外存储器

2022年秋

内容提要

- 硬盘存储
- RAID技术
- SSD存储

非易失性存储

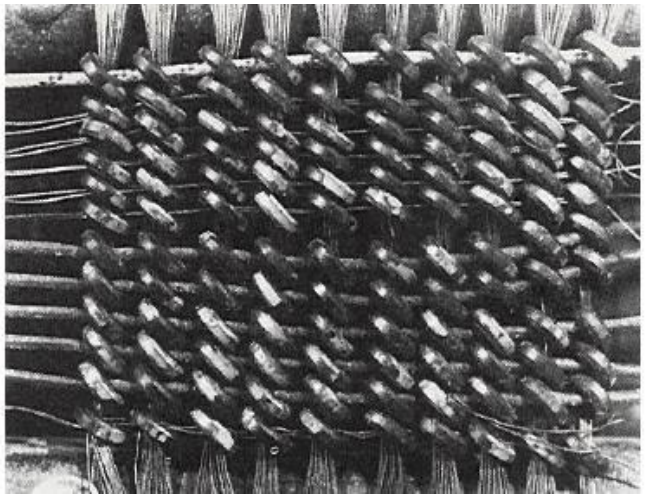
- 易失性存储：
  - 静态存储器：SRAM，Cache
  - 动态存储器：DRAM
  - 特点：快速，掉电后信息丢失，访问粒度小（字节，缓存块）
- 非易失性存储器：
  - 磁盘，磁带：磁表面存储器
  - 光盘
  - SSD，固态存储器
  - 特点：慢速，掉电后信息不丢失，访问粒度大（以数据块为访问单位）

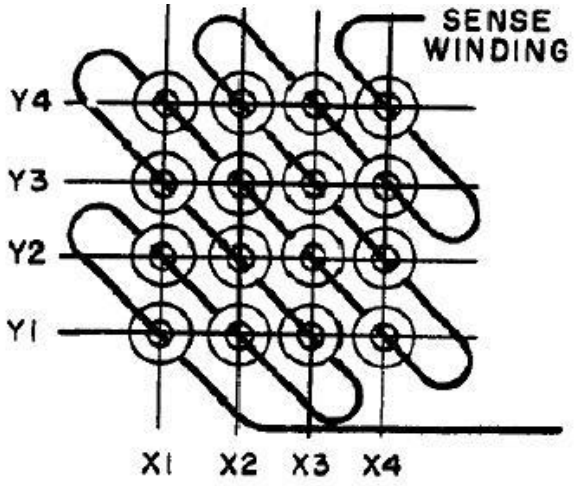
磁芯存储器

- 圆柱型陶瓷上涂磁粉
- 手工穿线，水手结
- 消磁后重写

- 存储原理简单
- 工艺复杂
- 可靠性低

- 大存储容量
- 成本低廉
- 断电后保存数据





磁表面存储设备

- 磁颗粒的不同偏转方向来区分不同的状态
- 主存中存放CPU要立即访问的程序和数据
- 辅助存储器中存放CPU不立即使用的信息，在需要时再调入主存中
  - 一般为磁盘、光盘等
  - 容量大、成本低、断电后还可以保存信息，能脱机保存信息，弥补了主存的不足
  - 串行访问、数据交换频率低、数据交换量大

随机访问和串行访问

- 随机访问
  - 随机访问任何单元，访问时间与信息存放位置无关
  - 每一位都有各自的读写设备
- 串行访问
  - 顺序地一位一位地进行，访问时间与存储位的物理位置有关
  - 共用一个读写设备
  - 顺序访问和直接访问

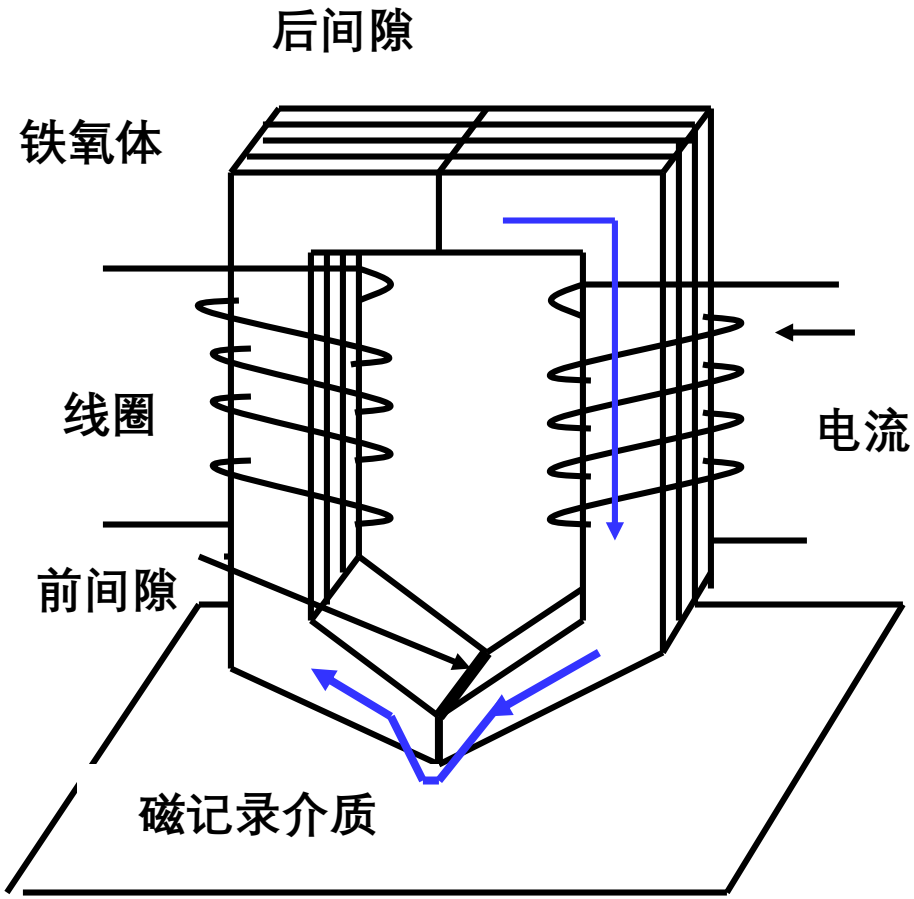
主要技术指标

- 存储密度
  - 单位长度（磁带）或单位面积（磁盘）磁层表面所存储的二进制信息量
- 存储容量
  - 磁表面存储器所能存储的二进制信息的总量，以字节为单位
- 寻址时间
- 数据传输率
- 误码率
- 价格

磁表面存储设备

- 如何保存?
  - 磁颗粒的不同磁化偏转方向
- 如何表示?
  - 磁记录方式
- 如何组织?
  - 扇区、磁道、柱面、硬盘
- 如何管理?
  - 操作系统的文件系统

磁记录原理



磁头，软磁材料
导磁率高，饱和磁感应强度大
矫顽力小，剩余磁感应强度小

磁记录材料，硬磁材料
记录密度高，记录信息时间长
输出信号幅度大，噪声低
表面组织紧密、光滑、无麻点
薄厚均匀，温度、湿度影响小

磁头结构和电磁转换示意图

磁记录方式

- 磁记录方式
  - 指一种编码方法，即如何将一串二进制信息，通过读写电路变换成磁层介质中的磁化翻转序列。
- 评价标准
  - 编码效率
    - 表示一个二进制位数据需要使用多少个磁颗粒？
  - 自同步能力
    - 读写时准确定位二进制数据位的能力
  - 读写可靠性

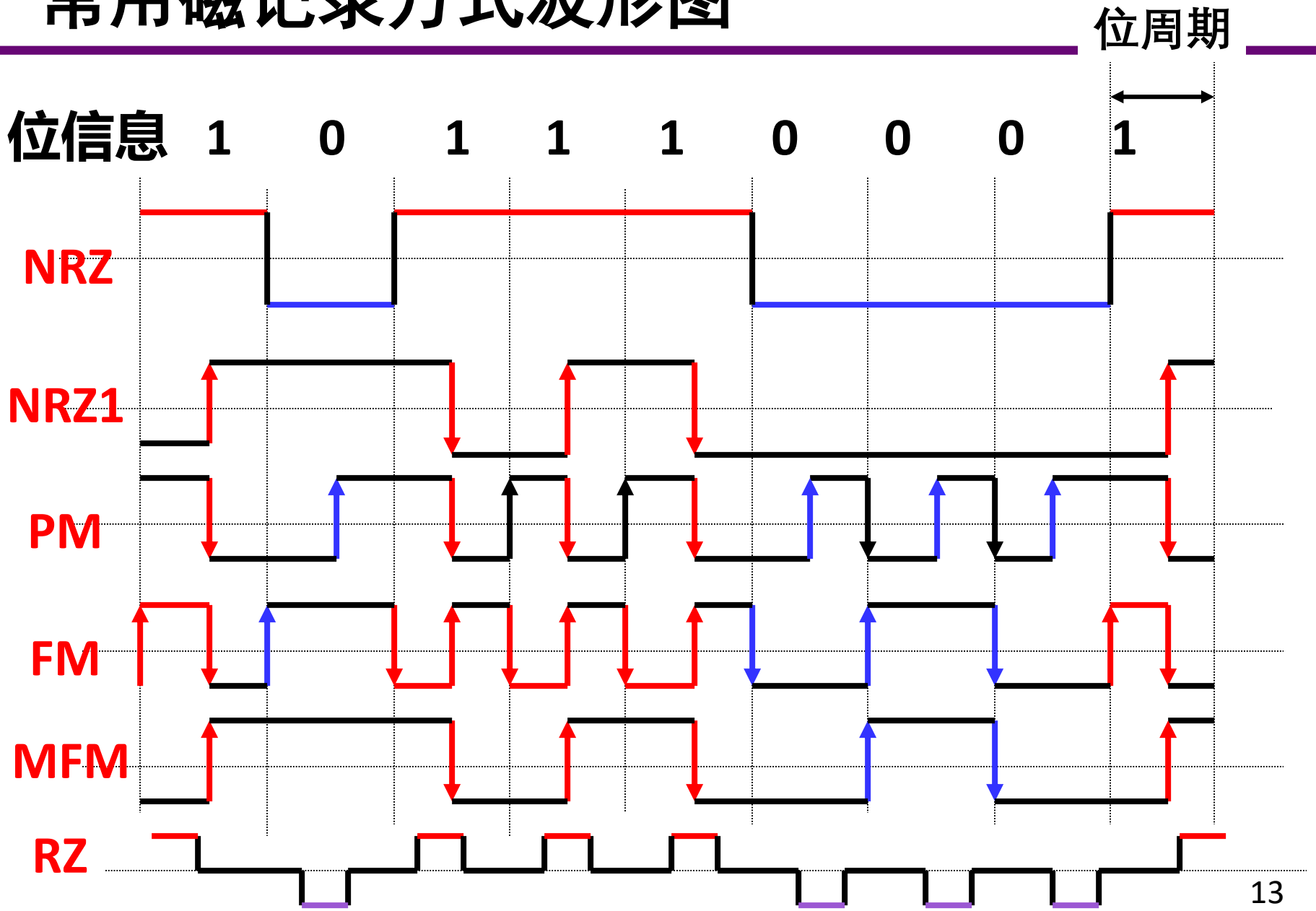
磁记录方式

- 归零制（RZ）
  - 线圈中正脉冲为“1”，负脉冲表示“0”，两位信息位之间线圈中电流为零。
- 不归零制（NRZ）
  - 线圈中一直有正或负脉冲（包括两位信息位之间）。
- 见1翻转的不归零制（NRZ1）
  - 只有见到“1”才改变电流的方向

磁记录方式

- 调相制（PM）
  - 用脉冲的边沿来表示“0”和“1”
- 调频制（FM）
  - “1”:位周期中心和位与位之间都翻转
  - “0”:位周期中心不翻转，位与位之间翻转
- 改进的调频制（MFM）
  - 只有连续两个或以上的“0”时，才在位周期的起始位置翻转

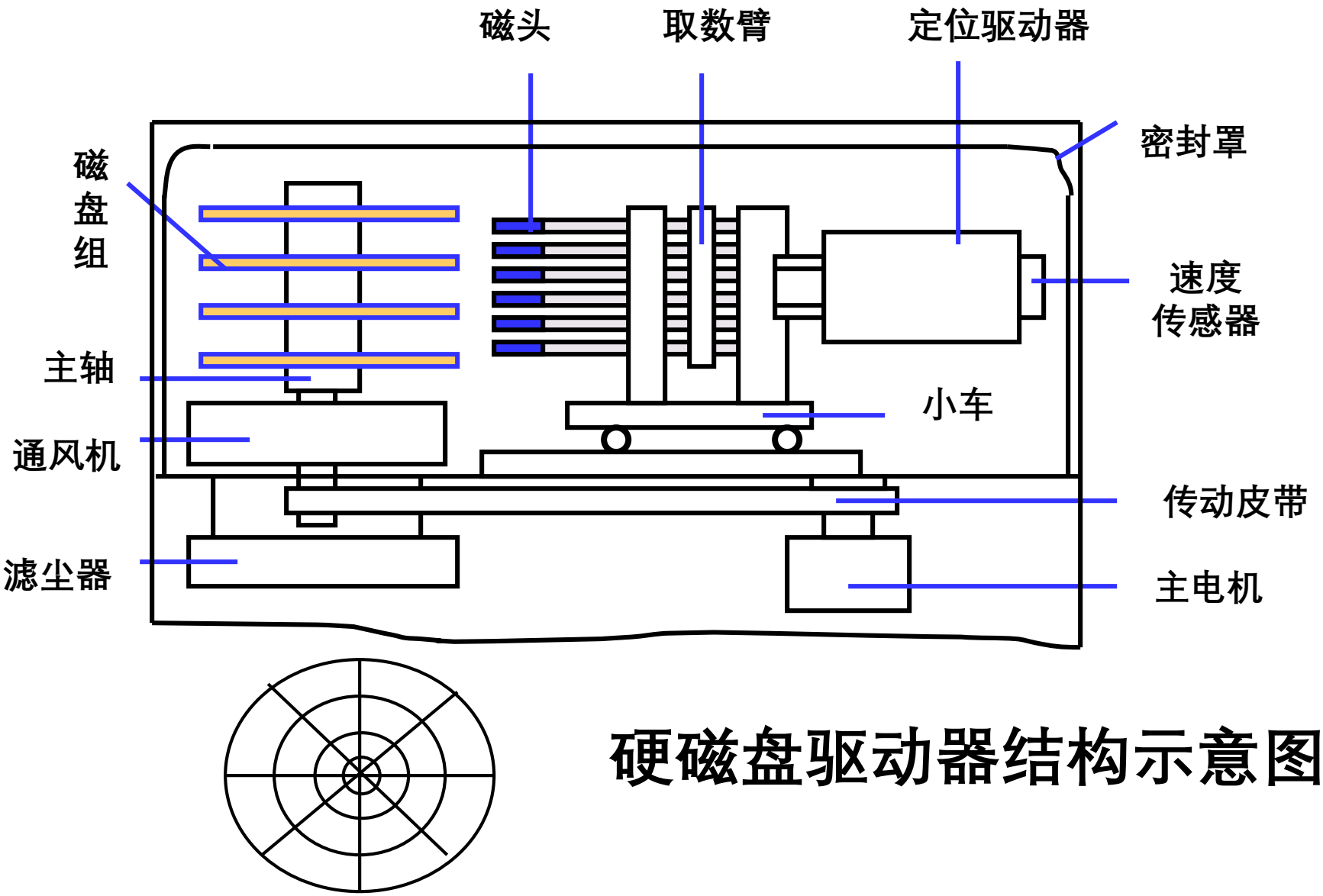
常用磁记录方式波形图



磁盘

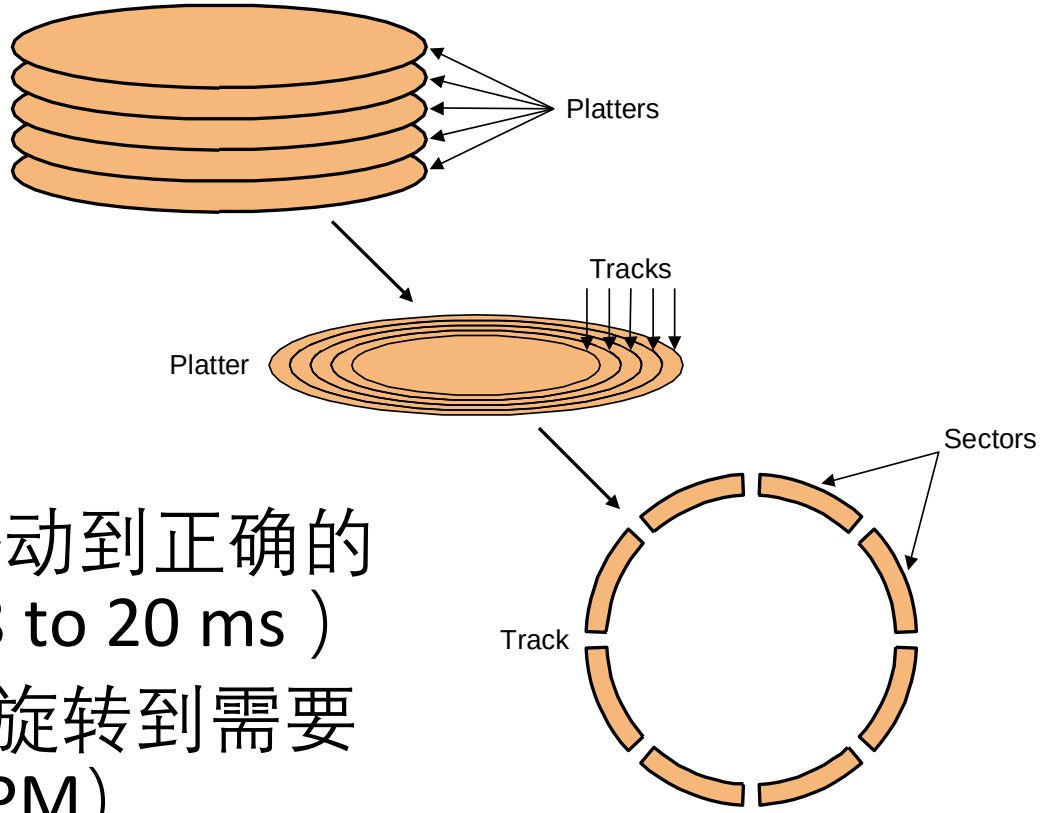
- 目的
  - 长期存储、断电后存储
  - 容量大、价格低廉，但速度慢
  - 可用在层次存储器的最底层
- 特点
  - 使用旋转托盘上的表面磁颗粒来存储数据
  - 可移动的读/写头来访问磁盘
- 硬盘、软盘比较
  - 硬质托盘（金属铝），面积可以比较大；
  - 由于可被精确控制，密度可以更高
  - 旋转速度快，传输率高
  - 可以多个盘片组合

硬磁盘设备



硬磁盘驱动器结构示意图

硬磁盘内部结构



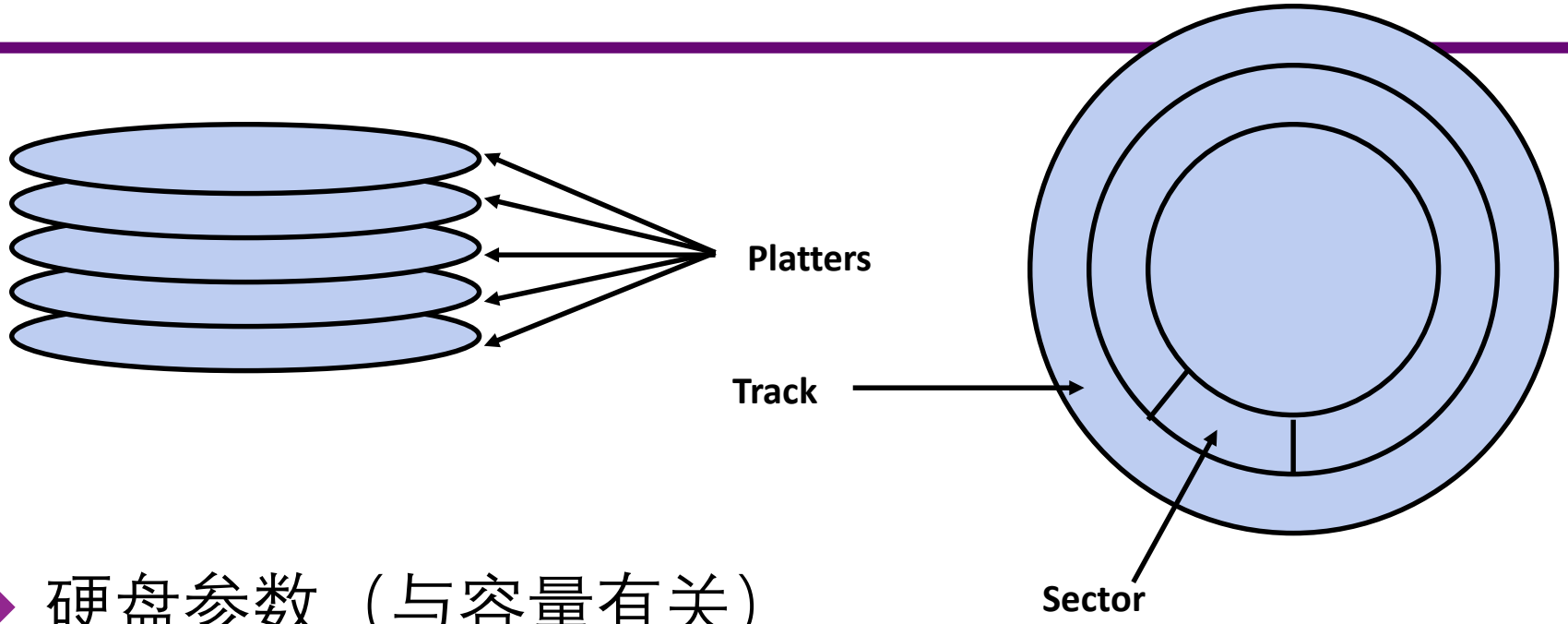
磁盘访问过程：

寻道：将读写磁头 移动到正确的磁道上（平均需要8 to 20 ms）

寻找扇区：等待磁盘旋转到需要访问的扇区（.5 / RPM）

数据传输：读写数据（1个或多个扇区）（2 to 15 MB/sec）

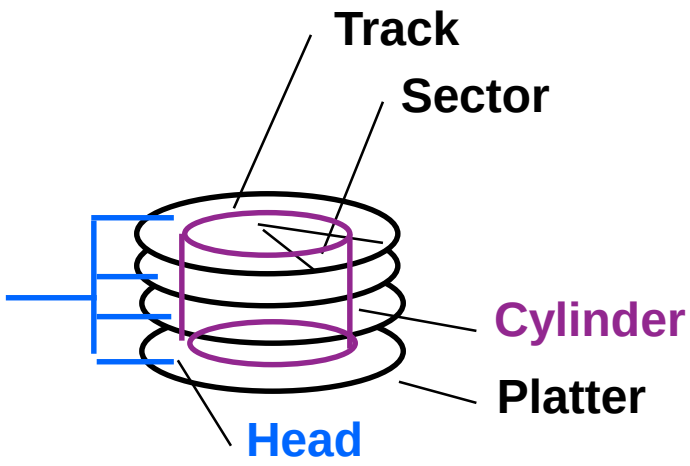
硬磁盘内部结构



- 硬盘参数（与容量有关）
  - 500 至 2,000 磁道（每面）
  - 32 至 128个扇区（每个磁道）
    - 扇区是磁盘访问的最小单位
- 早期硬盘上每个磁道上的扇区数是一样的
- 增加容量
  - 位密度不变：外磁道比内磁道扇区数多一些

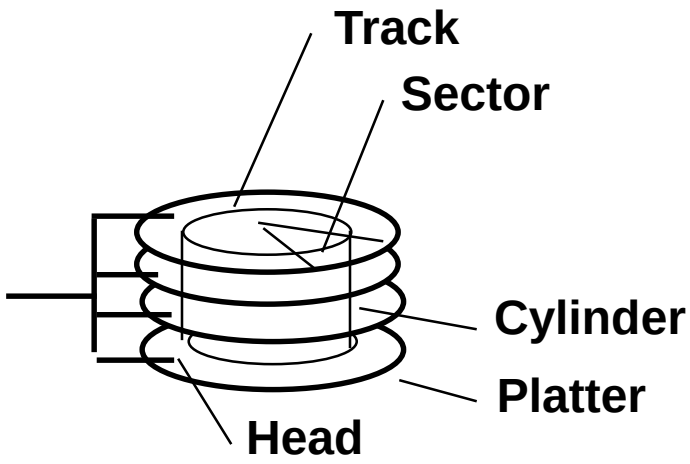
硬磁盘参数

- 柱面： 位于同一半径的磁道集合
- 读/写数据的三个步骤：
  - 寻道时间： 将磁头移动到正确的磁道上
  - 旋转延迟： 等待磁盘上扇区旋转到磁头下
  - 传输时间：真正的数据读/写时间
- 当前平均寻道时间：
  - 一般为 8 至12 ms



硬磁盘参数

- 旋转延迟:
  - 旋转速度: 3,600至7200 RPM
  - 旋转时间: 16 ms至8 ms每转
  - 平均寻址时间8ms至4ms
- 访问速度:
  - 数据量（通常为1个扇区）: 1 KB / sector
  - 旋转速度: 3600 RPM至7200 RPM
  - 存储密度: 磁道上单位长度存储的位数
  - 磁盘直径: 2.5至 5.25 in
  - 一般为: 2 至12 MB每秒



硬磁盘访问时间

- 磁盘访问时间 = 寻道时间 + 旋转延迟 + 传输时间 + 磁盘控制器延迟
- 举例:
  - 平均寻道时间 = 12ms;
  - 旋转速度 = 5400rpm
  - 磁盘控制器延迟： 2ms
  - 传输速度 = 5MB
  - 扇区大小 = 512 bytes
  - 读取一页（8KB）需要多少时间?

- 旋转延迟：平均旋转延迟应为磁盘旋转半周的时间。
- 旋转1 周 = 1/5400 minutes

  = 11.1ms => ½ 周： 5.6 ms
- 读1个扇区时间 = 12ms + 5.6ms + .5K/5MB + 2ms

  = 12 + 5.6 + .1ms + 2ms

  = 19.7 ms
- 读1页的时间=

  = 12 ms + 5.6ms + 8K/5MBpersec + 2ms

  = 12ms + 5.6ms + 1.6ms + 2ms

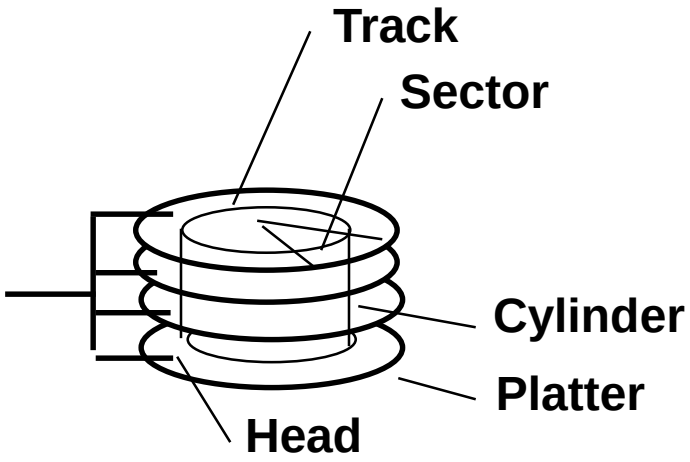
  = 21.2 ms

对磁盘访问的思考

页容量大，为什么扇区却如此小呢?

- 理由 #1：可用性。可以在扇区物理损坏时不再使用该扇区。
- 理由 #2：还是可用性。检错纠错码分布在每个扇区，扇区容量小，检错速度快，效率高。
- 理由 #3：灵活性。使用不同的操作系统，不同的页面大小。

- 采用并行方式和大容量传输方式克服磁盘控制器延迟
- 大容量传输：每次读取多个扇区，可以节约时间。
- 也可以分担部分总线延迟...
- 并行 #1：并行读多个层面
- 并行 #2：并行读多个磁盘



结论

- 应该记住以下两点：
  - 额外开销在总开销中比例较大 =>一次传输大量数据比较有效
  - 将页面存放在相邻扇区中可以避免额外的寻道开销

访问磁盘过程

- 对磁盘的访问总是由缺页引起的：
  - CPU给出地址，需要访问某存储单元；
  - 并行进行TLB查找和cache查找；
  - TLB查找后申明没有找到；
  - 停止并行查找，并通知操作系统处理；
  - 操作系统检查页表，发现该页不在内存中，需要从硬盘调入。应该如何进行呢？
- 操作系统从主存中选择一页准备换出，为调入的页安排存放空间；
- 若被换出的页是“脏”页，需要将其写回磁盘存储；
- 操作系统申请I/O总线；
- 获得批准后，发送写命令给I/O 设备（磁盘）。紧跟着传送需要写回的页的全部数据。
- I/O控制器发现发给自己的写命令，加入到握手协议，并接受数据。
- 根据数据要写入的地址，读/写头移动到正确的柱面，同时，将数据接收到缓冲区。
- 寻道结束后，等待相应的扇区旋转到磁头下面，将数据写入扇区中。
- 在写入数据间隙，计算校验码并写入扇区中。

磁盘访问过程

- 下一步，操作系统继续申请总线（如果还保持总线控制权，则不必申请）。
- 得到授权后，向磁盘发出读命令。
- 然后，磁盘识别地址，并转换为相应的地址段。
- 寻道，将读/写头移动到指定位置。
- 从指定扇区中读去数据，并进行校验。
- 磁盘申请I/O总线。
- 得到授权后，将数据通过总线送到内存。

可靠性和可用性

- 两个经常混淆的词汇：
  - 可靠性：设备出现故障的几率来衡量。
  - 可用性：系统能正常运行的几率来衡量。
- 可用性可以增加硬件冗余来提高：
  - 例如：在存储器中增加校验码。
- 可靠性只能通过下面途径提高：
  - 改善使用环境
  - 提高各部件的可靠性
  - 减少组成部件
    - 可用性的提高可能带来可靠性的降低

RAID的提出

- CPU性能在过去的十年中有了极大地提高，几乎是每18个月翻一番。但磁盘的性能却没能跟上。在70年代，小型机磁盘的平均查找时间为50到100毫秒，现在是10毫秒。在许多行业（如汽车或航空业），如果性能的提高能达到这个速度，即20年内提高5到10倍，那就会是头条新闻，但对计算机行业，这却成了一个障碍。因为CPU性能和磁盘性能间的差距这些年来越来越大。

RAID的提出

- 在提高CPU性能方面，并行处理技术已得到广泛使用。这些年来，许多人意识到，并行I/O也是一个提高磁盘性能的好办法。1988年，Patterson et al.在他的一篇文章中建议用6个特定的磁盘组织来提高磁盘的性能或可用性，或两方面都同时提高。这个建议很快就被采用，并导致了一种新的I/O设备的诞生，这就是**RAID盘**。Patterson et al.把RAID定义为**廉价磁盘的冗余阵列**（Redundant Array of Inexpensive Disks），但工业界把“I”由“廉价的（Inexpensive）”替换成“独立的（Independent）”。

RAID

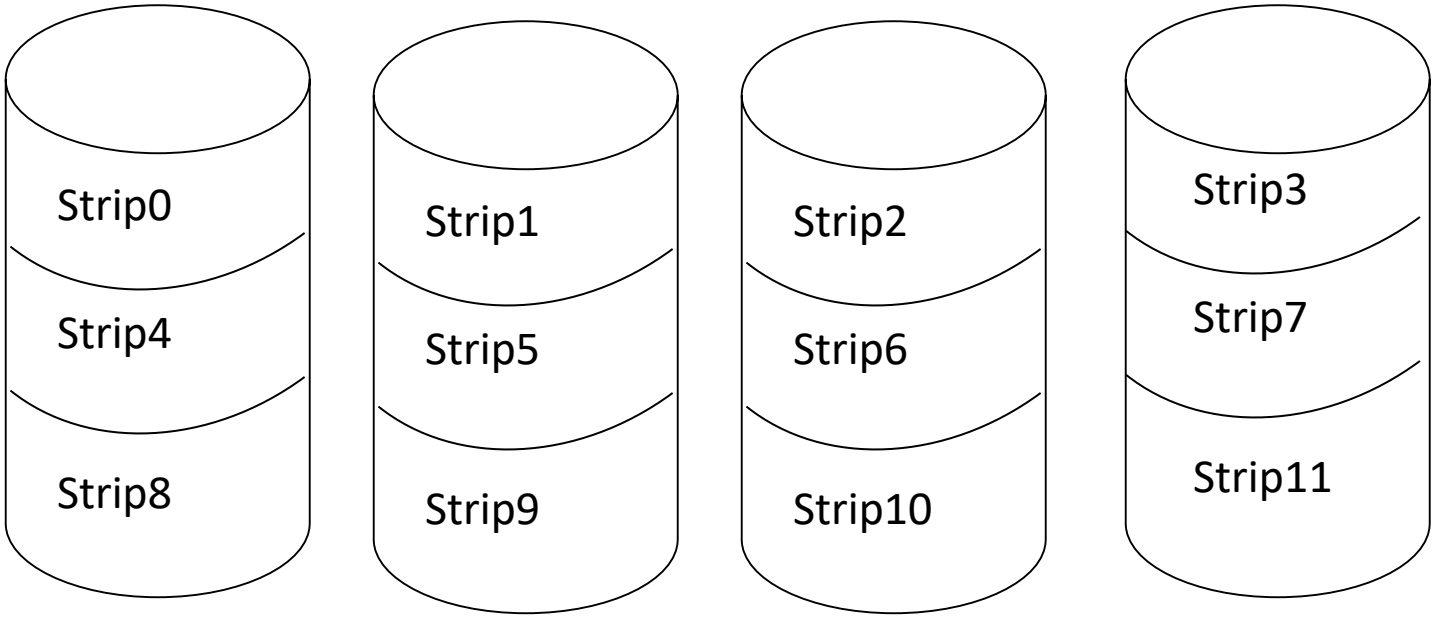
- RAID定义
  - 廉价磁盘的冗余阵列（Redundant Arrays of Inexpensive Disks）
  - 用N 个低价磁盘构成一个统一管理的阵列，以取代特贵单一磁盘
- RAID目标
  - N个磁盘的容量
  - 1/N的访问时间
  - 更高的性价比
  - 采用冗余技术提高存储信息的可用性

RAID0：DataStriping
RAID1：DriveMirroring
RAID4：Data Guarding
RAID5：Distributed Data Guarding

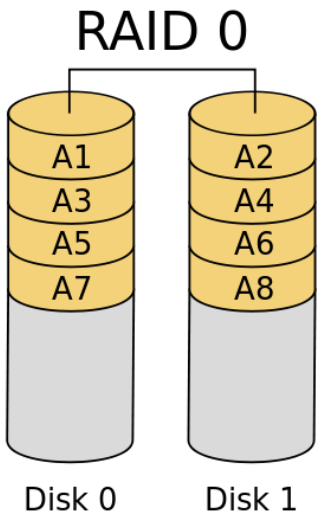
RAID0

- RAID0将由RAID模拟的单个虚拟磁盘划分成带（strip），每带k个扇区。第0带为第0到第$k-1$扇区，第1带为第k扇区到第$2k-1$扇区，等等。对k=1，每个带为1个扇区；对k=2，每带有2个扇区；等等。RAID 0以交叉循环的方式将数据写到连续的带中，下图描述的就是有4个磁盘驱动器的RAID盘。这种在多个驱动器上分布数据的方式叫作**分带**。如果软件发出从带的边界开始读四个连续带的数据块的命令，RAID控制器将把这个命令分解成四个单独的读命令，四个驱动器每个一个，让它们并行执行。这样，就实现了对软件透明的并行I/O操作。

RAID0



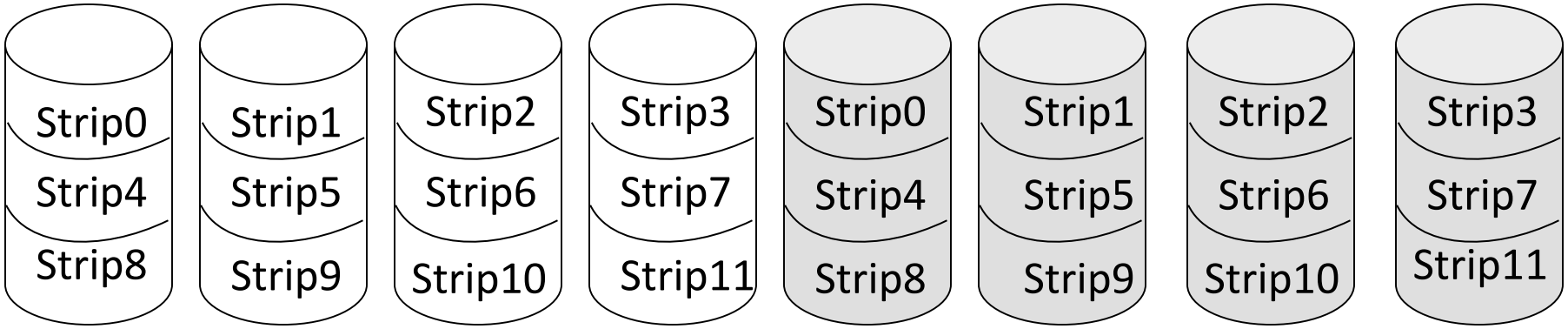
1. 适合数据请求量比较大的情况
2. 没有冗余，可靠性差，不算真正的RAID



RAID1

- 它复制了所有的磁盘，所以有四块主磁盘和四块辅助磁盘。每个对磁盘的写操作都进行两次，而每次读操作则可以读任意一个备份，把负载均衡分布到不同的驱动器上。这样，写操作的性能并不比单个磁盘好，但读磁盘的性能却比单个磁盘高了两倍。容错性能就更好了，如果一个驱动器崩溃的话，只要简单的用备份驱动器代替就行了。恢复整个磁盘的操作包括两个步骤：装上一个新的驱动器，然后将整个备份驱动器的内容拷贝到新的驱动器上。

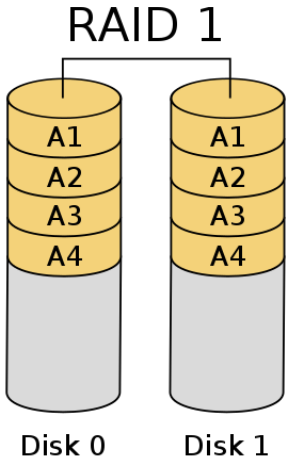
RAID1



1. 冗余备份，可靠性高

2. 写性能不高，但读性能却提高了两倍

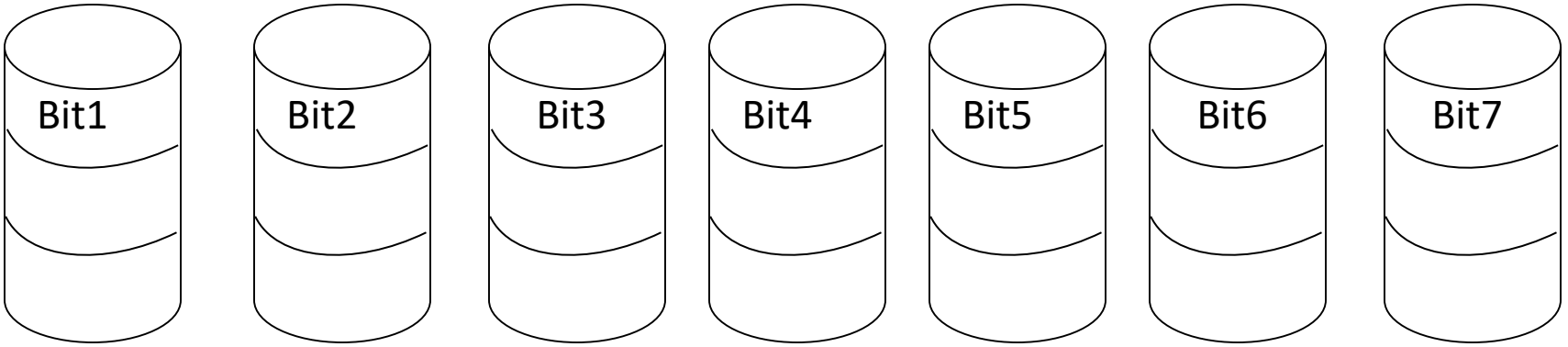
3. 成本较高



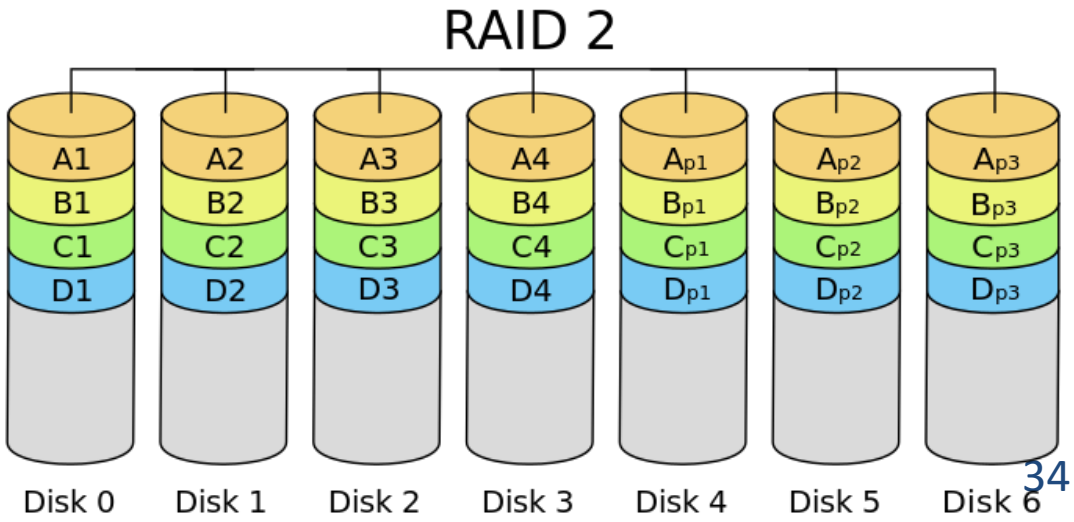
RAID2

- RAID 2的工作单位为字，可能的话甚至可以是字节。首先我们可以想象将单个虚拟磁盘上的字节分解成一对4位的半字节，对每个半字节加上3位海明码形成7位字，即其中1、2、4位做校验位。然后，用下图所示的七个驱动器的磁头和旋转同步，就可能将整个海明码字写在七个驱动器上，每个驱动器一位。

RAID2



1. 驱动器必须同步旋转
2. 驱动器个数要足够多
3. 需要多个控制器

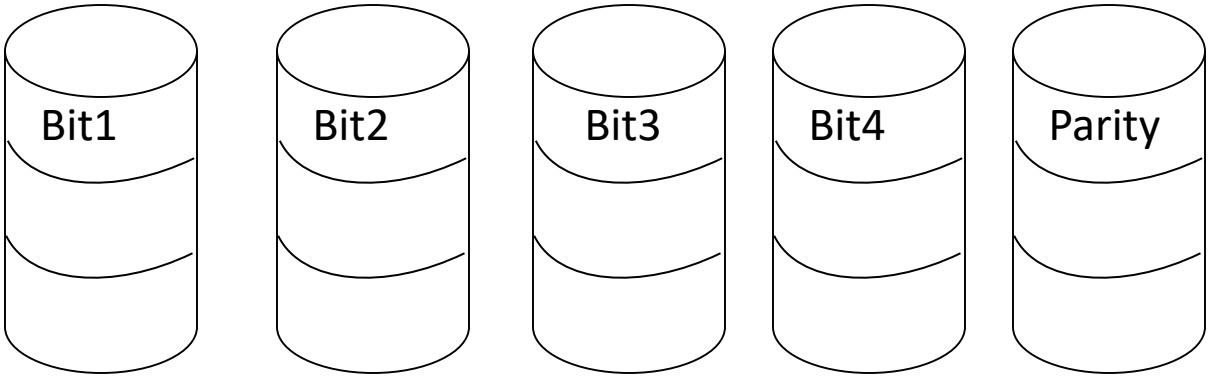


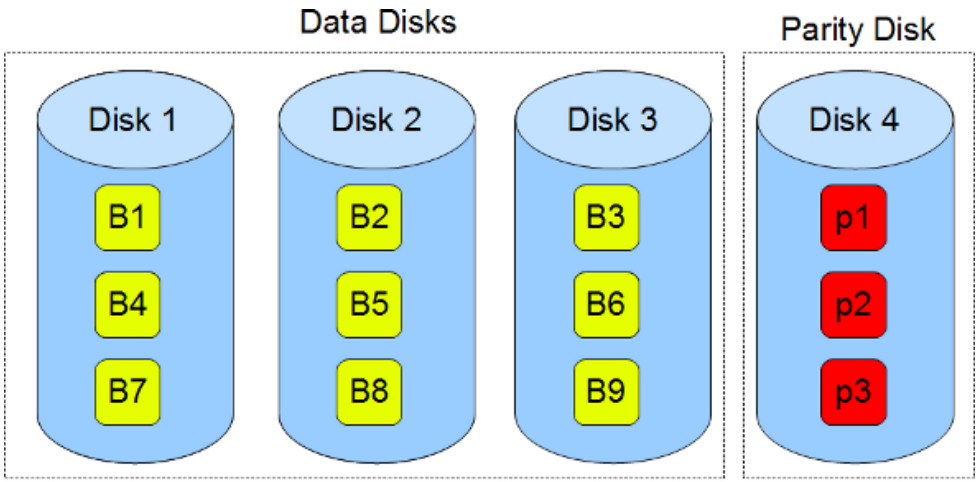
RAID3

- RAID 3是RAID 2的一个简化版本，它只需对每个字计算一个校验位，写到一个校验驱动器上。和RAID 2相同，驱动器之间必须严格同步，因为一个字被分布到多个驱动器中。

RAID3

1. 驱动器之间要严格同步
2. 对整个磁盘崩溃的错误，能够进行恢复





RAID 3 – Bytes Striped. (and Dedicated Parity Disk)

RAID4

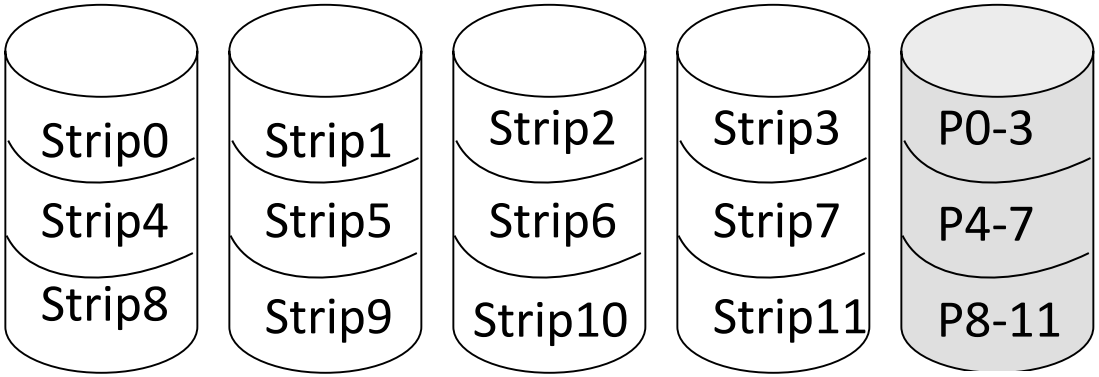
- RAID 4和RAID 0类似，将对带的校验写在额外的驱动器上。例如，若带的长度是*k*个字节，将所有的带异或到一起，产生一个*k*字节长的校验带。如果其中一块磁盘崩溃的话，它的内容可以从校验磁盘上重新计算出来。

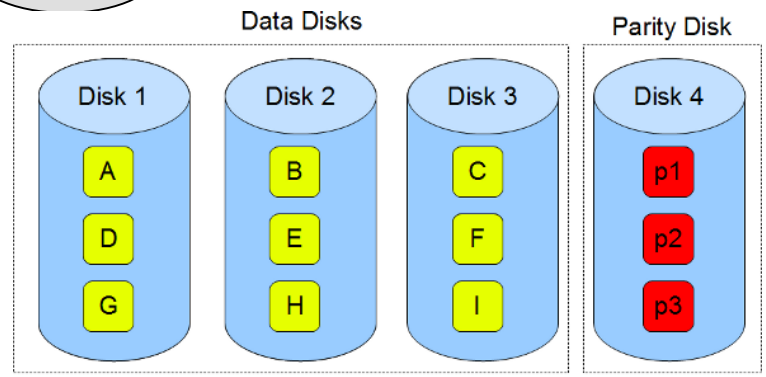
RAID4

1. 不对字进行校验，也不需要驱动器同步

2. 可以防止整块盘崩溃，但对盘上部分字节数据出错的纠错性能相当差

3. 校验盘负载沉重



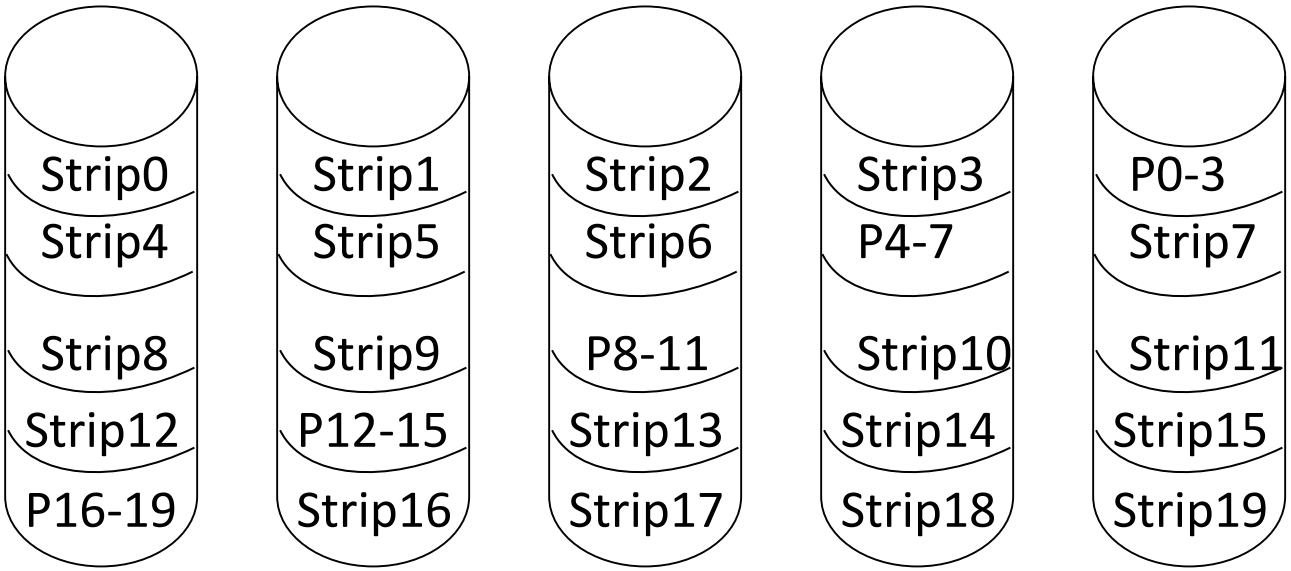


RAID 4 – Blocks Striped. (and Dedicated Parity Disk)

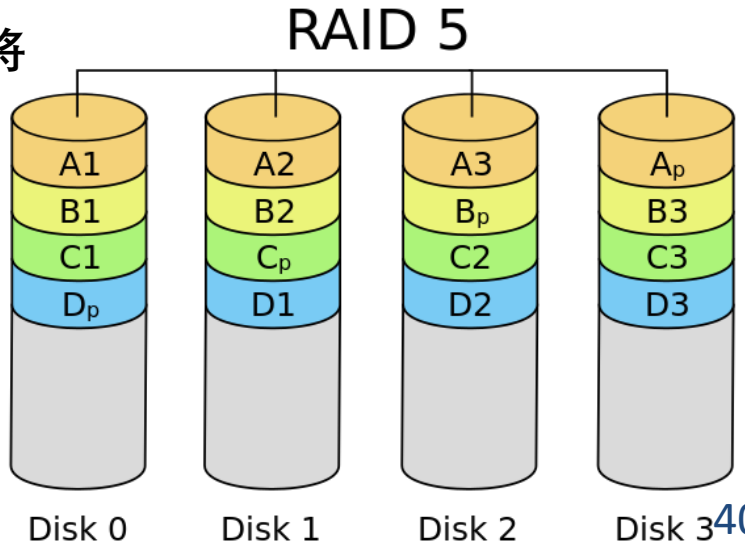
RAID5

- RAID 5为减少校验盘的负载，将校验位循环均匀分布到所有的驱动器上。

RAID5

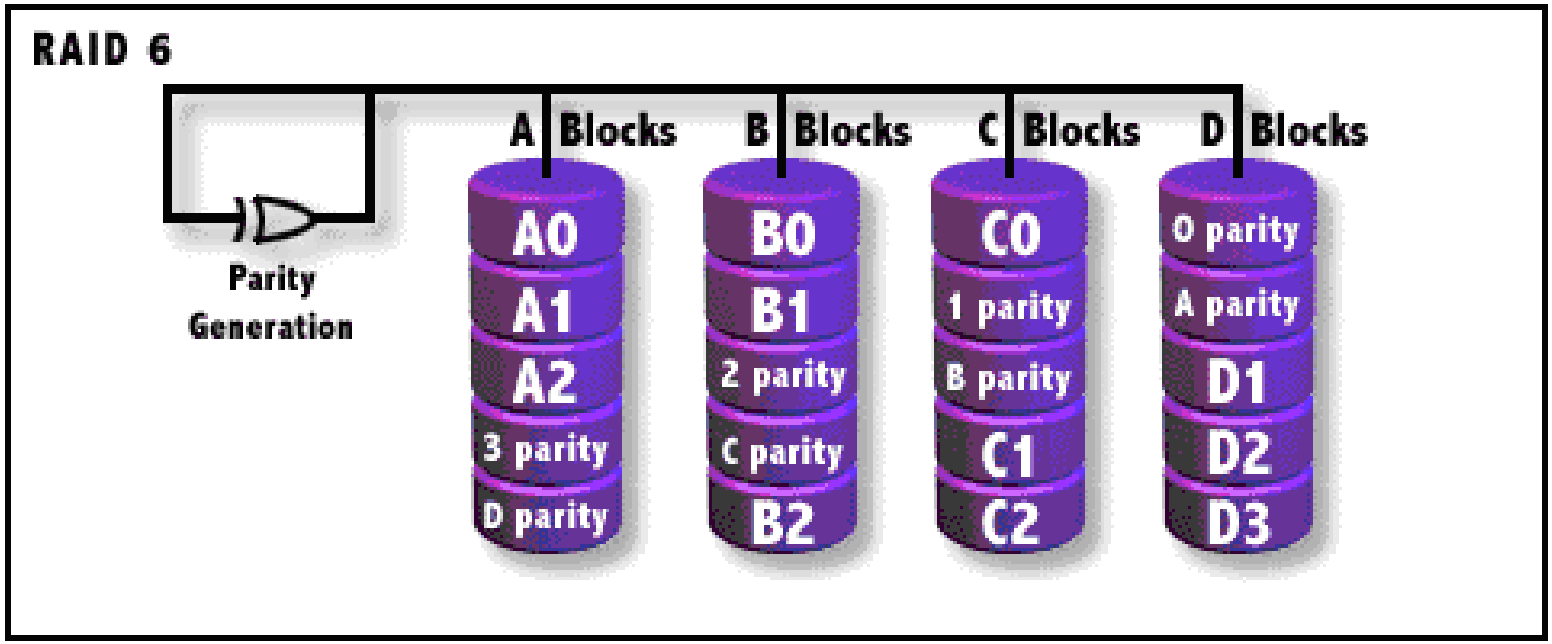


1.如果RAID 5的磁盘崩溃的话，修复磁盘内容的将是一个复杂的过程。

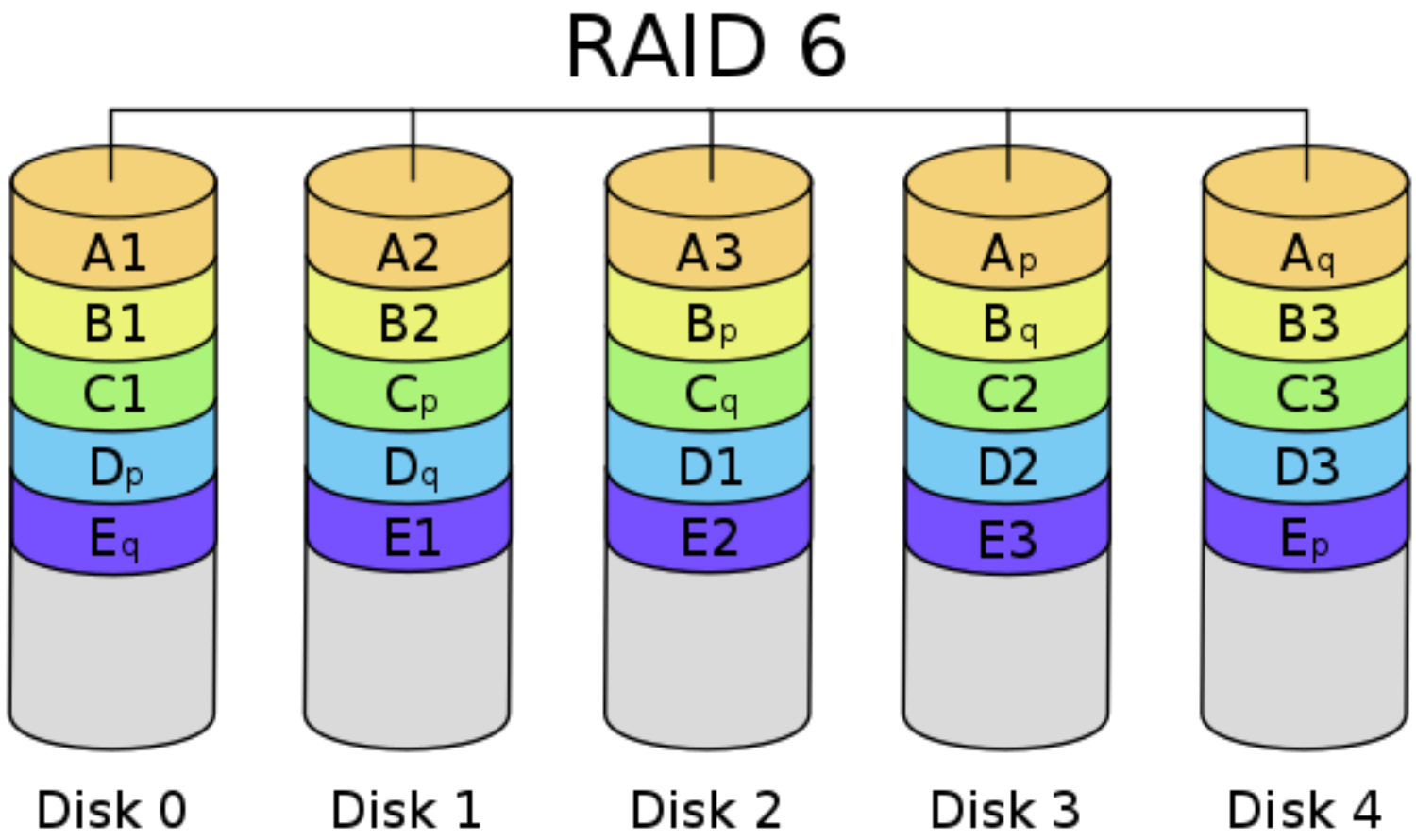


RAID6

- 如果两个磁盘出错呢？
  - Independent Data disks with two independent distributed parity schemes
  - 二维校验



RAID6

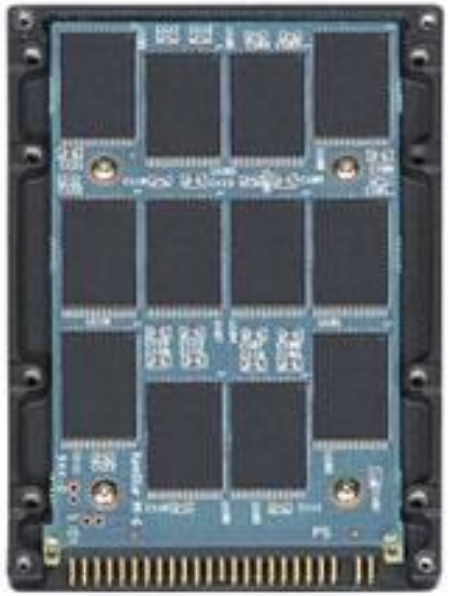

RAID 6
A1
B1
C1
Dp
Eq
A2
B2
Cp
Dq
E1
A3
Bp
Cq
D1
E2
Ap
Bq
C2
D2
E3
Aq
B3
C3
D3
Ep
Disk 0
Disk 1
Disk 2
Disk 3
Disk 4

磁存储小结

- 磁表面存储设备
  - 用磁颗粒的不同磁化方向表示0和1
  - 弥补了主存的不足
  - 磁盘存储原理及磁记录方式
- 磁盘的访问过程
  - 寻道、寻找扇区、访问
- RAID技术
  - 提高磁盘的可用性和性能

固态硬盘

- 固态硬盘没有机械结构，没有移动的部分
- 安静，低功耗，高性能，不怕摔，低发热
- 价格比硬盘高（将来会下降），有限的擦除次数

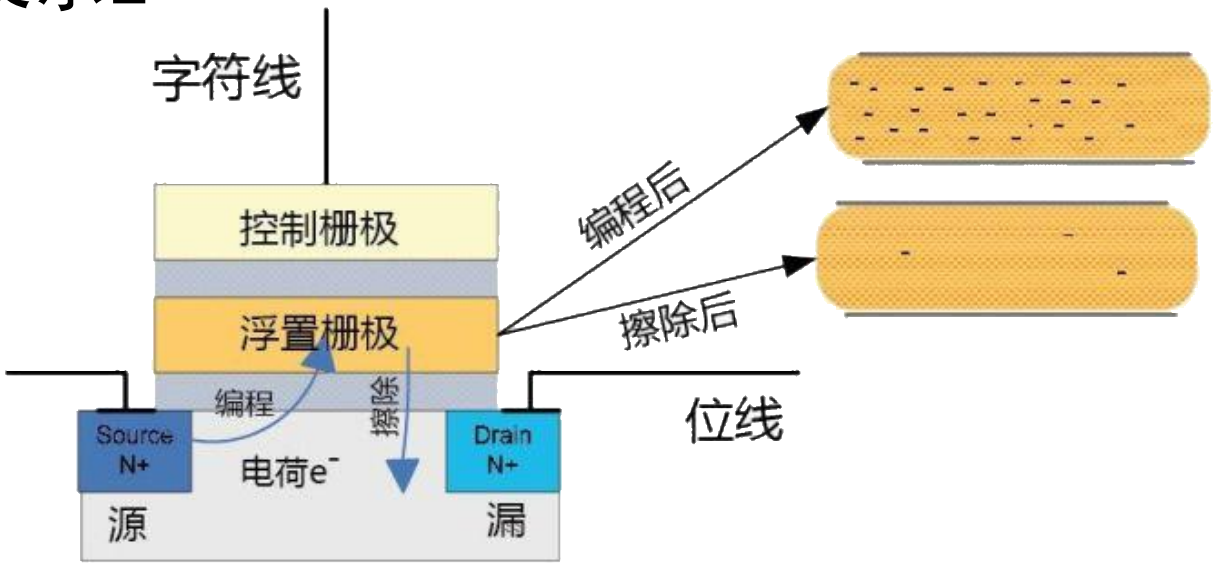


固态硬盘存储

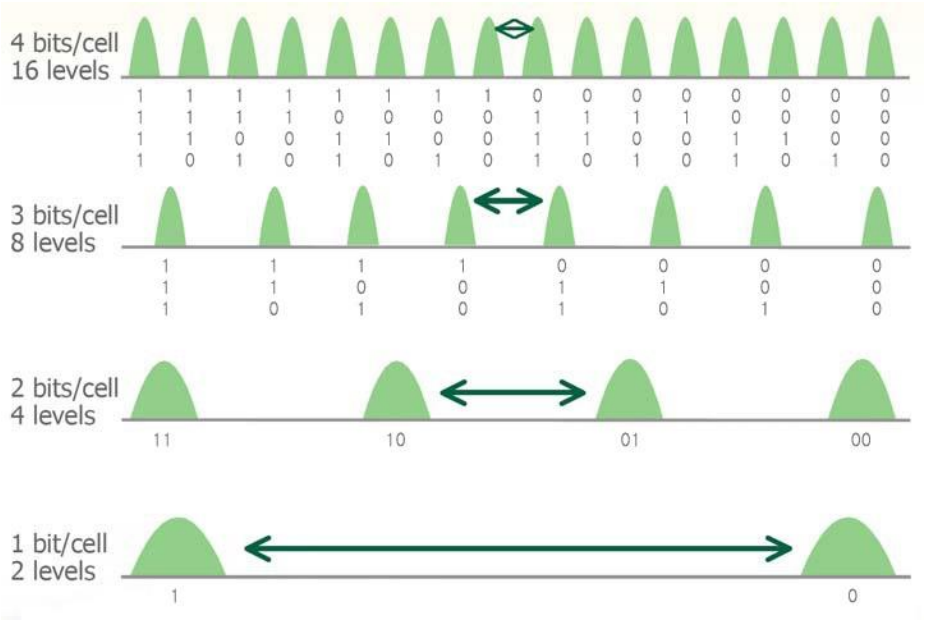
- 固态硬盘（Solid State Drives），用固态电子存储芯片阵列而制成的硬盘，由控制单元和存储单元（FLASH芯片、DRAM芯片）组成。

固态硬盘的存储单元

- 闪存的内部存储结构是金属-氧化层-半导体-场效应管(MOSFET)：源极，漏极和栅极，增加了浮置栅极
- 对于闪存的写入，即控制栅极去充电，对栅极加压，使得浮置栅极存储的电荷越多，超过阈值，就表示0
- 对于闪存的擦除，即对浮置栅极进行放电，低于阈值，就表示1



SLC, MLC, TLC, QLC



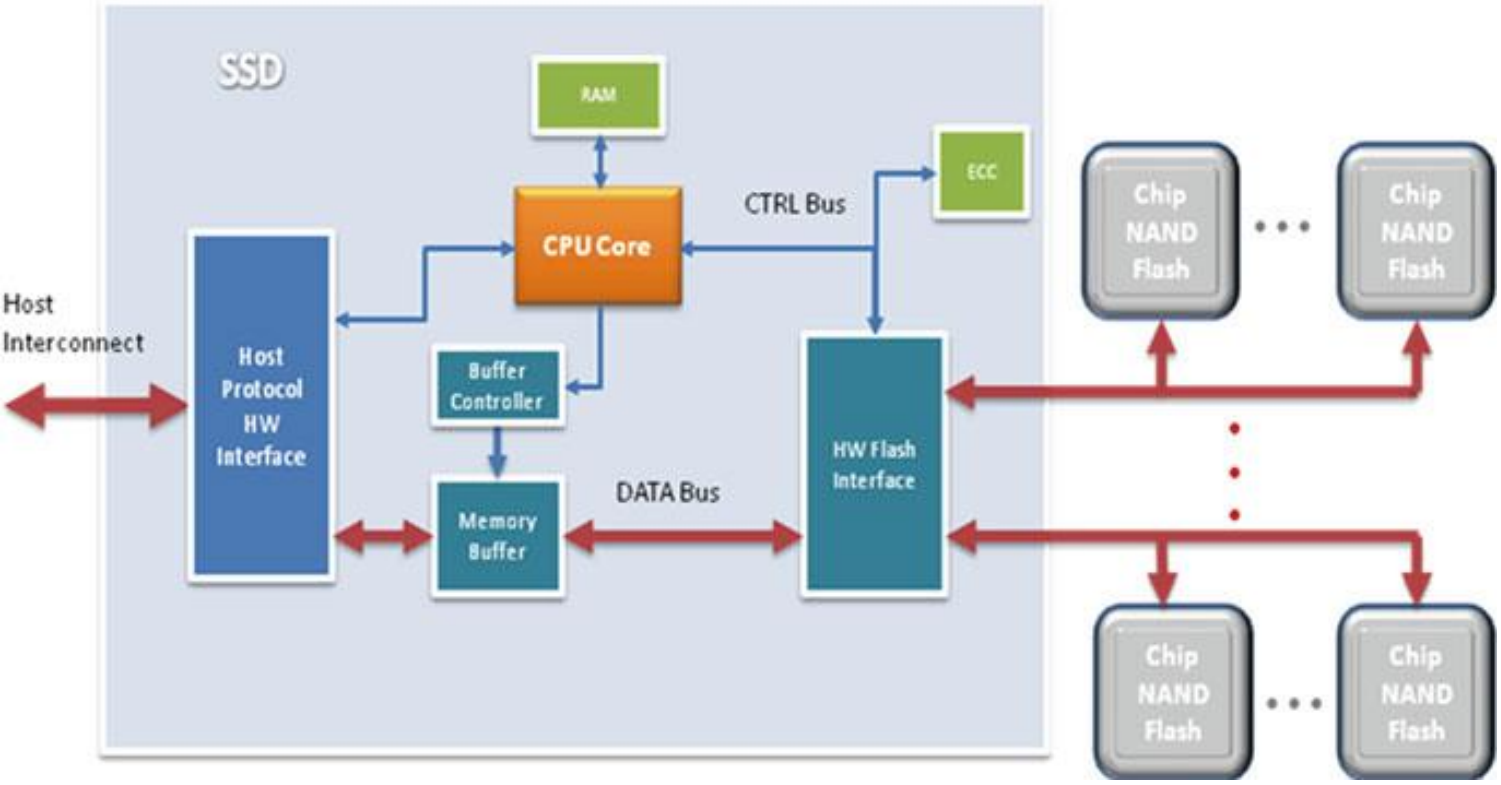
- 按照每个存储单元能够存储的位数分为SLC，MLC，TLC，QLC
- 多比特单元采用格雷码编码

固态硬盘存储单元的擦除次数

- 有限次擦除
  - 随着擦除次数的增加，存储单元不能可靠地保持状态（存储数据）
    - 耐久性 Endurance
    - 保持力 Retention
- SLC：100,000次
- MLC：10,000次
- TLC：1,000次

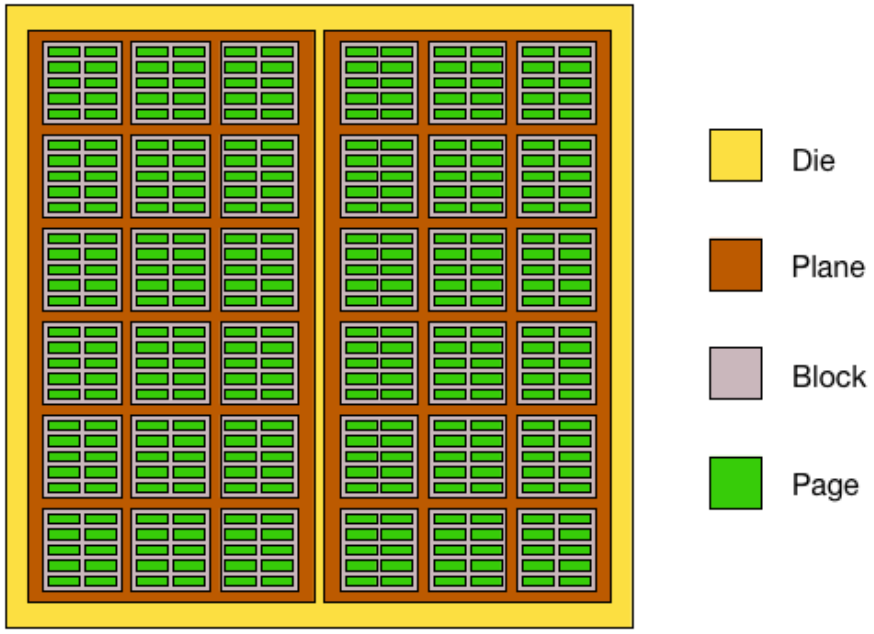
从存储器件到固态硬盘

- SSD主要由SSD控制器，FLASH存储阵列，板上DRAM（可选），以及跟HOST接口（诸如SATA, SAS, PCIe等）组成。
- 三个重要组成部分：主存芯片，闪存芯片，固件算法



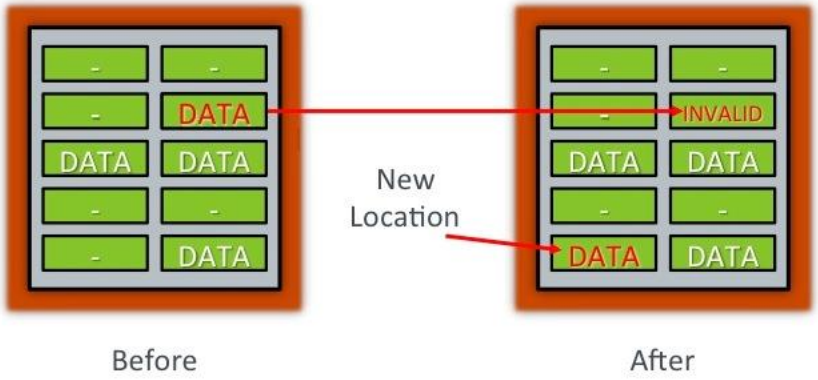
固态硬盘存储介质的组织

- 一个 package，即一个存储芯片，包含多个Die （典型1，2，4个）
- 一个Die包含1个或者2个plane，可并行操作
- 每一个plane包含多个block，block是最小的擦除单位
- 每一个block里面有多个页，页是最小的读写单位



- 闪存页 – 读写粒度
  - E.g., 4KB，8KB，16KB
  - us延迟
- 闪存块 – 擦除粒度
  - E.g., 2MB，4MB，8MB
  - ms延迟

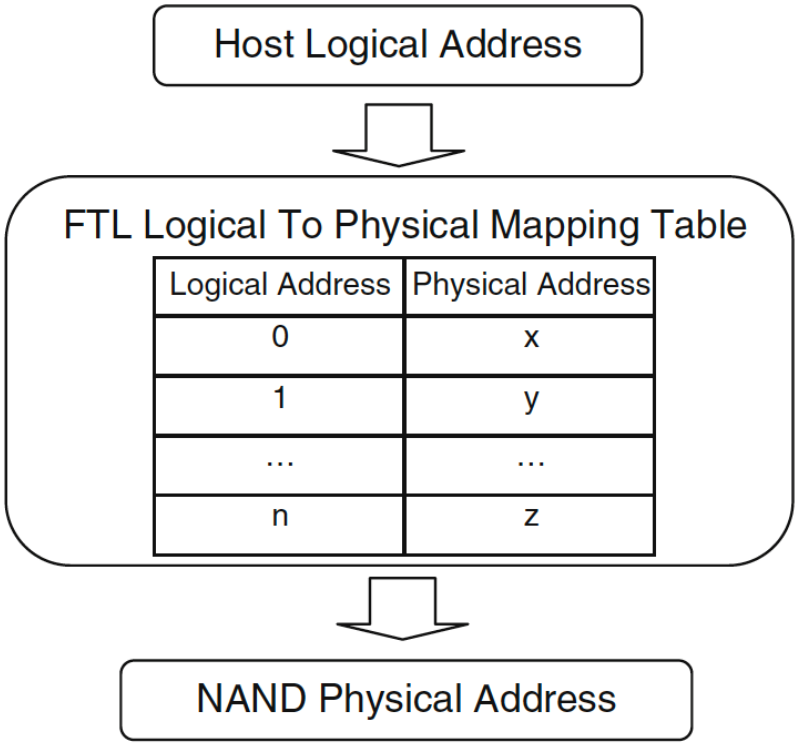
SSD的写入



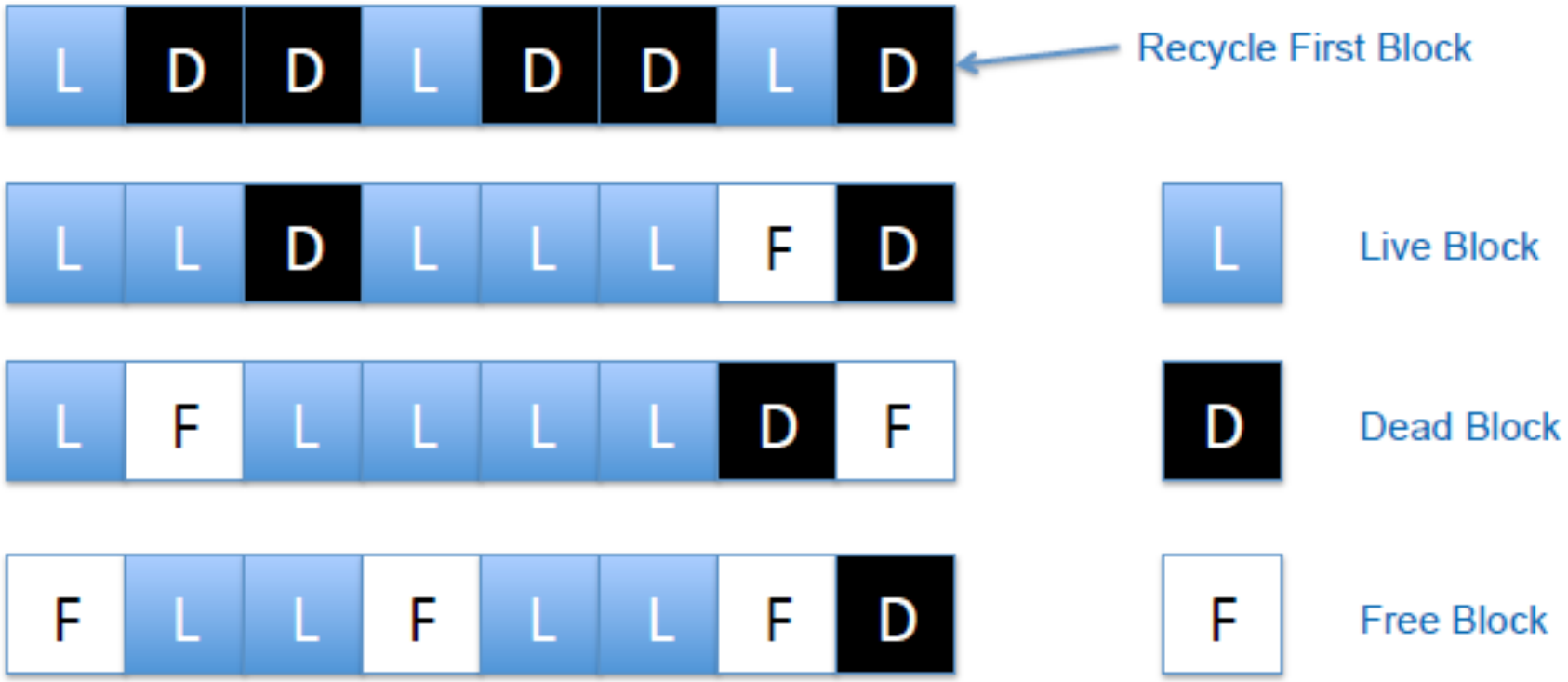
- 与磁盘不一样，不会写入到原来的page，写入之前需要进行擦除操作
- 写入的时候不会在原来的page中写入，会在一个新的页面（可以在同一个块，也可以在不同的块，可能不同的plane，甚至不同的die上面）
- 这样，需要维护上层管理软件的逻辑地址和底层的物理地址之间的映射关系
- 这个工作交给FTL层来完成

SSD的地址转换FTL

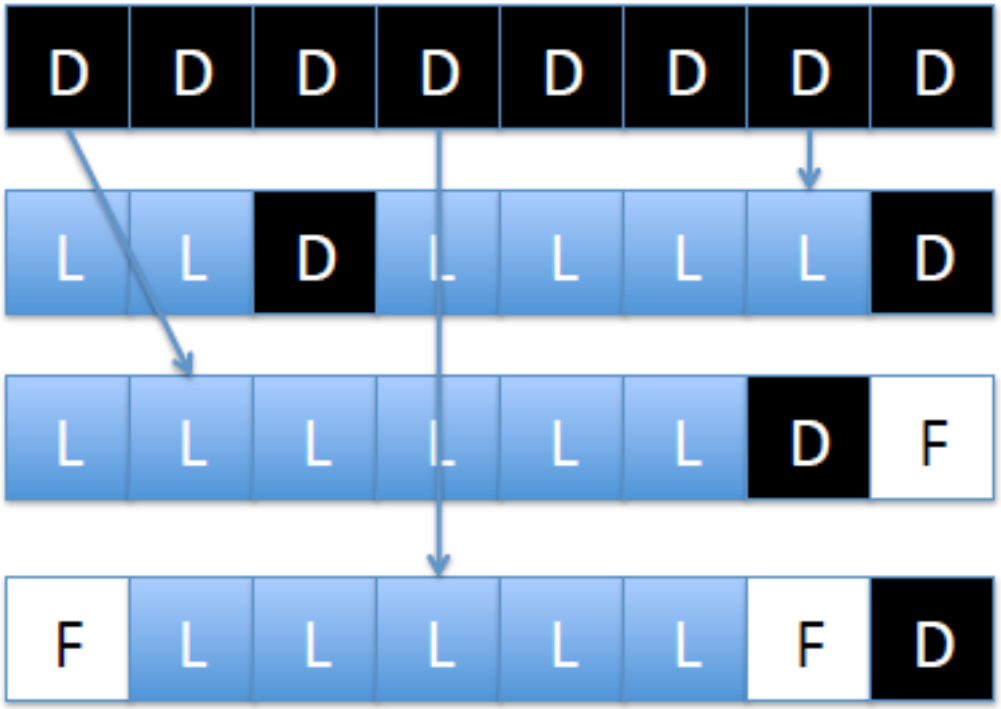
- FTL：Flash Translation Layer，做逻辑地址到物理地址的翻译
- 除了做地址转换，FTL还帮助完成磨损均衡
- 写入之前必须要进行擦除，但是每一个块的擦除的次数有限
- 写入的时候需要挑选位于擦除次数最少的块中的页面，完成磨损均衡



SSD上的垃圾搜集（1）

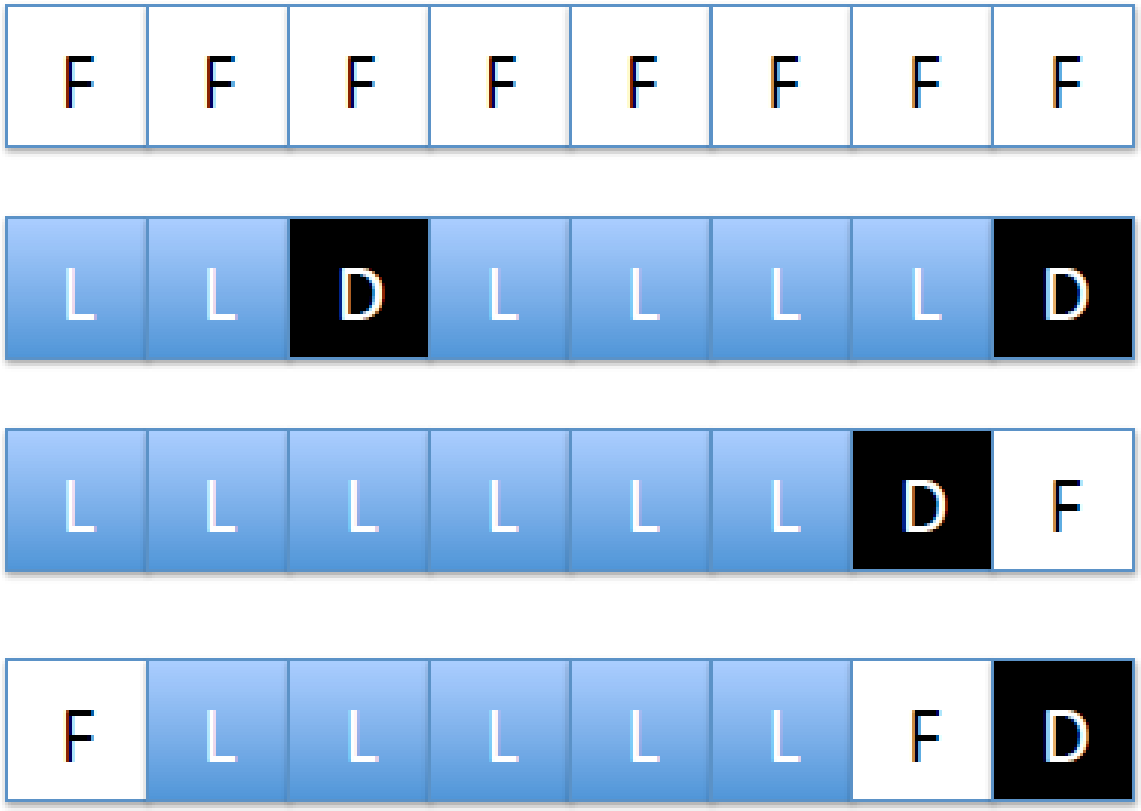

L D D L D D L D
Recycle First Block
L L D L L L F D
L
Live Block
L F L L L L D F
D
Dead Block
F L L F L L F D
F
Free Block

SSD上的垃圾搜集（2）



Copy Live Data somewhere else

SSD上的垃圾搜集（3）



Erase Data Block

Overhead:
-Copying of Live Data
-Block Erasing

典型的SSD性能数据

- 依据不同SSD的型号具有不同的性能数据
- 读数据的性能：
  - 20-100微秒的延迟
  - 100-500MB/s的带宽
- 擦除的性能：
  - 2ms，即毫秒级
- 写数据的性能：
  - 200微秒的延迟
  - 100-200MB/s的带宽

最新的性能数据

- DELTA MAX SSD硬盘为SATA 6Gbps接口
- 有250GB、500B、1TB及2TB容量
- 其中250GB的读取、写入速度分别是560MB/s、500MB/s
- 其他容量的读取、写入分别是560MB/s、510MB/s
- 4K随机读写最高90K、80K IOPS

SSD总结

- 机械式存储设备转化为电子式存储设备
- 不同的操作粒度
  - 以页的方式读写
  - 以块的方式擦除
- 擦除次数有限：需要磨损均衡
- FTL：
  - 逻辑地址到物理地址的转换
  - 磨损均衡
  - 垃圾回收

阅读与思考

- 阅读
  - 教材相关章节
  - SSD存储原理
- 思考
  - 磁表面存储设备的原理和特点，它在层次存储器系统中的地位和作用
  - 如何将文件分布在扇区使访问速度加快？
- 实践
  - 本单元作业

谢谢