《大话存储》

Notes

2.0版 清华大学出版社

PCI 总线

高速 PCI 2.0 直接接北桥

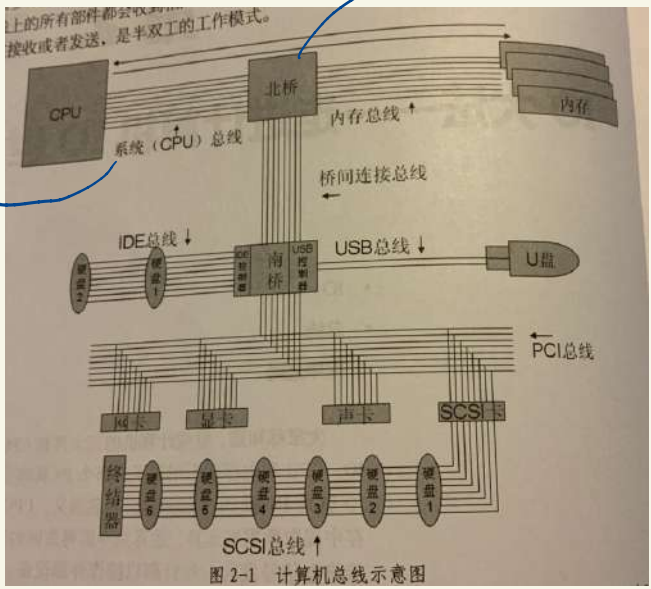


图 2-1 计算机总线示意图

前端总线（快）

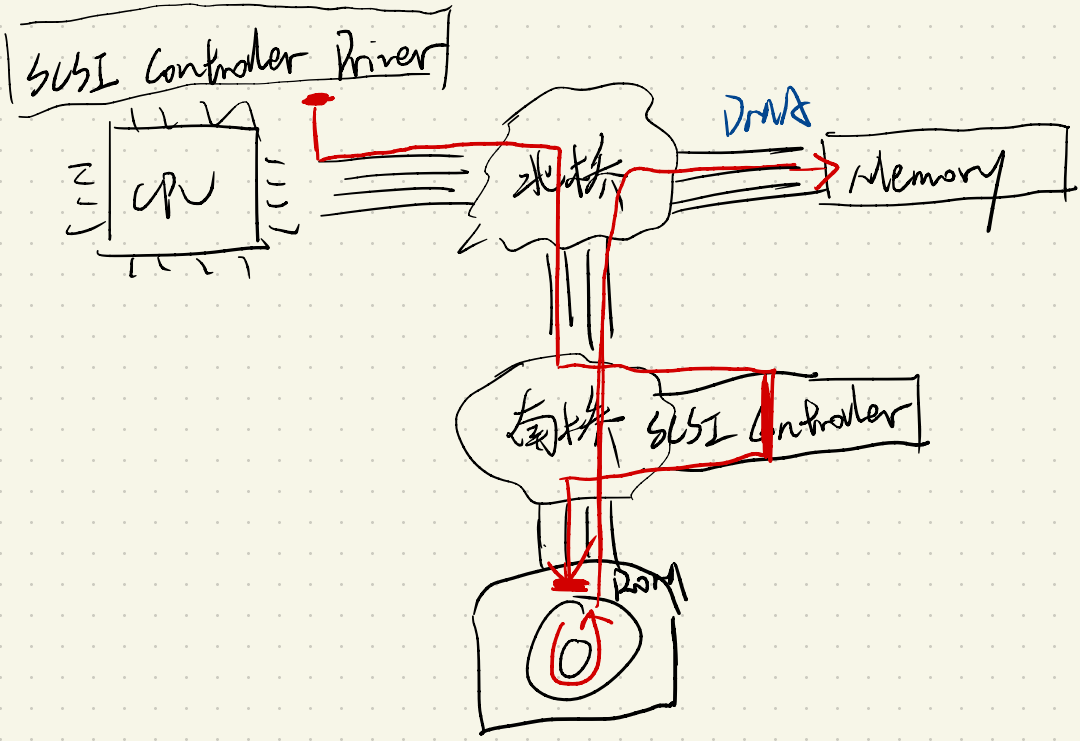
I/O总线（慢）

总线（bus）= 导线集合
每根导线 半双工

- 数据总线：决定位宽
- 控制总线
- 仲裁总线（中断）：决定某时刻某导线谁来使用
- 地址总线

PCI Multiplex + 中断共享

网络 = 连 + 找 + 发



向 SCSI 硬盘某地址读取的 Data flow.

P21 磁盘

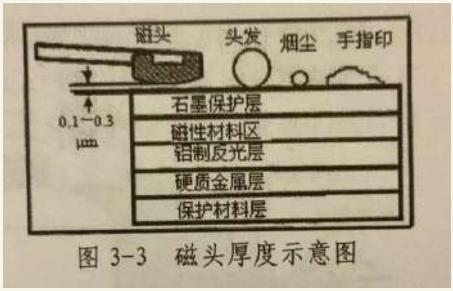


图 3-3 磁头厚度示意图

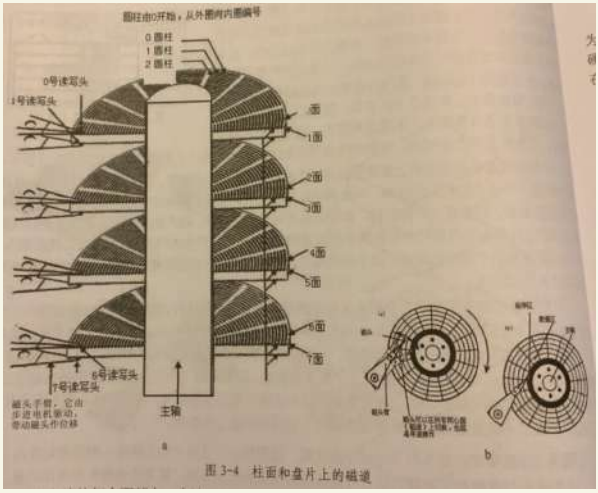


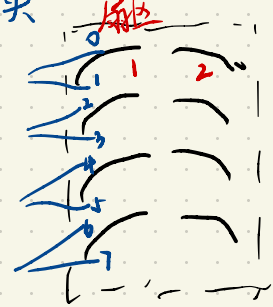
图 3-4 柱面和盘片上的磁道

引磁道(最外)存放的代码

读写最小单位为扇区 512B / 4KB

- 低级格式化：盘面上划分磁道-扇区
- 高级格式化：进行文件系统的标记

磁头



柱面

RW顺序：

柱面(磁道)⇒磁头(盘面)⇒扇区

$$Delay = T_{寻道} + T_{旋转} + T_{读取}$$

※

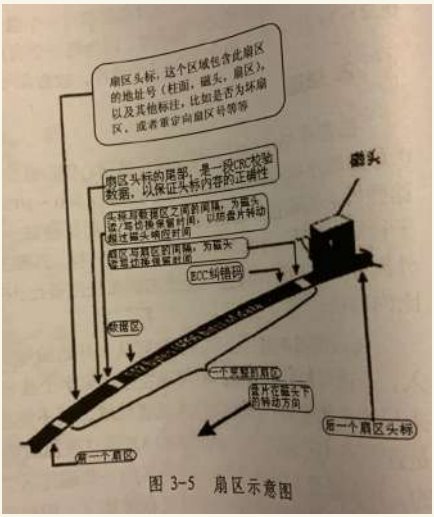


图 3-5 扇区示意图

CHS地址(内部)

LBA地址(对外线性)

Virtualization!

交叉因子编号：

防止扇区间隔间控制器来不及处理余留数据

P32 磁头 Scheduling

FCFS (First Come First Serve)

SSTF (Shortest Seek Time First)

SCAN / C-SCAN (回旋扫描)

LOOK / C-LOOK (回旋不到头)

P31 Drivers

BIOS中内置简化版可用驱动，启动完OS后再load OS中的复杂完整的驱动

$$IOPS = \frac{\text{Queue Depth}}{\text{IO Latency}}$$

n×的PCIE ≈ n·2.5 Gb/s.

高IOPS 在随机大量小RW时才优势明显

高Bandwidth 在传输大块连续数据时优势

二者不对立，可以兼有.

P57 SSD 固态硬盘

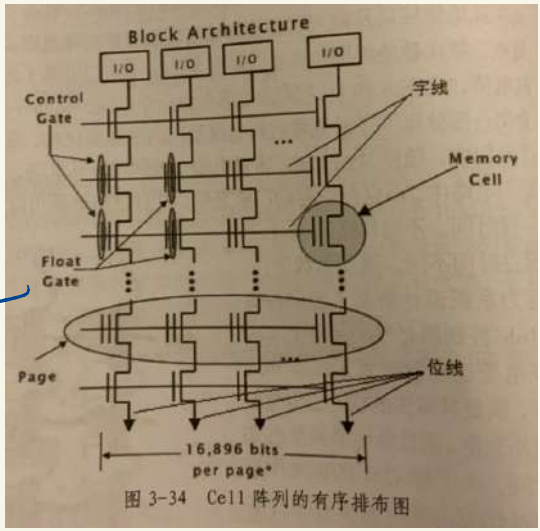


图 3-34 Cell 阵列的有序排布图

去突充电为"0"
平时保持"1"

* 最小R单位 1个Page
最小W单位 1个Block

MLC (Multi-level) 比 SLC (single-level) 单位成本低但更不robust

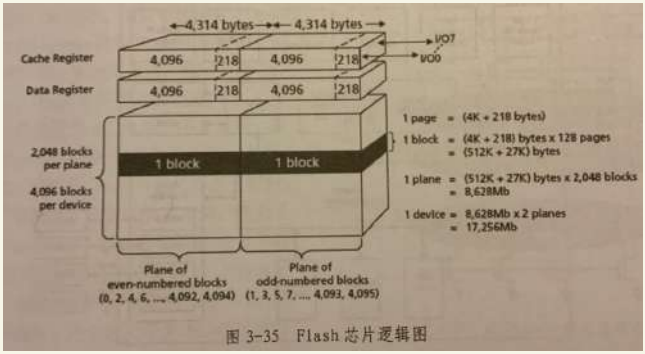


图 3-35 Flash 芯片逻辑图

示例规格：
一行为一个page = 4KB + 218B cells
一个Block = 128 Pages
奇/偶 Plane = 2,048 Blocks (各自)
一个SSD Device = 2 Planes + ...

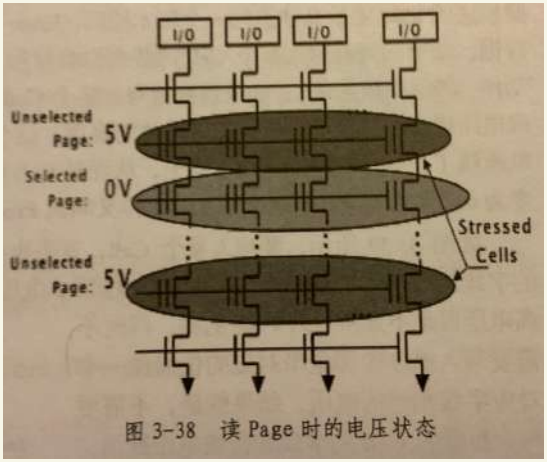


图 3-38 读 Page 时的电压状态

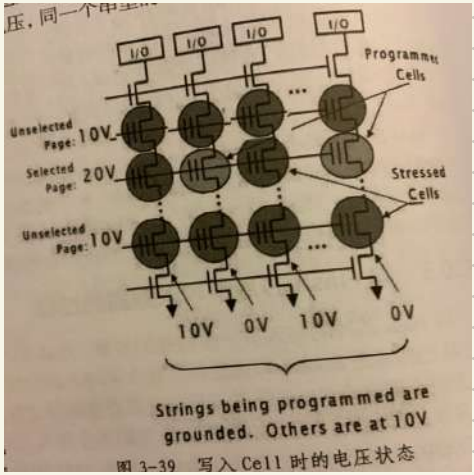


图 3-39 写入 Cell 时的电压状态

R

给位线 pre-charge ⇒ 让其自然漏电

⇒ 充了负电荷的 cell 漏电电流越小

⇒ 一段时间后电压较高 ⇒ $V_{ref} < V_{bit}$

⇒ 逻辑"0"

W

Erase 整个 Block ⇒ Program

(全部放电为1)

原因：寿命，1个 cell 擦写次数有限，

所以用 Append 写的方式平衡，R,W

使程序 (update) ⇒ Append + 重定向。

写放大 Write Amplification

缺点：随机大量小 W 时性能奇差

垃圾回收 GC

"就是钱的问题。做到SRAM那样的byte addressable会单位成本很贵。"(面积也大了)

NOR Flash即独立位线，可独立寻址但面积更大，写放率也低了。

Free space保留是缓解的方法之一(4%~20%)
Delayed Write等等 ---- Flash-aware FS ---

P68 处理cell损坏

参数 8b@512B ⇒ 每512B需配合可纠8bits错误的纠错码

P78 RAID Redundant Array of Inexpensive Disks / Independent

https://zh.m.wikipedia.org/zh/RAID

软件RAID vs. 硬件RAID卡
(集成于SCSI卡/南桥或独立成卡)

校验型RAID需要初始化

P140 卷管理 — LVM机制

物理盘（或RAID Controller暴露给OS的逻辑盘，OS还认为其为物理盘），称"物理卷"

⇓

多个物理卷组成一段逻辑上连续的 Volume Group

⇓

VG又被分割为数个"物理区块" Physical Partitions

⇓

PP被对应到"逻辑区块" Logical Partition

{ 1 PP → 1 LP
  多个PP → 1 LP

⇓

多个LP 组成一个逻辑卷 Logical Volume

mk的单位 ↓

好处：可以随时在线地增减LV大小或创建新的，不受限于物理磁盘或RAID配置

例：/boot 挂载到 sda 的前 20 个磁道
成为设备 sda1，不参与 LVM
而后 sda2，sdb，sdc --可捆在一起被分割
映射层级越来越多，更方便但也更复杂依赖
Trade off

P117 文件系统
在逻辑卷上又提供一层「文件」的抽象能力
比较熟悉，skip之.

P161 磁盘阵列 — storage node 独立

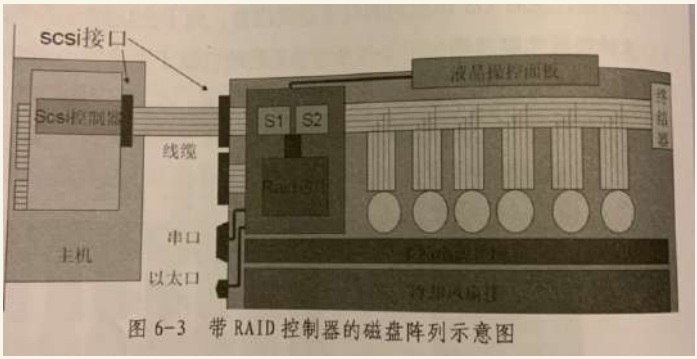


图 6-3 带 RAID 控制器的磁盘阵列示意图

· 前一后端；对外接口灵活性
· 人们把不论几块的硬件层面生成的虚拟磁盘均称为"LUN"；软件层面生成的则称"Volume"(卷)
· 双控制器⇒"脑裂"问题 (split brain)

P170 SAN (Storage Area Network)

独立的磁盘阵列和连上自己的复杂控制器，成为了一个单独于主机 cluster 之外的存储 cluster，即称 SAN.

P186 Fibre Channel, 略

引入包交换网络 ⇒ 扩展性↑

面向无连接 v.s. 面向有连接

+ 资源高效复用　　- 资源浪费，维护成本

- 数据丢失，鲁棒性　　- 可用性低

+ 保障到达

串行在高速底层连接上 > 并行传输

P228 SAS (Serial-Attached SCSI)

Speed Contest (平均速率)

- 机械 HDD
  - SCSI 1.2Gbps
  - SATA 1.5Gbps
- FC ~ SAS　　6G ~ 8Gbps
- 固态 SSD　　已上 3GB/s

ref

- IB 已上 40Gbps
- PCIe 4.0 16Gbps

P294 NFS 协议 ⇒ NAS (Network Attached Storage)

SAN over FC $\xrightarrow[\text{performance}]{\text{cost}}$ NAS over Ethernet

DAS 即原始的 Directly Attached Storage

P310 阶段总结，值得一看

P384 Virtualization 简读

带内 (In-band) v.s. 带外 (Out-band)

↓ 控制指令与数据共享线路

e.g. IP headers, NAS

↓ 控制指令通路独立

e.g. Bus, SanFS

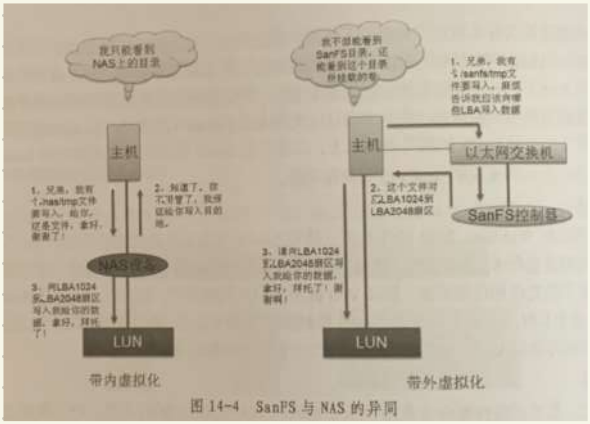


图 14-4 SanFS 与 NAS 的异同

P414 集群回顾

- 高可用性 HAC
- 负载均衡 LBC
- 高性能 HPC

P447 Scale-up v.s. Scale-out

↓ ↓

单机内堆砌 集群化，互联更多

更多更高的 较弱的单机

算力 ☆

P467 OSS 对象存储：数据库化的外置存储

P486 数据保护

· 块级保护 — 磁盘镜像

· 文件级保护 — 文件复制/备份

快照（Snapshot）技术

* clone

↓

只复制 metadata

（保证备份过程短，可用性个）

+

RoW / CoW

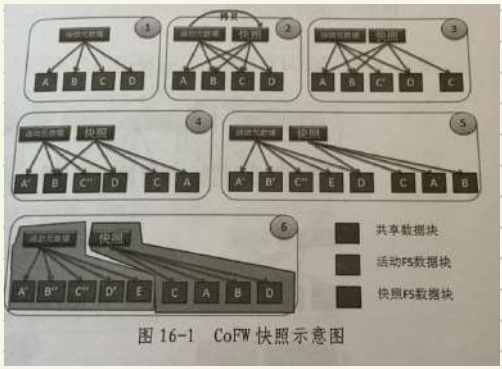


图 16-1 CoFW 快照示意图

P493 FS concepts： Bitmap, inode, ------

P5.14 连续数据保护 CDP

P5.16 备份系统总结
- 备份目标
- 备份通路
- 备份引擎
- 备份策略

P6.0
"容灾"不只包括安全的备份，还包括将备份数据尽快恢复生产状态。（同步？）

同步 vs. 异步复制，------

P6.8 数据管理 | 数据存储的软硬分离

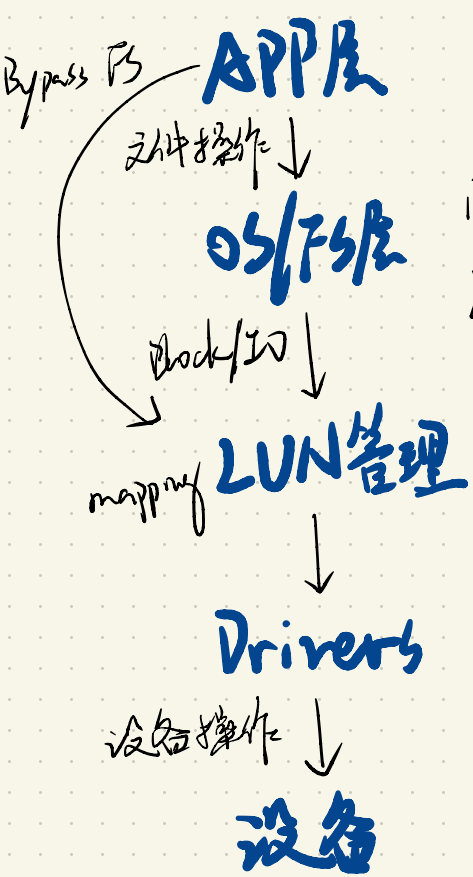
Data Center 之
包括
- snapshot, clone
- Recovery (sync)
- Virtualization
- Over subscription,
- De provision
- Deduplication (相关)
- ...

包括
- 介质
- 硬件冗余
- ...

Tiering

应用级 - 文件级 - 块级，充满Trade off

P748 IO路径：



（可看出各级均有缓存问题 trade off）

P806 IO性能诊断！值得一看并借鉴，定位Bottleneck在哪里

P8H 对"云"的定义显得过于宽泛了，
个人认为缺点些反而好，不妨作概念。
胜在灵活、节约、共用
弱在安全、性能（网络带宽之限）

P9H **分合** 的总结很到位

后记 一些补充和展望，值得一看。（2019）