附加内存空间来自磁盘

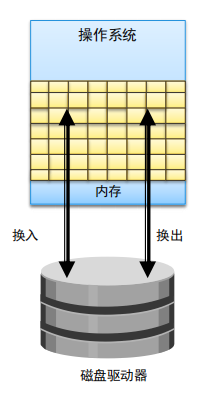
磁盘上虚拟内存使用的空间成为“交换空间/交换文件或页面文件”，虚拟内存效率比磁盘高

不活动的内存页面将从物理内存移动到交换文件

提供可用物理内存的高效使用

交换文件的数据访问较慢—交换空间使用闪存驱动器可提供最佳性能

示意图：



逻辑卷管理器LVM

负责创建和控制主机级逻辑存储

存储的物理视图转换成逻辑视图

逻辑数据块映射到物理数据块

一个或多个物理卷构成一个卷组

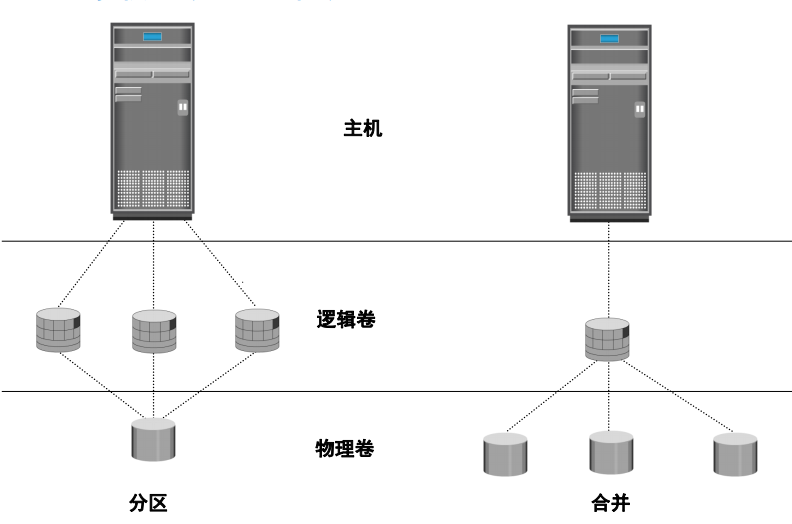
LVM将卷组作为单个实体进行管理

逻辑卷是从卷组中创建的

逻辑卷与卷组：

卷支撑着存多少数据，是服务器上的操作，一个存储需要很多的卷来完成  


分区和合并：



合并：将多个物理驱动器分组到逻辑驱动器

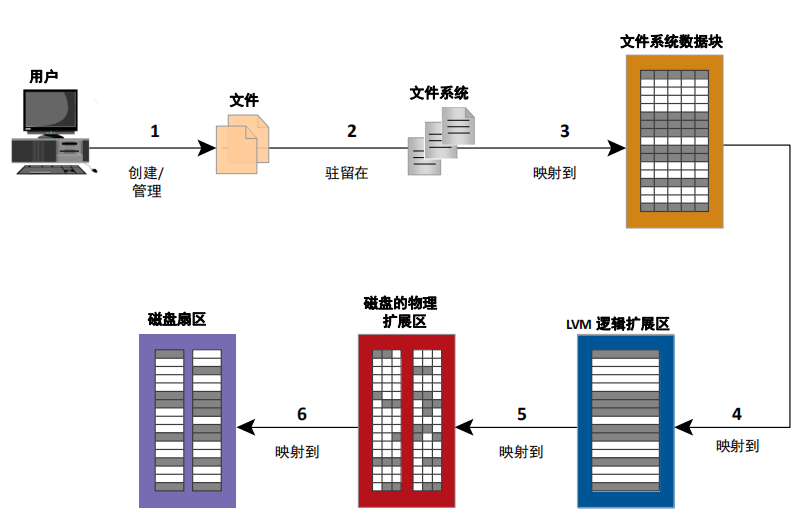
分区，将一个物理驱动器分为多个逻辑驱动器

文件系统则是用户和磁盘之间的一个系统，用于进行文件的传输

主机→逻辑卷→物理卷

磁盘→文件块→逻辑卷→物理卷→扇区

文件传输示意：



### 计算虚拟化

一项掩蔽或抽象化物理计算硬件的技术，支持对单个或群集物理机并行运行多个操作系统

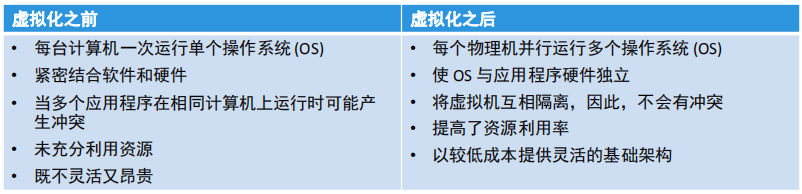
支持多台虚拟机VM，每台都运行一个OS和应用程序（虚拟机可以看作独立的文件组）

虚拟机是外观和行为类似于物理机的逻辑实体

虚拟化层位于硬件和虚拟机之间，也成为虚拟机管理程序

虚拟机提供有标准化的硬件资源

对计算虚拟化的需求，对比表格：



总结既是：并行多系统，OS与程序英阿锦独立，虚拟机隔离避免冲突，提高资源利用率，低成本提供灵活架构

### 桌面虚拟化

一项支持从中断设备断开用户状态，操作系统和应用程序的技术

支持组织托管和集中管理桌面：台式机在数据中心作为虚拟机运行

好处：启用瘦客户端访问更灵活，数据更安全，数据备份简化利于PC维护

## 2.2 连接有关

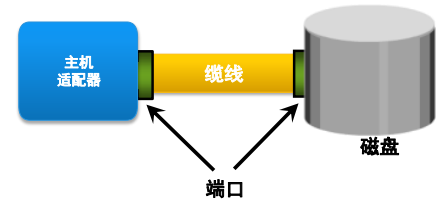
### 连接的定义

主机之间或主机与外围设备（如存储）之间的互联，磁盘和主机部分有连接，主机和总线之间有连接

连接的物理组件为：主机接口卡，端口和缆线

协议=发送设备与接收设备之间定义的通信格式，常用的存储接口协议：IDE/ATA和SCSI。

示意图：



IDE/ATA和串行ATA

IDE：集成的设备电子系统，ATA：高级技术附件

用于连接硬盘或CD-ROM驱动器的常用接口

可用于各种标准和名称

串行高级技术附件：SATA

取代ATA的IDE/ATA规范的串行版本

存储互联成本低，多用于内部连接

提供数据传输速度高，标准3.0，6G/s

SCSI和SAS

SCSI并行小型计算机系统接口

用于连接主机和外围设备的常用标准，常用于服务器中的存储连接

相比IDE/ATA成本高，PC不常用

可用于众多种类的相关技术和标准

一条总线最多支持16个设备

SAS串行连接SCSI

取代并行SCSI的点到点串行协议

支持的最大数据传输速度为6G/s（SAS2.0）

### 连接设备：光纤通道和IP

FC光纤通道

用于和存储设备进行高速通信的广泛使用的协议

可提供通过铜线或光纤操作的串行数据传输

最新版本的FC接口“16FC”允许的最大传输速度：16G/s

IP，用于传输主机到主机流量，提供现有基于IP的网络进行存储通信的机会，例如iSCSI和FCIP协议

## 2.3 存储有关

### 2.3.1 存储选项

磁带

低成本的长期解决方案

局限性：顺序存储和访问，一次一个应用程序，存在物理磨损

光盘

小型单用户计算环境中广泛用作分发介质

容量和速度有限

一次写入，多次读取（WORM）：CD-ROM，DVD-ROM

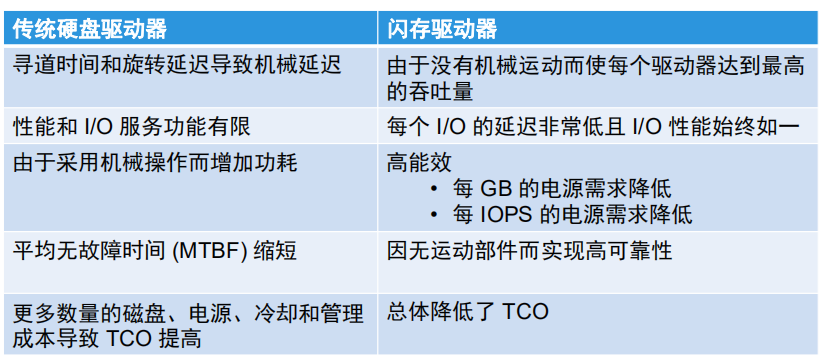
其他变体：CD-RW，Blu-ray磁盘

磁盘驱动器

最流行，很大存储容量，随机读写访问

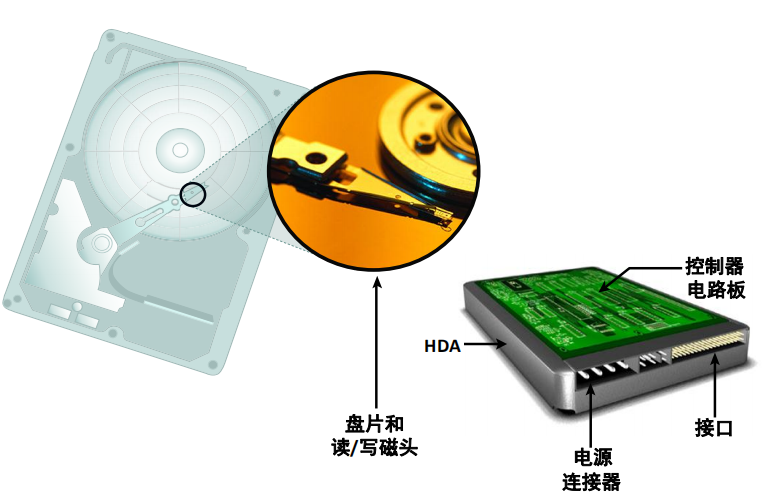
闪存驱动器

半导体介质，高性能低功耗



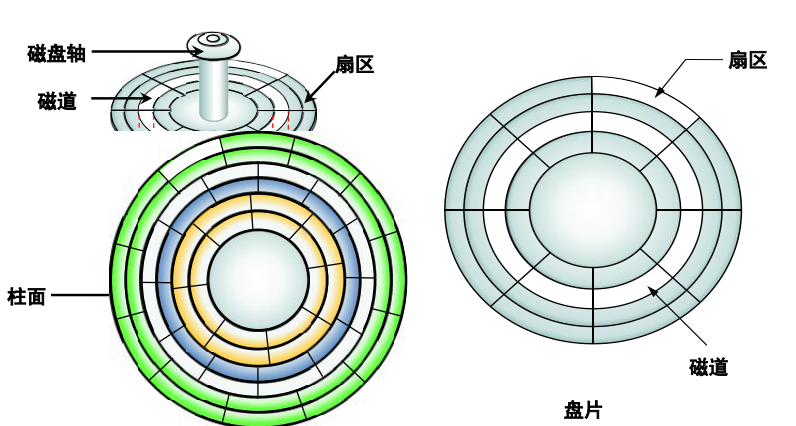
简述：无机械运动，每个驱动器最高吞吐量，IO延迟低且性能如一，高能效（每GB的电源需求降低，每IOPS（每秒读写次数）的电源需求降低），无运动部件高可靠性，整体降低TCO（总拥有成本）

磁盘驱动器组件示意图：



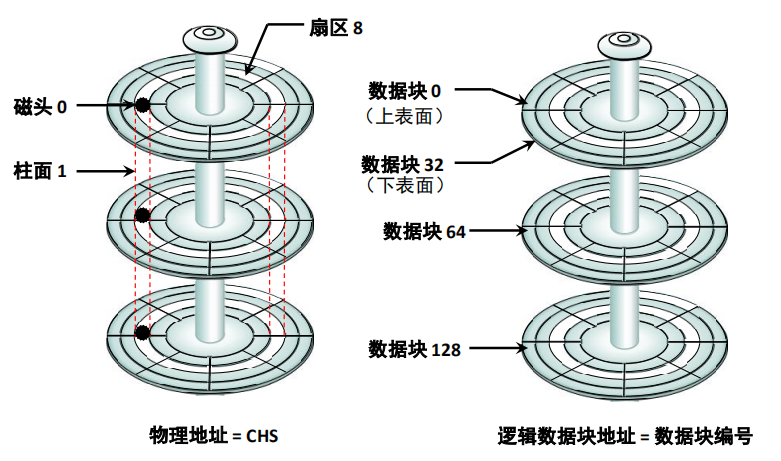
物理磁盘结构：

扇区是最小单位，扇区越多存的越多



逻辑数据块寻址

柱面正反相差32，两个柱面之间相差32



### 2.3.2 磁盘驱动器性能及有关计算

电子机械设备，影响存储系统的整体性能

磁盘服务时间，即磁盘完成IO请求所用的时间，等于寻道时间+旋转延迟+数据传输时间

寻道时间

定位读写磁头所用的时间，寻道时间越短IO操作越快

时间规范：全程，平均，道间

磁盘的寻道时间由驱动器制造商指定

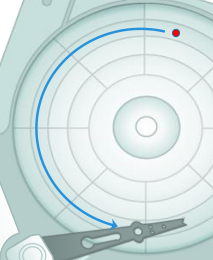


旋转延迟

盘片通过旋转将数据置于读写磁头下所用的时间

取决于磁盘轴的旋转速度

平均旋转延迟：旋转一周所用时间的一般，对于Xrpm的转速，驱动器延迟以毫秒为单位计算，公式为：(1/2)/(X/60)，转速对应于读写速率



数据传输时间

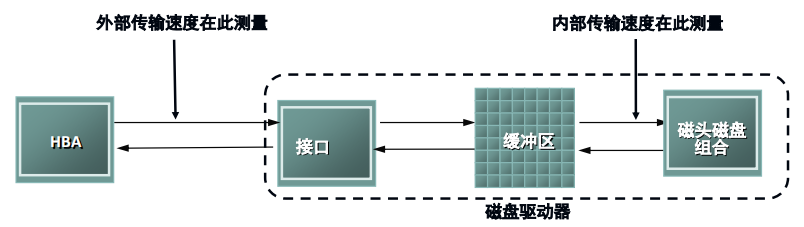
即由数据传输速度决定

数据传输速度：每单位时间驱动器可以向HBA输送的平均数据量（HBA：主机总线适配器Host Bus Adapter）

缓冲区：为减少对磁盘的损耗，将一部分用户数据放入其中

内部传输速度：数据从盘片表面移至内部缓冲区时的速度

外部：数据通过接口移交至HBA时的速度

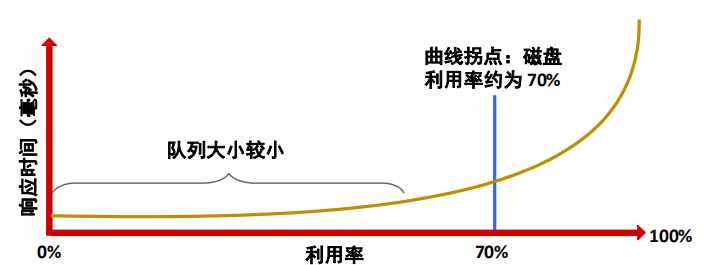


IO控制器的利用率与响应时间

磁盘驱动器性能的基本法则：平均响应时间=服务时间/(1-利用率)

服务时间就是控制器为IO提供服务所用的时间

根据公式，性能敏感型应用程序，磁盘的利用率通常低于其IO服务功能的70%，太高的话响应时间会极具上升。



### 2.3.3 基于应用程序要求和磁盘驱动器性能的存储设计

1.为满足应用程序容量所需的磁盘Dc

Dc=所需总容量/单个磁盘的容量

2.为满足应用程序性能需求所需的磁盘Dp

Dp=峰值工作负载时应用程序生成的IOPS/单个磁盘提供服务的IOPS

3.根据磁盘服务时间由磁盘提供服务的IOPS（Ts）

Ts=寻道时间+(0.5/(磁盘转速/60))+(数据块大小/数据传输速度)

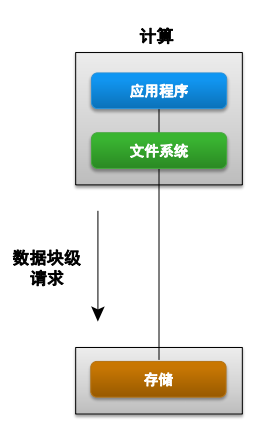
Ts是IO完成所用的时间，因此，有磁盘提供服务的IOPS等于1/Ts

对于性能敏感型应用程序(S)=0.7/Ts

应用程序所需的磁盘等于max(Dc,Dp)

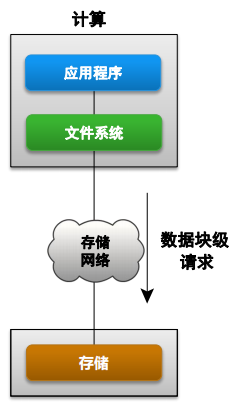
### 2.3.4 主机对存储的访问

直连存储DAS（可扩展性有限，有内部直连和外部直连）

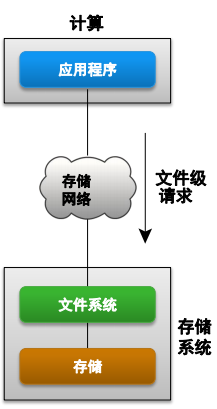


存储网络选项包括数据块级访问和文件级访问

数据块级访问：



文件级访问：



## 模块二总结

本模块涵盖以下要点：

• 数据中心关键元素

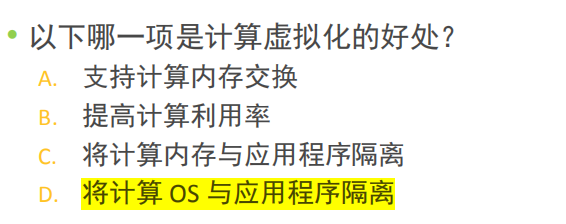
• 应用程序和计算虚拟化

• 磁盘驱动器的组件和性能

• 企业级闪存驱动器

• 主机对存储的访问

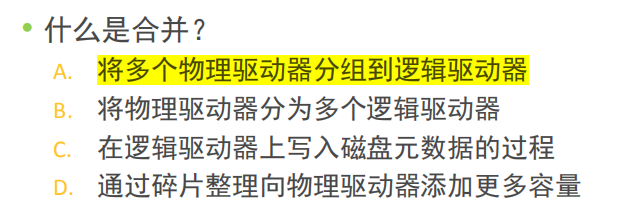
## 测验题目



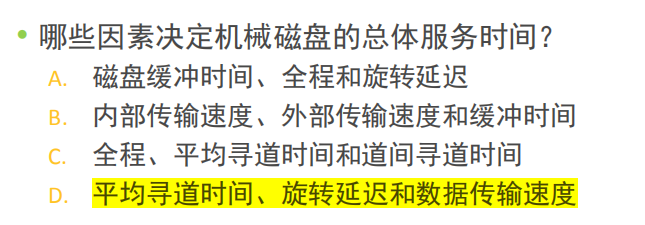
见计算虚拟化的好处



记住即可



可参考合并的示意图进行理解



三个主要的因素



见DAS简介

# 模块三：数据保护 – RAID

## 学习成果

• 描述 RAID 实现方法

• 描述三种 RAID 技术

• 描述常用 RAID 级别

• 描述 RAID 对性能的影响

• 根据 RAID 级别的成本、性能和保护能力比较各个级别

## 3.1 RAID基础介绍

### RAID

一项将多个磁盘驱动器合并到一个逻辑单元（RAID集）中，并提供保护和性能的技术

由于存在机械组件所以提供的性能有限

每个驱动器由特定的平均预期寿命，以MTBF为单位（MTBF：平均故障间隔时间，根据磁盘的预期寿命计算得出），例如一个驱动器的MTBF为750000小时，阵列中有1000台驱动器，则该阵列的MTBF为750小时

### RAID实现方法

软件实现

使用基于主机的软件提供RAID功能，即使用主机CPU周期执行RAID计算

限制

使用主机CPU周期执行RAID计算会影响整体性能

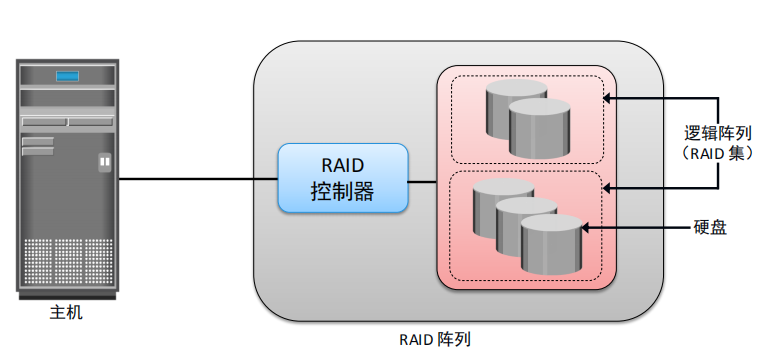
支持有限的RAID级别

仅当RAD软件和操作系统兼容时才能升级

硬件实现

使用安装在主机或阵列上的专用硬件控制器

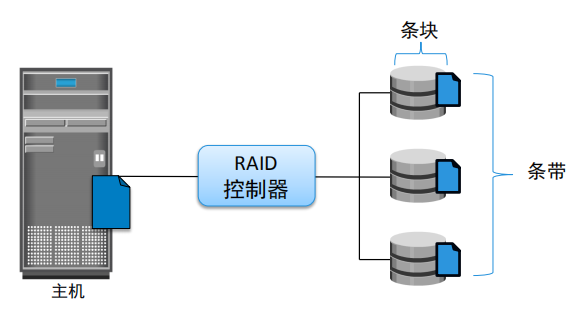
RAID阵列组件



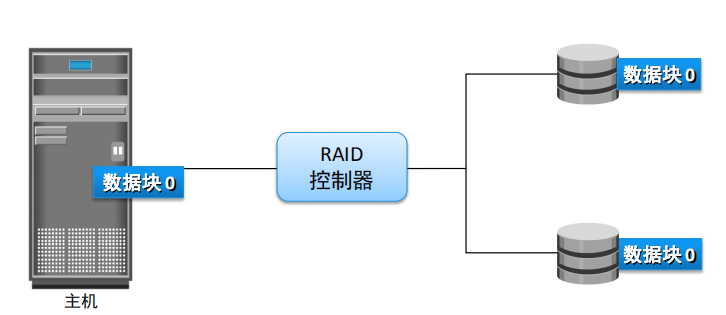
## 2.2 RAID所用技术介绍

分条，镜像，奇偶校验（分散校验）

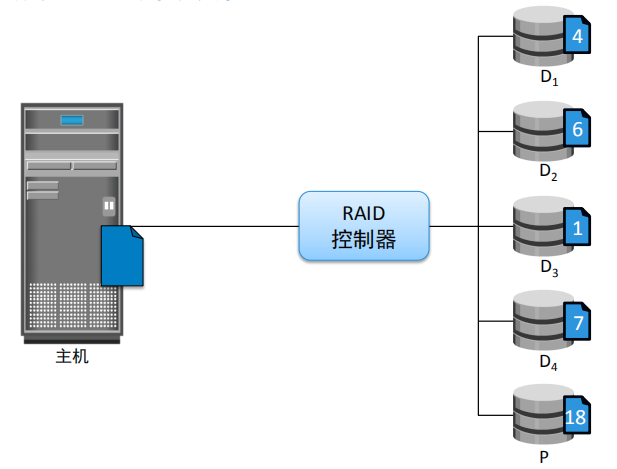
分条：将一个条带分为多块（分条后有木桶效应，按储存容量以最小的条为大小限制）



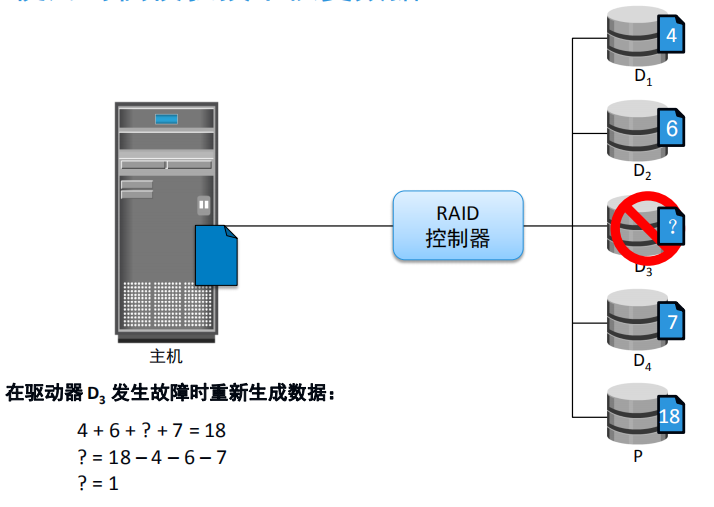
镜像：将一个数据块储存为两份



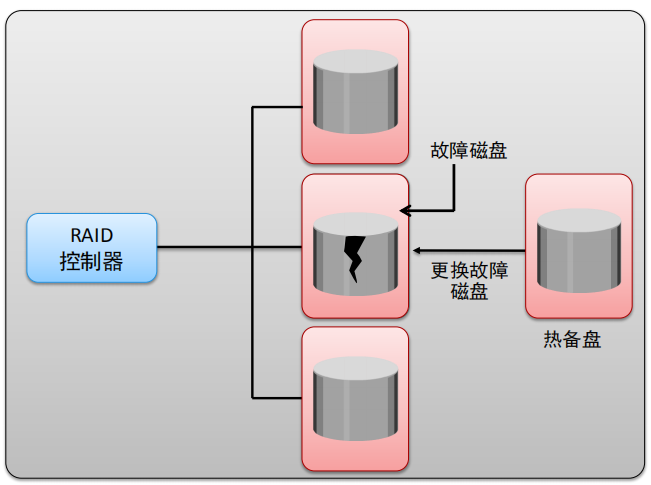
奇偶校验（分散校验）：将一个数据分开储存，合并时进行校验，通过校验可以发现错误和丢失的信息



分散校验可用于恢复数据，例如当有一个缺失，可通过计算找出：



### 热备盘



在RAID控制器中作为备用盘，用于替换故障的磁盘。

## 2.3 RAID各级别介绍

RAID 0 – 无容错能力的分条集（仅分条）

RAID 1 – 磁盘镜像（仅镜像）

RAID 1 + 0 – 表示一个嵌套 RAID（先分条再镜像，根据10出现的顺序）

RAID 3 – 具有并行访问和专用奇偶校验磁盘的分条集

RAID 5 – 具有独立磁盘访问和分布式奇偶校验的分条集

RAID 6 – 具有独立磁盘访问和双分布式奇偶校验的分条集

RAID 0 – 无容错能力的分条集（仅分条）

仅将ABC分开储存了

安全性最差



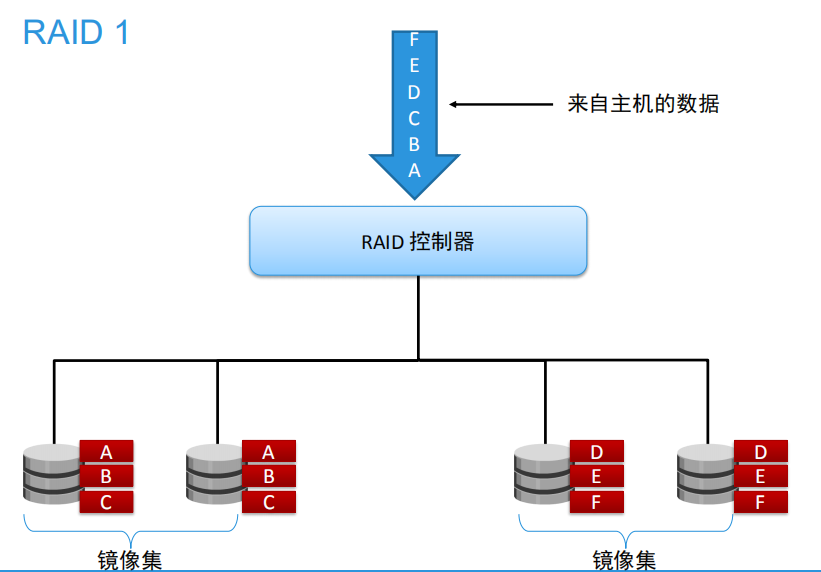
此时如果A1出错，那么B1C1是访问不到的，因为在结构上B1C1位于A1后，而且当A1出错后后面和A1位于一个条带的A2B2C2,A3B3C3等等也无法被访问到。

这两个属性在后面结合判断时会起到作用。

RAID 1 – 磁盘镜像（仅镜像）

只是所有数据都进行了备份

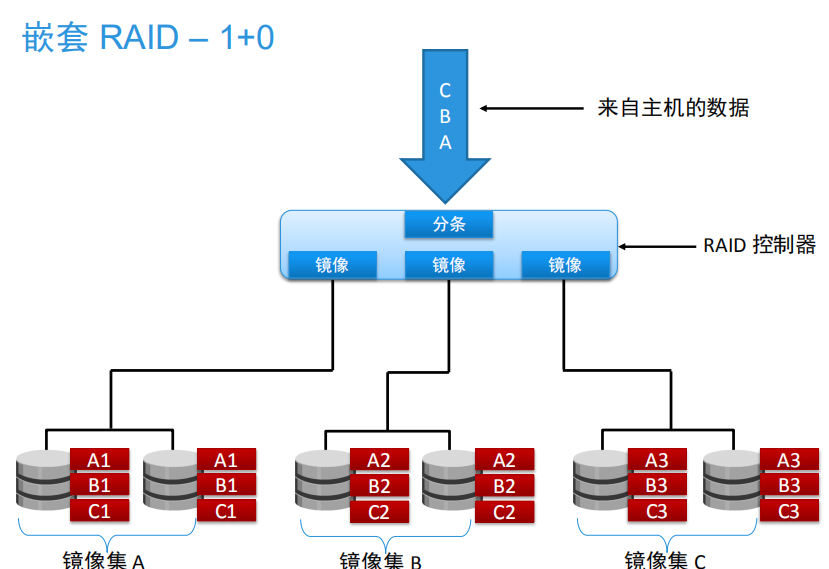
需要两倍的磁盘空间



RAID 1 + 0 – 嵌套 RAID（先分条再镜像），简写为RAID10，如果多个数字就是先前面那个模式然后在基础上应用后面那个数字的模式。

先分开储存再分别备份

如果在一个条带组中丢失驱动器，则必须从另一个条带组中访问数据，因为条带组没有奇偶性。RAID 1+0至少需要4个物理磁盘。



先将ABC分条分为三部分，然后再镜像分别储存为两份。

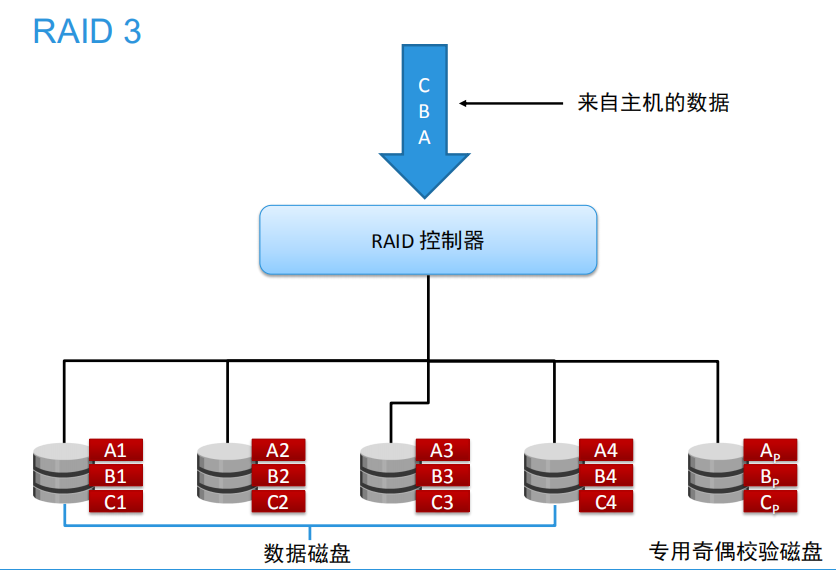
写A时会分别在三个逻辑磁盘上进行A各部分的获取。

当出现错误时，只要两个镜像没有同时出现错误就不会发生异常。

RAID 3 – 具有并行访问和专用奇偶校验磁盘的分条集

分条后专门建立奇偶校验的磁盘进行校验，和5对数据进行分类不太痛

大数据性能差，长时间连续数据传输表现良好，需要至少三个磁盘



分条后有一个单独磁盘作为专用的奇偶校验磁盘

优势：顺序读写能力强，即对于数据写入不经常变化的情况下有优势，随机读写即变化多时不占优势。

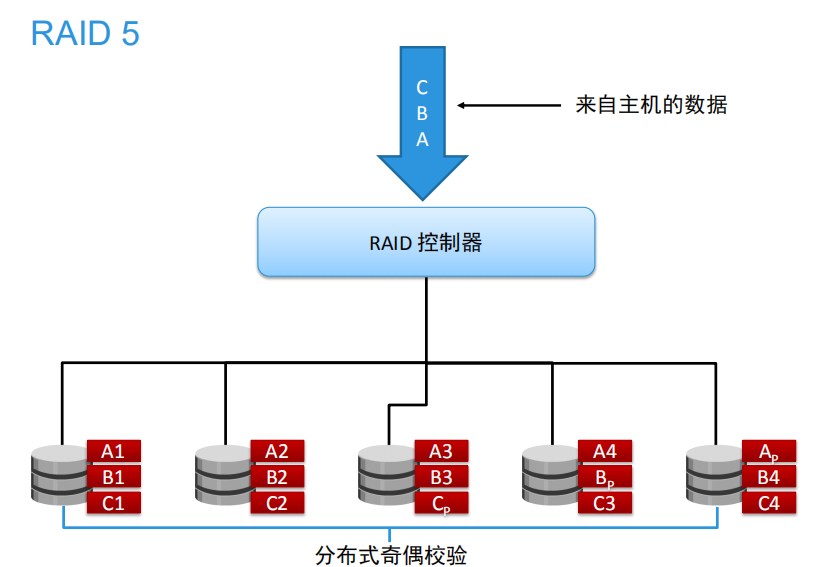
保护上作用好，但是性能不太优秀

RAID 5 – 具有独立磁盘访问和分布式奇偶校验的分条集

最常见

分条和奇偶校验混合一起储存，需要至少三个

所有磁盘上都有条带，并带有在磁盘故障时重建数据所需的奇偶性信息



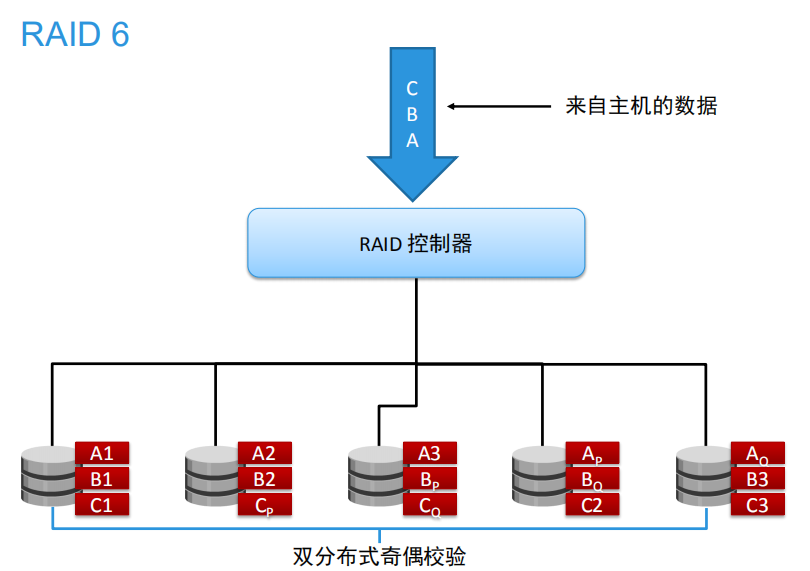
奇偶校验分布于各个磁盘内部，兼顾了容错和性能

应用广，对于随机读取有较好的性能

RAID 6 – 具有独立磁盘访问和双分布式奇偶校验的分条集

使用两个奇偶校验条带来提高可靠性

通常用于储存大量数据，在容量较大、重建时间较长的驱动器中更为常见。需要至少4个驱动器



随着用户对保护要求的增加，数据存储修改为三个存储块，两个校验信息，牺牲两块磁盘，性能下降，安全提高。

各级别对比图



效率最高的是RAID0，因为没有保护所以没有性能消耗。

不同RAID级别的适用：

RAID1+0

小型随机和写入密集型（写入量通常大于30%）的IO配置文件的应用程序

例如OLTP，RDBMS-临时空间

RAID3

大型，顺序读取，写入

例如：数据备份和多数据流

RAID5和6

小型，随机工作负载（写入量通常小于30%）

例如：电子邮件，RDBMS-数据输入

RAID5处理超小型数据时还不如单块磁盘性能高（例如1k1k的）

用于小型随机IO的RAID6配置中的写性能损失为6

## 2.4 RAID对性能的影响及有关计算

### 2.4.1 影响描述

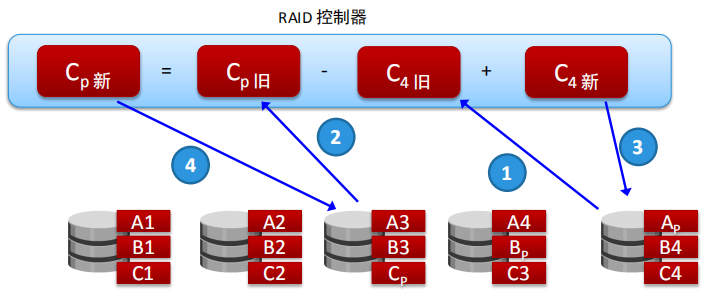
RAID 1中，每次写入都表现为两次IO操作（两次磁盘写入）

RAID5中，每次磁盘写入（更新）都表现为四次IO操作（两次磁盘读取和两次磁盘写入）

RAID6中，每次写入都表现为六次IO操作（三次读和三次写）

这些数字就是对应配置的读/写性能损失，例如RAID六的写性能损失即为6（六次IO操作）

RAID5示意图：



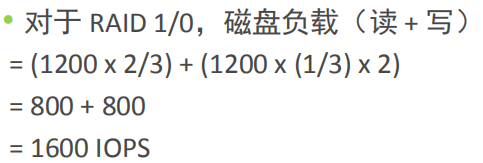
### 2.4.2 RAID性能损失计算示例

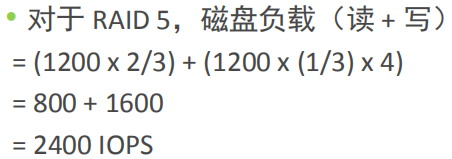
假设高峰工作负载时的IOPS为1200

读写比为2：1

针对RAID1/0和RAID5计算高峰活动时的磁盘负载

计算时需要分别计算读写





若要计算出磁盘个数，需要用结果除以磁盘总量得到实际需要存储数据的IOPS，若要计算出实际总共的数据还需要计算损耗，例如算出来要3个，对RAID5就得倒序(n-1)/n得出实际需要4个。

当多种级别结合，例如10结合需要考虑他们共同的性能影响，对于在读写上共同有惩罚的组合，则应将惩罚乘起来，例如15组合，写惩罚就是2\*4=8。

正常计算：

所有磁盘的IOPS=读+写

读=总IOPS\*读不命中率\*读操作的比例

写=总IOPS\*写不命中率\*写操作的比例

如果应用于RAID级别下，需要对读和写进行一个改变，因为有时候在进行读或写操作时有很多的子操作。

例如RAID5在一次写操作时会有四个IO，其中两次读两次写，这四次操作的比例是和写操作的比例一样的，不过读和写的命中率都是全局的。

RAID5下：

所有磁盘的IOPS=读+写

读=总IOPS\*读不命中率\*读操作的比例+总IOPS\*读不命中率\*写操作的比例\*2

写=总IOPS\*写不命中率\*写操作的比例\*2

例如RAID10在一次写操作是会有两个IO，两次写。

RAID10下：

正常计算：

所有磁盘的IOPS=读+写

读=总IOPS\*读不命中率\*读操作的比例

写=总IOPS\*写不命中率\*写操作的比例

其他的RAID级别的子操作如下：

等学习后会进行添加。

### 2.4.3 计算示例



## 模块三总结

本模块涵盖以下要点：

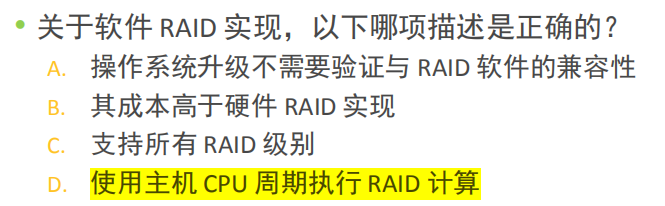
• RAID 实现方法和技术

• 常用 RAID 级别

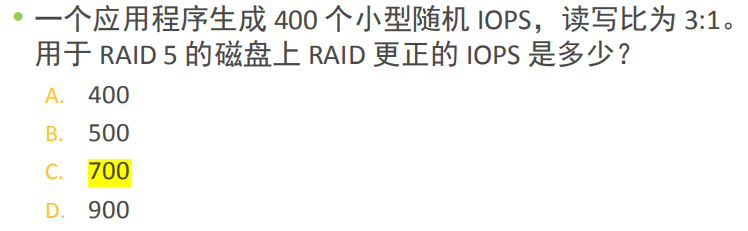
• RAID 写性能损失

• 根据 RAID 级别的成本和性能比较各个级别

## 测验题目



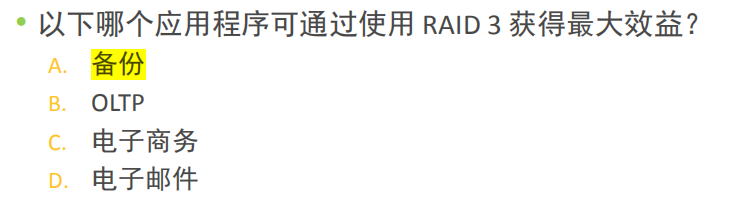
RAID的原理



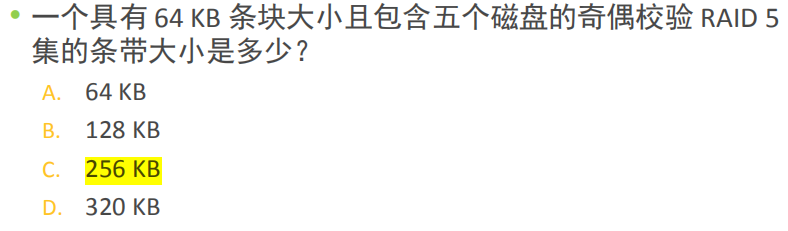
400\*(3/4)\*1+400\*(1/4)\*4 = 700



一次写实际为六次IO操作，因此 损失为6



备份多为直接的顺序存储，变化性小，所以适于大量顺序操作的RAID3



RAID5的磁盘应用量为(n-1)/n，所以五个磁盘能用上四个：4\*64 = 256

· 三块73G磁盘和两块153G磁盘，采用RAID5级别实际储存的内存为多少？

292G，因为此时磁盘大小不一，大磁盘能储存的容量和小磁盘一样，所以相当于五个73G磁盘，然后(5-1)/5\*5 = 4，4\*73=292G。

# 模块四 智能存储系统 ISS

## 学习成果

• 介绍智能存储系统的关键组件

• 介绍缓存管理和保护技术

• 介绍两个存储资源调配方法

• 介绍两种类型的智能存储系统

## 4.1 智能存储系统的关键组件

### 4.1.1 ISS概述

功能丰富的RAID阵列，提供高度优化的I/O处理功能

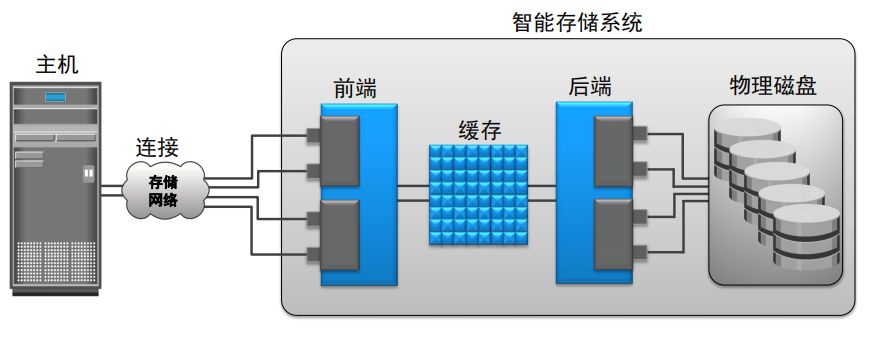
提供了可增强性能的大量缓存和多条I/O路径

具有提供如下功能的操作环境：

智能缓存管理，阵列资源管理，到异构主机的连接

支持山村驱动器，虚拟资源调配和自动存储分层。

### 4.1.2 ISS关键组件介绍



**前端**

包括端口和控制器

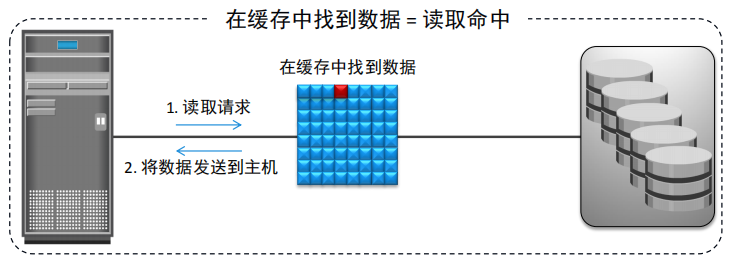
控制信息在后面如何传输。

**缓存**

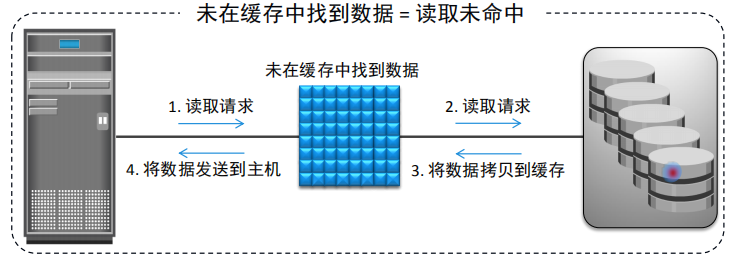
进行数据交互，包括读取和写入

读取

读取命中：此时速度快



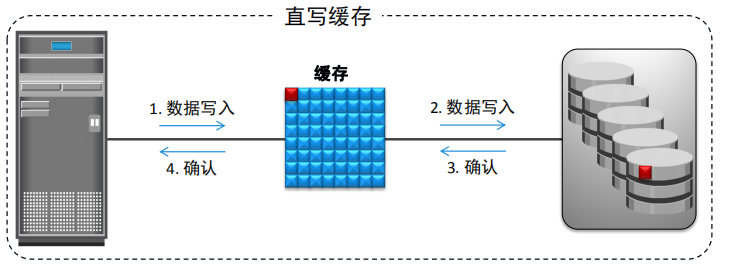
读取未命中：需要和后端联系获取物理磁盘上的数据，速度会变慢。



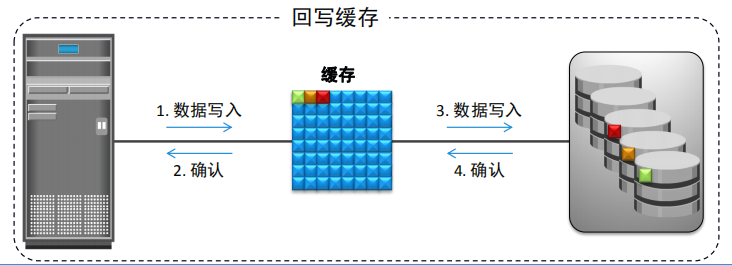
缓存上的时间是毫秒级，物理磁盘上是秒级。

写入

直写缓存：需要等到缓存存储到物理磁盘，速度较慢

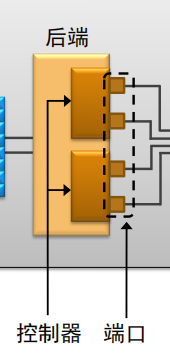


回写缓存：传输到缓存后便认为已经传输成功（此时主机可进行其他操作），空闲时后端再写数据，速度快



**后端**

同样包括端口和控制器



物理磁盘也是作为组件用于存储数据。

### 4.1.3 缓存有关

缓存管理：算法

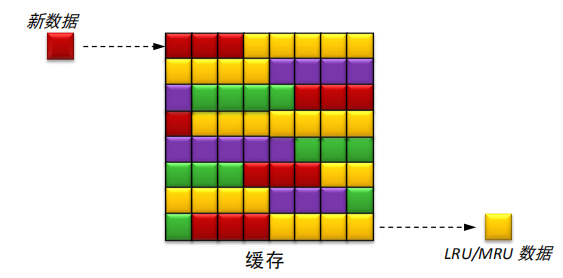
最近最少使用（LRU）

删除很长时间未访问的数据

最近最常使用（MRU）

删除最近最常访问的数据

当有新数据进来时，如果内存不够便会利用以上方法进行数据的删除：



缓存管理：水位线

通过刷新过程管理突发I/O

刷新是将缓存中的数据提交到磁盘的过程

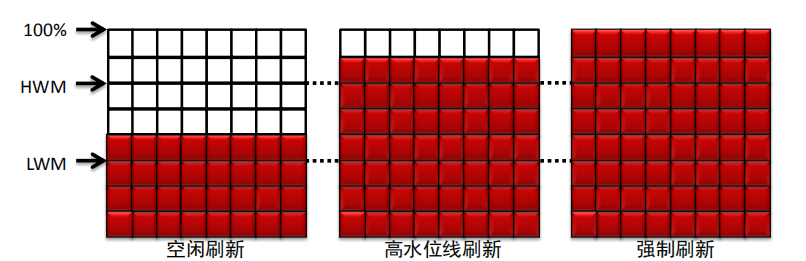
管理缓存利用率的三个刷新模式：

空闲刷新：未到达水位线时在空闲时进行刷新

高水位线刷新：当数据量达到设定的水位线时进行刷新

强制刷新：数据量突然超过水位线时强制进行刷新。

图示：



缓存数据保护

防止缓存中的数据受到电源或缓存故障的影响

1.缓存故障：缓存镜像

每次写入到缓存中的数据都保存在两个独立内存卡中的两个不同内存位置，一个出错时另一个也能正常储存。

2.电源故障：缓存保险存储

在出现电源故障时，将未提交的数据转储到称为“保险存储驱动器”的一组专用驱动器中

服务器闪存缓存技术

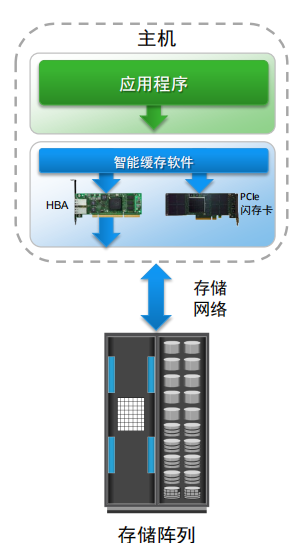
使用主机上的智能缓存软件和PCle内存卡，将数据放在服务器上的PCle闪存中，使用最少的CPU和内存资源（闪存管理减负到PCle卡上）

作用：显著提高应用程序性能

1.为读取密集型工作负载提供性能加速

2.避免与对存储阵列的I/O访问关联的网络延迟

原理图：



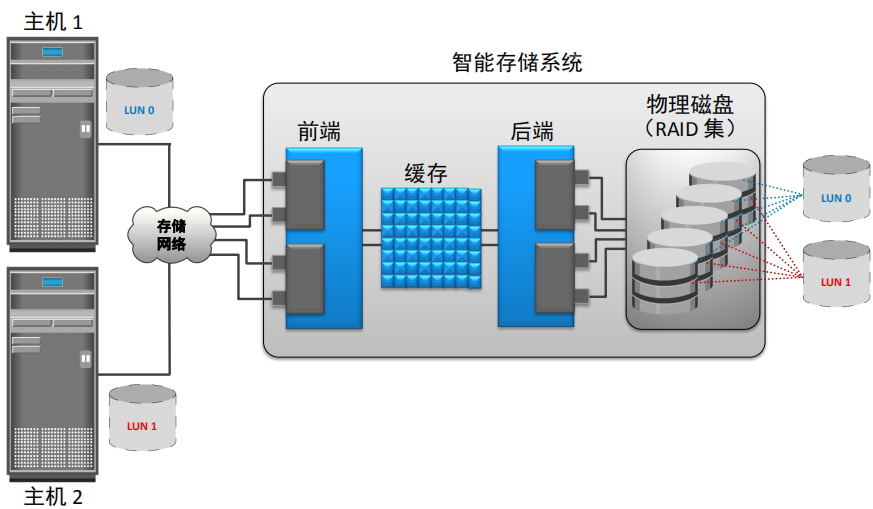
## 4.2 存储资源调配和ISS实施

向主机分配存储

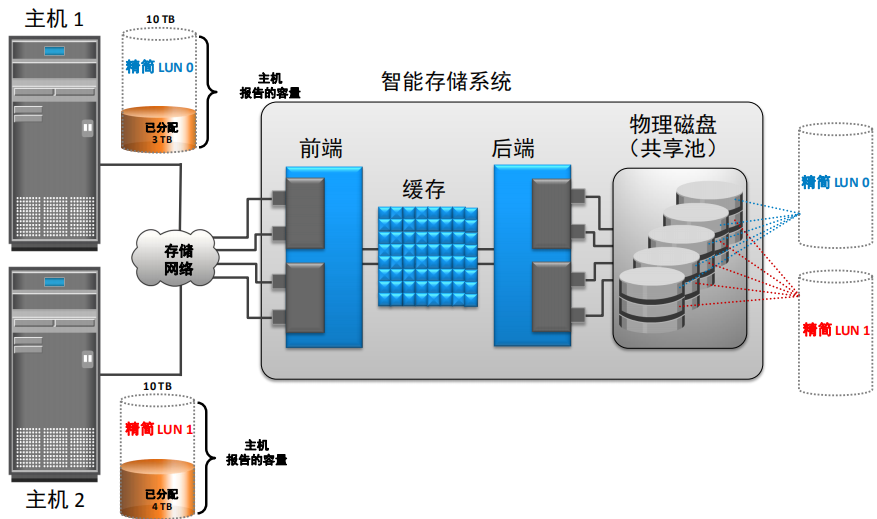
是根据主机上运行的应用程序的容量，可用性和性能要求向主机分配存储资源的过程，有两种方法可以执行：传统存储资源调配，虚拟存储资源调配。

### 4.2.1 存储资源调配方案

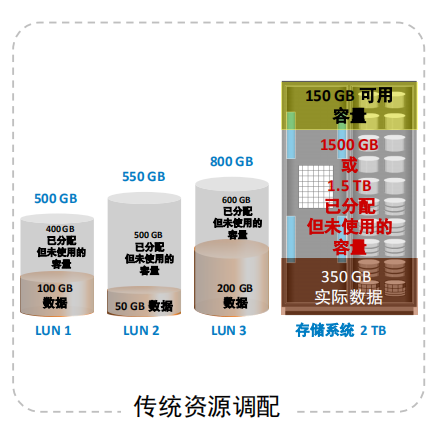
传统存储资源调配方案示意图：

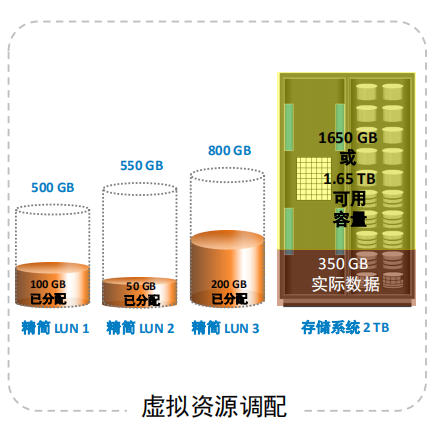


虚拟存储资源调配方案示意图：



对比：

全部分配不全部使用

分配即使用

### 4.2.2 LUN（逻辑盘）变形

**LUN扩展**

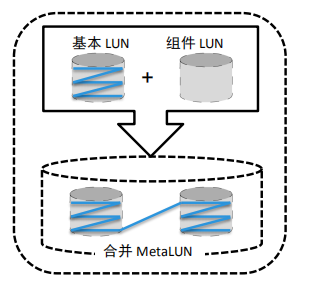
是扩展需要附加容量或性能的LUN的方法，通过组合多个LUN来创建一个MetaLUN，实现LUN的扩展。

MetaLUN可以是合并的，也可以是分条的。

合并metaLUN：

仅提供附加容量，而不提供信念

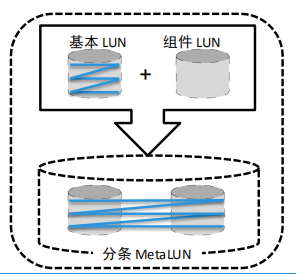
扩展很快，因为未重新条带化数据



分条metaLUN：

提供容量和性能

扩展很慢，因为会重新条带化数据



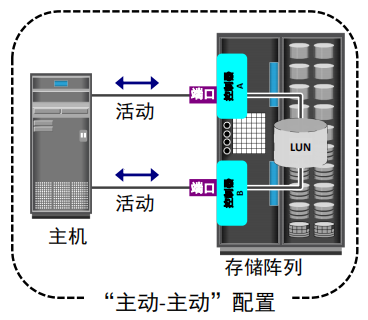
**LUN掩蔽**

是通过定义主机可以访问哪些LUN，来进行数据访问控制的过程

它在存储阵列上实施，防止在共享环境中未授权或意外使用LUN

### 4.2.3 ISS的类型

**高端存储系统：**



称为主动-主动阵列，通常面向大型企业应用程序，通过所有可用路径执行对LUN的I/O。

提供如下功能：

高存储容量和大型缓存

容错体系结构

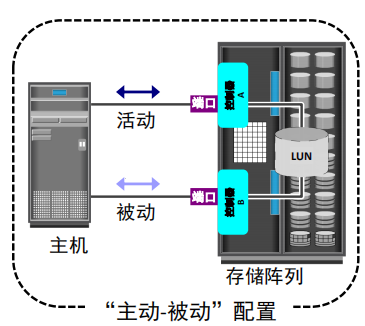
到大型机和开放系统的连接

多个前端端口和接口协议

能够处理大量并发 I/O

支持本地和远程数据复制

**中端存储系统：**



被称为主动-被动阵列，通常面向中小型企业应用程序，仅通过霍东阁路径执行对LUN的I/O

这些阵列通常有两个控制器，每个控制器都具有缓存，RAID控制器和磁盘驱动器接口

和高端阵列相比，前端端口、存储容量和缓存更少。

支持本地和远程数据复制

## 模块四总结

本模块涵盖以下要点：

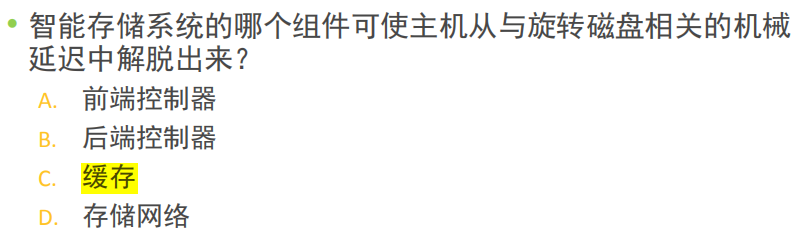
• 智能存储系统的主要组件

• 缓存管理和保护技术

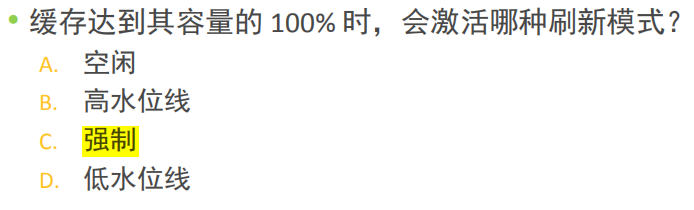
• 存储资源调配方法

• 智能存储系统的类型

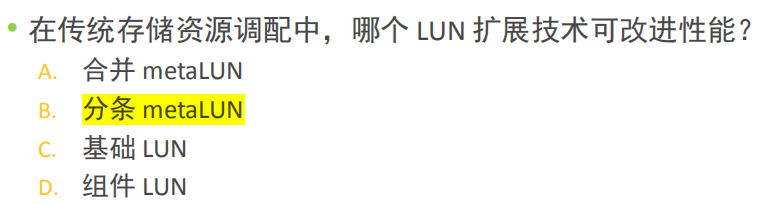
## 测验题目



数据存到缓存后便可认为数据已经到达，此时主机便可进行其他工作。（回写缓存）

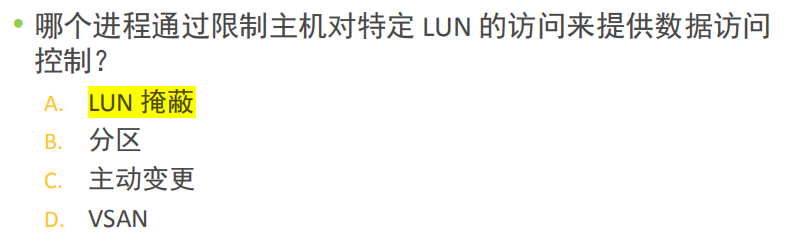


见缓存刷新，三种模式之一

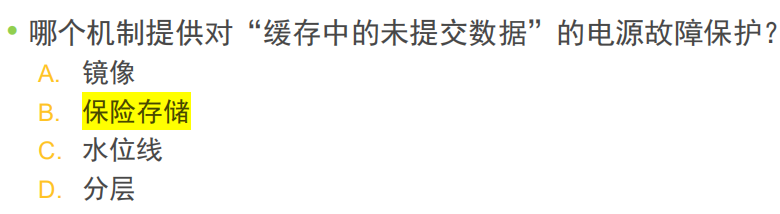


扩展慢，不代表不可改进性能

合并是仅提供容量，扩展快；分条是提供容量和性能，扩展慢



见LUN掩蔽



见缓存保护，两种方式之一

# 模块五 光纤通道存储区域网络 (FC SAN)

## 学习成果

• 介绍 FC SAN 及其组件

• 介绍 FC 体系结构

• 介绍 FC SAN 拓扑和分区

• 介绍 SAN 环境中的虚拟化

## 1.1 FC SAN概述

SAN：存储区域网络，可解决的挑战：

有项业务用户提供即时信息，提供灵活适用的存储基础架构

定义：SAN是服务器和共享存储设备的高速、专用网络。

集中化

FC体系结构概述

高性能低协议的开销，长距离的高可扩展性功能

利用FC网络可传输SCSI数据

FC中的层：

各层含义只需要了解即可

交换结构中的FC寻址：FC会分配到节点，节点用于在FC SAN中的节点之间进行通信。

domain id是结构中每格交换机的唯一编号（可用239个地址）

WWN：全球通用名

是唯一的64位标识符

相对于FC网络上的节点端口是固定的

FC数据科组职位三中结构，其中帧是数据传输到基本单元，包括五个部分（从n端口到f号端口建立会话）

连接结构服务不必记

链接中的登录类型：

连接结构登录FLOGI，通常的选择

发生在n到f端口之间

FC网络的拓扑和分区

拓扑：

完全互连和部分互连

核心－边缘拓扑

包括边缘和核心交换几层，网络通信遍历核心层或在核心层终止

分区：

FC交换机的一种功能，使连接节点的

和哪个节点联系就要和对方建立分区（画一个zone），有了分区才能真正通讯，主要是为了安全进行分区

分区方式

一种是利用WWN服务器，将存储阵列和服务器的WWN联合起来建立分区

一种是利用端口分区：将两者和FC交换器连接的switch domain id和端口id结合建立分区。（连接稳定时使用，因为依赖端口）

一种是结合进行，将一方的WWN和一方的switch+port建立分区

SAN中的虚拟化：

SAN中提供虚拟化层，抽象化数据块存储设备并通过聚合LUN创建存储池。

虚拟卷从存储池创建并分配到主机。

# 模块补充

虚拟化——计算

# 模块7 网络连接存储（NAS）

文件共享：

NAS设备，一般就用于文件共享，一个服务器红通常只需一组NAS便可实现，一般的服务器则需要至少两个。

网络文件系统

客户端-服务端。

# 模块8 基于对象的存储和统一存储

基于对象的存储设备OSD

关键组件：OSD节点，内部网络，存储