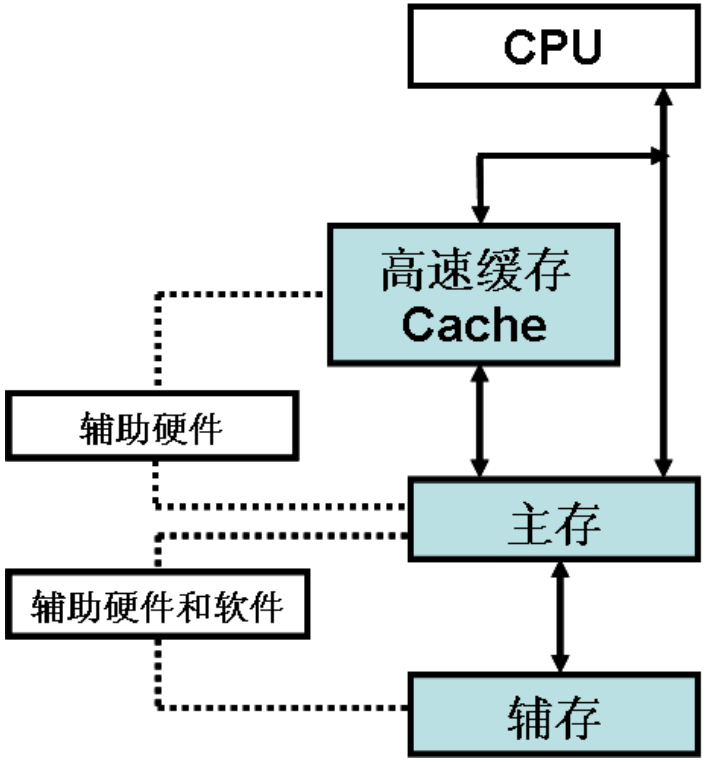
计算机组成原理

第四章 存储器

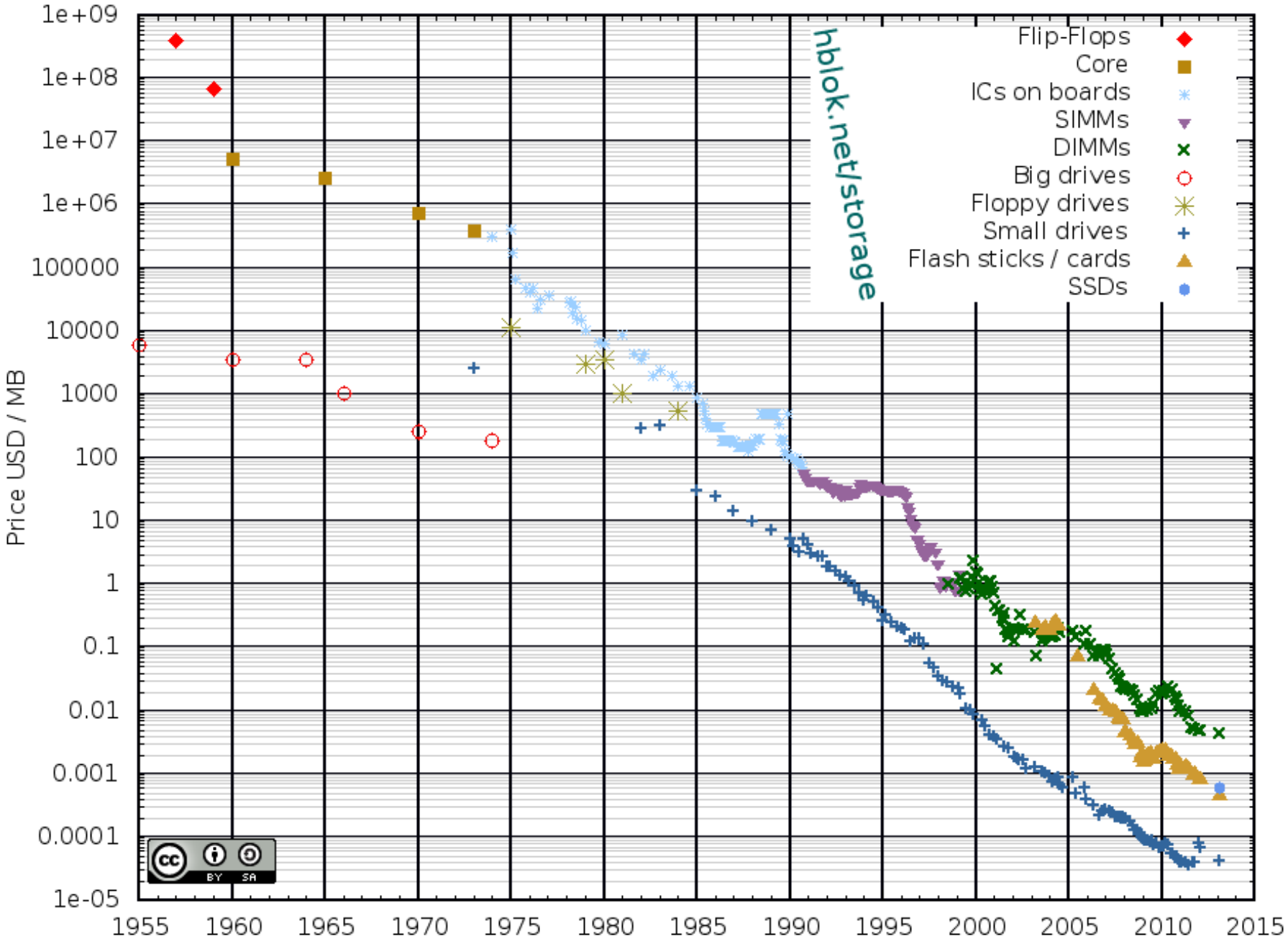
llxx@ustc.edu.cn
wjluo@ustc.edu.cn

本章内容

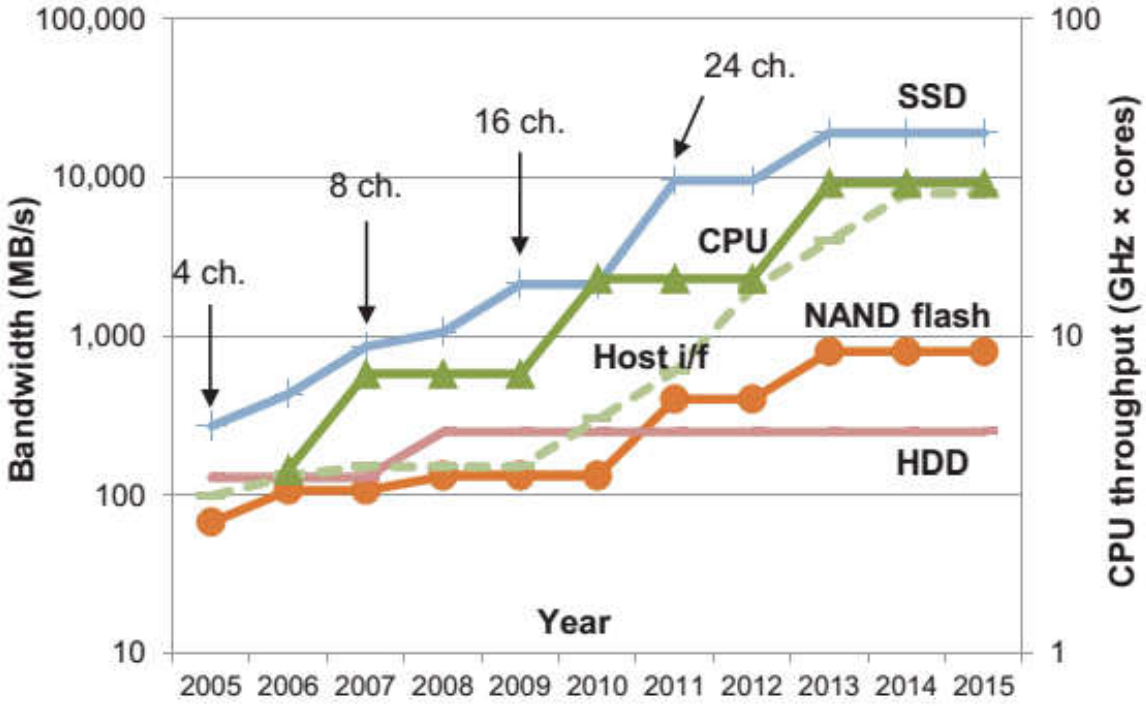
- ✓4.1 概述
- ✓4.2 主存储器
- 4.3 高速缓冲存储器
- 4.4 辅助存储器
  - 磁表面存储器
  - 硬盘
  - RAID技术：性能与可靠性
  - FlashDisk

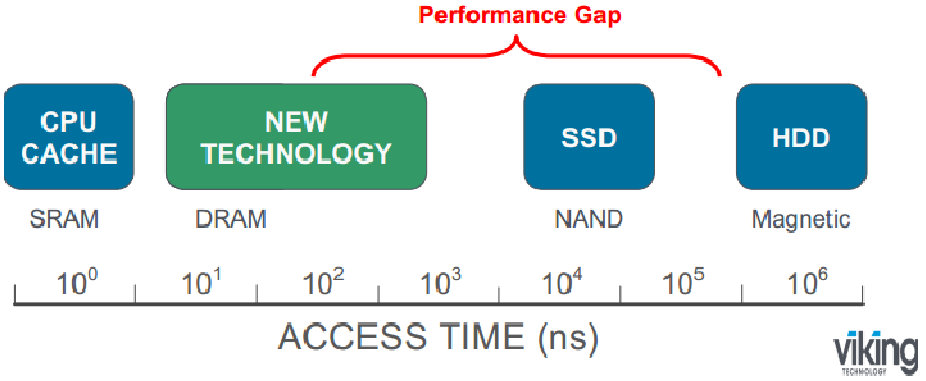


Historical Cost of Computer Memory and Storage



外存的性能


100,000
10,000
1,000
100
10
Bandwidth (MB/s)
100
10
1
CPU throughput (GHz × cores)
4 ch.
8 ch.
16 ch.
24 ch.
SSD
CPU
NAND flash
Host i/f
HDD
Year
2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015


Performance Gap
CPU CACHE
NEW TECHNOLOGY
SSD
HDD
SRAM
DRAM
NAND
Magnetic
10^0 10^1 10^2 10^3 10^4 10^5 10^6
ACCESS TIME (ns)
viking

4.4 辅助存储器

1. 磁表面存储器
   - 磁记录原理和记录方式
   - 硬磁盘存储器
   - *软盘存储器*
   - *磁带存储器*
2. *光盘存储器*
3. 循环冗余校验码、奇偶校验码

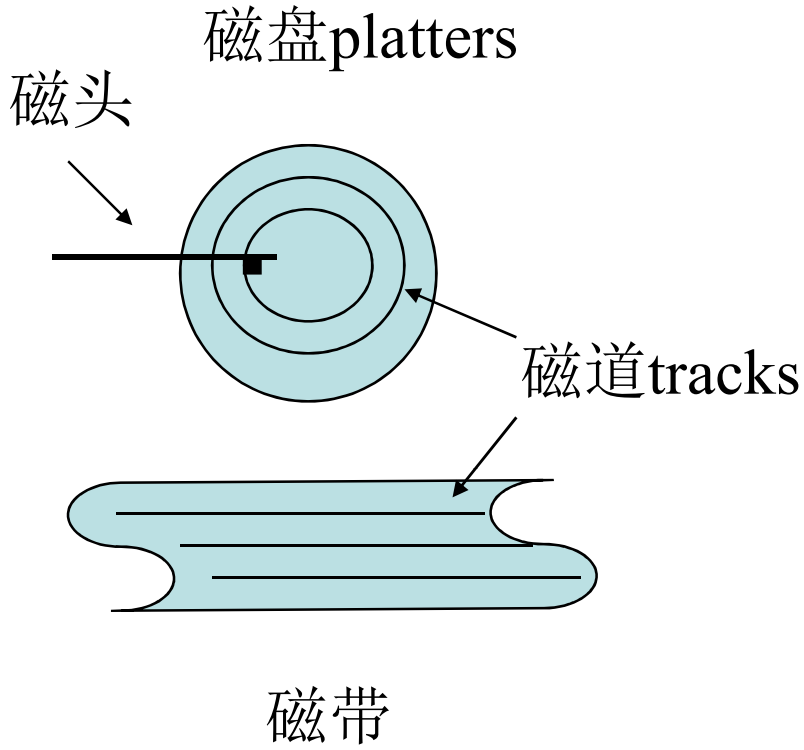
辅助存储器的特点

- 外存
  - 硬盘、软盘、磁带、光盘（CD ROM）
  - 容量大，GigaBytes
  - 速度慢, 7200转/min，速率<100Mb/s
    - RAM：几百兆(存取周期几十纳秒)
  - 价格低，80G/￥800.00
    - 内存：256M/￥400.00
  - 可脱机保存信息，具有非易失性的特点

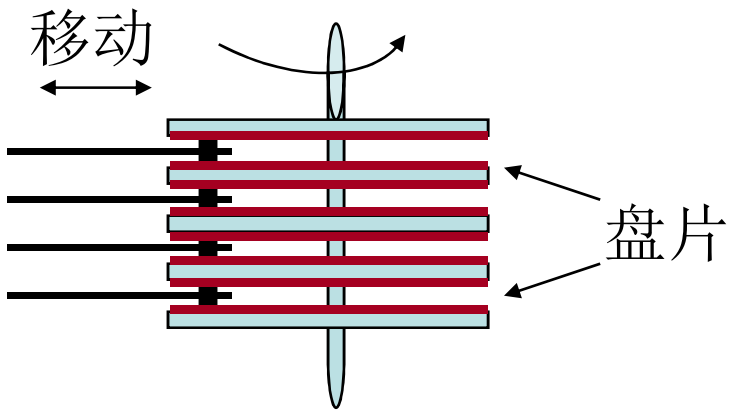
磁表面存储器

- 主要内容
  1. 技术指标
     - 记录密度、容量、寻址时间、传输率、误码率
  2. 磁记录原理
  3. 磁盘记录格式
  4. 评价记录方式的主要指标
  5. 硬磁盘存储器
  6. 软磁盘存储器
  7. 磁带存储器

磁记录设备



磁带



磁盘驱动器

- 设备读写方式
  - 随机方式：RAM
  - 顺序方式：磁带
  - 直接方式：磁盘（扇区的定位采用随机方式，依靠磁盘旋转可直接找到某一扇区，而扇区内则采用顺序读写方式）

技术指标 - 记录密度

- 记录密度：道密度（磁盘）、位密度（磁盘、磁带）
  - 道密度：沿半径方向单位长度磁道数
    - 单位：道/英寸（ TPI，Tracks Per Inch）　　P：道距

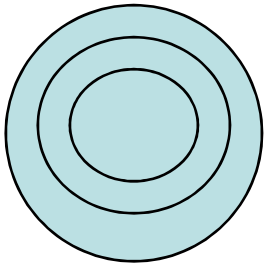
$$D_t = \frac{1}{P}$$

  - 位密度：单位长度磁道所记录的数据位数，单位为位/英寸（bpi）或位/毫米（bpm）

$$D_b = \frac{f_t}{\pi \bullet d_{\min}}$$

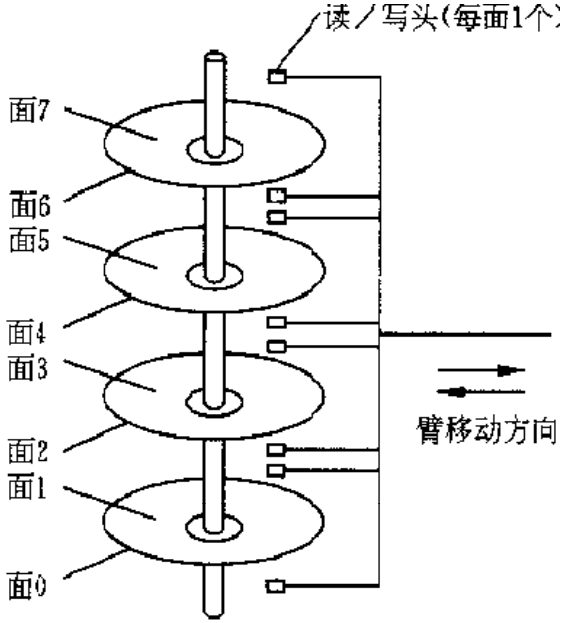
f_t：每道总位数，各道相同

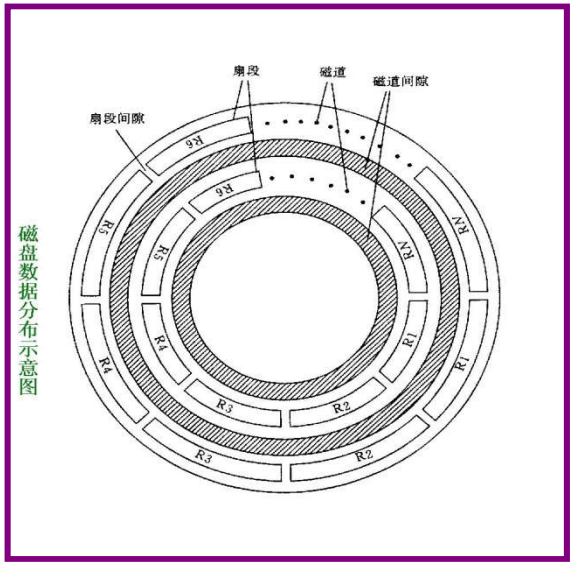
$d_{\min}$：同心圆最小直径



技术指标 - 容量

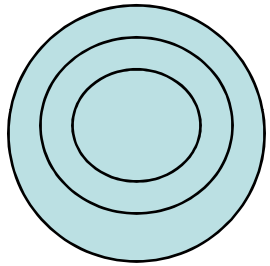
- 容量：存储的信息总量
  - 以磁盘为例
    - 磁盘总容量C = n×k×s
      - n：盘面数
      - k：每面磁道数
      - s：每道记录代码数
    - 非格式化容量：磁表面可以利用的磁化单元总数。
    - 格式化容量：按某种特定的记录格式所能存储信息的总量，约为非格式化容量的60%~70%





技术指标 - 寻址时间

- 磁盘寻址过程：直接存取
  - 随机寻道，顺序定位记录
  - 寻址时间 = 寻道时间（t_s）+ 等待时间（t_w）
  - 平均寻址时间
    - 寻道：最外、最内、相邻，各不相同
    - 等待时间：外道、内道长度不同



$$T_a = t_{sa} + t_{wa} = \frac{t_{s\max} + t_{s\min}}{2} + \frac{t_{w\max} + t_{w\min}}{2}$$

- 磁带寻址过程：顺序存取
  - 磁头不动，磁带空转到指定位置。
  - 寻址时间 = 空转时间

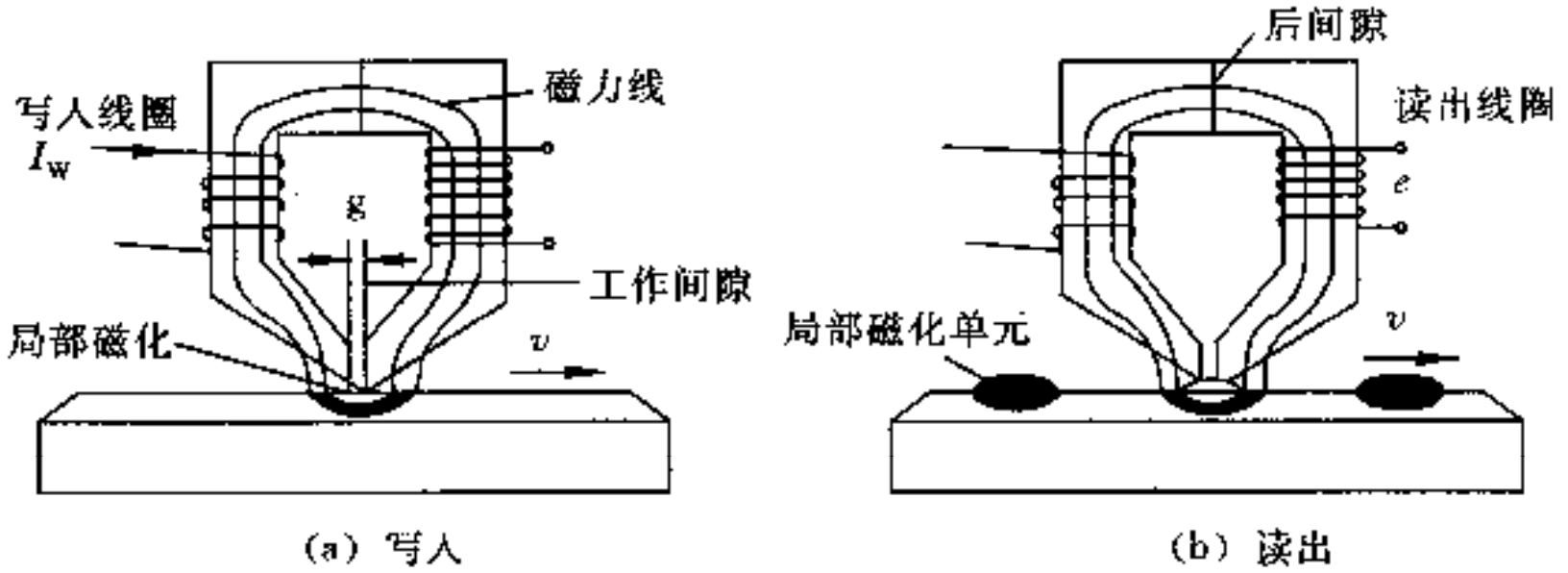
技术指标 - 传输率、误码率

- **传输率**：
  - 单位时间传输的数据量（字节、位）
  - D_r = 记录密度（D）×介质运行速度（V）
- **误码率**：
  - 读出时，出错位数/读出的总位数
  - 为了减少出错率，磁表面存储器通常采用循环冗余码CRC来发现并纠正错误。

磁记录原理

- 磁记录机制
  - 写：将磁层表面单元磁化，极性区别“0”、“1”
  - 读：磁化单元的磁通，产生感应电势，方向区别“0”、“1”
- 水平记录、垂直记录

磁记录原理—读、写过程

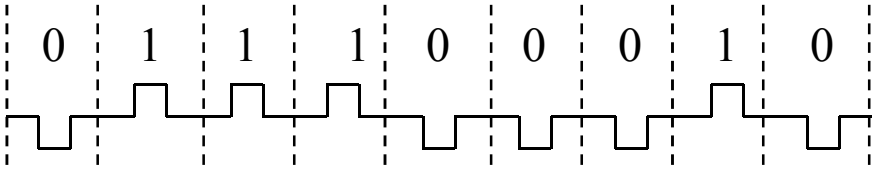


- **写入：**记录介质在磁头下匀速通过,磁头线圈中通入一定方向和大小的电流,则会在介质上形成一个磁化单元. 电流方向不同,则磁化方向也不同.一个磁化方向规定为“0”,另一个磁化方向就规定为“1”.
- **读出：**记录介质在磁头下匀速通过时,读出线圈会感应出电压,磁化方向不同,则感应电压就不同,对感应电压进行放大和整型,就可以读出“0”或“1”.

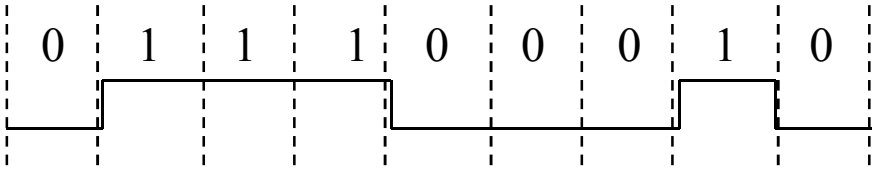
磁表面记忆原理—记录方式

- **磁记录方式：又称编码方式**
  - **即**按某种规律，将一串二进制数字信息变换成磁表面相应的磁化状态。
  - 对记录**密度**和**可靠性**有很大影响。
- 常用的编码方式有：
  1. 归零制（NZ）
  2. 不归零制（NRZ）
  3. 见 1 就翻的 NRZ1
  4. 调相制（PM）
  5. 调频制（FM）
  6. 改进调频制（MFM）

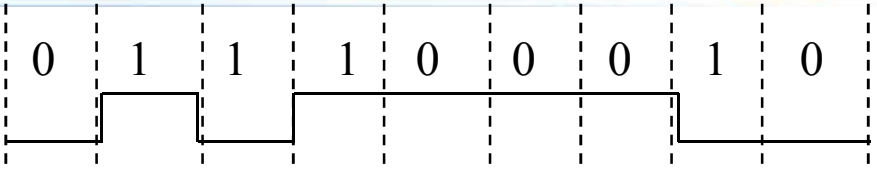
磁表面存储器的磁记录原理



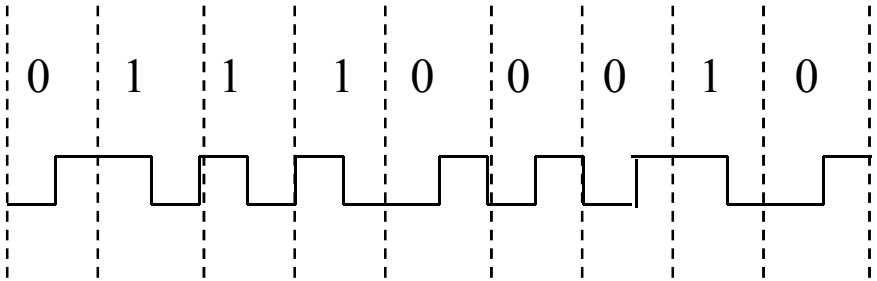
- （1） 归零制（RZ）
  - 正脉冲电流表示“1”，负脉冲电流表示“0”；
  - 不论记录“0”或“1”，在记录下一个信息前，记录电流恢复到零电流。
  - 简单易行，记录密度低
  - 改写磁层上的记录比较困难，一般是先去磁后写入。
  - 有自同步能力（能从磁头读出信号中分离获得同步信号）
- （2）不归零制（NRZ）
  - 磁头线圈始终有电流，电流方向“见变就翻”
  - 对连续记录的“1”和“0”，写电流的方向是不改变的。
  - 无自同步能力。



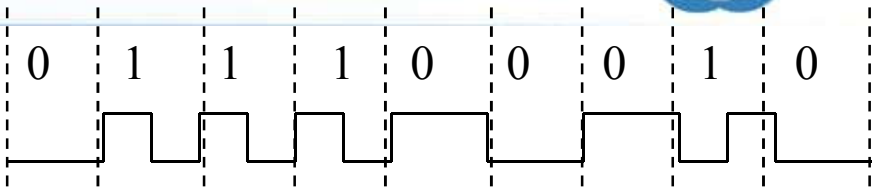
磁表面存储器的磁记录原理



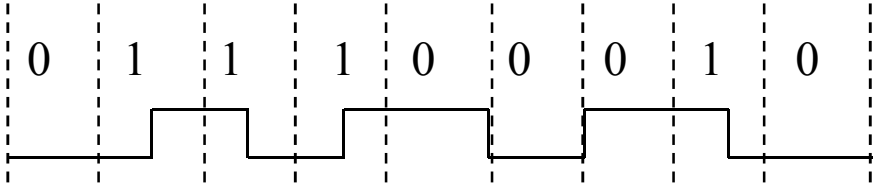
- （3）见“1”就翻的不归零制（NRZ1）
  - 磁头线圈始终有电流通过。
  - 在记录“1”时，电流改变方向，写“0”电流保持不变。
  - 不具备自同步能力，需要引用外同步信号
- （4）调相制（PM）：又称为相位编码（PE）
  - 记录数据“0”时，规定磁化翻转的方向由负变为正，记录数据“1”时从正变为负
  - “0”，“1”的读出信号相位不同，抗干扰能力强
    - 磁带多用此方式
  - 具有自同步能力



磁表面存储器的磁记录原理

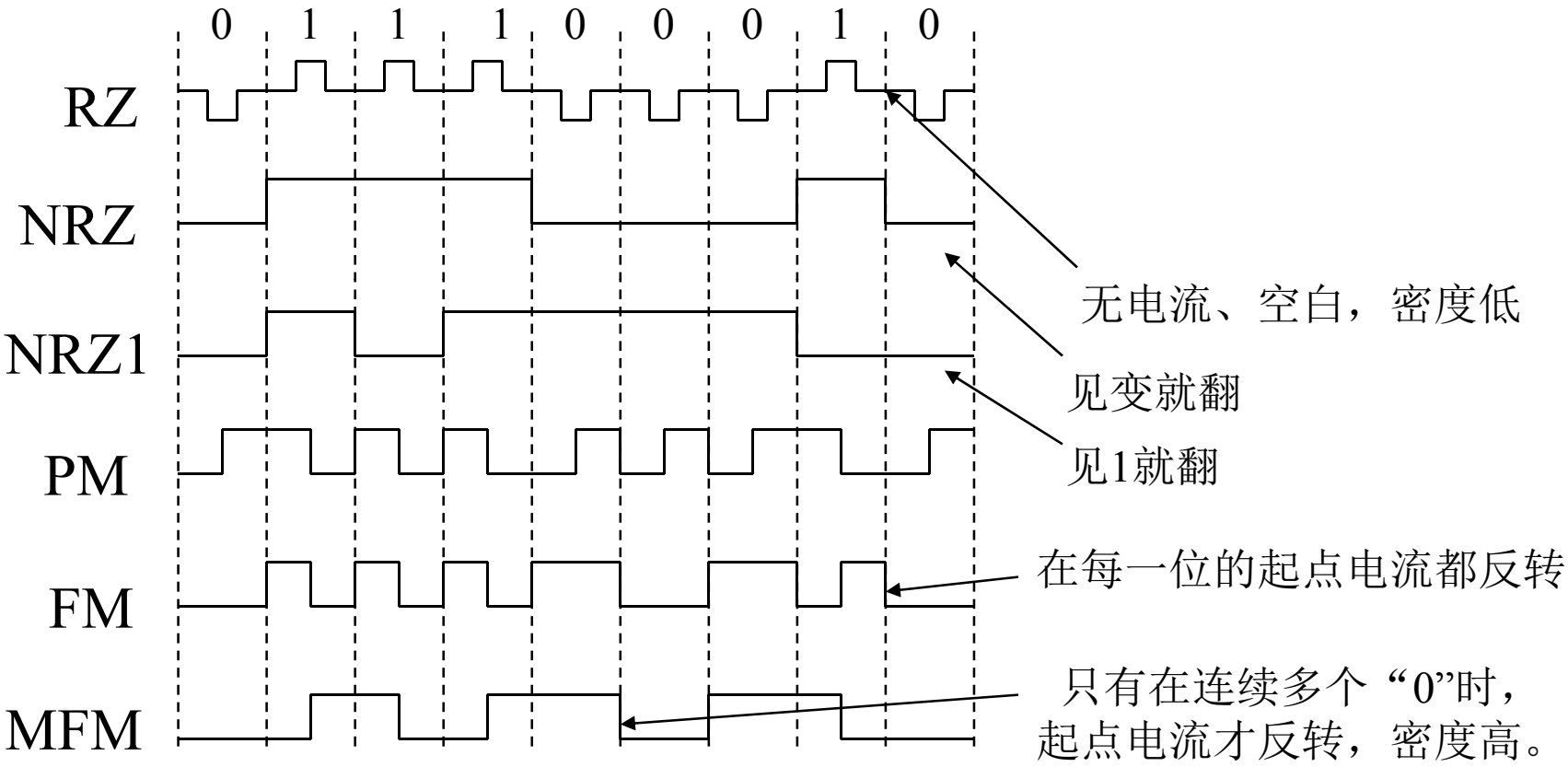


- （5）调频制（FM）
  - 频率变化（“1”的频率是“0”的两倍）
  - 在位与位之间的边界处都要翻转一次
  - 具有自同步能力。
  - 用于软硬磁盘
- （6）改进调频制（MFM）
  - 不是在每个位周期的起始处都翻转。当连续两个或两个以上“0”时，在位周期的起始位置翻转一次。
  - 具有自同步能力



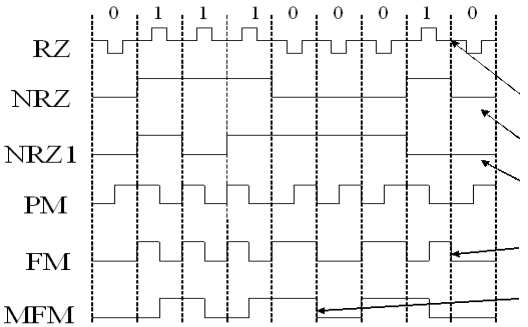
磁记录方式 - 编码方式

- 写电流波形的形式

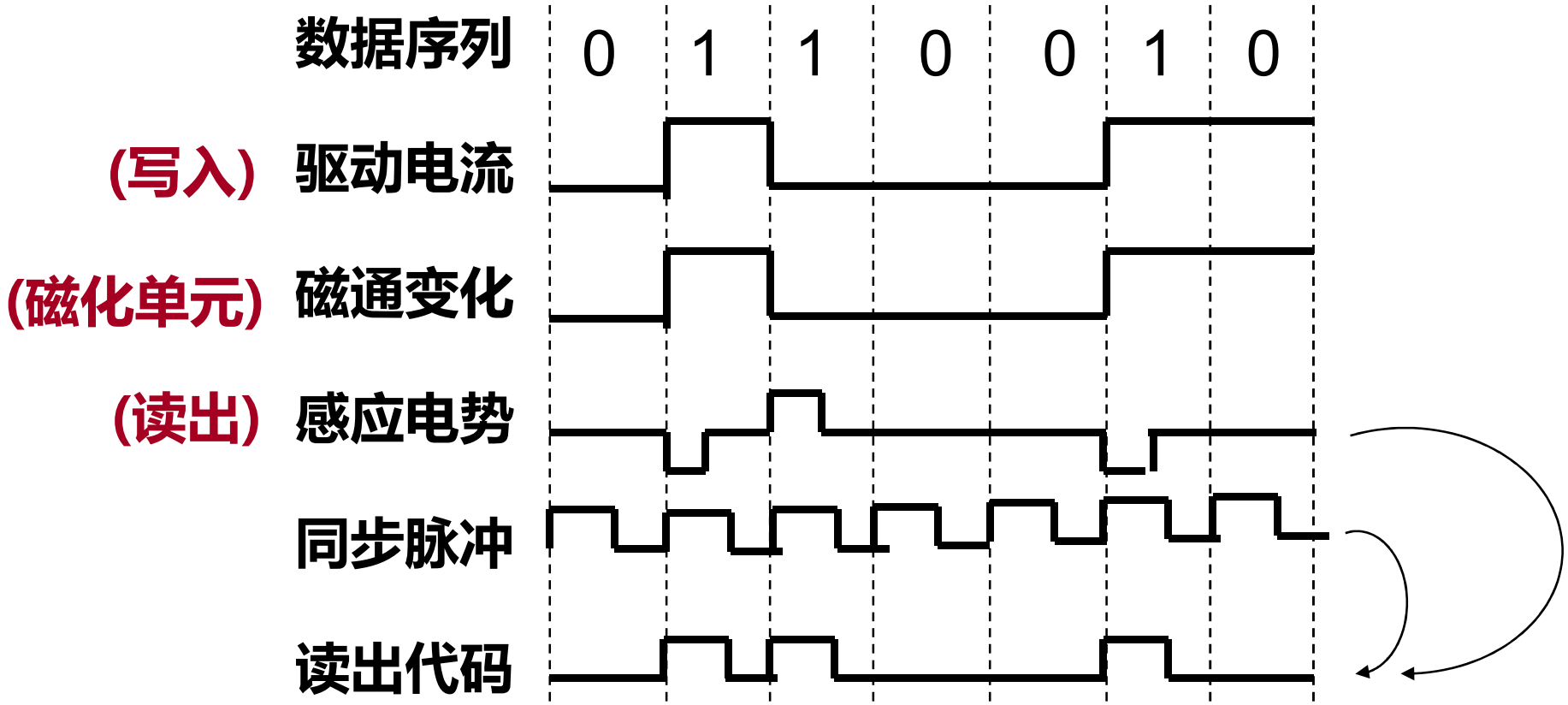


评价记录方式的主要指标

- **编码效率**：位密度与磁化翻转密度的比值，用记录一位信息的最大反转次数表示
  - FM、PM：最多需反转2次，效率50%
  - NRZ、NRZ1、MFM：最多只需反转1次，效率100%
- **自同步能力**
  - 指：从单个磁道读出的脉冲序列中提取同步脉冲的难易程度
  - 外同步：从专门设置的用来记录同步信号的磁道中取得同步脉冲。
    - NRZ、NRZ1
  - 自同步：记录方式中隐含同步信息
    - PM、FM、MFM
    - 自同步能力(R) = 最小反转间隔/最大反转间隔
      - FM：R=1/2



NRZ1的读出代码波形



•感应电势的负波要反向。与同步信号相“与”

硬磁盘存储器

硬盘的发展和几个指标

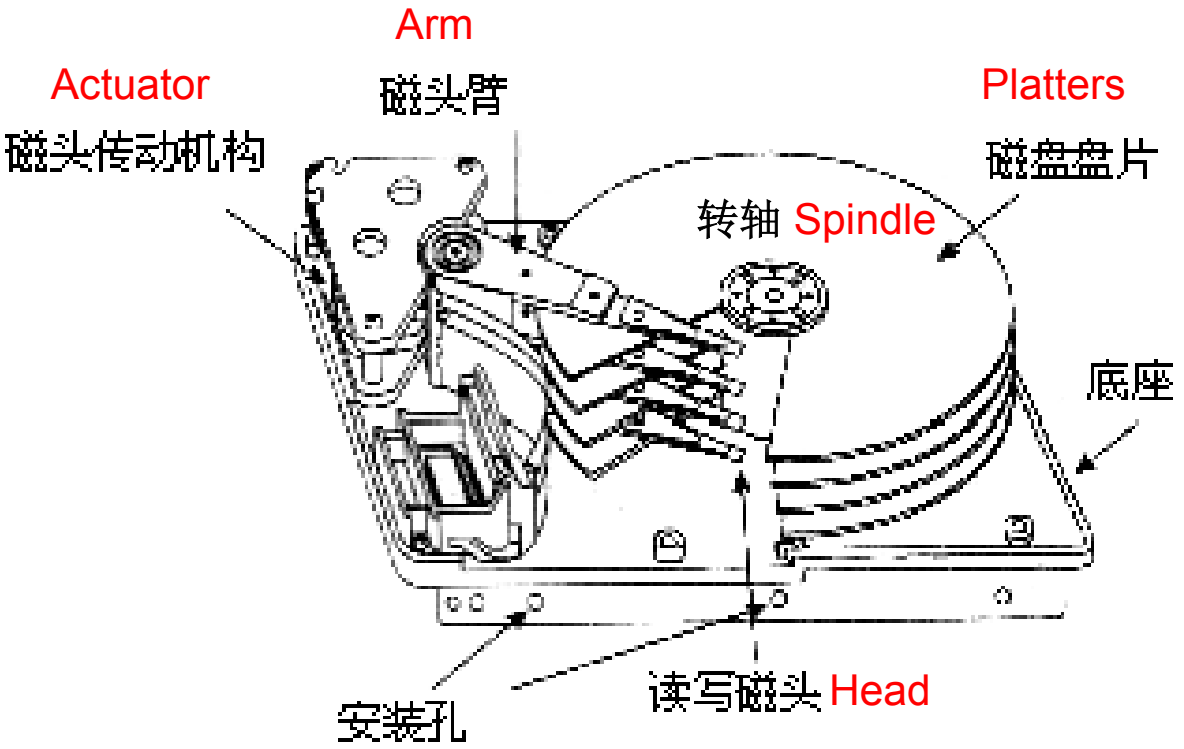
- 1956年，美国IBM公司研制成第一个商品化的硬磁盘
- 1973年，IBM发明了温彻斯特(温氏)磁盘，简称温盘
- 80年代以来，硬盘随微机的普及而广泛使用。

- 硬盘的几个指标：
- **体积：** 5.25英寸/全高、3.5英寸/半高(台式PC)；2.5英寸(笔记本PC)
- **容量：** 10~40MB(8086/286) → 120G
- **传输速率：** 100KB/s → 50MB/s
- **平均寻道时间：** 80ms → 5ms
- **转速：** 目前大约为7200转/s。

RAMAC (Random Access Method of Accounting and Control)

- IBM公司1956年推出的首台硬磁盘存储器
  - 50个直径为24英寸的盘片组成
  - 以每分钟1200转的速度旋转
  - 容量为5MB
  - 约有两个冰箱大

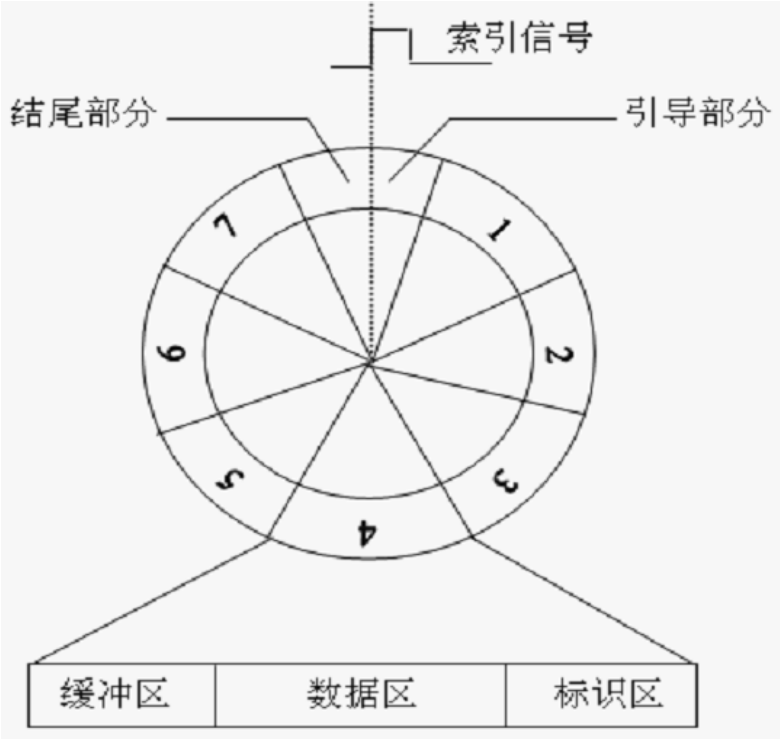
温盘（温彻斯特）

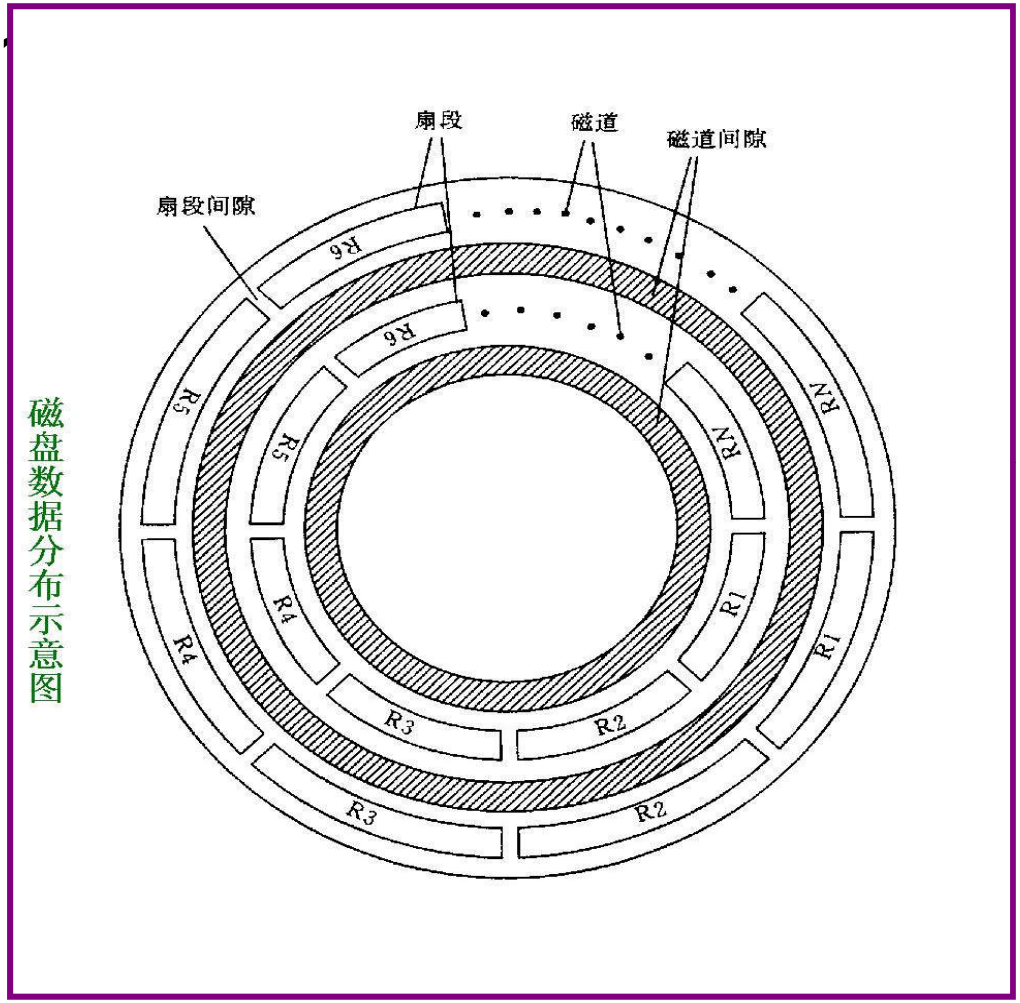


硬盘的内部结构

硬磁盘的磁道记录格式

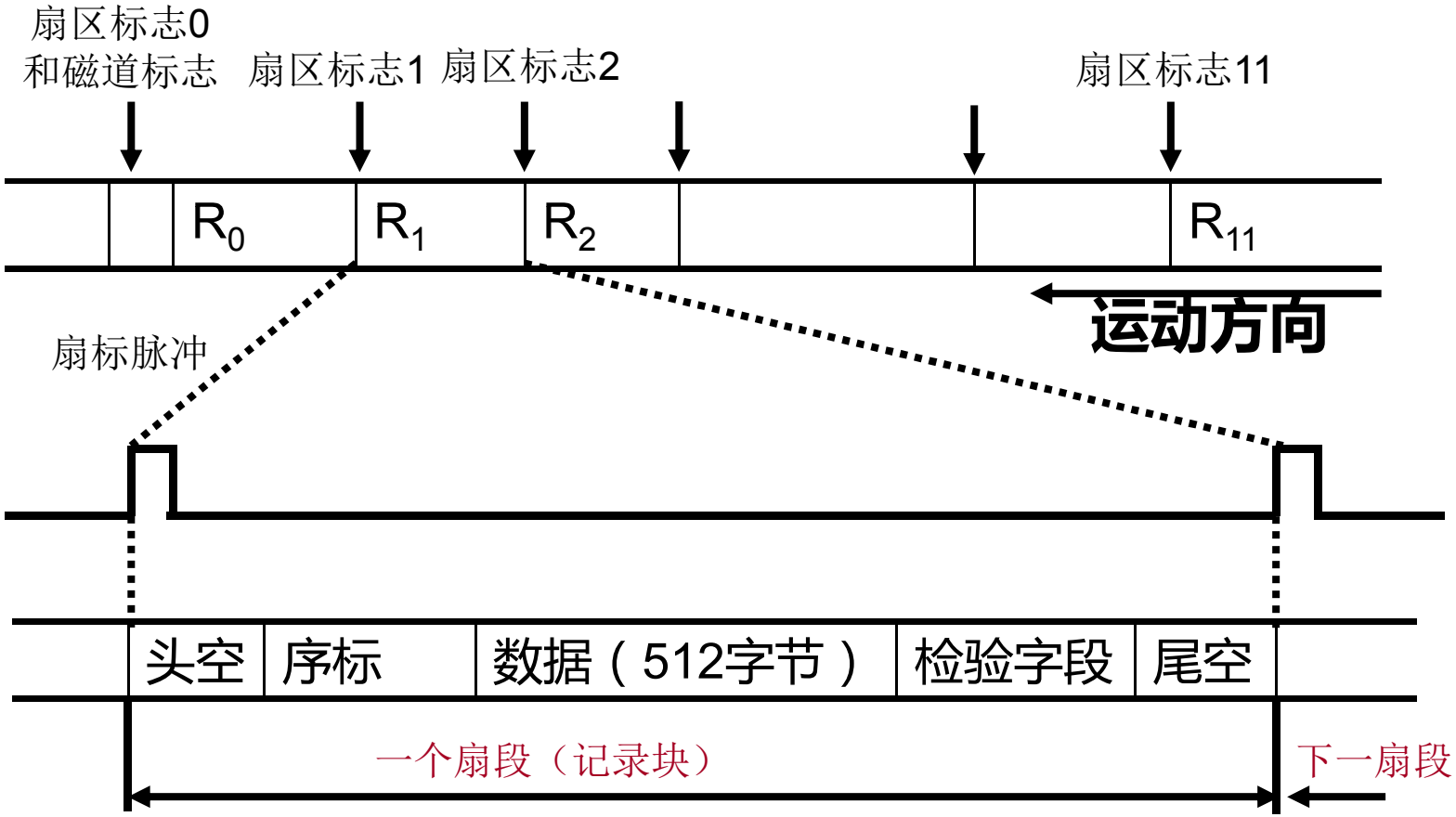
- 扇区sector的大小
  - 定长记录格式
  - 不定长记录格式



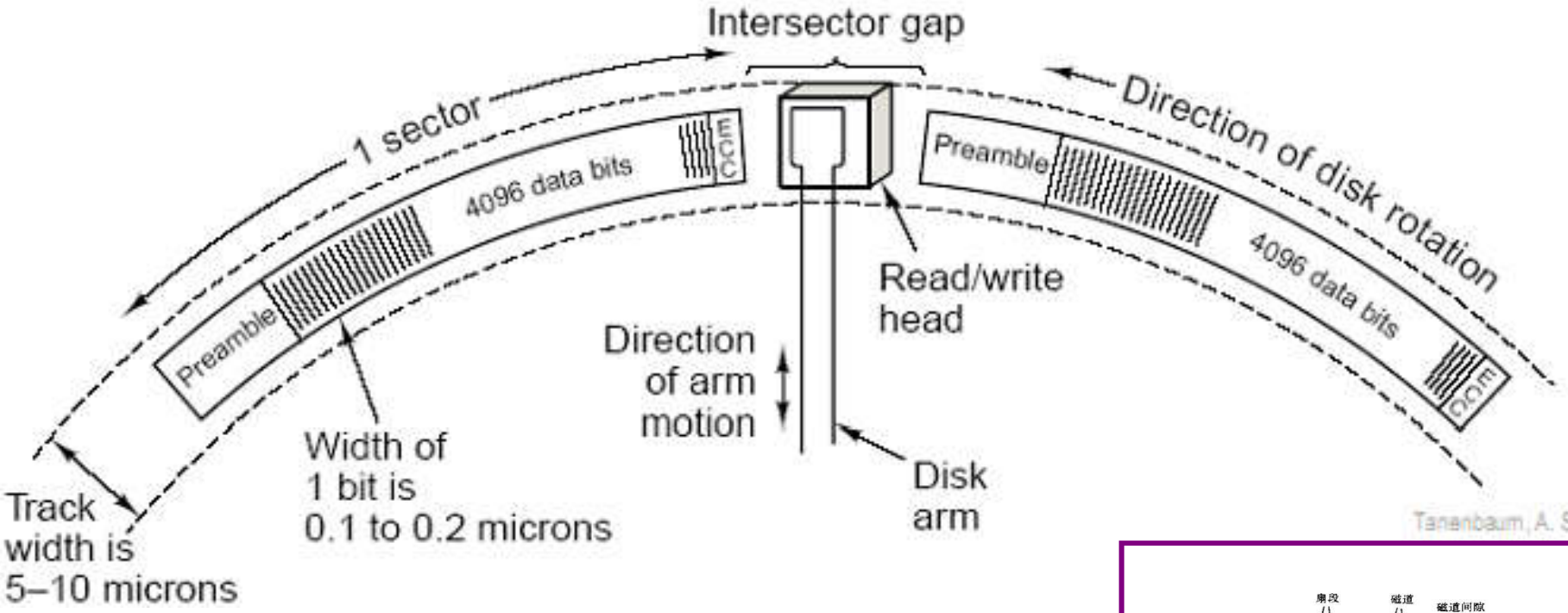


定长记录格式—ISOT型磁道记录格式

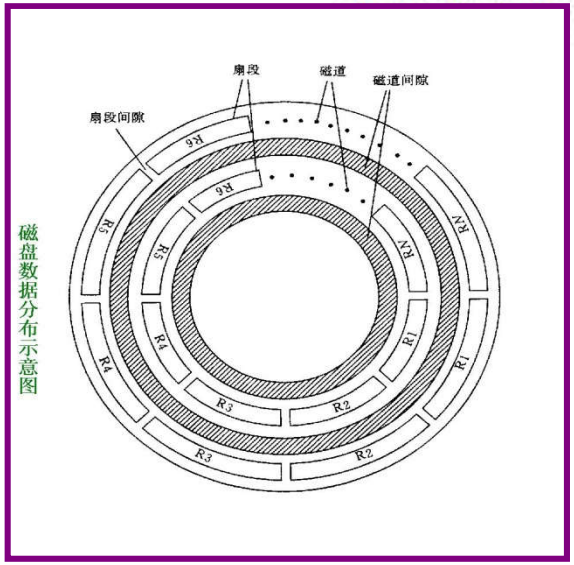
- 结构简单，可按柱面号、盘面号、扇段号进行直接寻址，但是记录区的利用率不高。



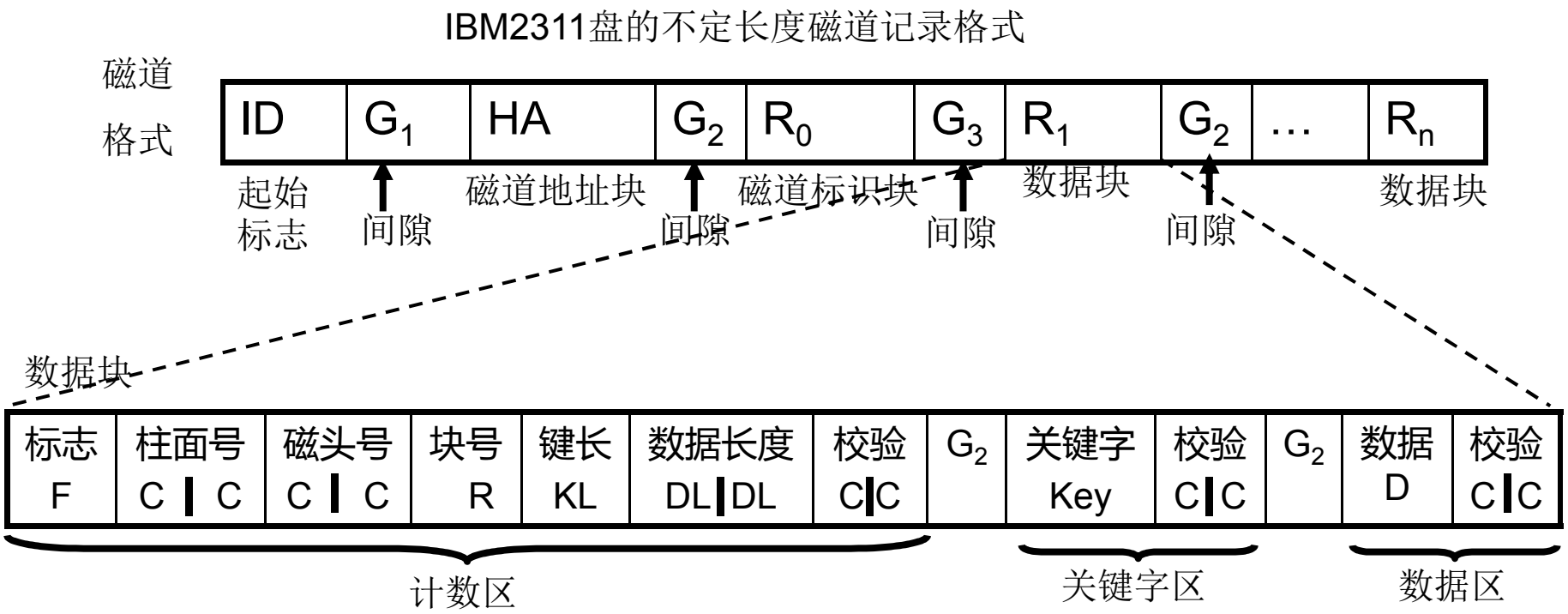
例：



Tanenbaum, A. S. (2006)

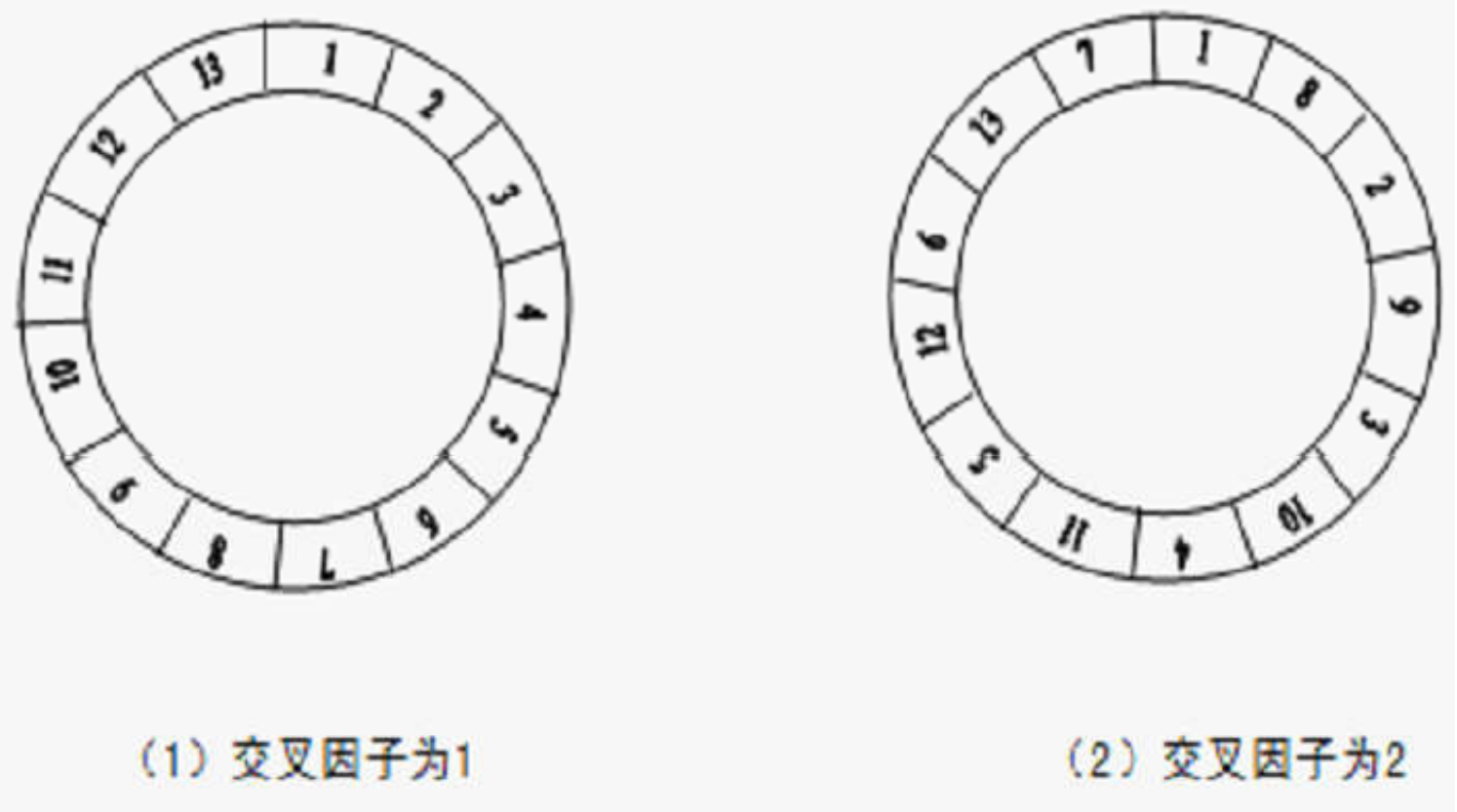


不定长记录格式



- 定长记录格式：若文件长度不是定长记录的整数倍时，往往造成记录块的浪费
- 不定长记录格式：根据需要来决定记录块的长度，如IBM2311、2314等磁盘驱动器。

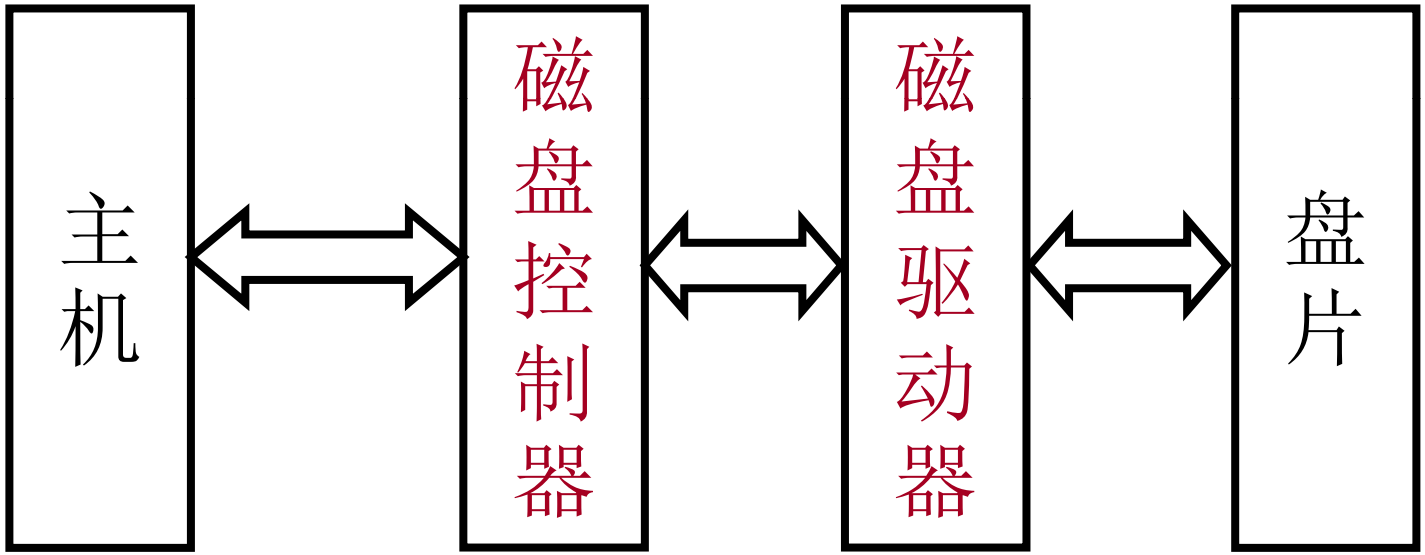
扇区的排列：交叉因子（Interleave）



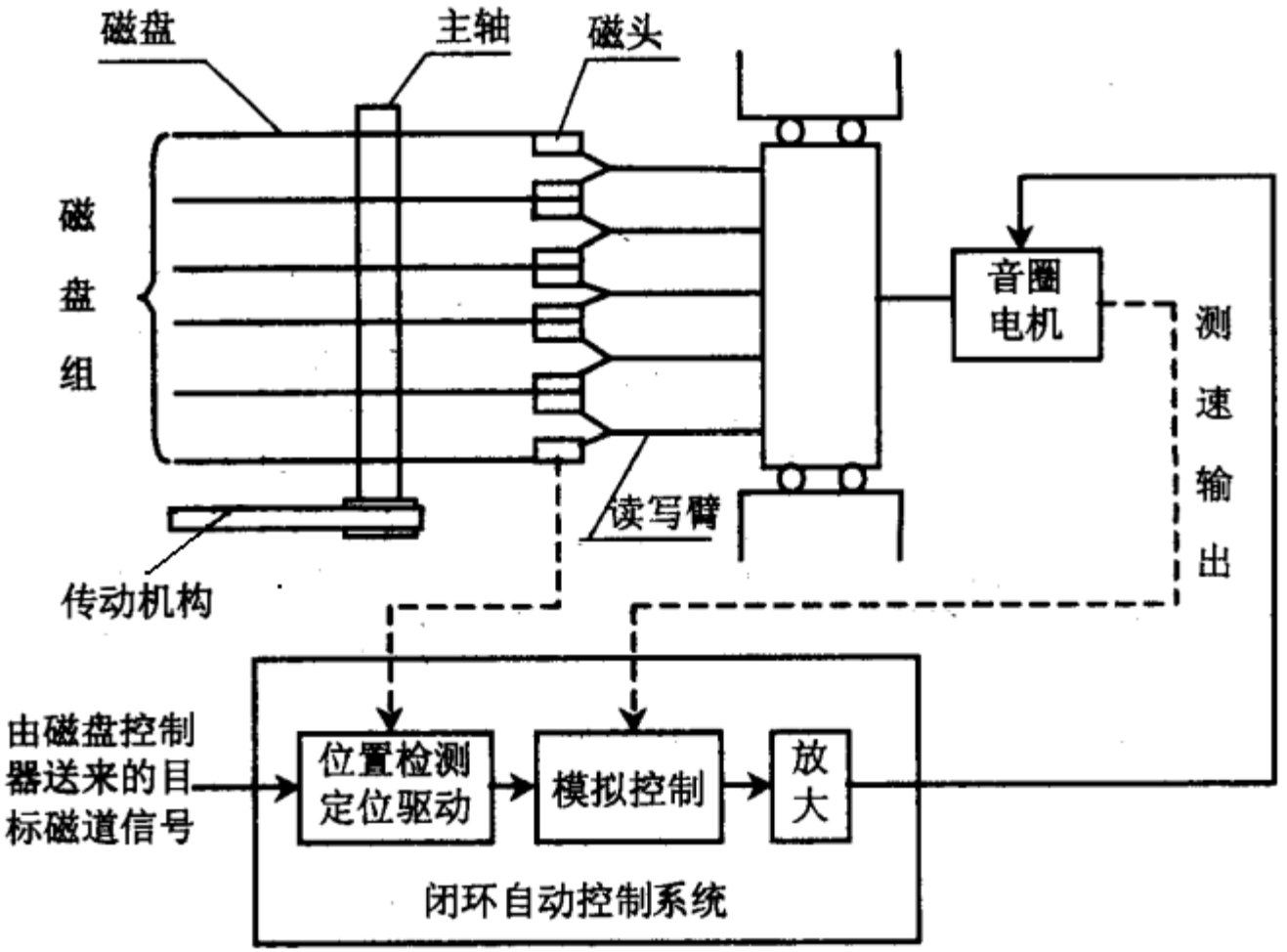
- 扇区交叉排列技术
  - 磁头读写反应速度低于盘片的旋转速度

硬磁盘存储器的结构

- 硬磁盘存储器由磁盘驱动器、磁盘控制器和盘片组成。



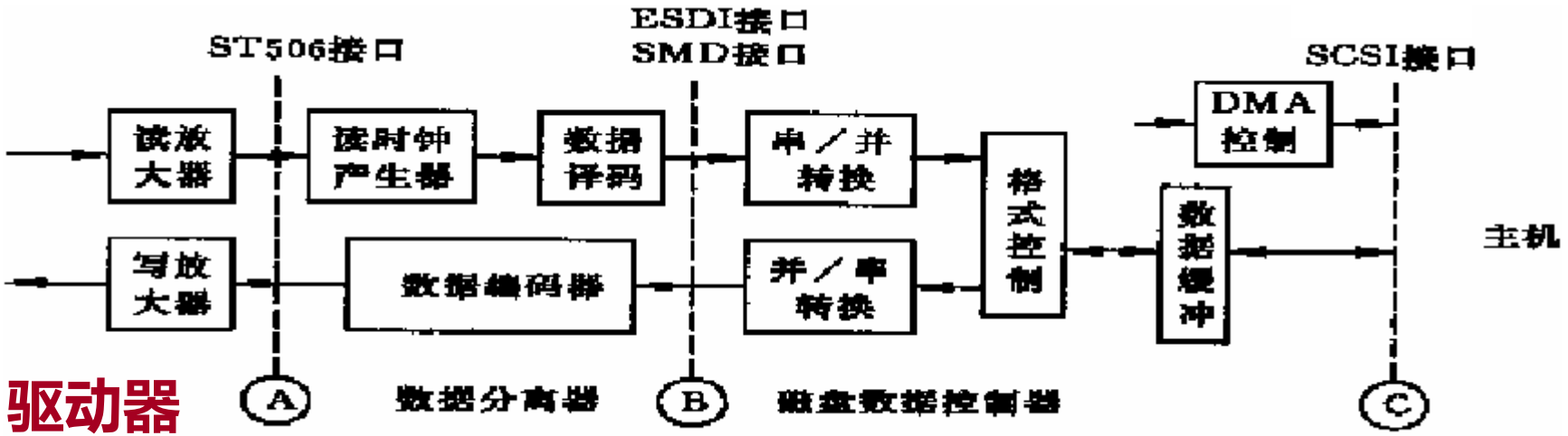
磁盘驱动器的结构及定位驱动系统



磁盘驱动器又称磁盘机，包括主轴、定位驱动系统和数据控制等.

磁盘控制器

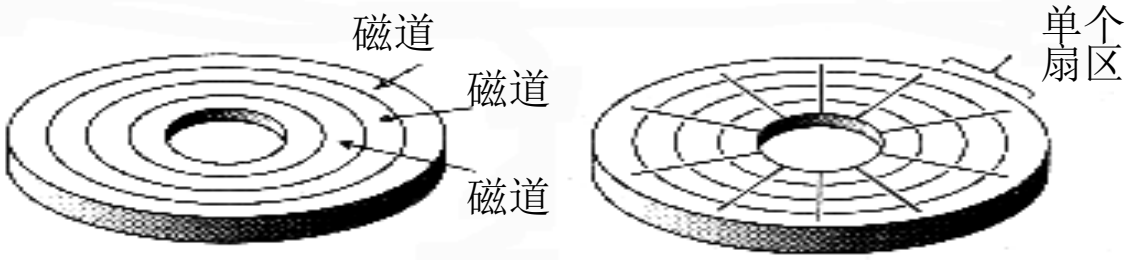
- 磁盘控制器是主机与磁盘驱动器之间的接口。
- 包含两个接口：
  - 对主机的接口，称作系统级接口
    - 界面比较清晰，只与主机的系统总线打交道，即数据的发送或接收，都是通过总线完成的。
  - 对硬盘（设备）的接口，称作设备级接口。
    - 可以放在多个不同的位置。

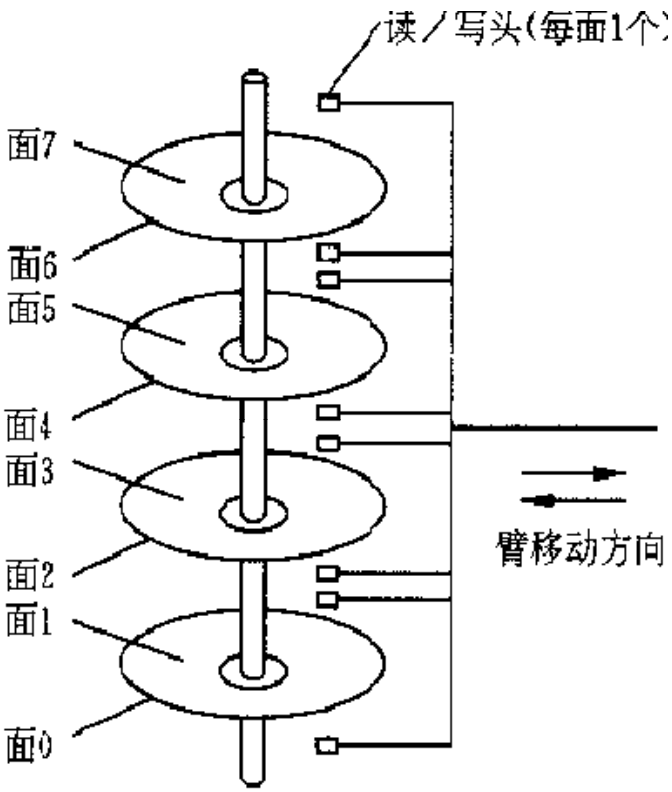


常见接口类型

- 台式机硬盘
  - IDE（Integrated Device Electronics）接口
    - 是一种类型的总称，采用16位数据并行传送方式，体积小，数据传输率可达到133Mb/s。一个IDE接口只能接两个外部设备。
    - 在实际应用中发展出多种类型，**如ATA、Ultra ATA、DMA、Ultra DMA等接口都属于IDE硬盘**。
  - SATA（Serial ATA）接口（“串口硬盘”）
    - 采用串行连接方式。具备了更强的纠错能力，支持热插拔
    - 数据传输率超过150Mb/s，新的接口规范达到600 Mb/s
- 工作站、服务器硬盘
  - SCSI接口（Small Computer System Interface）
    - 可以挂接7个设备。
  - 光纤通道

数据寻址—盘体、磁道、扇区和柱面



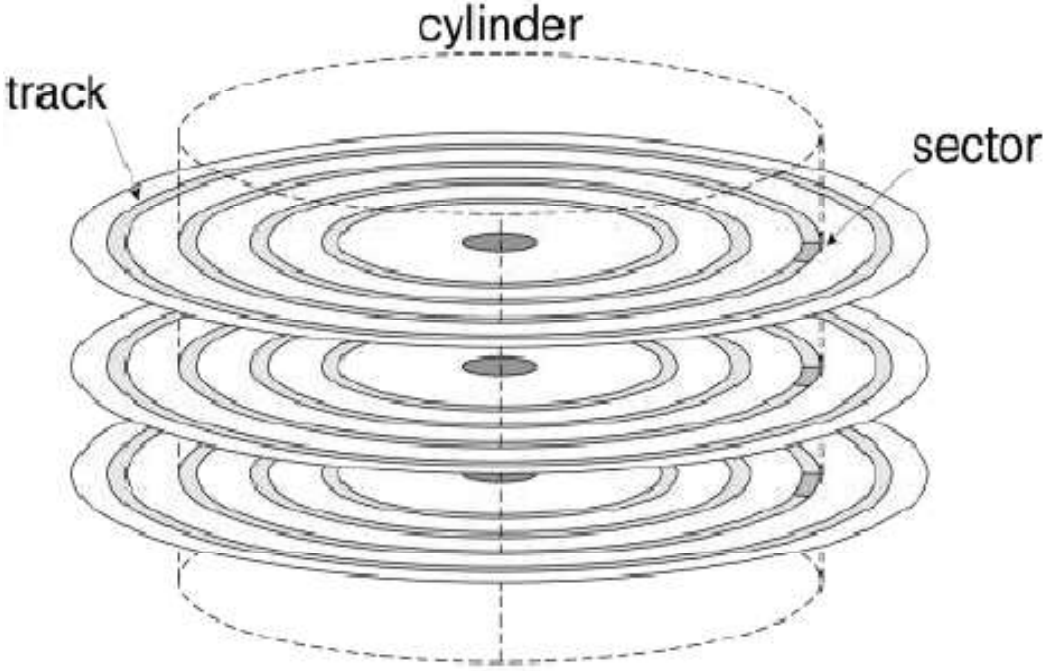


- **磁道**(Track)：磁面上均匀分布的同心圆存储轨迹。最外层为0磁道。
- **盘面**：磁盘组由多个同轴盘片组成，每个盘片都是双面存储，第一个盘片的第一面为0磁面，下一个为1磁面；第二个盘片的第一面为2磁面，以此类推。
  - 磁头号
- **扇区**(Sector)：磁道上等弧度划分的扇段。一般一个扇区的存储容量为512字节。
- **柱面**(Cylinder)：各个盘面上同一编号磁道的组合。
  - 柱面号

磁盘地址：

台号	磁道号	盘面号	扇段号

柱面



磁盘数据地址：

台号	磁道号	盘面号	扇段号

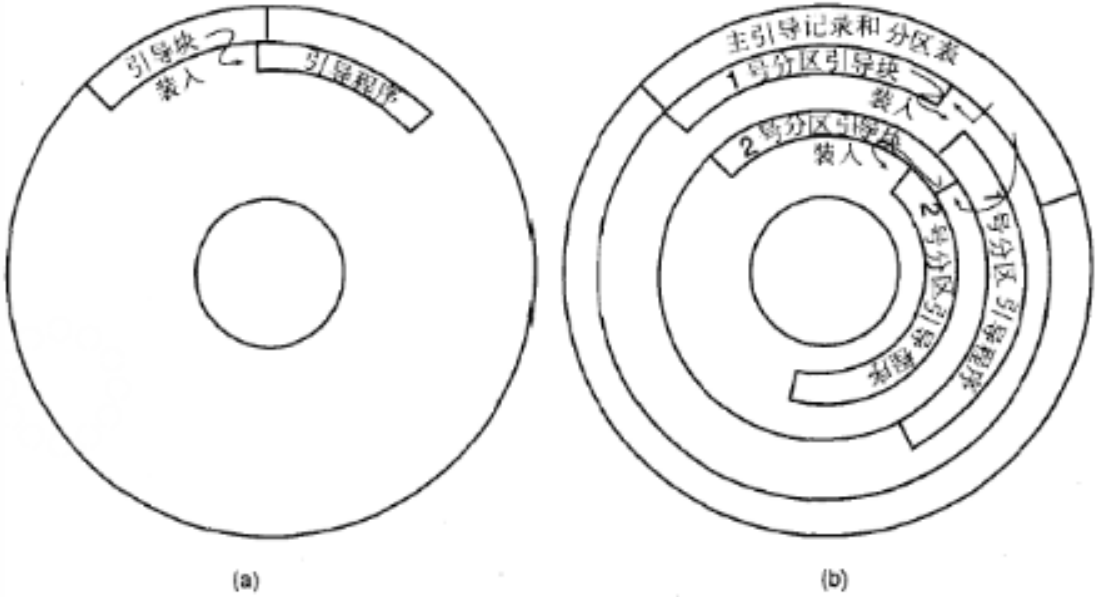
台号	柱面号	磁头号	扇段号

HDD寻址方式

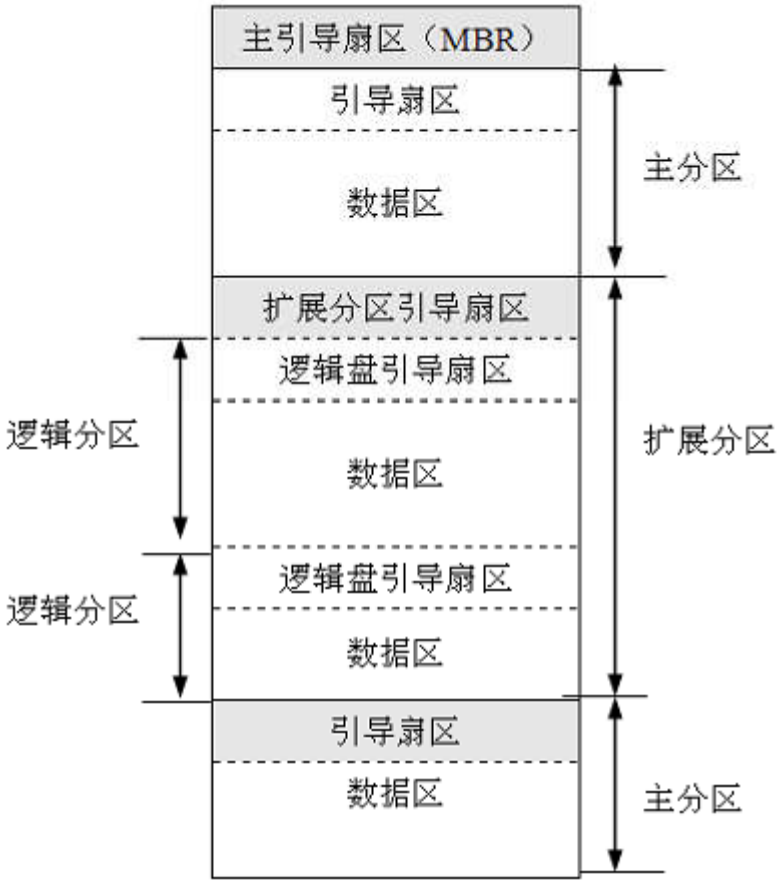
台号	柱面号	磁头号	扇段号

- C.H.S（Cylinder、Head、Sector）物理寻址方式
  - 硬盘容量 = 盘面数×柱面数×扇区数×512字节 。
  - 数据传输的开始地址写入4个8位寄存器：28位
    - 柱面地址16位[柱面低位寄存器（8位），柱面高位寄存器（8位）]，扇区地址8位，磁头地址4位（没有完全占用8位），最大容量为136.9GB
- LBA(Logical Block Addressing)逻辑块寻址模式（线性寻址模式）
  - 将磁盘上的所有扇区从0开始编号直到最大扇区数减1
  - 突破C.H.S模式的容量限制问题
    - 28位LBA硬盘寻址方式：137GB
    - 48位LBA硬盘寻址方式

逻辑盘：磁盘分区



引导使用的磁盘结构。(a)未分区的磁盘，第一扇区就是引导块。
(b)分过区的磁盘，第一个扇区是主引导记录



层次化信息记录结构

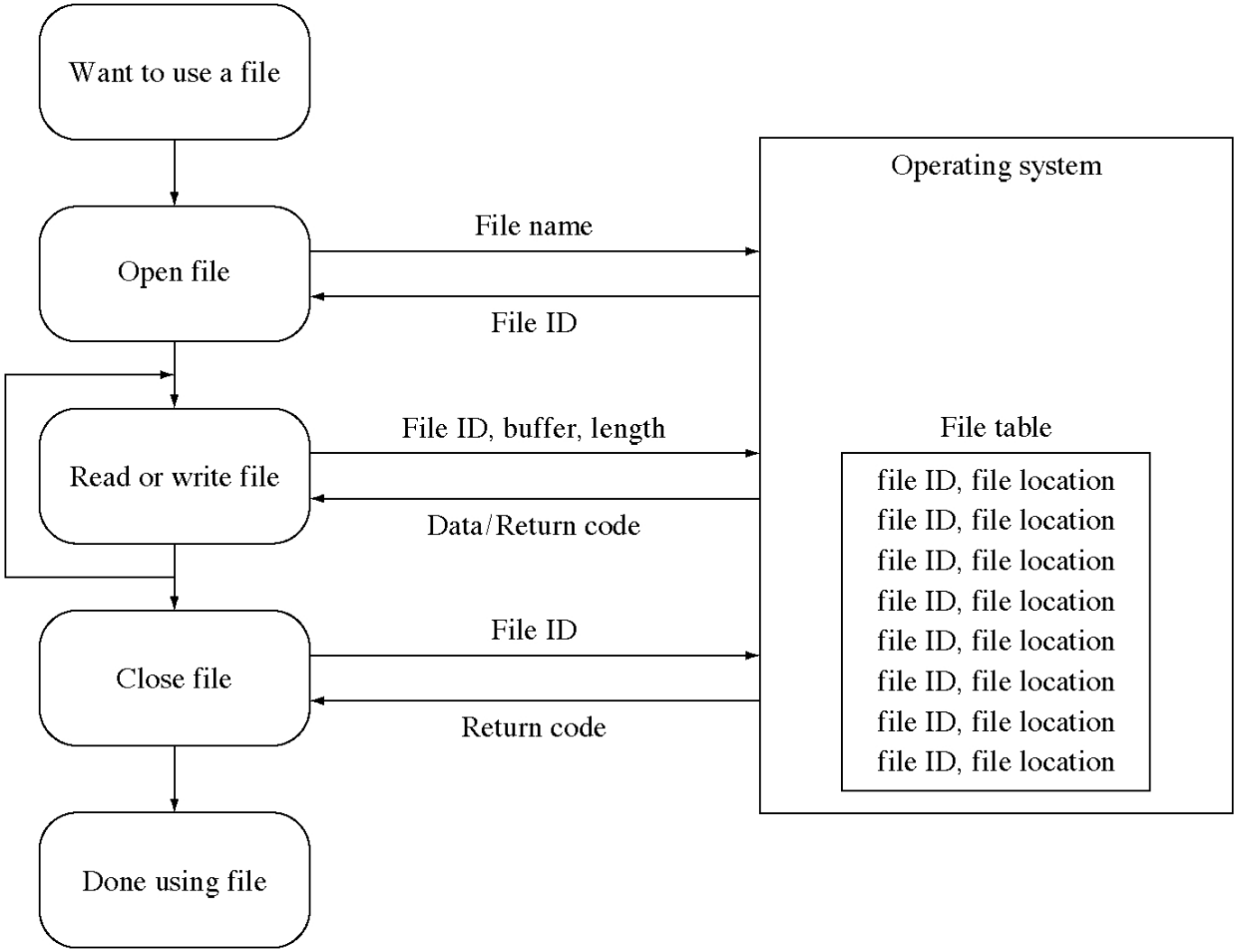
操作系统	文件、流数据
格式化记录	目录区、索引区、数据区（数据块）
物理层	磁道、扇区、位流

- 访问过程：
  - 应用程序->DOS功能（文件管理设备）->BIOS磁盘读写服务（INT 13中断）->IDE（ATA）接口->磁盘控制器
  - 传输数据：BIOS首先往IDE（ATA）的特定寄存器写入数据的开始地址和数据传输长度，再写入读/写命令

磁盘读写（BIOS int 0x13）

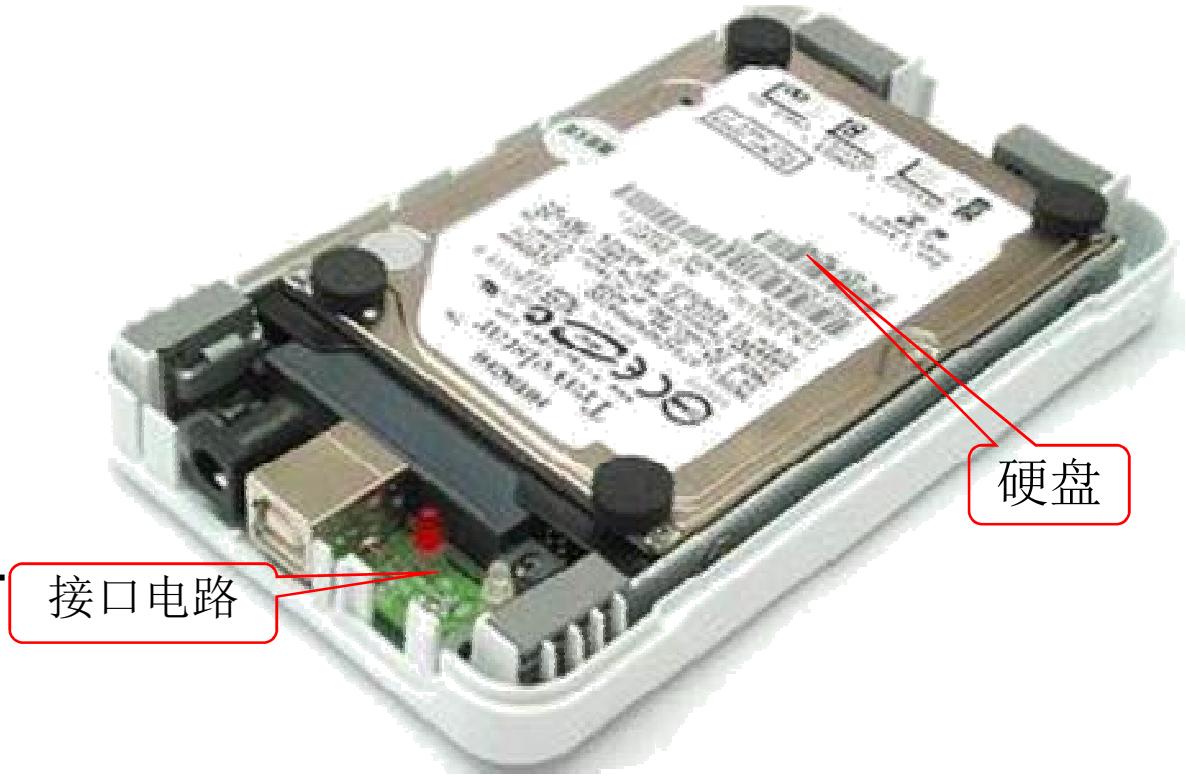
- 输入参数
  - AH=0x02(读盘)，=0x03(写盘) ,=0x04(校验）
  - AL=扇区数（同时处理连续的扇区）
  - CH=柱面号&0xff
  - CL=扇区号（0-5位）|(柱面号&0x300）>>2;
  - DH=磁头号
  - DL=驱动器号
  - ES：BX=缓冲区地址（校验寻道不使用）
- 返回值
  - FLACS.CF==0，没有错误，AH==0
  - FLAGS.CF==1，有错误，错误号存在AH内

Steps in using a file：syscall



移动硬盘

- 内装2.5英寸或1.8英寸的笔记本电脑硬盘
- 使用支持热插拔的USB或IEEE1394接口进行数据传输



移动硬盘的内部结构

硬磁盘存储器的发展：性能、可靠性

1. 提高磁盘记录密度
2. 提高传输率和缩短存取时间
   - 提高主轴转速，磁盘IO访问控制(“批处理”)，磁盘Cache
3. 硬盘数据保护技术
4. 半导体盘：用半导体材料制成的“盘”
   - 实际上并没有“盘”，而是以半导体芯片为核心，加上接口电路和其他控制电路组成的，在功能上模拟硬盘。
     - Flash Memory
5. 采用磁盘阵列RAID：性能、可靠性
   - 将并行处理技术引入磁盘系统。
     - 使用多台小型温盘构成同步化的磁盘阵列，数据分开存放。
   - 从外部看来为一个整体，可以像操作一台温盘那样操作，使数据传输时间为单台盘的1/n(n为并行驱动器的个数)。

磁盘IO访问控制(“批处理”)

- 多个进程共享DISK：决定块设备上IO操作提交的顺序
  - 提高IO吞吐量，降低IO响应时间（寻道时间）
- 软件方法：I/O调度器
  - FCFS
  - 在处理每一次I/O请求前，执行合并与排序的预处理操作
    - 读disk的第一个字节比读同一secter中的后续字节慢10万倍
    - 合并：访问多个相邻扇区的I/O请求被合并为一次I/O，只发给磁盘一条寻址命令，减少寻址次数
    - 排序：按照扇区增长排列I/O请求，一次旋转可访问更多扇区，缩短实际寻道时间
  - linux电梯算法：减小平均寻道时间
    - 假设有IO请求序列：100，500，101，10，56，1000
    - 按请求地址排序 ：100，101，500，1000，56 ，10
- 硬件方法：现代磁盘控制器Tagged command queuing（TCQ）优化
  - 可以通过由磁盘控制器对I/O请求进行重新排序来减少磁头的动作。
  - 通常，需要进行重组的I/O请求都会带有一个标识符，控制器在接收到这些I/O请求的时候会按照规则进行处理。

磁盘高速缓存(Disk Cache)

- 读写速度仍是限制系统整体速度的主要因素。
- 为减少对磁盘的频繁读写，在内存中开辟一块区域（磁盘缓冲区），一次尽可能多地将数据从磁盘读至该区，或将该区数据一次写入磁盘。
  - 逻辑上属于磁盘，而物理上是驻留在内存中的盘块。
- 写策略
  - write-through cache：磁盘启动频繁
  - 周期性地写回磁盘
    - UNIX SYNC的时间间隔定为30s
- 替换策略

硬盘数据保护技术

- 硬盘的MTBF已达30000～50000小时以上
  - 数据的完整性（ECC校验）
  - 防震保护技术
- S.M.A.R.T.技术 “Self-Monitoring，Analysis and Reporting Technology”
  - 可以通过监测指令对磁头、盘片、马达、电路的运行情况、历史记录及预设的安全值进行分析比较。当出现安全值范围以外的情况时，自动向用户发出警告。

RAID技术

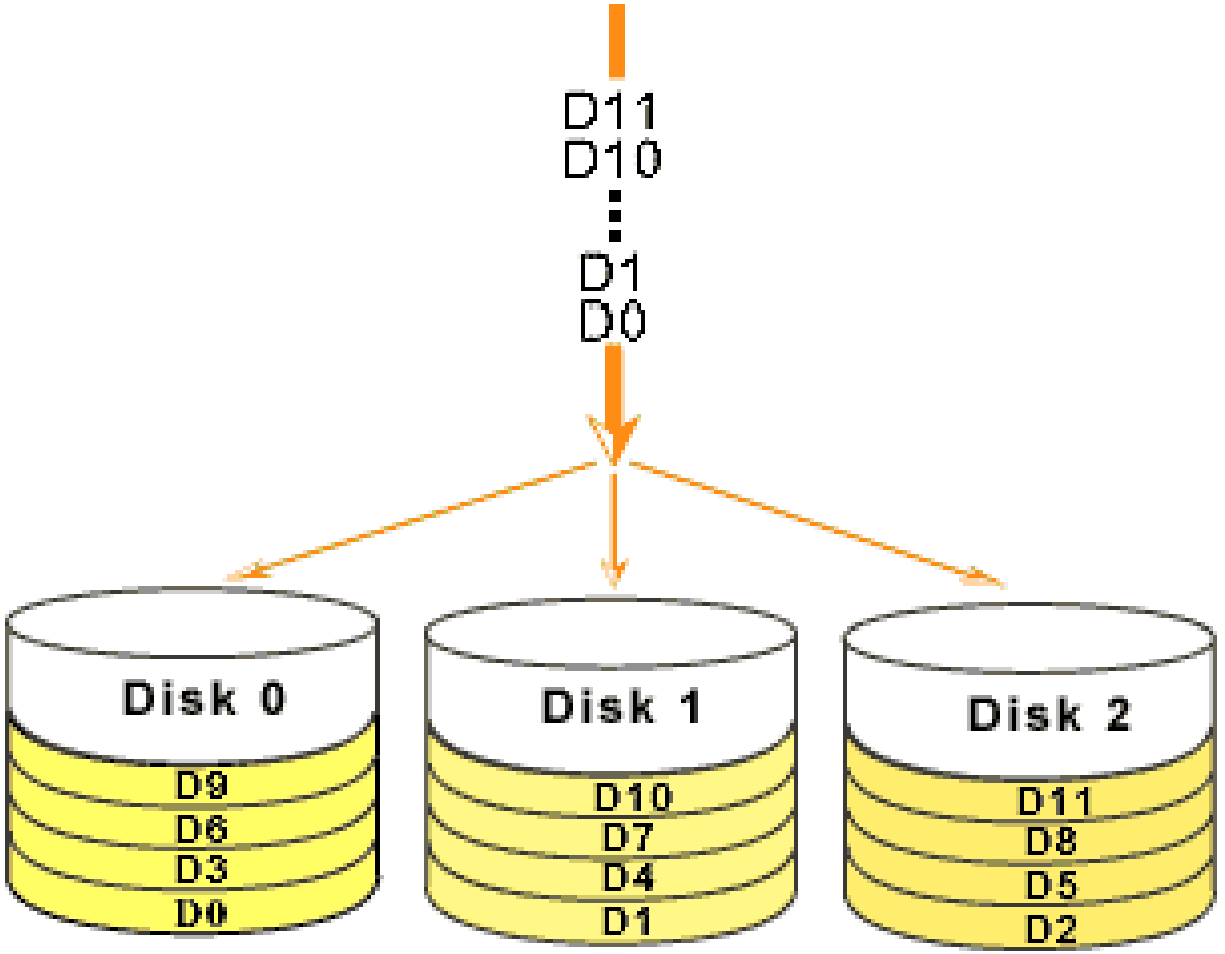
- 1987年，Patterson等@UCB
  - 将多只容量较小的、相对廉价的硬盘组合，使其性能超过一只昂贵的大硬盘
- Redundant Array of Independent Disk
  - 支持自动检测故障硬盘；
  - 支持重建硬盘坏轨信息；
  - 支持不须停机的硬盘备援(Hot Spare)
  - 支持不须停机的硬盘替换(Hot Swap)
  - 支持扩充硬盘容量

RAID级别

- RAID0：无差错控制的带区组
- RAID1：镜象结构
- RAID2：带海明码校验
- RAID3：带奇偶校验码的并行传送
- RAID4：带奇偶校验码的独立磁盘结构
- RAID5：分布式奇偶校验的独立磁盘结构
- RAID6：带有两种分布存储的奇偶校验码的独立磁盘结构
  - 对RAID5的扩展，主要是用于要求数据绝对不能出错的场合。
- RAID7：优化的高速数据传送磁盘结构
  - 采用并行和Cache技术

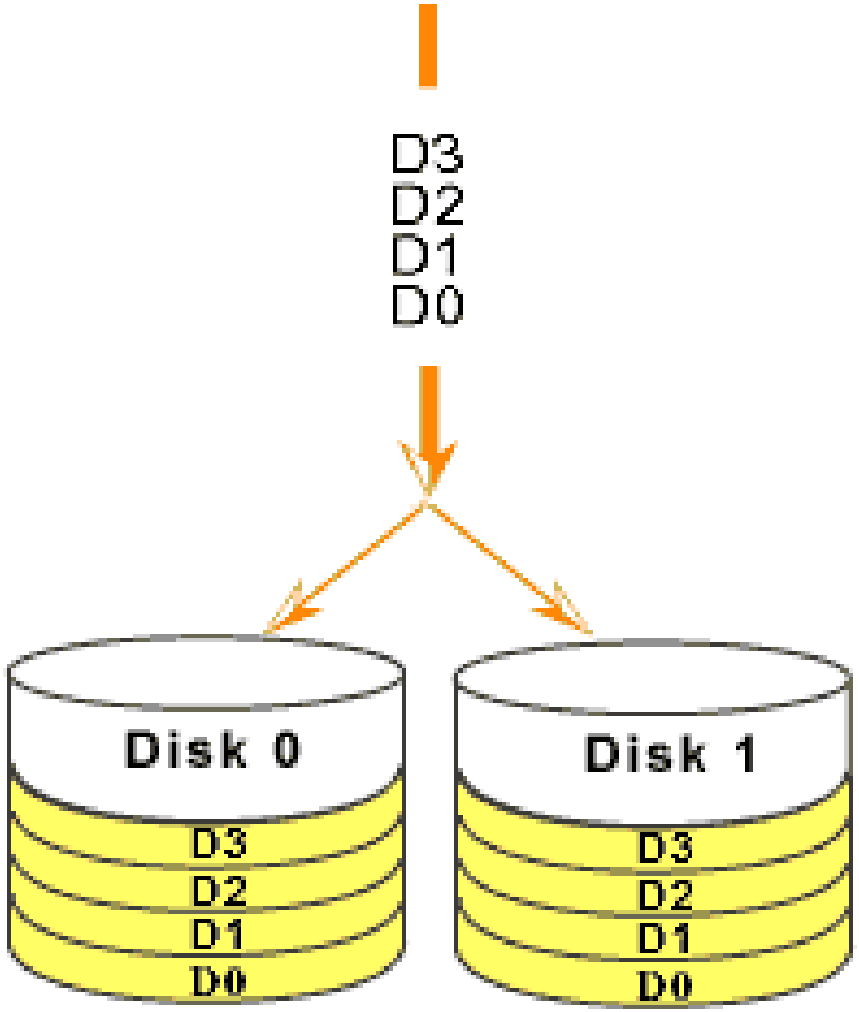
RAID 0（无差错控制的带区组）

- 目的：利用多体并行提高存储性能



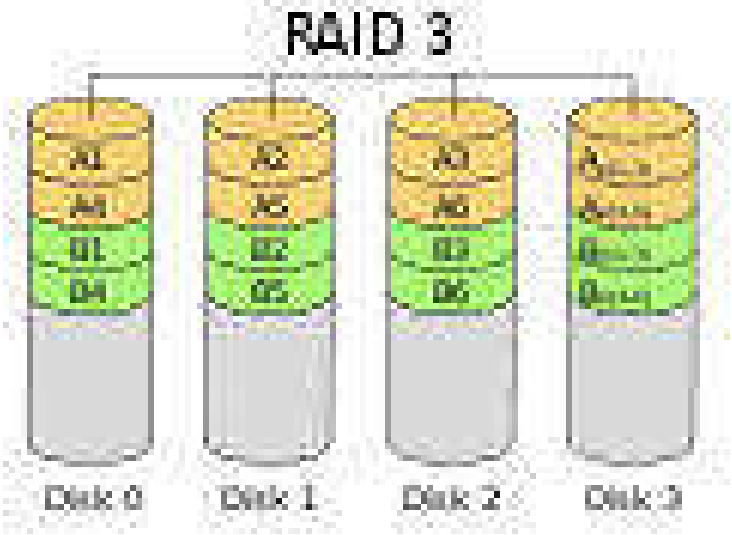
RAID 1（别名：镜像）

- 目标：保证数据的可用性和可修复性



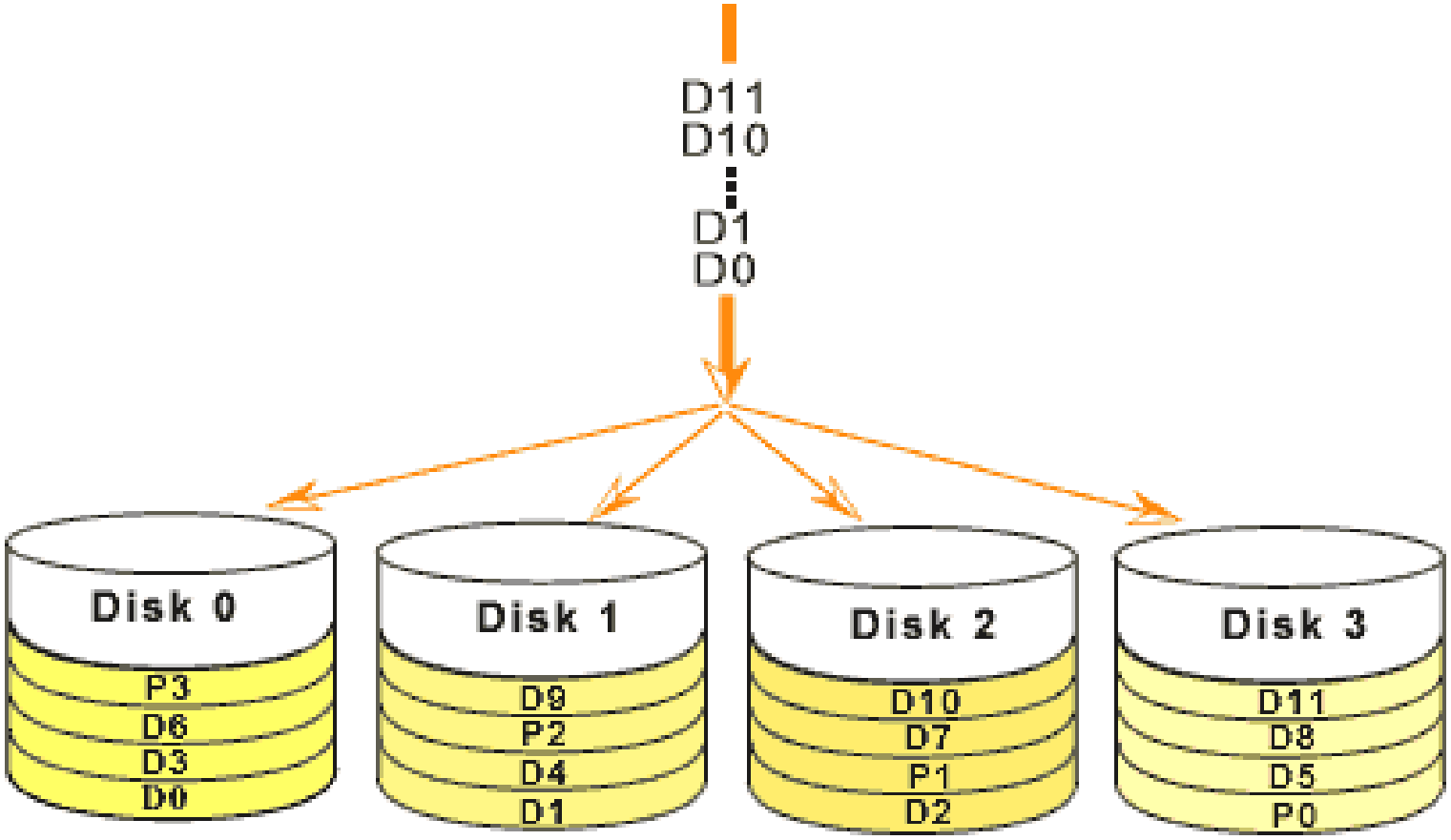
RAID3

- RAID3：带奇偶校验码的并行传送
  - 将数据条块化分布于不同的硬盘上
    - 条块单位为位或字节。
    - RAID4（少用）： 按数据块访问数据
    - RAID2（少用）：带海明码校验
  - 必须要有三个以上的驱动器
  - 校验码在写入数据时产生，保存在另一个磁盘上。

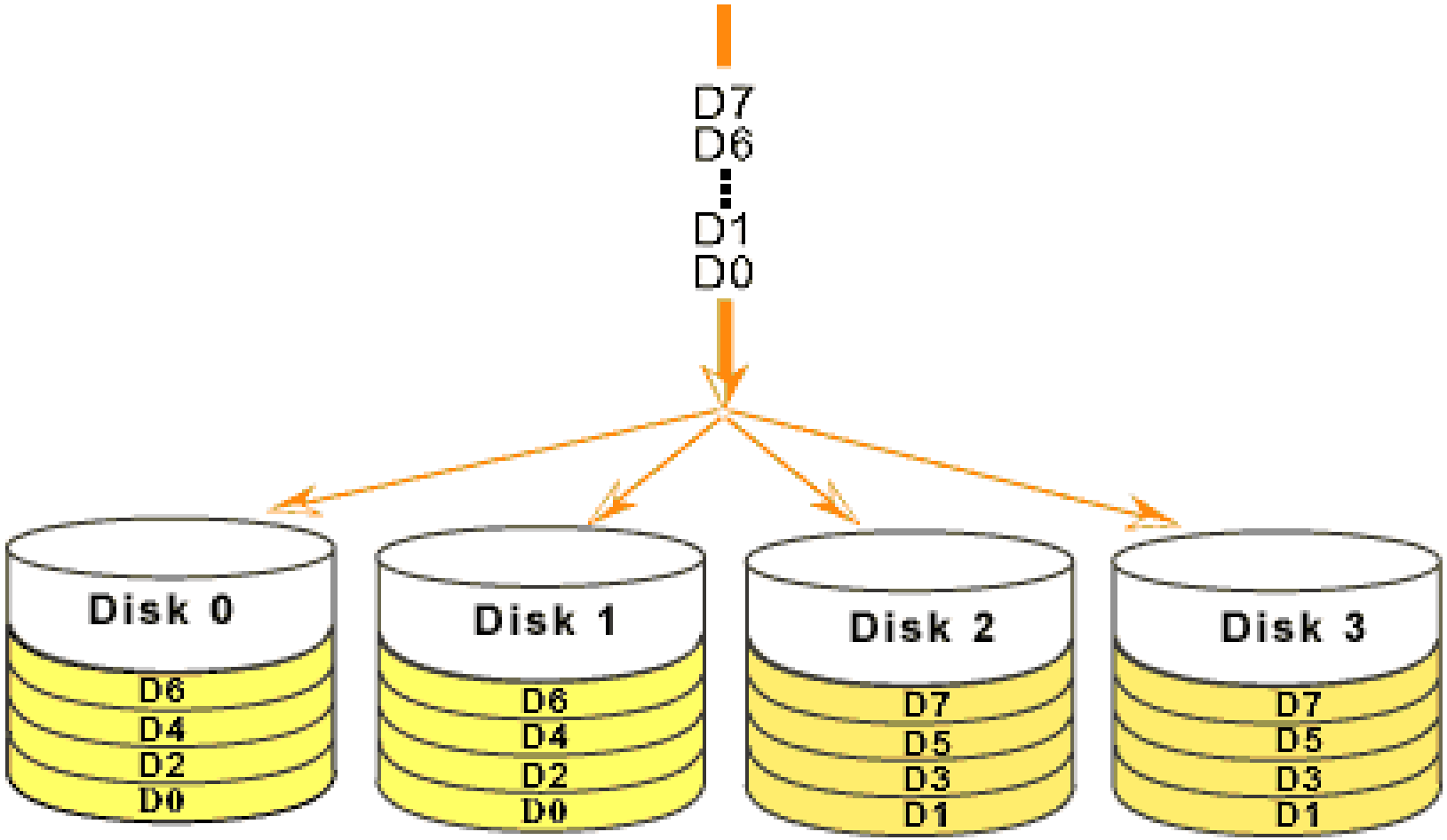


RAID 5

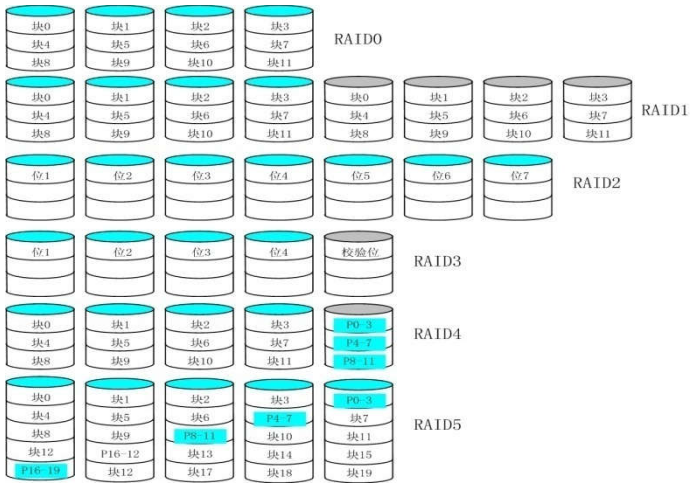
- 分布式奇偶校验的独立磁盘结构
  - 奇偶校验码存在于所有磁盘上



RAID 10 = RAID 0 + RAID 1



RAID选择

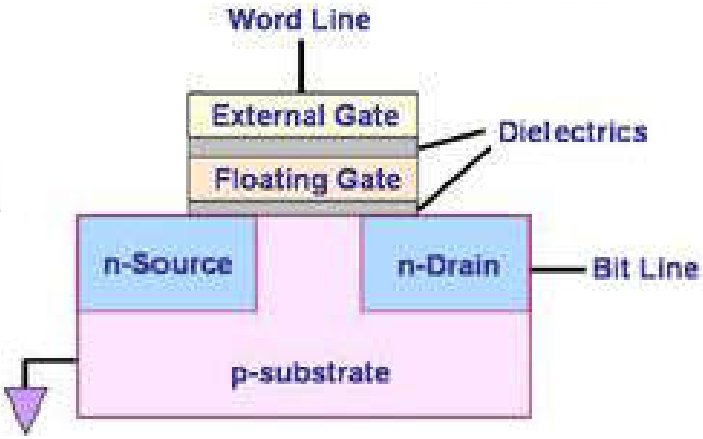


RAID级别	RAID-0	RAID-1	RAID-3	RAID-5	RAID-10
别名	条带	镜像	专用奇偶位条带	分布奇偶位条带	镜象阵列条带
容错性	没有	有	有	有	有
冗余类型	没有	复制	奇偶校验	奇偶校验	复制
热备盘选项	没有	有	有	有	有
读性能	高	低	高	高	中间
随机写性能	高	低	最低	低	中间
连续写性能	高	低	低	低	中间
需要的磁盘数	一个或多个	只需2个或2xN个	三个或更多	三个或更多	只需4个或4xN个
可用容量	总的磁盘的容量	只能用磁盘容量的50%	（n-1）/n的磁盘容量。其中n为磁盘数	（n-1）/n的总磁盘容量。其中n为磁盘数	磁盘容量的50%
典型应用	无故障的迅速读写，要求安全性不高，如图形工作站等	随机数据写入，要求安全性高，如服务器、数据库存储领域	连续数据传输，要求安全性高，如视频编辑，大型数据库等	随机数据传输要求安全性高，如金融，数据库，存储等	要求数据量大，安全性高，如银行，金融等领域

Flash存储器

- 移动存储设备
- 分为两种
  - 容量较小的NOR Flash
    - 可靠性较高、随机读取速度快
    - 擦除和编程速度较慢
  - 容量较大的NAND Flash
    - 体积小，零噪音，抗振动，容量大，成本低
    - 4G、8G、32GB乃至更大容量
    - 适合做文件系统

NAND Flash

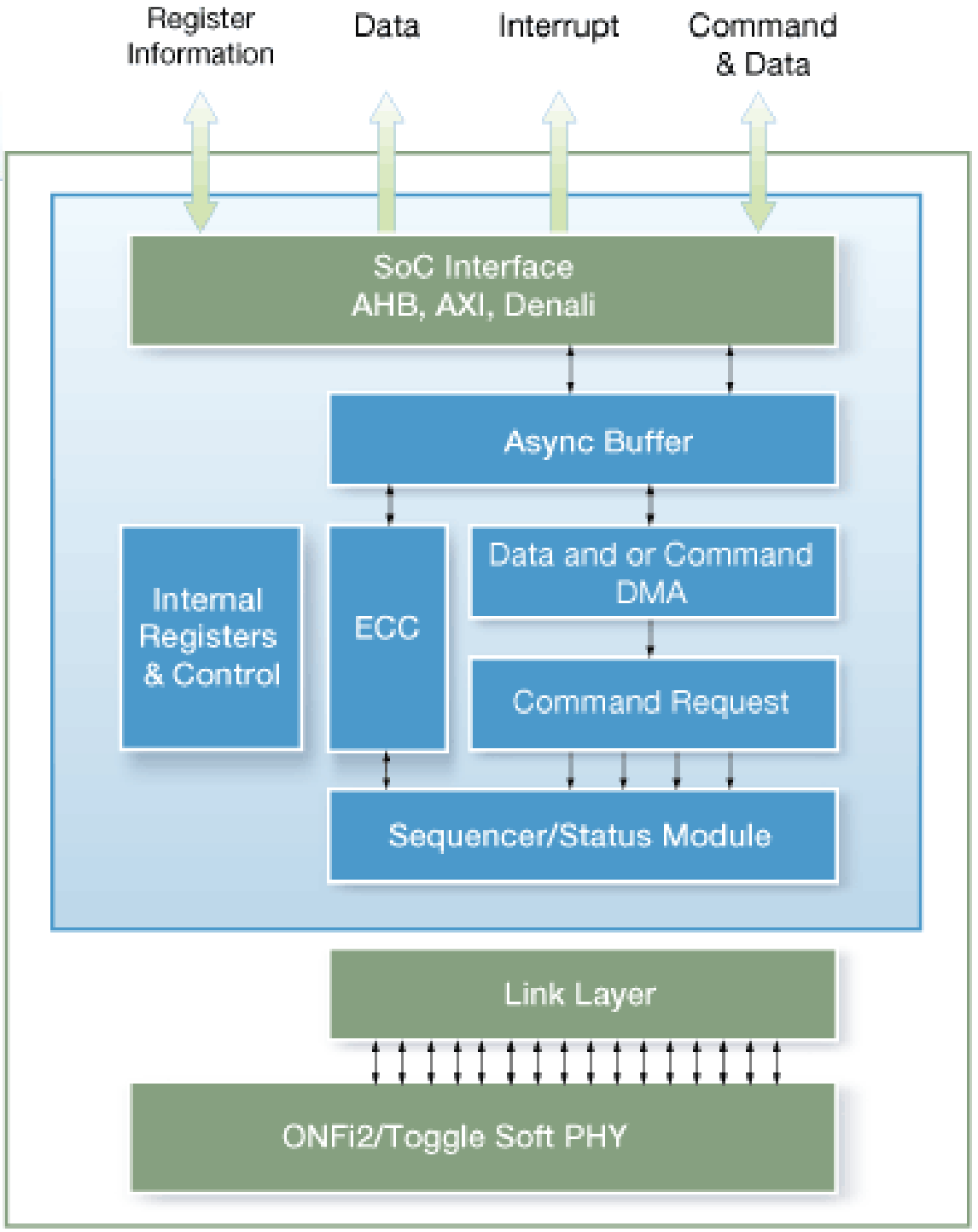


- 外部门（external gate）
  - 被施加的电压决定存储电荷的多少
  - 栅晶体管的结电容可长时间保存电压值，使之断电后能保存数据。
- 数据的表示
  - 以存储的电荷的电压是否超过一个特定的阈值Vth来表示
  - SLC（Single Level Cell）
    - 只存储一种状态：所存储电荷的电压如果大于阀值则为1，反之为0
  - MLC（Multi Level Cell）
    - 提升了NAND Flash的存储密度
- 擦写上限
  - 一般MLC型为几万次，SLC为几十万次

操作

- 读和写（编程）操作
  - 以页为单位，1页为512B、2KB、4KB或者更大
- 擦除操作
  - 以块为单位，1块为 64K、256K、512KB或者更大
- 写入操作无法实现原地覆盖，写入之前必须进行额外的擦除

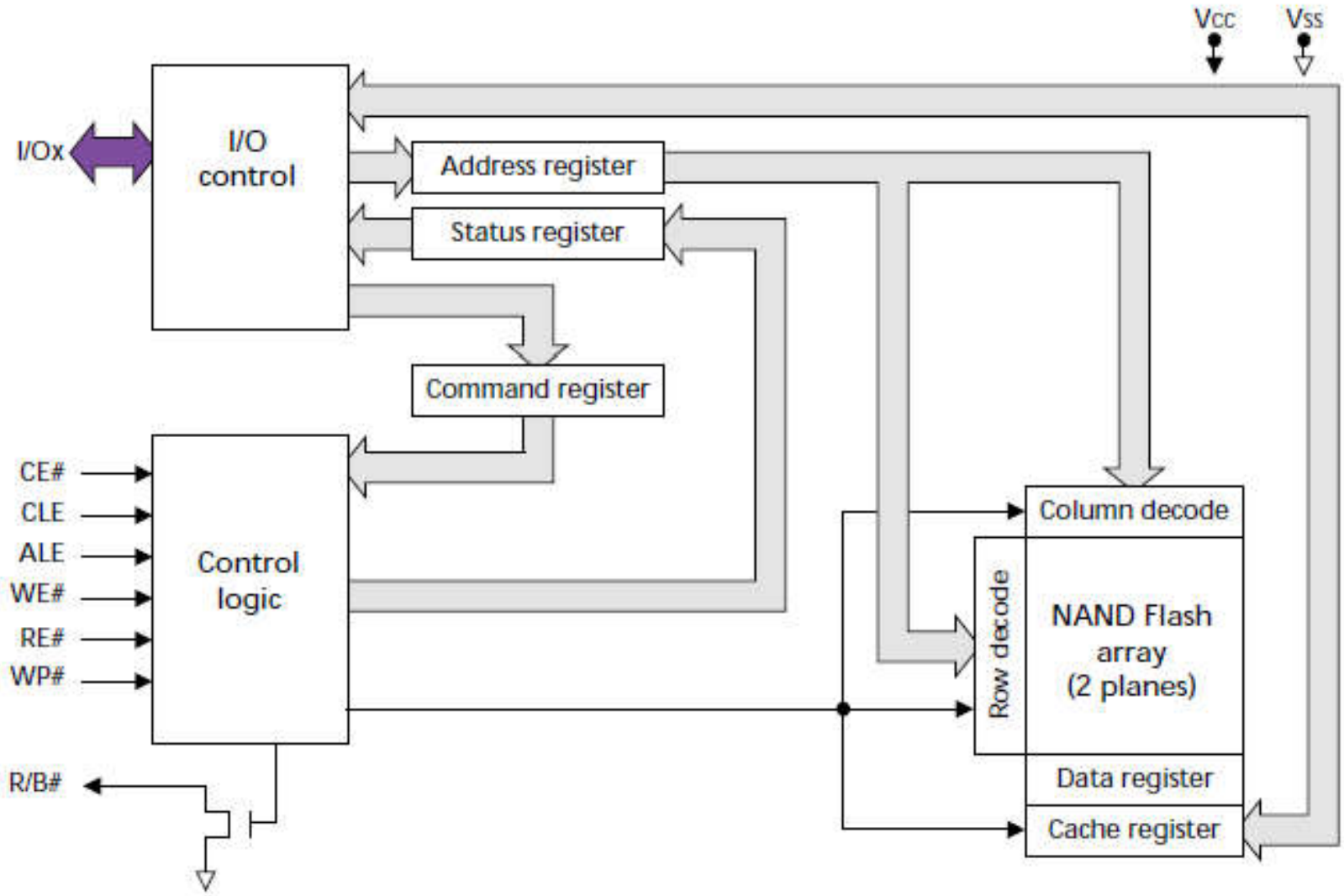
Flash控制器


Register Information
Data
Interrupt
Command & Data
SoC Interface AHB, AXI, Denali
Async Buffer
Internal Registers & Control
ECC
Data and or Command DMA
Command Request
Sequencer/Status Module
Link Layer
ONFi2/Toggle Soft PHY

NAND Flash 命令集

Command	Command Cycle 1	Number of Address Cycles	Data Cycles Required[1]	Command Cycle 2	Valid During Busy	Notes
PAGE READ	00h	5	No	30h	No	2
PAGE READ CACHE MODE SEQUENTIAL	31h	–	No	–	No	3
PAGE READ CACHE MODE RANDOM	00h	5	No	31h	No	4
PAGE READ CACHE MODE LAST	3Fh	–	No	–	No	
READ for INTERNAL DATA MOVE	00h	5	No	35h	No	2, 5
RANDOM DATA READ	05h	2	No	E0h	No	2
READ ID	90h	1	No	–	No	
READ PARAMETER PAGE	ECh	1	No	–	No	
READ STATUS	70h	–	No	–	Yes	
PROGRAM PAGE	80h	5	Yes	10h	No	2
PROGRAM PAGE CACHE MODE	80h	5	Yes	15h	No	2
PROGRAM for INTERNAL DATA MOVE	85h	5	Optional	10h	No	2, 5
RANDOM DATA INPUT	85h	2	Yes	–	No	2
BLOCK ERASE	60h	3	No	D0h	No	2
RESET	FFh	–	No	–	Yes	2
OTP DATA PROGRAM	A0h	5	Yes	10h	No	
OTP DATA PROTECT	A5h	5	No	10h	No	
OTP DATA READ	AFh	5	No	30h	No	
SET FEATURES	EFh	1	4	–	No	
GET FEATURES	EEh	1	No	–	No	

控制器与NAND芯片的连接



USB闪存盘（小容量）

- **填补了中国计算机存储领域20年来发明专利的空白**
- USB2.0接口传输速率480Mbps
  - 目前还无法达到（20-40MB/s）。
  - USB 3.0标准已经由Intel公司提出，其传输速率在5GBps以上。
- 数据传输率
  - 数据读取最大速率可达900KB/s
  - 数据写入速率最大可达700KB/s
- 数据类型：三种
  - 命令块包（CBW）、命令执行状态包（CSW）、数据包

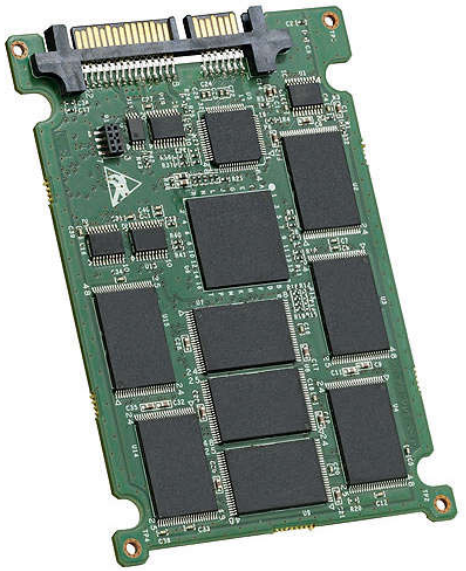
闪存盘的内部结构图

NAND Flash大容量存储解决方案

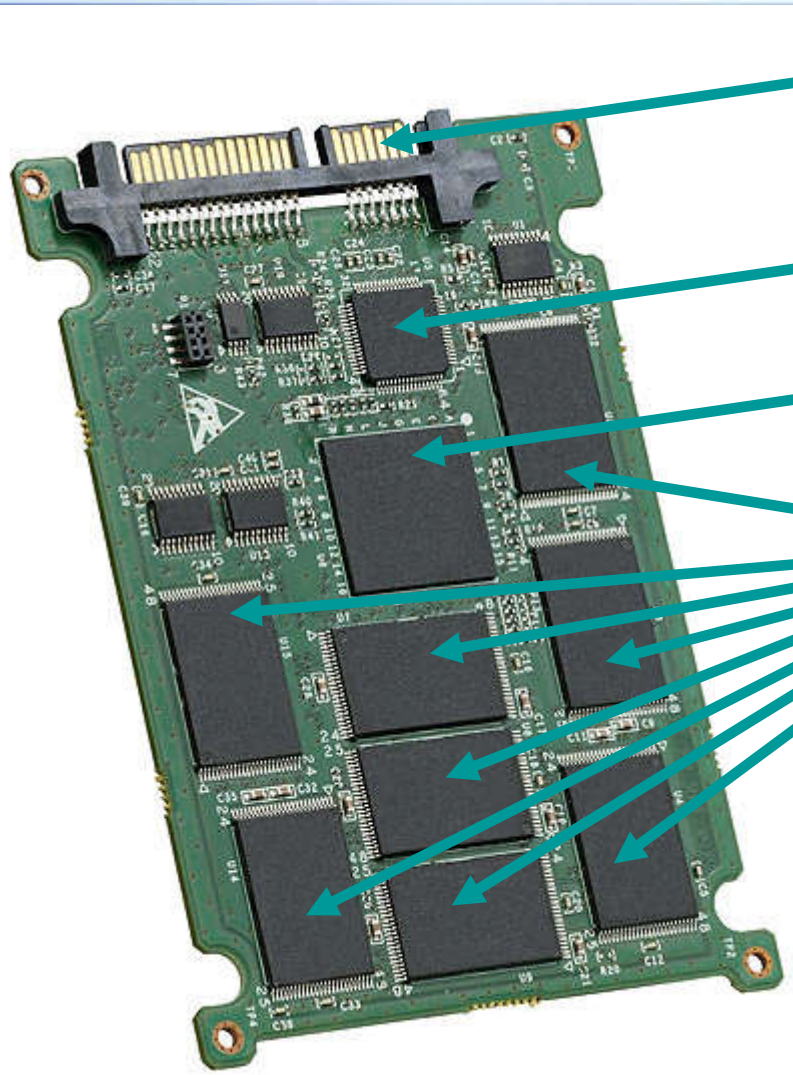
- 方案一： FTL（Flash Translation Layer）
  - 将NAND Flash通过闪存转换层FTL模拟成类似磁盘的块设备
    - FTL主要实现地址映射，磨损均衡，垃圾回收，坏块管理，ECC校验，缓存算法
  - 通过块设备层驱动提供对文件系统透明的块设备操作接口，使用传统的文件系统进行管理。
- 方案二：采用Flash存储器的专用文件系统
  - 对Flash硬件特性的处理直接交给Flash 文件系统去完成
    - JFFS (The Journaling Flash File System)、YAFFS (Yet Another Flash File System)、UBIFS (Unsorted Block Image File system) 等

方案一：固态硬盘 (SSD)

- 极其坚固耐用
  - 扩展的工作温度 （0ºC 到 70ºC）
  - 冲击和震动方面不存在任何问题
- 卓越的读取性能
  - > 50x SATA 随机读取性能
  - > 15x SAS 随机读取性能
  - 无需搜索时间，因此 IOPS 很高
  - 写入性能有限（与 15k SAS 相关）
- 增强的可靠性
  - 无移动部件：
  - 可消除对 RAID SSD 的需求
- 功耗降低 **10** 倍
  - 不到 2 瓦，而 15k 2.5 英寸 SAS 为 9 瓦
- 散热、尺寸和噪音优势
  - 无噪音，低热量
  - 轻盈小巧



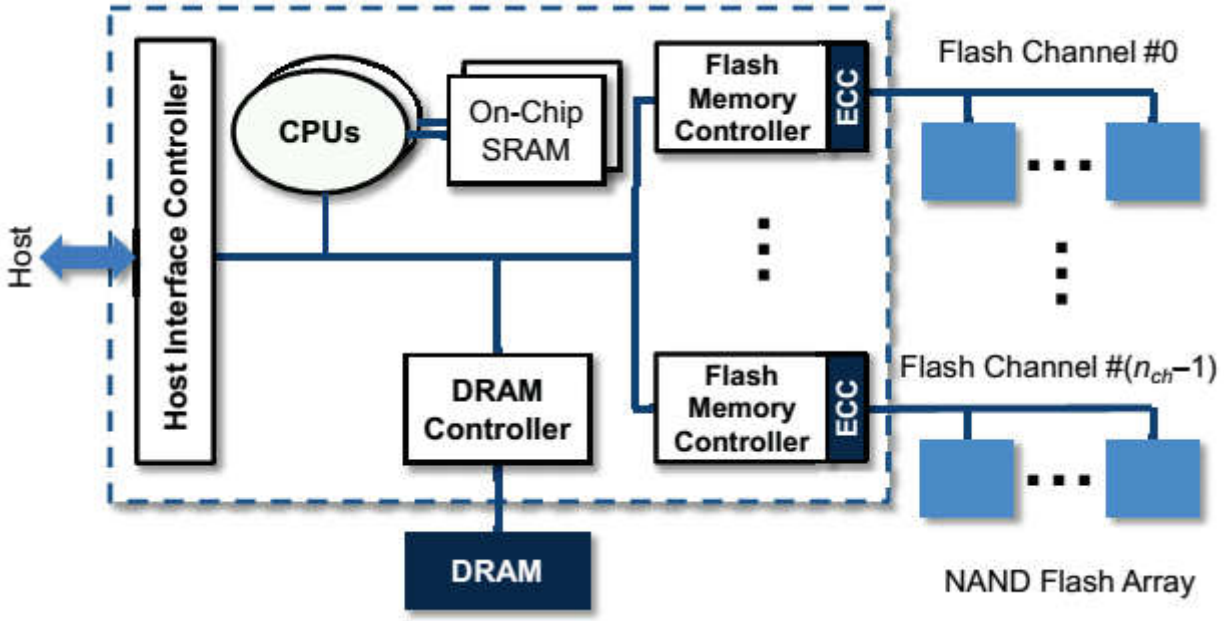
固态硬盘 (SSD，Solid State Disks)



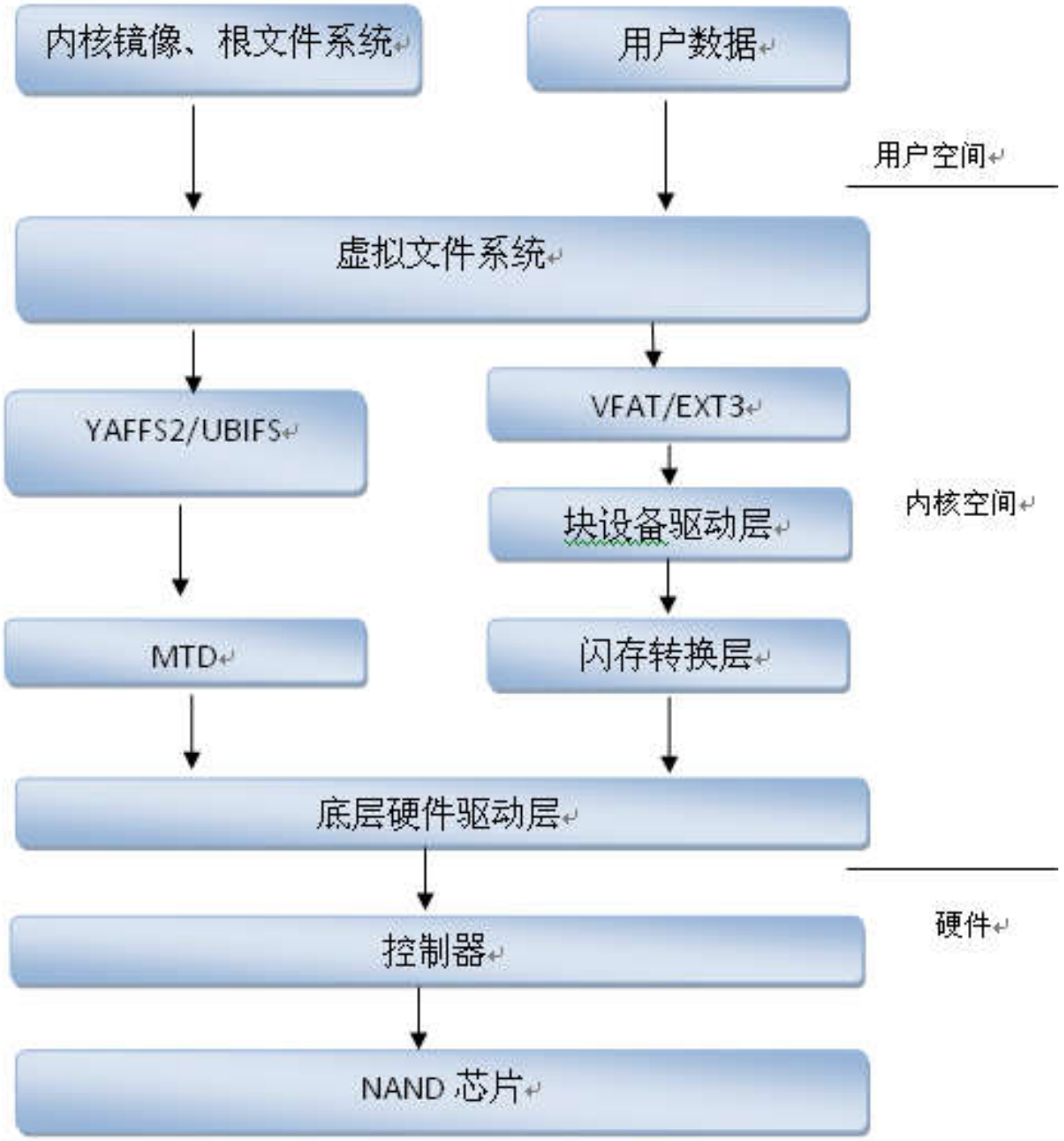
- **BL495c G5 专用 SSD：**
  - 1.5Gb SATA 接口
  - 非热插拔“裸板”
  - 容量为 32GB 和 64GB
  - BL495c G5 最多支持 2 个 SSD
  - 目标应用： 启动驱动器

the general architecture of an SSD

- 组成：host interface controller, embedded CPU(s), on-chip SRAM, DRAM and flash memory controllers connected to the flash chips.
- On top of the hardware substrate runs the SSD firmware commonly referred to as flash translation layer (or FTL)



方案二：

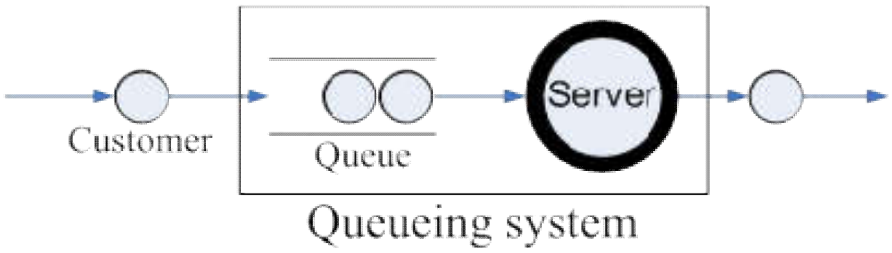

内核镜像、根文件系统
用户数据
用户空间
虚拟文件系统
YAFFS2/UBIFS
VFAT/EXT3
块设备驱动层
内核空间
MTD
闪存转换层
底层硬件驱动层
控制器
硬件
NAND 芯片

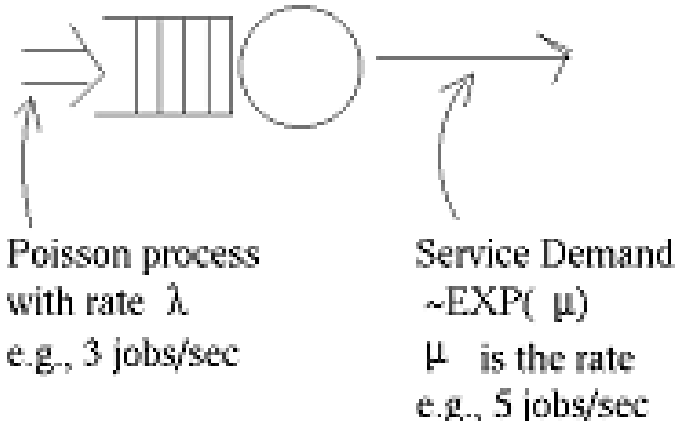
分析计算机外存系统性能

- 例：磁盘阵列的I/O响应时间？
- 性能评测方法
  - 建模（modeling）、仿真（simulation）、测量（measurement）

	建模	仿真	测量
待测系统	任何	任何	已有系统
测试所需时间	少	适中	不确定，受被测系统和测试工具的影响
所用工具	数学	计算机语言	设备
精确度	低	一般	不确定，受环境影响
评价折衷的能力*	强	适中	弱
开销	小	一般	大
市场信任度	低	一般	高

建模：排队论

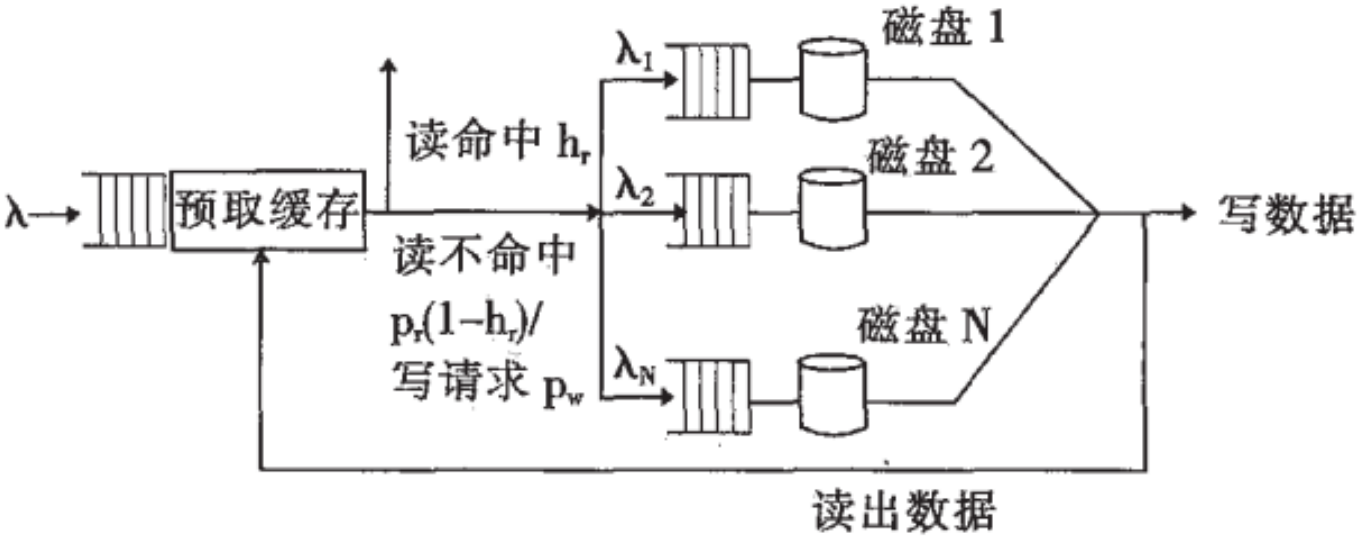




- 指标：顾客的平均等待时间，服务台的忙闲程度
- 两类经典模型：到达时间/服务时间/窗口数
  - 一个服务窗，顾客按参数为λ的泊松分布到达，到达的时间间隔为负指数分布
    - M/M/1：服务窗为每个顾客服务的时间为负指数分布M（马尔可夫），平均服务率为μ
    - M/G/1：服务窗为每个顾客服务的时间是一般分布G（随机）

例：磁盘阵列的I/O响应时间

- 从接收访问请求到完成服务所经历的时间
  - 由排队延迟、寻道时间、旋转等待时间、数据传输时间及调度软件开销决定
- 根据排队论
  - 在串联的排队系统中，假如到达过程为泊松分布，而每个服务台的服务时间服从互相独立的负指数分布，那么每个服务台可分开考虑，此时，串联排队系统可以简化为单独的排队系统。

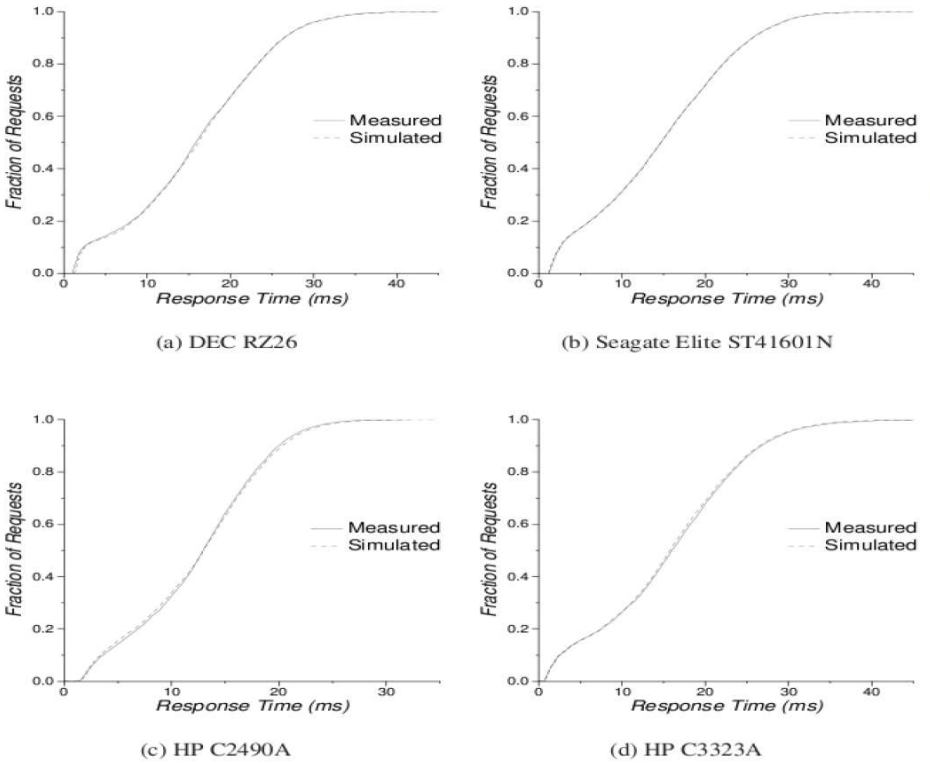


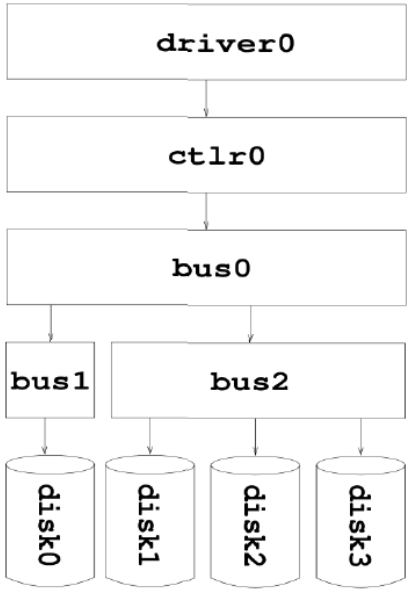
磁盘系统仿真器DiskSim

- Accurate, Highly-Configurable Storage Subsystem Simulator
  - Developed in Parrallel Data Labratory, CMU
  - delevoped in Linux environment
- Capabilities:
  - Simulate a hierarchy of storage components such as buses and controllers (e.g. RAID arrays) as well as disks
  - Using for perfomance evaluation
  - Can be integredet into full system simulators as a disk model
  - Model performance behaviour, but not actual data for each request.

示例

- I/O Driver Statistics
  - Idle time
  - Response time
- Disk/SSD Statistics
  - Idle time
  - Response time
  - IOPS
- Bus Statistics
  - Utilization time
  - #arbitrations
- Controller Statistics
  - Report disk cache subcomponent statistics
  - #misses/hits
  - #destages





SSD Extension

- Patches over DiskSim
  - Developed by Microsoft Research
  - Patched version Available in the DSN Lab. resource directory.
- Provide limited support for solid-state-disk (SSD) simulation.
- Not a simulator for any specific SSD, but rather a simulator for an idealized and parameterized SSD (was not Validated)

小结

- 硬盘、固盘、RAID、性能评估
  - 磁表面存储器原理与磁盘记录格式
    - 非格式化记录
    - 格式化：BOOT区、ROOT区、FAT表、数据区
  - 硬盘是机械设备，需进行针对性优化
    - 直接访问：OS的磁盘（臂）I/O调度器
    - 磁盘高速缓存
  - 磁盘的I/O响应时间？
  - 如何访问磁盘数据的过程？
- 作业
  - C语言读盘程序设计？（可选）：块方式，文件方式
  - 4.38、44


Thank You