计算机组成原理

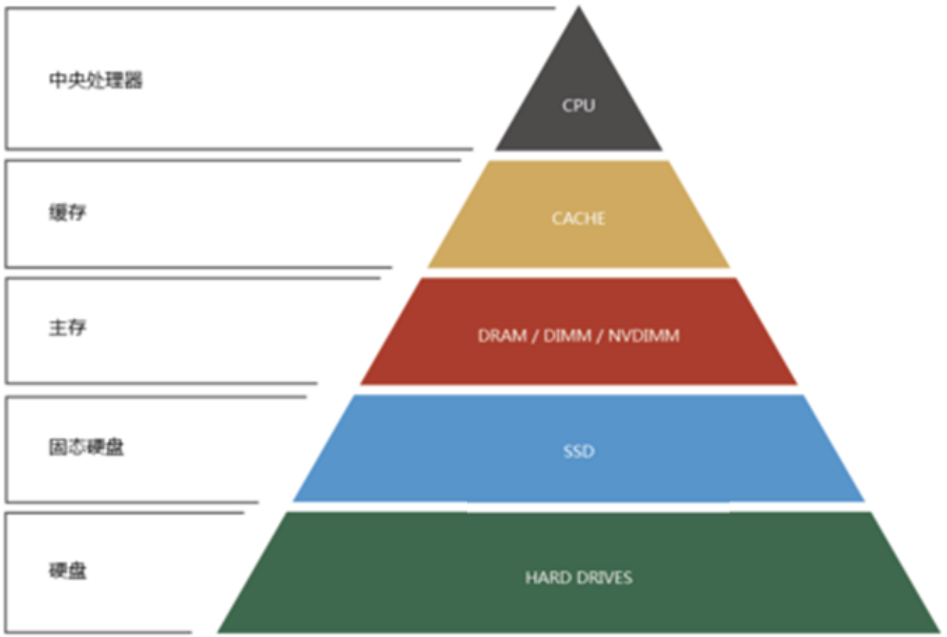
外存储器

唐$4.4，RV $5.2.4，$5.16

llxx@ustc.edu.cn

本章内容

- 辅助存储器
  - *半导体：Flash*
    - *U盘，SSD*
  - 磁表面存储器：Disk
    - 性能参数
    - 编码方式
  - 在线，*离线*
- 硬盘hard drives原理
  - 访问方式：直接访问
  - 控制器与接口
  - 记录格式
  - 磁盘I/O调度
- 信息容错



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

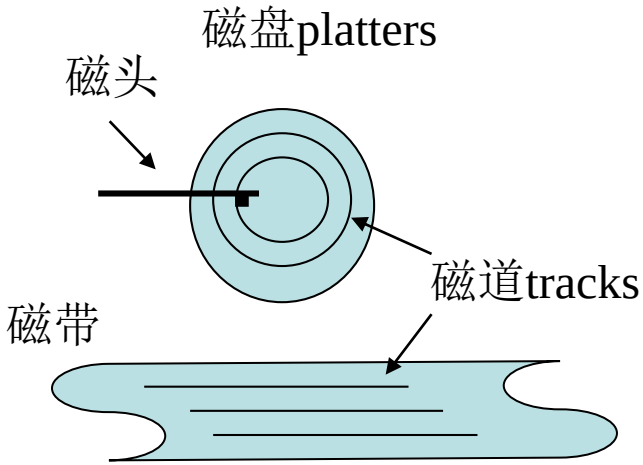
辅助存储器的特点

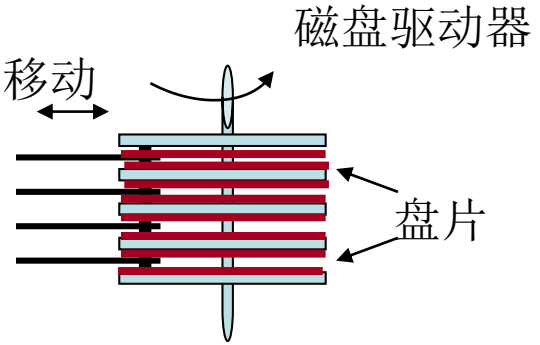
- 外存
  - 硬盘、软盘、磁带、光盘（CD ROM）
  - 容量大，GigaBytes
  - 速度慢, 7200转/min，速率<100Mb/s
    - RAM：几百兆(存取周期几十纳秒)
  - 价格低，80G/￥800.00
    - 内存：256M/￥400.00
  - 可脱机保存信息，具有非易失性的特点

磁表面存储器

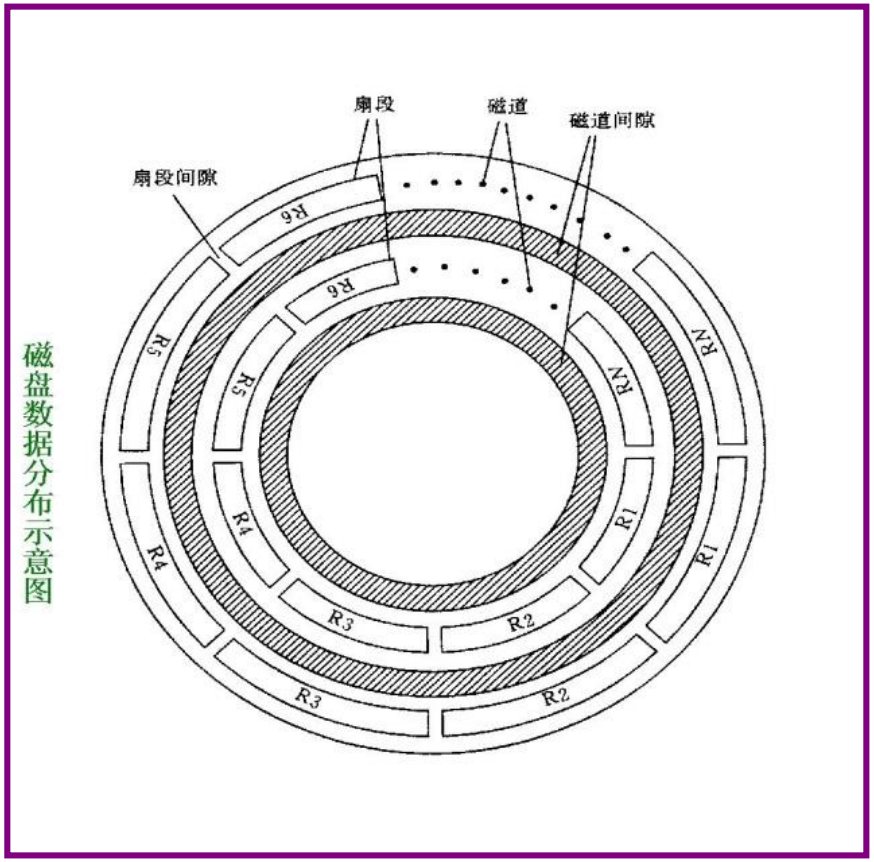
- 主要内容
  1. 技术指标
     - 记录密度、容量、寻址时间、传输率、误码率
  2. 磁记录原理
  3. 磁盘记录格式
  4. 评价记录方式的主要指标
  5. 硬磁盘存储器
  6. *软磁盘存储器*
  7. *磁带存储器*

磁记录设备





- 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

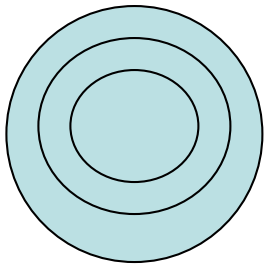
$$D_t = \frac{1}{P}$$

  - 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）

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

每道总位数，各道相同

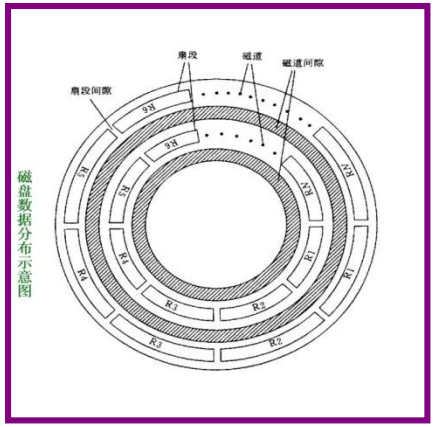
同心圆最小直径



技术指标 - 寻址时间

- 磁盘寻址过程：**直接**存取
  - **随机**寻道，扇段**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

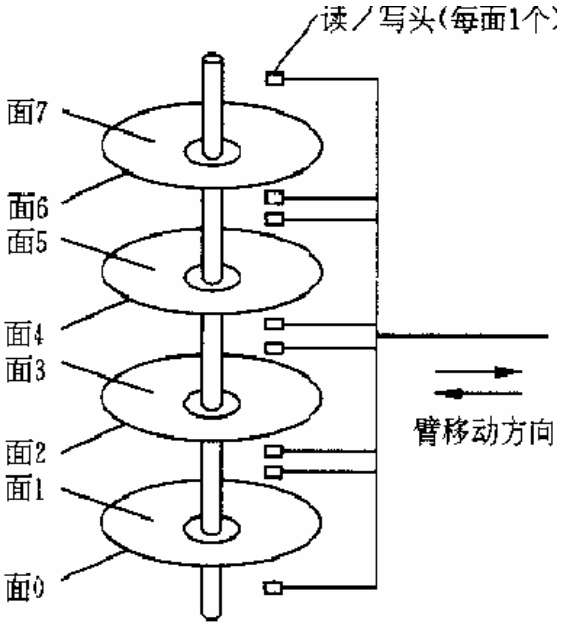
技术指标 - 传输率、误码率

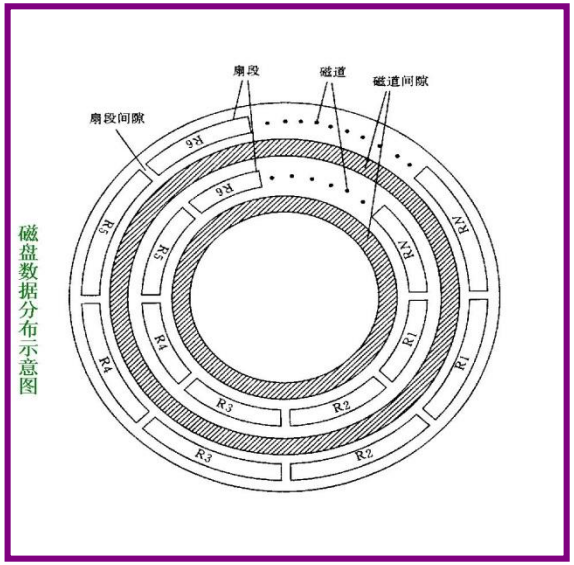
- **传输率**：
  - 单位时间传输的数据量（字节、位）
  - D_r = 记录密度（D）×介质运行速度（V）

- **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。

技术指标 - 容量

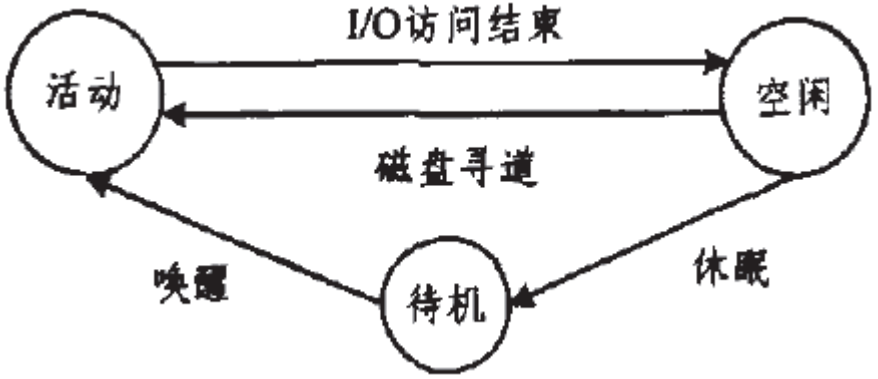
- 容量：存储的信息总量
  - 以磁盘为例
    - 磁盘总容量C = n×k×s
      - n：盘面数
      - k：每面磁道数
      - s：每道记录代码数
    - 非格式化容量：磁表面可以利用的磁化单元总数。
    - 格式化容量：按某种特定的记录格式所能存储信息的总量，约为非格式化容量的60%~70%





典型磁盘参数

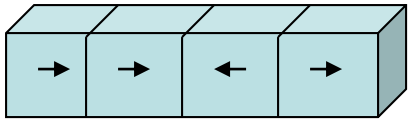
磁盘模型		企业级硬盘	笔记本硬盘	微硬盘
		IBM Ultrastar 36Z15	Toshiba MK5002MPL	IBM DSCM-11000
磁盘物理参数	容量	18.4GB	5GB	1GB
	转速	15000RPM	4000RPM	3600RPM
	平均寻道延迟	3.4ms	15ms	1ms
	平均旋转延迟	2ms	26ms	20ms
	持续传输率	55MB/s	66.7MB/s	4.3MB/s
	功率(活动/空闲/待机)	13.5W/10.2W/2.5W	1.2W/0.9W/0.2W	0.6W/0.5W/0.06W
磁盘能耗参数	休眠/唤醒能量	13J/135J	N/A	N/A
	休眠/唤醒耗时	1.5s/10.9s	15s	2s

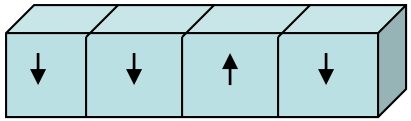


- Disks in mobile computers such as laptops consume 24 to 54 % of total system power, or more.

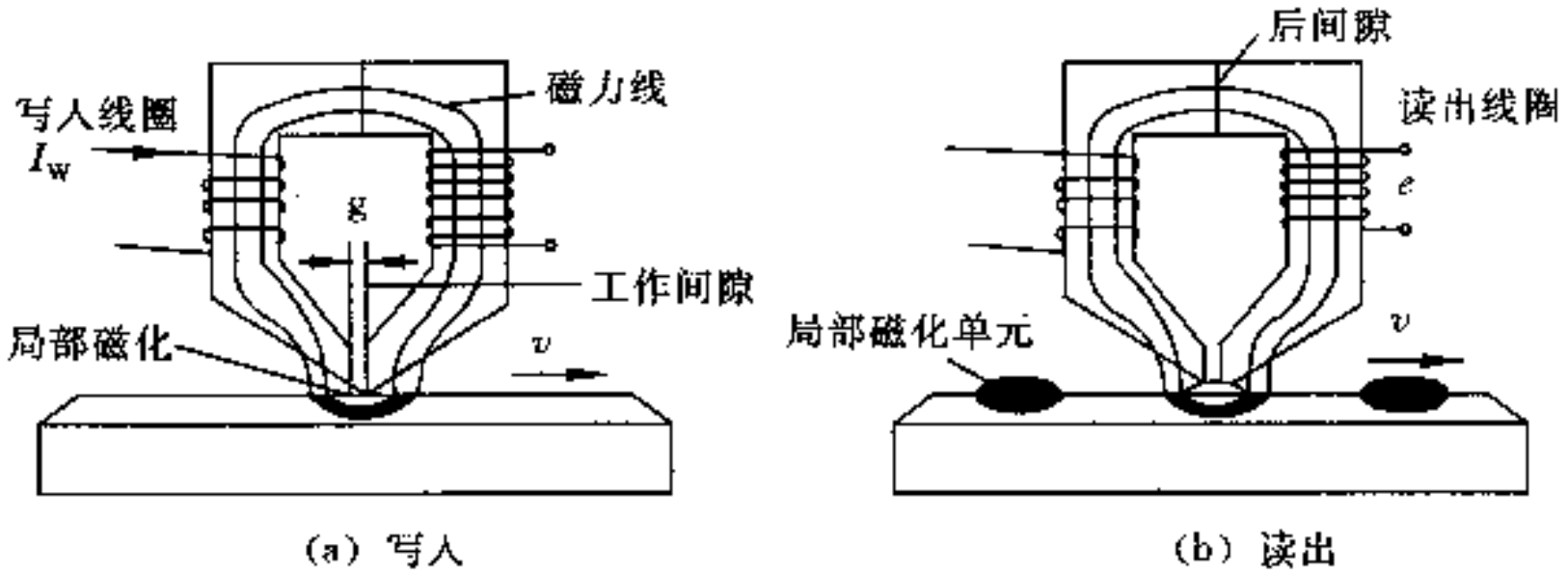
磁记录原理

- 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

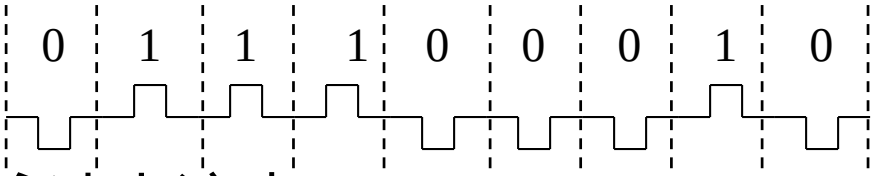


- **写入：** 记录介质在磁头下匀速通过,磁头线圈中通入一定方向和大小的电流,则会在介质上形成一个磁化单元. 电流方向不同,则磁化方向也不同.一个磁化方向规定为“0”,另一个磁化方向就规定为“1”.
- **读出：** 记录介质在磁头下匀速通过时,读出线圈会感应出电压,磁化方向不同,则感应电压就不同,对感应电压进行放大和整型,就可以读出“0”或“1”.

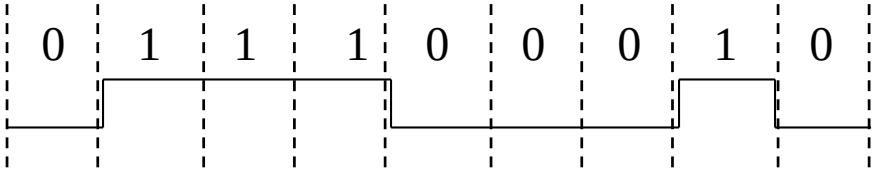
磁表面记录原理—记录（编码）方式

- 磁记录方式
  - 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）

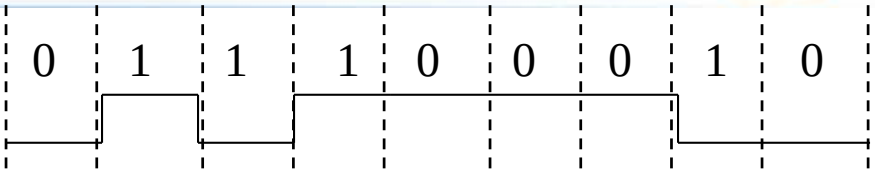
磁表面存储器的磁记录原理



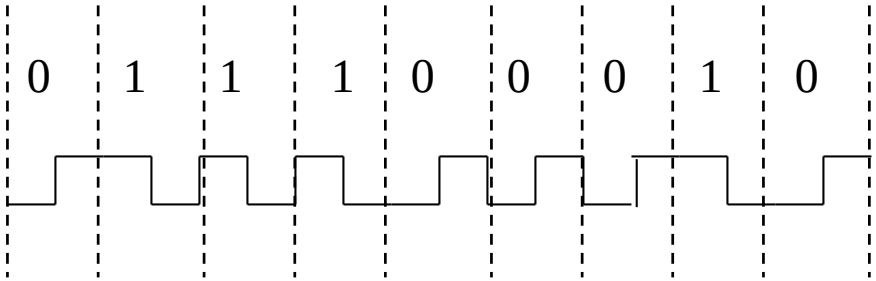
- （1） 归零制（RZ）
  - 正脉冲电流表示“1”，负脉冲电流表示“0”；
  - 不论记录“0”或“1”，在记录下一个信息前，记录电流恢复到零电流。
  - 简单易行，记录密度低
  - 改写磁层上的记录比较困难，一般是先去磁后写入。
  - 有自同步能力（能从磁头读出信号中分离获得同步信号）
- （2）不归零制（NRZ）
  - 磁头线圈始终有电流，电流方向“见变就翻”
  - 对连续记录的“1”和“0”，写电流的方向是不改变的。
  - 无自同步能力。



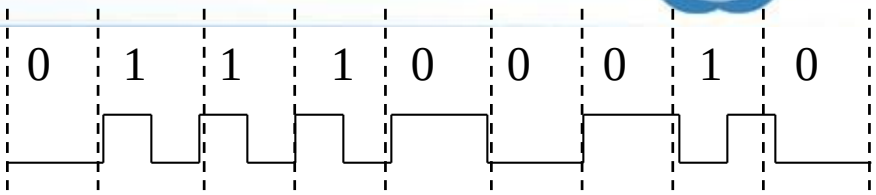
磁表面存储器的磁记录原理



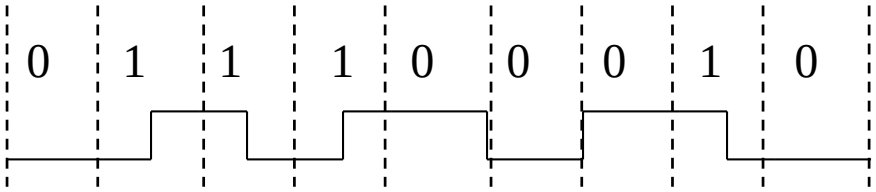
- （3）见“1”就翻的不归零制（NRZ1）
  - 磁头线圈始终有电流通过。
  - 在记录“1”时，电流改变方向，写“0”电流保持不变。
  - 不具备自同步能力，需要引用外同步信号
- （4）调相制（PM）：又称为相位编码（PE）
  - 记录数据“0”时，规定磁化翻转的方向由负变为正，记录数据“1”时从正变为负
  - “0”，“1”的读出信号相位不同，抗干扰能力强
    - 磁带多用此方式
  - 具有自同步能力



磁表面存储器的磁记录原理

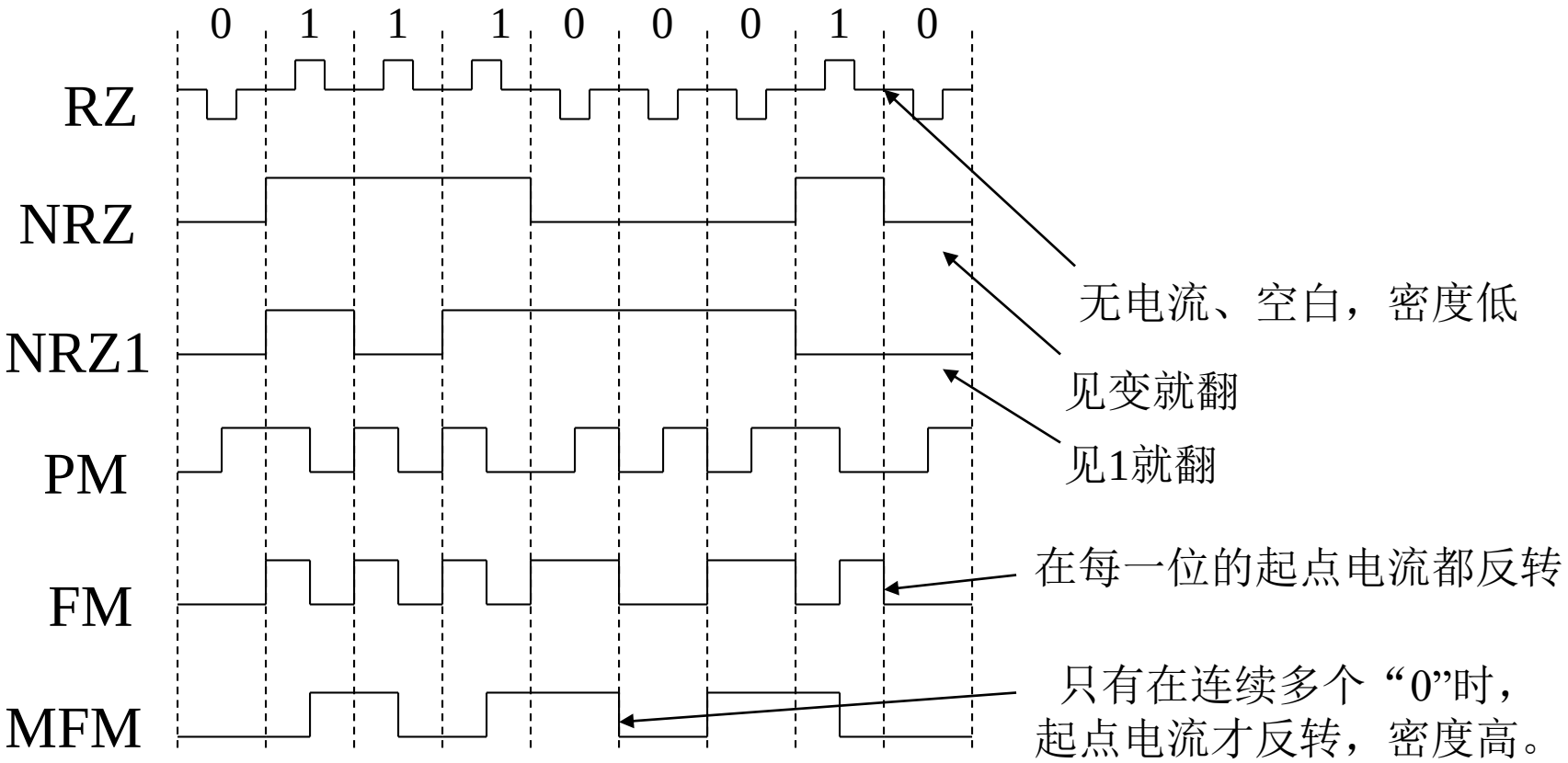


- （5）调频制（FM）
  - 频率变化（“1”的频率是“0”的两倍）
  - 在位与位之间的边界处都要翻转一次
  - 具有自同步能力。
  - 用于软硬磁盘
- （6）改进调频制（MFM）
  - 不是在每个位周期的起始处都翻转。只有当连续两个或两个以上“0”时，才在位周期的起始位置翻转一次。
  - 具有自同步能力



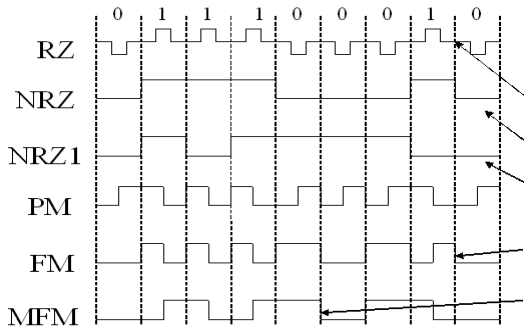
磁记录方式 - 编码方式小结

- 写电流波形的形式

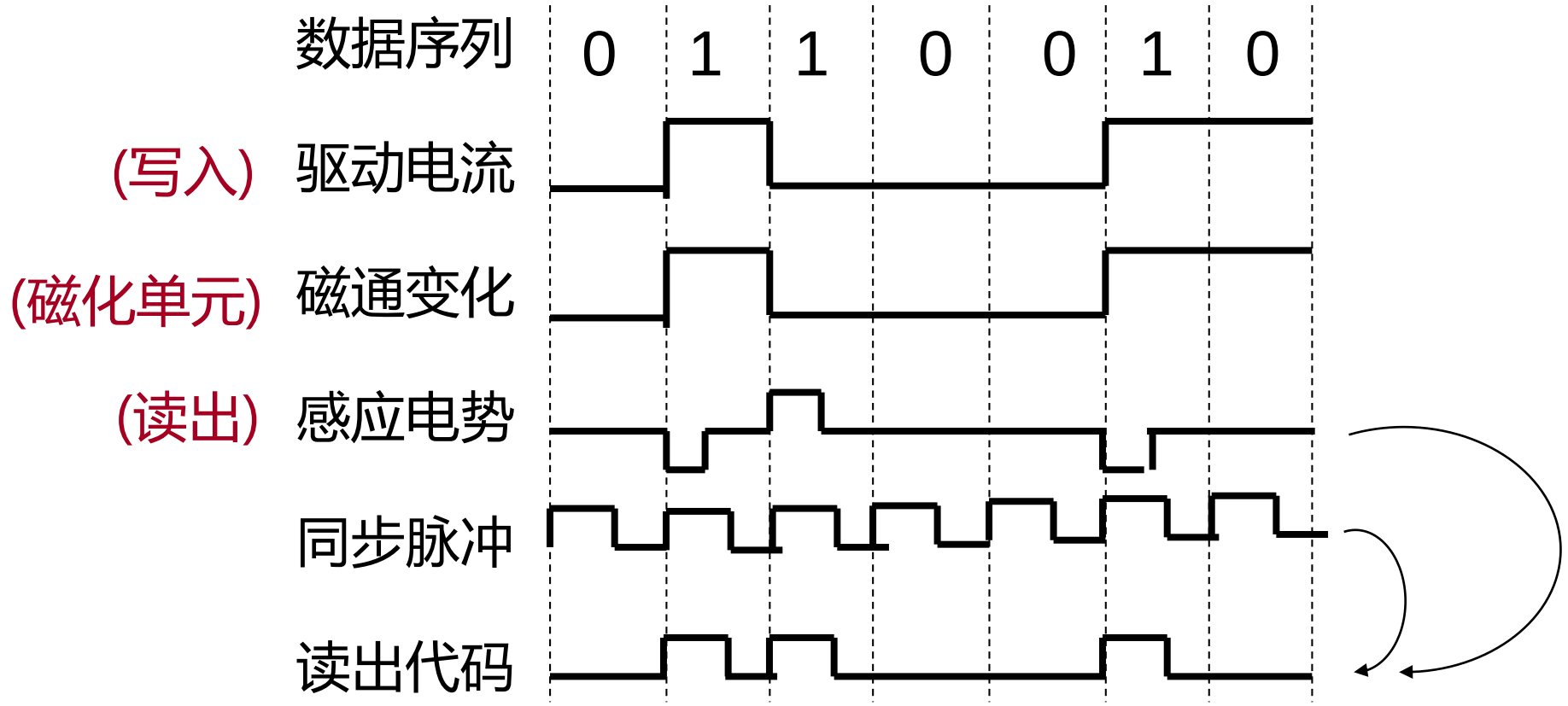


评价记录方式的主要指标

- 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

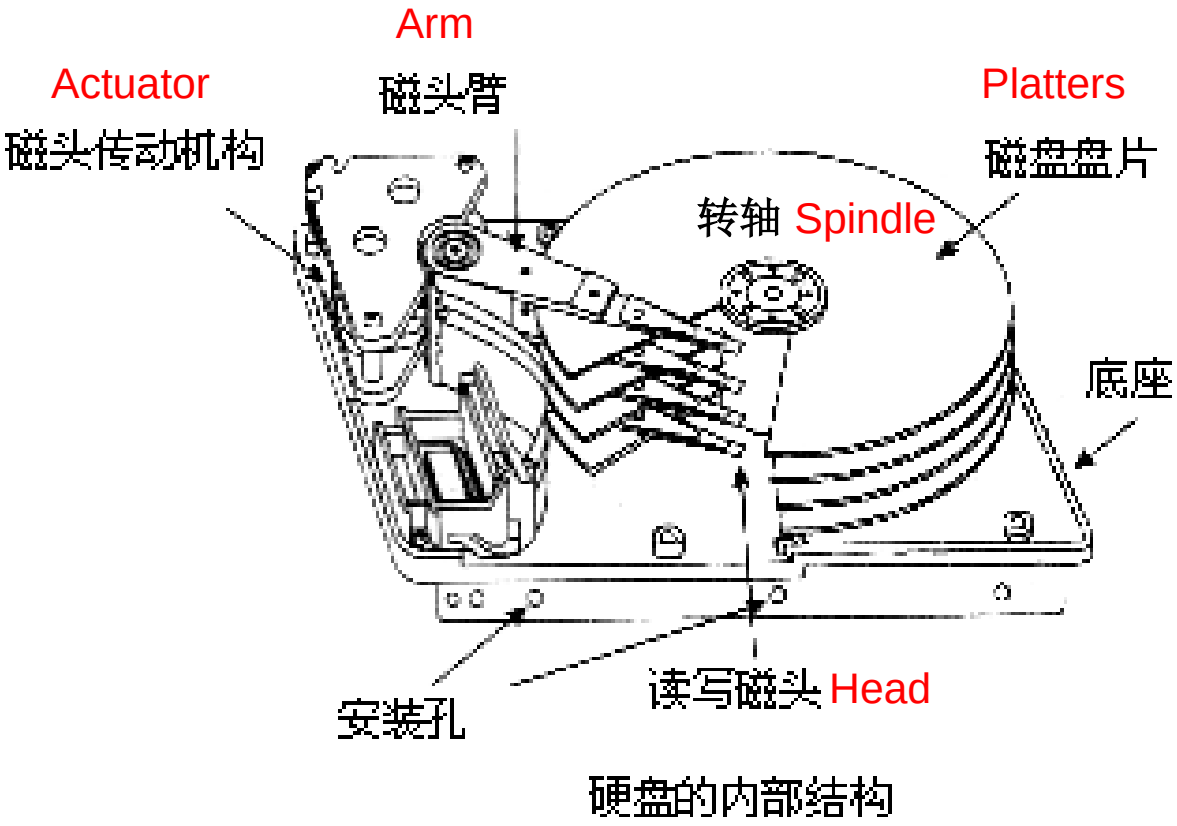
硬盘的发展和几个指标

- 1956年，IBM公司研制第一个商品化硬磁盘IBM350
  - –约有两个冰箱大
- 1973年，IBM发明了温彻斯特(温氏)磁盘，简称温盘
  - **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：** 3TB（盘片数3，单碟容量1T，磁头数6）
  - 缓存：64MB
  - **转速：** 7200rpm
  - 接口类型：SATA3.0，接口速率6Gb/秒
  - 平均寻道时间：读<8.5ms，写<9.5ms
  - 功率：运行8W，空闲5.4W，待机0.75W

IBM350磁盘存储单元

- IBM公司1956年推出的首台硬磁盘存储器
  - 用于RAMAC计算机：人口普查
    - Random Access Method of Accounting and Control
    - “使关系型数据库成为可能”
  - 50个直径为24英寸的盘片组成
  - 每分钟1200转
    - 比磁带机快200倍
    - 在几秒内找到数据
  - 容量5MB
  - 约有两个冰箱大
  - $100w

温盘（温彻斯特）

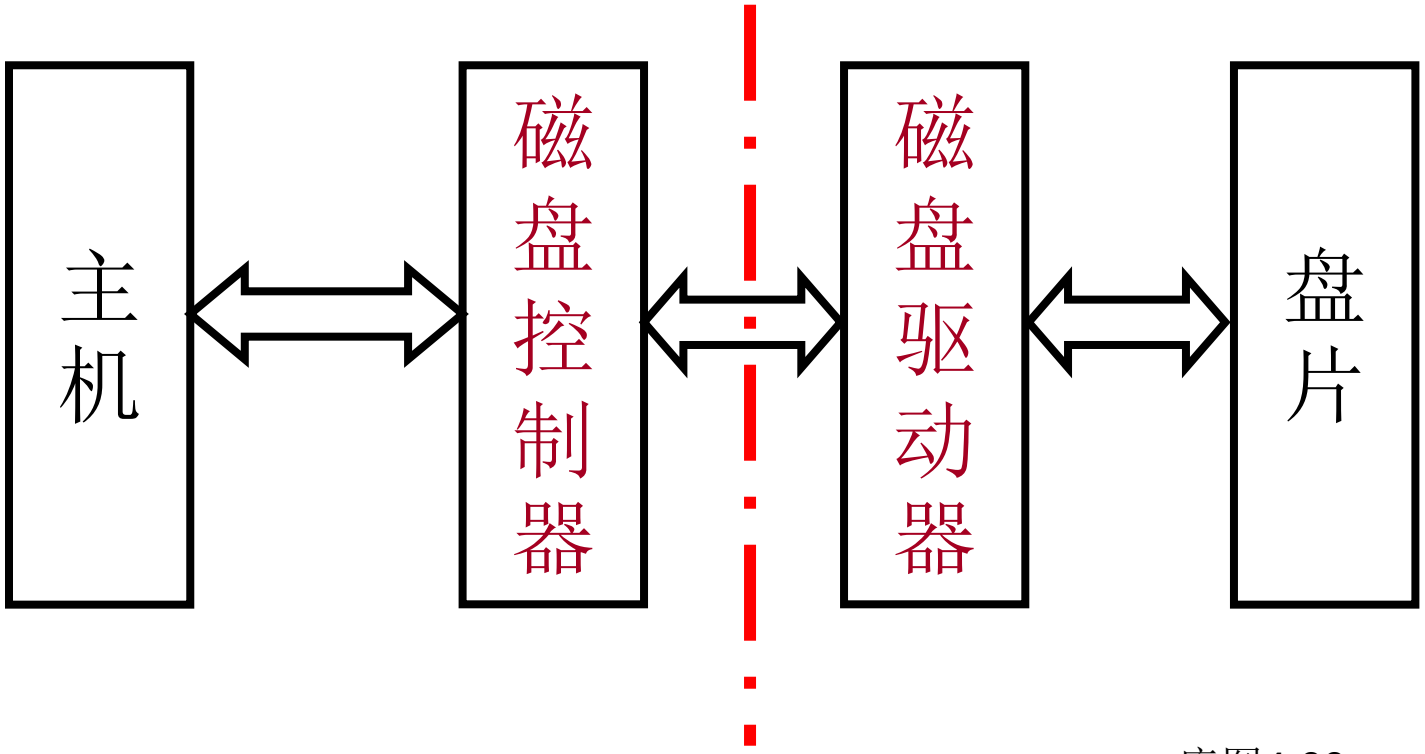


硬盘的内部结构

- **温彻斯特技术：IBM，1973**
  - **在高速旋转中，磁头与磁盘表面形成一层气泡**
  - **读写时间：< 100us**

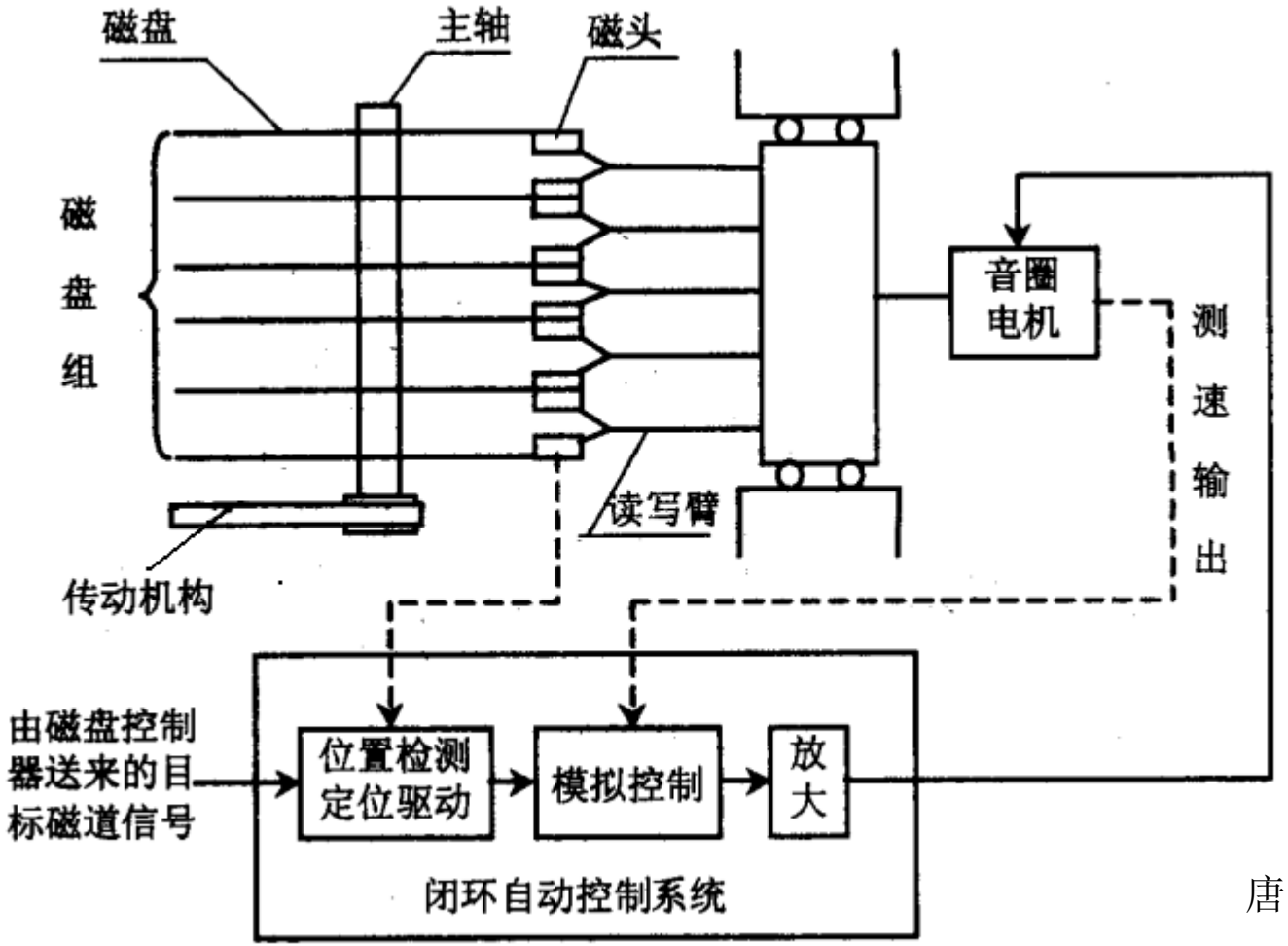
硬磁盘存储器的结构

- 硬磁盘存储器由磁盘驱动器、磁盘控制器和盘片组成。



唐图4.66

磁盘驱动器的结构及定位驱动系统

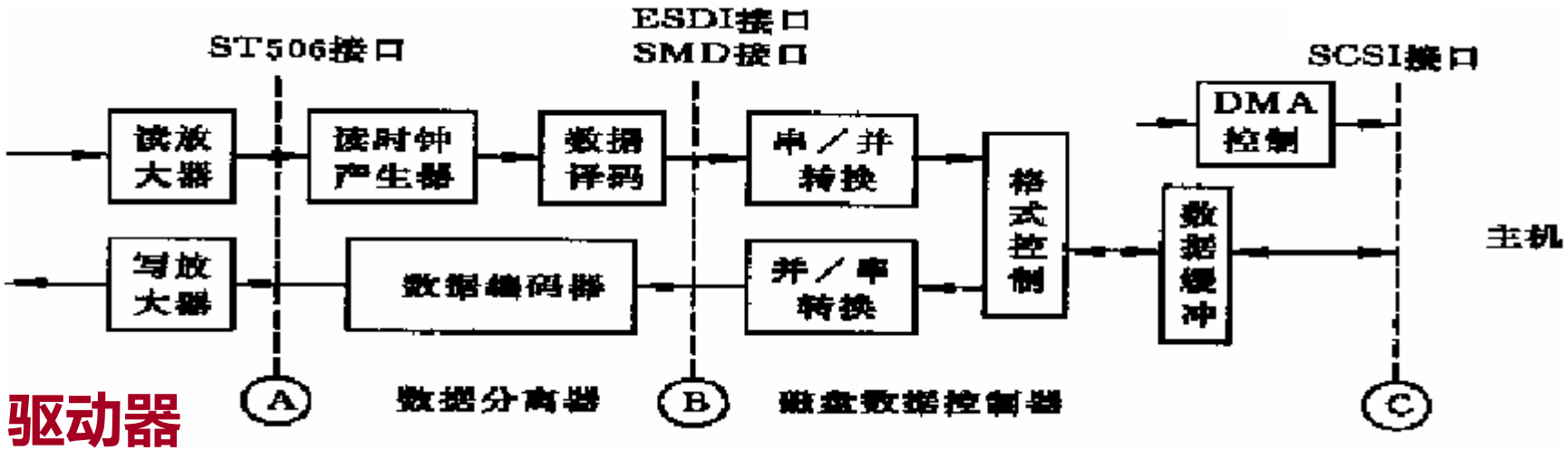


唐图4.67

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.

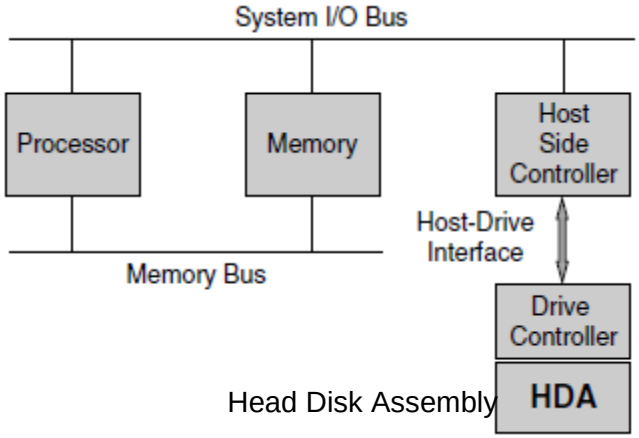
磁盘控制器

- 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（host controler）
    - 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（drive controler）
    - 可以放在多个不同的位置。

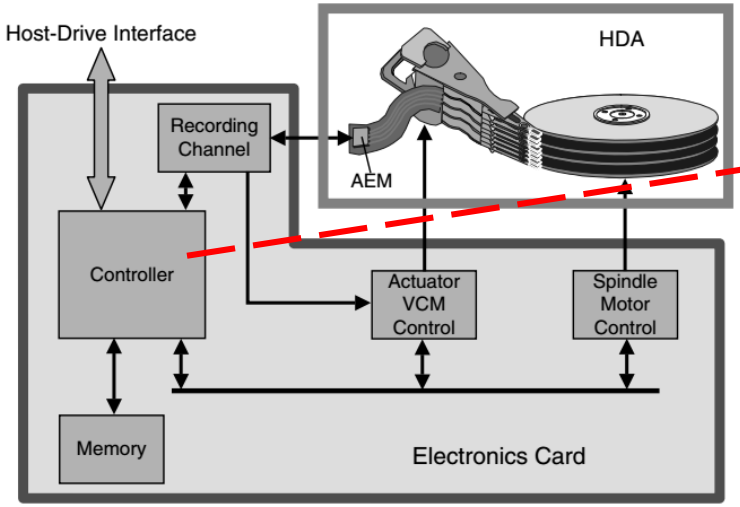


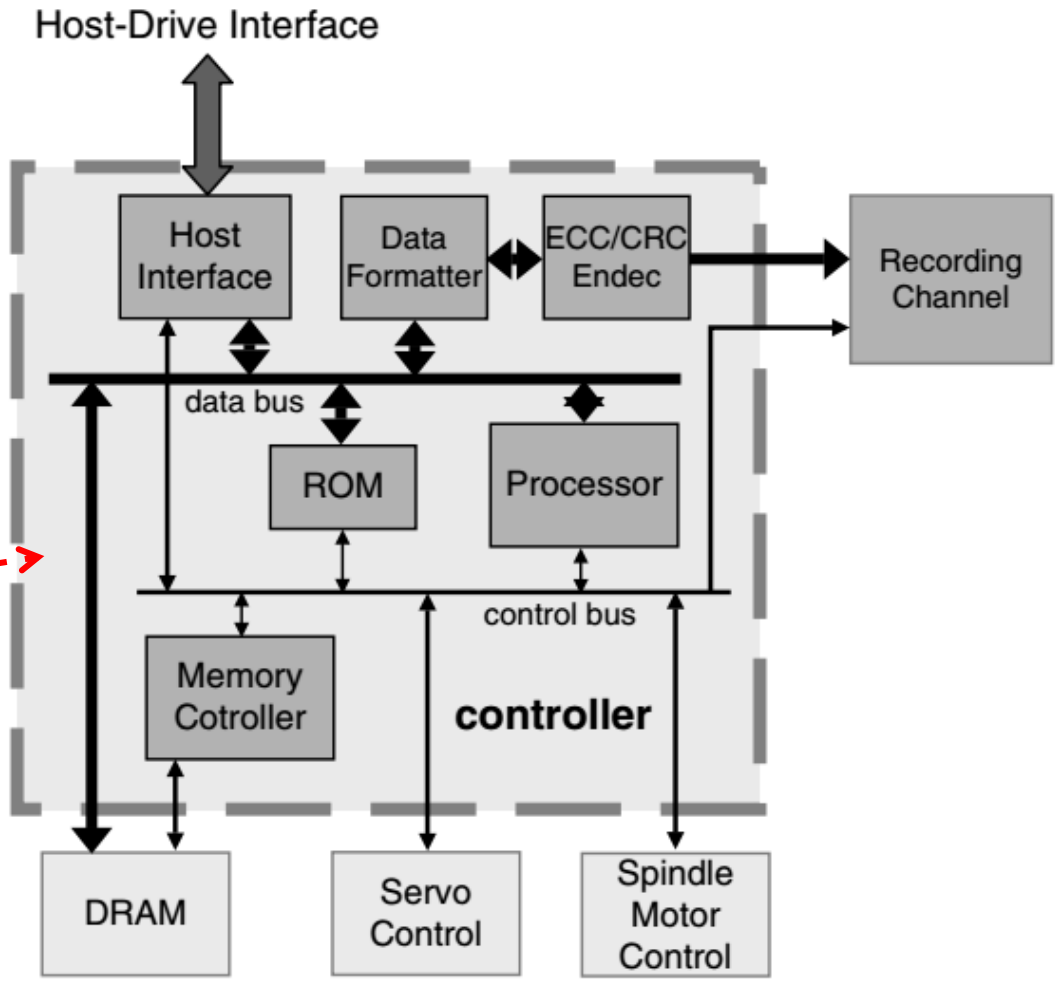
唐图4.68

Block diagram of the controller



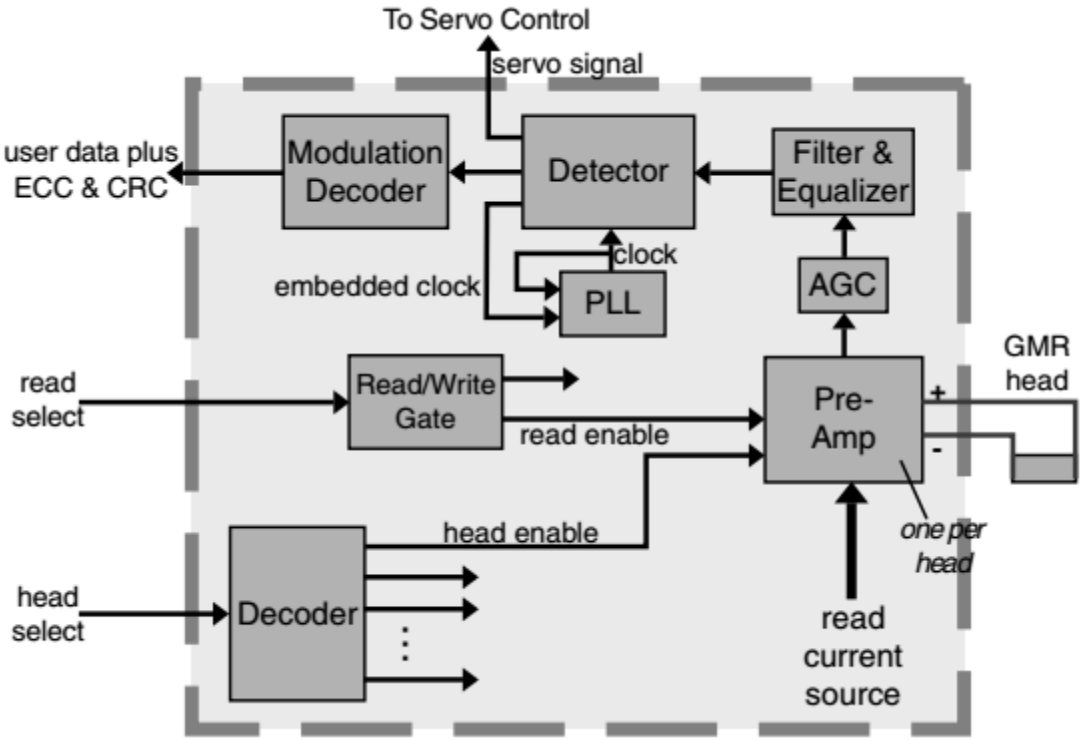
《Memory Systems: Cache, DRAM, Disk》
FIGURE 17.30



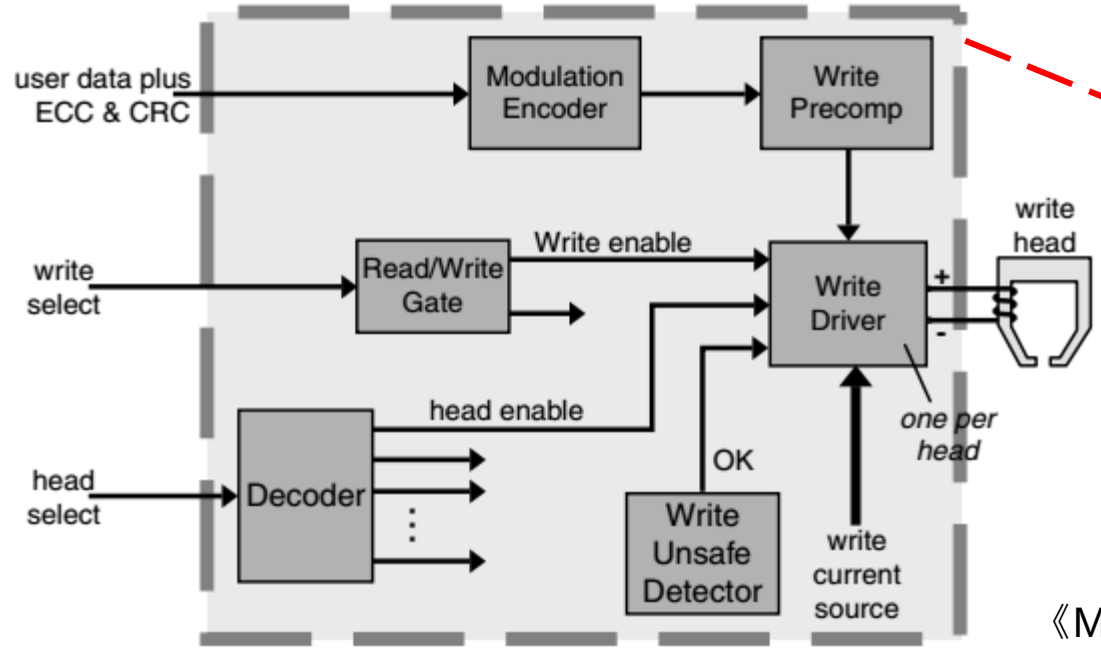


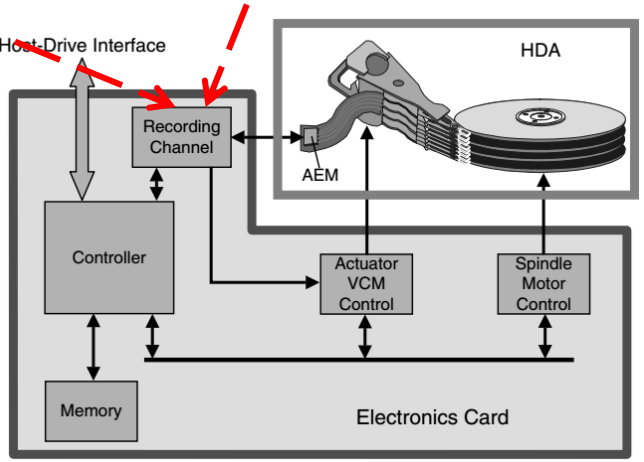
Recording Channel (驱动器)

Block diagram of a read channel.



Block diagram of a write channel.



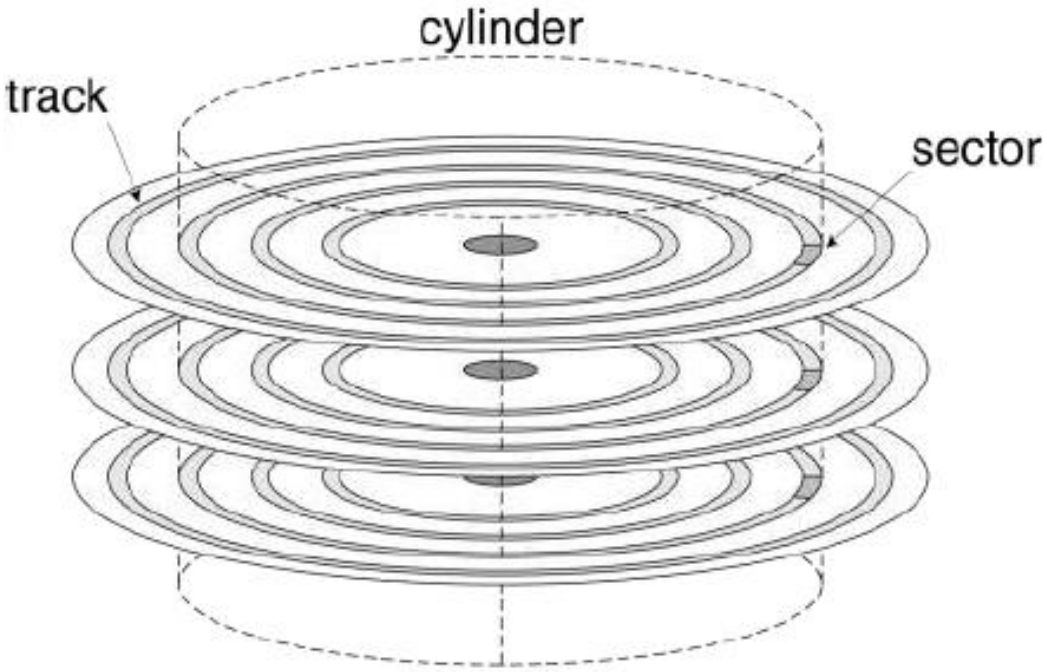


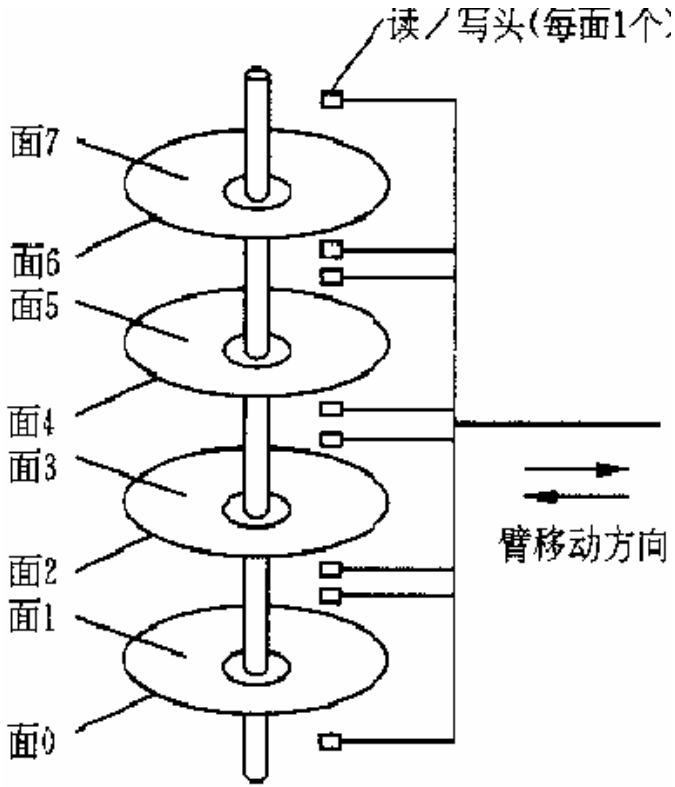
《Memory Systems: Cache, DRAM, Disk》

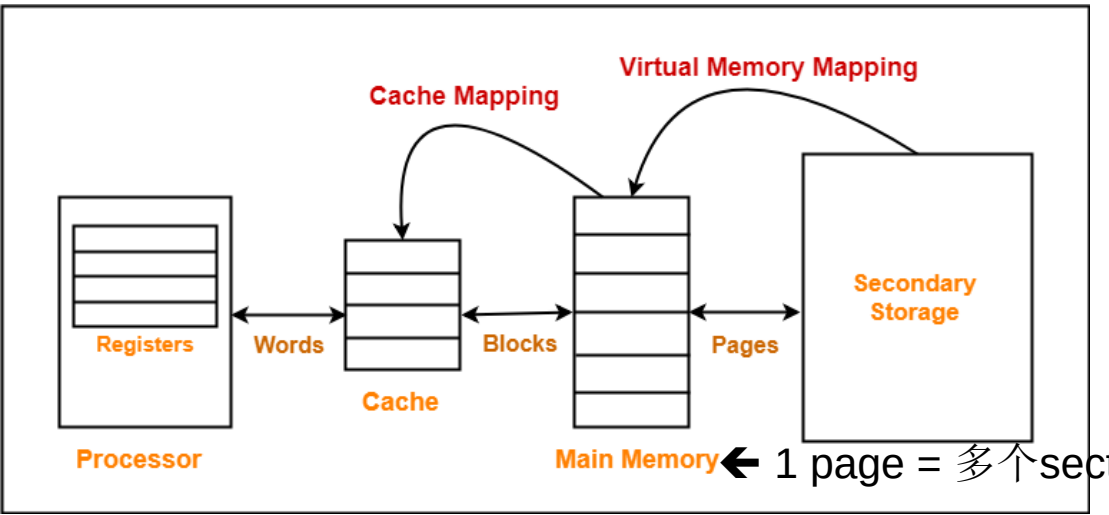
常见接口类型

- 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 (AT Attachment)）接口，“串口硬盘”
    - IDE升级：采用串行连接方式，具备更强的纠错能力，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
  - USB
- 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：一次访问一个扇段sector





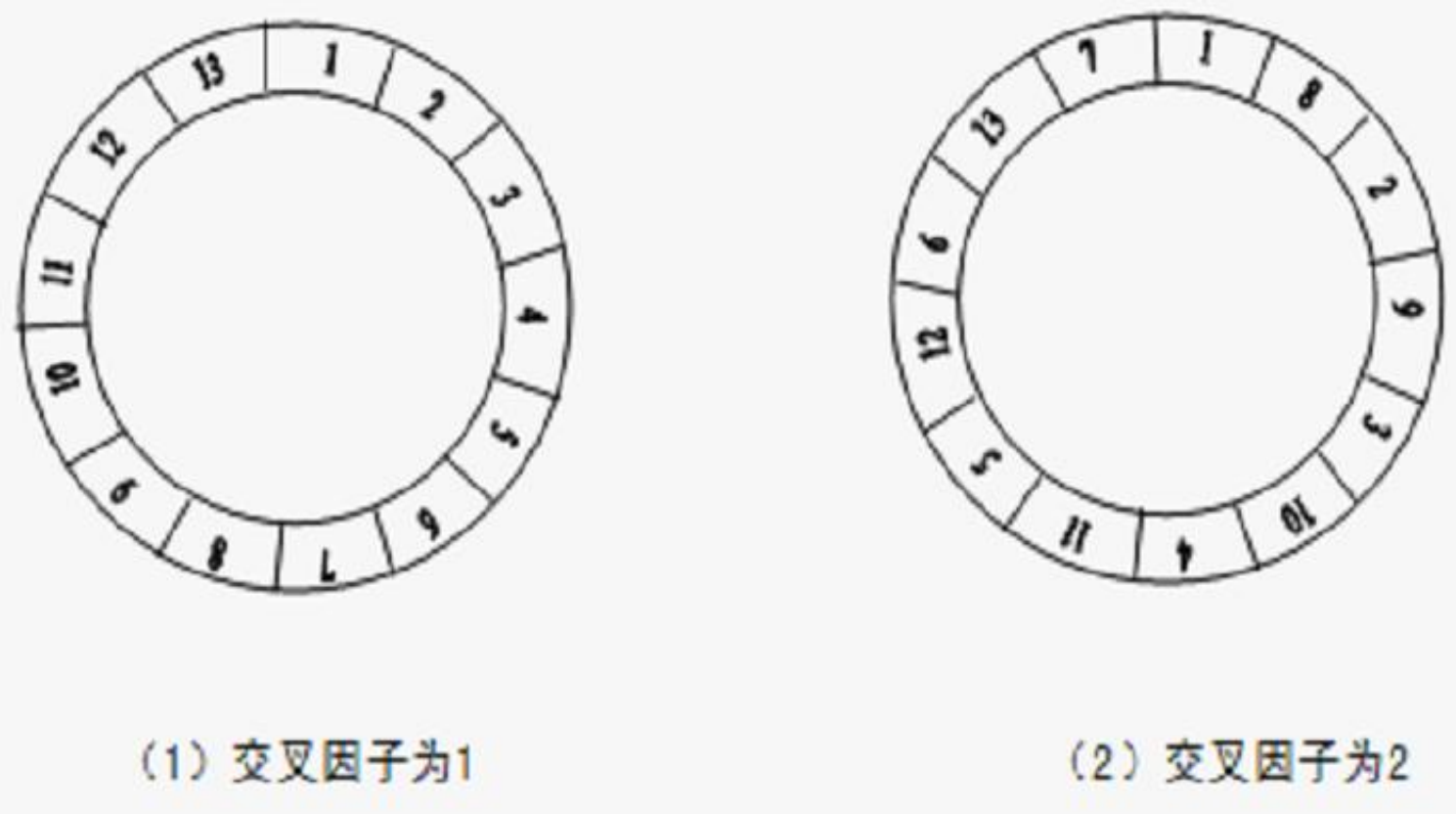


磁盘数据地址：

台号	磁道号	盘面号	扇段号
台号	柱面号	磁头号	扇段号

← 1 page = 多个sector？

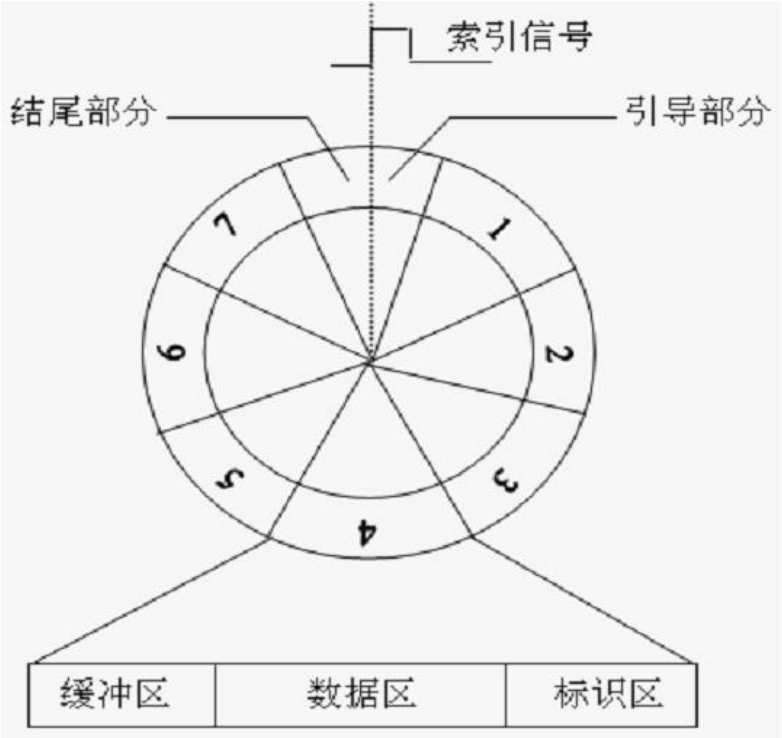
扇区的排列：交叉因子（Interleave）

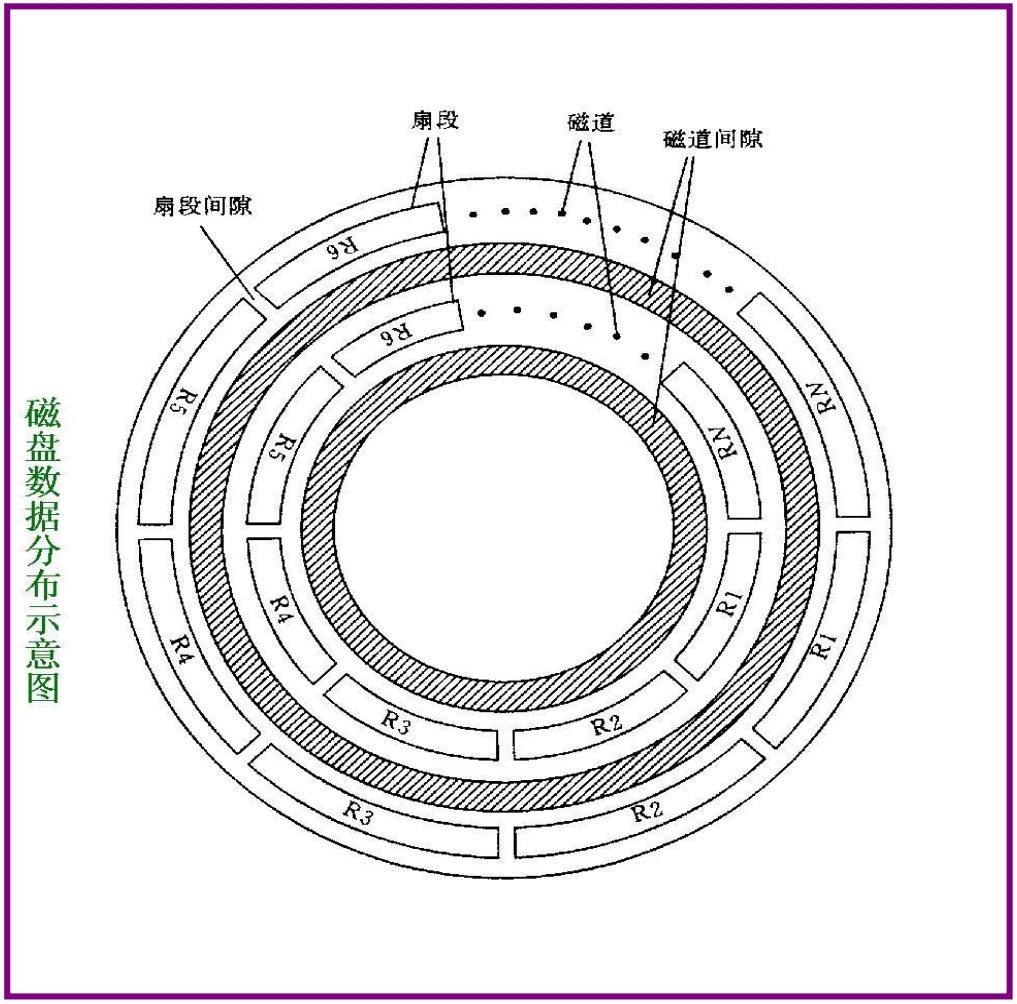


- 扇区交叉排列技术
  - 磁头读写反应速度低于盘片的旋转速度

硬磁盘的磁道记录格式

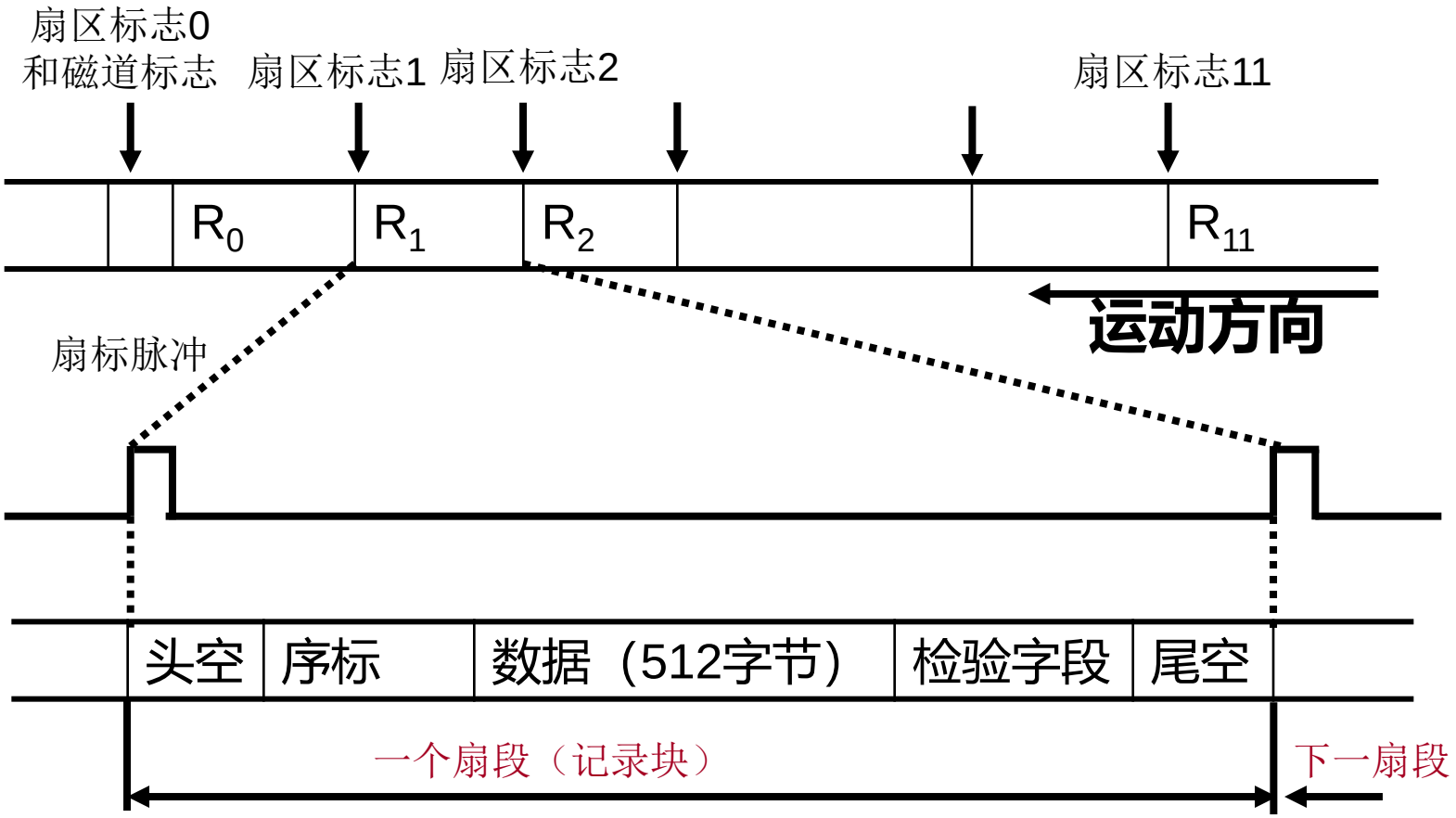
- 扇区sector的大小?
  - 定长记录格式
  - 不定长记录格式



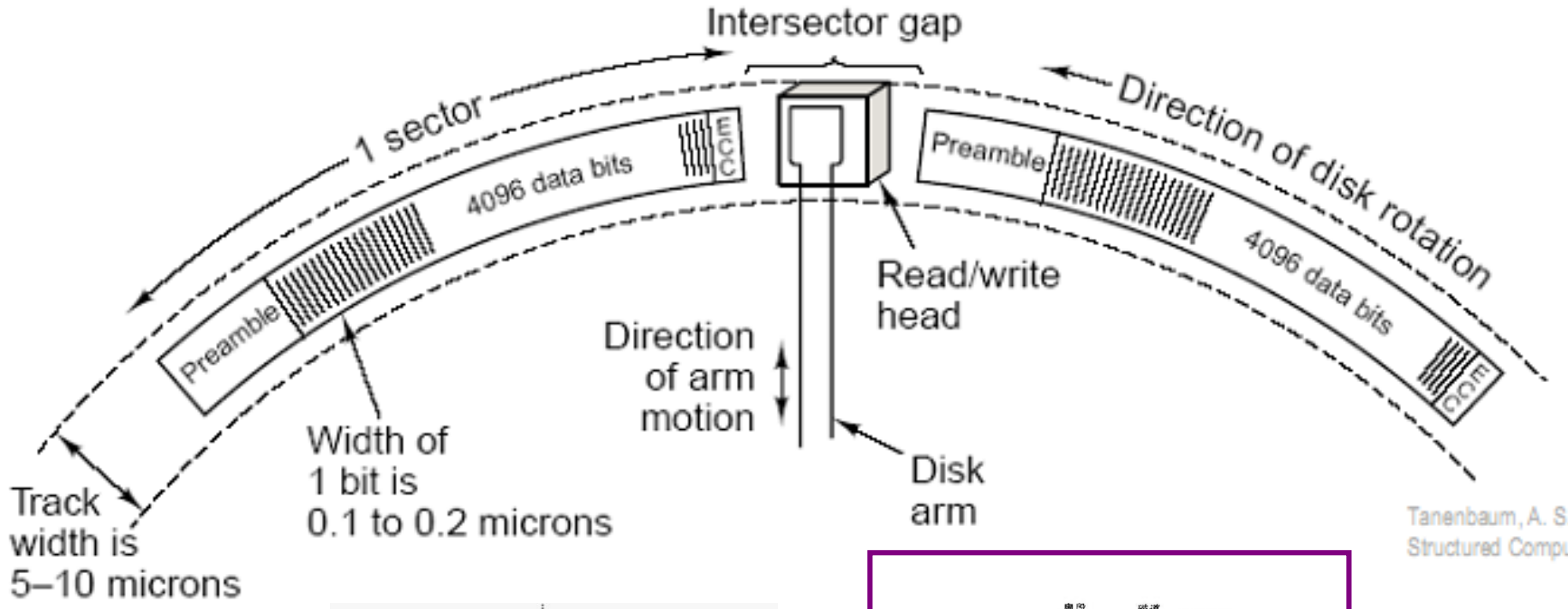


定长记录格式—ISO T型磁道记录格式

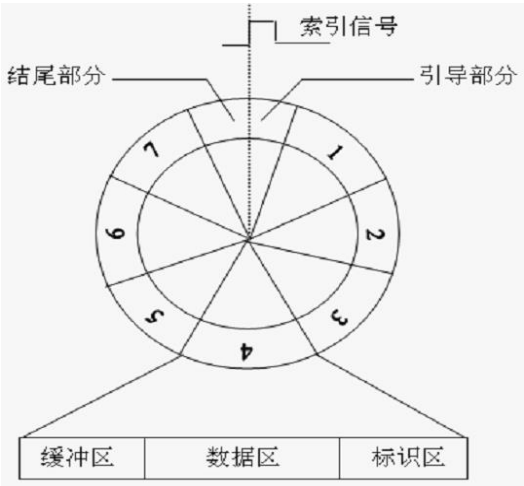
- 结构简单，可按柱面号、盘面号、扇段号进行直接寻址，但是记录区的利用率不高。

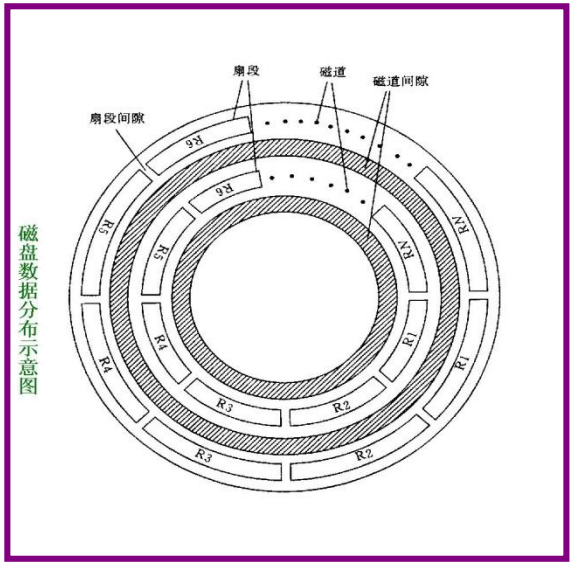


例：

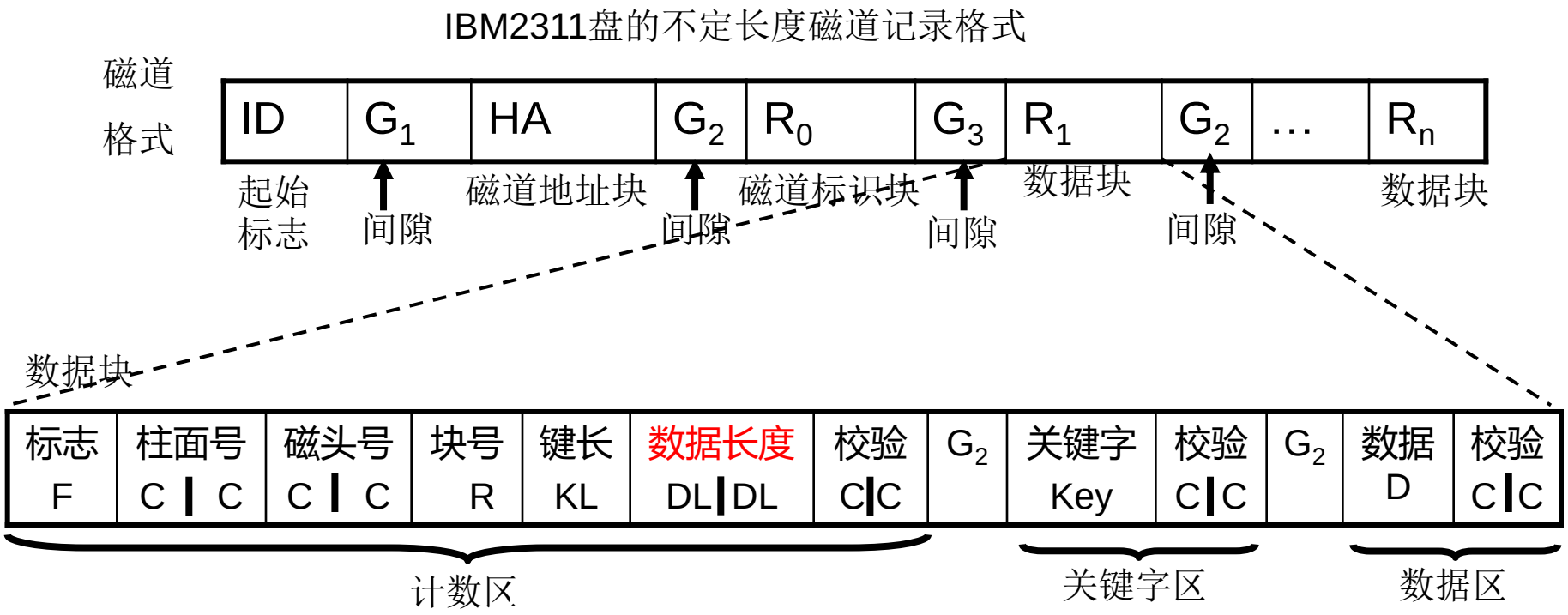


Tanenbaum, A. S. (2006)
Structured Computer Orga





不定长记录格式

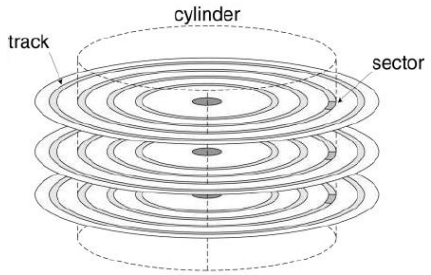


- 定长记录格式：若文件长度不是定长记录的整数倍时，往往造成记录块的浪费
- 不定长记录格式：根据需要来决定记录块的长度，如IBM2311、2314等磁盘驱动器。

HDD编址方式

台号	柱面号	磁头号	扇段号

- C.H.S（Cylinder/Head/Sector）物理寻址方式
  - 数据传输的开始地址28位：写入4个8位寄存器
    - 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）
  - 硬盘容量=盘面数×柱面数×每面扇区数×扇区容量（512B）
    - 最大容量=总扇区数×扇区容量=2^28个扇区*512Bytes=137GB
- LBA(Logical Block Addressing)逻辑块寻址模式
  - 线性寻址模式（COD$5.2.4）
    - 将磁盘上的所有扇区从0开始编号直到最大扇区数减1
    - 序号相邻的数据块可能在不同磁道上!
  - 突破C.H.S模式的容量限制问题
    - 28位LBA硬盘寻址方式
    - 48位LBA硬盘寻址方式：2^48扇区*512Bytes=?



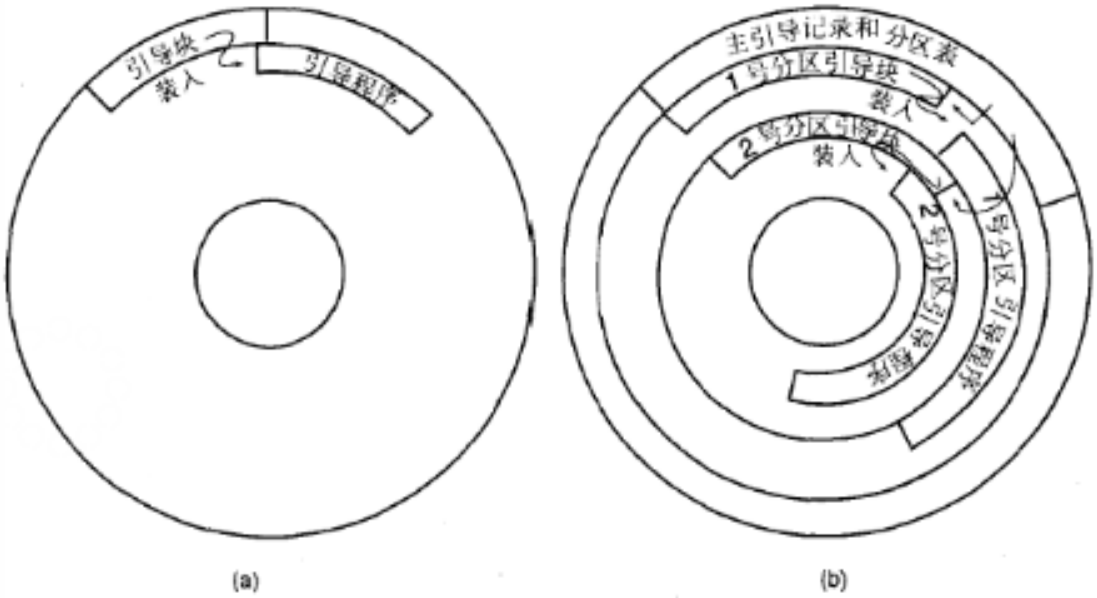
磁盘读写（BIOS int 0x13）

台号	柱面号	磁头号	扇段号

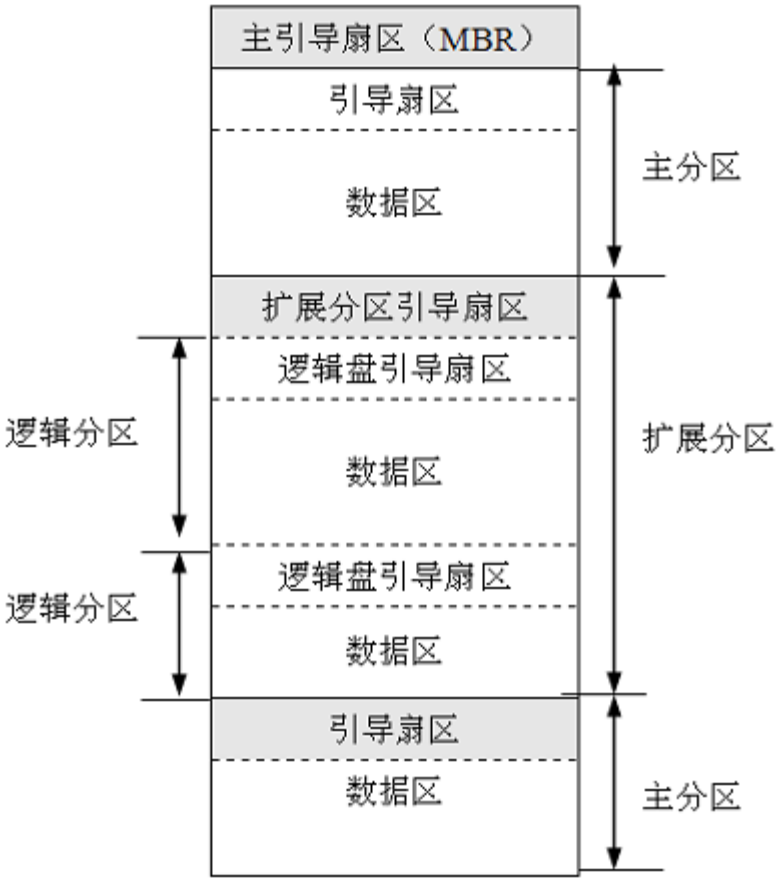
- 寻道 -> 扇区定位 -> 读写
- 输入参数
  - AH=0x02(读盘) , =0x03(写盘) ,=0x04(校验)
  - AL=扇区数（同时处理连续的扇区）
  - CH=柱面号&0xff
  - CL=扇区号（0-5位）|(柱面号&0x300）>>2;
  - DH=磁头号
  - DL=驱动器号（台号）
  - ES：BX=缓冲区地址（校验寻道不使用）
- 返回值
  - FLACS.CF==0，没有错误，AH==0
  - FLAGS.CF==1，有错误，错误号存在AH内

逻辑盘与磁盘分区

引导盘，非引导盘



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

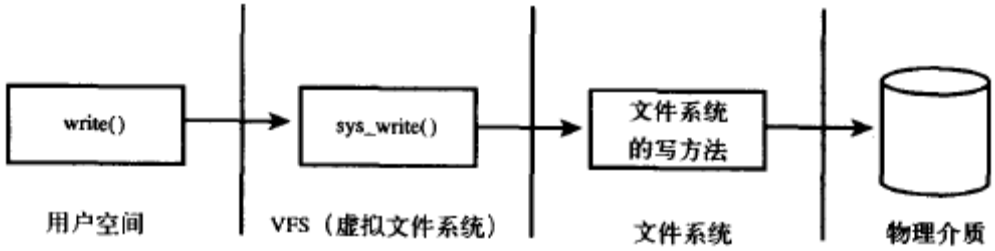
操作系统　　文件、流数据

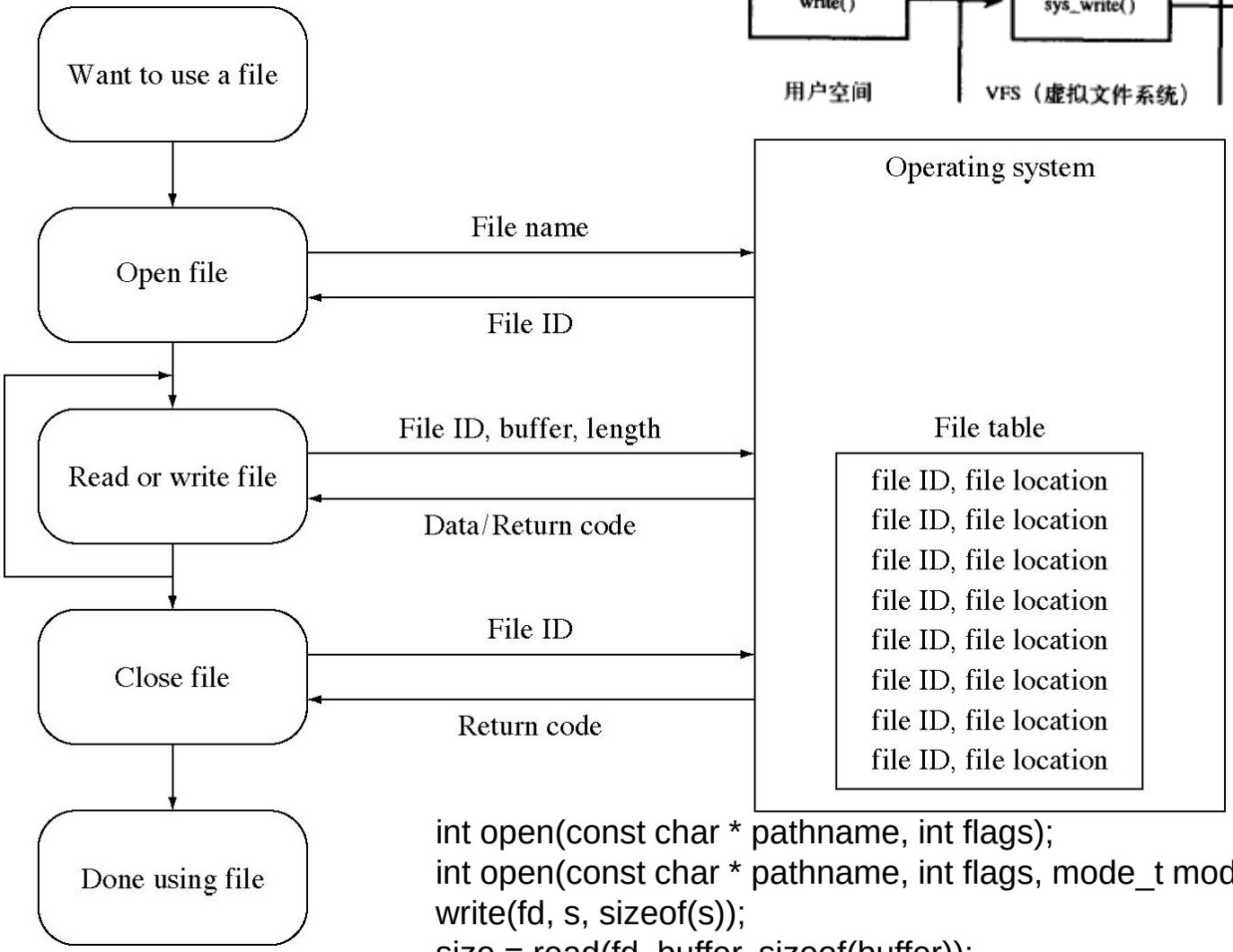
格式化记录　　目录区、索引区、数据区（数据块）

物理层　　磁道、扇区、位流

- 格式化记录：文件系统
  - FAT：File Allocation Table
    - 描述文件系统中存储单元的分配状态及文件内容的前后链接关系
      - 保留扇区：主引导扇区
      - FAT区：簇链表
      - 根目录区：文件和目录信息的目录表
      - 数据区：实际的文件和目录数据存储的区域
  - NTFS
  - EXT3、EXT4：一种日志式文件系统

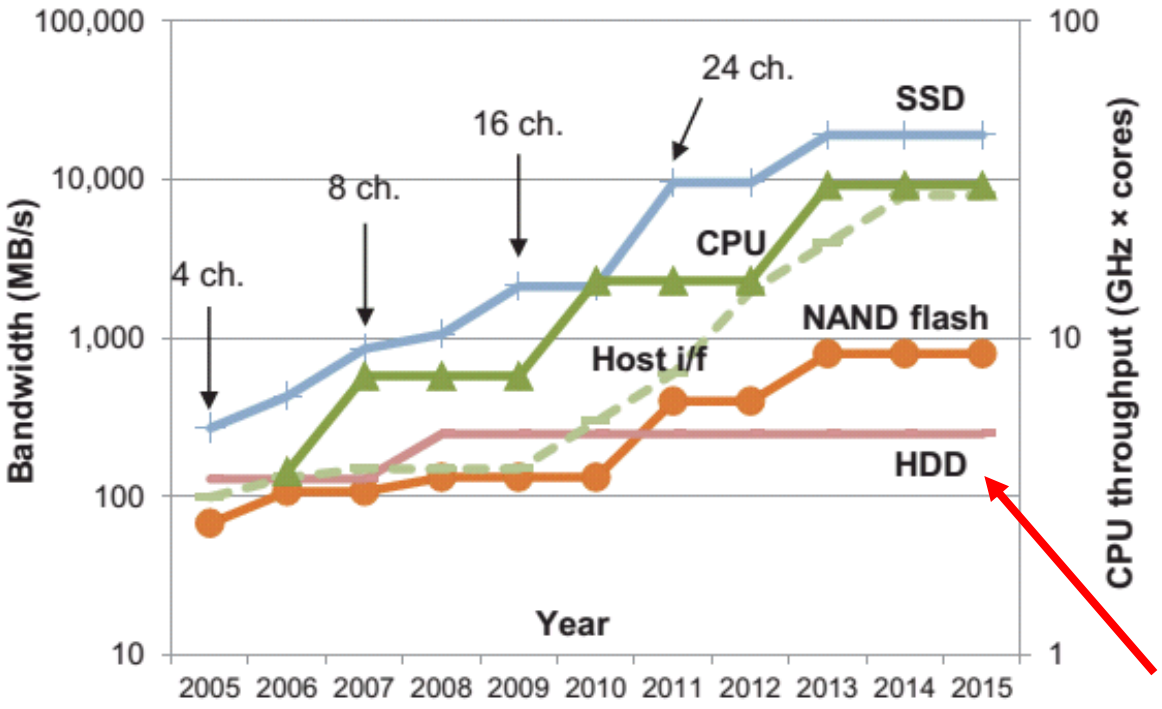
Steps in using a file：syscall

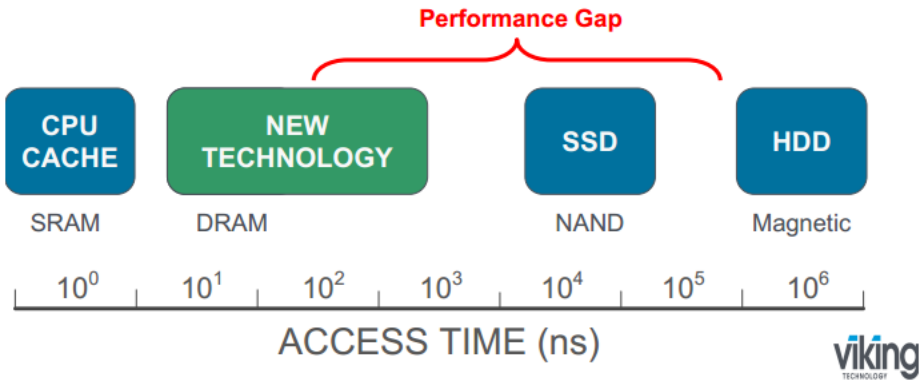





```
int open(const char * pathname, int flags);
int open(const char * pathname, int flags, mode_t mode);
write(fd, s, sizeof(s));
size = read(fd, buffer, sizeof(buffer));
```

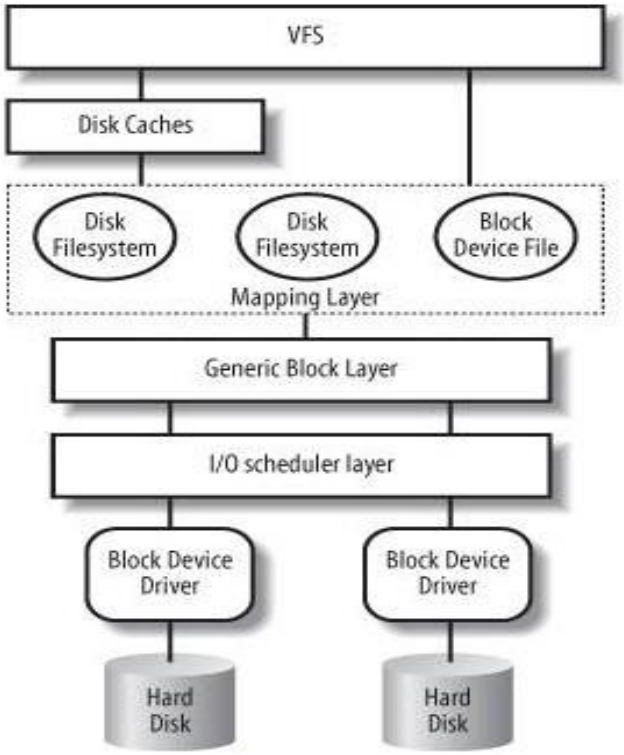
外存的性能


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

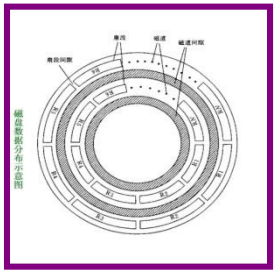
硬磁盘存储器的发展：性能、*可靠性*

- *提高HDD传输率和缩短存取时间*
  - *提高磁盘记录密度*
  - *提高主轴转速*
- *半导体盘SSD*
  - *在功能上模拟硬盘*
  - *以Flash Memory为核心，加上接口电路和其他控制电路组成*
- 磁盘Cache：驻留在内存中的盘块
  - 周期性地写回磁盘：UNIX时间间隔30s
- **磁盘IO访问调度：合并，排序**
  - 读第一个字节比读同一secter中的后续字节慢10万倍
  - OS：I/O调度器；
  - 硬件：磁盘控制器



磁盘IO访问控制：并发，“批处理”

- 多进程共享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）
    - 读disk的第一个字节比读同一secter中的后续字节慢10万倍
  - 在处理每一次I/O请求前，执行合并与排序的预处理操作
    - 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
- 软件方法： I/O调度器，OS“批处理”
  - FCFS：公平！
  - linux电梯算法：减小平均寻道时间
    - 假设有IO请求序列（磁道ID）：100，500，101，10，56，1000
    - 按请求地址排序：100，101，500，1000，56 ，10
    - 按逻辑地址调度可能更糟！！
- 硬件方法：磁盘控制器

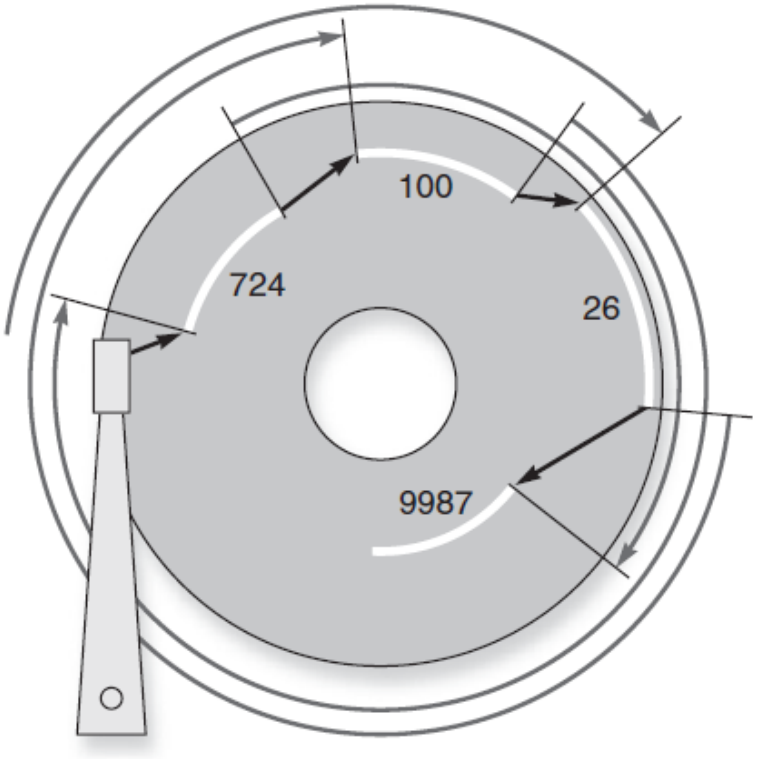


谁负责调度：OS or 磁盘控制器?

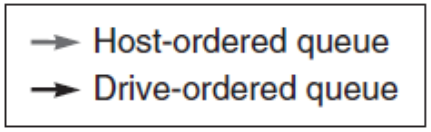
Operation	Starting LBA	Length
Read	724	8
Read	100	16
Read	9987	1
Read	26	128

OS
电梯

Operation	Starting LBA	Length
Read	26	128
Read	100	16
Read	724	8
Read	9987	1

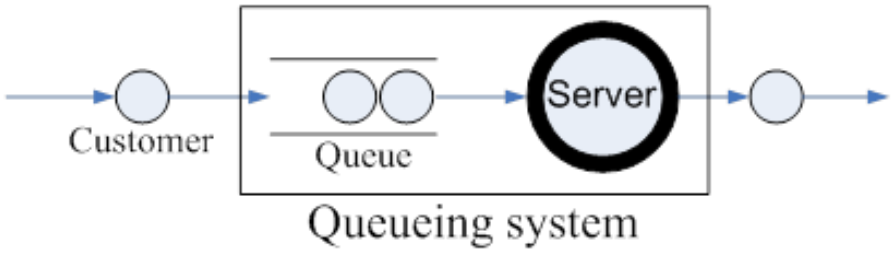


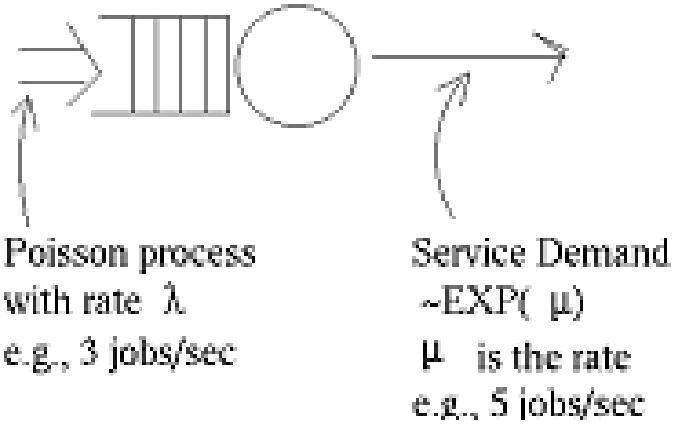
RV图5-50



- 从磁头位置开始
- OS调度：逻辑块地址，电梯，转3圈
  - Host-ordered
- 磁盘调度：物理地址，3/4圈
  - Drive-ordered：724，100，26，9987

建模：排队论





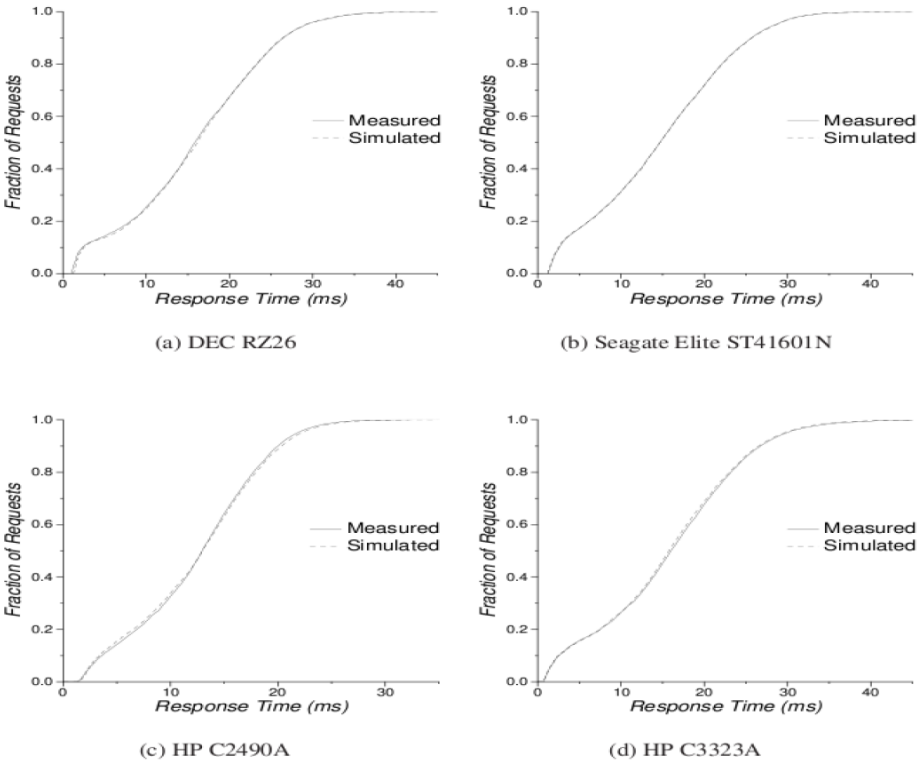
- 指标：顾客的平均等待时间，服务台的忙闲程度
- 两类经典模型：到达时间/服务时间/窗口数
  - 一个服务窗，顾客按参数为λ的泊松分布到达，到达的时间间隔为负指数分布
    - M/M/1：服务窗为每个顾客服务的时间为负指数分布M（马尔可夫），平均服务率为μ
    - M/G/1：服务窗为每个顾客服务的时间是一般分布G（随机）

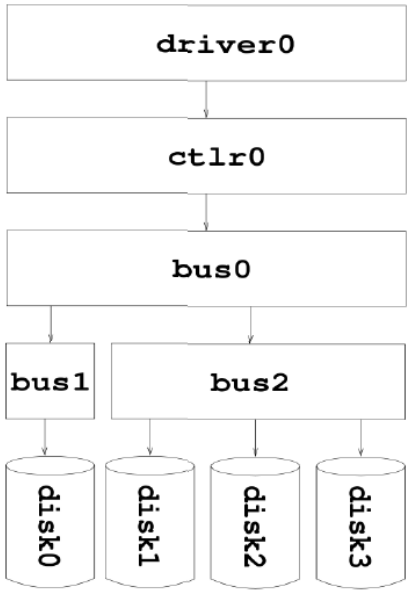
磁盘系统仿真器DiskSim

- 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





小结

- 结构，过程，指标：
  - 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
  - 访问磁盘数据的过程
  - 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，磁盘控制器
    - 磁盘高速缓存
  - 磁盘I/O响应时间?
- 作业
  - 唐4.38
  - （可选）C语言读盘程序设计？块方式，文件方式


Thank You