第12章 大容量存储器结构

目录

1、磁盘结构

磁盘组成

磁盘访问时间

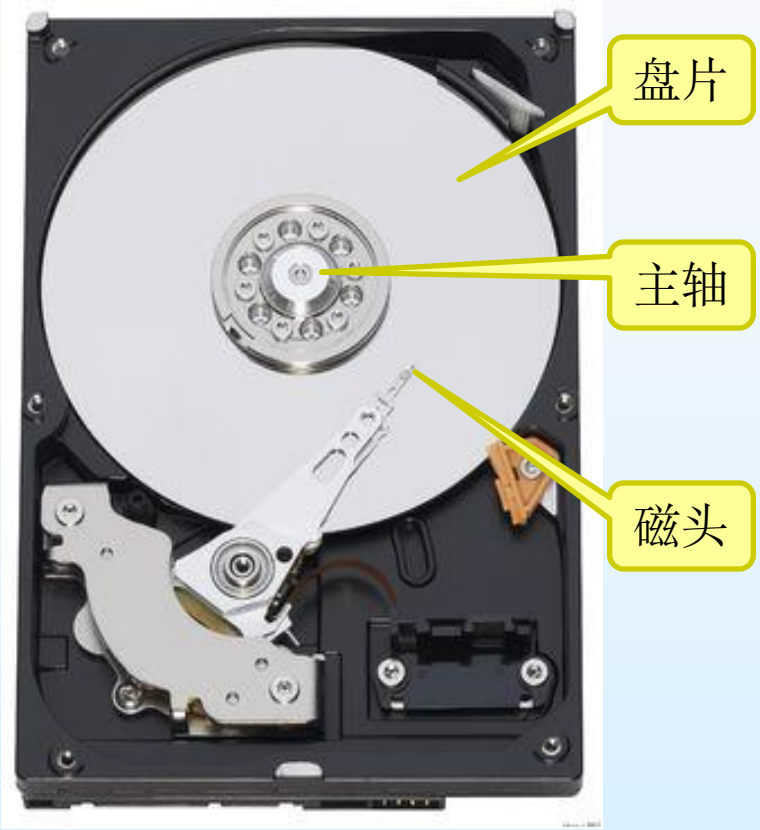
地址映射

磁盘管理

大容量存储设备

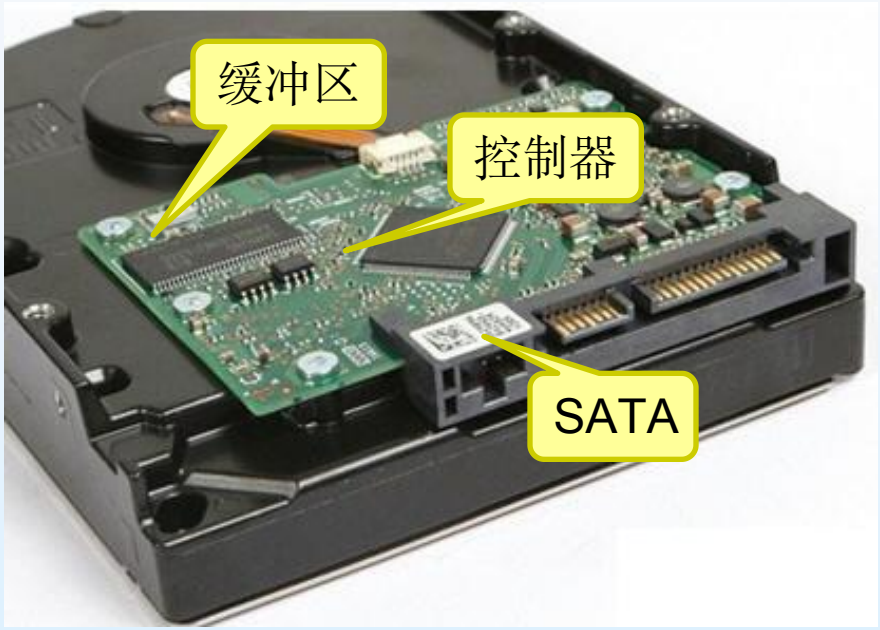
磁盘结构

- 盘片
  - 存储数据的介质
  - 正反两面可以存储数据
- 磁头
  - 读写数据，沿磁盘半径移动
  - 有多少盘面就有多少磁头
- 主轴
  - 马达驱动，使盘片旋转
  - 固定速度旋转



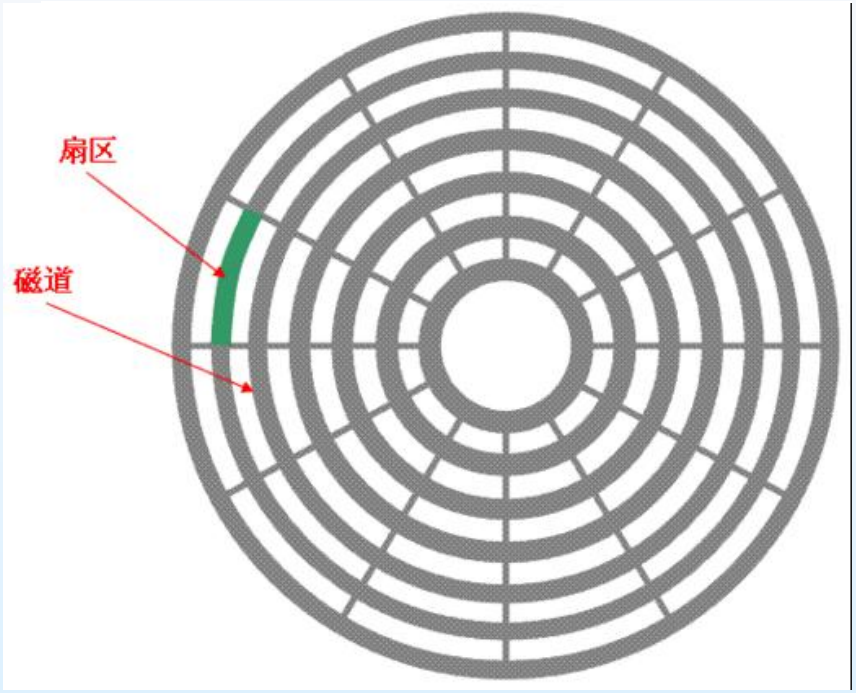
磁盘控制器和接口

- 接口
  - EIDE, ATA, SATA, USB, Fibre Channel, SCSI, SAS, Firewire
- 磁盘控制器
  - 控制磁盘的读写等操作
- 缓冲区
  - 利用磁盘缓冲区来暂存数据

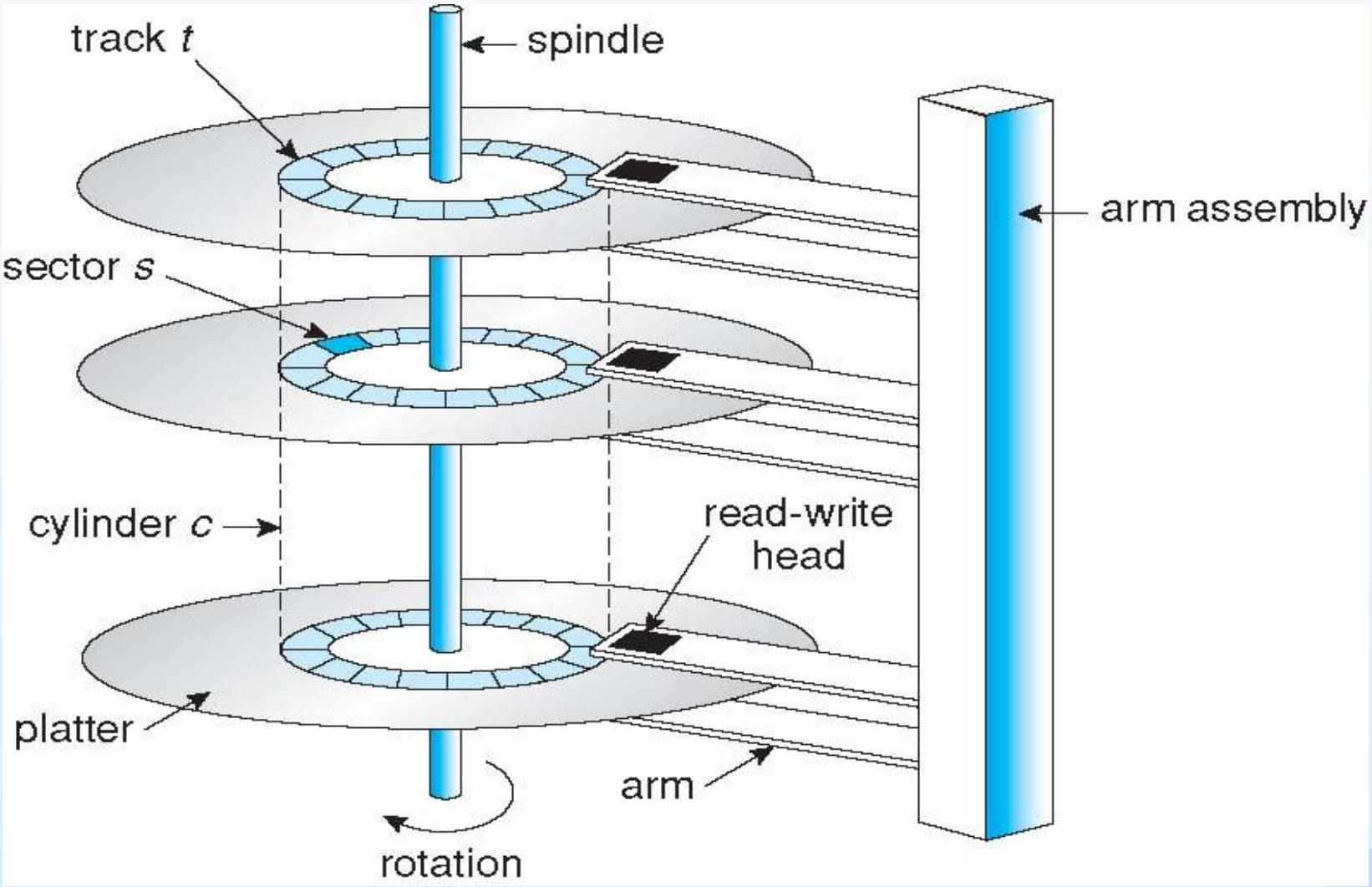


盘片结构

- 磁道
  - 磁头在盘片表面划出的圆形就是一个磁道
  - 每个盘面划分为数目相等的磁道
  - 从盘面外缘“0”开始编号
- 扇区
  - 磁道被等分为若干个弧段，称为扇区
  - 扇区大小：512字节
- 柱面
  - 每个盘面上具有相同编号的磁道形成一个圆柱，称为柱面
  - 一个盘面有几个磁道就有几个柱面



磁盘示意图


track t
spindle
arm assembly
sector s
cylinder c
read-write
head
platter
arm
rotation

地址映射关系

- 块号：LBA
- 磁盘地址(CHS)：(C,H,S)
  - Cylinder（柱面/道C）
  - Head（磁头/面H）
  - Sector（扇区S）

$$C = LBA \div (HPC \times SPT)$$

$$H = (LBA \div SPT) \bmod HPC$$

$$S = (LBA \bmod SPT) + 1$$

- SPT:每个磁道最大扇区数
- HPC:最大磁头（盘面）数

- 物理扇区号 = ((柱面号 × 磁头数) + 磁头号) × 每磁道扇区数 + 扇区偏移量 - 1
  - 柱面号从0开始编号；
  - 磁头号从0开始编号；
  - 扇区号从1开始编号；
  - 扇区偏移量指的是在某个磁头的某个柱面上的扇区位置偏移量，从1开始编号。
- 例如，假设某个磁盘有16384个柱面，16个磁头，每个磁头有63个扇区，每个扇区大小为512字节，要计算某个扇区的物理位置，其CHS地址为（1234, 5, 6），则计算过程如下：
  - 物理扇区号 = ((1234 × 16) + 5) × 63 + 6 - 1 = 1244192
  - 物理位置 = 1244192 × 512 = 637026304

- 一个磁盘有4个磁片，每个磁片划分为1024个磁道，每个磁道分为256个扇区。每个扇区容量512B，这个磁盘容量为（）

磁盘访问时间

- 定位时间/随机访问时间：
  - **寻道时间**：移动磁臂到所需磁道时间
    - 平均寻道时间：1/3 磁道移动（1-4ms)
  - **旋转延迟**：等待扇区移动到磁头下时间
    - 由磁盘的旋转速度决定
    - 磁盘旋转速度：60 – 250转/秒
    - RPM（Revolutions Per Minute）：每分钟旋转次数，如：7200RPM，即120转/秒
    - 平均旋转1/2圈时间：1/(2*RPM/60)
    - 平均延迟时间：1/(2*7200/60)=4.17毫秒

Spindle [rpm]	Average latency [ms]
4200	7.14
5400	5.56
7200	4.17
10000	3
15000	2

常用的RPM对应的平均延迟时间

- 传输时间
  - 传输的数据量乘以传输率
  - 传输率：传输总字节数除以传输时间
  - 例如： 6 Gb/sec➔1秒可以传输6G位数据

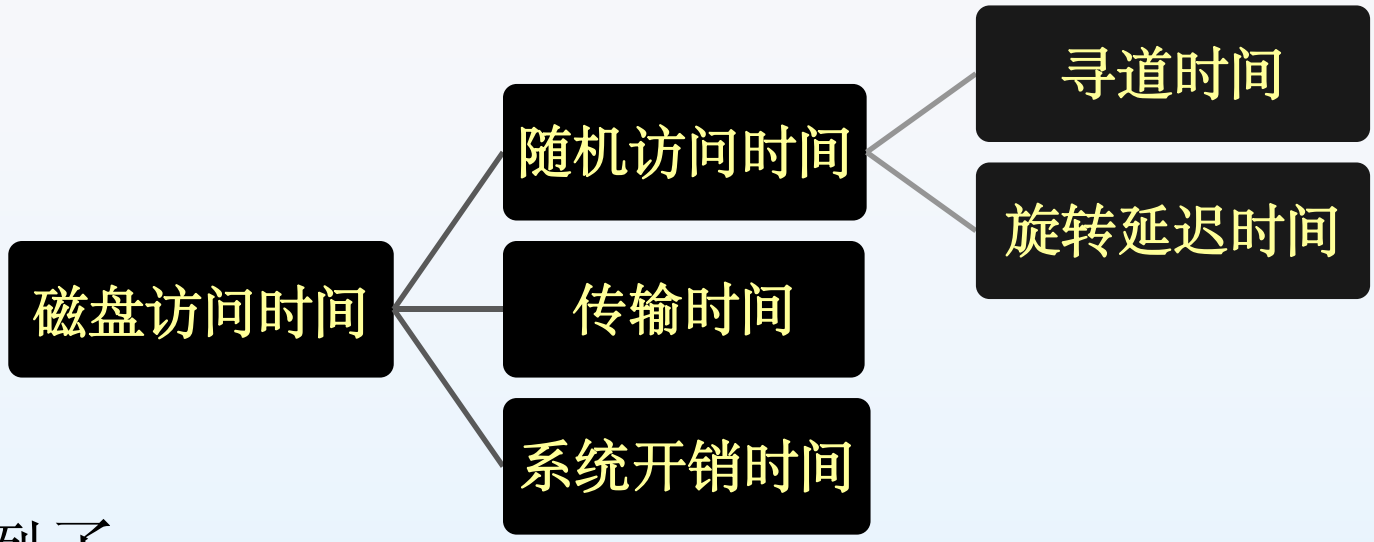
传输1KB数据的传输时间：

1K*8/6G=7.5微秒

磁盘访问时间

基本参数/ESSENTIAL PARAMETER	
型号：[redacted]	容量：6 TB
接口：SATA 6 Gb/s	规格：3.5 英寸
转速：5400 PPM	缓存：64 MB
尺寸（mm）：高*长*宽：26.1*147*101.6	
适用系统：台式机 / 一体机电脑	

磁盘访问时间



- 例子
  - 7200 RPM 转速， 5ms平均寻道时间， 1Gb/sec传输率，0.1ms 系统控制开销。那么读取4KB数据块的磁盘访问时间为：

5ms + ½*1/(7200/60)sec + 4KB / 1Gb/sec + 0.1ms

=5ms + 4.17ms + 0.03ms + 0.1ms

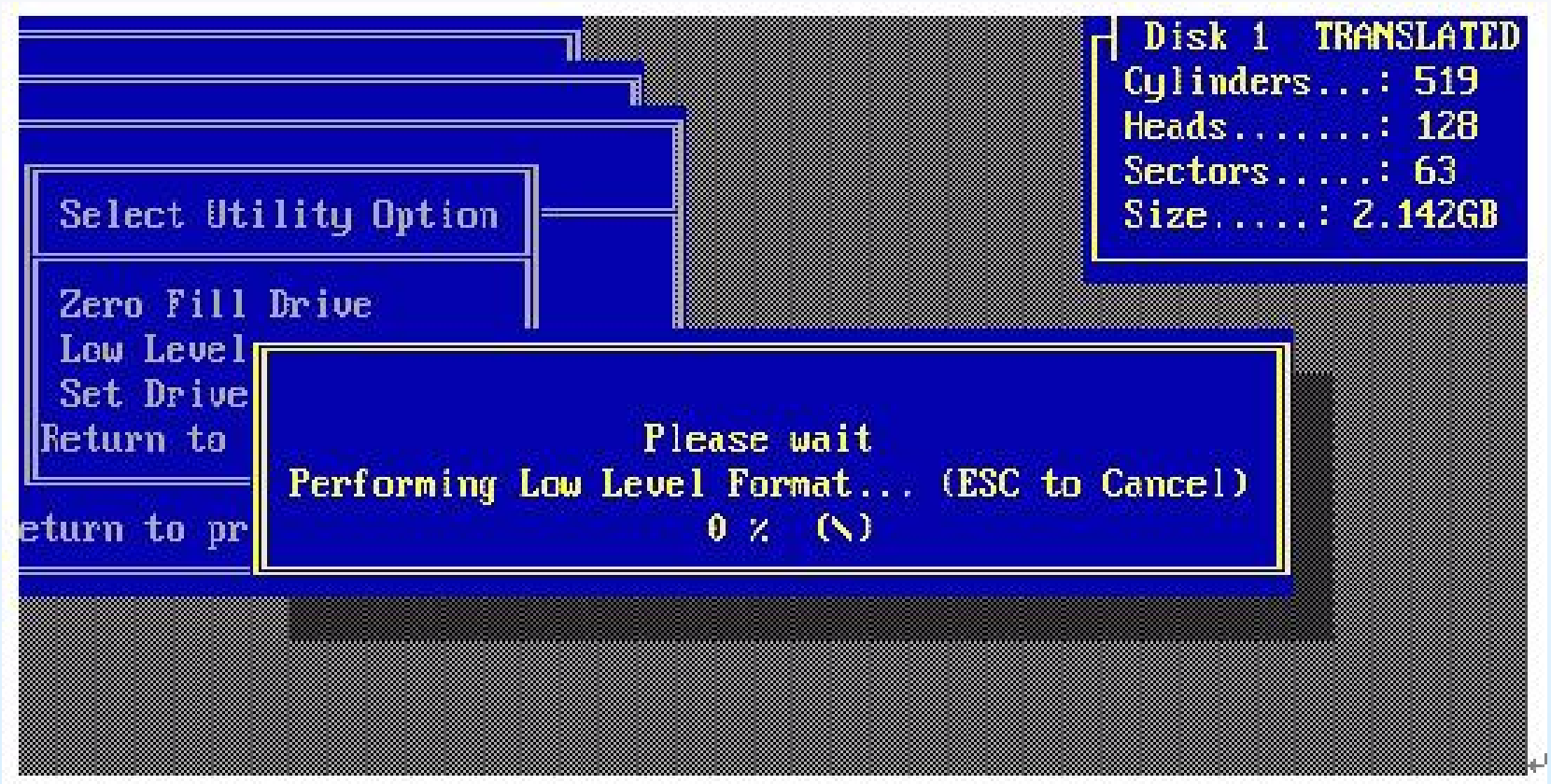
=9.3ms

- 一个磁盘的传输率为2Gb/s，那么传输1MB数据需要的传输时间是（）

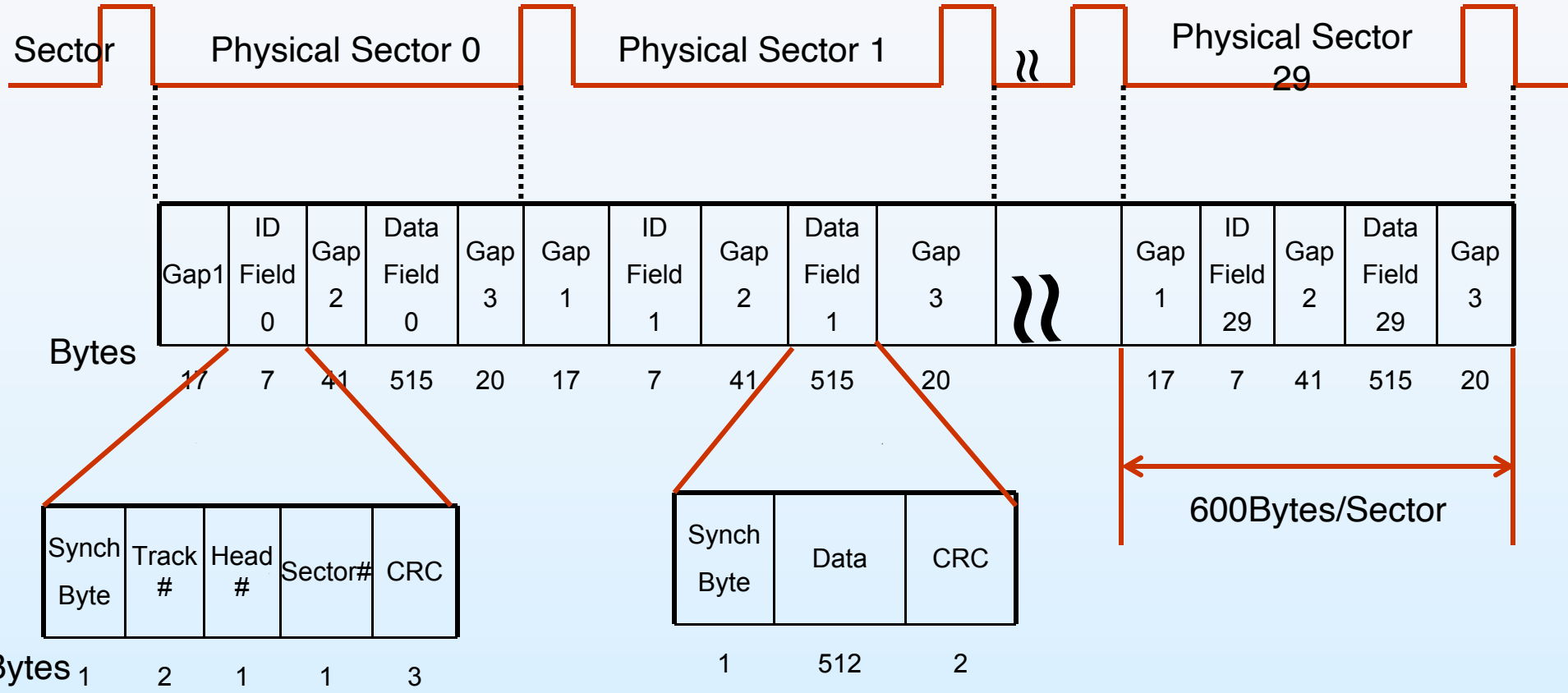
磁盘管理

- 低级格式化（物理格式化）
  - 将磁盘分成扇区，以便磁盘控制器读写
- 分区
  - 将磁盘分成分区
  - 主分区和扩展分区
- 高级格式化
  - 逻辑格式化，创建文件系统
- 引导块
  - 自举程序保存在ROM中
  - 自举程序装载引导块程序

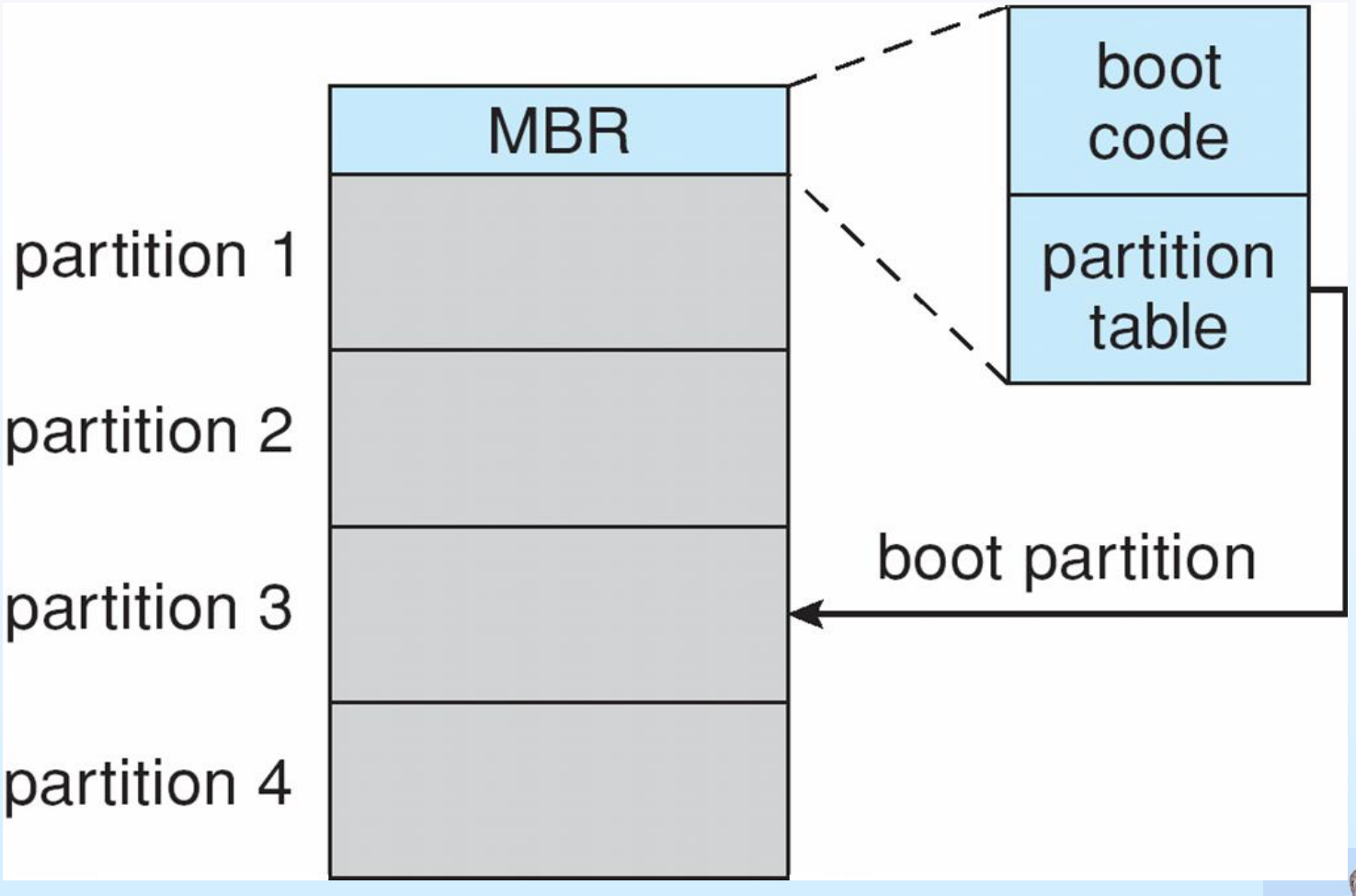
低级格式化


Disk 1 TRANSLATED
Cylinders...: 519
Heads.....: 128
Sectors.....: 63
Size.....: 2.142GB
Select Utility Option
Zero Fill Drive
Low Level
Set Drive
Return to
eturn to pr
Please wait
Performing Low Level Format... (ESC to Cancel)
0 % (\)

扇区格式


Sector
Physical Sector 0
Physical Sector 1
Physical Sector 29
Gap1
ID Field 0
Gap 2
Data Field 0
Gap 3
Gap 1
ID Field 1
Gap 2
Data Field 1
Gap 3
Gap 1
ID Field 29
Gap 2
Data Field 29
Gap 3
Bytes
17
7
41
515
20
600Bytes/Sector
Synch Byte
Track #
Head #
Sector#
CRC
Bytes
1
2
1
1
3
Synch Byte
Data
CRC
1
512
2

分区（Windows）


MBR
boot code
partition table
partition 1
partition 2
partition 3
partition 4
boot partition

引导区记录（MBR）

标准 MBR 结构					
地址			**描述**		**长度**
Hex	Oct	Dec			（字节）
0	0	0	代码区		**440** (最大 446)
01B8	670	440	选用软盘标志		4
01BC	674	444	一般为空值; 0x0000		2
01BE	676	446	**标准 MBR 分区表规划** (四个16 byte的主分区表入口)		**64**
01FE	776	510	55h	MBR 有效标志	**2**
01FF	777	511	AAh		
MBR, 总大小: 446 + 64 + 2 =					**512**

高级格式化

卷	布局	类型	文件系统	状态	容量	可用空间	% 可用	容错	开
(C:)	简单	基本	NTFS	状态良好 (启动, 页面文件, 故障转储, 主分区)	74.43 GB	7.22 GB	10 %	否	0
D3P_SCN (G:)	简单	基本	UDF	状态良好 (主分区)	4.26 GB	0 MB	0 %	否	0
Data (E:)	简单	基本	NTFS	状态良好 (主分区)	800.00 GB	249.75 GB	31 %	否	0
NLP (F:)	简单	基本	NTFS	状态良好 (主分区)	563.01 GB	439.10 GB	78 %	否	0
SDBACKUP (J:)	简单	基本	FAT32	状态良好 (主分区)	14.91 GB	3.40 GB	23 %	否	0
software (D:)	简单	基本	NTFS	状态良好 (主分区)	500.00 GB	456.27 GB	91 %	否	0
系统保留	简单	基本	NTFS	状态良好 (系统, 活动, 主分区)	100 MB	71 MB	71 %	否	0

磁盘 0
基本
74.53 GB
联机

系统保留
100 MB NTFS
状态良好 (系统, 活动,

(C:)
74.43 GB NTFS
状态良好 (启动, 页面文件, 故障转储, 主分区)

磁盘 1
基本
1863.02 GB
联机

software (D:)
500.00 GB NTFS
状态良好 (主分区)

Data (E:)
800.00 GB NTFS

NLP (F:)
563.01 GB NTFS
状态良好 (主分区)

磁盘 2
可移动 (I:)

无媒体

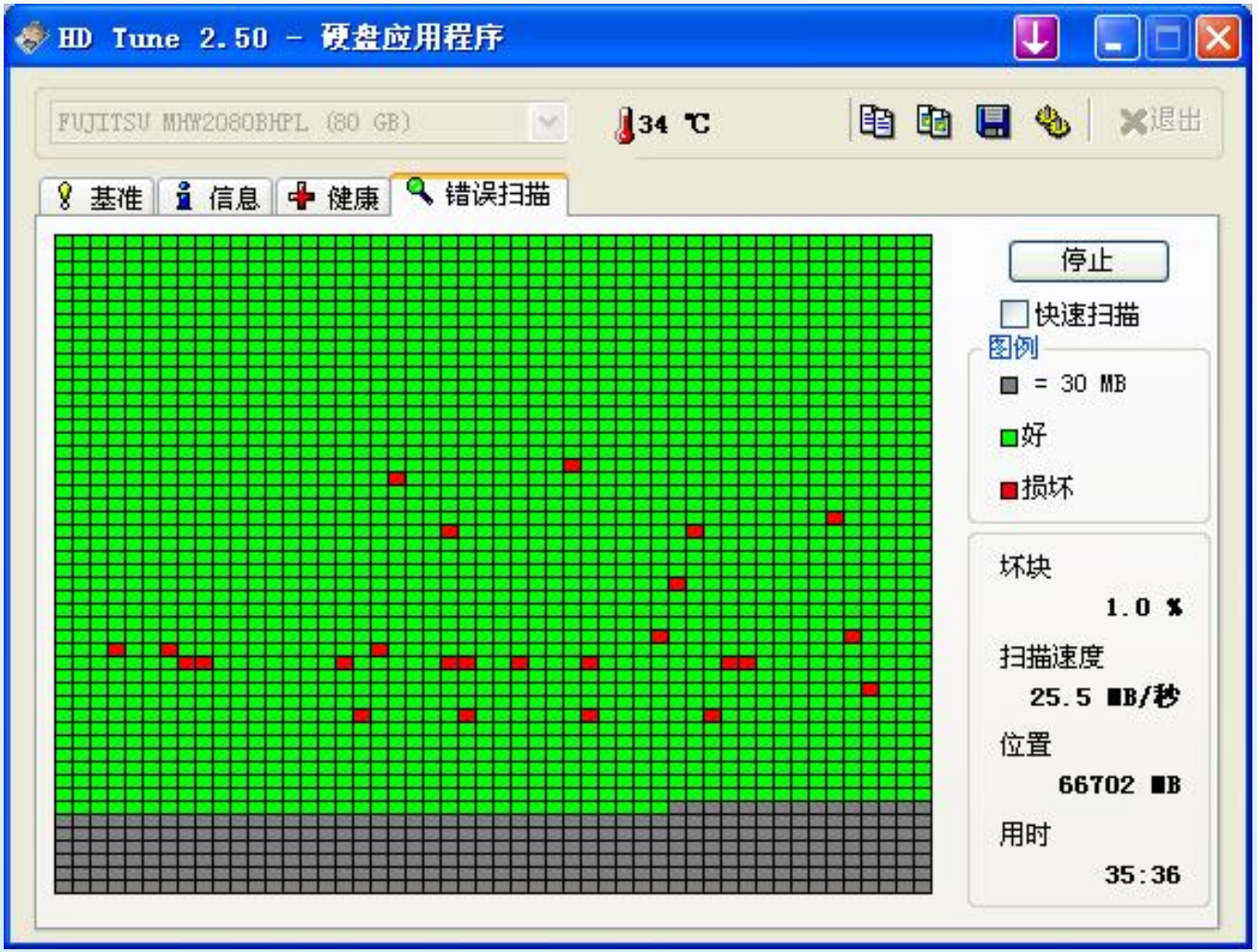
磁盘 3
可移动
14.93 GB
联机

SDBACKUP (J:)
14.92 GB FAT32

MB

打开(O)
资源管理器(E)
将分区标记为活动分区(M)
更改驱动器号和路径(C)...
格式化(F)...
扩展卷(X)...
压缩卷(H)...
添加镜像(A)...
删除卷(D)...
属性(P)

坏块检查


HD Tune 2.50 - 硬盘应用程序
FUJITSU MHW2080BHPL (80 GB)
34 ℃
退出
基准
信息
健康
错误扫描
停止
快速扫描
图例
= 30 MB
好
损坏
坏块
1.0 %
扫描速度
25.5 MB/秒
位置
66702 MB
用时
35:36

2、磁盘调度

内容

- 磁盘调度
  - 引入磁盘调度的目的是为了降低磁盘访问时间，提高文件系统的效率
- 先来先服务算法
- 最短寻道时间优先算法
- 扫描算法
- RAID
  - 引入RAID技术的目的是为了增强数据的可靠性和访问的并行性

磁盘调度

- 目标：减少磁盘访问时间
- 访问时间：
  - **寻道时间**：磁头移动到访问扇区所在磁道的时间
  - **旋转延迟时间**：将访问扇区转到磁头下的时间
  - **传输时间**：将数据从磁盘送到内存的时间
- **寻道时间最小化**
- 寻道时间 ≈ 寻道距离

请求系列

- 假定有一个请求序列(0-199道).:

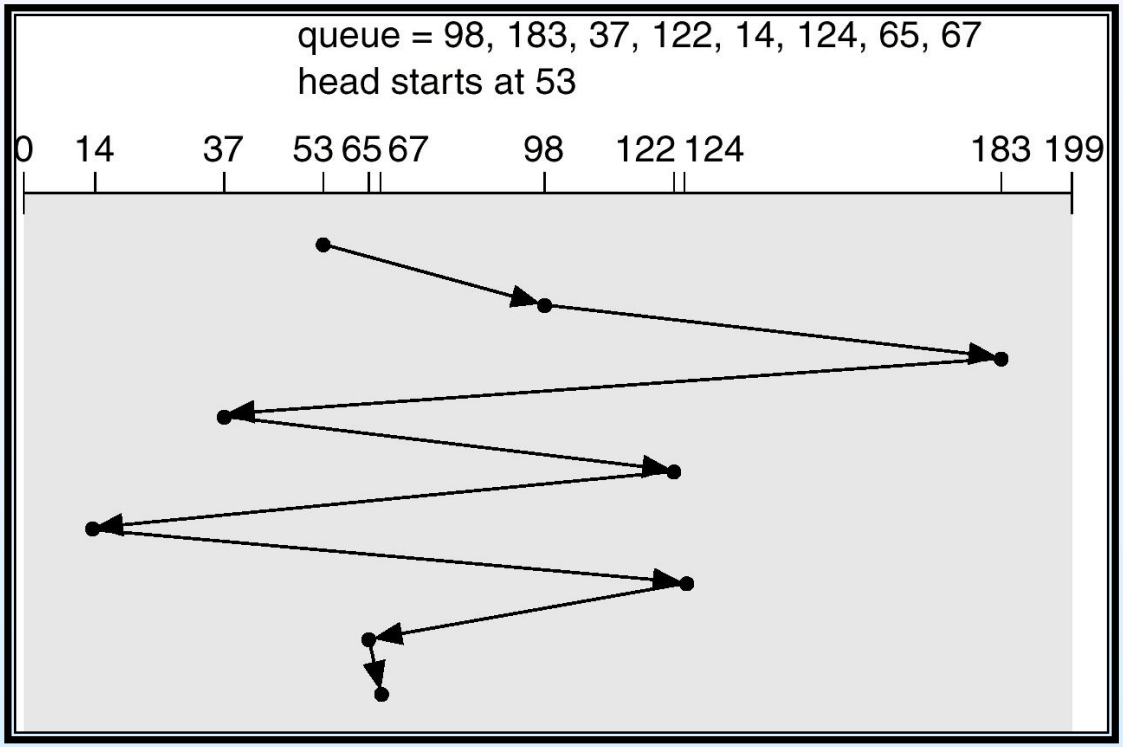
  98, 183, 37, 122, 14, 124, 65, 67

  磁头当前位置在53

目标：磁头移动距离最小，寻道时间最短

先来先服务算法FCFS

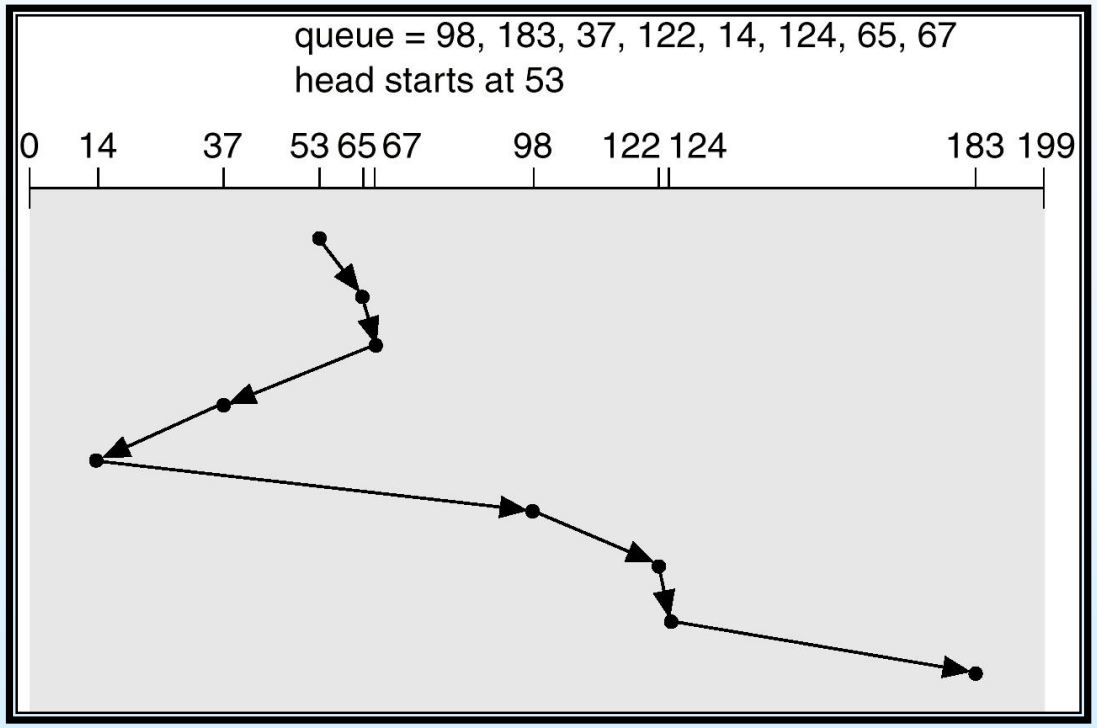
- First Come First Served
- 按照请求提交时间访问
  - 先提交先访问
  - 后提交后访问
- 优点
  - 简单、公平
  - 易实现
- 缺点
  - 寻道时间长
  - 效率低



总的磁头移动为640磁道

最短寻道时间优先算法SSTF

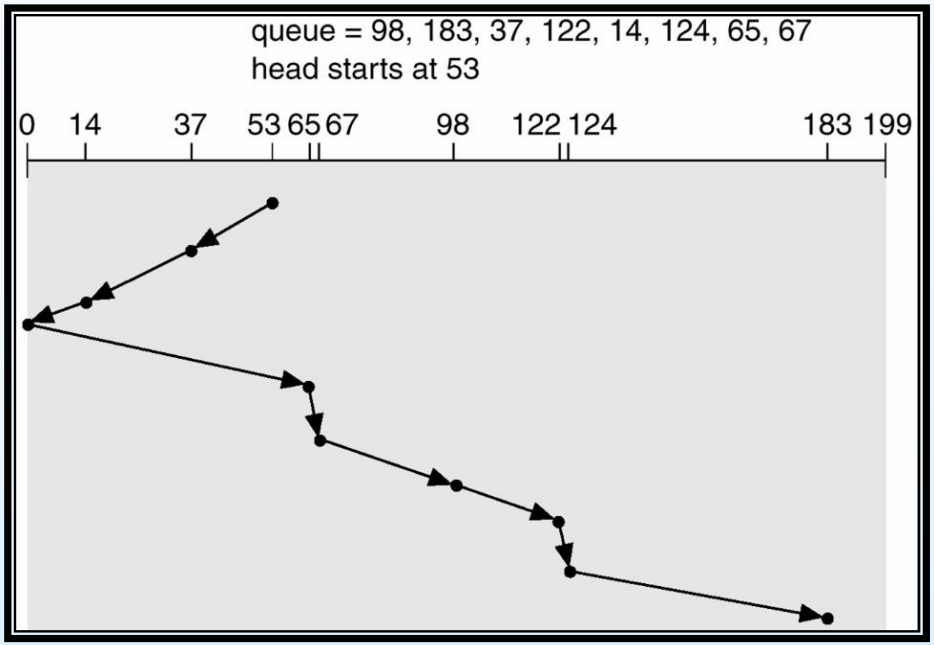
- Shortest Seek Time First
- 每次移动到离现在位置最近的磁道
  - 最短寻道时间
  - 最短作业优先（SJF）
- 优点
  - 寻道距离短
- 缺点
  - 存在饥饿
  - 磁头频繁变换移动方向
  - 增加寻道时间



总的磁头移动为208磁道

扫描算法SCAN

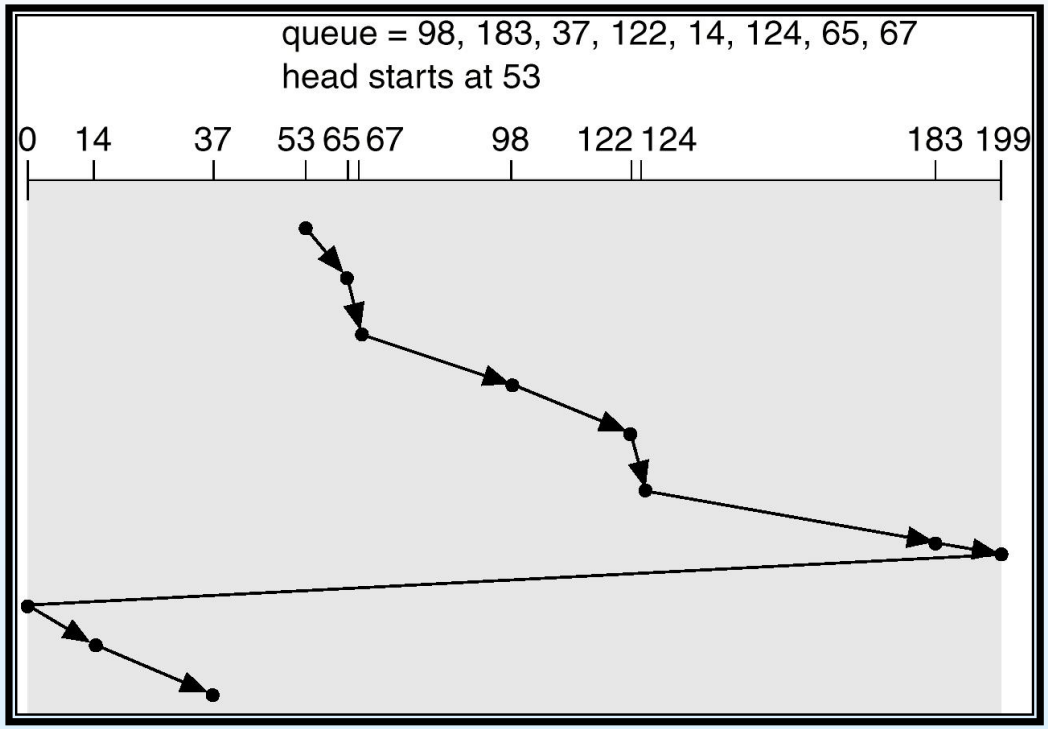
- 磁头从磁盘一端向另一段移动，沿途响应服务请求
  - 到达另一端时，磁头改变移动方向，继续处理
  - 磁头在磁盘上来回扫描
  - 又称为电梯算法
- 优点
  - 同一方向扫描，寻道时间短
  - 改变磁头方向少
- 缺点
  - 有的请求等待时间长



总的磁头移动为236磁道

循环扫描算法C-SCAN

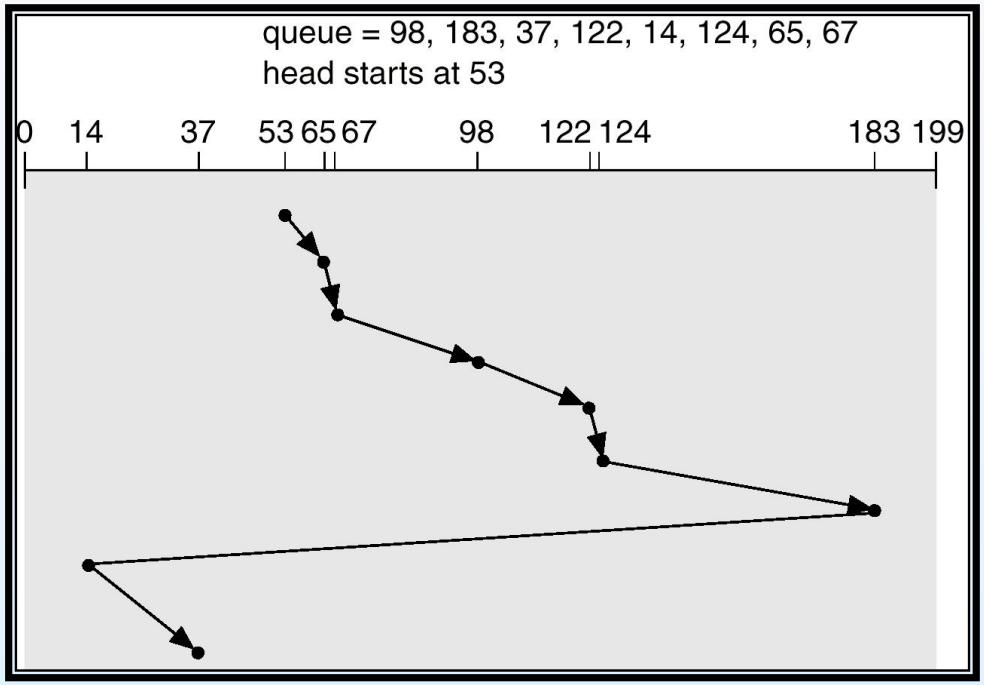
- Circular Scan
- 单向处理请求
  - 磁头从磁盘外道（0道）移到内道过程中处理请求
  - 内道移动到外道的过程中不处理请求
- 优点
  - 更均匀的等待时间
- **从磁道199移动到0的时间很短**



总的磁头移动为382磁道

循环Look算法C-LOOK

- C-SCAN变形
- 磁头只移动到一个方向上最远请求为止，而不是继续到磁盘尽头



总的磁头移动为322磁道

磁盘调度算法的选择

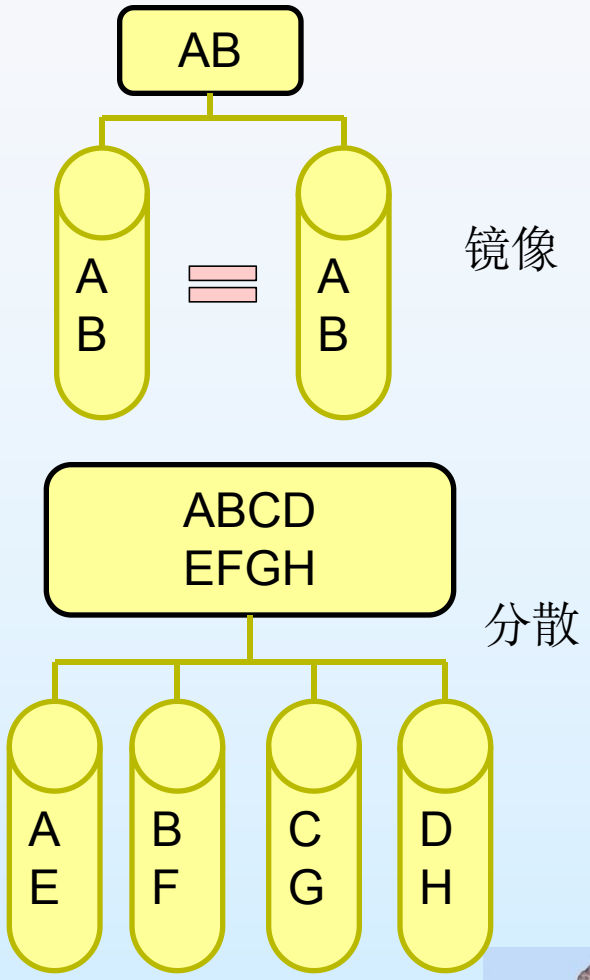
- 磁盘调度性能主要依赖于请求的数量和类型
  - 磁盘服务请求很大程度上受文件分配方法影响，例如隐式链接的服务请求数就会比较多
  - SSTF较为普遍且很有吸引力
  - SCAN和C-SCAN适合磁盘大负荷系统
- SSTF或LOOK是比较合理的缺省算法

RAID结构

- RAID
  - *Originally* Redundant Arrays of Inexpensive Disks（廉价磁盘冗余阵列）
  - *Now* Redundant Arrays of Independent Disks（独立磁盘冗余阵列）
  - RAID把很多价格较便宜的磁盘组合成一个大容量的磁盘组，利用个别磁盘提供数据所产生加成效果提升整个磁盘系统效能和可靠性。
- RAID卡（现代CPU集成RAID）
- RAID被分成了多个不同级别
  - RAID0-RAID7
  - RAID01，RAID10，RAID5E，RAID5EE，RAID50

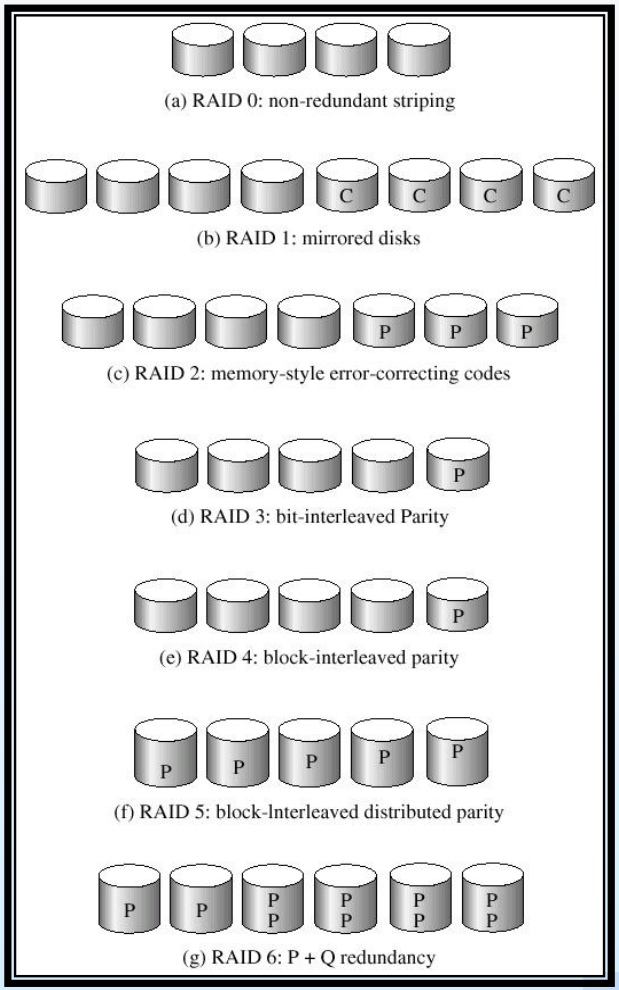
RAID性能

- 可靠性
  - 磁盘可靠性：要求存储在磁盘上的数据不易丢失
  - 引入冗余
  - 例如：镜像，把数据在两个磁盘上各存一次
- 性能（数据分散，并行读写）
  - 位级分散：数据每个字节的各个位分散在多个磁盘上
  - 块级分散：数据以块为单位分散在多个磁盘上



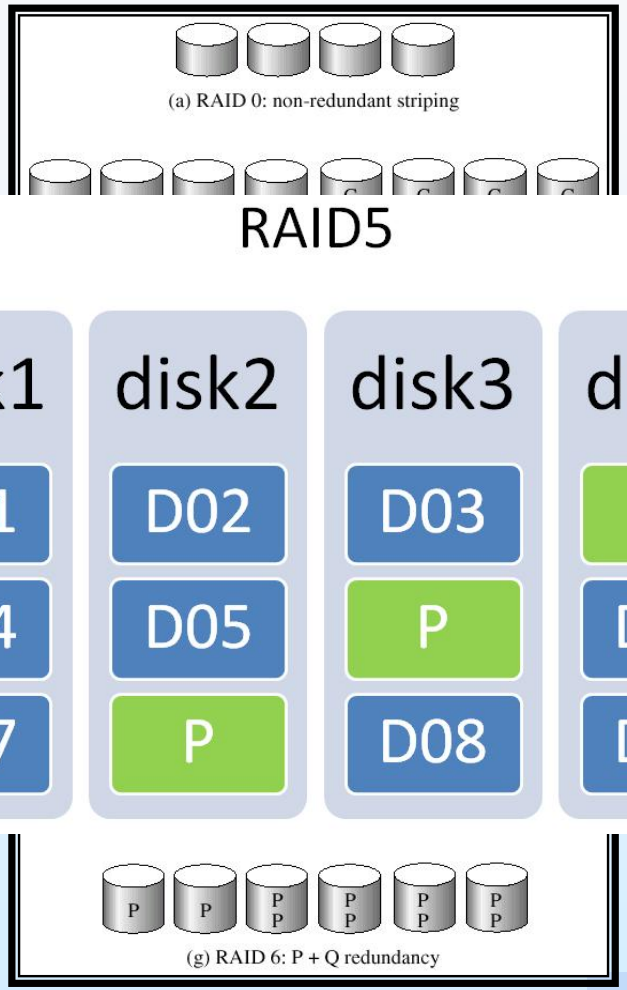
RAID级别

- RAID 0
  - 数据分散在多个磁盘上
  - 条状分散技术
  - 提高读写性能
- RAID 1
  - 磁盘镜像
  - 提高可靠性
- RAID 5
  - 分散+校验
  - 校验信息分散在各个磁盘避免对单个校验磁盘的过度使用



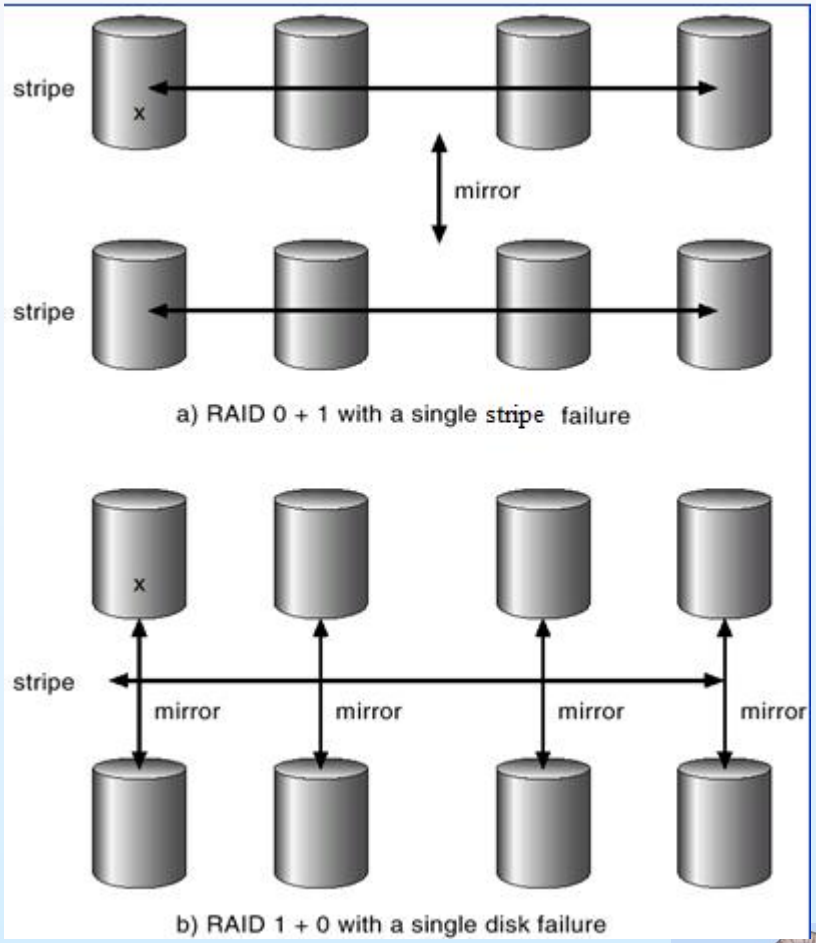
RAID级别

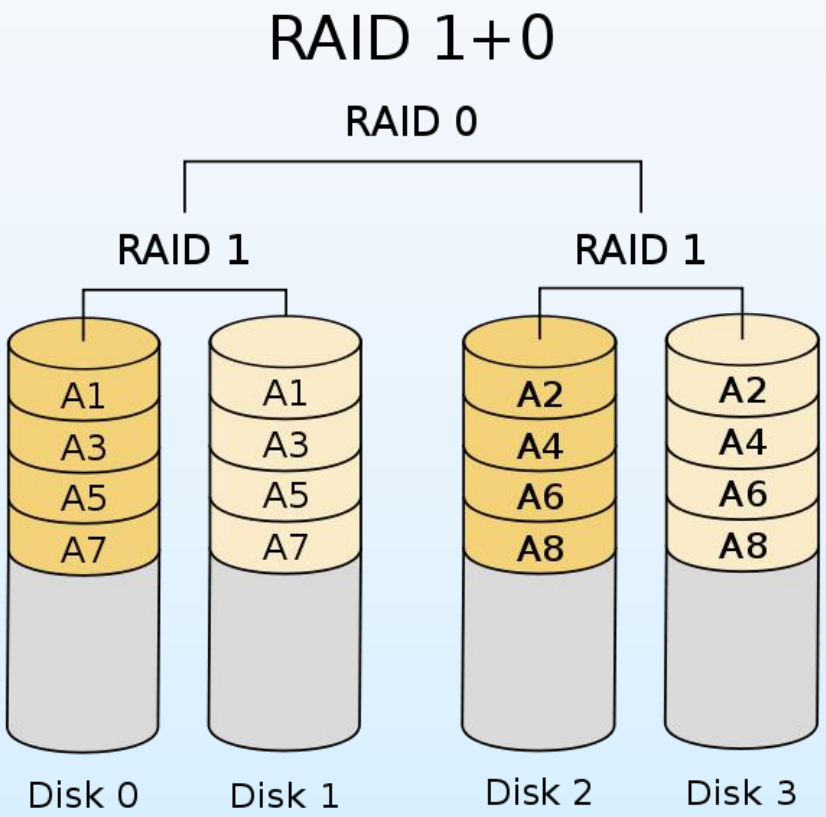
- RAID 0
  - 数据分散在多个磁盘上
  - 条状分散技术
  - 提高读写性能
- RAID 1
  - 磁盘镜像
  - 提高可靠性
- RAID 5
  - 分散+校验
  - 校验信息分散在各个磁盘避免对单个校验磁盘的过度使用

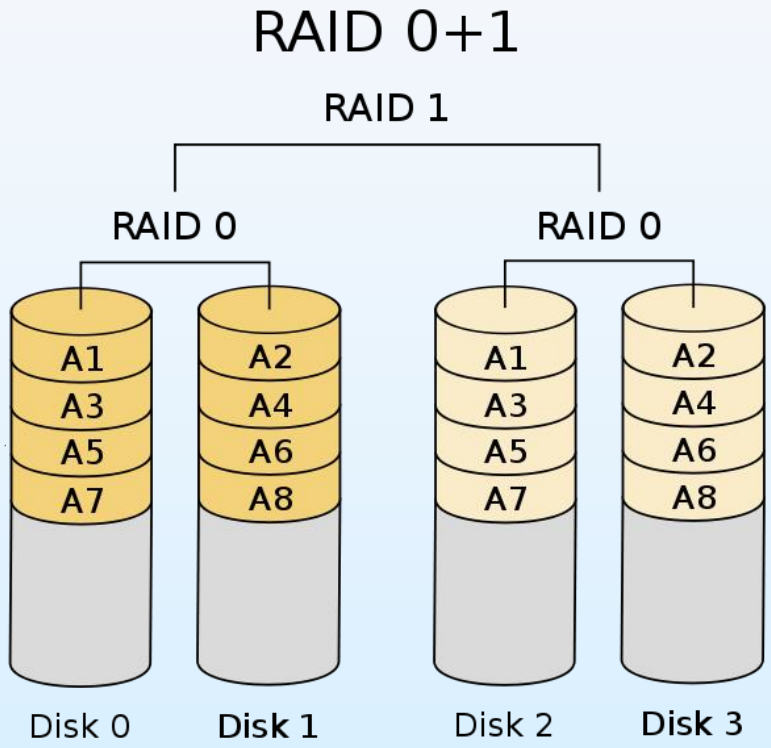


RAID (0 + 1) 和 (1 + 0)

- RAID0：性能
- RAID1：可靠性
- 二者兼备
- RAID01
  - 先做分散，再做镜像
  - 性能好
  - 但是一个磁盘的故障会导致一条磁盘带不能访问
- RAID10
  - 先做镜像，再做分散
  - 可靠性好，一个磁盘的故障不会影响其他磁盘




RAID 1+0
RAID 0
RAID 1
RAID 1
A1
A3
A5
A7
A1
A3
A5
A7
A2
A4
A6
A8
A2
A4
A6
A8
Disk 0
Disk 1
Disk 2
Disk 3


RAID 0+1
RAID 1
RAID 0
RAID 0
A1
A3
A5
A7
A2
A4
A6
A8
A1
A3
A5
A7
A2
A4
A6
A8
Disk 0
Disk 1
Disk 2
Disk 3